Nucleic Acids Research Advance Access published May 12, 2009

*Nucleic Acids Research, 2009,* **1–5**
*doi:10.1093/nar/gkp360*

# COMPASS server for homology detection: improved statistical accuracy, speed and functionality

**Ruslan I. Sadreyev[1],*, Ming Tang[1], Bong-Hyun Kim[2] and Nick V. Grishin[1,2]**

[1]Howard Hughes Medical Institute and [2]Department of Biochemistry, University of Texas Southwestern Medical Center, 5323 Harry Hines Blvd, Dallas, TX 75390-9050, USA

Received March 10, 2009; Revised and Accepted April 23, 2009

## ABSTRACT

**COMPASS is a profile-based method for the detection of remote sequence similarity and the prediction of protein structure. Here we describe a recently improved public web server of COMPASS, http://prodata.swmed.edu/compass. The server features three major developments: (i) improved statistical accuracy; (ii) increased speed from parallel implementation; and (iii) new functional features facilitating structure prediction. These features include visualization tools that allow the user to quickly and effectively analyze specific local structural region predictions suggested by COMPASS alignments. As an application example, we describe the structural, evolutionary and functional analysis of a protein with unknown function that served as a target in the recent CASP8 (Critical Assessment of Techniques for Protein Structure Prediction round 8). URL: http://prodata.swmed.edu/compass**

## INTRODUCTION

Comparison of multiple sequence alignments (MSA) leads to a significant increase in the quality of detecting relationships between remote protein homologs (1–3). Analysis of MSAs of two protein families, in the form of numerical profiles (4) or hidden Markov models (5), can reveal similar sequence patterns that reflect the evolutionary constraints dictated by common structural folds or functions. These patterns often remain detectable long after individual protein sequences diverge beyond recognition. In recent years, methods for profile–profile (2,6–11) and HMM-HMM (1,3,12) comparison, often featured on public web servers (13–17), have become a powerful addition to other web-based tools for predicting protein structure and function [e.g. (18–21)].

Suggested as a generalization of PSI-BLAST (22), COMPASS is a method for remote homology detection that generates numerical profiles, constructs optimal profile–profile alignments and estimates the statistical significance of the corresponding alignment scores (2). COMPASS has been successfully used by our group and other researchers for both prediction of protein structure and function [e.g. (23,24)] and evolutionary analysis [e.g. (12,25,26)].

