Candidate Fragments Prediction and their Assembly with a Greedy Algorithm and a Coarse-Grained Force Field to solve Protein Folding

Julien Maupetit[1], Frederic Guyon[1], Anne-Claude Camproux[1], Philippe Derreumaux[2], and Pierre Tufféry[1]

[1] Equipe de Bioinformatique Génomique et Moléculaire, INSERM E0346, Université Paris 7, Tour 53/54 1er Etage, 2 place Jussieu, 75251 Paris Cedex 05, France
{maupetit, guyon, camproux, tuffery}@ebgm.jussieu.fr
[2] Laboratoire de Biochimie Théorique, UPR 9080 CNRS, IBPC et Université Paris 7, 13 rue Pierre et Marie Curie, 75005 Paris, France
philippe.derreumaux@ibpc.fr

Abstract: *Prediction of proteins tertiary structure, starting from its amino acids sequence, still remains a challenge in computational biology. We are developing a method to fold proteins in silico, starting from a HMM based structural alphabet which consists of a local 3D description of the structure. Candidate protein fragments are selected by SAFrAN, a new original approach combining SA-Search and profile prediction conditionned by PSIPRED results. Selected fragments cover the target sequence with more than 90% and can approximate the native structure at high accuracy (less than 2 Å). Fragment assembly is performed by an improved greedy algorithm, and the relevance of the models is evaluated by a simplified version of the coarse-grained Optimized Potential for Efficient structure Prediction (sOPEP). We discuss the effectiveness of the approach to generate de novo 3D models of proteins.*

Keywords: candidate fragment prediction, fragment assembly, structural alphabet, greedy algorithm, coarse grained potential, protein structure prediction.

1 Introduction

Large scale genome sequencing projects bring sequences databases to grow exponentially. Even if the number of known proteins structures evolves in the same way, only few of the detected ORFs correspond to proteins having an experimentally resolved 3D structure. For such a reason, we need high throughtput methods to generate 3D protein structures. Moreover, for the proteins that fall out the range of application of comparative modelling, we need techniques able to generate new folds. This is the main purpose of *ab initio* and *de novo* structure prediction methods. *Ab initio* methods such as UNRES [1] are only established on protein physical properties : starting from a fully unfolded peptidic chain, their goal is to generate the native fold as it could happen in the natural protein environment. *De novo* methods are generally constructed on protein structure statistical properties. Most of these methods are based on selected protein fragments. To rebuild protein structure, fragment assembly methods are used, combined to an energy function. Presently, the most successfull methods are Rosetta [2,3] and TASSER [4,5]. The method we have developed can be related to them, but is based on the concept of structural alphabet. First, we predict candidate fragments from the amino acids sequence, using an HMM derived structural alphabet. Then, we assemble such fragments using

a folding potential to generate protein models. We will discuss the lessons of such approach applied during CASP7 (Critical Assessment of Techniques for Protein Structure Prediction).

2 Model generation

2.1 HMM-SA: HMM based Structural Alphabet

Structural alphabet properties. The present method is based on a Hidden Markov Model (HMM) learnt from 1429 PDB [6] structures [7,8]. Each letter of the alphabet is a 4 residues length protein fragment. Letters overlap on 3 residues. HMM is described by 4 distances between alpha carbons: C1C3, C1C4, C2C4, and the projection of C4 on the plane described by C1C2C3. This results in an optimal SA (Structural Alphabet) of 27 states or letters corresponding to 155 prototypes (or sub-conformations). The letters are labelled `[A-Z,a]`.

Structural alphabet encoding of proteins. Each state is described by a multi-gaussian density, and the transition matrix of the Markovian process quantifies the connections between the letters. Given such model, it is possible to encode each protein as a series of letters of the alphabet using the forward-backward algorithm for example. It computes, for each position, the probability that the structure is represented by each of the 27 states, then returns the most probable letter at each position.

