CHAPTER 1

PROTEIN FUNCTIONAL DYNAMICS: COMPUTATIONAL APPROACHES

Yves-Henri Sanejouand
Laboratoire de Physique, Ecole Normale Supérieure de Lyon
46 allée d’Italie, 69364 Lyon Cedex 07, France
E-mail: Yves-Henri.Sanejouand@ens-lyon.fr

As polymers, proteins have very unusual properties. The main one is that they are able to fold, that is, under physiological conditions, to adopt a very well defined three-dimensional structure. Such a folding process has been studied using simple lattice models that were shown to share many properties with natural proteins. Simple models, where a protein is represented as a network of harmonic springs, have also recently proved useful for studying the other functionally important motions of proteins, namely, their conformational changes. Both kinds of models could be used in order to test the hypothesis that non-linear energy localisation phenomena may occur in proteins and play a role in their functional dynamics.

1. Introduction

A question of interest is whether or not phenomena of non-linear energy localisation play a role in the way biological macromolecules achieve their function. As far as I know, in the case of proteins elements for discussing this question are still missing. The goal of this course is just to introduce the functional dynamics of proteins, focusing on the description of a few approaches and models that could be used in order to address this question.

2. Protein structure

Proteins are natural, unbranched hetero-polymers made of twenty different building blocks, the aminoacids, linked together according to a precisely defined sequence, which is encoded in the corresponding gene. For instance, the ten first aminoacids of human myoglobin (a protein responsible for
oxygen transport in muscles) are: glycine-leucine-serine-aspartate-glycine-glutamate-tryptophan-glutamine-leucine-valine (those are usually given using a one-letter code, namely, here: GLSDGEWQLV, where G stands for glycine, L for leucine etc). Human myoglobin is a 153 aminoacids protein while typical proteins have sequences two or three times longer, though there are some with less than fifty aminoacids and a few with more than ten thousands of them.

As exemplified in Fig. 1 for the case of the LAO binding protein, the most remarkable property of these polymers is that they have very well defined three-dimensional (3-D) structures. These are quite compact, as one can see by looking at an all-atoms representation (top of Fig. 1, where each atom is represented as a van der Waals sphere), though they are often sub-structured into "domains" (two of them, in the LAO binding protein case).

The all-atoms representation is quite realistic but it is rarely used. The reason is that the path followed by the polypeptidic chain, which is the information that is useful in order to discriminate the various kinds of protein structures, is hidden. Such a path is shown in the bottom of Fig. 1, drawn with the widely used Molscript program. This representation gives also informations concerning the series of hydrogen bonds within the polypeptidic chain that determine the local conformation: the polypeptidic chain is either in an helical configuration (the famous \(\alpha\)-helix) or extended in side-by-side stretches (the \(\beta\)-sheets). In this later case, represented as arrows in Fig. 1, the hydrogen bonds are made between atoms that can be far apart in the protein sequence.

Protein structures like the one shown in Fig. 1 are obtained using X-ray crystallography or nuclear magnetic resonance. While less than a dozen structures were known in 1975, more than 20000 are now available in the protein databank (PDB), a public repository for the processing and distribution of 3-D biological macromolecular structure data. These 20000 structures can be classified into nearly 1000 different folds, that is, 1000 clearly different paths of the polypeptidic chain.

Note that most proteins whose structure is presently known are soluble proteins, that is, non-membrane ones. Although the laters are much more difficult to produce in large amounts and to crystallize (sic) huge progress has recently been made in this field. For instance, in year 2000 seven new structures of membrane proteins were released in the PDB.
3. Energetics of protein stabilisation

When viewed at the level of description of the protein fold, that is, of the path followed by the polypeptidic chain within the structure, it seems that hydrogen bonding plays a major role in protein stability, through the formation of the secondary structure elements, namely, \( \alpha \)-helices and \( \beta \)-sheets. Moreover, because prediction of secondary structure elements from the knowledge of the protein sequence in the vicinity of a given sequence site is quite efficient (more than 75\% of the aminoacids are correctly predicted by current methods to be in one of the three following states: \( \alpha \)-helix, \( \beta \)-strand or coil\(^2\)), it appears that interactions between neighbouring aminoacids in the sequence also play an important role. Indeed,
during the folding process, secondary structure elements are often found to be formed quite rapidly. However, the key aspect of the protein folding process is related to the question of how the protein is able to "choose" between more than 1,000 different folds, given the fact that many of these folds are alternative packings with similar content of each of the secondary structure elements. For instance, in the last release (May 2003) of the SCOP protein classification database, there are 171 different folds of all-alpha proteins, that is, of proteins with no $\beta$-sheet in their structure.

