Nucleic Acids Research, 2006, Vol. 34, Web Server issue ***W15–W19***
doi:10.1093/nar/gkl254

The MIGenAS integrated bioinformatics toolkit for web-based sequence analysis

Markus Rampp*, Thomas Soddemann and Hermann Lederer

Rechenzentrum Garching der Max-Planck-Gesellschaft (RZG), am Max-Planck-Institut für Plasmaphysik, Boltzmannstrasse 2, 85748 Garching, Germany

Received February 14, 2006; Revised March 13, 2006; Accepted March 30, 2006

ABSTRACT

We describe a versatile and extensible integrated bioinformatics toolkit for the analysis of biological sequences over the Internet. The web portal offers convenient interactive access to a growing pool of chainable bioinformatics software tools and databases that are centrally installed and maintained by the RZG. Currently, supported tasks comprise sequence similarity searches in public or user-supplied databases, computation and validation of multiple sequence alignments, phylogenetic analysis and protein–structure prediction. Individual tools can be seamlessly chained into pipelines allowing the user to conveniently process complex workflows without the necessity to take care of any format conversions or tedious parsing of intermediate results. The toolkit is part of the Max-Planck Integrated Gene Analysis System (MIGenAS) of the Max Planck Society available at www.migenas.org (click 'Start Toolkit').

INTRODUCTION

A large pool of individual websites offering convenient access to basic bioinformatics software and data have certainly greatly helped to establish many computational methods as standard tools in life sciences. Meanwhile, almost any newly published bioinformatics software package which is distributed for installation on PCs is supplemented by a web server (hosted by the software developers and/or provided for download and local installation) in order to enhance usability, attract and guide users, and to promote visibility of the software in the scientific community. NCBI's BLAST services are the prototypical example.

Advanced analysis, however, most often requires the concerted interoperation of different tools and heterogeneous data. Processing the corresponding workflows by consecutively visiting websites dispersed over the Internet is apparently very cumbersome, if not impracticable. Apart from a small subset of well-defined applications which are well supported by existing special purpose software [e.g. the ARB package for sequence-based phylogenetic analysis (1)] surprisingly few integrated software environments for managing such workflows of basic analysis steps in a versatile and user-friendly way are publicly available. Existing client–server applications may be subdivided into classic web portals (2,3) and—emerging more recently—solutions based on so-called rich clients for harvesting services and data which are dispersed across the Internet [cf. Ref. (4) for an example and recent overview; see also www.kepler-project.org].

Owing to its service-oriented software architecture our system can serve both purposes: while in this article we shall mainly focus on functionalities offered by a powerful web interface, the MIGenAS infrastructure also provides SOAP-based web services that can be be utilized by third-party (remote) client applications.

FEATURES AND FUNCTIONALITIES

The MIGenAS bioinformatics toolkit is a new web application for processing basic bioinformatics tasks as well as orchestrating them into complex workflows within a single, coherent web interface. Target users are only assumed to be familiar with the basic functionality offered by the popular sequence analysis tools. Neither additional computational prerequisites (A modern version of one of the popular web browsers, Mozilla/Firefox, Opera or Internet Explorer is required with JavaScript enabled.) nor in-depth bioinformatics experience is considered to be necessary for working with the toolkit. The system has been developed with support of the MIGenAS consortium of the Max-Planck-Society. Founding members are the Max-Planck-Institute (MPI) of Biochemistry (Department of Oesterhelt), MPI for Computer Science (Department of Lengauer), MPI for Developmental Biology (Department of Lupas, and Group S.C. Schuster: presently at Pennsylvania State University, USA), MPI for Marine Microbiology (Department of Amann) and the RZG. Services are provided and hosted by

*To whom correspondence should be addressed. Tel: +49 0 89 3299 2176; Fax: +49 0 80 3299 1301; Email: markus.rampp@rzg.mpg.de

the Garching Computing Centre of the Max-Planck-Society (RZG), which maintains all software, hardware and data related to the MIGenAS toolkit.

Technology

Emphasis has been placed on designing a scalable and extensible, object-oriented software architecture (based on the Java2 Enterprise Edition platform). Details about architecture, design and implementation are described in Ref. (5). With a web application and web services as the main client interfaces a broad spectrum of use cases can be covered ranging from interactive, web-based workflow processing to the integration of (web) services into sophisticated remote applications.

