BIOINFORMATICS DISCOVERY NOTE Vol. 23 no. 19 2007, pages 2522–2527
doi:10.1093/bioinformatics/btm380

Structural bioinformatics

Comparative modeling without implicit sequence alignments

Andrzej Kolinski and Dominik Gront*

University of Warsaw, Faculty of Chemistry, Pasteura 1 02-093 Warsaw, Poland

Received on June 6, 2007; revised on June 25, 2007; accepted on July 14, 2007

Advance Access publication July 27, 2007

Associate Editor: Burkhard Rost

ABSTRACT

Motivation: The number of known protein sequences is about thousand times larger than the number of experimentally solved 3D structures. For more than half of the protein sequences a close or distant structural analog could be identified. The key starting point in a classical comparative modeling is to generate the best possible sequence alignment with a template or templates. With decreasing sequence similarity, the number of errors in the alignments increases and these errors are the main causes of the decreasing accuracy of the molecular models generated. Here we propose a new approach to comparative modeling, which does not require the implicit alignment — the model building phase explores geometric, evolutionary and physical properties of a template (or templates).

Results: The proposed method requires prior identification of a template, although the initial sequence alignment is ignored. The model is built using a very efficient reduced representation search engine CABS to find the best possible superposition of the query protein onto the template represented as a 3D multi-featured scaffold. The criteria used include: sequence similarity, predicted secondary structure consistency, local geometric features and hydrophobicity profile. For more difficult cases, the new method qualitatively outperforms existing schemes of comparative modeling. The algorithm unifies *de novo* modeling, 3D threading and sequence-based methods. The main idea is general and could be easily combined with other efficient modeling tools as Rosetta, UNRES and others.

Contact: dgront@chem.uw.edu.pl

1 INTRODUCTION

Comparative modeling remains the best established and the most powerful method for theoretical prediction of protein structures. Classical approach starts from sequence alignments of a query protein with sequences of template proteins for which their structures have been already solved experimentally (Tramontano, 2003). Instead of a straightforward sequence approaches various threading procedures (Lathrop and Smith, 1996; Madej *et al*., 1995) could also be applied, which is important for distantly homologous or analogous templates. The templates could then be used to build a consensus scaffold, which then needs to be 'decorated' with loops and other ambiguous fragment of the target putative structure. Alternatively, a set of distance restraints could be read from the templates and a consensus model could be built by means of a best possible satisfaction of these restraints. Such approach is extremely successful in Modeller (Sali and Blundell, 1993), which employs sophisticated distance geometry and various procedures for loop modeling. Another class of tools employing distance restraints are based on stochastic search of query protein conformational space guided by the restraints from templates. Typical examples are CABS (Kolinski, 2004) and Rosetta (Rohl *et al*.., 2004). These two modeling tools have proven to be very successful during the CASP experiments (Kolinski and Bujnicki, 2005). What is important, these stochastic search methods are applicable to comparative modeling as well as to *de novo* (template free) prediction of not too complex structures. Other reduced representation modeling tools as TASSER (Zhang *et al*., 2005) and UNRES (Liwo *et al*., 1997, 2005) also rely on complex strategies of conformational search. With increasing evolutionary distance (and consequently with decreasing sequence similarity), the sequence alignments and the threading-based (although to a somewhat lesser extent) alignments become more and more ambiguous. The errors are of a various nature, frequently making the subsequent molecular modeling very difficult or impossible (Marti-Renom *et al*., 2000). The problem is that all the alignment techniques ignore various geometric limitations. Suppose that there is a gap in the alignment of the query sequence and the flanking residues in the template are separated by a larger distance than the length of a single peptide bond. In such situations, the straightforward connection of the template fragments is not feasible and the alignment needs to be adjusted somehow. A number of problems of such type arise with gapped alignments.

