Using homology relations within a database markedly boosts protein sequence similarity search

Jing Tonga, Ruslan I. Sadreyevb,c, Jimin Peid, Lisa N. Kinchd, and Nick V. Grishinad,1

aDepartment of Molecular Biophysics, University of Texas Southwestern Medical Center, Dallas, TX 75390-9050; bDepartment of Molecular Biology, Massachusetts General Hospital, Boston, MA 02114; cDepartment of Pathology, Massachusetts General Hospital and Harvard Medical School, Boston, MA 02114; and dHoward Hughes Medical Institute, University of Texas Southwestern Medical Center, Dallas, TX 75390-9050

Edited by Andrej Sali, University of California, San Francisco, CA, and accepted by the Editorial Board April 28, 2015 (received for review December 19, 2014)

Inference of homology from protein sequences provides an essential tool for analyzing protein structure, function, and evolution. Current sequence-based homology search methods are still unable to detect many similarities evident from protein spatial structures. In computer science a search engine can be improved by considering networks of known relationships within the search database. Here, we apply this idea to protein-sequence–based homology search and show that it dramatically enhances the search accuracy. Our new method, COMPADRE (Comparison of Multiple Protein sequence Alignments using Database RELationships) assesses the relationship between the query sequence and a hit in the database by considering the similarity between the query and its known homologs. This approach increases detection quality, boosting the precision rate from 18% to 83% at half-coverage of all database homologs. The increased precision rate allows detection of a large fraction of protein structural relationships, thus providing structure and function predictions for previously uncharacterized proteins. Our results suggest that this general approach is applicable to a wide variety of methods for detection of biological similarities. The web server is available at prodata.swmed.edu/compadre.

Prediction of protein structure and function by sequence homology is among the most important problems in computational biology of proteins, perhaps next after the grand problem of de novo protein folding. The existing gap between the number of known protein sequences and the number of experimentally determined 3D structures is bound to grow with more genomes sequenced by high-throughput technologies (1, 2). Currently, the most reliable and effective way to predict the structure of an uncharacterized protein is to find a sequence homolog with available structural information (3, 4). The chance of finding such a template for a given protein sequence is increasing as sequence space is becoming more extensively covered by 3D structures (5). However, there is, and will be for a long time, a significant fraction of proteins for which finding experimentally characterized sequence homologs is challenging or impossible. The structures of many such proteins, when solved, reveal their remote homology to previously known structures that are undetectable by current sequence-based homology search methods (6). Therefore, the quality of sequence-based homology search remains key for accurate structure prediction, as consistently confirmed by multiple rounds of the Critical Assessment of protein Structure Prediction (7).

In the last several years, methods for sequence similarity search have been greatly improved by the analysis of sequence patterns reflecting evolutionary, structural, and functional constraints in protein families. Introduction of numerical profiles (3) and hidden Markov models (HMM) (8) has allowed comparing a sequence to a multiple sequence alignment (MSA) (8–10). Such work was later followed by methods for profile–profile (11–13) and HMM–HMM (14–16) comparison, aimed at detecting similarities between distant families. In addition to the residue substitution preferences at sequence positions, MSA can reveal highly informative patterns of interdependence between amino acid content at different positions, in the form of MSA motifs and secondary structure predictions (15, 17, 18).

Is it possible to further improve sequence-based homology search by considering nonsequence information? In a typical homology search aimed at 3D structure prediction, a sequence or family of interest is compared with a database of proteins with known structures. This knowledge allows confident establishment of evolutionary links within the database. In computer science, networks of relationships between database subjects have been successfully used to improve the quality of search methods, most notably web searchers (19). Here, we show that knowledge of the protein database homology network dramatically increases the accuracy of sequence-based search.

To capitalize on this idea, we modified PROCAIN (protein profile comparison with assisting information) (18), our sensitive method for sequence profile similarity search, by considering the template’s homologs within the database (Fig. 1). We designed the new similarity measure of COMPADRE (Comparison of Multiple Protein sequence Alignments using Database RELationships) as a linear combination of the original score for the given template with the scores for a set of its homologs identified by structure, function, and sequence. Consistent similarity of these homologs to the query elevates the original score, which can increase the significance of a marginal sequence-based similarity to a level above detection threshold. On the other hand, a favorable score for a spurious hit becomes less significant if the set of its homologs is consistently dissimilar from the query. Therefore, this measure improves both sensitivity and specificity of homology detection.

