Protein structure and function prediction by pairwise comparison of hidden Markov models

Johannes Söding
Max-Planck Institute for Developmental Biology, Tübingen

Abstract

Sequence similarity search methods that identify related proteins in large sequence databases are the most important application of bioinformatics in the biological sciences, since they allow to make predictions about a protein's function, structure, and evolution
I have developed the method HHsearch for the detection of remotely related proteins, which is three times more sensitive than standard methods like PSI-BLAST and considerably faster and more sensitive than the best alternative methods. To make the method accessible to a wider community, a web server based on HHsearch was set up (`hhpred.tuebingen.mpg.de`). It can search all popular protein family databases and returns ranked matches similar to PSI-BLAST in a matter of minutes. Several options assist in the functional analysis and 3D structure prediction. This server is complemented by a related server (`hhrep.tuebingen.mpg.de`) dedicated to the detection of internal repeats in protein sequences. Using HHrep, a clear sequence signal for the structural repeats in a number of common protein folds has been detected for the first time.

1 Introduction

The class of bioinformatic tools most often used by biologists are sequence similarity search methods, of which FASTA [1], BLAST [2], and PSI-BLAST

[3] are the most popular ones (with 7782, 20288, and 16411 citations to date). These methods identify *homologous* (i.e. related) proteins in sequence databases. In cases where an experimentally characterized, homologous protein can be identified, one can make inferences about the unknown protein, because closely related proteins (e.g. with $> 50\%$ identical amino acids in the pairwise sequence alignment) generally have the same or very similar functions. But for many proteins, no significant relationship to a protein of known function can be established, especially in the most interesting cases where the protein belongs to a family that has not yet been studied.

It is still not well known among biologists that, when conventional sequence search methods fail, recently developed, highly sensitive methods for remote homology detection [4, 5, 6] or structure prediction [7, 8, 9, 10, 11, 12, 13] quite often allow to make inferences from more distant relationships [14, 15, 16]. If the relationship is so remote that no common function can be assumed (e.g. if less than $\sim 30\%$ of amino acids are identical) one can generally still derive hypotheses about possible mechanisms, binding sites, functional residues, or the class of substrate bound [17].

When a homologous protein with known structure can be identified, it can be used as a *template* to model the 3D structure for the query protein [7], since even remotely homologous proteins generally have quite similar 3D structures [18]. The 3D model may then help to generate hypotheses to guide experiments.

2 Sequence alignments, sequence profiles, and HMMs

Sequence similarity search methods like FASTA or BLAST compare the sequence of a query protein with sequences of database proteins by *aligning* the two sequences, one below the other, in such a way that similar amino acids will preferably be in the same column. A *substitution matrix*, derived from the statistical analysis of many representative sequence alignments, quantifies the similarities between the twenty amino acids. The sequence similarity score is calculated as a sum over the substitution matrix elements of the aligned pairs of amino acid residues, minus penalties for gaps in the alignment. Clever heuristics speed up the calculation by a factor 10 to 100 with negligible loss in sensitivity.

The development of *profile*-to-sequence comparison methods such as PSI-BLAST [3] has lead to a vast improvement in sensitivity over these sequence–sequence comparison methods. A *sequence profile* is built from a multiple alignment of homologous sequences. It is a $20 \times L$ matrix that contains the fraction of each of the twenty amino acids in each of the L columns of the multiple alignment (Figure 1). The profile can be understood as a concise

statistical representation of the sequence family alignment, which contains more information about the sequence family than a single sequence. The profile allows to distinguish between conserved positions that are important for defining members of the family and nonconserved positions that are variable among the family members. More than that, it describes exactly how likely we are to find each of the amino acids at each position, which is why sequence profiles are sometimes called "position-specific substitution matrices". In practice, profile-based search methods work in an iterative fashion. After each search round, they add the significantly related sequences to the multiple alignment from which the sequence profile for the next round of database search is constructed.

```
HBA_human  ... W  G  K  V  G  A  -  -  H  A  G  E  ...
HBB_human  ... W  G  K  V  -  -  -  -  N  V  D  E  ...
MYG_phyca  ... W  G  K  V  E  A  -  -  D  V  A  G  ...
LGB2_luplu ... W  K  D  F  N  A  -  -  N  I  P  K  ...
GLB1_glydi ... W  E  E  I  A  G  A  D  N  G  A  G  ...
```

A	...	0	0	0	0	0.25	0.75		0	0.2	0.4	0	...
C	...	0	0	0	0	0	0		0	0	0	0	...
D	...	0	0	0.2	0	0	0		0.2	0	0.2	0	...
E	...	0	0.2	0.2	0	0.25	0		0	0	0	0.4	...
F	...	0	0	0	0.2	0	0		0	0	0	0	...
G	...	0	0.6	0	0	0.25	0.25		0	0.2	0.2	0.4	...
H	...	0	0	0	0	0	0		0.2	0	0	0	...
I	...	0	0	0	0.2	0	0		0	0.2	0	0	...
K	...	0	0.2	0.6	0	0	0		0	0	0	0.2	...
L	...	0	0	0	0	0	0		0	0	0	0	...
M	...	0	0	0	0	0	0		0	0	0	0	...
N	...	0	0	0	0	0.25	0		0.6	0	0	0	...
P	...	0	0	0	0	0	0		0	0	0.2	0	...
Q	...	0	0	0	0	0	0		0	0	0	0	...
R	...	0	0	0	0	0	0		0	0	0	0	...
S	...	0	0	0	0	0	0		0	0	0	0	...
T	...	0	0	0	0	0	0		0	0	0	0	...
V	...	0	0	0	0.6	0	0		0	0.4	0	0	...
W	...	1.0	0	0	0	0	0		0	0	0	0	...
Y	...	0	0	0	0	0	0		0	0	0	0	...

Fig. 1: A multiple sequence alignment and its associated sequence profile.

A significant improvement over profile–sequence based methods was made possible by comparing profiles to profiles. These methods use PSI-BLAST or a similar method to build a profile for a query sequence and compare this profile with a database of precomputed profiles. Several such programs for homology recognition have recently been developed: LAMA [4], PROF_SIM [5], and COMPASS [6]. They were shown to be significantly more sensitive than PSI-BLAST and have been applied to identify evolutionary links between protein families previously thought to be unrelated. In addition, almost all of the top structure prediction servers now rely on profile–profile com-

parison, as can be seen from the results of the blind, automated structure prediction contest CAFASP [19].

