A probabilistic approach to protein structure prediction: PHAISTOS in CASP8

M. Borg[1*], W. Boomsma[1], J. Ferkinghoff-Borg[2], J. Frellsen[1], T. Harder[1], K.V. Mardia [3], P. Røgen[4], K. Stovgaard[1] and T. Hamelryck[1]

[1]Department of Biology, University of Copenhagen
[2]DTU Electro, Technical University of Denmark
[3]Department of Statistics, University of Leeds
[4]Department of Mathematics, Technical University of Denmark

Introduction

The prediction of the 3-dimensional structure of a protein given its chemical composition (that is, its amino acid sequence), remains an elusive problem in spite of decades of research efforts. The rate of progress in the field is even stagnating, and much of the improvements seen in the last couple of years can be attributed to larger databases of experimentally determined protein structures, and to more powerful computers (Kryshtafovych *et al.*, 2007).

We recently developed a probabilistic model of local protein structure in continuous space, TorusDBN, which is an attractive alternative to the discrete, non-probabilistic fragment assembly methods (Boomsma *et al.*, 2008). Furthermore, the model allows for calculating the probability of a sampled conformation, which is essential in order to avoid biases in Monte Carlo sampling. However, in order to predict native protein structures, non-local interactions must also be taken into account.

Here we present our framework, PHAISTOS[5], and our initial attempts of predicting protein structure from sequence. We tested our approach rigorously by participating in the 8th Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction (CASP8); a biennial double-blind experiment in protein structure prediction (Moult, 2006). We submitted structure predictions for 5 different targets. Two of these targets turned out to be intrinsically unstructured proteins, without a fixed structure. Nonetheless, we managed to predict some important substructures present in these proteins. For the remaining three globular proteins, we successfully predicted the fold for two of them. These results are very encouraging, especially considering the preliminary state of the nonlocal energy function and the fact that we only use the protein sequence as input for the predictions.

Phaistos framework

Our long-term goal is to have a completely probabilistic description of protein folding. The connection between the underlying physics and the probability of finding a protein in a conformation j, comes from the Boltzmann distribution and is proportional to $g_j \exp(-E_j)$, where g_j and E_j are the multiplicity and energy of the conformation, respectively. According to the widely accepted Anfinsen hypothesis (Anfinsen *et al.*, 1961), the native state of a protein is the most probable state at thermodynamic equilibrium.

[5]http://www.phaistos.org

We represent the protein as a polymer with fixed bond lengths and bond angles. The amino acid side chain atoms are represented by one pseudo atom (SC) at the geometrical center for each residue, as shown in Fig. 1. The only degrees of freedom for the protein are thus the dihedral angles along the backbone; a conformation is specified by the set $\mathrm{X} = \{\omega_i, \phi_i, \psi_i\}$, where the index i runs over all amino acids and ω, ϕ, ψ are the usual backbone dihedral angles. In the current version of PHAISTOS, we fix the protein's secondary structure to the one predicted by PSIPRED (Jones, 1999).

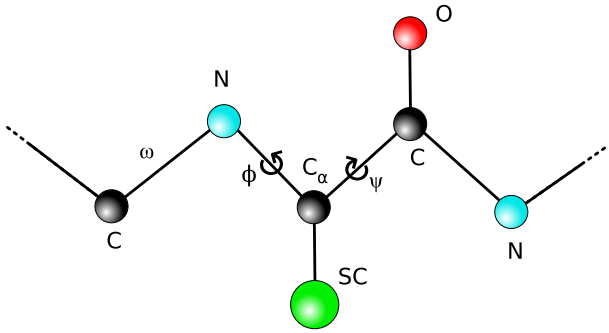

Figure 1: Protein representation. The protein backbone is modeled in atomic detail: each residue contains four backbone atoms (N, C_α, C and O). For the amino acid side chains we use a simplified representation: each side chain is represented by one pseudo-atom (SC).

