W532–W538 *Nucleic Acids Research, 2009, Vol. 37, Web Server issue* *Published online 5 May 2009*
doi:10.1093/nar/gkp328

HorA web server to infer homology between proteins using sequence and structural similarity

Bong-Hyun Kim[1], Hua Cheng[1,2] and Nick V. Grishin[1,2,*]

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

Received February 15, 2009; Revised April 9, 2009; Accepted April 20, 2009

ABSTRACT

The biological properties of proteins are often gleaned through comparative analysis of evolutionary relatives. Although protein structure similarity search methods detect more distant homologs than purely sequence-based methods, structural resemblance can result from either homology (common ancestry) or analogy (similarity without common ancestry). While many existing web servers detect structural neighbors, they do not explicitly address the question of homology versus analogy. Here, we present a web server named HorA (Homology or Analogy) that identifies likely homologs for a query protein structure. Unlike other servers, HorA combines sequence information from state-of-the-art profile methods with structure information from spatial similarity measures using an advanced computational technique. HorA aims to identify biologically meaningful connections rather than purely 3D-geometric similarities. The HorA method finds ~90% of remote homologs defined in the manually curated database SCOP. HorA will be especially useful for finding remote homologs that might be overlooked by other sequence or structural similarity search servers. The HorA server is available at http://prodata.swmed.edu/horaserver.

INTRODUCTION

Homology, or evolutionary relatedness, represents a key concept in studying protein sequence, structure, and function. Homologs can be inferred by sequence similarity search tools such as the popular sequence-profile comparison method PSI-BLAST (1) or the more sensitive profile-profile comparison methods COMPASS (2) and HHpred (3). Since protein three-dimensional (3D) structure is generally more conserved in evolution than sequence (4), structural similarity has been used to detect distant homologs (5–7). However, structural similarity may arise from factors other than descent from a common ancestor. Such 'analogous' similarity often comes from convergence to similar structures due to a limited number of energetically favorable ways to pack secondary structural elements (SSEs) (8,9). Thus, structure-based remote homology detection inevitably involves the challenging problem of discriminating between homologs and analogs. Currently, many servers are available for comparing protein structures, e.g. DALI (10), VAST (11), CE (12), SSM (13), MATRAS (14) and 3D-BLAST (15). Although strong structural similarity exemplified by the various different scores of these methods [e.g. DALI *Z*-score about 9 (16)] provides adequate evidence for homology, weak similarity often requires experts' knowledge and further analysis.

Here, we present a web server that combines sequence and structure information to detect remote homologs. This server is named HorA from 'Homology or Analogy' to reflect its goal: to identify remote homologs among structurally similar proteins lacking significant individual similarity scores (e.g. DALI *Z*-score ~5). To our knowledge, HorA represents the first web server to incorporate both sequence profile and structure information into its methodology. Previously, we used manually developed, reliable data sets of homologs (17) and analogs (18) to train a support vector machine (SVM) to discriminate homology from analogy (19). We improved over this method with the following approaches: (i) using transitive connections to identify remote homologs; (ii) employing a new negative filter to remove structurally dissimilar pairs; (iii) adding a positive filter incorporating a sensitive profile search to detect sequence homologs; and (iv) incorporating a new score standardization. The improved method (Cheng *et al.*, manuscript in preparation) recovered ~90% of manually defined remote homologs in SCOP (20). HorA implements the previously published method as a 'fast' procedure and the improved method as an 'accurate' procedure. We demonstrate the usefulness of the HorA server by an EF hand query example, where combining sequence and structural information found biologically more meaningful similarity (remote homology) than a structure-based method alone.

*To whom correspondence should be addressed. Tel: +214-645-5952; Fax: +214-645-5948; Email: grishin@chop.swmed.edu

HorA should be a useful tool for researchers interested in the biological implications of newly solved structures lacking close homologs.

DESCRIPTION OF THE HorA SERVER

Input

Users can upload protein structures in PDB file format or enter PDB IDs of previously deposited protein structures. PDB chain IDs also need to be specified if the PDB file contains more than one chain. Since using single protein domains frequently yields more accurate results than using complete chains with multiple domains, users can specify regions in the structure to be searched with residue ranges. The database search mode requires one input PDB file and the pairwise comparison mode requires two input PDB files. In both database search and pairwise comparison modes, users can choose either an 'accurate' procedure, which is slow, or a 'fast' procedure, which is less accurate.