Structural alphabet prediction from amino acids sequence. To predict SA letters from amino acids, we use the Markovian Model from the probability that SA letters emit amino acids sequences. This was learnt from a collection of HMM-SA encoded non redundant proteins. In addition, it is possible to constrain the forward-backward algorithm to a subset of the SA-letters at each position of the sequence. We use this possibility to constrain the prediction using PSIPRED [9]. PSIPRED evaluate the confidence level of the prediction between 0 and 9. So, for the regions of the sequence predicted with a confidence more than a given threshold (here 5), we apply the following constraints: regions predicted as

i – helices, we only consider `[a A V W Z B C D E]`,
ii – strands, the set of letters is `[L M N T X J K]`,
iii – coils, we consider `[B C D E F G H I J K L N O P Q R S T U Y Z]`,
iv – and others, the full alphabet is used (27 letters).

2.2 SAFrAN: Structural Alphabet candidate Fragments from AmiNo acid sequence

To determine the collection of candidate fragments, SAFrAN's method steps are : (i) predict HMM-SA profile from the amino acids sequence conditionally to PSIPRED [9]), (ii) search fragments compatible with the predicted profile in an HMM-encoded non-redundant PDB-derived profiles database (this method is derived from SA-Search [10]), and then (iii) apply filters to refine fragment selection (amino acids compatibility and PsiPred compatibility). The step (ii) and (iii) are iterated until the maximal percentage of covered sequence is reached (no accurate new fragment can be detected). Remaining non-covered parts are filled using direct HMM-SA prediction from sequence.

2.3 The greedy algorithm

The protein model is built linearly by overlapping predicted series of fragments. At each step of the model reconstruction, the greedy algorithm builds all possible combinations of fragments, ranks them according to an objective function, and keeps only the h best solutions for the next iteration (h is called the heap). The original algorithm proposed by Kolony and Levitt [12], has been improved by three ways [13]: (i) the algorithm is now stochastic (a part of the heap is randomly choosen), (ii) we add prefilters, consisting of mini-runs on small parts of the structure, to avoid some transitions, and (iii) the algorithm is now iterated : we build the structure from N to C terminus and then from C to N terminus.

2.4 The objective function

Algorithm validation. To test the capability of the algorithm to reconstruct protein structures, cRMSd (α carbons Root Mean Square Deviation) criterion was first used to drive the greedy algorithm, combined with a fuzzy description of the structure to rebuilt (HMM-SA letters with a probalility $p > 10^{-6}$). Reconstruction of protein structures are on average achieved at less than 1 Å cRMSd [13]. In a second time, the improved version of the greedy algorithm, driven by a Go based criterion, leads to structures differing by less than 4.8 Å cRMSd from the native structure [13]. Using only secondary structure information, and considering all the possible letters for non structured regions, the procedure could built 20 protein structures of 50-164 amino acids within 2.7 to 6.5 Å cRMSd [14] under Go criterion.

Coarse-grain force field. The greedy algorithm requires a computationaly non-expensive energetic function to drive the search during the folding simulations. In this context, the choice of coarse-grained force field seems appropriate. OPEP (Optimized Potential for Efficient structure Prediction) [15] is a coarse-grained force field established on a six-bead model per amino acid : the protein backbone is fully defined (N, Cα, C and O atoms are explicit), and the side-chains are represented by one single bead. It has been found that this force field is efficient to study protein folding and agregation [16,17,18,19,20,21,22,23]. We used a simplified version of the OPEP force field to skip all geometrical energetic terms (*i.e.* valence, bond lengths, dihedral angles except $\Phi > 0$). This version of the force field is named sOPEP for simplified Optimized Potential for Efficient structure Prediction. We are now optimizing sOPEP parameters by the same way as OPEP [24].

3 Results

3.1 Candidate fragments selection

We test SAFrAN on 228 representative selected PDB structures (see figure 1). It appears that, most of the time, only one fragment is selected by position of the target amino acids sequence (fig. 1A). The fragments lenghts are, on average, between 6 and 9 residues long (fig. 1B), but for some targets, fragments longer than 40 residues were obtained. Fragments accuracy has been assessed locally and globally. Local accuracy is checked by superposing each fragment onto the native structure of the target, and by deducing the cRMSd between the structures. Figure 1C shows that 75% of the fragments have a local cRMSd of less than 2 Å from the native structure. This result is quite identical to those obtained by Yang and Wang [11], but appears worse than Kolodny et al [12]. This makes sense: they used clusterized fragments of less than 7 residues long, and long fragments have a

