Short Specialist Review

Molecular simulations in structure prediction

Franca Fraternali
National Institute for Medical Research, London, UK

Jens Kleinjung
Vrije Universiteit, Amsterdam, The Netherlands

1. Introduction

Molecular simulations allow us to study properties of many-particle systems and, in particular, to extract those properties that are not easily accessible to experimental techniques (Frenkel and Smit, 1996). If we were to use the physicist approach, we have to admit that the basic laws of nature have the unpleasant feature to be expressed in equations that cannot be solved exactly (analytically), except for few cases. Even the description of the motion of a three-body system by means of the simple laws of Newtonian mechanics is analytically intractable. Computer simulations, on the other hand, use a number of algorithms that allow us to calculate properties of many-particles systems (more than 100 000 atoms) at the required degree of accuracy. All molecular simulations to which we will refer are based on the assumption that classical mechanics can be used to describe the motions of atoms and molecules. Physics-based force fields at different levels of accuracy are generally used to describe the topology and geometry of molecules (van Gunsteren and Berendsen, 1990).

On the other hand, Bioinformatics tools allow us to analyze and rationalize the information content of biological systems and, from the acquired knowledge, to extrapolate and predict molecular properties that are not yet experimentally available. Therefore, in terms of logical strategies, molecular simulations use the inductive argument to infer knowledge, while Bioinformatics uses a deductive strategy; they can be seen as working synergically for the aim of filling the gaps of current experimental knowledge. Computational biology embraces both approaches, and experience reveals that both are indeed necessary to shed light on the complexity of biomolecules.

Since the genomic projects have started their proliferation of sequences, the so-called sequence gap (the difference between the number of known sequences and the known protein 3D structures) has become one of the most prominent challenges in biology. In particular, structural genomics initiatives address the

investigation of the native fold of sequenced proteins as a fundamental step to a complete understanding of their biological function. Experimental techniques such as NMR and X-ray crystallography are very effective in providing atomic resolution structures of a large number of proteins, but large-scale initiatives of high-throughput structure determination are still at a different pace compared to genomic sequencing projects.

Until now, all structure prediction methods still suffer in accuracy and are not yet reliable enough to compete even with available experimental low-resolution structure determination methods. If one could predict the topology at a low-resolution level (3–6 Å) and use molecular simulations to refine the atomic details of the structure, the sequence gap could be filled more rapidly and be useful in the assignment of function. The bottleneck of such a procedure is that long simulations times are required with conventional MD simulations in order to refine accurately misfolded structures (Fan and Mark, 2004).

In this article, we review some of the recent efforts in the fields of structure prediction and molecular simulations, with the common aim to efficiently contribute to large-scale structural genomics initiatives.

2. Structure prediction

Structure prediction projects for the modeling of large proteins (Amodeo *et al*., 2001; Fraternali and Pastore, 1999) and entire genomes have been performed (Fischer and Eisenberg, 1997; Sanchez and Sali, 1998), and genome annotation has been shown to benefit from the use of structural homology information (Mayor *et al*., 2004). The qualitative progress in structure prediction methods is regularly assessed in the CASP experiments (Venclovas *et al*., 2001).

The field is traditionally separated into three disciplines, depending upon the level of identity between the query sequence and sequences of template structures: (1) comparative modeling (>30% sequence identity), (2) fold recognition or threading (<30% sequence identity), and (3) ab initio folding (no template). Independent of this classification, benchmarks show that the quality of the predicted structure, given as root mean square deviation from the native structure, decreases exponentially with decreasing sequence identity, that is, from about 2 Å at 95% to >5 Å below 25% identity (Contreras-Moreira and Bates, 2002). Even at relatively high sequence identity of about 50–60%, predicted structures achieve on average only medium resolution (∼3 Å) (Vitkup *et al*., 2001).

