Numéro d'ordre : 138-2005 Année 2005

T H E S E

présentée

devant l'UNIVERSITE CLAUDE BERNARD - LYON 1
UFR de Chimie - Biochimie

pour l'obtention

du **DIPLOME DE DOCTORAT**
(arrêté du 25 avril 2002)

présentée et soutenue publiquement le 6 octobre 2005

par

Monsieur Sébastien VIOLOT

ETUDES FONCTIONNELLES ET STRUCTURALES DE LA PROTEINE EED, PARTENAIRE CELLULAIRE DU VIRUS VIH-1 ET DE LA CELLULASE « FROIDE » CEL5G DE *PSEUDOALTEROMONAS HALOPLANKTIS*

Directeur de thèse : Monsieur le Docteur R. HASER

JURY :

Monsieur	le Professeur P. BOULANGER,	Président du jury
Monsieur	le Professeur C. GERDAY,	Rapporteur
Monsieur	le Docteur J.F. MOUSCADET,	Rapporteur
Monsieur	le Docteur F. REY,	Examinateur
Monsieur	le Docteur R. HASER,	Directeur de thèse
Monsieur	le Professeur P. GOUET,	Responsable scientifique

UNIVERSITE CLAUDE BERNARD - LYON I

Président de l'Université	**M. le Professeur D. DEBOUZIE**
Vice-Président du Conseil Scientifique	M. le Professeur J.F. MORNEX
Vice-Président du Conseil d'Administration	M. le Professeur R. GARRONE
Vice-Présidente du Conseil des Etudes et de la Vie Universitaire	M. le Professeur G. ANNAT
Secrétaire Général	M. J.P. BONHOTAL

SECTEUR SANTE

Composantes

UFR de Médecine Lyon R.T.H. Laënnec	Directeur : M. le Professeur D. VITAL-DURAND
UFR de Médecine Lyon Grange-Blanche	Directeur : M. le Professeur X. MARTIN
UFR de Médecine Lyon-Nord	Directeur : M. le Professeur F. MAUGUIERE
UFR de Médecine Lyon-Sud	Directeur : M. le Professeur F.N. GILLY
UFR d'Odontologie	Directeur : M. O. ROBIN
Institut des Sciences Pharmaceutiques et Biologiques	Directeur : M. le Professeur F. LOCHER
Institut Techniques de Réadaptation	Directeur : M. le Professeur L. COLLET
Département de Formation et Centre de Recherche en Biologie Humaine	Directeur : M. le Professeur P. FARGE
Département de Production et Réalisation Assistance Conseil en Technologie pour l'Education	Directrice : Mme. le Professeur M. HEYDE

SECTEUR SCIENCES

Composantes

UFR de Physique	Directeur : M. le Professeur A. HOAREAU
UFR de Biologie	Directeur : M. le Professeur H. PINON
UFR de Mécanique	Directeur : M. le Professeur H. BEN HADID
UFR de Génie Electrique et des Procédés	Directeur : M. le Professeur A. BRIGUET
UFR Sciences de la Terre	Directeur : M. le Professeur P. HANTZPERGUE
UFR de Mathématiques	Directeur : M. le Professeur M. CHAMARIE
UFR d'Informatique	Directeur : M. le Professeur M. EGEA
UFR de Chimie Biochimie	Directeur : M. le Professeur J.P. SCHARFF
UFR STAPS	Directeur : M. le Professeur R. MASSARELLI
Observatoire de Lyon	Directeur : M. le Professeur R. BACON
Institut des Sciences et des Techniques de l'Ingénieur de Lyon	Directeur : M. le Professeur J. LIETO
IUT A	Directeur : M. le Professeur M. C. COULET
IUT B	Directeur : M. le Professeur R. LAMARTINE
Institut de Science Financière et d'Assurances	Directeur : M. le Professeur J.C. AUGROS

Remerciements

Le travail présenté dans ce mémoire de thèse a été réalisé au sein de l'Institut de Biologie et Chimie des Protéines (UMR 5086, CNRS-UCB LYON I) dans le laboratoire de BioCristallographie.
Je tiens ainsi à remercier Monsieur le Professeur Alain Jean COZZONE pour m'avoir accueilli dans son laboratoire.

Je suis très reconnaissant à Monsieur Pierre BOULANGER, Professeur à l'Université Claude Bernard Lyon 1, pour avoir accepté de présider ce jury.
J'exprime toute ma gratitude à Monsieur Charles GERDAY, Professeur à l'Université de Liège, et à Monsieur Jean-François MOUSCADET, Directeur de Recherche au CNRS, pour avoir assuré la tâche de rapporteurs, ainsi qu'à Monsieur Félix Rey, Directeur de Recherche au CNRS, pour avoir assurer la tache d'examinateur.
J'adresse mes sincères remerciements à Monsieur Richard HASER, Directeur de Recherche au CNRS, pour avoir dirigé cette thèse, et à Monsieur Patrice GOUET, Professeur à l'Université Claude Bernard Lyon 1, pour l'avoir co-encadré.

Au terme de ce travail de recherche, je tiens à remercier l'ANRS pour leur financement de ma bourse de thèse.

Je tiens à associer à ces remerciements l'ensemble des membres du laboratoire de BioCristallographie, des doctorants, secrétaires et personnels pour l'ambiance chaleureuse de travail et pour nos échanges qui n'ont pas été que scientifiques.
Egalement un grand merci à tout le laboratoire de Virologie et Pathogenèse Virale pour leur accueil, leur collaboration et leur soutien.

Je ne saurais oublier de remercier toutes les personnes qui me sont chères, en particulier mes parents, mon «petit frère» Guillaume et ma belle famille qui m'ont toujours encouragé, pour l'aide, la confiance et le soutien dont ils ont fait preuve tout au long de ces dernières années.

Enfin, j'exprime ma dernière pensée à Stéphanie qui m'a soutenu et encouragé et à qui je dédicace ce travail.
Stéphanie, pour la confiance, le soutien et l'Amour que tu m'as témoignés ces dernières années.

A mon grand-père André et à ma grand-mère Marie,

A mon grand-père Robert,

A mes parents et mon petit frère,

A Stéphanie & Charlotte.

Contentons nous de faire réfléchir,
n'essayons pas de convaincre.

Georges Braque

Ma vie est aimantée par ce que je ne connais pas.
Si je savais où mes recherches doivent me mener,
je ne les entreprendrais pas !

Pierre Boulez

Table des matières

Abréviations et unités utilisées :

Å	:	Angström (1 Å ↔ 10^{-10} mètres)
ADN	:	Acide désoxyribonucléique
ADP	:	Adénosine diphosphate
AFMB	:	Laboratoire Architecture et Fonction des Macromolécules Biologiques
ALSV	:	*Avian Leukaemia and Sarcoma Viruses*
ANRS	:	Agence Nationale de Recherches sur le Sida
ARN	:	Acide ribonucléique
ATP	:	Adénosine triphosphate
BSA	:	*Bovine serum albumin* (albumine du sérum de bœuf)
°C	:	Degré Celsius
C-	:	Carboxy-
CBM	:	*Carbohydrates Binding Modules*
CCD	:	*Charge Coupled Device*
CCP4	:	*Collaborative Computational Project number 4*
CNRS	:	Centre National de la Recherche Scientifique
CNS	:	Logiciel «*Crystallographic & NMR Systems*»
Da	:	Dalton (kDa, Da)
DEA	:	Diplôme d'Etudes Approfondies
dNTP	:	Deoxynucleotide triphosphates
DO	:	Densité optique
DP(n)	:	Degré de polymérisation (n)
DTT	:	Dithiothréitol
E. coli	:	*Escherichia coli*
EDTA	:	Acide diamine-éthylène-tétraacétique
EED	:	*Embryonic ectoderm development*
ESRF	:	Synchrotron «*European Synchrotron Radiation Facility»*
EtOH	:	Ethanol
eV	:	Electron-volt (eV, Gev)
FIP (ligne)	:	*French beamline for Investigation of Proteins* (ESRF)
g	:	Accélération ($m.s^{-2}$)
g	:	Gramme (kg, g, mg, µg, ng)
GST	:	Glutathion-S-Transférase
h	:	Heure
HCl	:	Acide chlorhydrique
HEPES	:	Acide [(hydroxy-2-éthyl)-4-pipérazinyl-1]-2-éthane sulfonique
HTLV	:	Human T Leukemia Virus
IN	:	Intégrase
INRA	:	Institut National de la Recherche Agronomique
IPTG	:	Isopropyl thiogalactoside
K	:	Kelvin
kb	:	kilobase
L	:	Litre (L, mL, µL, nL)
LTR	:	*Long terminal repeat*
m	:	Mètre (m, mm, µm, nm)
M	:	Molaire (M, mM, µM, nM)
MA	:	Matrice
MAD	:	Méthode dite «*Multiple Anomalous Diffraction»*
MAR345	:	Détecteur MARresearch «*Image Plate*» de 345 mm de diamètre
MeOH	:	Méthanol
MES	:	Acide 2-(N-Morpholino) éthanesulfonique

min	:	minute
MIR	:	Méthode dite «*Multiple Isomorphous Replacement*»
MLV	:	*Murine Leukaemia Virus* (virus de la leucémie murine)
MM	:	Masse Moléculaire
MOPS	:	Acide 3-(N-Morpholino) propanesulfonique
MPD	:	2-methyl-2,4-pentanediol
N-	:	Amino-
ORF	:	*Open reading frame* (phase ouverte de lecture)
Pa	:	Pascal (Pa, MPa)
pb	:	Paire de bases
PBS	:	*Phosphate Buffer Saline*
PcG	:	*Polycomb Group*
PCR	:	*Polymerase Chain Reaction* (réaction de polymérisation en chaîne)
PEG	:	Polyéthylène glycol
PDB	:	Banque de structure «*Protein Data Bank*»
PIC	:	*Preintegration Complex* (complexe de pré-intégration)
PM	:	Poids moléculaire
p / v	:	poids / volume
RCSB	:	*Research Collaboratory for Structural Bioinformatics*
R_{libre} (R_{free})	:	Facteur d'accord R libre
RMN	:	Résonance Magnétique Nucléaire
r.m.s.d	:	*Root mean square deviation*
rpm	:	Rotation par minute
s	:	seconde (s, ms)
SAXS	:	*Small Angle X-ray Scattering* (diffraction aux rayons X aux petits angles)
SDS	:	Dodécylsulfate de sodium
ThApu	:	*Thermococcus hydrothermalis* Amylo-pullulanase
Tris	:	2-amino-2-(hydroxyméthyl)-1,3 propanediol
TrxG	:	Trithorax Group
UCBL	:	Université claude Bernard Lyon 1
UMR	:	Unité mixte de recherche
VIH-1	:	Virus de l'immunodéficience humaine de type 1
v / v	:	volume / volume

Préambule :

Ce travail de thèse, financé par une bourse de l'Agence Nationale de Recherches sur le Sida (ANRS) est le résultat d'une collaboration entre le laboratoire de BioCristallographie dirigé par le Dr R. Haser à l'Institut de Biologie et de Chimie des Protéines de Lyon (UMR 5086, CNRS-UCBL) et le laboratoire de Virologie et de Pathogenèse Virale (UMR 5537, CNRS-UCBL) dirigé par le Pr P. Boulanger. Il porte principalement sur la détermination de la structure cristallographique de la protéine humaine EED (Embryonic Ectoderm Development). Cette protéine semble jouer un rôle important au cours du cycle viral du virus VIH-1, en se liant avec sa protéine de Matrice (MA) et son Intégrase (IN). Cette étude structurale de EED et de son implication au cours du cycle viral a été motivée par le besoin de développer des médicaments différents de ceux actuellement dirigés contre la transcriptase inverse et la protéase du VIH-1 (Inhibiteurs Nucléosidiques de la Transcriptase Inverse tels l'AZT ou zidovudine (RETROVIR®), le ddI ou didanosine (VIDEX®), le 3TC ou lamivudine (EPIVIR®) ou encore le ténofovir (VIREAD®) / Inhibiteurs Non Nucléosidiques de la Transcriptase Inverse tels la névirapine (VIRAMUNE®), la delavirdine (RESCRIPTOR®) et l'efavirenz (SUSTIVA®) / Inhibiteurs de la protéase virale tel le saquinavir (INVIRASE®)).
Les zones d'interaction entre EED et ses partenaires viraux ont donc été identifiées dans le cadre de cette recherche, afin d'envisager la synthèse d'inhibiteurs peptidomimétiques performants.
Au cours de ce travail sur EED, nous avons été contacté par le Dr C. Ronfort dirigeant l'équipe «Rétrovirus et Intégration Rétrovirale» du laboratoire Rétrovirus et Pathologie Comparée, dirigé par le Pr J.F. Mornex, (UMR 754, INRA-UCBL-ENVL) afin de caractériser les relations structure-fonction de l'Intégrase aviaire de ALSV (Avian Leukemia and Sarcoma Viruses). Cette équipe avait exprimé et caractérisé 11 mutants ponctuels du domaine catalytique et 8 mutants ponctuels du domaine C-terminal de l'Intégrase aviaire de ALSV, et voulait confronter leurs résultats expérimentaux (tests d'activité *in vitro*) à la modélisation de ces mutants. Ces études ont donné lieu à deux publications (Moreau *et al.*, 2004 ; Moreau *et al.*, 2003) présentées dans ce manuscrit (Publication 2 et Publication 3, respectivement).

J'ai également travaillé durant cette thèse à l'étude cristallographique de deux glycosides hydrolases adaptées aux températures extrêmes : (i) l'amylo-pullulanase ThApu de l'organisme hyperthermophile *Thermococcus hydrothermalis* et (ii) l'endoglucanase Cel5G de l'organisme psychrophile *Pseudoalteromonas haloplanktis* en collaboration avec le laboratoire de Biochimie du Pr C. Gerday (Institut de Chimie B6, Université de Liège, B-

4000 Liège Sart-Tilman, Belgique) qui nous a fourni l'enzyme purifiée. Ce travail devait permettre de mieux comprendre les mécanismes moléculaires d'adaptation aux conditions extrêmes mis en œuvre par ces deux enzymes.

Des cristaux de ThApu ont été obtenus, mais ceux-ci étaient trop petits et n'ont jamais pu être optimisés afin de pouvoir être utilisés pour des expériences de diffraction aux rayons X. Ce sujet ne sera donc pas présenté dans le manuscrit de thèse.

Des cristaux du domaine catalytique de la «cellulase froide» Cel5G ont également été obtenus conduisant à la détermination de sa structure native à une résolution de 1,4 Å. Sa structure en complexe avec le cellobiose, produit de la réaction, a également été résolue à une résolution de 1,6 Å. Ce complexe met en évidence une fixation très spécifique du cellobiose dans les sous sites -2 et -3 du site catalytique. La comparaison structurale des domaines catalytiques de Cel5G et de Cel5A révèle des déterminants structuraux de l'adaptation de cette enzyme aux basses températures. Des données de diffusion des rayons X aux petits angles ont été obtenues pour l'enzyme entière en collaboration avec le laboratoire AFMB (UMR 6098, CNRS et Universités d'Aix-Marseille I et II). Ces résultats sont présentés dans ce manuscrit sous la forme de deux articles (Publication 4 et Publication 5).

Liste des publications jointes au manuscrit

Publication 1 Violot, S., Hong, S.S., Rakotobe, D., Petit, C., Gay, B., Moreau, K., Billaud, G., Priet, S., Sire, J., Schwartz, O., Mouscadet, J.F. and P. Boulanger. (2003) The human Polycomb group EED protein interacts with the integrase of human immunodeficiency virus type 1. *J. Virol.,* **77**, 12507-12522.

Publication 2 Moreau K., Faure C., Violot S., Verdier G. and Ronfort C. (2003) Mutations in the C-terminal domain of ALSV (Avian Leukaemia and Sarcoma Viruses) integrase alter the concerted DNA integration process *in vitro*. *Eur. J. Biochem.,* **270**, 4426-4438.

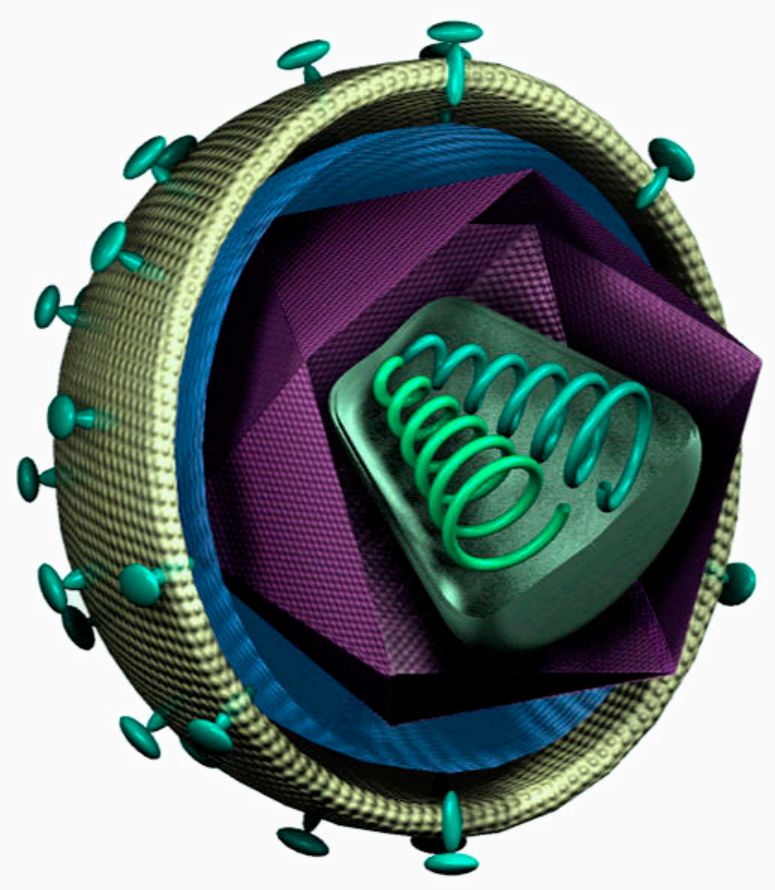
Publication 3 Moreau, K., Faure, C., Violot, S., Gouet, P., Verdier, G. and Ronfort, C. (2004) Mutational analyses of the core domain of Avian Leukaemia and Sarcoma Viruses integrase: critical residues for concerted integration and multimerization. *Virology,* **318**, 566-581.

Publication 4 Violot, S., Haser, R., Sonan, G., Georlette, D., Feller, G. and Aghajari, N. (2003) Expression, purification, crystallization and preliminary X-ray crystallographic studies of a psychrophilic cellulase from *Pseudoalteromonas haloplanktis*. *Acta Cryst.*, **D59**, 1256-1258.

Publication 5 Violot S., Aghajari N., Czjzek M., Feller G., Sonan G.K., Gouet P., Gerday C., Haser R., and Receveur-Brechot V. (2005) Structure of a full length psychrophilic cellulase from *Pseudoalteromonas haloplanktis* revealed by X-ray diffraction and small angle X-ray scattering. *J. Mol. Biol*., **348**, 1211-1224.

1ère Partie

Etude moléculaire et structurale de la protéine EED, partenaire cellulaire des protéines Matrice, Intégrase et Nef du virus VIH-1



Etude bibliographique

A/ LA PROTEINE CELLULAIRE HUMAINE EED

1. Caractérisation de EED :

La mise en place des profils d'expression génétique en fonction du type cellulaire implique l'activation ou la répression de l'expression de gènes codant des promoteurs, des activateurs et des facteurs de transcription. De nombreux niveaux de contrôle sont nécessaires pour le maintien de cette expression ou répression génique au fil des divisions cellulaires. Chez les eucaryotes, plusieurs complexes protéiques associés à la chromatine et impliqués dans la maintenance de la différenciation cellulaire ont été identifiés, dont le groupe des Polycomb (PcG). Les PcG sont des répresseurs de la transcription découverts initialement chez la drosophile (Struhl, 1981). Les PcG, ainsi que leurs antagonistes activateurs de la transcription du groupe trithorax (TrxG) prennent part aux mécanismes épigénétiques, c'est-à-dire au maintien d'un profil d'expression génétique au cours des divisions cellulaires successives. Au cours des phases précoces du développement embryonnaire, ces deux familles de protéines sont impliquées dans l'expression transitoire de gènes de la segmentation afin de maintenir un profil d'expression spatial de gènes homéotiques (*Hox*).

Les PcG ont été découverts initialement chez la drosophile, mais de nombreux homologues ont récemment été identifiés chez d'autres invertébrés et vertébrés, tels l'homme, la souris, le poulet, le xénope, la drosophile ou le nématode (Table 1).

Protéines PcG	Organisme	Interactants
Pc		
Pc	*Drosophila*	Psc, RING1
XPc	*Xenopus*	XBmi1
CHCB3	Poulet	
M33	Souris	Bmi1, Ring1A et B
MPc2	Souris	
Cbx2 / HPC1	Homme	RING1
HPC2	Homme	RING1, CtBP
Psc		
Psc	*Drosophila*	Pc, Ph

XBmi1	*Xenopus*	HPH1, HPH2, XPc
Bmi1	Souris	MPh2, Ring1B, M33
mel-18	Souris	
BMI1	Homme	RING1, BMI1
Ph		
Ph	*Drosophila*	Ph, Scm
Rae-28 / MPh1	Souris	Rae-28, Bmi1
MPh2	Souris	Bmi1, Ring1B
HPH1	Homme	HPH1, HPH2
HPH2	Homme	HPH1, HPH2, BMI1
Scm		
Scm	*Drosophila*	Scm
SCML1	Homme	
E(z)		
E(Z)	*Drosophila*	Esc
mes-2	*C. elegans*	mes-6
Enx1 / Ezh2	Souris	eed, Vav
Enx2 / Ezh1	Souris	eed
EZH1	Homme	
EZH2	Homme	EED
Esc		
Esc	*Drosophila*	E(Z)
Xeed	*Xenopus*	
mes-6	*C. elegans*	mes-2
eed	Souris	Enx1 / Ezh2, Enx2
EED	Homme	EZH2, SUZ12
Pho		
Pho	*Drosophila*	
YY1	Homme	EED
Enhancer of polycomb		
E(Pc)	*Drosophila*	
Epc1, Epc2	Souris	
EPC1, EPC2	Homme	

Table 1 : *Protéines homologues appartenant à la famille des PcG. Pour chaque protéine, ses interactants connus sont également mentionnés.*

Peu de choses sont connues sur les mécanismes moléculaires d'action des PcG. Toutefois, il a été suggéré qu'ils agissaient sous forme de complexes multi protéiques, en s'associant à la chromatine[1] et en modifiant la structure de cette dernière (Pirrotta *et al.*, 2003).

Le premier complexe PcG à avoir été isolé chez l'homme est PRC1 (Polycomb repressive complex 1). Il comporte les protéines BMI-1 (B cell-specific Mo-MLV integration site 1) HPC2 (Human Polycomb 2) HPH (Human Polyhomeotic) et RING1 (Ring-finger protein 1 ; Mager *et al.*, 2003). EED, pour Embryonic Ectoderm Developement, fait partie d'un autre complexe appelé PRC2. Celui-ci comporte également les protéines EZH2 (Enhancer of Zeste 2) SUZ12 (Supressor of Zeste 12) et les protéines de fixation aux histones RbAp46 et 48 (Cao *et al.*, 2002 ; Figure 1).

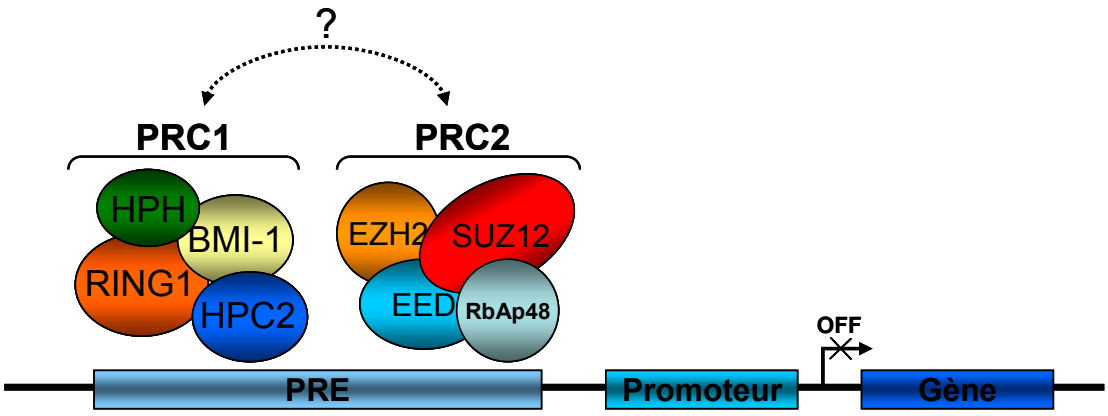


Figure 1 : *Modèle de deux complexes protéiques PcG humains. Ces deux complexes, PRC1 et PRC2, ne semblent pas interagir l'un avec l'autre. Par contre, chaque complexe interagit avec une séquence PRE (Polycomb Response Element) située à proximité du gène cible. Selon un mécanisme encore obscur, la région promotrice pourrait être masquée par le / les complexe(s) PcG, empêchant ainsi l'expression du gène cible (d'après Satijn et al., 1999).*

Il est intéressant de souligner que des travaux récents rapportent l'existence de quatre isoformes de EED (Figure 2) dans les cellules mammifères, chacune d'entre elles pouvant interagir avec EZH2 et SUZ12 et donnant lieu à la formation de complexes distincts appelés PRC2 et PRC3 (Kuzmichev *et al.*, 2004). PRC2 serait ainsi composé d'une protéine EED (EED isoforme 1) de 535 résidus débutant par la valine 1 et serait hyperphosphorylée. PRC3 serait quant à lui composé des protéines EED isoforme 3 ou 4, correspondant respectivement

[1]La chromatine constitue le support de l'information génétique chez les organismes eucaryotes. C'est une structure complexe constituée d'ADN et de protéines (histones) localisée dans le noyau cellulaire. On distingue l'hétérochromatine, qui représente la forme condensée de la chromatine et qui ne change pas d'état au cours du cycle cellulaire de l'euchromatine, qui représente la chromatine et qui apparaît relâchée pendant l'interphase.

à des produits d'initiation en Met95 et Met110. EED isoforme 2, qui correspond à une protéine de taille intermédiaire entre EED forme 1 et EED forme 3, pourrait quant à elle entrer dans la composition d'un nouveau complexe PRC4 (Kuzmichev *et al.*, 2005).

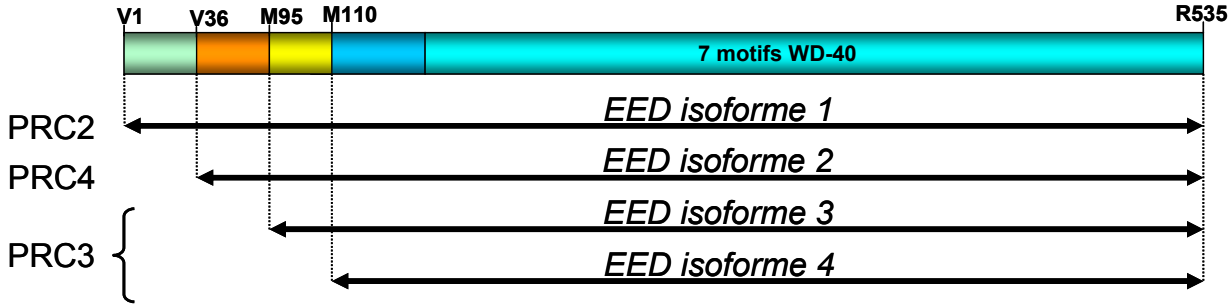


Figure 2 : *Caractérisation des 4 isoformes de EED. Le site d'initiation de la traduction (V1, V36, M95 et M110) de chaque isoforme est indiqué, ainsi que son appartenance à un complexe PRC. Dans cette étude sur la protéine EED, c'est l'isoforme 3 qui sera étudiée ; sa numérotation, résidus 95 à 535 (441 résidus) sur cette figure, sera notée des résidus 1 à 441 (441 résidus) dans la suite de ce manuscrit.*

La caractérisation des complexes PRC2 et PRC3 montre que chaque complexe comporte une activité histone lysine méthyltransférase (HMTase) intrinsèque portée par le domaine SET[2] de EZH2 (Kuzmichev *et al.*, 2002). *In vitro*, EZH2 dans PRC2/3 peut méthyler la lysine 9 (H3-K9) et la lysine 27 (H3-K27) de l'histone H3 ou la lysine 26 (H1-K26) de l'histone H1. L'efficacité avec laquelle ces deux histones peuvent être méthylées dépend de l'isoforme de EED présente dans le complexe PRC2/3.

EZH2 possède la même activité méthyl-transférase dans le complexe PRC4 et dans le complexe PRC2/3, à la seule différence qu'elle perd sa capacité à méthyler H3 en présence de H1.

1.a) Répression des gènes homéotiques : (Figure 3)

Les gènes *Hox* sont connus comme étant des cibles des PcG chez la drosophile, les vertébrés et les plantes. Des études récentes ont montré que ce système de régulation est également conservé chez le nématode *C.elegans* (Ross et Zarkower, 2003). La méthylation de la lysine

[2] Les protéines à domaine SET catalysent le transfert d'un groupement méthyle du cofacteur S-adénosylméthionine (AdoMet) vers des résidus lysines spécifiques de protéines substrats, telles que la queue N-terminale des histones H3 ou H4. Les gènes codant pour les protéines à domaine SET sont largement représentées parmi les génomes eucaryotes et ont été initialement classées en trois catégories, SU(VAR)3-9, E(Z) et Trithorax.

K27 de l'histone H3 par PRC2 pourrait constituer un site de reconnaissance pour le chromodomaine de PRC1. PRC1 pourrait ainsi bloquer le recrutement d'activateurs de la transcription tels SWI/SNF du groupe TrxG. De la même manière, il a été montré que la méthylation de la lysine K9 de l'histone H3 (H3-K9) crée un site de haute affinité pour la protéine de l'hétérochromatine HP1 (Heterochromatin Protein 1 ; Lachner *et al.*, 2001) laquelle semble constituer le maillon indispensable à la formation d'hétérochromatine, donc au maintien d'un état silencieux de la transcription.

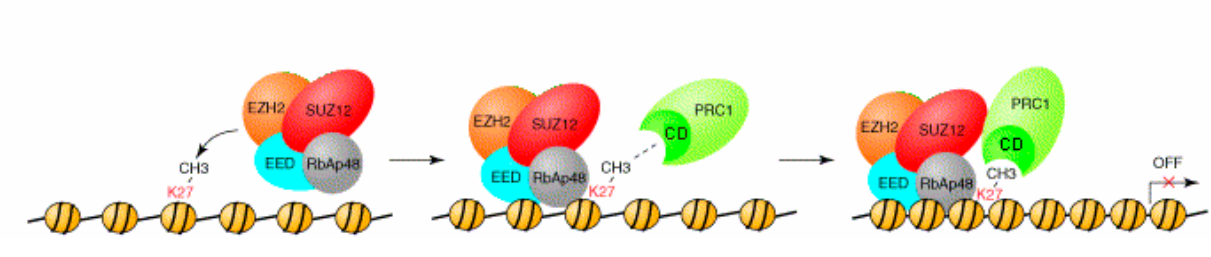


Figure 3 : *Au cours de la répression des gènes Hox, le complexe PRC2 (EED/EZH2/SUZ12/RbAp48) pourrait faciliter le recrutement du complexe PRC1 en assurant la méthylation de H3-K27. Le recrutement de PRC1, via son chromodomaine (CD) avec 3m-K27, pourrait participer à la condensation de la chromatine, réprimant ainsi l'expression des gènes cibles (d'après Cao et al., 2002).*

1.b) Inactivation du chromosome X : (Figure 4)

En plus de réguler le développement embryonnaire, les PcG mammifères sont également impliquées dans les phénomènes d'inactivation du chromosome X et dans le contrôle de la prolifération cellulaire.

En effet, malgré un nombre de copies différent du chromosome X selon le sexe (XX pour la femelle et XY pour le mâle) un mécanisme de compensation permet un taux d'expression des gènes quasiment identique. Chez les mammifères, ceci est rendu possible par le maintien silencieux de la transcription de la plupart des gènes d'un des deux chromosomes X très tôt au cours du développement : c'est le processus d'inactivation du chromosome X.

Durant ce processus d'inactivation, le chromosome X inactif (Xi) acquiert plusieurs caractéristiques physico-chimiques distinctes de son homologue actif. L'une de ces caractéristiques les plus remarquable est un profil de méthylation différentiel de l'histone H3 dès les premières étapes du processus d'inactivation du chromosome X (Cohen et Lee, 2002).

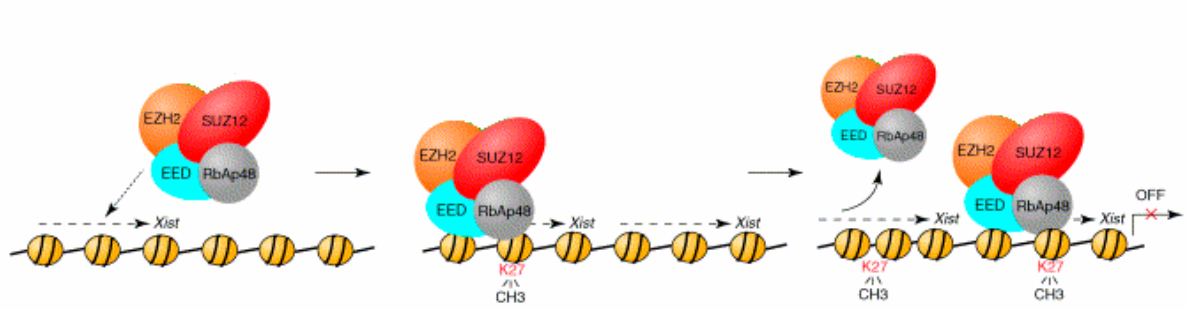


Figure 4 : *Au cours de l'inactivation du chromosome X, la transcription de l'ARN Xist à partir du chromosome X inactif (Xi) aide au recrutement du complexe PRC2 (EED/EZH2/SUZ12/RbAp48). La méthylation de H3-K27 par PRC2 pourrait contribuer à la répression de l'expression des gènes du Xi (d'après Cao et al., 2002).*

L'association transitoire du complexe PRC2 avec le chromosome X inactif intervient dès les premières étapes du processus d'inactivation et est corrélée avec son association en *Cis* avec un transcrit spécifique du X inactif (Xist) qui est le marqueur le plus précoce connu du processus d'inactivation du X. L'activité histone méthyltransférase (HMTase) intrinsèque de PRC2, portée par le domaine SET de Ezh2, permet alors la méthylation des lysines K9 et / ou K27 de l'histone H3. Cette méthylation des nucléosomes du Xi contribue au maintien silencieux de l'expression de ses gènes par l'adoption d'un état hétérochromatique.

1.c) Prolifération cellulaire et cancer : (Figure 5)

Les PcG et notamment le complexe PRC2 sont enfin impliqués dans des phénomènes de prolifération cellulaire. Les protéines Ezh2 et SUZ12 du complexe PRC2 sont surexprimées respectivement lors de cancers de la prostate (Varambally *et al.*, 2002) et dans plusieurs tumeurs humaines, notamment dans des tumeurs du colon, du sein et du foie (Kirmizis *et al.*, 2003). Des études montrent que EED, Ezh2 et SUZ12 sont sous le contrôle du facteur de transcription E2F (Bracken *et al.*, 2003). Comme de nombreux gènes cibles de E2F codent des protéines essentielles pour le contrôle de la prolifération cellulaire, il est possible qu'un dérèglement de l'expression de Ezh2 ou SUZ12 entraîne un mauvais contrôle des gènes sous le contrôle de PRC2, conduisant à une prolifération cellulaire incontrôlée.

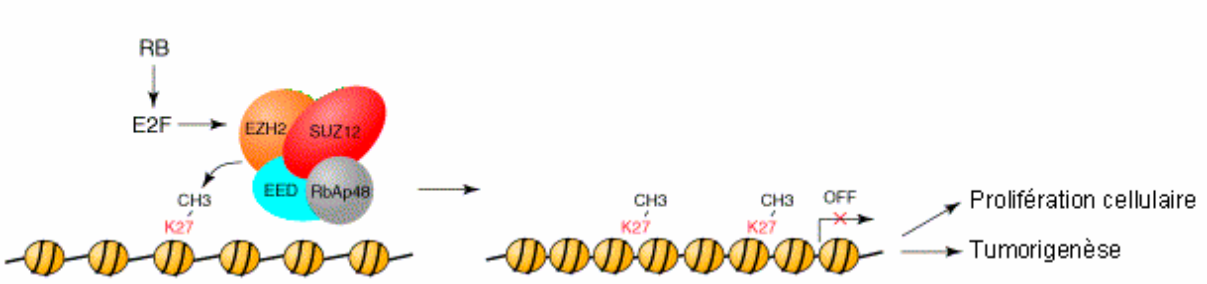


Figure 5 : *Au cours d'un cancer, une mauvaise expression de EZH2 ou SUZ12 sous l'influence de Rb-E2F peut modifier la stœchiométrie du complexe PRC2 (EED/EZH2/SUZ12/RbAp48) conduisant à une mauvaise expression de gènes impliqués (d'après Cao et al., 2002).*

2. Connaissances structurales sur EED :

Une recherche d'homologie éventuelle par rapport à l'ensemble des séquences répertoriées dans les banques de données de séquences protéiques SWISSPROT et TrEMBL (Boeckmann *et al.*, 2003) montre que la protéine EED humaine possède 99 % d'identité en acides aminés avec son homologue murin. Chez la souris, cette protéine est considérée comme un régulateur central de la segmentation de l'axe antéropostérieur. Son rôle apparaît indispensable dans ce mécanisme et certaines mutations ponctuelles sont létales pour l'embryon (Faust *et al.*, 1995 ; Holdener *et al.*, 1995). EED est par ailleurs identique à 55 % avec la protéine ESC (Extra Sex Combs) de la drosophile (Gutjahr *et al.*, 1995). ESC fait également partie de la grande famille des protéines Polycomb (Gutjahr *et al.*, 1995 ; Struhl, 1981).

Plus généralement, une recherche sur la banque de données de signatures protéiques PROSITE (Hulo *et al.*, 2004) montre que EED présente dans sa séquence des motifs caractéristiques des protéines de la famille WD-40. Les protéines de cette famille, en très grande majorité retrouvées chez des organismes eucaryotes, interviennent dans des mécanismes moléculaires très divers et se rencontrent dans tous les compartiments cellulaires (Table 2).

Fonction biologique	Protéine à motif WD-40 impliquée
Transduction du signal	Protéine Gβ, RbAp48
Synthèse et maturation des ARN	TUP1, Ski8p
Assemblage de la chromatine	CAF-1
Trafic vésiculaire	SEC13
Assemblage du cytosquelette	MAP, Aip1
Régulation du cycle cellulaire	Mad2, CDC4, Bub3
Programmation de la mort cellulaire	Apaf-1
Fonction inconnue	WDR1, WDR3, WDR4, WDR6, WDR10

Table 2 : *Diverses fonctions biologiques de protéines à motifs WD-40.*

Le motif WD-40 canonique, s'écrit $(X_{6-94}\text{-}[GH\text{-}X_{23-41}\text{-}WD])_{4-8}$ (Neer *et al.*, 1994). Il comprend 40 à 60 résidus et peut contenir le dipeptide GH, situé entre 10 et 20 résidus de son extrémité N-terminale, et le dipeptide WD situé à son extrémité C-terminale. Il est à noter que ni le dipeptide GH, ni le dipeptide WD ne sont absolument conservés.

Généralement, les protéines appartenant à cette famille comportent entre 4 et 8 répétitions de ce motif. Ce nombre varie pour EED selon les auteurs. Denisenko et collaborateurs considèrent EED comme une protéine à 6 motifs WD-40 (Denisenko et Bomsztyk, 1997) alors que Shumacher et collaborateurs la considère comme une protéine à 5 motifs (Shumacher *et al.*, 1996).

Pour notre part, nous préférons considérer que EED contient 7 motifs WD-40. Ces motifs WD-40, numérotés dans notre étude, II (131-176) III (179-219) IV (222-264) V (295-332) et VII (397-438) sont aussi répertoriés par les deux autres équipes. La séquence comprise entre les résidus 356 et 390 (motif WD-40 VI dans notre étude) est considérée comme un motif WD-40 par l'équipe de Bomsztyk (Denisenko et Bomsztyk, 1997) mais pas par celle de Magnuson (Shumacher *et al.*, 1996). Enfin, contrairement aux deux autres équipes, nous avons considéré un septième motif WD-40 entre les résidus 81 et 125 (domaine I dans notre étude).

A ce jour, neuf structures cristallographiques de protéines à motifs WD-40 sont connues (Figure 6):

(1) la sous-unité β des protéines G (code PDB :1TBG ; Gaudet *et al.*, 1996 ; Lambright *et al.*, 1996 ; Sondek *et al.*, 1996 ; Wall *et al.*, 1995).

(2) le domaine C-terminal du répresseur transcriptionnel Tup1 (code PDB : 1ERJ ; Sprague *et al.*, 2000).

(3) la protéine ARPC1 p40 (code PDB : 1K8K ; Robinson *et al.*, 2001).

(4) le domaine C-terminal du corépresseur transcriptionnel Groucho / TLE1 (code PDB : 1GXR ; Pickles *et al.*, 2002).

(5) la protéine Cdc4 (code PDB : 1NEX ; Orlicky *et al.*, 2003).

(6) la protéine Aip1 (code PDB : 1PI6 ; Voegtli *et al.*, 2003).

(7) la protéine TrCP1 (code PDB : 1P22 ; Wu *et al.*, 2003).

(8) la protéine Bub3 (code PDB : 1U4C ; Larsen et Harrison, 2004, non publié).

(9) la protéine Ski8p (code PDB : 1SQ9 ; Madrona et Wilson, 2004).

Toutes ces protéines adoptent une structure en turbine constituée de pales en feuillet β. Le nombre de pales est égal au nombre de motifs WD-40 présents dans leur séquence. Ceci renforce l'hypothèse que toutes les protéines à motifs WD-40 pourraient se structurer en turbine-β. En outre, si la plupart de ces protéines adoptent une structure en turbine-β à 7 pales, la protéine Cdc4 en présente 8. Enfin, la protéine Aip1, dont la séquence était prédite pour contenir 10 motifs répétés WD-40, présente une structure formée de 2 turbines-β à 7 pales connectées l'une à l'autre.

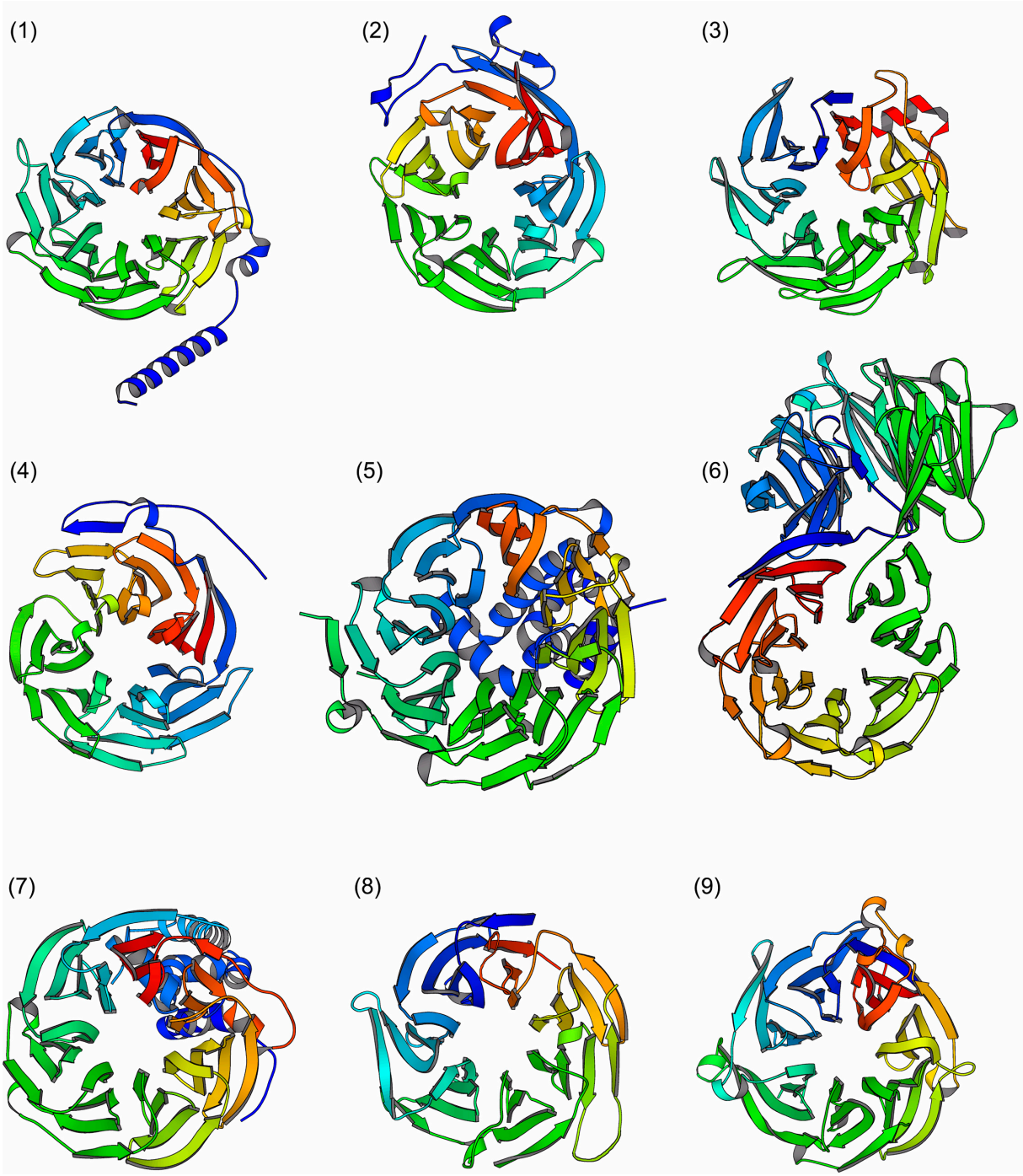


Figure 6 : *Diagramme en rubans des protéines à motifs WD-40 de structures connues. (1) protéine Gβ ; (2) domaine C-terminal de Tup1 ; (3) protéine ARPC1 p40 ; (4) domaine C-terminal de Groucho / TLE1 ; (5) protéine Cdc4 ; (6) protéine Aip1 ; (7) protéine TrCP1 ; (8) protéine Bub3 ; (9) protéine Ski8p.*

Chaque pale est constituée d'un feuillet β à 4 brins β antiparallèles disposés radialement autour d'un axe central. La stabilité de ces structures est due à des interactions hydrophobes entre les feuillets, et à un système de «fermeture du cercle» également appelé système «Velcro». En effet, la séquence répétée WD-40 ne correspond pas à une pale, mais au brin β

externe (brin β d) d'une pale et aux 3 brins β les plus proches de l'axe central (brins a, b et c) de la pale suivante (Figure 7). Ainsi, le brin β externe de la dernière pale de la turbine-β est fourni par la partie variable du premier motif WD-40 situé en N-terminal de la protéine, alors que les trois brins β internes sont fournis par le dernier motif WD-40 (Figure 7). Ceci permet d'obtenir une molécule circulaire. La nécessité apparente d'un tel mécanisme de fermeture suggère que de telles structures circulaires seraient instables en son absence, et que ce mécanisme les protège d'un mauvais repliement qui pourrait conduire à la formation de fibrilles amyloïdes (Chiti *et al.*, 1999).

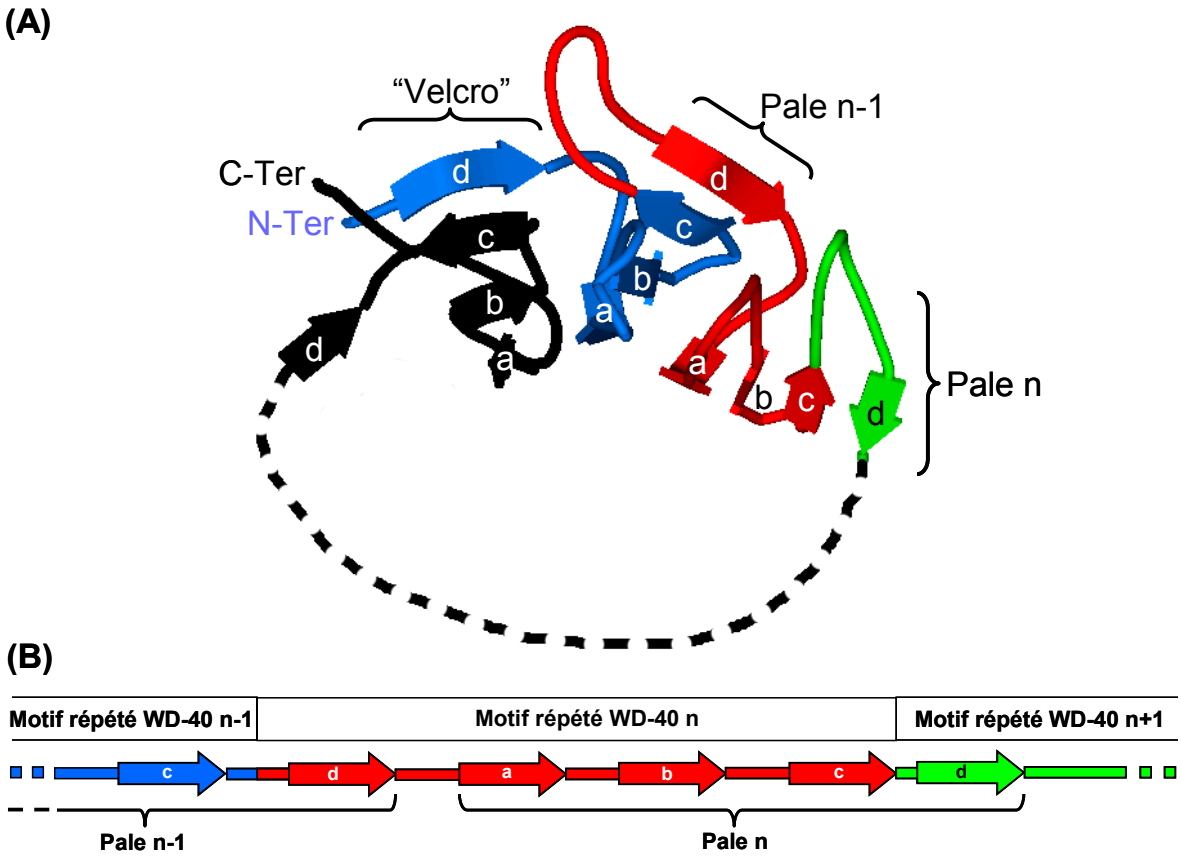


Figure 7 : (A) *Diagramme en ruban d'une portion de protéine en turbine-β montrant 3 pales constituées chacune de 4 brins β. La dernière pale (noire et bleue) montre le système «velcro» de fermeture permettant à la protéine d'acquérir une structure circulaire.* **(B)** *Représentation schématique montrant le décalage entre un motif répété WD-40 et la structure observée (la pale), le premier brin β (d) du motif répété WD-40 n constituant le brin β externe de la pale n-1.*

Ces protéines en turbines-β contiennent ainsi trois surfaces potentielles d'interaction ; la face supérieure constituée principalement des boucles reliant les brins β d aux brins β a (boucle d-a), la face inférieure constituée principalement des boucles reliant les brins β a aux brins β b

(boucles a-b) et celles reliant les brins β c aux brins β d (boucles c-d) et leur circonférence composée principalement des brins β d (Figure 8).

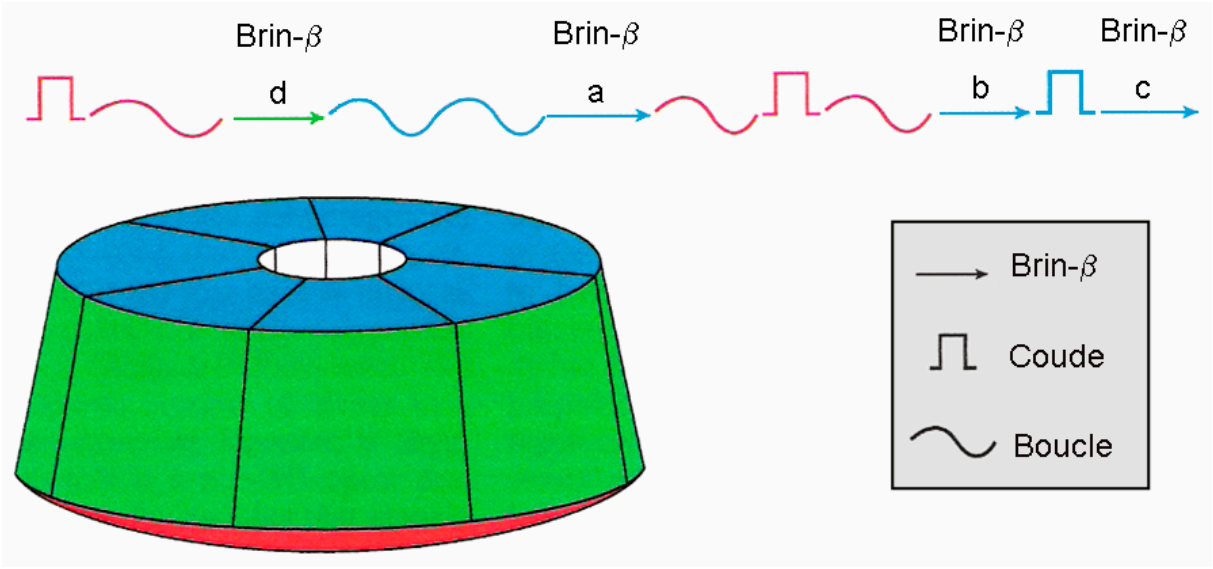


Figure 8 : (A) *Représentation schématique des éléments structuraux au sein d'un même motif répété WD-40.* **(B)** *Représentation schématique des positions des éléments structuraux présentés en (A) au sein de la structure tridimensionnelle de la turbine-β. Sa face supérieure est montrée en bleue, sa face inférieure en rouge et sa circonférence en vert (d'après Smith et al., 1999).*

Aucune protéine à motifs WD-40 n'est connue pour posséder une activité catalytique. Elles joueraient plutôt un rôle de «plateforme» stable pouvant donner lieu à la formation réversible de complexes multi protéiques Elles pourraient ainsi coordonner des interactions séquentielles ou simultanées de plusieurs jeux de protéines (Smith *et al.*, 1999). Plus récemment, Cooley et son équipe ont suggéré que les protéines en turbine-β pouvaient agir en temps qu'«organisateurs de complexes multimoléculaires», exploitant la grande variabilité de leurs boucles pour établir des contacts avec des partenaires moléculaires variés (Adams *et al.*, 2000).

B/ BIOLOGIE DU VIRUS DE L'IMMUNODEFICIENCE HUMAINE

1. Introduction :

Depuis sa découverte il y a plus de 20 ans, le virus du Sida est devenu la première cause de mortalité dans le monde. 40 millions de personnes sont à ce jour infectées par le VIH, alors que 3 millions de personnes ont été tué par la maladie en 2003.

Tout commence aux Etats-Unis en 1981, lorsque 5 cas d'une maladie rare, la pneumocystose pulmonaire, furent détectés. A la fin de l'année, la première étude épidémiologique indique qu'une maladie inconnue, provoquant une immunodéficience, se transmet par voie sexuelle et sanguine en touchant principalement les homosexuels. Le nom de SIDA (Syndrome de l'Immunodéficience Acquise) est créé.

Fin 1982, le nombre de cas de SIDA augmente considérablement et s'étend à l'ensemble de la population, notamment aux hémophiles, aux personnes à partenaires sexuels multiples et aux enfants nés de mère à risque.

En 1983, l'agent viral responsable de la maladie est isolé à l'Institut Pasteur par l'équipe du Pr Montagnier à partir de cellules lymphocytaires d'un patient atteint d'une lymphadénopathie (Barre-Sinoussi *et al.*, 1983). C'est un nouveau virus baptisé LAV (Lymphadenopathy Associated Virus).

En 1984 le Pr Gallo isole un virus qu'il nomme le HTLV-3 (Popovic *et al.*, 1984). Ce virus s'avérera être identique au virus identifié un an plus tôt par l'équipe Française.

La communauté scientifique s'accorde en 1986 pour leur donner le nom commun de virus de l'immunodéficience humaine (VIH). Un deuxième virus du Sida, donnant des symptômes légèrement différents du premier sérotype, est découvert à l'Institut Pasteur (Clavel *et al.*, 1986). Ce VIH-2 est présent majoritairement en Afrique de l'Ouest.

En 1990 naît l'idée d'associer plusieurs molécules thérapeutiques pour bloquer la réplication du virus. Cette multithérapie pourrait éviter l'apparition de souches virales résistantes aux médicaments. Le nombre de malades est alors estimé à environ 1 million.

Les premiers tests de vaccins contre le Sida ont lieu chez l'Homme en 1993. C'est un échec car le vaccin ne parvient pas à arrêter la prolifération du virus. Une étude publiée en 1995 conclut à l'efficacité des inhibiteurs de protéases pour lutter contre le Sida, ainsi qu'à l'effet positif des bi- et trithérapies. L'association de plusieurs molécules pour lutter contre le Sida devient rapidement la norme dans les pays industrialisés. Le nombre de morts liés au Sida diminue pour la première fois dans les pays occidentaux et l'état des patients s'améliore alors

malgré les effets secondaires des traitements. Le nombre de personnes infectées dans le monde est alors estimé à 20 millions.

En 1996, l'efficacité des trithérapies est confirmée. Cette année-là, le premier inhibiteur non nucléosidique de la transcriptase inverse apparaît aux USA. En 1998, l'effet protecteur de l'azidothymidine (AZT) est démontré dans la transmission de la mère à l'enfant.

En 1999 débutent des essais cliniques pour tester l'efficacité d'une nouvelle molécule (T-20) appartenant à une nouvelle classe de molécules anti-virales : les inhibiteurs de la fusion.

C'est durant l'année 2000 qu'est émise l'idée de faire des pauses dans les traitements anti-viraux pour tenter de limiter l'apparition de souches virales résistantes.

L'année 2001 est le vingtième anniversaire de la découverte du virus. Malgré d'énormes progrès réalisés dans la compréhension de la maladie et dans la prise en charge des malades, aucune solution à court terme ne semble se dessiner. Le Sida tue toujours.

2. Classification des rétrovirus :

Le VIH-1 est un lentivirus de la famille des *retroviridae*. Cette famille regroupe une cinquantaine de virus exogène retrouvés généralement chez les vertébrés et répartis d'après le comité international de taxonomie des virus (ICTV, 2000) en sept genres selon leur pathogénicité et leurs propriétés structurales (Table 3). Le cycle viral de tous ces virus se caractérise par une étape de réplication inverse de leur génome ARN en ADN grâce à une polymérase virale, la transcriptase inverse, puis par l'intégration de ce dernier dans le génome de l'hôte.

Genre	Espèce type
Alpharétrovirus	Virus de la leucémie aviaire (ALV)
Betarétrovirus	Virus de la tumeur mammaire murine (MMTV)
Gammarétrovirus	Virus de la leucémie murine (MLV)
Deltarétrovirus	Virus de la leucémie bovine (BLV)
Epsilonrétrovirus	Virus du sarcome dermique du saumon (WDSV)
Lentivirus	Virus de l'immunodéficience humaine de type 1 (VIH-1)
Spumavirus	Virus spumeux du Chimpanzé CFV)

Table 3 : *Classification des rétrovirus selon le comité international de taxonomie des virus (ICTV, 2000).*

Une autre classification basée sur la complexité des génomes a également été proposée (Cullen, 1991). Des rétrovirus simples et complexes ont ainsi été définis (Table 3). Les rétrovirus simples nécessitent la seule présence des gènes *gag*, *pol* et *env*, alors que les rétrovirus complexes (Lentivirus et Spumavirus) possèdent en plus des séquences codant des protéines accessoires de régulation.

3. Morphologie de la particule virale et organisation du génome :

Le VIH-1 est un virus enveloppé dont la particule virale sphérique présente un diamètre de 90 à 120 nm (Figure 9). Cette dernière est composée d'une enveloppe externe de nature protéo-lipidique, issue du bourgeonnement de la membrane plasmique de la cellule infectée. Y sont ancrées les glycoprotéines gp120 ou SU (Surface protein) et gp 41 ou TM (Transmembranar protein). Ces deux protéines associées de manière non covalente s'organisent en trimère, formant des spicules à la surface de la particule virale.
La membrane interne est constituée des protéines de Matrice (MA) organisées en trimères. Celles-ci tapissent la surface interne de l'enveloppe et sont étroitement associées à la bicouche lipidique grâce à leur extrémité N-terminale myristoylée.

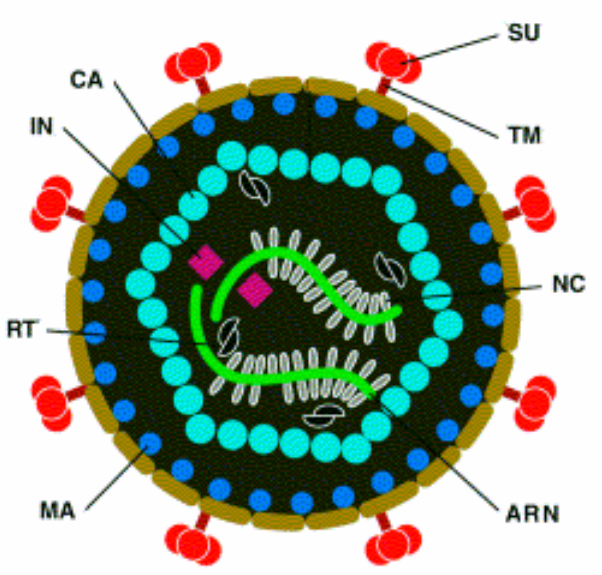


Figure 9 : *Morphologie de la particule virale. Les différents composants sont indiqués : CA (capside) ; IN (Intégrase) RT (transcriptase inverse) MA (Matrice) NC (nucléocapside) TM (gp41) et SU (gp 120 ; d'après Sherman et Greene, 2002).*

Les protéines de capside (CA) forment une capside virale de forme conique renfermant et protégeant le génome viral. Celui-ci est constitué d'un ARN bicaténaire associé à des protéines de nucléocapside (NC). Ces petites protéines basiques permettent une compaction de l'ARN au sein de la capside. Cette capside renferme également des enzymes virales libres (Intégrase, reverse transcriptase, protéase et protéines accessoires) et des molécules d'origine cellulaire (ARN de transfert, ARN ribosomaux).

Le génome viral comporte 9 cadres ouverts de lecture codant 15 protéines (Figure 10). Ceci est rendu possible par la synthèse de polyprotéines, par la présence de nombreux sites d'épissage et de sites différents d'initiation de la transcription.

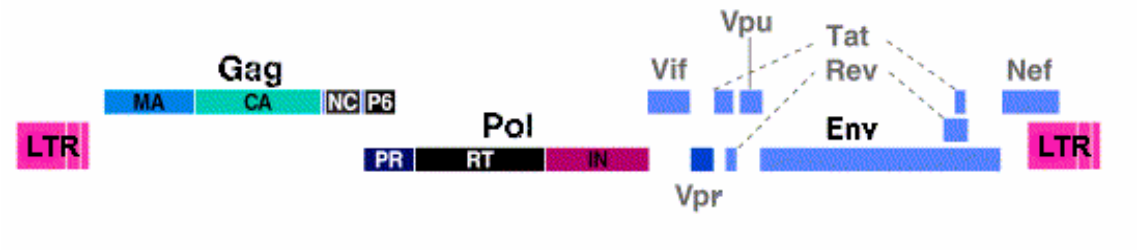


Figure 10 : *Organisation génomique de l'ADN du VIH-1. Le génome comporte 9 phases ouvertes de lecture encadrées par deux séquences LTR (Long Terminal Repeat). 3 ORF correspondent aux gènes gag, pol et env, et 6 codent les protéines accessoires Vif, Vpr, Vpu et Nef, ainsi que les protéines de régulation Tat et Rev (d'après Sherman et Greene, 2002).*

Les gènes *gag*, *pol* et *env* sont présents chez tous les rétrovirus. Le gène *gag* code le précurseur $Pr55^{gag}$, qui une fois clivé par PR donnera les protéines virales MA, CA, NC et p6[3]. Le gène *pol* code la polyprotéine $Pr160^{gag\text{-}pol}$ qui après maturation donne les enzymes rétrovirales PR, RT et IN. Les gènes *gag* et *pol* se chevauchent sur 241 nucléotides et l'expression du précurseur $Pr160^{gag\text{-}pol}$ est rendue possible par un glissement de type -1 du ribosome par rapport au cadre de lecture du gène *gag* (Jacks *et al.*, 1988). Ce changement de cadre de lecture ne survient qu'avec une faible fréquence (5 %). Cette régulation de l'expression des enzymes virales s'explique par le fait que la synthèse de nouvelles particules virales requiert plus de protéines de structure (MA, CA et NC) que de protéines enzymatiques (PR, RT et IN). Le gène *env* code les protéines d'enveloppe SU et TM. Elles sont issues de la

[3]La protéine p6 du VIH-1 est une protéine flexible de 52 résidus sans structure apparente. Elle contient 2 motifs critiques : un motif PTAPP impliqué dans le processus de bourgeonnement des particules virales et dans l'incorporation de Pol et du RNA, probablement par interaction avec des facteurs cellulaires pour l'instant non définis. Un deuxième motif de p6 joue un rôle dans l'incorporation de la protéine accessoire VPr

maturation du précurseur $Pr160^{env}$ par des protéases cellulaires au niveau du Golgi, avant leur adressage à la membrane en vue de leur incorporation au sein de particules virales néo-synthétisées.

Le VIH-1 présente 6 gènes supplémentaires codant pour des protéines de régulation (Tat et Rev) et des protéines dites accessoires (Nef, Vif, Vpr et Vpu). La conservation de ces gènes malgré la forte pression de sélection suggère qu'ils participent à des mécanismes importants lors de l'infection virale. Les protéines Tat, Rev et Nef sont issues d'ARN_m multi-épissés et les protéines Vif, Vpr et Vpu d'ARN_m mono-épissés. Ces protéines revêtent différentes fonctions, telle Tat qui a un rôle transactivateur des gènes viraux, Vif qui intervient dans l'assemblage et la maturation des virions, Rev qui est impliqué dans le transport des ARN_m non-épissés du noyau vers le cytoplasme, Vpr qui joue un rôle prépondérant dans le transport du complexe de pré-intégration (PIC) du cytoplasme vers le noyau, Nef et Vpu qui facilitent le transport des protéines d'enveloppe à la surface cellulaire en vue de la formation de nouvelles particules virales (Frankel et Young, 1998).

4. Le cycle viral du VIH-1 :

Le cycle viral s'étend de l'entrée de la particule virale dans la cellule hôte jusqu'à la production des nouvelles particules infectieuses (Figure 11). Il peut être divisé en plusieurs étapes successives regroupées en deux phases distinctes : la phase précoce et la phase tardive. La phase précoce est constituée des étapes comprises entre la pénétration du virus dans la cellule hôte et l'intégration du provirus dans son génome. La phase tardive débute avec l'expression des protéines virales et s'achève par le bourgeonnement des particules virales néo-synthétisées.

4.a) La phase précoce du cycle réplicatif :

La pénétration du virion dans le cytoplasme de la cellule hôte nécessite dans un premier temps la fixation de ce dernier sur sa cible. Ceci est réalisé par la reconnaissance de la glycoprotéine d'enveloppe SU par les molécules membranaires réceptrices CD4 de la cellule hôte. Cette interaction à haute affinité induit un changement conformationnel de SU, permettant la reconnaissance de régions particulières de cette protéine par d'autres protéines

cellulaires de surface, telles les récepteurs aux chimiokines CCR5 et CXCR4. La formation de ce complexe conduit, par l'intermédiaire de la protéine virale TM, à la fusion de l'enveloppe virale avec la membrane plasmique et permet la libération de la capside virale dans le cytoplasme de la cellule hôte.

Suite à la pénétration du virus, le virion subit une «décapsidation» conduisant à la formation d'un complexe nucléoprotéique, le complexe de rétro-transcription, composé des deux molécules d'ARN virales associées à des protéines et enzymes virales telles NC, MA, IN, RT et VPr

La transcription inverse est initiée par l'ARN de transfert Lys-3, coencapsidé dans la particule virale et utilisé comme amorce de polymérisation par la transcriptase inverse. La molécule d'ADN bicaténaire synthétisée s'associe alors fortement à IN au sein d'un complexe nucléoprotéique désigné sous le terme de complexe de pré-intégration (PIC). Sous l'action coordonnée des protéines MA, IN et Vpr, le PIC est transporté activement à l'intérieur du noyau cellulaire. L'ADN viral est alors intégré au sein du génome de la cellule hôte par action de l'Intégrase et est appelé provirus.

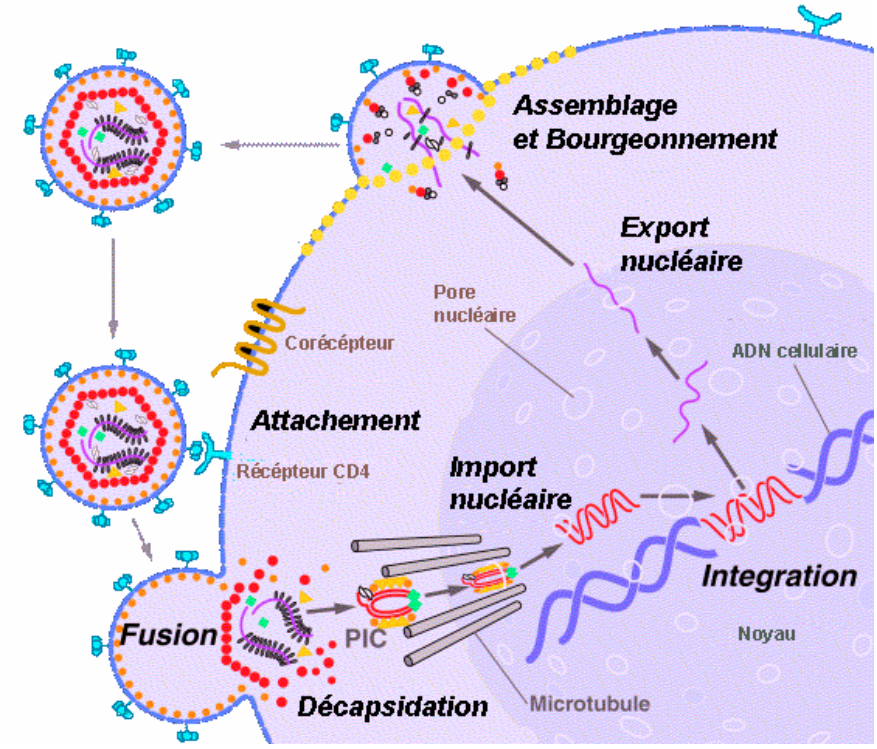


Figure 11 : *Le cycle viral du VIH-1 (d'après Sherman et Greene, 2002).*

4.b) La phase tardive du cycle réplicatif :

L'ARN viral est transcrit sous contrôle du promoteur localisé dans le LTR 5' du provirus. L'expression du génome viral est, quant à elle, initiée par la protéine Tat et régulée par des facteurs de transcription cellulaires.

Les ARN viraux, partiellement épissés et non épissés, sont exportés par l'intermédiaire de Rev vers le cytoplasme où ils sont traduits ou encapsidés.

Les ARN messagers des protéines d'enveloppe sont traduits dans le réticulum endoplasmique sous forme de précurseurs gp160, où ils s'associent avec les molécules CD4 nouvellement synthétisées. La protéine Vpu, en dirigeant la dégradation des CD4, permet la libération des gp160 vers le Golgi où ils seront glycosylés et maturés en TM et SU. Ces protéines d'enveloppe sont ensuite acheminées vers la membrane plasmique où elles sont protégées d'une éventuelle interaction avec les CD4 par la protéine Nef.

Dans le même temps, les ARN messagers *gag* et *pol* sont traduits sous forme de précurseurs polypeptidiques $Pr55^{gag}$ et $Pr160^{gag\text{-}pol}$ par les polysomes libres du cytoplasme. Leur adressage à la membrane plasmique *via* un mécanisme indépendant du trafic vésiculaire reste à définir. Les précurseurs s'assemblent au niveau de la membrane plasmique et recrutent les protéines d'enveloppe. L'ARN génomique est encapsidé et les particules virales bourgeonnent. Simultanément, les protéines virales sont maturées par protéolyse sous l'action de la protéase virale et de la protéine Vif.

C/ ROLE DE EED DANS LE CYCLE VIRAL DU VIH-1

1. Interaction avec la protéine Matrice du VIH-1 :

L'équipe du Professeur P. Boulanger avec qui nous collaborons a mis en évidence l'interaction de EED avec la Matrice du VIH-1 par le système du double hybride dans la levure (cf. Annexe A pour le principe). Ces résultats ont été confirmés *in vitro* par différentes techniques (Peytavi *et al.*, 1999). La zone d'interaction au niveau de EED a pu être localisée précisément par criblage d'une banque de phage filamenteux (technique du «phage-display» ; cf. Annexe B pour le principe) et vérifiée par mutagenèse dirigée. Elle se situe dans le motif WD-40 n°5 au sein d'une boucle reliant le premier et le second feuillet β. Ces boucles constituent chez les protéines à motifs WD-40 les principales zones d'interaction avec des partenaires protéiques. Il est intéressant de noter que cette région d'environ 20 résidus est conservée au cours de l'évolution, puisqu'elle présente 100 % d'identité avec son homologue ESC chez la drosophile (Ng *et al.*, 1997). Cette région semble donc importante pour la fonction biologique de EED.
Le fait que EED interagit avec la protéine de Matrice et non avec le précurseur $Pr55^{gag}$ laisse envisager que EED pourrait jouer intervenir pendant la phase précoce du cycle viral.

Structure de la Matrice du VIH-1

La protéine de Matrice du virus VIH-l a une masse moléculaire d'environ 15 kDa et est myristoylée sur sa glycine N-terminale. Sa structure 3D a été déterminée par RMN (Massiah *et al.*, 1994 ; Massiah *et al.*, 1996 ; Matthews *et al.*, 1994 ; Matthews *et al.*, 1995) et par diffraction aux rayons X (Hill *et al.*, 1996).
La protéine, constituée de 5 hélices α et de 3 brins β, est organisée autour de l'hélice centrale α4 (Figure 12). Celle-ci établit des contacts hydrophobes avec toutes les autres hélices, formant ainsi le coeur de la protéine. Les trois brins β forment une plate-forme basique à l'extérieur du coeur de la protéine près de son extrémité N-terminale. Ils portent les deux signaux de localisation nucléaire propres aux protéines de Matrice des lentivirus (Conte et Matthews, 1998).

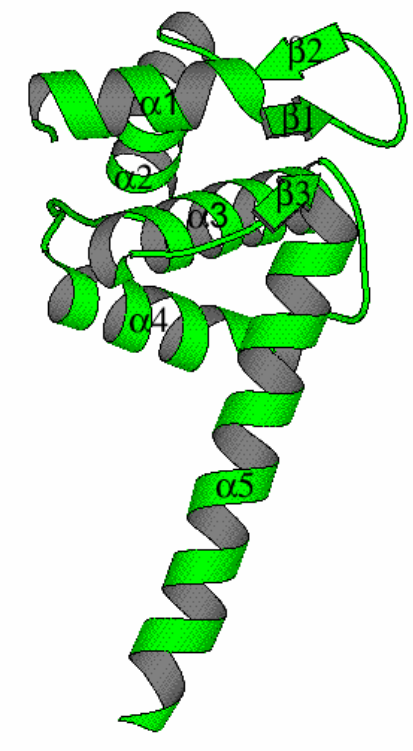


Figure 12 *: Représentation en ruban de la protéine de Matrice du VIH-1. Les structures secondaires sont numérotées respectivement de α1 à α5 pour les hélices α et de β1 à β3 pour les brins β.*

Les études cristallographiques de la Matrice du VIH-1 révèlent une oligomérisation de la protéine sous forme de trimères (Hill *et al.*, 1996). L'interaction entre les trois monomères a lieu au niveau des boucles reliant les hélice α3 et α4. Ce modèle trimérique permet d'expliquer comment les groupements myristates agissent en synergie avec les 31 résidus N-terminaux basiques du domaine Matrice pour stabiliser le précurseur à la membrane plasmique. Dans cette structure, l'orientation des 3 myristates leur permet de s'ancrer ensemble dans la bicouche lipidique. Les résidus N-terminaux basiques sont alors situés à l'apex du trimère et sont correctement positionnés pour former des liaisons ioniques avec les têtes acides des phospholipides (Hill *et al.*, 1996 ; Figure 13).

Lors de la phase tardive du cycle viral, le domaine Matrice dirige le précurseur $Pr55^{gag}$ à la membrane plasmique grâce à la présence de son double signal d'adressage membranaire (myristate et domaine polybasique N-terminal). Lors de la phase précoce du cycle, la Matrice participe au transport du complexe de pré-intégration dans le noyau grâce à sa séquence NLS N-terminale constituée par le même domaine polybasique (Gallay *et al.*, 1997). Or, le double signal d'adressage et d'ancrage à la membrane plasmique est présent à la fois dans les parties N-terminales de la Matrice et du $Pr55^{gag}$. La différence d'affinité pour la membrane plasmique

ne peut être expliquée que par une meilleure accessibilité de ce double signal dans le précurseur $Pr55^{gag}$ par rapport à la Matrice.

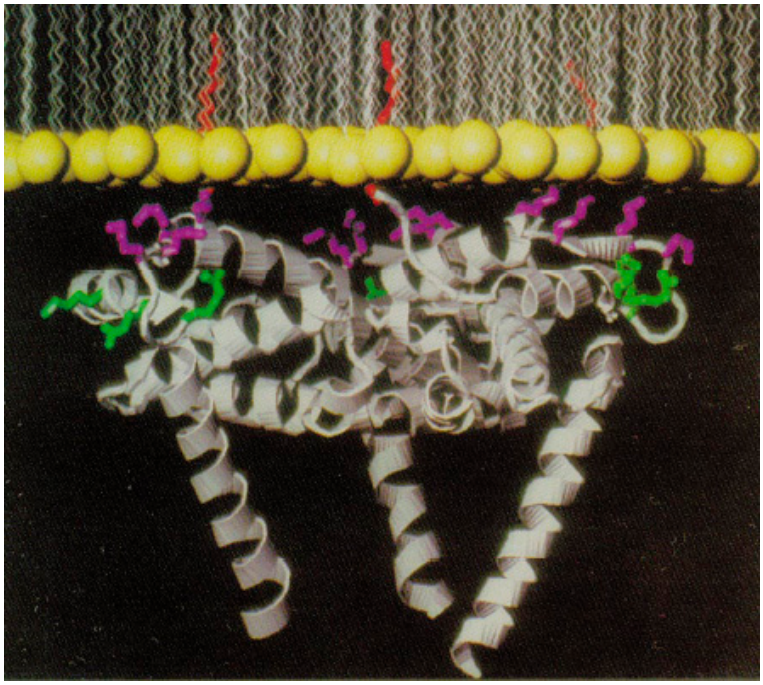


Figure 13 *: Modèle moléculaire d'un trimère de Matrice myristoylée ancrée au niveau de la membrane virale. Les radicaux myristates (en rouge) de chaque monomère sont insérés dans la bicouche phospholipidique (d'après Hill et al., 1996).*

Pour expliquer ce phénomène, un mécanisme de repliement et de séquestration du myristate à l'intérieur de la Matrice mature a été envisagé (Spearman *et al.*, 1997; Zhou et Resh, 1996). Récemment, Tang et collaborateurs ont confirmé cette hypothèse par la mise en évidence par RMN de deux états conformationnels de la protéine de Matrice myristoylée : une conformation exposant le myristate et une conformation séquestrant le myristate (Tang *et al.*, 2004).

Cette équipe a également pu démontrer par des études de sédimentation à l'équilibre que la multimérisation de la Matrice favorisait l'exposition du myristate. Lors de la phase tardive du cycle viral, la concentration de précurseurs $Pr55^{gag}$ au niveau de raft lipidique entraînerait leur multimérisation. L'exposition consécutive des myristates permettrait alors l'ancrage à la membrane. Après le bourgeonnement des nouveaux virions, $Pr55^{gag}$ est clivé par la protéase mais la Matrice reste fixée à la membrane dans un état trimérique. Lors de l'infection et de la phase de fusion, le contenu de la particule virale est dilué dans le cytoplasme de la cellule hôte, entraînant la séquestration du myristate et donc la libération de la Matrice dans le cytosol (Figure 14).

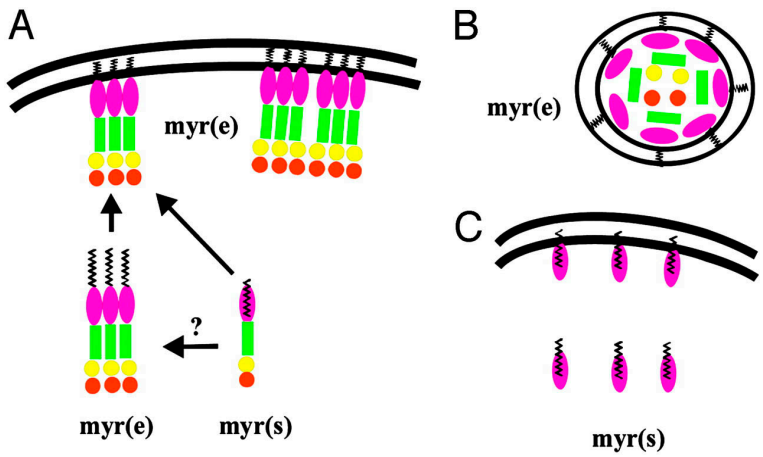


Figure 14 *: Régulation du «switch» du myristate au cours du cycle du VIH-1.Une représentation schématique de Pr55^{gag} est donnée au cours de la phase tardive de l'infection (A) dans le virion (B) et au cours de la phase précoce de l'infection (C). Au sein de Pr55^{gag}, la matrice est représentée en rose, la capside en vert, la nucléocapside en jaune et la protéine p6 en rouge (Resh, 2004).*

2. Interaction avec la protéine Intégrase du VIH-1 :

Comme nous venons de voir, l'équipe du Professeur P. Boulanger a mis en évidence une interaction entre EED et MA (Peytavi *et al.*, 1999). Au cours de ce travail, ils ont également suggéré une interaction entre EED et la protéine Intégrase par des études de co-localisation par immuno-électromicroscopie (Peytavi, 1999). Comme pour la Matrice, des co-localisations sont observées pour l'Intégrase et EED au niveau de régions du nucléoplasme. Des expériences de triples marquages des protéines de la Matrice, de l'Intégrase et de EED ont également été réalisées et des triples co-localisations peuvent être observées, toujours au niveau de régions du nucléoplasme. L'ensemble de ces résultats laisse supposer que EED pourrait lier l'Intégrase virale suite à son interaction avec la Matrice dans des régions denses du nucléoplasme. Ces régions pourraient correspondre aux régions de l'hétérochromatine condensée, au niveau desquelles EED agit naturellement.

Structure des différents domaines de la protéine Intégrase du VIH-1

L'Intégrase du VIH-1 est une protéine de 288 acides aminés, présentant trois domaines

fonctionnels bien définis : le domaine N-terminal constitué par les 50 premiers acides aminés, le domaine central long de 162 acides aminés et enfin le domaine C-terminal constitué de 76 acides aminés. Ces trois domaines sont nécessaires à la formation d'un complexe stable entre l'Intégrase et l'ADN et à la réalisation du processus d'intégration.

Le domaine N-terminal comporte 4 résidus très conservés communs à toutes les protéines Intégrase de rétrovirus (Khan *et al.*, 1991). Il s'agit de deux résidus histidine (en position 12 et 16 pour le VIH-1) et de deux résidus cystéine (en position 40 et 43 pour le VIH-1). Ces quatre résidus conservés sont impliqués dans la coordination d'ions Zn^{2+}, mais ils ne présentent pas une structure similaire aux structures en doigt à zinc classiques (Eijkelenboom *et al.*, 1997). La structure de ce domaine, composé de quatre hélices α, a été déterminée chez le VIH-1 par résonance magnétique nucléaire (Cai *et al.*, 1997). Il se présente sous forme dimérique avec une interface composée exclusivement de résidus hydrophobes provenant de la partie N-terminale de l'hélice α1 et des hélices α3 et α4. Ce domaine présente un motif hélice-tour-hélice, impliquant les hélices αl et α3, caractéristique des protéines se fixant à l'ADN.
Plus récemment, la structure du domaine «N-terminal + domaine catalytique» (Figure 15) a permis de définir une interface différente de celle définie avec le domaine N-terminal isolé (Wang *et al.*, 2001). Dans ces conditions, seules les extrémités N-terminales des hélices αl et α3 composent l'interface du dimère.

Figure 15 : *Représentation en ruban du dimère des domaines N-terminaux + domaines centraux de l'Intégrase du VIH-1. Les sphères orange représentent les ions zinc du domaine N-terminal.*

Le domaine central est le domaine le plus conservé entre les différentes protéines Intégrases (Khan *et al.*, 1991). Il possède notamment trois résidus invariables qui constituent le motif D, D(35)E, c'est à dire deux résidus acide Aspartique et un résidu acide Glutamique respectivement en position 64, 116 et 152 pour le VIH-1. Les deux derniers résidus conservés sont toujours séparés par 35 acides aminés.

La structure cristallographique de ce domaine a été obtenue pour le VIH-1 (Dyda *et al.*, 1994). Il se présente sous forme de dimère et se compose de 5 brins β et de 6 hélices α. La conformation tridimensionnelle de ce domaine montre que les trois résidus conservés et porteurs de l'activité catalytique sont très proches. En outre, cette étude a mis en évidence une boucle flexible composée par les résidus 141 à 148. La substitution des deux résidus G140 et G149 par des résidus Alanine rend cette boucle plus rigide. Les activités de cette protéine mutée, notamment l'activité de désintégration, sont fortement réduites, alors que la capacité à se fixer à l'ADN n'est que très faiblement affectée. Ils semble donc que la flexibilité de cette boucle joue un rôle important dans la structuration de la protéine IN active (Greenwald *et al.*, 1999).

Le domaine C-terminal est le domaine le moins conservé. La structure tridimensionnelle du domaine C-terminal de la protéine Intégrase du VIH-l a été obtenue par résonance magnétique nucléaire (Lodi *et al.*, 1995). Il comporte 5 brins β et présente une structure similaire à un domaine SH3[4]. Il est présent sous forme de dimères avec les brins β 2, 3 et 4 impliqués dans l'interface. Plus récemment, la structure d'une Intégrase comportant le domaine catalytique et le domaine C-terminal a été résolue par cristallographie aux rayons X (Figure 16 ; Chen *et al.*, 2000). Seuls les domaines centraux se dimèrisent, tandis que les deux domaines C-terminaux se retrouvent de part et d'autre des domaines centraux.

A ce jour, aucune structure de la protéine Intégrase entière n'a été obtenue. Comme nous venons de le voir, seules les structures des domaines pris séparément ou des domaines pris deux à deux (N-terminal + domaine catalytique ou domaine catalytique + C-terminal) ont été résolues.

[4] Les régions SH3 (Src Homology 3) sont des domaines d'interaction protéine-protéine présentant une affinité pour des régions riches en prolines.

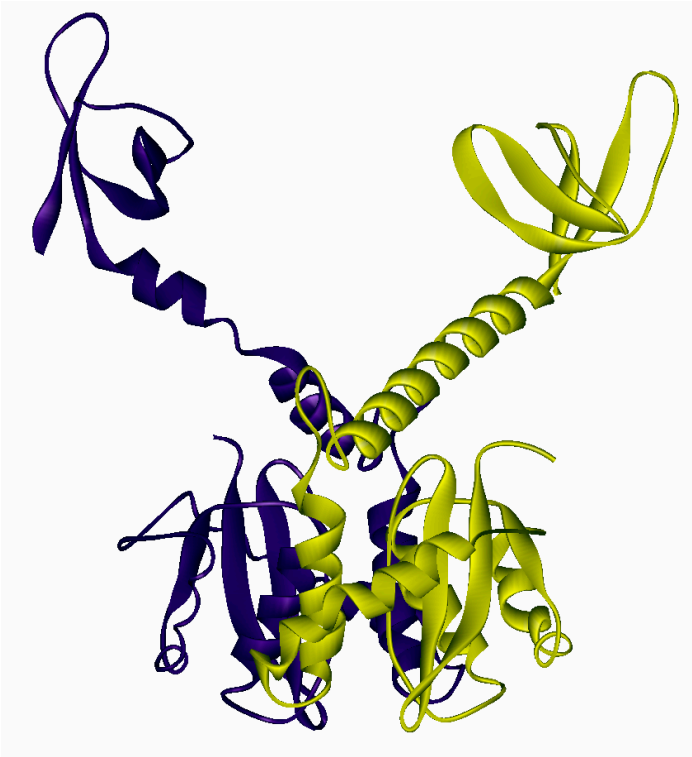


Figure 16 : *Représentation en ruban du dimère des domaines centraux + domaines C-terminaux de l'Intégrase du VIH-1.*

Ces études structurales ainsi que des expériences réalisées *in vitro* ont montré que l'Intégrase était fonctionnelle au minimum sous forme de dimères (Engelman *et al.*, 1993). Certains auteurs suggèrent qu'elle fonctionnerait plus probablement sous forme de tétramères (Faure *et al.*, 2005) voire sous la forme d'octamères (Deprez *et al.*, 2000).

3. Interaction avec la protéine Nef du VIH-1 :

La possibilité de recrutement de EED au niveau de la membrane plasmique par Nef a récemment été envisagée (Witte *et al.*, 2004).

La protéine EED a été identifiée comme étant un partenaire de la protéine Nef, en utilisant le système du double hybride avec l'hélice α du domaine N-terminal de Nef comme appât. Ce résultat posait problème, car EED était jusqu'alors considérée comme une protéine strictement nucléaire (Sewalt *et al.*, 1998) alors que Nef est cytoplasmique. La capacité de EED de faire la navette entre le noyau et le cytoplasme a été démontrée par micro-injection et «heterocaryon assays» par Witte et collaborateurs. Cette relocalisation semble être induite par l'expression de Nef. Ainsi selon cette équipe, la levée de l'inhibition de EED sur le LTR du

VIH-1 par Nef pourrait constituer un signal de dérépression permettant une transcription Tat-dépendante efficace (Figure 17).

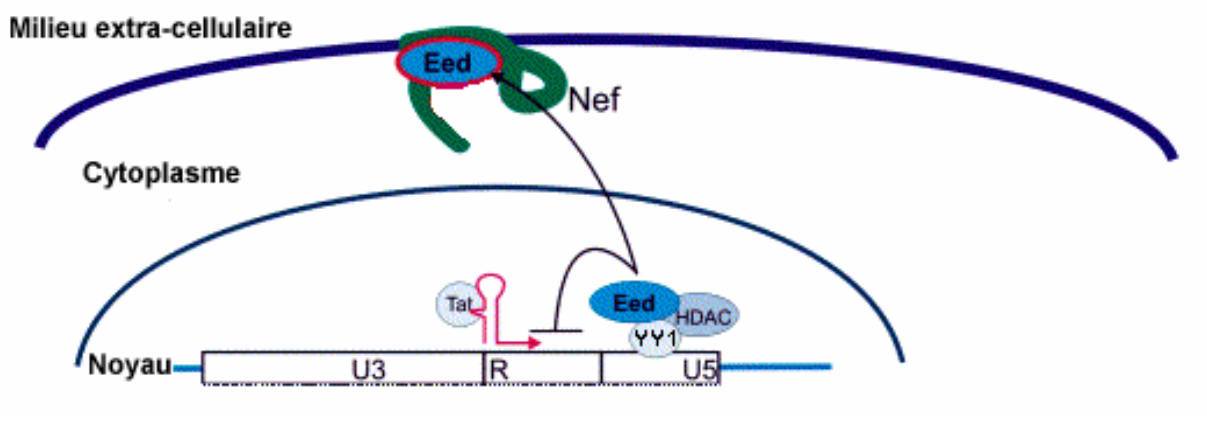


Figure 17: *Dérépression par Nef : A la membrane, le domaine N-terminal accessible de Nef recrute EED. Dans le noyau, EED exerce une répression sur le LTR en se liant à la région promotrice en formant un complexe multiprotéique avec d'autres facteurs tels la protéine YY1 (cf. Etude bibliographique) et l'histone désacétylase HDAC1. Ce complexe bloque l'élongation de la transcription par Tat. La sortie du noyau de EED entraîne le désassemblage du complexe et la levée de la répression de la transcription (d'après Witte et al., 2004).*

La protéine Nef est l'une des quatre protéines accessoires du virus de l'immunodéficience humaine (VIH-1 et VIH-2) et simienne (VIS). Ces protéines ne sont pas indispensables à la croissance virale et la protéine Nef avait initialement été proposée comme étant un facteur négatif (Nef = «Negatif factor») pour le développement viral. Néanmoins, plusieurs études ont depuis montré que Nef est un déterminant important du pouvoir pathogène *in vivo :* un mutant du virus VIS dont le gène *nef* a été supprimé ne cause pas de SIDA chez des macaques adultes ; une telle souche virale a été utilisée avec succès comme vaccin contre le virus sauvage infectieux (Daniel *et al.*, 1992). Le rôle critique de Nef est également confirmé par des études sur des porteurs sains, asymptomatiques à long terme, qui ont été infectés avec une souche virale caractérisée par des délétions dans le gène *nef* (Deacon *et al.*, 1995).

Il est maintenant établi que les effets biochimiques de Nef sont liés à son interaction fonctionnelle avec des molécules cellulaires. Plus d'une trentaine de cibles putatives de Nef ont été identifiées, la plupart de ces partenaires étant impliqués dans la transduction des signaux, l'activation des cellules T et la médiation de la réponse immunitaire.

Trois effets peuvent en effet être attribués à l'action de Nef: (i) la diminution de l'expression du récepteur membranaire CD4, qui est aussi le récepteur principal du virus et du complexe

majeur d'histocompatibilité de type 1 (CMH-1) (ii) l'augmentation de l'infectivité des virions et (iii) la perturbation des voies de signalisation dans les cellules infectées.

Structure de Nef du VIH-1

La protéine Nef du virus VIH-l a une masse moléculaire d'environ 27 kDa et est myristoylée sur sa glycine N-terminale (Figure 18). Elle peut-être clivée *in vitro* après le résidu W57 par la protéase virale (Welker *et al.*, 1996). La fonction de ce clivage n'est pas connue et ne semble pas être corrélée au pouvoir infectieux du virus (Chen *et al.*, 1998).

