Materials Physics and Mechanics 9 (2010) 162-166 Received: March 17, 2010

DE NOVO PROTEIN STRUCTURE PREDICTION BY SIMULATION OF FOLDING PATHWAYS

Sergey Feranchuk

United Institute of Informatics Problems of National Academy of Sciences of Belarus

Surganova 6, 220012, Minsk, Belarus

E-mail: s@feranchuk.linux.by

Abstract. We have developed a new de novo protein structure prediction algorithm, SKI-FOLDING. The search space of the algorithm is built from all the possible combinations of secondary structure elements in a given subset of protein structures databank. Search directions follow the best possible folding pathways, with a simple statistical score function. We tested the algorithm on the ability to distinguish between different families within the immunoglobulin-like type of fold, for 9 proteins with known structure. For the most of tested proteins the correct structure was presented in the program output and false-positive decoys were of worse quality than the correct structure.

1. Introduction

Significant results were achieved in template-based modeling of proteins, as the size of the protein structures databank increased. However, template-free modeling of a protein structure from a primary sequence still remains of some value for researchers [1, 2]. Aside from the need to deal with "hard" cases in structure prediction, where no significant homology can be found to the target protein, there is some interest in "demystification" of folding process and in explaining which energy balance lies behind a particular fold.

As it is commonly accepted [3, 4], the way how proteins get their structure in vitro and the way how to simulate it in silico is to assemble them from building blocks following some set of pathways. The problem in selecting the simulation way is how to choose the building blocks, how to combine them and how to select the best combinations. A lot of empirical potentials can be developed for the estimation of structure quality, but our experience reveals that the physical background and physical sense of such potentials is important.

In the protein structure the most physically clear difference between chain segments is the difference between secondary structure elements and coils. For these reasons we take the secondary structure elements as building blocks and try to combine them. We build the bank of possible combinations of secondary structure elements from some subset of protein data bank and put no restrictions on intermediate structures, except for the value of energy function. The problem is a choice of search strategy.

In our research we take the approach of folding pathway simulation, where the folding pathway is hierarchical and follows the primary structure order. By this approach, for each continuous region of the chain we can select a limited number of "best" pathways. The regions themselves can have different lengths and can overlap. The search strategy then consists of just two possibilities – expansion of region by new secondary structure elements and connection of adjacent regions. For the chain of length n we have $n(n-1)/2$ independent regions and a limited number of intermediate structures. As the energy function we take just hydro-

phobicity of residues multiplied by number of contacts (plus some value for coils and hydrogen bonds). Our simulation reveals that it is not enough to distinguish between different secondary structures types because of some extra tertiary limitations. But if we limit the search space to a particular fold type, we can distinguish between different families and super families.

2. Methods

Overview of tools in package. For prediction of a tertiary structure by the proposed approach one needs a database of secondary structure elements combinations. This database can be extracted from any subset of protein database. For each structure, first the secondary structure elements are defined (using KSDSSP program with the algorithm from [5]). The secondary structure element is described by its orientation and size, just like a rigid body. To build possible combinations of secondary structure elements for a particular structure, one needs to know the order of structure assembly. For this purpose possible folding pathways are reconstructed. After the list of secondary structure element combinations is ready, similar combinations are joined together by clustering.

After the search procedure, each predicted structure is just a list of secondary structure elements with absolute orientations and correspondences to the primary sequence. From these data, theoretically the detailed tertiary structure could be reconstructed, but for comparison we used only the skeleton of the structure. The search produces a number of "best" structures (usually 100).

