BMC Structural Biology

BioMed Central

Software

Open Access

PDBalert: automatic, recurrent remote homology tracking and protein structure prediction

Vatsal Agarwal[1], Michael Remmert[2], Andreas Biegert[2] and Johannes Söding*[2]

Address: [1]Department of Biotechnology, Indian Institute of Technology, Roorkee 247667, India and [2]Gene Center Munich and Center for Integrated Protein Science (CIPSM), Dept. of Chemistry and Biochemistry, Ludwig-Maximilians-Universtät München, Feodor-Lynen-Str. 25, 81377 Munich, Germany

Email: Vatsal Agarwal - agvatubt@iitr.ernet.in; Michael Remmert - remmert@lmb.uni-muenchen.de; Andreas Biegert - biegert@lmb.uni-muenchen.de; Johannes Söding* - soeding@lmb.uni-muenchen.de

* Corresponding author

Published: 25 November 2008

BMC Structural Biology 2008, **8**:51 doi:10.1186/1472-6807-8-51

Received: 29 July 2008

Accepted: 25 November 2008

This article is available from: http://www.biomedcentral.com/1472-6807/8/51

Abstract

Background: During the last years, methods for remote homology detection have grown more and more sensitive and reliable. Automatic structure prediction servers relying on these methods can generate useful 3D models even below 20% sequence identity between the protein of interest and the known structure (template). When no homologs can be found in the protein structure database (PDB), the user would need to rerun the same search at regular intervals in order to make timely use of a template once it becomes available.

Results: PDBalert is a web-based automatic system that sends an email alert as soon as a structure with homology to a protein in the user's watch list is released to the PDB database or appears among the sequences on hold. The mail contains links to the search results and to an automatically generated 3D homology model. The sequence search is performed with the same software as used by the very sensitive and reliable remote homology detection server HHpred, which is based on pairwise comparison of Hidden Markov models.

Conclusion: PDBalert will accelerate the information flow from the PDB database to all those who can profit from the newly released protein structures for predicting the 3D structure or function of their proteins of interest.

Background

With the advent of remote homology detection methods relying on the pairwise comparison of sequence profiles, automatic protein structure prediction has become reliable and sensitive enough to be of more general use[1]. For more than half of all proteins in representative genomes, at least one domain can be modelled with decent accuracy by fully automatic methods [[2]; J. Soeding, unpublished data]. When no template can be identified, the user will typically rely on keyword tracking services or regular manual checks of the PDB[3] to find out if a related structure has been released. But keyword searches will miss most of the useful templates, since paralogous proteins generally have different names while most will be sufficiently related to serve as templates for homology modelling or to generate hypotheses about possible functions.

Several freely available automatic systems have been developed to perform sequence searches periodically and to notify users about interesting hits. Earlier tools use

BLAST[4] to search Swiss-Prot[5] or the non-redundant sequence database at the NCBI: Swiss-Shop[6], DBWatcher[7], BLAST Search Updater[8], and Sequence Alerting System[9]. FastAlert[10] uses FASTA[11] to search the Swiss-Prot, EMBL data library and GenBank databases. Due to the limitations of the sequence search tools, these services are mainly useful for the detection of closely related sequences. ReHAB[12] and Re-searcher[13] employ the more sensitive method PSI-BLAST[14], but they need to be installed, configured and maintained locally. DbW[15] aims to update user-supplied alignments with homologous and functionally related sequences, using the HMMer method[16] to search Swiss-Prot and TREMBL. Except for Re-searcher, these tools do not provide an option to choose the target database or search parameters, and none except Swiss-shop allows to change preferences later. Most importantly, none of these tools allows to search the PDB database and none makes use of the reliable and considerably more powerful profile-profile comparison tools.