At this level there are strong indications that hydrophobic interactions play a key role. Indeed, when all known soluble protein structures are analyzed, most amino acids buried in their cores are found to be hydrophobic ones. In other words, the folding process seems analogous to the formation of an oil droplet in water, the hydrophobic parts being clustered, so as to minimize their surface accessible to the water environment. Moreover, when the genetic code is analyzed, it is found that its structure helps protecting amino acid hydrophobicity, by restricting the impact of errors (that is, of mutations) on this property.\(^3\) Correspondingly, when massive mutational experiments are performed, it is observed that a majority of sites along the protein sequence are generally tolerant to amino acid substitutions and that they belong to segments acting as spacers between hydrophobic amino acids.\(^4\) Some of these hydrophobic amino acids seem to be more important than others. For instance, when proteins of same fold but with very different sequences are compared, some buried positions, coined "topohydrophobic", are found to be always occupied by strong hydrophobic amino acids.\(^5\)

4. Protein folding

Because a protein is a long and flexible heteropolymer, it is not obvious how it is able to fold, that is, how starting from a more or less random coil it can reach its structure among the huge number of its possible conformations. Indeed, let us consider that each amino acid along the polypeptidic chain can choose between at least two conformations: either the one it has when it belongs to an $\alpha$-helix or the one it has when it belongs to a $\beta$-strand. Then for a small one hundred amino acids protein, a low estimate for the number of its possible conformations is: $2^{100} \lesssim 10^{30}$. As underlined by Cyrus Levinthal in 1969, a few years after the first protein structure was determined, this is a huge number (the so-called “Levinthal’s paradox”). In-
Indeed, if the folding process were a straightforward enumeration of all these conformations one after the other, and if one conformation were examined every, say, ten femtoseconds (one period of the fastest bond-length vibration in a protein), the whole enumeration would last nearly 400 millions years, while small proteins typically fold within milliseconds.

It would be easy to get rid of such a paradox if there were a clear path in the conformational space towards the native state of the protein. For instance, in the case of a long α-helix Levinthal’s argumentation also holds. Nevertheless, it is well known that the limiting step is the formation of the first turn of the helix, which is stabilised by a single hydrogen bond. Then, for each neighbouring aminoacid adopting the helical configuration, another hydrogen bond is gained, allowing for the fast formation of the rest of the helix.

So, following Levinthal’s conclusion, such paths were sought for, through the search of transient intermediaries during the folding process. There was however little, if any, success, as far as the way a given fold is reached by a given protein is concerned. Another way to get rid of Levinthal’s paradox is to suppose that there are some biological mechanisms able to guide the protein towards its native state. Though such mechanisms were indeed found, namely, an ensemble of proteins called chaperonins which facilitate folding of proteins that are otherwise destined to aggregate, the fact that they are not necessary for the proper folding of at least small proteins was most convincingly shown by total chemical synthesis of some of them. Moreover, in the case of a viral protease, the enantiomer of the natural protein, made with the same aminoacid sequence but with aminoacids of the non-natural D-type, was also synthesised. As expected, it exhibits reciprocal chiral specificity on peptide substrates, that is, each enzyme enantiomer cuts only the corresponding substrate enantiomer. This implies that the folded forms of the chemically synthesized D- and L-enzyme molecules are mirror images of one another in all elements of their 3-D structure.

4.1. On-lattice models

In order to gain a deeper understanding of Levinthal’s paradox, lattice protein models were studied in the late seventies, starting with the case of the 49-mer on a square lattice. In these models, the protein is represented by a chain of beads occupying the sites of a lattice in a self-avoiding way (see Fig. 2). The main advantage of lattice models over more detailed ones is that in many cases their whole conformational space can be examined.
However, even for such simple models the number of possible conformations is growing very quickly as the size of the polymer increases. For instance, on the square lattice, a 18-mer has \(5808335\) different conformations unrelated by symmetries except reverse-labeling.\(^{10}\) Simply enumerating them is tricky in the above case, while in the 49-mer case it is out of reach (there are \(\approx 10^{20}\) of them). However, as shown by Nobuhito Go and his collaborators, starting from a random conformation, the 49-mer can reach its ground state, that is, its lowest-energy configuration, within a few thousands steps of a Monte Carlo simulation, as long as the energy surface is defined as follows. First, the lowest-energy, compact \(7 \times 7\) conformation, is chosen \textit{a priori}. Then, for all pairs of monomers which are close neighbours in this conformation, the contact energy is assumed to be attractive, while for all others it is not.

![Fig. 2. A compact conformation of the 49-mer on the square lattice.](image)

In other words, when the ground-state is at the bottom of a deep funnel on the energy surface, then it is quite easy for a flexible polymer to find its way and reach it, through a random search biased by the average energy gradient. However, though the funnel picture is nowadays the preferred view for understanding the folding process,\(^{24}\) there is no indication that protein energy surfaces are as funneled and as deep as in a Go model. On the contrary, when the probabilities for two aminoacids of types \(k\) and \(l\) (\textit{e.g.}, glycine and leucine, leucine and leucine) to be found in contact in proteins of known 3-D structures are computed, it turns out that the effective interaction energy between them can be assumed to be nearly additive,\(^{11}\) that is:

\[
E_{kl} \approx E_k + E_l, \tag{1}
\]
where $E_k$ is a quantity correlated with the hydrophobicity of the aminoacid of type $k$. Later on, studies were performed with more realistic energy functions and, though models on the square lattice are still being considered, models on the cubic lattice were more and more often preferred, like the popular 27-mer.\textsuperscript{12,13} However, even in this rather simple case, the number of possible conformations is so large that only estimates can be given (nearly $10^{16}$).\textsuperscript{14} This is why most studies were performed with energy functions including a “compaction term”, that is, a bias towards the most compact conformations of the polymer.

For the 27-mer case, the most compact geometry is the $3\times3\times3$ sublattice, in which there are only 103346 self-avoiding possible conformations, unrelated by symmetry except reverse-labeling,\textsuperscript{12,13} that is, 103346 different ways for a chain to go through all 27 lattice sites, going from a site to a neighbouring one at each step. One of these so-called hamiltonian paths is shown in Fig. 3, for the following sequence: PHP$_4$HPHPHP$_{15}$H, where only two kinds of monomers, either polar (P) or hydrophobic (H), have been considered.

