Nucleic Acids Research, 2007, Vol. 35, Web Server issue ***W433–W437***
doi:10.1093/nar/gkm352

MyHits: improvements to an interactive resource for analyzing protein sequences

Marco Pagni[1,*], Vassilios Ioannidis[1,2], Lorenzo Cerutti[3], Monique Zahn-Zabal[4], C. Victor Jongeneel[1,2,4], Jörg Hau[5], Olivier Martin[1], Dmitri Kuznetsov[1] and Laurent Falquet[2]

[1]Swiss Institute of Bioinformatics (SIB), Vital-IT Group, UNIL-Génopode, CH-1015 Lausanne, [2]Swiss Institute of Bioinformatics (SIB), EMBnet Group, UNIL-Génopode, CH-1015 Lausanne, [3]Swiss Institute of Bioinformatics (SIB), Swiss-Prot Group, UNIGE-CMU, CH-1211 Genève 4, [4]Ludwig Institute for Cancer Research, UNIL-Génopode, CH-1015 Lausanne and [5]Nestlé Research Center, Department of BioAnalytical Science, PO Box 44, CH-1000 Lausanne 26, Switzerland

Received January 31, 2007; Revised and Accepted April 23, 2007

ABSTRACT

The MyHits web site (http://myhits.isb-sib.ch) is an integrated service dedicated to the analysis of protein sequences. Since its first description in 2004, both the user interface and the back end of the server were improved. A number of tools (e.g. MAFFT, Jacop, Dotlet, Jalview, ESTScan) were added or updated to improve the usability of the service. The MySQL schema and its associated API were revamped and the database engine (HitKeeper) was separated from the web interface. This paper summarizes the current status of the server, with an emphasis on the new services.

INTRODUCTION

The MyHits web site (http://myhits.isb-sib.ch) was first described in 2004 (1). Since then, a number of sequence and motif databases have been added. In addition, most web pages have been revised and upgraded, and new services were introduced (Table 1). Currently our server provides access to protein domain databases, something that can be obtained from other servers like the ones of SMART (2) or InterPro (3). However MyHits also offers many possibilities for users to submit their own sequences and multiple sequence alignment (MSA), and to interact with the different tools available. Recently the MPI Bioinformatics Toolkit (4) implemented a similar approach for their services.

One particular aspect of the initial project was the concept of a 'hub', an intermediary page that links the output of one service to the input of another one. This concept is well known on the command line, but rarely found on web pages. Two different hubs were designed: one for handling lists of unaligned proteins and another for handling a single multiple sequence alignment (MSA) at a time. The hubs dramatically improve the overall connectivity between most of the pages on this website (see http://myhits.isb-sib.ch/doc/connectivity.html). Consequently, a user can efficiently string together the different services, mine the available data and obtain answers to a broad range of questions related to protein sequences. Typical applications are:

(i) What domain signatures are present in an unknown sequence (using *Motif scan*)?
(ii) Starting from a single protein sequence, extract the repeated regions of homologies (using *Dotlet*), build a multiple sequence alignment (using *MAFFT*) and search for homologues (using *profile search*) in a given organism (using taxonomic filtering).
(iii) Starting from an MSA, retrieve all members of a protein family (using *PSI-Blast*), and classify the matched proteins automatically (using *Jacop*).

Another original aspect of the previous publication was the availability of 'private' databases. Hosting private data proved to be a successful concept, especially since it allows users to treat their 'private' data in the same workflow as the publicly available data, including simultaneous search, analysis and display; a redesign of the interface to private databases is in progress.

BIOLOGICAL DATABASES

The protein sequences that are currently included are obtained from *UniProt* (5), *RefSeq* (6), the peptide subsection of *ENSEMBL* (7) and *trGen, trEST, trome* which are locally produced databases of predicted protein sequences inferred from various genomic and High Throughput Genomic nucleotidic sequences (8).