Processing method

The primary goal of HorA is to find potential homologs for a protein structure of interest. To achieve this goal, HorA first computes various similarity measures between the structure of interest (or query protein) and every protein domain in a prepared database [less than 40% sequence identity representatives in SCOP v.1.69 downloaded from ASTRAL (21)]. Then, a decision is made about homology using three layers or components of the server: a negative filter, a positive filter, and an SVM model (Figure 1). The negative filter removes pairs lacking global structural similarity. The positive filter uses HHsearch probabilities (22) to identify close homologs. The SVM model combines a number of sequence, profile, and structure similarity measures into a single score. If a pair's SVM score is above a pre-defined threshold, it is classified as homologous. For cases where the direct SVM scores between proteins are too low to be confident for homology, HorA also finds homology using intermediate proteins (see 'Methods' section for more details). In addition to database search, HorA has a pairwise comparison mode that provides information about the likelihood of homology between two query proteins. The pairwise mode uses essentially the same procedure as the database mode.

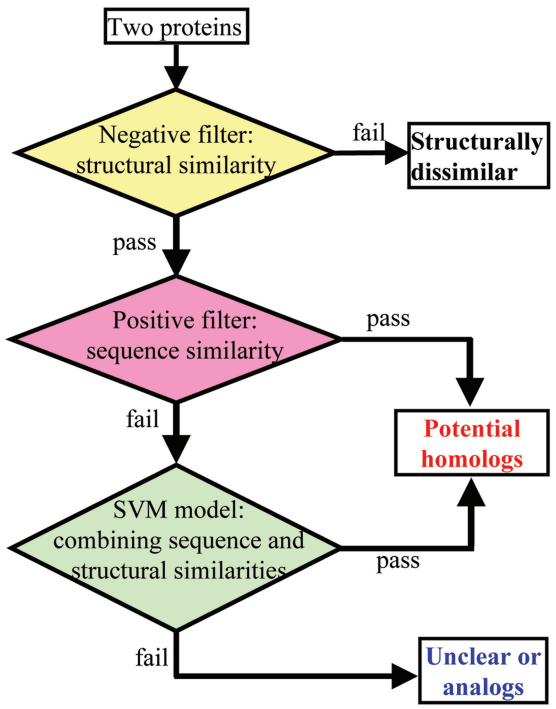


Figure 1. Overview of the HorA server 'accurate' procedure.

