IL NUOVO CIMENTO Vol. 32 C, N. 2 Marzo-Aprile 2009
DOI 10.1393/ncc/i2009-10395-6

Colloquia: CSFI 2008

Structural predictions of biomolecular systems

F. Fogolari(*)

Dipartimento di Scienze e Tecnologie Biomediche Università di Udine
Piazzale Kolbe, 4 33100 Udine, Italy

(ricevuto il 4 Agosto 2009; pubblicato online il 7 Settembre 2009)

Summary. — In this short paper we summarize the current landscape in structural predictions of biomolecular systems and underlying physical principles. The target molecules of predictions are shortly introduced and a summary of current methods for structural characterization is given. Basic principles and methods used in structural predictions are finally summarized.

PACS 87.14.E- – Proteins.
PACS 87.15.B- – Structure of biomolecules.
PACS 87.15.Cc – Folding: Thermodynamics, statistical mechanics, models, and pathways.

1. – Introduction [1-3]

Most important biomolecules, proteins and nucleic acids, are linear heteropolimers. Linearity and chirality allow to represent their chemical structure with a string of characters forming their so-called sequence or primary structure. Each character represents one of four nucleotides for DNA and RNA and one of 20 amino acids for proteins. Sequencing of DNA or RNA from different organisms and individuals (or different cells in different conditions), *i.e.* determining their chemical structure, has become in recent years fast and relatively cheap. The number of sequences deposited in the publicly available databases is now ca. 100 million entailing 100 billion characters.

The information coded in the DNA sequence is translated to assemble protein chains. In turn proteins perform their function due to their peculiar structure. For this reason knowledge of protein structure allows understanding their potential function and, when the latter is known, understanding of their biological mechanisms.

In this picture the double-helix structure formed by complementary strands of DNA has been mostly considered in connection with its capability of providing a template for

(*) E-mail: federico.fogolari@uniud.it

replication, although in recent years the role played by (mostly) RNA and DNA structure has been increasingly recognized.

Experimental characterization of biomolecular structure is time consuming and often difficult. The two main techniques employed for structure determination are X-ray crystallography contributing ca. 86% of all experimental structures deposited in the Protein Data Bank (PDB) and solution Nuclear Magnetic Resonance (NMR) contributing the remaining 14%. Although the size of system studied by NMR is limited compared to X-ray crystallography and structure determination is not free from artifacts, the technique is widely used to characterize also biomolecular dynamics.

When the number of known sequences is compared to the number of protein, RNA and DNA structures solved to date (ca. 60000 structures) the enormous gap between sequences and structures known is apparent. This is even more evident when some class of proteins (*e.g.*, membrane proteins) which are not amenable to crystallization or to NMR studies is considered.

It is worth to say that notwithstanding the structural genomics project already started, no big advance in automation for structure determination has been seen.

With the advent of more and more powerful and cheap computers however structural predictions and simulation could provide an alternative to structure determination. The fundamental challenges for computational chemists and physicists in the next years will be the prediction of the structure of biomolecules and their complexes, the prediction of their dynamics and the computation of kinetic and thermodynamic quantities, like kinetic constants and binding free energies, based on the predicted molecular models.

2. – Methods

The problem we address is the prediction of biomolecular structure, *i.e.* given the chemical structure of, say, a protein finding its structure. This is a reduced version of the problem of finding its folding pathway, bringing the protein from a disordered conformation to its more stable, active (native) form. This process takes place in a time larger than milliseconds even for small proteins, ruling out at the moment the usage of molecular dynamics simulations.

2˙1. *Homology modeling*. – The principles underlying molecular modeling are mostly based on recognition of similarity between the sequence whose structure is to be predicted and some sequence whose structure is known. It was observed very early that protein sequences sharing a large number of identical corresponding amino acids have very similar structures. This observation stems from the strong conservation of structure during evolution of proteins. When similarity beyond random expectance is found between two sequences, descendance from a common ancestor protein (homology) may be inferred and conservation of structure may be assumed [4]. This oservation constitutes the principle of homology modeling, where the common chemical parts, *e.g.*, backbone atoms and common side chains, are copied from the similar target and the rest is rebuilt using database fragments or information [5, 6].

2˙2. *Fold recognition* [5, 7]. – Unfortunately it is not always possible to find a homologous sequence whose structure is known. More refined schemes of comparison are preferred over simple sequence-sequence comparison. Profiles (*i.e.* common statistical features of a set of homologous sequences) are used instead of the sequences.

When no clearly homologous template structure can be found, it is still possible to find one or more template structures based on so-called fold recognition. The basic principle is that, although the number of different sequences may be very large, the number of folds that a protein sequence may assume is limited. The number of folds vary with definition and clustering methods and threshold, but most used current classifications define approximately 3000 different folds [8,9]. The rate of novel fold discovery is lower and lower, with no new fold superfamily deposited last year (see, *e.g.*, URL `http://www.rcsb.org/pdb/statistics/contentGrowthChart.do?content=supfam-scop`). Sequence and the propensity of the sequence to adopt local conformations may be compared to the features of each different fold in order to check whether the sequence could adopt that fold. Finding the best structure template and best correspondence between sequences is often difficult so that the output of different programs is combined in a consensus scheme and several different alternatives are tested.