Profile hidden Markov models (HMMs) are similar to simple sequence profiles, but in addition to the amino acid frequencies they contain the position-specific probabilities for inserts and deletions along the multiple sequence alignment. The logarithms of these probabilities are in fact equivalent to position-specific gap penalties [20]. Not surprisingly, profile HMMs perform better than sequence profiles in the detection of homologous proteins and in the quality of alignments [21,22,23], but despite the success of profile-profile alignment methods, the generalization to HMM-HMM comparison has not been done until recently.

3 Pairwise alignment of HMMs

A statistical theory of pairwise alignment of HMMs was independently developed by Lyngsgø *et al.* [24] and myself [25]. Both approaches start from the *co-emission probability* as a measure of similarity of two profile HMMs, i.e. the probability that the two aligned HMMs *emit* the same amino acid at each aligned position[1]. Lyngsgø *et al.* find the alignment that maximizes the logarithm of the co-emisssion probability, whereas our method sets the co-emission probability in relation to a *null model* probability describing the probability of emitting the same sequence under the assumption of unrelated proteins. More precisely, I look for the alignment that maximizes the score, defined as the logarithm of the sum over all co-emittable sequences of the ratio of co-emission probability to null model probability. It can be shown [25] that the use of a null model considerably improves performance by giving more weight to the co-emission of rarer amino acids[2]. Furthermore, by including a null model our HMM-HMM alignment score reduces to the successful log-odds score of HMM-to-sequence alignment in the case when one HMM is constructed from a single sequence.

A profile HMM contains in each column a match state M, a delete state D and an insert state I (Figure 2). The transition probabilities between states, symbolized by the arrows, are calculated from the insert and deletion frequencies at each position in the multiple alignment. To align two HMMs, they must be able to emit the same sequence of amino acids in aligned columns.

[1] When interpreting the HMM as a generative model, we say Match states *emit* amino acids according to the amino acid distribution of the corresponding multiple alignment column. Insert states emit amino acids with a probability distribution equal to some mean frequencies in a sequence database.

[2] I speculate that limited performance due to the lack of a null model as well as a relatively cumbersome algorithm may be reasons why the method of Lyngsgø was never developed further or made publicly available.

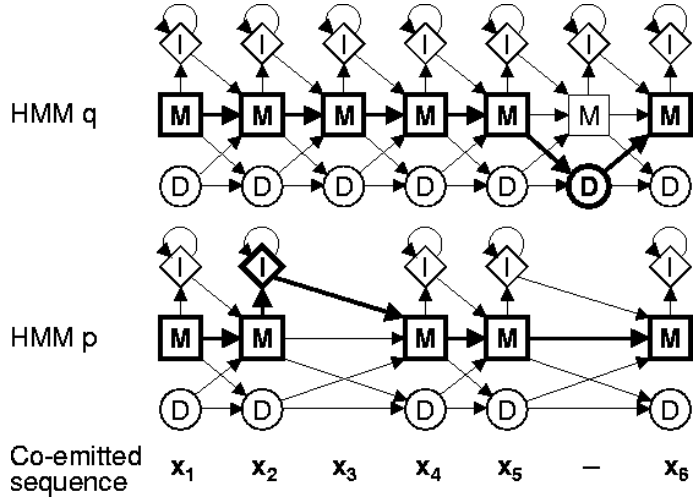

Fig. 2: Alignment of two HMMs. The path through the two HMMs corresponds to a sequence that is co-emitted by both HMMs. M: match states, D: delete states, I: insert states. Match states can emit amino acids with a probability distribution given by the corresponding column of the multiple alignment (see previous figure).

Match and insert states emit amino acids whereas delete states don't. Therefore, a match or insert state in one HMM can only be aligned with a match or insert state in the other HMM. Conversely, a delete state can only be aligned with a delete state or with nothing (i.e. with a *gap*) (Figure 2). As an example, in the third column of the alignment in Figure 2, HMM q emits a residue from its M state and HMM p emits a residue from the I state. In column six of the alignment, HMM q does not emit anything since it passes through the D state. HMM p does not emit anything either, since it has a gap in the alignment.

I use *dynamic programming* to iteratively solve the task of finding the highest-scoring alignment of two HMMs. Simply speaking, one calculates the score of the optimal alignment up to columns (i,j) from the optimal alignments up to ($i-1$,$j-1$), ($i-1$, j), and (i, $j-1$).

Since protein structures diverge much more slowly than sequences, it makes sense to include a comparison of secondary structures when aligning putative remotely homologous sequences. I have developed a statistical method akin to amino acid substitution matrices that also takes into account the confidence with which the secondary structure state (alpha helix, beta strand, or coil) at each position is predicted. The similarity scores between predicted states are then simply added to the amino acid-based scores of each column.

4 Benchmark comparison

I performed an all-against-all comparison with various similarity search tools to test their ability to detect remotely homologous proteins and to produce high-quality alignments below the twilight zone [26] of sequence similarity. I compared BLAST, PSI-BLAST, the HMM–sequence comparison package HMMER , the profile–profile alignment tools PROF_SIM and COMPASS, and our method HHsearch. In order to pinpoint the source of improvements, I benchmarked four versions of HHsearch. HHsearch 0 uses simple profile–profile comparison, HHsearch 1 is the basic HMM–HMM version, HHsearch 2 includes a novel correlation score [25], and HHsearch 3 and 4 additionally score secondary structure similarity (with predicted vs. predicted and predicted vs. actual secondary structure).

The SCOP hierarchical database [27] of structural domains was used as test set, since any pair of sequences from the same SCOP superfamily can safely be assumed to be homologous, whereas every pair with different folds are assumed to be unrelated. We will call these pairs *true positives* and *false positives*, respectively. SCOP (version 1.63) was filtered to obtain a set of 3691 sequences with a maximum pairwise *sequence identity* of 20% (i.e. having no more than 20% identical residues in a pairwise alignment). A multiple alignment was built from each sequence by using PSI-BLAST with up to eight iterations and an HMM was calculated from each of theses alignments.

