D252–D256 *Nucleic Acids Research, 2006, Vol. 34, Database issue*
doi:10.1093/nar/gkj106

SIMAP: the similarity matrix of proteins

Thomas Rattei[1],*, Roland Arnold[2], Patrick Tischler[2], Dominik Lindner[1], Volker Stümpflen[2] and H. Werner Mewes[1,2]

[1]Department of Genome Oriented Bioinformatics, Technical University of Munich, Wissenschaftszentrum Weihenstephan, 85350 Freising, Germany and [2]Institute for Bioinformatics, GSF-National Research Center for Environment and Health, Ingolstädter Landstrasse 1, 85764 Neuherberg, Germany

Received August 15, 2005; Revised and Accepted October 17, 2005

ABSTRACT

Similarity Matrix of Proteins (SIMAP) (http://mips.gsf.de/simap) provides a database based on a pre-computed similarity matrix covering the similarity space formed by >4 million amino acid sequences from public databases and completely sequenced genomes. The database is capable of handling very large datasets and is updated incrementally. For sequence similarity searches and pairwise alignments, we implemented a grid-enabled software system, which is based on FASTA heuristics and the Smith–Waterman algorithm. Our ProtInfo system allows querying by protein sequences covered by the SIMAP dataset as well as by fragments of these sequences, highly similar sequences and title words. Each sequence in the database is supplemented with pre-calculated features generated by detailed sequence analyses. By providing WWW interfaces as well as web-services, we offer the SIMAP resource as an efficient and comprehensive tool for sequence similarity searches.

INTRODUCTION

Sequence similarity searches, mostly performed by BLAST (1) or FASTA (2), are an essential step in the analysis of any protein sequence and by far the most intensively used bioinformatics methods. Sequence conservation as the basic evolutionary principle implies conservation of structure and function. Thus, structural and functional attributes that cannot be predicted from the sequence alone can be efficiently transferred from known to uncharacterized proteins. In general, for the coding segments of any genome, searches on the protein level are by far more sensitive than on the corresponding DNA-sequences owing to the better signal to noise ratio of the 20 amino acid alphabet in proteins (3).

The result of any sequence similarity search against a database is a list of significant matches ordered by the similarity score of the pairwise alignments. However, this list represents only a 1D view of the *n*-dimensional relation between a set of similar and probably evolutionarily conserved sequences. The complete similarity matrix (all-against-all) covers the complete 'protein similarity space'. Therefore, the information content of an exhaustive database of similarity scores increases substantially since it takes all relations of any similarity sub-graph into account. Employing subsequent analyses such as clustering allows for efficient computation of a number of essential genome analysis tasks applicable to the protein space. These include the systematic detection and identification of conserved domains (4), the analysis of protein families and super-families in large datasets (5), the detection of orthologs and paralogs for any pair of genomes (6), the identification of clusters of orthologous groups in any number of genomes (7) as well as the application of methods for functional prediction such as phylogenetic profiling (8), the Rosetta stone method (9) or the principle of conserved gene neighborhood (10). Several implementations of all-against-all matrices were reported (11–14). Most of these systems were built to support automatic annotation of proteins (15). However, none of the systems described earlier provides a comprehensive coverage with respect to the known sequence space nor does it allow for the searches by sub- or highly similar sequences.

The optimal solution to generate the similarity matrix would be the exhaustive application of the Smith–Waterman alignment algorithm (16) and the subsequent storage of any significant scores. Although efficient implementations (17) exist, the computational costs are beyond feasibility. Thus, heuristic approaches like BLAST (1) or FASTA (2) are used to speed up the search for biologically meaningful hits in a database and they became the most intensively used tools in sequence analysis.

