

# A New Scoring Function and Associated Statistical Significance for Structure Alignment by CE

YUTING JIA,<sup>1</sup> T. GREGORY DEWEY,<sup>1</sup> ILYA N. SHINDYALOV,<sup>2</sup> and  
PHILIP E. BOURNE<sup>2,3</sup>

## ABSTRACT

**A new scoring function for assessing the statistical significance of protein structure alignment has been developed. The new scores were tested empirically using the combinatorial extension (CE) algorithm. Specifically, the significance of a given score was given a  $p$ -value by curve-fitting the distribution of the scores generated by a random comparison of proteins taken from the PDB\_SELECT database and the structural classification of proteins (SCOP) database. Although the scoring function was developed based on the CE algorithm, it is portable to any other protein structure alignment algorithm. The new scoring function is examined by sensitivity, specificity, and ROC curves.**

**Key words:** scoring function, protein structure alignment, CE, statistical significance, sensitivity and specificity, ROC curves.

## 1. INTRODUCTION

**P**ROTEIN STRUCTURE ALIGNMENT IS USED to establish structural equivalences between amino acid positions based on the three-dimensional structures of two protein folds. Such alignments are important for the characterization of three-dimensional structures and for defining similarities and differences related to biological function. Structure alignment, like sequence alignment, has three components: the alignment algorithm, the scoring system, and the assessment of the statistical significance of the results. Currently, a number of different structure alignment algorithms exist including the sequential structure alignment program (SSAP; Taylor and Orengo, 1989), geometric hashing (GH; Nussinov and Wolfson, 1991), distance matrix alignment (DALI; Holm and Sander, 1993), vector alignment search tool (VAST, Gibrat *et al.*, 1996), dynamic programming (Yale alignment server) (Gerstein and Levitt, 1996, 1998) and combinatorial extension of the optimum path (CE; Shindyalov and Bourne, 1998). Although these different methods recognize similar folds, they do not provide identical alignments. Further, these different algorithms use different scoring systems. Gerstein and Levitt (1996, 1998) used the following scoring function:

$$S_{str} = M \left( \sum_{i,j} \frac{1}{1 + \left(\frac{d_{ij}}{d_0}\right)^2} - \frac{N_{gap}}{2} \right) \quad (1.1)$$

---

<sup>1</sup>Keck Graduate Institute of Applied Life Sciences, 535 Watson Drive, Claremont, CA 91711.

<sup>2</sup>San Diego Supercomputer Center, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92039.

<sup>3</sup>Department of Pharmacology, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92039.

where  $N_{gap}$  is the total number of gaps,  $d_{ij}$  is the distance between the aligned residues  $i$  and  $j$ ,  $M = 20$ , and  $d_0 = 5 \text{ \AA}$ . Because the dynamic programming step, which maximizes  $S_{str}$ , tries to match as many residues as possible, after an alignment is determined, a refinement step, which eliminates the worst-fitting pairs of aligned residues, and then refitting to get a new root mean square deviation (*rmsd*) are necessary.

**AU1**

DALI used the following elastic similarity score:

$$Score = \sum_{i,j} \phi^E(i, j);$$

$$\phi^E(i, j) = \begin{cases} \left( \theta^E - \frac{|d_{ij}^A - d_{ij}^B|}{d_{ij}^*} \right) w(d_{ij}^*), & i \neq j \\ \theta^E, & i = j \end{cases} \quad (1.2)$$

where  $i$  and  $j$  label pairs of aligned residues,  $d_{ij}^A$  and  $d_{ij}^B$  are  $C^\alpha - C^\alpha$  distances for chains  $A$  and  $B$ ,  $d_{ij}^*$  is the average of  $d_{ij}^A$  and  $d_{ij}^B$ ,  $\theta^E = 0.20$ , and  $w$  is an envelope function:  $w(r) = \exp(-r^2/400)$ . Besides the alignment scores, *rmsd* is the most common parameter reported in the alignment results. Other measures, such as aligned length, sequence identity, and number of gaps, are often reported.

