Fast and accurate automatic structure prediction with HHpred". Proteins

ARTICLE *in* PROTEINS STRUCTURE FUNCTION AND BIOINFORMATICS · JANUARY 2009

Impact Factor: 2.63 · DOI: 10.1002/prot.22499 · Source: PubMed

Michael Remmert
Ludwig-Maximilians-University of Munich

Andreas Biegert
Ludwig-Maximilians-University of Munich

Johannes Söding
Max Planck Institute for Biophysical Chemi…

Fast and accurate automatic structure prediction with HHpred

Andrea Hildebrand, Michael Remmert, Andreas Biegert, and Johannes Söding*

Gene Center and Center for Integrated Protein Science (CIPSM), Ludwig-Maximilians-University Munich, Feodor-Lynen-Str. 25, 81377 Munich, Germany

Abstract: *Automated protein structure prediction is becoming a mainstream tool for biological research. This has been fueled by steady improvements of publicly available automated servers over the last decade, in particular their ability to build good homology models for an increasing number of targets by reliably detecting and aligning more and more remotely homologous templates. Here, we describe the three fully automated versions of the HHpred server that participated in the community-wide blind protein structure prediction competition CASP8. What makes HHpred unique is the combination of usability, short response times – typically under 15 minutes – and a model accuracy that is competitive with those of the best servers in CASP8.*

Introduction

The potential for automated homology-based protein structure prediction is great and largely underexploited. We estimate that structural models for hundreds of unidentified domains in human proteins could be reliably modeled by state-of-the-art methods with an expected TM-score [1] greater than 0.5 (J. Söding *et al.*, unpublished data). This accuracy would be sufficient, for instance, to predict molecular functions and to guide site-directed mutagenesis experiments. We developed the HHpred server to fill the gap between the fast and widely used homology search programs such as BLAST, PSI-BLAST [2], or HMMer/Pfam [3], and the very sensitive and accurate but rather inflexible and slow automated protein structure prediction servers [4, 5, 6, 7, 8]. HHpred is therefore mainly meant to be used as interactive function and structure prediction server, allowing, for example, to search various databases, to select templates manually or to correct errors in the proposed target-template alignment. To test the accuracy of HHpred, we participated in CASP8 with three fully automated versions, available at http://toolkit.lmb.uni-muenchen.de/casp/hhpred5. The same protocols as used by these automatic versions are available on the interactive server at http://toolkit.lmb.uni-muenchen.de/hhpred (see the HHpred help pages). All HHpred Perl scripts are freely available upon request. HHsearch can be downloaded at ftp://toolkit.lmb.uni-muenchen.de/HHsearch/.

Methods

Three HHpred servers participated in CASP8: two single-template versions (HHpred2, HHpred4) and a multiple-template version (HHpred5). They follow a similar protocol:

1. *Build a multiple sequence alignment for the target sequence:* To build the target alignment, HHpred2 runs the `buildali.pl` script from the HHsearch 1.5.0 software package [9]. This script performs up to 8 iterative PSI-BLAST searches through filtered versions of the nonredundant (nr) database from the NCBI, each time jump-starting PSI-BLAST with the alignment extracted from the search results of the previous iteration. Since the most common source for corrupted PSI-BLAST alignments is the inclusion of non-homologous segments at the ends of local sequence matches, `buildali.pl` prunes the ends of each sequence separately if the similarity with the profile extracted after the first search iteration falls below 1/6 bit per column. HHpred4 and HHpred5 build their target alignment by a maximum of five iterated HMM searches through a filtered version of the nr database with a maximum of 30% pairwise sequence identity (M. Remmert and J. Söding, manuscript in preparation). In addition, they employ a preliminary version of context-specific pseudocounts to increase the sensitivity of these searches [10].

2. *Search for homologous templates*: A profile hidden Markov model (HMM) is calculated from the target alignment using the hhmake executable with default parameters. Homologous templates are identified by searching through HHpred's weekly updated PDB70 database using HHsearch, a method for pairwise comparison of HMMs [9]. The PDB70 database contains HMMs for a representative subset of PDB sequences. These HMMs are built in the same way as the target alignments in HHpred2. HHsearch employs the Viterbi algorithm for ranking the database matches but realigns the best matches with a local HMM-HMM alignment version of the more accurate Maximum Accuracy (MAC) algorithm [11, 12], which maximizes the expected number of correctly aligned residues. This algorithm improved the model quality in our unpublished benchmarks by several percent.