Typically, sequence similarity searches of individual sequences or genomes are repeated frequently since the available datasets change over time. In many analyses such

*To whom correspondence should be addressed. Tel: +49 8161 712132; Fax: +49 8161 712186; Email: t.rattei@wzw.tum.de as the detection of orthologous relationships (6), this re-computation is the most time consuming step and makes the analysis intractable for large numbers of datasets. Therefore, a pre-calculated all-against-all matrix becomes desirable, which stores the similarity-space in a database and allows rapid access to significant hits of interest.

Such a database must reduce redundancy generated by sequences that are conserved close to identity. It should provide useful interfaces for the user to allow for the extraction of biologically meaningful subsets and the application of different cut-offs. It should be regularly and frequently updated. Scores must therefore be independent of the database size and composition in order to ensure compatibility between different versions (use of probability values instead of expectation values). The time complexity for an all-against-all comparison to generate the sequence similarity space is $O(n^2)$ where n is the number of sequences in the database. In good approximation, the alignments and the alignment raw scores are symmetrical (therefore, we assume the score for an alignment formed by sequence A with sequence B to be the same as for B with A; this assumption is essential to be able to perform incremental updates). This property reduces the amount of computation required by half (18). Scores for any new sequences are saved and the result lists of the old sequences are updated without re-computation. In this paper, we present the Similarity Matrix of Proteins, SIMAP, as an implemented solution for a database representing the protein similarity space.

SYSTEM ARCHITECTURE

Import of data

SIMAP represents sequences extracted from heterogeneous data sources. For this reason we have implemented a flexible input layer which is based on the Data Access Object (DAO) design pattern. DAO classes are available for files using multiple FASTA and EMBL formats, databases like PEDANT (19) as well as for web-services as provided by plantsDB and Genome Research Environment (GenRE) projects at MIPS (20). The imported data is separated into three entities:

(i) Database (describes the context of the proteins),
(ii) Protein (describes a certain protein entry using references to database and protein sequence),
(iii) Sequence (contains the non-redundant protein sequences, checksums and self-scores).

As all similarity and feature calculations rely only on sequence information, the separation of protein and sequence information is necessary to avoid redundant calculations. All protein sequences are preprocessed for validation and low complexity filtering. In order to avoid loss of information, low complexity regions are not masked by ‘X’ but converted into lower case letters.

New databases to be included in SIMAP are added manually because some additional information, such as the taxonomy node ID is required. The protein sequence import and database update procedures run fully automatically. Update procedures may be triggered either by chronological jobs or manually. New sequences are scheduled for similarity calculation.

Similarity calculation

The central component of the SIMAP is the calculation module. Its concept is based on the heuristic search algorithm that pre-computes the sequence similarities. Because it was evaluated to be the best compromise between computational speed and sensitivity (21) we have chosen FASTA (2) for finding all putative hits. The FASTA parameter ktup = 1 and BLOSUM50 substitution matrix are used to adjust the calculations to optimal sensitivity. Before FASTA calculations all low complexity regions in the sequences are masked by seg (22). In order to store the correct alignment coordinates and scores into the hit database, every FASTA hit is recalculated without low complexity filtering using the Smith–Waterman algorithm and BLOSUM50 substitution matrix. If the final Smith–Waterman Score is $\geqslant 80$ the hit is accepted and stored. This score is independent from the query length and the database size as it is necessary for incremental updates. The score-threshold of 80 is a compromise between sensitivity and the amount of data to be handled in the database.

The calculation client runs as a command-line program e.g. in Sun Gridengine clusters (http://gridengine.sunsource.net) and also contains the BOINC core client to be used in BOINC based grid systems (http://boinc.berkeley.edu). The results are validated by the SIMAP server and encoded into the binary hit format. Every hit above the threshold to be stored in the databases contains

(i) Sequence ID,
(ii) Smith–Waterman score,
(iii) Identity score,
(iv) Similarity score,
(v) Overlap size of the pairwise alignment,
(vi) Start and Stop coordinates of the alignment in both proteins.