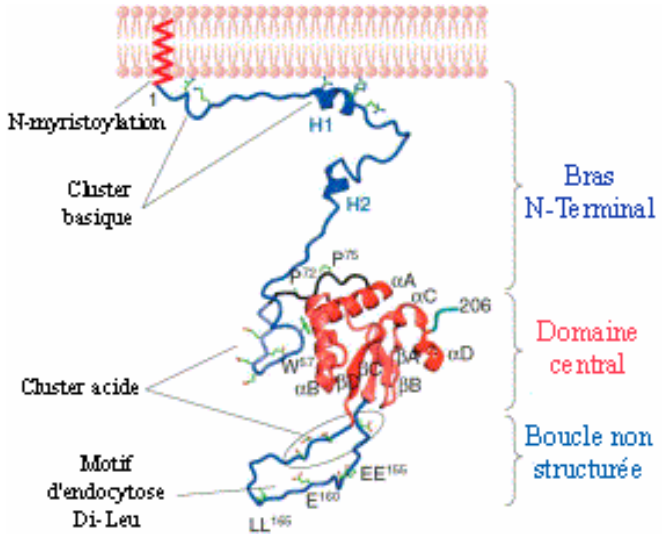


Figure 18 : *Modèle moléculaire de Nef entière ancrée à la membrane. Ce modèle a été construit à partir de la structure RMN du domaine N-terminal myristoylé (résidus 2-57) et de la structure cristallographique du domaine central (résidus 56-206 ; d'après Arold et al., 2001).*

La partie N-terminale comprend une grande boucle d'une centaine de résidus qui n'est pas structurée dans la forme libre de Nef. Un peptide correspondant aux résidus 1-25 de Nef a été caractérisé par RMN (Barnham *et al.*, 1997). Cette région n'est pas structurée en milieu aqueux, alors que les résidus 6-22 forment une hélice α dans le méthanol. Ceci pourrait indiquer une structuration partielle en milieu apolaire, par exemple en interaction avec une membrane (Figure 18).

La structure du coeur fonctionnel de Nef du VIH-l (résidus 58-206) a été déterminée par RMN (Grzesiek *et al.*, 1996 ; Grzesiek *et al.*, 1997) et par cristallographie (Arold *et al.*, 1997). La structure du coeur fonctionnel de Nef a également été résolue par cristallographie

en complexe avec le domaine SH3 d'une tyrosine kinase (Fyn ; Lee *et al.*, 1996). Un motif du type PxxP, strictement conservé dans tous les isolats, est impliqué dans l'interaction de Nef avec ce domaine SH3. Les modèles du coeur fonctionnel de Nef libre montrent que la moitié N-terminale de la région du motif PxxP (résidus 71-73) n'est pas structurée (modèle cristallographique) ou plus mobile que le reste du coeur (modèle RMN). Les différences dans les modèles sont probablement liées à la longueur différente des constructions utilisées. Dans les structures cristallographiques de Nef en complexe avec un domaine SH3 la région entière contenant le motif PxxP (résidus 71-77) est structurée en hélice gauche de type polyproline II (Lee *et al.*, 1996). Malgré une énergie entropique défavorable pour la structuration, un motif PxxP flexible de Nef pourrait avoir un avantage en permettant à la protéine virale de s'adapter à la surface des différents domaines SH3, permettant ainsi de cibler plusieurs kinases.

Une étude récente suggère une dimérisation de Nef grâce à des interactions électrostatiques impliquant l'Arginine 105 et l'acide Aspartique 123 (Liu *et al.*, 2000). L'analyse des structures cristallographiques fournit également une base moléculaire à cette dimérisation par l'intermédiaire de l'hélice α4 et de sa boucle adjacente (Arold *et al.*, 2000).

Résultats et discussion

A/ ETUDE FONCTIONNELLE DE EED DANS LE CYCLE VIRAL DU VIH-1

Publication 1 : The Human Polycomb Group EED protein interacts with the Integrase of Human Immunodeficiency Virus type 1

La première partie de mes recherches sur EED s'est déroulée dans le laboratoire de Virologie et de Pathogenèse Virale du Professeur P. Boulanger qui avait mis en évidence l'interaction entre EED et la Matrice (Peytavi *et al.*, 1999) et suggéré celle entre EED et l'Intégrase (Peytavi, 1999). Mon travail consistait à confirmer cette nouvelle interaction entre EED et IN et à en caractériser les modalités aussi bien *in vivo* que *in vitro*. Tous ces résultats ont fait l'objet d'une publication (Publication 1 ; Violot *et al.*, 2003).

1. Résumé de la Publication 1 :

Cet article décrit des expériences de double hybride en système levure qui permettent de confirmer *in vivo* l'existence d'une interaction EED-IN. De plus, des résultats obtenus *in vitro* par mutagenèse, essais «pull-down» et «phage-biopanning», suggèrent que le site d'interaction de l'Intégrase avec EED se situe dans domaine C-terminal de IN, entre les résidus 212 à 264. De même, ces expériences de «phage-display» montrent que les régions putatives d'interaction de EED avec IN se situent d'une part entre les résidus 96 à 105, et d'autre part entre les résidus 224 à 232. Ces deux zones d'interaction sont distinctes de celles mises en évidence pour la Matrice qui a été localisée entre les résidus 294 à 309 (Peytavi *et al.*, 1999). Enfin, il a été montré que l'interaction EED-IN nécessitait l'intégrité des deux motifs répétés WD-40 C-terminaux de EED.

J'ai plus particulièrement participé aux tests d'intégration *in vitro* (cf. Annexe C pour le principe) qui montrent que EED devrait avoir un effet activateur dose-dépendant sur la réaction d'intégration de l'ADN viral réalisée par l'Intégrase. Cet effet pourrait être indirect : en effet, EED pourrait par son interaction avec IN, faciliter l'oligomérisation de cette dernière et favoriser de ce fait le processus d'intégration.

Enfin, l'article décrit des études de la distribution cellulaire de l'Intégrase et de EED dans des cellules infectées par le VIH-1 (cellules HeLa $CD4^+$ ou cellules lymphoïdes MT4). Ces études ont été réalisées *in situ* par immuno-électromicroscopie et montrent une co-localisation de

EED et IN dans le noyau et à proximité des pores nucléaires. Cette co-localisation est observable au cours des phases précoces de l'infection, c'est-à-dire entre 1 h 30 et 6 h après l'infection des cellules. En outre, une triple co-localisation EED, IN et MA a pu être observée dans le nucléoplasme jusqu'à 6 h après l'infection, suggérant ainsi l'existence d'un complexe multiprotéique au stade précoce du cycle viral (complexe de pré-intégration ?). De tels phénomènes n'ont pas pu être observés avec un virus VIH-1 non infectieux et dépourvu de son enveloppe.

Le fait qu'aucune co-localisation EED / IN n'ait pu être observée en présence d'AZT au cours des études par immuno-électromicroscopie (Inhibiteurs Nucléosidiques de la Transcriptase Inverse) suggère également que le rôle joué par EED au cours du cycle viral nécessite au préalable l'étape de rétro-synthèse de l'ADN proviral.

2. Discussion sur la Publication 1 :

Le fait que EED soit un partenaire de deux protéines virales, (i) la Matrice qui possède un rôle structural et fonctionnel au cours des phases précoces du cycle viral (Kiernan *et al.*, 1998) et (ii) l'Intégrase qui est responsable de l'intégration du provirus au sein du génome hôte (Bushman et Craigie, 1991) laisse supposer que EED pourrait participer au moins à deux étapes majeurs du cycle viral :

- Le transport intracellulaire des virions
- L'intégration du provirus

2.a) Rôle possible de EED dans le transport intracellulaire des virions :

La présence en double marquage de EED et de IN à proximité du complexe du port nucléaire mis en évidence par nos expériences d'immuno-électromicroscopie, suggère un rôle de navette pour EED. Elle pourrait donc jouer un rôle dans le transport intracellulaire et la translocation nucléaire du complexe de pré-intégration (PIC). En effet, en tant que constituant putatif de ce complexe multifactoriel de transport, EED pourrait par son interaction avec IN et MA jouer le rôle d'une navette permettant le convoyage du PIC jusque dans le noyau.

A ce jour, seul un petit nombre de protéines non-virales semblent être impliquées dans le transport du PIC.

De manière intéressante, une protéine identique à EED, la protéine WAIT-1 (WD protein associated with integrin cytoplasmic tails-1) a été montrée comme interagissant avec les domaines cytoplasmiques des sous-unités des intégrines α4, αE et β7 (Rietzler *et al.*, 1998). Il a également été suggéré que WAIT-1 pouvait faire la navette entre les intégrines associées à la membrane et le noyau. Il est à noter que parmi la sous-famille β7, les intégrines αEβ7 ne se retrouvent que dans les lymphocytes T et les cellules dendritiques.

2.b) Rôle possible de EED dans le processus d'intégration du provirus :

Malgré de nombreuses études, les facteurs cellulaires et viraux contrôlant la réaction et les sites d'intégration de l'ADN proviral demeurent uniquement partiellement élucidés.

Les mécanismes moléculaires de l'intégration sont bien connus dans le cas des rétro-transposons. Par exemple Ty5 s'intègre chez la levure dans des régions silencieuses de la chromatine grâce à la protéine cellulaire Sir (Zou et Voytas, 1997). Comme cette protéine Sir chez la levure, de nombreuses protéines de la famille des Polycomb Group sont responsables chez les organismes eucaryotes supérieurs du maintien dans un état silencieux de la chromatine. Ceci est généralement réalisé par le recrutement d'histones désacétylases (HDCA). En effet, le rôle de répresseur transcriptionnel de EED a récemment été démontré comme impliquant une désacétylation des histones (cf. Etude bibliographique, (van der Vlag et Otte, 1999)). De même, EED a été montrée comme co-localisant avec l'histone H1 dans des régions transcriptionnellement inactives de l'hétérochromatine périnucléaire de neurones murins (Akhmanova *et al.*, 2000). Cette fonction de répresseur de la transcription des protéines du groupe des Polycomb a généralement été attribuée à une interaction avec des facteurs de transcription plutôt qu'à une interaction directe avec l'ADN. Par exemple, il a été montré que EED prend part à la fonction biologique du complexe PRC2 (Polycomb Repressive Complex 2) qui comporte les protéines EZH2 (Enhancer of Zeste 2) SUZ12 (Supressor of Zeste 12) et les protéines de fixation aux histones RbAp46 et 48 (Cao *et al.*, 2002). Ces résultats démontrent une fixation indirecte de EED sur les régions cible de la chromatine *via* le complexe PRC2.

Schroder et collaborateurs ont récemment localisé plus de 500 évènements d'intégration du VIH-1 dans des lignées de cellules T humaines infectées par le VIH-1, révélant ainsi que le processus d'intégration se produit préférentiellement dans des gènes hautement transcrits par

la polymérase III (Schroder *et al.*, 2002). Cette spécificité pourrait entraîner une transcription plus efficace des gènes viraux, favorisant la propagation virale au détriment de la survie de la cellule hôte. Un tel tropisme préférentiel le long du génome hôte, et ceci en l'absence de séquence spécifique, suggère que le processus d'intégration puisse être influencé soit par des interactions spécifiques entre des composants viraux et des protéines cellulaires soit par une architecture spécifique de la chromatine. Ainsi, plusieurs protéines cellulaires se fixant à l'ADN ont été décrites comme interagissant avec l'Intégrase et pourraient constituer de bons candidats permettant l'adressage du PIC vers les sites d'intégration. La protéine INI1 (Integrase Interactor 1, également appelée hSNF5) isolée en double hybride levure a été proposé comme favorisant *in vitro* l'étape de ligation et comme étant un facteur déterminant du ciblage du génome virale vers des sites privilégiés du génome hôte (Kalpana *et al.*, 1994). De la même manière, la protéine HMG-I (Y), une protéine chromosomique non-histone impliquée dans le contrôle transcriptionnel et l'architecture du chromosome, et la protéine BAF (Barrier-to-Autointegration Factor ; Farnet et Bushman, 1997 ; Mansharamani *et al.*, 2003), une protéine cellulaire impliquée dans la réorganisation post-méiotique du nucléole, ont été identifiées comme des partenaires de IN. Ces deux protéines semblent nécessaires à une intégration efficace *in vitro*, mais leur rôle respectif pour le ciblage du PIC reste non évalué.

Au vu de ces résultats, une hypothèse séduisante voudrait que l'interaction dans le noyau de EED avec IN et / ou MA puisse dérégler le maintien de l'état silencieux de la chromatine en libérant EED de son interaction avec le complexe PRC2 (ou tout autres complexes impliqués dans le processus de répression génique).

Publication 1

The Human *Polycomb* Group EED protein interacts with the Integrase of Human Immunodeficiency Virus type 1

Sébastien Violot, Saw See Hong, Dina Rakotobe, Caroline Petit, Bernard Gay, Karen Moreau, Geneviève Billaud, Stéphane Prillet, Joséphine Sire, Olivier Schwartz, Jean-François Mouscadet and Pierre Boulanger

JOURNAL OF VIROLOGY, Dec. 2003, p. 12507–12522 Vol. 77, No. 23
0022-538X/03/$08.00+0 DOI: 10.1128/JVI.77.23.12507–12522.2003

The Human *Polycomb* Group EED Protein Interacts with the Integrase of Human Immunodeficiency Virus Type 1

Sébastien Violot,[1] Saw See Hong,[1] Dina Rakotobe,[1] Caroline Petit,[2]† Bernard Gay,[1]‡ Karen Moreau,[1]§ Geneviève Billaud,[1] Stéphane Priet,[3] Joséphine Sire,[3] Olivier Schwartz,[2] Jean-François Mouscadet,[4] and Pierre Boulanger[1]*

Laboratoire de Virologie and Pathogénèse Virale, Faculté de Médecine RTH Laennec, CNRS UMR-5537 and Université Claude Bernard Lyon 1, 69372 Lyon Cedex 08,[1] Laboratoire Rétrovirus et Transfert Génétique, Institut Pasteur, 75724 Paris Cedex 15,[2] Unité de Pathogénie des Infections à Lentivirus, INSERM U-372, 13276 Marseille Cedex 09,[3] and CNRS UMR-8532, Ecole Normale Supérieure, 94235 Cachan Cedex,[4] France

Received 21 April 2003/Accepted 23 August 2003

Human EED, a member of the superfamily of WD-40 repeat proteins and of the *Polycomb* group proteins, has been identified as a cellular partner of the human immunodeficiency virus type 1 (HIV-1) matrix (MA) protein (R. Peytavi et al., J. Biol. Chem. 274:1635-1645, 1999). In the present study, EED was found to interact with HIV-1 integrase (IN) both in vitro and in vivo in yeast. In vitro, data from mutagenesis studies, pull-down assays, and phage biopanning suggested that EED-binding site(s) are located in the C-terminal domain of IN, between residues 212 and 264. In EED, two putative discrete IN-binding sites were mapped to its N-terminal moiety, at a distance from the MA-binding site, but EED-IN interaction also required the integrity of the EED last two WD repeats. EED showed an apparent positive effect on IN-mediated DNA integration reaction in vitro, in a dose-dependent manner. In situ analysis by immunoelectron microscopy (IEM) of cellular distribution of IN and EED in HIV-1-infected cells (HeLa CD4$^+$ cells or MT4 lymphoid cells) showed that IN and EED colocalized in the nucleus and near nuclear pores, with maximum colocalization events occurring at 6 h postinfection (p.i.). Triple colocalizations of IN, EED, and MA were also observed in the nucleoplasm of infected cells at 6 h p.i., suggesting the ocurrence of multiprotein complexes involving these three proteins at early steps of the HIV-1 virus life cycle. Such IEM patterns were not observed with a noninfectious, envelope deletion mutant of HIV-1.

Integration of the proviral DNA into the host cell genome has been considered an obligatory step of the human immunodeficiency virus type 1 (HIV-1) life cycle, as for all known retroviruses (20, 35). The host DNA sites for retroviral integration have been long debated. Although it has been reported that integration events are statistically more frequent in inactive than active chromatin (11, 81, 82), other data suggest that retrovirus integration preferentially occurs in transcriptionally active chromatin (65) and in regions of DNA distortion (57–59). Recent studies have confirmed that transcriptional units are preferred targets for both murine leukemia virus (MLV) and HIV-1 (66, 83), but refined mapping of integration sites has revealed significant differences between HIV-1 and MLV, the latter preferentially integrating near the start of transcription (83). Whatever the scenario followed by the different types of retroviruses in their host cells, it is reasonable to assume that proteins which control the state of cellular chromatin would play a key role in the integration of HIV proviral DNA and, possibly, in further steps of the virus life cycle. The viral RNA genome is retrotranscribed by the reverse transcriptase (RT) into a double-stranded DNA molecule, which is the substrate for the viral integrase (IN) in the integration reaction. The RT reaction starts within the central core of extracellular virions, before they bind to the surface of the host cell, and is completed after the virus enters the cell (reviewed in references 5, 30, and 35). The final reaction takes place after partial uncoating of the HIV-1 particle, within a protein-nucleic acid complex termed preintegration complex (PIC). PIC competent for integration reaction in vitro has been isolated from HIV-1-infected cells (20, 43, 48, 49).

It is generally recognized that PIC is composed of linear double-stranded DNA, RT, IN, matrix protein (MA), nucleocapsid protein (NC), and auxiliary protein Vpr (reviewed in references 13, 16, 30, and 71). One of these components, the MA protein, has been identified as an effector of various biological functions in the virus life cycle (reviewed in reference 23). Since the MA carries a potential nuclear localization signal, it has been hypothesized to mediate the transport of PIC to the cell nucleus and its traverse of the nuclear pore complex (6, 7, 26, 27). However, the observation that a mutant HIV-1 virus with most of the MA domain deleted was still capable of infecting nondividing cells (60) argued against a major role of the basic MA sequences in the nuclear import of the PIC, and other data seem to assign this function to the Vpr protein (13,

* Corresponding author. Mailing address: Laboratoire de Virologie and Pathogénèse Virale, Faculté de Médecine RTH Laennec de Lyon, 7, Rue Guillaume Paradin, 69372 Lyon Cedex 08, France. Phone: 33-4-7877-8621. Fax: 33-4-7877-8751. E-mail: Pierre.Boulanger@laennec.univ-lyon1.fr.

† Present address: Institut Cochin de Génétique Moléculaire, 75014 Paris, France.

‡ Present address: Laboratoire Infections Rétrovirales et Signalisation Cellulaire, CNRS UMR 5121, Institut de Biologie, 34060 Montpellier, France.

§ Present address: Laboratoire Génétique et Cancer, CNRS UMR 5641 and Université Claude Bernard Lyon 1, 69373 Lyon Cedex 08, France.

16, 22, 71), IN (2, 52), and the central DNA flap (18, 85). Several cellular proteins such as high-mobility-group proteins (19, 28), the barrier-to-autointegration factor (41), IN interactor 1 (Ini1) (40, 48, 84), and human lens epithelium-derived growth factor (LEDGF/p75) (45), have been identified as partners of IN and/or cofactors of integration reaction, and at least one of them, the barrier-to-autointegration factor, has been found to be associated with PIC of MLV and HIV-1 (43, 77). These proteins could participate, directly or indirectly, in the cellular trafficking of PIC and/or in provirus integration.

One of the cellular partners of the HIV-1 MA protein has been identified as EED (54), the human homolog of the mouse embryonic ectoderm development (*eed*) gene product (51, 67, 68, 74), and a member of the WD-40 repeat superfamily of proteins (50). The *eed* gene is highly conserved in humans and mice and is also homologous to the *Drosophila esc* gene. Both *eed* and *esc* belong to the family of widely conserved *Polycomb* group (*Pc*-G) of genes (17, 51). EED and many other proteins of the *Pc*-G family act as transcriptional repressors and gene silencers (1, 3, 17, 64, 67, 68, 74). The product of the *Pc*-G genes in eukaryotes are involved in chromatin remodeling and, in particular, in the maintenance of the silent state of chromatin (reviewed in reference 55) and the reduction of DNA accessibility (21). EED has also been found to colocalize with histone H1 in heterochromatin subdomains of neuronal cell nuclei containing inactive DNA (1).

Previous studies have indicated that MA, RT, and IN proteins are each a component of HIV-1 PIC (26, 27, 49). Since MA has been found to bind to cellular EED (54), we explored the possibility of a protein-protein interaction between EED and IN and of the occurrence of a ternary complex of EED, MA, and IN. We found that IN and EED proteins were indeed able to interact in vitro, as well as in vivo, in yeast. In human cells infected with HIV-1 (HeLa CD4$^+$ cells or MT4 lymphoid cells), in situ analysis by immunoelectron microscopy (IEM) showed that EED, IN, and MA proteins colocalized within the nucleus, with a maximum number of colocalization events occurring at 6 h after infection. Such patterns of nuclear colocalization were not observed in cells incubated with a noninfectious, envelope deletion mutant of HIV-1. Our data suggest that EED, a cellular partner of both MA and IN, would play a functional role in the HIV-1 life cycle, acting at early steps of virus infection.

MATERIALS AND METHODS

Cells, HIV-1 viruses and virus infection. Human embryonic kidney cells (HEK-293) and epithelial cells (P4P56) (HeLa CD4$^+$, *lck*+, LTR-*LacZ*) (52) were grown in Dulbecco modified Eagle medium (Gibco), supplemented with glutamine, antibiotics, and 10% fetal calf serum. MT4 lymphoid cells were grown in RPMI medium with the same supplements. The construction of HIV-1 BRU infectious molecular clone expressing the viral IN fused to the Flag epitope has been described in a previous study (53). It is abbreviated as BRU-FlagWT in the present study. BRU-FlagΔEnv mutant was derived from the BRU-Flag virus, by a deletion in the *env* gene (*Kpn*I-*Bgl*II, nucleotides 6343 to 7600 of the sequence of the NL43/2 clone). Viruses were produced by transfection of the corresponding plasmids as previously described (69). Supernatants were analyzed for HIV-1 p24 antigen content, and infectivity titers are expressed here in nanograms of p24 per milliliter (53). Aliquots of MT4 and P4P56 cell samples (4×10^6 cells) were infected at a multiplicity of infection of 150 ng of p24 antigen per 10^6 cells, i.e., ca. 1,000 virus particles per cell, considering the latest estimate of 4,000 to 5,000 copies of CA per HIV-1 virion. In negative control experiments, cells were treated with 15 μM zidovudine (AZT), which was added 3 h before infection and maintained throughout the infection period. Cells were harvested at 0 (mock-infected cells), 1.5, 6, and 24 h postinfection (p.i.); fixed in 0.5% glutaraldehyde–4% paraformaldehyde; and processed for IEM.

Yeast two-hybrid assays. (i) Generation of the DNA-binding LexA-IN hybrid (Fig. 1A). The portion of the HIV-1_{BRU} *pol* gene segment coding for the IN was cloned in frame with the DNA-binding *LexA* gene into the pBTM116 vector (47). The sequence of the construct was verified by DNA sequencing, and the expression of recombinant fusion protein in yeast was confirmed by gel electrophoresis and immunoblot analysis with anti-IN polyclonal antibody as described below. (ii) Generation of the pGAD-EED (Fig. 1A). The cDNA of wild-type (WT) EED (54) was inserted into the *Eco*RI and *Not*I sites of the Gal4 transcription activation domain vector pGAD3S2X, a modified version of the pGAD GH (Clontech) containing a *Not*I site in its polylinker. This resulted in a Gal4 activation domain AD-EED hybrid. In the hybrid mutant AD-EED-305A4 the tetrapeptide motif HNRY from positions 305 to 308 was replaced by the tetrapeptide AAAA, and in AD-EED-300AI the dipeptide motif ST at positions 300 and 301 was replaced by AI. (iii) The two-hybrid assays were performed as described in a previous study (54).

Bacterially expressed IN and EED recombinant proteins. The full-length, WT HIV-1 BRU IN, and various deletion mutants of IN were expressed in bacteria as glutathione *S*-transferase (GST) fusion proteins (GST-IN) (56). Their schematic structure and domains are presented in Fig. 1B, along with their acronyms. Likewise, full-length WT human EED was expressed as a GST fusion protein (GST-EED-441) by using the pGEX-KG plasmid (32). EED C-terminal deletion mutant was generated by insertion of a TGA stop codon at position 349 of GST-EED-441, yielding mutant protein GST-EED-C348. In the substitution mutant GST-EED-103A3, the tripeptide motif WHS from positions 103 to 105 was replaced by the tripeptide AAA. Various forms of EED and IN were also also expressed as His_6-tagged proteins (Fig. 1B) by using the pT7-7 IPTG-inducible promoter (14). This was the case for the full-length versions of EED (tagged at its C terminus and abbreviated EED-441-H6) and of IN (tagged at its N terminus and abbreviated H6-IN-WT) and for the N and C terminally deleted versions of IN (abbreviated H6-IN-ΔN and H6-IN-ΔC, respectively). Mutagenesis was performed by using PCR and splicing by overlap extension (34, 38). Mutant sequences were verified by DNA sequencing.

Purification of recombinant tagged proteins and affinity pull-down assays. Bacterial cells expressing the desired protein were centrifuged and resuspended at 1 to 2 g (wet weight) per 2 to 5 ml of lysis buffer (10 mM Tris-HCl [pH 8.0], 5 mM $MgCl_2$, 1 mM EDTA, 1% Triton X-100, 1 mM phenylmethylsulfonyl fluoride) and then disrupted by sonication. After incubation for 30 min in the presence of a broad-spectrum endonuclease (*Serratia marcescens* Benzonase; Sigma) at 100 U/ml of sample, the lysates were clarified by centrifugation at 100,000 × *g* for 45 min. GST fusion proteins were purified by adsorption on a glutathione-Sepharose gel and recovered by elution with a glutathione-containing buffer or by thrombin cleavage of the GST linker by using a commercial kit (Bulk GST Purification Module; Pharmacia Biotech). For purification of His-tagged proteins, the bacterial cell lysate supernatant was applied to a Ni^{2+}-nitrilotriacetic acid (NTA)-agarose column (Qiagen, Inc.) equilibrated in sodium phosphate buffer (SPB; 16 mM Na_2HPO_4, 4 mM NaH_2PO_4 [pH 6.8]) containing 0.5 M NaCl (SPB-500) and 1% Triton X-100. Unbound and weakly adsorbed proteins were washed from the affinity gel with 10 column volumes of SPB-500 containing 5 and 30 mM imidazole, successively, and then His-tagged proteins were eluted at 250 mM imidazole in SPB-500. For in vitro binding reactions, 100-ml cultures of *Escherichia coli* expressing GST fusion protein were centrifuged, resuspended in 400-μl aliquots of binding buffer (BB; phosphate-buffered saline containing 0.2% Nonidet P-40 and a cocktail of protease inhibitors [Boehringer Mannheim]), sonicated, and clarified by centrifugation. Aliquots (20 μl) of a glutathione-Sepharose bead suspension were added to the bacterial cell lysates, followed by incubation for 30 min at 4°C. After an extensive rinsing in BB, the affinity beads were mixed with 100 to 200 μl (100 μg of total protein) of cell lysates (from bacteria or insect cells) containing nontagged or differentially tagged partner protein and 900 μl of BB, followed by incubation for 2 h at room temperature. The beads were then washed three times with 500 μl of BB, resuspended, and boiled in 50 μl of sodium dodecyl sulfate (SDS) sample buffer. Alternatively, His-tagged proteins were retained on Ni^{2+}-NTA-agarose beads, and affinity beads were processed as described above by using a magnetic pull-down device (Qiagen). Protein pull-down experiments were also performed with affinity-purified proteins, e.g., GST-IN plus EED-441-H6 or H6-IN-WT plus GST-EED. Pull-down proteins were analyzed by SDS-polyacrylamide gel electrophoresis (PAGE) and immunoblotting.

Gel electrophoresis and immunoblotting analysis. PAGE of SDS-denatured protein samples and immunoblotting analysis have been described in detail in previous studies (8, 12, 54). Briefly, proteins were electrophoresed in SDS–

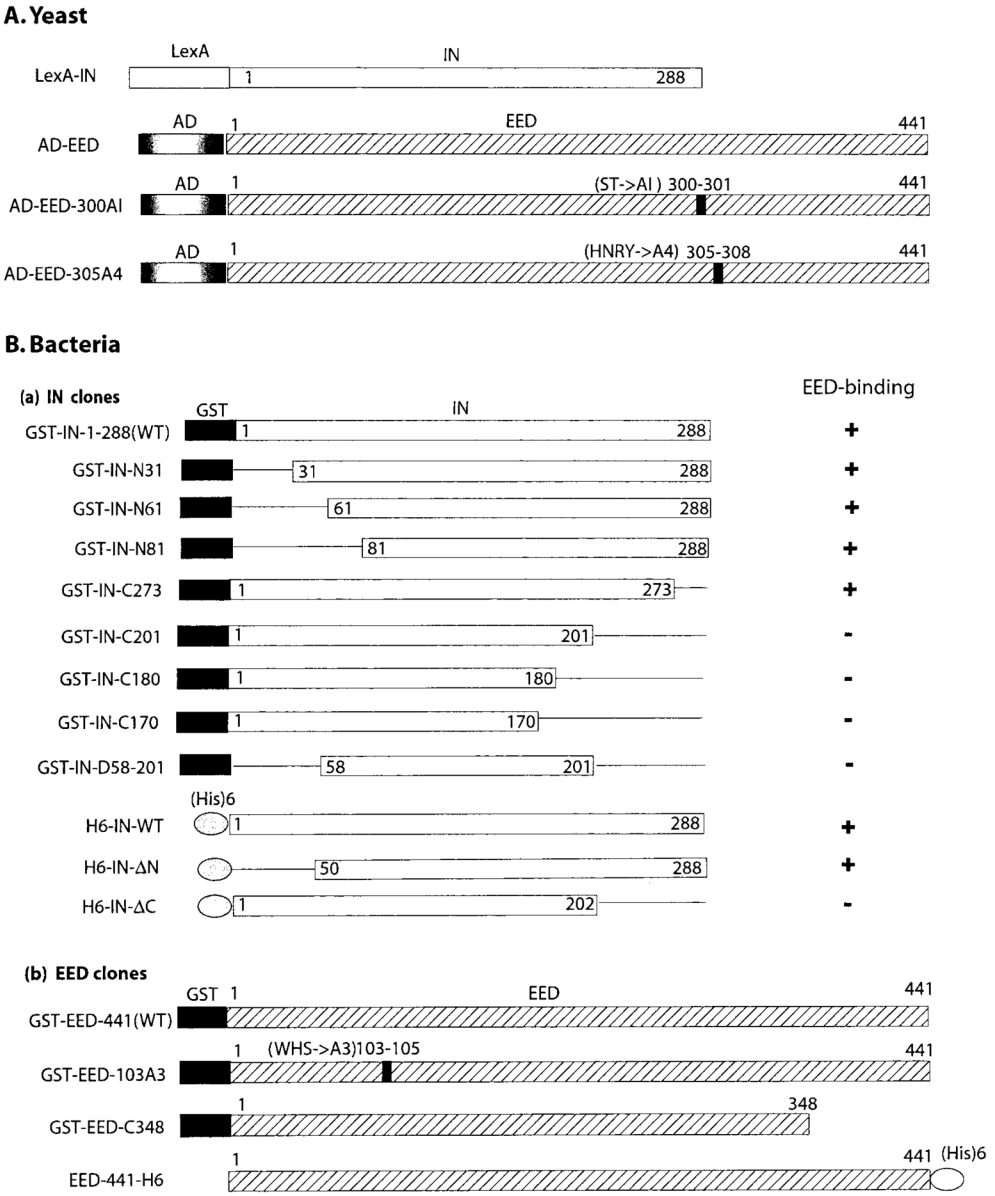


FIG. 1. Schematic representation of genetic constructs of HIV-1 IN and EED, expressed in yeast (A) or in bacteria (B). The LexA (DNA-binding) domain and the Gal4 activating domain (AD) are shown by gray boxes, the GST domain by a black box, the His_6 tag by a gray ellipse, the IN sequence by an open box, and EED by a hatched box. In panel Ba, deletions in the IN sequence are represented by a solid line. In panels A and Bb, the point mutations in the EED protein are shown as black bars, with the mutated amino acids indicated above. On the right of panel Ba, the binding of EED to IN protein (WT or deletants) is indicated by "+" or a "−."

denaturing, 10% polyacrylamide gel and electrically transferred to nitrocellulose membrane. Blots were blocked in 5% skimmed milk in Tris-buffered saline (TBS) containing 0.05% Tween 20 (TBS-T), rinsed in TBS-T, and then successively incubated with primary rabbit or mouse antibody (working dilutions of 1:1,000 to 1:2,000) and the complementary peroxidase- or phosphatase-labeled anti-immunoglobulin G (IgG) conjugate (1:1,000).

Antibodies. For detection of EED, IN, and MA proteins by blotting, immunofluorescence, or IEM, the following antibodies were used. Antiserum against EED was prepared in rabbits by injection of affinity-purified EED protein (54). Anti-IN rabbit polyclonal antibody was also laboratory-made (42), as well as our anti-GST rabbit antiserum. For Flag-tagged IN (53), mouse monoclonal antibody (MAb) anti-Flag M2 (Kodak) or rabbit polyclonal anti-Flag (Zymed Laboratories, South San Francisco, Calif.) were used. MA protein was detected with MAb anti-MAp17 (Epiclone 5003 [Cylex, Inc., Columbia, Md.] mapped to epitope 121-DTGHSSQVSQNY-132) (12) or rabbit polyclonal anti-MA antibody (laboratory-made) (39). His-tagged proteins were detected by using monoclonal anti-His.Tag antibody (Novagen) specific for N-terminal, C-terminal, and internal histidine clusters.

Phage-displayed peptide libraries and biopanning. Two filamentous phage-displayed peptide libraries were used, one consisting of random hexapeptides

(kindly provided by G. Smith [75]), the other of dodecapeptides (BioLabs). Biopanning of immobilized protein ligate and specific ligand elution of phages have been described in previous studies (36, 37, 39). Purified protein (e.g., IN) was immobilized on enzyme-linked immunosorbent assay plates used as the solid support. Elution of phages was carried out by using the purified partner protein as a soluble competing ligand (in this case, EED). In the reverse biopanning experiment, purified EED was coated onto the plate, and IN was used as the soluble competing ligand. The hexapeptide or dodecapeptide phagotopes were identified by manual DNA sequencing of the recombinant fUSE5 pIII protein (36, 37) by using the dideoxynucleotide chain termination method, the required oligonucleotide primers, and the Sequenase kit version 2.0 (Amersham Biosciences).

In vitro HIV-1 IN assays. Recombinant, bacterially expressed, N-terminally His-tagged HIV-1 IN protein was purified by affinity chromatography (44). IN-mediated strand transfer was assayed in vitro by using a restriction fragment from a modified version of plasmid pU3U5 ((9) as the IN substrate (donor DNA) and plasmid pBSK-*zeo* (a 3-kbp pBSK-derived construct in which the *amp* gene has been replaced by the *zeocine-resistance* gene) as target DNA. The donor DNA (300-bp fragment) contained the 20 terminal base pairs of the HIV-1 5′ and 3′ long terminal repeat (LTR) at each ends, with two-base 5′ overhangs generated by *Nde*I cleavage. Integration reactions (20 μl, final volume) were performed by using 10 ng of ^{32}P-labeled donor DNA, 100 ng of target DNA, and 2 to 5 pmol of IN (200 to 500 nM, final concentration) in 20 mM HEPES (pH 7.5) buffer containing 50 mM NaCl, 30 mM $MgCl_2$, 1 mM dithiothreitol, 15% dimethyl sulfoxide, and 8% PEG-8000. EED and IN proteins (diluted in reaction buffer) were mixed and preincubated at 0°C for 30 min in molar ratios varying from 1:0.5 to 1:16 in terms of IN:EED stoichiometry. Control reactions were performed with mock samples consisting of the same chromatographic fraction as the one containing IN but from bacterial lysates harboring a void plasmid. Donor and target DNAs were incubated with protein mix for an additional 30 min at 0°C. Reaction buffer was then added to the mixture to a final volume of 20 μl, and the reaction was allowed to proceed for 90 min at 37°C. Reaction products were phenol extracted, ethanol precipitated, and electrophoresed in 1.2% agarose gels. Agarose gels were dried, autoradiographed, and quantitatively analyzed by using a PhosphorImager SI-475 (Molecular Dynamics).

EM and IEM. Cell specimens were included in metacrylate resin, sectioned and processed for conventional electron microscopy (EM) or IEM according to previously described methods (8, 29, 54). Cellular localization and possible co-localization of EED, IN, and MA proteins were analyzed in HIV-infected cells at different times p.i. by using single, double, or triple labeling with the relevant primary antibody, followed by the corresponding (anti-mouse or anti-rabbit IgG) colloidal gold-conjugated complementary antibodies. For IEM, rabbit IgG from laboratory-made antisera was purified by affinity chromatography by using a protein A-Sepharose column. MA protein was detected by using mouse monoclonal or rabbit polyclonal anti-MA antibodies, EED with rabbit antibody, and IN with anti-Flag MAb or rabbit anti-IN. In double-labeling experiments, when a 5-nm-gold-conjugated anti-mouse IgG antibody was used, a 10-, a 15-, or a 20-nm-gold-conjugated anti-rabbit IgG antibody was simultaneously used on the same cell section. In the triple-labeling experiments, since two primary antibodies belonged to the same species (e.g., the two rabbit anti-MA and anti-EED antibodies), both sides of the specimens on the grids were successively reacted. One side was simultaneously incubated with anti-MA rabbit antibody and anti-Flag(IN) MAb, followed by 10-nm-gold-conjugated anti-rabbit IgG antibody and 5-nm-gold-conjugated anti-mouse IgG antibody. The other side was then incubated with anti-EED rabbit antibody, followed by 20-nm-gold-conjugated anti-rabbit IgG antibody. To verify that there was no undesired cross-reaction of the 20-nm-gold-conjugated anti-rabbit IgG antibody with primary rabbit IgG bound to antigens on the first side of the section, stereoscopic views were taken upon tilting of the microscope goniometer. The 20-nm-gold grains could be seen, by using stereoscopic glasses, in a plane different from that of 5- and 10-nm gold grains (as exemplified in Fig. 9). Specimens were examined under a Hitachi HU7001 or a JEOL 1200-EX electron microscope, the latter equipped with a MegaView II High-Resolution TEM camera and a Soft Imaging System of analysis (Eloïse, Roissy, France).

RESULTS

The larger isoform of human EED protein is now considered to be comprised of 441 residues. An ortholog of human and murine EED has been identified in *Xenopus* and is called Xeed (73). As in human EED, the initiator methionine in Xeed corresponds to the Met95 codon in the murine ortholog (mEED) sequence and terminates at residue Arg441, corresponding to Arg535 in mEED (17, 54). From sequence analogy with insect ESC proteins and mammalian Gβ protein (51), the full-length sequence of EED larger isoform contains seven potential WD repeats (54). Compared to WT EED, mutant EED-Ct348 had lost its last two WD repeats. (see Fig. 1 and the legend to Table 2).

Interaction of human EED protein with HIV-1 IN in yeast two-hybrid assays. The HIV-1_{BRU} IN, fused to the DNA-binding LexA protein (BD hybrid), was used in yeast two-hybrid assays to test the possible interaction with EED protein, coexpressed as the Gal4 transcription activation domain-fusion protein AD-EED. WT EED and IN proteins interacted in yeast with a significant affinity (Fig. 2A), which was ~2-fold lower than that shown by the strong interactors Raf and Ras used in positive control samples (Fig. 2B). A weaker, but still positive β-galactosidase signal (one-third to one-fourth of the Ras-Raf control level; see Fig. 2B) was also observed with AD-EED-394AI and AD-EED-399A4 hybrids, two AD-fused EED mutants defective in MA protein interaction (54). This finding suggested that the IN-binding region in EED differed from the MA-interacting site, mapped to residues 294 to 309 (54) (see also the Table 2 footnotes) but that mutations in the MA-binding domain had some negative effect on the binding of EED to IN in yeast cells. The possibility that EED could have activated transcription of the reporter gene independently of a bona fide interaction with IN-BD hybrid could be excluded since no β-galactosidase signal over the background was detected with the pair of hybrids AD-EED and BD-Ras (Fig. 2) (54).

Mapping of EED-IN interacting sites by phage biopanning. When phages are panned on immobilized EED protein and specific ligand-elution is performed with an excess of affinity-purified IN, the phagotopes isolated would theoretically be mimotopes of peptide motifs of IN, since IN acted as a competitor of phages bound to IN-binding sites on EED (36, 37, 39, 54). Two phage libraries were used—one displaying random hexapeptides and the other displaying random dodecapeptides—on the basis that longer peptides could adopt a conformational structure that would better mimic that of protein interacting sites. About one-third of the phagotopes recovered from both libraries (15 of 50) could be grouped according to recurrent peptide motifs that showed some homology with the C-terminal sequence of IN overlapping tryptophan-235, within residues 212 to 264 (Table 1). For example, motif VLPPK, homologous to 259-VVPRRK in IN, was found several times. However, we observed a high level of degeneration and scatter of all of the phagotopes isolated, suggesting that the EED-binding region in IN had a high degree of three-dimensional organization and/or was constituted of discontinuous epitopes.

Reverse biopanning experiments were then performed to obtain mimotopes of the IN-binding site(s) in EED. The 6-mer library was panned on immobilized IN, and specific ligand elution was performed with an excess of affinity-purified, His-tagged EED-441-H6 protein as the specific competitor for IN-bound phages. The most represented group of phages (14 phages out of 24 independent clones isolated) showed homology with hydrophobic and aromatic motifs within residues 96

TABLE 1. Mimotopes of IN in phages eluted from EED[a]

Library	Phagotope sequence[b]	No. of independent phages isolated	Homologous motif in IN
16-mer	WSNIVV	1	243-WKGEGAVV-250
	EVGPGW	1	229-DSRDPLW-235
	AQNFGQ	1	220-IQNFR-224
	LQTFRQ	1	220-IQNFR-224
12-mer	VLPPKPMRQPVA	3	259-VVPRRK-264
	GIQVANPPRLYG	2	257-IKVVPR-262
	TTGLPLWSFNPS	1	233-PLW-235
	SPWRLPTPLT	1	232-DPLWKGP-238
	QLPFKLGPARID	1	232-DPLWKGPAKL-241
	SHPWNAQRELSV	1	230-SRDPLWK-236
	FSHELSWKPRKA	1	234-LWKGPAKL-241
	VPTNVQLQTPRS	1	212-ELQKQIT-218

[a] Peptide motifs in EED-bound phagotopes showing some homology with the IN sequence are underlined. Note that in a few phages, e.g., SPWRLPTPLT, the phagotope is sometimes shorter than the expected 12-mer.

[b] The amino acid sequence of the HIV-1-BRU IN protein is shown below. The N-terminal domain is made up of the first 50 residues; the C-terminal domain is made up of last 88 residues. The residues involved in the N-terminal zinc finger-like domain are in boldface, the catalytic triad D, D, and E in the core domain is in boldface italics, and the putative region of EED binding (residues 212 to 264) is underlined.

FLDGIDKAQD E**H**EKY**H**SNWR AMASDFNLPP VVAKEIVAS**C** DKCQLKGEAM 50
HGQVDCSPGI WQL***D***CTHLEG KVILVAVHVA SGYIEAEVIP AETGQETAYF 100
LLKLAGRWPV KTIHT***D***NGSN FTSTTVKAAC WWAGIKQEFG IPYNPQSQGV 150
V***E***SMNKELKK IIGQVRDQAE HLKTAVQMAV FIHNFKRKGG IGGYSAGERI 200
VDIIATDIQT KELQKOITKI ONFRVYYRDS RDPLWKGPAK LLWKGEGAVV 250
IODNSDIKVV PRRKAKIIRD YGKQMAGDDC VASRQDED-(288)

to 104, and the other group (4 phages) showed homology with hydrophobic and aromatic motifs within residues 224 to 232 (Table 2). These two putative IN-binding regions were located in the N-terminal moiety of EED and apparently did not overlap with the MA-binding site at positions 294 to 309, which confirmed the results of our two-hybrid assays in yeast (Fig. 2).

EED-IN interaction in vitro analyzed by mutagenesis and pull-down assays. Samples of bacterial cell lysates containing various forms of GST-IN fusion protein (Fig. 3a) were incubated with EED-441-H6 protein, and the resulting complexes were isolated on a glutathione-agarose affinity gel. As shown in Fig. 3b, the binding of full-length WT IN to EED in vitro confirmed the data from our yeast two-hybrid assays and implied that IN and EED proteins could directly interact with each other, without any requirement for a third partner provided by the yeast cell. The possibility that some nucleic acid or protein bridge could link IN and EED in this assay was ruled out on the basis of the following experimental arguments. (i) IN interacted with EED in vitro not only in crude bacterial lysates but also as two affinity-purified recombinant proteins. (ii) In all cases, the bacterial cell lysates from which each protein was recovered were first treated with a broad-spectrum endonuclease from *S. marcescens* (Benzonase; Sigma) that is specific for both single- and double-stranded RNA and DNA substrates. (iii) UV spectrum analyses performed on EED and IN recombinant proteins revealed no detectable nucleic acid contamination, as shown by A280/A260 ratios consistently found to be greater than 1.9.

In order to map the EED-binding region(s) in IN, eight individual deletions scanning the N- and C-terminal structural domains of IN and a double mutant mimicking the IN central core (56) were generated in GST-tagged IN protein (Fig. 1Ba), and the EED-binding capacity of these mutants was evaluated in pull-down assays. Deletions in the N-terminal domain of IN had no negative effect on its binding to EED, and even a slight enhancing effect was observed (Fig. 3b, compare clones 1-288 and N-81), implying that the presence of the N-terminal domain negatively influenced the apparent affinity of IN for EED in vitro. The C-terminal deletion mutant GST-IN-C-273 bound to EED at WT levels, whereas further deletion of residues 273 to 202, as in mutant C-201, reduced the binding to background levels. Likewise, the double-mutant D58-201, which represented the IN core domain (80), bound to EED to insignificant levels (Fig. 3b). The C-terminal region from residues 202 to 273 included the two overlapping sequences from positions 220 to 250 and positions 212 to 264 defined by our biopanning data (Table 1), and this finding confirmed that the C-terminal domain of IN was essential for the IN–EED interaction in vitro.

Since the most represented phagotope retained on immobilized IN corresponded to the sequence 99-VQFNWH-104 in EED (Table 2), a nonconserved mutation in three prominent amino acid residues of this motif was constructed: the bulky tripeptide motif WHS at position 103 to 105 was replaced by a stretch of three alanine residues to generate the EED-103A3 mutant (see Fig. 1Bb). Since this mutation could alter the immunoreactivity of the resulting EED mutant protein, the EED-103A3 mutant was assayed for IN binding as the fusion protein GST-EED-103A3, in comparison with its GST-EED WT counterpart. H6-IN-WT was used as the bait, Ni^{2+}-agarose gel as the trap, and EED proteins were detected by their GST tag by using anti-GST antibodies. EED-103A3 mutant

TABLE 2. Mimotopes of EED in phages eluted from IN[a]

Group	Phagotope sequence[b]	No. of independent phages isolated	Homologous motif in EED
I	VAEWHG	5	99-VQFNWH-104
	FFGLTK	3	96-LFGV-99
	FGQVWS	2	96-LFGVQFNWHS-105
	GANWPS	2	102-NWHS-105
	QGWFWL	1	100-QFNW-103
	MFEVEF	1	96-LFGVQF-101
II	ATVLYG	1	224-TLVAIFG-230
	VRYGP-	1	226-VAIFG-230
	AIAYIG	1	227-AIFGG-231
	ALFLVV	1	227-AIFGGV-232

[a] Peptide motifs in IN-bound phagotopes showing homology with the EED sequence are underlined.

[b] The amino acid sequence of the human EED WT protein is shown below. The seven putative WD repeats are underlined. The N-terminal domain within residues 31 to 63 shows some sequence homology with consensus PEST motifs.

MSEREVSTAP AGTDMPAAKK QKLSSDENSN PDLSGDENDD AVSIESGTNT
ERPDTPTNTP NAPGRKSWGK GKWKSKKCKY 80
SFKCVNSLKE DHNOPLFGVO FNWHSKEGDP LVFATVGSNR VTLYECHSQG
EIRLLQSYVD ADADENFYTC AWTYDSNTSH 160
PLLAVAGSRG IIRIINPITM QCIKHYVGHG NAINELKFHP RDPNLLLSVS
KDHALRLWNI QTDTLVAIFG GVEGHRDEVL 240
SADYDLLGEK IMSCGMDHSL KLWRIYSKRM MNAIKESYDY NPNKTNRPFI
SQKIHFPDFS TRDIHRNYVD CVRWLGDLIL 320
SKSCENAIVC WKPGKMQDDI DKIKPSESNV TILGRFDYSO CDIWYMRFSM
DFWOKMLALG NOVGKLYVW DLEVEDPHKAK 400
CTTLTHHKCG AAIROTSFSR DSSILIAVCD DASIWRWDRL R-(441).

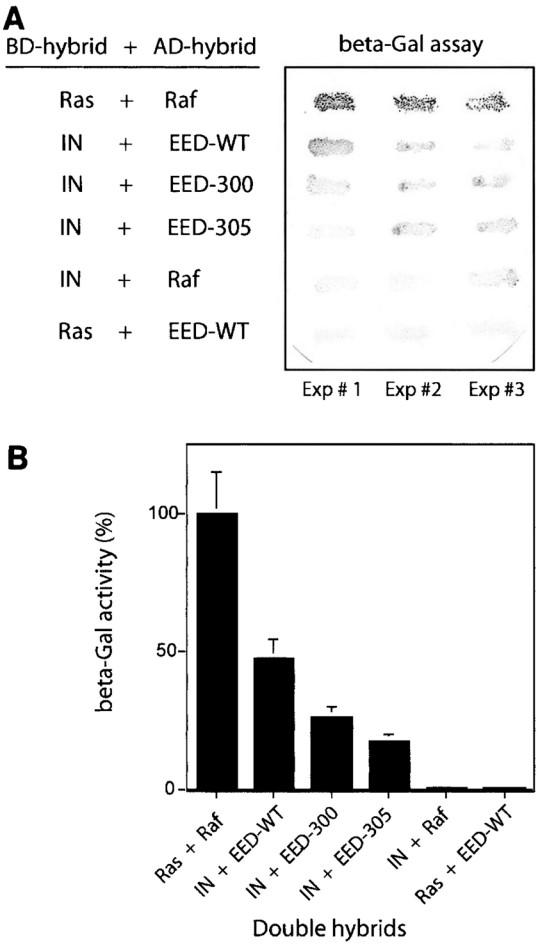


FIG. 2. Interaction of EED and HIV-1 IN in vivo. (A) The affinity of EED (WT or mutants EED-300 and EED-305) for IN was analyzed in yeast by using a two-hybrid assay. IN (or Ras) was cloned in fusion with LexA protein (BD hybrid), and EED (or Raf) was cloned in fusion with the Gal4 activation domain (AD hybrid). The β-galactosidase activity of yeast transformants plated on medium without histidine and replica plated on Whatman filters is shown. The results of three independent experiments are shown. Growth in the absence of His and blue color in the β-Gal assay are indicative of interaction between hybrid proteins in yeast cells. Positive controls involved BD-Ras and AD-Raf hybrids as strong interactors, and negative controls consisted of BD-IN and AD-Raf and of BD-Ras and AD-EED hybrids, respectively. (B) Quantification of β-galactosidase activity in yeast transformants harboring EED and IN hybrids is expressed as the percentage of the BD-Ras plus AD-Raf activity. beta-Gal, β-galactosidase.

was still capable of binding to IN, although with a lower efficiency (two- to threefold; Fig. 3d and e). This suggested that at least one of the IN-binding sites mapped to residues 103 to 105 in EED but that additional IN contact sites could also be present in other domains of EED, as suggested by our phage biopanning (Table 2).

One of the striking features of *Pc*-G gene *eed* products is their high degree of conservation at the protein level in mammals, implying that all amino acid residues of EED are important for their biological functions (51, 68). Thus, it has been shown that the integrity of the last WD repeat of *Pc*-G proteins is highly critical and that even short deletions at the C terminus are detrimental to their activity (51, 76). Consistent with this, the C-terminal deletion mutant EED-C348 was found to bind to the IN-affinity gel at very low levels (Fig. 3d), indicating that the C-terminal fourth of EED was crucial for its interaction with IN, directly or indirectly, via EED conformation and structure stabilization.

Influence of EED on HIV-1 IN activity in vitro. The enzymatic activity of IN was assayed in reactions of DNA integration in vitro in the presence or absence of affinity-purified recombinant EED protein. In IN assays, homologous integration (or autointegration) refers to the integration of one 5′+3′-LTR-containing DNA donor fragment into another (Fig. 4a, band i), whereas heterologous integration (or heterointegration) corresponds to the insertion of one (or more) 5′+3′-LTR-containing fragment(s) into one plasmid target (9). In the absence of EED, several discrete bands of heterointegration products were usually observed: a major band (Fig. 4a, band iii) corresponded to single-tagged circles, resulting from one single integration event per plasmid target, mediated by one LTR (one-ended integration) (10); faster-migrating minor band(s) represented concerted integration events involving two LTRs (Fig. 4a, band ii) and resulting from the insertion of two donors into one target DNA molecule with subsequent linearization (two one-ended concerted integration) and, more rarely, from the two-ended concerted integration of a single donor fragment into target DNA (10). When IN reactions were performed under conditions in which the major band of single-tagged circles (band iii) was the only detectable signal (Fig. 4a, control lane 1), the addition of WT EED provoked an increase in the intensity of autointegration products (band i) and heterointegration products corresponding to single-tagged circles (band iii) (Fig. 4a). The effect became detectable for EED/IN ratios over 2:1 and seemed to occur in a dose-dependent manner, with a twofold augmentation obtained at a ratio of EED to IN of 8:1 (Fig. 4c). However, a 20- to 30-fold augmentation of the putative two one-ended concerted integration products (band ii) was observed within EED/IN ratios ranging from 4:1 to 8:1 (Fig. 4a and c). The same patterns were observed with affinity-purified, WT EED-441-H6 and GST-EED-441, the latter being cleaved off the GST domain. No detectable enhancing effect was observed with corresponding chromatographic fractions from mock-expressing bacterial cells (not shown) or with the deletion mutant protein EED-C348, which provoked a slight and probably nonspecific negative effect at high concentrations (Fig. 4b). This pattern confirmed the interaction between EED and IN proteins in vitro, with a secondary influence on the enzymatic activity of IN.

In situ analysis of IN, EED, and MA proteins in HIV-1-infected cells. Human epithelial HeLa $CD4^+$ cells (P4P56 cells) (52) and lymphoid MT4 cells were infected with BRU-FlagWT, a subclone of $HIV\text{-}1_{BRU}$ expressing a Flag-tagged IN (53). Negative control samples consisted of mock-infected cells or cells incubated with BRU-FlagΔEnv, a noninfectious *env*-deleted mutant used as control for nonspecific HIV-cell interaction. Cells were harvested at 1.5, 6, or 24 h p.i. and processed

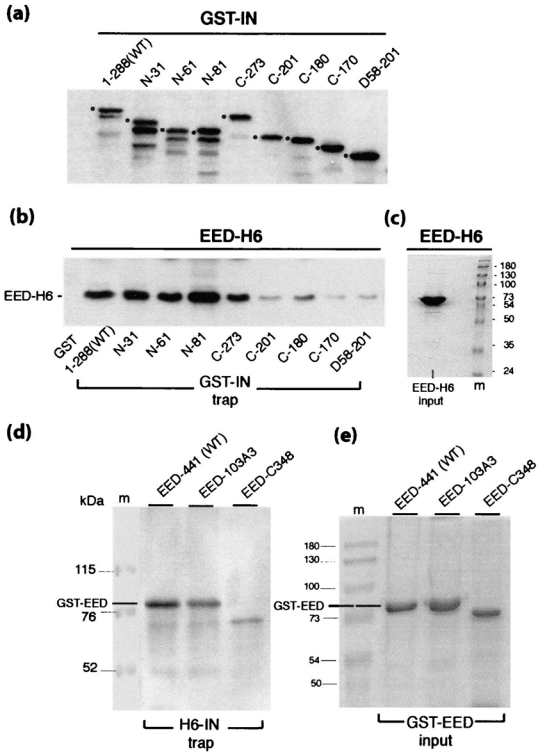


FIG. 3. Interaction of EED and HIV-1 IN in vitro. The affinity between EED and IN and their respective mutants was analyzed by GST and histidine pull-down assays with glutathione-agarose (a to c) and Ni^{2+}-NTA-agarose (d and e) as affinity gels. In panels a to c, aliquots of GST-fused full-length IN (i.e., positions 1 to 288), N-terminal (N), C-terminal (C), and double-deletion (D) mutants, as displayed in panel a, were immobilized on a glutathione-agarose gel, and the affinity gels were incubated with aliquots of bacterial cell lysates containing 10 μg each of full-length, His-tagged EED (EED-H6). The amount of EED-H6 protein retained on GST-IN-glutathione-agarose gel was then evaluated by SDS-PAGE and immunodetection on a blot with anti-His tag MAb as shown in panel b. In panel c is shown EED-H6 input (10-μg protein load; Coomassie blue staining). Note that in panel a, the pattern of anti-GST reacting bands is highly suggestive of a major proteolytic cleavage site located in the C-terminal domain of IN, within region 288-273: a prominent lower band migrating with a constant apparent molecular mass ~2 kDa lower than that of the original GST-IN gene products (marked by solid dots) is visible in samples of full-length and N-deleted forms of GST-IN, whereas this band is absent from the C-truncated or double-truncated GST-IN clones. In panels d and e, the GST-fused, full-length EED, point mutant EED-103A3 and C-terminal deletion mutant EED-C348 were incubated with aliquots of His-tagged IN (H6-IN) adsorbed onto an Ni^{2+}-agarose gel, and the amount of GST-EED protein retained on the H6-IN-Ni^{2+}-agarose gel was determined by SDS-PAGE and immunoblotting with anti-GST polyclonal antibody (panel d). (e) GST-EED protein input (2-μg protein load per well; Coomassie blue staining).

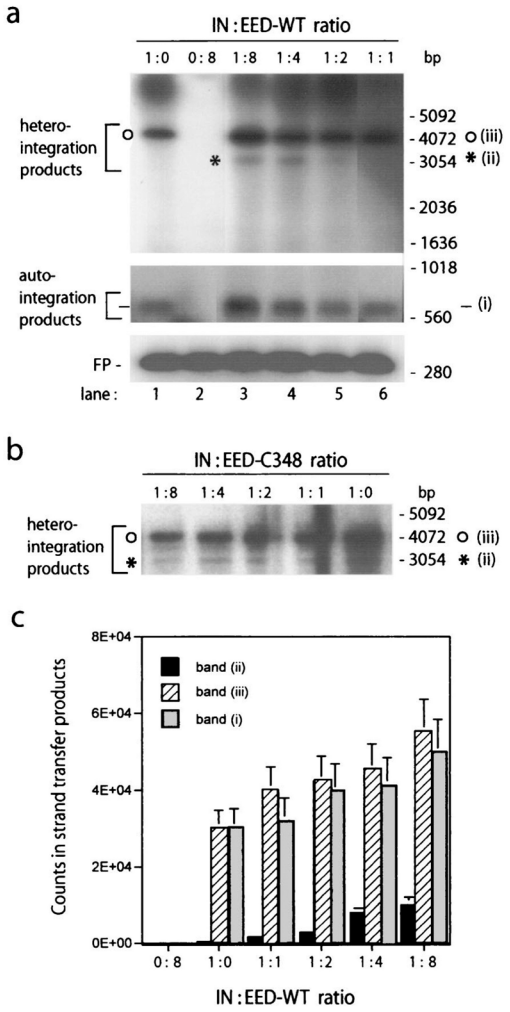


FIG. 4. Effect of EED WT (a and c) and deletion mutant EED-C348 (b) on IN-mediated integration reaction in vitro. The respective IN/EED stoichiometric ratios are indicated at the tops of panels a and b. Control reactions were performed in the absence of EED (lane 1) or in the absence of IN (lane 2). The positions of autointegration products (band i) and heterologous integration products (bands ii and iii) are indicated by brackets on the left. As inferred from previous studies (10, 31), the minor band (band ii [✱]) is likely to represent the product of two one-ended concerted integration followed by linearization, whereas the major band (band iii [○]) represents single-tagged target circles. (b) Integration reaction performed in the presence of EED-C348 mutant. Only heterointegration products are shown. (c) Quantification of counts recovered from auto- and heterointegration products of IN reactions performed in the presence of EED-WT (mean of three experiments ± the standard deviation).

for EM and IEM. For IEM analysis, cell specimens were reacted with monoclonal anti-Flag(IN), polyclonal anti-EED, and monoclonal anti-MA MAbs and, occasionally, with anti-CA monoclonal or anti-MA polyclonal antibodies, as indicated. Antibodies were used in single-, double-, or triple-immunolabeling reactions with their corresponding complementary gold-tagged anti-IgG antibody.

At 1.5 h p.i. in mock-infected or BRU-FlagΔEnv-reacted cells, the cytoplasm and nucleus were poorly labeled with anti-MA, anti-CA, and anti-Flag(IN) antibodies, and the immunolabeling corresponded to background reactivity (not shown). In

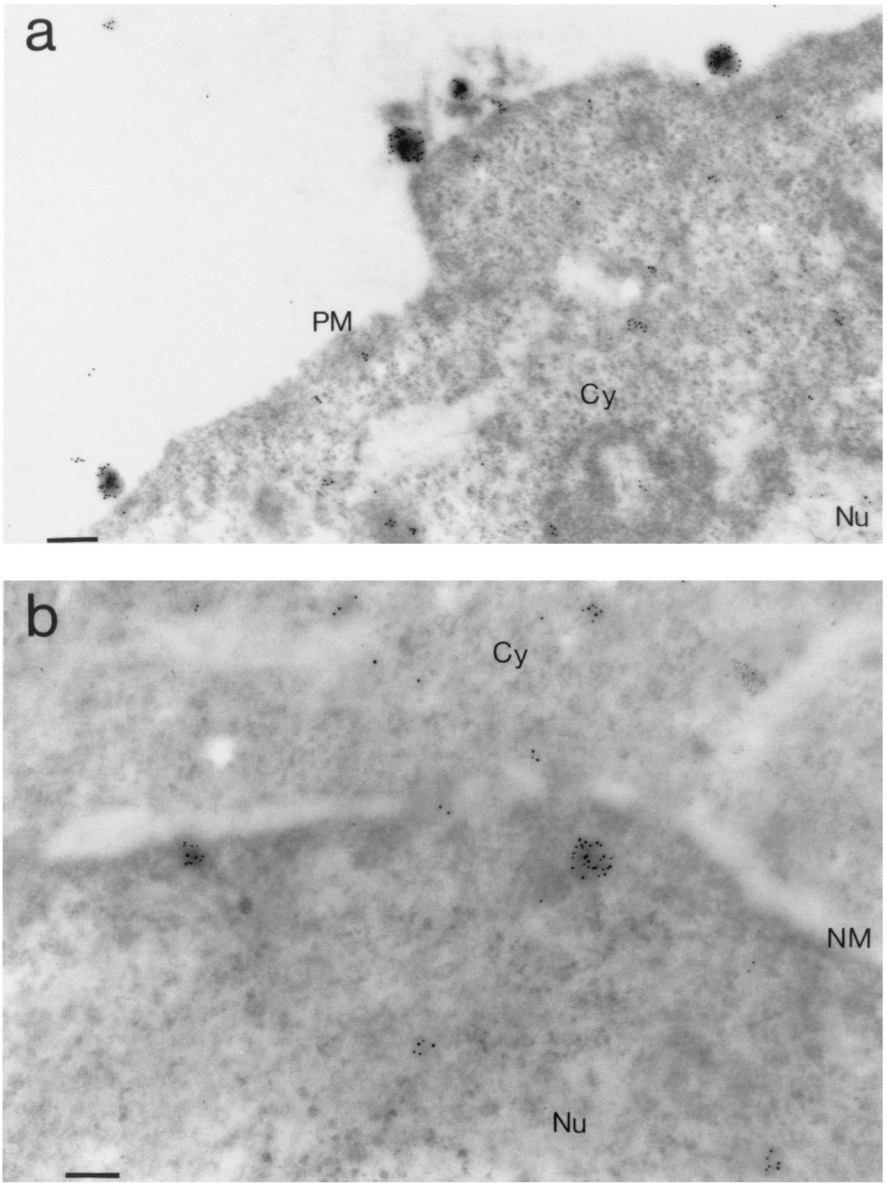


FIG. 5. IEM analysis of human MT4 (a) and HeLa-CD4$^+$ (P4P56) (b) cells infected with Flag-tagged IN-containing HIV-1 (BRU-FlagWT; multiplicity of infection of 1,000 virions/cell) obtained at 1.5 h p.i. In panel a, MA and CA proteins were detected by a mix of anti-MA and anti-CA MAbs, followed by 10-nm-colloidal-gold-tagged anti-mouse IgG antibody. The immunogold-labeled globular structures bound to the cell surface had diameters compatible to those of HIV-1 virions (110 to 130 nm). In P4P56 cells (panel b), the Flag-tagged IN protein was detected by using anti-FLAG MAb (M2), followed by 10-nm-colloidal-gold-tagged anti-mouse IgG antibody. Note that the two gold-labeled globular structures that were 60 to 80 nm in diameter visible in the nucleus in the vicinity of the nuclear membrane had dimensions compatible with those of the PIC. PM, plasma membrane; C, cytoplasm; N, nucleus; NM, nuclear membrane; PM, plasma membrane. Bars: 166 nm (a), 100 nm (b).

BRU-FlagWT-infected cells, however, virus particles were visible in numbers at the cell surface as roughly spherical structures of 120 to 140 nm in diameter that were double labeled with anti-MA and anti-CA antibodies (Fig. 5a). These plasma membrane-associated particles were no longer seen at later times, i.e., at 6 and 24 h p.i.. Within the cells, anti-Flag(IN) antibody-labeled globular structures that were 60 to 80 nm in diameter were visible in significant numbers in the nucleoplasm and frequently seen in the vicinity of nuclear pores (Fig. 5b). Their immunoreactivities and sizes were compatible with those of PICs, which have been reported to be ca. 56 nm in average diameter (49). At 1.5 h p.i., EED labeling was found to

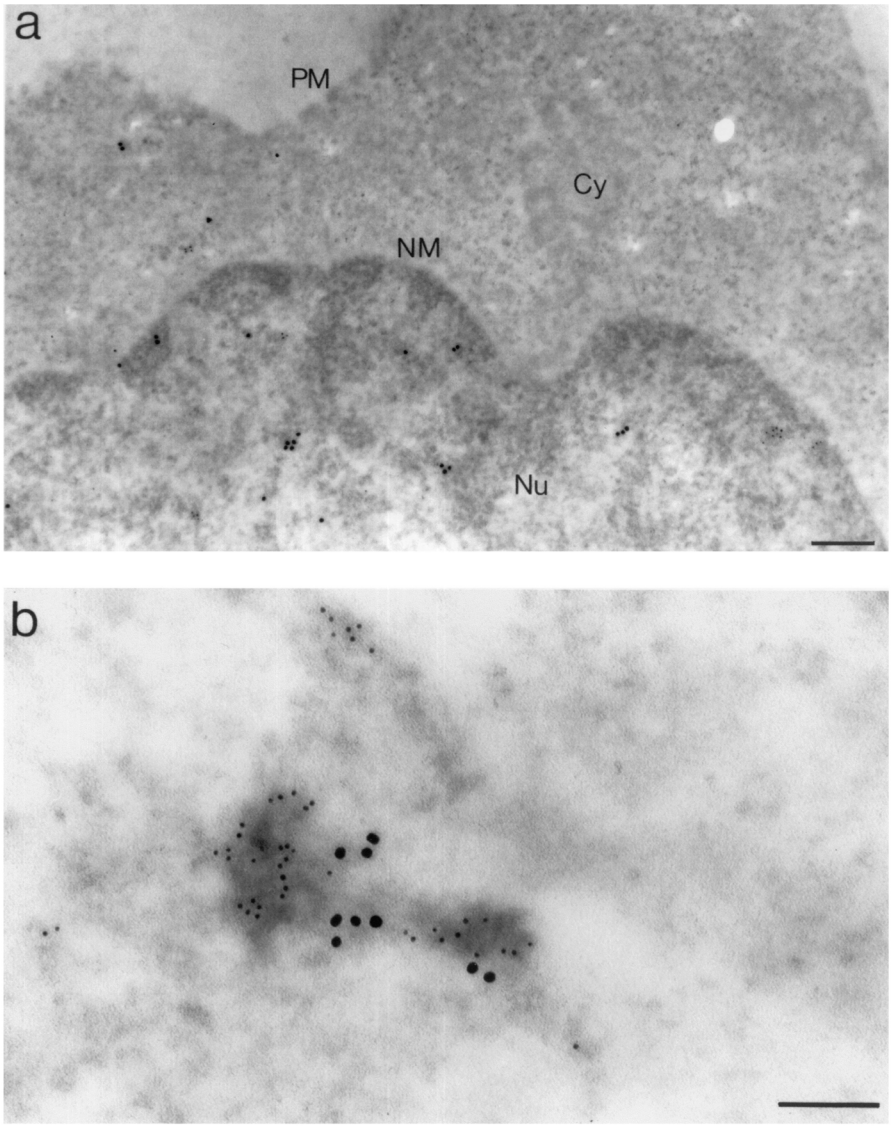


FIG. 6. In situ analysis by IEM of EED and MA proteins in HIV-1-infected MT4 cells taken at 6 h p.i. (a) Mock-infected cells; (b) BRU-FlagWT-infected cells. Cell sections were simultaneously labeled with rabbit anti-EED, detected by using a 10-nm-colloidal-gold-conjugated anti-rabbit IgG antibody, and with anti-MA MAb, detected by a 5-nm-colloidal-gold-conjugated anti-mouse IgG antibody. In panel a, a general view of the cell is shown. In panel b, an area of the nucleoplasm is presented, showing colocalization of 5-nm (MA) and 10-nm (EED) gold grains, associated with electron-dense material. Cy, cytoplasm; Nu, nucleus; PM, plasma membrane; NM, nuclear membrane. Bars: 200 nm (a), 100 nm (b).

be evenly distributed between the cytoplasm and the nucleus and randomly dispersed in both compartments, with no apparent difference between control cells, mock-infected cells, and BRU-FlagΔEnv-treated cells (IEM images not shown; refer to data in Fig. 10a and b).

Double labelings of MT4 and P4P56 cells were then performed to detect a possible colocalization of IN, EED, and MA, pairwise, in the cytoplasm or nucleus at later times p.i. In control cells (mock-infected or BRU-FlagΔEnv-treated cells), double immunolabeling with anti-EED (revealed by 10-nm-gold-tagged conjugate) and anti-IN (5-nm gold grains) or with anti-EED (10-nm gold grains) and anti-MA (5-nm gold grains), respectively, showed that grains of both diameters were randomly distributed in the cytoplasm and nucleoplasm, with no indication of a preferred localization or colocalization at 6 and 24 h p.i. (Fig. 6a and 7a).

The results were different in BRU-FlagWT-infected cells (Fig. 6b). Many patches of electron-dense material were double labeled with anti-MA and anti-EED antibodies in the nucleus of infected cells taken at 6 h p.i. This finding confirmed

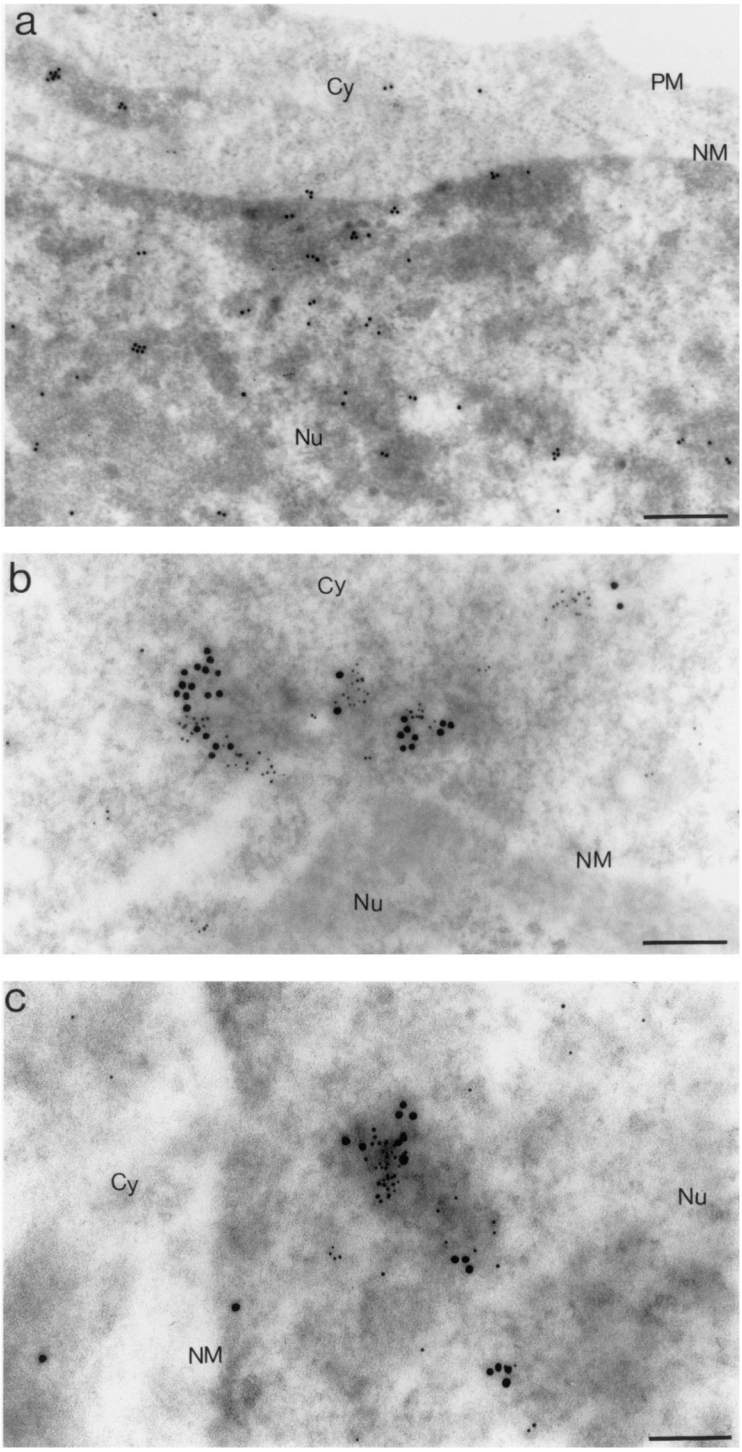


FIG. 7. In situ analysis by IEM of EED and IN proteins in HIV-1-infected human cells at 6 h p.i. Mock-infected MT4 cells (a) and BRU-FlagWT-infected MT4 cells (b) and HeLa-CD4$^+$ cells (P4P56) (c) are shown. Cell sections were simultaneously labeled with anti-EED rabbit antibody, detected by using a 10-nm-colloidal-gold-conjugated anti-rabbit IgG antibody, and with anti-Flag(IN) MAb (M2), detected by using a 5-nm-colloidal-gold-conjugated anti-mouse IgG antibody. Panel a presents a general view of the cell section. Panels b and c show enlargements of sections of nuclear membrane with flanking areas of nucleoplasm and cytoplasm. Cy, cytoplasm; Nu, nucleus; PM, plasma membrane; NM, nuclear membrane. Bars: 333 nm (a), 142 nm (b), 125 nm (c).

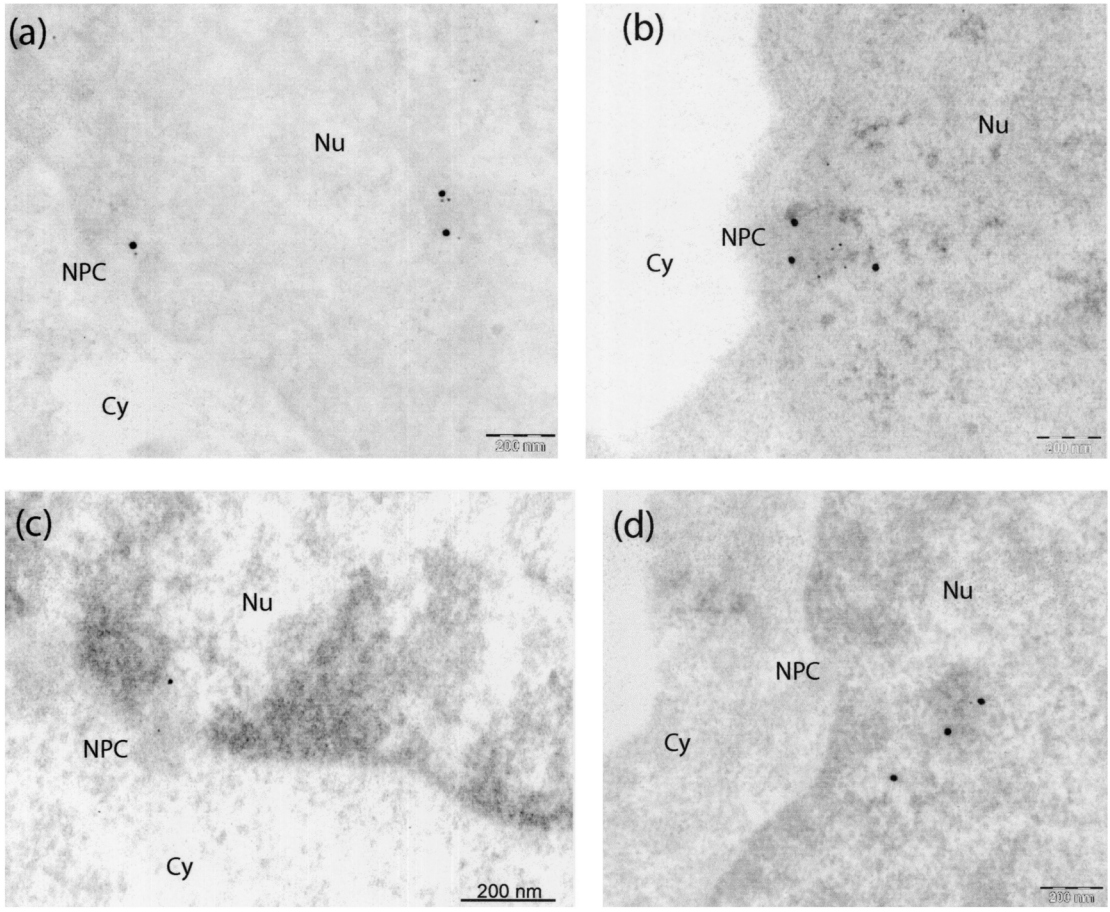
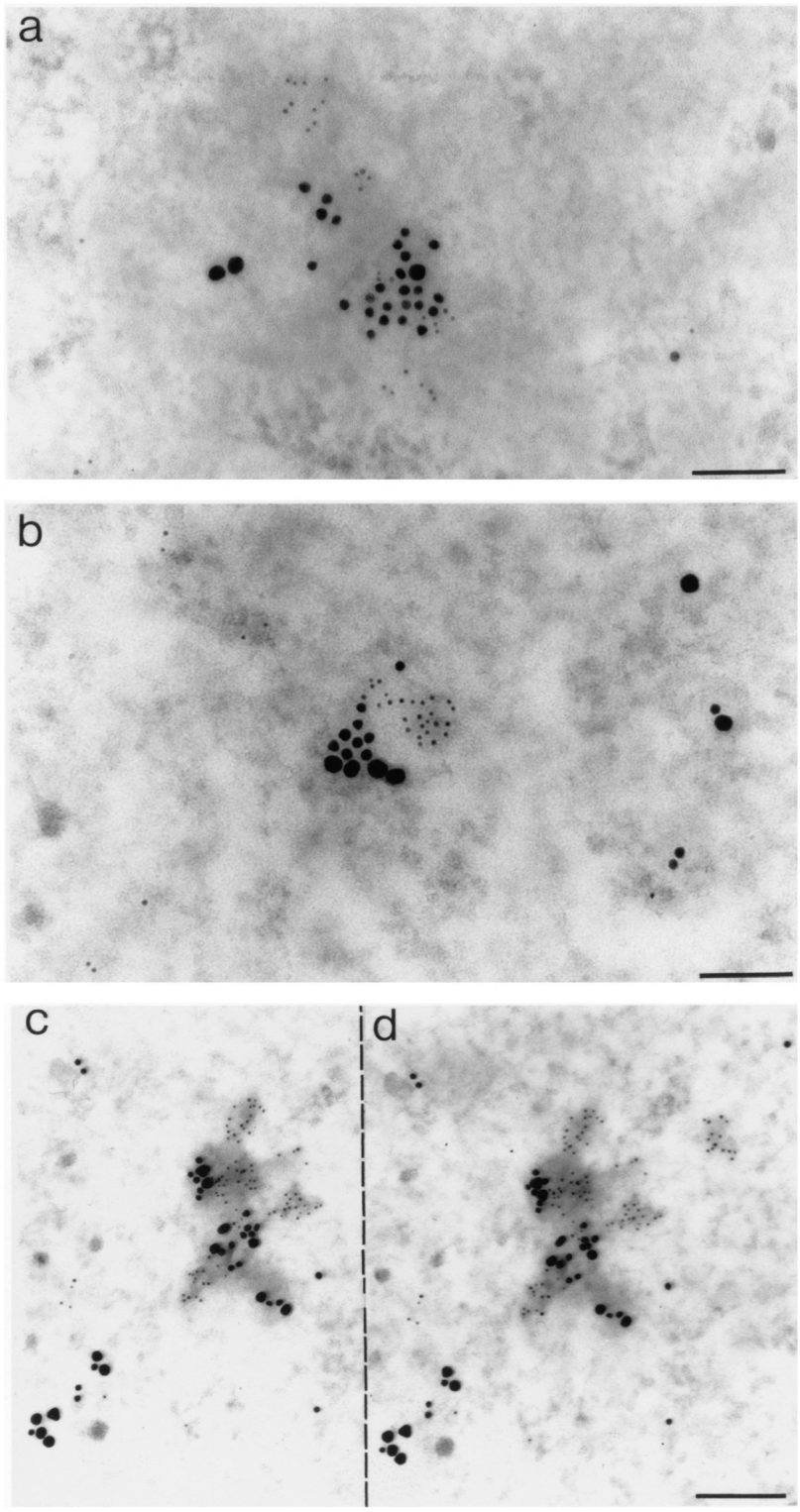


FIG. 8. IEM analysis of colocalization of EED and IN proteins at or near the nuclear pores of BRU-FlagWT-infected MT4 cells obtained at 6 h p.i. The different panels present various areas of cell sections showing nucleoplasm and nuclear pores. Specimens were labeled with rabbit anti-EED, detected by using a 20-nm-colloidal-gold-conjugated anti-rabbit IgG antibody, and anti-Flag(IN) MAb, detected by using a 5-nm-colloidal-gold-conjugated anti-mouse IgG antibody. Cy, cytoplasm; Nu, nucleus; NPC, nuclear pore complex. Bars: 200 nm.

the colocalization of EED and MA proteins already observed in recombinant protein coexpressing insect and human cells (54). Likewise, a pattern of colocalization was observed for IN and EED in BRU-FlagWT-infected MT4 (Fig. 7b) and P4P56 (Fig. 7c) cells taken at 6 h p.i. and double-immunolabeled with anti-IN and anti-EED antibodies. Colocalizations of IN and EED were mainly observed in the nucleus (Fig. 7c) but were also observed in the cytoplasm in close vicinity to the nuclear membrane (Fig. 7b). IN-EED colocalizations were frequently seen at or near the nuclear pores (Fig. 8). In most of these observations, the clusters of EED- and IN-bound grains were found to be associated with electron-dense material in the nucleoplasm or at the nuclear pores (Fig. 6 to 8). IN-EED colocalizations were no longer observed at 24 h p.i.

In order to detect a possible colocalization of the three proteins IN, EED, and MA, triple-immunolabeling experiments were also performed with anti-EED (revealed by 20-nm colloidal gold grain), anti-MA (10-nm colloidal gold grain), and anti-IN (5-nm colloidal gold grain) antibodies on sections of BRU-FlagWT-infected MT4 and P4P56 cells. Many clusters of gold grains of the three different diameters were found in nuclei at 6 h p.i. (Fig. 9). Such triple colocalizations were never observed in control samples, mock-infected cells, and BRU-FlagΔEnv-treated cells. Our IEM data therefore suggested that IN, MA, and EED proteins could form ternary complexes and/or participate together in higher-order complexes in vivo within the nucleus of HIV-1-infected cells at early times p.i.

Quantitative IEM analysis of the cellular distribution and colocalization of IN, EED, and MA proteins in HIV-1-infected cells. To further analyze the distribution of EED, IN, and MA proteins in the cellular compartments in a semiquantitative and kinetic manner, sections of control or HIV-1-infected P4P56 or MT4 cells were separately reacted with rabbit anti-EED, anti-IN and anti-MA IgG, followed by 10-nm-gold conjugated anti-rabbit IgG antibody in single immunogold labeling experiments. Gold grains were counted by using EM in several sections of independent cells taken at time intervals during HIV-1 infection. At least 400 grains were counted on 20 dif-


a
b
c
d

ferent cell sections, and the variations of the density of grains (number of grains per unit surface area of cell section) in the cytoplasm (Fig. 10a) and nucleus (Fig. 10b) were compared during the course of infection. Backgound labeling was given by mock-infected cells reacted with the same primary and gold-conjugated secondary antibodies (zero time point).

In HIV-1-infected cells, MA labeling was detected as early as at 1.5 h p.i. in both cytoplasm (Fig. 10a) and nucleus (Fig. 10b). The MA signal decreased to background levels at 6 h p.i. in the cytoplasm but was still observed at significant levels in the nucleus until 6 h p.i. In the cytoplasm, the IN labeling was not significantly higher than the background level at any time p.i. (Fig. 10a). In the nucleus however, IN signal was detected at 1.5 h p.i. and reached a maximum level at 6 h p.i., suggesting a higher number, or better accessibility, of IN molecules at this time of the virus cycle (Fig. 10b; see also Fig. 5b). The level of EED labeling remained almost constant throughout the time period in both cytoplasmic and nuclear compartments (Fig. 10a and b). No significant variation and increase over the background labeling was observed in BRU-FlagΔEnv-infected P4P56 or MT4 cells for the three markers, MA, IN, and EED, during the same time period (background line; Fig. 10c).

Colocalization of EED, IN, and MA proteins was also quantitavely evaluated by EM by counting neighboring grains on sections from cells taken at different times p.i. and subjected to triple immunolabeling with specific monoclonal and polyclonal antibodies (see, for example, Fig. 9). This quantitative assay was based on the following principle. The size of an antibody molecule under EM is ca. 15 nm and, in our IEM analyses, primary and secondary gold-labeled antibody molecules were used. Thus, if one antibody is labeled with a 10-nm gold grain and the other one is labeled with a 20-nm grain and if both gold conjugations occurred at their respective Fc domains, the two gold grains that are 90 nm apart (15 + 15 + 10 + 15 + 15 + 20 nm) might theoretically mark two adjacent epitopes carried by the same protein molecule or, alternatively, two epitopes belonging to two neighboring molecules. Since epitopes on interacting proteins could lie at a certain distance from each other, one could reasonably consider that immunogold grains that are seen within a distance range of 100 to 120 nm would mark molecules that colocalize in the same cell compartment.

Using this quantitative method, we observed no significant colocalization of grains marking EED, MA, and IN in the cytoplasm and nucleus of mock-infected cells or in cells incubated with BRU-FlagΔEnv (Fig. 10c). Likewise, colocalization of gold grains marking IN, MA, and EED was rarely observed in the cytoplasm of BRU-FlagWT-infected cells at early times p.i. (e.g., 1.5 h p.i.), and no statistical analysis could be performed (results not shown). The results were different for the nuclei of BRU-FlagWT-infected cells, which showed clusters of double and even triple colocalizations with a significant frequency (Fig. 10c). An average value of three colocalization events per square micrometer of nucleoplasm area was found for the pair IN-MA at 1.5 h p.i., but no other pairwise colocalization (IN-EED or MA-EED) was found in significant numbers. At later times, no more colocalization of IN and MA was observed. At 6 h p.i., a mean value of two colocalization events per square micrometer was found for IN and EED, and triple colocalization of EED, IN and MA occurred at an average frequency of one event per square micrometer of nucleus section area (Fig. 10c). No detectable double or triple colocalizations were detectable at 24 h p.i. This finding suggested that the time point of 6 h p.i. represented the phase of the virus cycle when significant numbers of EED, IN, and MA molecules were in close vicinity within the nucleus of HIV-1-infected cells. Alternatively, it could mean that the intranuclear microenvironment that allowed for a maximal immunoreactivity and accessibility of the three proteins took place at 6 h p.i.

DISCUSSION

We found here that human EED, a *Pc*-G protein, can interact with HIV-1 IN both in vitro and in vivo in yeast. Using deletion mutagenesis and phage biopanning, we mapped the major EED-binding sites to the C-terminal domain of IN. In vitro, we observed an apparent positive effect of EED on IN-mediated integration reaction. We hypothesize that this effect was indirect: the interaction of EED with IN could promote the oligomerization of IN molecules, which would in turn favor the integration process. In situ analysis of EED and IN cellular localization was performed on HIV-1-infected human epithelial (HeLa $CD4^+$) or lymphoid (MT4) cells by using IEM and differential immunogold labeling. We found that EED and IN colocalized within the nucleus of HIV-infected cells, a phenomenon that was mainly observed at early times p.i. (1.5 to 6 h). Triple-immunolabeling experiments showed that the MA protein, another viral protein partner of EED (54), was also detected in significantly frequent associations with both EED and IN proteins in the nucleus at early times p.i., suggesting the occurrence of ternary complexes involving EED, MA, and IN. Although the role of EED protein in the HIV-1 life cycle is not known, our data suggest that EED could be involved in some cellular function(s) necessary for and/or induced by early steps of the virus-host cell interaction. The fact that we never observed any cellular colocalization of EED, IN, and MA in cells infected with HIV-1 in the presence of virus inhibitor AZT

FIG. 9. In situ analysis by IEM of IN, EED, and MA proteins by using triple immunogold labeling of EM sections of BRU-FlagWT-infected human cells taken at 6 h p.i. (a) MT4 cell nucleoplasm; (b to d) HeLa $CD4^+$ cell (P4P56) nucleoplasm. In a first step, one side of the grid carrying the cell pellet section was simultaneously reacted with mouse anti-Flag(IN) MAb and rabbit anti-MA antibody and then with a mix of 5-nm-gold-conjugated anti-mouse IgG antibody and 10-nm-gold-conjugated anti-rabbit IgG antibody. In a second step, the other side of the grid was incubated with rabbit anti-EED antibody, detected by using a 20-nm-gold-conjugated anti-rabbit IgG antibody. (c and d) Stereoscopic view of a portion of P4P56 cell nucleus, taken upon tilting the microscope goniometer, showing clusters of grains 5 nm (IN), 10 nm (MA), and 20 nm (EED) in diameter, in association with electron-dense nucleoplasmic material. By using stereoscopic glasses and aiming at the dotted line separating panels c and d, one can see that the 20-nm gold grains belong to a plane different from that of the 5- and 10-nm grains and that there was no cross-labeling between anti-MA (top side of the specimen) and anti-EED primary IgG (bottom side of the specimen) by the secondary anti-IgG antibody. Bars: 100 nm (a and b), 166 nm (c and d).

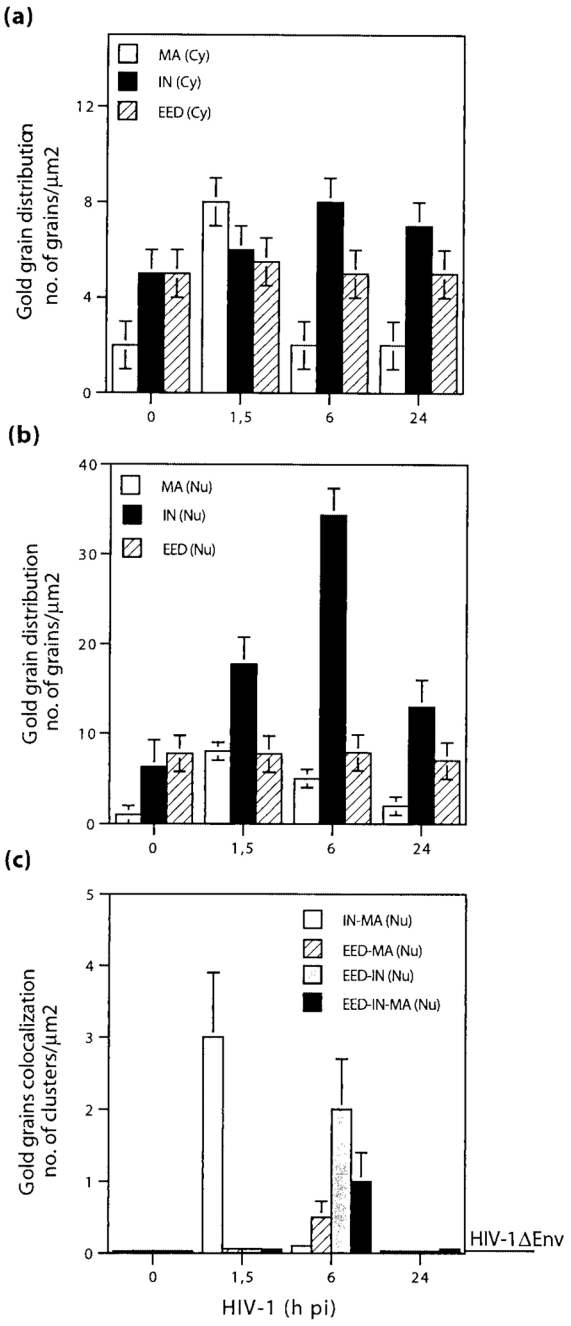


FIG. 10. Quantitative image analysis of intracellular distribution (a and b) and nuclear colocalization (c) of EED, MA, and IN proteins in HIV-1-infected MT4 cells. Cells infected with BRU-FlagWT, or its noninfectious, envelope-deleted version BRU-FLAGΔEnv, were harvested at different times p.i. as indicated on the *x* axis. (a) Cytoplasm; (b and c) nucleus. In panels a and b, gold grains were counted on cell sections after single labeling with affinity-purified anti-EED, anti-MA, or anti-IN rabbit antibodies, respectively, followed by 10-nm-gold-tagged anti-rabbit IgG antibody. Background labeling for IN and MA proteins was given by the number of grains per square micrometer on mock-infected cell sections (zero point). In panel c are shown the results of triple-labeling experiments (see, for example, Fig. 9). The baseline on the right corresponds to the data obtained with BRU-FLAGΔEnv (HIV-1ΔEnv).

(data not shown) suggests that the role played by EED in the virus cycle would require the step of proviral DNA synthesis. Since EED is a partner of two viral proteins—MA, which has both structural and functional roles at the early stages of the cycle (23), and IN, which is responsible for proviral integration—one might envisage EED as a possible participant in at least two major retroviral processes: the intracellular transport of incoming virions and/or the proviral integration.