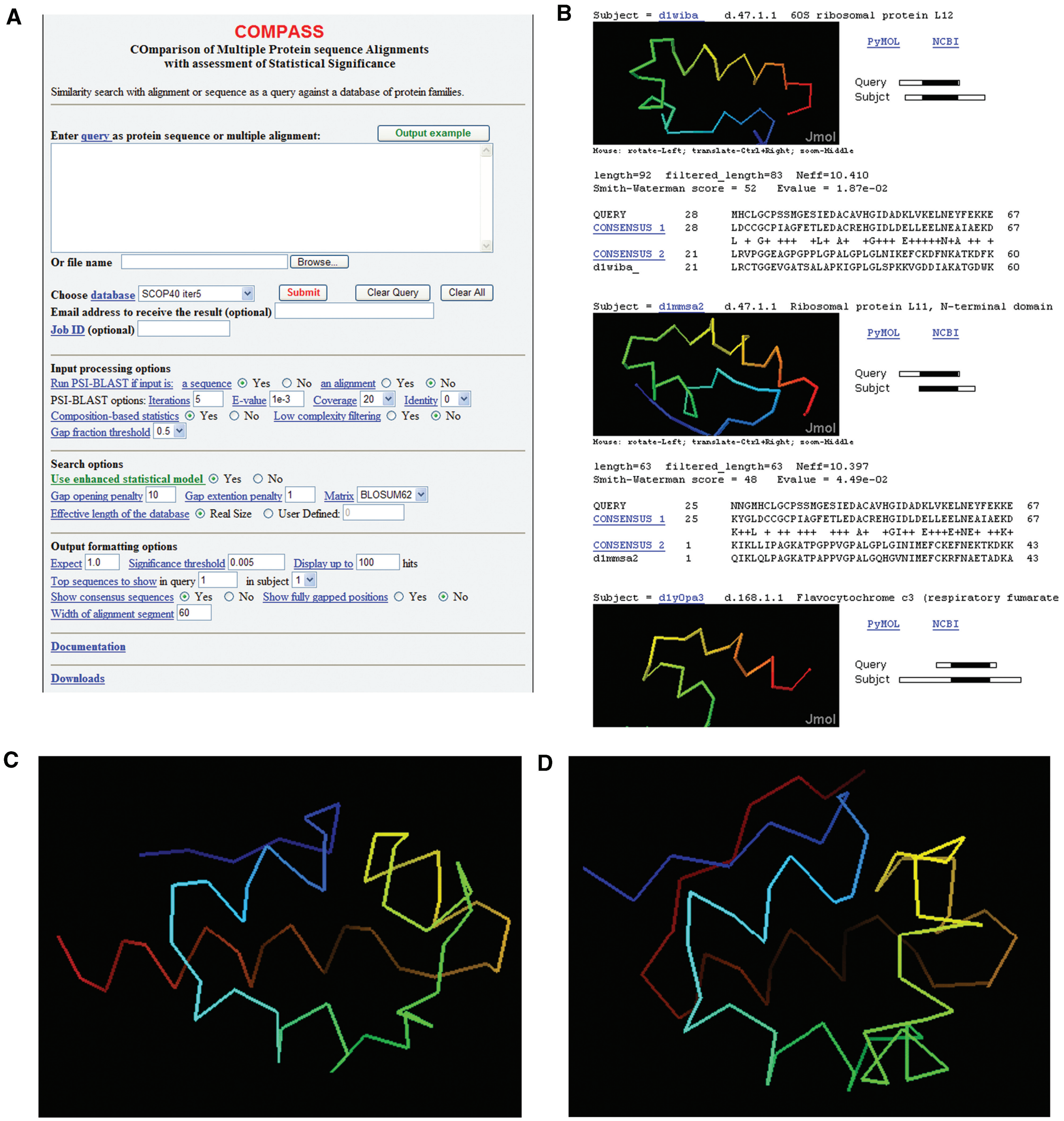
The COMPASS web server has been publicly available since 2007 (15). Based on user experiences, we have identified three major directions of further development in methodology, implementation and user interface. First, we improved the accuracy of statistical significance estimates (*E*-values) and the corresponding ranking of detected similarities by introducing a more realistic random model of profile comparison (27). This model is based on an accurate statistical description of MSA comparison that captures the essential features of protein families, as opposed to traditional models originally derived for the comparison of individual sequences. Second, we increased the computational speed by developing a parallelized version of COMPASS that runs on multiple processors. Third, we designed a new interface for a more comprehensive analysis of produced results. Unlike similar tools, the new COMPASS server allows for a fast and easy mapping and inspection of the specific regions of protein structure predicted by profile alignments. As a result, the presented server is a more accurate and fast tool for remote homology detection and structure prediction, with improved output that facilitates manual analysis and discovery of intricate protein relationships.

## METHODS

Given a sequence or MSA, COMPASS generates the query profile and compares it against a specified database of protein family profiles. The server's front page (Figure 1A) allows the user to upload the query, select the database [PFAM (28), COG, KOG (29), or PSI-BLAST alignments produced from PDB70 (30) or SCOP40 (31) representatives], and adjust the search parameters and output options.

The output of the search includes a list of the most significant hits and their corresponding optimal

*To whom correspondence should be addressed. Tel: 214 645 5951; Fax: 214 645 5948; Email: sadreyev@chop.swmed.edu



**COMPASS**
**COmparison of Multiple Protein sequence Alignments with assessment of Statistical Significance**

Similarity search with alignment or sequence as a query against a database of protein families.