Sensitivity: In order to assess the ability of the methods to distinguish true from false positives, we plot in Figure 3 the number of true positives versus the number of false positives detected above a score threshold. The ideal method would detect all homologous relationships before the first non-homologous pair is reported, yielding a vertically rising graph.

In short, HHsearch finds about twice as many homologous pairs at constant error rate of 10% (dashed diagonal line) as the next best method and more than three times as many as PSI-BLAST or HMMER. The improvement over the best alternative method (COMPASS) is due, to about one third, to the inclusion the statistical scoring scheme and the preparation of profiles (compare traces for COMPASS and HHsearch 0 in Figure 3). Another third is gained by using HMMs instead of simple profiles (compare HHsearch 0 with 1), and the last third is owed to the inclusion of secondary structure and correlation scoring (compare HHsearch 1 with HHsearch 3 or 4).

Alignment quality: In comparative structure modeling, the alignment quality between query and template sequence is the key determinant of model quality [28]. The quality of sequence alignments can be assessed by looking at the spatial distances between aligned pairs of residues upon superposition of their 3D structures. A measure for this structural fit is the MaxSub score [29]. It is 1.0 for a perfect structural fit between query and template

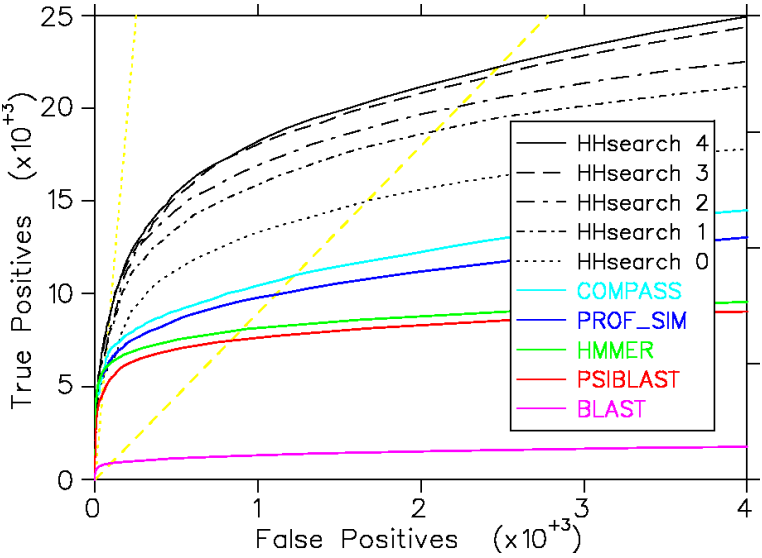

Fig. 3: Sensitivity of various homology detection tools, measured by how many true positives are detected at varying numbers of false positives. True positives are homologous pairs, false positives are unrelated. For definitions of HHsearch 0–4 please refer to the main text.

structures when all query residues are aligned at RMSD of 0 Å, and zero if the similarity is found to be insignificant.

Figure 4 plots the percentage of pairs found with scores in ten score bins, when all pairs with proteins from the same family (a) or the same superfamily (b) are considered. Pairs with MaxSub score of zero are omitted from the lowest bin. In conclusion, HHsearch is able to produce more alignments with scores above 0 than any of the other methods, and this difference increases for the more difficult inter-family alignments (b).

Speed: HHsearch scans a query sequence of 200 residues against 3691 domains in 33 s on an Athlon 64 3200+ PC. This is 10 times faster than PROF_SIM, 17 times faster than COMPASS, and only 2.5 times slower than the HMM-to-sequence comparison method HMMER. This speed was achieved through an efficient algorithm, rigorous profiling, and implementation of fast logarithm and power functions. In addition, HHsearch is parallelized (using POSIX threads) to run on multiple processors of an SMP machine, with good scaling properties (2 CPUs give a speed-up of $\times 1.7$ on a dual Athlon 64 3200+ machine under Linux).

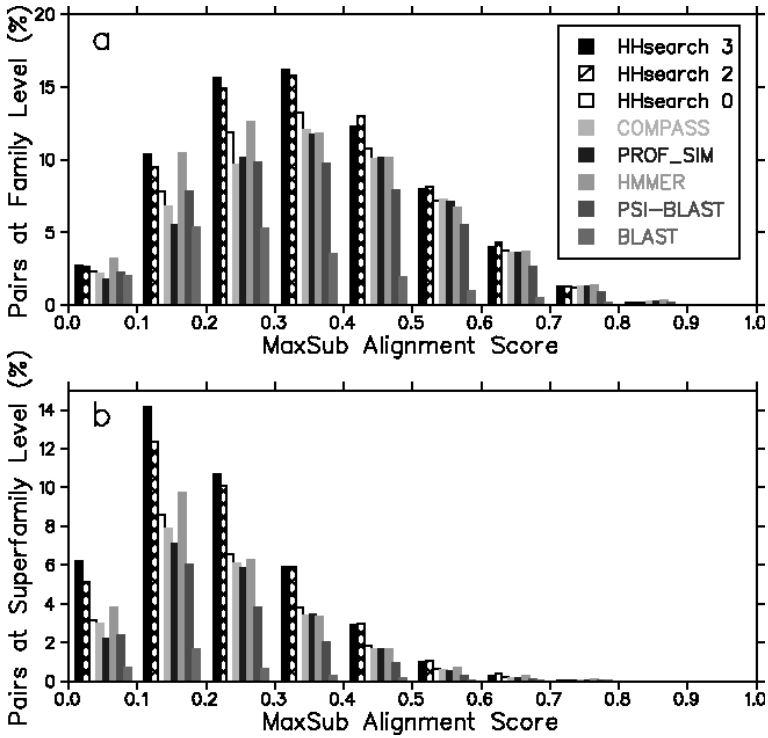

Fig. 4: Distribution of MaxSub scores for alignments of domain pairs related at the family or superfamily level in percent. Counts with MaxSub score of zero are not shown.

5 The HHpred server for structure and function prediction

The web server HHpred was developed to provide biologists with a method for sequence database searching that is as easy to use as BLAST or PSI-BLAST and yet competitive in sensitivity with the most powerful, specialized servers for structure prediction available.

