Comparative evaluation of word composition distances for the recognition of SCOP relationships

Susana Vinga\textsuperscript{1}, Rodrigo Gouveia-Oliveira\textsuperscript{1} and Jonas S. Almeida\textsuperscript{1,2,}\textsuperscript{*}

\textsuperscript{1}Biomathematics Group, ITQB, Universidade Nova de Lisboa, Rua da Quinta Grande, n. 6, 2780-156 Oeiras, Portugal and \textsuperscript{2}Department Biometry and Epidemiology, Medical University South Carolina, 135 Cannon Street, Suite 303, P.O. Box 250835, Charleston, SC 29425, USA

Received on April 23, 2003; revised on June 27, 2003; accepted on July 11, 2003

ABSTRACT

Motivation: Alignment-free metrics were recently reviewed by the authors, but have not until now been object of a comparative study. This paper compares the classification accuracy of word composition metrics therein reviewed. It also presents a new distance definition between protein sequences, the W-metric, which bridges between alignment metrics, such as scores produced by the Smith–Waterman algorithm, and methods based solely in L-tuple composition, such as Euclidean distance and information content.

Results: The comparative study reported here used the SCOP/ASTRAL protein structure hierarchical database and accessed the discriminant value of alternative sequence dissimilarity measures by calculating areas under the Receiver Operating Characteristic curves. Although alignment methods resulted in very good classification accuracy at family and superfamily levels, alignment-free distances, in particular Standard Euclidean Distance, are as good as alignment algorithms when sequence similarity is smaller, such as for recognition of fold or class relationships. This observation justifies its advantageous use to pre-filter homologous proteins since word statistics techniques are computed much faster than the alignment methods.

Availability: All MATLAB code used to generate the data is available upon request to the authors. Additional material available at http://bioinformatics.musc.edu/wmetric

Contact: svinga@itqb.unl.pt; almeidaj@musc.edu

INTRODUCTION

Bioinformatics applications rely heavily on sequence comparison techniques, from searching a database with a query DNA sequence to the classification of protein domains.

In most cases, alignments are performed between the target sequences and the resulting alignment scores are used to calculate a measure of dissimilarity. In protein comparison, the scoring methods depend on amino acid mutation rate information, represented as scoring matrices, and find optimal alignments between sequences by dynamic programming techniques. Alignment scores are particularly useful when sequences are known to be closely homologous since the more conserved regions are automatically detected. However, for remote homologues this approach tends to fail: proteins with $<20\%$ identity, a region sometimes referred to as the 'twilight zone', are not satisfactorily aligned neither its similarity detected (Pearson, 2000). It is also noteworthy that dynamic programming is computationally intensive and consequently unpractical for querying large datasets, which forces the use of some heuristics to reduce the running times, as exemplified by BLAST.

In a recent paper (Vinga and Almeida, 2003), the authors reviewed alignment-free methods for sequence comparison but did not compare them quantitatively. In that review metrics based in L-tuple composition, the focus of this report, emerged as the alignment-free technique most often proposed by other researchers. In these algorithms each sequence is mapped to an $n$-dimensional vector according to its word composition. Linear Algebra theory is further employed to define distances between sequences represented in those vector spaces, namely by using Euclidean (eu) distance and Information content (see Review for a full description and related references).

This report also presents a novel distance function between protein sequences, the W-metric (Wm), which tailors L-tuple composition methods with techniques based in alignment. This is accomplished by defining a function that includes both one-tuple composition information, more specifically the differences in amino acid content between two proteins, and weights from the scoring matrices used in alignment methods. Although these two concepts are not new, their conjugation
constitutes the novelty aspect of this metric. The weights correspond to the estimation of log-likelihood ratios between probabilities of symbols that best describe mutation rates in known homologous proteins, thus providing evolutionary information.

