J Comput Aided Mol Des (2008) 22:783–788
DOI 10.1007/s10822-008-9213-8

Fast and accurate methods for predicting short-range constraints in protein models

Dominik Gront · Andrzej Kolinski

Received: 12 July 2007 / Accepted: 25 March 2008 / Published online: 15 April 2008

Abstract Protein modeling tools utilize many kinds of structural information that may be predicted from amino acid sequence of a target protein or obtained from experiments. Such data provide geometrical constraints in a modeling process. The main aim is to generate the best possible consensus structure. The quality of models strictly depends on the imposed conditions. In this work we present an algorithm, which predicts short-range distances between $C\alpha$ atoms as well as a set of short structural fragments that possibly share structural similarity with a query sequence. The only input of the method is a query sequence profile. The algorithm searches for short protein fragments with high sequence similarity. As a result a statistics of distances observed in the similar fragments is returned. The method can be used also as a scoring function or a short-range knowledge-based potential based on the computed statistics.

Keywords Protein structure prediction · Spatial restraints · Comparative modeling · Profile alignment

Introduction

In many cases for a given protein sequence, homologous proteins (whose structures are already known) may be found in structural databases. The known structure may be used as a template in a comparative modeling procedure [1, 2]. Provided that the level of sequence similarity is high, one may expect accurate result. Moreover, homology based methods demand less computer time and are relatively easy to automate. Indeed, large-scale experiments have been started [3, 4] to model all known sequences for which it is possible to find a structural template.

Unfortunately in many cases sequence similarity observed between a query sequence and proteins from a database lies in a twilight zone. Then the evolutionary relationship between the query and the template becomes uncertain. This could be a serious limitation of comparative modeling since the probability of finding a template for a sequence randomly picked from a genome ranges about 50%, depending on the genome [5]. In the remaining cases the only way to obtain a 3D model for a query sequence is to do a de novo protein structure prediction. Although considerable progress has been made in this field, employing even sparse homology information still promises more accurate results.

If no global template structure can be identified for the target sequence, in many cases it is possible to find fragments of structures that can model some parts of the target. Such fragments, extracted from PDB files may be used as a source of spatial constraints to guide de novo modeling. There are also some other sources of information that can drive de novo search to a plausible topology or decrease the number of conformations that must be sampled.

The information predicted by sequence methods follows two main categories: long and short range. Starting from a protein sequence it is possible to predict contacts between residues [6, 7]. This data has a long-range nature, defining how a protein chain is folded and, as a consequence, describing its topology. On the contrary, short-range information describes local geometry of the backbone.

D. Gront (✉) · A. Kolinski
Faculty of Chemistry, University of Warsaw, Pasteura 1, 02-093 Warsaw, Poland
e-mail: dgront@chem.uw.edu.pl

A. Kolinski
e-mail: kolinski@chem.uw.edu.pl

Current methods [8–10] enable secondary structure prediction with accuracy (according to the Q3 measure) of about 75% [10]. Local geometry can also be described by short structural fragments (peptides) [11–13]. The advantage of such libraries of fragments is that they describe more precisely local conformation than the 3-state secondary structure alphabet does. Employing such 'building blocks' in a de novo modeling could be quite complicated. Sophisticated algorithms need to be designed for this purpose. An example is Rosetta method [14]. These algorithms are usually strictly related to specific databases of fragments characterized by their representation (e.g. Cartesian coordinates versus dihedral angles) and assumptions made in the derivation process (e.g. length and the size of fragments' database).

In our approach [15] the distributions of short-range distances between $C\alpha$ atoms are calculated. In this work we also show that a knowledge-based sequence-dependent scoring function and the most probable distances can also calculated from locally similar fragments. Our approach describes local conformation more accurately than predicted secondary structure does. In comparison with the methods that are based on explicit structural fragment assembly, the distance constraints are much easier to implement—they may be applied in virtually any modeling tool, e.g. in molecular dynamics, Monte Carlo, genetic algorithms and many hybrid methods.