In order to ensure privacy and security for users all communications are handled via the https protocol. Upon start of a new session with the MIGenAS toolkit (via anonymous login 'Guest') the user gets redirected to the secure (SSL/TLS encryption) https communication port. The web portal's identity is authenticated by a certificate issued by the Max-Planck Certificate Authority (http://ca.mpg.de/).

Tools

The web application supports the main categories of classic bioinformatics tasks (see Table 1). We have opted for a manageable selection of packages for each functional category rather than providing an anonymous collection of a large number of tools. Packages are carefully selected according to their performance, circulation and computational efficiency. New tools are scheduled for integration on request.

Databases

For efficient access by the MIGenAS server the following FASTA nucleic and amino acid sequence databases are mirrored locally at RZG with at least a weekly update interval (links to original resources are stated within parentheses): nr, env_nr, nt, sts, ESTs (www.ncbi.nih.gov), Swiss-Prot, TrEMBL (www.uniprot.org), PIR-NREF (pir.georgetown.edu), PDB (www.rcsb.org) and KEGG GENES (www.genome.ad.jp). A complete and up-to-date collection of organism-specific FASTA databases of the completed microbial genomes from NCBI is available together with a number of eukaryotic genomes. Clustered EST sequences are provided as FASTA databases for *Homo sapiens*, Mouse and Drosophila (http://genenest.molgen.mpg.de/). In addition, HMM libraries based on Pfam-A (http://pfam.wustl.edu/) can be searched. Uploading of user-supplied sequence databases is supported by the majority of tools. Such (private) data are not visible outside of the user's session.

Table 1. Overview of function categories with all tools currently supported by the MIGenAS toolkit

Sequence similarity search	Multiple sequence alignment	Phylogeny/ classification	Structure prediction
NCBI-BLAST (6)	ClustalW (7)	PHYLIP (8)	Arby (9)
HHSearch (10)	DIALIGN 2 (11)	seqboot	JNet (12)
HMMer (13)	MUSCLE (14)	protdist, neighbor	PsiPred (15)
PSI-BLAST (6)	PCMA (16)	consense	SignalP (17)
HMMAccel	POA (18)	drawgram	TMHMM (19)
	T-Coffee (20)	CLANS (21)	MODELLER (22)
	Blammer, CluCheck		

An up-to-date list of tools (and databases) with links to detailed documentation is maintained on the MIGenAS web portal. The tools named 'HMMAccel' (for performing accelerated HMMer searches; Frickey & Söding), 'Blammer' (for aligning BLAST hit sequences; Frickey & Lupas) and 'CluCheck' (for automatic assessment of alignment quality; Frickey & Lupas) are not yet published.

Basic user interface

The essential user interaction occurs in the large, central part of the web portal which displays the forms prompting the user for input data and parameters and renders the output of completed computations (Figures 1 and 2). The set of supported tools is arranged in a hierarchical tabbed structure. The user navigates between tools by first selecting the tab with the corresponding tool category and then clicking the particular tool. Basic controls for working with a tool are located in the narrow horizontal bar shown at the top of the page. This control bar hosts a number of pull-down menus which allow to switch between different runs with the same tool ('Runs'), to navigate between input form, documentation and output display ('View'), to redirect results to other tools ('Forward') and to download ('Export') results. The 'submit' button needs to be clicked for starting computations (see Figures 1 and 2). The user provides primary input data (e.g. protein sequences and multiple sequence alignments) to be analyzed by either pasting or uploading the data in one of the popular formats or by directly selecting output from a preceding computation performed within the toolkit (see below). Tool-specific parameters, such as *E*-value cutoffs, databases to be searched and so on, are defined by making selections in the corresponding form fields which are located below the aforementioned input-data fields (Figure 1). Small pop-up 'tooltips' with a brief explanation of a specific parameter are displayed when the user hovers over the corresponding hyperlink with the mouse pointer. Clicking the hyperlink redirects to a more detailed documentation of the tool and its parameters.

The parameter space of interest can be systematically explored by creating a new 'run' for each relevant combination of input parameters for a particular tool. Obtained results may be forwarded to another tool or downloaded in different formats to the user's PC by making the corresponding selection from the pull-down menu named 'Forward' or 'Export', respectively (see Figure 2).

The narrow vertical area on the right-hand side of the portal shows a status overview of computing tasks and facilitates quick navigation to all runs performed within a session. The upper part of this area is reserved for creating and managing persistent projects. This feature, which is currently available only to a core user community equipped with personalized accounts, will soon be released for public use.