The final step in protein structure alignment is assessing the significance of the alignment score. For sequence alignments, there has been considerable progress in assessing the statistical significance of alignment scores. In the BLAST family of programs, statistical significances are derived from a probability model of an arbitrary ungapped alignment based on a random walk technique (Karlin and Altschul, 1990). Levitt and Gerstein (1998) developed a simple empirical approach for calculating the significance of an alignment score using their dynamic programming algorithm (Gerstein and Levitt, 1996, 1998). They assessed the significance of a given score in terms of a  $p$ -value by curve-fitting the distribution of the scores generated by an all-versus-all comparison of proteins in the SCOP database. Alternatively, the statistical significance of an alignment can be assessed in a manner similar to Karlin and Altschul's classic work on the alignment of random sequences (Karlin and Altschul, 1990) by a "random model" of structure alignment. Recently, Jia and Dewey (2003) studied the statistical significance of structural alignment by developing a "random model" of structure alignment. In fact, they considered the alignment of random, ideal polymer chains defined as polymers whose end-to-end distribution function is Gaussian. The simple Gaussian chain model provided a good starting point for structural alignment statistical significance testing. Let  $\{\mathbf{R}_m\} \equiv (\mathbf{R}_1, \mathbf{R}_2 \cdots \mathbf{R}_M)$  denote the set of positions of polymer units on a chain with length  $M$ , where  $\mathbf{R}_m$  is the position vector of the  $m$ th unit. A Gaussian chain is characterized by a distribution of end-to-end vectors given by

$$\phi(\mathbf{R}_u - \mathbf{R}_v) = \left( \frac{3}{2\pi b^2 |u - v|} \right)^{\frac{3}{2}} \exp\left( -\frac{3(\mathbf{R}_u - \mathbf{R}_v)^2}{2b^2 |u - v|} \right) \quad (1.3)$$

for all possible  $u \neq v$  from 1 to  $M$ , where the parameter  $b$  is the expected bond length defined as

$$b^2 = \langle (R_m - R_{m-1})^2 \rangle. \quad (1.4)$$

Jia and Dewey (2003) found that if they superimposed two Gaussian chains with the same length (ungapped alignment),  $\frac{rmsd}{chain\_length^{1/2}}$  will be an appropriate score that follows an extreme value distribution (EVD) and the parameters of the *probability density function* (pdf) are independent of the chain length. Similarly, if we superimpose two proteins, then  $\frac{rmsd}{aligned\_length^{1/3}}$  will be a good similarity measure that follows an EVD and the pdf is independent of the aligned length. This result suggests that  $\frac{rmsd}{aligned\_length^\alpha}$  for any chain length is a good scoring function for an ungapped structural alignment.

In the current study, we define a scoring system for structure alignment based on *rmsd*, *aligned length*, and *number of gaps*. The scoring function is defined using the CE algorithm. The pdf of the scores generated from a nonredundant structure dataset is estimated, and the statistical significance to each observed score is inferred in the form of a  $p$ -value, which is the probability that a better score would occur by chance. This paper is organized as follows. In Section 2 a new scoring function for the CE program is defined. This scoring function can be easily adjusted for other structural alignment algorithms. In Section 3 the sensitivity and specificity of this new scoring function is evaluated.

The nonredundant dataset of protein structures used is chosen from PDB\_SELECT and ASTRAL SCOP 1.63. PDB\_SELECT is a representative list of PDB chain identifiers developed by Hobohm and Sander (1994). We use the most recent 25% and 90% lists (July 2003), which contain a total of 2,066 and 6,254 protein chains, respectively. We also use ASTRAL SCOP 1.63 protein domain sequence subsets, based on PDB SEQRES records (Brenner *et al.*, 2000; Chandonia *et al.*, 2002), with less than 90% identity to each other. We denote the datasets as PDB\_SELECT 25%, PDB\_SELECT 90%, and SCOP 90%, respectively.

