John von Neumann Institute for Computing

Modeling Protein Structure, Dynamics and Thermodynamics with Reduced Representation of Conformational Space

Andrzej Kolinski, Dominik Gront, Sebastian Kmiecik,
Mateusz Kurcinski, Dorota Latek

published in

NIC Workshop 2006,
From Computational Biophysics to Systems Biology,
Jan Meinke, Olav Zimmermann,
Sandipan Mohanty, Ulrich H.E. Hansmann (Editors)
John von Neumann Institute for Computing, Jülich,
NIC Series, Vol. **34**, ISBN-10: 3-9810843-0-6,
ISBN-13: 978-3-9810843-0-6, pp. 21-28 , 2006.

http://www.fz-juelich.de/nic-series/volume34

Modeling Protein Structure, Dynamics and Thermodynamics with Reduced Representation of Conformational Space

Andrzej Kolinski, Dominik Gront, Sebastian Kmiecik,
Mateusz Kurcinski, and Dorota Latek

Laboratory of Theory of Biopolymers, Faculty of Chemistry
Warsaw University, Pasteura 1, 02-093 Warsaw, Poland
E-mail: kolinski@chem.uw.edu.pl

In this contribution we describe a successful approach to protein modeling which is based on reduced representation of protein conformational space, all-atom-refinement, evaluation and selection of the best molecular models. During the sixth CASP (Critical Assessment of protein Structure Prediction) community-wide experiment our methodology (referred further as CABS) proven to be one of the best performing methods for protein structure prediction, applied both for comparative modeling and to *de novo* folding. The newest applications of the CABS modeling technology include: study of protein folding thermodynamic, dynamics in the denatured state and folding pathways, structure prediction based on sparse and inaccurate experimental data and prediction of protein-protein interactions or flexible ligand docking. The CABS reduced model could be easily integrated with the all-atom approaches providing solid starting point for reliable multiscale simulations of large biomolecular systems.

1 Introduction

Due to the systematic sequencing of numerous genomes[1] the number of recognizable gene products, and therefore protein sequences, grows exponentially in the recent years. The number of known protein sequences exceeds 30 millions. Understanding of protein biological function, from enzymatic activity, through transport and signaling, to mechanisms and thermodynamics of complex macromolecular assemblies, requires knowledge of proteins' three dimensional structures. For many purposes it is also important to know protein folding mechanism and the dynamics near the native state, as well as the dynamics at denatured state. Understanding protein dynamics may be even more challenging than theoretical prediction of protein structure. Nevertheless, at the moment protein structure prediction seems be the most urgent and the most advanced general goal of theoretical structural biology. It has an increasing impact on rational drug design, development of new biotechnologies, and on many other areas of biomedical sciences and technology.

Theoretical approaches are important regardless the great progress in experimental methods of protein structure determination. In spite of this progress the X-ray[2,3] or NMR based methods for structure determination remain very costly and time-consuming[4]. As a result, the number of good-quality experimental structures, although impressive, is still very small in comparison with the number of known sequences, and is now range of 30 thousands. It is not quite clear if we know already the majority of possible protein folds, i.e. loosely defined distinct three-dimensional structures. This is an important issue, since the most advanced and the most accurate methods of theoretical prediction of protein structures employ already determined structures as templates in the framework of comparative modeling techniques. Reasonable comparative modeling can be now done for some 50-60% of newly identified proteins, depending on a genome[5]. For the remaining proteins it is difficult, or even impossible, to find structural templates and different approaches need to be applied. In such cases various *de novo* modeling techniques could be helpful. A number of distinct approaches to the *de novo* structure prediction have been recently proposed and evaluated. In general, the most advanced techniques are quite successful when applied to relatively simple and small proteins, although dependable structure prediction of larger proteins remains elusive.