*To whom correspondence should be addressed. Tel: +41-21-692-40-38; Fax: +41-21-692-40-65; Email: marco.pagni@isb-sib.ch

Table 1. The different computational services that are integrated in MyHits

Search		
PSI-Blast (20)	Fast database searches using a multiple sequence alignment (MSA) as input, with taxonomic filter and clustering of the matched sequences by similarity.	Updated
Profile search	Slow database searches using an MSA as input, with taxonomic filter and clustering of the matched sequences by similarity.	New
Pattern search	Database searches using a regular expression as input, with taxonomic filter and clustering of the matched sequences if identical.	Updated
Scan		
Motif scan	Scan for motifs (pattern, profile, HMMs) in a sequence.	Updated
Query		
By motif	Query on pre-computed hit lists using motifs as input, with taxonomic filter, and AND/OR logic on the motif occurrences.	Updated
By protein	Query on pre-computed hit lists using sequences as input.	Updated
Align		
ClustalW (21)	Compute MSA.	–
T-Coffee	See External links below.	Replaced
MAFFT (17)	Compute MSA for a large number of sequences.	New
Classify		
Jacop (14)	Automated protein classification. Allows the detection of discriminant regions of sequences.	New
Translate		
Translator	Translate a DNA sequence into selectable ORFs, with the *transeq* program of EMBOSS (18).	New
ESTScan (19)	Detect coding regions in a DNA sequence for a given species.	New
Graphics and Others		
Catalogue factory	Create postscript and PDF document with the graphical representation of motifs and features of proteins.	New
Match viewer	Graphical representation of a local alignment between a protein sequence and a motif (PSI-Blast checkpoint, generalized profile, HMM).	New
Synonyms	Search for synonyms of an identifier in all databases.	New
Entry viewer	View the raw text of a database entry.	–
External servers linked with MyHits (http://t-coffee.vital-it.ch)		
T-coffee (22)	Compute high quality MSA.	Updated
M-Coffee (23)	Compute high quality MSA.	New
Expresso (24)	Compute high quality MSA using 3D structure information. Pretty slow.	New
Protogene (25)	Produce a DNA MSA based on a protein MSA (automatically extract DNA sequences from GenBank).	New
Core (22)	Consistency-based evaluation of an existing MSA.	New
Protein hub		
Selector	Allows the selection of representative members from a list of sequences.	–
Catalogue	Create graphical representation of the motifs and features of a list of proteins.	New
Dotlet (15)	Java Applet displaying a dot plot. Allows printing to a PDF, as well as the interactive selection of similarity regions.	Updated
SEView (26)	Visualize motifs and features from a Swiss-Prot entry.	–
MSA hub		
Selector	Allows the selection of representative members from a list of sequences.	–
Catalogue	Create graphical representation of the motifs and features of a list of proteins, starting from an MSA.	New
Jalview (16)	Java Applet allowing an MSA to be visualized and edited, outputs a PDF or an MSA.	Updated
BoxShade	External server to create a graphical MSA output for publication.	–

Note: When available, the reference for the program or algorithm is given, as well as the specificity of the deployment on our server. The status reported in the last column contains the differences relative to the previous publication (1).

The 'motif' databases include *Prosite* (9), *Pfam* (10), *HAMAP* (11), *InterPro* (3) and a collection of locally produced generalized profiles. Classification databases are currently limited to the *taxonomy of the NCBI* (12,13).

All these databases are updated frequently, in principle on a weekly basis. Hit lists are computed between pairs of databases (sequence *vs* motif), depending on the available CPU and storage space. The actual list is available online (http://myhits.isb-sib.ch/cgi-bin/compute).

HITKEEPER: A NEW RELATIONAL DATABASE

The website is mainly built as the front-end of a relational database that stores the primary data (sequences, motifs and classification data) as well as data derived from these (e.g. list of hits). The software related to this database system, HitKeeper, is now distributed as an Open Source Software package under GPL v2 license which is available at http://hitkeeper.sourceforge.net (27). Once set up, HitKeeper runs like a system daemon and takes care of the updates from external data sources (thus dealing with a 'continuous data flow'), and the incremental updates of the list of hits between sequences and motifs. In addition, HitKeeper features a powerful query language. Throughout the past two years, HitKeeper was improved and fine-tuned, and the existing MyHits services have gained both speed and new features. A key example is the introduction of taxonomic filtering. Some more advanced

features of HitKeeper are being progressively introduced into the system.