## 2. NEW SCORING FUNCTION FOR CE

Currently the CE program uses a  $z$ -score to measure the similarity of an alignment. The  $z$ -score is calculated based on the hypothesis that the number of gaps and the average distance,  $\bar{D}$ , which is defined as Equation (2.1), follows a normal distribution,

$$\bar{D} = \frac{1}{n^2} \sum_{i,j} D_{ij} \quad (2.1)$$

where

$$D_{ij} = \begin{cases} \frac{1}{m} \left( |d_{p_i^A, p_j^A}^A - d_{p_i^B, p_j^B}^B| + |d_{p_i^A+m-1, p_j^A+m-1}^A - d_{p_i^B+m-1, p_j^B+m-1}^B| \right. \\ \quad \left. + \sum_{k=1}^{m-2} |d_{p_i^A+k, p_j^A+m-1-k}^A - d_{p_i^B+k, p_j^B+m-1-k}^B| \right), & i \neq j \\ \frac{1}{m^2} \left( \sum_{k,l=1}^{m-1} |d_{p_i^A+k, p_i^A+l}^A - d_{p_i^B+k, p_i^B+l}^B| \right); & i = j \end{cases}$$

here,  $d_{ij}^A$  (respectively,  $d_{ij}^B$ ) denotes the distance between residues  $i$  and  $j$  in protein  $A$  (respectively,  $B$ ) based on the coordinates of  $C_\alpha$  atoms,  $n$  is the number of *aligned fragment pairs* (AFP),  $m$  is the fragment size ( $m = 8$  in CE) (see Shindyalov and Bourne [1998] for details). The  $z$ -score of a particular alignment is calculated by numerically solving Equation (2.2):

$$\rho(0, 1, -z) = \rho(D_i^{av}, D_i^{sd}, D^{obs}) \rho(G_i^{av}, G_i^{sd}, G^{obs}) \quad (2.2)$$

where  $\rho(\mu, \sigma, x)$  is a truncated normal distribution with mean  $\mu$  and standard deviation  $\sigma$ ,  $D^{obs}$  and  $G^{obs}$  are the observed distance score, and the number of gaps,  $D_i^{av}$  and  $D_i^{sd}$  (respectively,  $G_i^{av}$ ,  $G_i^{sd}$ ) are the sample mean and standard deviation for the distance score (respectively, number of gaps) for paths with length  $i$  in a random comparison. Although empirically the performance of the  $z$ -score is very good, the number of gaps and the average distance do not follow a normal distribution. Furthermore, it is hard to assign a  $p$ -value to an observed  $z$ -score.

For a structure alignment algorithm, we always try to get an alignment, which has a minimal *rmsd*, minimal number of gaps, and maximal aligned length that is defined as the number of matched residues in the alignment, with the understanding that higher sequence identity implies greater structural similarity. Thus, the most important variables needed to measure the similarity are *rmsd*, *aligned length*, and *number of gaps*. However, the percent sequence identity is not a critical measurement as seen by the *correlation coefficient* between  $z$ -score and percent sequence identity. We randomly choose 10,000 pairs of proteins from the PDB\_SELECT 25% and 90% lists, respectively, such that

$$\text{Corrcoef}(z\text{-score}, \text{sequence\_identity}) = \begin{cases} 0.19, & \text{PDB\_SELECT 90\%} \\ 0.07, & \text{PDB\_SELECT 25\%} \end{cases} \quad (2.3)$$

