Chapter 3

Methods for Sequence–Structure Alignment

Česlovas Venclovas

Abstract

Homology modeling is based on the observation that related protein sequences adopt similar three-dimensional structures. Hence, a homology model of a protein can be derived using related protein structure(s) as modeling template(s). A key step in this approach is the establishment of correspondence between residues of the protein to be modeled and those of modeling template(s). This step, often referred to as sequence–structure alignment, is one of the major determinants of the accuracy of a homology model.

This chapter gives an overview of methods for deriving sequence–structure alignments and discusses recent methodological developments leading to improved performance. However, no method is perfect. How to find alignment regions that may have errors and how to make improvements? This is another focus of this chapter. Finally, the chapter provides a practical guidance of how to get the most of the available tools in maximizing the accuracy of sequence–structure alignments.

Key words: Homology modeling, Protein structure, Sequence profiles, Hidden Markov models, Alignment accuracy, Model quality

1. Introduction

At present, homology or comparative modeling is the most accurate and therefore the most widely used protein structure prediction approach. Homology modeling is based on the empirical observation that evolutionary-related proteins (to be more precise—evolutionary-related protein domains) tend to have similar three-dimensional (3D) structures. Moreover, protein structural features often remain preserved long after the sequence signal is lost to mutations, insertions, and deletions. Therefore, 3D structure is considered to be the most robustly conserved feature of homologous proteins, certainly more conserved than the sequence or molecular function. Although there are some convincing exceptions to this rule (1), it still holds for the absolute majority of cases.

Andrew J.W. Orry and Ruben Abagyan (eds.), *Homology Modeling: Methods and Protocols*, Methods in Molecular Biology, vol. 857, DOI 10.1007/978-1-61779-588-6_3, © Springer Science+Business Media, LLC 2012

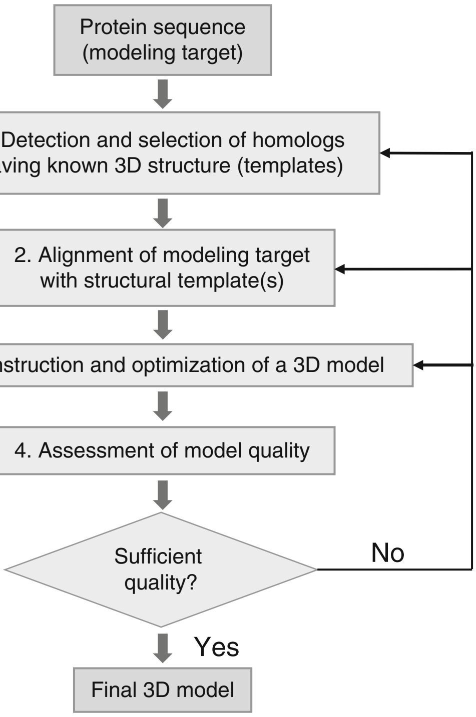

Protein sequence (modeling target)
1. Detection and selection of homologs having known 3D structure (templates)
2. Alignment of modeling target with structural template(s)
3. Construction and optimization of a 3D model
4. Assessment of model quality
Sufficient quality?
No
Yes
Final 3D model

Fig. 1. Homology modeling flowchart.

Homology modeling is used to build a 3D structural model of a protein (modeling target) on the basis of the alignment of its amino acid sequence with a related protein of known structure (template). Any homology modeling approach consists of four main steps: (1) identification of related proteins that have experimentally determined structures and therefore can be used as structural templates for modeling, (2) mapping corresponding residues between the target sequence and template structure, the process often referred to as sequence–structure alignment, (3) generating a 3D model of a target protein on the basis of the sequence–structure alignment, and (4) estimating the correctness of the resulting model. The whole process may be iterated (restarting at any of the steps) until the satisfactory estimated quality is obtained or until the model can no longer be improved (Fig. 1).

This chapter focuses on the second step in the homology modeling process—producing sequence–structure alignment—and will only touch upon other steps as necessary.

2. Sequence–Structure Alignment Problem

Once a suitable structural homolog (template) is identified, the accurate mapping of target sequence onto template structure becomes a major determinant of the resulting model quality.

What does it mean to produce an accurate sequence–structure mapping/alignment? Let us suppose that we know 3D structures of both the template and the target. If we superimpose those two structures, we will find out that for structurally similar regions of both proteins we can derive an unequivocal correspondence between residues. The sequence–structure alignment step in homology modeling aims to reproduce this correspondence as accurately as possible, but without the benefit of knowing the "real" (experimental) structure of the modeling target. Obviously, unless target and template are very closely related, there may be regions displaying significant structural differences between the two. These structurally dissimilar regions most often result from insertions, deletions, or extensive changes in the amino acid sequence. Therefore, in such regions, the assignment of residue correspondence is not always straightforward and sometimes plainly meaningless. In other words, an accurate sequence–structure alignment should include all the structurally and evolutionary equivalent residue pairs, at the same time leaving out structurally different regions. As the number of experimentally determined structures continues to grow steadily, in many cases a modeling target can be aligned not only to a single but also to a number (sometimes very large) of available structural templates. Often, an accurate alignment over the entire target length cannot be achieved with the same template; instead, different target regions (sometimes quite short) can be aligned to different templates. This provides opportunity for the model improvement but at the same time introduces additional complexity into the modeling procedure.

The sequence–structure alignment problem can be subdivided into the three subproblems: (1) generating initial sequence–structure alignment, (2) finding out which alignment regions may need adjustment, and (3) improving the alignment.

3. Sequence-Based Methods for Sequence–Structure Alignment

Usually, the construction of initial sequence alignment between the target and the template coincides with the first step in homology modeling (Fig. 1), template identification. Therefore, template identification will be discussed along with the sequence–structure alignment. Since for the modeling target only amino acid sequence is known to start with, sequence comparison is the primary means to detect related protein(s) having known experimental 3D structure. If aligned sequences share a statistically significant sequence similarity (the similarity which could not be expected by chance), it is considered that the sequences share common evolutionary origin. It further means that their 3D structures can also be expected to be similar.

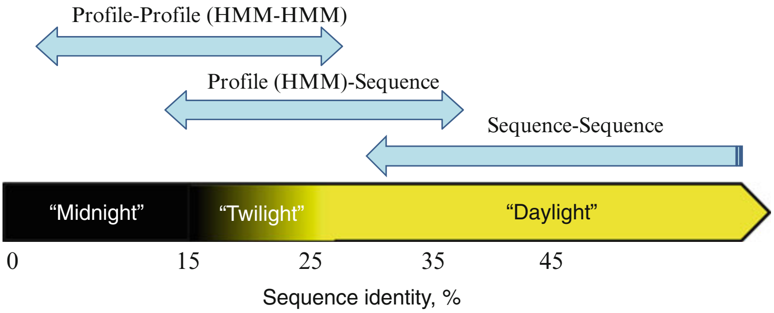


Fig. 2. Different types of homology detection and alignment methods are most effective for different sequence similarity ranges. Sequence similarity is partitioned into three approximate intervals corresponding to the decreasing difficulty of identifying homology from sequence: the "midnight" zone (<15% sequence identity), the "twilight" zone (~15–25%), and the "daylight" zone (>25%).

Depending on the evolutionary distance between proteins, sequence-based methods of different complexity may be required to detect their relationship (Fig. 2). These methods can be grouped on the basis of the increasingly complex sequence information they use:

1. Alignment of a pair of sequences
2. Profile–sequence and hidden Markov model (HMM)–sequence alignments
3. Profile–profile and HMM–HMM alignments.

3.1. Pairwise Sequence Alignment Methods

Methods that detect homology through the alignment of a pair of sequences (pairwise alignment) have emerged earliest and are conceptually the simplest. They use only amino acid sequences of two proteins, a scoring table for residue substitutions and an algorithm to produce an alignment. Usually, pairwise alignment methods report the statistical significance of the resulting alignments, allowing to use them for sequence database searches. Undoubtedly, the most popular database search tool based on pairwise alignment is BLAST (2, 3). It is very fast and has a solid statistical foundation for homology inference, provided by the incorporation of the Karlin–Altschul extreme value statistics (4). The integration of BLAST suite of programs together with major sequence databases at the National Center for Biotechnology Information (NCBI; http://www.ncbi.nlm.nih.gov/) is another important factor contributing to the popularity of BLAST. FASTA (5) and Ssearch (6, 7) are two other widely used pairwise alignment and database search methods. Pairwise sequence comparison programs can provide a fast initial estimate of the difficulty level of homology modeling. They can be adequate for detecting evolutionary-related proteins that share over 25–30% identical residues, the range of sequence similarity that may be called a “daylight” zone (Fig. 2). However, in many cases, corresponding alignments need improvements. Only if aligned sequences are over 40–50% identical to each other and have few or no gaps, it can be expected that alignments may be accurate in a structural sense.

Despite the limited and ever decreasing use of pairwise sequence comparison to obtain sequence–structure alignments for direct use in modeling, this is the initial step essentially in all of the more sophisticated sequence comparison techniques that utilize information from multiple related sequences. Therefore, the improvements in the initial pairwise comparison step may have a profound effect on the final results. Recently, a significant step forward was made by the development of the context-specific BLAST (CS-BLAST) (8). Unlike the original BLAST, which treats sequence positions independently of each other, CS-BLAST considers the substitution probability at a particular position to depend on the neighboring residues (sequence context). This methodological innovation led not only to a higher sensitivity in homology detection but also to a significant improvement of the alignment quality (8). CS-BLAST may be especially promising for application to *singleton sequences* (sequences without detectable homologs), because the lack of related sequences precludes the use of methods based on profile–sequence or profile–profile alignments that are discussed next.

3.2. Profile–Sequence and Hidden Markov Model–Sequence Alignment Methods

When the evolutionary relationship is more distant (sequence similarity is fading into the “twilight” zone; Fig. 2), the pairwise sequence comparison may not be sufficient to reliably identify homology and to produce an accurate alignment. In such cases, methods that use information from aligned multiple sequences represented by either sequence profiles (9) or HMMs (10) can be much more effective. The power of profiles and HMMs stems from a comprehensive statistical model generated for the aligned group of related sequences. This model indicates which positions are conserved and which are variable and where insertions or deletions are most likely to occur. Therefore, a comparison of a profile with database sequences can both provide more sensitive detection of homologs and generate more accurate alignments. Currently, the most widely used profile–sequence comparison method is position-specific iterated BLAST (PSI-BLAST) (3). PSI-BLAST uses a multiple alignment of the highest-scoring matches returned in an initial BLAST search to construct a position-specific scoring matrix (PSSM). The constructed PSSM replaces the generic substitution matrix (e.g., BLOSUM or PAM series) in a subsequent round of the BLAST search. This process can be repeated a number of times. Every time, new sequences detected above the predefined threshold are used to adjust the profile. Thus, with each iteration more and more distantly related sequences are included making the profile more inclusive yet still specific for the sequence family.