In spite of a number of studies (2, 13, 16, 24–27, 35, 40, 48, 49, 52, 84, 85), the cellular and viral factors that control the reaction and the site(s) of integration of HIV-1 proviral DNA still remain incompletely elucidated. The integration process is better understood at the molecular level for retrotransposons, such as Ty5, which integrates into regions of silent chromatin via yeast proteins Sir (86). Like the Sir proteins in yeast, the products of many *Pc*-G genes are responsible for the maintenance of the silent state of chromatin in upper eukaryotes, likely by recruitment of histone deacetylases (reviewed in reference 55). Thus, the transcriptional repression by EED has been recently reported to involve histone deacetylation (79). Likewise, in murine neurons, EED has been shown to colocalize with histone H1 in transcriptionally inactive domains of perinucleolar heterochromatin (1). The function of *Pc*-G proteins as transcriptional repressors and gene silencers (17, 33, 51, 63, 67, 74) has been attributed in several cases to direct interaction with transcription factors rather than to direct DNA binding (3, 64). For example, EED has been found to bind to YY1, a vertebrate DNA-binding protein (4), and a stable, if only transient, interaction between EED, EZH2 (enhancer of Zeste 2) (70), YY1, and histone deacetylase has been suggested (62). An indirect physical link has therefore been established between EED and the DNA of target chromatin regions via the DNA-binding protein YY1 (62).

In light of the latest results on HIV-1 integration into transcriptionally active regions of the host genome (66, 83), an attractive hypothesis would be that intranuclear interaction occurring between the viral proteins MA and IN on one side and the cellular protein EED on the other side might deregulate silent cellular genes by releasing the binding of EED to YY1 (or other factors involved in the transcription machinery), by modifying the acetylation state of histones, or by both of these processes. This would then activate the integration process of the proviral DNA. However, it is important to note that YY1 is a multifunctional transcription factor which, under certain circumstances, can act as a transcriptional repressor (72).

Alternatively, but not exclusively, there may be a role in the intracellular transport and nuclear translocation of the PIC for EED. As a component of a multifactor transport complex, EED might act as a shuttle protein to convoy the viral PIC to the nucleus via its binding to MA and IN. Only a few cellular proteins have been identified thus far as involved in the transport of HIV-1 PIC. A DNA targeting function has been assigned to the product of the *INI1* gene (40), a cellular inter-

actor of HIV-1 IN that seems to also act in late events in the viral life cycle (84). It has been shown that HIV-1 PIC recruits INI1 and PML proteins within the cytoplasm, and this complex could facilitate the integration via the recruitment of additional cell factors, such as the PML-binding CBP/p300 (78). Interestingly, a protein similar to EED, called WAIT-1 (WD protein associated with integrin cytoplasmic tails-1) (61), has been found to interact with the cytoplasmic domain of the integrin β7 subunits, and a shuttling of WAIT-1/EED between membrane-associated β7 integrins and the nucleus has been suggested (61). The β7 cytoplasmic domain is involved in major integrin functions such as receptor affinity and signaling (15, 46). It is noteworthy that, among the β7 subfamily members, αEβ7 integrin is restricted to a subset of gut-associated T lymphocytes and dendritic cells (61). The frequent occurrence of EED and IN double labeling at nuclear pore complexes, as observed in our IEM study, would support a shuttling function for EED. Further investigations by cell fractionation of HIV-1-infected cells, overexpression versus synthesis inhibition of EED in HIV-infected cells, and isolation of dominant-negative mutants of EED would help elucidate the role(s) of cellular EED protein in HIV-1 infection.

ACKNOWLEDGMENTS

This work was supported by the Agence Nationale de Recherche sur le SIDA (ANRS; AC14-2 "HIV-1 Integrase and Preintegration Complex"). S.V. received an ANRS fellowship, and S.P. received a fellowship from the French ECS Association (Ensemble contre le SIDA).

We are deeply grateful to Simone Peyrol and Isabelle Leparc-Goffart (Centre d'Imagerie de la Faculté de Médecine Laennec) and to Paul Paulet (Centre Régional d'Imagerie Cellulaire de Montpellier) for significant contributions to the EM specimen processing, image digitalization, and photography. We are also grateful to Jean-Claude Cortay for valuable advice on pT7-7 cloning, fast-performance liquid chromatography, and protein purification. We thank Roger Monier, Catherine Dargemont, Robert Vigne, Etienne Decroly, Gilles Quérat, and Corinne Ronfort for fruitful discussions during this study. We are indebted to François Grateau, Hospices Civils de Lyon, for the financing of the MegaView II TEM camera and automatic MT-X ultramicrotome (RMC EM Products Group, Ventana Medical Systems, Inc., Tucson, Ariz.).

REFERENCES

1. **Akhmanova, A., T. Verkerk, A. Langeveld, F. Grosveld, and N. Galijart.** 2000. Characterisation of transcriptionally active and inactive chromatin domains in neurons. J. Cell Sci. **113:**4463–4474.
2. **Bouyac-Bertoia, M., J. D. Dvorin, R. A. Fouchier, Y. Jenkins, B. E. Meyer, L. I. Wu, M. Emerman, and M. H. Malim.** 2001. HIV-1 infection requires a functional integrase NLS. Mol. Cell **7:**1025–1035.
3. **Breilling, A., B. M. Turner, M. E. Bianchi, and V. Orlando.** 2001. General transcription factors bind promoters repressed by *Polycomb* group proteins. Nature **412:**651–655.
4. **Brown, J. L., D. Mucci, M. Whiteley, M. L. Dirksen, and J. A. Kassis.** 1998. The *Drosophila Polycomb* group gene pleiohomeotic encodes a sequence-specific DNA binding protein with homology to the multifunctional mammalian transcription factor YY1. Mol. Cell **7:**1057–1064.
5. **Brown, P. O.** 1997. Integration. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.
6. **Bukrinsky, M. I., S. Haggerty, M. P. Dempsey, N. Sharova, A. Adzhubei, L. Spitz, P. Lewis, D. Goldfarb, M. Emerman, and M. Stevenson.** 1993. A nuclear localization signal within HIV-1 matrix protein that governs infection of non-dividing cells. Nature **365:**666–669.
7. **Bukrinsky, M. I., N. Sharova, T. L. McDonald, T. Pushkarskaya, W. G. Tarpley, and M. Stevenson.** 1993. Association of integrase, matrix, and reverse transcriptase antigens of human immunodeficiency virus type 1 with viral nucleic acids following acute infection. Proc. Natl. Acad. Sci. USA **90:**6125–6129.
8. **Carrière, C., B. Gay, N. Chazal, N. Morin, and P. Boulanger.** 1995. Sequence requirement for encapsidation of deletion mutants and chimeras of human immunodeficiency virus type 1 Gag precursor into retrovirus-like particles. J. Virol. **69:**2366–2377.
9. **Carteau, S., S. C. Batson, L. Poljak, J.-F. Mouscadet, H. de Rocquigny, J.-L. Darlix, B. P. Roques, E. Käs, and C. Auclair.** 1997. Human immunodeficiency virus type 1 nucleocapsid protein specifically stimulates Mg^{2+}-dependent DNA integration in vitro. J. Virol. **71:**6225–6229.
10. **Carteau, S., R. J. Gorelick, and F. D. Bushman.** 1999. Coupled integration of human immunodeficiency virus type 1 cDNA ends by purified integrase in vitro: stimulation by the viral nucleocapsid protein. J. Virol. **73:**6670–6679.
11. **Carteau, S., C. Hoffmann, and F. D. Bushman.** 1998. Chromosome structure and human immunodeficiency virus type 1 cDNA integration: centromeric alphoid repeats are a disfavored target. J. Virol. **72:**4005–4014.
12. **Chazal, N., B. Gay, C. Carrière, J. Tournier, and P. Boulanger.** 1995. Human immunodeficiency virus type 1 MAp17 deletion mutants expressed in baculovirus-infected cells: *cis* and *trans* effects on the Gag precursor assembly pathway. J. Virol. **69:**365–375.
13. **Chicurel, M.** 2000. Probing HIV's elusive activities within the host cell. Science **290:**1876–1879.
14. **Cortay, J.-C., D. Nègre, M. Scarabel, T. M. Ramseler, N. B. Vartak, J. Reizer, and A. J. Cozzone.** 1994. In vitro asymmetric binding of the pleiotropic regulatory protein, FruR, to the *ace* operator controlling glyoxylate shunt enzyme synthesis. J. Biol. Chem. **269:**14885–14891.
15. **Crowe, D. T., H. Chiu, S. Fong, and I. L. Weissmann.** 1994. Regulation of the avidity of integrin $\alpha_4\beta_7$ by the β_7 cytoplasmic domain. J. Biol. Chem. **269:**14411–14418.
16. **Cullen, B.** 2001. Journey to the center of the cell. Cell **105:**697–700.
17. **Denisenko, O. N., and K. Bomsztyk.** 1997. The product of the murine homolog of the *Drosophila extra sex combs* gene displays transcriptional repressor activity. Mol. Cell. Biol. **17:**4707–4717.
18. **Dvorin, J. D., P. Bell, G. G. Maul, M. Yamashita, M. Emerman, and M. H. Malim.** 2002. Reassessment of the roles of integrase and the central DNA flap in human immunodeficiency virus type 1 nuclear import. J. Virol. **76:**12087–12096.
19. **Farnet, C. M., and F. D. Bushman.** 1997. HIV-1 cDNA integration: requirement of HMG I(Y) protein for function of preintegration complexes in vitro. Cell **88:**483–492.
20. **Farnet, C. M., and W. A. Haseltine.** 1990. Integration of human immunodeficiency virus type 1 DNA in vitro. Proc. Natl. Acad. Sci. USA **87:**4164–4168.
21. **FitzGerald, D. P., and W. Bender.** 2001. Polycomb group repression reduces DNA accessibility. Mol. Cell. Biol. **21:**6585–6597.
22. **Fouchier, R. A., B. E. Meyer, J. H. Simon, U. Fischer, A. V. Albright, F. Gonzales-Scarano, and M. H. Malim.** 1998. Interaction of the human immunodeficiency virus type 1 Vpr protein with the nuclear pore complex. J. Virol. **72:**6004–6013.
23. **Freed, E. O.** 1998. HIV-1 Gag proteins: diverse functions in the virus life cycle. Virology **251:**1–15.
24. **Gallay, P., T. Hope, D. Chin, and D. Trono.** 1997. HIV-1 infection of nondividing cells through the recognition of integrase by the importin/karyopherin pathway. Proc. Natl. Acad. Sci. USA **94:**9825–9830.
25. **Gallay, P., V. Stitt, C. Mundy, M. Oettinger, and D. Trono.** 1996. Role of the karyopherin pathway in human immunodeficiency virus type 1 nuclear import. J. Virol. **70:**1027–1032.
26. **Gallay, P., S. Swingler, C. Aiken, and D. Trono.** 1995. HIV-1 infection of nondividing cells: C-terminal tyrosine phosphorylation of the viral matrix protein is a key regulator. Cell **80:**379–388.
27. **Gallay, P., S. Swingler, J. Song, F. Bushman, and D. Trono.** 1995. HIV nuclear import is governed by the phosphotyrosine-mediated binding of matrix to the core domain of integrase. Cell **83:**569–576.
28. **Gao, K., R. J. Gorelick, D. G. Johnson, and F. Bushman.** 2003. Cofactors for human immunodeficiency virus type 1 cDNA integration in vitro. J. Virol. **77:**1598–1603.
29. **Gay, B., J. Tournier, N. Chazal, C. Carrière, and P. Boulanger.** 1998. Morphopoietic determinants of HIV-1 GAG particles assembled in baculovirus-infected cells. Virology **247:**160–169.
30. **Goff, S. P.** 2001. Intracellular trafficking of retroviral genomes during the early phase of infection: viral exploitation of cellular pathways. J. Gene Med. **3:**517–528.
31. **Goodarzi, G., G. J. Im, K. Brackmann, and D. Grandgenett.** 1995. Concerted integration of retrovirus-like DNA by human immunodeficiency virus type 1 integrase. J. Virol. **69:**6090–6097.
32. **Guan, K. L., and J. E. Dixon.** 1991. Eukaryotic proteins expressed in *Escherichia coli*: an improved thrombin cleavage and purification procedure of fusion proteins with glutathione *S*-transferase. Anal. Biochem. **192:**262–267.
33. **Gutjarhr, T., E. Frei, S. C., S. Baumgatner, A. H. White, and M. Noll.** 1995. The polycomb-group gene, extra sex combs, encodes a nuclear member of the WD40 repeat family. EMBO J. **14:**4296–4306.
34. **Higuchi, R., B. Krummel, and R. K. Saiki.** 1988. A general method of in vitro preparation and specific mutagenesis of DNA fragments: study of protein and DNA interactions. Nucleic Acids Res. **16:**7351–7367.
35. **Hindmarsh, P., and J. Leis.** 1999. Retroviral DNA integration. Microbiol. Mol. Biol. Rev. **63:**836–843.
36. **Hong, S. S., and P. Boulanger.** 1995. Protein ligands of human adenovirus type 2 outer capsid identified by biopanning of a phage-displayed peptide

library on separate domains of WT and mutant penton capsomers. EMBO J. **14:**4714–4727.
37. **Hong, S. S., L. Karayan, J. Tournier, D. T. Curiel, and P. A. Boulanger.** 1997. Adenovirus type 5 fiber knob binds to MHC class I α2 domain at the surface of human epithelial and B lymphoblastoid cells. EMBO J. **16:**2294–2306.
38. **Horten, R. M., H. D. Hunt, S. N. Ho, J. K. Pullen, and P. L. R.** 1989. Engineering hybrid genes without the use of restrictions enzymes: gene splicing by overlap extension. Gene **77:**61–68.
39. **Huvent, I., S. S. Hong, C. Fournier, B. Gay, J. Tournier, C. Carriere, M. Courcoul, R. Vigne, B. Spire, and P. Boulanger.** 1998. Interaction and co-encapsidation of HIV-1 Vif and Gag recombinant proteins. J. Gen. Virol. **79:**1069–1081.
40. **Kalpana, G. V., S. Marmon, W. Wang, G. R. Crabtree, and S. P. Goff.** 1994. Binding and stimulation of HIV-1 integrase by a human homolog of yeast transcription factor SNF5. Science **266:**2002–2006.
41. **Lee, M. S., and R. Craigie.** 1998. A previously unidentified host protein protects retroviral DNA from autointegration. Proc. Natl. Acad. Sci. USA **95:**1528–1533.
42. **Leh, H., P. Brodin, J. Bischerour, E. Deprez, P. Tauc, J.-C. Brochon, E. LeCam, D. Coulaud, C. Auclair, and J.-F. Mouscadet.** 2000. Determinants of Mg^{2+}-dependent activities of recombinant human immunodeficiency virus type 1 integrase. Biochemistry **39:**9285–9294.
43. **Lin, C.-W., and A. Engelman.** 2003. The barrier-to-autointegration factor is a component of functional human immunodeficiency virus type 1 preintegration complexes. J. Virol. **77:**5030–5036.
44. **Lutzke, R. A., and R. H. Plasterk.** 1998. Structure-based mutational analysis of the C-terminal DNA-binding domain of the human immunodeficiency virus type 1 integrase: critical residues for protein oligomerization and DNA binding. J. Virol. 4841–4848.
45. **Maertens, G., P. Cherepanov, W. Pluymers, K. Busschots, E. De Clercq, Z. Debyser, and Y. Engelborghs.** 2003. LEDGF/p75 is essential for nuclear and chromosomal targeting of HIV-1 integrase in human cells. J. Biol. Chem. **278:**372–381.
46. **Manié, S. N., A. Astier, D. Wang, J. S. Phifer, J. Chen, A. I. Lazarovits, C. Morimoto, and A. S. Freedman.** 1996. Stimulation of tyrosine phosphorylation after ligation of β7 and β1 integrins on human B cells. Blood **87:**1855–1861.
47. **Margottin, F., S. P. Bour, H. Durand, L. Selig, S. Benichou, V. Richard, D. Thomas, K. Strebel, and R. Benarous.** 1998. A novel human WD protein, h-beta TrCP, that interacts with HIV-1 Vpu connects CD4 to the ER degradation pathway through an F-box motif. Mol. Cell **1:**565–574.
48. **Miller, M. D., and F. D. Bushman.** 1995. Ini1 for integration? Curr. Biol. **5:**368–370.
49. **Miller, M. D., C. M. Farnet, and F. D. Bushman.** 1997. Human immunodeficiency virus type 1 preintegration complexes: studies of organization and composition. J. Virol. **71:**5382–5390.
50. **Neer, E. J., C. J. Schmidt, R. Nambudripad, and T. F. Smith.** 1994. The ancient regulatory-protein family of WD-repeat proteins. Nature **371:**297–300.
51. **Ng, J., R. Li, K. Morgan, and J. Simon.** 1997. Evolutionary conservation and predicted structure of the *Drosophila* extra sex combs repressor protein. Mol. Cell. Biol. **17:**6663–6672.
52. **Petit, C., O. Schwartz, and F. Mammano.** 2000. The karyophilic properties of human immunodeficiency virus type 1 integrase are not required for nuclear import of proviral DNA. J. Virol. **74:**7119–7126.
53. **Petit, C., O. Schwartz, and F. Mammano.** 1999. Oligomerization within virions and subcellular localization of human immunodeficiency virus type 1 integrase. J. Virol. **73:**5079–5088.
54. **Peytavi, R., S. S. Hong, B. Gay, A. Dupuy d'Angeac, L. Selig, S. Bénichou, R. Benarous, and P. Boulanger.** 1999. H-EED, the product of the human homolog of the murine *eed* gene, binds to the matrix protein of HIV-1. J. Biol. Chem. **274:**1635–1645.
55. **Pirrotta, V.** 1998. Polycombing the genome: Pcg, trxG, and chromatin silencing. Cell **93:**333–336.
56. **Priet, S., J.-M. Navarro, N. Gros, G. Quérat, and J. Sire.** 2003. Functional role of HIV-1 virion-associated uracil DNA glycosylase 2 in the correction of G:U mispairs to G:C pairs. J. Biol. Chem. **278:**4566–4571.
57. **Pruss, D., F. D. Bushman, and A. P. Wolffe.** 1994. Human immunodeficiency virus integrase directs integration to sites of severe DNA distortion within the nucleosome core. Proc. Natl. Acad. Sci. USA **91:**5913–5917.
58. **Pruss, D., R. Reeves, F. D. Bushman, and A. P. Wolffe.** 1994. The influence of DNA and nucleosome structure on integration events directed by HIV integrase. J. Biol. Chem. **269:**25031–25041.
59. **Pryciak, P. M., and H. E. Varmus.** 1992. Nucleosomes, DNA-binding proteins, and DNA sequence modulate retroviral integration target site selection. Cell **69:**769–780.
60. **Reil, H., A. A. Bukovsky, H. R. Gelderblom, and H. G. Goettlinger.** 1998. Efficient HIV-1 replication can occur in the absence of the viral matrix protein. EMBO J. **17:**2699–2708.
61. **Rietzler, M., M. Bittner, W. Kolanus, A. Schuster, and B. Holzmann.** 1998. The human WD repeat protein WAIT-1 specifically interacts with the cytoplasmic tails of β7-integrins. J. Biol. Chem. **273:**27459–27466.
62. **Satjin, D. P. E., K. M. Hamer, J. den Blaauwen, and A. P. Otte.** 2001. The Polycomb group protein EED interacts with YY1, and both proteins induce neural tissue in *Xenopus* embryos. Mol. Cell. Biol. **21:**1360–1369.
63. **Satjin, D. P. E., and A. P. Otte.** 1999. RING1 interacts with multiple Polycomb group proteins and displays tumorigenic activity. Mol. Cell. Biol. **19:**57–68.
64. **Saurin, A. J., Z. Shao, H. Erdjument-Bromage, P. Tempst, and R. E. Kingston.** 2001. A *Drosophila* Polycomb group complex includes Zeste and dTAFII proteins. Nature **412:**655–660.
65. **Scherdin, U., K. Rhodes, and M. Breindl.** 1990. Transcriptionally active genome regions are preferred targets for retrovirus integration. J. Virol. **64:**907–912.
66. **Schroder, A. R., P. Shinn, H. Chen, C. Berry, J. R. Ecker, and F. Bushman.** 2002. HIV-1 integration in the human genome favors active genes and local hotspots. Cell **110:**521–529.
67. **Schumacher, A., C. Faust, and T. Magnuson.** 1996. Positional cloning of a global regulator of anterior-posterior patterning in mice. Nature **383:**250–253.
68. **Schumacher, A., O. Lichtarge, S. Schwartz, and T. Magnusson.** 1998. The murine polycomb-group *eed* and its human orthologue: functional implications of evolutionary conservation. Genomics **54:**79–88.
69. **Schwartz, O., A. Dautry-Varsat, B. Goud, V. Marechal, A. Subtil, J. M. Heard, and O. Danos.** 1995. Human immunodeficiency virus type 1 Nef induces accumulation of CD4 in early endosomes. J. Virol. **69:**528–533.
70. **Sewalt, R. G. A. B., J. van der Vlag, M. J. Gunster, K. M. Hamer, J. L. den Blaauwen, D. P. E. Satijn, T. Hendrix, R. van Driel, and A. P. Otte.** 1998. Characterization of interactions between the mammalian Polycomb-group proteins Enx1/EZH2 and EED suggests the existence of different mammalian Polycomb-group protein complexes. Mol. Cell. Biol. **18:**3586–3595.
71. **Sherman, M. P., and W. C. Greene.** 2002. Slipping through the door: HIV entry into the nucleus. Microbes Infect. **4:**67–73.
72. **Shi, Y., E. Seto, L.-S. Chang, and T. Shenk.** 1991. Transcriptional repression by YY1, a human GLI-Kruppel-related protein, and relief of repression by adenovirus E1A. Cell **67:**377–388.
73. **Showell, C., and V. T. Cunliffe.** 2002. Identification of putative interaction partners for the *Xenopus* Polycomb-group protein Xeed. Gene **291:**95–104.
74. **Simon, J.** 1995. Locking in stable states of gene expression: transcriptional control during *Drosophila* development. Curr. Opin. Cell Biol. **7:**376–385.
75. **Smith, G. P., and J. K. Scott.** 1993. Libraries of peptides and proteins displayed on filamentous phage. Methods Enzymol. **217:**228–257.
76. **Spillane, C., C. MacDouglas, C. Stock, C. Köhler, J.-P. Vielle-Calzada, S. M. Nunes, U. Grossniklaus, and J. Goodrich.** 2000. Interaction of the *Arabidopsis* Polycomb group proteins FIE and MEA mediates their common phenotypes. Curr. Biol. **10:**1535–1538.
77. **Suzuki, Y., and R. Craigie.** 2002. Regulatory mechanisms by which barrier-to-autointegration factor blocks autointegration and stimulates intermolecular integration of Moloney murine leukemia virus preintegration complexes. J. Virol. **76:**12376–12380.
78. **Turelli, P., V. Doucas, E. Craig, B. Mangeat, N. Klages, R. Evans, G. Kalpana, and D. Trono.** 2001. Cytoplasmic recruitment of INI1 and PML on incoming HIV preintegration complexes: interference with early steps of viral replication. Mol. Cell **7:**1245–1254.
79. **Van der Vlag, J., and A. P. Otte.** 1999. Transcriptional repression mediated by the human polycomb-group protein EED involves histone acetylation. Nat. Genet. **23:**474–478.
80. **Wang, J.-Y., H. Ling, W. Yang, and R. Craigie.** 2001. Structure of a two-domain fragment of HIV-1 integrase: implications for domain organization in the intact protein. EMBO J. **20:**7333–7343.
81. **Weidhaas, J. B., E. A. Angelichio, S. Fenner, and J. M. Coffin.** 2000. Relationship between retroviral integration and gene expression. J. Virol. **74:**8382–8389.
82. **Withers-Ward, E. S., Y. Kitamura, J. P. Barnes, and J. M. Coffin.** 1994. Distribution of targets for avian retrovirus DNA integration in vivo. Genes Dev. **8:**1473–1487.
83. **Wu, X., Y. Li, B. Crise, and S. M. Burgess.** 2003. Transcription start regions in the human genome are favored targets for MLV integration Science **300:**1749–1751.
84. **Yung, E., M. Sorin, A. Pal, E. Craig, A. Morozov, O. Delattre, J. Kappes, D. Ott, and G. V. Kalpana.** 2001. Inhibition of HIV-1 virion production by a transdominant mutant of integrase interactor 1. Nat. Med. **8:**920–926.
85. **Zennou, V., C. Petit, D. Guetard, U. Nerhbass, L. Montagnier, and P. Charneau.** 2000. HIV-1 genome nuclear import is mediated by a central DNA flap. Cell **101:**173–185.
86. **Zou, S., and D. F. Voytas.** 1997. Silent chromatin determines target preference of the *Saccharomyces* retrotransposon Ty5. Proc. Natl. Acad. Sci. USA **94:**7412–7416.

B/ SUREXPRESSION ET ESSAIS DE CRISTALLISATION DE EED SEULE ET EN COMPLEXE AVEC LA MATRICE, L'INTEGRASE ET NEF

Les moyens mis en œuvre afin de surexprimer la protéine EED et ses partenaires viraux seront analysés dans de ce chapitre.

La résolution de la structure d'une protéine par diffraction aux rayons X nécessite de surexprimer la protéine d'intérêt en quantité suffisante pour entreprendre des études de cristallogenèse. Les stratégies couramment employées comprennent l'expression en système bactérien des protéines d'intérêt (ce qui évite les modifications post-traductionnelles). Ces dernières sont exprimées avec des étiquettes d'affinité Histidine ou GST (Glutathion-S-Transférase) afin de faciliter leur purification avant cristallisation. *E. coli* est l'hôte le plus utilisé dans le cadre d'une expression hétérologue (Baneyx, 1999). Il est en effet le mieux caractérisé et il ne nécessite pas d'équipements lourds. Il présente un coût réduit, un nombre important de vecteurs disponibles et un temps de génération efficace (le temps de génération correspond au temps de doublement de la culture bactérienne ; il peut varier selon les souches utilisées et il est d'environ 20 min pour *E. coli)*. En général, *E. coli* surexprime les protéines recombinantes à un taux représentant 10 à 30 % de ses protéines totales.

L'obtention d'une solution protéique compatible avec des expériences de cristallogenèse implique une grande pureté de la protéine (99 %), une grande homogénéité (mêmes espèces moléculaires en présence), une monodispersité de la solution (état monomérique ou oligomérique unique) et généralement une grande quantité de matériel biologique (>10 mg). La protéine purifiée est portée à une concentration supérieure à 3 $mg.mL^{-1}$ pour être susceptible d'être cristallisée.

Dans le cadre de cette étude, une stratégie simple de production des protéines d'intérêt a été développée (Figure 19).

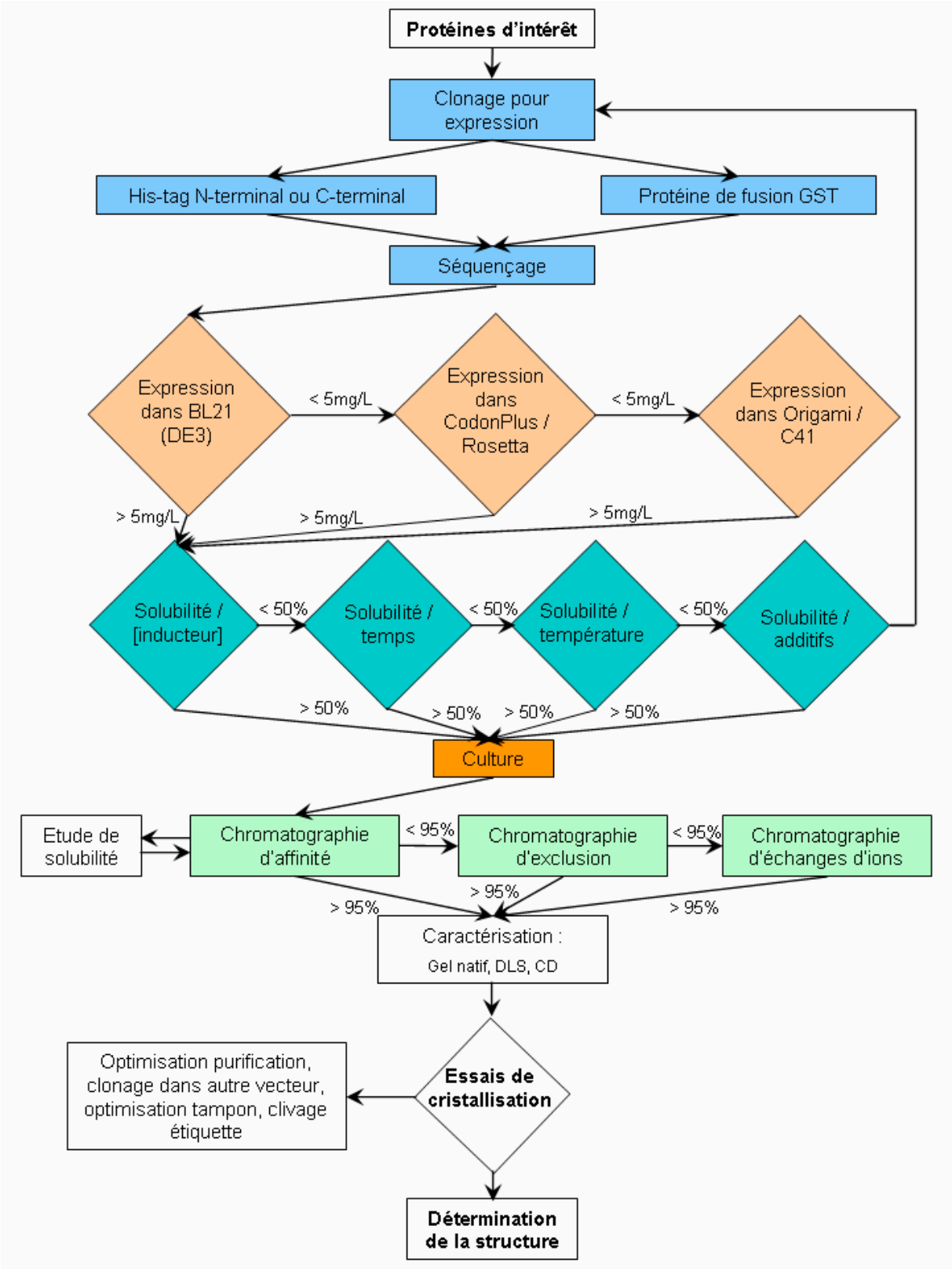


Figure 19: *Stratégie de production des protéines recombinantes chez E.coli.*

GST : Glutathion-S-Transférase ; DLS : Dynamic Light Scattering ; CD : Circular Dichroism.

1. Surexpression de la protéine EED dans *E. coli* :

Matériels et méthodes

- *Clonage*

Préparation plasmidique

Les préparations plasmidiques ont été réalisées selon la méthode de préparation quantitative «MIDI-PREP» (QIAGEN). Les bactéries sont cultivées sur la nuit à 37°C sous agitation dans 30 à 150 mL de milieu sélectif, puis centrifugées 15 min à 5000g. Le culot est lavé, repris dans 4 mL de tampon P1 (Tris-HCl 50 mM ; EDTA 10 mM). Les bactéries sont lysées par addition de 4 mL de tampon P2 (NaOH 0,2 M ; SDS 1 %). Après 5 min d'incubation à température ambiante, 4 mL de tampon P3 (acétate de potassium 2,55 M pH 4,8) sont ajoutés au mélange et le tube est centrifugé à 4°C pendant 30 min à 20000 g. Une deuxième centrifugation du surnageant à 2000 g permet d'éliminer au maximum les particules en suspension dans le lysat. Le surnageant est déposé sur une colonne Qiagen-100 pré-équilibrée par 3 mL de tampon QBT pH 7,0 (NaCl 750 mM ; MOPS 50 mM ; EtOH 15 % ; Triton-X100 0,15 %). La résine est lavée deux fois par 5 mL de tampon QC pH 7,0 (NaCl 1 M ; MOPS 50 mM ; EtOH 15 %) L'ADN est élué par 5 mL de tampon QF à pH 8,2 (NaCl 1,25 M ; MOPS 50 mM ; EtOH 15 %). L'ADN contenu dans l'éluat est concentré par précipitation à l'isopropanol (0,7 volume). Le culot, obtenu après centrifugation à température ambiante, est lavé, séché et enfin repris dans 50 µL d'eau stérile.

Dosage des acides nucléiques

Les solutions d'acides nucléiques sont dosées par spectrophotométrie à 260 nm. Une unité de DO correspond à 50 $\mu g.mL^{-1}$ d'ADN bicaténaire et à 40 $\mu g.mL^{-1}$ d'ARN ou d'ADN monocaténaire. Cette quantification n'est valable que dans la mesure où le rapport $DO_{260\ nm}$ / $DO_{280\ nm}$ est égal à 1,8 pour les ADN et 2 pour les ARN. Ce rapport est une estimation de la pureté de la solution par rapport aux contaminations protéiques.

Extraction de fragments d'ADN par agarose «low-melting»

Cette technique repose sur l'utilisation d'un type d'agarose pouvant fondre à une température relativement basse, environ 65°C, c'est à dire bien en dessous de la température de fusion des ADN double-brins. Cette propriété permet la récupération de l'ADN à partir des gels (Wieslander, 1979).

Des fragments d'ADN de 100 à 10000 pb peuvent être isolés avec un bon rendement à partir de bandes préalablement séparées sur gel d'agarose. L'électrophorèse doit avoir lieu à 4°C et à un voltage constant faible (30 volts). Après migration, le gel d'agarose est placé sur une table U.V. et incisé au scalpel autour de la bande à récupérer. Le morceau d'agarose prélevé est mis à fondre dans un tube Eppendorf à 65°C pendant 5 min en présence de 5 volumes de tampon Tris-HCl 20 mM pH 8,0 ; EDTA 1 mM.

Ligation des molécules d'ADN

L'ADN ligase du phage T4 (Boehringer) catalyse cette réaction qui est réalisée dans 10 µL de milieu réactionnel. Dans le cas d'extrémités cohésives, le mélange réactionnel ramené à 20 µL est d'abord chauffé 5 min à 45°C, puis refroidi dans la glace avant l'addition de l'enzyme et du tampon de ligation (Tris-HCl 20 mM ; $MgCl_2$ 10 mM ; EDTA 1 mM ; DTT 10 mM ; ATP 0,6 mM). La durée d'incubation diffère selon le type de molécule. Dans le cas d'extrémités franches, la ligation se déroule sur la nuit à température ambiante alors que dans le cas d'extrémités cohésives, la ligation se fait pendant 1 à 4h à 16°C ou 16h à 12°C. De façon générale et simplifiée, la formule suivante peut être appliquée :

$$\text{quantité d'insert} = 5 \times \text{quantité de vecteur} \times PM_{insert} / PM_{vecteur}.$$

Déphosphorylation des extrémités

La phosphatase alcaline de veau (Biolabs) permet de déphosphoryler les extrémités proéminentes 3' ou 5' phosphate des acides nucléiques ainsi que les extrémités franches. Elle est surtout utilisée sur les vecteurs de clonage pour empêcher la ligation de ces molécules sur elles-mêmes lorsque les extrémités sont cohésives. La réaction se déroule dans un volume de 10 µL pour 0,5 µg d'ADN directement dans le tampon TAB (évitant ainsi une précipitation intermédiaire du plasmide). L'enzyme est ajoutée à raison de 0,1 unité dans le cas des extrémités 5' et de 1 unité dans le cas des extrémités 3' ou franches. Après 1 h à 60°C, la réaction est arrêtée par addition d'EDTA 5 mM final suivie d'un chauffage de 15 min à 70°C.

- *Expression*

Vecteurs de clonage et d'expression

La protéine a été exprimée sous forme d'une protéine de fusion avec la GST (protéine GST-EED Full Lenght (GST-EED FL)). La séquence codante (résidus 1 à 441) a été clonée dans le vecteur d'expression pGEX-4T1 (Figure 20) entre les sites EcoR1 et Not1.

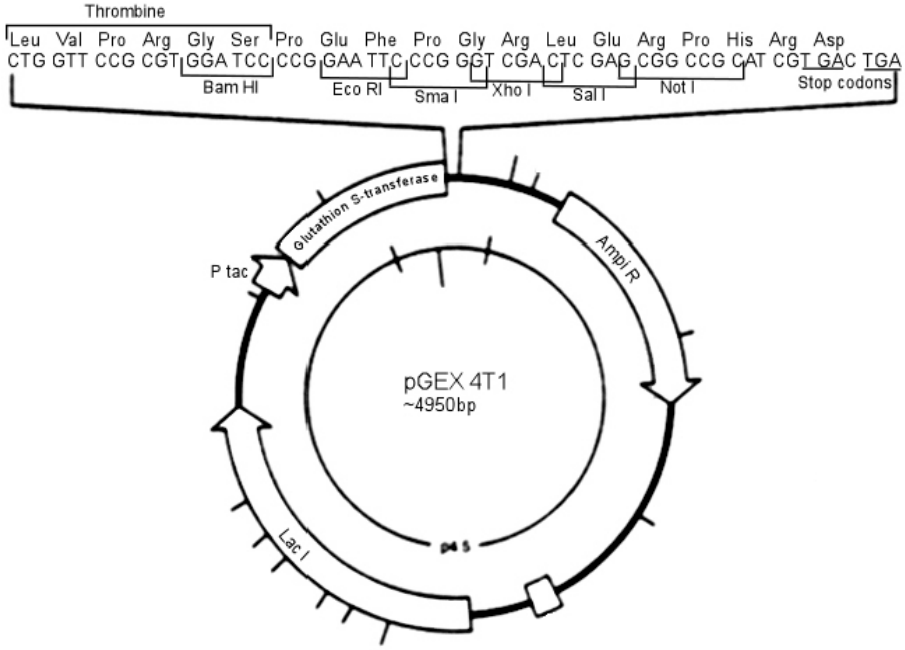


Figure 20: *Le plasmide pGEX-4T1 est un vecteur d'expression permettant la fusion de gène avec celui de la GST. Il possède un site multiple de clonage en aval de la séquence codant la GST et contenant 6 sites de restriction pour 6 endonucléases différentes, ainsi qu'une séquence de clivage protéique par la Thrombine. Le promoteur inductible en amont du gène GST ainsi que la séquence codant l'inhibiteur de l'opéron lactose permettent l'induction massive de l'expression du gène de fusion chez E.coli. Une origine de réplication et le gène de résistance à l'ampicilline complètent ce plasmide pour permettre sa multiplication, sa transmission et la sélection des bactéries transformées.*

La protéine a également été exprimée sous forme d'une protéine portant une étiquette histidine en C-terminal (EED-$(His)_6$). La séquence codante de EED (résidus 1 à 441) a été clonée dans le vecteur d'expression pT7-7_{His6} (Figure 21) entre les sites de restriction Pst1 et Nde1. Le plasmide pT7-7_{His6} est dérivé du plasmide pT7-7 où six codons Histidines ont été insérés entre les sites de restriction Pst1 et HindIII, avec un codon STOP suivant le dernier codon His.

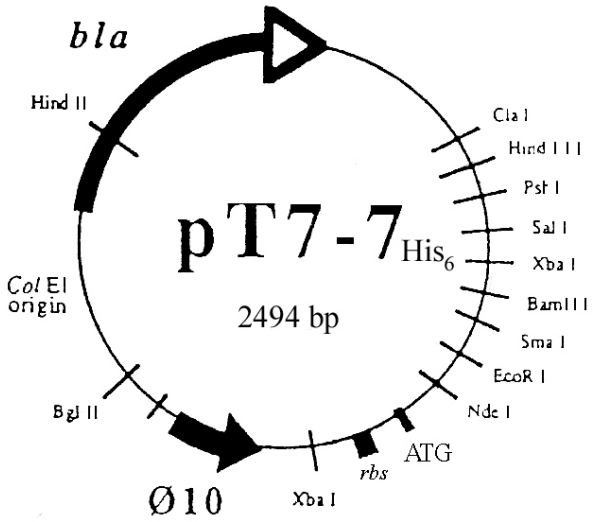


Figure 21 : *Le plasmide pT7-7$_{His6}$ est un vecteur d'expression permettant la fusion de gène avec une étiquette de six histidines en C-terminal de la séquence codante. Il possède un site multiple de clonage en amont de l'étiquette et contenant 7 sites de restriction pour 7 endonucléases différentes. Ce plasmide contient le promoteur Ø 10 de la RNA polymérase T7. Une origine de réplication et le gène de résistance à l'ampicilline complètent ce plasmide pour permettre sa multiplication, sa transmission et la sélection des bactéries transformées.*

Pour toutes les constructions, les régions des plasmides codant les protéines EED recombinantes ont été séquencées avant de débuter les tests d'expression.

Souches utilisées

Le génotype de chacune des souches utilisées au cours de cette étude est précisé ci-dessous (Table 4) :

Souche	Caractéristiques	Génotype
BL21 (DE3)	Exprime la polymérase T7 sous contrôle du promoteur lacUV5 inductible par l'IPTG	F^- ompT hsd S_B (rb^- mB^-) gal dcm (DE3)
BL21 (DE3) pLysS	Possède un plasmide additionnel codant pour le lysozyme T7 et permettant de supprimer avant induction l'expression de la polymérase T7	F^- ompT hsdSB (rb^- mB^-) gal dcm (DE3) pLysS (CmR)
BL21 star (DE3)	Dérive de BL21 (DE3) mais n'exprime pas la protéase *lon* permettant en association avec la mutation *ompT*, De réduire l'activité protéasique responsable de la dégradation aberrante des protéines recombinantes.	F^- ompT hsd S_B (r^-_B m^-_B) gal dcm rne 131(DE3)
Origami (DE3)	Possède une mutation dans les gènes de la thioredoxine réductase (trxB) et de la	ara-leu7697 dlacX74

	glutathion réductase (gor) facilitant la formation de ponts disulfures dans le cytoplasme	phoAPvuII phoR araD139 ahpC gale galK rspL F"[lac+ (lacIq) pro] gor522 Tn10 (TcR) trxB Kan (DE3)
Rosetta (DE3)	Possède un plasmide chloramphénicol résistant codant pour les ARN_t AGG, AGA, AUA, CUA, CCC et GGA, améliorant ainsi l'expression de protéines eucaryotes possédant ces codons rares	F^- ompT hsdSB (rb^- mB^-) gal dcm lacY1 (DE3) pRARE6 (CmR)
TOP10	hsdR permettant une transformation efficace d'ADN méthylé ou non méthylé	F^- mcrA (mrr^- hsdR MS^- mcrBC) 80lacZ M15 lacX74 recA1 ara 139 (ara-leu) 7697 galU galK rpsL (StrR) endA1 nupG
C41 (DE3)	Dérive de BL21 (DE3) mais présente au moins une mutation non caractérisée évitant la mort cellulaire lors de l'expression de protéine recombinante toxique	F^- ompT hsd S_B (rb^- mB^-) gal dcm (DE3)

Table 4 *: Caractéristiques des différentes souches bactériennes utilisées.*

Transformation des bactéries

Les bactéries réceptrices sont rendues compétentes par les sels de calcium. La souche est cultivée dans 30 mL de milieu Luria-Bertani en Erlenmeyer de 250 mL sous agitation à 37°C. Lorsque la DO à 600 nm de la culture en phase exponentielle est comprise entre 0,3 et 0,4, les bactéries sont centrifugées à 4°C pendant 5 min à 2800 g. Le culot est repris par la moitié du volume initial de la culture (environ 15 mL de $CaC1_2$ 50 mM) et incubé 1 h à 4°C dans la glace fondante. Les cellules sont de nouveau centrifugées à 4°C pendant 5 min à 2600 g et reprises soigneusement dans $1/10^{ème}$ de volume (environ 1,5 mL) de $CaCl_2$ 50 mM. Les cellules conservent leur compétence environ 48 h. L'ADN transformant (maximum 40 ng de plasmide circulaire) contenu dans un volume de 10 µL est mélangé à 100 µL de TCM (Tris-HCl 10 mM, pH 7,5 ; $CaCl_2$ 10 mM ; $MgCl_2$ 10 mM) et à 100 µL de cellules compétentes dans un tube de 1 mL stérile. La suspension est incubée 35 min dans la glace fondante puis soumise à un choc thermique de 2 min à 42°C. Après retour à température ambiante, 1,2 mL de LB sont rajoutés et les bactéries sont incubées pour une période de régénération de 45 min sous agitation douce à 37°C. Les bactéries sont alors récupérées par centrifugation 1 min à 2500 g et les culots sont repris par 100 µL de LB avant d'être étalées sur milieu sélectif. Les clones apparus après une nuit d'incubation à 37°C sont repiqués sur milieu sélectif ou conservées à - 80°C dans 50 % de glycérol.

Cultures bactériennes

Quelques colonies sont prélevées et mises dans 10 mL de préculture dans un milieu riche de Luria-Bertani (LB ; tryptone 1 %, extrait de levure 0,5 %, chlorure de sodium 0,5 %, SIGMA) additionné d'ampicilline à 100 $\mu g.mL^{-1}$ final (les milieux gélosés sont obtenus par addition de 15 $g.L^{-1}$ d'agar).à 37°C toute la nuit. Le lendemain, la préculture est ajoutée à 1 L de LB toujours avec ampicilline et la croissance bactérienne est poursuivie jusqu'à une densité optique à 600 nm ($D.O._{600}$) de 0,8. On ajoute ensuite de l'IPTG à 1 mM (Euromedex) et on induit à 37°C environ 3 h sous agitation à 200 rpm (incubateur Minitron, INFORS).

La concentration bactérienne est calculée par mesure de l'absorption optique (DO) à 600 nm la correspondance suivante (1 unité DO = 1 x 10^8 $cellules.mL^{-1}$). Tous les milieux sont stérilisés par autoclavage à 120°C pendant 20 min. L'antibiotique est ajouté stérilement au milieu refroidi (< 55°C).

Les souches transformées sont

Lyse bactérienne

Les bactéries issues d'une culture sont centrifugées à 4°C durant 30 min à 14000 g. Les culots sont conservés à -20°C après lavage à l'eau distillée, et ceci pour deux raisons :

- la congélation améliore l'étape de lyse bactérienne ultérieure

- des culots bactériens issus d'une même culture peuvent ainsi être utilisés pour plusieurs purifications successives, supprimant de ce fait un facteur d'hétérogénéité entre les différents lots de protéines purifiées.

Les bactéries sont alors remises en suspension dans un volume de tampon de lyse (Tris-HCl 50 mM pH 8,5 ; NaCl 300 mM ; Imidazole 10 mM ; glycérol 5 %) correspondant au $1/50^{ème}$ du volume de culture. La suspension est incubée 30 min à 4°C en présence d'un inhibiteur de protéase (Protease inhibitor cocktail tablets Complete™ EDTA-free, Roche) et de lysozyme à une concentration finale de 1 $mg.mL^{-1}$. Quatre cycles de 30 s de sonication, entrecoupés à chaque fois d'une incubation de 30 s à 4°C, sont effectués pour casser les bactéries. Les débris bactériens sont éliminés par centrifugation de 10 min à 10000 g. Les surnageants récupérés sont rassemblés et utilisés immédiatement.

- Purification par chromatographie d'affinité

Pour la protéine GST-EED (EED fusionnée à la GST (Glutathion-S-Transférase)) les chromatographies ont été réalisées à 15°C sur des colonnes GST-HiTrap de 1 mL (Amersham). Les tampons de fixation (PBS pH 7,4) et d'élution (Tris-HCl 50 mM pH 8 ; glutathion réduit 10 mM) sont filtrés et dégazés.

Pour la protéine comportant une étiquette histidine, les chromatographies ont été réalisées à 15°C sur des colonnes HiTrap de 1 mL (Amersham) préalablement chargée en nickel ($NiSO_4$ 0,1 M). Les tampons de fixation (Tris-HCl 50 mM pH 8,5 ; NaCl 300 mM ; Imidazole 10 mM ; glycérol 5 %) de lavage (Tris-HCl 50 mM pH 8,5 ; NaCl 1 M ; Imidazole10 mM) et d'élution (Tris-HCl 50 mM pH 8,5 ; Imidazole 1 M) sont filtrés et dégazés.

Les fractions les plus pures ont été rassemblées puis dialysées contre un tampon de stockage (Tris-HCl 50 mM pH 8,5) avant d'être concentrées jusqu'à 5 $mg.mL^{-1}$.

- Caractérisation des protéines

Electrophorèse en gel de polyacrylamide (SDS-PAGE)

La protéine est mise en présence de dodécylsulfate de sodium (SDS). Ce dernier forme un polyanion autour de la chaîne polypeptidique. Ces complexes présentent alors tous le même rapport charge / masse et ne diffèrent que par leur masse. Sous l'influence d'un champ électrique, ils vont migrer vers l'anode au travers des mailles d'un gel de polyacrylamide jouant le rôle de tamis moléculaire. Les chaînes de haut poids moléculaire migrent moins vite que celles de bas poids moléculaire (Laemmli, 1970).

Les électrophorèses sont réalisées avec le système Mini-PROTEAN 3 Electrophoresis System de BIO-RAD :

- Les gels de polyacrylamide ont une concentration en acrylamide de 8 % à 12 %
- Le tampon de migration est composé de Tris 50 mM pH 7,3 ; Glycine 192 mM et 0,1 % SDS.
- Le tampon de dépôt est composé de Tris 50 mM pH 8,5 ; SDS 2 % ; glycérol 1 M et marqueurs du front de migration (rouge de phénol et Serva Blue G250).

10 mM de DTT sont ajoutés extemporanément afin de travailler en conditions réductrices puis les échantillons sont placés 5 min au bain-marie à 95°C avant leur dépôt sur le gel de polyacrylamide.

Des marqueurs de masse moléculaire (Prestained SDS-PAGE Standards, BIO-RAD) sont déposés en même temps que les échantillons. La migration s'effectue pendant 1 h à 35 mA.

Le gel est coloré au bleu de Coomassie, puis décoloré dans une solution aqueuse d'éthanol (45 %) et d'acide acétique (7 %).

Electrotransfert sur membrane de nitrocellulose (Western blot)

Les protéines fractionnées en gel de polyacrylamide sont transférées sur nitrocellulose (Hybond-C Extra, Amersham Biosciences) par électrophorèse transverse (Mini Trans-Blot Electrophoretic Transfer Cell, BIO-RAD) pendant 1 h à 90 mA. Le transfert se fait dans un tampon Tris 25 mM pH 8,3 ; glycine 192 mM ; MeOH 20 % ; SDS 0,1 %. Les protéines transférées sur la feuille de nitrocellulose peuvent être visualisées par coloration réversible au rouge Ponceau. La feuille est immergée dans une solution de rouge Ponceau 0,2 %, acide trichloracétique 3 %, acide sulfosalicylique 3 % pendant 5 min puis rincée à l'eau distillée. Cette technique de coloration, peu sensible, est surtout utilisée pour contrôler l'efficacité du transfert et marquer la position des standards de masse moléculaire. Afin de réaliser une révélation par immunodétection, la membrane de nitrocellulose est saturée 1 h à température ambiante dans du PBS 1X (solution 10X: NaCl 80 g, KCl pH 7,2-7,4 80 g ; Na_2HPO_4 11,5 g ; KH_2PO_4 2 g ; H_2O qsp 1 L) et du lait écrémé à 5 % afin d'éliminer les interactions non spécifiques. La membrane est ensuite incubée 12 h à température ambiante avec l'anticorps primaire *ad hoc*. Après élimination de l'excès d'anticorps et 3 lavages successifs de 10 min dans 20 mL de TNT (Tris, NaCl, Tween : Tris-HCl 10 mM pH 8,0 ; NaCl 150 mM ; Tween-20 0,05 %) la membrane est incubée pendant 12 h à température ambiante avec l'anticorps secondaire *ad hoc* couplé à la phosphatase alcaline (SIGMA IgG AP conjugate). Après trois lavages successifs par 20 mL de TNT, la membrane est mise en présence des réactifs mélangés extemporanément : 33 µL de NBT (Nitro blue tetrazolium 50 $mg.mL^{-1}$ dans le diméthylformamide 70 %) et 16,5 µL de BCIP (5-bromo-4-chloro-3-indoyl phosphate 50 $mg.mL^{-1}$ dans le diméthylformamide 70 %) dans 5 mL de tampon AP (alkaline phosphatase : Tris-HCl 100 mM pH 9,5 ; NaCl 100 mM ; $MgCl_2$ 5 mM). Après une incubation d'environ 30 min à l'obscurité, la réaction colorée qui se développe est arrêtée par addition d'une solution stop: Tris-HCl 20 mM pH 8,0 ; EDTA 5 mM.

Dosage du matériel biologique

Les concentrations en protéines ont été mesurées en utilisant la méthode de Bradford (1976). Une courbe étalon, établie avec l'albumine du sérum de boeuf (BSA) permet la détermination de la concentration en protéine à partir de l'absorbance mesurée 595 nm.

Dans le cas de protéines pures, les concentrations ont également été déterminées par spectrométrie U.V. en mesurant leur absorption à 280 nm. L'application de la loi de Beer-Lambert (DO = ε.l.C) permet de calculer la concentration à partir du coefficient d'extinction molaire ε_{280} de la solution.

Le spectromètre utilisé est un BECKMAN™ DU 500. Le trajet optique des cellules utilisées est de 1 cm pour un volume de 1 mL ou 100 µL dans le cas des microcuves.

Diffusion de la lumière

Les études de DLS (Diffusion Light Scattering) ont été réalisées sur un appareil Zetasizer ZEN1600 de chez Malvern Instruments, à 25°C. Les protéines, concentrées autour de 1 $mg.mL^{-1}$, sont préalablement centrifugées 5 min à 5000 g. Leurs rayons hydrodynamiques sont déduits des coefficients de diffusion en utilisant l'équation de Stockes-Einstein. Les coefficients de diffusion sont eux-mêmes extrapolés de l'analyse de la décroissance de la fonction d'auto corrélation des intensités diffusées. Tous les calculs sont réalisés par le logiciel pilotant la machine.

Chrommatographie d'exclusion

0,5 mL de EED purifiée et concentrée à 5 $mg.mL^{-1}$ dans un tampon A (Tris-HCl 50 mM pH 8,5) sont chargés sur une colonne Superdex 200 10/300GL (Amersham). Les chromatographies sont réalisées à 15°C à un débit de 0,5 $mL.min^{-1}$. Les caractéristiques de la colonne, V_0 (volume d'exclusion) et V_t (volume total) sont respectivement déterminées avec du bleu dextran 2000 et de l'imidazole. Pour faire abstraction des dimensions de la colonne, chaque fraction est caractérisée par son Kav défini ainsi: $Kav = (V_e - V_0) / (V_t - V_0)$. Une représentation du Kav en fonction du log de la masse moléculaire ou du log du rayon hydrodynamique représente la droite d'étalonnage de la colonne. La colonne est calibrée avec différents standards de masses moléculaires solubilisées dans le tampon A : thyroglobuline (8,6 nm ; 669 kDa ± 15 %) ferritine (6,3 nm ; 440 kDa ± 15 %) catalase (5,2 nm ; 232 kDa ± 15 %) aldolase (4,6 nm ; 158 kDa ± 15 %) albumine du sérum de boeuf (3,5 nm ; 67 kDa ± 10 %) ovalbumine (2,8 nm ; 43 kDa ± 15 %) chymotrypsinogène (2,1 nm ; 25 kDa ± 25 %) et ribonuc1éase A (1,75 nm ; 13,7 kDa ± 15 %). La courbe d'étalonnage (Log Rs = -1,35 Kav +

1,05) et présentant un coefficient de corrélation R^2 = 0,99 a été utilisée afin d'extrapoler le rayon hydrodynamique de EED.

Résultats et discussion

La surexpression de EED chez *E.coli* doit permettre une production massive de cette protéine. Sa purification ultérieure doit en outre être la plus rapide et la plus efficace possible.
Pour ce faire, deux stratégies ont été envisagées :

- expression d'une protéine de fusion GST-EED possédant un site de clivage à la thrombine (séquence codante de la Glutathion-S-Transférase en 5' de la séquence codante de EED).

- expression d'une protéine recombinante flanquée d'une étiquette de six Histidines ($(His)_6$) à l'extrémité C-terminale de la séquence codante de EED.

Ces deux constructions répondent aux exigences d'un protocole de purification efficace par chromatographie d'affinité, respectivement sur colonne de glutathion et colonne de nickel.

Les moyens mis en œuvre afin de surexprimer la protéine de fusion GST-EED puis la protéine EED-$(His)_6$ seront analysés successivement en détail dans la suite de ce chapitre.

- *Clone GST-EED*

Une erreur de clonage introduisant un codon stop dans la séquence codante de EED a conduit dans un premier temps à l'expression d'un clone tronqué au résidu 349, clone appelé GST-EED DC349. Ce clone est cependant intéressant comme témoin négatif puisqu'il ne présente plus les propriétés d'interaction de EED avec MA et IN.

Ces deux clones présentent une surexpression des protéines de fusion GST-EED FL et GST-EED DC349 à un taux de surexpression représentant plus de 30 % des protéines totales de *E.coli* (Figure 22). Le maximum de protéine est obtenu 3 h après induction à 1 mM d'IPTG à une température de 37°C.

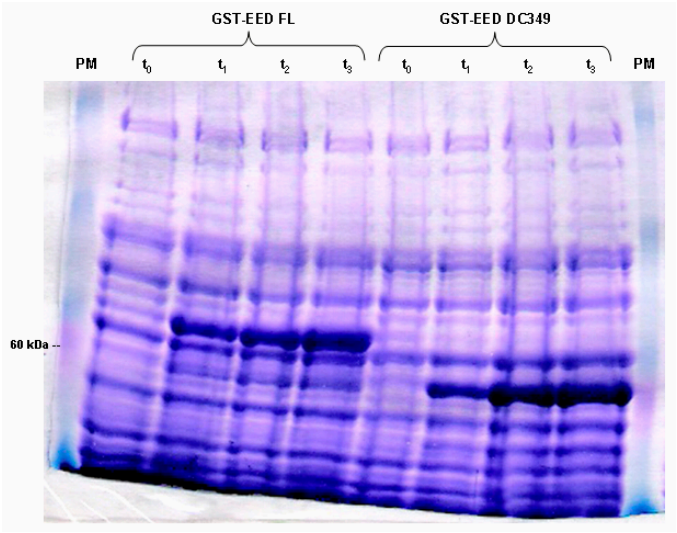


Figure 22 : *SDS-PAGE 10 % du profil de surexpression à 37°C des clones GST-EED FL et GST-EED DC349 à 0, 1, 2 et 3 h après induction à l'IPTG 1 mM. PM : poids moléculaire.*

Les rendements après purification sont faibles à cause d'une mauvaise solubilité de la protéine de fusion (cf. Figure 24 / Lysat). Ces résultats sont étonnant, car il a été reporté que la protéine GST pouvait amélioré la solubilité globale de la protéine de fusion (Purification of GST-Fusion Proteins by Magnetic Resin-Based MagneGST™ Particles. Marjeta *et al*., Promega Corporation).

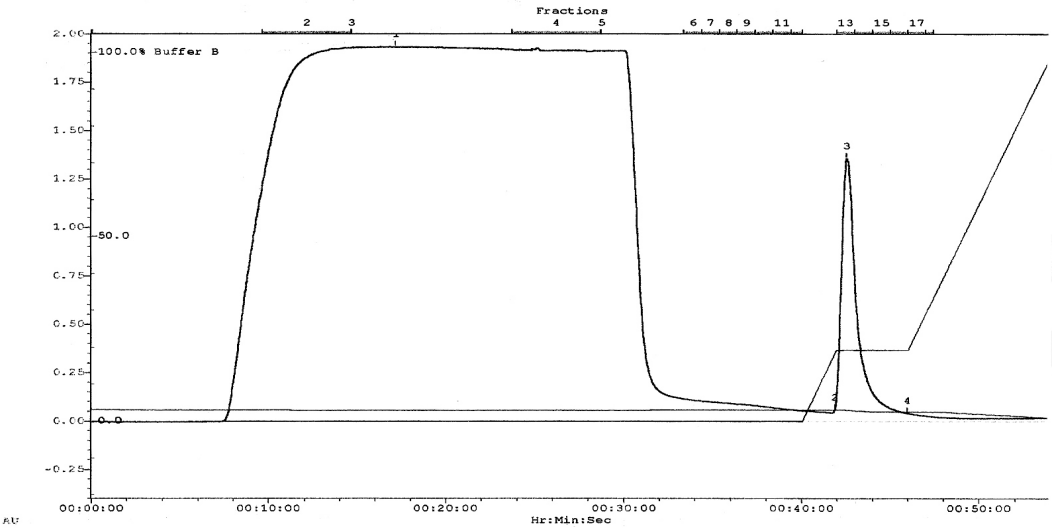


Figure 23 : *Chromatogramme de purification de GST-EED FL sur colonne de glutathion. L'absorption à 280 nm ainsi que le gradient du tampon d'élution sont représentés.*

Cette première étape de chromatographie d'affinité sur colonne de glutathion (Figure 23) n'a en outre pas pu être optimisée et les fractions les plus pures présentent de nombreux contaminants (Figure 24).

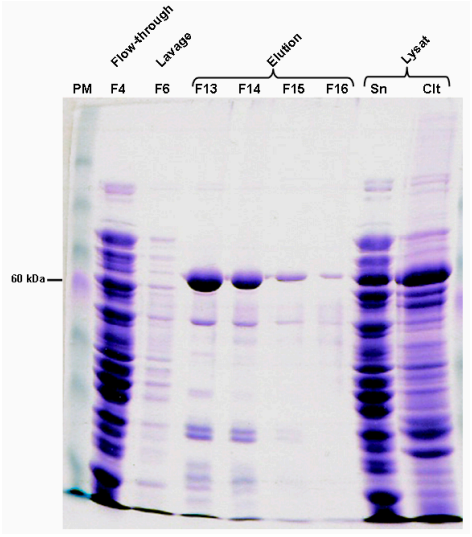


Figure 24 : *SDS-PAGE 10 % des fractions d'élution après chromatographie d'affinité sur colonne de glutathion (les fractions déposées correspondent au chromatogramme en figure 23). Clt : culot ; Sn : surnageant ; PM : poids moléculaire.*

Enfin, la partie GST de la protéine de fusion doit être éliminée pour l'étape ultérieure de cristallisation de la protéine EED. Les tests de clivage par la thrombine n'ont cependant jamais été concluants (résultats non présentés). Ceci est sans doute dû à la grande hétérogénéité de la solution avant clivage. En conclusion, l'utilisation du clone GST-EED FL a été abandonné au profit du clone EED-$(His)_6$.

- Clone EED-(His)$_6$

Les souches bactériennes utilisées pour la surexpression ont été dans un premier temps des souches *E.coli* BL21 (DE3) pLysS. Cette souche est connue pour permettre l'obtention de taux d'expression supérieurs à 10 mg.L^{-1} dans le cas de protéines toxiques pour l'hôte Cependant, nous avons observé que le taux de surexpression de EED-(His)$_6$ ne dépassait pas 5 à 10 % des protéines totales de *E.coli*.

Nous avons donc réalisé des essais avec les souches BL21 (DE3) CodonPlus-RP ou Rosetta (DE3) qui permettent l'expression de protéines dont la séquence codante présente des codons rares, ce qui est le cas pour EED. Ces souches co-expriment en effet les ARN de transferts pour ces codons rares. La présence de tels codons dans la séquence traduite diminue considérablement la vitesse d'élongation de la traduction et conduit à des taux d'expression inférieurs à 1 mg.L^{-1} (Wada *et al.*, 1992). Cependant, l'expression de EED dans les souches BL21 (DE3) CodonPlus-RP ou Rosetta (DE3) n'a jamais été concluante (expression indétectable, résultats non présentés).

L'expression de EED a finalement été satisfaisante par l'utilisation de la souche BL21 (DE3) qui permet un taux de surexpression représentant environ 30 % des protéines totales de *E.coli* (Figure 25).

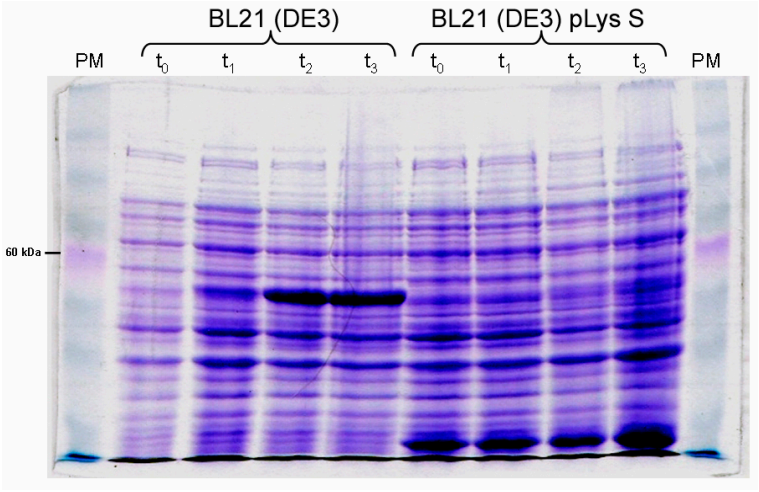


Figure 25 : *SDS-PAGE 10 % du profil de surexpression à 37°C du clone EED-(His)$_6$ à 0, 1, 2 et 3 h après induction à l'IPTG 1 mM dans les souches BL21 (DE3) ou BL21 (DE3) pLysS. PM : poids moléculaire.*

La purification par chromatographie d'affinité sur colonne de nickel (Figure 26) a été grandement améliorée par une étape de lavage à haute force ionique (NaCl 1 M). Ceci permet

d'éliminer plusieurs protéines contaminantes fixées de manière non spécifique sur la colonne de nickel (cf. Figure 27 / Lavage / F8 et F9).

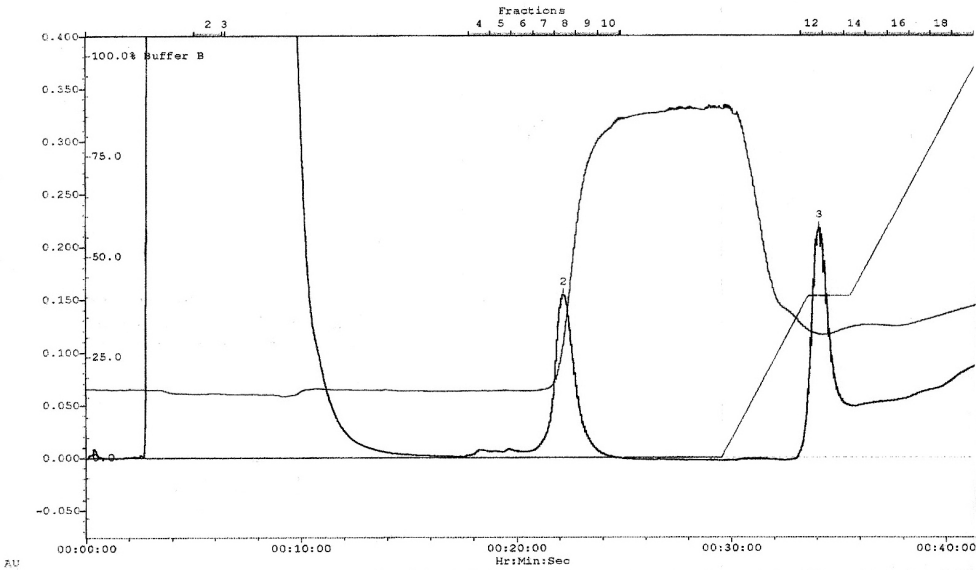


Figure 26 : *Chromatogramme de purification de EED-(His)$_6$ sur colonne de nickel. L'absorption à 280 nm, le gradient du tampon d'élution et la conductivité sont représentés.*

A l'inverse du clone GST-EED FL, EED-(His)$_6$ présente une bonne solubilité (cf. Figure 27 / Lysat). La protéine purifiée migre selon une seule bande en gel d'électrophorèse dénaturant (SDS-PAGE). Sa présence à la taille attendue a été vérifiée par Western blot en utilisant un anti-corps anti-His$_6$.

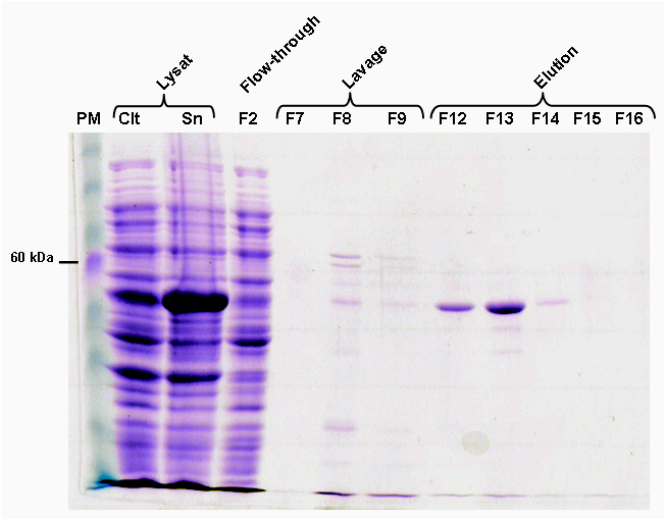


Figure 27 : *SDS-PAGE 10 % des fractions d'élution après chromatographie d'affinité sur colonne de nickel (les fractions déposées correspondent au chromatogramme en figure 26). Clt : culot ; Sn : surnageant ; PM : poids moléculaire.*

Les lots de protéine purifiée ont été analysés par diffusion de la lumière afin de caractériser l'état d'oligomérisation de EED (Figure 28). Il s'est avéré que EED-(His)$_6$ semble s'agréger

en solution compte tenu de son rayon hydrodynamique très élevé et calculé à 125 nm. L'indice de polydispersité de 0,35 de ces échantillons signifie en outre que les oligomères de EED sont très hétérogènes et présentent plusieurs assemblages de très haut poids moléculaire). Dans de telles conditions, la protéine EED-$(His)_6$ semble impropre à cristalliser.

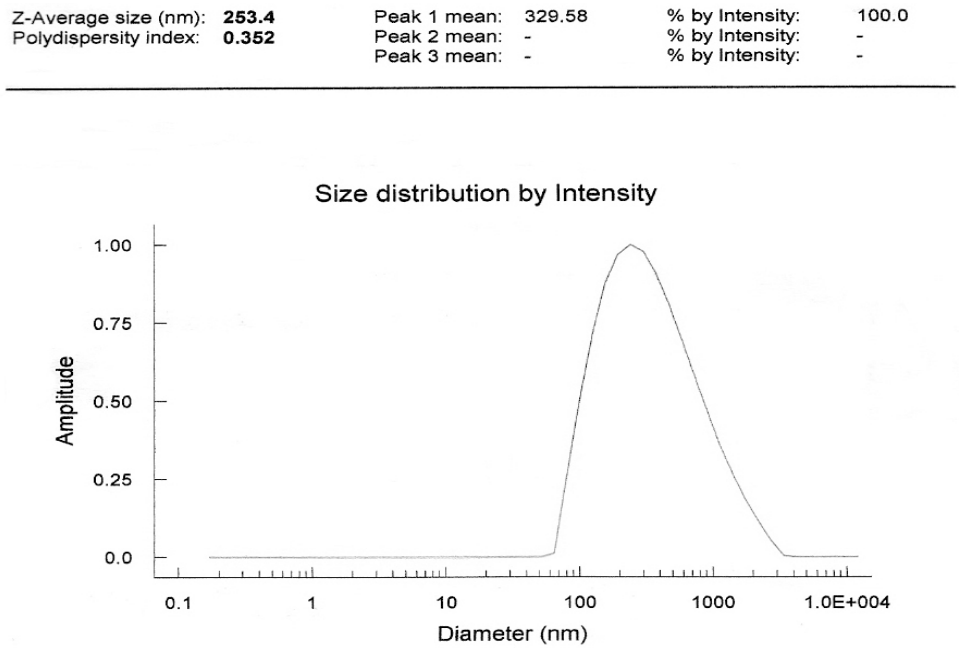


Figure 28 : *Spectrogramme de diffusion de la lumière pour EED-$(His)_6$.*

De plus ces lots purifiés présentent un rapport d'absorption 280 / 260 d'environ 1 (Figure 29) traduisant la présence importante d'acides nucléiques co-purifiés avec la protéine EED-$(His)_6$.

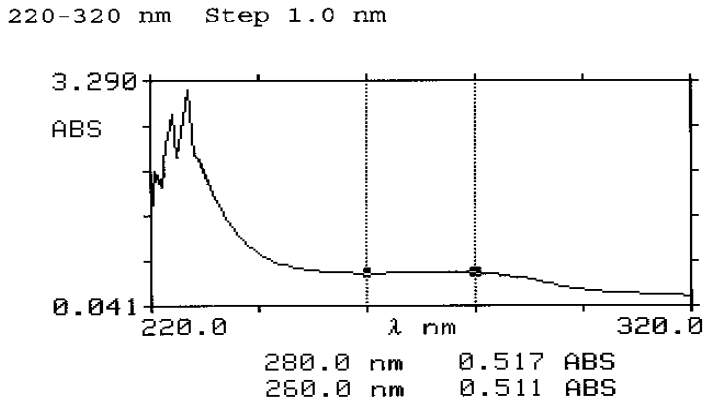


Figure 29 : *Spectrogramme d'absorption de la lumière par EED-$(His)_6$ entre 220 et 320 nm.*

Deux stratégies ont été envisagées afin de remédier à ces deux problèmes de polydispersité et de co-purification d'acides nucléiques :

- travailler sur une nouvelle construction de EED par reclonage

- établir un nouveau protocole de purification de EED-$(His)_6$

Reclonage de EED

Nos études de comparaison de séquences entre la protéine EED et des protéines homologues de structure connue montre que EED devrait présenter deux régions structurellement distinctes :

- une région N-terminale de 84 résidus qui ne serait pas impliquée dans les phénomènes d'interaction avec la Matrice et l'Intégrase (Peytavi *et al.*, 1999 ; Violot *et al.*, 2003).

- une région C-terminale présentant un repliement en turbine à 7 pales de brins β qui serait impliquée dans les phénomènes d'interaction avec MA et IN (Peytavi *et al.*, 1999 ; Violot *et al.*, 2003).

Une protéine EED tronquée, correspondant aux résidus 84 à 441 et conservant une étiquette histidine en C-terminal, a donc été reclonée dans le vecteur d'expression pT7-7 entre les sites Nde1 et Nco1. Après séquençage du nouveau clone, une étude complète de sa surexpression a été conduite par transformation de différentes souches bactériennes de *E.coli* (BL21 (DE3), BL21 (DE3) pLysS, BL21 (DE3) CodonPlus-RP, Origami (DE3), Rosetta (DE3) et C41 (DE3)). Des protocoles d'expression pour chacune de ces souches modulant les paramètres de temps d'induction (3 h ou 12 h) de température d'induction (37°C, 30°C, 23°C) et les concentrations en inducteur ([IPTG] = 0,5 ou 2 μM) ont été réalisées. Dans tous les tests effectués (6 souches x 2 temps x 3 températures x 2 [IPTG] = 72 conditions d'expression) la protéine $EED_{84\text{-}441}$-$(His)_6$ s'est toujours retrouvée dans la fraction insoluble après la lyse des culots bactériens. Ainsi, la purification de la protéine $EED_{84\text{-}441}$-$(His)_6$ en conditions natives n'a pu être réalisée du fait de son insolubilité. De même, des essais de purification en conditions dénaturantes dans l'urée 8 M n'ont pas abouti, car la protéine précipite lors de sa renaturation.

Nouveau protocole de purification de EED-(His)6

La région N-terminale de EED présente un grand nombre de résidus chargés positivement (3 arginines et 11 lysines) qui pourraient favoriser la formation de complexes nucléoprotéiques par des interactions électrostatiques avec les groupes phosphates des acides nucléiques. Sur la base de cette hypothèse, un protocole de purification à un pH plus basique (8,5 au lieu de 7) a été établi afin de diminuer cette interaction. La protéine recombinante entière EED-$(His)_6$ a donc été purifiée en une étape par chromatographie d'affinité sur colonne de nickel (Figure 30) après lyse ménagée et sonication, toutes ces étapes étant réalisées à pH 8,5.

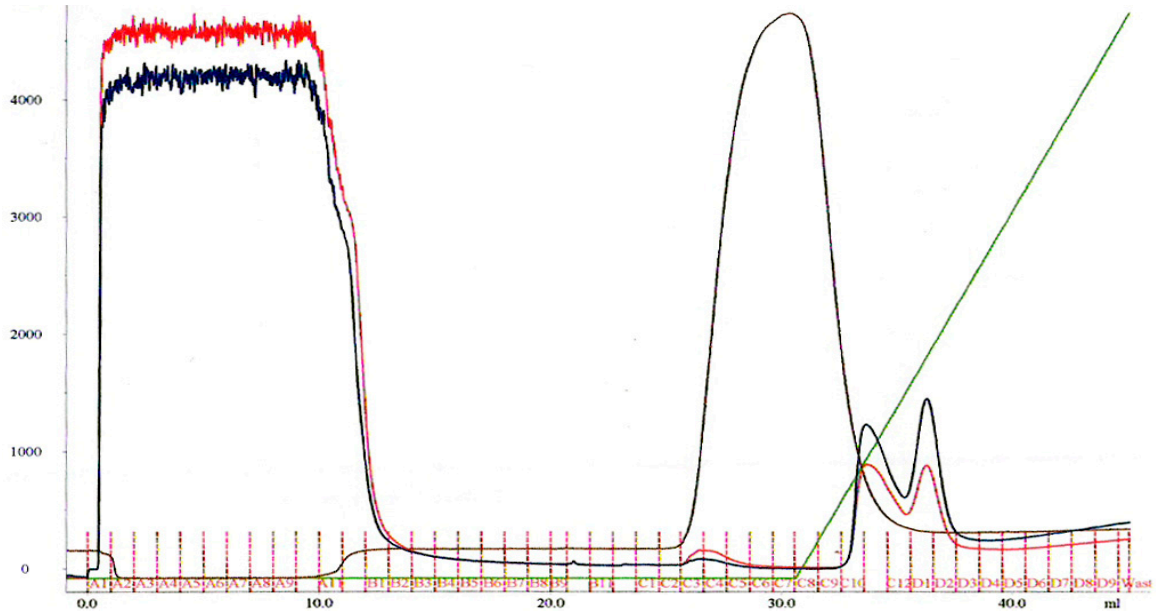


Figure 30 : *Chromatogramme de purification de EED-$(His)_6$ sur colonne de nickel. L'absorption à 280 nm (bleu foncé), l'absorption à 260 nm (rouge), la conductivité (noir) ainsi que le gradient du tampon d'élution (vert foncé) sont représentés.*

La protéine ainsi purifiée migre selon une seule bande en gel d'électrophorèse dénaturant (SDS-PAGE ; Figure 31). La présence de la protéine EED à la bonne taille a été vérifiée par Western blot en utilisant un anti-corps anti-His_6.

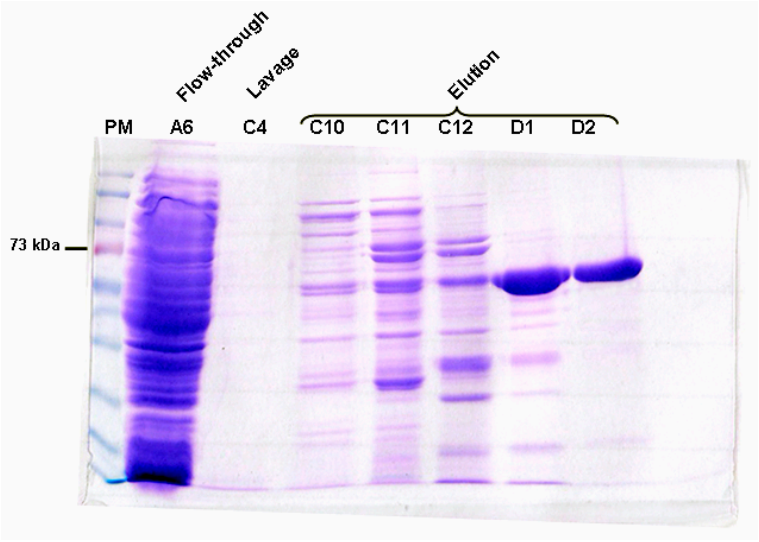


Figure 31 : *SDS-PAGE 10 % des fractions d'élution après chromatographie d'affinité sur colonne de nickel (les fractions déposées correspondent au chromatogramme en figure 30). PM : poids moléculaire.*

L'absorption de la lumière à 260 nm de ces lots purifiés est relativement faible (Figure 32) et ils présentent un rapport d'absorption 280 / 260 de 1,9, traduisant l'absence d'acide nucléique dans ces nouveaux lots purifiés de EED-$(His)_6$.

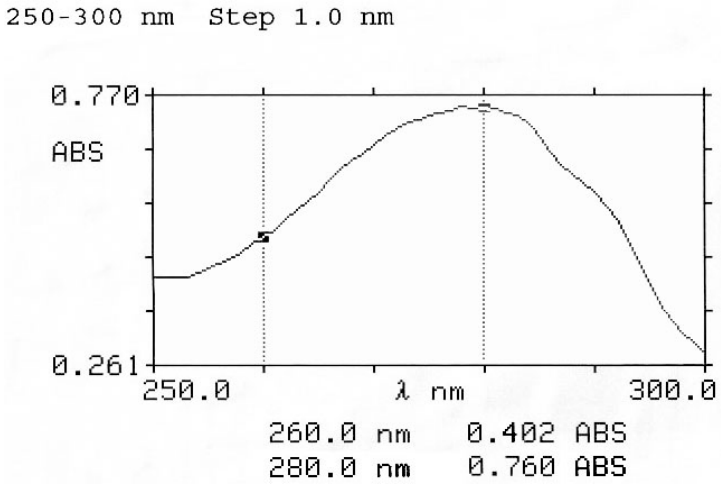


Figure 32 : *Spectrogramme d'absorption de la lumière par EED-$(His)_6$ entre 250 et 300 nm.*

En outre, ces lots de protéines purifiées caractérisés par diffusion de la lumière présentent un coefficient de polydispersité relativement faible (environ 20 %) traduisant la monodispersité de la protéine (Figure 33). Le rayon hydrodynamique déterminé à environ 3,2 nm est compatible avec une protéine de masse d'environ 50 kDa. Ces résultats suggèrent que EED serait sous forme monomérique en solution.

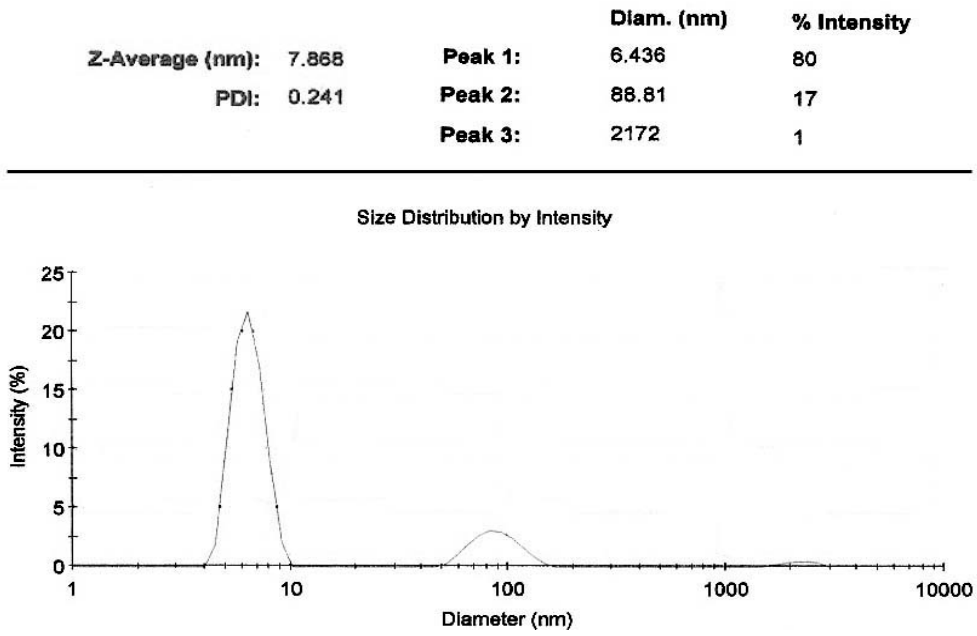


Figure 33 : *Spectrogramme de diffusion de la lumière de EED-(His)$_6$.*

Des études par gel filtration ont été conduites afin de confirmer l'existence d'un monomère de EED (Figure 34) en solution. La protéine est éluée pour un volume de 15,15 mL, soit un Kav de 0,43. Par report sur la droite de calibration de la colonne, ceci donne une masse moléculaire apparente d'environ 51,5 kDa et un rayon hydrodynamique d'environ 3 nm. Ces résultats permettent ainsi de confirmer les études par diffusion de la lumière (donnant un rayon hydrodynamique de 3,2 nm) et confirme l'existence de EED en solution sous forme d'un monomère.

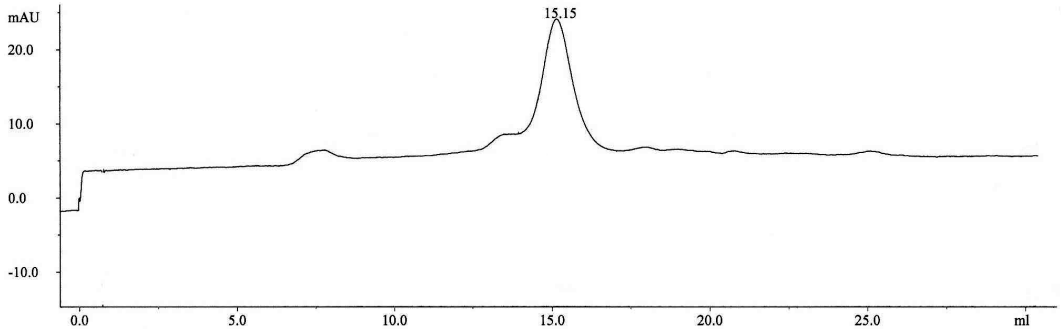


Figure 34 : *Chromatogramme d'élution de EED-(His)$_6$ sur colonne de gel filtration. L'absorption à 280 nm est représentée.*

Ce nouveau protocole de purification permet donc la purification à homogénéité de EED-(His)$_6$ et remplit les conditions nécessaires pour entreprendre des tests de cristallisation.

2. Surexpression de la protéine Matrice dans *E.coli* :

Matériels et méthodes

Les techniques utilisées sont identiques à celles décrites pour la surexpression de EED, à l'exception des points suivants :

Un clone fusionné avec la Glutathion-S-Transférase (GST) en N-terminal de la séquence codante de MA et possédant un site de clivage à la thrombine était disponible au laboratoire Virologie et de Pathogenèse Virale. La surexpression est effectuée dans des bactéries *E. coli* TOP10. Après lyse chimique et sonication, la protéine de fusion est fixée sur une colonne de glutathion puis mise à incuber en présence de thrombine. Après élution par du glutathion réduit (10 mM) la MA est séparée de la thrombine par une deuxième étape de purification sur colonne de benzamidine.

Résultats et discussion

La protéine ainsi purifiée migre selon une seule bande en gel d'électrophorèse dénaturant (SDS-PAGE ; Figure 35).

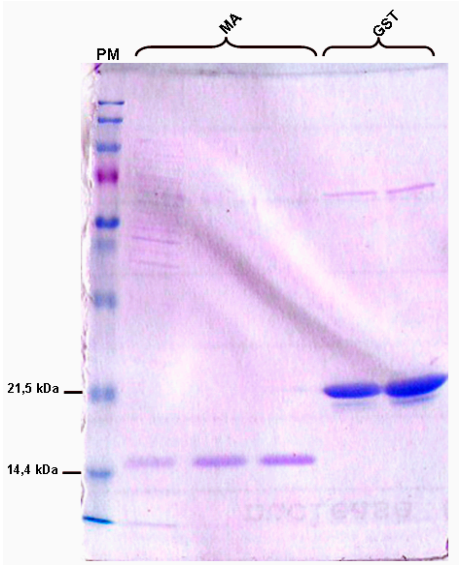


Figure 35 : *SDS PAGE 12 % des fractions d'élution de la protéine de fusion GST-MA après chromatographie d'affinité sur colonne de glutathion, clivage à la thrombine et élution sur colonne de benzamidine. PM : poids moléculaire ; MA : Matrice ; GST : Glutathion-S-Transférase.*

Cependant, les quantités de protéines pures obtenues sont très faibles (environ 100 µg par litre de culture) et ne permettent pas d'envisager des tests de co-cristallisation avec EED à l'heure actuelle.

3. Surexpression de la protéine Intégrase dans *E.coli* :

Matériels et méthodes

Les techniques utilisées sont identiques à celles décrites pour la surexpression de EED, à l'exception des points suivants :
La chromatographie a été réalisée à 15°C sur une colonne HiTrap de 1 mL (Amersham) préalablement chargée en nickel ($NiSO_4$ 0,1 M). Les tampons de fixation (Tris-HCl 25 mM pH 8 ; NaCl 150 mM ; Imidazole 5 mM ; DTT 1 mM) et d'élution (Tris-HCl 25 mM pH 8 ; NaCl 1 M ; Imidazole 1 M ; $ZnSO_4$ 50 µM) sont filtrés et dégazés.

Résultats et discussion

La recherche d'un protocole de surexpression et de purification à homogénéité de l'Intégrase du VIH-1 a été conduite afin de réaliser des expériences de co-cristallisation EED-IN.
Pour ce faire, trois clones de l'Intégrase du VIH-1 étaient disponibles dans le laboratoire de Virologie et de Pathogenèse Virale :

- IN complète (IN^{1-288})
- IN ΔC (IN^{1-202})
- IN ΔN (IN^{52-288})

Ces trois constructions possèdent une étiquette histidine en N-terminal de la séquence codante de l'Intégrase permettant une purification sur colonne de nickel. Il est à noter que seul les clones IN complète et IN ΔN sont intéressants, puisque le clone IN ΔC ne présente plus les propriétés d'interaction avec EED (Violot *et al.*, 2003).

Les protocoles de surexpression en système hétérologue (*E.coli* BL21 (DE3)) et de purification par chromatographie d'affinité sur colonne de nickel mis au point permettent l'obtention de quantités suffisantes de protéine pure soluble pour des études fonctionnelles (environ 300 µg par litre de culture). Ces rendements sont toutefois trop faibles pour des expériences de cristallisation. Ces mauvais rendements sont en partie dus à la très faible solubilité de l'Intégrase sauvage en solution.

Aussi, il a été décidé de travailler avec un mutant hyper-soluble de l'enzyme virale entière (F185K, C280S ; Jenkins *et al.*, 1996). La protéine recombinante $(His)_6$-IN (F185K, C280S) a été purifiée en une étape par chromatographie d'affinité sur colonne de nickel après lyse ménagée et sonication (Figure 36).

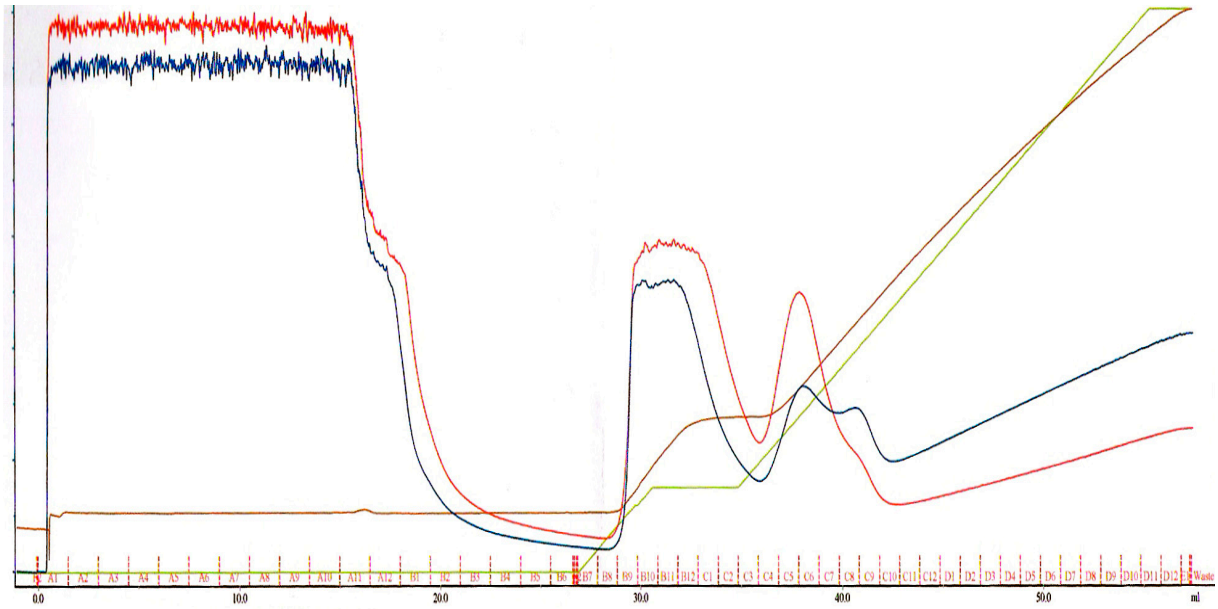


Figure 36 : *Chromatogramme de purification de EED-$(His)_6$ sur colonne de nickel. L'absorption à 280 nm (bleu foncé), l'absorption à 260 nm (rouge), la conductivité (noir) ainsi que le gradient du tampon d'élution (vert foncé) sont représentés.*

La protéine purifiée migre selon une seule bande en gel d'électrophorèse dénaturant (SDS-PAGE ; Figure 37). Sa présence à la taille attendue a été vérifiée par Western blot en utilisant un anti-corps anti-His_6.

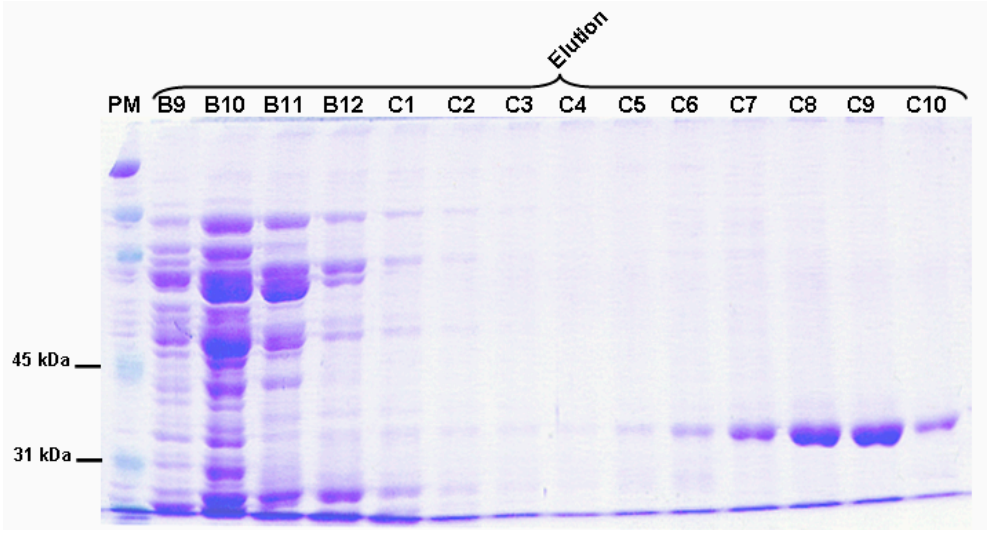


Figure 37 : *SDS PAGE 12 % des fractions d'élution après chromatographie d'affinité sur colonne de nickel (les fractions déposées correspondent au chromatogramme en figure 36). PM : poids moléculaire.*

Les fractions les plus pures ont été rassemblées puis dialysées contre un tampon de stockage (Tris-HCl 25 mM pH 8, NaCl 150 mM, β-mercaptoéthanol 1 mM) avant d'être concentrées à 5 $mg.mL^{-1}$.

