Heligeom: a multiscale approach to studying biomolecular helical assemblies with an application to RecA fillaments
Benjamin Boyer

To cite this version:
Heligeom
A multiscale approach to studying biomolecular helical assemblies with an application to RecA filaments

Thèse de doctorat de l'Université Pierre et Marie Curie

Benjamin Boyer

Under the supervision of Chantal PREVOST

20 june 2014

Jury committee

Pr Tâu Ha Duong Rapporteur
Dr Juan Cortés Rapporteur
Pr Martin Zacharias Examinateur
Pr Germain Trugnan Examinateur
Dr Jean-Marc Victor Examinateur
Dr Chantal Prévost Directrice de thèse
Note:
Pictures on the title page were made by David Goodsell. They represent from left to right microtubule filaments, recombinase filaments and actin filaments. More information can be found at http://mgl.scripps.edu/people/goodsell/.
Contents

I  Introduction  1

1  General introduction  3
   1.1  The cell  3
   1.2  General aim of the study  4
   1.3  Story of this work  6

2  Macromolecular assemblies  8
   2.1  Strength of association  11
   2.2  Stabilizing interactions  12
   2.3  Characteristics of macromolecular complexes  14
   2.4  Protein-DNA complexes  20
   2.5  Oligomers and polymers  23
   2.6  Towards higher level superstructures  28

3  Homologous Recombination  30
   3.1  Homologous recombination  31
   3.2  The RecA nucleoprotein filament  32
   3.3  ATP hydrolysis  39
   3.4  Conclusion  43

II  Methods  44

4  Introduction  46
   4.1  Methods and representation in PTools  46
       4.1.1  Macromolecular docking  46
       4.1.2  Coarse-grained models  47

5  PTools: overview and developments  50
   5.1  PTools  50
       5.1.1  Design  50
       5.1.2  PTools Objects and Functions  52
       5.1.3  Modules and Scripts  55
   5.2  PyAttract  56
<table>
<thead>
<tr>
<th>CONTENTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>6 Manipulation of flexible DNA in PTools</td>
</tr>
<tr>
<td>6.1 DNA in PTools</td>
</tr>
<tr>
<td>6.1.1 Editing DNA structures</td>
</tr>
<tr>
<td>6.1.2 Manipulating DNA structures</td>
</tr>
<tr>
<td>6.1.3 Deforming DNA structures</td>
</tr>
<tr>
<td>6.2 First application</td>
</tr>
<tr>
<td>7 Burdock</td>
</tr>
<tr>
<td>7.1 Overall strategy</td>
</tr>
<tr>
<td>7.1.1 Preliminary docking and pruning</td>
</tr>
<tr>
<td>7.1.2 From prediction to dot</td>
</tr>
<tr>
<td>7.1.3 Clustering</td>
</tr>
<tr>
<td>7.1.4 Axis construction</td>
</tr>
<tr>
<td>7.1.5 DNA construction</td>
</tr>
<tr>
<td>7.1.6 Optimization</td>
</tr>
<tr>
<td>7.2 Conclusion</td>
</tr>
<tr>
<td>8 Heligeom</td>
</tr>
<tr>
<td>8.1 Screw movement</td>
</tr>
<tr>
<td>8.2 Construction : 2GLS</td>
</tr>
<tr>
<td>8.3 Construction around a curved axis</td>
</tr>
<tr>
<td>8.4 Coupling Heligeom with PyAttract</td>
</tr>
<tr>
<td>8.5 Optimizing ring geometries</td>
</tr>
<tr>
<td>8.6 Measuring protein filament groove width</td>
</tr>
<tr>
<td>III Applications</td>
</tr>
<tr>
<td>9 Heligeom application : Construction</td>
</tr>
<tr>
<td>9.1 Extending modeling scales</td>
</tr>
<tr>
<td>9.2 Interactions between topologically linear assemblies</td>
</tr>
<tr>
<td>9.3 Structural interpretation of low resolution data</td>
</tr>
<tr>
<td>9.4 Transferring information from the local to the global levels</td>
</tr>
<tr>
<td>9.5 Beyond the topologically linear forms</td>
</tr>
<tr>
<td>10 Exploring protein filaments</td>
</tr>
<tr>
<td>10.1 Abstract</td>
</tr>
<tr>
<td>10.2 Author Summary</td>
</tr>
<tr>
<td>10.3 Introduction</td>
</tr>
<tr>
<td>10.4 Results</td>
</tr>
<tr>
<td>10.4.1 Characterizing known RecA assembly modes</td>
</tr>
<tr>
<td>10.4.2 Investigating geometries of RecA autoassembly</td>
</tr>
<tr>
<td>10.4.3 Variability within selected families of binding modes</td>
</tr>
<tr>
<td>10.4.4 Binding mode variations in a single fiber</td>
</tr>
</tbody>
</table>
List of Figures

1.1 Development of computing power of the most powerful computers. ........................................... 5
2.1 Lac repressor binding to DNA ......................................................... 10
2.2 Interfacial water molecules. .......................................................... 13
2.3 Definition of the buried surface area. ........................................... 14
2.4 Interface topology. ................................................................. 16
2.5 Surface remodeling: flexible loops. ........................................... 18
2.6 Proposed scheme for the thermodynamics of flexible protein-protein association. ..................... 19
2.7 Examples of protein-DNA complexes. ......................................... 21
2.8 Actin network in a keratocyte. ................................................. 25
2.9 Changes in the microtubule network morphology during the cell cycle. ................................. 26
2.10 Two assembly modes of DnaB cyclic hexamers. ....................................... 27
3.1 Scheme of the strand exchange process. .......................................... 31
3.2 ATP and its homologs. .............................................................. 33
3.3 Pitch variations of RecA-ATP and RecA-ADP filaments. ................. 33
3.4 Model of LexA binding to RecA. ................................................ 34
3.5 Crystal structures of the RecA filament. ....................................... 35
3.6 Atomic force microscopy images of hexameric RecA rings. ............... 36
3.7 Filament forms of recombinases from higher organisms. ................. 36
3.8 Variety in the oligomeric forms of recombinase proteins. .................... 37
3.9 Structure of RecA filaments with bound DNA. ................................ 37
3.10 RecA-bound DNA structure is intermediate between B and S DNA forms. ................................. 38
3.11 3D-SIM imaging of RecA bundles. ............................................. 39
3.12 Cox’s model of hydrolytic waves in a RecA filament. ..................... 40
3.13 Transitions between stretched and compressed forms of Rad51. ........ 41
3.14 Electron micrograph of mixed RecA filaments. ................................ 41
5.1 Overview of the PTools library. .................................................. 51
5.2 The force field equation used in pyAttract. .................................... 56
## LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.3</td>
<td>Coarse-grained representation in PyAttract.</td>
</tr>
<tr>
<td>5.4</td>
<td>The ATTRACT docking approach.</td>
</tr>
<tr>
<td>6.1</td>
<td>Coarse-grained model model of DNA.</td>
</tr>
<tr>
<td>6.2</td>
<td>Construction steps of a DNA oligomer.</td>
</tr>
<tr>
<td>6.3</td>
<td>The six inter base pair translation and orientation parameters.</td>
</tr>
<tr>
<td>6.4</td>
<td>Geometric criteria for stacking interactions.</td>
</tr>
<tr>
<td>6.5</td>
<td>Results of the Monte Carlo flexible docking procedure.</td>
</tr>
<tr>
<td>7.1</td>
<td>Schematic illustration of the Burdock approach.</td>
</tr>
<tr>
<td>7.2</td>
<td>Burdock step 1: Pruning.</td>
</tr>
<tr>
<td>7.3</td>
<td>Burdock step 2: Shaping clouds.</td>
</tr>
<tr>
<td>7.4</td>
<td>Burdock step 3: Clustering.</td>
</tr>
<tr>
<td>7.5</td>
<td>Burdock step 4: Axis extraction.</td>
</tr>
<tr>
<td>7.6</td>
<td>Burdock step 5: DNA reconstruction.</td>
</tr>
<tr>
<td>8.1</td>
<td>Scheme of the screw transformation between two monomers A and B.</td>
</tr>
<tr>
<td>8.2</td>
<td>Circular filament of 2GLS reconstructed with Heligeom.</td>
</tr>
<tr>
<td>8.3</td>
<td>Heligeom construction around a curved axis.</td>
</tr>
<tr>
<td>8.4</td>
<td>Optimizing near-cyclic geometries.</td>
</tr>
<tr>
<td>8.5</td>
<td>Groove width measurement.</td>
</tr>
<tr>
<td>9.1</td>
<td>Overlapping helical and cyclic forms of Dmc1.</td>
</tr>
<tr>
<td>9.3</td>
<td>Hyperstructure variability.</td>
</tr>
<tr>
<td>9.4</td>
<td>Small size complex between DNA and RecA filament.</td>
</tr>
<tr>
<td>9.5</td>
<td>Large size complex between DNA and RecA filament.</td>
</tr>
<tr>
<td>9.6</td>
<td>View of one turn of a RecA negatively supercoiled filament at atomic resolution.</td>
</tr>
<tr>
<td>9.7</td>
<td>Electrostatic potential around an actin filament.</td>
</tr>
<tr>
<td>9.8</td>
<td>Construction steps of a virus capsid.</td>
</tr>
<tr>
<td>10.1</td>
<td>Overview of the Heligeom/PyAttract integrated approach.</td>
</tr>
<tr>
<td>10.2</td>
<td>Scheme of geometric filtering.</td>
</tr>
<tr>
<td>10.3</td>
<td>RecA-RecA docking.</td>
</tr>
<tr>
<td>10.4</td>
<td>RecA auto-assemblage.</td>
</tr>
<tr>
<td>10.5</td>
<td>Exploration of the RecA-ADP and RecA-ATP structural families.</td>
</tr>
<tr>
<td>10.6</td>
<td>View of one turn of a RecA negatively supercoiled filament at atomic resolution.</td>
</tr>
<tr>
<td>10.7</td>
<td>Models of RecA filaments with alternating binding modes.</td>
</tr>
<tr>
<td>10.8</td>
<td>Accommodation of flexible regions in predicted RecA fiber forms.</td>
</tr>
<tr>
<td>Figure</td>
<td>Description</td>
</tr>
<tr>
<td>--------</td>
<td>-------------</td>
</tr>
<tr>
<td>10.9</td>
<td>Variability of the ADP and ATP fiber forms in the presence of flexible/mobile interaction component.</td>
</tr>
<tr>
<td>10.10</td>
<td>Helical parameters from Heligeom for regular assemblies obtained from ATTRACT docking simulations in RecA, RadA, and Dmc1.</td>
</tr>
<tr>
<td>10.11</td>
<td>Variation of the groove width along one filament turn.</td>
</tr>
<tr>
<td>A.1</td>
<td>Schéma d’une transformation de vissage entre deux monomères A et B.</td>
</tr>
<tr>
<td>A.2</td>
<td>Illustration de la construction d’une hélice par transformations de vissage.</td>
</tr>
<tr>
<td>A.3</td>
<td>Constructions sur un axe arbitraire.</td>
</tr>
<tr>
<td>A.4</td>
<td>Formes “voisines” en hélice et en anneau de Dmc1.</td>
</tr>
<tr>
<td>A.5</td>
<td>Potentiel électrostatique autour d’un filament d’actine.</td>
</tr>
<tr>
<td>A.6</td>
<td>Vue d’un tour de filament de RecA superenroulé, en résolution atomique.</td>
</tr>
<tr>
<td>A.7</td>
<td>Etapes de construction d’une capsïde de virus à partir de quatre unités en interaction.</td>
</tr>
<tr>
<td>A.8</td>
<td>Illustration de la diversité des formes possibles pour les oligomères de RecA.</td>
</tr>
</tbody>
</table>
Part I

Introduction
Chapter 1

General introduction

1.1 The cell

The name cell was proposed by R. Hooke in 1665 [1], following observations on cork conducted with a microscope of his own design. This first observation was soon followed by many others and eventually a theory was developed in 1939 [2], based on the observation of Matthias Jakob Schleiden on plants [3] and Theodor Schwann on notochord. They formulated the idea that the cell is a basic biological unit of life that can exist as an independent organism or as a brick for larger organisms.

The theory was soon completed by the idea that all living cells arise from pre-existing cells by division. This idea was proposed by Rudolf Virchow in 1858 [4], influenced by the work of Robert Remak [5] on embryology in 1855. By a famous series of experiments in 1861, Louis Pasteur [6, 7] refuted the concurrent theory of spontaneous generation, giving cell theory considerable momentum.

Since then, the study of the cell has been a corner stone of biology research.

The general organization of a cell is relatively simple: a plasma membrane mainly composed of a lipid bilayer separates the protoplasm from the outside environment. The protoplasm contains many biomolecules such as proteins and nucleic acids. Two types of cells exist: eukaryotic cells possess a nucleus (first observed by Antonie van Leeuwenhoek in 1722 [8]) that contains most of the genetic material of the cell, while in prokaryotic cells the genetic material is by contrast directly found in the protoplasm. This proposition of a fundamental division in the realm of life was formulated by Roger Stanier and C. B. van Niel in 1962 [9]. Further distinction, proposed by Carl Richard Woese [10], separates prokaryotes into the Bacteria and Archaea based on differences in the cell wall structure and genetics.

Most of the mechanical properties of the cell depend on the cytoskele-
ton\textsuperscript{1}, an ensemble of large protein filaments that structure and organize the cell. First thought unique to the eukaryotes, the presence of the cytoskeleton was later demonstrated in prokaryotes in 1991 \cite{12}. The cytoskeleton is composed of actin and microtubules in addition to diverse intermediate filaments in metazoan cells\textsuperscript{2}. Actin plays a role in muscle contraction, cell motility, cell division and cytokinesis and cell signalling. Microtubules play a role in the movement of secretory vesicles, organelles and intracellular substances, the internal structure of cilia and flagella and the cell division. Other filaments can play an important role in decisive functions of the cell, such as the mechanical support ensured by trichocytic keratins that make up hair or nails, or mechanisms fundamental for the cell survival such as homologous recombination performed by the RecA filaments.

The study of protein filamentous assemblies appears to be a key step in the understanding of the cell.

1.2 General aim of the study

These last years have testified quick progress in so-called “low resolution” techniques such as electron microscopy (EM) or small angle neutron or X-ray scattering (SAS), both for data generation and interpretation and for the range of biological systems that can be studied.

Impressive progress is also underway in the field of 3D cell imaging, notably using fluorescent proteins and high resolution apparatus (\textit{e.g.} 3D super-resolution microscopy and time lapse analysis), which permits for example to visualize the formation of supramolecular assemblies and their evolution along the cell cycle \cite{13}. Although still unsuited to detailed study at the atomic or even molecular level, these techniques offer a wealth of information at the level of supramolecular assembly.

Molecular modeling is one of the fields that can contribute to associating the observed global shapes with detailed structural, mechanical and dynamical characteristics at the atomic level, in an integrated vision of the cell.

However, molecular modeling only begins to take advantage of such important data mainly because the principal techniques, such as Molecular Dynamics (MD) or Normal Modes (NM), are bound by computing power dependent performance. As illustrated in fig. 1.1, the computing power of computers doubled every one year and a half since the early 1969’s and is predicted to continue in the same fashion for the foreseeable future, a prediction commonly known as the Moore law\textsuperscript{3}.

\textsuperscript{1}Term first proposed by Paul Wintrebert in 1931 \cite{11}.

\textsuperscript{2}The Metazoa subkingdom is formed by all animals whose bodies are made of differentiated cells arranged in tissues and organs.

\textsuperscript{3}The Moore law actually predicts that the number of transistors on computers doubles \cite{14}, which does not necessarily directly translate into computing power although it
1.2. GENERAL AIM OF THE STUDY

Although the size of protein filaments is very variable, one can roughly evaluate the average size of a RecA or an actin filament to $10^3$ to $10^4$ monomers, which in turn represents $10^5$ to $10^6$ residues, to finally end with a model of $10^6$ to $10^7$ atoms, reaching $10^8$ atoms with the water. In addition, as will be seen in the next chapter, the time scale of dynamic variations within polymeric suprastructures such as assembly/disassembly or structural reorganization, which are essential for many of the cell functions, extends beyond the second.

As a comparison, recent work [16] on the ribosome ($\cong 10^6$ atoms) reported MD simulation of a microsecond. As the cost of computing bigger systems roughly scales to $O(N\log N)$, where $N$ is the number of atoms, we should need 1000 times more power for a filament with $10^8$ atoms than for the ribosome. Extension of the MD run from a microsecond to a full second would further necessitate $10^6$ time more power for a full second finally totaling a $10^9$ ratio between the recent ribosome simulations and the planned one. We would thus need the power to double roughly 30 times ($2^{30} \cong 10^9$), which means we should wait forty five years before we have sufficient computer power, taking into account the Moore law.

As this rough calculation shows, it would be quite unreasonable to try

Figure 1.1: Development of computing power of the most powerful computers [15].

will be a sufficient approximation for the sake of this discussion.
to directly tackle the study of molecular assembly with today computing capacity using classical MD simulations. It appears therefore that there is a need for methods allowing the tools at the molecular level such as classical molecular dynamics and the data at the molecular assembly level to communicate.

The elaboration of such methods and associated tools was the object of my thesis.

1.3 Story of this work

When I started my thesis work, my main focus was on flexible docking, precisely on developing a docking method for DNA-protein complexes with a flexible representation of the DNA, based on the previous work of Pierre Poulain in the laboratory [17]. To develop such a method, I elaborated and implemented new tools in the PTools library (the PTools library and the improvements added during this thesis are presented in chapter 5, the early work on DNA-protein docking simulations in chapter 6). Eventually, these developments led me to propose a promising approach (chapter 7) but unfortunately, a very similar method [18] was published in the middle of its implementation.

Fortunately, the application part of the thesis was on the RecA filament, a nucleoproteic filament responsible for homologous recombination (chapter 3). This complex, formed by RecA polymerization on single-stranded DNA (ssDNA) resulting from a double strand break (DSB) lesion, promotes the search through the whole genome of a sequence identical to that of the ssDNA, then promotes the repair of the DSB via DNA strand exchange. In the nucleoprotein filament, the protein and the DNA share a common helical organization, which is thought to vary during the still unknown mechanism of sequence recognition and strand exchange. As we were stuck on the docking front, it appeared that many tools I had developed for the manipulation and simulation of the DNA helix were generic enough to be used on protein filaments. The results we obtained using this principle were very promising. For example, we found that the approach allows taking information obtained at the molecular or atomic level and projecting it to the level of protein filaments. We could also explore the relationship between local interaction geometries at the monomer-monomer level and global architecture of the resulting supra-assemblies. We decided to reorient the thesis towards this multiscale approach.

Chapter 8 will present the tools we developed (Heligeom suite) and chapter 9 will present the possibilities of Heligeom alone for the construction of various types of large macromolecular assemblies. Chapter 10 reports the coupling of Heligeom with docking simulations for the study of the RecA forms of self-association, which is the object of an article currently in the
1.3. *STORY OF THIS WORK*

submission process. Finally, chapter 11 will present a discussion on the possibilities opened by this method and some conclusion about this thesis work.

But first, the next chapter will present some information on our biological subject of interest, molecular assembly.
Chapter 2

Macromolecular assemblies

To exert their function in the cell, proteins as well as nucleic acids interact with other macromolecules. Interactions can be strong or weak, transient or obligate (section 2.3), fortuitous or productive. They can involve different types of macromolecules, for example within the huge molecular machineries responsible for replication, transcription, or signal transduction. They can also involve from two to thousands monomers of a same protein, notably to form the cytoskeleton responsible for vesicle transport and for the mechanical support of cell membranes. Macromolecular interactions are required to transmit information, to activate, control or inhibit biological events or to form quaternary superstructures with defined mechanical characteristics. Goodsell and Olson have quantified the occurrence of protein-protein complexes with strong enough interaction for the assembly to be detected by biophysical or biochemical methods [19]. Interestingly, these authors found that the majority of proteins in the cell are found in the form of oligomeric assemblies. Table 2.1 reproduces the data they extracted from the SwissProt databank for Escherichia coli.

Macromolecular association often contributes to exerting a particular function. An example of functional protein-DNA association is the binding of repressor proteins to selected DNA sequences in order to repress the transcription of genes. These proteins generally function as dimers or tetramers (fig. 2.1), which facilitates a large coverage of the double helix surface. Alternatively, some proteins mediate global structural transitions in the double helix. For example, spiral or helical oligomers of DnaA [20], DnaB [21, 22] and DnaC proteins [23] unwind and separate DNA strands to initiate and translocate replication forks.

Morphology is a crucial issue when the objective is the construction of mechanical support. In the cytoskeleton, the proteins that form microfilaments (actin), intermediate filaments (neurofilaments, keratin) or microtubules assemble to form long fibers. These fibers may in turn organize into higher order structures, either regular association, bundles or networks,
Table 2.1: Occurrence of oligomeric proteins in *E. coli* (data from SwissProt Databank, reproduced from the work of Goodsell and Olson, 2000 [19]).

<table>
<thead>
<tr>
<th>Oligomeric state</th>
<th>Number of homooligomers</th>
<th>Number of heterooligomers</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monomer</td>
<td>72</td>
<td></td>
<td>19.4</td>
</tr>
<tr>
<td>Dimer</td>
<td>115</td>
<td>27</td>
<td>38.2</td>
</tr>
<tr>
<td>Trimer</td>
<td>15</td>
<td>5</td>
<td>5.4</td>
</tr>
<tr>
<td>Tetramer</td>
<td>62</td>
<td>16</td>
<td>21.0</td>
</tr>
<tr>
<td>Pentamer</td>
<td>1</td>
<td>1</td>
<td>0.1</td>
</tr>
<tr>
<td>Hexamer</td>
<td>20</td>
<td>1</td>
<td>5.6</td>
</tr>
<tr>
<td>Heptamer</td>
<td>1</td>
<td>1</td>
<td>0.1</td>
</tr>
<tr>
<td>Octamer</td>
<td>3</td>
<td>6</td>
<td>2.4</td>
</tr>
<tr>
<td>Nonamer</td>
<td>0</td>
<td>0</td>
<td>0.0</td>
</tr>
<tr>
<td>Decamer</td>
<td>1</td>
<td>0</td>
<td>0.0</td>
</tr>
<tr>
<td>Undecamer</td>
<td>0</td>
<td>1</td>
<td>0.0</td>
</tr>
<tr>
<td>Dodecamer</td>
<td>4</td>
<td>2</td>
<td>1.6</td>
</tr>
<tr>
<td>Higher oligomers</td>
<td>8</td>
<td></td>
<td>2.2</td>
</tr>
<tr>
<td>Polymers</td>
<td>10</td>
<td></td>
<td>2.7</td>
</tr>
</tbody>
</table>

which increases their mechanical resistance. When proteins form filaments on DNA, like in the case of recombination, the shape of these filaments directly influences the structural characteristics of the bound DNA. The overall morphology and the global properties of the DNA molecule, itself a thin fiber, can also be regulated by the punctual association of proteins or protein assemblies, locally interrupting the persistence of its orientation (architectural proteins such as HU, IHF, TBP) or enabling its compression and storage (nucleosome). More generally, association/dissociation processes regulate the fiber length, modulation of the mode of association regulates the fiber architecture and these processes control the mechanical and dynamical fiber properties.

The three-dimensional structures of macromolecular assemblies have little been studied up to now due to difficulties in their experimental determination at the atomic level. The structures of complexes only occupy a small part of the Protein Data Bank. Filaments for example are too big to be studied by NMR and do not easily form crystals. Actin, like intermediate filaments, has only been crystallized in the form of monomers or dimers [24–26], tubulin as heterodimers of α, β tubulin or dimers of α, β dimers in interaction...
CHAPTER 2. MACROMOLECULAR ASSEMBLIES

Figure 2.1: Lac repressor binding to DNA. The lac repressor is formed by the association of two dimers, with each monomer being alternatively colored in dark green and light green. Binding of the lac repressor to the DNA in orange induced the formation of a large DNA loop together with local deformations characterized by the widening of consecutive grooves and the formation of bends. Representation by David Goodsell, Protein data bank, Molecule of the Month, http://www.rcsb.org/pdb/101/motm.do?momID=39

with drugs [27, 28]. Information on the fibers mainly comes from low resolution structure reconstruction based on electron or atomic force microscopy or from models built from the structure of monomers, which makes this information partial.

An important issue is to gain access to the shape variations of oligomeric assemblies: sixteen years of enduring efforts were necessary to solve the active form of the RecA filament after the structure of the inactive form was published [29, 30] (see chapter 3). In this context, theoretical approaches can offer complementary insights into the variability of structural assemblies. One important issue of my thesis work has been to develop flexible modelling methods specifically adapted to this exploration.

I will concentrate on two classes of macromolecular assemblies, the complexes between proteins or protein dimers and DNA and the oligomeric assemblies of protein monomers (as opposed to assemblies formed by a variety of proteins or nucleic acids such as the ribosome). Table 2.1 has shown that the homooligomeric forms of association are largely represented in the E. coli cell. Both assembly classes feature strongly interacting molecules, therefore factors that increase the binding strength will be of particular importance to determine the possible binding modes within these assemblies.

This chapter reviews the current knowledge on macromolecular interac-
2.1. STRENGTH OF ASSOCIATION

Cells present a highly crowded environment, where macromolecules are subject to numerous encounters. These encounters are principally diffusion driven, and many are non specific and do not result in the formation of complexes, with the partners quickly separating. Two types of driving forces can strengthen the association between macromolecules: long distance electrostatic interactions and short distance surface complementarity. The balance between these forces depends on the system and varies along the path of association.

Experimental characterization

The quantity that precisely measures the stability of a complex is the Gibbs free energy. This is the quantity of energy that is liberated during association performed at constant pressure and temperature. The free energy variation $\Delta G_{\text{binding}}$ during association can be decomposed into an enthalpic component $\Delta H$, which measures the internal energy of the system and its stability, and an entropic component $T \Delta S$ that increases with the number of conformational substates available to the system.

$$\Delta G_{\text{binding}} = \Delta H - T \Delta S$$

Enthalpic and entropic components of the activation energy of complex formation have been measured or calculated for the complex barnase-bastar, which presents a dominant electrostatic contribution, or for the lysozyme monoclonal antibodies HyHEL-5-HEL and HyHEL-10-HEL, which do not [31–33]. These studies showed a small entropic contribution in both cases. However, mutations of interface residues did not change the entropic contribution for barnase-bastar, while larger variations were observed for mutants of lysozyme antibodies. Entropic changes may result from different factors, such as desolvation (favorable contribution, see below) but also the decrease of the rotational and translational degrees of freedom together with restriction of side chain movement in the two macromolecular components (unfavorable contribution) [34].

The variation of the Gibbs free energy during complex formation can also be directly related to the kinetic constants of association $k_{\text{on}}$ and dissociation $k_{\text{off}}$, through the relation:
\[ \Delta G_{\text{binding}} = -RT \ln K_{eq} = -RT \ln \frac{k_{\text{on}}}{k_{\text{off}}} \]

The constants \( k_{\text{on}} \) and \( k_{\text{off}} \) can be determined experimentally by measuring the time evolution of the components and complex concentrations during complex formation (radio-labelling, spectroscopy, polarimetry, ...). The \( k_{\text{on}} \) constant is a good indicator of the importance of long distance electrostatics as a driving force for association. For systems highly dependent on long distance electrostatics, the association rate \( k_{\text{on}} \) exhibits variations with the ion concentration that can reach five orders of magnitude while for systems that little rely on electrostatics, the rate changes by less than one order [35, 36]. In fact, the \( k_{\text{on}} \) value itself is indicative of the electrostatic dependency. Camacho et al. [35] have shown that fast binding systems (\( k_{\text{on}} > 10^8 \text{ M}^{-1}\text{s}^{-1} \)) are strongly dependent on long distance electrostatics while slow binding systems (\( k_{\text{on}} < 10^7 \text{ M}^{-1}\text{s}^{-1} \)) are not.

The role of long distance electrostatics primarily concerns the phase of approach between the two entities, where it participates in orienting the two partners and accelerating their approach. Its effect is mostly kinetic [33]. Once the two components have encountered, either following electrostatic steering or as a result of diffusion followed by microcollisions, the complex will stabilize only if the macromolecules can optimize their short range interactions and their surface complementarity.

## 2.2 Stabilizing interactions

### Desolvation

Before they interact with their association partner, macromolecules are solvated by the surrounding water molecules. Polar and charged surface groups form hydrogen bonds or salt bridges with water molecules or ions, while the presence of hydrophobic groups restricts the movement of nearby water molecules. Upon formation of the complex, part of the solvent is released. Desolvation of charged and polar residues is energetically costly. However, the entropy of the solvent increases, which is particularly favorable when the solvated region was hydrophobic in the separate molecules. The hydrophobic effect has been proposed to be a major factor of complex stabilization.

### Non-bonded interactions

In the process of complex stabilization, new hydrogen bonds or salt bridges form between the residues of each partner situated at the interface, which compensates in part the desolvation of the charged and polar residues. When they remain partly solvated, these interactions do not contribute strongly to the complex stability. For example, previous exploration in the laboratory
of the driving forces for the association of specific protein-DNA complexes indicated that steric complementarity is sufficient to unambiguously distinguish a correctly assembled from a non-correctly assembled complex [17]. When they are buried in the interface and surrounded by hydrophobic interactions, interactions between polar and charged residues contribute much more strongly to the stability. The van der Waals interactions are important stabilizing interaction, provided that the interface atoms are densely packed. This requires high quality steric complementarity (see 2.3).

**Small molecules and cofactors**

Individual water molecules, ions or small molecules such as nucleotide triphosphate (NTP) may fully participate to the complex stabilization. In the example of the dimeric lambda repressor bound to its target DNA (fig. 2.2B), water molecules buried in the interface bridge interactions between polar or charged protein residues and DNA phosphates as well as DNA bases [38]. In the protein-protein complex represented in fig. 2.2C, a few water molecules are buried in the interface [39]. In addition, strongly bound water molecules commonly accumulate at the interface edges, where they may bridge interactions between polar or charged residues of each partner [39]. The interface between the TBP protein and its partner TATA-box is particularly hydrophobic, even at the sugar-binding locations (fig. 2.2A) [38]. In that case, strongly bound water molecules almost continuously line the DNA-binding region. Metal ions have also been reported in some structures, generally Mg$^{2+}$ or other divalent cation for protein-nucleic acid complexes. NTPs (either adenine triphosphate ATP or guanine triphosphate GTP) are other molecules of great interest that participate to interface strengthening. In addition to their high negative charge, NTP molecules store chemical energy in the form of a highly energetic bond between phosphate groups (see chapter 3 for a description of the ATP hydrolysis reaction). In cellular processes the chemical energy brought by NTP hydrolysis can be converted to mechanical energy, and this often involves allosteric mechanisms (described below). In some cases such as microtubules or the recombination filaments, NTPs bind at the interface between two consecutive protein monomers (tubulin for the
CHAPTER 2. MACROMOLECULAR ASSEMBLIES

microtubule, RecA or Rad51 for recombination) and may stabilize the association. We will come back in more detail to this aspect of protein-protein association.

