ACTA ABP BIOCHIMICA POLONICA

Vol. 57, No 1/2010
125–133
on-line at: www.actabp.pl

Regular paper

TRACER. A new approach to comparative modeling that combines threading with free-space conformational sampling

Sebastian Trojanowski, Aleksandra Rutkowska and Andrzej Kolinski✉

Laboratory of Theory of Biopolymers, Faculty of Chemistry, University of Warsaw, Warszawa, Poland

A new approach to comparative modeling of proteins, TRACER, is described and benchmarked against classical modeling procedures. The new method unifies true three-dimensional threading with coarse-grained sampling of query protein conformational space. The initial sequence alignment of a query protein with a template is not required, although a template needs to be somehow identified. The template is used as a multi-featured fuzzy three-dimensional scaffold. The conformational search for the query protein is guided by intrinsic force field of the coarse-grained modeling engine CABS and by compatibility with the template scaffold. During Replica Exchange Monte Carlo simulations the model chain representing the query protein finds the best possible structural alignment with the template chain, that also optimizes the intra-protein interactions as approximated by the knowledge based force field of CABS. The benchmark done for a representative set of query/template pairs of various degrees of sequence similarity showed that the new method allows meaningful comparative modeling also for the region of marginal, or non-existing, sequence similarity. Thus, the new approach significantly extends the applicability of comparative modeling.

Keywords: protein comparative modeling, coarse grained protein models, protein threading, Monte Carlo simulations, replica exchange Monte Carlo

Received: 20 January, 2010; revised: 10 March, 2010; accepted: 19 March, 2010; available on-line: 22 March, 2010

INTRODUCTION

Comparative modeling remains the most powerful and widely used methods for protein structure prediction. With the increasing number of experimentally solved protein structures the range of applicability of comparative modeling steadily increases, although with a slower pace than expected a few years ago. Still, 30–40% of newly sequenced proteins can not be modeled with an accuracy sufficient for practical applications, such as drug design, guiding of protein engineering or supporting X-ray or NMR structure determination. Thus, besides efforts in template-free, *de novo* modeling, it is important to develop new methods for comparative modeling, capable of generating reasonable models even for poor quality (remotely similar) structural templates.

Classical comparative modeling consists of two, to a large extent independent, tasks. The first is to find the best possible template, usually a homologous protein, and to build an as good as possible alignment of the query and template sequences. Most of contemporary comparative modeling methods can handle multiple templates, which usually leads to better models than does modeling based on a single template. Template structures, together with the assignments of the query protein residues to the template given in the alignments, are the source of spatial restraints for the assembly of the query protein structure. The assembly process, which constitutes the second main task of comparative modeling, may be based on several substantially different approaches. In the Modeller (Eswar *et al.,* 2008), a golden standard for comparative modeling, the restraints derived from templates are used to derive probability distributions for intra-protein (query) distances (Sali & Blundell, 1993). The model is constructed by means of distance geometry. In other methods the structure of a query protein is assembled from small protein fragments (as in Rosetta (Rohl *et al.,* 2004; Chivan & Baker, 2006) or folded from a random conformation (as in CABS (Koliński, 2004; Ekonomiuk *et al.,* 2005), using the distance restraints derived from templates.

In cases when a closely related template can be identified the alignment problem is relatively easy to solve. When sequence similarity is 50% or more the classical tools, such as PsiBlast (Altschul *et al.,* 1997), usually provide an error-free, optimal alignment. For low sequence the similarity alignments usually contain numerous errors. Threading (Rost *et al.,* 1997) or Fold Recognition, FR (Kosinski *et al.,* 2003) methods sometimes can detect remotely homologous and sometimes even evolutionarily unrelated but structurally analogous templates. This is possible due to tertiary information explored by FR algorithms — the alignments are scored not only by sequence similarity, but also by the three-dimensional context of the template structures (Koliński & Bujnicki, 2005). Unfortunately, true three-dimensional threading algorithms are of high complexity, and therefore computationally very expensive. Decreasing the level of similarity between the query and template structures leads not only to ambiguous alignments but also to differences in the geometry of the correctly aligned fragments. In

✉e-mail: kolinski@chem.uw.edu.pl

Abbreviations: cRMSD, coordinate root-mean-square deviation; DSSP, dictionary of secondary structure patterns; DOPE, discrete optimized protein energy; LCS, longest continuous segment; LGA_S, local-global alignment score; GDT, global distance test; GDT_TS, global distance test – total score; SCOP, structural classification of proteins

1knc, 2af7, 1cjx, 1sqi, 1qip, 1h5y, 1qo2, 1kk0, 1skq, 1gtd, 1ecm, 1umw, 1ijv, 1pcf, 1p8c, 2cwq, 2gmy, 1sqd, 1sp8, 1byl, 1qto, 1t47, 1jvn, 1h5y, 1jvn, 1ka9, 1thf, 1g7s, 1r5b, 1t3t, 1t4a, 2csm, 1mby, 1smy, 1l3a

other words, even for an optimal alignment the template may turn out to be of poor quality due to differences in the geometry of the scaffolds.