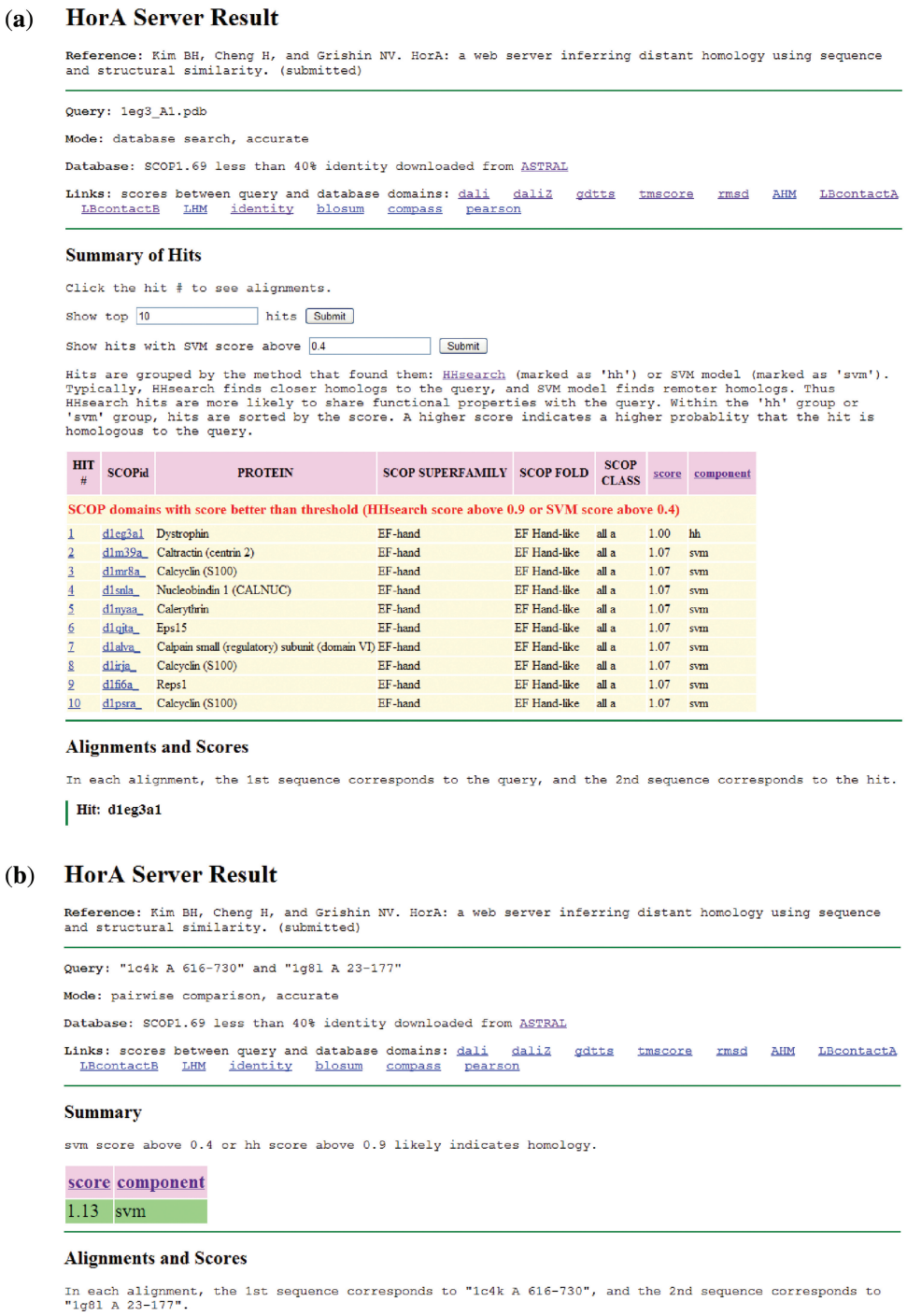
Although the above-mentioned procedure ('accurate' procedure) is sensitive and accurate in identifying remote homologs, extensive structural similarity comparisons make it very slow. Therefore, we also provide a less accurate 'fast' procedure that uses MAMMOTH (23) to compare the query with one representative structure from each SCOP fold. If the MAMMOTH *Z*-score between the query and a representative is above 4.0 (suggested by the authors), HorA aligns the query to every structure in that fold by FAST (24) and calculates similarity scores. These similarity scores are then combined by a less sophisticated and less sensitive SVM model published previously (19). The speed of the 'fast' procedure derives from three main aspects: (i) using MAMMOTH to quickly reduce the search space; (ii) employing FAST instead of DALI to build structural alignments; and (iii) using an SVM model that does not include modified database *Z*-scores for standardization (see 'Methods' section, Similarity scores and standardizations). Compared to the average running time (24 h) of the 'accurate' procedure, the 'fast' procedure significantly speeds up the process (~150 times), reducing the average running time to 10 min at the expense of compromised sensitivity.

Running time

Currently, the average running time of the 'accurate' procedure is about 24 hours. The running time is proportional to the size of the input protein, and some larger queries might take more than several days to complete. The long running time results from a need to compare the query with all proteins in the database to standardize scores (in both database search and pairwise comparison modes). To avoid the long running time, users can choose a less accurate and less sensitive 'fast' procedure that usually takes less than 10 min. However, queries that belong to highly populated folds such as doubly wound Rossmann-like or TIM-barrel may take longer (up to 1 h).

Output

The HorA server database search is designed to identify potential homologs among existing protein structures and facilitate further analysis of the found hits (Figure 2A). Results are summarized in a table that contains the SCOP classification, the component used in inferring homology ('hh' for the positive filter HHsearch and 'svm' for the SVM model), and the component-based score. Potential



(a) **HorA Server Result**

Reference: Kim BH, Cheng H, and Grishin NV. HorA: a web server inferring distant homology using sequence and structural similarity. (submitted)

Query: 1eg3_A1.pdb

Mode: database search, accurate

Database: SCOP1.69 less than 40% identity downloaded from ASTRAL

Links: scores between query and database domains: dali daliZ gdtts tmscore rmsd AHM LBcontactA LBcontactB LHM identity blosum compass pearson

Summary of Hits

Click the hit # to see alignments.