The new method proposed here avoids these alignment problems. The modeling is done directly onto a multi-featured scaffold and the protein target chain maintains its connectivity and protein-like geometry during the modeling procedure. The protein-like behavior is achieved due to the generic knowledge-based force field of the CABS (Kolinski, 2004) model. A template needs to be identified prior to the modeling. It could be done in various ways, including purely biological considerations. There is no need for the initial template-target alignment — it does not enter in any way into the input data. The multi-featured scaffold is built in the following fashion. First, the template reduced to the alpha-carbon trace is projected onto a fine 3D grid. Then, a set of selected properties

*To whom correspondence should be addressed.

is assigned to the nodes of the grid that are in proximities of particular alpha carbons of the template. These properties include: amino acid identity, assigned secondary structure (in the three letter code), amino acid hydrophobicity (the Kyte–Doolittle scale is used) and the local direction of the template coded by the alpha carbon–alpha carbon virtual bonds, using a discreet set of vectors. Obviously, other features of proteins could be used, although we found the above selection to be satisfactory. Also multiple templates could be used, provided a sensible structural alignment of these templates can be generated — the scaffolds need to be consistently oriented for the subsequent simulations. Here, for the sake of simplicity and clarity we limit ourselves to the single template case.

The conformational search on the template scaffolds is done by means of the Replica Exchange Monte Carlo (REMC) sampling (Geyer, 1991; Gront *et al*., 2000; Hansmann, 1997). The sampling is controlled by the CABS force field and guided by a scoring function describing the fit of the target structures to the template.

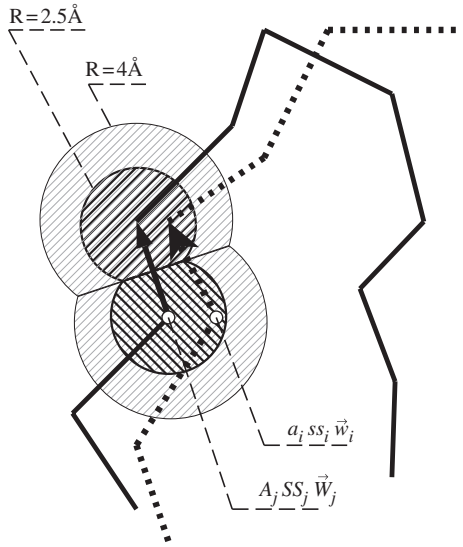

Fig. 1. Schematic illustration of the idea of a multi-featured scaffold for *de novo* comparative modeling of protein structures. The thick solid line represents the Cα-trace of a template. The spheres show the cutoff areas for comparison of various features of the template and the target. The 'running' target Cα-trace is shown in the thick dotted line. The two small open circles correspond to the alpha carbons being the reference centers for identification of the local properties of both structures: A — amino acid identity, SS — secondary structure (assignment for the template, and prediction for the target, respectively) and W — the local directions of the chains (see the text for more details).

2 METHODS

2.1 Definition of the multi-featured template scaffold

The alpha-carbon trace of the template chain is projected onto a fine lattice grid with the spacing of 0.61 Å and placed in the center of the Cartesian coordinate system. This is consistent with the CABS (Kolinski, 2004) representation, although different projections can be used, especially when alternative search engines are employed during the modeling phase. Then, to every point of the grid various features of the template are assigned, using two distinct cutoff distances. The secondary structure assignment (in the three-letter code: alpha, beta and coil) is done with 2.5 Å cutoff. This means that the grid points closest to an alpha carbon position and in the cutoff distance have assigned the secondary structure of the template's residue. Then, the separate arrays are used to store the template's residues identities and the local direction of the chain. Residues' identities are used to define the fitness of the sequences (according to a substitution matrix — BLOSUM62 has been used in the test presented in this work). Additionally, a hydrophobicity scale is used to smooth-out the alignment scoring landscape. We employed the Kyte–Doolittle scale (Kyte and Doolittle, 1982), because it appears to be a good consensus for the definition of the amino acids' hydrophobic properties. The same 4.0 Å cutoff is used for the spatial coding of these three properties of a template. The values of the cutoff distances were carefully adjusted using different targets, although we found out that the algorithm is not too sensitive to particular choices. It seems to be quite clear that the scaffold's role is to guide the topology assembly and perhaps to regularize a bit the secondary structure in the protein's core. The fine details of the target structure are defined by the CABS force field. Again, other search engines [Rosetta (Rohl *et al*., 2004), UNRES (Liwo *et al*., 1997), TASSER (Zhang *et al*., 2005)] can probably do an equally good job. Figure 1 explains in a schematic fashion the idea of the template scaffold.

