W374–W378 *Nucleic Acids Research, 2006, Vol. 34, Web Server issue*
doi:10.1093/nar/gkl195

HHsenser: exhaustive transitive profile search using HMM–HMM comparison

Johannes Söding*, Michael Remmert, Andreas Biegert and Andrei N. Lupas

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

Received February 14, 2006; Revised and Accepted February 21, 2006

ABSTRACT

HHsenser is the first server to offer exhaustive intermediate profile searches, which it combines with pairwise comparison of hidden Markov models. Starting from a single protein sequence or a multiple alignment, it can iteratively explore whole superfamilies, producing few or no false positives. The output is a multiple alignment of all detected homologs. HHsenser's sensitivity should make it a useful tool for evolutionary studies. It may also aid applications that rely on diverse multiple sequence alignments as input, such as homology-based structure and function prediction, or the determination of functional residues by conservation scoring and functional subtyping.

HHsenser can be accessed at http://hhsenser.tuebingen.mpg.de/. It has also been integrated into our structure and function prediction server HHpred (http://hhpred.tuebingen.mpg.de/) to improve predictions for near-singleton sequences.

INTRODUCTION

Most methods that predict properties about a protein from its sequence profit from the inclusion of evolutionary information in the form of a multiple sequence alignment. Examples are the prediction of secondary structure (1), solvent accessebility (2), disorder (3), transmembrane helices (4), intraprotein contacts (5), protein–protein interactions (6), subcellular localization (7), internal repeats (8), deleterious mutations (9) and conserved or subtype-specific functional residues (10–13). Finally, the sensitivity to detect remote homologs depends particularly on how much sequence information from distant homologs is used in a pairwise comparison (14,15). The ability to construct a diverse multiple alignment from a single query sequence can therefore critically influence the performance of a vast array of analyses and prediction methods.

To find as many homologs as possible, two approaches have been taken. In the first, exemplified by PSI-BLAST (16), a sequence profile or profile hidden Markov model (HMM) (17) is constructed iteratively by searching a sequence database and updating the profile with the statistically significant sequence matches after each round of search. In the second approach ('intermediate sequence search'), a search with BLAST or a similar method is performed and each of the significantly matched sequences is used as an intermediate sequence for a new search (18–20). Very similar to the second approach is intermediate profile search, where PSI-BLAST with a fixed number of iterations is used instead of BLAST (14,21–23). To keep computation times manageable, a maximum sequence identity between intermediate sequences is normally enforced. For the same reason, the search depth, i.e. the number of intermediate sequence links, is generally limited to one, two, or three.

An extension of the third approach was implemented in SENSER (24), where sequences need not constitute a significant match (E-value $< 10^{-3}$) in order to be used as intermediate sequences. It suffices if they are found in the trailing end of the last PSI-BLAST search, i.e. with an E-value below 10. These sequences are used as seeds for the construction of new alignments by PSI-BLAST. If, starting from a new seed sequence, PSI-BLAST finds the query or one of its already accepted homologs with E-value lower than 10, the seed and its homologs are accepted. This concept, referred to as 'back-validation', relies on the asymmetry inherent in profile-sequence comparison. Owing to its sensitivity, SENSER was quite successful in several CASP competitions, but the unpredictable risk of false positives made manual checking of results necessary. We ascribe this to the heuristic, non-statistical nature of the back-validation criterion.

HHsenser was inspired by this method and has adopted the concept of seeds and trailing ends. Instead of back-validation, we use HMM–HMM comparison (25) in combination with a score correlation analysis. This should make HHsenser substantially more sensitive than straightforward implementations of intermediate profile search (14,22,23) (see caption of Figure 2). At the same time, it increases selectivity as compared

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

with SENSER. Since HHsenser has not been described before, we will give a brief explanation in the following section.

METHOD

HHsenser takes a single query sequence or a multiple alignment as input and returns two multiple alignments, a strict alignment and a permissive one. The strict alignment normally contains only homologous sequences, whereas the permissive one occasionally includes unrelated sequences. On the other hand, it often contains homologs not present in the strict alignment. The strict alignment may be more suitable for automated analyses, whereas the permissive alignment can be very useful for further expert analysis, where the occurrence of a limited number of non-homologous subgroups poses no problems (26).