Show top 10 hits Submit

Show hits with SVM score above 0.4 Submit

Hits are grouped by the method that found them: HHsearch (marked as 'hh') or SVM model (marked as 'svm'). Typically, HHsearch finds closer homologs to the query, and SVM model finds remoter homologs. Thus HHsearch hits are more likely to share functional properties with the query. Within the 'hh' group or 'svm' group, hits are sorted by the score. A higher score indicates a higher probablity that the hit is homologous to the query.

HIT #	SCOPid	PROTEIN	SCOP SUPERFAMILY	SCOP FOLD	SCOP CLASS	score	component
SCOP domains with score better than threshold (HHsearch score above 0.9 or SVM score above 0.4)							
1	d1eg3a1	Dystrophin	EF-hand	EF Hand-like	all a	1.00	hh
2	d1m39a_	Caltractin (centrin 2)	EF-hand	EF Hand-like	all a	1.07	svm
3	d1mr8a_	Calcyclin (S100)	EF-hand	EF Hand-like	all a	1.07	svm
4	d1snla_	Nucleobindin 1 (CALNUC)	EF-hand	EF Hand-like	all a	1.07	svm
5	d1nyaa_	Calerythrin	EF-hand	EF Hand-like	all a	1.07	svm
6	d1qita_	Eps15	EF-hand	EF Hand-like	all a	1.07	svm
7	d1alva_	Calpain small (regulatory) subunit (domain VI)	EF-hand	EF Hand-like	all a	1.07	svm
8	d1irja_	Calcyclin (S100)	EF-hand	EF Hand-like	all a	1.07	svm
9	d1fi6a_	Reps1	EF-hand	EF Hand-like	all a	1.07	svm
10	d1psra_	Calcyclin (S100)	EF-hand	EF Hand-like	all a	1.07	svm

Alignments and Scores

In each alignment, the 1st sequence corresponds to the query, and the 2nd sequence corresponds to the hit.

Hit: d1eg3a1

(b) **HorA Server Result**

Reference: Kim BH, Cheng H, and Grishin NV. HorA: a web server inferring distant homology using sequence and structural similarity. (submitted)

Query: "1c4k A 616-730" and "1g8l A 23-177"

Mode: pairwise comparison, accurate

Database: SCOP1.69 less than 40% identity downloaded from ASTRAL

Links: scores between query and database domains: dali daliZ gdtts tmscore rmsd AHM LBcontactA LBcontactB LHM identity blosum compass pearson

Summary

svm score above 0.4 or hh score above 0.9 likely indicates homology.

score	component
1.13	svm

Alignments and Scores

In each alignment, the 1st sequence corresponds to "1c4k A 616-730", and the 2nd sequence corresponds to "1g8l A 23-177".

Figure 2. Result pages from the HorA server. (**a**) Result page of a database search. (**b**) Result page of a pairwise comparison.

homologs are ordered by closeness to the query protein: close homologs found by the positive filter using only sequence information appear first, and remote homologs found by the SVM model using both sequence and structure information follow. Within each category, HHsearch probabilities or SVM scores determine ranks. However, for close homologs whose HHsearch probabilities are indistinguishable (i.e. >95%), the BLOSUM62 score is used to improve ranks. Users can access additional information such as sequence and structural alignments and similarity scores by clicking the hit number in the table. Users can also change the number of hits shown in the result page by adjusting the threshold for hit display. The pairwise comparison output of the HorA server is similarly organized as the search output (Figure 2B), showing the information between the two query proteins.

Performance

We tested the method used in the 'accurate' procedure on SCOP (20) as well as on our manually prepared data sets

Table 1. Performance on different data sets

	MH	MA	FA	SF	FD	CL	RT
Total number of pairs	241	130	25 792	67 283	121 805	5 293 101	20 602 882
Accuracy (%)	96.3	90.8	98.2	92.0	27.4	89.0	99.7

SCOP1.69 domains with less than 40% sequence identity obtained from ASTRAL (21) are paired in an all-on-all fashion. These pairs are parsed into five subsets: FA (two domains are in the same SCOP family), SF (two domains are from different families but same superfamily), FD (two domains are from different superfamilies but same fold), CL (from different folds but same class) and RT (from different classes). Manual homologs (MH) (17) and manual analogs (MA) (18) are manually prepared data sets. Domain pairs in MH, FA and SF are labeled as 'homologs', while pairs in MA, FD, CL and RT are labeled as 'non-homologs'. Therefore, in calculating accuracies, classifying a MH, FA, or SF pair to be homologous is regarded as a 'correct' classification, while classifying a MA, FD, CL or RT pair to be homologous is regarded as a 'wrong' classification. The accuracy equals the number of 'correct' classifications divided by the total number of pairs in that data set. 3000 SF and 3000 FD pairs were used in training the SVM model (see 'Methods' section, SVM model).