PDBalert is a new web-based automatic system for protein homology detection, which checks the PDB database every week for templates homologous to the proteins in the users' watch lists. PDBalert performs searches with HHpred[17], a very sensitive and reliable remote homology detection server based on pairwise comparison of profile Hidden Markov models (HMMs)[18]. As soon as a homolog to a protein of interest is found in the PDB or among the sequences on-hold that will soon be released to the PDB, the user is notified with an email containing the link to the results page and to a 3D homology model.

Methods

The left part of the flow diagram in Fig. 1 illustrates the steps during and upon uploading of sequences to a user's

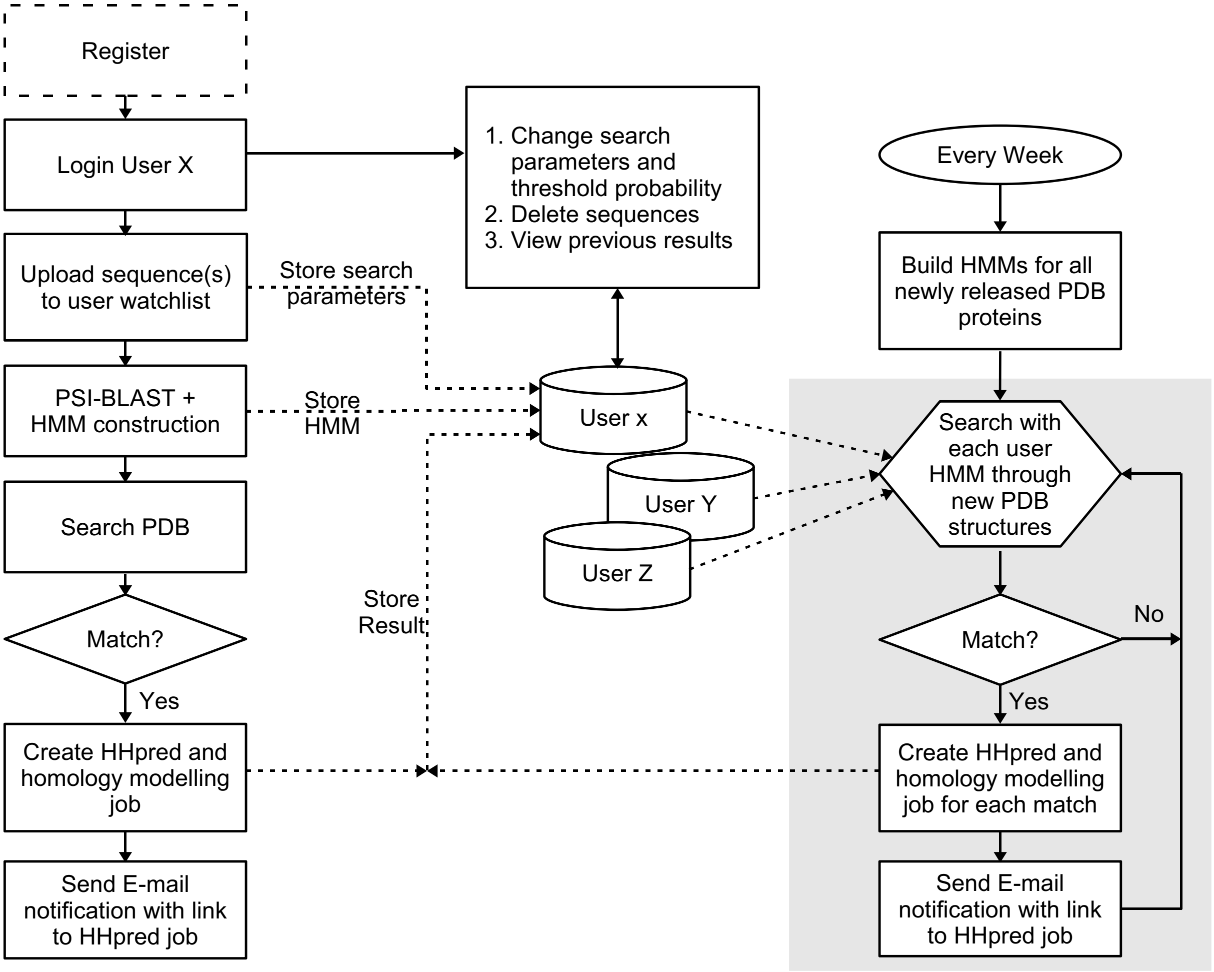
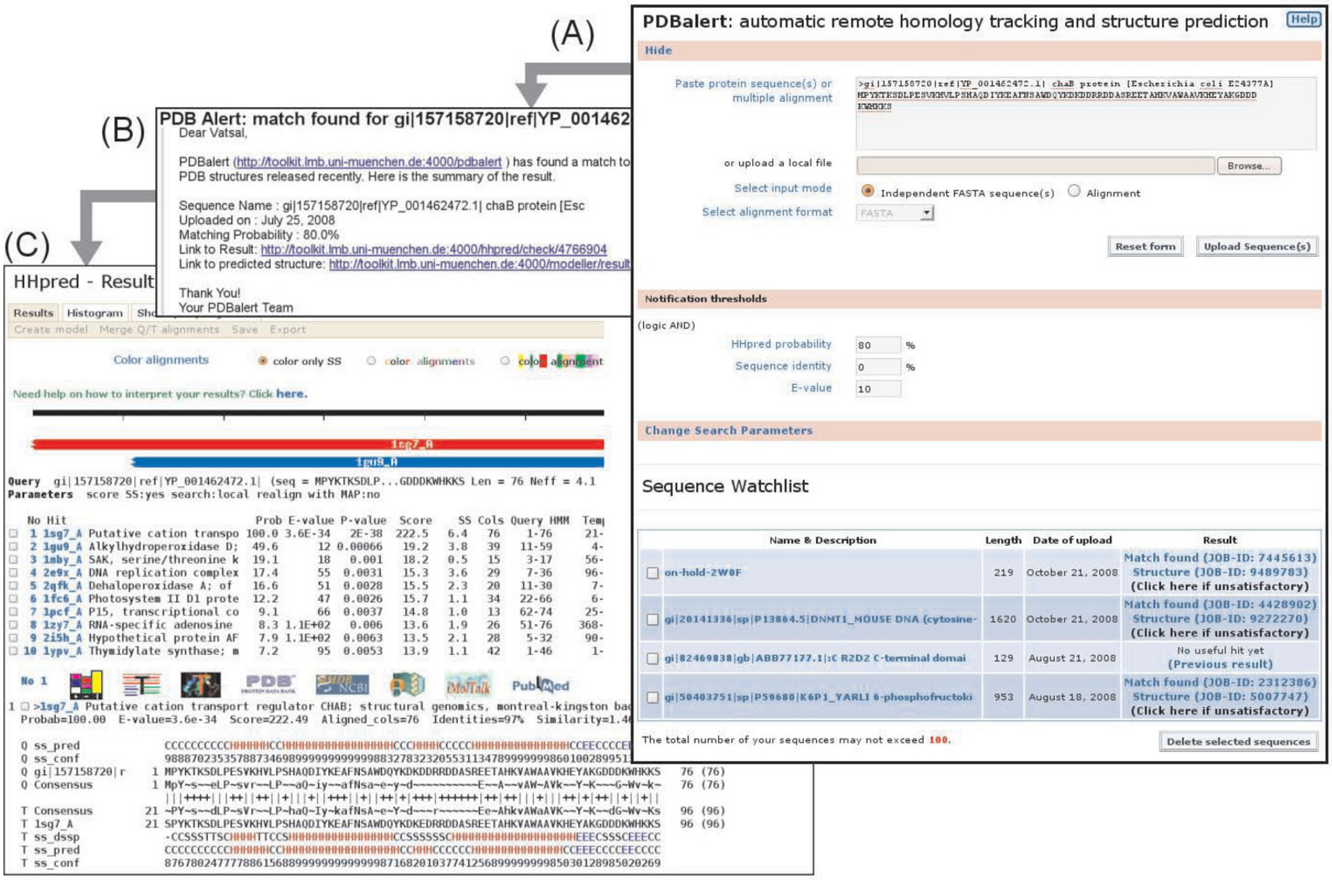