**Enter query as protein sequence or multiple alignment:**

**Or file name**

**Choose database** SCOP40 iter5
**Email address to receive the result (optional)**
**Job ID (optional)**

**Input processing options**
Run PSI-BLAST if input is: a sequence ● Yes ○ No an alignment ○ Yes ● No
PSI-BLAST options: Iterations 5 E-value 1e-3 Coverage 20 Identity 0
Composition-based statistics ● Yes ○ No Low complexity filtering ○ Yes ● No
Gap fraction threshold 0.5

**Search options**
Use enhanced statistical model ● Yes ○ No
Gap opening penalty 10 Gap extention penalty 1 Matrix BLOSUM62
Effective length of the database ● Real Size ○ User Defined: 0

**Output formatting options**
Expect 1.0 Significance threshold 0.005 Display up to 100 hits
Top sequences to show in query 1 in subject 1
Show consensus sequences ● Yes ○ No Show fully gapped positions ○ Yes ● No
Width of alignment segment 60

**Documentation**

**Downloads**

```
Subject = d1wiba_   d.47.1.1  60S ribosomal protein L12
Mouse: rotate-Left; translate-Ctrl+Right; zoom-Middle

length=92  filtered_length=83  Neff=10.410
Smith-Waterman score = 52   Evalue = 1.87e-02

QUERY          28     MHCLGCPSSMGESIEDACAVHGIDADKLVKELNEYFEKKE  67
CONSENSUS 1    28     LDCCGCPIAGFETLEDACREHGIDLDELLEELNEAIAEKD  67
                      L + G+ +++  +L+ A+  +G+++ E+++++N+A ++ +
CONSENSUS 2    21     LRVPGGEAGPGPPLGPALGPLGLNIKEFCKDFNKATKDFK  60
d1wiba_        21     LRCTGGEVGATSALAPKIGPLGLSPKKVGDDIAKATGDWK  60


Subject = d1mmsa2   d.47.1.1  Ribosomal protein L11, N-terminal domain
Mouse: rotate-Left; translate-Ctrl+Right; zoom-Middle

length=63  filtered_length=63  Neff=10.397
Smith-Waterman score = 48   Evalue = 4.49e-02

QUERY          25     NNGMHCLGCPSSMGESIEDACAVHGIDADKLVKELNEYFEKKE  67
CONSENSUS 1    25     KYGLDCCGCPIAGFETLEDACREHGIDLDELLEELNEAIAEKD  67
                      K++L + ++ +++  +++ A+  +GI++ E+++E+NE+ ++K+
CONSENSUS 2    1      KIKLLIPAGKATPGPPVGPALGPLGINIMEFCKEFNEKTKDKK  43
d1mmsa2        1      QIKLQLPAGKATPAPPVGPALGQHGVNIMEFCKRFNAETADKA  43


Subject = d1y0pa3   d.168.1.1  Flavocytochrome c3 (respiratory fumarate
```

**Figure 1.** (**A**) Submission form (front page) of the COMPASS server. The main section allows the user to submit the query (as MSA or a single sequence), choose the database, and enter the email address to receive results (optional). In the section of input processing, the user can require an additional PSI-BLAST run enriching the query profile and define the parameters of profile construction. The sections of search and output options contain controls of the main parameters of the search and of the output format. As a separate control, a choice of statistical models is available (see text for details). A brief explanation of each option can be viewed by clicking on the option's name. The links to more detailed documentation and to a downloadable standalone package are on the bottom of the page. (**B**) Example of output alignment sections with matched structural fragments displayed in Jmol, graphic schemas of alignments and additional links. Several sub-threshold SCOP hits for CASP target T0473 are shown. Although these domains belong to different SCOP folds, they all share an HTH motif that becomes apparent when matched fragments are displayed. (**C**) Ribbon diagram of CASP target T0473 (PDB ID 2k53, NMR model 1). The two C-terminal α-helices (yellow and orange-red) form the HTH motif. (**D**) Ribbon diagram of the only full-length template of T0473 (PDB ID 2fi0A), protein SP0561 of unknown function classified in SCOP as a unique SP0561-like fold.

profile–profile alignments. According to comprehensive evaluations by our group (32) and others (3,12,33), COMPASS is among the top-performing methods for sequence-based homology detection and local alignment construction. The presented web server features several major modifications to improve detection quality, speed and availability of specific structure predictions for visual analysis.