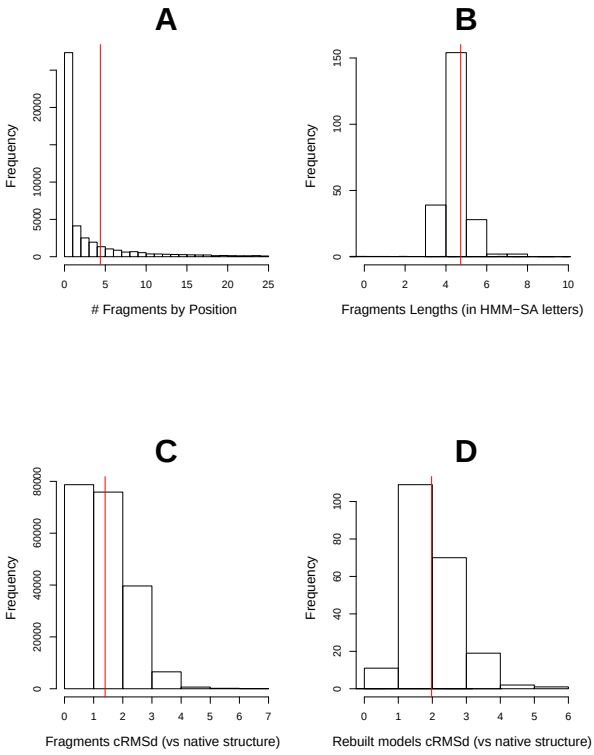

Figure 1. SAFrAN, candidate fragments properties. **A**: distribution of the number of fragments per position. **B**: distribution of the fragments lengths in HMM-SA letters (the x-axis is truncated at 10). **C**: distribution of the local-fit approximation. **D**: distribution of the global-fit approximation. Red vertical lines represent mean values.

higher probability to derive from the local structure than shorter ones, due to the decreasing number of degrees of freedom. Global accuracy of the candidate fragments is evaluated by calculating the cRMSd between the native structure and the model rebuilt using the HMM-SA letters of the fragments to feed the greedy algorithm under cRMSd criterion (see section 2.3). The distribution of these cRMSd is shown in figure 1D. Taken together, the candidate fragments could lead to a near-native structure, with an average cRMSd of 2 Å. Note that 60% of the models have a cRMSd value of less than 2 Å, and 90% of less than 3 Å when compared to the native structure. Such global approximation of the native structure is comparable to that obtained by Kolodny et al [12], but, here, starting from prediction.

3.2 CASP7

Last summer, we participated to the 7th edition of the Critical Assessment of Techniques for Protein Structure Prediction. We predicted the structure of 15 targets from the 3 CASP categories : 4 free modelling (FM) targets, 9 high accuracy template based modelling (HA-TBM) targets, and 2 template based modelling (TBM) targets. Only few targets were submitted due to concurrent development and improvement of the procedure during CASP. Emphase was put on FM targets. During this experiment, we started to develop a new homology modelling method starting from 3D-Jury [26] templates.

Assessing candidate fragments quality. For each kind of target, we analyse the fragments predicted by SAFrAN, in terms of (i) percentage of coverage of the target sequence, (ii) complexity

	FM	HA-TBM	TBM	TOT
% Coverage	97%	94%	88%	94%
Search complexity (1)	23.43 ±5.20	19.43 ±6.8	18.86 ±4.74	20.49 ±6.05
Search complexity (2)	14.11 ±2.61	12.19 ±3.51	12.31 ±2.02	12.72 ±3.09
# HMM-SA letters / Pos	7.44 ±4.84	5.99 ±4.82	5.61 ±4.56	6.20 ±4.83
Best Rebuilt cRMSd (1)	0.88 Å ±0.49	1.62 Å ±0.52	1.34 Å ±0.18	1.39 Å ±0.57
Best Rebuilt cRMSd (2)	1.12 Å ±0.41	1.96 Å ±0.54	1.68 Å ±0.33	1.70 Å ±0.59

Table 1. SAFrAN performance. (1) Using all prototypes by letter. (2) Using 3 prototypes maximum by letter. TOT = FM + HA-TBM + TBM.