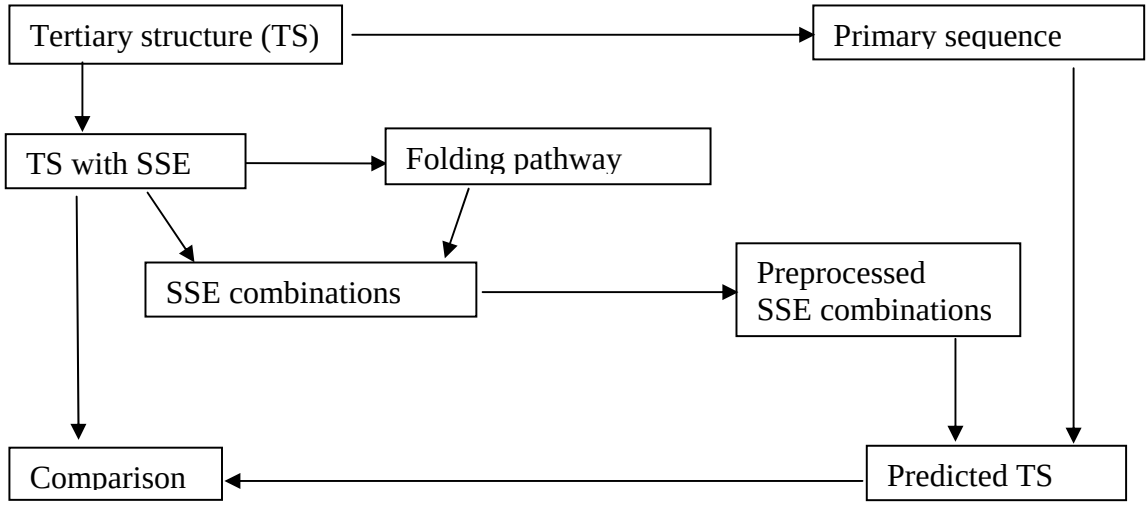

Fig. 1. Scheme of intermediate data types and data transformations in the algorithm. SSE (secondary structure elements).

Energy function. As we select a secondary structure element as a building block, we need describe the structure no deeper than a residue level. Inside a given secondary structure elements combination, the number of contacts between residues should not depend on a residue type. So, to define contacts we just build residue centers (in the direction of C-beta atoms) and check whether the distance between residues' centers is less then the threshold of 8 Angstroms. As we have noticed, the most robust way to derive energy from the number of contacts for a given residue is just to multiply this number by the hydrophobic coefficient of residue, according to [6]. So the energy function cold be written as

$$E = E_C + E_H + E_L + E_F$$

Here we use a reverse scale for energy, so that the optimal energy is the maximal energy E_C relating to the contact energy and is defined as

$$E_C = \sum_i n_i w(a_i), \quad w(j) = RT(\ln p(j) - s_0), s_0 = \sum_j p(j) \ln p(j), p(j) = \frac{N_c(j)}{N(j)N_0}, N_0 = \left\langle \frac{N_c(j)}{N(j)} \right\rangle = 6.51$$

Here n_i is the number of contacts for a given residue, $w(a_i)$ is the hydrophobic coefficient of a residue, based on statistics from a non-redundant set of the protein database (PDB); N_c is the number of contacts with a given residue, N is the number of occurrences of a given residue. E_H relates to the energy of hydrogen bonds.

Hydrogen bonds in SSE combinations were defined according to secondary structure definition by KSDSSP and their energy is just the number of hydrogen bonds in an intermediate structure multiplied by some factor. To choose this factor we started from the assumption that the hydrogen bond energy would contribute approximately 50% to the total energy. Another reason is that the energy of one hydrogen bond should be that of one hydrophobic contact. As the value of RT is 0.57 kKal/mol and $p(j)/\ln s_0$ is 2 for isoleucine, we take energy of one hydrogen bond equal to 0.3 kKal/mol.

E_L is the energy of coils and is defined statistically. We processed each coil in non-redundant PDB chains as a segment between two adjacent SSEs and defined for it the number of residues l, the distance between ends d and the angle a between directions of adjacent SSEs. Thus we obtained the distribution N(l,d,a). Then we assume that the average coil energy should not depend on a coil length. This assumption gave us the following formula for the energy of a particular coil:

$$E_L(l,d,a) = RT \ln N(l,d,a) - s(l), s(l) = \sum_{d,a} p(l,d,a) \ln p(l,d,a), p(l,d,a) = N(l,d,a) / \sum_{d,a} N(l,d,a)$$

E_F is the term for the loss of conformational freedom in SSE. It is the same for each SSE and is counted as the minimal SSE length multiplied by an average number of contacts and an average hydrophobic coefficient. We took E_F = -1.4 kKal/mol for each SSE.