The new method for comparative modeling of difficult structures, TRACER, described in this work, unifies true three-dimensional threading with unrestricted sampling of protein conformational space (Koliński & Gront, 2007). A template needs to be somehow identified, not necessarily by sequence or FR methods, but also, for instance, by purely biochemical considerations. The template is represented as a loosely defined three-dimensional object, with multi-featured spatial properties. The query protein chain samples the vicinity of such defined template scaffold using the mesoscopic representation of the CABS (Koliński, 2004) protein modeling software. The method does not require prior sequence alignment. The alignment is built in parallel to the process of structure assembly. The idea of TRACER is general. It does not need necessarily to be implemented in the context of CABS technology. It would be quite easy to use a different sampling engine, for instance Rosetta.

The Methods section contains a detailed description of the TRACER idea and its implementation. The new method is evaluated (Results section) and compared with the Modeller. The test set of modeled proteins contains cases of very different levels of similarity between the query and template structures, and therefore represents a broad range of difficulty for comparative modeling.

METHODS

TRACER employs CABS representation of protein conformational space and its knowledge-based force field. CABS is a coarse-grained model where protein residues are represented by up to four pseudo atoms: Cα — alpha carbon (CA), Cβ — beta carbon (B), center of mass of the side chain (S), and a pseudo atom at the center of a virtual Cα–Cα bond. The Cα trace is restricted to a simple cubic lattice with the mesh size of 0.61 Å, and provides a reference frame for the definition of the remaining pseudo atoms, which are not restricted to the lattice. The lattice representation facilitates rapid generation of local conformational transitions of the model chain and efficient calculation of the system energy. The force field of CABS consists of statistical knowledge-based potentials describing short-range conformational propensities, geometric context-dependent (multibody) potentials of side chains' interactions, and a cooperative model of main the chain hydrogen bonds. Solvent is treated in a highly idealized implicit fashion. Details of CABS representation, stochastic dynam-

Table 1. Summary of the test set

Protein name	Protein code (length)	Templatecode (length)	Sequence identity	cRMSD [Å] of structural alignment (length)
Antioxidant defense protein (AhpD)	1knc(68)	1p8c(69)	22.2%	2.43(68)
Antioxidant defense protein (AhpD)	1knc(68)	2cwq(70)	23.2%	4.24(67)
Gamma-carboxymuconolactone decarboxylase (CMD)	2af7(70)	2gmy(68)	22.2%	6.02(67)
4-Hydroxyphenylpyruvate dioxygenase (HppD)	1cjx(88)	1sqd(92)	29.0%	8.52(84)
4-Hydroxyphenylpyruvate dioxygenase (HppD)	1sqi(89)	1sp8(92)	40.4%	7.02(87)
Glyoxalase I	1qip(72)	1byl(76)	10.9%	11.93(69)
Glyoxalase I	1qip(72)	1qto(60)	15.6%	12.09(60)
Glyoxalase I	1qip(72)	1t47(85)	19.5%	12.61(70)
Histidine biosynthetic protein (HisF)	1h5y(116)	1jvn(136)	15.5%	10.12(110)
Phosphoribosylformimino-5-aminoimidazole carboxamide ribotite isomerase (HisA)	1qo2(116)	1h5y(116)	7.8%	4.63(105)
Phosphoribosylformimino-5-amino-imidazole carboxamide ribotite isomerase (HisA)	1qo2(116)	1jvn(136)	13.8%	10.46(111)
Phosphoribosylformimino-5-amino-imidazole carboxamide ribotite isomerase (HisA)	1qo2(116)	1ka9(114)	28.9%	7.94(110)
Phosphoribosylformimino-5-amino-imidazole carboxamide ribotite isomerase (HisA)	1qo2(116)	1thf(114)	23.7%	7.88(109)
Initiation factor eIF2	1kk0(48)	1g7s(52)	12.5%	10.50(48)
Elongation factor eEF-1alpha	1skq(62)	1r5b(61)	29.5%	1.17(61)
PurS subunit of FGAM synthetase	1gtd(60)	1t3t(58)	10.3%	5.37(58)
PurS subunit of FGAM synthetase	1gtd(60)	1t4a(57)	21.7%	2.56(56)
Chorismate mutase	1ecm(79)	2csm(99)	16.5%	3.27(76)
Serine/threonine protein kinase Plk	1umw(71)	1mby(76)	25.4%	9.59(69)
RNA polymerase	1ijv(60)	1smy(60)	93.3%	2.18(60)
Transcriptional coactivator PC4	1pcf(58)	1l3a(69)	8.6%	14.69(58)

ics and force field are described elsewhere (Kolinski, 2004), and parameters of the force field are available from the authors' homepage. Also a commercial version of CABS software is available with full documentation. Numerous applications of the CABS modeling technology include: protein structure prediction (from comparative modeling to *de novo* modeling of new folds), modeling of protein complexes (Kurcinski & Koliński, 2007), and study of protein dynamics (including mechanisms of protein folding) and thermodynamics (Kmiecik & Koliński, 2007; 2008). An early idea of TRACER has already been presented elsewhere (Koliński & Gront, 2007) as a proof of principle. The method described in this work employs an updated and optimized scoring scheme for the Replica Exchange Monte Carlo, different sampling procedures and different representation of the template scaffolds, that together significantly extend the range of efficient applications. In particular, the scaling factors in the template-query similarity function (Eqn. 1) have been optimized basing on a large set of modeling instances. Also the present version of TRACER contains a new, very efficient, subroutine for the detection of spatial proximity of the query and template Cα vertices. Additionally, the Monte Carlo sampling scheme of TRACER employs a ten-fold larger fraction of the 4–22 fragment moves compared with the original CABS algorithm (Koliński, 2004). This update enables a faster search for plausible alignments of the query and template chains.