The usefulness of the L-tuple composition approach is associated with its light computational load, which makes it very useful in pre-filtering relevant sequences, and then using alignment algorithms to refine the searches. This type of heuristic approach is already used in programs like BLAST (Altschul et al., 1990) and FASTA (Pearson and Lipman, 1988). Although the solution may not be the optimal, it drastically shortens processing speed to the point that the method can be used to query large databases. However, a comparative study of the effectiveness of alignment-free sequence dissimilarity measures is, to the authors’ best knowledge, absent from the literature. Consequently, it is difficult to decide at what similarity level are alignment methods required. Such a comparative study of how these different metrics perform is reported here. This is the main motivation for the present work, where alignment-free, linear algebra type methods are comparatively assessed. Some previous studies have reported comparative assessments of various methods (Brenner et al., 1998; Lindahl and Elofsson, 2000; Pearson, 1991, 1995), but not consistently for the same reference dataset. These studies showed, however, the importance of following an extensive protocol involving as many examples as possible in the assessment of any classification procedure. Only then it is possible to improve some heuristics commonly applied in sequence similarity searches and identify the best algorithmic choice for each problem category.

We compared L-tuple metrics with Smith–Waterman (SW) algorithm by Receiver Operating Characteristic (ROC) curves applying the algorithms to a subset of Structural Classification Of Proteins (SCOP)/ASTRAL database. This database constitutes the reference gold standard for protein secondary structure classification, which makes it a commonly used benchmark for protein structure prediction algorithms, a crucial field in Computational Biology applications. In addition, it has a hierarchical organization that can be browsed to assess classification accuracy for each of its levels.

SYSTEMS AND METHODS

In the section below the W-metric, a novel word-statistic distance between protein sequences is presented as well as additional background on alignment-free algorithms. In the subsequent sections, the reference protein datasets and the methods used to compare the distance measures are described. Finally, the last two sections describe the algorithms and protocol used and its implementation.

Word statistics

There is a large body of literature on Word Statistics (Reinert et al., 2000), where sequences are interpreted as a succession of symbols and are further analysed by first representing the frequencies of its small segments (L-tuples or words). This approach does not take into account any of the physico-chemical or structural properties of the amino acids or nucleotides. There are also an increasing number of studies focusing on distance definition in the frequency space of L-tuples. These definitions are a fundamental step for the subsequent application of exploratory analysis methods, such as cluster analysis and dimensionality reduction techniques. In a recent review (Vinga and Almeida, 2003), the authors overviewed these metrics and their application to biological sequences, both DNA and proteins. That review will be used as the main reference for description of the L-tuple distances and alignment-free algorithms that will be tested here. A protein X of length n is a sequence of symbols from the alphabet of all possible amino acids: 

\[ X = s_1 \ldots s_n, \quad s_i \in A = \{A, R, N, D, \ldots, V\}. \]

The mapping of X into the Euclidean space can be defined by representing X by its amino acid composition in counts, \( c_X \) and frequencies, \( f_X \) [Equation (1)]:

\[
  c_X = (c_X^A, c_X^R, c_X^N, c_X^D, \ldots, c_X^V) \\
  f_X = \frac{c_X}{n}
\]

For example, the peptide \( X = AARNNNDAA \) is mapped on to the vectors \( c_X = (4, 1, 2, 1, 0, 0, \ldots) \) and \( f_X = (0.5, 0.125, 0.25, 0.125, 0, 0, \ldots) \). Instead of single amino acid frequencies, longer fragments of length L could be considered (L-tuples) with resulting 20L long vector of frequencies. One can further define a distance or dissimilarity measure between two proteins X and Y, \( d(X, Y) \), based on their corresponding vectors \( f_X \) and \( f_Y \).

W-metric definition

The novel Wm hereby proposed to complement existing word composition methods is based on the quadratic form defined in Equation (2). The distance between two proteins X and Y, \( d^W(X, Y) \), is defined by their corresponding one-tuple frequencies, \( f_X \) and \( f_Y \), weighted by matrix W below described.

\[
  d^W(X, Y) = (f_X - f_Y)^T \cdot W \cdot (f_X - f_Y) \\
  = \sum_{i \in A} \sum_{j \in A} (f_i^X - f_i^Y) \cdot (f_j^X - f_j^Y) \cdot w_{ij}
\]

These quadratic forms play an important role in major theoretical and applied disciplines and scientific fields, from Linear Algebra to Econometrics. In Statistics they are used, e.g., in parameter estimation and statistical tests (Schott, 1997). They represent a scoring between conveniently weighted vectors of observations. It is noteworthy that other L-tuple distances are also based in quadratic forms [Equation (2)], e.g., when \( W \) is the covariance matrix of the data it represents Mahalanobis (ma) distance between the corresponding vectors and
the standard Euclidean (se) distance is obtained when taking only covariance matrix diagonal. The distance reduces to the squared Euclidean distance when $W$ is the identity matrix.

