J Mol Model (2006) 12: 125–139
DOI 10.1007/s00894-005-0003-0

ORIGINAL PAPER

Christer Levefelt · Dan Lundh

A fold-recognition approach to loop modeling

Received: 29 December 2004 / Accepted: 22 June 2005 / Published online: 8 November 2005

Abstract A novel approach is proposed for modeling loop regions in proteins. In this approach, a prerequisite sequence-structure alignment is examined for regions where the target sequence is not covered by the structural template. These regions, extended with a number of residues from adjacent stem regions, are submitted to fold recognition. The alignments produced by fold recognition are integrated into the initial alignment to create an alignment between the target sequence and several structures, where gaps in the main structural template are covered by local structural templates. This one-to-many (1: *N*) alignment is used to create a protein model by existing protein-modeling techniques. Several alternative approaches were evaluated using a set of ten proteins. One approach was selected and evaluated using another set of 31 proteins. The most promising result was for gap regions not located at the C-terminus or N-terminus of a protein, where the method produced an average RMSD 12% lower than the loop modeling provided with the program MODELLER. This improvement is shown to be statistically significant.

Keywords Protein structure prediction · Loop modeling · Fold recognition · Threading · Structurally variable regions

Introduction

A protein's function is enabled by its three-dimensional structure. The determination of protein structures from sequences of amino acids is essential for our understanding of the processes of life and is critical to many important areas such as drug design. Many methods for the prediction of protein structure align the target amino-acid sequence to a structural template derived from a known protein. The conformation of alignment regions where the target sequence is not covered by the template structure must be determined by an alternative approach, termed loop modeling. Together with alignment errors, loop modeling is a major limitation of protein-structure prediction methods [1].

Traditionally, computational methods for protein-structure prediction are divided into three categories [2]: (i) comparative or homology modeling, (ii) fold recognition or threading, and (iii) new fold methods or ab initio methods.

Comparative or homology modeling [3] predicts the structure primarily based on similarity between the sequence of the target protein and those of one or more template proteins of known structure. The comparative modeling approach is used in tools like MODELLER [1, 3, 4] and SWISS-MODEL [5].

Fold recognition or threading [6] is based on the observation that a large percentage of proteins adopt one of a limited number of folds. The result of a fold-recognition method is a ranking of the folds in a fold library according to the "goodness of fit" of the respective alignments, with the best fitting fold considered the most probable match. When a fold has been selected, the alignment can be passed to an automatic comparative modeling program for modeling loops and side chains, creating a complete three-dimensional model. Fold-recognition approaches are used in tools like THREADER [7–9], GenTHREADER [10, 11] 3D-PSSM [12–14], and LOOPP [15–17].

New fold methods [2, 18], are intended to construct structural models for a protein sequence without direct relationship to a known structure.

Alignments between the target sequence and a template structure are used in both comparative modeling and fold recognition. These approaches have shown to be useful for deriving an initial model. However, improvements are still necessary to overcome missing structural-template regions in the alignment. This is achieved by loop modeling.

C. Levefelt · D. Lundh (✉)
School of Humanities and Informatics, University of Skövde,
Box 408, 54128 Skövde, Sweden
E-mail: dan@ida.his.se
Tel.: +46-500-448315
Fax: +46-500-448399

The word "loop" is surrounded by some terminological confusion. At least two different meanings are applied to the term. According to van Vlijmen and Karplus [19], loops are segments that do not correspond to α-helical or β-strand secondary-structure elements. Moult [18] defines loops as regions, typically occurring between secondary-structure elements, where there are insertions and deletions in the target sequence relative to that of the template(s), or a local loss of sequence similarity. The presence of loops prevents these regions of the backbone being copied usefully from the template structure. The term "structurally variable region" (SVR) is used by Rohl et al. [20] for gaps, insertions, and regions of low-confidence alignment. This is similar to Moult's loop definition and the term is better suited for loops in the context of a sequence-structure alignment, as it is not burdened by any alternate meaning. When referring to loops and loop modeling, the definition of Moult [18] will be used in this paper. We will focus on the specific case where the target sequence is aligned to a gap in the template structure. This will be referred to as a "gap region." Nothing, however, prevents the method presented here from being used in the more general context of SVRs.

Loops are functionally important since they often contribute to binding sites [1]. Consequently, the impact of an accurate loop-modeling method would be great. The existing methods are reasonably accurate for modeling short loop regions, but modeling longer structurally divergent regions is an unsolved problem [20]. Fiser et al. [1] noted that in the first two Critical Assessment of Protein Structure Predictions (CASPs) [21, 22] there was no reliable method available for constructing loops longer than five residues, but that recently progress had been made. For example, van Vlijmen and Karplus [19] suggested an algorithm for loops of nine residues or less. Rohl et al. [20] presented a promising method for prediction of longer SVRs.

Loop-modeling methods can be grouped into [20] (i) knowledge-based methods, (ii) de novo or ab initio strategies, and (iii) combined approaches. Knowledge-based methods use known protein structures as a source of loop conformations. Likely conformations are generally selected based on evaluation using a knowledge-based potential or rule-based filter, evaluating criteria such as geometric fit and sequence similarity. In de novo strategies, loop conformations are generated by methods such as molecular dynamics, simulated annealing, exhaustive enumeration or heuristic sampling of a discrete set of ($\phi 1, \psi$) angles, random tweak, or analytical methods. Combined approaches are hybrids using both knowledge-based and de novo methods. The suggested method outlined in this paper shares similarity to knowledge-based methods in that the method uses knowledge of existing structures. However, the sequence-similarity concept traditionally used was expanded to include fold similarity by sequence threading.

The work presented in this paper expands the search for loop templates by fold recognition. Fold recognition is traditionally applied to entire protein chains, finding global folds. This work, however, was based on the hypothesis that the conformations of local folds could be analogous to local folds occurring in other known proteins. This allows a fold-recognition method to be applied locally to determine the conformation of regions in a sequence-structure alignment that are not covered by the main template structure.

Such an approach was used by Svensson et al. [23] in their prediction of the protein structure of a putative gene, flowering regulating factor (FRF), in the *Arabidopsis thaliana* genome. Fold recognition using the THREADER tool assigned a fold that had two long gaps in the alignment, 10 and 23 residues long, to the sequence. Fold recognition was applied to these sequences, assigning to them each an individual fold with the motivation that this constrained the loop regions in a state closer to the native energy state. In their study, the MODELLER tool was then used to build a model with the three folds as templates. Verification of models was performed using Ramachandran plots [24] and the software PROCHECK [25]. The best model had an energy profile similar to that of the template and no high-energy regions could be detected. The apparently high quality of the model indicated promise for a fold-recognition approach to loop modeling. Predictions of a local-domain structure can also be seen as related to the approach outlined in this paper in that a short fragment of sequence is searched for a local fold. Such an approach has been applied successfully assigning, e.g., the functionality and domain structure of the RB38 protein [26].

The aim of the work presented here was to investigate whether the fold recognition applied to gap regions could improve the quality of protein models. The basic principle underlying a fold-recognition approach to gap regions is described in Fig. 1. An initial sequence-structure alignment is used as input. Sequence regions that are not aligned to the template structure are extracted. These sequence segments are extended with residues from adjacent stem regions to facilitate subsequent modeling. The extended sequence segments are submitted to fold recognition and the alignments obtained are integrated into the initial alignment to create a set of sequence-to-structure equivalences, in this work called a one-to-many (1: *N*) alignment because of the suggested equivalence between one sequence to several structural templates (in different regions). This alignment can be used as input to a conventional comparative modeling tool to create a protein model. Results indicate that, for gap regions not located at the C-terminus or N-terminus of a chain, this is a promising approach.

Materials and methods

Several possible approaches to modeling gap regions by fold recognition were identified. Two sets of protein sequences were prepared for evaluation of these ap-

proaches. The first set consisted of ten protein sequences and was used to test all combinations of the identified approaches exhaustively. Based on this evaluation, a method for modeling gap regions by fold recognition was proposed. This method was applied to a second set of 31 protein sequences and the results were analyzed. The first protein set was named ''training set'' and the second ''test set,'' analogous to terms used in artificial intelligence.

The GenTHREADER tool was selected for the fold-recognition tasks in this work on the basis that it is a representative tool that has been used successfully in a number of fold-recognition assignments. It is very fast and reliable [10] and its neural network provides a combined quality measure.

To predict the gap regions, three different approaches were suggested for ranking local alignments: (i) by solvation energy, (ii) by alignment score, and (iii) by the GenTHREADER neural-network score [10]. Loops are usually located at the surface of a protein [1], corresponding to low solvation energy. This makes a ranking approach that favors low solvation energy a promising approach. A ranking favoring a high alignment score would lead to a fragment-based homology modeling approach. The score generated by the GenTHREADER neural network is based on several aspects of alignment quality including solvation energy, pairwise energy, and alignment score, possibly allowing for a more balanced ranking than relying on any single aspect of alignment quality. All of these features are obtained in the GenTHREADER output results. The proteins in the training set were explored using these features, identifying the best combination. In addition to these features, the length of the stem region, also called the anchor region, was explored. According to Martì-Renom et al. [3] the conformation of a given segment of a polypeptide chain must be calculated mainly from the sequence of the segment itself. However, they note that loops are generally too short to provide sufficient information about their local fold, and thus the conformation of a given segment is also influenced by the core stem regions that span the loop and by the structure of the rest of the protein that cradles the loop. The influence of stem regions could be accounted for by including additional residues on either side of the loop region in the sequence submitted to fold recognition. Influences from the rest of the protein structure are unfortunately not as easy to incorporate in a fold-recognition approach and were not taken into account. To determine the influence of stem regions on the results of fold recognition, three different approaches were suggested for generating the sequence to submit to fold recognition: one using no stem overlap and the other two using a stem overlap of three and ten residues. This allowed for testing the influence of stem regions on loop conformation.