(17,18). The testing results are summarized in Table 1. HorA detects ~90% of SCOP remote homologs (domains from different families but the same superfamily, 'SF' in Table I), while keeping high accuracies on non-homologs (~90% on manual analogs 'MA' and ~99% on SCOP domains from different classes, 'RT'). Since domains from different superfamilies but the same fold [e.g. superfamilies in the TIM-barrel fold (25)] or from different folds but the same class [e.g. Rossmann-like folds in the α/β class (26)] may be homologous, the accuracies on 'FD' and 'CL' are not as informative as those on the other data sets.

An example

The HorA server aims to detect evolutionarily meaningful hits rather than geometrically similar hits for a query structure. While overwhelming structural similarities do indicate evolutionary relationships, weak similarities remain problematic to interpret. In such ambiguous cases, servers considering only structural similarity may give misleading results. By analyzing both sequence and structure information, HorA should identify the most biologically relevant hits, as shown in the following example. Here, the query is one of two EF-hand domains in dystrophin (SCOP code: d1eg3a1) (27). A typical EF-hand consists of two tightly packed helix hairpins. The loop connecting the two helices in each hairpin displays a very characteristic conformation and binds calcium. The query domain deviates from typical EF-hands in several aspects: it does not bind calcium (27), the first hairpin loop adopts an atypical conformation, and the helices stack more parallel to one another (Figure 3, left structures). Nevertheless, the conformation of the second hairpin loop (marked by arrowheads in Figure 3) is largely the same as in a typical EF-hand, one of the calcium-binding residues (Asp187) is preserved, and the presence of a neighboring EF-hand in the structure implies duplication (=homology). In a search against a representative SCOP1.69 database (sequence identity below 40%), the first DALI hit to this query is a four-helical bundle in the prokaryotic signal recognition protein Ffh (SCOP code: d1ls1a1) (28). As shown in Figure 3a, all four helices in the query align to the hit with a reasonable DALI *Z*-score of 6.4. However, unlike the query, the hit lacks the characteristic loop of the EF-hand family and is classified in a different SCOP fold. Thus, the structural

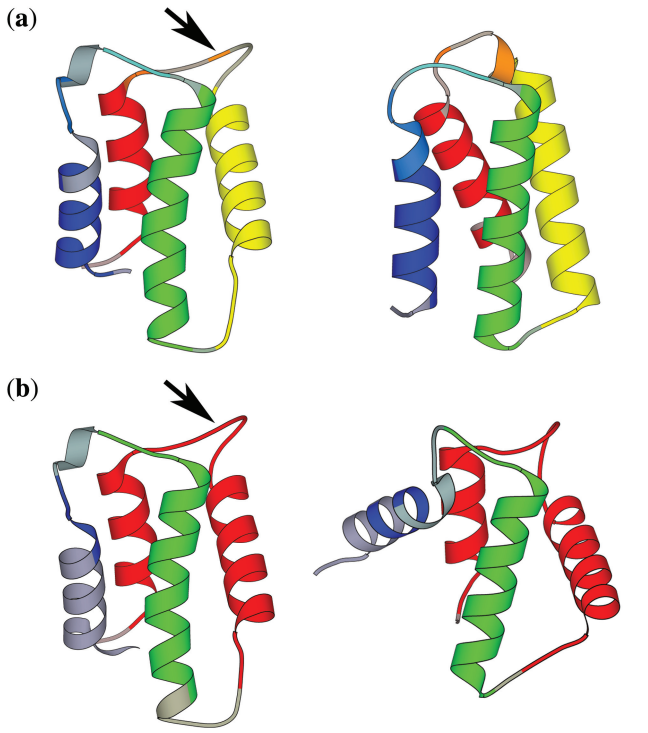


Figure 3. Comparison of the top DALI hit and the top HorA hit for an EF-hand query. (**a**) Left: query domain (PDB 1eg3, A124–A209). Right: first DALI hit (PDB 1ls1, A1–A88). According to the DALI alignment between these two domains, structurally equivalent parts are represented in the same color, while unaligned parts are in gray. Coloring starts from blue (N-terminus) and ends in red (C-terminus). (**b**) Left: the same query domain as in (a); Right: first HorA hit (PDB 1uhn, A118–A197). Colored as in (a).