![Fig. 3. One of the 120 remarkable compact conformations of the 27-mer.](image-url)
The fact that a sequence can be "threaded" in a given structure in two different ways, the position of the first aminoacid in the first way being the position of the last aminoacid in the other way, is certainly not a property of natural proteins, which are made with asymmetric building blocks, namely, aminoacids of the L-series. As a consequence, the $\alpha$-helix is right-handed, and left-handed helices have not yet been observed in structures of natural proteins. Note that the reverse-labeled sequence of a given protein could quite well be synthesized. If built with aminoacids of the D-series, its 3-D structure should be the same as that of the natural protein, as far as the positioning of the aminoacid sidechains is concerned. However, to our knowledge, up to now, following the seminal work of Shemyakin and its collaborators, only small reverse-labeled all-D peptides, now called "retro-inverso" peptides, have been synthesized.

Because the number of compact conformations of the 27-mer is small enough so that the energy, for a given sequence, can be calculated for all them, the lowest-energy compact conformation can be determined. Then, starting from any non-compact conformation on the cubic lattice, the folding of 27-mers can be studied, using Monte Carlo simulations. Doing so, it was shown that for certain sequences the lowest-energy conformation can be reached within $5 \cdot 10^7$ steps, during which only a few dozens compact conformations are sampled, among the 103346 possible ones, while the choice of the initial conformation has little consequence, if any, on the outcome of the simulation. Such special sequences, which are able to fold rapidly and, as such, appear to behave like natural proteins, all proved to bear a pronounced energy gap between the lowest-energy and the first excited compact state, that is, the second lowest-energy compact conformation. When only two kinds of monomers are considered, either hydrophobic or polar ones, the sequence-structure relationship for the model can be exhaustively examined, and compared to what it is in the case of actual proteins. The usual choice for the energy function has the following form:

$$H = \sum_{i<j} E_{ij} \Delta(r_i - r_j),$$

where $\Delta(r_i - r_j) = 1$ if monomers $i$ and $j$ are close neighbours in the lattice, $\Delta(r_i - r_j) = 0$ otherwise, and $E_{ij}$ depends on the nature of the interacting monomers. A popular choice for the $E_{ij}$ values has been $E_{ij} = E_{HH} = -\epsilon$, when monomers $i$ and $j$ are both hydrophobic, and $E_{ij} = 0$ otherwise.
When the additive, more realistic, case is considered, that is, when:

\[ E_{ij} \approx E_i + E_j, \tag{3} \]

with for instance \( E_i = E_H = -1 \) and \( E_i = E_P = 0 \), the sequence-structure relationship in the model exhibits remarkable features. Noteworthy, if the energy is determined for all sequences for each of the 103346 conformations, it is found that only 122750 of them (0.09%) have non-degenerate ground states, that is, for each of these sequences a given conformation is the lowest-energy one, while for all of them the energy gap \( \Delta \) between this conformation and the second lowest-energy one is: \( \Delta = 2 \). Strikingly, out of the 103346 possible conformations only 120 (0.11%) are found to be possible, non-degenerate, ground states. All these so-called “remarkable structures” are characterized by a large “designability” \(^{15}\) as measured by the number \( N_s \) of sequences of which they are the ground state (\( N_s \) ranges between 513 and 2306). The fact that the number of possible folds in the above 27-mer model is limited is consistent with the current belief that the number of protein folds seems also to be limited. Though nearly a thousand folds are now known, new folds are discovered less and less often and it is usually estimated that there may not exist more than 10000 protein folds in nature.\(^\text{19}\)

Moreover, when the \( N_s \) sequences with a given fold are analyzed, properties similar as that of ensembles of protein sequences with the same fold are recovered. For instance, in half of the sequence sites hydrophobic and polar monomers are equiprobable\(^\text{15}\) while, as a consequence, pairs of sequences as different as pairs of random sequences are found to adopt the same fold.\(^\text{18}\)