The weight matrices $W$ chosen in Equation (2) can be rationalized as being scoring-based amino acid substitution matrices, instead of covariance-based weights as in other distances. These matrices, such as Point Accepted Mutation (PAM) (Dayhoff et al., 1978) and BLOcks Substitution Matrices (BLOSUM) (Henikoff and Henikoff, 1992), are used in alignment-based methods and estimate the log-likelihood ratios between probabilities of symbols that best describe mutation rates in known homologous proteins. In particular, BLOSUMX matrix is estimated with ungapped aligned blocks of proteins sharing less than $X\%$ identity. PAM$n$ matrices account for evolutionary change in protein sequences and its estimation is based in the construction of phylogenetic trees, which are subsequently used to create a Markov Chain $n$-step transition matrix. This matrix is further transformed and normalized for conditional probabilities. For extensive description of this substitution matrices and some estimation examples, see Ewens and Grant (2001, section 6.5).

The key idea of Wm is to weight amino acid composition differences between two sequences, $f_i^X - f_i^Y$, according to its relative conservation in proteins known to be homologous. The overall distance between two proteins will be the sum of these weighted factors. For example, if an amino acid is highly conserved in known homologous sequences (high $w_{ij}$), two proteins with a very different frequency of this amino acid should be less similar than if the amino acids are ‘closer’ to each other in that statistical sense. If the opposite occurs, i.e. if an amino acid is known to have high mutational rates (low $w_{ij}$), the differences between its compositions in the two sequences being compared should be attenuated in the overall distance calculation. The same scheme applies to off-diagonal elements $w_{ij}$ ($i \neq j$); if there is a high mutation rate between these two amino acids, it means that $w_{ij}$ is higher than the corresponding weight of two amino acids very different, so this component should be weighted more. The main idea is thus weighting amino acid differences according to their similarity, given by known evolutionary information. The weighted metric hence includes both amino acid composition information, like other alignment-free techniques, and conserved homology information, as used to score the conventional alignment algorithms.

Some variations of this metric were also tested, namely using several normalization procedures. It is appealing the low computational load associated with the calculation expressed in Equation (2). It is not proven here, however, that the $W$ matrix associated with mutation information is the best in discriminating classification levels. This can be further accomplished by using Artificial Neural Networks (ANN) or other algorithms to optimize classification accuracy by finding a ‘better’ $W$ weighting matrix.

**ROC curve definition**

The methods that will be used here to assess and compare the accuracy of classification schemes and prediction algorithms are based on the analysis of ROC curves. This method goes back to signal detection and classification problems and is now widely applied in Medical diagnosis studies and psychometric analysis (Egan, 1975). This approach is employed in binary classification of data, usually categorized as positive (1) or negative (0) cases. The classification accuracy can be measured by plotting, for different threshold values, the number of true positives (TP), also named sensitivity or coverage versus false positives (FP), or (1 – specificity), encountered for each threshold, properly normalized [Equation (3)].

$$
\text{sensitivity} = \frac{\text{TruePositives}}{\text{Positives}} = \frac{TP}{TP + FN} \\
\text{specificity} = \frac{\text{TrueNegatives}}{\text{Negatives}} = \frac{TN}{TN + FP} \quad (3)
$$

A ROC curve is simply the plot of sensitivity versus (1 – specificity) for different threshold values. The area under a ROC curve (AUC) is a widely employed parameter to quantify the quality of a classifier because it is a threshold independent performance measure and is closely related to the Wilcoxon signed-rank test (Bradley, 1997). For a perfect classifier, the AUC is 1 and for a random classifier the AUC is 0.5. For additional results and comprehensive discussion of AUC measure, see Bradley (1997). Baldi et al. (2000); Brenner et al. (1998); Green and Brenner (2002) describe other possible classification accuracy measures not employed in this study.