2.2 Starting conformations of the target

In principle, there are three various ways to initiate the modeling. First, one can start from just a collection (replicas for REMC) of random coil chains, roughly imposed onto the template grid. This is the most straightforward, although rather expensive computational strategy. The chains need to adjust their orientation in respect to the template and fold into the proper structure. For complex folds it may take quite a time. Obviously, the cost of computations depends also on the quality (proximity to the target) of the template. A different way to generate the starting replicas is to take chains excised from the template in form of continuous fragments (with random starting points) supplemented with random tails at the N- or C-terminus, when necessary. This is a better strategy than the first one, since for particular replicas some fragments of the starting chains could be similar to the target structure. Computationally, the most effective is to build the starting models using other fast and easy to use modeling procedures (Metaservers). Then, the crude models needs to be structurally aligned onto the template scaffold. Even very crude models usually contain topologically correct structural fragments, what facilitates very fast convergence to the optimal structure of the target. In this work, we tested all three strategies and the final results were identical (in a range of statistical errors of the CABS representation), although significant differences in the computational costs were systematically observed.

2.3 Target to template fitness function

The scoring function for the REMC sampling consists of two parts. The first one is the generic force field of the CABS model. The CABS model has been described previously in detail (Kolinski, 2004), and used in various applications to protein structure prediction and study of protein dynamics (Kmiecik and Kolinski, 2007; Kmiecik *et al*., 2006) and thermodynamics (Gront and Kolinski, 2007a,b). Just for completeness, let us mention that the CABS interaction scheme contains a set of knowledge-based potentials derived from statistical analysis of the regularities seen in known protein structures. The potentials describe the short range conformational propensities, the context-dependent interactions between side groups, a model of excluded volume and a model of the main-chain hydrogen bonds network. The

second part of the scoring function quantifies the fitting of the modeled chain to the multi-featured template scaffolds (see Fig. 1). There are four components of this part of the score, outlined below.

2.3.1 A substitution matrix based scoring We employed the BLOSUM62 substitution matrix (Henikoff and Henikoff, 1992). It is implemented in a very simple way. When the glycine alpha carbon of the wiggling chain, representing the target protein, visits the 4 Å vicinity of a glycine of the template scaffold the score increases by 6, in the case of alanines imposed onto the glycine vicinity the score is 0, for serine it is equal to −3, etc. For not aligned residues the score is equal to −4, which corresponds to the minimal value of the substitution matrix's elements. Obviously, the problem of the gap penalty is irrelevant since the modeled chain maintains its connectivity.

2.3.2 Hydrophobicity fitness The Kyte–Doolittle scale is used (Kyte and Doolittle, 1982). The score for a single residue is equal to $\max(0, KD_{template} \times KD_{target})$.. Thus, for the ALA–ALA alignment the score is equal to 1.8×1.8, for HIS–HIS it is -3.2×-3.2, while for HIS–ALA pairing it is equal to 0. The ignoring of the negative products prevents a non-physical expansion of the model chains in cases of highly unfavorable local superimpositions. The distance cutoff is the same as in the previous case of the substitution matrix based scoring.

2.3.3 Local orientation of the template and the target chains The scoring is based on the dot products of the corresponding Cα–Cα virtual vectors. When they are directed in the same direction (product >0) the score is equal to 1, otherwise it is equal to 0. This criterion applies only to the pairs of residues having the same regular secondary structure (a helix or a strand) assignment (for the template) or prediction (for the target). The distance cutoff for the corresponding alpha carbons is again equal to 4 Å.

2.3.4 Secondary structure complementarities The alignment of a pair of residues having the same secondary structure (a helix, or a strand) is additionally awarded with the score equal to 1. In this case the cutoff distance was set to be equal to 2.5 Å. A smaller cutoff for this component of the scoring function facilitates more exact superimpositions of the regular secondary structure elements, presumably building a more conserved protein core.

