2240–2248 *Nucleic Acids Research, 2008, Vol. 36, No. 7* *Published online 19 February 2008*
doi:10.1093/nar/gkn065

Accurate statistical model of comparison between multiple sequence alignments

Ruslan I. Sadreyev[1,*] 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 January 2, 2008; Revised January 31, 2008; Accepted February 1, 2008

ABSTRACT

Comparison of multiple protein sequence alignments (MSA) reveals unexpected evolutionary relations between protein families and leads to exciting predictions of spatial structure and function. The power of MSA comparison critically depends on the quality of statistical model used to rank the similarities found in a database search, so that biologically relevant relationships are discriminated from spurious connections. Here, we develop an accurate statistical description of MSA comparison that does not originate from conventional models of single sequence comparison and captures essential features of protein families. As a final result, we compute *E*-values for the similarity between any two MSA using a mathematical function that depends on MSA lengths and sequence diversity. To develop these estimates of statistical significance, we first establish a procedure for generating realistic alignment decoys that reproduce natural patterns of sequence conservation dictated by protein secondary structure. Second, since similarity scores between these alignments do not follow the classic Gumbel extreme value distribution, we propose a novel distribution that yields statistically perfect agreement with the data. Third, we apply this random model to database searches and show that it surpasses conventional models in the accuracy of detecting remote protein similarities.

INTRODUCTION

Detection of remote sequence similarity is crucial for protein structure-functional prediction and evolutionary analysis. A growing consensus of independent surveys suggests that comparison of multiple sequence alignments (MSA) (1–8) is the best strategy for inference of distant evolutionary relations between protein families and for protein structure prediction (9–12). The strength of this approach stems from evolutionary conservation revealed by MSA, such as functionally and structurally important sequence motifs or amino acid periodicity in α-helices and β-strands. However, this conservation can misleadingly emphasize recurrent common patterns in unrelated proteins. In order to distinguish such spurious similarities from evolutionarily meaningful and predictive relationships, it is crucial to have an accurate null model, i.e. a method to simulate and statistically describe comparisons between unrelated families. Estimates of statistical significance based on such a model will provide high-quality detection of similarity between proteins.

For single sequence comparison, an elegant theoretical model (13) allows for a simple mathematical description. A local alignment of two sequences is characterized by a similarity score, which is assigned a statistical significance. Assuming the absence of correlation between amino acid content at individual sequence positions, the statistical significance can be estimated using extreme value distribution (EVD) (14,15). This EVD-based model, however, produces less accurate results when applied to the comparison of alignments (11,16,17). An alternative empirical approach, based on the comparisons of unrelated proteins (2,3,5–7), suffers from relatively small number of existing protein families, which restricts the size of sampled distributions of similarity scores. Thus, well-established models for sequence comparison are not fully adequate in the comparison of MSAs.

Construction of a null model, mathematical description of the resulting score distribution and final evaluation of the model are important parts in the development of methods for the remote homology detection (1,18–21). However, with respect to MSA comparison, the following central questions have not received sufficient systematic attention. What essential biophysical features of proteins should be captured by a null model of MSA comparison? Which mathematical approximation provides precise statistical agreement with a distribution of random

*To whom correspondence should be addressed. Tel: +1 214 645 5951; Fax: +1 214 645 5948; Email: sadreyev@chop.swmed.edu

alignment scores? Finally, is the model's precision important in practical applications?

The objective of this work is to develop null models of alignment comparison that accurately represent the properties of proteins. Specifically, we focus on modeling local sequence patterns dictated by protein secondary structure. As a result, we propose (i) a method to generate random alignments that captures essential features of protein families, and (ii) a precise mathematical description of the distributions of similarity scores between these alignments. We apply this model to protein similarity searches and conclude that relatively small changes in the statistics unexpectedly lead to significant improvement in the detection of similarities between protein families. The similarity detection tool based on this model is available for download from: ftp://iole.swmed.edu/pub/compass.

METHODS

Score distribution for the comparison of unrelated families

We select 1825 PFAM families that correspond to a known single-domain structure classified in SCOP database (22). Similarity relations between domains are assigned as described elsewhere (16), by complementing the SCOP superfamily classification with an SVM classifier based on measures of structure and sequence similarity. In this set, we randomly choose 200 PFAMs as queries and collect COMPASS similarity scores between queries and other unrelated PFAMs.

Randomization of alignments

Randomized PFAMs have the same length and thickness as real and the content generated according to one of 11 models (see Supplementary Data). We use a simple measure of alignment thickness, average count of different residue types per position. As a source of alignment fragments, we use PFAM alignments processed to purge redundant sequences and short sequence fragments. After randomizing 200 query alignments, we modify the remaining set, avoiding selection of MSA fragments already included in the randomized queries.

Generation of large samples of decoy profiles with given length and thickness

We concatenate the PSIPRED (23) SS predictions for PFAMs of known single-domain structure, forming a long template of SS states. To generate a decoy profile, we select a random template segment and fill it with randomly drawn profile fragments whose SS type matches the template (see also Supplementary Data). Based on pairs of these profiles, we produce double samples of 1 million COMPASS scores for various combinations of profile length and thickness ($50 \leq L \leq 1000$ and $2.0 \leq N \leq 17.0$).

Analytical approximation of the distribution dependency on profile length and thickness

Based on fitting individual score distributions with PEVD, we construct approximate analytical functions that involve 20 parameters for pair $m(L, N)$, $s(L, N)$ and 12 parameters for pair $\alpha(L, N)$, $\beta(L, N)$. We refine the values of these parameters by minimizing the sum of χ^2 values simultaneously produced for the corresponding PEVD fits to empirical distributions for various combinations (L_i, N_j) (Figure 4a, see Supplementary Data for details). The quality of the resulting fit is tested on independent samples of 10^5 scores for every of 464 combinations of profile length and thickness.

Performance evaluation

The testing set of PSI-BLAST alignments is a part of previously described set (16) and is produced from 2879 SCOP domains with <20% sequence identity. All-to-all comparisons are performed within this and PFAM-based set using COMPASS and other methods with default settings. Models based on comparison of a query to the database of real or reversed profiles were implemented as described elsewhere (7,20,30). Statistical accuracy plots are produced as follows. For each query, we record the lowest E-value of a false positive hit, calculate the corresponding P-value $P = 1\text{–}e^{-E}$, and plot the fraction of queries for which $P \leq x$ against x.

RESULTS

Currently, two approaches are used to model random MSA comparisons and to provide estimates of statistical significance for detected similarities. The first approach is based on randomized MSAs whose individual positions are randomly drawn from real alignments (4,19). This approach provides a potentially unlimited statistical sample and allows for a precise analysis of resulting distributions of scores, but uses an overly simplistic representation of real alignments. The second approach involves comparisons within a database of real MSAs (2,3,5–7), most of which are not similar to each other. This approach closely mimics real-life similarity search but has a limited precision due to a statistically small sample size. Here, we develop a new modeling approach that combines realistic representation of essential protein features with precise mathematical description.

First, we implement various random models and select the one that most closely reproduces comparisons between unrelated protein families. Second, we use this model to simulate random MSA comparisons and generate distributions of similarity scores. Third, we find precise mathematical description of these distributions depending on two MSA properties: alignment length and sequence diversity ('thickness'). Finally, we apply this description to estimate statistical significance in the searches for protein similarity, and compare the performance of the new model to other models. At all steps, similarity scores are generated by the COMPASS scoring function, our previously reported method for MSA comparison (4,17).

Selection of the best procedure for decoy generation

The first step is to generate randomized alignments that can be used as decoys to model random comparisons in a biologically meaningful way. We generate these decoys by randomly combining elementary blocks derived from real

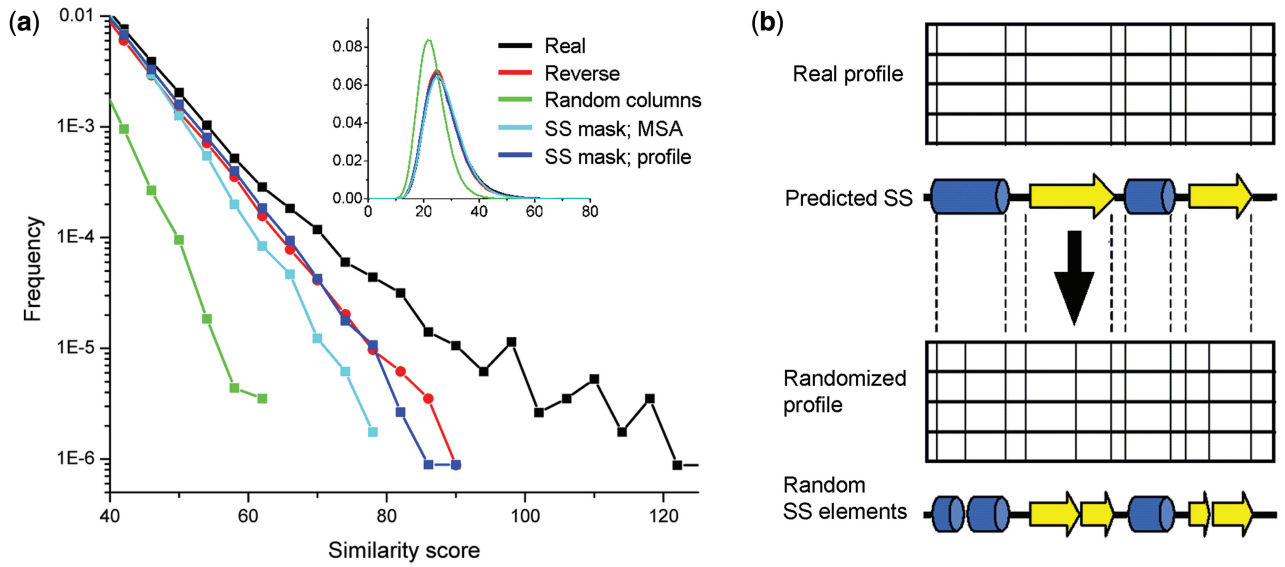

Figure 1. Distribution of similarity scores for unrelated protein families is most closely mimicked by randomized alignments with preserved secondary structure. (**a**) Score distribution for the comparison of real unrelated alignments ('Real') is plotted together with score distributions for the alignments randomized according to various models (Supplementary Figure S1). Distribution tails are shown in logarithmic scale. Plots for the following models are shown: comparison of real profiles with reversed directionality in one of the profiles ('Reverse'), profiles composed of randomly drawn profile positions ('Random columns'), and two models based on predicted secondary structure (SS). In these two models, alignment fragments ('SS mask, MSA') or profile fragments ('SS mask, profile') corresponding to original SS elements are replaced with random fragments corresponding to the elements of the same type. Note that the latter model closely reproduces the distribution for the most detailed 'Reverse' model. Inset: full plots in linear scale. (**b**) Schematic of profile randomization based on predicted secondary structure.

alignments. Our models vary in the definitions of these blocks and the rules for their combination. The simplest elementary block is a single column of MSA; the simplest rule is random draw. To better model native proteins, we (i) use predicted secondary structure elements (SSE) as elementary blocks; and (ii) arrange these SSE in native-like order (see Methods section for details). We use three types of predicted SSE (helix, strand and loop). For all models, we also implement their profile versions, with the random positions or fragments being drawn not from MSAs (columns of amino acid letters), but from corresponding numerical profiles. Profiles represent residue frequencies at MSA positions, taking into account redundancy of similar sequences and substitution propensities for different amino acids (4,24,25). As a source of alignment fragments, we use the PFAM database (26).

Our first goal is to select the random model that mimics most closely the distribution of similarity scores for unrelated protein families. As a set of families with determined relationships, we choose a set of PFAMs with solved protein structures. We compare 200 PFAM alignments designated as queries to unrelated PFAMs from a database of ~1500 families (see Methods section for details). The resulting similarity scores (Figure 1a) form a heavy-tailed distribution. This distribution serves as a reference for the distributions obtained by randomizing PFAM alignments according to various null models. Specifically, given a null model, the MSA for each PFAM is replaced with an alignment of the same length and thickness, and with the content generated according to the model (see Methods section).

The simplest model, randomly drawn MSA positions, fails to reproduce the heavy tail of the real distribution (Figure 1a, green line). On the other hand, comparison of real queries to MSAs of reversed directionality generates much closer distribution (Figure 1a, red line). This model preserves almost all features of real MSAs but provides only a limited number of scores due to the limited number of protein families. We aim to replace it with a model that produces a similar heavy-tailed distribution and generates a virtually unlimited statistical sample sufficient for precise analytical approximation.

Models based on combination of SSEs produce more realistic distributions than combining individual MSA positions (Figure 1a, see also Supplementary Figure S1). Among these models, we observe two notable trends. First, real MSA comparison is better mimicked by concatenating fragments of numerical profiles, as opposed to fragments of original MSAs (Figure 1a). Second, score distributions are closer to real when real profile segments are replaced with randomly drawn fragments that represent SSE of similar properties. Several SS properties are important to model: types of SSE, their length (number of MSA positions in each element), and thickness (sequence number and diversity). Reproducing more of these features results in more realistic score distributions (Supplementary Figure S1). The most elaborate model, which preserves SS type, length and thickness of profile segments, closely reproduces the distribution generated by the reversed profile comparisons (Figure 1a). We further consider this model in more detail.

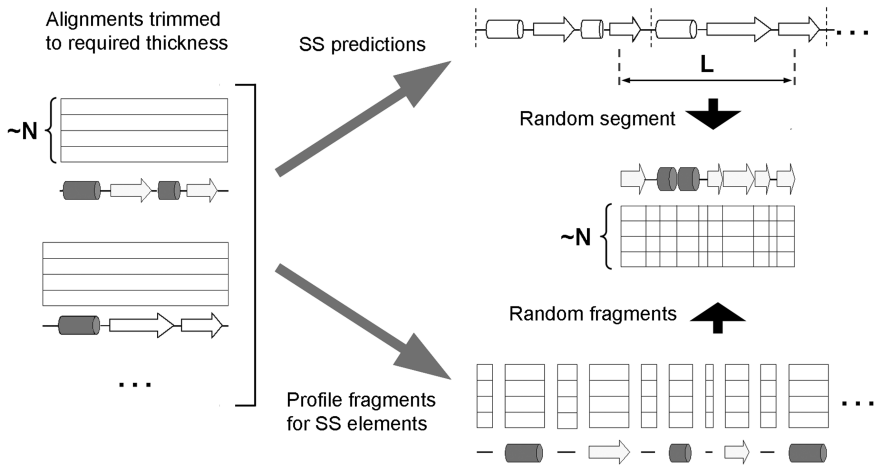

Figure 2. Large-scale generation of decoy profiles with native-like secondary structure (schema). First, we process the set of real profiles to make their thickness close to the desired value. Second, based on three-state SS predictions (helix, strand or loop), we define the profile fragments that will serve as elementary construction blocks of randomized profiles. Third, we concatenate SS predictions of real profiles, and prepare a long SS template. Finally, we select a random template segment of the desired length and fill this segment with randomly drawn profile fragments of matching SS types, splicing and cutting them to fit the length of the original SSE. The resulting profile reproduces a realistic SS arrangement and has the required length and thickness.

Analytical approximation of random score distributions

Our model of choice represents a profile as a combination of randomly drawn real profile fragments corresponding to SSEs that are arranged to mimic a real SS segment. This arrangement mimics the type (helix, strand or loop), order and lengths of predicted SSEs in real profiles, as well as reproduces the given profile thickness and length. Having this model, our second goal is to develop a mathematical function that precisely describes the score distribution produced by the model. To achieve this precise description, we generate large samples of randomized profiles, compare them to each other and search for a high-quality analytical approximation of the resulting distributions.

Statistical significance of the similarity between two profiles should be derived from the distribution attuned to the profiles' properties. Similarly to the comparison of individual sequences (27,28), the score distribution for random profile comparisons depends on protein length (4): longer profiles have a higher probability of containing similar segments by chance. We also find that the distribution depends on profile thickness (number and diversity of sequences): in thicker profiles, patterns of conservation are based on a more representative sampling, and the similarity between such patterns should receive higher scores. Thus, we generate multiple score samples for various values of profile length and thickness, and describe the dependency of the distribution's parameters on these profile properties.

The procedure for a large-scale generation of randomized profiles of a given length and thickness is schematically shown in Figure 2. In brief, we concatenate SS predictions for a set of PFAM profiles, forming a long SS template. To generate a randomized profile, we select a random template segment of the desired length and fill this segment with randomly drawn profile fragments for SSE that match the properties of the original SSE (Figure 2). We choose 464 combinations of profile length and thickness, produce 1 million pairs of such profiles for each combination, and calculate a similarity score for each pair (see Methods section for details).

The resulting score distributions (Figure 3a and b, see also Supplementary Table 1) deviate from the EVD. EVD emerges from the model assuming independence of residue content at different sequence positions, which corresponds to profiles generated by the random sampling of individual columns. As additionally confirmed by our results (Supplementary Figure S2), this model adequately describes the comparison of single sequences (13,28). In profile comparison, however, correlations between amino acid preferences at neighboring protein positions leads to a non-EVD shape of the score distribution.

In a search for the precise statistical description of the generated scores, we test several standard distributions (see Supplementary Data). None of these functions can provide an adequate fit of empirical data. EVD, a simple two-parameter distribution, fits the data comparably to other distributions involving up to four parameters. Thus, in order to improve the fitting precision, we attempt to modify EVD by introducing additional parameters. One of the tested EVD modifications fits the empirical data with high accuracy.

Probability density function (PDF) of EVD contains an exponential and a linear term in exponent, with the linear term defining the function's behavior at the tail, the area of our main interest:

$$f(x) = C_1 \exp\left(-e^{-\frac{x-m}{s}} - \frac{x-m}{s}\right) \quad \mathbf{1}$$

where C_1 is a normalizing constant. To approximate empirical distributions generated by our model, we add flexibility in the tail by replacing the linear term with a

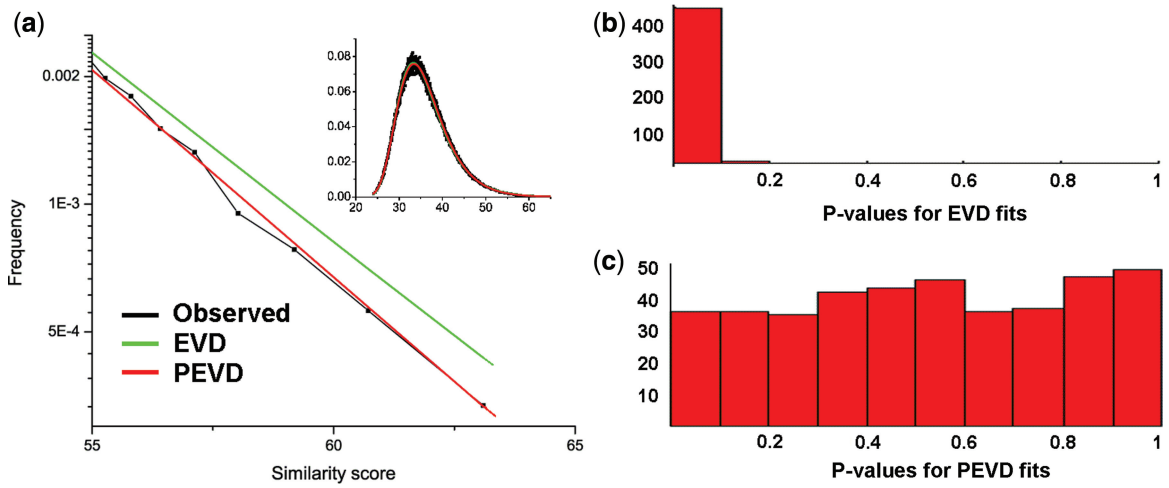

Figure 3. A new statistical distribution, PEVD, precisely fits the distributions of simulated profile similarity scores. (**a**) An example of fitting a simulated score distribution of 10^6 scores with EVD and with PEVD. The distribution's tail is shown in a logarithmic scale. Inset: full plot in a linear scale. (**b** and **c**) Histograms of the χ^2-goodness-of-fit *P*-values for multiple individual distributions. (b) EVD fits of the distributions are assigned mostly low *P*-values. (c) *P*-value histogram for PEVD is close to the ideally expected uniform distribution.

power function and a multiplicative parameter to balance the linear and exponential terms:

$$f(x) = C_2 \exp\left(-e^{-\frac{x-m}{s}} - \beta\frac{x^\alpha - m^\alpha}{s^\alpha}\right) \quad \mathbf{2}$$

where C_2 is a normalizing constant. This function (Figure 3a) fits the empirical data remarkably better than EVD, as confirmed by the χ^2-goodness-of-fit test (Figure 3b and c). Even the large score samples are statistically indistinguishable from random sampling from this distribution. To our knowledge, this type of distribution has not been previously discussed in the literature; we will further refer to it as 'power-EVD' or PEVD. PEVD is defined by four parameters: location parameter m, width parameter s, and two shape parameters α and β. For positive m, s, α and β, the function is defined on the half-axis $x \geq 0$, but since dynamic programming algorithms for the construction of profile alignments result in positive scores only, this limitation does not create a problem. The first term in the exponent rapidly decreases to zero for large x, while the power term defines the shape of the tail.

Fitting the generated score distributions with PEVD shows that its four parameters depend on profile length and thickness. We describe these dependencies with approximate empirical formulas that allow calculation of the PEVD parameters for comparison between any two profiles, given their length and thickness. These four formulas, one for each PEVD parameter, include a total of 32 parameters (see Methods section and Supplementary Data for details) whose values are optimized for the best fit on a set of score distributions for various profile properties. Specifically, we select a subset of empirical distributions for multiple combinations of profile length and thickness and minimize the sum of χ^2-values for their fits with PEVDs, whose parameters are determined by the formulas (Figure 4, see Supplementary Data for details). When we test the optimized formulas on the full set of distributions, a good quality of fit is confirmed by the χ^2 *P*-values (Figure 4b). As expected, the quality of this combined fit is worse than the quality of the PEVD fits individually optimized on each sample of a given length and thickness; however, it is considerably better than the fits achieved by EVD, even individually optimized (Figure 4c).

Model evaluation

Our final goal is to test whether the best null model, applied to the estimation of *E*-values, can improve the quality of similarity detection among proteins. Using the formulas for the dependency of PEVD parameters on the profile length and thickness, we implement the PEVD-based calculation of *E*-value and perform all-to-all comparisons within two different sets of alignments. The first set includes ~1800 PFAM alignments used for the generation of random profiles. The second, testing set includes ~2900 MSA produced by PSI-BLAST (25), with sequences of representative structural domains from SCOP (22) as queries (see Methods section for details). In both sets, the relation between proteins is judged by the similarity of their 3D structure (see Supplementary Data).

Given the COMPASS score for each alignment pair, we compare the new *E*-value estimates to those produced by the previously used null models: (i) model of randomly drawn profile positions and (ii) models based on the comparison of each individual query to the database of real profiles, or profiles with reversed directionality. In addition to these COMPASS-based estimates, we assess *E*-values produced by other methods for sequence and alignment comparison. Our main goal, however, is to evaluate the effect of different null models applied to the same set of scores that are produced by a single method for MSA comparison (COMPASS).

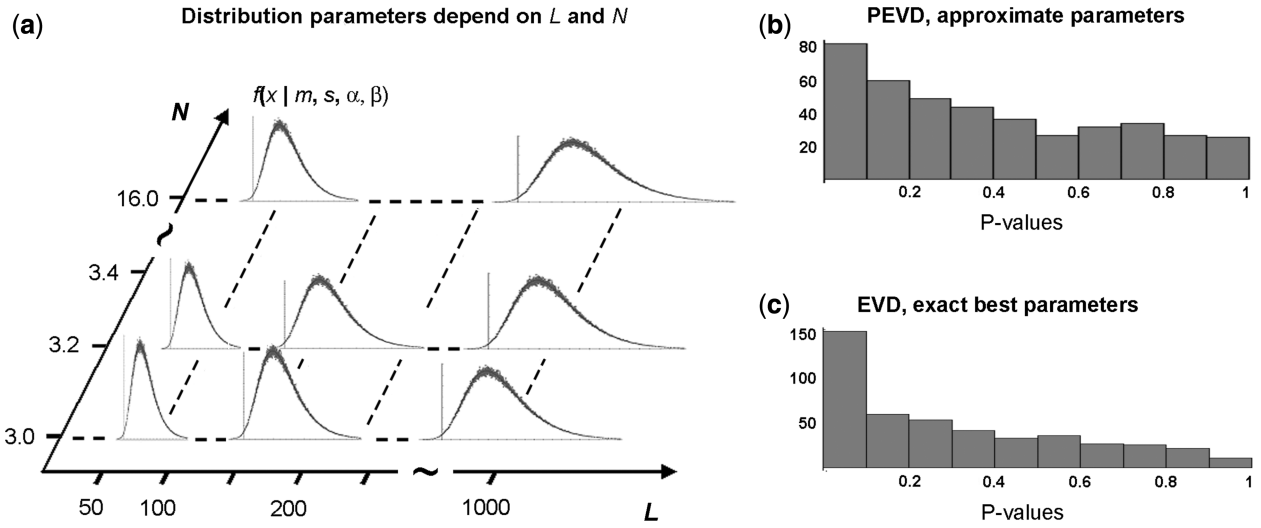

Figure 4. Approximating the dependency of the distribution parameters on profile length and thickness (sequence diversity). (**a**) Distribution of similarity scores depends on length L and thickness N of the compared profiles. A random sample of 10^6 scores is generated for multiple combinations of L and N. The resulting score distributions are used for constructing the formulas describing the dependency of the PEVD parameters (Equation (2)) on length and thickness (see also Supplementary Data). (**b** and **c**) The constructed approximate formulas provide good-quality PEVD fits for the testing set of distributions, as demonstrated by the χ^2-goodness-of-fit tests. (b) Histogram of the χ^2-test P-values for PEVD, with the parameters approximated by the formulas. (c) The same histogram for EVD, with the best possible parameters obtained by fitting each distribution individually.

We use two evaluation criteria. First, we assess the retrieval accuracy, i.e. the ability to discriminate between true and spurious hits (separate homologs from non-homologs). For this evaluation, we use ROC (receiver operating characteristic) (29). More specifically, we rank the list of all hits by ascending E-value and plot the number of true positives versus the number of false positives while moving from the top of this list. Second, we assess the statistical accuracy, i.e. the closeness between the predicted and actual numbers of false positives that are assigned a given E-value (11). Specifically, for the highest-ranking false positive hit of each query, we transform E-value into the predicted P-value and plot these P-values against the actual rate of top false positives (see Methods section for details). This plot reveals the accuracy of P-values in predicting the rate of false positives. The most accurate prediction should correspond to the identity line. Deviations of the curve from the identity line reveal biases in the E-value estimates for the top false positives.

The new statistical model improves both the retrieval and the statistical accuracies. Figure 5 shows the evaluation results for COMPASS E-values based on different models, as well as sequence–sequence comparison by SSEARCH (30,31), profile–sequence comparison by PSI-BLAST (28) and MSA comparison by HHsearch (7) (version comparable to COMPASS, using the sequence information not aided by SS prediction) and PRC (8,32) (evaluations of other models and methods for profile comparison are shown in Supplementary Figures S9 and S10). ROC curve for the PEVD-based model is significantly better than for the original EVD model (Figure 5a). The improvement is especially pronounced in the area of lower E-values (at the beginning of the ROC curves), consistent with the more accurate fit at the tails of empirical score distributions (Figure 3). The relative increase in performance is even higher on the PFAM-based testing set (Supplementary Figure S9a).

Most importantly, the new model surpasses conventional empirical models based on comparison of a query to the database of real or reversed profiles (Figure 5a, see also Supplementary Figure S10a). These empirical models closely capture the properties of both the query and the database, and provide a relatively high statistical accuracy for top false positives (Figure 5b). However, regardless of the function (EVD or PEVD) used to fit the empirical score distributions for each query, the retrieval accuracy of these models is inferior to the new model. The accuracy is also lower for the previously proposed model based on normalizing each individual score by the score of reversed comparison (20).

Similar to others (11,25,33), we find that the assessment of statistical accuracy (Figure 5b) does not necessarily correlate with retrieval accuracy. When comparing different methods, a higher quality of statistical prediction may coincide with a lower quality of hit ranking. For instance, SSEARCH most accurately predicts the actual numbers of false positives produced by the method, whereas the PSI-BLAST and the new COMPASS E-values tend to underestimate these numbers (Figure 5b). The retrieval accuracy of these three methods, however, has the reverse order: COMPASS > PSI-BLAST > SSEARCH (Figure 5a).

In contrast, for null models of the same family applied to the same method, COMPASS, retrieval accuracy follows statistical accuracy: in our set of null models, higher modeling precision provides better performance in both evaluations (Supplementary Figures S9 and S10). We also find that plots of statistical accuracy can be helpful in the method's development: since only top false positives are shown, such plots are sensitive to even rare cases of

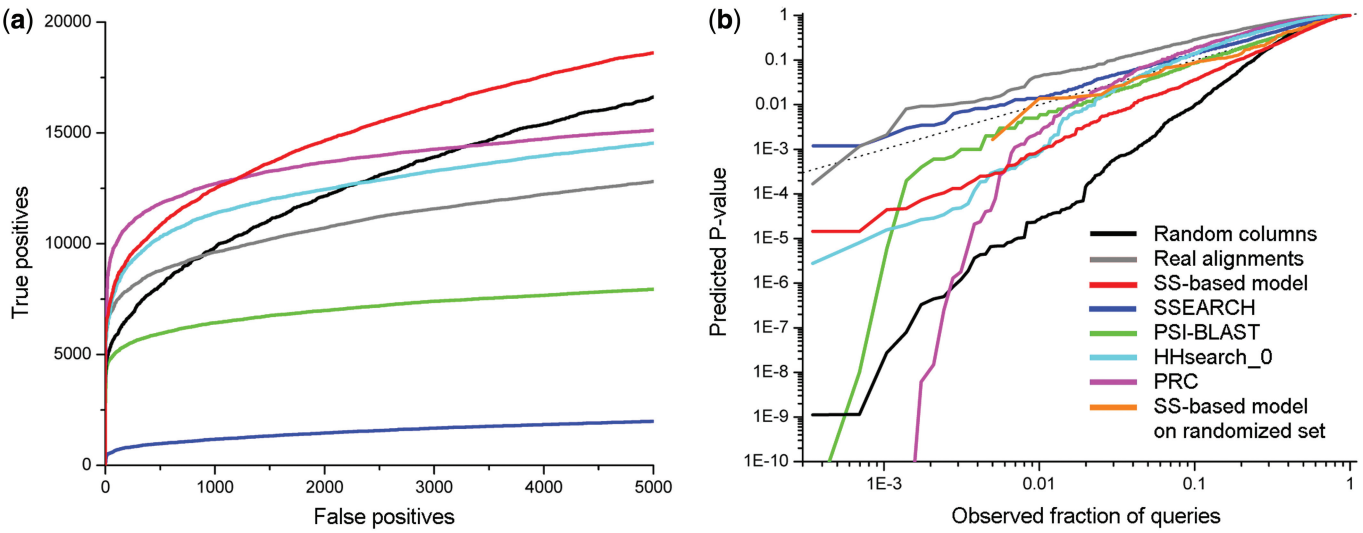

Figure 5. Assessment of model's performance in detecting protein similarities. Plots of retrieval and statistical accuracy for the comparison of single sequences (SSEARCH), alignments to sequences (PSI-BLAST), and alignments to alignments, HHsearch_0 (version of HHsearch comparing residue content of MSA but not predicted secondary structure), PRC and COMPASS, the latter based on three different null models: comparison of profiles composed of random profile positions ('Random columns'), comparison of each query to unrelated profiles from the database ('Real'), and the secondary structure model described in the text ('SS-based model'). Results are shown for the testing set of alignments produced by PSI-BLAST for 2900 representative SCOP domains as queries (see Supplementary Figure S10 for more results). (**a**) Retrieval accuracy (ROC curves). (**b**) Statistical accuracy (predicted *P*-value for a query's top false positive Vs observed rate of such false positives among the queries). As a control for the precision of the model's analytical description, this graph also includes the plot for SS-based model applied to the comparison of randomized alignments ('SS-based model on randomized set').

E-value underestimation. The analysis of such cases may reveal method's biases, for instance towards certain rare combinations of profile properties (length, thickness, residue composition, etc.).

Although the new random model improves statistical accuracy, the produced *P*-values still underestimate the actual number of false positives by ~1.5 orders of magnitude (Figure 5b). This underestimation may stem from two potential sources: (i) decoy profiles incompletely represent real MSA features, and (ii) generated distributions are imprecisely approximated. We rule out source (ii) by showing that the estimated *E*-values exactly correspond to the comparison of randomized profiles. Specifically, on the set of PFAMs randomized according to the model (Figure 1b, see Methods section), the produced *E*-values have high statistical accuracy (Figure 5b). Thus, we hypothesize that the observed relatively modest underestimation is caused by features of real MSAs that are not grasped by our model. Such features may include recurring nonlocal patterns of distant profile positions that reflect super-secondary and tertiary motifs in unrelated protein structures.

DISCUSSION

By considering patterns of evolutionary conservation in protein families, methods for MSA comparison produce similarity scores that allow for sensitive homology detection. It is unclear, however, whether the current simplistic statistical treatments of these scores can fully realize the potential of the methods. This practical motivation leads to a more general question: what is a realistic null model of a protein sequence family? Ultimately, can accurate modeling further improve discrimination between homologs and nonhomologs? These questions are not trivial: for example, in the case of comparing MSA to a sequence, more realistic null models had no significant effect on the quality of homology detection (20,25). To answer these questions, we set the goal of modeling, as closely as possible, the comparison between unrelated protein families. We then assess the effect of the improved model on the homology detection.

Modeled protein features

In this work we, for the first time, develop null models capturing, to various levels of detail, MSA patterns associated with protein secondary structure. First, we find that combining fragments of numerical profiles is more realistic than combining original MSA fragments. We attribute this effect to the distortion of evolutionary distances between member sequences in the latter case. When residue content at the MSA positions is calculated, the contribution of redundant sequences is down weighted. Combining random sequence fragments makes individual sequences more equidistant, which equalizes their contributions to the effective residue content. This distortion is avoided in the models that combine fragments of precomputed profiles.

Perhaps not surprisingly, simple random combination of profile fragments for SS elements is not the best way to reproduce the real score distribution. The majority of native proteins has a distinct SS arrangement, such as all α, all β, α/β and $\alpha + \beta$ classes in the SCOP classification (22). We find that reproducing these classes by mimicking the types and lengths of real SSE results in a more realistic random score distribution. A less expected

finding is the importance of thickness (sequence diversity) of profile fragments in randomized profiles. In COMPASS, profile similarity scores are explicitly affected by the sequence diversity, because both frequencies and counts of amino acids are considered (4). This effect might be less important when alignment comparison is based solely on residue frequencies, as in many other current methods (1–3,7,8).

The developed null model successfully reproduces the distribution of similarity scores generated by the highly accurate model based on reversed profile comparison, which mimics all features of native proteins except for directionality (20,34) (Figure 1).

Mathematical description of random score distribution

Based on the realistic procedure for decoy generation, we construct a precise analytical approximation of the resulting similarity score distributions. The high accuracy of the approximation is achieved by (i) using large statistical samples of scores; (ii) considering distributions for many combinations of the profile properties; and (iii) introducing a new analytical function, PEVD, that precisely fits empirical distributions. Even using approximate dependencies of the distribution parameters on profile length and thickness (Figure 4), PEVD can fit empirical data better than EVD with exact optimal parameters. High precision of this approximation is confirmed by the statistical accuracy of *E*-value estimates applied to the database of randomized profiles (Figure 5b).

Effect of the null model on homology detection

The main purpose of assigning *E*-values is the compatibility of the results between different queries, so that, for example, a universal *E*-value threshold can be established for a hit to be considered significant. Depending on the query's properties (length, thickness, composition, etc.), the significance of the same score value can vary dramatically. A common way of addressing this problem is to derive *E*-values from score distributions produced by comparisons of an individual query with either nonhomologous or reversed profiles from the database. However, relatively small sample size may lead to biases in analytical approximations of these distributions.

The null model proposed here reproduces important features of native proteins and yet allows for the generation of virtually unlimited statistical samples, providing an accurate analytical description. Remarkably, this model results in better homology detection than any other tested model. This result suggests that with the same similarity scores, homology detection can be improved by changing only the null model of profile comparison.

The quality of homology detection based on the new null model is also compared with the other methods for MSA comparison from the same class as COMPASS (Figure 5), HHsearch (version comparing residue content but not SS) (7) and PRC (8,32). Both HHsearch and PRC represent MSAs in the form of hidden Markov models (HMMs) rather than numerical profiles, which gives the advantage of adjustable gap penalties, depending on the content of individual MSA positions, in the construction of HMM–HMM alignments. These methods derive *E*-values from comparisons of each individual query with MSAs of nonhomologous families or reversed MSAs from the searched database. In the detection of close homologs (the area corresponding to the ROC plots near zero), PRC outperforms the new COMPASS-based model (Figure 5A). However, for more remote homologs, the performance of the new null model applied to the COMPASS profile alignments compares favorably with HMM-based methods, suggesting a practical value of improved statistical modeling.

Several sensitive methods have been proposed that involve the comparison of predicted SS, in addition to the residue content of the two MSAs (7,35,36). These methods are directly using the SS information in the construction and scoring of profile–profile or HMM–HMM alignments. Here we do not present another such method. We use the alignments and similarity scores based only on the residue content of the compared sequence families, and develop a formula for a fast calculation of corresponding *E*-values. Although the formula is based on the null model that generates decoys with realistic SS arrangements, the resulting *E*-values do not exploit SS of the specific compared protein pair. Applying the described modeling approach to the methods that involve the comparison of SS predictions may further improve the quality of homology detection by these methods.

Applicability of the model to other methods for homology detection

The proposed approach can be readily applied to any method of profile comparison. Native MSA profiles generated by a method of choice can be split into fragments corresponding to the predicted SS, and these fragments can be randomly concatenated according to a real SS template, as described earlier (Figure 2). The resulting decoys can be submitted to the method of interest, and random score distributions can be constructed and analytically approximated.

A potential complication of this process might be the dependence of the score distributions on the residue composition of the random profiles. For profiles with similar overall residue content, accidental high-scoring similarity should be more likely. In COMPASS, compositional biases are specifically addressed at the stage of score calculation (4). For a method without internal correction for compositional biases, the model might require considering additional dependence on the profile composition.

Protein features not captured by the model

Evaluation of statistical accuracy suggests that the new null model results in a modest underestimation of top false positive rates for individual queries (Figure 5b). Since the accuracy of the analytical approximation is confirmed in a control experiment on randomized profiles (Figure 5b), we hypothesize that the observed underestimation is caused by the shortcomings of the decoy generation itself. One of the native protein features not captured by

the model is the presence of nonlocal sequence patterns. Such patterns may correspond, for example, to distant sequence positions that belong to a 3D-structural motif, and are impossible to reproduce by local SS modeling. Another deviation introduced by the model is the alteration of native SS structure by replacing a single SSE with multiple concatenated SSEs, or with a trimmed SSE of the same type. This replacement (i) introduces local breaks in the residue periodicity at the boundaries of concatenated SSEs, and (ii) distorts common sequence patterns associated with ends of SSE, such as α-helical caps. These inaccuracies, however, appear to cause only second-order effects, compared to the general SS modeling.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

This study was supported by NIH grant GM67165 to NVG. 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, James Wrabl, Erik Nelson and Dorothee Staber for discussions and critical reading of the article. Funding to pay the Open Access publication charges for this article was provided by Howard Hughes Medical Institute.

Conflict of interest statement. None declared.

REFERENCES

1. Pietrokovski,S. (1996) Searching databases of conserved sequence regions by aligning protein multiple-alignments. *Nucleic Acids Res.*, **24**, 3836–3845.
2. 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.
3. 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.
4. Sadreyev,R. and Grishin,N. (2003) COMPASS: a tool for comparison of multiple protein alignments with assessment of statistical significance. *J. Mol. Biol.*, **326**, 317–336.
5. Ginalski,K., von Grotthuss,M., Grishin,N.V. and Rychlewski,L. (2004) Detecting distant homology with Meta-BASIC. *Nucleic Acids Res.*, **32**, W576–W581.
6. 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.
7. Soding,J. (2005) Protein homology detection by HMM-HMM comparison. *Bioinformatics*, **21**, 951–960.
8. 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.
9. Battey,J.N., Kopp,J., Bordoli,L., Read,R.J., Clarke,N.D. and Schwede,T. (2007) Automated server predictions in CASP7. *Proteins*, **69 (Suppl 8)**, 68–82.
10. 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.
11. Pearson,W.R. and Sierk,M.L. (2005) The limits of protein sequence comparison? *Curr. Opin. Struct. Biol.*, **15**, 254–260.
12. Wang,G. and Dunbrack,R.L. Jr. (2004) Scoring profile-to-profile sequence alignments. *Protein Sci.*, **13**, 1612–1626.
13. 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.
14. Gnedenko,B. (1943) Sur la distribution limite du terme maximum d'une serie aleatoire. *Ann. Math.*, **44**, 423–453.
15. Gumbel,E.J. (1958) *Statistics of Extremes*. Columbia University Press, New York, NY.
16. 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.
17. 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.
18. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
19. 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.
20. Karplus,K., Karchin,R., Shackelford,G. and Hughey,R. (2005) Calibrating E-values for hidden Markov models using reverse-sequence null models. *Bioinformatics*, **21**, 4107–4115.
21. Waterman,M.S. and Vingron,M. (1994) Rapid and accurate estimates of statistical significance for sequence data base searches. *Proc. Natl Acad. Sci. USA*, **91**, 4625–4628.
22. 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.
23. Jones,D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.*, **292**, 195–202.
24. Gribskov,M., McLachlan,A.D. and Eisenberg,D. (1987) Profile analysis: detection of distantly related proteins. *Proc. Natl Acad. Sci. USA*, **84**, 4355–4358.
25. Schaffer,A.A., Aravind,L., Madden,T.L., Shavirin,S., Spouge,J.L., Wolf,Y.I., Koonin,E.V. and Altschul,S.F. (2001) Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res.*, **29**, 2994–3005.
26. Finn,R.D., Mistry,J., Schuster-Bockler,B., Griffiths-Jones,S., Hollich,V., Lassmann,T., Moxon,S., Marshall,M., Khanna,A., Durbin,R. *et al.* (2006) Pfam: clans, web tools and services. *Nucleic Acids Res.*, **34**, D247–D251.
27. Altschul,S.F. and Gish,W. (1996) Local alignment statistics. *Methods Enzymol.*, **266**, 460–480.
28. 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.
29. Gribskov,M. and Robinson,N.L. (1996) Use of receiver operating characteristic (ROC) analysis to evaluate sequence matching. *Comput. Chem.*, **20**, 25–33.
30. Pearson,W.R. (1998) Empirical statistical estimates for sequence similarity searches. *J. Mol. Biol.*, **276**, 71–84.
31. Pearson,W.R. and Lipman,D.J. (1988) Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.
32. Madera,M. (2006) PRC – The profile comparer. *PhD Thesis*. University of Cambridge.
33. Yu,Y.K., Gertz,E.M., Agarwala,R., Schaffer,A.A. and Altschul,S.F. (2006) Retrieval accuracy, statistical significance and compositional similarity in protein sequence database searches. *Nucleic Acids Res.*, **34**, 5966–5973.
34. Taylor,W.R. (1986) Identification of protein sequence homology by consensus template alignment. *J. Mol. Biol.*, **188**, 233–258.
35. Chung,R. and Yona,G. (2004) Protein family comparison using statistical models and predicted structural information. *BMC Bioinformatics*, **5**, 183.
36. 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.