The flow chart of HHsenser is shown in Figure 1. The query is used as the first seed in an iterated PSI-BLAST search with an *E*-value threshold of $E = 10^{-3}$ (steps 0–4). Steps 5–7 are skipped in the first pass, since the strict alignment is still empty. The alignment parsed out from the PSI-BLAST results is copied into the strict and permissive alignment (steps 8 and 9) and all matched sequence segments up to $E = 1$ are appended to the list of seeds (step 10). If there are seeds left in the list (step 1), a new seed is taken (step 2) and compared with all seeds for which an alignment has already been built (step 3). If the pairwise sequence identity to all these seeds is below an adjustable threshold (e.g. 30%), this seed is used to generate a new alignment with PSI-BLAST (step 4). The alignment is compared with the strict alignment by pairwise comparison of HMMs (step 5) (25). A correlation analysis is performed and an effective *E*-value is calculated. This is

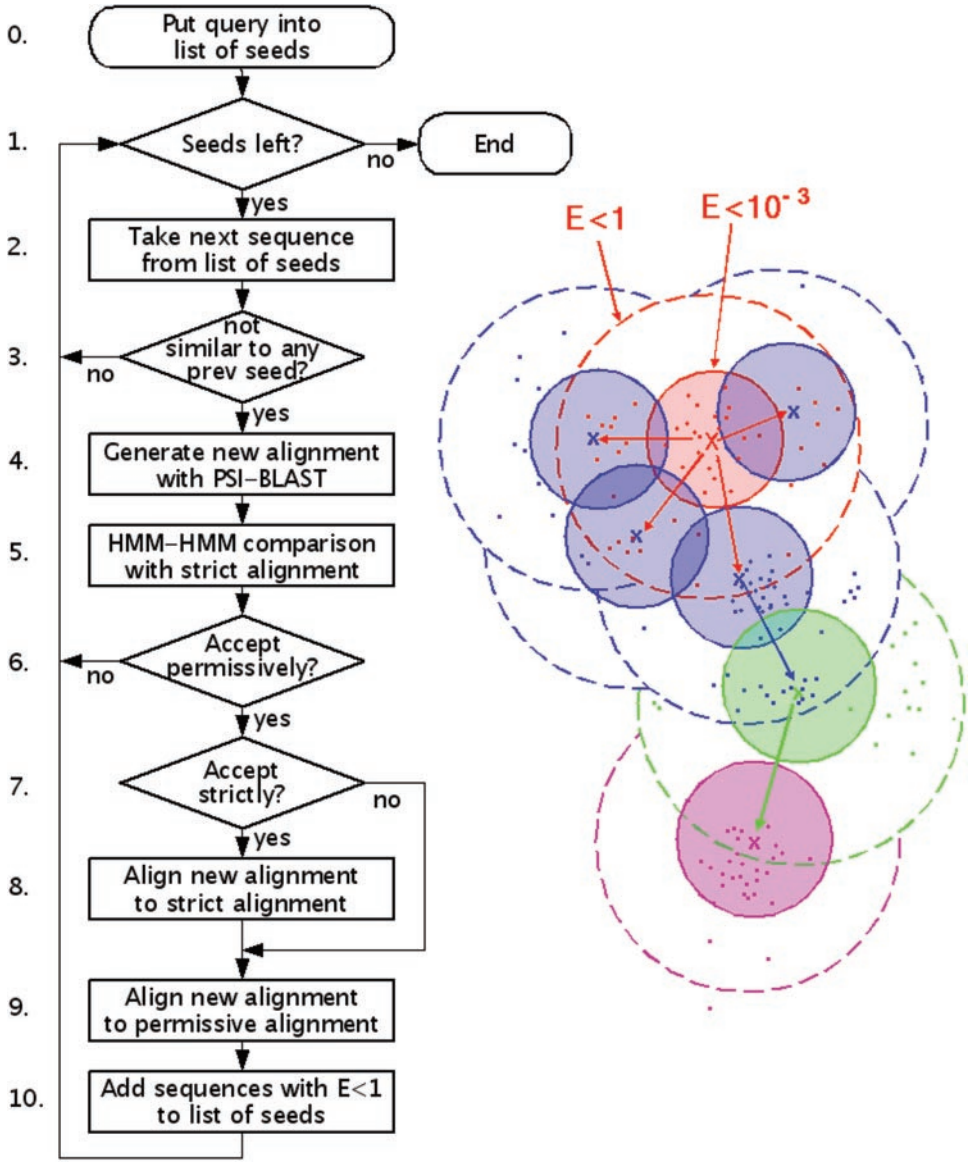

Figure 1. Simplified flow chart (left) and schematic diagram of HHsenser (right). The red X in the diagram is the query sequence, the other, smaller X's represent seed sequences from which new alignments are built (shaded disks). The large circles indicate the space from which new seeds are selected (arrows). The large search radius ($E < 1$) together with sensitive HMM–HMM comparison allows to jump wide gaps between related families (green arrow).