Pipelining

The notion of a 'run' with a tool is the central concept underlying the pipelining capabilities of this application: if output data of tool *A* can (in principle) be used as input for another tool *B*, all runs the user has already performed with tool *A* are offered as selectable input for tool *B*. For example, the target

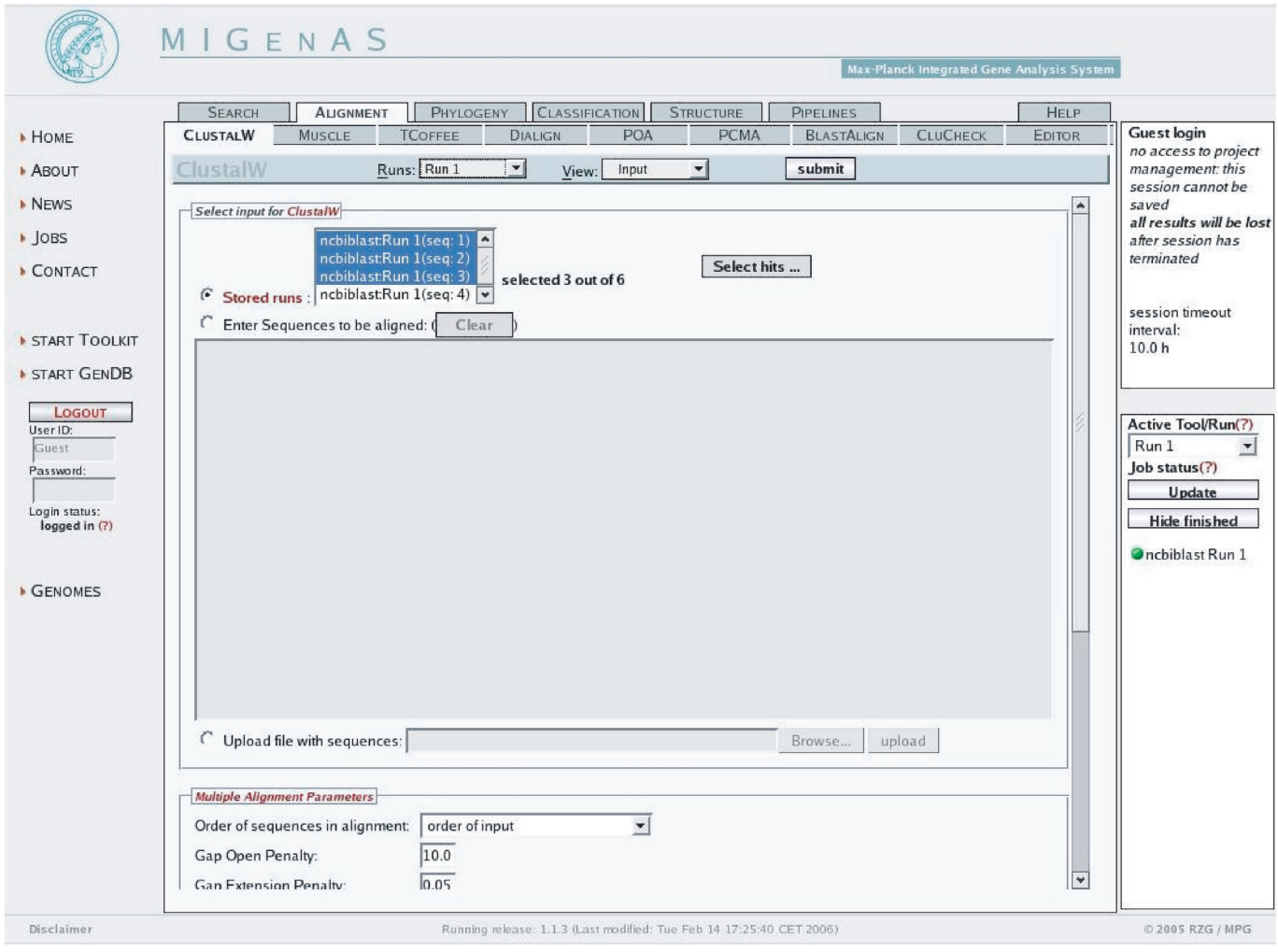

Figure 1. Selection of input data and parameters for multiple sequence alignment computation with the ClustalW tool. In this example three independent sets of target sequences identified by three different preceding BLAST searches will be subjected to multiple sequence alignment.

sequences found in a run with a search tool such as BLAST can be immediately used as input for an alignment tool such as ClustalW (see Figure 1). The above mentioned 'Forward' pull-down menu which is displayed when inspecting tool results facilitates the forwarding of results to another tool for further processing (Figure 2).

