W410–W415 *Nucleic Acids Research, 2016, Vol. 44, Web Server issue* *Published online 29 April 2016*
doi: 10.1093/nar/gkw348

The MPI bioinformatics Toolkit as an integrative platform for advanced protein sequence and structure analysis

Vikram Alva[1,*], Seung-Zin Nam[1], Johannes Söding[2] and Andrei N. Lupas[1,*]

[1]Department of Protein Evolution, Max Planck Institute for Developmental Biology, Tübingen D-72076, Germany and [2]Group for Quantitative and Computational Biology, Max Planck Institute for Biophysical Chemistry, Göttingen D-37077, Germany

Received February 14, 2016; Revised April 08, 2016; Accepted April 19, 2016

ABSTRACT

The MPI Bioinformatics Toolkit (http://toolkit.tuebingen.mpg.de) is an open, interactive web service for comprehensive and collaborative protein bioinformatic analysis. It offers a wide array of interconnected, state-of-the-art bioinformatics tools to experts and non-experts alike, developed both externally (e.g. BLAST+, HMMER3, MUSCLE) and internally (e.g. HHpred, HHblits, PCOILS). While a beta version of the Toolkit was released 10 years ago, the current production-level release has been available since 2008 and has serviced more than 1.6 million external user queries. The usage of the Toolkit has continued to increase linearly over the years, reaching more than 400 000 queries in 2015. In fact, through the breadth of its tools and their tight interconnection, the Toolkit has become an excellent platform for experimental scientists as well as a useful resource for teaching bioinformatic inquiry to students in the life sciences. In this article, we report on the evolution of the Toolkit over the last ten years, focusing on the expansion of the tool repertoire (e.g. CS-BLAST, HHblits) and on infrastructural work needed to remain operative in a changing web environment.

INTRODUCTION

Over the last two decades, bioinformatic analysis of proteins has become central to molecular biology research. In fact, several new stand-alone tools as well as specialized web services are developed and published every month. However, the large majority are hard to install and/or use, and therefore remain out of reach for most bench biologists. To alleviate this problem, a number of easy-to-use web services have been developed, including the NCBI Resources (http://www.ncbi.nlm.nih.gov/guide/all), the SIB Bioinformatics Resource Portal (ExPASy; http://www.expasy.org), the EMBL-EBI Bioinformatics Web Services [(1); http://www.ebi.ac.uk/services], the Protein Analysis Toolkit [PAT (2); http://pat.cbs.cnrs.fr], the PredictProtein server [(3); https://www.predictprotein.org] and the CBS Prediction Servers (http://www.cbs.dtu.dk/services). Motivated by the work of our department in both computational and experimental biology, we wished to provide our colleagues at the bench with access to cutting-edge bioinformatics tools; we therefore developed a simple web-based system that combined the most useful external and internal tools. This evolved into the MPI Bioinformatics Toolkit, which we made public in a beta version in 2005 with 30 tools and published in the 2006 Web Server issue of Nucleic Acids Research (4). The current production-level release was launched in 2008 with 37 tools and has been up and running reliably ever since, servicing over 1.6 million external queries. Over this time period, the number of queries, as well as of citations of the Toolkit and the tools developed by us, have increased fairly linearly from year to year (Figure 1).

Although the Toolkit has become an essential resource to experimental scientists mainly through its state-of-the-art remote homology detection tool HHpred (5), which is currently accessed more than 1000 times a day, its tools are in fact used broadly, with 18 of 53 called up more than five times a day on average (Tables 1 and 2). These highly accessed tools include both internal developments, such as CS-BLAST (6), HHblits (7) and PCOILS (8), and external tools, e.g. BLASTClust (9), Modeller (10) and MARCOIL (11) (Table 2). Interestingly, our 6FrameTranslation tool, which translates a given nucleotide sequence into the six possible frames, is also among the most accessed tools, suggesting that the Toolkit is not just a platform for advanced bioinformatic analyses, e.g. with HHpred, but that it has also grown into a general bioinformatics resource.