2.3 Characteristics of macromolecular complexes

Beyond the process of association itself, what keeps macromolecules strongly bound together is the complementarity of their interfaces [40]. This implies steric as well as electrostatic complementarity, and involves properties of the interfaces that will be described in this section.

Buried Surface Area

The degree of surface complementarity is often characterized by the buried surface area (BSA). The BSA is defined as the sum of the solvent accessible surface area (ASA) of the two macromolecules in their free form minus the solvent accessible surface area of the complex,

$$BSA_{AB} = ASA_A + ASA_B - ASA_{AB}$$

where A and B are the molecular components of the complex AB.

In fig. 2.3, this represents the cumulated area of the shaded regions on each schematically represented protein (for certain purposes, it is more convenient to define the BSA per protein, which is the area of the shaded region of one protein, see chapter 10).

Figure 2.3: Definition of the buried surface area (BSA). The solvent accessible surface area is delimited by the center of a sphere of typical radius 1.4 Å rolling over the surface [41] (thin black line). The BSA is the portion of the accessible surface area that becomes buried due to the formation of the complex (grey regions).

In a review on macromolecular complexes, Janin and collaborators have reported the range of BSA values obtained for different sets of non-obligate protein-protein complexes [39, 42]. Non-obligate complexes are formed by proteins that independently fold then assemble to perform a particular task, contrarily to proteins that fold only in the presence of their partner. These complexes can be considered as transient, yet their lifetime can vary from less than a second (redox proteins, $K_d \approx 10^{-6}$M, with $K_d = 1/K_{eq}$) to days (barnase-bastar, $K_d \approx 10^{-14}$M). Similar ranges of lifetimes can be found for protein-nucleic acids complexes [39]. In both cases, even the short lifetimes remain superior to the association lifetime resulting from random collisions between macromolecules that do not play a biological role together. The authors noted that contacts that form between macromolecules densely packed in a crystal are representative of random contacts resulting from
non-productive collision. Therefore, they compared the BSA values of non-obligate complexes to those characteristic of non-specific crystal contacts.

The majority of protein-protein complexes in a set of 70 protein-protein complexes examined by Lo Conte et al [42], then Chakrabarti and Janin [43], had a BSA comprised between 1100 Å² and 2000 Å², with typical interface size of 1600 ± 400 Å². This range of values is referred to as standard-size values as opposed to the small-size interfaces below 1200 Å² and the large-size interfaces beyond 2000 Å². Most of the antibody-antigen complexes, as well as many enzyme-inhibitors complexes, exhibit standard-size interfaces. The typical standard interface contains 57 amino acids (28 for each partner) but 16 interfacial amino acid per protein are sufficient to make a stable interface of about 1200 Å² [39]. More recently, analysis of a benchmark of non-redundant protein-protein complexes by Hwang and collaborators revealed complexes with interface size as low as 810 Å² [44].

Complexes with small-size interface, which constitute 7% of the Hwang benchmark, mostly correspond to very short-lived complexes such as redox complexes involved in electron transfer (half-lives of redox proteins are lower than 1s) or ubiquitin-bound protein. Nevertheless, the size of these interfaces is generally greater than that of non-specific crystal contacts.

Complexes with large-size interface are more abundant in the Hwang benchmark where they represent 38%. Homodimers for example contain representative complexes of these type, and a set of homodimers assembled by Bahadur et al. [45] showed an average BSA value of 3900 Å² (standard deviation 2200 Å²). Other members of this class are found among the signal transducing complexes. Large-size interfaces are usually associated to surface remodelling (see below).

BSA values have also been calculated for interfaces between proteins and nucleic acids [38, 39]. The values are larger in average than for protein-protein complexes: 2530 ± 1210 Å² for protein-RNA and 3100 ± 1050 Å² for protein-DNA. The range of values is very extended, the smallest interfaces being found for protein-RNA complexes. Proteins often bind to DNA in the form of multimers, which creates multiple interfaces with very large total buried surfaces. Even single proteins can interact with DNA via multiple interfaces. For example, the complex between the integration host factor and its DNA target structure (fig. 2.7, PDB code 1ihf) is characterized by a BSA value of 5120 Å².

Interface topology

Interfaces are composed by residues that contact the partner of association in the complex. The topology of the interface has been shown to play a crucial role for the affinity. In proteins, the interface residues are not necessarily consecutive within the polypeptide chains, but can provide from distant regions in the sequence, even from different domains. These residues
are spatially grouped to form patches on the protein surface. For nucleic acids, interfaces are formed by one or several patches of 6 to 15 base pairs in contact with 15 to up to 75 amino acids of the partner protein \[17, 38\]. For a given value of the surface area buried by the association, the spatial organization of interface residues generally distinguishes interfaces characterizing good affinity complexes from crystal packing interfaces, typically composed of small, non contiguous contact patches \[39\]. The packing in the good affinity interfaces is high, with no internal cavity. The interface has been shown to divide into two regions, a core region where the atoms are completely buried, exceeding no more than 50% of the interface, surrounded by a rim region where the atoms are only partially buried. Several descriptors distinguish these two regions: the amino acid propensity (the core being enriched in aromatic and long aliphatic residues), the conservation in evolution (maximum conservation for the core residues), the effect of site-directed mutagenesis (stronger effect for core residues) \[39\]. Small and standard-size interfaces exhibit a unique interface patch. To the contrary, large-size interfaces can be formed by the union of separate patches, each of them displaying the core/rim organization (fig. 2.4).

**Figure 2.4: Interface topology.** (left panel) The interface between the G−α subunit of the heterotrimeric G-protein transducin (surface representation, PDB code 1GOT \[46\]) and the Gβ, Gγ subunits (tube representation) can be divided between two separate patches, respectively colored in blue and red on the Gα surface; (middle panel) the same interface is shown as a Voronoi model, with strongly bound water molecules represented as spheres that surround the two patches \[37\] (from figure 4 of reference \[39\]); (right panel) similar representation for RecA-RecA interface (\[30\]). One monomer is represented in ribbon (dark grey), and the other in surface mode (light grey), with two distinct patches represented in blue and red. The red patch corresponds to the interface with a mobile N-terminal helix colored in green. ATP is in purple.

**Surface remodeling**

Association with other macromolecules can induce structural modifications in protein main chains, resulting in surface remodeling. This process is generally involved in the formation of large-size interfaces. Notably, flexibility may be useful to assemble multiple-patched interfaces. In such cases, one patch may serve as an anchor to keep the partners in contact during adaptation of the rest of the protein and optimization of the whole contact surface. For example, in the transducin system shown in fig. 2.4A, the patch colored in red is situated on a flexible helical domain of the Gα subunit. This domain is unstructured in the absence of the partner protein and only folds when the three subunits associate. For the RecA proteins (fig. 2.4B), a standart-size
2.3. CHARACTERISTICS OF MACROMOLECULAR COMPLEXES

A binding patch on the surface of the protein central core is completed by a second patch, of almost identical size (938 Å² for the central core, 868 Å² for the second patch) [47], located on a terminal helix attached to the core domain by a very flexible linker. This same pattern is found in both known modes of RecA oligomeric association (see fig. 10.9 in chapter 10).

In addition to domain movements, surface remodeling frequently involves loop movements or loop refolding. In the protein-protein benchmark that gathers non-redundant complex structures for which the structure of isolated association partner is available [48], important structural differences between bound and unbound proteins forms mainly involve loops, sometimes accompanied by domain movements. These cases represent 40 complexes among the 176 complexes of the benchmark. Loop refolding can completely remodel the surface of the protein accessible to partner binding, as seen in fig. 2.5A when comparing the free form of actin (top) to the form bound to the DNase-I protein (bottom). The presence of flexible loop at the protein-protein interface does not always lead to the formation of large-size interfaces. During her Master project in the laboratory, Justine Houndekon has analyzed the contribution of flexible interfacial loops to the characteristics of protein-protein complexes [49]. She identified at least two complexes with standard to small-size interface where a flexible loop contributes to almost 100% of the interface (fig. 2.5B). More than half of the complexes with flexible loops in the Hwang benchmark present large-size interfaces, with BSA values per monomer comprised between 1000 and 1800 Å². Cases were identified where several loops contribute to very large interfaces (up to 3000 Å² per monomer) with moderate individual contribution.

Finally, the whole tridimensional structure of a protein can be implied in the conformational change that leads to interface reshaping. One striking case can be found within the G-Protein Coupled Receptors (GPCR) membrane protein family. The seven helices of these proteins, which cross the plasma membrane, undergo a concerted rotation when a ligand binds outside the cell. This modifies the opposite surface of the protein inside the cell in such a way that binding G-proteins to this surface becomes favorable [50, 51]. This allosteric process implies concerted torsions and translations of the seven helices, resulting in a wide opening of the GPCR accessible surface inside the cell.

Situations where association partners modify their conformation upon binding another partner, in order to obtain a perfect steric fit, have been identified as early as 1958 by Koshland [52] and referred to as induced-fit. Koshland attributed the conformational change to a response to the exposure of the partner.
CHAPTER 2. MACROMOLECULAR ASSEMBLIES

Figure 2.5: Surface remodeling: flexible loops. (A) cartoon representation of the unbound (top) and the bound (bottom) structures of actin (PDB codes 1IJJ and 1ATN, respectively), with the flexible interface represented as a transparent surface; (B) superposition of the bound (blue and pink) and free (grey and white) forms of the protein components of PDB structure 1AK4; a 14 residue flexible loop (orange, Cα-RMSD = 4.0 Å between the free and the bound forms) contributes to 100% to the small-sized protein-protein interface (after figure 6 of reference [49]).

tion of each partner to the field of the other partner during the process of association. An alternative interpretation, which originated from the Monod-Wyman-Changeux model of allostery in 1963, propose that the bound state of each partner exists among substates accessible to the molecule even before association takes place [53]. This conformational selection scheme builds upon the fact that molecules exist in solution as an ensemble where diversely populated conformational substates coexist [54]. In this scheme, conformational change results from the selection of substates of each association partner that best fit each other. While the latter is presently the favored interpretation [55], both processes probably take place during association. Fig. 2.6 summarizes the model proposed by Grunberg and collaborators based on their explorative work on the substate selection model [56].

A combined process probably takes place in the case of multi-patch, large-size protein-protein interfaces, as discussed above, or for protein-DNA complexes that bury a large surface area. The helical character of the DNA helix sometimes makes the protein wrap around the DNA (fig. 2.7). In such cases, formation of a complex between the DNA and a protein already in the
2.3. **CHARACTERISTICS OF MACROMOLECULAR COMPLEXES**

Figure 2.6: Proposed scheme for the thermodynamics of flexible protein-protein association. Protein-protein association is governed by diffusion, selection of adequate substates and induced fit. \( R_f \) and \( L_f \) represent the ensemble of receptor and ligand substates in a free state. \( R_f^* \) and \( L_f^* \) are the bound states of the receptor and the ligand (recognition states). The height and lengths of the bars are only indicative and do not reflect real proportions. The middle and bottom panels indicate the thermodynamic and kinetic factors involved in the different steps of association (middle) and the resulting energy profile (bottom) (after reference [56]).

bound form would be sterically difficult, hence the necessity of coordinated steps of flexible association such as loose association followed by protein rearrangement.

**Control of association**

In the GPCR case, the conformational change is induced by the binding of a small molecule, *e.g.* hormone or neurotransmitter, to a specific site in the extracellular part of the protein. In that way, the ligand binding indirectly controls G-protein association. Other examples of controlled association mediated by conformational changes are commonly found in the process of formation or dissociation of protein fibers such as actin, tubulin or the nucleofilaments of homologous recombination (RecA or Rad51 proteins). Each monomer of these filaments binds a nucleotide triphosphate (NTP) as a cofactor. Hydrolysis of the NTP cofactor results in domain movements or in loop refolding, with consequences on the monomer/monomer contact surface that trigger association or dissociation events. For example, in the case of actin filaments, it has been suggested that the domain rotation that accompanies phosphate release and that closes an internal cleft in actin monomers [57] liberates the access for depolymerizing proteins like cofilin, ADF or destrin binding to actin [58,59]. In microtubules, structural change of an interfacial loop following the hydrolysis of the GTP cofactor has been proposed to modify the tubulin/tubulin geometry of association in such a way that it becomes incompatible with the maintenance of the microtubule quaternary structure [60]. While the mechanism of depolymerization remains to be determined for RecA nucleofilaments, it has been established that depolymerization in the presence of double stranded DNA requires the hydrolysis of the ATP cofactor and that the mode of association between RecA/ATP and RecA/ADP monomers is modified upon ATP hydrolysis (chapter 3).

Finally, surface remodeling can be directly controlled by the association of a binding partner. The DNase I-binding loop of actin is situated in a region of the protein that is poorly structured [58] and the loop modifies its
structure upon binding the DNase I protein (fig. 2.5A).

2.4 Protein-DNA complexes

As shown in Fig. 2.7, proteins associate to DNA in a variety of ways. Commonly found binding patterns have been classified by Luscombe and collaborators [61]. Several of these patterns feature a α-helix binding the DNA major groove where it probes the DNA sequence. In this category are found the helix-turn-helix motif (see for example complex 1o3q in Fig. 2.7), the helix-loop-helix motif, the zinc finger motif, where Zn$^{2+}$ ions forms four bounds with cystein or histidine residues or the leucine zipper. α-helices within High-Mobility-Group (HMG) domains (groups of three L-shaped helices, see Fig. 2.7, complex 2lef) can also interact with the DNA minor groove. Alternatively, the binding domain can be composed of β-sheets. In the 1ytb complex (Fig. 2.7), a β-sheet extensively interacts with the minor groove of the DNA. Finally, the binding motif is not always rigid, it can be formed for example by long, flexible loops such as the long loop of the integration host factor in 1ihf (Fig. 2.7).

Factors that favor protein-DNA association

As noticed earlier, protein-DNA interfaces generally bury more surface area than protein-protein complexes. Moreover, the DNA surface is particularly contrasted, with an alternation of deep grooves and protruding phosphodiester backbones. These two elements indicate that the steric complementarity between the protein and the DNA surfaces may play a particularly important role in the strength of association. Indeed, Poulain and collaborators [17] have observed during theoretical docking studies$^2$ that the correct geometry of association of a large range of specific protein-DNA complexes could be unambiguously identified on the sole basis of steric complementarity when the electrostatic interactions were artificially set to zero.

Electrostatics can however prove decisive for non-specific protein-DNA complexes. The small DNase I protein binds any DNA sequence and induces a cleavage of the DNA backbone. Guérout and collaborators have demonstrated the importance of two calcium and two magnesium ions to assist bovine pancreatic DNase I in binding DNA [62]. These ions strongly bind to four sites on the protein, three of them being situated on flexible loops distant from the DNA binding site. In this case, association and dissociation are directly regulated by the presence of the cations, the control that they exert is collective and it exclusively concerns the electrostatic fit.

---

$^2$The principles of macromolecular docking methods, which are developed to predicting the 3D structure of macromolecular complexes starting from the structure of their
2.4. PROTEIN-DNA COMPLEXES

Figure 2.7: Examples of protein-DNA complexes. The proteins, which present various DNA-binding motifs (see text), are shown in ribbon representation (orange); from left to right, lymphoid enhancer-binding factor (2lef), catabolite gene activator (1o3q), TATA-box-binding protein (1ytb), integration host factor (1ihf). The DNA is shown in surface representation (green).
CHAPTER 2. MACROMOLECULAR ASSEMBLIES

Mechanism of protein-DNA recognition: role of DNA flexibility

In many cases, the protein binds to specific DNA sequences. It has been established that in those cases, the fidelity of recognition results from the combination of two fundamentally different mechanisms [63, 64]. One is the direct or base readout mechanism. In this mode of recognition, specific amino acids are put in registration with DNA functional groups using a rigid support such as the $\alpha$-helices that are present in several major groove-binding motifs. This fingertip reading scheme involves the formation of hydrogen bonds or water-mediated hydrogen bonds and mainly takes place in the DNA major groove, where patterns of hydrogen bond donors and acceptors associated to a given sequence can unambiguously be distinguished. The second mechanism, either called indirect or shape readout, takes advantage of spontaneous or protein-induced sequence specific shape modulation. This can go from minor groove narrowing, which exacerbates the electrostatic field near the groove [64], to the creation of large deformations such as bends (Fig. 2.7, 1ihf) or local transitions to different substates of the double-helix [65,66] (Fig. 2.7, 1ytb). The protein-DNA interaction is then optimized by a combination of modifications of physical (e.g. electrostatic field) and mechanical (torsion or bending rigidity, ...) properties of the double helix, as well as modifications of its shape. For example in the complex between the TATA-Box and the TBP protein shown in Fig. 2.7 (1ytb), the local transition of the DNA region contacted by the protein towards the TA-DNA form, an unwound and stretch form with A-form character [65], results in a wide opening of the minor groove and offers a large and flat surface for protein binding. In the case of indirect readout, sequence discrimination results from the sequence-dependent energetic cost of the shape modulation. Both mechanisms generally contribute to some extent to the specificity of association and can be combined to reinforce the selectivity [66]. For example, optimal fit of the protein on a tailored DNA surface may position a readout motif in the exact position for directly probing nucleotide functional groups.

Prediction of the geometry of protein-DNA association

The characteristics of protein-DNA complexes, with often large size and contrasted interfaces, make them easier targets for theoretical structure prediction than protein-protein complexes. Indeed, the binding site on the protein side can be predicted unambiguously in numerous cases just by considering the steric adjustment, which does not even need to be strictly accurate. Poulain et al. [17] have shown that the protein interface can be predicted even when the DNA structure in the complex differs from canonical B-DNA, components, will be presented in chapter 6.
as far as the structural deviation is not too much important (less than 20° curvature).

When the DNA distortion is important, it is necessary to explicitly predict the induced DNA deformation simultaneously to predicting the relative position/orientation of the two macromolecules. Solutions for this problem have been first proposed by the group of Bonvin, based on the information-driven program HADDOCK [67]. In a first step, directions for internal DNA deformation are extracted from semi-rigid data-driven docking simulations starting from B-DNA. The detected deformations are then amplified using the 3D-DART DNA modelling server [68] and the distorted structures are used as new starting point for HADDOCK docking simulations. Another algorithm has recently been proposed by Banitt and Wolfson [18]. In this method, the protein surface is screened by rigid DNA fragments and a curved DNA structure is reconstructed on the identified consensus surface patches. This method only considers curvature for possible DNA deformation and does not support unwinding or stretching deformation nor possible local transitions to different substates but it is efficient as a first approximation treatment of shape deformation. In chapters 6 and 7, I will present the strategies I have developed to address the variety of possible DNA deformations that can be encountered in protein-DNA complexes.

In addition to detecting the DNA binding site on the protein and predicting the deformation of the bound DNA, complete prediction of the association geometry requires the phasing of the DNA sequence with respect to the protein. For van Dijk and collaborators [67], this question is inseparable from the detection of the overall geometry of association since the assembly is driven by experimental information on protein and DNA residues involved in the complex formation. Ab initio prediction is more tricky. In case the sequence is known, threading the sequence on a pre-deformed DNA structure should enable both the detection of sequence regions that minimize the deformation energy and sequence regions prone to direct recognition by DNA-binding motifs. One may also resort to the use of the ADAPT methodology [69,70], which was exactly developed to optimize the sequence of oligonucleotides given their structure.

2.5 Oligomers and polymers

At larger length scales, protein can self-associate as closed forms, such as rings or polyhedral envelopes (capsids), or as open forms like spiral filaments [19]. Polyhedral envelops enclose the genetic material of viruses. Rings are commonly found in processes involving processive DNA manipulation such as replication, repair or recombination [71]. As discussed earlier, filaments and filament assemblies (fibers) are the essential component of the cytoskeleton, responsible for the cell shape as well as for its mechanical
properties and also used as support for vesicle transport. Filaments can be used also to process DNA, in which case there is a coupling between the shape of the protein filament and the DNA form (chapter 3).

Due to their limited size, ring-shaped assemblies are often accessible to atomic structure determination. All forms can be observed using low resolution techniques such as EM, AFM or SAS. When combined with advanced image treatment, EM can determine the structure of regular forms at sub-nanometer resolutions.

**Dynamics of filament association/dissociation**

Open forms resulting from protein self-association need to be regulated to avoid ever-growing filaments. More fundamentally, the dynamic interplay between growth and dissociation is an essential element of the filament function in the cytoskeleton: it enables the cytoskeleton to exert mechanical forces on the cell membrane that are directly responsible for the cell shape; it is fundamental for the microtubule role in chromosomal segregation during mitosis. Particularly, the microtubule undergoes alternation of rapid growth and sudden disassembly events (known as “catastrophe”) that are inherent to its function. Antitumoral drugs such as taxol, colchicine or vinblastine that disregulate the microtubule assembly/disassembly cycle by artificially inhibiting its formation (colchicine, vinblastine) or overstabilizing the assembled microtubule (taxol), provoke an arrest of the cell division.

Several factors contribute to this regulation.

- **Accessory proteins** can reinforce the stability of filaments or promote their dissociation. This is the case of microtubule-binding proteins (MAPs), which initiate (γ-tubulin) or stabilize (Tau, MAP2, MAP4) the microtubule association, or promote (stathmin) its dissociation [60].

  The growth, stability and dissociation of actin filaments is also regulated by proteins such as Arp2/3 (initiation), capping proteins (growth termination), ADF/cofilin (promote the dissociation) or profilin (restore the growth) [72].

- **NTP cofactors** are another central factor that regulates filament self-assembly. As discussed earlier (section 2.3), the chemical energy stored by NTP molecules can be transmitted to the protein as a mechanical energy, enabling the transition between internal states of the protein and modifying the interface accessible to association partners. These partners whose association is controlled by the state of NTP hydrolysis can be other monomers, but they can as well be the accessory proteins described above. For example, cofilin associates along the sides of actin filaments or to actin monomers where ATP has been hydrolyzed [72].
2.5.  OLIGOMERS AND POLYMERS

Figure 2.8: Actin network in a keratocyte. (left) Electron micrograph of a keratocyte; detail of region delimited by a white square is shown in the right panel; (right) three zones of actin filament organization are shown; the schematic diagram below indicates the location of key proteins and the curves indicate actin filament assembly (red, actin subunits per unit time) and disassembly (blue). After figure 1 of [72].

In certain cases, the NTP molecule is located at the interface between two monomers. The presence of NTP or the products of its hydrolysis then chemically modifies the interface properties, notably its electrostatic properties. For example, in the protofilaments constitutive of the microtubules, the interface between the \( \beta \)-subunit of an \( \alpha,\beta \)-tubulin dimeric unit and the \( \alpha \)-subunit of the neighboring dimer hosts a GTP molecule [73], whose hydrolysis destabilizes the microtubule construction. In the active form of recombination filaments, an ATP molecule is sandwiched between each two consecutive monomers (fig. 2.4, right) [30]. Its presence is necessary for the ATPase activity as well as the recombinase activity, and its absence or its hydrolysis destabilizes the interface, leading to an alternative assembly mode for the filament [29]. The exact way these interfacial cofactors influence the stability of the assembly mode remains to be established. For microtubules, it has been proposed that hydrolysis of the GTP may lead to a preferentially curved form of the protofilament, analogous to the colchicine or stathmin-bound crystal form [28, 73]. This hypothesis has been questioned by the observation of lateral loop conformational change consecutive to GTP hydrolysis and by a recent study based on theoretical calculations and EM observations [74], which indicates that GTP hydrolysis would primarily affect the lateral contacts between protomers. For RecA, the question is still awaiting further studies.

In these cases where the NTP molecule is situated at the interface, some residues from the partner monomer participate to the hydrolysis reaction. This leaves the possibility of a coordinated interplay between assembly and hydrolysis\(^3\) that may fully take part in the filament growth control, albeit in a process that still has to be elucidated.

- **Super structuralization** of the fiber network has also been associated to the dynamics of association/dissociation. Reymann et al. [76] reported that the structural organization of the network of actin fibers influences the dynamics of polymerization/denpolymerization.

---

\(^3\)In the case of actin, it has been proposed that the conformational change induced by longitudinal monomer binding would position a key glutamine amino-acid close to the ATP in a way that may enhance the rate of ATP hydrolysis [75].
The fibers can be organized as branched entangled and dense networks, as antiparallel contractile bundles linked to the myosin motor, or as parallel bundles (fig. 2.8). The branch networks are associated to strong dynamics of actin polymerization/depolymerization. Microtubules are formed by the lateral assembly of between 11 and 16 quasi-straight protofilaments, formed by longitudinally assembled α, β dimers, which results in a tubular geometry. The dynamic turnover of microtubule association/dissociation varies by a factor of 4-100 depending on the phase of the cell cycle [77] (fig. 2.9). This change is associated with the formation of different spatial and temporal organization of the microtubules, such as spindle microtubules during mitosis.

Finally, the recombination filaments also have been shown to assemble in dynamic superstructures such as the RecA bundles recently observed by Lesterlin and collaborators [13] during the initiation phase of recombination (chapter 3, fig. 3.11). No accessory protein has been associated to this organization, yet ATP hydrolysis is required for the bundles to form.

Figure 2.9: Changes in the microtubule network morphology during the cell cycle. Microtubules are in red in metaphase and anaphase and in green in telophase and interphase; the chromosomal DNA is in blue. After Figure 2 of [77].

Beyond their apparent link with filament dynamics, variations in superstructure probably reflect varying states of the protein organization within the filaments, modulated by the presence of bound accessory proteins or with the state of ATP hydrolysis.

Multiple modes of association and filament morphology

As noted by Goodsell and Olson [19] and later theoretically justified by André and collaborators [78], the basal state of filamentous cellular structures such as protein filaments but also nucleic acids consists in a regular helical or cyclic organization. It is also under this basal state that the corresponding structures are more easily elucidated, based on crystal information on the symmetry state, on crystal structures of dimeric complexes or EM observations. Nevertheless, as discussed in the previous section, filament are also highly dynamic and subject to a high degree of variability. These properties are essential to their function, yet they are difficult to apprehend in a structural point of view.

[^4]: See chapter 8 for a description of underlying screw transformation.
X-ray or EM observations of DNA processing proteins such as the T7 gp4 helicase, the replication protein DnaB or the recombination proteins RecA, RadA, Rad51 or Dmc1 have revealed several forms for self-assembly, ranging from rings to helices with various morphologies. For example DnaB has been crystallized both as a helical ring [79] and as a spiral form [22], proposed to correspond to a translocation state of this helicase. The large diversity of the modes of recombinase self-organization will be detailed in chapter 10 (see for example fig. 3.8 of that chapter).

In these cases, the observed assembly forms were homogeneous, which means that the geometry was regular along the whole length of the oligomer. This is not always the case and non homogeneous oligomeric forms have been identified for the T7 gp4 and the DnaB helicases, both at low resolution [80, 81] and in crystal structures [82, 83]. In both cases, the hexamers could be observed as 6-fold symmetry rings with homogeneous modes of association, or as trimers of dimers (3-fold symmetry) combining two binding modes. The two symmetry states, either 6-fold or 3-fold, have been found to coexist in samples of DnaB hexamers analyzed by EM [80](Fig. 2.10. The cyclic structure of DnaB is a double-layered ring, with an upper ring containing the six C-terminal domains and a lower ring formed by the six N-terminal domains (Fig. 2.10). The two binding modes differ by the contacts between the upper and lower rings (stars in Fig. 2.10), with each N-terminal domain contacting either the C-terminal domain of its own monomer or that of the following monomer (arrows in Fig. 2.10). Heterogeneous structures have been proposed to help breaking the helicase ring when clamping the processed DNA. They have also been proposed to play a role as intermediates in the catalytic pathway leading to helicase unwinding and processive translocation [82, 83], together with the recently detected spiral assembly mode of DnaB [22]. Moreover, the simultaneous presence of different binding modes within the T7 gp4 helicase has been attributed to different states of hydrolysis of the nucleotide cofactor [82].

**Figure 2.10: Two assembly modes of DnaB cyclic hexamers.** Analysis of EM micrographs by Yang et al. [80] has revealed two symmetry forms for the cyclic self-assembly of DnaB hexamers in the presence of the AMP-PNP cofactor (a non-hydrolyzable ATP homolog) : a three-fold symmetry (a,c,e) and a six-fold symmetry (b,d,f). (a,b) and (e,f) are seen along the axis and are related by a 180°rotation; (c,d) represent a view perpendicular to the axis. The six monomers are labelled from 1 to 6, the helicase domain is labelled with a H and the N-terminal domain with a N; reproduced from [80], figure 7.

Helicases can therefore combine alternative modes of association without loosing their ring-shaped geometry, at the unique topology cost of a change in symmetry characteristics. Coexistence of alternative modes of
binding has also been reported in spiral oligomers (chapter 3, fig. 3.14), and it can reasonably be supposed that alternation of binding modes occurs in the course of their processive function. This may have implications in the filament morphology and the overall fiber organization.

2.6 Towards higher level superstructures

Oligomers tend to self-assemble as regular helical or cyclic structures. The presence at a given time of NTP cofactors in several hydrolysis states along the filament or the binding of auxiliary proteins can disrupt this regularity and introduce defects in the structure, with dynamic evolution. These irregularities are necessary to control the association/dissociation process in spiral filaments, they may also be necessary to perform processive tasks such as DNA unwinding or DNA strand exchange. They can result from internal structural variations of the monomers and subsequent modification of the surface accessible for binding a partner, like in actin fibers, they can also result from alterations of the binding surface or from association via completely different binding modes, like in recombinases. However, due to the difficulty of experimentally characterizing irregular and dynamically changing oligomers, there are presently few documented examples of these variations, which makes them poorly characterized. The principal observation here is that the quality of the interface and the strength of association locally associated to a given binding mode may play a role in determining the architecture of the whole oligomer.