To conduct experiments we utilized BioShell [16, 17] software package. Its most recent version provides modules that may be used in jython (python scripting language interpreter implemented in Java) scripts. Among numerous features, the package provides efficient routines for profile-to-profile alignment as well as statistical utilities. All algorithms described in this paper have been implemented as Jython (searching for fragments and GMM-EM estimation) or as bash (running Psi-Blast for a number of amino acid sequences in an automated manner) scripts. They can be downloaded from BioShell website: http://bioshell.chem.uw.edu.pl.

Methods

Overview of the method

To run a program, user should provide query and template sequence profiles [18] as well as template structures. During the search described below in details, local similarity between the two profiles (template and target) is assessed. Similarity of secondary structure, e.g. predicted by PsiPred [9], can also be assessed for higher specificity of the search. As a result, a set of short fragments cut out from the template structures is returned.

Fragment's search algorithm

The search method compares short gapless fragments of sequence profiles. The fixed length of the fragments is one of the parameters of the method, denoted as N_F, usually range 13–19 residues. Let P_i^Q and P_j^T denote the ith column from the target (query, denoted as **Q**) profile and the jth column from the template profile (denoted as **T**), respectively. The similarity score (denoted as S_{ij}) between the fragments starting at ith and jth amino acid in a target and template sequences, respectively, is a sum of the N_F scores calculated for corresponding positions in the target and in the template sequence profiles (Fig. 1):

$$\begin{aligned} S_{ij} &= \sum_{k=0}^{N_F-1} sim_{SP}(i+k, j+k) \\ &= \sum_{k=0}^{N_F-1} \left(sim_P\left(P_{i+k}^Q, P_{j+k}^T\right) \right. \\ &\quad \left. + sim_S\left(S_{i+k}^Q, S_{j+k}^T\right)\right) \end{aligned} \tag{1}$$

By $sim_P\left(P_i^Q, P_j^T\right)$ we denoted a similarity between two profile columns. Virtually any scoring scheme can be combined with our approach. Several most successful scoring schemes are implemented in BioShell library. In this work we used COMPASS [19] scoring function although some tests runs suggest that Picasso3 [20] would work similarly well. Additionally, similarity of secondary structure (based on three-state definition: H, E, C) for the target S^T and template S^Q is also assessed. Scoring function sim_S is defined as a simple similarity matrix based on H, E and C letters and taken from [21].

In practical implementation elements $sim_{SP}(i, j)$ comes from a matrix that contains scores that assess similarity between position i in a target and position j in a template protein. The size of the matrix is $N_Q \times N_P$. The matrix is computed as the first step of the program. In all the formulas (1–6) indices i and j denote a top-left corner of a

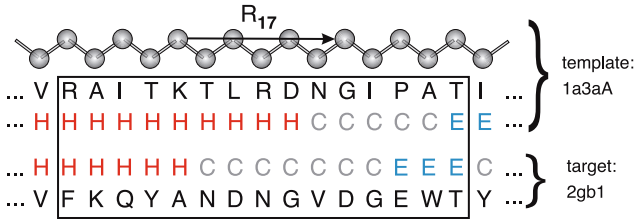

Fig. 1 Schematic illustration of the single step of the database search algorithm. A fragment of a template sequence profile is compared to a fragment of a target sequence profile (in the figure only single sequences are shown). If the similarity score is higher than a certain threshold, r13, r14, r15, r17 and r19 distances between $C\alpha$ atoms are measured from the central part of a template fragment (only r17 distance is shown for the sake of clarity)

certain submatrix of sim_{SP}. The submatrix contains scores calculated for a fragment from **Q** (starting at position i) and a fragment of **T** (starting at position j). The size of such submatrix is $N_F \times N_F$ and its elements are indexed by $i + k$ and $j + l$, where indices (k,l) change from 0 to $N_F - 1$