In principle, the simplest and the most straightforward approach to the *de novo* folding would be to use a detailed all-atom-representation of polypeptide chain immersed in a proper number of water molecules with a proper number of ions around. Such systems could be sampled using Molecular Dynamics or Monte Carlo techniques. This is unfortunately still limited to very small proteins or peptides. The conformational space of proteins is enormous, and in reality a typical folding time of globular proteins (in vivo as well as in vitrio) is range of milliseconds to minutes. This time frame is few orders of magnitude too large for the contemporary computers. Thus, the geometrical representation of proteins, their interactions and sampling techniques need to be simplified in order to make the task feasible. Up to the date the two distinct, reduced-space, approaches to the *de novo* folding simulations of proteins seem to be the most successful. The first one, adopted in the Rosetta model of Baker and collaborators[6,7] employs short structural fragments excised from the known protein structures. These fragments are 3-9 residues long and their representation in the algorithm is limited to the main chain and single united atoms for the side groups. During the sampling process the model chains undergoes a long series of deletion/insertions of such building blocks. The acceptance criteria for these modifications include short range geometrical restraints, a simplified model of side chain interactions and a model of main chain hydrogen bonds. The final models are subject to a clustering procedure, rebuilding of the atomic details and final selection of the best models. A different class of models assumes even a more simplified representation of the main chain backbone, where only the alpha carbons are treated in an explicit fashion, although the sampling process allows for a broader range of conformations, controlled by properly designed force fields. Good example of such model of protein folding is UNRES, designed by Scheraga and coworkers[8,9]. The UNRES force field is carefully designed and optimized, and is derived from basic physical principles. The conformational space of UNRES is continuous. In order to further speed-up the sampling process discretized space models have been developed. Below, we briefly describe the CABS model[10], which besides the grid-type representation employs knowledge-based force field derived from statistical analysis of structural regularities observed in known proteins. Rosetta and CABS based methods proven to be among the most successful[11,12] during the last round of the CASP (Critical Assessment of Protein Structure Prediction) community-vide experiment. The summary of the CASP results can be found in our homepage http://biocomp.chem.uw.edu.pl or at the CASP homepage.

The reduced representation of the CABS model[10] employs up to four interaction centers per residue (alpha carbon, beta carbon, the center of mass of the side group and the center of peptide bond). Carefully designed, tuned and tested force field[13] of CABS consists of several potentials of mean-force derived from statistical analysis of structural regularities seen in known protein structures. The sampling of conformational space of the model proteins employs various variants of the Monte Carlo method, including very efficient muticopy simulated tempering algorithms[14–16]. A number of supplementary bioinformatics tools have been developed to handle data processing and analysis of the large scale simulations of protein systems using the CABS modeling system. CABS methodology proven to be one of the best performing methods for protein structure prediction, from comparative modeling to de novo folding. It has been clearly demonstrated during the sixth CASP (Critical Assessment of protein Structure Prediction) community-wide experiment. The groups employing the CABS-based methodology scored among the best. The design of the CABS model has been recently described in great detail[10]. In the next section, the most typical recent applications in protein structure prediction, refinement of NMR data and study of protein dynamics and interactions are presented.

2 Structure Prediction Based on Sparse NMR Data

NMR based protein structure determination is a time consuming and costly process. However, some sparse NMR data as Chemical Shifts (CS), Residual Dipolar Coupling (RDC) and some sparse NOE's are relatively easy to obtain at early stages of structure determination process. Such data are usually insufficient for molecular model building using the standard computational procedures. The restraints derived from sparse experimental data (NMR in this case) are easy to implement in the CABS modeling tool. The first applications focused on the CS and RDC data[17]. The CS data provide loose restrictions on a fraction of Φ, Ψ angles of the protein backbone. Due to the reduced Cα representation of the main chain in the CABS algorithm it was necessary to develop a translating procedure, where the Φ, Ψ angles are transformed into θ, γ angles of the Cα -trace. θ is the planar angle of the trace, while γ is the dihedral angle defined by three consecutive Cα pseudobonds (Figure 1a). The ranges of experimental inaccuracies need also to be properly translated. The θ, γ restraints are easy to implement in the CABS simulations controlled by the simulated tempering MC scheme. It has been shown that for not too-complex structures CS experimental data are often sufficient for a moderate resolution structure prediction. It has been also demonstrated that a proper combination of the CS data with the artificial intelligence based secondary structure predictions significantly increases the accuracy of the generated molecular models and increase the range of applicability of the method - larger proteins could be modeled. Using simulated (extracted from known PDB structures) CS data it was demonstrated that increase of the fraction of assigned phi-psi angles from around 60% to some 80% improves qualitatively the accuracy of the obtained molecular models. It points onto a possible way of improving and speeding-up the NMR based structure prediction procedures. Increasing the accuracy of the CS measurements or/and supplementing the CS data with the local NOE's (or other measurements of the short range angular correlations) would be extremely beneficial for a fast computational model building with the CABS algorithm.