to account for the fact that the seed sequence may be a high-scoring false positive that was selected by PSI-BLAST for its chance similarity with the query profile from a large sequence database (see below). If the *P*-value from the HMM–HMM comparison is below 10^{-4} and the effective *E*-value is below 1 (step 6), the new alignment is appended to the permissive alignment by HMM–HMM comparison (step 9) (25). If the *P*-value is below 10^{-6} and the effective *E*-value is below 10^{-3}, the new alignment is appended also to the strict alignment (steps 7 and 8). Finally, all matches that scored better than $E = 1$ in the last PSI-BLAST search are added to the list of seeds. The process continues until all seeds have been processed. If a multiple alignment is given as input instead of a single sequence, this alignment is used to jump-start one round of PSI-BLAST, upon which the program proceeds directly to step 7.

To understand the necessity of calculating an effective *E*-value, assume, for example, that a false positive seed was found with an *E*-value of 0.1 and that this seed is a singleton, in other words the PSI-BLAST alignment contains only the seed itself. An HMM-HMM comparison between this single-sequence alignment and the strict query alignment would give approximately the same *P*-value as the profile-sequence comparison in PSI-BLAST. With an effective database size of 10^6 the *P*-value would therefore be $P \approx 10^{-7}$. Hence, the sequence would get a good *P*-value just because it was selected from a large database for its chance similarity with the query alignment.

Several measures have been devised to improve efficiency, sensitivity and selectivity:

- We have developed an 'end-pruning' procedure for PSI-BLAST that can significantly reduce the amount of non-homologous sequences creeping into the alignments (J. Söding, unpublished data).
- The maximum sequence identity threshold in step 3 is automatically adjusted according to the number of seeds in the list. It varies from 80% (<6 seeds) to 25% (>1000 seeds), thereby avoiding excessively long search times for large superfamilies.
- When extracting seeds in step 10, we add up to 50 residues on either side to the sequence segment matched by PSI-BLAST. In this way, we make sure that new seeds are not always shorter than their parent seed, since this would effectively limit the search depth.
- The first time that step 10 is performed, we extract a minimum number of four seeds from the PSI-BLAST results, even if they have *E*-values higher than 1. This increases chances of bridging the gap between singleton sequences and their closest homologs.
- Four databases can be selected: the non-redundant database (nr) at NCBI, and four filtered versions containing only eukaryotic, prokaryotic, bacterial, or archaeal sequences.
- We use versions of the sequence databases (e.g. nr90f and nr70f) that are clustered at 90 and 70% maximum pairwise sequence identity by CD-HIT (22). In PSI-BLAST searches, we start with the nr90f and automatically switch to the nr70f when >50 homologs are found. As a consequence, not all homologs contained in the complete nr database may appear in the alignments returned by HHsenser. If this is desired, however, the option 'Show representative sequences' can be disabled. In this case, the last PSI-BLAST search in step 4 will use the unclustered database.

In principle, our method works for single as well as multidomain sequences. However, it is recommended to break the sequence up into single domains (using HHpred or a similar method) since multi-domain sequences are at an increased risk of having non-homologous sequence segments included during the PSI-BLAST searches.

From the symmetry of HMM–HMM comparison one might be tempted to conclude that HHsenser will find the same sequences no matter with which particular sequence in a family it is started with. This is wrong for two reasons. First, even though the HMM–HMM comparison step is symmetric, the PSI-BLAST searches to detect new seed sequences are not. A sequence that has no homologs up to a BLAST *E*-value of 10 might very well come out as a significant match when the database is searched with a PSI-BLAST profile of related sequences. Second, the query sequence automatically defines the match state assignment for the strict and permissive HMMs used in the HMM–HMM comparisons in steps 5, 8 and 9: all aligment positions with a residue in the query sequence are assigned to match states, all others are inserts. Therefore, when starting from a different sequence, different positions will normally be assigned to match and insert states. In practice, we find that most of the times the choice of the start sequence does not influence the number of subgroups found and the sets of detected sequences do overlap to a high degree.

A downside of HHsenser, as of other intermediate profile search methods, are the long computation times involved, in particular when the query sequence is a member of a large superfamily. With ~1000 homologs in the nr90f database, the calculation time is typically <5 h, but for the largest superfamilies (like AAA+ ATPases, outer membrane beta barrels, TIM barrels or immunoglobulins) a search may take several days. To avoid straining our computational resources too much, we have therefore set a limit of 500 residues in the input sequence. Also, when a number of 5000 homologs is exceeded the search will terminate and the current results are returned. Last, the number of HHsenser jobs is limited to 10 and additional jobs will be queued. Users who would like to perform searches without these limitations are asked to contact us.