The S value is not particularly useful for scoring the similarity between two fragments. It highly depends on amino acid composition of compared profiles and on their length. The method that is frequently used in such situations is to calculate the z-score value:

$$z\text{-score} = \frac{S - <S>}{\sigma(S)} \tag{2}$$

The mean value of the score $< S >$ and the standard deviation of the score $\sigma(S)$ have to be estimated for all the permutations of the columns in both profiles. Thus, $< S >$ stands for an average alignment score for two profiles with a given length and amino acid composition, no matter what the amino acid order (column order in profiles) is. Usually, it is estimated via some random shuffling of one of the two profiles and recalculating the scoring function. Due to the assumptions mentioned above: fixed length of a fragment and its gapless nature, $< S >$ and $\sigma(S)$ can be calculated analytically in $O(N^2)$ and $O(N^4)$ time, respectively. The mean value is calculated as a sum of all possible scores for columns of the template and the query profiles, divided by N_F:

$$<S> = \frac{1}{N_F} \sum_{k=0}^{N_F-1} \sum_{l=0}^{N_F-1} sim_{SP}(i+k, j+l) \tag{3}$$

To calculate a variance from a well-known formula:

$$\sigma^2(S) = <S^2> - <S>^2 \tag{4}$$

an average value of squared-terms $< S^2 >$ must be calculated:

$$<S^2> = \frac{1}{N} \sum_{k=0}^{N_F-1} \sum_{l=0}^{N_F-1} (sim_{SP}(i+k, j+l))^2 + \frac{1}{N(N-1)} \sum_{k=0}^{N_F-1} \sum_{l=0}^{N_F-1} \mathcal{S}_{\mathcal{M}}(k,l) sim_{SP}(i+k, j+l) \tag{5}$$

where $\mathcal{S}_{\mathcal{M}}(i,j)$ denotes a minor of a considered sim_{SP} submatrix formed by eliminating row i and column j from sim_{SP}:

$$\mathcal{S}_{\mathcal{M}}(i,j) = \sum_{k=0 \neq i}^{N_F-1} \sum_{l=1 \neq j}^{N_F-1} sim_{SP}(i+k, j+l) \tag{6}$$

The distribution of S converges to the normal distribution when fragments are several residues long (see Fig. 2). Thus the use of z-score value is well-justified.

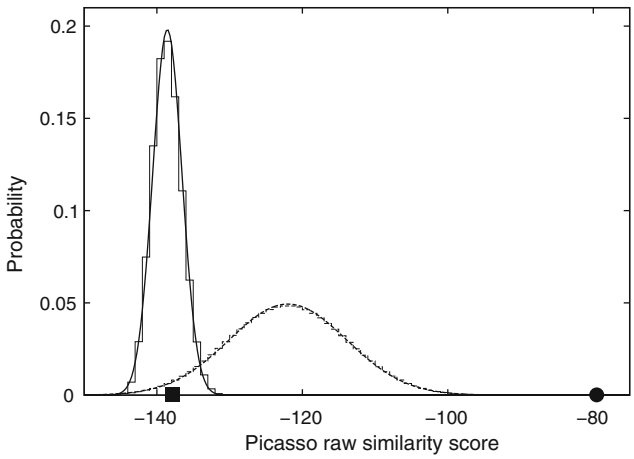

Fig. 2 Example distributions of Picasso similarity score between two fragments of evolutionary related (left peak) and randomly chosen pairs of sequence profiles (right peak). Stepwise plots show distributions obtained by 1,000,000 random shuffles of a sequence. Thick lines showing Gaussian approximation calculated by the algorithm described in this work perfectly agrees with the simulated data. Comparison score between a target and a template profiles is denoted by a square (for the pair of unrelated profiles) and by a circle (in the case of evolutionary related proteins). Z-score values for the two points are 0.01 and 5.11, respectively. The statistics were computed for fragments with 15 amino acid residues

When the total score is higher than a specified parameter, a proper structural fragment is cut out of a template structure and recorded. When the search is finished, a set of plausible structural fragments are returned. That structural information may be utilized in many ways. The first possibility, explored in our previous work, is to derive a sequence dependent scoring function, according to Boltzmann formalism. Here we apply Gaussian mixture model to derive a continuous scoring function and prediction of the most probable distance between two given Cα atoms.

It should be noted that our method does not predict fragments (and, as a consequence, distances and potentials) for all residues in a target sequence. Since the window for scoring the similarity between fragments of profiles always spans more residues ($N_F = 13$–19) than the measured fragment (usually 7–9 residues long), several N-terminal and several C-terminal residues (e.g. $(N_F - 7)/2$ or $(N_F - 7)/2$) are skipped. Moreover, for some sequence regions it is not possible to got a statistically significant hit. For example, in the experiment described in this work (see Results and discussion) nothing has been predicted for almost 30% of residues.