In Fig. 3 the “top structure” \(^\text{15}\) is shown, that is, the conformation which is the ground state of the largest number of sequences. Among these 2306 sequences, PHP\(_4\)HPHPHP\(_9\)H is the one with the smallest number of hydrophobic monomers. As a matter of fact, each remarkable structure is the ground state conformation of a single five-hydrophobic monomers sequence and the corresponding five monomers are always located as shown in Fig. 3, that is, one of them being at the cube center, the four other ones being at the center of facets which are not bonded to the monomer at the cube center. Furthermore, in all sequences sharing a given ground state, these five monomers are all hydrophobic.\(^\text{18}\) This later point helps to clarify why there are 120 remarkable conformations. Indeed, if a sequence like PHP\(_4\)HPHPHP\(_9\)H has a non-degenerate ground state, this means that, out of the 103346 possible ones, there is only one way to bring its five hydrophobic monomers close together, so that each of them can interact with
at least another hydrophobic one. In other words, it is from a topological point of view that the 120 remarkable conformations of the $3 \times 3 \times 3$ cubic lattice model are “atypical”. This is a quite satisfactory property of the model, since it means that the same 120 conformations are also expected to be remarkable when different $E_{ij}$ values are chosen. For instance, with $E_{ij} = E_{HH} = -2 - \gamma$, $E_{ij} = E_{HP} = -1$, $E_{ij} = E_{PP} = 0$, and $\gamma = 0.3$, the conformation shown in Fig. 3 is still the top structure, but with $N_s = 3794$ (see Ref. 15). The picture is now more complicated than in the case of the additive potential. The difference between additive and nearly additive cases comes from the fact that the departure from additivity lifts the degeneracy of many sequences. Now energy gaps of $\Delta = 0 + n\gamma$ and $\Delta = 2 \pm n\gamma$ are observed, with $n = 1, 2, \ldots$ (see Ref. 21). For instance, with $\gamma = 0.3$, nearly half of the sequences whose ground state is one of the remarkable conformations have energy gaps lower than 1.0 (see Ref. 18). Moreover, while 4.75% of the sequences have a unique ground state, for the majority of them it is not one of the remarkable conformations but the corresponding energy gap is small, being on average close to 0.3, the $\gamma$ value. When only sequences with energy gaps larger than 1.0 are considered, the picture obtained with the additive potential is restored, despite a few minor differences arising from the overlap of the energy gaps splitting around the $\Delta = 0$ and $\Delta = 2$ cases. An interesting protein-like property of the model allows for the determination of all large gap sequences (i.e., with $\Delta = 2$, in the case of the additive potential) whose ground state is a given remarkable conformation, without the need of any huge enumeration. To do so, starting from the corresponding remarkable five hydrophobic monomers sequence, all 27 singly-mutated sequences are generated. Then, those among them with the same single ground state are retained. Next, for each sequence of this subset all 27 singly-mutated sequences are generated and so on, until no new sequence can be retained. What is shown through the success of such a protocol is that the $N_s$ sequences of each of the 120 remarkable structures belong to a “neutral island” of the sequence space. Note that such a property has also been found in the case of square lattice models, and that it is expected to be shared by ensembles of protein sequences corresponding to a given fold. Thus, the sequence-structure relationship for the above cubic lattice model seems to have much in common with the sequence-structure relationship observed in the case of natural proteins. This suggests that topology may play an important role in the folding process, allowing for a sequence to choose between significantly different folds, each characterized by a topo-
logically unique “folding nucleus”.23
As far as Levinthal’s paradox is concerned, studies of lattice models have
yielded a so-called “new view” of the protein folding process in which folding
is seen as a parallel microscopic multi-pathway diffusion-like process. In
other words, it is nowadays assumed that folding occurs through funneling
to a single stable state by multiple routes in conformational space.24

4.2. Off-lattice models

However, despite their many successes, cubic lattice protein models have
several drawbacks. First, from a chemical point of view, the cubic geometry
is a very unlikely one for a polymer, since on this lattice par définition
the angle between three consecutive monomers is either 90 or 180 degrees.
Second, using the Monte Carlo Metropolis algorithm for studying dynamical
processes, like the folding process, has clear limits. Indeed, in order to study
such processes, using molecular dynamics methods, that is, standard step by
step integration of the equations of motion, would be a more obvious choice.
But then, since the conformational space in which dynamics take place is
continuous, if the energy function is not chosen in an ad hoc way, knowing
the lowest-energy conformation of a given polymer becomes difficult as soon
as it is not a very short or highly constrained one.

Hereafter, a class of 3-D, off-lattice, protein models is briefly described, for
which this can be done almost as easily as in the case of cubic lattice models,
while the folding process of the polymer towards its lowest-energy confor-
mation can be followed as a function of time using molecular dynamics
methods.

Like in the case of lattice protein models, in the following a N-mer is mod-
eled as a chain of N beads linked together. Moreover, in order to retain
some of the major advantages of lattice models, $D$, the diameter of the
bead, is chosen to be equal to the average length of the bonds between two
linked beads. Together with a compaction term in the energy function, such
a choice allows to know, a priori, which is the approximate geometry of the
lowest-energy conformations of the polymer, as long as $N$ is well chosen.

For instance, on a flat surface, the most compact conformation of a 7-mer
has the shape of an hexagon, the hexagonal packing being the most com-
 pact possible one for a set of disks of same size. For the corresponding 3-D
models, though the most compact packing of a set of $N$ spheres is not that
obvious, it is often very well defined.25 For instance, a maximally compact
set of thirteen spheres is known to have the shape of an icosahedron, while
for nineteen spheres, a compact conformation of the corresponding polymer is shown in Fig. 4. So, for such a family of cases also, all hamiltonian paths as well as the energy gap of any given sequence can be determined. Then, starting from random conformations, folding simulations can be performed using molecular dynamics methods.

From a practical point of view, the energy function associated to such models can be a simplified version of standard empirical energy functions used when protein dynamics is studied at the atomic level, like that implemented in the CHARMM program, namely:

$$H = \frac{1}{2} k_b \sum_{\text{bonds}} (l_{ij} - D)^2 + \sum_{i>j} (\epsilon_{ij} + E_c)(\frac{D_{12}^{ij}}{r_{ij}^{12}} - 2\frac{D_6^{ij}}{r_{ij}^6}),$$

where $l_{ij}$ is the distance between bonded monomers $i$ and $j$, $r_{ij}$ the distance between non-bonded ones, $E_c$ the compaction term, while $\epsilon_{ij}$ allows for defining the interactions between different types of monomers. In practice, values like those used in cubic lattice model studies could be chosen, assuming for instance an additive description (see Eq. 3).

