IL NUOVO CIMENTO Vol. 32 C, N. 2 Marzo-Aprile 2009
DOI 10.1393/ncc/i2009-10363-2

Colloquia: CSFI 2008

Protein folding: Can high-performance computing improve our understanding?

G. Tiana

Dipartimento di Fisica, Università di Milano and INFN
via Celoria 16, 20133 Milano, Italy

(ricevuto il 15 Maggio 2009; pubblicato online il 3 Agosto 2009)

Summary. — Proteins are complex physical systems of great biological and pharmaceutical interest. Computer simulations can be useful to understand how they fold to their biologically active conformation, but have to face two problems, namely the roughness of the energy landscape and the wide range of time scales associated with the folding process. Models at atomic detail are able to describe the protein with a high degree of realism, but are computationally very demanding and their results usually are difficult to analyse. Models with simplified degrees of freedom are less accurate but are good at highlighting the basic physical mechanism which controls protein dynamics. A combination of the two can be the right solution to the protein folding problem.

PACS 87.15.hm – Folding dynamics.

The study of proteins is of utmost interest because they control every aspect of cellular metabolism, and thus of life. For this reason, they are the target of most pharmaceutical drugs and are often studied with this goal. Proteins are polymers built out of twenty kinds of amino acids. At biological temperature (0–50°C), they can fold into a unique equilibrium conformation which is encoded in the specific sequence of amino acids of the protein. This is called native conformation and is responsible for the biological function of the protein. The folding process is very rapid, taking place on a time scale ranging typically from ms to s.

Given a protein sequence, one would like to predict what the native conformation is, and to understand the physical mechanism the protein follows to reach such a conformation. The source of all problems in the study of proteins is that they are frustrated systems, which means that the interactions between amino acids are so heterogeneous that some unfavourable interactions remain inevitably in the native conformation, which is the conformational ground state of the system. It is known from statistical mechanics that frustrated systems typically display a rough energy landscape [1]. In fact, the energy landscape of a random sequence of amino acids looks like that displayed in the upper right panel of fig. 1, with many minima at comparable energies. This is right the opposite of what one would expect from a protein: there is not a thermodynamically

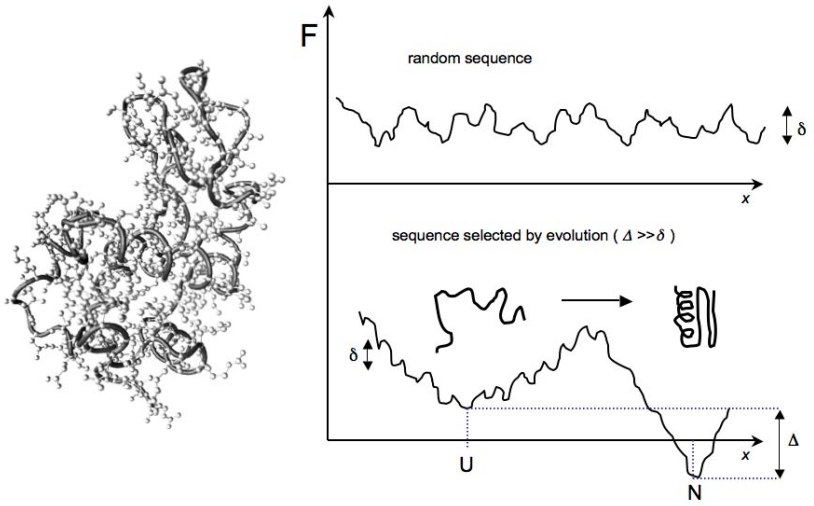

Fig. 1. – Left panel: the native state of lysozyme, a small protein of 129 amino acids; upper right panel: the energy landscape of a random sequence of amino acids, where the energy scale of roughness is δ; lower right panel: the energy landscape of a protein, in which evolution has sculpted a new energy scale Δ between the native conformation and the competing energy minima.

dominant state, and if one studies the system at a temperature comparable to the only energy scale of the system, that of roughness, the kinetics is slow. But a protein is not a random sequence. It has undergone a very long evolution, and the effect of this is to minimise its frustration, in such a way to display in the native conformation an energy lower than that of the competing energy minima (see lower right panel of fig. 1). The system has now two energy scales: δ, corresponding to residual frustration, and Δ, which is the energy difference between the native and the competing conformations. Consequently there exists a range of temperatures T with $\delta < kT < \Delta$ such that the native state is thermodynamically dominant and the kinetics is fast.