Search strategy. The total number of possible structures in enumeration will obviously be enormously large. In coiled state each segment of a chain is fluctuating around various types of local fold, but some local folds will be preferred for energy reasons. Here local fold can be compared with a node in a folding pathway. Best nodes survive and are joined together, so that the overall structure is assembled from building blocks. The key idea in our search strategy is that for each segment of the chain we consider as possible local folds for this segment only a limited number of "best" nodes. For a chain of length n we have $n(n-1)/2$ segments and the possible number of structures is always $O(n^2)$.

Competition between pathways starts at the level of two joined SSEs. This means, we starts from all possible single SSEs, enumerate all possible pairs of SSEs, then try to find segments of the chain with the best fit to a given pair. In this process, the pairs compete; so, if two pairs fit to the same segment, only the best is left. After SSEs are "settled" on the chain, possible actions are to prolong the given local fold by one SSE in any direction or to merge two segments. All these actions follow the rules from the database of possible SSE combinations and so the absolute positions of SSEs in the cluster are calculated. In each step we can determine the number of contacts for a given position in the SSE and can derive the energy of a local fold.

3. Results

The immunoglobulin-like type of fold is widely represented in the protein databank. It is constructed from two beta-sheets forming a sandwich-like structure. According to the structural classification of proteins (SCOP) classification [7], it has several super families, families and some deeper gradation. We describe this classification by four numbers, as it is used in the table below. Despite this type of fold being very stable to variations in a primary sequence,

the order of beta-segments in the sheets may vary between families and super families. In our experiment we check the ability of our program to distinguish these differences.

We built our database of SSE combinations from 43 random domains with the immunoglobulin-like type of fold from a non-redundant PDB subset. Then we predict pathways (pw) for 9 proteins with a known structure. As a result, for each protein we obtained a hundred "best" pathways. In this list we tried to find the best fit to a given PDB structure, using the TM-align program [8]. The criteria for comparison is so-called TM-score that ranges from 0 to 1 where 1 means identity. We compared each list of pathways with each structure. In the table below, the results of the comparison are listed. The second number relates to the order of the selected structure in the program output. True matches are shown with boldface.

Table 1. Results of structure prediction. TM-score of the best predicted structure compared with a known structure.

	1ac6.pw b-1-1-1-5	1acx b-1-7-1-1	1nci b-1-6-1-1	1ah1 b-1-1-1-6	1b0w b-1-1-1-1	1cd0 b-1-1-1-2	1axi b-1-2-1-5	1a3r b-1-1-2-1	1c5c b-1-1-1-3
1ac6.pdb	**0.38 6**	0.36 32	0.33 36	0.32 1	0.32 68	0.29 26	0.28 3	0.33 96	0.28 28
1acx	0.32 58	**0.33 4**	0.34 36	0.30 13	0.32 68	0.28 5	0.25 1	0.28 96	0.28 25
1nci	0.30 41	0.33 76 ***(1D)***	**0.36 2** ***(1P)***	0.31 9	0.30 31	0.30 13	0.24 88	0.31 96	0.27 83
1ah1	0.34 6	0.34 8	0.31 23	**0.27 29**	0.29 68	0.26 44	0.24 3	0.30 96	0.26 45
1b0w	0.38 6 ***(2D)***	0.36 8	0.36 36	0.31 20	**0.33 68** ***(2P)***	0.29 6	0.30 45	0.35 96	0.29 32
1cd0	0.37 6	0.35 5	0.35 36	0.30 74	0.32 12	**0.29 81**	0.28 99	0.35 96	0.29 67
1axi	0.41 58	0.42 40	0.34 2	0.34 25	0.33 68	0.35 93	**0.30 10**	0.37 96	0.31 31
1a3r	0.34 41	0.35 67	0.32 10	0.31 81	0.31 31	0.32 3	0.28 49	**0.30 96**	0.30 31
1c5c	0.34 41	0.36 97	0.33 10	0.29 4	0.32 98	0.33 53	0.29 5	0.28 96	**0.31 31**

The selected methodology of comparison using structure alignment is not obvious. For example, in a recent CASP experiment de novo predictions were compared using distance matrices [2]. In figure 2 we also took several distance matrices in order to compare them visually – a true structure, a predicted structure and a decoy. We mark the appropriate cells in table 1. As one can see from the distance matrix, the positions of beta-hairpins in the native structure in most cases are significantly closer to the prediction than to the decoy, even if the score for the decoy is higher. We also note from the visual comparison of the structures that the "logic" of folding is closer to true prediction than to decoy.