This makes PSI-BLAST a very powerful sequence search and comparison tool that can often detect and align homologs having sequence identities of 15% or even lower (both "twilight" and "midnight" zones of sequence similarity). Since the elementary step in PSI-BLAST is based on BLAST, it also treats positions as being independent from each other. Just like CS-BLAST, context-specific iterated BLAST (CSI-BLAST) (8) has been shown to outperform PSI-BLAST, suggesting that the incorporation of sequence context into sequence or profile comparisons is a promising avenue for improvements.

HMMER (11) and sequence alignment and modeling (SAM) (12) tool suites are the best known HMM–sequence comparison methods. HMMs are similar to sequence profiles, but they use probability theory to guide how all the scoring parameters should be set. HMMs also have additional probabilities for insertions and deletions at each position of the profile. The latter feature of HMMs is important in trying to better represent properties of protein sequence evolution. It is obvious that the probability of insertions and deletions within the protein sequence is very much position-dependent because of varying structural and/or functional constraints. While insertions/deletions may be detrimental within the structural core, they are more likely to be tolerated within solvent-exposed structurally variable regions such as loops. HMMs, however, have important limitations too. Just like sequence profiles (PSSMs), HMMs treat a particular position independent of all the other positions, and thus are not able to capture any higher-order correlations that may exist (and we know that they do!) in protein sequences. Despite seeming methodological advantages, HMM–sequence-based methods have not been used as widely as PSI-BLAST. Why so? For one, so far HMM–sequence comparison methods have been much slower than PSI-BLAST. Besides, it has been difficult to devise an iteration procedure for HMMs that would work as smoothly and seamlessly as in PSI-BLAST. However, the HMM field has made significant advances. For example, SAM-T08 (13), the latest protein structure prediction method based on SAM tool suite, features several iterative procedures. The use of heuristics has also recently helped to achieve a significant speedup and to introduce an iterative search protocol for HMMER (14). Reportedly, HMMER is now roughly on a par with BLAST according to the speed of database search, and its iterative search procedure (jackhmmer) rivals PSI-BLAST in sensitivity and alignment accuracy.

3.3. Profile–Profile and HMM–HMM Alignment Methods

Evolutionary relationships that are too distant to be detected either by pairwise sequence or by profile–sequence (HMM–sequence) comparisons ("midnight" zone; Fig. 2) may still be identified by methods that are based on profile–profile or HMM–HMM alignments. These methods add another level of complexity by comparing two sequence profiles (HMMs) instead of a profile (HMM) with a single sequence. In other words, instead of asking the question of whether a sequence belongs to the family, these methods are asking the question of whether two sequence families are evolutionary related. This generalization brought about a previously unseen sensitivity of homology detection and, albeit less dramatic, an improvement in the alignment accuracy (15–20). Although in sensitivity and alignment accuracy they still lag behind the methods based on 3D structure comparison such as DALI (21), it is possible to see examples of the opposite (17). Some of the best performers among methods based on HMM–HMM comparison include HHsearch (16) and PRC (19), while COMPASS (15), COMA (17), and PROCAIN (22) represent those based on profile–profile comparison. At present, both methodologies (profile and HMM-based) are being actively developed, and it is not clear whether one of the two will be dominating in the future. There are pros and cons on both sides. Traditionally, sequence profile–profile alignments have been using fixed gap penalties, while the HMM framework naturally accommodates more biologically relevant position-dependent gap penalties. Nonetheless, position-dependent gap penalties can be successfully implemented in profile–profile methods, as recently has been demonstrated in COMA (17). The Karlin–Altschul statistics introduced in BLAST and PSI-BLAST can be more easily extended for profile–profile than for the HMM–HMM comparison. On the other hand, recently a probabilistic model of local sequence alignment amenable to the Karlin–Altschul statistics has been introduced in HMMER. This has significantly reduced the computational cost for statistical significance estimation without sacrificing the accuracy (23). Both profile–profile and HMM–HMM methods consider sequence positions to be independent of each other, but as demonstrated by the success of CS/CSI-BLAST (8), this is clearly a non-optimal representation of protein sequence information. Indirectly, the importance of positional context in the profile–profile (HMM–HMM) comparison has been demonstrated by a boost in performance with the incorporation of additional information (16, 22). The largest impact has been observed by the inclusion of the secondary structure (SS) information, which may be considered as a particular representation of context dependency. Thus, a further improvement of the context-specific scoring may be a promising direction for increasing homology detection sensitivity and alignment accuracy.

A brief summary of different types of alignment methods is provided in Table 1.

3.4. Multiple Sequence Alignment Methods

Multiple sequence alignment (MSA) methods represent a distinct case as they are not designed to detect homologous sequences. Instead, they align a set of homologous sequences already identified by other methods, such as those discussed above. MSA methods may be useful in at least two different ways. First, these methods

Table 1
Sequence-based methods for homology detection and sequence–structure alignment construction

Method	Type	Address
BLAST	Sequence–Sequence	http://blast.ncbi.nlm.nih.gov/
FASTA/Ssearch	Sequence–Sequence	http://fasta.bioch.virginia.edu/ http://www.ebi.ac.uk/Tools/sss/fasta/
CS-BLAST	Sequence (profile)–Sequence	http://toolkit.lmb.uni-muenchen.de/cs_blast/
PSI-BLAST	Profile–Sequence	http://blast.ncbi.nlm.nih.gov/
CSI-BLAST	Profile–Sequence	http://toolkit.lmb.uni-muenchen.de/cs_blast/
HMMER	HMM–Sequence	http://hmmer.org/
SAM	HMM–Sequence	http://compbio.soe.ucsc.edu/HMM-apps/
COMPASS	Profile–Profile	http://prodata.swmed.edu/compass/
PROCAIN	Profile–Profile + additional sequence features + SS[a]	http://prodata.swmed.edu/procain/
COMA	Profile–Profile	http://www.ibt.lt/bioinformatics/coma/
HHsearch	HMM–HMM + SS[a]	http://toolkit.lmb.uni-muenchen.de/hhpred/
PRC	HMM–HMM	http://supfam.org/PRC http://www.ibi.vu.nl/programs/prcwww/

[a]Secondary structure

may be used to improve the quality of MSAs, from which profiles (HMMs) for homology search and alignment are constructed. Second, if both target and template are in the set of sequences to be aligned, target-template alignment can be directly obtained in the context of resulting MSA.

Given a set of sequences, MSA methods aim to construct an alignment in which columns represent evolutionary (structurally) equivalent residues. Although in theory dynamic programming algorithms for pairwise alignment can be extended for computing an optimal alignment of multiple sequences, they are too computationally demanding to be practically useful. As a result, most current techniques use various approximations and heuristics. These methods are not guaranteed to derive an optimal MSA, but in practice they can often produce good alignments using modest computational resources. Most of the modern MSA tools use heuristics known as *progressive alignment*. In this strategy, an approximate alignment guide tree is first constructed based on pairwise sequence similarities. Using this guide tree, the most closely related sequences are aligned first. Next, these subalignments are aligned to each other until all sequences are incorporated into MSA.

Thus, the progressive alignment substitutes the task of MSA into a series of pairwise alignments. ClustalW (24), one of the earliest programs and still a very popular choice, is a representative of progressive alignment methods. The main drawback of the progressive alignment strategy is that errors made early on in the construction of guide trees or pairwise alignments (especially in the initial stages) cannot be corrected and tend to propagate in the entire alignment. Thus, ClustalW can produce good alignments for closely related sequences, but alignments for divergent sequence sets may be poor. Therefore, a number of approaches have been devised to avoid the problems associated with an application of progressive alignment. For more details on recent methodological and algorithmic improvements, the reader is referred to recent reviews (25–27). Here, only several methods that had been reported to perform well in various benchmarks are briefly discussed.

One of the strategies to deal with errors in progressive alignments is to perform an iterative refinement. MAFFT (28) and MUSCLE (29) are two representative MSA methods that use such an iterative refinement strategy. Both are very fast and flexible: depending on the number of sequences the balance between the accuracy and speed can be easily adjusted.

Another strategy to improve initial progressive alignments is to use consistency information. The consistency concept is very simple. Let us suppose that we have three sequences (A, B, and C) and the corresponding pairwise alignments. If residue A_i is aligned to residue B_j and residue B_j is aligned to residue C_k, this implies that in A-C alignment A_i should be aligned with C_k. In other words, pairwise alignments induced by multiple alignments should be consistent. This transitivity condition is taken into account in scoring the alignment of two sequences (or group of sequences) by considering the information of their alignment to other sequences not involved in pairwise merge. T-coffee (30) and ProbCons (31) are examples of methods that make use of consistency-based scoring. In general, consistency-based methods are more accurate than those based on iterative refinement, but are more computationally demanding. However, in some cases, such as in recent versions of MAFFT (32), a simpler version of consistency measure has helped to keep the program fast. While being much faster, MAFFT now rivals the accuracy of both T-coffee and ProbCons (33).

Other strategies to improve the alignment accuracy include combination of several methods, as in M-coffee (34), or the incorporation of additional information. The additional information may be evolutionary (e.g., additional homologous sequences) or structural, since a 3D structure evolves more slowly than a sequence. For example, the MAFFT package has an option to add close homologs (35) detected using a BLAST search to improve the alignment accuracy of the initially submitted set of multiple sequences. One of the recently developed programs, PROMALS (36), uses a number of sources for additional information. First, it detects

Table 2
Multiple sequence alignment methods

Method	Type of information used	Address
ClustalW	Sequence	http://www.clustal.org/
MAFFT MAFFT-homologs	Sequence Sequence + homologs	http://mafft.cbrc.jp/alignment/software/
MUSCLE	Sequence	http://www.drive5.com/muscle/, http://www.ebi.ac.uk/Tools/muscle/index.html
ProbCons	Sequence	http://probcons.stanford.edu/
PROMALS	Sequence + homologs + SS[a]	http://prodata.swemd.edu/promals/
PROMALS3D	Sequence + homologs + SS[a] + 3D[b]	http://prodata.swemd.edu/promals3d/
T-coffee M-coffee 3DCoffee/Expresso	Sequence Consensus Sequence + 3D[b]	http://www.tcoffee.org/

[a] Secondary structure
[b] Three-dimensional structure

sequence homologs with PSI-BLAST and uses the obtained profiles to predict secondary structure. Next, profile–profile comparisons enhanced with predicted secondary structures are used in the alignment processes. If the 3D structural information is available, it can also be combined with sequence data within the consistency framework to improve accuracy of MSAs. The automatic incorporation of the available 3D structural information has been implemented in programs such as PROMALS3D (37), a successor of PROMALS, and 3DCoffee/Expresso (38, 39).