While the CS data are of the local nature the RDC data provide global, although very

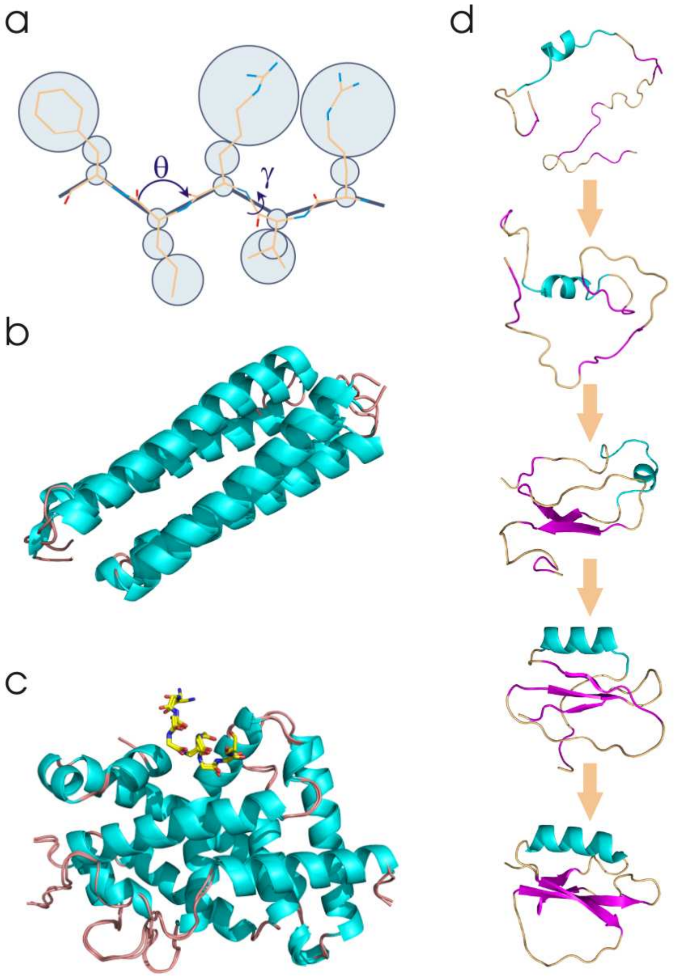

Figure 1. **(a)** Translation from Φ, Ψ angles (obtained from NMR chemical shifts) to the CABS representation (θ, γ angles). **(b)** De-novo assembly of ROP homodimer: theoretical model superimposed on NMR structure (pdb code: 1RPR) with cRMSD equall to 2.97 Å . **(c)** Predicted model of vitamin D receptor bound with short peptide co-activator fragment, superimposed on crystallographic structure (pdb code: 1RJK). Accuracy of obtained model is 0.61 Å . **(d)** Example snapshots from folding simulation, illustrating the folding pathway of chymotrypsin inhibitor 2. The portions of the protein that form β-strands in the native state are highlited in magenta, the α helix is colored in cyan.

inaccurate, restraints. The RDC data need to be also properly translated onto the CABS geometry. Preliminary computational studies show that the RDC based restraints are very useful in guiding the simulations for more complex structures - they provide a bias towards the right fold topology. The local details are tuned by the CABS force field, eventually supplemented by the CS-based restraints. Finally, even very small number of the long range restraints from NOE's (range of N/12, where N is the number residues in the chain) allows for the proper fold assembly even for quite complex structures.