4. Surexpression de la protéine Nef dans *E.coli* :

Matériels et méthodes

Les techniques utilisées sont identiques à celles décrites pour la surexpression de EED, à l'exception des points suivants :

La chromatographie d'affinité a été réalisée à 15°C sur une colonne HiTrap de 1 mL (Amersham) préalablement chargée en nickel ($NiSO_4$ 0,1 M). Les tampons de fixation (Tris-HCl 50 mM pH 8 ; NaCl 150 mM ; Imidazole 20 mM) et d'élution (Tris-HCl 50 mM pH 8 ; NaCl 150 mM ; Imidazole 1 M) sont filtrés et dégazés.

Résultats et discussion

Un protocole de surexpression et de purification à homogénéité de Nef a également été élaboré afin de réaliser des expériences de co-cristallisation EED-Nef.

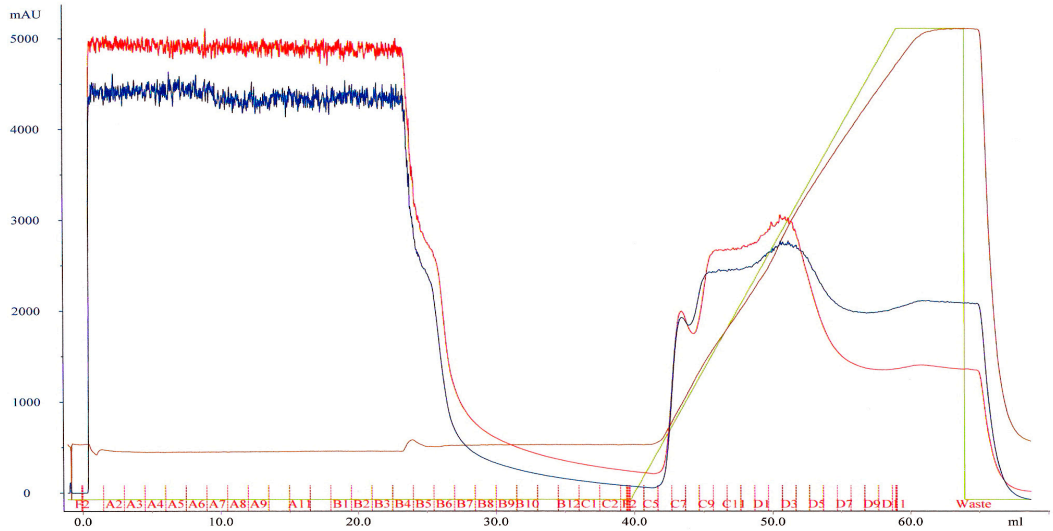


Figure 38 : *Chromatogramme de purification de $(His)_6$-Nef sur colonne de nickel. L'absorption à 280 nm (bleu foncé), l'absorption à 260 nm (rouge), la conductivité (noir) ainsi que le gradient du tampon d'élution (vert foncé) sont représentés.*

La protéine $(His)_6$-Nef est purifiée par chromatographie d'affinité sur colonne de nickel (Figure 38). Elle migre selon une seule bande en gel d'électrophorèse dénaturant (SDS-PAGE ; Figure 39).

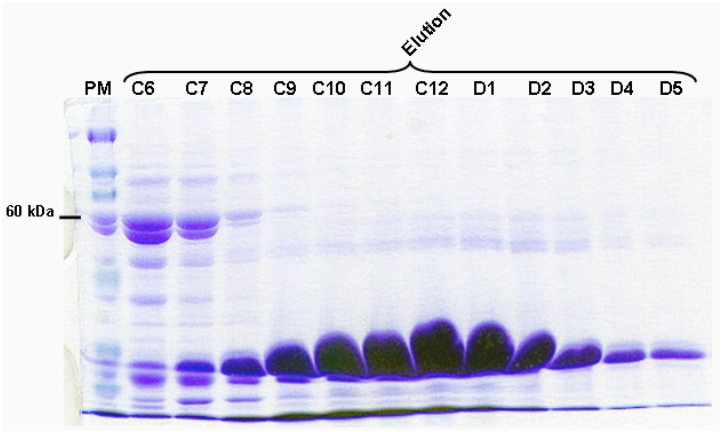


Figure 39 : *SDS-PAGE 12 % des fractions d'élution après chromatographie d'affinité sur colonne de nickel (les fractions déposées correspondent au chromatogramme en figure 38). PM : poids moléculaire.*

Les fractions les plus pures ont été rassemblées puis dialysées contre un tampon de stockage (Tris-HCl 50 mM pH 8,5 ; NaCl 150 mM) avant d'être concentrées à 10 $mg.mL^{-1}$.

5. Cristallisation de EED seule et en complexe :

5.a) Principe de la cristallogenèse et de ses techniques associées :

La cristallogenèse

La cristallisation est le processus qui permet le passage d'une molécule d'un état soluble à un état solide ordonné. Une augmentation progressive de la concentration en protéine (par exemple en diminuant sa solubilité) conduit progressivement la solution à être sursaturée en protéine (Figure 40). Cet état est nécessaire à la nucléation, c'est à dire à la formation de germes cristallins.

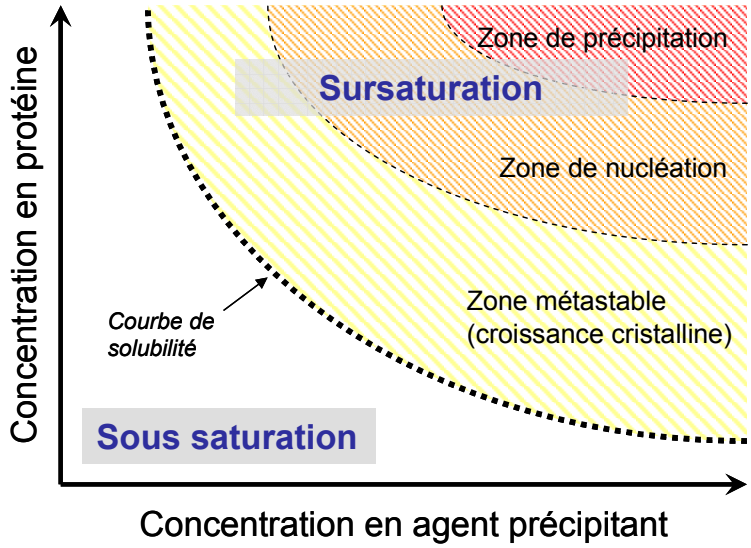


Figure 40 : *Diagramme de solubilité suivant la concentration en protéine et la concentration en agent précipitant. La courbe de solubilité marque une séparation entre une zone où la protéine est en solution (zone de sous saturation) et une zone de sursaturation (zones métastable, de nucléation ou de précipitation).*

Les méthodes les plus couramment utilisées sont la microdialyse et la diffusion en phase vapeur (Wlodaver et Hodgson, 1975). Seule la diffusion en phase vapeur a été utilisée pour cette étude.

La diffusion en phase vapeur – technique de la goutte suspendue

En pratique, nous avons utilisé les techniques de la goutte suspendue et de la goutte assise.
Pour les deux techniques, on considère un système clos composé d'une goutte contenant la protéine en solution et un agent de cristallisation de concentration initiale C. Le second composant du système est un réservoir contenant le même agent de cristallisation dont la concentration C est généralement supérieure d'un facteur 2 à sa concentration C' dans la goutte (Figure 41). La diffusion de vapeur à l'intérieur du système clos provoque un transfert net d'eau (et des autres espèces chimiques volatiles) de la goutte (moins concentrée en agent de cristallisation) vers le réservoir (plus concentré) et ce jusqu'à ce que la concentration en agent de cristallisation soit la même dans les deux compartiments. Ceci induit une augmentation lente de la concentration de la protéine dans la goutte et sa sortie de la zone de solubilité (Figure 41).
L'avantage de cette méthode réside essentiellement dans la faible quantité de matériel protéique nécessaire (de 0,5 à 4 µl de protéine par essai à une concentration usuelle pouvant aller de 5 à 20 $mg.mL^{-1}$).

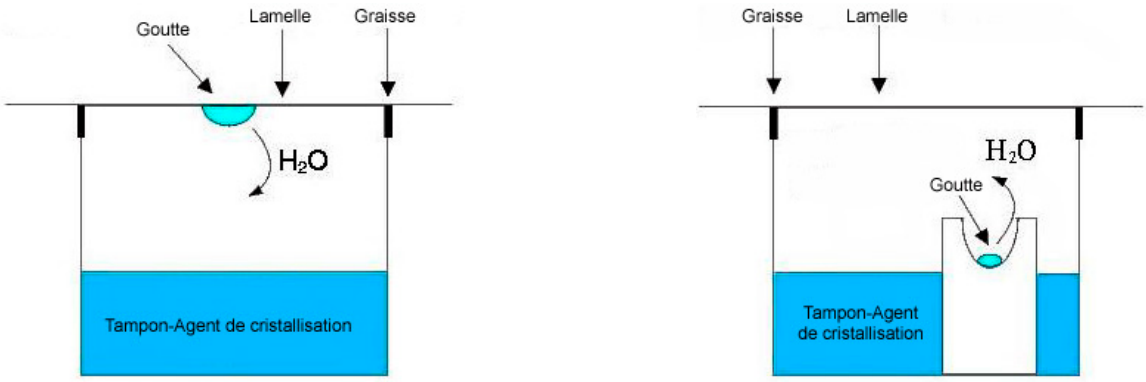


Figure 41 : *Croissance cristalline par les techniques de la goutte suspendue / assise : la goutte (contenant la protéine, un agent de cristallisation, et éventuellement des additifs, ligands, etc.) est suspendue sous une lamelle siliconée fermant hermétiquement, avec un joint de graisse, un puit contenant le même agent de cristallisation que la goutte, à la même concentration.*

Comme les paramètres permettant l'apparition d'un cristal de protéine sont au départ inconnues (nature et concentration de l'agent de cristallisation, concentration de la protéine, pH, température, force ionique, présence d'additifs et / ou de ligands) le but de l'étape de cristallisation est de cribler un maximum de conditions (Ducruix et Giégé, 1992). Puis si l'une ou plusieurs de ces conditions semblent donner un résultat, continuer d'affiner les paramètres

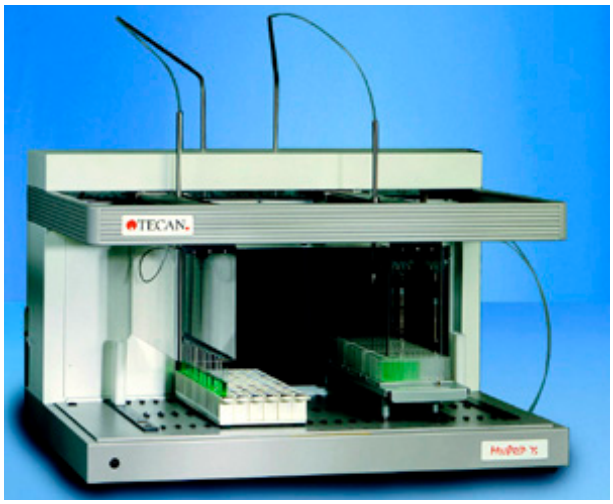
retenus jusqu'à l'obtention de monocristaux de qualité et de taille satisfaisante pour une étude aux rayons X.

Les données de cristallisation de protéines homologues peuvent représenter une piste intéressante pour l'expérimentateur. Ainsi, des banques de données disponibles sur Internet (http://wwwbmcd.nist.gov:8080/bmcd/bmcd.html) classent par famille de protéines les conditions ayant permis d'obtenir des cristaux (Gilliland, 1994) Des cribles sont aussi disponibles, basés sur les 50 conditions les plus utilisées (Jancarik et Kim, 1991). Des kits commerciaux sont proposés par de nombreuses firmes (Molecular Dimension Limited, Nextal, Hampton research).

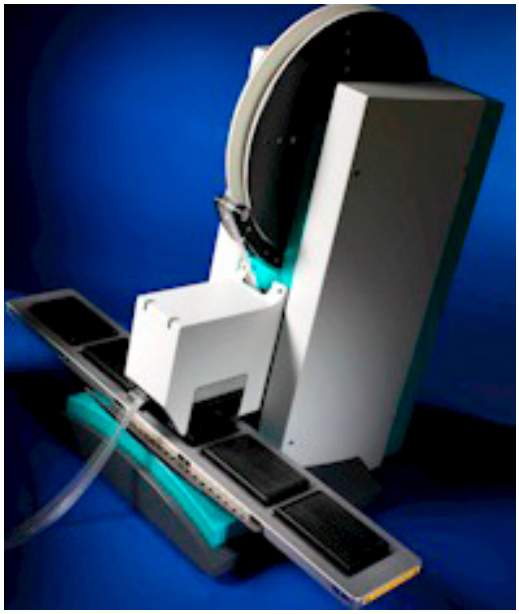
Des techniques ont été développées afin de tester un maximum de conditions de cristallisation avec une consommation minimum de protéine purifiée. Dans le milieu des années 80, l'utilisation de boites de cristallisation à 24 puits a permis des criblages rapides et efficaces des conditions de cristallisation par réalisation de gouttes d'environ 1 µL (volume minimum pour un pipetage manuel). Des robots sont maintenant disponibles, permettant la délivrance de volume de l'ordre du nanolitre.

La plupart des expériences de cristallisation réalisées au cours de cette thèse ont été effectuées de manière robotisée. Deux robots disponibles au laboratoire ont été utilisés :

- un Tecan MiniPrep 75 (Tecan Systems, Inc., USA) pour la re-distribution de conditions de cristallisation commerciales en tubes de 15 mL vers des boites de cristallisation de format 96 puits.



- un Mosquito (TTP Labtech, UK) permettant de délivrer des volumes de l'ordre du nanolitre (entre 50 nL et 1,2 µL) du réservoir vers les puits de cristallisation adjacents, et d'ajouter un volume d'échantillon équivalent à chacun de ces puits.



Trempage des cristaux («crystal soaking»)

Cette technique permet de faire diffuser au sein de l'édifice cristallin, *via* des canaux de solvant, des substances d'intérêt (substrats, analogues de substrat, ligands ou dérivés lourds). Le trempage consiste à immerger un cristal de protéine dans une solution contenant le ligand choisi. La concentration de cette substance et le temps de contact avec le cristal sont déterminés expérimentalement. Le risque majeur de ce type d'expérience est que le cristal s'abîme ou se désagrège sous l'effet de la diffusion de la substance utilisée dans l'édifice cristallin. Dans ce cas, une diminution du temps de contact ou de la concentration du ligand permet généralement d'éviter le problème.

5.b) Cristallisation de EED-$(His)_6$:

Matériels et méthodes

- Préparation du matériel biologique

Les protéines purifiées ont été concentrées par ultrafiltration grâce au système Ultrafree® de chez MILLIPORE. Les solutions sont conservées à 4°C ou à –18°C afin d'éviter leur dénaturation. Les solutions protéiques sont centrifugées avant chaque utilisation pendant 5 min à 3000 g, afin de culotter les éventuels agrégats protéiques ou contaminants, tels des poussières.

- Préparation des solutions tampons

Tous les produits chimiques utilisés pour préparer les solutions sont de grade analytique. Les solutions sont ajustées au pH désiré par ajout de HCl 1 M ou de NaOH 1 M à l'aide d'un pH-mètre (meterLAB™ de chez RADIOMETER). Les solutions sont ensuite filtrées sur des filtres de 0,22 µm ou 0,45 µm afin d'éliminer les poussières. Elles sont conservées à 4°C. Par convention, ces solutions tamponnées additionnées d'un agent de cristallisation sont appelées solution de cristallisation. Elles sont à distinguer des solutions contenant les protéines et appelées solutions protéiques.

- Cristallisation manuelle

Les essais préliminaires de cristallisation de la protéine EED-$(His)_6$ ont été réalisés en utilisant des kits commerciaux de Hampton Research (Crystal Screen Cryo) et Molecular Dimension Limited (Structure Screen 1 et Structure Screen 2). Ce criblage a été conduit en boîte 24 puits selon la méthode de la diffusion de vapeur par la technique de la goutte suspendue. Les gouttes ont un volume de 4 µL (2 µL de solution protéique + 2 µl de solution de cristallisation) et sont mises à équilibrer contre un réservoir de 700 µL de solution de cristallisation à la température constante de 19°C ou de 4°C. Ces essais sont réalisés pour une concentration en protéine de 5 $mg.mL^{-1}$ ou de 10 $mg.mL^{-1}$.

- Cristallisation robotisée

Les essais préliminaires de cristallisation du complexe EED-IN ont été réalisés en utilisant des kits commerciaux de Hampton Research (Crystal Screen 1 et 2 ; Grid Screen Ammonium Sulfate ; Grid Screen Sodium Chloride ; Grid Screen MPD ; Grid Screen PEG 6000 ; Grid

Screen PEG / LiCl ; Natrix ; SaltRX ; Index et PEG / Ion Screen) et Nextal (PEG Suite ; Anions et Cations). Ce criblage a été conduit en boîte 96 puits Greiner Bio-one selon la méthode de la diffusion de vapeur par la technique de la goutte assise. Les réservoirs sont remplis manuellement par 100 µL de solution de cristallisation. Les gouttes (250 nL de complexe + 150 nL de solution de cristallisation) sont réalisées par le robot Mosquito LabTech (MDL) piloté par le logiciel fourni. Les boîtes sont placées à la température constante de 19°C ou de 4°C. Ces essais sont réalisés pour une concentration en protéine de 3 $mg.mL^{-1}$ ou de 10 $mg.mL^{-1}$.

Résultats et discussion

Dans un premier temps, un criblage de différents polyéthylènes glycols (PEG) et de différents sels, tels le sulfate d'ammonium, le chlorure de sodium ou le chlorure de lithium, a été réalisé afin d'analyser les caractéristiques de solubilité de EED.

Ces expériences préliminaires, réalisées à 4°C et à 19°C avec des concentrations en protéine de 5 $mg.mL^{-1}$ ou de 10 $mg.mL^{-1}$ n'ont jamais permis de définir une règle générale sur la solubilité de la protéine EED en fonction des différents agents de cristallisation testés. Effectivement, nous n'avons pas obtenu de résultats reproductibles suivant les lots de protéines préparées.

Nous avons toutefois poursuivi nos essais avec de nouveaux kits commerciaux afin de cribler un maximum de conditions de cristallisation. Des plaquettes (Figure 42) de dimension 0,1 x 0,07 x 0,01 mm^3 sont apparues au bout de 1 mois à 19°C pour la condition suivante : MES 0,1 M pH 6, MPD 40 % (v / v).

Des tests de diffraction aux rayons X ont été réalisés au laboratoire afin d'essayer de les caractériser. Cependant, les clichés ne présentaient pas de tâches de diffraction.

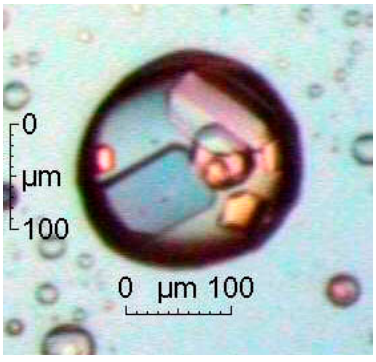


Figure 42 : *Cristaux de EED obtenus en goutte assise avec le robot Mosquito.*

Une plaquette a été dissoute et caractérisée sur un gel d'électrophorèse (Figure 43) afin de vérifier si il s'agissait bien de protéine. Elle a été pêchée puis lavée successivement dans trois solutions de cristallisation dont la concentration en agent de cristallisation est à chaque fois légèrement inférieure à la précédente : ceci permet une légère dissolution du cristal et un décapage de sa surface. Toute trace de molécules de protéines en solution éventuellement adsorbées est ainsi éliminée ; le produit de dissolution des cristaux est donc pur. Cette solution a été déposée sur un gel de polyacrylamide à 12 % en même temps qu'une solution de EED témoin.

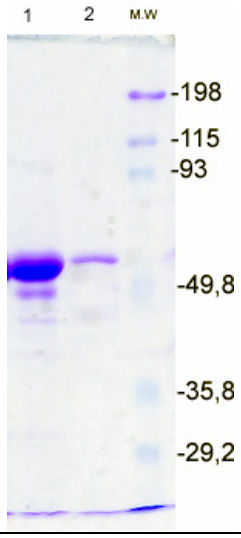


Figure 43 : *SDS-PAGE 12 % d'un cristal de EED dissout. La piste 1 correspond à un dépôt témoin de la protéine EED purifiée. La piste 2 correspond au dépôt de la solution dans laquelle le cristal a été dissout. MW : poids moléculaire.*

La visualisation d'une bande pour la piste 2 au même niveau que le dépôt témoin de la piste 1 permet de conclure à la présence de EED dans le produit de dissolution des cristaux.
En conclusion, les cristaux obtenus semblent être des cristaux de EED mais de taille trop faible pour une étude aux rayons X.
Une optimisation des conditions de cristallisation a été conduite afin d'essayer d'obtenir des cristaux de plus grande dimension en faisant varier le pourcentage en MPD (10 % à 65 %) et le pH (5,2 à 7,2). Ces essais d'optimisation n'ont cependant pas été concluants, car aucun nouveau cristal n'a pu être obtenu.

A ce jour, plus d'un millier de conditions de cristallisation ont été testées, ce qui représente une consommation de 30 mg de protéine purifiée.

Il arrive que les protéines qui ne cristallisent pas dans leur état natif puissent cristalliser en complexe. En effet, les changements conformationnels résultant de la formation de ces complexes peuvent être favorable à la cristallisation par l'établissement de nouveaux contacts cristallins ou par la stabilisation de la protéine.

Ainsi, des essais de co-cristallisation de EED avec ses partenaires IN et Nef ont été conduits.

5.c) Cristallisation du complexe EED-IN :

Matériels et méthodes

Les essais préliminaires de cristallisation du complexe EED-IN ont été réalisés en utilisant des kits commerciaux de Hampton Research (Crystal Screen 1 et 2) et Nextal (PEG Suite). Ce criblage a été conduit en boîte 96 puits Greiner Bio-one selon la méthode de la diffusion de vapeur par la technique de la goutte assise. Les réservoirs sont remplis manuellement par 100 µL de solution de cristallisation. Les gouttes (250nL de complexe + 150nL de solution de cristallisation) sont réalisées par le robot Mosquito LabTech (MDL) piloté par le logiciel fourni. Les boîtes sont placées à la température constante de 19°C ou de 4°C.
Les complexes sont réalisés pour un rapport molaire EED:IN de 1:2 à partir d'un stock de protéine EED concentrée à 5 $mg.mL^{-1}$, soit environ 100 µM, et d'un stock de protéine Intégrase concentrée à 5 $mg.mL^{-1}$, soit environ 160 µM.

Résultats et discussion

L'Intégrase semble se comporter en solution comme un dimère (cf. Etude bibliographique) alors que nous avons montré que EED devrait être sous forme monomérique. L'hypothèse d'une interaction mole à mole entre les deux partenaires nous a conduit à réaliser dans un premier temps des études de cristallisation pour un complexe EED-IN dans un rapport molaire de 1:2.

La purification de l'Intégrase, bien qu'il s'agisse du mutant soluble F185K, C280S, pose le problème d'un rendement faible (environ 0,5 mg par litre de culture). Ces quantités ne permettent pas un criblage exhaustif des conditions de cristallisation.

Dans un premier temps, les kits commerciaux Crystal Screen 1 et 2 de chez Hampton Research qui sont basés sur les travaux de Jancarik et Kim (Jancarik et Kim, 1991) ont été utilisés.

Dans un deuxième temps, le kit commercial PEG Suite de chez Nextal a été utilisé. En effet, le PEG (polyéthylène glycol) semble induire des interactions attractives entre les macromolécules de haut poids moléculaire. Pour des protéines de grande taille (ATCase : 306 kDa, Apoferritine : 443 kDa), la présence d'un polymère non absorbant (PEG) est indispensable pour augmenter les interactions attractives et induire la cristallisation, un sel seul n'étant pas efficace (Finet *et al.*, 2003 ; Tardieu *et al.*, 2002 ; Vivares *et al.*, 2002). Ce résultat est observé pour 71 % des complexes protéine-protéine dont les conditions de cristallisation reportées dans la littérature sont composées de PEG plutôt que de sulfate d'ammonium ou d'autres solutions salines de haute force ionique (Radaev et Sun, 2002).

Cependant, tous ces essais n'ont donné que des gouttes limpides.

Ainsi, les quelques 200 conditions testées jusqu'à maintenant n'ont pas permis d'obtenir des cristaux du complexe EED-IN.

5.d) Cristallisation du complexe EED-Nef :

Matériels et méthodes

Les techniques utilisées sont identiques à celles décrites pour la cristallisation du complexe EED-IN, à l'exception des points suivants :
Les complexes sont réalisés pour un rapport EED:Nef de 1:2 à partir d'un stock de protéine EED concentrée à 5 $mg.mL^{-1}$, soit environ 100 µM, et d'un stock de protéine Nef concentrée à 10 $mg.mL^{-1}$, soit environ 900 µM.

Résultats et discussion

Le complexe EED-Nef a été réalisé pour un rapport molaire de 1:2. En effet, Nef semble également se comporter en solution comme un dimère (cf. Etude bibliographique).

Comme pour le complexe EED-IN, la majorité des essais de cristallisation du complexe EED-Nef ont été réalisés dans un premier temps grâce à l'utilisation du kit commercial PEG Suite de chez Nextal.

Un petit cristal est apparu à 19°C au bout de 2 semaines pour les conditions suivantes : HEPES 0,1 M pH 7,5 ; PEG 6000 25 % (p / v). Ce cristal présente une taille d'environ 0,075 x 0,075 x 0,075 mm^3 (Figure 44).

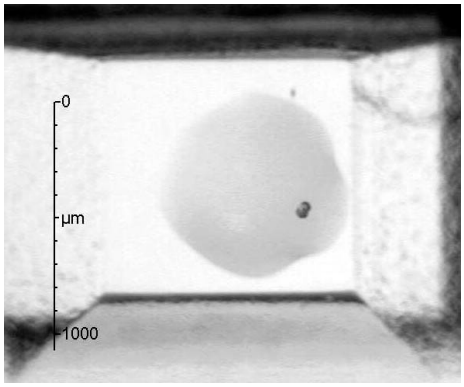


Figure 44 : *Cristal du complexe EED-Nef obtenu en goutte assise avec le robot Mosquito.*

Afin d'optimiser ces conditions de cristallisation et d'obtenir des cristaux de plus grande dimension, des expériences faisant varier le pourcentage en PEG 6000 (15 % à 40 %) et le pH (7,0 à 8,1) ont été conduites à 4°C et à 19°C (Tableau 5).

Toutes les gouttes sont limpides à 4°C. Des précipités sont observés à 19°C pour les gouttes de pH inférieur à 7,5. Des précipités se forment dans les gouttes de pH supérieur à 7,5 présentant une concentration en PEG 6000 dans le réservoir supérieur à 30 %. Ces expériences recensées dans le tableau 5 suggèrent que le complexe pourrait être cristallisé pour des concentrations en PEG 6000 comprise entre 25 et 27,5 % à pH basique.

	0,1 M Hepes	pH 7,0	pH 7,1	pH 7,2	pH 7,3	pH 7,4	pH 7,5	pH 7,6	pH 7,7	pH 7,8	pH 7,9	pH 8,0	pH 8,1
PEG 6000		1	2	3	4	5	6	7	8	9	10	11	12
15%	A												
20%	B												
22,5%	C												
25%	D												
27,5%	E												
30%	F												
35%	G												
40%	H												

Tableau 5: *Matrice d'optimisation des conditions de cristallisation du complexe EED-Nef. Les cellules grisées indiquent la présence de précipités et les cellules blanches la présence de gouttes limpides.*

Ces essais d'optimisation n'ont cependant pas été concluants car aucun nouveau cristal n'a été obtenu.

Un essai d'ensemencement macroscopique a également été réalisé. Trois lavages successifs du petit cristal précédemment obtenu ont été réalisés pendant 5 min dans une goutte pré-équilibrée (HEPES 0,1 M pH 7,5 ; PEG 6000 20 % (p / v)). Le cristal a ensuite été ensemencé dans une solution pré-équilibrée (HEPES 0,1 M pH 7,5 ; PEG 6000 25 % (p / v)). Il s'est malheureusement dissous au bout de quelques jours. Nous n'avons donc pas pu vérifier par des expériences de dissolution et de caractérisation sur gel d'électrophorèse que ce cristal correspondait bien au complexe EED-Nef.

Nous avons enfin utilisé les kits commerciaux Crystal Screen 1 et 2 de chez Hampton Research sans succès.

500 conditions de cristallisation ont été testées pour ces essais de cristallisation de EED en complexe avec Nef, ce qui représente une consommation de 7,5 à 10 mg de chacune des deux protéines purifiées.

C/ CONSTRUCTION D'UN MODELE MOLECULAIRE DE EED ET VALIDATION PAR DES ETUDES DE «PHAGE-DISPLAY»

En parallèle à ces essais de cristallisation de EED, il nous est paru utile de réaliser un modèle moléculaire afin de corroborer les résultats obtenus par «phage-display» sur les zones d'interaction de la protéine EED avec ses partenaires viraux (Peytavi *et al.*, 1999 ; Violot *et al.*, 2003).

Matériels et méthodes

- *Modélisation par homologie de EED*

Le choix de l'empreinte s'est fait par une recherche d'homologues structuraux de EED, après alignement de sa séquence contre les séquences des protéines de structure connue répertoriées dans la Protein Data Bank (PDB ; Berman *et al.*, 2000) grâce au logiciel CLUSTALW (Thompson *et al.*, 1994). Les coordonnées de la protéine Gβ dont la séquence présente des motifs répétés WD-40 ont été obtenues (code PDB 1TBG).

L'alignement des séquences de EED avec Gβ a été réalisé en utilisant le programme CLUSTALW. Cet alignement initial a ensuite été optimisé manuellement afin d'améliorer la correspondance des motifs répétés WD-40. Pour cela, les éléments de structure secondaire de EED ont été prédits grâce aux logiciels MLRC (Guermeur *et al.*, 1999) DSC (King et Sternberg, 1996) et PHD (Rost et Sander, 1993) accessibles sur le serveur NPS@ (Combet *et al.*, 2000).

La construction du modèle de EED à partir de Gβ a été obtenue par substitution des chaînes latérales par le programme CALPHA (Esnouf, 1997). La réorientation manuelle de ces chaînes ainsi que la construction des fragments d'insertions ont été réalisées avec le logiciel TURBO-FRODO (Roussel et Cambillau, 1989).

L'optimisation du modèle a été réalisée par minimisation d'énergie selon la méthode du gradient conjugué par le programme CNS (Brunger *et al.*, 1998).

- *«Phage-displa»y et «phage-biopanning»*

Deux librairies de phages filamenteux ont été utilisées, une consistuée d'hexapeptides et une de dodécapeptides. Les protéines purifiées (IN ou EED) ont été immobilisées sur des plaques. L'élution des phages a été obtenue par compétition avec la protéine partenaire purifiée. Les séquences d'interaction (phagotopes) ont été identifiés par séquençage de la protéine

recombinante fUSE5 pIII par la méthode des didésoxynucléotides, des amorces *ad hoc* et le kit de séquençage Sequenase kit version 2.0 (Amersham).

Résultats et discussion

La protéine EED appartient à la superfamille des protéines à motifs répétés WD-40 (Neer *et al.*, 1994). Un modèle de EED a été construit grâce à la structure de la sous-unité β de la protéine-G bovine (Lambright *et al.*, 1996 ; Sondek *et al.*, 1996) qui présente une identité de séquence de 20 % avec EED (Figure 45).

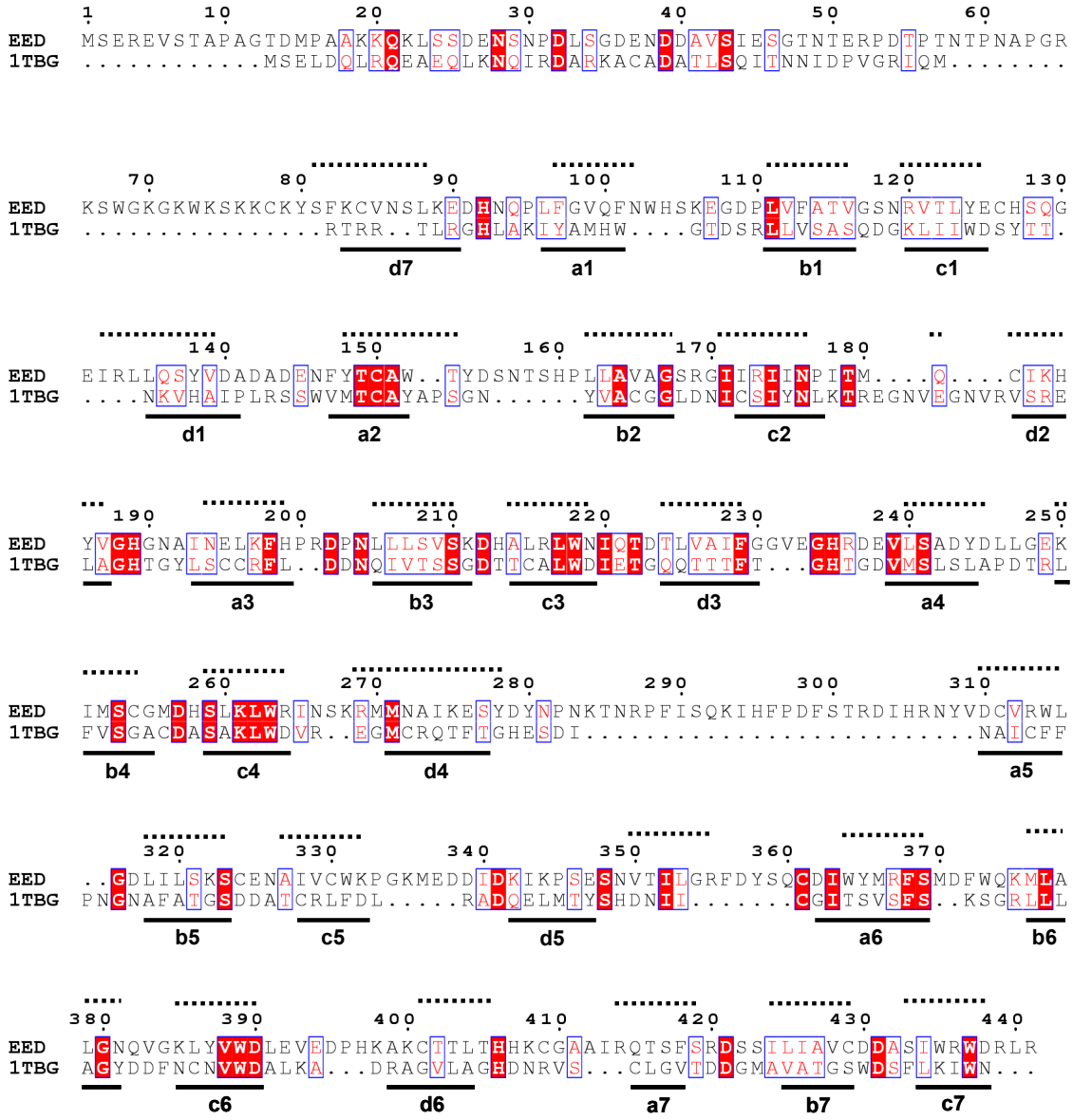


Figure 45 : *Alignement des structures primaires de EED et de la sous-unité β de la protéine-G bovine (PDB : 1TBG). La localisation des brins β chez 1TBG est décrite (soulignée) ainsi que la prédiction des brins β chez EED (pointillés). L'alignement a été optimisé afin de faire correspondre les régions prédites et structurées en brins β.*

L'alignement des séquences a été optimisé en tenant compte de la correspondance entre les résidus conservés des motifs WD-40 (Figure 46).

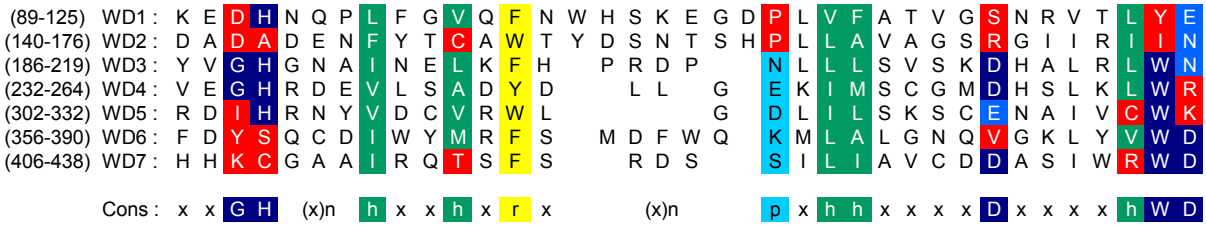


Figure 46: *Les 7 motifs répétés WD-40 de la protéine EED alignés avec la séquence consensus décrite par Neer (Neer et Smith, 1996) ; h, résidu hydrophobe ; r, résidu aromatique ; p, résidu polaire.*

Le modèle (Figure 47 A) donne une idée du repliement global du domaine C-terminal de la protéine (résidus 84 à 441). Chaque motif répété WD-40 se replie en fait selon 3 brins β antiparallèles a, b et c. La portion de séquence reliant chaque motif répété se structure également en un brin β d additionnel. Un motif WD-40 forme une unité structurale composée de 4 brins β antiparallèles appelée pale β (Figure 47 B). Les pales se disposent pour former une structure en turbine-β.

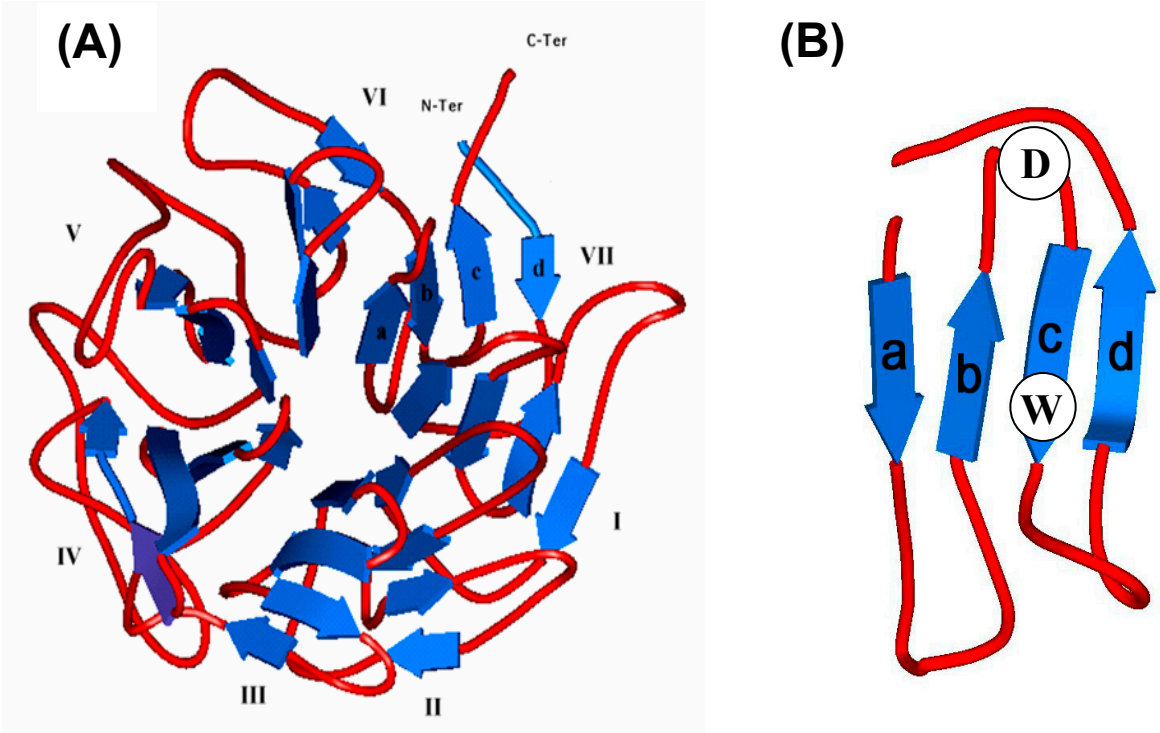


Figure 47 : ***(A)*** *Structure modélisée de EED (résidus 84 à 441) montrant les 7 pales β (I à VII) chacune constituée de 4 brins β (a à d).* ***(B)*** *Structuration de la pale β I. Les résidus Tryptophane et Aspartate conservés du motif WD-40 sont indiqués (cercles blancs).*

Au cours des études d'interaction de EED avec l'Intégrase (Publication 1), il a été montré que cette interaction nécessitait l'intégrité des sept motifs répétés WD-40. En effet, la délétion des deux motifs WD-40 C-terminaux entraînait l'abolition de cette interaction. De manière similaire, des délétions du dernier motif WD-40 C-terminal du répresseur de transcription Tup1 (Williams et Trumbly, 1990) ou du facteur d'épissage Prp4 de *Saccharomyces cerivisiae* (Hu *et al.*, 1994) sont délétères. La perte d'interaction suite à la délétion des ces motifs répétés WD-40 C-terminaux peut s'expliquer par l'incapacité de ces protéines à se structurer correctement. La fermeture de la turbine-β implique en effet une interaction entre le premier et le dernier motif répété WD-40. En outre, la délétion d'un motif répété WD-40 interne abolit les fonctions de Prp4 (Hu *et al.*, 1994). Ainsi, l'intégrité des motifs WD-40 est nécessaire à la fonction de ces protéines.

Les surfaces accessibles de structure en turbine-β se réduisent ainsi aux brins β d externes, ainsi qu'aux boucles reliant chaque brin β d'une même pale. Ces zones accessibles sont d'ailleurs les zones d'interaction, mises en évidence par cristallisation dans les structures de complexes d'autres protéines en turbine-β (Chen *et al.*, 1992 ; Wall *et al.*, 1995). Par exemple, dans le cas de la sous-unité β de la protéine-G bovine, plusieurs résidus appartenant aux boucles d-a et b-c sont impliqués dans des contacts hydrophobes avec la sous-unité α (Wall *et al.*, 1995).

Cependant, compte tenu de la faible homologie de séquence de EED avec des protéines de structure connue, il existe pour ce modèle de EED une grande incertitude sur la conformation de chacune des boucles. Nous avons vérifié expérimentalement les résidus impliqués dans ces boucles accessibles d'EED par analyse immunologique des motifs peptidiques représentant les épitopes immunodominants de la molécule d'EED grâce à deux antisérums polyclonal et monoclonal anti-EED. Les premiers résultats de l'analyse des épitopes reconnus par ces différentes classes d'anticorps ont été effectués en utilisant la technique du «phage-display» (Hong et Boulanger, 1995 ; Table 6). Ils permettent de confirmer que les boucles décrites par le modèle sont bien des régions accessibles de EED (Figure 48).

Classe d'anticorps	Séquence phagotopique	Motif homologue chez EED
Monoclonal	WWGMSY	98-GVQF-101
	ARLVYL	132-IRLL-135
	ARLFYS	132-IRL-134
	LWNVLH	217-LWNIQT-222
	HMPHMK	295-HFPDFS-300
	MLRYDH	352-ILGRFDY-358
	DFMSML	371-DFWQKML-377
Polyclonal	GVSFIN	98-GVQFNW-103
	CLFYWH	101-FNWH-104
	MNWSSS	101-FNWHSK-106
	SLGEIS	128-SQGEIR-133
	DVLAAF	223-DTLVAI-228
	KRLQNT	268-KRMMNA-273
	SHFPLL	294-IHFPDF-299
	VPTHIH	299-FSTRDIH-305
	WLGEIS	314-WLGDLI-319
	PVGRGY	382-QVGKLY-387
	LFRWSR	434IWRWDR-439

Table 6 : *Mimotopes de EED dans les phages élués par les anticorps anti-EED monoclonal et polyclonal.*

La plupart des régions mise en évidence correspondent à des séquences modélisées comme étant des boucles chez EED. Trois séquences modélisées comme des brins β sont également mise en évidence. Si le brin β IIId (Figure 48) semble bien être accessible dans le modèle, les brins β Ia et VIIc semblent être moins accessibles. Une telle accessibilité pour le brin VIIc pourrait s'expliquer par un défaut de fermeture de la turbine-β (cf. Etude bibliographique).
Le cas du brin β Ia est très intéressant, car ce dernier a été décrit comme étant un motif d'interaction potentiel avec l'Intégrase (Violot *et al.*, 2003). Il semble donc que cette région soit en fait bien plus accessible que ne le montre le modèle, puisqu'elle représente une région immunodominante tant pour l'anticorps monoclonal que pour l'anticorps polyclonal.

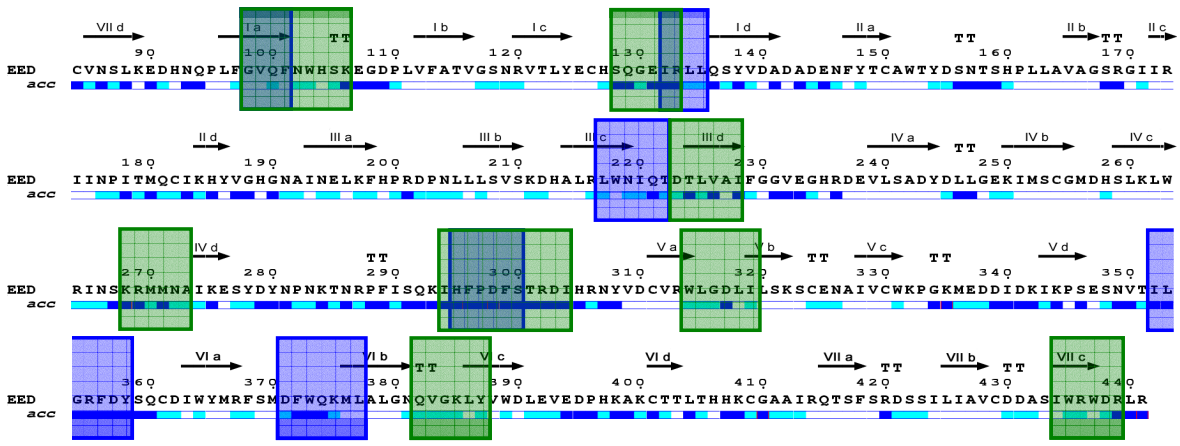


Figure 48 : *Structures primaire et secondaire de EED. Les motifs homologues chez EED aux phagotopes mis en évidence par élution de phages sur l'anticorps anti-EED monoclonal (en bleu) et sur l'anticorps anti-EED polyclonal (en vert) sont montrés. L'accessibilité (acc) des résidus chez EED est également représenté (bleu foncé : très accessible ; bleu clair : accessible ; blanc : enfoui).*

Certaines de ces régions immunodominantes précédemment décrites se superposent exactement aux régions putatives d'interaction entre EED et ses partenaires viraux (Figure 49). C'est en effet le cas pour les régions d'interaction putative avec l'Intégrase (brin β Ia et brin β IIId) qui sont également des régions immunodominantes, aussi bien pour l'anticorps monoclonal que pour l'anticorps polyclonal. C'est également le cas pour la région d'interaction putative avec la Matrice (boucle IVd-Va) qui est aussi une région immudominante pour les deux types d'anticorps.

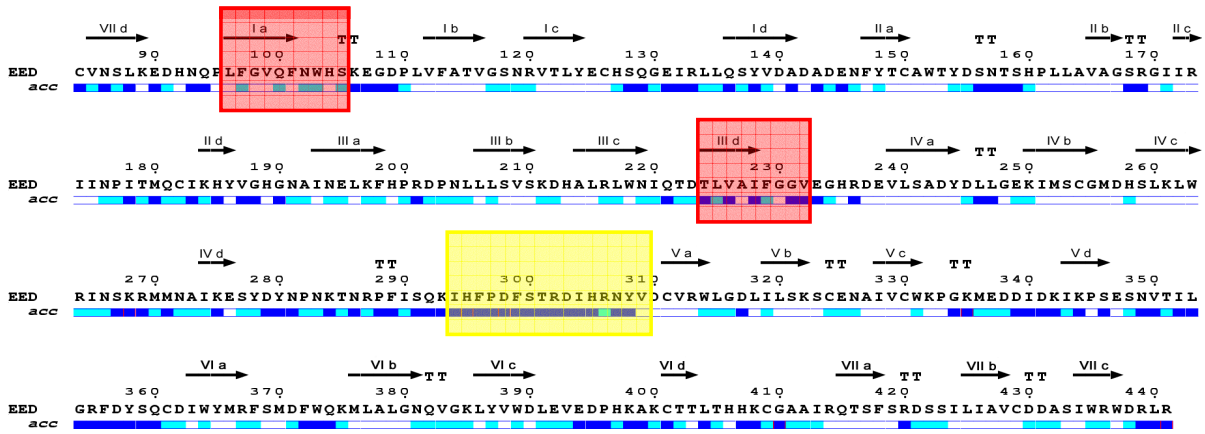


Figure 49 : *Structures primaire et secondaire de EED. Les zones d'interaction potentielles avec MA (jaune) et IN (rouge) sont montrées. L'accessibilité (acc) des résidus chez EED est également représenté (bleu foncé : très accessible ; bleu clair : accessible ; blanc : enfoui).*

Ainsi, malgré les incertitudes d'interprétation dues à la faible similarité de séquence entre EED et la sous-unité β de la protéine G bovine, ce modèle reste une bonne base pour caractériser les zones d'interaction entre EED et ses partenaires viraux MA et IN. Nos expériences de «phage-display» (Peytavi *et al.*, 1999 ; Violot *et al.*, 2003) permettent de proposer le modèle structural présenté sur la figure 50.

Figure 50 : *Modèle structural de EED présentant les zones d'interaction potentielles avec MA (jaune) et IN (rouge).*

Ce modèle montre que la zone d'interaction avec la Matrice se situerait sur la face supérieure de la turbine-β, dans une boucle d-a très accessible (boucle IIId-IVa). Les zones d'interaction avec l'Intégrase se situeraient pour la région 96-105 en partie dans une boucle a-b (boucle Ia-Ib) située sur la face inférieure de la turbine-β, et pour la région 224-232 dans le brin β IIId situé sur la circonférence de la turbine-β.

Les trois régions d'interaction mises en évidence avec la Matrice et l'Intégrase sont donc structurellement distinctes. Ceci implique que EED pourrait avoir la capacité d'interagir en même temps avec la Matrice et avec l'Intégrase.

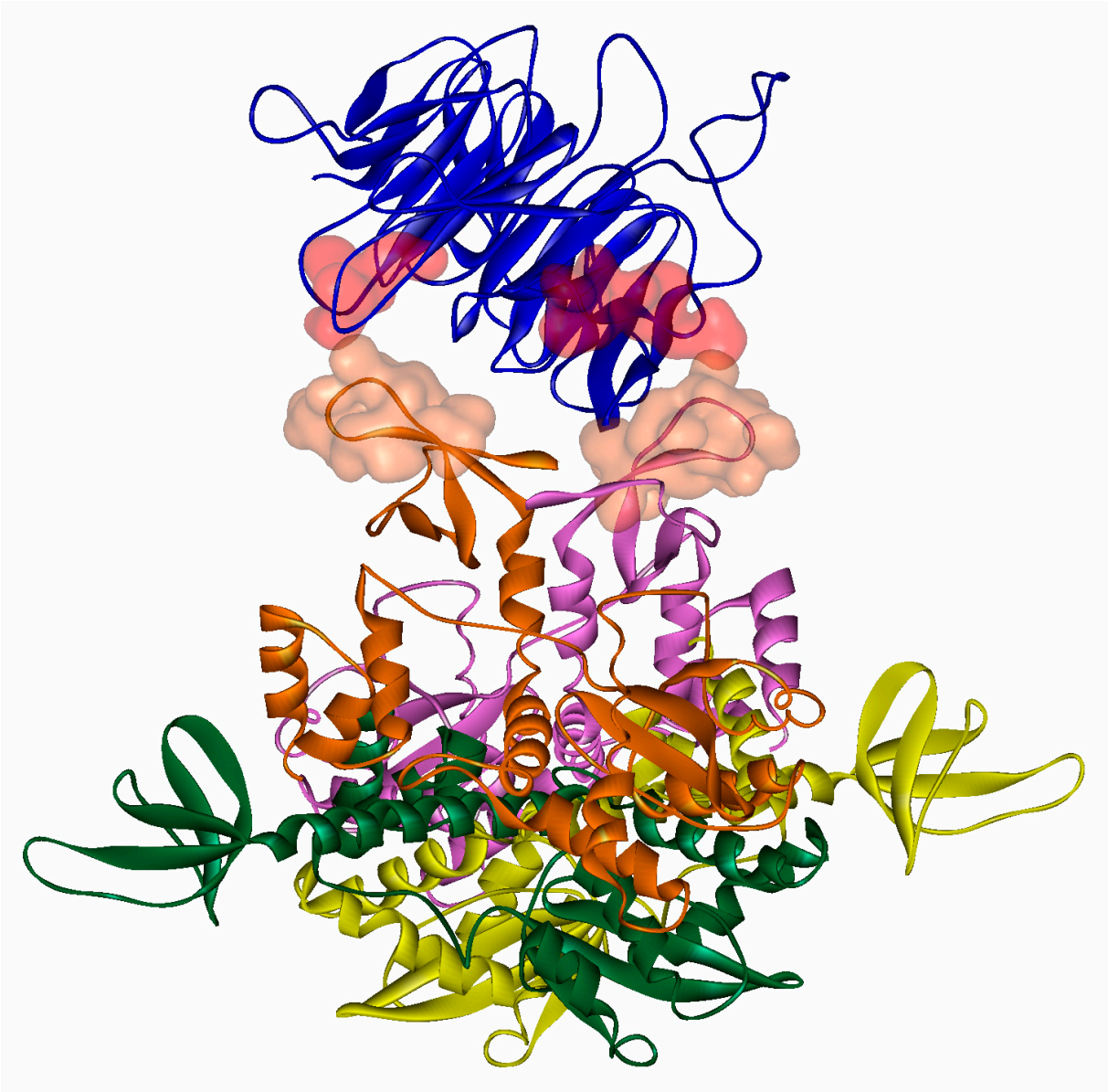


Figure 51 : *Modèle structural du complexe EED-IN présentant les deux zones d'interactions potentielles chez EED (rouge) et les zones d'interaction sur deux monomères de IN (orange). Pour le modèle tétramérique de IN, chacun des monomères est représenté en une couleur différente (vert, orange, rose et jaune).*

La mise en évidence de deux zones d'interaction avec l'Intégrase pourrait également indiquer que cette dernière se retrouve sous une forme oligomérique lors de son interaction avec EED. En effet, chacune de ces deux zones chez EED devrait être en contact avec une zone mise en évidence chez IN et se situant dans la région 224-264 (Violot *et al.*, 2003). Or la modélisation du complexe putatif EED-IN montre que seule la forme tétramérique[5] de IN peut permettre

[5] Le modèle tétramérique de l'Intégrase a été obtenu sur la base de contacts cristallographiques par la superposition des domaines centraux communs aux structures des domaines N-terminal et central (Wang, J.Y., Ling, H., Yang, W. and Craigie, R. (2001) Structure of a two-domain fragment of HIV-1 integrase: implications for domain organization in the intact protein. *Embo J*, **20**, 7333-7343. central et

l'établissement de contacts entre deux monomères d'IN et les deux zones d'interaction chez EED (Figure 51). Dans le modèle tétramérique de IN, les domaines C-terminaux, au sein desquels la zone putative d'interaction avec EED a été mise en évidence sont distants de 35 Å pour les deux monomères les plus proches (monomères orange et rose sur la figure 51). Les autres distances entre domaines C-terminaux sont au moins deux fois supérieures. Les deux régions d'interaction chez EED sont également distantes de 35 Å.

Selon cette hypothèse, EED pourrait par ce type d'interaction avec IN faciliter l'oligomérisation de cette dernière sous forme de tétramère et favoriser de ce fait le processus d'intégration. Ceci permettrait de corroborer nos résultats obtenus par le test d'intégration *in vitro* qui montrent que EED devrait avoir un effet activateur dose-dépendant sur la réaction d'intégration de l'ADN viral réalisée par l'Intégrase.

terminal Chen, J.C., Krucinski, J., Miercke, L.J., Finer-Moore, J.S., Tang, A.H., Leavitt, A.D. and Stroud, R.M. (2000) Crystal structure of the HIV-1 integrase catalytic core and C-terminal domains: a model for viral DNA binding. *Proc Natl Acad Sci U S A*, **97**, 8233-8238). Ce modèle constitué de deux dimères est plausible car les différents domaines ne s'entrecoupent pas.

Conclusion et perspectives sur EED et ses partenaires viraux

La protéine humaine EED semble jouer un rôle important dans le cycle du virus VIH-1, en se liant avec les protéines virales MA (Peytavi *et al.*, 1999) IN (Peytavi, 1999) et Nef (Witte *et al.*, 2004). EED appartient à la super-famille des protéines à motifs répétés WD-40, qui interviennent dans la structuration de la chromatine et l'extinction des gènes. Cette protéine pourrait agir sur le pouvoir d'infection du virus, en participant à son processus d'intégration tout en agissant sur l'activité des gênes cellulaires.

Le travail présenté dans ce manuscrit a permis d'obtenir des données importantes à propos de la protéine EED, tant d'un point de vue fonctionnel que structural.
Tout d'abord, d'un point de vu fonctionnel, l'interaction de EED avec l'Intégrase au cours des phases précoces du cycle viral a pu être démontrée et a fait l'objet d'une première publication (Violot *et al.*, 2003). Les principaux résultats sont ici brièvement résumés.
In vitro, les résultats obtenus par mutagenèse, essais «pull-down» et «phage-biopanning», montrent que l'interaction EED-IN nécessite l'intégrité des deux motifs répétés WD-40 C-terminaux de EED. De plus, il a été montré que EED semblait avoir *in vitro* un effet activateur dose-dépendant sur la réaction d'intégration d'ADN réalisée par l'Intégrase.
Des études réalisées par immuno-électromicroscopie de la distribution cellulaire de l'Intégrase et de EED dans des cellules infectées par le VIH-1 ont montré *in situ* une co-localisation de EED et IN dans le noyau à proximité des pores nucléaires. Une triple co-localisation EED, IN et MA a également pu être observée dans le nucléoplasme. Ceci suggère la présence de complexes multi protéiques impliquant ces trois protéines au stade précoce du cycle viral.

Ensuite, d'un point de vue structural, quatre protocoles de purification en vue d'études cristallographiques de EED et de ses complexes avec ses partenaires viraux Matrice, Intégrase et Nef ont été mis au point. Des essais de cristallisation sont en cours. Les essais les plus prometteurs concernent le complexe EED-Nef, qui revêt un intérêt particulier puisque sa résolution permettrait d'obtenir la première structure cristallographique de la protéine Nef entière. L'implication de Nef VIH-1 dans le cycle viral et son rôle important en font actuellement une cible thérapeutique de choix.

EED présente au niveau de sa séquence une signature de 7 motifs répétés WD-40 et devrait posséder un repliement en turbine de brins- β homologue à celui de la sous unité β de la protéine G (Sondek *et al.*, 1996). La structure de cette sous unité a été utilisée comme empreinte afin d'effectuer une modélisation par homologie de EED. L'obtention de ce modèle

moléculaire de EED a permis des études préliminaires sur la relation structure-fonction de EED et a permis de valider les résultats obtenus sur la cartographie des zones d'interaction avec ses partenaires viraux MA et IN. Les régions antigéniques localisées sur des boucles susceptibles d'interagir avec ses partenaires viraux ont été confirmées par la technique de «phage-display». L'analyse de ce modèle montre également que EED pourrait stabiliser l'Intégrase sous une forme tétramérique.

Il est clair que la détermination des structures de EED seule et des complexes EED-MA, EED-IN et EED-Nef permettrait une approche rationnelle pour la mise au point de nouveaux inhibiteurs anti-VIH. Ceux-ci permettrait d'inhiber spécifiquement deux étapes majeurs du cycle viral du VIH-1, à savoir la migration du complexe de pré-intégration dans le noyau et / ou l'étape d'intégration du provirus dans le génome de l'hôte. Ces inhibiteurs seraient soit des peptides suicides présentant une affinité pour les sites de liaison EED-MA, EED-IN et EED-Nef, soit des inhibiteurs péptido-mimétiques des interactions EED-MA, EED-IN et EED-Nef.

Au cours de ce travail sur EED, nous avons été contacté par le Dr C. Ronfort dirigeant l'équipe «Rétrovirus et Intégration Rétrovirale» du laboratoire Rétrovirus et Pathologie Comparée (UMR 754, INRA-UCBL-ENVL) afin de caractériser les relations structure-fonction de l'Intégrase aviaire de ALSV (Avian Leukemia and Sarcoma Viruses) par modélisation structurale. Ces études ont donné lieu à deux publications (Moreau *et al.*, 2004 ; Moreau *et al.*, 2003) présentées ci-après.

Développement d'inhibiteurs de l'Intégrase

L'Intégrase, en tant que troisième enzyme rétrovirale (à côté de la transcriptase inverse et de la protéase) représente une nouvelle cible thérapeutique anti-VIH importante. Elle est responsable de l'intégration du génome rétroviral dans l'ADN de la cellule hôte, étape capitale du cycle viral sans laquelle celui-ci ne peut pas se poursuivre. En outre, l'Intégrase, contrairement aux deux autres enzymes rétrovirales, ne possède pas d'homologues cellulaires connus. Il est donc probable qu'un inhibiteur contre l'Intégrase soit moins toxique que les inhibiteurs actuellement utilisés. Enfin, cette protéine est extrêmement conservée structurellement et des inhibiteurs développés contre l'Intégrase du VIH-1 seront probablement efficaces chez d'autres types viraux, comme le rétrovirus humain HTLV-1 (Human T Leukemia Virus type 1).

La conception d'inhibiteurs contre l'Intégrase a longtemps été retardée en raison de l'absence de structures tridimensionnelle de cette protéine. Néanmoins, deux inhibiteurs de l'Intégrase dérivés d'acides β-dicétoniques sont actuellement en phase d'étude préclinique (Anthony, 2004 ; Dayam et Neamati, 2003 ; Johnson *et al.*, 2004). De plus, des souches virales résistantes contre ces inhibiteurs ont déjà été décrites dans des expériences réalisées *in vitro* (Fikkert *et al.*, 2003).

L'équipe de C. Ronfort avait mis en évidence par mutagenèse dirigée des résidus impliqués dans la formation d'oligomères et d'autres impliqués dans la reconnaissance de l'ADN viral. Nous avons effectué un travail de modélisation et montré le rôle majeur de certains résidus, tant du domaine catalytique que du domaine C-terminal. Ces résidus interviennent dans le repliement correct de l'enzyme et dans l'interface dimérique de l'Intégrase.
D'après ces travaux, il apparaît que la mutation d'un seul résidu peut affecter la multimérisation de la protéine et inhiber totalement le processus d'intégration. Ainsi, l'inhibition de la multimérisation de l'Intégrase apparaît comme une stratégie privilégiée pour le blocage de l'étape d'intégration au cours du cycle infectieux.

Afin de poursuivre cette étude, des demandes de financement ont été déposées auprès de l'ANRS et de la région Rhône-Alpes. Le projet porte sur l'étude des relations structure-fonction de l'Intégrase du VIH-1 et le développement d'inhibiteurs antiviraux. Il vise également à résoudre la structure cristallographique entière de l'Intégrase de ALSV. En effet, les Intégrases du VIH-1 et de ALSV sont très proches : leurs structures tridimensionnelle sont très proches au niveau de domaine central mais diffèrent dans la position du domaine C-terminal par rapport au domaine central (Chen *et al.*, 2000 ; Yang *et al.*, 2000).

L'étude structurale de l'Intégrase d'ALSV dont le repliement tridimensionnel du domaine N-terminal est inconnu pourrait permettre de résoudre la première structure tétramérique d'Intégrase. En effet, deux mutants ponctuels présentant des activités supérieures à la protéine sauvage ont été identifiés au cours de nos travaux sur la caractérisation des relations structure-fonction de l'Intégrase aviaire de ALSV (Moreau *et al.*, 2004). Ces deux Intégrases mutées sont particulièrement intéressantes. En effet, dans la mesure où l'activité de ces protéines est supérieure à l'activité de la protéine sauvage, on peut supposer que les structures oligomériques formées par ces Intégrases sont plus stables que les structures formées par l'Intégrase sauvage et donc plus propice à cristalliser. Les essais de surproduction de ces deux mutants ont commencé en collaboration avec le Dr K. Moreau de l'équipe de C. Ronfort.

Publication 2

Mutations in the C-terminal domain of ALSV (Avian Leukemia and Sarcoma Viruses) integrase alter the concerted DNA integration process *in vitro*

Karen Moreau, Claudine Faure, Sébastien Violot, Gérard Verdier and Corinne Ronfort

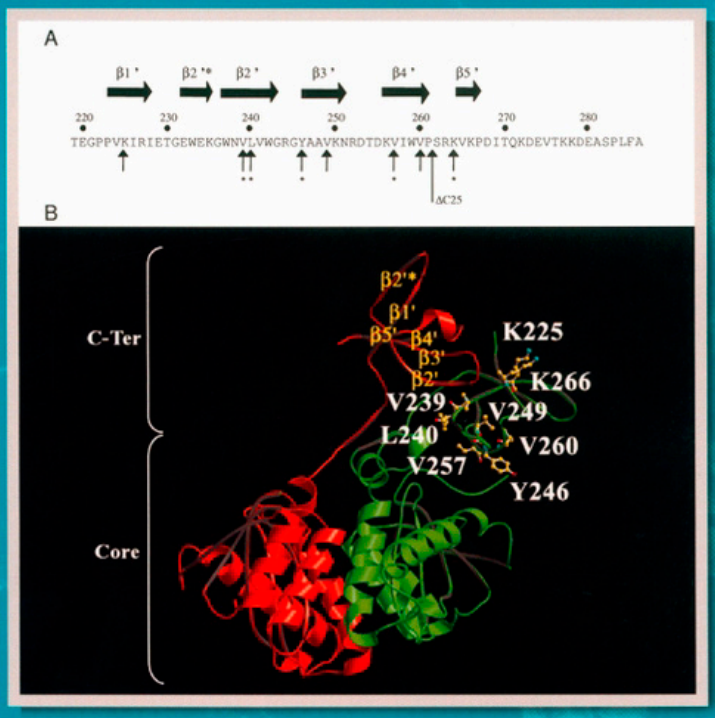
ISSN 0014-2956
http://www.ejbiochem.com

Volume 270 | Number 22 | November 2003

EJB The FEBS Journal

Review Article

Catalytic mechanism of lactoperoxidase and myeloperoxidase



Blackwell Publishing

Eur. J. Biochem. **270**, 4426–4438 (2003) © FEBS 2003 doi:10.1046/j.1432-1033.2003.03833.x

Mutations in the C-terminal domain of ALSV (Avian Leukemia and Sarcoma Viruses) integrase alter the concerted DNA integration process *in vitro*

Karen Moreau[1], Claudine Faure[1], Sébastien Violot[2,3], Gérard Verdier[1,3] and Corinne Ronfort[1,3]

[1]*Université Claude Bernard, Centre National de la Recherche Scientifique, Institut National de la Recherche Agronomique, Lyon, France;* [2]*Institut de Biologie et Chimie des Protéines, Centre National de la Recherche Scientifique, Laboratoire de Bio-Cristallographie, Université Claude Bernard, France;* [3]*IFR 128 'BioSciences Lyon Gerland', Lyon, France*

Integrase (IN) is the retroviral enzyme responsible for the integration of the DNA copy of the retroviral genome into the host cell DNA. The C-terminal domain of IN is involved in DNA binding and enzyme multimerization. We previously performed single amino acid substitutions in the C-terminal domain of the avian leukemia and sarcoma viruses (ALSV) IN [Moreau *et al.* (2002). *Arch. Virol.* **147,** 1761–1778]. Here, we modelled these IN mutants and analysed their ability to mediate concerted DNA integration (in an *in vitro* assay) as well as to form dimers (by size exclusion chromatography and protein–protein cross-linking). Mutations of residues located at the dimer interface (V239, L240, Y246, V257 and K266) have the greatest effects on the activity of the IN. Among them: (a) the L240A mutation resulted in a decrease of integration efficiency that was concomitant with a decrease of IN dimerization; (b) the V239A, V249A and K266A mutants preferentially mediated non-concerted DNA integration rather than concerted DNA integration although they were found as dimers. Other mutations (V260E and Y246W/ΔC25) highlight the role of the C-terminal domain in the general folding of the enzyme and, hence, on its activity. This study points to the important role of residues at the IN C-terminal domain in the folding and dimerization of the enzyme as well as in the concerted DNA integration of viral DNA ends.

Keywords: concerted DNA integration; integrase; multimerization; mutations; retroviruses.

Integration of the retrotranscribed viral DNA into a host cell chromosome, an essential requirement for viral gene expression and hence retroviral replication, is mediated by the viral integrase (IN). Integration also requires short specific DNA sequences at the viral DNA ends, designated *att* sequences [1]. Using *in vitro* assays, it has been shown that the integration process occurs in three steps as illustrated in Fig. 1A. Firstly, two terminal nucleotides are removed from both 3′ viral ends to generate the CA-3′OH ends, with a two-base 5′ overhang (3′-processing step). Secondly, during the strand transfer reaction, the 3′ viral ends are linked to the host DNA in a single cleavage–ligation reaction. The host DNA is asymmetrically cleaved and the insertion of the two viral DNA ends typically occurs 4–6 bp apart, according to the retrovirus [1]. In the third step (gap filling), the 5′ overhanging dinucleotides of the viral DNA ends are removed and single-stranded DNA gaps are repaired, creating a short duplication (4–6 bp) of host sequence. The integration process is defined as concerted because it enables the concomitant integration of two viral DNA ends at the same site of the host cell DNA generating a complete provirus flanked by short host DNA repeats [1]. Steps of 3′-processing and strand transfer are catalysed by the viral IN enzyme whereas repairing DNA gaps most probably involves cellular enzymes [2–5].

Concerted DNA integration has been reconstituted *in vitro* using a short linear DNA flanked by viral *att* sequences at its ends as donor DNA, a suitable plasmid as acceptor DNA and the IN enzyme supplied either as preintegration complex purified from infected cells or as a recombinant protein. This system has been developed with Avian Leukaemia and Sarcoma Viruses (ALSV) [6–13], HIV [12,14–19], Simian Immunodeficiency Virus [20] and more recently Murine Leukemia Virus [21] integrases. Such an *in vitro* assay has allowed reproduction of the integration process as observed *in vivo,* with the cleavage of the two terminal nucleotides of viral DNA ends and the duplication of a short acceptor DNA sequence.

The IN enzyme, which consists of three domains, is rather well conserved among the different retroviruses [22–24]. The C-terminal domain is the least conserved and contains no recognizable active site, but is necessary for both 3′-processing and strand transfer activities *in vitro* [25,26]. It is involved in binding to both viral DNA and nonspecific target DNA [27–29]. Several experiments have shown that the C-terminal domain is also involved in the oligomerization of IN. Indeed, ASLV and HIV INs are present as

Correspondence to C. Ronfort, Laboratoire 'Retrovirus et Pathologie Comparée', UCBL-INRA-ENVL, Université Claude Bernard. 50, avenue Tony Garnier, 69366 Lyon cedex 07, France.
Fax: +33 437 287 605, Tel.: +33 437 287 629,
E-mail: ronfort@univ-lyon1.fr

Abbreviations: ALSV, Avian Leukemia and Sarcoma Viruses; *att*, attachment sequence; HMG, high mobility group; IN, integrase; RSV, Rous Sarcoma Virus; RF, recombinant form; DSS, disuccinimidyl suberate.

(Received 25 July 2003, revised 9 September 2003, accepted 12 September 2003)

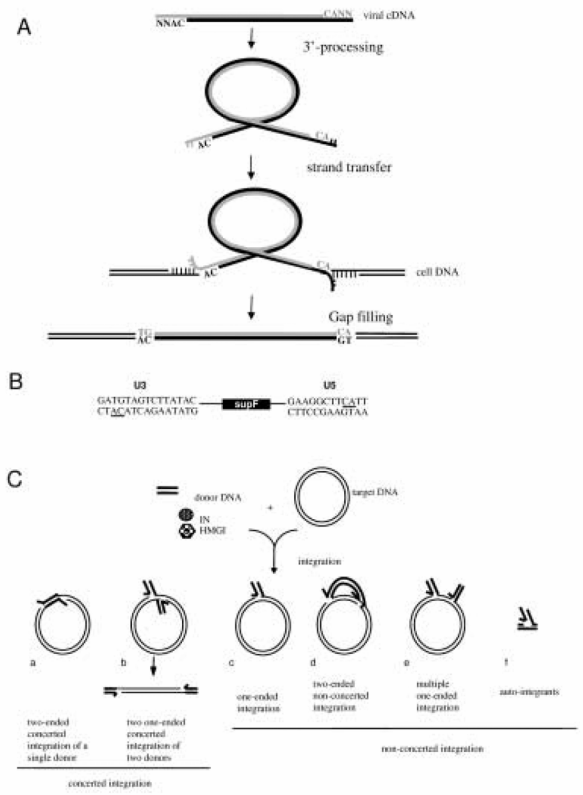


Fig. 1. Schematic representation of the retroviral integration process and principle of the *in vitro* concerted DNA integration assay. (A) Retroviral integration. The viral DNA made by reverse transcription is linear and blunt-ended. In the first step of integration (3′-processing), two nucleotides are removed from each 3′ end of the viral DNA. In the second step (strand transfer), the hydroxyl groups at the 3′ ends of the processed viral DNA attack a pair of phosphodiester bonds in the target DNA. In the last step (gap filling), completion of the integration process requires removal of the two unpaired nucleotides at the 5′ ends of the viral DNA and filling in the gaps between target and viral DNAs, generating a duplication of target DNA. (B) *In vitro* assay. Representation of the donor DNA with 15 bp of the U3 viral end and 12 bp of the U5 viral end. The highly conserved CA dinucleotides are underlined. The closed rectangle represents the *sup*F tRNA transcription unit. (C) *In vitro* assay. Schematic representation of the reconstituted integration reaction with the donor DNA, acceptor plasmid, purified integrase and HMGI proteins. Concerted DNA integration products include those that result from use of both ends from a single donor (product **a**) and from use of different ends from two donors (product **b**). Note that when two donors are inserted at the same site, a linear product is synthesized. Non-concerted DNA integration products result from one-ended integration of a single donor (product **c**), or two-ended integration of a single donor with insertion at different sites on the acceptor DNA (product **d**), or one-ended integration of two or more donors at different sites on the acceptor DNA (product **e**). Auto-integrants result from integration of a donor DNA in a second donor DNA (product **f**). Adapted from [13].

monomers, dimers and tetramers in solution, as shown by exclusion chromatography and analytical ultracentrifugation [30–38]. Within the C-terminal domain, deletion of residues 208–286 of ALSV IN or residues 218–288 of HIV IN proteins result in proteins deficient in multimerization [31,34] and specific mutations in the C-terminal domain inhibit the oligomerization of HIV-1 IN [39,40]. Conversely, the ALSV IN 201–286 fragment was shown to self-associate [31] and NMR analysis revealed that the C-terminal domain of HIV IN form dimers in solution [41]. The formation of multimeric molecules is essential for correct IN function, as shown by *trans*-complementation experiments *in vitro* [25,26] and *in vivo* [42,43]. It has been suggested that IN may function as a dimer, a tetramer or even as an octamer complex during the integration process [23,32–35,37,38,44].

We have previously introduced specific changes in selected amino acid in the C-terminal domain of an ALSV IN [24] and analysed the effects of these mutations on the catalytic activities of the resulting proteins [3′-processing, strand transfer and disintegration (reversal of strand transfer)]. These assays of catalytic activities relied on the use of short oligonucleotides carrying a unique viral end. In the present study, our aim was to test effects of several mutations on integration of two viral ends (concerted DNA integration) in an *in vitro* assay, as well as on oligomerization of IN. Recently, a two-domain structure of the Rous Sarcoma Virus (RSV) IN was published [23]. We used this structure to model the structure of the mutants. Our analyses focussed on proteins mutated at conserved residues or on residues shown to be involved in the dimer interface. These analyses allow us to identify the important role of specific residues within the C-terminal domain of ALSV IN.

Experimental procedures

DNA manipulation

The DNA pBSK-Zeo acceptor plasmid was constructed as follows: Plasmid pBSK + (Stratagene) was digested with *Sma*I and *Sac*II restriction enzymes, treated with Klenow DNA polymerase and reclosed by ligation to generate plasmid pBSK + Δ*Bam*HI. This was then digested with *Hin*dIII and *Eco*RV, filled by Klenow enzyme and reclosed by ligation to generate plasmid pBSK + Δ2. These manipulations removed the *Bam*HI and *Eco*RV restriction sites, respectively. Then, plasmid pBSK + Δ2 was amplified by PCR using *Pfu* turbo polymerase (Stratagene) and primers BU (5′-CCGATATCATACTCTTCC-3′) and BL (5′-CC GATATCAGACCAAGTTTAC-3′). In the same way, the *zeo* gene was amplified from plasmid pHook (Invitrogen) using primers Z1 (5-CCGATATCGTGTTGACAATT AATC-3′) and Z2 (5′-CCGATATCCAGACATGATAA GATAC-3′). All primers contain an *Eco*RV restriction site and resulting PCR products, pBSK + Δ2 and *zeo* gene, were digested by the *Eco*RV restriction enzyme and ligated together to produce plasmid pBSK-zeo. This plasmid, which carries the zeocin resistance gene, was amplified in *E. coli* DH5α (Invitrogen).

The donor DNA was obtained as follows: *sup*F gene was amplified by PCR from piVX plasmid (ATCC) using primers H-sup1 (5′-GAGAAGCTTAACGTTGCCCGG ATCCGGTC-3′) and P-sup2 (5′-GAGCTGCAGTAGTC CTGTCGGGTTTCGCC-3′) containing *Hin*dIII and *Pst*I restriction sites, respectively. The amplification product was digested with *Hin*dIII and *Pst*I restriction enzymes and ligated into the pBSK + plasmid digested by the same

restriction enzymes, giving pBSK-supF plasmid. The donor DNA was then amplified from pBSK-supF plasmid by PCR using *pfu* turbo polymerase, and primers U3 (5′-GATGTAGTCTTATACGTTGCCCGGATCCGG-3′) and U5bi (5′-AATGAAGCCTTCTGCTTTGAGCGTCGATTTTTG-3′). The PCR product was purified from agarose gel using the Qiaex II kit (Qiagen). The final donor DNA contained 15 bp of the U3 end sequence of Avian Erythroblatosis Virus and 12 bp of the U5 end.

Modelling of the mutants

Construction of IN mutants has been reported elsewhere [24]. Two two-domain structures of RSV IN, containing the core and the C-terminal region, have been solved in space groups $P2_1$ and P1 at 3.1-Å and 2.5-Å resolution, respectively [23]. No structure containing also the N-terminal domain has yet been published. In consequence, the 2.5 Å two-domain structure of RSV IN was used to model the structure of mutants. Modelling was performed on the dimer. Each structure of a single mutant was generated using the program CALPHA [45] and minimized with the program CNS using a conjugate gradient method [46]. Resulting models were displayed and analysed on a graphic station using the program TURBO-FRODO [47]. Contact distances were computed with CNS around each mutated residue. In parallel, a BLAST search [48] was performed against the SWISS-PROT and the TrEMBL sequences databases [49] to detect homologous proteins. A multiple sequence alignment was performed in turn with CLUSTAL [50]: the eight studied substitutions are unique in retrovirus as well as in lentivirus integrases.

Purification of proteins

IN mutants [24] were expressed in BL21 bacteria (Invitrogen) and purified as described by others [40].

The HMGI(Y) proteins (high mobility group; now referred as HMGa1) consist of two proteins (HMGI and HMGY) which are expressed from the same gene and differ by altenative mRNA splicing. The pET15b-HMGI vector (generously donated by T. H. Kim, Harvard University, Cambridge, MA, USA) expresses HMGI [51]. The HMGI protein was expressed in BL21(DE3) pLysS bacteria (Invitrogen) in the presence of 100 $\mu g \cdot mL^{-1}$ ampicillin and 34 $\mu g \cdot mL^{-1}$ chloramphenicol upon induction with 1 mM of isopropyl-thio-β-D-galactopyranoside for 3 h. Purification was carried out as follows. The bacterial pellet was resuspended in NaCl/P_i containing 0.1% Triton X-100 and sonicated. Then 5% of perchloric acid was added and the solution was incubated for 30 min at 4 °C. The lysate was then centrifuged for 10 min at 12 000 ***g***. A total of 25% of trichloroacetic acid was added to the supernatant which was incubated for 1 h on ice. After 10 min centrifugation at 12 000 ***g***, the pellet was rinsed once with acetone and 0.2% HCl (−20 °C), twice with acetone 70%/ethanol 20%/20 mM Tris/HCl pH 8 (−20 °C), and once with acetone (−20 °C). The pellet was dried at room temperature before being resuspended in 250 µL Tris/EDTA, pH 8.0. The solution was passed through a Hitrap Heparin column (Pharmacia), which had been equilibrated with 0.5 M NaCl, 50 mM NaH_2PO_4 pH 7.4. The column was washed with 0.5 M NaCl, 50 mM NaH_2PO_4 pH 7.4 and the proteins were eluted with a gradient of 0.5–1.5 M NaCl. Each fraction was analysed by Bradford quantification and Western blot.

Integration reaction

Purified IN protein (60 ng) was incubated overnight at 4 °C with 100 ng pBSK-zeo plasmid, 10 ng donor DNA and 100 ng purified HMGI protein in a final volume of 5 µL. The volume of reaction was then increased to 20 µL with a final concentration of 20 mM Hepes, pH 7.5, 1 mM dithiothreitol, 30 mM $MgCl_2$, 15% dimethyl sulfoxide, 8% PEG 8000 and 50 mM NaCl, and the integration mixture was incubated at 37 °C for 90 min.

Gel analysis of the integration reaction

For gel analysis of the integration reaction, the DNA donor was radiolabelled by including 8 µCi [^{32}P]dCTP[αP] in the PCR amplification mixture. After the integration reaction was performed, the volume was increased to 50 µL by the addition of 4.25 mM EDTA, 0.44% SDS and 20 ng proteinase K (Roche Diagnostics) and samples were incubated for 1 h at 55 °C. The DNAs were deproteinized by phenol/chloroform extraction and purified by ethanol precipitation. Samples were then loaded on 1.2% agarose gel in 0.5 × Tris/borate/EDTA electrophoresis buffer. After electrophoresis, the gels were fixed in 5% trichloroacetic acid for 30 min and dried for 3 h at 45 °C. Lastly, the gels were exposed to autoradiographic film overnight at −80 °C. Integration products were quantified using a phosphoimager (Biorad).

Cloning and sequencing of two-ended integration products

To clone integration products for sequencing, products of the integration reaction were purified on a Qiaquick column (Qiagen) as described by the supplier. The whole reaction was introduced into MC1060/P3 *E. coli* (Invitrogen) as described by others [9]. MC1061/P3 *E. coli* carry ampicillin, tetracyclin and kanamycin resistance genes. Both ampicillin and tetracyclin resistance genes carry an *amb* mutation. These proteins are thus expressed only in the presence of the *supF* gene products. Integration clones carrying both zeocin-resistant and *supF* genes were therefore selected in the presence of 40 $\mu g \cdot mL^{-1}$ ampicillin, 10 $\mu g \cdot mL^{-1}$ tetracyclin, 15 $\mu g \cdot mL^{-1}$ kanamycin and 25 $\mu g \cdot mL^{-1}$ zeocin. Plasmids were isolated from quadruply resistant colonies and donor–acceptor DNA junctions were sequenced using SL primer (5′-ACTCTAAATCTGCCGTCATCG-3′) for the U3 junction and SU primer (5′-ATCATATCAAATGACGCGCCG-3′) for the U5 junction. SL and SU primers are located on the donor DNA.

Size exclusion chromatography

All proteins were centrifuged for 10 min at 14 000 r.p.m. to remove IN aggregates. A total of 100 µL integrase solution at a concentration of 30 µM was loaded on a Superoz 12 column (Pharmacia) equilibrated previously with 1 M NaCl, 25 mM Hepes pH 7.5, 0.1 mM EDTA, 1 mM β-mercaptoethanol. Size exclusion chromatography was performed at

4 °C. The column was calibrated with molecular mass markers. Protein elution was monitored at A_{280} nm at a flow rate of 0.3 mL·min^{-1}.

Protein–protein cross-linking

Wild-type or mutant integrases were treated with 40 μg·mL^{-1} disuccinimidyl suberate (Pierce). Reactions included 2 μg protein in a final volume of 10 μL 20 mM Hepes pH 7.5, 60 mM NaCl, 0.7 mM EDTA, 10% glycerol, 4.5 mM Chaps. After 30 min at 22 °C reactions were quenched by the addition of 3 mM lysine and 25 mM Tris/HCl pH 8. After a further 10 min at 22 °C, reactions were boiled for 10 min in sample buffer and separated by SDS/PAGE (10% acrylamide). Products were revealed by Western blot using anti-His-tag Ig (Roche Diagnostics).

Results

Reconstitution of the concerted DNA integration assay *in vitro*

The *in vitro* retroviral concerted DNA integration system (Fig. 1B,C) has previously been described by others [9,12,13]. It is composed of a linear donor DNA, a plasmid acceptor DNA and recombinant IN. HMGI protein is added to the reaction because it has been found to enhance the concerted DNA integration reaction [12]. HMGI is a component of the HMGI(Y) protein (now referred as HMGa1). HMGI(Y) is a DNA binding protein that has been found in HIV preintegration complexes isolated from infected cells [52]. HMGI(Y) might stimulate concerted DNA integration by bending the donor DNA and helping to bring the two ends into close proximity; alternatively, the unwinding activity of HMG proteins could facilitate binding of IN proteins to DNA ends and their subsequent distortion [12,53].

In the present report, we used the IN protein from Rous Associated Virus type 1, and a donor DNA of 326 bp containing 15 bp of the U3 *att* sequence at one end and 12 bp of the U5 *att* sequence at the other end (Fig. 1B). Products of the integration reaction can arise from concerted or non-concerted DNA integration (Fig. 1C) [9,12,13]. Two-ended concerted DNA integration products include those that result from integration of both viral ends from a single donor (product **a**) or those that result from integration of two viral ends from two donors at the same integration site (generating the linear product **b**). Non-concerted DNA integration products result from one-ended integration of a single donor (product **c**), from two-ended integration of a single donor with insertion at different sites on the acceptor DNA (product **d**), or from one-ended integration of two or more donors at different sites on the acceptor DNA (product **e**). Auto-integration products, which are the results of the integration of donor DNA in a second donor DNA are also observed (product **f**).

By using labelled donor DNA, the integration of the small donor DNA into larger acceptor DNA can be visualized by autoradiography after separation on agarose gel. Under these conditions, three characteristic bands were revealed in presence of IN (Fig. 3A, lane 3). As described previously by others [7,8,16], the slowest band correspond to a mix of circular forms (Recombinant Form RFII products: **a**, **c** and **d**), the middle band correspond to the linear form **b** (RFIII products) and the fastest band correspond to auto-integration products (form **f**). Product **e**, which migrates more slowly because two or more donors are inserted into the target, is observed on some gels, but not all. A recombinant, identified by an asterisk in Fig. 3A, and which migrated slightly faster than the RFII recombinants has been observed by others [6,10,16,18,19]; its structure is unknown [18]. Total integration products were cleaved with either *Bam*HI (which cleaves the donor DNA) or *Xho*I (which cleaves in the acceptor DNA). Structures of digestion products were fully consistent with the above assignment of the DNA forms (data not shown). As controls, reactions were performed in the absence of IN (Fig. 3A, lane 1) or with an IN mutated in its catalytic site, the D121E mutant (lane 2) [24]. No integration product was observed demonstrating that the products observed with wild-type protein resulted from IN enzymatic activity.

Gel analysis permits the quantification of integration efficiency but does not distinguish one-ended from two-ended integration products, as product **a** is not resolved independently of other RFII forms (**c** and **d** products). However, integration products can also be cloned into MC1061/P3 *E. coli*, which contain drug resistance markers with amber mutations. Only DNA products carrying the amber mutation suppressor gene (*sup*F) should be able to replicate and form colonies under drug selection. Among the different integration products, one-ended or multiple one-ended donor integration products (**c** and **e**) and linear product (**b**) should be lost upon cloning into *E. coli*. Only the circular two-ended integration products (forms **a** and **d**) should be able to replicate into bacteria [14,15]. Thus, the cloning analysis enables estimation of the efficiency of IN proteins to perform two-ended donor integration (concerted, form **a**; or not, form **d**). Following cloning, donor DNA–acceptor plasmid junctions of isolated integration products have to be sequenced in order to check the accuracy of the integration reaction (cleavage of viral ends and duplication of short acceptor DNA sequence).

Integration products generated with wild-type IN were cloned. Between 98 and 324 colonies were observed according to the experiments. Thirty-one clones were isolated and sequenced (Table 1). Sixteen clones exhibited a target DNA duplication of 6 bp and 11 clones a duplication of different size (from 4 or 5 bp). *In vivo*, the 6-bp duplication is a hallmark of ALSV viruses [54,55] although some size variations have been reported [56]. In *in vitro* assays, shorter duplications have often been observed [9,12,13]. Four clones exhibited a deletion of acceptor DNA. However, as these clones were correctly cleaved at both ends and integrated between the canonical TG and CA viral dinucleotides, they were interpreted as the result of an IN mediated process but with incorrect cleavage of the acceptor DNA. Integration products with acceptor DNA deletion could arise from either two independent one-ended donor integration events (form **e**) or from nonconcerted DNA integration of the two ends of one donor (form **d**). Assuming that only circular integration products (**a** and **d**, Fig. 1C) could be amplified in bacteria [14,15], we speculate that these clones were most probably the result of a non-concerted DNA

Table 1. Sequencing of donor–target junctions from clones produced by wild-type and V239A INs. Square brackets, number of clones harbouring incorrect cleavage of *att* sequences (deletion of more than the 2 nucleotides expected).

	Products obtained with [n (%)]	
	WT	V239A
Duplication size		
7 bp	0	1 (3.5)
6 bp	16 (51.5) [2]	18 (60) [1]
5 bp	8 (26) [1]	3 (10) [1]
4 bp	3 (9.5) [1]	1 (3.5) [1]
Deletion	4 (13)[a]	7 (23)[b]
Total	31	30

[a] Deletions range from 150 to 948 bp, [b] deletions range from 33 to 503 bp.

integration of the two viral ends of one donor DNA at two different sites on acceptor plasmid DNA (form **d**, Fig. 1C). Deletion in the acceptor DNA by two-ended nonconcerted DNA integration events has already been described [12–15]. Regarding the viral DNA ends, we observed deletion of more than the two expected nucleotides at one or the other *att* sequence in four clones. These four clones exhibited a duplication of acceptor DNA at the integration site, which led us to conclude that they have indeed arisen from a mechanism of integration mediated by IN. In other works describing the ALSV concerted DNA integration, neither deletions of acceptor DNA nor the use of internal cleavage sites on the donor DNA were observed using the wild-type enzyme unless the viral sequences were mutated [13]. These assays with wild-type IN have typically used 15 bp of viral sequence at each end, while we used 12 bp of U5 instead of 15. Therefore, it is possible that the structures we observed were generated due to the small U5 att site. IN may have used a larger U5 att site, recognizing the nonviral sequence covalently linked to the att site, which would represent a mutant att site. However, the number of such clones is rather low and does not impair the following analyses since all mutants were systematically compared with wild-type IN.

Description and modelling of IN mutants

Arrangement of the C-terminal domain. We have previously constructed mutants, each containing single amino acid substitutions in the C-terminal domain of the Rous Associated Virus type 1 IN [24] (Fig. 2A). In the meantime, a 2.5-Å structure of the closely related RSV IN was published [23] containing both the core and C-terminal domains. In this structure, the two core domains are related by a twofold symmetry axes, whereas the two C-terminal domains have a similar fold but associate asymmetrically, giving rise to a 'proximal' and a 'distal' domain (close to the core domain or away from it, respectively; Fig. 2B). Therefore, equivalent residues of the 'proximal' and 'distal' domains have a different environment at interface regions [23]. The C-terminal domain is composed of six strands forming a β-barrel fold resembling an SH3 domain (Fig. 2)

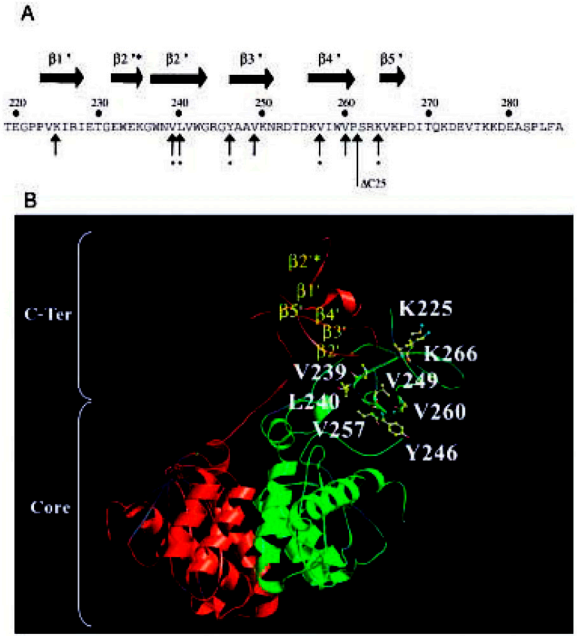


Fig. 2. Description of the IN mutants analysed. (A) Sequence of the ALSV C-terminal domain (residues 219–286) is shown. Above are indicated β-strands (large arrows) [23]. Arrows indicate residues mutated in the present study. Arrows with asterisks indicate residues at the C dimer interface. Longer arrow at position 261 indicates end of Y246W/ΔC25 IN mutant. (B) Ribbon representation of the dimeric two-domain structure of RSV integrase (residues 54–268). Green and red molecules represent 'proximal' and 'distal' subunits, respectively. Labels on the green subunit correspond to the eight mutated residues discussed in this paper. Labels on the red subunit indicate β-strands, strand β2′ being designated as two shorter strands β2′* and β2′ (adapted from [23]).

[23]. Strands β1′, β2′ and β5′ of the proximal monomer and strands β2′, β3′ and β4′ of the distal monomer are involved in the dimer interface (Fig. 2B). It is noteworthy that the two-domain structures of HIV-1 and Simian Immunodeficiency Virus INs [57,58] show different arrangements of the C-terminal domains. The biological relevance of this is unclear: it may indicate considerable flexibility in the linkage between the core and C-terminal domains [59].

Modelling of IN mutants. We used the two-domain structure of RSV [23] to model the mutants studied here.

First, five of the mutants studied herein carried mutations on residues involved in the dimer interface (Fig. 2, Table 2). This includes the V239 and K266 residues of the proximal monomer and the L240, Y246, V257 residues of the distal monomer. Mutations of these residues were supposed to affect the dimeric interface.

V239 is located in strand β2′ at the dimeric interface. Based on multiple sequence alignments, this residue is well conserved among INs [24]. The proximal V239 residue is involved in the interface with the second C-terminal domain and has an intermolecular long contact distance (4.1 Å) with residues V241 and W259 of the distal domain. The

Table 2. Contacts between residues in the monomers and dimers of the wild-type and mutants INs. The two-domain structure [23] was used to model the mutants. #, Residues of the distal subunit. In bold, residues at the interface of the dimer. Maximum contact distance is 5 Å, residues in italic type have contact distances < 3.2 Å.