Template representation. TRACER requires a template for the modeling. The template can be identified by any bioinformatics method, or by purely biochemical/genetic considerations. Only Cα trace of the template is taken into consideration. The Cα trace is projected onto a simple cubic lattice with the spacing of 0.61 Å, consistent with the CABS representation. In the vicinity of the template trace a three-dimensional object is defined in such a way that to each point of the lattice assigned are the amino acid identity and its characteristics read from the template. The amino acid characteristics include: values of the Blosum62 substitution matrix (Henikoff & Henikoff, 1992), hydrophobicity according to Kyte–Doolittle (Kyte & Doolittle, 1982) scale, and secondary structure according to the three-letter DSSP assignment (Kabsch & Sander, 1983). An imprint of a residue is the cloud of lattice vertices which are closest to a template residue and within a certain cut-off distance, which, after careful calibration, has been set at 4 Å.

TRACER sampling scheme. The stochastic dynamics of the query chain is executed as a long sequence of local conformational transitions controlled by the CABS algorithm. Conformational updates include: two-bond kink motions of the Cα trace, three-bond motions, four-bond motions, and small random translations and "reptation-like" movements of longer (4–22 residues) fragments that do not break the chain connectivity. Sampling is executed according to the Replica Exchange Monte Carlo (REMC) scheme (Swendesen & Wang, 1986; Geyer, 1992; Hukushima & Nemoto, 1996; Hansmann, 1997) (REMC), where a number (20) of query protein chains are placed at various temperatures. The temperatures of the replicas are uniformly distributed and the stack of temperatures is gradually lowered during the simulations. Each replica is controlled by asymmetric Metropolis scheme, where the system energy is the sum of CABS energy E_{CABS} and the query-template similarity pseudo energy $E_{TEMPLATE}$. The latter can be written as (see Fig. 1 for reference):

$$E_{TEMPLATE} = \Sigma(0.25E_{subst} + 0.25E_{hp} + E_{sec} + E_{align}) \quad (1)$$

where the summation is done for the entire query protein chain, and:
$E_{subst} = -a_{i,j}$ for superposition of the template and query residues at distances smaller than 4 Å, $a_{i,j}$ – the value of BLOSUM62 substitution matrix
$E_{hp} = -\max(0, H_i \cdot H_j)$ for superposition of the template and query residues at distances smaller than 4 Å, H_i, H_j are values of the Kyte–Doolittle hydrophobicity indexes
$E_{sec} = -1$ for identical (helical or extended) secondary structures of the template (assigned) and the query (predicted), and superposition at distances smaller than 2.5 Å
$E_{align} = -1$ for both chains (template and query) having locally (in the vicinity of two residues superimposed at a distance smaller than 4 Å) similar orientations and directions (the angle between pairs of flanking Cα–Cα vectors smaller than 90 degrees).

The term measuring the amino acid similarity is typical for various FR methods. The hydrophobicity term has been included for two reasons. First, the Kyte–Doolittle scale is less specific and reflects the hydrophobic/hydrophilic patterns of protein sequences. Thus, the energy landscape is smoothened and hydrophobicity patterns (abstracted from the sequence patterns) influence the alignments of the query chains onto the template scaffold. Second, hydrophobicity profiles of globular proteins exhibit a relatively well-defined hydrophobic core surrounded by a more hydrophilic surface. Thus, the early stages of the TRACER sampling are to some extent driven by an overall shape of the template, and therefore the energy landscape becomes funnel-like, even at relatively large distance from the target structure. The term measuring the similarity of secondary structures (assigned according to DSSP for the template and predicted for the query sequence) is defined for a smaller distance cut-off than the three remaining terms. This reflects the higher structural conservation of the regular secondary/supersecondary structure elements in comparison with entire folds. Finally, the alignment term favors monotonic assignments of pairs of residues (query and template) along the both sequences.

In principle, the starting conformations of the query chains (replicas) could be generated in a random fashion and then placed in the center of gravity of the template scaffold. This is, however, not the best choice, since reorientation of the query chains and their folding from a random conformation within the CABS sampling scheme is computationally quite expensive. It is better to build the starting chains on the scaffold of the template using a crude threading procedure and a simple search algorithm for random conformations of missing loops. Then the TRACER runs are shorter, since the algorithms only does a search for the best superposition of the query chains onto the fuzzy scaffold of the template, satisfying intrachain interactions.

The conformational updates in TRACER simulations are the same as in CABS and include small local modifications of two, three and four pseudo-bonds of the Cα trace, with appropriate rearrangements of the side chains. Additionally, the CABS algorithm employs, although less frequently, rearrangements of larger fragments, 4–22 residues long. There are two types of these larger-scale modifications. The first one is a rigid body translation, provided the required changes of the valence angles and lengths of the end bonds of the affected segment permit the move. The second one is a "reptation-like" move, where the chain units slide along the Cα-trace, removing

