Available online at www.sciencedirect.com
ScienceDirect
Current Opinion in Structural Biology

Protein sequence comparison and fold recognition: progress and good-practice benchmarking

Johannes Söding and Michael Remmert

Protein sequence comparison methods have grown increasingly sensitive during the last decade and can often identify distantly related proteins sharing a common ancestor some 3 billion years ago. Although cellular function is not conserved so long, molecular functions and structures of protein domains often are. In combination with a domain-centered approach to function and structure prediction, modern remote homology detection methods have a great and largely underexploited potential for elucidating protein functions and evolution. Advances during the last few years include nonlinear scoring functions combining various sequence features, the use of sequence context information, and powerful new software packages. Since progress depends on realistically assessing new and existing methods and published benchmarks are often hard to compare, we propose 10 rules of good-practice benchmarking.

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

Corresponding author: Söding, Johannes (soeding@genzentrum.lmu.de)

Current Opinion in Structural Biology 2011, **21**:404–411

This review comes from a themed issue on
Sequences and topology
Edited by Julian Gough and Keith Dunker

Available online 31 March 2011

0959-440X/$ – see front matter

DOI 10.1016/j.sbi.2011.03.005

Introduction

Protein sequence search methods are among the most widely used computational tools, with BLAST and PSI-BLAST alone having been cited over 60 000 times [1,2]. These methods help to infer the functions and structures of proteins from those of homologous (i.e., related) proteins found in the sequence databases. A homologous relationship is assumed if the sequence similarity is sufficiently high to exclude a chance similarity. Below about 40% sequence identity cellular function is usually not conserved. However, proteins are composed of one or several domains, the structural, functional, and evolutionary units of proteins. Their molecular functions (e.g. ATP-driven motor activity or RNA binding) and structures are often conserved over billions of years, below the limit of detectable sequence similarity [3,4]. This high degree of conservation makes the inference of structures and functions of protein domains from homologous domains potentially very powerful.

In combination with this domain-centered approach to function and structure prediction, the growing sensitivity of sequence comparison methods has considerably expanded the scope of applications in recent years. First, in homology modeling, still the only practical method for routine protein structure prediction, the identification of suited template proteins (*fold recognition*) and the generation of optimal query-template alignments constitute the major quality bottlenecks. Second, the functional annotation of proteins based on their predicted domains is having a profound impact. Numerous domain family databases have been developed that contain multiple sequence alignments (MSAs) for thousands of domain families together with detailed annotations about their functions, critical residues, structures, interactions, phylogenetic spread, etc.[5–7]. All major sequence databases use these domain databases for annotation. Third, Building MSAs by iterative sequence searches has become one of the most essential bioinformatic applications, since multiple alignments are a key intermediate step for almost all sequence-based predictions (e.g. secondary structure, tertiary structure, membrane helices, functional residues, and interaction motifs).

This review summarizes progress in sequence searching and pairwise alignment during the last two to three years (see [8,9] for earlier reviews). Since information is power, the first four sections introduce the various sources of information used to detect remote homologies. Promising algorithmic and technical advances are described next, and in the last two sections, we discuss rules for good practice benchmarking that we hope can serve as a guideline in the field.

Evolutionary information and sequence profiles

The sensitivity and alignment quality depend crucially on the amount of information which is used to compare proteins. Today, the most sensitive methods for fold recognition use sequence profiles to represent both the query and the database proteins. Sequence profiles contain position-specific substitution scores that are computed from the frequencies of amino acids at each position of a multiple alignment of related sequences. Modern methods employ a standard two-step approach. They build up a profile using an iterative profile-to-sequence search method such as PSI-BLAST [2] and

then search with this query profile through a precomputed database of profiles using profile–profile comparison. In the two most recent CASP (Critical Assessment of techniques for protein Structure Prediction) competitions, all 20 top-ranked servers (out of ~80) used such an approach to identify and align suitable template proteins [10] (for CASP9 results see http://predictioncenter.org/casp9/groups_analysis.cgi?type=server). Further improvements were realized during the last decade by extending sequence profiles to *profile hidden Markov models* (profile HMMs), which include position-specific gap penalties derived from frequencies of insertions and deletions [11••–14].

1D structural properties

In addition to sequence profiles, top-performing structure prediction methods compare the predicted secondary structure of the query protein with the actual secondary structure of the template protein. Such *1D properties* defined for each position have a big advantage: their similarity scores can be combined with the similarity score between profile columns in the dynamic programming algorithms that calculate the optimal alignment [15]. Hence 1D similarity scores may improve both the sensitivity of fold recognition and the alignment quality. Although secondary structure has had the largest impact, many other 1D scores have been proposed. To increase information content over the three secondary structure states, SAM-T2K and SP5 define finer alphabets of backbone structure [12,16,17]. Other scores that have become widely used recently are predicted solvent accessibility, predicted number of tertiary residue–residue contacts (coordination number) [12, 18, 19, 20••, 21], and 1D environmental fitness scores [22], which evaluate how well the amino acid distribution at each query position would fit into the structural environment at each template position.