The oligomers in turn participate in higher level structures, examples of which are shown in fig. 2.8 of this chapter or fig. 3.11 of chapter 3. These higher level structures are diverse, with distinct properties in terms for example of association/dissociation turnover, and they are the hallmarks of specific functions. In a 2007 review, Norris and collaborators [84] described and classified what they call “hyperstructures”, defined as the intermediate level of organization between the macromolecule and the cell. These hyperstructures are generally defined as heteromeric protein associations, yet several hyperstructure classes as defined by Norris and coll. are centered on oligomeric structures, for example the cytoskeletal hyperstructures or the cell cycle hyperstructures responsible for DNA replication, sequestration of newly replicated origins, segregation, compaction, and division.

Experimental studies of these defined and specialized hyperstructures are rapidly progressing, with the development of new techniques, apparatus and protocols adapted to their size and dynamics and dedicated to their study within the cells [13] or in vitro [76]. There is a growing need to understanding the physical rules governing these assemblies, as they are the key towards an integrated vision of the cell functioning.

Molecular modeling methods have proven their efficiency to fill spatial or
dynamic gaps in the experimental study of biomacromolecules. Expanding these methods to studying cellular super or hyperstructures should help rationalize the new observations. This is inherently a multiscale problem that requires considering the detail of interfaces together with large scale protein superstructure organisation. A great part of my work presented in chapters 8, 10 has been devoted to developing computational tools and modeling strategies to this aim.
Chapter 3

Homologous Recombination

As described in the previous chapter, many important biological processes require the formation of huge protein or protein/DNA assemblies, which organize the constitutive macromolecules in such a way that specific mechanisms can take place. In order to understand how such systems function, it is necessary to consider both the architectural organization (at the macromolecular level) and the local interactions (at the atomic level). In the case of homologous recombination (HR), the process takes place within long helical filaments made of recombinase proteins assembled on single-stranded DNA (ssDNA), formed at sites of DNA damage. These filaments can reach several kilobases of DNA. They can identify and incorporate regions of the genomic DNA (dsDNA) that are homologous to the constitutive ssDNA and they promote strand exchange between the three DNA strands. This means that the complementary strand of the dsDNA switches its Watson-Crick partner between the two homologous strands, giving rise to a new heterologous dsDNA bound to the recombinase filament (see Fig.3.1).

The mechanism of this reaction has been investigated for more than two decades, leading to fundamental advances but still without any clue to the detailed mechanism up to now. How sequence homology is recognized, how strands are exchanged and how the whole genome can be scanned in minutes remain completely open questions. The principal problems encountered when studying the reaction are the size of the system, its dynamics and the processivity of the reaction in time and space, which makes intermediate species very difficult to isolate. Indeed, there is increased evidence of a strong coupling between the geometry and dynamics of each part of the system (DNA and protein) and the local creation and breaking of hydrogen bonds that characterize base pairing exchange. This coupling must be taken into account for the problem to be finally solved.

The next chapters of this thesis will present the tools I have developed to cope with the large scale part of this multi-scale problem, and their use to investigate the mechanism of homologous recombination. The present
3.1 Homologous recombination

Double-strand breaks are very dangerous to the cell since they can lead to genome rearrangement. They cannot be repaired by most of the repair processes, which use sequence information provided by the opposite strand to reconstitute the damaged strand. There are two mechanisms to repair double-strand breaks, non-homologous or microhomology-mediated end joining, and homologous recombination. End joining processes lack fidelity but they can take place even when no copy of the sequence is available (e.g., before replication).

When both DNA strands break in a genome, the 3’-ends of each strand are partially digested by nucleases and, as mentioned above, recombinase proteins polymerize on the remaining single-stranded region, forming a nucleoprotein filament (NpF) [85–90]. NpFs are active for ATP hydrolysis (recombinases are ATPase) and induce the SOS response which consists of a cascade of protein expression events [91]. In E. coli, the SOS response starts with the auto-cleavage of protein LexA after its binding to a NpF. NpFs are also active for recombination, including both sequence recognition of the
template dsDNA within a sister chromatid or a homologous chromosome - depending on the species - and strand exchange.

This whole process is common to organisms from all realms of life, from prokaryotes (where RecA is the recombinase) to yeast (RadA) and eukaryotes (Rad51, Dmc1). In eukaryotes, Dmc1 is specific to genetic recombination that takes place during meiosis. There, double-strand breaks are deliberately produced to induce chromosomal crossing-over and, further, genetic diversity. In prokaryotes, homologous recombination is also involved in horizontal gene transfer, where bacteria of similar species exchange genetic material during conjugation (through direct cell contact) or transduction (via a virus), again producing diversity.

RecA-induced HR has been widely studied and most available data on HR has been obtained using this system. In higher organisms, the recombinase proteins RadA or Rad51 require the assistance of accessory proteins to reach the same level of activity as RecA [92, 93]. In in vitro studies performed without these accessory proteins, the rate of HR reaches no more than 10% of the rate observed with RecA. Very recent findings strongly suggest that among all accessory proteins that take part in HR in yeast or eukaryotes, a unique dimeric protein, Swi5-Sfr1, is capable of restoring the activity to the RecA level [94–96]. However, it is too early to conclude on the possible existence of a common principle for this class of proteins, which would be promoted by Swi5-Sfr1 in higher organisms. I will therefore limit the following description to E. coli and RecA.

### 3.2 The RecA nucleoprotein filament

NpFs have been observed and characterized at various levels of resolution using atomic force spectroscopy (AFM), electron microscopy (EM) and X-ray crystallography. The filament which is active for recombination in vivo is formed in the presence of ATP and it hydrolyzes ATP at an important rate of about 30 min$^{-1}$ per RecA monomer [97]. The active form of RecA is therefore a polymer with both ATP and ADP bound and with a dynamic turnover of these nucleotides. However, since filaments made with non-hydrolyzable ATP analogs are able to induce DNA sequence recognition and strand exchange in vitro while filaments made with ADP are not (see below), the term “active” is usually applied to filaments with homogeneously bound ATP, or rather with non-hydrolyzable ATP analogs. Examples of these ATP analogs are ATP$\gamma$S (slightly hydrolyzable), AMP-PNP or ADP-AlF$_4$ (Fig. 3.2).

**RecA-ADP and RecA-ATP filament forms.**

The formation of the active filament requires the presence of ATP or one of its non-hydrolyzable analogs and DNA. Formation on ssDNA is more favorable, but the filament can also form on dsDNA. RecA monomers with
3.2. THE RECA NUCLEOPROTEIN FILAMENT

Figure 3.2: ATP and its homologs. (left) Chemical formula of (a) adenosine-5'-diphosphate (ADP), (b) adenosine-5'-triphosphate (ATP), (c) phosphotheiophosphoric acid-adenylate ester (ATPγS), (d) adenosine-5'-diphosphate tetrafluoroaluminate (ADP.AlF₄⁻), (e) phosphoaminophosphonic acid-adenosine monophosphate (AMP-PNP); (right) the ATP hydrolysis reaction; exothermicity results from the release of interaction between negative charges and internal resonance within the Pi group.

bound ADP or even with no cofactor also form filaments on ssDNA. These filaments differ from the so-called active filaments in terms of their helical characteristics: while both forms are right-handed and present 6 or slightly more monomers per turn, RecA-ADP filaments display lower pitch values (generally reported around 70-80 Å) than filaments built on ATP analogs (around 95 Å). For this reason, the RecA-ATP form is also called the “extended” form and the RecA-ADP form the “compressed” or “inactive” form. It can be noted that, as Yu and Egelman reported from the analysis of EM data, the range of pitch values characterizing each of the ATP and the ADP forms extends over up to 50 Å (from 80 to 130 Å for the ATP-form) [98]. Filaments of the two forms with overlapping pitch values are represented in Fig. 3.3, b and c.

Figure 3.3: Pitch variations of RecA-ATP and RecA-ADP filaments. Representative structures of the so-called compressed (a, b) and extended (c, d) filament forms of RecA are represented by their surface envelope. (a) Reconstruction from the 2REB PDB structure [29]; (b, c, d) reconstruction from electron microscopy. (e) The two filament forms can be distinguished by the orientation of the lobe (arrow) formed by the C-terminal domain. After Figure 4 of reference [99].

The rather small differences in helical characteristics reflect more pronounced differences in the local organization of monomers within the filament. EM observation distinguish the ADP and ATP forms by the orientation of a pending lobe positioned at the entrance of the filament groove and formed by the C-terminal (Cter) domain (Fig. 3.3e). The orientation of the two forms varies by 35°. It was long considered that the Cter domain rotated as a consequence of ATP hydrolysis independently from the rest of the protein. However, it is now established that not only the Cter domain, but the whole monomer presents a 35° orientation difference with respect to the axis in the two filament forms. [30, 100]. Chapters 9 and 10 will present a more detailed study of the consequences of this rotation and of the resulting interface modification.

Filament shape and interactions.
CHAPTER 3. HOMOLOGOUS RECOMBINATION

Figure 3.4: Model of LexA binding to RecA. The model was built based on evolutionary trace analysis, which identified conserved residues (red) associated to LexA (blue) binding. After Figure 8 of reference [101].

The overall shape that results from the monomer organization around the filament influences the binding of partner macromolecules such as DNA (see below) or the LexA protein. In the absence of DNA damage, LexA represses the SOS response genes coding for DNA repair proteins. Active RecA filaments bind LexA and induce its autocleavage, which liberates the expression of these proteins [102]. It has been shown that the LexA binding region in NpFs is the filament groove, where LexA is deeply inserted (Fig. 3.4) and interacts with seven consecutive RecA monomers according to recent analysis [101,103]. The overall organization of the NpFs therefore determines not only the particular amino acids that will be accessible in the groove for lexA binding, but also the volume and surface properties that will be necessary to fit LexA shape. In chapter 8, I will present the tool I have developed to measure the accessibility of filament grooves.

Crystal structures.

The crystal structure of the RecA-ADP filament form was solved in 1992 by Story and collaborators [29,104]. More exactly, two structures were published, one obtained without any cofactor at a resolution of 2.3 Å (PDB code 2REB) and one obtained with ADP as a cofactor at 2.7 Å resolution (PDB code 1REA). The two geometries are very similar and the Ca-RMSD (root mean squared deviation) between the two entries is 0.3 Å. Since the 1REA PDB file only contains Ca atoms, structure 2REB is generally taken.
as representative of the RecA-ADP form (see Fig. 3.5, left). The structure was solved with a $P6_1$ symmetry space group, which means that the helix presents exactly 6 monomers per turn. The pitch value is 82.7 Å. Despite great effort, no DNA-bound form of the filament could be obtained under the conditions used to obtain the RecA-ADP form.

Figure 3.5: Crystal structures of the RecA filament. The crystal structures of the RecA-ADP (left, PDB ID 2REB) [29] and RecA-ATP (right, PDB ID 3CMW) [30] filaments are shown in surface representation. The monomers are alternatively represented in light and dark shades. The pitch value and the number of monomers per turn are indicated for each structure.

The crystal structure of RecA-(ADP-AlF$_4$Mg)-DNA was elucidated only sixteen years later in 2008, using an ingenious protocol in which five to six RecA monomers, linked to each other by covalently attached segments, were simultaneously expressed. The two terminal monomers of this oligomer had their monomer-monomer binding region deleted in order to avoid the formation of long oligomers during the crystallization process. Using this procedure, it was possible to the Pavletich group to obtain crystals of RecA-(ADP-AlF$_4$Mg) helices bound to one (PDB code 3CMW) or two (3CMX) DNA strands and to solve the crystal structures at resolutions of 2.8 Å (3CMW) and 3.4 Å (3CMX) [30] (Fig. 3.5, right). This work, which was a real “tour de force”, illustrates the difficulty to obtain structures of filamentous protein assemblies at the atomic level. Indeed, no such structure is available for actin, microtubule or intermediate filaments and the structural information on these systems, when available, comes from a combination of
atomic resolution structures at the monomer or the dimer levels and lower resolution EM data. In the case of the HR filaments, a crystal structure with bound DNA was a mandatory step towards unravelling the HR mechanism. This explains why long standing effort have been focussed on obtaining that crystal form.

**Diversity of recombinase assemblies.**

In addition to the well characterized ATP and ADP-forms of the filament, RecA has been found to self assemble following other geometric motifs such as dimers or hexameric rings, notably when no DNA is bound [105,106](Fig. 3.6).

**Figure 3.6: Atomic force microscopy images of hexameric RecA rings.** After Figure 1B of reference [106].

Similar polymorphic character exists with RecA homologs. All known recombinase proteins can polymerize on DNA under “compressed” or “extended” filament forms depending on the cofactor, ADP or ATP (Fig. 3.7). These NpFs present helical characteristics similar to those of RecA-ATP and RecA-ADP filaments [107].

**Figure 3.7: Filament forms of recombinases from higher organisms.** These electron microscopy reconstruction images subsequently represent extended and compressed forms of the RadA (left) and Rad51 (right) proteins. Like in RecA, a terminal domain (here, the N-terminal domain) appearing as a lobe shows different orientations in both forms (arrow). After Figure 4 of reference [108].

Like RecA, recombinases from higher organisms also form rings but these rings are octameric or heptameric instead of hexameric for RecA [108–110](Fig. 3.8b). Interestingly, Dmc1 as well as RadA have been shown to polymerize on DNA in the form of stacked octameric rings as well as right-handed helical filaments (Fig. 3.8a). Other motifs have been reported, such as RadA left-handed helices obtained at low pH in the presence of DNA (EM observation and x-ray structures) or overwound helices with three monomers per turn and no bound DNA [111](Fig. 3.8c). All these motifs were observed via a large panel of methods including EM [99,105], AFM [106,110] or X-ray crystallography [29,30,111,112].

In chapter 10, I will show how these oligomeric forms can be organized and rationalized based on pairwise mode of monomer association. In what follows, I will concentrate on the NpFs, the helical filaments built on DNA,
3.2. THE RECA NUCLEOPROTEIN FILAMENT

Figure 3.8: Variety in the oligomeric forms of recombinase proteins. (a) Different forms of RadA self association depending on the bound cofactor of DNA, after Figure 1 of reference [108]. (b) Three dimensional reconstruction of RadA octameric rings from the upper left images shown in (a) (after Figure 2 of [108]). (c) Crystal structures of left-handed (left) and overwound right-handed (right) Sulfolobus solfataricus (Sso) RadA filaments, after Figure 3 of reference [111]. The pitch value is indicated. Regions important for DNA binding (loops L1 and L2, HhH motif) are indicated by different colors (resp. pink, green and blue).

Figure 3.9: Structure of RecA filaments with bound DNA. The structures with PDB code 3CMW (6 monomers on ssDNA) and 3CMX (5 monomers on dsDNA) are respectively represented in the left and right panels. Monomers of the RecA protein are shown in cartoon representation with alternating colors. The poly-(dT) DNA strand in site I is in red (left and right) and the complementary poly-(dA) strand is in mauve (right). The structure of the (dT) strand is very similar whether or not it is paired with its complementary strand. After figures 1 and 4 of reference [30].

and their role during the recombination reaction.

NpFs and DNA.

As mentioned above, NpFs result from the polymerization of RecA monomers on DNA in the presence of ATP or ADP. RecA-ADP filaments only form on single-stranded DNA and dissociate in the presence of double-stranded DNA. RecA-ATP filaments can form either on ssDNA or on dsDNA.

In both cases, the NpF formation induces large amplitude deformations in the structure of bound DNA, which is stretched by about 50% [115] (40 to 45% for the ADP-form) and unwound by 40% [116] with respect to standard B-DNA. The reason is that in NpFs, the DNA is located at the filament center and is coaxial to the protein filament. In order to maintain favorable contacts along the densely packed filament interior, the DNA structure needs to adapt to the NpF helical characteristics (pitch and winding). With a stoichiometry of 3 nucleotides per RecA monomer, the result is a DNA with close to 18 nucleotides per helix turn and showing an average value of 5.1 Å for the inter base or base pair separation (rise) in the extended RecA-ATP form.

Indeed, the DNA in the NpFs presents a very specific conformation, revealed by the crystal structure of Chen and collaborators [30] (Fig. 3.9). In this structure, the stretching and unwinding deformations are not uniformly distributed. Rather, they concentrate in intercalation sites periodically found every three base pair steps and characterized by very large rise
Figure 3.10: RecA-bound DNA structure is intermediate between B and S DNA forms. Simulating the application of a pulling force on the 3'-extremity of a B-DNA (left) until the elongation reaches 50% of the initial size produces a stretched DNA form (right) that is very close to the RecA-bound crystal structure (the RMSD on phosphate atoms is 1.5 Å) [90,113]. This form corresponds to a metastable state between B-DNA and S-DNA (not represented) [114], as illustrated by a schematic free energy curve. The DNA are shown in van der Waals representation, with purines in orange and pyrimidines in cyan.

The structure of NpF-bound double-stranded DNA has been predicted in the laboratory as soon as 1999 as a 50% extension intermediate when going from B-DNA to overstretched S-DNA (70 to 80% extension). The structure has been obtained using molecular mechanics by simulating the application of a force on both 3'-extremities of the DNA [90,113,117]. A recent study using single molecule manipulation confirmed that when extended by 50%, DNA reaches a metastable state intermediate between the B-DNA and the S-DNA forms [114] (energy scheme in Fig. 3.10).

The structure of NpF-bound DNA is at the heart of the recognition/strand exchange process. However, the two crystal forms 3CMW and 3CMX corre-
3.3 ATP HYDROLYSIS

spond to NpF states before or after strand exchange: several intermediate structures are still missing and are the object of present research in the laboratory. Due to the dense packing within NpFs, it can be expected that any DNA structure variation that may accompany the strand exchange process will have some degree of coupling with the overall filament geometry and its possible variations. In addition to thermal fluctuations of the NpFs, one possible factor for NpF structural variations is ATP hydrolysis. We report below what is presently known on the role of ATP hydrolysis in the mechanism of homologous recombination.

3.3 ATP hydrolysis

In *E. coli*, the influence of ATP hydrolysis has been demonstrated both upstream and downstream of the recognition and strand exchange process that constitutes the heart of HR.

Upstream, very recent in vivo observations by 3D Super Illumination Microscopy (3D-SIM) indicate that ATP hydrolysis helps coordinating the search for sister chromatid within the cell via the formation of large RecA bundles at the location of double strand breaks (Figure 3.11), even before the SOS mechanism takes place. The bundles only form if ATP is hydrolyzed and their formation conditions the double strand break repair [13].

Figure 3.11: 3D-SIM imaging of RecA bundles. The left panel shows cells with double-strand breaks-induced bundles of RecA-GFP, alone (a), with the position of the double-strand break ends in red (b), the DNA in blue (c) and the membrane in red (d); the right panel indicates the typical dimensions of the bundle (e) and shows the movement at the bundle extremities (f); from [13].

Downstream, ATP hydrolysis is necessary for strand exchange to go to completion after initiation has successfully taken place [118, 119]. It allows strand exchange to proceed unidirectionally from the DNA 5'- to 3'-ends, to propagate over kilobases of genomic DNA and to bypass regions of heterologous insertions [120] or perform strand exchange between two double-stranded DNAs [120]. It is necessary for net filament disassembly after strand exchange [121]. However, and this remains very intriguing, ATP hydrolysis is NOT necessary for the central molecular events of DNA sequence recognition and strand exchange [122]. Not only are these steps successfully performed when ATP is replaced by a non-hydrolyzable analog (strand exchange can propagate up to 3 kilobases under strict conditions of homology) but the kinetics of strand exchange performed with either ATP or ATP\(_\gamma\)S are identical for the three early steps of recombination [123], characterizing the incorporation of the double-stranded DNA and the internal rearrangement
of the three strands [124, 125] (in these studies, the steps have been identified by monitoring dsDNA strand separation via fluorescence resonance energy transfer - FRET - between labeled bases). Only the final fourth step showed noticeable difference in the Gumb and Shaner's study, with a 4-fold reduction of its rate constant in the presence of ATPγS with respect to ATP.

Figure 3.12: Cox’s model of hydrolytic waves in a RecA filament. In Cox’s model, monomers periodically distributed every six monomers simultaneously enter an hydrolytic step (represented by black spheres). Displacement of the hydrolytic wave to the adjacent monomers occur every 0.5 s. Reproduced from Figure 10 of reference [126].

During the phase of strand exchange propagation, the RecA NpF acts as a molecular motor fueled by a high rate of ATP hydrolysis. Several models have tried to capture the mechanism associated to this motor function. Radding has proposed a system of concerted rotations in opposite directions for the two exchanging DNA species within the filament [127]. The group of Kowalczykowski suggested a mechanism of local redistribution, where the RecA subunits would dissociate in their RecA/ADP form and re-associate as RecA/ATP [119] but such polymerization/depolymerization mechanism could never be observed experimentally. Finally, the most accomplished model has been proposed by the Cox group as a “facilitated rotation model”, where the rotation of DNA would take place outside the protein filament [97, 128]. This model is backed on precise measurements of the rate of ATP hydrolysis in the different phases of the HR process, which are compared to the rate of filament disassembly. Cox’s results indicate the presence of waves of hydrolysis propagating along the filament, where one ATP molecule would be hydrolyzed every six monomers [126].

All these models lack a translation in terms of structure. One reason is that they were proposed before the structure of the RecA-ATP form was solved. However, even now that both RecA-ADP and RecA-ATP filament
structures are available, little is known about the dynamic evolution of NpFs in the presence of ATP hydrolysis. What are the structural repercussions of the constant ATP/ADP turnover on the active NpFs is an open question.

**Figure 3.13: Transitions between stretched and compressed forms of Rad51.** At the beginning of each experience A-C, the Rad51 filaments were assembled on DNA with ATP, resulting in an increase of the DNA length (vertical axis). Washing with a cofactor-free buffer decreases the size of the filament but does not lead to dissociation in 0 mM NaCl (A and B). C. Addition of AMP-PNP after Rad51 dissociation with 500 mM NaCl does not increase the DNA size. After Figure 6 of reference [129].

Given the difference between the monomer-monomer interfaces in the ADP and ATP forms, it was first proposed that no interconversion was possible between the two filament forms [98], which means that the passage from one filament form to the other would require complete dissociation followed by re-association. In this context, Klapstein and Bruinsma proposed a model where ATP hydrolysis would generate an elastic stress along the filament globally retaining its ATP form, due to the supposed impossibility for RecA/ADP monomers to adopt the equilibrium form they occupy in the compressed form [130]. This stress would be capable of inducing a force on DNA at the point of strand exchange. Building on Cox’s work identifying waves of concerted monomer hydrolysis, these authors proposed that NpFs would act as rotary motors propagating along the length of the filament.

Interconversion between the two filament forms has nevertheless been detected in recent studies using single molecule manipulation [129, 131, 132]. In these studies, the length of the filament shows discontinuous changes of its size during force-driven stretching/compression or cofactor exchange cycles where no RecA dissociation is detected (see for example fig. 3.13).

**Figure 3.14: Electron micrograph of mixed E96D RecA filaments.** The filaments formed on dsDNA in the presence of ATP present two different conformations, a compressed one (white arrows) and an extended one (black arrows), which often coexist within a same filament. It is proposed that the extended regions may be formed by ATP-bound monomers, while the compressed one would include nucleotide-free monomers. Filaments constructed in the presence of ADP only show the compressed form. Reproduced from reference [133], figure 1.

The reported observations indicate a highly cooperative transformation between the RecA/ATP and the RecA/ADP forms, which confirms former observation obtained with mutated E96D RecA proteins that do not hydrolyze ATP [133]. Filaments of the mutated monomers on dsDNA, formed in the presence of ATP, clearly show two conformations that may coexist
within a single filament but where the two distinct assembly modes undergo a phase separation (Fig. 3.14). Even that way, these new facts allow considering other scenarii for the molecular motor activity, where the ATP- and ADP-modes of monomer-monomer association would coexist within the filament. Such alternation of distinct binding geometries has been reported in cyclic oligomers as described in chapter 2. In some of those cases, variability within a single oligomer was proposed to be associated to the molecular motor function of the oligomer. Chapter 10 proposes some models of RecA filament presenting irregular monomer organization.
3.4 Conclusion

In addition to the obvious interest of understanding a process that is fundamental for cell survival, incentives for unravelling the mechanism of HR come from its interesting perspectives from a medical point of view. The ability of RecA to incorporate exogenous genes make it a widely used tool for gene targeting and a promising tool for gene therapy, providing an increase in its efficiency. In higher organisms, failures in the HR process in meiosis are associated with infertility, stillbirths and congenital diseases and in meiotic program with malfunction in maintaining the fidelity of segregation and in creating genetic diversity. In addition, Rad51 has been associated to cancer cell development.

In my thesis, I have developed tools for handling large helical oligomeric assemblies, which enable linking the local level of protein-protein and protein-DNA interactions to the level of the global architecture of the recombinase nucleofilaments. In chapter 10, I will present examples of exploratory work that I performed using these tools on the RecA system, which illustrate how accounting for the global filament architecture is essential to study any aspect of the HR mechanism. These studies intend to provide support for assembling pieces of the HR mechanistic puzzle. The aim of my work on RecA was thus to create new tools that were formerly missing to simultaneously cope with all dimensions of the HR system, then to contribute with these tools to investigations that are presently performed in the laboratory on the HR mechanism.
Part II

Methods
Chapter 4

Introduction

In the following four chapter, we will present PTools, the library that was the software basis for this whole thesis. PTools was designed to facilitate the development of methods or algorithms that manipulate macromolecular objects such as proteins or DNA [134] (chapter 5). Its first implementation already hosted the docking suite PyAttract, based on the ATTRACT docking method developed by Martin Zacharias [135, 136]. In the present work, I have extended the scope of the PTools library in two ways. First, I have implemented new functions to handle the internal flexibility of DNA, initially treated as a rigid body (chapter 6). I have also added a new module to address the global architecture of macromolecular assemblies with helix-based organization (chapter 8). Before detailing the structure and the possibilities of PTools and its new implementations, I will rapidly discuss the choices that have guided these developments in terms of macromolecule representation, level of resolution and modeling methods.

4.1 Methods and representation in PTools

The PTools library has initially been developed to address large scale exploration of possible assembly modes between macromolecules, which is at the heart of macromolecular docking methods.

4.1.1 Macromolecular docking

Docking methods aim at rapidly and efficiently predicting the three-dimensional structure of macromolecular complexes starting from available coordinates of the complex components. The coordinates are generally taken from the Protein Data Bank (PDB; http://www.rcsb.org/pdb) and correspond to structures of the isolated components determined by X-ray crystallography or Nuclear Magnetic Resonance (NMR) or they can be modelled from homology. The component structures may therefore differ from the structures
in the complex, notably due to surface rearrangement upon association as described in chapter 2 (section 2.3).

The development of macromolecular docking methods has been accompanied since year 2002 by the Capri experience [137], a blind prediction exercise where groups who develop docking programs are challenged for the prediction of solved but not yet released geometries of complexes, starting from the structure of their components. Evaluation of Capri round results enables tracking the progress in the field as well as identifying remaining bottlenecks.

Typical docking methods need to perform extensive exploration of the possible geometries of partner association and to identify favorable association modes among the generated geometries, using a scoring function. Both of these tasks can be performed at various degrees of resolution. For example, the macromolecules can be considered as rigid bodies (with no internal variability), which corresponds to a search in six-dimensional space (three translational and three rotational degrees of freedom). The search process can also explicitly take into account interface adjustment, thus increasing the dimension of the searched space. Depending on the degree of allowed flexibility, the dimension can jump from six to 3N degrees of freedom, where N is the number of atoms in the system. In the same way, the scoring function can vary from simple residue-based statistical function or geometric-fit measurement to approximations of the Gibbs free energy, which ideally is the accurate measure of the strength of association (chapter 2, section 2.1). The most precise of these options (full atomic flexibility and free-energy calculation) would necessitate intense sampling of the 3N degrees of freedom (where N may include solvation water molecules and ions) and therefore do not meet the requirement for rapidity of docking methods.

The development of efficient docking methods requires to make choices regarding the precision of the macromolecular representation used in the geometry search, the sampling method used for the search and the accuracy of the scoring function. Often, docking methods use successive search phases, performed at increasing levels of precision for the molecule representation and the scoring evaluation. Detailed information on the various approaches that have been proposed up to now can be found in several reviews or books [138–142]. We will focus here on the choices associated to the ATTRACT method.

4.1.2 Coarse-grained models

The use of low resolution, coarse-grained representation for macromolecules, together with a simplified energy function to score the predicted geometries, has appeared to be a good compromise between speed and accuracy in docking methods [135]. Indeed, ATTRACT showed very good performance in the Capri experience [143–145]. The coarse-grained model associated to AT-
TRACT groups from four to five heavy atoms in each bead, thus decreasing the number of particles in the system by a factor of $\approx 4$. In addition, using a reduced representation results in smoothing the molecule surface, thus making the model more tolerant to small conformational changes that may result from movements of small amino acid side chains. In addition, tolerance to small movement may reveal useful when studying systems with multiple binding modes like in chapter 10 of the present work. In such systems, too precise representation of the association geometry according to one specific mode may considerably weaken the association according to other modes unless the interfaces in these other modes are precisely adjusted, which generally involves lengthy exploration of the side chain conformations. This aspect is mentioned in chapter 10, section 10.5.

The coarse-grained representation used in ATTRACT does not handle the calculation of internal energy. In this representation, the macromolecule is a collection of packed beads which do not have any explicit connection or interaction. Nevertheless, it is possible to introduce flexibility in the model. For example, normal mode analysis can be performed by superposing a gaussian network model on the representation, where pairs of spatially close beads are linked by a spring function. Displacements along the lowest frequency modes\footnote{Two or three modes are generally sufficient to capture most of the possible internal deformation of the molecule [146].} can be added as supplementary variables during the docking search. As shown by May and Zacharias [146,147], this significantly improved the docking prediction in terms of ranking and in terms of the deviation of the predicted geometries from experiment. Flexibility can also be introduced by pre-generating (in atomic resolution) ensembles of conformers of a flexible region (e.g. long amino acids or flexible loops) and taking this ensemble into account during the docking search, each conformer being taken into account according to its interaction energy with the approaching partner [135,147,148].