The deviation of predicted structures from their native counterpart has chiefly two reasons: alignment errors and the template structure approximation. Matching a sequence onto a template structure is a computationally hard problem (Lathrop, 1994; Kolodny and Linial, 2004), which is generally solved using heuristic approximations that lead frequently to suboptimal solutions, in particular for difficult alignments of distant homologs. The template structure approximation is due to the simple fact that two homologous structures are similar but usually not identical. Taking one structure as the template for the other naturally causes deviations. Moreover, structure prediction programs use simplified potentials and heuristic methods to generate target structures at reasonable computational expense.

Therefore, subsequent refinement is imperative to evolve the modeled structure toward the native conformation.

3. Conformational sampling

The "sampling problem" is a well-known obstacle in molecular simulations of biomolecules, and it is particularly relevant for endeavors to detect the native fold. It is primarily caused by the enormous dimension of the conformational space, which is the collection of all accessible internal coordinate combinations. The backbone of an average size protein has about 10^{100} degrees of freedom, far beyond our systematic searching capabilities. Fortunately, biological evolution has selected for biological macromolecules that adopt a distinct fold, so that in many cases conformational sampling can be restricted to a subspace around a limited number of template folds.

However, the energy profile of a biological macromolecule with respect to internal motion is rugged, with barriers above the background energy kT between neighboring conformers. Thermal fluctuations are rarely high enough to drive the system over these barriers, and energy minimization calculations lead to local minima instead to the global minimum. Therefore, structure prediction methods use Monte Carlo protocols to assemble a large variety of low-resolution conformers, from which the best ones are chosen for further calculations (Simons *et al*., 1997). Approaches to overcome the sampling problem in refinement by molecular simulations are discussed in the last paragraph.

4. Identification of the native state

The central paradigm of structure prediction and refinement by molecular simulations is the identification of the native structure as the conformer at the global free energy minimum, or alternatively at the probability maximum. The underlying thermodynamic principle is the notion that a system (here a macromolecule in solution) in weak contact with a thermal bath evolves spontaneously to the state of minimal free energy (we use here the more common notation of the free enthalpy G). Thus, successful structure prediction and refinement tools need a target function that discriminates effectively between native and nonnative folds (Park *et al*., 1997; Mirny and Shakhnovich, 1998). The free enthalpy of a protein can be decomposed into the intramolecular van der Waals and electrostatic contributions and an additional solvent term:

$$\Delta G = \Delta G_{\mathrm{vdW}} + \Delta G_{\mathrm{ele}} + \Delta G_{\mathrm{sol}} \tag{1}$$

However, in computer simulations the entropic component of the free enthalpy is often neglected, leading to a description in terms of intramolecular enthalpy:

$$\Delta G \simeq \Delta H_{\mathrm{vdW}} + \Delta H_{\mathrm{ele}} + \Delta G_{\mathrm{sol}} \tag{2}$$

Intramolecular enthalpies are calculated as the sum of all pairwise (and sometimes triple-wise) atomic interaction energies. These energies are based on carefully

parametrized interaction functions, most importantly the Lennard–Jones potential for van der Waals interactions and the Coulomb potential for electrostatic interactions. The ensemble of parametrized interaction functions defines a "force field". We distinguish between two types of force fields: physics-based force fields and knowledge-based force fields.

Physics-based force fields are derived from quantum-molecular or molecular dynamics simulations of small molecules. Prominent examples for physics-based force fields are AMBER (Wang *et al*., 2000), CHARMM (Brooks *et al*., 1983), and GROMOS (van Gunsteren *et al*., 1996). Specific energy parameters have been designed for RNA molecules, owing to their particular intramolecular interactions and secondary structure patterns (Freier *et al*., 1986; Jaeger *et al*., 1989).

Knowledge-based or statistical force fields are derived from probabilities of states of known structures (Bowie *et al*., 1991; Jones *et al*., 1992; Madej *et al*., 1995; Huber and Torda, 1999; Lu *et al*., 2003). The transformation function between the (observed) probability of a state (for example the distance between two residues i and j) and the associated interaction energy is the Boltzmann term:

$$p(X_{ij}(r)) = \mathrm{e}^{-\frac{E(X_{ij}(r))}{kT}} \tag{3}$$

where $X_{ij}(r)$ is the state X as a function of the distance r between i and j, $E(X_{ij}(r))$ is the interaction energy (potential) associated with $X_{ij}(r)$, k is the Boltzmann constant, and T is the temperature. Transformation of equation (3) yields the knowledge-based potential

$$E(X_{ij}(r)) = -kT \log p(X_{ij}(r)) \tag{4}$$

often referred to as "potential of mean force", because the interaction energy represents a collection of states.