Register
Login User X
Upload sequence(s) to user watchlist
PSI-BLAST + HMM construction
Search PDB
Match?
Yes
Create HHpred and homology modelling job
Send E-mail notification with link to HHpred job
Store search parameters
Store HMM
Store Result
1. Change search parameters and threshold probability
2. Delete sequences
3. View previous results
User x
User Y
User Z
Every Week
Build HMMs for all newly released PDB proteins
Search with each user HMM through new PDB structures
Match?
No
Yes
Create HHpred and homology modelling job for each match
Send E-mail notification with link to HHpred job

Figure 1
PDBalert flow chart (see Methods section).

"watch list", while the right part details the weekly procedure of checking for new hits among the newly released structures. After registering and logging in to the Bioinformatics Toolkit[19] (Fig. 1, left), users can upload protein sequences to their watch lists kept in their accounts (Fig. 2A for a screenshot). Input can be one or more independent FASTA sequences, or a multiple sequence alignment in one of ten common formats. Search parameters may be modified and are kept in a central MySQL database (Fig. 1, middle). Upon uploading a query sequence or alignment, an alignment of homologs is built by the buildali.pl script from the HHsearch package[18], which is also employed in HHpred. Next, a profile HMM is generated from the multiple alignment. The query HMM is then compared using HHsearch with HMMs representatives of all PDB structures and all sequences currently on hold (downloaded from http://www.rcsb.org/pdb/searcsearchStatusDoSearch.do?newSearch=yes&full=true&for mat=SEQ). Three thesholds can be specified by the user to decide when an e-mail notification should be sent (HHpred probability, sequence identity, E-value). If the query protein matches a protein in the PDB (or among the on-hold sequences) according to all three threshold criteria, the user will be notified with an e-mail (see Fig. 2B) containing a link to the results page and to a 3D homology model created by the MODELER package[20] using the HHpred alignment with the newly identified template (Fig. 2C). All results are also stored in the database for at least 6 months. They can be accessed via links in the user's watch list (Fig. 2A), which also allows to add or delete sequences and to change search parameters and threshold probabilities.

Every week, newly released PDB structures are obtained and an HMM for each of them is generated (Fig. 1, right). They are then compared with all sequences in the users' watch lists, and email notifications are sent to those users whose sequences get hits that meet the user-definable threshold criteria.

Whenever possible, users should upload sequences of single protein domains, since sensitivity increases and the



PDBalert: automatic remote homology tracking and structure prediction

Paste protein sequence(s) or multiple alignment:

```
>gi|157158720|ref|YP_001462472.1| chaB protein [Escherichia coli E24377A]
MPYKTKSDLPESVKHVLPSHAQDIYKEAFNSAWDQYKDKDDRRDDASREETAHKVAWAAVKHEYAKGDDD
KWHKKS
```

or upload a local file

Select input mode: ● Independent FASTA sequence(s) ○ Alignment

Select alignment format: FASTA

Notification thresholds

(logic AND)

HHpred probability: 80 %

Sequence identity: 0 %

E-value: 10

Sequence Watchlist

	Name & Description	Length	Date of upload	Result
□	on-hold-2W0F	219	October 21, 2008	Match found (JOB-ID: 7445613) Structure (JOB-ID: 9489783) (Click here if unsatisfactory)
□	gi|20141336|sp|P13864.5|DNMT1_MOUSE DNA (cytosine-	1620	October 21, 2008	Match found (JOB-ID: 4428902) Structure (JOB-ID: 9272270) (Click here if unsatisfactory)
□	gi|82469838|gb|ABB77177.1|:C R2D2 C-terminal domai	129	August 21, 2008	No useful hit yet (Previous result)
□	gi|50403751|sp|P59680|K6P1_YARLI 6-phosphofructoki	953	August 18, 2008	Match found (JOB-ID: 2312386) Structure (JOB-ID: 5007747) (Click here if unsatisfactory)

The total number of your sequences may not exceed 100.