Upon submission of a query sequence (or alignment), HHpred proceeds in three steps. First, an alignment of homologs is built for the query sequence by multiple iterations of PSI-BLAST searches. In the next step, a profile HMM is generated from the multiple alignment that also includes the information about predicted secondary structure. In the last step, the query HMM is compared to each HMM in the selected database. The database HMMs have been precalculated and also contain secondary structure information, either predicted by PSIPRED, or assigned from 3D structure by DSSP [30]. The server then presents the results organized into three sections: a graphical overview

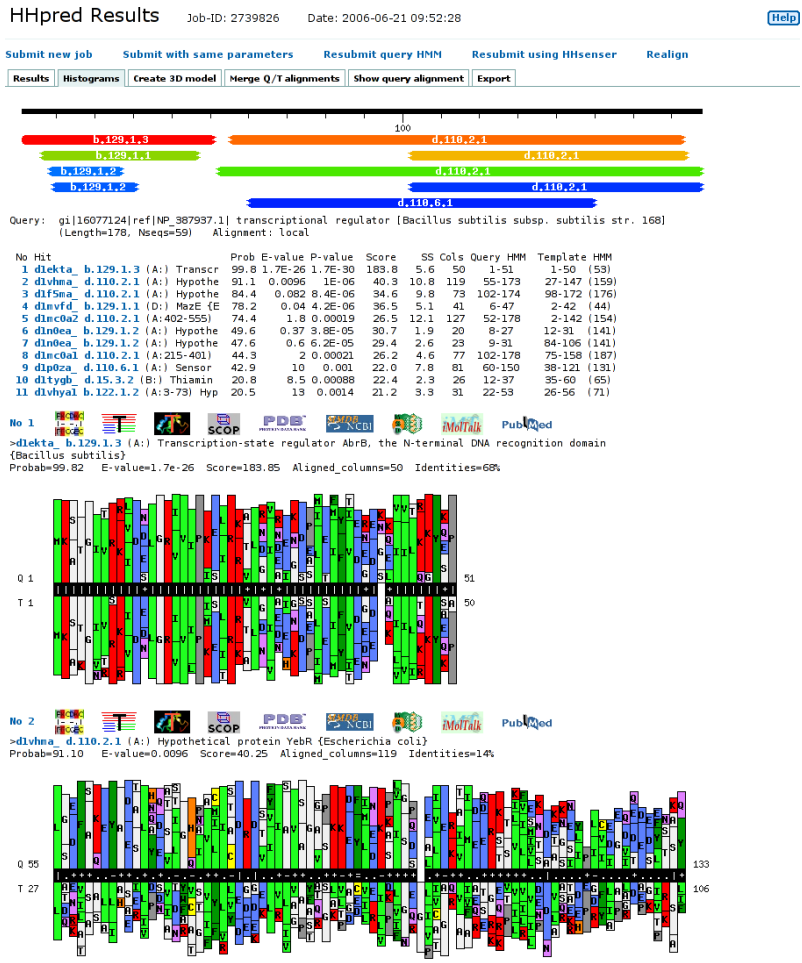

Fig. 5: Search results for HHpred at the example of transition state regulator SpoVT. The bar graph and summary hit list at the top show that SpoVT consists of two domains: the N-terminal domain is very similar to AbrB (rank 1) and clearly homologous to MazE (rank 4) and the C-terminal domain is similar to GAF and PAS domains (rank 2, 3, 5 etc). In the summary table, column 'Prob' lists the probability that the hit is homologous to the query. The alignments can be either displayed as annotated sequence alignments, or, like here, as histograms giving the amino acid distributions in the profile HMM columns. Amino acids with less than 10% are not shown. The coloring indicates the class, e.g. aliphatic, aromatic, polar etc. Various links and options provide means to further analyze results or generate a structural model.

of location and significance of the best matches, a summary table, and detailed pairwise query-template alignments (Figure 5). I believe that HHpred is unique in the advantages it offers:

Databases: In addition to the PDB and SCOP databases, all of the standard protein family databases can be searched and are automatically updated: Pfam [31], SMART [32], COG/KOG [33], CDD [34], InterPro [35], TIGRFAM [36], Panther [37], PIRSF [38], and CATH/Gene3d [39]. This sets HHpred apart from most other servers able to detect remote homologies, which are generally more specialized for protein structure prediction and only offer searches of the PDB.

User-friendliness: Search results are presented in an easy-to-read format similar to BLAST. Alignments contain annotation about secondary structure, consensus sequences, and position-specific reliability, and a histogram view of the HMM-HMM alignments permit to quickly identify functional motifs (Figure 5).

Flexibility: We try to offer the user maximum control and flexibility. One can paste one's own query alignment, search in local or global alignment mode, realign with other parameters, edit the query-template (multiple) alignment with which to launch the comparative modeling, merge the query HMM with database HMMs for *intermediate profile search*, view structures of templates or computed models, and so forth. Furthermore, HHpred is embedded in our web-based MPI Bioinformatics toolkit, which integrates many in-house and public tools in one convenient environment.

Selectivity: High-scoring false positives have systematically been reduced by developing a protocol for building query and database alignments that suppresses non-homologous sequences (J. Söding, to be published).

Sensitivity: HHpred is among the most sensitive servers for remote homology detection. A comparison of the new version with the servers that took part in the structure prediction benchmark CAFASP4 [19] can be viewed at `http://protevo.eb.tuebingen.mpg.de/hhpred/hhpred_in_CAFASP4.html`. Recently, we have integrated a new method for automatized exhaustive intermediate profile search, HHsenser [40], which can be called from within HHpred, to further increase sensitivity.

Documentation: Detailed help pages (>13000 words) are available.

6 Applications

HHpred now processes over 2000 external queries per month. A year after publication of HHsearch and HHpred, I found 36 articles citing them. Of these, three applied HHsearch on a large scale as a main method of analysis [41, 42, 43] and another 17 employed HHsearch or HHpred for detecting a remote homology relationship or predicting a 3D protein structure [44, 45,

46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60]. In the following, four examples that illustrate the use of HHsearch and HHpred are presented.