*To whom correspondence should be addressed. Tel: +49 7071 601 341; Fax: +49 7071 601 352; Email: andrei.lupas@tuebingen.mpg.de
Correspondence may also be addressed to Vikram Alva. Tel: +49 7071 601 451; Fax: +49 7071 601 352; Email: vikram.alva@tuebingen.mpg.de

Table 1. An overview of tools available in the Toolkit

Category	Tools
Search	**CS-BLAST** (6), **HHblits** (7), **HHpred** (5), **HHsenser** (45), HMMER3 (46), **PatternSearch**, ProtBLAST (9), ProtBLAST+ (15), PSI-BLAST (14), PSI-BLAST+ (15), SimShiftDB (47), **PDBalert** (37)
Alignment	AlignmentViewer, **Blammer**, Clustal Omega (38), GLProbs (42), **HHalign**, Kalign (39), MAFFT (48), MSAProbs (40), MUSCLE (49), ProbCons (50), TCoffee (41)
Sequence Analysis	**Aln2Plot**, **COILS/PCOILS** (8), **FRpred** (51), **HHrep** (25), **HHrepID** (13), MARCOIL (11), **REPPER** (52), **TPRpred** (53)
Secondary Structure	**Ali2D**, **HHomp** (30), **Quick2D**
Tertiary Structure	**bFit** (54), **HHfrag** (55), **HHpred** (5), Modeller (10), **SamCC** (31)
Classification	ANCESCON (12), BLASTClust (9), **CLANS** (56), **ClubSub-P** (57), **daTAA** (28), **GCView** (58), **HHcluster** (27), PHYLIP-NEIGHBOR (59)
Utilities	**6FrameTranslation**, **Backtranslator**, **Extract GIs**, **GI2Promoter**, **HHfilter**, **Reformat**, **RetrieveSeq**

The categories listed here correspond to the section tabs in the menu bar located at the top of the Toolkit page. Section-specific tools are listed in a submenu within each tab. All tools can also be accessed through the tool index displayed on the homepage of the Toolkit. Tools developed by us are shown in boldface and tools added since our last publication on the Toolkit in 2006 are underlined.

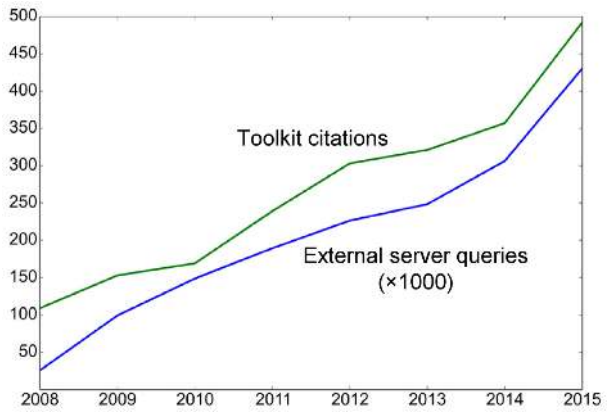

Figure 1. Toolkit usage and citations. The number of queries from external IP addresses in thousands is shown by the blue line and the citations of the Toolkit framework and the tools developed by us, as indicated by Google Scholar, is shown by the green line.

Table 2. Tools used more than five times a day on average: usage in 2008 and 2015

Tool	2008	2015
Ali2D	136	2620
AlignmentViewer	339	4997
BLASTClust	980	6531
Clustal Omega	575	3416
CS-BLAST	67	2063
HHalign	205	1906
HHblits	–	4353
HHpred	12562	328973
MAFFT	203	3637
MARCOIL	–	4289
Modeller	2401	20134
MUSCLE	414	2706
PCOILS	584	5542
ProbCons	245	2343
PSI-BLAST	1302	3049
Quick2D	811	5783
6FrameTranslation	328	5298
TPRpred	364	4350
Total (all 53 tools)	**25 611**	**430 296**

Tools developed in our group are shown in boldface and tools added since our last publication on the Toolkit in 2006 are underlined.

KEY FEATURES

Our primary motivation behind developing and maintaining the Toolkit has always been the desire to provide experimental biologists (starting with our colleagues in our own department) with a simple web-based, one-stop platform that integrates a limited number of highly useful bioinformatic tools for the analysis of protein sequences and structures. In our opinion, the following features make our Toolkit useful to experts and non-experts alike:

Ease of use

We offer easy, web-based access to a number of tools that are otherwise only accessible from the command line and are often hard to install and get to work for a non-expert user [e.g. BLASTClust (9), ANCESCON (12), HHpred (5) or HHrepID (13)]. For many external tools, we also offer enhanced functionalities; for instance, our BLAST tools allow searches against nonstandard databases, such as the nonredundant (nr) protein sequence database clustered down to a pairwise sequence identity of 90% (nr90) or 70% (nr70), or personal databases uploaded by the user. Further, our implementations of PSI-BLAST (14) and PSI-BLAST+ (15) allow users to change the database between iterations. Using this feature, users can, for instance, train a profile on a large database (e.g. nr70) and then search for all homologs of the query protein in a specific genome or in a database uploaded by the user, with a much higher sensitivity than would be available if the profile had been trained only on the genome or the user database in question.