### Improved accuracy of estimates of statistical significance

The server features our novel approach to statistical modeling of MSA comparison (27) that provides a more accurate discrimination between biologically meaningful protein similarities and spurious hits.

The elegant theoretical model (34) developed for single sequence comparison assumes an absence of correlation between amino-acid content at individual positions. As a result, the statistical significance can be estimated using extreme value distribution (EVD) (35,36). This model corresponds to a randomization of profiles by shuffling positions, which allows for generation of potentially unlimited statistical samples and, therefore, precise analysis of resulting score distributions. However, this representation of real alignments is overly simplistic. An alternative empirical approach of comparing real unrelated proteins (3,6,7,10,11) requires an additional calibration on a database of diverse protein representatives, and can generate only a limited random sample of similarity scores, which hinders the precise analytical fitting of the empirical score distribution. We developed a new modeling approach that combines a realistic representation of essential protein features with a precise mathematical description. This approach is based on (i) a biologically meaningful modeling of the secondary structure and of the resulting correlations between sequence positions and (ii) a precise analytical description of the empirical statistics (27).

In particular, we found that the score distribution for the comparison of unrelated profiles can be realistically modeled by reproducing major classes of real proteins (such as all $\alpha$, all $\beta$, $\alpha/\beta$ and $\alpha + \beta$ classes in the SCOP classification) and by mimicking the types, lengths and sequence diversity (thickness) of real profile fragments corresponding to secondary structure elements (27). Since the resulting simulated score distributions do not follow EVD, we proposed a novel distribution, 'power EVD' (27), that yields statistically perfect agreement with the data. In an evaluation based on a set of existing protein families with known structure, our model surpassed currently available models in the accuracy of detecting remote protein similarities. In addition, this model has a realistic statistical accuracy, i.e. the closeness between predicted and actual numbers of false positives that are assigned a given *E*-value (27).

The server uses the new statistical model by default; however, the old model is also available. The type of statistics is selected by switching the corresponding radio button on the front page.

### Faster parallelized implementation

To speed up the computationally costly process of constructing profile–profile alignments, we developed a parallelized version of COMPASS based on the MPI platform. In brief, comparisons of the query to database entries are distributed among multiple processors, so that each processor works on a subset of the database profiles. Data about the most significant hits are sent to the manager node, which accumulates the hit list. When the size of the growing list exceeds the user-defined number of hits to display, the manager modifies the *E*-value cutoff used by the worker nodes to decide whether the hit is significant enough to be sent to the manager. This cutoff is set to the worst *E*-value in the current hit list, so that the worker nodes do not spend time on reporting hits that will not be displayed. This implementation increases the computational speed by approximately an order of magnitude, so that a typical running time is within a minute. This time may increase when the server is heavily loaded or when the user requires generation of the query profile by PSI-BLAST search, which may take longer for queries with a large number of homologs in the sequence database.

### Improved interface to facilitate expert analysis of structure predictions

In structure prediction, it is important to visualize the exact structural region of the potential homolog that is predicted to be similar to the query. In most currently available public servers, the structure of the detected homolog is not displayed at all; when it is displayed, the full protein chain is shown. However, local sequence alignments often reveal similarity restricted to a single structural domain, subdomain, or even functional motif. This information is especially important when a query detects multiple homologs with highly divergent structures that share only locally similar regions (see an example below). In such cases, the consistency of structure predictions would not be apparent if one views only full-chain homolog structures.