EXAMPLE APPLICATIONS

Pei and Grishin recently found by manual transitive PSI-BLAST searches and sequence analysis that the putative endopeptidase P5 from bacteriophage phi-6 and the family of lytic transglycosilases that cleave bacterial peptidoglycans are distantly related (27). When HHsenser is launched with P5 (sp|P07582|VLYS_BPPH6) and default parameters, it returns 1591 sequences in the strict alignment and 1991 sequences in the permissive one (Figure 2). A clustering analysis with CLANS (28) reveals that P5 from phi-6 is indeed a distant member of a superfamily containing various families of lytic transglycosilases (PilT, LysM, SLT, mltC/mltE, resuscitation-promoting factor Rpf), a group of putative periplasmic-binding transport proteins, lysozyme C and G, and several as yet undescribed groups of hypothetical proteins. We

Figure 2. Sample output of HHsenser showing a part of the stringent alignment of lytic transglycosilases obtained with protein P5 from bacteriophage phi-6 as starting sequence. The overlaid window shows a JalView applet (30). Tabs allow to switch to the stringent alignment, the permissive alignment, the sequences rejected in the course of the transitive search and the various intermediate aligments accepted as homologs. Files can be downloaded via links, either with or without inserts relative to the query sequence.

found no obvious false positives in either the strict or permissive alignment. This search took ~3 h to complete on a 2.2 GHz AMD64 PC.

A further example is the family of AbrB-like transcription factors which we have studied recently (26). When the structure of AbrB was solved, we were surprised to learn that it adopted a new fold even though previous sequence analyses conducted in our group had indicated that it should be related to MazE and MraZ, two proteins which fold into swapped-hairpin barrels. We therefore decided to study AbrB and its homologs in more detail. HHsenser was started with the sequence of AbrB in *Bacillus subtilis* (ABRB_BACSU) and after ~5 h it returned a permissive alignment with 724 sequences that contained no identifiable false positives. Both MazE and the two domains of MraZ showed up in our alignments. We therefore decided to redetermine the structure of AbrB and found indeed that it is very similar to the swapped-hairpin barrel of MazE and MraZ. Clustering of the sequences with CLANS resulted in eight major groups of bacterial transcription factors, six of which had been described before and two of which were groups of hypothetical proteins from cyanobacteria and proteobacteria. With the default

parameters in our web server, all but the proteobacterial group of hypothetical proteins are recovered without false positives.

CONCLUSION

Starting from a single sequence, HHsenser is able to explore the space of homologous sequences and align these to the query sequence by HMM–HMM comparison. Its distinctive feature is its ability to jump between distantly related families while producing few or no false positives. In combination with an interactive clustering method such as the publicly available CLANS program (28), HHsenser is in our experience a powerful tool for the functional annotation and evolutionary analysis of whole protein superfamilies (26).

Its main drawback is the long computation times necessary for these exhaustive searches. Since we believe that HHsenser can be particularly useful for singletons or sequences with only few PSI-BLAST-detectable homologs, we offer a quick, non-exhaustive version that can be called from the results page of our structure and function prediction server HHpred (29). The search will terminate as soon as 100 homologs have been found, which will normally take <15 min.

ACKNOWLEDGEMENTS

We would like to thank Kristin Koretke for many fruitful discussions related to this work and for her help in testing HHsenser. Funding from the Max-Planck Society to pay for the open access publication charge is acknowledged.

Conflict of interest statement. None declared.