PDB Alert: match found for gi|157158720|ref|YP_001462

Dear Vatsal,

PDBalert (http://toolkit.lmb.uni-muenchen.de:4000/pdbalert) has found a match to
PDB structures released recently. Here is the summary of the result.

Sequence Name : gi|157158720|ref|YP_001462472.1| chaB protein [Esc
Uploaded on : July 25, 2008
Matching Probability : 80.0%
Link to Result: http://toolkit.lmb.uni-muenchen.de:4000/hhpred/check/4766904
Link to predicted structure: http://toolkit.lmb.uni-muenchen.de:4000/modeller/result

Thank You!
Your PDBalert Team

HHpred - Result

Color alignments ● color only SS ○ color alignments ○ color alignment

Need help on how to interpret your results? Click here.

```
Query  gi|157158720|ref|YP_001462472.1| (seq = MPYKTKSDLP...GDDDKWHKKS Len = 76 Neff = 4.1
Parameters  score SS:yes search:local realign with MAP:no

   No Hit                           Prob E-value P-value  Score    SS Cols Query HMM  Tem
□  1 1sg7_A Putative cation transpo 100.0 3.6E-34   2E-38  222.5   6.4   76    1-76   21-
□  2 1gu9_A Alkylhydroperoxidase D;  49.6      12 0.00066   19.2   3.8   39   11-59    4-
□  3 1mby_A SAK, serine/threonine k  19.1      18   0.001   18.2   0.5   15    3-17   56-
□  4 2e9x_A DNA replication complex  17.4      55  0.0031   15.3   3.6   29    7-36   96-
□  5 2qfk_A Dehaloperoxidase A; of   16.6      51  0.0028   15.5   2.3   20   11-30    7-
□  6 1fc6_A Photosystem II D1 prote  12.2      47  0.0026   15.7   1.1   34   22-66    6-
□  7 1pcf_A P15, transcriptional co   9.1      66  0.0037   14.8   1.0   13   62-74   25-
□  8 1zy7_A RNA-specific adenosine    8.3 1.1E+02   0.006   13.6   1.9   26   51-76  368-
□  9 2i5h_A Hypothetical protein AF   7.9 1.1E+02  0.0063   13.5   2.1   28    5-32   90-
□ 10 1ypv_A Thymidylate synthase; m   7.2      95  0.0053   13.9   1.1   42    1-46    1-

1 □ >1sg7_A Putative cation transport regulator CHAB; structural genomics, montreal-kingston ba
  Probab=100.00  E-value=3.6e-34  Score=222.49  Aligned_cols=76  Identities=97%  Similarity=1.4

 Q ss_pred             CCCCCCCCCHHHHHHCCHHHHHHHHHHHHHHHHHHHCCCHHHHCCCCCHHHHHHHHHHHHHHHHCCEECCCCE
 Q ss_conf             98887023535788734698999999999999988327832320553113478999999998601002899951
 Q gi|157158720|r    1 MPYKTKSDLPESVKHVLPSHAQDIYKEAFNSAWDQYKDKDDRRDDASREETAHKVAWAAVKHEYAKGDDDKWHKKS   76 (76)
 Q Consensus         1 MpY~s~~eLP~svr~~LP~~aQ~iy~~afNsa~e~y~d~~~~~~~~~~~~E~~A~~vAW~AVk~~Y~K~~~G~Wv~k~   76 (76)
                       |||++++|||++||++||+|||+||+++||+|++|+|+++|++++++|++|++||||+|+||++|+|+|+|||
 T Consensus        21 ~PY~s~~dLP~sVr~~LP~haQ~Iy~kafNsA~e~Y~d~~~r~~~~~~Ee~AhkvAWaAVK~~Y~K~~dG~Wv~Ks   96 (96)
 T 1sg7_A           21 SPYKTKSDLPESVKHVLPSHAQDIYKEAFNSAWDQYKDKEDRRDDASREETAHKVAWAAVKHEYAKGDDDKWHKKS   96 (96)
 T ss_dssp             -CCSSSTTSCHHHHTTCCSHHHHHHHHHHHHHHHHHCCSSSSSSCHHHHHHHHHHHHHHHHHHHEEECSSSCEEECC
 T ss_pred             CCCCCCCCCHHHHHHCCHHHHHHHHHHHHHHHHHHHHCCHHHCCCCCHHHHHHHHHHHHHHHCCEEECCCCEECCCC
 T ss_conf             8767802477778861568899999999999998716820103774125689999999985030128985020269
```