Recall that the correlation coefficient,  $\text{Corrcoef}(X, Y)$ , between two variables,  $X$  and  $Y$ , is defined by

$$\text{Corrcoef}(X, Y) = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{\mu_{XY} - \mu_X \mu_Y}{\sigma_X \sigma_Y} \quad (2.4)$$

where  $\sigma_Z$  and  $\mu_Z$  are the standard deviation and the mean of the random variable,  $Z$ . The correlation coefficient has the following properties: (i) for two independent variables  $X$  and  $Y$ ,  $Corrcoef(X, Y) = 0$ , (ii) if the variables are correlated in some way, then their correlation coefficient will be nonzero. In fact, if  $Corrcoef(X, Y) > 0$ , then  $Y$  tends to increase as  $X$  increases. For weakly correlated variables, their correlation coefficient will be closer to zero; on the other hand, for strongly correlated variables their correlation coefficient will be closer to  $\pm 1$ . Thus, Equation (2.4) shows that the  $z$ -score and the sequence identity have only very weak correlation. Hence, percent sequence identity is not considered in our new scoring system.

It is clear that optimal structure alignment requires a multiobjective optimization procedure, which is concerned with the minimization of a vector,  $F(x) = (rmsd, num\_gap, -aligned\_length)$ , of objectives that can be the subject of a number of constraints or bounds. We can convert the multiobjective problem of minimizing the vector  $F(x)$  into a scalar problem by constructing a weighted sum of all the objectives:

$$\min f(x) = w_1 * rmsd + w_2 * num\_gap + w_3 * aligned\_length. \quad (2.5)$$

However, Equation (2.5) lacks biological meaning, and it is hard to select the weights  $w_i$ . Hence, we must define a scoring scheme as a function of  $rmsd$ ,  $num\_gap$ , and  $aligned\_length$ .

Jia and Dewey's result (Jia and Dewey, 2003) suggests the following scoring function:

$$ce\_score = \frac{rmsd}{aligned\_length^\alpha} \left( 1 + \frac{num\_gap}{aligned\_length^\beta} \right), \quad \alpha, \beta > 0. \quad (2.6)$$

The smaller the  $ce\_score$  is, the better the alignment. The term  $\frac{rmsd}{aligned\_length^\alpha}$  serves as the measurement for goodness of ungapped alignment. Several  $\alpha$  values were tested to choose the optimum. As shown in Fig. 1(a), when  $\alpha = \frac{1}{2}$  or  $\frac{1}{3}$ , the best scores are always from alignments of small length, which is not what is expected intuitively. Subsequent study showed that  $\alpha = 1$  is a good choice. The reason why  $\alpha = \frac{1}{3}$  or  $\frac{1}{2}$  were not good choices for the new score is that there is no linear relationship between the aligned length and minimal protein length of the protein pair (see Fig. 2(a)). On the other hand, Fig. 2(b) shows that there is a linear relationship between  $\frac{MPL}{mean(aligned\_length)}$  and  $MPL$ ; here,  $MPL$  is the minimal protein length in a random pair. The performance for different values is examined subsequently.

Gerstein and Levitt (1998) also defined a *normalized rmsd* as

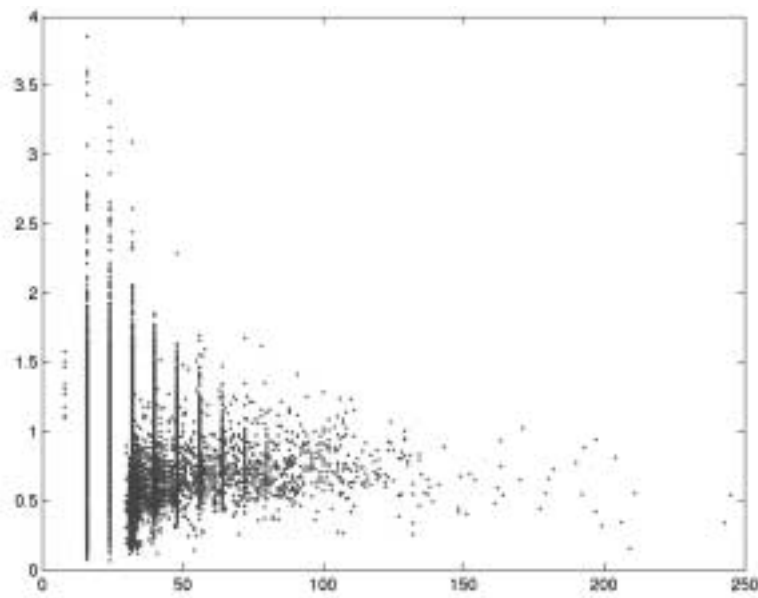
$$normalized\_rmsd = \frac{225 * rmsd}{aligned\_length + 135}, \quad (2.7)$$