Gaussian mixture approximation

Originally the potentials that can be derived from fragments [15] were designed to work with a lattice modeling tool CABS [22]. Therefore the scoring functions are

represented in a form of histograms. This could be however not the best choice in other applications. In order to make our results more general, we also approximated the distributions of the distances by continuous functions. We tested several approaches and decided that Gaussian mixture (GM) is the most suitable for this task.

The unknown probability distribution function $f(x)$ is approximated as a linear combination of normal distributions:

$$f(x) = \sum_{i=1}^{k} \alpha_i N(x; \mu_i, \sigma_i) \tag{7}$$

where k is the number of components, α_i, μ_i and σ_i are mixing factor, mean value and standard deviation, respectively, for ith normal distribution, denoted as $N(x)$. The unknown parameters (α_i, μ_i, σ_i) are calculated via Expectation Maximization (EM) algorithm. Unfortunately, the EM procedure cannot determine optimal number of components (k). We employed the following iterative approach to find the best value of k. The procedure starts from a large k and for this value the best GM approximation is found according to the EM algorithm. After that, for each pair of normal distributions a dissimilarity measure is calculated. We choose symmetric Kullback–Leibler KL(i, j) divergence for this purpose—the standard dissimilarity measure between probability densities. In the case of Gaussian densities, the symmetric KL distance has a closed form:

$$\mathrm{KL}(i,j) = \frac{\sigma_i^4 + \sigma_j^4 + (\mu_i - \mu_j)^2(\sigma_i^2 + \sigma_j^2)}{2\sigma_i^2\sigma_j^2} + 1 \tag{8}$$

Equation (8) may be obtained directly from the definition of the symmetric Kullback–Leibler KL(i, j) divergence:

$$\mathrm{KL}(p_i, p_j) = \int p_i(x) \log\left(\frac{p_i(x)}{p_j(x)}\right) \mathrm{d}x + \int p_j(x) \log\left(\frac{p_j(x)}{p_i(x)}\right) \mathrm{d}x \tag{9}$$

in several steps of integrations. Analytical formulas describing symmetric KL divergence for some common univariate distributions may be found in [23]. A sketch of a derivation for the case of multivariate normal distribution may be found e.g. in [24].

If any two components are closer to each other than a certain threshold, k is decreased by one and the EM step is repeated for the new k. The procedure stops when any two components are less similar to each other than a threshold or when only one component remains.

Once the observed probability distribution $f(x)$ has been described as a mixture of Normal components one can select the most probable component (i.e. that with the highest value of α_i). The most probable distance may be introduced into a modeling protocol as a harmonic force.

Results and discussion

Sequence specificity of our knowledge-based scoring function has been already compared with the specificity of simple statistical potentials using gapless threading [15]. We also tested the ability to predict specific geometry of protein fragments. Significant improvement in threading sensitivity and increased ability to generate sequence-specific protein-like conformations has been achieved. Moreover, we found, that the new scoring function implemented in a Monte Carlo sampling scheme semi-quantitatively reproduces conformational properties of denatured proteins [25]. In this work, we investigate further applications of gapless comparison of short fragments of sequence profiles. We also bring a user-friendly and highly portable implementation of our algorithm to the public.

Results presented in this work show that the set of potentials calculated by Frags approach is able to assess the quality of a large group of decoys. In order to make test difficult, we focused on targets from the last CASP competition, denoted by organizers as "new folds." That means that no protein structurally similar to T0201, T0209 (second domain), T0216, T0238, T0241, T0242, T0248 (second domain) and T0273 can be found in the PDB database. We used PISCES_30_res2.0 [26] as a set of templates. This database contains only these protein structures culled from PDB, whose sequences are less similar to each other than 30% and whose resolution is not worse than 2.0 Å. That is the same database version that we used during the CASP6 experiment. We used PSI-BLAST [27] to calculate sequence profiles both for target and template sequences. We also employed secondary structure in our test. For targets it was predicted with Psi-Pred [9] and for template structure it was calculated with DSSP [28].