2.4 Model building procedure

The total score is a linear combination of the above four subcomponents. The two first are taken with a weight equal to 0.25, while the weight factor for the two remaining components are equal to 1. The weighting is arbitrary, although carefully adjusted by a trial and error procedure. Fortunately, the method is not sensitive to even significant variation of the weighting scheme. Taken with the minus (−) sign, the target to template fitness scoring is added to the CABS energy, and used together to control the REMC sampling. Total 10–20 replicas were used with the temperatures equally distributed around the estimated collapse transition temperatures for the CABS models of the target chains. Couple of thousands of snapshots from the simulations was subject to a hierarchical clustering procedure (Gront and Kolinski, 2005). For the cluster centroids, the atomic details were reconstructed (Gront *et al.*, 2007) and the resulting structures were briefly energy minimized using an all-atom force field with an implicit solvent. The lowest energy structure of the target was taken as a final result. The details of such multiscale approach to the fold evaluation and selection could be found elsewhere (Kmiecik *et al.*, 2007). Since the results of the CABS simulations for the method presented here are very consistent, the clustering and the minimization play a marginal role — the improvements of the generated structures are usually incremental.

3 EXAMPLE RESULTS AND DISCUSSION

Three example systems represent various levels of difficulty for comparative modeling procedures. The systems are characterized in Table 1. A classical modeling was also performed for these systems. To make the test as difficult as possible, the alignments for the classical modeling using Modeller (Sali and Blundell, 1993) were optimized. The best possible alignments (leading to the lowest possible errors of the structural alignment of the targets and templates) were generated running the alignment program with a wide range of parameters. The following procedure was applied to build the best possible models:

At the first step, we calculated an optimal profile-to-profile alignment for a given target/template pair. For this purpose, we used PRAline tool from BioShell (Gront and Kolinski, 2006) package. During optimization, we were varying gap opening/continuation penalties and profile–profile scoring system. We tried all combinations of the four features: gap opening penalty (from −11.0 to −3.0 with step −1.0), gap continuation penalty (from −0.1 to −2.0 with step −0.1), profile–profile similarity scoring system: [dot-product, outer-product, COMPASS (Sadreyev and Grishin, 2003)] and alignment method (global or local). To calculate the outer products of two profile columns, we used BLOSUM62 matrix (Henikoff and Henikoff, 1992). Then we employed Modeller ver. 9.1 to build models from the best alignments. The best alignments are very close to the structural alignments (due to the optimization procedures). Thus, models very close to the best possible models are built. Each model was optimized in all-atom CHARMM force field (Brooks *et al.*, 1983). We tried also several methods of loops refinement that are implemented in Modeller, but in all cases the results were worse that without loops remodeling. In all cases, significantly better models were obtained with the method proposed in the present work. The results are summarized in Table 1. Other classical methods of comparative modeling lead to significantly worse models than these obtained with Modeller. Very interesting is the case of protein G modeling using 1ubq as a template. For this system,

Table 1. Summary of the modeled systems

Target/Template	2gb1/1ubq	2pcy/1paz	2pcy/2azaA
Target length	56	99	99
Template length	76	123	129
Optimization for classical modeling method	DP1	OPg	DPg
Gap penalty	11	5	6
Gap extension	1.5	0.2	0.4
Classical modeling cRMSD	8.20 Å	3.51 Å	5.81 Å
This method cRMSD			
via clustering	1.67 Å	2.86 Å	3.44 Å
the best structure observed	1.37 Å	2.55 Å	3.19 Å

Optimization for classical modeling: rows describe the parameters that resulted in the structurally optimal alignment: DPl = dot-product scoring with local alignment, OPg = outer-product global alignment, DPg dot-product global alignment. (Detailed definition of profile–profile scoring may be found in (Wang and Dunbrack, 2004.)