**Protein test datasets—SCOP/ASTRAL classification**

The sequences used to perform the tests and compare different metrics are proteins from the SCOP database (Lo Conte et al., 2002; Murzin et al., 1995). This database consists of Protein Data Bank (PDB) entries and provides a detailed and reliable description of protein structure relationships and homology. The three-dimensional (3D) structure analysis allows the detection of more remote homologous, since structure is typically more conserved than sequence. The fundamental unit of classification is the protein domain, which is the basic element of protein structure and evolution. The ASTRAL compendium provides additional tools and data-sets (Brenner et al., 2000; Chandonia et al., 2002), namely the possibility to filter sequence sets where two different proteins have less than a chosen percentage identity to each other. This classification is a hierarchical description of proteins (Fig. 1). The first two levels, family ($fa$) and superfamly ($sf$), describe evolutionary relationships; the third one, fold
Fig. 1. SCOP/ASTRAL db—hierarchical classification of proteins. Example of Fibroblast growth factor receptor (FGFR2) classification in each of the four levels.

(cf), describes geometrical relationships or major structural similarity; and the fourth one represents protein structural class (cl). This will allow the study of each classifier for different levels of similarity.

Two different datasets were tested in order to assess the accuracy of each metric. The basic protein set, PDB40-B, was extracted directly from the ASTRAL web site and corresponds to SCOP database release 1.61 (November 2002). This subset includes all the sequences that share <40% identity to each other and has become a benchmark test set in the evaluation of methods to detect remote protein homologies (Brenner et al., 1998; Dubchak et al., 1999; Karwath and King, 2002; Lindahl and Elofsson, 2000; Luo et al., 2002; Park et al., 1997; Webb et al., 2002). This dataset was subsequently trimmed to exclude sequences with unknown amino acids and those belonging to families with <5 elements, thus obtaining the protein group named PDB40-v (Table 1). For example, there are 232 families with only one sequence, which is not informative regarding intra-family dissimilarity, which makes these domains insufficiently representative of a family. The effect of trimming the dataset was in this way also studied. Only the four major classes were included, namely all-α class, constituted mainly by proteins with α helix; all-β class, essentially formed by β-sheet structures; α/β class, proteins with mixtures of α-helices and β-strands; and α + β class, those where α-helices and β-strands are largely segregated. Other SCOP classes include multi-domain proteins, small proteins, theoretical models and other types, and were not included in this study. See Chothia et al. (1997) and SCOP documentation for description of protein folds and classification.

This study also considered separately another protein set from an outdated release of the SCOP database (1.35), the PDB40-b, due to the large amount of literature already published with those sequences (Luo et al., 2002 and corresponding references). Table 1 summarizes all the sequences sets examined in this paper.

Protocol for comparative assessment

The comparative test procedure followed in this report was based on a binary classification of each protein pair, where 1 corresponds to the two proteins sharing the same group in SCOP database, 0 otherwise. The group can be defined at one of the four different levels of the database: family (fa), superfamily (sf), class fold (cf) or class (cl), exploring the hierarchical organization of the proteins in that structure. Therefore, each protein pair is associated to four binary classifications, one for each level.

In order to compute the ROC curves, we calculated the distances between all possible protein pairs, according to the different metrics referred and briefly described below.

The similarity measure based on alignment tested was the Smith–Waterman raw score, with no correction for statistical significance, using score matrix BLOSUM50 and a linear gaping penalty scheme, with a gap penalty of 8. The distances based in L-tuple composition evaluated were W-metric, Euclidean, standard Euclidean, Kullback-Leibler (ku) discrepancy, Cosine (co) and (Mahalanobis). For the corresponding complete definitions and properties, see Vinga and Almeida (2003). In Wm calculations some alternative weighting matrices W [Equation (2)] were used: these included the scoring matrices BLOSUM50, BLOSUM40, BLOSUM62 and PAM250. The following normalization procedures were also applied: take only the diagonal of W, pass all its negative values to zero, use the exponential function of the original matrix and normalize by minimum and range. However, in this printed report only the results obtained with BLOSUM50 will be presented. The variations described are documented on the online annex.

For each metric, the distances between all proteins pairs were subsequently sorted, from maximum to minimum similarity, i.e. from the closest to the farthest pair. A perfect metric would completely separate negative from positive relationships, i.e. the maximum similarity would correspond always to the same group and the binary classification obtained after this distance sorting would be the vector (1, ..., 1, 0, ..., 0). Of course, this does not happen in practice, and the classes are interspersed. The ROC curves permit to assess the level of accuracy of this separation without choosing any distance threshold for the separation point. In particular, the AUC will give us a unique number of the relative accuracy of each metric and level, according to the SCOP classification scheme. We also tested each of the four classes separately with the same procedure, to evaluate hypothetical differences between the structural classes.