However, up to now, such Lennard-Jones chains have been considered in the case of homopolymers only, focusing on the collapse transition eventually as a function of an external stretching constraint, and often still using Monte Carlo methods for the sampling of the configurational space.
4.3. More detailed models

Of course, one could think of trying to study the protein folding process using brute force, by solving the equations of motion for a protein model at an all-atoms level of description, surrounded by explicit water molecules, as done in standard molecular dynamics studies of biological macromolecules. Here, the main limitation is the time step required for solving accurately enough the equations of atomic motion, which is in the femtosecond range. In practice, this limits the time scale accessible with such approaches to nanoseconds, though a few microseconds long simulations have been performed in the case of very small, fast-folding, proteins. Indeed, recently, interesting results were obtained in the case of a 36 aminoacids protein, using an “implicit solvent” model and a worldwide distributed computing network of tens of thousands of PCs. However, for such small proteins, results of a similar quality were also obtained using a much simpler protein description, Monte Carlo methods, and a single PC.

5. Protein conformational changes

5.1. Functional motions

The precise positioning of the aminoacids in space is vital for the ability of proteins to function and, noteworthy, for the specificity of the kind of function they perform. In particular, if enzymes are efficient catalysts, their main property, as far as their role in the very existence of the life phenomenon is concerned, is to be able to catalyze very narrowly defined chemical reactions. However, in order to perform their function, most, if not all, protein structures have to be flexible. For instance, when the first structure of a protein (the myoglobin) was obtained at a level of resolution allowing for the identification of the aminoacid sidechains, it was found that the ligand (the oxygen molecule) is buried deeply into the structure and that it would have no possibility to escape the binding pocket, if the protein were a perfectly rigid object.

More recently, using hydrogen/deuterium exchange as a tool for evaluating protein flexibility, it was found that the 3-isopropylmalate dehydrogenase enzyme from a thermophilic bacterium, which is hardly active at room temperature, is much more rigid at this temperature than the same enzyme from a normal, mesophilic bacterium, whereas these enzymes have nearly identical flexibilities under their respective optimal working conditions.
5.2. Collective motions

In many proteins, large conformational transitions involve the relative movement of almost rigid structural elements: a loop, a helix, a whole domain. Such motions are important for a variety of protein functions, including catalysis, regulation of activity, etc. For example, the binding of the coenzyme A on the citrate synthase enzyme, a two-domain protein, induces a $18^\circ$ rotation of the smaller domain around an axis close to aminoacid 274, which functions as the hinge. One consequence of this motion is the closure of the cleft between the two domains, in which the binding site lies.
Protein Functional Dynamics

allowing for the catalyzed chemical reaction to occur in a non-water, organic (sic) environment. As in most other cases (hexokinase, phage T4 lysozyme, etc) this motion was probed by X-ray crystallography. Fig. 5 shows one such simple large-amplitude, collective, motion, called “hinge-bending”, in which two structural domains of similar size move quite rigidly, so as to close a binding pocket. This protein binds charged aminoacids, namely, lysine (represented as van der Waals spheres), arginine or ornithine. Hence its name: the Lysine-Arginine-Ornithine (LAO) binding protein. The two conformations shown can be found in the Brookhaven protein data bank under the codes 2lao (top) and 1lst (bottom). In this case, distances between pairs of aminoacids vary by up to 20 Å.

A more spectacular case is that of calmodulin, a protein of the EF-hand family, whose members act as Ca$^{2+}$ sensors. Calmodulin reacts to transient increases in cellular calcium concentration occurring during a variety of cellular activities, ranging from cell cycle control to muscle contraction. As shown in the top part of Fig. 6, its Ca$^{2+}$-bound conformation has a

![Calmodulin](image-url)
dumbell shape, with two globular domains linked by a long helix. (Here, van der Waals spheres stand for calcium ions.) One of the effects of Ca\(^{2+}\) binding is the exposure to the solvent of hydrophobic patches that form the binding site of target proteins. Indeed, Ca\(^{2+}\)-bound calmodulin is able to bind to more than one hundred different target proteins, and it does so by breaking its long central helix, allowing for the hydrophobic patches to grasp an hydrophobic part of the target protein, usually an \(\alpha\)-helix, as shown in bottom part of Fig. 6 (the black moiety). In this case, distances between pairs of aminoacids vary by up to 50 Å.

Because it is that large, the conformational change of calmodulin can be observed out of the crystal state, namely, in solution, by X-ray scattering, and it is one of the few cases where direct measurements of the kinetics...
of the conformational change have been performed. For instance, using fluorescence resonance energy transfer (FRET) between two chromophores, each attached to one domain, it was possible to show that the timescale of domains separation, when the bound peptide is released (the "off-rate"), is in the tens of a second range.\textsuperscript{38}

Except in a few cases, conformational changes take place in proteins without any modification of the secondary structural elements: the lengths of the $\alpha$-helices and of the $\beta$-strands are the same before and after the motion. Also, most interactions between secondary structural elements are preserved and the motion can often be described as a relative motion of two “blocks” of structural elements, the “domains”, with respect to each other, like in Figs. 5 or 6. An interesting exception is the recently discovered “domain-swapping” conformational change, which can occur between members of a dimer, that is, a pair of identical proteins associated together at physiological concentrations. Here, as shown in Fig. 7 in the case of the CD2 protein, the association can either be a standard, simple contact between the two monomers (top), or the two monomers can be intertwined (bottom). While the packing of the secondary structural elements is the same, it is obtained through contacts between elements belonging to different monomers, the difference between the two conformations being a change in the configuration of a single, flexible, loop which in the later case “jumps” from a monomer to the other.