6.1 Structure and function prediction for Rtv from the fruit fly

The retroactive (rtv) gene of *Drosophila melanogaster* was identified at our institute in a genetic screen for chitin-associated developmental defects. No annotated homologous sequence were found with standard methods. A manual sequence search with PSI-BLAST and submission of the resulting alignment to HHpred resulted in the detection of a family of snake toxins and a family of extracellular receptor domains as distant homologs. From the latter relationship we could conclude that Rtv is an extracellular protein anchored to the cell membrane by a GPI linkage. Modeling the structure of the Rtv protein with a set of diverse templates resulted in a structure with three long, floppy, exposed loops that are held together by five disulphide bridges. The length of the loops is unique among relatives of Rtv, and each carries two exposed, aromatic residues at the end. Exposed, aromatic residues are known to bind sugar-derivatives like chitin. This lead us to the hypothesis that Rtv is involved in binding and organizing the chitin fibers emerging from the epithelial cell surface [60]. Recent preliminary experimental evidence confirms this prediction.

6.2 Detection of tandem BRCT domains in human Nbs1

Human Nbs1 (and its homolog Xrs2 from yeast) are part of the conserved MRN complex which plays a crucial role in maintaining genomic stability. NBS1 corresponds to the gene mutated in the Nijmegen breakage syndrome known as a radiation hyper-sensitive disease. Despite the importance of the MRN complex, the high sequence divergence between Nbs1 and Xrs2 prevented the identification of common domains downstream of the N-terminal Fork-Head Associated (FHA) domain. Using HHpred, a team from Saclay could identify three as yet undetected BRCT domains in Nbs1 and Xrs2 downstream of FHA [47]. Based on the hand-refined HHpred alignments, a structural model of FHA with the two BRCT domains was built, leading to the prediction that the duplicated BRCT domain acts as phospho-serine binding module in phosphorylation-dependent protein-protein interactions and to the identification of the binding surface for the phospho-serine carrying interaction partner. The model further shows that the phospho-binding sites of FHA and BRCT are at least 45 Å apart and, surprisingly, that there is not a single residue of linker between FHA and BRCT in any of the homologs, which hints at a tight coupling between the phospho-binding functions of the FHA and BRCT domains.

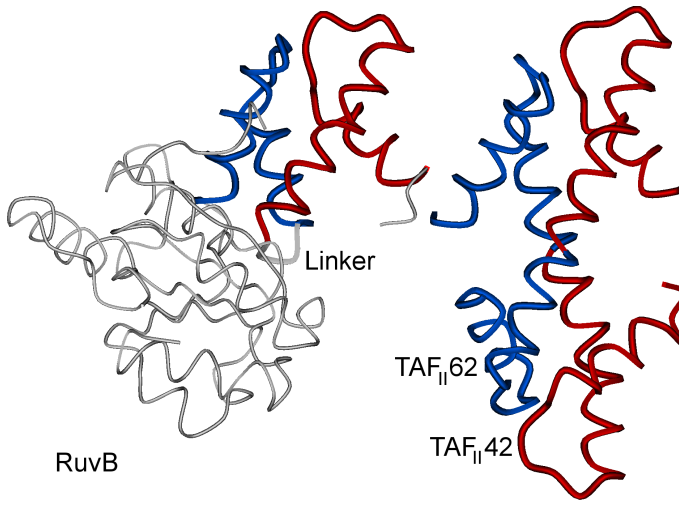

Fig. 6: The histone fold, represented here by the heterodimer TAF$_{II}$62/TAF$_{II}$42, evolved from the C-terminal subdomain of AAA+ ATPases like RuvB by the deletion of a linker and a 3D domain swap.

6.3 Novel members of the PD(D/E)XK nuclease superfamily

The PD(D/E)XK nuclease superfamily of Mg^{2+}-dependent nucleases groups together protein domains found in diverse enzymes involved in DNA replication, repair, and recombination. Typically, the sequence similarity between these proteins is so low that most members of this superfamily could be classified as PD-(D/E)XK nuclease only after their structures were determined experimentally. To find new members of this superfamily, Kosinski *et al.* [41] used the HHsearch package to build HMMs for all known members and to search the Pfam and COG databases for significant similarities. They report the identification of a PD(D/E)XK nuclease domain in numerous proteins implicated in interactions with DNA, but with unknown structure and function. The work will help to jump-start the experimental characterization of new nucleases, of which many will be important for the understanding of mechanisms that govern the evolution and stability of the genome.

6.4 Evolution of histones from a subdomain of AAA+ ATPases

In an all-against-all screen of homologous relationships between members of different folds (see section 4), we identified a striking similarity both in sequence and in structure between the histone proteins and the small helical subdomain of extended ATPase domains in AAA+ proteins (Figure 6). This

relationship is remarkable since it is conventionally assumed that different folds share no common ancestors. We conclude that the histones evolved from the ATPase subdomain, consisting of two alpha-helical hairpins connected by a short linker, by deletion of the linker and merging of the two inner helices into a long straight helix, and subsequent dimerization in order to preserve the tertiary interactions ("3D domain swap") [61].

7 The HHrep server for *de novo* repeat detection

Six out of the ten most populated folds possess an approximate structural symmetry [62, 63]. Most proteins that adopt one of these folds have no symmetry detectable in their sequences, however, and it is unclear for most domain families in these folds whether their structural symmetry has its cause in an origin through duplication. The ability to detect these structural repeats by their sequences would open a window to study hypotheses about the origin of these domains by duplication of simpler fragments. Furthermore, the detection of structural repeat patterns could help to predict the fold and function of sequences for which no detectable homolog with known structure can be found.

There are two general classes of methods to detect repeats in protein sequences. The first use their own database of profile HMMs or sequence profiles which are constructed from known repeat families, and they compare these profiles one by one with the query sequence. The second class is called *de novo* repeat detection methods: They do not rely on *a priori* knowledge about repeat families. Instead, they look for internal similarities by comparing the protein sequence to itself with standard sequence-sequence alignment techniques.

HHrep [64] is a web server for *de novo* identification of repeats in protein sequences, which is based on the pairwise comparison of HMMs. Its main strength is its sensitivity, allowing it to detect highly divergent repeat units in protein sequences whose repeats could as yet only be detected from their structures. Examples include sequences with β-propeller fold, ferredoxin-like fold, double psi barrels, or $(\beta\alpha)_8$ (TIM) barrels.