But frustration has not completely disappeared: it is the residual frustration that makes the study of protein folding complicated, especially when carried out through computer simulations. For example, as in all frustrated systems, it is not possible to predict the conformational ground state (*i.e.*, the native conformation) with any fast algorithm [2]. Moreover, the stability of the protein is anyway marginal, the value of Δ amounting typically to $10kT$, and the residual energy minima can sometimes trap the chain in metastable states.

Another computational problem is associated with the wide range of time scales involved in protein folding. The overall folding, starting from a disordered conformation, takes place on the time scale of ms to s; the diffusion of the different segments of the chain and the formation of secondary structures take place in μs; the hydrophobic collapse of the chain in ns, the motion of the side chains of the amino acids in ps, the atomic vibrations in fs. Simulating protein folding using, for example, a molecular dynamics (MD) algorithm requires to use a time step which is smaller than the frequency of the fastest degree of freedom one wants to describe, in this case femtoseconds. Simulating the full protein folding mechanism needs then 10^{12}–10^{15} time steps.

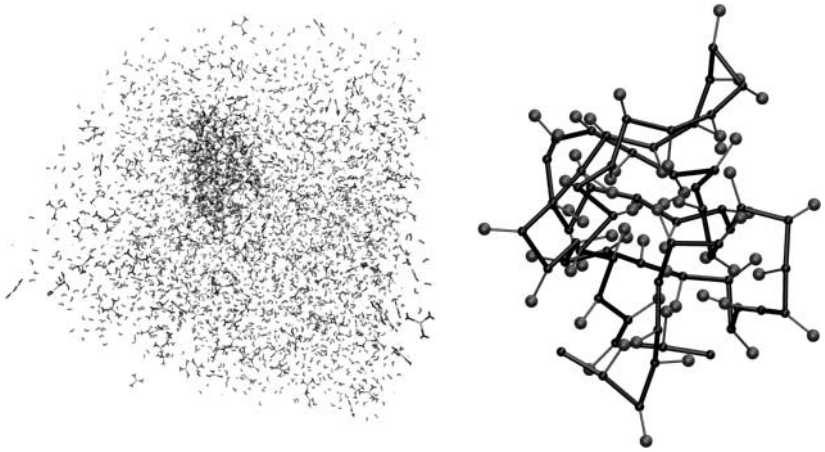

Fig. 2. – Left panel: An all-atom protein model in explicit solvent; right panel: a simplified model with reduced degrees of freedom.

All-atom models with explicit solvent are often used to simulate protein folding because they reproduce quite realistically their degrees of freedom and the interaction among amino acids, but suffer the problems described above. Since the first simulations of few picoseconds done in the sixties [3], to refine the structure obtained by crystallographic experiments, things have greatly improved, due to the availability of fast computers and algorithms. At the end of the nineties, it was possible to simulate approximately one tenth of the folding time of villin, the smallest known protein, built out of 36 amino acids [4]. This took approximately two months on 256 processors of a Cray T3E. But even if one is able to simulate once the full folding trajectories of a protein, this gives very little information. In fact, protein folding is a stochastic process and one has to average over many trajectories to obtain data which are statistically sound. Recently, within the folding@home project [5], consisting of a world-wide network of volunteering fast PCs, it was possible to generate 410 trajectories of $\sim 1\,\mu$s each, thus solving the problem of statistics. One should anyway note that a protein of 36 amino acids is not representative of typical proteins, which are typically built of several domains of hundreds of amino acids each.