Computation

All the algorithms were implemented in MATLAB language (version 6 release 13). The code is available upon request to the authors.
RESULTS AND DISCUSSION

In the following sections, we present some of the results obtained. For extensive and additional results regarding all metrics and datasets see also the web page http://bioinformatics.musc.edu/wmetric, where the complete graphs and tables are shown (data not shown due to space limitations).

Complete dataset

ROC curves and AUC values The ROC curves obtained for the complete dataset (Table 1) are presented in Figures 2 (PDB40-v) and 3 (PDB40-b). As overviewed in the Systems and Methods section, a random classifier would have identical values of sensitivity and (1−specificity) for any threshold value considered (dashed diagonal).

Figures 4 and 5 provide graphs with the areas under ROC curves (AUC) obtained for both datasets and each SCOP level. The AUC values are typically used as a measure of overall discrimination accuracy.

As would be expected, Figures 4 and 5 show that the AUC decreases from family to class level for both datasets. The sequence similarity between proteins sharing the same family is still well recognized. Consequently, all the distances achieve their best discrimination accuracy at this level. At class level, classification relationships reflect similar structures, which can have completely different sequences and amino acid compositions. This underlies the observation that sequence similarity is lost, regardless of the metric, from family to class. The comparative discriminant value of the different metrics (Figs 4 and 5) shows two clear trends. First, at family level, alignment has a clear advantage, with AUC values of 0.86 and 0.81 (PDB40b and PDB40v sets), whereas all word-statistics metrics perform at or under 0.75 and 0.68, respectively. The most discriminant word-statistics metric at family level is the novel Wm introduced by this report (see Availability). It is interesting to note that, although defining a different score for each domain pair, the different matrices W produce the same score ordering. Similarly, all the normalization procedures did not lead to improved discrimination, producing worse classification results but are still made available in the same web page.

Variations in the Wm definition The Wm AUC values in the previous graphics were obtained using the scoring matrix BLOSUM50. The results using BLOSUM40, BLOSUM62 and PAM250 are virtually the same and will be omitted. Nevertheless, those results were compiled and are made available at the support web page (see Availability). It is interesting to note that, although defining a different score for each domain pair, the different matrices W produce the same score ordering. Similarly, all the normalization procedures did not lead to improved discrimination, producing worse classification results but are still made available in the same web page.

Table 1. Protein datasets used in this study

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Classes</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All-α</td>
<td></td>
</tr>
<tr>
<td></td>
<td>All-β</td>
<td></td>
</tr>
<tr>
<td></td>
<td>a/b</td>
<td></td>
</tr>
<tr>
<td></td>
<td>a+β</td>
<td></td>
</tr>
<tr>
<td>PDB40-B (1.61)</td>
<td>867 409 257 151 1051 362 213 111 1237 467 190 117 1065 487 307 212 4220</td>
<td></td>
</tr>
<tr>
<td>PDB40-v (1.61)</td>
<td>285 35 28 27 517 43 30 24 542 58 40 31 339 39 37 33 1683</td>
<td></td>
</tr>
<tr>
<td>PDB40-b (1.35)</td>
<td>220 128 97 73 309 150 115 54 285 154 98 66 240 147 115 80 1054</td>
<td></td>
</tr>
</tbody>
</table>

For each protein set, number of sequences or domains (do), families (fa), superfamilies (sf) and folds (cf), in each class. PDB40-B: sequences that share <40% to each other, current release (1.61) of SCOP/ASTRAL (not tested). PDB40-v: set derived from PDB40-B (1.61) by excluding sequences with unknown amino acids and families with <5 domains. PDB40-b: sequence dataset used by Luo et al. (2002), corresponds to previous release (1.35) of the same database.
Comparative evaluation of word composition distances

**Fig. 2.** ROC curves for PDB40-v dataset. Sensitivity (sen) versus 1—specificity (spe). SCOP levels: family (fa), superfamily (sf), class fold (cf) and class (cl). Metrics: Smith–Waterman (SW), W-metric (Wm), standard Euclidean (se), cosine (co), Kullback–Leibler (ku), Euclidean (eu) and Mahalanobis (ma). A random classifier would generate equal proportions of FP and TP classifications, which corresponds to the ROC diagonal (dashed line). Correspondingly, the better classification schemes have plots with higher values of sensitivity for equal values of specificity, resulting in higher values for the areas under the curve (AUC, see Text). SW is the best at family and superfamily levels. Wm and se outperform other alignment-free metrics. Standard Euclidean is the best at fold level for high sensitivity/low specificity values. For class level all metrics have similar results, slightly above random guessing.

**Higher order tuples** We also tested higher order word composition metrics, calculating 2- and 3-tuple distances between the domains for eu, se, ku and co. Somewhat intriguing was the fact that for all levels of classification discrimination worsened (see web page). However, it should be noted that the high dimension of the frequency vectors in these cases (respectively 400 and 8000) and the relative low dimension of the sequences length itself (mean values around 175 amino acids), caused the frequency vector \( f \) to be very sparse. Additional problems arising from this increased dimensionality of data are the need to increase sampling size in order to maintain accuracy, which goes along with the ‘curse of dimensionality’ (Donoho, 2000). Consequently, only the results obtained for one-tuples were presented in this report. The weighting proposed, as observed before for the one-tuple scenario, might not be the best for the recognition of the relationships. One idea worth exploring would be to extract some effective higher order tuples, by adequate selection of the weights, thus optimizing the classification accuracy and avoiding hopefully the dimensionality problem. However, this would lead to discriminatory and optimization procedures, which are out of the scope of this exploratory study.
Fig. 3. ROC curves for PDB40-b dataset. Sensitivity (sen) versus 1 − specificity (spe). SCOP levels: fa, sf, cf and cl. Metrics: Smith–Waterman (SW), W-metric (Wm), standard Euclidean (se), cosine (co), Kullback–Leibler (ku), Euclidean (eu) and Mahalanobis (ma). The classification accuracies for this dataset are slightly better than for the PDB40-v dataset (Fig. 2). The qualitative relation between the metrics is maintained.

Computational performance. It is noteworthy that the SW algorithm is computationally intensive. Its running times can be 1000-fold longer than that of the other metrics here compared. For example in PDB40-v dataset, SW took ~80 h and Wm just 5 min, using a 700 MHz PentiumIII with 1 GB total memory. The other word composition metrics themselves have varied computation implementation efficiencies (Vinga and Almeida, 2003).

Stratified analysis by class

AUC values. In order to compare the metrics, we also conducted additional studies for each of the four classes (all-α, all-β, α/β and α + β) separately. The AUC values are represented in Figure 6, for SW alignment scores and se distance, the two metrics that emerged as the most discriminant in the previous analysis (Figs 2–5) (see web page for similar analysis for the other metrics).

It is easier to recognize family relationships by alignment (Fig. 6, black symbols) for proteins belonging to class all-α, where values are above the overall accuracy (AUC values ranging from 0.70 to 0.87) and for α + β class (AUC from 0.70 to 0.91). The class where these relationships seem more difficult to detect was the class all-β, where we obtained the lowest AUC values for this level (0.60–0.77). For superfamily level, class α + β enables a surprising accuracy for both metrics (AUC from 0.70 to 0.90) as opposed to class all-β, where the superfamily relationships are still harder to detect only by sequence inspection (AUC between 0.55 and 0.64).
AUC values for fold level are much lower for all-α qualitatively the same (see web page) with a difference: the both metrics (0.69–0.81). The graph obtained for PDB40-b is Fig. 5.

At fold level, all-α class retains the higher AUC values for both metrics (0.69–0.81). The graph obtained for PDB40-b is qualitatively the same (see web page) with a difference: the AUC values for fold level are much lower for all-α and α + β classes for both metrics.

PDB40 version datasets comparison There is a significant improvement of discrimination accuracy for α + β class, in PDB40-v dataset. The difference in AUC values is constantly positive, for different metrics and levels, reaching a value as high as 0.21 at fold level with the SW alignment scores. It seems that the trimming procedure taken when obtaining PDB40-v set (see Systems and Methods) affected particularly all-α, all-β, α / β and α + β. Smith–Waterman is generally a better classification scheme—higher AUC values. At family level the best results are for proteins belonging to classes all-α and α + β; the lowest AUC values where obtained for class all-β. At superfAMILY level class α + β enables a surprising accuracy for both metrics as opposed to class all-β, which has the worse results. At fold level, all-α class retains the higher AUC values for both metrics.