Significance

In the field of protein structure prediction, identifying homology to known folds offers the most successful and practically useful strategy to provide protein spatial structure models. For protein sequence targets without closely related structures, even very distant homologs can provide reasonable templates for modeling. Despite significant progress in the field, remote sequence similarity search is far from perfection and fresh ideas are needed to extend detection limits. In computer science, the concept of utilizing internal relations within a database to improve similarity search was key to the success of search engines such as Google. Here, we show that similar consideration of the homology network within a protein database of structure templates can dramatically improve the accuracy of homology search.

Author contributions: J.T., R.I.S., and N.V.G. designed research; J.T. and R.I.S. performed research; J.T., J.P., L.N.K., and N.V.G. analyzed data; and J.T., R.I.S., J.P., and N.V.G. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission. A.S. is a guest editor invited by the Editorial Board.

1To whom correspondence should be addressed. Email: grishin@chop.swmed.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1424324112/-/DCSupplemental.
would be meaningful templates for 3D structure modeling.

**Methods**

For details). This SVM classifier finds similarities between structure similarity scores to establish homology (see SI Appendix, Methods for details). This SVM classifier finds similarities between structure domains pairs that are most likely evolutionarily related and should be ranked higher than more remote homologs. The closest query homolog can often be identified as the top hit that has the highest PROCAIN score: $H_{all} = 1$ if the homology is to the top hit, $H_{all} = 0$ if it is nonhomologous, and $h_{top} = 0$ if the homology is unclear (see SI Appendix, Methods for details).

**Results**

**Defining Close and Remote Template Homology Networks.** To establish the network of homologous relationships in the structure database, we defined homologous and nonhomologous pairs in a set of 5,116 representative structural domains (also known as templates) from the Structural Classification of Proteins (SCOP) database (20). Because a good homology search method should rank homologs according to their distance from the query as well as discriminate homologs from nonhomologs, we considered homologous relationships between protein pairs at two different levels. The first level represents “close homologs,” and was defined as domain pairs assigned to the same superfamily in SCOP. These templates are typically closely related to each other by 3D structure, tend to share similarities in sequence and function, and should be ranked higher than more remote homologs. The second level represents “all homologs,” and includes more distantly homologous protein pairs. To define all homologs, the SCOP classification was supplemented with a support vector machine (SVM) classifier (21) that uses a number of sequence and structure-based similarity scores to establish homology (see SI Appendix, Methods for details). This SVM classifier finds similarities between domain pairs that are most likely evolutionarily related and would be meaningful templates for 3D structure modeling.

**Establishing a Scoring Scheme by Using Homology Networks in the Search Database.** The original PROCAIN E values reflecting the sequence-based similarity between the query and templates were first log-transformed into similarity scores $S_0$ (see Methods for details). The closest query homolog can often be identified as the top hit by this direct score. To further improve PROCAIN scoring, similarity scores on a particular template ($S_0^T$) were boosted by the similarity scores of the template’s homologs ($S_0^H$) according to the following equation:

$$S_1 = w_T S_0^T + (1 - w_T) \Sigma S_0^H,$$  \hspace{1cm} [1]$$

where $S_0^T$ and $S_0^H$ are the similarity scores $S_0$ for the given template and for a set $H$ of its structure-based homologs within a certain evolutionary distance from the template, respectively. The $S_0^H$ similarity score can be calculated using either the close homolog level ($H_{close}$) or the all-homolog level ($H_{all}$) described above. $w_T$ is a weight optimized for the performance ($w_T = 0.8$).

Additional information about the query’s top hit may help detect the query’s homologs in the database. Indeed, we find (SI Appendix, Fig. S1 and Tables S1 and S2) that performance can be additionally improved by transforming the measure in Eq. 1 to boost scores for the templates that are homologous to the top hit and to reduce scores for nonhomologous templates:

$$S_2 = S_1 + h_{top} (\alpha \cdot S_1 + \beta),$$  \hspace{1cm} [2]$$

where $S_1$ is the measure defined by Eq. 1, $\alpha$ and $\beta$ are optimized parameters, and $h_{top}$ depends on the homology of the template to the top hit that has the highest PROCAIN score: $h_{top} = 1$ if the template is homologous to the top hit, $h_{top} = -1$ if it is nonhomologous, and $h_{top} = 0$ if the homology is unclear.

**Improvement of COMPARE Scoring Scheme by the Choice of Homology Network.** The choice of homolog set $H$ in Eqs. 1 and 2 has a dramatic influence on the method’s behavior. Including scores for a template’s all homologs ($H_{all}$) results in a wider sampling of protein space around the template and thus should provide more representative information for homolog/nonhomolog discrimination. Such a wide sample, however, may lead to a scrambled ranking among detected homologs, with the closest ones being placed below the more distant. For a query’s close homologs, the strong direct similarity signal from the first scoring term of Eq. 1 may be diluted by the contribution of a diverse set of template’s homologs from the second scoring term of Eq. 1. Restricting the set’s diversity to close homologs ($H_{close}$) should improve the ranking of close homologs, but may limit the sensitivity of detection at remote homology levels. Therefore, the size of homolog set $H$ may require adjustment for different evolutionary distances between query and template.

Indeed, applying different sets $H$ ($H_{all}$ or $H_{close}$) to generate $S_1$ and $S_2$ results in a very different performance of our scoring scheme. We used receiver operator characteristic (ROC) curves to evaluate the homology detection performance of Eq. 2 for all query homologs designated as true positives (Fig. 2A) and for only close homologs designated as true positives (Fig. 2B). An ROC curve plots true positive numbers against false positive numbers at various E-value thresholds. Both plots (Fig. 2A and B) include the curves for several published sequence-based searchers: the popular PSI-BLAST (position-specific iterated basic local alignment search tool) method (22), PROCAIN (18), and a comparable state-of-the-art method HHSearch (15) (see Methods for details). These plots are shown together with those produced by our scoring scheme using two definitions of $H$. When contribution from all template homologs ($H_{all}$) is allowed in determining $S_1$ and $S_2$, the quality of homolog/nonhomolog discrimination is dramatically higher than in other methods (Fig. 2).

![Fig. 1. Context of template’s relationships within the database is used to modify the original score ($S_0^T$) for sequence-based similarity between query and template. The modified measure is a linear combination of $S_0^T$ and scores ($\Sigma S_0^H$) for the similarity between query and template’s homologs.](image)

![Fig. 2. Detection quality depends on the range of template’s homologs included in the similarity scoring. ROC plots for (A) inclusion of all homologs and (B) retrieval of close homologs (relationships outside SCOP superfamily are considered false positives). The scoring based on the inclusion of only close template homologs $H_{close}$ (orange) is compared with the unrestricted inclusion of all homologs $H_{all}$ (brown).](image)
2A and SI Appendix, Tables S1 and S3). However, when the contribution is limited to close homologs (H_close), the discrimination between homologs and nonhomologs becomes worse, especially for more distant homologs, in the area further from the plot’s origin (Fig. 2A and SI Appendix, Tables S1 and S3). The situation is opposite when only close homology detection is evaluated (Fig. 2B and SI Appendix, Tables S2 and S4). Inclusion of all template homologs (H_all) in determining S1 and S2 of Eq. 2 results in extremely poor identification of close homologs, suggesting that more distant relationships are often erroneously assigned higher significance. Limiting the set H to the close homology level (H_close) leads to an accurate close homology detection far surpassing the original PROCAIN performance (Fig. 2B and SI Appendix, Tables S2 and S4). These effects are consistent among all major protein secondary structure classes (SI Appendix, Figs. S2–S5).

The results shown in Fig. 2 suggest that to improve both the detection of remote homology and the ranking by evolutionary distance to the query, we can adjust the contribution from more distant homologs in Eqs. 1 and 2 according to the template’s distance from the query. Both goals may be achieved if set H is kept relatively narrow for close query–template relationships (the left part of the orange curve in Fig. 2 A and B), and the input from remote template homologs is added only for templates more distant from the query (the right part of the brown curve in Fig. 2 A and B). We construct a combined scoring function for such an adjustment:

\[ S_3 = w_c (S_{2}^{1}) * S_{2}^{3} + (1 - w_c) (S_{2}^{1}) * S_{2}^{3}, \]

where \( S_{2}^{1} \) and \( S_{2}^{3} \) are determined by Eq. 2 with different definitions of set H: only close homologs (H_close) for \( S_{2}^{1} \) and all database homologs (H_all) for \( S_{2}^{3} \). The weight \( w_c \) is a variable depending on score \( S_{2} \) as a measure of closeness to template to query. For high \( S_{2} \) values [closely similar to the query, when \( S_{2} \) is above an upper boundary \( S_{2}^{4} \)] this weight is set to 1, so that the final score includes only the score of close homologs of a template (\( S_3 = S_{2}^{1} \)), whereas for low \( S_{2} \) values [distantly similar to the query, when \( S_{2} \) is less than a lower boundary \( S_{2}^{3} \)] the weight is set to 0 so that the score includes only the score for all template homologs (\( S_3 = S_{2}^{3} \)). For the intermediate values of \( S_{2} \), i.e., \( S_{2}^{4} < S_{2} < S_{2}^{3} \), the weight monotonically grows from 0 to 1, to gradually mix \( S_{2}^{1} \) with \( S_{2}^{3} \). Testing several functions, we find that exponential dependency of \( w_c \) on \( S_{2} \) (SI Appendix, Fig. S11) provides the best performance (see Methods for details).

Although consideration of a template’s homologs in Eqs. 1–3 can boost scores of marginally detectable homologs, it can also reduce the significance of original PROCAIN E values for highly confident homologs. Thus, we construct a second combined scoring function:

\[ S_4 = w_3 (\ln E_{p}^{2}) * S_{3} + (1 - w_3) (\ln E_{p}^{2}) * S_{p}, \]

where \( S_3 \) is determined by Eq. 3 and \( S_p \) is the score obtained from the original PROCAIN E value \( E_{p} \) using the Gumbel extreme value distribution (EVD) (23), which approximates a distribution of sequence similarity scores of random comparisons (24–26). Since it was introduced in sequence analysis by BLAST (24), this distribution has been widely used to estimate statistical significance of sequence and profile similarity scores (i.e., to compute E value from a score) in many applications (9, 18, 27). Karlin and Altschul and Dembo et al. (25, 26) estimated parameters of EVD and suggested a formula to transform a sequence score to E value. E values can be back-transformed to scores using this approximation. The weight \( w_3 \) is a function of \( \ln E_{p} \), the logarithm of the PROCAIN E value. For low \( \ln E_{p} \) values [highly confident PROCAIN hits, when \( \ln E_{p} \) is less than a lower boundary \( \ln E_{p}^{2} \) ] this weight is set to 0, so that the final score is only determined by the original PROCAIN score (\( S_4 = S_p \)), whereas for high \( \ln E_{p} \) values [marginal PROCAIN hits, when \( \ln E_{p} \) is above an upper boundary \( \ln E_{p}^{2} \) ] this weight is set to 1 so that the final score \( S_4 \) is equal to the new score \( S_p \). For the intermediate values of \( \ln E_{p} \), i.e., \( \ln E_{p} \leq \ln E_{p} \leq \ln E_{p}^{2} \), the weight monotonically increases from 0 to 1, to gradually mix \( S_p \) with \( S_3 \). Testing several functions, we find that exponential dependency of \( w_3 \) on \( \ln E_{p} \) (SI Appendix, Fig. S12) gives the best performance (see Methods for details). Based on the score \( S_4 \) for a given template, statistical significance of the detected similarity is provided in the form of E value estimated by transforming the score using the EVD approximation.

The final scoring function \( S_4 \) offers best performance both in remote homology detection and in ranking by evolutionary distance to a query. Performance of the resulting measure is compared with several methods in Fig. 3 A and B and SI Appendix, Figs. S6–S8. The inclusion of all template homologs (H_all) in the set H leads to highly sensitive and accurate retrieval of homology relationships (Fig. 3A and SI Appendix, Tables S1 and S3). At the same time, using a restricted set H for shorter ranges of query–template distance [only close homologs (H_close)] leads to the correct placement of the close query homologs above others (Fig. 3B and SI Appendix, Tables S2 and S4). One of the most important characteristics of this scoring scheme is precision rate, i.e., the expected proportion of true positives among top hits. As shown in SI Appendix, Table S5, the scoring function \( S_4 \) achieves the precision rate of 83% at half-coverage of all homologs, more than quadruple that of the original PROCAIN rate of 18%. Thus, the combined measure \( S_4 \) by far exceeds the current state-of-the-art performance levels in both capturing remote protein relationships and ranking homologs consistently with evolutionary distance. We refer to the resulting detection method as COMPADRE, for COMparison of Multiple Protein sequence Alignments using Database RELationships.

A web server based on this method (prodata.swmed.edu/compadre) is developed to detect homologs for the query sequence provided by the user. The output page shows ranked significant hits, their sequence alignments to the query, and their 3D structures. The sequence alignments between the query and each template are generated by PROMALS (profile multiple alignment with predicted local structures) (28), which produces better alignments than PROCAIN.

**Comparison with Structure Similarity Score.** A more detailed analysis of the COMPADRE results suggests that it accurately captures a large fraction of structural similarities that are only weakly reflected in sequence, and at the same time highlights the similarity of local functional motifs that may be missed by an automatic structure comparison method. As an illustration,
Fig. 4 shows the comparison of protein groupings based on COMPADRE E value and on the structural similarity measured by DALI (distance-matrix alignment) (29) Z score. We use 1,313 representative protein domains from the α/β class in our database to perform hierarchical clustering by all-to-all COMPADRE scores (logarithm of COMPADRE E values) (Fig. 4A) and by DALI Z scores (Fig. 4B). The resulting matrices of scores for domain pairs are represented as colored maps, with the scores used for grouping shown above diagonal and the corresponding scores by the other method shown below diagonal.

These groupings provide several notable observations. First, there is a strong general correlation in clustering by both scores, including the identity of major protein groups such as TIM (triosephosphate isomerase)-barrels, Rossmann fold, and SAM (S-adenosylmethionine)-dependent methyltransferases (Fig. 4A and B). Second, as expected, a fraction of structure-based relationships still remain undetected by sequence. These relationships include both similarities outside major clusters (off-diagonal area of the matrices, Fig. 4B) and links within clusters. For example, TIM-barrels have uniformly high DALI Z scores, whereas the coverage by COMPADRE scores is more fragmented (Fig. 4A and B). Third, COMPADRE produces several clusters of pronounced similarity that stand out from the background (red in Fig. 4A) but are not produced by DALI. These clusters correspond to local functional sequence motifs whose presence is less obvious from structure comparison alone. The most notable example is P-loop nucleoside triphosphatases that are accurately placed together by COMPADRE but split apart by structure similarity (Fig. 4A). As another example, COMPADRE grouping within the TIM-barrel fold highlights a superfamily of (trans)glycosidases that share similar phosphate-binding sites, which is challenging for DALI-based clustering (Fig. 4A).

**Detecting More Homologs at the SCOP Superfamily Level.** Compared with the original PROCAIN scoring, COMPADRE detects more homologs at the SCOP superfamily level. As an example, in bacterial lysozyme [Protein Data Bank (PDB) ID code 1jfx, chain A, SCOP family 1, α-β-N-acetylmuraminidase, Fig. 5A], the last α/β unit of 1jfx is atypical for TIM-barrel folds: it has an antiparallel hairpin replacing the typical parallel α/β unit. A typical TIM-barrel structure (PDB ID code 1bqc, chain A) is shown in Fig. 5B. Due to this alteration, it is difficult for PROCAIN to detect homologs of 1jfx, and the PROCAIN E values for 1jfx vs. other TIM-barrel examples are high. The exception is an N-terminal domain of endolysin (PDB ID code 2j8g, chain A, domain 2) and an uncharacterized bacterial protein (PDB ID code 1sf, chain A), which are in the same SCOP family 1, α-β-N-acetylmuraminidase. A scatter plot of E value vs. DALI Z score shows COMPADRE E values shifted lower to significant E values (E value = 0.005 line displayed), while keeping roughly the same ranking as PROCAIN. SCOP classifies all of these structures (dots in Fig. 5C) in the same superfamly: (trans)glycosidases, and they catalyze reactions with similar chemistry, suggesting they should be homologous. PROCAIN only detects two closest sequences at the SCOP family level (PDB ID code 2j8g, chain A, and PDB ID code 1sf, chain A) with a significant E value, whereas COMPADRE detects all of the most distant structures with significant E values.

**Detecting Homology Relationship for a Recently Determined Structure.** For the inference of structure, function, and evolution of a given uncharacterized protein, COMPADRE improves the opportunity for detection of experimentally characterized homologs. As another example, the zinc-finger antiviral protein (ZAP) is a host factor that specifically inhibits the replication of certain viruses, such as HIV-1 (30). The N-terminal part of ZAP is the major functional region that binds target RNA and recruits the mRNA degradation machinery. The structure of the N-terminal region of ZAP was determined recently (31) (NZAP225, PDB ID code 3u9g, chain A), consisting of an uncharacterized top “cockpit” layer (1-65aa) and four connected zinc fingers (66-225aa). Fig. 6 shows how the confidence of an originally marginal sequence
similarity can be verified for the uncharacterized top docking domain by using our COMPADRE method. For the query sequence of the first 65 residues from NZAP225, no confident homology to proteins with known 3D structure can be detected by direct query–template sequence comparisons. PSI-BLAST search in NCBI (National Center for Biotechnology Information) nr database converges after three iterations with no hits of known structures, whereas all HHpred and PROCAIN hits in structural databases are outside the significance threshold (best probability of 82.4% for HHpred and best E value of 76.1 for PROCAIN). COMPADRE, however, is able to assign significant E values to the similarities between the N-terminal domain of the query and multiple DNA-binding helix–turn–helix (HTH) domains, with the top E value as low as 2e-3 (Fig. 6). Detected similarity to HTH method, COMPADRE, shows a dramatic increase in the performance of 82.4% for HHpred and best E value of 76.1 for PROCAIN). COMPADRE, however, is able to assign significant E values to the similarities between the N-terminal domain of the query and multiple DNA-binding helix–turn–helix (HTH) domains, with the top E value as low as 2e-3 (Fig. 6). Detected similarity to HTH (Fig. 6B) suggests that the first 65-residue domain of ZAP may also play an important role in recognizing target viral RNA. Sequence alignment to the template (Fig. 6A) points to specific residues that may be involved in RNA recognition and binding, providing potential targets for mutagenesis studies.

Conclusions

Our findings show that defined homology relationships within the search database contain essential information for the improvement of sequence-based homology search. Although this concept is not new to computer science in general, to our knowledge, it has not been successfully applied to protein sequence searches before. Our method, COMPADRE, shows a dramatic increase in the performance of sequence search compared with current methods that are based on traditional query–template similarity measures. COMPADRE detects a large fraction of structural protein relationships and allows for predictions of structure and function in previously uncharacterized proteins. Furthermore, our results suggest that this approach may be applicable to a wide variety of methods for the detection of biological similarities.

Methods

Protein Database. To perform homology searches, we use the set of 5,116 SCOP domain representatives (18, 21), with relationships defined by the combination of manual expert classification in SCOP with an automated SVM-based analysis of multiple scores for structure and sequence similarity (SI Appendix, Methods). PSI-BLAST alignments of homologs for each protein are used for PSI-BLAST, HSSearch, PROCAIN, and COMPADRE comparisons with default settings. PROCAIN is a sequence profile search method that, in addition to sequences, incorporates similarity in secondary structure, positional conservation, and sequence motifs into profile–profile scoring. When combined with an improved estimation of statistical significance of hits, this scoring results in better performance compared with other methods (19).

Scoring Function. The original similarity scores (Sc) are produced by logarithmetic scaling of PROCAIN E values: Sc = log (E/E0), where E is the E value of PROCAIN hit and E0 is a constant offset [log (E0) = 6]. In Eq. 1, the optimal weight is empirically determined as wA = 0.8. In Eq. 2, optimal parameters are A = 0.3, B = 5.5 for the scores based on closer template homologs (same SCOP superfamily) and A = 0.8, B = 8.0 for the scores based on all homologs. In Eq. 3, scores Sc1 and Sc2 are defined by Eqs. 1 and 2, with Sc additionally rescaled to be comparable to Sc in the area of mixing (SI Appendix, Methods). Weight w1(Sc) is defined as follows (SI Appendix, Figs. 39 and 51): w1 = 0 if Sc ≤ Sc1(1); w1 = 1 if Sc > Sc1(1). In Eq. 4, Sc is determined by Eq. 3 and Sc is the equivalent score converted by original PROCAIN E value based on the EVD estimation (25) of Sc, respectively. The weight of Sc is defined as follows (SI Appendix, Figs. 31 and 52): w2 = 0 if ln(E1) < ln(5); w2 = 1 if ln(E1) > ln(5) and w2 = a + be(E1−ln(5)), where a, b are derived from boundary conditions at Sc1(2) and Sc2(2). In Eq. 4, Sc is determined by Eq. 3 and Sc is the equivalent score converted by original PROCAIN E value based on the EVD estimation (25) of Sc, respectively. The weight of Sc is defined as follows (SI Appendix, Figs. 31 and 52): w2 = 0 if ln(E1) < ln(5); w2 = 1 if ln(E1) > ln(5) and w2 = a + be(E1−ln(5)), where a, b are derived from boundary conditions at Sc1(2) and Sc2(2). Based on the final scores in Eq. 4, E values are calculated and similarities are ranked by statistical significance. Hierarchical clustering is performed by Cluster 3.0 (32).

ACKNOWLEDGMENTS. The authors acknowledge the Texas Advanced Computing Center (TACC) at The University of Texas at Austin for providing grid resources that have contributed to the research results reported within this paper. This work was supported in part by the National Institutes of Health (GM094575 to N.V.G.) and the Welch Foundation (I-1505 to N.V.G.).