Patterns and sequence context

The 1D scores discussed above are limited to searching for templates with known structure. If 1D predictions are compared to 1D predictions, structures are not needed and, surprisingly, this works almost as well [13,23]. But how can comparing profile-based 1D predictions add information to the profile–profile comparison? Profile column scores ignore correlations between columns. In contrast, 1D predictions are done on context windows. Comparing 1D predictions therefore amounts to scoring the similarity of local amino acid patterns which may contain strong inter-column correlations.

A more direct method to include information from sequence contexts was developed in our group. It generates sequence profiles from single sequences by predicting amino acid substitution probabilities given the sequence context (13 positions) around each residue. We could show that both sensitivity and alignment quality of our context-specific version of BLAST were increased by a large margin with negligible computational overhead [24••]. The idea is quite general and can be applied to sequence–profile and profile–profile alignment schemes, models of molecular evolution, and alignment of oligonucleotide sequences.

Taxonomy, function, and gene context

Other types of information have been explored. Matched database sequences whose taxa are distant with respect to the taxa in the query profile are more likely to be false positives than sequences from closely related taxa [25]. Other studies use cataloged protein–protein interactions [26,27], but the approach is only effective for proteins with a sufficient number of known interactions. In bacteria conserved genomic context and sequence length can help to improve sensitivity of marginally significant matches [28]. This concept of conserved genomic context could be transferred to protein domain context, which is also conserved to some degree [29].

Comparing profile columns

What is the best way to score the similarity of profile columns during profile–profile comparison? Several studies compared heuristic column scores and found little variation in their performance (see e.g. [30]). However, the sensitivity of profile–profile alignment methods can be improved by rewarding the alignment of low-entropy columns to each other [31–33]. This may indicate that the way profile columns are compared may not be optimal, in particular how pseudocounts are added [34]. Altschul *et al.* propose a generalization of the log-odds score for pairwise sequence comparison [35••] to the pairwise comparison of MSA columns. They suggest to calculate the logarithm of the probability that both alignment columns were generated by the same underlying ancestral distribution, divided by the probabilities that the two alignment columns were generated by their own independent ancestral distributions. Instead of using a heuristic to add pseudocounts before the comparison, pseudocounts emerge naturally from Dirichlet priors on the ancestral amino acid distributions. It will be interesting to see more thorough tests of the proposed scheme with state-of-the-art profile–profile aligners.

Nonlinear scoring functions

Most methods for fold recognition calculate a total per-residue score by adding the profile column score and various structure-derived 1D scores independently of each other, using constant weights. Batzoglou and coworkers report substantial improvements in pairwise sequence alignment quality by learning scores and gap penalties for each combination of features, such as amino acids, secondary structure, solvent accessibility, and hydropathy index of sequence context. Their software CONTRalign models the probability of an alignment given the two sequences and their features as a pair

conditional random field (CRF). They learn the model parameters by maximizing the log likelihood of a set of training alignments [36•]. Another study confirms the advantage of making match scores and gap penalties dependent on secondary structure and learning optimal score function parameters [37]. MICalign extend this idea to using sequence profiles for both sequences and adding (predicted or actual) secondary structure and solvent accessibility [38].

These methods have a fixed score for each combination of features and they add these scores linearly, allowing them to learn dependencies between features. In this linear formulation, the CRF's parameters can be efficiently optimized by convex quadratic programming, avoiding the risk of running into local optima. Peng and Xu observe that exhaustively learning a score for all feature combinations limits the number of features and discretization levels while requiring huge numbers of parameters. They take a conceptionally big step in formulating the score as a true nonlinear function of the sequence features whose number of parameters can be much lower than the number of all feature combinations. The parameters are optimized using the gradient boost algorithm [20••,39]. The success of their method is confirmed in the recent CASP9 benchmark, where their RAPTOR servers were ranked third to fifth out of 78 participating servers.

Homologous overextension

Most high-scoring false positive matches in PSI-BLAST are caused by corrupted alignments, which contain non-homologous sequence stretches at the ends of correctly aligned homologous regions. These stretches can recruit many more unrelated segments in further search iterations. Several solutions have been proposed. The buildali.pl script in the HHsearch package [13] prunes away the ends of sequences to be included in the alignment if their score per column is below a specified threshold. In [40], a statistical framework is developed to calculate *P*-values for the alignment ends and to remove overaligned ends. Another possibility is to combine *E*-values from the last search iteration with those of the second search iteration, which does not usually suffer much from overextension [41]. Pearson and coworkers show that by freezing the alignment of a sequence that has already been found previously, the specificity increases four-fold to eight-fold [42]. Finally, the iterative search methods HHMER3 and HHblits (see below) replace the Smith–Waterman algorithm by the maximum accuracy algorithm [43,44], which is much less prone to overextension.

Iterative profile HMM searches

A major practical and technical advance has been accomplished by speeding up the popular HMM-based search method HMMER by a factor ~100. The new version HMMER3 is now the standard search tool for Pfam [5]. It is only three to four times slower than NCBI PSI-BLAST while possessing much better sensitivity and alignment quality [11••]. HMMER3 is fast enough for iterative searches through a comprehensive sequence database. The drastic speed improvement is mainly due to a fast prefilter based on SIMD (single instruction multiple data) technology. With SIMD instructions, the 128-bit arithmetic–logic units that form a part of modern CPU cores can compute 16 parallel operations on single-byte variables per clock cycle [45•,46]. Sensitivity was improved further by using the Forward instead of the Viterbi algorithm for scoring matches and replacing the time-consuming sampling-based *E*-value calculation by a heuristic approximation [47]. A similar SIMD-based strategy was pursued in our group to speed up searches with our HHsearch software for pairwise comparison of profile HMMs [13]. The new iterative version of HHsearch, HHblits, runs faster than PSI-BLAST and achieves sensitivities and alignment qualities much superior to HMMER3 (Remmert *et al.*, unpublished). An overview of free software and web servers for sensitive sequence searching is given in Table 1.

Protein similarity networks

When similarities between database members are known, the homology of the query Q with a database protein T can be supported by paths through the similarity network that lead from node Q via intermediate nodes X and Y to node T (Figure 1 top).

The *empirical feature map* uses the kernel function $k(X, Y) = \sum_Z S(X, Z)S(Y, Z)$, where $S(X, Y)$ is a matrix of pairwise similarities. This kernel has been employed by numerous machine learning approaches. They train a support vector machine (SVM) for every fold in the training set and predict the fold whose SVM yields the highest score [48]. Two simple methods similar to the empirical feature map have been proposed [49,50] which calculate network-based similarities $S_{net}(X, Y)$ using a heuristic similarity measure between the empirical feature vectors $S(X, \cdot)$ and $S(Y, \cdot)$.

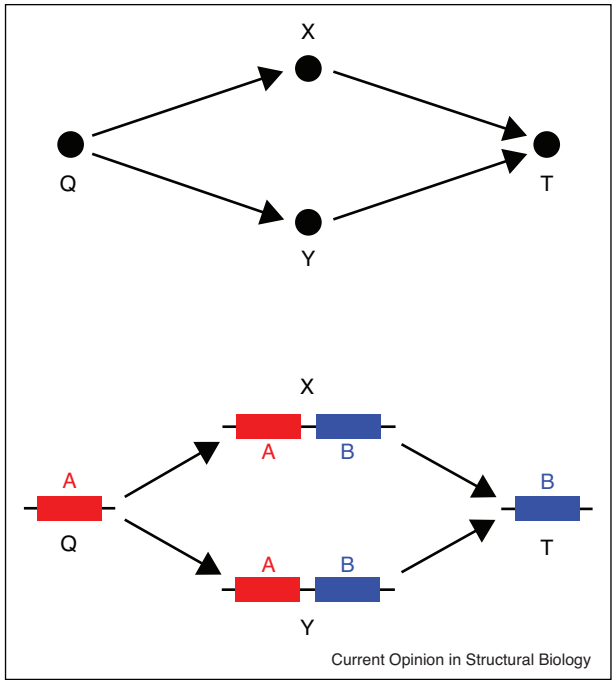
These SVM-based approaches can only predict folds they were trained with. Two recent algorithms, RANKPROP[51,52] and ProtEmbed [53•], are independent of any fold definitions. They can make use of network information to better rank database matches and can therefore also be applied to regular sequences searches. RANKPROP propagates in an iterative fashion virtual activities from the query through the network of database members. After convergence, the database members are ranked by their activity. ProtEmbed learns a large-scale embedding of the empirical feature vectors obtained by HHsearch all-against-all comparison into a lower-dimensional 'semantic space' while trying to conserve pairwise distances as well as possible [53•].

Table 1

Tools and web servers for protein remote homology detection

Name	Description	Software/server	Ref.
PSI-BLAST	General-purpose, iterative profile-to-sequence search; very fast	http://blast.ncbi.nlm.nih.gov/Blast.cgi ftp://ftp.ncbi.nih.gov/blast/	[2]
SAM	Iterative sequence search based on HMM-to-sequence comparison for remote homology detection and protein structure prediction	http://compbio.soe.ucsc.edu/sam.html http://compbio.soe.ucsc.edu/SAM_T08/T08-query.html	[12]
HMMER3	General-purpose iterative sequence search based on HMM-to-sequence comparison; fast, better sensitivity and alignment quality than PSI-BLAST	http://hmmer.janelia.org/ http://hmmer.janelia.org/search	[11••]
HHblits	General-purpose iterative sequence search based on HMM–HMM comparison; very fast, better sensitivity and alignment quality than PSI-BLAST and HMMER3	ftp://toolkit.lmb.uni-muenchen.de/HHblits/ http://toolkit.lmb.uni-muenchen.de/hhblits	–
COMPASS	Remote homology detection and fold recognition based on profile–profile comparison	ftp://iole.swmed.edu/pub/compass/ http://prodata.swmed.edu/compass/	[62]
PROCAIN	Remote homology detection and fold recognition based on profile–profile comparison	http://prodata.swmed.edu/procain/download/ http://prodata.swmed.edu/procain/	[32]
COMA	Remote homology detection and fold recognition based on profile–profile comparison	http://www.ibt.lt/bioinformatics/coma/ http://bioinformatics.ibt.lt:8085/coma/	[63]
PRC	Remote homology detection and fold recognition based on HMM–HMM comparison	http://supfam.mrc-lmb.cam.ac.uk/PRC/ http://www.ibi.vu.nl/programs/prcwww/	[14]
HHsearch/HHpred	Remote homology detection, fold recognition, and structure prediction based on HMM–HMM comparison	ftp://toolkit.lmb.uni-muenchen.de/HHsearch/ http://toolkit.lmb.uni-muenchen.de/hhpred	[13]

Figure 1



Transitive paths of similarity may help to predict remote relationship between proteins *Q* and *T*. When *Q* is similar to *X* and *Y*, and *X*, *Y* are similar to *T*, this is taken as an indication that *Q* is homologous to *T*. This assumption is not valid for multi-domain proteins: *Q* and *T* have no domains in common although the transitive paths via *X* or *Y* have high pairwise similarities due to shared domains.

All discussed network-based methods rely on the assumption of transitivity (Figure 1). This limits their general applicability since most proteins, particularly in eukaryotes, are composed of multiple structural domains. An elegant approach for transitive homology search that circumvents these problems was presented by Heger *et al.* [54]. They associate the nodes in the similarity network with the residues of all database proteins and draw edges whose weights reflect how likely two residues are to be aligned with each other.

Analysing alignment quality and detection sensitivity

Published benchmarks on sequence comparison methods differ much in their setup and often seem to come to divergent conclusions. A standardization of the benchmark procedures and analyses would help to make benchmarks more comparable and to facilitate progress in the field. We need to be able to objectively assess, firstly, the quality of the produced alignments and, secondly, the sensitivity for detecting homologous protein pairs.

Alignment quality is usually measured by comparing the sequence-based alignments to a set of 'gold standard' alignments generated by structural alignment. But because structural alignment is non-trivial and different programs often produce quite different alignments, we advocate the reference-free method [55], which assesses the alignment quality by how well the aligned pairs are structurally superposable. *Alignment sensitivity* is measured as the weighted fraction of query residues that are superposable to the template. The complementary measure, *alignment precision*, is the weighted fraction of aligned

residue pairs that are superposable. Here, the residue weights go from 1 for a perfect superposition to 0 for large spatial divergence. The precise distance dependence differs between programs [56,57].

To measure the sensitivity for detecting homologous proteins, a 'gold standard' set of known homologous (true positive, TP) and non-homologous (false positive, FP) protein pairs is needed. Since structure is conserved better than sequence, benchmark sets are usually drawn from a database of domains of known structure. Although CATH [58], Pfam [5], and COPS [59] are sometimes used, the SCOP database [60] has emerged as *de facto* standard. In SCOP, members of a superfamily are assumed to be homologous. Pairs from the same fold but different superfamilies are treated as 'unknown'. Members of different folds are non-homologous, except for Rossman-like folds (c.2–c.5, c.27 and 28, c.30 and 31) and the four- to eight-bladed β-propellers (b.66–b.70). Pairs from these groups of folds should be treated as 'unknown'. An alternative that avoids a catalog of specific exceptions is the reference-free scheme proposed by Qi [55]. There, the classification of a pair as TP or FP depends on the structurally evaluated quality of the method's suggested alignment. The drawback is that the sensitivity cannot be evaluated independently of alignments, and methods producing longer alignments will be at an advantage.

A Receiver Operating Characteristic (ROC) plot measures how well a method ranks all protein pairs with respect to each other *over all searches*. It shows the number of TP pairs (homologous pairs above score threshold) as a function of the number of FP pairs (non-homologous pairs above threshold) (Figure 2a). Various scaled and inverted versions exist, such as sensitivity versus selectivity, or recall versus precision.

A ROC5 plot assesses how well a method ranks the matched proteins *within each search*. For each query protein (or family), one calculates a ROC plot and computes the ROC5 value, that is the area under the ROC curve up to the fifth FP. The ROC5 plot shows the fraction of queries for which the ROC5 value is above the threshold on the x-axis (Figure 2d). A sensitive method will achieve high ROC5 values for a high fraction of queries.

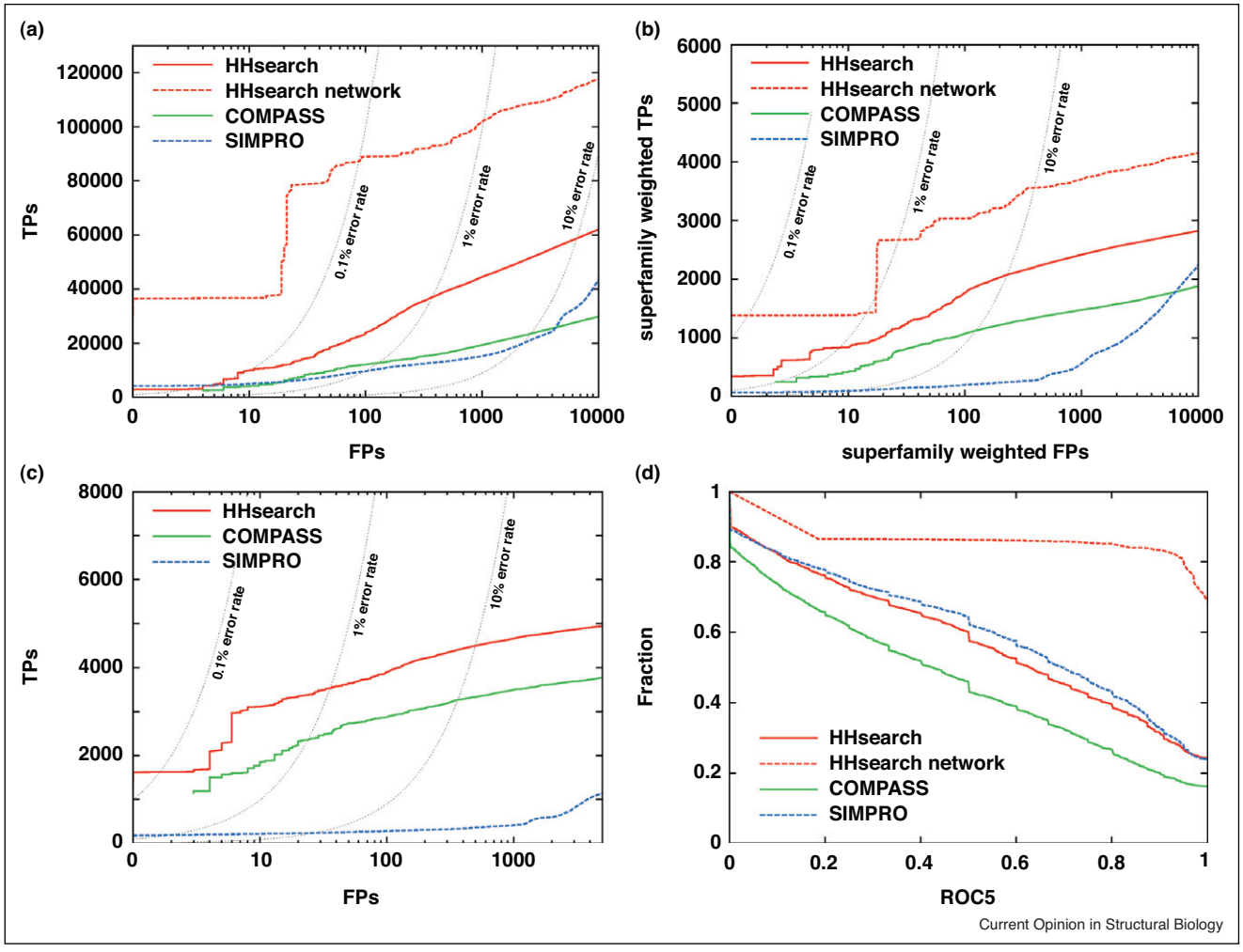
The ROC5 analysis is complimentary to the ROC plot. Whereas the ROC analysis is more relevant for automatic methods that rely on the score for deciding whether to accept a match as homologous or not, the ROC5 type of analysis may be more relevant for a human user who is willing to look through a certain number of matches to his query protein (e.g. 5) to decide which of these to follow up.

Ten rules of good-practice benchmarking

Benchmarking the sensitivity of sequence search methods in a way that results can be compared among studies is by no means straightforward. We describe here what we consider as the ten rules of good-practice benchmarking for these methods.

1. Ensure that training and test proteins are not too similar. Machine learning methods typically train thousands of free parameter and can learn fold classifications of training proteins by heart. Therefore, correct fold predictions for test proteins that are very similar to training proteins can be trivial. As an example, a maximum pairwise sequence identity of 25% between training and test *sequences* is not sufficient, since at that similarity the *profiles* built from these sequences can be very similar. When using SCOP, training and test sets should rather be split along superfamily boundaries.
2. Optimize parameters on the training set or an independent validation set, not on the test set. ROC plot analyses are particularly prone to overtraining parameters on the test set, since the ROC plot depends on only a few tens or hundreds of high-scoring false positives in the relevant FDR range.
3. Consult authors on good parameter settings for their methods. An alternative is to optimize the most important parameter(s) of all tested methods.
4. Choose a standard benchmark set to facilitate comparison with other studies.
5. Choose a benchmark set of maximum possible size. Differences on small test sets can be due to chance rather than to significant differences. When training and test data are scarce, 10-fold cross validation is advisable.
6. Apply superfamily weights in ROC analyses. Because the number of homologous pairs scales as the number of members squared, large superfamilies would otherwise dominate the ROC plot (Figure 2b), in particular since the few highest scoring FPs strongly influence the ROC plot. Changing the weighting from family-based to superfamily-based decreased the sensitivity gain of our method CS-BLAST [24••] over BLAST from 140% to 40%.
7. Select the FP range in the ROC plot that corresponds to relevant FDRs. The relevant FDR range is roughly between 0.1% and 10%, since an FDR of 10% typically corresponds to a *marginal FDR* of ~50%. (The marginal FDR measures what fraction of predictions with scores between a threshold S and some $S + \Delta S$ are correct.).
8. Check that the tested methods use only information that is actually available in a practical setting.
9. Check that the benchmark simulates the methods' envisaged use. Example: suppose we want to test a fold recognition method that uses similarities between SCOP members (e.g. [49,50,48]). We should only evaluate it on the top match per query, since that match determines the superfamily or fold to predict. Whether the 10th match, say, also has the correct fold

Figure 2



Analysis of detection sensitivity. The test set consists of all SCOP domains (v. 1.75) with 20% maximum pairwise sequence identity. SIMPRO is a network-based method, COMPASS (3.1) and HHsearch (1.6.0) are methods based on profile–profile and HMM–HMM comparison, respectively. HHsearch_network is a trivial extension that ranks members from the superfamily of the best HHsearch hit on top. **(a)** A ROC plot shows the number of true positive (TP) over false positive (FP) protein pairs detected above the score threshold. Superfamilies of size n contribute n^2 TP pairs, therefore large superfamilies dominate the performance. Dotted grey lines indicate false discovery rates of 0.1%, 1% and 10%. **(b)** Same as a, each pair is weighted by 1 over the size of the query's superfamily. **(c)** Same as b, only the best match per query is evaluated. **(d)** ROC5 plot: fraction of query proteins whose ROC5 value is above the ROC5 threshold on the x-axis. This ROC5 analysis is much less prone to overfitting than the ROC analyses in a–c. The analyses in a, b, and d are not suited to evaluate fold recognition methods that use network information, such as SIMPRO and HHsearch_network (see point 9 in 'Ten rules of good-practice benchmarking' section). (SIMPRO was implemented according to the description in [50]. All data and scripts can be downloaded at ftp://toolkit.lmb.uni-muenchen.de/COSB-2011-Seq-comparison).

is irrelevant in practice. Figure 2b and c shows how the performance of a network-based method (SIMPRO [50]) drops in comparison to non-network based methods when evaluated only on best matches.

10. It should become good practice to offer for download all the data and scripts that are necessary to reproduce published benchmark results.

Conclusion

A great number of advances in sequence searching have been made during the last decade, many in the context of protein structure prediction. Yet most have not been applied generally in sequence searching. It is time to consolidate these ideas and turn them into practically useful, fast, and user-friendly tools for general purpose sequence searching that are much more powerful than the tools of the last millennium that still dominate the field. Protein structure prediction and domain annotation are becoming mainstream applications, and the growing importance of domain family databases such as Pfam, InterPro, and CDD [5–7] for functional domain annotation is accelerating this trend. We envisage that in the future raw sequence databases will be superseded by

highly structured domain family databases that cover the entire sequence space. They will be searched with sensitive methods that can combine information from diverse sources. Such databases would represent the vast and fast growing amount of information in a more economic and efficient way than today's databases of raw sequences [61]. We believe that many of the ideas highlighted here will be integrated into mainstream sequence search tools that will better realize the great potential benefits of the fast-growing sequence databases [61].

Acknowledgements

We acknowledge financial support by the Deutsche Forschungsgemeinschaft and SFB646. We thank Julian Gough, Ceslovas Venclovas, and Jinbo Xu for comments and discussions.

References and recommended reading

Papers of special interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool**. *J Mol Biol* 1990, **215**:403-410.

2. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs**. *Nucleic Acid Res* 1997, **25**:3389-3402.

3. Murzin AG: **How far divergent evolution goes in proteins**. *Curr Opin Struct Biol* 1998, **8**:380-387.

4. Alva V, Koretke KK, Coles M, Lupas AN: **Cradle-loop barrels and the concept of metafolds in protein classification by natural descent**. *Curr Opin Struct Biol* 2008, **18**:358-365.

5. Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington J, Gavin OL, Gunasekaran P, Ceric G, Forslund K *et al.*: **The Pfam protein families database**. *Nucleic Acids Res* 2010, **38**:D211-D222.

6. Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Das U, Daugherty L, Duquenne L *et al.*: **InterPro: the integrative protein signature database**. *Nucleic Acids Res* 2009, **37**:D205-D211.

7. Marchler-Bauer A, Lu S, Anderson JB, Chitsaz F, Derbyshire MK, DeWeese-Scott C, Fong JH, Geer LY, Geer RC, Gonzales NR *et al.*: **CDD: a Conserved Domain Database for the functional annotation of proteins**. *Nucleic Acids Res* 2011, **39**:D225-D229.

8. Dunbrack RL: **Sequence comparison and protein structure prediction**. *Curr Opin Struct Biol* 2006, **16**:374-384.

9. Xu D: **Computational methods for protein sequence comparison and search**. *Curr Protoc Prot Sci* 2009, **56**: 2.1.1–2.1.27.

10. Cozzetto D, Kryshtafovych A, Fidelis K, Moult J, Rost B, Tramontano A: **Evaluation of template-based models in CASP8 with standard measures**. *Proteins* 2009, **77**:18-28.

11. •• Eddy S: **A new generation of homology search tools based on probabilistic inference**. *Genome Inform* 2009, **23**:205-211.
The article describes the software package HMMER3 for HMM-based sequence search, which is about 100 times faster than its popular predecessor HMMER2. HMMER3 is fast enough for iterative searches but it has much better sensitivity and alignment quality than PSI-BLAST.

12. Karplus K: **SAM-T08 HMM-based protein structure prediction**. *Nucleic Acids Res* 2009, **37**:W492-W497.

13. Söding J: **Protein homology detection by HMM–HMM comparison**. *Bioinformatics* 2005, **21**:951-960.

14. Madera M: **Profile Comparer: a program for scoring and aligning profile hidden Markov models**. *Bioinformatics* 2008, **24**:2630.

15. Rost B, Schneider R, Sander C: **Protein fold recognition by prediction-based threading**. *J Mol Biol* 1997, **270**:471-480.

16. Karchin R, Cline M, Mandel-Gutfreund Y, Karplus K: **Hidden Markov models that use predicted local structure for fold recognition: alphabets of backbone geometry**. *Proteins* 2003, **51**:504-514.

17. Zhang Y: **I-TASSER server for protein 3D structure prediction**. *BMC Bioinformatics* 2008, **9**:40.

18. Liu S, Zhang C, Liang S, Zhou Y: **Fold recognition by concurrent use of solvent accessibility and residue depth**. *Proteins* 2007, **68**:636-645.

19. Wu S, Zhang Y: **MUSTER: improving protein sequence profile-profile alignments by using multiple sources of structure information**. *Proteins* 2008, **72**:547-556.

20. •• Peng J, Xu J: **Boosting protein threading accuracy**. *Lect Notes Comput Sci* 2009, **5541**:31-45.
The authors train a nonlinear score function on various per-residue sequence features using a conditional random field. They demonstrate considerable improvements in alignment quality in particular for very remote homologies.

21. Teichert F, Minning J, Bastolla U, Porto M: **High quality protein sequence alignment by combining structural profile prediction and profile alignment using SABER-TOOTH**. *BMC Bioinformatics* 2010, **11**:251-1251.

22. Rice DW, Eisenberg D: **A 3D-1D substitution matrix for protein fold recognition that includes predicted secondary structure of the sequence**. *J Mol Biol* 1997, **267**:1026-1038.

23. Przybylski D, Rost B: **Improving fold recognition without folds**. *J Mol Biol* 2004, **341**:255-269.

24. •• Biegert A, Söding J: **Sequence context-specific profiles for homology searching**. *Proc Natl Acad Sci U S A* 2009, **106**:3770-3775.
Substitution matrices use only the residue itself to predict likely mutations. The study shows that by including all 13 positions around each residue the probabilities of mutations are predicted much better.

25. Abeln S, Teubner C, Deane CM: **Using phylogeny to improve genome-wide distant homology recognition**. *PLoS Comput Biol* 2007, **3**:e3.

26. Fornes O, Aragues R, Espadaler J, Marti-Renom MA, Sali A, Oliva B: **ModLink+: improving fold recognition by using protein-protein interactions**. *Bioinformatics* 2009, **25**:1506-1512.

27. Fokkens L, Botelho SM, Boekhorst J, Snel B: **Enrichment of homologs in insignificant BLAST hits by co-complex network alignment**. *BMC Bioinformatics* 2010, **11**:86-186.

28. Boekhorst J, Snel B: **Identification of homologs in insignificant blast hits by exploiting extrinsic gene properties**. *BMC Bioinformatics* 2007, **8**:356-1356.

29. Kummerfeld SK, Teichmann SA: **Protein domain organisation: adding order**. *BMC Bioinformatics* 2009, **10**:39.

30. Mittelman D, Sadreyev R, Grishin NV: **Probabilistic scoring measures for profile–profile comparison yields more accurate short seed alignments**. *Bioinformatics* 2003, **19**:1531-1539.

31. Dlakic M: **HHsvm: fast and accurate classification of profile-profile matches identified by HHsearch**. *Bioinformatics* 2009, **25**:3071-3076.

32. Wang Y, Sadreyev RI, Grishin NV: **PROCAIN: protein profile comparison with assisting information**. *Nucleic Acids Res* 2009, **37**:3522-3530.

33. Laganeckas M., Margelevicius M., Venclovas C. Identification of new homologs of PD-(D/E)XK nucleases by support vector machines trained on data derived from profile-profile alignments. Nucleic Acids Res (2010); doi:10.1093/nar/gkq958.

34. Altschul SF, Gertz EM, Agarwala R, Schäffer AA, Yu YK: **PSI-BLAST pseudocounts and the minimum description length principle**. *Nucleic Acids Res* 2009, **37**:815-824.

35. Altschul SF, Wootton JC, Zaslavsky E, Yu YK: **The construction and use of log-odds substitution scores for multiple sequence alignment**. *PLoS Comput Biol* 2010, **6**:e1000852.
In hierarchical multiple sequence alignment and profile-profile comparison, one needs to compare two columns of amino acid counts or frequencies with each other. The article proposes a very natural log-odds score: the logarithm of the probability to generate the two count distributions from a single ancestral distribution divided by the probability to generate the two count distributions independently from two ancestral distributions.

36. Do C, Gross S, Batzoglou S: **CONTRAlign: discriminative training for protein sequence alignment**. *Lect Notes Comput Sci* 2006:160-174.
This article introduces the idea to combine the score from different per-residue features and to optimize the parameters by maximizing the likelihood on a set of training alignments.

37. Kim E, Wheeler T, Kececioglu J: **Learning models for aligning protein sequences with predicted secondary structure**. *J Comput Biol* 2010, **17**:561-580.

38. Xia X, Zhang S, Su Y, Sun Z: **MICAlign: a sequence-to-structure alignment tool integrating multiple sources of information in conditional random fields**. *Bioinformatics* 2009, **25**:1433-1434.

39. Peng J, Xu J: **Low-homology protein threading**. *Bioinformatics* 2010, **26**:i294-i300.

40. Frith MC, Park Y, Sheetlin SL, Spouge JL: **The whole alignment and nothing but the alignment: the problem of spurious alignment flanks**. *Nucleic Acids Res* 2008, **36**:5863-5871.

41. Lee MM, Chan MK, Bundschuh R: **Simple is beautiful: a straightforward approach to improve the delineation of true and false positives in PSI-BLAST earches**. *Bioinformatics* 2008, **24**:1339-1343.

42. Gonzalez MW, Pearson WR: **Homologous over-extension: a challenge for iterative similarity searches**. *Nucleic Acids Res* 2010, **38**:2177-2189.

43. Holmes I, Durbin R: **Dynamic programming alignment accuracy**. *J Comput Biol* 1998, **5**:493-504.

44. Biegert A, Söding J: **De novo identification of highly diverged protein repeats by probabilistic consistency**. *Bioinformatics* 2008, **24**:807-814.

45. Farrar M: **Striped Smith-Waterman speeds database searches six times over other SIMD implementations**. *Bioinformatics* 2007, **23**:156-161.
A very efficient implementation of Smith-Waterman alignment is described that uses the SSE2 instruction set of Intel and AMD CPUs. The article is the basis for drastic speed improvements in HMM-to-sequence and HMM–HMM alignment realized in HMMER3 [11••] and HHblits (M Remmert and J Söding, unpublished).

46. Johnson LS, Eddy SR, Portugaly E: **Hidden Markov model speed heuristic and iterative HMM search procedure**. *BMC Bioinformatics* 2010, **11**:431-1431.

47. Eddy SR: **A probabilistic model of local sequence alignment that simplifies statistical significance estimation**. *PLoS Comput Biol* 2008, **4**:e1000069.

48. Rangwala H, Karypis G: **Profile-based direct kernels for remote homology detection and fold recognition**. *Bioinformatics* 2005, **21**:4239-4247.

49. Bateman A, Finn RD: **SCOOP: a simple method for identification of novel protein superfamily relationships**. *Bioinformatics* 2007, **23**:809-814.

50. Jung I, Kim D: **SIMPRO: simple protein homology detection method by using indirect signals**. *Bioinformatics* 2009, **25**:729-735.

51. Weston J, Elisseeff A, Zhou D, Leslie CS, Noble WS: **Protein ranking: from local to global structure in the protein similarity network**. *Proc Natl Acad Sci U S A* 2004, **101**:6559-6563.

52. Melvin I, Weston J, Leslie C, Noble WS: **RANKPROP: a web server for protein remote homology detection**. *Bioinformatics* 2009, **25**:121-122.

53. Melvin I, Weston J, Noble WS, Leslie C: **Detecting remote evolutionary relationships among proteins by large-scale semantic embedding**. *PLoS Comput Biol* 2011, **7**:e1001047.
The authors describe a network method for remote homology detection that uses all pairwise similarities in the sequence database. The empirical feature vectors of database proteins are embedded in a lower-dimensional virtual space while conserving the pairwise distances as much as possible. For searching, the query position in this virtual space is calculated from its similarities to all database proteins.

54. Heger A, Mallick S, Wilton C, Holm L: **The global trace graph, a novel paradigm for searching protein sequence databases**. *Bioinformatics* 2007, **23**:2361-2367.

55. Qi Y, Sadreyev RI, Wang Y, Kim BH, Grishin NV: **A comprehensive system for evaluation of remote sequence similarity detection**. *BMC Bioinformatics* 2007, **8**:314-1314.

56. Zhang Y, Skolnick J: **Scoring function for automated assessment of protein structure template quality**. *Proteins* 2004, **57**:702-710.

57. Zemla A, Venclovas , Moult J, Fidelis K: **Processing and evaluation of predictions in CASP4**. *Proteins Struct Funct Bioinformatics* 2001, **45**:13-21.

58. Cuff AL, Sillitoe I, Lewis T, Clegg AB, Rentzsch R, Furnham N, Pellegrini-Calace M, Jones D, Thornton J, Orengo CA: **Extending CATH: increasing coverage of the protein structure universe and linking structure with function**. *Nucleic Acids Res* 2011, **39**:D420-D426.

59. Frank K, Gruber M, Sippl MJ: **COPS Benchmark: interactive analysis of database search methods**. *Bioinformatics* 2010, **26**:574-575.

60. Andreeva A, Howorth D, Chandonia JM, Brenner SE, Hubbard TJ, Chothia C, Murzin AG: **Data growth and its impact on the scop database: new developments**. *Nucleic Acids Res* 2008, **36**:D419-D425.

61. Chubb D, Jefferys BR, Sternberg MJ, Kelley LA: **Sequencing delivers diminishing returns for homology detection: implications for mapping the protein universe**. *Bioinformatics* 2010, **26**:2664-2671.

62. Sadreyev RI, Tang M, Kim BH, Grishin NV: **Compass server for homology detection: improved statistical accuracy, speed and functionality**. *Nucleic Acids Res* 2009, **37**:90-94.
(Web Server issue)

63. Margelevicius M, Venclovas C: **Detection of distant evolutionary relationships between protein families using theory of sequence profile-profile comparison**. *BMC Bioinformatics* 2010, **11**:89.