5.3. \textit{Low-frequency normal modes}

One of the best suited theoretical methods for studying collective motions in proteins is the normal mode analysis (NMA), which leads to the expression of the dynamics in terms of a superposition of collective variables, namely the normal mode coordinates. The principles underlying NMA are briefly recalled below.

5.3.1. \textit{Normal mode analysis}

This kind of analysis is based on the following ideas.\textsuperscript{39} In the vicinity of a stationary point, the potential energy $V$ of a system of $N$ atoms, can be approximated by:

$$V = \frac{1}{2} \sum_{i=1}^{3N} \sum_{j>i}^{3N} k_{ij} (r_i - r_i^s)(r_j - r_j^s),$$  \hspace{1cm} (5)
where $k_{ij}$ is the mass-weighted second derivative of the potential energy with respect to coordinates $r_i$ and $r_j$, and where $r_i^s$ and $r_j^s$ are the $i$ and $j$ coordinates of the stationary structure under study. Within the frame of this approximation, the equations of motion can be solved analytically, leading to the following solutions:

$$r_i(t) = r_i^s + \frac{1}{\sqrt{m_i}} \sum_{j=1}^{3N} a_{ij} q_j(t),$$

with:

$$q_i(t) = C_j \cos(\omega_j t + \phi_j),$$

where $m_i$ is the atomic mass, $a_{ij}$ the $i^{th}$ coordinate of normal mode $j$ and $\nu_j = \frac{\omega_j^2}{2\pi}$ the corresponding frequency. $\omega_j^2$ is the $j^{th}$ eigenvalue of the Hessian, that is, the $3N \times 3N$ matrix built with the $k_{ij}$'s, while $a_{ij}$ is the $i^{th}$ coordinate of the corresponding eigenvector. $C_j$ and $\phi_j$, the amplitude and phase of mode $j$, are determined once the coordinates and the velocities of the system at $t = 0$ are known. Noteworthy, $C_j$ is such that:

$$C_j = \sqrt{\frac{2E_j}{\omega_j}},$$

where $E_j$ is the amount of energy in mode $j$, which according to the equipartition principle is expected to be at a given temperature $T$: $E_j \simeq k_B T$, where $k_B$ is the Boltzmann constant. One can see that at a given temperature the lower the frequency of a mode, the larger its amplitude. Typically, the normal modes whose frequencies lie under $30-100$ cm$^{-1}$ are found to be responsible for most of the amplitude of the atomic displacements in proteins.$^{30,41}$

From a practical point of view, the potential energy of the studied protein is first minimized, using standard algorithms that seek for minima close to the starting structure. Such are for example the Powell or the Newton-Raphson algorithms. Note that during this process each atom in the structure drifts by $1-2$ Å on average. Then, the Hessian is diagonalized. Due to its size, this step used to be the technically limiting one. Indeed, though the normal mode analysis of the small, 58 aminoacids BPTI protein was performed as soon as 1982 (see Ref. 42) ten years later the largest protein studied at the atomic level of description was still the 153 aminoacids myoglobin,$^{43}$ while most interesting proteins are much larger. Since then, efficient algorithms were designed (e.g. DIMB, Ref. 44) or adapted to the case of macromolecular assemblies (e.g. the block Lanczos approach, Ref. 45) in
order to compute the lowest-frequency normal modes, which proved to be the most informative ones.

5.3.2. The RTB approximation

Instead of diagonalizing the Hessian, $H$, as in standard NMA, the principle of the RTB method is to diagonalize $H_b$, a $6n_b \times 6n_b$ matrix build as follows:

$$H_b = P^t HP,$$

where $P$ is an orthogonal $3N \times 6n_b$ matrix composed of the vectors describing the six rigid-body rotations and translations of each of the $n_b$ blocks the protein is split into. For instance, each block can contain a single aminoacid. $A_p$, the $3N \times 6n_b$ matrix with the $6n_b$ approximate lowest-frequency normal modes of the protein, is then obtained as follows:

$$A_p = P A_b,$$

where $A_b$ is the matrix diagonalising $H_b$, $A_b$ being obtained with standard diagonalisation routines.

Of course, the RTB approximation can only be used for calculating modes in which aminoacids behave almost rigidly. Even in that case, calculated frequencies are found to be higher than exact ones, reflecting the fact that aminoacids can not adapt their conformation so as to make the whole motion easier. However, for frequencies lower than 40 cm$^{-1}$ it was shown that, when one aminoacid is put in each block, a linear relationship holds between approximated and exact frequencies, that is:

$$\nu_{rtb} \cong d_P \cdot \nu_s,$$

where $\nu_s$ and $\nu_{rtb}$ are the frequencies obtained using, respectively, standard approaches or the RTB approximation. In the case of a set of proteins of various sizes, using the standard CHARMM force field and a 8.5 Å cutoff for electrostatic interactions, it was found that $d_P$ does not depend upon protein size or fold type ($d_P \approx 1.7$). This allows to get fair estimates for exact frequencies, once the approximated ones are known.

5.3.3. Comparison with crystallographic B-factors

Frequencies are necessary in order to obtain atomic fluctuations, which themselves can be compared to crystallographic B-factors (the “temperature factors”). Indeed, according to NMA, $B_i$, the crystallographic B-factor
for coordinate $i$ is as follows:

$$B_i = \frac{8\pi^2 k_B T}{3 m_i} \sum_{j=1}^{n} \frac{a_{ij}^2}{\omega_j^2},$$

(7)

where $n$ is the number of low-frequency normal modes retained for the calculation. Usually, the six rotational and translational rigid-body motions of the whole system are not included in the calculation. However, it is well known that crystal static disorder contributes significantly to the experimental value of B-factors. This contribution can be partly described in terms of rigid-body rotations and translations of the whole system within the crystal cell, assuming for instance the corresponding frequencies to be identical, very low, but non-zero.