We combine our local model with a non-local energy function, and obtain low energy structures by generalized Monte Carlo sampling. Then, a subset of these samples are selected based on their energies and by clustering structures. The final predictions are obtained by filling in all the missing atomic details (that is, the side chain atoms and the hydrogen atoms) and ranking the samples using an all-atom force-field.

Our non-local energy function consists of contributions from excluded volume interactions, hydrogen bonds between main chain atoms, compactness and a multi-body contact term. The excluded volumes of atoms were taken into account in the form of a simple repulsion when two atoms are closer than the sum of their van der Waals radii, of the form

$$E_{\text{clash}} \sim H(r_c - r_{ij})(1 - r_{ij}/r_c),$$

where H is the Heaviside step function, r_{ij} is the interatomic distance and r_c is the clash distance.

The hydrogen bond term was taken from the literature, and is a function of various distances and angles between the hydrogen bond partners (Fabiola *et al.*, 2002). We only include hydrogen bonds between $\beta-$strands, as $\alpha-$helix hydrogen bonds are taken into account by the local model. The hydrogen bond energy term is of the form

$$E_{hb} = \epsilon \, \left((\sigma/r_{N-O})^6 - (\sigma/r_{N-O})^4 \right) \cos^4(\theta - \theta_0) \, \mathrm{SW}(r_{N-O}),$$

where ϵ is a scaling factor, σ is a parameter determining the optimal hydrogen bond N-O distance, r_{N-O} is the distance between the N and O atoms, θ is the N-O-C angle (θ lies between 0 and 180 degrees), θ_0 is the ideal hydrogen bond angle and SW is a smoothing function.

Folded protein molecules are in general compact, and we include a term for biasing sampling towards compact structures. This term is based on a spatial Poisson process. Such a Poisson

process can be used to assign a probability to observing a number of points in a given three-dimensional volume. We use the probability $\mathrm{P}(N|R_g)$, where N is the length and R_g is the radius of gyration of the protein. The expression for $\mathrm{P}(N|R_g)$ according to the spatial Poisson process is:

$$\mathrm{P}(N \mid R_g) = \frac{1}{N!} \left(\rho R_g^c\right)^N \exp\left(-\rho R_g^c\right).$$

The values for the parameters ρ and c of the spatial Poisson process were obtained by maximum likelihood estimation (resulting in $\rho = 0.334, c = 2.274$), using a database of (N, R_g) pairs derived from protein structures in the Top500 database [6], which is a standard reference database of high quality protein structures. As far as we know, this is the first application of the spatial Poisson process for constructing a probabilistic model of the compactness of proteins.

The last non-local energy term in our model is a multibody contact potential which captures the preferences of the different amino acid types with respect to neighboring amino acids and the total number of neighbors. To determine the neighbors of an amino acid, we made use of the half sphere solvent exposure construction (Hamelryck, 2005), which separates the spherical environment of an amino acid into two separate domes, one of which contains the side chain. A simple probabilistic model (essentially based on a dice model) is then used to model the amino acid type content of the domes. Details of this model will be described in a separate publication.

We sample the conformational space from the distribution

$$\mathrm{P} \sim \mathrm{P}(\mathrm{X}|\mathrm{S},\mathrm{A})\mathrm{P}_{\mathrm{clash}}\mathrm{P}_{\mathrm{hb}}\mathrm{P}(N|R_g)\mathrm{P}_{\mathrm{multibody}},$$

where $\mathrm{P}(\mathrm{X}|\mathrm{S},\mathrm{A})$ is the probability of the structure according to TorusDBN, included by using it as a proposal distribution in a Markov Chain Monte Carlo (MCMC) scheme. The remaining factors are probabilities from our non-local terms, where energy terms have been converted to probabilities in the form of Boltzmann factors $\mathrm{P_i} \sim e^{-\lambda_i E_i}$, where λ_i is a scaling factor for energy term i. The scaling factors were determined manually using a small set of known protein structures. By using our local model as proposal distribution, only conformations with a realistic local structure are considered, allowing for an efficient search in conformational space.

