The HHpred interactive server for protein homology detection and structure prediction

Johannes Söding*, Andreas Biegert and Andrei N. Lupas

Department of Protein Evolution, Max-Planck-Institute for Developmental Biology, Spemannstrasse 35, 72076 Tübingen, Germany

Received February 14, 2005; Revised and Accepted March 21, 2005

ABSTRACT

HHpred is a fast server for remote protein homology detection and structure prediction and is the first to implement pairwise comparison of profile hidden Markov models (HMMs). It allows to search a wide choice of databases, such as the PDB, SCOP, Pfam, SMART, COGs and CDD. It accepts a single query sequence or a multiple alignment as input. Within only a few minutes it returns the search results in a user-friendly format similar to that of PSI-BLAST. Search options include local or global alignment and scoring secondary structure similarity. HHpred can produce pairwise query-template alignments, multiple alignments of the query with a set of templates selected from the search results, as well as 3D structural models that are calculated by the MODELLER software from these alignments. A detailed help facility is available. As a demonstration, we analyze the sequence of SpoVT, a transcriptional regulator from Bacillus subtilis. HHpred can be accessed at http://protevo.eb.tuebingen.mpg.de/hhpred.

INTRODUCTION

It is well known that sequence search methods such as BLAST, FASTA or PSI-BLAST (1–3) are of prime importance for biological research because functional information of a protein or gene can be inferred from homologous proteins or genes identified in a sequence search. But quite often no significant relationship to a protein of known function can be established. This is certainly the case for the most interesting group of proteins, those for which no ortholog has yet been studied.

It is less well known that in cases where conventional sequence search methods fail, the recently developed, highly sensitive methods for homology detection or structure prediction (confer, e.g. (4–11) and descriptions and links at http://bioinfo.pl/Meta/servers.html) quite often allow to make inferences from more remotely homologous relationships (12–17). If the relationship is so remote that no common function can be assumed, one can generally still derive hypotheses about possible mechanisms, active site positions and residues, or the class of substrate bound (18,19). When a homologous protein with known structure can be identified, its structure can be used as a template to model the 3D structure for the protein of interest (5), since even remotely homologous proteins generally have quite similar 3D structure (20). The 3D model may then help to generate hypotheses to guide experiments.

The primary aim in developing HHpred was to provide biologists with a method for sequence database searching that is as easy to use as BLAST or PSI-BLAST and yet competitive in sensitivity with the most powerful servers for structure prediction that are currently available. We believe that HHpred is unique in the advantages it offers:

- **Speed**: A search with a 300 residue sequence through the Protein Data Bank (PDB) (≈9000 HMMs) takes ~1 min.
- **Databases**: A wide range of regularly updated structure and protein family databases can be searched: the PDB (21), SCOP (22), Pfam (23), SMART (24), COG (25) and CDD (26).
- **User-friendliness**: Search results are presented in an easy-to-read format that is similar to PSI-BLAST. The summary hit list includes E-values and true probabilities. Alignments contain annotation about secondary structure, consensus sequences and position-specific reliability and they can be augmented by representative sequences from the underlying multiple alignments.
- **Flexibility**: We try to offer the user maximum control and flexibility. He can paste his own input query alignment, search in local or global alignment mode, realign alignments with other parameters and edit the query-template (multiple) alignment with which to launch the comparative modelling.
- **Multi-domain proteins**: HHpred has been designed to work equally well for single-domain and multi-domain query sequences. It can therefore be used to predict domain boundaries.
- **Documentation**: A comprehensive help facility is available.

*To whom correspondence should be addressed. Tel: +49 7071 601 451; Fax: +49 7071 601 349; Email: johannes.soeding@tuebingen.mpg.de

© The Author 2005. Published by Oxford University Press. All rights reserved.

The online version of this article has been published under an open access model. Users are entitled to use, reproduce, disseminate, or display the open access version of this article for non-commercial purposes provided that: the original authorship is properly and fully attributed; the Journal and Oxford University Press are attributed as the original place of publication with the correct citation details given; if an article is subsequently reproduced or disseminated not in its entirety but only in part or as a derivative work this must be clearly indicated. For commercial re-use, please contact journals.permissions@oupjournals.org
Selectivity: High-scoring false positives have systematically been reduced by developing a protocol for building query and database alignments that supresses non-homologous sequences (J. Söding, to be published).

Sensitivity: HHpred is among the most sensitive servers for remote homology detection. A comparison of the new version HHpred2.1 with the servers that took part in the recent structure prediction benchmark CAFASP4 (27) can be viewed at http://protevo.eb.tuebingen.mpg.de/ hhpred/hhpred_in_CAFASP4.html. In a recent study (28), in which we benchmarked HHsearch, the method for HMM–HMM comparison employed by our server, together with PSI-BLAST, HMMER, PROF_SIM and COMPASS (3,6,7,29), HHsearch was found to possess the highest sensitivity and alignment accuracy.