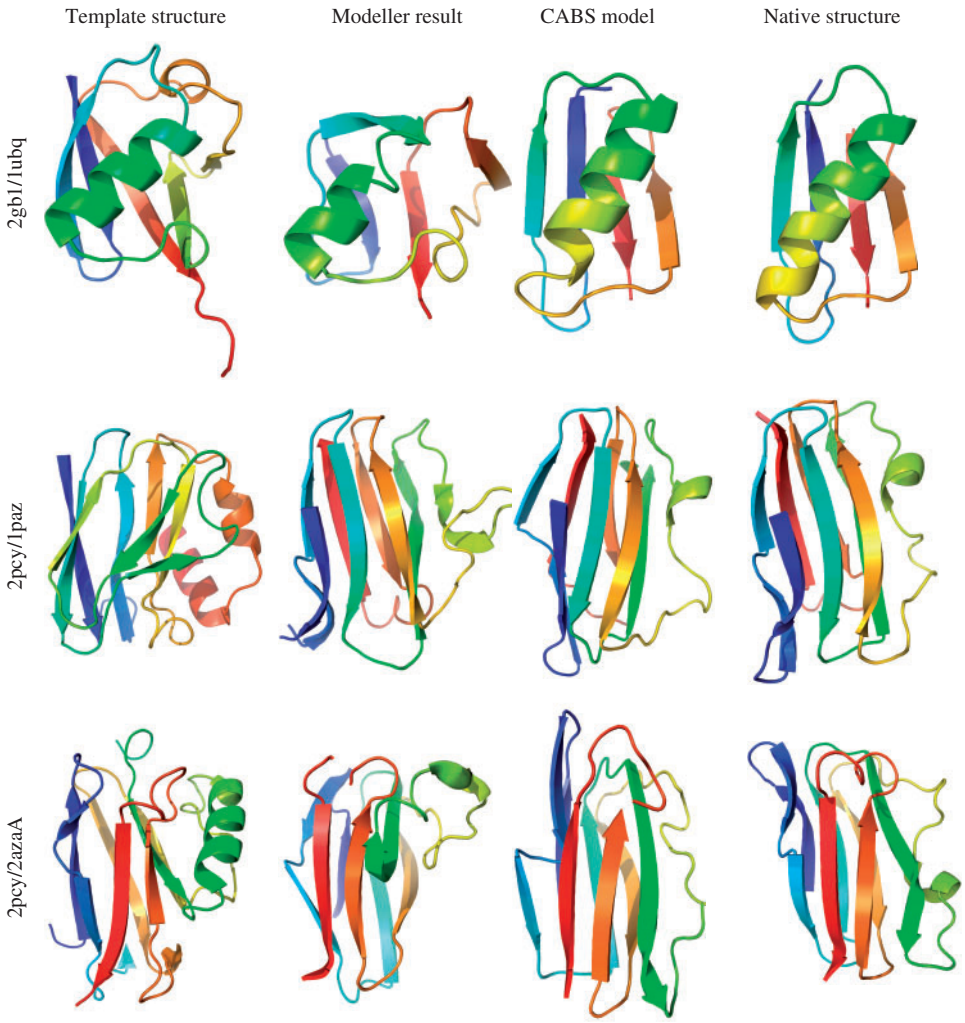

Fig. 2. Ribbon drawings of the example models. Top row: 2gb1 modeled on1ubq as template, middle row: 2pcy modeled on 1paz and bottom row: 2pcy on 2azaA. First column: the template structures. Second column: models obtained with Modeller. Third column: structures from the present method and Fourth column: native structures.

the sequence similarity between the target and template is at a random level, although some threading programs can identify possible remote structural similarity, though the obtained alignments lead to very bad models, barely resembling the target topology (but not its secondary structure). Yet the present method leads to a very good model. It builds on the structural similarity, and hydrophobic pattern similarity of the N-terminal fragment consisting of a helix and a β-hairpin. This fragment could be structurally aligned with a high fidelity. This fragment of the model is even better than the model resulting from the structural alignment. For the fairness, it should be mentioned that the 2gb1 domain can be folded by CABS in a pure *de novo* fashion. Such *de novo* simulations produce a number of alternative conformations, including the topological mirror image structures. Sophisticated model selection procedures are needed to select the proper fold from *de novo* trajectories. On contrary, the comparative modeling by the present method produces only the near-native structures, and the model selection procedures are used only to select the best model from the set of good models. In Figure 2 the modeled structures are compared with the native structures of the targets. The method proposed here has some features of our earlier methods for improvement of threading-based models. For example, the GeneComp algorithm (Kolinski *et al*., 2001) is based on a *de novo* modeling of a target using a 'tube' scaffold built on a template. The present method goes much further, and unifies a true 3D threading with *de novo* modeling. It is easy to see that the idea of the multi-featured template scaffold is general and can be (with possible modifications of the grid, properties coded in the scaffold, etc.) combined with other efficient conformational search tools as Rosetta (Rohl *et al*., 2004), UNRES (Liwo *et al*., 1997, 2005), SICHO (Kolinski and Skolnick, 1998) or TASSER (Zhang *et al*., 2005) — a close relative of CABS. Very importantly, the proposed method does