The energy landscape of conformational space is very rugged. Therefore, the task of finding the global minimum among a large number of local minima separated by high barriers is a formidable challenge which requires advanced sampling techniques. The sampling algorithm that we use is a generalized multi-histogram method that estimate the density of states as a function of energy, $g(E)$, from the sampling histogram (Hesselbo and Stinchcombe, 1995; Ferkinghoff-Borg, 2002). The density of states is used to sample conformational space with sampling weights

$$\omega(E) = \left(\sum_{E' \leq E} g(E')\right)^{-1}.$$

The sampling algorithm thus biases the sampling towards low energy conformations, and at the same time avoids getting trapped in local energy minima.

Using our coarse-grained model and efficient sampling we can generate large numbers of conformations. In order to select conformations for further analysis we use a clustering algorithm

[6] Available from http://kinemage.biochem.duke.edu/databases/top500.php.

that is based on Gauss integrals (Røgen, 2005; Røgen and Fain, 2003). Briefly, each conformation is mapped to a point in $\mathbb{R}^{30}$ using 30 different topological invariants of the polygonal curve connecting the C_α atoms of the structure. The points in $\mathbb{R}^{30}$ are then clustered using standard Euclidean metrics and k-means clustering.

We select a subset of structures based on the clustering and the coarse-grained energies for further analysis. For each structure, we generate an all-atom model from the backbone coordinates (Hartmann *et al.*, 2007) and calculate the energy of the structure using a physics-based all-atom force field with implicit solvent (Dominy and Brooks, 1999). The all-atom structures are energy-minimized using steepest descent and adopted-basis Newton-Raphson minimization. The final selection of structures is based on all-atom energies, radius of gyration, hydrophobic radius of gyration, and physical properties.

Results

We selected 5 target sequences from the CASP8 experiment for prediction and used our protocol to generate 5 candidates for each target. Two of the selected targets turned out to be largely disordered proteins. One of the disordered targets, T0474, was a dimer with some ordered parts which included a helix-turn-helix motif, which was also present in our prediction (Fig. 2). The other disordered protein, T0480, contained a zinc finger with four cystein residues in close proximity, which we also correctly predicted (results not shown). For two of the remaining targets, T0469 and T0473, we predicted the native structures with C_α root mean square deviations (RMSDs) of 5.0 and 5.4 Å, respectively (Fig. 3).

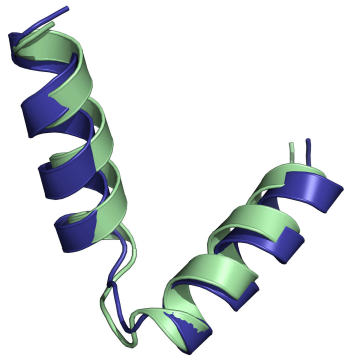

Figure 2: Experimental (light grey/green) and predicted (dark grey/blue) structures for residues 17-49 of CASP target T0474. C_α RMSD is 1.9 Å. (Colour figures will be included in the proceedings available online).

Conclusions and outlook

Our results are very encouraging, given that our energy terms for non-local interactions were just initial versions of what will eventually become a fully probabilistic framework for protein structure prediction. In particular, we did not use any templates from homologous structures in our predictions, suggesting that our approach is useful in *de novo* structure prediction where template-based methods can not be used.