3. *Re-rank the potential templates with a neural network:* HHsearch ranks database matches by the probability of the match to be homologous to the target sequence. This is useful to distinguish homologous from nonhomologous matches, but it is not most appropriate for ranking homologous templates according to the expected quality of the homology models they would yield. We therefore train a neural network to predict the TM-score of the homology model. Based on this prediction we re-rank the database matches. The following three features proved to be most informative: the raw HHsearch score, HHsearch's secondary structure similarity score divided by target length, and the expected number of correctly aligned target residues divided by target length. The expected number of

*Corresponding author: Johannes Söding, email: soeding@lmb.uni-muenchen.de

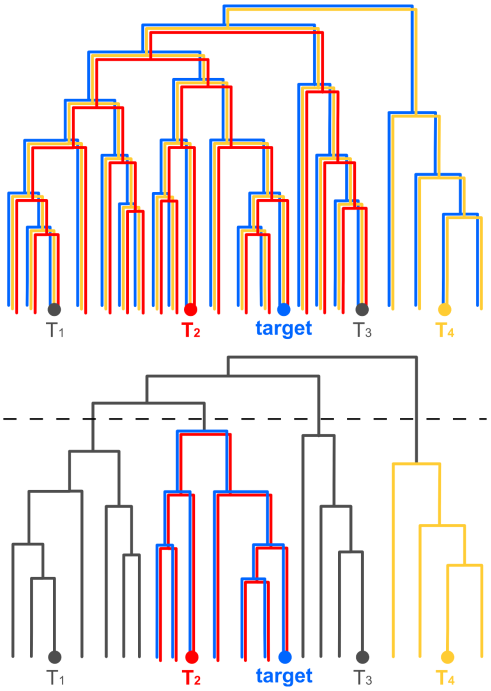

Figure 1: Narrowing down the diversity of target and template alignments facilitates finding the most closely related template. The idealized phylogenetic tree shows a target and four possible templates T_1 - T_4. The sequences contained in the alignments of the target and of templates T_2 and T_4 are shown in red, blue and yellow, respectively. Without filtering (top), the sequences of the target and template alignments largely overlap, making it impossible to discern the most closely related template T_2. At the optimal filter threshold (bottom), the HMM of template T_2 will be most similar to the target HMM.

correctly aligned residues is calculated by the MAC algorithm and is part of the output of HHsearch version 1.5 and above ("Sum_probs"). We used the Stuttgart Neural Network Simulator (http://www.ra.cs.uni-tuebingen.de/SNNS/) to train a feed-forward neural network with a single hidden layer of three nodes and a single linear output node. As training set we picked 4293 target-template alignments constructed by HHsearch for 507 targets. For each target-template alignment, we built a homology model using the MODELLER software package [13] and calculated the TM-score of the model with respect to the actual target structure. The neural network was then trained to predict these TM-scores.

4. *Generate sets of multiple alignments with successively lower sequence diversities for the target sequence and the templates:* Often, several templates can be detected with high probabilities. If these templates are more closely related with each other than with some of the sequences contained in the multiple alignments from which their HMMs are calculated, their HMMs will all be very similar to each other (Fig. 1 top). Hence HHpred's probabilities will not reflect the true degree of relatedness to the target sequence. To decide which of them is in fact most closely related to the target, we need to narrow down the diversity of the target and template alignments. HHpred generates 10 sets of alignments with successively lower diversity for the target sequence and for all database matches with at least 80% probability. For this purpose, we employ `hhfilter` from the HHsearch package with option `-qsc`. We remove all sequences from the multiple sequence alignments which have a Gonnet matrix score per column with the target or template sequence of less than the given threshold. The similarity threshold is increased in ten steps from 0.1 to 1.0 bits per column. At each threshold value, target and template alignments are filtered and the filtered target is used to search with HHsearch through an HMM database built from the filtered template alignments.

5. *Rank target-template alignments of various alignment diversities with neural network:* All in all, one unfiltered set (down to a probability of 10%) and ten filtered sets (down to a probability of 80%) of target-template alignments are generated by HHpred in this way. For each of these alignments, we predict the expected TM-score of the resulting structural model with the neural network and rank the templates according to this score. This procedure has two advantages: First, it allows to pick the template most closely related to the target (see Fig. 1). Second, it allows to choose the alignment diversity that maximizes the expected number of correctly aligned residues. For example, sequences in the template multiple alignment that are more distantly related to the template than to the target will in general impair the target-template alignment quality and will be filtered out.