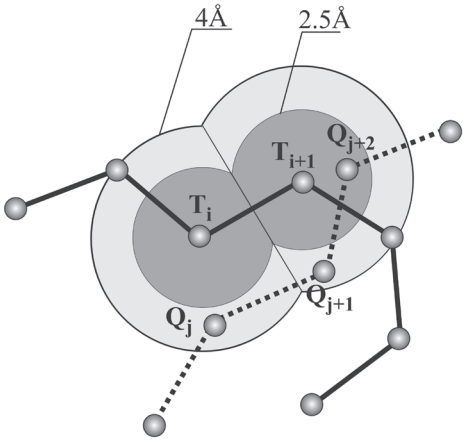

Figure 1. Explanation of TRACER idea.
Template scaffold is drawn in solid line and marked T, with spheres showing residue locations as seen by moving query chain (in dashed lines), marked Q. Pairs of residues (T_i, Q_j) and (T_{i+1}, Q_{j+1}) satisfy three of the four template/query compatibility criteria (see text) while (T_{i+1}, Q_{j+2}) satisfies all four criteria, including the secondary structure compatibility criterion. For clarity side chains are not shown.

a "wave" at one end of the sliding fragment and creating one at the opposite end. These intermediate distance moves facilitate rapid sampling of the template scaffold, frequently changing the "alignment" of the two chains. Obviously, all conformational updates are controlled by a pseudo random mechanism.

In the test modeling described here, the TRACER trajectories from relatively long runs (several hours on a single LINUX box) were clustered using K-means method and the medoids of the largest clusters (top clusters) were used for all-atom reconstruction and structural analysis. Also the best clusters, for which the cRMSD from the target structure was the lowest, were recorded.

The idea of TRACER is illustrated in Fig. 1.

Test set for modeling with TRACER

TRACER is designed for "difficult" cases of comparative modeling, where the sequence similarity between the query protein and the template is low, although it is also important to see how the proposed method performs for easier cases. For these reasons, we selected the SISYPHUS (Andreeva *et al.,* 2007) database, which contains structural alignments of proteins or protein fragments taken from the SCOP (Murzin *et al.,* 1995) classification scheme. The test set used in this work is summarized in Table 1. Pairs of the test proteins (query/template) are small single-domain structures (or their fragments) with very different degrees of sequence similarity.

RESULTS AND DISCUSSION

The results of test simulations are summarized in Table 2, where various measures of the model quality are given for 21 modeling experiments. The second column of Table 2 gives values of coordinate root-mean square deviation (cRMSD) of the models after the best superposition with the target structure. The third column contains values of GDT_TS, which is frequently used for benchmarking modeling procedures. GDT (Zemla *et al.,* 1992) measures fractions of residues that could be aligned with a certain cRMSD cut-off. Additionally, the Table (columns 4–6) contains numbers of residues in the longest continuous segments (LCS) (Zemla *et al.,* 1999) of the models that could be superimposed with the target structure with an accuracy of 1 Å, 2 Å and 5 Å, respectively. Finally, the last column contains LGA_S (Zemla, 2003; Zemla *et al.,* 2005), a combination of LCS and GDT measures. The lines abbreviated TRACER (top) correspond to the top (largest) clusters' medoid from the clustering of the trajectories from TRACER simulations. Lines abbreviated TRACER (best) contain data for the best clusters (with the rank of the best cluster given in brackets). The rank of the best clusters are given in parentheses. The results of modeling with TRACER (without prior sequence alignment) is compared with classical modeling, starting from a sequence alignment. The latter was done using Modeller, version 9v5, a standard for comparative modeling.

Global *versus* local quality of models

A crude comparison of the results obtained with the new modeling method with those of classical modeling shows that models from TRACER are on average more accurate globally. Of the 21 models generated by TRACER, 14 have lower cRMSD when compared with the target structures. When local accuracy of the models, as measured by LCS, is compared, an opposite tendency is observed. Namely, only in 8 of the 21 cases the LCS (1 Å) is better or the same for TRACER, when compared with Modeller. Values of LCS (2 Å) are on average very similar for both methods — in 10 of the 21 cases the TRACER models are better or the same than for Modeller. When the LCS (5 Å) are considered, the results for TRACER are much better — 19 per 21 models have longer continuous segments that can be aligned with the target assuming a 5 Å tolerance. Thus, the classical models from Modeller are on average more accurate locally, while TRACER yields models that have a better overall structure. This is illustrated in Fig. 2 and Fig. 3, showing snapshots of best superpositions of the models with the target structures and GDT plots for two selected, typical, cases. In the GDT plots the cRMSD cut-offs of the best segments of the models are plotted against the length of the segment, usually given in percentage of the total length of the modeled protein. Analysis of the results of modeling illustrated in Fig. 2 and Fig. 3 explains the profoundly different performance of the two methods. In one case (Fig. 2) the initial alignment used in the classical method was qualitatively correct. This led to a very good quality of the model built by Modeller, as quantified by the GDT plot showing that the model is actually better than the template used along the entire length of the query protein. The model from TRACER was globally correct, although with lower local accuracy — the GDT plots show that the model is of the same accuracy as the template used. A different situations is shown in Fig. 2. Here the sequence alignment employed by Modeller was incomplete and the resulting model contained qualitatively wrong fragments. Due to its three-dimensional scoring of query/template superpositions TRACER found an optimal alignment and produced globally correct model. Thanks to intrinsic force field of CABS the resulting model was much closer to the target structure than to the template used, as quantified in the GDT plot. These examples (and similar results observed in other cases) illustrate the qualitative differ-