METHODS AND INPUT PARAMETERS

In the first step, an alignment of homologs is built for the query sequence by multiple iterations of PSI-BLAST searches against the non-redundant database from NCBI. The maximum number of PSI-BLAST iterations and the E-value threshold can be specified on the start page (Figure 1). Instead of a single sequence, the user may also enter a multiple alignment to jumpstart PSI-BLAST, or he can choose to skip the PSI-BLAST iterations altogether by choosing zero for the maximum number of PSI-BLAST iterations.

The user can further specify a minimum coverage of the query by the PSI-BLAST matches. With a value of 50%, at least half of the query residues must be aligned (‘covered’) with residues from the matched sequence in order for it to enter into the profile. Similarly, a minimum sequence identity of the PSI-BLAST match to the query sequence can be demanded. Our benchmarks (data not published) have shown that a value between 20 and 25% improves selectivity without compromising sensitivity. The final alignment from PSI-BLAST is annotated with the predicted secondary structure and confidence values from PSIPRED (30).

In the next step, a profile HMM is generated from the multiple alignment that includes the information about predicted secondary structure. A profile HMM is a concise statistical description of the underlying alignment. For each column in the multiple alignment that has a residue in the query sequence, an HMM column is created that contains the probabilities of each of the 20 amino acids, plus 4 probabilities that describe how often amino acids are inserted and deleted at this position (insert open/extend, delete open/extend). These insert/delete probabilities are translated into position-specific gap penalties when an HMM is aligned to a sequence or to another HMM.

Figure 1. Start page for the HHpred server at http://protevo.eb.tuebingen.mpg.de/hhpred with part of a help window overlaid.
The query HMM is then compared with each HMM in the selected database. The database HMMs have been precalculated and also contain secondary structure information, either predicted by PSIPRED, or assigned from 3D structure by DSSP (31). The database search is performed with the HHsearch software for HMM–HMM comparison (28). Compared to methods that rely on pairwise comparison of simple sequence profiles, HHsearch gains sensitivity by using position-specific gap penalties. If the default setting ‘Score secondary structure’ is active, a score for the secondary structure similarity is added to the total score. This increases the sensitivity for homologous proteins considerably (28). As a possible drawback, it may lead to marginally significant scores for structurally analogous, but non-homologous proteins.

The user can choose between local and global alignment mode. In global mode alignments extend in both directions up to the end of either the query or the database HMM. No penalties are charged for end gaps. In local mode, the highest-scoring local alignment is determined, which can start and end anywhere with respect to the compared HMMs. It is recommended to use the local alignment mode as a default setting since it has been shown in our benchmarks to be on average more sensitive in detecting remote relationships as well as being more robust in the estimation of statistical significance values. A global search might be appropriate when one expects the database entries to be (at least marginally) similar over their full length with the query sequence. In most cases it will be advisable to run a search in both modes to gain confidence in one’s results.

EXAMPLE ANALYSIS

As an example we analyze the sequence of Stage V sporulation protein T (SpoVT) from Bacillus subtilis that is known to regulate forespore-specific σ43-dependent transcription (32) (annotated as ‘transcriptional regulator’ in GenBank). Input parameters are set as shown in Figure 1. The results consist of two parts (Figure 2): a summary list with matching database sequences (‘templates’) and a list of query–template alignments below.

The first column of the summary hit list has indices that link to the corresponding alignment further down. Next are the first 30 characters from the description of the HMM. The ‘Prob’ column lists the probability in percent that the database match is a true positive, i.e. that it is homologous to the query sequence at least in some core part. This is the most relevant statistical measure of significance and can be interpreted quite literally. The true-positive probability is a conservative measure in the sense that it corrects for occasional high-scoring false positives. (The major cause for high-scoring false positives are corrupted alignments that contain non-homologous sequences which slipped in during the automatized alignment-building with PSI-BLAST.) [See (28) for details.] The E-values in HHpred are defined in the same way as in BLAST or PSI-BLAST. (The E-value for a sequence match is the expected number of false positives per database search with a score at least as good as the score of this sequence match.) But it is important to note that, in contrast to the true-positive probability, HHpred E-values do not take into account the secondary structure similarity. Hits can therefore be significant by the true-positive probability criterion even when the E-value is ~1. The P-value is equal to the E-value divided by the number of HMMs in the searched database. The ‘Score’ column gives the total score that includes the score from the secondary structure comparison which is listed in the next column (‘SS’). ‘Cols’ contains the total number of matched columns in the query–template alignment and the remaining columns describe the range of aligned residues in the query and template.