ALSV	Location	Contacts between side chains in wild-type	Contacts between side chains in mutant
Proximal			
K225H	Strand β1′	W233, *K235*, K266, D268	W233, *K235*, *D268*
V239A	Strand β2′	P222, V224, #**V241**, W242, A247, V249, #**W259,** V265	P222, V224, W242, V265
L240A	Strand β2′	L55, L218, V241, A248, K250, V257	L55, V241, A248
Y246W	Strand β3′	R53, W259, P261	R53, W259, *P261*
V249A	Strand β3′	V224, I226, W237, V239, A247, I258, V260, V265	I226, W237, I258
V257A	Strand β4′	L55, L240, A248, K250, W259	L55, K250
V260E	Strand β4′	I226, A247, V249, I258, P261, K264, V265	*I226*, W246, I258, P261, K264
K266A	Strand β5′	K225, W233, P267, #**R244**	W233, P267, #**R244**
Distal			
#K225H	Strand β1′	#W233, #K235, #K266, #D268	#W233, #*K235*, #*D268*
#V239A	Strand β2′	#P222, #V224, #W242, #A247, #V249, #V265	#P222, #V224, #W242
#L240A	Strand β2′	**L218,** #E220, **P222,** #V241, #A248, #K250, #V257	**L218, P222,** #V241, #A248,
#Y246W	Strand β3′	#R244, #W259, #P261,#S262, **P267**	#R244, #W259, #*P261*, #S262, **P267**
#V249A	Strand β3′	#V224, #I226, #W237, #V239, #A247, #I258, #V260, #V265	#V224, #I226, #W237, #I258
#V257A	Strand β4′	#L240, #K250, **P223**	#K250
#V260E	Strand β4′	#I226, #V249, #P261, #K264, #V265	#*I226*, #W237, #V249, #I258, #P261, #V265
#K266A	Strand β5′	#K225, #W233, #P267	#W233, #P267

V239A mutation removes these two intermolecular contacts as well as several other intramolecular contacts within each monomer (with A247 and V249).

L240 is also located in strand β2′. This residue is well conserved in retroviruses [24]. The distal L240 is at the dimeric interface between C-terminal domains, and its side chain makes van der Waals' contacts with residues L218 and P222 in strand β1′ of the proximal monomer. The L240A mutation does not remove these contacts at the interface of the dimer. However, the mutation decreases the number of intramolecular contacts within both monomers.

Y246 is located at the beginning of strand β3′. The distal Y246 is involved in intermolecular contacts in the dimer through an interaction with P267 in strand β5′ of the proximal monomer. Nevertheless, the mutation Y246W does not remove this contact. It only reinforces the contact with P261 in each monomer.

V257 is located at the beginning of strand β4′. The distal V257 is involved in an intermolecular contact in the dimer through an interaction with P223 in strand β1′ of the proximal monomer. Residue V257 is also involved in a contact with the above-mentioned L240 residue within each monomer. The V257A mutation removes the intermolecular contact between monomers as well as several intramolecular contacts.

K266 is a well conserved residue located in strand β5′. At the dimeric interface, the proximal K266 is in contact with R244 of the distal monomer, a residue located in a turn between strands β2′ and β3′. Nevertheless, the K266A mutation does not remove this contact in the dimer. The mutation only removes a contact with K225 within each monomer.

Secondly, mutant Y246W/ΔC25, missing the 25 C-terminal residues was studied as well to evaluate the effect of deleting the terminal end of the C-terminal domain. The protein ends at P261, just after strand β4′ (Fig. 2A) and lacks the β5′ strand.

Finally, three other mutants were studied too:

K225 is a nonconserved residue of the β1′ strand. The conservative K225H mutation makes closer contact with the D268 residue within the monomer and removes an intramolecular contact with the K266 residue (Table 2).

V249 is a moderately well conserved residue of strand β3′ which is not involved in intermolecular contacts. Mutation V249A removes several contacts in the monomers, especially with the V260 residue.

V260 is a highly conserved residue of strand β4′. V260 in HIV-1 IN is potentially involved in the formation of multimeric complexes [39]. The V260E mutation was the same as that performed on HIV IN [39]. The V260E mutation replaces several contacts inside both monomers (W246 instead of A247, V249 and V265 in the proximal monomer, W237 and I258 instead of K264 in the distal monomer). It also makes closer contact with the I226 residue in each monomer (Table 2).

Catalytic activities of IN mutants. In the preliminary study [24], 3′-processing and strand transfer catalytic activities of wild-type protein and of each mutant were examined *in vitro*, using a 15-bp long oligonucleotide corresponding to the U5 *att* terminal sequence (Table 3). Briefly, mutant K266A was as efficient as wild-type protein for both activities. K225H, V239A, L240A and V249A mutants displayed a slightly reduced efficiency for 3′ processing while strand transfer activity was close to that of the wild-type protein. Y246W, Y246W/ΔC25, V257A and V260E mutants had 3′-processing activity that was drastically reduced compared to that of wild-type IN, while strand transfer activity was either correct or reduced (V260E). With the exception of V260E, all other mutants displayed a correct disintegration activity. Furthermore, mutants bound DNA with an efficiency similar to that of the wild-type protein [24] (Table 3).

Table 3. Compilation of data obtained for each mutant. Catalytic activity data from unpublished observations and from [24]. DNA binding data from [24]. Integration efficiency results from Fig. 3B (integration efficiencies as revealed on gel, and in comparison with wild-type IN efficiency). 1- and 2-ended results from Fig. 3C. Oligomeric status results from Figs 4 and 5. C, residues conserved among INs (as shown by sequence alignments [24] and checked by comparing crystallographic structures of the INs); 3′-P, 3′-processing; S.t., strand transfer; dis, disintegration; +, 0–30% activity of the wild-type IN; + +, 30–60% activity of the wild-type IN; + + +, 60–90% activity of the wild-type IN; + + + + > 90% activity of the wild-type IN; 1 = 2, level of 1- and 2-ended DNA integration events comparable to those of wild-type IN; 1 > 2, 1-ended DNA integration events are favoured over 2-ended DNA integration events, as revealed in *E. coli*; D, dimers; M, monomers; mis, misfolded; ND, not determined.

		Catalytic activities			DNA	Concerted integration		
Mutation	Conservation	3′-P	S.t.	dis	binding	Integration efficiency	1- and 2- ended	Oligomeric status
K225H	–	+ +	+ + +	+ + + +	+ + + +	Same	1 = 2	D
V239A[a]	C	+ +	+ + +	+ + + +	+ + + +	Increased	1 ≫ 2	D
L240A[a]	C	+ +	+ + +	+ + +	+ + + +	Reduced		D + M
Y246W[a]	–	+	+ + +	+ + + +	+ + + +	Reduced		D
V249A	C	+ +	+ + +	+ + + +	+ + + +	Same	1 > 2	D
V257A[a]	–	+	+ + +	+ + + +	+ + + +	Reduced		D
V260E	C	+	+ +	+ +	+ + + +	Reduced		mis
K266A[a]	C	+ + +	+ + + +	+ + + +	+ + + +	Same	1 ≫ 2	D
Y246W/Δ25	–	+	+ + +	+ + +	ND	Reduced		mis

[a] Residues at the dimer interface.

Analysis of integration efficiency of IN mutants

The IN mutants were analysed in the context of the concerted DNA integration assay *in vitro*. Integration reactions were performed in the presence of labelled donor DNA, and integration products were separated by electrophoresis (Fig. 3B). For each mutant, integration efficiency (Fig. 3C, black bars) was determined by calculating the intensity of bands corresponding to RFII and RFIII integration products (forms **a** + **b** + **c** + **d**) and in comparison with the wild-type protein. The experiment was repeated at least twice (according to the mutants) and results (integration efficiencies relative to wild-type IN) were similar in these experiments. The integration activity of V249A (lane 6) and K266A (lane 9) mutants was roughly similar to that of the wild-type IN (lane 1). The K225H (lane 2) and V239A (lane 3) mutants were slightly more efficient than wild-type IN. L240A (lane 4), Y246W (lane 5), V257A (lane 7), V260E (lane 8) and Y246W/ΔC25 (lane 10) mutants exhibited low activities as deduced from gel analyses (Fig. 3B,C). It is noteworthy that mutants which displayed 3′-processing reduced to a level 30% of that of wild-type IN (e.g. K225H, V239A, V249A) were nevertheless able to perform concerted DNA integration with high efficiency (Table 3). Only mutants displaying a strong reduction in 3′-processing activity (< 20% that of wild-type IN) such as Y246W, V257A and V260E did not perform concerted DNA integration with high efficiency. Afterwards, we focussed on the ability of IN mutants to perform two-ended integration.

First, the RFIII products containing the linear **b** form were quantified as this form was supposed to result from one event of two-ended concerted DNA integration. For each mutant, results are given as percentage of **b** products relative to total integration products (RFII/RFII + RFIII) (Fig. 3B, bottom). Product **b** represents 28% of total integration products generated by wild-type IN. For four mutants (L240A, Y246W, V260E and Y246W/ΔC25), product **b** was too low and was not quantified. For all others, product **b** represents 21–35% according to the mutants, which led us to conclude that there were no relevant differences between these mutants and wild-type IN regarding the ratio of the product **b**.

Second, integration products were cloned into *E. coli*. Integration efficiency was determined by comparing the number of clones obtained for each tested mutant to the one obtained with wild-type IN (Fig. 3C). For each mutant, the experiment was repeated at least twice and the independent experiments gave similar results (integration efficiencies relative to that of wild-type IN). The K225H mutant had an activity close to that of the wild-type protein and the V249A mutant presented a slightly reduced activity (118 and 62%, respectively). V260E and Y246W/ΔC25 mutants were totally defective (< 2% of the wild-type IN activity). All other mutants (V239A, L240A, Y246W, V257A and K266A) exhibited reduced activity, from 10 to 40% of wild-type IN activity.

For some mutants, the gel analysis (black bars) was in agreement with cloning analysis (white bars). Thus, the K255H mutation did not modify the integration efficiency as observed by electrophoresis and after cloning into *E. coli*. L240A, Y246W, V257A, V260E and Y246W/ΔC25 mutations modified the integration efficiency both on gels and after cloning into *E. coli*. On the contrary, V239A, K266A mutants, and to a lesser extent V249A, were found to be at least as efficient as the wild-type protein for integration by electrophoresis but they were less efficient for two-ended donor integration, as revealed by cloning. For these mutants, this result suggests that among the integration products observed on the gels, there was a lower proportion of two-ended integration products as compared to the wild-type protein (Table 3). Thus, these three mutations (V239A, K266A and V249A) appear to alter specifically the two-ended integration process.

Molecular characterization of integration products

After cloning, we sequenced integration products of mutants which displayed a reduced efficiency for two-ended

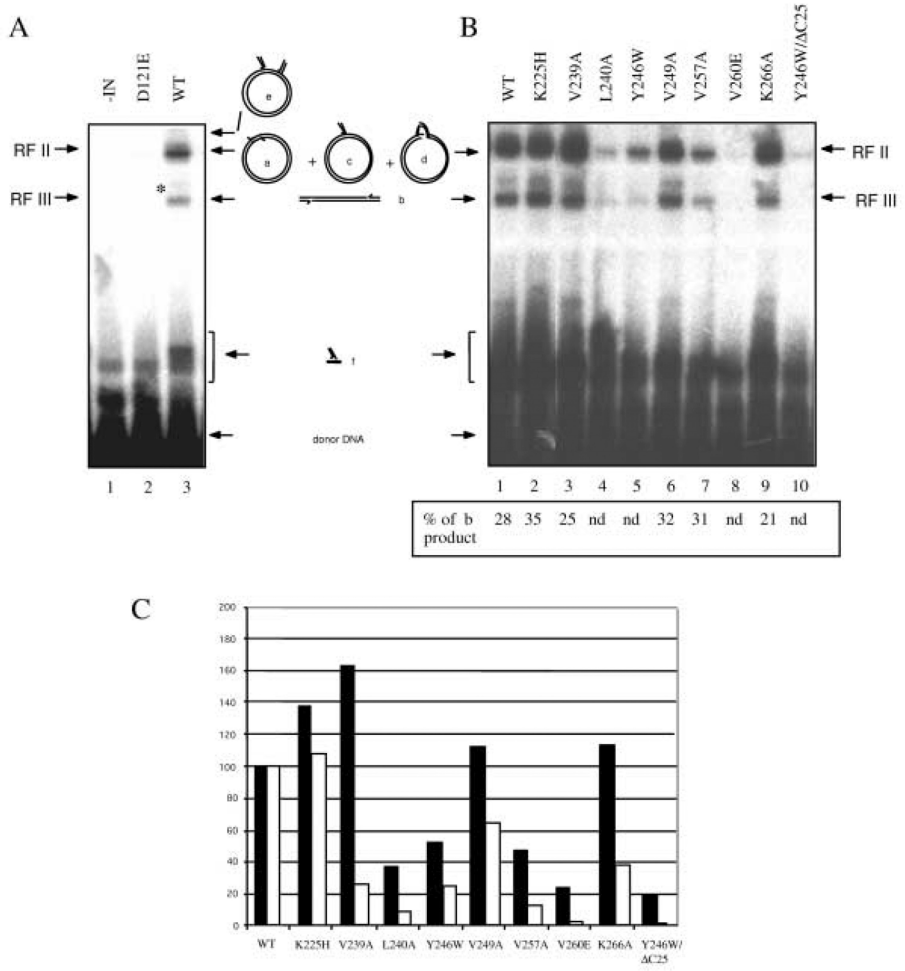


Fig. 3. Analysis of the integration products. (A) Integration reactions performed in absence of IN, with D121E IN mutant and the wild-type IN. –IN, reaction without IN. DNA products were analysed by gel electrophoresis. *Structure of this recombinant is unknown. (B) Integration reactions performed with wild-type IN and the C-terminal domain mutants. Letters above indicate the mutation: the first letter is the original residue, the number its position in the protein, and the second letter the residue that it was substituted into. Bottom: percentage of **b** product [RFIII forms relative to total integration products (RFII + RFIII)]. Nd, not determined. (C) Quantification of integration products shown in (B), corresponding to RFII plus RFIII products (in black) and total number of colonies recovered after the reaction products were introduced into bacteria (in white). Integration efficiency of wild-type protein was set as 100%. For cloning analyses, 100% correspond to 98–320 colonies per plate (according to the experiments) derived from reaction products with wild-type IN. Results for mutants are the mean of at least two experiments.

integration (Fig. 3C). For V239A, 30 clones were sequenced (Table 1). Eighteen clones exhibited a 6-bp duplication of acceptor DNA, and five a duplication of another size (4–7 bp). Among these clones, three exhibited incorrect cleavage of the U3 *att* sequence with more than two nucleotides deleted, although they exhibited short duplication of acceptor DNA. These structures were also observed with the wild-type IN in similar proportion and therefore were not characteristics of this mutant. Seven clones exhibited acceptor DNA deletion. As previously suggested for wild-type IN, these structures might be the result of a nonconcerted DNA integration of both viral ends at two different sites of acceptor DNA (form **d**, Fig. 1). Nevertheless, these structures seemed to be generated more by the V239A mutant (23%) than by the wild-type IN (13%), but the difference was not statistically significant ($P < 0.05$).

For the two other mutants specifically defective in two-ended integration (Fig. 3C) (K266A and V249A), about 10 clones were sequenced (data not shown). These clones did not display any differences with products obtained from wild-type protein. For them, the sequencing analysis was not extended.

In conclusion, these analyses show that V239A, V249A and K266A mutants performed a correct integration process, roughly comparable to that of the wild-type protein, with correct cleavage of viral ends and small size duplication of acceptor DNA.

Multimeric forms of IN proteins

It has been reported that IN acts as a multimeric complex during integration, and this complex is at least a

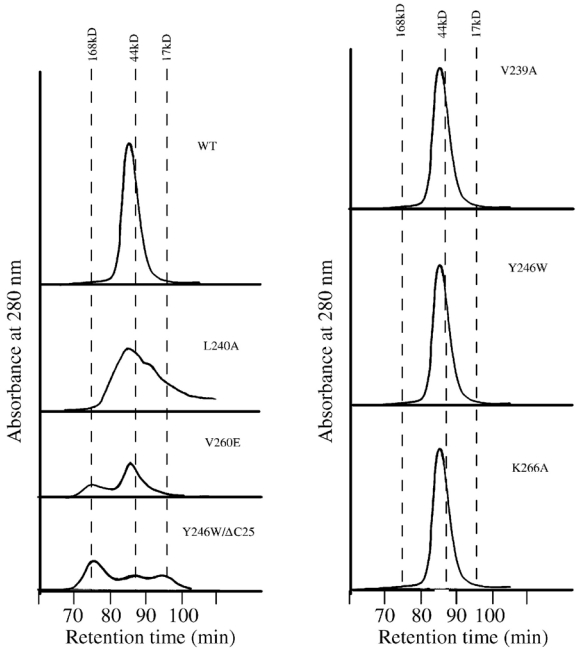


Fig. 4. Size exclusion chromatography of wild-type integrase and mutants. Elution profiles of wild-type IN as well as V239A, L240A, Y246W, V260E, K266A and Y246W/ΔC25 mutants are shown. The molecular size of monomeric form of all INs is 36.7 kDa except for the Y246W/ΔC25 mutant which is 33.9 kDa. For reference, the elution positions of three globular standard proteins are indicated by dotted vertical lines. Retention times in minutes are indicated on *x*-axis. Other mutants (K225H, V249A, V257A, which had the same profiles than the wild-type protein) are not shown.

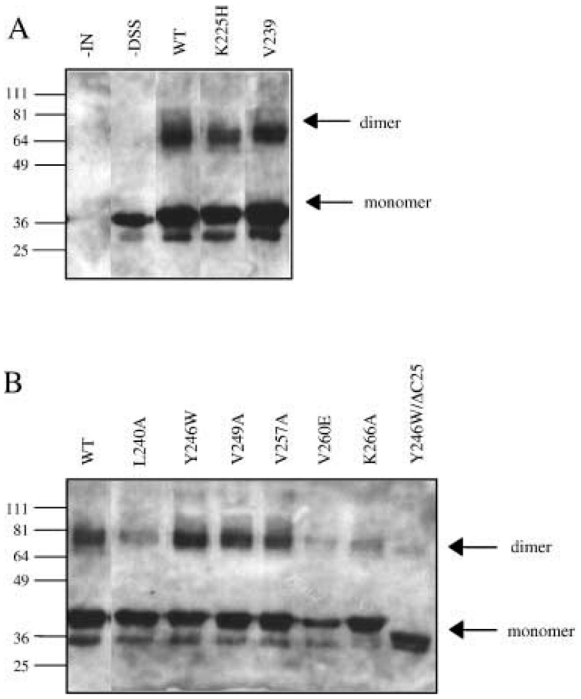


Fig. 5. Protein–protein cross-linking of wild-type integrase and mutants. Proteins were incubated in the presence of disuccinimidyl suberate (DSS). Reaction products were analysed on 10% polyacrylamide gels and revealed by Western blotting using anti-His-tag Ig. The migration of cross-linked species, monomers and dimers, are marked. (A) Controls, mutants K225H and V239A. –IN, Without integrase; –DSS, without DSS. (B) Other mutants.

dimer [25,26]. Some mutations studied herein involved residues at the dimer interface (Table 2). To test whether these substitutions altered the ability of IN to form dimers, the wild-type and IN mutants were analysed by size exclusion chromatography and protein–protein cross-linking.

In size exclusion chromatography (Fig. 4), wild-type protein eluted at a position consistent with the molecular size of a dimer. In similar conditions, others [31] also observed dimers of ALSV IN. Mutants V239A, Y246W and K266A (Fig. 4) as well as mutants K225H, V249A and V257A (data not shown) had the same elution profiles as wild-type protein and were complexed in a dimeric form. Conversely, L240A, V260E and Y246W/ΔC25 exhibited different profiles. The elution peaks were smaller. The L240A profile exhibited a large and a small peak, which could correspond to a mix of dimers and monomers. The V260E profile exhibited two peaks consistent with dimer and higher-molecular forms, while the Y246W/ΔC25 elution profile exhibited three peaks which correspond to monomers, dimers and higher molecular size products (Fig. 4). However, regarding size of the peaks, we interpreted these two last mutants as being misfolded rather than structured as stable dimers and tetramers. The same interpretation has been made previously for the counterpart V260E mutation of HIV IN [39,40].

In protein–protein cross-linking experiments (Fig. 5), INs were incubated with the disuccinimidyl suberate (DSS) cross-linker. Reaction products were separated by SDS/PAGE and revealed by Western blot. As expected, in the absence of IN, we did not observed any product (Fig. 5, lane 1); in the absence of cross-linker, we observed only the monomeric form of IN (lane 2). With wild-type IN and in presence of DSS, we detected products at the expected molecular mass of integrase monomers and dimers (lane 3).

K225H (lane 4), V239A (lane 5), Y246W (lane 7), V249A (lane 8) and V257A (lane 9) mutants were observed as monomeric and dimeric forms in similar proportions to that of the wild-type protein (lane 3). On the contrary, L240A (lane 6), V260E (lane 10), K266A (lane 11) and Y246W/ΔC25 (lane 12) mutants were not cross-linked as efficiently as wild-type protein by DSS and the dimeric form was less represented for mutants than for the wild-type protein. These results confirm those from size exclusion chromatography analysis for L240A mutants. For V260E and Y246W/ΔC25, these analyses are in accordance with our hypothesis that these two mutants have a misfolded structure rather than being formed of stable dimers and tetramers. By contrast, the K266A mutant was able to form dimers as shown by size exclusion chromatography. However, DSS is reactive towards amino groups. Therefore, the most likely explanation is that the lysine to alanine mutation

renders the mutant unable to be cross-linked by DSS in this position, although it was associated as a dimer.

Discussion

The C-terminal domain of IN is able to bind DNA [27–29], is required for the 3′-processing and strand transfer activities of IN [25,26], and is essential for the formation of IN oligomers [30–38]. In this study, we analysed several points mutants in the C-terminal domain of ALSV IN and examined their ability to mediate the concerted DNA integration in an *in vitro* assay as well as to form dimers. Our analysis focused on mutations at the C-terminal dimer interface. Similar analyses have been performed on residues of the core domain [60].

In the concerted DNA integration assay, we could evaluate the ability of IN to catalyse the two-ended concerted DNA integration in two ways: (a) by quantifying the linear product b, since this product is supposed to be generated by a two-ended concerted DNA integration of two DNA donors [9,12–15]; and (b) by quantifying the number of colonies recovered after cloning of integration products into bacteria which allow selective amplification of two-ended circular integration products [**a** (concerted) and **d** (nonconcerted)]. The products **a** and **d** are subsequently distinguished by sequencing the integration products, and gross deletions of target DNA are assigned to the two-ended nonconcerted DNA integration (class **d**) [13–15]. In our experiments with wild-type IN, most products (87%) were of type **a** (without deletion of target DNA) (Table 1). Therefore, cloning of integration reactions into bacteria give a relevant estimation of the product **a** and, subsequently, of the two-ended concerted DNA integration events. According to these assays, if an IN mutant performed two-ended integration less efficiently than wild-type IN, we would expect a concomitant decrease both in the proportion of product **b** among the total integration products and in the number of recovered colonies from bacteria. Unexpectedly, we found that the quantity of product **b** did not systematically match the recovered number of colonies (Fig. 3B,C). This is particularly striking for mutant V239A which produced total integration products (RFII plus RFIII) in ratios as high as 170% that of wild-type, and the ratio of product **b** was found close to that of wild-type proteins (25 and 28% of product b, respectively). By contrast, the proportion of two-ended integration products amplified in bacteria was reduced to less than 30% that of wild-type IN. Such a discrepancy is also evident for the mutant K266A and, to a lesser extent, for mutant V249A. Similar observations have been made previously by others [6,13–15]. For example, the ability of a U5 mutated-donor DNA to undergo concerted DNA integration *in vitro* was 1.5–2-fold greater than observed with a wild-type donor substrate. This stimulation of integration concerned both the RFII (**a** + **c** + **d**) and RFIII products (**b**). However, when integrants were introduced into bacteria, the number of colonies recovered was reduced to 25% relative to the wild-type donor. Even more, a reduction to 4% was observed in the presence of HMGI despite an increase in the RF products on gels [13]. Altogether, these independent observations show that: (a) when the quantity of the total integration products increases, the quantity of product **b** increases in a similar proportion; (b) whereas, in the same reaction, the quantity of product **a** (and product **d**) may decrease in an independent manner. Therefore, discrepancies between gels and bacteria may be due to an increase in one-ended integration events (which are not amplified in bacteria) or to a specific decrease in two-ended integration events, or to both. Further, these observations strongly suggest that product **b** and product **a** are generated by different mechanisms. We propose that product **b** should be considered as the result of two non-independent events of one-ended DNA integration with two donors rather than the result of two-ended integration with two donors. Alternatively, product **b** could be a mix of several products: the expected product **b** and other products generated by non-concerted events of integration whose structures are unknown. Thus, to estimate the two-ended concerted DNA integration efficiency, quantification of product **b** on a gel would not be as stringent as quantification of product **a** by cloning and sequencing.

Data obtained for each C-terminal domain mutant studied here and in the previous study [24] are shown in Table 3.

We observed that V260E and Y246W/ΔC25 mutants were drastically misfolded and completely defective in the concerted DNA integration assay. In the case of the Y246W/ΔC25 mutant, this misfolding was most probably due to deletion of the last 25 residues of the C domain, as the single Y246W mutant was not so significantly impaired. The loss of strand β5′ could locally destabilize the C domain by disrupting intramolecular interactions with strand β1′ (Fig. 2B). Alternatively, this defect might be due to the combination of both the Y246W mutation and the deletion of the 25 terminal residues. Regarding the V260E mutation, it has been shown previously that mutant V260E in HIV-1 IN was mainly misfolded as well [40]. V260 is a highly conserved residue of strand β4′. The V260E mutation could prevent the formation of this strand as glutamate acts as a strand breaker [61]. Altogether, these data suggest a strong structural role for the terminal part of the C-terminal domain of ALSV integrase in the general folding of the enzyme and, hence, in its activity in the concerted DNA integration assay.

According to the structure proposed by Yang *et al.* [23], residues V239 and K266 of the proximal monomer and residues L240, Y246 and V257 of the distal monomer are directly involved in the C domain dimer interface (Table 2). Three mutations at this dimer interface (L240A, Y246W and K266A) do not remove contacts between monomers (Table 2). Accordingly, mutants Y246W and K266A were present exclusively in dimeric forms (Figs 4 and 5). However, and to our surprise, the L240A mutant had a reduced ability to form dimers. As mutating this residue reduces intramolecular interactions within the monomers (Table 2), it is possible that the conformation of the whole monomeric molecule is destabilized rendering the monomer unable to associate as dimers. Alternatively, it is noteworthy that this residue is well conserved among INs and that the homologue HIV IN residue (L242) has been involved in the formation of tetramers [40]. Therefore, it is possible that this residue is involved in other intermolecular interactions not seen in the dimeric structure proposed for ALSV IN. The two other mutations of residues at the dimer interface

(V239A and V257A) abrogate a contact between the two monomers (Table 2) but mutants were not impaired in dimer formation (Figs 4 and 5). For these last two mutants, it is possible that mutating these two residues was not sufficient by itself to impair the formation of the dimer.

All the mutations of residues at the dimer interface caused a decrease in the concerted DNA integration process (Fig. 3; Table 3). For the Y246W and V257A mutants, this decrease in concerted DNA integration is most probably due to a strong defect in 3′-processing activity (Table 3). It is possible that these mutations induce local conformational changes in the region of the β3′ strand rendering the molecule less efficient in 3′-processing.

The L240A mutant is less efficient than wild-type IN in performing all types of integration events (one- and two-ended, concerted and not) as revealed on gels and in bacteria. We speculate that the decrease in integration efficiency is directly related to the decrease in the proportion of dimers that this mutant is able to form.

We observed that mutations V239A, K266A and, to a lesser extent, V249A affected the two-ended integration process specifically, but did not reduce (K266 and V249A), and even enhanced (V239A) the one-ended integration process (Fig. 3, Table 3). However, when mutants catalysed concerted DNA integration, it was performed correctly as revealed by sequencing of integration products. These observations suggest that a few molecules are able to assemble in a complex competent in performing concerted DNA integration and that most molecules performed one-ended non-concerted DNA integrations rather than two-ended concerted DNA integration. Many data have suggested that at least a tetramer is necessary to catalyse concerted DNA integration [23,33,35–38,62]. DNAse protection studies [8] have even suggested that a dimer of ALSV IN is required for the one-ended donor insertion reaction and that for two-ended donor concerted DNA integration a tetramer is assembled on the ALSV U5 end and a higher-order multimer forms on the ALSV U3 end. Therefore, as we observed that K266A, V239A and V249A were less efficient at performing two-ended concerted DNA integration in the absence of alterations in dimer formation of IN (Figs 4 and 5), it is tempting to speculate that mutations of the K266, V239 and V249 residues might prevent the formation of a higher molecular size complex such as a tetramer. It is possible that these mutations either induce local conformational changes that prevent the formation of a tetramer or have distal effects in the protein affecting its global structure. Alternatively, these residues could be directly involved in the formation of the tetramer. Accordingly, the HIV L241A IN mutant (L241 of HIV IN is homologous to V239 of ALSV IN), has been shown to be unable to form tetramers [40]. Furthermore, in the tetrameric model of HIV-1, residue L241 is located at the interface between dimers [59]. Unfortunately, this mutant has not been yet tested in the concerted DNA integration assay. However, it is tempting to speculate that the distal V239, which is accessible and is located away from the dimeric interface (Fig. 2), could be part of the putative tetrameric interface in the ALSV IN.

Our results provide news insights into the multiple structure–function relationships of IN for concerted DNA integration. They show a strong structural role of the most C-terminal part of this C-terminal domain in the general folding of the enzyme. They reinforce the role of the IN dimers, as a mutant deficient in dimerization is similarly deficient in concerted DNA integration. Even more, they predict that high-order IN complexes are required to perform two-ended concerted DNA integration. Finally, they confirm the importance of residues within the C-terminal domain dimer interface in concerted DNA integration. This part of the protein may constitute a new target for the development of antiviral drugs against integrases.

Acknowledgements

This work was supported by research grants from the Centre National de la Recherche Scientifique and the Institut National de la Recherche Agronomique. We acknowledge the French Ministry of Research and the Agence Nationale de Recherche contre le SIDA (ANRS) for fellowships (K.M. and S.V.). We thank Dr T.H. Kim (Cambridge) for providing the pET15b-HMGI plasmid. Special thanks to Dr S. Carteau for helpful discussions and to Dr E. Derrigton for helpful discussions and for correcting the English. We also thank Dr P. Gouet for valuable scientific support and Pr J. L. Darlix for critical comment on the manuscript. Thanks are also due to Dr S. Arnaud and M.-F. Grasset (Dr G. Mouchiroud's laboratory) for their help with size exclusion chromatography. We gratefully acknowledge Pr P. Boulanger for putting his laboratory at our disposal for some parts of this work.

References

1. Brown, P.O. (1997) Integration. *Retroviruses* (Coffin, J.M., Hugues, S.H. & Varmus, H.E., eds), pp. 161–204. Cold Spring. Harbor Laboratory Press, Cold spring Harbor, New York.
2. Daniel, R., Katz, R.A. & Skalka, A.M. (1999) A role for DNA-PK in retroviral DNA integration. *Science* **284**, 644–647.
3. Daniel, R., Kao, G., Taganov, K., Greger, J.G., Favorova, O., Merkel, G., Yen, T.J., Katz, R.A. & Skalka, A.M. (2003) Evidence that the retroviral DNA integration process triggers an ATR-dependent DNA damage response. *Proc. Natl Acad. Sci. USA* **100**, 4778–4783.
4. Gaken, J.A., Tavassoli, M., Gan, S.U., Vallian, S., Giddings, I., Darling, D.C., Galea-Lauri, J., Thomas, M.G., Abedi, H., Schreiber, V., Menissier-de Murcia, J., Collins, M.K., Shall, S. & Farzaneh, F. (1996) Efficient retroviral infection of mammalian cells is blocked by inhibition of poly (ADP-ribose) polymerase activity. *J. Virol.* **70**, 3992–4000.
5. Siva, A.C. & Bushman, F. (2002) Poly (ADP-ribose) polymerase 1 is not strictly required for infection of murine cells by retroviruses. *J. Virol.* **76**, 11904–11910.
6. Vora, A.C., Chiu, R., McCord, M., Goodarzi, G., Stahl, S.J., Mueser, T.C., Hyde, C.C. & Grandgenett, D.P. (1997) Avian retrovirus U3 and U5 DNA inverted repeats. Role of non-symmetrical nucleotides in promoting full-site integration by purified virion and bacterial recombinant integrases. *J. Biol. Chem.* **272**, 23938–23945.
7. Vora, A.C., McCord, M., Fitzgerald, M.L., Inman, R.B. & Grandgenett, D.P. (1994) Efficient concerted integration of retrovirus-like DNA *in vitro* by avian myeloblastosis virus integrase. *Nucleic Acids Res.* **22**, 4454–4461.
8. Vora, A. & Grandgenett, D.P. (2001) DNase protection analysis of retrovirus integrase at the viral DNA ends for full-site integration *in vitro*. *J. Virol.* **75**, 3556–3567.
9. Aiyar, A., Hindmarsh, P., Skalka, A.M. & Leis, J. (1996) Concerted integration of linear retroviral DNA by the avian sarcoma virus integrase *in vitro*: dependence on both long terminal repeat termini. *J. Virol.* **70**, 3571–3580.

10. Chiu, R. & Grandgenett, D.P. (2000) Avian retrovirus DNA internal attachment site requirements for full- site integration *in vitro*. *J. Virol.* **74**, 8292–8298.
11. Chiu, R. & Grandgenett, D.P. (2003) Molecular and genetic determinants of rous sarcoma virus integrase for concerted DNA integration. *J. Virol.* **77**, 6482–6492.
12. Hindmarsh, P., Ridky, T., Reeves, R., Andrake, M., Skalka, A.M. & Leis, J. (1999) HMG protein family members stimulate human immunodeficiency virus type 1 and avian sarcoma virus concerted DNA integration *in vitro*. *J. Virol.* **73**, 2994–3003.
13. Hindmarsh, P., Johnson, M., Reeves, R. & Leis, J. (2001) Base-pair substitutions in avian sarcoma virus U5 and U3 long terminal repeat sequences alter the process of DNA integration *in vitro*. *J. Virol.* **75**, 1132–1141.
14. Brin, E. & Leis, J. (2002a) Changes in the mechanism of DNA integration *in vitro* induced by base substitutions in the HIV-1 U5 and U3 terminal sequences. *J. Biol. Chem.* **277**, 10938–10948.
15. Brin, E. & Leis, J. (2002b) HIV−1 integrase interaction with U3 and U5 terminal sequences *in vitro* defined using substrates with random sequences. *J. Biol. Chem.* **15**, 15.
16. Carteau, S., Gorelick, R.J. & Bushman, F.D. (1999) Coupled integration of human immunodeficiency virus type 1 cDNA ends by purified integrase *in vitro*: stimulation by the viral nucleocapsid protein. *J. Virol.* **73**, 6670–6679.
17. Gao, K., Gorelick, R.J., Johnson, D.G. & Bushman, F. (2003) Cofactors for human immunodeficiency virus type 1 cDNA integration *in vitro*. *J. Virol.* **77**, 1598–1603.
18. Goodarzi, G., Im, G.J., Brackmann, K. & Grandgenett, D. (1995) Concerted integration of retrovirus-like DNA by human immunodeficiency virus type 1 integrase. *J. Virol.* **69**, 6090–6097.
19. Sinha, S., Pursley, M.H. & Grandgenett, D.P. (2002) Efficient concerted integration by recombinant human immunodeficiency virus type 1 integrase without cellular or viral cofactors. *J. Virol.* **76**, 3105–3113.
20. Goodarzi, G., Pursley, M., Felock, P., Witmer, M., Hazuda, D., Brackmann, K. & Grandgenett, D. (1999) Efficiency and fidelity of full-site integration reactions using recombinant simian immunodeficiency virus integrase. *J. Virol.* **73**, 8104–8111.
21. Yang, F. & Roth, M.J. (2001) Assembly and catalysis of concerted two-end integration events by Moloney murine leukemia virus integrase. *J. Virol.* **75**, 9561–9570.
22. Bujacz, G., Jaskolski, M., Alexandratos, J., Wlodawer, A., Merkel, G., Katz, R.A. & Skalka, A.M. (1995) High-resolution structure of the catalytic domain of avian sarcoma virus integrase. *J. Mol. Biol.* **253**, 333–346.
23. Yang, Z.N., Mueser, T.C., Bushman, F.D. & Hyde, C.C. (2000) Crystal structure of an active two-domain derivative of Rous sarcoma virus integrase. *J. Mol. Biol.* **296**, 535–548.
24. Moreau, K., Faure, C., Verdier, G. & Ronfort, C. (2002) Analysis of conserved and non-conserved amino acids critical for ALSV (Avian leukemia and sarcoma viruses) integrase functions *in vitro*. *Arch. Virol.* **147**, 1761–1778.
25. Engelman, A., Bushman, F.D. & Craigie, R. (1993) Identification of discrete functional domains of HIV-1 integrase and their organization within an active multimeric complex. *EMBO J.* **12**, 3269–3275.
26. van Gent, D.C., Vink, C., Groeneger, A.A. & Plasterk, R.H. (1993) Complementation between HIV integrase proteins mutated in different domains. *EMBO J.* **12**, 3261–3267.
27. Esposito, D. & Craigie, R. (1998) Sequence specificity of viral end DNA binding by HIV-1 integrase reveals critical regions for protein–DNA interaction. *EMBO J.* **17**, 5832–5843.
28. Lutzke, R.A., Vink, C. & Plasterk, R.H. (1994) Characterization of the minimal DNA-binding domain of the HIV integrase protein. *Nucleic Acids Res.* **22**, 4125–4131.
29. Mumm, S.R. & Grandgenett, D.P. (1991) Defining nucleic acid-binding properties of avian retrovirus integrase by deletion analysis. *J. Virol.* **65**, 1160–1167.
30. Jones, K.S., Coleman, J., Merkel, G.W., Laue, T.M. & Skalka, A.M. (1992) Retroviral integrase functions as a multimer and can turn over catalytically. *J. Biol. Chem.* **267**, 16037–16040.
31. Andrake, M.D. & Skalka, A.M. (1995) Multimerization determinants reside in both the catalytic core and C terminus of avian sarcoma virus integrase. *J. Biol. Chem.* **270**, 29299–29306.
32. Coleman, J., Eaton, S., Merkel, G., Skalka, A.M. & Laue, T. (1999) Characterization of the self association of Avian sarcoma virus integrase by analytical ultracentrifugation. *J. Biol. Chem.* **274**, 32842–32846.
33. Bao, K.K., Wang, H., Miller, J.K., Erie, D.A., Skalka, A.M. & Wong, I. (2003) Functional oligomeric state of avian sarcoma virus integrase. *J. Biol. Chem.* **278**, 1323–1327.
34. Jenkins, T.M., Engelman, A., Ghirlando, R. & Craigie, R. (1996) A soluble active mutant of HIV-1 integrase: involvement of both the core and carboxyl-terminal domains in multimerization. *J. Biol. Chem.* **271**, 7712–7718.
35. Lee, S.P., Xiao, J., Knutson, J.R., Lewis, M.S. & Han, M.K. (1997) Zn2+ promotes the self association of human immunodeficiency virus type-1 integrase *in vitro*. *Biochemistry* **36**, 173–180.
36. Deprez, E., Tauc, P., Leh, H., Mouscadet, J.F., Auclair, C. & Brochon, J.C. (2000) Oligomeric states of the HIV-1 integrase as measured by time-resolved fluorescence anisotropy. *Biochemistry* **39**, 9275–9284.
37. Vercammen, J., Maertens, G., Gerard, M., De Clercq, E., Debyser, Z. & Engelborghs, Y. (2002) DNA-induced polymerization of HIV-1 integrase analyzed with fluorescence fluctuation spectroscopy. *J. Biol. Chem.* **277**, 38045–38052.
38. Cherepanov, P., Maertens, G., Proost, P., Devreese, B., Van Beeumen, J., Engelborghs, Y., De Clercq, E. & Debyser, Z. (2003) HIV-1 integrase forms stable tetramers and associates with LEDGF/p75 protein in human cells. *J. Biol. Chem.* **278**, 372–381.
39. Kalpana, G.V., Reicin, A., Cheng, G.S., Sorin, M., Paik, S. & Goff, S.P. (1999) Isolation and characterization of an oligomerization-negative mutant of HIV-1 integrase. *Virology* **259**, 274–285.
40. Puras Lutzke, R.A. & Plasterk, R.H. (1998) Structure-based mutational analysis of the C-terminal DNA-binding domain of human immunodeficiency virus type 1 integrase: critical residues for protein oligomerization and DNA binding. *J. Virol.* **72**, 4841–4848.
41. Eijkelenboom, A.P., Lutzke, R.A., Boelens, R., Plasterk, R.H., Kaptein, R. & Hard, K. (1995) The DNA-binding domain of HIV-1 integrase has an SH3-like fold. *Nat. Struct. Biol.* **2**, 807–810.
42. Fletcher, T.M.R., Soares, M.A., McPhearson, S., Hui, H., Wiskerchen, M., Muesing, M.A., Shaw, G.M., Leavitt, A.D., Boeke, J.D. & Hahn, B.H. (1997) Complementation of integrase function in HIV-1 virions. *EMBO J.* **16**, 5123–5138.
43. Holmes-Son, M.L. & Chow, S.A. (2000) Integrase-lexA fusion proteins incorporated into human immunodeficiency virus type 1 that contains a catalytically inactive integrase gene are functional to mediate integration. *J. Virol.* **74**, 11548–11556.
44. Heuer, T.S. & Brown, P.O. (1998) Photo-cross-linking studies suggest a model for the architecture of an active human immunodeficiency virus type 1 integrase-DNA complex. *Biochemistry* **37**, 6667–6678.
45. Esnouf, R. (1997) Polyalanine reconstruction from Calpha positions using the program CALPHA can aid initial phasing of data by molecular replacement procedures. *Acta Crystallogr. D Biol. Crystallogr.* **53**, 665–672.

46. Brunger, A.T., Adams, P.D., Clore, G.M., DeLano, W.L., Gros, P., Grosse-Kunstleve, R.W., Jiang, J.S., Kuszewski, J., Nilges, M., Pannu, N.S., Read, R.J., Rice, L.M., Simonson, T. & Warren, G.L. (1998) Crystallography & NMR system: a new software suite for macromolecular structure determination. *Acta Crystallogr. D Biol. Crystallogr.* **54**, 905–921.
47. Roussel, A. & Cambillau, C. (1989) TURBO-FRODO. *Silicon Graphics Geometry Partner Directory* (Graphics, S., ed.). Silicon Graphics, Moutain View, CA.
48. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl Acids Res.* **25**, 3389–3402.
49. Bairoch, A. & Apweiler, R. (2000) The SWISS-PROT protein sequence database and its (Suppl.)TrEMBL in 2000. *Nucleic Acids Res.* **28**, 45–48.
50. Thompson, J.D., Higgins, D.G. & Gibson, T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**, 4673–4680.
51. Thanos, D. & Maniatis, T. (1992) The high mobility group protein HMG I (Y) is required for NF-kappa B-dependent virus induction of the human IFN-beta gene. *Cell* **71**, 777–789.
52. Farnet, C.M. & Bushman, F.D. (1997) HIV-1 cDNA integration: requirement of HMG I (Y) protein for function of preintegration complexes *in vitro*. *Cell* **88**, 483–492.
53. Li, L., Yoder, K., Hansen, M.S., Olvera, J., Miller, M.D. & Bushman, F.D. (2000) Retroviral cDNA integration: stimulation by HMG I family proteins. *J. Virol.* **74**, 10965–10974.
54. Hughes, S.H., Mutschler, A., Bishop, J.M. & Varmus, H.E. (1981) A Rous sarcoma virus provirus is flanked by short direct repeats of a cellular DNA sequence present in only one copy prior to integration. *Proc. Natl Acad. Sci. USA* **78**, 4299–4303.
55. Ju, G., Boone, L. & Skalka, A.M. (1980) Isolation and characterization of recombinant DNA clones of avian retroviruses: size heterogeneity and instability of the direct repeat. *J. Virol.* **33**, 1026–1033.
56. Moreau, K., Torne-Celer, C., Faure, C., Verdier, G. & Ronfort, C. (2000) In vivo retroviral integration: fidelity to size of the host DNA duplication might be reduced when integration occurs near sequences homologous to LTR ends. *Virology*. **278**, 133–136.
57. Chen, J.C., Krucinski, J., Miercke, L.J., Finer-Moore, J.S., Tang, A.H., Leavitt, A.D. & Stroud, R.M. (2000) Crystal structure of the HIV-1 integrase catalytic core and C-terminal domains: a model for viral DNA binding. *Proc. Natl Acad. Sci. USA* **97**, 8233–8238.
58. Chen, Z., Yan, Y., Munshi, S., Li, Y., Zugay-Murphy, J., Xu, B., Witmer, M., Felock, P., Wolfe, A., Sardana, V., Emini, E.A., Hazuda, D. & Kuo, L.C. (2000) X-ray structure of simian immunodeficiency virus integrase containing the core and C-terminal domain (residues 50–293) -an initial glance of the viral DNA binding platform. *J. Mol. Biol.* **296**, 521–533.
59. Wang, J.Y., Ling, H., Yang, W. & Craigie, R. (2001) Structure of a two-domain fragment of HIV-1 integrase: implications for domain organization in the intact protein. *EMBO J.* **20**, 7333–7343.
60. Moreau, K., Faure, C., Violot, S., Gouet, P., Verdier, G. & Ronfort, C. (2003) Mutational analyses of the core domain of Avian Leukemia and Sarcoma Viruses integrase: critical residues for concerted integration and multimerization. *Virology*, in press.
61. Chou, P.Y. & Fasman, G.D. (1979) Prediction of beta-turns. *Biophys. J.* **26**, 367–373.
62. Deprez, E., Tauc, P., Leh, H., Mouscadet, J.F., Auclair, C., Hawkins, M.E. & Brochon, J.C. (2001) DNA binding induces dissociation of the multimeric form of HIV-1 integrase: a time-resolved fluorescence anisotropy study. *Proc. Natl Acad. Sci. USA* **98**, 10090–10095.

Publication 3

Mutational analyses of the core domain of the Avian Leukemia and Sarcoma Viruses integrase : critical residues for concerted integration and multimerization

Karen Moreau, Claudine Faure, Sébastien Violot, Patrice Gouet, Gérard Verdier and Corinne Ronfort

Available online at www.sciencedirect.com
SCIENCE DIRECT®
Virology 318 (2004) 566–581
VIROLOGY
www.elsevier.com/locate/yviro

Mutational analyses of the core domain of Avian Leukemia and Sarcoma Viruses integrase: critical residues for concerted integration and multimerization

Karen Moreau,[a] Claudine Faure,[a] Sébastien Violot,[b,c] Patrice Gouet,[b,c] Gérard Verdier,[a] and Corinne Ronfort[a,c,*]

[a] *Centre National de la Recherche Scientifique, Institut National de la Recherche Agronomique, Université Claude Bernard, Lyon, France*
[b] *Laboratoire de Bio-Cristallographie, Centre National de la Recherche Scientifique, Institut de Biologie et Chimie des Protéines, Université Claude Bernard, Lyon, France*
[c] *IFR128 "Biosciences Lyon Gerland", Lyon, France*

Received 8 August 2003; returned to author for revision 25 September 2003; accepted 25 September 2003

Abstract

During replicative cycle of retroviruses, the reverse-transcribed viral DNA is integrated into the cell DNA by the viral integrase (IN) enzyme. The central core domain of IN contains the catalytic site of the enzyme and is involved in binding viral ends and cell DNA as well as dimerization. We previously performed single amino acid substitutions in the core domain of an Avian Leukemia and Sarcoma Virus (ALSV) IN [Arch. Virol. 147 (2002) 1761]. Here, we modeled the resulting IN mutants and analyzed the ability of these mutants to mediate concerted DNA integration in an in vitro assay, and to form dimers by protein–protein cross-linking and size exclusion chromatography. The N197C mutation resulted in the inability of the mutant to perform concerted integration that was concomitant with a loss of IN dimerization. Surprisingly, mutations Q102G and A106V at the dimer interface resulted in mutants with higher efficiencies than the wild-type IN in performing two-ended concerted integration of viral DNA ends. The G139D and A195V mutants had a trend to perform one-ended DNA integration of viral ends instead of two-ended integration. More drastically, the I88L and L135G mutants preferentially mediated nonconcerted DNA integration although the proteins form dimers. Therefore, these mutations may alter the formation of IN complexes of higher molecular size than a dimer that would be required for concerted integration. This study points to the important role of core domain residues in the concerted integration of viral DNA ends as well as in the oligomerization of the enzyme.

Keywords: ALSV; Retrovirus; Integrase; Concerted integration

Introduction

During the life cycle of retroviruses, the reverse-transcribed viral DNA is integrated into the host cell DNA by viral integrase protein (IN). The integration step is essential for the replicative cycle of retroviruses. The integration process can be subdivided into three steps: (i) the 3′-processing during which the two 3′ terminal nucleotides of each viral end are cleaved creating the characteristic CA-OH 3′ ends; (ii) the strand transfer or the joining of 3′ viral ends to host DNA, and finally (iii) the gap filling which allows the reparation of single-stranded DNA gaps and the ligation of 5′ viral ends to host DNA. The first two steps require only two viral elements which are the specific sequences at viral ends designated as *att* sequences and the viral integrase protein (Brown, 1997). By contrast, the gap-filling step may involve cellular enzymes (Daniel et al., 1999, 2003; Ha et al., 2001; Yoder and Bushman, 2000). The integration process is designated as "concerted integration" because it allows the coordinated integration of both viral ends at the same site of host DNA. As the result of this process, the integrated provirus is invariably flanked by the canonical 5′ TG. . .CA 3′ viral ends and the duplica-

* Corresponding author. Present address: Laboratoire "Rétrovirus et Pathologie Comparée", UCBL-INRA-ENVL, Université Claude Bernard, 50, avenue Tony Garnier, 69366, Lyon cedex 07, France. Fax: +33-437287605.
E-mail address: ronfort@univ-lyon1.fr (C. Ronfort).

0042-6822/$ - see front matter © 2003 Elsevier Inc. All rights reserved.
doi:10.1016/j.virol.2003.09.037

tion of a short host DNA sequence (4–6 bp) whose size is virus specific (Brown, 1997).

The retroviral integrase protein contains three domains, all necessary for its activity (Engelman et al., 1993; van Gent et al., 1993). Among them, the central core domain is the most highly conserved (Bujacz et al., 1995; Khan et al., 1991; Moreau et al., 2002). It contains a highly conserved motif composed of two aspartic acids and one glutamic acid, with a second aspartic acid separated from glutamic acid invariably by 35 residues. This D, D (35) E motif constitutes the active site of the protein. Mutation of one of these residues drastically inhibits catalytic activities of 3′-processing, strand transfer, and disintegration (reverse reaction of strand transfer) of INs tested in vitro (Engelman and Craigie, 1992; van Gent et al., 1992), as well as infectivity of viruses in vivo (Engelman, 1999). In addition to catalytic activities, the core domain is involved in recognition of viral ends. Indeed, using chimeric IN proteins, it has been shown that the core domain together with the C-terminal domain is responsible for *att* sequence-specific recognition during 3′-processing and strand transfer experiments in vitro (Berger et al., 2001). Furthermore, photo cross-linking experiments have demonstrated that the two terminal nucleotides and the invariable CA dinucleotide of the viral ends interact with residues of the core domain of Human Immunodeficiency Virus (HIV) IN protein (Esposito and Craigie, 1998; Gerton et al., 1998) and that two lysine residues within the core domain are critical for interaction of integrase with viral DNA (Jenkins et al., 1997). Finally, the core domain is also involved in binding host DNA because it influences the choice of integration site as observed in in vitro assays with chimeric enzymes (Appa et al., 2001; Katzman and Sudol, 1995). The crystallographic structure of the core domain has been determined for both Avian Leukemia and Sarcoma Viruses (ALSV) and HIV proteins. Both IN core domains are dimeric and are composed of 5 β strands and either six (HIV IN) or five (ALSV IN) α helices (Bujacz et al., 1995, 1996; Chen et al., 2000; Dyda et al., 1994; Goldgur et al., 1998; Lubkowski et al., 1998a, 1999; Wang et al., 2001; Yang et al., 2000).

We previously introduced 11 single amino acid substitutions in the ALSV Rous Associated Virus type 1 (RAV-1) IN core domain and analyzed the 3′ processing, strand transfer, and disintegration catalytic activities of the resulting mutants (Moreau et al., 2002). In the present study, we examined the effect of these mutations on the concerted integration process of two viral ends using an in vitro integration assay. We also examined the oligomeric state of the resulting proteins. Furthermore, we used the two-domain structure of Rous Sarcoma Virus (RSV) IN (Yang et al., 2000) to model the structure of core mutants analyzed here. Our analyses focused on residues that were at or close to either the dimer interface or the catalytic site, as well as on residues conserved among integrases. These analyzes allow us to identify specific residues within the core domain important for concerted integration and multimerization of IN.

Results

Reconstitution of the concerted DNA integration assay in vitro

The in vitro retroviral concerted integration assay has been previously described by others for ALSV (Aiyar et al., 1996; Chiu and Grandgenett, 2000, 2003; Hindmarsh et al., 1999, 2001; Vora and Grandgenett, 1995, 2001; Vora et al., 1994, 1997) and other retroviruses such as HIV and Simian Immunodeficiency Virus (SIV) (Brin and Leis, 2002a, 2002b; Carteau et al., 1999; Gao et al., 2003; Goodarzi et al., 1995, 1999; Sinha et al., 2002) and Murine Leukemia Virus (MLV) (Yang and Roth, 2001). It is composed of a linear donor DNA, a plasmid acceptor DNA, and recombinant IN. High Mobility Group I protein (HMGI) [now referred as HMGa1] is added to the reaction because it has been found to enhance the concerted integration reaction (Hindmarsh et al., 1999). In the present report, we used a donor DNA of 326 bp containing 15 bp of the terminal U3 *att* sequence at one end and 12 bp of the U5 *att* sequence at the other end (Fig. 1A).

Products of the integration reaction can arise from concerted or nonconcerted integration processes (Fig. 1B) (Aiyar et al., 1996; Brin and Leis, 2002a, 2002b; Carteau et al., 1999; Goodarzi et al., 1995; Hindmarsh et al., 1999, 2001; Vora and Grandgenett, 1995, 2001; Vora et al., 1994, 1997). Concerted integration products include those that result from two-ended integration of both viral ends from a single donor (product *a*) or from two one-ended integration of two viral ends from two donors at the same integration site (generating the linear product *b*). Nonconcerted integration products result from one-ended donor integration of a single donor (product *c*), from two-ended integration of a single donor with insertion at different sites on the acceptor DNA (product *d*), or from one-ended donor integration of two or more donors at different sites on the acceptor DNA (product *e*). Auto-integration products, which are the result of the integration of donor DNA in a second donor DNA, are also obtained (product *f*). By using labeled donor DNA, the integration products were separated on agarose gel and visualized by autoradiography, and three characteristic bands were revealed (see Fig. 3A, lane 6). As previously described by others (Carteau et al., 1999; Vora and Grandgenett, 2001; Vora et al., 1994), the slowest band corresponds to a mix of circular forms (RFII products: *a*, *c* and *d*), the middle band corresponds to the linear *b* form (RFIII products), and the fastest band corresponds to autointegration products (form *f*). Product *e*, which migrates more slowly because two or more donors are inserted into the target, is observed on some gels, but not all. A recombinant which migrated slightly faster than RFII recombinants (identified with an asterisk in Fig. 3A) has been observed by others (Brin and Leis, 2002a,

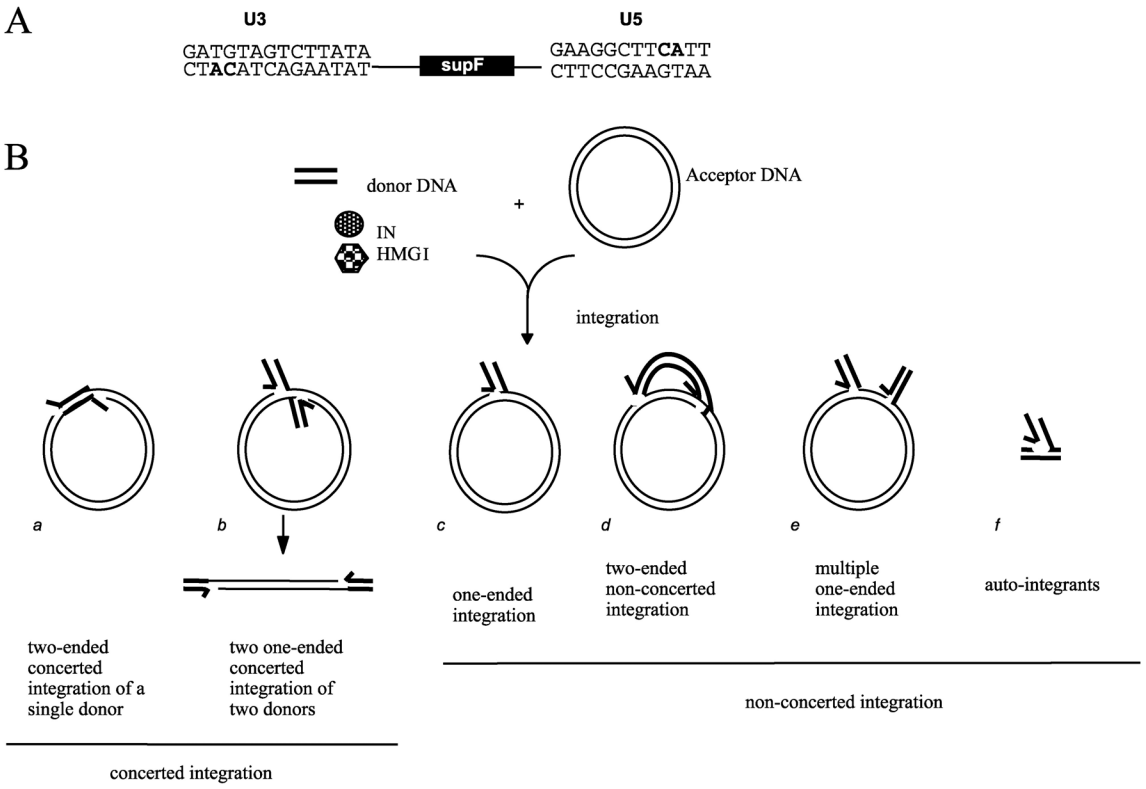


Fig. 1. Principle of the concerted integration assay. (A) Representation of the donor DNA. It contains 15 bp of the U3 viral end and 12 bp of the viral U5 end. The highly conserved CA dinucleotides are in bold. The closed rectangle represents the *sup*F tRNA transcription unit. (B) Schematic representation of the reconstituted integration reaction with the donor DNA, acceptor plasmid, purified integrase, and HMGI proteins. Concerted integration products include those that result from use of both ends from a single donor (product *a*) and from use of different ends from two donors (product *b*). Note that when two donors are inserted at the same site, a linear product is synthesized. Nonconcerted integration products result from one-ended integration of a single donor (product *c*), or two-ended integration of a single donor with insertion at different sites on the acceptor DNA (product *d*), or one-ended integration of two donors at different sites on the acceptor DNA (product *e*). Autointegrants result from integration of a donor DNA in a second donor DNA (product *f*). Adapted from Hindmarsh et al. (2001).

2002b; Carteau et al., 1999; Chiu and Grandgenett, 2000; Goodarzi et al., 1995; Sinha et al., 2002; Vora et al., 1997); its structure is unknown (Goodarzi et al., 1995). As a control, the reaction was performed without IN: no integration products were observed (data not shown), leading to the conclusion that these bands corresponded to integration products and resulted from IN enzymatic activity. Integration products were cleaved with either *Bam*HI (which cleaves in the donor DNA) or *Xho*I enzymes (which cleaves in the acceptor DNA). Structures of digestion products were fully consistent with assignment of the DNA forms (data not shown).

We also cloned the integration products into MC1061/P3 *Escherichia coli* bacteria. These bacteria contain drug-resistance markers with amber mutations. Only DNA products carrying the amber mutation suppressor gene (*sup*F) should be able to replicate and form colonies under drug selection. Among the different integration products, one-ended or multiple one-ended donor integration products (*c* and *e*) and linear products (*b*) should be lost upon cloning into *E. coli*. Only circular two-ended forms of integration products (forms *a* and *d*) should be able to replicate (Brin and Leis, 2002a, 2002b). Thus, the cloning analyze enables estimation of the efficiency of IN proteins to mediate the two-ended integration process (concerted (form *a*) or not (form *d*)). Between 98 and 324 resistant colonies were obtained according to the experiment with the wild-type IN. Following cloning, isolated integration products were analyzed by sequencing. Donor DNA–acceptor plasmid junctions were sequenced to check the accuracy of the integration reaction (cleavage of viral ends and duplication of short acceptor DNA sequence). In the present study, 19 clones obtained with the wild-type protein were sequenced (Table 1). Ten clones exhibited a duplication of 6 bp and eight clones a duplication of different size (from 4 to 7 bp). In vivo, the 6-bp duplication is a hallmark of ALSV viruses (Ju et al., 1980) although some variations have been observed (Moreau et al., 2000). In vitro, duplications of other size than 6 bp have previously been observed (Aiyar et al., 1996; Hindmarsh et al., 1999). One clone exhibited a deletion of

Table 1
Sequence analyses of donor–acceptor junction sites produced by wild-type IN protein

Protein	Characteristics	Number of recombinants
Wt	Duplications of 6 bp	10
	Other duplications (4–7 bp)	8
	Deletion in acceptor DNA (150 bp)	1
	Total	19
	Incorrect cleavage of *att* sequence[a]	3

[a] Among all clone studies, few of them were deleted of more than 2 bp at one or the other *att* viral end.

acceptor DNA. However, as this clone was correctly cleaved of two nucleotides at both viral ends and integrated between the canonical TG and CA dinucleotides, it was interpreted as the result of an integration process but with incorrect cleavage of the acceptor DNA. Such integration products with acceptor DNA deletion could arise from either two independent one-ended donor integration events (form *e*) or from nonconcerted integration of the two ends of one donor DNA (form *d*). Assuming that only circular integration products (*a* and *d*, Fig. 1) are amplified in bacteria, we speculate that this clone was most probably the result of a nonconcerted integration of the two viral ends at two different sites on acceptor plasmid DNA (form *d*, Fig. 1). Such structures have been previously described by others (Brin and Leis, 2002a, 2002b; Hindmarsh et al., 1999, 2001). Finally, regarding viral DNA ends, we observed deletion of more than two expected nucleotides at one or the other *att* sequence in three clones.

Description of IN mutants

We previously constructed mutants, each containing single amino acid substitutions in the core of the RAV-1 IN (Moreau et al., 2002) (Fig. 2). Meanwhile, a 2.5-Å structure of the closely related RSV (Rous Sarcoma Virus) IN has been published (Yang et al., 2000) containing both the core and the C-terminal domains. In this structure, the two core domains are related by a 2-fold symmetry axes, whereas C-terminal domains have a similar fold but associate asymmetrically. The core domain of ALSV INs consists of five stranded mixed β sheet flanked by five α helices (Fig. 2) (Bujacz et al., 1995; Yang et al., 2000). Interface of the dimer is created between the α1 helix in one monomer and the α5 helix in the complementary molecule. Further, residues from the β3 strand, which is located deeper within the monomer, contribute to the middle part of the dimer interface (Bujacz et al., 1995; Yang et al., 2000).

The 11 single mutants studied here can be separated into three structural groups: (1) Q102G, A106V, G186P, A195V, and N197C were mutated on residues at the dimeric molecular interface or close to it; (2) D121E, G123P, and F126I were mutated on residues at or close to the catalytic site; (3) and three other mutants I88L, L135G, and G139D were studied as well (Fig. 2 and Table 2).

(1) Q102 is in helix α1, which is in contact with helix α5 of the symmetry related monomer at the dimeric interface (Fig. 2B). The side chain of Q102 makes strong hydrogen bonds with residues S130 and T131 of the same monomer (Table 2). The Q102 residue has an intermolecular contact with E187 of the other monomer. The mutation Q102G cancels this network of interactions.
A106 is also located at the surface of helix α1 at the dimeric interface. Its side chain makes Van der Waals contacts with residue E187 and A190 of the symmetry related monomer (Table 2). The A106V mutation creates a new contact in the dimer (with K191) (Table 2). A valine can well be accommodated at this site (for instance, a leucine is found in equivalent position in HIV IN).
G186 and A195 are part of helix α5, which is in contact with helix α1 of the symmetry related monomer. Unlike Q102 and A106, these two residues do not face helix α1 (Fig. 2). The two mutations G186P and A195V create new contacts inside the monomer (with Q185 and H103 in mutant G186P and R166 in mutant A195V). Note that proline is known as helix breaker. In consequence, helix α5 may start at residue 187 in the mutant G186P, and not at residue 182 as in the wild type.
N197 is located at the end of helix α5, but it is not involved in intermolecular contacts in the dimer (Table 2). This asparagine is strictly conserved in integrases and, hence, is supposed to participate in important IN function. This residue displays an alternate conformation in the high-resolutions structures of ASVIN_AS and of ASVIN_HEP (Lubkowski et al., 1999). The mutation N197C reduces the number of contacts inside the monomer.
(2) D121 is one of the three fundamental acidic residues of the catalytic site, D64 and E157 being the others (Bujacz et al., 1995). In structures, a magnesium ion is present in the catalytic site and interacts with the carboxylate groups of the fundamental aspartates D64 and D121. Four water molecules complete the octahedral coordination (Bujacz et al., 1996). The mutation D121E removes a contact within the monomer (with F126) and prevents the correct coordination of the divalent cation.
G123 is strictly conserved in all retroviruses bearing the catalytic triad DDE. Main chain angles (φ, ψ) of G123 are characteristic of an allowed β-strand conformation. Thus, the mutant G123P may retain the overall fold of ALSV IN. However, the main chain nitrogen atom N of G123 establishes a strong hydrogen bond with the carboxylic group of the fundamental residue D121 (note that such contacts with the carbon skeleton are not mentioned in the table). This liaison can be determinant

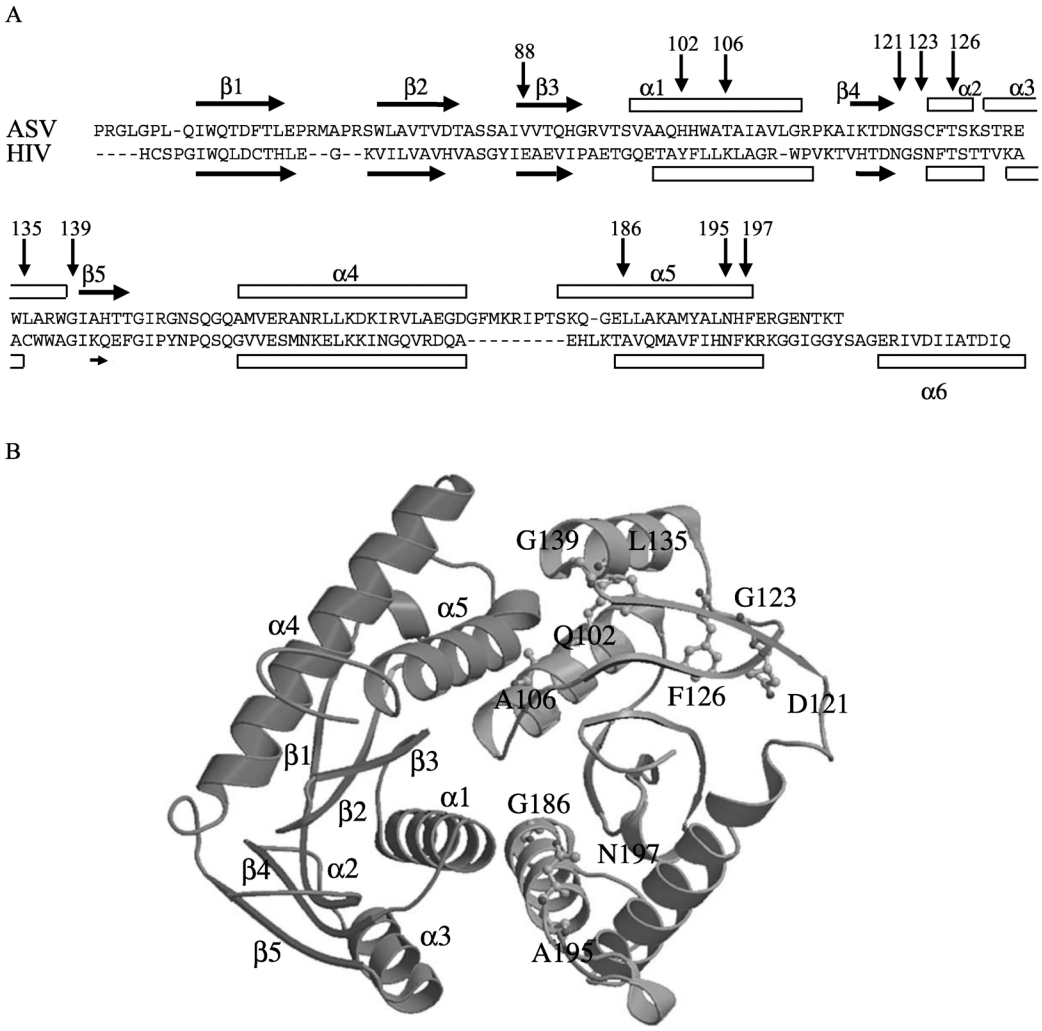


Fig. 2. IN core domain mutants. (A) Alignment of amino acid sequences of the ASV and HIV-1 integrases catalytic domains, with the elements of secondary structure indicated (adapted from Bujacz et al., 1995; Yang et al., 2000). Arrows indicate the amino acids mutated in the present study. (B) Ribbon representation of the dimeric core domain of ALSV integrase (residue 54–207) (adapted from (Yang et al., 2000)). To increase clarity of the figure, C-terminal domains (residues 208–268) have been omitted. Labels on the right subunit correspond to the 11 mutated residues discussed in this paper. Labels on the left subunit indicate α-helix and β-strands.

for the orientation of D121 side chain, and in consequence, essential for the catalytic mechanism. In mutant G123P, this contact is replaced by a contact with the side chain of D121. This could modify the orientation of D121. The mutation also creates two new contacts with C125 and F126.

F126 is a highly conserved residue of helix α2. It is part of the catalytic region and positioned at 8 Å away from the catalytic residue D121. F126 is accessible to solvent and, thus, could allow an important hydrophobic stacking of the protein with the DNA at the active site. The mutation F126I reduces contacts within the monomer, reinforces a contact (with H142), and may cancel the above-mentioned hydrophobic stacking of the protein with the DNA.

(3) I88 is a well-conserved residue in strand β3. The residue is involved in several intramolecular contacts. The mutation I88L induces changes in these contacts (M193 instead of V81 and L163).

L135 is comprised in helix α3, which is not involved in intermolecular interactions, and is positioned away from the active site. The mutation L135G removes all contacts of L135 within the monomer.

G139 is a well-conserved residue that makes the transition between helix α3 and strand β5. Its main chain angles (φ, ψ) are characteristic of a left handed α-

Table 2
Contacts between residues in monomers and dimers of the wt and mutant INs

ASV	Location	Contacts between side chains in wild type[a]	Contacts between side chains in mutant[a]
I88L	Strand β3	V81, L163, L196, N197, R201	L196, M193, N197, R201
Q102G	Helix α1	S98, *S130*, *T131*, W134, **E187**[b]	/
A106V	Helix α1	**E187**[b], **A190**[b]	***E187***[b], ***A190***[b], **K191**[b]
D121E	Active site	F126, *P147*, Q153	*P147*, Q153
G123P	Loop between strand β4 and helix α2	/	D121, C125, *F126*
F126I	Helix α2	V101, W105, T120, D121, C125, H142	T120, C125, *H142*
L135G	Helix α3	W105, T131, W134, I140, H142	/
G139D	Turn between helix α3 and strand β5	/	/
G186P	Helix α5	/	Q185, *H103*
A195V	Helix α5	/	*R166*
N197C	Helix α5	A87, I88, R201, P208	I88

[a] Maximum contact distance is 4 Å, residues in italic show contact distances <3.2 Å.
[b] Indicates residues of the second subunit in the dimer.

helix which is rare in proteins. Thus, G139 is most probably a critical residue for the correct folding of ALSV IN. This residue is not in contact with another residue neither in the wild-type protein nor in the mutant G139D (Table 2).

In a preliminary study (Moreau et al., 2002), mutants were tested for their 3′-processing and strand transfer activities using oligonucleotides that mimic the final 15 bp of *att* sequence (see Table 4). The activities of mutants were compared to that of the wild-type protein. I88L, A106V, L135G, G139D, G186P, and A195V mutants displayed levels of activity similar or close to that of the wild-type protein. G123P and N197C displayed reduced levels of activities (30–60% that of wild-type protein). The D121E mutant displayed a reduced level of 3′-processing activity (30–60% that of wild type) and a strongly reduced strand transfer activity (0–30%). Finally, mutations Q102G and F126I displayed reduced efficiency to perform the 3′-processing activity (30–60%) whereas strand transfer activity was close to that of the wild-type protein (Moreau et al., 2002) (Table 4).

Analyses of IN mutants in the concerted integration assay

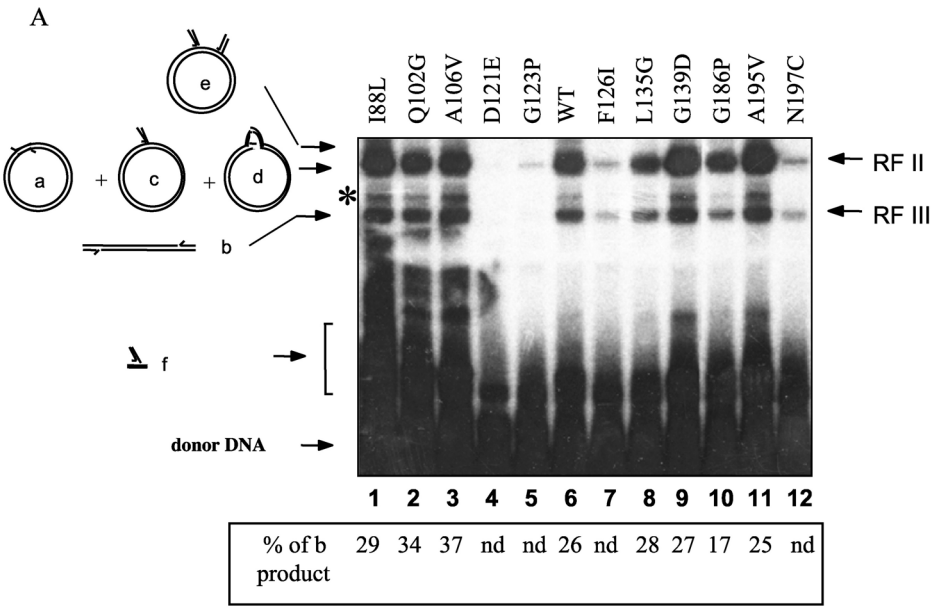
Mutants were analyzed with respect to their ability to perform an integration process using the reconstituted in vitro assay (Fig. 3A). For each mutant, integration efficiency was determined on gel by the intensity of bands corresponding to RFII and RFIII products (forms *a* + *b* + *c* + *d*) and in comparison to the wild-type protein (Fig. 3B, black bars). The experiments were repeated at least two times.

Gel analyses demonstrated that four mutations (D121E, G123P, F126I, and N197C) reduced significantly the integration process of the resulting mutants (Figs. 3A and 3B, lanes 4, 5, 7, and 12; and Table 4, column 6). By contrast, the Q102G mutant (lane 2) showed an integration efficiency similar to that of the wild-type protein (lane 6). Finally, some mutants displayed an integration efficiency slightly lower as compared to the wild-type protein (L135G and G186P, Fig. 3A, lanes 8 and 10) or slightly higher (I88L, A106V, G139D, and A195V; lanes 1, 3, 9, and 11) (Fig. 3B and Table 4).

Afterwards, we focussed on the ability of IN mutants to perform two-ended integration.

Firstly, integration products were cloned into MC1061/P3 *E. coli* bacteria to amplify the two-ended integration products. For each mutant, the two-ended integration efficiency was determined by the number of clones obtained relatively to the number of clones obtained with the wild-type protein. The experiment was performed at least two times. G139D, G186P, and A195V mutants had an efficiency slightly reduced when compared with that of the wild-type IN (between 70% and 90%) for two-ended donor integration. Two other mutants (I88L and F126I) were less efficient than wild-type IN (approximately 40% of the efficiency of the wild-type IN) (Fig. 3B). Mutations of four residues (D121, G123, L135, and N197) drastically inhibited the two-ended integration process. Unexpectedly, two mutants, Q102G and A106V, appeared to be more efficient than the wild-type IN in catalyzing the two-ended integration process.

For some mutants, the gel analyze (Fig. 3B, black bars) was in good agreement with the cloning analyze (grey bars). Indeed, (i) mutants D121E, G123P, F126I, G186P, and N197C were less efficient than the wild-type protein in performing integration as revealed on gel and after introduction of reactions into the bacteria. This suggests that these mutants are less efficient than wild-type IN in performing all kinds of integration (1- and 2-ended) (noted "1 and 2" in Table 4, column 7), (ii) the A106V mutant was more efficient than the wild-type protein in performing integration as revealed in both tests, which suggests that this mutant is more efficient than the wild-type IN in performing all kinds of integration (1- and 2-ended) (1 and 2, Table 4). By contrast, for mutant Q102G, the number of colonies recovered from bacteria was greater than the number of colonies recovered with the wild-type IN, while efficiency of integration revealed on gels was similar to that of the wild-type IN. This strongly suggests that mutant Q102G is specifically more efficient than the wild-type IN in performing two-ended DNA integration (noted 2 > 1 in Table 4). For mutants I88L, G139D, and A195V, the concerted integration reaction was enhanced as revealed



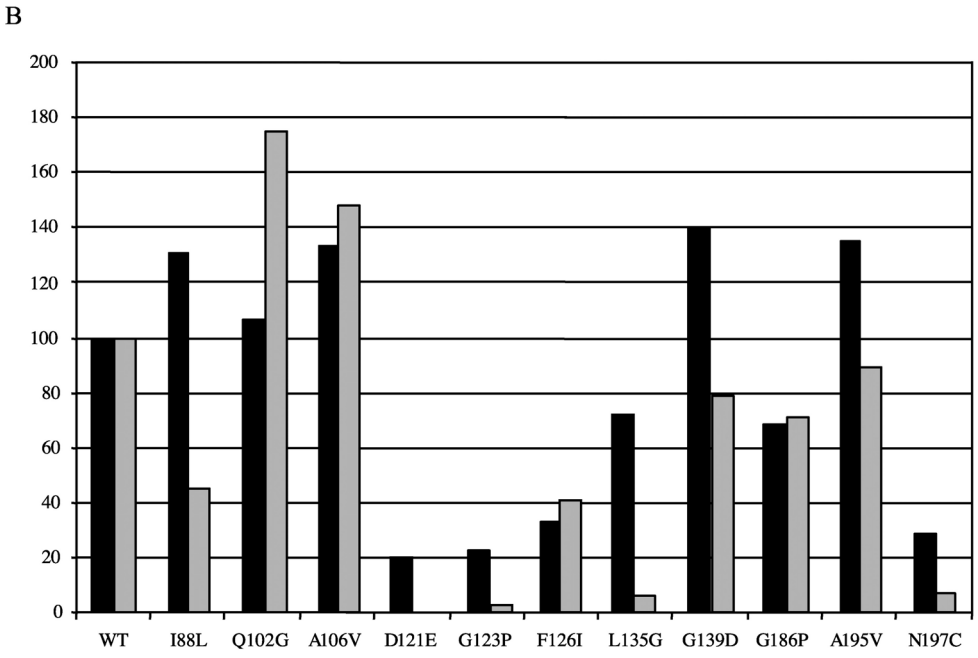


Fig. 3. Analyses of integration products obtained with the wild-type IN protein and mutants. (A) Using radiolabelled donor DNA, integration products obtained with wild-type IN protein and mutants were analyzed on 1.2% agarose gel. The letters above indicate the mutation: the first letter is the original residue, the number its position in the protein, and the second letter the residue that it was substituted into. Next to the picture are schematized integration products. (Bottom) Percentage of b product (RFIII forms relative to total integration products RFIII/(RFII + RFIII)). * structure of this recombinant is unknown. (B) Quantification of integration products (forms *a* + *b* + *c* + *d*) shown in A (in black) and total number of colonies recovered after the reaction products were introduced into bacteria (in grey). Integration efficiency of wild-type protein was set as 100%. 100% is defined as 98–324 colonies per plates (according to experiments), derived from reactions with wild-type IN. Results are from at least two experiments.

on the gel, but was less or slightly less efficient than the wild-type IN on bacteria (40%, 80%, and 90% of the efficiency of the wild-type IN, respectively). This suggests that these three mutants are more efficient than wild-type IN in performing the one-ended integration process, but are less efficient in performing the two-ended integration process.

Thus, these mutants have a trend to perform one-ended DNA integration instead of two-ended DNA integration (this is noted 1 > 2 in Table 4). More drastically, mutant L135G was slightly less efficient than wild-type IN in performing the reaction as observed on gel (70%), but it was much less efficient for two-ended integration as revealed on bacteria. These results suggest that among the integration products generated by this mutant and observed on gels, there was a lesser proportion of circular two-ended integration products amplified on bacteria as compared with the wild-type protein. Thus, the L135G mutant appears to be specifically altered in performing the two-ended integration process (noted 1 > 2 in Table 4).

Secondly, the RFIII products containing the linear *b* form were quantified separately (Fig. 3A, bottom) as this form was supposed to result from one event of two-ended concerted DNA integration (Fig. 1B). For each mutant, results are given as the percentage of product *b* relative to the total integration products (RFIII/RFII + RFIII) (Fig. 3A, bottom). Product *b* represents 26% of the total integration products generated by wild-type IN. It represents from 25% to 29% for mutants I88L, L135G, G139D, and A195V. Therefore, there are no relevant differences between these mutants and the wild-type IN regarding the ratio of product *b* among total integration products. Only mutants G186P in one hand, and mutants Q102G and A106V in the other hand, displayed reduced and increased proportion of product *b* (17% and 34–37%, respectively).

Overall, results of quantification of the two-ended integration events (cloning efficiency on bacteria and product *b*) are not in accordance. Only two mutants (Q102G and A106V) displayed concomitantly an enhanced ability in performing two-ended DNA integration, as revealed by both experiments. Mutant G186P displayed a concomitant reduced activity in both tests. For others, a decrease in two-ended integration on gels was not accompanied by a decrease of product b (see Discussion).

Molecular characterization of integration products

After cloning, integration products obtained with some of the mutants were analyzed. We sequenced products of two mutants with reduced efficiency in performing the two-ended integration (I88L and G139D) as well as products of two mutants with enhanced efficiency (Q102G and A106V) (Table 3).

Between 13 and 16 clones, depending of the mutant, were sequenced. Overall, the integration products were similar to those obtained with the wild-type protein. Most of them presented a short duplication of acceptor DNA whose sizes ranged from 4 to 9 bp. As for wild-type protein, a deletion of acceptor DNA (ranging from 14 to 504 bp) was observed in a few clones (in one clone for A106V and G139D mutants and in two clones for Q102G mutant). Several clones showed incorrect cleavage of the *att* sequence: three clones for I88L, five clones for A106V mutant, and seven clones for G139D mutants. In conclusion, these analyses show that mutants I88L, Q102G, A106V, and G139D performed a correct integration process, roughly comparable to that of the wild-type IN, with correct cleavage of viral ends and small-size duplication of acceptor DNA. We can note a higher number of clones displaying incorrect cleavage with the G139D mutant than with other mutants.

Table 3
Sequence analyses of donor–acceptor junction sites produced by IN mutant proteins

Protein	Characteristics	Number of recombinants
I88L	Duplications of 6 bp	9
	Other duplications (4–7 bp)	4
	Deletion in acceptor DNA	0
	Total	13
	Incorrect cleavage of *att* sequence[a]	3
Q102G	Duplications of 6 bp	6
	Other duplications (4–7 bp)	7
	Deletion in acceptor DNA (316–504 bp)	2
	Total	15
	Incorrect cleavage of *att* sequence[a]	0
A106V	Duplications of 6 bp	6
	Other duplications (5–9 bp)	9
	Deletion in acceptor DNA (14 bp)	1
	Total	16
	Incorrect cleavage of *att* sequence[a]	5
G139D	Duplications of 6 bp	9
	Other duplications (5–9 bp)	6
	Deletion in acceptor DNA (292 bp)	1
	Total	16
	Incorrect cleavage of *att* sequence[a]	7

[a] Among all clone studies, few of them were deleted of more than 2 bp at one or the other *att* viral end.

Oligomeric state of IN

It has been reported that IN acts as a multimeric complex during integration. It was proposed to act as a dimer (Vincent et al., 1993), a tetramer (van Gent et al., 1993; Wang et al., 2001), or even as an octamer (Bao et al., 2003; Cherepanov et al., 2003; Heuer and Brown, 1998) during the integration process. ALSV and HIV INs are present as mono-, di-, and tetramers in solution, as shown by size exclusion chromatography, analytical ultracentrifugation, and cross-linking (Andrake and Skalka, 1995; Bao et al., 2003; Cherepanov et al., 2003; Coleman et al., 1999; Deprez et al., 2000; Jenkins et al., 1996; Jones et al., 1992; Vercammen et al., 2002). Mutants which are less efficient than the wild-type protein for two-ended concerted integration (I88L, D121E, G123P, F126I, L135G and N197C) were therefore tested for their ability to form a complex by protein–protein cross-linking and size exclusion chromatography.

In the protein–protein cross-linking experiment, the proteins were incubated with the cross-linker disuccinimydyl suberate (DSS) and reaction products were analyzed

on SDS PAGE and revealed by Western blot. As expected, in the absence of IN, we did not observed any product (Fig. 4A, lane 1); in the absence of cross-linker, we observed only the monomeric form of IN (Fig. 4A, lane 2). With wild-type IN and in the presence of DSS, we detected products at the expected molecular weight of integrase monomers and dimers (Fig. 4A, lane 3). With the exception of N197C mutant (lane 9), all mutants gave a profile similar to the wild-type protein profile and were revealed as monomeric and dimeric complexes. Furthermore, the intensity of the cross-linked product was roughly the same as the one observed with wild-type protein. On the contrary, the N197C protein was not cross-linked by DSS and only the monomeric form was observed on gel. All other mutants described in the present study (Q102G, A106V, G139D, G186P, and A195V) have also been tested by cross-linking, and all of them presented a profile similar to that of the wild-type protein (data not shown).

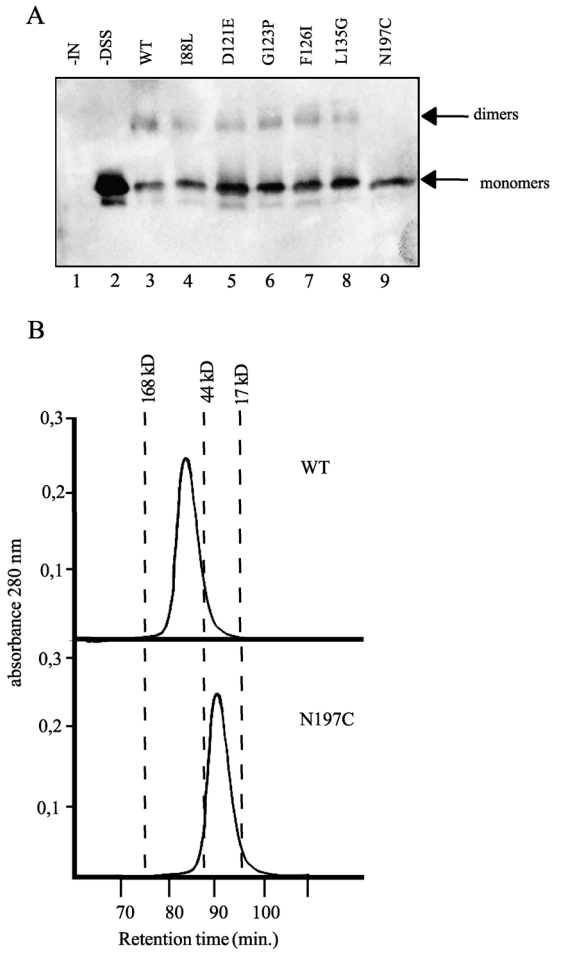


Fig. 4. Analyze of oligomeric state of wild-type integrase and mutants. (A) Protein–protein cross-linking experiments. Proteins were incubated in the presence of disuccinimidyl suberate (DSS). Reaction products were analyzed on 10% polyacrylamide gel and revealed by western blot with an anti-Tag Histidine antibody. The migration of cross-linked species, consistent with the molecular size of monomers and dimers, are marked. −IN: without integrase, −DSS: without disuccinimidyl suberate. (B) Size exclusion chromatography of wild-type IN protein and N197C mutant. Elution profiles of wild-type (WT) IN and N197C mutant are shown. The molecular size of monomer protein is 36.7 kDa. For reference, the elution positions of three globular molecular mass standards are indicated with dotted vertical lines. Retention times, given in minutes, are indicated on *x*-axis.

The size exclusion chromatography analyses confirmed the results of cross-linking analyses. Under the conditions used, the wild-type protein eluted at a position consistent with a molecular size of dimers (Fig. 4B). The N197C mutant presented a different profile and eluted at a position consistent with the molecular size of monomers (Fig. 4B). All other mutants tested by size exclusion chromatography (D121E, G123P, F126I, L135G) eluted at a position consistent with a dimeric complex (data not shown).

Results of both experiments demonstrated that the N197C mutation rendered the IN enzyme unable to form dimers. All other mutants studied here were able to form dimeric complexes.

Discussion

In this study, we analyzed both the functional activities of several points mutants in the core domain of an ALSV IN in a concerted DNA integration assay in vitro and the oligomeric state of these mutants. Our analysis focused on residues at or close either to the active site of the enzyme or to the core domain dimer interface as well as on highly conserved residues. Similar analyses have been performed on residues in the C-terminal domain (Moreau et al., in press).

A compilation of data obtained for each mutant in the core domain studied here and previously (Moreau et al., 2002) is reported in Table 4. In general, results of concerted integration on gel were in good agreement with catalytic activities of 3′-processing and strand transfer. Thus, mutants which act efficiently in the concerted integration test such as I88L, A106V, G139D, and A195V (which are slightly more efficient than the wild-type IN) as well as mutants L135G and G186P (which are slightly less efficient than the wild-type IN) had high activities of 3′-processing and strand transfer. For four other mutants (D121E, G123P, F126I, and N197C), a reduced activity of concerted integration on gel is correlated with a decrease either of the 3′-processing activity or of the strand transfer activity or of both activities. Surprisingly, mutant Q102G, which displayed a reduced activity of 3′-processing (55% that of the wild-type IN), was nevertheless able to perform concerted DNA integration with the same efficiency of the wild-type IN.

In the in vitro concerted DNA integration assay, we could evaluate the ability of IN to catalyze two-ended concerted DNA integration by two ways: (i) by quantifying the linear product *b*, because this product is supposed to be generated

Table 4
Compilation of data for each mutant

Mutation[a]	Conservation[b]	Catalytic activities[c]		DNA binding[d]	IC test		Oligomeric status[e]
		3′-proc	S. transf.		Total integr.[f]	Bacteria[g]	
I88L	C	++++	+++	++++	↗	1 > 2	D
Q102G*	–	++	+++	++++	→	2 > 1	D
A106V*	–	++++	++++	++++	↗	1 and 2	D
D121E	C	++	+	++++	↘ ↘	1 and 2	D
G123P	C	++	++	++++	↘ ↘	1 and 2	D
F126I	C	++	+++	++++	↘ ↘	1 and 2	D
L135G	–	+++	+++	++++	↘	1 ≫ 2	D
G139D	C	++++	++++	++++	↗	1 >2	D
G186P	–	++++	++++	++++	↘	1 and 2	D
A195V	–	+++	++++	++++	↗	1 > 2	D
N197C	C	++	++	++	↘ ↘	1 and 2	M

[a] The * indicates residues involved in the dimer interface.
[b] C: residues conserved among INs, as shown by sequences alignments (Moreau et al., 2002) and crystallographic data.
[c] Data not shown and data from Moreau et al. (2002). 3′-P, 3-processing; S.t., strand transfer; +, 0–30% activity of the wt IN; ++, 30–60%; +++, 60–90%; ++++ >90%.
[d] Data from Moreau et al. (2002).
[e] Results from Fig. 4. D: dimers. M: monomers.
[f] Results from Fig. 3B. Changes (increase or decrease) in integration efficiencies as revealed on gel, and in comparison with wild-type IN efficiency.
[g] Results from Fig. 3C. 1 and 2: the 1- and 2-ended DNA integration events increase or decrease concomitantly by comparison with the 1- and 2-ended DNA integration events of the wild-type IN. 1 > 2:1-ended DNA integration events are favored more than the 2-ended DNA integration events. 2 > 1: 2-ended DNA integration events are favored more than the 1-ended DNA integration events.

by a two-ended concerted DNA integration of two DNA donors (Aiyar et al., 1996; Brin and Leis, 2002a, 2002b; Hindmarsh et al., 1999, 2001) (Fig. 1B); (ii) by quantifying the number of colonies recovered after cloning of integration products into the bacteria, which allow selective amplification of the two-ended circular integration products [*a* (concerted) and *d* (nonconcerted)]. The products *a* and *d* were subsequently distinguished by sequencing of integration products, and gross deletions of acceptor DNA were assigned to two-ended nonconcerted DNA integration events (class *d)* (Brin and Leis, 2002a, 2002b; Hindmarsh et al., 2001). In our experiments, most products were of type *a* (without deletion of acceptor DNA) (Tables 1 and 3). Therefore, cloning of integration reactions into bacteria gives a relevant estimation of product *a*, and subsequently, of the two-ended concerted DNA integration events. According to these assays, if an IN mutant performs two-ended integration less efficiently than wild-type IN, we expect a concomitant decrease both in the ratio of product *b* among the total integration products and in the number of recovered colonies from bacteria. Unexpectedly, we found that the quantity of product *b* did not systematically match the recovered number of colonies (Fig. 3B). This is particularly striking for mutant I88L (lane 1), for which the ratio of product *b* was similar to that of wild-type proteins (29% and 26%, respectively), whereas the ratio of two-ended integration products amplified in bacteria was reduced to about 40% of that observed with wild-type IN. Such a discrepancy is also evident for mutant L135G (lane 8). Similar observations have previously been made by us when analyzing mutants of IN in the C-terminal domain (Moreau et al., in press) and by others (Hindmarsh et al., 2001; Vora et al., 1997). For instance, the ability of a U5 mutated donor to undergo concerted DNA integration in vitro was 1.5–2-fold greater than that observed with a wild-type donor substrate. This stimulation of integration concerned both the RFII (*a* + *c* + *d*) and RFIII (*b*) products. However, when integrants were introduced into bacteria, the number of colonies recovered was reduced to 25% relative to the wild-type donor (Hindmarsh et al., 2001). Altogether, these independent observations show that: (i) when the quantity of the total integration products revealed on gel increases, the quantity of product *b* increases in a similar proportion, (ii) whereas, in the same reaction, the quantity of product *a* (and *d*) may decrease in an independent manner. Furthermore, only a significant increase in the two-ended integration reaction quantified on gel (34–37% of product *b*), as observed for Q102G and A106V, was accompanied by an increase in the recovered number of clones from bacteria. Therefore, we speculate that product *b* is a mix of several linear products resulting from both nonconcerted and concerted events of integration and, hence, that quantification of product *b* is not strictly relevant to the two-ended integration products.