4. Discussion

Our results show that the accepted search strategy can effectively select best structures. It is possible because different segments of a chain are fluctuating independently, and we can apply the 'divide and dominate' principle to dynamic programming. While the energy function serves as a fairly good estimate for structure quality, it seems to be too rough to find precisely a true solution. When we used larger datasets, the program made wrong predictions. Our first idea to improve the situation is to use independent secondary structure predictions as an additional filter. Our second idea is to use the energy function at atomic level and to analyze rotamer conformations for each predicted position of a residue. It is possible because the algorithm itself is fast enough, and we have a reserve of time for such analysis.

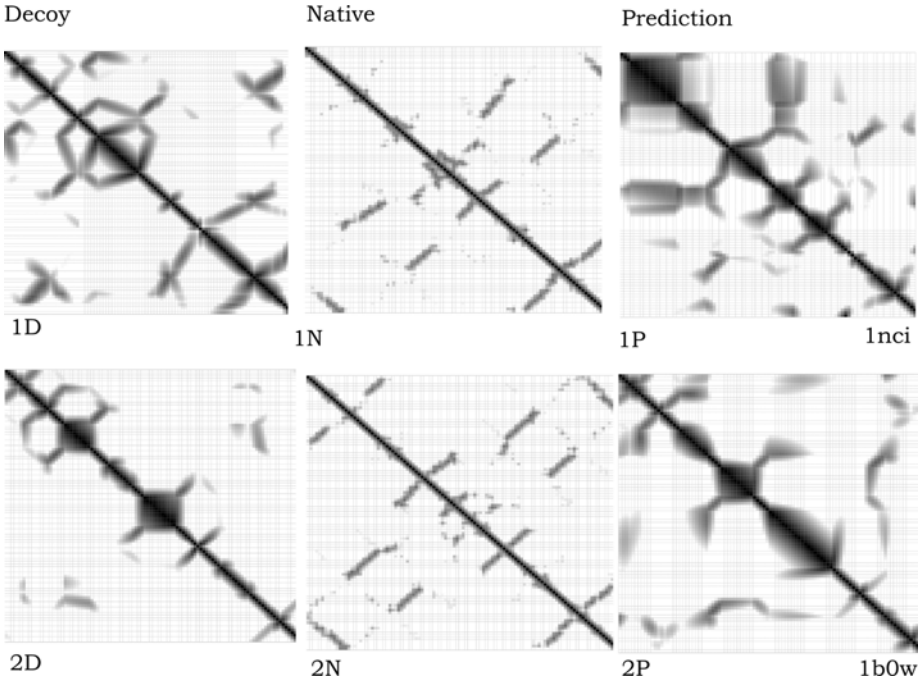

Fig. 2. Comparison of distance matrices for selected cases.

Acknowledgments

The research was made in the framework of the SKIF-GRID program and the BalticGrid project, with use of grid resources for time-consuming calculations. The author thanks Alexander Tuzikov for support.

References

[1] Y. Zhang // DOI: **10.1002/prot.21702** (2007)
[2] R. Jauch, H.C. Yeo, P.R. Kolatkar, N.D. Clarke // DOI: **10.1002/prot.21771** (2007)
[3] A.V. Finkelstein, O.V. Galzitskaya // *Physics Life Rev.* **1** (2004) 23.
[4] C.-J. Tsai, J.V. Maizel, R. Nussinov // *Proc. Natl. Acad. Sci. USA* **97** (2000) 12038.
[5]W. Kabsch and C. Sander // *Biopolymers* **22** (1983) 2577.
[6] P. Pokarowski, A. Kloczkowski, R.L. Jernigan, N.S. Kothari, M. Pokarowska, A. Kolinski // *Proteins* **59** (2005) 49.
[7] A.G. Murzin, S.E. Brenner, T. Hubbard, C. Chothia // *J. Mol. Biol.* **247** (1995) 536.
[8] Y. Zhang, J. Skolnick // *Nucleic Acids Research* **33** 2302–2309 DOI:**10.1093/nar/gki524** (2005)