Physics-based force fields model the physical reality starting from first principles, which renders them suitable for simulating accurately a wide range of molecules. However, early force fields were based on relatively small training sets, so that knowledge-based force fields were superior in describing specific classes of molecules at considerably reduced computational cost. On the other hand, knowledge-based potentials are by definition biased toward the training set, yielding questionable results for molecules with features far outside the range of those in the training set. Over the last decades, continuous refinement has improved the performance of physics-based force fields to a level equaling knowledge-based potentials (Lazaridis and Karplus, 1999a).

As an alternative to the conversion of state probabilities to energies as exemplified by equations (3) and (4), evaluation of predicted structures can be performed entirely in probability space, in which case the target function is the total probability maximum instead of the free energy minimum. It is common practice in Bioinformatics not to use the raw observed probabilities of states, but to convert those to a scoring function by normalization and transformation to a logarithmic scale:

$$S(X_{ij}(r)) = \log \frac{p(X_{ij}(r))}{p(X_{ij}^{\mathrm{rand}}(r))} \tag{5}$$

Here, $p(X_{ij}^{\text{rand}}(r))$ is the probability of observing state $X_{ij}(r)$ in a random set of conformers, which normalizes the observation probability with the expectation probability. The score S is often referred to as "log odds" score or "relative entropy", and it has the property to be additive.

If the logarithm is taken as $\log_2$, the score S in equation (5) is equivalent with the amount of information in units of "bit" that is gained when state $X_{ij}(r)$ is observed.

5. Refinement by molecular simulations

Structure refinement has originally been developed for experimental structure determination, that is, in NMR or X-ray resolution of proteins and DNA/RNA molecules. Experimental data are converted into distance restraints and combined with a classical molecular force field. Shortly after the invention of restrained molecular dynamics for (experimental) de novo structure calculation of proteins (Brünger *et al*., 1986), the "distance-geometry" method was developed, in which the classical force field was replaced with a simplified interaction function to increase computational speed, and simulated annealing was performed to enhance conformational sampling (Nilges *et al*., 1988). These developments led to the structure refinement program XPLOR (Brünger, 1992). Another approach was implemented in the program DIANA, where conformational sampling starts from random conformations and converges to the final structure by target function minimization in angular space (Güntert and Wüthrich, 1991).

A different situation emerges with regard to the refinement of predicted structures. The accuracy of generated conformers relies entirely on the precision of the force field and the efficiency in sampling around the native state. The detailed physics-based force fields and the associated extensive interaction calculations in molecular dynamics provide a much finer resolution than structure prediction methods and thus should be superior in defining the native state. Reports about the success of molecular dynamics in structure refinement imply that simplified representations or limited sampling fail to improve the predicted structure (Schonbrun *et al*., 2002), but carefully defined simulations can yield significantly better conformers (Lee *et al*., 2001; Flohil *et al*., 2002; Fan and Mark, 2004).

However, long simulation times are prohibitive for large-scale structure prediction applications. An effective means to shorten simulation time without sacrificing the solvent contribution to the free energy is the usage of implicit solvation (Fraternali and van Gunsteren, 1996; Lazaridis and Karplus, 1999b Ferrara *et al*., 2002). Numerous protocols have been developed to improve the sampling efficiency of molecular simulations. The extension of molecular dynamics to four dimension allows the molecule to bypass 3D barriers (van Schaik *et al*., 1993), and "local elevation" disfavors already visited conformations in order to enhance exploration of conformational space (Huber *et al*., 1994). Leap dynamics is a combination of Monte Carlo and molecular dynamics methods, in which limited conformational changes are induced and the new conformer performs a local search (Kleinjung *et al*., 2000; Kleinjung *et al*., 2003). An example for a purely stochastic protocol is the optimal-bias Monte Carlo method, in which the global energy minimum is searched within the internal coordinate space (Abagyan and Totrov, 1999).