TOOLS FOR EXPLORATORY DATA MINING OF PROTEIN SETS

When looking for a conserved domain, a particular signature or attempting to delineate protein families, it is usually more efficient to deal with a whole set of related proteins at once, as opposed to dealing with them one by one. Many tools in MyHits have been designed with this in mind, and accept a set of proteins as input. These include the *protein Hub*, the *Selector* and the *Catalogue*, as well as the different programs that produce an MSA (Table 1). When it is not possible to meaningfully arrange a set of sequences as a single MSA, usually because of architectural reorganization of the sequences (i.e. fusion, duplication, swapping, shuffling of domains), one can try using the *Jacop* tool which is presented below. A typical case is the sequences that can be retrieved from a single *PSI-Blast* search which typically contain a 'common domain' (the PSI-Blast-targeted region) that occurs one or more times per sequence. This domain is usually flanked by diverse but often related sequences.

Jacop (14) is a simple and robust approach for the automated classification of protein sequences with no prerequisite for the sequences to be organized as an MSA. To our knowledge it is the only service currently available online where a collection of sequences in FASTA format can be pasted and an automated classification of these is returned. The method applied is an unsupervised classification method: no prior knowledge is required. Unsupervised classification methods can often be outperformed by supervised ones on the condition that a trusted reference classification is available, which is usually not the case when doing exploratory data mining. *Jacop* can also be used to extract a representative subset of a collection of sequences, complementing the *Selector* tool. In addition, a specifically designed version of the *Catalogue factory* that can be accessed from the *Jacop* output page provides a graphical representation of the local sequence homologies distributed over the classification from *Jacop*. Figure 1 shows an example of graphics that can be produced using this interface.

DOTLET AND JALVIEW

Dotlet is a Java applet for the interactive visualization of dot plots (15). The applet was updated (current version is 1.5), correcting some bugs (mostly in the scrolling process) and adding several new features. Since we had to write a printing library for *Jalview*, it was also incorporated into *Dotlet*, allowing a PDF file of the dot plot to be exported. In addition, it is now possible to select several matching diagonals (i.e. regions of similarity between two sequences) and to send them back to the browser (*Protein Hub*) for further analysis. This feature is unique to *Dotlet*.

Jalview is a Java applet to edit and visualize an MSA (16). In the frame of MyHits we offer a locally modified extension of the public version 2.1, which has the ability to export a PDF for both the alignment and the associated trees. It also allows the alignment to be sent back to the browser (*MSA hub*) for further analysis. These options are not present in the standard applet; we will maintain these options for future versions in collaboration with the *Jalview* development team.

MISCELLANEOUS IMPROVEMENTS AND ADDITIONS

Interactivity was improved for many web pages. For *PSI-Blast*, *Pattern search*, *Profile search* and the *Query* pages, buttons have been added to the output to facilitate the selection of candidate matches for further processing through one of the two hubs. In addition, for these search tools, an interactive taxonomic filter has been implemented: the user can customize the menu by introducing one or several TaxID, to select a species (e.g. 37349 for the woolly mammoth) or a higher taxonomical rank (e.g., 8782 for all birds).

MAFFT is a novel method for calculating MSA based on the fast Fourier Transform (17). We provide access to this tool to deal with large number of sequences.

The *Translator* tool uses the EMBOSS *transeq* program (18) to translate a DNA sequence into the corresponding peptide sequence in any of the six frames. It can also translate specified regions corresponding to the coding regions of the sequence of interest. *ESTScan* is another program that can detect coding regions in DNA sequences, even if they are of low quality (19). It will also detect and correct sequencing errors that lead to a frameshift. The result depends on the species selected. Both *Translator* and *ESTScan* tools can direct their resulting protein sequences to the *Motif scan* for further analysis.

In order to avoid the multiplication of the web pages to maintain, the *T-coffee* page was replaced by calls to an external server that offers the *T-Coffee*, *M-Coffee*, *Expresso*, *Protogene* and *Core* programs (http://tcoffee.vital-it.ch). The resulting MSA can be sent back to the *MSA hub* with a single click.