which is very similar to the term  $\frac{rmsd}{aligned\_length}$  in our score. The term  $1 + \frac{num\_gap}{aligned\_length^\beta}$  is used to take gaps in the alignment into account. If there are no gaps, the score is  $\frac{rmsd}{aligned\_length}$ ; otherwise, the positive term  $\frac{num\_gap}{aligned\_length^\beta}$  serves as the gap penalty, which results in a larger score;  $\beta = 1$  was chosen in this study. The definition of gap penalty depends on the linear relationship between  $num\_gap$  and  $aligned\_length$ . Indeed, there is a clear linear relationship between them (see Fig. 3). On the other hand, we also tested different values of  $\beta$  as described subsequently, and we find that the performance of the new scoring system exhibits no significant differences for different choices of  $\beta$ . Hence,  $\alpha = 1$ ,  $\beta = 1$  was used for the definition of  $ce\_score$ , the name given to the new scoring function.

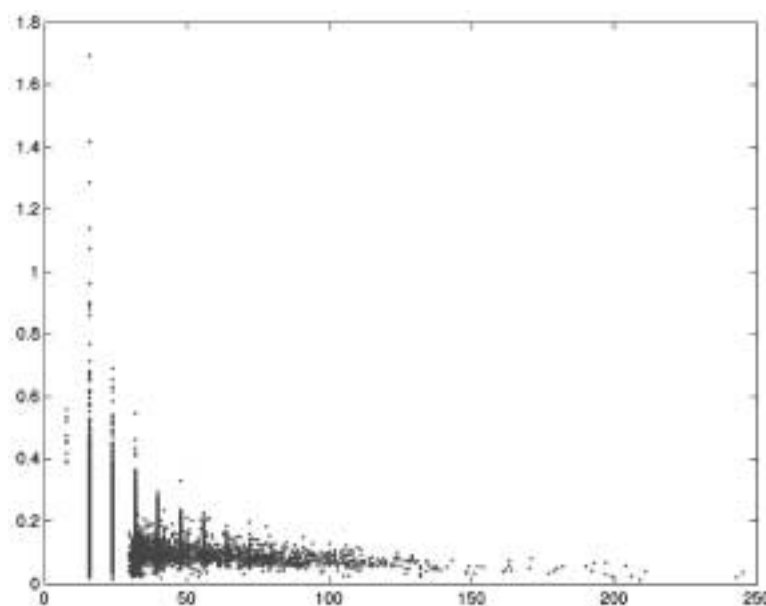
To obtain the statistical significance of an observed score, the probability density function for the random scores needs to be deduced. While it is impossible to give an analytical pdf for the new scores defined here, simulation can be performed on the nonredundant datasets. To obtain a large number of scores from randomly chosen structures, CE was applied to 10,000 pairs of chains randomly chosen from the PDB\_SELECT 25% list, the PDB\_SELECT 90% list, and the SCOP 90% list, respectively. As shown in Fig. 4, the density functions for these scores follow an extreme value distribution (EVD):

$$f(ce\_score) = \lambda \exp(-\lambda(ce\_score - \mu)) \exp(-\exp(-\lambda(ce\_score - \mu))). \quad (2.8)$$

To derive specific values for the  $\lambda$  and  $\mu$  parameters, we fit the above formula to the observed density distribution of scores obtained from PDB\_SELECT 25%, PDB\_SELECT 90%, and SCOP 90%, respectively,