Systematic studies about the predictive power of these refinement methods in combination with large-scale structure prediction projects should be undertaken to provide high-resolution models for the majority of sequenced biomolecules.

References

Abagyan R and Totrov M (1999) *Ab initio* folding of peptides by the optimal-bias Monte Carlo minimization procedure. *Journal of Comparative Physica*, **151**, 402–421.

Amodeo P, Fraternali F, Lesk AM and Pastore A (2001) Modularity and homology: modelling of the type I modules and their interfaces. *Journal of Molecular Biology*, **311**, 283–296.

Bowie JU, Lüthy R and Eisenberg D (1991) A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, **12**, 164–170.

Brooks BR, Brucoleri RE, Olafson BD, States DJ, Swaminathan S and Karplus M (1983) CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. *Journal of Comparative Chemistry*, **4**, 187–217.

Brünger AT (1992) *X-PLOR, Version 3.1. A System for X-ray Crystallography and NMR*, Yale University Press: New Haven.

Brünger AT, Clore GM, Gronenborn AM and Karplus M (1986) Three-dimensional structures of proteins determined by molecular dynamics with interproton distance restraints: application to crambin. *Proceedings of the National Academy of Sciences of the United States*, **83**, 3801–3805.

Contreras-Moreira B, Fitzjohn PW and Bates PA (2002) Comparative modelling: an essential methodology for protein structure prediction in the post-genomic era. *Applied Bioinformatics*, **1**, 177–190.

Fan H and Mark AE (2004) Refinement of homology-based protein structures by molecular dynamics simulation techniques. *Protein Science*, **13**, 211–220.

Ferrara P, Apostolakis J and Caflisch A (2002) Evaluation of a fast implicit solvent model for molecular dynamics simulations. *Proteins*, **46**, 24–33.

Fischer D and Eisenberg D (1997) Assigning folds to the proteins encoded by the genome of *Mycoplasma genitalium*. *Proceedings of the National Academy of Sciences of the United States*, **94**, 11929–11934.

Flohil JA, Vriend G and Berendsen HJ (2002) Completion and refinement of 3-D homology models with restricted molecular dynamics: application to targets 47, 58, and 111 in the CASP modeling competition and posterior analysis. *Proteins*, **48**, 593–604.

Fraternali F and Pastore A (1999) Modularity and homology: modelling of the type II module family from titin. *Journal of Molecular Biology*, **290**, 581–593.

Fraternali F and van Gunsteren WF (1996) An efficient mean solvation force model for use in molecular dynamics simulations of proteins in aqueous solution. *Journal of Molecular Biology*, **256**, 939–948.

Freier SM, Kierzek R, Jaeger JA, Sugimoto N, Caruthers MM, Neilson T and Turner DH (1986) Improved free-energy parameters for predictions of RNA duplex stability. *Proceedings of the National Academy of Sciences of the United States*, **83**, 9373–9377.

Frenkel D and Smit B (1996) *Understanding Molecular Simulations*, Academic Press: London.

Güntert P and Wüthrich K (1991) Improved efficiency of protein structure calculations from NMR data using the program DIANA with redundant dihedral angle constraints. *Journal of Biomolecular NMR*, **1**, 446–456.

Huber T and Torda AE (1999) Protein sequence threading, the alignment problem and a two step strategy. *Journal of Computational Chemistry*, **20**, 1455–1467.

Huber T, Torda AE and van Gunsteren WF (1994) Local elevation: a method for improving the searching properties of molecular dynamics simulation. *Journal of Computer-aided Molecular Design*, **8**, 695–708.

Jaeger JA, Turner DH and Zuker M (1989) Improved predictions of secondary structures for RNA. *Proceedings of the National Academy of Sciences of the United States*, **86**, 7706–7710.

Jones DT, Taylor WR and Thornton JM (1992) A new approach to protein fold recognition. *Nature*, **358**, 86–89.

