HAL
open science

Approches bioinformatiques et structurales des replicases virales

Francois Ferron

To cite this version:

Francois Ferron. Approches bioinformatiques et structurales des replicases virales. Autre [q-bio.OT]. Université de la Méditerranée - Aix-Marseille II, 2005. Français. NNT : . tel-00010419v1

HAL Id: tel-00010419

https://theses.hal.science/tel-00010419v1

Submitted on 5 Oct 2005 (v1), last revised 7 Oct 2005 (v2)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Université de la Méditerranée

Aix- Marseille II

Faculté des Sciences de Luminy

THESE DE DOCTORAT

Ecole Doctorale des Sciences de la Vie et de la Santé

Présentée par

François-Patrice FERRON

en vue d'obtenir le grade de docteur de l'Université de la Méditerranée

---

Approches bioinformatiques et structurales des réplicases virales

---

Soutenance soutenue le 04 février 2005 devant la commission d'examen:

Pr Jean-Louis ROMETTE	Président de Jury
Dr Olivier POCH	Rapporteur
Dr Félix REY	Rapporteur
Dr Denis GERLIER	Examinateur
Dr Sonia LONGHI	Co-Directeur de thèse
Dr Bruno CANARD	Directeur de thèse

Ils ne savaient pas que c'était impossible, alors ils l'ont fait.

Mark Twain

Remerciements

Je tiens à remercier vivement les membres du jury, M. Olivier Poch, M. Félix Rey, M. Denis Gerlier et M. Jean-Louis Romette pour l'honneur qu'ils me font de juger cette thèse.

Je tiens à remercier la Commission de la Communauté Européenne pour son soutien financier dans le cadre des programmes VirRNADrugPolTarget et SPINE.

Je remercie Christian Cambillau et Bernard Henrissat de m'avoir accueilli au sein du laboratoire et soutenu toutes ces années.

Je voudrais exprimer ma gratitude à Bruno Canard pour m'avoir accueilli dans son équipe, pour sa confiance, sa compréhension, son encadrement et tous ces morceaux de musiques partagés. Merci Bruno, j'ai apprécié le degré de liberté que tu m'as laissé tout en étant là quand j'en avais besoin !

Ma reconnaissance va également à Sonia Longhi pour son encadrement, sa patience (dont j'ai si souvent abusé), sa gentillesse, sa disponibilité, son sens pédagogique et son aide constante. Cela été un réel plaisir de travailler avec toi !
Joelle Boretto pour son aide, ses conseils et sa bonne humeur constante qui donne la pêche. Joelle surtout ne change pas.
Corinne Rancurel pour sa gentillesse, sa disponibilité et sa patience pendant ces années durant lesquelles elle a dû me subir. Corinne, sans toi rien n'aurait été possible !
David Karlin pour ses milliers d'idées (qui fusent dans tous les sens), pour son enthousiasme, sa disponibilité, sa volonté pédagogique et son soutien.
Barbara Sélisko pour sa gentillesse, sa disponibilité, son soutien pendant ces quatre années et pour toutes les discussions passionnées et passionnantes qu'on a eu. Karine Alvarez pour sa disponibilité, son soutien, son sourire et son enthousiasme fantastique. Kenth Hallberg (précédemment Johansson) pour tous nos moments de travail et les autres … . Jean-Marie Bourhis pour son soutien, son humour, sa bonne humeur et sa vivacité d'esprit. Ce fût un plaisir de partager tous ces moments qui resteront gravés dans ma mémoire (je vais acheter un bloc de marbre au cas où !). Un grand merci à Isabelle Imbert pour son aide, son dynamisme, sa joie de vivre, et son soutien constant. Toute cette bonne humeur qui fait s'envoler les

soucis, Isa change pas ! Je remercie Céline De Michelis pour son soutien et sa gentillesse au cours de ces années, bonne chance pour la suite. Claire Debarnot pour son humour, et son dynamisme. Marie-Pierre Egloff pour son soutien constant au cours de ces quatre années et son encadrement précieux au synchrotron. Hélène Dutartre pour ses conseils et sa disponibilité. Jean-Claude Guillemot pour son aide et ses conseils. Delphine Benarroch, pour sa disponibilité et sa gentillesse tout au long de cette thèse. Antoine Frangeul pour son soutien, je suis sûr que bientôt tu vas avoir la main verte pour faire pousser les cristaux ! Cécile Bussetta pour son soutien, sa vivacité d'esprit, sa disponibilité et ces longues discussions. Karine Barral pour sa gentillesse et sa disponibilité.

A tous ceux qui sont et sont passés dans le bureau 8, merci pour tous ces moments, vous avez contribué à changer cette sombre pièce en bureau convivial.

Valérie Campanacci pour son aide précieuse au cours de cette dernière année. Merci à toute l'équipe de glycobiologie du laboratoire de nous avoir parrainé et aidé dans cette aventure.

Je remercie tous les membres de l'équipe, passés ou présents. Jean-Louis, Karine, Boulbaba, Jérôme, Magali, Coralie, Marielle, Ling, Nadia, Deborah, Fred, Philippe. Merci à tous pour votre bonne humeur.

Merci à Eric Blanc, Philippe Cantau, Luciana de Stéphanis, France Chassary, Aurore, pour leur gentillesse et efficacité.

Merci à toutes les personnes de la plateforme SPINE, qui m'ont aidé et gentiment accueilli parmi eux.

Merci à Jérôme Courcambeck pour son soutien et son aide. Merci à Emmanuel Courcelle et Patrice Gouet qui pour leurs aides.

Merci, pour leurs encouragements, et leur soutien indéfectible Jean-Christophe, Sandrine, Jean-Bernard, Latifa, Jérôme, Céline, Fabienne, Olivier, Raphaël, Cent, Emir, Anne, Emeline, David, Maria, Jean-Pierre, YoYo, Valérie, Marie. Toutes ces tranches de vie ou folles soirées ne seront jamais oubliées, c'est que du bonheur et le plus beau reste à venir !!!

Merci, à Rofia pour son soutien et ses encouragements !

Merci à ma famille pour son soutien inconditionnel. Vous avez partagé tous ces moments d'excitations scientifiques et les passages à vides. Ce travail vous est dédié…

Liste des Abréviations

aa	Acide aminé
Å	Angström
ADN	Acide désoxyribonucléique
ARN	Acide ribonucléique
ATP	Adénosine tri phosphate
BLAST	Basic Local Alignment Search Tool. (Altschul et al.) : Algorithme de comparaison de séquences utilisé pour effectuer des recherches dans les banques en utilisant des alignements locaux.
BLOSUM	Matrice de substitution dans laquelle les valeurs sont dérivées de l'observation des fréquences de substitution dans des familles de protéines alignées.
C	Protéine de Capside
C-terminal	Carboxy-terminal
DC	Dichroïsme Circulaire
Dis	Désordonné
DLS	(Dynamic light Scattering) Diffusion dynamique de la lumière
E-value	(Expect value) Nombre d'alignements différents ayant un score égal ou supérieur S que l'on peut espérer trouver par hasard dans les banques. Plus la E-value est basse, plus le score est significatif (BLAST).
F	Protéine de Fusion
FASTA	premier algorithme de recherche de similarité couramment utilisé.
GTP	Guanosine triphosphate
H	Protéine d'attachement
HCA	(Hydrophobic Cluster analysis) Analyse des amas hydrophobes
HD	Domaine hydrophobe
Indels	Espace introduit dans un alignement pour contre balancer une insertion dans une autre séquence.
kb	Kilo bases
kDa	Kilo Dalton
L	Protéine L incluant la polymérase et une méthyltransférase
M	Protéine de Matrice
Mb	Mega bases
MD	Domaine de multimérisation
N	Nucléoprotéine
NCBI	National Center for Biotechnology Information
NS	Protéine non structurale
N-terminal	Amino-terminal
NTP	Nucléotide triphosphate
OMS	Organisation Mondiale de la Santé
P	Phosphoprotéine
PAGE	Poly Acrylamide Gel Electrophoresis
PAM	(Percent Accepted Mutation) Unité introduite par Dayhoff pour quantifier les changements dus à l'évolution dans une séquence protéique.
PSI-BLAST	(Position-Specific Iterated BLAST) : algorithme permettant une recherche itérative en utilisant BLAST.
P-value	Probabilité qu'un alignement apparaisse avec un score égal ou supérieur à un score donné. Les P-value les plus significatives sont celles proches de zéro. Les P-values et les E-values diffèrent dans la manière de représenter la significativité de l'alignement.
RN	Réseaux neuronaux
Se	Score élémentaire
SeV	Virus de Sendai
Sp	Score de pénalité
SRAS	Syndrome Respiratoire Aiguë Sévère
TM	Domaine transmembranaire
UNK	Domaine inconnu
VSV	Virus de la Stomatite Vésiculaire

Résumé

La virologie bénéficie de plus en plus du développement de la bioinformatique. Mon projet de thèse recouvre la gestion de séquences de protéines se rapportant aux virus à ARN simple brin, de polarité négative et positive. J'ai mis en place une base de données VaZyMolO, qui gère les informations structurales et fonctionnelles des protéines, en définissant et en classant des modules. Cette approche a permis l'identification d'un domaine méthyltransférase sur la protéine L des Mononegavirales, et la définition de la modularité des protéines N et P des *Paramyxoviridae*. La cartographie du génome du virus du Syndrome Respiratoire Aigue Sévère réalisée à l'aide de VaZyMolO, a contribué à la résolution structurale de la protéine nsp9 de ce virus. Enfin, je présente une étude incluant évolution et analyse structurale des polymérases des *Flaviviridae*. Dans cette dernière, je propose un modèle de la polymérase du virus GBV-C et un mécanisme d'initiation de la synthèse d'ARN.

Abstract

Virology is increasingly benefiting from bio-informatics approaches. My thesis project concerns the study of protein sequences from single stranded, negative and positive RNA viruses. I have set up VaZyMolO, a database that deals with proteins at a functional and structural level. Based on similarity, it is a tool that defines and classifies protein modules. This approach contributed to the identification of a methyltransferase domain on L proteins of Mononegavirales, and to the definition of the N and P modularity in *Paramyxoviridae*. The mapping of the Severe Acute Respiratory Syndrome virus genome using VaZyMolO contributed to the structural definition of the nsp9 protein of this virus. My thesis also presents an analysis of the *Flaviviridae* polymerases. It combines structural analyses and methods that infer evolutionary relationships. In this latter study, I propose a structural model for the GBV-C polymerase, and suggest a new mechanism of initiation of RNA synthesis.

Mots clés

Bioinformatique, virologie, structure, base de données, polymérase, Coronavirus, *Paramyxoviridae, Flaviviridae*.

Avant-propos

Durant ma thèse, je me suis intéressé à l'étude bioinformatique et structurale des enzymes impliquées dans la réplication virale. Comme le projet VaZyMolO, ce manuscrit est non conventionnel dans sa conception. Il n'y a pas d'introduction bibliographique en tant que telle. Par souci de lisibilité, la recherche bibliographique est intégré à chacune des parties. Ainsi, le manuscrit s'articule autour de cinq chapitres. Les deux premiers chapitres se veulent être le reflet du bilan qu'il a été nécessaire de dresser avant de se lancer dans le projet de la construction de la base de données VaZyMolO. Le premier chapitre replace la biologie structurale dans le contexte du projet et la description de l'organisation générale des bases de données biologiques et leur spécificité. Dans le second chapitre, je fais un rappel sur les outils et les méthodes d'analyses bioinformatiques. Dans la troisième partie, je présente les méthodes d'analyse du désordre structural intrinsèque des protéines et notre approche qui combine différentes techniques.

Je présente ensuite, dans le quatrième chapitre un projet que nous avons développé, VaZyMolO. C'est à la fois une base de données et un outil pour la définition et la classification d'entités structurales et fonctionnelles (modules) de protéines virales, en vue de leur étude tridimensionnelle. Les critères de définition de ces derniers reposent sur la comparaison de séquences, l'analyse des données structurales existantes et l'intégration de données bibliographiques.

Enfin, le cinquième chapitre rapporte les principaux résultats issus de ces analyses. J'ai travaillé en particulier sur la caractérisation de protéines impliquées dans le complexe de réplication de différents virus. Pour chaque article présenté, j'effectuerai un bref rappel sur les généralités de chaque virus étudié. J'ai notamment réalisé l'annotation de la protéine «□L□ (polymérase) des *Mononegavirales*, des protéines N (Nucléoprotéine) et P (Phosphoprotéine) des *Paramyxoviridae*. J'ai aussi participé à l'analyse bioinformatique de la nucléoprotéine du Hantaan Virus appartenant à la famille des *Bunyaviridae*. Lors de l'épidémie du virus du syndrome respiratoire aigu sévère (SRAS), VaZyMolO a prouvé son utilité en permettant l'annotation complète et rapide du génome de ce nouveau membre des *Coronaviridae*. Grâce à ces annotations, nous avons pu cristalliser et résoudre la structure de nsp9, l'une des protéines du complexe de réplication de ce virus. Cette protéine ne possède aucun homologue dans les banques de séquences et sa structure présente un nouveau repliement. Nous verrons comment par une analyse de la structure, nous avons pu postuler la fonction de cette protéine.

Je finirai par la modélisation de la polymérase du virus GBV-C. Lors de cette étude, j'ai travaillé sur la comparaison structurale des polymérases des *Flaviviridae*. J'ai également proposé un mécanisme pour l'initiation de la réplication au sein de cette famille.

Chapitre 1

La biologie structurale et ses bases de données

1. Structure des protéines 16

1.1. Le repliement protéique 16

1.2. Les structures secondaires 17

2.2.1. Les hélices α 17

2.2.2. Les feuillets β 19

1.3. Structure tertiaire 19

1.4. Structure quaternaire 21

1.5. Classification des protéines en fonction de leur repliement 21

2. Stockage et organisation de l'information biologique 22

2.1. But, organisation et historique 22

2.2. Les données brutes 23

2.2.1. Les banques nucléotidiques 23

2.2.2. Les banques protéiques 24

2.2.3. Les bases de données structurales 27

2.3. Données à forte valeur ajoutée 30

2.3.1. Classification et analyses structurales 30

2.3.2. Bases de données de motifs 33

2.3.3. Les bases de données de domaines de protéines 34

2.3.4. Les bases de données fonctionnelles et taxonomiques 37

2.3.5. Bases de données virales 38

Chapitre 2

Principes et méthodes de l'analyse de séquences

1. Introduction 40

2. Méthode d'analyse de la structure primaire 41

2.1. Généralités sur l'alignement de séquences 41

2.2. Alignement de deux séquences et algorithme de recherche dans les bases de données 44

2.2.1. Méthode globale 46

2.2.2. Méthode locale 47

2.3. Alignements multiples 49

2.4. Domaines et motifs 51

2.4.1. Les domaines 51

2.4.2. Les motifs 51

2.5. Méthodes de construction d'arbres phylogénétiques 52

2.5.1. UPGMA (Unweight Pair Group Method with Arithmetic mean) 52

2.5.2. NJ (Neighbor-Joining) 52

2.5.3. Parcimonie 53

2.5.4. Maximum de vraissemblance (Maximum Likelihood) 53

3. Méthodes pour la prédiction de structures secondaires 55

3.1. Approches de première génération 55

3.2. Les réseaux neuronaux 55

3.4. Méthodes pour l'identification des régions transmembranaires et peptides signaux 57

3.5. La méthode d'analyse des amas hydrophobes (HCA) 57

4. Méthode pour la prédiction de structure tertiaire 59

4.1. Méthode de « Threading » (enfilage) 59

Chapitre 3

Méthodes de prédiction du désordre structural

ARTICLE 1 Combining prediction methods to achieve accurate recognition of disordered regions in proteins.

Chapitre 4

La base de données VaZyMolO

1. Introduction 81

1.1. Fiabilité des bases de données publiques 81

1.1.1. Perte d'information dans la base 81

1.1.2. Problème du suivi des mises à jour 81

1.1.3. Transfert systématique de l'information sans esprit critique 81

1.2. Absence de bases de données virales 82

1.3. VaZyMolO« Viral enZyme Module lOcalisation » 83

1.3.1. Objectif de la base de données 83

1.3.2. VaZyMolO : base de données relationnelle 83

1.3.3. Organisation des données 84

2. Structuration de l'information 85

2.1. Les séquences 85

2.1.1. Les séquences de protéines 87

2.1.2. Les séquences de polyprotéines 87

2.2. La modularité 89

2.2.1. Les familles de modules 89

2.2.2. Identification des modules 90

2.2.3. Modules de référence 96

2.4. Composition actuelle de VaZyMolO 98

2.4.1. Les génomes 98

2.4.2. Les bibliothèques de séquences de VaZyMolO et de bases annexes 102

3. L'interface de VaZyMolO : le serveur public 102

ARTICLE 2 VaZyMolO: a tool to define and classify modularity in viral proteins.

Chapitre 5

Exemples d'application des méthodes d'analyse de séquences à l'étude des réplicases virales.

1. Etudes des protéines du complexe réplicatif de virus à ARN négatif 123

1.1. Les Mononegavirales 123

1.1.1.Généralités 123

1.1.2. La particule virale des *Paramyxoviridae* 123

1.1.3. La structure de la nucléocapside 124

1.1.4. Organisation du génome viral 124

1.1.5. Description des protéines du complexe réplicatif 125

ARTICLE 3 Viral RNA-Polymerases-A predicted 2'-o-ribose methyltransferase domain shared by all *Mononegavirales*

-Commentaires 131

ARTICLE 4 Structural disorder and modular organization in Paramyxovirinae N and P

-Commentaires 150

1.2. Les Hantavirus 151

1.2.1. Généralités 151

1.2.2. Organisation de la particule virale et du génome des *Bunyaviridae* 151

ARTICLE 5 Essential amino acids of the hantaan virus N protein in its interaction with RNA

2. Etudes des protéines du complexe réplicatif de virus à ARN positif 196

2.1. Les Coronavirus 196

2.1.1. Généralités 196

2.1.2. Organisation de la particule virale et du génome 196

ARTICLE 6 Structural genomics of the SARS coronavirus: cloning, expression, crystallization and preliminary crystallographic study of the Nsp9 protein

ARTICLE 7 The severe acute respiratory syndrome-coronavirus replicative protein nsp9 is a single-stranded RNA-binding subunit unique in the RNA virus world.

-Commentaires 210

2.2. Les virus de l'hépatite G ou virus GB 213

2.2.1. Généralités 213

2.2.2. Organisation de la particule virale et du génome 213

ARTICLE 8 The modeled structure of the RNA dependent RNA polymerase of the GBV-C virus suggests a role for motif E in *Flaviridae* RNA polymerases

-Commentaires 247

Conclusion générale 250

Références 254

Chapitre 1

La biologie structurale et ses bases de données

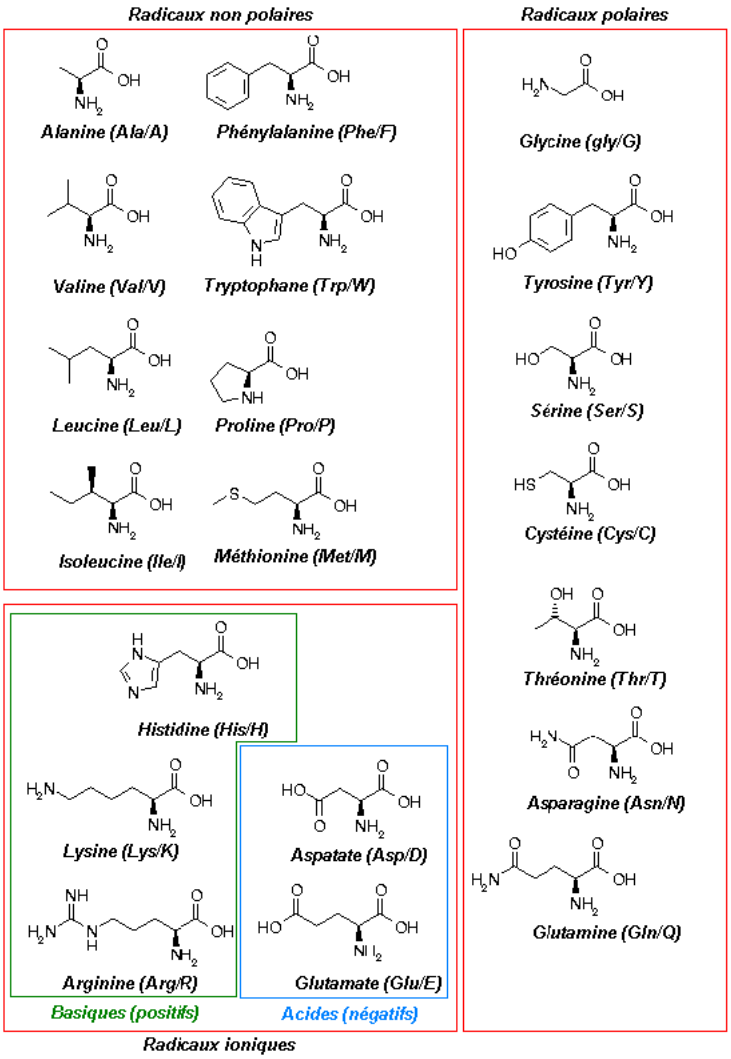


Tableau 1 : Structures des 20 acides aminés, regroupés selon leur propriété.

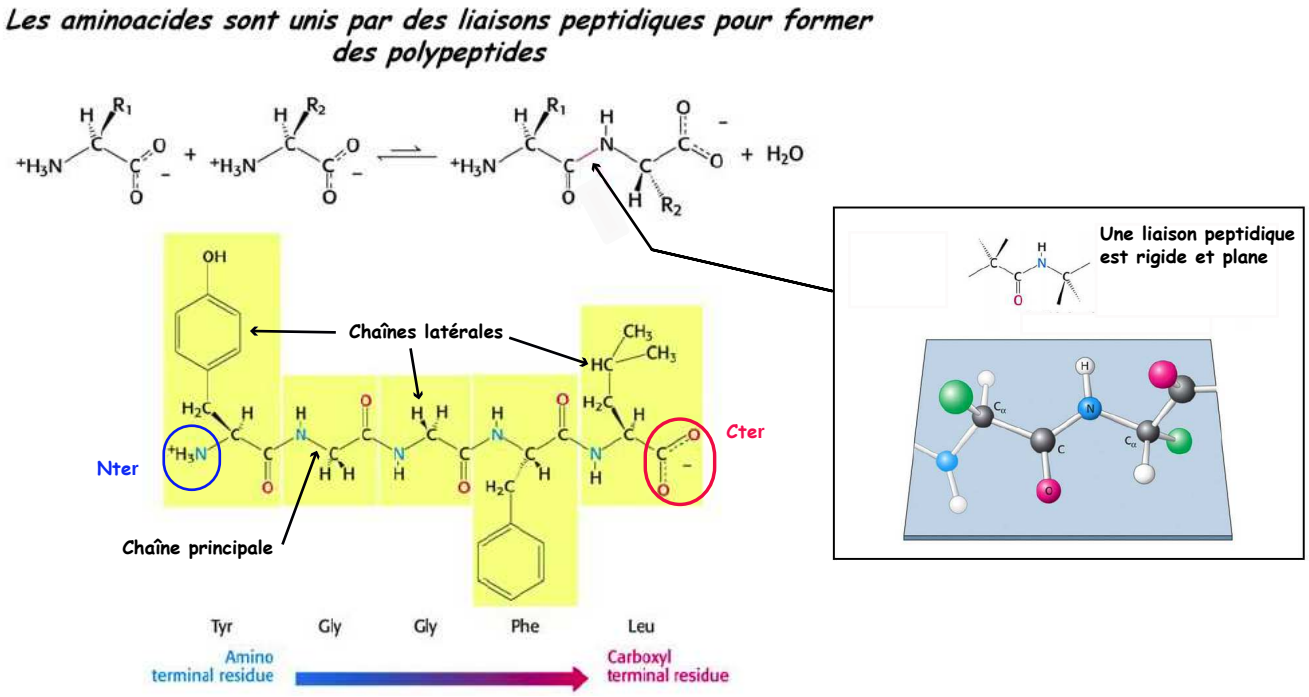


Figure 1 : Organisation des liaisons peptidiques, chaînes principales et latérales.

1. Structure des protéines

Si l'ADN est le support physique de l'information biologique, la protéine en est le reflet, et à l'échelle moléculaire c'est déjà une véritable machine fonctionnelle, assurant à la fois des fonctions vitales aussi bien structurales que dynamiques. Ainsi une protéine est un polymère linéaire constitué de différentes unités de base, les acides aminés (aa), disposées les unes à la suite des autres. Un aa est constitué d'un carbone central (carbone alpha ou Cα) lié à un groupement carboxyle (COOH), à un groupement amine (NH_2), à un atome d'hydrogène (H) et à un radical R. Il existe vingt aa naturels (Tableau 1). Ils se différencient les uns des autres par la nature même de ce radical qui leur confère différentes propriétés telles que, entre autres, la charge, la flexibilité, l'encombrement stérique ou bien encore l'hydrophobicité. Les aa sont reliés entre eux par une liaison peptidique plane formée entre le groupement carboxyle COOH d'un résidu et le groupement amine NH_2 du résidu suivant. La chaîne ainsi formée est appelée « la chaîne principale ou squelette », alors que les radicaux sont désignés sous le terme de « chaînes latérales ». La synthèse des protéines débute par l'extrémité N-terminale (Nter) appelée ainsi car le premier atome de la chaîne est un atome d'azote, et elle se termine par l'extrémité C-terminale (Cter) car le dernier atome de la chaîne est le carbone du groupement carboxyle (Figure 1).

La chaîne principale subit de la part de son environnement des contraintes qui vont entraîner son réarrangement spatial. La succession des aa constituant la chaîne principale définit la séquence protéique dans laquelle sont contenues toutes les informations qui vont déterminer le repliement de la protéine. Cette succession correspond à la structure primaire des protéines, qui s'organise ensuite en quatre niveaux successifs (Figure 2).

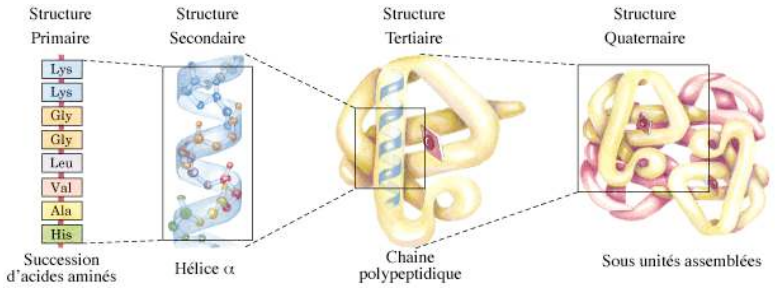


Figure 2 : Repliement des protéines selon les quatre niveaux de structuration.

1.1. Le repliement protéique

L'environnement des protéines est majoritairement aqueux. Certains acides aminés ont une chaîne latérale hydrophobe (Tableau 1). Ceux-là vont avoir tendance à se regrouper pour

masquer leurs chaînes latérales. Ce phénomène est la « compaction hydrophobe » et c'est le moteur essentiel du repliement des protéines globulaires.

Ainsi la chaîne polypeptidique va se replier de façon spontanée, afin d'enfouir les chaînes latérales de ses résidus hydrophobes au cœur de la protéine et à présenter à la surface les chaînes polaires. Cela dit, du fait de la compaction, la chaîne principale passe aussi dans ce cœur hydrophobe, or cette chaîne est fortement hydrophile avec pour chaque unité peptidique un « donneur » de liaisons hydrogènes (NH) et un « accepteur » de liaisons hydrogènes (C=O) (Figure 1). Pour faire passer une chaîne hydrophile au milieu d'un environnement hydrophobe, la solution repose sur la neutralisation des groupements polaires en formant des liaisons hydrogènes (liaisons H). Ces dernières sont optimisées, par l'intermédiaire de la formation de structures secondaires régulières : les hélices alpha (α) et les brins bêta (β).

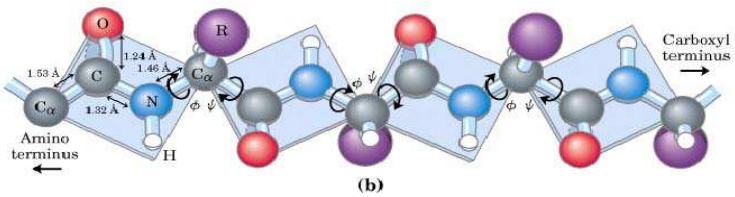
Une fois la protéine repliée, les forces de Van der Waals entre les chaînes latérales ainsi regroupées vont contribuer à la stabilité de la protéine. Il est aussi important de noter que les cystéines tiennent une place particulière dans l'assemblage tridimensionnel des protéines. En effet, deux cystéines éloignées sur la chaîne principale, mais proches dans l'espace tridimensionnel peuvent former une liaison covalente entre les deux atomes de soufre des deux chaînes latérales. Ces liaisons sont appelées « ponts disulfure » et jouent un rôle dans la stabilisation de la structure de la protéine.

1.2. Les structures secondaires

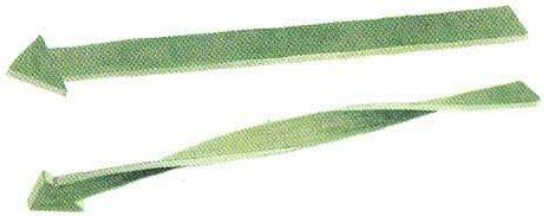
1.2.1. Hélices α

L'hélice α est caractérisée par des liaisons hydrogènes internes entre le groupe CO du résidu et le groupe NH du résidu situé quatre résidus plus loin (Figure 3). Chaque tour comporte 3,6 résidus ce qui correspond à un pas de 5,4 Å (soit 1,5 Å par aa). On notera que les hélices font intervenir des résidus proches les uns des autres le long de la structure primaire.

A **Brin β** **B**



Géométrie du Brin β

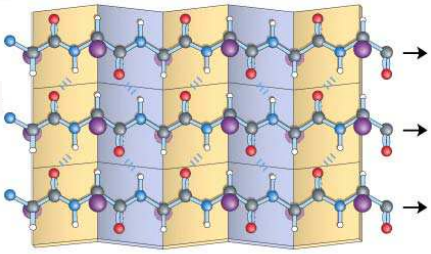


Représentation schématique du Brins β

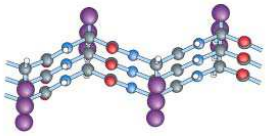
C **Feuillet β importance des chaînes Latérales**

Feuillet Parallèle

Vue de dessus

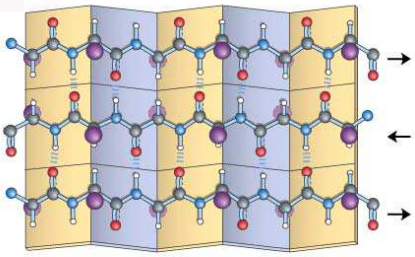


Vue latérale

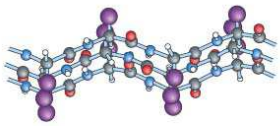


Feuillet antiparallèle

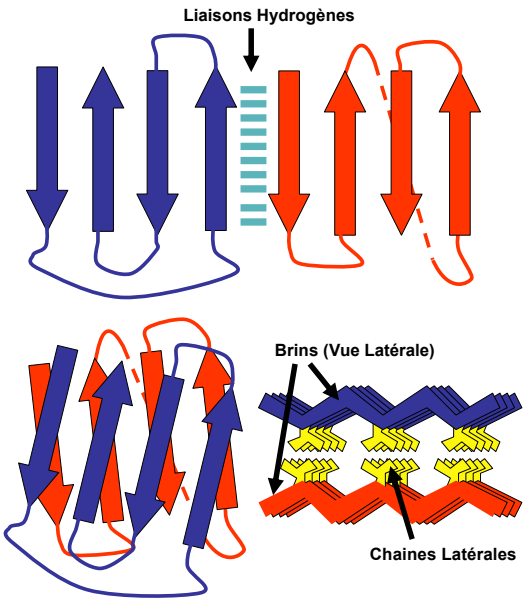
Vue de dessus



Vue latérale



D



Interactions brin-brin
via la chaîne principale
(Liaisons hydrogènes)

Associations de brins β:

Face-à-face
via les chaînes latérales
(Interactions Hydrophobes)

Figure 4 : A) Atomes du squelette constituant le brin. B) Représentation schématique du brin en ruban. C) Formation des feuillets. D) Structuration des feuillets.

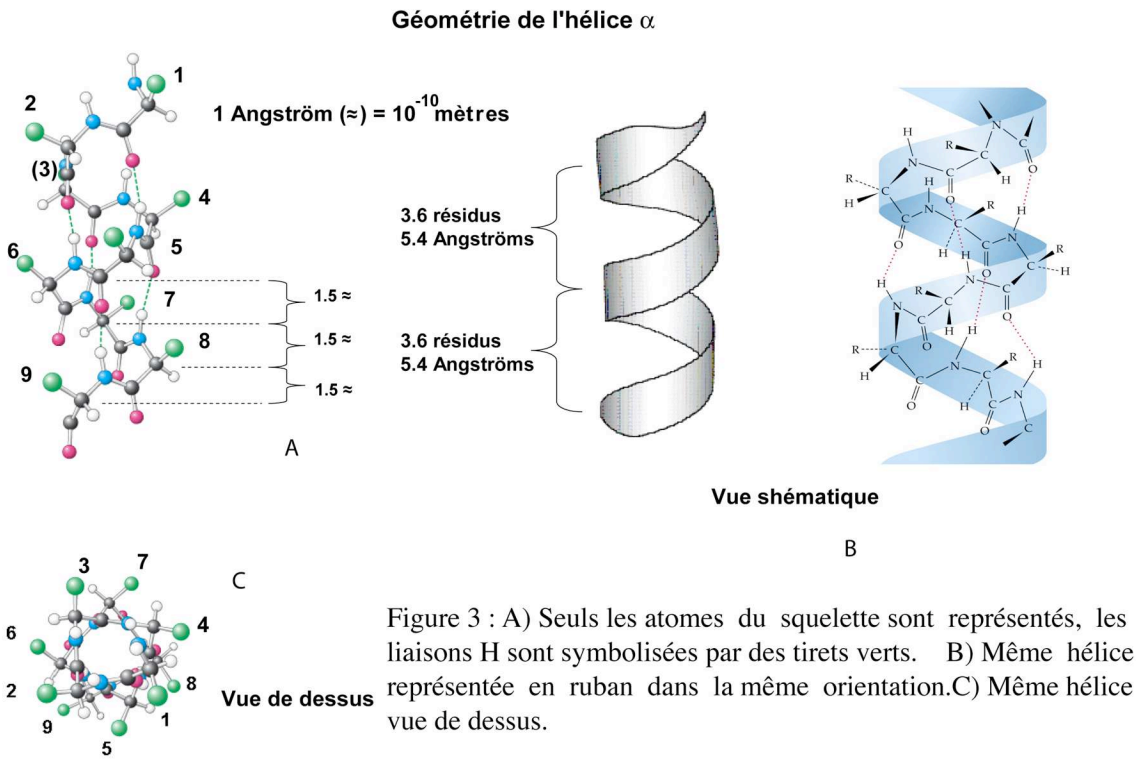


Figure 3 : A) Seuls les atomes du squelette sont représentés, les liaisons H sont symbolisées par des tirets verts. B) Même hélice représentée en ruban dans la même orientation.C) Même hélice vue de dessus.

1.2.2. Les feuillets β

Contrairement aux hélices, les feuillets β font intervenir des régions éloignées les unes des autres le long de la structure primaire, appelée brins β. Les feuillets β sont composés de plusieurs brins, alignés les uns à côtés des autres entre lesquels s'établissent des liaisons hydrogènes. Cet alignement peut s'organiser de trois façons différentes : si deux brins adjacents sont orientés dans le même sens (de Nter vers Cter) on dit alors que le feuillet β est parallèle, si les deux brins sont dans des sens opposés, on dit que le feuillet β est antiparallèle. La troisième façon étant des feuillets mixtes contenant des brins orientés de manière parallèle et antiparallèle (Figure 4).

La plupart des protéines sont composées à partir de combinaisons d'hélices et/ou de brins qui contiennent environ les deux tiers des résidus. Le dernier tiers correspond aux boucles qui relient les hélices et les brins. Ces boucles sont généralement situées à la surface des protéines.

1.3. Structure tertiaire

La structure tertiaire est l'architecture d'une protéine. Elle correspond à l'agencement des différentes structures secondaires entre elles. L'architecture adoptée permet à des aa éloignés

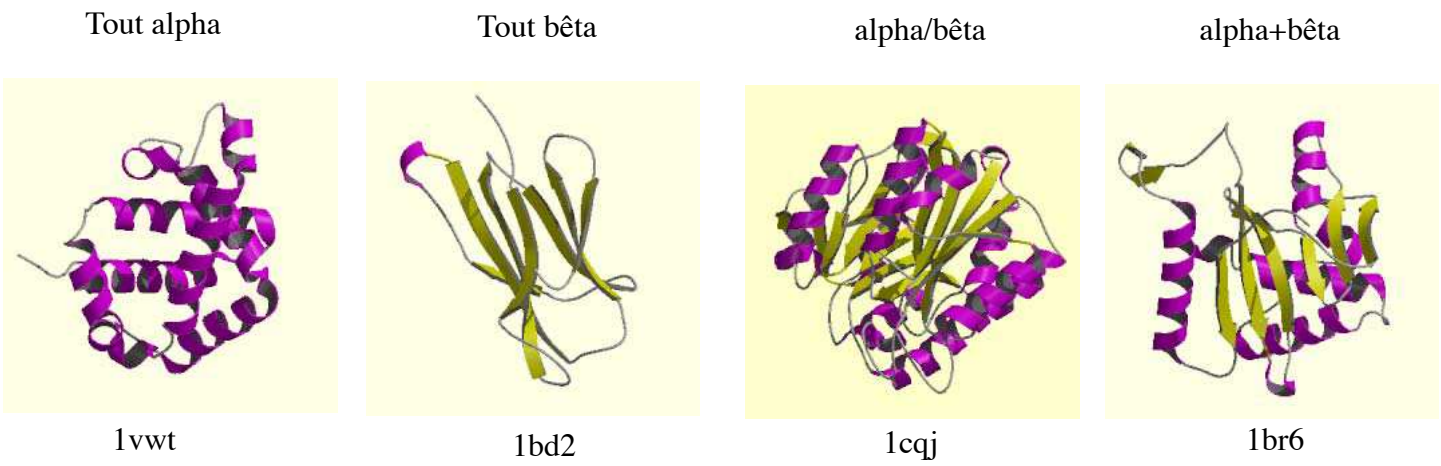


Figure 5 : Différents types de repliement illustrant la variabilité de la structure tertiaire.

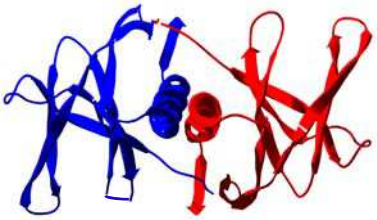


Figure 6 : Assemblage de deux monomères formant la structure quaternaire.

en séquence d'être proches dans l'espace alors que la stabilité de l'ensemble est assurée par différents types de relations entre ces résidus : liaisons ioniques, liaisons hydrogènes, forces de Van der Waals, compaction hydrophobe ou bien encore ponts disulfures. Pour des raisons énergétiques, les types de repliement ne sont pas infinis, il est donc possible de décrire des familles de repliement. Un type particulier de repliement est caractérisé par la disposition relative des structures secondaires régulières entre elles, par leurs connexions et leur succession le long de la séquence (Figure 5).

Il arrive que les protéines possèdent plusieurs fonctions, dans ce cas, pour chaque fonction est associée une structure et les séquences peuvent être découpées en unités de base, que l'on appelle domaines protéiques. Ces domaines sont des structures globulaires et compactes qui adoptent des repliements indépendants. Ces domaines structuraux sont les constituants de l'architecture des protéines et de leurs fonctions. Ainsi, les combinaisons peuvent être variées tant par la nature que par l'ordre. Néanmoins, il est possible de distinguer deux grandes catégories : les domaines continus qui sont formés par des régions peptidiques consécutives qui se replient pour former un domaine indépendant et unique, et les domaines discontinus qui sont composés de régions peptidiques non consécutives. Il est possible de caractériser structurellement un domaine et ce, indépendamment du reste de la protéine. Cette approche en domaines est largement utilisée dans la caractérisation structurale.

1.4. Structure quaternaire

Si une protéine est constituée d'une seule chaîne polypeptidique, on parle de protéine monomérique. Il arrive souvent que l'ultime étape du repliement implique l'assemblage de plusieurs monomères pour former un multimère. Dans ce cas, les monomères peuvent être identiques, on parle alors d'homodimère. S'ils sont différents, on parle d'hétérodimère (Figure 6).

1.5. Classification des protéines en fonction de leur repliement

Sur le plan structural, les protéines peuvent être réparties en cinq classes. La classe « tout α » regroupe les structures qui comme l'hémoglobine sont essentiellement constituées d'hélices (au moins 90% de leurs structures secondaires régulières). La classe « tout β » contient toutes les protéines constituées essentiellement de brins β, comme les immunoglobulines. La classe « α/β » rassemble les protéines dont les structures sont construites par l'alternance d'hélices α et de brins β alors que la classe « α+β » correspond aux structures comportant des hélices et

des brins répartis dans des régions plus ou moins distinctes (l'alternance n'est plus observée). La dernière catégorie rassemble les protéines de petites tailles (constituées souvent de moins de 70 aa) comportant généralement peu de structures secondaires régulières.

2. Stockage et organisation de l'information biologique

2.1. But, organisation et historique

L'émergence et le développement des banques puis des bases de données sont consécutifs au besoin des chercheurs de stocker et d'organiser des quantités de données qu'ils produisent de façon exponentielle. Ces données sont de cinq natures différentes et les exemples suivants renvoient aux bases de données correspondantes

a. ***Les séquences***

Données de séquences nucléiques et protéiques (Exemple : Entrez Proteins, SWISS-PROT…).

b. ***Les données cartographiques*** (non traitées)

Données concernant la localisation des gènes et marqueurs sur les chromosomes. Données issues de la cartographie génétique et transcriptionnelle (Exemple : GDB, GENATLAS…).

c. ***Les données structurales***

Données correspondant aux **structures 3D** des biomolécules des séquences nucléiques (Exemple : NDB) ou des séquences protéiques (Exemple : PDB).

d. ***Les données d'expression*** (non traitées)

Données issues de l'analyse de l'expression génique (analyse des ARNm, ADNc et des EST).

e. ***Les données fonctionnelles, relationnelles et taxonomiques***.

Données métaboliques et relationnelles (Exemple : ENZYME, Biochemical pathway map, ICTV).

La séquence, qu'elle soit nucléique ou protéique, est l'élément central autour duquel les banques de données se sont constituées. Les séquences biologiques, dès leur apparition, ont fait très rapidement l'objet de compilation dans ce que l'on appellera plus tard des banques de données. Ainsi la première compilation de séquences de protéines apparaît dans les années 1960 par Margaret Dayhoff (Dayhoff 1965; 1969) : « ATLAS of protein sequence and structure ». Au début des années 1980, avec la découverte et la mise au point de la technique de séquençage des acides nucléiques, les premières grandes banques de données généralistes de séquences ont vu le jour.

On peut distinguer trois périodes majeures en biologie :

-La période pré-génomique, qui rassemble quelques essais de banques de séquences, ainsi que l'émergence des banques de références bibliographiques ;

-La période génomique, qui a vu le stockage de séquences d'acides nucléiques, de protéines et de structures ;

-La période actuelle post-génomique, qui bien sûr continue à stocker des séquences, mais qui tend aussi à organiser l'information, la recouper, l'enrichir.

Aujourd'hui, il y a près de 540 bases de données disponibles sur l'internet (Galperin 2004). Ce mémoire ne fait pas une description systématique de ces banques mais décrit seulement celles qui présentent le plus de pertinence par rapport au développement de notre projet.

2.2. Les données brutes

2.2.1. Les banques nucléotidiques

Aucun journal scientifique n'accepte de publier un travail décrivant le génome d'un nouvel organisme (ou d'un gène) sans que la séquence ne soit au préalable déposée dans l'une des trois grandes banques (GenBanK (Benson et al. 1994), EMBL (Uchida 1986), DDBJ (Uchida 1986)). Cette clause éditoriale assure indirectement un contrôle sur la qualité des séquences déposées dans les banques. Elles forment aujourd'hui un réseau collaboratif et unifié qui collecte et assure l'accessibilité des séquences à la communauté scientifique. L'information originellement stockée dans l'une d'entre elles peut être retrouvé dans les trois, sous le même identifiant.

Les séquences nucléotidiques des banques de données peuvent être de trois natures différentes :

ADN génomique : compilation de séquences correspondant à la totalité de l'ADN d'une espèce.

ARN : compilation de séquences correspondant à l'ensemble des ARN messagers transcrits à partir du génome (transcriptome). Pour certain virus c'est aussi leur génome.

ADNc : compilation de séquences d'ADN complémentaire.

Les banques nucléotidiques sont essentiellement des banques d'archives qui servent principalement à stocker les séquences déposées. Ces dernières sont sous la responsabilité de la personne qui effectue le dépôt. La banque n'a pas autorité pour modifier les informations déposées. En effet, le travail de vérification ou « curation » se limite à vérifier que chaque entrée a correctement été formatée et qu'il n'y a pas de contamination par des séquences

étrangères, par exemple de vecteur utilisé pour le clonage. Si des erreurs sont commises lors du dépôt ou lorsque des corrections apparaissent, seule la personne qui a déposé la séquence peut éditer l'entrée. Ce règlement permet de préserver l'intégrité du dépôt original. La contrepartie de celui-ci est la remise en cause de la fiabilité des données déposées.

2.2.2. Les banques protéiques

Les premières banques de séquences protéiques contenaient essentiellement des séquences dont la fonction était caractérisée. Elles étaient surtout destinées à l'étude de la variabilité de séquences de protéines homologues. Après l'apparition des méthodes « rapides » de séquençage de l'ADN, le nombre de gènes séquencés, ainsi que de fragments d'ADN ou d'ARNm non caractérisés a considérablement augmenté. La principale source de séquences de protéines est la traduction directe des séquences déposées dans les banques de séquences nucléiques. Les trois grandes banques précédemment citées fournissent ce service. Néanmoins, la croissance rapide de la taille des banques de données et la masse d'informations nouvellement générée, a entraîné *de facto* une baisse de la qualité des données. Pour éviter que les informations collectées deviennent inutilisables et maintenir une certaine qualité des données, des bases de données de séquences de protéines plus ou moins spécialisées ont vu le jour.

<u>- NCBI « protein database ou Entrez Proteins »</u>

Cette base offre le plus grand échantillon de séquences protéiques « déduites ». Chaque séquence de protéine est assignée à un gène unique identifié par un numéro (GI). Si la séquence est modifiée, la nouvelle séquence reçoit un nouvel identifiant (de même si la séquence nucléotidique dont elle est issue est modifiée). Cette méthode de référence permet de ne perdre aucune information sur « l'histoire » de la séquence dans la base, mais le corollaire est une base de donnée excessivement large et redondante. Le problème de la taille et de la redondance est d'autant plus vrai que « Entrez Proteins » intègre les entrées SWISS-PROT, TrEMBL et PIR (voir paragraphes suivant). Bien que cette base soit la plus complète, elle devient très vite inutilisable. Le NCBI, conscient de ce problème, maintient en parallèle une banque de données de séquences non-redondantes (NR). Dans NR toutes les séquences identiques provenant du même organisme et tous les fragments sont répertoriés comme une seule entrée. De fait NR inclus tous les variants (même ceux qui sont dus aux erreurs de séquençage ou aux erreurs de traitement dans les bases de données). Les séquences dans

« Entrez Proteins » sont très souvent reliées à d'autres bases de données, soit externes (telles que SWISS-PROT, ICTV, PIR), soit internes au NCBI (Wheeler et al. 2001). Ces dernières assurent des services utiles pour les utilisateurs tels que : les informations bibliographiques (PubMED, MEDline), la classification taxonomique (NCBI Taxonomy database), l'information structurale (MMDB (Marchler-Bauer et al. 1999)) ou modulaire (CDD).

- SWISS-PROT

C'est la base de données de référence en ce qui concerne les séquences protéiques. Contrairement au NCBI où il n'y a pas de contrôle sur les séquences déposées (hormis par la personne qui dépose la séquence), la SWISS-PROT (Bairoch and Boeckmann 1991) est « validée » par des opérateurs spécialisés. Le projet initié par Amos Bairoch (Université de Genève) est maintenant pris en charge par l'Institut Européen de Bioinformatique (EBI). Le groupe SWISS-PROT s'efforce d'annoter chaque entrée par une analyse de séquence minutieuse. L'annotation inclut la description de la fonction de la protéine, l'indication de sa structure en domaines et les éventuelles modifications post-traductionnelles. S'ajoutent à cela la listes des variants, la description de la réaction catalysée et les similarités de séquences. Les nouvelles séquences obtiennent un numéro d'entrée seulement après une analyse par des biologistes. Dans le cas où il existerait des contradictions entre différentes bases de données pour une même protéine, une séquence consensus est incluse dans la base de données et les variants sont listés sous l'entrée consensus.

A chaque entrée est assigné un numéro « ENZYME », base qui répertorie la nomenclature des fonctions enzymatiques. Lorsque c'est possible des liens vers d'autres banques (EMBL,GenBank, DDBJ) et bases de données (interpro, prosite, Pfam, PDB, Blocks, UNIPROT) sont faits afin d'enrichir l'entrée par des données croisées.

A l'heure actuelle, il y a près de 480 000 commentaires et 739 307 caractéristiques de séquences répertoriées dans SWISS-PROT version 42.6.

- TrEMBL (ou Translated EMBL)

C'est l'antichambre de la SWISS-PROT. Cette banque de séquences est utilisée pour faire face à la masse de séquences protéiques qui découle des grands projets de séquençage, mais qui n'ont pas encore reçu le « label » SWISS-PROT. Pour chaque entrée TrEMBL (Bairoch and Apweiler 1996), est assigné un numéro temporaire SWISS-PROT qui le suit jusqu'à ce que l'entrée soit définitivement acceptée. Pour faciliter le travail de curation, le format adopté par TrEMBL est le même que celui de la SWIS-PROT. Cela dit, les entrées TrEMBL sont

générées de façon automatique, ce qui les rend sujettes à caution étant donné que les annotations générées vont découler directement des informations fournies lors du dépôt de la séquence nucléique. De plus, contrairement à « Entrez Proteins », qui est mis à jour quotidiennement, TrEMBL est mis à jour trimestriellement, ce qui dans certains cas ne permet pas à l'utilisateur de bénéficier des dernières séquences et entraîne une perte d'information.

- PIR (Protein Information Ressource)

Historiquement, PIR (Barker et al. 1987a; b; 1988; Barker et al. 1992; Wu et al. 2004) est le prolongement de la première banque de données crée par Margaret Dayhoff. C'est une base de données qui comme la SWISS-PROT est sujette à un travail de curation, mais moins précis que celui de son homologue Suisse car non systématique. L'intérêt de PIR réside dans son organisation hiérarchique en familles et superfamilles de protéines. Originellement cette classification repose sur la similarité de séquences entre protéines entières. Cette approche permet un regroupement complet et non-chevauchant des séquences contenues dans la base de données tout en reflétant leurs liens évolutifs. Ce système de classification permet une hiérarchisation spécifique et générique des fonctions biochimiques. Cette approche permet une classification des protéines les moins bien caractérisées. Le premier niveau d'organisation, c'est-à-dire celui de « la famille homéomorphe », consiste en un petit nombre de séquences de protéines qui sont à la fois homologues (similarité de séquence détectable) et possédant une taille et une organisation en domaines similaires. Une organisation commune en domaines va se traduire par le même type, le même nombre de domaines organisés de la même façon. La classification tolère des variations pour les domaines répétés et les domaines auxiliaires qui glissent, apparaissent, disparaissent, ou sont remplacés au cours de l'évolution. Si la variation est trop forte au sein d'une famille et ce, malgré un ensemble de critères généraux communs, les séquences sont alors classées en sous-familles. Celles-ci reflètent une spécialisation fonctionnelle. Les niveaux supérieurs de classification correspondent aux superfamilles qui regroupent les familles distantes. PIR a rejoint l'EBI et l'institut de bioinformatique Suisse (SIB) pour contribuer au projet UniProt.

- UniProt

Ces dernières années, les bases de données se sont multipliées chacune avec leurs propres numéro d'accès, leurs spécificités, se référençant souvent les unes les autres mais sans standardisation de l'information ce qui peut transformer une simple recherche en véritable cauchemar pour l'utilisateur. Le but d'UniProt est de palier ce problème, en centralisant toute l'information sous une seule entrée, et en garantissant une information de qualité. UniProt(Apweiler et al. 2004) est organisée en trois niveaux :

- Archives (UniParc) : constituées d'un ensemble de séquences stables et non redondantes organisées sur le mode des bases de séquences publiques. Ce niveau attribue un numéro « universel » d'accès à la séquence.
- Connaissances (UniProt) : il s'agit de la base de données proprement dite, constituée des séquences et de leurs annotations précises et pertinentes. Elle correspond à la réunion des trois bases SWISS-PROT, TrEMBL et PIR. Cette mise en commun a pour but de recouvrir les zones qui n'étaient pas référencées individuellement par ces banques. Elle subit une curation plus sévère que la SWISS-PROT basée majoritairement sur l'intervention manuelle de curateurs et un usage extensif de la littérature.
- Références (UniRef) : ensemble des données non-redondantes, obtenues à partir du niveau « connaissance » d'UniProt, afin d'obtenir une couverture complète des connaissances. UniRef est constitué de trois bases de séquences NREF100, NREF90 et NREF50. NREF100 constitue une collection de séquences regroupées en fonction de leur identité et taxonomie. Les séquences strictement identiques et les sous-fragments d'un même organisme sont présentés sous la même entrée « NREF ». NREF90 et 50 sont construites à partir de NREF100 afin de fournir un échantillon de travail qui rend plus facile la recherche par homologie. Les séquences de tous les organismes qui ont au moins une identité de séquence de 90% (NREF90) et 50%(NREF50), sont regroupées sous la même entrée. Cette compaction des données permet une réduction de 40% de la taille de la base de données dans le cas de NRF90 et de 65% pour NREF50.

2.2.3. Les bases de données structurales

Les informations qu'il est possible d'obtenir à partir d'une structure (architecture du site actif, organisation des éléments de structures secondaires, surfaces exposées, régions d'interactions,

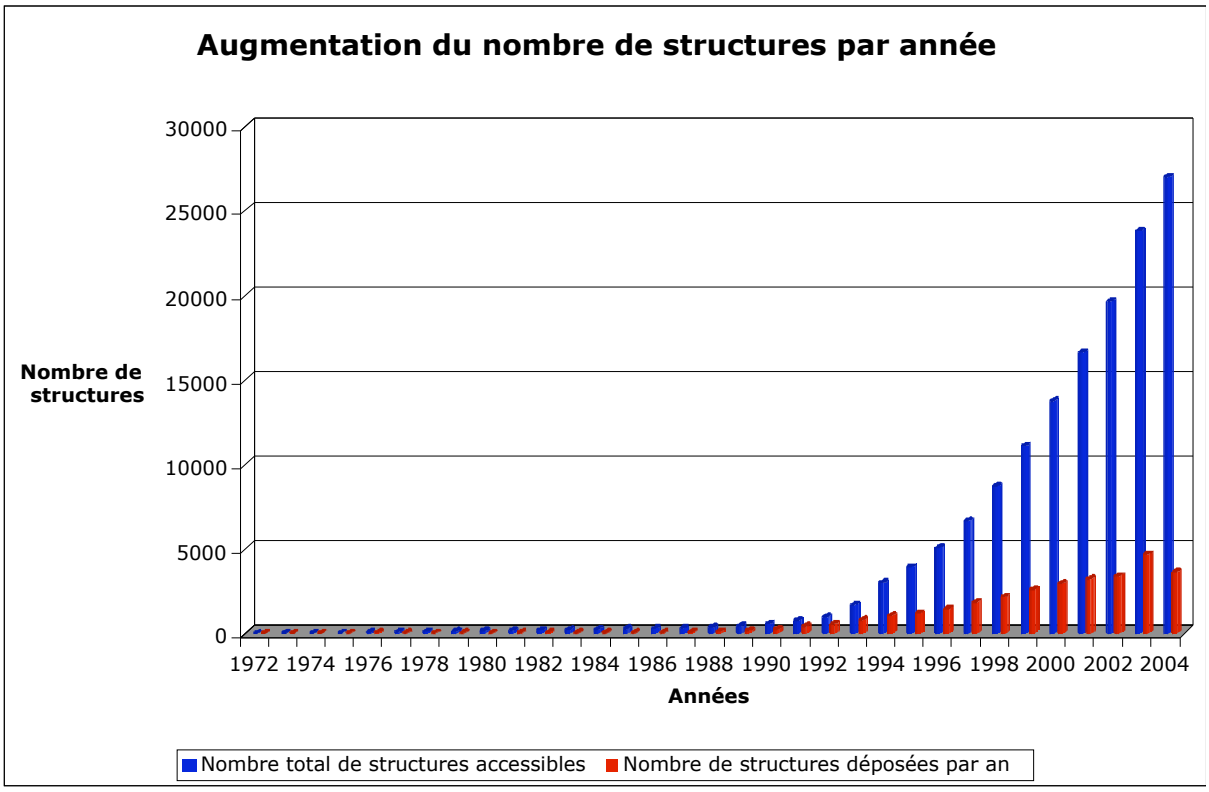


Figure 7: Croissance annuelle du nombre de structures disponibles (en bleu) et du nombre de structures déposées (en rouge).

Données du 21 Septembre 2004.

Technique expérimentale	Type de Molécule				
	Protéines, Peptides, Virus	Complèxes Protéine/Acides nucléiques	Acides Nucléiques	Carbohydrates	Total
Diffraction rayons-X et autres	21523	1062	747	14	23346
RMN	3260	103	608	4	3975
Total	24783	1165	1355	18	27321

Tableau 2: Composition de la PDB en fonction des différentes techniques de résolution structurale.

compréhension du mécanisme moléculaire…) sont bien plus informatives pour le biologiste que la séquence primaire.

- Protein Data Bank (PDB)

La Protein Data Bank (Bernstein et al. 1977 ; Berman et al. 2000) est une banque de données publique où sont déposées les structures tridimensionnelles de protéines et d'acides nucléiques, seuls ou en complexes. Elles contiennent les structures résolues par cristallographie aux rayons X, RMN, microscopie électronique, et les modèles.
Comme c'est le cas pour GenBank, les coordonnées atomiques doivent être déposées avant publication. Les structures sont vérifiées lors de leur dépôt afin de s'assurer que les coordonnées déposées soient conformes aux standards établis. Une des particularités de cette banque est de laisser le choix à la personne qui dépose de ne pas rendre les données accessibles pendant une année. Le taux de structures retenues est autour de 22%. Le contenu et la croissance de la PDB sont résumés par le Tableau 2 et la Figure 7.

- E-MSD (European Macromolecular Structure Database)

C'est la banque Européenne de structures tridimensionnelles de macromolécules biologiques, maintenue par l'EBI. Elle dérive de la PDB mais contrairement à cette dernière c'est une banque relationnelle. Il y a donc pour chaque entrée des liens croisés vers les bases de données (structurales, modulaires ou de séquences) à information ajoutée. Comme dans le cas de la PDB, il est possible d'y déposer de nouvelles structures (20% des nouvelles structures résolues). Un des atouts majeurs de E-MSD (Boutselakis et al. 2003 ; Golovin et al. 2004) par rapport à la PDB est de générer une banque de ligands (chempdb) à partir des structures résolues en complexes. En plus des informations générales et structurales propres au ligand, elle fournit des détails assez précis de l'environnement chimique des sites de liaison de ce dernier.

- NDB Nucleic acid structures Database

Les structures disponibles dans la NDB (Berman et al. 1996) incluent des structures d'ARN et d'oligonucléotides d'ADN composés d'au moins deux bases. Ces molécules peuvent être seules ou complexées avec des protéines ou de petits ligands. Les archives stockent des informations primaires et dérivées des structures. Les données primaires incluent les coordonnées atomiques, les facteurs de structure pour les structures aux rayons X ou les

contraintes pour les structures RMN, et le détail des expériences (la condition de cristallisation, l'empilement cristallin, la collecte de données, et les statistiques d'affinement). L'information dérivée correspond à l'analyse de chaque structure. On retrouve des informations telles que la géométrie de valence, des angles de torsion et des contacts intermoléculaires. La NDB est partiellement redondante avec la PDB lorsqu'il s'agit des complexes avec des protéines.

2.3. Données à forte valeur ajoutée

2.3.1. Classification et analyses structurales

- FSSP (Fold classification based on Structure-Structure alignment of Protein database)

C'est une base de données de classification de repliements structuraux basée sur des superpositions de structures (générant un alignement totalement indépendant de la similarité de séquence). La comparaison des structures de protéines est faite de façon automatique en utilisant le programme DALI (Holm et al. 1992; Holm and Sander 1997). DALI aligne selon le critère du RMSD minimal pour la chaîne principale. Les structures très proches sont regroupées sous une même entrée. Seules les structures présentant des différences significatives au niveau du repliement sont comparées.

- SCOP (Structural Classification Of Proteins)

SCOP (Brenner et al. 1998) est une base de données de classifications structurales de protéines qui est manuellement vérifiée. Elle utilise une classification hiérarchique en trois niveaux, à savoir, les classes structurales, les superfamilles structurales et les familles de protéines proches. Pour SCOP il y a huit classes structurales : tout alpha (α), tout bêta (β), alpha-bêta (α/β), alpha et bêta ($\alpha+\beta$), protéines membranaires, protéines de surface, petites protéines, et les protéines « coiled-coil ». A chaque classe est associé un repliement, qui est ensuite divisé en superfamilles. Lors de la curation de la base de données, les comparaisons de séquences, les motifs et les données fonctionnelles sont pris en compte.
SCOP est souvent considérée comme étant la base de référence pour ce qui est de la classification structurale.

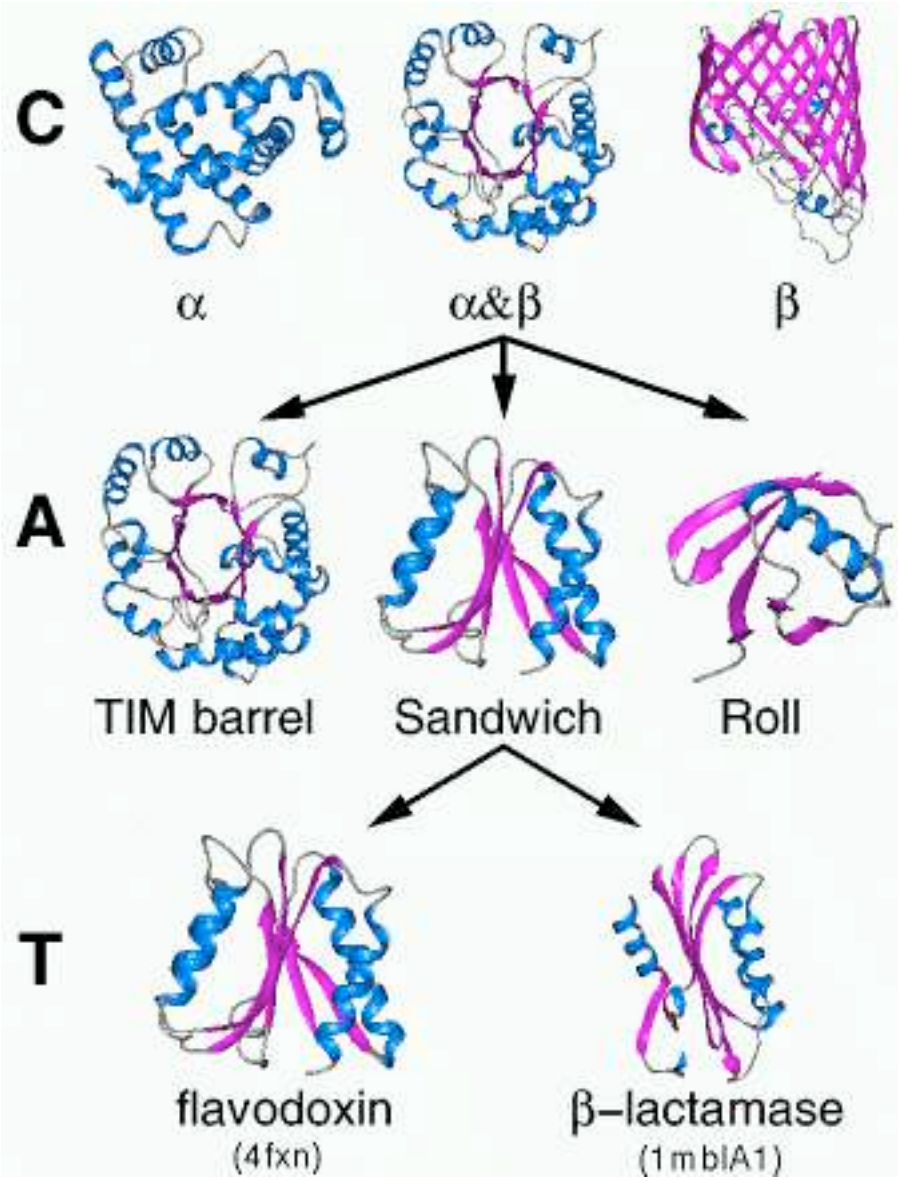


Figure 8 : Classification des structures de protéines dans CATH (Pearl et al. 2001).

- CATH (acronyme de Class, Architecture, Topology et Homologous superfamily)

CATH (Pearl et al. 2001) est une base de données de structures qui organise les domaines de protéines résolues dans la PDB. Elle contient exclusivement des protéines résolues par cristallographie et RMN, dont la limite de résolution est au minimum de 3Å. CATH est organisée en quatre grands niveaux hiérarchiques : Classe, Architecture, Topologie (famille de repliement) et superfamilles Homologues. Chaque niveau est décrit dans le paragraphe suivant. La classification dans CATH est exclusivement basée sur des domaines individuels de protéines. Les protéines-mosaïques ne sont pas traitées en tant que telles. Elles sont sub-divisées en domaines et ces derniers sont classés individuellement. Par ce mode de classification, CATH se pose en alternative à SCOP, permettant une annotation et une classification plus précise pour chacun des domaines.

La hiérarchie dans CATH est organisée autour de quatre niveaux (Figure 8) qui vont du général au particulier :

- La classe (niveau C) : Ce niveau est déterminé en fonction de la composition des éléments de structures secondaires et de leur arrangement. Pour les protéines structurées, trois grandes classes de protéines se distinguent : tout alpha (α), tout bêta (β) et alpha-bêta (α/β) (Levitt and Chothia 1976). Une quatrième classe a été ajoutée pour traiter des domaines dépourvus ou avec peu de structures secondaires.
- L'architecture (niveau A) : ce niveau décrit de façon générale la forme de la structure par l'analyse de l'orientation des éléments de structures secondaires mais sans tenir compte de leurs connectivités.
- La topologie ou famille de repliement (niveau T) : ce niveau est une description plus approfondie qui tient compte à la fois de l'architecture et de la connectivité. Les comparaisons de structures sont menées à l'aide de l'algorithme SSAP (Orengo and Taylor 1996) et les critères de classification ont été déterminés de façon empirique.
- Les superfamilles Homologues (niveau H et S) : Ce niveau regroupe les domaines de protéines qui possèdent un ancêtre commun ou qui sont considérés comme homologues.

Le sous-niveau S correspond à la sous-classification des structures du niveau H sur la base de l'identité de séquence. Cette sous-classification permet de faire des liens structure-fonction.

2.3.2. Bases de données de motifs

- PROSITE

PROSITE est la base de données historique de motifs (Bairoch 1991) et elle est intégrée à SWISS-PROT. Cette base de données contient des motifs de séquences, stockés comme des expressions régulières. Les motifs sont constitués d'une série plus ou moins courte d'acides aminés. Ils décrivent la variabilité des acides aminés pour une position donnée au sein de la série, et leur agencement les uns par rapport aux autres. Cette méthode de description permet une recherche rapide de motifs, mais l'inconvénient, c'est qu'elle ne peut pas refléter l'entière diversité des séquences et entraîne l'exclusion des séquences déviantes. Pour les motifs complexes, la description est associée à un commentaire.

- BLOCKS

Cette base de données développe une approche complètement différente de PROSITE, même si originellement BLOCKS (Henikoff and Henikoff 1994) dérive de cette dernière. L'approche choisie est celle du « bloc de séquences ». Pour la base de données, le « bloc » est un court alignement multiple de séquences, sans insertion ni délétion, d'une région conservée dans une famille de protéines. BLOCKS inclut des blocs provenant de pfam, ProDom, Prints (voir paragraphes suivants). C'est à partir de ces alignements que les matrices de substitution « BLOSUM » ont été dérivées.

- PRINTS

Comme BLOCKS, PRINTS (Attwood and Beck 1994) est une base de données qui collecte des fragments de séquences protéiques. Cependant, pour une protéine, PRINTS liste l'intégralité des blocs conservés. Cette classification à pour effet de réduire la taille de la base de données, et de mettre en relation les différents motifs isolés. Cela permet l'identification de relations évolutives entre des séquences de protéines distantes. Les données de PRINTS sont intégrées et gérées par l'EBI, via UNIPROT et InterPro, une base de données qui regroupe les informations de domaines, de motifs et de séquences.

Les bases de données de motifs collectionnent des portions de séquences dont les résidus sont impliqués dans l'activité des protéines. Collectionner ces motifs revient à collectionner des

signatures fonctionnelles. L'interrogation de ces banques de motifs permet d'assigner rapidement une fonction sur une séquence de protéine.

2.3.3. Les bases de données de domaines de protéines

- Pfam

A la différence des bases de séquences qui sont des banques de séquences entières, pfam (Bateman et al. 2000) est une base de données de domaines. Pfam est une base de données contenant des alignements de séquences de domaines conservés. Ces alignements sont construits à partir de modèles de Markov cachés (HMM voir chapitre suivant). Les alignements sont ensuite manuellement vérifiés et annotés. Les séquences des alignements proviennent de SWISS-PROT (UNIPROT). Une entrée pfam correspond non pas à une séquence mais à un alignement de séquences d'une région conservée de la protéine. Cette région conservée est assignée en « domaine » par pfam. Un alignement classique pfam est constitué de 20 à 30 séquences. Pfam fournit des alignements minima qui contiennent un nombre réduit de séquences suffisamment divergentes pour refléter la diversité des membres de la famille pfam. Les séquences de cet alignement permettent de reconstituer l'alignement intégral. L'entrée pfam, en plus de l'alignement, fournit le modèle de Markov qui a servi pour la constituer. Les modèles se composent d'une matrice de positions spécifiques pondérées (PSSM en anglais) combinée à une probabilité d'apparition, pour un acide aminé donné à une position donnée, suite à une mutation. Cette combinaison rend les modèles très sensibles, ce qui permet de retrouver rapidement à quelle famille appartient la protéine. Pfam inclut pfam-B qui joue le même rôle que TrEMBL pour SWISS-PROT. Les domaines de pfam-B n'ont pas été manuellement vérifiés. Pfam à été incorporée à InterPro, et elle est gérée par l'EBI (Figure 9).

- SMART

SMART (Ponting et al. 1999) est une base de données de domaines qui comme pfam, contient des alignements issus de modèles de Markov. SMART n'est pas aussi généraliste que pfam et concerne uniquement les domaines impliqués dans les mécanismes de signalisation en cascades (Figure 9).

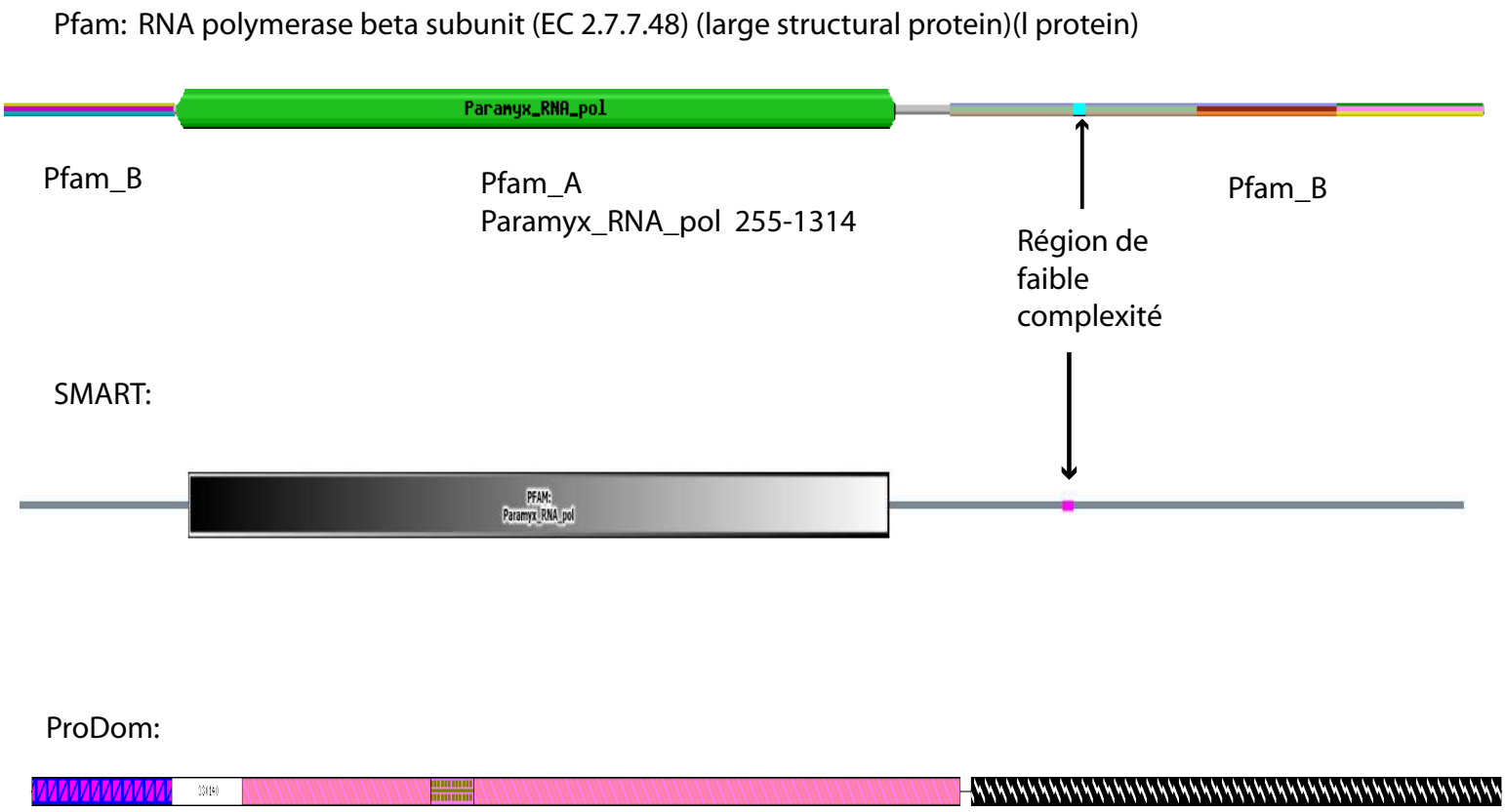


Figure 9 : Mode de représentation graphique pour les bases de donnnées de domaines pour l'entrée P3975 ☐ correspondant à la protéine L du virus de la Rougeole.

- ProDom

ProDom (Corpet 1988; Corpet et al. 1999a; b; Corpet et al. 2000) est une base de domaines. Contrairement à pfam et à SMART qui sont manuellement vérifiées, ProDom est en large partie générée de façon automatique, à partir de résultats de psi-BLAST sur SWISS-PROT et TrEMBL. Il s'agit donc d'une compilation de domaines homologues qui repose sur des critères stringents des résultats de psi-Blast pour l'assignement des domaines. Si ce critère garantit des domaines conservés, il arrive aussi que des séquences homologues mais ne répondant pas au critère de psi-BLAST se retrouvent assignées en deux familles distinctes (Galperin and Grishin 2000). Le résultat du découpage est donné sous forme graphique, ce qui permet d'appréhender très facilement la modularité de la protéine. ProDom permet de visualiser aussi toutes les protéines possédant le domaine sélectionné, même si l'organisation générale de la modularité est totalement différente. ProDom est inclus dans UniProt et les résultats sont aussi consultables via InterPro (Figure 9).

- NCBI CDD (conserved domain database)

CDD (Marchler-Bauer et al. 2002) est une compilation des alignements de pfam et de SMART auxquels s'ajoutent quelques alignements provenant directement de chercheurs du NCBI. Les domaines définis par CDD sont référencés selon la nomenclature utilisée par le NCBI et présentent des liens vers les bases de données dont sont issus les alignements.

- CAZyModO

CaZyModo (Coutinho and B. 1999a; Coutinho and B. 1999b) est une base de données spécialisée dans les protéines d'intérêts pour la glycobiologie. Les « glycosyltransférases » sont des protéines souvent composées de plusieurs domaines fonctionnels. La classification que propose CAZyModO est originale et diffère sensiblement des autres bases de domaines. Cette base de données tend à définir des modules et non des domaines. Les modules sont définis comme des unités de repliement autonomes et toujours fonctionnels, mais le module à la différence du domaine peut être constitué de plusieurs domaines. Dans cette approche, les modules sont regroupés par similarité de séquences en utilisant BLAST avec un critère d'une E-value $< 10^{-3}$, afin de constituer des familles modulaires. Les modules sont analysés à l'aide de la méthode « d'analyse des clusters hydrophobes » (HCA). C'est sur ce regroupement en modules que repose la classification des « glycosyltransférases ». CAZyModO intègre toutes les séquences provenant de GenBank, SWISS-PROT et PDB.

2.3.4. Les bases de données fonctionnelles et taxonomiques

- MEROPS

La base de données MEROPS (Barrett 2004) tend à intégrer l'ensemble des informations (séquences, site actif, publications, structure…) se rapportant aux protéases et à leurs inhibiteurs. L'organisation de cette base de données repose sur une classification hiérarchique, pour laquelle les protéines homologues sont regroupées en familles et les familles en clans. La base de données classe également des inhibiteurs de protéases. Le mode de classification est identique. Des liens relationnels sont établis entre les deux classifications ainsi que vers d'autres bases de données, telles que PDB et PUBMED. La répartition phylogénique et les alignements pour chaque famille et clan sont générés, ainsi que les informations sur les sites de clivage. Pour une protéase donnée, un seul numéro d'accès permet de retrouver l'intégralité de ces informations.

- NCBI Taxonomie

Le NCBI maintient sa propre base de données taxonomiques (Wheeler et al. 2001). Son but est de maintenir une certaine cohérence entre les données obtenues à partir des séquences et leur classification phylogénique. Ainsi elle contient l'ensemble des noms de tous les organismes présents dans GenBank. La classification fournit un consensus basé sur les données bibliographiques, les bases de données spécialisées, les experts en taxonomie et les informations recueillies lors des dépôts de séquences. La base de données est donc une base universelle, hiérarchisée en six niveaux : Archebactéries, Eubactéries, Eucaryotes, Viroïdes, Virus, et les non-classés.

- ICTV

L'ICTV (Maniloff 1995; Pringle 1995) est la base de référence taxonomique pour les virus. Elle fournit des descriptions complètes de virus. La classification est organisée en cinq catégories : Ordre, Famille, Genre, Espèce, Isolat. Elle fournit des images de microscopie électronique des virus, ainsi que la reconstitution de leur structure et des liens vers des bases de données de séquences génomiques et protéiques. C'est aussi la première base de données taxonomiques à donner une information sur la biologie des virus

2.3.5. Bases de données Virales

- VIDA

VIDA (Alba et al. 2001) est une base de données spécialisée qui organise et annote les protéines de trois familles de virus : *Herpesviridae, Coronaviridae* et *Arteriviridae*. Les protéines sont organisées en familles sur la base de la similarité de séquences. Les régions conservées sont identifiées et les alignements correspondants peuvent être consultés. VIDA fournit une information structurale lorsqu'elle existe, mais pas de données fonctionnelles.

- NCBI Viral Genomes Project (VGP)

Le « projet de génomes viraux » (Bao et al. 2004) du NCBI a pour but de combler la carence en annotations dont souffrent les entrées relatives aux génomes viraux. C'est un projet collaboratif qui implique le NCBI et des chercheurs externes spécialisés. Le VGP rassemble seulement des séquences de génomes complets issues de GenBank. Ces séquences sont ensuite manuellement vérifiées, corrigées et mises à jour, contrairement aux séquences déposées par les utilisateurs (voir plus haut). De ce fait la base de données produite par VGP est très peu redondante et comprend des séquences de très bonne qualité. Le VGP prend aussi en charge l'annotation taxonomique. Cette dernière reprend en majeure partie les annotations de ICTV, néanmoins certains virus dont les séquences n'ont pas été publiées sont classés (par exemple le bactériophage MX8). Le VGP fournit une analyse comparative des génomes viraux et classe les protéines en fonction de leur profil fonctionnel ou phylogénique. La classification des séquences se fait par regroupement en populations de protéines homologues à l'aide de BLASTP. A terme, l'annotation des séquences sera complétée par la gestion des informations découlant des polyprotéines, notamment l'incorporation des sites de clivage et des sites de décalage de cadre de lecture.

Chapitre 2

Principes et méthodes de l'analyse de séquences

1. Introduction

Nous avons vu dans le premier chapitre que les protéines sont une succession d'acides aminés formant des polymères complexes qui, sous contraintes, vont se réorganiser en structures tridimensionnelles. Nous avons aussi vu comment ces informations se rapportant à ces différents états sont stockées et classées. La bioinformatique ne s'emploie pas seulement à la gestion de ces données mais aussi à leur traitement. Les deux contributions majeures de la bioinformatique concernent la « biologie fonctionnelle et structurale », et les mécanismes de l'évolution. Nous allons ici rappeler les principales méthodes d'analyses qui permettent d'obtenir un maximum d'informations à partir des séquences et/ou des structures.
Les protéines sont schématiquement classées en quatre niveaux d'organisation. Pour chaque niveau la bioinformatique génère de l'information.

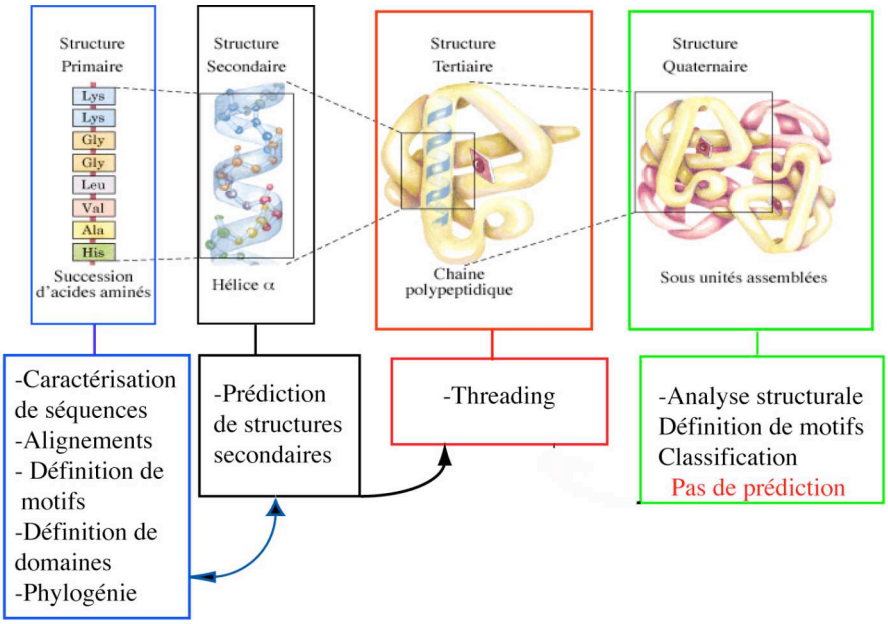


Figure 10 : Les quatre niveaux hiérarchiques traités par la bioinformatique et le type d'information disponible à chaque niveau.