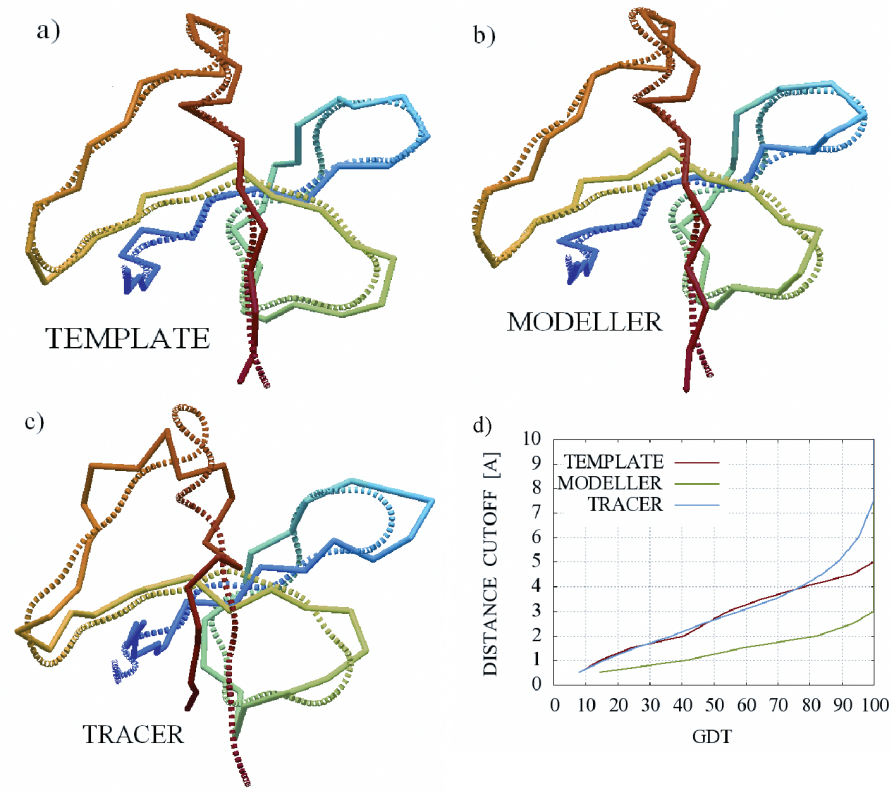

Figure 2. Results of modeling of 1skq structure using 1r5b as template.
From top: **(a)** superposition of template (solid Cα trace) with target structure (smoothened dashed Cα trace), **(b)** superposition of Modeller model with target structure, **(c)** superposition of TRACER model with target structure, and **(d)** GDT plot for template, Modeller and TRACER structures. Overall accuracy of Modeller and TRACER models measured by cRMSD after superposition with the target structure are 1.63 Å and 4.07 Å, respectively.

ences between the TRACER and classical methods of comparative modeling.

Effect of sequence identity on modeling accuracy

The accuracy of comparative modeling depends on sequence similarity between the query and template proteins. In Fig. 4 the overall accuracy of the models built by TRACER (the top ranking models) and Modeller is plotted against the percentage of sequence identity. Figure 4 clearly shows that modeling by TRACER, as measured by overall similarity of the target and query structures, is superior in the region of very low sequence identity, while for "easier" cases the performances of the classical scheme and TRACER are similar, with a clearly better performance of Modeller for

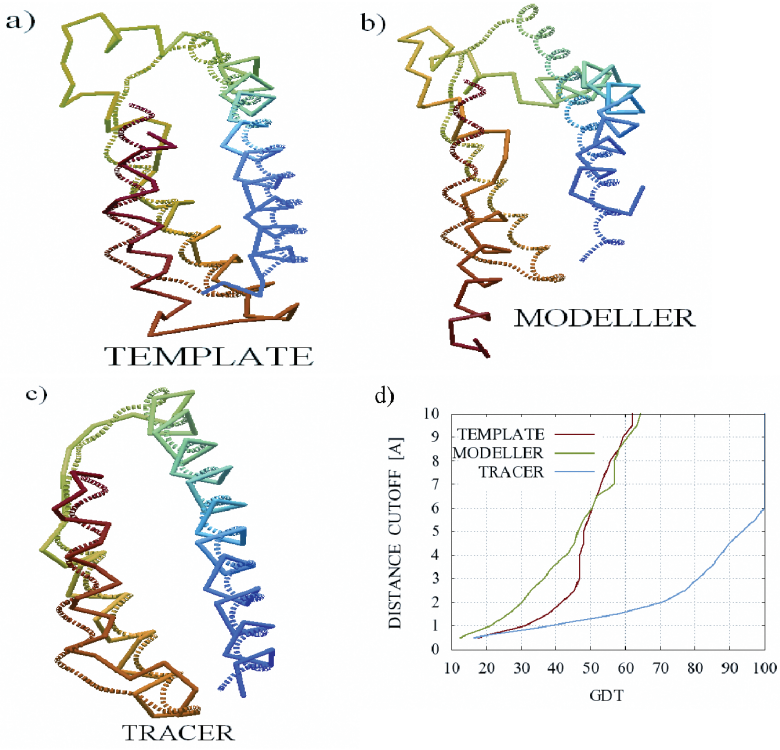

Figure 3. Results of modeling of 1ecm structure using 2cms as template.
From top: **(a)** superposition of template (solid Cα trace) with target structure (smoothened dashed Cα trace), **(b)** superposition of Modeller model with target structure, **(c)** superposition of TRACER model with target structure, and **(d)** GDT plot for template, Modeller and TRACER structures. Overall accuracy of Modeller and TRACER models measured by cRMSD after superposition with target structure are 13.81 Å and 2.82 Å, respectively.