This is illustrated in Figure 7 at the example of the $(\beta\alpha)_8$ barrel structure of KDPG aldolase, by revealing a clear fourfold symmetry which we detect solely from sequence information. This symmetry points to an ancient origin through duplication of a $\beta\alpha\beta\alpha$ unit [64] and not, as previously hypothesized, by duplication of a half-barrel [65].

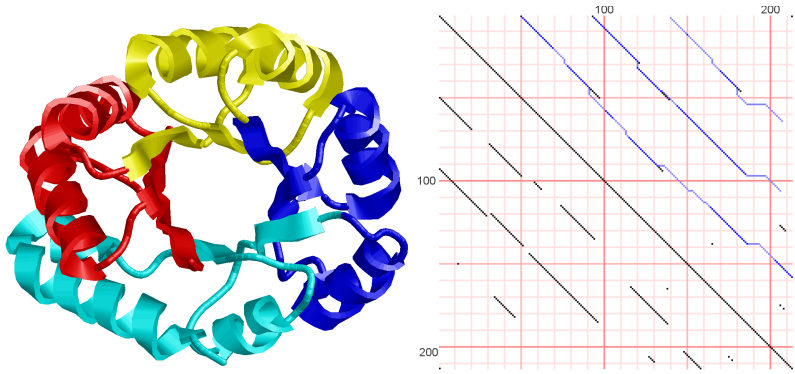

Fig. 7: The structure of $(\beta\alpha)_8$ barrels like KDPG aldolase (1fq0_A) is made up of four structural repeat units (left). The dot plot generated by HHrep shows for the first time a clear fourfold repeat structure in the sequence of a $(\beta\alpha)_8$ barrel (right). A black dot at position (i, j) indicates that profile columns i and j are similar.

8 The HHsearch software package

The HHsearch software package is programmed in C++ with a modular and object-oriented design. It consists of a program `hhmake` to generate HMMs from multiple alignments, `hhsearch` to search a database of HMMs (simple obtained by concatenating single HMMs), and `hhalign` to perform pairwise optimal or stochastically sampled alignment of HMMs and to generate dot plots as used by HHrep. In addition, a fast utility `hhfilter` can select a representative set of sequences by a maximum pairwise sequence identity criterion. With the software, users can download twelve standard, free family alignment databases in HHsearch-readable format, including our own pdb70 and scop70 databases.

Several scripts written in perl are offered with the package: `reformat.pl` can transform many standard multiple alignment formats into each other, `alignblast.pl` can parse a multiple alignment from PSI-BLAST output, `addpsipred.pl` adds predicted secondary structure to FASTA-formatted alignments or HMMs, and `hhmakemodel.pl` can parse the output of hhsearch or hhalign and generate merged multiple alignments in various formats or create rough 3D models.

The web servers have been set up as part of our MPI bioinformatics toolbox for protein sequence analysis [66] in a model-view-controller web framework. The new version is completely rewritten in Ruby on Rails (`http://www.rubyonrails.org/`) and will soon be released. It will also be made freely available under the GLPL license.

9 Summary and outlook

The HHsearch software for remote homology detection through pairwise comparison of HMMs has already found numerous applications in protein structure prediction, protein function prediction, and protein evolution, of which only a few could be mentioned in section 6. The HHpred web server was developed to make this method accessible to a wider community and to greatly enhance its functionality and usability for structure and function prediction. The HHrep server, which is based on the same method for HMM-HMM comparison, represents the most sensitive tool for *de novo* repeat detection in proteins.

I envisage many developments for protein function and structure prediction that build on the present methods. (1) A planned extension to HHpred is a PDBalert system. Users can enter a list of proteins in a web form, which will be automatically checked every week for similarity with the newly released protein structures. (2) We are working on a new method for comparative modeling that employs Bayesian statistics and advanced Markov chain Monte Carlo sampling techniques to simultaneously determine an optimal structural model and an improved query-template alignment (In collaboration with M. Habeck). (3) I will explore a way to speed up HHsearch by a factor of 10–100 by condensing the information contained in a single profile column into a discrete state alphabet and using fast heuristics developed for sequence-sequence comparison [1, 2] to pre-screen for potential homologs. This could enable HHsearch to reach the speed of PSI-BLAST at much higher sensitivity.

Acknowledgements

I would like to thank Andreas Biegert and Michael Remmert for their invaluable help in setting up the web servers. I am indebted to Andrei Lupas for his advice in the design of the servers, as well as for his constant support. Last, I thank all users who gave us feedback to improve our software.

References

[1] Pearson, W. R. and Lipman, D. J. (1988) Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. U. S. A.,* **85**, 2444–2448.

[2] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990) Basic local alignment search tool. *J Mol Biol,* **215**, 403–410.

[3] 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 Acid Research,* **25**, 3389–3402.

[4] Pietrokovski, S. (1996) Searching databases of conserved sequence regions by aligning protein multiple-alignments. *Nucleic Acids Res.,* **24**, 3836–3845.

[5] Yona, G. and Levitt, M. (2002) Within the twilight zone: a sensitive profile–profile comparison tool based on information theory. *J. Mol. Biol.,* **315**, 1257–1275.

[6] Sadreyev, R. I. and Grishin, N. V. (2003) COMPASS: A tool for comparison of multiple protein alignments with assessment of statistical significance. *J. Mol. Biol.,* **326**, 317–336.