A further important feature of our Toolkit is the tight interconnection between most tools on offer, allowing the results of one tool to be forwarded as input to several others. The users could, for instance, start a sequence search with a protein of interest against a database of choice, using the sensitive search tool CS-BLAST, and then forward the results to Blammer to parse out a multiple alignment of the obtained sequence hits (Figure 2). This multiple alignment could then be forwarded further to Quick2D, to obtain an overview of secondary structure features such as α-helices,

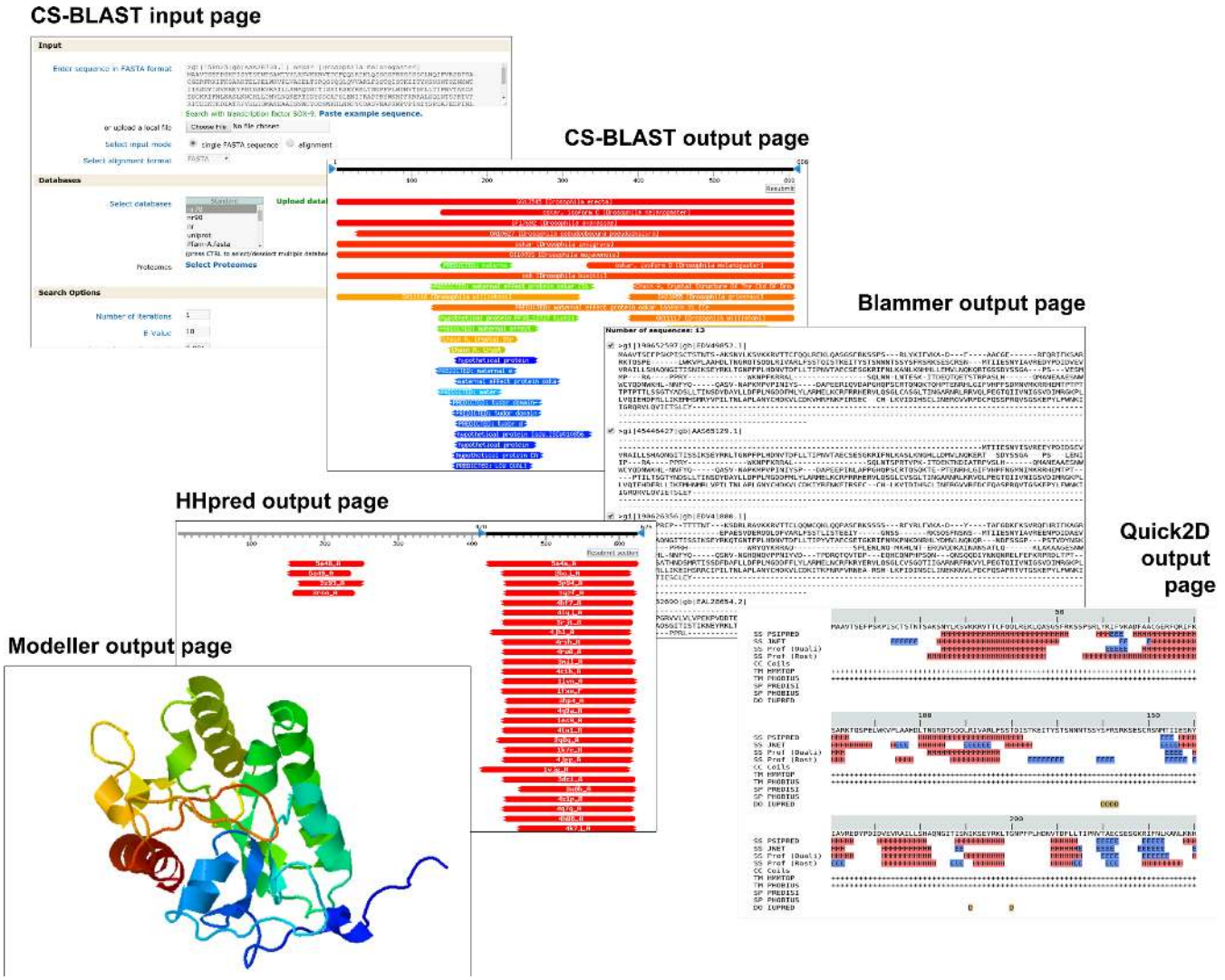

Figure 2. Interconnection of the tools in the Toolkit. The output of most tools can be forwarded as input to many other tools. One such possible forwarding pipeline is shown, wherein the output of the sensitive search method CS-BLAST is forwarded to Blammer to parse out a multiple alignment, which is subsequently forwarded to Quick2D for secondary structure prediction and HHpred for the identification of remote homologs. The output of HHpred is then forwarded to Modeller in order to obtain a structural model.

β-strands, coiled coils, transmembrane helices and intrinsically disordered regions. In parallel, the alignment could be forwarded to HHpred in order to detect distant homologs of known structure. Subsequently, the hits obtained from HHpred could be forwarded to Modeller (10) as templates for comparative modelling, resulting in a structural model for the protein of interest. This tight interconnection of the tools thus allows for complex bioinformatic analyses in a simple and straightforward manner, starting with just a single protein sequence.

High quality, up-to-date tools and databases

Since we rely heavily on the Toolkit for our own research into the structure, function and evolution of proteins, we have been able to maintain its tools and databases at a high level and to detect and fix bugs rapidly. We update the databases regularly and ensure that the various internal and external tools are up-to-date as well. In addition to standard databases, such as the nonredundant protein sequence database (nr), the Protein Data Bank [PDB; (16)], the Pfam database (17) and the UniProt database (18) (including their variants, such as nr70 or nr90), we provide the databases of profile HMMs needed for HHblits [UniProt20] and HHpred [e.g. PDB70, SCOPe95 (19), Pfam (17), CDD (20), representative proteomes] and also allow users to upload their own sequence databases. We strive to react to bug reports and update/feature requests sent to us by our users in a timely manner and are currently engaged in revising and expanding our help pages.

Job management and personal workspace

One of the main design goals during the development of the Toolkit was to provide users with easy access to their jobs and the possibility to share them, in order to facilitate