However, even without taking static disorder into account, B-factors calculated according to Eq. 7 are often found to be well correlated with experimental values. A typical example is shown in Fig. 8, in the case of the open form of the LAO binding protein (see top of Fig. 5), where calculated (dotted line) and experimental (plain line) values are given as a function of the aminoacid number. Here, calculated values were scaled so as to have...
same average and root-mean-square than experimental ones.

![Graph showing correlation between amino-acid displacements and normal mode amplitudes.](image)

**Fig. 9.** LAO binding protein: the lowest mode versus the conformational change.

### 5.3.4. Comparison with conformational changes

Normal mode theory has also been shown to give a fair picture of X-ray diffusive scattering of protein crystals.\(^48\) However, the idea that NMA may be an accurate tool for studying protein conformational changes comes from the fact that in several cases, e.g., hexokinase,\(^49\) lysozyme,\(^50,51\) citrate synthase,\(^45\) *etc.*, the largest amplitude motion obtained with this theory, that is, the one with the lowest frequency, was found to compare well with the conformational change observed by crystallographers in these proteins upon ligand binding. Indeed, as shown in Fig. 9, in the case of the LAO binding protein conformational change shown in Fig. 5, there is a clear correlation between the observed aminoacid displacements and the amplitude of the aminoacid motions in the lowest-frequency normal mode of this protein, as calculated for the “open” form of the LAO binding protein (top of Fig. 5). The calculation was performed in the absence of the lysine ligand, whose position in the open form is not known, at variance with what suggests Fig. 5, since it induces the closure motion. Here, \(C_j\), the amplitude of the normal mode (see Eq. 6), was set to one, and data are shown for \(\alpha\)-carbons only, as representatives of aminoacid motions.
In order to quantify how well a conformational change is described by a given normal mode, one can calculate $I_j$, the overlap between $\vec{\Delta r} = \{\Delta r_1, \cdots, \Delta r_i, \cdots, \Delta r_{3N}\}$, the conformational change observed by crystallographers, and $a_j$, the $j^{th}$ normal mode of the protein. This is a measure of the similarity between the direction of the conformational change and the one given by mode $j$. It is obtained as follows:\[45\]

$$I_j = \frac{\sum_{i=1}^{3N} \Delta r_i a_{ij}}{\left(\sum_{i=1}^{3N} \Delta r_i^2 \sum_{j=1}^{3N} a_{ij}^2\right)^{1/2}}$$

where $\Delta r_i = r^o_i - r^c_i$, $r^o_i$ and $r^c_i$ being, respectively, the $i^{th}$ atomic coordinate of the protein in the “open” crystallographic structure and in the “closed” one. A value of one for the overlap means that the direction given by $a_j$ is identical to $\vec{\Delta r}$. From a practical point of view, $\vec{\Delta r}$ is calculated after both crystallographic conformations of the protein were superimposed, using standard fitting procedures. These pairs of conformations are often referred to as “open” or “closed”, because many known conformational changes involve the closure of a binding pocket site. Note that $O_s(n)$, the cumulative square overlap, calculated as:

$$O_s(n) = \sum_{j=1}^{n} I_j^2$$

is equal to one when $n = 3N$, that is, when all modes are taken into account, since the $3N$ modes form a basis set.\[39\] In the case of the LAO binding protein conformational change, $I_1$, the overlap with the lowest-frequency motion is: $I_1 = 0.90$, and $O_s(50) = 0.90$, which means that 90% of the atomic displacements observed during the conformational change can be described with 50 coordinates only, namely, the 50 normal coordinates corresponding to the 50 lowest-frequency normal modes. Note that there are 2253 atoms in the system considered here, that is, 6759 coordinates. The fact that atomic displacements corresponding to a protein motion with a high “collective” character, that is, a motion in which many atoms are involved, can be accurately described with a small subset of low-frequency normal coordinates is not a surprising result, because low-frequency normal coordinate themselves have such a collective character. However, the fact that one, or a few, of them may prove enough for obtaining a fair description of a conformational change was not \textit{a priori} expected. On the contrary, many good reasons can be found in favor of the opinion that it should not be so. For instance, from a physical point of view, the energy function
used to compute protein normal modes is a very approximate one, and it is quite sure that frequency values would be significantly different, if it were possible to compute them at an ab initio level. Moreover, the low-frequency part of protein normal mode spectra is not characterized by clear gaps, as shown in Fig. 10, in the case of the open form of the LAO binding protein.

![Fig. 10. The low-frequency normal mode spectrum of the LAO binding protein.](image)

Indeed, in the case of such spectra, it was shown that $g(\omega)$, the density of states, follows a characteristic, universal curve with, for the lowest-frequency values:

$$g(\omega) \sim \omega$$

Furthermore, from a biological point of view, proteins are known to fold and function in a water environment, within a narrow range of pH, temperature, ionic strength, etc, while NMA is performed in vacuo. Also, standard NMA requires a preliminary energy minimization which drifts the atoms of the protein up to a few Å away from their position in the crystallographic structure. As a consequence, the structure studied with standard NMA is always a distorted one. More generally, NMA is based on a severe small displacements approximation, which amounts to suppose that a protein behaves like a solid does at low temperature, while it is well known that a protein is a somewhat flexible polymer, undergoing many local conformational transitions at room temperature.