(a)



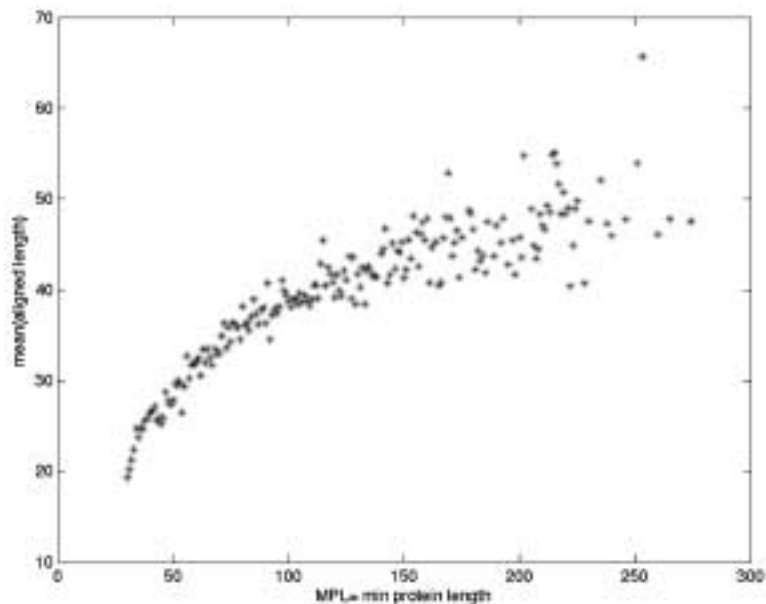
(b)

**FIG. 1.** CE runs on 10,000 random pairs from PDB\_SELECT 25%. (a) The  $ce\_score$  versus aligned length for  $\alpha = 1/2$ ; (b) the  $ce\_score$  versus aligned length for  $\alpha = 1$ .

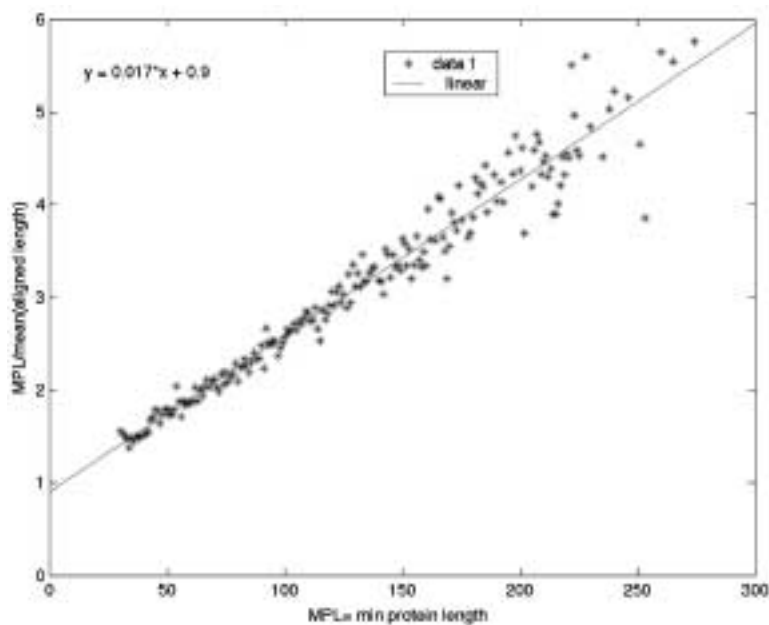
by maximum likelihood fitting (Eddy, 1997). The estimated parameters are as follows:

$$\begin{cases} \lambda = 18.31, & \mu = 0.13; & PDB\_SELECT\ 25\% \\ \lambda = 19.37, & \mu = 0.12; & PDB\_SELECT\ 90\% \\ \lambda = 19.07, & \mu = 0.12; & SCOP\ 90\% \end{cases}$$

The derived parameters for PDB\_SELECT 90% and SCOP 90% are almost the same, while a discrepancy was observed between these and PDB\_SELECT 25% since PDB\_SELECT 90% and SCOP 90% contain more similar chains than does PDB\_SELECT 25%.