collaborative research. In line with this, every job submitted to the Toolkit is assigned either an automatic or a user-specified identifier, which upon submission of the job is displayed along with its current status in a sidebar on the left side of the browser. Users can click on previously submitted jobs to check their current status, get to the results page or return to an earlier job. Furthermore, users can share these job identifiers with their collaborators, allowing them to see the same output page as the user. For uploading custom sequence databases and to preserve jobs for a longer period of time, we offer users the possibility to create a personal account; their jobs are then stored for two months, rather than for two weeks, as without a log-in. Jobs in a personal account are private and cannot be viewed by other users. For a more detailed description of job submission and management, please refer to our previous paper on the Toolkit (4).

Links to external resources

Complementary to the tools offered in the Toolkit, we collate links to external tools that we think are particularly useful; these links are found on the front pages of the individual sections. For example, in the 'Sequence Analysis' section, we provide links to the function prediction servers The Seed (21) and String (22), and to the de-novo repeat detections servers RADAR (23) and TRUST (24). In this we emulate other highly used bioinformatic platforms such as ExPASy.

NEW TOOLS

Since our previous article in 2006, the Toolkit has grown from 30 to 53 tools (Table 1), more than half of which were developed internally. New developments concern sensitive sequence searching, address the classification of domains or are structure-based tools.

Sensitive sequence comparison tools

We have included two new sequence comparisons tools, CS-BLAST (6) and HHblits (7). CS-BLAST is a BLAST-like tool that gains sensitivity by including context-specific pseudocounts and finds twice as many homologs as BLAST at the same error rate and a comparable runtime (6). This tool can also be used iteratively and two iterations of it are typically more sensitive than five iterations of PSI-BLAST (6). HHblits is a remote homology detection tool based on iterative HMM–HMM comparison. In the first step, it converts the input sequence or multiple sequence alignment (MSA) to a profile HMM, which it then uses to iteratively search through profile HMMs in the UniProt20 database, employing an algorithm similar to the one used by HHsearch. Target sequences found to be significantly similar in each iteration are added to the query profile HMM for the next iteration. Compared to PSI-BLAST, HHblits is twice as sensitive, faster and produces alignments that are more accurate (7). We therefore now use HHblits as the preferred method for the MSA generation steps of HHpred (5), HHrep (25) and HHrepID (13). The latter is also a new addition to the Toolkit, built for the *de novo* detection of highly divergent repeats based on profile HMM comparison. We have previously used this tool to detect evidence for the homology of structural repeats in outer membrane proteins [OMPs; (26)] and TIM barrels (13).