In addition to such semi-automatic workflow management where the user interactively coordinates the succession of tools it is also possible to preconfigure a custom 'Meta'-tool (tab-group 'Pipelines') as a pipeline of individual tools and intermediate filters. The same pipeline can then be employed for conveniently processing different sets of input data and parameters. For example, such a tool pipeline could start by a sequence similarity search with the target sequences being filtered according to a chosen *E*-value cut-off, subsequently being subjected to multiple alignment, automatic validation and finally phylogenetic tree-building.

Customization of results, data integration

All relevant results of computations are internally interpreted ('parsed') by the server. This is not only a fundamental prerequisite for the pipelining capabilities described above but also allows us to add value to the raw results delivered by the underlying software packages. Figure 2, for example, shows a color-coded version of a scored multiple sequence alignment as computed by 'ClustalW' together with a ruler for residue-position numbers. As an example for a more advanced feature we point out the capability for comprehensive and reliable annotation of sequences by species and gene names, protein names as well as possible synonyms and accession codes in various sequence databases. This is based on the PIR-NREF (23) and UniProt (24) databases (since recently, PIR-NREF has been superseded by UniProt) and applies to all sequences which have been extracted from one of the major protein sequence databases. We also show literature links to PubMed (www.ncbi.nlm.nih.gov), which are related (according to the information provided by PIR-NREF/UniProt) to the protein under consideration. The complete text of PubMed abstracts gets asynchronously retrieved and is displayed in a small frame when the user hovers the mouse pointer over the PubMed icon, which is displayed next to, e.g. a BLAST hit.

Tasks for display and post processing of results, which require a higher degree of interactivity than an HTML-based