The accuracy of the Frags method depends on the quality of the sequence profile created in the first step of our method. For example, for T0201 no homologous sequence could be found and therefore PsiPred failed to build a sequence profile. Therefore, this protein was excluded from the further analysis.

For all targets except T0201 we derived our scoring function and measured the drmsd (distance root mean square deviation) between target structures (which are currently known) and the structures sent by CASP6 participants. In our analysis we took into account all groups i.e. both human and server predictors, taking all models (up to 5) submitted by each group. All the structures have been downloaded from CASP website. Since our potentials

scores only local geometry of a protein structure and does not recognize the topology, we took only short range distances (up to r19) into account in the drmsd calculations. Results are plotted on Fig. 3. The x axis refers to the total harmonic energy (r13 + r14 + r15 + r17 + r19 sequence-dependent potentials) for a nine-residue fragments. The y axis refers to the drmsd distance between fragments from the templates and the target structures. Overall correlation coefficient calculated for all models from all groups is 0.3. The four outliers in the upper middle in the Fig. 3 correspond to two CASP targets (T0216 and T0241) for which a very limited number of sequence homologs were found. Resulting low-quality sequence profile lead to bad prediction results.

Our results show that applicability of Frags approach depends the number of homologs that may be found for a given target in sequence databases. The quality of both target and template profiles affects the sensitivity of profile-profile scoring scheme. It also influences secondary structure prediction and as a consequence, the second term of our scoring system. After initial optimization we decided to combine the two scoring terms with equal weights and to use 15 residues long fragments ($N_F = 15$). There is however a trade-off between optimization of sequence-based and secondary structure related terms. The use of longer protein fragments allow for more accurate detection of locally conserved sequence motifs. Unfortunately, the longer protein fragment is, the greater chance that it spans two or even more secondary structure elements. In such a case a high score for secondary structure match may be gained, although the fragments may be not structurally similar because secondary structure elements may have different spatial orientation in the two protein structures.

The method described in this work allow prediction of local geometry of protein backbones starting from sequence information and selected structures of the Protein Data Bank. The local distances between the alpha carbon atoms could be used as a supporting distances in various methodologies of protein structure prediction and ranking of protein models. To test the predictive strength of the proposed method the set of the models submitted in the "new fold" (and fold-analogy) category of the CASP6 [29, 30] experiment has been evaluated by the Frags approach. The data collected in Fig. 3 clearly show strong correlation between the quality of the local geometry of the models and the score returned by the Frags method. Interestingly, almost all models submitted by the two best groups show better (or much better) local geometry than the average for all predictions. This leads to two more general observations. Firstly, the CASP6 [31] experiment indicates that the fidelity of the local geometry is strongly correlated with overall quality of the molecular models. Secondly, from comparison of the results of the two best groups it appears that the first observation is independent of the type of methodology employed in de novo modeling. This is somewhat surprising, since one of these groups (i.e. the Baker group) employed a fragment assembly technique [29] while the second one simulated protein folding controlled by a force-field derived from statistical analysis of general structural regularities seen in the experimentally solved protein structures [22, 30]. The potentials analogical to these provided by Frags were one of the most important components of the knowledge-based force field of the Monte Carlo folding algorithm CABS [22]. Thus we hope, that the proposed tool and the method will find several applications also in context of other protein modeling schemes.