CONCLUSION

MyHits is an evolving resource. New developments will take advantage of the relational database in the back-end. A layer of web services for programmatic access will be proposed in the near future. A better integration with other external web sites, successfully achieved with the T-Coffee server, is planned.

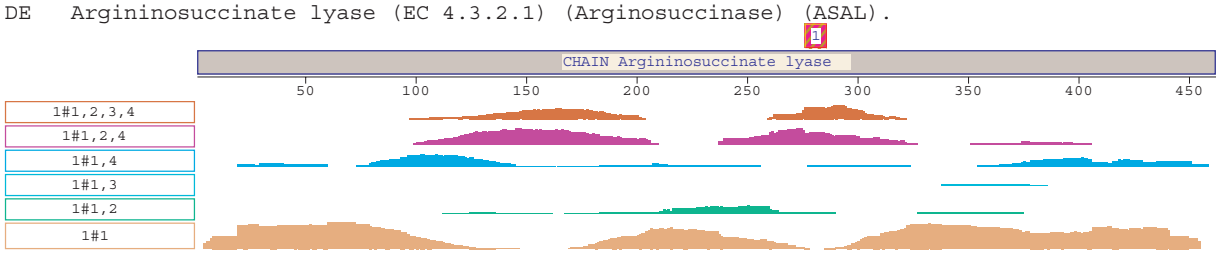
ACKNOWLEDGEMENTS

We thank many people that have contributed, be it with suggestions, testing and technical support to the maintenance and development of MyHits: Volker Flegel, Roberto Fabbretti, Christian Iseli, Peter Sperisen, Heinz Stockinger, Sébastien Moretti, Cédric Notredame. M.P. acknowledges financial support from EMBRACE.

Jacop Group1

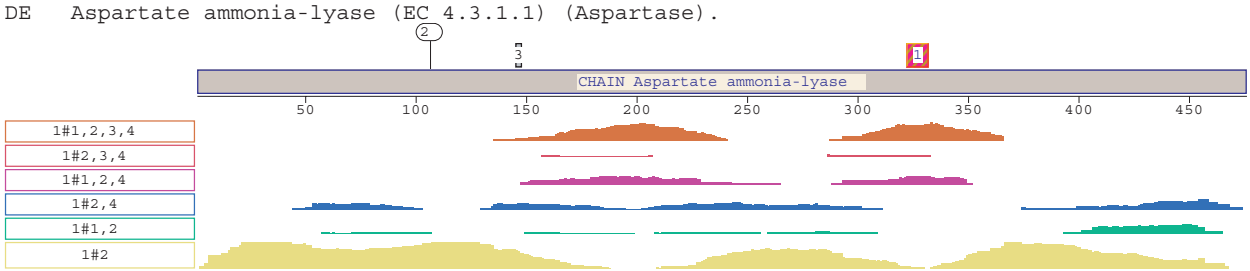
Jacop Subgroup 1#1

sw:ARLY_BACSU
DE Argininosuccinate lyase (EC 4.3.2.1) (Arginosuccinase) (ASAL).



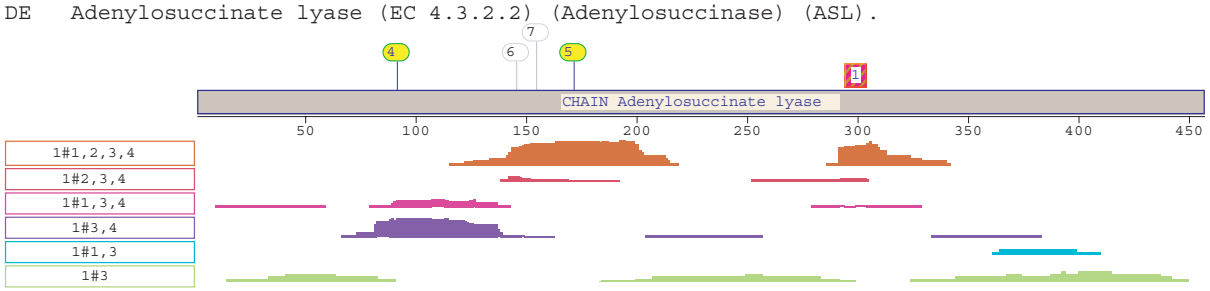
Jacop Subgroup 1#2

sw:ASPA_HAEIN
DE Aspartate ammonia-lyase (EC 4.3.1.1) (Aspartase).



Jacop Subgroup 1#3

sw:PUR8_ECOLI
DE Adenylosuccinate lyase (EC 4.3.2.2) (Adenylosuccinase) (ASL).



Jacop Subgroup 1#4

sw:PUR8_BACSU
DE Adenylosuccinate lyase (EC 4.3.2.2) (Adenylosuccinase) (ASL)
DE (Glutamyl-tRNA synthetase regulatory factor).

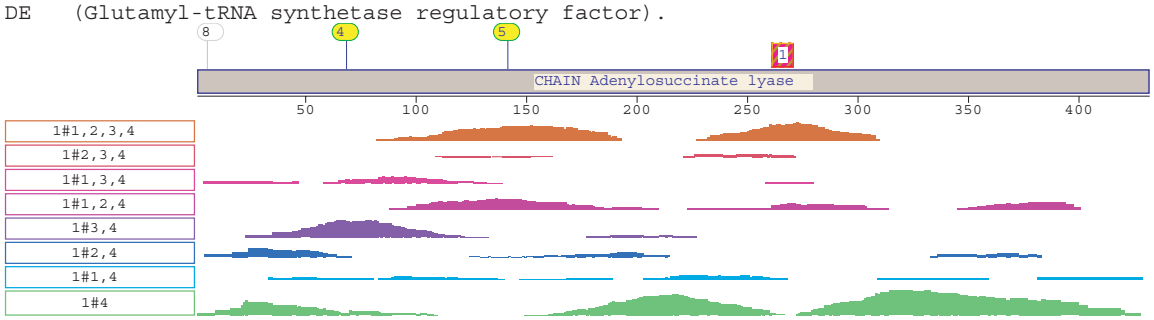


Figure 1. PDF generated using the Catalogue factory from an output of Jacop. The data set is the one designated as 'ARLY' in (14). A representative member of every four sub-groups of the automatically generated classification was picked for graphical display. The histograms indicate where the regions of homology are located. The color scheme corresponds to the sub-groups, i.e. 1#2,4 means that this region of homology matches with a sufficient score proteins of sub-groups 1#2 and 1#4 exclusively. Annotation from the FT lines of the original Swiss-Prot entries, as well as matches from the pre-computed hit list (FUMARATE_LYASES pattern), are represented; numbers refer to: **1**, pat:FUMARATE_LYASES [!]; **2**, BINDING substrate (by similarity); **3**, REGION substrate binding (by similarity); **4**, ACT_SITE Proton donor (by similarity); **5**, ACT_SITE Proton acceptor (by similarity); **6**, CONFLICT P -> A (in Ref. 1); **7**, CONFLICT I -> L (in Ref. 1); **8**, CONFLICT Y -> K (in Ref. 3).

The EMBRACE project is funded by the European Commission within its FP6 Program, under the thematic area 'Life sciences, genomics and biotechnology for health', contract number LHSG-CT-2004-512092. L.C. is funded by the Swiss National Science Foundation (grant no. 3152A0-103922). M.Z.Z. acknowledges financial support from the Cancer Research Institute. Funding to pay the Open Access publication charges for this article was provided by the Swiss Institute of Bioinformatics.

Conflict of interest statement. None declared.