In summary, CABS based modeling could be a powerful tool for structure determination from sparse (and of low accuracy) experimental data. It could be also used as a part of the full process of structure determination using the NMR techniques. Namely, the initial data (CS, RDC, etc.) could be used for building an initial model, or a number of alternative models. These models could be subsequently used as a guide in the assignments of the NOE signals.

3 Flexible Docking with CABS

At present, the CABS force field is applicable only to proteins and peptides. This allows studies of protein complexes and docking of peptide ligands (co-factors, etc) to proteins. The docking procedure can be performed in various regimes; from completely unrestrained simulations of two or more polypeptide (peptide) chains to simulations with partially restrained internal coordinates of the interacting macromolecules. It has been shown that unrestrained simulations of protein dimers (GCN4 leucine zippers, ROP-dimer, etc) lead to a rapid assembly of the correct dimeric structures (Figure 1b). Weak restraints superimposed onto internal coordinates of the interacting molecules lead to a higher resolution of the obtained assemblies.
A number of test docking experiments have been performed assuming a limited flexibility of a receptor and the full conformational mobility of the interacting peptides[21]. No knowledge of the binding sites was assumed. At the beginning of simulations peptides were placed at random position at a large distance from the receptors. In all studied cases the peptides attached correctly to various nuclear receptors (Figure 1c). In about half of cases the accuracy was range of 1Å RMSD, as measured for the alpha carbons of the peptide after the best superposition of the model receptor structure onto its crystallographic structure. In other cases the resolution was range of 2-2.5 Å . This implies, that always the pattern of the side chain interactions was correctly predicted. Thus, it has been demonstrated that the new modeling tool is already capable of producing correct high-resolution structures of protein - peptide complexes and of proper assembly of protein mutimers (in the last case the process is computationally expensive for larger structures). This should be very important for better understanding of protein interactions, signaling pathways and for computer aided design of new drugs.

4 Modeling of Protein Dynamics and Folding Pathways

The force field of CABS has been designed basing on statistical analysis of structural regularities observed in known protein structures[10, 13]. Thus, it may appear that the model is applicable only to structure prediction, but not to study of protein dynamics in the denatured state, or modeling the folding mechanisms. This is however not the case. Apparently, the nature of the mean field interactions in the denatured state is very similar to these in the compact native structures. For a number of small proteins a detailed data describing the nature of the denatured state and the folding mechanism are available. Using the CABS model isothermal simulations just above the folding temperature were performed and the degree of exposure to the solvent for all residues computed from the resulting trajectories[19]. These correlated extremely well with the measurements of the protection factors for the burst intermediates. The unfolded model proteins exhibited the features of the molten globule state[20], believed to be a common folding intermediate for globular proteins. Interestingly, the most frequently observed long range contacts in the simulated denatured conditions overlapped with the folding nuclei observed in experiments. Thus, it seems to be safe to conclude, that the CABS model could be used not only in studies of protein folded structures, but also in semiquantitative modeling of protein dynamics, folding pathways (Figure 1d) and mechanism of macromolecular assembly.

5 Conclusions

CABS is a high resolution, lattice based model of protein structure and protein stochastic dynamics. The force field of the model consists of several statistical potentials of mean force, derived from the regularities seen in the known protein structures. The solvent in this force field is treated in an implicit fashion. Due to computational speed CABS can be used in a large scale protein modeling (large scale in respect to the size of the modeled system, as well as in respect to the number of proteins that could be structurally annotated in a reasonable time). Recent applications include protein structure prediction (*de novo*, as well as supported by sparse experimental data), modeling of protein interactions in macromolecular assemblies and study of protein dynamics and folding mechanisms. Finally, it should be noted that CABS is compatible with the classical all-atom modeling tools. The spatial accuracy of the CABS models is sufficient for a meaningful reconstruction of the atomic details. Thus multiscale simulations at various levels of resolution become feasible. Future extensions of the model will include interactions with non-peptide ligands, membranes and nucleic acids.