Figure 2

Representative screen shots. (A) PDBalert web interface with sequence upload section and personal watch list. (B) Email alert sent when significant hit is detected. (C) HHpred Results page containing alignment to PDBalert match.

false discovery is rate reduced compared to multiple domains. When PDBalert confidently predicts a domain in a longer sequence, it is therefore recommended to split the sequence at the boundaries of the discovered domain and upload the segments separately to PDBalert. In practice, it may be useful to leave some overlap of up to 30 residues between the segments when domain boundaries are not precisely known.

The web-interface of PDBalert is built on a Ruby on Rails[21] architecture on a Linux platform together with a MySQL[22] database for storing user inputs and preferences. Users do not require anything except a web-browser. PDBalert is integrated into the Bioinformatics Toolkit, a user-friendly web system of interlinked tools for protein sequence analysis and structure prediction.

Discussion

The biannual CASP benchmarks[1] as well as the many studies employing state-of-the-art remote homology detection and structure prediction servers such as FFAS[23], HHpred[17], SAM-T2K[24], 3DJury[25], and I-TASSER[26] testify to the usefulness of these automatic methods. However, we believe that their full potential is far from being fully exploited. The principle reasons are that (1) innovations take time to spread; (2) most servers do not have user-friendly interfaces nor help pages; (3) Only few servers provide reliable significance estimates; (4) The servers are generally too slow to allow one to wait for the results on-line, taking hours or days to finish and discouraging usage on a regular basis. PDBalert addresses the last point in particular, by noting that most biologists and biochemists will have a fairly limited and conserved set of proteins in the focus of their attention. PDBalert saves these users the time to periodically redo searches for new templates to these proteins.

Conclusion

The usefulness of PDBalert is owed to a large extent to the power of its underlying remote homology detection and structure prediction protocols, borrowed from HHpred. Two fully automated versions of HHpred that use the same homology detection method as PDBalert were ranked 2nd (HHpred2, multiple template modelling) and 8th (HHpred1, single template modelling, used by PDBalert to build a model with the detected template) out of a total of 68 automatic servers in the last community-wide protein structure prediction benchmark CASP7[1], while being more than 50 times faster than the other top servers. This speed allows to offer remote homology detection and structure prediction services for an automatic recurrent search to a wider community. We hope that PDBalert will encourage many more biologists to profit from recent advances in remote homology detection and structure prediction.

Availability and requirements

- **Project name**: PDBalert

- **Project home page**: http://toolkit.lmb.uni-muenchen.de/pdbalert/

- **Operating system(s)**: Platform independent (web service)

- **Programming language**: Ruby

- **Licence**: None (Freely available to all academic and non-academic users)

Authors' contributions

VA developed the PDBalert system, MR integrated the on-hold sequence database, AB, MR and JS coordinated the development and tested the application, JS conceived of the project, and VA and JS wrote the mansucript. All authors read and approved the final manuscript.

Acknowledgements

VA would like to thank the Deutscher Akademischer Austausch Dienst for financing his internship at the Ludwig-Maximilians-University Munich.