[7] Rychlewski, L., Zhang, B., and Godzik, A. (1998) Fold and function predictions for Mycoplasma genitalium proteins. *Fold Des,* **3**, 229–238.
[8] vonÖhsen, N., Sommer, I., and Zimmer, R. (2003) Profile–profile alignment: a powerful tool for protein structure prediction. *Pac. Symp. Biocomput.,* pp. 252–263.
[9] Panchenko, A. R. (2003) Finding weak similarities between proteins by sequence profile comparison. *Nucleic Acids Res.,* **31**, 683–689.
[10] Fischer, D. (2003) 3D-SHOTGUN: a novel, cooperative, fold-recognition meta-predictor. *Proteins,* **51**, 434–441.
[11] Ginalski, K., Pas, J., Wyrwicz, L. S., vonGrotthus, M., Bujnicki, J. M., and Rychlewski, L. (2003) ORFeus: detection of distant homology using sequence profiles and predicted secondary structure. *Nucleic Acid Research,* **31**, 3804–3807.
[12] Ginalski, K., Elofsson, A., Fischer, D., and Rychlewski, L. (2003) 3D-Jury: a simple approach to improve protein structure predictions. *Bioinformatics,* **19**, 1015–1018.
[13] Tang, C. L., Xie, L., Koh, I. Y., Posy, S., Alexov, E., and Honig, B. (2003) On the role of structural information in remote homology detection and sequence alignment: new methods using hybrid sequence profiles. *J. Mol. Biol.,* **334**, 1043–1062.
[14] Ginalski, K., Rychlewski, L., Baker, D., and Grishin, N. V. (2004) Protein structure prediction for the male-specific region of the human Y chromosome. *Proc Natl Acad Sci USA,* **101**, 2305–2310.
[15] Pawlak, S. D., Radlinska, M., Chmiel, A. A., Bujnicki, J. M., and Skowronek, K. J. (2005) Inference of relationships in the 'twilight zone' of homology using a combination of bioinformatics and site-directed mutagenesis: a case study of restriction endonucleases Bsp6I and PvuII. *Nucleic Acids Res,* **33**, 661–671.
[16] Kihara, D. and Skolnick, J. (2004) Microbial genomes have over 72% structure assignment by the threading algorithm PROSPECTOR_Q. *Proteins,* **55**, 464–473.
[17] Todd, A. E., Orengo, C. A., and Thornton, J. M. (2001) Evolution of function in protein superfamilies, from a structural perspective. *J. Mol. Biol.,* **307**, 1113–1143.
[18] Kinch, L. and Grishin, N. (2002) Evolution of protein structures and functions. *Curr. Opin. Struct. Biol.,* **12**, 400–408.
[19] Fischer, D., Rychlewski, L., Dunbrack, R. L. J., Ortiz, A. R., and Elofsson, A. (2003) CAFASP3: the third critical assessment of fully automated structure prediction methods. *Proteins,* **53**, 503–516.
[20] Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. (1998) Biological sequence analysis: Probabilistic models of proteins and nucleic acids, Cambridge University Press, Cambridge, .
[21] Krogh, A., Brown, M., Mian, I. S., Sjölander, K., and Haussler, D. (1994) Hidden markov models in computational biology. Applications to protein modeling. *J. Mol. Biol.,* **235**, 1501–1531.
[22] Eddy, S. R. (1998) Profile hidden markov models. *Bioinformatics,* **14**, 755–763.
[23] Karplus, K., Karchin, R., Barrett, C., Tu, S., Cline, M., Diekhans, M., Grate, L., Casper, J., and Hughey, R. (2001) What is the value added by human intervention in protein structure prediction. *Proteins,* **45**, 86–91.
[24] Lyngsø, R. B., Pedersen, C. N. S., and Nielsen, H. (1999) Metrics and similarity measures for hidden markov models. *Proc. Int. Conf. Intell. Syst. Mol. Biol.,* pp. 178–186.
[25] Söding, J. (2005) Protein homology detection by HMM-HMM comparison. *Bioinformatics,* **21**, 951–960.
[26] Doolittle, R. F. (1981) Similar amino acid sequences: chance or common ancestry. *Science,* **214**, 149–159.
[27] Murzin, A. G., Brenner, S. E., Hubbard, T., and Chothia, C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.,* **247**, 536–540.
[28] Venclovas, C. (2003) Comparative modeling in CASP5: progress is evident, but alignment errors remain a significant hindrance. *Proteins,* **53**, 380–388.

[29] Siew, N., Elofsson, A., Rychlewski, L., and Fischer, D. (2000) MaxSub: an automated measure for the assessment of protein structure prediction quality. *Bioinformatics,* **16**, 776–85.

[30] Kabsch, W. and Sander, C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers,* **22**, 2577–2637.

[31] Sonnhammer, E. L., Eddy, S. R., Birney, E., Bateman, A., and Durbin, R. (1998) Pfam: multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Res,* **26**, 320–322.

[32] Ponting, C. P., Schultz, J., Milpetz, F., and Bork, P. (1999) SMART: identification and annotation of domains from signalling and extracellular protein sequences. *Nucleic Acids Res,* **27**, 229–232.

[33] Tatusov, R. L., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Kiryutin, B., Koonin, E. V., Krylov, D. M., Mazumder, R., Mekhedov, S. L., Nikolskaya, A. N., Rao, B. S., Smirnov, S., Sverdlov, A. V., Vasudevan, S., Wolf, Y. I., Yin, J. J., and Natale, D. A. (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics,* **4**, 41–41.

[34] Marchler-Bauer, A., Panchenko, A., Shoemaker, B., Thiessen, P., Geer, L., and Bryant, S. (2002) CDD: a database of conserved domain alignments with links to domain three-dimensional structure. *Nucleic Acids Res.,* **30**, 281–283.

[35] Mulder, N. J., Apweiler, R., and /it et al., A. (2005) InterPro, progress and status in 2005. *Nucleic. Acids. Res.,* **33**, D201–D205.

[36] Haft, D. H., Selengut, J. D., and White, O. (2003) The TIGRFAMs database of protein families. *Nucleic. Acids. Res.,* **31**, 371–373.

[37] Mi, H., Lazareva-Ulitsky, B., Loo, R., Kejariwal, A., Vandergriff, J., Rabkin, S., Guo, N., Muruganujan, A., Doremieux, O., Campbell, M. J., Kitano, H., and Thomas, P. D. (2005) The PANTHER database of protein families, subfamilies, functions and pathways. *Nucleic. Acids. Res.,* **33**, D284–D288.

[38] Wu, C. H., Nikolskaya, A., and Huang, H. *et al.*. (2004) PIRSF: family classification system at the Protein Information Resource. *Nucleic. Acids. Res.,* **32**, D112–D114.

[39] Pearl, F., Todd, A., and Sillitoe, I. *et al.*. (2005) The CATH Domain Structure Database and related resources Gene3D and DHS provide comprehensive domain family information for genome analysis. *Nucleic. Acids. Res.,* **33**, D247–D251.

[40] Söding, J., Remmert, M., Biegert, A., and Lupas, A. N. (2006) HHsenser: detection of remotely homologous protein sequences by intermediate profile search and HMM-HMM comparison. *Nucleic Acids Res.,* **34**, in press.