The MSA methods discussed here are summarized in Table 2. It should be emphasized that, depending on the situation, different MSA methods may be optimal. In general, when sequences to be aligned are fairly similar (over 35% sequence identity; the “daylight” zone), any method is likely to produce an accurate alignment. The alignment accuracy starts deteriorating when sequence similarity falls into the “twilight” zone (<25%) and/or the number of sequences is small. In such cases, despite being slower, methods that use additional sequence and/or structure information may be more suitable.

4. Hybrid Methods, Fully Integrated Automatic Servers and Meta-servers

A growing number of contemporary modeling methods derive sequence–structure mapping (alignment) by combining multiple sequence and structure features. Moreover, often a number of alignments with multiple templates or their fragments are considered simultaneously in deriving protein models based on homology. Even the concept of sequence–structure alignment sometimes becomes blurred because the derived final model cannot be easily attributed to one or more explicit sequence–structure alignments. Another popular trend is the use of meta-approaches. By combining results of different algorithms, these approaches attempt to identify the closest structural templates and the most accurate sequence–structure alignments. It would be impossible to provide an in-depth description for each of the multitude of methods presently available. Therefore, here only several popular methods that performed well in recent international blind trials of protein structure prediction known as CASP (40), and at the time of writing were accessible as public Web servers on the Internet (Table 3), are briefly discussed.

I-TASSER (41), one of the top hybrid protein structure modeling methods, uses combined results from multiple profile–profile comparison algorithms to detect suitable structural templates and to generate sequence–structure alignments. During next steps, the continuous fragments of initial alignments are reassembled into full-length models using iterative rounds of structure construction, model assessment, and refinement. In a sense, I-TASSER represents a meta-server for distant homology detection combined with techniques for structure simulation and evaluation. A similar approach is used in pro-Sp3-TASSER (42) with the difference being mostly in the methods used for the construction of initial sequence–structure alignments and model evaluation. The SAM-T08 server (13) uses the HMM-based sequence comparison

Table 3
Hybrid methods, fully integrated protein modeling servers and meta-servers

Method	Type	Address
I-TASSER	Server	http://zhanglab.ccmb.med.umich.edu/I-TASSER/
Pro-sp3-TASSER	Server	http://cssb.biology.gatech.edu/skolnick/webservice/pro-sp3-TASSER/
Robbeta	Server	http://robetta.bakerlab.org/
Phyre	Server	http://www.sbg.bio.ic.ac.uk/~phyre/
MULTICOM	Server	http://casp.rnet.missouri.edu/multicom_3d.html
SAM-T08	Server	http://compbio.soe.ucsc.edu/SAM_T08/T08-query.html
pGenTHREADER	Server	http://bioinf.cs.ucl.ac.uk/psipred/
GeneSilico	Meta-server	http://genesilico.pl/meta2/
Pcons.net	Meta-server	http://pcons.net/

enriched with predicted local structural features to detect templates and to generate several alignments with each of them. Models are then assembled using the templates, the local structure predictions, the distance constraints, and the contact predictions. Robetta (43) in the homology modeling regime uses profile-based methods to detect templates. Next, an ensemble of sequence–structure alignments is generated, followed by structure simulation and refinement. Perhaps the most important difference between Robetta and other methods discussed here is that in structure simulation it uses extensive conformational sampling coupled with physics-based all atom refinement. However, this means that much larger computational resources are needed. Phyre (44) is based on an ensemble of algorithmic variants for remote homology detection (essentially an in-house meta-server) combined with model construction and selection. MULTICOM (45) implements a combination of data at multiple modeling levels including templates, alignments, and models. pGenTHREADER (46), the latest implementation of GenTHREADER (47), the classical threading method, uses a linear combination of profile–profile alignments with secondary-structure-specific gap-penalties and classic pair- and solvation potentials.

There are also a number of meta-servers that apply a consensus approach either to select a best model or to construct a consensus model using the results obtained from different methods. GeneSilico (48) and Pcons.net (49) are among those meta-servers that are being continuously developed and updated.

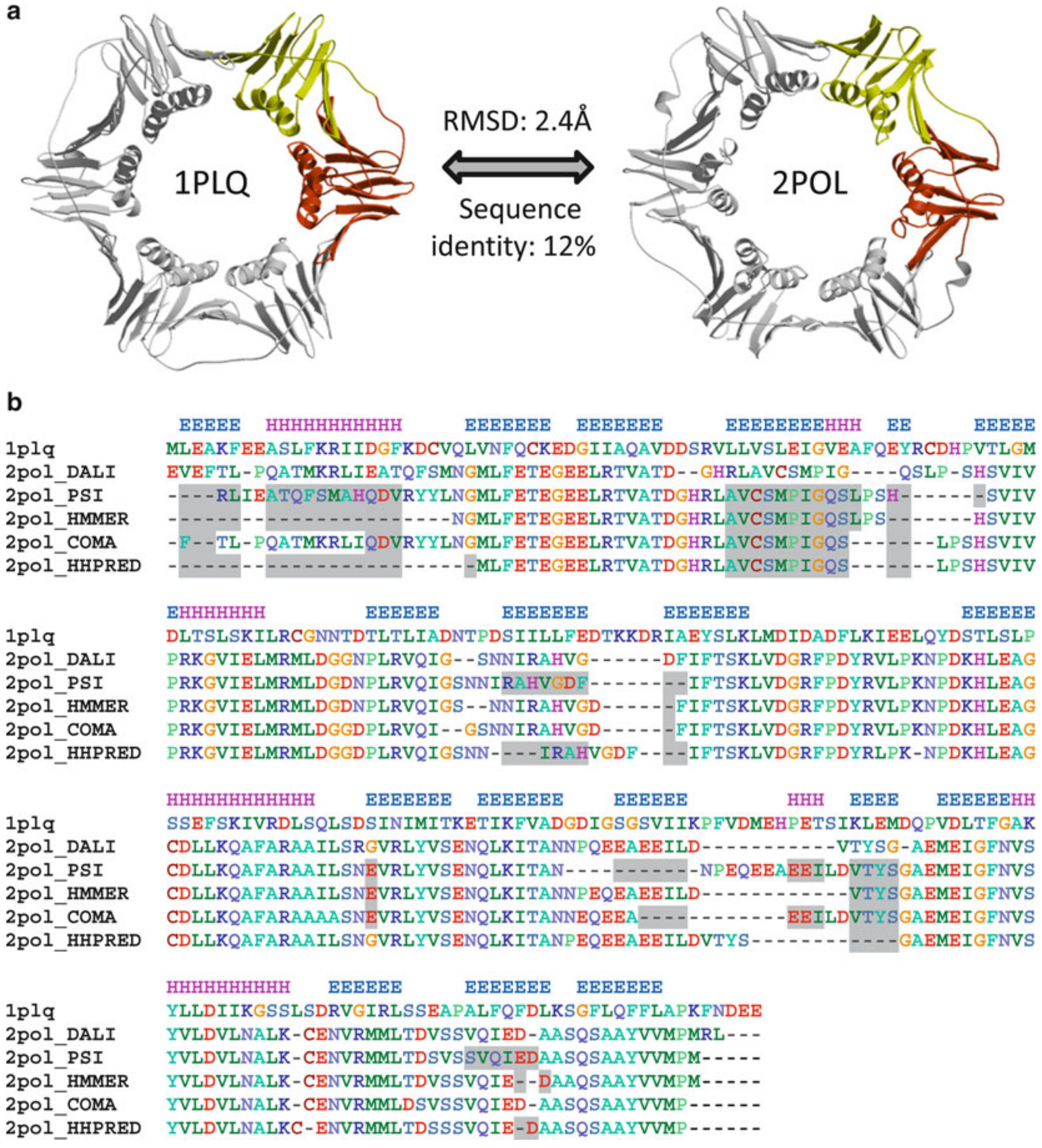
Although now there are a large number of fully automated methods for homology modeling, one should keep in mind that the use of a more sophisticated procedure does not necessarily guarantee a better quality of the final model. It has been observed over and over again that no matter which template-based techniques are used to arrive at the final model, the largest contribution to its quality comes from the optimal template selection and the improvement of sequence–structure alignment (50). Therefore, a method that generates accurate alignments may sometimes outperform those with multiple layers of complexity. A vivid example of that was provided in CASP8 (51) by HHpred (52), a server implementation of the HHsearch method (16). HHpred was ranked among top servers despite the fact that it was neither exploring alternative alignments, nor reassembling structures from fragments, nor using additional structural features and optimization procedures. At the same time, HHpred was orders of magnitude faster than any other of the top servers. When just single domain targets were considered, it was second to only I-TASSER (52). This example clearly shows that the optimal selection of template(s) and especially the accuracy of the sequence–structure alignment are of paramount importance.

5. Accuracy of the Sequence–Structure Mapping

The construction of the initial sequence–structure alignment either through database searching or by using MSA methods on a predefined set of sequences is usually straightforward. However, unless the alignment between the modeling target and the structural template(s) is trivial (sequence identity over 40–50% and no or only few gaps), its reliability should be carefully evaluated.