References

1. Battey JN, Kopp J, Bordoli L, Read RJ, Clarke ND, Schwede T: **Automated server predictions in CASP7.** *Proteins* 2007, **69(Suppl 8):**68-82.
2. Pawlowski K, Zhang B, Godzik A: **The Helicobacter pylori genome: from sequence analysis to structural and functional predictions.** *Proteins* 1999, **36:**20-30.
3. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: **The Protein Data Bank.** *Nucl Acids Res* 2000, **28:**235-242.
4. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215:**403-410.
5. Bairoch A, Boeckmann B, Ferro S, Gasteiger E: **Swiss-Prot: juggling between evolution and stability.** *Brief Bioinform* 2004, **5:**39-55.
6. **Swiss-Shop** [http://www.expasy.org/swiss-shop/]
7. **DBWatcher** [ftp://ftp-igbmc.u-strasbg.fr/pub/DBWatcher/]
8. Boone M, Upton C: **BLAST Search Updater: a notification system for new database matches.** *Bioinformatics* 2000, **16:**1054-1055.
9. Hegyi H, Lai JM, Bork P: **The Sequence Alerting Server – a new WEB server.** *Comput Appl Biosci* 1997, **13:**619-620.
10. Eggenberger F, Redaschi N, Doelz R: **FastAlert – an automatic search system to alert about new entries in biological sequence databanks.** *Comput Appl Biosci* 1996, **12:**129-133.
11. Pearson WR: **Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms.** *Genomics* 1991, **11:**635-650.
12. Whitney J, Esteban DJ, Upton C: **Recent Hits Acquired by BLAST (ReHAB): a tool to identify new hits in sequence similarity searches.** *BMC Bioinformatics* 2005, **6:**23.
13. Repsys V, Margelevicius M, Venclovas C: **Re-searcher: a system for recurrent detection of homologous protein sequences.** *BMC Bioinformatics* 2008, **9:**296.
14. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman D: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucl Acids Res* 1997, **25:**3389-3402.
15. Prigent V, Thierry JC, Poch O, Plewniak F: **DbW: automatic update of a functional family-specific multiple alignment.** *Bioinformatics* 2005, **21:**1437-1442.
16. Eddy SR: **Profile hidden Markov models.** *Bioinformatics* 1998, **14:**755-763.

17. Söding J, Biegert A, Lupas AN: **The HHpred interactive server for protein homology detection and structure prediction.** *Nucl Acids Res* 2005, **33:**W244-248.
18. Söding J: **Protein homology detection by HMM-HMM comparison.** *Bioinformatics* 2005, **21:**951-960.
19. Biegert A, Mayer C, Remmert M, Söding J, Lupas AN: **The MPI Bioinformatics Toolkit for protein sequence analysis.** *Nucl Acids Res* 2006, **34:**W335-339 [http://toolkit.lmb.uni-muenchen.de/].
20. Sali A, Blundell TL: **Comparative protein modelling by satisfaction of spatial restraints.** *J Mol Biol* 1993, **234:**779-815.
21. **Ruby on rails** [http://www.rubyonrails.org/]
22. **MySQL** [http://www.mysql.com/]
23. Jaroszewski L, Rychlewski L, Li Z, Li W, Godzik A: **FFAS03: a server for profile – profile sequence alignments.** *Nucl Acids Res* 2005, **33:**W284-288.
24. Karplus K, Karchin R, Barrett C, Tu S, Cline M, Diekhans M, Grate L, Casper J, Hughey R: **What is the value added by human intervention in protein structure prediction?** *Proteins* 2001:86-91.
25. Ginalski K, Elofsson A, Fischer D, Rychlewski L: **3D-Jury: a simple approach to improve protein structure predictions.** *Bioinformatics* 2003, **19:**1015-1018.
26. Zhang Y: **I-TASSER server for protein 3D structure prediction.** *BMC Bioinformatics* 2008, **9:**40.