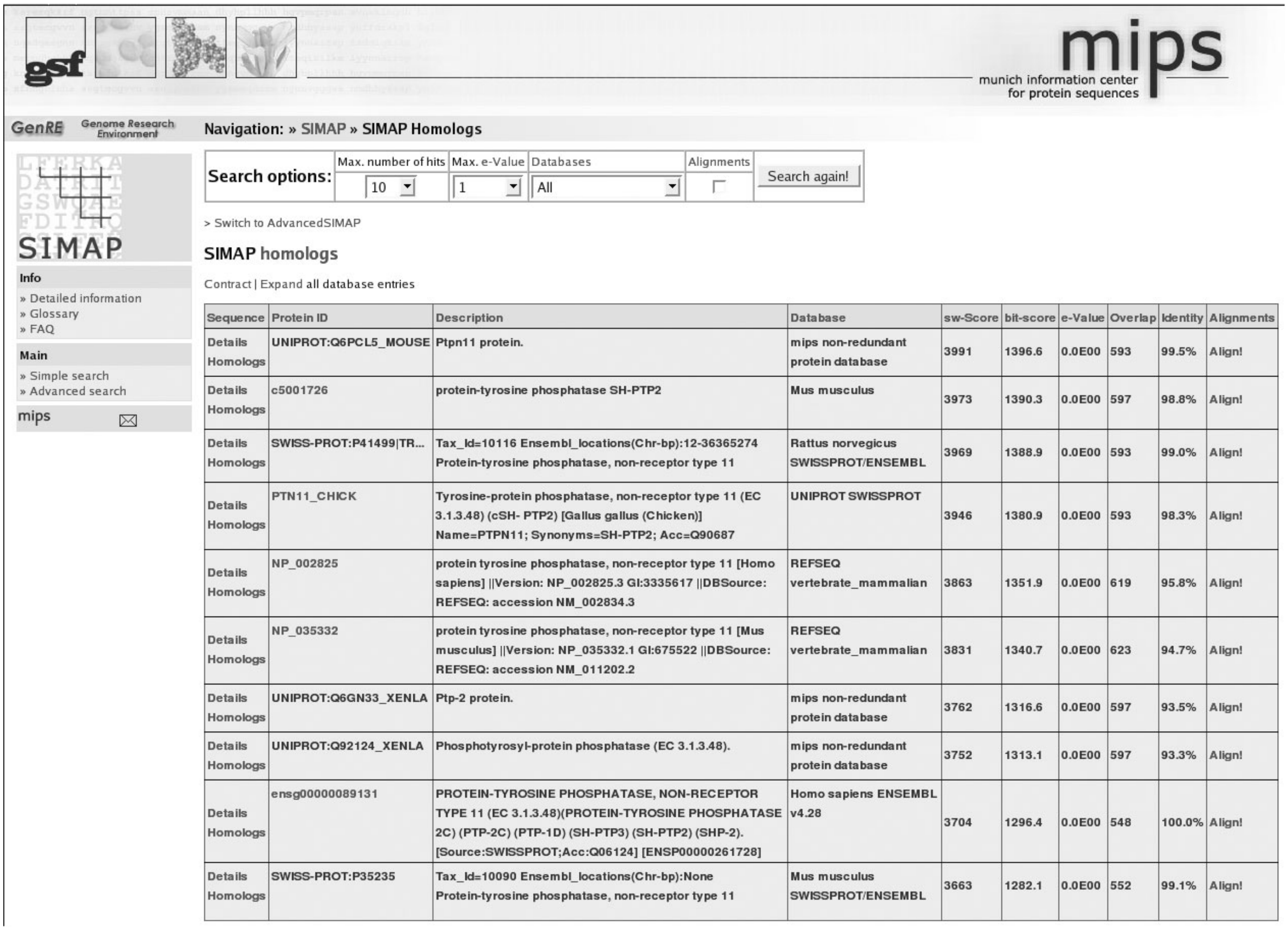
To provide retrieval-optimized data structures, all hits are sorted descending by score and organized in a hash-like structure that is stored in one binary hitfile per sequence:

(i) The key (sequence ID) is encoded by pathname and filename,
(ii) The value (sorted list of hit data blocks as described above) is stored within the file content.