Domain annotation/classification tools

Over the last years, we have developed and included further classification tools into the Toolkit. One, HHcluster, allows users to explore homologous connections between superfamilies with different structures in our galaxy of folds, which is a two-dimensional map of sequence relationships in protein fold space (27). We constructed this map by performing pairwise HMM-HMM comparisons for all domains in the SCOP database filtered to a maximum of 20% sequence identity and subsequently clustering them by a force-directed procedure using the statistical significance of these comparisons.

Two other tools address the detection of domains belonging to specific superfamilies. daTAA (28) provides a platform for the annotation of trimeric autotransporter adhesins (TAAs), an important family of pathogenic determinants in Gram-negative bacteria. TAAs present special challenges for automated domain annotation due to their high sequence diversity, mosaic-like arrangement of constituent domains, fuzzy domain boundaries and the frequent presence of extended regions of low sequence complexity, some of which we recognized as compositionally unusual coiled coils (29). daTAA meets these challenges through a combination of knowledge-based rules and HMM-based sequence analyses against manually curated alignments. The second, HHomp (30), is a tool for the prediction and classification of outer membrane proteins (OMPs), which are a major component of the outer membranes of Gram-negative bacteria, mitochondria and plastids. The transmembrane domains of OMPs comprise 4–12 β-hairpins that organize themselves around a central pore to form a β-barrel. We have previously shown that the β-barrels of all bacterial OMPs share a common ancestry and that they may have evolved by amplification of a single, ancestral β-hairpin (26). HHomp exploits this evolutionary observation; for a given input sequence, it builds a profile HMM and compares it with a database containing profile HMMs for ∼20 000 OMPs, in order to detect and classify new members.

Structure-based tools

We have also extended the repertoire of structure-based tools: (i) SamCC (31) measures the local structural parameters of parallel and antiparallel four-helical bundles, and compares these with the ideal values of four-helical coiled coils. We developed it in order to quantify departures from the ideal state and thus make variants of one domain comparable to each other in a quantitative way. Based on SamCC analyses of HAMP domains, we proposed a model for transmembrane signal transduction in TCST receptors (32–34). (ii) Ali2D is a tool that annotates multiple sequence alignments. It accepts MSAs as input, predicts the secondary structure of the constituent sequences with PSIPRED (35) and their membrane propensity with MEMSAT2 (36), and maps the results onto the MSA. This gives a

consensus overview of secondary structure and membrane insertion in a given protein family, and alerts the user to potentially misaligned regions. Ali2D has become quite popular and was among the top third most accessed tools last year (Table 2). (iii) PDBalert (37), finally, is a tool that notifies users of the availability of PDB structures (released or on hold) with homology to a given protein of interest. This tool is only accessible from a personal user account (see above).

External tools

Of the external tools added to the Toolkit in recent years, most are multiple protein sequence alignment methods and include Clustal Omega (38), Kalign (39), MSAProbs (40), TCoffee (41) and GLProbs (42). Other newly incorporated external tools are the NCBI tools BLAST+ and PSI-BLAST+ (15) and the coiled coil detection tool MARCOIL (11).

TEACHING

In addition to establishing itself as a resource for protein bioinformatic analysis, the Toolkit has also become a useful platform for teaching bioinformatic enquiry to students in the life sciences. Due to its broad array of tools and their tight interconnection, its simple web interface, and its intuitive job management features that allow to pre-compute and share jobs, the Toolkit empowers students to efficiently progress to the scientific aspects of bioinformatic analysis, without the need to install programs and learn how to connect these with scripts. We ourselves use it as a primary resource to teach the 'Bioinformatics for Biochemists' practical course at the University of Tübingen and the graduate students of our institute. We are currently striving to make it more attractive for teaching purposes by including more detailed help pages and tutorials.