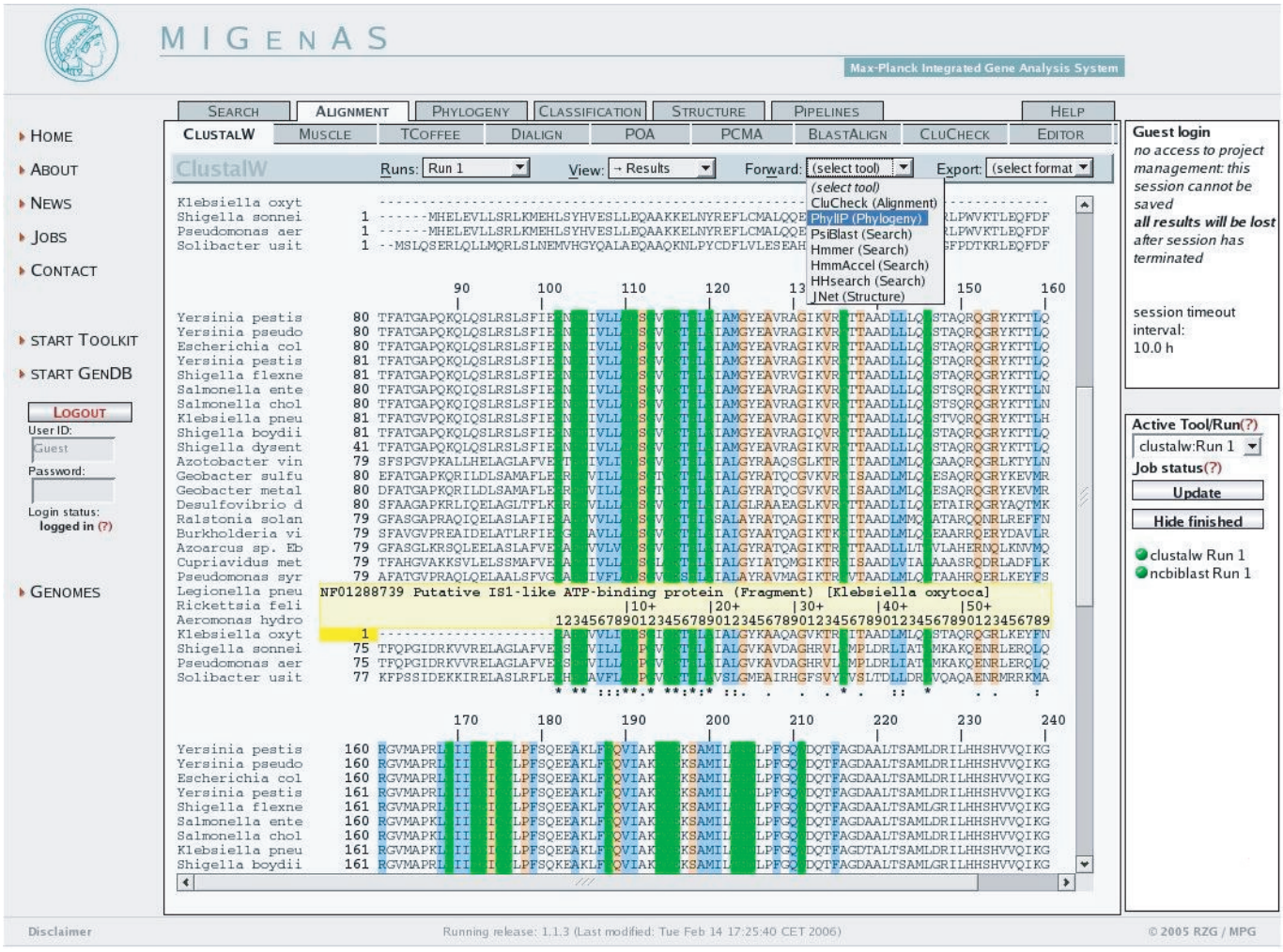

Figure 2. Result of a multiple sequence alignment computation with the ClustalW tool. The pulled-down menu named 'Forward' (top right) offers a selection of tools suitable for subsequent processing of the alignment.

web application conceivably can offer, are delegated to Java Applets. Examples are the applets named 'ATV' (25) for treeviewing, 'JalView' (26) for editing alignments, 'Jmol' (www.jmol.org) for rendering 3D protein structures and 'CLANS' (21) for interactive visualization of pairwise sequence similarities.

Parallel processing

The majority of tools supported by the MIGenAS toolkit allow parallel processing of multiple, mutually independent input data. When pasting or uploading a set of protein sequences, for example, or selecting multiple output from a preceding run for further processing with another tool, a new run with this tool is created automatically and executed in parallel for each individual input with only a single step of user interaction.

SOAP-based web services

Naturally, not all conceivable sorts of analysis and post-processing procedures for tool results can be anticipated and implemented into a web application. In order to allow advanced users to take advantage of existing MIGenAS services, yet exert maximum control (e.g. by embedding them in their own scripts), programmatic access to individual tool interfaces is exported in the form of SOAP-based web services (cf. 27). This, in particular, allows integration with other third-party remote applications [see Ref. (28) and references cited therein]. Example code written in the Perl or Java programming language for a number of web service clients of the MIGenAS toolkit is distributed on request.

FUTURE DIRECTIONS

Development of the MIGenAS toolkit which we introduced in this article has been user-driven from the beginning. The functionalities of the toolkit are continually being updated and extended in response to requests and suggestions, which are emerging from the core user community of the MIGenAS consortium. According to the consortium's original focus on microbial genome research the majority of studies conducted so far has been dealing with microbial genes. Although the toolkit in principle is not limited to these types of analysis, the

current selection of tools, databases and especially supported use-cases is probably slightly biased.

Accordingly, we plan to extend and generalize scope and functionality of the server, and would like to encourage prospective users to provide us with feedback, in particular on usability of the system and desirable new features.

In addition, a comprehensive set of SOAP-based web services with corresponding client codes and workflow tools will be made available on the MIGenAS web portal in the near future.

ACKNOWLEDGEMENTS

We are indebted to the members of the MIGenAS consortium for sharing their software and expertise with us. An anonymous referee is gratefully acknowledged for valuable criticism which helped us to improve the usability of the system. Funding to pay the Open Access publication charges for this article was provided by the Max-Planck-Society.

Conflict of interest statement. None declared.