The core domain contains the catalytic site of the IN enzyme (Engelman and Craigie, 1992; van Gent et al., 1992). Three mutations examined were at residues of the catalytic site of the core domain (D121E) or close to it (G123P and F126I). As expected, we observed that the D121E mutation completely inhibits concerted integration in vitro, as deduced from gel analyses and cloning experiment (Table 4). Structurally, D121 interacts with magnesium ion. The mutant

D121E is certainly inactive because the substitution prevents the correct coordination of the divalent cation. Mutation of the G123 residue, which is very close to the D121 active site residue and is highly conserved among INs (Bujacz et al., 1995; Moreau et al., 2002) (Fig. 2A), also completely inhibited the concerted integration process. On the basis of structural data, the mutant G123P may retain the overall fold of ALSV IN. However, mutation G123P, which changes the liaison between residues G123 and D121, may disturb the orientation of the D121 residue, and hence, may render the protein inactive. Mutation of the nearby highly conserved F126 residue also induced a reduction in integration efficiency. The mutation F126I may cancel an important hydrophobic stacking of the protein with the DNA at the active site. Consequently, the nucleotide segment may not be firmly bound in the mutant, which could impair the activity of the mutant.

The core domain is also involved in dimerization of the protein and the isolated domain has been found to dimerize (Bujacz et al., 1995, 1996; Lubkowski et al., 1999). Some mutants studied here [Q102 and A106V (α1) and G186P, A195V, and N197C (α5)] are part of the secondary structures involved in the dimer interface (Fig. 2).

Mutation of the N197 residue drastically impaired dimer formation (Fig. 4). The N197 residue is positioned at the end of the α5 helix and is close to the dimer interface but is not involved in contacting with the other monomer (Fig. 2B, Table 2). It is possible that a cystein residue cannot adopt the alternate conformation observed for the asparagine (Lubkowski et al., 1999). This alternate conformation may be necessary for the correct folding of helix α5 and, thus, for the formation of the dimeric interface. Alternatively, it is possible that the monomer of N197C IN is not folded correctly, which may impair formation of the dimer. Indeed, the N197 residue makes contact with the A87 residue of the β3 strand, which the mutation N197C cancels. Concomitantly to dimer impairment, the N197C protein was unable to perform concerted integration. We had also previously shown that the N197C mutant interacts with DNA with a weaker efficiency than does the wild-type protein (Table 4, Moreau et al., 2002). Thus, defect of the N197C mutant for concerted integration may be due to its inability to associate as a dimer or to its weak ability to bind DNA, or both.

By contrast, the mutation Q102G removes contact in the dimer within the α1 helix (Fig. 2B). Nevertheless, size exclusion chromatography and cross-linking experiments conducted with Q102G show that it had the same elution profile as the wild-type protein and was folded as a dimer. It is possible that mutating this residue was not sufficient by itself to impair the formation of the dimer. Conversely, mutation A106V in the α1 helix reinforces contacts within the dimer (Table 4). Accordingly, this mutant is folded as a dimer. To our surprise, both mutants were more efficient than the wild-type protein in performing two-ended concerted integration as deduced from both gel and cloning analyzes of integration products (Fig. 3). This result suggests that among the integration products observed on gel, there is a high proportion of two-ended concerted integration products able to be amplified in bacteria. Accordingly, the ratio of product *b* was found to be enhanced for both mutants (Fig. 3B). Furthermore, the processes of integration were correct as revealed by sequence analysis of integration products (Table 3). Mutation Q102G, which reduces the number of contacts within the monomer, is expected to generate a more flexible protein while the A106V mutation, which increases the number of contacts within the monomer and the dimer, is supposed to generate a more rigid protein (Table 2). As these two mutations have the same effect in increasing the two ended-integration events, changes in plasticity of the protein do not explain the increased ability for concerted integration. It is well known that IN protein has other functions in addition to integration in the replicative cycle of retroviruses, such as precursor polyprotein processing, particle assembly and release, DNA synthesis, and nuclear import (Bouyac-Bertoia et al., 2001; Engelman et al., 1995; Shin et al., 1994). Therefore, it may be possible that these two residues in the α1 helix, whose mutations induce a more active protein for integration, are essential for other functions of IN in the replicative cycle and are not the best one for integration. To test this hypothesis, we are analyzing these mutants in the context of a replicative cycle.

G186 and A195 are part of helix α5, which is in contact with the above-mentioned helix α1 of the symmetry related monomer. However, these two residues do not face helix α1, unlike Q102 and A106 which face α5. Mutations G186P and A195V affect only slightly the property of the protein (A195V only increases the one-ended events of integration over the two-ended integration). Thus, it seems that mutations at the dimer interface between α1 and α5 alter more significantly the concerted mechanism than mutations at the sides of the interface.

The highly conserved G139 makes the transition between helix α3 and strand β5 and appears to be a critical residue for the correct folding of ALSV IN. The mutant G139D performed concerted integration with high efficiency. However, almost half of the clones (7/16) generated by this mutant were incorrectly cleaved and contained deletions of *att* sequences (Table 3). This suggests that the overall conformation of the mutant G139D may be different from that of the wild type. This may cause a mispositionning of the substrate resulting in ectopic cleavages of *att* extremities but without reducing efficiency of the integration process.

Finally, mutant L135G and mutant I88L to a lesser extent presented interesting profiles. Both mutants are active for 3′-processing and strand transfer and both are also able to bind DNA with the same efficiency as wild-type protein (Moreau et al., 2002) (Table 4). In the concerted integration assay, they exhibited an integration efficiency close to that of the wild-type protein, as observed on gel (130% and 75%, respectively), but there was a significant reduction in the proportion of integration products recovered when introduced into bacteria (about 40% for I88L and 1% for

L135G). These results reveal that among the integration products observed on gel, there was a lower proportion of two-ended integration products (able to replicate in bacteria) for I88L and L135G mutants than for wild-type protein. Therefore, these mutations appear to reduce more specifically the two-ended integration process. For the I88L mutant, sequencing of integration products revealed that, when two-ended integration occurred, it was correct (Table 3). Furthermore, size exclusion chromatography and cross-linking experiments demonstrated that these mutated proteins were present as a dimer complex, like the wild-type protein. Altogether, these observations are consistent with a disruption of a higher-order IN complex necessary for two-ended integration. Indeed, different authors have suggested that the minimal IN structure for concerted integration may be a tetramer (Bao et al., 2003; Cherepanov et al., 2003; Heuer and Brown, 1998; Vora and Grandgenett, 2001; Yang et al., 2000). Therefore, we suggest that I88L and L135G mutations may alter the formation of higher molecular size complexes than dimers resulting specifically in a reduction in two-ended concerted DNA integration. It is possible that these residues are directly involved in the formation of the tetramer. However, this assumption is not in accordance with structural data. Indeed, I88L and L135G residues are part of the interior core of the protein, and therefore are unlikely to be involved in tetramer formation. Therefore, these mutations most likely induce conformational changes that either have local effects that prevent the formation of a tetramer or have distal effects that affect the global structure of the protein.

In summary, our results show the importance of residues of the core domain in IN oligomerization and concerted integration. Our data show a strong structural role of residue N197 in ALSV IN dimerization, and the importance of residues Q102 and A106 at the dimer interface in the efficiency of two-ended concerted integration. They also predict a potential role for the L135 residue, and to a lesser extent the I88 residue, in the architecture of the molecule. These residues of the protein may constitute new targets for the development of antiviral drugs against integrase.

Materials and methods

DNA manipulation

The DNA pBSK-Zeo acceptor plasmid was constructed as follows. The pBSK + plasmid (Stratagene) was digested by *Sma*I and *Sac*II restriction enzymes, subjected to the *Klenow* enzyme, and reclosed by ligation to generate pBSK + Δ*Bam*HI plasmid. pBSK + Δ*Bam*HI plasmid was then digested by *Hin*dIII and *Eco*RV, subjected to the *Klenow* enzyme, and reclosed by ligation to generate pBSK + Δ2 plasmid. These manipulations removed the *Bam*HI and *Eco*RV restriction sites. Afterwards, pBSK + Δ2 plasmid was amplified by PCR using the *Pfu* turbo polymerase enzyme (Stratagene) and BU (5′CCGATATCATACTC-TTCC3′) and BL (5′CCGATATCAGACCAAGTTTAC3′) primers. In the same way, the *zeo* gene was amplified from the pHook.3 plasmid (Invitrogen) using Z1 (5′CCGATAT-CGTGTTGACAATTAATC3′) and Z2 (5′CCGATATCC-AGACATGATAAGATAC3′) primers. All primers contain an *Eco*RV restriction site and the resulting PCR products, pBSK + Δ2 and the *zeo* gene, were digested by the *Eco*RV restriction enzyme and ligated together giving the pBSK-zeo plasmid. This plasmid, which is zeocin resistant, was amplified in DH5α bacteria (Invitrogen).

The donor DNA was obtained as follows. The *sup*F gene was amplified by PCR from piVX plasmid (ATCC) using H-sup1 (5′ GAGAAGCTTAACGTTGCCCGGATCCGGTC 3′) and P-sup2 (5′ GAGCTGCAGTAGTCCTGTCGGGTT-TCGCC 3′) primers containing *Hin*dIII and *Pst*I restriction sites, respectively. The amplification product was digested with *Hin*dIII and *Pst*I restriction enzymes and ligated into the pBSK + plasmid digested by the same restriction enzymes giving pBSK-*sup*F plasmid. The donor DNA was then amplified from pBSK-*sup*F plasmid by PCR using the pfu turbo polymerase enzyme, U3 (GATGTAGTCTTATAC-GTTGCCCGGATCCGG 3′) and U5 bis (5′ AATGAAG-CCTTCTGCTTTGAGCGTCGATTTTTG 3′) primers. The PCR product was purified from agarose gel using the Qiaex II kit (Qiagen). The final donor DNA contained 15 bp of the U3 end sequence of the Avian Erythroblatosis Virus (AEV) and 12 bp of the U5 end sequence.

Modeling of the mutants

Construction of IN mutants has been reported elsewhere (Moreau et al., 2002). Numerous crystallographic structures of the core domain of ASV IN have been solved under different crystallization conditions (changes in pH and in precipitant agents) (Bujacz et al., 1995, 1996; Lubkowski et al., 1998a, 1998b, 1999). Several residues in alternate conformation have been observed in the two highest resolution structures, determined at 1.02 Å resolution in presence of ammonium sulphate (named ASVI-N_AS) and at 1.06 Å resolution in presence of HEPES (named ASVIN_HEP) (Lubkowski et al., 1999). Two two-domain structures of Rous Sarcoma Virus (RSV) IN, containing the core and the C-terminal region, have been solved in space groups $P2_1$ and P1 at 3.1 and 2.5 Å resolution, respectively (Yang et al., 2000). No structure containing also the N-terminal domain has been published yet. In consequence, the 2.5-Å two-domain structure of RSV IN was used to model the structure of mutants. Modeling has been performed on the dimer. Each structure of single mutant has been generated using the program CALPHA (Esnouf, 1997) and minimized with the program CNS using a conjugate gradient method (Brunger et al., 1998). Resulting models were displayed and analyzed on a graphic station using the program TURBO-FRODO (Rous-

sel and Cambillau, 1989). Contact distances were computed with CNS around each mutated residue. During structural analysis, the high-resolution structures of ASVIN_AS and ASVIN_HEP have been superimposed with the models and displayed on graphic station so as to check if the structures of ASVIN cores share a similar topology at the mutated regions. In parallel, a BLAST search (Altschul et al., 1997) was performed against the SWISS-PROT and the TrEMBL sequence databases (Bairoch and Apweiler, 2000) to detect homologous proteins. A multiple sequence alignment was performed in turn with CLUSTAL (Thompson et al., 1994): the 11 studied substitutions are unique in retrovirus as well as in lentivirus integrases.

Purification of proteins

The RAV-1 IN coding sequences (wild type or mutants) were cloned into the pET30a plasmid (Invitrogen) beyond a six histidine tag. IN proteins were expressed in BL21 bacteria (Invitrogen) and purified as described by others (Puras Lutzke and Plasterk, 1998).

The pET15b-HMGI expressing vector was generously donated by T.H. Kim, Cambridge, USA (Thanos and Maniatis, 1992). The High Mobility group type I (HMGI) proteins were expressed in BL21 (DE3) pLysS bacteria (Invitrogen) in the presence of 100 μg/ml of ampicillin and 34 μg/ml of chloramphenicol upon induction with 1 mM of IPTG for 3 h. Purification was conducted as follows. The bacteria pellet was resuspended in PBS plus 0.1% of triton before sonication. Then 5% of perchloric acid was added and the solution was incubated for 30 min at 4 °C. The lysate was then centrifuged for 10 min at 12 000 × *g*. A total of 25% of trichloroacetic acid was added to the supernatant, which was then incubated for an hour on ice. After 10 min centrifugation at 12 000 × *g*, the pellet was rinsed once with acetone and 0.2% HC1 (−20 °C), twice with acetone 70%, ethanol 20%, 20 mM Tris−HCl, pH 8 (−20 °C), and once with acetone (−20 °C). The pellet was then dried at room temperature before being resuspended in 250 μl of Tris−EDTA. The solution was then passed through a Hitrap Heparin column (Pharmacia), which was previously equilibrated with 0.5 M NaCl, 50 mM $NaH_2\ _PO_4$, pH 7.4 solution. The column was washed with 0.5 M NaCl, 50 mM $NaH_2\ _PO_4$, pH 7.4 solution and the proteins were eluted with a gradient from 0.5 to 1.5 M of NaCl. Each fraction was analyzed by Bradford quantification and Western blot.

Integration reaction

A total of 60 ng of purified IN protein was incubated overnight at 4 °C with 100 ng of pBSK-zeo plasmid, 10 ng of donor DNA, and 100 ng of purified HMGI protein in a final volume of 5 μl. The volume of reaction was then increased to 20 μl with a final concentration of 20 mM HEPES, pH 7.5, 1 mM DTT, 30 mM $MgCl_2$, 15% DMSO, 8% PEG 8000, and 50 mM NaCl, and the integration mixture was incubated at 37 °C for an hour and a half.

Gel analyses of the integration reaction

For gel analyses of the integration reaction, the DNA donor was radiolabelled by including 8 μCi of dCTPα^{32}P in the PCR amplification mixture. After the integration reaction was performed, the volume was increased to 50 μl with the addition of 4.25 mM of EDTA, 0.44% of SDS, and 20 ng of proteinase K (Roche diagnostics) and samples were incubated for 1 h at 55 °C. The DNAs were deproteinized by phenol-chloroform extraction and purified by ethanol precipitation. Samples were then loaded on 1.2% agarose gel in 0.5× TBE. After electrophoresis, the gels were fixed in a 5% TCA solution for 30 min and dried for 3 h at 45 °C. Lastly, the gels were exposed to autoradiographic film overnight at −80 °C. Integration products were quantified using a phosphoimager (Biorad).

Cloning and sequencing of two-ended integration products

To clone integration products for sequencing, products of the integration reaction were purified on a Qiaquick column (Qiagen) as described by the supplier. The whole reaction was introduced into MC10610/P3 *E. coli* bacteria (Invitrogen) as described by others (Aiyar et al., 1996). MC1061/P3 *E. coli* bacteria contain ampicillin-, tetracyclin-, and kanamycin-resistant genes. Both ampicillin- and tetracyclin-resistant genes carry an *amb* mutation. These proteins are thus expressed only in the presence of the *sup*F gene products. Among the different integration products, only circular plasmids carrying the *sup*F gene were able to replicate and form colonies. Integration clones carrying both zeocin-resistant and *sup*F genes were therefore selected in the presence of 40 μg/ml of ampicillin, 10 μg/ml of tetracyclin, 15 μg/ml of kanamycin, and 25 μg/ml of zeocin. Under these conditions, we detected no colony when donor and acceptor DNAs were transfected to cell in the absence of IN. According to the experiment, we were able to detect between 100 and 300 colonies with the wild-type IN protein. Plasmids were isolated from quadruply resistant colonies and donor−acceptor DNA junctions were sequenced using SL primer (5′ ACTCTAAATCTGCCGTCATCG 3′) for the U3 junction and SU primer (5′ ATCATATCAAATGACGCGCCG 3′) for the U5 junction. SL and SU primers are located on the donor DNA.

Size exclusion chromatography

All proteins were centrifuged for 10 min at 14 000 rpm to remove IN aggregates. A total of 100 μl of integrase solution at a concentration of 30 μM was loaded on a Superoz 12 column (Pharmacia) previously equilibrated with 1 M NaCl, 25 mM HEPES, pH 7.5, 0.1 mM EDTA,

and 1 mM β-mercaptoethanol buffer. Size exclusion chromatography was performed at 4 °C. The column was calibrated with molecular mass markers. Protein elution was monitored at A280 nm at a flow rate of 0.3 ml/min.

Protein–protein cross-linking

Wild-type or mutant integrases were treated with 40 μg/ml disuccinimidyl suberate (Pierce). Reactions included 2 μg of protein in a final volume of 10 μl of 20 mM HEPES, pH7.5, 60 mM NaCl, 0.7 mM EDTA, 10% glycerol, and 4.5 mM CHAPS. Following 30 min at 22 °C, reactions were quenched by the addition of 3 mM of Lysine and 25 mM of Tris–HCl (pH 8). After 10 min at 22 °C, reactions were boiled for 10 min in sample buffer and electrophoresed in SDS page (10%). Products were revealed by Western blot using anti-His-tag antibody (Roche Diagnostics).

Acknowledgments

This work was supported by research grants from the Centre National de la Recherche Scientifique and the Institut de la Recherche Agronomique. We acknowledge the French Ministry of Research and the Agence Nationale de Recherche contre le SIDA (ANRS) for fellowships (KM and SV). We thank T.H. Kim (Cambridge) for providing the pET15b-HMGI plasmid. Special thanks to Dr. S. Carteau for helpfull discussions and to Dr. E. Derrigton for correcting of the English. Thanks are also due to Dr. S. Arnaud and M.-F. Grasset for their help with the size exclusion chromatography assay.

References

Aiyar, A., Hindmarsh, P., Skalka, A.M., Leis, J., 1996. Concerted integration of linear retroviral DNA by the avian sarcoma virus integrase in vitro: dependence on both long terminal repeat termini. J. Virol. 70, 3571–3580.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402.

Andrake, M.D., Skalka, A.M., 1995. Multimerization determinants reside in both the catalytic core and C terminus of avian sarcoma virus integrase. J. Biol. Chem. 270, 29299–29306.

Appa, R.S., Shin, C.G., Lee, P., Chow, S.A., 2001. Role of the nonspecific DNA-binding region and alpha helices within the core domain of retroviral integrase in selecting target DNA sites for integration. J. Biol. Chem. 276, 45848–45855.

Bairoch, A., Apweiler, R., 2000. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res. 28, 45–48.

Bao, K.K., Wang, H., Miller, J.K., Erie, D.A., Skalka, A.M., Wong, I., 2003. Functional oligomeric state of avian sarcoma virus integrase. J. Biol. Chem. 278, 1323–1327.

Berger, N., Heller, A.E., Stormann, K.D., Pfaff, E., 2001. Characterization of chimeric enzymes between caprine arthritis–encephalitis virus, maedi-visna virus and human immunodeficiency virus type 1 integrases expressed in *Escherichia coli*. J. Gen. Virol. 82, 139–148.

Bouyac-Bertoia, M., Dvorin, J.D., Fouchier, R.A., Jenkins, Y., Meyer, B.E., Wu, L.I., Emerman, M., Malim, M.H., 2001. HIV-1 infection requires a functional integrase NLS. Mol. Cell 7, 1025–1035.

Brin, E., Leis, J., 2002a. Changes in the mechanism of DNA integration in vitro induced by base substitutions in the HIV-1 U5 and U3 terminal sequences. J. Biol. Chem. 277, 10938–10948.

Brin, E., Leis, J., 2002b. HIV-1 integrase interaction with U3 and U5 terminal sequences in vitro defined using substrates with random sequences. J. Biol. Chem. 277, 18357–18364.

Brown, P.O., 1997. Integration. In: Coffin, J.M., Hugues, S.H., Varmus, H.E. (Eds.), Retroviruses. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, pp. 161–204.

Brunger, A.T., Adams, P.D., Clore, G.M., DeLano, W.L., Gros, P., Grosse-Kunstleve, R.W., Jiang, J.S., Kuszewski, J., Nilges, M., Pannu, N.S., Read, R.J., Rice, L.M., Simonson, T., Warren, G.L., 1998. Crystallography and NMR system: a new software suite for macromolecular structure determination. Acta Crystallogr., D Biol. Crystallogr. 54, 905–921.

Bujacz, G., Jaskolski, M., Alexandratos, J., Wlodawer, A., Merkel, G., Katz, R.A., Skalka, A.M., 1995. High-resolution structure of the catalytic domain of avian sarcoma virus integrase. J. Mol. Biol. 253, 333–346.

Bujacz, G., Jaskolski, M., Alexandratos, J., Wlodawer, A., Merkel, G., Katz, R.A., Skalka, A.M., 1996. The catalytic domain of avian sarcoma virus integrase: conformation of the active-site residues in the presence of divalent cations. Structure 4, 89–96.

Carteau, S., Gorelick, R.J., Bushman, F.D., 1999. Coupled integration of human immunodeficiency virus type 1 cDNA ends by purified integrase in vitro: stimulation by the viral nucleocapsid protein. J. Virol. 73, 6670–6679.

Chen, J.C., Krucinski, J., Miercke, L.J., Finer-Moore, J.S., Tang, A.H., Leavitt, A.D., Stroud, R.M., 2000. Crystal structure of the HIV-1 integrase catalytic core and C-terminal domains: a model for viral DNA binding. Proc. Natl. Acad. Sci. U.S.A. 97, 8233–8238.

Cherepanov, P., Maertens, G., Proost, P., Devreese, B., Van Beeumen, J., Engelborghs, Y., De Clercq, E., Debyser, Z., 2003. HIV-1 integrase forms stable tetramers and associates with LEDGF/p75 protein in human cells. J. Biol. Chem. 278, 372–381.

Chiu, R., Grandgenett, D.P., 2000. Avian retrovirus DNA internal attachment site requirements for full-site integration in vitro. J. Virol. 74, 8292–8298.

Chiu, R., Grandgenett, D.P., 2003. Molecular and genetic determinants of rous sarcoma virus integrase for concerted DNA integration. J. Virol. 77, 6482–6492.

Coleman, J., Eaton, S., Merkel, G., Skalka, A.M., Laue, T., 1999. Characterization of the self association of Avian sarcoma virus integrase by analytical ultracentrifugation. J. Biol. Chem. 274, 32842–32846.

Daniel, R., Katz, R.A., Skalka, A.M., 1999. A role for DNA-PK in retroviral DNA integration. Science 284, 644–647.

Daniel, R., Kao, G., Taganov, K., Greger, J.G., Favorova, O., Merkel, G., Yen, T.J., Katz, R.A., Skalka, A.M., 2003. Evidence that the retroviral DNA integration process triggers an ATR-dependent DNA damage response. Proc. Natl. Acad. Sci. U.S.A. 100, 4778–4783.

Deprez, E., Tauc, P., Leh, H., Mouscadet, J.F., Auclair, C., Brochon, J.C., 2000. Oligomeric states of the HIV-1 integrase as measured by time-resolved fluorescence anisotropy. Biochemistry 39, 9275–9284.

Dyda, F., Hickman, A.B., Jenkins, T.M., Engelman, A., Craigie, R., Davies, D.R., 1994. Crystal structure of the catalytic domain of HIV-1 integrase: similarity to other polynucleotidyl transferases. Science 266 (5193), 1981–1986.

Engelman, A., 1999. In vivo analysis of retroviral integrase structure and function. Adv. Virus Res. 52, 411–426.

Engelman, A., Craigie, R., 1992. Identification of conserved amino acid

residues critical for human immunodeficiency virus type 1 integrase function in vitro. J. Virol. 66, 6361–6369.

Engelman, A., Bushman, F.D., Craigie, R., 1993. Identification of discrete functional domains of HIV-1 integrase and their organization within an active multimeric complex. EMBO J. 12, 3269–3275.

Engelman, A., Englund, G., Orenstein, J.M., Martin, M.A., Craigie, R., 1995. Multiple effects of mutations in human immunodeficiency virus type 1 integrase on viral replication. J. Virol. 69, 2729–2736.

Esnouf, R., 1997. Polyalanine reconstruction from Calpha positions using the program CALPHA can aid initial phasing of data by molecular replacement procedures. Acta Crystallogr., D Biol. Crystallogr. 53, 665–672.

Esposito, D., Craigie, R., 1998. Sequence specificity of viral end DNA binding by HIV-1 integrase reveals critical regions for protein–DNA interaction. EMBO J. 17, 5832–5843.

Gao, K., Gorelick, R.J., Johnson, D.G., Bushman, F., 2003. Cofactors for human immunodeficiency virus type 1 cDNA integration in vitro. J. Virol. 77, 1598–1603.

Gerton, J.L., Ohgi, S., Olsen, M., DeRisi, J., Brown, P.O., 1998. Effects of mutations in residues near the active site of human immunodeficiency virus type 1 integrase on specific enzyme–substrate interactions. J. Virol. 72, 5046–5055.

Goldgur, Y., Dyda, F., Hickman, A.B., Jenkins, T.M., Craigie, R., Davies, D.R., 1998. Three new structures of the core domain of HIV-1 integrase: an active site that binds magnesium. Proc. Natl. Acad. Sci. U.S.A. 95, 9150–9154.

Goodarzi, G., Im, G.J., Brackmann, K., Grandgenett, D., 1995. Concerted integration of retrovirus-like DNA by human immunodeficiency virus type 1 integrase. J. Virol. 69, 6090–6097.

Goodarzi, G., Pursley, M., Felock, P., Witmer, M., Hazuda, D., Brackmann, K., Grandgenett, D., 1999. Efficiency and fidelity of full-site integration reactions using recombinant simian immunodeficiency virus integrase. J. Virol. 73, 8104–8111.

Ha, H.C., Juluri, K., Zhou, Y., Leung, S., Hermankova, M., Snyder, S.H., 2001. Poly(ADP-ribose) polymerase-1 is required for efficient HIV-1 integration. Proc. Natl. Acad. Sci. U.S.A. 98, 3364–3368.

Heuer, T.S., Brown, P.O., 1998. Photo-cross-linking studies suggest a model for the architecture of an active human immunodeficiency virus type 1 integrase-DNA complex. Biochemistry 37, 6667–6678.

Hindmarsh, P., Ridky, T., Reeves, R., Andrake, M., Skalka, A.M., Leis, J., 1999. HMG protein family members stimulate human immunodeficiency virus type 1 and avian sarcoma virus concerted DNA integration in vitro. J. Virol. 73, 2994–3003.

Hindmarsh, P., Johnson, M., Reeves, R., Leis, J., 2001. Base-pair substitutions in avian sarcoma virus U5 and U3 long terminal repeat sequences alter the process of DNA integration in vitro. J. Virol. 75, 1132–1141.

Jenkins, T.M., Engelman, A., Ghirlando, R., Craigie, R., 1996. A soluble active mutant of HIV-1 integrase: involvement of both the core and carboxyl-terminal domains in multimerization. J. Biol. Chem. 271, 7712–7718.

Jenkins, T.M., Esposito, D., Engelman, A., Craigie, R., 1997. Critical contacts between HIV-1 integrase and viral DNA identified by structure-based analysis and photo-crosslinking. EMBO J. 16, 6849–6859.

Jones, K.S., Coleman, J., Merkel, G.W., Laue, T.M., Skalka, A.M., 1992. Retroviral integrase functions as a multimer and can turn over catalytically. J. Biol. Chem. 267, 16037–16040.

Ju, G., Boone, L., Skalka, A.M., 1980. Isolation and characterization of recombinant DNA clones of avian retroviruses: size heterogeneity and instability of the direct repeat. J. Virol. 33, 1026–1033.

Katzman, M., Sudol, M., 1995. Mapping domains of retroviral integrase responsible for viral DNA specificity and target site selection by analysis of chimeras between human immunodeficiency virus type 1 and visna virus integrases. J. Virol. 69, 5687–5696.

Khan, E., Mack, J.P., Katz, R.A., Kulkosky, J., Skalka, A.M., 1991. Retroviral integrase domains: DNA binding and the recognition of LTR sequences. Nucleic Acids Res. 19, 851–860.

Lubkowski, J., Yang, F., Alexandratos, J., Wlodawer, A., Zhao, H., Burke, T.R.J., Neamati, N., Pommier, Y., Merkel, G., Skalka, A.M., 1998a. Structure of the catalytic domain of avian sarcoma virus integrase with a bound HIV-1 integrase-targeted inhibitor. Proc. Natl. Acad. Sci. U.S.A. 95, 4831–4836.

Lubkowski, J., Yang, F., Alexandratos, J., Merkel, G., Katz, R.A., Gravuer, K., Skalka, A.M., Wlodawer, A., 1998b. Structural basis for inactivating mutations and pH-dependent activity of avian sarcoma virus integrase. J. Biol. Chem. 273, 32685–32689.

Lubkowski, J., Dauter, Z., Yang, F., Alexandratos, J., Merkel, G., Skalka, A.M., Wlodawer, A., 1999. Atomic resolution structures of the core domain of avian sarcoma virus integrase and its D64N mutant. Biochemistry 38, 13512–13522.

Moreau, K., Torne-Celer, C., Faure, C., Verdier, G., Ronfort, C., 2000. In vivo retroviral integration: fidelity to size of the host DNA duplication might Be reduced when integration occurs near sequences homologous to LTR ends. Virology 278, 133–136.

Moreau, K., Faure, C., Verdier, G., Ronfort, C., 2002. Analysis of conserved and non-conserved amino acids critical for ALSV (Avian leukemia and sarcoma viruses) integrase functions in vitro. Arch. Virol. 147, 1761–1778.

Moreau, K., Faure, C., Verdier, G., Ronfort, C., in press. Mutations in the C-terminal domain of ALSV (Avian Leukemia and Sarcoma Viruses) integrase alter the concerted DNA integration process in vitro. Eur. J. Biochem.

Puras Lutzke, R.A., Plasterk, R.H., 1998. Structure-based mutational analysis of the C-terminal DNA-binding domain of human immunodeficiency virus type 1 integrase: critical residues for protein oligomerization and DNA binding. J. Virol. 72, 4841–4848.

Roussel, A., Cambillau, C., 1989. TURBO-FRODO. In: Graphics, S. (Ed.), Silicon Graphics Geometry Partner Directory. Silicon Graphics, Mountain View, CA, pp. 77–78.

Shin, C.G., Taddeo, B., Haseltine, W.A., Farnet, C.M., 1994. Genetic analysis of the human immunodeficiency virus type 1 integrase protein. J. Virol. 68, 1633–1642.

Sinha, S., Pursley, M.H., Grandgenett, D.P., 2002. Efficient concerted integration by recombinant human immunodeficiency virus type 1 integrase without cellular or viral cofactors. J. Virol. 76, 3105–3113.

Thanos, D., Maniatis, T., 1992. The high mobility group protein HMG I(Y) is required for NF-kappa B-dependent virus induction of the human IFN-beta gene. Cell 71, 777–789.

Thompson, J.D., Higgins, D.G., Gibson, T.J., 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22, 4673–4680.

van Gent, D.C., Groeneger, A.A., Plasterk, R.H., 1992. Mutational analysis of the integrase protein of human immunodeficiency virus type 2. Proc. Natl. Acad. Sci. U.S.A. 89, 9598–9602.

van Gent, D.C., Vink, C., Groeneger, A.A., Plasterk, R.H., 1993. Complementation between HIV integrase proteins mutated in different domains. EMBO J. 12, 3261–3267.

Vercammen, J., Maertens, G., Gerard, M., De Clercq, E., Debyser, Z., Engelborghs, Y., 2002. DNA-induced polymerization of HIV-1 integrase analyzed with fluorescence fluctuation spectroscopy. J. Biol. Chem. 277, 38045–38052.

Vincent, K.A., Ellison, V., Chow, S.A., Brown, P.O., 1993. Characterization of human immunodeficiency virus type 1 integrase expressed in *Escherichia coli* and analysis of variants with amino-terminal mutations. J. Virol. 67, 425–437.

Vora, A.C., Grandgenett, D.P., 1995. Assembly and catalytic properties of retrovirus integrase-DNA complexes capable of efficiently performing concerted integration. J. Virol. 69, 7483–7488.

Vora, A., Grandgenett, D.P., 2001. DNase protection analysis of retrovirus integrase at the viral DNA ends for full-site integration in vitro. J. Virol. 75, 3556–3567.

Vora, A.C., McCord, M., Fitzgerald, M.L., Inman, R.B., Grandgenett, D.P., 1994. Efficient concerted integration of retrovirus-like DNA in vitro by avian myeloblastosis virus integrase. Nucleic Acids Res. 22, 4454–4461.

Vora, A.C., Chiu, R., McCord, M., Goodarzi, G., Stahl, S.J., Mueser, T.C., Hyde, C.C., Grandgenett, D.P., 1997. Avian retrovirus U3 and U5 DNA inverted repeats. Role Of nonsymmetrical nucleotides in promoting full-site integration by purified virion and bacterial recombinant integrases. J. Biol. Chem. 272, 23938–23945.

Wang, J.Y., Ling, H., Yang, W., Craigie, R., 2001. Structure of a two-domain fragment of HIV-1 integrase: implications for domain organization in the intact protein. EMBO J. 20, 7333–7343.

Yang, F., Roth, M.J., 2001. Assembly and catalysis of concerted two-end integration events by Moloney murine leukemia virus integrase. J. Virol. 75, 9561–9570.

Yang, Z.N., Mueser, T.C., Bushman, F.D., Hyde, C.C., 2000. Crystal structure of an active two-domain derivative of Rous sarcoma virus integrase. J. Mol. Biol. 296, 535–548.

Yoder, K.E., Bushman, F.D., 2000. Repair of gaps in retroviral DNA integration intermediates. J. Virol. 74, 11191–20000.

2ème Partie

Détermination de la structure cristallographique de la cellulase Cel5G isolée de la souche psychrophile *Pseudoalteromonas haloplanktis*

Etude bibliographique

Préambule :

La vie sur terre présente une immense capacité d'adaptation. Les limitations physiques compatibles avec la biologie vont en effet de – 40°C à + 115°C en température (respectivement dans la stratosphère et les fosses hydrothermales sous-marines) jusqu'à 120 MPa en pression (pression hydrostatique des fosses hadales) de 1 à 11 en pH et jusqu'à 4 M en concentration saline (Jaenicke et Böhm, 1998).

Les organismes vivant et prospérant dans de telles conditions extrêmes sont regroupés en différentes classes telles les thermophiles, les psychrophiles, les barophiles, les acidophiles et alkalophiles ou encore les halophiles. Ces organismes sont respectivement adaptés à des températures élevées (supérieures à 55°C) des températures basses (autour de 0°C) des pressions élevées, des conditions de pH acide ou basique et des concentrations salines importantes (Figure 1).

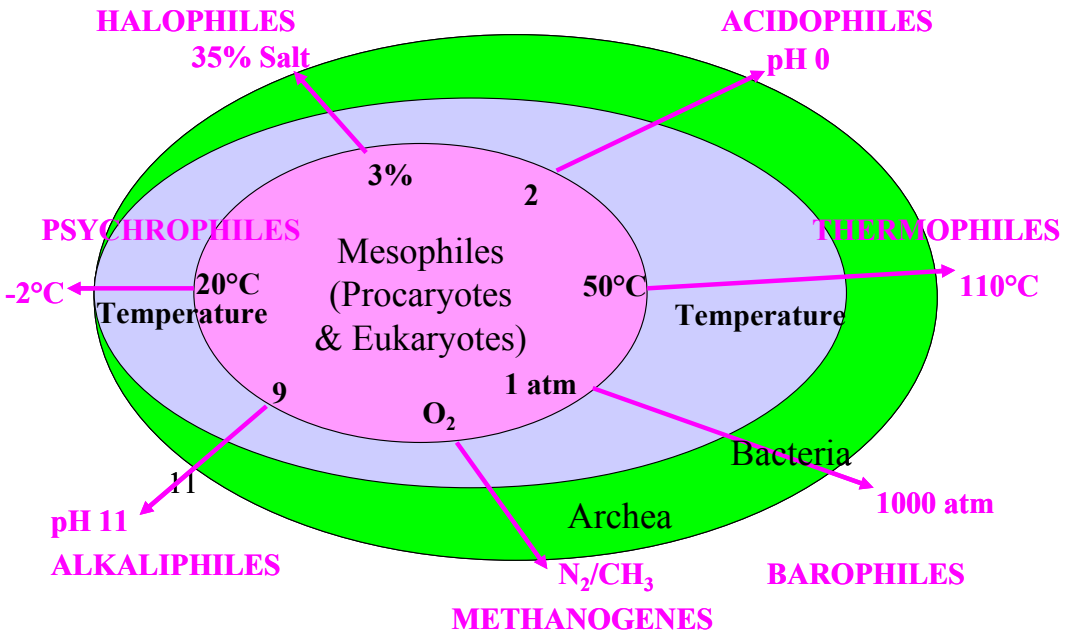


Figure 1 : *Schéma représentant les différentes classes d'extrêmophiles.*

A/ MICROORGANISMES PSYCHROPHILES ET «ENZYMES FROIDES»

1. Adaptations aux basses températures :

Les basses températures ont pour conséquence de diminuer considérablement la vitesse des réactions chimiques. L'équation d'Arrhenius[6] montre qu'une diminution de température entraîne une diminution exponentielle de la vitesse de réaction. Ainsi, une baisse de la température de 10°C induit une diminution par un facteur 2 à 3 de la vitesse des réactions chimiques. Les organismes psychrophiles doivent donc réussir à maintenir des vitesses réactionnelles adéquates, avec la réalisation des mécanismes enzymatiques nécessaires aux processi cellulaires essentiels à leur survie. Ceci est réalisé par la synthèse par ces organismes d'enzymes adaptées aux basses températures. Ces «enzymes froides» sont généralement caractérisées (i) par un k_{cat} et un k_{cat}/K_m supérieurs à leurs homologues mésophiles et thermophiles pour des températures comprises entre 0°C et 30°C, (ii) une stabilité thermique limitée se traduisant par une dénaturation à des températures moyennes, (iii) une courbe d'activité déplacée vers les basses températures par rapport à leurs homologues mésophiles (Feller *et al.*, 1996). L'hypothèse selon laquelle une flexibilité et une plasticité accrues de ces enzymes pourraient leur permettre des changements conformationnels à basse température durant la catalyse, est compatible avec les valeurs de leur efficacité catalytique et de leur «turn-over».

2. Activité, flexibilité, stabilité :

Contrairement aux organismes thermophiles, le problème majeur auquel les organismes psychrophiles sont confrontés n'est pas la stabilité mais l'efficacité catalytique de leurs enzymes (cf. § précédent).

La clef du fonctionnement des enzymes aux températures extrêmes semble donc résulter d'un compromis entre la flexibilité (permettant à l'enzyme d'assurer sa fonction catalytique à une vitesse raisonnable d'un point de vue métabolique) et la stabilité (permettant d'éviter toute

[6] $k = A.e^{-Ea/RT}$, avec k la constante de vitesse de la réaction considérée, A une constante, E_a l'énergie d'activation, R la constante des gaz parfaits (8,314 $kJ.mol^{-1}$) et T la température absolue en Kelvin.

dénaturation et ainsi d'assurer le maintien d'une conformation tridimensionnelle fonctionnelle).

A l'inverse des thermophiles, les enzymes psychrophiles pourraient augmenter la plasticité de leur édifice moléculaire (localement ou totalement) en diminuant le nombre ou la force des liaisons intramoléculaires. Ceci aurait pour conséquence de diminuer le coût énergétique des mécanismes d'interaction dû aux changements conformationnels nécessaires pour une interaction efficace avec le substrat (Zavodszky *et al.*, 1998). Leur structure serait donc moins compacte et maintenue par un faible nombre de liaisons intramoléculaires. Cependant, la conséquence d'une telle flexibilité serait une sensibilité accrue à la chaleur et à la dénaturation chimique (Feller et Gerday, 1997).

Jusqu'à présent, la corrélation entre la flexibilité conformationnelle et l'activité enzymatique est l'hypothèse la plus largement acceptée afin d'expliquer les propriétés catalytiques des enzymes psychrophiles (Gerday *et al.*, 1997 ; Zavodszky *et al.*, 1998).

Toutefois, des études récentes ont mis en évidence de nouvelles stratégies relatives à la thermostabilité. Ainsi, la résistance à la dénaturation thermique de la β-glycosidase de *Sulfolobus sulfataricus* (Aguilar *et al.*, 1997) serait acquise non pas grâce à une rigidité accrue mais grâce à son «élasticité» obtenue par l'enfouissement de molécules de solvants. Ces résultats contrastent avec l'hypothèse selon laquelle la thermostabilité est associée à la rigidité accrue de l'édifie moléculaire. De plus, des études de mutagenèse dirigée et d'évolution dirigée ont mis en évidence la possibilité d'augmenter la thermostabilité des enzymes sans perte d'activité spécifique, voire avec un gain d'activité (Miyazaki *et al.*, 2000; Narinx *et al.*, 1997) illustrant ainsi qu'il n'existe pas de rapport absolu entre ces deux caractéristiques. La faible thermostabilité des enzymes psychrophiles serait donc plutôt le résultat d'un manque de pression de sélection que celui des contraintes physiques ou chimiques imposées nécessaires à une activité catalytique efficace à basse température.

3. Déterminants structuraux de l'adaptation au froid :

La catalyse à basse température nécessiterait une plasticité accrue de l'édifice moléculaire, à l'inverse des thermophiles qui arborent cette flexibilité à leur optimum de température de fonctionnement. En fonction de la température de l'environnement dans lequel elles évoluent, la flexibilité des enzymes serait donc «ajustée», notamment *via* la modification

du nombre ou de la force de liaisons intramoléculaires.

Plusieurs structures cristallographiques d'enzymes bactériennes ou eucaryotes issues d'organismes psychrophiles ont été résolues : α-amylase (Aghajari *et al.*, 1998a ; Aghajari *et al.*, 1998b) triose phosphate isomérase (Alvarez *et al.*, 1998) citrate synthase (Russel *et al.*, 1998) malate dehydrogénase (Kim *et al.*, 1999) phosphatase alkaline (de Backer *et al.*, 2002) uracile-glycosylase (Leiros *et al.*, 2003) calcium-zinc protéase (Aghajari *et al.*, 2003) xylanase (Van Petegem *et al.*, 2003) et adénylate kinase (Bae et Phillips, 2004). Leur analyse permet de révéler que les résidus impliqués dans le mécanisme catalytique, ainsi que ceux pointant vers la cavité catalytique et / ou impliqués dans la liaison au substrat, sont généralement conservés par rapport à leur homologue mésophile. Cette observation corroborerait l'hypothèse selon laquelle les enzymes psychrophiles ne mettent pas en œuvre de nouveaux mécanismes catalytiques mais modifient celui existant afin de le rendre efficient à basse température. Ceci est probablement obtenu par un accroissement de la plasticité de la protéine. Ainsi, la première structure cristallographique obtenue pour une «enzyme froide», l'α-amylase de *Pseudoalteromonas haloplanktis* (Aghajari *et al.*, 1998a ; Aghajari *et al.*, 1998b) présente une flexibilité accrue de son édifice moléculaire. Le nombre réduit de résidus chargés en surface diminue le nombre des interactions par liaisons hydrogènes, conduisant à une plus grande flexibilité de la surface externe de la molécule. Ceci n'est cependant pas suffisant pour accroître la flexibilité globale de la protéine si le cœur de celle-ci reste rigide et compact. Cette étude montre donc qu'au moins quatre facteurs structuraux pourraient être impliqués dans cet accroissement de la flexibilité. Premièrement, le nombre réduit de résidus prolines au sein des boucles et des tours peut induire un degré de liberté supérieur aux segments reliant les éléments de structures secondaires. Deuxièmement, une diminution du nombre de résidus arginines réduit le nombre d'interactions électrostatiques responsables du maintien de l'édifice moléculaire. Troisièmement, cette structure est caractérisée par des contacts inter-domaines plus faibles, ceci étant dû sur la base de la comparaison avec ses homologues mésophiles, à une diminution du nombre de ponts salins, à l'absence d'un pont disulfure et à une coordination plus lâche de l'ion calcium présent entre les domaines A et B. Enfin, l'exposition au solvant de résidus hydrophobes par cette enzyme constitue certainement un facteur majeur de ce gain en flexibilité.

Des études menées sur diverses enzymes extrêmophiles (D'Amico *et al.*, 2001 ; Van den Burg *et al.*, 1998) ont montré que seuls quelques changements de conformation suffisent à modifier la stabilité des enzymes. Ainsi, la sélection naturelle pourrait rendre psychrophile (ou thermophile) une enzyme *via* un petit nombre de mutations, voire une seule (Merz *et al.*,

1999 ; Narinx *et al.*, 1997). Cependant, et d'après l'analyse des données structurales, la majorité des enzymes étudiées jusqu'à présent ont révélé de nombreuses modifications d'interactions faibles (Table 1). Tous ces facteurs ont été discutés par Feller et collaborateurs (Feller et Gerday, 1997 ; Feller et Gerday, 2003) et Gerday et collaborateurs (Gerday *et al.*, 1997).

Facteurs stabilisant/déstabilisant	Rôles/propriétés	Modification chez les psychrophiles
Liaisons ioniques		Diminution du nombre de ponts salins en surface ou au sein des structures secondaires
Liaisons hydrogènes	Contribution individuelle faible mais contribution globale importante compte tenu de leur nombre	Diminution du nombre de liaisons hydrogènes
Interactions hydrophobes	Formation de cluster au cœur de la protéine	Affaiblissement des clusters par substitution dans le cœur hydrophobe Augmentation des résidus hydrophobes en surface
Interactions aromatiques		Diminution du nombre des interactions
Interactions avec le solvant		Augmentation des résidus chargés en surface Augmentation des résidus hydrophobes en surface
Stabilisation dipolaire des hélices α	Stabilisation par des résidus de charges opposés aux extrémités	Diminution du nombre d'éléments stabilisant
Fixation d'ions	Stabilisation des structures	Perte de la fixation ou fixation moins forte
Résidus arginine	Stabilisation par formation de plusieurs ponts salins ou liaisons hydrogènes	Diminution du nombre de résidus arginine
Insertions et délétions		Insertions au niveau des boucles Délétions facilitant l'accès au site actif
Facteurs entropiques	Ponts disulfures et proline réduisent le degré de liberté conformationnels alors que glycine l'augmente	Diminution du nombre de résidus proline Augmentation du nombre de résidus glycine
Extrémités N et C-terminales	Sites d'initiation de déstructuration	Extrémités plus longues et plus flexibles

Table 1 : *Modifications moléculaires suspectées comme étant impliquées dans l'adaptation aux basses températures des «enzymes froides».*

Ainsi, chaque enzyme adopte une ou plusieurs de ces altérations structurales afin de s'adapter au mieux à la température de son environnement.

4. Intérêts et applications biotechnologiques ou industrielles :

Les enzymes sont déjà largement utilisées tant dans des applications industrielles que dans des produits ménagers tels les lessives. Mais presque toutes ces enzymes sont dérivées d'organismes qui vivent à des températures relativement élevées et agissent généralement plus efficacement à une température supérieure à 40° C.

Ainsi, les microorganismes psychrophiles constituent un potentiel énorme pour la découverte de «nouveaux» biocatalyseurs (Demirjian *et al.*, 2001 ; Gerday *et al.*, 2000 ; Russell, 1998). Etant donné la gamme de température à laquelle elles travaillent (0-20°C) les enzymes froides présentent une valeur économique potentiellement élevée et leur utilisation pourrait être envisagée dans des applications très diverses. De plus, leur thermolabilité accrue pourrait être exploitée pour arrêter un traitement ou pour inactiver sélectivement une (des) enzyme(s) dans un milieu réactionnel complexe. Dans l'industrie textile par exemple, l'utilisation de cellulases psychrophiles pour le délavage des jeans permettrait une diminution de la température de traitement et une diminution de la concentration nécessaire en enzymes. De plus, un contrôle des traitements par inactivation thermique de l'enzyme garantirait le maintien des propriétés mécaniques du produit final. Dans l'industrie alimentaire, le traitement de la nourriture à basse température permet d'éviter toute détérioration, de préserver les composés volatils et labiles et de prévenir toute contamination par des microorganismes mésophiles. Mais l'un des plus grands marchés reste probablement celui des détergents, le lavage à basse température étant synonyme de gain d'énergie et d'argent.

Cependant, le transfert technologique de ces «enzymes froides» vers le secteur industriel pourrait prendre plusieurs années. Leur coût de production est en effet un facteur primordial à prendre en compte en vue de leur mise sur le marché. Les difficultés d'adaptation de techniques de production satisfaisantes et la réticence des industriels à remplacer des systèmes qui fonctionnent déjà raisonnablement bien sont autant de facteurs qui pourraient ralentir le développement industriel de ces «enzymes froides».

B/ CELLULOSE ET CELLULASES

La cellulose est un composé présent chez de nombreux organismes (bactéries, champignons) et plus particulièrement dans le règne végétal où elle constitue le principal composant de la paroi cellulaire. Ainsi, la cellulose peut-être considérée comme le polymère le plus abondant sur terre, sa synthèse par les plantes étant estimée à 4 x 10^7 tonnes / an (Singh et Hayashi, 1995).

1. La cellulose :

La cellulose est un homopolymère linéaire composé de sous-unités de glucose reliées entre elles par des liaisons glucosidiques β-1,4 dont le degré de polymérisation varie entre 100 et 14 000. Les propriétés physico-chimiques de la cellulose (insolubilité, structure rigide) responsables de sa résistance naturelle à la dégradation sont très différentes de celles de l'amidon, un autre polymère de glucose aux liaisons glucosidiques α-1,4. Ces différences peuvent s'expliquer par le rôle de l'amidon de stockage de l'énergie, qui peut être relibéré sous forme de glucose par son hydrolyse, et par le rôle structural de la cellulose, qui confère à la paroi des cellules végétales les qualités mécaniques nécessaires pour résister au stress osmotique ou mécanique. D'un point de vue structural, la nature même de la liaison glucosidique β-1,4 implique une conformation étendue du polymère de cellulose, chaque résidu glucose étant orienté à 180° par rapport au résidu voisin. Ainsi, l'unité répétitive de la cellulose n'est pas le glucose, mais le cellobiose (Figure 2A). Des liaisons hydrogène intra-moléculaires entre résidus adjacents confèrent rigidité à la cellulose. En fait, deux conformations stables de la cellulose ont pu être observées (Atalla et Van der Hart, 1984). Dans la première conformation dite *k1* (Figure 2Aa) l'oxygène O6 ainsi que l'oxygène du cycle d'un même résidu sont donneurs d'électrons pour le proton de l'hydroxyle en C3 d'un résidu adjacent. La conformation *k2* (Figure 2Ab) ne présente qu'une interaction hydrogène simple entre l'oxygène du cycle d'un résidu et le proton de l'hydroxyle en C3 d'un résidu adjacent.

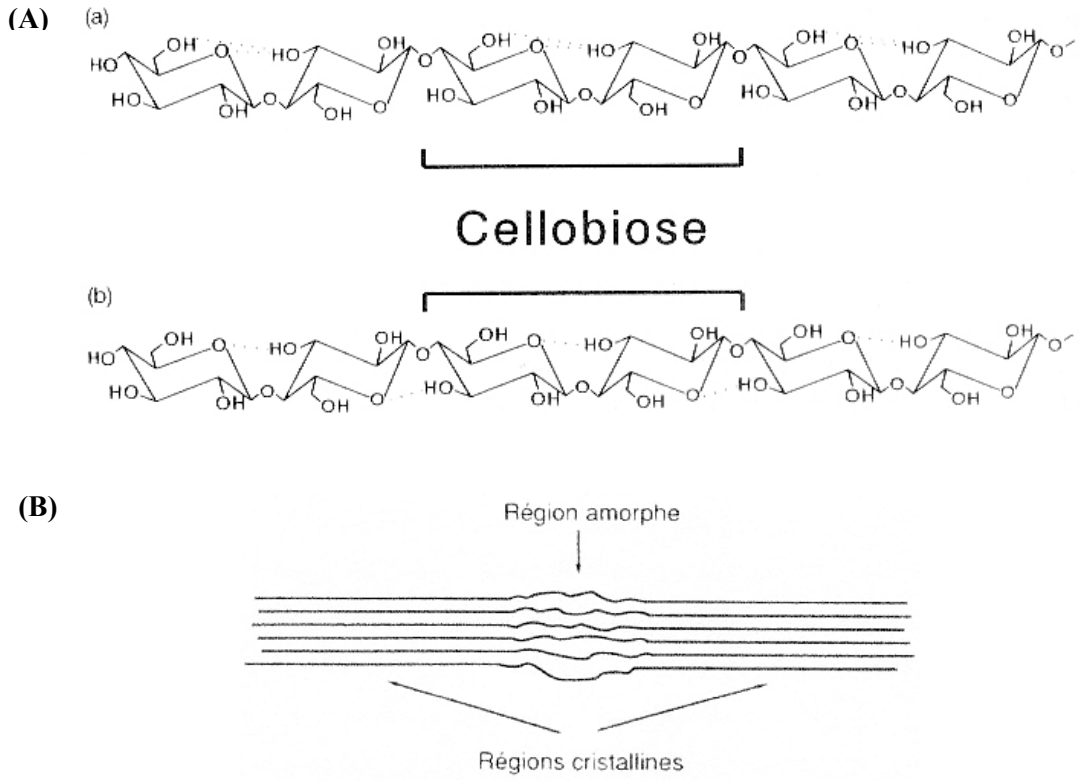


Figure 2 : *Structure de la cellulose (A) : composition et liaisons hydrogène de la cellulose k1 (a) et k2 (b) ; (B) : schéma d'une microfibrille de cellulose.*

A l'état naturel, les chaînes de cellulose s'associent généralement en microfibrilles stabilisées par des liaisons hydrogène intermoléculaires. L'organisation de ces microfibrilles reste cependant très complexe et n'est pas encore totalement élucidée. Elles présentent en effet soit des zones cristallines, majoritairement composées de chaînes de cellulose en conformation *k1*, soit des zones amorphes composées à la fois de cellulose en conformation *k1* et *k2* (Figure 2B). Les parties cristallines ont un rôle structural alors que les parties amorphes ont des propriétés viscoélastiques.

Le degré de cristallinité, le degré de polymérisation et la largeur des microfibrilles varient selon leur origine et leur âge. Ainsi, la cellulose gonflée à l'acide est entièrement amorphe, alors que celle isolée de l'algue *Valonia macrophysa* est presque entièrement cristalline (Henrissat, 1985). La cellulose commerciale telle l'Avicel présente un degré de cristallinité de 47 % environ et celle du coton de70 % (Wood, 1998). Dans les parois végétales, la cellulose de la paroi primaire présente un degré de polymérisation relativement faible et est polydisperse (DP(2000) à DP(6000)) alors que celui de la paroi secondaire est plus élevé et plus homogène (DP(14000)).

2. Les cellulases :

Dans les milieux naturels, la dégradation de la cellulose résulte essentiellement de l'action de microorganismes, bactéries ou champignons.
Cependant, ces organismes ne possèdent pas d'enzymes capables d'hydrolyser à elles seules la cellulose native (Singh et Hayashi, 1995). Ainsi, tous les systèmes cellulolytiques efficaces mettent en jeu plusieurs enzymes de spécificités différentes, organisées ou non en complexes protéiques structurés, tels les cellulosomes (Bayer *et al.*, 1998a). D'autre part, l'extrême insolubilité de la lignocellulose et de la cellulose exige que les enzymes capables de les dégrader soient sécrétées dans le milieu ou exposées à la surface de la cellule.

2.a) Les différents types de cellulases :

Sous le terme général de cellulase est désignée une large classe d'enzymes capables d'hydrolyser à différents degrés, divers substrats cellulosiques et apparentés contenant des liaisons glucosidiques β-1,4. Les hydrolases constituent la grande majorité des cellulases, mais on trouve également des enzymes coupant les liaisons β-1,4 par phosphorolyse (phosphorylases).

Trois classes majeures d'enzymes sont distinguées parmi les cellulases : les endoglucanases, les cellobiohydrolases et les β-glucosidases. Leur mode d'action sur la cellulose est schématisé sur la figure 3.

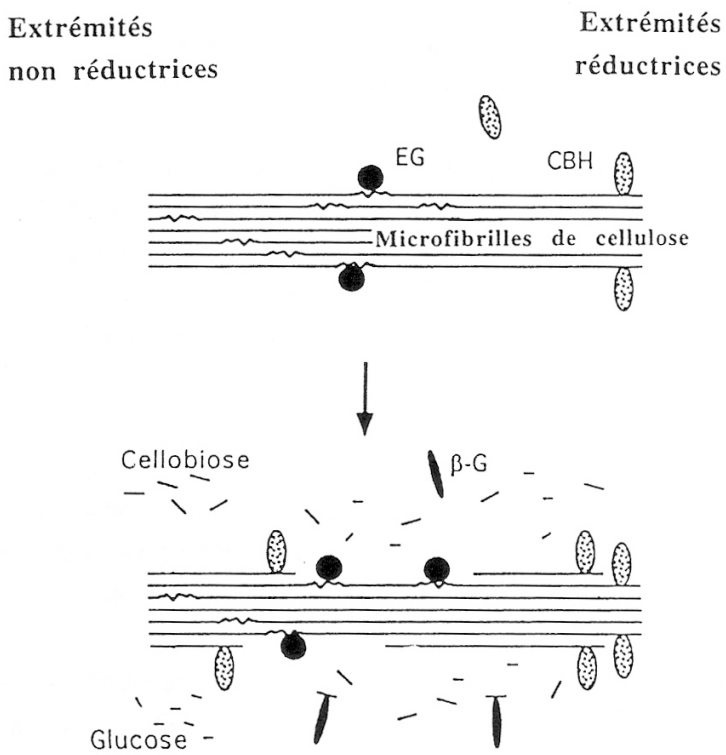


Figure 3: *Mode d'action des différents types de cellulases proposé par Wood et McCrae (1979): les endoglucanases (EG) en attaquant les régions amorphes de la cellulose, produisent de nouveaux sites d'action pour les cellobiohydrolases (CBH) et les β-glucosidases (β-G) éliminent le cellobiose, inhibiteur des cellobiohydrolases.*

- Les **endoglucanases** (EG) ou endo-β-1,4-D-glucane glucanohydrolase (E.C. 3.2.1.4) hydrolysent de façon aléatoire les liaisons β-1,4 glycosidiques situées à l'intérieur des chaînes de cellulose amorphe. En général, ces enzymes dégradent exclusivement la cellulose amorphe et possèdent une activité importante vis-à-vis de la cellulose soluble, cette activité étant d'autant plus élevée que le degré de polymérisation est élevé. Les endoglucanases sont très actives sur la Carboxy-Methyl-Cellulose soluble (CMC) alors que leur activité reste très faible sur la cellulose cristalline. L'attaque par des EGs engendre de nouvelles extrémités non réductrices, cibles d'hydrolyse pour les cellobiohydrolases (CBHs).

- Les **cellobiohydrolases** (CBHs) ou exo-β-l,4- D-glucane cellobiohydrolase (E.C. 3.2.1.91) ou encore exoglucanases dégradent séquentiellement la cellulose à partir de l'extrémité non réductrice de la chaîne, en libérant dans la plupart des cas du cellobiose. On parle d'exo-β-l,4-D-glucosidases ou glucohydrolases (E.C. 3.2.1.74) lorsqu'il y a libération de glucose.

Contrairement aux EGs, les CBHs sont généralement actives vis-à-vis de la cellulose cristalline (Avicel, fibres de coton) mais bien plus efficacement en synergie avec des endoglucanases). En revanche, elles ne sont en général pas actives vis-à-vis de la CMC.

Cette distinction entre endoglucanases et cellobiohydrolases n'est cependant pas absolue et il est préférable de considérer que certaines enzymes présentent une activité préférentiellement endo tandis que d'autres présentent une activité préférentiellement exo.

- Les **β-glucosidases** ou β-D-glucoside glucohydrolases, ou encore cellobiases (EC 3.2.1.21), hydrolysent le cellobiose en glucose. Les cellobiases ne dégradent pas directement la cellulose, mais empêchent une rétroinhibition des cellobiohydrolases par le cellobiose (Gilbert et Hazlewood, 1993).

Grâce au clonage et au séquençage d'un grand nombre de gènes, un nombre élevé de séquences de cellulases sont actuellement connues. L'analyse de ces séquences montre que la majorité d'entre elles sont composées de segments que l'on retrouve sous une forme plus ou moins similaire dans plusieurs cellulases et qui peuvent être recombinés dans des ordres différents. Ces segments sont fréquemment reliés entre eux par des éléments de séquence riches en sérine et thréonine appelés «linkers». Cette organisation modulaire est également constatée chez d'autres enzymes hydrolysant des polymères glucidiques insolubles, comme les amylases ou les chitinases. Des expériences de protéolyse ou de délétion génétique ont montré que les segments identifiables par alignement de séquences représentent le plus souvent des domaines indépendants d'un point de vu structural et fonctionnel (Bayer *et al.*, 1998b). Plusieurs types de domaines ont été identifiés :

Domaines catalytiques

Toutes les cellulases possèdent un domaine catalytique, dont la séquence détermine la famille à laquelle l'enzyme appartient. Ces domaines comportent de 300 à 600 résidus environ. Bien que les cas de similarité fortes (>50 %) soient peu fréquents, la comparaison des séquences par la technique dite HCA (Hydrophobic Cluster Analysis ; Gaboriaud *et al.*, 1987) a montré que toutes les cellulases actuellement connues appartiennent à 12 des 66 familles distinctes de glycoside-hydrolases (http://www.expasy.org/cgi-bin/lists?glycosid.txt).

Deux conclusions peuvent être tirées de l'examen des enzymes faisant partie de chaque famille : premièrement, la corrélation est loin d'être absolue entre la similarité de séquences des protéines et la parenté phylogénétique des organismes qui les produisent. La famille 7 (ou C) contient des enzymes issues d'organismes eucaryotes uniquement, tandis que les enzymes des familles 8 et 48 (D et L) sont toutes bactériennes. Certaines familles (5, 6, 9, 12 et 45) présentent cependant des enzymes de diverses origines. Il existe ainsi des similitudes frappantes entre des EGs provenant de bactéries, de plantes ou de *Dictyostelium discoideum* (famille 9). De même, les membres d'une famille peuvent issus de champignons et de bactéries, d'organismes aérobies et anaérobies, de mésophiles et de thermophiles.

De plus au sein d'une famille, la taxonomie basée sur la séquence n'est pas toujours en accord avec la phylogénie bactérienne, impliquant un transfert horizontal extensif des gènes. Ainsi, certaines familles (5 et 8) contiennent des gènes à la fois d'organismes Gram + et Gram -. Ceci suggère qu'au cours de l'évolution, les différents organismes cellulolytiques se sont constitués un jeu de cellulases adaptées à leurs besoins à partir d'un pool de gènes ancestraux. Ces gènes se seraient répandus non seulement par transmission verticale, mais aussi sur une large échelle par transfert horizontal (domaine «shuffling»).

Deuxièmement, l'appartenance d'une cellulase à l'une ou l'autre famille ne permet pas de définir de façon stricte ses propriétés enzymatiques. Certaines corrélations peuvent bien entendu être observées: les enzymes des familles 5, 8, 9, 12 et 45 (A, D, E, H et K) sont essentiellement des EGs et, pour toutes les enzymes d'une même famille, la stéréochimie de la réaction (inversion ou rétention de la configuration) est identique. Trois familles 6, 7 et 48 (B, C et L) comportent cependant à la fois des EGs et des CBHs. De plus, on observe souvent à l'intérieur d'une même famille des variations importantes de la spécificité de substrat (activité relative en fonction du degré de polymérisation des cellodextrines, site de clivage de celles-ci, activité secondaire sur des polymères apparentés à la cellulose tels que le xylane ou la chitine par exemple).

Modules de liaison aux sucres (CBM)

De nombreuses études ont établi la présence dans la majorité des cellulases et dans plusieurs hémicellulases, d'un domaine indépendant du domaine catalytique et responsable de la fixation à la cellulose (Gilkes *et al.*, 1991). Ce type de domaine appelé CBM accroît généralement l'activité vis à vis de la cellulose cristalline. Ce domaine peut être localisé à l'extrémité C ou N-terminale de l'enzyme et est généralement séparé du domaine catalytique

par un domaine de liaison ou «linker». Certaines enzymes peuvent présenter plusieurs CBMs (Irwin *et al.*, 1998) tandis que d'autres n'en ont pas (Henriksson *et al.*, 1999).
La suppression de ce domaine par protéolyse génère des domaines catalytiques qui conservent leur activité vis-à-vis de substrats amorphes et solubles, mais perdent la majeure partie, voire la totalité, de l'activité de l'holoenzyme vis-à-vis de la cellulose cristalline (Tomme *et al.*, 1995).
Le rôle du CBM dans l'amélioration de l'activité enzymatique vis à vis de substrats insolubles n'est pas encore bien compris, mais deux hypothèses sont généralement proposées (Linder et Teeri, 1997 ; Reinikainen *et al.*, 1992) :

(i) ils pourraient servir à augmenter la concentration locale des cellulases à la surface du substrat et de ce fait, contribuer à accroître leur efficacité.

(ii) ils pourraient contribuer à la déstructuration du substrat et faciliter alors, son accessibilité aux enzymes de dégradation. Le CBD agirait comme un coin ou un rabot pour détacher les chaînes de cellulose du réseau cristallin. Plutôt qu'une dislocation des cristaux de cellulose, il s'agirait alors d'une rupture des interactions faibles existantes entre les microfibrilles (Din *et al.*, 1994 ; Gilkes *et al.*, 1993).

Tous ces domaines furent initialement classés en familles (type I à IX) différentes des familles de domaine catalytique, sur la base de similarités de séquence et de longueur de chaîne polypeptidique (Tomme *et al.*, 1995). Près d'une centaine de CBM furent ainsi classés en neuf types. Actuellement, près d'un millier de CBMs sont connus, non seulement chez des endo- et exo- glucanases, mais aussi chez d'autres glycoside-hydrolases ou protéines liant des sucres (cellulose, xylose, mannose, chitine, amidon,…). Ces CBMs sont classés en 42 familles, les familles 1 à 13 correspondant aux types I-XIII de l'ancienne classification en temps que Cellulose Binding Domain (CBDs). Cette classification peut être consultée sur le site http://afmb.cnrs-mrs.fr/CAZY/index.html.

Contrairement aux domaines catalytiques, les CBMs forment des familles avec une affiliation plus stricte à certains groupes taxonomiques. Ainsi, les CBMs fongiques appartiennent uniquement à la famille 1 Les CBMs de la famille 2 sont essentiellement bactériens et généralement situés à l'extrémité N-terminale alors que ceux de la famille 3 sont principalement issus de bactéries produisant des cellulosomes (Linder et Teeri, 1997). Mais les propriétés des CBMs peuvent varier au sein d'une même famille notamment en ce qui concerne leur affinité pour le substrat. Les CBMs de champignons, constitués de 30 à 40 résidus, sont très différents des CBMs de bactéries présentant généralement 100 à 150

résidus.

L'analyse des séquences de CBM a permis de mettre en évidence un certain nombre de caractéristiques communes : ils contiennent peu de résidus chargés, ils sont riches en résidus hydroxylés (sérine, thréonine) et aromatiques (tyrosine, tryptophane) et possèdent généralement des résidus cystéine à leurs extrémités. A l'intérieur d'une famille, les positions des résidus aromatiques (tyrosine, tryptophane) sont très conservées. Des structures tridimensionnelles de CBM sont actuellement disponibles pour plusieurs familles. La structure tridimensionnelle du CBD de Cel5A *d'Erwinia chrysanthemi* (famille 5 ; Brun *et al.*, 1997) comporte ainsi principalement des brins β et deux résidus tryptophanes qui pourraient participer à la liaison à la cellulose.

Domaines de liaison («linker»)

Les séquences, qui constituent généralement une région charnière entre domaines catalytiques et CBM, sont présentes chez de nombreuses cellulases et xylanases (Gilkes *et al.*, 1991). La plupart de ces régions sont riches en proline, sérine, thréonine, alanine et glycine. Ces régions charnières sont souvent O-glycosylées (Ong *et al.*, 1994). Il est probable que cette O-glycosylation induise une plus grande stabilité dans un environnement aqueux et protège les enzymes de la protéolyse (Langsford *et al.*, 1987).

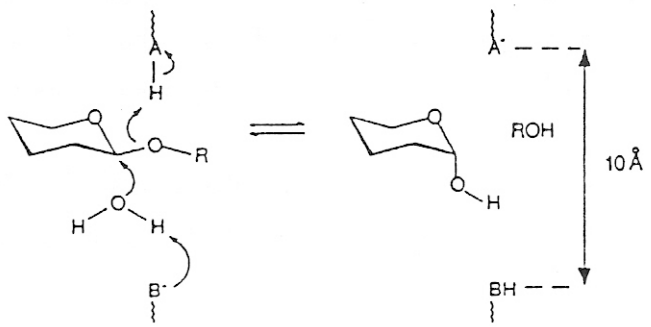
Le rôle des régions charnières serait de permettre une certaine flexibilité entre les domaines fonctionnels tout en les maintenant dans un positionnement correct les uns par rapport aux autres, ceci afin d'assurer une efficacité catalytique optimale. Ces régions charnières ne semblent pas indispensables à l'activité, mais dans certains cas, elles semblent nécessaires pour une activité optimale. C'est notamment ce qui sera développé et discuté dans la suite de ce manuscrit grâce à la Publication 3.

2.b) Mécanisme catalytique :

Il est généralement admis que l'hydrolyse des liaisons glycosidiques par les cellulases procède selon un mécanisme acide / base, nécessitant deux résidus (en général, Asp et / ou Glu) : un donneur de proton et un nucléophile / base (Henrissat et Davies, 1997). L'hydrolyse de la liaison glycosidique s'effectue *via* deux mécanismes majeurs, donnant lieu soit à une inversion soit à une rétention de la configuration anomérique (Koshland, 1953 ; Sinnott,

1990). Le détail de ces deux mécanismes est présenté dans la figure 4. Le positionnement du donneur de proton est identique dans les deux cas, alors que celui du nucléophile diffère. Dans les enzymes procédant par inversion de configuration, la distance qui sépare le nucléophile du donneur de proton est plus importante que dans les enzymes procédant par rétention de configuration puisqu'une molécule d'eau doit être logée entre la base et le sucre. La distance moyenne entre les groupements carboxyliques des deux résidus catalytiques est de 5,5 Å environ pour le mécanisme de rétention (Figure 4b). Cette distance peut varier pour le mécanisme d'inversion mais elle se situe plus fréquemment aux environs de 10 Å (Figure 4a).

a) **Mécanisme d'inversion de configuration**



b) **Mécanisme de rétention de configuration**

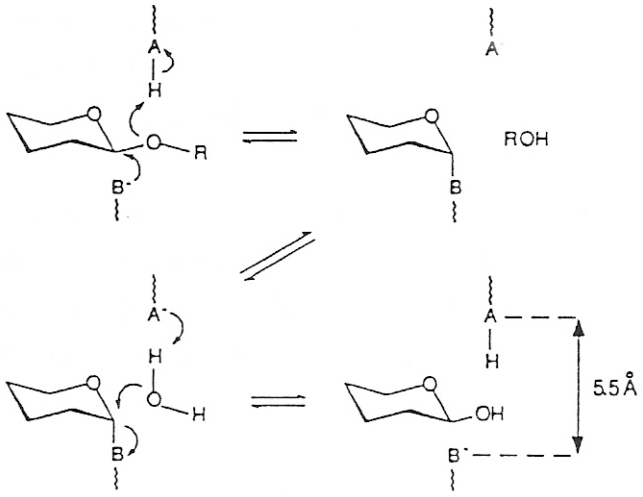


Figure 4 : *Mécanismes d'hydrolyse des liaisons β-l,4 glycosidiques :*

a) L'oxygène de la liaison β-glycosidique est protoné par un groupement carboxylique, appelé catalyseur acide / base (K). Simultanément, un groupement carboxylate, appelé catalyseur nucléophile (B^-) et situé de l'autre côté de la liaison à couper, permet l'ionisation d'une molécule d'eau et la formation d'un anion hydroxyle. Cet anion se

substitue au groupe partant du carbone anomérique. Cette substitution nucléophile simple conduit à l'inversion de la configuration du carbone anomérique.

b) Le départ du groupement partant est également induit par la protonation de la liaison glycosidique, mais dans ce cas, le groupement nucléophile est suffisamment proche de la liaison glycosidique pour agir directement comme substituant nucléophile. Ceci conduit à la formation d'un intermédiaire glycosyl / enzyme. Cet intermédiaire est ensuite hydrolysé au cours d'une deuxième substitution par l'ion hydroxyle d'une molécule d'eau dont l'ionisation est promue par la forme basique du résidu catalytique acide / base. La double inversion impliquée dans ce mécanisme se traduit, du point de vue stéréochimique, par la rétention de la configuration du carbone anomérique.

Un mécanisme général pour les glycosidases retenant la configuration du carbone anomérique est maintenant bien reconnu : mécanisme à double déplacement et formation d'un intermédiaire covalent glycosyl / enzyme (McCarter et Withers, 1994). Une paire d'acides aminés carboxyliques (distants de 5Å) est trouvée dans le site actif : l'un agit comme catalyseur acide (donneur de proton) tandis que l'autre se comporte comme nucléophile (base) favorisant la stabilisation d'un ion oxocarbonium (*via* la charge négative) et la diffusion du groupe partant.

2.c) Classification :

La comparaison des séquences des cellulases par les méthodes classiques ne détecte généralement d'homologies significatives qu'entre cellulases d'un même organisme. Seules quelques parentés entre champignons et bactéries ont pu être mises en évidence.

En 1989, Henrissat et collaborateurs ont utilisé une nouvelle méthode de comparaison des séquences d'acides aminés exploitant la méthode d'analyse des amas hydrophobes (Hydrophobic Cluster Analysis ou HCA ; Gaboriaud *et al.*, 1987). Ils ont ainsi développé un schéma logique de classification en familles. L'HCA permet de détecter des régions homologues au niveau de la structure tertiaire de protéines présentant une faible identité de séquence d'un point de vue composition et longueur. Grâce à cet outil, une série de similarités entre des cellulases bactériennes et fongiques ont pour la première fois été mise en évidence et une classification des domaines catalytiques des cellulases en 6 familles (A-F) a été établie.

Avec l'arrivée de nouvelles séquences et de nouvelles méthodes de comparaison de

séquences, une nouvelle classification combinant HCA et homologie de séquence a permis la classification des glycoside-hydrolases en 35 familles (Henrissat, 1991).

Une nouvelle classification a été établie afin de : (i) refléter les caractéristiques structurales des enzymes plutôt que leur spécificité de substrat uniquement, (ii) aider à révéler l'histoire évolutive des enzymes et (iii) fournir un outil pour prédire le mécanisme d'action de nouvelles enzymes.

Actuellement, la classification comporte 70 familles (74 avant suppression des familles 21, 40, 41 et 60) et plus une centaine d'enzymes non classifiées (Henrissat et Davies, 1997 ; http://www.expasy.ch/cgi-bin/lists?glycosid.txt). Au moins 10 de ces familles contiennent des cellulases (5-9, 12, 44, 45, 48 et 61 + 1 et 3 si on inclut également les β-glucosidases).

La structure tridimensionnelle des protéines étant mieux conservée que leurs séquences, certaines familles peuvent être regroupées en clans, soit quand la sensibilité des méthodes de comparaison de séquences révèle de nouvelles homologies ou quand les déterminations structurales démontrent une ressemblance entre des membres de diverses familles (Henrissat et Bairoch, 1996) partageant un même repliement tridimensionnel, une même architecture du site actif et un même mécanisme catalytique. Huit clans ont actuellement été définis. Ceux-ci peuvent contenir des enzymes hydrolysant une variété de substrats différents. Le clan GH-A est actuellement le plus important, avec plus de 250 membres représentant au moins 18 spécificités de substrat différentes.

Une superfamille 4/7 a également été constituée. Elle regroupe des enzymes chez lesquelles l'architecture en tonneau $(\beta/\alpha)_8$ et le mécanisme catalytique avec rétention de configuration sont conservés, impliquant la présence d'un résidu Glu à la fin des feuillets β4 (donneur de proton) et β7 (nucléophile). Cette superfamille inclut des enzymes sans homologie de séquence détectable, ainsi que des enzymes présentant des activités différentes.

2.d) Structure tridimensionnelle et mécanisme de dégradation :

La cristallisation de diverses cellulases a permis d'établir la structure 3D des domaines catalytiques (Chapon *et al.*, 2001 ; Ducros *et al.*, 1995 ; Parsiegla *et al.*, 1998 ; Sulzenbacher *et al.*, 1996) et de certains CBDs (Mattinen *et al.*, 1998 ; Tormo *et al.*, 1996) ce qui ouvre des possibilités de modélisation pour les cellulases faisant partie des mêmes groupes.

Les cellulases peuvent être subdivisées en deux catégories selon leur mécanisme de dégradation de la cellulose : (i) les enzymes non processives aussi appelées endocellulases et (ii) les enzymes processives, comprenant les exocellulases et les nouvelles endocellulases.

Ces dernières restent attachées à la chaîne de cellulose, libérant principalement des unités cellobiose ou cellotétraose. Le fait de rester attachée au substrat est dû à des différences structurales entre ces enzymes et les enzymes non processives.

La géométrie du site actif des diverses enzymes fournit donc une explication élégante de leur spécificité endo ou exo :

- Dans le cas des endocellulases, le substrat pénètre dans un sillon ouvert. Cette **crevasse** (Figure 5) constitue un site actif ouvert en surface de la molécule pouvant accommoder la cellulose n'importe où le long de sa chaîne, ceci en accord avec leur mode d'action endo (au hasard de la chaîne de cellulose amorphe).

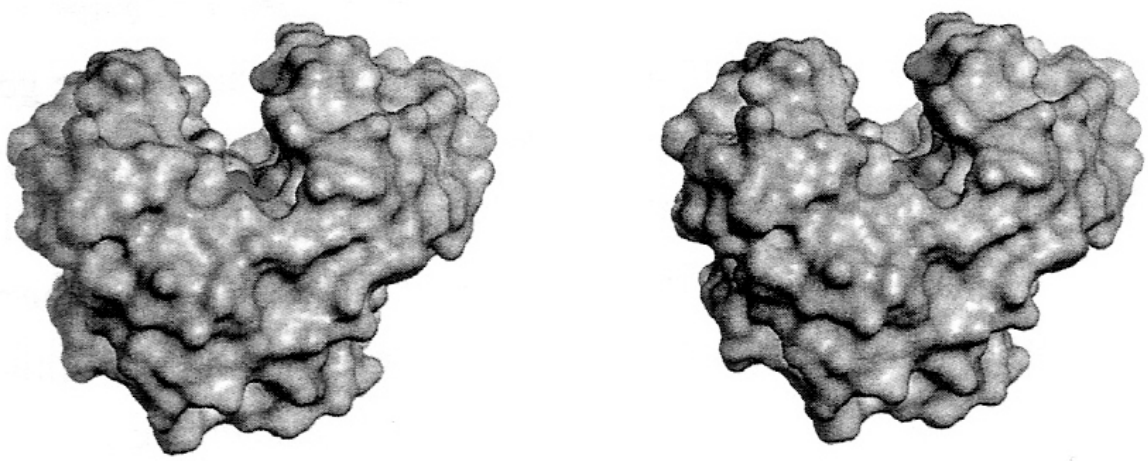


Figure 5 : *Représentation en surface du site catalytique au fond d'une crevasse de l'endoglucanase E2 de T. fusca (d'après Davies et Henrissat, 1995).*

- En ce qui concerne les exoenzymes, la topologie du site actif est généralement une poche en forme de tunnel. Ce **tunnel** est formé par de longues boucles en surface de la molécule qui viennent recouvrir presque parfaitement le sillon catalytique (Figure 6).

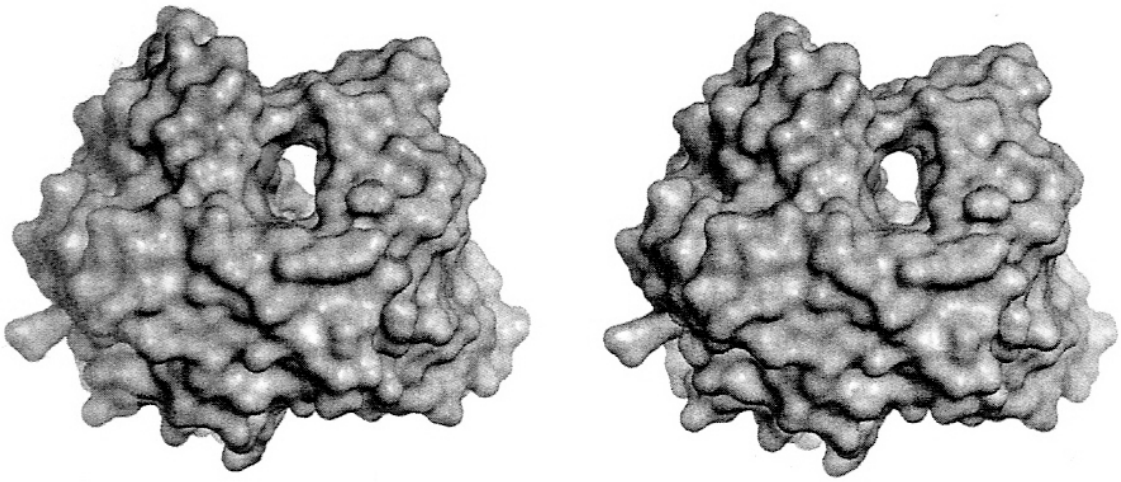


Figure 6 : *Représentation en surface du site catalytique en tunnel de la cellobiohydrolase II de T. reesei (d'après Davies et Henrissat, 1995).*

- Pour les CBHs, le site actif est généralement logé dans une **poche** parfaitement adaptée pour la reconnaissance des extrémités non réductrices des sucres (Figure 7). La chaîne de cellulose cristalline doit ainsi se faufiler à partir de son extrémité non réductrice, permettant uniquement des attaques à cette extrémité de la chaîne cellulosique. On parle alors d'enzymes processives, c'est-à-dire pouvant effectuer des clivages successifs.

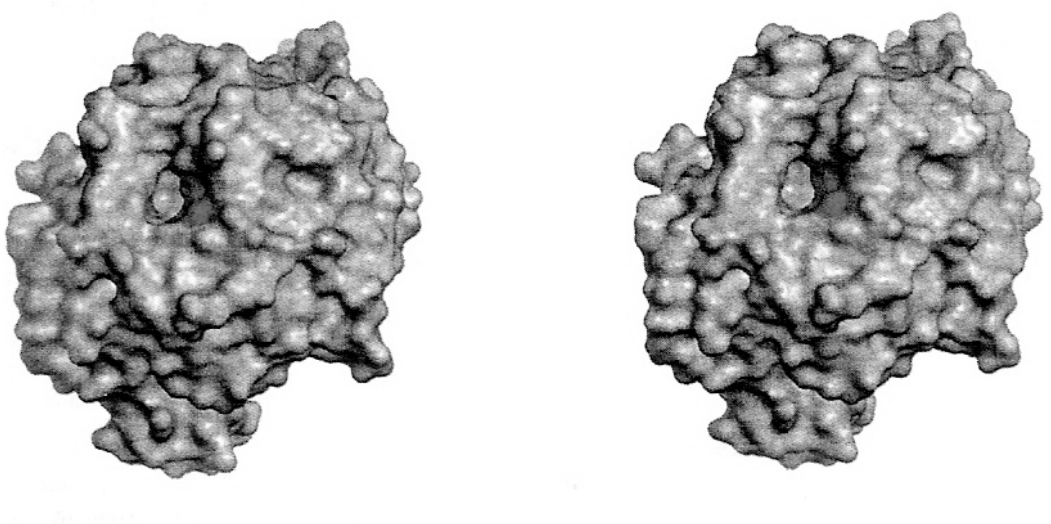


Figure 7 : *Représentation en surface du site catalytique au fond d'une poche de la glucoamylase d'Aspergillus awamor (d'après Davies et Henrissat, 1995).*

De nouvelles endocellulases processives ont été mises récemment en évidence. Ainsi, l'endocellulase CelF de *Clostridium cellulolyticum* (Parsiegla *et al.*, 1998) présente une

structure mixte avec un tunnel à une extrémité et une crevasse à l'autre extrémité, le site actif se situant à la jonction des deux régions (Figure 8).

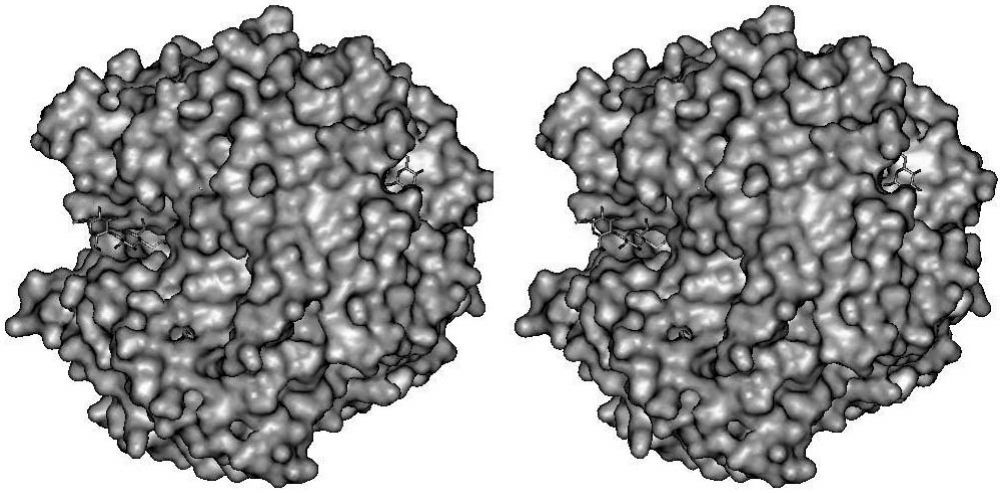


Figure 8 : *Représentation en surface de Cel48F de Clostridium cellulolyticum montrant le substrat dans la crevasse et dans le tunnel.*

L'enzyme digère la cellulose en exerçant une attaque endo à un endroit exposé de la chaîne (comme une endocellulose) puis continue par un clivage processif libérant des unités cellobiose depuis l'extrémité non réductrice d'une des chaînes de cellulose nouvellement produites.

Résultats et discussion

A/ ORIGINE DE LA SOUCHE *PSEUDOALTEROMONAS HALOPLANKTIS*

La bactérie psychrophile *Pseudoalteromonas haloplanktis* a été isolée d'eau de mer dans la région côtière de la station française Antarctique de J.S. Dumont d'Urville (66°40' S, 140°01' E) en terre Adélie (Feller *et al.*, 1992). Celle-ci a été sélectionnée pour son activité cellulolytique.

La souche *Pseudoalteromonas haloplanktis* TAB23 est une bactérie Gram négative. Cette souche isolée en Antarctique dans un milieu dont la température est en moyenne de -1°C, -1,5°C, est capable de se développer de 0°C à 25°C environ. Sa limite de croissance étant de 25°C, *Pseudoalteromonas haloplanktis* TAB23 peut être considérée comme une souche psychrophile.

Pseudoalteromonas haloplanktis secrète l'endoglucanase Cel5G appartenant à la famille 5-2 des glycoside-hydrolases. Cette enzyme modulaire est composée d'un domaine catalytique de 292 résidus (dénommé ci-après $Cel5G_{CM}$) d'un domaine de liaison («linker» de 109 résidus et d'un module de fixation aux sucres de la famille 5 (CBM) situé à l'extrémité C-terminal de l'enzyme et composée de 61 résidus (Figure 9).

Figure 9 : *Représentation schématique des endoglucanases Cel5G de Pseudoalteromonas haloplanktis et de Cel5A de Erwinia chrysanthemi. CBM (carbohydrates bniding module) : module de fixation aux sucres.*

Cel5G présente une homologie de séquence maximale avec l'endoglucanase Cel5A de *Erwinia chrysanthemi*, tant pour le domaine catalytique que pour le CBM. L'identité de séquence est en effet de 64 % (similarité de 79 %) pour leur domaine catalytique et de 57 % d'identité (similarité de 73 %) pour leur CBM. Les deux enzymes exprimées sous forme de précurseurs, présentent également en N-terminal un peptide signal qui est ultérieurement éliminé par clivage enzymatique. Cette très forte homologie de séquence fait de Cel5A l'homologue mésophile de Cel5G.

De récentes études comparatives d'un point de vue enzymatique de ces deux endoglucanases démontre très clairement une adaptation de Cel5G aux basses températures (Garsoux, 2002 ; Garsoux *et al.*, 2004). Afin de mieux comprendre cette adaptation et d'en mettre en évidence les déterminants structuraux, des études cristallographiques ont été entreprises tant sur le domaine catalytique seul que sur Cel5G entière. Cependant, si des cristaux du domaine catalytique ont été obtenus, Cel5G entière n'a pour l'instant pas pu être cristallisée. On peut supposer que la cellulase entière composée de deux domaines distincts (catalytique et CBM) reliés par un long domaine de liaison est relativement flexible en solution, empêchant ainsi sa cristallisation.

B/ CRISTALLISATION DE Cel5G DE *PSEUDOALTEROMONAS HALOPLANKTIS*

Publication 4 : Expression, purification, crystallization and preliminary X-ray crystallographic studies of a psychrophilic cellulase from Pseudoalteromonas haloplanktis

Résumé de la publication 4 :

Un criblage des conditions de cristallisation du domaine catalytique de Cel5G ($Cel5G_{CM}$), correspondant à une chaîne polypeptidique de 293 résidus pour 31,9 kDa, a été effectué par la méthode de la diffusion de vapeur et par la technique de la goutte suspendue.

Des cristaux ont été obtenus au bout de deux semaines à la température de 19°C (Figure 10) pour une concentration en protéine de 9 $mg.mL^{-1}$ et pour les conditions de cristallisation suivantes : HEPES 0,1 M pH 7,5 ; citrate de sodium 1,3 M ; glycérol 10 %.

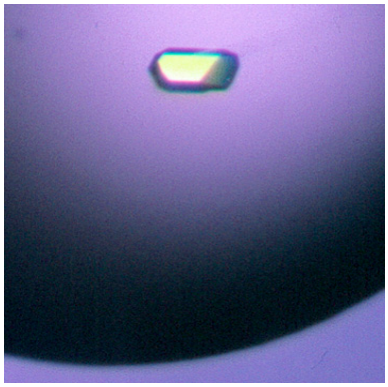


Figure 10 : *Cristal du domaine catalytique de la cellulase Cel5G de Pseudoalteromonas haloplanktis obtenu par la technique de la goutte suspendue en présence de citrate de sodium à pH 7,5.*

Ces cristaux appartiennent au groupe d'espace orthorhombique $P2_12_12_1$ avec des paramètres de maille : a = 135,6 Å ; b = 78,8 Å ; c = 44,2 Å.
Ces cristaux diffractent à une résolution maximale de 1,8 Å sur l'anode tournante du laboratoire (cf. Annexe D pour le principe).

Publication 4

Expression, purification, crystallization and preliminary X-ray crystallographic studies of a psychrophilic cellulase from *Pseudoalteromonas haloplanktis*

Sébastien Violot, Richard Haser**,** Guillaume Sonan, Daphné Georlette, Georges Feller and Nushin Aghajari

crystallization papers

Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Expression, purification, crystallization and preliminary X-ray crystallographic studies of a psychrophilic cellulase from *Pseudoalteromonas haloplanktis*

Sébastien Violot,[a] Richard Haser,[a]* Guillaume Sonan,[b] Daphné Georlette,[b] Georges Feller[b] and Nushin Aghajari[a]

[a]Laboratoire de BioCristallographie, Institut de Biologie et Chimie des Protéines, CNRS et Université Claude Bernard Lyon 1, UMR 5086, F-69367 Lyon CEDEX 07, France, and [b]Laboratoire de Biochimie, Institut de Chimie B6, Université de Liège, B-4000 Liège Sart-Tilman, Belgium

Correspondence e-mail: r.haser@ibcp.fr

Received 12 February 2003
Accepted 17 April 2003

The Antarctic psychrophile *Pseudoalteromonas haloplanktis* produces a cold-active cellulase. To date, a three-dimensional structure of a psychrophilic cellulase has been lacking. Crystallographic studies of this cold-adapted enzyme have therefore been initiated in order to contribute to the understanding of the molecular basis of the cold adaptation and the high catalytic efficiency of the enzyme at low and moderate temperatures. The catalytic core domain of the psychrophilic cellulase CelG from *P. haloplanktis* has been expressed, purified and crystallized and a complete diffraction data set to 1.8 Å has been collected. The space group was found to be $P2_12_12_1$, with unit-cell parameters a = 135.1, b = 78.4, c = 44.1 Å. A molecular-replacement solution, using the structure of the mesophilic counterpart Cel5A from *Erwinia chrysanthemi* as a search model, has been found.

1. Introduction

Life on earth displays a wide capacity for adaptation. Physical limits consistent with biology range from 233 to 388 K in temperature (in the stratosphere and in hydrothermal vents, respectively), up to 120 MPa in pressure (hydrostatic pressures in the deep sea), from pH 1 to 11 and up to 4 *M* in salt concentration (Jaenicke & Böhm, 1998).

Organisms growing under such conditions have been classified as thermophiles, psychrophiles, barophiles (or piezophiles), acidophiles, alkalophiles and halophiles. These organisms present adaptations to high temperatures (>328 K), cold temperatures (around 273 K), high pressures, acidic or alkaline conditions and high ionic strength, respectively.

'Cold enzymes' from psychrophilic microorganisms are generally characterized (i) by having a higher catalytic activity and catalytic efficiency than their mesophilic counterparts in the temperature range 273–303 K, (ii) by a limited thermostability owing to denaturation at moderate and high temperatures and (iii) by an activity curve displaced towards low temperatures compared with mesophilic counterparts (Feller *et al.*, 1996).

Cellulases catalyze the hydrolysis of cellulose, an unbranched homopolymer of β-1,4-linked glucose, which is the major polysaccharidic component of plant biomass.

Cellulases have been classified according to their activity on the substrate into endocellulases (EC 3.2.1.4) and exocellulases (EC 3.2.1.91), which attack the cellulose chain randomly or at the non-reducing extremity, respectively.