La structure quaternaire est le niveau ultime d'organisation. C'est certainement le niveau le plus informatif car il permet de comprendre les mécanismes moléculaires d'assemblage, mais malheureusement il n'est pas encore accessible à la bioinformatique prédictive. La structure tertiaire est dans certains cas prédictible. Il a été estimé que la modélisation par homologie de séquence est une technique envisageable dans 30% des cas où la séquence d'intérêt possède un homologue structural (Rost et Schneider 1998). La fiabilité de la technique est fonction de la similarité de séquence entre les deux protéines. Les derniers résultats de CASP5 ont montré

qu'à partir de 25% d'identité, la méthode de modélisation par homologie de séquence donne des résultats validés par des approches expérimentales (Tramontano and Morea 2003).
Il est possible de prédire les structures secondaires et ces prédictions sont devenues performantes (environ 75% de fiabilité). La structure primaire d'une protéine est le niveau qui contient le plus de potentialité de données. A partir d'une simple séquence il est possible d'identifier les caractéristiques physico-chimiques d'une protéine (hydrophobicité, pI, accessibilité au solvant, ...) et d'en déduire les structures secondaires. A partir d'un alignement multiple, il est possible, de cartographier la protéine, définir si elle s'organise en domaines, de déterminer des motifs conservés, de retracer son évolution et ses liens phylogénétiques et éventuellement d'identifier sa fonction.

2. Méthode d'analyse de la structure primaire

2.1. Généralités sur l'alignement de séquences

La recherche de similarité de séquence à pour but d'identifier la fonction portée par celle-ci en la comparant à d'autres séquences. Si la similarité est significative pour ne pas être le fruit du hasard, alors deux postulats peuvent être posés. Premièrement, la similarité de séquence peut être considéré comme le reflet de l'homologie entre ces deux séquences, autrement dit elle traduit leur lien phylogénétique. Deuxièmement, l'homologie entre deux séquences peut laisser supposer que les protéines ont la même structure et la même fonction (Rost 1999).
La recherche de similarité au sein d'un échantillon de séquences passe par l'alignement de ces séquences, et par la vérification que l'alignement obtenu est statistiquement significatif. Quand on cherche à mettre en évidence que ces séquences possèdent un ancêtre commun et qu'elles sont donc homologues, on considère non seulement les mutations ponctuelles (substitutions), mais aussi la possibilité d'insertions et/ou de délétions (indels). Ainsi, les critères de la recherche par similarité sont :

- le type d'alignement
- le système de « scores » pour pondérer les opérations d'édition
- l'algorithme pour trouver l'alignement optimal
- les méthodes statistiques d'évaluation de la qualité de l'alignement

De façon générale, un bon alignement de séquences de protéines homologues doit révéler des caractéristiques communes conservées qui sont importantes pour la fonction et/ou la structure de chacune de ces protéines. Mais, il doit aussi mettre en relief des régions peu conservées qui

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	X
A	2																				
R	-2	6																			
N	0	0	2																		
D	0	-1	2	4																	
C	-2	-4	-4	-5	12																
Q	0	1	1	2	-5	4															
E	0	-1	1	3	-5	2	4														
G	1	-3	0	1	-3	-1	0	5													
H	-1	2	2	1	-3	3	1	-2	6												
I	-1	-2	-2	-2	-2	-2	-2	-3	-2	5											
L	-2	-3	-3	-4	-6	-2	-3	-4	-2	2	6										
K	-1	3	1	0	-5	1	0	-2	0	-2	-3	5									
M	-1	0	-2	-3	-5	-1	-2	-3	-2	2	4	0	6								
F	-4	-4	-4	-6	-4	-5	-5	-5	-2	1	2	-5	0	9							
P	1	0	-1	-1	-3	0	-1	-1	0	-2	-3	-1	-2	-5	6						
S	1	0	1	0	0	-1	0	1	-1	-1	-3	0	-2	-3	1	2					
T	1	-1	0	0	-2	-1	0	0	-1	0	-2	0	-1	-3	0	1	3				
W	-6	2	-4	-7	-8	-5	-7	-7	-3	-5	-2	-3	-4	0	-6	-2	-5	17			
Y	-3	-4	-2	-4	0	-4	-4	-5	0	-1	-1	-4	-2	7	-5	-3	-3	0	10		
V	0	-2	-2	-2	-2	-2	-2	-1	-2	4	2	-2	2	-1	-1	-1	0	-6	-2	4	
X	-15	-15	-15	-15	-15	-15	-15	-15	-15	-15	-15	-15	-15	-15	-15	-15	-15	-15	-15	-15	0

A) Matrice PAM 250

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	X
A	4																				
R	-1	5																			
N	-2	0	6																		
D	-2	-2	1	6																	
C	0	-3	-3	-3	9																
Q	-1	1	0	0	-3	5															
E	-1	0	0	2	-4	2	5														
G	0	-2	0	-1	-3	-2	-2	6													
H	-2	0	1	-1	-3	0	0	-2	8												
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4											
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4										
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5									
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5								
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6							
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7						
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4					
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5				
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11			
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7		
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4	
X	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	1

B) Matrice BLOSUM 62

Figure 11: A) Matrice de substitution PAM 250. Les nombres indiquent les scores attribués pour la substitution d'un acide aminé par un autre. Plus le nombre est élevé, moins la pénalité est importante pour la substitution correspondante. On notera que les pénalités sont élevées pour remplacer une cystéine ou des acides aminés aromatiques contre n'importe quel autre acide aminé. A l'inverse leur conservation est distinguée par de hauts scores. B) Matrice de substitution BLOSUM 62. X correspond à un indel.

sont moins importantes pour la fonction commune, mais porteuses d'une spécificité pour chacune d'entre elles. Pour remplir ces deux critères, il faut impérativement travailler avec un échantillon de séquences de bonne qualité (choix de la banque la plus appropriée). En pratique une fois que le choix de l'échantillon est fait, l'opérateur doit être capable de quantifier la similitude de son alignement et d'évaluer sa qualité. Pour quantifier la similitude, des « scores » sont calculés à partir de matrices de scores (Figure 11). Ces matrices permettent de pondérer les modifications observées au sein des séquences. Il existe différentes matrices qui reposent sur des critères tels que : l'évolution (Blosum (Henikoff and Henikoff 1992)-PAM (Dayhoff MO et al. 1978)), les caractéristiques physico-chimiques (hydrophobicité (Levitt and Chothia 1976), ou le repliement (Overington et al. 1992; Johnson and Overington 1993)). Le choix de la matrice (Figure 12) utilisée lors de l'alignement est important car selon celle choisie, le résultat produira un alignement guidé par un critère « évolutif » ou « structural ».

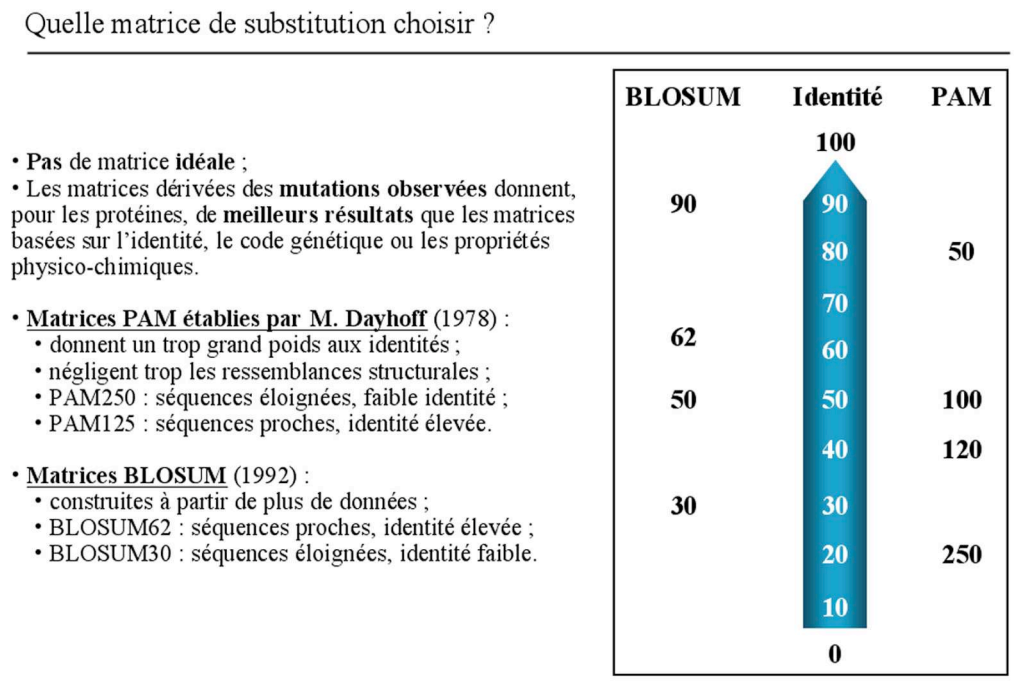


Figure 12 : Choix de la matrice de substitutions.

Il faut distinguer deux types de « score ». Le « score » élémentaire « Se » qui est la valeur donnée directement dans la matrice, et le « score » global « S » qui est calculé à partir des « scores » élémentaires, après soustraction des pénalités pour les indels « Sp ».

$$S=\sum Se-\sum Sp$$

L'évaluation du « score » se fait par rapport à une valeur seuil de probabilité relative à l'alignement considéré.

2.2. Alignement de deux séquences et algorithme de recherche dans les bases de données

Deux méthodes ont été développées pour aligner un groupe de séquences entre elles. La méthode globale, qui aligne les séquences sur leur longueur totale et la méthode locale qui tend à aligner spécifiquement des régions conservées sans aligner les régions variables.

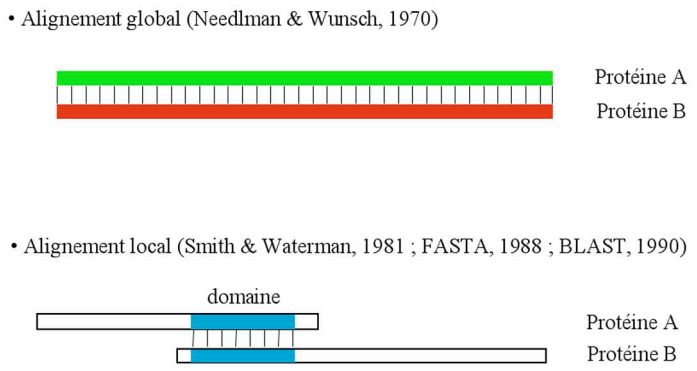


Figure 13 : Alignement global et local.

Les deux méthodes sont fondées sur des approches différentes, qui doivent être utilisées à bon escient en fonction de l'échantillon de séquences que l'on souhaite étudier. La méthode globale donnera de très bons (et fiables) alignements pour un échantillon de séquences de longueur homogène, alors que la méthode locale sera plus efficace pour aligner des régions conservées au sein d'un groupe de séquences de longueur variable. Les deux approches mathématiques qui sont à la base de ces méthodes sont la programmation dynamique (Needleman and Wunsch 1970) et les modèles de Markov cachés (HMM) (pour revue voir (Eddy 1996)). La programmation dynamique est une approche déterministe (pour laquelle le meilleur alignement est celui qui a le « score » le plus haut) alors que le HMM est stochastique, c'est à dire descriptible en termes de probabilité. Dans le cas de HMM, un alignement multiple se construit en trois étapes que l'on considère comme des paliers (aligné, inséré, et enlevé). A chaque palier est associé une probabilité de transition, qui est dépendante de l'état précédent. L'alignement issu d'un HMM sera de grande qualité et incluera des « scores » de position pour chaque acide aminé. La contrepartie de ce type d'alignement est un temps de calcul et de paramétrage assez long. Le temps nécessaire à la comparaison de deux séquences de longueur X est proportionnel à X^2. L'évaluation d'une éventuelle insertion pour chaque position de chaque séquence augmente d'un facteur 2 le temps de calcul. Au

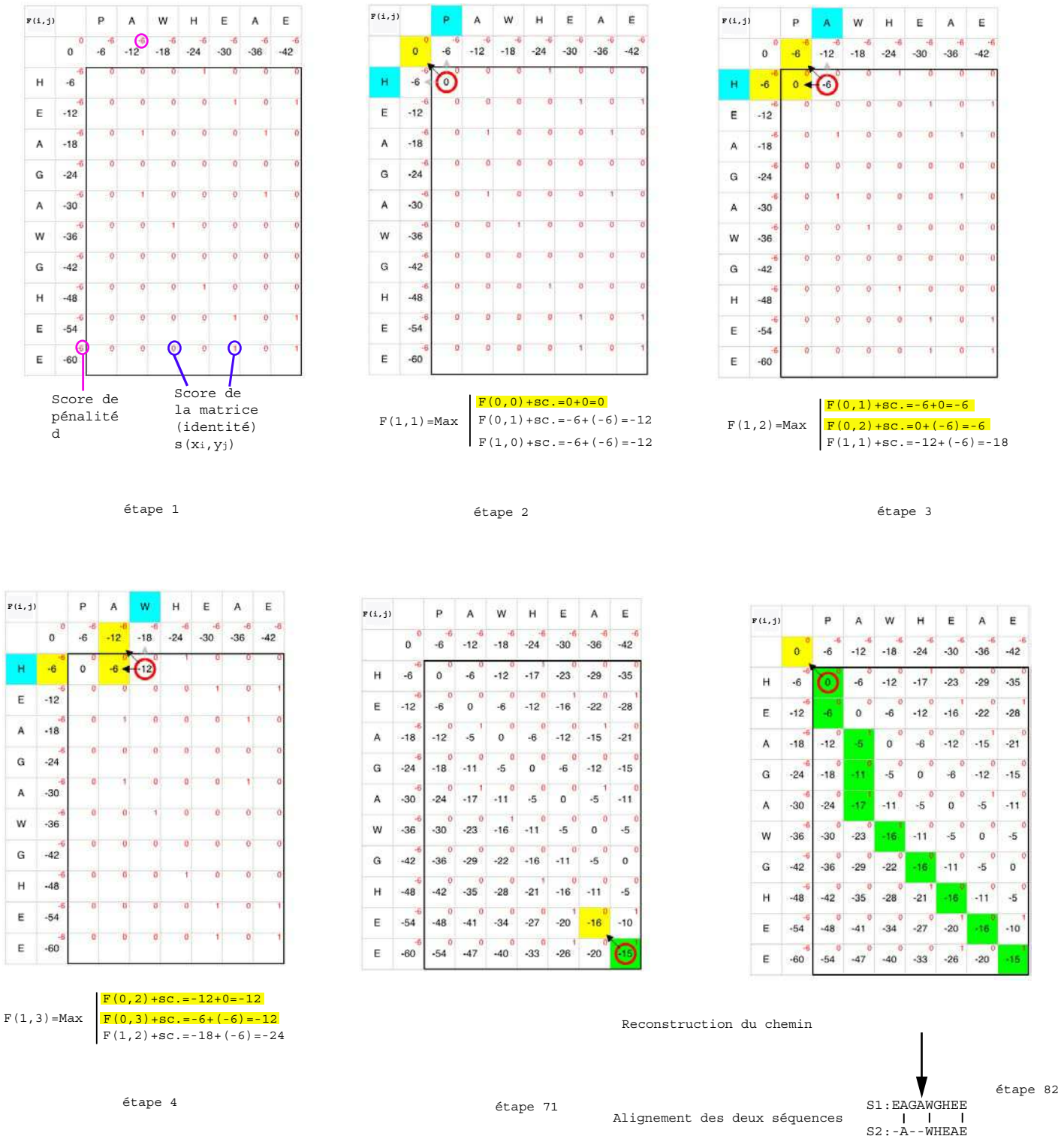


Figure 14 : Principe de la construction d'un alignement par la méthode globale. La première étape présente la matrice identité et les pénalités ont été fixées uniformément à –6. (chiffres rouges).Les étapes suivantes (2 à 71) montrent la construction de la matrice selon les différentes possibilités d'alignement.Les équations de construction sont reportées sous la matrice. Les choix possibles sont surlignés en jaune. Les étapes 71 et 82 montrent la construction de l'alignement en fonction des meilleurs scores.

final, les méthodes de programmation dynamique sont les plus efficaces et pratiques, car elles permettent de limiter le temps de calcul au seul facteur X^2.

2.2.1. Méthode globale

Comme nous l'avons précédemment présenté, l'algorithme de Needleman-Wunsch fournit une méthode optimale pour aligner deux séquences de longueur comparable (Needleman and Wunsch 1970). L'algorithme reflète la nature récurrente de la définition de la ressemblance.

L'algorithme calcule les « scores » maximaux pour un alignement en fonction d'une évaluation des substitutions pour un acide aminé donné et pondéré par des pénalités associées aux insertions et délétions. Pour cela construisons une matrice F indexée par i et j, correspondant à chaque séquence et où F(i,j) sera le « score » du meilleur alignement entre les segments de séquences $x_{1...i}$ et $y_{1...j}$. F(i,j) peut être construit de façon récursive. Par convention, F(0,0)=0. Le remplissage de la matrice va se faire alors du coin en haut à gauche vers le coin en bas à droite. Si F(i-1,j-1), F(i-1,j) et F(i,j-1) sont connus alors il est possible de calculer le "score" F(i,j) (Figure 16). Il y a trois façons d'aligner les séquences afin d'obtenir le meilleur « score ». Soit x_i s'aligne sur y_j, et dans ce cas $F(i,j)=F(i-1,j-1)+Se(x_i,y_j)$, soit x_i est aligné contre un indel et dans ce cas F(i,j)=F(i-1,j)-Sp, soit c'est y_j qui s'aligne contre un indel et alors F(i,j)=F(i,j-1)-Sp. Le « score » maximal d'un alignement entre x_i et y_j peut donc s'exprimer comme :

$$F(i,j)=\max \begin{cases} F(i,j)=F(i-1,j-1)+Se(x_i,y_j), \\ F(i,j)=F(i-1,j)-Sp, \\ F(i,j)=F(i,j-1)-Sp. \end{cases}$$

Au fur et à mesure que l'on construit la matrice, on indique l'origine de la cellule qui est à l'origine du « score » de la cellule suivante (Figure 14). La valeur de la cellule finale de la matrice correspond nécessairement au meilleur « score » de l'alignement de $x_{1...i}$ à $y_{1...j}$. Pour trouver l'alignement lui-même il faut trouver le chemin des choix qui ont été à l'origine de cette valeur finale. Pour cela, il suffit de suivre les indications laissées dans chaque cellule au moment de la construction de la matrice.

2.2.2. Méthode locale

La méthode de l'alignement local est utilisée pour aligner de façon optimale, des séquences divergentes qui possèdent quelques régions communes. Le reste de la séquence contient des mutations et des indels empêchant l'alignement. C'est en utilisant cette potentialité d'aligner ces régions conservées que les programmes de recherche de séquences homologues dans les bases de données ont été développés. La méthode locale repose sur l'algorithme de Smith-Waterman (Smith and Waterman 1981) qui reprend l'algorithme de Needleman-Wunsch, incluant deux changements.

$$F(i,j)=\max\begin{cases} 0\ , \\ F(i,j)=F(i-1,j-1)+Se(x_i,y_j), \\ F(i,j)=F(i-1,j)-Sp, \\ F(i,j)=F(i,j-1)-Sp. \end{cases}$$

Premièrement, si le « score » est négatif on le force à zéro. Cette option entraîne l'arrêt de l'alignement et le commencement d'un nouveau, considérant qu'il est inutile de persévérer dans un mauvais choix. La seconde différence tient au fait que l'alignement puisse s'arrêter n'importe où. De fait, le meilleur « score » reflétant le meilleur alignement peut se trouver dans n'importe quelle cellule de la matrice. Ainsi, l'alignement local maximal est borné à gauche et à droite par la première cellule contenant un zéro (Smith and Waterman 1981).
Comme nous l'avons vu précédemment, le temps de calcul dans la programmation dynamique est de l'ordre X^2 (X étant la longueur de la séquence). Dans le cas de recherche contre les banques de séquences, la programmation dynamique conduit à des temps de calcul parfois importants. Pour cette raison, des heuristiques ont été développées. Les heuristiques fonctionnent par des réévaluations successives de plus en plus rapprochées. Ces programmes sont beaucoup plus rapides, mais au prix de la garantie d'optimalité de l'alignement. Le but des heuristiques est la recherche de la fraction la plus petite possible des cellules de F, en évitant de perdre tous les alignements de plus grand « score ».

FASTA (Pearson and Lipman 1988) et surtout BLAST (Altschul et al. 1990) sont les familles de programmes les plus populaires dans cette catégorie. FASTA fut le premier programme de recherche rapide dans les bases de données donnant des résultats comparables à ceux fournis par la programmation dynamique. FASTA considère uniquement les séquences présentant une région de forte similitude avec la séquence recherchée. Pour cela, il recherche des petites

portions de séquences identiques appelées « mots ». L'idée est que des paires de séquences homologues doivent avoir au moins un « mot » en commun. Les séquences de la base de données ne répondant pas à ce critère ne sont pas traitées. Dans un deuxième temps FASTA applique localement à chacune de ces meilleures zones de ressemblance un algorithme d'alignement optimal.

BLAST (de Basic Local Alignement Search Tool) est une méthode heuristique destinée à trouver les alignements optimaux locaux, de meilleurs « scores » entre la séquence requête et les séquences déposées de la banque (Altschul et al. 1990). Le principe général est le même que pour FASTA, mais le traitement statistique est différent. BLAST recherche des « mots » de taille définie (M) constituant des points d'ancrage. Un « score » (L) est attribué à chaque mot. C'est à partir de ces points d'ancrage que l'alignement s'étend. Une valeur statistique est attribuée au « score » par comparaison avec la séquence d'origine.

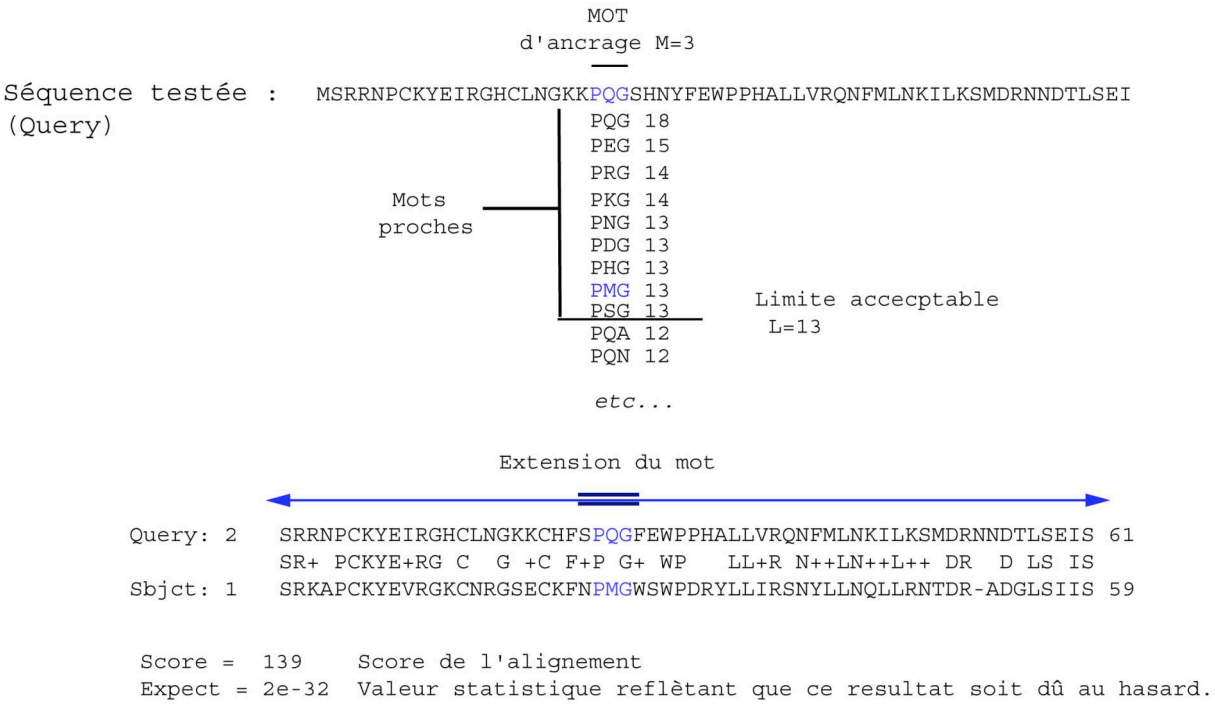


Figure 15 : Principe du BLAST.

L'algorithme initial de BLAST ne permet ni les insertions ni les délétions (Karlin and Altschul 1993). Bien que performant et rapide, la sensibilité de BLAST est moindre que celle de l'algorithme Smith-Watermann ou de FASTA. Les nouvelles versions de BLAST, WU-BLAST et Psi-BLAST (Altschul et al. 1997), dont l'algorithme initial a été modifié,

présentent les mêmes performances que FASTA. De plus, des filtres, tels que SEG (Wootton JC. and S. 1993) ou XNU (Claverie J-M. and D. 1993) ont été implémentés pour éliminer les régions répétitives qui conduisent à des résultats statistiquement significatifs, mais sans intérêt biologique.

2.3. Alignements multiples

Les algorithmes de programmation dynamique donnent un alignement optimal (global ou local selon le cas) pour n'importe quelle comparaison impliquant deux séquences. En revanche pour les alignements multiples comprenant plus de trois séquences, il n'existe pas de méthode optimale. Pour éviter des temps de calcul qui seraient de l'ordre de X^K (X étant la longueur de la séquence et K le nombre de séquences à aligner), les méthodes d'alignement multiple sont des méthodes heuristiques. Globalement, elles regroupent les séquences d'une façon plus ou moins hiérarchique, afin de guider la construction de l'alignement (Feng et Doolittle 1997). Les séquences sont d'abord comparées en utilisant une méthode rapide (globale ou locale) et regroupées en fonction des « scores » de similarité. Ces « scores » permettent la création d'un arbre qui guide la construction de l'alignement. Les séquences sont ensuite alignées séquentiellement en commençant par les séquences regroupées aux extrémités de l'arbre et redescendant vers la racine. Une fois que deux séquences sont alignées, leur alignement est alors fixé et considéré par l'algorithme comme une séquence simple. Ce mode de calcul permet de limiter les temps de calcul au seul facteur X^2. Pour éviter de compromettre la qualité de l'alignement, les séquences de plus forte similarité sont alignées en premier, dans l'espoir de pondérer favorablement les résidus conservés dès les premières étapes de la construction de l'alignement. Il faut aussi noter que le récent développement d'outils de correction des alignements multiples (Thompson et al. 2003) permet une optimisation sensible des alignements produits.

Les programmes d'alignement multiple sont aujourd'hui nombreux ; les plus populaires et fiables (Thompson et al. 1999; Edgar 2004b) sont indiqués dans le tableau suivant :

Programme	Commentaires	Références
Multalign	Utilise la méthode globale et la méthode UPGMA	(Corpet 1988)
Clustal_W	Utilise la méthode globale et la méthode NJ	(Thompson et al. 1994)
Clustal_X	Le programme X bénéficie d'une interface graphique	(Thompson et al. 1997)
Dialign	Utilise la méthode locale et une méthode itérative pour optimiser l'alignement	(Morgenstern et al. 1998)
Dialign2	Combine les méthodes locale et globale ainsi qu'une méthode itérative pour optimiser l'alignement	(Morgenstern 1999) (Morgenstern 2004)
Db-clustal	Modification de Clustal_W, incorporant des données issues d'alignements locaux permettant la préservation des motifs lors de la construction de l'alignement multiple malgré les indels	(Thompson et al. 2000)
T-Coffee	Utilise une bibliothèque d'alignements (locale et globale) pondérés et optimisés par un algorithme génétique	(Notredame et al. 2000)
MAFFT	Utilise une transformée de Fourrier et la méthode UPGMA pour guider l'alignement Des transformées de Fourrier affinent les résultats	(Katoh et al. 2002)
MUSCLE	Utilise la méthode d'évaluation des distances *k*mer et de Kimura. La méthode UPGMA guide l'alignement. Le programme réalise l'alignement progressivement en trois étapes : -Premier alignement -Amélioration de l'alignement -Affinement de l'alignement	(Edgar 2004a)

Tableau 3 : Programmes d'alignements multiples.

Aligner « à l'œil » (Troy et al. 2001) reste toujours possible, mais selon le nombre de séquences considérées cela peut être un travail long et fastidieux.

2.4. Domaines et motifs

Les termes de domaines et motifs sont souvent employés dans des contextes différents de la biologie, ce qui a pour effet d'entraîner une certaine confusion au niveau de leur définition.

2.4.1. Les domaines

La définition du « domaine » est légèrement variable selon le degré d'information qui est pris en compte. D'un point de vue structural, un « domaine » est un fragment de séquence de protéine contiguë, conservé dans une ou plusieurs familles de protéines et qui se replie indépendamment. Il peut être dupliqué et réutilisé par des protéines de fonctions différentes (cas des gènes « mosaïques »). D'autre part par analyse bioinformatique, il est possible de mettre en évidence des régions conservées au sein de séquences de protéines. Sans informations structurales disponibles, il est impossible de prédire si ces régions peuvent se structurer de façon autonome. Malgré cette incertitude, ces régions sont aussi (souvent à tort) définies comme des « domaines ».

2.4.2. Les motifs

La définition du « motif » est sujette à confusion. Au sens strict, le motif est un segment de séquence identifiable, court, continu et non ambigu. Cela dit, la définition de « motif » en bioinformatique est plus étendue. Sera appelé motif un segment de séquence dégénéré mais conservant malgré tout des éléments reconnaissables. Par extension, le terme motif peut désigner l'ensemble de plusieurs motifs distribués le long de la séquence. Par l'analyse d'un alignement multiple, il est possible de déterminer ces motifs, mais l'assignement de leur rôle ou fonction nécessite une analyse structurale. La recherche de motifs est fondamentale pour au moins deux raisons. Premièrement elle participe à caractériser de nouvelles séquences (identification de fonction ou sites de reconnaissances…) de nouveaux organismes. Deuxièmement elle permet la classification des protéines et leur assignement en famille sur le critère d'une « signature » commune. La dérive d'un motif se fait visuellement, en revanche, des programmes tels que ScanProsite (Bairoch 1991) permettent de retrouver des motifs.

2.5. Méthode de construction d'arbres phylogénétiques

L'un des aspects de l'analyse de séquences est la mise en évidence de leur évolution. Jusqu'aux années 1960, les comparaisons morphologiques, comportementales et l'analyse de la répartition géographique des espèces étaient les seuls moyens disponibles pour construire des classifications d'espèces. La découverte que des protéines homologues ont des séquences en acides aminés variables d'une espèce à l'autre, et que cette variabilité est fonction de l'évolution, a fourni un nouveau moyen d'étude à la phylogénie. Ce paragraphe ne prétend pas présenter tous les aspects de la phylogénie, mais il se limite à rappeler les principales méthodes permettant la reconstruction d'arbres phylogénétiques :

- Méthodes basées sur les mesures de distances dans un couple de séquences

Ces méthodes déterminent le nombre de substitutions de nucléotides ou d'acides aminés entre deux séquences protéiques. Elles permettent la reconstruction d'arbres phylogénétiques sans racine. La construction de l'arbre repose sur la recherche des séquences les plus proches et ceci à chaque étape de regroupement. Ces méthodes sont rapides et donnent de bons résultats pour des séquences ayant une forte similarité.

2.5.1. UPGMA (Unweight Pair Group Method with Arithmetic mean)

Cette méthode est utilisée pour reconstruire des arbres phylogénétiques si les séquences ne sont pas trop divergentes. UPGMA (Sokal and Sneath 1963) utilise un algorithme de regroupement des séquences. Le regroupement des séquences se fait par un « score » attribué en fonction de la similarité entre les séquences. La reconstruction de l'arbre se fait pas à pas grâce à ce « score ». Schématiquement, la première étape passe par l'identification des deux séquences les plus proches. Ce groupe est ensuite traité comme un tout, puis on recherche la séquence la plus proche et ainsi de suite jusqu'à ce qu'il n'y ait plus que deux groupes. L'inconvénient majeur de la méthode est sa sensibilité. Si les taux de mutations sont différents sur les différentes branches, cela peut mener à des erreurs lors de la reconstruction de l'arbre.

2.5.2. NJ (Neighbor-Joining)

Cette méthode développée par Saitou et Nei (Saitou and Nei 1987) tente de corriger le biais introduit par la méthode UPGMA en tenant compte d'un taux de mutation différent sur les branches. Les données initiales permettent de construire une matrice qui donne un arbre en

étoile. Cette matrice de distances est ensuite corrigée afin de prendre en compte la divergence moyenne de chacune des séquences avec les autres.
L'arbre est alors reconstruit en reliant les séquences les plus proches dans cette nouvelle matrice. Lorsque deux séquences sont liées, le nœud représentant leur ancêtre commun est ajouté à l'arbre tandis que les deux feuilles sont enlevées. Ce processus convertit l'ancêtre commun en un nœud terminal dans un arbre de taille réduite.

-Méthodes basées sur les caractères

Ces méthodes s'intéressent au nombre de mutations (substitutions ou indels) qui affectent chaque position de la séquence.

2.5.3. Parcimonie

La parcimonie consiste à minimiser le nombre de « pas » (mutations ou substitutions) nécessaires pour passer d'une séquence à une autre dans une topologie de l'arbre.
Ainsi la méthode considère que les changements au sein de la séquence se font indépendamment les uns des autres, autrement dit la séquence peut être considérée comme une suite de caractères aléatoires. Le second postulat considère que la vitesse d'évolution est lente et constante au cours du temps. Cette méthode, quand elle est appliquée à des séquences protéiques, utilise le code génétique pour comptabiliser le nombre de substitutions nécessaires (changements de bases) pour passer d'une séquence à l'autre. La méthode est particulièrement adaptée aux séquences relativement éloignées.

2.5.4. Maximum de vraissemblance (Maximum Likelihood)

Cette méthode de reconstruction phylogénétique évalue, en termes de probabilités, l'ordre des branchements et la longueur des branches d'un arbre sous un modèle évolutif donné. La méthode considère que la probabilité de chaque changement est indépendante des changements précédents (Modèle de Markov). Les probabilités de substitution ne changent pas au cours du temps (le long de l'arbre) et les changements sont réversibles. Ce modèle peut être pondéré afin d'accroître son réalisme en utilisant des taux de substitutions différents pour chaque remplacement (matrice de substitution), une correction pour le nombre de sites susceptibles de muter ou des taux de substitutions variables pour ces sites. La méthode est particulièrement adaptée aux séquences éloignées.

Serveur et Programme	Commentaires	Références
Jpred	Centralise la requête, et interroge en parallèle : PHD, Predator, DSC, NNSSP, Zpred, Mulpred, Jnet, COILS, Multicoil Le résultat présenté est le consensus de ces prédictions	(Cuff et al. 1998)
Predict Protein	Permet d'interroger PHD, Prof, PSIpred, PSSP, SAM-T99, SSPro pour les prédictions de structures secondaires et DAS, TMHMM, TopPred pour les prédictions de région Transmemebranaires. Les résultats sont rendus de façon indépendante.	(Rost and Sander 1994)
NPS@	Permet d'interroger PHD, Predator et des programmes plus anciens tels que la suite GOR. Le résultat présenté n'est pas forcément un consensus.	
PHD	Combine les informations des alignements multiples avec trois réseaux neuronaux -Prédiction de structures secondaires (1) -Prédiction de structures secondaires (2) -Ré-évaluation et corrections des prédictions (3)	(Rost and Sander 1993)
PSIpred	Utilise des matrices de position générées par PSI-BLAST (3 itérations) et les utilise pour son réseau neuronal comme données. Il faut impérativement avoir des homologues dans la base de donnée pour avoir une prédiction fiable.	(McGuffin et al. 2000)
Predator	Méthode particulièrement efficace pour prédire les brins Utilise des alignements locaux qui sont par la suite pondérés. Prends en compte les potentiels d'interactions à longue distance	(Frishman and Argos 1995)
Target-99 /02	Recherche itératives contre des bibliothèques HMM de structures 3D de protéines	(Karplus et al. 1997; Karplus et al. 1998; Karplus and Hu 2001)
NNSSP	Prédiction de la structure secondaire par homologie avec des séquences de structure connue. Le choix de la matrice d'évaluation est critique, un réseau neuronal combine les résultats de six matrices différentes.	(Salamov and Solovyev 1995)

Tableau 4 : Programmes de prédictions de structures secondaires et les méthodes correspondantes.

3. Méthodes pour la prédiction de structures secondaires

Seule, la prédiction de structures secondaires n'apporte pas beaucoup d'informations sur la fonction ou sur le repliement de la protéine. Néanmoins, les prédictions de structures secondaires combinées à d'autres informations (alignements, données biochimiques et structurales) peuvent se révéler très informatives. Les protéines évoluent par des insertions ou des délétions dans des régions qui subissent une moindre pression de sélection (boucles et régions désordonnées). *A contrario,* les éléments de structures secondaires, qui constituent les domaines structurés des protéines, tendent à être conservés. Ainsi une prédiction fiable des structures secondaires aide à affiner un alignement composé de séquences distantes, ou peut parfois permettre de mettre en évidence des similarités de séquences subtiles (motifs dégénérés).

3.1. Approches de première génération

Les premières méthodes de prédiction de structures secondaires, développées par Nagano (Nagano 1973), Chou-Fasman (Chou and Fasman 1974b; a), Garnier, Osguthorpe et Robson (Garnier et al. 1978) reposaient sur une analyse statistique de la propension des acides aminés à se trouver dans des éléments de structures secondaires (α, β). Le nombre de structures disponibles était limité, leur taux de fiabilité est autour de 50%. La méthode de Lim (Lim 1974a; b) fut la première à prendre en compte des règles stéréochimiques dans la prédiction, ce qui a permis de repousser la limite symbolique des 50% à 56%.

Parmi ces premières approches, la méthode GOR (Garnier et al. 1978) est encore utilisée et reste accessible *via* des serveurs de prédictions multiples (Tableau 4). La méthode GOR est basée sur la prédiction de trois états (hélices, brins et coudes). Il s'agit d'une analyse statistique de la fréquence des résidus observés dans une fenêtre de 17 aa. Les éléments de structure secondaire sont assignés en fonction de la composition et de la fréquence des résidus. La prédiction est affinée par la suite, en tenant compte du fait que le nombre minimum de résidus est respectivement quatre ou deux pour les hélices et les brins.

3.2. Les réseaux neuronaux

Les approches de première génération dernièrement développées tels que GORIV (Garnier et al. 1996) arrivaient à un taux de fiabilité de 64,4%. Les nouvelles méthodes de prédiction de structures secondaires utilisent toutes des réseaux neuronaux (RN). Ces méthodes

Méthode	informations			Références
	Limites en aa	Hydrophobicité	Topologie	
TMHMM	OUI	OUI	OUI	(Krogh et al. 2001)
TopPred2	OUI	OUI	OUI	(Claros and von Heijne 1994)
PhDhtm;PhD Topology	OUI	NON	OUI	(Rost et al. 1996)
PSORT; PSORT II	OUI	NON	OUI	(Nakai and Horton 1999)
DAS	OUI	OUI	NON	(Cserzo et al. 1997)
TMPred	OUI	OUI	OUI	(Hofmann and Soffel 1993)
HMMTop	OUI	NON	OUI	(Tusnady and Simon 2001)
TMAP	OUI	NON	OUI	(Persson and Argos 1997)
SOSUI	OUI	OUI	OUI	(Mitaku et al. 2002) (Hirokawa et al. 1998)

Tableau 5 : Programmes de prédiction de régions transmembranaires.

partitionnent la séquence de la protéine en trois états (hélices, brins et boucles) et leur taux de fiabilité est autour de 75%. Les RN sont des systèmes d'apprentissage reposant sur des modèles statistiques complexes. Ils s'organisent en couches d'informations interconnectées. Chaque couche est composée d'unité traitant des informations entrantes ou sortantes. Chaque unité d'une couche reçoit de l'information d'une ou plusieurs autres unités. La réponse de l'unité sera fonction de la pondération de chaque information reçue. Pour être opérationnel, le RN doit être préalablement entraîné, afin d'optimiser les connexions de chaque unité et d'ajuster les facteurs de pondération de chaque connexion. Des approches différentes sont utilisées selon les programmes de prédiction de structures secondaires, ce qui conduit parfois à des résultats contradictoires. Il est nécessaire de comparer plusieurs de ces méthodes afin d'obtenir un consensus. Le tableau 4 présente quelques méthodes (non exhaustives) de prédiction de structures secondaires.

3.4. Méthodes pour l'identification des régions transmembranaires et peptides signaux

De nombreux programmes existent pour identifier des régions transmembranaires (Tableau 6). Tous reposent sur les profils d'hydrophobicités de la chaîne polypeptidique impliquant 16 à 25 résidus et postulent que leur structuration sera en hélice α. Ce dernier paramètre induit un biais dans les méthodes de prédiction puisque des domaines transmembranaires structurés en tonneau β ont été identifiés (Barrow CJ and MG. 1991; Talafous et al. 1994). Par conséquent, aucune méthode automatique de prédiction n'arrive à prédire ces segments transmembranaires structurés en brin (Tableau 5).

Les peptides signaux sont des courts segments hydrophobes composés approximativement d'une vingtaine de résidus se situant à l'une ou l'autre des extrémités de la protéine. De nombreuses études ont été faites (von Heijne 1984; 1985; Claros et al. 1997) et ont conduit au développement du programme SignalIP (Nielsen et al. 1997). Ce dernier est un réseau neuronal qui a été entraîné à reconnaître des séquences de peptides caractérisés, provenant d'eucaryotes ou de bactéries.

3.5. La méthode d'analyse des amas hydrophobes (HCA)

Selon la méthode, les amas hydrophobes reflètent l'environnement de chaque acide aminé dans la structure. Les acides aminés hydrophobes ne sont pas distribués aléatoirement mais ont tendance à se regrouper en amas. Il a été montré statistiquement que ces amas se

Programme	Commentaires	Références
SAM-T99	Ce n'est pas du threading mais un mode de reconnaissance de repliement. Recherche de façon itérative contre une bibliothèque d'alignement HMM, construit un nouveau profil HMM afin d'optimiser la recherche contre la PDB	(Karplus et al. 1997; Karplus et al. 1998; Karplus and Hu 2001)
InBGU	Comparaison des profils de séquences et des structures secondaires pour la séquence cherchée et pour les structures. Il combine cinq méthodes différentes pour produire un consensus	(Fischer 2000)
UCLA/DOE Fold Server	Même principe que INBGU mais la bibliothèque de repliements est différente	(Jaroszewski et al. 2000; Mallick et al. 2000)
GenThreader	Établit un profil à partir de trois itérations de PSI-BLAST avec la séquence. Le profil est utilisé pour faire la recherche de structure. Performant mais nécessite des séquences homologues dans la base de données	(Jones 1999)
3D-PSSM	Comparaison de profils 1D et 3D couplés avec les prédictions de structures secondaires et les potentiels de solvatation.	(Kelley et al. 2000)
FFAS	Etablit un profil à partir de PSI-BLAST avec la séquence. Le profil est comparé au profil de familles de protéines de la PDB	(Rychlewski et al. 2000)
FUGUE	Recherche des séquences homologues à partir d'une bibliothèque d'alignements structuraux. Il génère des alignements globaux ou locaux selon les différences de longueur de séquences	(Shi et al. 2001)
SUPER Family	Ce n'est pas du threading mais un mode de reconnaissance de repliement. Compare la séquence contre une bibliothèque de HMM des superfamilles de SCOP	(Gough et al. 2001; Gough and Chothia 2002)
LOOPP	Effectue la comparaison d'alignements obtenus de façon classique et à partir de profil, avec et sans indels, de façon à décrire la meilleure fonction de “score” pour chaque cas	(Meller and Elber 2001)
123D+	Enfilage des structures secondaires prédites par NNSSP sur une bibliothèque de repliements. L'utilisateur peut choisir le type d'alignement, la matrice de substitution, et la pénalité pour les indels	(Alexandrov and Luethy 1998)

Tableau 6 : Les programmes et les méthodes correspondantes de « Threading » et de reconnaissance de repliement.

retrouvent sur la face interne des éléments de structures secondaires. La forme des amas hydrophobes est caractéristique des éléments de structures secondaires (pour revue, (Callebaut et al. 1997)).

4. Méthode pour la prédiction de structures tertiaires

4.1. Méthode de « Threading » (enfilage)

La technique d'enfilage de séquences de protéines sur une structure tridimensionnelle (« threading ») permet de définir, parmi toutes les structures 3D connues, la structure qui a le repliement le plus compatible avec celui prédit pour la séquence étudiée (Madej et al. 1995; Sippl and Flockner 1996; Smith et al. 1997). La séquence est enfilée sur les structures et le programme calcule le meilleur « enfilage » (Bryant and Altschul 1995; Jones and Thornton 1996; Mirny et al. 2000; Panchenko et al. 2000; Panchenko and Bryant 2002). C'est une méthode qui permet d'identifier des homologues distants. L'idée sous-jacente de cette technique est qu'au cours de l'évolution la structure est mieux conservée que la séquence. La méthode nécessite la comparaison des structures secondaires prédites de la séquence d'intérêt contre des bibliothèques de repliements, sans tenir compte de la similarité de séquences. En général, la méthode implique le calcul de l'énergie de contact des résidus de la séquence superposée sur chaque structure de la bibliothèque. Le classement des résultats est fondé sur un critère d'énergie décroissante. La structure de plus basse énergie sera la structure retenue par le programme (Bryant and Altschul 1995; Jones and Thornton 1996; Miyazawa and Jernigan 1999; Zhang and Kim 2000). Plusieurs modèles statistiques ont été développés afin d'estimer la probabilité de la validité du choix du repliement (Bryant and Altschul 1995; Sunyaev et al. 1997; Mirny et al. 2000). La combinaison du calcul des énergies de contacts avec des profils de séquences et des prédictions de structures secondaires améliore grandement les chances de succès de cette méthode (Jones et al. 1999; Panchenko et al. 1999; Panchenko et al. 2000; McGuffin and Jones 2002). Le tableau 6 liste les principaux programmes de « threading ».

Malgré l'efficacité de la méthode, il faut toujours remettre les résultats obtenus dans le contexte d'une analyse globale (séquences, motifs, définition de domaines, analyse de structures).

Chapitre 3

Méthodes de prédiction du désordre structural

Il y a quelques années maintenant, le paradigme structure–fonction fut quelque peu remis en question par la suggestion faite de l'existence de protéines (ou de domaines protéiques) fonctionnelles dépourvues de structures secondaires et tertiaires stables en conditions physiologiques (Wright and Dyson 1999; Liu et al. 2002; Tompa 2002). En particulier, la composition en acides aminés d'une protéine peut être évaluée en terme de complexité, une faible complexité étant l'un des marqueurs de ces régions non structurées (Wootton 1994).

Ces régions de faible complexité sont un véritable problème lors de la recherche de séquences homologues dans les bases de données, car elles ne sont pas prises en compte lors de l'évaluation statistique de l'alignement. De fait, la présence d'une région de basse complexité dans la requête peut aboutir à un résultat biaisé. En particulier, des régions de faible complexité comme les boucles peuvent êtres alignées ou servir de point d'ancrage pour étendre un alignement sans que cet alignement soit informatif d'un point de vue fonctionnel. Pour compenser ce problème l'algorithme SEG (Wootton and Federhen 1996) permet de filtrer ces régions et donc d'éviter de faire remonter le score de résultats biologiquement non significatifs.

De plus, l'identification des régions désordonnées permet de délimiter les régions globulaires et donc les domaines. Comme déjà mentionné, les domaines se replient de façon autonome et sont souvent séparés par des régions flexibles (non structurées) plus ou moins courtes. De fait, si une protéine contient de nombreuses régions de faible complexité il est possible d'identifier rapidement les limites des domaines globulaires. De nombreuses protéines de pathogènes humains possèdent ce genre d'organisation modulaire. C'est le cas chez les Plasmodium ou les Trypanosomes qui possèdent des boucles désordonnées hypervariables en séquences et qui sont à la base de leur mécanisme d'échappement au système immunitaire de l'hôte (Pizzi and Frontali 2001). Ou encore chez les *Paramyxoviridae,* où les régions non structurées jouent un rôle mécanistique important dans le complexe de réplication (Karlin 2002; Karlin et al. 2003).

Le découpage des séquences en modules globulaires facilite grandement l'étude structurale des protéines. Le problème des régions désordonnées est connu de longue date par les cristallographes. D'un point de vue pratique, ces régions à forte mobilité empêchent souvent la cristallisation des protéines. Dans les cas où des protéines ayant des régions désordonnées cristallisent, leurs structures comportent des régions de densités électroniques manquantes qui reflètent la mobilité de ces régions. S'en débarrasser optimise souvent la qualité du cristal et donc la résolution.

Dans la revue suivante, nous décrivons les principales techniques d'identification du désordre intrinsèque en donnant un exemple de leurs applications et nous présentons notre méthode originale d'identification du désordre.

ARTICLE 1

Combining prediction methods to achieve accurate recognition of disordered regions in proteins.

F. Ferron, S. Longhi, B. Canard and D. Karlin

Revue en préparation pour Protein Sciences

Combining prediction methods to achieve accurate recognition of disordered regions in proteins

François Ferron[1], Sonia Longhi[1*], Bruno Canard[1] and David Karlin[2*]

[1] Architecture et Fonction des Macromolécules Biologiques, UMR 6098 CNRS et Université Aix-Marseille I et II, ESIL, Campus de Luminy, 13288 Marseille Cedex 09, France

[2]Ecole de l'ADN, INMED, Parc de Luminy 13273 Marseille cedex 09

(*) to whom requests for reprints should be addressed
e-mail: **longhi@afmb.cnrs-mrs.fr** and **ecoleadn@agl.univ-mrs.fr**
(2) Tel: (33) (491) 82 86 47
Fax: (33) (491) 82 86 46

Summary

In the last years there has been a growing awareness that a large number of proteins contain long disordered (unstructured) regions that often play a functional role. However, these disordered regions are still widely under-detected. Recognition of disordered regions in a protein is important for several reasons: to avoid bias in sequence similarity analysis, and to help delineate the boundaries of protein domains, so as to guide structural and functional studies. Presently, no fully reliable automated method for disorder prediction is available. We present an overview of the methods currently employed and show how they can be combined to achieve accurate predictions.

Introduction

No current common definition of disorder exists today. We will use a practical definition : a protein region is disordered if it is devoid of stable secondary structure and it has a large number of conformations as seen using X-ray crystallography, NMR, CD (circular dichroism) and a variety of hydrodynamic volume measurements (Tompa 2002). However, this definition embraces several categories of disorder: molten globules, partially unstructured proteins (pre-molten globules), and random coils (by increasing mobility and decreasing residual secondary structure content; see (Uversky 2002)). In this review, we deal only with disordered regions of length >20 aa since the identification of shorter regions is generally not of practical interest to structural biologists.

What is the practical interest of identifying disordered regions ?

a) Disordered regions often have a biased composition that can lead to spurious sequence similarity with unrelated proteins. Such mistakes can be easily avoided by noticing that the query sequence has been aligned with that of a globular protein sequence (for examples, see (Iyer et al. 2001) in Table 2). Therefore, disorder prediction is an essential prerequisite to protein sequence analysis. Moreover, identifying disordered regions narrows the search for short linear functional motifs (called ELMs, because 70% ELMs are found in disordered regions (e.g. SH3, PDZ, phosphorylation sites (Puntervoll et al. 2003)).

b) Disordered regions often prevent crystallization of proteins, or the obtention of interpretable NMR data. Therefore, structural biologists may use disorder predictions to delineate compact domains in order to solve their 3D structure, or to dissect target sequences into a set of independently folded domains to facilitate tertiary structure prediction (Friedberg et al. 2004).

Like in other areas of bioinformatics, the reliability of disorder prediction benefits from the use of several methods based on different concepts, different physico-chemical parameters, or different implementations. From our experience, using a single disorder predictor to achieve predictions good enough to decipher the modular organization of a protein is not realistic, nor is it good bio-computing practice. Herein, we briefly review the sequence features of disordered proteins. Disorder prediction methods are described in Table 1, and informative articles making use of these methods are listed in Table 2. We present an in-depth analysis example of the well-characterized nucleoprotein of measles virus (MV), which illustrates how all prediction methods need to, and can be combined, to achieve accurate disorder prediction.

Sequence features of disordered proteins

Sequence composition

Intrinsically disordered proteins (IDPs) generally have a biased amino acid composition. A consensus of two studies, focusing respectively on the amino acids preferred at the surface of globular proteins or on those found less frequently in secondary structures (Dunker et al. 2001; Linding et al. 2003b) establishes the following list : G, S, P are disorder-promoting amino acids, W, F, I, Y, V, L are order-promoting amino acids, while H, T are considered neutral with respect to disorder. The role of other amino acids varies depending upon their context. From our experience, sequence composition by itself is insufficient at present to be used as a predictive parameter of disorder, in the absence of a rigorous method. For instance, the RNA cap 2'-O-methyltransferase domain of dengue virus polymerase, whose structure has been solved (Egloff et al. 2002), is heavily depleted in some order-promoting residues while at the same time markedly enriched in some disorder-promoting residues (unpublished observations). However, Weathers and co-workers recently reported that amino acid composition alone could allow recognition of proteins with a good accuracy (Weathers et al. 2004). Developments of their method are awaited. In any case, it is good bio-computing practice to always analyse the sequence composition of proteins prior to any sequence analysis (Koonin E and Galperin 2003).

Low sequence complexity

Low Complexity Regions (LCRs) are regions with a biased composition (homopolymeric runs, short-period repeats, and more subtle overrepresentation of a few residues). To put it simply, they make use of fewer types of amino acids. Unstructured proteins tend to have a low sequence complexity (Wootton 1994; Romero et al. 2001).

Some special cases of low-complexity sequences are proteins with a certain amino acid periodicity (such as coiled-coiled region) and other non-globular, yet ordered proteins (collagen for example). As for sequence composition, it is good biocomputing practice to always look for LCRs, coiled-coils, and repeats in a protein prior to any sequence analysis (using programs such as (Lupas et al. 1991), Paircoil (Berger et al. 1995), Multicoil (Wolf et al. 1997)). More subtle parameters to discriminate between globular and non globular proteins using SEG are discussed in (Wootton 1994; Koonin E and Galperin 2003) (see Table 1).

Low secondary structure content

Long (> 70aa) regions devoid of predicted secondary structure elements (as judged by using a combination of methods) are generally disordered. There are a few exceptions, called "loopy proteins", which have no regular secondary structure yet are ordered, like the Kringle domain, a triple-looped, disulphide-linked domain, found in some serine proteases and in some plasma proteins (Liu et al. 2002).

Sequence variability

Disordered regions are on average much more variable than ordered ones (Dunker et al. 2000). The reason why they evolve faster is not clear at present. High sequence variability is not by itself an evidence of disorder, but only an indicator. The relationship between sequence variability and flexibility is well known by crystallographers: when a protein does not crystallize despite repeated attempts, they are used to remove hypervaviable regions, presumed to be flexible linkers. A simple method to appreciate sequence variability is visual inspection of a multiple sequence alignment. However, it can be misleading at times. Programs that rely on nucleotide substitution rate (like in (Dunker et al. 2000)) can be very informative and should be used for a more rigorous analysis (Hurst 2002)

Predictors and indicators of disorder

Several programs have been developed to predict disordered regions using the sequence features reviewed above. Because there is no consensus on what disorder means, it is necessary to know precisely what is predicted by each of these predictors. They are summarized with their philosophy in Table 1. How to combine them to achieve good predictions is described in the following chapter.

The authors of all these predictors warned us that a general error rate for their predictor was difficult to evaluate, since it depends on the training set and on their own definition of disorder. Awaiting for a better standardization of predictors, we have not given error rates in Table 1. In general, predictors are more reliable in predicting order than in predicting disorder, for the simple reason that ordered sequences comprise only a very narrow portion of sequence space, i.e. their sequence properties are much more recognizable. From the authors' personal communications, a conservative error rate for *ab initio* methods, such as Disopred (Jones and Ward 2003), Disembl (Linding et al. 2003a) and PONDR (Iakoucheva et al. 2001) is around 60-70% for predicting disorder, and about 80% for predicting order. The method described by Uversky has the best overall accuracy (83%), menaing that its success rate in predicting order must therefore be higher (about 90%) (V. Uversky, personal communication)

(Garbuzynskiy et al. 2004). However, this method requires prior knowledge of the domain boundaries (Uversky et al. 2000). A novel, promising method based on the predicted number of contacts per residue claims an accuracy of 89% (Garbuzynskiy et al. 2004). We have not had access to it before submission of this manuscript, however.

Combining the different approaches

The recognition of globular and disordered regions requires a systematic sequence analysis combining different methods. As a first step, one should perform an analysis of sequence composition (Wilkins et al. 1999) and complexity (Wootton 1994), a search for signal peptides, transmembrane regions (Krogh et al. 2001) and coiled-coil regions(Lupas et al. 1991; Lupas 1996; 1997), to pre-mark regions of biased composition. HCA is a precious aid for the identification of transmembrane and coiled-coil rgions (Callebaut et al. 1997). Then *ab initio* methods, such as Globplot (Linding et al. 2003b), Disembl (Linding et al. 2003a), PONDR (Iakoucheva et al. 2001), Disopred (Jones and Ward 2003), and NORSp (Liu and Rost 2003)) can be combined to define a consensus on globular and unstructured regions, where Globplot is particularly indicated to highlight globular regions. regions. Of course other information, when available, such as sequence similarity of one region to multi-domain proteins, are precious in terms of domain boundary definition. Once reasonable domain architecture for the target protein is established, the Uversky's method, which has a quite low error rate (see above), can be used to confirm the domain boundaries. However, it should be pointed out that this method should not be applied to regions containing disulphide bridges because this can lead to over-prediction of disorder. Indeed, disulphide bridges can stabilize an otherwise disordered protein sequence, thereby leading to a non random-coil conformation.

Figure 1 and 2 illustrate the approach we used to study the domain organization of the nucleoprotein (N) ofMV, a protein that encapsidates the viral RNA. As shown in Fig. 1, most *ab initio* methods converge to show the presence of a disordered region at its C-terminus (consensus is aa 437-484), and of a globular core (aa 145-344). Interestingly, Foldindex (Zeev-Ben-Mordehai et al. 2003) highlights a very hydrophilic region (aa 100-150) which is visible as a short plateau (aa 131-149) in the output of Disembl Remark 465 predictor and is also predicted by Disopred. Noticeably, this region is hypervariable in sequence (not shown). Finally, as changes in slope of Globplot often correspond to domain boundaries, one would suspect the following domain organization: a first domain or sub-domain encompassing residues 1-130, which is not confidently predicted but might be ordered (*cfr*. negative slope of globplot together with PONDR prediction), an exposed loop spanning aa 131-149, a

second, more compact domain comprised between aa 150 and –400 (*cfr*.steep negative slope), and a disordered domain encompassing aa 401-525. Finally, the Uversky's method predicts that both suspected sub-domains are ordered and confirms that the C-terminal domain is disordered (not shown).

HCA helps refine these predictions. As shown in Fig. 2, the density of hydrophobic clusters indicates without ambiguity that both sub-domains identified by the combination of previous methods are ordered, and the lack of hydrophobic clusters within the region 422-525 indicates that it cannot be ordered by itself (the hydrophobic clusters spanning residues 494-525 being too short to form a compact domain).

In fact, experimental data available indicate that N is organized into 2 regions, N_{CORE} (aa 1-399) and N_{TAIL} (aa 400-525), respectively ordered (Karlin et al. 2002) and disordered (Longhi et al. 2003; Bourhis et al. 2004). The hypervariable region (aa 131-149) is indeed an antigenic loop (Giraudon et al. 1988) that is likely exposed to the solvent(Karlin et al. 2003). No single method nor even a combination of two predictors could successfully highlight this organization of MV N, whereas the combined use of all predictors proved to be much more powerful in terms of domain boundary rcognition. achieve much better results. However, a wealth of mutational data indicates that N_{CORE} cannot be divided into independent modules, but rather than the sub-domains indicated above (aa 1-130 and aa 145-400) probably fold cooperatively. Thus, the exposed region (aa 131-149) is probably a loop and not a linker that would connect the two domains. Whether it is disordered or not is not known. However, as it is not sensitive to proteolysis (Karlin et al. 2002) it is probably at least partially ordered. These unsolved issues nicely illustrate the present limits of disorder prediction.

Refining the analysis

HCA stands aside from other predictors, since they only give insights on the extent of disorder/order, but do not correlate this information with the sequence by itself. Furthermore, there is little one can actually *learn* from comparing the output of these predictors for homologous proteins. HCA fills this gap, giving a more qualitative information on the short range spatial organisation of each amino acid, based on hydrophobic clusters organization. Each plot provides a “texture” of the sequence that can give information not only on order/disorder but also on the folding potential. Indeed, the study of hydrophobic clusters and of secondary structures is of major interest for studying induced folding, because it has long been recognized that burial of hydrophobic residues provides the major driving force in

protein folding. This force is in turn regulated by secondary structures that play a role in guiding the folding pathway. In some cases, hydrophobic clusters are found within secondary structure elements that are unstable in the native protein, but can stably fold upon binding a partner (Johansson et al. 2003; Kingston et al. 2004). As an example, we suspected that the isolated hydrophobic cluster with a predicted alpha-helix at the disordered C-terminus of N (Figure 2) corresponded to a binding region for one partner of N, the viral phosphoprotein P, and that N would undergo induced folding upon binding P. Later, this hypothesis was proven experimentally (Johansson et al. 2003; Bourhis et al. 2004; Kingston et al. 2004) **(**see region highlighted by a purple bar in Fig. 2).

Conclusion

As we've seen from the MV N example, no single predictor can reveal the structural organization of a protein, yet in combination they provide relatively accurate results. There is obvious room for improvement of predictors by combining features of several programs. For instance, they could include not only the density of certain hydrophobic residues, but also their spacing (to unveil part of the information encoded by hydrophobic clusters). Including information on predicted secondary structure elements might also improve some predictors. Other improvements will come from a better understanding, and a standardization of the notion of disorder. As a last note, it would be of major interest to check whether known regions of induced folding correlate well with isolated hydrophobic clusters, corresponding to predicted α-helices, within disordered regions.

Acknowledgements

We thank K. Dunker, P. Romero, J. Ward, R. Linding, V. Uversky, J. Wootton, I. Callebaut, for their useful comments on their respective predictors.

References

Albrecht, M., and Lengauer, T. 2004. Novel Sm-like proteins with long C-terminal tails and associated methyltransferases. *FEBS Lett* **569:** 18-26.

Berger, B., Wilson, D.B., Wolf, E., Tonchev, T., Milla, M., and Kim, P.S. 1995. Predicting coiled coils by use of pairwise residue correlations. *Proc Natl Acad Sci U S A* **92:** 8259-8263.

Bourhis, J.M., Johansson, K., Receveur-Brechot, V., Oldfield, C.J., Dunker, K.A., Canard, B., and Longhi, S. 2004. The C-terminal domain of measles virus nucleoprotein belongs to the class of intrinsically disordered proteins that fold upon binding to their physiological partner. *Virus Res* **99:** 157-167.

Callaghan, A.J., Aurikko, J.P., Ilag, L.L., Gunter Grossmann, J., Chandran, V., Kuhnel, K., Poljak, L., Carpousis, A.J., Robinson, C.V., Symmons, M.F., et al. 2004. Studies of the RNA degradosome-organizing domain of the Escherichia coli ribonuclease RNase E. *J Mol Biol* **340:** 965-979.

Callebaut, I., Courvalin, J.C., Worman, H.J., and Mornon, J.P. 1997. Hydrophobic cluster analysis reveals a third chromodomain in the Tetrahymena Pdd1p protein of the chromo superfamily. *Biochem Biophys Res Commun* **235:** 103-107.

Callebaut, I., Labesse, G., Durand, P., Poupon, A., Canard, L., Chomilier, J., Henrissat, B., and Mornon, J.P. 1997. Deciphering protein sequence information through hydrophobic cluster analysis (HCA): current status and perspectives. *Cell Mol Life Sci* **53:** 621-645.

Dunker, A.K., Lawson, J.D., Brown, C.J., Williams, R.M., Romero, P., Oh, J.S., Oldfield, C.J., Campen, A.M., Ratliff, C.M., Hipps, K.W., et al. 2001. Intrinsically disordered protein. *J Mol Graph Model* **19:** 26-59.

Dunker, A.K., Obradovic, Z., Romero, P., Garner, E.C., and Brown, C.J. 2000. Intrinsic protein disorder in complete genomes. *Genome Inform Ser Workshop Genome Inform* **11:** 161-171.

Egloff, M.P., Benarroch, D., Selisko, B., Romette, J.L., and Canard, B. 2002. An RNA cap (nucleoside-2'-O-)-methyltransferase in the flavivirus RNA polymerase NS5: crystal structure and functional characterization. *Embo J* **21:** 2757-2768.

Friedberg, I., Jaroszewski, L., Ye, Y., and Godzik, A. 2004. The interplay of fold recognition and experimental structure determination in structural genomics. *Curr Opin Struct Biol* **14:** 307-312.

Garbuzynskiy, S.O., Lobanov, M.Y., and Galzitskaya, O.V. 2004. To be folded or to be unfolded? *Protein Sci* **13:** 2871-2877.

Garner, E., Romero, P., Dunker, A.K., Brown, C., and Obradovic, Z. 1999. Predicting Binding Regions within Disordered Proteins. *Genome Inform Ser Workshop Genome Inform* **10:** 41-50.

Giraudon, P., Jacquier, M.F., and Wild, T.F. 1988. Antigenic analysis of African measles virus field isolates: identification and localisation of one conserved and two variable epitope sites on the NP protein. *Virus Res* **10:** 137-152.

Hurst, L.D. 2002. The Ka/Ks ratio: diagnosing the form of sequence evolution. *Trends Genet* **18:** 486.

Iakoucheva, L.M., Kimzey, A.L., Masselon, C.D., Bruce, J.E., Garner, E.C., Brown, C.J., Dunker, A.K., Smith, R.D., and Ackerman, E.J. 2001. Identification of intrinsic order and disorder in the DNA repair protein XPA. *Protein Sci* **10:** 560-571.

Iyer, L.M., Aravind, L., Bork, P., Hofmann, K., Mushegian, A.R., Zhulin, I.B., and Koonin, E.V. 2001. Quoderat demonstrandum? The mystery of experimental validation of apparently erroneous computational analyses of protein sequences. *Genome Biol* **2:** RESEARCH0051.

Johansson, K., Bourhis, J.M., Campanacci, V., Cambillau, C., Canard, B., and Longhi, S. 2003. Crystal structure of the measles virus phosphoprotein domain responsible for the induced folding of the C-terminal domain of the nucleoprotein. *J Biol Chem* **278:** 44567-44573.

Jones, D.T., and Ward, J.J. 2003. Prediction of disordered regions in proteins from position specific score matrices. *Proteins* **53 Suppl 6:** 573-578.

Karlin, D., Ferron, F., Canard, B., and Longhi, S. 2003. Structural disorder and modular organization in Paramyxovirinae N and P. *J Gen Virol* **84:** 3239-3252.

Karlin, D., Longhi, S., and Canard, B. 2002. Substitution of two residues in the measles virus nucleoprotein results in an impaired self-association. *Virology* **302:** 420-432.

Kingston, R.L., Baase, W.A., and Gay, L.S. 2004. Characterization of nucleocapsid binding by the measles virus and mumps virus phosphoproteins. *J Virol* **78:** 8630-8640.

Koonin E, V., and Galperin, M. 2003. *Sequence-Evolution-Function: Computational Approaches in Comparative Genomics*. Kluwer Academic Publishers, pp. 488.

Linding, R., Jensen, L.J., Diella, F., Bork, P., Gibson, T.J., and Russell, R.B. 2003a. Protein disorder prediction: implications for structural proteomics. *Structure (Camb)* **11:** 1453-1459.

Linding, R., Russell, R.B., Neduva, V., and Gibson, T.J. 2003b. GlobPlot: Exploring protein sequences for globularity and disorder. *Nucleic Acids Res* **31:** 3701-3708.

Liu, J., and Rost, B. 2003. NORSp: Predictions of long regions without regular secondary structure. *Nucleic Acids Res* **31:** 3833-3835.

Liu, J., Tan, H., and Rost, B. 2002. Loopy proteins appear conserved in evolution. *J Mol Biol* **322:** 53-64.

Longhi, S., Receveur-Brechot, V., Karlin, D., Johansson, K., Darbon, H., Bhella, D., Yeo, R., Finet, S., and Canard, B. 2003. The C-terminal domain of the measles virus nucleoprotein is intrinsically disordered and folds upon binding to the C-terminal moiety of the phosphoprotein. *J Biol Chem* **278:** 18638-18648.

Lupas, A. 1996. Prediction and analysis of coiled-coil structures. *Methods Enzymol* **266:** 513-525.

Lupas, A. 1997. Predicting coiled-coil regions in proteins. *Curr Opin Struct Biol* **7:** 388-393.

Lupas, A., Van Dyke, M., and Stock, J. 1991. Predicting coiled coils from protein sequences. *Science* **252:** 1162-1164.

Puntervoll, P., Linding, R., Gemund, C., Chabanis-Davidson, S., Mattingsdal, M., Cameron, S., Martin, D.M., Ausiello, G., Brannetti, B., Costantini, A., et al. 2003. ELM server: A new resource for investigating short functional sites in modular eukaryotic proteins. *Nucleic Acids Res* **31:** 3625-3630.

Rabitsch, K.P., Gregan, J., Schleiffer, A., Javerzat, J.P., Eisenhaber, F., and Nasmyth, K. 2004. Two fission yeast homologs of Drosophila Mei-S332 are required for chromosome segregation during meiosis I and II. *Curr Biol* **14:** 287-301.

Romero, P., Obradovic, Z., Li, X., Garner, E.C., Brown, C.J., and Dunker, A.K. 2001. Sequence complexity of disordered protein. *Proteins* **42:** 38-48.

Tompa, P. 2002. Intrinsically unstructured proteins. *Trends Biochem Sci* **27:** 527-533.

Uversky, V.N. 2002. Natively unfolded proteins: a point where biology waits for physics. *Protein Sci* **11:** 739-756.

Uversky, V.N., Gillespie, J.R., and Fink, A.L. 2000. Why are "natively unfolded" proteins unstructured under physiologic conditions? *Proteins* **41:** 415-427.

Weathers, E.A., Paulaitis, M.E., Woolf, T.B., and Hoh, J.H. 2004. Reduced amino acid alphabet is sufficient to accurately recognize intrinsically disordered protein. *FEBS Lett* **576:** 348-352.

Wilkins, M.R., Gasteiger, E., Bairoch, A., Sanchez, J.C., Williams, K.L., Appel, R.D., and Hochstrasser, D.F. 1999. Protein identification and analysis tools in the ExPASy server. *Methods Mol Biol* **112:** 531-552.

Wolf, E., Kim, P.S., and Berger, B. 1997. MultiCoil: a program for predicting two- and three-stranded coiled coils. *Protein Sci* **6:** 1179-1189.

Wootton, J.C. 1994. Non-globular domains in protein sequences: automated segmentation using complexity measures. *Comput Chem* **18:** 269-285.

Zeev-Ben-Mordehai, T., Rydberg, E.H., Solomon, A., Toker, L., Auld, V.J., Silman, I., Botti, S., and Sussman, J.L. 2003. The intracellular domain of the Drosophila cholinesterase-like neural adhesion protein, gliotactin, is natively unfolded. *Proteins* **53:** 758-767.

Legends

Table 1: Software tools for disorder prediction and related methods. at the time of writing, we have not had access to a promising novel method based on the predicted number of contacts per residue (Garbuzynskiy et al. 2004). Therefore we have not included it in the table.

Table 2: Predictor success stories.

Figure 1: Measles Virus nucleoprotein analysed with different predictors. The output of each method and the corresponding interpretation is shown.

Figure 2: HCA plot of Measles Nucleoprotein.

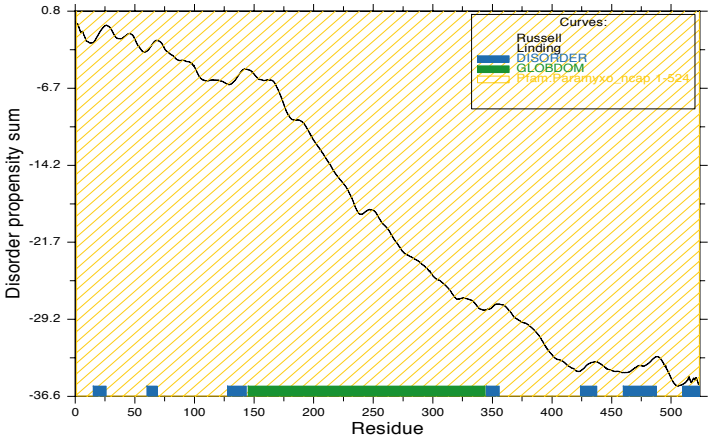
Conventions are explicited in the caption. Globular regions (framed) are characterized by a thick distribution of hydrophobic clusters, while unstructured regions are poor or devoid of hydrophobic clusters. Long disordered regions and predicted secondary structure elements are shown. There are no low-complexity region in MV N. The induced folding region is underligned in purple, and the corresponding structure (purple α-helix) is presented in complex with the C-terminal domain of MV P (PDB code : 1T6O). The image was obtained using Pymol.

Table 1

	PONDR	SEG	Disopred	Globplot	Disembl	NORSp	FoldIndex	Uversky	HCA
What is predicted	All regions that are not rigid including random coils, partially unstructured regions, and molten globules	Low-complexity segments i.e. "simple sequences" or "compositionally-biased regions"	Regions devoid of ordered regular secondary structure	Regions with high propensity for globularity on the Russell/Linding scale (see below)	LOOPS (regions devoid of regular secondary structure); HOT LOOPS (highly mobile loops); REMARK465 (regions lacking electron density in crystal structure)	Regions with No Ordered Regular Secondary structure (NORS). Most, but not all, are highly flexible	Regions that have a low hydrophobicity and high net charge	Regions with an elusive property of either disorder or order	Hydrophobic clusters, which tend to form secondary structure elements
Methods	Local aa composition, flexibility, hydropathy, etc	Locally-optimized low-complexity segments are produced at defined levels of stringency, based on formal definitions of local compositional complexity (Wootton & Federhen, 1993). The segment lengths and the number of segments per sequence are determined automatically by the algorithm	Cascaded support vector machines trained on PSI-BLAST profiles	Russell/Linding scale of disorder (propensities for secondary structures and random coils)	Neural networks trained on X-ray structure data	Secondary structure and solvent accessibility	Sequence composition (hydrophobicity versus net charge)	Sequence composition (hydrophobicity versus net charge)	Helical visualization of amino acid sequence
Generates and use multiple sequence alignment?	No	No	Yes	No	No	Yes	No	No	No
Minimal protein length (aa) for reliable prediction	40	-	60	40	-	50 (default is 70)	40	40	-
Remarks	Standard predictor is VL-XT. Sharp drops in the middle of long disordered regions might indicate binding sites. Another PONDR predictor, VL3, gives smoother predictions and can be used to delineate domains	The stringency of the search for low-complexity segments is determined by 3 user-defined parameters: trigger window length [W], trigger complexity [K(1)] and extension complexity [K(2)] Parameters for disorder prediction: - for long non-globular domains, use long window lengths, typically: seg sequence 45 3.4 3.75 - for shorter non-globular domains, typically use: seg sequence 25 3.0 3.3	Prediction accuracy is lower if there are few homo-logues available	Built-in SMART, PFAM and low complexity predictions Gives easy overview of modular organization of large proteins Changes of slope often correspond to domain boundaries	Use the Loop predictor only as a filter to remove false disorder predictions of Hot Loops and Remark465	Beware: some NORS are rigid, whereas some highly mobile regions have predicted secondary structure Prediction accuracy is lower if there are few homologues available	This is a derivative of the Uversky's method used with a sliding window. Long regions give the same result as Uversky. For shorter regions it has not been validated Highlights some regions that are probably short loops better than a hydrophobicity plot.	Requires prior knowledge of modular organization of protein. Applicable only to proteins without disulfide bonds	Scientist's interpretation required Highlights regions with potential for induced folding Highlights very short potential globular domains Allows meaningful comparison with related proteins

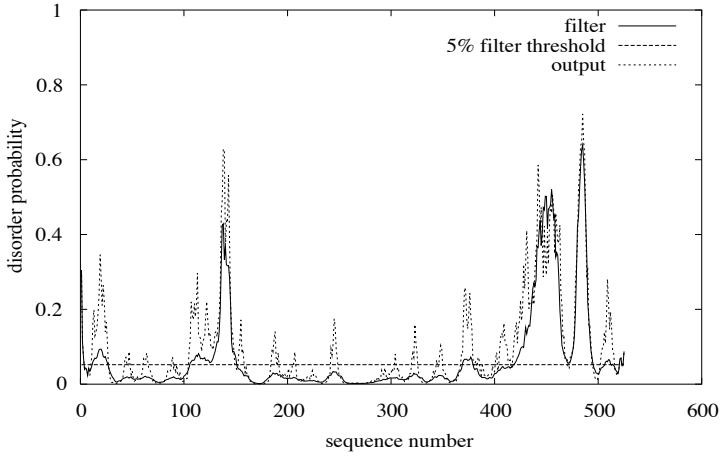
Table 2.

(Iyer et al. 2001)	Two examples in which SEG in combination with multiple alignment and secondary structure prediction invalidates previous functional assignments for 2 proteins, ATF-2 and PIF3, made on the basis of distant sequence similarity to 2 domains (respectively histone acetyltransferase (HAT) and PAS domain). The HAT and PAS domains are globular, whereas the similar regions of ATF-2 and PIF3 are confidently predicted to be unstructured, casting a strong doubt on their suspected homology.
(Callaghan et al. 2004)	Globplot and PONDR are used to define disordered regions. A fine analysis of PONDR plots (Garner et al. 1999) led to the successful identification of segments of increased structural propensity (i.e. prone to induced folding) in the RNA degradosome-organizing domain of the *E. coli* ribonuclease RNase E.
(Albrecht and Lengauer 2004)	This article identifies 5 novel groups of Lsm domain proteins. The architecture of target proteins, which guides the research for sequence similarities, is elucidated using the consensus of Globplot, Disembl, Norsp, and PONDR. The authors identify conserved motifs described as "stable islands in a large sea of intrinsically unstructured sequence regions". This is probably true of many large human proteins for which very short conserved motifs in the middle of long disordered regions remain to be discovered.
(Rabitsch et al. 2004)	This article illustrates how to perform search for homologs of proteins (Sgo1 and Sgo2) composed mostly of unstructured regions. The authors searched for candidate proteins having short stretches of sequence similarity or structural similarity (coiled-coil) to Sgo1 and Sgo2, distributed in a similar fashion (at the N- and C-terminus).
(Karlin et al. 2003)	Dissection of the architecture of 2 proteins of the virus family *Paramyxoviridae*. The presence of unstructured regions and other predictions of induced folding have been later confirmed experimentally. Using a combination of HCA, PONDR, Uversky's method, and biological insight, the P proteins of *Paramyxovirinae* were found to have a common architecture comprising 6 modules (structured or not), despite the lack of sequence similarity. The protein N is discussed in the present review.
(Callebaut et al. 1997)	Identification of short (60-70 aa), globular domains (called chromodomains) often located within long, disordered regions using HCA.
Personal communications and unpublished observations	
Dr. Charles S. Bond MRSC Division of Biological Chemistry and Molecular Microbiology University of Dundee C.S.Bond@dundee.ac.uk	All the constructs of a heterocomplex of two human nucleic-binding proteins were poorly expressed and/or insoluble in *E. coli*. These proteins were truncated on the basis of a combination of inspection of crystal structures of distant proteins and of the output of GLOBPLOT. The complex could then be efficiently expressed.



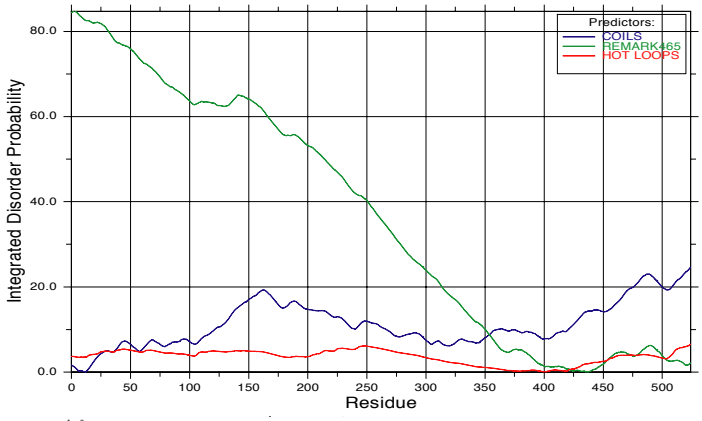
Globplot of MV Nucleoprotein

Conclusions: A globular domains spanning residues 145-344 (▬) is predicted. Other regions are not reliably predicted.



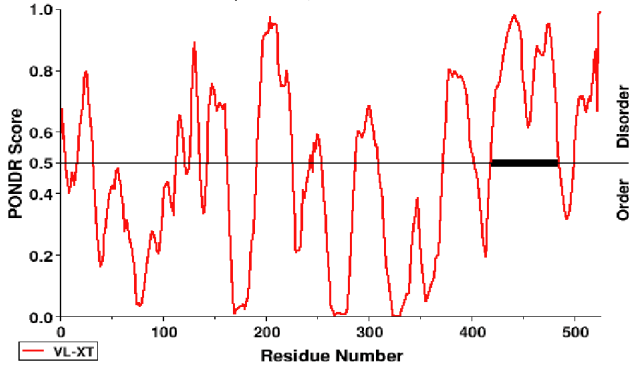
Disopred of MV Nucleoprotein

Conclusions: Disordered regions spanning residues 131-149 and 426-494. Other regions are predicted as globular.



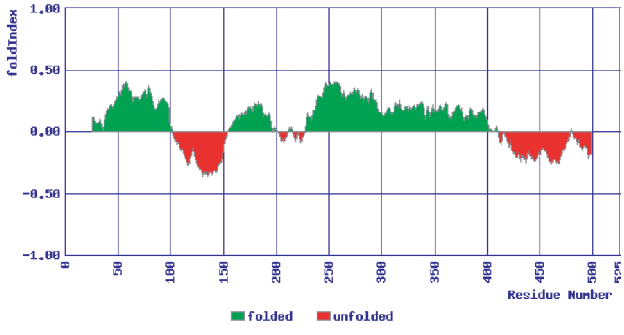
DisEMBL of MV Nucleoprotein

Conclusions: Disordered regions spanning residues 131-144, 437-466, and 478-490. Other regions are predicted as globular. The interpretation is based on remark 465, and depends on the down-ward slope.



PONDR® of MV Nucleoprotein

Conclusions: A single disordered region (aa 419-484) is predicted (thick black line).



FoldIndex of Nucleoprotein MV

Conclusions: Disordered regions (▪) spanning residues 100-150 and 420-494. Other regions are predicted as globular (▬).

Figure 1

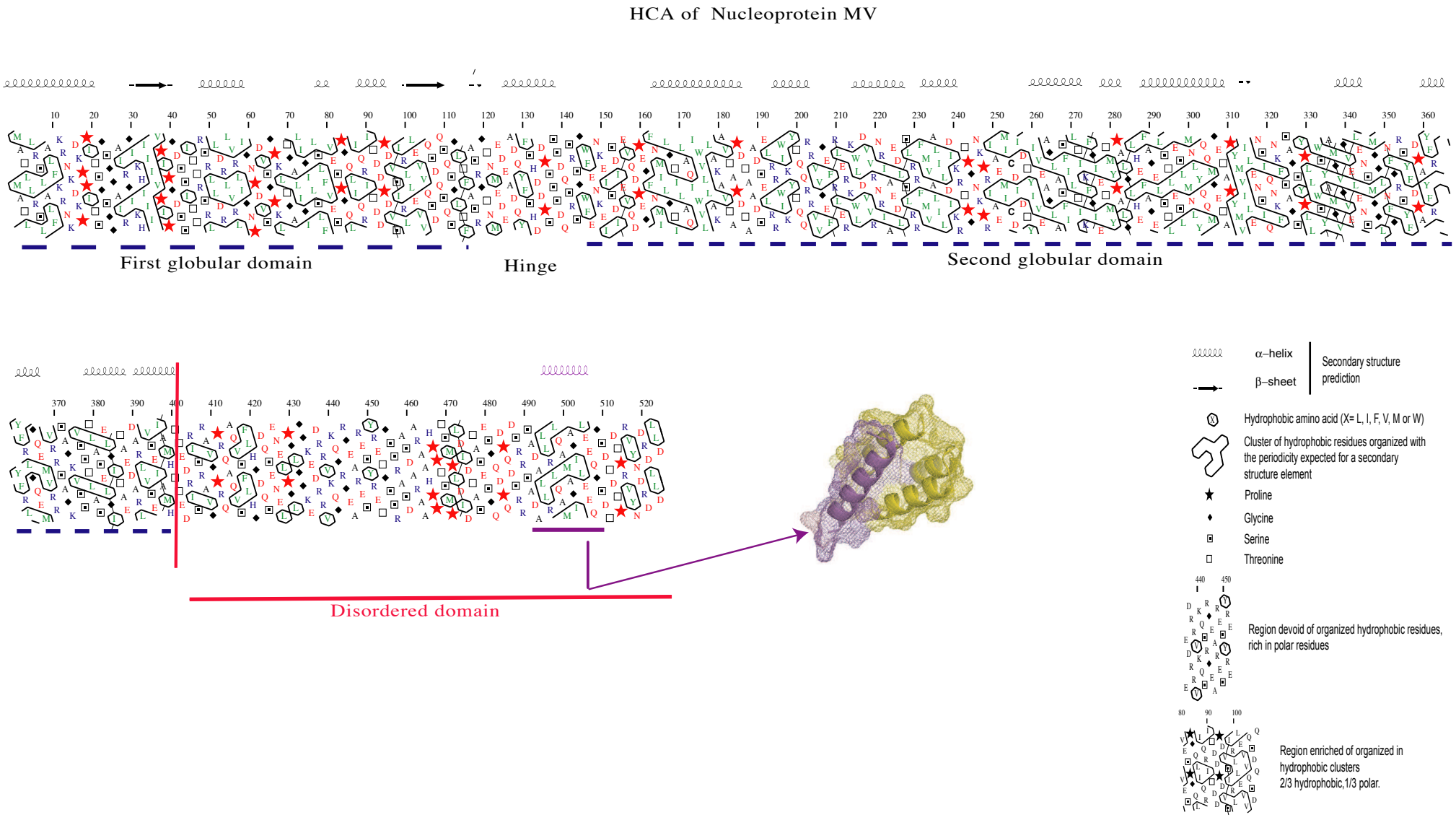

HCA of Nucleoprotein MV
First globular domain
Hinge
Second globular domain
Disordered domain
α–helix
β–sheet
Secondary structure prediction
Hydrophobic amino acid (X= L, I, F, V, M or W)
Cluster of hydrophobic residues organized with the periodicity expected for a secondary structure element
Proline
Glycine
Serine
Threonine
Region devoid of organized hydrophobic residues, rich in polar residues
Region enriched of organized in hydrophobic clusters
2/3 hydrophobic,1/3 polar.