Table 2. Summary of modeling results

Method	RMSD	GDT_TS	LCS(1A)	LCS(2A)	LCS(5A)	LGA_S
	1knc / 1p8c		(target / template)			
MODELLER	2.50	74.3	19	38	68	71.7
TRACER(top)	2.98	74.3	26	61	68	81.2
TRACER(best) (9)	2.11	76.5	20	67	68	86.9
			1knc / 2cwq			
MODELLER	5.87	41.9	14	15	45	35.5
TRACER(top)	6.15	42.6	14	15	44	34.2
TRACER(best) (9)	3.10	76.5	33	55	65	79.5
			2af7 / 2gmy			
MODELLER	4.91	70.0	41	54	70	71.4
TRACER(top)	4.04	63.2	18	51	70	67.9
TRACER(best) (11)	3.82	67.4	30	52	70	70.2
			1cjx / 1sqd			
MODELLER	3.94	64.5	39	50	88	61.1
TRACER(top)	3.88	63.4	20	30	88	57.9
TRACER(best) (9)	3.22	64.8	22	46	88	62.8
			1sqi / 1sp8			
MODELLER	2.80	78.7	40	42	89	70.4
TRACER(top)	2.15	75.0	17	61	89	78.6
TRACER(best) (10)	1.96	78.4	22	89	89	89.0
			1qip / 1byl			
MODELLER	12.26	36.8	13	23	29	34.3
TRACER(top)	9.40	38.9	13	16	54	35.6
TRACER(best) (13)	7.46	45.5	13	19	56	41.1
			1qip / 1qto			
MODELLER	12.55	38.2	15	23	29	34.6
TRACER(top)	9.94	52.8	11	21	49	48.1
TRACER(best) (12)	7.95	59.0	12	42	51	56.9
			1qip / 1t47			
MODELLER	12.19	20.8	8	9	19	18.2
TRACER(top)	10.02	55.9	10	20	49	50.8
TRACER(best) (1)	10.02	55.9	10	20	49	50.8
			1h5y / 1jvn			
MODELLER	9.45	48.9	23	25	72	40.6
TRACER(top)	4.75	55.6	14	28	116	50.0
TRACER(best) (5)	4.19	55.8	15	18	116	48.2
			1qo2 / 1h5y			
MODELLER	4.67	57.3	22	44	116	50.9
TRACER(top)	3.60	62.1	26	41	116	56.8
TRACER(best) (1)	3.60	62.1	26	41	116	56.8

1qo2 / 1jvn						
MODELLER	10.29	45.0	23	25	65	37.9
TRACER(top)	7.46	54.7	17	34	94	50.3
TRACER(best) (1)	7.46	54.7	17	34	94	50.3
1qo2 / 1ka9						
MODELLER	5.46	73.9	43	61	94	67.7
TRACER(top)	5.51	64.7	14	42	96	58.3
TRACER(best) (1)	5.51	64.7	14	42	96	58.3
1qo2 / 1thf						
MODELLER	5.34	72.8	43	60	94	66.5
TRACER(top)	5.62	58.2	16	37	95	53.3
TRACER(best) (1)	5.62	58.2	16	37	95	53.3
1kk0 / 1g7s						
MODELLER	10.15	33.3	11	12	17	28.8
TRACER(top)	8.11	37.0	7	10	27	31.9
TRACER(best) (3)	5.56	51.1	8	12	47	45.4
1skq / 1r5b						
MODELLER	1.63	81.0	26	62	62	90.7
TRACER(top)	4.07	52.2	10	19	60	47.3
TRACER(best) (11)	4.00	57.3	7	20	62	53.6
1gtd / 1t3t						
MODELLER	4.07	56.7	13	16	60	51.1
TRACER(top)	3.43	65.4	15	28	60	62.5
TRACER(best) (1)	3.43	65.4	15	28	60	62.5
1gtd / 1t4a						
MODELLER	3.51	67.5	20	36	60	67.8
TRACER(top)	3.73	65.4	18	37	60	66.2
TRACER(best) (1)	3.73	65.4	18	37	60	66.2
1ecm / 2csm						
MODELLER	13.81	38.0	21	29	45	36.0
TRACER(top)	2.82	74.4	33	62	79	76.9
TRACER(best) (10)	2.38	75.3	33	61	79	77.9
1umw / 1mby						
MODELLER	6.05	59.5	16	35	64	54.2
TRACER(top)	4.59	62.7	14	39	71	60.0
TRACER(best) (9)	2.74	70.1	14	50	71	72.2
1ijv / 1smy						
MODELLER	2.24	83.3	18	39	60	78.7
TRACER(top)	4.20	65.8	17	22	60	58.6
TRACER(best) (5)	3.54	67.9	17	22	60	59.4
1pcf / 1l3a						
MODELLER	9.27	45.7	15	20	40	38.3
TRACER(top)	5.87	48.7	16	18	48	46.0
TRACER(best) (7)	4.75	54.7	14	21	58	50.2