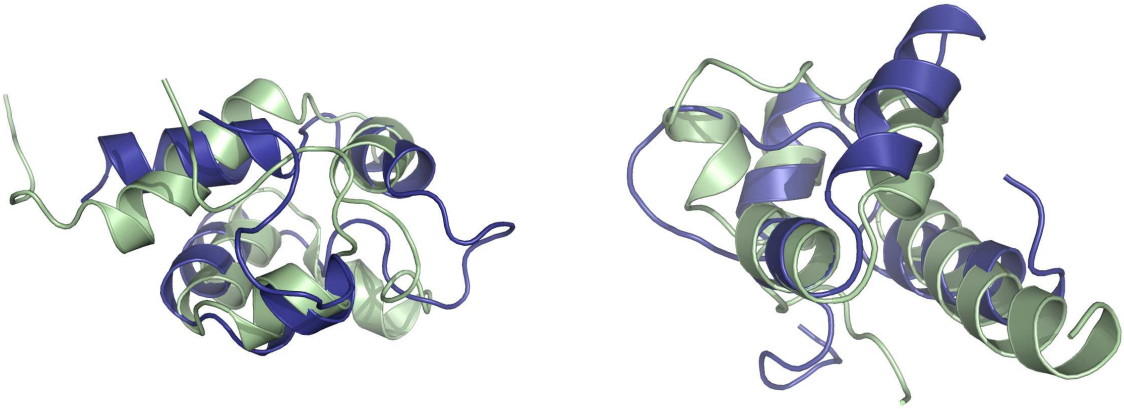

Figure 3: Experimental (light grey/green) and predicted (dark grey/blue) structures for CASP targets T0469 (left) and T0473 (right). C_α RMSDs are 5.0 and 5.4 Å, respectively. (Colour figures will be included in the proceedings available online).

Comparing the experimental results with our sampled conformations, it is apparent that we need to improve the non-local energy terms considerably. Furthermore, the secondary structure predictions that we used to restrict the search in conformational space turned out to be detrimental to the results in some cases. We are currently working on an improved description of non-local features in folded proteins which does not rely on a fixed secondary structure and we expect considerable improvements.

Acknowledgements

We acknowledge funding by the Danish *Program Commission on Nanoscience, Biotechnology and IT (NABIIT)* (project: simulating proteins on a millisecond time-scale) and the *Danish Research Council for Technology and Production Sciences (FTP)* (project: data driven protein structure prediction).

References

Anfinsen, C. B., Haber, E., Sela, M. and White, F. H. (1961). The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proc. Natl. Acad. Sci. U.S.A.*, **47**, 1309–1314.

Boomsma. W., Mardia, K. V., Taylor, C. C., Ferkinghoff-Borg, J., Krogh, A. and Hamelryck, T. (2008). A generative, probabilistic model of local protein structure. *Proc. Natl. Acad. Sci. U.S.A.*, **105**, 8932–8937.

Dominy, B. N. and Brooks, C. L. (1999). Development of a generalized born model parametrization for proteins and nucleic acids. *The Journal of Physical Chemistry B*, **103**, 3765–3773.

Fabiola, F., Bertram, R., Korostelev, A. and Chapman, M. S. (2002). An improved hydrogen bond potential: impact on medium resolution protein structures. *Protein Sci.*, **11**, 1415–1423.

Ferkinghoff-Borg, J. (2002). Optimized monte carlo analysis for generalized ensembles. *The European Physical Journal B*, **29**, 481–484.

Hamelryck, T. (2005). An amino acid has two sides: a new 2D measure provides a different view of solvent exposure. *Proteins*, **59**, 38–48.

Hartmann, C., Antes, I. and Lengauer, T. (2007). IRECS: a new algorithm for the selection of most probable ensembles of side-chain conformations in protein models. *Protein Sci.*, **16**, 1294–1307.

Hesselbo, B. and Stinchcombe, R. B. (1995). Monte carlo simulation and global optimization without parameters. *Phys. Rev. Lett.*, **74**, 2151–2155.

Jones, D. T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.*, **292**, 195–202.

Kryshtafovych, A., Fidelis, K. and Moult, J. (2007). Progress from CASP6 to CASP7. *Proteins*, **69**, 194–207.

Moult, J. (2006). Rigorous performance evaluation in protein structure modelling and implications for computational biology. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.*, **361**, 453–458.

Røgen, P. (2005). Evaluating protein structure descriptors and tuning Gauss integral based descriptors. *Journal of Physics: Condensed Matter*, **17**, S1523–S1538.

Røgen, P. and Fain, B. (2003). Automatic classification of protein structure by using Gauss integrals. *Proc. Natl. Acad. Sci. U.S.A.*, **100**, 119–124.