From the summary list in Figure 2 it is evident that the SpoVT protein consists of two domains, one from residue 1 to ~51 and the other from residue 52 to 178. The N-terminal domain has two significant hits in SCOP at rank 1 and 3. The first hit is the DNA-binding domain of transition-state regulator AbrB (33), a known close homolog of SpoVT. AbrB is a protein that is broadly represented in bacterial species and is involved in switching from exponential growth to stationary phase by integrating a great number of environmental factors. The second hit is to MazE, the antidote of the antidote-toxin addiction module MazEF (34). How can both AbrB and MazE be homologous to the query if they are not even classified into the same class, let alone fold or superfamily, by the SCOP database? Can the match with MazE be a false positive despite the rather significant 84% probability?

To elucidate this, we can look at the SpoVT–MazE alignment below. Five representative (i.e. maximally diverse) sequences from each of the two underlying alignments are shown for each HMM. (Their amino acids can be colored by biochemical properties by pressing one of the radio buttons entitled ‘color alignments’ above the summary hit list.) First we note that the predicted secondary structure of SpoVT (sequence ‘Q ss_pred’) agrees very well with the actual secondary structure of MazE determined by the program DSSP (sequence ‘T ss_dssp’). Second, the hydrophobicity pattern in the aligned HMMs looks quite similar, which is especially evident with the coloring. Third, the HMM–HMM alignment contains a single gap in MazE at a position where also some sequences in SpoVT exhibit a gap. All in all, the alignment looks very much like what one would expect for a distant homologous relationship.

The conflict posed by the manifest homology between MazE and AbrB and their grossly different structural topology prompted us to undertake a thorough bioinformatic investigation of the AbrB-like superfAMILY and to re-determine the AbrB structure by NMR (M. Coles and S. Djuranovic et al., manuscript submitted, PDB ID: 1YFB). Indeed, we found that the published structure of AbrB (PDB ID: 1EKI) is incorrect and that the correct structure for AbrB places it in the same superfamily as MazE.

Hits 2 and 4–9 in the summary list are all proteins from the same SCOP fold d.110. Clicking on the SCOP family IDs opens a window with the corresponding entry in SCOP. Irrespective of the specific significance values, the fact that so many quite divergent members from the same two superfamilies d.110.2 (GAF-domain) and d.110.3 (PAS-domain) appear among the best hits strongly indicates that these are not high-scoring chance hits but true homologs. Whether the C-terminal domain looks more like a GAF or a PAS domain, we can now generate an approximate structural model that could help us to guide experiments to investigate what regulatory substrate this domain may actually bind (32).

By clicking ‘Create CM Model’ one can select the templates to be used for comparative modelling. HHpred then returns a
multiple alignment in PIR format with the query sequence and the selected templates. This alignment may be edited by the user and then fed to the MODELLER software (35), accessible via the MPI toolkit for users of HHpred.

A very useful feature is the possibility to view and manually improve the query alignment that was used to generate the query HMM; via the tab ‘Edit Query Alignment’ the user can modify the query alignment that appears in a text field and start a new search with the modified alignment. By pressing ‘Realign’ at the top, the user may also realign the identified templates in the summary hit list with different parameters without the need to rerun the database search.

**CONCLUSION**

Whenever biologists cannot get satisfactory results from BLAST, PSI-BLAST or other database searches due to insignificant matches with proteins of known structure or function,
they should consider using one of the recently developed sensitive structure prediction and homology detection servers (4–11) that are listed, for instance, on the LiveBench/CAFASP site at http://bioinfo.pl/Meta/servers.html. Among these servers, HHpred offers a high degree of flexibility and user-friendliness combined with excellent sensitivity. In contrast to methods based on profile–profile comparison, HHpred exploits the information that is contained in insert and delete probabilities by including them in a statistical framework. But the speed of HHpred is perhaps the most important advantage, considering that the best-ranked servers in CAFASP4 generally take hours or even days to return a prediction. The speed enables the user to tweak the performance and gain confidence in the results by modifying input alignments, search parameters or selected databases on a trial and error basis.

ACKNOWLEDGEMENTS

We would like to thank Michael Remmert for his valuable help in setting up the web interface. We thank Sergej Djuranovic for first pointing out the HHpred prediction for the SpoVT C-terminal domain. J.S. is indebted to Alex Diemand for assistance in preparing the screenshots. Many thanks to all users who helped to improve our software with their questions and feedback. Funding to pay the Open Access publication charges for this article was provided by the Max-Planck Society.

Conflict of interest statement. None declared.

REFERENCES