REFERENCES

1. Ludwig,W., Strunk,O., Westram,R., Richter,L., Meier,H., Yadhukumar, Buchner,A., Lai,T., Steppi,S., Jobb,G. *et al.* (2004) ARB: a software environment for sequence data. *Nucleic Acids Res.*, **32**, 1363–1371.
2. Crass,T., Antes,I., Basekow,R., Bork,P., Buning,C., Christensen,M., Claussen,H., Ebeling,C., Ernst,P., Gailus-Durner,V. *et al.* (2004) The Helmholtz Network for Bioinformatics: an integrative web portal for bioinformatics resources. *Bioinformatics*, **20**, 268–270.
3. Gracy,J. and Chiche,L. (2005) PAT: a protein analysis toolkit for integrated biocomputing on the web. *Nucleic Acids Res.*, **33**, W65–W71.
4. Navas-Delgado,I., Rojano-Muñoz,M., Ramírez,S., Pérez,A., Andrés León,E., Aldana-Montes,J. and Trelles,O. (2006) Intelligent client for integrating bioinformatics services. *Bioinformatics*, **22**, 106–111.
5. Rampp,M. and Soddemann,T. (2005) A work flow engine for microbial genome research. In Kremer,K. and Macho,V. (eds), *Forschung und Wissenschaftliches Rechnen 2004,* Volume 68 of GWDG-Reports (ISSN 0176-2516). Ges. für wissenschaftliche Datenverarbeitung, Göttingen, Germany, pp. 17–46.
6. Altschul,S., Madden,T., Schaffer,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.
7. Chenna,R., Sugawara,H., Koike,T., Lopez,R., Gibson,T.J., Higgins,D.G. and Thompson,J.D. (2003) Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res.*, **31**, 3497–3500.
8. Felsenstein,J. (1989) Phylip—phylogeny inference package. *Cladistics*, **5**, 164–166.
9. Von Öhsen,N., Sommer,I., Zimmer,R. and Lengauer,T. (2004) Arby: automatic protein structure prediction using profile–profile alignment and confidence measures. *Bioinformatics*, **20**, 2228–2235.
10. Söding,J. (2004) Protein homology detection by HMM–HMM comparison. *Bioinformatics*, **199**, 133–154.
11. Morgenstern,B. (1999) Dialign 2: improvement of the segment-to-segment approach to multiple sequence alignment. *Bioinformatics*, **15**, 211–218.
12. Cuff,J. and Barton,G. (1999) Application of enhanced multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins*, **40**, 502–511.
13. Eddy,S. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.
14. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
15. Jones,D. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.*, **292**, 195–202.
16. Pei,J., Sadreyev,R. and Grishin,N. (2003) PCMA: fast and accurate multiple sequence alignment based on profile consistency. *Bioinformatics*, **19**, 427–428.
17. Bendtsen,J., Nielsen,H., von Heijne,G. and Brunak,S. (2004) Improved prediction of signal peptides: Signalp 3. *J. Mol. Biol.*, **340**, 783–795.
18. Lee,C., Grasso,C. and Sharlow,M. (2002) Multiple sequence alignment using partial order graphs. *Bioinformatics*, **18**, 452–464.
19. Krogh,A., Larsson,B., Heijne,G.v. and Sonnhammer,E. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.*, **305**, 567–580.
20. Notredame,C., Higgins,D. and Heringa,J. (2000) T-Coffee: a novel method for multiple sequence alignments. *J. Mol. Biol.*, **302**, 205–217.
21. Frickey,T. and Lupas,A. (2004) CLANS: a Java application for visualizing protein families based on pairwise similarity. *Bioinformatics*, **20**, 3702–3704.
22. Fiser,A. and Sali,A. (2003) Modeller: generation and refinement of homology-based protein structure models. *Methods Enzymol.*, **374**, 461–491.
23. Wu,C.H., Yeh,L.S., Huang,H., Arminski,L., Castro-Alvear,J., Chen,Y., Hu,Z., Kourtesis,P., Ledley,R.S., Suzek,B.E. *et al.* (2003) The Protein Information Resource. *Nucleic Acids Res.*, **31**, 345–347.
24. Apweiler,R., Bairoch,A., Wu,C.H., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R., Magrane,M. *et al.* (2005) UniProt: the universal protein knowledgebase. *Nucleic Acids Res.*, **32**, D115–D119.
25. Zmasek,C.M. and Eddy,S.R. (2001) ATV: display and manipulation of annotated phylogenetic trees. *Bioinformatics*, **17**, 383–384.
26. Clamp,M., Cuff,J., Searle,S.M. and Barton,G.J. (2004) The Jalview Java alignment editor. *Bioinformatics*, **12**, 426–427.
27. Pillai,S., Silventoinen,V., Kallio,K., Senger,M., Sobhany,S., Tate,J., Velankar,S., Golovin,A., Henrick,K., Rice,P. *et al.* (2005) SOAP-based services provided by the European Bioinformatics Institute. *Nucleic Acids Res.*, **33**, W25–W28.
28. Oinn,T., Addis,M., Ferris,J., Marvin,D., Senger,M., Greenwood,M., Carver,T., Glover,K., Pocock,M.R., Wipat,A. *et al.* (2004) Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics*, **20**, 3045–3054.