6. *Choose template(s):* HHpred2 and HHpred4 pick only a single template per domain. They always pick the top-ranked target-template alignment from the list containing all 11 sets of target-template alignments. They then move down the list to find templates for potential further domains not covered by the top-ranked alignment. They select an additional alignment if it does not overlap more than 20 residues with any of the already selected alignments and if it covers at least 40 target residues not already covered. All target-template alignments thus selected are combined into a target-template multiple alignment. HHpred5 models each domain with multiple templates if possible. However, we want to avoid including too remote templates that could negatively affect the model through alignment errors or strong structural divergence. Therefore, after selecting the top-ranked target-template alignment from the list, HHpred5 goes down this list and selects only those target-template alignments whose templates are more similar to the top template (up to 0.1 bits per column) than this top template is to the target sequence. Similarity is again calculated as sequence-sequence similarity using the Gonnet matrix with gap open and extend scores of 6 and 1 bits, respectively. As HHpred2 and HHpred4, HHpred5 also selects target-template alignments that cover at least 40 target residues not already covered in the selected alignments (regardless of overlaps with already selected alignments).

7. *Run MODELLER [13]:* The standard `automodel` script is run with the target-template multiple alignment from step 6. In CASP8, we calculated three models and submitted the model with the highest MODELLER score.

Results and Discussion

Figure 2A shows the performance of automatic tertiary structure prediction servers that participated in CASP8, measured by the sum of positive Z-scores $\sum_t \max\{Z_t, 0\}$ for all

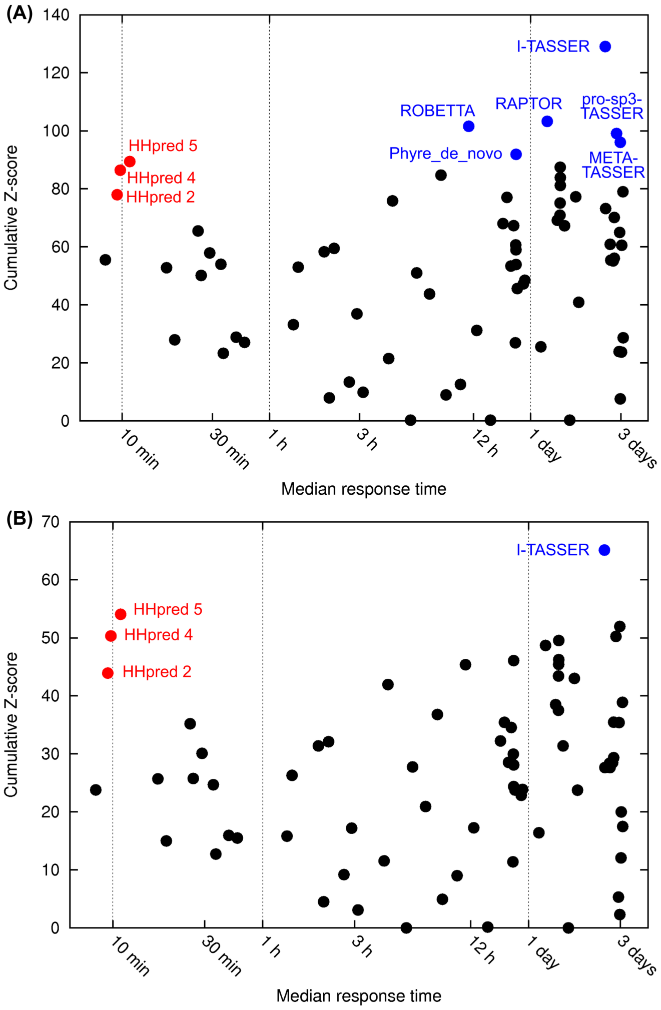

Figure 2: Model accuracy versus response time for all 72 servers that participated in the tertiary structure prediction category of CASP8 (A) on all 164 target domains and (B) on the 85 single-domain targets. The servers with better Z-score than HHpred5 are shown in blue (I-TASSER is the Zhang server). The improvement of the HHpred servers from A to B indicate that the domain parsing severely limited their performance in CASP8.

164 target domains t. The scores were downloaded from the CASP8 web site (`http://www.predictioncenter.org/casp8/groups_analysis.cgi`) and are based on domain-specific GDT-TS scores. On the x-axis, the median response time of the servers in CASP8 is shown. At around 10 minutes, the three HHpred servers (red) provide an excellent combination of accuracy and speed. The six servers that score better than HHpred5 all took more than 50 times longer to respond (blue dots). Except for the best-scoring server (I-TASSER), their accuracies are comparable to HHpred5. In practice, response times may differ considerably from the times recorded for CASP8. According to our own recent tests on the six best-performing servers, most have either queuing times much in excess of these times or were not yet publicly available at the time of writing.

The ranking and domain parsing procedures in the HHpred servers possess two shortcomings: First, target-template alignments are ranked better by the neural network the more target residues are covered. This makes sense in single-domain proteins, since unaligned target residues cannot be reasonably modeled in our approach. In multi-domain proteins, however, this is problematic. Consider the example of a two-domain target, for which both domains can be modeled independently with two closely related templates. If a remotely related template exists which covers both domains, we risk ranking this template higher because it covers more target residues than any of the two more closely related templates. Second, the present domain parsing (step 6) can fail when the target-template alignment of one target domain is extended too much and overlaps a neighboring domain by more than 20 residues. In this case, we would reject alignments covering the neighboring domain.