The selection of protein sets was made based on the following requirements. First, given an alignment between each target sequence and a template structure, each alignment should contain at least one gap region of ten or more amino acids. This restriction was chosen since traditional loop-modeling methods are already capable of accurate prediction of loops up to nine residues in length [19]. Furthermore, fold recognition is not suited for short sequences, i.e., a sequence of one residue in length can be threaded onto any structure. Second, the proteins should have experimentally determined structures available. These were needed for the evaluation of protein models. Third, the protein sets should be sufficiently large to draw conclusions with reasonable confidence. In related studies van Vlijmen and Karplus [19] used two protein sets of 13 and 8 proteins, defined by Leszczynski and Rose [27] and Tramontano and Lesk [28], respectively. Based on the sizes of these sets, the minimum size of each set was set to ten proteins. Fourth, the protein sets should be representative of prediction targets known to have been used previously in protein-structure prediction.

Given these requirements, the prediction targets [29, 30] used in CASP4 and CASP5 were selected (see Table 1). In order to find additional native structures not known during the publication of the CASP targets, a FASTA [31] search was carried out for each sequence against the Protein Data Bank (PDB) [32], using PDB SearchFields (http://www.rcsb.org/pdb/cgi/queryForm.cgi?Fasta = 1) with the default scoring matrix. For many of the targets, more than one structure was found. The search also revealed that ''native'' structures did not always have a 100% sequence identity to the CASP target sequences. The sequence identity between CASP target sequence and PDB structure was generally either above 90 or below 70%. Structures with a sequence identity above 85% were treated as ''native.'' The lowest sequence identity of a ''native'' structure was 87% and the highest sequence identity for a ''non-native'' structure 70%. The sequence identity of the best alignment for each target is shown in Table 1. For proteins where several PDB entries were identified as native, one structure was selected based on *E*-value and resolution. One specific chain in each structure was selected for model evaluation. PDB files were downloaded from the online version of the PDB (http://www.rcsb.org/pdb/) during June and July 2004. All structures used are present in PDB release 200F. Of the original CASP targets, 110 in total, 15 were removed since no native structure was found.

Following the method outlined in Fig. 1, fold recognition was initially applied to the entire sequence, followed by local template fold recognition of the gap regions. The PDB files searched also included files not present at the time of the release of the CASP4/5 sequences. This dilutes the competitive issue as present in CASP.

The reasons for applying the initial fold recognition, instead of integrating loop conformations into the native PDB structure with deleted loop regions, were the following. First, as the core of the protein structure affects the loop configuration, we assume a worst-case scenario of using little or no knowledge of the native fold. This

Table 1 CASP targets considered for inclusion in the protein sets

Target[a]	Length[b]	Native structure[c]	Native seq. ID (%)[d]	Additional native structures[e]	Template[f]	Template seq. ID (%)[g]	Gap regions of length >9[h]
T0086	164	1G1B:A	100.0	1FW9, 1G81, 1JD3	1UAE	20.7	0
T0087[j]	310	1I74:A	98.4		1IR6:A	16.5	2
T0088	156	1OIO:A	100.0	1O9W, 1O9V, 1O9Z	1GQ8:A	16.7	0
T0089[k]	419	1E4F:T	100.0	1E4G	1BA1	15.3	3
T0090[k]	209	1G0S:A	100.0	1G9Q, 1GA7, 1KHZ, 1VIQ	1VIU:A	28.3	2
T0091	109	1PUG:A	87.2	1J8B	1MOJ:A	15.6	0
T0092	241	1IM8:A	97.5		1XVA:A	13.3	0
T0093[j]	160	1MXI:A	100.0	1J85	1IPA:A	21.9	1
T0094	181	1JH6:A	100.0	1JH7, 1FSI	1REC	11.0	0
T0095	244	1H6G:A	97.9	1L7C	1VHN:A	13.1	0
T0096[k]	239	1HW1:A	100.0	1E2X, 1H9G, 1H9T, 1HW2	1J5Y:A	13.4	2
T0097	105	1G7D:A	100.0		1K6K:A	12.4	0
T0098	121	1FC3:A	100.0	1LQ1	1VI0:A	9.9	0
T0099	56	None	n/a		n/a	n/a	n/a
T0100[k]	342	1QJV:A	100.0		1GQ8:A	31.7	2
T0101[k]	400	1RU4:A	100.0		1RMG	9.8	1
T0102	70	1O82:A	100.0	1E68, 1O83, 1O84	1KV8:A	7.1	0
T0103	372	1GA6:A	100.0	1GA1, 1GA4, 1KDV, 1KDY, 1KDZ, 1KE1, 1KE2, 1NLU	1SIO:A	30.3	0
T0104[k]	158	1HTW:A	100.0	1FL9	1RZ3:A	12.7	2
T0105	94	1H5P:A	100.0		1OQJ:A	26.7	0
T0106	128	1IJX:A	100.0		1GVF:A	14.8	0
T0107	188	1I82:A	99.5	1I8A, 1I8U	1ATG	6.4	0
T0108	206	1J84:A	98.9	1J83	1QEX:A	20.4	0
T0109	182	1J9A:A	97.8		1UOC:A	13.2	0
T0110[k]	128	1JOS:A	100.0		1PA4:A	17.7	1
T0111	431	1E9I:A	100.0		4ENL	50.1	0
T0112	352	1E3J:A	100.0		1LLU:A	24.0	0
T0113	261	1E3W:B	100.0	1E3S, 1E6W, 1SO8	1H5Q:A	23.1	0
T0114	87	1GH5:A	100.0	1G6E	1QCS:A	9.2	0
T0115[k]	300	1H72:C	100.0	1FWK, 1FWL, 1H73, 1H74	1S4E:B	19.7	1
T0116[k]	811	1NNE:A	100.0	1EWQ, 1EWR, 1FW6	1TAQ	15.5	6
T0117[k]	250	1OT3:A	100.0	1J90, 1OE0	1QHI:A	13.6	1
T0118[k]	149	1M0D:A	100.0	1FZR, 1M0I	1KNY:A	26.2	1
T0119	338	1KRH:A	100.0		1CQX:A	19.8	0
T0120[j]	336	1IK9:A	98.1	1FU1	1O5Z:A	11.3	2
T0121[k]	372	1G29:1	98.7		1B0U:A	26.7	1
T0122[k]	248	1GEQ:A	100.0		2TYS:A	31.0	1
T0123	160	1EXS:A	100.0		1BEB:A	65.4	0
T0124[k]	242	1JAD:A	97.1		1CUN:A	16.0	2
T0125	141	1GAK:A	100.0		1LIS	16.0	0
T0126	163	1JOB:A	100.0	1F35, 1JOD, 1JYT	1CBY	24.5	0
T0127[j]	350	1G8P:A	100.0		1FNN:A	11.1	2
T0128[k]	222	1P7G:A	98.6		1AVM:A	49.3	1
T0129	182	1IZM:A	98.9		1AOX:A	13.2	0
T0130[j]	114	1NO5:A	100.0		1J0L:A	28.6	1
T0131	100	None	n/a		n/a	n/a	n/a
T0132[k]	154	1NNG:A	100.0		1NJK:A	14.3	1
T0133	312	None	n/a		n/a	n/a	n/a
T0134	251	None	n/a		n/a	n/a	n/a
T0135	108	None	n/a		n/a	n/a	n/a
T0136	523	1ON3:A	100.0	1ON9	1UYR:A	17.4	0
T0137	133	1O8V:A	99.2		1MDC	22.5	0
T0138	135	1M2E:A	100.0	1M2F, 1R8J	1PEY:A	18.5	0
T0139	83	1IYR:A	100.0	1KOY	1S3J:A	18.1	0
T0140	103	1MJC	100.0	3MEF	1LCL	15.5	0
T0141[k]	187	1J3G:A	100.0		1LBA	23.3	2
T0142	282	1NZH:A	100.0	1NTF	1I9Y:A	24.5	0
T0143	216	1WCZ:A	99.5	1QY6	1P3C:A	22.3	0
T0144	172	None	n/a		n/a	n/a	n/a
T0145	216	None	n/a		n/a	n/a	n/a
T0146[k]	325	1NRK:A	97.2		1PJ5:A	11.7	1
T0147	245	1M65:A	100.0	1M68, 1PB0	1J6O:A	13.9	0
T0148	163	1IN0:A	100.0		1DD5:A	15.3	0
T0149[k]	318	1NIJ:A	100.0		1O5Z:A	9.1	4

Table 1 (Contd.)