The fact that such approaches were successfully implemented in reduced representation guided the elaboration of our approaches to treating DNA flexibility during protein-DNA docking, which will be exposed in chapters 6 and 7. Previous works in the laboratory by Poulain \textit{et al.} [17] had already identified the importance of shape recognition for protein/DNA assembly and had validated the appropriate use of coarse-grained models to probe shape recognition. As already reported in chapter 2, it was shown during this work that the correct geometry of sequence-selective protein-DNA complexes, with diverse modes of association, could be identified using the steric component of the interaction energy alone, without considering electrostatics. This proved true even when the bound DNA structure moderately deviated from a canonical B-DNA structure. The conclusion was that for more important deviation (bending greater than 20$^\circ$), it would be necessary
to explicitly take into account the protein-induced DNA deformations. I will show in chapter 6 how I increased the complexity of the DNA representation in PTools while conserving a coarse-grained representation.

Another important aspect when using coarse-grained representation is that it is particularly well adapted to treating huge systems while retaining the granular character of the molecule (which makes it easier to go back to atomic representation). In such representation, it is possible to perform extensive search on large systems with a reasonable calculation time. For example, Saladin et al. [134] used ATTRACT rigid body docking as a sampling method to identify all possible ways of DNA association to a helical turn of the RecA filament. The same approach was recently used to identify possible association geometries of the dystrophin molecular assembly to actin filaments, which were then refined using interactive docking simulation (O. Delalande, personal communication, manuscript submitted to Faraday Discussions). These first applications make us confident about the pertinence of using a coarse-grained level of resolution to treating very large assemblies. Chapter 8 will present the Heligeom suite of geometric transformations that open the way to treating such large regular or irregular assemblies. An application combining the advantages of coarse-grained representations in ATTRACT and Heligeom to sample both the local and the global geometries of macromolecular organization will be presented in chapter 10.
Chapter 5

PTools: overview and developments

This chapter will describe the tools we used both to develop the methods presented in chapters 6, 7 and 8 and to perform the applications described in chapters 9 and 10. It will start by a presentation of the PTools library, with an overview of its functionalities. The second part of this chapter will focus on PyAttract, a docking application that is part of PTools.

5.1 PTools

PTools is a docking library that relies on a modular, object-oriented implementation based on Python/C++ coupling. PTools was designed to perform methodological investigations on macromolecular assemblies as well as implement and test new methodological developments. To this aim, it is sufficiently flexible to easily allow the addition of new functionalities in an efficient and rigorous fashion.

PTools can handle coarse-grained as well as atomic macromolecular objects that can be compared or superposed for the purpose of analysis, or that can be docked using multiple energy minimizations in coarse-grained representation according to the ATTRACT protocol [135] (see section 5.2).

5.1.1 Design

C++ is a low level programming language, meaning that it is made to communicate easily with the hardware. It allows for fast computation, but the language can be cumbersome from the human side. The fact that it is compiled and strongly typed reinforces its performance.

Python is a high level language, meaning that it is easy to read and write for a human operator but it needs to go through some transformations to be understood by the hardware. It is easy to read through block indentation
and also very comprehensive, with numerous additional modules such as command-line option parsing, matrix handling, multithreading, interactivity with others programs, etc. Finally, Python, as an interpreted language, does not require compilation after each modification thus allowing rapid and flexible development of new features. However Python is considered rather slow compared to C++ and in itself unadapted to scientific computation.

Both are object oriented languages, a way of organizing code that allows for greater compartmentation. This means that a lesser part of a library is accessible to the end user, which is a good thing as it allows modifying a part of the library without disturbing implementations that use it.

PTools is implemented with a core in C++ interfaced with Python bindings (see fig.5.1), which allows for the best of the two worlds. Indeed, fundamental functionalities exist either as objects (Rigidbody, AtomSelection, transformation matrices, Forcefield, DNA, BasePair) or functions (superpose, automatic derivative calculation, DNA associated displacement –rise, shift, slide,... –, stacking) in the C++ part of the library, readily available in Python with minimal overhead. On the other hand, scripts (which form the PyAttract suite or the Heligeom suite) are written in Python for ease of use and modification. If they prove useful enough or if they lack speed, some of the fundamental functions can eventually be implemented in C++, allowing for both faster computation and new opportunity in methodological exploration. As an example, this is what happened to the stacking function during this thesis.

The following section will offer an overview of PTools functionalities.

<table>
<thead>
<tr>
<th>C++</th>
<th>Python</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classes</td>
<td>Classes</td>
</tr>
<tr>
<td>Rigidbody</td>
<td>atomic</td>
</tr>
<tr>
<td>AtomSelection</td>
<td>coarse grain</td>
</tr>
<tr>
<td>Transformation matrices</td>
<td>coarse grain</td>
</tr>
<tr>
<td>ForceField</td>
<td>AttractForceField</td>
</tr>
<tr>
<td>* DNA</td>
<td>* DNA</td>
</tr>
<tr>
<td>* BasePair</td>
<td>* BasePair</td>
</tr>
<tr>
<td>Functions</td>
<td>Functions</td>
</tr>
<tr>
<td>Superposition</td>
<td>Superposition</td>
</tr>
<tr>
<td>Automatic derivative</td>
<td>Automatic derivative</td>
</tr>
<tr>
<td>calculation</td>
<td>calculation</td>
</tr>
<tr>
<td>* Rise, Shift, Slide,Tilt, Twist, Roll</td>
<td>* Rise, Shift, Slide,Tilt, Twist, Roll</td>
</tr>
<tr>
<td>+ Stacking</td>
<td>+ Stacking</td>
</tr>
</tbody>
</table>

Figure 5.1: Overview of the PTools library. Additions made through this thesis are in red and will be present in the next release of PTools.
5.1.2 PTools Objects and Functions

C++ objects and functions (and corresponding Python interfaces) are the core of the PTools library. Methods that are either basic (useful to build on), generic (used by a large array of methods) and/or critical (speed is necessary for a viable implementation) are preferentially implemented that way. As this represents a cost both in implementation time and maintenance, methods that are too complex, too specific, fast enough or subject to change are implemented in Python.

Rigidbody and AtomSelection objects

Loading a PDB file into a Rigidbody object requires a single line of code. In the following example written in C++ the 1GC1.pdb file is loaded into the Rigidbody object prot. Then we select the chain A of the protein and write it into a new PDB file.

```cpp
Rigidbody prot("1GC1.pdb");
AtomSelection selA = prot.SelectChainId("A");
Rigidbody chainA = selA.CreateRigid();
WritePDB(chainA,"1CG1_A.pdb");
```

The equivalent Python code is:

```python
prot = Rigidbody("1GC1.pdb")
selA = prot.SelectChainId("A")
chainA = selA.CreateRigid()
WritePDB(chainA,"1CG1_A.pdb")
```

The similarities between the two languages result in near-identical Application Programming Interfaces (APIs) in the library, the main difference in the example above being that the type of a new variable is not declared in Python. Further examples will be only given in Python.

The Rigidbody object contains a vector of atom objects, each atom object grouping a set of atomic properties. In the following example, the second atom of the protein is extracted (indexed as 1 since the first atom is numbered 0), followed by modification of some of its atomic properties (coordinates, residue identifier and name) using low level methods.

```cpp
atom = prot.CopyAtom(1)
new_xyz = Coord3D(2.23,6.12,8.56)
atoms.SetCoords(new_xyz)
atoms.SetResidId(1)
atoms.SetResidType("LEU")
```

```python
atom = prot.CopyAtom(1)
new_xyz = Coord3D(2.23,6.12,8.56)
atoms.SetCoords(new_xyz)
atoms.SetResidId(1)
atoms.SetResidType("LEU")
```
5.1. PTOOLS

The class \texttt{AtomSelection} implements a convenient method for selecting atoms from a protein or DNA molecule. The user can filter atoms on properties like atom types, residue name, residue number, backbone or side chain and combine the selections using ensemble operators \texttt{AND}, \texttt{OR} and \texttt{NOT}, which gives full control over which atoms are included, in an efficient and intuitive way.

As an example, the following code creates a selection (\texttt{result}) containing non-C\textalpha atoms of residues 5-36 and 40-52 of the \texttt{Rigidbody prot}.

\begin{verbatim}
    sel1 = prot.SelectResRange(8,30)
    sel2 = prot.SelectResRange(45,52)
    sel3 = prot.CA()

    result = (sel1 | sel2) & !sel3
\end{verbatim}

This selection can then be converted into a \texttt{Rigidbody} object, which is a subset of the initial protein

\begin{verbatim}
    subprot = result.CreateRigid()
\end{verbatim}

In addition to the \texttt{Rigidbody} object, a \texttt{DNA} object was added during this thesis to address the specificity of the nucleic acids (represented in red print in fig. 5.1). More details can be found in chapter 6.

Transformation matrix

Translations and rotations of molecules are internally stored into a $4 \times 4$ homogeneous coordinate matrix. This is combined with a lazy evaluation of atom coordinates, which means that atom coordinates are evaluated only upon request (\texttt{WritePDB} function). When the macromolecule is submitted to a series of transformations, only the resulting $4 \times 4$ matrix is calculated. An advantage of these matrices is the storage of docking results. A docking simulation typically generates thousands of geometries (section 5.2), and replacing final ligand coordinates by a matrix saves a lot of disk space. In addition, this enables high performance for the construction and manipulation of large size macromolecular complexes, as will be seen in the next chapters (chapters 8, 9 and 10).

The following C++ code shows an example of a $\pi/4$ rotation of RigidBody \texttt{prot} around axis (AB), followed by a translation:

\begin{verbatim}
    PI = 3.141592653
    ptA = Coord3D(3.0, 4.0, 5.0)
    ptB = Coord3D(12, -5, 2)
    prot.ABRotate(ptA, ptB, PI/4.0)
    tr = Coord3D(6,7,8)
    prot.Translate(tr)
\end{verbatim}
The above code runs in constant time with respect to the number of atoms because, due to the lazy evaluation, only a $4 \times 4$ matrix has been modified.

PTools can use alternative, although equivalent, representation of molecular displacements. One of these is the **screw** transformation, where two positions of a molecule are related by a rotation around an axis and a translation along this same axis. The definition of the screw object and the function to convert a $4 \times 4$ homogeneous coordinate matrix into a screw object were included in the initial PTools release. The screw transformation is the basis of the new Heligeom module that was developed for the present work (see description in chapter 8). Another equivalent representation, called **Movement**, is used in the DNA module, also developed for this thesis. It enables the user to apply physically meaningful displacements of DNA base pairs in terms of helical parameters (see chapter 6).

**Superposition**

Another application where PTools can prove very useful is superposition, particularly when combined with the selection methods described above. Users can superpose two molecules in various ways provided that the two selections have the same size. The result of a superposition is an object which contains the Root Mean Square Deviation (RMSD) after superposition and a $4 \times 4$ homogeneous matrix to be applied to the mobile element to obtain this calculated RMSD.

The following code shows a superposition of two **Rigidbody** objects `prot1` and `prot2`, which have the same number of atoms.

```python
sup = superpose(prot1, prot2)
rmsd_best = sup.rmsd
mat = sup.matrix
```

Variable `mat` now contains the matrix that has to be applied to `prot2` in order to minimize its RMSD with respect to `prot1`.

**Force field**

The PTools library contains by default a force field associated to the ATTRACT docking method (fig. 5.2), described in the next section. PTools is coded in a way that makes the addition of a new forcefield easy. Indeed, two new force fields will be available in the next release of PTools, planned for the end of 2014.

- **ATTRACT force field 2.** The group of M. Zacharias has developed a new force field for ATTRACT, which incorporates knowledge-based
parameters in such a way that native geometries are better distinguished from near-native ones in docking simulation results. The functional forms as well as the coarse-grained representation are also modified in this force field. A description can be found in reference [136]. Although it generally leads to improved docking results with respect to the original force field, this force field was not used in the present work.

- **Scorpion.** The Scorpion force field was developed by N. Basdevant and T. Ha-Duong following a bottom-up strategy based on MD simulations in atomic coordinates, both to determine the parameters and the functional forms of the energy function. More detail can be found in reference [149].

Currently, PTools is interfaced with the L-BFGS minimizer that was written in FORTRAN by Jorge Nocedal [150, 151]. This minimizer can be used with any force field implemented in PTools. It is used in PyAttract.

### 5.1.3 Modules and Scripts

Python scripts complement the PTools library. They are the preferred mode of implementation of methods, as they present a more versatile solution to frequent changes or adaptations to specific problems than hard coded implementation in C++.

**PyAttract suite**

The PyAttract suite groups the script `PyAttract.py` which implements the ATTRACT docking method, the script `reduce.py` that transforms a molecule from all atom representation to coarse-grained representation and scripts useful to analyze the results of the docking simulation such as: `Extract.py`, that allows obtaining a PDB file from a prediction, `cluster.py` that allows clusterizing the docking results, `fnat.py` or `irmsd.py` that permit the calculation of the fraction of predicted native contacts and the interface Cα-RMSD, respectively, when a reference structure of the native complex is available. The docking procedure will be presented in more detail below (see also an example in Appendix B).

**Heligeom suite**

Added during the thesis, the Heligeom suite allows building and manipulating molecular construction with a helical character. It will be presented in greater detail in chapter 8.
5.2 PyAttract

The PyAttract.py script contains the implementation of the ATTRACT docking method in PTools. It performs extensive search of the possible geometries of association between two macromolecular partners, using a multi-minimization strategy and a reduced representation for the target proteins or nucleic acids (the reduction factor is about four heavy atoms per grain). The force field governing the docking process is composed of a Coulombic term screened with a distance-dependent dielectric function and a smooth, 6-8 van der Waals terms 5.2. No internal energy is calculated.

\[
E = \sum_{i \in I, j \in J} \left( \frac{B_{ij}}{r_{ij}^8} - \frac{C_{ij}}{r_{ij}^6} \right) + \sum_{i \in I, j \in J} \left( \frac{q_i q_j \epsilon_{ij}}{r_{ij}} \right)
\]

Figure 5.2: The force field equation used in pyAttract.

The principle that guides the modeling choices in ATTRACT is to trade off some details of the macromolecule representation against a quicker computation, as not only these details are not essential to identifying correct geometries of association but they may obfuscate the problem by introducing energetic local minima irrelevant to the prediction of the complex geometry. One factor that reduces the level of detail is the coarse-grained representation used for both molecules (fig. 5.3).

Figure 5.3: Coarse-grained representation in PyAttract. Atomic representations of tyrosine (left) and thymine (right) residues in stick mode are superposed on coarse-grained representation in transparent spheres.

Computation wise, as the number of particles is smaller, less time is spent computing the energy. The functional form of the van der Waals energy, with its 8-6 exponential terms instead of the commonly used 12-6 terms, also contributes to producing smoother landscapes. This is actually an advantage in our simulation as unnecessary energetic local minima are ignored from the search.

In the following description, the molecules are represented as rigid bodies. One partner (arbitrarily chosen) is called the receptor, and the other partner is the ligand. The receptor will be held fixed in the docking run.
The ligand can modify its position following six degrees of freedom (three rotational and three translational variables). This means a quicker simulation, as the minimization only concerns six variables.

Preparation of an ATTRACT docking run necessitates the creation of reduced structures for the docking partners using the `reduce.py` script (where `receptor.pdb` and `receptor.red` contain the coordinates of each receptor particle in the atomic and reduced representation, respectively)

```
python reduce.py --prot receptor.pdb > receptor.red
python reduce.py --prot ligand.pdb > ligand.red
```

Starting points are then distributed around the (fixed) receptor using the `translate.py` script

```
python translate.py receptor.red ligand.red > translation.dat
```

From each starting point and for each of about 200 predefined ligand orientations, the `Attract.py` script runs a series of six minimizations of the interaction energy between the receptor and the ligand. When a reference structure (`reference.red`) is known for the ligand, the RMSD with respect to this structure is calculated using the Cα atoms

```
python attract.py -r receptor.red -l ligand.red
               --ref=reference.red > attract.att
```

Fig. 5.4 illustrates the method. The PyAttract.py script is used in chapter 10 of the present work. A detailed example can be found in Annex B.
**Figure 5.4: The ATTRACT docking approach.** Scheme of the ATTRACT docking strategy in the case of a protein-DNA complex. The following steps are illustrated: (top, from left to right) conversion from reduced to coarse-grained representation; generation of starting points distributed around the receptor; (bottom, from right to left) multiple minimization step – the center of mass of the ligand is successively positioned on each starting point and for each pre-defined orientation, a series of six minimizations of the interaction energy is performed; analysis – plot of energy versus Ca-RMSD with respect to the reference structure when available; resulting complex.
Chapter 6

Manipulation of flexible DNA in PTools

6.1 DNA in PTools

Added during this thesis, the DNA PTools class enables the construction of double-stranded DNAs of any sequence and presenting a variety of shapes. PTools/DNA shares several functionalities with the 3DNA software for DNA construction and analysis [152,153], or with 3D-DART which generates customized DNA structures [68] or GraphiteLifeExplorer for easy graphical and interactive construction [154]. However, a notable characteristic of the PTools implementation is that it offers a multi-level representation of DNA. In addition, the flexibility of Python environment enables users to tailor their own application scripts.

Two representations are proposed, either a detailed atomic representation or a coarse-grained representation with a degree of reduction of about four heavy atoms per grain [17] (fig.5.3 (right) in chapter 5). Fig. 6.1 describes the partition into grains of atoms from the four types of nucleotides.

In PTools, the nucleic acids are stored in DNA objects. There are several ways to create a DNA object. It can be created empty:

\[ d = \text{DNA}() \]

Or from another DNA object:

\[ d = \text{DNA}(\text{old}_d) \]

It can also be generated from a PDB file. In that case, a base pair data file is required, which contains the reference structure of each base pair. Two reference files are presently available in PTools, one of them, pb.ato.pdb,
Figure 6.1: Definition of the five (pyrimidines) or six (purines) grains representing each of the four nucleotides.

corresponds to an all-atom representation and the other one, pb.red.pdb, to a coarse-grained one. For example, the command

\[
d = \text{DNA("pb.red.pdb","mydna.pdb")}
\]

will create a DNA in reduced representation, following the internal DNA geometry of the PDB file mydna.pdb.
6.1. DNA IN PTOOLS

Unlike the Rigidbody objects, the DNA objects can be subdivided into fragments. The object BasePair is a convenient way to hold the base pair data of the DNA object. Individual base pairs can be easily accessed for edition as well as to control internal deformation. Below is an example of base pair extraction

\[
d = \text{DNA("pb.ato.pdb","AAAA",BDNA())}
\]

\[
\text{basePairNumber2} = d[1] \quad \# \text{numeration starts at 0}
\]

In the first line, BDNA() indicates that the DNA of sequence d(A)₅ is constructed with a B-DNA geometry (see below).

6.1.1 Editing DNA structures

DNA structures can be constructed from scratch or edited in several ways, such as sequence modification, fragment extraction or concatenation.

These possibilities will be illustrated here by reporting a set of commands used to construct a composite DNA, starting from the structure of a TBP-bound DNA (PDB code 1QN4; named 1QN4-DNA.pdb, fig. 6.2d) \[155\]. Snapshots of the construction are represented in fig. 6.2.

Figure 6.2: Construction steps of a DNA oligomer. The labels d to d⁷ refer to the PTools/DNA objects described in the text, constructed using functionalities of the PTools/DNA toolkit. d-d₄: construction steps; the eight base pairs in d₁, which have been extracted from the 1QN4 crystal structure (d), then concatenated with new DNA parts (d₃, d₄), are lined with a red triangle; d₅-d₇: edition and deformation steps; in d₅, the sequence has been changed to d(A)₃₀ and the representation has been set to coarse-grained; in d₆, the base pair where a 45° Roll movement has been applied is indicated by a red line and the direction of the rotation is shown with a red arrow (see text). The DNA oligomers are represented using van der Waals representations, with the backbone colored in grey and the bases respectively colored in orange for A, G, cyan for T, C and using the VMD software \[156\].
CHAPTER 6. MANIPULATION OF FLEXIBLE DNA IN PTOOLS

The following Python code snippet reads the PDB file, adjusts the atomic coordinates to conform to the library geometry (object \texttt{d}), and writes the resulting file into the PDB file \texttt{D.pdb}.

```python
from ptools import *

d = DNA("bp.ato.pdb","1QN4-DNA.pdb")
d.WritePDB("D.pdb")

Let us first extract the eight central base pairs of the structure (DNA object \texttt{d1}). This can be done with the command

\texttt{d1 = d.SubDNA(2,9)}

We now create two new structures, one of a B-form DNA of sequence \texttt{d(GCGAAC)} (object \texttt{d0}) and the other an A-form DNA of sequence \texttt{d(GGCGGCGCGCGCGCGG)} (object \texttt{d2}). These DNA fragments are then concatenated with \texttt{d1} in such a way that \texttt{d1} is positioned at the 3' extremity of \texttt{d0} and \texttt{d2} at the 3'-extremity of \texttt{d1}. Object \texttt{d3} results from the concatenation of \texttt{d0} and \texttt{d1} and object \texttt{d4} includes the three fragments;

\texttt{d0 = DNA("bp.ato.pdb","GCGAAC",BDNA())}
\texttt{d2 = DNA("bp.ato.pdb","GGCGGCGCGCGCGCGG",ADNA())}
\texttt{d3 = DNA(d0)}
\texttt{d3.Add(d1)}
\texttt{d4 = DNA(d3)}
\texttt{d4.Add(d2)}

The \texttt{BDNA()} and \texttt{ADNA()} objects in the first two lines contain predefined DNA internal geometries for B-DNA and A-DNA and will be described below.

It must be noted that the bases in the final structure follow the usual sequential order from 5' to 3' extremities.

The \texttt{changeType()} method also gives the possibility of changing the sequence of the whole oligomer to any desired sequence. To change the \texttt{d4} sequence to \texttt{d(A)_{30}} (object \texttt{d5}), the base pairs in the molecule are modified sequentially

\texttt{d5 = DNA(d4)}
\texttt{for i,c in enumerate ("AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA"):}
\texttt{    d5.ChangeType(i,c,"bp.red.pdb")}
```
In this example, not only the identity (using the letter c taken from the provided chain of characters), but also the representation (to coarse-grained, by specifying the library file \texttt{bp.red.pdb}) are modified for each basepair \( i \).

### 6.1.2 Manipulating DNA structures

DNA or DNA fragments can be manipulated as rigid bodies, using six degrees of freedom stored in \( 4 \times 4 \) homogeneous matrices as described in chapter 5. For convenience when manipulating DNA objects, PTools also provides a \texttt{Movement} object together with special objects inherited from the \texttt{Movement} object, which define rotation or translation that can be directly related to the helical character of DNA. Rotations are given in degree. The next lines characterize a rigid body movement with 2 Å translation in all three directions and 15° rotation in all three rotational directions.

\[
\begin{align*}
\text{mov} &= \text{Shift}(2) \ #\text{translation along } x \ (\text{short axis}) \\
\text{mov} &= \text{Slide}(2) \ #\text{translation along } y \ (\text{long axis}) \\
\text{mov} &= \text{Rise}(2) \ #\text{translation along } z \ (\text{perpendicular axis}) \\
\text{mov} &= \text{Twist}(15) \ #\text{rotation around } z \ (\text{perpendicular axis}) \\
\text{mov} &= \text{Roll}(15) \ #\text{rotation around } y \ (\text{long axis}) \\
\text{mov} &= \text{Tilt}(15) \ #\text{rotation around } x \ (\text{short axis})
\end{align*}
\]

Specific \texttt{Movement} objects can be used for example to position a second object with respect to the first one.

```
\text{mov} = \text{Movement(matrix)} \\
\text{matrix} = \text{mov.getMatrix()}
```

Individual \texttt{Movement} objects can be combined to form a new movement, as shown below. Note that there is a special order to mix the translations and each of the three rotations:

\[
\begin{align*}
\text{mov} &= \text{Twist}(31.1) + \text{Roll}(2.0) + \text{Tilt}(2.1) + \text{Rise}(3.3) \\
&\quad + \text{Slide}(-2.4) + \text{Shift}(-0.5)
\end{align*}
\]

### 6.1.3 Deforming DNA structures

In addition to rigid body displacements, it is possible to modify the internal arrangement of the base pairs. To this aim, the \texttt{Movement} objects defined above can be applied to individual base pairs. They can be directly related
to the six inter base pair parameters commonly used to characterize the internal geometry of DNA (fig. 6.3).

Standard DNA conformations such as B-DNA and A-DNA can easily be generated using either of two pre-defined Movement objects. The first object, BDNA, is for DNA in B conformation. The second one, ADNA is, as can be guessed, for a A-form DNA.

To obtain the parameters of a specific DNA conformation with available structure, one can use the "computeParametersOfDNA.py" script provided in the DNA/ directory of the PTools library.

```
python computeParametersOfDNA.py mydna.pdb
```

Modifying the internal arrangement of the base pairs can be done either locally, by modifying the position/orientation of a base pair relative to the preceding one (in which case the following bases are also displaced by default), or globally by applying a given conformational change simultaneously to all base pairs. The first transformation can be made through the use of the ApplyLocal() method, the second one uses the ApplyGlobal() method.
These two types of deformation are illustrated in the following lines, which continue the transformations of DNA structure in fig. 6.2, starting back from the d4 object:

\[
\begin{align*}
    \text{d6} &= \text{DNA(d4)} \\
    \text{d6}.\text{ApplyLocal(Roll(45.),14)} \\
    \text{d7} &= \text{DNA(d4)} \\
    \text{d7}.\text{ApplyGlobal(Twist(20.))} \\
    \text{d6}.\text{WritePDB("D6.pdb")} \\
    \text{d7}.\text{WritePDB("D7.pdb")}
\end{align*}
\]

In this sequence of operations, a 45° Roll movement is locally applied to base pair 14 of d4 (resulting in object d6), then a 20° Twist movement is globally applied to d4 (resulting in d7).

## 6.2 First application

The results of Poulain et al. [17] were the starting point of my work on flexible protein-DNA docking. They demonstrated that using a B-form DNA in coarse grain representation as a ligand for rigid body docking was sufficient in many cases to predict a high percentage of native protein-DNA contacts. Even more interestingly, partial recovery of the native interface could be observed for difficult cases where the bound DNA conformation was very different from B-DNA.

Thus our initial idea for developing a flexible docking strategy was to first proceed with a rigid docking simulation and use the interface information issued from the rigid body docking as a starting point for further exploration of both the internal conformation of the DNA and its position with respect to the protein. By using this approach, we could focus the search on a limited part of the protein. This restriction of the search space meant that we could use more complicated methods to address flexibility and still conserve reasonable computational time.

As described in the previous section, we chose to model the flexibility of double-stranded DNA by treating the base pairs as separate rigid bodies that can move independently from each other in the six classical degrees of freedom (three degrees of rotation and three of translation). The base pairs are in the Poulain coarse-grained model, and they are initially disposed as in a B-DNA step. Their position is internally recorded as a 4 × 4 homogeneous matrix but can also be accessed in PTools using the Movement objects.

The challenge was thus to go from a B-form DNA spatially close to its binding position with respect to the partner protein to the conformation that would best fit the binding surface. There are many ways to proceed, for
example by computing the normal modes of the DNA and minimizing the interaction energy along these modes (which is the strategy used by Zacharias for protein-protein [145] or partly by Bonvin [67]) but this approach is dependent on the initial conformation and does not account for anharmonic deformations such as kinks, or too important deviation. Another way would be to use molecular dynamics but this is too computationally expensive to be included in a docking method that we want as rapid as possible. We therefore opted for a worm-like chain model, in a coarse-grained force field and with a Monte Carlo exploration method.

To predict a correct conformation we implemented a Monte Carlo (MC) exploration where the random DNA conformational change was obtained by modifying the value of one of the 6 degrees of freedom of a randomly chosen continuous set of base pairs, with a size comprised between one and the length of the DNA. In our preliminary tests [157], all the bases of the set were simultaneously modified using a “global” type of movement (see above). Acceptance or rejection followed a Metropolis criteria, based on the new interaction energy of the modified DNA.

Of course, some control on the internal geometry of the DNA had to be added to the initial force field, where only the interaction energy between protein and DNA is considered. Otherwise, impossible conformation would soon have emerged such as all the base pairs collapsing on the most attractive part of the interface. To prevent this we imposed constraints to the system. We first limited the amplitude of the base pair translations and rotations. We also imposed minimum and maximum distances between consecutive base pairs. Finally, we implemented a stacking measurement based on geometric criteria [158] and imposed a minimum percentage of stacking.

**Figure 6.4: Geometric criteria for stacking interactions.** Two aromatic cycles of DNA bases are considered stacked if (i) the distance $d_{ij}$ between the geometric centers of the cycles is less than 4.5 Å; (ii) the angle between two consecutive $n_i$, $n_j$ vectors, normal to the cycle plane, is less than $23^\circ$; (iii) the angle $\tau_{ij}$ between vector $n_i$ and the vector joining the geometric centers is less than $40^\circ$. The limiting values were tuned against a benchmark of regular and irregular DNA structures [158].

As shown in fig. 6.5 the use of this method improved our docking prediction compared to rigid docking of B-DNA. However, it became soon clear that the method needed improvement. As an example if the bases presented favorable interactions with non contiguous patches of the protein surface in the initial rigid body docking pose or during the MC search, and if these interactions were not correctly phased (e.g., in terms of the number of base pairs between two patches), it was extremely hard to make a base pair break
Figure 6.5: Results of the Monte Carlo flexible docking procedure.
free from its good interface to allow the correct base pair to take its place. More generally, it appeared that the method as it was unable to go through the energy barriers at reasonable computational speed. In addition, even when the protein surface was correctly matched by the DNA contacting base pairs, the generated internal DNA conformations did not always reproduce the global features of the experimental distortions (notably distortions resulting from cooperative movements such as minor groove opening).

At this preliminary stage of the method development, there was still a lot of space for improvement. At the level of the internal distortions, we could integrate internal energy terms in the coarse-grained DNA force field (at the cost of parameterizing these energy terms and tuning them against the interaction terms), thus allowing more efficient sampling than by using a simple geometrical set of constraints, as we would explore less unnatural conformations. The sampling could also be improved using more efficient methods than a simple MC such as replica exchange or tabu search [159] that would more largely sample the potential energy surface. However, it is not clear that the increase in sampling efficiency would compensate for the additional computational cost. Moreover, the method relied on the hypothesis that the correct interface would be detected as the result of rigid body docking using B-DNA conformation, which remains uncertain. Even if the interface was detected, the correct positioning may not necessarily rank first in terms of interaction energy, which would multiply the number of flexible search simulations starting from different docking poses.