Figure 2

Chapitre 4

La base de données VaZyMolO

1. Introduction

Le projet « VaZyMolO » est la construction d'une base de données qui gère et organise des familles de séquences de protéines virales ainsi que les informations qui s'y rapportent. Après avoir présenté le leitmotiv du projet, je développerai les différents aspects de la base de données, ses outils d'annotation et son interface publique.

1.1. Fiabilité des bases de données publiques

L'un des moteur du projet « VaZyMolO » vient du constat qu'il existe un problème de fiabilité concernant les données que l'on trouve dans les bases de données. Bien sûr, selon que l'on travaille avec une banque de données (archives et de stockage) ou une base de données maintenue par des « experts », le degré de fiabilité est différent. Je vais présenter quelques exemples concernant les protéines virales qui illustrent ce problème et qui concernent toutes les bases de données.

Les banques d'archives ne peuvent être éditées que par les découvreurs de la séquence, ce qui génère un nombre considérable de fausses informations. A titre d'exemple, l'entrée NC_002534 du NCBI du génome du « lactate dehydrogenase-elevating » virus contient des « w ». Ce genre d'erreur rend impossible l'utilisation de la séquence.

D'une certaine façon, les bases de données reflètent plus notre ignorance que notre savoir. En effet, seule une petite minorité de produits géniques de quelques organismes ont été caractérisés directement. La plupart des informations retrouvées dans les banques de données sont en fait issues d'un transfert d'informations d'une séquence caractérisée vers des séquences homologues. Les bases de données qui sont maintenues et annotées par des experts sont fiables dans leur ensemble. Néanmoins, il est fondamental de conserver un esprit critique vis à vis de l'information trouvée et ce, pour au moins trois raisons.

1.1.1. Perte d'information dans la base

Du fait du très grand nombre de séquences dans les bases de données, l'analyse relationnelle structure-séquence est rarement poussée à son maximum, ainsi les annotations se retrouvent diluées dans la masse de nouvelles séquences. C'est, par exemple, le cas de l'entrée P35975 dans Pfam pour laquelle le domaine méthyltransférase n'est pas référencé.

1.1.2. Problème du suivi des mises à jour

Dans la plupart des cas, les transferts d'informations de séquence à séquence se font de façon raisonnable, sur la base de critères objectifs (conservation des motifs, homologie de séquence, significative, homologie structurale). Mais il arrive que la biologie contredise ces annotations. Par exemple, l'entrée UNIprot Q54527 est assignée comme une O-méthyltransférase, mais les récentes données de Jansson (Jansson et al. 2003) ont montré qu'il s'agit en fait d'une hydrolase. La correction des bases de données peut prendre des années.

1.1.3. Transfert systématique de l'information sans esprit critique

L'exemple suivant, illustre à lui seul le problème du transfert systématique de l'information, il concerne l'entrée AF12241 du NCBI. Elle se rapporte au bactériophage SPP1 qui possède un gène codant pour une protéine structurale, appelé « tête du bactériophage ». Cette protéine possède une certaine homologie avec une protéine d'une bactérie. Cette dernière a été annotée comme protéine impliquée dans la formation de la tête (ce qui est plutôt surprenant pour une bactérie).

1.2. Absence de bases de données virales

L'idée de construire une base de données tels que « VaZyMolO » est venue du constat de l'absence de ressources spécifiques disponibles pour les virologistes. Comme nous l'avons vu précédemment dans le premier chapitre, il existe des bases de données généralistes ou spécialisées qui assurent la mise à disposition des séquences, des domaines, des structures de tous les organismes. Actuellement il existe peu de bases de données dédiées aux virus. VIDA (Alba et al. 2001) offre des alignements de séquences de quelques familles de virus (cf. Chapitre 1) et le « comité international de classification taxonomique virale » (ICTV) (Pringle 1995) a mis en ligne sa propre classification. D'autres développements ponctuels de bases de données ont été tentés (Hiscock and Upton 2000), mais ces efforts ne répondent pas à une approche globale (séquences, structures, fonctions). Les virus méritent cependant d'être traités spécifiquement. Les virus doivent s'adapter en permanence et rapidement à leur environnement sous peine de disparaître. Cette pression de sélection se traduit par un taux de mutation élevé chez beaucoup de virus et en particulier chez les virus à ARN (de l'ordre de 10^{-4} par cycle de réplication). La conséquence de ces taux de mutations élevés se traduit par un rythme d'évolution rapide et par une divergence des séquences protéiques. Néanmoins, cette divergence, n'est pas forcément retrouvée au niveau structural. De plus, les séquences

virales sont noyées parmi celles des autres organismes. Elles constituent un bruit de fond car bien souvent elles ne correspondent à rien d'autre. Le récent projet du NCBI (Bao et al. 2004) tend à rattraper ce retard. Cependant, même s'il progresse vite dans la structuration de l'information, l'annotation de nombreuses protéines reste succincte.

1.3. VaZyMolO « Viral enZyme Module lOcalisation »

1.3.1. Objectif de la base de données

Ces quelques exemples, qui peuvent faire sourire, reflètent malheureusement le peu d'effort qui a été porté jusqu'ici sur les protéines virales. Il a fallu une menace majeure sur le monde (épidémie du virus du SRAS en 2003, agent pathogène issu d'une famille particulièrement négligée, les *Coronavirus*) pour que les grands organismes, comme le NCBI commencent à constituer de vrais groupes spécialisés dans l'annotation des protéines virales. La faible qualité des données disponibles et l'absence de bases spécialisées nous ont poussé à développer une base de données interne au laboratoire.

Cela dit, le but de VaZyMolO n'est pas de se substituer aux grandes bases de données, mais de proposer une alternative à la façon qu'ont les bases de données d'aborder le travail sur les séquences virales.

-Le premier objectif est de cartographier systématiquement les protéines codées par les génomes viraux. Cette cartographie a deux perspectives : structurale et fonctionnelle d'une part, et évolutive d'autre part.

-Le second objectif de la base de données est de faciliter les recherches des utilisateurs, en leur fournissant le plus d'informations possibles sur la(es) protéine(s) d'intérêt, mais aussi sur des protéines apparentées sur lesquelles beaucoup d'informations seraient déjà disponibles.

1.3.2. VaZyMolO : base de données relationnelle

Pour stocker et interroger le savoir contextuel rassemblé sur les protéines virales, nous avions besoin de définir un cadre de travail. Le choix qui s'est imposé à nous fut celui de la base de données relationnelle. Les bases de données relationnelles organisent et gèrent les données selon un schéma structuré, et le langage utilisé est le « langage de questionnement structuré » ou SQL (structured query language).

Si le choix du cadre de travail à été facile à définir, en revanche la conception de VaZyMolO n'a pas été triviale. D'une part, l'ensemble du type de données que l'on souhaitait pouvoir

consulter n'était pas clairement défini au début du projet. D'autre part, VaZyMolO dérive (et j'insiste sur le fait qu'il ne s'agit pas d'un clone) de la base de données CAZyModO (cf. Chapitre 1). Le noyau CAZY_3.1 (incluant architecture et interface) sur lequel nous avons commencé à travailler n'était pas optimal pour la gestion des séquences virales et la gestion d'utilisateurs multiples. A décharge, les premières phases de création d'une base de données sont souvent des essais de projection et de prédiction afin de trouver des exemples de données que l'on souhaiterait pouvoir consulter ou qui sont nécessaires pour effectuer une annotation complète et de qualité.

1.3.3. Organisation des données

D'une part, VaZyMolO tend à organiser de façon hiérarchique les informations se rapportant aux protéines virales, qu'elles soient séquentielles, structurales, ou fonctionnelles. D'autre part, VaZyMolO est aussi un outil qui permet de définir des modules protéiques et à les classer en familles de modules. Ces familles peuvent être regroupées en clan ou subdivisées en sous-familles. Pour notre classification, nous avons repris l'approche développée dans la base de données « CAZyModO ». L'approche modulaire a été choisie compte tenu des exigences dictées par les projets de génomique structurale développés au laboratoire et de la nécessité d'obtenir à grande échelle des protéines solubles et cristallisables, en vue de leur étude par cristallographie aux rayons X. Par conséquent, le découpage en modules est guidé par la structure et reflète l'organisation structurale. Il reste néanmoins cohérent pour l'ensemble des disciplines de la biologie, un biologiste moléculaire ou un virologiste peut ainsi rapidement identifier et reconnaître à quelle famille appartient une nouvelle séquence virale, il peut proposer très rapidement une classification. Le classement des modules en famille permet aussi de repérer les régions variables et conservées au sein du génome du virus, ce qui permet d'identifier les zones qui sont particulièrement soumises à la pression de sélection. Les biologistes structuraux ont ainsi une vision instantanée de ce qui est connu en terme structural au sein des familles. Cela leur permet de définir leurs priorités et d'établir facilement des plans d'expériences. Les enzymologistes peuvent rationaliser leurs expériences et leurs interprétations en reliant facilement leurs données mécanistiques aux données structurales.

Lorsque des informations fonctionnelles ou structurales sont disponibles, nous établissons systématiquement les liens vers des bases de données factuelles correspondantes.

Enfin le dernier point d'importance est le choix des séquences de travail. Dans un premier temps, nous avons considéré la possibilité d'utiliser comme source de données SWISS-PROT (UNIprot n'existait pas encore au début du projet), puisque cette dernière est la base de référence en terme de qualité des données. Cependant, nous avons constaté que pour un certain nombre de virus, SWISS-PROT ne répertorie pas toutes les séquences protéiques. La redondance d'information qui caractérise cette base peut constituer un problème non négligeable et qui a fini par nous dissuader de l'utiliser. Ainsi, notre choix s'est porté sur le NCBI qui fournit la liste et les séquences des génomes viraux complètement séquencés. La redondance est faible et la représentation du nombre de virus par familles virales est relativement homogène. Chaque gène du génome bénéficie d'un numéro d'accès (NC) ce qui permet de les identifier rapidement. Conscients des limites que cela implique eu égard aux problèmes d'erreurs lors des dépôts, ce choix nous est apparu comme le meilleur.

Dans le but de simplifier la consultation des informations, les modules sont assignés en trois catégories de protéines : les protéines de surface (S), les protéines de soutien ou matrices (M) et les protéines virales ayant un caractère fonctionnel (F). C'est au sein de ces catégories que sont créées les familles modulaires, et de fait, les familles peuvent être transversales par rapport à la classification taxonomique. Cette organisation simplifie la recherche pour l'utilisateur, mais n'interfère pas dans le travail d'annotation de VaZyMolO.

2. Structuration de l'information

2.1. Les séquences

VaZyMolO gère les séquences protéiques codées par les gènes viraux. Deux cas de figure se présentent alors:

- gène codant pour une protéine ;
- gène codant pour une polyprotéine.

Selon les virus, les polyprotéines peuvent être composées de deux à dix-huit protéines, qui sont ensuite produites par clivage protéolytique. Dans VaZyMolO, la gestion de la séquence est différente suivant qu'il s'agisse d'une séquence protéine ou d'une polyprotéine. Cette distinction est nécessaire pour permettre au curateur, comme à l'utilisateur, de ne pas être submergé d'informations lors de l'annotation ou de la consultation.

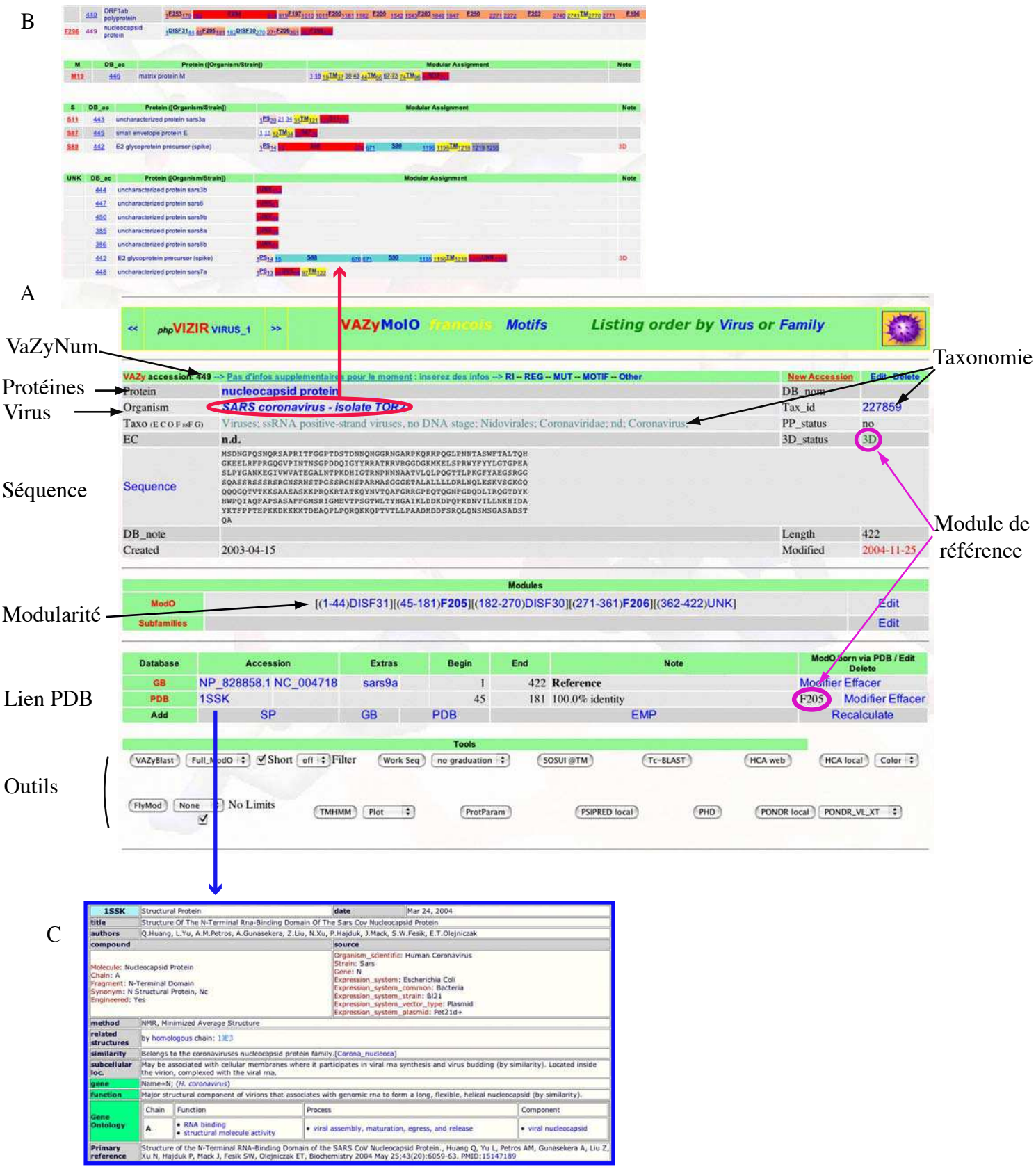


Figure 16 : Interface de gestion d'une protéine A) Entrée complète de la protéine. B) Représentation modulaire de toutes les protéines du virus. C) Fiche technique des données structurales se rapportant au module référant.

2.1.1. Les séquences de protéines

A chaque séquence entrée dans la base de données est attribué un numéro unique, le « VaZyNum » qui va servir de lien avec toutes les informations qui seront rattachées à la séquence par la suite. Pour conserver une trace quant à l'origine de la séquence, au numéro VaZyNum sont associés le numéro GenBank (qui correspond à l'entrée NCBI originale de la séquence en base) et le numéro Genpep (qui correspond à l'entrée NCBI originale de la séquence en acides aminés). Le numéro de suivi taxonomique (NCBI-ICTV) est aussi lié au VaZyNum (Figure 16).

Dans VaZyMolO, chaque séquence est caractérisée par : nom générique de la protéine, nom du virus, classification taxonomique, longueur de la séquence, date de création de l'entrée, date de la dernière modification et éventuellement un nom d'usage courant interne au laboratoire. A chaque champ est associé un lien vers une information complémentaire, par exemple au nom du virus est associé un lien permettant de retrouver l'intégralité des protéines de ce dernier (Figure 16).

Nous vérifions ensuite si la structure de la protéine (ou d'une partie) est connue. Pour cela, Nous effectuons un BLAST contre la PDB. Si la recherche aboutie à l'identification d'une structure deux cas de figure peuvent se présenter :

- il s'agit de la séquence entière, *i.e.* , la protéine sera définie comme un module unique, même si elle peut être composée de plusieurs domaines
- il s'agit d'une partie de la séquence, *i.e.* , la protéine a une organisation modulaire.

Dans tous les cas, le code PDB, la référence de la chaîne ainsi que les limites de la partie de la séquence correspondant au(x) module(s) sont reportés sur l'entrée correspondante. Le module ainsi défini est assigné comme module de référence. Une mention « 3D » indique que cette protéine possède au moins un module référent. Le découpage modulaire est reporté sur l'entrée, sous la forme linéaire des bornes du/des modules, et du nom du/des modules. Cette description reprend l'intégralité de la séquence (Figure 16). Nous verrons un peu plus loin le détail de la procédure d'assignation modulaire. Pour chaque entrée, nous avons la possibilité d'attribuer des informations spécifiques à la protéine, par exemple la description de motifs et/ou de mutants.

2.1.2. Les séquences de polyprotéines

D'un point de vue bioinformatique et selon la philosophie de VaZyMolO, une polyprotéine est une protéine qui possède plusieurs modules de structure et de fonction différentes. Il nous

A

VaZyNum

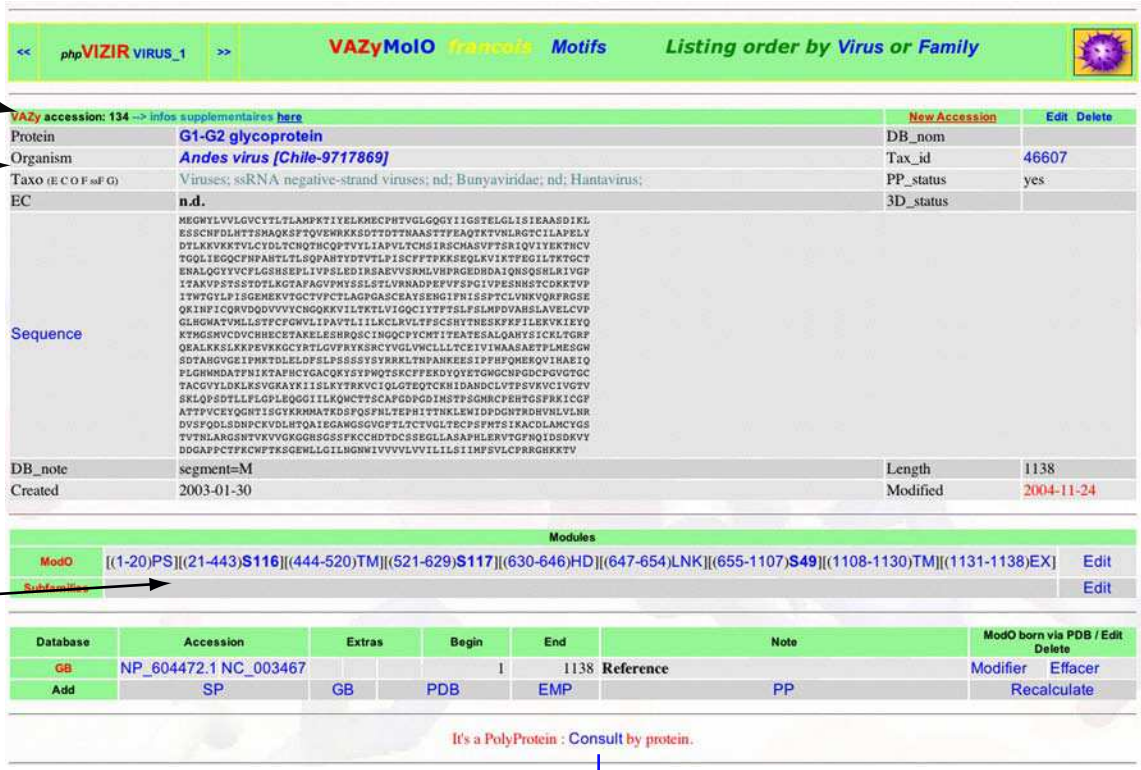
Protéines

Virus

Séquence de la Polyprotéine

Date de création et d'édition

Modularité de la polyprotéine



B

VaZyNum_PP

Découpage en protéine

C

Séquence de la protéine

Outils

Modularité de la protéine

Numérotation globale

Numérotation spécifique

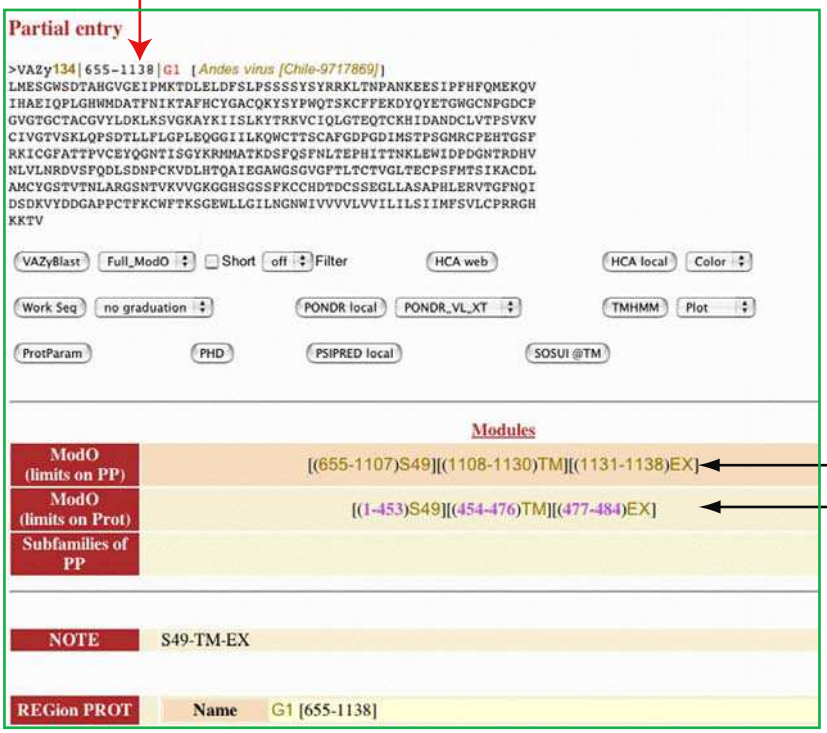


Figure 17 : Les interfaces de gestion des polyprotéines : A) Entrée complète de la polyprotéine. B) Liste des protéines. C) Entrée de la protéine.

est donc possible de traiter ce type de séquence de la même façon qu'une protéine « simple ». Cela dit, la gestion des polyprotéines implique le traitement des séquences de plus de deux mille acides aminés avec plusieurs fonctions. La complexité qui est sous-tendue a nécessité une adaptation dans le mode de gestion de ces entrées, afin de limiter le nombre de données à manipuler tant lors du processus d'annotation que lors de leur consultation. A chaque polyprotéine est assigné un numéro VaZyNum pour lequel sont renseignés les informations générales du virus tels que : les identifiants GenBank, GenPep, le suivi taxonomique, le nom générique du virus, la longueur de la séquence de la polyprotéine, la date de création de l'entrée, la date de la dernière modification et éventuellement un nom d'usage courant interne au laboratoire. Puis nous découpons cette polyprotéine en protéines suivant les bornes décrites dans la littérature. Ce processsus de découpage nous permet d'éventuellement corriger des bornes erronées. Pour chaque protéine ainsi définie, nous attribuons un numéro d'accès VaZyNum_PP qui permet à la fois de relier chaque protéine à l'entrée principale de la polyprotéine de la base de données et d'éviter ainsi la perte et la redondance de l'information. A chaque protéine, nous assignons la fonction (lorsqu'elle est connue) et l'identifiant Genpep. Le processus d'assignation de modules est ensuite effectué sur chacune de ces protéines. Nous fournissons le découpage modulaire à la fois localement sur l'entrée de la protéine, mais aussi sur l'entrée de la polyprotéine (Figure 17).

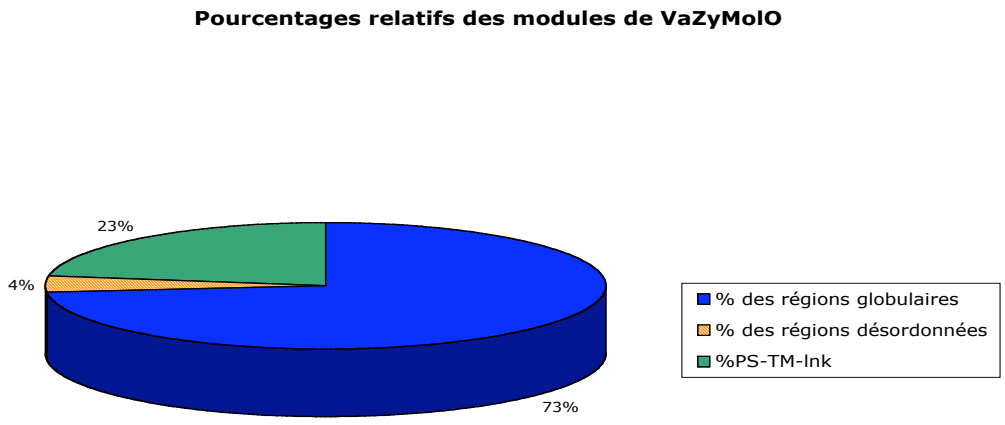
2.2. La modularité

2.2.1. Les familles de modules

Nous définissons des modules sur les séquences de protéines virales et les regroupons en familles. Donc lorsque nous parlons de « module », nous faisons référence à une partie de séquence, alors que le terme de « famille » correspond à l'ensemble des modules définis comme étant homologues sur plusieurs séquences.

Comme nous l'avons évoqué précédemment, nous assignons les modules globulaires en trois grandes catégories de modules (S, M, F). Au-delà de ces trois catégories nous nous attachons à identifier les modules correspondant aux peptides signaux (module PS), les régions hydrophobes impliquées dans les interactions protéine-protéine (modules HD) et les domaines transmembranaires (modules TM). Les régions courtes et à séquence variable entre modules sont assignées en modules de liaisons (module LNK). Les régions de plus de vingt acides aminés dépourvues ou « pauvres » en résidus hydrophobes sont assignées comme modules désordonnés (module DIS). Afin de pouvoir retrouver les régions désordonnées au sein des

protéines, la nomenclature de ces modules reprend celles des modules principaux : DISF, DISM et DISS. Ces derniers représentent 4% du nombre total de modules dans la base.



Graphique 1 : Composition de VaZyMolO.

Il est aussi à noter que ce pourcentage est inférieur aux prédictions réalisées chez la bactérie ou chez les eukaryotes, et ceci, malgré le fait que ces dernières études concernaient des segments de plus de quarante acides aminés (Dunker and Obradovic 2001). Dans les cas où il ne nous a pas été possible d'identifier à quelle catégorie se rapporte la protéine, les modules la composant sont assignés comme inconnus (module UNK). Les familles de modules PS, HD, TM, LNK ne font pas parties de nos bibliothèques de séquences modulaires. Plus de 70% des modules répertoriés dans VaZyMolO sont des candidats potentiels pour des études structurales.

2.2.2. Identification des modules

Du fait du peu de structures disponibles, la définition de modules « *ab initio* » constitue la majeure partie de notre travail. La définition modulaire repose sur une utilisation stringente du PSI-BLAST et du BLAST, de la recherche de motifs, de l'utilisation extensive des programmes de prédiction de structures secondaires, de prédiction des régions hydrophobes et d'une analyse fine des graphes issus de la méthode HCA.

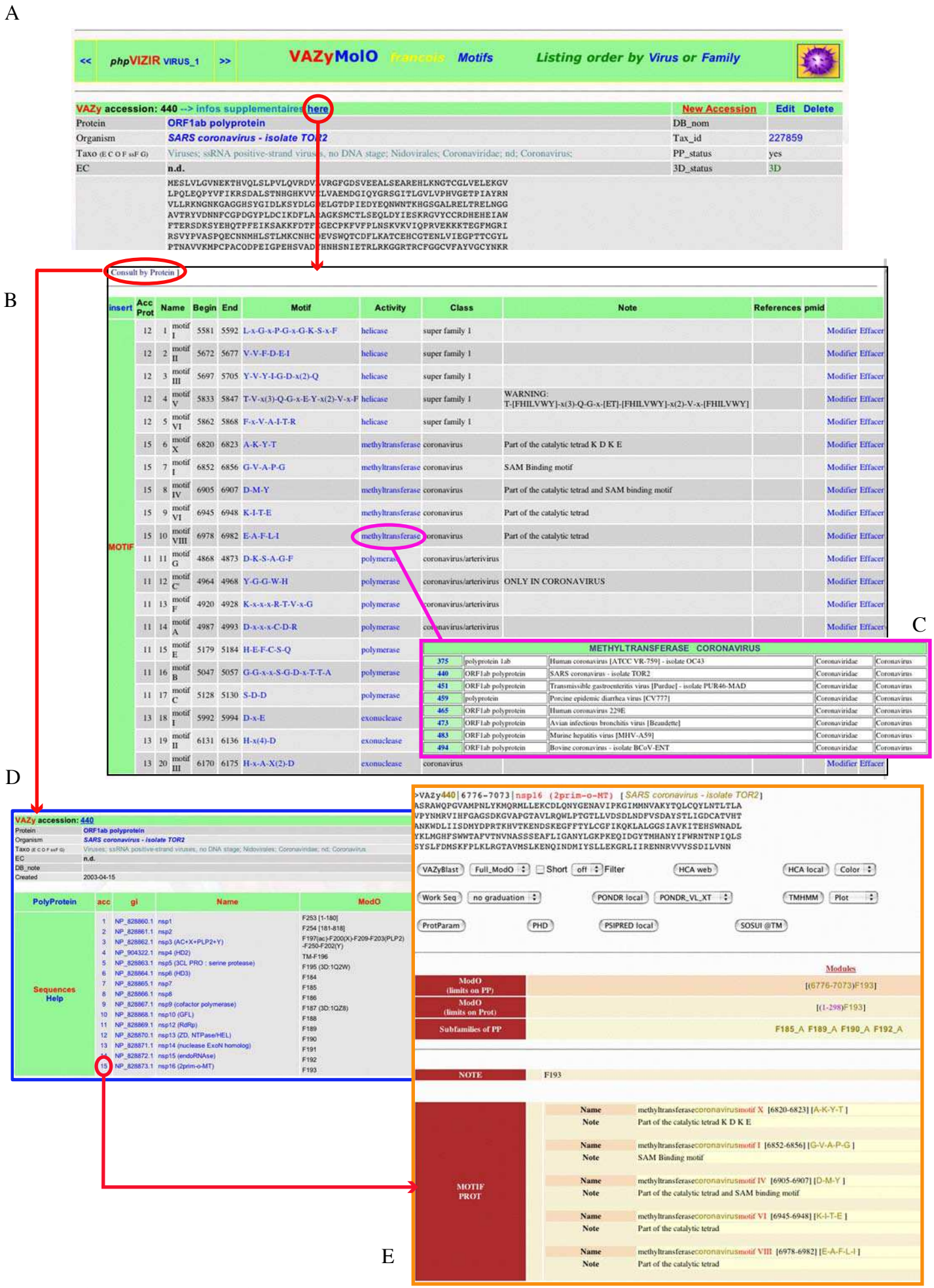


Figure 18 :Gestion des motifs. A) Informations complémentaires disponibles sur l'entrée principale. B) Archivage des motifs identifiés. C) Liste des virus possédant les motifs. D) Informations complémentaires disponibles sur l'entrée secondaire. E) Localisation du motif sur la protéine.

- Définition des modules par recherche de similarité

Le PSI-BLAST permet d'identifier un ensemble de séquences homologues et de délimiter grossièrement les contours des modules. L'annotation commence par la recherche d'un groupe de séquences homologues à une séquence inconnue à l'aide d'un PSI-BLAST interne. L'utilisation du PSI-BLAST permet d'identifier le nombre de séquences qui va constituer la famille. Nous utilisons un critère de PSI-BLAST relativement strict avec une E-value limite de 10^{-3}. Chaque séquence qui a été trouvée par PSI-BLAST est ensuite utilisée comme séquence d'interrogation afin de vérifier que toutes les séquences appartiennent bien à la même famille. Cette étape de validation permet d'éviter les faux positifs pour les petits modules. En fonction du résultat du PSI-BLAST, une délimitation grossière du module est proposée. Lorsque la famille modulaire est grande (contenant un grand nombre de séquences) nous analysons les modules de la famille par BLAST afin de définir éventuellement des sous-familles.

- Définition ou recherche de motifs

Pour chaque famille, nous recherchons, identifions et répertorions des motifs fonctionnels. Nous effectuons des alignements multiples (à l'aide MULTALIGN) des séquences d'une même famille. Ces alignements sont ensuite affinés manuellement en utilisant SEAVIEW (Galtier et al. 1996). La plupart des motifs que nous recherchons ont été décrits il y a une dizaine d'années (Gorbalenya et al. 1989; Poch et al. 1990; Gorbalenya and Koonin 1991). A cette époque, le nombre de séquences virales disponibles étaient limité. Sur la base de ces travaux, nous avons précisé ces motifs, spécifiques aux familles virales. Nous utilisons la nomenclature PROSITE. Actuellement nous disposons d'une vingtaine de motifs. La collection de motifs joue un rôle important dans l'annotation. En effet, les motifs ont été définis en tenant compte de l'évolution virale, nous permettant ainsi d'effectuer rapidement la recherche de fonctions sur les nouveaux génomes entrés. La nomenclature PROSITE permet d'être suffisamment flexible pour pouvoir retrouver des fonctions sur des séquences distantes. Les bibliothèques de motifs sont aussi utiles lors de la consultation, nous avons mis en place des liens qui permettent de retrouver instantanément la position du motif sur la séquence entière et ce sur chaque entrée déjà assignée (Figure 18).

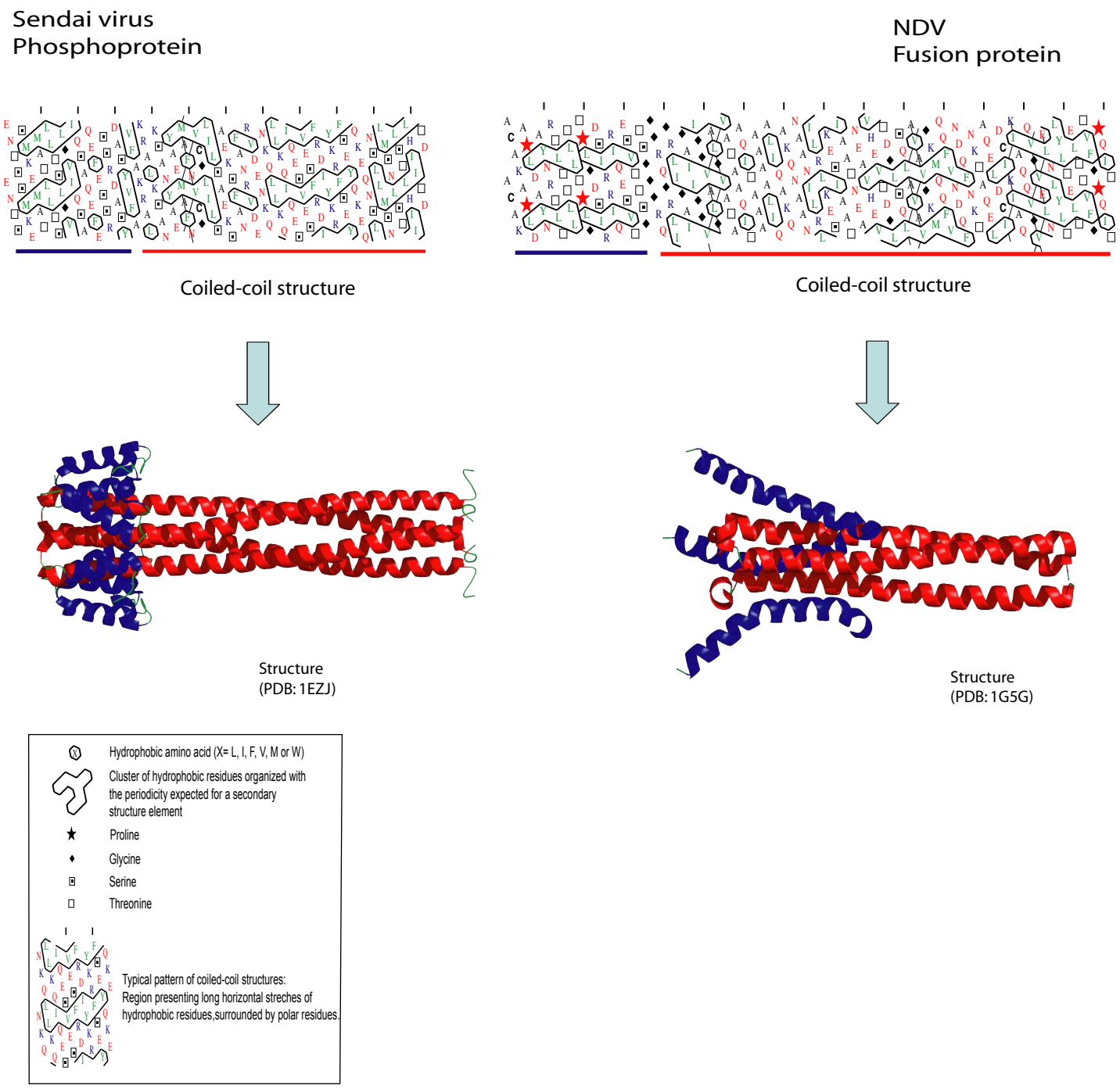


Figure 19 : Exemple de motif HCA caractéristiques d'élements structuraux.

- Outils d'analyses complémentaires

Pour identifier des régions transmembranaires et les peptides signaux nous utilisons les outils présentés dans le chapitre 2. Pour valider ces prédictions, la littérature est souvent une aide précieuse, en termes de données biochimiques. Les outils et méthodes d'étude du désordre ont été présentés dans la revue du chapitre 3.

- Délimitation modulaire par la méthode HCA et identification d'homologie distante ou de repliement commun.

Comme nous l'avons vu dans le premier chapitre, l'un des moteur du repliement des protéines est le regroupement des résidus hydrophobes. La méthode HCA permet de visualiser sous forme graphique linéaire la répartition des résidus hydrophobes. Une densité élevée des résidus hydrophobes permet de définir les domaines globulaires, les domaines transmembranaires et les peptides signaux. À l'inverse, une faible densité, voir une absence d'amas hydrophobes traduira selon le cas, soit une région de liaison entre deux domaines, soit une région désordonnée (voir revue chapitre 3). Nous utilisons cette méthode pour affiner et valider les bornes des modules que nous avons défini par la recherche de similarité de séquence et ce pour l'ensemble des séquences de la famille.

De plus, la méthode HCA est une méthode alternative pour détecter des homologies distantes et en particulier les structures de type « coiled-coil ». Ces dernières constituent une classe particulière de domaines structuraux. Ils sont souvent prédits comme des régions désordonnées par les programmes prédisant le désordre (voir revue chapitre 2). Les structures « Coiled-coil » sont de longues hélices α qui se caractérisent au niveau de la séquence par une faible complexité, accompagnée par une périodicité de résidus hydrophobes (périodicité en général de sept, mais elle peut être plus courte). Un motif caractéristique a aussi été proposé (Wolf et al. 1997; Alfadhli et al. 2001). Cela dit, si effectivement le motif se retrouve dans ce type de domaine, pour autant il n'est ni exclusif, ni systématique. Même s'il existe de très bons programmes (Lupas et al. 1991; Wootton 1994; Berger et al. 1995; Lupas 1996b; a; 1997; Wolf et al. 1997) pour prédire les domaines « coiled-coil », la méthode HCA reste la seule technique qui permet de détecter ces domaines. Le résultat graphique que propose le programme permet de visualiser des motifs caractéristiques de ce type d'éléments structuraux. (Figure 19).

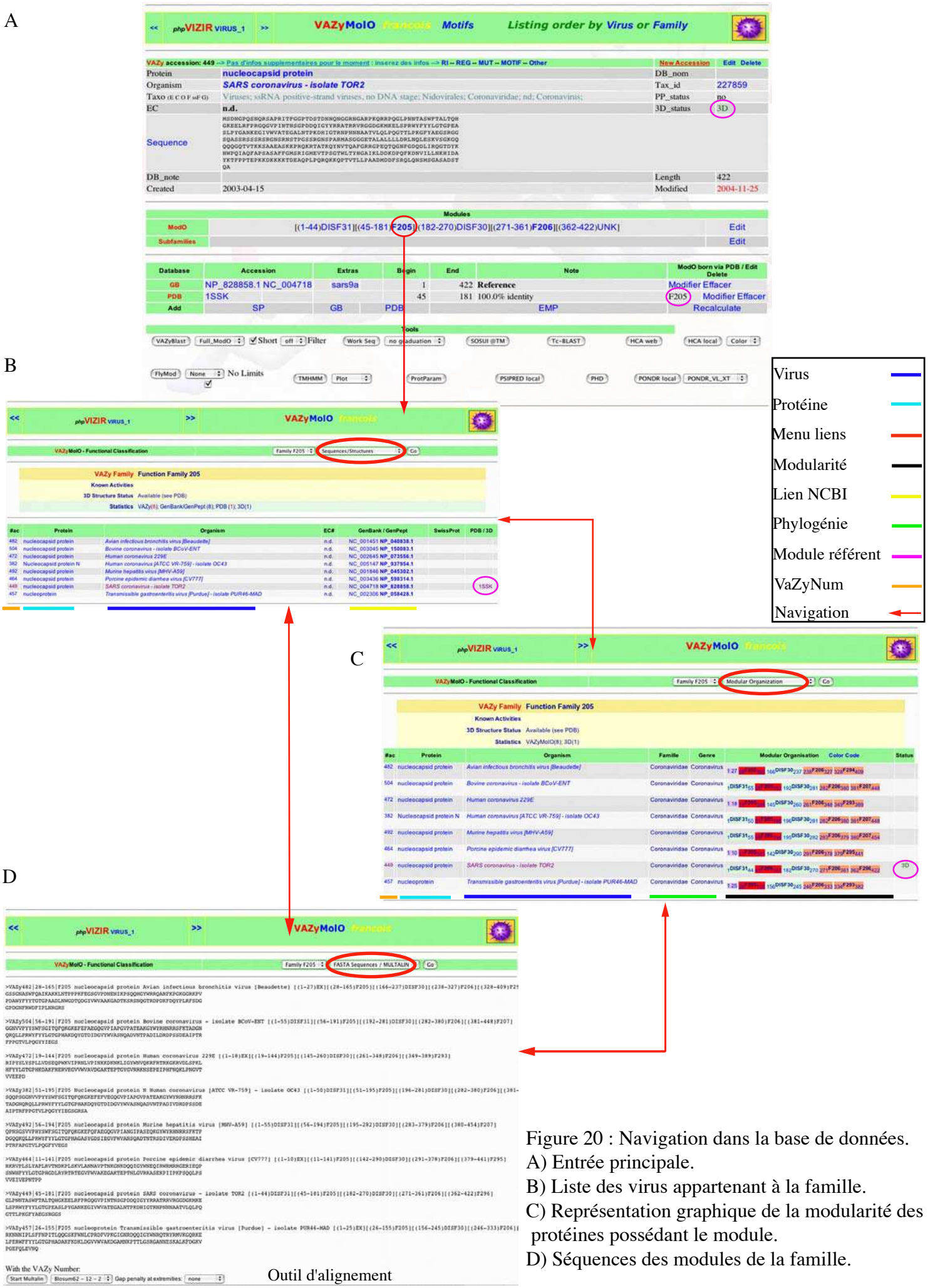
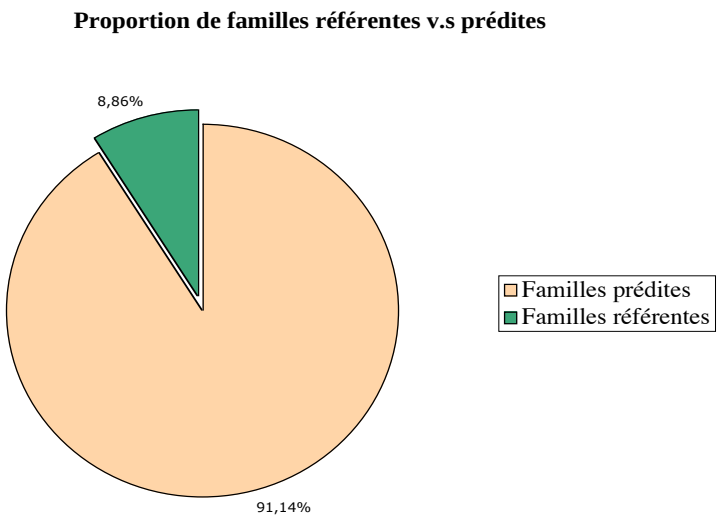


Figure 20 : Navigation dans la base de données.
A) Entrée principale.
B) Liste des virus appartenant à la famille.
C) Représentation graphique de la modularité des protéines possédant le module.
D) Séquences des modules de la famille.

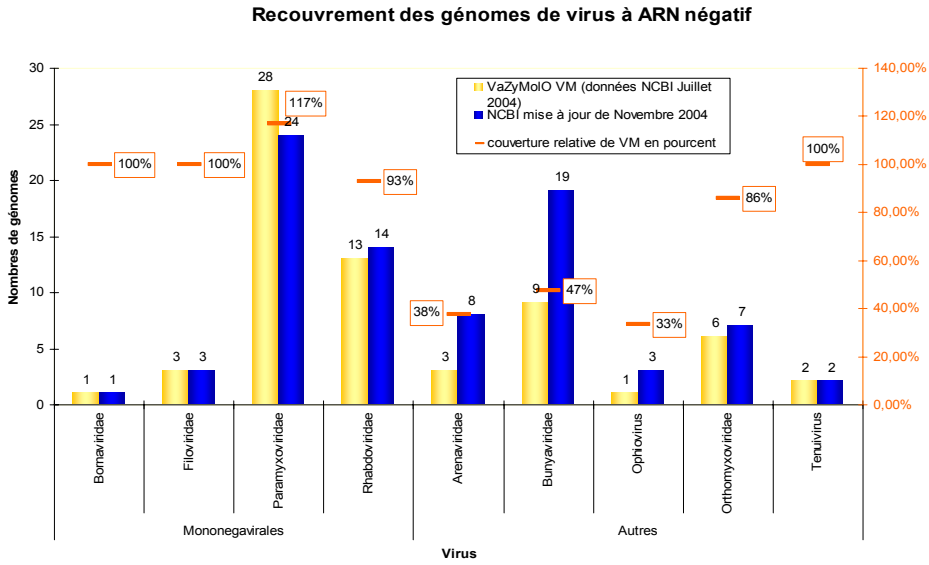
Une fois l'ensemble de ces étapes accomplies la modularité s'exprime, par exemple, de la façon suivante : [(1-29)PS][(30-504)S12][(505-527)TM], et est reportée sur chaque entrée (Figures 16 & 17). Depuis la modularité, il est possible de naviguer dans la base et d'accéder à la liste des virus appartenant à la famille (Figure 20).

2.2.3. Modules de référence

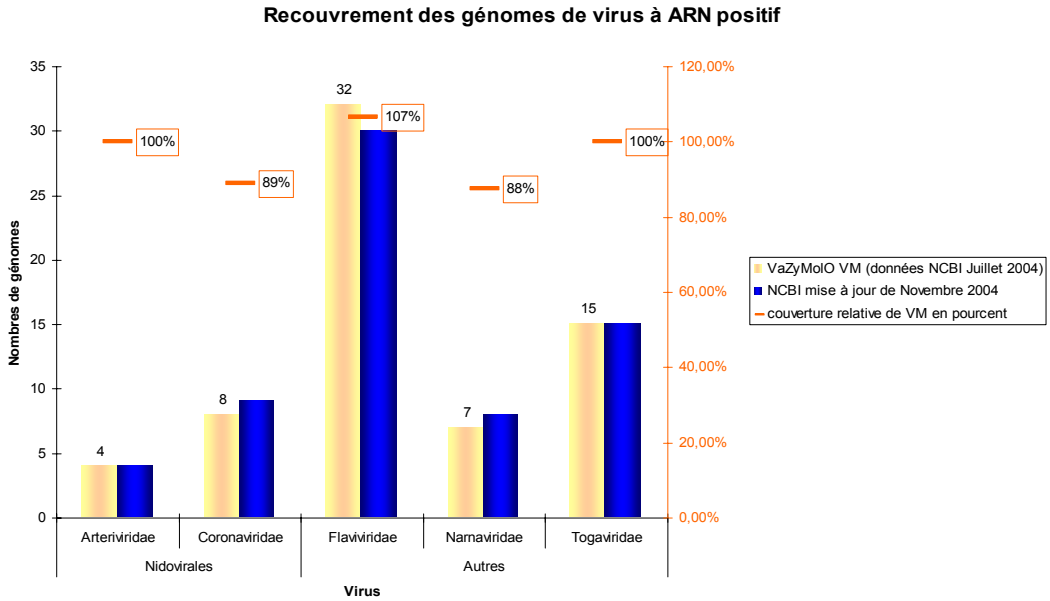
Par défaut, nous considérons comme module de référence toute construction qui a mené ou menant à la résolution d'une structure RMN ou cristallographique, d'une meilleure résolution (inférieure ou égale à 3Å). Ces modules servent de point d'ancrage pour le début de notre analyse, ils sont assignés en priorité lorsqu'ils existent, puis sont étendus aux autres séquences virales selon la méthode précédemment exposée. La proportion de modules référents est de l'ordre de 10% (Graphique 2). Bien que ce taux soit encore faible, nous commençons à avoir des exemples de prédictions modulaires qui se sont avérées fiables et ont aboutit à la résolution de structures tridimensionnelles (Johansson et al. 2003; Yang et al. 2003; Blanchard et al. 2004; Cevik et al. 2004; Egloff et al. 2004; Huang et al. 2004; Mavrakis et al. 2004; Nelson et al. 2004; Xu et al. 2004a; Xu et al. 2004b).



Graphique 2 : Proportion du nombres de familles prédites et de références.



Graphique 3 : Recouvrement des génomes de virus à ARN négatif.



Graphique 4 : Recouvrement des génomes de virus à ARN positif.

2.4. Composition actuelle de VaZyMolO

2.4.1. Les génomes

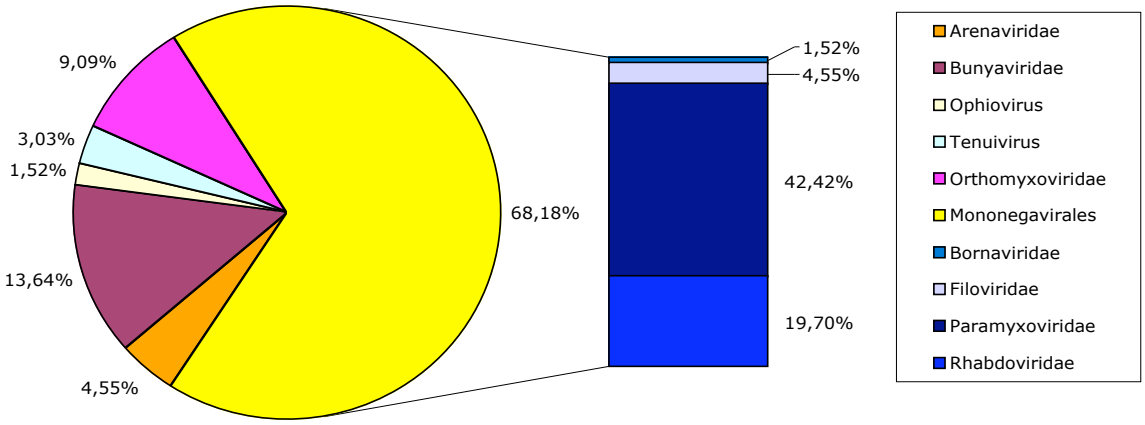
La base de données contient actuellement 67 génomes annotés de virus à ARN simple brin négatif et 66 génomes annotés de virus à ARN simple brin positif. Suite à la nouvelle mise à jour de Novembre 2004 (81 génomes disponibles pour les ARN négatifs), nous couvrons toujours l'intégralité des familles, mais plus que 81% des génomes complets disponibles au NCBI (Graphique 3).

La couverture des génomes de virus à ARN simple brin positif est plus faible (14%), et ce du fait que nous avons commencé à annoter les virus à ARN négatif, et que le nombre de virus à ARN positif séquencés est plus important.
Nous avons donc sélectionné les premiers génomes à ARN positif en fonction des sujets d'études internes au laboratoire. Nous couvrons actuellement 5 familles virales complètement annotées (Graphique 4) et nous sommes sur le point de finir l'annotation d'une nouvelle famille les *Picornaviridae*. Avec seulement six familles sur les trente-quatre disponibles au NCBI nous couvrons près du quart des génomes complets (en terme de séquences disponibles). Comme l'annotation de la famille des *Picornaviridae* est en cours, nous ne la prenons pas en compte dans les statistiques de la base de données. Ces statistiques montrées dans le graphe sont donc le reflet de la base au 9 septembre 2004.

<u>- Virus à ARN simple brin de polarité négative</u>

Le fait que toutes les familles de virus à ARN négatif soient traitées dans VaZyMolO, nous permet des mises à jour très rapides pour cette classe de virus. Les *Paramyxoviridae*, les *Rhabdoviridae* et les *Bunyaviridae* représentent les familles majoritaires de cette classe (Graphique 5).

Composition en pourcentage des differentes familles de virus à ARN négatif

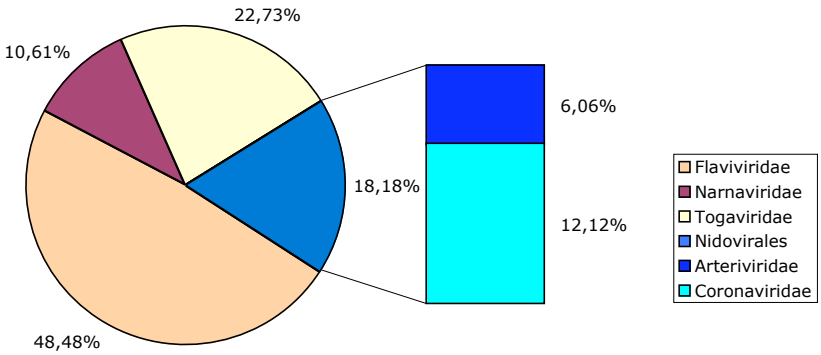


Graphique 5 : Composition des virus à ARN négatif.

<u>- Virus à ARN simple brin de polarité positive</u>

Nous disposons de 14% de la totalité des génomes complets annotés. Les *Flaviviridae* représentent la famille majoritaire de cette classe (Graphique 6).

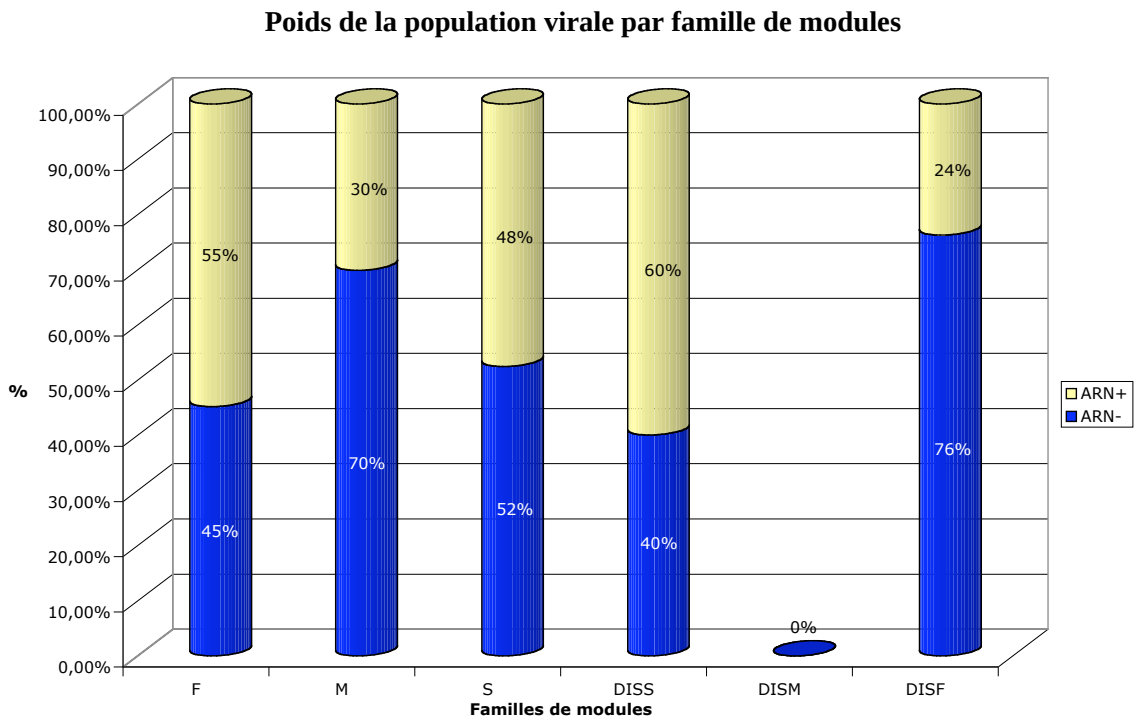
Composition en pourcentage des différentes familles de virus à ARN positif



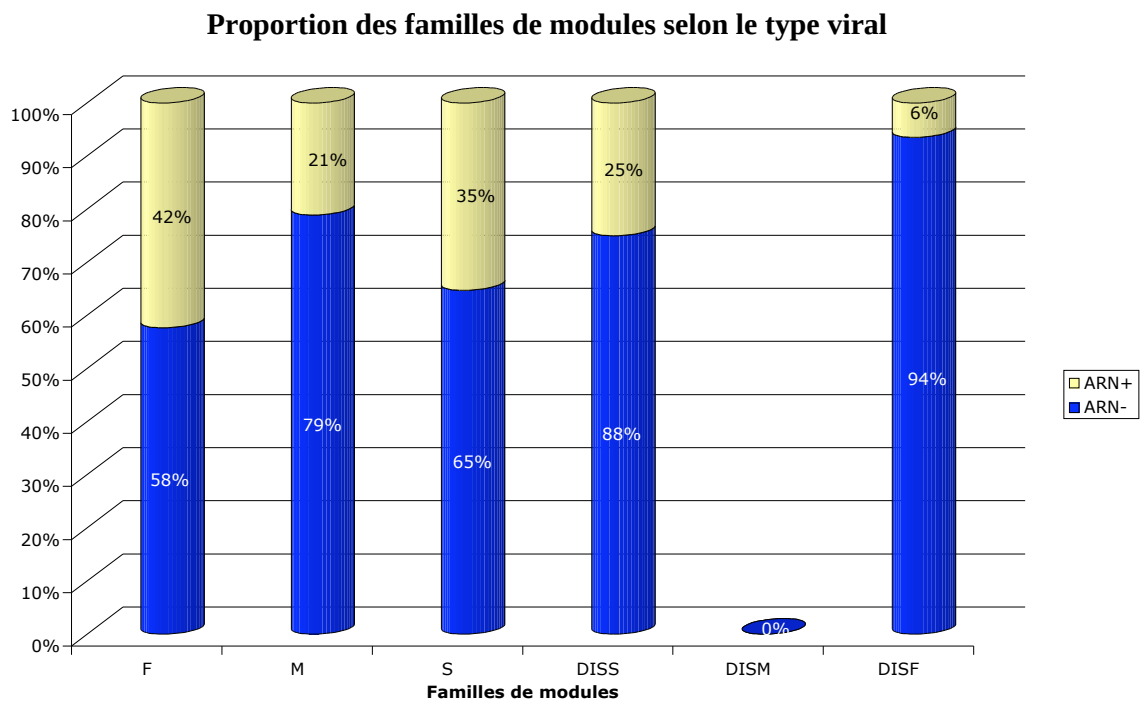
Graphique 6 : Composition des virus à ARN positif.

L'analyse de l'origine des modules de VaZyMolO montre que nos bibliothèques de modules sont encore majoritairement dominées par un nombre de familles issues de virus à ARN simple brin négatif. Cependant il est à noter que les familles F et DISS issues des virus à ARN

négatif, bien que plus nombreuses, possèdent moins de représentants (virus) par famille (Graphiques 7& 8).



Graphique 7 : Poids relatif des familles de virus.



Graphique 8 : Proportion des familles de modules.

A

B

C

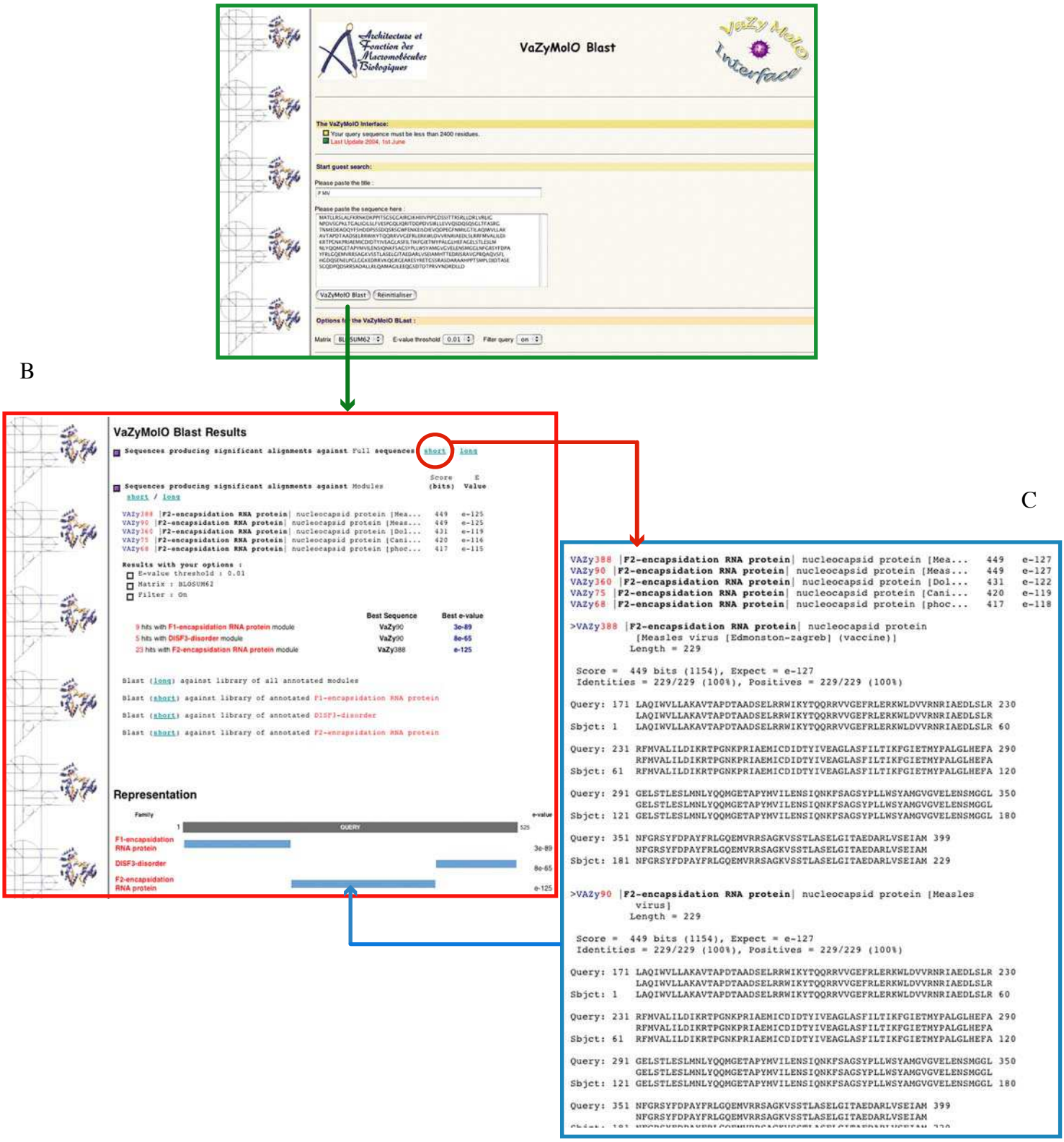


Figure 21 : Présentation de l'interface publique VaZyMolO. A) Serveur BLAST B) Résultats graphiques proposant la modularité et les liens vers les alignements. C) Alignements de la séquence requête contre nos modules.

2.4.2. Les bibliothèques de séquences de VaZyMolO et de bases annexes

Nous maintenons des banques (au format Fasta) de séquences complètes et issues du découpage modulaire, de toutes les protéines de la base. Ce sont ces bibliothèques qui sont accessibles sur le site public. De plus, nous effectuons une mise à jour régulière des bases annexes permettant l'identification modulaire, à savoir une base des séquences de la PDB et une base des séquences de MEROPS.

3. L'interface de VaZyMolO : le serveur public.

L'interface de VaZyMolO se compose principalement d'un serveur BLAST qui est mis à disposition à l'adresse suivante http://www.vazymolo.org. Il permet à chaque utilisateur de comparer la séquence de protéine « requête » contre nos bibliothèques de séquences complètes et de modules. Les résultats apparaissent sous la forme graphique et sous la forme classique de BLAST (Figure 21). Nous avons mis aussi à disposition la classification des virus disponibles dans VaZyMolO. Nous fournissons des liens vers les séquences des génomes et des protéines disponibles au NCBI, d'où proviennent originellement les séquences, ainsi que vers l'information taxonomique de l'ICTV. Nous proposons la liste complète des structures avec leurs codes PDB, qui nous ont aidés à définir les modules de références. Pour chaque entrée PDB, nous fournissons un lien vers une fiche détaillée de la structure. Nous proposons aussi à chaque utilisateur d'analyser la structure linéairement à l'aide d'ENDscript. L'article suivant présente ce travail et reprend succinctement la procédure d'annotation. L'interface de VaZyMolO est un prototype d'un futur site web qui devrait être plus complet, et qui devrait tendre vers une description plus complète pour chaque entrée.

ARTICLE 2

VaZyMolO : a tool to define and classify modularity in viral proteins.

F. Ferron & C. Rancurel, S. Longhi, C. Cambillau, B. Henrissat and B. Canard

Article accepté dans Journal of General Virology

VaZyMoLO: a tool to define and classify modularity in viral proteins

François Ferron[α] & Corinne Rancurel[α], Sonia Longhi, Christian Cambillau, Bernard Henrissat and Bruno Canard*.

Architecture et Fonction des Macromolécules Biologiques, UMR 6098, CNRS and Universités Aix-Marseille I and II, ESIL, 163, Avenue de Luminy, Case 925, F-13288 Marseille Cedex 9, France

[α] These authors have equally contributed to this work

* To whom correspondence should be sent

ESIL-CNRS-AFMB Case 925
163, avenue de Luminy, 13288 Marseille Cedex 09, France
E-mail bruno.canard@afmb.cnrs-mrs.fr
Tel: (33) 4 91 82 86 44
Fax: (33) 4 91 82 86 46

E-mail: bruno.canard@afmb.cnrs-mrs.fr

Running title: Modularity in viral proteins

Total number of words in the text: 2176

Total number of words in the summary: 150

2 Figures and 1 Table

Summary

Viral structural genomics projects aim at unveiling the function of unknown viral proteins by employing high-throughput approaches to determine their three-dimensional structure and to identify their function through fold-homology studies. We have developed the "viral enzyme module localisation" (VaZyMolO) tool, which aims at defining within viral proteins modules that might be expressed in a soluble and functionally active form, thereby identifying candidates for crystallisation studies. VaZyMolO includes 114 complete viral genome sequences of both negative and positive-sense, single-stranded RNA viruses available from NCBI. In VaZyMolO, a module is defined as a structural and/or functional unit. Modules are first identified by homology search and then validated by the convergence of results from sequence composition analysis, motif search, transmembrane region search, and domain definitions, as found in the literature. The public interface of VaZyMolO, which is accessible from http://www.vazymolo.org, allows comparison of a query sequence to all VaZyMolO modules of known function.

Keywords

Modularity; structural genomics; sequence analysis; virus; database.

INTRODUCTION

The advent of the genomic era implies that biologists are now confronted with vast quantities of raw sequence data. The number of available viral genome sequences at the NCBI has increased by 7,1% between October, 2003 and March, 2004. The last release is composed of a total set of 32308 viral proteins, of which 7472 proteins have been biochemically assessed, 22099 have a postulated function, and 2737 are classified as unknown. Such large volumes of data require the development of tools capable of distilling this information, thereby aiding scientists to devise a rational approach to the study of viral proteins. Viral structural genomics projects attempt to assign functions to uncharacterised proteins, by solving their structures and identifying function through fold homology. High-throughput structure determination combines computer-based analysis of proteins, automated expression and purification of gene products, and determination of their three-dimensional structure. One of the bottlenecks in this integrative approach is the production of pure soluble proteins suitable for crystallogenesis. We have previously reported that many viral proteins have a modular organisation, containing regions (hydrophobic or disordered) that are often not compatible with the crystallisation process (Ferron et al., 2002, Karlin et al., 2003). To increase the chance of producing protein domains suitable for crystallisation, we have developed the "viral enzyme module localisation" (VaZyMolO) tool which serves to define and classify viral protein modularity.

VaZyMolO enables the handling of viral sequences at the protein level in order to define their modularity. Sequence analysis is made possible by implementation of softwares such as BLASTp (Altschul et al., 1997), multalin (Corpet, 1988), and HCA (Callebaut et al., 1997). The two main pillars of VaZyMolO are the «protein sequence motif» and the «protein domain» definition. We define a "protein sequence motif" as a set of conserved amino acids

located within a short distance from one another that are both important for function and structure. A “protein domain” is a structurally compact, autonomously folding unit that forms a stable structure and shows a certain level of evolutionary conservation. In VaZyMolO, a "module" is defined as a structural and/or functional unit, which may contain one or several protein domains. VaZyMolO organises information about modularity on viral ORFs from complete genome sequences derived from GenBank (Benson et al., 2002, Pruitt et al., 2003). We focused on single-stranded (both negative and positive-sense) RNA viruses. We used an approach derived from that used by Coutinho & al. (Couthino & Henrissat, 1999a, Couthino & Henrissat, 1999b) to construct the Carbohydrate-active enzymes modular organisation database. In VaZyMolO modules rely on structure definition and sequence similarity, and also on sequence properties and biological data (e.g. membrane anchor, protein-protein interactions, solubility…). Moreover, our classification system allows us to overcome divergence due to both orthologous and paralogous origin of sequences, as long as they have significant sequence similarity. Those modules whose structures are known or that display high solubility, serve as a reference. To allow comparison between viruses, we take into account basic taxonomic information. We have also developed a module-function classification. The identification of conserved or missing modules is a valuable tool in the comparison of different virus genomes, helping us to derive information about the viral life cycle. The final information are stored in a library of modules, which can be interrogated using the VaZyMolO interface, *via* a BLAST engine.

CONSTRUCTION AND ORGANISATION OF VaZyMolO

Three layers in VaZyMolO

Virions are organised into three layers: surface proteins, matrix proteins, and non-structural proteins. The organisation of VaZyMolO has been directly inspired by this organisation and is therefore organised into three layers reflecting surface (layer S), matrix (layer M), and non-structural proteins (layer F). This first classification is a way to detect and highlight the common modularity between two proteins belonging to different layers.

Global annotation procedure

Virus sequences and basic information are collected from complete viral genome sequences deposited at the NCBI. Complete genomes are identified from the 'Viruses.ids' file available from the NCBI ftp site (ftp://ftp.ncbi.nih.gov/genomes/IDS/Viruses.ids). Each virus file is downloaded and parsed by a semi-automatic data-processing program. Thus, each entry in the database contains multiple information, such as NCBI accession number, GenBank Identifier (GI number), taxon number, virus name, taxonomy, product name, gene name, sequence length, and fasta sequence (Fig. 1B). The use of complete genomes avoids redundancy in these data. Sometimes however, files do not refer to proteins that are not coded in frame, such as when there is a frame shift (e.g. V protein of measles virus, NC_001498). To bridge this gap, such protein sequences are manually processed.

A library of full fasta sequences (full library) is then built from VaZyMolO. Analysis of related proteins is based on sequence similarities using BLASTp. A two-step clustering-bounding procedure is followed. First, each protein sequence is compared with the full library. The conservation of at least one region between sequences leads to an initial classification into subgroups. Then, the degree of similarity between the query and target sequences is analysed from the BLAST results. This procedure yields a first classification, leading to a kernel of families based on the strongest similarity.

The modular annotation process for each sequence includes an analysis procedure that is detailed below and summarised in Fig. 1A.

After cross-validating the information as described in Fig. 1A, modules of sequences are annotated. At this stage, a modular library of fasta sequences is built into the ModulO library.

Protein structure leads to module definition

Structural data are the basis of our module definition. Whenever possible, we have included structural information on viral proteins. This information is extended to homologous protein families using an internal BLAST procedure. The retrieval of structural information is done by searching all viral proteins against the PDB (Berman et al., 2000) using BLAST. When the result is in the twilight zone of the BLAST (i.e. according to our criteria, when the E-value is $>10^{-3}$) we consider the candidate as a distant protein and we perform threading analysis using a combination of 3D-PSSM (Kelley et al., 2000), mGenthreader (Jones, 1999) and InBGU (Fischer, 2000).

To be considered further, results from the different threading analyses should converge with high scores (according to the criteria of the selected program). The hit is then analysed. It should present the same function and key motif residues as the query. We perform a secondary structure prediction on the query using Predict Protein (Rost, 1996), in order to check for structural compatibility with the hit. A protein region is defined as a module only after this cross-validation procedure has been completed.

Disorder, globularity and transmembrane region search leads to module definition

In order to producemodules suitable for crystallisation, we attempt to precisely defineprotein regions that may contain hydrophobic (peptide signal, hydrophobic domain, transmembrane), or natively disordered (Uversky et al., 2000) patterns. In the absence of three-dimensional

data, we perform a systematic sequence analysis, to define globular and disordered regions. Disordered regions are defined by combining the results from the analysis of the mean hydrophobicity/mean charge ratio (Uversky et al., 2000), as well as from PONDR® (Iakoucheva et al., 2001) and DISEMBL (Linding et al., 2003a). We use hydrophobic cluster analysis (HCA)(Callebaut et al., 1997) to refine the boundaries of the modules and to identify linker regions. To define globularity we combine two approaches: HCA and GLOBplot (Linding et al., 2003b). HCA plots give patterns reflecting structural elements, such as coiled-coils, for example. It is known that structural homology may not be reflected in terms of sequence similarity. For this reason, each module for which a 3D structure is known is analysed using HCA in order to define the corresponding HCA pattern. These patterns allow grouping of solved or predicted distantly related modules. The power of HCA in deciphering structural homology in the absence of significant sequence similarity is well illustrated in the case of the P multimerization domain (PMD) of Sendai and measles viruses (Fig. 2). The structure of Sendai virus PMD has been solved by X-ray crystallography and consists of a coiled-coil (Tarbouriech et al., 2000). The sequence similarity between the Sendai virus PMD and the corresponding region in measles virus P (aa 304-375) is 11%, which is not high enough to be detected by Psi-Blast. However, this measles virus P region exhibits a HCA profile similar to that of the Sendai virus PMD, thus designating it as a promising candidate for crystallographic studies. Indeed, this region turned out to be expressed in a soluble form in *E. coli*, with purification yields suitable for crystallographic studies.

Finally, transmembrane regions are predicted by both TMHMM (Krogh et al., 2001) and HCA.

Functional annotation and module validation

We have developed a simple functional classification to assign proteins to broad functional classes that reflect typical viral processes. So far we have defined the following classes: structural proteins, proteases, helicases, replicases, and capping enzymes.

The different homologous protein families have been manually assigned to these classes and given a short functional description. We start with the original NCBI annotations in the "NC_xxx" files to assign the protein to a functional group. Whenever possible, we correlate the findings of this "in-house" procedure with experimental data retrieved by literature search using "ENTREZ" (http://www.ncbi.nlm.nih.gov/Entrez/index.html). Moreover, experts working on *Paramyxoviridae* within a collaborative network in which we are involved (http://virrnapoldrugtarget.univ-lyon1.fr/jdc_publicHomePage.htm) contribute to functional annotation of viral proteins. We strongly encourage virologists to provide us with functional data which will greatly help us in defining module boundaries. These data can be deposited by completing a form that is available on-line (http://afmb.cnrs-mrs.fr/stgen/vazymolo.html).

Virus taxonomy

All modules defined in VaZyMolO are related to taxon and virus name. This allows assessment of viral phylogenetic distribution of each module. We have used the nomenclature used in The International Committee on Taxonomy of Viruses (ICTV) (http://ictvdb.mirror.ac.cn/index.htm) to name species, genera and subfamilies of each virus entry. The search by virus name is facilitated by a list of standard virus names.

The VaZyMolO interface

The VaZyMolO modular assignment is accessible on-line through a web interface (http://www.vazymolo.org). The VaZyMolO interface lists the number of complete genomes in the current release as well as taxonomic and structural information (Table1). It contains a link to a listing of the complete genome sequences of viruses sorted by virus name, and family. The protein module library as defined by VaZyMolO, can be queried by a sequence search *via* a BLAST server. In future development we are planning to integrate an interactive graphical interface allowing for each entry an easy navigation between schematic domain modularity, protein information, alignment, structure and phylogeny.

Relevance of the VaZyMolO approach in structural genomics and comparison with other databases

VaZyMolO makes use of a novel approach to define protein modularity, thus rendering it complementary and not redundant with other modular databases. Indeed, most of the modular databases are based on extensive and mainly automated annotation procedures(Bateman et al., 2000, Corpet, 1988). Conversely, VaZyMolO annotations are based on a stringent manual checking, specially concerning the boundaries of the modules, and it benefits of virologists' knowledge. As for motif definition, the fact that VaZyMolO deals only with viral protein sequences allows to overcome the problems of bias that can be found in other motif databases, and enables to derive motifs reflecting the evolution of viral proteins. The VaZyMolO interface allows the fast and easy retrieval of information on the modular organisation of a query sequence, which represents a critical step in view of structural studies. Indeed, this tool is the keystone in the selection of the best targets in the SPINE structural genomics project (http://www.ebi.ac.uk/msd-srv/msdtarget) in which our laboratory is engaged. Targets defined in this way are processed by an "in-house" high-throughput platform for expression, purification, and crystallisation. Feedback of the behaviour of each tested protein allows

biochemical validation of module boundaries. Since this structural genomics project began in 2002, VaZyMolO analysis has proven to be crucial for the structural and functional characterisation of the 2'-O-methyltransferase domain of dengue virus NS3 (Egloff et al., 2002), the X domain of measles virus phosphoprotein (Johansson et al., 2003, Karlin et al., 2003) and Nsp9 of SARS virus (Egloff et al., 2004).

CONCLUSIONS

VaZyMolO is a tool devoted to the modular description and classification of both non-structural and structural viral proteins. Viral sequences are retrieved from different plant and animal virus families. Non-redundant complete genome sequences derived from NCBI are automatically clustered into homologous protein families, following a process of pre-classification, and modules are then defined. The primary basis of this classification are structural motifs detected by a variety of complementary methods. Protein families are a rich source of information for functional and evolutionary studies. Sequence alignments of conserved regions highlight important conserved amino acids, allowing the definition of new motifs within proteins. VaZyMolO is presently tailored to studies of Mononegavirales, coronaviruses and flaviviruses. It will be updated with each new GenBank release, and we are currently incorporating other animal virus families. Functional annotation should benefit from contributions and feedback from collaborating experts in the field, via an on-line form. This comprehensive analysis facilitates the identification of many previously undetected module families of unknown function, thereby paving the way for their structural and functional analysis.

Acknowledgements

We want to thank Pedro Couthino, Eric Blanc, Emeline Deleury for their support and contribution. We also thank Denis Gerlier for useful discussion. We are grateful to David Bhella and Alexander E. Gorbalenya for critical reading of the manuscript. This work was funded by the European Commission as "SPINE" (contract no. QLG2-CT-2002-00988) and "VirRNApoldrugtarget" ("Towards the design of new potent antiviral drug: structures-function analysis of *Paramyxoviridae* RNApolymerase", contract no. QLK2-CT2001-01225) under the specific RTD programme "Quality of Life and Management of Living Resources". It does not necessarily reflect its views and in no way anticipates the Commission's future policy in this area.

Figure Legends

Fig. 1: (A)General work scheme of the annotation process of VaZyMolO.. (B) Snapshot of the annotator interface of VaZyMolO. An example of Newcastle disease virus fusion protein entry (NP_071469.1).

Fig. 2: The multimerization domain (PMD) of Sendai virus phosphoprotein (P) is a coiled-coil structure (Tarbouriech et al., 2000). The corresponding HCA pattern is interpreted as long hydrophobic stretches (underlined region). We have found the same kind of pattern in Measles virus P, despite no significant sequence similarity. In VaZyMolO, the two modules are thus considered as belonging to the same class. The measles virus PMD module was indeed expressed and purified from the soluble fraction of *E. coli*. Its crystallization is in progress.

References

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**, 3389-402.

Bateman, A., Birney, E., Durbin, R., Eddy, S. R., Howe, K. L. & Sonnhammer, E. L. (2000). The Pfam protein families database. *Nucleic Acids Res* **28**, 263-6.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., Rapp, B. A. & Wheeler, D. L. (2002). GenBank. *Nucleic Acids Res* **30**, 17-20.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Res* **28**, 235-42.

Callebaut, I., Labesse, G., Durand, P., Poupon, A., Canard, L., Chomilier, J., Henrissat, B. & Mornon, J. P. (1997). Deciphering protein sequence information through hydrophobic cluster analysis (HCA): current status and perspectives. *Cell Mol Life Sci* **53**, 621-45.

Corpet, F. (1988). Multiple sequence alignment with hierarchical clustering. *Nucleic Acids Res* **16**, 10881-90.

Couthino, P. & Henrissat, B. (1999a). Carbohydrate-active enzyme: an integrated approach. In *Recent advances in Carbohydrate Bioengineereing*, pp. 3-12. Edited by H. Gilbert, G. Davies, B. Henrissat & B. Svensson. Cambridge: The Royal Society of Chemistry.

Couthino, P. & Henrissat, B. (1999b). The modular structure of cellulases and other carbohydrate-active enzymes: an integrated database approach. In *Genetics, Biochemistry and Ecology of Cellulose Degradation*, pp. 15-23. Edited by K. Ohmiya, K. Hayashi, K. Sakka, Y. Kobayashi, S. Karita & T. Kimura. Tokyo: Uni Publishers Co.

Egloff, M. P., Benarroch, D., Selisko, B., Romette, J. L. & Canard, B. (2002). An RNA cap (nucleoside-2'-O-)-methyltransferase in the flavivirus RNA polymerase NS5: crystal structure and functional characterization. *Embo J* **21**, 2757-68.

Egloff, M. P., Ferron, F., Campanacci, V., Longhi, S., Rancurel, C., Dutartre, H., Snijder, E. J., Gorbalenya, A. E., Cambillau, C. & Canard, B. (2004). The severe acute respiratory syndrome-coronavirus replicative protein nsp9 is a single-stranded RNA-binding subunit unique in the RNA virus world. *Proc Natl Acad Sci U S A* **101**, 3792-6.

Ferron, F., Longhi, S., Henrissat, B. & Canard, B. (2002). Viral RNA-polymerases -- a predicted 2'-O-ribose methyltransferase domain shared by all Mononegavirales. *Trends Biochem Sci* **27**, 222-4.

Fischer, D. (2000). Hybrid fold recognition: combining sequence derived properties with evolutionary information. *Pac Symp Biocomput*, 119-30.

Iakoucheva, L. M., Kimzey, A. L., Masselon, C. D., Bruce, J. E., Garner, E. C., Brown, C. J., Dunker, A. K., Smith, R. D. & Ackerman, E. J. (2001). Identification of intrinsic order and disorder in the DNA repair protein XPA. *Protein Sci* **10**, 560-71.

Johansson, K., Bourhis, J. M., Campanacci, V., Cambillau, C., Canard, B. & Longhi, S. (2003). Crystal structure of the measles virus phosphoprotein domain responsible for the induced folding of the C-terminal domain of the nucleoprotein. *J Biol Chem* **278**, 44567-73.

Jones, D. T. (1999). GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. *J Mol Biol* **287**, 797-815.

Karlin, D., Ferron, F., Canard, B. & Longhi, S. (2003). Structural disorder and modular organization in Paramyxovirinae N and P. *J Gen Virol* **84**, 3239-52.

Kelley, L. A., MacCallum, R. M. & Sternberg, M. J. (2000). Enhanced genome annotation using structural profiles in the program 3D- PSSM. *J Mol Biol* **299**, 499-520.

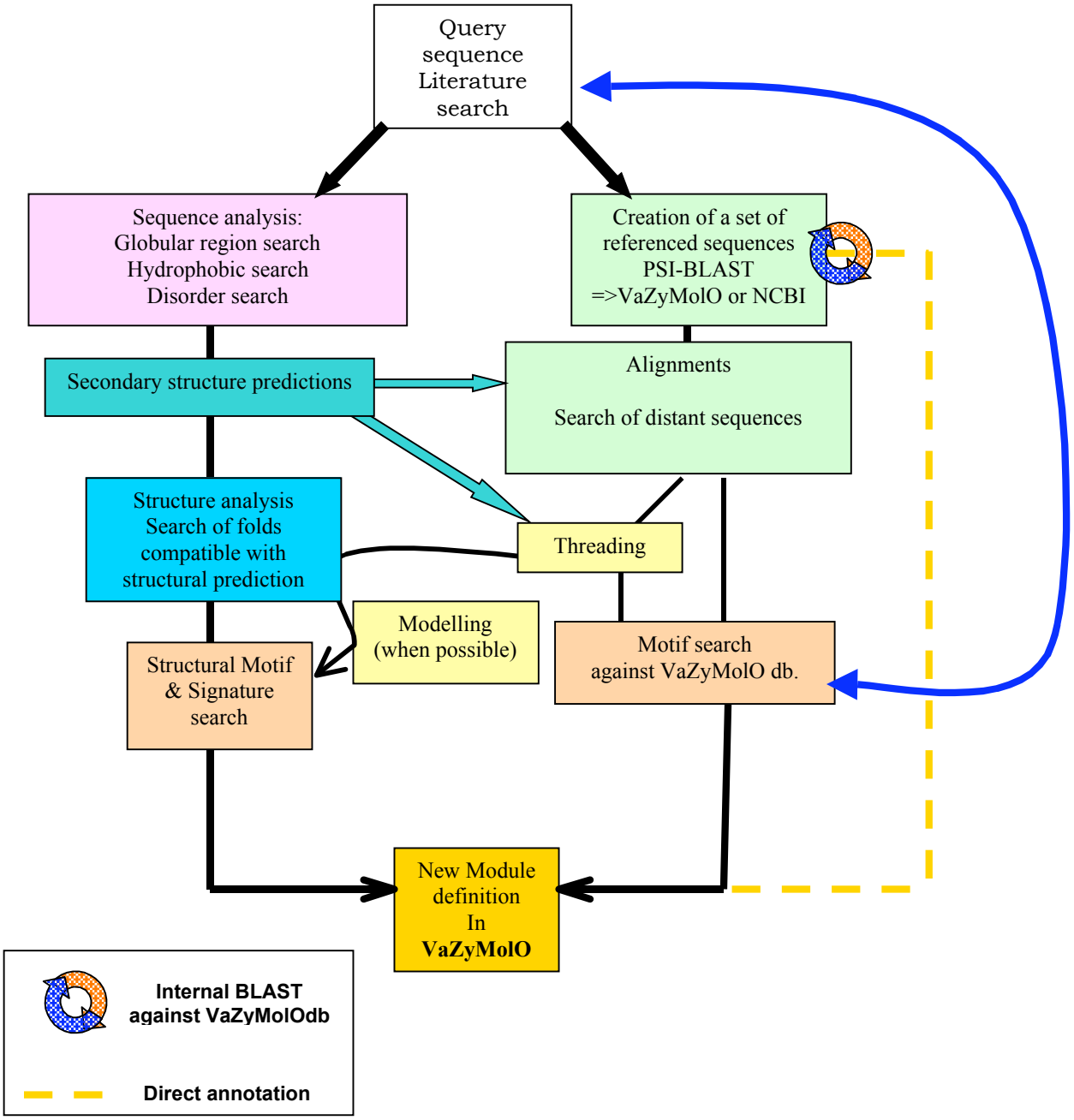
Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol* **305**, 567-80.
Linding, R., Jensen, L. J., Diella, F., Bork, P., Gibson, T. J. & Russell, R. B. (2003a). Protein disorder prediction: implications for structural proteomics. *Structure (Camb)* **11**, 1453-9.
Linding, R., Russell, R. B., Neduva, V. & Gibson, T. J. (2003b). GlobPlot: Exploring protein sequences for globularity and disorder. *Nucleic Acids Res* **31**, 3701-8.
Pruitt, K. D., Tatusova, T. & Maglott, D. R. (2003). NCBI Reference Sequence project: update and current status. *Nucleic Acids Res* **31**, 34-7.
Rost, B. (1996). PHD: predicting one-dimensional protein structure by profile-based neural networks. *Methods Enzymol* **266**, 525-39.
Tarbouriech, N., Curran, J., Ruigrok, R. W. & Burmeister, W. P. (2000). Tetrameric coiled coil domain of Sendai virus phosphoprotein. *Nat Struct Biol* **7**, 777-81.
Uversky, V. N., Gillespie, J. R. & Fink, A. L. (2000). Why are "natively unfolded" proteins unstructured under physiologic conditions? *Proteins* **41**, 415-27.

CLASS ORDER FAMILY *GENUS*	**numb**
ssRNA negative-strand viruses	**64**
Mononegavirales	**43**
Bornaviridae	**1**
Bornavirus	*1*
Filoviridae	**3**
Ebola-like viruses	*2*
Marburg-like viruses	*1*
Paramyxoviridae	**27**
Avulavirus	*3*
Henipavirus	*2*
Metapneumovirus	*1*
Morbillivirus	*5*
Pneumovirus	*3*
Respirovirus	*4*
Rubulavirus	*6*
Unclassified Paramyxovirinae	*3*
Rhabdoviridae	**12**
Cytorhabdovirus	*1*
Ephemerovirus	*1*
Lyssavirus	*2*
Novirhabdovirus	*4*
Nucleorhabdovirus	*2*
Vesiculovirus	*2*
Arenaviridae	**3**
Arenaviruses	*3*
Bunyaviridae	**9**
Hantavirus	*1*
Nairovirus	*1*
Orthobunyavirus	*2*
Phlebovirus	*1*
Tospovirus	*4*
Orthomyxoviridae	**6**
Influenzavirus A	*3*
Influenzavirus B	*3*
Ophiovirus	**1**
Tenuivirus	**2**

ssRNA positive-strand viruses, no DNA stage	**50**
Nidovirales	**12**
Arteriviridae	**4**
Arterivirus	*4*
Coronaviridae	**8**
Coronavirus	*8*
Flaviviridae	**32**
Flavivirus	*20*
Hepacivirus	*2*
Pestivirus	*6*
Unclassified GB virus	*4*
Narnaviridae	**6**
Mitovirus	*5*
Narnavirus	*1*

Table. 1.

A


Query sequence Literature search
Sequence analysis: Globular region search Hydrophobic search Disorder search
Creation of a set of referenced sequences PSI-BLAST =>VaZyMolO or NCBI
Secondary structure predictions
Alignments
Search of distant sequences
Structure analysis Search of folds compatible with structural prediction
Threading
Modelling (when possible)
Structural Motif & Signature search
Motif search against VaZyMolO db.
New Module definition In VaZyMolO
Internal BLAST against VaZyMolOdb
Direct annotation

B

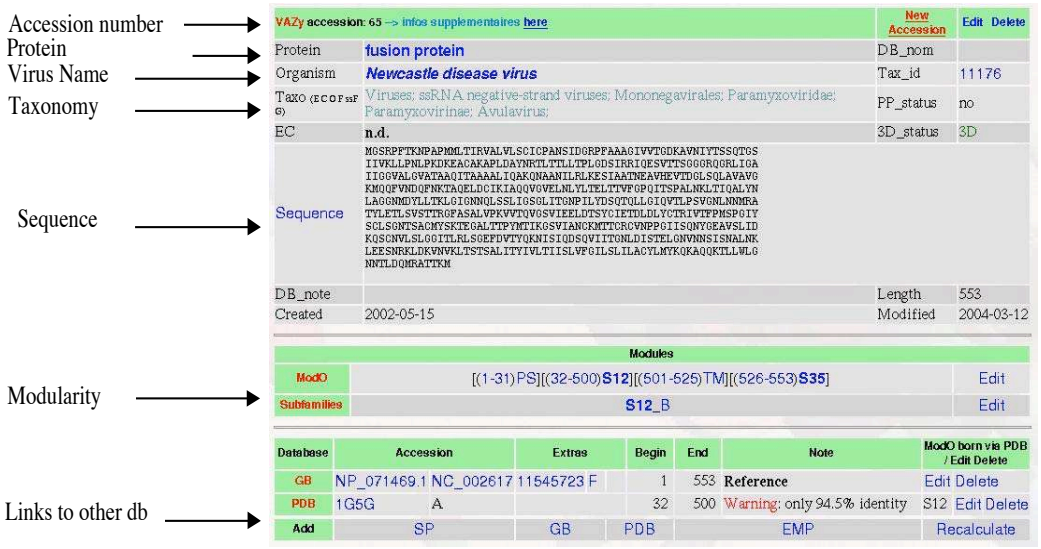

Accession number
Protein
Virus Name
Taxonomy
Sequence
Modularity
Links to other db
VAZy accession: 65 —> infos supplementaires here
New Accession
Edit Delete
Protein
fusion protein
DB_nom
Organism
Newcastle disease virus
Tax_id
11176
Taxo (E C O F ssF G)
Viruses; ssRNA negative-strand viruses; Mononegavirales; Paramyxoviridae; Paramyxovirinae; Avulavirus;
PP_status
no
EC
n.d.
3D_status
3D
Sequence
MGSRPFTKNPAPMMLTIRVALVLSCICPANSIDGRPFAAAGIVVTGDKAVNIYTSSQTGS
IIVKLLPNLPKDKEACAKAPLDAYNKTLTTLLTPLGDSIRRIQESVTTSGGGRQGRLIGA
IIGGVALGVATAAQITAAAALIQAKQNAANILRLKESIAATNEAVHEVTDGLSQLAVAVG
KMQQFVNDQFNKTAQELDCIKIAQQVGVELNLYLTELTTVFGPQITSPALNKLTIQALYN
LAGGNMDYLLTKLGIGNNQLSSLIGSGLITGNPILVDSQTQLLGIGVTLPSVGNLNNMRA
TYLETLSVSTTRGFASALVPKVVTQVGSVIEELDTSYCIETDLDLYCTRIVTFPMSPGIY
SGLSGNTSACMYSKTEGALTTPYMTIKGSVIANCKMTTCRCVNPPGIISQNYGEAVSLID
KQSCNVLSLGGITLRLSGEFDVTYQKNISIQDSQVITTGNLDISTELGNVNNSISNALNK
LEESNRKLDKVNVKLTSTSALITYIVLTIISLVFGILSLILACYLMYKQKAQQKTLLWLG
NNTLDQMRATTKM
DB_note
Length
553
Created
2002-05-15
Modified
2004-03-12
Modules
ModO
[(1-31)PS][(32-500)S12][(501-525)TM][(526-553)S35]
Edit
Subfamilies
S12_B
Edit
Database
Accession
Extras
Begin
End
Note
ModO born via PDB / Edit Delete
GB
NP_071469.1 NC_002617 11545723 F
1
553
Reference
Edit Delete
PDB
1G5G
A
32
500
Warning: only 94.5% identity
S12
Edit Delete
Add
SP
GB
PDB
EMP
Recalculate

Fig.1

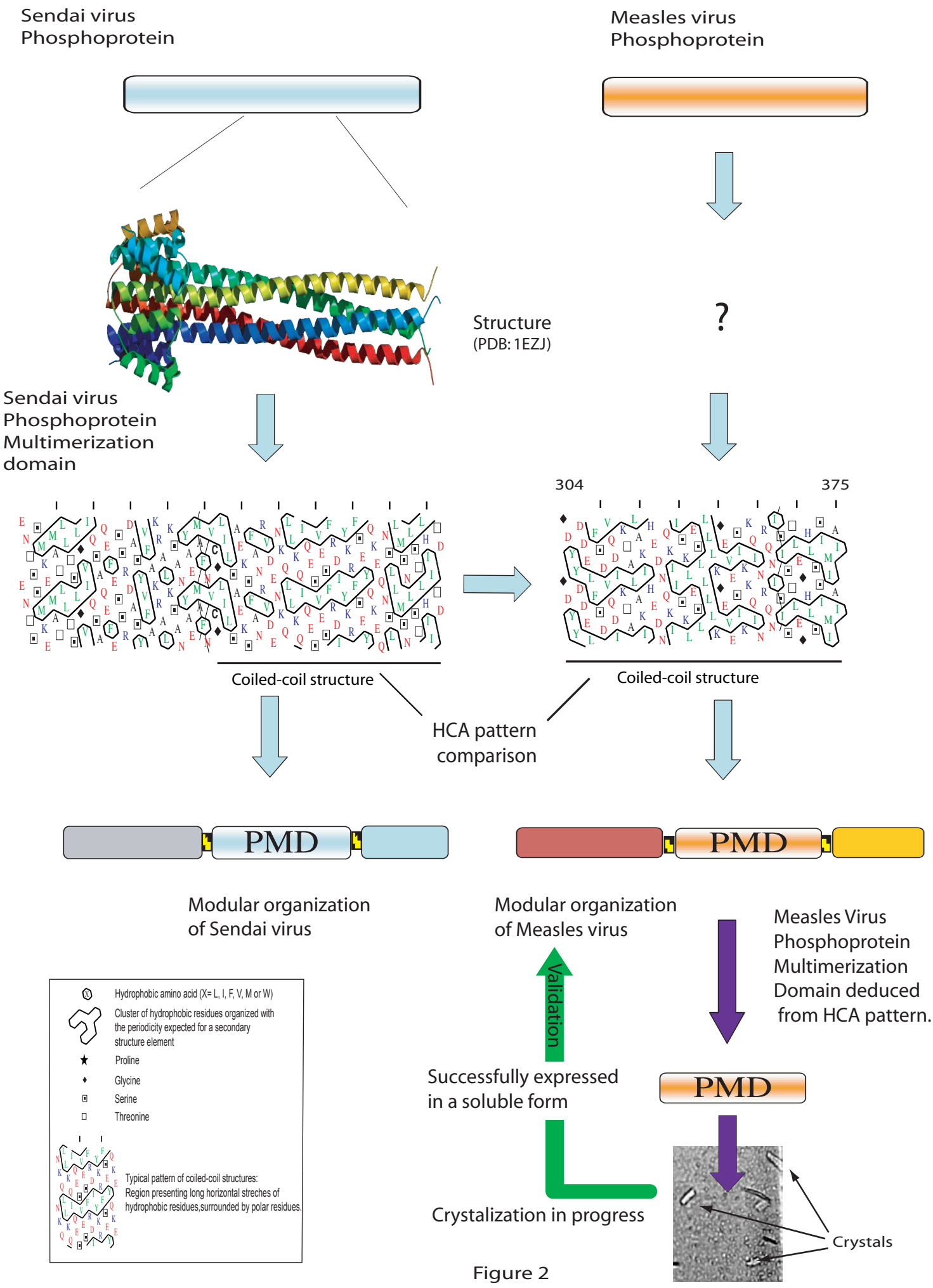
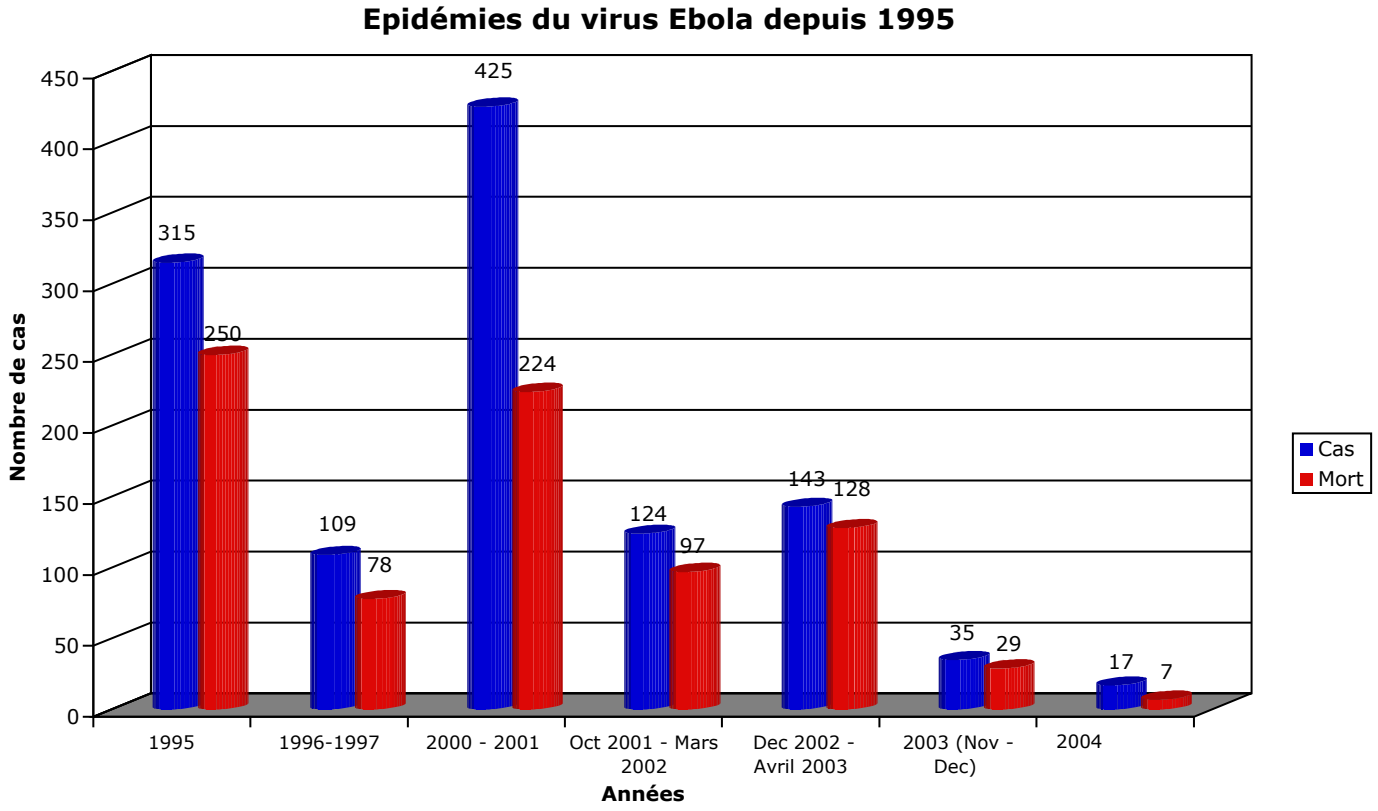

Sendai virus
Phosphoprotein
Measles virus
Phosphoprotein
Structure
(PDB: 1EZJ)
?
Sendai virus
Phosphoprotein
Multimerization
domain
304
375
Coiled-coil structure
Coiled-coil structure
HCA pattern
comparison
PMD
PMD
Modular organization
of Sendai virus
Modular organization
of Measles virus
Measles Virus
Phosphoprotein
Multimerization
Domain deduced
from HCA pattern.
Validation
Successfully expressed
in a soluble form
PMD
Crystalization in progress
Crystals
Hydrophobic amino acid (X= L, I, F, V, M or W)
Cluster of hydrophobic residues organized with the periodicity expected for a secondary structure element
Proline
Glycine
Serine
Threonine
Typical pattern of coiled-coil structures:
Region presenting long horizontal streches of hydrophobic residues,surrounded by polar residues.

Figure 2

Chapitre 5

Exemples d'application des méthodes d'analyse de séquences à l'étude des réplicases virales.



Graphique 9 : Statistique des dernières épidémies du virus Ebola.

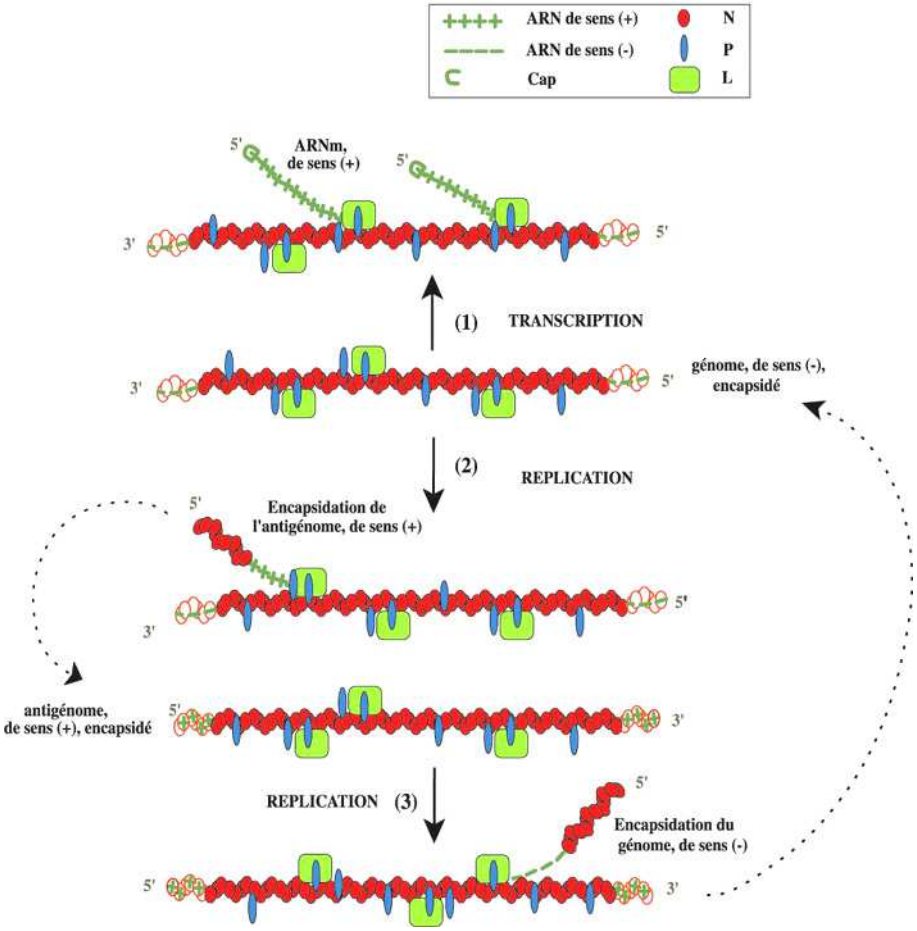


Figure 22 : Cycle réplicatif des *Paramyxoviridae*.

Dans ce chapitre, j'effectue un bref rappel concernant les virus étudiés. Je vais vous présenter dans un premier temps les virus à ARN négatif appartenant à l'ordre des Mononegavirales, puis me focaliser sur la famille des *Paramyxoviridae* et la sous-famille des *Paramyxovirinae*. J'enchaînerai brièvement sur les Hantavirus qui appartiennent à la famille des *Bunyaviridae*. Dans un deuxième temps je vous présenterai deux genres de virus à ARN positif, appartenant au Coronavirus et au GB virus.

1. Etudes des protéines du complexe réplicatif de virus à ARN négatif

1.1. Les Mononegavirales

1.1.1.Généralités

L'ordre des Mononegavirales se compose des *Paramyxoviridae*, *Bornaviridae*, *Rhabdoviridae* et *Filoviridae*. Cet ordre comprend parmi les plus virulents et contagieux pathogènes humains, comme les virus para-influenza, le virus de la rougeole (Source OMS 2004 : 614000 morts par an), le virus des oreillons, le virus de la rage, et des virus responsables de fièvres hémorragiques, comme les virus Ebola (Graphique 9), Marburg, ou les virus émergents Tioman et Hendra (Mackenzie et al. 2001; Snell 2004). Hormis les *Bornaviridae*, les virus de cet ordre ont un cycle réplicatif intégralement cytoplasmique (Lamb and Kolakofsky 2001). Il s'agit de virus enveloppés avec un génome non segmenté constitué par un ARN simple brin de polarité négative (-). L'ARN génomique sert à la fois à la synthèse des ARN messagers viraux de polarité positive (+), et à la production de brins complémentaires entiers (+) appelés antigénomes. Ces derniers sont répliqués dans leur intégralité pour donner des génomes (-). Les génomes des *Paramyxoviridae*, *Rhabdoviridae*, et *Filoviridae* ont une taille comprise entre 11 et 18 Kb et possèdent une structure commune. Le génome comprend en 3' une séquence unique renfermant le promoteur pour la transcription et la réplication suivi de 5 à 10 gènes, séparés par des séquences intergéniques conservées et servant de site d'initiation et de terminaison dans la synthèse des ARNs messagers (pour revue, voir (Kolakofsky et al. 2004)).

1.1.2. La particule virale des *Paramyxoviridae*

Les *Paramyxoviridae* sont des virus ayant généralement une forme sphérique avec un diamètre compris entre 150 et 300 nm. Leur enveloppe est constituée par une bicouche lipidique qui provient de la membrane plasmique de la cellule infectée et dans laquelle sont

insérées des glycoprotéines virales (fusion et attachement). Sous l'enveloppe, on retrouve une couche constituée par la protéine de matrice. A l'intérieur du virus on trouve la nucléocapside.

1.1.3. La structure de la nucléocapside

L'ARN viral n'est pas nu, mais il est étroitement associé à la nucléoprotéine (N) pour former une structure hélicoïdale appelée nucléocapside sur laquelle viennent se fixer la phosphoprotéine (P) et la protéine « large » (L) ou ARN polymérase ARN-dépendante. Le complexe L-P constitue le complexe de transcription-réplication.

La transcription et la réplication du génome viral ne nécessitent pas d'étape de décapsidation de la nucléocapside. Chez les *Paramyxoviridae,* il a été montré que la nucléocapside peut adopter des états morphologiques plus ou moins compacts (Heggeness et al. 1980; Bhella et al. 2002; Longhi et al. 2003). Il est raisonnable de penser que les différents états morphologiques puissent correspondre à des niveaux différents d'exposition de l'ARN viral au solvant et donc à des formes plus au moins transcriptionnellement et réplicativement actives.

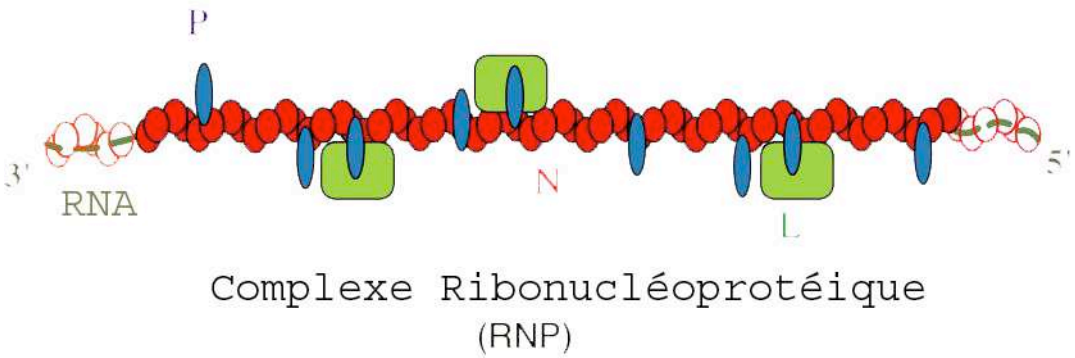


Figure 23: Complexe Ribonucléoprotéique ou nucléocapside.

1.1.4. Organisation du génome viral

Le génome des *Paramyxoviridae*, contient un nombre de gènes variable entre sept et dix selon les différents genres (Lamb and Kolakofsky 2001). Cela dit, chez les *Paramyxovirinae,* le nombre de cadres ouverts de lecture chevauchant à l'intérieur du gène P peut augmenter de un à trois le nombre de protéines finalement codées (Lamb and Kolakofsky 2001).

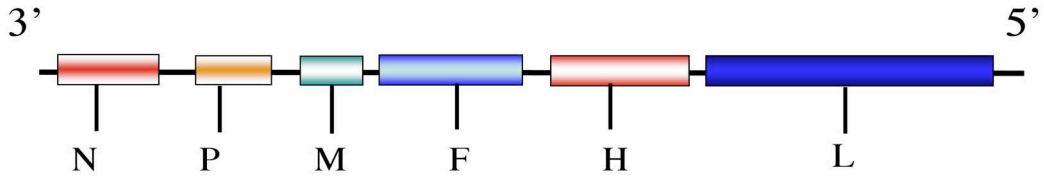


Figure 24 : Organisation du génome du virus de la rougeole.

1.1.5. Description des protéines du complexe réplicatif

-La nucléoprotéine

La nucléoprotéine à un rôle structural majeur puisque elle est responsable de l'encapsidation du génome viral. Pendant la transcription et la réplication de l'ARN viral, elle interagit avec la phosphoprotéine, ce qui permet le recrutement du complexe polymérase. Lors de l'assemblage, elle interagit avec la protéine de matrice. Suite à l'apoptose ou à la nécrose de cellules infectées, la nucléocapside est rélarguée dans le compartiment extracellulaire où la nucléoprotéine interagit avec le récepteur FcγRII (Marie et al. 2001; Marie et al. 2004) et aussi avec un récepteur encore non caractérisé (Laine et al. 2003). L'interaction avec ces deux récepteurs est impliquée dans l'immunosuppression transitoire provoquée par le virus de la rougeole (Laine et al. 2003).

-La phosphoprotéine

La protéine P est un co-facteur de la polymérase (L). Cette protéine joue probablement un rôle clé dans le mécanisme de réplication-transcription. Cette protéine se lie à la fois à N et à L. Il a été montré que P empêche l'auto-assemblage de N en formant un complexe (N-P). La nucléoprotéine ainsi complexée est maintenue transitoirement soluble jusqu'à ce que l'encapsidation de l'ARN commence. Chez le virus de la rougeole les sites de liaison à L ont été cartographiés (Cevik et al. 2004).

-L'ARN polymérase ARN-dépendante, ou protéine Large, L

De toutes les protéines des *Mononegavirales*, la protéine L est la protéine dont la séquence est la mieux conservée, mais aussi la protéine la moins bien caractérisée. Cette enzyme de plus de 250 KDa, faiblement exprimée à l'état naturel, n'a jamais pu être purifiée, ni étudiée *in vitro*. Les activités de cette protéine proviennent de preuves indirectes. La similarité de séquence

des L n'est significative qu'à l'intérieur des sous-familles. Malgré cela, l'identification des motifs A, B, C, communs à d'autres polymérases à ARN, est à l'origine de la réconnaissance de L en tant que ARN polymérase ARN-dépendante (Poch et al. 1989; Delarue et al. 1990; O'Reilly and Kao 1998). Cela dit, bien d'autres fonctions lui sont classiquement (et souvent hypothétiquement) attribuées comme les activités impliquées dans la coiffe de l'ARN (phosphatase, guanyltransférase et 2'-O-méthyltransférase), des activités de polyadénylation, d'édition, kinase et hélicase. Cette dernière activité pourrait ne pas être nécessaire dans le cas des *Mononegavirales*. Du fait que N encapside l'ARN au cours de la synthèse de ce dernier, la présence du double brin pourrait n'être de ce fait que transitoire. Comme déjà mentionné, L interagit avec la phosphoprotéine. L'importance de la région N-terminale de L dans l'interaction avec P a été montré chez les virus Sendai et de la rougeole (Chandrika et al. 1995; Cevik et al. 2004).

Les protéines L, N et P ont fait l'objet d'études bioinformatiques, dont les résultats sont exposés dans les deux articles suivants.

ARTICLE 3

Viral RNA-Polymerases-A predicted 2'-o-ribose methyltransferase domain shared by all *Mononegavirales*

F. Ferron, S. Longhi, B. Henrissat and B. Canard

Trends in Biochemical Sciences. Vol. 27, 5 May 2002 , 222-224

Viral RNA-polymerases – a predicted 2′-O-ribose methyltransferase domain shared by all *Mononegavirales*

François Ferron, Sonia Longhi, Bernard Henrissat and Bruno Canard

The *Mononegavirales* virus group comprises several major human pathogens, including measles, rabies and Ebola viruses. This article reports a computational analysis of the C-terminal region of RNA-dependent RNA-polymerases from *Mononegavirales*. Using a combination of sequence similarity and threading analysis, a 2′-O-ribose methyltransferase domain was identified that is involved in the capping of viral mRNAs.

Mononegavirales constitute a group of viruses with a single-stranded (–) RNA genome. Many members of this group have a strong impact on human health and are responsible for severe diseases, such as measles, rabies or haemorrhagic fevers. Emerging viruses, such as Hendra and Nipah, cause encephalitis associated with high (>50%) mortality. Despite their significant socio–economical impact, very little is known about their molecular biology. In particular, reactions implicated in the replication cycle and in the mRNA capping processes are still poorly understood.

The cap is a unique structure found at the 5′ termini of cellular eukaryotic mRNAs. It plays a crucial role in mRNA stability and ribosome binding for translation [1]. mRNA capping is a co-transcriptional modification resulting from a series of four chemical reactions that occur in the nucleus [2] . The 5′-triphosphate of the mRNA is first converted into a diphosphate by an RNA triphosphatase. A GMP moiety is then transferred from GTP to the 5′-diphosphate RNA by a guanylyltransferase to yield $G^{5'}$-$ppp^{5'}$-N. In a third reaction using S-adenosyl-L-methionine (SAM) as the methyl donor, the guanosine moiety is methylated by a methyltransferase (MTase) at its N7 position to yield $^{7Me}G^{5'}$-$ppp^{5'}$-N (cap 0 structure). A second methyl transfer reaction then methylates the 2′-OH of the first nucleotide, 3′ of the triphosphate bridge, to yield $^{7Me}G^{5'}$-$ppp^{5'}$-$N_{2'0Me}$ (cap 1 structure).

Viruses adopt different capping strategies according to their replication cycle and/or host, but the large majority adopts the same cap structure as cellular mRNAs [2], even if the order of the guanylyl- and methyltransfer reactions is variable. Viruses having a cytoplasmic replication cycle, such as the *Mononegavirales,* have to synthesize their own mRNA 5′-terminal cap structure.

Genetic studies have shown that the polymerase of vesicular stomatitis virus (VSV), a *Mononegavirales*, possesses a MTase activity involved in the making of a cap 1 structure [3]. However, neither the N7 nor the 2′-O-ribose MTases involved in the capping reaction have yet been biochemically characterized. A previous computational analysis on two families of *Mononegavirales* RNA-dependent RNA-polymerase (L) assigned an ATP-binding site to residues 1785–1799 (numbering refers to measles sequence) [4,5]. In this paper, we have analysed all available *Mononegavirales* polymerase sequences. We report the presence of a MTase domain rather than a NTP-binding site.

The assignment of an ATP-binding site on the L protein was based on the identification of a signature composed of a Gly-x-Gly-x-Gly motif, followed by a lysine-rich sequence and by a (F,Y)-Y-N motif prevalent in protein kinases [4]. It has since been suggested that it could act as a general NTP-binding site [5]. However, the assignment of a GTP-binding site to the Gly-rich motif has been a matter of discussion since 1983 [6]. Keeping this in mind, we noticed that one of the canonical motifs of MTases is a Gly-rich motif, shared by all members of the MTase superfamily. This observation suggests that the putative ATP-binding site might, rather, be part of a MTase domain.

The C-terminal region (residues 1755–1960) of the measles virus L protein was subjected to BLAST and PSI–BLAST analysis [7], with an expected E-value of 10^{-4}. Only viral sequences were retained. The ProDom database [8] was also scanned. This led to the identification of 26 polymerase sequences encompassing three viral families of single negative stranded RNA viruses, belonging to the *Mononegavirales* order.

After suppression of redundant sequences, remaining sequences were aligned using CLUSTAL W [9] and manually adjusted with SEAVIEW [10].

Simultaneously, a fold recognition analysis was performed on the delimited measles virus sequence using the threading programs* 3D-PSSM [11], GENthreader [12] and INBGU [13]. Strikingly, results obtained by these three methods converged to the same protein, namely RrmJ (pdb: 1EJ0), a bacterial 2′-O-ribose MTase responsible for rRNA methylation [14]. 3D-PSSM and GENthreader gave an E-value of $2.70 \times e^{-03}$ and a probability of >99%, respectively. INBGU proposed this protein with the highest score (9.2) in five rounds. Notably, when the RrmJ sequence was used as a query in a reciprocal PSI–BLAST search, a similarity with the C-terminal region of *Mononegavirales* polymerases was detected (E-value 10^{-12} for Marburg virus), thus providing further support for the functional prediction.

We have superimposed the secondary structure elements of RrmJ and Vp39 on the multiple sequence alignment (Fig. 1). Different secondary structural predictions have also been performed on all viral sequences combining different programs, such as PSIPRED* [15] and PHD [16]. All *Mononegavirales* sequences share the same predicted secondary structure organization, thus allowing us to derive a consensus of structural prediction (Fig. 1). A good agreement between the prediction and the structural organization of the reference structures is observed (Fig. 1). β Strands and α helices are almost perfectly superimposed, although a small shift of one or two amino acids can be observed in some cases. Relative to RrmJ,

*PSIPRED&GENthreader server is available at http://insulin.brunel.ac.uk/psipred; 3D-PSSM server is available at http://www.bmm.icnet.uk/~3dpssm/html/ffrecog.html; INBGU server is available at http://www.cs.bgu.ac.il/~bioinbgu/form.html

http://tibs.trends.com 0968-0004/02/$ – see front matter © 2002 Elsevier Science Ltd. All rights reserved. PII: S0968-0004(02)02091-1

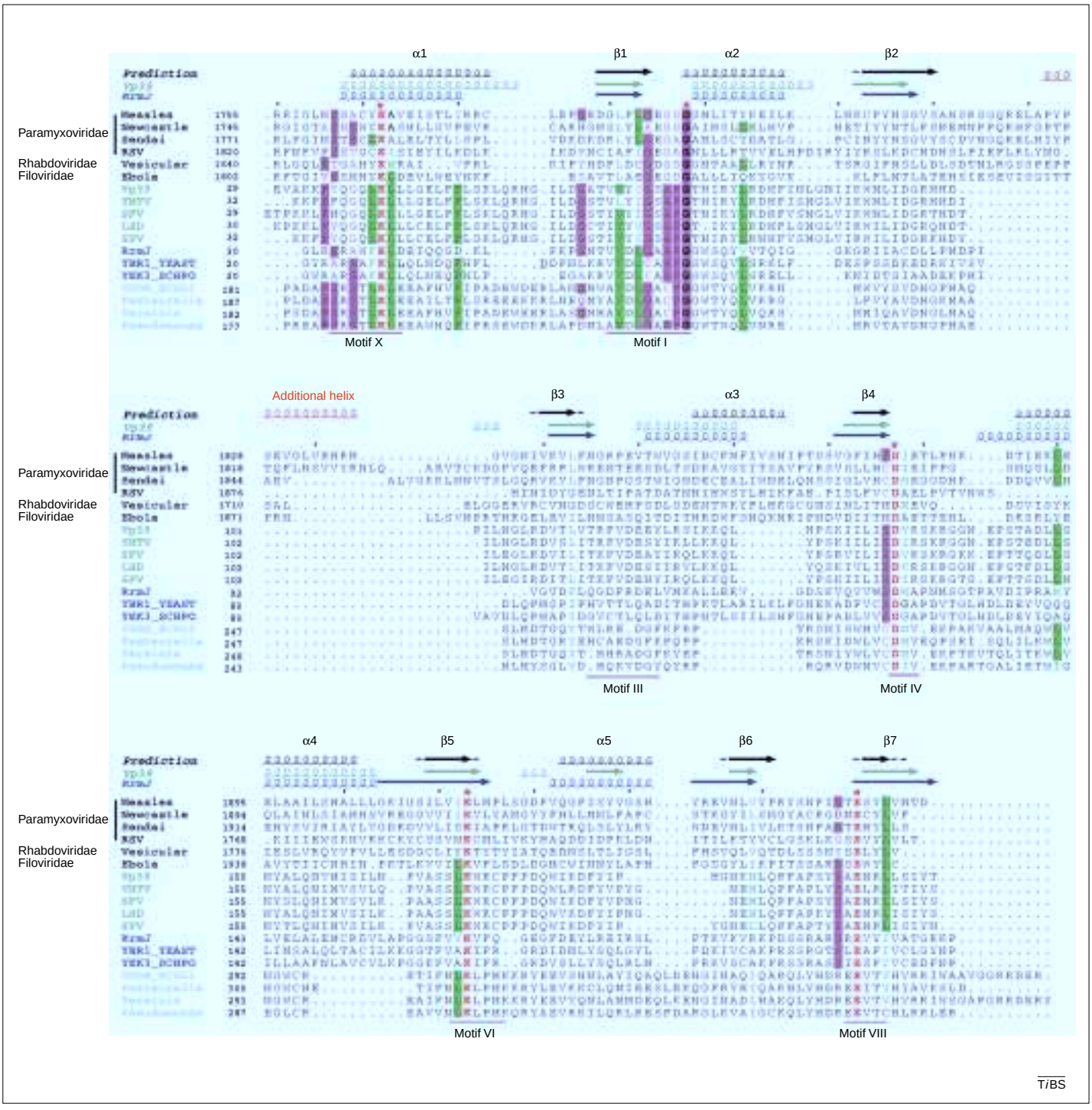


Fig. 1. Multiple sequence alignment of the predicted cap 1 methyltransferase (MTase) domain of *Mononegavirales* polymerases, featuring one *Rhabdoviridae* member, one *Filoviridae* member and four *Paramyxoviridae* members. Three subfamilies of 2′-O-ribose MTases are also presented, sequences of genuine RrmJ in dark-blue, Vp39 in green and putative YgdE in light-blue. To underline each subfamily, representative members are also shown in the corresponding color. Accession numbers and e-values obtained with different BLAST queries are as follows: measles virus, P35975, query for the BLAST; newcastle virus, P11205, e-value: 6.57×10^{-57}; sendai virus, P27566, e-value: 5×10^{-33}; respiratory syncytial virus (RSV), P28887, e-value: 2×10^{-4}; vesicular stomatitis virus (VSV), P13615, e-value: 4×10^{-15}; ebola virus, Q66802, e-value: 8×10^{-36}; Vp39, P21033 (pdb code 1BKY), query for the BLAST; lumpy skin disease virus (LSD) Q91MU2, e-value: 1.7×10^{-82}; swinepox virus (SPV), AAL69804, e-value: 1.3×10^{-79}; shope fibroma virus (SFV), Q9Q906, e-value: 9.6×10^{-82}; yaba monkey tumor virus (YMT), Q9QBB3, e-value: 1.9×10^{-83}; RrmJ, P28692 (pdb code: 1EJ0), query for the BLAST; YBR1_Yeast, P38238, e-value: 2.4×10^{-21}; YEK3_SCHPO, O36015 , e-value: 1.7×10^{-22}; YgdE, P32066, query for the BLAST; *Pasteurella multocida*, Q9CN71, e-value: 3.3×10^{-56}; *Yersina pestis*, Q8ZH78, 39×10^{-78}; *Pseudomonas aeruginosa*, Q9I3F4, e-value: 3.3×10^{-56}. The secondary structure elements of RrmJ (blue) are numbered according to Ref. [14] and those of Vp39 (green) to Ref. [24]. The consensus secondary structure prediction is shown in black, and the additional helix is shown in red. The six motifs of MTases are numbered according to Ref. [17]. In the multiple sequence alignment, the catalytic tetrad represented by the conserved residues K, D, K, E and also the conserved Gly in Motif I, playing a crucial role in S-adenosyl-L-methionine binding, are in bold and indicated by a red star above the alignment. The amino acids involved with 80% homology are highlighted as follows: conservation of aliphatic, polar and hydrophobic residues is indicated by lilac, red and green letters, respectively. Positions corresponding to consensus (>50% identity) are boxed with the corresponding color, except for the residues implicated in the catalytic motif. Dots above the alignment indicate intervals of ten residues. This multiple sequence alignment (alignment number ALIGN_000317) has been deposited with the European Bioinformatics Institute (ftp://ftp.ebi.ac.uk/pub/databases/embl/align/ALIGN_000317.dat).

an extra helix (in red) between β2 and β3 strands of RrmJ is predicted by the consensus in all viral sequences. Nevertheless, this helix is in accordance with the MTase fold, and has been observed in the crystal structure of other MTases [17].

A 2′-O-ribose- or an N7-guanosine methyltransferase?

The multiple sequence alignment reveals the presence of six motifs characteristic of methyltransferases, involved either in SAM binding (motifs I, III, IV) or in the catalytic reaction (motifs IV, VI, VIII, X) (Fig. 1). After establishing the presence of a MTase domain within *Mononegavirales* polymerases, we addressed the question as to whether it methylates the N7-guanine, the 2′-O-ribose hydroxyl, or both.

The alignment reveals conservation of the residues constituting the 2′-O-ribose MTase catalytic tetrad (KDKE), already observed in other predicted and genuine 2′-O-ribose MTases [18,19], such as RrmJ and the vaccinia virus VP39 protein responsible for 2′-O-ribose methylation in the cap 1 structure [20].

Moreover, according to the model of cap 0 N7-guanine MTases, developed by Bujnicki [21], these MTases possess an additional three-stranded β sheet with respect to the canonical MTase fold. This additional motif consists of a 50 amino acid insertion between the strand β_5 and the helix α_5 of the MTase fold. It is required for guanosine stabilization during the N7-guanine methylation. The structural similarity between VP39 and RrmJ, as well as the absence of such additional structural elements either in genuine 2′-O-ribose MTases or in the MTase domain of *Mononegavirales* polymerases, allow us to postulate that an N7-guanosine MTase would not fit in this domain.

Finally, if we benchmark our domain sequence to the SwissProt data bank with a BLAST in standard conditions, the best hits, beyond the viral proteins themselves, are 2′-O-ribose MTases, not adenine or guanine MTases. A scan in the Pfam database [22] is consistent with this hypothesis because it assigned this domain as a FtsJ-like methyltransferase (Pfam: PF01728). Nevertheless, we could not identify in the *Mononegavirales* sequences any aromatic residues that would stabilize the guanine moiety such as those found in the cap binding site of VP39 [23]. In conclusion, all the above lines of evidence lead us to propose that the MTase domain of *Mononegavirales* polymerase methylates the 2′-O-ribose specifically, and the physical mapping of this domain to the RNA polymerase suggests that it is involved in capping of the viral mRNAs.

Acknowledgements

We thank Barbara Selisko for helpful discussions and critical reading of the manuscript. This work was supported by the CNRS. This study has been conducted with financial support from the Commission of the European Communities, specific RTD programme 'Quality of Life and Management of Living Resources', QLK2-CT2001-01225, 'Towards the design of new potent antiviral drugs: structure–function analysis of Paramyxoviridae RNA polymerase'. It does not necessarily reflect its views and in no way anticipates the Commission's future policy in this area.

References

1 Beelman, C.A. *et al.* (1996) An essential component of the decapping enzyme required for normal rates of mRNA turnover. *Nature* 382, 642–646
2 Furuichi, Y. and Shatkin, A.J. (2000) Viral and cellular mRNA capping: past and prospects. *Advances in Virus Research* 55, 135–184
3 Hercyk, N. *et al.* (1988) The vesicular stomatitis virus L protein possesses the mRNA methyltransferase activities. *Virology* 163, 222–225
4 Poch, O. *et al.* (1990) Sequence comparison of five polymerases (L proteins) of unsegmented negative-strand RNA viruses: theoretical assignment of functional domains. *J. Gen. Virol.* 71, 1153–1162
5 McIlhatton, M.A. *et al.* (1997) Nucleotide sequence analysis of the large (L) genes of phocine distemper virus and canine distemper virus (corrected sequence). *J. Gen. Virol.* 78, 571–576
6 Halliday, K.R. (1983) Regional homology in GTP-binding proto-oncogene products and elongation factors. *J. Cyclic Nucleotide Protein Phosphor Res.* 9, 435–448
7 Altschul, S.F. *et al.* (1997) Gapped BLAST and PSI–BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402
8 Corpet, F. *et al.* (2000) ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons. *Nucleic Acids Res.* 28, 267–269
9 Thompson, J.D. *et al.* (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
10 Galtier, N. *et al.* (1996) SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. *Comput. Appl. Biosci.* 12, 543–548
11 Kelley, L.A. *et al.* (2000) Enhanced genome annotation using structural profiles in the program 3D-PSSM. *J. Mol. Biol.* 299, 499–520
12 Jones, D.T. (1999) GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. *J. Mol. Biol.* 287, 797–815
13 Fischer, D. (2000) Hybrid fold recognition: combining sequence derived properties with evolutionary information. *Pac. Symp. Biocomput.*, 119–130
14 Bugl, H. *et al.* (2000) RNA methylation under heat shock control. *Mol. Cell* 6, 349–360
15 McGuffin, L.J. *et al.* (2000) The PSIPRED protein structure prediction server. *Bioinformatics* 16, 404–405
16 Rost, B. (1996) PHD: predicting one-dimensional protein structure by profile-based neural networks. *Methods Enzymol.* 266, 525–539
17 Schluckebier, G. *et al.* (1995) Universal catalytic domain structure of AdoMet-dependent methyltransferases. *J. Mol. Biol.* 247, 16–20
18 Bujnicki, J.M. and Rychlewski, L. (2001) Reassignment of specificities of two cap methyltransferase domains in the reovirus lambda2 protein. *Genome Biol.* 2, 0038
19 Bujnicki, J.M. and Rychlewski, L. (2000) Prediction of a novel RNA 2′-O-ribose methyltransferase subfamily encoded by the *Escherichia coli* YgdE open reading frame and its orthologs. *Acta Microbiol. Pol.* 49, 253–260
20 Hodel, A.E. *et al.* (1996) The 1.85 Å structure of vaccinia protein VP39: a bifunctional enzyme that participates in the modification of both mRNA ends. *Cell* 85, 247–256
21 Bujnicki, J.M. *et al.* (2001) mRNA:guanine-N7 cap methyltransferases: identification of novel members of the family, evolutionary analysis, homology modeling, and analysis of sequence-structure–function relationships. *BMC Bioinformatics* 2, 2
22 Bateman, A. *et al.* (2002) The Pfam protein families database. *Nucleic Acids Res.* 30, 276–280
23 Hodel, A.E. *et al.* (1998) Structural basis for sequence-nonspecific recognition of 5′-capped mRNA by a cap-modifying enzyme. *Mol. Cell* 1, 443–447
24 Hu, G. *et al.* (1999) mRNA cap recognition: dominant role of enhanced stacking interactions between methylated bases and protein aromatic side chains. *Proc. Natl. Acad. Sci. U. S. A.* 96, 7149–7154

François Ferron
Sonia Longhi
Bernard Henrissat
Bruno Canard*
Architecture et Fonction des Macromolécules Biologiques, UMR 6098, CNRS, and Universitées Aix-Marseille I and II, ESIL, 163, Avenue de Luminy, Case 925, F-13288 Marseille Cedex 9, France.
*e-mail: bruno@esil.univ-mrs.fr

-Commentaires

L'analyse bioinformatique a permis d'affiner la cartographie de la protéine L et d'identifier un domaine 2'-O-méthyltransférase sur cette protéine. La définition de ce domaine constitue une découverte intéressante sur le mode de fonctionnement des polymérases des Mononegavirales. Comme c'est le cas chez le virus de la Dengue dont cette activité a été découverte au laboratoire (Egloff et al. 2002), il est logique de penser que cette méthyltransférase puisse jouer un rôle dans le mécanisme de coiffe des ARN messagers viraux. Le cycle réplicatif des Mononegavirales est intégralement cytoplasmique, et leur impose donc d'assurer cette fonction avec des protéines virales. La présence du domaine est donc cohérente avec la nécessité pour le virus de synthétiser des ARNm coiffés. Toutefois, le parallèle ne doit pas être poussé plus loin : en effet si la 2'-O-méthyltransférase a été délimité sur la polymérase, aucun domaine de fixation au GTP n'est présent sur la protéine L (Gorbalenya and Koonin 1989), à la différence de la polymérase du virus de la Dengue (Egloff et al. 2002). Ceci suggère l'existance un mécanisme de coiffe propre aux Mononegavirales, ou pouvant faire intervenir des protéines cellulaires, comme cela a été déjà suggéré pour le virus de la stomatite vesiculaire (VSV) (Das et al. 1998).

La mise en évidence de l'activité 2'-O-méthyltransférase dans la polymérase du VSV (Hercyk et al. 1988) confirme la fiabilité de l'annotation. Toutefois, chez le VSV et le virus Sendaï, il semblerait que la protéine L porte à la fois la fonction N-7-méthyltransférase et 2'-O-méthyltransférase (Spadafora and Perrault 2003 ; Ogino et al. 2004). Les deux fonctions pourraient impliquer soit deux régions différentes de la protéine L, comme suggéré par Spadafora et coll. (Spadafora and Perrault 2003), soit une seule et même région en C-terminale de la protéine, comme postulé par Ogino et coll. (Ogino et al. 2004)

Nous avons tenté d'exprimer le domaine méthyltransférase de trois *Paramyxoviridae*, à savoir le virus de la rougeole (MV), le virus de Sendai (SV) et le virus respiratoire syncitial (RSV). Sur la base de notre assignation, nous avons cloné et tenté d'exprimer chez *E. coli* de nombreuses constructions faisant varier les bornes du domaine en N et C-terminal.

Aucun de ces domaines s'est avéré soluble. Dans le but de solubiliser ce domaine, nous avons essayé d'optimiser les conditions de purification en cherchant des tampons idoines (Lindwall et al. 2000), mais nos efforts furent vains.

Dans le but de comprendre les bases structurales de cette faible solubilité, nous avons construit un modèle du domaine méthyltransférase des Mononegavirales. Pour cela nous avons utilisé la structure de la RrmJ (voir article) comme référence. Nous avons ensuite

A

Protéine L

Domaine méthyltransférase

1 — 2183

Constructions Virus de la Rougeole

* : Expression du domaine seul et et en co-expression avec P

1675		2183
1718	*	2183
1740	*	1960
1755		1960
1740		2008

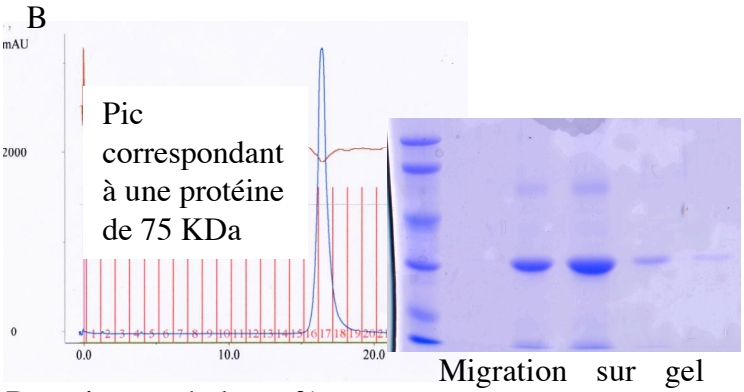
Constructions SeV

1771		1979

Constructions RSV

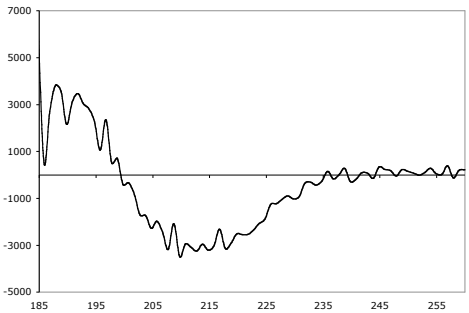
1820		2009

B



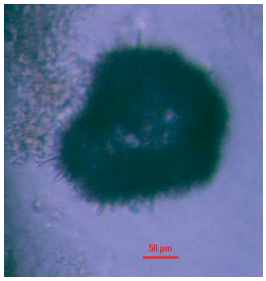
Domaine methyltransférase de SeV purifié par chromatographie d'affinité sous forme de trimère

Migration sur gel SDS PAGE correspondant à une protéine de 25 KDa



Spectre DC du domaine méthyltransférase de SeV

C



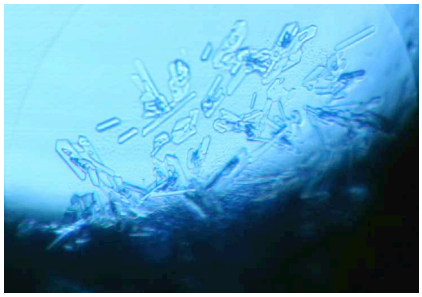


Figure 25 : A) Résumé des différentes constructions du domaine méthyltransférase. B) Purification, gel SDS PAGE et Spectre DC du domaine renaturé du du virus de SeV. C) Exemple de cristaux obtenus.

analysé la structure du domaine méthyltransférase ainsi modélisé et constaté la présence de régions hydrophobes exposées au solvant, qui pourraient être à l'origine des problèmes de solubilité que nous avons rencontrés.
En ultime approche, nous avons donc décidé d'exprimer et de purifier le domaine méthyltransférase du virus Sendaï (résidus 1771-1979) en conditions dénaturantes et de le renaturer selon un protocole mis au point au laboratoire (Figure 25).

Cette approche nous a permis effectivement d'obtenir le domaine purifié et en quantité suffisante pour des études fonctionnelles et cristallographiques. Le domaine obtenu est stable et apparaît sous forme de trimère. Il est concevable que lors de la renaturation les régions hydrophobes ne soient plus exposées au solvant, mais impliquées dans des interactions protéine-protéine. Nous avons vérifié au préalable par dichroïsme circulaire que la protéine avait un repliement conforme à celui des méthyltransférases et par diffusion dynamique de la lumière (DLS) quelle n'était pas agrégée. Ces résultats, nous ont rendu particulièrement confiants pour la suite des expériences de cristallogenèse (voir paragraphe suivant).

-Essais de cristallogenèse du domaine méthyltransférase du virus Sendai.

Nous avons testé plus de deux mille conditions de cristallisation, en utilisant la technique de diffusion de la vapeur. Deux conditions ont donné lieu à de petits cristaux, et les optimisations sont en cours. Des expériences de DLS ont montré que la protéine s'agrège au cours du temps, ce qui pourrait expliquer la difficulté que l'on rencontre pour obtenir, reproduire et optimiser les cristaux. Il semblerait donc que nous n'ayons pas encore trouvé la définition optimale du domaine en vue de sa cristallisation. Les résultats obtenus par Ogino et coll. montrent que la région délimitée par les résidus 1756-2228, et possédant l'activité N-7-méthyltransférase, est exprimée sous une forme soluble et stable chez les cellules d'insectes. Nous envisageons donc d'exprimer cette partie de la protéine chez la bactérie et/ou d'établir une collaboration avec ce groupe dans le but de résoudre la structure tridimensionnelle de ce domaine (Figure 25).

-Un domaine clé pour la structure de L ?

Le dernier point que je souhaite discuter concerne le rôle structural probable de ce domaine. Contrairement à ce que laisse penser le titre de l'article, il y a une exception parmi les Mononegavirales. Les *Bornaviridae* sont apparemment dépourvus de la fonction méthyltransférase. Bien qu'appartenant à l'ordre des Mononegavirales, ces virus possèdent un

cycle mixte cytoplasmique et nucléaire. L'analyse de la protéine L des *Bornaviridae* nous apprend beaucoup sur les protéines L des Mononegavirales. Bien que plus petite, elle reste comparable aux autres protéines L de l'ordre. Il est intéressant de noter que, si les résidus catalytiques (K-D-K-E) sont absents, il est encore possible de distinguer la signature du motif I. La conservation de ce motif chez les polymérases des *Bornaviridae* suggère que ces virus ont probablement évolué à partir d'un ancêtre possédant un domaine méthyltransférase fonctionnel. La perte de la fonction est probablement liée a une adaptation qui a permis au complexe réplicatif de ces virus de passer dans le noyau, où des enzymes cellulaires pourraient assurer la coiffe des messagers viraux. Cependant, le fait que le domaine méthyltransférase soit toujours présent laisse supposer que celui-ci pourrait jouer un rôle fondamental dans l'architecture de la protéine. Alternativement, il est possible d'imaginer que ce domaine soit destiné à disparaître et que sa présence ne traduise qu'un intermédiaire au cours de l'évolution.

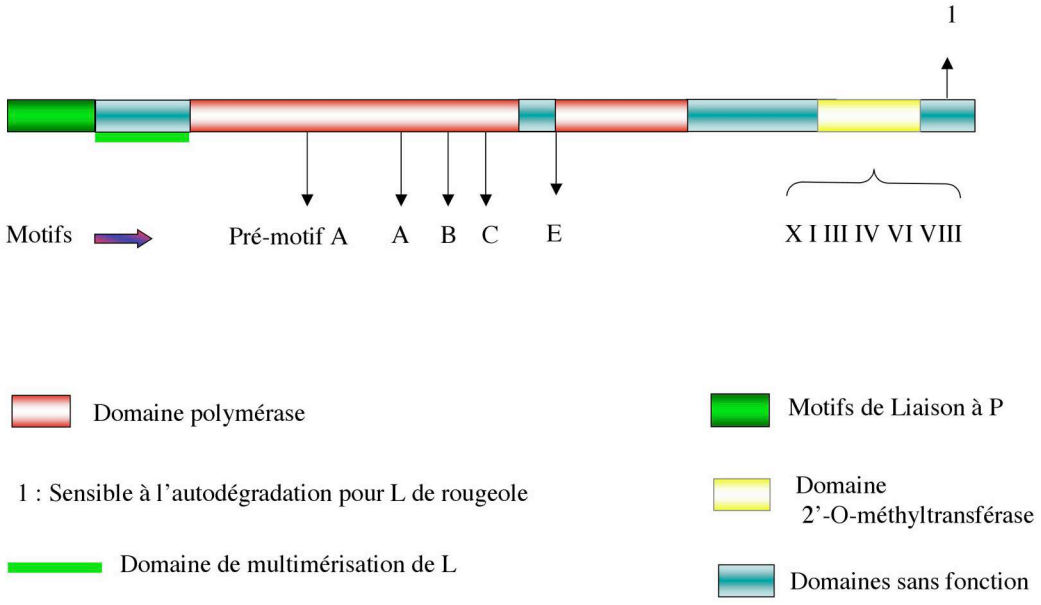


Figure 26 : Cartographie bilan de la protéine L.

ARTICLE 4

Structural disorder and modular organization in Paramyxovirinae N and P

D. Karlin, F. Ferron, B. Canard, and S. Longhi

Journal of General Virology ; Dec 1 , 2003, 84 (12) : 3239-3252

Journal of General Virology (2003), **84**, 3239–3252 DOI 10.1099/vir.0.19451-0

Structural disorder and modular organization in *Paramyxovirinae* N and P

David Karlin,† François Ferron, Bruno Canard and Sonia Longhi

Architecture et Fonction des Macromolécules Biologiques, UMR 6098 CNRS et Université Aix-Marseille I et II, ESIL, Campus de Luminy, 13288 Marseille Cedex 09, France

Correspondence
Sonia Longhi
longhi@afmb.cnrs-mrs.fr

Received 23 June 2003
Accepted 29 August 2003

The existence and extent of disorder within the replicative complex (N, P and the polymerase, L) of *Paramyxovirinae* were investigated, drawing on the discovery that the N-terminal moiety of the phosphoprotein (P) and the C-terminal moiety of the nucleoprotein (N) of measles virus are intrinsically unstructured. We show that intrinsic disorder is a widespread property within *Paramyxovirinae* N and P, using a combination of different computational approaches relying on different physico-chemical concepts. Notably, experimental support that has often gone unnoticed for most of the predictions has been found in the literature. Identification of disordered regions allows the unveiling of a common organization in all *Paramyxovirinae* P, which are composed of six modules defined on the basis of structure or sequence conservation. The possible functional significance of intrinsic disorder is discussed in the light of experimental data, which show that unstructured regions of P and N are involved in numerous interactions with several protein and protein–RNA partners. This study provides a contribution to the rather poorly investigated field of intrinsically disordered proteins and helps in targeting protein domains for structural studies.

INTRODUCTION

Paramyxovirinae, which include major human pathogens such as parainfluenza virus and measles virus (MV), are enveloped viruses with a non-segmented, negative, single-stranded RNA genome encapsidated by the nucleoprotein (N) within a helical nucleocapsid. Transcription and replication are carried out on this (N : RNA) template by a viral RNA-dependent RNA polymerase complex, made of the phosphoprotein (P) and the large protein (L) (reviewed by Lamb & Kolakofsky, 2001). Association of P with the soluble, monomeric form of N (N^o) prevents its illegitimate self-assembly onto cellular RNA. The assembled form of N (N^{NUC}) also forms complexes with P and P–L during transcription and replication (Lamb & Kolakofsky, 2001).

N consists of two regions: an N-terminal moiety, well conserved in sequence, N_{CORE}, and a hypervariable, C-terminal moiety, N_{TAIL}. N_{CORE} contains all the regions necessary for self-assembly and RNA binding. N_{TAIL} binds P within both N^{NUC} and N^o and is required for N : RNA to act as a template for viral RNA synthesis (Bankamp *et al.*, 1996; Buchholz *et al.*, 1994; Curran *et al.*, 1993; Harty & Palese, 1995; Nishio *et al.*, 1999).

From a structural point of view, P is the best-characterized protein of the replicative complex. P is organized into two moieties that are functionally and structurally distinct: a C-terminal moiety (PCT) and an N-terminal moiety (PNT). PCT is the most conserved in sequence and contains all regions required for virus transcription, whereas PNT, which is poorly conserved, provides several additional functions required for replication (Curran & Kolakofsky, 1999). P forms oligomers through a coiled-coil motif located within PCT. PCT also contains the region responsible for binding to L (Liston *et al.*, 1995; Smallwood *et al.*, 1994), as well as the regions necessary for binding N^{NUC} (Harty & Palese, 1995; Ryan *et al.*, 1991). The extreme C-terminal domain of PCT (called XD, 'X domain') is responsible for binding to N^{NUC}, as well as for stable binding to N^o (Curran *et al.*, 1995b; Nishio *et al.*, 1996; Shaji & Shaila, 1999). The structure of the Sendai virus (SeV) P multimerization domain (PMD) has been solved by X-ray crystallography. It is composed of a short bundle of α-helices located upstream of the coiled-coil (Tarbouriech *et al.*, 2000b). The only structural information available on PNT concerns MV PNT, which is unstructured *in vitro* (Karlin *et al.*, 2002b). PNT prevents the illegitimate self-assembly of N^o by binding to it. The main N^o-binding site has been mapped to the N terminus of PNT in all *Paramyxovirinae* (Curran *et al.*, 1995b; Nishio *et al.*, 1996; Precious *et al.*, 1995; Shaji & Shaila, 1999; Tober *et al.*, 1998). Beyond P, the P mRNA encodes a variety of proteins, including proteins consisting of either PNT alone (proteins W and R) or PNT fused to a zinc-binding region (protein V) (Lamb & Kolakofsky, 2001).

L is thought to be a multifunctional enzyme carrying most catalytic functions necessary for synthesis of viral RNA, such as RNA-dependent RNA polymerase (Poch *et al.*, 1990;

†Present address: Ecole de l'ADN, Association Grand Luminy, Case 922, Bât. CCIMP, 13288 Marseille Cedex 09, France.

0001-9451 © 2003 SGM *Printed in Great Britain*

Svenda *et al.*, 1997) and 2′-*O*-methyltransferase (Ferron *et al.*, 2002). However, very little is known about its functional organization and almost nothing about its structural organization. The stable P-binding site of L is located in the N-terminal moiety of L (Holmes & Moyer, 2002; Malur *et al.*, 2002; Parks, 1994).

Although the roles of N, P and L within the replicative complex of *Paramyxovirinae* have been partially clarified, very limited three-dimensional information on the replicative machinery is available. The lack of structural data stems from several facts: (i) the difficulty of obtaining homogeneous polymers of N suitable for X-ray analysis (Karlin *et al.*, 2002a; Schoehn *et al.*, 2001); (ii) the low abundance of L in virions and its very large size that renders its heterologous expression difficult; and (iii) the structural flexibility of N and P. Indeed, we have reported recently that MV PNT (Karlin *et al.*, 2002b) and N_{TAIL} (Longhi *et al.*, 2003) are intrinsically disordered. The terms intrinsically disordered (or natively unfolded) designate proteins or protein domains that are unstructured *in vitro* under physiological conditions of salt and pH, in the absence of a binding partner (reviewed by Dunker *et al.*, 2001; Uversky, 2002b; Wright & Dyson, 1999). In recent years, it has been discovered that they are usually distinguished from globular proteins by common sequence features. Intrinsically disordered proteins (IDPs) tend to have a low sequence complexity (i.e. they make use of fewer types of amino acids) (Romero *et al.*, 2001). They are generally enriched in amino acids preferred at the surface of globular proteins (A, R, G, Q, S, P, E and K) (termed 'disorder-promoting amino acids') and are depleted in W, C, F, I, Y, V, L and N ('order-promoting amino acids') (Williams *et al.*, 2001).

Their distinct sequence properties allow disordered regions to be predicted with good accuracy. A neural network-based predictor of naturally disordered regions (PONDR) allows predictions of long disordered regions (>40 aa) (LDRs) of proteins with very good confidence (>99·6 %) (Li *et al.*, 1999; Romero *et al.*, 1997). Long ordered regions (>40 aa) are predicted with a similar confidence. PONDR, however, tends to underpredict disordered regions and therefore its disorder predictions can be considered as conservative (Dunker *et al.*, 2002b).

Another quantitative method to characterize disordered proteins or protein domains relies on their mean net charge/mean hydrophobicity ratio, which is distinctly higher than that of their structured counterparts. This allows the two classes of proteins to be discriminated with a very good accuracy (Uversky, 2002b; Uversky *et al.*, 2000). However, contrary to PONDR, the hydrophobicity/net charge method can only be applied to modular regions and therefore requires prior knowledge of the organization of the protein under study.

A more qualitative method is hydrophobic cluster analysis (HCA) (Callebaut *et al.*, 1997). A HCA plot consists of a two-dimensional, helical representation of a protein sequence, allowing an intuitive visualization of clusters of hydrophobic amino acids (generally corresponding to secondary structure elements in globular proteins). Because the hydrophobic cluster information is plotted directly on the primary sequence, globular regions can be visualized, owing to their typical, thick distribution of hydrophobic clusters. On the contrary, non-globular regions are generally poor in hydrophobic residues and rich in polar residues.

Combining these different computational methods, we have analysed the presence and extent of structural disorder within *Paramyxovirinae* N, P and L. We focused on the three best-characterized genera (*Morbillivirus*, *Respirovirus* and *Rubulavirus*), calling in other viruses when they present informative differences. We used PONDR as a first guide to delineate disordered regions within N, P and L. Then, the precise boundaries of disordered regions were manually refined with HCA. This approach allowed the identification of modular regions to which we applied the net/charge hydrophobicity method. In our analysis, we have also taken into account other indicators of structural disorder, such as low sequence complexity (Romero *et al.*, 2001), lack of predicted secondary structure (Liu *et al.*, 2002) and sequence variability (Brown *et al.*, 2002).

We show that spectacularly long unstructured regions are found in two (out of three) actors of the replicative complex of *Paramyxovirinae*, namely N and P. These disordered regions are conserved in the different genera, implying functional significance. Beyond providing a contribution to the study of the rather poorly investigated field of IDPs, the identification of disordered regions within these proteins facilitates their study at the structural and functional level.

METHODS

Sequence retrieval. Sequences for this study were obtained from the NCBI. Sequence accession numbers for P are: MV, CAA91364; SeV, P04860; canine distemper virus (CDV), AAG15481; Nipah virus (NiV), NP_112022; Menangle virus (MeV), AAK62280; Newcastle disease virus (NDV), NP_071467; Hendra virus (HeV), NP_047107; goose paramyxovirus (GPV), AAN04252; human parainfluenza virus type 2 (hPIV-2), NP_598402; hPIV-4, A43685; avian paramyxovirus (APV), NP_150058; simian virus type 5 (SV5), P11208; Tioman virus (TiV), NP_665865; mumps virus (MuV), P16072; and La-Piedad-Michoacan-Mexico virus (LPMV), AAL09693. Sequence accession numbers for N are: bovine parainfluenza virus type 3 (bPIV-3), AAF28254; MV, P35972; SeV, Q07097; and hPIV-2, P21737.

Plotting mean net charge against mean hydrophobicity to assess whether a protein is intrinsically disordered. The mean net charge (R) and the mean hydrophobicity (H) of a protein were calculated as described in Karlin *et al.* (2002b) and Uversky *et al.* (2000). For a given protein, R is then plotted against H. The charge/hydrophobicity diagram is divided into two regions by a line, which corresponds to the equation $H=(R+1{\cdot}151)/2{\cdot}785$. In the left part of the diagram [where $H<(R+1{\cdot}151)/2{\cdot}785$], a protein is predicted as disordered, whereas it is predicted as ordered in the right part. The net charge-hydrophobicity method is only applicable to a protein (or protein region) provided it is not composed of shorter,

structurally independent modules. It might otherwise give conflicting results. It was only validated for regions >50 aa (Uversky *et al.*, 2000). An estimation of its error rate can be drawn from Uversky (2002b). In that study, no globular protein was found to have a ratio located on the left side of the line, indicating that the positive error rate for the prediction of disordered proteins must be very low. However, five unfolded proteins out of 105 – which were all borderline – were wrongly assigned as being globular, indicating a negative error rate of about 5 %.

PONDR prediction of unstructured regions. Sequences were submitted to the PONDR server (http://www.pondr.com/) using the default integrated predictor VL-XT (Li *et al.*, 1999; Romero *et al.*, 2001). The threshold for reliable (>99·6 %) predictions of disorder, or of order, is set to 40 residues. Access to PONDR was provided by Molecular Kinetics (Pullman, WA, USA) under licence from the WSU Research Foundation. PONDR is copyright ©1999 by the WSU Research Foundation, all rights reserved.

HCA and amino acid composition analysis. HCA was carried out with the program DRAWHCA (Callebaut *et al.*, 1997). The average sequence composition of globular proteins was taken from Tompa (2002). If the average composition of an amino acid X in globular proteins is CG_X, and CP_X is the composition in X of a protein P, deviation from the composition in X of globular proteins was defined for P as $(CP_X-CG_X)/CG_X$.

Identification of low sequence complexity segments and secondary structure predictions. Low sequence complexity segments were identified using the 'low complexity filter' of the BLAST program at the NCBI (http://ncbi.nlm.nih.gov), based on the program SEG (Wootton, 1994). Secondary structure predictions were performed with PSI-PRED (McGuffin *et al.*, 2000) and the PREDICT protein server (Rost, 1996). The results presented are a consensus of both methods.

Multiple sequence alignment of N and P. The sequences of *Paramyxovirinae* N_{CORE} were aligned using CLUSTALW and manually refined with SEAVIEW (Galtier *et al.*, 1996). The sequences of N_{TAIL} could not be aligned among different genera. The sequence alignment of *Paramyxovirinae* PCT was generated in the same way and was essentially the same as that reported in Curran *et al.* (1995a), with the exception of the central region of PCT, which is not presented in Curran *et al.* (1995a). The sequences of PNT among different genera could not be aligned, with the exception of the N-terminal regions of *Rubulavirus*, *Henipavirus* and *Avulavirus* PNT (see below).

Multiple sequence alignment of the N-termini of PNT. The sequences related to the N terminus of *Rubulavirus* PNT (aa 1–100) were retrieved by PSI-BLAST (Altschul *et al.*, 1997) from SWISS-PROT (Bairoch & Apweiler, 2000), PDB (Berman *et al.*, 2000) and translated GenBank (Benson *et al.*, 2002). Fragments and duplicates were discarded. PSI-BLAST converged after seven iterations. The most distant hit is given with a significant (4×10^{-8}) E-value. All positive hits were used as subsequent PSI-BLAST queries and cross-validated.

As the sequences are not closely related, the first alignment was done using CLUSTALW (Thompson *et al.*, 1994) with the slow algorithm, an identity matrix, a window of 3 aa and the standard gaps penalties. The alignment was manually refined with SEAVIEW (Galtier *et al.*, 1996) using predicted structural information. The alignment was drawn using ESPript 2.0 (Gouet *et al.*, 1999).

RESULTS

MV PNT and N_{TAIL} are intrinsically disordered (Karlin *et al.*, 2002b; Longhi *et al.*, 2003). We have analysed the sequence properties of *Paramyxovirinae* N, P and L in order to determine whether such intrinsic disorder is a conserved feature in these viruses. The identification of disordered regions is expected to help decipher their modular organization and thereby facilitate their structural characterization.

All *Paramyxovirinae* N_{TAIL} are intrinsically disordered

PONDR predicts at least one LDR (>40 aa) in N_{TAIL} in the three genera but none in N_{CORE} (Fig. 1A). The HCA plot of MV N shows the presence within N_{CORE} of two large regions rich in hydrophobic clusters (Fig. 2). An antigenic region of N_{CORE} (Giraudon *et al.*, 1988) is clearly visible as a short interruption of hydrophobic clusters, indicating that it may form a loop exposed to the solvent. Conversely, MV N_{TAIL} contains a strikingly long region totally devoid of hydrophobic clusters (aa 421–494) (Fig. 2), which correlates well with the PONDR prediction. In all *Paramyxovirinae*, N_{TAIL} has little or no predicted secondary structure (shown for MV on Fig. 2). Furthermore, in the three genera, N_{TAIL} possesses a combination of low hydrophobicity and relatively high net charge (due to the presence of numerous acidic residues) typical of IDPs (Fig. 3). Finally, N_{TAIL} is also greatly variable in sequence. Thus, the sequence properties of *Paramyxovirinae* N_{TAIL} converge to show that they are intrinsically disordered.

A flexible linker between the coiled-coil and XD in all *Paramyxovirinae*?

PONDR predictions point to the possible presence of a disordered linker between the coiled-coil and XD. In all respiroviruses, with the exception of SeV, this region is predicted to be a LDR (data not shown). In SeV, this region displays a pattern of predicted disorder alternating with borderline order (Fig. 1B, circled). A similar pattern of alternating order and disorder can be seen in all morbilliviruses (circled in Fig. 1B for CDV). In almost all rubulaviruses, the corresponding region is predicted as a LDR (shown in Fig. 1B for hPIV-2). HCA predictions are not conclusive but show that this region exhibits peculiar properties in all viruses. In particular, the linker region of SeV P, while not being as poor in hydrophobic clusters as LDRs found in P, contains distinctly fewer such clusters than globular regions, such as PMD or XD (Fig. 4). At the same time, it is enriched in disorder-promoting residues (data not shown). A region of analogous composition can be found between PMD and PX in *Morbillivirus* (data not shown) and *Rubulavirus* (shown in Fig. 5 for hPIV-2). Finally, contrary to *Rubulavirus* (Fig. 5) and *Morbillvirus* (data not shown), the linker of *Respirovirus* PCT is predicted to lack any secondary structure (shown in Fig. 4 for SeV).

In conclusion, a disordered linker is very probably found in *Respirovirus* and *Rubulavirus* PCT. It is probably found in *Morbillivirus*, too, but we could not reach the same degree of confidence as for our other predictions.

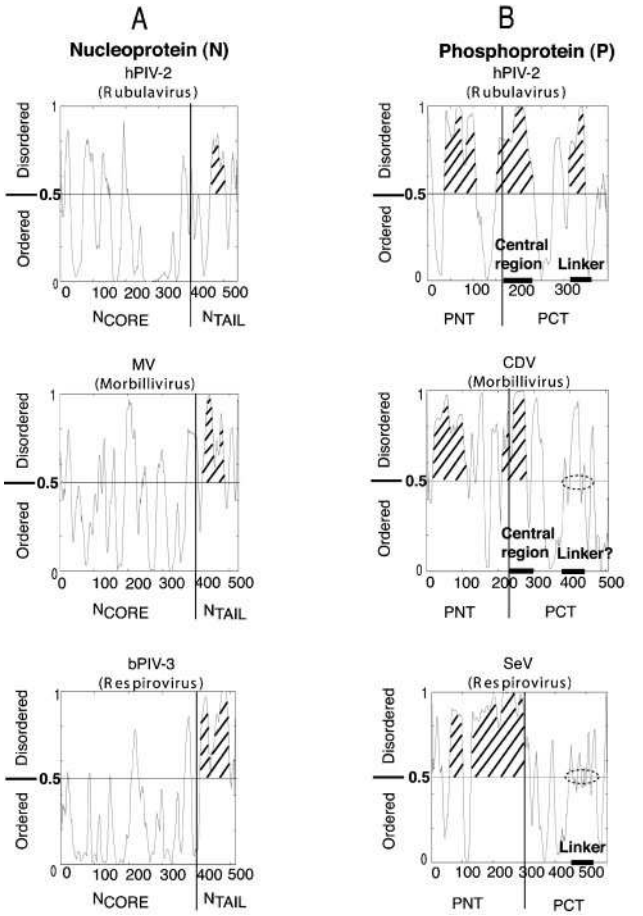


Fig. 1. PONDR predictions of structural disorder in *Paramyxovirinae* N (A) and P (B). Disorder prediction values for a given residue are plotted against the residue number. The significance threshold, above which residues are considered to be disordered, set to 0·5, is shown. LDRs (>40 residues) are hatched. PONDR predictions are qualitatively similar for the N and P proteins of other viruses in each genus (same number of LDR and approximately same position) (data not shown), with the exception of the linker region, which is discussed in the text. N_{CORE} and N_{TAIL}, as well as PNT and PCT, are separated by a vertical line. In the latter case, the line is placed at the border between the region shared by P and V (PNT) and the region unique to P (PCT). The central regions (see text) of *Rubulavirus* and *Morbillivirus* PCT, as well as the linker within *Respirovirus* PCT, are underlined in bold. Predictions of disorder alternating with borderline order in regions corresponding to the linker of *Respirovirus* and *Morbillivirus* PCT are circled (see text).

A disordered central region in *Rubulavirus* and *Morbillivirus* PCT

The region of P located downstream of PNT and upstream of the coiled-coil, herein referred to as the central region, has different properties between *Respirovirus* and the rest of *Paramyxovirinae*. In *Respirovirus*, the central region is composed of a bundle of α-helices (A to C in SeV PMD), buttressing the coiled-coil (Tarbouriech *et al.*, 2000b). The central regions of other genera share little or no sequence similarity among them but all have the same peculiar composition (being rich in G, S and A). They are depleted in most 'order-promoting' residues and enriched in most 'disorder-promoting' residues (Fig. 6), suggesting that they might be disordered (Williams *et al.*, 2001). In agreement with these sequence features, PONDR predicts a LDR in the central regions of *Rubulavirus* and *Morbillivirus* P (Fig. 1B). Furthermore, their HCA plots are typical of disordered regions, they lack predicted secondary structure (shown for hPIV-2 in Fig. 5) and contain low sequence complexity segments (Fig. 5). In both genera, their net charge/hydrophobicity ratios are those of globular proteins, although they are borderline (Fig. 3). In conclusion, the central region is likely intrinsically disordered. Interestingly, the central region of P overlaps the V ORF. Similarly, the C ORF overlaps PNT (Lamb & Kolakofsky, 2001), which is unstructured (see below). This suggests that the presence of unstructured regions might be a common feature of proteins encoded by overlapping reading frames.

The PNT moiety

There are functional and structural differences between *Morbillivirus* and *Respirovirus* PNT, which are acidic and 230–320 aa in length, and *Rubulavirus* PNT, which are shorter (about 160 aa), basic and have a short stretch of sequence similarity with the C proteins of *Respiroviruses* (Lamb & Kolakofsky, 2001). Moreover, *Rubulavirus* V is found in virions (Paterson *et al.*, 1995) and binds RNA through a stretch of basic residues located within PNT (Lin *et al.*, 1997), contrary to the V proteins of the other genera. Consequently, we have analysed separately *Morbillivirus* and *Respirovirus* PNT on one hand and *Rubulavirus* PNT on the other hand.

Respirovirus and *Morbillivirus* PNT are disordered but are composed of two distinct regions

We reported previously that *Morbillivirus* and *Respirovirus* PNT are predicted to be largely disordered using both PONDR (Fig. 1B) and the hydrophobicity/net charge method (Fig. 3) (Karlin *et al.*, 2002b). Further analysis using HCA and secondary structure predictions reveals that PNT is in fact divided into two regions: an N-terminal region rich in hydrophobic clusters associated with a clear α-helical propensity followed by a region devoid of hydrophobic clusters and of predicted secondary structure (shown for SeV in Fig. 4). The α-helical propensity of the extreme N-terminal region of PNT is in agreement with data available in the literature on MV PNT (Karlin *et al.*, 2002b).

Modular organization of *Rubulavirus* PNT

We found previously that *Rubulavirus* PNT has the hydrophobicity/net charge ratio typical of globular proteins (Fig. 3) (Karlin *et al.*, 2002b). However, we present evidence indicating that *Rubulavirus* PNT is composed of at least two

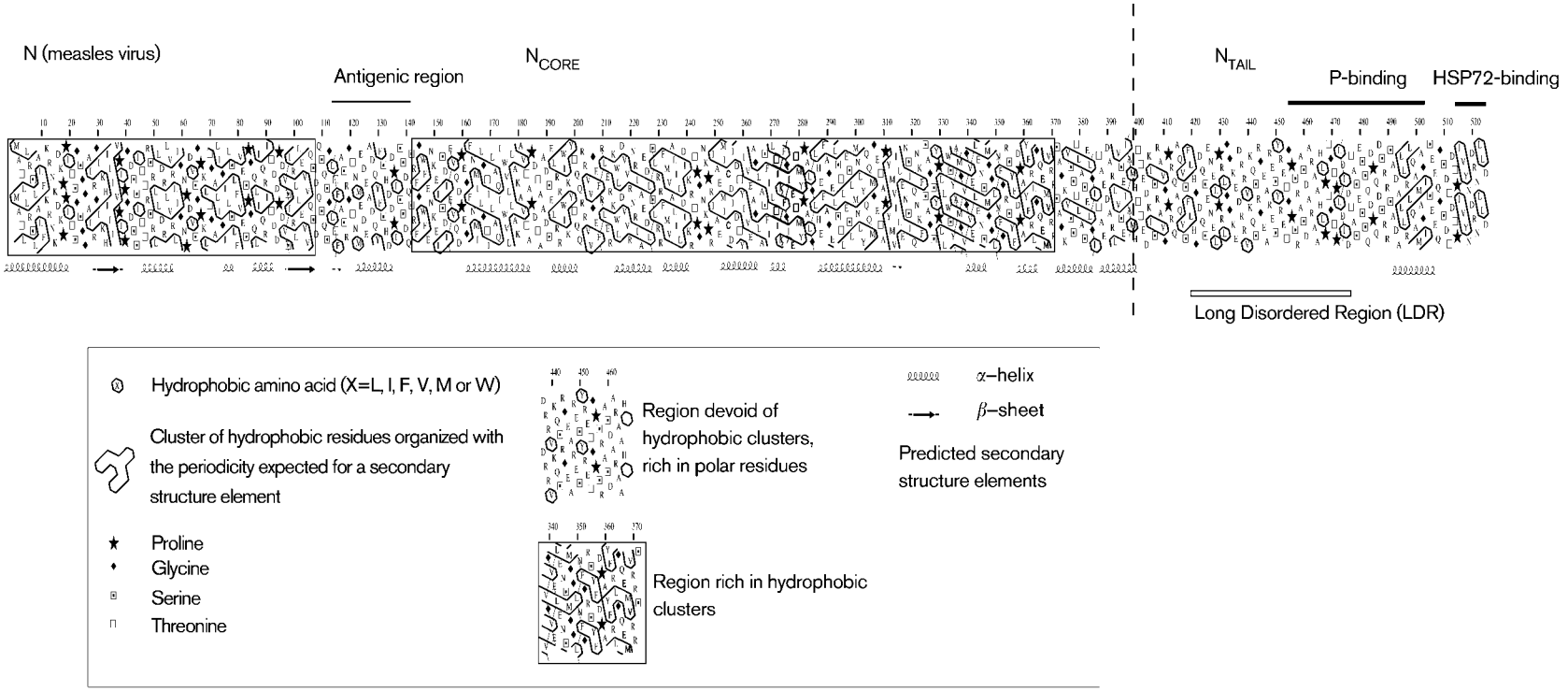


Fig. 2. HCA plot of MV N. Conventions are given in the caption. Globular regions (framed) are characterized by a thick distribution of hydrophobic clusters, while unstructured regions are poor or devoid of hydrophobic clusters. LDR and predicted secondary structure elements are shown. An antigenic region of N_{CORE} (see text) (Giraudon *et al.*, 1988) is highlighted above the diagram. There are no low complexity regions in MV N.

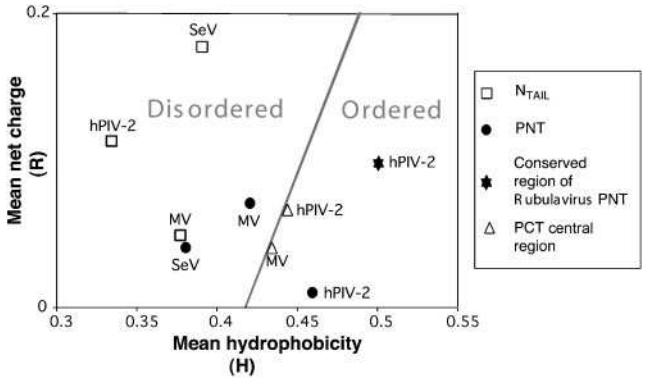


Fig. 3. Net charge/hydrophobicity plot of different regions of *Paramyxovirinae* N and P. The mean net charge (R) of a protein region is plotted against its mean hydrophobicity (H). In the left part of the diagram, a protein is predicted to be intrinsically disordered, whereas it is predicted to be structured in the right part (see Methods).

modular regions (see below). Since the hydrophobicity/net charge method can be applied only to modules, the predictions of globularity obtained on the whole PNT cannot be considered reliable. Using PSI-BLAST (Altschul *et al.*, 1997), we have identified in *Rubulavirus* PNT a conserved N-terminal region with a previously unreported sequence identity with the N-termini of *Henipavirus* and NDV PNT (Fig. 7). Conversely, there is no detectable sequence identity among the corresponding regions of *Morbillivirus* and *Respirovirus* PNT. The conserved N-terminal region of *Rubulavirus* is distinguished by numerous hydrophobic clusters and by a high α-helix-forming potential (shown for hPIV-2 in Fig. 5). Notably, it has the hydrophobicity/net charge ratio typical of globular proteins (shown for hPIV-2 in Fig. 3). It contains an N^o-binding region and a nuclear localization signal, which can both function in isolation, arguing for some degree of functional independence of this region (Watanabe *et al.*, 1996).

The region downstream of this conserved module is mostly disordered, as estimated by PONDR (Fig. 1B), HCA (Fig. 5) and as suggested by the lack of predicted secondary structure (Fig. 5). Nevertheless, two subgroups of *Rubulavirus* can be distinguished. The first subgroup includes hPIV-2 and closely related viruses such as SV5, SV41 and MuV, while the second comprises more distant viruses such as MeV, TiV or LPMV. In the first subgroup, the disordered region downstream of the conserved region contains a lysine/arginine-rich RNA-binding region (Fig. 5) (Lin *et al.*, 1997), which is noteworthy because unstructured regions of proteins (that fold upon binding RNA) are a recurring theme in RNA–protein interactions (Dyson & Wright, 2002; Leulliot & Varani, 2001). The region downstream of the RNA-binding site contains a hydrophobic cluster (aa 111–143) corresponding to a short sequence reportedly homologous to *Respirovirus* C (Fig. 5) (Lamb & Kolakofsky, 2001). However, a motif derived from the corresponding alignment retrieves numerous unrelated, non-viral sequences (data not shown), thus reducing the reliability of the inferred relationship. In the second subgroup of *Rubulavirus*, the region downstream of the conserved module is widely variable and neither the Lys/Arg-rich motif nor the region of identity with *Respirovirus* C can be found. It is consistently predicted to be disordered by PONDR and lacks predicted secondary structure (data not shown).

Because of the shortness of the N-terminal-conserved module and of its position upstream of a disordered region, we cannot conclude whether or not it can fold alone. However, its α-helical potential and high hydrophobicity/net charge ratio indicate that it has the potential to fold in cooperation with another part of P or with another protein.

A common modular organization in all *Paramyxovirinae* P

Identification of disordered regions has allowed us to unveil a common modular organization of P. A summary of our findings on P is presented in Fig. 8. In particular, Fig. 8(A) shows the general organization of P in all *Paramyxovirinae*, whereas Fig. 8(B) shows the organization of P in a representative member of each of the three main genera. P has an even further modular organization than was thought previously, consisting of six distinct regions: a hydrophobic, N^o-binding region with α-helical potential, a disordered region of greatly variable length, a central region (overlapping the V ORF) than can be either ordered or disordered, a coiled-coil, a disordered linker and the XD.

Structural flexibility: a widespread property of *Mononegavirales* N and P?

The last 80 C-terminal residues of *Pneumovirinae* P, as well as the N-termini of *Rhabdoviridae* and *Bornaviridae* P, are predicted to be in large part disordered using PONDR and HCA (data not shown). In the same vein, *Filoviridae* N are grossly organized into an N-terminal moiety homologous to *Rhabdoviridae* and *Paramyxoviridae* N_{CORE} (Barr *et al.*, 1991), and a C-terminal moiety that is hypervariable, very acidic (Sanchez *et al.*, 2001), has a low sequence complexity and contains large predicted disordered regions (data not shown). It might thus be a structural equivalent of *Paramyxoviridae* N_{TAIL}. Likewise, the C-termini of *Bornaviridae* N are predicted to be in large part disordered using PONDR (data not shown). Taken together, these observations suggest that structural flexibility is a widespread property within the replicative complex of non-segmented, negative-stranded viruses.

DISCUSSION

Disorder in the replicative complex of *Paramyxovirinae*

Using complementary biocomputing methods, we have identified disordered regions of N and P and unveiled a

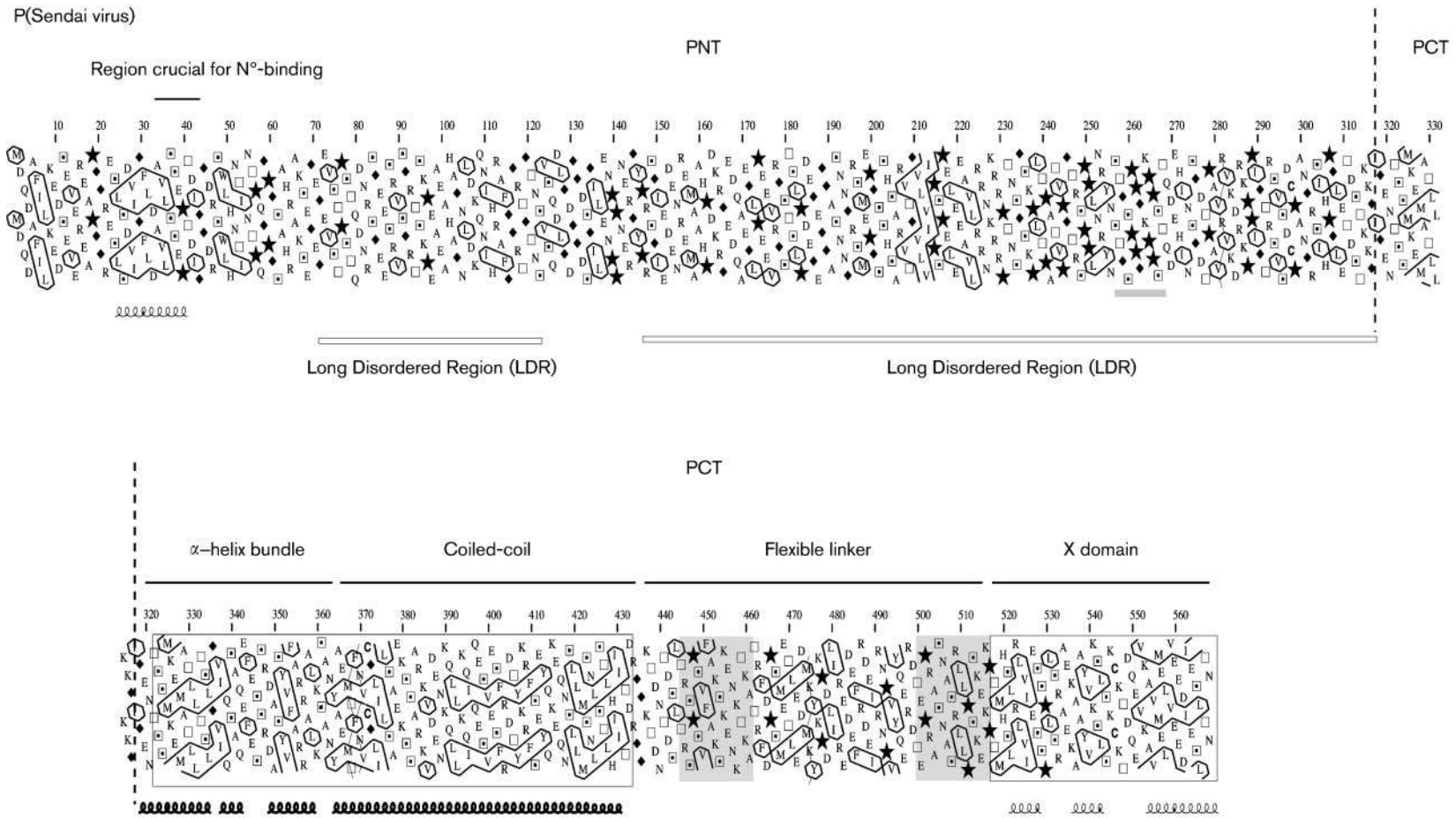


Fig. 4. HCA plot of SeV P. The conventions are the same as in Fig. 2. Low sequence complexity regions are underlined in light grey. Predicted or actual secondary structure elements as observed in the three-dimensional structure (Tarbouriech *et al.*, 2000b) are shown in regular or bold style, respectively. Regions enriched in disorder-promoting residues are shaded.

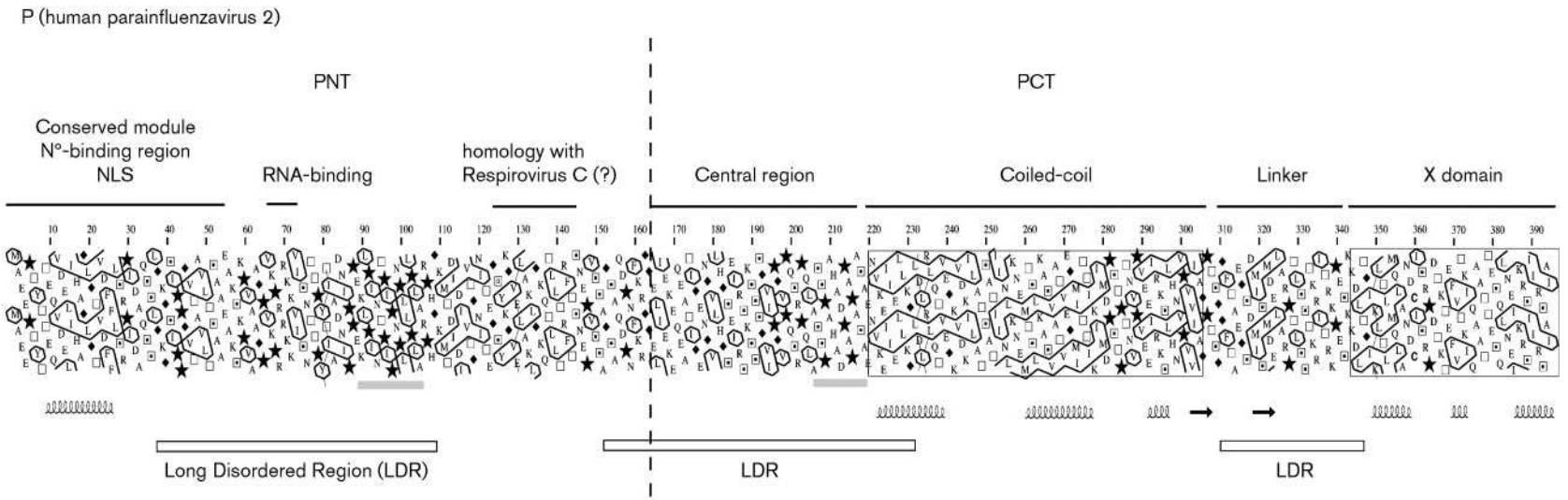


Fig. 5. HCA plot of hPIV-2 P. The conventions are the same as in Figs 2 and 4. Low sequence complexity regions are underlined in light grey. Predicted secondary structure elements are shown. Note that LDRs correspond to regions poor in hydrophobic clusters, with the first and second one being strikingly rich in proline residues.

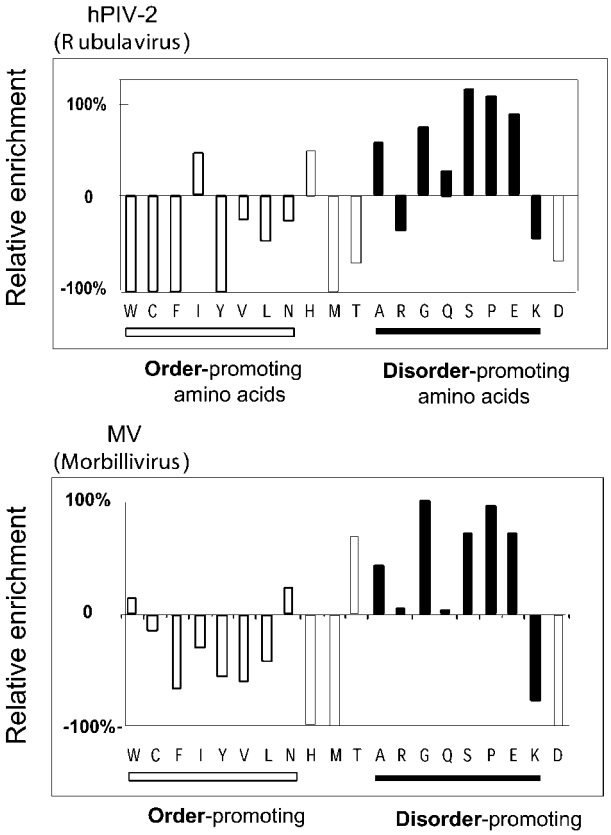


Fig. 6. Amino acid composition of the central region of *Rubulavirus* and *Morbillivirus* PCT. Deviation in sequence composition from globular proteins. Order-promoting and disorder-promoting amino acids are indicated by empty and black bars, respectively. Amino acids that are indifferently enriched or depleted in disordered regions of proteins are represented by empty bars with a thin contour.

common modular organization in all *Paramyxovirinae* P (Fig. 8). The identification of previously undetected modular regions is expected to guide functional studies. For instance, we discovered a homologous N-terminal module in the P and V proteins of *Rubulavirus*, *Henipavirus* and NDV. So far, the only identity shared by these V proteins was thought to concern their zinc-binding domain. The conserved module could be an interesting target for mutational studies aimed at elucidating the different mechanisms by which V counters interferon transduction (reviewed by Gotoh *et al.*, 2002).

Of note is that *Paramyxovirinae* L contains no predicted LDR (data not shown). However, this does not exclude the presence of disordered regions shorter than the 40 aa threshold of PONDR. Indeed, the presence of a flexible hinge region in *Morbillivirus* L (aa 1695–1717 of MV L) has been suggested on the basis of sequence variability (McIlhatton *et al.*, 1997). Enhanced green fluorescent protein could be inserted at this position without interfering with the function of L, which suggests that the C-terminal moiety, located downstream of the hinge, enjoys a certain degree of conformational independence (Duprex *et al.*, 2002). This hinge region contains no low sequence complexity segments, is not visible as an interruption of hydrophobic clusters using HCA and contains predicted secondary structure elements (data not shown). This suggests that some other short flexible regions of L might escape detection using the current prediction methods.

Although no direct biochemical evidence is available, the L protein is supposed to bear most enzymatic activities required for transcription and replication, such as RNA-dependent RNA polymerase (Svenda *et al.*, 1997) and 2′-*O*-methyltransferase (Ferron *et al.*, 2002). The absence of structural disorder in the L protein might be related to the fact that a precise protein scaffold is required for these enzymatic activities.

Experimental support for disorder predictions

In most cases, experimental support exists for our predictions of disorder, although they are not always recognized as such. In retrospect, studies carried out 20 years ago on

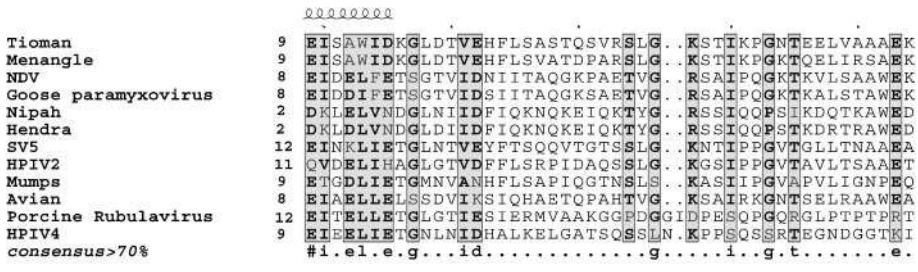


Fig. 7. Multiple sequence alignment of the N terminus of PNT in *Rubulavirus*, *Henipavirus* and *Avulavirus*. The consensus sequence (identity cut-off of >70 %) is shown under the multiple sequence alignment. Dots and residues shown in lower-case correspond to residues under and above the cut-off, respectively, while positions marked by # correspond to any of the N, D, Q or E residues. Residues corresponding to an identity of >70 % are boxed. For a given position, only residues homologous to the consensus are in bold. The front numbers correspond to the amino acid position in sequence. Dots above the alignment indicate intervals of 10 residues. Predicted secondary structure elements are shown above the alignment.

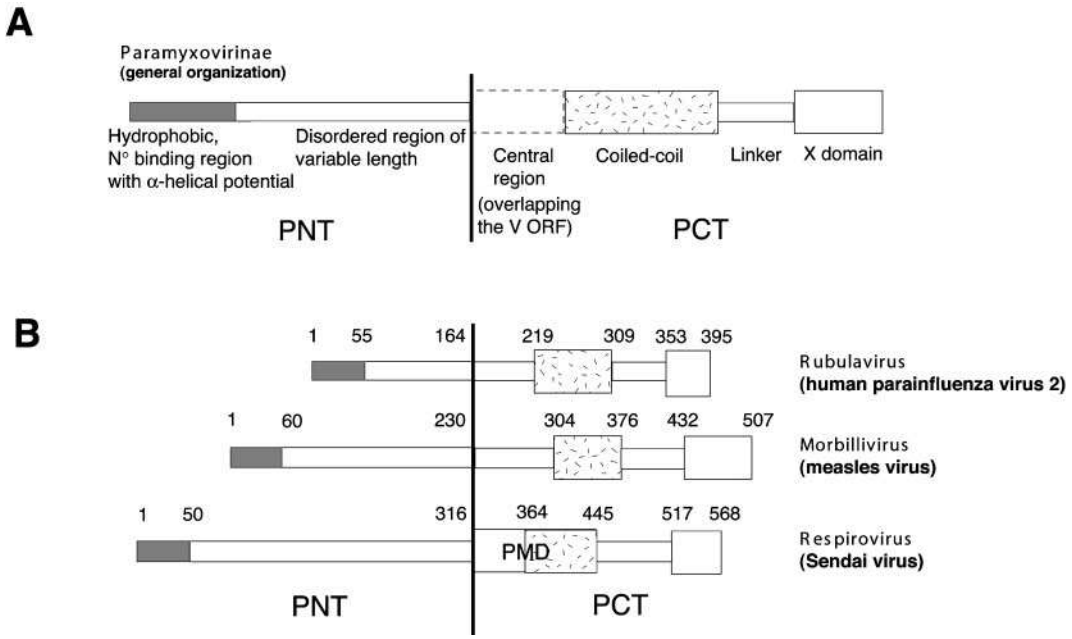


Fig. 8. Functional and structural organization of P in *Paramyxovirinae*. (A) General organization of *Paramyxovirinae* P. Globular and disordered regions are represented by large and narrow boxes, respectively. The region overlapping the V ORF is represented by dotted lines to indicate that it can be either disordered or structured (see lower panel). The line separating PNT and PCT is located as defined in Fig. 1. (B) Organization of P in the prototype member of each genus. In *Rubulavirus* and *Morbillivirus*, the central region, overlapping the V ORF, is disordered, whereas it is globular in the case of *Respirovirus*. The hydrophobic, N°-binding region within the PNT moiety is shaded.

SeV and NDV PNT indicate that they are in large part unstructured within the viral ribonucleoprotein complex. Indeed, SeV P bound to nucleocapsids is composed of a 40 kDa C-terminal core resistant to proteolysis, while the remaining N-terminal region (extending to at least aa 221 out of 320 for PNT) is hypersensitive to proteolysis (Chinchar & Portner, 1981a; Deshpande & Portner, 1985). Likewise, NDV P is composed of a core resistant to proteolysis and an acidic region hypersensitive to proteolysis (Chinchar & Portner, 1981b). From the present knowledge of the sequence of NDV P, we can conclude that the resistant core is PCT, while it is PNT that is degraded and thus mostly disordered. Interestingly, like *Rubulavirus*, NDV PNT is composed of the conserved module followed by a predicted disordered region (data not shown). However, one still cannot reliably conclude whether the conserved module is disordered or not, since a putative 5 kDa globular peptide (the expected molecular mass of the module) might escape detection using SDS-PAGE.

Likewise, the hypersensitivity to proteolysis of *Paramyxovirinae* N_{TAIL} clearly suggests that it is mostly unstructured (Heggeness *et al.*, 1981; Karlin *et al.*, 2002a). Indeed, while this computational analysis was in progress, the intrinsic disorder of MV N_{TAIL} has been experimentally assessed (Longhi *et al.*, 2003). In the same vein, the presence of a flexible linker in P, for which we could not reach the same degree of confidence than for our other predictions, is supported by the protease sensitivity observed within MV (Longhi *et al.*, 2003) and SeV (Tarbouriech *et al.*, 2000a) PCT and by spectroscopic studies on SeV PCT (Marion *et al.*, 2001).

Functional implications of structural disorder in the replicative complex

Beyond *Paramyxovirinae*, we found that structural disorder is a widespread property of *Mononegavirales* N and P, suggesting a functional significance. In particular, since unstructured regions are considerably more extended than globular ones, the very long reach of N and P might enable them to act as linkers. Moreover, the presence of flexible regions at the surface of the viral nucleocapsid enables transient interactions with several, structurally distinct partners (Dunker *et al.*, 1998, 2001; Dunker & Obradovic, 2001; Liu *et al.*, 2002; Uversky, 2002a; Wright & Dyson, 1999). The pattern of interactions of N_{TAIL} and PNT (Curran *et al.*, 1994) are consistent with this hypothesis. Indeed, MV N_{TAIL} takes part in numerous interactions with different protein partners, including P (both within N^{o}–P and N^{NUC}–P), polymerase complex P–L, interferon regulatory factor 3 (tenOever *et al.*, 2002) and heat-shock protein Hsp72 (which modulates the level of viral RNA synthesis) (Zhang *et al.*, 2002). Likewise, PNT interacts not only with N^{o} and L but also with several cellular proteins (Liston *et al.*, 1995). Therefore, disordered regions of N and

P are involved in manifold interactions essential for RNA transcription and replication.

Induced folding in N and P?

The experimental evidence mentioned above suggests that some disordered regions we describe are unstructured *in vitro* not only as isolated domains but also in the context of full-length proteins. SeV and NDV PNT are mostly unstructured even within P bound to nucleocapsids (Chinchar & Portner, 1981a, b; Deshpande & Portner, 1985), while the linker region of SeV P is unstructured in the context of PCT (Tarbouriech *et al.*, 2000a). However, the possibility that these regions may fold *in vivo* in the presence of appropriate solute concentrations or of their physiological partner(s) (a process called 'induced folding') (Dyson & Wright, 2002; Uversky, 2002b) cannot be ruled out. While this manuscript was in preparation, we found that MV N_{TAIL} undergoes such an unstructured-to-structured transition upon binding to PCT (Longhi *et al.*, 2003). With respect to PNT, we note that the extreme N terminus of *Paramyxovirinae* P (especially the conserved module in *Rubulavirus*), which is involved in binding to N^o (Curran *et al.*, 1995b; Nishio *et al.*, 1996; Precious *et al.*, 1995; Shaji & Shaila, 1999; Tober *et al.*, 1998), contains hydrophobic clusters associated with α-helical potential (Figs. 4, 5, 7 and 8). Such an α-helix could be actually induced in MV PNT in the presence of the solvent trifluoroethanol, which is used to unveil disordered regions with a propensity to undergo induced folding (Karlin *et al.*, 2002b). Another region likely to undergo induced folding upon binding its target is the arginine-rich, RNA-binding region of *Rubulavirus* PNT, reminiscent of the arginine-rich motif in the disordered bacteriophage anti-termination protein N, which folds upon binding to RNA (Mogridge *et al.*, 1998).

Phosphorylation occurs on disordered regions of N and P

The role of phosphorylation of *Paramyxoviridae* N and P is still unclear (Lamb & Kolakofsky, 2001). Remarkably, phosphorylation of SeV N occurs within N_{TAIL} and that of *Morbillivirus* and *Respirovirus* P occurs within PNT (Byrappa & Gupta, 1999; Byrappa *et al.*, 1996; Das *et al.*, 1995; Hsu & Kingsbury, 1982; Jonscher & Yates, 1997; Vidal *et al.*, 1988). Further studies will tell whether the occurrence of phosphorylation on disordered regions of proteins is coincidental or whether it is a widespread property, as suggested by Dunker *et al.* (2002a). Interestingly, Zetina (2001) has recently observed that a number of intrinsically disordered proteins share a common motif, called the 'helix-unfolding motif', which might control the unfolding of intrinsically disordered proteins in response to cellular events, perhaps by means of phosphorylation. Although no such motif is found in *Paramyxovirinae* N or P, a hint that phosphorylation of N_{TAIL} and PNT might modulate their structural state comes from data available in the literature. Indeed, MV N^o and N^{NUC}, which have different conformations (Gombart *et al.*, 1995), are phosphorylated with a different pattern (Gombart *et al.*, 1995). In the same vein, Byrappa *et al.* (1996) showed that all potential phosphorylation sites of SeV PNT are equally accessible to kinases, whereas once phosphorylated they have a different accessibility to phosphatases, thus suggesting that phosphorylation may affect the conformation of PNT. However, much caution is required because no study has so far been able to elucidate the tantalizing function(s) of phosphorylation of P.

Preliminary insights from the comparison of HCA and PONDR

The present study shows that HCA is an invaluable tool, very intuitive in qualitatively highlighting disordered regions, owing to its easy visualization of periodical, hydrophobic features directly on the primary sequence. Although its usefulness in elucidating the modular organization of proteins (which implies recognition of disordered linkers) has been proved (Callebaut *et al.*, 1997), HCA is not widely used for the specific purpose of predicting disordered regions and is in fact not mentioned in the reviews on disorder referenced herein. We show that an HCA plot can serve as a convenient support to plot other information, such as PONDR LDRs, secondary structure predictions and low sequence complexity segments. Comparison of HCA plots and PONDR LDRs shows that, although by no means absolute, there seems to be an inverse correlation between PONDR predictions of disorder and the presence of hydrophobic clusters (Figs. 2, 4 and 5). We hope that our study will open the way to a more quantitative comparison of disorder predictors, and perhaps to their future refinement, a point crucial for current structural genomics projects.

Disorder and lack of predicted secondary structure

Another hallmark of disordered regions, i.e. their lack of predicted secondary structure, has not been the subject of systematic studies at the sequence level. Interestingly, however, Liu *et al.* (2002) recently showed the wide occurrence in proteins of long (>70 aa) regions with little or no predicted secondary structure, called NORS (for 'no ordered regular structure'). NORS are defined as protein regions that comprise more than 70 residues with less than 12 % predicted secondary structure and have at least one segment >10 aa predicted to be accessible to the solvent, as estimated by PHD (Rost, 1996). Liu *et al.* (2002) found that NORS overlap only partially with LDRs (i.e. some NORS are structured) and their structural significance is not well established yet. Analysis of all disordered regions >70 aa identified herein (N_{TAIL}, PNT and the central regions) reveals that they fulfil only the first criterion, i.e. they have less than 12 % predicted secondary structure. However, the accessibility prediction values of PHD for these regions, including those the disorder of which has been biochemically proved (Karlin *et al.*, 2002b; Longhi *et al.*, 2003), are under the threshold of reliability of this

program. Although preliminary, this suggests that the methods of accessibility prediction might not give reliable results on disordered regions, causing some of these disordered regions to escape the classification as NORS.

Implications for structural studies of N and P

The identification of disordered regions within proteins should avoid numerous fruitless attempts to crystallize proteins (or protein domains) containing such large unstructured regions. Furthermore, some of these regions are probably unstructured even when complexed with their partner(s), thus preventing crystallization. The information on the modular organization of PCT derived from the present study led to the resolution of the crystal structure of the extreme C-terminal domain (XD) of MV P (Johansson *et al.*, 2003), thus validating the reliability of the prediction approach described in this paper.

Once the structure of globular domains of N and P has been solved, structural advances in the study of the replicative complex will either (i) need removal of several dispensable, unstructured regions for crystallization of protein complexes, a trial-and-error process likely to be very time-consuming, (ii) rely on techniques that can deal with both ordered and disordered regions of proteins (such as small angle X-ray scattering) and (iii) rely on biocomputing methods to identify further functional regions in poorly conserved, unstructured parts of N and P.

ACKNOWLEDGEMENTS

This work was supported by a grant from the Fondation pour la Recherche Médicale (FRM) to D. K. This study has been carried out with financial support from the Commission of the European Communities, specific RTD programme 'Quality of Life and Management of Living Resources', QLK2-CT2001-01225, 'Towards the design of new potent antiviral drugs: structure–function analysis of *Paramyxoviridae* RNA polymerase'. It does not necessarily reflect its views and in no way anticipates the Commission's future policy in this area. We wish to thank K. Dunker, J. Curran, L. Roux and D. Gerlier for useful remarks. We also thank B. Henrissat and I. Callebaut for critical advice on HCA.

REFERENCES

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.

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

Bankamp, B., Horikami, S. M., Thompson, P. D., Huber, M., Billeter, M. & Moyer, S. A. (1996). Domains of the measles virus N protein required for binding to P protein and self-assembly. *Virology* **216**, 272–277.

Barr, J., Chambers, P., Pringle, C. R. & Easton, A. J. (1991). Sequence of the major nucleocapsid protein gene of pneumonia virus of mice: sequence comparisons suggest structural homology between nucleocapsid proteins of pneumoviruses, paramyxoviruses, rhabdoviruses and filoviruses. *J Gen Virol* **72**, 677–685.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., Rapp, B. A. & Wheeler, D. L. (2002). GenBank. *Nucleic Acids Res* **30**, 17–20.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). The protein data bank. *Nucleic Acids Res* **28**, 235–242.

Brown, C. J., Takayama, S., Campen, A. M., Vise, P., Marshall, T. W., Oldfield, C. J., Williams, C. J. & Dunker, A. K. (2002). Evolutionary rate heterogeneity in proteins with long disordered regions. *J Mol Evol* **55**, 104–110.

Buchholz, C. J., Retzler, C., Homann, H. E. & Neubert, W. J. (1994). The carboxy-terminal domain of Sendai virus nucleocapsid protein is involved in complex formation between phosphoprotein and nucleocapsid-like particles. *Virology* **204**, 770–776.

Byrappa, S. & Gupta, K. C. (1999). Human parainfluenza virus type 1 phosphoprotein is constitutively phosphorylated at Ser-120 and Ser-184. *J Gen Virol* **80**, 1199–1209.

Byrappa, S., Pan, Y. B. & Gupta, K. C. (1996). Sendai virus P protein is constitutively phosphorylated at serine249: high phosphorylation potential of the P protein. *Virology* **216**, 228–234.

Callebaut, I., Labesse, G., Durand, P., Poupon, A., Canard, L., Chomilier, J., Henrissat, B. & Mornon, J. P. (1997). Deciphering protein sequence information through hydrophobic cluster analysis (HCA): current status and perspectives. *Cell Mol Life Sci* **53**, 621–645.

Chinchar, V. G. & Portner, A. (1981a). Functions of Sendai virus nucleocapsid polypeptides: enzymatic activities in nucleocapsids following cleavage of polypeptide P by *Staphylococcus aureus* protease V8. *Virology* **109**, 59–71.

Chinchar, V. G. & Portner, A. (1981b). Inhibition of RNA synthesis following proteolytic cleavage of Newcastle disease virus P protein. *Virology* **115**, 192–202.

Curran, J. & Kolakofsky, D. (1999). Replication of paramyxoviruses. *Adv Virus Res* **54**, 403–422.

Curran, J., Homann, H., Buchholz, C., Rochat, S., Neubert, W. & Kolakofsky, D. (1993). The hypervariable C-terminal tail of the Sendai paramyxovirus nucleocapsid protein is required for template function but not for RNA encapsidation. *J Virol* **67**, 4358–4364.

Curran, J., Pelet, T. & Kolakofsky, D. (1994). An acidic activation-like domain of the Sendai virus P protein is required for RNA synthesis and encapsidation. *Virology* **202**, 875–884.

Curran, J., Boeck, R., Lin-Marq, N., Lupas, A. & Kolakofsky, D. (1995a). Paramyxovirus phosphoproteins form homotrimers as determined by an epitope dilution assay, via predicted coiled coils. *Virology* **214**, 139–149.

Curran, J., Marq, J. B. & Kolakofsky, D. (1995b). An N-terminal domain of the Sendai paramyxovirus P protein acts as a chaperone for the NP protein during the nascent chain assembly step of genome replication. *J Virol* **69**, 849–855.

Das, T., Schuster, A., Schneider-Schaulies, S. & Banerjee, A. K. (1995). Involvement of cellular casein kinase II in the phosphorylation of measles virus P protein: identification of phosphorylation sites. *Virology* **211**, 218–226.

Deshpande, K. L. & Portner, A. (1985). Monoclonal antibodies to the P protein of Sendai virus define its structure and role in transcription. *Virology* **140**, 125–134.

Dunker, A. K. & Obradovic, Z. (2001). The protein trinity: linking function and disorder. *Nat Biotechnol* **19**, 805–806.

Dunker, A. K., Garner, E., Guilliot, S., Romero, P., Albrecht, K., Hart, J., Obradovic, Z., Kissinger, C. & Villafranca, J. E. (1998).

Protein disorder and the evolution of molecular recognition: theory, predictions and observations. *Pac Symp Biocomput*, 473–484.

Dunker, A. K., Lawson, J. D., Brown, C. J. & 17 other authors (2001). Intrinsically disordered protein. *J Mol Graph Model* **19**, 26–59.

Dunker, A. K., Brown, C. J., Lawson, J. D., Iakoucheva, L. M. & Obradovic, Z. (2002a). Intrinsic disorder and protein function. *Biochemistry* **41**, 6573–6582.

Dunker, A. K., Brown, C. J. & Obradovic, Z. (2002b). Identification and functions of usefully disordered proteins. *Adv Protein Chem* **62**, 25–49.

Duprex, W. P., Collins, F. M. & Rima, B. K. (2002). Modulating the function of the measles virus RNA-dependent RNA polymerase by insertion of green fluorescent protein into the open reading frame. *J Virol* **76**, 7322–7328.

Dyson, H. J. & Wright, P. E. (2002). Coupling of folding and binding for unstructured proteins. *Curr Opin Struct Biol* **12**, 54–60.

Ferron, F., Longhi, S., Henrissat, B. & Canard, B. (2002). Viral RNA-polymerases: a predicted 2′-*O*-ribose methyltransferase domain shared by all *Mononegavirales*. *Trends Biochem Sci* **27**, 222–224.

Galtier, N., Gouy, M. & Gautier, C. (1996). SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. *Comput Appl Biosci* **12**, 543–548.

Giraudon, P., Jacquier, M. F. & Wild, T. F. (1988). Antigenic analysis of African measles virus field isolates: identification and localisation of one conserved and two variable epitope sites on the NP protein. *Virus Res* **10**, 137–152.

Gombart, A. F., Hirano, A. & Wong, T. C. (1995). Nucleoprotein phosphorylated on both serine and threonine is preferentially assembled into the nucleocapsids of measles virus. *Virus Res* **37**, 63–73.

Gotoh, B., Komatsu, T., Takeuchi, K. & Yokoo, J. (2002). Paramyxovirus strategies for evading the interferon response. *Rev Med Virol* **12**, 337–357.

Gouet, P., Courcelle, E., Stuart, D. I. & Metoz, F. (1999). ESPript: analysis of multiple sequence alignments in PostScript. *Bioinformatics* **15**, 305–308.

Harty, R. N. & Palese, P. (1995). Measles virus phosphoprotein (P) requires the NH_2- and COOH-terminal domains for interactions with the nucleoprotein (N) but only the COOH terminus for interactions with itself. *J Gen Virol* **76**, 2863–2867.

Heggeness, M. H., Scheid, A. & Choppin, P. W. (1981). The relationship of conformational changes in the Sendai virus nucleocapsid to proteolytic cleavage of the NP polypeptide. *Virology* **114**, 555–562.

Holmes, D. E. & Moyer, S. A. (2002). The phosphoprotein (P) binding site resides in the N terminus of the L polymerase subunit of sendai virus. *J Virol* **76**, 3078–3083.

Hsu, C. H. & Kingsbury, D. W. (1982). Topography of phosphate residues in Sendai virus proteins. *Virology* **120**, 225–234.

Johansson, K., Bourhis, J. M., Campanacci, V., Cambillau, C., Canard, B. & Longhi, S. (2003). Crystal structure of the measles virus phosphoprotein domain responsible for the induced folding of the C-terminal domain of the nucleoprotein. *J Biol Chem* (in press).

Jonscher, K. R., Yates, J. R., III (1997). Matrix-assisted laser desorption ionization/quadrupole ion trap mass spectrometry of peptides. Application to the localization of phosphorylation sites on the P protein from Sendai virus. *J Biol Chem* **272**, 1735–1741.

Karlin, D., Longhi, S. & Canard, B. (2002a). Substitution of two residues in the measles virus nucleoprotein results in an impaired self-association. *Virology* **302**, 420–432.

Karlin, D., Longhi, S., Receveur, V. & Canard, B. (2002b). The N-terminal domain of the phosphoprotein of morbilliviruses belongs to the natively unfolded class of proteins. *Virology* **296**, 251–262.

Lamb, R. A. & Kolakofsky, D. (2001). *Paramyxoviridae*: the viruses and their replication. In *Fields Virology*, 4th edn, pp. 1305–1340. Edited by B. N. Fields, D. M. Knipe & P. M. Howley. Philadelphia, PA: Lippincott Williams & Wilkins.

Leulliot, N. & Varani, G. (2001). Current topics in RNA–protein recognition: control of specificity and biological function through induced fit and conformational capture. *Biochemistry* **40**, 7947–7956.

Li, X., Romero, P., Rani, M., Dunker, A. K. & Obradovic, Z. (1999). Predicting protein disorder for N-, C- and internal regions. *Genome Inform Ser Workshop Genome Inform* **10**, 30–40.

Lin, G. Y., Paterson, R. G. & Lamb, R. A. (1997). The RNA binding region of the paramyxovirus SV5 V and P proteins. *Virology* **238**, 460–469.

Liston, P., DiFlumeri, C. & Briedis, D. J. (1995). Protein interactions entered into by the measles virus P, V, and C proteins. *Virus Res* **38**, 241–259.

Liu, J., Tan, H. & Rost, B. (2002). Loopy proteins appear conserved in evolution. *J Mol Biol* **322**, 53–64.

Longhi, S., Receveur-Brechot, V., Karlin, D., Johansson, K., Darbon, H., Bhella, D., Yeo, R., Finet, S. & Canard, B. (2003). The C-terminal domain of the measles virus nucleoprotein is intrinsically disordered and folds upon binding to the C-terminal moiety of the phosphoprotein. *J Biol Chem* **278**, 18638–18648.

Malur, A. G., Choudhary, S. K., De, B. P. & Banerjee, A. K. (2002). Role of a highly conserved NH_2-terminal domain of the human parainfluenza virus type 3 RNA polymerase. *J Virol* **76**, 8101–8109.

Marion, D., Tarbouriech, N., Ruigrok, R. W., Burmeister, W. P. & Blanchard, L. (2001). Assignment of the 1H, 15N and 13C resonances of the nucleocapsid-binding domain of the Sendai virus phosphoprotein. *J Biomol NMR* **21**, 75–76.

McGuffin, L. J., Bryson, K. & Jones, D. T. (2000). The PSIPRED protein structure prediction server. *Bioinformatics* **16**, 404–405.

McIlhatton, M. A., Curran, M. D. & Rima, B. K. (1997). Nucleotide sequence analysis of the large (L) genes of phocine distemper virus and canine distemper virus (corrected sequence). *J Gen Virol* **78**, 571–576.

Mogridge, J., Legault, P., Li, J., Van Oene, M. D., Kay, L. E. & Greenblatt, J. (1998). Independent ligand-induced folding of the RNA-binding domain and two functionally distinct antitermination regions in the phage lambda N protein. *Mol Cell* **1**, 265–275.

Nishio, M., Tsurudome, M., Kawano, M., Watanabe, N., Ohgimoto, S., Ito, M., Komada, H. & Ito, Y. (1996). Interaction between nucleocapsid protein (NP) and phosphoprotein (P) of human parainfluenza virus type 2: one of the two NP binding sites on P is essential for granule formation. *J Gen Virol* **77**, 2457–2463.

Nishio, M., Tsurudome, M., Ito, M., Kawano, M., Kusagawa, S., Komada, H. & Ito, Y. (1999). Mapping of domains on the human parainfluenza virus type 2 nucleocapsid protein (NP) required for NP–phosphoprotein or NP–NP interaction. *J Gen Virol* **80**, 2017–2022.

Parks, G. D. (1994). Mapping of a region of the paramyxovirus L protein required for the formation of a stable complex with the viral phosphoprotein P. *J Virol* **68**, 4862–4872.

Paterson, R. G., Leser, G. P., Shaughnessy, M. A. & Lamb, R. A. (1995). The paramyxovirus SV5 V protein binds two atoms of zinc and is a structural component of virions. *Virology* **208**, 121–131.

Poch, O., Blumberg, B. M., Bougueleret, L. & Tordo, N. (1990). Sequence comparison of five polymerases (L proteins) of unsegmented negative-strand RNA viruses: theoretical assignment of functional domains. *J Gen Virol* **71**, 1153–1162.

Precious, B., Young, D. F., Bermingham, A., Fearns, R., Ryan, M. & Randall, R. E. (1995). Inducible expression of the P, V, and NP genes of the paramyxovirus simian virus 5 in cell lines and an examination of NP–P and NP–V interactions. *J Virol* **69**, 8001–8010.

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 IEEE International Conference on Neural Networks*, vol. **1**, 90–95.

Romero, P., Obradovic, Z., Li, X., Garner, E. C., Brown, C. J. & Dunker, A. K. (2001). Sequence complexity of disordered proteins. *Proteins* **42**, 38–48.

Rost, B. (1996). PHD: predicting one-dimensional protein structure by profile-based neural networks. *Methods Enzymol* **266**, 525–539.

Ryan, K. W., Morgan, E. M. & Portner, A. (1991). Two noncontiguous regions of Sendai virus P protein combine to form a single nucleocapsid binding domain. *Virology* **180**, 126–134.

Sanchez, A., Khan, A. S., Zaki, S. R., Nabel, G. J., Ksiazek, T. G. & Peters, C. J. (2001). *Filoviridae*: Marburg and Ebola viruses. In *Fields Virology*, 4th edn, pp. 1279–1304. Edited by B. N. Fields, D. M. Knipe & P. M. Howley. Philadelphia, PA: Lippincott Williams & Wilkins.

Schoehn, G., Iseni, F., Mavrakis, M., Blondel, D. & Ruigrok, R. W. (2001). Structure of recombinant rabies virus nucleoprotein–RNA complex and identification of the phosphoprotein binding site. *J Virol* **75**, 490–498.

Shaji, D. & Shaila, M. S. (1999). Domains of Rinderpest virus phosphoprotein involved in interaction with itself and the nucleocapsid protein. *Virology* **258**, 415–424.

Smallwood, S., Ryan, K. W. & Moyer, S. A. (1994). Deletion analysis defines a carboxyl-proximal region of Sendai virus P protein that binds to the polymerase L protein. *Virology* **202**, 154–163.

Svenda, M., Berg, M., Moreno-Lopez, J. & Linne, T. (1997). Analysis of the large (L) protein gene of the porcine rubulavirus LPMV: identification of possible functional domains. *Virus Res* **48**, 57–70.

Tarbouriech, N., Curran, J., Ebel, C., Ruigrok, R. W. & Burmeister, W. P. (2000a). On the domain structure and the polymerization state of the sendai virus P protein. *Virology* **266**, 99–109.

Tarbouriech, N., Curran, J., Ruigrok, R. W. & Burmeister, W. P. (2000b). Tetrameric coiled coil domain of Sendai virus phosphoprotein. *Nat Struct Biol* **7**, 777–781.

tenOever, B. R., Servant, M. J., Grandvaux, N., Lin, R. & Hiscott, J. (2002). Recognition of the measles virus nucleocapsid as a mechanism of IRF-3 activation. *J Virol* **76**, 3659–3669; erratum **76**, 6413.

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.

Tober, C., Seufert, M., Schneider, H., Billeter, M. A., Johnston, I. C., Niewiesk, S., ter Meulen, V. & Schneider-Schaulies, S. (1998). Expression of measles virus V protein is associated with pathogenicity and control of viral RNA synthesis. *J Virol* **72**, 8124–8132.

Tompa, P. (2002). Intrinsically unstructured proteins. *Trends Biochem Sci* **27**, 527–533.

Uversky, V. N. (2002a). Natively unfolded proteins: a point where biology waits for physics. *Protein Sci* **11**, 739–756.

Uversky, V. N. (2002b). What does it mean to be natively unfolded? *Eur J Biochem* **269**, 2–12.

Uversky, V. N., Gillespie, J. R. & Fink, A. L. (2000). Why are 'natively unfolded' proteins unstructured under physiologic conditions? *Proteins* **41**, 415–427.

Vidal, S., Curran, J., Orvell, C. & Kolakofsky, D. (1988). Mapping of monoclonal antibodies to the Sendai virus P protein and the location of its phosphates. *J Virol* **62**, 2200–2203.

Watanabe, N., Kawano, M., Tsurudome, M., Kusagawa, S., Nishio, M., Komada, H., Shima, T. & Ito, Y. (1996). Identification of the sequences responsible for nuclear targeting of the V protein of human parainfluenza virus type 2. *J Gen Virol* **77**, 327–338.

Williams, R. M., Obradovi, Z., Mathura, V., Braun, W., Garner, E. C., Young, J., Takayama, S., Brown, C. J. & Dunker, A. K. (2001). The protein non-folding problem: amino acid determinants of intrinsic order and disorder. *Pac Symp Biocomput,* 89–100.

Wootton, J. C. (1994). Non-globular domains in protein sequences: automated segmentation using complexity measures. *Comput Chem* **18**, 269–285.

Wright, P. E. & Dyson, H. J. (1999). Intrinsically unstructured proteins: re-assessing the protein structure–function paradigm. *J Mol Biol* **293**, 321–331.

Zetina, C. R. (2001). A conserved helix-unfolding motif in the naturally unfolded proteins. *Proteins* **44**, 479–483.

Zhang, X., Glendening, C., Linke, H., Parks, C. L., Brooks, C., Udem, S. A. & Oglesbee, M. (2002). Identification and characterization of a regulatory domain on the carboxyl terminus of the measles virus nucleocapsid protein. *J Virol* **76**, 8737–8746.

-Commentaires

Dans cet article, nous nous sommes intéressés aux deux autres protéines du complexe de réplication des *Paramyxoviridae*, à savoir les protéines N et P. Les protéines N et P du virus de la rougeole sont étudiées depuis longtemps au laboratoire. Ces études avaient montré que le domaine C-terminal de la nucléoprotéine et le domaine N-terminal de la phosphoprotéine sont intrinsèquement désordonnés (Karlin et al. 2002; Longhi et al. 2003; Bourhis et al. 2004). Dans cette étude nous avons étendu ces résultats aux nucléoprotéines et phosphoprotéines des membres de la famille des *Paramyxovirinae*. L'utilisation de la méthode HCA, combinée aux autres méthodes d'étude du désordre (*cf.* article 1), s'est révélée la clé de voûte de notre système d'analyse. Nous avons montré qu'il est possible en croisant les données structurales sur les tracés HCA de définir des motifs caractéristiques identifiables en l'absence de toute similarité de séquence. Actuellement, aucune analyse automatisée n'est disponible. L'analyse structurale systématique des structures des grandes familles de repliement, couplée à l'analyse des tracés HCA, devrait permettre de définir des « signatures » HCA caractéristiques et identifiables pour les différents motifs structuraux. Ces motifs HCA pourraient servir par la suite comme facteur de pondération dans des techniques de reconnaissance de repliement tels que le « threading ». Ce facteur devrait permettre entre autres de pouvoir rendre les techniques de « threading » plus sensibles.

Dans cette étude nous avons aussi montré que l'utilisation de HCA permet de détecter au sein de domaines désordonnés des régions pouvant subir un repliement induit suite à l'interaction avec un partenaire (Longhi et al. 2003; Mavrakis et al. 2004). Ce dernier point fait l'objet d'une collaboration avec Isabelle Callebault, qui a développé la méthode et qui est spécialiste dans l'interprétation des tracés.

1.2. Les Hantavirus

1.2.1. Généralités

Les Hantavirus appartiennent à la famille des *Bunyaviridae* et sont responsables chez l'homme de fièvres hémorragiques avec syndrome rénal (Asie et Europe), ou pulmonaire (Amérique). A la différence des autres *Bunyaviridae*, ils ne sont pas transmis par l'intermédiaire d'insectes mais ont pour réservoir des rongeurs sauvages.
Les Hantavirus sont des virus enveloppés dont le génome est constitué d'un ARN simple brin tri-segmenté, à polarité négative.

1.2.2. Organisation de la particule virale et du génome des *Bunyaviridae*

Les virions sont sphériques ou ovales, avec un diamètre de 90 à 120 nm, et contiennent trois nucléocapsides circulaires, chacune constituée d'un segment d'ARN encapsidé par la nucléoprotéine (N), auquel sont liées des polymérase. Les segments d'ARN sont désignés L (large), M (medium), et S (small). Le segment L comprend environs 6 Kb et code pour l'ARN polymérase ARN-dépendante. Le segment M, d'environ 3,6 Kb, code pour les glycoprotéines G1 et G2 qui forment une structure en damier à la surface de l'enveloppe lipidique. Le segment S, de longueur comprise ente 1,7 et 2,1 Kb, code pour la nucléoprotéine. Les trois segments ont des séquences conservées et complémentaires en 3' et 5'. Ces séquences sont conservées au sein des virus de même genre, mais diffèrent d'un genre à l'autre (Schmaljohn C and JW. 2001). L'appariement de ces séquences terminales permet la cyclisation de l'ARN. Le complexe N RNA forme des nucléocapsides de structure hélicoïdale (Pettersson and von Bonsdorff 1975; Obijeski et al. 1976a; Obijeski et al. 1976b; Raju and Kolakofsky 1989).

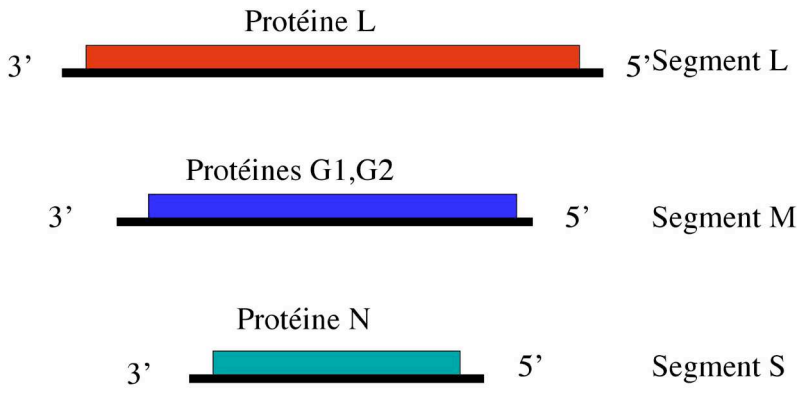


Figure 27 : Organisation du génome des *Bunyaviridae*.

Pendant ma thèse je me suis intéressé à l'analyse bioinformatique de la nucléoprotéine de hantaan virus, dans le but d'identifier les sites potentiels de liaison à l'ARN et de repliement induit. Ce travail est présenté dans l'article suivant.

ARTICLE 5

Essential amino acids of the hantaan virus N protein in its interaction with RNA

William Severson, Xiaolin Xu, Michaela Kuhn, Nina Senutovitch, Mercy Thokala, François Ferron, Sonia Longhi, Bruno Canard and Colleen B. Jonsson

Article soumis à Journal of Virology

ESSENTIAL AMINO ACIDS OF THE HANTAAN VIRUS N PROTEIN IN ITS INTERACTION WITH RNA

William Severson[1], Xiaolin Xu[2], Michaela Kuhn[3],

Nina Senutovitch[4], Mercy Thokala[5], François Ferron[6], Sonia Longhi[6],

Bruno Canard[6] and Colleen B. Jonsson[7*]

Department of Chemistry and Biochemistry, New Mexico State University, Las Cruces, NM 88003, Architecture et Fonction des Macromolécules Biologiques, UMR 6098 CNRS et Université Aix-Marseille I et II, ESIL, Campus de Luminy, 163 Avenue de Luminy, case 925, 13288 Marseille Cedex 09, France, Department of Biochemistry and Molecular Biology, Southern Research Institute, Birmingham, AL 35205

Running Title: RNA Binding Domain of the Hantaan N protein

*Correspondent Footnote:
Dr. Colleen B. Jonsson
Department of Biochemistry and Molecular Biology
2000 9th Avenue South
Southern Research Institute, Birmingham, AL 35205
Phone: 205-581-2681, Fax: 205-581-2093, Email: jonsson@sri.org

---

[1] Department of Entomology, Plant Pathology and Weed Science, P.O. Box 30003, MSC 3BE, New Mexico State University, Las Cruces, NM 88003
[2] Department of Medical Microbiology and Immunology, University of Alberta, Canada
[3] Diplomand, Institute of Microbiology and Virology, University of Witten/Herdeck, Stockumer Str. 10, 58448 Witten, GERMANY
[4] NIH RISE Undergraduate Program, Department of Biology, New Mexico State University, Las Cruces, NM 88003
[5] Graduate Program in Biochemistry, Department of Entomology Plant Pathology and Weed Science, P.O. Box 30003, MSC 3BE, New Mexico State University, Las Cruces, NM 88003
[6] Architecture et Fonction des Macromolécules Biologiques,UMR 6098 CNRS et Université Aix-Marseille I et II, ESIL, Campus de Luminy, 163 Avenue de Luminy, case 925, 13288 Marseille Cedex 09, France
[7] Department of Biochemistry and Molecular Biology, 2000 9th Ave S, Southern Research Institute, Birmingham, AL 35205

ABSTRACT

The nucleocapsid (N) protein of hantavirus encapsidates viral genomic and antigenomic RNAs. Previously, deletion mapping identified a central, conserved region (amino acids 175- 217) within the Hantaan virus (HTNV) N protein that interacts with a high affinity with these viral RNAs. To further define the boundaries of the RNA binding domain (RBD), several peptides were synthesized and examined for their ability to bind full-length S-segment vRNA. Peptide 195-217 retained ninety-four percent of the vRNA as compared to HTNV N protein, while peptides 175-186 and 205-217 bound only one percent of the vRNA. To further explore which residues were essential for binding vRNA, we made a comprehensive mutational analysis of the amino acids in the RBD. Single and double ala substitutions were constructed for 18 amino acids from amino acids 175-217 in the full-length N protein. In addition, ala substitutions were made for the three arg residues in peptide 185-217. Analysis of protein-RNA interactions by electrophoretic mobility shift assays implicated E192, Y206 and S217 to be important for binding. Chemical modification experiments showed that lys, but not arg or cys residues, contribute to RNA binding. Overall, these data implicate lys residues dispersed from amino acids 175-429 of the protein and three amino acids located in the RBD as essential for RNA binding.

INTRODUCTION

Hantaviruses, classified as emerging viruses, can cause two diseases when transmitted to humans, hemorrhagic fever with renal syndrome (HFRS) and hantavirus pulmonary syndrome (HPS) (20, 23). The prototype virus for HFRS is the Hantaan virus (HTNV) and for HPS is the Sin Nombre virus (SNV). These viruses are often referred to as Old World and New World viruses, respectively, because of their geographical distribution, which is limited by the ecological habitats of their rodent reservoirs (27). *Hantaviruses*, a genus within the *Bunyaviridae,* have a negative-strand, tri-partite genome that encodes an RNA dependent RNA polymerase (RdRp) (L-segment), the nucleocapsid (N) protein (S-segment), and the G1 and G2 glycoproteins (M-segment) (28). G1 and G2, which are post-translationally processed through the ER and the Golgi, are presented on the external face of the virion and facilitate entry of the virus into the host cell. The three genomic viral RNAs (vRNAs) are complexed with N and possibly the RdRp in the virion. These ribonucleocapsids, not naked viral RNA, serve as templates for transcription and replication by the RdRp in the cytoplasm of an infected host cell.

For many RNA viruses, assembly initiates with the binding of N protein or a core protein to a unique encapsidation signal within the viral genome. This interaction promotes the oligomerization of the nucleocapsid, interaction with other viral proteins, and the subsequent formation of the virus particle. Over a decade ago it was suggested that sequences or structures present in the 5' ends of the hantaviral genome could provide a point of nucleation for encapsidation by the N protein (24). Shortly thereafter, studies of the

Bunyawera virus (BUNV), a member of the *Bunyavirus* genus within the family *Bunyaviridae* suggested that the N protein encapsidates vRNA and cRNA (antigenomic RNA), but not mRNA or other nonviral RNAs (10). Using *in vitro* binding assays, we have shown that the HTNV N protein has a strong preference for the vRNA as compared to the viral cRNA, mRNAs and nonspecific RNAs (29). A similar result was reported for the BUNV N protein (21). Both HTNV and BUNV N proteins have similar levels of affinity for their S-segment vRNA substrates with dissociation constants of 53 nM and 80 nM, respectively (29). The *cis*-acting elements that show the highest binding affinity to HTNV and BUNV N proteins map to the 5'-end of the vRNAs (21, 30). Furthermore, we have identified a RNA binding domain (RBD) within the HTNV N protein that maps to a central, conserved region (amino acids 175-217) (35). In addition to interactions with viral RNA, the N protein has been shown to form trimers in vitro (13, 14) and in virions (1). It has been proposed the N protein trimers may assemble onto the viral RNA and this is followed by protein-protein interactions that promote the encapsidation of the entire vRNA or cRNA (13). Recent experiments also suggest the involvement of a host cellular protein in assembly, the small ubiquitin-like modifier-1 (SUMO-1) conjugating enzymes 9 (Ubc9) (15, 17). Additional macromolecular interactions are likely to be required for the assembly of the three nucleocapsids into the virion and the budding of this complex into the Golgi (28).

Our previous studies have outlined the location of an RNA binding domain (35) and *cis*-acting element involved in HTNV nucleocapsid encapsidation and/or assembly (30). To further define the boundaries and essential amino acids involved in the interaction of the HTNV N protein with RNA, single alanine substitutions were made within amino acids 175-

217 of the full-length protein and in peptide mimics of the RBD. In addition, several different chemicals were used to modify specific functional groups of amino acids in the full-length HTNV N protein to identify amino acids that contribute to RNA binding. The results delineate the presence of a major RNA binding domain with additional interactions being mediated by lysine residues scattered in the terminal half of the N protein.

MATERIALS AND METHODS

Site-directed mutagenesis. The Quick-Change Site Directed Mutagenesis kit (Stratagene, La Jolla, CA), was used to introduce single and double ala substitutions into the full length HTNV N protein, which is cloned into the pET23b bacterial expression vector (29). Single mutations were directly introduced into this plasmid with two synthetic oligonucleotide primers (Table 1). Each primer was designed complementary to the opposite strand of the vector. An extension reaction was made from each primer with *Pfu* polymerase to incorporate the desired changes. Wild type template was removed with *Dpn*1 and the reaction was transformed into XL-1 Blue competent cells. The nucleotide sequences of all clones were confirmed with the LiCOR 4200 IR^2 Automated Sequencer (Lincoln, NE). Each clone was constructed with a hexahistidine tag at the C-terminus, which has been previously shown to not affect the RNA binding ability of the protein (29).

Protein purification. HTNV N protein and N protein mutants were purified as described previously (11, 35). Briefly, the wild type HTNV N protein and mutants were transformed into BL21(DE3) cells. Cells were grown overnight in 100 ml LB medium containing 200 μg/ml ampicillin. After 12-14 h, the cells were centrifuged for 5 min at 5,000 rpm in a Sorval GS3 rotor. The pellet was resuspended in 40 ml of LB medium containing 200 μg/ml ampicillin, and inoculated into 1 L of LB medium containing 200 μg/ml ampicillin. Cultures were grown for 2 h at 30° C before IPTG (isopropylthiogalactoside) was added to a final concentration of 1.2 mM to induce protein expression. After an induction time of 2-4 h, the cells were harvested (centrifugation for 5 min at 5,000 rpm in a Sorval GS3

rotor) and resuspended in 40 ml of Extraction Buffer pH 8.0 (EB) (0.1 M sodium phosphate, 0.5 M sodium chloride, 0.1% TWEEN 20, 6 M guanidine-HCl). The resulting suspension was sonicated twice on ice using a Branson Sonifier until the solution was clear and not viscous. The extraction was continued for 2 h with shaking at room temperature. Soluble and insoluble material was separated by centrifugation at 30,000 g for 30 min. The supernatant was applied to a 5 ml column containing 2 ml of nickel-nitriloacetate agarose resin (Qiagen, Valencia, CA), which had been preequilibrated with 10 column volumes of EB pH 8.0. The column was washed with 10 column volumes of Buffer A (0.1 M sodium phosphate, 0.5 M sodium chloride, 10% glycerol, 8 M urea) pH 8.0, pH 6.3, pH 5.9. The N protein was eluted from the column using 10 column volumes of Buffer A pH 4.5. Three 1 ml fractions of each mutant protein, and five 1 ml fractions of the wild type protein were collected. Twenty microliters of each fraction were analyzed for protein content on a 12% SDS PAGE. The fractions containing the desired protein were dialyzed over five days into Buffer C (0.02 M HEPES pH 7.5, 0.5 M sodium chloride, 10% glycerol) with the urea concentration decreasing daily by half.

Oligoribonucleotides. Oligoribonucleotide vRNA 1-39 (5′-UAgUAgUAUgCUCCCUAAAAAgACAAUCAAggAgCAAUC-3′) was synthesized on a 1 μM scale and HPLC purified by Integrated DNA Technologies, Inc (Coralville, IA). The synthetic RNA was labeled at the 5' terminus with [γ-^{32}P] ATP and T4 polynucleotide kinase (New England Biolabs, Boston, MA), and purified on Quick Spin™ Columns (Roche, Indianapolis, IN).

ABSTRACT

The nucleocapsid (N) protein of hantavirus encapsidates viral genomic and antigenomic RNAs. Previously, deletion mapping identified a central, conserved region (amino acids 175- 217) within the Hantaan virus (HTNV) N protein that interacts with a high affinity with these viral RNAs. To further define the boundaries of the RNA binding domain (RBD), several peptides were synthesized and examined for their ability to bind full-length S-segment vRNA. Peptide 195-217 retained ninety-four percent of the vRNA as compared to HTNV N protein, while peptides 175-186 and 205-217 bound only one percent of the vRNA. To further explore which residues were essential for binding vRNA, we made a comprehensive mutational analysis of the amino acids in the RBD. Single and double ala substitutions were constructed for 18 amino acids from amino acids 175-217 in the full-length N protein. In addition, ala substitutions were made for the three arg residues in peptide 185-217. Analysis of protein-RNA interactions by electrophoretic mobility shift assays implicated E192, Y206 and S217 to be important for binding. Chemical modification experiments showed that lys, but not arg or cys residues, contribute to RNA binding. Overall, these data implicate lys residues dispersed from amino acids 175-429 of the protein and three amino acids located in the RBD as essential for RNA binding.

In vitro transcription of viral and non-viral RNA. Radiolabeled HTNV S-segment RNA transcripts was produced from the linear HTNV S/ pGEM1 using a MaxiScript Sp6 RNA Transcription kit protocol (Ambion, Austin, TX) as described previously (29). pGEM7Zf was used for *in vitro* transcription of a nonviral 67-nucleotide RNA used as a control RNA. The RNeasy Kit (Qiagen) was used to purify transcripts.

Filter binding assay. Binding reactions were done as previously described (29, 30). Briefly, purified HTNV N proteins or N deletions were serially diluted (22.7, 7.57, 2.52, 0.84 and 0.28 μM) in a final volume of 20 μl binding buffer (40 mM HEPES, pH 7.4, 100 mM NaCl, 5% Glycerol). Samples were incubated at 37 C for 5 min, 1 ng of [α-32P] UTP-labeled RNA was added to each reaction, and the reactions incubated for an additional 10 min at 37 C. Signals were quantitated with a Storm Molecular Dynamics PhosphorImager and analyzed by ImageQuaNT™ version 4.2 software.

Gel electrophoresis mobility shift assay (GEMSA). One ng of ^{32}P-radiolabeled vRNA S-segment, prepared as described above, was incubated with 56 μM of purified N protein in binding buffer [40 mM HEPES pH 7.4, 100 mM NaCl, 2.5 mM $MgCl_2$, 5% glycerol] in a final reaction volume of 20 μl, and incubated at 37 C for 10 min. Fifty units of RNase T1 (Ambion, Austin, TX) were added and the reactions incubated for an additional 10 min at 37 C. One μg of heparin was then added, and reactions were incubated for an additional 15 min at 37 C. One μl of sample buffer (30 % glycerol and 0.2 % bromophenol blue) was added to each reaction, the reactions were loaded onto 6% acrylamide gel,

separated by electrophoresis in 0.5X TBE buffer at 200V (constant voltage) for 3.0 hr, and visualized by autoradiography.

Chemical modification of HTNV N protein. To neutralize positive charges on lysine residues or block lysine residues, sulfosuccinimidyl acetate (NHS) (Pierce, Rockford, IL) was used. Chemical modification of arg side chains was accomplished with 4-hydroxyphenylglyoxal (HPG) (Pierce). N-succinimidyl iodoacetate (SIA) (Pierce) was used as a cross-linking reagent to modify amine and sulfhydryl groups in close proximity. Lastly, N-ethylmaleimide (NEM) (Pierce) in combination with DTT was used to modify cys residues as described previously (12). To modify the appropriate residues, the chemical was added to a final concentration of 10 mM to aliquots of full-length HTNV N protein and incubated on ice for 30 min. To modify cys residues with NEM, aliquots of HTNV N protein were treated with 10 mM NEM for 30 min on ice. After the 30 min incubation, DTT was added to a final concentration of 50 mM. Likewise, NEM was added to final concentration of 10 mM to HTNV N protein treated with 50 mM DTT. Preincubated HTNV N proteins (56 μM) were then added to reaction mixtures as described above and the complexes were analyzed by GEMSA.

Sequence retrieval and hydrophobic cluster analysis (HCA). Sequences for this study were obtained using BLASTP (3) against SWISSPROT and Trembl data bases (4). N sequence accession numbers are as follow: P17881 Seoul virus (strain SR-11) (Sapporo rat virus), Q8UY38 Hantavirus Tchoupitoulas, Q9DJK9 Dobrava virus, Q910R1 Amur virus and Hantaan virus, Q9WMG3 Topografov hantavirus, O36307 Andes virus, Q8QRN0

Lechiguanas virus, Q82761 Isla Vista virus, Q82953 Khabarovsk virus, Q8QRN1 Bermejo virus, Q9WJ32 Puumala virus, Q98590 Prospect Hill virus (PHV), Q8QRM9 Oran virus, Q8QRM7 Pergamino virus, O41357 Rio Mamore hantavirus, O12369 Laguna Negra virus, Q83889 New York hantavirus, Q98620 Convict Creek 107 virus (CC107V). Hydrophobic cluster analysis was carried out with the program DRAWHCA (5).

Plotting mean net charge against mean hydrophobicity to assess whether a protein is intrinsically disordered. The Mean Net Charge (R) of a protein is determined 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 (http://www.expasy.ch/tools). The mean hydrophobicity (H) is the sum of normalized hydrophobicities of individual residues divided by the total number of amino acid residues minus 4 residues (to take into account fringe effects in the calculation of hydrophobicity). Individual hydrophobicities were determined using the Protscale program at the EXPASY server (http://www.expasy.ch/tools), 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 4 to yield (H).

For a given protein, R is then plotted against H. The charge/hydrophobicity diagram is divided into 2 regions by a line, which corresponds to the equation $H = (R+1.151)/2.785$. In the left part of the diagram (where $H < (R+1.151)/2.785$), a protein is predicted as disordered, whereas it is predicted as ordered in the right part. The net charge-hydrophobicity method is only applicable to a protein or protein region provided it cannot be subdivided into shorter,

structurally independent modules, otherwise it might give conflicting results. It was only validated for regions > 50 aa (33). An estimation of the error rate of the method can be drawn from (32).

PONDR prediction of unstructured regions. Sequences of proteins were submitted to the PONDR server (http://www.pondr.com/) using the default integrated predictor VL-XT (16, 25). The threshold for reliable predictions of disorder, or of order, is set to 40 residues (>99.6%). Access to PONDR was provided by Molecular Kinetics (IUETC, 351 west 10th Street, Suite 318, Indianapolis, IN 46202; E-mail: main@molecularkinetics.com) under license from the WSU Research Foundation. PONDR is copyright ©1999 by the WSU Research Foundation, all rights reserved.

Amino acid composition analysis. The average sequence composition of globular proteins was taken from the "globular 3D" dataset at http://www.disorder.chem.wsu.edu. If the average composition of an amino acid X in globular proteins is CG_X, and CP_X is the composition in X of a protein P, deviation from the composition in X of globular proteins was defined for P as $(CP_X - CG_X)/ CG_X$.

Multiple sequence alignment and secondary structure predictions. The sequences of the hantavirus nucleoproteins were aligned using CLUSTAL W (31) and manually refined with SEAVIEW (8). The alignment was drawn using ESPrit 2.0 (9). The secondary structure predictions were performed with PSI-PRED (18) and the Predict Protein Server (26). The results presented are consensus of both methods.

RESULTS

Reconstitution of the RNA binding activity of the Hantaan virus N protein in a peptide. In our previous work, HTNV N protein deletions were prepared in the N-terminal, C-terminal or both regions of the N protein (35) and used to map a minimal RBD between amino acid residues 175 to 217. To further confirm the function of this region in RNA binding, we designed and purchased synthetic peptides of the RBD (Fig. 1 and Table 2). The largest peptide that Sigma-Genosys was able to synthesize was from amino acids 185-217 (Table 2). Filter binding analysis of peptide 185-217 with the 5'end vRNA (1-39) substrate showed it to have a K_d of ~70 nM (data not shown). This value is similar to the K_d published for the full-length HTNV N protein with S-segment viral RNA substrates (53 ± 8 nM) (29). Therefore, three arginine to alanine substitutions (R197A, R199A, R213A) were made in this peptide (Table 3). Arg residues within the RBD were targeted for mutational studies given their importance to a number of proteins in their RNA binding interactions.

The RNA binding ability of each peptide was measured by filter binding (Table 2). The percentage of full-length HTNV vRNA that was bound by each peptide is given as compared to the amount bound by the full length protein (Tables 2 and 3). Peptide 195-217 showed the greatest binding ability, with 94% of the RNA bound to the peptide relative to the N protein. Peptides 185-217 and 175-206 showed an approximate 2-fold reduction in binding, while approximately 1% of the vRNA was bound by peptides 205-217 and 175-186. These results suggested that amino acids 195-217 contain essential

amino acids for RNA binding (Table 2). While, peptides 185-217 (R197A) and 185-217 (R199A) showed a five-fold reduction in binding compared to wild-type HTNV N protein (Tables 2 and 3), they only showed a 2.5 fold reduction as compared to the 185-217 peptide. Hence, none of the arg to ala substitutions, individually, had a significant effect on RNA binding in the filter binding assays.

Chemical modification of HTNV N protein. To further examine the contribution of specific classes of amino acids in the full-length HTNV N protein as they relate to HTNV vRNA binding, we used chemical mutagenesis of the full length protein. RNA-binding activity of HTNV N protein after chemical modification of cys, lys, arg or a cross-linking agent were assayed by GEMSA. To investigate the effect of modification of cys residues on RNA binding, we treated the HTNV N protein with NEM. HTNV N protein contains five cysteine residues, one of which is in the RBD (C203). The other four are located at aa positions 244, 293, 315 and 319. We noted no inhibition of RNA binding by the HTNV N protein when treated with NEM (Fig. 2, lanes 3 and 4). The levels of RNA-N protein complex were similar with NEM-treated and untreated HTNV N protein.

Eleven percent of the aa residues in the full-length HTNV N protein are either lys or arg. Within the HTNV N protein RBD (175-217), 6 of the 43 aa (14%) are lys or arg residues. When we treated HTNV N protein with NHS, which neutralizes the positive charge on lys residues, there was complete inhibition of RNA binding (Fig. 2, lanes 5 and 6). However, HPG treatment of HTNV N protein, which reacts with the guanidyl group of arg, had no effect on RNA binding (Fig. 2, lanes 7 and 8). Similarly, chemical modification with

succinyimidyl iodoacetate (Fig. 2, lanes 9 and 10), which cross-links proteins via an amine to thio-linkage showed no reduction in band shift intensity.

To further map the region where critical lys residues reside, five previously prepared truncated constructs that removed conserved regions in the N-terminal, C-terminal or both regions of the N protein (35) were subjected to chemical modification by NHS and examined by GEMSA analysis (Fig. 3). NΔ209 is devoid of the first 209 N-terminal amino acids, while CΔ255 is devoid from the last 255 C-terminal amino acids. NP175-217, NP175-270 and NP175-300 had both N-terminal and C-terminal deletions. In the presence of the chemical modifier, NHS, there was complete inhibition of RNA binding of HTNV N protein, NΔ209 (Fig. 3, lane 10), NP175-217 (lane 11), NP175-270 (lane 12), and NP175-300 (lane 13) as compared to non-treated proteins (Fig. 3, lanes 3-6). However, there was only a 6.5 fold reduction in band shift intensity with CΔ255 (Fig. 3, lane 14 as compared to lane 7) suggesting that lys residues located between 175-429 contribute to RNA binding.

What are the critical amino acids in the Hantaan virus N protein that are used in its interaction with vRNA and cRNA? To map the critical amino acids within the RBD, we created single amino acid substitutions in the full-length N protein from aa 175-217 (Table 1). We targeted amino acids that are known to interact with nucleic acids, such as lys, arg, thr, ser, asp, and glu. The 31 amino acid region encompassing residues 175-217 does not have any trp or phe residues, but does have several tyr, which can interact through the aromatic ring via stacking interactions. Hence these were also targeted for mutational analysis. Mutant

N proteins were purified and examined for their ability to bind viral and nonviral RNA as compared to *wt* N protein.

There was a reduction in binding of the single amino acid substitutions E192A, Y206A and S217A as compared to *wt* N protein (Fig 4). The percentage of HTNV vRNA bound to mutant N proteins E192A and S217 was 16%, a 6.2-fold decrease as compared to *wt* N protein (Fig. 4, lanes 10 and 20, Table 4). The mutant Y206A showed a 4.5-fold reduction in signal intensity (Fig. 4, lane 16 and Table 4). The remaining fifteen mutant N proteins had approximately the same signal intensities as compared to *wt* N protein (Fig. 4, and Table 4). Three amino acids, a glutamic acid (E192), a tyrosine (Y206) and a serine (S217) in the RBD were identified as being essential for RNA recognition.

While single mutations in E192, Y206 and S217 reduced RNA binding, none of the mutants were devoid of binding activity. Therefore, two double mutants were produced and examined by GEMSA (Fig. 5). One double mutant protein was made with Y178A / Y206A. Mutant protein Y178A maintained 100 % signal intensity as compared to *wt* HTNV N protein. However, the double mutant showed only 5% of the RNA bound as compared to 22% for mutant Y206A protein alone (Fig. 5, compare lanes 4 and 5 and Table 4). Strikingly, the Tyr to Ala changes at positions 206 and 217 showed almost complete inhibition with only 2 % of the signal present as measured by phosphorimaging (Fig. 5, lane 7 and Table 4). To examine specificity, purified Rous Sarcoma Virus (RSV) nucleocapsid protein (NC) was included in the GEMSA to determine its ability to bind to HTNV vRNA. There was no gel shift with the RSV NC protein (Fig. 5, lane 9). In the same vein, the HTNV N protein did not

bind to the packaging signal of the RSV RNA (data not shown). Finally, GEMSA analysis of single ala substitutions did not produce a band/gel shift when tested in the presence of a nonviral RNA substrate (data not shown). In summary, this supports a specific interaction between the RBD of HNTV N and its vRNA.

Bioinformatic analysis of the hantavirus N protein. Intrinsically disordered regions have been shown to be involved in protein-RNA and protein-protein interactions (6) (34). To determine whether this was a characteristic within the hantavirus N protein, we carried out a bioinformatics analysis of the protein. We used BLASTP to retrieve 19 hantavirus N sequences homologous to that of HTNV N, with an overall identity of 48% and an overall similarity of 78%. Examination of the HTNV N sequence using HCA points out the presence of three globular domains (aa 95-141), (aa 156-279) and (aa 295-429) separated by two linkers (aa 142-155 and aa 280-294) and by a large disordered N-terminal domain (aa 1-94). Globular regions are characterized by a typical thick distribution of hydrophobic clusters contrasting with disordered regions that contain scattered hydrophobic amino acids. The region encompassing aa 23-47 is enriched in Lys and Arg residues (R22, K25, R26, K30, K44, K41, R42) (Fig. 6) and is depleted in hydrophobic clusters (Fig. 7). Nevertheless, two small stretched hydrophobic clusters, encompassing aa 1-20 and 49-69, are found, thus suggesting that these regions may have a potentiality to fold. HCA carried out on all the hantavirus N sequences shown in Fig. 7, reveals that they possess the same overall modular organization as HTVN, with a large disordered N-terminal domain exhibiting a clear folding potential. Analysis of the first 100 residues of HTNV and Andes Virus (ANDV) N by the method of the mean hydrophobicity/mean net charge ratio shows that this domain is predicted

to be natively unfolded. PONDR analysis on HTNV, ANDV and Bermejo virus N shows a consensus of disorder for the first 52 amino acids (Fig. 7). Noticeably, in the case of HTNVN, the prediction of disorder further extends to the region encompassing residues 1-94. Analysis of the deviation in amino acid composition of the regions encompassing amino acids 1-50 and 1-100 of HTNV and ANDV N, indicates that they are both depleted in order promoting residues and enriched in Q which is considered as a disorder promoting residue (Fig. 6 and data not shown). Although PONDR, the method of the mean hydrophobicity/mean net charge ratio and the analysis of the aa composition converge all to show that the region spanning aa 1-94 of HTVN is intrinsically disordered, secondary structure predictions indicate the presence of two long α-helices in the first 70 amino acids (Fig. 7). The occurrence within the N-terminal domain of predicted α-helices, together with the presence of two small hydrophobic clusters (aa 1-20 and 49-69) suggests that the region encompassing residues 1-70 may have a propensity to undergo induced folding in the presence of a partner/ligand. The relative enrichment of this region in Lys residues is reminiscent of RNA binding regions of RNA dependent RNA polymerases. In conclusion, the sequence properties of the region spanning residues 1-70 of HTVN converge to suggest that it is an intrinsically disordered region that may undergo induced folding upon binding to a partner. The presence of numerous lys residues suggests that the ligand may be RNA.

The 175-220 region. The region 175-220 does not have the sequence features typical of intrinsically disordered regions. Although no secondary structure element is consistently predicted in this region (Fig. 7), PSI-PRED and Predict Protein both predict an α -helix encompassing residues 200-220 of HTVN (even if with a reliability lower than 50%). This prediction is in agreement with the HCA plot of HTVN, which shows a potential α-helix in

this region (data not shown). Therefore, the 175-220 region of HTVN might be an unstructured region possibly involved in induced folding upon binding to RNA.

DISCUSSION

The hantavirus N protein is instrumental in encapsidation of the viral RNAs for formation of the nucleocapsid and assembly of the virion. In addition it has been suggested to play a role in the replication and transcription activities of the virus. Each of these activities requires interaction with viral RNA. Previously, we mapped an RNA binding domain within the HTNV N protein to amino acids 175-217 (35). This region was mapped by its preferred interaction with full-length S-segment vRNA and oligoribonucleotides of the vRNA sequence. Herein, we refined the mapping of the RBD and identified the region spanning amino acids 194-204 as the minimal region for binding. GEMSA analysis showed that chemical modification of Lys residues within 175-429 completely abolished RNA-binding activity (Fig. 2). In contrast, chemical modification of Arg residues did not affect RNA binding activity (Fig. 2). In support of these findings, two Ala substitutions at Lys197 and Lys199 in the RBD reduced RNA-binding by five-fold as compared to *wt* HTNV N protein (Table 4), while Arg substitutions reduced binding in peptides only 2 fold (Table 3). Three additional amino acids were mapped within the context of the full length N protein, namely Glu192, Tyr206 and Ser217 (Figure 4). The double mutant protein, Y206A / S217A, showed almost complete inhibition of RNA binding (Figure 5 and Table 4). In conclusion, our work has demonstrated that specific amino acids located in the RBD (175-217) contributed to RNA binding.

Studies with other RNA viruses have shown there are specific RNA-binding

domains in their respective N proteins. In the mouse hepatitis virus nucleocapsid protein, a 55 amino acid region from amino acids 177 to 231 bound vRNA with a dissociation constant of 32 nM (19). In Sindbis virus, a region in the nucleocapsid from residues 97 to 106 dictates specific encapsidation (22). In contrast, Elton et al., determined that multiple regions of the influenza virus nucleocapsid protein were essential for RNA binding (7). In their study, the authors show by fluorescence spectroscopy that five tryptophan residues, one phenylalanine and two arginines, distributed throughout the protein, are critical for high-affinity RNA binding (7). Comparatively, our data implicates three residues, a glutamic acid, a tyrosine and a serine in the RBD that may be essential for the specificity of the RNA-N protein interaction. Further, chemical modification of lysines, arginines and cysteines showed that lysine residues, scattered throughout the protein, additionally contribute to the binding of the viral RNA; particularly in the region spanning 175-429. Further, lysine residues within the 1-70 region, predicted to be disordered and to have a propensity for induced folding, may also play a crucial role in RNA binding. Further studies are required to define the possible functional interactions of this region, which could be involved in subsequent assembly steps such as the multimerization of N and interactions of the viral genomic ends that can form panhandles.

In addition to the role the RBD plays in encapsidation of the RNA, studies aimed at determining if the RBD motif plays a role in N protein multimerization and subsequent assembly processes have been inconclusive. For example, in a study by Alfadhli et al., the C-terminal region of the SNV and PHV N proteins were required for N protein- N

protein interaction by a yeast two-hybrid assay (2). Furthermore, in a recent study by Yoshimatsu et al., HTNV, SEOV, and DOBV N protein multimerization was observed in a competitive enzyme-linked immunosorbent assay (ELISA) even when 49 amino acids were deleted in the N-terminal region (36). Additionally, using a yeast two-hybrid assay, the authors found that amino acids 100 to 125 and 404 to 429 were required for N protein multimerization. Collectively, these data indicate that at least two regions are required for multimerization of hantavirus N protein. However, neither study speculated on the role of RNA in the assembly process. Overall, these data suggest that there may be an interdependence between nucleocapsid multimerization and N protein-RNA binding.

Our previous studies show the HTNV N protein – vRNA complexes are stable over a wide range of ionic strength conditions suggesting that the N protein interaction with viral RNA relies on specific structural and/or sequence determinants in the 5' end of the genome (29, 30). One hypothesis that Elton et al. propose is that influenza RNA binding by N protein is mediated by a combination of electrostatic and planar interactions through multiple regions of the protein (7). In contrast, we have recently found evidence from thermodynamic analysis that the major contributions of the N protein –RNA interaction lay in van der Waals forces and hydrogen bonding (data not shown). It is intriguing to note, that similar to the case of the influenza virus(7), we found positively charged amino acids, specifically lysine residues, distributed from amino acids 175-429 of the N protein, were absolutely required for the N protein –RNA interaction. While mutation of a negatively charged, an aromatic and a hydroxyl containing side chain amino acid in the RBD led to an 80 % decrease in RNA binding affinity. At a minimum,

the evidence indicates that there are overlapping specificities between the lysine residues and the key amino acids in the RBD.

ACKNOWLEDGEMENTS

This work was supported by a grant to C.J. from the Department of Defense, DAMD17-00-1-0513 and by the National Institutes of Health RISE Grant GM61222-01. We thank Salvador Cisneros, Samantha Claw and Stephanie Chapman, interns in the NIH MBRS/RISE Program at New Mexico State University, for technical assistance. We thank Dr. Maxine Linial at Fred Hutchinson Cancer Center for Rous Sarcoma Virus nucleocapid protein.

FIGURE LEGENDS

FIG. 1. Clustal W sequence alignment of hantavirus N proteins containing the putative minimal RNA binding domain.

FIG. 2. GEMSA of the chemical modification of amino acid residues in HTNV N. Binding affinities were examined by GEMSA, in duplicate, with in vitro-transcribed full-length HTNV S-segment vRNA and the expressed wild type N protein (lane 2). Samples were treated with NEM (lanes 3 and 4), NHS (lanes 5 and 6), HPG (lanes 7 and 8), and SIA (lanes 9 and 10). A no protein control is presented in lane 1. A 56 μM concentration of HTNV N protein was treated with a 10 mM final concentration of each chemical modification reagent and incubated with radiolabeled vRNA S-segment. The protein-vRNA complexes were separated from free RNA by 6% non-denaturing PAGE as described in Materials and Methods.

FIG. 3. GEMSA of the chemical modification of wt and truncated forms of HTNV N. Binding affinities were examined by GEMSA with *in vitro*-transcribed full-length HTNV S-segment vRNA and the expressed wt N protein (lanes 2 and 9), NΔ209 (lanes 3 and 10), NP175-217 (lanes 4 and 11), NP175-270 (lanes 5 and 12), NP175-300 (lanes 6 and 13), and CΔ255 (lanes 7 and 14). A no protein control is presented in lanes 1 and 8. A 56 μM concentration of HTNV N protein and deletions (lanes 9-14) were treated with a 10 mM final concentration of NHS and incubated with radiolabeled vRNA S-segment. The reaction mixtures were loaded onto a 6% non-denaturing PAG and the protein-vRNA

complexes were separated from free RNA by gel electrophoresis as described in Materials and Methods.

FIG. 4. GEMSA of the *wt* HTNV N protein and single amino acid substitutions. Binding affinities were examined by GEMSA with *in vitro*-transcribed full-length HTNV S-segment vRNA and the expressed *wt* N protein (lane 2), K175A (lane 3), Y178A (lane 4), Q185A (lane 5), S186A (lane 6), S187A (lane 7), K189A (lane 8), E191A (lane 9), E192A (lane 10), T194A (lane 11), R197A (lane 12), Y198A (lane 13), R199A (lane 14), T200A (lane 15), Y206A (lane 16), Q209A (lane 17), R213A (lane 18), Q214A (lane 19), and S217A (lane 20). A no protein control is presented in lane 1. A 56 μM concentration of wild- type N protein and each single amino acid substitution was incubated with radiolabeled vRNA S-segment. The reaction mixtures were loaded onto a 6% non-denaturing PAG and the protein-vRNA complexes were separated from free RNA by gel electrophoresis as described in Materials and Methods.

FIG. 5. GEMSA of the *wt* HTNV N protein and single and double amino acid substitutions. Binding affinities were examined by GEMSA with *in vitro*-transcribed full-length HTNV S-segment vRNA and the expressed *wt* N protein (lane 2), Y178A (lane 3), Y206A (lane 4), Y178A / Y206A (lane 5), S217A (lane 6), Y206A / S217A (lane 7), E192A (lane 8), and RSV NC (lane 9). A no protein control is presented in lane 1. A 56 μM concentration of *wt* N protein and each single and double amino acid substitution was incubated with radiolabeled vRNA S-segment. The reaction mixtures

were loaded onto a 6% non-denaturing PAG and the protein-vRNA complexes were separated from free RNA by gel electrophoresis as described in Materials and Methods.

FIG. 6. Sequence composition. Deviation in aa composition of the 1-50 region of HTVN from the average values of globular proteins. Order-promoting, disorder-promoting and amino acids that are indifferently enriched or depleted in disordered regions of proteins are indicated.

FIG 7. Multiple sequence alignment. Multiple sequence alignment of hantavirus nucleoproteins. The consensus sequence (identity cutoff >70%) is shown under the multiple sequence alignment. Dots and residues shown in lowercase correspond to residues under and above the cutoff, respectively, while positions marked by # correspond to any of NDQE. Residues corresponding to an homology >70%, are boxed. For a given position only residues homologous to the consensus are in bold. The front numbers correspond to the amino acid position in sequence. Dots above the alignment indicate intervals of 10 residues. Secondary structure elements consistently predicted are shown above the alignment. The consensus of the PONDR prediction of disorder is also shown (aa 1-50). The hydrophobic cluster plot of the 1-100 region is shown above the multiple sequence alignment. Conventions are explicited in the caption. Globular regions are characterized by a thick distribution of hydrophobic clusters, while unstructured regions are poor or devoid of hydrophobic clusters. The putative RNA binding region and the regions potentially involved in induced folding are indicated.

REFERENCES

1. **Alfadhli, A., Z. Love, B. Arvidson, J. Seeds, J. Willey, and E. Barklis.** 2001. Hantavirus nucleocapsid protein oligomerization. J Virol **75:**2019-23.
2. **Alfadhli, A., E. Steel, L. Finlay, H. P. Bachinger, and E. Barklis.** 2002. Hantavirus nucleocapsid protein coiled-coil domains. J Biol Chem **277:**27103-8.
3. **Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman.** 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res **25:**3389-402.
4. **Bairoch, A., and R. Apweiler.** 2000. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res **28:**45-8.
5. **Callebaut, I., G. Labesse, P. Durand, A. Poupon, L. Canard, J. Chomilier, B. Henrissat, and J. P. Mornon.** 1997. Deciphering protein sequence information through hydrophobic cluster analysis (HCA): current status and perspectives. Cell Mol Life Sci **53:**621-45.
6. **Dunker, A., J. Lawson, C. Brown, R. Williams, P. Romero, J. Oh, C. Oldfield, A. Campen, R. CM, K. Hipps, J. Ausio, M. Nissen, R. Reeves, C. Kang, C. Kissinger, R. Bailey, M. Griswold, W. Chiu, E. Garner, and Z. Obradovic.** 2001. Intrinsically disordered protein. J Mol Graph Model. **19:**26-59.
7. **Elton, D., L. Medcalf, K. Bishop, D. Harrison, and P. Digard.** 1999. Identification of amino acid residues of influenza virus nucleoprotein essential for RNA binding. J Virol **73:**7357-67.

8. **Galtier, N., M. Gouy, and C. Gautier.** 1996. SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. Comput Appl Biosci **12:**543-8.

9. **Gouet, P., E. Courcelle, D. I. Stuart, and F. Metoz.** 1999. ESPript: analysis of multiple sequence alignments in PostScript. Bioinformatics **15:**305-8.

10. **Jin, H., and R. M. Elliott.** 1993. Characterization of Bunyamwera virus S RNA that is transcribed and replicated by the L protein expressed from recombinant vaccinia virus. J Virol **67:**1396-404.

11. **Jonsson, C. B., J. Gallegos, P. Ferro, W. Severson, X. Xu, C. S. Schmaljohn, and P. Fero.** 2001. Purification and characterization of the Sin Nombre virus nucleocapsid protein expressed in Escherichia coli. Protein Expr Purif **23:**134-41.

12. **Jonsson, C. B., and M. J. Roth.** 1993. Role of the His-Cys finger of Moloney murine leukemia virus integrase protein in integration and disintegration. J Virol **67:**5562-71.

13. **Kaukinen, P., V. Koistinen, O. Vapalahti, A. Vaheri, and A. Plyusnin.** 2001. Interaction between molecules of hantavirus nucleocapsid protein. J Gen Virol **82:**1845-53.

14. **Kaukinen, P., A. Vaheri, and A. Plyusnin.** 2003. Mapping of the regions involved in homotypic interactions of Tula hantavirus N protein. J Virol **77:**10910-6.

15. **Kaukinen, P., A. Vaheri, and A. Plyusnin.** 2003. Non-covalent interaction between nucleocapsid protein of Tula hantavirus and small ubiquitin-related modifier-1, SUMO-1. Virus Res **92:**37-45.

16. **Li, X., P. Romero, M. Rani, A. K. Dunker, and Z. Obradovic.** 1999. Predicting Protein Disorder for N-, C-, and Internal Regions. Genome Inform. Ser. Workshop Genome Inform. **10:**30-40.

17. **Maeda, A., B. H. Lee, K. Yoshimatsu, M. Saijo, I. Kurane, J. Arikawa, and S. Morikawa.** 2003. The intracellular association of the nucleocapsid protein (NP) of hantaan virus (HTNV) with small ubiquitin-like modifier-1 (SUMO-1) conjugating enzyme 9 (Ubc9). Virology **305:**288-97.

18. **McGuffin, L. J., K. Bryson, and D. T. Jones.** 2000. The PSIPRED protein structure prediction server. Bioinformatics **16:**404-5.

19. **Nelson, G. W., S. A. Stohlman, and S. M. Tahara.** 2000. High affinity interaction between nucleocapsid protein and leader/intergenic sequence of mouse hepatitis virus RNA. J Gen Virol **81:**181-8.

20. **Nichol, S. T., J. Arikawa, and Y. Kawaoka.** 2000. Emerging viral diseases. Proc Natl Acad Sci U S A **97:**12411-2.

21. **Osborne, J. C., and R. M. Elliott.** 2000. RNA binding properties of bunyamwera virus nucleocapsid protein and selective binding to an element in the 5' terminus of the negative-sense S segment. J Virol **74:**9946-52.

22. **Owen, K. E., and R. J. Kuhn.** 1996. Identification of a region in the Sindbis virus nucleocapsid protein that is involved in specificity of RNA encapsidation. J Virol **70:**2757-63.

23. **Peters, C. J., G. L. Simpson, and H. Levy.** 1999. Spectrum of hantavirus infection: hemorrhagic fever with renal syndrome and hantavirus pulmonary syndrome. Annu Rev Med **50:**531-45.

24. **Raju, R., and D. Kolakofsky.** 1989. The ends of La Crosse virus genome and antigenome RNAs within nucleocapsids are base paired. J Virol **63:**122-8.

25. **Romero, P., Z. Obradovic, X. Li, E. C. Garner, C. J. Brown, and A. K. Dunker.** 2001. Sequence complexity of disordered proteins. Proteins **42:**38-48.

26. **Rost, B.** 1996. PHD: predicting one-dimensional protein structure by profile-based neural networks. Methods Enzymol **266:**525-39.

27. **Schmaljohn, C., and B. Hjelle.** 1997. Hantaviruses: a global disease problem. Emerg Infect Dis **3:**95-104.

28. **Schmaljohn, C. S., and J. W. Hooper.** 2001. Bunyaviridae: The Viruses and Their Replication, p. 1581-1602. *In* D. M. Knipe and P. M. Howley (ed.), Fields Virology, 4 ed, vol. 2. Lippincott Williams and Williams, Philadelphia.

29. **Severson, W., L. Partin, C. S. Schmaljohn, and C. B. Jonsson.** 1999. Characterization of the Hantaan nucleocapsid protein-ribonucleic acid interaction. J Biol Chem **274:**33732-9.

30. **Severson, W. E., X. Xu, and C. B. Jonsson.** 2001. cis-Acting signals in encapsidation of Hantaan virus S-segment viral genomic RNA by its N protein. J Virol **75:**2646-52.

31. **Thompson, J. D., D. G. Higgins, and T. J. Gibson.** 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:**4673-80.

32. **Uversky, V. N.** 2002. What does it mean to be natively unfolded? Eur J Biochem **269:**2-12.

33. **Uversky, V. N., J. R. Gillespie, and A. L. Fink.** 2000. Why are "natively unfolded" proteins unstructured under physiologic conditions? Proteins **41:**415-427.

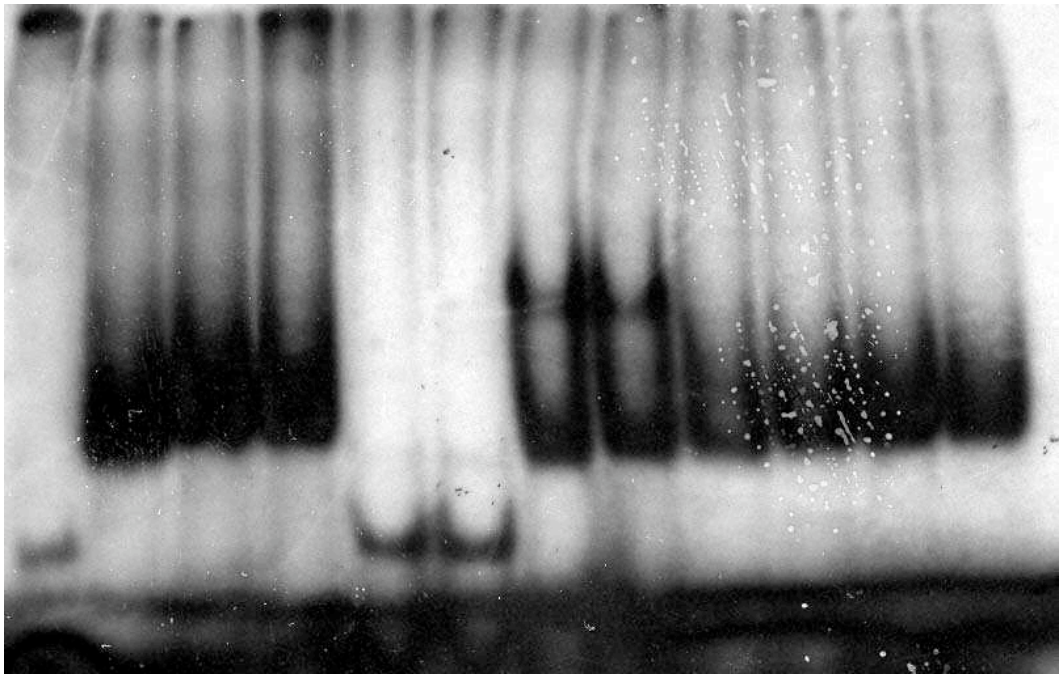
34. **Wright, P., and H. Dyson.** 1999. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. J Mol Biol. **293:**321-31.

35. **Xu, X., W. Severson, N. Villegas, C. S. Schmaljohn, and C. B. Jonsson.** 2002. The RNA binding domain of the hantaan virus N protein maps to a central, conserved region. J Virol **76:**3301-8.

36. **Yoshimatsu, K., B. H. Lee, K. Araki, M. Morimatsu, M. Ogino, H. Ebihara, and J. Arikawa.** 2003. The multimerization of hantavirus nucleocapsid protein depends on type-specific epitopes. J Virol **77:**943-52.

Fig. 1

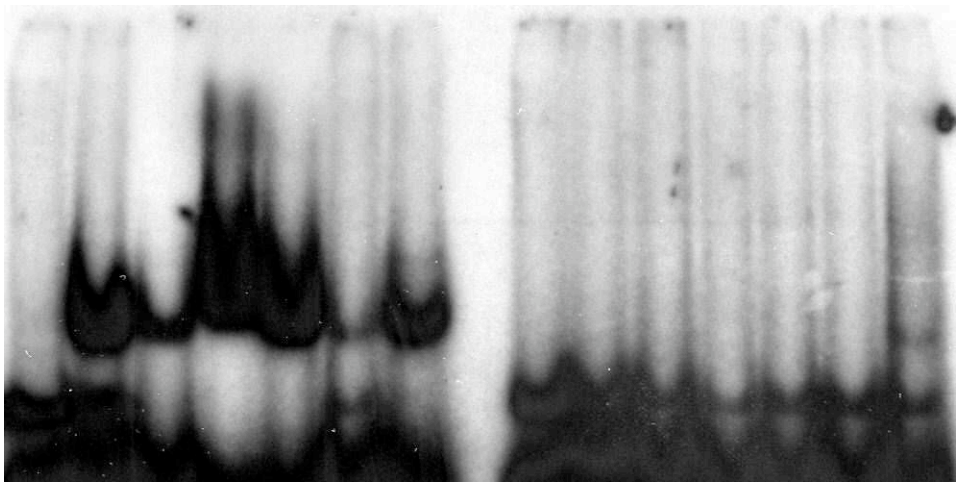
```
____________187        195        205      215

Topogravof TMKAEELT PGRFRTIVCG LFPAQIQARN IMS
Puumala    TMKAEELT PGRFRTIVCG LFPTQIQVRN IMS
Isla       TMKADELT PGRFRTIVCG LFPAQIMNRN IIS
Prospect   TMKAEELT PGRFRTIVCG LFPAQIMARN IIS
CC107      TMKADEIT PGRFRTIACG LFPAQVKARN IIS
Hantaan    SMKAEEIT PGRYRTAVCG LYPAQIKARQ MIS
IDENTITY   -MKA-E-T PGR-RT--CG L-P-Q---R- --S
SIMILARITY ******** ******__** ***_**__** ***
```



1 2 3 4 5 6 7 8 9 10 11 12

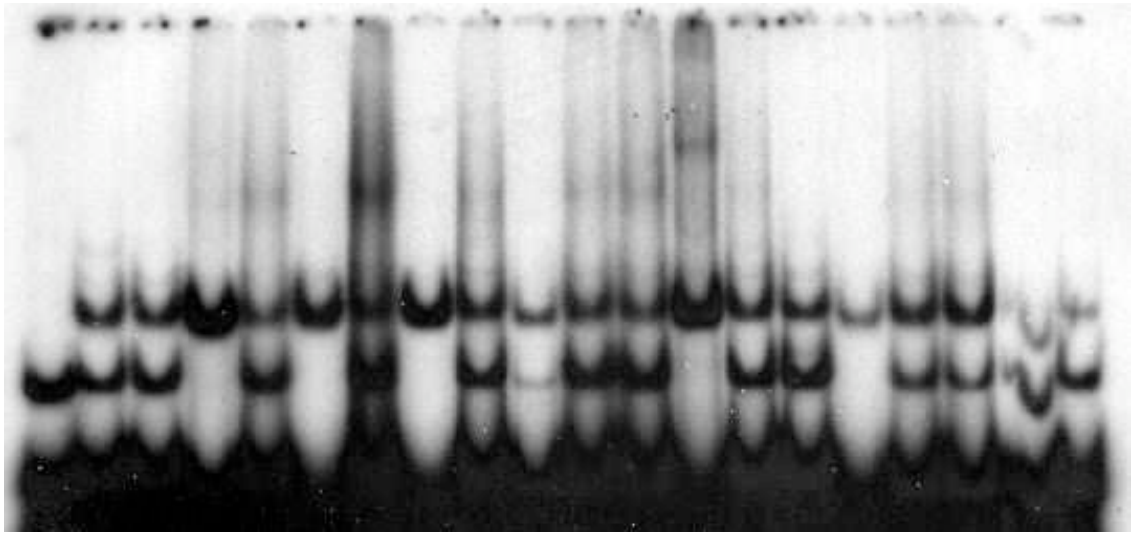
Fig. 2. GEMSA of the chemical modification of amino acid residues in HTNV N protein.

Fig. 3.



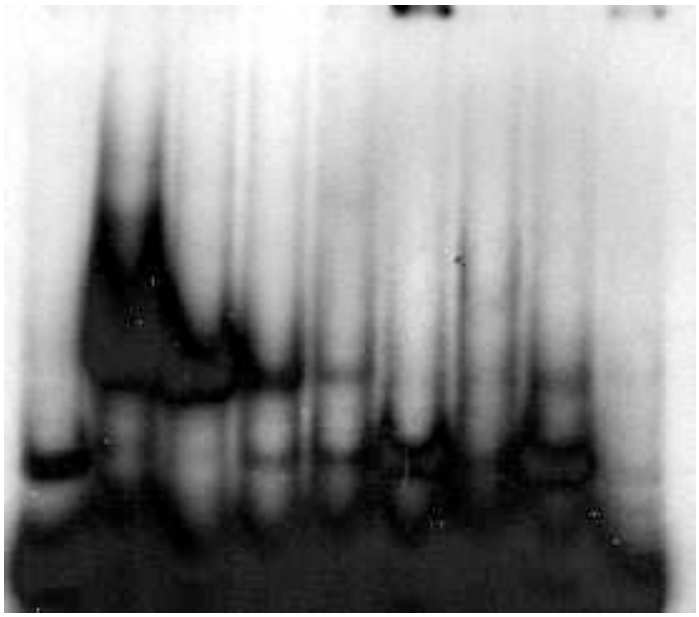
1 2 3 4 5 6 7 8 9 10 11 12 13 14

Fig. 4.



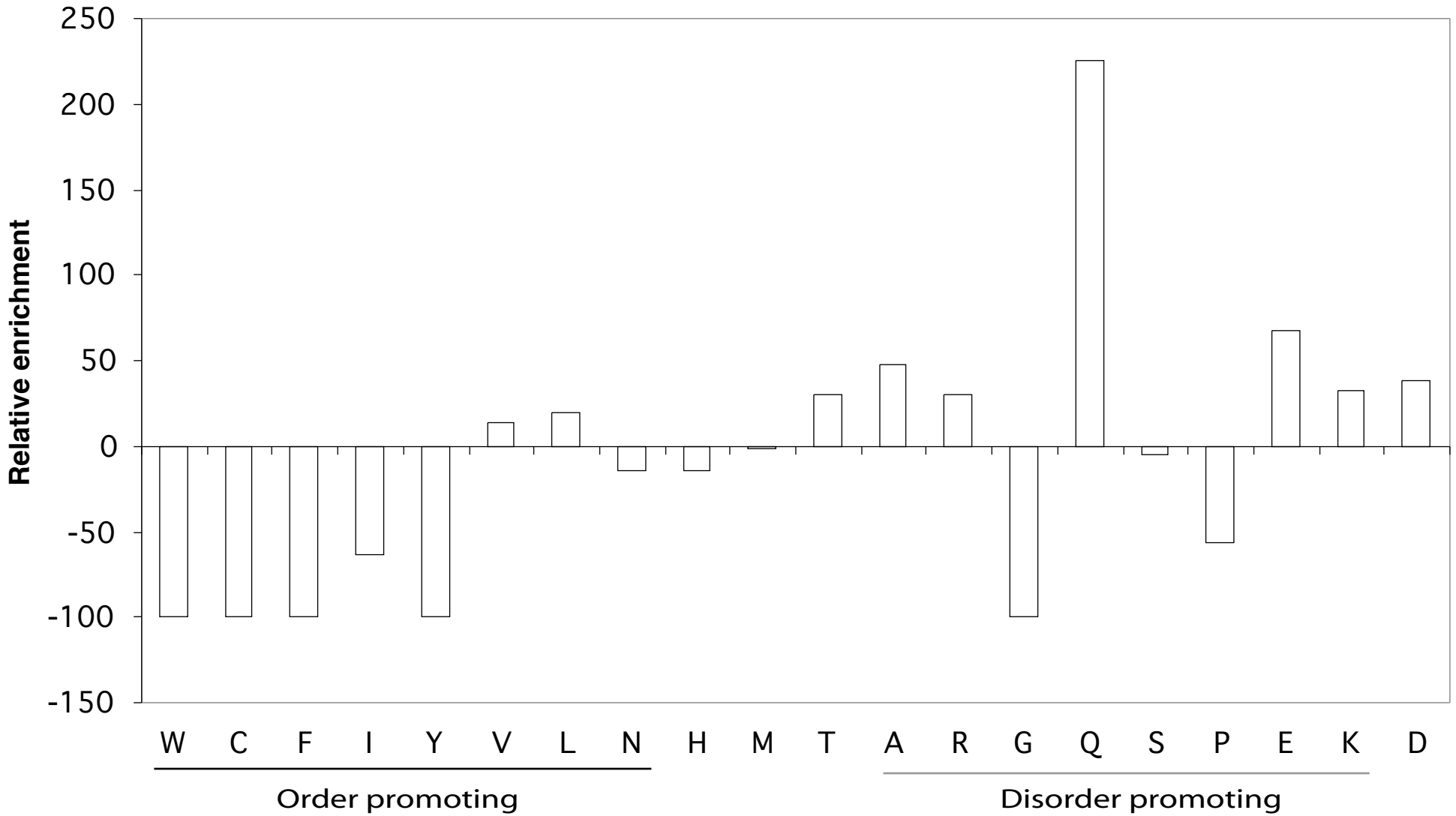
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20

Fig. 5.



1 2 3 4 5 6 7 8 9

Deviation in amino acid composition from the average values in the PDB of the 1-50 region of Hantaan Virus N (strain 76-118)



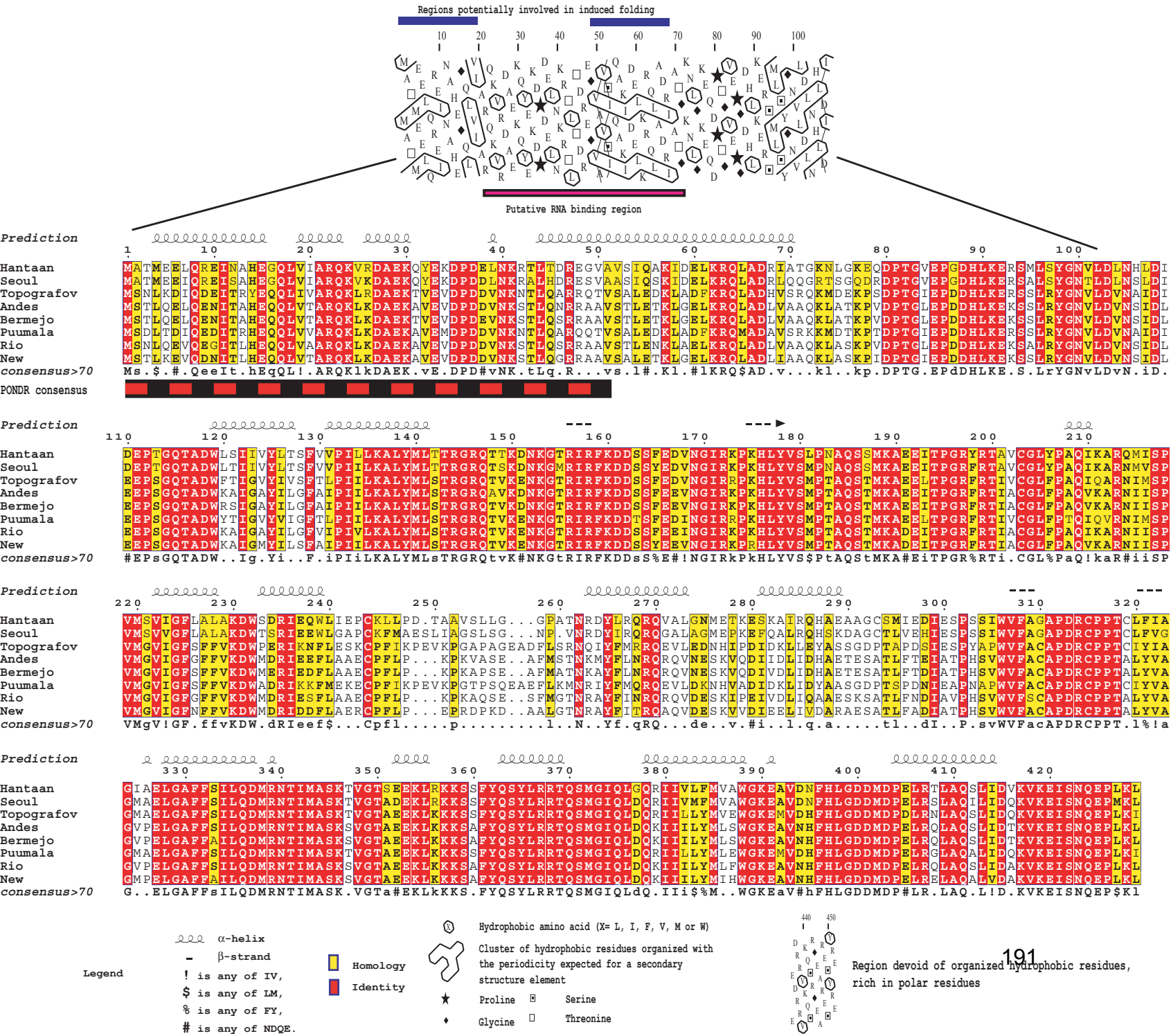

Regions potentially involved in induced folding
Putative RNA binding region
Prediction
Hantaan
Seoul
Topografov
Andes
Bermejo
Puumala
Rio
New
consensus>70
PONDR consensus
Legend
α-helix
β-strand
! is any of IV,
$ is any of LM,
% is any of FY,
# is any of NDQE.
Homology
Identity
Hydrophobic amino acid (X= L, I, F, V, M or W)
Cluster of hydrophobic residues organized with the periodicity expected for a secondary structure element
Proline
Glycine
Serine
Threonine
Region devoid of organized hydrophobic residues, rich in polar residues

TABLE 1. Primer design for site-directed mutagenesis

Mutant	*Primer 1*
K175A	c ggt atc cgg aaa cca gca cat ctt tac gtg tcc ttg
Y178A	cgg aaa cca aaa cat ctt GCc gtg tcc ttg cca aat gc
Q185A	gtg tcc ttg cca aat gca GCg tca agc atg aag gca g
S186A	c ttg cca aat gca cag Gca agc atg aag gca g
S187A	c ttg cca aat gca cag tca GCc atg aag gca gaa gag
K189A	gca cag tca agc atg GCg gca gaa gag att aca cc
E191A	g tca agc atg aag gca gCa gag att aca cct gg
E192A	gc atg aag gca gaa gCg att aca cct ggt aga tat ag
T194A	g aag gca gaa gag att Gca cct ggt aga tat aga ac
R197A	gaa gag att aca cct ggt GCa tat aga aca gca gtc tgt gg
Y198A	gag att aca cct ggt aga GCt aga aca gca gtc tgt ggg
R199A	g att aca cct ggt aga tat GCa aca gca gtc tgt ggg ctc
T200A	cct ggt aga tat aga Gca gca gtc tgt ggg c
Y206A	gca gtc tgt ggg ctc GCc cct gca cag att aag
Q209A	ggg ctc tac cct gca GCg att aag gca cgg cag
R213A	ct gca cag att aag gca GCg cag atg atc agt cca g
Q214A	gca cag att aag gca cgg GCg atg atc agt cca gtt atg
S217A	g gca cgg cag atg atc GCt cca gtt atg agt gta att g

TABLE 2. Percent of S-segment vRNA bound to HTNV N peptides as measured by filter binding

N Peptide					% vRNA bound
185-217		QSSMKAEEIT	PGRYRTAVCG	LYPAQIKARQ MIS	50 ± 5
195-217			PGRYRTAVCG	LYPAQIKARQ MIS	94 ± 3
205-217				LYPAQIKARQ MIS	1 ± 0.1
175-206	KHLYVSLPNA	QSSMKAEEIT	PGRYRTAVCG	LYP	44 ± 2
175-186	KHLYVSLPNA	QSS			0.9 ± 0

[a]56 μM of each peptide was used in the analysis. Bold amino acid indicates amino acid substitution.

[b]Percentage of vRNA bound to N peptide is from three separate experiments performed in duplicate (± SD).

TABLE 3. Percent of S-segment vRNA bound to HTNV N peptides as measured by filter binding

PEPTIDE[a]	SEQUENCE	% BINDING[b]
185-217	QSSMKAEEIT PGRYRTAVCG LYPAQIKARQ MIS	50 ± 5
P-R197A	QSSMKAEEIT PG**A**YRTAVCG LYPAQIKARQ MIS	20 ± 2
P-R199A	QSSMKAEEIT PGRY**A**TAVCG LYPAQIKARQ MIS	20 ± 1
P-R213A	QSSMKAEEIT PGRYRTAVCG LYPAQIKA**A**Q MIS	47 ± 5

[a]56 μM of each peptide was used in the analysis. Bold amino acid indicates amino acid substitution.

[b]Percentage of vRNA bound to N peptide is from three separate experiments performed in duplicate (± SD).

Table 4. GEMSA analysis of amino acid substitutions in the HTNV N protein RBD

Amino Acid(s)	*% vRNA bound*
Y178A	100
E192A	16
Y206A	22
Y178A/Y206A	5
S217A	16
Y206A/Y217A	2
RSV NC	0

Binding affinities of single and double amino acid mutant proteins were examined by GEMSA with in vitro transcribed full-length HTNV S-segment vRNA and compared to wild-type HTNV N protein.

2. Etudes des protéines du complexe réplicatif de virus à ARN positif

2.1. Les Coronavirus

2.1.1. Généralités

Les virus du genre Coronavirus appartiennent à la famille des *Coronaviridae*. Ce sont de gros virus enveloppés, dont le génome est constitué par de l'ARN simple brin, à polarité positive (Lai and Holmes 2001). Les Coronavirus infectent les oiseaux et beaucoup de mammifères, y compris les humains. Jusqu'à l'épidémie du syndrome respiratoire aigu sévère (SRAS) en 2003, les Coronavirus étaient considérés comme des virus ne présentant pas une forte menace pour l'homme (Holmes 2001). Les Coronavirus sont relativement restreints dans leur spectre d'hôte : ils infectent en effet seulement leur hôte naturel et des espèces animales relativement proche (Siddell 1995). Occasionnellement, les Coronavirus peuvent franchir la barrière d'espèce, comme illustré dans les cas de l'infection de dindon par le coronavirus bovin (BCoV) (Ismail et al. 2001), ou l'infection expérimentale de chien par le TGEV (« transmissible gastroenteritis virus ») (Larson et al. 1979). Malheureusement, les vecteurs biologiques restent inconnus.

2.1.2. Organisation de la particule virale et du génome

Les Coronavirus doivent leur nom à leur forme caractéristique en couronne ("corona" en latin). Ce sont des virus enveloppés ayant généralement une forme sphérique, avec un diamètre compris entre 100 et 120 nm. Leur enveloppe est constituée par une bicouche lipidique qui provient de la membrane plasmidique de la cellule infectée et dans laquelle sont insérées une lipoprotéine E (protéine d'enveloppe) et deux (ou trois) glycoprotéines virales S (communes à tous les Coronavirus), HE (propre aux Coronavirus bovins et murins) et M qui est intégralement membranaire. A l'intérieur du virus on trouve la nucléocapside associée à l'ARN génomique. La nucléocapside à une structure hélicoïdale (Figure 28).

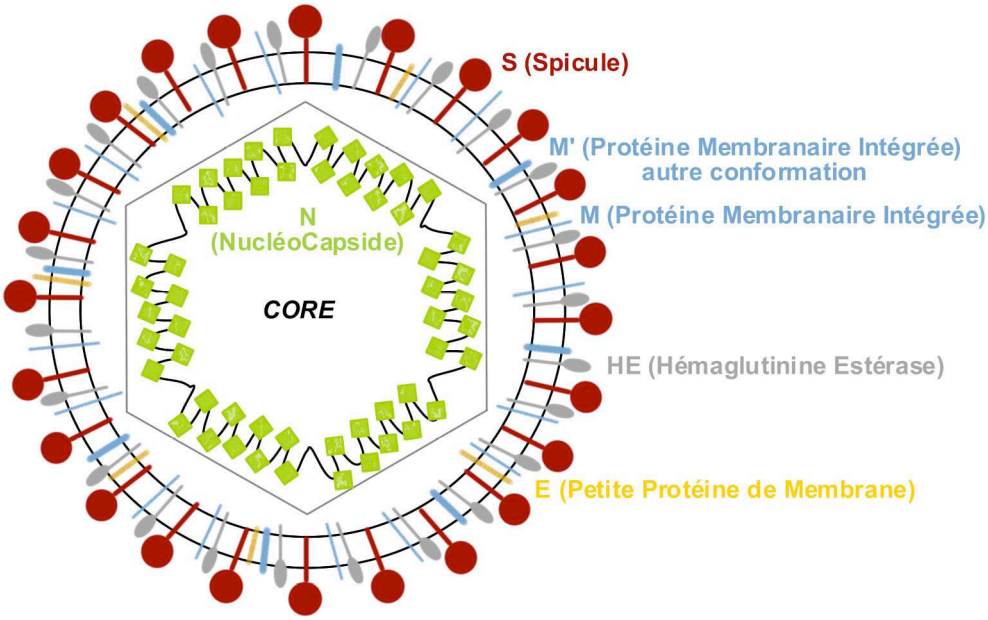


Figure 28 : Particule virale des Coronavirus.

Le génome des Coronavirus est non segmenté, constitué par de l' ARN simple brin de polarité positive, coiffé et polyadénylé. La taille du génome de ces virus est comprise entre 27 et 32 Kb. Le génome se compose d'un gène qui code pour une polyprotéine appelée « réplicase » et des quatre gènes codant pour des protéines structurales (S, E, M, et N) communes à tous les Coronavirus. Le génome des Coronavirus murins possèdent une protéine de surface additionnelle (HE), dont le gène se situe entre celui de la polyprotéine et celui de la protéine S. A cela s'ajoute une série de gènes non conservés (entre deux et quatre) au sein des Coronavirus et de fonction inconnue. Dans le génome, les gènes sont séparées par une région intergénique conservée qui permet la formation des ARN subgénomiques.

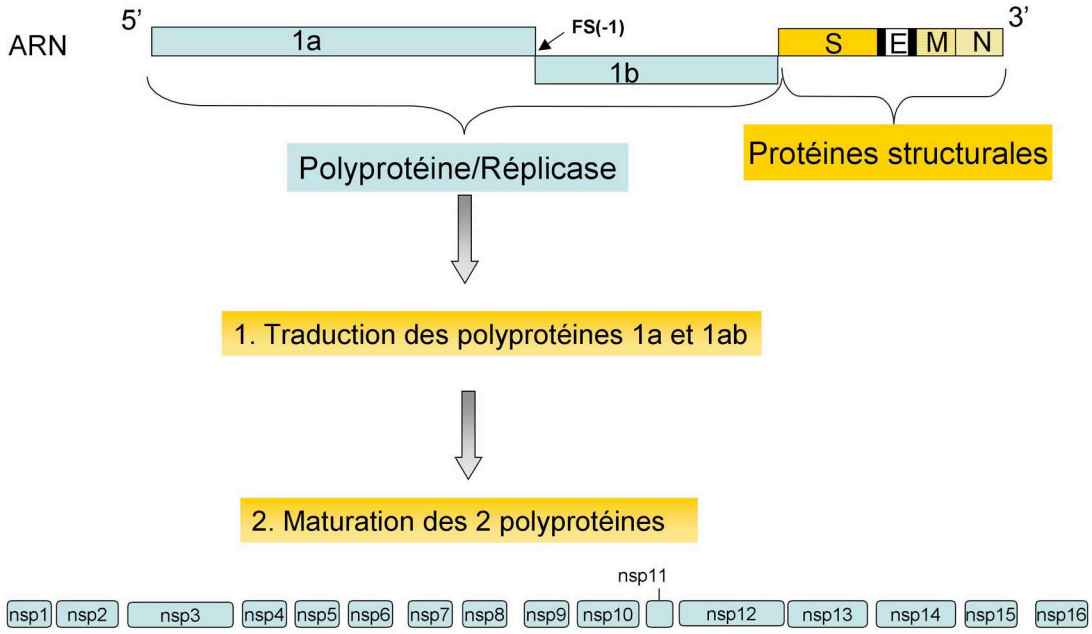


Figure 29 : Organisation du génome des Coronavirus

Au cours de ma thèse, j'ai réalisé l'annotation du génome du virus SRAS, et le découpage en modules des protéines de ce virus. Ce découpage fut la première étape qui a mené à la résolution de la structure de la protéine nsp9 issue de la polyprotéine. Ce travail est décrit dans les deux articles suivants.

ARTICLE 6

Structural genomics of the SARS coronavirus: cloning, expression, crystallization and preliminary crystallographic study of the Nsp9 protein

Campanacci V, Egloff MP, Longhi S, Ferron F, Rancurel C, Salomoni A, Durousseau C, Tocque F, Bremond N, Dobbe JC, Snijder EJ, Canard B, Cambillau C.

Acta Crystallogr D Biol Crystallogr. 2003 Sep;59(Pt 9):1628-31.

Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Structural genomics of the SARS coronavirus: cloning, expression, crystallization and preliminary crystallographic study of the Nsp9 protein

Valérie Campanacci,[a] Marie-Pierre Egloff,[a] Sonia Longhi,[a] François Ferron,[a] Corinne Rancurel,[a] Aurelia Salomoni,[a] Cécile Durousseau,[a] Fabienne Tocque,[a] Nicolas Brémond,[a] Jessika C. Dobbe,[b] Eric J. Snijder,[b] Bruno Canard[a]* and Christian Cambillau[a]*

[a]Architecture et Fonction des Macromolécules Biologiques, UMR 6098 CNRS and Universités Aix-Marseille I and II, 31 Chemin Joseph Aiguier, 13402 Marseille CEDEX 20, France, and [b]Molecular Virology Laboratory, Department of Medical Microbiology, Center of Infectious Diseases, Leiden University Medical Center, Leiden, The Netherlands

Correspondence e-mail: cambillau@afmb.cnrs-mrs.fr, canard@afmb.cnrs-mrs.fr

Received 12 July 2003
Accepted 30 July 2003

The aetiologic agent of the recent epidemics of Severe Acute Respiratory Syndrome (SARS) is a positive-stranded RNA virus (SARS-CoV) belonging to the *Coronaviridae* family and its genome differs substantially from those of other known coronaviruses. SARS-CoV is transmissible mainly by the respiratory route and to date there is no vaccine and no prophylactic or therapeutic treatments against this agent. A SARS-CoV whole-genome approach has been developed aimed at determining the crystal structure of all of its proteins or domains. These studies are expected to greatly facilitate drug design. The genomes of coronaviruses are between 27 and 31.5 kbp in length, the largest of the known RNA viruses, and encode 20–30 mature proteins. The functions of many of these polypeptides, including the Nsp9–Nsp10 replicase-cleavage products, are still unknown. Here, the cloning, *Escherichia coli* expression, purification and crystallization of the SARS-CoV Nsp9 protein, the first SARS-CoV protein to be crystallized, are reported. Nsp9 crystals diffract to 2.8 Å resolution and belong to space group $P6_{1/5}22$, with unit-cell parameters $a = b = 89.7$, $c = 136.7$ Å. With two molecules in the asymmetric unit, the solvent content is 60% ($V_M = 3.1$ Å^3 Da^{-1}).

1. Introduction

The recent epidemics of Severe Acute Respiratory Syndrome (SARS) represent a real paradigm for emerging viral pathogens, as well as an example of worldwide coordinated efforts to control a serious viral outbreak, a test of the reaction time of the scientific community. The first cases of Severe Acute Respiratory Syndrome originated from the Guangdong province in South East China. The number of cases reported and our current knowledge regarding this illness are still currently evolving, but a number of basic facts have been firmly established. The aetiologic agent of SARS is a positive-stranded RNA virus belonging to the *Coronaviridae* family and its genome differs substantially from those of previously identified coronaviruses, including two other human coronaviruses (Peiris *et al.*, 2003; Ksiazek *et al.*, 2003; Drosten *et al.*, 2003; Snijder *et al.*, 2003). The virus, whose name SARS-CoV is now currently accepted, is mainly transmitted by the respiratory route. However, evidence for a secondary faeco–oral route of transmission has also been presented. The viral strain probably primarily infected wild animals traded in Asian markets and crossed the species barrier to infect humans.

There is to date no vaccine and no prophylactic or therapeutic treatments against this agent. A prophylactic treatment would have been useful to combat the epidemics; the only effective measure available to prevent the spread of

© 2003 International Union of Crystallography
Printed in Denmark – all rights reserved

electronic reprint

the virus is to quarantine all persons that have been exposed to SARS-CoV. The number of antiviral molecules that can be used to treat patients infected by RNA viruses is incredibly low. Accordingly, it is important to search for efficient antiviral drugs for a large number of RNA viruses, while giving priority to viruses transmitted by the respiratory route because they have the highest potential for causing pandemic outbreaks.

The scientific community has reacted promptly and efficiently to identify and characterize this new infectious agent, as well as to develop methods for SARS-CoV detection and containment protocols. In the meantime, a wide effort is being made to design drugs active against SARS-CoV. Ribavirin has been used in the absence of other candidates, but its intrinsic efficiency against SARS-CoV appears to be low (Koren *et al.*, 2003).

To select drugs active against a viral pathogen, one usually relies on screening candidate drugs for their efficacy in virus-infected cell cultures and/or animal models. However, during the current research on drugs for treating hepatitis C virus (HCV) infections, a novel and promising approach has been introduced. The RNA-dependent RNA polymerase of HCV has been purified and crystallized and enzymatic tests have been used to find potent nucleoside and non-nucleoside inhibitors of the virus, the structure–activity relationships of which allow further testing and clinical developments (de Francesco *et al.*, 2003). This approach is gaining momentum owing to a concomitant increase in the power of new technologies and technological developments. Among those, genomics approaches are being conducted to solve the crystal structures of large sets of clinically relevant proteins, which will become the subjects of future structure–function relationship studies.

A crystal structure has not yet been determined for any of the 28 predicted mature SARS-CoV proteins. The crystal structure of the main (or 3CL) protease of transmissible gastroenteritis virus, a related coronavirus, has been determined and was used to construct a model of the SARS-CoV 3CL protease, facilitating future drug design against this important target (Anand *et al.*, 2003). The putative coronavirus RNA-dependent RNA polymerase has been purified, but is inactive *in vitro* (Grotzinger *et al.*, 1996).

In this context, we have developed a SARS-CoV whole-genome approach aimed at determining the crystal structure of all SARS-CoV proteins. We anticipate that this will greatly facilitate drug design as well as the study of many other aspects related to the biology of these complex viruses. Coronaviruses are enveloped viruses with a single-stranded RNA genome of positive polarity (Lai & Holmes, 2001). Their genome is between 27 and 31.5 kbp in length, the largest of the known RNA viruses. Like other coronaviruses, the SARS-CoV genome is known to encode two large replicase polyproteins (the ORF1a and ORF1ab proteins), which are processed into a set of mature non-structural proteins (Nsps) by internal viral proteases (Snijder *et al.*, 2003). The functions of many of these products, such as the Nsp9–Nsp10 polypeptides produced from the C-terminal domain of the ORF1a-encoded polyprotein, are still unknown. In the related mouse hepatitis virus, which is a group 2 coronavirus, the SARS-CoV Nsp9 corresponds to a 12 kDa cleavage product (P1a-12) that is found preferentially in the perinuclear region of infected cells, where it co-localizes with other components of the viral replication complex (Bost *et al.*, 2000). No clues to the function of the Nsp9 equivalent of any coronavirus have been obtained thus far. Here, we report the cloning, expression, purification and crystallization of the SARS-CoV Nsp9 protein, a 113-residue protein (Fig. 1), which is the first SARS-CoV protein to be crystallized.

2. Material and methods

2.1. Infection and RNA isolation

Vero cells were infected with SARS-CoV (Frankfurt-1 strain; NCBI Accession No. AY291315; Drosten *et al.*, 2003) at a multiplicity of infection of 0.01. At the onset of the cytopathogenic effect (approximately 40 h post-infection), intracellular RNA was isolated by cell lysis for 10 min at room temperature with 5% lithium dodecyl sulfate in LET buffer (100 m*M* LiCl, 1 m*M* EDTA, 10 m*M* Tris–HCl pH 7.4) containing 20 μg ml^{-1} of proteinase K. After shearing of the cellular DNA using a syringe, lysates were incubated at 315 K for 15 min, extracted with phenol (pH 4.0) and chloroform and the RNA was ethanol-precipitated. cDNA was obtained by reverse transcription using primer SAV009 (5′-GGACAG-CAACCGCTGGACAATC-3′), complementary to nucleotides 13644–13665 of the Frankfurt-1 genome, using Thermoscript reverse transcriptase (Invitrogen).

2.2. Subcloning, *Escherichia coli* protein expression and purification

The SARS-CoV Nsp9-coding sequence was amplified by PCR from the cDNA prepared above using two primers containing the attB sites of the Gateway recombination system (Invitrogen). At the 5′ end of the gene, a sequence encoding a hexahistidine tag was attached. The cDNA was then subcloned in the pDest14 plasmid (Invitrogen). The open reading frame of the final construct (referred to as pDest14/Nsp9-HN and encoding an N-terminally His-tagged version of SARS-CoV orf1a polyprotein residues

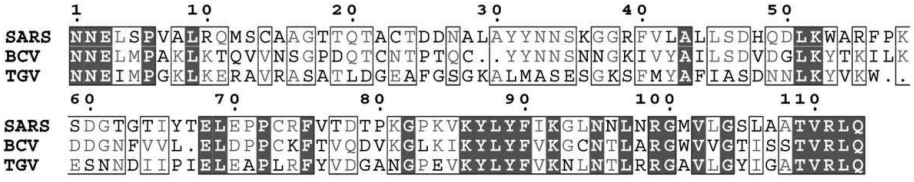


Figure 1
Alignment of the sequence of the SARS-CoV Nsp9 protein with that of bovine coronavirus (BCV) and of the transmissible gastroenterisis virus (TGV). Conserved residues are identified with a black background. Homologous residues are boxed.

electronic reprint

4118–4230) was checked by sequencing (MilleGen, Toulouse, France). Expression was performed in *E. coli* strain C41(DE3) (Avidis SA, France) transformed with the pLysS plasmid (Novagen). This plasmid carries the lysozyme gene, allowing tight regulation of the expression, and supplies the tRNAs for six rare codons used with a very low frequency in *E. coli*. Cultures were grown at 310 K until OD_{600} reached 0.6 and were then stored for 2 h on ice; 2% ethanol was added for the induction of stress chaperones (Gong & Shuman, 2002). Expression was induced by adding 50 μ*M* IPTG and cells were incubated for 16 h at 290 K. Cells were collected by centrifugation and the bacterial pellets were resuspended and frozen in 50 m*M* Tris–HCl, 150 m*M* NaCl, 10 m*M* imidazole pH 8.0.

Cellular suspensions were thawed with 0.25 mg ml^{-1} lysozyme, 0.1 μg ml^{-1} DNase and 20 m*M* $MgSO_4$ and were centrifuged at 12 000*g*. The supernatant was applied onto an Ni-affinity column connected to an FPLC system (Amersham Pharmacia Biotech). The protein was eluted with 50 m*M* Tris–HCl, 150 m*M* NaCl, 250 m*M* imidazole pH 8.0 and then applied onto a preparative Superdex 200 gel-filtration column pre-equilibrated in 10 m*M* Tris–HCl, 300 m*M* NaCl pH 8.0. The recombinant protein was characterized by N-terminal sequencing, mass spectroscopy, dynamic light scattering (DLS) and circular dichroism (CD).

2.3. Protein characterization

DLS was performed with a Dynapro Microsampler (Protein Solutions) using a protein solution at 5.8 mg ml^{-1} in 10 m*M* Tris–HCl, 300 m*M* NaCl pH 8.0. The CD spectrum of the final purified product was recorded between 185 and 260 nm on a JASCO J810 spectrometer using a protein solution at 0.1 mg ml^{-1} in sodium phosphate buffer pH 7.0 containing 25 m*M* NaCl.

2.4. Crystallization

Crystallization screening was performed by vapour diffusion with nanodrops using a Cartesian robot as described previously (Sulzenbacher *et al.*, 2002; Vincentelli *et al.*, 2003). Briefly, three commercial kits were used: Wizard Screens 1 and 2 (Emerald BioStructures), Structure Screens 1 and 2 and Stura Footprint screen (Molecular Dimensions Ltd). The crystals were obtained in 2.0 *M* ammonium sulfate, 0.1 *M* phosphate–citrate pH 4.2 and with a protein concentration of 5.8 mg ml^{-1} in the gel-filtration buffer. The optimization of the crystallogenesis was performed with nanodrops in a two-dimensional matrix (Lartigue *et al.*, 2003) with a precipitant range of 1.8–2.2 *M* ammonium sulfate and a pH range of 4.0–4.5 (0.1 *M* phosphate–citrate), leading to a crystal size of ~100 × 100 × 80 μm (Fig. 2).

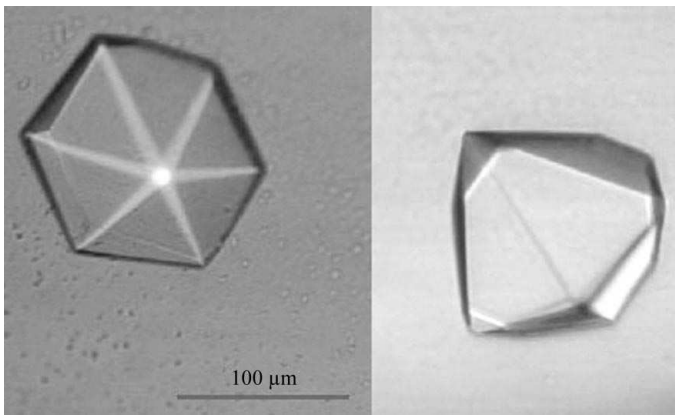


Figure 2
Optimized crystals of the SARS-CoV Nsp9 protein. The scale bar is 100 μm.

Table 1
Crystal parameters and data-reduction statistics of the Nsp9 protein crystals.

Values in parentheses are for the last resolution shell.

Space group	$P6_{1/5}22$
Unit-cell parameters (Å)	$a = b = 89.7$, $c = 136.7$
Beamline	ID14-EH1 at ESRF (λ = 0.934 Å)
Resolution (Å)	26.0–2.8 (2.94–2.8)
R_{sym} (%)	5.3 (28.1)
$I/\sigma(I)$	9.9 (2.5)
No. reflections	90899 (11486)
No. unique reflections	8395 (1166)
Completeness	98.7 (98.7)
Multiplicity	10.8 (9.9)

2.5. Data collection

The crystals were cryocooled in a pure solution of silicone oil DC200. They were exposed at beamline ID14-EH1, ESRF, Grenoble using a Quantum ADSC Q4R detector. A total of 110 1° oscillations were recorded with a crystal-to-detector distance of 180 mm and a collection time of 9 s per frame. Diffraction data were integrated with *DENZO* (Otwinowski & Minor, 1997) and were reduced with *SCALA* (Collaborative Computational Project, Number 4, 1994).

3. Results and discussion

3.1. *E. coli* protein expression and purification

We have subcloned 35 SARS-CoV targets in the Gateway system, including 20 full-length proteins and 15 protein domains. To date, 70 constructs have been generated, of which 28 were expressed, 14 were soluble and five were purified. Four of them led to small crystals, among which were those of the Nsp9 protein described in this report. Expression of selenomethionine-substituted Nsp9 was performed using the method of methionine-biosynthesis pathway inhibition (Doublié, 1997). Purification of the selenomethionine protein was performed as described above and crystal optimization is under way.

3.2. Data collection and reduction

Nsp9 crystals diffract to 2.8 Å at ID14-EH1 (ESRF, Grenoble). Data integration and reduction indicate that they belong to the *P*622 space group. R_{sym} is 5.3%, an excellent value considering the redundancy of the data (Table 1). Reflections are observed at multiples of six along the *c* axis

electronic reprint

(00*l*), indicating that the space group is either $P6_122$ or its enantiomorph $P6_522$. The unit-cell parameters are $a = b = 89.7$, $c = 136.7$ Å, which lead to a V_M value of 3.1 $Å^3\,Da^{-1}$ (60% solvent) with two molecules in the asymmetric unit (Matthews, 1968). The observed distribution of centric or acentric intensities overlaps with the theoretical curve, an indication that merohedral twinning, a feature that is often observed in trigonal or hexagonal crystals, is not present.

3.3. Characterization

SARS-CoV Nsp9 has been purified to homogeneity in two steps. The identity of the final product has been confirmed by N-terminal sequencing. The oligomeric status of Nsp9 has been checked using gel filtration and DLS. The former technique indicates that the protein is monomeric, while the DLS analysis is consistent with a monodisperse species with an apparent Stokes radius of 26 Å and an equivalent mass of 31 kDa, which corresponds to a dimer. This discrepancy might be related to the concentration differences between the two techniques.

A PSI-Blast search retrieved seven homologous sequences, all belonging to members of the *Coronaviridae* family. They were aligned using *MULTALIGN* (Corpet, 1988) with standard options. The consensus of the secondary-structure predictions obtained with *JPRED* (Cuff *et al.*, 1998), *PSI-PRED* (McGuffin *et al.*, 2000) and *PREDICT PROTEIN* (Rost, 1996) converges to a fold of seven β-strands. A fold-recognition analysis was performed with the threading programs *3D-PSSM* (Kelley *et al.*, 2000) and *INBGU* (Fischer, 2000). Both programs fail to detect any protein homologue to Nsp9, but converge to a fold of two seven-stranded β-sheets. In agreement, the CD spectrum of purified Nsp9 reveals a structured protein formed by a majority of β-strands (35%) and β-turns (18%), but which also contains 15% α-helix. Random-coil segments account for 32% of the total.

4. Conclusion

The SARC-CoV Nsp9 protein expressed in *E. coli* was readily crystallized using the nanodrop screening (Sulzenbacher *et al.*, 2002) and optimization (Lartigue *et al.*, 2003) approaches. Crystals diffract to 2.8 Å resolution and are amenable to structure determination using SeMet substitution and MAD methods (Hendrickson, 1991) at synchrotrons.

This study was funded by the SPINE project of the European Union 6th PCRDT (QLRT-2001-00988), by the French Genopole programme and by the Conseil General of the Bouches-du-Rhone. We thank H. W. Doerr and H. Rabenau (Institute for Medical Virology, Johan Wolfgang Goethe University, Frankfurt-am-Main, Germany) for providing us with the virus and P. Bredenbeek, S. Gorbalenya and W. Spaan for technical assistance and helpful discussions/suggestions.

References

Anand, K., Ziebuhr, J., Wadhwani, P., Mesters, J. R. & Hilgenfeld, R. (2003). *Science*, **300**, 1763–1767.
Bost, A. G., Carnahan, R. H., Lu, X. T. & Denison, M. R. (2000). *J. Virol.* **74**, 3379–3387.
Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.
Corpet, F. (1988). *Nucleic Acids Res.* **16**, 10881–10890.
Cuff, J. A., Clamp, M. E., Siddiqui, A. S., Finlay, M. & Barton, G. J. (1998). *Bioinformatics*, **14,** 892–893.
De Francesco, R., Tomei, L., Altamura, S., Summa, V. & Migliaccio, G. (2003). *Antivir. Res.* **58**, 1–16.
Doublié, S. (1997). *Methods Enzymol.* **276**, 523–530.
Drosten, C. *et al.* (2003). *N. Engl. J. Med.* **348**, 1967–1976.
Fischer, D. (2000). *Pac. Symp. Biocomput.* **5**, 119–130.
Gong, C. & Shuman, S. (2002). *J. Biol. Chem.* **277**, 15317–24.
Grotzinger, C., Heusipp, G., Ziebuhr, J., Harms, U., Suss, J. & Siddell, S. G. (1996). *Virology*, **222**, 227–235.
Hendrickson, W. A. (1991). *Science*, **254**, 51–58.
Kelley, L. A., MacCallum, R. M. & Sternberg, M. J. (2000). *J. Mol. Biol.* **299**, 499–520.
Koren, G., King, S., Knowles, S. & Phillips, E. (2003). *CMAJ*, **168**, 1289–1292.
Ksiazek, T. G. *et al.* (2003). *N. Engl. J. Med.* **348**, 1953–1966.
Lai, M. M.C. & Holmes, K. V. (2001). *Fields Virology*, 4th ed., edited by D. M. Knipe & P. M. Howley, pp. 1163–1185. Philadelphia: Lippincott Williams & Wilkins.
Lartigue, A., Rivière, S., Brossut, R., Tegoni, M. & Cambillau, C. (2003). *Acta Cryst.* D**59**, 916–918.
McGuffin, L. J., Bryson, K. & Jones, D. T. (2000). *Bioinformatics*, **16**, 404–405.
Matthews, B. W. (1968). *J. Mol. Biol.* **33**, 491–497.
Otwinowski, Z. & Minor, W. (1997). *Methods Enzymol.* **276**, 307–326.
Peiris, J. S., Lai, S. T., Poon, L. L., Guan, Y., Yam, L. Y., Lim, W., Nicholls, J., Yee, W. K., Yan, W. W., Cheung, M. T., Cheng, V. C., Chan, K. H., Tsang, D. N., Yung, R. W., Ng, T. K. & Yuen, K. Y. (2003). *Lancet*, **361**, 1319–1325.
Rost, B. (1996). *Methods Enzymol.* **266**, 525–539.
Snijder, E. J., Bredenbeek, P. J., Dobbe, J. C., Thiel, V., Ziebuhr, J., Poon, L. L. M., Guan, Y., Rozanov, M., Spaan, W. J. M. & Gorbalenya, A. E. (2003). In the press.
Sulzenbacher, G. *et al.* (2002). *Acta Cryst.* D**58**, 2109–2115.
Vincentelli, R., Bignon, C., Gruez, A., Sulzenbacher, G., Canaan, S., Tegoni, M., Campanacci, V. & Cambillau, C. (2003). *Acc. Chem. Res.* **36**, 165–172.

electronic reprint

ARTICLE 7

The severe acute respiratory syndrome-coronavirus replicative protein nsp9 is a single-stranded RNA-binding subunit unique in the RNA virus world.

Egloff MP, Ferron F, Campanacci V, Longhi S, Rancurel C, Dutartre H, Snijder EJ, Gorbalenya AE, Cambillau C, Canard B.

Proc. Natl. Acad. Sci. U S A. 2004 Mar 16;101(11):3792-6.

The severe acute respiratory syndrome-coronavirus replicative protein nsp9 is a single-stranded RNA-binding subunit unique in the RNA virus world

Marie-Pierre Egloff*, François Ferron*, Valérie Campanacci*, Sonia Longhi*, Corinne Rancurel*, Hélène Dutartre*, Eric J. Snijder†, Alexander E. Gorbalenya†, Christian Cambillau*‡, and Bruno Canard*‡

*Architecture et Fonction des Macromolécules Biologiques, Unité Mixte de Recherche 6098 Centre National de la Recherche Scientifique and Universités Aix-Marseille I et II, 31 Chemin Joseph Aiguier, 13402 Marseille Cedex 20, France; and †Molecular Virology Laboratory, Department of Medical Microbiology, Center of Infectious Diseases, Leiden University Medical Center, P.O. Box 9600, 2300 RC, Leiden, The Netherlands

Edited by Stephen C. Harrison, Harvard Medical School, Boston, MA, and approved January 28, 2004 (received for review November 26, 2003)

The recently identified etiological agent of the severe acute respiratory syndrome (SARS) belongs to *Coronaviridae* (CoV), a family of viruses replicating by a poorly understood mechanism. Here, we report the crystal structure at 2.7-Å resolution of nsp9, a hitherto uncharacterized subunit of the SARS-CoV replicative polyproteins. We show that SARS-CoV nsp9 is a single-stranded RNA-binding protein displaying a previously unreported, oligosaccharide/oligonucleotide fold-like fold. The presence of this type of protein has not been detected in the replicative complexes of RNA viruses, and its presence may reflect the unique and complex CoV viral replication/transcription machinery.

In 2003, a human coronavirus (CoV) was identified as the causative agent of a form of atypical pneumonia: severe acute respiratory syndrome-CoV (SARS-CoV) (1–5). *Coronaviridae* have the longest known single-stranded (ss)RNA genome (27–31.5 kb), with a complex genetic organization and sophisticated replication/transcription cycle (6, 7). Twenty-eight proteins are predicted to be encoded by the SARS-CoV genome (8, 9). The nonstructural (nsp) or "replicase" proteins of CoVs are derived from an unusually large replicase gene of >20 kb that consists of two large ORFs (ORFs 1a and 1b). Translation of this replicase gene from the incoming genomic RNA is the first step in CoV genome expression and includes a −1 ribosomal frameshift to express the ORF1b-encoded polypeptide. Translation products are the pp1a polyprotein (>4,000 amino acids) and the C-terminally extended pp1ab polyprotein (>7,000 amino acids), which are both cleaved by two or three ORF1a-encoded viral proteinases (10). Most of these replicase cleavage products assemble into a membrane-associated viral replication/transcription complex. Among other components, this complex includes a set of relatively small polypeptides (nsp6 to nsp11) encoded by the 3′ region of ORF1a, for which no predicted nor proven function has been assigned. For the mouse hepatitis CoV, several of these cleavage products were reported to colocalize with other components of the viral replication complex in the perinuclear region of the infected cell (11), suggesting their involvement (directly or indirectly) in viral RNA metabolism.

As part of a viral structural genomics program (12), we have cloned the 28 gene products of SARS-CoV and expressed them either as full-length proteins or as (predicted) functional domains. The determination of the three-dimensional structures of these gene products is expected to facilitate and accelerate discovery of drugs against this emerging and life-threatening pathogen. Furthermore, structural homology search is becoming a powerful method to infer biochemical and/or biological function of previously uncharacterized proteins. We report here the crystal structure of nsp9, one the SARS-CoV uncharacterized nonstructural protein, as well as evidence for its function as an ssDNA/RNA-binding protein.

Materials and Methods

Crystallization, Structure Determination, and Refinement. SARS-CoV nsp9 has been expressed, purified, and crystallized as described (12). X-ray diffraction data were collected at 100 K at the European Synchrotron Radiation Facility, Grenoble, France. Native data were collected on beamline ID14–1 by using a Quantum ADSC Q4R charge-coupled device detector. Crystals diffracted X-rays to 2.7-Å resolution and belonged to space group $P6_122$ with unit cell dimensions $a = b = 89.7$ Å, $c = 136.7$ Å. There are two molecules per asymmetric unit, leading to a solvent content of ≈60%. The structure was solved by using single-wavelength anomalous dispersion data (13) collected at the peak wavelength of selenium on beamline ID14–4.

Data were integrated with MOSFLM and were scaled by using SCALA (14). The four expected selenium sites (two in each of the two molecules of the asymmetric unit) were identified and refined by using SOLVE (15). Density modification of the experimental maps and initial fragment building was performed with RESOLVE (16). Model building was carried out by using TURBO-FRODO (17) and maximum-likelihood refinement was performed with REFMAC5 (18) using NCS restraints. Residues 3–113 could be modeled and refined in molecule A, and residues 4–113 in molecule B. Four sulfate ions and 31 water molecules were added manually during refinement. Overall geometric quality of the model was assessed by using PROCHECK (19). A total of 86.1% of the residues were found in the most favored regions of the Ramachandran plot, and 13.4% were in additionally allowed regions. The solvent-accessible surface of nsp9 was calculated and displayed by using GRASP (20) and SWISS-PDBVIEWER (21). Fig. 1 *A* and *B* was generated by using MOLSCRIPT (22) and was rendered by using RASTER3D (23). The alignment of Fig. 1*C* has been displayed by using the program ESPRIPT (24).

Surface Plasmon Resonance and Fluorescence Spectroscopy. DNA and RNA oligonucleotides were purchased from Life Technologies (Grand Island, NY) and Amersham Pharmacia, respectively. Surface plasmon resonance measurements were performed on a BIAcore apparatus (Pharmacia Biosensor) by using the BIAlogue kinetics evaluation program (BIAEVALUATION V.3.1, Pharmacia Biosensor). Biotinylated oligonucleotides were

This paper was submitted directly (Track II) to the PNAS office.

Abbreviations: SARS, severe acute respiratory syndrome; CoV, coronavirus; ss, single-stranded; SSB, ssDNA-binding protein; OB, oligosaccharide/oligonucleotide-binding; K_{Dapp}, dissociation equilibrium constant; NCBI, National Center for Biotechnology Information.

Data deposition: The atomic coordinates and structure factors have been deposited in the Protein Data Bank, www.pdb.org (PDB ID code 1QZ8).

‡To whom correspondence should be addressed. E-mail: cambillau@afmb.cnrs-mrs.fr or bruno@afmb.cnrs-mrs.fr.

© 2004 by The National Academy of Sciences of the USA

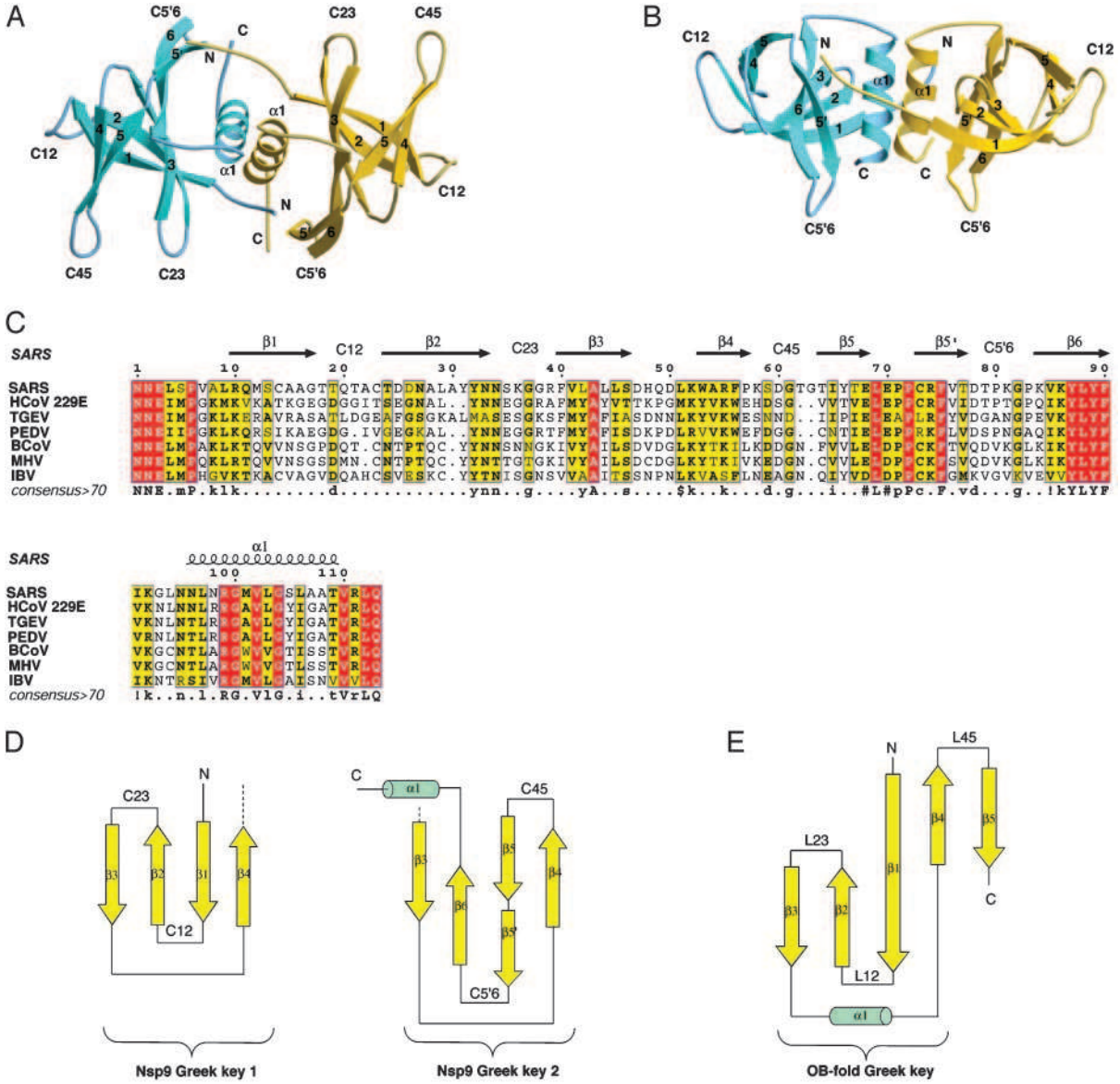


Fig. 1. Crystal structure, sequence, and topology of SARS-CoV nsp9. (*A*) Ribbon representation of SARS-CoV nsp9. One molecule of the dimer is gold and the other is cyan. Loops between strands x and y are labeled C. (*B*) A 90° view of *A*. (*C*) Multiple alignment of nsp9 sequences from SARS-CoV National Center for Biotechnology Information (NCBI) accession no. AY291315, and several related CoVs: HCoV 229E, human CoV 229E, NCBI accession no. NP_073550; TGEV, transmissible gastroenteritis virus, NCBI accession no. NP_058423; PEDV, porcine epidemic diarrhea virus CV777, NCBI accession no. NP_598309, BCoV, bovine CoV, NCBI accession no. NP_150074; MHV, mouse hepatitis virus MHV-A59, NCBI accession no. NP_045298; and IBV, avian infectious bronchitis virus, NCBI accession no. NP_040829). The consensus sequence (identity cutoff >70%) is displayed under the multiple sequence alignment. Dots and residues in lowercase correspond to positions for which the residue conservation is under and above the cutoff value, respectively; positions marked by # correspond to either Asn, Asp, Glu, or Gln; positions marked by ! correspond to either Ile or Val, and $ corresponds to Leu or Met. Residues that are conserved in all sequences are boxed in red, and those for which conservation is >70% are boxed in yellow. For a given position, only residues homologous to the consensus are bold. The top numbers correspond to the amino acid sequence of SARS-CoV nsp9. Secondary structure elements and loops of nsp9 SARS-CoV are numbered according to Fig. 1 and are indicated above the alignment. (*D*) Schematic representation of nsp9 topology. nsp9 SARS-CoV β-barrel structure is a concatenation of two Greek key motifs, Greek key 1 having a g− topology and Greek key 2 a g+ topology (30), resulting in a six-stranded RH-g− to g+ topology. β-strands and α-helices are symbolized by arrows and cylinders, respectively, and they are numbered consistently with the sequence alignment. (*E*) Schematic representation of the typical Greek key (g− topology) motif found in the OB fold.

immobilized on a Sensor Chip SA according to the manufacturer's instructions (BIAcore).

Fluorescence quenching of the single tryptophan in nsp9 was measured by using a Cary Eclipse (Varian) equipped with a front-face fluorescence accessory at 20°C, by using 2.5-nm excitation and 10-nm emission bandwidths. The excitation wavelength was 280 nm and the emission spectra were measured between 290 and 540 nm. Titrations were performed in a 1-ml quartz fluorescence cuvette containing 1 μM protein in 10 mM Tris·HCl buffer/300 mM NaCl, pH 8.0, and by the successive addition of aliquots of appropriate nucleic acids stock solutions (1 mM). Experimental fluorescence intensities were corrected for dilution. Data were analyzed by plotting the relative fluorescence intensities at 340 nm at increasing concentrations of quencher. Dissociation equilibrium constant (K_{Dapp}) values were determined from data fitted to a single exponential equation, by

BIOCHEMISTRY

Table 1. Summary of crystallographic data

	SeMet	Native
Data collection		
Wavelength, Å	0.9793	0.9340
Resolution range, Å†	30–3.0 (3.11–3.0)	3.0–2.7 (2.84–2.7)
No. of unique reflections	6,831	9,345
No. of measured reflections	97,058	100,668
$I/\sigma I$	7.4 (6)	9.4 (1.6)
Multiplicity	14.2 (14.2)	10.8 (10.2)
Completeness, %	99.9 (99.9 anomalous)	99.1 (99.0)
R_{merge}*, %	8.4 (47.9)	5.6 (43.3)
Refinement		
Resolution limits, Å		15.0–2.7
R factor†, R_{free}‡, %		23.4/27.1
rms deviation: bonds, Å/angles, °		0.018/1.88

SeMet, selenomethionyl single-wavelength anomalous dispersion data set. Values in parentheses are for the highest-resolution shell.

*R_{merge}, $\Sigma\Sigma i\,|Ih-Ii|/\Sigma\Sigma Ih$, where Ih is the mean intensity for reflection h.

†R factor, $\Sigma|F_o - F_c|/\Sigma\,|F_o|$, where F_o and F_c are measured and calculated structure factors, respectively.

‡R_{free} was calculated over 5% of reflections not used in the refinement.

using the PRISM 3.02 nonlinear regression tool (GraphPad, San Diego).

Results and Discussion

The crystal structure of SARS-CoV nsp9 is reported (Fig. 1 and Table 1). Crystals contain a dimer in the asymmetric unit (Fig. 1 *A* and *B*); in each monomer, seven β-strands and one α-helix (Fig. 1 *A–C*) are arranged into a single compact domain and form a cone-shaped β-barrel flanked by the C-terminal α-helix. The latter makes a 45° angle with the axis of the β-barrel and has a high content of hydrophobic residues, yielding two hydrophobic sides. One faces the β-barrel and the other interacts with the α-helix of the second crystallographic monomer (Fig. 1 *A* and *B*). This dimer is therefore assembled by hydrophobic interactions and is further stabilized by four long hydrogen bonds involving main-chain atoms. Comparing the buried dimerization surface of 1,632 Å^2 with the few other crystallographic contacts suggests that this crystallographic dimer is also present in solution, which is in agreement with dynamic light scattering and gel permeation experiments (12). This surface of 1,632 Å^2 is among standard interfacial areas found in biologically relevant dimers (25). The two molecules of the dimer are spatially similar (rms deviations value of 0.99 Å over the 109 C^α atoms of the structure). This deviation is further reduced after exclusion of N and C termini together with the tips of three long loops (L23, L45, and L5′6) emerging from the barrel (Fig. 1 *A*, rms deviation of 0.45 Å over the 87 C^α atoms).

Screening public protein databases with BLAST or PSI-BLAST (26) failed to identify any sequence homologue of CoV nsp9 proteins, which is consistent with their yet unknown function. No structural homologues of nsp9 were found when scanning either the Protein Data Bank with the DALI server (27) or the CATH database (28) with the GRATH server. Visual inspection of the Structural Classification of Protein database, version 1.63 (29), revealed some common features between SARS-CoV nsp9 and four different existing folds (trypsin-like proteases, the C-terminal domain of μ-transposase, α- and β-subunits of F1-ATP synthase-like, and small protein B). Only the N-terminal six-stranded β-barrel of the trypsin-like proteases displays the same connectivity and spatial arrangement as nsp9 monomers, and can be significantly superimposed [rms deviation of 1.73 Å over the

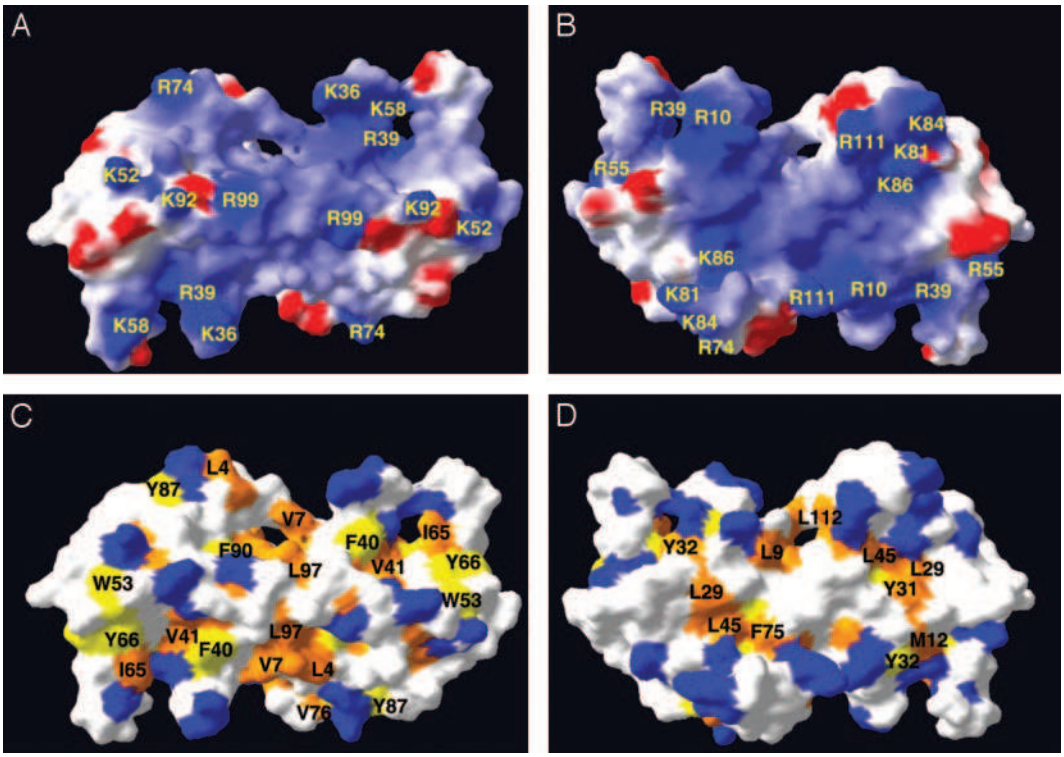


Fig. 2. Surface analysis of SARS-CoV nsp9. (*A*) Electrostatic surface potential of nsp9 viewed from the same orientation as in Fig. 1 *A*. Potential values range from −5 *kT* (red) to 0 (white) and to +5 *kT* (blue), where *k* is the Boltzman constant and *T* is the temperature. Accessible Lys and Arg residues are indicated. (*B*) Back view from *A*, with the same color code. (*C*) Accessible surface colored according to the following: Lys or Arg are blue; Tyr, Phe, and Trp are yellow; and Val, Leu, and Ile are orange. The orientation is the same than in *A*. (*D*) Same coloring as in *C* with the same orientation as in *B*.

74 C^{α} atoms when superimposing nsp9 and thrombin structure (PDB code 1A2C)]. This homology has no functional relevance, however, because nsp9 lacks a trypsin-like catalytic triad. This visual inspection did, however, reveal two other interesting features. First, six-stranded closed β-barrels have been identified in several proteins that interact with RNA, such as the small protein B (30) and the domain III of EF-Tu (31). Second, the short six-stranded β-barrel of nsp9 includes an opened five-stranded barrel reminiscent of the five-stranded β-barrel of the oligosaccharide/oligonucleotide-binding (OB)-fold proteins. The latter proteins form a superfamily in which two-thirds of the members are nucleic acid-binding proteins (32).

Structural homology between nsp9 and small protein B (or domain III of EF-Tu) is not strong enough to allow a reliable superimposition, which would have provided an indication about the localization of a putative nucleic acid-binding site. Likewise, nsp9 and OB-fold proteins cannot easily be compared, because Greek key 1 of nsp9 (Fig. 1*D*) and the classical OB-fold Greek key motif (Fig. 1*E*) cannot be superimposed. Although there is a structural equivalence between the Greek key 2 motif of nsp9 (Fig. 1*D*) and the OB-fold Greek key, the connectivity between strands is different. When both motifs are superimposed, the canonical binding face observed in the OB fold is buried in the dimer interface of nsp9. For these reasons, nsp9 may be considered as a new variant within the OB superfamily. Nonetheless, nsp9 displays the same features that OB-fold proteins use to bind nucleic acids: a network of positively charged amino acids defines a positive track suitable for binding the phosphate backbone to the protein surface (Fig. 2 *A* and *B*), whereas exposed aromatic residues might provide stacking interactions with nucleobases (Fig. 2 *C* and *D*). These residues are conserved in all CoV nsp9 sequences (Arg-10, Lys-52, Trp-53, Arg-55, Arg-74, Phe-75, Lys-86, Tyr-87, Phe-90, Lys-92, Arg-99, and Arg-111 in Figs. 1*C* and 2*A–D*), further suggesting that nsp9 is a nucleic acid-binding protein. In addition, two extended loops L23 and L45 display weak electron density associated with high *B* factor values, indicating that they are flexible and/or mobile. They line the positively charged track, and they may clamp nucleic acids on the nsp9 surface after conformational change, as observed in other OB-fold proteins (32). In other members of the OB-fold superfamily, each monomer has its own, autonomous single-stranded nucleic acid-binding site. For example, replication protein A trimerizes by means of its C-terminal α-helix, each monomer keeping an individual ssDNA-binding site acting cooperatively with other units of the trimer (33, 34). In nsp9, it is the dimeric form that provides a single, uninterrupted nucleic acid-binding site.

Surface plasmon resonance was used to demonstrate the function of nsp9 as a nucleic acid-binding protein. Biotinylated oligonucleotides bound to a streptavidin-coated solid support are able to bind nsp9 (Fig. 3*A*). This function was confirmed by fluorescence experiments. As a fluorophore, nsp9 monomer has a single Trp residue (Trp-53), which is partially exposed to the solvent. The Trp-53 indole moiety is in a polar environment comprising side chains of Gln-20, Gly-66, and more remotely, Lys-52. Interactions of Trp-53 with ligand might therefore quench its fluorescence. This occurrence was indeed observed by using ssDNA and ssRNA oligonucleotides of defined sequence. The quenching efficiency increased steadily when the probe size was increased from 6-mer to 45-mer and then reached a plateau (Fig. 3*B*). The occurrence of this plateau suggests that the nsp9 tryptophans in the dimer achieve an optimal energy transfer (reflecting optimal molecular interactions) only through the tight packing of probes equal or longer than the 45-mer. In contrast, shorter probes do not result in optimal transfer, probably due to remote or loose contacts with the tryptophans. With both ssDNA and ssRNA, a large decrease in tryptophan fluorescence was observed (Fig. 3*C*), but the emission maximum

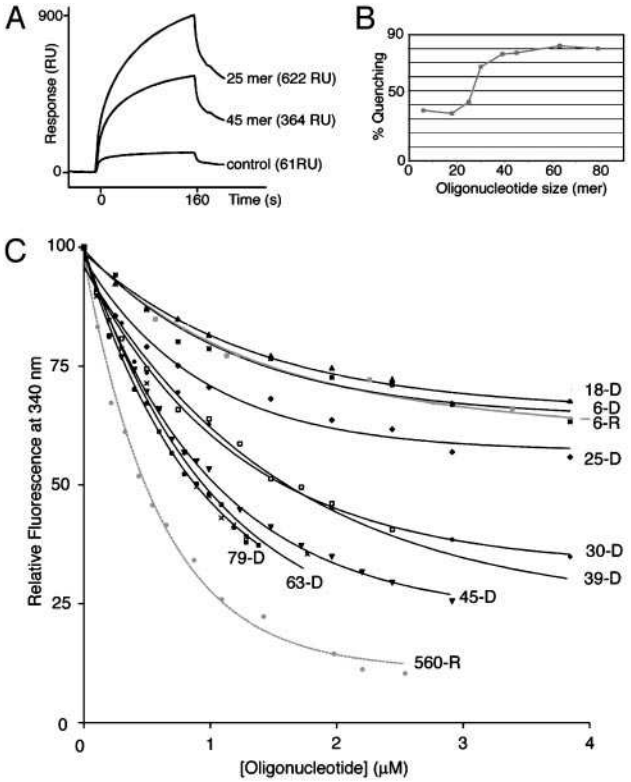


Fig. 3. SARS-CoV nsp9 is an oligonucleotide-binding protein. (*A*) BIAcore analysis of nsp9 binding to immobilized DNA oligonucleotides. The protein (16 μM) was injected at a flow rate of 5 μl/min in HBS buffer on dextran layers containing 550 and 850 resonance units (RU) of the 25- and 45-mer oligonucleotides, respectively. The sensorgrams are the result of two independent experiments. RU values at 125 s (5 s after the end of the injection) are indicated. (*B*) Tryptophan fluorescence quenching study on SARS-CoV nsp9. The tryptophan fluorescence quenching at the plateau (in percent) is plotted versus the length of ssDNA probes. The K_{Dapp} values are extracted from the plot displayed in *C* and are discussed in the text. (*C*) The relative fluorescence of nsp9 at 340 nm is plotted as a function of the oligonucleotide concentration for ssDNA, ranging from 6- to 79-mer and for a 6, and 560-mer ssRNA. D, DNA; R, RNA. The K_{Dapp} values (discussed in the text), result from the fitting of the data to a single exponential (GraphPad).

wavelength was unchanged, indicating (*i*) that the interaction is nonspecific and may involve the sugar-phosphate backbone rather than the bases, and (*ii*) that the environment of Trp-53 remains polar. The apparent affinity does not depend on the ssDNA length, because K_{Dapp} values fall between 0.63 and 1.1 μM when data are fitted to a single exponential (Fig. 3*C*). The strongest quenching (90%) and the best K_{Dapp} (0.4 μM) are observed with the 560-mer ssRNA, suggesting that each ssRNA binds several nsp9 dimers and that each nsp9 dimer can bind two distinct single-stranded segments.

The binding of both ssDNA and ssRNA of unrelated defined sequences, together with K_{Dapp} values in the micromolar range, suggests that the nucleic acid-binding activity of nsp9 is not sequence-specific. Much like the human CoV 229E helicase, which has RNA and DNA duplex-unwinding activities (35), nsp9 is able to bind ssDNA or ssRNA equally, although binding of the latter is expected to be the native function. In the infected cell, the coupling/compartimentation of the viral RNA synthesis with the RNA-binding function of nsp9 might render RNA versus DNA specificity unnecessary. The wrapping of ssRNA around the nsp9 dimer is an interesting possibility that is compatible with the structural characteristics of the nsp9 dimer described here. An ssRNA binding-function of nsp9 is also consistent with its

natural abundance in the replication complex. Due to a ribosomal frameshifting mechanism (36), nsp9 and other ORF1a-encoded CoV replicase subunits are produced in 3- to 5-fold excess relative to the "core" replicative enzymes [such as the RNA-dependent RNA polymerase and helicase (9) produced from replicase ORF1b]. For example, nsp9 might stabilize nascent nucleic acid during replication or transcription, thus providing protection from nucleases. The amount of nsp9 may not be enough to cover the entire ssRNA genome. The latter may not be entirely single-stranded, however, due to secondary RNA structure. Whether the ssRNA-binding function of nsp9 may be restricted to specific segments of the genome, or be complemented with other proteins is still an open question.

The complexity of the RNA synthesis machinery of CoVs has long been predicted, considering the size of the pp1a- and pp1ab-replicative polyproteins and the number of cleavage products produced from these precursors. Recently, Snijder *et al.* (8) described a set of putative RNA processing enzymes in the replicase complex of CoVs, including SARS-CoV. In addition to mere RNA replication, nsp9 could also participate in such a base-pairing-driven process as RNA processing. An informative parallel in the virus world is observed with bacteriophage T7: its gene 2.5 ssDNA-binding protein binds substrates with similar affinity as SARS-CoV nsp9 does [k_d in the μM range (37)] and is involved in replication/recombination/homologous base-pairing events (38, 39). The structural and functional characterization of nsp9 may also be relevant to SARS-CoV control: the SARS epidemics as well as previous work on CoVs have shown that genome plasticity (evolution by mutation and recombination) relate to pathogenicity and probably also to drug resistance. Because many viral and cellular single-stranded nucleic acid-binding proteins are essential (40), nsp9 is to be added to the list of potential targets for anti-CoV drug design.

We thank the staff at the European Synchrotron Radiation Facility for technical assistance and, in particular, Joanne Mac Carthy and Edward Mitchell for their assistance with data collection; Laurent Gauthier, Aurelia Salomoni, Cécile Durousseau, Fabienne Tocque, Nicolas Brémond, Willy Spaan, Peter Bredenbeek, Jessika Dobbe, and Sylvie Doublié for their contribution to this work; and Holli Conway for correcting the English. This work was supported by the Structural Proteomics in Europe project of the European Union 5th framework research program (Grant QLRT-2001-00988), by the French Genopole program, and by the Conseil Général des Bouches-du-Rhône.

1. Drosten, C., Gunther, S., Preiser, W., van der Werf, S., Brodt, H. R., Becker, S., Rabenau, H., Panning, M., Kolesnikova, L., Fouchier, R. A., *et al.* (2003) *N. Engl. J. Med.* **348,** 1967–1976.
2. Ksiazek, T. G., Erdman, D., Goldsmith, C. S., Zaki, S. R., Peret, T., Emery, S., Tong, S., Urbani, C., Comer, J. A., Lim, W., *et al.* (2003) *N. Engl. J. Med.* **348,** 1953–1966.
3. Marra, M. A., Jones, S. J., Astell, C. R., Holt, R. A., Brooks-Wilson, A., Butterfield, Y. S., Khattra, J., Asano, J. K., Barber, S. A., Chan, S. Y., *et al.* (2003) *Science* **300,** 1399–1404.
4. Peiris, J. S., Lai, S. T., Poon, L. L., Guan, Y., Yam, L. Y., Lim, W., Nicholls, J., Yee, W. K., Yan, W. W., Cheung, M. T., *et al.* (2003) *Lancet* **361,** 1319–1325.
5. Rota, P. A., Oberste, M. S., Monroe, S. S., Nix, W. A., Campagnoli, R., Icenogle, J. P., Penaranda, S., Bankamp, B., Maher, K., Chen, M. H., *et al.* (2003) *Science* **300,** 1394–1399.
6. Siddell, S. G. (1995) *The Coronaviridae* (Plenum, New York).
7. Lai, M. M. C. & Holmes, K. V. (2001) in *Fields Virology*, eds. Knipe, D. M. & Howley, P. M. (Lippincott, Williams & Wilkins, Philadelphia), Vol. 1, pp. 1163–1185.
8. Snijder, E. J., Bredenbeek, P. J., Dobbe, J. C., Thiel, V., Ziebuhr, J., Poon, L. L., Guan, Y., Rozanov, M., Spaan, W. J. & Gorbalenya, A. E. (2003) *J. Mol. Biol.* **331,** 991–1004.
9. Thiel, V., Ivanov, K. A., Putics, A., Hertzig, T., Schelle, B., Bayer, S., Weissbrich, B., Snijder, E. J., Rabenau, H., Doerr, H. W., *et al.* (2003) *J. Gen. Virol.* **84,** 2305–2315.
10. Ziebuhr, J., Snijder, E. J. & Gorbalenya, A. E. (2000) *J. Gen. Virol.* **81,** 853–879.
11. Bost, A. G., Carnahan, R. H., Lu, X. T. & Denison, M. R. (2000) *J. Virol.* **74,** 3379–3387.
12. Campanacci, V., Egloff, M. P., Longhi, S., Ferron, F., Rancurel, C., Salomoni, A., Durousseau, C., Tocque, F., Brémond, N., Dobbe, J. C., *et al.* (2003) *Acta Crystallogr. D* **59,** 1628–1631.
13. Dauter, Z., Dauter, M. & Dodson, E. (2002) *Acta Crystallogr. D* **58,** 494–506.
14. Collaborative Computational Project 4 (1994) *Acta Crystallogr. D* **50,** 760–763.
15. Terwilliger, T. C. & Berendzen, J. (1999) *Acta Crystallogr. D* **55,** 849–861.
16. Terwilliger, T. C. (2002) *Acta Crystallogr. D* **58,** 1937–1940.
17. Roussel, A. & Cambillau, C. (1991) *Silicon Graphics Directory* (Silicon Graphics, Mountain View, CA).
18. Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997) *Acta Crystallogr. D* **53,** 240–255.
19. Vaguine, A. A., Richelle, J. & Wodak, S. J. (1999) *Acta Crystallogr. D* **55,** 191–205.
20. Nicholls, A., Sharp, K. A. & Honig, B. (1991) *Proteins* **11,** 281–296.
21. Guex, N. & Peitsch, M. C. (1997) *Electrophoresis* **18,** 2714–2723.
22. Kraulis, P. J. (1991) *J. Appl. Crystallogr.* **24,** 946–950.
23. Merritt, E. A. & Bacon, D. J. (1997) *Methods Enzymol.* **277,** 505–524.
24. Gouet, P., Courcelle, E., Stuart, D. I. & Metoz, F. (1999) *Bioinformatics* **15,** 305–308.
25. Janin, J. & Rodier, F. (1995) *Proteins* **23,** 580–587.
26. Altschul, S. F. & Koonin, E. V. (1998) *Trends Biochem. Sci.* **23,** 444–447.
27. Holm, L. & Sander, C. (1995) *Trends Biochem. Sci.* **20,** 478–480.
28. Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B. & Thornton, J. M. (1997) *Structure (London)* **5,** 1093–1108.
29. Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995) *J. Mol. Biol.* **247,** 536–540.
30. Gutmann, S., Haebel, P. W., Metzinger, L., Sutter, M., Felden, B. & Ban, N. (2003) *Nature* **424,** 699–703.
31. Nissen, P., Kjeldgaard, M., Thirup, S., Clark, B. F. & Nyborg, J. (1996) *Biochimie* **78,** 921–933.
32. Arcus, V. (2002) *Curr. Opin. Struct. Biol.* **12,** 794–801.
33. Bochkarev, A., Bochkareva, E., Frappier, L. & Edwards, A. M. (1999) *EMBO J.* **18,** 4498–4504.
34. Bochkareva, E., Korolev, S., Lees-Miller, S. P. & Bochkarev, A. (2002) *EMBO J.* **21,** 1855–1863.
35. Seybert, A., Hegyi, A., Siddell, S. G. & Ziebuhr, J. (2000) *RNA* **6,** 1056–1068.
36. Brierley, I., Boursnell, M. E., Binns, M. M., Bilimoria, B., Blok, V. C., Brown, T. D. & Inglis, S. C. (1987) *EMBO J.* **6,** 3779–3785.
37. Kim, Y. T., Tabor, S., Bortner, C., Griffith, J. D. & Richardson, C. C. (1992) *J. Biol. Chem.* **267,** 15022–15031.
38. Kong, D. & Richardson, C. C. (1996) *EMBO J.* **15,** 2010–2019.
39. Kong, D., Nossal, N. G. & Richardson, C. C. (1997) *J. Biol. Chem.* **272,** 8380–8387.
40. Glassberg, J., Meyer, R. R. & Kornberg, A. (1979) *J. Bacteriol.* **140,** 14–19.

-Commentaires

Ces deux articles sont le résultat de l'analyse complète du génome du virus du SRAS à l'aide de VaZyMolO. Lorsque l'épidémie du SRAS a éclaté, nous n'avions annoté aucun *Coronavirus* dans la base de données. En revanche, nous avions déjà procédé à l'annotation des *Flaviviridae* et *Togaviridae,* ce qui nous avait déjà permis de mettre en évidence des signatures pour les hélicases, les protéases, les méthyltransférases, ainsi qu'une série de motifs de site de liaison du zinc. L'utilisation du BLAST interne couplée à l'analyse des tracés HCA s'est montrée déterminante pour délimiter les domaines hydrophobes et désordonnés. A l'aide de la bibliographie, nous avons validé la cartographie effectuée et optimisé quelques bornes. L'extension de cette cartographie aux protéines du virus SRAS a été ensuite très rapide. Néanmoins, nous avons constaté des divergences au sein des réplicases de Coronavirus, dans la partie amino-terminale de la « réplicase ». Les deux premières protéines, nsp1 et nsp2, sont propres au virus du SRAS, et nsp3, une protéine de plus de 1900 aa, contient un domaine additionnel que l'on ne retrouve pas chez les autres Coronavirus. L'utilisation de la bibliothèque de motifs de VaZyMolO, nous a permis d'identifier les sites catalytiques et d'attribuer les fonctions aux modules correspondants.

Nous avons donc identifié 28 protéines dont 16 issues de la « réplicase ». Le découpage modulaire se traduit par 62 modules répartis en 8TM, 2HD, 3PS, 35F, 4S, 1M, 1DISF, et 8UNK. La "réplicase" compte 22 modules F. Nous avons réussi à exprimer de façon soluble dix de ces modules, dont nsp9 (Figure 30).

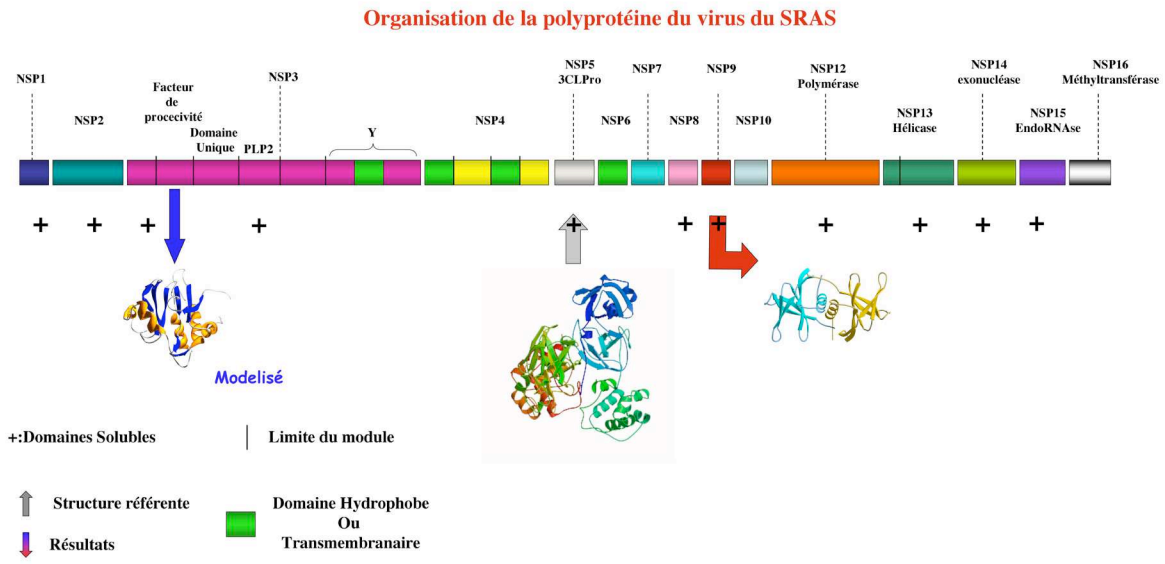


Figure 30 : Organisation de la réplicase.

Le cas de NSP9 est un parfait exemple de ce que peut apporter une approche bioinformatique combinée à une approche de génomique structurale. La protéine nsp9 du SRAS fait partie des protéines orphelines ne présentant aucune similarité de séquence en dehors du genre Coronavirus. La résolution de la structure de nsp9 a révélé un nouveau repliement structural. L'inspection de la base de données CATH à permis ensuite de mettre en évidence une architecture rappelant celle des protéines de repliement de type OB (« Oligonucleotide Binding »).
Cette analyse a permis d'orienter les investigations biochimiques vers des essais de liaisons à l'ARN et à l'ADN.

Ainsi, nous avons pu montrer que NSP9 est une protéine de liaison à l'ARN, bien que son rôle n'ait pas été encore clairement défini. Elle participe probablement à la stabilisation de l'ARN au sein du complexe de réplication (Prentice et al. 2004; Sutton et al. 2004). Dans l'article, nous proposons un parallèle avec le bactériophage T7 et nous proposons un rôle pour NSP9 similaire à celui du produit du gène 2.5. Cette hypothèse s'appuie sur le fait que NSP9 possède un repliement similaire, et elle se fixe à l'ARN/ADN avec une affinité comparable à celle du produit du gène 2.5. Ce dernier joue un rôle essentiel lors de la réplication et des événements de recombinaisons et réparation de l'ADN. Cette observation soulève donc une question sur le plan de l'évolution qui mériterait d'être ultérieurement approfondie. Avec un génome de 30 Kb, les Coronavirus possédent parmi les plus grands génomes viraux (hormis les génomes du Mimivirus 1.2 Mb (Raoult et al. 2004) et de bactériophages 0.93 Mb (Lobocka et al. 2004)). Mis à part la fonction ligase et celle de reconnaissance de mesappariement, la "réplicase" porte presque l'intégralité des fonctions permettant de corriger les erreurs sur le génome et laisse penser qu'un mécanisme de réparation des bases incorrectement insérées existe, et rappelle celui de la bactérie (système « mismatch repair MutHLS »). Il est donc permis de penser que les Coronavirus pourraient posséder un système de réparation de leur ARN. La capacité de nsp9 à se lier à d'autres protéines du complexe de réplication (Sutton et al. 2004) suggère un rôle de coordination des différentes réactions lors d'éventuelles étapes de réparation du génome. Elle pourrait notamment permettre le positionnement correct du complexe réplicatif sur les segments d'ARN devant être réparés. Si cette hypothèse venait à se confirmer ce serait remarquable du fait que ce mécanisme n'a jamais été montré chez aucun virus à ARN.

Dans ce cas, les Coronavirus ne seraient-ils pas des virus ancestraux qui dériveraient d'un organisme plus complexe ? La question reste en suspens, mais l'étude structurale et fonctionnelle en vue de la compréhension de leur mécanisme de réplication pourrait permettre de répondre en partie à cette question.

2.2. Les virus de l'hépatite G, ou virus GB

2.2.1. Généralités

Sur la base de la comparaison de séquences et de la structure du génome, les virus GB ont été classés au sein des *Flaviviridae*. Cette famille comprend quatre genres : les Pestivirus (exemple : virus de la diarrhée virale bovine (BVDV)), les Hepacivirus (exemple : virus de l'hépatite C (HCV)), les Flavivirus (exemple : Virus de la Dengue) et les non classés (exemple : les virus GB) (Lindenbach and Rice 2001). Le plus proche homologue des virus GB est le virus de l'hépatite C (Lindenbach and Rice 2001). Le pseudo-genre GB se compose de trois type de virus GBV-A, GBV-B et GBV-C. Les deux premiers infectent les singes, alors que GBV-C infecte l'homme. Ce dernier fut découvert en 1995 dans des sérums de malades ayant présenté une hépatite inexpliquée (Simons et al. 1995a; Simons et al. 1995b; Linnen et al. 1996). Depuis 1998, diverses équipes ont montré que l'évolution des sujets infectés par le virus de l'immunodéficience humaine (VIH) vers le SIDA était plus lente lorsqu'ils étaient coinfectés par le virus dit « de l'hépatite G » (GBV-C ou VHG) (Heringlake et al. 1998; Toyoda et al. 1998; Lefrere et al. 1999b). GBV-C, en dépit de son appellation, n'est en rien un virus d'hépatite et demeure, malgré de nombreuses recherches, un virus orphelin de maladie. A ce jour, l'infection qu'il entraîne apparaît totalement asymptomatique. Elle se traduit par une virémie élevée (attestant un haut niveau de réplication), qui perdure durant plusieurs années (Lefrere et al. 1996; Lefrere et al. 1997; Lefrere et al. 1999a), pour s'éteindre ensuite brusquement, laissant place à des anticorps sériques dirigés contre l'enveloppe virale (Dille et al. 1997; Tacke et al. 1997).

2.2.2. Organisation de la particule virale et du génome

Ce sont des virus enveloppés ayant généralement une forme sphérique avec un diamètre compris entre 40 et 60 nm. Leur enveloppe est constituée par une bicouche lipidique qui provient de la membrane plasmique de la cellule infectée et dans laquelle sont insérées les protéines d'enveloppe (E1 et E2). A l'intérieur du virus on trouve une nucléocapside se composant de l'ARN génomique associé (GBV-B) ou non (GBV-A et GBV-C) à la protéine de capside C (Linnen et al. 1996).

Le génome du GBV-C est constitué par de l'ARN simple brin, à polarité positive, d'environ 9400 nucléotides. Il code pour une polyprotéine contenant deux protéines structurales et pour six protéines non structurales (Muerhoff et al. 1995; Linnen et al. 1996).

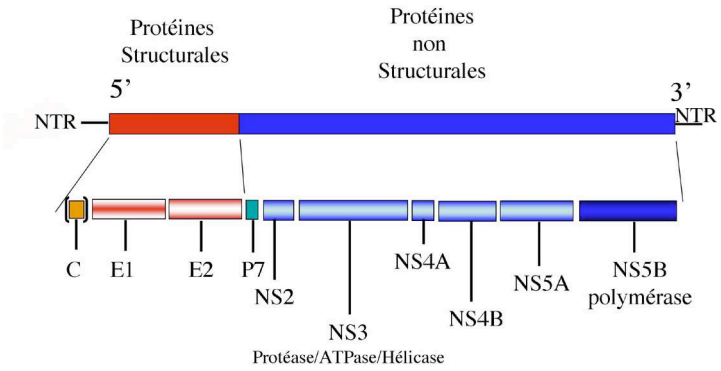


Figure 31 : Organisation du génome des GB virus.

Au cours de ma thèse, j'ai travaillé sur la modélisation de la polymérase du virus GBV-C, la protéine NS5b. La comparaison des structures des polymérases de l'hépatite C (HCV) et du virus de la diarrhée virale bovine (BVDV) m'a permis de proposer un mécanisme pour l'initiation de la réplication. Ce travail est présenté dans l'article suivant.

ARTICLE 8

The modeled structure of the RNA dependent RNA polymerase of the GBV-C virus suggests a role for motif E in *Flaviridae* RNA polymerases

François Ferron, Cécile Bussetta, Hélène Dutartre, and Bruno Canard

Article en préparation pour BMC bioinformatics

The modeled structure of the RNA dependent RNA polymerase of the GBV-C Virus suggests a role for motif E in *Flaviviridae* RNA polymerases.

François Ferron, Cécile Bussetta, Hélène Dutartre, and Bruno Canard*

Architecture et Fonction des Macromolécules Biologiques, UMR 6098 CNRS et Université Aix-Marseille I et II, ESIL, Campus de Luminy, 13288 Marseille Cedex 09, France
*e-mail : bruno.canard@afmb.cnrs-mrs.fr

Abstract

Background: The *Flaviviridae* virus family includes major human and animal pathogens. The RNA dependent RNA polymerase (RdRP) plays a central role in the replication process, and is thus a validated target for antiviral drugs. Despite the increasing structural and enzymatic characterization of viral RdRps, detailed replication mechanisms remain unclear at the molecular level. Hepatitis C virus (HCV) is a major human pathogen difficult to study in cultured cells, Bovine viral diarrhea virus (BVDV) is often used as a surrogate model to screen antiviral drugs against HCV. The structure of BVDV RdRP was recently published. This latter structure presents several differences with that of HCV. These differences raise questions about the relevance of the BVDV model, and cast a novel interest on GBV-C. Indeed, GBV-C is genetically closer to HCV than BVDV, and can lead to productive infection of culture cells. There is no structural data for the GBV-C polymerase.

Results: We show in this study that the GBV-C RNA polymerase is the closest representative to the HCV RNA polymerase. We report a 3D model of the GBV-C RNA polymerase, developed using sequence-to-structure threading and comparative modelling based on the atomic coordinates of the HCV polymerase structure. Analysis of the predicted structural features in the phylogenetic context of the RNA polymerase family allows us to rationalize most of the experimental data available. Available structures and our model are explored to examine the catalytic cleft, allosteric and substrate binding sites.

Conclusions: Computational methods were used to infer evolutionary relationships

and to predict the structure of a viral RNA polymerase. Docking a GTP molecule into the structure allows us to define a GTP binding pocket in GBV-C, such as that of BVDV. Our structural model suggests an alternative proposition to [1] for mechanism of initiation of RNA synthesis, and may prove useful to design new experiments to implement our knowledge on the mechanism of RNA polymerases.

Background

The *Flaviviridae* virus family comprises three genera pestivirus, hepacivirus, and the large group of flavivirus. Hepatitis C virus (HCV) causes acute and chronic hepatitis that may lead to cirrhosis and/or liver cancer. Nowadays, HCV is a major human pathogen, with 170 million people infected over the world and 3 to 4 million of people newly infected each year. Despite its large socio-economic impact, there is neither a vaccine nor an efficient, side-effect free therapy against this virus. Thus, the identification of potent drugs would be a major public health achievement. However, convenient small-animal models or productively infected cell systems to study HCV are still lacking. Consequently, compounds are often directly validated in HCV infected chimpanzees, or in cultured cells infected with related, surrogate viruses such as pestiviruses. The latter are animal pathogens showing similarity to hepaciviruses and flaviviruses [2] in genome structure, translation strategy, and individual gene products.

The RNA-dependent RNA polymerase (RdRP) is an enzyme playing a key role in the RNA replication process. Despite the increasing number of studies on the characterization of RdRP's activity and structure, the polymerase mechanism remains unclear at the molecular level. The postulated RNA replication process is a two-step mechanism. First, the initiation step of RNA synthesis begins at or near the 3' end of the (+) RNA template by means of a primer-independent (*de novo*) mechanism [3]. The *de novo* initiation consists in the addition of a nucleotide tri-phosphate (NTP) to the 3'-OH of the first initiating NTP. At the following so-called elongation phase, this nucleotidyl transfer reaction is repeated with subsequent NTPs to generate the complementary RNA product [3-6].

The structure of the RdRP of HCV (NS5B) has been determined [7, 8] and serves as a reference in elaborating mechanism and links between structure and biochemical data for RNA polymerases [9]. The HCV polymerase shape is a semi-closed right hand made of three subdomains : fingers, palm and thumb. Computational and structural analysis of viral RdRP sequences has identified five universal motifs (F, A, B, C, E) located in (or close to) the palm (additional file 1). These motifs are both catalytic and structural. Fingers are made of a β-strand subdomain (four strands β1, β2, β4, β5 and an α-helix α1) and an α-helix rich subdomain (seven helices αA, αB, αC, αD, αE, αF, αH). The palm is made of three stranded antiparallel β-sheet (β3, β6, β7), and three helices (αG, αJ, αK). The thumb is mainly made of α-helices (αN, αM, αL, αQ, αO, αP, αR) and a two-stranded antiparallel β-sheet (β10, β11) forming an extra structure called "the flap". The flap is supposed to play a role in the initiation mechanism, allowing only ssRNA to access the active site, and helping the correct positioning of the first two nucleotides[7].

Based on the polymerase structure solved in complex with NTPs [7], several GTP and NTP binding sites have been proposed. One is located behind the thumb, in a pocket on the surface of the structure, and has been called the allosteric (or surface) GTP binding site. The second one is in the catalytic cavity, where NTP can bind at different sites called P (priming), C (catalytic), and I (interrogating) sites. Recently, the crystal structure of the RdRP of Bovine viral diarrhea virus (BVDV) was published [10]. Another GTP binding site was found in the catalytic site, distinct from the P, C, and I sites of HCV NS5B. In the latter structure, this site corresponds to a cavity filled with water.

BVDV and HCV polymerases share a similar fold (Figure 1), but exhibit differences in the fingers and thumb subdomains due to a difference in the number of secondary structure elements. As for the HCV polymerase, the shape of the BVDV polymerase is a semi-closed right hand made of fingers, palm, and thumb. Fingers are made of eleven β-strands, and twelve α-helices. The palm domain shows great conservation with the HCV palm domain. It consists of four strands forming a central β-sheet surrounded by three α-helices. The thumb contains height α-helices and five β-strands. The flap is lacking in BVDV RNA polymerase although Choi & al [10] proposed that two β-strands with their connecting loops play the same role.

A number of structural differences in the flap and other subdomains raise the question of the relevance of BVDV as a surrogate model to discover HCV RNA polymerase inhibitors. Few years ago, "GB" viruses were identified and characterized as *Flaviviridae* agent leading to hepatitis [2] but not belonging to hepacivirus. Previous phylogenetic studies of GBV viruses were based on NS3 sequence comparisons [2]. Out of the three GB viruses identified so far, namely GBV-A, -B, and -C, two of them (GBV-A and GBV-B) are most likely monkey viruses while GBV-C can infect humans. HCV and GB virus genomes organisation are similar [11, 12]. This similarity has been extended to the functional level with the characterisation of the polymerase activity carried out by NS5B [13, 14]. GBV-C virus can lead to a productive infection of cultured cells, which makes it a relevant alternate virus to be used as a model for HCV antiviral drug screenings. In this study, we show using an NS5B-based phylogenetic analysis that GB viruses indeed carry the closest known RNA polymerases to HCV in *Flaviviridae*. We have built a structural model for the GBV-C polymerase which allows comparative analysis with HCV and BVDV polymerase. Results presented in this paper suggest a novel model for the initiation of RNA synthesis in *Flaviviridae*. Due to its phylogenetic and nearness to HCV, GBV-C might be an alternate and more relevant surrogate viral system to HCV than BVDV. Finally, the GBV-C polymerase model proposed in this study might help drug discovery and facilitate the characterization of the RNA polymerization mechanism.

Results and Discussion

Sequence analysis and phylogenic distribution

To compare *Flaviviridae* RdRPs, we have used the set of sequences defined in VaZyMolO [15], which include all sequences of completly sequenced viral genomes (Table1). The polymerase gene product alignment is based both on motif conservation and structural superimposition or conservation of secondary structures. We observe a great disparity depending on the genera of the compared sequences. Based on the alignment, a tree is derived (Figure 2 and additional file 1). Three major groups appear reflecting respective genus. Pestiviruses form a clear group distant from hepacivirus and flavivirus. This latter is the largest group of the family. It may be

divided in several groups and isolated viruses reflecting adaptation. GB viruses cluster with hepacivirus in one group. This phylogenic distribution suggests that, in terms of a most relevant model polymerase useful in the screening of anti-viral drugs, GBV-C is closer to HCV than BVDV. The PSI-BLAST [16] search against non-redundant data bases (nrdb) using the GBV-C polymerase as an input sequence converges after one iteration and retrieves the HCV polymerase only, with an E-value of 9×10^{-59}.

Homology modeling of the GBV-C Virus RNA Polymerase.

A sequence alignment of GBV-C and HCV polymerases is presented in Figure 3. It is based on sequence and structure comparison taking into account the prediction of secondary structure for GBV-C. In order to validate our method to predict the secondary structure, we have first submitted the HCV polymerase NS5B as a test sequence. Using the software Predict Protein, 50% of the β-sheets and 84% of the α-helix are correctly predicted in the HCV polymerase, and we obtain 87.5% of correctly predicted structural elements using PSI-PRED [17]. Such prediction results make us confident with respect to the reliability of the GBV-C prediction. The secondary structure elements of HCV polymerase and the structural prediction of the GBV-C polymerase are superimposed on the sequence alignment shown in Figure 3. The comparison between the secondary structure elements observed in the HCV crystal structure and the prediction made for GBV-C polymerase (Figure 3) shows that β-strands and α-helices are almost perfectly superimposed, albeit small gaps are located in few α-helices or loops. The alignment shows 32% identity and 72% similarity. Insertions and deletion localized primarily in loops. The amino acid conservation in the fingers and palm is close to 40% identity. Both motifs (F, A, B, C, E) and the residues involved in the I site (Arg 45, Lys 48, Lys 145, and Arg 151) are matching very well. As in the crystal structure of the HCV polymerase where the 55 C-terminal amino acids are deleted, we did not include the last 47 amino acids at the GBV-C polymerase C-terminus.

The sequence alignment and predicted secondary structure shown in Figure 3 were used in Swissmodeller (see methods) to build the GBV-C polymerase model (Figure 4). The modelled structure was then evaluated using by PROCHECK [18] and "What if" [19]. Results are shown in Table 2A/B. According to this last program, the Ramachandran plot is correct and the score is within expected ranges for well-refined

structures. Nevertheless, several residues located in flexible loops fall into disallowed regions of the Ramachandran plot (additional file 2A): Ser 100, Val 255, Thr 256 and Cys 215. The Ramachandran Plot statistics given by PROCHECK (additional file 2B) shows clearly that 99% of the residues are in allowed regions. The score corresponding to the chi-1/chi-2 angles of all residues is within expected ranges for well-refined structures (Table 2A/B). The model has a normal distribution of residue types over the inside and the outside of the protein. Again, the backbone conformation analysis gives a score that is normal for correctly refined protein structures. The RMS Z-score given in Table 2A is expected to be around 1.0 for a normally restrained data set, and this is indeed observed as in the case of high resolution X-ray structures. In the GBV-C polymerase model, bond angles and lengths can be considered to deviate normally from the mean standard bond angles.

As expected with such good scores, the model of the GBV-C polymerase is similar to that of HCV, and displays the essential features of the typical RNA dependent RNA polymerase fold (Figure 4A and 4B). However, we noticed two small differences between the HCV structure and the GBV-C model. First, we observe that Cys 283 and Cys 308 are spatially close enough to model a disulphide bridge (Figure 3 and additional file 3). This bond connects the fingers and the palm, and may stabilize the protein. Second, the superimposition of the GBV-C model and the HCV structure (Figure 4C) shows little but notable differences in the palm and thumb. The secondary structure elements are conserved in place and type, but they are shorter in the model than in the structure. However, these secondary structure elements should have similar functions. For example His 428 overlaps Tyr 448 of the HCV flap (Figure 4D) and replacement of the aromatic ring of the tyrosine by the histidine ring could play the same stacking role during initiation (see discussion below).

Surface and NTP-binding sites

We note several differences in the surface shape (Figure 5) of the GBV-C model. As the two backbones are superimposed theses differences are only due to the variability of side chains. The sequence conservation reported for the GBV-C model (additional file 3) shows that amino acids oriented toward the inner side of the protein are conserved whereas the amino acid which are pointing to the surface show low identity. This surface variability may be explained by the fact that the GBV-C

polymerase form a complex with other viral proteins, as it is the case for the HCV polymerase which interacts with NS3 or NS5A proteins, or as observed in the case of the poliovirus polymerase [20]. These other viral proteins may differ in their NS5B binding domain between HCV and GBV-C. Moreover, it has been shown that the HCV polymerase dimerizes and can form higher order structures after oligomerization. This multimerization is required for the HCV polymerase activity [21]. As the GBV-C polymerase is similar to HCV polymerase, the same oligomerization may also occur in the case of the GBV-C polymerase. Surface amino-acids have then to be specific to the virus to allow correct dimerization of the polymerase and/or interaction with the other components of the replicative complex. The electrostatic potential comparison is presented in Figure 5. It shows that the repartition of the charges on the surface of the model is globally equivalent to those located on the surface of HCV polymerase. We observe that the thumb in both cases is negatively charged (Figure 5B and E). The positive channel supposed to guide the RNA template to the catalytic site is very well conserved, and the flap is partially obstructing this cavity. The difference appears near the NTP tunnel (Figure 5C and F). In the HVC polymerase structure, the surface is clearly positively charged whereas in the GBV-C polymerase model the positive charge is less apparent.

In the HCV polymerase, the allosteric site forms a pocket where GTP binds. Such a pocket does exist in GBV-C despite sequence variability (Figure 3), and is located behind the thumb subdomain. The surface analysis shows that the pocket has an hydrophobic nature, except for the side chains of Asp 30 and Lys 473 that may however participate in the binding of a GTP molecule (see below).

In the HCV structure, several NTP molecules can bind to the catalytic site at P, C, and I sites. Indeed, up to 9 phosphate moieties can be seen in the crystal structure [7]. Only the nucleotide bound at the C site is well defined, although its nucleobase is probably incorrectly located in the absence of the RNA template. Clearly, a better definition of nucleotides and template is needed in order to understand the RNA synthesis process. On the other hand, the BVDV polymerase structure in complex with GTP in the catalytic cavity suggests a role for this nucleotide in the initiation of RNA synthesis, as proposed below.

Docking of GTP GBV-C

The analysis of the thumb in terms of structure and sequence comparison proved to be informative to propose an RNA synthesis initiation mechanism. Previously, in HCV polymerase the E motif has been proposed as part of the site which accommodates the first NTP incorporated during initiation of RNA synthesis (P site). Motif E is defined by the CS-18X-R signature (Figure 3 and additional file 1) [7]. In the case of BVDV, the polymerase structure has also been solved in complex with GTP [10]. This GTP is found in a binding pocket that is mainly constituted by amino acids within motif E. Their side chains in addition to an Arginine (Arg 529) further away in the sequence, are effectively stabilizing the phosphate chains of GTP. Indeed, the NS5B sequence comparison of *Flaviviridae* showed that motif E could be extended CS-18X-[RKT]-x(8)-[RK] signature (Figure 6). In the BVDV polymerase structure, the GTP molecule has been compared to a vestigial RNA molecule acting as a primer [10]. In the HCV polymerase structure, this GTP position corresponds to a cavity filled with water molecules. In the GBV-C model such a pocket exists, but its shape is different. Based on the GTP localization in BVDV structure we have docked a GTP molecule in GBV-C polymerase to see if this binding pocket could accommodate a GTP molecule in the same manner. Docking results show a perfect fitting of the molecule into the binding pocket (Figure 7). Because the pocket is somehow smaller than in the case of BVDV, the GTP ribose is flipped and the phosphate chain bends to follow the surface of the pocket. Thr 367 and Arg 371 of GBV-C E motif are involved in the stabilization of the phosphate chain. The Ser 349 of the CS motif forms the bottom of cavity. In both the structure and model, the cavity is obstructed by a proline (Pro 189 GBV-C; Pro 321 BVDV). Amino acids stabilizing the guanine base are different. While the base is stabilized only by hydrogen bonds to Tyr 187 in GBV-C, it is stabilized by Thr 320 and the Tyr 581 in the case of BVDV. In GBV-C the His 448 of the flap is forming the top of the cavity stabilizing the cycle of the base (Figure 7). In all cases the cavity is positively charged contributing to stabilize GTP in the pocket. Based on our model, we propose that motif E is the signature sequence of a GTP binding site in which GTP is required to hold the initiation complex tight. In our structural model, the GTP itself is too remote to act as a platform for the nucleotide positioned at the P site. The modelled GTP binding site together with the observed position of the flap lead us to suggest a mechanism for *de novo* initiation (Figure 8). We propose that once the first reaction of initiation is achieved (Figure 8C and D), the initiated template enters the

pocket where the motif E GTP is situated, and stacks against the guanine base (Figure 8E). This stacking induces a rearrangement of the base, which now contacts the flap. This latter interaction induces the opening of the flap leading to GTP release and further major structural changes within the polymerase (Figure 8F and G). The movement of the flap is supposed to occur to open the cavity allowing the elongation of the neo-synthesized RNA. The opening of the cavity implies that the thumb moves. It has been already observed that the fingers and the palm rotate as rigid body around the axis against the thumb domain [8]. In our model, the flap is spatially conserved suggesting that the same movement may occur during the elongation step of the GBV-C polymerization. Additionally, the position of the amino acid closing the cavity of the polymerase (flap in the case of GBV-C and HCV, or the β-sheet in the case of BVDV) suggests that the opening movement is specific for each virus. This movement would be best described as an opening from the top for HCV and GBV-C and, lateral for BVDV. Recently, we have characterized the initiation steps of RNA synthesis kinetically [22]. It is interesting to note that our present model is in agreement with the kinetic data showing that the N_2 to N_3 polymerization reaction is strongly rate limiting, and corresponds to the first partial opening of the flap to release GTP as proposed in Figure 8 panel F, whereas the other rate-limiting step from N_4 to N_6 corresponds to the other complete flap opening allowing dsRNA to exit from the active site as proposed in panel G.

Conclusions

The recently published high-resolution three-dimensional structure of BVDV and HCV polymerase has allowed the structural comparison of the two polymerases. Major differences in fingers and thumb suggest that molecular interactions during the initiation mechanism are different. BVDV has been used as model in studying hepaciviruses. However, phylogenic analysis shows that GBV-C is more closely related to HCV than BVDV. We propose here a reliable model of the GBV-C polymerase structure.

The model of the GBV-C polymerase is poorly defined in loopy regions where most of the gaps have been introduced. Despite this imprecision, the very good scores of the structural indicators make us very confident on the reliability of our model. Moreover, the model is consistent with the known three-dimensional structure of

RNA dependent RNA polymerases, and show conservation of all structural elements involved in polymerization (catalytic site, RNA positive channel, NTP tunnel). As expected after the alignment and prediction study, the GBV-C model is very close to the HCV structure, even with an allosteric GTP binding site being conserved. Based on the BVDV polymerase/GTP complex structure, we generated a model of a corresponding complex of GBV-C. We propose a role for the GTP molecule bound at a site implicated in the initiation of RNA synthesis associated with structural re-arrangements. Our study provides useful information of the location of residues involved in the polymerization process and hence should offer a useful tool for future biochemical analysis and drug discovery.

Methods

Sequence Retrieval

The sequences related to the different kind of polymerase were retrieved with a PSI-BLAST[16] with standard parameters from the public available protein database Swiss-Prot[23], Protein Data Bank (PDB)[24] and VaZyMolO[15] (http://www.vazymolo.org). For this study we have used different structures of HCV (PDB code : 1GX5, 1GX6), and BVDV(PDB code : 1S48,1S49) .

Sequence alignment comparison

Alignment of representative sequences from several members of *Flaviviridae* were performed using ClustalW [25] with the following parameter. Slow algorithm Identity matrix for pairwise alignment and Blosum series matrix for multiple alignments. Then the alignment is carefully analysed and optimised with Seaview[26] ; taking into account the secondary structure prediction and structural elements when it does exist.

The secondary structure predictions were carried out by JPRED^2 [27], PSI-PRED[17] and Predict Protein Server[28]. We used Predict Protein with a window of 150 amino acids in order to increase the sensibility of the prediction. 20 amino acids overlap with each common superimposed windows. The results presented are consensus. Sequence alignment with structural information (structure or predictions) and the comparison of the structure in 1D of the known viral polymerases had been performed using ESPript 2.0 [29] and ENDscript 1.0 [30].

To visualised conserved region in amino acids composition on the reference structure, we have used BOBSRCIPT[31]. The similarity scores are calculated from the Clustal W[25] alignment and they are shown on this structure by a white (low score) to red (identity) colour ramp .

Phylogenetic analysis

The sampling variance of the distance values was estimated from 1000 bootstrap resamplings of the alignment columns. The evolutionary inference was performed according to the Neighbor-joining method. And multiple runs were conducted with randomized sequence input order to avoid the tree being caught in a local statistical minimum. Tree is generated using Phylodendron (©1997 Gilbert)

Model building, refinement and evaluation

The resulting multiple sequence alignment with the consensus secondary structure prediction was used as template to generate the threading alignment. The derived pairwise alignment serve as reference for preparing the file for model. Swiss-PDB Viewer[32] is used to generate a first threading model. The three dimensional model of GBV-C RNA-polymerase was constructed using the crystal structure coordinates of HCV Polymerase [7, 8] (PDB code: 1GX5). Main gaps appear in loops and smaller one in some helix. This alignment and threading model serve as template file for Swiss Modeller. The non-modelled loops are manually build scanning loop data-base. The model is then minimized with a cut off of 10 A with 40 cycles of steepest descent until the gradient fell below 10Kcal/mol and 20 cycles of conjugate gradient. The computations were done in vacuo with using GROMOS 96 [33, 34] force field.

Surface comparison of the template and the model are performed with GRASP[35]. The model produced is checked using PROCHECK[18] and “WHAT IF”[19].

Docking GTP molecule in GBV-C

The 3D model of GBV-C RNA polymerase was used as a target for the docking of GTP. First, we superimpose the structure of BVDV RNA polymerase/GTP complex (PDB code 1S49) with our 3D model. This step was performed with the program Turbo-Frodo[36]. A docking study was performed to explore the presence or not of a GTP binding pocket like it was described in BVDV polymerase. For the docking procedure, the programme AutoDock 3.0.5 [37] was used with a grid spacing of 0.375 Å and 40 x 40 x 40 number of points. The grid was centered on the mass center of

GTP molecule. The GA-LS method was adopted using the default settings. Amber united atoms were assigned to the protein using the program AutoDock Tools. 250 possible binding conformations were generated. The results of AutoDock run were clustered using a RMSD tolerance of 1.0 Å. We consider the structure of the first cluster. To validate the use of the AutoDock program, the docking study was performed on the reference compounds BVDV polymerase and GTP. This program successfully reproduced the experimental binding conformation with acceptable root-mean-square deviation (RMSD) of atom coordinates. Finally, the interaction models of GTP with the binding pocket were produced using the LIGPLOT program[38].

Authors' contributions.

FF carried out the sequence retrieval, alignments, modellization, phylogenic studies, and, the structure and docking analysis. CB performed the docking and structure analysis. BC conceived of the study, and participates in its design and coordination. FF, CB, HD and BC all contributed to writing the final manuscript and interpretation of data.

Acknowledgements

The authors thanks Dr. Barbara Selisko, Dr. Sonia Longhi and Jean-Marie Bourhis for critical reading of the manuscript.

References

1. Ranjith-Kumar, C.T., et al., *De novo initiation pocket mutations have multiple effects on hepatitis C virus RNA-dependent RNA polymerase activities.* J Virol, 2004. **78**(22): p. 12207-17.
2. Lindenbach, B.D. and C.M. Rice, *Flaviridae : The viruses and their replication*, in *Fields virology*, D.M. Knipe and P.M. Howley, Editors. 2001. p. 991-1042.
3. Kao, C.C., P. Singh, and D.J. Ecker, *De novo initiation of viral RNA-dependent RNA synthesis.* Virology, 2001. **287**(2): p. 251-60.
4. Kao, C.C., A.M. Del Vecchio, and W. Zhong, *De novo initiation of RNA synthesis by a recombinant flaviviridae RNA-dependent RNA polymerase.* Virology, 1999. **253**(1): p. 1-7.
5. Kim, M.J., et al., *Template nucleotide moieties required for de novo initiation of RNA synthesis by a recombinant viral RNA-dependent RNA polymerase.* J Virol, 2000. **74**(22): p. 10312-22.

6. Luo, G., et al., *De novo initiation of RNA synthesis by the RNA-dependent RNA polymerase (NS5B) of hepatitis C virus.* J Virol, 2000. **74**(2): p. 851-63.
7. Bressanelli, S., et al., *Structural analysis of the hepatitis C virus RNA polymerase in complex with ribonucleotides.* J Virol, 2002. **76**(7): p. 3482-92.
8. Adachi, T., et al., *The essential role of C-terminal residues in regulating the activity of hepatitis C virus RNA-dependent RNA polymerase.* Biochim Biophys Acta, 2002. **1601**(1): p. 38-48.
9. Zhong, W., et al., *RNA-dependent RNA polymerase activity encoded by GB virus-B non- structural protein 5B.* J Viral Hepat, 2000. **7**(5): p. 335-42.
10. Choi, K.H., et al., *The structure of the RNA-dependent RNA polymerase from bovine viral diarrhea virus establishes the role of GTP in de novo initiation.* Proc Natl Acad Sci U S A, 2004. **101**(13): p. 4425-30.
11. Muerhoff, A.S., et al., *Genomic organization of GB viruses A and B: two new members of the Flaviviridae associated with GB agent hepatitis.* J Virol, 1995. **69**(9): p. 5621-30.
12. Simons, J.N., et al., *Isolation of novel virus-like sequences associated with human hepatitis.* Nat Med, 1995. **1**(6): p. 564-9.
13. Ranjith-Kumar, C.T., et al., *Requirements for de novo initiation of RNA synthesis by recombinant flaviviral RNA-dependent RNA polymerases.* J Virol, 2002. **76**(24): p. 12526-36.
14. Ranjith-Kumar, C.T., et al., *Enzymatic activities of the GB virus-B RNA-dependent RNA polymerase.* Virology, 2003. **312**(2): p. 270-80.
15. Ferron, F., Rancurel, C., Longhi, S., Cambillau, C., Henrissat, B. and Canard, B., *VaZyMolO: A tool to define modularity in viral proteins*. Journal of General Virology, 2004. in press.
16. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic Acids Res, 1997. **25**(17): p. 3389-402.
17. McGuffin, L.J., K. Bryson, and D.T. Jones, *The PSIPRED protein structure prediction server.* Bioinformatics, 2000. **16**(4): p. 404-5.
18. Laskowski R A, et al., *PROCHECK: A program to check the stereochemical quality of protein structures.* J. Appl. Cryst, 1993. **26**: p. 283-291.
19. Vriend, G., *WHAT IF: a molecular modeling and drug design program.* J Mol Graph, 1990. **8**(1): p. 52-6, 29.
20. Lyle, J.M., et al., *Visualization and functional analysis of RNA-dependent RNA polymerase lattices.* Science, 2002. **296**(5576): p. 2218-22.
21. Qin, W., et al., *Oligomeric interaction of hepatitis C virus NS5B is critical for catalytic activity of RNA-dependent RNA polymerase.* J Biol Chem, 2002. **277**(3): p. 2132-7.
22. Dutartre, H., et al., *A relaxed discrimination of 2'-O-methyl GTP relative to GTP between de novo and elongative RNA synthesis by the hepatitis C RNA-dependent RNA polymerase NS5B.* J Biol Chem, 2004.
23. Bairoch, A. and R. Apweiler, *The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000.* Nucleic Acids Res, 2000. **28**(1): p. 45-8.
24. Berman, H.M., et al., *The Protein Data Bank.* Nucleic Acids Res, 2000. **28**(1): p. 235-42.
25. Thompson, J.D., D.G. Higgins, and T.J. Gibson, *CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.* Nucleic Acids Res, 1994. **22**(22): p. 4673-80.

26. Galtier, N., M. Gouy, and C. Gautier, *SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny.* Comput Appl Biosci, 1996. **12**(6): p. 543-8.
27. Cuff, J.A. and G.J. Barton, *Application of multiple sequence alignment profiles to improve protein secondary structure prediction.* Proteins, 2000. **40**(3): p. 502-11.
28. Rost, B., *PHD: predicting one-dimensional protein structure by profile-based neural networks.* Methods Enzymol, 1996. **266**: p. 525-39.
29. Gouet, P., et al., *ESPript: analysis of multiple sequence alignments in PostScript.* Bioinformatics, 1999. **15**(4): p. 305-8.
30. Gouet, P. and E. Courcelle, *ENDscript: a workflow to display sequence and structure information.* Bioinformatics, 2002. **18**(5): p. 767-8.
31. Esnouf, R.M., *An extensively modified version of MolScript that includes greatly enhanced coloring capabilities.* J Mol Graph Model, 1997. **15**(2): p. 132-4, 112-3.
32. Guex, N. and M.C. Peitsch, *SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling.* Electrophoresis, 1997. **18**(15): p. 2714-23.
33. Gunsteren, W.F.v. and B. H.J.C, *Computer Simulation of Molecular Dynamics: Methodology, Applications and Perspectives in Chemistry*, in *Angew. Chem. Int.*, E. Engl., Editor. 1990. p. 992-1023.
34. Gunsteren, W.F.v., et al., *Computer simulation of protein motion.* Computer Phys. Communications, 1995. **91**: p. 305-319.
35. Sharp, K., Fine, R., and Honig, B, *Computer simulations of the diffusion of a substrate to an active site of an enzyme.* Science, 1987. **236**: p. 1460 -1463.
36. Roussel, A. and C. Cambillau, *TURBO-FRODO*, M.V. Silicon Graphics, CA, Editor. 1991, In Silicon Graphics Geometry Partners Directory,. p. 86.
37. Morris, G.M., et al., *"Automated Docking Using a Lamarckian Genetic Algorithm and and Empirical Binding Free Energy Function".* Journal of Computational Chemistry, 1998. **19**: p. 1639-1662.
38. Wallace, A.C., R.A. Laskowski, and J.M. Thornton, *LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions.* Protein Eng, 1995. **8**(2): p. 127-34.
39. Ago, H., et al., *Crystal structure of the RNA-dependent RNA polymerase of hepatitis C virus.* Structure Fold Des, 1999. **7**(11): p. 1417-26.

Figures

Figure 1- Ribbon representation of the RNA polymerase structure

HCV-C and BVDV are presented with the different subunits and domains. The thumb is colored in dark blue and yellow, fingers are colored in green and purple and the palm is colored in red. The image are generated using Pymol

Figure 2 - The pylogenetic tree of *Flaviviridae* RNA polymerase.

The numbers at the nodes indicates the statistical support of the branching order by bootstrap criterion. The bar at the bottom of the phylogram indicates the evolutionary distance, to which the branch lengths are scaled based on the estimated divergence.

Figure 3 - Alignment of the structural template (HCV) and the sequence of GBV-C

Sequence alignment of the HCV polymerase and GBV-C polymerase. Identical amino acids are boxed in red. We superimposed the secondary structure elements from HCV polymerase in pink, the predicted structural elements of GBV-C in blue and the secondary structure element of our final model in black. We present the HCV numbering according to [39] in pink. The numbering in dark red corresponds to the structure elements, which have been observed with a better resolution. The dots in the alignment and structural elements (predicted or average) symbolise gaps. Green letters shows the described universal motifs of RNA polymerase. The green arrows indicate the amino acids implied in the NTP binding. The green star marks the amino acid supposed to stack the priming base. The numbering is for GBV-C. The Numbers in dark green indicates in our model the two pair (1-1) of cysteines involved in disulfide bridge. Residues forming the allosteric GTP binding site are underlined in black

Figure 4

A: The model of GBV-C is presented with a front view putting in relief the Flap and the Histidine pointing to the catalytic site. α-helices are represented in Blue and β-sheets are in yellow. The image are generated using Pov-Ray.

B: a 180° rotation view of the GBV-C model show in A.

C: Superimposition of the X-ray structure of the HCV polymerase (in red) and the GBV-C polymerase model (in purple and yellow). A zoom view of the Flap region is presented in the upper side box in order to highlight the perfect superimposition of the cycle of the histidine found in the GBV-C polymerase and the tyrosine found in the HCV polymerase. The image are generated using Pov-Ray.

Figure 5 - Surface comparison between HCV and GBV-C

A, B, C correspond to different views of GBV-C polymerase surfaces calculated with Grasp. The surface is colored according to the electrostatic potential. The red correspond to negative charges, the white is neutral and the blue color correspond the positive charges. **D, E, F** correspond to the surface of HCV polymerase in the respective orientation. The color ramp is the same as for GBV-C polymerase surface.

Figure 6 - General alignment of the E motif in *Flaviviridae*

The conserved motif is labelled according to the nomenclature described for the RNA polymerase family. Invariant residues are highlighted in red, while conserved residues are in yellow box highlighted in bold. Consensus sequence with 70% similarity is shown down the alignment. The sequences are sorted by their genera.

Figure 7 - Docking GTP in the binding pocket of GBV-C

A: View of the BVDV's GTP pocket, where a GTP molecule is co-crystalized. The surface is coloured according to the electrostatic potential nomenclature.

B: Ligplot of the BVDV's GTP pocket presenting residues involved in the stabilization of the GTP. Hydrogen bonds are indicated in dotted lines and numbering refers to the distance between amino acids.

C: View of the proposed GTP pocket in GBV-C with a docked GTP molecule. The surface is coloured according to the electrostatic potential nomenclature. Hydrogen bonding are indicated in dotted lines and numbering refers to the distance between amino acids.

D: Ligplot of the GBV-C's GTP pocket presenting residues involved in the stabilization of the GTP. For all hydrogen bonds are indicated in dotted lines and numbering refers to the distance between amino acids.

The images were generated using Pymol.

Figure 8 - A model for *de novo* RNA synthesis at the hepacivirus NS5B active site

A: The polymerase is represented schematically to illustrate key points in the reaction mechanism.

B: The RNA matrix is represented in clear blue squares. NTP are in red squares, allosteric GTP is in dark blue square, and the bound GTP is in green blue square.

C: Binding of the first NTP in the active site.

D: The initiation reaction is presented with a yellow flash. Upon incorporation of the third NTP, the matrix and the neo synthesized RNA slide to the cavity pushing the GTP to the flap.

E: intermediate position where the flap, GTP and RNA matrix are stacked.

F: Opening of the flap and release of GTP.

G: The polymerase shift to the elongation mode; the thumb moves to fully open the cavity, and the elongation resumes.

Tables

Table 1 - Listing of *Flaviviridae*

Viruses used in the study with their VaZyMolO and NCBI accession numbers correspondance.

Table 2 - Quality of the model

A: Parameters reflecting the quality of the model checked by «What if» [19]. B: Quality of chain of the model. The model is verified at 2Å resolution. The parameters value in the table represents observed value for the GBV-C polymerase model compared with typical value obtained for well refined structures at the same resolution [18].

Additional files

Additional file 1 – Multiple alignment of the Palm subdomain between *Flaviviridae* RNA polymerase.

The conserved motifs are labelled according to the nomenclature described for the RNA polymerase family. Invariant residues are highlighted in red, while conserved

residues are in yellow box highlighted in bold. Consensus sequence with 70% similarity is shown down the alignment. The sequences are sorted by their genera.

Additional file 2 – Ramachandran plot of the GBV-C Model with Procheck Statistics

A: Ramachandran plot of GBV-C polymerase model. Favoured and allowed region are respectively in red and yellow. All residues are symbolized by black box (■) except Glycine (▲). Red box (■) highlight residues in not permitted region.

B: Ramachandran plot statistics.

Additional file 3 – Residues conservation plotted on the structure.

Bodscript figure of the calculated homology based on the superimposition of the structure of HCV polymerase on the GBV-C polymerase model. The similarity is shown on this structure by a white (low score) to red (identity) colour ramp. The green doted line indicates the position of the disulfide bridges.

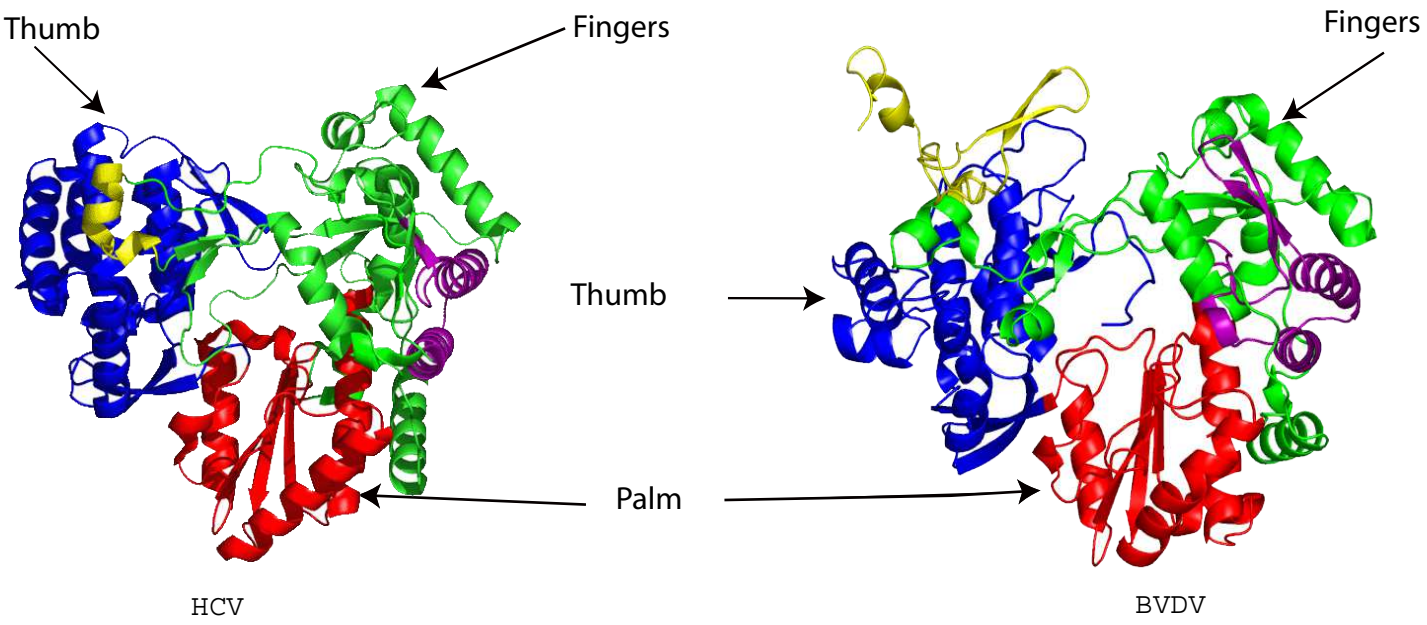


Figure 1

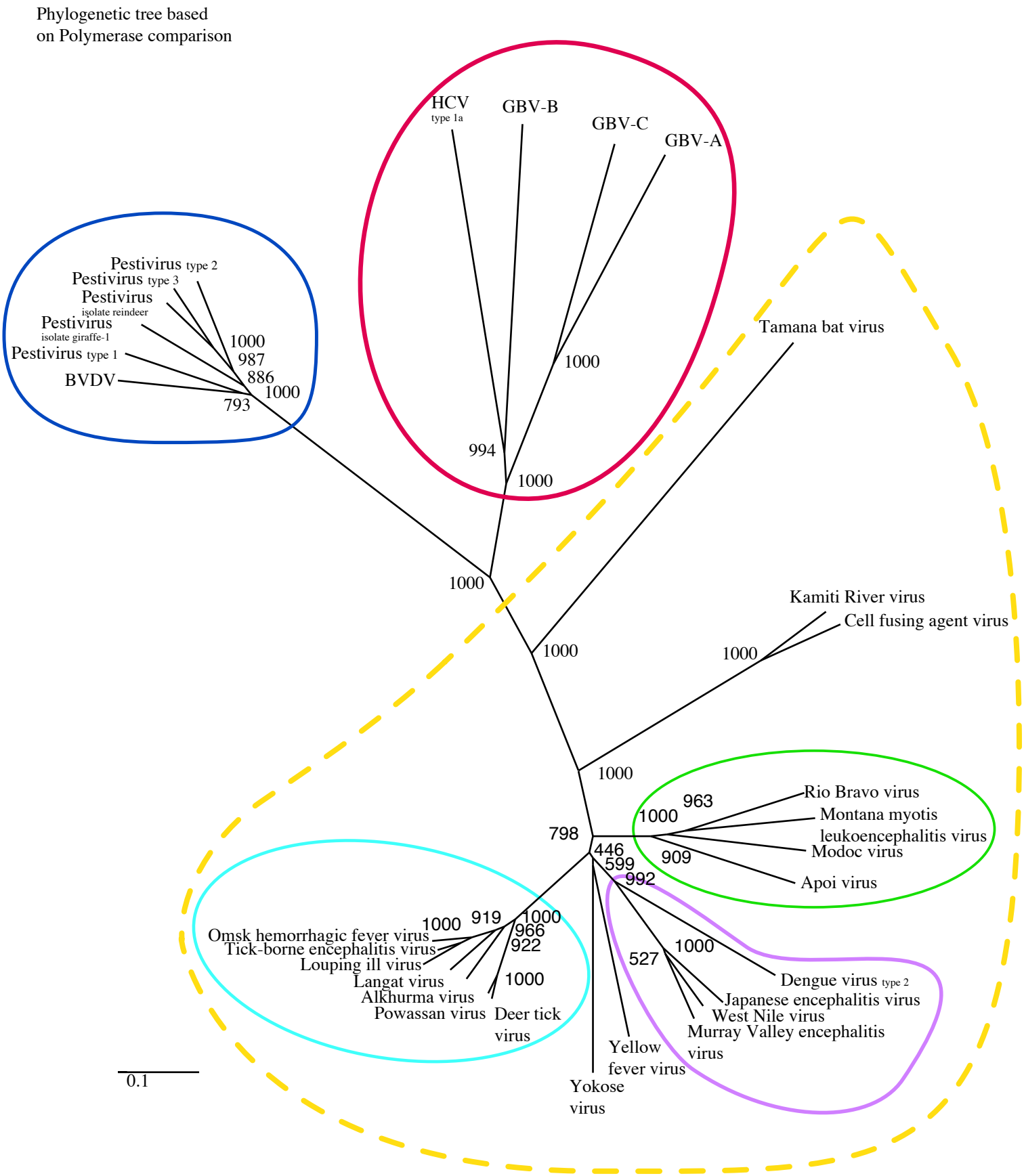

Phylogenetic tree based
on Polymerase comparison
HCV
type 1a
GBV-B
GBV-C
GBV-A
Pestivirus type 2
Pestivirus type 3
Pestivirus
isolate reindeer
Pestivirus
isolate giraffe-1
Pestivirus type 1
BVDV
1000
987
886
1000
793
1000
994
1000
1000
Tamana bat virus
Kamiti River virus
Cell fusing agent virus
1000
1000
1000
963
1000
Rio Bravo virus
Montana myotis
leukoencephalitis virus
Modoc virus
Apoi virus
909
798
446
599
992
1000
919
1000
966
922
1000
Omsk hemorrhagic fever virus
Tick-borne encephalitis virus
Louping ill virus
Langat virus
Alkhurma virus
Powassan virus
Deer tick
virus
527
1000
Dengue virus type 2
Japanese encephalitis virus
West Nile virus
Murray Valley encephalitis
virus
Yellow
fever virus
Yokose
virus
0.1

Figure 2

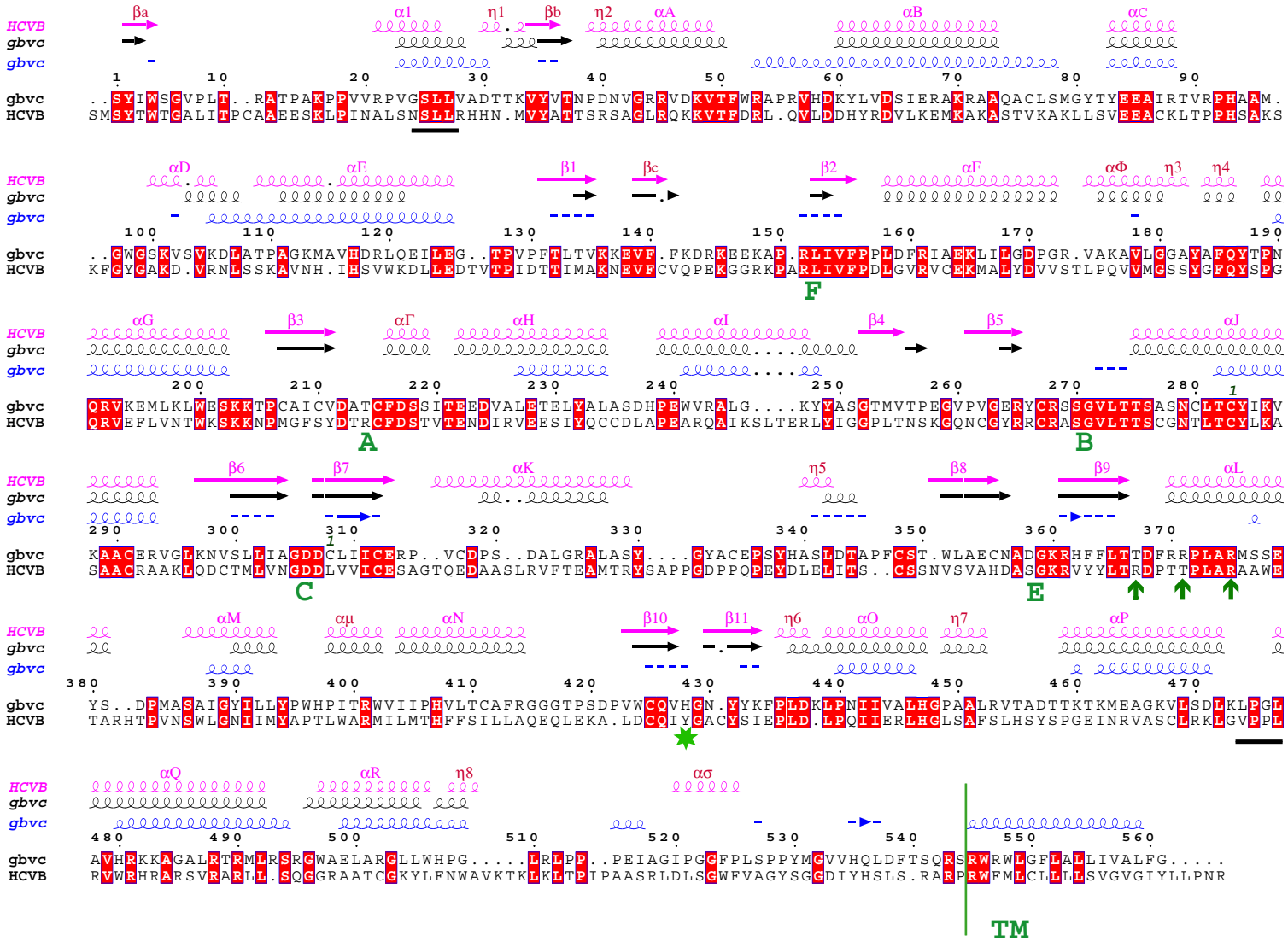


Figure 3

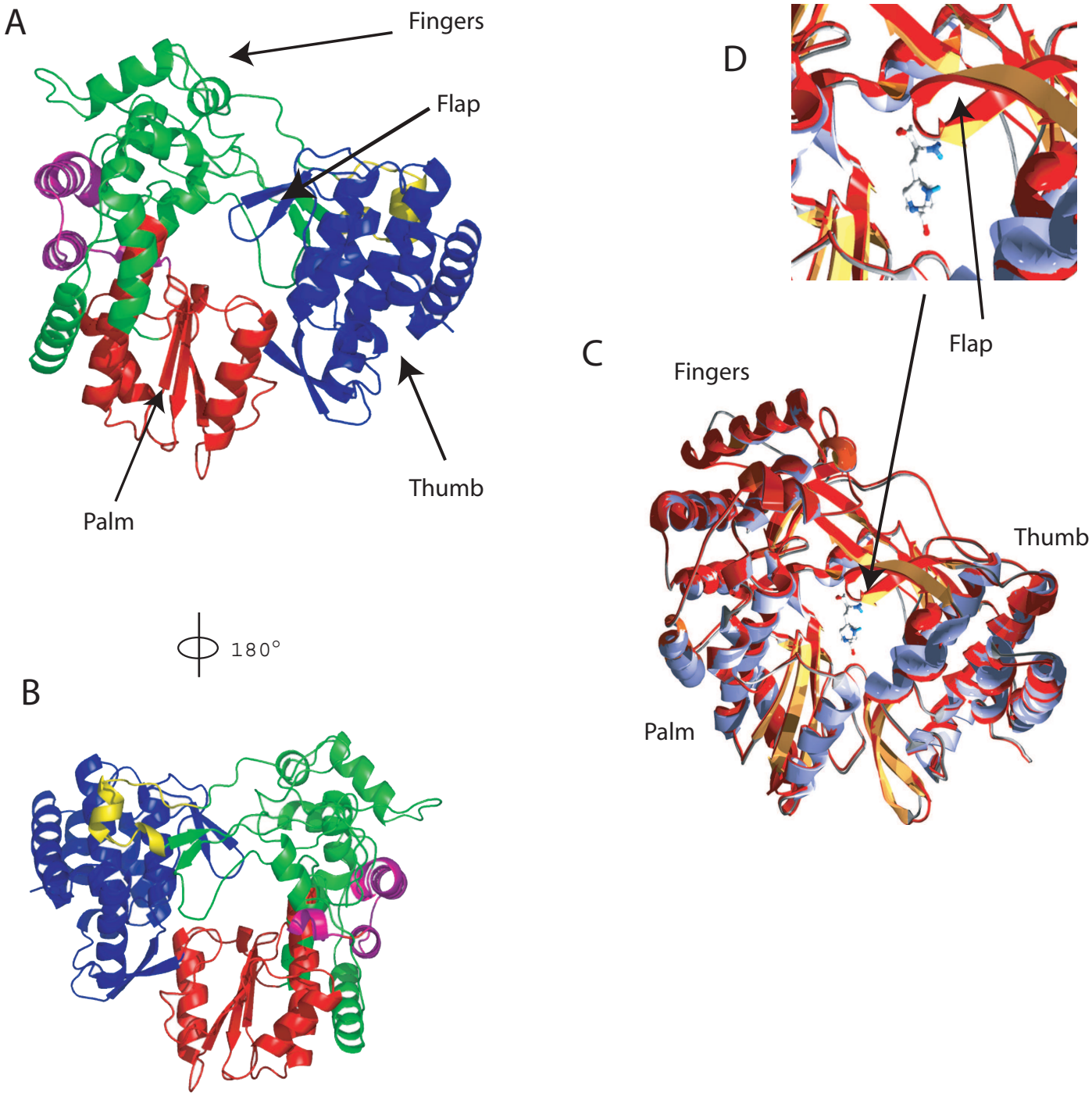


Figure 4

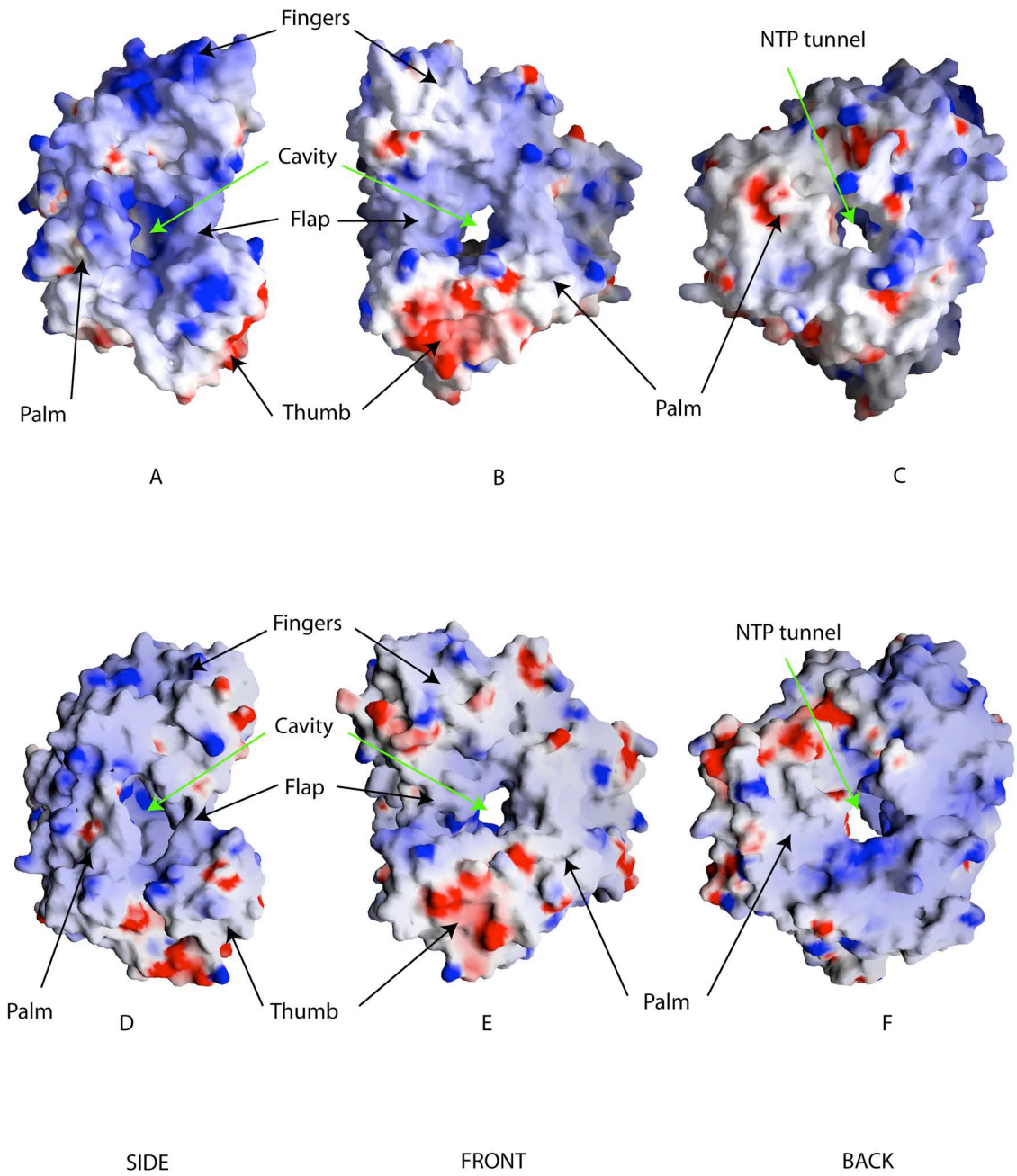

Fingers
NTP tunnel
Cavity
Flap
Palm
Thumb
Palm
A
B
C
Fingers
NTP tunnel
Cavity
Flap
Palm
D
Thumb
E
Palm
F
SIDE
FRONT
BACK