REFERENCES

1. Pagni,M., Ioannidis,V., Cerutti,L., Zahn-Zabal,M., Jongeneel,C.V. and Falquet,L. (2004) MyHits: a new interactive resource for protein annotation and domain identification. *Nucleic Acids Res.*, **32**, W332–W335.
2. Letunic,I., Copley,R.R., Pils,B., Pinkert,S., Schultz,J. and Bork,P. (2006) SMART 5: domains in the context of genomes and networks. *Nucleic Acids Res.*, **34**, D257–D260.
3. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Binns,D., Bradley,P., Bork,P., Bucher,P. *et al.* (2005) InterPro, progress and status in 2005. *Nucleic Acids Res.*, **33**, D201–D205.
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. Wu,C.H., Apweiler,R., Bairoch,A., Natale,D.A., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H. *et al.* (2006) The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res.*, **34**, D187–D191.
6. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **35**, D61–D65.
7. Hubbard,T.J., Aken,B.L., Beal,K., Ballester,B., Caccamo,M., Chen,Y., Clarke,L., Coates,G., Cunningham,F. *et al.* (2007) Ensembl 2007. *Nucleic Acids Res.*, **35**, D610–D617.
8. Sperisen,P., Iseli,C., Pagni,M., Stevenson,B.J., Bucher,P. and Jongeneel,C.V. (2004) trome, trEST and trGEN: databases of predicted protein sequences. *Nucleic Acids Res.*, **32**, D509–D511.
9. Hulo,N., Bairoch,A., Bulliard,V., Cerutti,L., De Castro,E., Langendijk-Genevaux,P.S., Pagni,M. and Sigrist,C.J. (2006) The PROSITE database. *Nucleic Acids Res.*, **34**, D227–D230.
10. Bateman,A., Coin,L., Durbin,R., Finn,R.D., Hollich,V., Griffiths-Jones,S., Khanna,A., Marshall,M., Moxon,S. *et al.* (2004) The Pfam protein families database. *Nucleic Acids Res.*, **32**, D138–D141.
11. Gattiker,A., Michoud,K., Rivoire,C., Auchincloss,A.H., Coudert,E., Lima,T., Kersey,P., Pagni,M., Sigrist,C.J. *et al.* (2003) Automated annotation of microbial proteomes in SWISS-PROT. *Comput. Biol. Chem.*, **27**, 49–58.
12. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J., Rapp,B.A. and Wheeler,D.L. (2000) GenBank. *Nucleic Acids Res.*, **28**, 15–18.
13. Wheeler,D.L., Chappey,C., Lash,A.E., Leipe,D.D., Madden,T.L., Schuler,G.D., Tatusova,T.A. and Rapp,B.A. (2000) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **28**, 10–14.
14. Sperisen,P. and Pagni,M. (2005) JACOP: a simple and robust method for the automated classification of protein sequences with modular architecture. *BMC Bioinformatics*, **6**, 216.
15. Junier,T. and Pagni,M. (2000) Dotlet: diagonal plots in a web browser. *Bioinformatics*, **16**, 178–179.
16. Clamp,M., Cuff,J., Searle,S.M. and Barton,G.J. (2004) The Jalview Java alignment editor. *Bioinformatics*, **20**, 426–427.
17. Katoh,K., Kuma,K., Toh,H. and Miyata,T. (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.*, **33**, 511–518.
18. Rice,P., Longden,I. and Bleasby,A. (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.*, **16**, 276–277.
19. Iseli,C., Jongeneel,C.V. and Bucher,P. (1999) ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol.,* 138–148.
20. 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.
21. 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.
22. 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.
23. Wallace,I.M., O'Sullivan,O., Higgins,D.G. and Notredame,C. (2006) M-Coffee: combining multiple sequence alignment methods with T-Coffee. *Nucleic Acids Res.*, **34**, 1692–1699.
24. Armougom,F., Moretti,S., Poirot,O., Audic,S., Dumas,P., Schaeli,B., Keduas,V. and Notredame,C. (2006) Expresso: automatic incorporation of structural information in multiple sequence alignments using 3D-Coffee. *Nucleic Acids Res.*, **34**, W604–W608.
25. Moretti,S., Reinier,F., Poirot,O., Armougom,F., Audic,S., Keduas,V. and Notredame,C. (2006) PROTOGENE: turning amino acid alignments into bona fide CDS nucleotide alignments. *Nucleic Acids Res.*, **34**, W600–W603.
26. Junier,T. and Bucher,P. (1998) SEView: a Java applet for browsing molecular sequence data. *In Silico Biol.*, **1**, 13–20.
27. Hau,J., Muller,M. and Pagni,M. (2007) HitKeeper, a generic software package for hit list management. *Source code for biology and medicine*, **2**, 2.