REFERENCES

1. Przybylski,D. and Rost,B. (2002) Alignments grow, secondary structure prediction improves. *Proteins*, **46**, 197–205.
2. Garg,A., Kaur,H. and Raghava,G.P. (2005) Real value prediction of solvent accessibility in proteins using multiple sequence alignment and secondary structure. *Proteins*, **61**, 318–324.
3. Peng,K., Vucetic,S., Radivojac,P., Brown,C.J., Dunker,A.K. and Obradovic,Z. (2005) Optimizing long intrinsic disorder predictors with protein evolutionary information. *J. Bioinform. Comput. Biol.*, **3**, 35–60.
4. Viklund,H. and Elofsson,A. (2004) Best alpha-helical transmembrane protein topology predictions are achieved using hidden Markov models and evolutionary information. *Protein Sci.*, **13**, 1908–1917.
5. Punta,M. and Rost,B. (2005) PROFcon: novel prediction of long-range contacts. *Bioinformatics*, **21**, 2960–2968.
6. Res,I., Mihalek,I. and Lichtarge,O. (2005) An evolution based classifier for prediction of protein interfaces without using protein structures. *Bioinformatics*, **21**, 2496–2501.
7. Nair,R. and Rost,B. (2003) Better prediction of sub-cellular localization by combining evolutionary and structural information. *Proteins*, **53**, 917–930.
8. Söding,J., Remmert,M., Biegert,A. and Lupas,A. (2006) HHrep: *de novo* protein repeat detection and the origin of TIM barrels. *Nucleic Acids Res.*
9. Saunders,C.T. and Baker,D. (2002) Evaluation of structural and evolutionary contributions to deleterious mutation prediction. *J. Mol. Biol.*, **322**, 891–901.
10. Lichtarge,O., Bourne,H.R. and Cohen,F.E. (1996) An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.*, **257**, 342–358.
11. Hannenhalli,S.S. and Russell,R.B. (2000) Analysis and prediction of functional sub-types from protein sequence alignments. *J. Mol. Biol.*, **303**, 61–76.
12. delSol Mesa,A., Pazos,F. and Valencia,A. (2003) Automatic methods for predicting functionally important residues. *J. Mol. Biol.*, **326**, 1289–1302.
13. Nimrod,G., Glaser,F., Steinberg,D., Ben-Tal,N. and Pupko,T. (2005) *In silico* identification of functional regions in proteins. *Bioinformatics*, **21**, i328.
14. Wallner,B., Fang,H., Ohlson,T., Frey-Skott,J. and Elofsson,A. (2004) Using evolutionary information for the query and target improves fold recognition. *Proteins*, **54**, 342–350.
15. 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.
16. Altschul,S.F., Madden,T.L., Schäffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
17. Krogh,A., Brown,M., Mian,I.S., Sjolander,K. and Haussler,D. (1994) Hidden Markov models in computational biology. Applications to protein modeling. *J. Mol. Biol.*, **235**, 1501–1531.
18. Park,J., Teichmann,S.A., Hubbard,T. and Chothia,C. (1997) Intermediate sequences increase the detection of homology between sequences. *J. Mol. Biol.*, **273**, 349–354.
19. Gerstein,M. (1998) Measurement of the effectiveness of transitive sequence comparison, through a third 'intermediate' sequence. *Bioinformatics*, **14**, 707–714.
20. Li,W., Pio,F., Pawlowski,K. and Godzik,A. (2000) Saturated BLAST: an automated multiple intermediate sequence search used to detect distant homology. *Bioinformatics*, **16**, 1105–1110.
21. Margelevicius,M. and Venclovas,C. (2005) PSI-BLAST-ISS: an intermediate sequence search tool for estimation of the position-specific alignment reliability. *BMC. Bioinformatics*, **6**, 185.
22. Li,W., Jaroszewski,L. and Godzik,A. (2002) Sequence clustering strategies improve remote homology recognitions while reducing search times. *Protein. Eng.*, **15**, 643–649.
23. Sandhya,S., Chakrabarti,S., Abhinandan,K.R., Sowdhamini,R. and Srinivasan,N. (2005) Assessment of a rigorous transitive profile based search method to detect remotely similar proteins. *J. Biomol. Struct. Dyn.*, **23**, 283–298.
24. Koretke,K.K., Russell,R.B. and Lupas,A.N. (2002) Fold recognition without folds. *Protein Sci.*, **11**, 1575–1579.
25. Söding,J. (2005) Protein homology detection by HMM–HMM comparison. *Bioinformatics*, **21**, 951–960.
26. Coles,M., Djuranovic,S., Söding,J., Frickey,T., Koretke,K., Truffault,V., Martin,J. and Lupas,A.N. (2005) AbrB-like transcription factors assume a swapped hairpin fold that is evolutionarily related to double-psi beta barrels. *Structure*, **13**, 919–928.
27. Pei,J. and Grishin,N.V. (2005) The P5 protein from bacteriophage phi-6 is a distant homolog of lytic transglycosylases. *Protein Sci.*, **14**, 1370–1374.
28. Frickey,T. and Lupas,A. (2004) CLANS: a Java application for visualizing protein families based on pairwise similarity. *Bioinformatics*, **20**, 3702–3704.
29. 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–W248.
30. Clamp,M., Cuff,J., Searle,S.M. and Barton,G.J. (2004) The Jalview Java alignment editor. *Bioinformatics*, **20**, 426–427.