The situation is even worse if one is interested in obtaining from MD simulations the equilibrium thermodynamics of a protein. In this case, the simulation should be able to sample exhaustively most of conformational space, something which is computationally more demanding than just simulating few folding trajectories. For this purpose, it is the availability of new algorithms [6] rather than the increase in computational power, that allows to obtain equilibrium quantities of, at least, small proteins like villin.

Another approach is to use simplified models, with reduced degrees of freedom. An example is that of fig. 2, where each amino acid is represented with two beads, one accounting for its main chain and the other for its side chain. The main problem with this kind of models is the difficulty to write a potential function which reflects the physics of the different amino acids and allows the model protein to fold. A way to overcome it is to design a potential function to reproduce some experimental facts associated with the specific protein one wants to study. A common choice is to use a Lennard-Jones

potential in the form

$$U = \sum_{ij} \left(\frac{a_{ij}}{r_{ij}}\right)^{12} - \left(\frac{b_{ij}}{r_{ij}}\right)^{6}, \tag{1}$$

where the parameters a_{ij} and b_{ij} are tuned to reproduce the experimental data available (like, *e.g.*, the experimental native conformation as ground state of the model, the energetic effects of point mutations, the folding temperature, etc.).

This kind of models is computationally so light that can be easily run on a PC. Nonetheless they can provide an important insight into the folding mechanism of proteins. For example, they indicate that a consistent amount of local native structure is usually present already at the beginning of the folding process, in the unfolded state, and that this local structure is critical to drive the protein to its native conformation [7]. This fact has been exploited to design a new kind of drugs which block protein folding. This result is obtained using as Trojan horses some peptides with the same sequence as the segments which are structured in the unfolded state [8]. This strategy has been used with success against HIV protease, a protein which is critical in the replication of HIV [9].

Of course, this kind of simplified models misses many details of the folding process, but can provide the basic physics which controls protein folding, something which is necessary to understand the general picture associated with it. Massive simulations with realistic models cannot do it because they are still computationally too demanding. But even if they could simulate repeatedly the folding of realistic-size proteins, it would be very difficult to extract from the dynamics of hundreds of thousands of atoms some physical insight. Simulating a phenomenon is not equivalent to understanding it. The ferromagnetic phase transition, for example, can be understood through the simplified Ising model, not through detailed quantum simulations of the ferromagnet. The use of computationally demanding, detailed models is useful afterwards, when a physical insight has already been reached, in order to refine the understanding of the behaviour of specific proteins.

It is not unreasonable to expect that the combination of the two approaches can magnify their potentialities. This is the case, for example of multiscale methods, in which proteins are modelled at the same time at multiple resolution [10], or of metadynamics, in which the insight coming from simplified models is used to define the collective variables to control the sampling in the detailed model [6, 11].

REFERENCES

[1] Vannimenus J. and Toulouse G., *J. Phys. C*, **10** (1977) L537.
[2] Mezard M., Parisi G. and Virasoro M. A., *Spin Glasses and Beyond* (World Scientific) 1987.
[3] Levinthal C., *Sci. Am.*, **214** (1966) 42.
[4] Duan Y. and Kollman P. A., *Science*, **282** (1998) 740.
[5] Shirts M. and Pande V., *Science*, **290** (2000) 1903.
[6] Laio A. and Parrinello M., *Proc. Natl. Acad. Sci. U.S.A.*, **99** (2002) 12562.
[7] Sutto L., Tiana G. and Broglia R. A., *Protein Sci.*, **15** (2006) 1638.
[8] Broglia R., Tiana G. and Berera R., *J. Chem. Phys.*, **118** (2003) 4754.
[9] Broglia R. A., Provasi D., Vasile F., Ottolina G., Longhi R. and Tiana G., *Proteins*, **62** (2006) 928.
[10] de Mori G., Colombo G. and Micheletti C., *Proteins*, **58** (2005) 459.
[11] Camilloni C., Provasi D., Tiana G. and Broglia R. A., *Proteins*, **71** (2008) 1647.