[41] Kosinski, J., Feder, M., and Bujnicki, J. M. (2005) The PD-(D/E)XK superfamily revisited: identification of new members among proteins involved in DNA metabolism and functional predictions for domains of (hitherto) unknown function. *BMC Bioinformatics,* **6**, 172.

[42] Suhre, K. (2005) Gene and genome duplication in Acanthamoeba polyphaga Mimivirus. *J. Virology,* **79**, 14095–14101.

[43] Liu, J., Glazko, G., and Mushegian, A. (2006) Protein repertoire of double-stranded DNA bacteriophages. *Virus Res.,* **117**, 68–80.

[44] Ammelburg, M., Frickey, T., and Lupas, A. N. (2006) Classification of AAA+ proteins. *J. Struc. Biol.,* in press.

[45] Djuranovic, S., Rockel, B., Lupas, A. N., and Martin, J. (2006) Characterization of AMA, a new AAA protein from Archaeoglobus and methanogenic archaea. *J. Struc. Biol.,* in press.

[46] Diemand, A. and Lupas, A. N. (2006) Modeling AAA+ ring complexes from monomeric structures. *J. Struc. Biol.,* in press.

[47] Becker, E., Meyer, V., Madaoui, H., and Guerois, R. (2006) Detection of a tandem BRCT in Nbs1 and Xrs2 with functional implications in the DNA damage response. *Bioinformatics,* **22**, 1289–1292.

[48] Tkaczuk, K. L., Obarska, A., and Bujnicki, J. M. (2006) Molecular phylogenetics and comparative modeling of HEN1, a methyltransferase involved in plant microRNA biogenesis. *BMC Evo. Biol.,* **6**, 6.

[49] Boekhorst, J., Helmer, Q., Kleerebezem, M., and Siezen, R. J. (2006) Comparative analysis of proteins with a mucus-binding domain found exclusively in lactic acid bacteria. *Microbiology-SGM,* **152**, 273–280.

[50] Albrecht, R., Zeth, K., Söding, J., Lupas, A. N., and Linke, D. (2006) Expression, crystallization and preliminary X-ray crystallographic studies of the outer membrane protein OmpW from Escherichia coli. *Acta. Crystallograph. Sect. F. Struct. Biol. Cryst. Commun.,* **62**, 415–418.

[51] Devos, D., Dokudovskaya, S., Williams, R., Alber, F., Eswar, N., Chait, B. T., Rout, M. P., and Sali, A. (2006) Simple fold composition and modular architecture of the nuclear pore complex. *PNAS,* **103**, 2172–2177.

[52] Dokudovskaya, S., Williams, R., Devos, D., Sali, A., Chait, B. T., and Rout, M. P. (2006) Protease accessibility laddering: a proteomic tool for probing protein structure. *Structure,* **14**, 653–660.

[53] Neugebauer, H., Herrmann, C., Kammer, W., Schwarz, G., Nordheim, A., and Braun, V. (2005) ExbBD-dependent transport of maltodextrins through the novel MalA protein across the outer membrane of Caulobacter crescentus. *J. Bact.,* **187**, 8300–8311.

[54] Minakhin, L., Semenova, E., Liu, J., Vasilov, A., Severinova, E., Gabisonia, T., Inman, R., Mushegian, A., and Severinov, K. (2005) Genome sequence and gene expression of Bacillus anthracis bacteriophage Fah. *J. Mol. Biol.,* **354**, 1–15.

[55] Gibson, A., Lewis, A. P., Affleck, K., Aitken, A. J., Meldrum, E., and Thompson, N. (2005) hCLCA1 and mCLCA3 are secreted non-integral membrane proteins and therefore are not ion channels. *J. Biol. Chem.,* **280**, 27205–27212.

[56] Jin, J., Cai, Y., Yao, T., Gottschalk, A. J., Florens, L., Swanson, S. K., Gutierrez, J. L., Coleman, M. K., Workman, J. L., Mushegian, A., Washburn, M. P., Conaway, R. C., and Conaway, J. W. (2005) A mammalian chromatin remodeling complex with similarities to the yeast INO80 complex. *J. Biol. Chem.,* **280**, 41207–41212.

[57] Tilburn, J., Sanchez-Ferrero, J. C., Reoyo, E., Arst, H. N., and Penalva, M. A. (2005) Mutational analysis of the pH signal transduction component PalC of Aspergillus nidulans supports distant similarity to BRO1 domain family members. *Genetics,* **171**, 393–401.

[58] Chatterjee, I., Richmond, A., Putiri, E., Shakes, D. C., and Singson, A. (2005) The Caenorhabditis elegans spe-38 gene encodes a novel four-pass integral membrane protein required for sperm function at fertilization. *Development,* **132**, 2795–2808.

[59] 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.

[60] Moussian, B., Söding, J., Schwarz, H., and Nüsslein-Volhard, C. (2005) Retroactive, a membrane-anchored extracellular protein related to vertebrate snake neurotoxin-like proteins, is required for cuticle organization in the larva of Drosophila melanogaster. *Dev. Dyn.,* **233**, 1056–1063.

[61] Alva Kullanja, V., Ammelburg, M., Söding, J., and Lupas, A. N. (2006) The origin of the histone fold. In preparation.

[62] Salem, G. M., Hutchinson, E. G., Orengo, C. A., and Thornton, J. M. (1999) Correlation of observed fold frequency with the occurrence of local structural motifs. *J. Mol. Biol.,* **287**, 969–981.

[63] Söding, J. and Lupas, A. N. (2003) More than the sum of their parts: on the evolution of proteins from peptides. *Bioessays,* **25**, 837–846.

[64] Söding, J., Remmert, M., and Biegert (2006) HHrep: *de novo* protein repeat detection and the origin of TIM barrels. *Nucleic Acids Res.,* **34**, in press.

[65] Lang, D., Thoma, R., Henn-Sax, M., Sterner, R., and Wilmanns, M. (2000) Structural evidence for evolution of the β/α barrel scaffold by gene duplication and fusion. *Science,* **289**, 1546–1550.

[66] Biegert, A., Remmert, M., Söding, J., and Lupas, A. N. (2006) The MPI Bioinformatics Toolkit for protein sequence analysis. *Nucleic Acids Res.,* **34**, in press.