Target[a]	Length[b]	Native structure[c]	Native seq. ID (%)[d]	Additional native structures[e]	Template[f]	Template seq. ID (%)[g]	Gap regions of length >9[h]
T0150	102	1H7M:A	100.0	1GO0, 1GO1	1CK9:A	34.0	0
T0151[k]	164	1UE1:A	100.0	1UE5, 1UE6, 1UE7	1QVC:A	29.0	1
T0152	210	None	n/a		n/a	n/a	n/a
T0153[j]	154	1MQ7:A	100.0	1SIX, 1SJN, 1SLH, 1SM8, 1SMC, 1SNF	1DUP:A	31.6	1
T0154[k]	309	1MOP:A	99.3	1N2B, 1N2E, 1N2G, 1N2H, 1N2I, 1N2J, 1N2O	1IHO:A	42.9	2
T0155[k]	133	1NBU:A	100.0		1DHN	33.1	1
T0156	157	1NXJ:A	99.4		1VI4:A	45.9	0
T0157	138	1NMN:A	100.0	1NU0, 1OVQ	1VHX:A	32.6	0
T0158	319	None	n/a		n/a	n/a	n/a
T0159	309	1R9L:A	99.4	1R9Q	4MBP	12.6	0
T0160	128	None	n/a		n/a	n/a	n/a
T0161[k]	156	1MW5:A	98.1		1QSP:A	15.4	1
T0162[k]	286	1IZN:A	100.0		1WER	12.2	1
T0163	369	1NG4:A	99.7	1NG3	1L9F:A	19.0	0
T0164[k]	166	1IO0:A	100.0		1UW4:B	13.9	1
T0165[k]	318	1ODT:C	99.4	1L7A, 1ODS	1EVQ:A	17.9	2
T0166[k]	150	1LJ9:A	100.0		1S3J:A	16.1	1
T0167	185	1M3S:A	100.0	1VIV	1JEO:A	35.6	0
T0168[j]	327	1MKI:A	96.3		1BTL	11.8	3
T0169	156	1MK4:A	100.0		1GHE:B	16.0	0
T0170	69	1UZC:A	98.6		1J7N:A	8.7	0
T0171	256	1M33:A	97.3		1MT3:A	15.6	0
T0172[k]	299	1M6Y:A	98.7	1N2X	1B74:A	11.5	3
T0173	303	1Q74:A	100.0	1Q7T	1UAE	17.8	0
T0174[j]	417	1MG7:A	100.0		1UAE	12.2	8
T0175	248	1NKV:A	97.6		1KPH:A	10.1	0
T0176	100	1N91:A	100.0		1BX4:A	8.0	0
T0177	240	1MW7:A	100.0		1LFP:A	32.1	0
T0178	219	1MZH:A	100.0		1JCL:B	26.9	0
T0179	276	1IY9:A	100.0		1JQ3:C	43.1	0
T0180	53	None	n/a		n/a	n/a	n/a
T0181	111	1NYN:A	100.0		1KC6:A	26.1	0
T0182	250	1O0X:A	100.0		1QXW:A	33.3	0
T0183[j]	248	1O0Y:A	100.0		1MZH:A	41.3	1
T0184[k]	240	1O0W:A	100.0		1TGO:A	17.5	1
T0185[j]	457	1J6U:A	98.5		1GQQ:A	29.2	2
T0186	364	1O12:A	97.8		1GKP:A	15.1	0
T0187[k]	417	1O0U:A	100.0		1UAE	16.8	4
T0188	124	1O13:A	99.2		1EO1:A	27.4	0
T0189	319	1O14:A	100.0		1RKD	15.4	0
T0190	114	None	n/a		n/a	n/a	n/a
T0191[k]	282	1NVT:A	100.0		1NPD:A	33.0	1
T0192	171	None	n/a		n/a	n/a	n/a
T0193	211	1R72:A	100.0		1R9L:A	12.3	0
T0194	237	None	n/a		n/a	n/a	n/a
T0195	299	None	n/a		n/a	n/a	n/a

[a] CASP target ID
[b] Number of residues in target sequence
[c] Structure and chain identified by FASTA and used as reference for model evaluation
[d] Sequence identity between the CASP target sequence and the native structure identified by FASTA
[e] Additional structures considered as native and thus not used as templates
[f] Template structure identified by GenTHREADER and used in initial alignment
[g] Sequence identity between the CASP target sequence and the template structure identified by GenTHREADER
[h] Number of gap regions longer than nine residues occurring in the initial alignment generated by GenTHREADER between the CASP target sequence and the template structure
[j] Target included in the training set
[k] Target included in the test set

prevents prediction bias from the true fold. Second, initial threading reveals gap regions not necessarily corresponding to true loops, while deleted loop regions in a PDB structure would restrain the gap regions. The gap regions may include terminal amino acids in a secondary structure or include additional elements. Initial threading truly reflects the situations that may occur in a real prediction situation. Third, using unrestrained gap

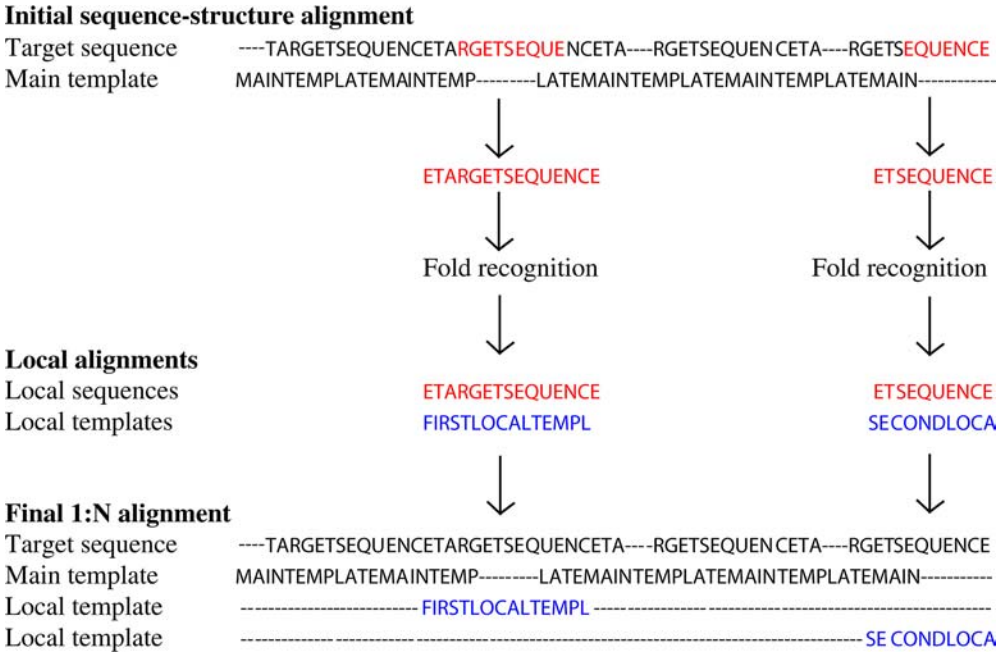

Fig. 1 The proposed approach to loop modeling by fold recognition

regions allows the local fold to influence the core fold in modeling, which could be considered as more appropriate due to the content in the gap.

Treating native protein structures as unknown allows the method to be implemented as an automatic service. However, the implication for this in the case of multi-domain proteins was that the best scored alignment for any of the domains was used and evaluated, omitting the other domains. Thus, for multi-domain proteins, models were only built and evaluated for one domain. Furthermore, fold recognition cannot always find a correct template, i.e., a false positive fold may be introduced. Templates with low structural similarity to the native fold, such as for target T0091; template 1MOJ, and for target T0161; template 1QSP, will influence the measured performance negatively in this study since the entire structure RMSD (see below) was measured.

For each sequence with a known native structure, the initial sequence-structure alignment was created using the fold recognition web service GenTHREADER at the PSIPRED Protein Structure Prediction Server [33]. Filtering options were left at their default settings (masking of low-complexity regions). The best scoring alignment for each sequence was selected unless that alignment was made to one of the native or alternative native structures for that sequence, in which case the next best alignment would be chosen. The reasons for ignoring native structures were to simulate fold recognition of a typical sequence, for which no native structure would be available and to produce alignments containing gap regions. Each alignment was examined for regions where the target sequence was aligned to a gap in the template structure. Fifty-four target sequences without such gap regions of at least ten residues were removed. The final number of targets was 41, each containing at least one gap region of at least ten residues in length. These were divided randomly into a training set of ten proteins (see Table 2) and a test set of 31 proteins (see Table 3).

All gap-region sequences in the training set were extracted and submitted to GenTHREADER. All of GenTHREADER's filtering options were disabled. Three sequences were submitted for each gap region, one for each stem-overlap approach. From the results returned by GenTHREADER, the different ranking approaches were used to create three alternative local alignments for each sequence. Alignments to native structures were ignored. The local alignments of the gap-region sequences were integrated into the initial sequence-structure alignment. This way, 1: *N* alignments were created consisting of the target sequence, the main template structure and one local template structure for each gap region. The geometric transformation for integrating the local template structure of the gap region into the main template was enabled using MODELLER spatial restraints. That is, MODELLER works by satisfying the restraints provided by the template structures (sequence-structure equivalences; the 1: *N* alignment), i.e., both from the main template and local template. Unsatisfied restraints will create a poor model in this way. Such an instance will be revealed by the evaluation (see below). In total, nine such 1: *N* alignments were created from each initial alignment, one for each combination of stem overlap and alignment ranking. The MODELLER tool (version 6v2) was used to build models for all ten alignments for each target (the initial alignment and the nine 1: *N* alignments). The newly released MODELLER version 8v0 adopts the same technique as the version used here, i.e., it is applicable for the framework performed in this study. Five models were built from each alignment. The models differed from each other as a result of MODELLER's default modeling protocol, i.e., a uniform randomization of the

Table 2 Gap regions in training set

Target[a]	Region[b]	Length[c]	Missing residues[d]
T0087	62–71	10	0
T0087	152–167	16	0
T0093	149–160[e]	12	4
T0120	205–215	11	4
T0120	283–336[e]	54	54
T0127	95–112	18	0
T0127	245–271	27	0
T0130	102–114[e]	13	9
T0153	131–154[e]	24	19
T0168	1–32[e]	32	8
T0168	131–146	16	0
T0168	153–165	13	0
T0174	27–36	10	8
T0174	75–89	15	0
T0174	198–211	14	0
T0174	228–250	23	0
T0174	272–281	10	0
T0174	303–314	12	0
T0174	353–372	20	0
T0174	403–417[e]	15	15
T0183	1–27[e]	27	0
T0185	130–145	16	1
T0185	445–457[e]	13	11

[a] CASP target ID
[b] Residue sequence numbers included in gap region
[c] Length of gap region
[d] The number of residues from the target sequence gap region whose positions were not available in the native structure PDB file
[e] Terminal gap region

Table 3 Gap regions in test set

Target[a]	Region[b]	Length[c]	Missing residues[d]
T0089	107–121	15	0
T0089	357–369	13	0
T0089	394–419[e]	26	26
T0090	1–13[e]	13	0
T0090	150–159	10	5
T0096	3–12	10	2
T0096	203–239[e]	37	9
T0100	115–128	14	0
T0100	249–262	14	0
T0101	168–197	30	0
T0104	62–73	12	0
T0104	130–139	10	0
T0110	107–128[e]	22	22
T0115	1–10[e]	10	4
T0116	55–71	17	0
T0116	549–563	15	0
T0116	571–580	10	0
T0116	606–621	16	0
T0116	673–683	11	0
T0116	687–811[e]	125	46
T0117	1–16[e]	16	11
T0118	140–149[e]	10	4
T0121	248–372[e]	125	0
T0122	165–177	13	7
T0124	1–15[e]	15	1
T0124	227–242[e]	16	2
T0128	1–15[e]	15	11
T0132	1–15[e]	15	10
T0141	1–20[e]	20	0
T0141	48–58	11	0
T0146	310–325[e]	16	1
T0149	37–53	17	0
T0149	81–90	10	0
T0149	114–125	12	0
T0149	287–296	10	0
T0151	144–164[e]	21	21
T0154	1–10[e]	10	2
T0154	288–309[e]	22	19
T0155	121–133[e]	13	13
T0161	147–156[e]	10	1
T0162	246–255	10	0
T0164	72–82	11	0
T0165	30–40	11	0
T0165	92–103	12	0
T0166	139–150[e]	12	5
T0172	105–123	19	0
T0172	235–244	10	0
T0172	257–299[e]	43	5
T0184	14–24	11	0
T0187	48–59	12	0
T0187	84–95	12	0
T0187	253–281	29	0
T0187	289–304	16	0
T0191	158–167	10	0

[a] CASP target ID
[b] Residue sequence numbers included in gap region
[c] Length of gap region
[d] The number of residues from the target sequence gap region whose positions were not available in the native structure PDB file
[e] Terminal gap region

Cartesian coordinates before terminating energy minimization to rectify bad stereochemistry (200 cycles of molecular dynamics; default in MODELLER). Here we used a randomization of 4 Å as recommended in the MODELLER manual, which typically results in models after energy minimization within 1 Å of each other. That is, the default parameters of MODELLER were used, with a random seed of −12312. The number of models created for each target sequence, i.e., five, generated an approximation of models in or close to the energy minimum given the sequence to template equivalences. That is, in order to reduce the effect of outliers, the evaluation was based on the approximation over five models. For the purpose of this study we generated 50 models for the training set and 145 models for the test set.

Models built using initial alignments containing gap regions (initial models) and models built using 1: N alignments created by one of the proposed approaches (final models) were evaluated using three different measures of model quality; (i) root mean square deviation (RMSD) for all C_α atoms of the created model from the native protein chain, (ii) RMSD from model to native chain for all C_α atoms in each gap region, and (iii) Ramachandran plot [24] for each model. RMSD values were obtained by fitting using the McLachlan algorithm [34] as implemented in the program ProFit (Martin, ACR, http://www.bioinf.org.uk/software/profit/). Since we wanted to evaluate the entire fold, the RMSD value was calculated for the entire protein chain, not only the core regions. This might thus introduce some difficulty when validating the models since the loop configuration might be mobile. However, using the entire chain enables measuring the influence of the local loop

conformation on the entire structure. Ramachandran plots were created with the program PROCHECK [25]. While analyzing native structures, only the selected chain of the native structure was used (see Table 1). Non-standard atom groups ("HETATM" entries) in structure files as well as residues containing multiple alternative locations for atoms were ignored. All structure files were examined to create an alignment between model and native structure for fitting manually.

Table 4 shows average RMSDs obtained for the training set from modeling using 1: *N* alignments created by all combinations of stem overlap and alignment ranking, as well as for each of the stem overlap and ranking approaches. Shown for comparison are average RMSDs for initial models, built from the initial alignment with gap regions modeled using MODELLER's own loop modeling. The averages were calculated over all proteins in the training set, and for models, over all five models created per protein. Average RMSDs are presented for entire structures and for gap regions.

This initial prestudy of the training set showed that using no stem overlap produced the lowest-quality models while there was no great difference between models produced using overlaps of three and ten residues. For stem overlaps of three and ten residues, average RMSDs of entire structures were higher than those of the main templates, but lower than those of the initial models, i.e., final models were improved. The most significant difference between alignment-ranking approaches was that when combined with no stem overlap, solvation energy produced models of higher quality than the other two rankings. However, these models were still of lower quality than models created using stem overlap. Though the difference was not great, ranking by alignment score produced the lowest-quality models for all stem overlaps.

In terms of structure RMSD, a stem overlap of ten residues and ranking by GenTHREADER score produced best results, while in terms of gap-region RMSD a stem overlap of three residues and ranking by alignment score proved best. However, there were no great differences in average quality between any of the combinations using a stem overlap of three or ten residues.

Although the PROCHECK program suite reports several structural features, e.g., Ramachandran plots, stereochemical parameters, hydrogen bonding (packing), etc., in this study we focused on the Ramachandran. A Ramachandran plot indicates whether the backbone of a structure has a configuration that is unusual and therefore likely to be incorrect. Since the model building assumes that the backbone is correct, this a major quality measure for models. In this study, five models were generated for each target, within approximately 1 Å of each other (see above). To estimate the average quality of the backbone in these structures, averages were computed for the percentage of residues in the different regions of the Ramachandran plots, i.e., most favorable (core), allowed, generously allowed and disallowed. That is, the five models for each modeling approach resulted in four averages.

The Ramachandran plots for the training set (not shown) were in line with RMSD values, in that the approaches using no stem overlap produced the worst values (lower percentage of residues in most favored regions and higher percentage of residues in disallowed regions). There were no significant differences in average Ramachandran plot values between different alignment rankings. Stem overlaps of three and ten residues produced average Ramachandran values similar to those of initial models, but slightly higher than native structures and main template structures.

When selecting a combination of stem overlap and ranking approaches, it was decided to favor good values for entire-structure RMSD over gap-region RMSD values. The reason for this was that even though the focus of loop modeling is on the loops themselves, the final goal is a high-quality model of the entire protein. Favoring the entire-structure RMSD may enable the gap regions to influence the fold, while using an approach based on favoring the gap- region RMSD does not. Furthermore, using the gap-region RMSD might be misleading; loop configurations may alter due to environmental factors, e.g., temperature. Thus, a method for loop modeling was proposed using a stem overlap of ten residues and ranking of alignments by the GenTHREADER neural-network score. This method was then applied to the test set.

Table 4 Average RMSD for training set

Category of structures	Structure RMSD		Gap region RMSD	
	Average (Å)[a]	SD[b]	Average (Å)[c]	SD[d]
Main template structures	11.73	9.7	n/a	n/a
Initial models	14.75	8.3	6.21	3.6
Solvation energy, no overlap	16.14	9.5	7.38	4.4
Solvation energy, overlap 3	13.39	8.4	5.68	3.4
Solvation energy, overlap 10	12.93	8.1	5.16	3.2
Alignment score, no overlap	22.51	18.4	8.98	10.2
Alignment score, overlap 3	13.83	8.2	4.61	3.1
Alignment score, overlap 10	13.32	8.3	4.63	3.5
GenTHREADER score, no overlap	21.77	18.6	9.70	10.0
GenTHREADER score, overlap 3	13.70	8.1	4.97	2.7
GenTHREADER score, overlap 10	12.57	8.4	5.13	3.0

[a] Average RMSD from native structures for all proteins in training set
[b] SD of RMSD from native structures
[c] Average RMSD from native conformations for all gap regions in training set
[d] SD of RMSD from native conformations of gap regions

All gap-region sequences in the test set were extracted. To each sequence segment was added the ten immediately preceding and succeeding residues, where possible. The extended sequence was then submitted to GenTHREADER. All of GenTHREADER's filtering options were disabled. The highest scoring alignment as determined by GenTHREADER's neural network was selected to model the gap region. Alignments to native structures were ignored. The local alignments produced by GenTHREADER were integrated into the initial alignment between the target sequence and main template structure to create a 1: *N* alignment, each gap region adding one sequence to the alignment. Figure 2 shows the method applied to CASP target T0191.

The MODELLER program was used to build five models for the initial alignment (initial models) as well as for each 1: *N* alignment (final models). The models from the test set were evaluated using the same three measures as with the training set. The average RMSD between the five models and the native structure for each target sequence was used to reflect the general outcome of each modeling approach. Results were evaluated both over all gap regions and over terminal gap regions (located at the C-terminus or N-terminus of the protein chain) and non-terminal gap regions separately. Because of significantly better results for non-terminal gap regions, these were evaluated further. For some of the most interesting results a Student's *t*-test was performed to determine the statistical significance. For this, Microsoft Excel's statistical functions (TTEST and TINV) were used to do a paired, two-tailed *t*-test with initial data and final data as input sets. That is, the degrees of freedom (*df*) were calculated according to Eq. 1:

$$df = (n_1 - 1) + (n_2 - 1) \quad (1)$$

where n_1 and n_2 represent the number (population) of initial and final average models, e.g., the change in RMSD between initial and final structures results in $n_1 = 29$ and $n_2 = 29$ (omitting target sequences/native structures where RMSD could not be computed; see Table 5). The critical values for rejecting the hypothesis that the two populations are the same at probability P (here $P = 0.05$) are obtained from the *t*-distribution given the degrees of freedom, e.g., for $df = 56$ the critical value is 2.00 (ob-

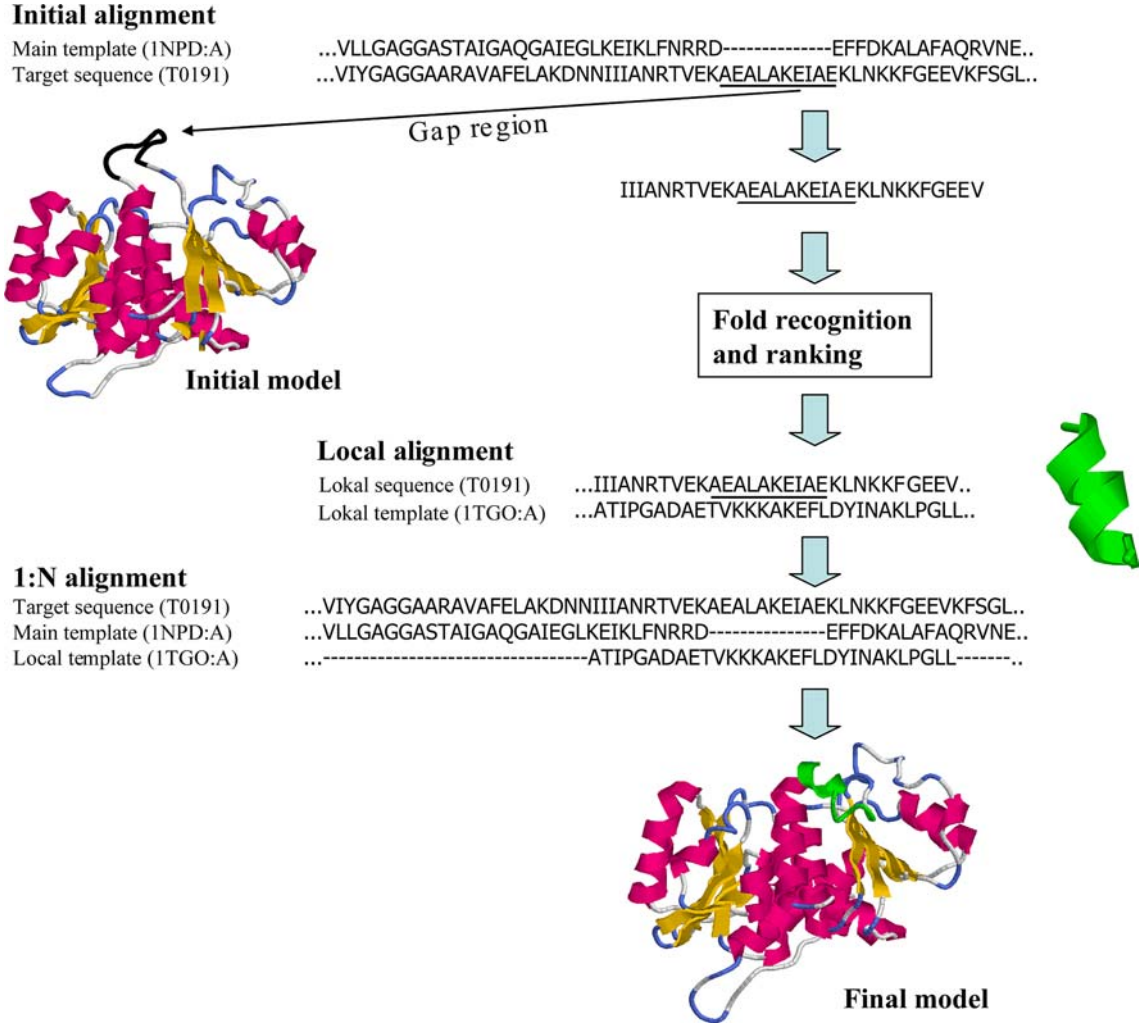

Fig. 2 The method derived from the training set applied to CASP target T0191

Table 5 Average RMSD from native structures

Target[a]	Structure RMSD						
	Template (Å)[b]	Initial models[c]		Final models[d]		Change[e]	
		Average (Å)	SD	Average (Å)	SD	(Å)	(%)
T0089	17.38	17.74	0.2	17.49	0.1	−0.25	−1
T0090	1.90	5.10	0.6	6.83	0.7	1.73	34
T0096	15.47	23.69	0.7	18.24	0.3	−5.45	−23
T0100	3.00	6.86	0.4	7.02	0.3	0.16	2
T0101	14.62	16.46	0.2	n/a[f]	n/a[f]	n/a[f]	n/a[f]
T0104	12.96	15.27	0.2	14.85	0.4	−0.42	−3
T0110	6.94	6.58	0.3	7.23	0.3	0.65	10
T0115	7.72	9.54	0.2	9.57	0.1	0.03	0
T0116	38.10	46.34	0.5	38.99	0.2	−7.35	−16
T0117	6.81	7.57	0.2	7.46	0.2	−0.11	−1
T0118	23.44	20.38	0.8	19.80	0.6	−0.58	−3
T0121	4.99	97.18	1.3	14.11	0.1	−83.07	−85
T0122	2.77	3.60	0.2	3.00	0.1	−0.60	−17
T0124	31.57	44.32	1.1	42.42	0.9	−1.90	−4
T0128	1.24	4.48	0.3	4.41	0.4	−0.07	−2
T0132	3.93	5.60	0.4	5.15	0.6	−0.45	−8
T0141	8.56	15.52	1.1	13.09	0.3	−2.43	−16
T0146	9.80	12.95	0.6	10.94	0.1	−2.01	−16
T0149	21.84	21.23	0.3	20.69	0.2	−0.54	−3
T0151	5.38	5.58	0.5	5.38	0.3	−0.20	−4
T0154	3.72	4.64	0.4	4.44	0.3	−0.20	−4
T0155	0.91	0.85	0.0	0.86	0.0	0.01	1
T0161	18.11	19.02	0.4	18.27	0.1	−0.75	−4
T0162	21.67	22.50	0.6	22.64	0.5	0.14	1
T0164	14.61	14.90	0.3	14.51	0.4	−0.39	−3
T0165	14.26	16.29	0.2	15.73	0.2	−0.56	−3
T0166	3.61	4.89	0.6	4.98	0.5	0.09	2
T0172	21.28	31.89	0.7	22.53	0.2	−9.36	−29
T0184	19.85	16.97	0.3	17.02	0.4	0.05	0
T0187	20.48	21.43	0.2	n/a[f]	n/a[f]	n/a[f]	n/a[f]
T0191	6.00	7.24	0.1	6.51	0.1	−0.73	−10

[a] CASP target ID
[b] RMSD between the native structure and the main template structure used to model the protein
[c] Average RMSD and SD for initial models
[d] Average RMSD and SD for final models
[e] Change in average RMSD from initial models to final models. Negative values indicate improvements
[f] Modeling from 1:N alignment failed

tained by the Microsoft Excel TINV function; critical values for the *t*-test are available in most textbooks on statistics). The true critical value $T_{\mathrm{RMSD(initial\text{-}final)}}$ was calculated by Eq. 2:

$$t = \frac{\overline{Y_1} - \overline{Y_2}}{\sqrt{((s_1)^2/n_1) + ((s_2)^2/n_2)}} \quad (2)$$

where $\overline{Y_i}$ represents the mean in each population i, and s_i is the standard deviation (SD) in each population. Using Eq. 2 results in $T_{\mathrm{RMSD(initial\text{-}final)}} = 1.36$. Thus, the hypothesis is not rejected since the critical value is not exceeded; $P > 0.05$. This is also shown by using the Microsoft Excel TTEST function, which results in a probability of 0.18, i.e., greater than 0.05. The implication was that the change in RMSD between initial and final models is not significant.

Results

The initial study of the training-set proteins suggested that a method for loop modeling of gap regions using a stem overlap of ten residues and ranking of alignments by the GenTHREADER neural-network score produced the best results. This setup was applied to the test set.

Table 5 shows the average RMSDs for initial models and final models. The change in RMSD from initial models to final models is shown both in Å ngströms and as a percentage of initial model RMSD. The averages were calculated over all five models built from each alignment. Two of the targets failed modeling (see Discussion). The change in RMSD between initial and final models ranged between 85% lower to 34% higher, compared to the initial models. For 21 of the 31 targets (68%), the final models had a better average RMSD than the initial models. Of the remaining targets, two failed modeling and eight produced final models of lower quality than the initial models. The average change for final models was 7% lower RMSD compared to initial models.

For the 54 gap regions in the test set, the average change in RMSD from initial to final models ranged between 84% lower RMSD to 160% higher RMSD (see Table 6). Five gap regions could not be evaluated since they belonged to one of the proteins that failed modeling, and 21 gap regions in the final models showed an average RMSD higher than the initial models. Four gap regions could not be evaluated because of too many

Table 6 Average RMSD from native conformation of gap regions

Target[a]	Region[b]	Gap region RMSD						
		Template[c]	Initial models[d]		Final models[e]		Change[f]	
			Average (Å)	SD	Average (Å)	SD	(Å)	(%)
T0089	107–121	1AUA	7.41	0.5	4.73	0.1	−2.68	−36
T0089	357–369	7ODC:A	6.37	1.5	6.33	0.2	−0.04	−1
T0089	394–419[i]	1AUK	n/a[g]	n/a[g]	n/a[g]	n/a[g]	n/a[g]	n/a[g]
T0090	1–13[i]	1F82:A	3.48	0.7	4.03	0.9	0.55	16
T0090	150–159	1CM3:A	1.70	0.5	1.81	0.1	0.11	6
T0096	3–12	1VJT:A	3.25	0.4	3.21	0.3	−0.04	−1
T0096	203–239[i]	1JCU:A	10.82	0.8	10.48	0.4	−0.34	−3
T0100	115–128	1V86:A	3.77	0.5	4.35	0.2	0.58	15
T0100	249–262	1OAC:A	7.21	0.5	4.78	0.3	−2.43	−34
T0101	168–197	1PFO	n/a[h]	n/a[h]	n/a[h]	n/a[h]	n/a[h]	n/a[h]
T0104	62–73	1PFO	3.60	0.2	4.02	0.3	0.42	12
T0104	130–139	1TBM:A	3.02	0.5	3.24	0.2	0.22	7
T0110	107–128[i]	1FUI:A	n/a[g]	n/a[g]	n/a[g]	n/a[g]	n/a[g]	n/a[g]
T0115	1–10[i]	1PIE:A	1.90	0.9	1.85	0.4	−0.05	−3
T0116	55–71	1K9D:A	5.03	0.7	5.92	0.1	0.89	18
T0116	549–563	1GKR:A	7.73	0.2	7.28	0.1	−0.45	−6
T0116	571–580	1E9S:E	2.81	0.1	2.64	0.2	−0.17	−6
T0116	606–621	1AYL	6.02	1.0	6.58	0.4	0.56	9
T0116	673–683	16PK	5.47	0.2	1.91	0.7	−3.56	−65
T0116	687–811[i]	1MOJ:A	58.95	2.2	16.61	0.1	−42.34	−72
T0117	1–16[i]	1HZ4:A	1.61	0.6	2.68	0.0	1.07	67
T0118	140–149[i]	1PFO	1.82	0.3	3.20	0.1	1.38	76
T0121	248–372[i]	1GVF:A	102.68	1.8	16.14	0.1	−86.54	−84
T0122	165–177	1IWG:A	2.75	0.4	2.79	0.3	0.04	1
T0124	1–15[i]	1GM5:A	5.31	1.0	7.54	0.1	2.23	42
T0124	227–242[i]	1BE3:B	5.65	0.3	3.97	0.9	−1.68	−30
T0128	1–15[i]	1JB0:B	0.46	0.4	1.20	0.6	0.74	160
T0132	1–15[i]	1BYB	0.99	0.4	1.34	0.5	0.35	35
T0141	1–20[i]	1DN1:A	6.50	0.6	6.74	0.4	0.24	4
T0141	48–58	1TVF.A	3.69	0.4	4.14	0.1	0.45	12
T0146	310–325[i]	1SF9:A	5.43	1.8	7.57	0.1	2.14	39
T0149	37–53	1ACC	4.78	1.0	6.40	0.2	1.62	34
T0149	81–90	1LDJ:A	3.92	0.2	2.74	1.1	−1.18	−30
T0149	114–125	1LRW:A	5.08	0.4	5.58	0.2	0.50	10
T0149	287–296	1DF0:A	4.16	0.9	3.75	0.2	−0.41	−10
T0151	144–164[i]	1FIQ:C	n/a[g]	n/a[g]	n/a[g]	n/a[g]	n/a[g]	n/a[g]
T0154	1–10[i]	1OYG:A	2.35	0.5	2.37	0.7	0.02	1
T0154	288–309[i]	1H80:A	0.10	0.1	0.25	0.2	0.15	157
T0155	121–133[i]	1L5J:A	n/a[k]	n/a[g]	n/a[g]	n/a[g]	n/a[g]	n/a[g]
T0161	147–156[i]	1YGE	3.28	0.7	5.08	0.0	1.80	55
T0162	246–255	1AV1:A	4.18	0.4	3.17	0.2	−1.01	−24
T0164	72–82	1TVF:A	4.90	0.4	4.76	0.2	−0.14	−3
T0165	30–40	1VK3:A	5.13	0.3	4.40	0.5	−0.73	−14
T0165	92–103	1I3Q:A	5.42	0.3	5.20	0.2	−0.22	−4
T0166	139–150[i]	1BYB	3.57	0.2	2.90	0.6	−0.67	−19
T0172	105–123	1GPR	6.75	0.2	4.84	0.4	−1.91	−28
T0172	235–244	1UFK:A	5.81	0.9	1.65	0.3	−4.16	−72
T0172	257–299[i]	1C3C:A	25.85	1.3	13.93	1.0	−11.92	−46
T0184	14–24	1I4S:A	3.69	0.6	0.81	0.1	−2.88	−78
T0187	48–59	1DQR:A	n/a[h]	n/a[h]	n/a[h]	n/a[h]	n/a[h]	n/a[h]
T0187	84–95	1IWG:A	n/a[h]	n/a[h]	n/a[h]	n/a[h]	n/a[h]	n/a[h]
T0187	253–281	2ACY	n/a[h]	n/a[h]	n/a[h]	n/a[h]	n/a[h]	n/a[h]
T0187	289–304	1FNO:A	n/a[h]	n/a[h]	n/a[h]	n/a[h]	n/a[h]	n/a[h]
T0191	158–167	1TGO:A	4.56	0.2	2.41	0.1	−2.15	−47

[a] CASP target ID
[b] Residue sequence numbers included in gap region
[c] Local template for gap region
[d] Average gap region RMSD and SD for initial models
[e] Average gap region RMSD and SD for final models
[f] Change in average gap region RMSD from initial models to final models. Negative values indicate improvements
[g] Too many residues were missing in native structure file for RMSD to be calculated.
[h] Modeling from 1:N alignment failed
[i] Terminal gap region

missing residues in native structure files. The remaining 24 gap regions had an improved RMSD. This is 44% of all gap regions and 53% of those, which could be evaluated. The average change for all gap regions was a 1% higher RMSD.

There was a large difference between the results for terminal and non-terminal gap regions; terminal gap regions ranged from 84% lower to 160% higher RMSD. Thirty-two percent of terminal gap regions showed an improvement in RMSD and the average change was a 22% increase in RMSD. The non-terminal gap regions ranged from 78% lower to 34% higher RMSD. Fifty-three percent of the non-terminal gap regions showed a decrease in RMSD. Discounting gap regions, which could not be evaluated, this figure was 63%. The average RMSD improvement for non-terminal gap regions was 12%.

The average RMSD for all non-terminal gap regions in initial and final models of the proteins in the test set is shown in Fig. 3. The relation between the change in gap-region RMSD and the length of the region is shown in Fig. 4 for both terminal and non-terminal regions.

The GenTHREADER confidence measure revealed that 14 of the template structures used in the initial alignments had a confidence level of "certain," 11 template structures had a confidence level of "high" and six structures had a confidence level of "medium." The average RMSD for the entire structure for confidence level "certain" was 11.3 Å, for "high" it was 12.8 Å and for "medium" it was 14.1 Å . The average improvements to the final models were 1.1 Å ; for confidence level "certain" it was 0.8 Å, for "high" it was 1.7 Å, and for "medium" it was 0.7 Å . When a good initial template (with low RMSD to the native structure) was identified, the method outlined in this paper preformed better than MODELLER's loop modeling. The change in RMSD for non-terminal gap regions showed that 9 out of 27 loop regions using the outlined approach had improvements larger than 1 Å (average 2.4 Å) compared to initial models. Only one gap region in the initial models had more than 1 Å (1.6 Å) lower RMSD than in final models. The remaining 17 gap regions resulted in changes less than 1 Å between initial and final models.

The change in Ramachandran plots from initial to final models (not shown) showed no clear correlation to RMSD values. According to the number of residues in the most favored regions of the Ramachandran plots, there was on average some degradation in final models compared to initial models. The average number of residues in disallowed regions, however, remained unchanged.

Significance tests were performed and evaluated using a two-tailed *t*-test with a significance level of 0.05. The *t*-test (see Material and methods) for the improvement in average RMSD of entire structures between initial and final models was not statistically significant ($T_{\text{RMSD-structure(initial-final)}} = 1.36$, $P > 0.05$, $df = 56$, *t*-test probability = 0.18). Using the *t*-test for the average gap region, the RMSD improvement between initial and final models in the entire test set was not significant ($T_{\text{RMSD-gaps(initial-final)}} = 1.57$, $P > 0.05$, $df = 88$, *t*-test probability = 0.12). Significance tests for the change of terminal gap regions and non-terminal gap regions revealed that the terminal gap region change in RMSD was not significant ($T_{\text{RMSD-gaps(initial-final)-terminal}} = 1.37$, $P > 0.05$, $df = 34$, *t*-test probability = 0.18). The improvement for non-terminal gap regions, however, was statistically significant ($T_{\text{RMSD-}}$

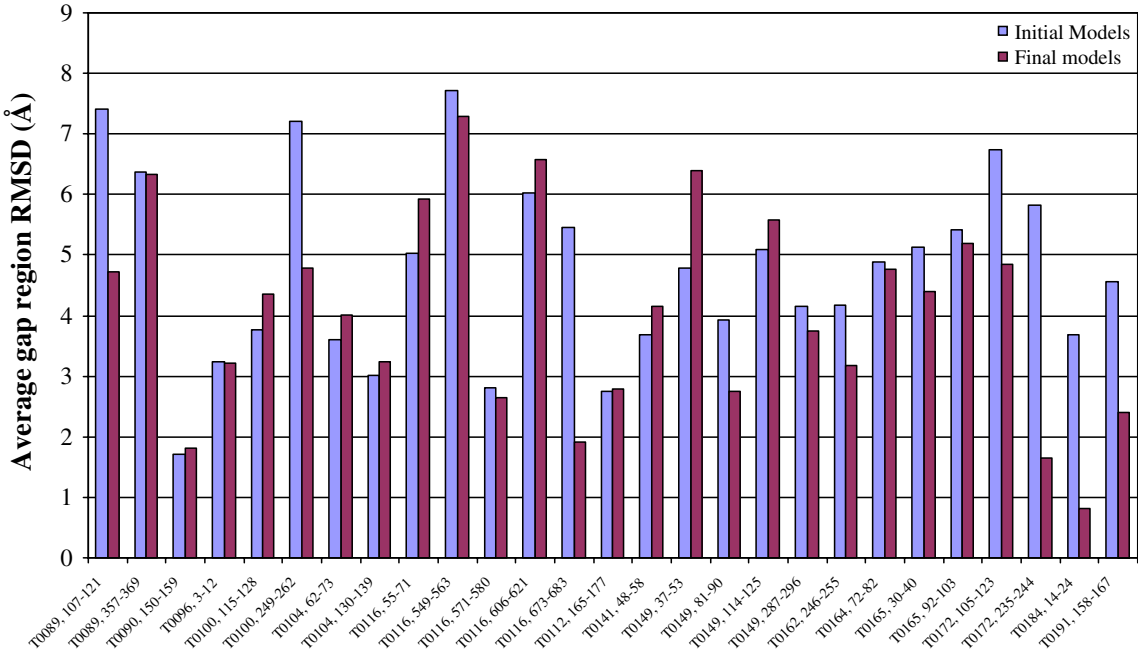

Fig. 3 Average RMSDs for non-terminal gap regions in initial and final models

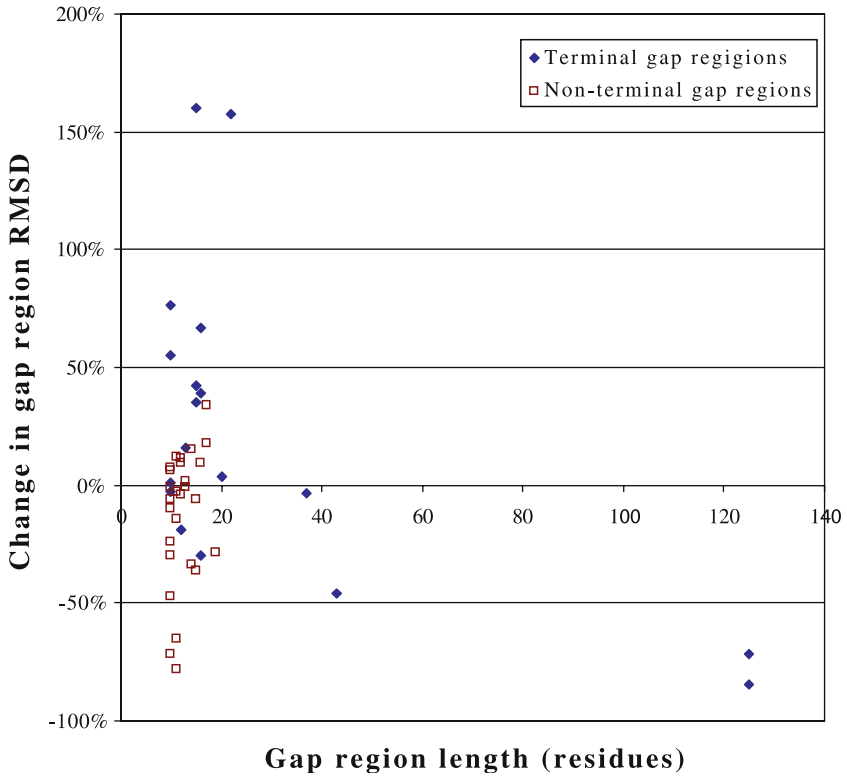

Fig. 4 Change in gap region RMSD by gap region length and terminalness

$_{\text{gaps(initial-final)-non-terminal}} = 2.41$, $P < 0.05$, $df = 52$, t-test probability = 0.020). For the nine proteins in the test set that contained no terminal gap regions, the average change in entire-structure RMSD was an improvement of 4%. The t-test probability was statistically significant ($T_{\text{RMSD-structure(initial-final)-non-terminal gaps}} = 2.51$, $P < 0.05$, $df = 16$, t-test probability = 0.023).

Discussion

There is now general agreement [2] that changes in the nature of structure modeling have made these categories (comparative modeling, fold recognition, and new fold methods) outdated, e.g., improved sequence-comparison techniques have blurred the boundary between comparative modeling and fold recognition.

The aim of the work presented was to determine whether a fold-recognition approach to loop modeling could improve the quality of protein models. In particular, when loop regions are long and suitable conformations are difficult to find, alternative approaches must be investigated. This approach applied fold recognition to sequence regions in a sequence-structure alignment, which were not covered by the structural template. The result of this was a 1: N alignment created from the initial sequence-structure alignment through the addition of local template structures. For this purpose, we have used GenTHREADER as a working tool. The framework for the performed work treated the protein structure to be unknown.

To identify local templates with respect to finding the possible solutions, three different parameters were identified: alignment score, solvation energy, and GenTHREADER's combined neural-network score. These measures were here believed to capture features that made it possible to find a good solution when modeling the structure. An addition to these features was the length of stem overlap, which has previously been identified as a major feature for determining the loop configuration. A short overlap is challenging to integrate with the rest of the model, while a long stem overlap can introduce difficulties in modeling by having two completely different template structures for one sequence of residues. The impact of this is that different modeling algorithms may require different lengths of stem overlap for optimal results. Another aspect of varying stem-overlap length is that it changes the sequence for fold recognition. This could change the template structure suggested for the gap region, which in turn can affect the model quality. Using no stem overlap produced severe impact on the results, confirming that stem regions do have an influence on the conformation of gap regions. Stem overlaps of three and ten residues produced results of similar quality to each other, and were better than no overlap. It is possible that better results could be achieved for some length between three and ten, or greater than ten.

The approach selected for the proposed method was made from evaluation of the training set. There were no significant differences between the best-performing approaches. The best approaches according to the struc-

tures created from the training set were a stem overlap of ten residues and ranking by GenTHREADER score (for entire structures), and a stem overlap of three residues and ranking by alignment score (for gap regions). When deciding which approach to use, it was decided to favor good results for entire-structure RMSD, and thus a ten residues overlap and ranking by GenTHREADER score were selected. A useful measure for ranking of alignments is distinguished by a high correlation to the change in RMSD. As expected, there was a general tendency for gap regions with low RMSD to have low solvation energy, a high alignment score and a high GenTHREADER score. However, the GenTHREADER score selected for use in the proposed method did not seem to show a higher correlation to the change in RMSD of gap regions than ranking by solvation energy or alignment score, indicating that gap-region alignment ranking could be improved. Ranking by alignment score, the approach most similar to a traditional homology loop modeling produced the lowest quality models for the training set for all stem overlaps. This indicates that a fold recognition approach is able to capture additional information to create better models.

The data set for this study was chosen from the prediction targets used in CASP4 and CASP5. Structures obtained from the PDB occasionally revealed discrepancies from associated CASP sequences, e.g., missing residues in structures, as shown in Table. 2 and 3. The missing residues often corresponded to gap regions in the sequence-structure alignment, especially for terminal gap regions. The missing residues may have introduced unwanted bias on the model building (template structure alignments) and the evaluation of the models (comparing to native structures). A decrease in quality and reliability is naturally assumed with increasing number of missing residues, especially in terminal regions. To overcome the missing-residue problem, sequences could have been extracted from PDB structure files to ensure sequences and structures would match. However, this would not reflect a realistic process where the protein structure is unknown.

Model building failed for two targets, T0101 and T0187, due to an alignment of the target sequence to two or more very different structures. Because of nearby gap regions and the stem overlap of ten residues, the alignment for T0187 contained three template structures covering the same sequence. The alignment for T0101 contained a short gap region (less than nine residues) near a longer gap region (more than nine residues). Here, the short gap region may have interfered with the integration between the long gap region's template and the main template structure. One possible solution to these modeling problems could be to treat gap regions separated by a small number of residues as one sequence for threading. This would replace the template structure or structures with one template covering both gap regions.

The approach outlined was evaluated based on the average of five generated models for each alignment. For each model, RMSD for both the entire structure and the separate loop regions were measured in addition to Ramachandran values. It might be argued that the evaluation should be done on the best model generated from each alignment. The problem then shifts to identifying the best model for each alignment. However, determining the best model is difficult and usually based on human intervention. To eliminate this, the average was used.

It might be argued that loop configuration is dynamic. That is, loops are generally believed to alter configuration due to environmental factors, e.g., B-factor. The influence of this mobility is difficult to evaluate in structure prediction. This is reflected by the RMSD value in the evaluation. However, even if loops contained dynamics allowing them to alter the configuration, the decreased RMSD shows significant improvement to at least one possible loop configuration, namely the one contained by the native structure.

A simple test of examining the gap-region energy by the Gromacs molecular dynamics software [35, 36], performed by taking the initial potential energy at time zero and ignoring the remaining time steps, revealed for the target sequence T0191, sequence residues 158–167, that the gap-region energy of final models was on average 135 kJ mol^{-1} from the native structure (SD of 84.5), while that of initial models was on average 889 kJ mol^{-1} from the native structure (SD of 221). Thus, the improvement in RMSD also reflected decreased gap-region energy. However, if this is a general feature is yet to be determined and should be investigated further.

Applying the approach to a larger dataset, i.e., the test set, the results indicated that modeling gap regions by fold recognition is indeed a promising approach. While the method did not perform well on gap regions located at the C-terminus or N-terminus of a chain, so called dangling regions, non-terminal gap regions were modeled more accurately. For the two very long-terminal gap regions (125 residues each), the fold-recognition approach showed great improvement from the models calculated by MODELLER. However, this is rather because MODELLER is not suited for modeling long-terminal gap regions than because of merits of the fold-recognition approach. Disregarding these artificially positive results, performance for terminal gap regions appears even worse.

The longest non-terminal gap regions successfully modeled in the test was 19 residues long. T0101 and T0187, which failed modeling, contained the two longest non-terminal gap regions in the test set (30 and 29 residues, respectively), which may indicate that the method is not suited for longer gap regions. In the training set, however, non-terminal gap regions of 20, 23, and 27 residues were modeled successfully.

One potential problem was the presence of alignment gaps in local alignments. These were not specifically dealt with, but were left for MODELLER's loop-modeling function to handle. It is possible that restricting the

use of local alignments with too many gaps could lead to better models.

As the only input needed is a sequence-structure alignment, the method can be applied to alignments created through either comparative modeling or fold recognition. In the integration of local alignments into the main sequence-structure alignment, a completely mechanistic approach was taken. The 1:N alignments could likely have been improved by further intensive fine-tuning. Since the method as presented does not need any human intervention, it could be implemented as an automated server.

The approach used here could probably be improved by further adjustment of parameters such as stem overlap or development of better criteria for the ranking of alignments. However, the features used, i.e., GenTHREADER score, solvation energy, and alignment score, demonstrated that they certainly play a role in the prediction of the gap regions. This could also involve identifying better strategies or tools for ranking possible local folds. In this study GenTHREADER was used. However, there exist a number of other tools that can be regarded as suitable for this approach, e.g., THREADER, 3D-PSSM, and LOOPP. Also, as with traditionally applied fold recognition, performance will improve with time as more experimentally determined structures are made available and added to the fold database.

While average results of final models were better than those of MODELLER's initial loop modeling, results were not consistently better. However, it is argued here that as a rule of thumb it is better to use the approach outlined here than not using any gap-region template. At the very least it provides a complement to existing techniques. It is hoped that further work may improve the method and indicate scenarios where the method can be expected to produce improved results.

While the method proposed has been compared to the loop modeling in the program MODELLER, it would be of great interest to perform a more thorough comparison to other loop-modeling methods. Also, the performance for longer gap regions should be investigated.

Models created by the proposed method were not universally improved. However, it was shown that for non-terminal gap regions in the test set and for the proteins that contained them, the average RMSD was improved. These improvements were shown to be statistically significant.

References

1. Fiser A, Do RKG, Šali A (2000) Protein Sci 9:1753–1773
2. Moult J, Fidelis K, Zemla A, Hubbard T (2003) Proteins 6(Suppl):334–339
3. Martí-Renom MA, Stuart AC, Fiser A, Sánchez R, Melo F, Šali A (2000) Annu Rev Biophys Biomol Struct 29:291–325
4. Šali A, Blundell TL (1993) J Mol Biol 234:779–815
5. Schwede T, Kopp J, Guex N, Peitsch MC (2003) Nucleic Acids Res 31:3381–3385
6. Jones D, Hadley C (2000) Threading methods for protein structure prediction. In: Higgins D, Taylor W (eds) Bioinformatics: sequence, structure and databanks. Oxford University Press, Oxford
7. Jones DT, Taylor WR, Thornton JM (1992) Nature 358:86–89
8. Jones DT, Miller RT, Thornton JM (1995) Proteins 23:387–397
9. Jones DT (1998) THREADER: protein sequence threading by double dynamic programming. In: Salzberg SL, Searls DB, Kasif S (eds) Computational methods in molecular biology. Elsevier Science, Amsterdam
10. Jones DT (1999) J Mol Biol 287:797–815
11. McGuffin LJ, Jones DT (2003) Bioinformatics 19:874–881
12. Fischer D, Barret C, Bryson K, Elofsson A, Godzik A, Jones D, Karplus KJ, Kelley LA, MacCallum RM, Pawowski K, Rost B, Rychlewski L, Sternberg M (1999) Proteins 3(Suppl):209–217
13. Kelley LA, MacCallum RM, Sternberg MJE (1999) Recognition of remote protein homologies using three-dimensional information to generate a position specific scoring matrix in the program 3D-PSSM. In: Istrail S, Pevzner P, Waterman M (eds) RECOMB99, Proceedings of the third annual international conference on computational molecular biology. The Association for Computing Machinery, New York
14. Kelley LA, MacCallum RM, Sternberg MJE (2000) J Mol Biol 299:499–520
15. Tobi D, Elber R (2000) Proteins 41:40–46
16. Meller J, Elber R (2001) Proteins 45:241–261
17. Teodorescu O, Galor T, Pillardy J, Elber R (2004) Proteins 54:41–48
18. Moult J (1999) Curr Opin Biotechnol 10:583–588
19. van Vlijmen HWT, Karplus M (1997) J Mol Biol 267:975–1001
20. Rohl CA, Strauss CEM, Chivian D, Baker D (2004) Proteins 55:656–677
21. Moult J, Pedersen JT, Judson R, Fidelis K (1995) Proteins 23:ii–iv
22. Moult J, Hubbard T, Bryant SH, Fidelis K, Pedersen JT (1997) Proteins 1(Suppl):2–6
23. Svensson M, Lundh D, Ejdebäck M, Mandal A (2004) J Mol Model 10:130–138
24. Ramachandran GN, Ramakrishnan C, Sasisekharan V (1963) J Mol Biol 7:95–99
25. Laskowski RA, MacArthur MW, Moss DS, Thornton JM (1993) J Appl Cryst 26:283–291
26. Barnes D, Cohen A, Bruick RK, Kantardjieff K, Fowler S, Efuet E, Mayfield SP (2004) Biochemistry 43:8541–8550
27. Leszczynski JF, Rose GD (1986) Science 234:849–855
28. Tramontano A, Lesk AM (1992) Proteins 13:231–245
29. Murzin A, Hubbard TJP (2001) Proteins Suppl 5:8–12
30. Kinch LN, Qi Y, Hubbard TJP, Grishin NV (2003) Proteins 6(Suppl):340–351
31. Pearson WR, Lipman DJ (1988) Proc Natl Acad Sci USA 85:2444–2448
32. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) Nucleic Acids Res 28:235–242
33. McGuffin LJ, Bryson K, Jones DT (2000) Bioinformatics 16:404–405
34. McLachlan AD (1982) Acta Crystallogr A 38:871–873
35. Lindahl E, Hess B, van der Spoel D (2001) J Mol Model 7:306–317
36. Berndsen HJC, van der Spoel D, van Drumen R (1995) Comp Phys Comm 91:43–56