similarity between the query and the hit probably results from convergent evolution. In other words, these two domains are structural analogs. On the contrary, the first HorA hit is classified in the EF-hand family (SCOP code: d1uhna_) (29). As shown in Figure 3b, although the first helix in the query is barely aligned to the hit, the characteristic loop of the EF-hand family is aligned (DALI *Z*-score 4.8). In this example, the top DALI hit is a structural analog, while the top HorA hit is a remote homolog. DALI, which is a purely geometric method, scores the overall structural similarity between the analog and the query higher than that of the homolog and the query. However, because the homolog retains a

more similar sequence conservation pattern to the query and HorA considers not only structural similarity but also sequence profile similarity, HorA correctly picks the homolog as the first hit.

CONCLUSIONS

We present a new web server that finds remote homologs of a query protein structure or quantifies the likelihood of homology between two query protein structures. In addition to decisions about homology, the HorA server provides helpful sequence and structural similarity scores and alignments for further analysis. As demonstrated by the EF-hand domain example, HorA is able to identify biologically meaningful protein as the first hit in contrast to commonly used structural similarity search methods based solely on geometry.

METHODS

The method used in the 'accurate' procedure is described below. For method used in the 'fast' procedure, see Cheng *et al*. (19).

Similarity scores and standardizations

For a pair of protein structures, 26 sequence and structure scores (13 similarity scores times two different standardization schemes) are calculated. The 13 similarity scores come from four categories: pairwise sequence scores (sequence identity and BLOSUM score), profile sequence scores (COMPASS-like and Pearson's correlation coefficient), intra-molecular structure scores (DALI score, DALI *Z*-score, LiveBench contact score A and LiveBench contact score B) and inter-molecular structure scores (TM score, RMSD, GDT_TS, Alignment-based Hausdorff measure and loop-based Hausdorff measure). See Cheng *et al*. (19) for equations and references of these scores. All scores are calculated based on the structural alignment between a pair of domains. The structural alignments are most often from DALI. However, FAST or TM-align alignments substitute in cases where DALI fails.

The sequence and structure scores are standardized in two different schemes: pair-specific scaling and modified database *Z*-scores, producing 26 scores in total. The two standardization schemes are conceptually complementary: scaled scores only consider a specific pair, while modified *Z*-scores take the information of the whole database into consideration. In scaling, $S = (S_{12} - S_{\text{random}})/(S_{\text{self}} - S_{\text{random}})$, where S is the scaled score. S_{12} is the raw score calculated from the structural alignment between domain 1 and domain 2. S_{random} is the random score generated by circularly permuting domain 1 relative to domain 2. S_{self} is the average of the two self scores S_{11} and S_{22}, which are calculated from domain 1 aligned to itself and domain 2 aligned to itself, respectively. In the modified *Z*-score standardization, $Z = (S_{12} - M_{12})/STD_{12}$, where Z is the modified *Z*-score. S_{12} is the raw score between domains 1 and 2. $M_{12} = (M_1 + M_2)/2$, where M_1 is the mean of the score distribution generated by comparing domain 1 to every domain in the database, and M_2 is the mean for domain 2. $STD_{12} = \sqrt{(VAR_1 + VAR_2)/2}$, where VAR_1 is the variance of the score distribution generated by comparing domain 1 to every domain in the database, and VAR_2 is the variance for domain 2. Z is transformed by $1/(1 + e^{-Z})$ to make it between 0 and 1 (e is the base of the natural logarithm).