PDB40-v dataset. The difference in AUC values is constantly positive, for different metrics and levels, reaching a value as high as 0.21 at fold level with the SW alignment scores. It seems that the trimming procedure taken when obtaining PDB40-v set (see Systems and Methods) affected particularly all-α and α + β classes. It is noteworthy these quantitatively differences obtained for the two datasets.

The α-helix and β-sheet content Judging from published reports, protein class classification is controversial. Some studies based class classification in the percentages of α-helix and β-sheets content of each chain. In a recent report, a schematic table was presented with different definitions (Eisenhaber et al., 1996). As noted in that study, there are some regions of the space defined by those percentages that are not clearly classifiable. It is in this uncertainty context that SCOP offers a classification that is a global measure and takes into account all the structural information of all chains in a protein.

In order to assess the correct assignment to classes, and avoid arbitrary classification, we extracted the α and β content for each SCOP domain tested from the PDB web page (http://www.rcsb.org/pdb/). In Figure 7 we present the α and β percentages for each domain, grouped by the corresponding SCOP class classification, obtained for the PDB40-b dataset.

From Figure 7, it is apparent that some domains have arguable classifications. For example protein with PDB iden-tification 1HYM–Trypsin inhibitor V [species: pumpkin (Cucurbita maxima)], has two chains that correspond to two SCOP domains. Domain 1HYM:A has 24.44% of α-helix and 0% of β-sheet (labelled * symbol close to the x-axis in
two chains, 1HYM:A and 1HYM:B, have contrasting α-helix and β-sheet content.

Fig. 7) and domain 1HYM:B has 0% of α-helix and 33.33% of β-sheet (labelled * symbol close to the y-axis in Fig. 7). Nevertheless the whole protein was classified in the α + β class, in spite of the fact that each of its chains taken individually would be classified in other classes. The SCOP classification is global in the sense that looks to the whole protein rather than to a particular domain; therefore, classifying chains of 1HYM as α + β is formally correct. Interestingly, a multivariate analysis of variance (MANOVA) of the amino acid composition in the four classes leads to similar results (see web page), showing that class α + β is clearly intermixed with the others in terms of α and β content.

CONCLUSION

In this report, we quantitatively compared several protein dissimilarity measures based in L-tuple composition with alignment scores obtained with Smith–Waterman algorithm. A new metric, the W-metric, which combines both approaches by including word-statistics information weighted by scoring matrices is described.

The accuracy of each metric to detected protein relationships was assessed through the four hierarchical levels of the SCOP/ASTRAL database. The comparative protocol employed the AUCs, which are a good measure of overall accuracy of a classification scheme.

The SW alignment score was shown to be the most discriminant at family and superfamily levels. At family level, the Wm is clearly more discriminant than the other L-tuple distances for sensitivity values between 0.5 and 0.8. From superfamily to class levels, all metrics lose discriminant power and converge to similar AUC values, which makes it counterproductive to use computational intensive alignment algorithms to detect those relationships. At fold level standard Euclidean distance outperforms most of the metrics, achieving an unexpected accuracy for high sensitivity/low sensibility range. This important result anticipates its use in providing a conservative pre-screening procedure for this problem category. In fact, since L-tuple methods are computationally much lighter, they can be useful to pre-select similar proteins before applying the alignment algorithms, thus combining the powerful aspects of each technique and greatly improving heuristic methods in sequence similarity searches.

The graph showing α-helix and β-sheet content for each domain shows that class classification cannot be inferred directly from that information, at least for mixed classes. Therefore, it might be advantageous in some applications to reconsider protein class classification of each domain by exploring the distribution of sequence distances by unsupervised learning algorithms.

ACKNOWLEDGEMENTS

The authors thank John Schwacke, of the Medical University of South Carolina, for providing streamlined MATLAB code for Smith–Waterman alignment and Steven Brenner of the University of California at Berkley for precious advice in the use of the PDB40-B set. The authors thankfully acknowledge the financial support by grants SFRH/BD/3134/2000 to S.V. and SAPIENS/34794/99 from Fundação para a Ciência e a Tecnologia (FCT) of the Portuguese Ministério da Ciência e do Ensino Superior. R.G.-O. thankfully acknowledges grant QLK2-CT-2000-01020 (EURIS) from the European Commission. This work was also supported in part by the NHLBI Proteomics Initiative through contract N01-HV-28181, and a Cancer Center grant from the Department of Energy (C.E. Reed, PI).

REFERENCES


Comparative evaluation of word composition distances