To check if the domain parsing procedure was impairing prediction results on multi-domain proteins, we calculated the cumulative positive Z-scores for all servers on the subset of the 85 single-domain proteins in CASP8 (Fig.2B). Indeed, on this target subset, the HHpred servers perform better, improving their ranks from 7th to 2nd (HHpred5), 9th to 4th (HHpred4), and 14th to 12th (HHpred2).

We measured the modeling accuracy on 507 targets from the PDB database to test the effect of the various changes. Relative to the old HHpred2 server, which ranked second in CASP7 [14], the improvements are split up as follows: (1) The removal of bugs in the old server led to an increase in the cumulative GDT_TS score of 0.5%. (2) The neural network-based reranking yielded another 1.3%. (3) Filtering the target and template alignments and picking the alignments with optimal diversity resulted in a gain of an additional 0.9%. Finally, using multiple templates improved the cumulative score by another 2.1%. In total, the improvement over the old HHpred2 of CASP7 was 4.9%. The preliminary HMM-based procedure for generating the target alignments had not been evaluated but seemed to be of minor effect in CASP8. The improvement of 4.9% from the old HHpred2 to HHpred5 is in accord with the performance of servers that ran unchanged in CASP7 and CASP8, such as SAM-T2K.

It is remarkable that HHpred is competitive in model accuracy with servers that need many times more CPU time. To keep it fast, no alternative alignments are explored [4, 5, 6, 7, 15], no side chain optimization performed [4, 5, 6, 7], no contacts predicted [4, 5, 6, 7], no loops modeled [4, 5, 6, 7], and no model quality assessment or structural clustering employed to pick the best of several models [4, 5, 6, 7, 8]. Our philosophy has been to dispense of everything that needs computational resources of more than a few minutes. At the moment, HHpred executes around 200 jobs per day, each on a single CPU, with negligible queuing time and typical response times of a few minutes. This is in contrast to most other highly ranked servers in CASP7/CASP8, which in practice take days to months to return a 3D model.

For the future, we are working on improving homology detection and alignment methods which are both crucial, accuracy-limiting steps in homology modeling [15]. We plan to make better use of the information from multiple templates, which will also help to resolve the current problems with multi-domain proteins. Furthermore, we would like to correct alignment errors during the homology modeling stage and to include local structural preferences as additional modeling restraints.

Acknowledgments

Many thanks to Andriy Kryshtafovych for the response time data, to the CASP organizers and assessors for their valuable service to the community, to a reviewer for suggesting the single-domain analysis, and to Andrej Sali and David Jones for making their software freely available.

References

[1] Zhang Y, Skolnick J (2004) Scoring function for automated assessment of protein structure template quality. Proteins 57:702–710.

[2] Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acid Research 25:3389–3402.

[3] Eddy SR (1998) Profile hidden markov models. Bioinformatics 14:755–763.

[4] Zhang Y (2008) I-TASSER server for protein 3D structure prediction. BMC Bioinformatics 9:40.

[5] Chivian D, Kim DE, Malmström L, Schonbrun J, Rohl CA, et al. (2005) Prediction of CASP6 structures using automated Robetta protocols. Proteins 61:157–166.

[6] Zhou H, Skolnick J (2009) Protein structure prediction by pro-Sp3-TASSER. Biophys J 96:2119–2127.

[7] Zhou H, Pandit SB, Lee SY, Borreguero J, Chen H, et al. (2007) Analysis of TASSER-based CASP7 protein structure prediction results. Proteins 69:90–97.

[8] Xu J, Jiao F, Yu L (2008) Protein structure prediction using threading. Methods Mol Biol 413:91–121.

[9] Soding J (2005) Protein homology detection by HMM-HMM comparison. Bioinformatics 21:951–960.

[10] Biegert A, Söding J (2009) Sequence context-specific profiles for homology searching. Proc Natl Acad Sci U S A 106:3770–3775.

[11] Holmes I, Durbin R (1998) Dynamic programming alignment accuracy. J Comput Biol 5:493–504.

[12] Biegert A, Söding J (2008) De novo identification of highly diverged protein repeats by probabilistic consistency. Bioinformatics 24:807–814.

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

[14] Battey JN, Kopp J, Bordoli L, Read RJ, Clarke ND, et al. (2007) Automated server predictions in CASP7. Proteins 69:68–82.

[15] Jaroszewski L, Li W, Godzik A (2002) In search for more accurate alignments in the twilight zone. Protein Sci 11:1702–1713.