Negative filter

If two domains share global structural similarity, their aligned regions usually have many long-range contacts, and their similarity tends to be consistently captured by different structure comparison programs. Based on these observations, we design the negative filter as a function of two numbers: a long-range contact c and an agreement between structural aligners a. Contact c measures the number of long-range contacts contained within a structural alignment. Suppose residues Ai and Aj in domain A are aligned to residues Bi and Bj in domain B, respectively. If Ai and Aj (and Bi and Bj) are separated by at least 10 amino acids in primary sequence and are within 14 Å in the 3D structure, we consider that there is one long-range contact. By scanning all possible residue pairs in the aligned region, we sum up the total contact number c. Agreement a measures to what extent the alignments generated by different programs [DALI (30), TM-align (31) and FAST (24)] agree with one another. We calculate the agreement between every pair of programs by counting the number of residues identically aligned by the two aligners. Then we take the maximum of the resulting three agreement numbers, and divide it by the shorter one of the two domains' lengths. We optimize the negative filter so that it filters out as many structurally dissimilar pairs as possible while keeping as many similar pairs as possible. Cheng *et al*. (32) contains a more detailed description of the negative filter idea.

Positive filter

The positive filter is designed to detect homologs with sequence information alone. Although structures are generally more conserved than sequences, sometimes sequences can be more helpful than structures in homology detection, e.g. large conformational changes may occur upon ligand binding. We use HHsearch (22) as the positive filter. Specifically, a pair is classified as homologous if its HHsearch score is above a conservative threshold (HHsearch probability 0.9).

SVM model

We use SVM^{light} (version 6.01, downloaded from http://svmlight.joachims.org/) to discriminate homology and analogy. Following Hsu *et al*. (33), we use the radial basis function (RBF) kernel and carry out a 'grid search' to optimize parameters C and γ. The SVM model is trained to discriminate remote homologs and structural analogs. The training set consists of 3000 domain pairs from different SCOP families but the same superfamily as remote homologs and 3000 domain pairs from different superfamilies but the same fold as structural analogs. These pairs are selected from a data set that has one

representative for every SCOP family in the four major classes (all α, all β, α/β, and α + β). In preparing the analog set, we try not to include putative homologs by avoiding pairs that belong to those folds whose superfamilies are known to be homologous, e.g. TIM-barrel (25), and pairs that are classified as homologous by the best linear classifier trained on the manual data sets as described in the previous publication (32). The SVM model combines the 26 input scores (see 'Similarity scores and standardizations' section) into a single prediction score. The default prediction score threshold in SVM classification is zero, i.e. a pair is classified as homologous if its SVM score is above zero or analogous if its score is below zero. We empirically chose a more conservative threshold of 0.4 to balance the classifier's performance on homologous and non-homologous sets. Specifically, domains within the same SCOP superfamily should be classified as homologs, while domains from different SCOP classes (e.g. all alpha versus all beta) should be classified as non-homologs. At the same time, the manually constructed, reliable data sets of homologs (17) and analogs (18) should be classified with high accuracy.

Transitivity with intermediates

Two domains A and B can be directly linked (classified as homologous) if the SVM score between them is above the pre-defined threshold. Additionally, A and B can be linked through an intermediate domain C if the SVM scores between A and C and between B and C are both above the threshold. Due to the extensive computing time associated with considering more intermediates, we limit the server to a single intermediate. Transitivity is also used in the positive filter.

ACKNOWLEDGEMENTS

The authors are very grateful to Lisa Kinch for many helpful comments about the server and the manuscript. B.H.K. thanks Seung-Jae Lee for critical reading of the manuscript. H.C. is grateful to Dr David Baker for accommodating her as a visiting scientist in University of Washington.

FUNDING

National Institutes of Health grant GM67 165; Welch foundation grant I1505 (to N.V.G.). Funding for open access charge: Howard Hughes Medical Institute.

Conflict of interest statement. None declared.