2˙3. *Ab-initio modeling* [10, 11, 7]. – When template-based methods fail it is possible to model the structure using *ab initio* methods. Most successful methods assemble the structure to be modeled from fragments of real structures deposited in the PDB. Candidate fragments are identified using sequence similarity. Monte Carlo methods are used to this aim using simplified molecular representations and adhoc moves. Even for moderate size proteins the search for native conformation may take several days and the methods are still not practical for sequences longer than 100 amino acids. For smaller sequences the accuracy of the models has been sometimes impressive.

2˙4. *Quality assessment of the models*. – Based on Anfinsen's hypothesis the native structure of the protein is the one that minimizes Gibbs' free energy of the system. All scoring systems for ranking the accuracy of predictive models can be related to free-energy estimation.

It is worth to remind that here we should in principle estimate Gibbs' free energy of the macrostate corresponding to each model of the protein relative to the macrostate corresponding to the unfolded state. Although such states cannot theoretically be clearly defined, in practice there are many, mainly spectroscopic, techniques which can distinguish the two states. In practice when different models for the same protein are given we can simply compare the free energy of models relative to each other in order to find the most stable one.

Gibbs' free-energy estimation is difficult because it involves entropic and solvation effects. Two main classes of energy functions are used: physical effective energy functions (PEEFs) and statistical effective energy functions (SEEFs) which are based on different ideas and approximations [12].

PEEFs estimate Gibbs' free energy making typically some approximations:

1) the potential of mean force involving only the protein coordinates is written as the sum of an intrasolute term, estimated using one of the available forcefield for molecular dynamics simulations, and a solvation term. The solvation term is estimated from the solvent accessible surface area [13] and electrostatic free-energy calculations [14].

2) Each model is considered as representative of a macrostate corresponding to an ensemble of models with the same energy as the model itself. Furthermore it is assumed that the entropic unfavorable contributions to the folding free energy for

each macrostate is constant or sometimes that it can be estimated based on solvent exposure of the atoms defining each torsional degree of freedom.

The approach has been sometimes successful but it is flawed with many problems ranging from inaccuracy of available forcefield and of solvation models to the detrimental effect that minor building errors have on the overall quality estimation of the whole model. Moreover energy differences of few kcal/mol are obtained as difference between energies as large as thousand of kcal/mol.

In this respect SEEFs appear more robust. Here the information found in the PDB is used to select descriptors of the structures which range from presence or absence of contacts to interatomic distances [15]. A scoring system (or potential) may be defined using different criteria, *e.g.*, heuristically in order to maximize the gap between real structures and wrong models, or by taking the logarithm of the ratio of the observed frequency of that feature over the expected one. In the latter case a proper reference state must be defined.

3. – Future perspectives

Current limitations to the structural prediction methods described above reside mainly in the limited size of the systems amenable to such studies. Membrane protein structures, which are of capital interest for pharmaceutical industry, are still difficult to model in general due to the paucity of experimentally solved structures. It is somehow surprising that structural modeling has not entered clinical genetics practice as yet where it could explain or at least provide clues to the effect of a single mutation on a protein's stability and its interactions.

Although the structure-function relationship of proteins (and nucleic acids) is central to understanding molecular biology, the recent applications of genomics, transcriptomics and proteomics approaches to the study of living systems calls for the long-term challenge of predicting entire networks of interactions. Outstanding advancements have been obtained in predicting protein-protein complexes and in modeling molecular complexes rather than single molecules. In the next years we could see a scale change in structural predictions.

* * *

This work was supported by Italian Ministry of University and research through grants FIRB-RBNE03PX83 and PRIN-2007M3E2T2_003.

REFERENCES

[1] Nelson D. L. and Cox M. M, *Lehninger Principles of Biochemistry*, 5th ed. (W.H. Freeman, San Francisco) 2008.
[2] Brown T. A., *Genomes*, 3rd ed. (Garland Science, New York) 2006.
[3] Branden C. and Tooze J., *Introduction to Protein Structure* (Garland Science, New York) 1999.
[4] Chothia C. and Lesk A. M., *EMBO J.*, **5** (1986) 823.
[5] Martì-Renom M. A., Stuart A. C., Fiser A., Sanchez R., Melo F. and Sali A., *Annu. Rev. Biophys. Biomol. Struct.*, **29** (2000) 291.
[6] Xiang Z., *Curr. Protein Pept. Sci.*, **7** (2006) 217.
[7] Zhang Y., *Curr. Op. Struct. Biol.*, **18** (2008) 342.

[8] Andreeva A., Howorth D., Chandonia J. M., Brenner S. E., Hubbard T. J., Chothia C. and Murzin A. G., *Nucl. Acids. Res.*, **36** (2008) D419.
[9] Cuff R. L., Sillitoe I., Lewis T., Redfern O. C., Garratt R., Thornton J. and Orengo C. A., *Nucl. Acids. Res.*, **37** (2009) D310.
[10] Baker D. and Das R., *Annu. Rev. Biochem.*, **77** (2008) 363.
[11] Wu S., Skolnick J. and Zhang Y., *BMC Biol.*, **5** (2007) 17.
[12] Lazaridis T. and Karplus M., *Curr. Op. Struct. Biol.*, **10** (2000) 139.
[13] Nicholls A., Sharp K. A. and Honig B., *Proteins*, **11** (1991) 281.
[14] Fogolari F., Brigo A. and Molinari H., *J. Mol. Recogn.*, **15** (2002) 377.
[15] Skolnick J., *Curr. Op. Struct. Biol.*, **16** (2006) 166.