Pseudoalteromonas haloplanktis, a psychrophilic Gram-negative bacterium collected in Antarctic seawater, produces the endoglucanase CelG, the gene sequence (Genbank accession No. Y17552) of which shows that the mature CelG protein is made up of 494 residues and comprises a catalytic domain, a proline/serine/threonine-rich linker and a carbohydrate-binding module. Sequence alignments clearly show that the catalytic domain of CelG belongs to family 5 of the glycoside hydrolases from the 90 known families which have so far been classified (Henrissat & Bairoch, 1993; Carbohydrate-Active Enzymes server, http://afmb.cnrs-mrs.fr/~cazy/CAZY/index.html). In family 5, which contains more than 170 glycoside hydrolases, only eight residues are found to be strictly conserved (Wang *et al.*, 1993). Alignment of the primary structures within this family revealed that the cellulase Cel5A from *Erwinia chrysanthemi* (Py *et al.*, 1991) is a mesophilic homologue of the psychrophilic CelG, with which it displays 64% sequence identity.

Determination of the three-dimensional structure of this psychrophilic cellulase will allow detailed analyses and comparative studies with the three-dimensional structure of the mesophilic cellulase (Chapon *et al.*, 2001) in order to obtain insights into protein adaptation to temperature on the molecular level.

2. Materials and methods

2.1. Construction and expression of the recombinant gene

Construction of the *celG* gene lacking the coding sequence for the linker and the

cellulose-binding domain was carried out in two main steps. In order to insert the gene in the expression vector pET22b (Novagen), the nucleotide sequence including the start codon of the wild-type gene (accession No. Y17552) was modified by PCR using Vent DNA polymerase to CATATG, creating an *Nde*I restriction site. In the second step, the catalytic domain was amplified by reverse PCR using a mutating antisense primer which introduces the stop codon TGA at nucleotide 975. The silent sense primer corresponded to the polylinker sequence of the plasmid, therefore allowing the removal of the linker and CBD coding region after amplification and circularization of the product. The nucleotide sequence of the truncated gene was checked on an Amersham Biosciences ALF DNA sequencer.

Escherichia coli Epicurian BL21 (DE3) cells (Stratagene) carrying the recombinant plasmid were grown in LB–ampicillin medium at 291 K. Expression was induced at an OD_{550nm} of ∼3 by 0.1 m*M* isopropyl thio-β-D-galactoside (IPTG) and the culture was grown for an additional 7 h.

2.2. Purification of the catalytic domain of CelG endoglucanase

After centrifugation of the culture at 277 K, periplasmic proteins were extracted by osmotic shock of the cells (Ausubel *et al.*, 1989) using 100 m*M* Tris–HCl, 0.5 m*M* EDTA, 0.5 *M* saccharose, 0.1 m*M* PMSF (phenylmethylsulfonyl fluoride) pH 8.0 as the hypertonic buffer and 1 m*M* $MgCl_2$, 1 m*M* PMSF as the hypotonic solution. When present, nucleic acids were removed by precipitation after overnight stirring in the presence of 0.1%(*w*/*v*) protamine sulfate. Ammonium sulfate (75% saturation) was added to the supernatant and the resulting protein precipitate, collected by centrifugation, was dissolved in a minimal volume of 25 m*M* PIPES pH 6.5. This sample was loaded onto a phenyl-Sepharose CL-4B (Pharmacia) column (20 × 3 cm) equilibrated in 25 m*M* PIPES, 25% saturation $(NH_4)_2SO_4$ pH 6.5 (buffer *A*) and the column was then washed with 2 l buffer *A*. The elution was performed with a decreasing $(NH_4)_2SO_4$ gradient in buffer *A*. The active fractions were pooled and buffer exchange was carried out by ultrafiltration on an Amicon concentrator fitted with a PTGC membrane with sequential addition of 5 volumes of 10 m*M* HEPES pH 7.5. The sample was loaded onto a Macro-prep high Q (Biorad) column (20 × 3 cm) and elution was carried out with a linear gradient from 0 to 0.35 *M* KCl in 10 m*M* HEPES pH 7.5. In the final step, the active fractions were buffer-exchanged with 10 m*M* HEPES pH 7.5 and loaded onto an FPLC system (Pharmacia) equipped with a Mono-Q HR 5/5 column. The enzyme was eluted with a linear gradient from 0 to 0.35 *M* KCl in 10 m*M* HEPES pH 7.5. The purified catalytic domain was buffer-exchanged with 10 m*M* HEPES, 0.04% NaN_3 pH 7.5, concentrated to 18 mg ml^{-1} and stored at 203 K.

Table 1
X-ray diffraction data.

Values in parentheses are for the highest resolution shell.

Space group	$P2_12_12_1$
Unit-cell parameters (Å)	a = 135.1, b = 78.4, c = 44.1
Resolution (Å)	25–1.8
Measured reflections	102004
Unique reflections	42621
Redundancy	2.4 (2.1)
Completeness (%)	94 (94)
$I/\sigma(I)$	8.2 (3.7)
R_{sym}† (%)	5.8 (19)

† $R_{sym} = \sum_{hkl}\sum_i |I(hkl)_i - \langle I(hkl)\rangle| / \sum_{hkl}\sum_i I(hkl)$, where $I(hkl)$ are the intensities of symmetry-redundant reflections and $\langle I(hkl)\rangle$ is the average intensity over all observations.

2.3. Crystallization

Initial crystallization trials were performed using the sparse-matrix sampling method. The screening was conducted using the hanging-drop vapour-diffusion technique in 24-well Linbro plates, employing Crystal Screen kits I and II (Hampton Research, Laguna, CA, USA; Jancarik & Kim, 1991). 4 µl droplets were equilibrated against 500 µl of reservoir solution at 277 and at 292 K.

Crystals suitable for X-ray diffraction studies (Fig. 1) were obtained in 1.3 *M* trisodium citrate dihydrate, 10%(*v*/*v*) glycerol and 0.1 *M* HEPES buffer pH 7.5 at 292 K. The protein-to-mother liquor ratio was 2:1 in 4 µl drops and the initial protein concentration was 9 mg ml^{-1}. Crystals grew within two weeks to dimensions of 0.4 × 0.2 × 0.2 mm.

2.4. X-ray data collection and processing

Diffraction data were collected on a 345 mm MAR Research image-plate system and the X-ray radiation used was Cu *Kα* radiation from a Nonius FR 591 rotating-anode generator operated at 44 kV and 90 mA and equipped with Osmic confocal mirrors. The crystal was flash-frozen in supercooled N_2 gas produced by an Oxford Cryosystems Cryostream (600 series) and maintained at 100 K during the data collection.

The crystal-to-detector distance was 140 mm, the oscillation range per image was 1°, the total oscillation angle was 60° and the exposure time per image was 10 min (Fig. 2).

Determinations of unit-cell parameters and the integration of reflections were performed with the program *MOSFLM* (Leslie, 1991), whereas scaling was performed with the program *SCALA* from the *CCP*4 suite (Collaborative Computational Project, Number 4, 1994). The diffraction data for the crystal of the native enzyme, which were 94% complete to 1.8 Å resolution, display good statistics (see Table 1).

3. Results

The nucleotide sequence of the truncated *celG* gene encodes a polypeptide of 325 amino acids including the signal peptide and

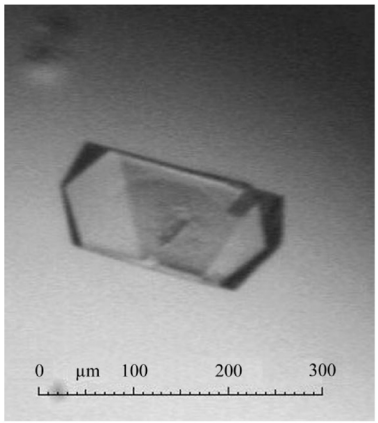


Figure 1
Crystal of the catalytic domain of *P. haloplanktis* CelG.

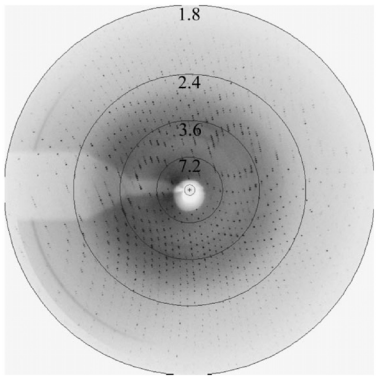


Figure 2
A 1° oscillation image collected in-house at 100 K on a native CelG crystal.

the catalytic domain. N-terminal amino-acid sequencing of the purified gene product shows that the signal peptide (32 amino acids) has been correctly processed in *E. coli*. Accordingly, the isolated catalytic domain contains 293 amino acids, with a predicted mass of 31 890 Da. This truncated protein retains the catalytic properties of the full-length protein, since its specific activity towards carboxymethylcellulose and *p*-nitrophenyl β-D-cellobioside is unmodified.

Crystals of this catalytic domain were obtained and diffraction data were collected. The space group of the crystals was unambiguously determined to be $P2_12_12_1$ owing to systematic extinctions along the three twofold axes. The refined unit-cell parameters are $a = 135.1$, $b = 78.4$, $c = 44.1$ Å. Assuming a molecular weight of 31 890 Da, this gives a solvent content of 67 or 33% and a volume-to-mass ratio, V_M, of 3.7 or 1.8 Å^3 Da^{-1} for one or two molecules in the asymmetric unit, respectively (Matthews, 1968). The 1.8 Å resolution data (Table 1) were used for molecular replacement with the program *AMoRe* (Navaza, 1994). The refined coordinates of the cellulase Cel5A from *E. chrysanthemi* (PDB code 1egz), with which CelG displays 64% sequence identity (Chapon *et al.*, 2001), were used as a search model. Diffraction data in the resolution range 15–2.8 Å were used throughout the search. After the rotation-function search, two peaks with correlation factors of 14.8 and 14.7%, and *R* factors of 56.3 and 56.4%, respectively, were found (the next solution gave a correlation coefficient of 10% and an *R* factor of 57.6%). A correctly positioned molecule of Cel5A corresponding to the first solution after the translation-function search (correlation coefficient of 40% and an *R* factor of 51.2%, compared with the next solution which had a correlation factor of 29.8% and an *R* factor of 55.5%) was used to locate the second molecule. The correlation coefficient subsequently increased to 53.7%. The two Cel5A molecules were then subjected to rigid-body refinement, resulting in a correlation coefficient of 56.4% and an *R* factor of 41.2% for the two molecules in the asymmetric unit.

Refinement of the core domain of CelG is in progress.

This work was partly financially supported by EU Contract No. BI04-CT97-0131; support from the CNRS (Centre National de la Recherche Scientifique) is also gratefully acknowledged.

References

Ausubel, F. M., Brent, R., Kingston, R. E., Moore, D. D., Seidman, J. G., Smith, J. A. & Struhl, K. (1989). Editors. *Current Protocols in Molecular Biology*. New York: John Wiley & Sons.

Chapon, V., Czjzek, M., El Hassouni, M., Py, B., Juy, M. & Barras, F. (2001). *J. Mol. Biol.* **310**, 1055–1066.

Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.

Feller, G., Narinx, E., Arpigny, J. L., Aittaleb, M., Baise, E., Genicot, S. & Gerday, C. (1996). *FEMS Microbiol. Rev.* **18**, 189–202.

Henrissat, B. & Bairoch, A. (1993). *Biochem. J.* **293**, 781–788.

Jaenicke, R. & Böhm, G. (1998). *Curr. Opin. Struct. Biol.* **8**, 738–748.

Jancarik, J. & Kim, S.-H. (1991). *J. Appl. Cryst.* **24**, 409–411.

Leslie, A. G. W. (1991). *Crystallographic Computing 5. From Chemistry to Biology*, edited by D. Moras, A. D. Podjarny & J.-C. Thierry, pp. 50–61. IUCR/Oxford University Press.

Matthews, B. W. (1968). *J. Mol. Biol.* **33**, 491–497.

Navaza, J. (1994). *Acta Cryst.* A**50**, 157–163.

Py, B., Bortoli-German, I., Haiech, J., Chippaux, M. & Barras, F. (1991). *Protein Eng.* **4**, 325–333.

Wang, Q., Tull, D., Meinke, A., Gilkes, N. R., Warren, R. A. J., Aebersold, R. & Withers, S. G. (1993). *J. Biol. Chem.* **268**, 14096–14102.

C/ RESOLUTION DE LA STRUCTURE : DETERMINATION DES STRUCTURES NATIVE ET EN COMPLEXE AVEC LE CELLOBIOSE

Publication 5 : Structural features of cold adaptation in a psychrophilic cellulase revealed by X-ray diffraction and Small Angle X-ray Scattering

1. Résumé de la publication 5 :

L'obtention de cristaux du domaine catalytique de Cel5G de *Pseudoalteromonas haloplanktis* ainsi que la connaissance de leurs conditions de cryo-protection a conduit à la détermination de la structure native de l'enzyme, puis à celle de son complexe avec le cellobiose.

Un jeu complet de données de diffraction aux rayons X a été enregistré à 1,4 Å de résolution en utilisant le rayonnement synchrotron de l'ESRF à Grenoble. La résolution de la structure a été obtenue par la méthode du remplacement moléculaire en utilisant comme modèle-guide la structure cristallographique de son homologue mésophile la cellulase Cel5A d'*Erwinia chrysanthemi.* De plus, la structure tridimensionnelle d'un complexe Cel5G / cellobiose a pu être établie à 1,6 Å de résolution.

Ces résultats cristallographiques, combinées à des données obtenues par diffusion des rayons X aux petits angles, ont permis de suggérer que l'adaptation fonctionnelle de Cel5G pouvait être due aux propriétés structurales originales de son long domaine de liaison connectant le module catalytique au module de fixation du substrat.

Les résultats issus de cette publication démontrent le pouvoir et les avantages de la combinaison des deux techniques de diffraction aux rayons X et de diffusion des rayons X aux petits angles, lorsque la détermination des structures cristallographiques est limitée en raison de la flexibilité ou de l'hétérogénéité des édifices macromoléculaires d'intérêt.

L'application d'une telle stratégie nous a ainsi permis, non seulement de déterminer à une échelle atomique les paramètres moléculaires impliqués dans l'adaptation au froid du domaine catalytique de Cel5G, mais également d'obtenir des informations sur la conformation de l'enzyme entière qui *à priori* ne pourraient être obtenues par la seule approche cristallographique, compte tenu de la nature et de la grande flexibilité du «linker».

2. Déterminants structuraux de l'adaptation au froid du domaine catalytique :

La très forte identité de séquence et de structure 3D entre Cel5G et Cel5A permet une comparaison détaillée des relations structure / fonction ainsi que des propositions de stratégies adaptatives de l'enzyme psychrophile. Une étude attentive du site actif du module catalytique de Cel5G ne révèle pas de caractéristiques structurales pouvant être de façon non ambiguë liées à sa plus forte activité à basse température. Par exemple, les 9 résidus interagissant avec le cellobiose dans le complexe ainsi que les 19 résidus formant la crevasse catalytique sont strictement conservés chez l'enzyme psychrophile et son homologue mésophile. Plusieurs facteurs structuraux affectant potentiellement la dynamique des résidus du site actif, peuvent être mis en évidence. Les deux tours-β (β-turns) additionnels des boucles α2-β2 et α8-β8 pourraient ainsi agir comme des bras de levier, favorisant la déformation de la crevasse catalytique au cours de l'hydrolyse de la cellulose. De même, l'absence de 3 prolines et la présence supplémentaire de 5 glycines (comparativement à Cel5A) pourraient induire une flexibilité accrue de la chaîne polypeptidique de $Cel5G_{CM,}$ alors que l'absence de 3 ponts ioniques pourrait diminuer sa stabilité conformationnelle. De plus, l'augmentation des charges en surface (principalement due à des résidus Aspartate supplémentaires) peut réduire la stabilité de l'enzyme, tandis que l'exposition au solvant de 4 résidus apolaires peut diminuer la compacité et donc déstabiliser la couche externe de la protéine.
En ce qui concerne les interactions entre Cel5G et le cellobiose, la Thr 66 du sous site -2 établit des liaisons hydrogène *via* 2 molécules d'eau, alors que chez Cel5A, ce contact établi par la Ser 69 n'est relayé que par une seule molécule d'eau.

3. Caractéristiques insolites du «linker» chez Cel5G :

La structure de Cel5G entière a également été étudiée par diffusion des rayons X aux petits angles en collaboration avec le laboratoire AFMB. Les résultats obtenus montre que le «linker» ne possède pas une structure régulière et qu'il présente une grande flexibilité. En conséquence, il peut adopter de nombreuses conformations, parmi lesquelles une conformation étendue permettant aux deux modules globulaires compacts que sont le module catalytique et le CBM d'être séparés d'une distance maximale de 140 Å. La possibilité pour le «linker» d'adopter une telle conformation peut s'explique par sa structure primaire. Il contient en effet 23 résidus chargés négativement, aucun résidu chargé positivement et un nombre très

faible de résidus hydrophobes. Les charges négatives sont réparties de manière homogène tout le long de la séquence, évitant ainsi, grâce aux répulsions électrostatiques, le repliement de la chaîne polypeptidique en une structure globulaire. En outre, l'existence de 3 boucles de 13 résidus chacune, résultant de l'établissement de ponts disulfures entre les 6 cystéines du «linker», pourraient stabiliser encore un peu plus la conformation étendue du «linker» par des contraintes stériques.

Le «linker» de son homologue mésophile Cel5A de *Erwinia chrysanthemi* ne présente quant à lui que 30 résidus et ne possède aucune cystéine. Il est également intéressant de constater que de longs «linker» présentant des cystéines sont retrouvées dans d'autres cellulases psychrophiles, comme par exemple CelG de la bactérie psychrophile marine WP-1 (84 % d'identité avec Cel5G de *P. haloplanktis*) et CelA de *Pseudoalteromonas sp. MB-1*. Ce type de «linker» a également été observé chez une α-agarase de *Alteromonas agarilytica GJ1B*. Enfin, une version interne de la base de données CAZy (http://afmb.cnrs-mrs.fr/CAZY/) indique que les «linkers» tendent à se raccourcir quand la thermostabilité augmente, les enzymes hyperthermophiles ne possédant généralement aucun «linker» du tout (Henrissat et Coutinho, communication personnelle).

Toutes ces données suggèrent que l'adaptation fonctionnelle et structurale de Cel5G pour l'hydrolyse de la cellulose aux basses températures pourrait être due à son «linker» anormalement long. Ainsi, considérant (i) un CBM fixé sur une chaîne de cellulose, (ii) le «linker» d'une longueur de 140 Å à 4°C, (iii) la répétition d'un motif cellobiose le long de la chaîne tous les 10,4 Å, il résulte que Cel5G peut hydrolyser 13 à 14 liaisons glycosidiques (Figure 11) le long d'une même chaîne de cellulose. Cette valeur est 3 à 5 fois supérieure à celle reportée pour la mésophile Cel45 de *Humicola insolens* à 20°C. Celle-ci possède un «linker» de 40 Å de long pour 36 résidus, c'est-à-dire environ la taille de celui de la cellulase mésophile Cel5A d'*E. chrysanthemi*.

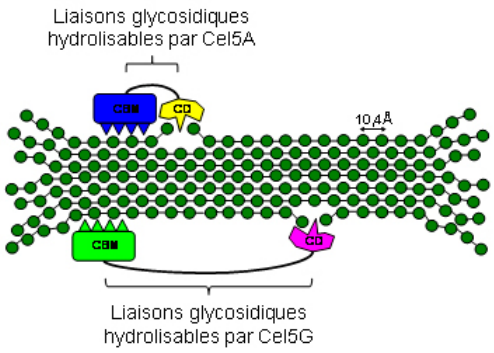


Figure 11 : *Représentation schématique de l'accessibilité au substrat de l'enzyme psychrophile Cel5G et de l'enzyme mésophile Cel5A. Chaque boule verte représente une unité cellobiose.*

Si on considère que le «linker» peut tourner librement autour du CBM, et qu'ainsi le domaine catalytique peut avoir accès à une surface circulaire de 280 Å de diamètre centrée sur le CBM, la superficie en substrat accessible est d'environ $60*10^3$ $Å^2$, soit 40 fois supérieure à son homologue mésophile Cel5A (Figure 12).

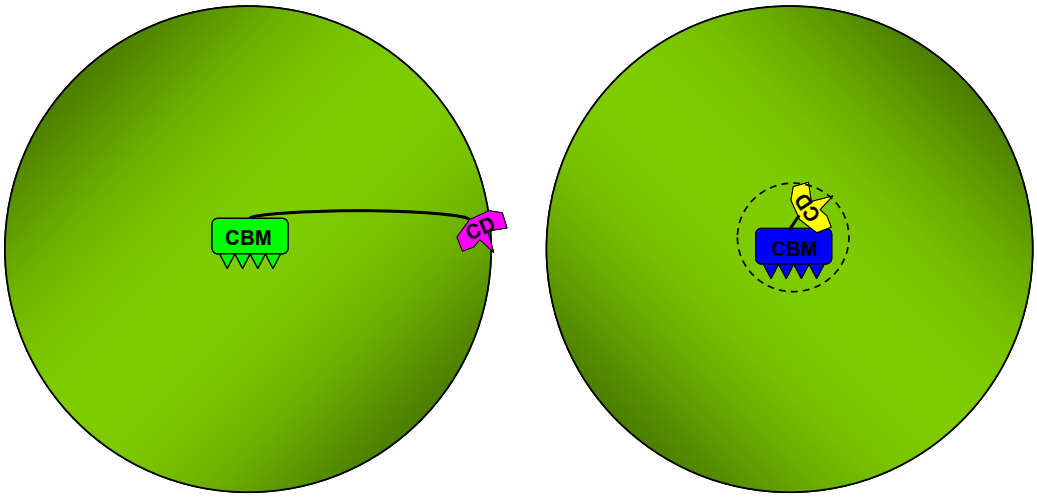


Figure 12 : *Représentation schématique de la surface accessible de substrat pour l'enzyme psychrophile (domaine catalytique en rose) et pour l'enzyme mésophile (domaine catalytique en jaune). La surface accessible pour l'enzyme mésophile est celle comprise à l'intérieur du cercle en pointillé ; CBM : Carbohydrates Binding Domain, CD : Catalytic Domain.*

Publication 5

Structural features of cold adaptation in a psychrophilic cellulase revealed by X-ray diffraction and Small Angle X-ray Scattering

Sébastien Violot, Nushin Aghajari, Georges Feller, Mirjam Czjzek, Patrice Gouet, Guillaume K. Sonan, Charles Gerday, Richard Haser and Véronique Receveur-Bréchot

doi:10.1016/j.jmb.2005.03.026 *J. Mol. Biol.* (2005) **348**, 1211–1224

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Structure of a Full Length Psychrophilic Cellulase from *Pseudoalteromonas haloplanktis* revealed by X-ray Diffraction and Small Angle X-ray Scattering

Sébastien Violot[1], Nushin Aghajari[1]*, Mirjam Czjzek[2], Georges Feller[3] Guillaume K. Sonan[3], Patrice Gouet[1], Charles Gerday[3], Richard Haser[1] and Véronique Receveur-Bréchot[2]*

[1]*Laboratoire de BioCristallographie, Institut de Biologie et Chimie des Protéines CNRS et Université Claude Bernard Lyon 1, UMR 5086 IFR 128 "Biosciences Lyon-Gerland", 7 Passage du Vercors, F-69367 Lyon Cedex 07 France*

[2]*Architecture et Fonction des Macromolécules Biologiques UMR 6098, CNRS et Universités d'Aix-Marseille I and II, 31 Chemin Joseph Aiguier, F-13402 Marseille cedex 20, France*

[3]*Laboratoire de Biochimie Institut de Chimie B6 Université de Liège, B-4000 Liège Sart-Tilman, Belgium*

**Corresponding authors*

Pseudoalteromonas haloplanktis is a psychrophilic Gram-negative bacterium isolated in Antarctica, that lives on organic remains of algae. This bacterium converts the cellulose, highly constitutive of algae, into an immediate nutritive form by biodegrading this biopolymer. To understand the mechanisms of cold adaptation of its enzymatic components, we studied the structural properties of an endoglucanase, Cel5G, by complementary methods, X-ray crystallography and small angle X-ray scattering. Using X-ray crystallography, we determined the structure of the catalytic core module of this family 5 endoglucanase, at 1.4 Å resolution in its native form and at 1.6 Å in the cellobiose-bound form. The catalytic module of Cel5G presents the $(\beta/\alpha)_8$-barrel structure typical of clan GH-A of glycoside hydrolase families. The structural comparison of the catalytic core of Cel5G with the mesophilic catalytic core of Cel5A from *Erwinia chrysanthemi* revealed modifications at the atomic level leading to higher flexibility and thermolability, which might account for the higher activity of Cel5G at low temperatures. Using small angle X-ray scattering we further explored the structure at the entire enzyme level. We analyzed the dimensions, shape, and conformation of Cel5G full length in solution and especially of the linker between the catalytic module and the cellulose-binding module. The results showed that the linker is unstructured, and unusually long and flexible, a peculiarity that distinguishes it from its mesophilic counterpart. Loops formed at the base by disulfide bridges presumably add constraints to stabilize the most extended conformations. These results suggest that the linker plays a major role in cold adaptation of this psychrophilic enzyme, allowing steric optimization of substrate accessibility.

Keywords: cold adaptation; protein disorder; cellulose; protein flexibility; TSP3

Introduction

Psychrophilic micro-organisms grow at temperatures below 4 °C where most of the other organisms cannot grow. Psychrophiles produce cold-adapted enzymes which efficiently catalyze reactions at low temperatures and which are characterized by (i) a higher catalytic activity with k_{cat} values up to ten times higher than their mesophilic counterparts in a temperature range from 0 °C to 30 °C, (ii) a limited thermostability due to denaturation at moderate and high temperatures, and (iii) an activity curve displaced towards low temperatures.[1] It is assumed

Present address: M. Czjzek, Station Biologique de Roscoff, Végétaux Marins et Biomolécules UMR7139, Place George Teissier, BP 74, F-29682 Roscoff cedex, France.

Abbreviations used: CBM, cellulose-binding module; $Cel5G_{CM}$, catalytic module of Cel5G; $Cel5A_{CM}$, catalytic module of Cel5A; SAXS, small angle X-ray scattering; *R*g, radius of gyration; D_{max}, maximum dimension; pNPC, 4-nitrophenyl β-D-cellobioside; LDR, long disordered region.

E-mail addresses of the corresponding authors: n.aghajari@ibcp.fr; receveur@afmb.cnrs-mrs.fr

that these enzymes are more flexible with notably less stabilizing interactions as compared to mesophilic and thermophilic enzymes, allowing conformational changes necessary for activity at low temperature.[2] However, while they have generated considerable interest, not only to improve industrial processes, but also for environmental applications, very little is known on the structural properties of psychrophilic enzymes, with only seven reported structures from bacterial enzymes[3–9] and five structures from cold adapted eukaryotes.[10–12]

Pseudoalteromonas haloplanktis is a psychrophilic Gram-negative bacterium isolated in Antarctica and thriving permanently at temperatures close to the freezing point of water. It is therefore expected that its cellular components are adapted to life in the cold.[13] *P. haloplanktis* secretes a multi-modular endocellulase Cel5G, belonging to GH5 subfamily 2 (GH5-2) and is composed of an N-terminal catalytic module of 292 residues (hereinafter referred to as $Cel5G_{CM}$), a linker region of 109 residues and a cellulose-binding module (CBM) from family 5 of 61 residues at the C-terminal end. The catalytic module of the psychrophilic enzyme shares a highest sequence identity of 64% (79% similarity) with that of endoglucanase Cel5A from *Erwinia chrysanthemi*,[14] its mesophilic homologue, while the corresponding CBM5s display 57% identity (73% similarity). The recent enzymatic comparison of both full-length enzymes Cel5G and Cel5A clearly indicates adaptations to temperature through the kinetic parameters.[15]

Cellulases catalyze the hydrolysis of cellulose, an unbranched homopolymer of β-1,4-linked glucose, which is the major polysaccharidic component of plant bio-mass. They have been classified according to their activity into *endo*- (EC 3.2.1.4) and *exo*-cellulases (EC 3.2.1.91), attacking the cellulose chain randomly or at the non-reducing extremity, respectively. The catalytic modules of cellulases are found in 14 families of glycoside hydrolases, among the 91 families of the CAZy database[16] (families 5 to 10, 12, 26, 44, 45, 48, 51, 61 and 74). Crystal structures are available for families 5 to 10, 12, 26, 45, 48 and 51, and display different folds such as $(\beta/\alpha)_8$, $(\alpha/\alpha)_6$, β-barrel and β-jelly roll.[17] Within family 5, encompassing more than 460 bacterial and fungal members†, all 146 cellulases proceed by the mechanism that retains the configuration of the anomeric carbon atom although only eight residues are invariant.[18] The catalytic modules of eight family 5 endocellulases with known X-ray structures display a common $(\beta/\alpha)_8$-barrel fold.

Most cellulases exhibit a multi-modular organization where the two functional modules are separated by a long flexible linker. This modularity is important for enhanced synergy between the catalytic module and the CBM on its natural substrate, since the shortening or the deletion of the linker drastically reduces enzymatic activity on crystalline cellulose.[19,20] Crystallographic studies have only given access to the structure of the isolated globular modules, and information on the structure of full-length enzymes is only acquired by the use of small angle X-ray scattering (SAXS), which determines the conformational shape of entire molecules in solution. SAXS reveals to be a valuable complementary tool when used together with X-ray diffraction as it can provide missing information from the crystal structure with the low-resolution structure of a protein in solution.[21,22] It has proven to be successful on full length cellulases from the fungus *Humicola insolens*.[23] Recent other examples have shown the power of combining high and low resolution X-ray scattering techniques[24,25] especially in enzymes where flexibility is an important feature.

Here we report the combined structural analysis of the catalytic module of Cel5G solved by X-ray crystallography to focus on the catalytic active site and of the entire enzyme by SAXS studies. Structural adaptations to catalysis at low temperatures have been identified in the catalytic module by analysis of the X-ray structures, while the SAXS experiments suggest a new and unsuspected mechanism of cold adaptation through the modular structure of the cellulase and the peculiarities of its linker region.

Results

Cel5G catalytic module: overall structure

Two molecules, A and B, are present in the asymmetric unit, which superimpose with a root-mean-square deviation (rmsd) of 0.26 Å based on 291 C^{α} atoms.

The catalytic module of Cel5G displays an overall globular form with a long cleft which corresponds to the active site region. Its main structural motif is a $(\beta/\alpha)_8$-barrel, typical for the catalytic module of family 5 glycoside hydrolases. As in the other four subfamily 5–2 structures,[26–28,50] the β-barrel is closed at its N-terminal side by a small additional antiparallel β-sheet (residues 1–20).

Superimposition of the three-dimensional structures of $Cel5G_{CM}$ and $Cel5A_{CM}$ gives an rmsd value of 0.6 Å, based on 291 C^{α} atoms (Figure 1(a)). As expected the two structures differ mainly in loop regions. Structure-based sequence alignment, performed with MAPS‡, shows that $Cel5G_{CM}$ has two two-residue insertions (T64-S65 and W274-N275) compared to $Cel5A_{CM}$ (Figure 1(b)). These insertions occur in the loops between β2 and α2 and β8 and α8, which surround the catalytic cleft at one extremity (Figure 1(a)). The loop connecting β2 with α2 is situated approximately 15 Å from the catalytic site and appears to be the longest among the known

† See the continuously updated CAZY web server at http://afmb.cnrs-mrs.fr/CAZY

‡ http://bioinfo1.mbfys.lu.se/TOP/maps.html

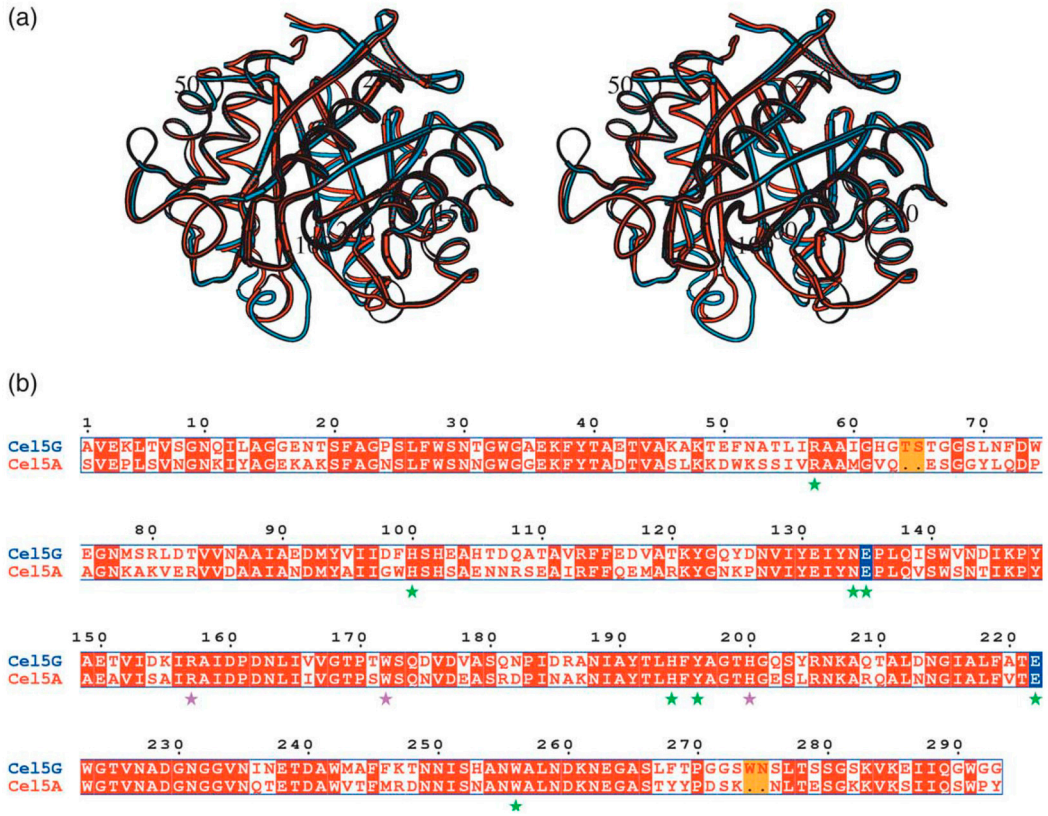


Figure 1. (a) Stereo view of the superimposition of the catalytic module of Cel5G (blue) on that of Cel5A from *E. chrysanthemi* (red) catalytic module. The Figure was prepared with MOLSCRIPT[60] and BOBSCRIPT.[61,62] (b) Sequence comparison of the catalytic modules of Cel5G (blue) with Cel5A *E. chrysanthemi* (red), respectively. Catalytic residues are indicated (blue font), as well as supplementary residues in $Cel5G_{CM}$ *versus* $Cel5A_{CM}$ (golden font). The eight conserved residues in family 5 are indicated (green stars), as well as the three conserved residues in subfamily 5–2 (pink stars). Sequence alignment was performed with the program CLUSTAL[62] and color coded with ESPript.[63]

structures of family 5 glycoside hydrolases. On the other hand, the loop between β8 and α8, which is about 25 Å from the catalytic cavity, is one of the shortest. The insertion of two residues in each loop allows the formation of β-turns, which are unique to $Cel5G_{CM}$.

Comparison of the catalytic modules of the psychrophilic Cel5G and the mesophilic Cel5A cellulases: structural determinants for cold adaptation

Flexibility

It is believed that the catalytic efficiency of psychrophilic enzymes at low temperatures is primarily due to an increased flexibility of the psychrophilic molecules, compared to meso- and thermophilic homologues. In crystallographic studies, the parameters affected by the flexibility of the protein structure are the *B*-factors. The temperature factors of $Cel5G_{CM}$ and $Cel5A_{CM}$ have been compared in order to analyze the flexibility of both enzymes. $Cel5G_{CM}$ has an average *B*-factor of 9 $Å^2$ for all atoms, significantly lower than that of $Cel5A_{CM}$, being 21 $Å^2$. To remove the bias due to crystal quality, data collection or refinement procedures, relative *B*-factors have been used to compare both molecules. Nearly 50% of the residues in $Cel5G_{CM}$ display a higher relative *B*-factor than in $Cel5A_{CM}$. In particular, one stretch in $Cel5G_{CM}$ from residues 227 to 233, located in the loop connecting β7 with α7, displays high relative *B*-factors (up to 200%), whereas they are low for the corresponding region in $Cel5A_{CM}$ (Figure 3). It has been reported that this loop displays an induced movement upon binding for $Cel5A_{CM}$ from *B. agaradhaerens*.[27] Indeed the loop moves towards the catalytic site, bringing the side-chain of Tyr202 (Tyr196 in $Cel5G_{CM}$) into hydrogen bonding distances of sugar binding subsite −1.[29] Similarly in the mesophilic $Cel5A_{CM}$ from *E. chrysanthemi* and $Cel5G_{CM}$, this loop and Tyr196 are in conformations allowing the formation of a hydrogen bond to the substrate.

Thermolability

It is generally considered that high flexibility of proteins is tightly correlated to an increase in thermolability.[30] Several factors contribute to decrease enzyme thermal stability: (i) a lower

Table 1. Amino acid composition for Cel5G and Cel5A catalytic modules

	$Cel5G_{CM}$		$Cel5A_{CM}$		Swiss-Prot[a]
Residue	Number	Frequency	Number	Frequency	Frequency
Asp	18	6.1	14	4.8	5.28
Phe	13	4.4	8	2.8	4.05
Gly	27	9.2	22	7.6	6.90
Lys	12	4.1	19	6.6	5.96
Asn	25	8.5	30	10.4	4.28
Pro	7	2.4	10	3.5	4.86
Arg	6	2.0	10	3.5	5.25
Ser	19	6.5	25	8.7	6.97
Thr	25	8.5	13	4.5	5.55

[a] Data from http://us.expasy.org/sprot

number and special locations of proline/arginine residues; (ii) a higher number of glycine residues; (iii) less aromatic, hydrophobic and charge-mediated interactions; (iv) a larger accessible surface area; and (v) a decreased number of hydrogen bonds.[9] Pair-wise comparison of these factors between the catalytic core of Cel5G and of the mesophilic homologue Cel5A from *E. chrysanthemi* are given in Tables 1–3. A higher flexibility of the catalytic core module of Cel5G may be attained by less proline and arginine residues than in its mesophilic homologue $Cel5A_{CM}$ (see Table 1) and an increase in the number of glycine residues by 21%. Similar trends are frequently observed in psychrophilic enzymes.[31] Furthermore, the surface of $Cel5G_{CM}$ contains more hydrophilic residues, which are supposed to increase flexibility through an increase of interaction with the solvent.[31]

Electrostatic potential

$Cel5G_{CM}$ contains 52 charged residues (34 negatively charged and 18 positively charged), whereas the mesophilic enzyme $Cel5A_{CM}$ contains 60 charged residues (31 negatively charged and 29 positively charged), giving rise to differences in their surface charge distributions. While no remarkable differences can be noticed around the catalytic site, the face opposite to the catalytic site of the molecule is dominated by positive charges in $Cel5A_{CM}$ and has significantly more negative potential in $Cel5G_{CM}$. Whereas the same environment is maintained around the catalytic site on both homologous enzymes, the increased negative potential of the surface of the reverse side of the molecule could allow electrostatic repulsion with the linker, which contains a high number of acidic residues. This suggests that the molecule might adopt an extended conformation, as is further confirmed by the SAXS experiments (see below: Dimensions of full length Cel5G).

Active site and cellobiose binding

The active site is formed by a cleft approximately 35 Å long and 5–10 Å broad. Several aromatic residues line up along the walls of this cleft forming the substrate-binding subsite. Glycoside hydrolases from family 5 use a pair of carboxylic acids in order to cleave the glycosidic bond with net retention of configuration of the anomeric carbon atom in a

Table 2. Intramolecular salt bridges in Cel5G and Cel5A catalytic modules

$Cel5G_{CM}$	$Cel5A_{CM}$	Localization
K4-D126 (S)	–	N-terminal
–	K17-E3 (S)	N-terminal
–	K19-E3 (S)	N-terminal
K37-D260 (S)	K37-D258 (S)	α1-loop β8α8
K46-E42 (S)	–	α1-α1
R57-D98/R57-E131/R57-E222 (I)	R57-E129/R57-E220 (I)	β2-β3/β2-β4/β2-β7
–	K76-E115 (S)	α2-α3
R80-E36 (S)	K78-E36 (S)	α2-α1
K121-D82/K121-D117 (S)	K119-E80/K119-E115 (S)	α3-α2/α3-α3
K146-E150 (S)	K144-E148 (S)	α4-α4
–	R179-D175 (S)	α5-α5
R186-D154 (S)	–	Loop α5β6-α4
R205-E238 (S)	R203-E200/R203-E236 (S)	α6-α6/α6-α7
K207-D175/K207-D177 (S)	K205-D177 (S)	α6-α5
–	R207-E200 (S)	α6-α6
–	K278-E275 (S)	α8-α8
K283-D240/K283-E286 (S)	K279-D238 (S)	α8-α7/α8-α8
K285-E50 (S)	K281-D50 (S)	α8-α1

Letters in parentheses refer to: S, where the salt bridge is located at the surface; and I, in the interior of the molecule.

Table 3. Parameters affecting stability and flexibility in Cel5G and Cel5A catalytic modules

	$Cel5G_{CM}$	$Cel5A_{CM}$
No. of ion pairs	9	12
No. of hydrogen bond/residues	1.1	1.0
Accessible surface area (Å^2)	10,776	10,937
Hydrophobic surface (%)	18	16
Polar surface (%)	44	39
Charged surface (%)	31	41
No. glycine residues	27	22
No. proline residues	7	10
No. arginine residues	6	10

double displacement mechanism.[32] By analogy with other family 5 enzymes, the acid/base and the nucleophile have been identified as being residues Glu135 and Glu222, respectively.

The electron density map of the complex between $Cel5G_{CM}$ and cellobiose revealed a cellobiose molecule in the active site cleft of only one of the molecules in the asymmetric unit (molecule B). The absence of binding in molecule A can be explained by the crystal packing, where molecule B hinders the access to the active site of molecule A. In molecule B, the two sugar moieties occupy binding subsites −2 and −3 according to the nomenclature established for glycoside hydrolases.[33] The interactions between cellobiose and $Cel5G_{CM}$ are shown in Figure 2. The sugar moiety in subsite −3 binds primarily through a hydrophobic stacking on Trp33 and makes a direct hydrogen bond with Lys261. The second glucose unit in subsite −2 forms hydrogen bonds with Trp28, Trp256 and Lys261. A *cis*-peptide bond between Trp256 and Ala257 allows Trp256 (strictly conserved among family 5) to make two hydrogen bonds through its NE1 with the O-1 and O-2 hydroxyl (3 Å) of the −2 subsite sugar (Figure 2). In order to analyze the possible role of sugar-binding subsites in cold adaptation, structures of $Cel5G_{CM}$ and $Cel5A_{CM}$ from

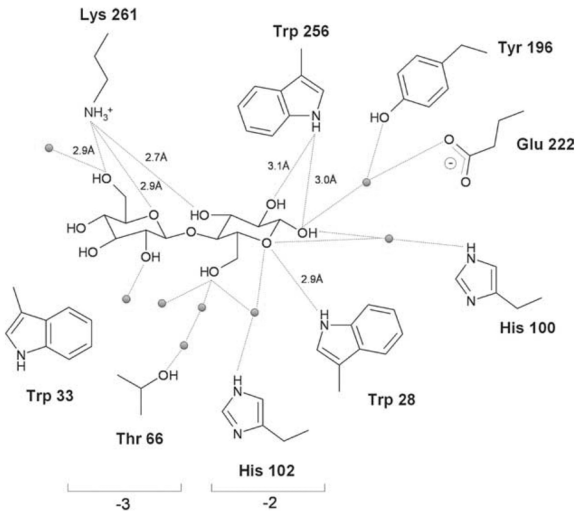


Figure 2. Schematic representation of $Cel5G_{CM}$–cellobiose interactions in subsites −3 and −2.

E. chrysanthemi have been superimposed with the structure of $Cel5A_{CM}$ from *B. agaradhaerens* in complex with a thiopentasaccharide (Figure 4). The only notable differences observed for $Cel5G_{CM}$ with respect to the other family 5 enzymes are located in subsite −1 and subsite +2. In subsite −1 His100 is stabilized by a supplementary hydrogen bond donated by Asp98 that is not present in enzymes from subfamily 5–2. In subsite +2 a tyrosine (Tyr204) in $Cel5G_{CM}$ replaces a leucine or valine in all other structures of enzymes from subfamily 5–2. Besides these minor differences, binding of the substrate is highly similar in $Cel5G_{CM}$ and the binding mode therefore does not seem to play a major role in cold adaptation.

Computational analysis of the sequence of the linker

The very long primary structure of the Cel5G linker allowed a specific analysis of its sequence as an individual module. We considered that the linker starts from the first residue in the sequence after the C-terminal residue of the catalytic module, and ends with the N-terminal residue of the CBM model. Thus, the linker contains 109 residues from residue 293 to 401.

According to secondary structure prediction, the region from residue 290 to 425 comprising the linker is given as random coil. The prediction of a low level of secondary structure is a feature that has recently been noticed in protein regions with "no ordered regular structure".[34]

To fold properly, proteins need a hydrophobic core and a relative small amount of repulsive electrostatic charges. Uversky[35] noticed that the primary structure of natively unfolded domains of proteins differs significantly from globular folded proteins and is characterized by a lack of hydrophobic amino acid residues and an excess of electrostatic residues. He established an equation relating the mean hydrophobicity and the mean net charge, and predicting the natively folded/unfolded state of proteins. According to Uversky's method, the individual Cel5G linker is predicted as natively unfolded, with a mean hydrophobicity of 0.36 and a mean net charge of 0.21.

The group of K. Dunker developed algorithms for the identification of regions within proteins, that lack a fixed tertiary structure, which are referred to as "disordered regions" and are partially or fully unfolded.[36,37] These programs are interesting because they can deal with proteins having both ordered and disordered regions. This collection of Predictors of Natural Disordered Regions is termed PONDR. The reliability of the prediction increases with the length of the region concerned. A borderline length of 40 residues has been chosen by Dunker *et al*. in their extensive study to define long disordered regions (LDRs) of proteins with very good confidence.[38] The error rate for prediction of LDRs consisting of 40 amino acid residues is 0.4% and falls below distinguishable levels for LDRs

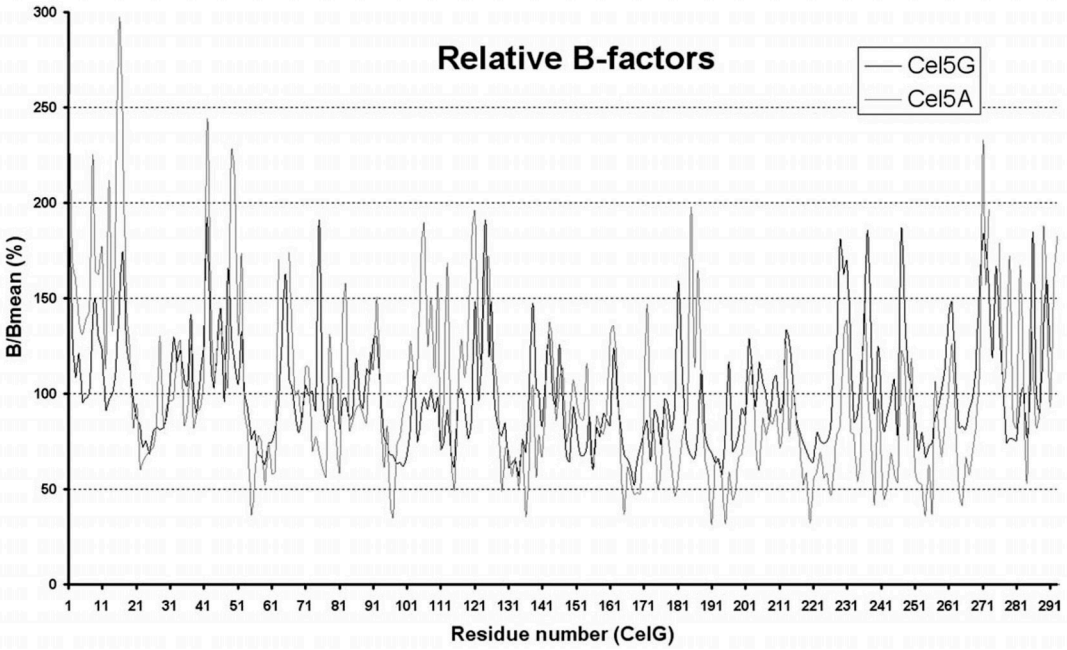


Figure 3. Relative *B*-factors for Cel5G$_{CM}$ and Cel5A$_{CM}$ from *E. chrysanthemi* (%). The relative *B*-factors were obtained by dividing the mean *B*-factor of the residue by the overall *B*-factor value. The native enzyme structure was used for Cel5G$_{CM}$, and the residue numbering corresponds to that of Cel5G (Figure 1(b)).

longer than 60 amino acid residues.[37] The program PONDR was run on full length Cel5G and predicted a LDR within Cel5G, ranging from residues 339 to 382, and several disordered segments in the linker region, including residues 389–391.

This reinforces the prediction that was made on the basis of the hydrophobicity and mean net charge and on secondary structure prediction. This sequence analysis using these different methods strongly suggests that the linker is disordered within Cel5G in solution.

Dimensions of full length Cel5G determined by SAXS

The average size of full length Cel5G is estimated by the measure of its radius of gyration at 20 °C and 5 °C. At low angles, the scattered intensities are very

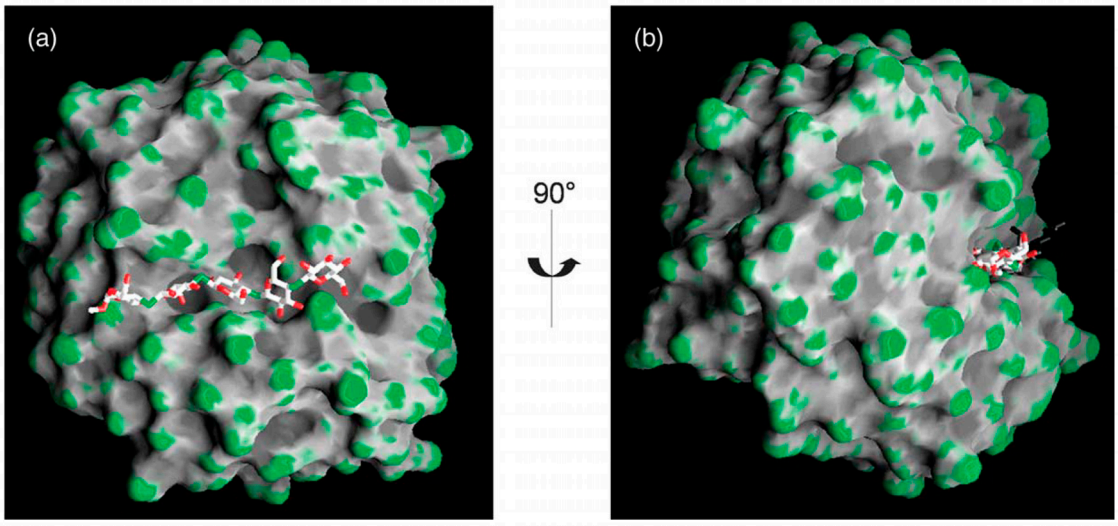


Figure 4. Hypothetical Cel5G$_{CM}$–thiopentasaccharide complex obtained by structural superimposition with Cel5A$_{CM}$ from *B. agaradhaerens* (PDB, 2A3H). No additional fitting was performed on the position of the substrate. The Cel5G$_{CM}$ surface curvature is colored from green (convex) to dark gray (concave) as drawn by GRASP, and is viewed into the active site (a) and after rotation by 90° (b).

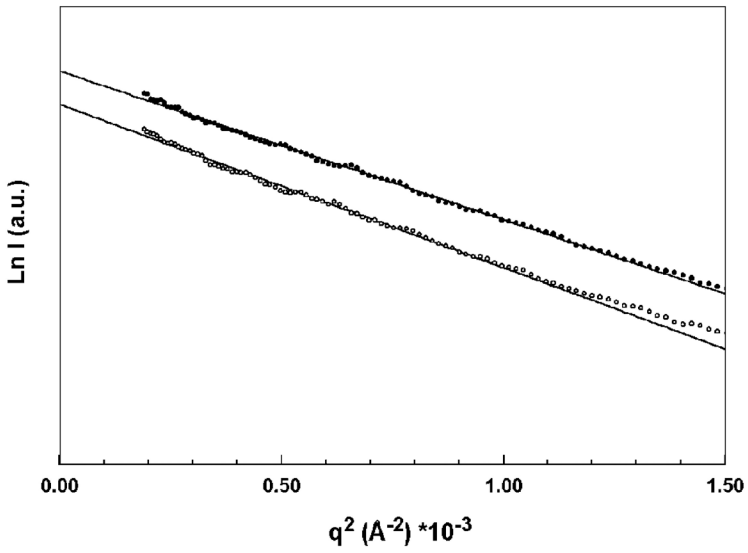


Figure 5. Guinier plot of the scattered intensities of *P. haloplanktis* Cel5G at 5 °C (filled circles) and 20 °C (open circles). The straight lines were obtained from linear regression in the q-region verifying $q \times R_g < 1.0$ for each curve. Data are offset on the vertical axis for better visibility.

well approximated by the Guinier law (Figure 5), and reveal some slightly repulsive inter-particle interactions. The radii of gyration extrapolated at zero concentration are 53.2 (±1.2) Å at 20 °C and 50.7 (±1.1) Å at 5 °C. The distance distribution functions are similar at both temperatures, and typical of an elongated protein, leading to D_{max} values of 218 (±2) Å and 210 (±2) Å at 20 °C and 5 °C, respectively. This pattern is similar to the one observed for the cellulase Cel45 and its variants from *H. insolens*[23] and indicates that the linker is extremely extended between the catalytic module and the CBM. The differences in dimensions of Cel5G at the two temperatures are consistent with a higher Brownian movement of the linker at higher temperature.

3-D modeling of full length Cel5G

The shape of Cel5G has been restored *ab initio* with the program GASBOR,[39] which is well adapted to proteins containing regions with no defined surface such as disordered linkers. Moreover, GASBOR can account for the information contained in the data at high resolution. Several independent calculations were run on the scattering data at both temperatures, in order to reduce the probability of inferring an erroneous structure due the inherent inverse scattering problem. The shapes obtained are reproducible and all present a highly elongated particle, with a large globular spherical part at one end (Figure 6(a)). The best fit gives a χ^2 value of 2.2. Once the catalytic module and the CBM are placed into the shape at their corresponding location at each extremity, the maximum dimension of the linker can be measured, giving values of 148 Å at 20 °C and 140 Å at 5 °C.

As the linker is not glycosylated, the C^{α} trace of the linker can also be modeled with the program package CREDO, which models missing loops or domains in crystal structures from the scattering profile of the entire protein. GLOOPY, which takes into account the sequence of the protein, has been run ten times on the experimental scattering data at each temperature. Very similar results were obtained for the data at 20 °C and 5 °C. All models with equivalent χ^2 values, obtained at 5 °C, are superimposed onto the calculated shape and the atomic structure of the individual globular modules (Figure 6(a)). The best fit of one individual model yields a χ^2 value not better than 4.3.

The large variety of different models whose form factors fit the data clearly shows that the linker can adopt a wide range of conformations and is flexible. Thus, one single model cannot be representative of all the conformations present in the examined solution. To check whether a mixture of Cel5G molecules in all these conformations better describes the scattering properties of the protein solution, we have averaged the form factors of all the models provided by GLOOPY. With this average form factor, the fit to the experimental data is considerably improved, leading to a χ^2 value of 1.0 at 20 °C and 1.6 at 5 °C (Figure 6(c)). This further supports the assumption that the linker is highly mobile in solution.

We are aware that, due to the high flexibility of the linker, more compact conformations may also exist in solution, where the CBM would be closer to the catalytic module. Taking into account these other numerous possible conformations would certainly further improve the quality of the fit. This last result also explains why none of the modeled linkers individually fits into the calculated shape of Cel5G. Indeed, GASBOR calculates the average molecular envelope observed in solution, and therefore calculates the shape of the linker region in its average position. This average shape yields a better χ^2 value than any individual model

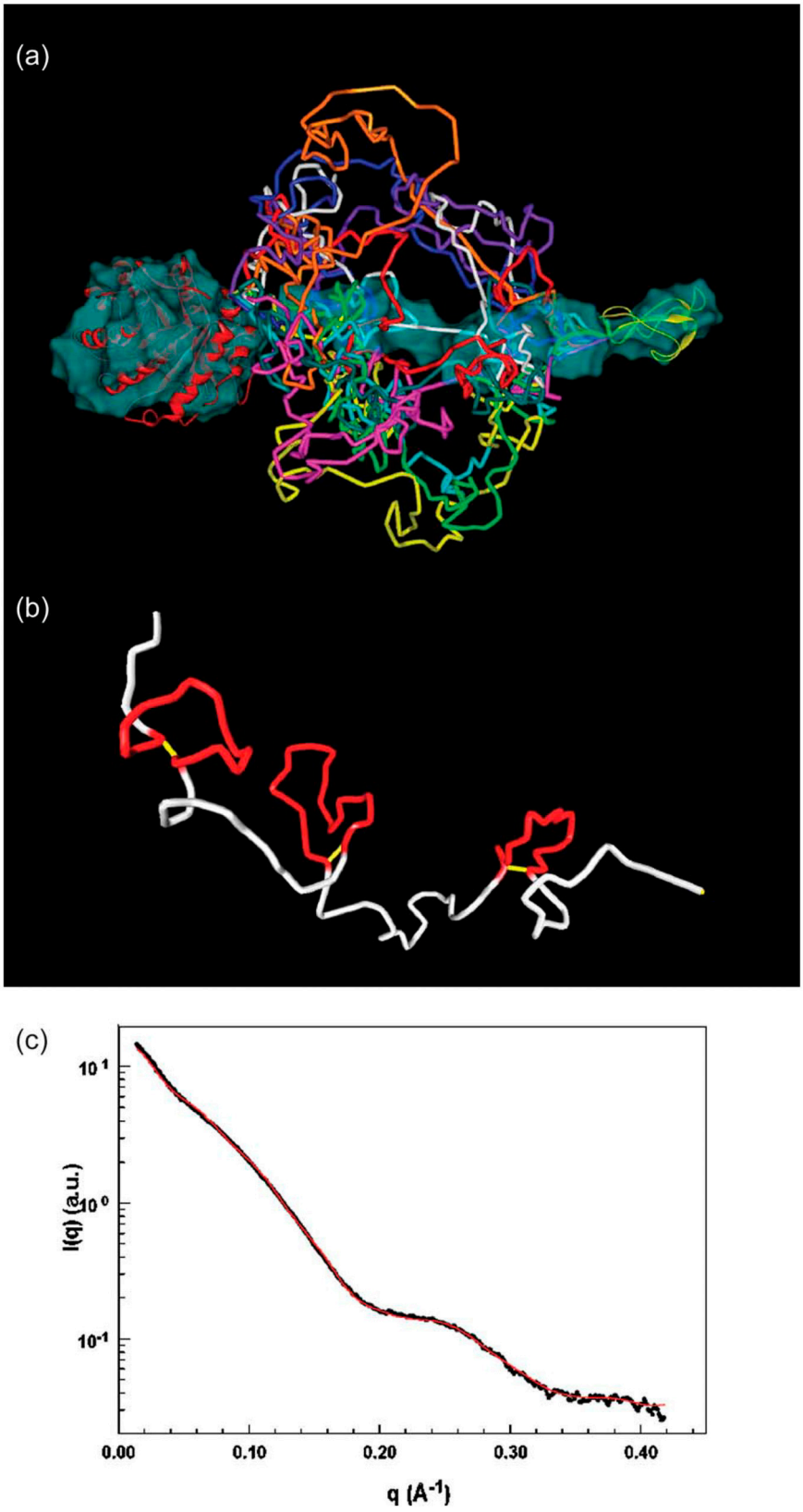


Figure 6. (a) Shape calculated with GASBOR (blue) superimposed with ten different models of Cel5G provided by GLOOPY represented by secondary structure element type. (b) C^{α} trace of a typical linker modelled by GLOOPY exhibiting loops (red) putatively closed by disulfide bonds (yellow). (c) Fit on the experimental scattering curve obtained with the average form factor of the different models provided by GLOOPY.

of Cel5G, but a worse χ^2 value than the mixture of all the models provided by GLOOPY.

Inspection of the linker sequence shows that it contains six cysteine residues. The lack of reaction of 5,5′-dithio-bis-2-nitrobenzoic acid (Ellman's reagent) with Cel5G, either in the native or unfolded (6 M GdmCl) forms, indicates that all cysteine residues in the linker are engaged in disulfide bonds. Although not determined experimentally, a covalent linkage between the adjacent cysteine residues Cys314–Cys328, Cys345–Cys359 and Cys382–Cys396 in the linker can be proposed because other cross-linking combinations would not be compatible with the extended linker conformation recorded by SAXS. Interestingly, these three disulfide linkages result in the formation of three loops, each comprising 13 residues with five conserved positions. More interestingly, such loops can be observed in the models calculated by GLOOPY (Figure 6(b)), thus supporting the above-mentioned formation of loops, which could act as spacers to stabilize the extended conformations by generating steric hindrance.

Discussion

Here we report the analysis of the structural properties of full length cellulase Cel5G from the psychrophilic bacterium *P. haloplanktis* using both X-ray diffraction and small angle X-ray scattering methods. This work again demonstrates the power and advantages of the combination of these two techniques when high resolution techniques alone are likely to be limited by flexibility or heterogeneity in the primary, tertiary and/or quaternary structure.[22] The application of this procedure allowed us not only to gain insights into detailed features at the atomic level of the two crystal structures of $Cel5G_{CM}$, native and complexed to cellobiose, but also to obtain very substantial information on the entire system of Cel5G, which *a priori* could not have been obtained by X-ray crystallography, primarily due to the nature and high flexibility of the linker.

Structural determinants of cold adaptation in the catalytic module

The high level of sequence and structure identity between the cold-adapted Cel5G and the mesophilic Cel5A allows a detailed comparison of the structure–function relationships in the psychrophilic enzyme and of its adaptive strategies. Careful examination of the active site in the catalytic module of Cel5G does not reveal structural features that can be unambiguously related to the higher activity. For instance, the nine residues interacting with cellobiose in the complex as well as the 19 residues forming the catalytic cleft are strictly conserved in both $Cel5G_{CM}$ and $Cel5A_{CM}$, with the exception of the additional Tyr204 in $Cel5G_{CM}$. This suggests that the underlying structural elements bearing the conserved active site residues of $Cel5G_{CM}$ may possess distinct properties conferring a higher activity to the cold adapted enzyme at low temperature as compared to the mesophilic Cel5A, together with a weak substrate binding. This is reminiscent of the high active site flexibility proposed for cold-active enzymes.[13,40] Several structural factors, potentially affecting the dynamics of the active site residues, are found in the crystal structures. The two additional β-turn-containing loops may act as rigid lever arms, assisting deformations of the catalytic cleft during cellulose hydrolysis. The lack of three proline residues and the addition of five glycine residues (compared to $Cel5A_{CM}$) may confer an increased flexibility to the main-chain of the Cel5G catalytic module and the lack of three ion pairs may reduce its conformational stability. In addition, the increase of surface charges (mainly through additional aspartic acid residues) can reduce the domain stability, according to the adhesive–cohesive model[41] and the exposure of four additional non-polar residues to the solvent (entropically unfavorable) can reduce the compactness of the external shell. The active site residues, although conserved in both psychrophilic and mesophilic cellulases, may have distinct properties, as suggested by weaker hydrogen bonding interactions between the cold adapted enzyme and cellobiose (these distances are longer than in Cel5A/cellobiose; data not shown) and by higher relative *B*-factors in the substrate binding cleft. As concerns the interactions, Thr66 in subsite −2 in $Cel5G_{CM}$ establishes hydrogen bonds with cellobiose through two water molecules, while in $Cel5A_{CM}$ this contact is mediated between Ser69 and just one water molecule. The relative *B*-factors in the substrate binding cleft are 92% *versus* 76% for His200 at subsite +2 for $Cel5G_{CM}$ and $Cel5A_{CM}$, respectively, and at subsite −2 the values are 125% *versus* 83% for Tyr196 and 72% *versus* 51% for Tyr256 (Figure 3).

Unusual features of the linker in Cel5G

The structure of Cel5G determined by SAXS shows that the linker does not possess any regular structure and exhibits a high flexibility, as predicted from the sequence. Consequently, it can adopt a wide range of conformations between the two compact globular modules and can separate them up to the maximal distance of 140 Å. The ability of the linker to adopt highly extended conformations arises from its peculiar sequence. The *P. haloplanktis* Cel5G linker contains 23 negatively charged residues, no positively charged residues, and very few hydrophobic residues. The negative charges are widespread along the sequence, thus generating electrostatic repulsions and preventing the polypeptide chain from folding into a globular structure in the absence of counter-ions. Moreover, the existence of three loops each comprising 13 residues and being a result of disulfide linkages between the

six cysteine residues of the linker may certainly further stabilize the most extended conformations through steric hindrance. Indeed, the Cel5G linker can reach far more extended conformations than the mesophilic *H. insolens* Cel45 and its variants previously studied by SAXS. In the latter case, we showed that the linker in its most extended conformation had a density of about 0.7 residue/Å. Here, the density is as low as 0.5 residue/Å (72 residues distributed on 140 Å at 5 °C), calculated without the 39 residues constituting the three loops formed between the disulfide bridges. One may also notice that the linker itself, and not the putative loops (only four out of 23 aspartate or glutamate residues are located in the three loops) of the linker, is rich in charged residues. The combination of numerous repulsive charges and of steric hindrance generated by transversal loops most probably enables the extremely extended conformation of the linker.

Remarkably, 15 of these charged residues form repeats of the TSP3 motif (DxDxDGxxDxxD) flanking the putative cysteine loops. A recent structural study of the C-terminal region of thrombospondin[42] showed that the TSP3 motifs are devoid of regular secondary structure and form irregular loops coordinating a series of calcium ions, while they are random-like when calcium depleted. Although neither calcium nor EDTA influences Cel5G activity on model substrates, these motifs possibly are able to bind calcium *in vivo*. A putative physiological role of these motifs in a cellulase is unexplained but might be related to adaptation to marine conditions (i.e. cold and high salt concentration). Interestingly, the linker of its mesophilic counterpart Cel5A from *E. chrysanthemi* is much shorter (30 amino acid residues) and shows the typical amino acid composition, rich in threonine and serine. Unlike the psychrophilic linker, the mesophilic linker does not contain any cysteine residues or TSP3 motifs. Strikingly, this is the only major difference between the entire sequence of the psychrophilic and of the mesophilic enzyme, since both the catalytic modules and the CBMs are highly homologous (64% and 57% identity, respectively). Furthermore, similar long linkers containing these motifs and cysteine residues are found in other psychrophilic cellulases, like in Cel5G from the psychrophilic marine bacterium WP-1 (84% identity with *P. haloplanktis* Cel5G) and Cel5A from *Pseudoalteromonas sp. MB-1* as well as in an α-agarase from the marine *Alteromonas agarilytica GJ1B*. In addition, a multimodular psychrophilic chitobiase was found to be significantly larger than its mesophilic homologue as a result of a long linker and an additional binding module.[43] Finally, an overview of the linkers of glycoside hydrolases using a modular annotation of the database of CAZy (internal version of CAZy; B. Henrissat & P. Coutinho, personal communication) indicates that linkers of multi-modular enzymes tend to shorten with higher thermostability, while hyperthermophile enzymes have hardly any linker at all. An enhanced flexibility allowed by a long, possibly highly extended linker, with a typical sequence most probably is a new key factor of the cold adaptation of Cel5G. This would allow a higher entropy for the unbound enzyme as in entropic springs found in natively disordered proteins.[44] These molecular springs enable the regulation of distances between the different modules of the enzyme, with a high orientational freedom allowing an efficient targeting, crucial for the catalytic degradation of inaccessible, insoluble cellulolytic substrates.

All this evidence leads us to suggest that a very important functional and structural adaptation to cellulose hydrolysis at low temperatures might arise from the original structural properties of the unusually long linker and in the extended shape of the entire molecule. Such drastic differences in the linker of Cel5G compared to Cel5A should optimize the cellulose hydrolysis by the enzyme in its natural (cold, salted and with a poorly accessible substrate) environment. This has also to be replaced in the context of the caterpillar-like model of cellulase motion along the cellulose fiber, allowing random hydrolysis of neighboring β-1,4 bonds in cellobiose units. Considering (i) a fixed CBM on a cellulose chain, (ii) the 140 Å long linker at 4 °C and (iii) an accessible cellobiose repeat unit being 10.4 Å long, it follows that Cel5G would hydrolyze a maximum of 13–14 glycosidic bonds of similar orientation on a single chain. This is a 3.5-fold higher value than that reported for the mesophilic Cel45 from *H. insolens* at 20 °C[23] which has 36 amino acid residues distributed on a 40 Å long linker. If considering that the linker is able to rotate around the CBM, and that the catalytic domain may thus reach a circular area of diameter 280 Å centered on the CBM, this gives an accessible substrate surface of approximately 60×10^3 Å^2, corresponding to a 40-fold higher value than for the mesophilic Cel45. Accordingly, the unusually long linker of Cel5G would provide a significant steric optimization of the cellulolytic activity by increasing substrate accessibility. Furthermore, it has been proposed that the free energy increase of the cellulase, resulting from linker bending, can be released *via* CBM translation along the cellulose fiber.[23] Consequently, the characteristic linker of Cel5G may be regarded as an improved free energy reservoir, assisting CBM sliding along the substrate at low temperatures that impair diffusion processes: in this model, the disulfide-bonded loops associated with fine tuning through low-affinity binding and release of calcium ions by the TSP3 motifs would store the energy, liberated by spring motions. Thorough investigations to enlighten the precise role of the TSP3 motifs of glycoside hydrolase linkers in the above-described process are presently underway. Finally, studies of Cel5G variants with modified and significantly shorter linker regions will further shed light on the role of the linker in cold adaptation.

Materials and Methods

Sequence analyses

The mean net charge (R)[1] of a protein is defined as the absolute value of the difference between the number of positively and negatively charged residues divided by the total number of amino acid residues. It was calculated using the program ProtParam at the EXPASY server†. The mean hydrophobicity (H)[1] is the sum of normalized hydrophobicities of individual residues divided by the total number of amino acid residues minus four residues (to take into account fringe effects in the calculation of hydrophobicity). Individual hydrophobicities were determined using the Protscale program at the EXPASY server, using the options "Hphob/Kyte & Doolittle", a window size of 5, and normalizing the scale from 0 to 1. The values computed for individual residues were then exported to a spreadsheet, summed and divided by the total number of residues minus four to yield (H). $H_{Boundary}$ was computed according to:[45] $H_{Boundary}=(R+1.15)/2.785$.

Secondary structure predictions were carried out using the PSIPRED program‡[46] on the entire sequence of Cel5G.

Crystallization and data collection

Crystals of recombinant $Cel5G_{CM}$ were obtained by the hanging drop vapor diffusion method at 19 °C, where the drops were equilibrated against reservoirs of 1.3 M trisodium citrate dihydrate, 10%(v/v) glycerol, 0.1 M Hepes (pH 7.5) as described.[47] They belong to the orthorhombic space group $P2_12_12_1$ with unit cell parameters $a=135.1$ Å, $b=78.4$ Å, $c=44.1$ Å (Table 4). The solvent content is 33% for two molecules in the asymmetric unit. The cellobiose complex was obtained by soaking crystals of $Cel5G_{CM}$ in a reservoir solution containing 20 mM cellobiose for 30 minutes. Crystals were flash-cooled in a stream of nitrogen gas prior to data collection at the beamline BM14 at ESRF, Grenoble. A complete data set was collected to 1.4 Å resolution for native $Cel5G_{CM}$ on a MarCCD detector using synchrotron radiation with a wavelength, $\lambda=0.96$ Å. Data on the complex with cellobiose have been collected to 1.6 Å resolution. All data were processed with MOSFLM[48] and reduced with SCALA.[49]

Structure determination and refinement

The structure of native $Cel5G_{CM}$ was solved by molecular replacement, using the structure of its mesophilic homologue (64% identity), the endoglucanase Cel5A from *E. chrysanthemi*,[50] as a search model (PDB idcode, 1EGZ) (Figure 1).

A solution was obtained with the program AMoRe,[51] corresponding to two molecules in the asymmetric unit with a correlation factor of 56% and an R_{factor} of 41% at 2.8 Å resolution.[47] The two molecules in the asymmetric unit can be superimposed by a rotation of 127°. The crystallographic refinement was performed with the program CNS.[52] Five percent of randomly distributed reflections were set apart for the calculation of R_{free} values. The calculated $2F_o-F_c$ and F_o-F_c electron densities were displayed on a Silicon Graphics station with the program TURBO-FRODO.[53] Electron density was absent for C-terminal residues 292–293 in molecule B in the asymmetric unit, and were therefore omitted from the model. Residues 11, 42, 75, 78, 79, 142 and 250 for molecule A and residues 17, 142 and 203 for molecule B have been refined with alternate conformations. The final structure of the native enzyme has been refined to an R_{factor} value of 15.1% and an R_{free} factor of 17.7%, respectively.

The final structure of the native enzyme without water molecules was taken as a starting model to refine the cellobiose complex. The initial R_{factor} of 23% dropped to 18% after refinement with CNS. Electron density maps were calculated and additional density corresponding to a molecule of cellobiose was clearly visible in the $2F_o-F_c$ and F_o-F_c maps in the active site of molecule B of the asymmetric unit. Residues 227–232 and C-terminal residues 292–293 are disordered for both molecules in the asymmetric unit and were therefore omitted from the model. Residues 11 and 117 for molecule A and residue 203 for molecule B have been refined in double conformations. The final structure of the complex, refined to an R_{factor} of 15.8% and an R_{free} factor of 18.9%.

To compare the B-factors from the different structures, relative B-factors were calculated by taking the mean B-factor of every residue and dividing it by the mean B-factor of the whole protein.

Hydrogen bonds were calculated with HBPLUS,[54] while surface features were analyzed with GRASP.[55]

Table 4. Crystal data and refinements statistics

	Native[a]	Cellobiose[a]
A. *Data collection*		
Unit-cell parameters (Å)	135.4 78.9 44.1	135.5 78.8 44.0
Resolution (Å)	15–1.4	20–1.6
Measured reflections	209,086	122,317
Unique reflections	87,830	53,768
Redundancy	2.4 (1.8)	2.3 (1.9)
Completeness (%)	95.1 (95.1)	85.4 (85.4)
$I/\sigma(I)$	10.0 (3.7)	4.8 (3.3)
R_{sym} (%)	4.3 (17.5)	8.0 (18.1)
B. *Refinement statistics*		
Resolution range (Å)	15–1.4	15–1.6
No. of protein atoms	4586	4531
No. of solvent water molecules	776 H_2O	698 H_2O
	–	1 Cellobiose
	–	1 Hepes
R_{factor} (%)[b]	15.1	15.8
R_{free} (%)[c]	17.7	18.9
C. *rms deviation from ideal*		
Bond distances (Å)	0.012	0.007
Bond angles (deg.)	1.8	1.5
D. *Average* B *factor (Å²)*		
Overall	11.05	12.5
Molecule A	8	10
Molecule B	10	12
Solvent	22	24

Values in parenthesis are for highest resolution shell (0.1 Å slice).

[a] X-ray source was beamline BM14 at the ESRF in Grenoble.

[b] $R_{factor}=\sum_{hkl}|F_o-F_c|/\sum_{hkl}|F_o|$ where F_o and F_c are the observed and calculated structure factor amplitudes, respectively.

[c] R_{free} is calculated as the R_{factor}, using F_o terms that were excluded from the refinement (5% of the data).

Sample preparation for SAXS experiments

Full-length Cel5G was expressed in *Escherichia coli* and purified essentially as described[47] except that the plasmid encoding the complete cellulose was used. The enzyme

† http://www.expasy.ch/tools

‡ http://bioinf.cs.ucl.ac.uk/psipred/

was stored in 10 mM Hepes–NaOH (pH 7.5), 0.04% (w/v) NaN_3, 20 μM 4-(2-amino-ethyl)-benzene-sulfonyl-fluoride hydrochloride. Glycerol was added to the buffer as a radiation scavenger. The protein sample was filtered through a Millex®-GV filter with a 0.22 μm cut-off and a non-cellulosic membrane (PVDF) before each measurement. The protein concentrations were measured by absorbance at 280 nm using the extinction coefficient $\varepsilon = 1.91$ ml mg^{-1} cm^{-1} calculated from the sequence.

SAXS experiments

SAXS experiments were carried out at the European Synchrotron Radiation Facility (E.S.R.F., Grenoble, France), on beam-line ID02. The wavelength was 1.0 Å. The sample-to-detector distances were set at 4.0 m and 1.0 m, leading to scattering vectors q ranging from 0.015 $Å^{-1}$ to 0.15 $Å^{-1}$ and from 0.030 $Å^{-1}$ to 0.45 $Å^{-1}$, respectively. The scattering vector is defined as $q = 4\pi/\lambda \sin\theta$, where 2θ is the scattering angle. The detector was a Thomson X-ray image intensified optically coupled to an ESRF developed FReLoN CCD camera. Collections made of 40 successive frames of 0.1 s with 5 s intervals (dead time) between each frame were recorded for each sample. During the dead-time, fresh protein solution was pushed in a 1.5 mm Lindemann-type quartz capillary by using a remote-controlled syringe coupled with the data acquisition program. Therefore no protein solution was irradiated longer than 100 ms and no radiation-induced aggregation was observed. The protein concentration was varied from 0.8 mg/ml to 5.2 mg/ml at each temperature in order to check for inter-particle interactions. Background scattering was measured after each protein sample using the buffer solution and then subtracted from the protein scattering patterns after proper normalization and correction from detector response. Absolute calibration was made with a Lupolen sample. Experiments were carried out at two temperatures, 20 °C and 5 °C. The data acquired at both sample-to-detector distances of 4 m and 1 m were merged for the calculations using the entire scattering spectrum.

Scattering data analysis

The values of radii of gyration (R_g) were derived from the Guinier approximation:[56] $I(q) = I(0) \exp(-q^2 R_g^2/3)$, where $I(q)$ is the scattered intensity and $I(0)$ is the forward scattered intensity. The radius of gyration and $I(0)$ are inferred, respectively, from the slope and the intercept of the linear fit of $\ln[I(q)]$ *versus* q^2 in the q-range $q \times R_g < 1.0$. The distance distribution function $P(r)$ was calculated on the merged curve by the Fourier inversion of the scattering intensity $I(q)$ using GNOM[57] and GIFT.[58]

3-D modeling of full length Cel5G

The low-resolution shape of the full length cellulase was determined *ab initio* from the scattering curve using the program GASBOR.[39] This program restores low-resolution shapes of proteins and calculates a volume filled with densely packed spheres (dummy residues of 3.8 Å diameter) fitting the experimental scattering curve by a simulated annealing minimization procedure with a nearest-neighbor distribution constraint. Several independent fits were run with no symmetry restriction and the stability of the solution was checked. The atomic structures of the catalytic module presented herein and the model of the CBM were then positioned in the envelope using TURBO-FRODO.[53] The structural model of $Cel5G_{CBM}$ was constructed using atomic coordinates of the cellulose-binding module of the cellulase Cel5 from *E. chrysanthemi* obtained by RMN (PDB idcode, 1AIW). Subsequent to the model building with HOMOLOGY (MSI, San Diego, CA) an energy minimization was achieved by the DISCOVER program (MSI) in order to avoid bad molecular contacts. Then the low-resolution model of the linker was determined using GLOOPY from the program package CREDO[59] employing the adequately positioned atomic structures of the isolated modules as template. CREDO is an extension of the original program GASBOR. It calculates the structure of missing portions of crystal structures and represents them by an ensemble of dummy residues forming a chain-compatible model. GLOOPY also accounts for the residue-specific information contained in the primary structure of the model, so that two adjacent residues in the sequence are distant by 3.8 Å in the model. This allows employment of further restrains to generate native-like folds configuration of the missing loop or domain.

Protein Data Bank accession codes

The coordinates of native $Cel5G_{CM}$ (PDB idcode 1TVN) and its cellobiose complex (PDB idcode 1TVP) have been deposited in the Protein Data Bank at Research Collaboratory for Structural Bioinformatics Protein Data Bank.

Acknowledgements

This work was supported by the EU, contract no BI04-CT97-0131. Support from the CNRS (Centre National de la Recherche Scientifique), the Fonds National de la Recherche Scientifique (Belgium) and the Institut Polaire Français are also gratefully acknowledged. We acknowledge Stéphanie Finet and staff for the use of beamline ID02, as well as Philippe Carpentier at beamline BM14, both at the ESRF in Grenoble. We also thank Pedro Coutinho for his help in using CAZy and Bernard Henrissat for reading the manuscript and his constructive comments.

References

1. Feller, G., Narinx, E., Arpigny, J. L., Aittaleb, M., Baise, E., Genicot, S. & Gerday, C. (1996). Enzymes from psychrophilic organisms. *FEMS Microbiol. Rev.* **18**, 189–202.
2. Zecchinon, L., Claverie, P., Collins, T., D'Amico, S., Delille, D., Feller, G. *et al*. (2001). Did psychrophilic enzymes really win the challenge? *Extremophiles*, **5**, 313–321.
3. Aghajari, N., Feller, G., Gerday, C. & Haser, R. (1998). Structures of the psychrophilic *Alteromonas haloplanctis* alpha-amylase give insights into cold adaptation at a molecular level. *Structure*, **6**, 1503–1516.
4. Russel, R. J., Gerike, U., Danson, M. J., Hough, D. W. & Taylor, G. L. (1998). Structural adaptations of the cold-active citrate synthase from an antarctic bacterium. *Structure*, **6**, 351–361.
5. Kim, S. Y., Hwang, K. Y., Kim, S. H., Sung, H. C., Han, Y. S. & Cho, Y. (1999). Structural basis for cold

adaptation. Sequence, biochemical properties, and crystal structure of malate dehydrogenase from a psychrophile *Aquaspirillium arcticum*. *J. Biol. Chem.* **274**, 11761–11767.

6. Alvarez, M., Zeelen, J. P., Mainfroid, V., Rentier-Delrue, F., Martial, J. A., Wyns, L. *et al*. (1998). Triosephosphate isomerase (TIM) of the psychrophilic bacterium *Vibrio marinus*. Kinetic and structural properties. *J. Biol. Chem.* **273**, 2199–2206.
7. Aghajari, N., Van Petegem, F., Villeret, V., Chessa, J. P., Gerday, C., Haser, R. & Van Beeumen, J. (2003). Crystal structures of a psychrophilic metalloprotease reveal new insights into catalysis by cold-adapted proteases. *Proteins: Struct. Funct. Genet.* **50**, 636–647.
8. Van Petegem, F., Collins, T., Meuwis, M. A., Gerday, C., Feller, G. & Van Beeumen, J. (2003). The structure of a cold-adapted family 8 xylanase at 1.3 Å resolution. Structural adaptations to cold and investigation of the active site. *J. Biol. Chem.* **278**, 7531–7539.
9. Bae, E. & Phillips, G. N., Jr (2004). Structures and analysis of highly homologous psychrophilic, mesophilic, and thermophilic adenylate kinases. *J. Biol. Chem.* **279**, 28202–28208.
10. Smalås, A. O., Leiros, H. K., Os, V. & Willassen, N. P. (2000). Cold adapted enzymes. *Biotechnol. Annu. Rev.* **6**, 1–57.
11. Leiros, I., Moe, E., Lanes, O., Smalås, A. O. & Willassen, N. P. (2003). The structure of uracil-DNA glycosylase from Atlantic cod (*Gadus morhua*) reveals cold-adaptation features. *Acta Crystallog. sect. D*, **59**, 1357–1365.
12. de Backer, M., McSweeney, S., Rasmussen, H. B., Riise, B. W., Lindley, P. & Hough, E. (2002). The 1.9 Å crystal structure of heat-labile shrimp alkaline phosphatase. *J. Mol. Biol.* **318**, 1265–1274.
13. Feller, G. & Gerday, C. (2003). Psychrophilic enzymes: hot topics in cold adaptation. *Nature Rev. Microbiol.* **1**, 200–208.
14. Py, B., Bortoili-German, I., Haiech, J., Chippaux, M. & Barras, F. (1991). Cellulase EGZ of *Erwinia chrysanthemi*: structural organization and importance of His98 and Glu133 residues for catalysis. *Protein Eng.* **4**, 325–333.
15. Garsoux, G., Lamotte, J., Gerday, C. & Feller, G. (2004). Kinetic and structural optimization to catalysis at low temperatures in a psychrophilic cellulase from the Antarctic bacterium *Pseudoalteromonas haloplanktis*. *Biochem. J.* **384**, 247–253.
16. Henrissat, B. & Bairoch, A. (1993). New families in the classification of glycosyl hydrolases based on amino acid sequence similarities. *Biochem. J.* **293**, 781–788.
17. Bayer, E. A., Chanzy, H., Lamed, R. & Shoham, Y. (1998). Cellulose, cellulases and cellulosomes. *Curr. Opin. Struct. Biol.* **8**, 548–557.
18. Wang, Q., Tull, D., Meinke, A., Gilkes, N. R., Warren, R. A. J., Aebersold, R. & Withers, S. G. (1993). Glu280 is the nucleophile in the active site of *Clostridium thermocellum* CelC, a family A endo-beta-1,4-glucanase. *J. Biol. Chem.* **268**, 14096–14102.
19. Srisodsuk, M., Reinikainen, T., Penttila, M. & Teeri, T. T. (1993). Role of the interdomain linker peptide of *Trichoderma reesei* cellobiohydrolase I in its interaction with crystalline cellulose. *J. Biol. Chem.* **268**, 20756–20761.
20. Shen, H., Schmuck, M., Pilz, I., Gilkes, N. R., Kilburn, D. G., Miller, R. C. & Warren, R. A. J. (1991). Deletion of the linker connecting the catalytic and cellulose-binding domains of endoglucanase A (CenA) of *Cellulomonas fimi* alters its conformation and catalytic activity. *J. Biol. Chem.* **266**, 11335–11340.
21. Svergun, D. I., Petoukhov, M. V., Koch, M. H. & Konig, S. (2000). Crystal *versus* solution structures of thiamine diphosphate-dependent enzymes. *J. Biol. Chem.* **275**, 297–302.
22. Svergun, D. I. & Koch, M. H. (2002). Advances in structure analysis using small-angle scattering in solution. *Curr. Opin. Struct. Biol.* **12**, 654–660.
23. Receveur, V., Czjzek, M., Schulein, M., Panine, P. & Henrissat, B. (2002). Dimension, shape, and conformational flexibility of a two domain fungal cellulase in solution probed by small angle X-ray scattering. *J. Biol. Chem.* **277**, 40887–40892.
24. Morth, J. P., Feng, V., Perry, L. J., Svergun, D. I. & Tucker, P. A. (2004). The crystal and solution structure of a putative transcriptional antiterminator from *Mycobacterium tuberculosis*. *Structure (Camb)*, **12**, 1595–1605.
25. Hammel, M., Fierobe, H., Czjzek, M., Finet, S. & Receveur-Brechot, V. (2004). Structural insights into the mechanism of formation of cellulosomes probed by small angle X-ray scattering. *J. Biol. Chem.* **279**, 55985–55994.
26. Shaw, A., Bott, R., Vonrhein, C., Bricogne, G., Power, S. & Day, A. G. (2002). A novel combination of two classic catalytic schemes. *J. Mol. Biol.* **320**, 303–309.
27. Davies, G. J., Dauter, M., Brzozowski, A. M., Bjornvad, M. E., Andersen, K. V. & Schulein, M. (1998). Structure of the *Bacillus agaradherans* family 5 endoglucanase at 1.6 Å and its cellobiose complex at 2.0 Å resolution. *Biochemistry*, **37**, 1926–1932.
28. Shirai, T., Ishida, H., Noda, J., Yamane, T., Ozaki, K., Hakamada, Y. & Ito, S. (2001). Crystal structure of alkaline cellulase K: insight into the alkaline adaptation of an industrial enzyme. *J. Mol. Biol.* **310**, 1079–1087.
29. Varrot, A., Schulein, M. & Davies, G. J. (2000). Insights into ligand-induced conformational change in Cel5A from *Bacillus agaradhaerens* revealed by a catalytically active crystal form. *J. Mol. Biol.* **297**, 819–828.
30. Tehei, M., Franzetti, B., Madern, D., Ginzburg, M., Ginzburg, B. Z., Giudici-Orticoni, M. T. *et al*. (2004). Adaptation to extreme environments: macromolecular dynamics in bacteria compared *in vivo* by neutron scattering. *EMBO Rep.* **5**, 66–70.
31. Schiffer, C. A. & Dotsch, V. (1996). The role of protein-solvent interactions in protein unfolding. *Curr. Opin. Biotechnol.* **7**, 428–432.
32. Koshland, D. E. (1953). Stereochemistry and the mechanism of enzymatic reactions. *Biol. Rev.* **28**, 416–436.
33. Davies, G. J., Wilson, K. S. & Henrissat, B. (1997). Nomenclature for sugar-binding subsites in glycosyl hydrolases. *Biochem. J.* **321**, 557–559.
34. Liu, J., Tan, H. & Rost, B. (2002). Loopy proteins appear conserved in evolution. *J. Mol. Biol.* **322**, 53–64.
35. Uversky, V. N., Gillespie, J. R. & Fink, A. L. (2000). Why are "natively unfolded" proteins unstructured under physiologic conditions? *Proteins: Struct. Funct. Genet.* **41**, 415–427.
36. Romero, P., Obradovic, Z., Kissinger, C. R., Villafranca, J. E., &Dunker, A. K. (1997). Identifying disordered regions in proteins from amino acid sequences. In *Proceedings of the I.E.E.E. International Conference on Neural Networks*, vol. 1 pp. 90–95.
37. Romero, P., Obradovic, Z., Li, X., Garner, E., Brown, C.

& Dunker, A. K. (2001). Sequence complexity of disordered protein. *Proteins: Struct. Funct. Genet.* **42**, 38–48.
38. Romero, P., Obradovic, Z., Kissinger, C. R., Villafranca, J. E., Garner, E., Guilliot, S. & Dunker, A. K. (1998). Thousands of proteins likely to have long disordered regions. *Pacific Symp. Biocomput.* **3**, 435–446.
39. Svergun, D. I., Petoukhov, M. V. & Koch, M. H. (2001). Determination of domain structure of proteins from X-ray solution scattering. *Biophys. J.* **80**, 2946–2953.
40. D'Amico, S., Marx, J. C., Gerday, C. & Feller, G. (2003). Activity–stability relationships in extremophilic enzymes. *J. Biol. Chem.* **278**, 7891–7896.
41. Dadarlat, V. M. & Post, C. B. (2003). Adhesive-cohesive model for protein compressibility: an alternative perspective on stability. *Proc. Natl Acad. Sci. USA*, **100**, 14778–14783.
42. Kvansakul, M., Adams, J. C. & Hohenester, E. (2004). Structure of a thrombospondin C-terminal fragment reveals a novel calcium core in the type 3 repeats. *EMBO J.* **23**, 1223–1233.
43. Lonhienne, T., Zoidakis, J., Vorgias, C. E., Feller, G., Gerday, C. & Bouriotis, V. (2001). Modular structure, local flexibility and cold-activity of a novel chitobiase from a psychrophilic *Antarctic bacterium*. *J. Mol. Biol.* **310**, 291–297.
44. Tompa, P. (2003). The functional benefits of disorder. *J. Mol. Struct. (Theochem)*, **666–67**, 361–371.
45. Uversky, V. N. (2002). Natively unfolded proteins: a point where biology waits for physics. *Protein Sci.* **11**, 739–756.
46. McGuffin, L. J., Bryson, K. & Jones, D. T. (2000). The PSIPRED protein structure prediction server. *Bioinformatics*, **16**, 404–405.
47. Violot, S., Haser, R., Sonan, G., Georlette, D., Feller, G. & Aghajari, N. (2003). Expression, purification, crystallization and preliminary X-ray crystallographic studies of a psychrophilic cellulase from *Pseudoalteromonas haloplanktis*. *Acta Crystallog. sect. D*, **59**, 1256–1258.
48. Leslie, A. G. W. (1990). *Crystallographic Computing*, Oxford University Press, Oxford, UK.
49. Collaborative Computing Project number 4. (1994). The CCP4 suite: programs for proteins crystallography. *Acta Crystallog. sect. D*, **50**, 760–763.
50. Chapon, V., Czjzek, M., El Hassouni, M., Py, B., Juy, M. & Barras, F. (2001). Type II protein secretion in gram-negative pathogenic bacteria: the study of the structure/secretion relationships of the cellulase Cel5 (formerly EGZ) from *Erwinia chrysanthemi*. *J. Mol. Biol.* **310**, 1055–1066.
51. Navaza, J. (1994). AMoRe: an automated package for molecular replacement. *Acta Crystallog. sect. A*, **50**, 157–163.
52. Brünger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W. *et al*. (1998). Crystallography & NMR system: a new software suite for macromolecular structure determination. *Acta Crystallog. sect. D*, **54**, 905–921.
53. Roussel, A. & Cambillau, C. (1989). TURBO-FRODO. In *Silicon Graphics Geometry Partners* (Committee, S. G., ed.), pp. 77–78, Silicon Graphics, Mountain View, CA.
54. McDonald, I. K. & Thornton, J. M. (1994). Satisfying hydrogen bonding potential in proteins. *J. Mol. Biol.* **238**, 777–793.
55. Nicholls, A., Sharp, K. A. & Honig, B. (1991). Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins: Struct. Funct. Genet.* **11**, 281–296.
56. Guinier, A. & Fournet, F. (1955). *Small Angle Scattering of X-rays*, Wiley Interscience, New York.
57. Svergun, D. (1992). Determination of the regularization Parameter in Indirect-Transform Methods using perceptual criteria. *J. Appl. Crystallog.* **25**, 495–503.
58. Bergmann, A., Fritz, G. & Glatter, O. (2000). Solving the generalized indirect Fourier transformation (GIFT) by Boltzmann simplex simulated annealing (BSSA). *J. Appl. Crystallog.* **33**, 1212–1216.
59. Petoukhov, M. V., Eady, N. A., Brown, K. A. & Svergun, D. I. (2002). Addition of missing loops and domains to protein models by X-ray solution scattering. *Biophys. J.* **83**, 3113–3125.
60. Kraulis, P. J. (1991). MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallog.* **24**, 946–950.
61. Esnouf, R. M. (1997). An extensively modified version of MolScript that includes greatly enhanced coloring capabilities. *J. Mol. Graph. Model.* **15**, 132–134.
62. Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994). CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucl. Acids Res.* **22**, 4673–4680.
63. Gouet, P., Courcelle, E., Stuart, D. I. & Metoz, F. (1999). ESPript: analysis of multiple sequence alignments in PostScript. *Bioinformatics*, **15**, 305–308.