REFERENCES

1. 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.
2. 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.
3. 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.
4. Chothia,C. and Lesk,A.M. (1986) The relation between the divergence of sequence and structure in proteins. *EMBO J.*, **5**, 823–826.
5. Holm,L. and Sander,C. (1995) DNA polymerase beta belongs to an ancient nucleotidyltransferase superfamily. *Trends Biochem. Sci.*, **20**, 345–347.
6. Grishin,N.V. (2001) Mh1 domain of Smad is a degraded homing endonuclease. *J. Mol. Biol.*, **307**, 31–37.
7. Holm,L. and Sander,C. (1997) An evolutionary treasure: unification of a broad set of amidohydrolases related to urease. *Proteins*, **28**, 72–82.
8. Finkelstein,A.V. and Ptitsyn,O.B. (1987) Why do globular proteins fit the limited set of folding patterns? *Prog. Biophys. Mol. Biol.*, **50**, 171–190.
9. Krishna,S.S. and Grishin,N.V. (2004) Structurally analogous proteins do exist! *Structure*, **12**, 1125–1127.
10. Holm,L., Kaariainen,S., Rosenstrom,P. and Schenkel,A. (2008) Searching protein structure databases with DaliLite v.3. *Bioinformatics*, **24**, 2780–2781.
11. Gibrat,J.F., Madej,T. and Bryant,S.H. (1996) Surprising similarities in structure comparison. *Curr. Opin. Struct. Biol.*, **6**, 377–385.
12. Shindyalov,I.N. and Bourne,P.E. (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.*, **11**, 739–747.
13. Krissinel,E. and Henrick,K. (2004) Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallogr. D Biol. Crystallogr*, **60**, 2256–2268.
14. Kawabata,T. (2003) MATRAS: a program for protein 3D structure comparison. *Nucleic Acids Res.*, **31**, 3367–3369.
15. Yang,J.M. and Tung,C.H. (2006) Protein structure database search and evolutionary classification. *Nucleic Acids Res.*, **34**, 3646–3659.
16. Dokholyan,N.V., Shakhnovich,B. and Shakhnovich,E.I. (2002) Expanding protein universe and its origin from the biological Big Bang. *Proc. Natl Acad. Sci. USA*, **99**, 14132–14136.
17. Cheng,H., Kim,B.H. and Grishin,N.V. (2008) MALIDUP: a database of manually constructed structure alignments for duplicated domain pairs. *Proteins*, **70**, 1162–1166.
18. Cheng,H., Kim,B.H. and Grishin,N.V. (2008) MALISAM: a database of structurally analogous motifs in proteins. *Nucleic Acids Res.*, **36**, D211–D217.
19. 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.
20. Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.
21. Brenner,S.E., Koehl,P. and Levitt,M. (2000) The ASTRAL compendium for protein structure and sequence analysis. *Nucleic Acids Res.*, **28**, 254–256.
22. Soding,J. (2005) Protein homology detection by HMM-HMM comparison. *Bioinformatics*, **21**, 951–960.
23. Ortiz,A.R., Strauss,C.E. and Olmea,O. (2002) MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. *Protein Sci.*, **11**, 2606–2621.
24. Zhu,J. and Weng,Z. (2005) FAST: a novel protein structure alignment algorithm. *Proteins*, **58**, 618–627.
25. Copley,R.R. and Bork,P. (2000) Homology among (betaalpha)(8) barrels: implications for the evolution of metabolic pathways. *J. Mol. Biol.*, **303**, 627–641.
26. Burroughs,A.M., Allen,K.N., Dunaway-Mariano,D. and Aravind,L. (2006) Evolutionary genomics of the HAD superfamily: understanding the structural adaptations and catalytic diversity in a superfamily of phosphoesterases and allied enzymes. *J. Mol. Biol.*, **361**, 1003–1034.
27. Huang,X., Poy,F., Zhang,R., Joachimiak,A., Sudol,M. and Eck,M.J. (2000) Structure of a WW domain containing fragment of dystrophin in complex with beta-dystroglycan. *Nat. Struct. Biol.*, **7**, 634–638.

28. Ramirez,U.D., Minasov,G., Focia,P.J., Stroud,R.M., Walter,P., Kuhn,P. and Freymann,D.M. (2002) Structural basis for mobility in the 1.1 A crystal structure of the NG domain of Thermus aquaticus Ffh. *J. Mol. Biol.*, **320**, 783–799.
29. Nagae,M., Nozawa,A., Koizumi,N., Sano,H., Hashimoto,H., Sato,M. and Shimizu,T. (2003) The crystal structure of the novel calcium-binding protein AtCBL2 from Arabidopsis thaliana. *J. Biol. Chem.*, **278**, 42240–42246.
30. Holm,L. and Sander,C. (1993) Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.*, **233**, 123–138.
31. Zhang,Y. and Skolnick,J. (2005) TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.*, **33**, 2302–2309.
32. Cheng,H. (2007) Classification and differentiation of homologs and structural analogs. *Ph.D. Dissertation*, The University of Texas Southwestern Medical Center at Dallas, Dallas. http://www4.utsouthwestern.edu/library/ETD/etd.cfm.
33. Hsu,C.-W., Chang,C.-C. and Lin,C.-J. A practical guide to support vector classification. http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf.