HHblits: Lightning-fast iterative protein sequence searching by HMM-HMM alignment

Michael Remmert, Andreas Biegert, Andreas Hauser & Johannes Söding

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

Abstract

The success of sequence-based protein function and structure prediction depends crucially on the sensitivity of sequence searches and the accuracy of the resulting multiple sequence alignments. HHblits (`http://hhblits.genzentrum.lmu.de`) is a general-purpose sequence search tool that represents both query and database sequences by profile hidden Markov models (HMMs). It is faster than PSI-BLAST thanks to its discretized-profile prefilter, has 50%–100% higher sensitivity, and generates more accurate alignments.

Building protein multiple sequence alignments (MSAs) by iterative sequence searches is of fundamental importance in computational biology, since MSAs are a key intermediate step in the sequence-based prediction of evolutionarily conserved properties, such as secondary or tertiary structure, disorder, catalytic sites, post-translational modifications, short linear motifs, or interaction interfaces. Sequence profiles and profile HMMs are condensed representations of MSAs that contain for each sequence position the probabilities to observe each of the 20 amino acids in related proteins. PSI-BLAST[1], the most popular iterative search tool, progressively refines a query sequence profile by adding significant sequence matches to the profile for the next search iteration. SAM2K[2] and HMMER3[3] work similarly but employ profile HMMs for better sensitivity and alignment quality. Fast heuristic prefilters in PSI-BLAST and HMMER3 speed up the iterative searches.

Profile-profile and HMM-HMM alignment methods are the most sensitive class of sequence search methods and are the methods of choice for identifying and aligning templates for 3D homology modeling[4]. Our HMM-HMM alignment method HHsearch[5] is used by many of the best protein structure prediction servers, among them HHpred[6], the best-scoring server in template-based structure prediction during the last CASP9 blind structure prediction benchmark (`http://toolkit.genzentrum.lmu.de/CASP9`). However, profile-profile alignment methods have been much too slow for iteratively searching through large, representative sequence databases such as UniProt or the nonredundant (nr) database from NCBI. Here, we present HHblits (H̲MM-H̲MM-b̲ased l̲ightning-fast i̲terative sequence s̲earch), which extends HHsearch to enable fast, iterative sequence searches. Its profile-profile alignment prefilter reduces the number of full HMM-HMM alignments to a few thousand, making it faster than PSI-BLAST yet as sensitive as HHsearch (**Supplementary Fig. 1**).

In order to perform HMM-HMM comparisons, the sequence database is clustered into sequence sets alignable over nearly their full length and MSAs and HMMs are generated for each. We devised a very fast method (kClust: Hauser, Mayer, and Söding, to be published) for clustering large sequence databases such as UniProt down to 20-30% maximum pairwise sequence identity ~1000 times faster than BLAST. We use kClust to regularly update the clustered UniProt and nr databases. The clustered UniProt (07/2011) contains 15M sequences in 2.6M HMMs with 5.5 sequences on average per cluster. The requirement of full-length alignability ($>$ 80% of longest sequence) ensures that clusters are mainly composed of functionally similar, orthologous sequences[7].

HHblits first converts the query sequence (or MSA) to an HMM. To increase sensitivity, we add sequence context-specific pseudocounts to the observed amino acid counts[8]. HHblits then searches the HMM database and adds the sequences from HMMs below an E-value threshold to the query MSA, from which the HMM for the next search iteration is built (**Fig. 1a**). For speed and sensitivity, the prefilter is critical. The key idea is to effectively reduce profile-profile comparison to sequence-to-profile comparison by discretizing the vectors of 20 amino acid probabilities in each HMM column into an alphabet of 219 letters. Each letter represents a typical profile column, shown in **Supplementary Fig. 2**. The database HMMs are approximated by sequences over this extended alphabet (ignoring the HMMs' insertion and deletion probabilities). Prior to prefiltering, the score of each query HMM column with each of the 219 letters is calculated, resulting in a 219-row extended sequence profile. The prefiltering consists of two steps. A very fast gap-less local alignment between the extended query profile and the extended database sequences is followed by a gapped local alignment, for which we modified code by Michael Farrar[9]. Each step lets 1%–5% of sequences pass. Both filters are implemented with SSE2 (Streaming SIMD Extensions 2) instructions, available on all modern Intel and AMD CPUs, which process 16 single-byte operations in parallel per clock cycle[9]. The database HMMs whose extended sequences passed the prefilter are Viterbi-aligned to the query HMM, and E-values and probabilities are calculated (Online Methods). Significant matches are realigned using the maximum accuracy algorithm generalized to local HMM-HMM alignment[10].

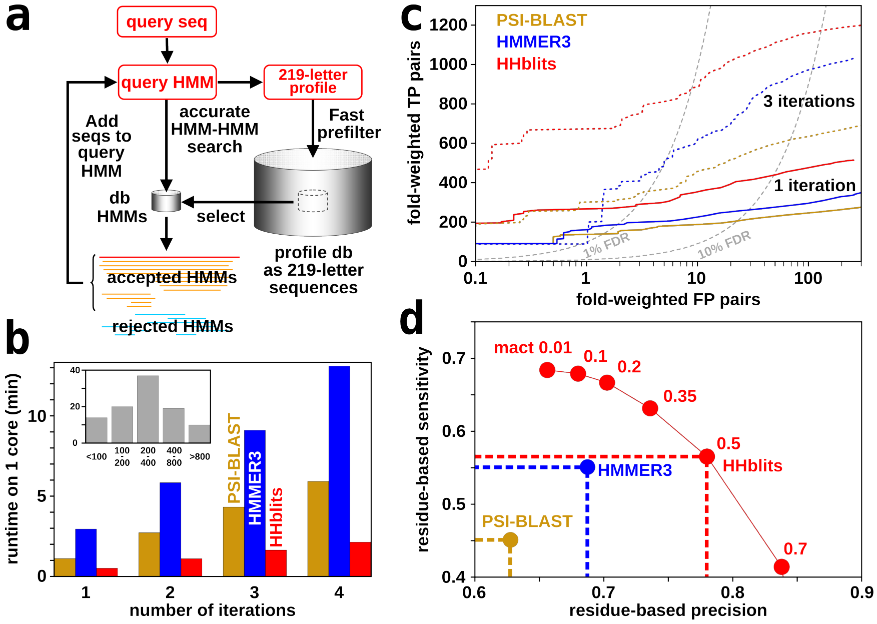

Figure 1: Work flow and benchmark comparison. (**a**) HHblits uses iterative HMM-HMM alignment to search for homologous sequences in large sequence databases such as UniProt. The HHblits database is a clustered version in which each set of full-length alignable sequences is represented by an HMM. Sequences from matched HMMs with significant E-value are added to the query MSA, from which a new HMM is calculated for the next search iteration. A prefilter reduces the number of full HMM-HMM alignments ~2500-fold. (**b**) Median run times for searches through the UniProt database (inset: sequence length distribution). (**c**) True positive pairs (TPs, same SCOP fold) versus false positives (FPs, different fold) for 1 and 3 search iterations in an all-against-all comparison. FDR: false discovery rate. (**d**) Mean fraction of correctly aligned residue pairs out of all structurally alignable pairs (sensitivity) versus the fraction of correctly aligned pairs out of aligned pairs (precision) on a set of 4128 pairs of SCOP sequences. The mact parameter in HHblits controls alignment greediness.

An HHblits search through UniProt takes 31s (1m13s), measured as median (average) over 100 randomly selected query sequences on a single Xeon 2.9 GHz core (**Fig. 1b**). PSI-BLAST needs 1m7s (1m26s) and HMMER3 2m57s (5m8s). Further iterations take roughly the same time, hence HHblits is twice (15%) faster than PSI-BLAST and 4-5 (6) times faster than HMMER3. All three tools scale similarly well on 2 to 8 cores (**Supplementary Fig. 3, Supplementary Data 1**).

In **Fig. 1c** we compare sensitivities to detect homologs, i.e., to rank homologous pairs (TP) above unrelated pairs (FP) of proteins. We performed an all-against-all comparison of 5287 representative domain sequences from SCOP[11]. HHblits parameters were optimized on a separate set of SCOP folds (Online Methods). After one iteration, HHblits detects 112% (68%) more TPs than PSI-BLAST (HMMER3) at 10% false discovery rate, after three iterations the improvement is 90% (20%). Similar values are obtained in a ROC5 analysis (**Supplementary Fig. 4d**). On multi-domain proteins, multiple PSI-BLAST iterations often lead to corrupted alignments through homologous over-extension[12], whereas HHblits is robust against this effect (**Supplementary Fig. 5**).

To assess the quality of pairwise alignments (**Fig. 1d**), we chose 4128 query-template pairs by randomly selecting from each SCOP superfamily up to 10 pairs of domains with $< 30\%$ sequence identity and TM-align score > 0.6 (**Supplementary Data 2**). With each method we built MSAs for the queries using two search iterations through UniProt and aligned the resulting query MSAs with their corresponding templates. (For HHblits, we selected the template HMMs from the clustered UniProt that contained the SCOP template sequence.) We determined correctly aligned residues by comparison with structural alignments from TM-align. For default parameters (mact 0.5), HHblits alignments have 12% higher sensitivity and 15% higher precision per residue than those of PSI-BLAST (2% and 10% for HMMER3, respectively). The higher precision of HHblits alignments explains its robustness against homologous over-extension[12].

As a further measure of MSA quality, we compared the accuracy of secondary structure prediction by PSIPRED[13] using the PSIPRED procedure to generate sequence profiles (three iterations of PSI-BLAST on a filtered database) with the accuracy of PSIPRED run on profiles built from MSAs generated by HHblits. Even though PSIPRED was trained with its own MSAs, HHblits MSAs improved the Q3 score on proteins from PDBselect 2007 from 80.4% to 81.3% and the SOV (segment overlap) score from 77.5% to 78.6% (**Supplementary Table 1**). These results, obtained without training a large parameter set, are among the best achieved so far[14].

To demonstrate the utility of HHblits, we sought to predict structures for Pfam families[15] for which no homologous template is known, and neither for any family from the same Pfam clan (**Fig. 2a**). We jump-started with the Pfam seed alignment two HHblits iterations through UniProt and then searched HHpred's PDB70 database (`ftp://toolkit.genzentrum.lmu.de/HHsearch/databases`). HHblits assigns templates to 620 families with E-value $< 10^{-3}$, only 226 of which can be as-

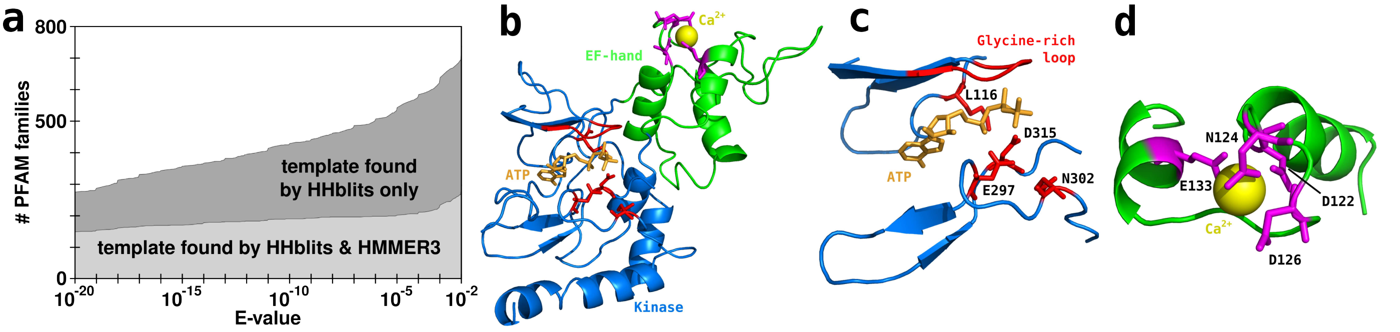

Figure 2: Structure predictions for Pfam families and modeling of human Pip49/FAM69B. (**a**) Pfam families to which only HHblits and both HHblits and HMMER3 assign a structural template below a given E-value. (**b**) Homology model of hPip49 kinase domain (blue) with inserted EF hand (green). (**c**) Catalytic center showing conserved residues (red) for protein kinase activity. (**d**) EF hand insertion with conserved residues (magenta) for predicted Ca^{2+}-dependent activation.

signed by HMMER3 (41 are found only by HMMER3). For a complete list of HHblits-detected templates for Pfam, see **Supplementary Table 2**.

We illustrate the practical relevance of these predictions at the example of Pip49_C, the "Pancreatitis induced protein 49 C-terminal", a domain of unknown structure and function with an N-terminal transmembrane helix. The 100 best HHblits matches in PDB70 are all with protein kinases (best E-value 2×10^{-20}), even though the Pfam MSA is missing the N-terminal part. An HHblits search started with full-length human Pip49/FAM69B (2 iterations through UniProt, 1 through PDB70) detected many protein kinase domains, and, interestingly, a tandem Ca^{2+}-binding EF hand (E-value 0.09) inserted after the small N-terminal β sheet of the kinase domain. Although many protein kinases contain EF hands downstream of their kinase domains [15], Pip49 is the first case where an EF hand is inserted within the kinase domain. The kinase domain is framed by two short domains with four or more highly conserved cysteines each that are likely to form disulfide bonds. Based on our homology models (**Fig. 2b-d, Supplementary Data 3**) and the conservation of critical residues, we predict that Pip49/FAM69B and FAM69A are ER membrane bound protein kinases in the ER lumen that are activated by Ca^{2+} through structural rearrangement of their EF hand. Residue conservation suggests that metazoan FAM69C will also possess protein kinase activity.

In conclusion, HHblits is a robust, general-purpose protein sequence search tool based on HMM-HMM alignment that is faster than PSI-BLAST, gives more reliable E-value estimates (**Supplementary Fig. 6**), is considerably more sensitive, and produces alignments of much better quality.

ACKNOWLEDGEMENTS

We acknowledge financial support by the Deutsche Forschungsgemeinschaft and by the LMU through the Excellence Initiative of the BMBF.

AUTHOR CONTRIBUTIONS

M.R. performed research, J.S. initiated and guided research, A.B. generated the profile-column alphabet, A.H. contributed code for fast file access, M.R. and J.S. wrote the manuscript.

COMPETING INTERESTS

The authors declare that they have no competing financial interests.

CORRESPONDENCE

Correspondence should be addressed to J.S. (soeding@genzentrum.lmu.de).

ONLINE METHODS

Discretized profile-column alphabet. We discretize profile columns into an alphabet of 219 states (number of printable ASCII characters), where each letter represents a typical profile column. This allows us to approximate any sequence profile by a sequence over this 219-letter, extended alphabet. To compare two profiles, we first calculate the score of each query profile column with each of the 219 letters, using the formula $\log_2 \sum_{a=1}^{20} q_i(a)p_k(a)/f(a)$, where $q_i(a)$ denotes the query profile at position i, $p_k(a)$ is the profile column represented by letter $k \in \{1,\ldots,219\}$, and $f(a)$ is the background frequency of residue a. We thus obtain a 219-row extended sequence profile, which can be aligned to extended sequences representing the other profile using fast, standard dynamic programming techniques. We generate the 219-letter alphabet using the same method employed for learning an optimal set of sequence context profiles[8], but we set the window size from 13 to 1 residues. We also set the window weights w_j to 100 in order to obtain a hard clustering. We initialized the 219 states randomly and maximized the likelihood that the 10M training sequence profile columns were generated by the 219 profile columns. The best of several trials was used. The 10M profile columns were randomly sampled from MSAs in our clustered nr database.

Prefiltering. In the three prefilter steps, the extended query sequence profile is aligned to the extended database sequences. The first step calculates the score of the largest ungapped alignment. To pass this filter, the score has to be larger than $2.5 + \log_2(L_Q L_T)$ bits, where L_Q and L_T are the lengths of the query profile and database sequence. The log term is a standard length correction. The second step calculates a Smith-Waterman alignment with affine gap penalties (gap open: 5 bits, gap extend: 1 bit). From the bit score S, an approximate E-value is calculated: $E = N_{db} L_Q L_T 2^{-S}$, where N_{db} is the number of sequences/HMMs in the database, and sequences pass if $E < E_{pre} = 1000$. Each filter step leads to a ~50-fold reduction of database sequences.

Both filters are implemented with SSE2 (Streaming SIMD extension 2) instructions that process 16 single bytes in parallel on 128-bit SIMD units present on each CPU core. Each byte holds the score in units of 1/4 bits plus an offset of 50, allowing to represent a score range between -12.5 and +51.5 bits. The algorithms are programmed such that the scores will saturate at 255 upon overflow. Since any score larger than 51 bits will pass the filter anyway, this range is sufficient for prefiltering. The first step processes 4 to 5 cells of the dynamic programming matrix per CPU clock cycle, the second step ∼1.3 cells per clock cycle. The clustered UniProt database (07/2011) contains 2.6M sequences of average length 320, hence the first prefilter search with a query profile of length 300 through UniProt takes about $300 \times 320 \times 2.6 \times 10^6/(4.5 \times 2.9\text{GHz}) = 18s$, which is about 25% of the average time needed for the entire HHblits search.

Sequences that pass the first two steps are aligned in a third step using SSE2 instructions to determine the region likely to contain the true alignments. For back-tracing the alignments, we need to prevent the score from saturating. Therefore, each score is held in 2 bytes in this step (again in units of 1/4 bits), yielding a score range of -12.5 to 16371.5 bits. Up to 10 suboptimal alignments are extracted by setting all cells at a distance of $<$ 150 residues from the previously extracted alignments to 0 until the prefilter E-value rises above E_{pre}.

Viterbi alignment and E-value calculation. To speed up the time-consuming HMM-HMM alignment steps, all cells with a distance of $>$ 200 to all alignment identified in the previous step are flagged as inactivated. An HMM-HMM alignment is performed on the active cells using the Viterbi algorithm of HHsearch. From the Viterbi score S, a P-value is calculated using an extreme value distribution (EVD): $P = 1 - \exp(-\exp[-\lambda(S - \mu)])$. The EVD parameters μ and λ are estimated from the four features $(L_Q, L_T, N_Q^{eff}, N_T^{eff})$ using two standard, two-layer neural networks with four hidden nodes each. Here, N_Q^{eff} and N_T^{eff} are the numbers of effective sequences in the query and template HMMs, respectively, defined in[5]. The Viterbi E-value is calculated from the P-value using $E = N_{db} P \times (E_{pre}/N_{db})^\alpha$, where $\alpha = 0.4 + 0.02 \times (N_T^{eff} - 1) \times (1 - 0.1 \times (N_Q^{eff} - 1))$. The term $(E_{pre}/N_{db})^\alpha$ is an empirical correction for the correlation between the prefiltering and Viterbi scores (α=0: perfect correlation, α=1: no correlation). The three coefficients were optimized to yield accurate E-values (**Supplementary Fig. 6**).

Further speed-ups. Viterbi alignments are performed in the order of decreasing prefilter E-value. We stop the time-consuming HMM-HMM comparisons in cases when very few homologs are likely to have been observed among the last 200 HMM-HMM alignments. A coarse estimate for the probability for a match to be a true homolog is $1/(1 + E)$ for Viterbi E-value E. We average $1/(1 + E)$ over the last 200 processed Viterbi alignments and skip all further database HMMs when this average drops below 0.01.

Maximum accuracy alignment. Whereas the Viterbi algorithm calculates the alignment with the best score, the maximum accuracy alignment, proposed in[16], yields the global alignment with the maximum possible accuracy defined by the sum of probabilities for each residue pair to be correctly aligned: $\sum_{(i,j)\in\text{alignment}} P(i \text{ aligned to } j) \rightarrow \text{max}$. We extended this algorithm to the case of local HMM-HMM comparison[10], which produces the *local* alignment that maximizes the sum of probabilities for each residue pair to be correctly aligned minus a penalty (mact): $\sum_{(i,j)\in\text{alignment}} [P(i \text{ aligned to } j) - \text{mact}] \rightarrow \text{max}$. With the mact parameter, the alignment greediness can be controlled, from nearly global, long, greedy alignments (mact near 0) to very precise and short (mact near 1).

Adding sequences from significant matches to query HMM. Sequences from all HMMs below the Viterbi E-value inclusion threshold (default 10^{-3}) are read from the alignment files of the clustered database and are aligned to the query MSA according to the HMM-HMM maximum accuracy alignment. The query HMM is calculated from the query MSA.

Parameter optimization. We optimized parameters (filter thresholds, gap costs, amino acid and transition pseudocount strengths, E-value inclusion threshold) on a optimization set which has no member from the same fold as sequences in the test set (see next paragraph). We varied parameters in discrete steps one after the other, performed an all-against-all search on the optimization set and tried to maximize the mean ROC5 value (see next paragraph). For prefilter settings, we chose the best trade-off between efficiency and sensitivity.

Sensitivity benchmarks. We filtered the sequences from SCOP 1.73 [17] to a maximum pairwise sequence identity of 20%. We assigned every fifth fold to the optimization set (1329 sequences in 215 folds) and the others to the test set (5287 sequences in 862 folds, **Supplementary Data 4**). SCOP is a hierarchically ordered database of protein domain sequences with known structure. We consider domains from the same fold as TP, domains from different folds as FP, i.e., non-homologous. Exceptions are members of 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), which are probably related and treated as "unknown". To prevent a few large folds from dominating the benchmark, we weight each hit with one over the number of members in the query SCOP fold ("fold-weighted TPs and FPs"). All but the last search iteration are performed against the UniProt database. The final iteration of PSI-BLAST and HMMER searches are performed against all UniProt and

SCOP sequences. For HHblits, the final iteration is against the clustered UniProt. Each SCOP sequence from the test set was mapped to its UniProt cluster containing the test sequence or added as singleton cluster to UniProt if no matching cluster was found. All pairs of domains were ranked by E-value for each of the tools, and TPs versus FPs below a given E-value were plotted. The ROC5 plots in **Supplementary Figs. 4d and 5b** assess how well a method ranks the matched proteins within each search. They show the fraction of queries with ROC5 scores above the threshold on the x-axis. The ROC5 score is the area under the TP-versus-FP ROC (receiver operating characteristic) curve up to the 5'th FP, divided by the area under the optimal ROC curve.

Sensitivity benchmark for multi-domain proteins. Since multi-domain protein sequences present particular challenges such as homologous over-extension [12] to iterative sequence search methods, we tested the tools on a benchmark set of multi-domain proteins. For each of the 5287 sequences in our test set, we searched for a sequence in the non-redundant (nr) database that has a BLAST match to the SCOP sequence with an E-value $< 10^{-40}$, a sequence coverage $> 95\%$, a sequence identity $> 60\%$ and whose full-length sequence contains at least 100 additional residues. This procedure lead to 2343 multi-domain proteins. For all extracted multi-domain proteins we proceeded as described in the previous paragraph (2 iterations through UniProt, 1 iteration through UniProt/SCOP). We counted TPs and FPs only if the alignment covers at least 50 residues of the SCOP domain in the nr query sequence.

Improving PSIPRED secondary structure prediction. For the secondary structure prediction by PSIPRED we used PDBselect 2007, which contains 3649 sequences ranging from 30 to 1040 amino acids length. We built MSAs for each sequence using 2 and 3 iterations of PSI-BLAST and 1, 2, and 3 iterations of HHblits. HHblits alignments with diversity 7 were generated by applying hhfilter from the HHblits package with option `-neff 7`. For all MSAs we performed PSIPRED with the default parameters and calculated the Q3 and SOV score based on the known DSSP sequences (mapping E and B to strand, H, G, and I to helix, S, T, and C to coil states).

Fold prediction for Pfam. For nearly half of all Pfam families in version 24.0 (5716 out of 11913), no structure is known and neither for any of the remotely related families in their Pfam clan. We generated MSAs for the 5716 Pfam families by using their seed alignments as input and performing two iterations with HHblits through the UniProt database. The PDB70 database of HHpred is searched with the resulting MSAs. For HMMER3, we scan the PDB70 sequence database with the HMMER3 models provided by Pfam.

Pip49/FAM69B modeling. We built an MSA for human Pip49/FAM69B (UniProt-ID: Q5VUD6) by running two iterations of HHblits through the clustered UniProt database and adding the secondary structure prediction from PSIPRED to this MSA (**Supplementary Data 5**). To identify structural homologs, the PDB database was scanned by HHblits with this MSA with a mact-value of 0.2. From the list of PDB matches we chose a protein kinase with bound ATP (PDB-ID: 1RDQ) and a Ca^{2+}-bound EF hand (PDB-ID: 3C1V) as templates and use the corresponding HHblits alignments to create a homology model with MODELLER[18] (**Supplementary Data 3**). We confirmed the presence of the EF hand insertion by building an MSA with two iterations of HHblits starting from the presumed inserted sequence and searched the PDB70. This yielded very significant matches with EF hands (best E-value 4×10^{-5}). The previously reported transmembrane helix from position 31 to 51 could be confirmed by HMMTOP, MEMSAT-SVM and PHOBIUS.

References

[1] Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).

[2] Karplus, K., Barrett, C. & R., H. Hidden Markov models for detecting remote protein homologies. *Bioinformatics* **14**, 846–856 (1998).

[3] Eddy, S. A new generation of homology search tools based on probabilistic inference. *Genome Inform.* **23**, 205–211 (2009).

[4] Söding, J. & Remmert, M. Protein sequence comparison and fold recognition: progress and good-practice benchmarking. *Curr. Opin. Struct. Biol.* **21**, 401–411 (2011).

[5] Söding, J. Protein homology detection by HMM-HMM comparison. *Bioinformatics* **21**, 951–960 (2005).

[6] Söding, J., Biegert, A. & Lupas, A. N. The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res.* **33**, W244–W248 (2005).

[7] Hegyi, H. & Gerstein, M. Annotation transfer for genomics: measuring functional divergence in multi-domain proteins. *Genome Res.* **11**, 1632–1640 (2001).

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

[9] Farrar, M. Striped Smith-Waterman speeds database searches six times over other SIMD implementations. *Bioinformatics* **23**, 156–161 (2007).

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

[11] Murzin, A. G. How far divergent evolution goes in proteins. *Curr. Opin. Struct. Biol.* **8**, 380–387 (1998).

[12] Gonzalez, M. W. & Pearson, W. R. Homologous over-extension: a challenge for iterative similarity searches. *Nucleic Acids Res.* **38**, 2177–2189 (2010).

[13] Jones, D. T. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* **292**, 195–202 (1999).

[14] Aydin, Z., Singh, A., Bilmes, J. & Noble, W. Learning sparse models for a dynamic bayesian network classifier of protein secondary structure. *BMC Bioinformatics* **12**, 154 (2011).

[15] Finn, R. D. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **38**, D211–222 (2010).

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

[17] Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536–540 (1995).

[18] Sali, A. & Blundell, T. L. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **234**, 779–815 (1993).

HHblits: Lightning-fast iterative protein sequence searching by HMM-HMM alignment

Supplementary Information

Michael Remmert, Andreas Biegert, Andreas Hauser and Johannes Söding

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

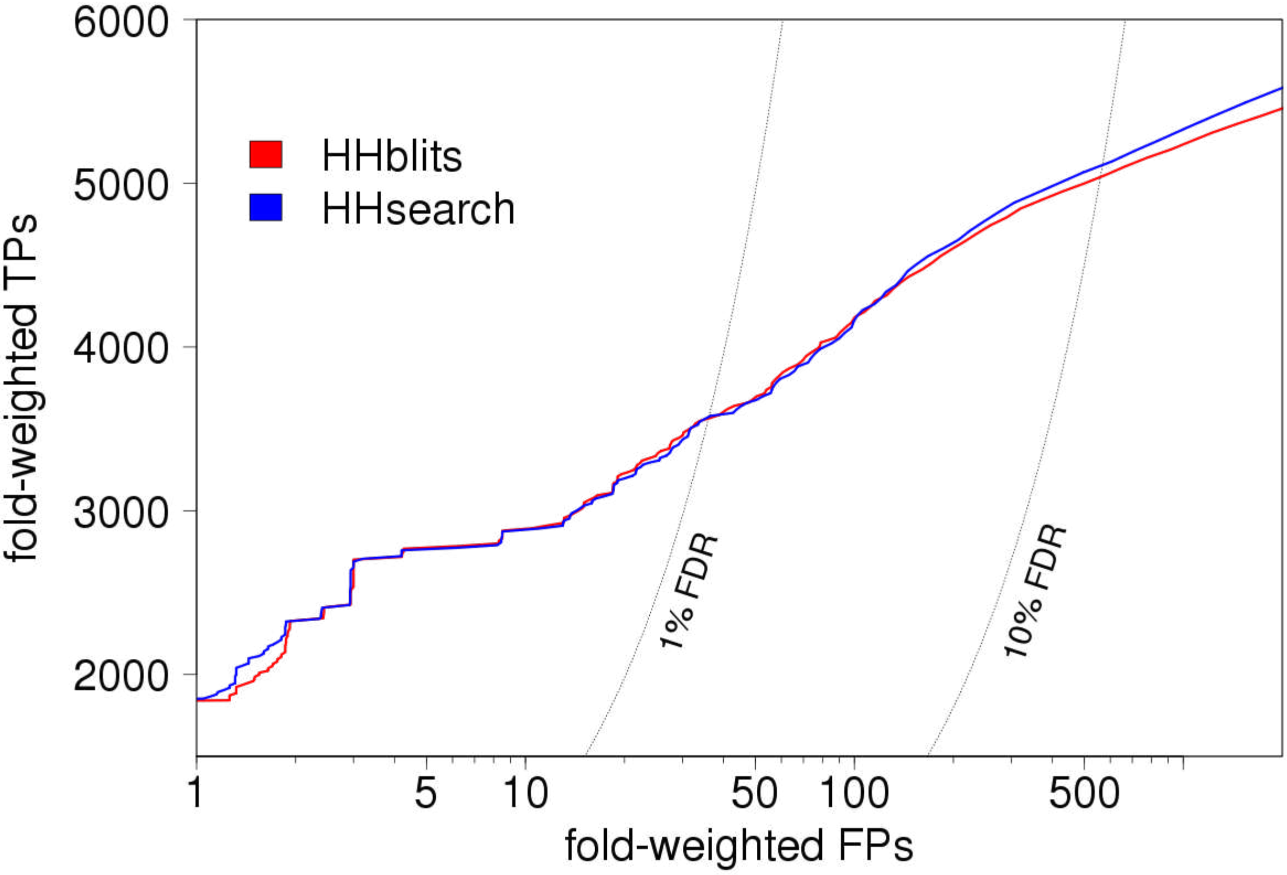

Figure S1: Sensitivity / selectivity comparison between HHblits and HHsearch on the SCOP test set for a single search. The prefiltering in HHblits leads only to a very slight performance decrease with respect to HHsearch, even though the run time of HHblits is decreased dramatically.

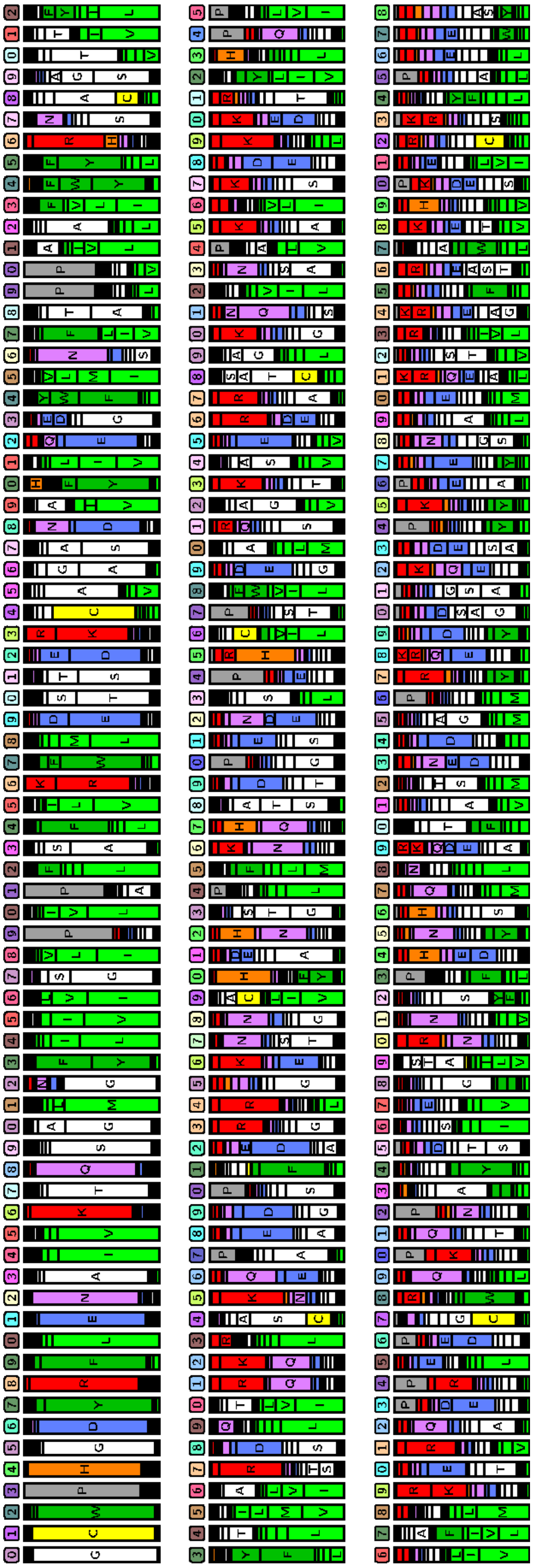

Figure S2: Histogram representation of the amino acid distributions in the 219 profile columns of the extended profile column alphabet. The column vectors are ordered by entropy, starting with the almost pure states and ending with near-random background distributions.

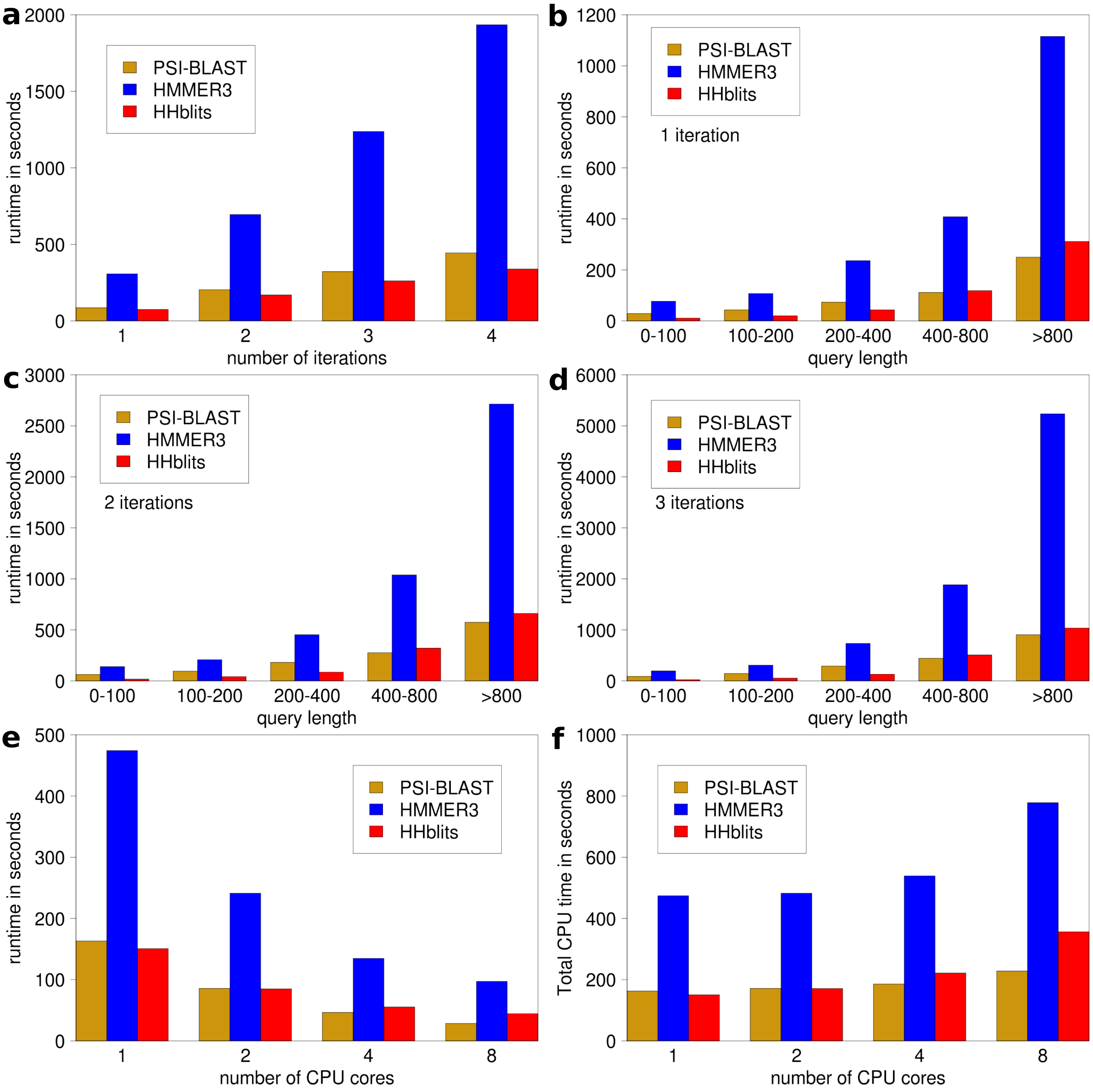

Figure S3: Average run times on 100 randomly selected sequences from the nr database, measured on an Intel Xeon X5570 at 2.93 GHz **(a)** Average run times for 1 to 4 iterations. **(b)-(d)** Run times for various bins of query length for (b) one, (c) two, and (d) three search iterations. HHblits has a very good run time for queries with a sequence length below 400 residues and clearly outperforms PSI-BLAST in this range of query lengths. In the range of 400 to 800 residues both methods have a similar run time and only for proteins with a length $>$ 800 residues the run time of HHblits is slightly worse to that of PSI-BLAST. HMMER3 scales in a similar way as HHblits with the query length, always by a factor 3 to 5 slower. **(e), (f)** Run time for two search iterations on 1 to 8 CPU cores. (e) shows the wall time, whereas (f) gives the total CPU time, equal to the wall time times the number of CPUs. All three tools scale well with an increasing number of CPUs.

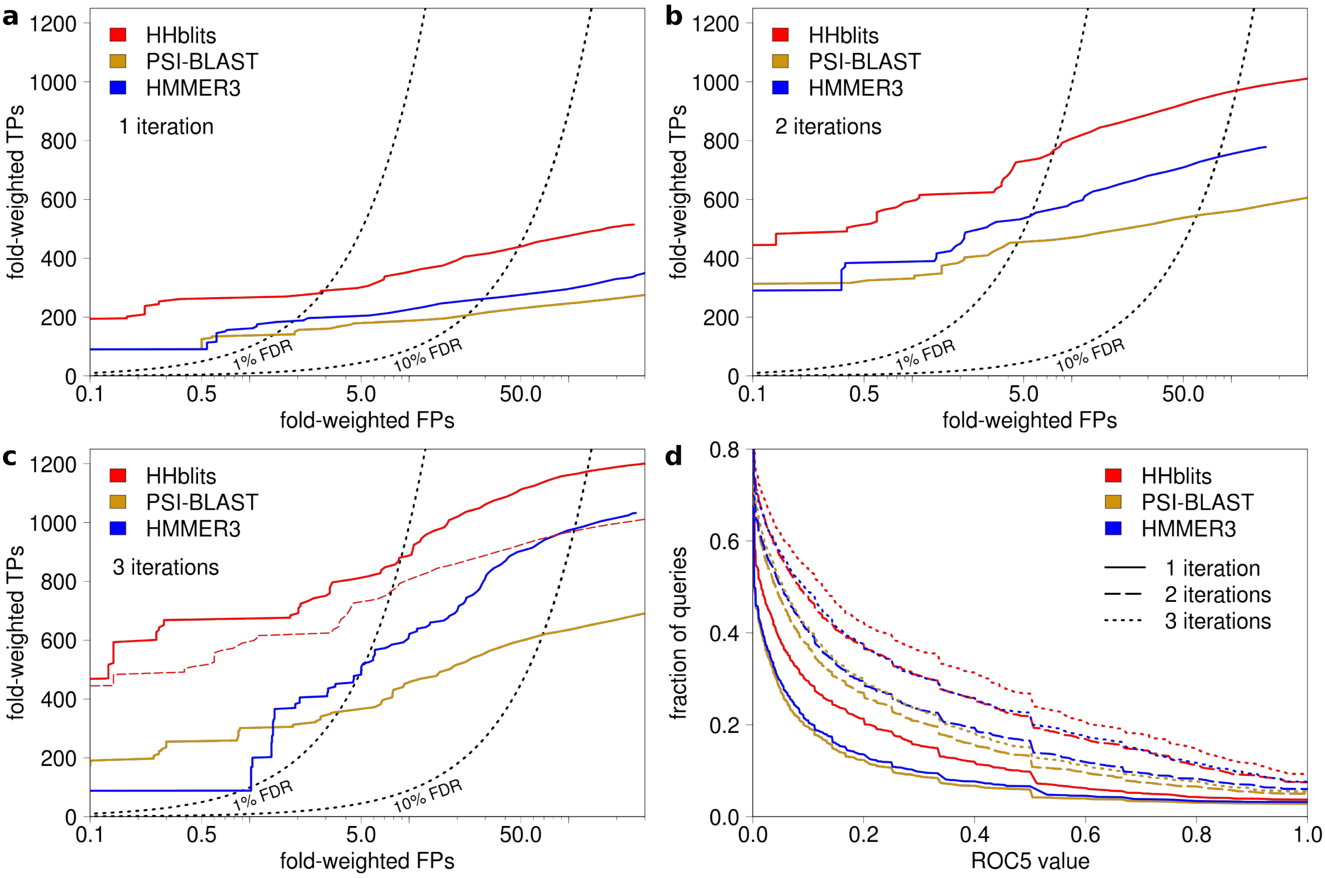

Figure S4: Sensitivity and selectivity of homology detection. **(a)-(c)** ROC (receiver operating characteristic) plots for (a) one, (b) two, and (C) three iterations on the test set (5287 sequences from the SCOP 1.73 database). All but the last search iteration are performed against the UniProt. The last search iteration is done through a combined database containing the UniProt and the SCOP sequences (See Online Methods). TPs are defined as pairs from the same SCOP folds, FPs as pairs from different folds, with the exception of Rossman folds and β propellers. At a false discovery rate (FDR) of 10% HHblits detects in the first iteration twice as many TPs as PSI-BLAST and 68% more than HMMER3. In (c), the light red curve (two iterations of HHblits) shows a clear improvement over three iterations of PSI-BLAST and HMMER3. **(d)** Fraction of queries with ROC5 value above the threshold on the x-axis. The ROC5 value is the area under the ROC curve up to the 5'th FP, normalized to yield a theoretical maximum of 1. The ROC5 plot is more robust to overfitting than the ROC curves in (a-c). (see Söding & Remmert, Curr. Opin. Struct. Biol. 2011).

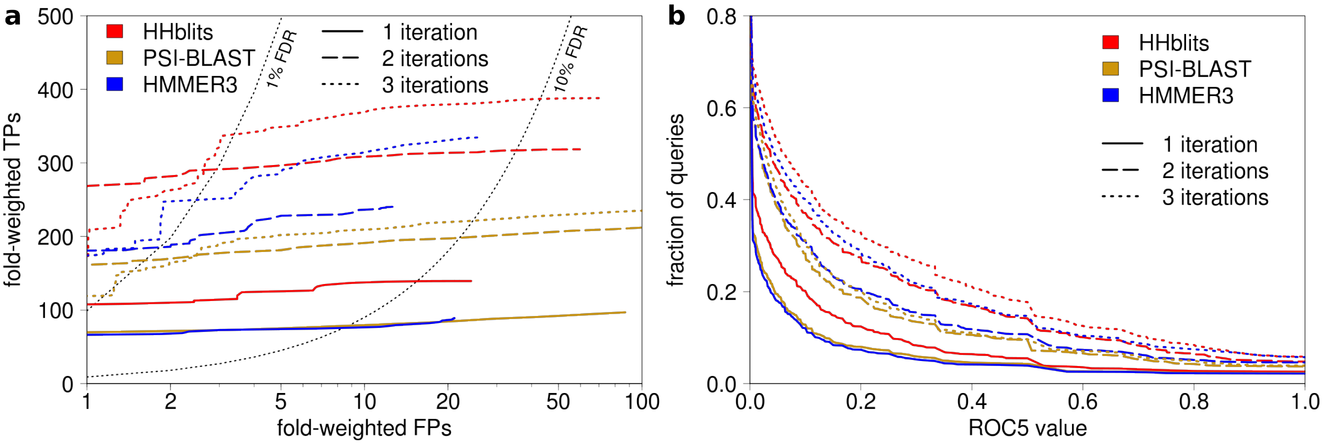

Figure S5: Sensitivity and selectivity of homology detection for multi-domain proteins. True positive pairs (TPs) and false positive pairs (FPs) are counted only if the alignment covers at least 50 residues of the SCOP domain in the query NR protein. **(a)** ROC (receiver operating characteristic) plot showing TPs versus FPs detected at the same E-value thresholds, for 1, 2 and 3 search iterations. After three iterations, HHblits has significantly fewer FPs at high confidence (false discovery rate FDR $< 1\%$) than PSI-BLAST and HMMER3. **(b)** Fraction of queries with ROC5 value above the threshold on the x-axis. The ROC5 value is the area under the ROC curve up to the 5'th FP, normalized to yield a theoretical maximum of 1.

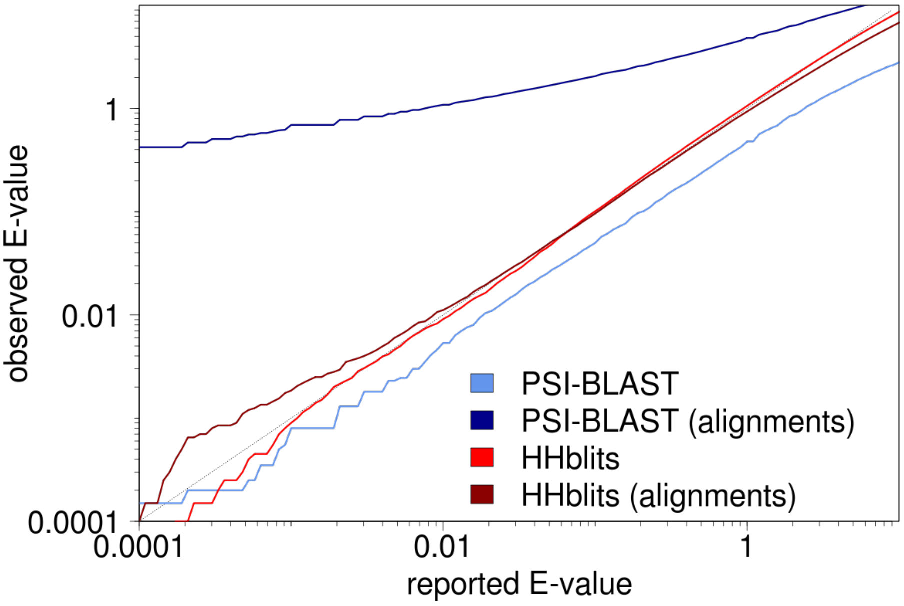

Figure S6: Accuracy of E-value estimation by HHblits and PSI-BLAST. We generated a version of UniProt (and the clustered UniProt) with randomly shuffled residues (MSA columns) and randomly selected 20 000 proteins from the nr database to search through the scrambled database. Any match is therefore a false positive. We counted the number of matches below a given reported E-value. Dividing this number through the number of total searches (20 000) yields the empirical, observed E-value. Reported and observed E-values should be similar. Both PSI-BLAST (light blue) and HHblits (red) report reliable E-values when started with a single sequence. When searches are jump-started with a MSAs (obtained from one iteration of HHblits), PSI-BLAST produces a great excess of false positive matches at E-values below 1 (dark blue).

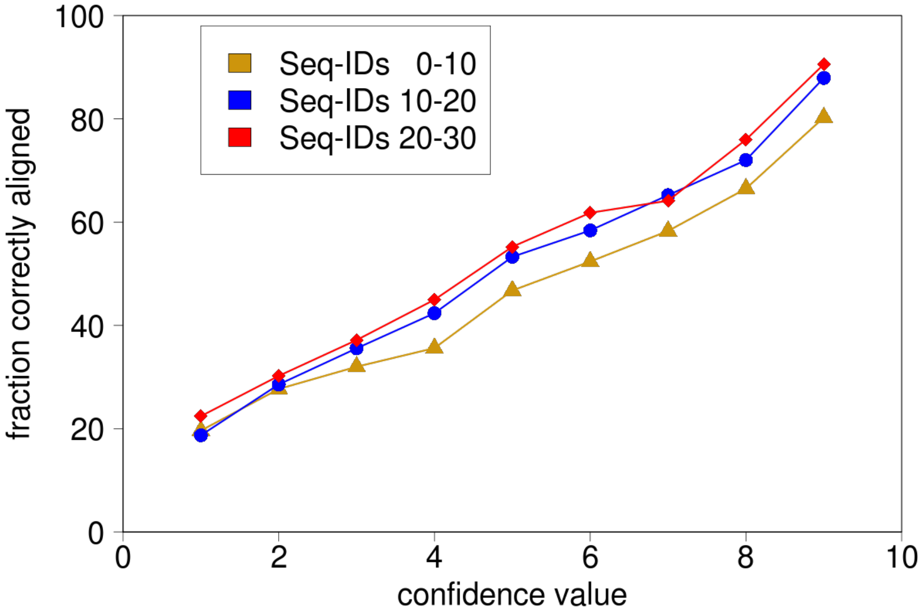

Figure S7: Relationship between HHblits/HHsearch confidence estimates from the maximum accuracy algorithm and the probability for a residue pair to be correctly aligned. The confidence values have an excellent correlation with the fraction of correctly aligned columns, and are nearl independent of the sequence identity between query and template sequences.

Table S1: Improvement of PSIPRED secondary structure prediction accuracy through HHblits multiple sequence alignments (MSAs). Performance is measured on the sequences from the PDBselect 2007 data set by Segment Overlap score (SOV) and 3-state accuracy (Q3). For each sequence in this set, MSAs are generated by 2 and 3 iterations of PSI-BLAST and 1, 2 and 3 iterations HHblits. Even 1 iteration of HHblits yield better performance than the standard PSIPRED, which uses 3 iterations of PSI-BLAST. The best results with an improvement of more than 1% are achieved by performing up to 3 iterations of HHblits and filtering the generated MSAs to a diversity of $N_{eff} = 7$.

input alignments	**SOV**	**Q3**
2 iterations PSI-BLAST	74.64%	77.31%
3 iterations PSI-BLAST	77.52%	80.38%
1 iteration HHblits	77.87%	80.71%
2 iterations HHblits	78.31%	80.99%
3 iterations HHblits	78.12%	80.83%
HHblits diversity 7	**78.62%**	**81.31%**

Table S2: List of 394 PFAM families for which no homologous template is known, HMMER3 has no match in the PDB below an E-value of $< 10^{-3}$ and for which HHblits has a match in the PDB database with E-value $< 10^{-3}$. In each row, the best HHblits match is given with its E-value and the coverage of the PFAM query. The last column specifies the HMMER3 E-value for the best match, an '-' indicates that HMMER3 has no matches up to the default reporting E-value threshold of 10.

PFAM-ID		**HHblits**		**HMMER**
	hit	**E-value**	**cov(%)**	**E-value**
PF03115	3hag	2e-101	53.24	-
PF11838	2xdt	1.1e-55	99.77	1
PF09562	2oa9	1.2e-54	98.85	-
PF10991	1je5	4e-54	96.02	-
PF04412	1c96	2.6e-53	83.81	0.006
PF09863	3iv3	1.2e-50	95.77	0.0048
PF12043	3c5n	1.9e-49	98.81	0.26
PF11047	3cxb	2e-49	73.10	0.049
PF07718	1r4x	3.8e-49	86.38	0.051
PF07632	2mas	2.1e-47	95.94	-
PF08010	2b3w	4.2e-45	96.58	0.14
PF11329	3eu8	9.5e-44	98.64	0.31
PF03813	1q79	1e-43	47.19	0.017
PF05316	3bbn	3.4e-41	90.71	-
PF09520	1wte	5.9e-40	97.33	-
PF12264	1qqp	7.5e-38	90.26	0.043
PF11161	2ra9	7.2e-37	77.46	-
PF09739	3f8t	2.8e-36	70.51	-
PF10963	3fgx	4.5e-36	100.00	0.0027
PF05864	2waq	5.8e-36	87.30	0.022
PF04486	1tuw	8.2e-36	77.39	-
PF05714	1w33	1e-35	86.26	0.078
PF05428	3kq4	1e-34	92.90	0.06
PF07588	1koe	2.1e-34	95.83	0.059
PF09536	2kii	1.1e-33	95.60	-
PF03662	1qw9	1.5e-33	99.38	-
PF11443	2nvo	1.6e-33	92.88	-
PF05538	2odj	1.7e-33	95.05	-
PF07520	1yuw	2.9e-33	58.91	-
PF06124	2r41	4.1e-33	100.00	-
PF05291	2ilr	4.7e-33	60.86	0.073
PF06045	1nkg	2.3e-31	92.65	-
PF06787	2a9s	4.6e-31	99.38	-
PF10023	1z5h	8.3e-31	93.51	0.061
PF05986	3ghm	1.4e-30	100.00	-
PF05482	1tr2	4e-30	85.30	-
PF11686	1se7	1.6e-29	100.00	-
PF07307	3nf2	1.8e-29	80.95	0.027
PF07528	1px5	2e-29	94.19	0.46
PF03254	2de0	2.4e-29	74.48	0.5
PF07395	1lrz	5.9e-29	98.86	-
PF10287	3iln	6.3e-29	91.81	-
PF01531	2hhc	7.1e-29	92.88	0.12
PF06074	3kdr	1.8e-28	56.66	0.14
PF06128	1yyh	2e-28	99.65	0.17
PF02088	1dec	1.2e-27	100.00	0.011
PF05136	3kdr	1.2e-27	90.17	-
PF07756	2qc0	1.5e-27	97.69	-
PF04681	1z3q	1.8e-27	86.45	-
PF09865	2w7q	1.8e-27	96.79	-
PF08189	2b5b	3e-27	94.87	0.031
PF10770	2plg	5.3e-27	82.88	0.014
PF11841	3dad	6e-27	99.36	-
PF06245	1vk1	6.1e-27	52.51	0.78
PF05060	1fo8	7.7e-27	75.42	0.062
PF12243	3d9j	1.1e-26	99.27	-
PF09960	2w3z	1.6e-26	43.59	1.2
PF10962	1v9m	2.1e-26	87.63	-
PF07014	2nw8	2.3e-26	95.20	-
PF10100	3c7a	2.7e-26	92.56	0.25
PF04841	1got	7.2e-26	70.80	-
PF07608	2yzy	1e-25	96.71	-
PF07379	3ci0	1.3e-25	79.43	-
PF03687	3bry	3.2e-25	87.35	0.11
PF07592	3hot	3.6e-25	93.57	-
PF05651	2a2l	7.2e-25	82.22	0.0057
PF12260	2acx	1.1e-24	89.76	-
PF06230	1zcz	1.5e-24	69.67	-
PF05213	1vgj	3.9e-24	66.67	0.0084
PF11768	1vyh	4.8e-24	61.72	0.097
PF04551	1tx2	4.9e-24	71.68	0.07
PF05550	2wur	5.5e-24	74.40	-
PF03290	1th0	7.4e-24	53.85	1.5
PF09927	2jxp	8.9e-24	99.15	6.2
PF09807	3bs4	9.5e-24	97.60	0.087
PF06199	2k4q	1.1e-23	100.00	0.013
PF08472	1tp6	1.2e-23	84.21	-
PF08553	1got	1.7e-23	42.52	0.15
PF11340	3cz8	1.8e-23	86.14	0.17
PF09892	3na6	2.3e-23	84.54	0.091
PF09674	1kea	5.4e-23	81.06	0.012
PF05551	1a73	5.6e-23	52.24	0.91
PF08695	2ciu	7.4e-23	76.56	0.053
PF04405	2k5e	7.9e-23	98.21	0.0013
PF10179	1fnh	8.7e-23	95.62	0.66
PF02677	1wy5	9.2e-23	94.12	9.9
PF04708	3guv	1e-22	62.02	-
PF07006	2k3d	1.2e-22	66.40	-
PF10250	2hhc	1.8e-22	94.17	0.099
PF11813	3h2d	1.9e-22	71.67	-
PF08928	2fef	2.3e-22	100.00	0.029
PF09859	3itq	2.5e-22	83.73	-
PF10107	3fov	8.5e-22	48.75	0.89
PF09843	2zws	8.6e-22	73.60	-
PF08470	2vxr	1.2e-21	59.88	-
PF11017	2hcy	2.4e-21	98.75	0.061
PF04937	2bw3	2.6e-21	99.35	0.21
PF10222	1h54	2.7e-21	56.50	-
PF03336	2jig	2.9e-21	46.15	0.23
PF10677	2f1c	3.4e-21	97.85	0.48
PF06420	1h2i	3.9e-21	56.50	-
PF06674	3dtd	4.7e-21	46.13	3
PF09796	2fyu	7.8e-21	87.10	0.33
PF06951	1lwb	1e-20	50.56	-
PF06437	2fue	1.2e-20	64.90	-
PF10748	2ivw	2.2e-20	72.39	-
PF07894	1byr	2.7e-20	58.80	0.0045
PF00609	2bon	4e-20	100.00	-
PF09517	1yd6	5.6e-20	69.01	0.77
PF11039	2vzy	5.9e-20	98.01	0.11

Table S2 continue

PFAM-ID		HHblits		HMMER
	hit	E-value	cov(%)	E-value
PF03351	1d7b	9.8e-20	96.80	0.0047
PF12541	1ogo	1e-19	79.50	-
PF11680	3k44	1.2e-19	63.64	0.9
PF12439	1v7w	1.5e-19	99.55	-
PF05176	3gkn	2e-19	62.70	-
PF09810	3l0a	2.1e-19	54.39	0.011
PF10738	3lyd	2.3e-19	55.51	0.0022
PF05046	2ogh	2.7e-19	88.89	0.043
PF05342	3n6z	3.4e-19	48.98	-
PF10288	1ni5	3.4e-19	98.95	0.14
PF09778	3erv	4.4e-19	99.12	0.0024
PF04788	3bk5	5.1e-19	84.50	-
PF03214	1qg8	6.2e-19	34.00	-
PF06544	2iyg	6.3e-19	93.42	0.39
PF11854	2guf	7.5e-19	82.01	-
PF01973	2p2v	1e-18	85.88	0.0029
PF07845	1m2d	1e-18	78.95	0.012
PF08885	1ivn	1.1e-18	79.34	0.051
PF08642	1jbi	1.4e-18	85.27	3.5
PF07607	1z5h	1.4e-18	96.00	0.39
PF06241	1lnq	1.4e-18	78.16	-
PF07076	2qcp	1.4e-18	83.54	0.26
PF10365	3km5	1.5e-18	89.51	0.49
PF07959	1yp2	1.5e-18	65.83	0.38
PF10934	2ia7	1.7e-18	89.32	-
PF02411	2h3o	3.5e-18	54.78	-
PF10012	3h96	3.7e-18	80.00	0.028
PF07115	2ia7	4.1e-18	90.09	0.0018
PF07293	3i9v	6.3e-18	97.44	-
PF06477	2ag4	7.4e-18	97.32	-
PF06021	1sqh	1.1e-17	99.51	-
PF03018	2brj	1.2e-17	77.78	0.0043
PF04114	3k9t	1.3e-17	44.98	-
PF05565	2p2u	1.3e-17	74.84	0.016
PF07905	2ioj	1.3e-17	91.06	0.012
PF11751	3bry	1.4e-17	89.78	-
PF09565	2c1l	1.5e-17	61.20	-
PF08734	2zbc	1.8e-17	82.42	0.0045
PF11019	3mc1	3.2e-17	75.37	0.0029
PF06044	2jne	3.4e-17	18.43	-
PF04865	3h2t	5e-17	74.60	-
PF12362	2aya	5.2e-17	84.62	-
PF05991	1exn	6.8e-17	98.14	-
PF06622	1o9y	7.3e-17	24.59	-
PF08521	2kse	1.3e-16	97.95	0.05
PF08480	1ru4	1.4e-16	99.46	-
PF03302	1yy9	1.8e-16	77.92	-
PF07506	1zx4	2.1e-16	98.83	0.0047
PF01185	2fmc	5.1e-16	79.63	3.1
PF10087	2iw1	6.2e-16	95.65	0.057
PF11814	3erv	6.2e-16	90.00	0.15
PF03452	1xhb	7.3e-16	93.31	-
PF10703	2p8g	8.9e-16	30.65	0.12
PF02413	2kz6	1.2e-15	57.14	0.71
PF07327	1wqj	1.3e-15	55.14	0.35
PF11356	2ivw	1.5e-15	54.17	0.22
PF12055	1k1x	2.3e-15	58.25	-
PF09337	3nnq	5.1e-15	100.00	0.022
PF08424	3dss	5.1e-15	90.61	0.031
PF08170	3gir	8.5e-15	88.78	0.046
PF08379	3isr	1.1e-14	100.00	-
PF06805	3dwg	1.3e-14	42.70	-
PF09363	1umd	2.4e-14	72.41	-
PF10826	2fe3	2.7e-14	85.19	0.46
PF07611	3bma	3.4e-14	80.00	0.021
PF08757	3dnu	3.5e-14	58.54	-
PF07461	1tvg	4.2e-14	31.39	0.029
PF09941	1vet	7.7e-14	91.67	-
PF10141	2zxr	8.4e-14	75.13	0.11
PF09366	2pcs	9.6e-14	93.71	-
PF05272	2dhr	1e-13	64.00	1.2
PF07618	1y6u	1.2e-13	98.25	0.46
PF11824	2okx	1.3e-13	77.90	0.038
PF10483	3bs4	3.4e-13	79.18	-
PF10029	3c12	3.5e-13	81.51	0.095
PF11863	3hxl	4e-13	95.55	0.016
PF06147	3g27	4.2e-13	42.93	0.1
PF10743	1y6u	4.2e-13	70.93	0.047
PF10246	1k0r	4.3e-13	82.86	-
PF11288	3fak	5.7e-13	66.99	0.0043
PF05227	1vls	8.2e-13	97.10	0.17
PF11845	3b42	8.3e-13	59.88	2.2
PF10302	2bps	8.8e-13	34.29	0.019
PF11312	3mgg	9.4e-13	55.44	-
PF11071	1s2d	1.8e-12	99.29	0.3
PF10037	1xi4	2.3e-12	75.06	-
PF04781	1elw	2.9e-12	96.72	0.27
PF09530	2i71	4.1e-12	74.59	0.023
PF07800	3knv	4.9e-12	80.62	0.1
PF08156	3id6	6.1e-12	100.00	2.8
PF06381	3kdr	8.8e-12	95.51	0.05
PF04244	2wq7	9.2e-12	69.78	-
PF10127	3c18	1.2e-11	79.20	-
PF08130	1w9n	1.2e-11	51.79	-
PF11959	3hft	1.4e-11	84.09	0.27
PF02066	1m0j	1.5e-11	51.85	3.2
PF04986	1omh	1.6e-11	33.68	-
PF04082	2veq	1.6e-11	29.30	-
PF10138	3ibz	1.6e-11	78.49	0.14
PF06075	2b29	2.1e-11	20.55	-
PF03490	2plc	2.5e-11	72.55	-
PF01927	3ga8	2.9e-11	37.58	0.035
PF09345	1h4x	2.9e-11	84.00	0.13
PF11761	3eeq	2.9e-11	100.00	0.1
PF06881	2e31	3.4e-11	70.09	0.011
PF04492	3e6c	3.5e-11	82.00	0.02
PF06890	2p5z	4.2e-11	46.03	-
PF08303	1yj5	4.7e-11	98.82	0.011
PF03281	1px5	6.3e-11	97.49	0.12
PF08497	2yxb	1.5e-10	47.75	-
PF10030	2jyx	1.5e-10	63.83	-
PF11997	3c48	1.7e-10	99.64	-
PF02697	3fmt	1.7e-10	86.67	0.0022
PF02474	1m4i	2.1e-10	73.10	-

Table S2 continue

PFAM-ID		HHblits		HMMER
	hit	E-value	cov(%)	E-value
PF12226	3iyo	2.5e-10	43.04	-
PF10567	1l3k	2.7e-10	69.36	-
PF09826	1fwx	4.3e-10	81.37	0.014
PF03158	2xeh	5.9e-10	75.65	-
PF08371	3hsi	6.6e-10	86.42	-
PF02666	2gpr	9.4e-10	90.23	-
PF06676	2waq	1.4e-09	37.86	0.59
PF06322	3e7l	1.5e-09	67.19	0.09
PF08685	1z3u	1.9e-09	24.00	-
PF07508	2r0q	1.9e-09	56.14	-
PF09317	2z1q	2e-09	48.75	-
PF07328	2ba3	2.2e-09	27.89	-
PF10908	1zat	2.9e-09	75.70	-
PF10474	2fji	3e-09	96.58	-
PF02521	3jty	3.5e-09	57.42	-
PF08499	3g4g	3.7e-09	88.52	0.025
PF06669	3d9x	3.8e-09	97.14	-
PF11954	1ei5	4e-09	63.87	0.028
PF07202	1h3i	4.5e-09	71.98	0.22
PF12010	1j1n	5.1e-09	94.20	0.12
PF04936	3hot	7.7e-09	78.49	8.1
PF08116	1c6w	7.8e-09	87.10	0.017
PF12000	3fro	8.5e-09	69.59	-
PF11766	1n67	8.6e-09	97.18	-
PF11711	2qv7	1.2e-08	32.09	-
PF06883	1twf	1.7e-08	100.00	2
PF08074	1ofc	1.7e-08	41.07	1.3
PF08465	1p6x	2.1e-08	96.97	-
PF05263	2o8x	2.5e-08	48.89	0.095
PF12128	1w1w	3.2e-08	6.19	0.0075
PF09889	1lv3	3.3e-08	46.55	0.059
PF04450	1z5h	4.3e-08	86.00	0.023
PF11308	2zxq	4.5e-08	49.71	-
PF09597	2e8n	4.7e-08	98.25	2.3
PF09824	2p4w	5.3e-08	78.75	0.0017
PF10780	1s3a	5.3e-08	100.00	-
PF11658	3lxq	6.1e-08	58.70	-
PF06011	1nep	7.2e-08	22.92	-
PF04407	3dcm	8.4e-08	95.43	0.0061
PF11853	2odj	8.5e-08	69.22	0.049
PF09582	1p90	9.6e-08	50.46	-
PF07610	2qsv	1e-07	100.00	0.015
PF11325	2vw9	1.2e-07	98.85	-
PF10686	2nx2	1.2e-07	88.73	0.017
PF01439	2kak	1.3e-07	93.67	0.34
PF08737	2fau	1.3e-07	77.88	-
PF05380	1rw3	1.7e-07	86.67	-
PF07699	2hey	1.9e-07	100.00	0.0045
PF06378	1h2i	2.5e-07	80.50	6.1
PF05610	2apn	3.6e-07	89.47	0.19
PF10758	3hxl	4.1e-07	99.45	-
PF04189	1yb2	4.4e-07	77.70	-
PF09855	2k4x	5.5e-07	82.81	0.22
PF06355	1gwy	6.1e-07	76.52	-
PF05895	2fl8	6.6e-07	28.31	-
PF04917	1oqw	6.8e-07	15.71	0.0036
PF04572	2vk9	7e-07	80.00	0.087
PF10144	3b42	7.5e-07	54.29	0.061
PF08405	3i86	8.4e-07	12.01	-
PF07087	1lwb	9.7e-07	77.17	-
PF09970	2fcl	9.7e-07	77.30	-
PF06977	1npe	1e-06	89.81	0.0024
PF07429	2gek	1.1e-06	80.06	-
PF10349	2hth	1.3e-06	32.41	-
PF10116	3e20	1.3e-06	98.57	-
PF08417	3gke	1.3e-06	59.43	-
PF10711	2vxz	1.9e-06	82.65	0.014
PF12303	2wg3	2.1e-06	61.70	0.0061
PF07813	3epv	2.2e-06	88.46	0.0022
PF10373	1ya0	3.5e-06	72.60	-
PF02681	2ipb	4.2e-06	90.54	0.27
PF11897	2gj4	4.2e-06	42.86	0.008
PF10987	3h35	4.5e-06	72.65	0.1
PF07617	3ia8	4.7e-06	98.18	-
PF03345	2gk3	4.9e-06	54.17	-
PF03850	3ibs	5.2e-06	80.92	0.011
PF07919	2icn	5.3e-06	62.70	0.069
PF10781	1whg	5.3e-06	98.39	-
PF07107	3ec9	5.5e-06	69.39	0.0012
PF05782	1kxp	5.5e-06	52.22	-
PF12578	1lw3	6.2e-06	66.29	0.29
PF04377	3gkr	6.4e-06	95.35	-
PF11903	1baz	6.6e-06	47.95	-
PF12073	1pjr	6.8e-06	94.23	0.26
PF04312	1hjr	8.3e-06	76.98	0.025
PF07480	3epv	8.6e-06	94.74	0.58
PF12525	1v9n	8.7e-06	86.67	0.28
PF04413	1vgv	8.8e-06	86.34	-
PF05091	3fqi	9.4e-06	53.89	0.39
PF10367	1chc	9.7e-06	29.36	6.2
PF06353	2fph	1e-05	36.81	-
PF06239	1xi4	1e-05	65.67	-
PF08192	1hpg	1.1e-05	13.75	-
PF08579	1xi4	1.2e-05	82.50	-
PF05510	1u2c	1.3e-05	39.78	-
PF11833	1faf	1.3e-05	22.06	3.9
PF06823	2l1s	1.4e-05	80.33	-
PF09984	3b42	1.5e-05	93.29	-
PF09352	2qgp	1.6e-05	43.68	-
PF06375	2g30	1.9e-05	34.78	-
PF03406	1h6w	2.1e-05	90.70	-
PF12340	3ly5	2.1e-05	77.73	0.036
PF10952	2fbn	2.2e-05	88.03	-
PF10240	2qp2	2.2e-05	37.80	-
PF02754	3cf4	2.5e-05	80.95	0.91
PF10941	1uf3	2.5e-05	55.08	0.054
PF05689	1f00	2.5e-05	99.45	-
PF01941	2p02	2.6e-05	91.86	0.62
PF11839	1jcd	2.7e-05	39.76	-
PF06956	1xmx	2.7e-05	75.94	-
PF07585	2x55	2.8e-05	62.71	-
PF06448	1lsh	3.2e-05	39.33	-
PF10865	1ilo	3.2e-05	51.28	1
PF10505	3fqi	3.5e-05	72.90	-

Table S2 continue

PFAM-ID		HHblits		HMMER
	hit	E-value	cov(%)	E-value
PF11865	2qk1	3.9e-05	96.89	0.65
PF12222	1pgs	4e-05	69.09	0.21
PF00746	2ww8	4.1e-05	92.50	0.36
PF09759	1xqr	4.9e-05	72.63	1.2
PF11834	2dnf	5.1e-05	98.48	5.4
PF10126	3dfe	5.5e-05	92.86	0.068
PF07878	1nla	5.7e-05	92.00	-
PF07505	3c8f	7.2e-05	81.89	0.1
PF04155	1yo3	7.6e-05	72.97	-
PF10904	1j8b	7.6e-05	63.37	-
PF10706	2fcl	8.7e-05	88.51	-
PF01963	2g5g	8.9e-05	97.76	0.32
PF11336	2o4v	9.3e-05	77.48	0.014
PF03249	1p4t	9.4e-05	17.98	-
PF04599	1rxw	9.8e-05	65.00	-
PF01696	1pcl	0.0001	49.61	-
PF06702	1cja	0.00011	54.02	-
PF10122	2jr6	0.00013	80.39	10
PF11112	1z4h	0.00015	86.84	-
PF11849	3e0y	0.00016	94.77	-
PF06904	1lbu	0.00018	54.82	0.015
PF05918	1b3u	0.00019	60.90	0.0019
PF09854	2qgp	0.00019	23.82	0.021
PF10497	1wil	0.00025	61.76	0.71
PF07802	2k3j	0.00028	81.43	0.64
PF04305	3chh	0.00028	73.09	0.092
PF08736	2i1j	0.00028	51.06	-
PF06974	2jgp	0.0003	98.04	-
PF08288	3fro	0.0003	82.22	-
PF00242	1wz4	0.00031	7.33	-
PF09538	1vd4	0.00036	23.02	2.9
PF10673	3lub	0.00037	66.21	-
PF12215	2cqs	0.00038	63.13	-
PF11379	2cqy	0.00038	22.10	-
PF03258	2w7a	0.00045	46.67	0.094
PF10115	2kon	0.00046	76.34	-
PF08498	1wg8	0.00049	76.12	-
PF05444	3laq	0.00049	87.82	2.9
PF05869	3lkd	0.00055	81.40	0.06
PF11001	1wij	0.00058	65.61	0.0013
PF10309	3d45	0.00063	96.67	0.036
PF04049	3kae	0.00065	86.05	-
PF11006	2pxg	0.00067	87.21	1.6
PF07295	1lko	0.00068	25.85	0.096
PF10407	2ns5	0.00069	94.67	2.9
PF10165	1xm9	0.00073	77.28	-
PF08749	2plg	0.00074	92.41	-
PF04904	1rg6	0.00076	78.05	0.13
PF07409	2ia7	0.00078	63.25	0.049
PF12416	2dmh	0.00085	97.09	-
PF09576	1v54	0.00086	91.23	-
PF08004	2cob	0.00086	44.27	-
PF09415	1b67	0.00087	91.78	0.028
PF09894	1iru	0.0009	92.23	4.3
PF07855	2vfx	0.00091	95.73	0.16
PF10790	2al3	0.00095	92.11	0.019