Another alternative was to decouple the characterization of the DNA-binding interface on the protein and the generation of the DNA internal deformation that would best fit this interface. The next chapter will expose how we proceeded.
Chapter 7

Burdock

As we have seen in the preceding chapter, our first investigation suggested that a DNA in the canonical B-form correctly positioned with respect to the protein surface was not necessarily a good starting point for our method when using a DNA representation with uncorrelated base pairs positions. We chose to look for a starting conformation where the global DNA form would be closer to the correct solution. One of our advantage was that, as we had an optimization method that worked well for small adjustment of the DNA conformation, only a rough approximation of the overall structure was needed.

But how could we obtain this rough approximation? Assuming that we could obtain the correct interface from docking, directly predicting a DNA from this information would be difficult and expose us to some of the previous problems (namely, the misplacement of base pairs). Thus we decided to use an intermediary representation of the DNA, in the form of its axis. Once we have an axis situated in the interface, we can easily generate a set of starting conformations with different base pairs in contact, simply by rotating the DNA around this axis. Additionally, the axis offers a global representation of the DNA that our worm-like chain DNA model lacks, and it can be used as a global restraint when exploring the base pair positions/orientations with respect to the protein. This can be inexpensively implemented in the method. Finally, the search for the axis can involve the whole protein surface, thus freeing us from the assumption that the correct interface has been identified following rigid B-DNA docking.

I will first present a summary of the method I elaborated on this basis, and then examine in detail each step. Then, the main tools that were developed during the implementation will be presented, followed by a word on the unexpected turn this work took.
7.1 Overall strategy

This method, that was known in the lab as “Burdock” from the name of the plant that inspired Velcro\(^1\), can be summarized as follows: first, we do a docking simulation of small double-stranded B-DNA fragment (docking); using the best prediction results, we obtain a set of points in space (pruning); we clusterize this set into groups of points (clustering) that are coherent with the goal of creating an axis (axis building); once the axis is generated, we ”build” a DNA around it and we optimize its conformation with the method presented in the preceding chapter (DNA building).

\(^1\)Velcro was conceptualized by George de Mestral, a Swiss engineer, after a trip in the Alps with his dog in 1941 where both of them got covered with burdock burrs; https://en.wikipedia.org/wiki/George_de_Mestral#cite_note-Better-3.
7.1. OVERALL STRATEGY

7.1.1 Preliminary docking and pruning

The first step consists to dock small double stranded B-DNA fragments on the protein receptor. We tested different DNA sizes (from 3 to 9 base pairs) on the protein-DNA docking benchmark proposed by Van Dijk et al. [160]. The benchmark is composed of 47 protein-DNA complexes, both in bound and unbound conformation, organized by difficulty of prediction (as evaluated by the RMSD difference between the bound and unbound conformations).

In a majority of cases, the different DNA fragments performed well, with an average of 90% of the interface recovered by the cumulated contact of the 500 best predictions. Additionally, 20% on average (or one in five) of the 500 best predictions were in contact with the correct interface. In most cases the different fragments performed undifferentially, but the 3 base pair long DNA fragments produced better results than all others in cases were the electrostatics recognition was of prime importance, and conversely the 9 base pair long DNA fragments performed better when steric recognition was the prime factor.

In the end, we decided to keep the 250 best predictions from the docking simulation of a 3 base pair DNA fragment and the 250 best predictions from a 9 base pair fragment docking simulation.

Figure 7.2: Burdock step 1: Pruning.
CHAPTER 7. BURDOCK

7.1.2 From prediction to dot

Once a set of predictions has been selected, the structure of the short DNA fragments is not necessary any more. Two approaches were tested: either replace each predicted DNA fragment by its center of mass, or replace each base pair of these predicted fragments by its center of mass. Although the second approach produces more points and is thus more computationally expensive in later steps, it also enables to conserve information on the directionality of the predicted geometry and so was favored.

Figure 7.3: Burdock step 2: Shaping clouds.
7.1.3 Clustering

The points that remain after the pruning step generally occupy several
regions in space, corresponding to regions of the protein interface that can
favorably bind the DNA. The size and shape of each region will determine
the feasibility of a DNA binding to that region: the resulting axis must
not be bifurcated and must be long enough to allow the final formation
of protein-DNA interface with sufficiently large size (see chapter 2 for the
typical size of protein-DNA interfaces). The existence of more than one re-
gion of interaction can cause issues in later steps, especially during the axis
construction. To circumvate this problem, we perform a clustering of the
points to ensure that the set of points that will be used in the following
steps will be coherent (topologically linear, with reasonable curvature). The
choice of the clustering algorithm was still in discussion, but the necessity
of an algorithm able to determine by itself the number of clusters was es-

![Figure 7.4: Burdock step 3: Clustering.](image)

The points that remain after the pruning step generally occupy several
regions in space, corresponding to regions of the protein interface that can
favorably bind the DNA. The size and shape of each region will determine
the feasibility of a DNA binding to that region: the resulting axis must
not be bifurcated and must be long enough to allow the final formation
of protein-DNA interface with sufficiently large size (see chapter 2 for the
typical size of protein-DNA interfaces). The existence of more than one re-
gion of interaction can cause issues in later steps, especially during the axis
construction. To circumvate this problem, we perform a clustering of the
points to ensure that the set of points that will be used in the following
steps will be coherent (topologically linear, with reasonable curvature). The
choice of the clustering algorithm was still in discussion, but the necessity
of an algorithm able to determine by itself the number of clusters was es-

![Figure 7.4: Burdock step 3: Clustering.](image)

Figure 7.4: Burdock step 3: Clustering.

The points that remain after the pruning step generally occupy several
regions in space, corresponding to regions of the protein interface that can
favorably bind the DNA. The size and shape of each region will determine
the feasibility of a DNA binding to that region: the resulting axis must
not be bifurcated and must be long enough to allow the final formation
of protein-DNA interface with sufficiently large size (see chapter 2 for the
typical size of protein-DNA interfaces). The existence of more than one re-
gion of interaction can cause issues in later steps, especially during the axis
construction. To circumvate this problem, we perform a clustering of the
points to ensure that the set of points that will be used in the following
steps will be coherent (topologically linear, with reasonable curvature). The
choice of the clustering algorithm was still in discussion, but the necessity
of an algorithm able to determine by itself the number of clusters was es-

Figure 7.4: Burdock step 3: Clustering.

The points that remain after the pruning step generally occupy several
regions in space, corresponding to regions of the protein interface that can
favorably bind the DNA. The size and shape of each region will determine
the feasibility of a DNA binding to that region: the resulting axis must
not be bifurcated and must be long enough to allow the final formation
of protein-DNA interface with sufficiently large size (see chapter 2 for the
typical size of protein-DNA interfaces). The existence of more than one re-
gion of interaction can cause issues in later steps, especially during the axis
construction. To circumvate this problem, we perform a clustering of the
points to ensure that the set of points that will be used in the following
steps will be coherent (topologically linear, with reasonable curvature). The
choice of the clustering algorithm was still in discussion, but the necessity
of an algorithm able to determine by itself the number of clusters was es-

Figure 7.4: Burdock step 3: Clustering.

The points that remain after the pruning step generally occupy several
regions in space, corresponding to regions of the protein interface that can
favorably bind the DNA. The size and shape of each region will determine
the feasibility of a DNA binding to that region: the resulting axis must
not be bifurcated and must be long enough to allow the final formation
of protein-DNA interface with sufficiently large size (see chapter 2 for the
typical size of protein-DNA interfaces). The existence of more than one re-
gion of interaction can cause issues in later steps, especially during the axis
construction. To circumvate this problem, we perform a clustering of the
points to ensure that the set of points that will be used in the following
steps will be coherent (topologically linear, with reasonable curvature). The
choice of the clustering algorithm was still in discussion, but the necessity
of an algorithm able to determine by itself the number of clusters was es-

Figure 7.4: Burdock step 3: Clustering.

The points that remain after the pruning step generally occupy several
regions in space, corresponding to regions of the protein interface that can
favorably bind the DNA. The size and shape of each region will determine
the feasibility of a DNA binding to that region: the resulting axis must
not be bifurcated and must be long enough to allow the final formation
of protein-DNA interface with sufficiently large size (see chapter 2 for the
typical size of protein-DNA interfaces). The existence of more than one re-
gion of interaction can cause issues in later steps, especially during the axis
construction. To circumvate this problem, we perform a clustering of the
points to ensure that the set of points that will be used in the following
steps will be coherent (topologically linear, with reasonable curvature). The
choice of the clustering algorithm was still in discussion, but the necessity
of an algorithm able to determine by itself the number of clusters was es-

7.1.4 Axis construction

Given one set of points determined in the preceding step, an axis can be built according to the following steps: (i) first for N iterations, a random point in the set is selected, and the positions of every points situated within a given radius (the including radius), are averaged; these points are then replaced by a unique point at the averaged position; this allows linearising the set of points and reducing the influence of lone outliers; in addition, the number of points is reduced, which results in faster computation; (ii) in the next step, a random point is selected among the remaining points, and all points of the set within a given excluding radius are discarded; we then average all the unit vectors starting from this point and pointing at other points of the set within a given radius (the detecting radius, naturally larger than the excluding radius) and we place a new point at a given distance along this averaged vector. We repeat this step until there is no point left to build the axis. Typically, the detecting radius is taken between 5 and 10 Å, with larger values corresponding to straighter axis. The excluding radius is taken as half the detecting radius. We then proceed by starting a new search from the first selected point of the axis to make sure the full extent of the set was used.

The result is a collection of points (green in fig. 7.5) slightly shorter than the set that would be used to describe a full axis (red points in fig. 7.5).
7.1.5 DNA construction

The exact algorithm used for DNA construction given an axis will be presented in the following chapter on Heligeom, as it applies not only to DNA but to any molecular object with helical shape.

Few things should be noted though:

- during DNA reconstruction, the transformation that calculates the position of a base pair from the preceding one is taken as a classic B-DNA step (before deformation guided by the axis);
- reconstruction naturally form open grooves or kinks in the DNA structure if the axis present a curve or a brutal change of direction, respectively;
- at this point, the possibility to discard the DNAs formed on small size axes was considered;
- as the orientation of the first base pair is randomly chosen, the DNA is not guaranteed to be in phase with the protein; further exploration needs to be done to optimize the phase.
7.1.6 Optimization

The strategy for optimization was twofold: first place the DNA in an optimal phase with respect to the protein without modifying the internal geometry resulting from construction, then proceed to Monte Carlo optimization of the internal DNA conformation in the presence of the protein as described in the Application section of chapter 6.

We first approached the optimization phase with a simple solution: rotating the DNA on its axis, in order to present different orientations of the base pairs to the protein. An extra step was added where the base pairs could translate along the axis. Still, this was not enough to get rid of steric clashes that formed between the protein and the DNA base pairs, even when further adjusting the base pair orientation.

In a second approach, we first slightly separated the constructed DNA from the protein and then proceeded to a series of energy minimization cycles starting from 360 structures obtained by rotation around the axis by steps of $1^\circ$, the minimization thus taking care of all the small adjustments of both the axis position and the DNA phasing around the axis necessary for optimal protein-DNA fitting. This method was successful and was further improved by the possibility to insert pre-deformed DNA fragments during the DNA construction around the axis in such a way as to make conformational transitions possible. These types of deformations, like the B-form to TA-form transition [65], are very hard to reach using uncorrelated variations of the base pair parameters [67].

7.2 Conclusion

In parallel with my work on docking, I have been involved during my thesis in a long term project going on in our group at the LBT, concerning the study of the RecA filament and the mechanism of homologous recombination [90, 113, 117, 163]. It appeared that the tools I developed to treating DNA internal flexibility for flexible DNA docking, which are generic with regards to the molecule considered, were very well adapted to handling the global or local variations of the RecA fibers. Far from being a quick hack to work out some minor and specific aspect in the construction of RecA-DNA systems, these tools and the underlying approach in general proved to be thought provoking in regard to the modelisation of protein filaments. When the group of Haim Wolfson published their ParaDOCK approach to flexible protein-DNA docking [18] that shared some of the key features I had imagined for Burdock, we decided to concentrate on this unexpected and promising new lead. The next chapter will present our results in this endeavor.
Chapter 8

Heligeom

Heligeom is a module made with and for the PTools library. It is based on the Python script heligeom.py and contains a number of associated scripts that will be described below. By extension the name Heligeom can refer to the overall integrated approach that has been developed concurrently with the script.

8.1 Screw movement

Heligeom is based on the concept of screw movement that appeared with the early developments of the theory of rigid body kinematics in the nineteenth century [164](for example in the Chasles theorem [165]). Specifically, the movement of a rigid body, generally described as the combination of a translation and a rotation, can be represented in an equivalent way as a screw transformation, i.e. a rotation of the rigid body around an axis, combined with a translation along this same axis (fig. 8.1) [164]. In the extreme case where there is no translation, the movement becomes an in-plane rotation around the axis; alternatively, if the rotation angle is null, the rigid body is displaced along a straight line. Regular repeats of screw transformations generate helices, rings or straight segments that will be referred to as screw-type organization.

The screw transformation relating one helical unit to the following one may be calculated following a variety of geometric methods. Its implementation by Adrien Saladin in the original release of PTools\(^1\) followed the elegant analytical method published by Angelidis in 2004, based on his concept of hexanion [166]. In this method, the screw parameters are directly derived from the $4 \times 4$ homogeneous matrices that store displacements in PTools (see chapter 5).

\(^1\)The screw object and function initially implemented in PTools had not been used until the present work.
Figure 8.1: Scheme of the screw transformation between two monomers A and B. The screw axis is defined by point P and vector Ω. The transformation from monomer A to monomer B is the combination of a rotation $\theta$ around the axis and a translation $\text{trans}$ along the axis.

If $M$ is a $4 \times 4$ homogeneous matrix storing the displacement of a rigid body,

$$
M = \begin{pmatrix}
  x_x & x_y & x_z & q_x \\
  y_x & y_y & y_z & q_y \\
  z_x & z_y & z_z & q_z \\
  0 & 0 & 0 & 1
\end{pmatrix}
$$

the unit vector $\vec{\Omega}$ of the axis in fig. 8.1, which is invariant when applying the screw transformation, can be directly calculated from matrix $M$ as its unit eigen-vector by solving the equation:

$$
M \vec{\Omega} = \vec{\Omega}
$$

Once $\vec{\Omega}$ is known, the value of the translation $\text{trans}$ can be obtained as follows

$$
\text{trans} = \overrightarrow{OQ} \cdot \vec{\Omega}
$$

where $\overrightarrow{OQ}$ is the vector with coordinates $(q_x, q_y, q_z)$ and O is the reference frame center.

Point P (or vector $\overrightarrow{OP}$) can then be computed using the relation

$$
\overrightarrow{MOP} = \overrightarrow{OP} + \text{trans} \cdot \vec{\Omega}
$$

Finally, the rotation angle $\theta$ is computed as follows, where $\vec{u}$ represents any
vector non aligned with $\vec{\Omega}$, and $\vec{v}$ is defined as

$$v = \frac{\vec{u} - (\vec{u} \cdot \vec{\Omega}) \vec{\Omega}}{\|\vec{u} - (\vec{u} \cdot \vec{\Omega}) \vec{\Omega}\|}$$

$$\cos \theta = \vec{v} \cdot M \vec{v}$$

$$\sin \theta = (\vec{\Omega} \times \vec{v}) \cdot M \vec{v}$$

Interestingly, Angelidis’ motivation in developing the geometry of hexanions\(^2\) was to be able to easily and accurately interpolate between two states of a transformation. This is used in several occasions during this work, for example to calculate the width of filament grooves (see section 8.6 in this chapter).

In PTools, screw transformations are stored in `screw` objects that contain a point and a unit vector to define the axis, a rotation angle and a translation value. A screw transformation $hp$ can be obtained from a $4 \times 4$ matrix using the `MatTrans2screw` function within a python script.

```python
from ptools import *
hp = MatTrans2screw(matrix)
```

Alternatively, application of the screw transformation $hp$ to a given molecule `mol` typically follows the python command lines:

```python
from ptools import *
mol.ABrotate(hp.point,hp.point+hp.unitVector,hp.angle)
mol.Translate(hp.unitVector * hp.normtranslation)
```

The analytical description of Angelidis conveys a high precision to these transformations. For example, we ran a verification test on two consecutive monomers `2GLS_A.pdb` and `2GLS_B.pdb` of the cyclic hexamer with PDB code 2GLS, which consisted to calculate the screw parameters relating `2GLS_B.pdb` to `2GLS_A.pdb` and apply them back to `2GLS_A.pdb` (using the above written python commands) to generate a calculated geometry. The result confirmed the method accuracy to $10^{-13}$ Å RMSD between the original and the calculated monomer Cα coordinates.

### 8.2 Construction : 2GLS

Both analysis and construction with Heligeom require having the three-dimensional structure of two successive monomers of the filament under

\(^2\)Hexanions, which specifically represent screws, are defined by Angelidis as the logarithm of the rigid body transformation matrix; algebraic properties in the (real) 6D hexanion space are described in reference [166], available at http://www.dgp.toronto.edu/~silex/Publications/hexanions.pdf.
study. In this example, we analyze the interface between two successive
chains A and B from hexameric glutamine synthetase, with PDB code 2GLS
[167]), which were first extracted into two individual files 2GLS_A.pdb and
2GLS_B.pdb.

The components of the screw transformation relating these two monomers
can be obtained using the command:

```
python heligeom.py 2GLS_A.pdb 2GLS_B.pdb
```

which outputs the following information:

- \( P: 0.000 -0.000 0.000 \)
- \( \omega: -0.000 0.000 1.000 \)
- \( \theta: \text{radian:} -1.047 \text{degree:} -60.000 \)
- \( \text{trans:} 0.000 \)

- monomers per turn: 6.000
- pitch: 0.002
- radius: internal: 12.200 external: 73.590
- direction: \( R \)

The four initial lines describe the screw motion. The point \( P \) and the
vector \( \Omega \) define the rotation axis; the rotation angle \( \theta \) (in degree and radian)
and the translation \( \text{trans} \) (in Å) define the transformation that needs to be
applied along the axis to go from one monomer to the next one. These four
components of the screw transformation are illustrated in Fig. 8.1.

In addition, Heligeom provides information on the global geometry of
the resulting filament: the number of monomers per turn, the pitch (the
distance between the monomers separated by a complete turn) and the
direction of rotation, which is either left-handed (\( L \)) or right-handed (\( R \)).

These parameters are calculated from the screw parameters as follows

\[
N = \frac{360}{\theta}
\]

\[
pitch = N \times \text{trans}
\]

\[
dir = \begin{cases} 
R & \text{if } (\theta \times \text{trans}) > 0 \\
L & \text{if } (\theta \times \text{trans}) < 0 
\end{cases}
\]

where \( N \) is the number of monomers per turn and \( dir \) the helical hand-
edness.

Once the helical information has been extracted, constructing an ex-
tended oligomer of any desired length is straightforward by using the pre-
ceding command with the total number of desired helical units (here, 6
8.3. CONSTRUCTION AROUND A CURVED AXIS

monomers) added at the end of the command line, followed by output redirection to create a new PDB file:

```
python heligeom.py 2GLS_A.pdb 2GLS_B.pdb 6 > 2GLS_new.pdb
```

The result for the ring-shaped 2GLS example is shown in Fig. 8.2.

![Circular filament of 2GLS reconstructed with Heligeom](image)

**Figure 8.2:** Circular filament of 2GLS reconstructed with Heligeom. The hexameric ring structure of the glutamine synthetase (PDB code 2GLS) has been reconstructed from two monomers A and B (in cyan and red respectively).

By default, the filament/ring is constructed around the axis defined by the point \(P\) and the direction \(\Omega\). It is possible however to align the helix of the generated filament along the \(Z\) axis by specifying the \(-Z\) option. Note that there is no theoretical limit to the number of generated units.

8.3 Construction around a curved axis

Heligeom offers the possibility to form an helix along an arbitrary axis. For example, let us take two molecular units \(\text{mol1.pdb}\) and \(\text{mol2.pdb}\) that share a common interface. The molecular units may be monomers of a same protein, or else adjacent nucleotides, whose regular association following the binding geometry defined by \(\text{mol1.pdb}\) and \(\text{mol2.pdb}\) is a helix.

First, an arbitrary axis has to be generated, formed by a collection of points distributed every \(\sim 1\ \text{Å}\) and stored in a PDB file. Below is an example of a PTools script to generate such axis, called \(\text{genAxis.py}\):

```
from ptools import *

r = RigidBody()

r.AddAtom(Atom(Atomproperty(), Coord3D(20, 0, 0)))
```

Figure 8.3: Heligeom construction around a curved axis.

\[ m = (\text{Rise}(0.7) + \text{Roll}(0.1)).\text{getMatrix()} \]

for i in \text{xrange}(0, 2000):
    r.\text{ApplyMatrix}(m)
    print r.\text{PrintPDB()},

The axis file can be generated with the command

\text{python genAxis.py > axis.pdb}

Construction of the oligomer along this axis following the binding geometry defined by the two given molecular units, via local application of the corresponding screw transformation along locally linearized axis segments, can be achieved using the script \text{buildProteinAlongAnAxis.py}

\text{python buildProteinAlongAnAxis.py axis.pdb mol1.pdb mol2.pdb > new.pdb}

Optionally, an angle can be specified to operate a global rotation around the axis (which means here rotating \text{mol1.pdb} before building the deformed helix). The angle is given in degrees.

\text{python buildProteinAlongAnAxis.py axis.pdb mol1.pdb mol2.pdb 180 > new2.pdb}

Fig. 8.3 shows the construction result on a curved axis (left) when the molecular units are DNA base pairs in B-form reference geometry (middle) or monomers of the RecA protein (right).

8.4 Coupling Heligeom with PyAttract

Heligeom is in itself a very convenient construction and analysis tool for molecular assemblies with already known helical organization; chapter 9 will present some examples where it can bring new insight into biological systems.

It can also become the central element of an integrative approach to predicting and modeling filamentous assemblies when coupled to the exploration of interaction geometries at the monomer-monomer level. In the work presented in chapter 10, we have coupled Heligeom to the PyAttract module of PTools described in chapter 5. The coupling implied the development of a number of scripts that have been incorporated into the Heligeom suite. For example, the script `extractHelicalParameters.py` extracts the screw parameters from each PyAttract docking output; the script `filterHelicalParameters.py` filters out docking geometries that fulfill user-defined criteria in terms of helical parameters; the script `extractHelicalModel.py` constructs one or more helical turns of the assembly corresponding to a given PyAttract output. Detailed description of the use of these scripts in the context of the study of RecA filaments (chapter 10) can be found in Appendix B.

The PyAttract docking output can also be post-processed to select only those interaction geometries that give rise to a sterically viable regular assembly, and to optimize near-cyclic geometries to the closest ring geometries (see next section and fig. 10.2 for our definition of near-cyclic geometries). This is done using the script `extractAndFilter.py`. The next section describes the method used to optimize the near-cyclic geometries.

8.5 Optimizing ring geometries

We consider that binding geometries resulting from docking simulations correspond to a cyclic organization when the number of monomers per turn $N$ differs by less than 0.1 Å from an integer, and when the pitch value per interface pitch / $(N-1)$ is lower than 0.5 Å.

When $N$ and pitch do not meet these conditions, but the pitch value per interface pitch / $(N-1)$ is lower than 5 Å, we attempt that the helix geometry resulting from regular monomer association following this binding geometry is sufficiently close to a ring geometry that it can be adjusted to perfect circularity with only minor adjustment (“Near-cyclic” organization).

Adjustment is systematically attempted to match each of two cyclic organizations, comprising either $M$ or $M+1$ monomers, where $M$ is an integer such that $M \leq N < M+1$. It is readily performed within the formalism associated with screw transformations. Given the proximity of the target ring structure to the initial screw transformation, we conserve the axis $\Delta$.
associated with that screw transformation. We call receptor the reference monomer in the docking simulation, ligand its binding monomer in the considered binding geometry. \( C_0 \) is the center of mass of the receptor and \( C_1 \) that of the ligand. \( R_i \) is the initial distance between \( C_0 \) and the axis, and \( \theta_i \) is the initial value of the rotation around the axis to go from the receptor to the ligand. The cyclic geometry is obtained simply by setting the \( \text{trans} \) value to 0 and the target angle \( \theta_T \) to \( 2\pi/M \) for a \( M \)-member ring or \( 2\pi/(M+1) \) for a \( M+1 \)-member ring. Distance \( R_i \) is accordingly modified to value \( R_T \) in such a way that the \((C_0, C_1)\) distance is conserved between the initial and target geometries (see scheme in fig. 8.4).

\[
R_T \sin(\theta_T/2) = R_i \sin(\theta_i/2)
\]

therefore, \( R_T = R_i \sin(\theta_i/2) / \sin(\theta_T/2) \)

Starting from the new distance \( R_T \), we adjust the position and orientation of the receptor with respect to the fixed axis by performing a series of 1000 steps of Monte Carlo simulation where the receptor is displaced along the radial axis and rotated around three orthogonal axes centered on \( C_0 \). Trial radial displacements are taken uniformly in the range \([-3 \text{ Å}, 3 \text{ Å}] \) and angular variations in the range \([-5^\circ, 5^\circ] \) in each of the three directions.

### 8.6 Measuring protein filament groove width

The groove of a protein filament can be defined as the solvent accessible volume between two consecutive helix turns. There are several ways to characterize the groove, either as a helical cavity with measurable dimensions or by focusing on its function within the filament. Indeed, the groove
8.6. MEASURING PROTEIN FILAMENT GROOVE WIDTH

constitutes a privileged binding site for accessory proteins or ligands and its topology regulates the accessibility to the filament core. Here, we define the groove width as the smallest distance permitting a protein or ligand to penetrate inside the helix. Since the groove width presents local variations, we perform the measurement along a whole helix turn. The procedure is detailed below.

We define the outer $R$ and inside $r$ radii of the protein filament as the maximum and minimum distances of protein atoms to the helix axis, respectively (fig. 8.5). First, a line is drawn between the center of mass of a monomer and its projection on the helix axis. On this line, reference points are taken every 1 Å between the outer radius $R$ and the mid point between the inner and outer radii, $(r+R)/2$. The set of reference points is translated by half a pitch so that they are positioned approximately at the center of the groove. For each reference point, a new line is drawn parallel to the helix axis. The maximum diameter of the sphere centered on this line is then computed. The groove width is locally defined as the minimum value of the set of diameters computed from the set of reference points. It represents the maximum possible size of a locally inserted sphere. We construct a similar set of reference points at a next step by interpolating the screw movement of the helix to a half degree. Fig. 8.5 shows a complete set of reference points that will be used to measure the groove width of RecA fibers in chapter 10. In that way, we obtain a measure of the groove width for each half degree of a complete turn. This data can then be displayed graphically as a simple line plot (an example can be seen in fig. 10.11, chapter 10). This method has been implemented and will be found in the next release of PTools.
Figure 8.5: Groove width measurement. (A) construction of a set of reference points used to measure the groove width of a protein filament. $r$ and $R$ are respectively the inside and outer radii of the filament, reference points are taken every 1 Å between $(r + R)/2$ and $R$; (B) representation of the set of reference points (red points) used to compute the groove width of a RecA filament. Groove width values are calculated every half degree.
Part III

Applications
Chapter 9

Heligeom application : Construction

One of the opportunities offered by Heligeom is to build, analyze and investigate bigger molecular models than usual. This chapter will present some of the advantages coming with this change of scale.

9.1 Extending modeling scales

One of the first interests in extending molecular models to bigger sizes is that this allows to immediately notice the effects of locally modifying the geometry of monomer-monomer association on the global shape of the resulting assembly. Even small variations of the geometry of association within a regular assembly can lead to dramatic changes. Fig. 9.1 shows an example where two close binding geometries give rise to either a ring or a helix shapes.

The geometry of a given helical form can also be modified by locally introducing irregularities, for example in the form of an alternative binding modes. As shown in fig. 9.2 (left) the effect of such introduction may seem limited when only a small size filament is considered. Once it is extended to ~100 monomers (right), this filament will form a superhelix (a helix that is itself coiled into a helix).

Furthermore, knowing the shape resulting from a specific arrangement offers little to no insight on which geometry close arrangements will adopt, as illustrated in fig. 9.3
Figure 9.1: Overlapping helical and cyclic forms of Dmc1. Comparison between the cyclic octameric form of Dmc1 (1V5W in green), and a helical form obtained using ATTRACT/Heligeom (violet). Cartoon representations show the “receptor” and “ligand” monomers, both in yellow for the cyclic octamer and in yellow and violet respectively for the helical form. The two forms are superimposed using the receptor monomer. The two assembly modes share 61% of monomer-monomer contact pairs. Their interaction energies are within 1 RT, and Co-RMSD calculated for their ligand monomers is 5.5 Å. The number of monomers per turn $N = 7.4$ and pitch $P = 79.0$ Å characterizing the right-handed helical form are compatible with geometries seen in filaments active in homologous recombination.
Figure 9.2: Changing scales: Models of RecA filament with alternating ATP- and ADP-form interfaces. The monomers with an ADP-type upper interface are represented in green, those with an ATP-type upper interface in cyan. The filament, principally in the ATP-form, presents ADP-type interfaces periodically distributed every 6 monomers. represented on the left is a fragment of about ten monomers and on the right a filament of about a hundred monomers.
Figure 9.3: Hyperstructure variability. Three nucleoprotein superhelices were constructed from helical units with close composition. The central filament (T5D1 in the table) corresponds to the filament shown in fig. 9.2. The filaments on the left and right follow the same organization, i.e. they are principally built in the ATP-form, but present an ADP-type interface periodically distributed every 5 (left, T4D1) and 7 (right, T6D1) monomers instead of 6 in T5D1. In all three cases, the left view shows the whole superhelix while the right view shows one helical unit of the protein component together with the single-stranded DNA the filament is built on. The helical characteristics of each form are given in the bottom table.
9.2 Interactions between topologically linear assemblies

Another interesting property of big models is that their construction by itself sometimes allows discovering unexpected results, without any extra work. As an example, fig. 9.4 shows the result of a docking simulation of a bent double-stranded DNA in a coarse-grained representation and a RecA filament of five monomers directly taken from the 3CMW PDB file.