OUTLOOK

The growing use of the Toolkit gives us confidence that providing easy access to state-of-the-art bioinformatic tools will remain an important endeavor. In order to continue this and meet the software challenges of the next decade, our current focus is on replacing Java Applets with JavaScript-based solutions, to ensure the usability of our Toolkit on all different browsers. For instance, we now use JSmol (43), a JavaScript-based molecular viewer, to display protein structures, the BioJS MSA viewer (44) to display multiple sequence alignments, and the BioJS Tree viewer (44) to display phylogenetic trees. We are currently making the transition to accession.version identifiers for tools that use sequence GI identifiers, because NCBI is phasing out GIs this September. As mentioned earlier, Toolkit usage has grown linearly over the years, passing the 400 000 mark last year. This year we expect to cross the 500 000 mark and in anticipation of this and further growth in the future, we are upgrading our computational resources and are migrating to a more scalable architecture.

ACKNOWLEDGEMENTS

We would like to thank Andreas Biegert, Alexander Diemand, Klaus Faidt, Klaus O. Kopec, Jörn Marialke, Christian Mayer, Markus Meier, Andre Noll, Michael Remmert, Tina Streich, Christina Wassermann and Johannes Wörner for their contributions to the development and maintenance of the Toolkit over the years, as well as our current undergraduate students working on the Toolkit: Andrew Stephens, Jonas Kübler and Lukas Zimmermann. We would also like to thank all our users and members of our department for helping us to improve the Toolkit through their bug reports and feature requests. AL gratefully acknowledges Kristin K. Brown (GlaxoSmithKline) for many discussions, particularly in the early stages of the Toolkit.

FUNDING

Institutional funds of the Max Planck Society; German Federal Ministry of Education and Research (BMBF) (to J.S.) within the framework of e:Med [e:AtheroSysMed, 01ZX1313A-2014] and e:bio [SysCore]. Funding for open access charge: Institutional funds of the Max Planck Society.

Conflict of interest statement. None declared.

REFERENCES

1. Li,W., Cowley,A., Uludag,M., Gur,T., McWilliam,H., Squizzato,S., Park,Y.M., Buso,N. and Lopez,R. (2015) The EMBL-EBI bioinformatics web and programmatic tools framework. *Nucleic Acids Res.*, **43**, W580–W584.
2. Gracy,J. and Chiche,L. (2005) PAT: a protein analysis toolkit for integrated biocomputing on the web. *Nucleic Acids Res.*, **33**, W65–W71.
3. Yachdav,G., Kloppmann,E., Kajan,L., Hecht,M., Goldberg,T., Hamp,T., Honigschmid,P., Schafferhans,A., Roos,M., Bernhofer,M. *et al.* (2014) PredictProtein–an open resource for online prediction of protein structural and functional features. *Nucleic Acids Res.*, **42**, W337–W343.
4. Biegert,A., Mayer,C., Remmert,M., Soding,J. and Lupas,A.N. (2006) The MPI Bioinformatics Toolkit for protein sequence analysis. *Nucleic Acids Res.*, **34**, W335–W339.
5. Soding,J., Biegert,A. and Lupas,A.N. (2005) The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res.*, **33**, W244–W248.
6. Biegert,A. and Soding,J. (2009) Sequence context-specific profiles for homology searching. *Proc. Natl. Acad. Sci. U.S.A.*, **106**, 3770–3775.
7. Remmert,M., Biegert,A., Hauser,A. and Soding,J. (2012) HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods*, **9**, 173–175.
8. Gruber,M., Soding,J. and Lupas,A.N. (2006) Comparative analysis of coiled-coil prediction methods. *J. Struct. Biol.*, **155**, 140–145.
9. 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.
10. Sali,A. and Blundell,T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.*, **234**, 779–815.
11. Delorenzi,M. and Speed,T. (2002) An HMM model for coiled-coil domains and a comparison with PSSM-based predictions. *Bioinformatics*, **18**, 617–625.
12. Cai,W., Pei,J. and Grishin,N.V. (2004) Reconstruction of ancestral protein sequences and its applications. *BMC Evol. Biol.*, **4**, 33.
13. Biegert,A. and Soding,J. (2008) De novo identification of highly diverged protein repeats by probabilistic consistency. *Bioinformatics*, **24**, 807–814.
14. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