5.3.5. *Simplified potentials*

Recent results have shed some light on this paradox. Noteworthy, it was shown that using a single parameter hookean potential for taking into ac-
count pairwise interactions between neighbouring atoms yields results in good agreement with those obtained when NMA is performed with standard semi-empirical potentials, as far as low-frequency normal modes are concerned.\textsuperscript{53,54} More specifically, within the frame of the approach proposed by M. Tirion, the standard detailed potential energy function is replaced by:

\[ E_p = \sum_{d_{ij} < R_c} C(d_{ij} - d_{ij}^0)^2, \]

where \( d_{ij} \) is the distance between atoms \( i \) and \( j \), \( d_{ij}^0 \) being the distance between these two atoms in the given studied crystallographic structure. The strength of the potential \( C \) is a phenomenological constant assumed to be the same for all interacting pairs. This energy function was designed so that for any chosen configuration the total potential energy \( E_p \) is a minimum of the function. Thus, with such an approach \textit{par définition} NMA does not require any prior energy minimization.

Note that in Eq. 8, the sum is restricted to atom pairs separated by less than \( R_c \), which is an arbitrary cut-off parameter. When, as proposed by Ivett Bahar and her collaborators, only the \( C_\alpha \) atoms are taken into account,\textsuperscript{55} a cut-off of 8-10 Å can for instance be used. Such a network of harmonic springs, shown for the case of the open form of LAO binding protein in Fig. 11, is enough to study backbone motions, which in turn proves sufficient
for characterizing low-frequency normal modes of proteins. Moreover, it allows for studying proteins of large size on common workstations, using small amounts of CPU time, since, with this simple model, the matrix to be diagonalized is a $3N_r \times 3N_r$ one, where $N_r$ is the number of aminoacids of the protein.

Using this kind of highly simplified potential yields low-frequency normal modes in good agreement with those obtained using standard NMA. Moreover, when the interactions between closely located $\alpha$-carbon pairs are described by a gaussian network model, crytallographic temperature B-factors are also found to be accurately predicted. Again, this means that low-frequency normal modes of proteins are well described with such models. Indeed, like when detailed semi-empirical energy functions are used, a few low-frequency normal modes are often found to yield a good description of the functional motion of the protein, especially when the corresponding conformational change has a highly collective character.

Thus, results obtained with NMA in the field of low-frequency protein dynamics seem to be of a very good quality even when most atomic details are simply ignored. This means that the low-frequency normal modes of a protein depend mainly upon its shape, that is, upon the distribution of its aminoacids in space. Reciprocally, because protein conformational changes are often found to be well described by one or by a few such modes, it seems that during the course of evolution proteins have taken advantage of these solid-like motions for finding paths in configuration space that allow for conformational changes to proceed within reasonable time spans.

6. Dissipation of energy in proteins

Though protein dynamics at room temperature is certainly not an harmonic one, the harmonic picture has often proved useful for analyzing it. For instance, when temperature jumps are performed during a protein molecular dynamics simulation, if the total kinetic energy of the protein is monitored during the following hundreds of femtoseconds, “echoes” are observed (coined “temperature echoes”), which can be explained in terms of the dynamics of a set of damped oscillators. Unfortunately, the question of how well protein dynamics can be described as the dynamics of a set of coupled harmonic oscillators has not yet been fully worked out, maybe because of a too important focus on slow motions, whose behaviors are known to be highly anharmonic, involving energy barrier crossings, etc, or because
of a reluctance to study in depth high-frequency protein motions, which are expected to behave quantum-mechanically at room temperature. However, non-linear energy localisation may well have already been observed in proteins, in the course of molecular dynamics simulations. Indeed, when analysing the very first simulation of a protein at an all-atoms level of description, McCammon, Gelin and Karplus noticed that the kinetic energy of a few aminoacids remained higher than expected from the equipartition principle by 10-20% for more than 9 picoseconds, concluding that “vibrational energy can localise for relatively long time periods in proteins”.

Experimentally, though there have been attempts to follow how the energy given to a protein, usually through laser impulses on an aromatic moiety, is redistributed within the protein, probing energy paths within such large and complex systems remains problematic. In this respect, studies of model peptides may prove useful. For instance, recently the kinetics of energy release within a cyclic peptide was followed, taking advantage of the insertion in the peptide of a bi-aromatic bistable ring with clearly different absorption spectra in each of its two possible conformations.

7. Conclusion

Simple models have proved useful for studying the two major kinds of protein functional motions, namely their folding and their conformational changes. Up to now, the folding process has mainly been studied using lattice models. However, Lennard-Jones chains are an obvious generalisation for such models if one wishes to characterise dynamical aspects of the process, like the way energy is redistributed in the structure, as a function of its progress towards lower and lower energy configurations.

Models considered for studying conformational changes, in which a protein is figured as a network of harmonic springs, could also be generalised in a straightforward way, by including non-linear terms, so as to explore possible links between no-linear energy localisation phenomenons and functional protein motions, which have been shown to be often very well described with a few of their lowest-frequency normal modes.

Acknowledgments

I thank the LOCNET members for their interest as well as for their numerous and stimulating questions. I have tried to answer some of them above. For several of the many remaining ones: work is still in progress...
References