of the search (*i.e.*, the average number of rigid fragment used per residue during model generation), (iii) the average number of SA letters describing each position of the structure (max is 27, *i.e.* no prediction), and (iv) best built cRMSd. The results are presented in table 1. Despite of the small number of analysed proteins (15 targets), previous observed SAFrAN's properties seems conserved: coverage of the target sequence is good (about 90%), with a surprisingly high level for FM targets, and the average best rebuilt cRMSd is lower than 2 Å, with the lowest value for FM targets (0.88 Å). During CASP7, we note that the search complexity could be decreased by only considering the 3 most populated prototypes of a HMM-SA fragment. This simplification, on average, only increases the best rebuilt cRMSd by 0.3 Å (from 0 to 1 Å), and decrease the computation time necessary to generate a model by a factor 2 (74 prototypes are considered, instead of the 155 original one)[3]. This observed complexity is in accordance with Kolodny et al [12], and Micheletti et al [27] prior works, when considering a library size of 155. The number of predicted HMM-SA letters by position vary from 5 to 8, depending of the query sequence structuration signature. To conclude, the SAFrAN method seems efficient, since we have the solution in the selected fragments.

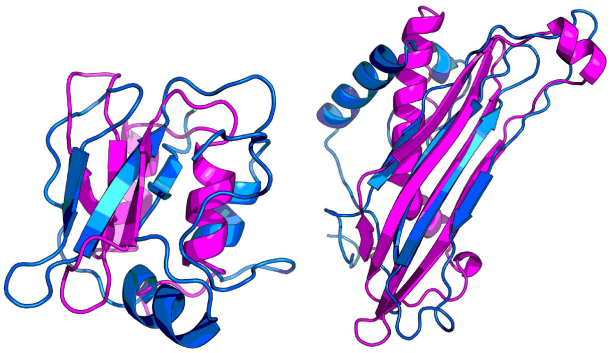

Figure 2. T0358, FM *(left)* and **T0383, FM** *(right)*. The native structure is colored in magenta, and our superposed model in blue. Pictures generated using the PyMol software [25].

Assessing model generation procedure In some cases, the complete procedure, starting from sequence only (without template structure), was able to produce topologically near native models as pertinent as the Robetta models [28]. We show here two *de novo* predicted models for the targets T0358 and T0383 (see figure 2). For homology modelling, results were more problematic: it seems

[3] Greedy algorithm is implemented in C language ; a typical simulation for a protein of 100 residues length takes about fifteen hours simulation on a single modern CPU (AMD Opteron 2.4 GHz).

that assembly procedure is too rigid. Finally, we observe that the procedure needs a final model refinement able to relax the whole structure.

4 Perspectives

We have described a new original approach to predict 3D candidate fragments from an amino acids sequence. Assembled by a greedy algorithm, these fragments can, in some cases, generate relevant all atom protein models. The method is still under development, but it has revealed some promising results for protein structure prediction during the first test phase. The secondary structure prediction performance has to be further examined. We will now focus on the improvement of *de novo* folding simulations and the derived homology modeling method. Outside from the model generation problem, SAFrAN could be helpfull for experimentators to determine the local conformation of experimentaly non resolved protein regions (NMR and X-Ray analysis).

Acknowledgements

We want to thanks J. Martin and L. Regad for their precious help during the development of the method and the CASP experiment. Thanks to G. Debret for the relecture of the manuscript. Some of the computing was made at RPBS (`http://bioserv.rpbs.jussieu.fr`).

References

[1] Rojas AV, Liwo A, Scheraga HA. Molecular dynamics with the United-residue force field: ab initio folding simulations of multichain proteins. *J Phys Chem B.*, 111(1):293-309, 2007.

[2] Simons KT, Bonneau R, Ruczinski I, Baker D. Ab initio protein structure prediction of CASP III targets using ROSETTA. *Proteins.*, Suppl 3:171-6, 1999.

[3] Misura KM, Chivian D, Rohl CA, Kim DE, Baker D. Physically realistic homology models built with ROSETTA can be more accurate than their templates. *Proc Natl Acad Sci U S A.*, 103(14):5361-6, 2006.

[4] Zhang Y, Arakaki AK, Skolnick J. TASSER: an automated method for the prediction of protein tertiary structures in CASP6. *Proteins.*, 61 Suppl 7:91-8, 2005.