Kleinjung J, Bayley PM and Fraternali F (2000) Leap-dynamics: efficient sampling of conformational space of proteins and peptides in solution. *FEBS Letters*, **470**, 257–262.

Kleinjung J, Fraternali F, Martin SR and Bayley PM (2003) Thermal unfolding simulations of apo-calmodulin using Leap-dynamics. *Proteins*, **50**, 648–656.

Kolodny R and Linial R (2004) Approximate protein structural alignment in polynomial time. *Proceedings of the National Academy of Sciences of the United States*, **101**, 12201–12206.

Lathrop RH (1994) The protein threading problem with sequence amino acid interaction preferences is NP-complete. *Protein. Engineering*, **7**, 1059–1068.

Lazaridis T and Karplus M (1999a) Discrimination of the native from mis-folded protein models with an energy function including implicit solvation. *Journal of Molecular Biology*, **288**, 477–487.

Lazaridis T and Karplus M (1999b) Effective energy functions for proteins in solutions. *Proteins*, **35**, 133–152.

Lee MR, Tsai J, Baker D and Kollman PA (2001) Molecular dynamics in the endgame of protein structure prediction. *Journal of Molecular Biology*, **313**, 417–430.

Lu H, Lu L and Skolnick J (2003) Development of unified statistical potentials describing protein-protein interactions. *Biophysical Journal*, **84**, 1895–1901.

Madej T, Gibrat JF and Bryant SH (1995) Threading a database of protein cores. *Proteins*, **23**, 356–369.

Mayor LR, Fleming KP, Muller A, Balding DJ and Sternberg MJ (2004) Clustering of protein domains in the human genome. *Journal of Molecular Biology*, **340**, 991–1004.

Mirny LA and Shakhnovich EI (1998) Protein structure prediction by threading. Why it works and why it does not. *Journal of Molecular Biology*, **283**, 507–526.

Nilges M, Clore GM and Gronenborn AM (1988) Determination of three-dimensional structures of proteins from interproton distance data by dynamical simulated annealing from a random array of atoms. Circumventing problems associated with folding. *FEBS Letters*, **239**, 129–136.

Park BH, Huang ES and Levitt M (1997) Factors affecting the ability of energy functions to discriminate correct from incorrect folds. *Journal of Molecular Biology*, **266**, 831–846.

Sanchez R and Sali A (1998) Large-scale protein structure modeling of the *Saccharomyces cerevisiae* genome. *Proceedings of the National Academy of Sciences of the United States*, **95**, 13597–13602.

Schonbrun J and Wedemeyer WJ (2002) Protein structure prediction in 2002. *Current Opinion in Structural Biology*, **12**, 348–354.

Simons KT, Kooperberg C, Huang E and Baker D (1997) Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *Journal of Molecular Biology*, **268**, 209–225.

van Gunsteren WF and Berendsen HJC (1990) Computer simulations of molecular dynamics: methodology, applications and perspectives in chemistry. *Angewandte Chemie (International ed. in English)*, **29**, 992–1023.

van Gunsteren WF, Billeter SR, Eising AA, Hünenberger PH, Krüger P, Mark AE, Scott W and Tironi I (1996) *Biomolecular Simulations: The GROMOS96 Manual and User Guide*. BIOMOS b.v. Laboratory of Physical Chemistry: ETH Zentrum, CH-8092 Zürich vdf Hochschulverlag AG, Zürich, ISBN 3 7281 2422 2.

van Schaik RC, Berendsen HJ, Torda AE and van Gunsteren WF (1993) A structure refinement method based on molecular dynamics in four spatial dimensions. *Journal of Molecular Biology*, **234**, 751–762.

Venclovas C, Zemla A, Fidelis K and Moult J (2001) Comparison of performance in successive CASP experiments. *Proteins*, **45**(Suppl 5), 163–170.

Vitkup D, Melamud E, Moult J and Sander C (2001) Completeness in structural genomics. *Nature Structural Biology*, **8**, 559–566.

Wang J, Cieplak P and Kollman PA (2000) How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *Journal of Comparative Chemistry*, **21**, 1049–1074.