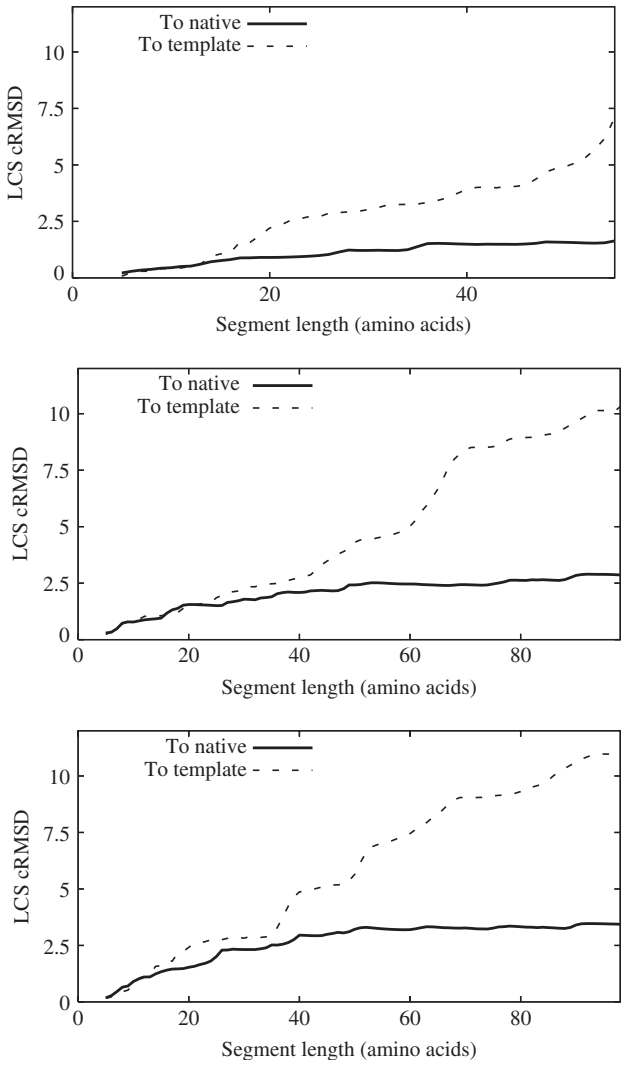

Fig. 3. Coordinate root-mean-square deviation (cRMSD — in Å) of the longest continuous segment (LCS) of the model after the best superimposition with the native structures (solid lines) and with the template structures (dashed lines) for the 2gb1/1ubq, 2pcy/1paz and 2pcy/2azaA systems from top to bottom, respectively. The plots clearly demonstrate that the obtained models are much closer to the target than to the templates used.

not need a starting sequence alignment. Consequently, one can focus just on the best template identification, disregarding the quality of the global alignment. 'Partial' templates could be also used. As a result, the fraction of proteins for which good molecular models could be predicted *in silico* (in an automated fashion) increases significantly. What is very important the obtained models are much better than the templates used — the models are much closer to the target structures, regardless of the alignment used. This is illustrated in Figure 3, where the distances from the target and template structures of the longest continuous segments of the models are plotted as a function of the segment length. Here, only three representative examples are shown for an illustration of the essence of the discovery. A massive benchmark using this method and other methods of protein structure prediction is now underway. We hope that the multi-featured template scaffold idea finds its applications in other labs using different modeling tools.

ACKNOWLEDGEMENT

The computational part of this work was done using the computer cluster at the Computing Center of the Faculty of Chemistry, University of Warsaw.

Conflict of Interest: none declared.

REFERENCES

Brooks,B.R. *et al.* (1983) CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.*, **4**, 187–217.

Geyer,C.J. (1991) Markov chain Monte Carlo maximum likelihood. In *Computing Science and Statistics: Proceedings of 23rd Symposium on the Interface Interface Foundation*. Fairfax Station, pp. 156–163.

Gront,D. and Kolinski,A. (2005) HCPM–program for hierarchical clustering of protein models. *Bioinformatics*, **21**, 3179–3180.

Gront,D. and Kolinski,A. (2006) BioShell – a package of tools for structural biology computations. *Bioinformatics*, **22**, 621–622.