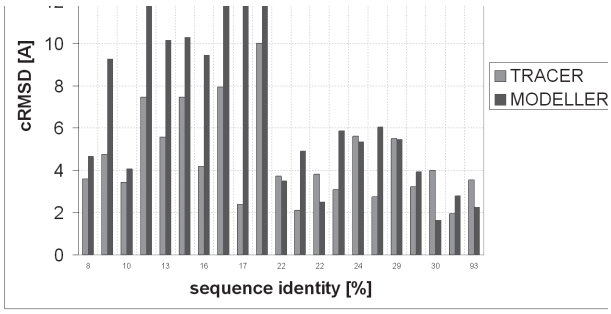

Figure 4. Comparison of modeling accuracy by TRACER and Modeller as measured by cRMSD from target structure, for various degrees of sequence similarities of query/template pairs.

the cases of very high sequence similarity. The effect of sequence similarity on the quality of the resulting models for both types of methods is further illustrated in Fig. 5 and Fig. 6, where values of LCS (1 Å) and LCS (5 Å) plotted as a function of sequence identity. The length of very accurate continuous segments (LCS 1 Å) ranges between 10% and 60% of the total length of the query proteins. In the region of low sequence similarity both methods give very good segments of similar length, typically between 10% and 20% of the chain length. In the range of higher sequence similarity the results of classical modeling with Modeller, as measured by LCS 1 Å, are on average much better. When the longest continuous segments are measured with lower accuracy (LCS 5 Å), a different picture emerges. TRACER produces complete, or almost complete, models of such defined low accuracy in all cases, outperforming the classical method in the range of low sequence similarity (below 22%). Such low resolution models have qualitatively correct patterns of side chain interactions (contacts) and therefore could be quite useful in identifying the type of a protein's function and in guiding site-directed mutation experiments.

The above analysis leads to a very simple practical conclusion: TRACER should be always employed in cases of low sequence similarity, where in almost all cases it outperforms the classical method. In the region of higher sequence similarity the models generated by TRACER are on average slightly less accurate, although there are numerous exceptions from this general rule.

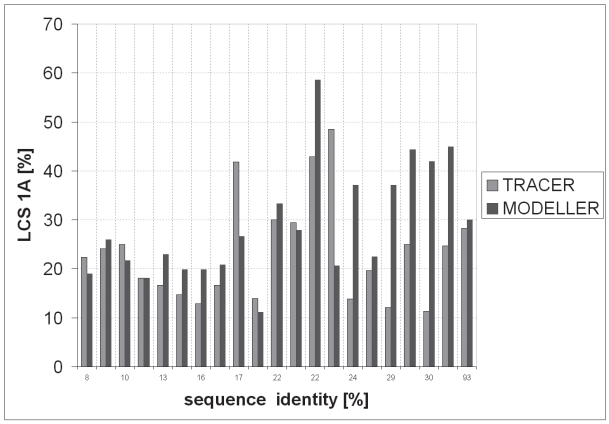

Figure 5. Comparison of modeling accuracy by TRACER and Modeller as measured by LCS (1 Å), for various sequence degrees of similarities of query/template pairs.

Model ranking and selection

Selection of the top model from TRACER simulations was done by means of cluster analysis. As could be seen from Table 2, in many cases the top model, the medoid from the largest cluster, was actually the best. There are, however, cases when lower-rank clusters provide better models. We checked the performance of different model ranking procedures. Unfortunately, none of them, including an all-atom reconstruction of the clusters' medoids followed by scoring with DOPE statistical potentials (Shen & Sali, 2006) gave clearly better results. Selection of the best models from coarse-grained simulations and their refinement is still a challenging problem of computational protein structure prediction. This is true not only for *de novo*, template-free, methods but also for various hierarchical comparative modeling procedures.

A very interesting observation comes from an analysis of the width of the distribution of the clusters' medoids (or centroids), as a measure of variance of the pairwise cRMSD distances between the clusters. For the 21 test cases studied, this width of the distribution of the simulation results varied from approx. 1.5 Å to 5–10 Å. In all the cases (except one) where this distribution was broad (variance larger than 4 Å) the models generated by TRACER were significantly better than those obtained by classical modeling with Modeller. Thus, there is a means to make quite a dependable choice between the two methods. It is actually not surprising. Sampling of a broad spectrum of models by TRACER indicates ambiguity of alignments, and consequently a higher probability that modeling based on prior alignment, without taking a good account of the tertiary structure context, would lead to less accurate models. In such cases TRACER, which is actually a search engine for true three-dimensional threading with a fuzzy template scaffold, often leads to results that are not accessible for any other combination of FR and modeling.

CONCLUSIONS

The new method for protein structure prediction, TRACER, described in this work unifies three-dimensional threading (taking an explicit account of tertiary interactions) with an efficient open-space conformational sampling. The method needs a template for modeling. The template needs not necessarily to be identified by

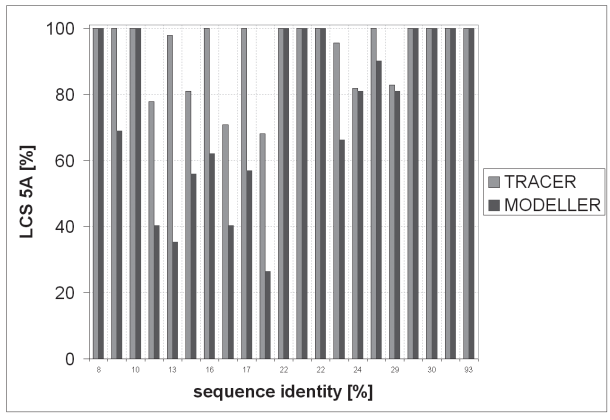

Figure 6. Comparison of modeling accuracy by TRACER and Modeller as measured by LCS (5 Å), for various degrees of sequence similarities of query/template pairs.

any sequence alignment or Fold Recognition method, it could be also defined by other (e.g., biochemical) considerations. TRACER does not use prior sequence alignment, although a crude alignment could be used to build starting structures for simulations. A proper alignment is built simultaneously with the search of the query protein conformational space.