In all current servers for homology detection, a user wishing to visualize the actual region of predicted structural similarity has to retrieve the full 3D structure of the detected homolog and then manually map the aligned sequence part on the structure. The improved COMPASS server automates this important yet tedious process. For the top significant hits in databases of proteins with known 3D structure, an additional panel is displayed in the alignment section (Figure 1B). This section includes several items. First, a Jmol (http://jmol.sourceforge.net) panel is used to interactively display the C-alpha trace of the structural fragment covered by the COMPASS alignment. The user can rotate, move and zoom this fragment; position numbers and residue names can be viewed by moving the mouse over the residues. Second, the user can analyze the structure in more detail, either by downloading the fragment as an all-atom PDB file, or by clicking on the 'PyMOL' link, which generates and launches a PyMOL (http://pymol.sourceforge.net) script to show, in a separate window, the full structure of the potential homolog, with the aligned region highlighted. In this window, the user can employ the full functionality of PyMOL to view sidechains, ligands, protein surface, atom–atom distances, etc. Third, the additional links to databases allow viewing the information about the protein structure in PDB, SCOP, CATH and NCBI repository. Finally, a simple graphical schema shows the location of aligned sequence regions with respect to the full sequences of the query and subject. Full query and subject MSAs can be viewed by clicking on either the bars representing the profiles in the schema or the highlighted sequence names in the profile–profile alignment.

The ability to view and compare 'on the fly' multiple predicted regions of structural similarity simplifies the manual analysis of search results. As shown in the example below, this feature helps visualize important structural and functional features of the query that could be missed otherwise.

## EXAMPLE OF STRUCTURE PREDICTION AND EVOLUTIONARY ANALYSIS

As an illustration, we describe detection of distant sequence similarities that leads to structure prediction and reveals an evolutionarily conserved functional motif in a target protein from the recent CASP8 (Critical Assessment of Techniques for Protein Structure Prediction). Target T0473, an all-alpha protein from *Clostridium thermocellum*, is a member of a tight sequence family with unknown function. PSI-BLAST, up to five iterations, could not detect significant sequence similarities to proteins with solved 3D structures. When a PSI-BLAST alignment of T0473 was used as a query in a COMPASS search against the database of SCOP representatives, a domain of known structure (PDB ID 2fi0) was easily detected with a low *E*-value of $10^{-20}$ (Figure 1D). This homolog, aligned to the query over the whole domain length, is a hypothetical protein SP0561 of unknown function classified in SCOP as a unique SP0561-like fold. This prediction was additionally confirmed by other profile–profile comparison servers, such as HHSearch (assigned a high probability of 99%). After the structure of this CASP target was released (PDB ID 2k53), it proved to be highly similar to the template (Figure 1C).

Although this structure prediction is relatively straightforward, it does not provide much information about the potential function and evolutionary history of this protein: the only template covering the whole domain is a hypothetical protein. However, this domain contains an important functional motif that can be easily detected with the new COMPASS feature of displaying precise structural fragments that match the query in potential homologs.

The presence of the motif becomes apparent when viewing multiple hits in the SCOP database with *E*-values below and slightly above the significance cutoff (Figure 1B). In these otherwise highly divergent domains from different SCOP folds, the query consistently matches a helix-turn-helix (HTH) motif that is easily recognized in structure panels of the corresponding alignment sections (Figure 1B). All structural fragments have the characteristic helix orientation and positioning of the connecting turn, and include conserved residues typical of HTH (Figure 1B). Inspection of the query's structure confirmed that the last two α-helices (yellow and orange-red, Figure 1C) form the HTH motif, readily recognizable by an almost perpendicular orientation of the two helices and a long extended 'turn' that positions the N-terminus of the second helix close to the middle of the first helix. The 49-Gly-Ile-Asp-51 in T0473 is a very typical sequence for an HTH turn, and Ala44, four residues before Gly49, is a conserved small residue to make room for packing of the second HTH helix (37).

However, the similarity is restricted to only the HTH part: the way it is completed is quite different among more distant homologs. The rest of the structure is unique and shared only with the closest homologs, the detected full-chain template 2fi0 and another CASP8 target T0469 (PDB ID 2k5e). This N-terminal segment is structured as three helices and completes the HTH core in a manner similar to the packing of a single N-terminal helix against the HTH in a classic three-helical bundle fold.