Edited by R. Huber

(Received 11 January 2005; received in revised form 9 March 2005; accepted 10 March 2005)

Conclusion et perspectives sur Cel5G et son adaptation aux basses températures

La cristallisation d'une forme recombinante tronquée (domaine catalytique) de Cel5G de *Pseudoalteromonas halolanktis* nous a permis d'obtenir la première structure cristallographique d'une «cellulase froide», et ceci à une résolution de 1,4 Å. La comparaison des structures natives des domaines catalytiques de Cel5G et de son homologue mésophile Cel5A de *Erwinia chrysanthemi* n'a pu mettre en évidence que de subtiles différences n'expliquant pas, a *priori*, les dissemblances des deux enzymes sur le plan de leurs propriétés physico-chimiques et enzymatiques en fonction de la température.
La structure native de Cel5G nous a tout de même permis de valider l'hypothèse d'une flexibilité accrue de l'édifice moléculaire, comme attendu pour des enzymes issues d'organismes psychrophiles. En effet, plusieurs déterminants structuraux responsables de cette adaptation au froid ont pu être mis en évidence chez Cel5G.
Afin d'aller plus avant dans la compréhension de la reconnaissance des substrats par Cel5G et de ses mécanismes d'action, la structure de son complexe avec le cellobiose (produit de la réaction d'hydrolyse de la cellulose) a été résolue à 1,6 Å de résolution.
L'étude détaillée et comparée de ce complexe nous a permis de définir structurellement les sous-sites de fixation -3 et -2.

En outre, un des résultats qui nous paraît primordial est la mise en évidence chez Cel5G du rôle essentiel de son «linker» dans le processus d'adaptation aux basses températures. Les différences majeures entre les «linkers» de Cel5G et Cel5A pourraient en effet permettre l'optimisation de l'activité catalytique de l'«enzyme froide» au sein de son environnement naturel.

Malgré ces nouveaux apports dans la connaissance des relations structure / fonction de Cel5G et de son adaptation moléculaire aux basse températures, plusieurs questions n'ont pas trouvé de réponse claire et devront faire l'objet d'études ultérieures. Il en est notamment de l'influence du solvant dans les mécanismes structuraux de l'adaptation aux basses températures. Bien que la structure de Cel5G ait été obtenue à haute résolution, une étude comparative de l'effet du solvant avec son homologue mésophile n'a pu être conduite du fait d'une moins bonne résolution de cette dernière.

En outre, d'importants travaux d'ingénierie protéique rationnelle, conduits en partenariat avec l'équipe de Charles Gerday (*Laboratoire de Biochimie*, Liège, Belgique) ont permis de modifier les propriétés du linker de Cel5G. Ainsi, plusieurs mutants ont pu être clonés, surexprimés et purifiés. Les modifications introduites tendent à réduire la longueur du

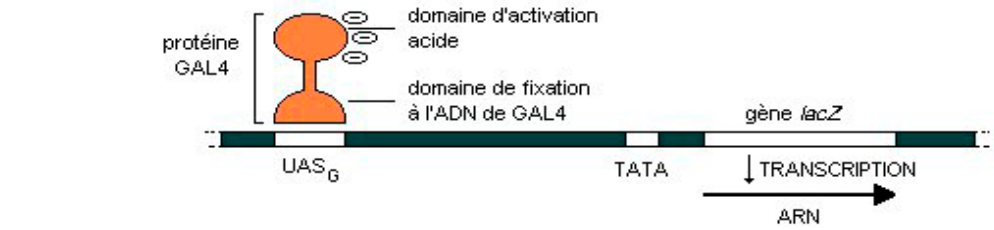
«linker» et la résolution des structures des mutants appropriés apportera de précieux renseignements additionnels sur son rôle.

Enfin, ces travaux montrent qu'une étude structurale n'est pas l'aboutissement d'un projet de recherche mais, au contraire, qu'elle est souvent à l'origine de nouvelles investigations. Elle montre aussi que la mise en commun des moyens et des compétences (biochimie, enzymologie, biologie moléculaire, synthèse organique, biophysique, bioinformatique et biologie structurale, biologie animale, etc.) et les allers-retours entre ces disciplines sont plus que nécessaires, non seulement pour les aspects fondamentaux de l'étude, mais aussi pour toutes les applications potentielles qui peuvent en découler (ingénierie protéique, biotechnologie, conception de principes actifs, etc.).

Annexes

A/ SYSTEME DOUBLE HYBRIDE DANS LA LEVURE

Le système double hybride permet d'identifier *in vivo* une interaction protéine-protéine (Fields et Song, 1989). Il est basé sur la nature modulaire de l'activateur de la transcription GAL4, constitué d'un domaine de liaison à l'ADN qui lie spécifiquement une séquence UAS_G (*upstream activated sequence for the yeast Gal genes*) et d'un domaine d'activation de la transcription contenant une région acide (Keegan *et al.*, 1986 ; Figure 1).



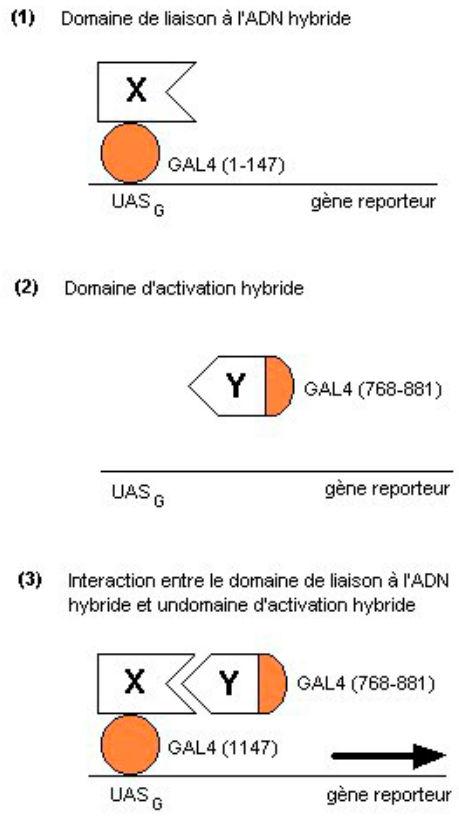


Figure 1 *: Stratégie de détection des interactions entre protéines par le système double hybride.* UAS_G : *upstream activated sequence for the yeast Gal genes*.

Ainsi, les plasmides codant pour deux protéines hybrides, une représentant le domaine de GAL4 liant l'ADN fusionné à une protéine X appelée «appât» (l'appât est généralement une protéine dont on recherche les partenaires cellulaires) et l'autre représentant le domaine de GAL4 activateur de la transcription, fusionné à la protéine Y appelée «proie» (la proie est souvent inconnue et provient d'une banque issue d'un type cellulaire donné) sont construits et introduits à l'intérieur de la levure. L'interaction entre les deux protéines X et Y conduit à la reconstitution d'une protéine GAL4 fonctionnelle et donc à l'activation de la transcription d'un gène rapporteur contenant un site de liaison à GAL4 (Figure 1). Le gène rapporteur le plus utilisé est le gène lacZ, qui code une enzyme, la β-galactosidase. L'expression de lacZ peut-être mesurée par un

test colorimétrique basée sur l'activité de la β-galactosidase. Cette activité peut être facilement visualisée *in vivo* grâce à un test coloré, en remplaçant le substrat naturel de l'enzyme (galactose) par du X-gal (5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside). Le X-gal incolore donne un produit bleu lorsqu'il est clivé par l'enzyme. Il suffit donc de l'ajouter au milieu de culture des bactéries pour visualiser la présence d'une enzyme β-galactosidase active.

B/ LA TECHNIQUE DU «PHAGE-DISPLAY»

La technique du «phage-display» (Figure 2) c'est à dire la présentation de peptides à la surface de phages filamenteux, est devenue un outil très puissant de synthèse combinatoire et de sélection des peptides.

Les phages filamenteux sont utilisés pour présenter à leur surface, en fusion avec le domaine N-terminal de leurs protéines pIII ou pVIII, des molécules telles que des peptides aléatoires, des fragments d'anticorps ou d'autres protéines. Les phages recombinants sont ensuite sélectionnés pour leur capacité de liaison à une cible (biopanning). Après de nombreux lavages, les phages fixés sont élués puis isolés et amplifiés par infection de bactéries. Les phages amplifiés sont sélectionnés à nouveau sur la même cible. Les phages sélectionnés sont analysés et testés pour l'activité recherchée après 3 ou 4 cycles de sélection-amplification.

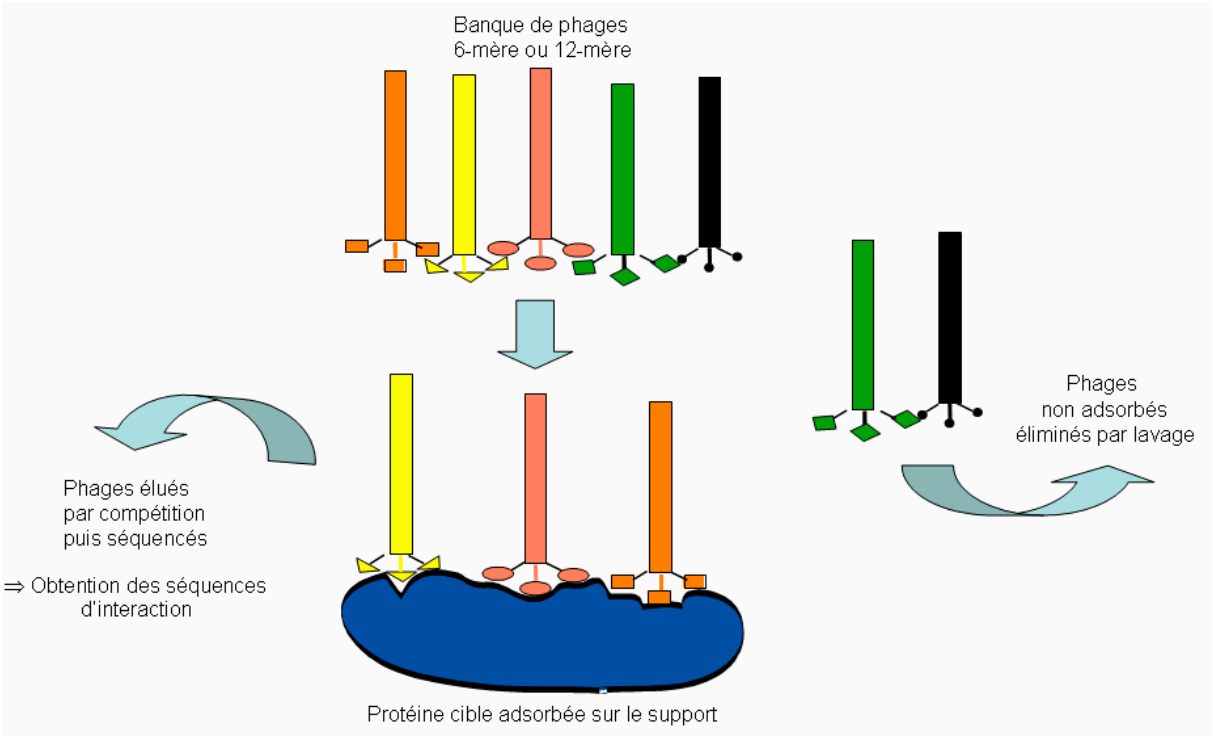


Figure 2 : *Représentation schématique du principe de la méthode du «phage-display».*

Cette stratégie fondée sur la sélection est nettement plus puissante qu'une stratégie de criblage classique qui nécessite de nombreuses manipulations. Il est en effet possible de cribler 10^6 à 10^{10} molécules recombinantes différentes dans un volume réduit de quelques microlitres. De plus, l'association de la protéine exposée en surface (phénotype) avec son ADN codé par le

phage (génotype) permet d'accéder rapidement aux séquences des molécules sélectionnées car l'ADN est directement isolé avec la protéine pour laquelle il code. Cette méthode est très efficace puisqu'il est possible de sélectionner un phage dont la fréquence était de $1/10^8$ dans la banque originale.

C/ REACTION D'INTEGRATION *IN VITRO*

Le système d'intégration rétrovirale *in vitro* (Figure 3) consiste à mettre en présence :

- un ADN donneur linéaire qui comporte à ces extrémités 15 à 30 nucléotides correspondant aux extrémités virales U3 et U5.
- un ADN receveur (plasmide pBSK-zeo).
- la protéine IN purifiée.

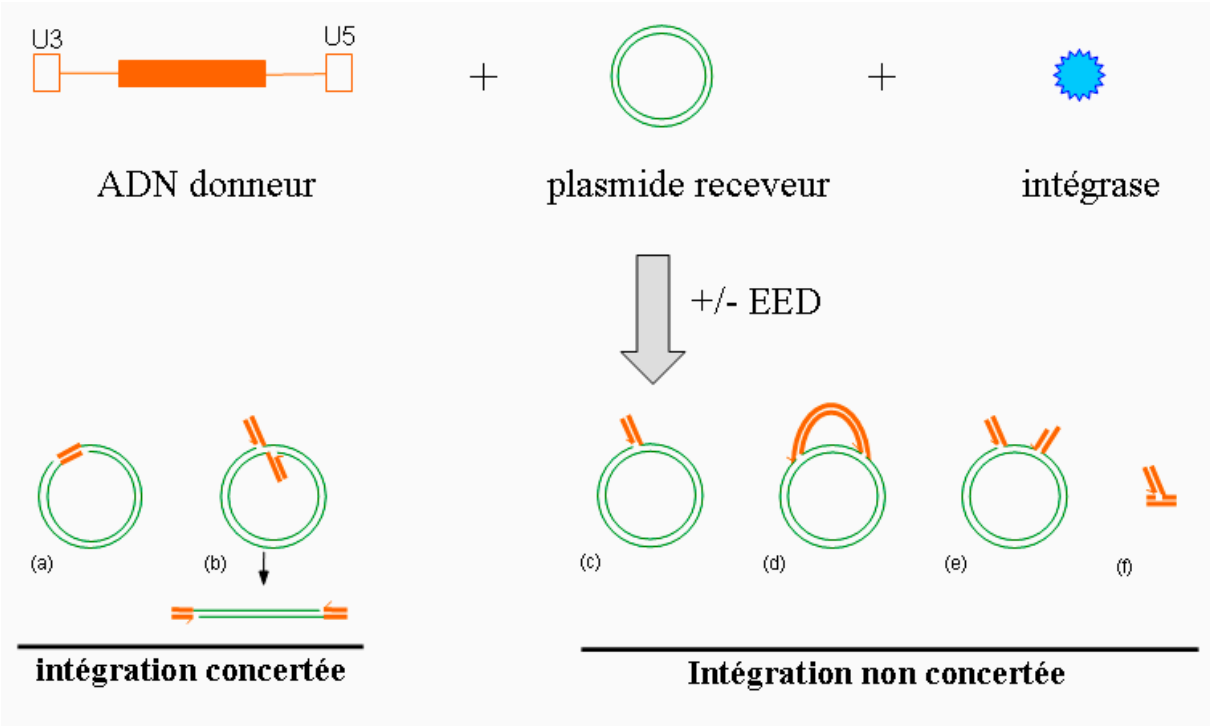


Figure 3 : *Représentation schématique du principe de la réaction d'intégration in vitro.*

Différents produits d'intégration sont obtenus (Figure 3) :

- des produits issus de l'intégration concertée de deux extrémités virales provenant soit d'une seule molécule d'ADN donneur (a) on obtient alors un ADN de forme circulaire, soit de deux molécules (b) on obtient alors une forme linéaire.

- des produits d'intégration non-concertée correspondant à l'intégration d'une seule extrémité virale (c) ou de plusieurs extrémités virales (d et e).

- l'ADN donneur peut aussi s'intégrer dans un second ADN donneur (f).

En présence d'un ADN donneur marqué radioactivement, les produits de la réaction d'intégration peuvent être analysés sur gel d'électrophorèse par autoradiographie. Ils se caractérisent alors par trois bandes spécifiques qui correspondent respectivement aux formes circulaires, linéaires et d'auto-intégration (Figure 3).

D/ ENREGISTREMENT DES DONNEES DE DIFFRACTION

Certaines données exploitées au cours de cette thèse ont été enregistrées au laboratoire sur un détecteur bidimensionnel de type «*Image Plate*» (Mar345 de *MARresearch*) couplé à un générateur de rayons X à anode tournante (FR581 de *Nonius*) associé à des miroirs confocaux *Osmic* (Figure 4). Le recours au rayonnement synchrotron de l'ESRF à Grenoble a permis d'obtenir des données à plus haute résolution.

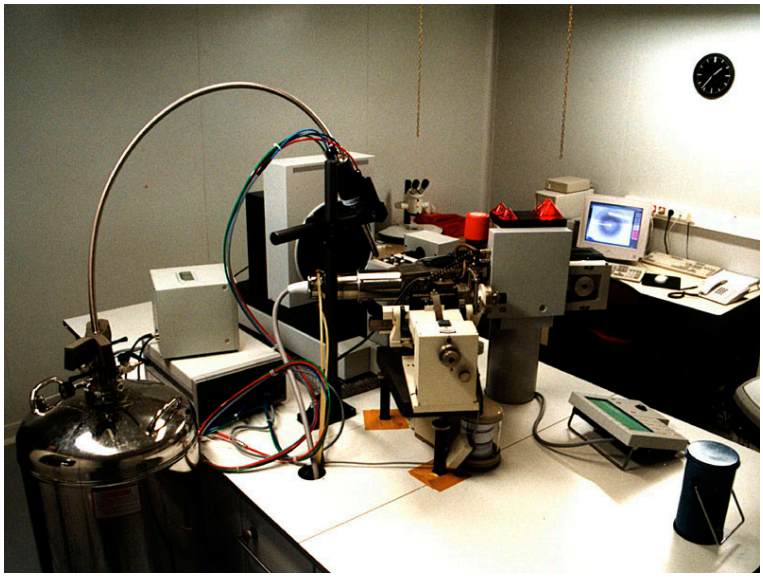


Figure 4 : *Installation rayons X du laboratoire de BioCristallographie de Lyon.*

1. Le générateur de rayons X à anode tournante :

L'anode tournante produit des rayons X en accélérant des électrons dans le vide à un très haut potentiel (entre 40 et 50 kV) sur une cible en métal (anticathode). Le ralentissement des électrons par les atomes de la cible se traduit par une transformation de leur énergie cinétique en majeure partie en chaleur (environ 99,9 %) mais également sous forme d'un rayonnement continu de freinage (ou *Bremsstrahlung*). Des rayons X font parti de ces rayonnements de freinage (Figure 5). A ce spectre continu se superpose un spectre de raies caractéristiques qui ne dépend que de la nature de l'anticathode. Ces raies observées sont des raies d'émission des atomes de l'anticathode. Elles correspondent à des transitions électroniques au niveau de leurs couches profondes. L'éjection d'un électron de la couche K

d'un atome de l'anticathode par un électron projectile détermine l'apparition de deux radiations, K_α et K_β.

Le choix du métal de l'anticathode détermine la longueur d'onde caractéristique à laquelle les rayons X sont émis. L'anode tournante du laboratoire étant équipée d'une cible en cuivre, la longueur d'onde émise est de 1,5418 Å (CuK_α). La monochromie du faisceau est produite grâce aux systèmes optiques de miroirs confocaux placés à la sortie du générateur.

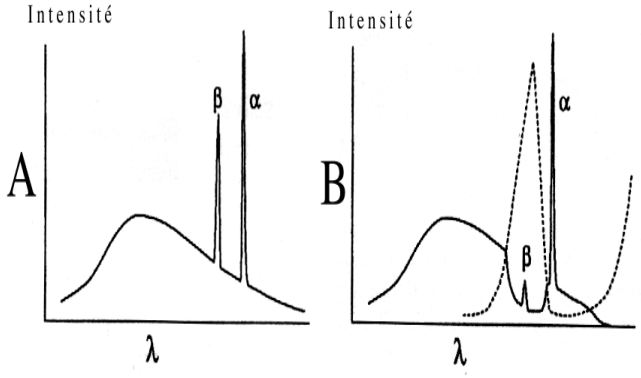


Figure 5 : *Production de rayons X : mise en évidence des raies K_α et K_β superposées au spectre continu (**A**). La ligne en pointillée (**B**) représente le spectre d'absorption du filtre utilisé pour rendre l'émission la plus monochromatique possible.*

2. Le rayonnement synchrotron :

Les électrons de haute énergie dont la trajectoire est infléchie par un champ magnétique intense émettent des ondes électromagnétiques : c'est le rayonnement synchrotron. Couvrant toute la gamme du spectre, des micro-ondes jusqu'aux rayons X durs ($0,1 < \lambda < 5$ Å) la lumière issue de ce rayonnement prend la forme d'un faisceau ultrafin et très intense. L'ESRF («*European Synchrotron Radiation Facility*», Grenoble, France ; http://www.esrf.fr) a été le premier synchrotron de troisième génération mis en fonction dans le monde. Les faisceaux de rayons X produits y sont approximativement 10^{12} fois plus brillants que ceux conventionnellement utilisés en laboratoire ou dans les services de radiographie.

Au cours de l'étude sur la cellulase, nous avons obtenu des temps d'expériences sur les lignes FIP BM30A et ID14-EH1 de l'ESRF qui sont brièvement présentées ci-dessous :

- **Ligne FIP BM30A** : La ligne FIP («*French beamLine for Investigation of Proteins*») est située sur une section d'aimant de courbure. Elle est spécialement dédiée à la cristallographie des macromolécules biologiques. Elle peut être utilisée pour des expériences de diffraction à longueurs d'ondes fixes ou multiples (dispersion anomale, MAD). L'optique de la ligne délivre un faisceau avec un large éventail d'énergie accessible (7 à 21 keV). La ligne possède un détecteur MarCCD de 165 mm de diamètre, un diffractomètre avec un goniomètre 5 cercles et un équipement cryogénique complet (*Oxford system 600 Series Cryostream Cooler*).

- **Ligne ID14** : La ligne ID14 est dédiée à la cristallographie macromoléculaire. Elle possède une brillance moyenne de 13,3 keV. Cette ligne possède 4 sous-stations indépendantes. Chacune est équipée d'un détecteur CCD (*ADSC Q4R CCD*) et d'un goniomètre à axe φ simple. La longueur d'onde fixe de 0,934 Å (13,27 keV) permet d'obtenir une résolution maximale de 0,97 Å. Enfin, un système cryogénique complet est également installé.

3. Les expériences en conditions cryogéniques :

Une fois le cristal exposé au faisceau de rayons X, une série de réactions chimiques, formation de radicaux libres et phénomènes d'ionisation, peuvent l'endommager. Ces dégradations causées par le rayonnement X sont importantes si l'on utilise des sources de rayonnement synchrotron.

La conduite des expériences à 100 K est une solution très efficace pour résoudre ce problème ; le cristal est instantanément gelé (étape de «*flash cooling*») afin de prévenir la formation de glace dans le milieu aqueux. Celui-ci est alors maintenu à une température de 100 K (sous un flux d'azote gazeux) durant toute la phase d'enregistrement des données. Ces conditions cryogéniques impliquent l'utilisation de cryo-protégeants qui empêchent la formation de cristaux d'eau au sein du cristal de protéine. Les cryo-protégeants couramment utilisés sont l'éthylène glycol, le PEG (polyéthylène glycol) le MPD (2-methyl-2,4-pentanediol) le glycérol, certaines huiles à base de silicone ou encore des sucres simples (glucose, par exemple) ou des alcools.

E/ TRAITEMENT DES DONNEES : CARTES DE DENSITE ELECTRONIQUE

1. La diffraction :

Sous un faisceau de rayons X, les phénomènes d'interférences dus à la périodicité du cristal déterminent les directions des ondes diffractées. L'amplitude diffusée par les atomes dépend des facteurs de diffusion de chaque type d'atome (relié au nombre d'électrons) et des déphasages de ces amplitudes. On montre que le facteur de structure F(hkl) s'écrit :

$$F(hkl) = |F(hkl)|.e^{i\varphi_{hkl}}$$

avec |F(hkl)| amplitude (ou module) de l'onde diffractée et φ_{hkl}, phase de l'onde diffractée.

La connaissance simultanée des modules F(hkl) et des phases φ_{hkl} permet le calcul de la densité électronique en tout point (x,y,z) de l'espace réel du cristal, selon la transformée de Fourier inverse :

$$\rho(x,y,z) = \frac{1}{V}\sum_{h=-\infty}^{+\infty}\sum_{k=-\infty}^{+\infty}\sum_{l=-\infty}^{+\infty}|F(hkl)|.e^{i\varphi_{hkl}}.e^{-2\pi i(hx+ky+lz)}$$

avec hkl indices de Miller ou coordonnées dans l'espace réciproque et V le volume de la maille.

Expérimentalement, seules sont enregistrées les intensités I(hkl) des ondes diffractées égales au produit scalaire du facteur de structure $\vec{F}_{hkl}$ et de son conjugué $\vec{F}^*_{hkl}$:

$$I(hkl) \cong \vec{F}_{hkl} . \vec{F}^*_{hkl} \cong |F(hkl)|^2$$

Ainsi, seuls les modules des facteurs de structure sont connus après enregistrement. La mesure obtenue est donc partielle puisque l'information sur les phases est perdue. Celles-ci doivent donc être déterminées par des méthodes indirectes.

2. Le problème des phases :

Comme nous l'avons vu, les données de diffraction donnent directement accès à l'intensité I(hkl) et par conséquent à l'amplitude du facteur de structure F(hkl). Cependant, l'information concernant les phases est perdue, d'où l'impossibilité de calculer une densité électronique.

Trois méthodes principales (MIR, MAD et remplacement moléculaire) sont utilisées en cristallographie biologique afin de résoudre ce problème des phases.

La méthode du remplacement isomorphe (Harker, 1956) et la méthode de la diffusion anomale multiple (Hendrickson, 1991) permettent d'accéder à l'information de phase dans le cas de protéines pour lesquelles aucunes structure apparentée n'est connue.

La méthode MIR («Multiple Isomorphous Replacement») :

La méthode MIR consiste à fixer de manière spécifique dans l'édifice protéique, des atomes ayant un grand nombre d'électrons (atomes lourds). La contribution des atomes lourds, beaucoup plus riches en électrons que les atomes de carbone, d'azote ou d'oxygène, détermine un pouvoir de diffraction plus important et va permettre de les localiser (Harker, 1956). La difficulté majeure de cette technique est d'obtenir des cristaux dérivés isomorphes aux cristaux natifs ; cet isomorphisme implique que les paramètres de maille ainsi que la position des macromolécules n'ont pas subi de variations significatives.

Ainsi, la densité électronique du dérivé PH est égale à la densité électronique de la protéine P plus celle de l'atome lourd H :

$$F_{PH}hkl = F_{P}hkl + F_{H}hkl$$

Dans cette équation, seuls les modules F_P et F_{PH} sont connus expérimentalement. La phase φ_P peut prendre deux valeurs possibles :

$$\varphi_P = \varphi_H \pm \cos^{-1} \frac{(F^2{}_{PH} - F^2{}_P - F^2{}_H)}{2F_P F_H} \qquad (1)$$

La fonction de Patterson permet de remonter à la position (x,y,z) des atomes lourds dans la maille. Elle est notée :

$$P(u,v,w) = \frac{2}{V} \sum_{h=-\infty}^{+\infty} \sum_{k=-\infty}^{+\infty} \sum_{l=-\infty}^{+\infty} (F_{PH} - F_P)^2 . \cos 2\pi(hu + kv + lw)$$

Cette fonction est une fonction réelle qui est la transformée de Fourier inverse des carrés des facteurs de structures. Elle représente les vecteurs différence, ramenés à l'origine, entre les atomes présents dans la maille. La position de l'atome lourd peut alors être déterminée en déconvoluant cette fonction. Le calcul de F_H et φ_H est alors possible et par conséquent celui des phases pour la protéine entière également.
Il existe cependant une ambiguïté puisque l'équation (1) admet deux solutions. L'utilisation d'un deuxième dérivé lourd, ayant un site de fixation différent, permet de lever cette ambiguïté.

La méthode MAD («Multiple Anomalous Diffraction») :

Si la protéine contient un diffuseur anomal, la différence d'intensité entre les paires de Bijvoet $|F_h(+)|^2$ et $|F_h(-)|^2$ peut être utilisée pour résoudre le problème des phases. Le signal anomal provient de l'excitation des électrons des couches internes des atomes. Si la longueur d'onde utilisée est proche de la longueur d'onde d'excitation de ces électrons, alors ceux-ci entrent en résonance et il apparaît un changement dans les rayons X diffusés : c'est la diffusion anomale. Les facteurs de diffusion des atomes anomaux sont décrits par un nombre complexe qui tient compte de la composante de résonance. Le principe physique de cette méthode est décrit par Hendrickson (Hendrickson, 1991).

3. La méthode du remplacement moléculaire :

La méthode du remplacement moléculaire nécessite une structure connue présentant plus de 30% de similarité de séquence avec la macromolécule étudiée. Cette structure homologue (appelée modèle-guide) est placée dans la maille de la structure à déterminer et les phases de la macromolécule étudiée sont estimées à partir des facteurs de structure calculés à partir du modèle-guide. Une recherche sur six dimensions est nécessaire afin de trouver le meilleur placement du modèle-guide, à savoir une Matrice de rotation et un vecteur de translation. En fait, il est possible de diviser cette recherche en deux étapes à trois dimensions : un recherche de rotation suivie par une recherche de translation.

La première étape du remplacement moléculaire est le calcul des facteurs de structure du modèle-guide placé dans une maille P1 artificielle. La maille est plus grande que le modèle-guide dans les trois directions de l'espace, de façon à ce qu'il n'y ait pas de vecteurs intermoléculaires dans le rayon de Patterson utilisé pour la recherche de la fonction de rotation. Cette fonction est une fonction de corrélation qui rend compte de la superposition des deux jeux de vecteurs obtenus à partir des F_{obs} et F_{calc}. La fonction de rotation (Rossmann et Blow, 1962) est calculée sur un volume d'intégration sphérique dont le rayon est choisi suffisamment petit pour exclure la majorité des vecteurs intermoléculaires. Cette fonction de rotation calculée dans l'espace de Patterson s'écrit :

$$R = \int_u P1(x1)P2(x2)d_{x1}$$

avec P1 et P2 les cartes de Patterson calculées respectivement à partir des F_{obs} et des F_{calc} et x1, x2 les vecteurs interatomiques de l'espace de Patterson.

Une fois la solution de rotation déterminée, il faut chercher la solution de translation représentée par un vecteur de translation T. Son calcul consiste à rechercher par des mouvements de translation dans l'unité asymétrique, les positions qui correspondent aux meilleures corrélations entre les F_{obs} et le F_{calc}. Plusieurs méthodes de calcul peuvent être utilisées pour cette recherche. Celles utilisées par le programme AMoRe sont présentées ci-dessous.

Le programme AMoRe (Navaza, 1994) contraction de Automatic. Molecular Replacement, correspond à une suite de programmes permettant de rechercher automatiquement les solutions de rotation puis de translation présentant la meilleure corrélation. Les principaux programmes qui le composent sont:

- Sorting : trie et convertit les données de diffraction au format utilisable par le programme.
- Tabling : positionne le modèle-guide dans une maille triclinique et calcule les facteurs de structure correspondants (F_{calc}).
- Roting : recherche les solutions de la fonction de rotation dans l'espace réciproque selon la méthode de Crowther (Crowther et Blow, 1967) et les classe par ordre de corrélation décroissante. Par défaut, seules les solutions dont la corrélation est au moins égale à 50 % de la meilleure solution sont retenues par le programme.

- Traing : recherche les fonctions de translation sur les solutions de la fonction de rotations retenues à l'étape précédente. Plusieurs méthodes peuvent être utilisées par le programme mais dans notre cas, la méthode de Crowther & Blow (1967) méthode la plus rapide qui est utilisée par défaut par le programme, a donnée des solutions suffisamment contrastées.
- Fitting : effectue un affinement en corps rigide (Castellano *et al.*, 1992) sur les solutions de translation retenues et calcule les coefficients de corrélation et les facteurs R pour chaque solution.

Le facteur R représente le désaccord existant entre le modèle (F_{calc}) et les données cristallographiques (F_{obs}). Il sera suivi tout au long de l'affinement afin de juger la vraisemblance du modèle. Ce facteur se définit comme suit:

$$R = \frac{\sum_h \left\| F_{obs}(h) \right| - k \left| F_{calc}(h) \right\|}{\sum_h \left| F_{obs}(h) \right|} \quad \text{avec k, facteur d'échelle.}$$

La méthode du remplacement moléculaire est de plus en plus utilisée, compte tenu du nombre grandissant de structures connues (Rossmann, 1990). Elle a été mise en oeuvre pour la résolution de la structure native de Cel5G, du fait de la connaissance antérieure de la structure de son homologue mésophile Cel5A (code d'accession *Protein Data Bank* : 1EGZ (Chapon *et al.*, 2001).

F/ AFFINEMENT DU MODELE : VERS LA STRUCTURE FINALE

Le modèle de la protéine étudiée (F_{calc}) doit correspondre au maximum aux données expérimentales (F_{obs}). Afin d'obtenir le meilleur accord entre les données expérimentales et calculées, il est nécessaire, pour chaque atome, de modifier sa position spatiale (x,y,z) son facteur d'agitation thermique B et son occupation. La phase d'affinement des données permet de minimiser l'énergie totale du système définie par :

$$\mathbf{E_{totale}} = E_{\text{cristallographique}} + E_{\text{empirique}}$$

Le terme $E_{empirique}$ représente les termes d'énergie liés aux contraintes stéréochimiques. Elle est égale à la somme des énergies liées aux interactions covalentes (longueur E_{bond}, angles de liaison E_{angl}, angles impropres E_{impr}) et non covalentes (potentiels électrostatiques E_{elec} et interactions de Van der Waals E_{vdw}). L'énergie empirique est évaluée en considérant la différence entre la valeur actuelle et la valeur idéale. Ainsi, si les atomes dévient de la géométrie idéale, la quantité d'énergie augmente.

Le terme $E_{cristallographique}$ est la différence pondérée existant entre les facteurs de structure observés et calculés. Il s'écrit :

$$E_{\text{cristallographique}} = \frac{W_a}{N_a} \sum_{hkl} W_{hkl} \left[\left| F_{obs}(hkl) \right| - k. \left| F_{calc}(hkl) \right| \right]$$

Le facteur W_a est un facteur de pondération appliqué à toutes les réflexions de manière à ce que le gradient de l'énergie soit égal au gradient de l'énergie empirique. Il est déterminé par une étape de dynamique moléculaire. Le facteur N_a est un facteur de normalisation, W_{hkl} le poids associé à une réflexion donnée, et k est le facteur de mise à l'échelle entre les facteurs de structure observés et calculés.

L'affinement va donc par des variations de coordonnées atomiques, faire évoluer le système vers une conformation plus stable et donc de moindre énergie. Il existe deux protocoles principaux de minimisation d'énergie. Ces derniers ont été utilisés au cours de notre étude, tels qu'ils sont implémentés dans le logiciel CNS version 1.1 (Brunger *et al.*, 1998).

1. L'affinement en corps rigide («*rigid body refinement*») :

Dans cette méthode d'affinement, le nombre de degrés de liberté de la protéine à affiner est réduit au maximum. Le rapport entre le nombre de paramètres à affiner et le nombre de paramètres observés est donc diminué.

Dans la pratique, le modèle est divisé en une ou plusieurs parties, appelées «corps» qui correspondent aux différentes sous-unités ou domaines de la structure. Pour chacun de ces «corps», il est possible d'affiner jusqu'à six paramètres (trois degrés de rotation et trois degrés de translation). Les changements de positions des «corps» sont calculés en combinant les changements de tous les atomes les composant. Cette méthode est très utile lorsque, entre deux structures homologues, les positions des domaines varient les uns par rapport aux autres.

2. L'affinement par recuit simulé (Brunger *et al.*, 1987) :

Les champs de force définissent une surface multidimensionnelle, nommée surface de potentiel d'énergie ou, plus couramment, espace conformationnel. Ce dernier possède une topologie complexe où deux types de points ont un intérêt particulier : les minima et les maxima énergétiques locaux, qui correspondent respectivement à des structures stables et à des points de transition dans l'espace conformationnel (Figure 6).

Pour une protéine, la surface d'énergie est extrêmement accidentée. Il existe un grand nombre de minima accessibles, chacun d'eux ayant approximativement la même énergie, mais pouvant correspondre à des structures «différentes».

L'affinement par recuit simulé permet d'explorer un très large espace conformationnel. Des vitesses aléatoires sont attribuées aux atomes du système, qui est porté à une température élevée (typiquement jusqu'à 3000 K), puis refroidi progressivement généralement par pas de 20 K. Les trajectoires des atomes sont calculées entre chacun de ces pas. L'énergie cinétique ainsi fournie par augmentation de température permet de franchir des maxima énergétiques locaux (Figure 6). Le refroidissement qui s'ensuit a pour objectif de figer le système dans son état de plus basse énergie (donc, de plus grande stabilité).

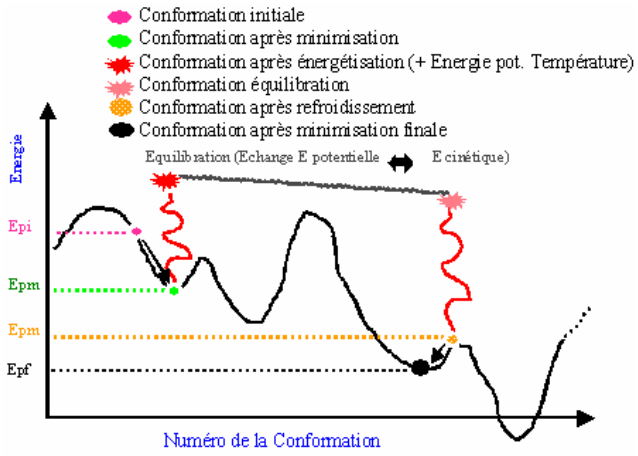


Figure 6 : *Représentation schématique de l'évolution de l'énergie d'un système au cours d'un affinement par recuit simulé.*

L'utilisation des méthodes de «vraisemblance maximum» («*maximum likelihood*» et «*cross-validated maximum likelihood*» ; Adams *et al.*, 1997) qui sont incluses dans le programme CNS (Brunger *et al.*, 1998) ont permis d'améliorer l'affinement des structures. Leur but est de déterminer la vraisemblance du modèle à partir de l'estimation des erreurs de ce dernier et des intensités mesurées. Cette méthode combinée à une validation croisée avec le facteur R_{libre} (Brunger, 1992) permet de réduire les risques de sur-affinement des données. L'estimation du solvant («*bulk solvent*») a également été incluse dans la procédure de recuit simulé (Brunger *et al.*, 1990 ; Brunger *et al.*, 1987 ; Jiang et Brunger, 1994). Ceci permet de calculer une enveloppe moléculaire séparant la protéine du solvant (Leslie, 1987). Ainsi, par itération, l'amélioration des phases contribue à l'amélioration des cartes de densité électronique.

Quand la résolution et la qualité des données affinées l'ont rendu possible, un affinement individuel des facteurs de température (facteurs B) est réalisé. Enfin les cartes de densité électroniques $2F_o$-F_c et F_o-F_c sont classiquement calculées (Kleywegt et Brunger, 1996).

Le suivi qualitatif de la phase d'affinement est conduit grâce au calcul des facteurs d'accord, à savoir le facteur R et le facteur R_{libre} (R_{free}). Le facteur R_{libre} est défini comme suit :

$$R_{libre} = \frac{\sum_{h \in T} \left| |F_{obs}(h)| - k|F_{calc}(h)| \right|}{\sum_{h \in T} |F_{obs}(h)|}$$

où k est un facteur d'échelle et T, un jeu de réflexions qui comprend, dans cette étude, une sélection aléatoire de 5 % de toutes les réflexions observées (Brunger, 1992).

3. Agitation thermique :

Les atomes du cristal subissent, du fait de la température, une vibration permanente dans les trois dimensions de l'espace autour de leur position d'équilibre : c'est l'agitation thermique. Dans le cas général, chaque atome se déplace en fonction de son environnement et des liaisons auxquelles il participe, de façon à ce que sa densité électronique se trouve répartie dans un ellipsoïde. Cette agitation est prise en compte avec le facteur de température atomique isotrope T :

$$T = \exp(-B.\sin^2\theta / \lambda^2)$$

avec λ la longueur d'onde, θ l'angle de Bragg et B le facteur de température exprimé en $Å^2$.

Le facteur de température B est relié à l'écart quadratique moyen du déplacement (μ^2) de l'atome par rapport à sa position moyenne :

$$B = 8\pi^2\mu^2$$

Les expériences effectuées à basse température permettent d'avoir de plus forte intensité à grand angle de diffraction puisque le facteur d'agitation thermique joue moins sur la contribution à la diffusion.

L'examen des facteurs d'agitation thermique B permet en outre de détecter des erreurs dans la structure. Ainsi, des facteurs B très bas (< 10 $Å^2$) peuvent montrer que les atomes considérés sont très peu mobiles ou qu'il y a un problème de sur-affinement des données. A l'inverse, des facteurs B trop élevés (> 50 $Å^2$) peuvent montrer que les atomes ne sont pas à leurs bonnes positions ou bien que leurs occupations ne sont pas égales à 1.

4. Logiciels utilisés dans l'analyse structurale

La qualité des structures tridimensionnelles des protéines étudiées a été contrôlée à l'aide des logiciels PROCHECK (Laskowski *et al.*, 1993) et WHATCHECK (Vriend, 1990) permettant de calculer, notamment, un diagramme de Ramachandran (Ramakrishnan et Ramachandran, 1965).

Les alignements de séquences ont été réalisés en utilisant le programme CLUSTALW (Thompson *et al.*, 1994). Les structures secondaires obtenues à partir des fichiers de coordonnées atomiques PDB ont été déduites par l'algorithme DSSP (Kabsch et Sander, 1983). Les alignements de séquences incluant les superpositions de structures secondaires déduites des fichiers de coordonnées atomiques ont été réalisés avec le serveur ESPript (http://espript.ibcp.fr/ESPript/ESPript/ ; Gouet *et al.*, 1999).

Les figures illustrant ce travail ont été réalisées avec les logiciels TURBO-FRODO (Roussel et Cambillau, 1989), MOLSCRIPT (Kraulis, 1991), BOBSCRIPT (Esnouf, 1997), GRASP (Nicholls *et al.*, 1991), VIEWERLITE® (Accelrys, 2001) et ISIS / Draw (MDL, 1990-2001).

G/ PUBLICATIONS ET COMMUNICATIONS

1. Publications :

P1-2003 - Violot, S., Haser, R., Sonan, G., Georlette, D., Feller, G. and Aghajari, N. (2003) Expression, purification, crystallization and preliminary X-ray crystallographic studies of a psychrophilic cellulase from *Pseudoalteromonas haloplanktis*. *Acta Cryst.,* **D59**, 1256-1258.

P2-2003 - Violot, S., Hong, S.S., Rakotobe, D., Petit, C., Gay, B., Moreau, K., Billaud, G., Priet, S., Sire, J., Schwartz, O., Mouscadet, J.F. and P. Boulanger. (2003) The human Polycomb group EED protein interacts with the integrase of human immunodeficiency virus type 1. *J. Virol.,* **77**, 12507-12522.

P3-2003 - Moreau K., Faure C., Violot S., Verdier G. and Ronfort C. (2003) Mutations in the C-terminal domain of ALSV (Avian Leukemia and Sarcoma Viruses) integrase alter the concerted DNA integration process *in vitro*. *Eur. J. Biochem*., **270**, 4426-4438.

P4-2004 - Moreau, K., Faure, C., Violot, S., Gouet, P., Verdier, G. and Ronfort, C. (2004) Mutational analyses of the core domain of Avian Leukemia and Sarcoma Viruses integrase : critical residues for concerted integration and multimerization. *Virology*, **318**, 566-581.

P5-2005 - Godoy S., Violot S., Boullanger P., Bouchu M.N., Leca-Bouvier B.D., Blum L.J., and Girard-Egrot A.P. (2005) Kinetics study of Bungarus fasciatus venom acetylcholinesterase immobilised on a Langmuir–Blodgett proteo-glycolipidic bilayer. *ChemBioChem.*, **6**, 395-404.

P6-2005 - Violot S., Aghajari N., Czjzek M., Feller G., Sonan G.K., Gouet P., Gerday C., Haser R., and Receveur-Brechot V. (2005) Structure of a full length psychrophilic cellulase from Pseudoalteromonas haloplanktis revealed by X-ray diffraction and small angle X-ray scattering. *J. Mol. Biol*., **348**, 1211-1224.

2. Communications :

2.a) Communications orales :

CO1-2003 - S. Violot, P. Boulanger, S.S. Hong, D. Rakotobe, P. Gouet et R. Haser.
Etude cristallographique de la protéine EED, partenaire cellulaire de la Matrice et de l'Intégrase du virus VIH-1
7ème Journée Scientifique de l'Ecole Doctorale Interdisciplinaire Sciences-Santé (EDISS) Lyon, 16 avril 2003.

CO2-2004 - S. Violot, P. Boulanger, S.S. Hong, D. Rakotobe, P. Gouet et R. Haser.
Etude fonctionnelle et structurale de la protéine cellulaire EED, partenaire potentiel des protéines Matrice et Intégrase du virus VIH-1
Séminaire interne, Lyon, 19 janvier 2004.

CO3-2004 - S. Violot, P. Gouet, P. Boulanger et R. Haser.
Protéines virales Matrice et Intégrase du virus VIH-1 et protéine cellulaire EED : un ménage à 3 ?
Congrès GTBIO 2004, Lyon, 22 au 25 juin 2004.

2.b) Communications par affiche:

CA1-2001 - S.Violot, G. Parsiegla, A. Belaïch, J.P. Belaïch et R.Haser.
Structure cristallographique de Cel9M de *Clostridium cellulolyticum*
Congrès de l'Association Française de Cristallographie, Paris / Orsay, 3 au 6 juillet 2001.

CA2-2002 - M. Foucault, S. Violot et R. Haser.
Les cellulases : étude des enzymes clès associées à la conversion de la biomasse
6ème Journée Scientifique de l'Ecole Doctorale Interdisciplinaire Sciences-Santé (EDISS) Lyon, 22 mars 2002.

CA3-2002 - S. Violot, P. Gouet, R. Haser, C. Gerday, G. Feller et N. Aghajari.
Cellulase d'un microorganisme psychrophile : structures 3D de ses formes natives et en complexe avec du cellobiose
Congrès GTBIO 2002, Marseille, 16 au 19 octobre 2002.

CA4-2002 - S. Violot, P. Gouet, R. Haser, C. Gerday, G. Feller et N. Aghajari.
A cellulase from a psychrophilic microorganism : 3D structures of its native form and its complex with cellobiose
The 4th congress on extremophiles, Naples (Italie) 22 au 26 septembre 2002.

CA5-2004 - S. Violot, V. Receveur-Bréchot, M. Czjzek, P. Gouet, G. Feller, R. Haser et N. Aghajari.
L'adaptation au froid d'une cellulase psychrophile étudiée par cristallographie et diffusion des rayons X aux petits angles
Congrès GTBIO 2004, Lyon, 22 au 25 juin 2004.

CA6-2005 - Sébastien Violot, Patrice Gouet, Pierre Boulanger, Richard Haser.
EED, a cellular partner of the viral proteins MA, IN and Nef from HIV-1
XX IUCr Congress, Florence (Italie) 23 au 31 Août 2005.

H/ PARTICIPATIONS A DES COLLOQUES ET ATELIERS

1. Colloques :

C1-2001 - "**1st Symposium on the Alpha-Amylase Family**" : Smolenice Castle, Slovaquie, 30 septembre au 4 octobre 2001.

C2-2002 - "**The 4th congress on extremophiles**" : Naples, Italie, 22 au 26 septembre 2002.

2. Ateliers :

A1-2002 - "**High-throughput Structure Determination CCP4 Study Weekend**" : York, Royaume-Uni, 4 au 5 janvier 2002.

Références bibliographiques

[A]

Adams, J., Kelso, R. and Cooley, L. (2000) The kelch repeat superfamily of proteins: propellers of cell function. *Trends Cell Biol*, **10**, 17-24.

Adams, P.D., Pannu, N.S., Read, R.J. and Brunger, A.T. (1997) Cross-validated maximum likelihood enhances crystallographic simulated annealing refinement. *Proc Natl Acad Sci U S A*, **94**, 5018-5023.

Aghajari, N., Feller, G., Gerday, C. and Haser, R. (1998a) Crystal structures of the psychrophilic alpha-amylase from Alteromonas haloplanctis in its native form and complexed with an inhibitor. *Protein Sci.*, **7**, 564-572.

Aghajari, N., Feller, G., Gerday, C. and Haser, R. (1998b) Structures of the psychrophilic Alteromonas haloplanctis alpha-amylase give insights into cold adaptation at a molecular level. *Structure*, **6**, 1503-1516.

Aghajari, N., Van Petegem, F., Villeret, V., Chessa, J.P., Gerday, C., Haser, R. and Van Beeumen, J. (2003) Crystal structures of a psychrophilic metalloprotease reveal new insights into catalysis by cold-adapted proteases. *Proteins.*, **50**, 636-647.

Aguilar, C.F., Sanderson, I., Moracci, M., Ciaramella, M., Nucci, R., Rossi, M. and Pearl, L.H. (1997) Crystal structure of the beta-glycosidase from the hyperthermophilic archeon Sulfolobus solfataricus: resilience as a key factor in thermostability. *J Mol Biol*, **271**, 789-802.

Akhmanova, A., Verkerk, T., Langeveld, A., Grosveld, F. and Galjart, N. (2000) Characterisation of transcriptionally active and inactive chromatin domains in neurons. *J Cell Sci*, **113 Pt 24**, 4463-4474.

Alvarez, M., Zeelen, J.P., Mainfroid, V., Rentier-Delrue, F., Martial, J.A., Wyns, L., Wierenga, R.K. and Maes, D. (1998) Triose-phosphate isomerase (TIM) of the psychrophilic bacterium Vibrio marinus. Kinetic and structural properties. *J Biol Chem.*, **273**, 2199-2206.

Anthony, N.J. (2004) HIV-1 integrase: a target for new AIDS chemotherapeutics. *Curr Top Med Chem*, **4**, 979-990.

Arold, S., Franken, P., Strub, M.P., Hoh, F., Benichou, S., Benarous, R. and Dumas, C. (1997) The crystal structure of HIV-1 Nef protein bound to the Fyn kinase SH3 domain suggests a role for this complex in altered T cell receptor signaling. *Structure*, **5**, 1361-1372.

Arold, S., Hoh, F., Domergue, S., Birck, C., Delsuc, M.A., Jullien, M. and Dumas, C. (2000) Characterization and molecular basis of the oligomeric structure of HIV-1 nef protein. *Protein Sci*, **9**, 1137-1148.

Atalla, R.H. and Van der Hart, D.L. (1984) "Native cellulose: a composite of two distinct crystalline forms". *Science*, **223**, 283-285.

[B]

Bae, E. and Phillips, G.N., Jr. (2004) Structures and analysis of highly homologous psychrophilic, mesophilic, and thermophilic adenylate kinases. *J Biol Chem*, **279**, 28202-28208.

Baneyx, F. (1999) Recombinant protein expression in Escherichia coli. *Curr Opin Biotechnol*, **10**, 411-421.

Barnham, K.J., Monks, S.A., Hinds, M.G., Azad, A.A. and Norton, R.S. (1997) Solution structure of a polypeptide from the N terminus of the HIV protein Nef. *Biochemistry*, **36**, 5970-5980.

Barre-Sinoussi, F., Chermann, J.C., Rey, F., Nugeyre, M.T., Chamaret, S., Gruest, J., Dauguet, C., Axler-Blin, C., Vezinet-Brun, F., Rouzioux, C., Rozenbaum, W. and Montagnier, L. (1983) Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). *Science*, **220**, 868-871.

Bayer, E.A., Chanzy, H., Lamed, R. and Shoham, Y. (1998a) Cellulose, cellulases and cellulosomes. *Curr Opin Struct Biol*, **8**, 548-557.

Bayer, E.A., Shimon, L.J., Shoham, Y. and Lamed, R. (1998b) Cellulosomes-structure and ultrastructure. *J Struct Biol*, **124**, 221-234.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. (2000) The Protein Data Bank. *Nucleic Acids Res*, **28**, 235-242.

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.C., Estreicher, A., Gasteiger, E., Martin, M.J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S. and Schneider, M. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res*, **31**, 365-370.

Bracken, A.P., Pasini, D., Capra, M., Prosperini, E., Colli, E. and Helin, K. (2003) EZH2 is downstream of the pRB-E2F pathway, essential for proliferation and amplified in cancer. *Embo J*, **22**, 5323-5335.

Brun, E., Moriaud, F., Gans, P., Blackledge, M.J., Barras, F. and Marion, D. (1997) Solution structure of the cellulose-binding domain of the endoglucanase Z secreted by Erwinia chrysanthemi. *Biochemistry*, **36**, 16074-16086.

Brunger, A.T. (1992) Free R value : a novel statistical quantity for assessing the accuracy of crystal structures. *Nature*, **355**, 472-475.

Brunger, A.T., Adams, P.D., Clore, G.M., DeLano, W.L., Gros, P., Grosse-Kunstleve, R.W., Jiang, J.S., Kuszewski, J., Nilges, M., Pannu, N.S., Read, R.J., Rice, L.M., Simonson, T. and Warren, G.L. (1998) Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Cryst. D*, **54**, 905-921.

Brunger, A.T., Krukowski, A. and Erickson, J.W. (1990) Slow-cooling protocols for crystallographic refinement by simulated annealing. *Acta Crystallogr A*, **46 (Pt 7)**, 585-593.

Brunger, A.T., Kuriyan, M. and Karplus, M. (1987) Crystallographic R factor refinement by molecular dynamics. *Science*, **235**, 458-460.

Bushman, F.D. and Craigie, R. (1991) Activities of human immunodeficiency virus (HIV) integration protein in vitro: specific cleavage and integration of HIV DNA. *Proc Natl Acad Sci U S A*, **88**, 1339-1343.

[C]

Cai, M., Zheng, R., Caffrey, M., Craigie, R., Clore, G.M. and Gronenborn, A.M. (1997) Solution structure of the N-terminal zinc binding domain of HIV-1 integrase. *Nat Struct Biol*, **4**, 567-577.

Cao, R., Wang, L., Wang, H., Xia, L., Erdjument-Bromage, H., Tempst, P., Jones, R.S. and Zhang, Y. (2002) Role of histone H3 lysine 27 methylation in Polycomb-group silencing. *Science*, **298**, 1039-1043.

Castellano, E., Olivia, G. and Navaza, J. (1992) Fast rigid-body for molecular replacement techniques. *J. Appl. Cryst.*, **25**, 281-284.

Chapon, V., Czjzek, M., El Hassouni, M., Py, B., Juy, M. and Barras, F. (2001) Type II protein secretion in gram-negative pathogenic bacteria: the study of the structure/secretion relationships of the cellulase Cel5 (formerly EGZ) from Erwinia chrysanthemi. *J. Mol. Biol.*, **310**, 1055-1066.

Chen, J.C., Krucinski, J., Miercke, L.J., Finer-Moore, J.S., Tang, A.H., Leavitt, A.D. and Stroud, R.M. (2000) Crystal structure of the HIV-1 integrase catalytic core and C-terminal domains: a model for viral DNA binding. *Proc Natl Acad Sci U S A*, **97**, 8233-8238.

Chen, L., Mathews, F.S., Davidson, V.L., Huizinga, E.G., Vellieux, F.M. and Hol, W.G. (1992) Three-dimensional structure of the quinoprotein methylamine dehydrogenase from Paracoccus denitrificans determined by molecular replacement at 2.8 A resolution. *Proteins*, **14**, 288-299.

Chen, Y.L., Trono, D. and Camaur, D. (1998) The proteolytic cleavage of human immunodeficiency virus type 1 Nef does not correlate with its ability to stimulate virion infectivity. *J Virol*, **72**, 3178-3184.

Chiti, F., Webster, P., Taddei, N., Clark, A., Stefani, M., Ramponi, G. and Dobson, C.M. (1999) Designing conditions for in vitro formation of amyloid protofilaments and fibrils. *Proc Natl Acad Sci U S A*, **96**, 3590-3594.

Clavel, F., Guetard, D., Brun-Vezinet, F., Chamaret, S., Rey, M.A., Santos-Ferreira, M.O., Laurent, A.G., Dauguet, C., Katlama, C., Rouzioux, C. and et al. (1986) Isolation of a new human retrovirus from West African patients with AIDS. *Science*, **233**, 343-346.

Cohen, D.E. and Lee, J.T. (2002) X-chromosome inactivation and the search for chromosome-wide silencers. *Curr Opin Genet Dev*, **12**, 219-224.

Combet, C., Blanchet, C., Geourjon, C. and Deleage, G. (2000) NPS@: network protein sequence analysis. *Trends Biochem Sci*, **25**, 147-150.

Conte, M.R. and Matthews, S. (1998) Retroviral matrix proteins: a structural perspective. *Virology*, **246**, 191-198.

Crowther, R.A. and Blow, D.M. (1967) A method for positioning a known molecule in an unknown crystal structure. *Acta Cryst.*, **23**, 544-548.

Cullen, B.R. (1991) Human immunodeficiency virus as a prototypic complex retrovirus. *J Virol*, **65**, 1053-1056.

[D]

D'Amico, S., Gerday, C. and Feller, G. (2001) Structural determinants of cold adaptation and stability in a large protein. *J Biol Chem*, **276**, 25791-25796.

Daniel, M.D., Kirchhoff, F., Czajak, S.C., Sehgal, P.K. and Desrosiers, R.C. (1992) Protective effects of a live attenuated SIV vaccine with a deletion in the nef gene. *Science*, **258**, 1938-1941.

Dayam, R. and Neamati, N. (2003) Small-molecule HIV-1 integrase inhibitors: the 2001-2002 update. *Curr Pharm Des*, **9**, 1789-1802.

de Backer, M., McSweeney, S., Rasmussen, H.B., Riise, B.W., Lindley, P. and Hough, E. (2002) The 1.9 A crystal structure of heat-labile shrimp alkaline phosphatase. *J Mol Biol*, **318**, 1265-1274.

Deacon, N.J., Tsykin, A., Solomon, A., Smith, K., Ludford-Menting, M., Hooker, D.J., McPhee, D.A., Greenway, A.L., Ellett, A., Chatfield, C. and et al. (1995) Genomic structure of an attenuated quasi species of HIV-1 from a blood transfusion donor and recipients. *Science*, **270**, 988-991.

Demirjian, D.C., Moris-Varas, F. and Cassidy, C.S. (2001) Enzymes from extremophiles. *Curr Opin Chem Biol*, **5**, 144-151.

Denisenko, O.N. and Bomsztyk, K. (1997) The product of the murine homolog of the Drosophila extra sex combs gene displays transcriptional repressor activity. *Mol Cell Biol*, **17**, 4707-4717.

Deprez, E., Tauc, P., Leh, H., Mouscadet, J.F., Auclair, C. and Brochon, J.C. (2000) Oligomeric states of the HIV-1 integrase as measured by time-resolved fluorescence anisotropy. *Biochemistry*, **39**, 9275-9284.

Din, N., Forsythe, I.J., Burtnick, L.D., Gilkes, N.R., Miller, R.C., Jr., Warren, R.A. and Kilburn, D.G. (1994) The cellulose-binding domain of endoglucanase A (CenA) from Cellulomonas fimi: evidence for the involvement of tryptophan residues in binding. *Mol Microbiol*, **11**, 747-755.

Ducros, V., Czjzek, M., Belaich, A., Gaudin, C., Fierobe, H.P., Belaich, J.P., Davies, G.J. and Haser, R. (1995) Crystal structure of the catalytic domain of a bacterial cellulase belonging to family 5. *Structure.*, **3**, 939-949.

Ducruix, A. and Giégé, R. (1992) Crystallisation of nucleic acids and proteins : a practical approch. In Press, I. (ed.). Oxford University Press.

Dyda, F., Hickman, A.B., Jenkins, T.M., Engelman, A., Craigie, R. and Davies, D.R. (1994) Crystal structure of the catalytic domain of HIV-1 integrase: similarity to other polynucleotidyl transferases. *Science*, **266**, 1981-1986.

[E]

Eijkelenboom, A.P., van den Ent, F.M., Vos, A., Doreleijers, J.F., Hard, K., Tullius, T.D., Plasterk, R.H., Kaptein, R. and Boelens, R. (1997) The solution structure of the amino-terminal HHCC domain of HIV-2 integrase: a three-helix bundle stabilized by zinc. *Curr Biol*, **7**, 739-746.

Engelman, A., Bushman, F.D. and Craigie, R. (1993) Identification of discrete functional domains of HIV-1 integrase and their organization within an active multimeric complex. *Embo J*, **12**, 3269-3275.

Esnouf, R.M. (1997) Polyalanine reconstruction from Calpha positions using the program CALPHA can aid initial phasing of data by molecular replacement procedures. *Acta Cryst. D*, **53**, 665-672.

[F]

Farnet, C.M. and Bushman, F.D. (1997) HIV-1 cDNA integration: requirement of HMG I(Y) protein for function of preintegration complexes in vitro. *Cell*, **88**, 483-492.

Faure, A., Calmels, C., Desjobert, C., Castroviejo, M., Caumont-Sarcos, A., Tarrago-Litvak, L., Litvak, S. and Parissi, V. (2005) HIV-1 integrase crosslinked oligomers are active in vitro. *Nucleic Acids Res*, **33**, 977-986.

Faust, C., Schumacher, A., Holdener, B. and Magnuson, T. (1995) The eed mutation disrupts anterior mesoderm production in mice. *Development*, **121**, 273-285.

Feller, G. and Gerday, C. (1997) Psychrophilic enzymes: molecular basis of cold adaptation. *Cell Mol Life Sci*, **53**, 830-841.

Feller, G. and Gerday, C. (2003) Psychrophilic enzymes : hot topics in cold adaptation. *Nat. Rev. Microbiol.*, **1**, 200-208.

Feller, G., Lonhienne, T., Deroanne, C., Libioulle, C., Van Beeumen, J. and Gerday, C. (1992) Purification, characterization, and nucleotide sequence of the thermolabile alpha-amylase from the antarctic psychrotroph Alteromonas haloplanctis A23. *J Biol Chem*, **267**, 5217-5221.

Feller, G., Narinx, E., Arpigny, J.L., Aittaleb, M., Baise, E., Genicot, S. and Gerday, C. (1996) Enzymes from psychrophilic organisms. *FEMS Microbiol. Rev.*, **18**, 189-202.

Fields, S. and Song, O. (1989) A novel genetic system to detect protein-protein interactions. *Nature*, **340**, 245-246.

Fikkert, V., Van Maele, B., Vercammen, J., Hantson, A., Van Remoortel, B., Michiels, M., Gurnari, C., Pannecouque, C., De Maeyer, M., Engelborghs, Y., De Clercq, E., Debyser, Z. and Witvrouw, M. (2003) Development of resistance against diketo derivatives of human immunodeficiency virus type 1 by progressive accumulation of integrase mutations. *J Virol*, **77**, 11459-11470.

Finet, S., Vivares, D., Bonnete, F. and Tardieu, A. (2003) Controlling biomolecular crystallization by understanding the distinct effects of PEGs and salts on solubility. *Methods Enzymol*, **368**, 105-129.

Frankel, A.D. and Young, J.A. (1998) HIV-1: fifteen proteins and an RNA. *Annu Rev Biochem*, **67**, 1-25.

[G]

Gaboriaud, C., Bissery, V., Benchetrit, T. and Mornon, J.P. (1987) Hydrophobic cluster analysis: an efficient new way to compare and analyse amino acid sequences. *FEBS Lett*, **224**, 149-155.

Gallay, P., Hope, T., Chin, D. and Trono, D. (1997) HIV-1 infection of nondividing cells through the recognition of integrase by the importin/karyopherin pathway. *Proc Natl Acad Sci U S A*, **94**, 9825-9830.

Garsoux, G. (2002) Adaptations des enzymes psychrophiles aux basses températures : etude de l'endoglucanase isolée de la souche antarctique Pseudoalteromonas haloplanktis TAB23. Université de Liège, Liège, p. 200.

Garsoux, G., Lamotte, J., Gerday, C. and Feller, G. (2004) Kinetic and structural optimization to catalysis at low temperatures in a psychrophilic cellulase from the Antarctic bacterium Pseudoalteromonas haloplanktis. *Biochem J*, **384**, 247-253.

Gaudet, R., Bohm, A. and Sigler, P.B. (1996) Crystal structure at 2.4 angstroms resolution of the complex of transducin betagamma and its regulator, phosducin. *Cell*, **87**, 577-588.

Gerday, C., Aittaleb, M., Arpigny, J.L., Baise, E., Chessa, J.P., Garsoux, G., Petrescu, I. and Feller, G. (1997) Psychrophilic enzymes: a thermodynamic challenge. *Biochim Biophys Acta*, **1342**, 119-131.

Gerday, C., Aittaleb, M., Bentahir, M., Chessa, J.P., Claverie, P., Collins, T., D'Amico, S., Dumont, J., Garsoux, G., Georlette, D., Hoyoux, A., Lonhienne, T., Meuwis, M.A. and Feller, G. (2000) Cold-adapted enzymes: from fundamentals to biotechnology. *Trends Biotechnol*, **18**, 103-107.

Gilbert, H.G. and Hazlewood, G.P. (1993) Bacterial cellulase and xylanases. *J. Gen. Microbiol.*, **139**, 187-194.

Gilkes, N.R., Henrissat, B., Kilburn, D.G., Miller, R.C., Jr. and Warren, R.A. (1991) Domains in microbial beta-1, 4-glycanases: sequence conservation, function, and enzyme families. *Microbiol Rev*, **55**, 303-315.

Gilkes, N.R., Kilburn, D.G., Miller, R.C., Jr., Warren, R.A., Sugiyama, J., Chanzy, H. and Henrissat, B. (1993) Visualization of the adsorption of a bacterial endo-beta-1,4-glucanase and its isolated cellulose-binding domain to crystalline cellulose. *Int J Biol Macromol*, **15**, 347-351.

Gilliland, G.L. (1994) Biological Macromolecule Crystallization Database, Version 3.0: new features, data and the NASA archive for protein crystal growth data. *Acta Crystallogr D Biol Crystallogr*, **50**, 408-413.

Gouet, P., Courcelle, E., Stuart, D.I. and Metoz, F. (1999) ESPript: analysis of multiple sequence alignments in PostScript. *Bioinformatics*, **15**, 305-308.

Greenwald, J., Le, V., Butler, S.L., Bushman, F.D. and Choe, S. (1999) The mobility of an HIV-1 integrase active site loop is correlated with catalytic activity. *Biochemistry*, **38**, 8892-8898.

Grzesiek, S., Bax, A., Clore, G.M., Gronenborn, A.M., Hu, J.S., Kaufman, J., Palmer, I., Stahl, S.J. and Wingfield, P.T. (1996) The solution structure of HIV-1 Nef reveals an unexpected fold and permits delineation of the binding surface for the SH3 domain of Hck tyrosine protein kinase. *Nat Struct Biol*, **3**, 340-345.

Grzesiek, S., Bax, A., Hu, J.S., Kaufman, J., Palmer, I., Stahl, S.J., Tjandra, N. and Wingfield, P.T. (1997) Refined solution structure and backbone dynamics of HIV-1 Nef. *Protein Sci*, **6**, 1248-1263.

Guermeur, Y., Geourjon, C., Gallinari, P. and Deleage, G. (1999) Improved Performance in Protein Secondary Structure Prediction by Inhomogeneous Score Combination. *Bioinformatics*, **15**, 413-421.

Gutjahr, T., Frei, E., Spicer, C., Baumgartner, S., White, R.A. and Noll, M. (1995) The Polycomb-group gene, extra sex combs, encodes a nuclear member of the WD-40 repeat family. *Embo J*, **14**, 4296-4306.

[H]

Harker, D. (1956) The detremination of the phases of the structure factors on non centrosymmetric crystals by the method of double isomorphous replacement. *Acta Crystallogr.*, **9**, 1-7.

Hendrickson, W.A. (1991) Determination of macromolecular structures from anomalous diffraction of synchrotron radiation. *Science*, **254**, 51-58.

Henriksson, G., Nutt, A., Henriksson, H., Pettersson, B., Stahlberg, J., Johansson, G. and Pettersson, G. (1999) Endoglucanase 28 (Cel12A), a new Phanerochaete chrysosporium cellulase. *Eur J Biochem*, **259**, 88-95.

Henrissat, B. (1985) Structure et réactivité enzymatique de la cellulase. Université de Grenoble, Grenoble.

Henrissat, B. (1991) A classification of glycosyl hydrolases based on amino acid sequence similarities. *Biochem J*, **280 (Pt 2)**, 309-316.

Henrissat, B. and Bairoch, A. (1996) Updating the sequence-based classification of glycosyl hydrolases. *Biochem J*, **316 (Pt 2)**, 695-696.

Henrissat, B. and Davies, G. (1997) Structural and sequence-based classification of glycoside hydrolases. *Curr Opin Struct Biol*, **7**, 637-644.

Hill, C.P., Worthylake, D., Bancroft, D.P., Christensen, A.M. and Sundquist, W.I. (1996) Crystal structures of the trimeric human immunodeficiency virus type 1 matrix protein: implications for membrane association and assembly. *Proc Natl Acad Sci U S A*, **93**, 3099-3104.

Holdener, B.C., Thomas, J.W., Schumacher, A., Potter, M.D., Rinchik, E.M., Sharan, S.K. and Magnuson, T. (1995) Physical localization of eed: a region of mouse chromosome 7 required for gastrulation. *Genomics*, **27**, 447-456.

Hong, S.S. and Boulanger, P. (1995) Protein ligands of the human adenovirus type 2 outer capsid identified by biopanning of a phage-displayed peptide library on separate domains of wild-type and mutant penton capsomers. *Embo J*, **14**, 4714-4727.

Hu, J., Xu, Y., Schappert, K., Harrington, T., Wang, A., Braga, R., Mogridge, J. and Friesen, J.D. (1994) Mutational analysis of the PRP4 protein of Saccharomyces cerevisiae suggests domain structure and snRNP interactions. *Nucleic Acids Res*, **22**, 1724-1734.

Hulo, N., Sigrist, C.J., Le Saux, V., Langendijk-Genevaux, P.S., Bordoli, L., Gattiker, A., De Castro, E., Bucher, P. and Bairoch, A. (2004) Recent improvements to the PROSITE database. *Nucleic Acids Res*, **32**, 134-137.

[I]

Irwin, D., Shin, D.H., Zhang, S., Barr, B.K., Sakon, J., Karplus, P.A. and Wilson, D.B. (1998) Roles of the catalytic domain and two cellulose binding domains of Thermomonospora fusca E4 in cellulose hydrolysis. *J Bacteriol*, **180**, 1709-1714.

[J]

Jacks, T., Power, M.D., Masiarz, F.R., Luciw, P.A., Barr, P.J. and Varmus, H.E. (1988) Characterization of ribosomal frameshifting in HIV-1 gag-pol expression. *Nature*, **331**, 280-283.

Jaenicke, R. and Böhm, G. (1998) The stability of proteins in extreme environnements. *Current Opinion in Structural Biology*, **8**, 738-748.

Jancarik, J. and Kim, S.H. (1991) Sparse matrix sampling : A screening method for crystallisation of proteins. *J. Appl. Cryst.*, **24**, 409-411.

Jenkins, T.M., Engelman, A., Ghirlando, R. and Craigie, R. (1996) A soluble active mutant of HIV-1 integrase: involvement of both the core and carboxyl-terminal domains in multimerization. *J Biol Chem*, **271**, 7712-7718.

Jiang, J.S. and Brunger, A.T. (1994) Protein hydration observed by X-ray diffraction. Solvation properties of penicillopepsin and neuraminidase crystal structures. *J Mol Biol*, **243**, 100-115.

Johnson, A.A., Marchand, C. and Pommier, Y. (2004) HIV-1 integrase inhibitors: a decade of research and two drugs in clinical trial. *Curr Top Med Chem*, **4**, 1059-1077.

[K]

Kabsch, W. and Sander, C. (1983) Dictionnary of protein secondary structure : pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577-2637.

Kalpana, G.V., Marmon, S., Wang, W., Crabtree, G.R. and Goff, S.P. (1994) Binding and stimulation of HIV-1 integrase by a human homolog of yeast transcription factor SNF5. *Science*, **266**, 2002-2006.

Keegan, L., Gill, G. and Ptashne, M. (1986) Separation of DNA binding from the transcription-activating function of a eukaryotic regulatory protein. *Science*, **231**, 699-704.

Khan, E., Mack, J.P., Katz, R.A., Kulkosky, J. and Skalka, A.M. (1991) Retroviral integrase domains: DNA binding and the recognition of LTR sequences. *Nucleic Acids Res*, **19**, 851-860.

Kiernan, R.E., Ono, A., Englund, G. and Freed, E.O. (1998) Role of matrix in an early postentry step in the human immunodeficiency virus type 1 life cycle. *J Virol*, **72**, 4116-4126.

Kim, S.Y., Hwang, K.Y., Kim, S.H., Sung, H.C., Han, Y.S. and Cho, Y. (1999) Structural basis for cold adaptation. Sequence, biochemical properties, and crystal structure of malate dehydrogenase from a psychrophile Aquaspirillium arcticum. *J Biol Chem.*, **274**, 11761-11767.

King, R.D. and Sternberg, M.J. (1996) Identification and application of the concepts important for accurate and reliable protein secondary structure prediction. *Protein Sci*, **5**, 2298-2310.

Kirmizis, A., Bartley, S.M. and Farnham, P.J. (2003) Identification of the polycomb group protein SU(Z)12 as a potential molecular target for human cancer therapy. *Mol Cancer Ther*, **2**, 113-121.

Kleywegt, G.J. and Brunger, A.T. (1996) Checking your imagination : applications of the R free value. *Structure*, **4**, 897-904.

Koshland, D.E. (1953) Stereochemistry and the mechanism of enzymatic reactions. *Biol. Rev. Camb. Philos. Soc.*, **28**, 416-436.

Kraulis, P.J. (1991) MOLSCRIPT : a program to produce both detailed and schematic plots of protein structures. *J. Appl. Cryst.*, **24**, 946-950.

Kuzmichev, A., Jenuwein, T., Tempst, P. and Reinberg, D. (2004) Different EZH2-containing complexes target methylation of histone H1 or nucleosomal histone H3. *Mol Cell*, **14**, 183-193.

Kuzmichev, A., Margueron, R., Vaquero, A., Preissner, T.S., Scher, M., Kirmizis, A., Ouyang, X., Brockdorff, N., Abate-Shen, C., Farnham, P. and Reinberg, D. (2005) Composition and histone substrates of polycomb repressive group complexes change during cellular differentiation. *Proc Natl Acad Sci U S A*, **102**, 1859-1864.

Kuzmichev, A., Nishioka, K., Erdjument-Bromage, H., Tempst, P. and Reinberg, D. (2002) Histone methyltransferase activity associated with a human multiprotein complex containing the Enhancer of Zeste protein. *Genes Dev*, **16**, 2893-2905.

[L]

Lachner, M., O'Carroll, D., Rea, S., Mechtler, K. and Jenuwein, T. (2001) Methylation of histone H3 lysine 9 creates a binding site for HP1 proteins. *Nature*, **410**, 116-120.

Laemmli, U.K. (1970) Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature*, **227**, 680-685.

Lambright, D.G., Sondek, J., Bohm, A., Skiba, N.P., Hamm, H.E. and Sigler, P.B. (1996) The 2.0 A crystal structure of a heterotrimeric G protein. *Nature*, **379**, 311-319.

Langsford, M.L., Gilkes, N.R., Singh, B., Moser, B., Miller, R.C., Jr., Warren, R.A. and Kilburn, D.G. (1987) Glycosylation of bacterial cellulases prevents proteolytic cleavage between functional domains. *FEBS Lett*, **225**, 163-167.

Laskowski, R.A., MacArthur, M.W. and Moss, D.S. (1993) PROCHECK : a program to check the stereochemical quality of proteins structures. *J. Appl. Cryst.*, **26**, 283-291.

Lee, C.H., Saksela, K., Mirza, U.A., Chait, B.T. and Kuriyan, J. (1996) Crystal structure of the conserved core of HIV-1 Nef complexed with a Src family SH3 domain. *Cell*, **85**, 931-942.

Leiros, I., Moe, E., Lanes, O., Smalas, A.O. and Willassen, N.P. (2003) The structure of uracil-DNA glycosylase from Atlantic cod (Gadus morhua) reveals cold-adaptation features. *Acta Crystallogr D Biol Crystallogr*, **59**, 1357-1365.

Leslie, A.G.W. (1987) A reciprocal space method for calculating a molecular envelope using the algorithm of B.C. Wang. *Acta Cryst. A*, **43**, 134-136.

Linder, M. and Teeri, T.T. (1997) The roles and function of cellulose-binding domains. *J. Biotechnol.*, **57**, 15-28.

Liu, L.X., Heveker, N., Fackler, O.T., Arold, S., Le Gall, S., Janvier, K., Peterlin, B.M., Dumas, C., Schwartz, O., Benichou, S. and Benarous, R. (2000) Mutation of a conserved residue (D123) required for oligomerization of human immunodeficiency virus type 1 Nef protein abolishes interaction with human thioesterase and results in impairment of Nef biological functions. *J Virol*, **74**, 5310-5319.

Lodi, P.J., Ernst, J.A., Kuszewski, J., Hickman, A.B., Engelman, A., Craigie, R., Clore, G.M. and Gronenborn, A.M. (1995) Solution structure of the DNA binding domain of HIV-1 integrase. *Biochemistry*, **34**, 9826-9833.

[M]

Madrona, A.Y. and Wilson, D.K. (2004) The structure of Ski8p, a protein regulating mRNA degradation: Implications for WD protein structure. *Protein Sci*, **13**, 1557-1565.

Mager, J., Montgomery, N.D., de Villena, F.P. and Magnuson, T. (2003) Genome imprinting regulated by the mouse Polycomb group protein Eed. *Nat Genet*, **33**, 502-507.

Mansharamani, M., Graham, D.R., Monie, D., Lee, K.K., Hildreth, J.E., Siliciano, R.F. and Wilson, K.L. (2003) Barrier-to-autointegration factor BAF binds p55 Gag and matrix and is a host component of human immunodeficiency virus type 1 virions. *J Virol*, **77**, 13084-13092.

Massiah, M.A., Starich, M.R., Paschall, C., Summers, M.F., Christensen, A.M. and Sundquist, W.I. (1994) Three-dimensional structure of the human immunodeficiency virus type 1 matrix protein. *J Mol Biol*, **244**, 198-223.

Massiah, M.A., Worthylake, D., Christensen, A.M., Sundquist, W.I., Hill, C.P. and Summers, M.F. (1996) Comparison of the NMR and X-ray structures of the HIV-1 matrix protein: evidence for conformational changes during viral assembly. *Protein Sci*, **5**, 2391-2398.

Matthews, S., Barlow, P., Boyd, J., Barton, G., Russell, R., Mills, H., Cunningham, M., Meyers, N., Burns, N., Clark, N. and et al. (1994) Structural similarity between the p17 matrix protein of HIV-1 and interferon-gamma. *Nature*, **370**, 666-668.

Matthews, S., Barlow, P., Clark, N., Kingsman, S., Kingsman, A. and Campbell, I. (1995) Refined solution structure of p17, the HIV matrix protein. *Biochem Soc Trans*, **23**, 725-729.

Mattinen, M.L., Linder, M., Drakenberg, T. and Annila, A. (1998) Solution structure of the cellulose-binding domain of endoglucanase I from Trichoderma reesei and its interaction with cello-oligosaccharides. *Eur J Biochem*, **256**, 279-286.

McCarter, J.D. and Withers, S.G. (1994) Mechanisms of enzymatic glycoside hydrolysis. *Curr Opin Struct Biol*, **4**, 885-892.

Merz, A., Knochel, T., Jansonius, J.N. and Kirschner, K. (1999) The hyperthermostable indoleglycerol phosphate synthase from Thermotoga maritima is destabilized by mutational disruption of two solvent-exposed salt bridges. *J Mol Biol*, **288**, 753-763.

Miyazaki, K., Wintrode, P.L., Grayling, R.A., Rubingh, D.N. and Arnold, F.H. (2000) Directed evolution study of temperature adaptation in a psychrophilic enzyme. *J Mol Biol*, **297**, 1015-1026.

Moreau, K., Faure, C., Violot, S., Gouet, P., Verdier, G. and Ronfort, C. (2004) Mutational analyses of the core domain of Avian Leukemia and Sarcoma Viruses integrase: critical residues for concerted integration and multimerization. *Virology*, **318**, 566-581.

Moreau, K., Faure, C., Violot, S., Verdier, G. and Ronfort, C. (2003) Mutations in the C-terminal domain of ALSV (Avian Leukemia and Sarcoma Viruses) integrase alter the concerted DNA integration process in vitro. *Eur J Biochem*, **270**, 4426-4438.

[N]

Narinx, E., Baise, E. and Gerday, C. (1997) Subtilisin from psychrophilic antarctic bacteria: characterization and site-directed mutagenesis of residues possibly involved in the adaptation to cold. *Protein Eng*, **10**, 1271-1279.

Navaza, J. (1994) AMoRe: An automated package for molecular replacement. *Acta Cryst. A*, **50**, 157-163.

Neer, E.J., Schmidt, C.J., Nambudripad, R. and Smith, T.F. (1994) The ancient regulatory-protein family of WD-repeat proteins. *Nature*, **371**, 297-300.

Neer, E.J. and Smith, T.F. (1996) G protein heterodimers: new structures propel new questions. *Cell*, **84**, 175-178.

Ng, J., Li, R., Morgan, K. and Simon, J. (1997) Evolutionary conservation and predicted structure of the Drosophila extra sex combs repressor protein. *Mol Cell Biol*, **17**, 6663-6672.

Nicholls, A., Sharp, K.A. and Honig, B. (1991) Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins.*, **11**, 281-296.

[O]

Ong, E., Kilburn, D.G., Miller, R.C., Jr. and Warren, R.A. (1994) Streptomyces lividans glycosylates the linker region of a beta-1,4-glycanase from Cellulomonas fimi. *J Bacteriol*, **176**, 999-1008.

Orlicky, S., Tang, X., Willems, A., Tyers, M. and Sicheri, F. (2003) Structural basis for phosphodependent substrate selection and orientation by the SCFCdc4 ubiquitin ligase. *Cell*, **112**, 243-256.

[P]

Parsiegla, G., Juy, M., Reverbel-Leroy, C., Tardif, C., Belaich, J.P., Driguez, H. and Haser, R. (1998) The crystal structure of the processive endocellulase CelF of Clostridium cellulolyticum in complex with a thiooligosaccharide inhibitor at 2.0 A resolution. *Embo J*, **17**, 5551-5562.

Peytavi, R. (1999) Mise en évidence et caractérisation d'un partenaire cellulaire de la matrice du HIV-1. *Biochimie, Biologie moléculaire et cellulaire*. Université Montpellier II, Montpellier, p. 146.

Peytavi, R., Hong, S.S., Gay, B., d'Angeac, A.D., Selig, L., Benichou, S., Benarous, R. and Boulanger, P. (1999) HEED, the product of the human homolog of the murine eed gene, binds to the matrix protein of HIV-1. *J Biol Chem*, **274**, 1635-1645.

Pickles, L.M., Roe, S.M., Hemingway, E.J., Stifani, S. and Pearl, L.H. (2002) Crystal structure of the C-terminal WD40 repeat domain of the human Groucho/TLE1 transcriptional corepressor. *Structure (Camb)*, **10**, 751-761.

Pirrotta, V., Poux, S., Melfi, R. and Pilyugin, M. (2003) Assembly of Polycomb complexes and silencing mechanisms. *Genetica*, **117**, 191-197.

Popovic, M., Sarngadharan, M.G., Read, E. and Gallo, R.C. (1984) Detection, isolation, and continuous production of cytopathic retroviruses (HTLV-III) from patients with AIDS and pre-AIDS. *Science*, **224**, 497-500.

[R]

Radaev, S. and Sun, P.D. (2002) Crystallization of protein-protein complexes. *J. Appl. Cryst.*, **35**, 674-676.

Ramakrishnan, C. and Ramachandran, G.N. (1965) Stereochemical criteria for polypeptide and protein chain conformations. *Biophys J.*, **5**, 909-933.

Reinikainen, T., Ruohonen, L., Nevanen, T., Laaksonen, L., Kraulis, P., Jones, T.A., Knowles, J.K. and Teeri, T.T. (1992) Investigation of the function of mutated cellulose-binding domains of Trichoderma reesei cellobiohydrolase I. *Proteins*, **14**, 475-482.

Rietzler, M., Bittner, M., Kolanus, W., Schuster, A. and Holzmann, B. (1998) The human WD repeat protein WAIT-1 specifically interacts with the cytoplasmic tails of beta7-integrins. *J Biol Chem*, **273**, 27459-27466.

Robinson, R.C., Turbedsky, K., Kaiser, D.A., Marchand, J.B., Higgs, H.N., Choe, S. and Pollard, T.D. (2001) Crystal structure of Arp2/3 complex. *Science*, **294**, 1679-1684.

Ross, J.M. and Zarkower, D. (2003) Polycomb group regulation of Hox gene expression in C. elegans. *Dev Cell*, **4**, 891-901.

Rossmann, M.G. (1990) The molecular replacement method. *Acta Cryst. A*, **46**, 73-82.

Rossmann, M.G. and Blow, D.M. (1962) The detection of subunits within the asymmetric unit. *Acta Crystallogr.*, **12**, 24-38.

Rost, B. and Sander, C. (1993) Prediction of protein secondary structure at better than 70% accuracy. *J Mol Biol*, **232**, 584-599.

Roussel, A. and Cambillau, C. (1989) TURBO-FRODO. In Committee, S.G. (ed.), *Silicon Graphics Geometry Partners*. Silicon Graphics, Mountain View, California, pp. 77-78.

Russel, R.J., Gerike, U., Danson, M.J., Hough, D.W. and Taylor, G.L. (1998) Structural adaptations of the cold-active citrate synthase from an antarctic bacterium. *Structure*, **6**, 351-361.

Russell, N.J. (1998) Molecular adaptations in psychrophilic bacteria: potential for biotechnological applications. *Adv Biochem Eng Biotechnol*, **61**, 1-21.

[S]

Schroder, A.R., Shinn, P., Chen, H., Berry, C., Ecker, J.R. and Bushman, F. (2002) HIV-1 integration in the human genome favors active genes and local hotspots. *Cell*, **110**, 521-529.

Sewalt, R.G., van der Vlag, J., Gunster, M.J., Hamer, K.M., den Blaauwen, J.L., Satijn, D.P., Hendrix, T., van Driel, R. and Otte, A.P. (1998) Characterization of interactions between the mammalian polycomb-group proteins Enx1/EZH2 and EED suggests the existence of different mammalian polycomb-group protein complexes. *Mol Cell Biol*, **18**, 3586-3595.

Sherman, M.P. and Greene, W.C. (2002) Slipping through the door: HIV entry into the nucleus. *Microbes Infect*, **4**, 67-73.

Shumacher, A., Faust, C. and Magnuson, T. (1996) Positional cloning of a global regulator of anterior-posterior patterning in mice. *Nature*, **383**, 250-253.

Singh, A. and Hayashi, K. (1995) Microbial cellulases: protein architecture, molecular properties, and biosynthesis. *Adv Appl Microbiol*, **40**, 1-44.

Sinnott, M.L. (1990) Catalytic mechanisms of enzymic glycosyl transfer. *Chem. Rev.*, **90**, 1171-1202.

Smith, T.F., Gaitatzes, C., Saxena, K. and Neer, E.J. (1999) The WD repeat: a common architecture for diverse functions. *Trends Biochem Sci*, **24**, 181-185.

Sondek, J., Bohm, A., Lambright, D.G., Hamm, H.E. and Sigler, P.B. (1996) Crystal structure of a G-protein beta gamma dimer at 2.1A resolution. *Nature*, **379**, 369-374.

Spearman, P., Horton, R., Ratner, L. and Kuli-Zade, I. (1997) Membrane binding of human immunodeficiency virus type 1 matrix protein in vivo supports a conformational myristyl switch mechanism. *J Virol*, **71**, 6582-6592.

Sprague, E.R., Redd, M.J., Johnson, A.D. and Wolberger, C. (2000) Structure of the C-terminal domain of Tup1, a corepressor of transcription in yeast. *Embo J*, **19**, 3016-3027.

Struhl, G. (1981) A gene product required for correct initiation of segmental determination in Drosophila. *Nature*, **293**, 36-41.

Sulzenbacher, G., Driguez, H., Henrissat, B., Schulein, M. and Davies, G.J. (1996) Structure of the Fusarium oxysporum endoglucanase I with a nonhydrolyzable substrate analogue: substrate distortion gives rise to the preferred axial orientation for the leaving group. *Biochemistry*, **35**, 15280-15287.

[T]

Tang, C., Loeliger, E., Luncsford, P., Kinde, I., Beckett, D. and Summers, M.F. (2004) Entropic switch regulates myristate exposure in the HIV-1 matrix protein. *Proc Natl Acad Sci U S A*, **101**, 517-522.

Tardieu, A., Bonnete, F., Finet, S. and Vivares, D. (2002) Understanding salt or PEG induced attractive interactions to crystallize biological macromolecules. *Acta Crystallogr D Biol Crystallogr*, **58**, 1549-1553.

Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) CLUSTALW : improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673-4680.

Tomme, P., Warren, R.A. and Gilkes, N.R. (1995) Cellulose hydrolysis by bacteria and fungi. *Adv Microb Physiol*, **37**, 1-81.

Tormo, J., Lamed, R., Chirino, A.J., Morag, E., Bayer, E.A., Shoham, Y. and Steitz, T.A. (1996) Crystal structure of a bacterial family-III cellulose-binding domain: a general mechanism for attachment to cellulose. *Embo J*, **15**, 5739-5751.

[V]

Van den Burg, B., Vriend, G., Veltman, O.R., Venema, G. and Eijsink, V.G. (1998) Engineering an enzyme to resist boiling. *Proc Natl Acad Sci U S A*, **95**, 2056-2060.

van der Vlag, J. and Otte, A.P. (1999) Transcriptional repression mediated by the human polycomb-group protein EED involves histone deacetylation. *Nat Genet*, **23**, 474-478.

Van Petegem, F., Collins, T., Meuwis, M.A., Gerday, C., Feller, G. and Van Beeumen, J. (2003) The structure of a cold-adapted family 8 xylanase at 1.3 A resolution. Structural adaptations to cold and investgation of the active site. *J Biol Chem.*, **278**, 7531-7539.

Varambally, S., Dhanasekaran, S.M., Zhou, M., Barrette, T.R., Kumar-Sinha, C., Sanda, M.G., Ghosh, D., Pienta, K.J., Sewalt, R.G., Otte, A.P., Rubin, M.A. and Chinnaiyan, A.M. (2002) The polycomb group protein EZH2 is involved in progression of prostate cancer. *Nature*, **419**, 624-629.

Violot, S., Hong, S.S., Rakotobe, D., Petit, C., Gay, B., Moreau, K., Billaud, G., Priet, S., Sire, J., Schwartz, O., Mouscadet, J.F. and Boulanger, P. (2003) The human polycomb group EED protein interacts with the integrase of human immunodeficiency virus type 1. *J Virol*, **77**, 12507-12522.

Vivares, D., Belloni, L., Tardieu, A. and Bonnete, F. (2002) Catching the PEG-induced attractive interaction between proteins. *Eur Phys J E Soft Matter*, **9**, 15-25.

Voegtli, W.C., Madrona, A.Y. and Wilson, D.K. (2003) The structure of Aip1p, a WD repeat protein that regulates Cofilin-mediated actin depolymerization. *J Biol Chem*, **278**, 34373-34379.

Vriend, G. (1990) WHAT IF : a molecular modelling and drug design program. *J. Mol. Graph.*, **8**, 52-56.

[W]

Wada, K., Wada, Y., Ishibashi, F., Gojobori, T. and Ikemura, T. (1992) Codon usage tabulated from the GenBank genetic sequence data. *Nucleic Acids Res*, **20 Suppl**, 2111-2118.

Wall, M.A., Coleman, D.E., Lee, E., Iniguez-Lluhi, J.A., Posner, B.A., Gilman, A.G. and Sprang, S.R. (1995) The structure of the G protein heterotrimer Gi alpha 1 beta 1 gamma 2. *Cell*, **83**, 1047-1058.

Wang, J.Y., Ling, H., Yang, W. and Craigie, R. (2001) Structure of a two-domain fragment of HIV-1 integrase: implications for domain organization in the intact protein. *Embo J*, **20**, 7333-7343.

Welker, R., Kottler, H., Kalbitzer, H.R. and Krausslich, H.G. (1996) Human immunodeficiency virus type 1 Nef protein is incorporated into virus particles and specifically cleaved by the viral proteinase. *Virology*, **219**, 228-236.

Williams, F.E. and Trumbly, R.J. (1990) Characterization of TUP1, a mediator of glucose repression in Saccharomyces cerevisiae. *Mol Cell Biol*, **10**, 6500-6511.

Witte, V., Laffert, B., Rosorius, O., Lischka, P., Blume, K., Galler, G., Stilper, A., Willbold, D., D'Aloja, P., Sixt, M., Kolanus, J., Ott, M., Kolanus, W., Schuler, G. and Baur, A.S. (2004) HIV-1 Nef mimics an integrin receptor signal that recruits the polycomb group protein Eed to the plasma membrane. *Mol Cell*, **13**, 179-190.

Wlodaver, A. and Hodgson, K.O. (1975) Crystallization and crystal data for monellin. *Proc. Natl. Acad. Sci. U.S.A*, **72**, 398-399.

Wood, T.M. (1998) Preparation of crystalline, amorphous, and dyed cellulase substrates. *Methods Enzymol.*, **160**, 19-25.

Wu, G., Xu, G., Schulman, B.A., Jeffrey, P.D., Harper, J.W. and Pavletich, N.P. (2003) Structure of a beta-TrCP1-Skp1-beta-catenin complex: destruction motif binding and lysine specificity of the SCF(beta-TrCP1) ubiquitin ligase. *Mol Cell*, **11**, 1445-1456.

[Y]

Yang, Z.N., Mueser, T.C., Bushman, F.D. and Hyde, C.C. (2000) Crystal structure of an active two-domain derivative of Rous sarcoma virus integrase. *J Mol Biol*, **296**, 535-548.

[Z]

Zavodszky, P., Kardos, J., Svingor and Petsko, G.A. (1998) Adjustment of conformational flexibility is a key event in the thermal adaptation of proteins. *Proc Natl Acad Sci U S A*, **95**, 7406-7411.

Zhou, W. and Resh, M.D. (1996) Differential membrane binding of the human immunodeficiency virus type 1 matrix protein. *J Virol*, **70**, 8540-8548.

Zou, S. and Voytas, D.F. (1997) Silent chromatin determines target preference of the Saccharomyces retrotransposon Ty5. *Proc Natl Acad Sci U S A*, **94**, 7412-7416.

Sébastien VIOLOT – Institut de Biologie et Chimie des Protéines
CNRS UMR 5086 – Université Claude bernard Lyon 1
Laboratoire de BioCristallographie (R. Haser)
7 passage du Vercors – 69367 Lyon cedex 07 – FRANCE
s.violot@ibcp.fr – http://www.ibcp.fr – http://www.ibcp.fr/rhaser/