15. Camacho,C., Coulouris,G., Avagyan,V., Ma,N., Papadopoulos,J., Bealer,K. and Madden,T.L. (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.
16. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
17. Finn,R.D., Coggill,P., Eberhardt,R.Y., Eddy,S.R., Mistry,J., Mitchell,A.L., Potter,S.C., Punta,M., Qureshi,M., Sangrador-Vegas,A. *et al.* (2016) The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.*, **44**, D279–D285.
18. UniProt Consortium. (2015) UniProt: a hub for protein information. *Nucleic Acids Res.*, **43**, D204–D212.
19. Fox,N.K., Brenner,S.E. and Chandonia,J.M. (2014) SCOPe: Structural Classification of Proteins–extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Res.*, **42**, D304–D309.
20. Marchler-Bauer,A., Derbyshire,M.K., Gonzales,N.R., Lu,S., Chitsaz,F., Geer,L.Y., Geer,R.C., He,J., Gwadz,M., Hurwitz,D.I. *et al.* (2015) CDD: NCBI's conserved domain database. *Nucleic Acids Res.*, **43**, D222–D226.
21. Overbeek,R., Begley,T., Butler,R.M., Choudhuri,J.V., Chuang,H.Y., Cohoon,M., de Crecy-Lagard,V., Diaz,N., Disz,T., Edwards,R. *et al.* (2005) The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res.*, **33**, 5691–5702.
22. Szklarczyk,D., Franceschini,A., Wyder,S., Forslund,K., Heller,D., Huerta-Cepas,J., Simonovic,M., Roth,A., Santos,A., Tsafou,K.P. *et al.* (2015) STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.*, **43**, D447–D452.
23. Heger,A. and Holm,L. (2000) Rapid automatic detection and alignment of repeats in protein sequences. *Proteins*, **41**, 224–237.
24. Szklarczyk,R. and Heringa,J. (2004) Tracking repeats using significance and transitivity. *Bioinformatics*, **20**(Suppl. 1), i311–i317.
25. Soding,J., Remmert,M. and Biegert,A. (2006) HHrep: de novo protein repeat detection and the origin of TIM barrels. *Nucleic Acids Res.*, **34**, W137–W142.
26. Remmert,M., Biegert,A., Linke,D., Lupas,A.N. and Soding,J. (2010) Evolution of outer membrane beta-barrels from an ancestral beta beta hairpin. *Mol. Biol. Evol.*, **27**, 1348–1358.
27. Alva,V., Remmert,M., Biegert,A., Lupas,A.N. and Soding,J. (2010) A galaxy of folds. *Protein Sci.*, **19**, 124–130.
28. Szczesny,P. and Lupas,A. (2008) Domain annotation of trimeric autotransporter adhesins–daTAA. *Bioinformatics*, **24**, 1251–1256.
29. Bassler,J., Hernandez Alvarez,B., Hartmann,M.D. and Lupas,A.N. (2015) A domain dictionary of trimeric autotransporter adhesins. *Int. J. Med. Microbiol.*, **305**, 265–275.
30. Remmert,M., Linke,D., Lupas,A.N. and Soding,J. (2009) HHomp–prediction and classification of outer membrane proteins. *Nucleic Acids Res.*, **37**, W446–W451.
31. Dunin-Horkawicz,S. and Lupas,A.N. (2010) Measuring the conformational space of square four-helical bundles with the program samCC. *J. Struct. Biol.*, **170**, 226–235.
32. Ferris,H.U., Zeth,K., Hulko,M., Dunin-Horkawicz,S. and Lupas,A.N. (2014) Axial helix rotation as a mechanism for signal regulation inferred from the crystallographic analysis of the E. coli serine chemoreceptor. *J. Struct. Biol.*, **186**, 349–356.
33. Ferris,H.U., Dunin-Horkawicz,S., Hornig,N., Hulko,M., Martin,J., Schultz,J.E., Zeth,K., Lupas,A.N. and Coles,M. (2012) Mechanism of regulation of receptor histidine kinases. *Structure*, **20**, 56–66.
34. Ferris,H.U., Dunin-Horkawicz,S., Mondejar,L.G., Hulko,M., Hantke,K., Martin,J., Schultz,J.E., Zeth,K., Lupas,A.N. and Coles,M. (2011) The mechanisms of HAMP-mediated signaling in transmembrane receptors. *Structure*, **19**, 378–385.
35. Jones,D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.*, **292**, 195–202.