Thus T0473 is an unusual helix-turn-helix (HTH) containing protein, which suggests its functional role in nucleic acid or protein binding. As another indication of a potential regulatory function, COMPASS detects profile similarity to COG2846 in the COG database (not shown), which is predicted to be a 'Regulator of cell morphogenesis and NO signaling'.

## ACKNOWLEDGEMENTS

The authors acknowledge the Texas Advanced Computing Center (TACC) at The University of Texas at Austin for providing high-performance computing resources. We would like to thank Lisa Kinch for discussions and critical reading of the manuscript.

## FUNDING

National Institutes of Health [grant GM67165 to NVG]. Funding for open access charge: Howard Hughes Medical Institute.

*Conflict of interest statement*. None declared.

## REFERENCES

1. Madera,M. (2008) Profile comparer: a program for scoring and aligning profile hidden Markov models. *Bioinformatics*, **24**, 2630–2631.
2. Sadreyev,R.I. and Grishin,N.V. (2003) COMPASS: a tool for comparison of multiple protein alignments with assessment of statistical significance. *J. Mol. Biol.*, **326**, 317–336.
3. Soding,J. (2005) Protein homology detection by HMM-HMM comparison. *Bioinformatics*, **21**, 951–960.
4. Gribskov,M., McLachlan,A.D. and Eisenberg,D. (1987) Profile analysis: detection of distantly related proteins. *Proc. Natl Acad. Sci. USA*, **84**, 4355–4358.
5. Durbin,R.E., Krogh,A., Mitchison,G. and Eddy,S. (1999) *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge, UK.
6. Ginalski,K., Pas,J., Wyrwicz,L.S., von Grotthuss,M., Bujnicki,J.M. and Rychlewski,L. (2003) ORFeus: Detection of distant homology using sequence profiles and predicted secondary structure. *Nucleic Acids Res.*, **31**, 3804–3807.
7. Kahsay,R.Y., Wang,G., Gao,G., Liao,L. and Dunbrack,R. (2005) Quasi-consensus-based comparison of profile hidden Markov models for protein sequences. *Bioinformatics*, **21**, 2287–2293.
8. Ohlson,T., Wallner,B. and Elofsson,A. (2004) Profile-profile methods provide improved fold-recognition: a study of different profile-profile alignment methods. *Proteins*, **57**, 188–197.
9. Pietrokovski,S. (1996) Searching databases of conserved sequence regions by aligning protein multiple-alignments. *Nucleic Acids Res.*, **24**, 3836–3845.