![Small size complex between DNA and RecA filament.](image)

**Figure 9.4:** Small size complex between DNA and RecA filament. The RecA filament structure (five monomers) is directly taken from PDB file 3CMW; the DNA has been constructed from the PDB file 1HRY and B-form DNA fragments have been added at both extremities.

Simply by extending the RecA filament (as shown on fig. 9.5), a new contact appears between the docked DNA and a new region of the filament.
(highlighted lysine in fig. 9.5). This contact was judged of interest by our experimentalists collaborators.

**Figure 9.5:** Large size complex between DNA and RecA filament. Same as fig. 9.4 except that the size of the RecA filament has been increased using Heligeom. The newly formed protein-DNA contact is shown within a red circle.

### 9.3 Structural interpretation of low resolution data

It is also possible to build models of big macromolecular assemblies based on experimental “low resolution” data. Fig. 9.6 shows a picture made by atomic force microscopy and the resulting molecular model. As for now, the two strands result from simple construction following the geometric characteristics specified by Shi and collaborators [106], but the optimization of the
Figure 9.6: View of one turn of a RecA negatively supercoiled filament at atomic resolution. The structure has been constructed following the geometric characteristics described by Shi and collaborators [106], with a pitch of 160 nm for the supercoil (the script is written in Appendix C). The axes of the two strands (in cyan and green, respectively) are separated by 110 Å. The DNA incorporated in each of the two RecA filaments is in red.
interface between the two strands should be the object of future work.

9.4 Transferring information from the local to the global levels

The construction of long polymers can also be used to extrapolate locally calculated properties to long filaments by taking advantage of helical symmetry. Fig. 9.7 shows the electrostatic potential around an actin filament calculated using this approach. Both the filament and the electrostatic potential map have been generated using screw transformations, starting from the refined central units of a model of F-actin [168] and the electrostatic potential calculated on points distributed with helical symmetry around the axis (fig. 9.7).

![Figure 9.7: Electrostatic potential around an actin filament obtained using Heligeom. The Y-axis slice is colored from blue to red as a function of potential values ranging from -8 to 0 kcal/mol.](image)

9.5 Beyond the topologically linear forms

The PTools/Heligeom suite hosts all necessary functions to construct oligomers with composite interfaces, using a combination of screw transformations. This includes symmetric (i.e. capsids) or completely dissymmetric assemblies. An example of a reconstruction of the virus capsid with PDB code 1F2N, starting from four selected protein units, is shown in fig. 9.8. The figure illustrates the corresponding construction steps, given that the four selected units define three distinct interfaces in terms of screw transformations, all of which are cyclic symmetries.
Figure 9.8: Construction steps of a virus capsid starting from four interacting units. (left) the four units are shown in surface representation. Each of the three interfaces (schematized by colored lines, cyan, green and purple) correspond to different screw transformations. The characteristics of these transformations in terms of number of monomers per turn (N) and pitch (P) are indicated. (from left to right) the construction steps labelled (i) to (iv) respectively correspond to applying (i) a 3-order cyclic symmetry; (ii) a 5-order cyclic symmetry applied on the trimer resulting from (i); (iii) a 2-order cyclic symmetry applied on the pentamer resulting from (ii); (iv) the same transformation successfully applied to each interface represented by broken purple lines. To ease the understanding, the trimer that has been constructed in (i) is lined by a black triangle. Note that the result from step (iv) has been scaled down.
Chapter 10

Exploring protein filaments

The following chapter presents results submitted to the journal PLOS computational biology. In this work, we have coupled the Heligeom construction and analysis method to methods for sampling the binding geometries between two monomers of the RecA protein. This permits to relate the local level of monomer-monomer association to the global level of supramolecular assembly. Figs. 10.1 and 10.2 give an overview of the approach.

Note that the supplementary information of the submitted article have been included in different parts of the present manuscript. Specific methods have been included in the Methods part (chapter 8), complementary results are reported in section 10.8 at the end of this chapter and protocols are given in Appendix B and C. Footnotes indicate where references to supplementary information can be found.
Figure 10.1: Overview of the Heligeom/PyAttract integrated approach. Binding geometries generated by a PTools/ATTRACT coarse-grained docking simulation are analyzed with Heligeom in terms of the helical parameters of regular assemblies that can be generated with them. The results are filtered based on relative energies and geometry considerations (see fig. 10.2). Binding geometries leading to near-cyclic organizations with steric clashes are submitted to an optimization process in which they are adjusted towards the two closest cyclic geometries (section 8.5 in chapter 8). Binding geometries leading to steric clashes that are also not in the near-cyclic category are currently not analyzed further but will be adjusted to helical organization in future developments. Heligeom may be used for any of the final structures to construct filamentous or cyclic structures for further analysis.
Figure 10.2: Scheme of geometric filtering. After energy filtering, binding geometries issued from ATTRACT docking simulations are characterized by Heligeom in terms of the number of monomers per turn \( (N) \) and pitch value \( (P) \) of associated regular assemblies, and are separated into “Filament” (blue), “Cyclic” (green), “Near-cyclic” (orange) or Near-helical” (red) categories based on their corresponding position in the plot of \( P \) versus \( N \) schematized here. In the present work, the horizontal dotted line separating the “Filament” from the other categories has been set at \( P = 2R_M \), where \( R_M \) is the maximum radius of the monomer. The green boxed areas correspond to “Cyclic” geometries centered on integral values of \( N \) 0.1 with allowed values of \( P \leq 0.5 \) Å. “Near-cyclic” geometries are defined by accepting a pitch error of up to 5 Å per interface, and thus \( P < (N - 1) \times 5 \) Å. These geometries are shunted to an automated Monte Carlo energy-minimization (adjustment) procedure in which cyclic geometry is enforced.
10.1 Abstract

Oligomeric macromolecules in the cell self-organize into a wide variety of geometrical motifs such as helices, rings or linear filaments. The recombinase proteins involved in homologous recombination present many such assembly motifs. Here, we examine in particular the polymorphic characteristics of RecA, the most studied member of the recombinase family, using an integrative approach that relates local modes of monomer/monomer association to the global architecture of their screw-type organization. In our approach, local modes of association are sampled via docking or Monte Carlo simulation. This enables shedding new light on fiber morphologies that may be adopted by the RecA protein. Two distinct RecA helical morphologies, the so-called “extended” and “compressed” forms, are known to play a role in homologous recombination. We investigate the variability within each form in terms of helical parameters and steric accessibility including groove width. We also address possible helical discontinuities in RecA filaments due to multiple monomer-monomer association modes. By relating the local level of interface organization to the global level of filament morphology, the strategies developed here to study RecA self-assembly are particularly well suited to other DNA-binding proteins and to filamentous protein assemblies in general.
10.2 Author Summary

Organisms rely on proteins organized into regular assemblies such as helices in order to carry out different functions. For the individual proteins themselves, structural information is often obtained at the atomic level through experimental techniques such as crystallography or electron microscopy. Docking methodologies have been developed to extrapolate these structures to higher level organizations, but are typically applied to predicting binary complexes. In this work, we describe an approach to modeling regular fiber assemblies that incorporates docking of monomeric protein structures, along with optional Monte Carlo energy-minimization, and a coupled analysis of fiber morphology through the use of a new analytical tool, Heligeom. This approach is applied to the intriguing case of RecA, representative of a family of proteins essential for DNA repair that are found throughout the biological kingdoms. Individual RecA proteins self-interact in various helical or possibly ring assemblies that can interact with the DNA, and which exhibit both regular and mixed modes of association. Leveraging results from both crystallography and docking simulations, Heligeom allows us to reproduce and interpret known high-resolution information on RecA modes of association and offers new insights into observations obtained at a variety of resolutions.

10.3 Introduction

The organization of biological objects as multimers, and specifically as symmetric multimers, is the norm rather than the exception in cells. In an instructive review [19], Goodsell and Olson listed possible ways that proteins self-organize in cells and suggested why such association modes provide favorable options for proteins to exert their function.

Among the possible organizations, helical symmetries are particularly well represented. This type of organization, which incorporates the characteristics of a rigid body displacement, can be described as a rotation of the body around a particular axis combined with a translation along this axis, and is referred to as screw movement [164]. When repeatedly applied to positioning a monomer with respect to the preceding one, screw transformations produce helices, as well as cyclic assemblies or linear arrangements in case the translation or the rotation, respectively, is null. Helical organizations appeared naturally in the search for regular structures in biological macromolecules, and led Linus Pauling in 1951 to predict how proteins are organized in terms of helices [169] or quasi-straight segments such as β-sheets [170]. Soon after, the helical structures of DNA in the A, B or Z forms provided additional examples [171,172].

Beyond the level of secondary structure, screw organization is widely encountered in the world of homo-oligomeric or polymeric protein associ-
10.3. **INTRODUCTION**

RecA and RecB are the best characterized examples of RecA-family proteins, where copies of the same protein assemble in organized quaternary structures that can attain impressive sizes. Recombinase proteins are a particularly relevant example. These proteins are involved in homologous recombination (HR) [85], where they catalyze the faithful repair of DNA double strand breaks in a process that is common to all realms of life [86]. For this purpose they interact with DNA molecules, either in the form of cyclic assemblies [108,110] or as long filaments [173]. In the latter case, their organization as right-handed helices reflects the secondary structure of associated DNA, which is stretched by 50% [115] and unwound by 40% [116] with respect to standard B-DNA. Observed pitch variations between filaments of RecA-ATP (often called “active”, or “extended” form) and RecA-ADP (so-called “inactive”, or “compressed” form) have fueled the debate on the role of DNA stretching in the HR mechanism [107, 174, 175]. DNA-free forms of association have also been observed, involving dimers or hexameric rings for the prokaryotic RecA [105, 106], octameric rings for eukaryotic recombinases Rad51 or Dmc1 [108, 110] or even left-handed helices for the yeast RadA [111]. These observations were obtained using a large panel of methods including electron microscopy (EM) [99, 105]), atomic force microscopy (AFM) [106, 110] or X-ray crystallography [29,30,111,112].

Attempts have been made to establish a relationship between these diverse forms of RecA association. For example, interconvertibility between the extended and compressed forms of the RecA filament, which long appeared unlikely [98], has been demonstrated in recent years using single molecule experiments [131,176]. In the same way, identification of two possible DNA-binding forms for RadA and Dmc1, as stacked octamers or as helical filaments [108,110], raises the question of a possible interconversion between them. Wang and collaborators have proposed that the passage between different forms of recombinase filaments may play a role in the HR mechanism of DNA strand exchange [177]. Their discussion centered on the N-terminal domain of RecA as a possible key element to relate the different forms, based on structural observations and biochemical analysis revealing its fundamental role in the HR process [178].

In principle, known crystal structures of the protein components involved in such complex biological processes should help us model the structures of the different assemblies that have been either proposed or observed at low resolution. Multidocking methods have been developed to treat just this type of problem. In general, however, the problem is highly complex, as supramolecular assemblies may be characterized by multiple, simultaneous component interactions, i.e. in which each component presents different interfaces with different partners. Depending on the number and types of interfaces, this may lead to a quaternary structure of fixed size, or to an elongating system as seen for example in bundled actin protofilaments. Homo-oligomeric structures have been obtained via specific multiple docking methods [179–184]. These predictions generally take advantage of known
Cₙ or Dₙ symmetry, either directly, by applying symmetry conditions during the conformational search (thus reducing the number of degrees of freedom searched), or indirectly, by filtering out the results that fulfilled the desired symmetry criteria. More complex assemblies, which may not be symmetrical, have also been approached [134, 179, 185–188]. In each of these cases the overall architecture is highly constrained by the steric requirements of the multiple interfaces.

On the other hand, the recombinase systems we address here are not as constrained as in the general case, as they tend to form filamentous assemblies that grow using only a single monomer interface. Beyond recombination, such organization is found in other DNA processing systems such as replication (DnaA) but also in the protofilament building blocks of cytoskeleton fibers. In line with early work of Eisenstein and coll. [179], we address here the question of fiber assembly from the point of view of two-component modes of interaction. We note in particular that each favorable monomer-monomer binding geometry defines a unique helical or cyclic organization, or mode of self-assembly. Such filaments may demonstrate considerable diversity, in which each distinct interface geometry gives rise to a structural family. Within a family, slight modifications of the interface can result in significant changes at the level of the overall filament morphology. Further, different interface geometries may be found in different regions of a filament.

The integrated approach we present allows investigating these different sources of morphological variation, and combines interface sampling and construction using simple mathematical concepts such as screw transformations to characterize putative binding modes. This makes use of a set of recently developed computational tools, called Heligeom, which aim at characterizing, manipulating and assembling structural units with a screw organization, and in which the structural units may be individual proteins or protein hetero-multimers (Boyer et al., in preparation). Heligeom relies on the structures of monomer-monomer interfaces both for deriving the transformations and for filament construction; for the latter it is thus complementary to other packages that apply known space group symmetries to obtain the structures of supra-assemblies (see for example the web servers PQS [189], PITA [190] and PISA [191]). Because Heligeom is bundled with the freely available Python/C++ library PTools [134, 136], our approach can therefore benefit from existing PTools functionalities such as coarse graining, energy calculations, and diverse sampling protocols, which can be combined to arrive at novel strategies for investigating helical assemblies.

Our target in this article is RecA protein assemblies, where multiple association states have been documented. Although more complex systems may also be treated using these tools, we show that this approach is well adapted for modelling a variety of fiber morphologies that correspond remarkably well to both known and proposed forms of recombinase protein
assemblies.

10.4 Results

Two distinct forms of RecA filaments have been solved by crystallography. Both of them are right-handed helices, but they differ by pitch and monomer orientation within the filament. In particular, the so-called *compressed* form (*inactive* or RecA-ADP form, see Methods) (PDB code 2REB) has a pitch of 85 Å and 6 monomers per turn [29], while the *extended* form (*active* or RecA-ATP form) (PDB code 3CMW), which was co-crystallized with a homolog of ATP and with DNA, assembles with a pitch of 94 Å and 6.2 monomers per turn [30].

RecA monomers in the compressed and extended forms of the RecA fiber present notable geometry differences in the N-terminal domain (residues 1-37, called here the Nter flexible region). Also, loops L1 and L2 (residues 156-165 and 194-210, respectively) are absent in the RecA-ADP form. The \( \text{C}^\alpha \text{RMSD} \) between the two structures is 6.7 Å, but this mainly reflects the movement of the Nter flexible region. On excluding this region, the RMSD falls to 0.89 Å. This will be termed the “rigid core” region in the following (Fig. 10.3; see also Fig. 10.9A in section 10.8).

10.4.1 Characterizing known RecA assembly modes

The RecA-ADP and RecA-ATP crystal structures represent two distinct modes of RecA assembly. Indeed, after superimposing one monomer from each crystal structure dimer form, the \( \text{C}^\alpha \text{RMSD} \) of the second monomer is 22.6 Å. In order to place our exploration of RecA fiber morphologies in context, we characterized the assembly modes seen in these two structures. The surface area buried by different regions of the RecA protein in the compressed and extended structures is shown in Table 10.1.

In comparing the Buried Surface Area (BSA) values of the RecA-ADP (2REB) and RecA-ATP (3CMW) forms, it can be seen that the interface comprises both rigid and flexible regions of the interacting monomers. Table 10.1 also indicates that loops L1 and L2 contribute significantly to the interface area in the RecA-ATP form (and almost equally, data not shown), while in the RecA-ADP form the fact that the loops are disordered suggests that they do not stably contribute to the interface. The Cter domain was not seen to contribute to the interaction in either the compressed or extended forms. The contribution of the Nter flexible region is large, on the order of the surface buried by the rigid core region. We point out that the total area buried in the RecA fiber is quite high – 2800 to 4400 Å\(^2\). The BSA of the Nter flexible region alone, or of the core region alone, corresponds to the surface buried by one partner in a typical protein-protein interface (1600±400 Å\(^2\)) [192]. Thus, the formation of native interactions involving
either the Nter or the rigid core region alone may suffice to stabilize the initial form of the complex.

The similarity of two assembly modes can be quantified using the complementary measures $f_{\text{NAT}}$ and $f_{\text{IR}}$ (see Methods). The $f_{\text{NAT}}$ measures the fraction of contacting residue pairs (one residue from each monomer) that are shared between the two assembly modes. Considering the complete RecA protein, the $f_{\text{NAT}}$ between the RecA-ADP and RecA-ATP forms is 0.34; that is, 34% of the interface residue pairs are shared between the two modes. A complementary measure is $f_{\text{IR}}$, defined by the fraction of interface residues (and not residue pairs) on a given monomer that are shared between the two modes, and which is thus less stringent than $f_{\text{NAT}}$. The $f_{\text{IR}}$ values of the interface overlap between the two modes is between 63% and 96% for the complete RecA monomers.

We next determined the relative contributions of the rigid core region and the Nter flexible region to the $f_{\text{NAT}}$ and $f_{\text{IR}}$ values. In Fig. 10.3, panel
Table 10.1: Comparison of the 2REB and 3CMW interfaces.

<table>
<thead>
<tr>
<th>Form</th>
<th>Total</th>
<th>Flexible</th>
<th>Rigid</th>
<th>( f_{IR} )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Nter Loops</td>
<td>Central core</td>
<td></td>
</tr>
<tr>
<td>ADP</td>
<td>1400.3</td>
<td>709.0</td>
<td>691.3</td>
<td>0.71</td>
</tr>
<tr>
<td></td>
<td>1377.8</td>
<td>664.5</td>
<td>713.3</td>
<td>0.61</td>
</tr>
<tr>
<td>ATP</td>
<td>2225.5</td>
<td>895.1</td>
<td>943.1</td>
<td>0.59</td>
</tr>
<tr>
<td></td>
<td>2171.9</td>
<td>840.3</td>
<td>933.1</td>
<td>0.44</td>
</tr>
</tbody>
</table>

Contribution of flexible and rigid regions to monomer/monomer buried surface area (BSA) is calculated for the 2REB and 3CMW filament forms. Region boundaries were defined in order to minimize the RMSD difference between the two superposed rigid cores. The buried interface areas (\( \text{Å}^2 \)) of half/interfaces associated with each monomer A and B are reported in columns 2 to 5 and divided in Flexible (columns 3-4) or Rigid (column 5-6) interface components. Differences with respect to the Supplementary material of reference [30] reflect the different region definitions, in addition to the neglect of cofactor contributions here. Column 6 displays the fraction of residues \( f_{IR} \) on monomer A or B which belong to the interface of the rigid component in both 2REB and 3CMW forms.

A shows the rigid core interface in the RecA-ADP form in blue and panel B that of the RecA-ATP form in orange. The substantial overlap of the binding regions of the two forms seen in this figure corresponds to elevated \( f_{IR} \) values (44–71%) given in Table 10.1. However, no pairwise contacts between the rigid cores of adjoining monomers are conserved (\( f_{NAT} = 0 \)). This is due to the different relative orientations of the neighboring monomers in the two forms.

In the crystal structures, the N-terminal helix (residues 6-23), because of its flexible attachment to the core, can be seen to maintain essentially the same interactions with the adjacent monomer in both the compressed and extended forms of the fiber [30, 100, 178]. Wang and collaborators [111] have observed similar characteristics in crystal structures of RecA homologs. More exactly, comparing the N-terminal helix interaction between RecA fiber forms, the calculated fraction of pairwise contacts (\( f_{NAT} \)) is 0.9. We note that this region alone accounts for the overall total \( f_{NAT} \) value of 0.34. Because the relative orientation of the core region changes between the compressed and the extended forms of RecA, the long segment linking the N-terminal helix to the rigid core modifies its conformation in adapting to the geometry change.

Based on these observations, one can make the hypothesis that the quaternary organization of RecA oligomers, to a first approximation, relies on the interface between the rigid cores of adjacent monomers. In this hypoth-
Figure 10.4: RecA auto-assemblage. Modes of regular association of RecA monomers resulting from docking simulations with ATTRACT. Structures labelled A to E represent cyclic assemblies, structure F is a quasi-straight assembly, structure G is a left-handed helix and structures H, I are right-handed helices, very close to the 2REB [29] (H) and 3CMW [30] (I) crystal structures. Complete Heligeom characterization of all of the structures represented here are provided in Table 10.2.

esis, the flexible parts, along with mobile elements such as ATP, ADP or bound DNA, would enter as complementary components of the interaction to modulate the association affinity. We therefore limited the investigations presented below to the rigid core of RecA.

10.4.2 Investigating geometries of RecA autoassembly

In this work we have taken an integrative approach to the study of RecA filaments. This approach consists first in identifying favorable monomer-monomer binding modes and their possible deviations using docking simulations, followed by the construction and analysis of corresponding helical assemblies with our new tool Heligeom. An overview of the procedure can be found in Methods\(^1\).

\(^1\)The procedure is also detailed in Appendix B.

We first carried out docking simulations to explore the diversity of RecA association modes using the ATTRACT method [134–136]. The docking was restricted to the rigid core of the RecA monomers, which as we noted above does not vary more than 1 Å between the different known helix morphologies. Two docking runs were carried out using the rigid cores obtained from the two known RecA crystal structures 2REB or 3CMW, and
the results merged for analysis (see Methods). According to our hypothesis, exploring the interface between rigid cores should permit recovering the known RecA-ADP and RecA-ATP forms of association. We first checked that this was the case. The overall results of the docking are represented graphically in Fig. 10.3(C), in which each residue is colored as a function of the best interaction energy among all predicted interfaces involving it. The docking simulations predicted favorable interface regions that largely overlapped those characterizing the known interaction geometries (Fig. 10.3A and B), while sampling nearby alternatives as well. Indeed, the RecA-ADP and RecA-ATP forms were accurately predicted by the simulations, with \( f_{\text{NAT}} = 94\% \) and 89\% native pairs of amino acids recovered respectively in each case. In addition, the corresponding \( \text{C}_{\alpha}\)-RMSD, calculated after superposition of the first monomer of the corresponding dimers, was 1.0 Å for the RecA-ADP form (2REB) and 2.4 Å for the ATP form (3CMW). Moreover, these geometries were ranked among the best predictions of the docking simulations in terms of energy. Similar results were found for the rigid domains of Dmc1 and RadA, although with less favorable interface energies (data not shown). These results validated our approach and led us to examine in more detail alternative association geometries predicted by the docking simulation.

Each pairwise interface geometry resulting from the docking simulation corresponds to a unique form of RecA oligomer assuming regular association. We examined these via Heligeom, which was employed to automatically characterize the geometry of the oligomers in terms of pitch, direction of rotation and number of monomers per turn, and to construct the corresponding fiber of arbitrary, specified length. This allowed ascertaining if the association mode was consistent with regular helical or cyclic morphology, or if a steric clash was produced; near-cyclic assemblies corresponding to the latter case were optimized using symmetry constraints, while other conflicted geometries were set aside for the current study. Details of the procedures are given in Methods. The raw results for RecA and two other recombinase proteins, Dmc1 and RadA, are shown in Fig. 10.10 (section 10.8). Each point in these plots represents the number of monomers per turn \( N \) and pitch \( P \) for a distinct monomer-monomer association geometry; the optimized cyclic geometries are seen along the horizontal axis. Although globally similar results are obtained for the three systems, each presents a different overall signature.

Fig. 10.4 displays a gallery of fiber organizations for RecA association obtained from the simulations following the protocol in SI\(^2\). Complete Heligeom characterization of these binding modes, as well as of the crystal structures 2REB and 3CMW, is provided in Table 10.2. The assemblies represented in Fig. 10.4 were chosen to illustrate how variations in the mode

\(^2\)The protocol can be found in Appendix B.
CHAPTER 10. EXPLORING PROTEIN FILAMENTS

of RecA association with reasonable energies can result in drastically different geometric characteristics for the assembly, with all types of screw transformations, either cyclic (A-E), quasi-straight (F) or helical (G-I), being represented. Docking results characterized by a quasi-null pitch (A-D) gave way to dimers, trimers, pentamers or hexamers. An 18-mer quasi-ring structure (with an axial closure defect of only about 1.8 Å) was also obtained (structure E). Each of the forms A-I was verified to accommodate the pruned flexible regions, i.e. the two loops and the Nter flexible region, without steric clash (Fig. 10.8 and accompanying text in section 10.8). In addition, we examined whether the monomer region that binds the N-terminal helix in both 2REB and 3CMW was accessible for binding.

It can be emphasized that evidence for several of these predicted geometries has been observed experimentally – through atomic force microscopy [106] (forms A and D), and electronic microscopy [99,105] (forms D, H and I), in addition to crystallography [29, 30] for forms H and I as described above. We also identified left-handed helices with very good interaction energy values (geometry labeled G), which can be related to left-handed forms of RadA observed by electron microscopy and crystallography [111]. Wang and coworkers [79] have suggested that left-handed assembly may be a general property for RecA family proteins. On the other hand, the quasi-straight geometry F, ranked second in terms of interaction energy, has not been observed in natural fibers, perhaps due to its lack of compaction.

We note that helices H and I rank among the best energy predictions, which in some sense conforms to their similarity to the known binding modes. However, the interaction energy alone here cannot be used to rank the most likely structures in terms of probability, which is a function of the change in free energy of the interaction. As we have emphasized, only the rigid-core regions of the RecA monomers are used in these simulations. Also, as our interest is principally in filamentous assemblies, we focus on the interface energy only. Yet, for a given complex, the binding energy must include contributions from all interfaces present in the assemblage. Relatedly, locally favorable monomer-monomer association geometries may be inconsistent with viable regular fiber geometries. For example, the lowest interface energy obtained from the pairwise monomer docking would produce steric clashes between monomers of successive turns in a regular helix. Steric clashes in such cases could be resolved using multidocking techniques, just as we adjusted near-cyclic geometries to cyclic ones using energy minimization and symmetry constraints. However, even without adjustment, shorter stretches of fiber employing such interfaces could also play a role in mixed-mode fibers, as will be discussed below.

We compared the monomer interfaces for the different association modes in Fig. 10.4 in a pairwise manner. The $f_{\text{Nat}}$ comparison showed that only

3Structure X in Appendix B.
the interfaces associated with ring morphologies C and E overlap somewhat, sharing 15% of their residue contacts, while those associated with the quasi-
straight and left-handed helical filaments F and G share only about 3% of the
residue contacts. All the other interfaces are perfectly distinct (0% shared
contacts). The variability of the observed shapes presented in Fig. 10.4
therefore arises from the employment of distinct interfaces. However, it
may also be the case that slight modifications in the interface can produce
substantially different filament morphologies. An example of this is seen in
the case of Dmc1, in which a helical form shares 61% of interface contacts
with the octameric ring observed crystallographically, and with almost the
same interface energy\(^4\). This result is particularly interesting in the context
of possible interconversion between such forms in Dmc1 and RadA [108,110].

\begin{table}[h]
\centering
\caption{Comparison of RecA assembly modes}
\label{tab:reca_modes}
\begin{tabular}{|l|l|l|l|l|l|l|}
\hline
Mode\(^a\) & pitch\(^b\) & N & \text{radius}^b & \text{E}_{\text{int}}^c & f_{\text{NAT}}^d & \\
 & & & int : ext & & & 2REB 3CMW \\
\hline
A & 0.0 & 2.0 & 0.0 : 46.6 & -22.5 & 0.0 & 0.0 \\
B & 0.0 & 3.1 & 0.0 : 53.6 & -22.7 & 0.0 & 0.0 \\
C & 0.3 & 5.0 & 16.6 : 58.2 & -33.9 & 0.0 & 0.0 \\
D & 0.1 & 6.0 & 20.7 : 62.9 & -30.7 & 0.0 & 0.0 \\
E & 1.8 & 18.0 & 107.0 : 149.8 & -30.2 & 0.0 & 0.0 \\
F & 52.5 & 2.0 & 0.0 : 46.3 & -44.8 & 0.0 & 0.0 \\
G & 106.3 & 5.8 & 13.3 : 48.7 & -40.2 & 0.0 & 0.0 \\
H & 72.8 & 5.8 & 10.9 : 61.6 & -39.4 & 0.94 & 0.0 \\
I & 90.3 & 6.4 & 8.0 : 60.1 & -41.1 & 0.0 & 0.89 \\
2REB & 82.7 & 6.0 & 11.7 : 62.9 & -41.3\(^*\) & 1.0 & 0.0 \\
3CMW & 94.2 & 6.2 & 6.2 : 58.9 & -41.2\(^*\) & 0.0 & 1.0 \\
\hline
\end{tabular}
\end{table}

\(^a\) Each mode A-I describes one assembly type from ATTRACT docking experi-
ments restricted to the rigid core of the RecA monomer. Last two lines provide

\(^b\) Value in Å

\(^c\) Per-interface energy (RT units)

\(^d\) measured with respect to the rigid cores in 2REB and 3CMW, respectively.

\footnote{A representation of the two forms can be seen in fig. 9.1, chapter 9.}

10.4.3 Variability within selected families of binding modes
Coupling Heligeom to targeted docking and Monte Carlo (MC) exploration, we explored the variability of each of the two known binding modes RecA-ADP and RecA-ATP (see Methods) in the vicinity of particular interface geometries.

Figure 10.5: Exploration of the RecA-ADP and RecA-ATP structural families. Helical characteristics corresponding to the members of the ADP (left) and ATP (right) structural families sampled via targeted docking simulations and MC explorations are represented by the number of monomers per turn (N, vertical axis) versus the pitch values (horizontal axis, in Å). The results are colored according to energy values $E_{(RT)}$, with $E \leq -42$ (orange); $-42 < E \leq -40$ (red); $-40 < E \leq -38$ (green); $-38 < E \leq -36$ (blue); $E > -36$ (cyan). The arrows indicate MC sampling results starting from the 2REB (left) and 3CMW (right) binding modes as extracted from the crystal structures. Inserts show representative sampled structures for the RecA-ADP (left, cyan) and the RecA-ATP families (right, pink); their location in the $(P, N)$ space are indicated (numbers). Angular deviations (calculated using Heligeom) between the reference RecA-ADP binding geometry and the geometries labelled 1, 2, 3 of the same structural family (left) are respectively 14.2, 5.2 and 23.3°; in the case of RecA-ATP, the angular deviations between the reference geometry and the geometries labelled 1, 2, 3 (right) are respectively 5.1, 11.0 and 9.3°.

Within the results of the targeted docking simulations, 69 structures for RecA-ADP and 4 structures for RecA-ATP were obtained close to the experimentally known geometry; the structures are represented as blue crosses.
in Fig. 10.8, section 10.8. These structures were used as starting points for subsequent MC simulations. For RecA-ADP, 18 starting points out of the 69 selected structures sufficed to fully cover the entire space defined by the 69 structures. For RecA-ATP, all four selected structures were used as starting points for MC sampling.

Fig. 10.5 shows the sampled regions in terms of pitch and number of monomers per turn. For comparison, the figure also displays the results of MC simulations performed under the same conditions starting from the exact 2REB (left) and 3CMW (right) binding modes (arrows in Fig. 10.5). This procedure defined two families of right-handed helices whose members present a similar interface but with pitch values ranging from 45 to 160 Å for the ADP family or from 70 to 140 Å for the ATP family. The angular deviation of the “ligand” monomer in a given MC-sampled geometry was calculated by determining the screw transformation relating it to its starting position; in RecA-ADP this deviation reached 23° while in RecA-ATP it reached 11° (caption, Fig. 10.5). Selected members of the ADP and ATP families are represented in Fig. 10.5 (inserts).

The results are compatible with EM observations obtained by the Egelman group through three-dimensional reconstruction specific to helical polymers [99], in which the authors observed a large range of pitch values for helix families related to the compressed (ADP) or the extended (ATP) forms, with overlapping pitch values. It is interesting to note that the region sampled when starting from the ADP binding mode presents a minimum at a pitch value of 77.3 Å, (N= 5.98, E= −40.2 RT, $F_{\text{NAT}} = 0.98$), which is closer to EM observations [99] than the pitch of the crystal structure (82.7 Å). More generally, the most stable elements of the ADP family present pitch values below 80 Å.

Unexpectedly, the modes of helix distortion revealed by Fig. 10.5 within the ADP or the ATP families notably differ. For the ADP family, increase of pitch is accompanied by a roughly regular increase in the number of monomers per turn, indicating global unrolling/stretching of the fiber form (1 → 2 → 3 in Fig. 10.5, left panel). No such regularity is observed for the ATP family: the right panel shows steeper variations, with a slope that can be positive (pitch values above 100 Å) or negative (below 100 Å). This indicates that stretching (2 → 1 in Fig. 10.5, right panel) as well as compression (2 → 3) of RecA-ATP fiber forms with 100 Å pitch are accompanied by an increase of the number of monomers per turn— that is, helix unwinding.

10.4.4 Binding mode variations in a single fiber

Variations in the binding mode within a single fiber can also lead to a variety of changes in filament morphology. For example, in the negatively supercoiled filament shown in Fig. 10.6, consecutive interfaces differ slightly
Figure 10.6: View of one turn of a RecA negatively supercoiled filament at atomic resolution. The structure has been constructed following the geometric characteristics described by Shi and collaborators [106], with a pitch $P = 160$ nm for the the supercoil (see text). The axis of the two strands (in cyan and green, respectively) are separated by 110 Å. The DNA incorporated in each of the two RecA filaments is in red.

from each other, allowing the torsional deformation to be regularly distributed along the whole structure. The pictured filament was constructed using PTools/Heligem following observations obtained by atomic force microscopy [106]. Another example is the crystal structure of the RecA human homolog Rad51 determined by Conway and collaborators [193], where two slightly different binding modes have been reported to alternate along the helix.

In the same way, it may be envisioned that binding modes belonging to different structural families coexist within a single RecA fiber. Alternation between fiber regions presenting extended (RecA-ATP) or compressed (RecA-ADP) forms has been observed by electron microscopy [100]. Fig. 10.7 (A) shows a model of a RecA-ATP and RecA-ADP junction, obtained with PTools/Heligem by simply appending monomers using the different binding modes. The junction results in a $\sim 35^\circ$ kink, corresponding to the $\sim 35^\circ$ difference between monomer orientations with respect to the helix axis in structures 2REB and 3CMW. We note that kinks can indeed be observed in Fig. 1 of reference [100].

In Fig. 10.7 (B-C), we explore the effects of different combinations of RecA-ADP and RecA-ATP binding modes on the overall RecA fiber morphology. In the chimeric structure shown in Fig. 10.7 (B), a RecA-ADP interface is periodically inserted every six monomers in a filament otherwise in RecA-ATP form. This corresponds to a situation in which ATP molecules would be hydrolyzed every six monomers (which appears to be the case in active RecA filaments [126, 194]), with the monomer-monomer binding modes modified accordingly. The result is a negative superhelix with a pitch of 312 Å, an external radius of 182 Å and 59 monomers per turn. On the other hand, a regular fiber is obtained when RecA-ATP and RecA-ADP binding modes alternate evenly along the filament (Fig. 10.7 C). In

---

5Details of the construction are given in Appendix C.
10.5 DISCUSSION

Figure 10.7: Models of RecA filaments with alternating binding modes. Consecutive monomers are represented in surface representation with different colors. A. The model features a junction between ATP-form regions (monomers 1-18) and ADP-form regions (18-36). B. The filament, principally in the ATP-form, presents an ADP-type interface between the 5th and 6th monomers. C. 24-mer filament with alternating ATP- and ADP-form interfaces. The monomers with an ADP-type upper interface are represented in green, those with an ATP-type upper interface in cyan.

this case, the resulting geometry is intermediate between 2REB and 3CMW helices (82 Å pitch and 6.4 monomers per turn). The main characteristic of this structure is a locally reduced and strongly variable groove width with respect to the RecA-ADP and RecA-ATP forms (Fig. 10.11, section 10.8)). We note from Fig. 10.11 that the ATP form itself presents large and regular groove accessibility, which may be related to its co-protease function during the SOS response [101,102,194]. Although no direct proof of the existence of filaments with alternating RecA-ATP and RecA-ADP interfaces has been published for RecA, such filaments represent putative transient intermediates during interconversion between ATP and ADP-forms of RecA filaments.

10.5 Discussion

In this study, we have taken an integrated approach to investigating RecA supramolecular assemblies, by sampling possible modes of association via docking and/or Monte Carlo simulations coupled with helical analysis us-
ing a new software tool, Heligeom, in the PTools modelling package. The first result is that interactions between rigid core regions of successive RecA monomers suffice to account for all known oligomeric forms of RecA assembly, including those that have been postulated but never solved at atomic resolution such as left-handed helices. An important role for the rigid core was expected because of its large contribution to the interface area in the 2REB and 3CMW forms (Table 10.1). However, because flexible regions contribute by the same amount, the central role of the rigid core needed to be established. The building-block role for the RecA rigid core region gives latitude to flexible and mobile regions of the protein, as well as external factors (ATP, ions, DNA), to modulate energetic preferences among binding modes. We have shown for example that the ATP cofactor stabilizes the 3CMW form of RecA association (Dataset S1, Supporting information, Figure S5). Some flexible regions may even actively control the passage between binding modes, as postulated by Chen and collaborators [111].

Our results also provide new understanding of the relationship between modes of RecA self-association. For example, we identified families of structures which can be considered close in terms of binding mode since they share interacting pairs of amino-acids (structures C and E in Fig. 10.4). The combined use of Monte Carlo sampling with Heligeom allowed us to explore the detailed variability of RecA filament morphology within structural families, each defined as an ensemble of binding modes with low interaction energy and a fraction of common contact pairs \( f_{\text{NAT}} \) higher than 50%. Again we caution, however, that our approach does not permit ordering the structural families in terms of their probability of formation, which would require evaluation of the free energy of interaction, and would thus necessitate taking into account not only the flexible regions but also the system composition (ion, cofactor and monomer concentrations).

Variability in the helical pitch of RecA filaments may also play an important role for binding DNA molecules. During the process of homologous recombination, RecA filaments are found bound to DNA in an extended (by 50%) and unwound (by 40%) form. This corresponds to a helix with 18 base pairs per turn and about 94 Å pitch. The recent work of Bosaeus and collaborators establishes that the DNA form in the RecA filaments corresponds to a metastable state of the double-stranded DNA stabilized by its interaction with the recombinase filament [114]. Small modifications of the DNA stretching may then direct the DNA into either a B-DNA state or a S-DNA state. The DNA characteristics can also be modified during the course of base pairing exchange. Adaptability of the protein filament geometry within a given family may allow the conservation of key interactions with the DNA during such events.

More generally, the variability within each binding mode enables an oligomer to absorb distortions arising from the environment. In any biological setting, RecA filaments are not expected to conserve perfect symmetry.
10.5. DISCUSSION

Thermal effects or external constraints due to dense packing in the cell introduce some degree of disorder. For example, in the 3CMW crystal structure, the binding modes locally defined by the five monomers (four interfaces) differ somewhat, with associated pitch values respectively comprised between 89.2 and 98.8 Å (Heligem analysis; the average value is 94.7 Å). RecA filaments also bend due to thermal fluctuations; the persistence length is estimated to be about 900 nm for RecA/DNA filaments [195] and 95 nm for DNA-free fibers [196]. In many cases, distortion induced by external constraints can be smoothly distributed over the whole structure, taking advantage of the local variations within the binding modes. An illustration is the response to torsional stress transmitted by bound DNA, as observed by Shi et al. [106] and modeled in Fig. 10.6. When the external stress is very large, or if regulating elements (cofactors, conformations of flexible regions) are not uniformly distributed along the fiber [194], different binding modes can be expected to coexist within single RecA oligomers. Examples of such multi-modal association have been documented in the case of ring assemblies [80], where they have been attributed to the effect of non-uniform cofactor hydrolysis [82]. As shown in Fig. 10.7 in the case of RecA, a large range of variation in terms of filament morphology can result simply from multi-modal association. In this respect we note that in several EM images, kinks, strong curvature and other irregularities can be observed (see for example Fig.3B in [132], Fig.1 in [100] or Fig.1 in [99]).

An advantage of using PTools/ATTRACT in this work is the possibility of easily incorporating coarse grained representations. In addition to sparing computer time, coarse-graining offers a simple way of smoothing the potential energy surface and implicitly accounting for small conformational changes at the interface. In earlier work it has been shown that the ATTRACT docking performance largely tolerates conformational changes of small and medium-sized side chains [135]. Although not used here, ATTRACT’s handling of side chain or even loop flexibility at the coarse-grained level allows selecting side-chain rotamers or loop substates during the docking process [135,148]. Side chain rotamer changes are particularly frequent upon formation of protein-protein interfaces [197]. If the conformations of long interfacial side-chains need to be optimized at the atomic level, one may resort to the methods recently developed by the Baker [198,199] or Redon [200] groups, which simultaneously optimize side chain and even main chain geometries within protein oligomers, together with the relative positioning of monomers. The two methods are specific to symmetric assemblies and take advantage of the symmetry to reduce the number of degrees of freedom. When symmetry is disrupted (for example due to axis curvature), one can use more general methods for optimization of interface packing as described for example in references [201,202]. Elements that have been locally optimized can then be re-injected into the supra-assembly using simple PTools superposition commands. Coarse grained representa-
tions are also compatible with the exploration of internal deformation of the monomers or concerted deformation of dimers, for example by following vibrational modes [145,146]. This aspect has not been directly considered in the present study but can be easily coupled to docking or Monte Carlo explorations in order to evaluate such effects on the global helical form. Such internal deformations may contribute significantly to pitch variations. Methods for taking into account flexibility may prove important in studying recombinases such as RadA or Dmc1 from higher organisms because of the proportionately larger contribution of their flexible N-terminal domains compared to RecA.

Our focus on interfaces makes it possible to construct non symmetrical morphologies and notably those combining different binding modes. This can be important given that biological processes often involve symmetry disruption. In this vein, we are currently investigating the morphology of RecA filaments presenting both ADP- and ATP-like interfaces in more detail. Finally, the results obtained for RecA support the coupled use of docking simulations and Heligeom processing for the interpretation of low resolution observations from EM or AFM on regular or irregular oligomeric assemblies. In its present state, the method is limited to studying oligomeric assemblies characterized by only a single monomer interface. Further addition of techniques from the multimolecular docking [134,179] should enable its future application to study the morphology of protofilament assemblies such as cytoskeleton fibers.

10.6 Methods

10.6.1 Overall approach

The approach we present combines the PyATTRACT and Heligeom modules of the PTools Python/C++ library [134], along with other functions of the library, in order to investigate the geometry of open oligomeric filaments (a schematic view of the overall approach is presented in Fig. 10.1). The process starts with the structure of a unique monomer of the system under study, which is reduced to coarse grain representation and then docked against itself using the ATTRACT method. The docking results are then processed by the Heligeom module to provide the helical parameters corresponding to regular assemblies based on each binding modes (see below). Post-processing allows conserving only the most favorable binding modes for regular self-assembly of the monomers. The generated structural families can then be investigated in terms of their internal variability. To this aim, the interface is sampled at a finer scale using a combination of targeted docking and Monte Carlo exploration.
10.6.2 Screw transformations with Heligeom

Heligeom is a Python module that interfaces with the Python/C++ library PTools [134, 136]. It is packaged with the latest version of PTools and includes a variety of scripts of varying complexity, including those developed for the present study (Boyer et al., manuscript in preparation).

The fundamental operation of Heligeom centers on the definition of the screw transformation. In general, the coordinates of a given monomeric unit can be derived from those of another through such a transformation [166], defined by the position $O$ and direction $\Omega$ of a screw axis, a rotation of angle $\theta$ around this axis and a translation value $\text{trans}$ parallel to the axis$^6$. Regular repetition of the screw transformation generically leads to a helical shape. Global parameters describing helix shape, i.e. the pitch ($P$), the number of monomers per turn ($N$) and the direction of rotation ($\text{dir}$), are derived from the screw parameters as follows:

\[
N = \frac{360}{\theta} \\
P = N \times \text{trans} \\
dir = \begin{cases} 
R & \text{if } (\theta \times \text{trans}) > 0 \\
L & \text{if } (\theta \times \text{trans}) < 0
\end{cases}
\]

The pitch and the number of monomers per turn can be directly compared to values extracted from electron microscopy images where available. Analysis of a monomer-monomer pair with Heligeom typically consists of using a single command to automatically extract the screw parameters (helix axis, rotation angle and translation) from the coordinates of two interacting protein monomers, via an analytical geometric calculation [134, 166]. The structural data for the interaction may have been obtained experimentally or else through modeling or docking studies. In addition to extracting screw parameters, the same Heligeom command can generate fiber structures of arbitrary length.

Heligeom can also be used to simulate the assembly of monomers along a non-linear path. Once the screw transformation has been defined from the structures of two interacting monomers, oligomeric assemblies can be reconstructed along any given curved axis$^7$.

In the present study, scripts using Heligeom have been developed, either as post-processing tools to extract screw parameters and build fiber models from ATTRACT docking simulation output, or to directly incorporate screw analysis into sampling procedures such as Monte Carlo exploration (see below).

$^6$See Figure 8.1 in chapter 8.
$^7$Details can be found in chapter 8.
10.6.3 PDB files

Coordinate data was obtained from the Protein Data Bank [203]. Two crystal structures of RecA were used, PDB code 2REB (space group $P_{6_1}$) with no bound DNA or cofactor, solved at a resolution of 2.3 Å [29] and 3CMW (5 monomers expressed as a single fusion protein, space group $P_{2_1}2_12$ crystallized with ADP-AlF$_4$-Mg, a non-hydrolyzable analog of ATP, in the presence of DNA, resolution of 2.8 Å) [30]. Structure 2REB is very similar to the structure with PDB code 1REA, which was obtained with ADP as a cofactor but only contains C$\alpha$ atoms; the Co-RMSD (root mean squared deviation) between the two entries is 0.3 Å. By extension, structure 2REB is referred to as the RecA-ADP (inactive, compressed) form and structure 3CMW as the RecA-ATP (active, extended) form. We used the 2REB asymmetric unit and the crystal symmetry information provided in the PDB file to construct the RecA-ADP filament form. In 3CMW, the five RecA units of the fusion protein are indexed by residue number: 1–333 for the first unit, 1001–1333 for the second, and so on. We separated these units into monomers 1 to 5 for this work. Since their 3D structures differ slightly from one another, we used monomers 2 and 3 to define the interaction mode of the RecA-ATP fiber, with monomer 3 as the reference structure.

RecA in the two crystal structures 2REB and 3CMW presents large geometry differences at three locations: the N-terminal domain (residues 1-37) and the L1 (residues 156-165) and the L2 (residues 194-210) loops. The two loops are disordered in the 2REB structure. These regions, together with terminal residues 1-5 and 329-333 which are disordered in 2REB, were pruned before rigid body docking was performed (next section). The remaining monomer core structures differ by less than 1 Å C$\alpha$-RMSD.

For Dmc1, we used PDB structure 1V5W [112] (resolution 3.2 Å, two monomer chains, space group $I4_{22}$). Terminal residues 1-83 and flexible loop residues 271-289 are missing in this structure, and the flexible N-terminal domain extends up to residue 98. Analyses were performed using the rigid core (residues 99-270, 290-340) of chain A.

For RadA ($Sulfolobus solfataricus$) we used two different filament morphologies, PDB codes 2Z43 [111] and 2ZUB [177]. In 2Z43 (resolution 1.93 Å, space group $P_{3_1}$) chain A was used to define helix geometry along with the crystal symmetry operations using the PISA server [191], recovering the right-handed filament with 3.0 monomers per turn and a pitch value of 99.4 Å. The rigid core residues (86-221,232-256,282-309) were used for the analyses. In structure 2ZUB (resolution 2.90 Å, space group $P_{2_1}2_12_1$), chain A was used, again with PISA, to define the filament geometry and again truncated to its rigid core region. The corresponding left-handed filament presents 5.7 monomers per turn and a helical pitch of 193.3 Å.
10.6.4 Calculating interface area contributions

The buried surface area (BSA) between two interacting monomers A and B was calculated using atomic solvent accessible surface area (ASA) values computed by NACCESS [204] with a default solvent probe radius of 1.4 Å. The total BSA = BSA_A + BSA_B includes the ASA lost from the two monomer surfaces upon formation of the complex. For monomer A, BSA_A = ASA_A - ASA_A(B), in which the subscript A(B) indicates monomer A in the presence of monomer B in the complex; BSA_B is defined correspondingly. Where indicated, individual residue i contributions BSA_i were calculated in the same way. The contributions of the different regions of the protein were obtained by repeating the BSA calculations before and after truncation of the concerned region and subtracting.

10.6.5 Sampling protein modes of interaction

Interaction modes were sampled via docking studies of the monomers, which produce solutions representing candidate pairwise interfaces ranked in terms of a scoring function or interaction energy [137, 138]. When combined with Heligeom, interface sampling enables the generation and exploration of a large number of filament morphologies corresponding to different interaction modes. We used the PTools/ATTRACT suite for performing and analyzing docking simulations using the ATTRACT protocol [135, 136]. ATTRACT uses a multi-minimization strategy and a reduced (coarse grained) representation (about four heavy atoms per grain) for target proteins and/or nucleic acids. The force field governing the docking process is composed of a Coulombic term screened with a distance-dependent dielectric function and a smooth, 6-8 van der Waals terms [135]. In the case of RecA monomers, a typical PTools/ATTRACT run took 7 hours on a single Intel Core 2 Duo running at 3 GHz, or minutes when the run was distributed on tens of processors (since PTools facilitates breaking the docking into independent jobs, the speed-up is essentially equal to the number of processors). Corresponding to analysis described in this study, several new functionalities were added to PTools/ATTRACT, coupling it to Heligeom analysis.

In the interface sampling study, the quasi-rigid core region of the RecA monomers was exclusively considered and treated as a rigid body, as discussed above and in Results, although limited flexible docking is possible with ATTRACT [136]. No qualitative differences were observed in simulation results obtained using the monomer structures from 2REB or 3CMW and the results of both simulations were merged for analysis. We performed two different types of coarse-grained docking simulations. In the first type of simulation, one RecA monomer (the ligand) was initially placed in a

---

8 see chapter 8 for a description of these functionalities and Appendix B for an example of their utilization.
quasi-spherical array of regularly distributed starting points and orientations around the other (the receptor), at a constant distance between the two proteins’ surfaces and with a 10 Å distance between neighboring points. This allowed global identification of potential modes of association. Additionally, targeted docking simulations were performed in order to explore the variability of the two known geometries of RecA-RecA association; here the starting points were positioned in a dense grid (2 Å separation between neighboring points) no farther than 20 Å from the known interfaces. The targeted docking simulation provided starting points for detailed characterization of selected binding modes (see below).

Two residues on different sides of the interface are considered to be in contact if any of their pseudo-atoms are within 7 Å [148]. Comparison of two modes of interaction was characterized by $f_{\text{NAT}}$, which is the fraction of residue-residue interface contacts in the first mode that are also present in the second and $f_{\text{IR}}$, calculated for each of the partner proteins, which is the fraction of interface residues in the first mode that are also in the interface in the second. Since we have two reference monomers 2REB and 3CMW whose side chain conformations may vary at the interface, we used either 2REB or 3CMW as reference depending on the starting monomer in the docking simulation. For more detailed graphical comparisons we also attributed to each residue of the receptor or ligand the best interaction energy of the interface or interfaces in which it was involved in the docking.

We note that the starting monomer structures in our simulations consist of the rigid core of the extended (RecA-ATP) or compressed (RecA-ADP) forms; our results could thus be construed globally as being “bound-bound”. However, the interface regions in these two assembly modes share no native contacts (see Table 10.2), and thus are largely “unbound” with respect to the other form. Indeed, when starting from the RecA-ADP structure we were able to generate binding modes close to the RecA-ATP one (RMSD of 4.8 Å, interface RMSD 4.2 Å, $f_{\text{NAT}} = 0.5$), and vice versa (RMSD of 4.8 Å, interface RMSD 2.8 Å, $f_{\text{NAT}} = 0.8$), with an energy difference of less than 4.2 RT from that of the true binding mode in both cases.

10.6.6 Processing and filtering the sampling results
Automatic extraction of screw parameters from a PTools/ATTRACT docking output file (about 50,000 poses) was readily performed with Heligem and took about 5 minutes on a 3 GHz Intel Core 2 Duo processor. Additional filtering of the binding modes was performed to select modes of association that are compatible with the formation of helical assemblies. A first selection criterion (local filter) is the docking interaction energy. We retained modes of association up to 20 RT above the reference form interaction energy, which allows uncertainty due to the absence of the flexible regions of
the monomer. We also eliminated docking results where the residues bordering flexible regions (RecA residues 38, 156, 165, 194, 210) were found in either of the two interfaces corresponding to a given binding mode (interfaces with the preceding and following monomers).

After initial screening, we applied a screen designed to identify binding modes suffering from steric constraints arising from self-excluding geometries. Because the binding geometries output by ATTRACT are characterized by Heligeom in terms of the number of monomers per turn (N) and pitch value (P) of associated regular assemblies, these parameters are used for the classification. Structures with essentially no steric constraints are classified into the “Filament” category. To identify these geometries, the current version of our screen uses a simple radius-based scheme, allowing $P > 2R_M$, where $R_M$ is the maximum radius of a monomer. Similarly, a “Cyclic” geometry is defined if N falls in a range centered on an integral value ±0.1 and P is in the range $0 - 0.5 \, \text{Å}$. Both “Filament” and “Cyclic” structures are passed through the screen with no further modification. On the other hand, “Near-cyclic” geometries are defined by accepting a pitch error of up to 5 Å per interface, and thus structures for which $P < (N - 1) \times 5\,\text{Å}$. These geometries are shunted to an automated Monte Carlo energy-minimization or “adjustment” procedure in which cyclic geometry is enforced\(^9\). The remaining monomer-monomer binding geometries, classified as “Near helical”, are currently tested for steric clash, as identified by the interaction energy computed between any monomer $i$ and the two closest monomers from the next helical turn, $i+M$ and $i+M+1$, where $M$ is the largest integer not greater than the number N of monomers per turn. Suitable adjustment methodologies to optimize multiple interfaces that may appear between adjacent helix turns will be applied in a future version of our approach.

The automatic filtering and cyclic adjustment procedure for the entire set of RecA docking poses took approximately 3h on a single 3GHz Intel CPU and retained 9% of the initial 90,489 ATTRACT-generated poses.

10.6.7 Exploring the variability of binding modes

In addition to identifying binding modes using unbiased docking, the variability of a given binding mode was explored as follows. First, targeted docking simulations were performed near the structure of the desired binding mode. The results from the targeted docking simulation were then post-processed to associate screw parameters (pitch values, number of monomers per turn) to each docking result, together with the $f_{\text{NAT}}$ values with respect to the reference binding mode. Then, results with $f_{\text{NAT}}$ values higher than 50%, interface Cα-RMSD < 3.5 Å and energies lower than -37 RT (native interface energy + 5 RT) were filtered out and used as starting geometries.

\(^9\)The adjustment procedure is described in chapter 8, section 8.5
in the second stage of exploration. This second stage consisted of $10^5$ steps of Monte Carlo simulation at a temperature of 300K, using six variables (three translations and three rotations) for the rigid body displacement of the sampled monomer with respect to its fixed monomer partner (i.e., the receptor in the docking simulations). The Monte Carlo trajectory was confined to $f_{\text{NAT}}$ values higher than 0.5 with respect to the reference geometry. Sampling of variables was uniformly performed in intervals of ± 5 degrees for the rotational variables and ± 3 Å for the translations. Acceptance varied between 0.2 and 0.4. Parameters such as the pitch, the number of monomers per turn, the Co-RMSD, the $f_{\text{NAT}}$ and the $f_{\text{IR}}$ values with respect to the starting geometry were output at each simulation step. The number of Monte Carlo simulations was varied to ensure coverage of the entire set of targeted docking solutions that satisfied the filtering conditions.

10.7 Acknowledgments

The authors thank B. Hartmann and M. Prentiss for stimulating discussions.
Figure 10.8: Accommodation of flexible regions in predicted RecA fiber forms. Accommodation of flexible regions in predicted RecA fiber forms. The fibers forms A to I (Fig. 2, main manuscript) are displayed in ribbon representation, with monomers alternatively colored in white or gray. The extremities of the flexible regions are shown in van der Waals representation, with the extremities of loop L1 (residues 156 and 165) in blue, those of loop L2 (194, 210) in red and the linker extremity (residue 38) in green. The N-terminal helix binding region (residues 89, 124, 127, 128, 131, 132, 135 to 138) is in orange.
10.8 Additional results

The results presented in this section complete and detail some points addressed in the article. They have been included in the supplementary information that accompanies the submitted article.

10.8.1 Flexible RecA fragments

In the study presented above, the RecA monomers were truncated to their rigid core to explore whether the rigid core of the interface can determine the putative modes of monomer-monomer association. This section will come back on two aspects where these regions may play a role in the filament formation.

Binding geometries and flexible regions

Figure 10.8 displays the same oligomers A-I shown in fig. 10.4 of the article, with the regions corresponding to the extremities of pruned flexible regions represented as color patches. It can be verified that in all cases, these regions are situated outside the monomer-monomer interfaces, which means that the oligomeric form can accommodate the flexible regions. We also represented in orange the amino-acids of the rigid core that participate in binding the N-terminal helix (1-23) in both the 2REB and 3CMW crystal structures. Although the helix does not necessarily bind to that region, the conservation of the helix binding region between the two known RecA structures indicates that it strongly stabilizes the association. In all cases except binding form B (the RecA cyclic trimer), this conserved region was found accessible to the helix binding.

Influence of RecA flexible/mobile components on its binding modes

The results presented in the above study suggested that the flexible regions may modulate the mode of association and/or the preference towards one or another binding mode. Here, we further investigated this idea by exploring whether the presence of the ATP cofactor (for the RecA-ATP binding mode) or the flexible N-terminal linker (for RecA-ADP) modifies the association landscape characterized by the pitch and the number of monomers per turn in fig. 10.5. In the 3CMW crystal structure [30], the ATP cofactor is located at the interface between two monomers (represented in purple in fig. 10.9, right). In the 2REB crystal [29], the same region of the interface that contacts the ATP in 3CMW is occupied by a fraction of the N-terminal linker that folds upon the monomer it belongs to (residues 30 to 37, orange region in fig. 10.9, left).

We performed targeted docking simulations in the presence of the (30-37) linker segment for the RecA-ADP form (left) and of ATP for the RecA-
Figure 10.9: Variability of the ADP and ATP fiber forms in the presence of flexible/mobile interaction component. (A) Comparison of RecA-ADP (left) and RecA-ATP (right) binding modes. Two consecutive monomers issued from the PDB files 2REB (left) and 3CMW (right) are represented. In both cases, the top monomer is represented in surface mode, in white. The bottom monomer is shown in a ribbon representation and with an orientation that is common to both panels left and right. The rigid core is in grey; the L1 loop (right) or its extremities (left) are in blue; the L2 loop (right) or its extremities (left) are in red; the N-terminal domain, including a terminal helix and a flexible linker, are in green (right) or green and orange (left), the orange region corresponding to the fraction of the linker which folds back on its own monomer in structure 2REB (residues 30-37). The ATP cofactor in the right panel is in purple. (B) Results from targeted docking simulations on RecA-ADP (left) and RecA-ATP forms (right), characterized by their pitch P and number of monomers per turn N; the results were obtained in the presence (red +) or in the absence (blue ×) of flexible or mobile elements, the 30-37 linker segment for RecA-ADP (left) and the ATP cofactor for RecA-ATP (right).
ATP form (right), and we compared the results with those obtained in the absence of the segment or ATP. Only the results that fulfill the conditions given in the Methods section of the main article (energy lower than -37 RT, $f_{\text{NAT}}$ values greater than 0.5 and interface $\alpha$-RMSD lower than 3.5 Å) were selected. The comparison is displayed in fig. 10.9, in terms of pitch and number of monomers per turn where the blue crosses, corresponding to the results obtained without the linker (left) or without ATP (right), were used as starting points for the Monte Carlo simulations represented in fig. 10.5. While the linker fragment only slightly modifies the sampled regions of the RecA-ADP interface (fig. 10.9, left) and does not modify the interaction energy (not shown), the presence of ATP confines the sampled region to pitch regions below 100 Å and number of monomers per turn between 6.2 and 6.5 (fig. 10.9, right). The presence of ATP also lowers the energy of the selected members of the RecA-ATP family by $\sim$ 5 RT, which stabilizes that binding mode with respect to the RecA-ADP binding mode. As seen in Table 10.2 (article), the two binding modes present equivalent values of the interaction energy in the absence of ATP.

10.8.2 Near-cyclic to cyclic adjustment

<table>
<thead>
<tr>
<th>N</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>#adjusted</td>
<td>1</td>
<td>77</td>
<td>88</td>
<td>72</td>
<td>51</td>
<td>36</td>
<td>31</td>
<td>27</td>
<td>13</td>
<td>11</td>
<td>6</td>
</tr>
<tr>
<td>#families</td>
<td>1</td>
<td>52</td>
<td>56</td>
<td>47</td>
<td>42</td>
<td>31</td>
<td>26</td>
<td>22</td>
<td>11</td>
<td>10</td>
<td>6</td>
</tr>
</tbody>
</table>

For the RecA docking output reported in the article, 640 binding geometries out of 90489 docking poses were considered as “Near-cyclic” and their adjustment to cyclic geometries, following the method described in chapter 8, section 8.5, gave rise to 1280 cyclic geometries with an average deviation resulting from the starting state of 10 Å $\alpha$-RMSD. Among these, 437 cyclic geometries successfully passed the filtering process based on steric and energetic conditions (section 10.6.6 above). The 437 cyclic geometries varied from homodimers (1 occurrence) to 38-mers (1 occurrence),
with the highest populated cyclic N-mers being found for N between 3 and 11 monomers per turn (>10 occurrence). The number of occurrences for a given N value reaches 88 (N=4), corresponding to 56 distinct binding geometries (characterized by $f_{\text{NAT}}$ values greater than 50%)(table 10.3).

10.8.3 Results from docking simulation on RecA, RadA and Dmc1

The results of ATTRACT simulations, after post processing with the filtering/adjustment protocol described in section 10.6.6, are displayed in fig. 10.10 as a plot of the pitch versus the number of monomers per turn. Results for which regular association would lead to steric clashes have been discarded or adjusted to cyclic geometries (almost empty horizontal band). Cyclic geometries can be observed for pitch values equal to zero. The three plots represent three different signatures for the homolog proteins, where left-handed and right-handed helical organization are similarly represented.
Figure 10.10: Helical parameters from Heligeom for regular assemblies obtained from ATTRACT docking simulations in RecA, RadA, and Dmc1. Each circle on the P versus N plot, where P is the pitch (Å) and N the number of monomers per turn, represents one docking pose geometry after filtering and post-processing as described in Methods, and Fig. 10.1. Pitch values for left-handed helical geometries (parameter dir in Methods) are given as negative values. Points are colored by the interface energy of the corresponding association geometry, with darker values indicating more favorable energies.
10.8.4 Groove width

Fig. 10.11 shows the groove width variation for three different RecA helices, the two known forms 2REB [29] and 3CMW [30] and a chimeric form obtained by alternating the 2REB and 3CMW binding modes, represented in Fig. 10.7 of the main article.

Figure 10.11: Variation of the groove width along one filament turn. Groove width variations along one filament turn (360°) are represented for structures 2REB (black line), 3CMW (red line) and the model with alternate 2REB and 3CMW interfaces (green line).
Part IV

Conclusion
Chapter 11

Discussion and conclusion

11.1 Discussion

The study on Heligeom applications was separated in two parts. First we explored the possibility of Heligeom alone and then we coupled it with docking simulations or Monte Carlo exploration to explore the binding modes of RecA.

We used Heligeom alone to build regular geometries from a given binding mode (chapter 9). This can be useful to assist in the interpretation of data obtained experimentally at low resolution as well as precise data on local interaction obtained via experimental or theoretical studies, as demonstrated on figs. 9.2, 9.4, 9.5. We also demonstrated the possibilities to build irregular assembly, both in a continuous deformation fashion (along an axis, like the plectoneme fig. A.6) or discontinuously (as the RecA superhelix, fig. 9.2).

Finally we established that Heligeom can be used as a two way communication tool between the atomic level and the level of protein filament: from the atomic to the molecular assembly level, as demonstrated with the construction of the electrostatic potential of actin (fig. 9.7) and from the molecular assembly to the atomic level, like in the construction of the RecA plectoneme (fig. A.6) where we used experimental data to determine the axis of the two strands.

Coupling Heligeom to exploration methods of monomer-monomer binding geometries such as docking or Monte Carlo simulations (chapter 10) resulted in the generation of a large diversity of self-association modes for the RecA filament. The plausibility of the generated geometries was supported by the presence of known or suggested forms among them (fig. 10.4). We also explored the variability in the vicinity of the two best characterised binding modes and expressed it by the variation of helical parameters (pitch and number of monomers by turn, fig. 10.5) that can be directly related to experimental observation. This allows to look at results not only at the local level but also at the assembly level. With more detailed exploration...
11.1. DISCUSSION

(for example reintroducing the flexible regions that have been neglected in the present study and more extensive sampling), this approach could help shedding light on the mechanism of conformational changes in the studied filaments.

Heligeom is presently dedicated to studying topologically linear filaments, where each monomer interacts at most with two other monomers, the preceding and the following ones. We also implemented tools to specifically optimise the ring geometries using the possibilities offered by the screw representation. Such optimisation methods could be developed to investigate other geometrical organisation with the tools available in PTools. A polymeric shape of particular interest is the stacked filament as it can be found in various systems like for example microtubules. The construction of stacked filament presents the challenge of simultaneously optimising two interfaces for a given monomer \( n \), the interface with the following monomer \( n+1 \) and that with the monomer on top of it, the \( (n + \text{number of monomer by turn}(N)) \) monomer. However, in a regular helix, Heligeom can produce the later as long as we dispose of the former. Thus, any given mode of interaction between two monomers defines the position of the monomer atop of them (fig. 11.1), allowing computation of the interaction energy for both interfaces at minimal computational cost. In addition, exploring the possibilities of a stacked conformation between consecutive helix turns only necessitates moving one monomer. This can be done for example via Monte Carlo optimisation.

Another interesting oligomeric organisation is the interlaced filament (such as actin). The construction of the RecA plectoneme shows that it is already possible to build interlaced filaments with Heligeom. Further progress in the modelisation could be done, for example by optimising the phase presented by one filament to the other. This could be done by rotating a section of one of the filaments around its axis (a function already present in PTools) and computing the interaction energy for every step of the rotation.

A further step could be to optimise the internal conformation of the
filaments when interlaced. Again using the existing tools to quickly build a section of the interlaced filaments and computing the interaction energy would permit optimising the conformation.

The improvements discussed here present logical next steps in the extension of the potentialities of the PTools library and are ready to be implemented as the tools necessary for their implementation already exist.

An important point is that the Heligeom method is neutral to the source of information considered at either local or global levels (and the implementation corresponds to that neutrality, both by the opensource implementation and the modular nature of PTools). We chose a coarse grain representation and a rigid body docking process to explore the binding modes at the monomer level because it suits well for preliminary exploration and for a global view of the possible assembly modes, independently of external factors favoring one or the other binding geometry. Nevertheless, if precise information is needed, more detailed methods can be used.

Atomistic level can be explored with Monte Carlo, molecular dynamics, normal modes or any other methods of choice. We could even imagine using quantum simulation to work out some details of a molecular interface and then project this information on an object of hundreds of thousand atoms to study their consequences.

On the other hand, information at the filament level can come from any method such as low resolution methods or molecular imagery. One could also use physical simulation of filament such as a wormlike chain simulation to explore the kinetics of the filament, and combine this information with a full atomic model of the filament.

The work we realised on DNA (chapter 6), beyond the extension of PTools capabilities and the possible complementarity with the study of RecA, was a source of inspiration for the conception and improvement of Heligeom. One of the main insights during this work was that DNA and protein filaments could be modeled in a very similar way and the methods developed for one could be used for the other. There is a wealth of methods for the modelisation and study of nucleic acids and when stumbling on a new problems with protein filaments, one should first check that a turn-key solution does not exist for DNA or RNA, ready to be used.

**Perspective**

There are many directions in which Heligeom could be extended. The construction of irregular geometries of filaments is important to model conformational changes that often accompany biological processes. This represents a logical next step after the study of binding modes as we have done for RecA. The tools for constructing irregular filaments are only in their infancy in the present release of PTools/Heligeom and new scripts generalising
11.1. DISCUSSION

and easing this kind of modelisation should be implemented.

As previously discussed, modelisation tools for addressing more complex shapes than a topologically linear filament should be implemented. Most of the cytoskeleton should be open for study once the tools for stacked or interlaced filament will be developed, and, as illustrated by the construction of the virus capsid, the method could be extended to other forms as long as a pattern of recurring interfaces can be found.

We developed Heligeom in the context of the cell, but this approach is viable to any kind of molecular filament. It could complement the design of synthetic filaments for industrial applications, opening new opportunities for the method.

During this thesis, I was able to show few of the possibilities of Heligeom and to realise many proofs of concept. Furthermore, by offering to the scientific community a fast and open implementation of Heligeom with the coming release of PTools, it is my hope that this approach will be used and extended by future users.
Part V

Appendix
Appendix A

Résumé

A.1 Introduction

Les assemblages moléculaire de grande taille jouent des rôles aussi bien divers que primordiaux dans la cellule. C’est le cas par exemple de l’ADN, support de l’information génétique; des microtubules ou de l’actine, qui jouent un rôle structurel pour la cellule en formant le cytosquelette mais aussi un rôle important de transport des vesicules; ou encore des recombinases, telles que le filament de RecA, qui jouent un rôle dans la réparation de l’ADN.

Ces assemblages sont souvent sujets à d’importantes réorganisations structurelles (polymérisation/ dépolymerisation, transition de formes, ...) qui peuvent se produire à l’échelle de la seconde.

Longtemps difficiles d’étude de par leur taille, les assemblages supra-moléculaires sont aujourd’hui des objets d’étude privilégiés. En effet, ces dernières années ont vu l’essor d’un ensemble de techniques expérimentales dites de “basse résolution”, spécialement adaptées pour les objets de grande taille. On peut citer la diffusion aux petits angles des rayons X (SAXS de l’anglais : Small Angle X-rays Scattering) ou des neutrons (SANS de l’anglais Small-angle neutron scattering) qui permet d’étudier les propriétés structurelles des matériaux à une échelle allant de 1 à 100 nm ou la microscopie électronique.

Ces dernières décennies, de nombreuses techniques de modélisation moléculaire ont vu le jour, visant à compléter les techniques expérimentales de biochimie. Ces développements ont été reconnus par l’attribution du prix Nobel de chimie en 2013. Parmi ces techniques, on retrouve par exemple la dynamique moléculaire, qui permet de simuler les mouvement physiques de particules (atomes et/ou molécules) selon les lois de Newton, ou le calcul des mode normaux qui permet de prédire les directions des mouvements de plus grande amplitude d’une molécule. Néanmoins, ces techniques devenues classiques pour l’étude de molécules se revèlent à l’heure actuelle trop coûteuses en temps de calcul pour explorer les modes d’organisation et les
propriétés mécaniques des assemblages moléculaires de grande taille.

Le but de cette thèse est de proposer des approches méthodologiques et des outils de modélisation pour faire le lien entre l’échelle atomique et les techniques de simulation moléculaire d’une part, et l’échelle des assemblages moléculaires et les techniques “basse resolution” d’autre part.

A.2 Heligeom

Pour ce faire, nous avons utilisé un paradigme particulier pour notre représentation du déplacement des molécules: la transformation de vissage (en anglais: “screw movement”). Cette représentation est très simple: pour un axe donné (défini par un vecteur et un point), elle associe une rotation et une translation (voir figure A.1).

Figure A.1: Schéma d’une transformation de vissage entre deux monomères A et B. L’axe est défini par le point P et le vecteur Ω. La transformation du monomère A au monomère B est la combinaison de la rotation θ autour de l’axe et de la translation trans le long de l’axe.

Il est possible de définir de cette manière tous les mouvements dans l’espace de type “corps rigide”. L’intérêt de cette représentation est qu’elle se prête très facilement à la construction d’hélices (voir figure A.2).

Cette représentation nous permet notamment de construire de manière efficace l’assemblage helicoïdal correspondant à une géométrie d’interaction donnée entre deux monomères.

Nous avons choisi d’implémenter cette représentation dans la bibliothèque PTools. PTools est une bibliothèque de modélisation moléculaire distribuée de façon libre. Elle est organisée autour d’un coeur en langage C++ et d’une interface en langage Python, ce qui la rend performante en terme de calcul et simple d’utilisation pour le développement méthodologique.

Nous avons développé, dans la bibliothèque PTools, une suite d’outils
Figure A.2: Illustration de la construction d’une hélice par transformations de vissage. La répétition de la même rotation \( R \) et translation \( t \) autour d’un axe \( \Delta \) sur un monomère donné permet la construction d’un assemblage de monomères qui prend la forme d’une hélice.

appelée Heligeom. Cette suite d’outils logiciels inclut une implémentation des transformations de vissage, un programme pour obtenir une transformation de vissage à partir de deux monomères et un certain nombre de programmes affiliés permettant d’exploiter ces informations dans le cadre de simulations d’amarrage moléculaire (“docking”). Nous avons également réalisé un programme permettant d’utiliser une transformation de vissage pour construire une hélice le long d’un axe arbitraire (figure A.3).

A.3 Application

Notre capacité à construire de manière efficace l’assemblage hélicoïdal correspondant à une géométrie donnée d’association entre deux monomères nous ouvre plusieurs possibilités.

Nous pouvons par exemple comparer deux assemblages de formes globales très différentes en terme de mode d’association local. De façon surprenante, les deux formes représentées en violet et vert sur la figure A.4 sont localement similaires.

Cela nous permet aussi de transférer des informations calculées au niveau atomique jusqu’au niveau de supra-assemblage. Ainsi nous avons pu représenter le potentiel électrostatique du filament d’actine (figure A.5) à partir de calculs de potentiel effectués sur un nombre limité de monomères.

Inversement, nous avons aussi utilisé des informations expérimentales pour reconstruire un modèle de plectonème de RecA, c’est-à-dire un filament
Figure A.3: Constructions sur un axe arbitraire. Il est possible de construire une hélice macromoléculaire sur un axe de forme arbitraire (à gauche), indépendamment des éléments qui la composent, ADN (au centre) ou protéine (ici RecA, à droite)
Figure A.4: Formes “voisines” en hélice et en anneau de Dmc1. Comparaison entre la forme octamérique cyclique de DMC1 (PDB code 1V5W, en vert) et une forme hélicoïdale obtenue construite par Heligeom à partir d’une géométrie d’association monomère-monomère très voisine. La représentation en ruban montre géométries d’association pour deux monomères, en jaune pour la forme cyclique, en jaune et violet pour la forme hélicoïdale. Les deux formes ont été superposées sur le monomère de gauche représenté en jaune. les deux géométries d’assemblage partagent 61% de paires de contact monomère-monomère. Leurs énergies d’interaction diffèrent de moins d’1 RT, et la déviation entre les coordonnées des atomes Co des deux monomères de droite en ruban (respectivement jaune et violet) est de 5.5 Å. Le nombre de monomères par tour (7,4) et le pas d’hélice (79,0 Å) de la forme en hélice droite sont compatibles avec les géométries caractéristiques des filaments actifs de recombinaison homologue.
Figure A.5: Potentiel électrostatique autour d’un filament d’actine. Le potentiel a été calculé localement en tous les points d’une grille hélicoïdale, qui a été ensuite étendue à deux tours d’hélice de F-actine par transformation de vissage. La figure est colorée du bleu au rouge en fonction des valeurs de potentiel allant de -8 à 0 kcal/mol.

entrelacé avec lui même (figure A.6).

Nous avons également montré que le principe d’Heligeom pouvait être généralisé à des assemblages très différents des hélices, avec la reconstruction d’une capsade de virus (figure A.7).

A.4 Application à l’étude de RecA

Si on associe Heligeom à une méthode d’échantillonnage des modes d’auto-association d’une protéine donnée, on peut explorer les géométries possibles pour les assemblages réguliers de cette protéine. Nous avons choisi de coupler Heligeom avec PyAttract, une suite logicielle de “docking” gros grain présente dans la bibliothèque PTools, pour étudier le filament de RecA, responsable de la recombinaison homologue.

En effectuant une simulation d’amarrage moléculaire d’un monomère de RecA sur lui même, puis en explorant ces résultats à l’aide d’Heligeom, nous avons étudié plusieurs aspects de l’auto-association de RecA.

Nous avons généraè la diversité des formes possibles pour les oligomères de RecA et nous avons pu retrouver parmi ces résultats des filaments présentant des paramètres hélicoïdaux similaires à ceux des structures de filaments identifiées expérimentalement (figure A.8).

Pour certaines géométries d’association d’intérêt biologique, nous avons exploré la variabilité des formes possibles à leur voisinage. Nous avons pu représenter cette variabilité en terme de déviation des paramètre hélicoïdaux (pas d’hélice et nombre de monomères par tour).

Finalement nous avons pu examiner les conséquences structurales de l’introduction d’irrégularités dans ces oligomères.

Pour réaliser ces études, nous avons dû mettre en place des outils complémentaires, tel qu’un système de filtres pour identifier les géométries d’association compatibles avec une organisation régulière en hélice ou en anneau.
Figure A.6: Vue d’un tour de filament de RecA superenroulé, en résolution atomique. La structure a été construite selon les caractéristiques géométriques décrites par Shi et al. [106], avec un pas d’hélice de 160 nm pour le super enroulement gauche. Les axes des deux brins (en cyan et vert) sont éloignés de 110Å. L’ADN inclus dans chacun des deux filaments de RecA est représenté en rouge.
Figure A.7: Étapes de construction d’une capsid de virus à partir de quatre unités en interaction. À gauche, les quatre unités sont représentées en mode surface. Chacune des trois interfaces ainsi définies (schematisées par les lignes colorées en cyan, vert et mauve) correspond à un mouvement de vissage différent. Les caractéristiques de ces transformations en terme de nombre de monomères par tour (N) et de pas d’hélice (P) sont indiquées. De droite à gauche, les étapes de construction étiqueté de (i) à (iv) correspondent respectivement à l’application en (i) d’une symétrie cyclique d’ordre 3; en (ii) d’une symétrie cyclique d’ordre 5 appliquée au trimère obtenu en (i); en (iii) d’une symétrie cyclique d’ordre 2 appliquée au pentamère obtenu en (ii); en (iv) la même transformation utilisée à chaque interface représentée par les lignes pointillées mauves. Pour faciliter la compréhension, le trimère qui a été construit en (i) est entouré par un triangle noir.
Figure A.8: Illustration de la diversité des formes possibles pour les oligomères de RecA. Partant de prédictions d’amarrage moléculaire présentant une énergie d’interaction suffisamment basse, nous avons utilisé Heligeom pour construire les assemblages correspondants. Les formes d’hélice droite H et I correspondent respectivement aux formes sans et avec ATP observées par microscopie électronique et résolues par cristallographie. Des formes dimériques ou hexamériques ont également été détectées à partir d’observations à basse résolution.
A.5 Conclusion

Heligeom permet d’établir une communication entre l’échelle atomique et l’échelle des gros assemblages moléculaires. Son utilisation prend tout son sens lorsqu’elle est couplée avec des méthodes de modélisation ou avec des données expérimentales. Dans ce travail, il a été couplé à des explorations d’amarrage moléculaire dans le cadre d’une étude sur le filament de RecA. Cependant, son implémentation dans la bibliothèque PTools a été réalisée de manière à rendre aisé son couplage avec d’autres méthodes. Le filament de RecA est topologiquement linéaire, mais j’ai pu montrer que notre approche se prête à d’autres types d’assemblage: des formes d’hélice diverses, telles que les hélices entrelacées (plectonème), ou plus généralement n’importe quel assemblage présentant des motifs d’interface récurrents, comme les capsides de virus.

Les techniques mises en œuvre ne sont pas nouvelles mais l’application dans ce contexte est originale. Durant cette thèse, j’ai pu montrer un aperçu des possibilités offertes par cette approche. D’autre part, en offrant à la communauté scientifique une implémentation efficace et ouverte dans le cadre de la bibliothèque PTools, mon espoir est que cette approche sera reprise et améliorée pour des utilisations ultérieures.
Appendix B

Exploring RecA interfaces with PTools/Heligeom

This section details the PTools/PyAttact and PTools/Heligeom commands that were used for the generation of various modes of auto-association of the RecA monomer, the analysis of each of these modes in terms of the corresponding screw transformations and the construction of corresponding regular assemblies, either helix, ring or straight oligomer.

B.1 Docking procedure

Preparation of a PTools/ATTRACT coarse-grained docking run necessitates the creation of reduced structures for the docking partners using the reduce.py script. Here, reduction of the rigid core of the RecA protein gave rise to 595 grains, starting from 2021 atoms. In the following description, one partner (arbitrarily chosen) will be called the receptor and the other the ligand; the receptor is held fixed in the docking run. A total of 244 starting points were distributed around the (fixed) receptor using the translate.py script. From each starting point and for each of 228 predefined ligand orientations (rotations), the Attract.py script ran a series of six minimizations of the interaction energy between the receptor and the ligand. In the present example, two docking simulations were run starting from the two RecA protein structures 2REB [29] and 3CMW [30]. Four reference structures were given for comparison, obtained from the two different modes of filamentous association seen in the PDB entries 2REB and 3CMW. Each filament structure furnished two ligand orientations, corresponding to the n+1 (upper) and n-1 (lower) interfaces with respect to the receptor. For each of the two PDB entries, then, two reference files were obtained as follows. First, three consecutive (and non-terminal) monomers were extracted from the filament structure and superposed on the central monomer that was used as receptor in the docking simulation, using the PTools selection and superposition util-
B.1. DOCKING PROCEDURE

Ities. This defined the preceding and following monomer positions relative to the receptor, onto which the ligand molecule in the docking simulation was superimposed in order to create the corresponding reference structures. For each ATTRACT output, root mean square deviation (RMSD) values with respect to each of the four reference structures, taken on the Ca atoms, were calculated. We first prepared the protein by pruning the flexible regions (N-terminal domain, L1 and L2 loops, see Methods in chapter 10). We then reduced the rigid core (reca-rigid.pdb) to coarse grain resolution (reca.red) using the PTools reduce.py script.

```
python reduce.py --prot reca-rigid.pdb > reca.red
```

Docking was then performed using the ATTRACT function, which performs energy minimizations between two partner macromolecules with respect to translation and rotation degrees of freedom, starting from thousands of initial configurations, in which the ligand is distributed around the receptor at different positions and orientations.

```
python attract.py reca.red reca.red --ref=reca_n1.red --ref=reca_n2.red --ref=reca_p1.red --ref=reca_p2.red > docking.att
```

In this command, the two molecules to be docked are identical (reca.red). Also, four separate reference structures have been provided for root mean square deviation (RMSD) calculations (calculated using Ca coordinates): two of them with the next monomer (reca_n1.red for structure 2REB, reca_n2.red for structure 3CMW), and two with the preceding monomer (reca_p1.red and reca_p2.red) (see Docking Procedure). Selected and annotated output lines from the ATTRACT simulation are shown here (annotation is indicated by "#”).

```
#  i  j   E   rmsd1  rmsd2  rmsd3  rmsd4
==  9 186 -46.48  25.66  34.04  64.49  54.34 # X
== 42 113 -45.00  55.09  49.00  73.41  72.06 # A
== 86 95  -44.78  63.68  68.37  37.08  35.44 # F
== 39 21  -41.14  28.03  2.43  70.45  67.62 # I
== 77 85  -40.21  52.23  59.15  31.22  19.62 # G
== 20 204 -39.41  0.96  26.52  71.06  59.94 # H
== 99 117 -33.91  79.33  73.69  52.21  58.13 # C
== 49 129 -30.69  65.99  54.31  70.20  72.65 # D
== 215 131 -30.16  73.53  75.56  58.45  54.67 # E
== 241 17  -22.68  79.85  65.17  60.28  71.69 # B
```

Each line represents an interface geometry after energy minimization carried out by ATTRACT, and contains, from left to right, two indices for the starting orientation in terms of translation and rotation (see Methods), the value of the interaction energy (in RT units) and the RMSD values corresponding to the four reference structures (in Å). On the right-hand side we have added structural labels that correspond to the labels discussed in the main article and displayed in figs.10.4 and 10.8 in chapter 10.
APPENDIX B. EXPLORING RECA INTERFACES WITH PTOOLS/HELIGEOM

B.2 Heligeom Analysis

The Heligeom utility `extractHelicalParameters.py`, which couples the ATTRACT output to the screw analysis performed by `heligeom.py`, was used to compute and list the pitch, the number of monomers per turn and the direction of rotation for each docking geometry, together with interaction energy values.

```
python extractHelicalParameters.py docking.att reca.red > screw.txt
```

The results here are redirected to the file `screw.txt`. Selected lines output are shown below.

```
# i  j  N/turn  pitch  hand   E
  9 186  4.54  12.35  L  -46.48 # X
 42 113  2.00  0.01  L  -45.00 # A
 86  95  2.03  52.50  L  -44.78 # F
 39  21  6.41  90.26  R  -41.14 # I
 77  85  5.78 106.29  L  -40.21 # G
 20 204  5.80  72.77  R  -39.41 # H
 99 117  5.05  0.28  R  -33.91 # C
 49 129  6.01  0.09  R  -30.69 # D
215 131 18.04  1.78  R  -30.16 # E
241  17  3.07  0.01  L  -22.68 # B
```

Alternatively, an automatic filtering/adjustment post-processing script can be run in order to extract the docking results corresponding to the “Filament” or “Cyclic” categories defined in Figs. 10.1 and 10.2 (chapter 10) and to filter them as described in Methods (“Processing and filtering the sampling results”, chapter 10). The command and its output are similar to the above description of the generic extraction process.

```
python extractAndFilter.py docking.att reca.red > filtered_screw.txt
```

From the files `screw.txt` or `filtered_screw.txt`, it is possible to select particular geometries, using ranges of values for the pitch and the number of monomers per turn as selection criteria. For example, the Heligeom command

```
python filterHelicalParameters.py screw.txt -p 70 95 -n 5.5 6.5 -d R
```

accepts screw transformations leading to right-handed (-d flag) helices with pitch values between 70 and 95 Å (-p flag) and comprising between 5.5 and 6.5 monomers per turn (-n flag). The output is here

```
 39  21  6.41  90.26  R  -41.14
 20 204  5.80  72.77  R  -39.41
```
In the same way, geometries consistent with hexameric ring arrangements can be selected using

```python
python filterHelicalParameters.py screw.txt -p 0 0.1 -n 5.9 6.1
```

Note that a range of values was indicated in these commands in order to allow for flexibility in the ring closure conditions and the desired overall pitch. Finally, we used Heligeom to build one or more turns of ring/helix specified by the rotation and translation indices in the docking.att output file. For example, the commands

```python
python extractHelicalModel.py docking.att reca.pdb 49 129  > D.pdb
python extractHelicalModel.py docking.att reca.pdb 39 21 2 > I.pdb
```

were used to build a ring and a helical fiber, respectively. In the first command, the docking result corresponding to translation index 49 and rotation index 129, which we labeled D in the above output, was used. This geometry corresponds to an hexameric ring (6.01 monomers/turn, with a nearly vanishing pitch of 0.09 Å). In the second command, the result labeled I is a helix with 6.4 monomers per turn and a 90.3 Å pitch. In the latter command, the desired number of helical turns to be output (here 2) was indicated at the end of the command line. Cyclic or helical fibers shown in fig.10.4 were constructed for the A-I binding modes using that same procedure.
Appendix C

Construction of the RecA plectoneme

Helical oligomers can be constructed around any axis provided by the user, which can be input to Heligeom as a set of atoms forming a line and positioned at around 1 Å from each other. Generation of the filament is performed by successively applying the given screw transformation with respect to the axis at each monomer level.

Below are reported the PTools commands that were used to construct the supercoiled RecA-DNA filament represented in fig. 9.6 of chapter 9. The construction involves two intertwined regular helical assemblies of RecA proteins, each formed along a separate curved axis forming a negatively supercoiled turn. Such structures have been observed by electron microscopy and described in reference [106]. The following Python code generates the two axis portions (here stored as axis1.pdb, axis2.pdb), each with a pitch of 160 nm and differing in the direction of the extension: in the \( +Z \) direction for one and the \( -Z \) direction for the other. The pitch value used represents the lower limit of the pitch distribution observed in reference [106].

To generate the first axis portion, a point \( r_1 \), initially situated 54 Å from the origin in the X direction (54 Å is approximately the radius of the RecA filament with PDB code 3CMW), is incrementally displaced along a superhelical path about the Z axis in 1600 steps. The desired rotation at each step is produced using the PTools \texttt{A Rotate} function, where points A and B define the Z axis. The 1601 generated points are stored in the \texttt{Rigidbody} object \( R_1 \). The second axis portion is generated from the first by rotating by \( \pi \) about the Z axis and flipping the result.

```python
from ptools import *
import math

# generating first axis portion
```
r1 = RigidBody()
r1.AddAtom(Atom(Atomproperty(),Coord3D(54,0,0)))
R1 = r1
A = Coord3D(0,0,0)
B = Coord3D(0,0,1)

dphi = 2.0*math.pi/1600
for i in xrange(0,1600):
    r1.ABrotate(A, B, dphi)
    r1.Translate(B)
    R1 = R1 + r1
WritePDB(R1, "axis1.pdb")

# generating second axis portion

R1.ABrotate(A, B, math.pi)
R2=RigidBody()

for i in range(1600,1,-1):
    ato = R1.CopyAtom(i)
    R2.AddAtom(ato)
WritePDB(R2, "axis2.pdb")

A RecA helix can then be built along each of these two axes using Heligeom’s
buildProteinAlongAnAxis.py script.

```
python buildProteinAlongAxis.py axis1.pdb reca.pdb reca-n1.pdb > RecAsupercoil1.pdb
python buildProteinAlongAxis.py axis2.pdb reca.pdb reca-n1.pdb > RecAsupercoil2.pdb
```

This tool first obtains the local screw transformation parameters from
the two structures provided (reca.pdb and reca-n1.pdb, successive monomers
in 3CMW). The filament is obtained by sequentially generating monomer n
from monomer n-1, using a composition of the screw transformation with
the transformations between appropriate sections of axis1.pdb. Construction
can be accompanied by simultaneous global rotation of the oligomer
around the curved axis. This may be necessary since the filament is no
longer symmetric.
Bibliography


BIBLIOGRAPHY


BIBLIOGRAPHY


118] W. Rosselli and A. Stasiak, “Energetics of RecA-mediated recombination reactions. Without ATP hydrolysis RecA can mediate po-


Résumé


La suite logicielle Heligeom est disponible dans la bibliothèque libre PTools.