36. Nugent,T. and Jones,D.T. (2009) Transmembrane protein topology prediction using support vector machines. *BMC Bioinformatics*, **10**, 159.
37. Agarwal,V., Remmert,M., Biegert,A. and Soding,J. (2008) PDBalert: automatic, recurrent remote homology tracking and protein structure prediction. *BMC Struct. Biol.*, **8**, 51.
38. Sievers,F., Wilm,A., Dineen,D., Gibson,T.J., Karplus,K., Li,W., Lopez,R., McWilliam,H., Remmert,M., Soding,J. *et al.* (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.*, **7**, 539.
39. Lassmann,T. and Sonnhammer,E.L. (2005) Kalign–an accurate and fast multiple sequence alignment algorithm. *BMC Bioinformatics*, **6**, 298.
40. Liu,Y. and Schmidt,B. (2014) Multiple protein sequence alignment with MSAProbs. *Methods Mol. Biol.*, **1079**, 211–218.
41. Notredame,C., Higgins,D.G. and Heringa,J. (2000) T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.*, **302**, 205–217.
42. Ye,Y., Cheung,D.W., Wang,Y., Yiu,S.M., Zhan,Q., Lam,T.W. and Ting,H.F. (2015) GLProbs: aligning multiple sequences adaptively. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **12**, 67–78.
43. Hanson,R.M., Prilusky,J., Renjian,Z., Nakane,T. and Sussman,J.L. (2013) JSmol and the next-generation web-based representation of 3D molecular structure as applied to proteopedia. *Isr. J. Chem.*, **53**, 207–216.
44. Yachdav,G., Goldberg,T., Wilzbach,S., Dao,D., Shih,I., Choudhary,S., Crouch,S., Franz,M., Garcia,A., Garcia,L.J. *et al.* (2015) Anatomy of BioJS, an open source community for the life sciences. *Elife*, **4**, doi:10.7554/eLife.07009.
45. Soding,J., Remmert,M., Biegert,A. and Lupas,A.N. (2006) HHsenser: exhaustive transitive profile search using HMM-HMM comparison. *Nucleic Acids Res.*, **34**, W374–W378.
46. Eddy,S.R. (2011) Accelerated Profile HMM Searches. *PLoS Comput. Biol.*, **7**, e1002195.
47. Ginzinger,S.W. and Coles,M. (2009) SimShiftDB; local conformational restraints derived from chemical shift similarity searches on a large synthetic database. *J. Biomol. NMR*, **43**, 179–185.
48. Katoh,K., Misawa,K., Kuma,K. and Miyata,T. (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.*, **30**, 3059–3066.
49. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
50. Do,C.B., Mahabhashyam,M.S., Brudno,M. and Batzoglou,S. (2005) ProbCons: Probabilistic consistency-based multiple sequence alignment. *Genome Res.*, **15**, 330–340.
51. Fischer,J.D., Mayer,C.E. and Soding,J. (2008) Prediction of protein functional residues from sequence by probability density estimation. *Bioinformatics*, **24**, 613–620.
52. Gruber,M., Soding,J. and Lupas,A.N. (2005) REPPER–repeats and their periodicities in fibrous proteins. *Nucleic Acids Res.*, **33**, W239–W243.
53. Karpenahalli,M.R., Lupas,A.N. and Soding,J. (2007) TPRpred: a tool for prediction of TPR-, PPR- and SEL1-like repeats from protein sequences. *BMC Bioinformatics*, **8**, 2.
54. Mechelke,M. and Habeck,M. (2010) Robust probabilistic superposition and comparison of protein structures. *BMC Bioinformatics*, **11**, 363.
55. Kalev,I. and Habeck,M. (2011) HHfrag: HMM-based fragment detection using HHpred. *Bioinformatics*, **27**, 3110–3116.
56. Frickey,T. and Lupas,A. (2004) CLANS: a Java application for visualizing protein families based on pairwise similarity. *Bioinformatics*, **20**, 3702–3704.
57. Paramasivam,N. and Linke,D. (2011) ClubSub-P: cluster-based subcellular localization prediction for Gram-negative bacteria and archaea. *Front. Microbiol*, **2**, 218.
58. Grin,I. and Linke,D. (2011) GCView: the genomic context viewer for protein homology searches. *Nucleic Acids Res.*, **39**, W353–W356.
59. Felsenstein,J. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.*, **17**, 368–376.