The benchmark set of query/template pairs used in this work contains folds of various structural classes and represents a broad range of sequence similarity within pairs. The results of modeling with TRACER were compared with those from classical comparative modeling, where one starts from a sequence alignment of the query protein with the template, or templates. For the latter we selected a golden standard for comparative modeling, the Modeller. The comparison showed that TRACER significantly extends the range of applications of comparative modeling, allowing building of meaningful molecular models for very weak templates with marginal, or nonexisting, sequence similarity to the query proteins.

Future work will be aimed at achieving better means of selection of the best models from the TRACER pseudo-trajectories, and more effective methods of model refinement. These task remains to be very challenging for almost all protein prediction methods that are based on coarse-grained protein representations. It is also very important to design a multi-template version of TRACER. This is, however, significantly more complex than implementation of multi-template schemes driven by simple sets of distance restraints.

Acknowledgements

This work was supported by the NIH grant No 1R01GM081680 and the Ministry of Science and Higher Education (Poland), grant No NN301465634.

Computational part of this work was done using the computer cluster at the Computing Center of the Faculty of Chemistry, University of Warsaw. A commercial version of CABS-based modeling software was used http://www.selvita.com/selvita-protein-modeling-platform.html

REFERENCES

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25:** 3389–3402.

Andreeva A, Prlic A, Hubbard TJ, Murzin AG (2007) SISYPHUS — structural alignments for proteins with non-trivial relationships. *Nucleic Acids Res* **35** (Database issue)**:** D253–D259.

Chivian D, Baker D (2006) Homology modeling using parametric alignment ensemble generation with consensus and energy-based model selection. *Nucleic Acids Res* **34:** e112.

Ekonomiuk D, Kielbasinski M, Kolinski A (2005) Protein modeling with reduced representation: statistical potentials and protein folding mechanism. *Acta Biochim Pol* **52:** 741–758.

Eswar N, Eramian D, Webb B, Shen MY, Sali A (2008) Protein structure modeling with MODELLER. *Methods Mol Biol* **426:** 145–159.

Geyer JC (1992) Practical Markov Chain Monte Carlo. *Stat Sci* **7:** 473–483.

Hansmann UHE (1997) Parallel tempering algorithm for conformational studies of biological molecules. *Chem Phys Lett* **281:** 140–150.

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

Hukushima K, Nemoto K (1996) Exchange Monte Carlo Method and Application to Spin Glass Simulations. *J Phys Soc Jpn* **65:** 1604–1608.

Kabsch W, Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22:** 2577–2637.

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

Kmiecik S, Kolinski A (2008) Folding pathway of the b1 domain of protein G explored by multiscale modeling. *Biophys J* **94:** 726–736.

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

Kolinski A, Bujnicki JM (2005) Generalized protein structure prediction based on combination of fold-recognition with de novo folding and evaluation of models. *Proteins* **61** (Suppl 7)**:** 84–90.

Kolinski A, Gront D (2007) Comparative modeling without implicit sequence alignments. *Bioinformatics* **23:** 2522–2527.

Kosinski J, Cymerman IA, Feder M, Kurowski MA, Sasin JM, Bujnicki JM (2003) A "FRankenstein's monster" approach to comparative modeling: merging the finest fragments of Fold-Recognition models and iterative model refinement aided by 3D structure evaluation. *Proteins* **53** (Suppl 6)**:** 369–379.

Kurcinski M, Kolinski A (2007) Steps towards flexible docking: modeling of three-dimensional structures of the nuclear receptors bound with peptide ligands mimicking co-activators' sequences. *J Steroid Biochem Mol Biol* **103:** 357–360.

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

Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* **247:** 536–540.

Rohl CA, Strauss CE, Chivian D, Baker D (2004) Modeling structurally variable regions in homologous proteins with Rosetta. *Proteins* **55:** 656–677.

Rost B, Schneider R, Sander C (1997) Protein fold recognition by prediction-based threading. *J Mol Biol* **270:** 471–480.

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

Shen MY, Sali A (2006) Statistical potential for assessment and prediction of protein structures. *Protein Sci* **15:** 2507–2524.

Swendsen RH, Wang JS (1986) Replica Monte Carlo Simulation of Spin Glasses. *Phys Rev Lett* **57:** 2607–2609.

Zemla A (2003) LGA: A method for finding 3D similarities in protein structures. *Nucleic Acids Res* **31:** 3370–3374.

Zemla A, Venclovas C, Moult J, Fidelis K (1999) Processing and analysis of CASP3 protein structure predictions. *Proteins* (Suppl 3)**:** 22–29.

Zemla A, Zhou CE, Slezak T, Kuczmarski T, Rama D, Torres C, Sawicka D, Barsky D (2005) AS2TS system for protein structure modeling and analysis. *Nucleic Acids Res* **33** (Web Server issue)**:** W111–115.