Acknowledgments

This work was partially supported by the grant # PBZ-KBN-088/P04/2003

References

1. R. D. Fleischmann, M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J. F. Tomb, B. A. Dougherty, and J. M. Merrick. Whole-genome random sequencing and assembly of haemophilus influenzae rd. *Science*, 269(5223):496–512, July 1995.
2. W. A. Hendrickson. Synchrotron crystallography. *Trends Biochem Sci*, 25(12):637–643, December 2000.
3. A. Schmidt and V. S. Lamzin. Veni, vidi, vici - atomic resolution unravelling the mysteries of protein function. *Curr Opin Struct Biol*, 12(6):698–703, December 2002.
4. D. Vitkup, E. Melamud, J. Moult, and C. Sander. Completeness in structural genomics. *Nat Struct Biol*, 8(6):559–566, June 2001.
5. M. A. Marti-Renom, A. C. Stuart, A. Fiser, R. Sanchez, F. Melo, and A. Sali. Comparative protein structure modeling of genes and genomes. *Annu Rev Biophys Biomol Struct*, 29:291–325, 2000.

6. C. Bystroff and D. Baker. Prediction of local structure in proteins using a library of sequence-structure motifs. *J Mol Biol*, 281(3):565–577, August 1998.
7. C. A. Rohl, C. E. Strauss, K. M. Misura, and D. Baker. Protein structure prediction using rosetta. *Methods Enzymol*, 383:66–93, 2004.
8. A. Liwo, S. Oldziej, M. R. Pincus, R. J. Wawak, S. Rackovsky, and H. A. Scheraga. A united-residue force field for off-lattice protein-structure simulations. i. functional forms and parameters of long-range side-chain interaction potentials from protein crystal data. *Journal of Computational Chemistry*, 18(7):849–873, 1997.
9. A. Liwo, C. Czaplewski, J. Pillardy, and H. A. Scheraga. Cumulant-based expressions for the multibody terms for the correlation between local and electrostatic interactions in the united-residue force field. *The Journal of Chemical Physics*, 115(5):2323–2347, 2001.
10. A. Kolinski. Protein modeling and structure prediction with a reduced representation. *Acta Biochim Pol*, 51(2):349–371, 2004.
11. J. Moult. A decade of casp: progress, bottlenecks and prognosis in protein structure prediction. *Curr Opin Struct Biol*, June 2005.
12. A. Kolinski and Janusz M. M. Bujnicki. Generalized protein structure prediction based on combination of fold-recognition with de novo folding and evaluation of models. *Proteins*, September 2005.
13. D. Gront and A. Kolinski. A new approach to prediction of short-range conformational propensities in proteins. *Bioinformatics*, 21(7):981–987, April 2005.
14. C. J. Geyer. Markov chain monte carlo maximum likelihood. In *Computing Science and Statistics: Proceedings of 23rd Symposium on the Interface Interface Foundation*, pages 156–163. Fairfax Station, 1991.
15. U. H. E. Hansmann. Parallel tempering algorithm for conformational studies of biological molecules, *Chem. Phys. Lett*, 281:140 1997.
16. D. Gront, A. Kolinski, and J. Skolnick. Comparison of three monte carlo conformational search strategies for a proteinlike homopolymer model: Folding thermodynamics and identification of low-energy structures. *The Journal of Chemical Physics*, 113(12):5065–5071, 2000.
17. D. Plewczynska and A. Kolinski. Protein folding with a reduced model and inaccurate short-range restraints. *Macromolecular Theory and Simulations*, 14(7):444–451, 2005.
18. H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The protein data bank. *Nucleic Acids Res*, 28(1):235–242, January 2000.
19. S. Kmiecik, M. Kurcinski, A. Rutkowska, D. Gront, and A. Kolinski. Denatured proteins and early folding intermediates simulated in a reduced conformational space. *Acta Biochim Pol*, December 2005.
20. D. Ekonomiuk, M. Kielbasinski, and A. Kolinski. Protein modeling with reduced representation: statistical potentials and protein folding mechanism. *Acta Biochim Pol.*, 52(4):741–748, 2005.
21. M. Kurcinski and A. Kolinski. Steps towards flexible docking: Modeling of threedimensional structures of the nuclear receptors bound with peptide ligands mimicking co-activators' sequences. *Journal of Steroid Biochemistry and Molecular Biology*, (submitted)