This approach trades time for disc-space, so every hit is stored redundantly in two hitfiles according to the two sequences of the pair. Nevertheless, this turned out to be a simple and straightforward implementation providing the necessary retrieval speed and scalability with respect to the expected growth of public sequence databases.

Data access and retrieval

A server based retrieval layer was implemented using Enterprise Java Beans (EJB). It operates as a database abstraction layer and hides the internal structure of SIMAP for clients. The EJBs are server side components designed for distributed access and information management. They allow easy integration of SIMAP in any kind of application within the MIPS Genome Research Environment GenRE (http://mips.gsf.de/genre/proj/genre) used for our various genome and protein interaction databases. Direct access to SIMAP is not restricted to internal applications but the same functionality is offered for external access through the web. We have developed


gsf
mips
munich information center for protein sequences
GenRE Genome Research Environment
SIMAP
Info
» Detailed information
» Glossary
» FAQ
Main
» Simple search
» Advanced search
mips
Search again!
> Switch to AdvancedSIMAP
Contract | Expand all database entries

Navigation: » SIMAP » SIMAP Homologs

Search options: Max. number of hits: 10; Max. e-Value: 1; Databases: All; Alignments: □

SIMAP homologs

Sequence	Protein ID	Description	Database	sw-Score	bit-score	e-Value	Overlap	Identity	Alignments
Details Homologs	UNIPROT:Q6PCL5_MOUSE	Ptpn11 protein.	mips non-redundant protein database	3991	1396.6	0.0E00	593	99.5%	Align!
Details Homologs	c5001726	protein-tyrosine phosphatase SH-PTP2	Mus musculus	3973	1390.3	0.0E00	597	98.8%	Align!
Details Homologs	SWISS-PROT:P41499|TR...	Tax_Id=10116 Ensembl_locations(Chr-bp):12-36365274 Protein-tyrosine phosphatase, non-receptor type 11	Rattus norvegicus SWISSPROT/ENSEMBL	3969	1388.9	0.0E00	593	99.0%	Align!
Details Homologs	PTN11_CHICK	Tyrosine-protein phosphatase, non-receptor type 11 (EC 3.1.3.48) (cSH- PTP2) [Gallus gallus (Chicken)] Name=PTPN11; Synonyms=SH-PTP2; Acc=Q90687	UNIPROT SWISSPROT	3946	1380.9	0.0E00	593	98.3%	Align!
Details Homologs	NP_002825	protein tyrosine phosphatase, non-receptor type 11 [Homo sapiens] ||Version: NP_002825.3 GI:3335617 ||DBSource: REFSEQ: accession NM_002834.3	REFSEQ vertebrate_mammalian	3863	1351.9	0.0E00	619	95.8%	Align!
Details Homologs	NP_035332	protein tyrosine phosphatase, non-receptor type 11 [Mus musculus] ||Version: NP_035332.1 GI:675522 ||DBSource: REFSEQ: accession NM_011202.2	REFSEQ vertebrate_mammalian	3831	1340.7	0.0E00	623	94.7%	Align!
Details Homologs	UNIPROT:Q6GN33_XENLA	Ptp-2 protein.	mips non-redundant protein database	3762	1316.6	0.0E00	597	93.5%	Align!
Details Homologs	UNIPROT:Q92124_XENLA	Phosphotyrosyl-protein phosphatase (EC 3.1.3.48).	mips non-redundant protein database	3752	1313.1	0.0E00	597	93.3%	Align!
Details Homologs	ensg00000089131	PROTEIN-TYROSINE PHOSPHATASE, NON-RECEPTOR TYPE 11 (EC 3.1.3.48)(PROTEIN-TYROSINE PHOSPHATASE 2C) (PTP-2C) (PTP-1D) (SH-PTP3) (SH-PTP2) (SHP-2). [Source:SWISSPROT;Acc:Q06124] [ENSP00000261728]	Homo sapiens ENSEMBL v4.28	3704	1296.4	0.0E00	548	100.0%	Align!
Details Homologs	SWISS-PROT:P35235	Tax_Id=10090 Ensembl_locations(Chr-bp):None Protein-tyrosine phosphatase, non-receptor type 11	Mus musculus SWISSPROT/ENSEMBL	3663	1282.1	0.0E00	552	99.1%	Align!

Figure 1. Illustration of the list of homologs for the UNIPROT protein Q06124, the human protein-tyrosine phosphatase, non-receptor type 11. Starting from the fulltext search using ProtInfo, the list of homologs can be accessed. From a list of homologs, for every hit links to the pairwise alignment, the report page and the list of its homologs are provided. Additionally the filter options and search scope for the list of homologs can be modified.

additionally a Helmholtz Open Bioinformatics Information Technology (HOBIT) service layer (http://hobit.gsf.de) based on the web service technology to open SIMAP for programming language independent access.

DATA CONTENT

Data from the prominent public protein databases and completely sequenced genomes was imported into SIMAP. At present SIMAP contains the recent versions of these databases:

(i) UNIPROT TrEMBL (23)
(ii) UNIPROT Swissprot (23)
(iii) mips nonredH
(iv) PDB (24)
(v) All genomes included in PEDANT (http://pedant.gsf.de) (19)
(vi) All genome databases at MIPS, e.g. CYGD and MatDB (20,25)
(vii) Several project specific databases.

The total number of ~8 million protein entries corresponds to ~4 million non-redundant protein sequences. The hit files contain ~10 billion single hits.

Most of the databases (UNIPROT, PDB and PEDANT) are weekly checked for updated entries. These updates are performed by a fully automated procedure that also triggers the similarity calculations for new sequences.

SEARCH CAPABILITIES

We have developed ProtInfo to allow for searching sequence homologs for sequences and proteins in SIMAP by using complete sequences but also sequence fragments, similar sequences and keywords. The query sequences are searched within the SIMAP sequences using an indexing structure that allows fast searches for similar or partial sequences in large databases. Each ProtInfo query yields a result list of the identical, containing, contained and most similar SIMAP sequences and their related protein entries. Full text queries are searched in protein IDs and descriptions. Using ProtInfo SIMAP serves as a comprehensive protein information system that provides quickly all proteins that share same or very similar sequences. For every sequence displayed in the search result a link to the list of homologs is provided.

SEQUENCE FEATURES

The non-redundant sequence set of SIMAP is supplemented with protein feature information and cross-references to secondary databases of protein domains and families. The database of associated information is updated automatically whenever new sequences are imported into SIMAP. Currently both calculated and imported features are contained:

(i) General protein information like isoelectric point and molecular weight,
(ii) Transmembrane domains from TMHMM (26),
(iii) Signal peptides from SignalP (27),
(iv) Protein localization from TargetP (28),
(v) Protein domains from InterPro and its member databases (4).

Except on InterPro these features are calculated for the complete amount of sequences. Owing to the computationally expensive hidden Markov Model (HMM) searches for InterPro calculations we import the InterPro hits for all UNIPROT sequences which are provided by the EBI. Additionally we have started to calculate InterPro domains for sequences that are not yet contained in UNIPROT.

WWW INTERFACES

The public SIMAP WWW server (http://mips.gsf.de/simap) offers three entry points for users:

(i) ProtInfo (protein information system),
(ii) SimpleSIMAP (simple retrieval of homologs using a predefined set of parameters), and
(iii) AdvancedSIMAP (flexible retrieval of homologs that provides a wide variety of parameters, sorting and filtering capabilities).

SimpleSIMAP and AdvancedSIMAP retrieve homologs for given protein sequences that need to be contained in the SIMAP database. SimpleSIMAP provides only selected parameters and preconfigured search spaces; it includes the precalculated sequence features. In SimpleSIMAP, *E*-values are computed on-the-fly according to the search space of the query (Figure 1). AdvancedSIMAP allows the user to specify search space, filtering and sorting parameters in a flexible manner. Both types of queries return lists of similar sequences that are recursively linked to their own homologs. Both types of queries provide Smith–Waterman alignments that are computed on-the-fly. Thus, the web interfaces allow users to explore the protein space by sequence similarity, starting with any user defined protein sequence. The retrieved sequences may be downloaded for post-processing, e.g. multiple alignments or reconstruction of phylogenetic trees. The AdvancedSIMAP system provides integrated tools for clustering, multiple alignments and the construction of HMMs.

WEB-SERVICES

Web-services provide open access to SIMAP databases and applications. They are platform independent and may be connected from many programming languages as Perl, Java, C/C++ and Python. Currently methods for the retrieval of homologs by a given sequence are offered.

The web-services are part of the HOBIT project (http://hobit.gsf.de) and can be accessed through http://mips.gsf.de/proj/hobitws/services/RPCSimapService?wsdl and http://mips.gsf.de/proj/hobitws/services/DocSimapService?wsdl.

CONCLUSION AND FURTHER DIRECTIONS

We implemented SIMAP, a database containing the similarity space formed by ~4 million amino acid sequences from >400 organisms by exhaustive similarity searches using the FASTA heuristics. The efficient backbone for computation in addition to the FASTA heuristics and the incremental update process enables us to keep up with the ever-increasing amount of data by using our in-house resources in an efficient way. Powerful search capabilities and the additional sequence feature database allow users to explore the protein space by sequence similarity, starting with a user defined protein sequence or keyword. SIMAP will be continuously updated and expanded to include all publicly available proteomes and major sequence data collections.

ACKNOWLEDGEMENTS

Funding to pay the Open Access publication charges for this article was provided by the GSF-National Research Center for Environment and Health, Neuherberg.

Conflict of interest statement. None declared.

REFERENCES

1. Altschul,S.F., Gish,W., Miller,W., Myers,G. and Lipman,D.J. (1990) A basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
2. Pearson,W.R. (2000) Flexible sequence similarity searching with the FASTA3 program package. *Methods Mol. Biol.*, **132**, 185–219.
3. Gojobori,T., Li,W.H. and Graur,D. (1982) Patterns of nucleotide substitution in pseudogenes and functional genes. *J. Mol. Evol.*, **18**, 360–369.
4. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Binns,D., Bradley,P., Bork,P., Bucher,P., Cerutti,L. *et al.* (2005) InterPro, progress and status in 2005. *Nucleic Acids Res.*, **33**, D201–D205.
5. Krause,A., Haas,S.A., Coward,E. and Vingron,M. (2002) SYSTERS, GeneNest, SpliceNest: exploring sequence space from genome to protein. *Nucleic Acids Res.*, **30**, 299–300.
6. O'Brien,K.P., Remm,M. and Sonnhammer,E.L. (2005) Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Res.*, **33**, D476–D480.
7. Li,L., Stoeckert,C.J.,Jr and Roos,D.S. (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.*, **13**, 2178–2189.
8. Pellegrini,M., Marcotte,E.M., Thompson,M.J., Eisenberg,D. and Yeates,T.O. (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl Acad. Sci. USA*, **96**, 4285–4288.
9. Marcotte,C.J. and Marcotte,E.M. (2002) Predicting functional linkages from gene fusions with confidence. *Appl. Bioinformatics*, **1**, 93–100.
10. Rogozin,I.B., Makarova,K.S., Murvai,J., Czabarka,E., Wolf,Y.I., Tatusov,R.L., Szekely,L.A. and Koonin,E.V. (2002) Connected gene neighborhoods in prokaryotic genomes. *Nucleic Acids Res.*, **30**, 2212–2223.
11. Gonnet,G.H., Cohen,M.A. and Brenner,S.A. (1992) Exhaustive matching of the entire protein sequence database. *Science*, **18**, 1609–1610.
12. Kriventseva,E.V., Fleischmann,W., Zdobnov,E.M. and Apweiler,R. (2001) CluSTr: a database of clusters of SWISS-PROT+TrEMBL proteins. *Nucleic Acids Res.*, **29**, 33–36.
13. Sasson,O., Vaaknin,A., Fleischer,H., Portugaly,E., Bilu,Y., Linial,N. and Linial,M. (2003) ProtoNet: hierarchical classification of the protein space. *Nucleic Acids Res.*, **31**, 348–352.

14. Michalickova,K., Bader,G.D., Dumontier,M., Lieu,H., Betel,D., Isserlin,R. and Hogue,C.W. (2002) Seqhound: biological sequence and structure database as a platform for bioinformatics research. *BMC Bioinformatics*, **3**, 32.
15. Petryszak,R., Kretschmann,E., Wieser,D. and Apweiler,R. (2005) The predictive power of the CluSTr database. *Bioinformatics*, **21**, 3604–3609.
16. Smith,T. and Waterman,M. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.
17. Rognes,T. and Seeberg,E. (2000) Six-fold speed-up of Smith–Waterman sequence database searches using parallel processing on common microprocessors. *Bioinformatics*, **16**, 699–706.
18. Dumontier,M. and Hogue,C.W. (2002) NBLAST: a cluster variant of BLAST for NxN comparisons. *BMC Bioinformatics*, **3**, 13.
19. Riley,M.L., Schmidt,T., Wagner,C., Mewes,H.W. and Frishman,D. (2005) The PEDANT genome database in 2005. *Nucleic Acids Res.*, **33**, D308–D310.
20. Schoof,H., Ernst,R., Nazarov,V., Pfeifer,L., Mewes,H.W. and Mayer,K. (2004) MIPS *Arabidopsis thaliana* Database (MAtDB): an integrated biological knowledge resource for plant genomics. *Nucleic Acids Res.*, **32**, D373–D376.
21. Pearson,W.R. (1991) Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith–Waterman and FASTA algorithms. *Genomics*, **11**, 635–650.
22. Wootton,J.C. (1994) Non-globular domains in protein sequences: automated segmentation using complexity measures. *Comput. Chem.*, **18**, 269–285.
23. Bairoch,A., Apweiler,R., Wu,C.H., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R., Magrane,M. *et al.* (2005) The Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **33**, D154–D159.
24. Westbrook,J., Feng,Z., Chen,L., Yang,H. and Berman,H. (2003) The Protein Data Bank and structural genomics. *Nucleic Acids Res.*, **31**, 489–491.
25. Güldener,U., Münsterkötter,M., Kastenmüller,G., Strack,N., van Helden,J., Lemer,C., Richelles,J., Wodak,S., García-Martínez,J., Pérez-Ortín,J. *et al.* (2005) CYGD: the Comprehensive Yeast Genome Database. *Nucleic Acids Res.*, **33**, D364–D368.
26. Krogh,A., Larsson,B., von Heijne,G. and Sonnhammer,E.L. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.*, **305**, 567–580.
27. Bendtsen,J.D., Nielsen,H., von Heijne,G. and Brunak,S. (2004) Improved prediction of signal peptides: SignalP 3.0. *J. Mol. Biol.*, **340**, 783–795.
28. Emanuelsson,O., Nielsen,H., Brunak,S. and von Heijne,G. (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J. Mol. Biol.*, **300**, 1005–1016.