5.1. Non-trivial Relationship Between Sequence Similarity, Statistical Significance, and Alignment Accuracy

In general, with the increase of evolutionary distance, both structures and sequences of homologous proteins become less similar, making homology detection more challenging. Intuition suggests that a lower sequence similarity might also be expected to result in the decreased accuracy of sequence–structure mapping. However, it turns out that the relationship between sequence similarity, statistical significance of the alignment, and its accuracy is not simple. In distant homology cases, sequence similarity between the target and template by itself is a poor predictor of alignment accuracy, because most commonly, the target-template pairwise alignment is derived in the context of multiple aligned sequences (sequence profiles, HMMs, or explicitly derived MSAs). Therefore, the number and the similarity distribution of additional homologous sequences seem to play a major role in determining both the sensitivity of homology detection and the overall alignment accuracy. As in crossing a river by hopping from one stone to the next, intermediate homologs may serve as "bridging stones" helping to link the target and the template (53). It is apparent that the more intermediate sequences are available and the smoother is their similarity transition, the more accurate alignment may be expected. A higher statistical significance of an alignment usually means a higher alignment accuracy. However, in distant homology cases, it would be a big mistake to think that highly statistically significant alignments are always highly accurate. This is illustrated in Fig. 3 with a distantly homologous pair of DNA sliding clamps. While BLAST is not able to detect this relationship at all, PSI-BLAST, HMMER, COMA, and HHpred, representing both profile- and HMM-based methods, detect it with a very high confidence. However, all of the corresponding alignments show significant discrepancies with the "gold standard" alignment derived from structure comparison with DaliLite (54). In other words, there is no strict dependency between alignment accuracy and homology detection ability. At the same time, this example seems to support observations (e.g., refs. 17, 55) that profile–profile alignments are in general more accurate than profile–sequence alignments. Alignment accuracy may also depend on inherent properties of a protein family. In particular, it has been observed that families with a high diversity of confident homologs tend to produce lower quality profile–profile alignments



```
              EEEEE  HHHHHHHHHHH     EEEEEEE  EEEEEEE     EEEEEEEEEHHH EE     EEEEE
1plq         MLEAKFEEASLFKRIIDGFKDCVQLVNFQCKEDGIIAQAVDDSRVLLVSLEIGVEAFQEYRCDHPVTLGM
2pol_DALI    EVEFTL-PQATMKRLIEATQFSMNGMLFETEGEELRTVATD--GHRLAVCSMPIG----QSLP-SHSVIV
2pol_PSI     ----RLIEATQFSMAHQDVRYYLNGMLFETEGEELRTVATDGHRLAVCSMPIGQSLPSH-------SVIV
2pol_HMMER   -----------------------NGMLFETEGEELRTVATDGHRLAVCSMPIGQSLPS-------HSVIV
2pol_COMA    -F--TL-PQATMKRLIQDVRYYLNGMLFETEGEELRTVATDGHRLAVCSMPIGQS-------LPSHSVIV
2pol_HHPRED  -------------------------MLFETEGEELRTVATDGHRLAVCSMPIGQS-------LPSHSVIV

             EHHHHHHH        EEEEEE     EEEEEEE      EEEEEEE                 EEEEEE
1plq         DLTSLSKILRCGNNTDTLTLIADNTPDSIILLFEDTKKDRIAEYSLKLMDIDADFLKIEELQYDSTLSLP
2pol_DALI    PRKGVIELMRMLDGGNPLRVQIG--SNNIRAHVG------DFIFTSKLVDGRFPDYRVLPKNPDKHLEAG
2pol_PSI     PRKGVIELMRMLDGDNPLRVQIGSNNIRAHVGDF--------IFTSKLVDGRFPDYRVLPKNPDKHLEAG
2pol_HMMER   PRKGVIELMRMLDGDNPLRVQIGS--NNIRAHVGD------FIFTSKLVDGRFPDYRVLPKNPDKHLEAG
2pol_COMA    PRKGVIELMRMLDGGDPLRVQI--GSNNIRAHVGD------FIFTSKLVDGRFPDYRVLPKNPDKHLEAG
2pol_HHPRED  PRKGVIELMRMLDGGDPLRVQIGSNN----IRAHVGDF----IFTSKLVDGRFPDYRLPK-NPDKHLEAG

             HHHHHHHHHHHH    EEEEEEE  EEEEEEE    EEEEEE        HHH  EEEE   EEEEEEHH
1plq         SSEFSKIVRDLSQLSDSINIMITKETIKFVADGDIGSGSVIIKPFVDMEHPETSIKLEMDQPVDLTFGAK
2pol_DALI    CDLLKQAFARAAILSRGVRLYVSENQLKITANNPQEEAEEILD-----------VTYSG-AEMEIGFNVS
2pol_PSI     CDLLKQAFARAAILSNEVRLYVSENQLKITAN-----------NPEQEEAEEILDVTYSGAEMEIGFNVS
2pol_HMMER   CDLLKQAFARAAILSNEVRLYVSENQLKITANNPEQEAEEILD------------VTYSGAEMEIGFNVS
2pol_COMA    CDLLKQAFARAAAASNEVRLYVSENQLKITANNEQEEA------------EEILDVTYSGAEMEIGFNVS
2pol_HHPRED  CDLLKQAFARAAILSNGVRLYVSENQLKITANPEQEEAEEILDVTYS-----------GAEMEIGFNVS

             HHHHHHHHHH   EEEEEE     EEEEEEE  EEEEEEE
1plq         YLLDIIKGSSLSDRVGIRLSSEAPALFQFDLKSGFLQFFLAPKFNDEE
2pol_DALI    YVLDVLNALK-CENVRMMLTDVSSVQIED-AASQSAAYVVMPMRL---
2pol_PSI     YVLDVLNALK-CENVRMMLTDSVSSVQIEDAASQSAAYVVMPM-----
2pol_HMMER   YVLDVLNALK-CENVRMMLTDVSSVQIE--DAAQSAAYVVMPM-----
2pol_COMA    YVLDVLNALK-CENVRMMLDSVSSVQIED-AASQSAAYVVMP------
2pol_HHPRED  YVLDVLNALKC-ENVRMMLTDSSSVQIE-DAASQSAAYVVMP------
```

Fig. 3. Structure and sequence comparison of distantly homologous DNA sliding clamps from yeast (PDB code: 1plq) and *E. coli* (2pol). (**a**) Their 3D structures are similar despite sharing only 12% identical residues. (**b**) Comparison of DaliLite (DALI) structure-based alignment between 1plq and 2pol with the alignments produced by PSI-BLAST (PSI; *E* value = 3e–30), HHMER (*E* value = 2e–32), COMA (*E* value = 3e–13), and HHpred (probability = 99%). Alignments were obtained by searching PDB with 1plq sequence profiles (HMMs) that were obtained by running up to five iterations of PSI-BLAST (jackhmmer in the case of HMMER) with the 1plq sequence as a query against the filtered "nr" database. For easier comparison, columns corresponding to gaps in 1plq sequence were removed from all the alignments. Alignment positions showing discrepancies between DaliLite and each of the methods are shaded. Only positions corresponding to secondary structure elements ("H," helix, "E," strand) in 1plq were considered. The best agreement with the DaliLite alignment is shown by COMA, followed by HMMER, HHsearch, and PSI-BLAST.

with their remote relatives (56). However, this lower alignment accuracy cannot be improved when the most distant members of these families are excluded from their profiles. On the contrary, the presence of more diverse members has been found to result in more accurate alignments. This implies that the growth of the sequence databases should automatically result in more accurate alignments for the same level of sequence identities. However, this conclusion appears to hold only for confident high-quality homologous sequences. The inclusion of spurious contaminating sequences or even low-quality metagenomic sequences may negatively impact the target-template alignment accuracy (57).

5.2. Estimation of the Region-Specific Alignment Reliability

Sequence–structure alignment by itself does not tell which regions are aligned reliably (provide the correct residue mapping) and which ones may require adjustment. Therefore, to improve an alignment, the first task is to identify those alignment regions that can be trusted. Once the reliable regions are identified, the remaining alignment stretches can be either subjected to refinement or (if a significant conformational change is anticipated) rebuilding using different templates or template fragments.

The earliest methods for identification of reliable alignment regions (58–60) were focusing on pairwise sequence alignments that are largely irrelevant for the present day comparative modeling approaches. For target-template alignments constructed in the context of sequence profile- (or HMM)-based methods, several approaches were shown to be useful. Perhaps the simplest approach is based on the scores of individual positions within the profile–profile alignment. It was shown that the regions containing high scoring positions correlate well with the correctness of their alignment (61). More commonly, the positional reliability of sequence–structure alignments is estimated by assessing the region-specific alignment stability. There are two general strategies to generate sufficient alignment variability from which stable alignment regions can then be identified. The first strategy relies on a single method to generate alignment variability. This has been done either by using suboptimal alignments derived from the same sequence data (62, 63) or by diversifying alignments through the sampling of the available sequence space of homologs as in PSI-BLAST-ISS (64). The second strategy is based on the use of multiple methods to generate corresponding alignments followed by the analysis of alignment regions that do or do not agree between these different methods (65). Independently of which strategy is used, a strong consensus is considered to indicate reliably aligned regions. The lack of consensus may be caused by different reasons such as weak sequence conservation, insertions/deletions, or a significant conformational change. Figures 4 and 5 illustrate two typical situations resulting in unreliable alignment regions delineated with PSI-BLAST-ISS (64). In Fig. 4, the region of unreliable alignment coincides with a significant difference in orientation of corresponding α-helices.

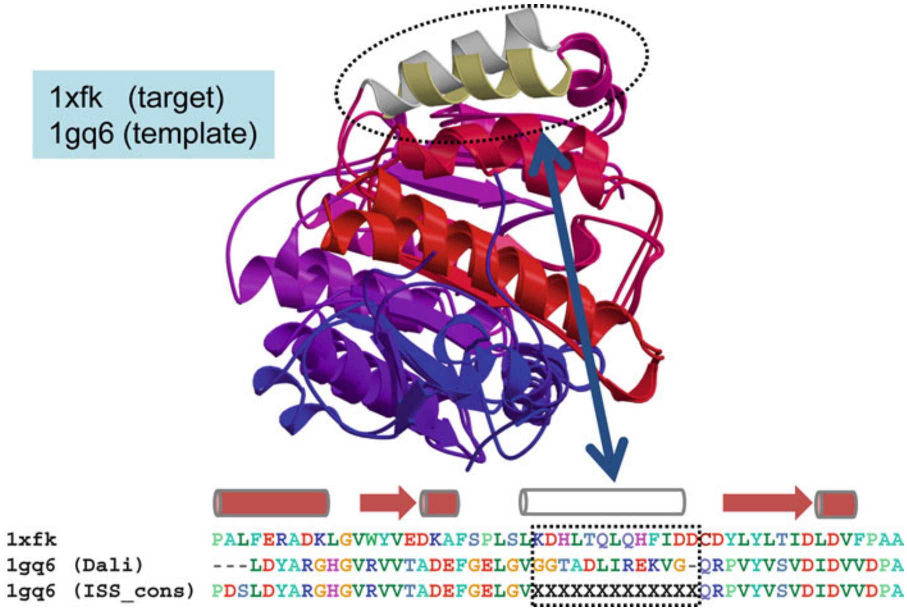
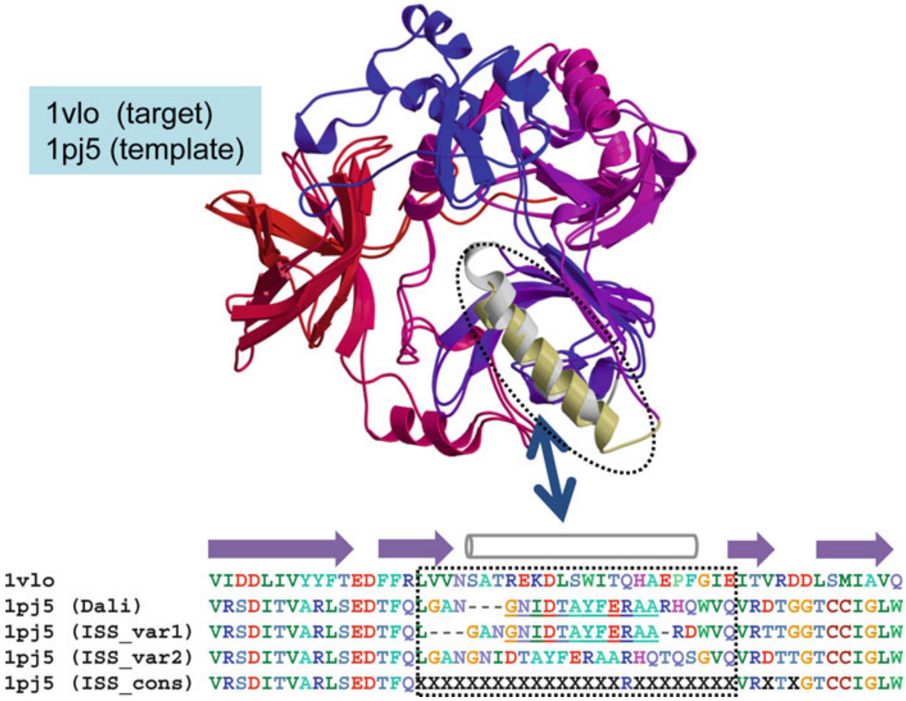

1xfk (target)
1gq6 (template)
1xfk PALFERADKLGVWYVEDKAFSPLSLKDHLTQLQHFIDDCDYLYLTIDLDVFPAA
1gq6 (Dali) ---LDYARGHGVRVVTADEFGELGVGGTADLIREKVG-QRPVYVSVDIDVVDPA
1gq6 (ISS_cons) PDSLDYARGHGVRVVTADEFGELGVXXXXXXXXXXXXXQRPVYVSVDIDVVDPA

Fig. 4. Example of an unreliable alignment region corresponding to a structurally divergent motif. This motif is represented by an α-helix shown in *light colors* (enclosed in an *ellipse*) in superimposed structures of the modeling target (PDB code: 1xfk) and the template (1gq6). Below, the 1xfk is aligned with 1gq6 according to both structural correspondence (Dali) and a consensus alignment produced by PSI-BLAST-ISS (ISS_cons). "X" denotes positions lacking the consensus. The secondary structure of the 1xft is shown above the alignment. Figure adopted from ref. 64.

The unreliable region in Fig. 5 corresponds to a structurally conserved α-helix, which, however, has an insertion at one end and a deletion at the other end. Aligning this region correctly for sequence-based methods is difficult because of their tendency to cancel out the insertion and the deletion adjacent to the α-helix by shifting (incorrectly) its sequence. Yet, among individual alignment variants suggested by PSI-BLAST-ISS, there is one that corresponds to the structurally accurate alignment.

5.3. Improvement of Sequence–Structure Alignments

Although it is useful to know which regions in the model may be misaligned, the desirable goal is to achieve the highest possible sequence–structure alignment accuracy. Since sequence features alone are of little help in resolving alignment ambiguities, the often used recipe is to apply the assessment of alternative alignments in the context of a corresponding 3D model. To do this, one needs some sort of diagnostic tool for evaluating model quality in a region-specific way. Until recently, there were only few such tools available for performing the task. For quite some time, classical methods, ProSA (66) and Verify3D (67), have been popular choices for both the overall (global) and the position-specific (local) protein structure quality assessment. An important stimulus for development of new methods has appeared a few years back with the introduction



```
1vlo            VIDDLIVYYFTEDFFRLVVNSATREKDLSWITQHAEPFGIEITVRDDLSMIAVQ
1pj5 (Dali)     VRSDITVARLSEDTFQLGAN---GNIDTAYFERAARHQWVQVRDTGGTCCIGLW
1pj5 (ISS_var1) VRSDITVARLSEDTFQL---GANGNIDTAYFERAA-RDWVQVRTTGGTCCIGLW
1pj5 (ISS_var2) VRSDITVARLSEDTFQLGANGNIDTAYFERAARHQTQSGVQVRDTTGTCCIGLW
1pj5 (ISS_cons) VRSDITVARLSEDTFQXXXXXXXXXXXXXXXXRXXXXXXXXVRXTXGTCCIGLW
```

Fig. 5. Example of an unreliable alignment region corresponding to a structurally conserved motif surrounded with variable adjacent regions. The motif includes a structurally conserved α-helix (shown in *light color* and marked by an *ellipse*) in superimposed structures of the modeling target (PDB code: 1vlo) and the template (1pj5). However, one of the adjacent loops has an insertion and the other one has a deletion. The alignment shows structural correspondence (Dali), the PSI-BLAST-ISS consensus alignment (cons), and two individual variants (var1 and var2). "X" denotes positions lacking the consensus. One of the variants (var1) reproduces most of the structure-based mapping for the conserved α-helix (sequence underlined). Figure adopted from ref. 64.

of the model quality assessment category in CASP experiments (68). Quite a few approaches for estimating both the global and the local quality of a protein model have been developed since. Clustering- or consensus-based methods currently are the most accurate and the best such methods show a respectable accuracy in predicting global model quality (69). However, to work well, they require a large ensemble of models generated by different methods. Unfortunately, while this setting is natural for CASP, it has little to do with real modeling projects. In addition, even clustering-based methods perform significantly worse in the local model quality assessment mode, which is critical for the alignment improvement task. Nevertheless, promising new methods such as QMEAN (70, 71) that are capable of assessing position-specific quality of individual models have also emerged.

CASP results revealed that the systematic identification of correct alignment variants in unreliable regions is still difficult. Analysis of common alignment failures showed that the error-prone regions often share similar traits (72, 73). These regions often correspond to peripheral secondary structure elements (β-strands at the edge of β-sheets, highly solvent-exposed α-helices) that are under lesser structural/energy constraints than the structural core. Another feature that frequently correlates with alignment errors is the appearance or disappearance of small structural "defects" such as β-bulges. Arguably, alternative alignment variants in such error-prone regions have subtle energy differences and therefore are difficult to rank correctly. In addition, template structure is just an approximation of the native structure of modeling target. Inevitably, this introduces additional error during the evaluation of alternative alignments, and because of that even an effective assessment technique might fail. It is intuitively apparent that the more accurately is the protein main chain modeled, the easier it should be to distinguish the correct residue mapping from the erroneous one. In other words, perhaps the most effective, although computationally expensive, way to identify the native alignment would be to test an ensemble of alignments by performing simultaneous refinement for each of the corresponding models. In fact, the sampling of alignment variants coupled with all-atom refinement has been tested at CASP, with impressive results for some modeling targets (74). Less successful results were attributed to insufficient sampling and imperfect energy estimation (74).

Thus, the accurate mapping of sequence onto structure remains one of the important bottlenecks in homology modeling. Although there are signs of improvement, a lot more will have to be done in developing more effective approaches for sampling alignments and conformations, together with better methods for the local model quality estimation.

6. Practical Guide for Sequence–Structure Alignment

The following is a brief description of practical steps for aligning a sequence to known structure(s), estimating the reliability of alignment regions and selecting the best alignment. To a large degree, this rough guide is based on an updated protocol (73) used to achieve the top-ranked results in the homology (template-based) modeling category during the CASP8 experiment (75). The flowchart depicting main steps in sequence–structure alignment is presented in Fig. 6.

6.1. Searching for Structural Templates and Constructing Initial Alignments

First, it is useful to find out what is the level of difficulty for generating accurate sequence–structure alignment. The initial estimate can be made, once it is known if there are closely related experimental 3D structures available. If so, how similar their sequences are to the protein of interest? How many structures are available? How many additional homologs can be detected in sequence databases and how closely they are related to the target?

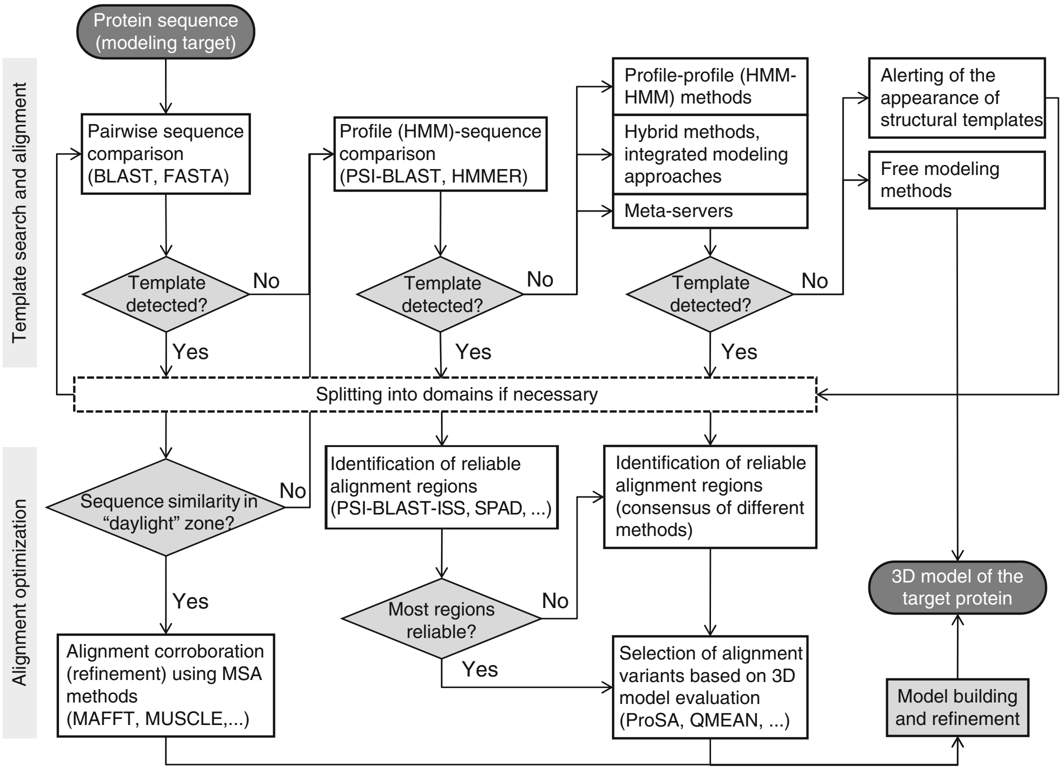

Protein sequence (modeling target)
Template search and alignment
Pairwise sequence comparison (BLAST, FASTA)
Profile (HMM)-sequence comparison (PSI-BLAST, HMMER)
Profile-profile (HMM-HMM) methods
Hybrid methods, integrated modeling approaches
Meta-servers
Alerting of the appearance of structural templates
Free modeling methods
Template detected?
No
Yes
Template detected?
No
Yes
Template detected?
No
Yes
Splitting into domains if necessary
Alignment optimization
Sequence similarity in "daylight" zone?
No
Yes
Identification of reliable alignment regions (PSI-BLAST-ISS, SPAD, ...)
Identification of reliable alignment regions (consensus of different methods)
Most regions reliable?
No
Yes
3D model of the target protein
Alignment corroboration (refinement) using MSA methods (MAFFT, MUSCLE,...)
Selection of alignment variants based on 3D model evaluation (ProSA, QMEAN, ...)
Model building and refinement

Fig. 6. Flowchart of major steps in sequence to structure alignment.

The best idea is to start with a simple sequence search using BLAST (3). It is useful to have the BLAST suite of programs including both BLAST and PSI-BLAST as well as protein sequence databases installed locally. This provides an increased flexibility in using these programs. The BLAST program suite and sequence databases can be obtained from the NCBI FTP site at ftp://ftp.ncbi.nlm.nih.gov/blast/. Sequence databases at NCBI are updated daily and can be retrieved automatically using the *update_blastdb.pl* script, which is provided freely as part of the BLAST documentation at NCBI. For the local installation, it is important to have at least two protein sequence databases: nonredundant sequence database (nr) containing all nonredundant protein sequences (except those from metagenomic projects) and the PDB sequence database (pdbaa), which contains protein sequences of known 3D structures. The latter sequences are also available for downloading directly from PDB (http://www.pdb.org). Since the nonredundant (nr) sequence database is huge and continues to grow fast, it is advisable to have several smaller versions of this database with very similar sequences removed. It is a common practice to remove sequences up to 90, 80, and 70% identical to each other. This helps to reduce the database size significantly without negatively affecting homology search results. The filtering of sequence databases can be done with clustering tools such as CD-HIT (76). If the filtering of the locally installed "nr" database turns out to be too computationally expensive, the user may choose to download preprocessed UniRef sequence databases with the reduced levels of redundancy from UniProt (http://www.uniprot.org/). These sequence databases are also aiming at a complete coverage of sequence space. At present, UniRef100, UniRef90, and UniRef50 filtered correspondingly at 100, 90, and 50% sequence identity, are available. Alternatively, the user can run both BLAST and PSI-BLAST sequence searches using web servers either at NCBI (http://blast.ncbi.nlm.nih.gov/), EBI (http://www.ebi.ac.uk/Tools/sss/), or at many other locations on the Internet.

The results of BLAST search against PDB sequences give an approximate estimate of the difficulty to derive an accurate sequence–structure alignment. During the simplest scenario, BLAST search detects a PDB sequence with a statistically significant expectation value (E value < 0.001) and a relatively high sequence similarity (over 40% sequence identity) to the modeling target. In such case, the homologous relationship is obvious and the alignment may be structurally optimal. However, even if such pairwise alignment does not have any gaps, it is still recommended to substantiate the alignment with methods that rely on information derived from multiple sequences. This can be done by collecting additional close sequence homologs with BLAST, pooling them together with target and template sequences and aligning with one of the fast MSA methods such as MAFFT (28) or MUSCLE (29). If sequence identity is lower than 40% and there are gaps, the alignment almost certainly will need some adjustments such as the placement of the gaps or their boundaries. In such case, an MSA might also help to refine the target-template alignment. However, if the sequence similarity enters the "twilight" zone, MSA methods that use additional information (predicted secondary structure, 3D structural information) such as PROMALS/PROMALS3D (36, 37) and 3DCoffee/Expresso (38, 39) might be more appropriate. The use of PSI-BLAST and other profile (HMM)-based methods is also recommended in more distant homology cases (see below).

If no PDB sequences with statistically significant E values are detected with BLAST, more sensitive methods such as PSI-BLAST should be used next. The power of PSI-BLAST is in rich sequence profiles generated from aligned multiple homologous sequences. The PDB sequence database is too small to perform the iterative PSI-BLAST searches against it directly. Usually, potential structural templates are detected and aligned with the target sequence using the so-called PDB-BLAST procedure. It involves performing several iterations of PSI-BLAST search against a large sequence database (e.g., "nr" or its derivatives) and then using the constructed profile to run the last iteration against the PDB sequence database.

It is worthwhile to make several PDB-BLAST runs, every time generating a more inclusive profile by increasing the number of iterations against the "nr" database or its derivatives. The change in the number of detected PDB sequences and the corresponding *E* values will give an approximate estimate of evolutionary distance between the target sequence and the confidently (*E* value < 0.001) detected structures. If PSI-BLAST and sequence databases are not installed locally, it is still possible to perform PDB-BLAST-like searches using the NCBI BLAST server through several manual steps. Automatic PDB-BLAST searches can be performed both locally and remotely (at NCBI) using Re-searcher (77). Note that PSI-BLAST is not the only available option. Recently, an iterative procedure similar to that in PSI-BLAST was implemented in HMMER (http://hmmer.org/). With the reported high speed and sensitivity, the iterative HMMER3 procedure (jackhmmer) is at least as good as PSI-BLAST.

If sequence searches with profiles (PSI-BLAST) or HMMs (e.g., HMMER) do not reveal any obvious structural homologs, it does not necessarily mean that they are absent from the PDB. It may be that the evolutionary relationship is too distant to be detected by profile (HMM)–sequence comparisons. In such case the obvious next step is to turn to the even more sensitive profile–profile, HMM–HMM, or hybrid sequence–structure methods. There are now a large number of such methods available and only a small fraction is listed in Tables 2 and 3. One of the best choices to start with is HHsearch (16), a very fast and one of the most sensitive homology detection methods. Based on HMM–HMM comparison, HHsearch is available both as a standalone toolkit and as part of the HHpred web server (78). Other sensitive alternatives to HHsearch include PRC (19, 79), COMA (17, 80), COMPASS (15, 81), and PROCAIN (22, 82). Both HHpred and COMA servers also have a useful option to produce 3D models based on the reported sequence–structure alignments. Among the fully integrated modeling approaches I-TASSER (41) at present is clearly the best choice. As many other integrated hybrid modeling methods it will return the final 3D model, which may not necessarily correspond to any of the initial sequence–structure alignments used. Meta-servers such as Genesilico (48) or Pcons.net (49) may also be useful, since they provide results from several methods simultaneously. In general, many new methods are continuously reported, making it difficult to select the best methods at a given time. It may be instructive to check the server results during latest CASP experiments (http://www.predictioncenter.org/). However, not always well-performing methods at CASP are available as public servers and not all well-performing methods take part in CASP. Independently of which servers you use, check when the databases were last updated; even the best methods will likely perform poorly on old sequence and structure databases.

Initial template search results usually reveal the domain composition of the modeling target. If it is a multidomain protein, it may be beneficial or even necessary to partition the sequence into chunks corresponding to individual domains. First, individual protein domains may have a closer relationship with different structural templates. In such case, treating domains individually may improve the selection of templates and/or the accuracy of sequence–structure alignments. Second, the partition of the sequence into domains may help to avoid homologous over-extension (HOE), an important source of errors in iterative profile-based searches (83). This error occurs when the alignment initially covering only homologous domains over the course of iterations is extended into nonhomologous regions.

6.2. Estimation of Position-Dependent Alignment Reliability

Typically, sequence–structure alignments produced within the "twilight" or "midnight" zones of sequence similarity will have inaccuracies. However, a visual inspection at this level of sequence similarity is virtually useless in spotting them. How then to distinguish alignment regions that are reliable from those that may be incorrect and will likely require refinement? One of the options is to use alignment stability as an indicator of reliability. One of the available tools that use this idea is PSI-BLAST-ISS (64). It is based on multiple PSI-BLAST searches with different yet related queries. PSI-BLAST-ISS results simultaneously provide several types of information: (1) automatically detected structural templates and corresponding alignments, (2) data suggesting which one of the templates may be the closest to the target, and (3) the region-specific alignment reliability indication for each of the templates. The drawback of PSI-BLAST-ISS is that it takes time to run all the PSI-BLAST searches (typically 50–100) and that parameter settings may need adjustment depending on the target. PSI-BLAST-ISS is also useless in cases of very distant homology, when PSI-BLAST is not sensitive enough to detect templates. In such cases, perhaps the simplest way to estimate regional alignment reliability is to use the agreement between the sequence–structure alignments produced by different methods. However, different methods may provide alignments or build models using different templates. To cope with this potential heterogeneity of results, it is useful to convert all the outputs into a common format such as 3D structure. Nowadays, many methods generate 3D models as the final output or at least provide an option to construct models using the resulting alignments. However, if models are unavailable, they can be easily constructed from sequence–structure alignments using one of the modeling tools such as MODELLER (84), Nest (85), and Swiss-PdbViewer (86). There are also web servers for converting sequence–structure alignments to structural models. For example, "alignment mode" of SwissModel (86), one of the popular modeling servers, can be used for this purpose. Comparison of the resulting models with one of the representative templates provides the underlying sequence–structure mappings. After that, all the pairwise alignments can be merged into a single PSI-BLAST-ISS-like alignment, in which a template is aligned to the target sequence variants corresponding to different models. Both pairwise structure comparisons and merging of the corresponding alignments can be easily performed in one step using the *dali_sp.pl* wrapper (http://www.ibt.lt/bioinformatics/software/) for DaliLite (54). Just like in the case of PSI-BLAST-ISS, the agreement between different methods tends to indicate reliable regions of the alignment, while the lack of consistency points to the need of further analysis.

6.3. Improving Alignments

If the sequence of the modeling target is aligned reliably with all the structurally conserved regions of the template(s) the sequence–structure mapping is done. In such case, the final quality of the homology model will be determined by other steps such as the ability to accurately model variable regions and to drive the model structure closer to the native one. The tricky part begins with the regions that are not reliably aligned, because first it is important to understand whether the uncertainty is caused by the conformational change or simply by the lack of sequence conservation. Only if there are hints from available template(s) that the region is structurally conserved, there is a good chance to identify structurally/evolutionary meaningful alignment for this region without modifying the template backbone. In that case, the assessment of sequence–structure mapping within the context of 3D structure (i.e., assessing a structural model based on a particular sequence–structure alignment) perhaps is the most promising. Structure quality evaluation methods such as ProSA (66, 87) or QMEAN (70, 71) can help identify the correct alignment by estimating both the overall and region-specific model quality. Often, the problem with the evaluation of models based on alternative alignment variants is the noisiness of the results. More often than not, the evaluation results do not show a clear preference towards a particular alignment variant. One way to deal with the noisy signal is to include additional homologs of the target sequence into the analysis. The homologs should be selected such that their alignment with the target sequence would be unambiguous. The consensus of evaluation results of models based on alternative sequence–structure alignments for multiple family members may help rank the alignment variants more effectively. However, the consistent improvement of the sequence–structure mapping based on model evaluation is still an unresolved problem.

6.4. What Can Be Done If No Template Is Detected Reliably?

If none of the most sensitive profile (HMM)-based methods can reliably detect any structural template it may mean that indeed there is no related template in the PDB. Alternatively, the relationship might be too distant, beyond the sensitivity limits of current methods. In both cases, there are at least two ways to approach the problem.

If obtaining the 3D model is not the most urgent task, the first option is to use alerting systems such as Re-searcher (77) or PDBalert (88) for performing automatic recurrent searches of homologous structures in PDB. Re-searcher uses PSI-BLAST as the search engine, and PDBalert is based on even more sensitive method, HHsearch. Usually the confident detection of a modeling template is the result of new homologous structure being deposited into PDB. However, in some cases, merely an increase of the number of sequence homologs may be sufficient to reliably detect templates that have already been present in PDB. This may happen because additional sequences help to build more representative sequence profiles (or HMMs). The serious drawback of this option is the unpredictability of the time frame when the suitable template will be detected. It may happen within days, but it may also happen years later, when the structure of a homolog is solved and deposited into PDB.

The second option is to use free modeling (FM) methods that do not have to rely on explicit templates and sequence–structure alignments to construct 3D models. Currently, there are a number of methods that would automatically shift to the free modeling mode if no suitable templates could be detected. Some of the most effective such methods include Robetta (43), an automatic server based on Rosetta, a highly successful fragment-based approach (89), I-TASSER (41, 90) and its relative Pro-sp3-TASSER (42, 91), SAM-T08 (13), MULTICOM (45). As it has been observed in CASP trials, these approaches can produce models of reasonable quality for small proteins (up to ~100 residues) having simple topology. However, at present, it would be too optimistic to expect consistently good models from FM approaches. Therefore, the confident detection of even remotely homologous structural template may help to improve modeling results considerably.

7. Conclusions

A steady growth of experimentally determined protein structures coupled with a dramatic increase of sequence data has made homology modeling both widely applicable and practically useful. In recent years, there have also been significant advances in distant homology detection and sequence alignment. The largest progress has been made mainly due to the application of sequence profiles and HMMs. At the same time, there are a number of remaining issues. In particular, there is a great need for improvement of the sequence–structure alignment accuracy, which is a key factor determining the quality of a homology model. This issue is tightly linked with the ability to accurately estimate local errors in protein models. As indicated by CASP blind trials this is a notoriously difficult problem. However, with the recent emphasis within the modeler community on the accurate model quality estimates there is hope for significant breakthroughs in this area. On the other hand, even currently available tools provide users with a lot of possibilities to construct, assess, and improve sequence–structure alignments for homology modeling.

Acknowledgments

Ana Vencloviené and members of Venclovas' lab are gratefully acknowledged for useful comments and suggestions.

References

1. Grishin, N. V. (2001) Fold change in evolution of protein structures, *J Struct Biol* ***134***, 167–185.
2. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990) Basic local alignment search tool, *J Mol Biol* ***215***, 403–410.
3. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res* ***25***, 3389–3402.
4. Karlin, S., and Altschul, S. F. (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes, *Proc Natl Acad Sci U S A* ***87***, 2264–2268.
5. Pearson, W. R., and Lipman, D. J. (1988) Improved tools for biological sequence comparison, *Proc Natl Acad Sci U S A* ***85***, 2444–2448.
6. Smith, T. F., and Waterman, M. S. (1981) Identification of common molecular subsequences, *J Mol Biol* ***147***, 195–197.
7. Pearson, W. R. (1991) Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms, *Genomics* ***11***, 635–650.
8. Biegert, A., and Söding, J. (2009) Sequence context-specific profiles for homology searching, *Proc Natl Acad Sci U S A* ***106***, 3770–3775.
9. Gribskov, M., McLachlan, A. D., and Eisenberg, D. (1987) Profile analysis: detection of distantly related proteins, *Proc Natl Acad Sci U S A* ***84***, 4355–4358.
10. Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. (1999) *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, Cambridge University Press.
11. Eddy, S. R. (1998) Profile hidden Markov models, *Bioinformatics* ***14***, 755–763.
12. Hughey, R., and Krogh, A. (1996) Hidden Markov models for sequence analysis: extension and analysis of the basic method, *Comput Appl Biosci* ***12***, 95–107.
13. Karplus, K. (2009) SAM-T08, HMM-based protein structure prediction, *Nucleic Acids Res* ***37***, W492–497.
14. Johnson, L. S., Eddy, S. R., and Portugaly, E. (2010) Hidden Markov model speed heuristic and iterative HMM search procedure, *BMC Bioinformatics* ***11***, 431.
15. Sadreyev, R., and Grishin, N. (2003) COMPASS: a tool for comparison of multiple protein alignments with assessment of statistical significance, *J Mol Biol* ***326***, 317–336.
16. Söding, J. (2005) Protein homology detection by HMM-HMM comparison, *Bioinformatics* ***21***, 951–960.
17. Margelevičius, M., and Venclovas, Č. (2010) Detection of distant evolutionary relationships between protein families using theory of sequence profile-profile comparison, *BMC Bioinformatics* ***11***, 89.
18. Yona, G., and Levitt, M. (2002) Within the twilight zone: a sensitive profile-profile comparison tool based on information theory, *J Mol Biol* ***315***, 1257–1275.
19. Madera, M. (2008) Profile Comparer: a program for scoring and aligning profile hidden Markov models, *Bioinformatics* ***24***, 2630–2631.
20. Rychlewski, L., Jaroszewski, L., Li, W., and Godzik, A. (2000) Comparison of sequence profiles. Strategies for structural predictions using sequence information, *Protein Sci* ***9***, 232–241.

21. Holm, L., and Sander, C. (1993) Protein structure comparison by alignment of distance matrices, *J Mol Biol* ***233***, 123–138.
22. Wang, Y., Sadreyev, R. I., and Grishin, N. V. (2009) PROCAIN: protein profile comparison with assisting information, *Nucleic Acids Res* ***37***, 3522–3530.
23. Eddy, S. R. (2008) A probabilistic model of local sequence alignment that simplifies statistical significance estimation, *PLoS Comput Biol* ***4***, e1000069.
24. Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, *Nucleic Acids Res* ***22***, 4673–4680.
25. Do, C. B., and Katoh, K. (2008) Protein multiple sequence alignment, *Methods Mol Biol* ***484***, 379–413.
26. Pei, J. (2008) Multiple protein sequence alignment, *Curr Opin Struct Biol* ***18***, 382–386.
27. Kemena, C., and Notredame, C. (2009) Upcoming challenges for multiple sequence alignment methods in the high-throughput era, *Bioinformatics* ***25***, 2455–2465.
28. Katoh, K., Misawa, K., Kuma, K., and Miyata, T. (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform, *Nucleic Acids Res* ***30***, 3059–3066.
29. Edgar, R. C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput, *Nucleic Acids Res* ***32***, 1792–1797.
30. Notredame, C., Higgins, D. G., and Heringa, J. (2000) T-Coffee: A novel method for fast and accurate multiple sequence alignment, *J Mol Biol* ***302***, 205–217.
31. Do, C. B., Mahabhashyam, M. S., Brudno, M., and Batzoglou, S. (2005) ProbCons: Probabilistic consistency-based multiple sequence alignment, *Genome Res* ***15***, 330–340.
32. Katoh, K., Kuma, K., Toh, H., and Miyata, T. (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment, *Nucleic Acids Res* ***33***, 511–518.
33. Edgar, R. C., and Batzoglou, S. (2006) Multiple sequence alignment, *Curr Opin Struct Biol* ***16***, 368–373.
34. Wallace, I. M., O'Sullivan, O., Higgins, D. G., and Notredame, C. (2006) M-Coffee: combining multiple sequence alignment methods with T-Coffee, *Nucleic Acids Res* ***34***, 1692–1699.
35. Katoh, K., Kuma, K., Miyata, T., and Toh, H. (2005) Improvement in the accuracy of multiple sequence alignment program MAFFT, *Genome Inform* ***16***, 22–33.
36. Pei, J., and Grishin, N. V. (2007) PROMALS: towards accurate multiple sequence alignments of distantly related proteins, *Bioinformatics* ***23***, 802–808.
37. Pei, J., Kim, B. H., and Grishin, N. V. (2008) PROMALS3D: a tool for multiple protein sequence and structure alignments, *Nucleic Acids Res* ***36***, 2295–2300.
38. O'Sullivan, O., Suhre, K., Abergel, C., Higgins, D. G., and Notredame, C. (2004) 3DCoffee: combining protein sequences and structures within multiple sequence alignments, *J Mol Biol* ***340***, 385–395.
39. Armougom, F., Moretti, S., Poirot, O., Audic, S., Dumas, P., Schaeli, B., Keduas, V., and Notredame, C. (2006) Expresso: automatic incorporation of structural information in multiple sequence alignments using 3D-Coffee, *Nucleic Acids Res* ***34***, W604–608.
40. Moult, J. (2005) A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction, *Curr Opin Struct Biol* ***15***, 285–289.
41. Roy, A., Kucukural, A., and Zhang, Y. (2010) I-TASSER: a unified platform for automated protein structure and function prediction, *Nat Protoc* ***5***, 725–738.
42. Zhou, H., and Skolnick, J. (2009) Protein structure prediction by pro-Sp3-TASSER, *Biophys J* ***96***, 2119–2127.
43. Kim, D. E., Chivian, D., and Baker, D. (2004) Protein structure prediction and analysis using the Robetta server, *Nucleic Acids Res* ***32***, W526–531.
44. Kelley, L. A., and Sternberg, M. J. (2009) Protein structure prediction on the Web: a case study using the Phyre server, *Nat Protoc* ***4***, 363–371.
45. Wang, Z., Eickholt, J., and Cheng, J. (2010) MULTICOM: a multi-level combination approach to protein structure prediction and its assessments in CASP8, *Bioinformatics* ***26***, 882–888.
46. Lobley, A., Sadowski, M. I., and Jones, D. T. (2009) pGenTHREADER and pDomTHREADER: new methods for improved protein fold recognition and superfamily discrimination, *Bioinformatics* ***25***, 1761–1767.
47. Jones, D. T. (1999) GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences, *J Mol Biol* ***287***, 797–815.
48. Kurowski, M. A., and Bujnicki, J. M. (2003) GeneSilico protein structure prediction meta-server, *Nucleic Acids Res* ***31***, 3305–3307.
49. Wallner, B., Larsson, P., and Elofsson, A. (2007) Pcons.net: protein structure prediction meta server, *Nucleic Acids Res* ***35***, W369–374.