Figure 5

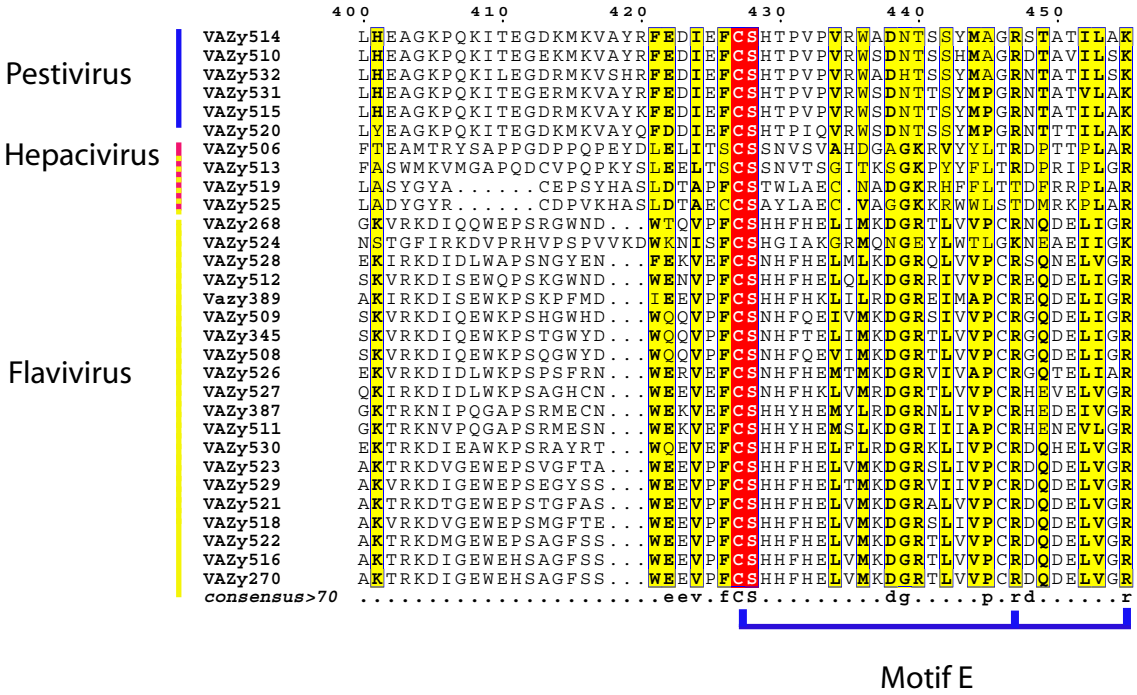
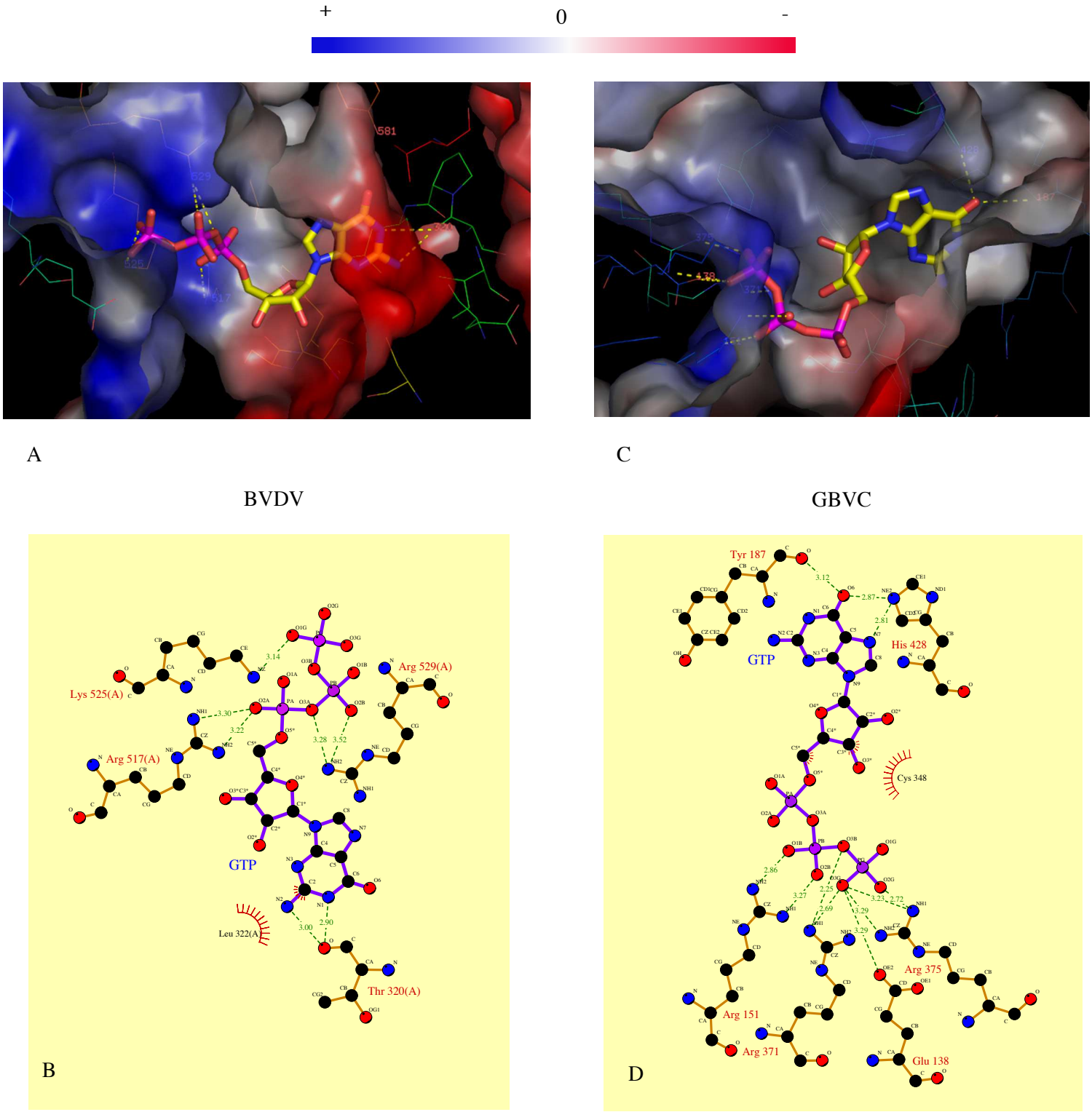

400 410 420 430 440 450
Pestivirus
Hepacivirus
Flavivirus
VAZy514 LHEAGKPQKITEGDKMKVAYRFEDIEFCSHTPVPVRWADNTSSYMAGRSTATILAK
VAZy510 LHEAGKPQKITEGEKMKVAYRFEDIEFCSHTPVPVRWSDNTSSHMAGRDTAVILSK
VAZy532 LHEAGKPQKILEGDRMKVSHRFEDIEFCSHTPVPVRWADHTSSYMAGRNTATILSK
VAZy531 LHEAGKPQKITEGERMKVAYRFEDIEFCSHTPVPVRWSDNTTSYMPGRNTATVLAK
VAZy515 LHEAGKPQKITEGDRMKVAYKFEDIEFCSHTPVPVRWSDNTTSYMPGRNTATILAK
VAZy520 LYEAGKPQKITEGDKMKVAYQFDDIEFCSHTPIQVRWSDNTSSYMPGRNTTTILAK
VAZy506 FTEAMTRYSAPPGDPPQPEYDLELITSCSSNVSVAHDGAGKRVYYLTRDPTTPLAR
VAZy513 FASWMKVMGAPQDCVPQPKYSLEELTSCSSNVTSGITKSGKPYYFLTRDPRIPLGR
VAZy519 LASYGYA.....CEPSYHASLDTAPFCSTWLAECC.NADGKRHFFLTTDFRRPLAR
VAZy525 LADYGYR.....CDPVKHASLDTAECCSAYLAECC.VAGGKKRWWLSTDMRKPLAR
VAZy268 GKVRKDIQQWEPSRGWND...WTQVPFCSHHFHELIMKDGRTLVVPCRNQDELIGR
VAZy524 NSTGFIRKDVPRHVPSPVVKDWKNISFCSHGIAKGRMQNGEYLWTLGKNEAEIIGK
VAZy528 EKIRKDIDLWAPSNGYEN...FEKVEFCSNHFHELMLKDGRQLVVPCRSQNELVGR
VAZy512 SKVRKDISEWQPSKGWND...WENVPFCSHHFHELQLKDGRRIVVPCREQDELIGR
Vazy389 AKIRKDISEWKPSKPFMD...IEEVPFCSHHFHKLILRDGREIMAPCREQDELIGR
VAZy509 SKVRKDIQEWKPSHGWHD...WQQVPFCSNHFQEIVMKDGRSIVVPCRGQDELIGR
VAZy345 SKVRKDIQEWKPSTGWYD...WQQVPFCSNHFTELIMKDGRTLVVPCRGQDELIGR
VAZy508 SKVRKDIQEWKPSQGWYD...WQQVPFCSNHFQEVIMKDGRTLVVPCRGQDELIGR
VAZy526 EKVRKDIDLWKPSPSFRN...WERVEFCSNHFHEMTMKDGRVIVAPCRGQTELIAR
VAZy527 QKIRKDIDLWKPSAGHCN...WEEVEFCSNHFHKLVMRDGRTLVVPCRHEVELVGR
VAZy387 GKTRKNIPQGAPSRMECN...WEKVEFCSHHYHEMYLRDGRNLIVPCRHEDEIVGR
VAZy511 GKTRKNVPQGAPSRMESN...WEKVEFCSHHYHEMSLKDGRIIIAPCRHENEVLGR
VAZy530 EKTRKDIEAWKPSRAYRT...WQEVEFCSHHFHELFLRDGRKLIVPCRDQHELVGR
VAZy523 AKTRKDVGEWEPSVGFTA...WEEVPFCSHHFHELVMKDGRSLIVPCRDQDELVGR
VAZy529 AKVRKDIGEWEPSEGYSS...WEEVPFCSHHFHELTMKDGRVIIVPCRDQDELVGR
VAZy521 AKTRKDTGEWEPSTGFAS...WEEVPFCSHHFHELVMKDGRALVVPCRDQDELVGR
VAZy518 AKVRKDVGEWEPSMGFTE...WEEVPFCSHHFHELVMKDGRSLIVPCRDQDELVGR
VAZy522 AKTRKDMGEWEPSAGFSS...WEEVPFCSHHFHELVMKDGRTLVVPCRDQDELVGR
VAZy516 AKTRKDIGEWEHSAGFSS...WEEVPFCSHHFHELVMKDGRTLVVPCRDQDELVGR
VAZy270 AKTRKDIGEWEHSAGFSS...WEEVPFCSHHFHELVMKDGRTLVVPCRDQDELVGR
consensus>70 .....eev.fCS.....dg.....p.rd.....r
Motif E

Figure 6

Figure 7


+
0
-
A
C
BVDV
GBVC
Lys 525(A)
Arg 517(A)
Arg 529(A)
GTP
Leu 322(A)
Thr 320(A)
B
Tyr 187
His 428
Cys 348
Arg 375
Arg 151
Arg 371
Glu 138
D
Key
Ligand bond
Non-ligand bond
Hydrogen bond and its length
His 53
Non-ligand residues involved in hydrophobic contact(s)
Corresponding atoms involved in hydrophobic contact(s)

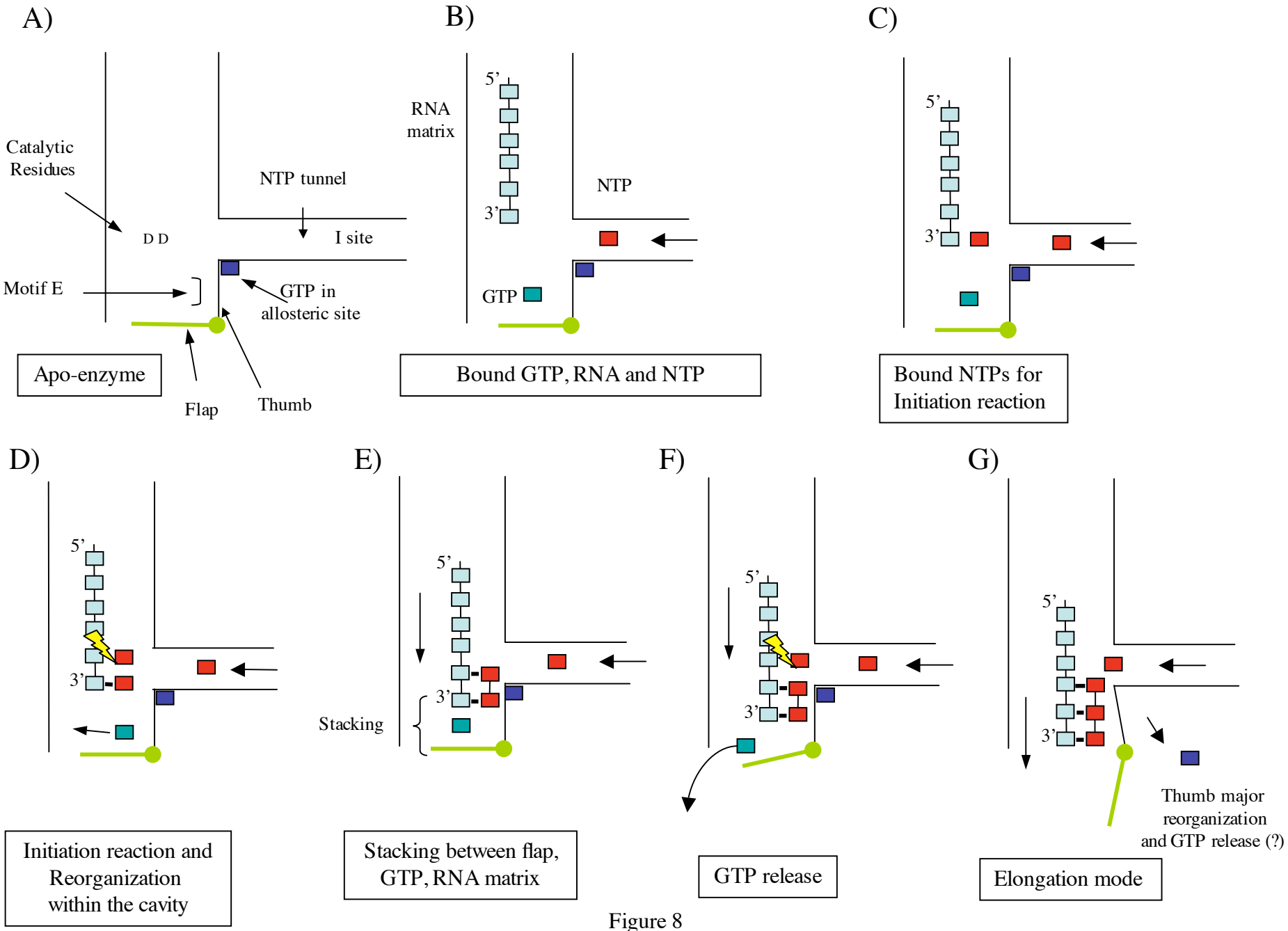

A)
Catalytic Residues
NTP tunnel
D D
I site
Motif E
GTP in allosteric site
Flap
Thumb
Apo-enzyme
B)
RNA matrix
5'
3'
NTP
GTP
Bound GTP, RNA and NTP
C)
5'
3'
Bound NTPs for Initiation reaction
D)
5'
3'
Initiation reaction and Reorganization within the cavity
E)
5'
3'
Stacking
Stacking between flap, GTP, RNA matrix
F)
5'
3'
GTP release
G)
5'
3'
Thumb major reorganization and GTP release (?)
Elongation mode

Figure 8

Listing of Flaviridae RNA polymerase

Flaviviridae			
DataBase Accession	Virus	NCBI Acc Protein	Genus
VaZy 268	Dengue virus type 2	NP_056776.1	
VaZy 270	Omsk hemorrhagic fever virus [Bogoluvovska]	NP_878909.1	
VaZy 345	West Nile virus	NP_041724.2	
VaZy 387	Kamiti River virus - isolate SR-82	NP_891560.1	
VaZy 389	Yokose virus [Oita 36]	NP_872627.1	
VaZy 506	Hepatitis C virus type 1a - isolate H77	NP_671491.1	
VaZy 508	Murray Valley encephalitis virus	NP_051124.1	
VaZy 509	Japanese encephalitis virus	NP_059434.1	
VaZy 510	Pestivirus type 1 [NADL]	NP_040937.1	
VaZy 511	Cell fusing agent virus	NP_041725.1	
VaZy 512	Yellow fever virus [Flavivirus (mosquito-borne)]	NP_041726.1	
VaZy 513	Hepatitis GB virus B	NP_056931.1	
VaZy 514	Bovine viral diarrhea virus genotype 2 [C413]	NP_044731.1	
VaZy 515	Pestivirus type 3 [X818; Clover Lane]	NP_620062.1	
VaZy 516	Tick-borne encephalitis virus	NP_043135.1	
VaZy 518	Powassan virus [LB]	NP_620099.1	
VaZy 519	Hepatitis GB virus C	NP_043570.1	
VaZy 520	Pestivirus type 2 [Eystrup]	NP_075354.1	
VaZy 521	Langat virus [TP21]	NP_620108.1	
VaZy 522	Louping ill virus [369/T2]	NP_044677.1	
VaZy 523	Deer tick virus [ctb30] - isolate CT95	NP_476520.1	
VaZy 524	Tamana bat virus	NP_658908.1	
VaZy 525	Hepatitis GB virus A	NP_045010.1	
VaZy 526	Modoc virus [M544]	NP_619758.1	
VaZy 527	Montana myotis leukoencephalitis virus	NP_689391.1	
VaZy 528	Rio Bravo virus [RiMAR]	NP_620044.1	
VaZy 529	Alkhurma virus [1176]	NP_722551.1	
VaZy 530	Apoi virus [ApMAR]	NP_620045.1	
VaZy 531	Pestivirus - isolate reindeer-1 V60-Krefeld	NP_620051.1	
VaZy 532	Pestivirus - isolate giraffe-1 H138	NP_620053.1	
	GB group		
	Pestivirus		
	Hepacivirus		
	Flavivirus		
		sub group	

Table 1

General parameters checked by «What if»

Structure Z-scores :	
1st generation packing quality	-1.577
2nd generation packing quality	-2.94
Ramachandran plot appearance	-0.74
chi-1/chi-2 rotamer normality	-0.224
Backbone conformation	-0.904

RMS Z-scores, should be close to 1.0:	
Bond lengths	0.950
Bond angles	1.426
Omega angle restraints	-0.923
Inside/Outside distribution	1.096

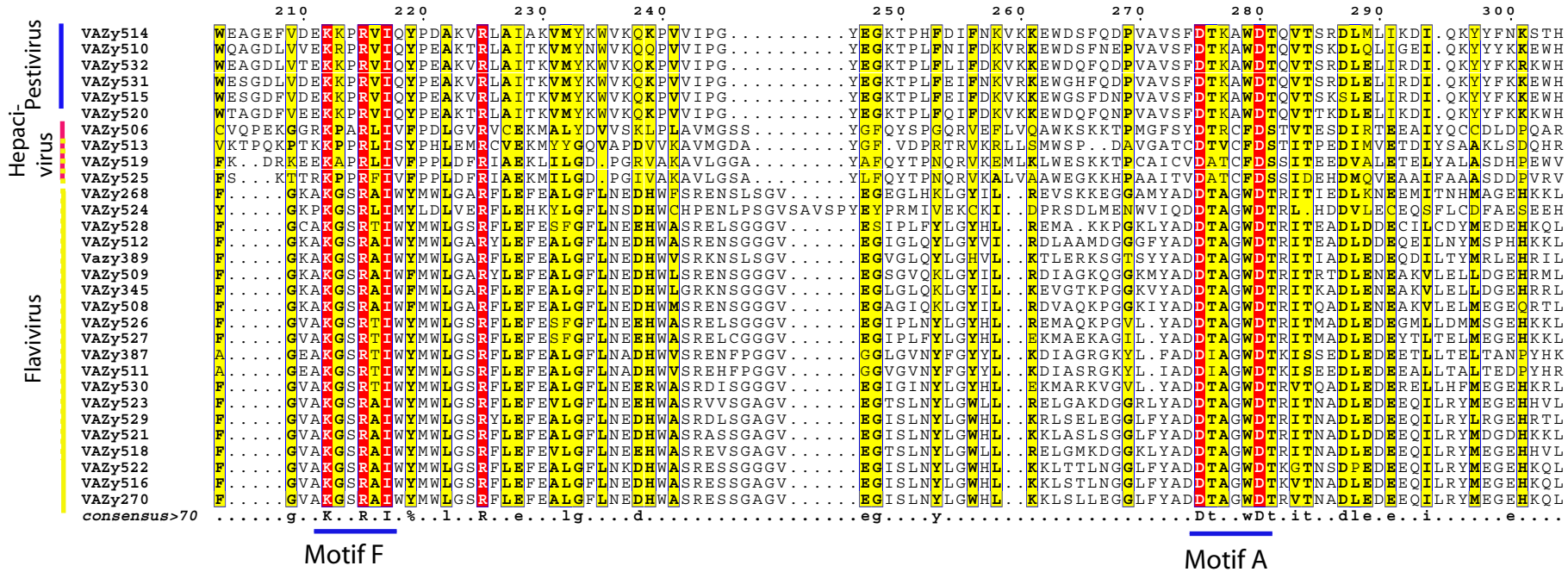
A

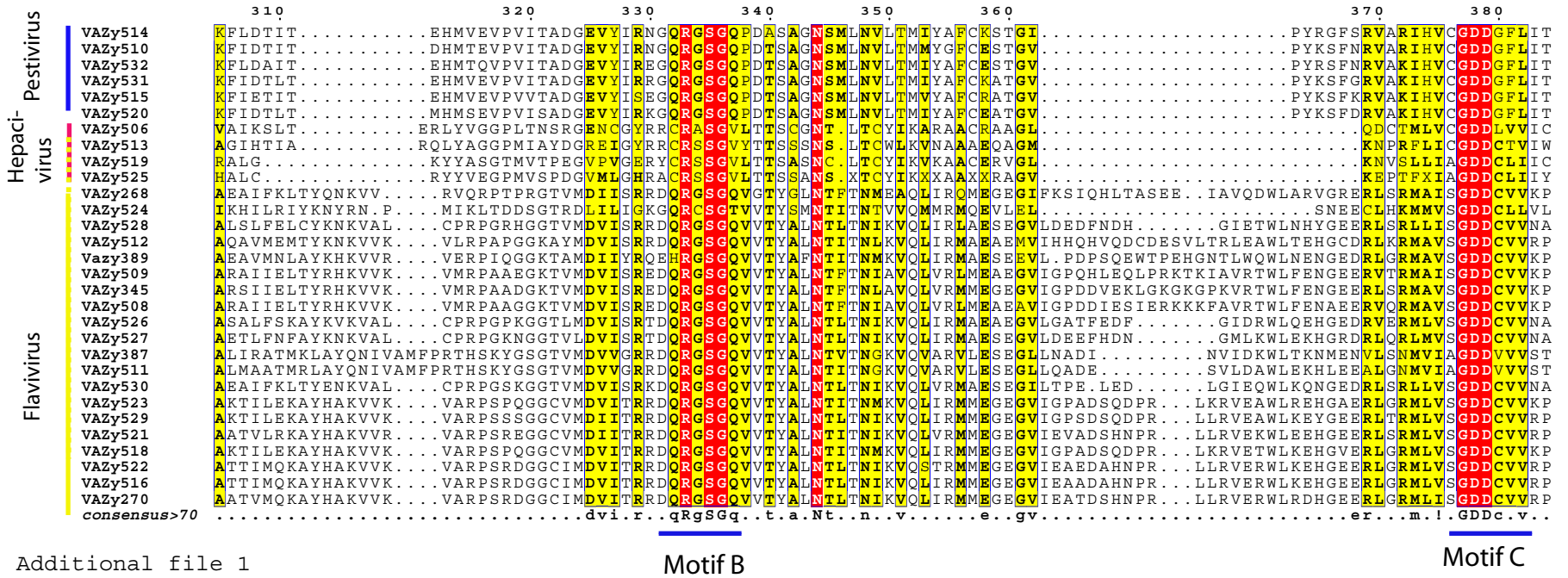
Quality of main-chain and side chain parameters of modeled RNA polymerase of GBV-c

Stereochemical parameter	N° of data points	Parameter value	Typical value	Band width	N° of bandwidths from mean
Stereochemistry of main-chain					
Percentage residues in A, B, L	438	89	83.8	10	0.5
Omega angle S.D	507	6.8	6	3	0.3
Bad contacts 100 residues	3	0.6	4.2	10	-0.4
Zeta angle S.D.	476	2.8	3.1	1.6	-0.2
Hydrogen bond energy S.D.	305	0.7	0.8	0.2	-0.4
Stereochemistry of side-chain					
Chi-1 gauche minus S.D.	79	15.7	18.1	6.5	-0.4
Chi-1 trans S.D.	103	13.2	19	5.3	-1.1
Chi-1 gauche plus S.D.	205	11.4	17.5	4.9	-1.2
Chi-1 pooled S.D.	387	13	18.2	4.8	-1.1
Chi-2 trans S.D.	114	15.7	20.4	5	-0.9

B

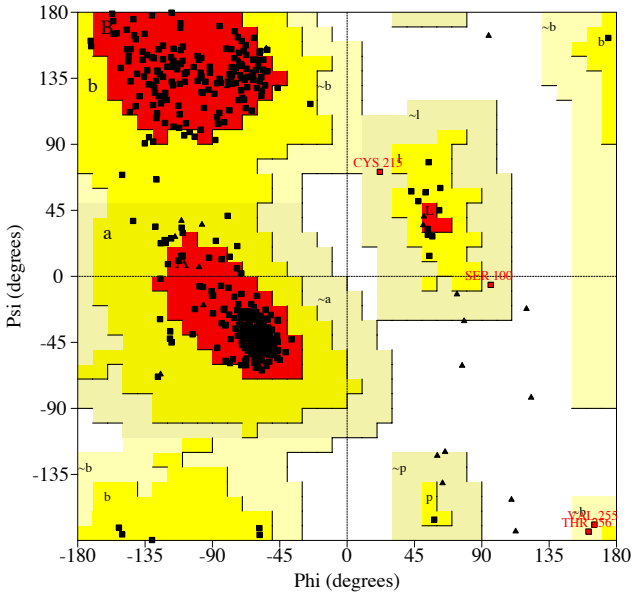
Table 2.





Additional file 1

Ramachandran Plot of the GBV-C polymerase model



A

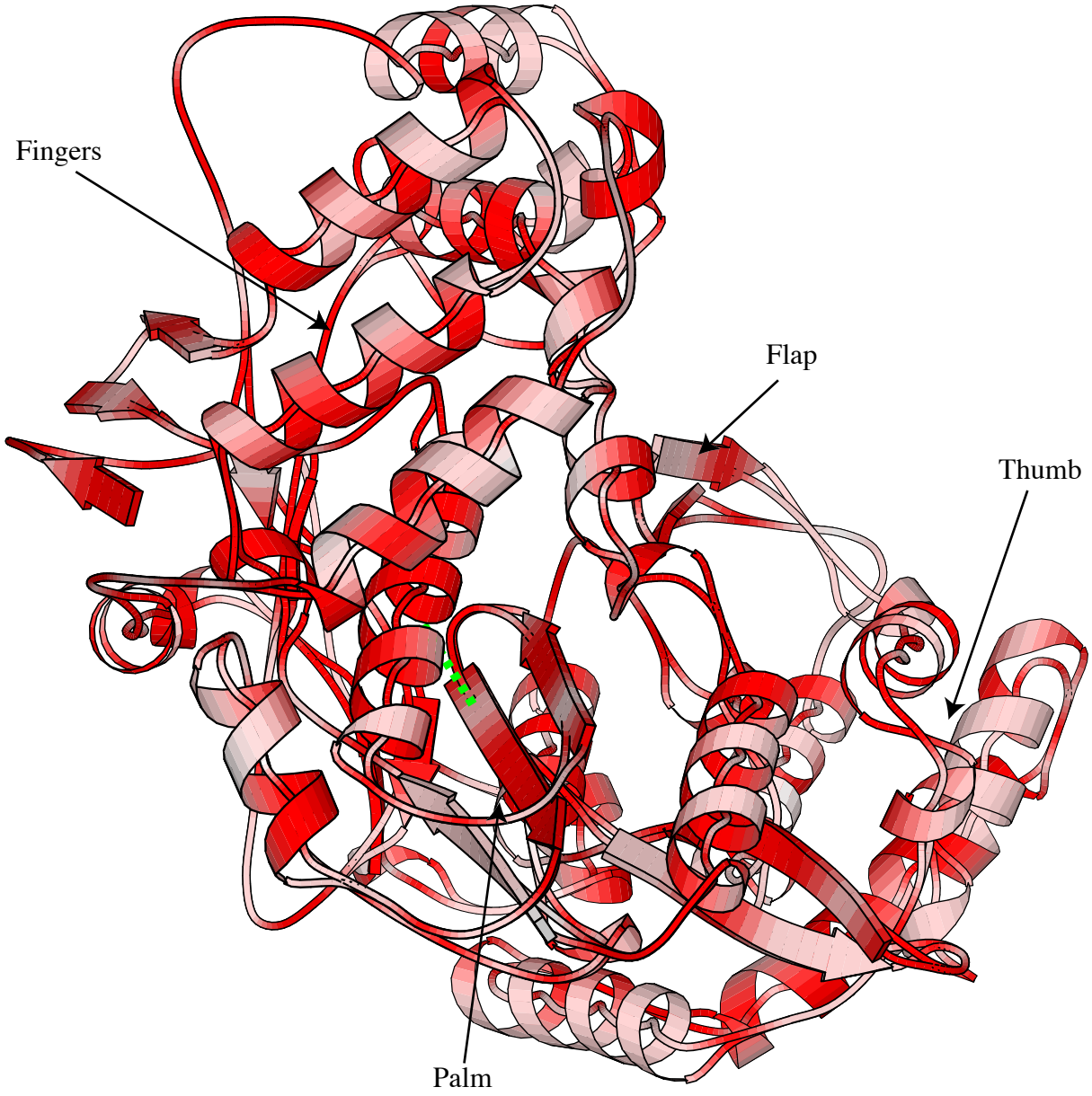
Plot statistics

Residues in most favoured regions [A,B,L]	390	89.0%
Residues in additional allowed regions [a,b,l,p]	44	10.0%
Residues in generously allowed regions [~a,~b,~l,~p]	4	0.9%
Residues in disallowed regions	0	0.0%
	---	------
Number of non-glycine and non-proline residues	438	100.0%
Number of end-residues (excl. Gly and Pro)	2	
Number of glycine residues (shown as triangles)	34	
Number of proline residues	36	

Total number of residues	510	

B

Additional file 2



Additional file 3

-Commentaires

Dans cet article, nous nous sommes attaché à la compréhension du mécanisme d'initiation de la synthèse d'ARN par la polymérase de GBV-C.

L'étude que nous avons menée nous a conduits à analyser de nouveau les séquences des polymérases des *Flaviviridae* à la lumière des dernières données structurales et fonctionnelles disponibles. Nous avons ainsi pu mettre en évidence la proche parenté entre les polymérases des GBV et des Hepacivirus. Ces résultats sont novateurs dans la mesure où les dernières études bioinformatiques concernant les GBV concernaient les hélicases (Simons et al. 1995a) et la dernière étude de séquences des polymérases de *Flavidiridae* est antérieure à la découverte des virus GB (Koonin 1993). La résolution de la structure des polymérases de HCV et de BVDV a permis de constater des différences structurales notamment au niveau du pouce.

Il semblerait que le mécanisme d'initiation soit commun à toutes les polymérases de la famille, ce qui implique la conservation d'un certain nombre d'étapes consensuelles (Kao et al. 2001; Choi et al. 2004). L'analyse de la structure des polymérases de deux virus appartenant à des genres différents, couplée aux données enzymatiques, nous a permis de proposer un mécanisme d'initiation commun aux Pestivirus, Hepacivirus, et GB virus.

-Compatibilité du mécanisme proposé à partir du modèle avec la structure de la polymérase de HCV

Dans l'article, nous décrivons le modèle de la polymérase du virus de GBV-C obtenu à partir de la structure de la polymérase de HCV. Les deux chaînes principales se superposent et les différences majeures observées au niveau de la surface et des cavités concernent le positionnement des chaînes latérales. Nous avons montré que la poche qui stabilise le GTP prés du site catalytique dans la structure de BVDV est conservée chez GBV-C et permet d'accommoder une molécule de GTP. Chez la polymérase de HCV, cette poche est caractérisée par une densité élevée de molécules d'eau (Bressanelli et al. 2002). Ce résultat découle de l'observation de la présence de densité électronique résiduelle dans la poche. Si cette densité résiduelle est mal définie, elle pourrait néanmoins correspondre à une molécule de GTP. Or, la poche est essentiellement composée de résidus amphiphiles qui, s'ils peuvent accommoder des molécules d'eau, pourraient également stabiliser une molécule de GTP. De plus, dans la structure de la polymérase du bactériophage Phi6 (Butcher et al. 2001), à la place de la cavité on retrouve une tyrosine qui stabilise le premier nucléotide à incorporer

(Bressanelli et al. 2002). La présence d'une molécule de GTP dans la poche de la polymérase de GBV-C pourrait jouer un rôle similaire et intervenir donc dans la stabilisation du premier nucléotide. De plus, cette hypothèse est particulièrement séduisante à la lumière des récents résultats du Dr. Hélène Dutartre, qui montrent que dans le cas de la polymérase de HCV, la transition de N2 à N3 est limitante et de fait compatible avec le modèle que nous proposons. Pour valider cette hypothèse, nous envisageons de dessiner et produire des mutants ponctuels des résidus impliqués dans l'accommodation du GTP. Ces substitutions sont censées empêcher la stabilisation du GTP et bloquer - ou réduire fortement - l'initiation, mais pas l'élongation de la synthèse d 'ARN.

Conclusion générale

Au cours de ma thèse, j'ai été amené à étudier de nombreuses protéines impliquées dans la réplication virale de différents virus. J'ai opté pour une approche bioinformatique capable de gérer, d'organiser et hiérarchiser l'ensemble de ces informations. La phase préparatoire de ce travail a consisté à dresser un bilan de l'ensemble des ressources déjà disponibles. Ce travail était nécessaire afin de pouvoir optimiser mon travail de recherche tout en évitant la redondance. Concernant la virologie, le bilan de ces recherches a mis en évidence une absence quasi complète de données spécifiques. J'ai pu ainsi évaluer les avantages et les inconvénients d'utiliser telle ou telle banques de données, chacune présentant des avantages et des inconvénients en termes de fiabilité, annotations, disponibilité et mises à jour. J'ai pu évaluer par rapport au besoin du projet, les limites des bases de données thématiques (séquences annotées, domaines, motifs) qui souffrent de ne pouvoir couvrir complètement un immense champ de connaissances. Leur conception remontent presque pour la plupart à une époque où le nombre de séquences était gérable (bien que conséquent), et provenait majoritairement de quelques organismes. Aujourd'hui la maintenance de ces bases est un véritable casse-tête, pour lequel se pose le dilemme d'entrer de nouvelles séquences ou de mettre à jour celles qui ont déjà été annotées. Les bases de données spécialisées structurales (PDB, CATH) ou fonctionnelles (MEROPS) sont pour la plupart factuelles (une structure ou une fonction décrite) et nous les avons considérées comme étant une source d'information complémentaire.
Le second volet de cette phase préparatoire a été de se familiariser avec les outils d'annotation afin d'en comprendre les avantages et les limites. La compréhension de leur mode de fonctionnement m'a permis de mettre au point ma méthode d'analyse tout en la confrontant en permanence aux nouveaux programmes ou méthodes décrites dans la littérature. La pratique de la veille technologique m'a permis d'utiliser ces nouvelles approches permettant ainsi l'optimisation des résultats ou un gain de temps. L'émergence de nouveaux concepts tels que le « désordre intrinsèque » des protéines m'a permis en particulier d'affiner ma méthode de travail et de détourner avec profit la méthode HCA de son utilisation première.
Fort de ces expériences, je me suis attelé à la conception de la base de données VaZyMolO. Partant d'une base de données interne au laboratoire, il fallut adapter cette dernière à nos besoins spécifiques (taxonomie, motifs, informations complémentaires, outils d'analyses, interface multi-utilisateurs, gestion des polyprotéines). Le travail préparatoire m'a permis aussi de lister une série d'écueils à éviter comme, par exemple une trop grande redondance ou une annotation automatique et non suivie.

Dans le cadre de la base de données VaZyMolO, j'ai privilégié la qualité (séquence complète et sans erreurs) des séquences, dont l'annotation a été faite et vérifiée manuellement. En général, une seule séquence par genre suffit pour annoter n'importe quelle nouvelle séquence.
La base de données est aussi un outil d'annotations qui s'est avéré un élément fondamental pour l'étude à grande échelle de séquences de protéines virales. Les premiers virus qui ont été étudiés dans VaZyMolO sont les virus à ARN négatif. Les premières analyses menées ont bénéficié du travail de caractérisation biochimique qui était mené en parallèle au sein du laboratoire. Ces informations ont permis d'ajuster nos prédictions et nous ont permis de délimiter nos premiers modules avec une certaine confiance. J'ai pu mettre en évidence un domaine méthyltransférase sur la protéine L des Mononegavirales. Des études ont depuis prouvé cette activité, et les essais de cristallisation sont prometteurs. Les modules délimités sur la phosphoprotéine et les nucléoprotéines ont fait par la suite l'objet d'études structurales qui pour certaines d'entre elles ont validé notre travail.
Le projet VaZyMolO a aussi montré son potentiel lors de l'analyse du génome du virus du SRAS. J'ai effectué l'annotation du génome de ce nouveau Coronavirus, et cela nous a permis d'obtenir une avance stratégique pour résoudre la structure cristallographique de nsp9. Au cours de cette étude, j'ai été amené à m'intéresser à l'évolution des repliements « OB », et à proposer une nouvelle famille appartenant à ce repliement.

de l'utilisation d'un outil tel que VaZyMolO pour faire une étude comparative de différentes protéines. Par son truchement, il a été possible de générer ces résultats très rapidement. J'ai pu me familiariser avec les outils de modélisation. Cette étude inclut l'évolution et l'analyse structurale des polymérases des *Flaviviridae* et nous a permis de proposer un mécanisme d'initiation de la synthèse d'ARN.

VaZyMolO vers d'autres développements ?

VaZyMolO par sa taille et sa spécialisation reste une base de données à taille humaine gérable par un petit nombre d'opérateurs. Le noyau développé mériterait d'être encore amélioré, il reste, notamment de nombreuses phases non automatisées dans l'implémentation des informations qui devront êtres développées. Ces développements devront avoir lieu pour une raison pratique : le nombre de séquences augmentant, le temps passé à rentrer manuellement l'information augmente de façon exponentielle. L'échantillon de séquences qui a été annoté est une base solide pour des extensions d'annotation. Le choix de se concentrer uniquement sur des génomes complets est, et doit rester la règle, pour éviter les écueils dont nous avons

déjà parlé. Il est vrai, et c'est la contrepartie de ce choix, que nous ne couvrons pas la totalité de l'espace viral. D'une part parce que nous n'avons pas fini d'intégrer la totalité de nos génomes, et d'autre part parce que tous les génomes ne sont pas accessibles. Néanmoins, et cela a déjà été fait par le passé, rien n'empêche de faire des annotations indépendantes, selon nos critères de qualité, qui pourront être intégrés par la suite dans la base de données le jour où il sera opportun de rentrer l'information.

Le nombre de séquences virales est finalement négligeable au regard de celui des autres séquences. L'implication des virologistes dans le domaine émergent de la bioinformatique sera décisive pour l'avenir de la discipline. D'ores et déjà, de part le réseau de virologistes impliqués dès l'origine dans ce projet, il y a une opportunité unique de créer un pôle de référence dans le monde de la bioinformatique virale. Cela dit, s'il n'est pas question de diffuser l'architecture de VaZyMolO vers l'extérieur, un rapprochement vers le projet du NCBI (VGP) ou la base de données Américano-Européenne UniProt serait néanmoins souhaitable. D'un coté, le VGP présente l'avantage d'être déjà constitué en équipe spécialisée en virologie. Leurs critères d'annotation et leur mode opératoire sont parfaitement compatibles avec ceux de VaZyMolO. Le projet bénéficie en plus du concours de spécialistes reconnus, comme E. Koonin ou A.E. Gorbalenya, et d'infrastructures uniques grâce à son intégration au NCBI. D'un autre coté, UniProt a intégré les nouveaux critères d'annotation qui sont aussi compatibles avec les nôtres. Dans les deux cas, il est possible d'envisager de mettre à disposition une partie des données, sous format échangeable, qui seront par la suite prises en charge par le NCBI ou UniProt. La contrepartie serait évidemment d'alléger la part de gestion publique qui incombe aux administrateurs.

En conclusion, la bioinformatique est de nos jours l'étape primordiale dans la recherche et le suivi de projets ayant pour thèmes, entre autres, la virologie et la biologie structurale. Les outils, méthodes et résultats exposés dans cette thèse devraient contribuer à définir et suivre de nouvelles voies innovantes dans ce domaine.

Références

Alba, M.M., Lee, D., Pearl, F.M., Shepherd, A.J., Martin, N., Orengo, C.A., and Kellam, P. 2001. VIDA: a virus database system for the organization of animal virus genome open reading frames. *Nucleic Acids Res* **29:** 133-136.

Alexandrov, N.N., and Luethy, R. 1998. Alignment algorithm for homology modeling and threading. *Protein Sci* **7:** 254-258.

Alfadhli, A., Love, Z., Arvidson, B., Seeds, J., Willey, J., and Barklis, E. 2001. Hantavirus nucleocapsid protein oligomerization. *J Virol* **75:** 2019-2023.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. 1990. Basic local alignment search tool. *J Mol Biol* **215:** 403-410.

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

Apweiler, R., Bairoch, A., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., et al. 2004. UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res* **32 Database issue:** D115-119.

Attwood, T.K., and Beck, M.E. 1994. PRINTS--a protein motif fingerprint database. *Protein Eng* **7:** 841-848.

Bairoch, A. 1991. PROSITE: a dictionary of sites and patterns in proteins. *Nucleic Acids Res* **19 Suppl:** 2241-2245.

Bairoch, A., and Apweiler, R. 1996. The SWISS-PROT protein sequence data bank and its new supplement TREMBL. *Nucleic Acids Res* **24:** 21-25.

Bairoch, A., and Boeckmann, B. 1991. The SWISS-PROT protein sequence data bank. *Nucleic Acids Res* **19 Suppl:** 2247-2249.

Bao, Y., Federhen, S., Leipe, D., Pham, V., Resenchuk, S., Rozanov, M., Tatusov, R., and Tatusova, T. 2004. National center for biotechnology information viral genomes project. *J Virol* **78:** 7291-7298.

Barker, W.C., George, D.G., Mewes, H.W., and Tsugita, A. 1992. The PIR-International Protein Sequence Database. *Nucleic Acids Res* **20 Suppl:** 2023-2026.

Barker, W.C., Hunt, L.T., George, D.G., Yeh, L.S., Chen, H.R., Blomquist, M.C., Seibel-Ross, E.I., Elzanowski, A., Bair, J.K., and Ferrick, D.A. 1987a. Protein sequence database of the protein identification resource (PIR). *Protein Seq Data Anal* **1:** 43-98.

Barker, W.C., Hunt, L.T., George, D.G., Yeh, L.S., Chen, H.R., Blomquist, M.C., Seibel-Ross, E.I., Elzanowski, A., Bair, J.K., and Ferrick, D.A. 1987b. Protein sequence database of the protein identification resource (PIR). *Protein Seq Data Anal* **1:** 129-175.

Barker, W.C., Hunt, L.T., George, D.G., Yeh, L.S., Chen, H.R., Blomquist, M.C., Seibel-Ross, E.I., Elzanowski, A., Bair, J.K., and Ferrick, D.A. 1988. Protein Sequence Database of the Protein Identification Resource (PIR). *Protein Seq Data Anal* **1:** 195-250.

Barrett, A.J. 2004. Bioinformatics of proteases in the MEROPS database. *Curr Opin Drug Discov Devel* **7:** 334-341.

Barrow CJ, and MG., Z. 1991. Solution structures of beta peptide and its constituent fragments: relation to amyloid deposition. *Science* **12:** 179-182.

Bateman, A., Birney, E., Durbin, R., Eddy, S.R., Howe, K.L., and Sonnhammer, E.L. 2000. The Pfam protein families database. *Nucleic Acids Res* **28:** 263-266.

Benson, D.A., Boguski, M., Lipman, D.J., and Ostell, J. 1994. GenBank. *Nucleic Acids Res* **22:** 3441-3444.

Berger, B., Wilson, D.B., Wolf, E., Tonchev, T., Milla, M., and Kim, P.S. 1995. Predicting coiled coils by use of pairwise residue correlations. *Proc Natl Acad Sci U S A* **92:** 8259-8263.

Berman, H.M., Gelbin, A., and Westbrook, J. 1996. Nucleic acid crystallography: a view from the nucleic acid database. *Prog Biophys Mol Biol* **66:** 255-288.

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.

Bernstein, F.C., Koetzle, T.F., Williams, G.J., Meyer, E.F., Jr., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., and Tasumi, M. 1977. The Protein Data Bank. A computer-based archival file for macromolecular structures. *Eur J Biochem* **80:** 319-324.

Bhella, D., Ralph, A., Murphy, L.B., and Yeo, R.P. 2002. Significant differences in nucleocapsid morphology within the Paramyxoviridae. *J Gen Virol* **83:** 1831-1839.

Blanchard, L., Tarbouriech, N., Blackledge, M., Timmins, P., Burmeister, W.P., Ruigrok, R.W., and Marion, D. 2004. Structure and dynamics of the nucleocapsid-binding domain of the Sendai virus phosphoprotein in solution. *Virology* **319:** 201-211.

Bourhis, J.M., Johansson, K., Receveur-Brechot, V., Oldfield, C.J., Dunker, K.A., Canard, B., and Longhi, S. 2004. The C-terminal domain of measles virus nucleoprotein belongs to the class of intrinsically disordered proteins that fold upon binding to their physiological partner. *Virus Res* **99:** 157-167.

Boutselakis, H., Dimitropoulos, D., Fillon, J., Golovin, A., Henrick, K., Hussain, A., Ionides, J., John, M., Keller, P.A., Krissinel, E., et al. 2003. E-MSD: the European Bioinformatics Institute Macromolecular Structure Database. *Nucleic Acids Res* **31:** 458-462.

Brenner, S.E., Chothia, C., and Hubbard, T.J. 1998. Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc Natl Acad Sci U S A* **95:** 6073-6078.

Bressanelli, S., Tomei, L., Rey, F.A., and De Francesco, R. 2002. Structural analysis of the hepatitis C virus RNA polymerase in complex with ribonucleotides. *J Virol* **76:** 3482-3492.

Bryant, S.H., and Altschul, S.F. 1995. Statistics of sequence-structure threading. *Curr Opin Struct Biol* **5:** 236-244.

Butcher, S.J., Grimes, J.M., Makeyev, E.V., Bamford, D.H., and Stuart, D.I. 2001. A mechanism for initiating RNA-dependent RNA polymerization. *Nature* **410:** 235-240.

Callebaut, I., Labesse, G., Durand, P., Poupon, A., Canard, L., Chomilier, J., Henrissat, B., and Mornon, J.P. 1997. Deciphering protein sequence information through hydrophobic cluster analysis (HCA): current status and perspectives. *Cell Mol Life Sci* **53:** 621-645.

Cevik, B., Holmes, D.E., Vrotsos, E., Feller, J.A., Smallwood, S., and Moyer, S.A. 2004. The phosphoprotein (P) and L binding sites reside in the N-terminus of the L subunit of the measles virus RNA polymerase. *Virology* **327:** 297-306.

Chandrika, R., Horikami, S.M., Smallwood, S., and Moyer, S.A. 1995. Mutations in conserved domain I of the Sendai virus L polymerase protein uncouple transcription and replication. *Virology* **213:** 352-363.

Choi, K.H., Groarke, J.M., Young, D.C., Kuhn, R.J., Smith, J.L., Pevear, D.C., and Rossmann, M.G. 2004. The structure of the RNA-dependent RNA polymerase from bovine viral diarrhea virus establishes the role of GTP in de novo initiation. *Proc Natl Acad Sci U S A* **101:** 4425-4430.

Chou, P.Y., and Fasman, G.D. 1974a. Conformational parameters for amino acids in helical, beta-sheet, and random coil regions calculated from proteins. *Biochemistry* **13:** 211-222.

Chou, P.Y., and Fasman, G.D. 1974b. Prediction of protein conformation. *Biochemistry* **13:** 222-245.

Claros, M.G., Brunak, S., and von Heijne, G. 1997. Prediction of N-terminal protein sorting signals. *Curr Opin Struct Biol* **7:** 394-398.

Claros, M.G., and von Heijne, G. 1994. TopPred II: an improved software for membrane protein structure predictions. *Comput Appl Biosci* **10:** 685-686.

Claverie J-M., and D., S. 1993. Information enhancement methods for large scale sequence analysis. *Computers and Chemistry* **17:** 191-201.

Corpet, F. 1988. Multiple sequence alignment with hierarchical clustering. *Nucleic Acids Res* **16:** 10881-10890.

Corpet, F., Gouzy, J., and Kahn, D. 1999a. Browsing protein families via the 'Rich Family Description' format. *Bioinformatics* **15:** 1020-1027.

Corpet, F., Gouzy, J., and Kahn, D. 1999b. Recent improvements of the ProDom database of protein domain families. *Nucleic Acids Res* **27:** 263-267.

Corpet, F., Servant, F., Gouzy, J., and Kahn, D. 2000. ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons. *Nucleic Acids Res* **28:** 267-269.

Coutinho, P.M., and B., H. 1999a. Carbohydrate-active enzymes: an integrated database approach. In "*Recent Advances in Carbohydrate Bioengineering*". (eds. G.D. H.J. Gilbert, B. Henrissat and, and B. Svensson), pp. 3-12. The Royal Society of Chemistry, Cambridge.

Coutinho, P.M., and B., H. 1999b. The modular structure of cellulases and other carbohydrate-active enzymes: an integrated database approach. In *Genetics, Biochemistry and Ecology of Cellulose Degradation*. (ed. K.H. K. Ohmiya, K. Sakka, Y. Kobayashi,S. Karita & T. Kimura.), pp. 15-23. Uni Publishers Co., Tokyo.

Cserzo, M., Wallin, E., Simon, I., von Heijne, G., and Elofsson, A. 1997. Prediction of transmembrane alpha-helices in prokaryotic membrane proteins: the dense alignment surface method. *Protein Eng* **10:** 673-676.

Cuff, J.A., Clamp, M.E., Siddiqui, A.S., Finlay, M., and Barton, G.J. 1998. JPred: a consensus secondary structure prediction server. *Bioinformatics* **14:** 892-893.

Das, T., Mathur, M., Gupta, A.K., Janssen, G.M., and Banerjee, A.K. 1998. RNA polymerase of vesicular stomatitis virus specifically associates with translation elongation factor-1 alphabetagamma for its activity. *Proc Natl Acad Sci U S A* **95:** 1449-1454.

Dayhoff, M.O. 1965. Computer aids to protein sequence determination. *J Theor Biol* **8:** 97-112.

Dayhoff, M.O. 1969. Computer analysis of protein evolution. *Sci Am* **221:** 86-95.

Dayhoff MO, Schwartz RM, and BC., O. 1978. A model of evolutionary change in proteins. In *Atlas of Protein Sequence and Structure*. (ed. M.O. Dayhoff), pp. 345–352. National Biomedical Research Foundation, Washington, DC.

Delarue, M., Poch, O., Tordo, N., Moras, D. & Argos, P. (1990). An attempt to unify the structure of polymerases. Protein *Eng* **3**, 416-7

Dille, B.J., Surowy, T.K., Gutierrez, R.A., Coleman, P.F., Knigge, M.F., Carrick, R.J., Aach, R.D., Hollinger, F.B., Stevens, C.E., Barbosa, L.H., et al. 1997. An ELISA for detection of antibodies to the E2 protein of GB virus C. *J Infect Dis* **175:** 458-461.

Dunker, A.K., and Obradovic, Z. 2001. The protein trinity--linking function and disorder. *Nat Biotechnol* **19:** 805-806.

Eddy, S.R. 1996. Hidden Markov models. *Curr Opin Struct Biol* **6:** 361-365.

Edgar, R.C. 2004a. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5:** 113.

Edgar, R.C. 2004b. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32:** 1792-1797.

Egloff, M.P., Benarroch, D., Selisko, B., Romette, J.L., and Canard, B. 2002. An RNA cap (nucleoside-2'-O-)-methyltransferase in the flavivirus RNA polymerase NS5: crystal structure and functional characterization. *Embo J* **21:** 2757-2768.

Egloff, M.P., Ferron, F., Campanacci, V., Longhi, S., Rancurel, C., Dutartre, H., Snijder, E.J., Gorbalenya, A.E., Cambillau, C., and Canard, B. 2004. The severe acute respiratory syndrome-coronavirus replicative protein nsp9 is a single-stranded RNA-binding subunit unique in the RNA virus world. *Proc Natl Acad Sci U S A* **101:** 3792-3796.

Feng, D.F., and Doolittle, R.F. 1997. Converting amino acid alignment scores into measures of evolutionary time: a simulation study of various relationships. *J Mol Evol* **44:** 361-370.

Fischer, D. 2000. Hybrid fold recognition: combining sequence derived properties with evolutionary information. *Pac Symp Biocomput***:** 119-130.

Frishman, D., and Argos, P. 1995. Knowledge-based protein secondary structure assignment. *Proteins* **23:** 566-579.

Galperin, M.Y. 2004. The Molecular Biology Database Collection: 2004 update. *Nucleic Acids Res* **32 Database issue:** D3-22.

Galperin, M.Y., and Grishin, N.V. 2000. The synthetase domains of cobalamin biosynthesis amidotransferases cobB and cobQ belong to a new family of ATP-dependent amidoligases, related to dethiobiotin synthetase. *Proteins* **41:** 238-247.

Galtier, N., Gouy, M., and Gautier, C. 1996. SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. *Comput Appl Biosci* **12:** 543-548.

Garnier, J., Gibrat, J.F., and Robson, B. 1996. GOR method for predicting protein secondary structure from amino acid sequence. *Methods Enzymol* **266:** 540-553.

Garnier, J., Osguthorpe, D.J., and Robson, B. 1978. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J Mol Biol* **120:** 97-120.

Golovin, A., Oldfield, T.J., Tate, J.G., Velankar, S., Barton, G.J., Boutselakis, H., Dimitropoulos, D., Fillon, J., Hussain, A., Ionides, J.M., et al. 2004. E-MSD: an integrated data resource for bioinformatics. *Nucleic Acids Res* **32 Database issue:** D211-216.

Gorbalenya, A.E., and Koonin, E.V. 1989. Viral proteins containing the purine NTP-binding sequence pattern. *Nucleic Acids Res* **17:** 8413-8440.

Gorbalenya, A.E., and Koonin, E.V. 1991. Endonuclease (R) subunits of type-I and type-III restriction-modification enzymes contain a helicase-like domain. *FEBS Lett* **291:** 277-281.

Gorbalenya, A.E., Koonin, E.V., Donchenko, A.P., and Blinov, V.M. 1989. Two related superfamilies of putative helicases involved in replication, recombination, repair and expression of DNA and RNA genomes. *Nucleic Acids Res* **17:** 4713-4730.

Gough, J., and Chothia, C. 2002. SUPERFAMILY: HMMs representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments. *Nucleic Acids Res* **30:** 268-272.

Gough, J., Karplus, K., Hughey, R., and Chothia, C. 2001. Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *J Mol Biol* **313:** 903-919.

Heggeness, M.H., Scheid, A., and Choppin, P.W. 1980. Conformation of the helical nucleocapsids of paramyxoviruses and vesicular stomatitis virus: reversible coiling and uncoiling induced by changes in salt concentration. *Proc Natl Acad Sci U S A* **77:** 2631-2635.

Henikoff, S., and Henikoff, J.G. 1992. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A* **89:** 10915-10919.

Henikoff, S., and Henikoff, J.G. 1994. Protein family classification based on searching a database of blocks. *Genomics* **19:** 97-107.

Hercyk, N., Horikami, S.M., and Moyer, S.A. 1988. The vesicular stomatitis virus L protein possesses the mRNA methyltransferase activities. *Virology* **163:** 222-225.

Heringlake, S., Ockenga, J., Tillmann, H.L., Trautwein, C., Meissner, D., Stoll, M., Hunt, J., Jou, C., Solomon, N., Schmidt, R.E., et al. 1998. GB virus C/hepatitis G virus infection: a favorable prognostic factor in human immunodeficiency virus-infected patients? *J Infect Dis* **177:** 1723-1726.

Hirokawa, T., Boon-Chieng, S., and Mitaku, S. 1998. SOSUI: classification and secondary structure prediction system for membrane proteins. *Bioinformatics* **14:** 378-379.

Hiscock, D., and Upton, C. 2000. Viral Genome DataBase: storing and analyzing genes and proteins from complete viral genomes. *Bioinformatics* **16:** 484-485.

Hofmann, K., and Soffel, W. 1993. TM-Base A database of membrane spanning proteins segment. In *Biol. Chem.* (ed. Hoppe-Seyler), pp. 166.

Holm, L., Ouzounis, C., Sander, C., Tuparev, G., and Vriend, G. 1992. A database of protein structure families with common folding motifs. *Protein Sci* **1:** 1691-1698.

Holm, L., and Sander, C. 1997. Dali/FSSP classification of three-dimensional protein folds. *Nucleic Acids Res* **25:** 231-234.

Holmes, K.V. 2001. Coronaviruses. In *Fields VIROLOGY.* (eds. D.M. Knipe, and P.M. Howley), pp. 1187-1204. Lippincott Williams & Wilkins, Philadelphia.

Huang, Q., Yu, L., Petros, A.M., Gunasekera, A., Liu, Z., Xu, N., Hajduk, P., Mack, J., Fesik, S.W., and Olejniczak, E.T. 2004. Structure of the N-terminal RNA-binding domain of the SARS CoV nucleocapsid protein. *Biochemistry* **43:** 6059-6063.

Ismail, M.M., Cho, K.O., Ward, L.A., Saif, L.J., and Saif, Y.M. 2001. Experimental bovine coronavirus in turkey poults and young chickens. *Avian Dis* **45:** 157-163.

Jansson, A., Niemi, J., Lindqvist, Y., Mantsala, P., and Schneider, G. 2003. Crystal structure of aclacinomycin-10-hydroxylase, a S-adenosyl-L-methionine-dependent methyltransferase homolog involved in anthracycline biosynthesis in Streptomyces purpurascens. *J Mol Biol* **334:** 269-280.

Jaroszewski, L., Rychlewski, L., and Godzik, A. 2000. Improving the quality of twilight-zone alignments. *Protein Sci* **9:** 1487-1496.

Johansson, K., Bourhis, J.M., Campanacci, V., Cambillau, C., Canard, B., and Longhi, S. 2003. Crystal structure of the measles virus phosphoprotein domain responsible for the induced folding of the C-terminal domain of the nucleoprotein. *J Biol Chem* **278:** 44567-44573.

Johnson, M.S., and Overington, J.P. 1993. A structural basis for sequence comparisons. An evaluation of scoring methodologies. *J Mol Biol* **233:** 716-738.

Jones, D.T. 1999. GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. *J Mol Biol* **287:** 797-815.

Jones, D.T., and Thornton, J.M. 1996. Potential energy functions for threading. *Curr Opin Struct Biol* **6:** 210-216.

Jones, D.T., Tress, M., Bryson, K., and Hadley, C. 1999. Successful recognition of protein folds using threading methods biased by sequence similarity and predicted secondary structure. *Proteins* **Suppl 3:** 104-111.

Kao, C.C., Singh, P., and Ecker, D.J. 2001. De novo initiation of viral RNA-dependent RNA synthesis. *Virology* **287:** 251-260.

Karlin, D. 2002. Etude structurale des protéines du complexe réplicatif du virus de la rougeole. In *Ecole doctorale des Sciences de la Vie et de la Santé*, pp. 198. UIVERSITE DE LA MEDITERRANEE, Marseille.

Karlin, D., Ferron, F., Canard, B., and Longhi, S. 2003. Structural disorder and modular organization in Paramyxovirinae N and P. *J Gen Virol* **84:** 3239-3252.

Karlin, D., Longhi, S., Receveur, V., and Canard, B. 2002. The N-terminal domain of the phosphoprotein of Morbilliviruses belongs to the natively unfolded class of proteins. *Virology* **296:** 251-262.

Karlin, S., and Altschul, S.F. 1993. Applications and statistics for multiple high-scoring segments in molecular sequences. *Proc Natl Acad Sci U S A* **90:** 5873-5877.

Karplus, K., Barrett, C., and Hughey, R. 1998. Hidden Markov models for detecting remote protein homologies. *Bioinformatics* **14:** 846-856.

Karplus, K., and Hu, B. 2001. Evaluation of protein multiple alignments by SAM-T99 using the BAliBASE multiple alignment test set. *Bioinformatics* **17:** 713-720.

Karplus, K., Sjolander, K., Barrett, C., Cline, M., Haussler, D., Hughey, R., Holm, L., and Sander, C. 1997. Predicting protein structure using hidden Markov models. *Proteins* **Suppl 1:** 134-139.

Katoh, K., Misawa, K., Kuma, K., and Miyata, T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* **30:** 3059-3066.

Kelley, L.A., MacCallum, R.M., and Sternberg, M.J. 2000. Enhanced genome annotation using structural profiles in the program 3D- PSSM. *J Mol Biol* **299:** 499-520.

Kolakofsky, D., Le Mercier, P., Iseni, F., and Garcin, D. 2004. Viral DNA polymerase scanning and the gymnastics of Sendai virus RNA synthesis. *Virology* **318:** 463-473.

Koonin, E.V. 1993. Computer-assisted identification of a putative methyltransferase domain in NS5 protein of flaviviruses and lambda 2 protein of reovirus. *J Gen Virol* **74 (Pt 4):** 733-740.

Krogh, A., Larsson, B., von Heijne, G., and Sonnhammer, E.L. 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol* **305:** 567-580.

Lai, M.C., and Holmes, K.V. 2001. *Coronaviridae*: The Viruses and their Replication. In *Fields VIROLOGY*, Fourth ed. (eds. D.M. Knipe, and P.M. Howley), pp. 1163-1186. Lippincott Williams & Wilkins, Philadelphia.

Laine, D., Trescol-Biemont, M.C., Longhi, S., Libeau, G., Marie, J.C., Vidalain, P.O., Azocar, O., Diallo, A., Canard, B., Rabourdin-Combe, C., et al. 2003. Measles virus (MV) nucleoprotein binds to a novel cell surface receptor distinct from FcgammaRII via its C-terminal domain: role in MV-induced immunosuppression. *J Virol* **77:** 11332-11346.

Lamb, R., and Kolakofsky, D. 2001. Paramyxoviridae: the viruses and their replication. In *Fields VIROLOGY*. (eds. D.M. Knipe, and P.M. Howley), pp. 1305-1340. Lippincott Williams & Wilkins, Philadelphia.

Larson, D.J., Morehouse, L.G., Solorzano, R.F., and Kinden, D.A. 1979. Transmissible gastroenteritis in neonatal dogs: experimental intestinal infection with transmissible gastroenteritis virus. *Am J Vet Res* **40:** 477-486.

Lefrere, J.J., Ferec, C., Roudot-Thoraval, F., Loiseau, P., Cantaloube, J.F., Biagini, P., Mariotti, M., LeGac, G., and Mercier, B. 1999a. GBV-C/hepatitis G virus (HGV) RNA load in immunodeficient individuals and in immunocompetent individuals. *J Med Virol* **59:** 32-37.

Lefrere, J.J., Loiseau, P., Maury, J., Lasserre, J., Mariotti, M., Ravera, N., Lerable, J., Lefevre, G., Morand-Joubert, L., and Girot, R. 1997. Natural history of GBV-C/hepatitis G virus infection through the follow-up of GBV-C/hepatitis G virus-infected blood donors and recipients studied by RNA polymerase chain reaction and anti-E2 serology. *Blood* **90:** 3776-3780.

Lefrere, J.J., Mariotti, M., Lerable, J., Thauvin, M., and Girot, R. 1996. Long-term persistence of hepatitis G virus in immunocompetent patients. *Lancet* **348:** 1174-1175.

Lefrere, J.J., Roudot-Thoraval, F., Morand-Joubert, L., Petit, J.C., Lerable, J., Thauvin, M., and Mariotti, M. 1999b. Carriage of GB virus C/hepatitis G virus RNA is associated with a slower immunologic, virologic, and clinical progression of human immunodeficiency virus disease in coinfected persons. *J Infect Dis* **179:** 783-789.

Levitt, M., and Chothia, C. 1976. Structural patterns in globular proteins. *Nature* **261:** 552-558.

Lim, V.I. 1974a. Algorithms for prediction of alpha-helical and beta-structural regions in globular proteins. *J Mol Biol* **88:** 873-894.

Lim, V.I. 1974b. Structural principles of the globular organization of protein chains. A stereochemical theory of globular protein secondary structure. *J Mol Biol* **88:** 857-872.

Lindenbach, B.D., and Rice, C.M. 2001. Flaviridae : The viruses and their replication. In *Fields virology*, Fourth ed. (eds. D.M. Knipe, and P.M. Howley), pp. 991-1042.

Lindwall, G., Chau, M., Gardner, S.R., and Kohlstaedt, L.A. 2000. A sparse matrix approach to the solubilization of overexpressed proteins. *Protein Eng* **13:** 67-71.

Linnen, J., Wages, J., Jr., Zhang-Keck, Z.Y., Fry, K.E., Krawczynski, K.Z., Alter, H., Koonin, E., Gallagher, M., Alter, M., Hadziyannis, S., et al. 1996. Molecular cloning and disease association of hepatitis G virus: a transfusion-transmissible agent. *Science* **271:** 505-508.

Liu, J., Tan, H., and Rost, B. 2002. Loopy proteins appear conserved in evolution. *J Mol Biol* **322:** 53-64.

Lobocka, M.B., Rose, D.J., Plunkett, G., 3rd, Rusin, M., Samojedny, A., Lehnherr, H., Yarmolinsky, M.B., and Blattner, F.R. 2004. Genome of bacteriophage P1. *J Bacteriol* **186:** 7032-7068.

Longhi, S., Receveur-Brechot, V., Karlin, D., Johansson, K., Darbon, H., Bhella, D., Yeo, R., Finet, S., and Canard, B. 2003. The C-terminal domain of the measles virus nucleoprotein is intrinsically disordered and folds upon binding to the C-terminal moiety of the phosphoprotein. *J Biol Chem* **278:** 18638-18648.

Lupas, A. 1996a. Coiled coils: new structures and new functions. *Trends Biochem Sci* **21:** 375-382.

Lupas, A. 1996b. Prediction and analysis of coiled-coil structures. *Methods Enzymol* **266:** 513-525.

Lupas, A. 1997. Predicting coiled-coil regions in proteins. *Curr Opin Struct Biol* **7:** 388-393.

Lupas, A., Van Dyke, M., and Stock, J. 1991. Predicting coiled coils from protein sequences. *Science* **252:** 1162-1164.

Mackenzie, J.S., Chua, K.B., Daniels, P.W., Eaton, B.T., Field, H.E., Hall, R.A., Halpin, K., Johansen, C.A., Kirkland, P.D., Lam, S.K., et al. 2001. Emerging viral diseases of Southeast Asia and the Western Pacific. *Emerg Infect Dis* **7:** 497-504.

Madej, T., Gibrat, J.F., and Bryant, S.H. 1995. Threading a database of protein cores. *Proteins* **23:** 356-369.

Mallick, P., Goodwill, K.E., Fitz-Gibbon, S., Miller, J.H., and Eisenberg, D. 2000. Selecting protein targets for structural genomics of Pyrobaculum aerophilum: validating automated fold assignment methods by using binary hypothesis testing. *Proc Natl Acad Sci U S A* **97:** 2450-2455.

Maniloff, J. 1995. Identification and classification of viruses that have not been propagated. *Arch Virol* **140:** 1515-1520.

Marchler-Bauer, A., Addess, K.J., Chappey, C., Geer, L., Madej, T., Matsuo, Y., Wang, Y., and Bryant, S.H. 1999. MMDB: Entrez's 3D structure database. *Nucleic Acids Res* **27:** 240-243.

Marchler-Bauer, A., Panchenko, A.R., Shoemaker, B.A., Thiessen, P.A., Geer, L.Y., and Bryant, S.H. 2002. CDD: a database of conserved domain alignments with links to domain three-dimensional structure. *Nucleic Acids Res* **30:** 281-283.

Marie, J.C., Kehren, J., Trescol-Biemont, M.C., Evlashev, A., Valentin, H., Walzer, T., Tedone, R., Loveland, B., Nicolas, J.F., Rabourdin-Combe, C., et al. 2001. Mechanism of measles virus-induced suppression of inflammatory immune responses. *Immunity* **14:** 69-79.

Marie, J.C., Saltel, F., Escola, J.M., Jurdic, P., Wild, T.F., and Horvat, B. 2004. Cell surface delivery of the measles virus nucleoprotein: a viral strategy to induce immunosuppression. *J Virol* **78:** 11952-11961.

Mavrakis, M., McCarthy, A.A., Roche, S., Blondel, D., and Ruigrok, R.W. 2004. Structure and function of the C-terminal domain of the polymerase cofactor of rabies virus. *J Mol Biol* **343:** 819-831.

McGuffin, L.J., Bryson, K., and Jones, D.T. 2000. The PSIPRED protein structure prediction server. *Bioinformatics* **16:** 404-405.

McGuffin, L.J., and Jones, D.T. 2002. Targeting novel folds for structural genomics. *Proteins* **48:** 44-52.

Meller, J., and Elber, R. 2001. Linear programming optimization and a double statistical filter for protein threading protocols. *Proteins* **45:** 241-261.

Mirny, L.A., Finkelstein, A.V., and Shakhnovich, E.I. 2000. Statistical significance of protein structure prediction by threading. *Proc Natl Acad Sci U S A* **97:** 9978-9983.

Mitaku, S., Hirokawa, T., and Tsuji, T. 2002. Amphiphilicity index of polar amino acids as an aid in the characterization of amino acid preference at membrane-water interfaces. *Bioinformatics* **18:** 608-616.

Miyazawa, S., and Jernigan, R.L. 1999. An empirical energy potential with a reference state for protein fold and sequence recognition. *Proteins* **36:** 357-369.

Morgenstern, B. 1999. DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment. *Bioinformatics* **15:** 211-218.

Morgenstern, B. 2004. DIALIGN: multiple DNA and protein sequence alignment at BiBiServ. *Nucleic Acids Res* **32:** W33-36.

Morgenstern, B., Frech, K., Dress, A., and Werner, T. 1998. DIALIGN: finding local similarities by multiple sequence alignment. *Bioinformatics* **14:** 290-294.

Muerhoff, A.S., Leary, T.P., Simons, J.N., Pilot-Matias, T.J., Dawson, G.J., Erker, J.C., Chalmers, M.L., Schlauder, G.G., Desai, S.M., and Mushahwar, I.K. 1995. Genomic organization of GB viruses A and B: two new members of the Flaviviridae associated with GB agent hepatitis. *J Virol* **69:** 5621-5630.

Nagano, K. 1973. Logical analysis of the mechanism of protein folding. I. Predictions of helices, loops and beta-structures from primary structure. *J Mol Biol* **75:** 401-420.

Nakai, K., and Horton, P. 1999. PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends Biochem Sci* **24:** 34-36.

Needleman, S.B., and Wunsch, C.D. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* **48:** 443-453.

Nelson, C.A., Pekosz, A.S., Lee, C.A., Diamond, M.S., and Fremont, D.H. 2004. Structure and Intracellular Targeting of the Sars-Coronavirus Orf7A Accessory Protein.

Nielsen, H., Engelbrecht, J., Brunak, S., and von Heijne, G. 1997. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng* **10:** 1-6.

Notredame, C., Higgins, D.G., and Heringa, J. 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol* **302:** 205-217.

O'Reilly, E.K., and Kao, C.C. 1998. Analysis of RNA-dependent RNA polymerase structure and function as guided by known polymerase structures and computer predictions of secondary structure. *Virology* **252:** 287-303.

Obijeski, J.F., Bishop, D.H., Murphy, F.A., and Palmer, E.L. 1976a. Structural proteins of La Crosse virus. *J Virol* **19:** 985-997.

Obijeski, J.F., Bishop, D.H., Palmer, E.L., and Murphy, F.A. 1976b. Segmented genome and nucleocapsid of La Crosse virus. *J Virol* **20:** 664-675.

Ogino, T., Masaki Kobayashi, Minako Iwama, and Mizumoto, a.K. 2004. Sendai virus RNA-dependent RNA polymerase L protein catalyzes cap methylation of virus-specific mRNA. *Journal of of Biological Chemistry* **in press**.

Orengo, C.A., and Taylor, W.R. 1996. SSAP: sequential structure alignment program for protein structure comparison. *Methods Enzymol* **266:** 617-635.

Overington, J., Donnelly, D., Johnson, M.S., Sali, A., and Blundell, T.L. 1992. Environment-specific amino acid substitution tables: tertiary templates and prediction of protein folds. *Protein Sci* **1:** 216-226.

Panchenko, A., Marchler-Bauer, A., and Bryant, S.H. 1999. Threading with explicit models for evolutionary conservation of structure and sequence. *Proteins* **Suppl 3:** 133-140.

Panchenko, A.R., and Bryant, S.H. 2002. A comparison of position-specific score matrices based on sequence and structure alignments. *Protein Sci* **11:** 361-370.

Panchenko, A.R., Marchler-Bauer, A., and Bryant, S.H. 2000. Combination of threading potentials and sequence profiles improves fold recognition. *J Mol Biol* **296:** 1319-1331.

Pearl, F.M., Martin, N., Bray, J.E., Buchan, D.W., Harrison, A.P., Lee, D., Reeves, G.A., Shepherd, A.J., Sillitoe, I., Todd, A.E., et al. 2001. A rapid classification protocol for the CATH Domain Database to support structural genomics. *Nucleic Acids Res* **29:** 223-227.

Pearson, W.R., and Lipman, D.J. 1988. Improved tools for biological sequence comparison. *Proc Natl Acad Sci U S A* **85:** 2444-2448.

Persson, B., and Argos, P. 1997. Prediction of membrane protein topology utilizing multiple sequence alignments. *J Protein Chem* **16:** 453-457.

Pettersson, R.F., and von Bonsdorff, C.H. 1975. Ribonucleoproteins of Uukuniemi virus are circular. *J Virol* **15:** 386-392.

Pizzi, E., and Frontali, C. 2001. Low-complexity regions in Plasmodium falciparum proteins. *Genome Res* **11:** 218-229.

Poch, O., Blumberg, B.M., Bougueleret, L., and Tordo, N. 1990. Sequence comparison of five polymerases (L proteins) of unsegmented negative-strand RNA viruses: theoretical assignment of functional domains. *J Gen Virol* **71 (Pt 5):** 1153-1162.

Poch, O., Sauvaget, I., Delarue, M., and Tordo, N. 1989. Identification of four conserved motifs among the RNA-dependent polymerase encoding elements. *Embo J* **8:** 3867-3874.

Ponting, C.P., Schultz, J., Milpetz, F., and Bork, P. 1999. SMART: identification and annotation of domains from signalling and extracellular protein sequences. *Nucleic Acids Res* **27:** 229-232.

Prentice, E., McAuliffe, J., Lu, X., Subbarao, K., and Denison, M.R. 2004. Identification and characterization of severe acute respiratory syndrome coronavirus replicase proteins. *J Virol* **78:** 9977-9986.

Pringle, C.R. 1995. International Committee on Taxonomy of Viruses (ICTV) activity report. *Arch Virol* **140:** 2100-2103.

Raju, R., and Kolakofsky, D. 1989. The ends of La Crosse virus genome and antigenome RNAs within nucleocapsids are base paired. *J Virol* **63:** 122-128.

Raoult, D., Audic, S., Robert, C., Abergel, C., Renesto, P., Ogata, H., La Scola, B., Suzan, M., and Claverie, J.M. 2004. The 1.2-megabase genome sequence of Mimivirus. *Science* **306:** 1344-1350.

Rost, B. 1999. Twilight zone of protein sequence alignments. *Protein Eng* **12:** 85-94.

Rost, B., Fariselli, P., and Casadio, R. 1996. Topology prediction for helical transmembrane proteins at 86% accuracy. *Protein Sci* **5:** 1704-1718.

Rost, B., and Sander, C. 1993. Prediction of protein secondary structure at better than 70% accuracy. *J Mol Biol* **232:** 584-599.

Rost, B., and Sander, C. 1994. Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins* **19:** 55-72.

Rost, B., and Schneider, R. 1998. Pedestrian guide to analysing sequence databases. In *Core techniques in Biochemistry*. (ed. K. Ashman). Springer, Heidelberg.

Rychlewski, L., Jaroszewski, L., Li, W., and Godzik, A. 2000. Comparison of sequence profiles. Strategies for structural predictions using sequence information. *Protein Sci* **9:** 232-241.

Saitou, N., and Nei, M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* **4:** 406-425.

Salamov, A.A., and Solovyev, V.V. 1995. Prediction of protein secondary structure by combining nearest-neighbor algorithms and multiple sequence alignments. *J Mol Biol* **247:** 11-15.

Schmaljohn C, S., and JW., H. 2001. *Bunyaviridae* : The virus and their replication. In *Fields VIROLOGY*, Fourth ed. (eds. K. D., and H. P.), pp. 1581-1602. Lippincott Williams & Wilkins, Philadelphia.

Shi, J., Blundell, T.L., and Mizuguchi, K. 2001. FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *J Mol Biol* **310:** 243-257.

Siddell, S.G. 1995. The Coronaviridae: an introduction. In *The Coronaviridae*. (ed. S.G. Siddell), pp. 1-10. Plenum Press, New York.

Simons, J.N., Leary, T.P., Dawson, G.J., Pilot-Matias, T.J., Muerhoff, A.S., Schlauder, G.G., Desai, S.M., and Mushahwar, I.K. 1995a. Isolation of novel virus-like sequences associated with human hepatitis. *Nat Med* **1:** 564-569.

Simons, J.N., Pilot-Matias, T.J., Leary, T.P., Dawson, G.J., Desai, S.M., Schlauder, G.G., Muerhoff, A.S., Erker, J.C., Buijk, S.L., Chalmers, M.L., et al. 1995b. Identification of two flavivirus-like genomes in the GB hepatitis agent. *Proc Natl Acad Sci U S A* **92:** 3401-3405.

Sippl, M.J., and Flockner, H. 1996. Threading thrills and threats. *Structure* **4:** 15-19.

Smith, T.F., Lo Conte, L., Bienkowska, J., Gaitatzes, C., Rogers, R.G., Jr., and Lathrop, R. 1997. Current limitations to protein threading approaches. *J Comput Biol* **4:** 217-225.

Smith, T.F., and Waterman, M.S. 1981. Identification of common molecular subsequences. *J Mol Biol* **147:** 195-197.

Snell, N.J. 2004. Novel and re-emerging respiratory infections. *Expert Rev Anti Infect Ther* **2:** 405-412.

Sokal, R.R., and Sneath, P.H.A. 1963. *Principle of numerical taxonomy.*, San Fransisco.

Spadafora, D., and Perrault, J. 2003. Gly-rich motif of vesicular stomatitis virus L polymerase protein is a SAM-binding site for viral mRNA cap methylations and reulates polymerase activity. In *XII international conference on negative strand viruses*. (eds. B. Mahy, WJ., D. Kolakofsky, and M. Bendinelli), pp. 94. NSV, Pisa, Italy.

Sunyaev, S., Kuznetsov, E., Rodchenkov, I., and Tumanyan, V. 1997. Protein sequence-structure compatibility criteria in terms of statistical hypothesis testing. *Protein Eng* **10:** 635-646.

Sutton, G., Fry, E., Carter, L., Sainsbury, S., Walter, T., Nettleship, J., Berrow, N., Owens, R., Gilbert, R., Davidson, A., et al. 2004. The nsp9 replicase protein of SARS-coronavirus, structure and functional insights. *Structure (Camb)* **12:** 341-353.

Tacke, M., Kiyosawa, K., Stark, K., Schlueter, V., Ofenloch-Haehnle, B., Hess, G., and Engel, A.M. 1997. Detection of antibodies to a putative hepatitis G virus envelope protein. *Lancet* **349:** 318-320.

Talafous, J., Marcinowski, K.J., Klopman, G., and Zagorski, M.G. 1994. Solution structure of residues 1-28 of the amyloid beta-peptide. *Biochemistry* **33:** 7788-7796.

Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F., and Higgins, D.G. 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res* **25:** 4876-4882.

Thompson, J.D., Higgins, D.G., and 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.

Thompson, J.D., Plewniak, F., and Poch, O. 1999. A comprehensive comparison of multiple sequence alignment programs. *Nucleic Acids Res* **27:** 2682-2690.

Thompson, J.D., Plewniak, F., Thierry, J., and Poch, O. 2000. DbClustal: rapid and reliable global multiple alignments of protein sequences detected by database searches. *Nucleic Acids Res* **28:** 2919-2926.

Thompson, J.D., Thierry, J.C., and Poch, O. 2003. RASCAL: rapid scanning and correction of multiple sequence alignments. *Bioinformatics* **19:** 1155-1161.

Tompa, P. 2002. Intrinsically unstructured proteins. *Trends Biochem Sci* **27:** 527-533.

Toyoda, H., Fukuda, Y., Hayakawa, T., Takamatsu, J., and Saito, H. 1998. Effect of GB virus C/hepatitis G virus coinfection on the course of HIV infection in hemophilia patients in Japan. *J Acquir Immune Defic Syndr Hum Retrovirol* **17:** 209-213.

Tramontano, A., and Morea, V. 2003. Assessment of homology-based predictions in CASP5. *Proteins* **53 Suppl 6:** 352-368.

Troy, C.S., MacHugh, D.E., Bailey, J.F., Magee, D.A., Loftus, R.T., Cunningham, P., Chamberlain, A.T., Sykes, B.C., and Bradley, D.G. 2001. Genetic evidence for Near-Eastern origins of European cattle. *Nature* **410:** 1088-1091.

Tusnady, G.E., and Simon, I. 2001. The HMMTOP transmembrane topology prediction server. *Bioinformatics* **17:** 849-850.

Uchida, H. 1986. [DNA Data Bank of Japan]. *Tanpakushitsu Kakusan Koso***:** 159-162.

von Heijne, G. 1984. How signal sequences maintain cleavage specificity. *J Mol Biol* **173:** 243-251.

von Heijne, G. 1985. Signal sequences. The limits of variation. *J Mol Biol* **184:** 99-105.

Wheeler, D.L., Church, D.M., Lash, A.E., Leipe, D.D., Madden, T.L., Pontius, J.U., Schuler, G.D., Schriml, L.M., Tatusova, T.A., Wagner, L., et al. 2001. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **29:** 11-16.

Wolf, E., Kim, P.S., and Berger, B. 1997. MultiCoil: a program for predicting two- and three-stranded coiled coils. *Protein Sci* **6:** 1179-1189.

Wootton, J.C. 1994. Non-globular domains in protein sequences: automated segmentation using complexity measures. *Comput Chem* **18:** 269-285.

Wootton, J.C., and Federhen, S. 1996. Analysis of compositionally biased regions in sequence databases. *Methods Enzymol* **266:** 554-571.

Wootton JC., and S., F. 1993. Statistics of local complexity in amino acid sequences and sequence databases. *Computers and Chemistry* **17:** 149-163.

Wright, P.E., and Dyson, H.J. 1999. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J Mol Biol* **293:** 321-331.

Wu, C.H., Nikolskaya, A., Huang, H., Yeh, L.S., Natale, D.A., Vinayaka, C.R., Hu, Z.Z., Mazumder, R., Kumar, S., Kourtesis, P., et al. 2004. PIRSF: family classification system at the Protein Information Resource. *Nucleic Acids Res* **32 Database issue:** D112-114.

Xu, Y., Lou, Z., Liu, Y., Pang, H., Tien, P., Gao, G.F., and Rao, Z. 2004a. Crystal Structure of Severe Acute Respiratory Syndrome Coronavirus Spike Protein Fusion Core. *J Biol Chem* **279:** 49414-49419.

Xu, Y., Zhu, J., Liu, Y., Lou, Z., Yuan, F., Cole, D.K., Ni, L., Su, N., Qin, L., Li, X., et al. 2004b. Characterization of the heptad repeat regions, HR1 and HR2, and design of a fusion core structure model of the spike protein from severe acute respiratory syndrome (SARS) coronavirus. *Biochemistry* **43:** 14064-14071.

Yang, H., Yang, M., Ding, Y., Liu, Y., Lou, Z., Zhou, Z., Sun, L., Mo, L., Ye, S., Pang, H., et al. 2003. The crystal structures of severe acute respiratory syndrome virus main protease and its complex with an inhibitor. *Proc Natl Acad Sci U S A* **100:** 13190-13195.

Zhang, C., and Kim, S.H. 2000. Environment-dependent residue contact energies for proteins. *Proc Natl Acad Sci U S A* **97:** 2550-2555.