Gront,D. and Kolinski,A. (2007a) Efficient scheme for optimization of parallel tempering Monte Carlo method. *J. Phys. Condens. Matter*, **19**, 036225–036234.

Gront,D. and Kolinski,A. (2007b) T-Pile – a package for thermodynamic calculations for biomolecules. *Bioinformatics*, doi:10.1093/bioinformatics/btm259.10.

Gront,D. *et al.* (2000) Comparison of three Monte Carlo conformational sea strategies for a proteinlike homopolymer model: folding thermodynamics and identification of low-energy structures. *J. Chem. Phys.*, **113**, 5065–5071.

Gront,D. *et al.* (2007) Backbone building from quadrilaterals: a fast and accurate algorithm for protein backbone reconstruction from alpha carbon coordinates. *J. Comput. Chem.*, **28**, 1593–1597.

Hansmann,U.H.E. (1997) Parallel tempering algorithm for conformational studies of biological molecules. *Chem. Phys. Lett.*, **281**, 140–150.

Henikoff,S. and Henikoff,J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA*, **89**, 10915–10919.

Kmiecik,S. and Kolinski,A. (2007) The characterization of protein folding pathways by reduced-space modeling. *Proc. Natl Acad. Sci. USA*, **104**, 12330–12335.

Kmiecik,S. *et al.* (2006) Denatured proteins and early folding intermediates simulated in a reduced conformational space. *Acta Biochim. Pol.*, **53**, 131–143.

Kmiecik,S. *et al.* (2007) Towards the high-resolution protein structure prediction. Fast refinement of reduced models with all-atom force field. *BMC Struct. Biol.*, **7**, 43.

Kolinski,A. (2004) Protein modeling and structure prediction with a reduced representation. *Acta Biochim. Pol.*, **51**, 349–371.

Kolinski,A. and Bujnicki,J.M. (2005) Generalized protein structure prediction based on combination of fold-recognition with de novo folding and evaluation of models. *Proteins*, **61**, 84–90.

Kolinski,A. and Skolnick,J. (1998) Assembly of protein structure from sparse experimental data: an efficient Monte Carlo model. *Proteins*, **32**, 475–494.

Kolinski,A. *et al.* (2001) Generalized comparative modeling (GENECOMP): a combination of sequence comparison, threading, and lattice modeling for protein structure prediction and refinement. *Proteins*, **44**, 133–149.

Kyte,J. and Doolittle,R.F. (1982) A simple method for displaying the hydropathic character of a protein. *J. Mol Biol.*, **157**, 105–132.

Lathrop,R.H. and Smith,T.F. (1996) Global optimum protein threading with gapped alignment and empirical pair score functions. *Mol J. Biol.*, **255**, 641–665.

Liwo,A. *et al.* (1997) 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. *J. Comput Chem.*, **18**, 849–873.

Liwo,A. *et al.* (2005) Ab initio simulations of protein-folding pathways by molecular dynamics with the united-residue model of polypeptide chains. *Proc. Natl Acad. Sci. USA*, **102**, 2362–2367.

Madej,T. *et al.* (1995) Threading a database of protein cores. *Proteins*, **23**, 356–369.

Marti-Renom,M.A. *et al.* (2000) Comparative protein structure modeling of genes and genomes. *Annu. Rev. Biophys. Biomol. Struct.*, **29**, 291–325.

Rohl,C.A. *et al.* (2004) Protein structure prediction using rosetta. *Methods Enzymol.*, **383**, 66–93.

Sadreyev,R. and Grishin,N. (2003) Compass: a tool for comparison of multiple protein alignments with assessment of statistical significance. *J. Mol. Biol.*, **326**, 317–336.

Sali,A. and Blundell,T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.*, **234**, 779–815.

Tramontano,A. (2003) Comparative modelling techniques: where are we? *Comp. Funct. Genomics*, **4**, 402–405.

Wang,G. and Dunbrack,R.L. (2004) Scoring profile-to-profile sequence alignments. *Protein Sci*, **13**, 1612–1626.

Zhang,Y. *et al.* (2005) Tasser: an automated method for the prediction of protein tertiary structures in casp6. *Proteins*, **61**, 91–98.