50. Ginalski, K. (2006) Comparative modeling for protein structure prediction, *Curr Opin Struct Biol 16*, 172–177.

51. Moult, J., Fidelis, K., Kryshtafovych, A., Rost, B., and Tramontano, A. (2009) Critical assessment of methods of protein structure prediction - Round VIII, *Proteins* 77 Suppl *9*, 1–4.

52. Hildebrand, A., Remmert, M., Biegert, A., and Söding, J. (2009) Fast and accurate automatic structure prediction with HHpred, *Proteins* 77 Suppl *9*, 128–132.

53. Cozzetto, D., and Tramontano, A. (2005) Relationship between multiple sequence alignments and quality of protein comparative models, *Proteins 58*, 151–157.

54. Holm, L., Kaariainen, S., Rosenstrom, P., and Schenkel, A. (2008) Searching protein structure databases with DaliLite v.3, *Bioinformatics 24*, 2780–2781.

55. Qi, Y., Sadreyev, R. I., Wang, Y., Kim, B. H., and Grishin, N. V. (2007) A comprehensive system for evaluation of remote sequence similarity detection, *BMC Bioinformatics 8*, 314.

56. Sadreyev, R. I., and Grishin, N. V. (2004) Quality of alignment comparison by COMPASS improves with inclusion of diverse confident homologs, *Bioinformatics 20*, 818–828.

57. Tress, M. L., Cozzetto, D., Tramontano, A., and Valencia, A. (2006) An analysis of the Sargasso Sea resource and the consequences for database composition, *BMC Bioinformatics 7*, 213.

58. Chao, K. M., Hardison, R. C., and Miller, W. (1993) Locating well-conserved regions within a pairwise alignment, *Comput Appl Biosci 9*, 387–396.

59. Vingron, M., and Argos, P. (1990) Determination of reliable regions in protein sequence alignments, *Protein Eng 3*, 565–569.

60. Mevissen, H. T., and Vingron, M. (1996) Quantifying the local reliability of a sequence alignment, *Protein Eng 9*, 127–132.

61. Tress, M. L., Jones, D., and Valencia, A. (2003) Predicting reliable regions in protein alignments from sequence profiles, *J Mol Biol 330*, 705–718.

62. Cline, M., Hughey, R., and Karplus, K. (2002) Predicting reliable regions in protein sequence alignments, *Bioinformatics 18*, 306–314.

63. Chen, H., and Kihara, D. (2008) Estimating quality of template-based protein models by alignment stability, *Proteins 71*, 1255–1274.

64. Margelevičius, M., and Venclovas, Č. (2005) PSI-BLAST-ISS: an intermediate sequence search tool for estimation of the position-specific alignment reliability, *BMC Bioinformatics 6*, 185.

65. Prasad, J. C., Comeau, S. R., Vajda, S., and Camacho, C. J. (2003) Consensus alignment for reliable framework prediction in homology modeling, *Bioinformatics 19*, 1682–1691.

66. Sippl, M. J. (1993) Recognition of errors in three-dimensional structures of proteins, *Proteins 17*, 355–362.

67. Eisenberg, D., Luthy, R., and Bowie, J. U. (1997) VERIFY3D: assessment of protein models with three-dimensional profiles, *Methods Enzymol 277*, 396–404.

68. Cozzetto, D., Kryshtafovych, A., Ceriani, M., and Tramontano, A. (2007) Assessment of predictions in the model quality assessment category, *Proteins* 69 Suppl *8*, 175–183.

69. Cozzetto, D., Kryshtafovych, A., and Tramontano, A. (2009) Evaluation of CASP8 model quality predictions, *Proteins* 77 Suppl *9*, 157–166.

70. Benkert, P., Kunzli, M., and Schwede, T. (2009) QMEAN server for protein model quality estimation, *Nucleic Acids Res 37*, W510–514.

71. Benkert, P., Tosatto, S. C., and Schomburg, D. (2008) QMEAN: A comprehensive scoring function for model quality assessment, *Proteins 71*, 261–277.

72. Venclovas, Č. (2003) Comparative modeling in CASP5: progress is evident, but alignment errors remain a significant hindrance, *Proteins* 53 Suppl *6*, 380–388.

73. Venclovas, Č., and Margelevičius, M. (2009) The use of automatic tools and human expertise in template-based modeling of CASP8 target proteins, *Proteins* 77 Suppl *9*, 81–88.

74. Raman, S., Vernon, R., Thompson, J., Tyka, M., Sadreyev, R., Pei, J., Kim, D., Kellogg, E., DiMaio, F., Lange, O., Kinch, L., Sheffler, W., Kim, B. H., Das, R., Grishin, N. V., and Baker, D. (2009) Structure prediction for CASP8 with all-atom refinement using Rosetta, *Proteins* 77 Suppl *9*, 89–99.

75. Cozzetto, D., Kryshtafovych, A., Fidelis, K., Moult, J., Rost, B., and Tramontano, A. (2009) Evaluation of template-based models in CASP8 with standard measures, *Proteins* 77 Suppl *9*, 18–28.

76. Li, W., and Godzik, A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences, *Bioinformatics 22*, 1658–1659.

77. Repšys, V., Margelevičius, M., and Venclovas, Č. (2008) Re-searcher: a system for recurrent detection of homologous protein sequences, *BMC Bioinformatics 9*, 296.

78. Söding, J., Biegert, A., and Lupas, A. N. (2005) The HHpred interactive server for protein homology detection and structure prediction, *Nucleic Acids Res 33*, W244–248.

79. Brandt, B. W., and Heringa, J. (2009) web-PRC: the Profile Comparer for alignment-based

searching of public domain databases, *Nucleic Acids Res 37*, W48–52.

80. Margelevičius, M., Laganeckas, M., and Venclovas, Č. (2010) COMA server for protein distant homology search, *Bioinformatics 26*, 1905–1906.
81. Sadreyev, R. I., Tang, M., Kim, B. H., and Grishin, N. V. (2007) COMPASS server for remote homology inference, *Nucleic Acids Res 35*, W653–658.
82. Wang, Y., Sadreyev, R. I., and Grishin, N. V. (2009) PROCAIN server for remote protein sequence similarity search, *Bioinformatics 25*, 2076–2077.
83. Gonzalez, M. W., and Pearson, W. R. (2010) Homologous over-extension: a challenge for iterative similarity searches, *Nucleic Acids Res 38*, 2177–2189.
84. Sali, A., and Blundell, T. L. (1993) Comparative protein modelling by satisfaction of spatial restraints, *J Mol Biol 234*, 779–815.
85. Petrey, D., Xiang, Z., Tang, C. L., Xie, L., Gimpelev, M., Mitros, T., Soto, C. S., Goldsmith-Fischman, S., Kernytsky, A., Schlessinger, A., Koh, I. Y., Alexov, E., and Honig, B. (2003) Using multiple structure alignments, fast model building, and energetic analysis in fold recognition and homology modeling, *Proteins* 53 Suppl *6*, 430–435.
86. Guex, N., Peitsch, M. C., and Schwede, T. (2009) Automated comparative protein structure modeling with SWISS-MODEL and Swiss-PdbViewer: a historical perspective, *Electrophoresis* 30 Suppl *1*, S162–173.
87. Wiederstein, M., and Sippl, M. J. (2007) ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins, *Nucleic Acids Res 35*, W407–410.
88. Agarwal, V., Remmert, M., Biegert, A., and Söding, J. (2008) PDBalert: automatic, recurrent remote homology tracking and protein structure prediction, *BMC Struct Biol 8*, 51.
89. Bradley, P., Malmstrom, L., Qian, B., Schonbrun, J., Chivian, D., Kim, D. E., Meiler, J., Misura, K. M., and Baker, D. (2005) Free modeling with Rosetta in CASP6, *Proteins 61* Suppl *7*, 128–134.
90. Zhang, Y. (2009) I-TASSER: fully automated protein structure prediction in CASP8, *Proteins* 77 Suppl *9*, 100–113.
91. Zhou, H., Pandit, S. B., and Skolnick, J. (2009) Performance of the Pro-sp3-TASSER server in CASP8, *Proteins* 77 Suppl *9*, 123–127.