[5] Pandit SB, Zhang Y, Skolnick J. TASSER-Lite: an automated tool for protein comparative modeling. *Biophys J.*, 91(11):4180-90, 2006.

[6] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. *Nucleic Acids Res.*, 28(1):235-42, 2000.

[7] Camproux AC, Tuffery P, Chevrolat JP, Boisvieux JF, Hazout S. Hidden Markov model approach for identifying the modular framework of the protein backbone. *Protein Eng.*, 12(12):1063-73, 1999.

[8] Camproux AC, Gautier R, Tuffery P. A hidden markov model derived structural alphabet for proteins. *J Mol Biol.*, 339(3):591605, 2004.

[9] Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.*, 292:195-202, 1999.

[10] Guyon F, Camproux AC, Hochez J, Tuffery P. SA-Search: a web tool for protein structure mining based on a Structural Alphabet. *Nucleic Acids Res.*, 32(Web Server issue):W545-8, 2004.

[11] Yang AS, Wang LY. Local structure prediction with local structure-based sequence profiles. *Bioinformatics.*, 19(10):1267-74, 2003.

[12] Kolodny R, Koehl P, Guibas L, Levitt M. Small libraries of protein fragments model native protein structures accurately. *J Mol Biol.*, 323(2):297-307, 2002.

[13] Tuffery P, Guyon F, Derreumaux P. Improved greedy algorithm for protein structure reconstruction. *J Comput Chem.*, 26(5):50613, 2005.

[14] Tuffery P, Derreumaux P. Dependency between consecutive local conformations helps assemble protein structures from secondary structures using Go potential and greedy algorithm. *Proteins.*, 61(4):73240, 2005.

[15] Santini S, Wei G, Mousseau N, and Derreumaux P. Exploring the Folding Pathways of Proteins through Energy Landscape Sampling: Application to Alzheimer's Amyloid Peptide. *Internet Electron. J. Mol. Des.*, 2, 564577, 2003.
[16] Derreumaux P. From polypeptide sequences to structures using Monte Carlo simulations and an optimized potential. *J. Chem. Phys.*, 111:2301-2310, 1999.
[17] Derreumaux P. Generating ensemble averages for small proteins from extended conformations by Monte Carlo simulations. *Phys Rev Lett.*, 85:206-209, 2000.
[18] Forcellino F., Derreumaux P. Computer simulations aimed at structure predictions of supersecondary motifs in proteins. *Proteins.*, 45:159-166, 2001.
[19] Derreumaux P. Insights into protein topology from Monte Carlo simulations. *J Chem Phys.*, 117:3499-3503, 2002.
[20] Malek R, Mousseau N. Dynamics of lennard-jones clusters: A characterization of the activation-relaxation technique. *Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics.*, 62(6 Pt A):7723-8, 2000.
[21] Wei G, Mousseau N, Derreumaux P. Exploring the energy landscape of proteins: a characterization of the activation relaxation technique. *J Chem Phys.*, 117:11379-11387, 2002.
[22] Wei G, Mousseau N, Derreumaux P. Complex folding pathways in a beta-hairpin. *Proteins.*, 56:464-474, 2004.
[23] Mousseau N, Derreumaux P. Exploring the early steps of amyloid peptide aggregation by computers. *Acc Chem Res.*, 38:885-91, 2005.
[24] Maupetit J, Tuffery P and Derreumaux P. A coarse-grained protein force field for folding and structure prediction. *Proteins.* In press.
[25] DeLano WL. The PyMOL Molecular Graphics System. *San Carlos, CA, USA: DeLano Scientific*, 2002.
[26] Ginalski K, Elofsson A, Fischer D, Rychlewski L. 3D-Jury: a simple approach to improve protein structure predictions. *Bioinformatics.*, 19(8):1015-8,2003.
[27] Micheletti C, Seno F, Maritan A. Recurrent oligomers in proteins: an optimal scheme reconciling accurate and concise backbone representations in automated folding and design studies. *Proteins.*, 40(4):662-74, 2000.
[28] Kim DE, Chivian D, Baker D. Protein structure prediction and analysis using the Robetta server. *Nucleic Acids Res.*, 32(Web Server issue):W526-31, 2004.