10. Rychlewski,L., Jaroszewski,L., Li,W. and Godzik,A. (2000) Comparison of sequence profiles. Strategies for structural predictions using sequence information. *Protein Sci.*, **9**, 232–241.
11. Yona,G. and Levitt,M. (2002) Within the twilight zone: a sensitive profile-profile comparison tool based on information theory. *J. Mol. Biol.*, **315**, 1257–1275.
12. Reid,A.J., Yeats,C. and Orengo,C.A. (2007) Methods of remote homology detection can be combined to increase coverage by 10% in the midnight zone. *Bioinformatics*, **23**, 2353–2360.
13. Frenkel-Morgenstern,M., Singer,A., Bronfeld,H. and Pietrokovski,S. (2005) One-Block CYRCA: an automated procedure for identifying multiple-block alignments from single block queries. *Nucleic Acids Res.*, **33**, W281–W283.
14. Jaroszewski,L., Rychlewski,L., Li,Z., Li,W. and Godzik,A. (2005) FFAS03: a server for profile–profile sequence alignments. *Nucleic Acids Res*, **33**, W284–W288.
15. Sadreyev,R.I., Tang,M., Kim,B.H. and Grishin,N.V. (2007) COMPASS server for remote homology inference. *Nucleic Acids Res.*, **35**, W653–W658.
16. 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.
17. 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.
18. Chivian,D., Kim,D.E., Malmstrom,L., Schonbrun,J., Rohl,C.A. and Baker,D. (2005) Prediction of CASP6 structures using automated Robetta protocols. *Proteins*, **61(Suppl. 7)**, 157–166.
19. Ginalski,K., Elofsson,A., Fischer,D. and Rychlewski,L. (2003) 3D-Jury: a simple approach to improve protein structure predictions. *Bioinformatics*, **19**, 1015–1018.
20. Kelley,L.A., MacCallum,R.M. and Sternberg,M.J. (2000) Enhanced genome annotation using structural profiles in the program 3D-PSSM. *J. Mol. Biol.*, **299**, 499–520.
21. Zhang,Y. (2008) I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics*, **9**, 40.
22. 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.
23. Raman,S., Vernon,R., Thompson,J., Tyka,M., Sadreyev,R.I., Pei,J., Kim,D., Kellogg,E., DiMaio,F., Lange,O. *et al.* (2009) Structure prediction for CASP8 with all-atom refinement using Rosetta3. *Proteins* (in press).
24. Yarbrough,M.L., Li,Y., Kinch,L.N., Grishin,N.V., Ball,H.L. and Orth,K. (2009) AMPylation of Rho GTPases by Vibrio VopS disrupts effector binding and downstream signaling. *Science*, **323**, 269–272.
25. Cheng,H., Kim,B.H. and Grishin,N.V. (2008) Discrimination between distant homologs and structural analogs: lessons from manually constructed, reliable data sets. *J. Mol. Biol.*, **377**, 1265–1278.
26. Theobald,D.L. and Wuttke,D.S. (2005) Divergent evolution within protein superfolds inferred from profile-based phylogenetics. *J. Mol. Biol.*, **354**, 722–737.
27. Sadreyev,R.I. and Grishin,N.V. (2008) Accurate statistical model of comparison between multiple sequence alignments. *Nucleic Acids Res.*, **36**, 2240–2248.
28. Finn,R.D., Tate,J., Mistry,J., Coggill,P.C., Sammut,S.J., Hotz,H.R., Ceric,G., Forslund,K., Eddy,S.R., Sonnhammer,E.L. *et al.* (2008) The Pfam protein families database. *Nucleic Acids Res*, **36**, D281–D288.
29. Tatusov,R.L., Natale,D.A., Garkavtsev,I.V., Tatusova,T.A., Shankavaram,U.T., Rao,B.S., Kiryutin,B., Galperin,M.Y., Fedorova,N.D. and Koonin,E.V. (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.*, **29**, 22–28.
30. 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.
31. Andreeva,A., Howorth,D., Chandonia,J.M., Brenner,S.E., Hubbard,T.J., Chothia,C. and Murzin,A.G. (2008) Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res.*, **36**, D419–D425.
32. Qi,Y., Sadreyev,R.I., Wang,Y., Kim,B.H. and Grishin,N.V. (2007) A comprehensive system for evaluation of remote sequence similarity detection. *BMC Bioinformatics*, **8**, 314.
33. Frenkel-Morgenstern,M., Voet,H. and Pietrokovski,S. (2005) Enhanced statistics for local alignment of multiple alignments improves prediction of protein function and structure. *Bioinformatics*, **21**, 2950–2956.
34. Karlin,S. and Altschul,S.F. (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl Acad. Sci. USA*, **87**, 2264–2268.
35. Gnedenko,B. (1943) Sur la distribution limite du terme maximum d'une serie aleatoire. *Ann. Mathematics*, **44**, 423–453.
36. Gumbel,E.J. (1958) *Statistics of Extremes*. Columbia University Press, New York, NY.
37. Aravind,L., Anantharaman,V., Balaji,S., Babu,M.M. and Iyer,L.M. (2005) The many faces of the helix-turn-helix domain: transcription regulation and beyond. *FEMS Microbiol. Rev.*, **29**, 231–262.