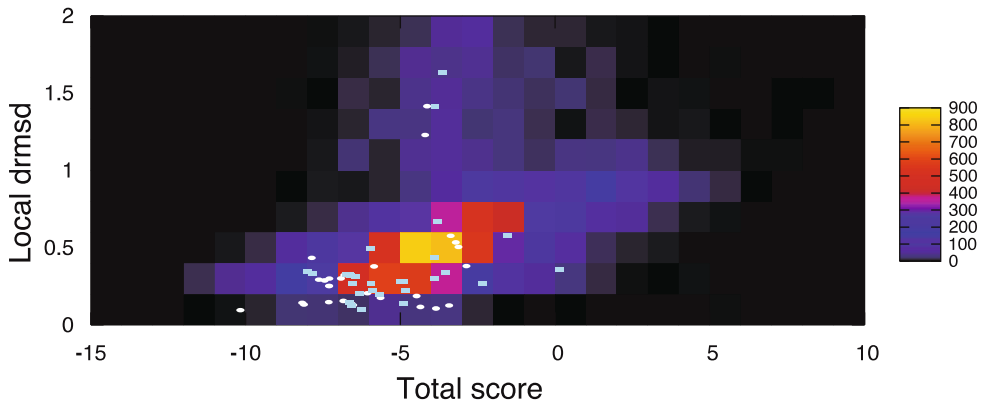

Fig. 3 Benchmark of the short-range scoring function. The diagram shows the two-dimensional histogram of local harmonic energy for structures submitted by all human predictors (x axis) and local drmsd (y axis) calculated for all models submitted during CASP6 experiment. The color scale describes the number of models of a given quality. Structures submitted for two groups: Baker and coworkers [29] and Kolinski and Bujnicki [30] in the "new fold" category are shown separately as white ellipses and blue rectangles, respectively

Acknowledgement This work was partially supported from NIH grant 1R01GM081680-01.

References

1. Plewczynska D, Kolinski A (2005) Macromol Theory Simul 14:444
2. Tramontano A (2003) Comp Funct Genomics 4:402
3. Pieper U, Eswar N, Braberg H, Madhusudhan MS, Davis FP, Stuart AC, Mirkovic N, Rossi A, Marti-Renom MA, Fiser A, Webb B, Greenblatt D, Huang CC, Ferrin TE, Sali A (2004) Nucleic Acids Res 32:D217
4. Kopp J, Schwede T (2004) Nucleic Acids Res 32:D230
5. Marti-Renom MA, Stuart AC, Fiser A, Sanchez R, Melo F, Sali A (2000) Annu Rev Biophys Biomol Struct 29:291
6. Goebel U, Sander C, Schneider R, Valencia A (1994) Proteins 18:309
7. Punta M, Rost B (2005) Bioinformatics 21:2960
8. Cuff JA, Clamp ME, Siddiqui AS, Finlay M, Barton GJ (1998) Bioinformatics 14:892
9. Jones DT (1999) J Mol Biol 292:195
10. Rost B (2001) J Struct Biol 134:204
11. de Brevern AG, Etchebest C, Hazout S (2000) Proteins 41:271
12. Fetrow JS, Palumbo MJ, Berg G (1997) Proteins 27:249
13. Bystroff C, Baker D (1998) J Mol Biol 281:565
14. Rohl CA, Strauss CE, Misura KM, Baker D (2004) Methods Enzymol 383:66
15. Gront D, Kolinski A (2005) Bioinformatics 21:981
16. Gront D, Kolinski A (2006) Bioinformatics 22:621
17. Gront D, Kolinski A (2008) Bioinformatics 24:584
18. Gribskov M, McLachlan AD, Eisenberg D (1987) Proc Natl Acad Sci USA 84:4355
19. Sadreyev R, Grishin N (2003) J Mol Biol 326:317
20. Ohlson T, Wallner B, Elofsson A (2004) Proteins 57:188
21. Prlic A, Domingues FS, Sippl MJ (2000) Protein Eng 13:545
22. Kolinski A (2004) Acta Biochim Pol 51:349
23. Dragalin V, Fedorov V, Patterson S, Jones B (2003) Stat Med 22:913
24. Davis JV, Dhillon I (2006) Adv Neural Inform Process Syst 19:89
25. Kmiecik S, Kurcinski M, Rutkowska A, Gront D, Kolinski A (2005) Acta Biochim Pol 53:131
26. Wang G, Dunbrack RLJ (2003) Bioinformatics 19:1589
27. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Nucleic Acids Res 25:3389
28. Kabsch W, Sander C (1983) Biopolymers 22:2577
29. Bradley P, Malmstrom L, Qian B, Schonbrun J, Chivian D, Kim DEE, Meiler J, Misura KMSM, Baker D (2005) Proteins 61:124
30. Kolinski A, Bujnicki JM (2005) Proteins 61:84
31. Moult J, Fidelis K, Rost B, Hubbard T, Tramontano A (2005) Proteins 61:3