(a)



(b)

**FIG. 2.** CE runs on 10,000 random pairs from PDB\_SELECT 25%. (a) Mean (aligned length) versus MPL; (b)  $\frac{MPL}{\text{mean}(\text{aligned\_length})}$  versus MPL; here, MPL is the minimal protein length in a random pair.

The statistical significance of a particular comparison can then be derived by the cumulative distribution function of the EVD, i.e., the  $p$ -value of an observed CE\_SCORE is

$$P(\text{ce\_score} < \text{CE\_SCORE}) = \exp(-\exp(-\lambda(\text{CE\_SCORE} - \mu))). \quad (2.9)$$

A plot of  $p$ -values against scores between randomly chosen pairs from the whole PDB database is shown in Fig. 5. The parameters from the SCOP 90% datasets were used ( $\lambda = 19.07$ ,  $\mu = 0.12$ ). We can easily

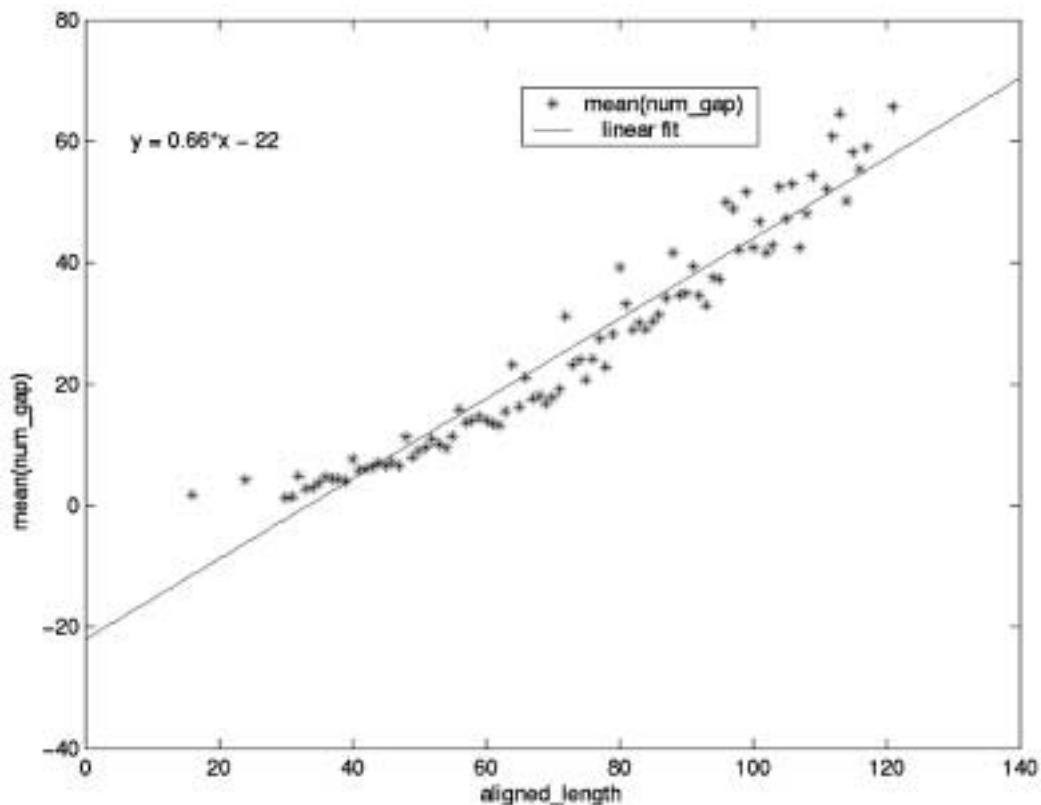


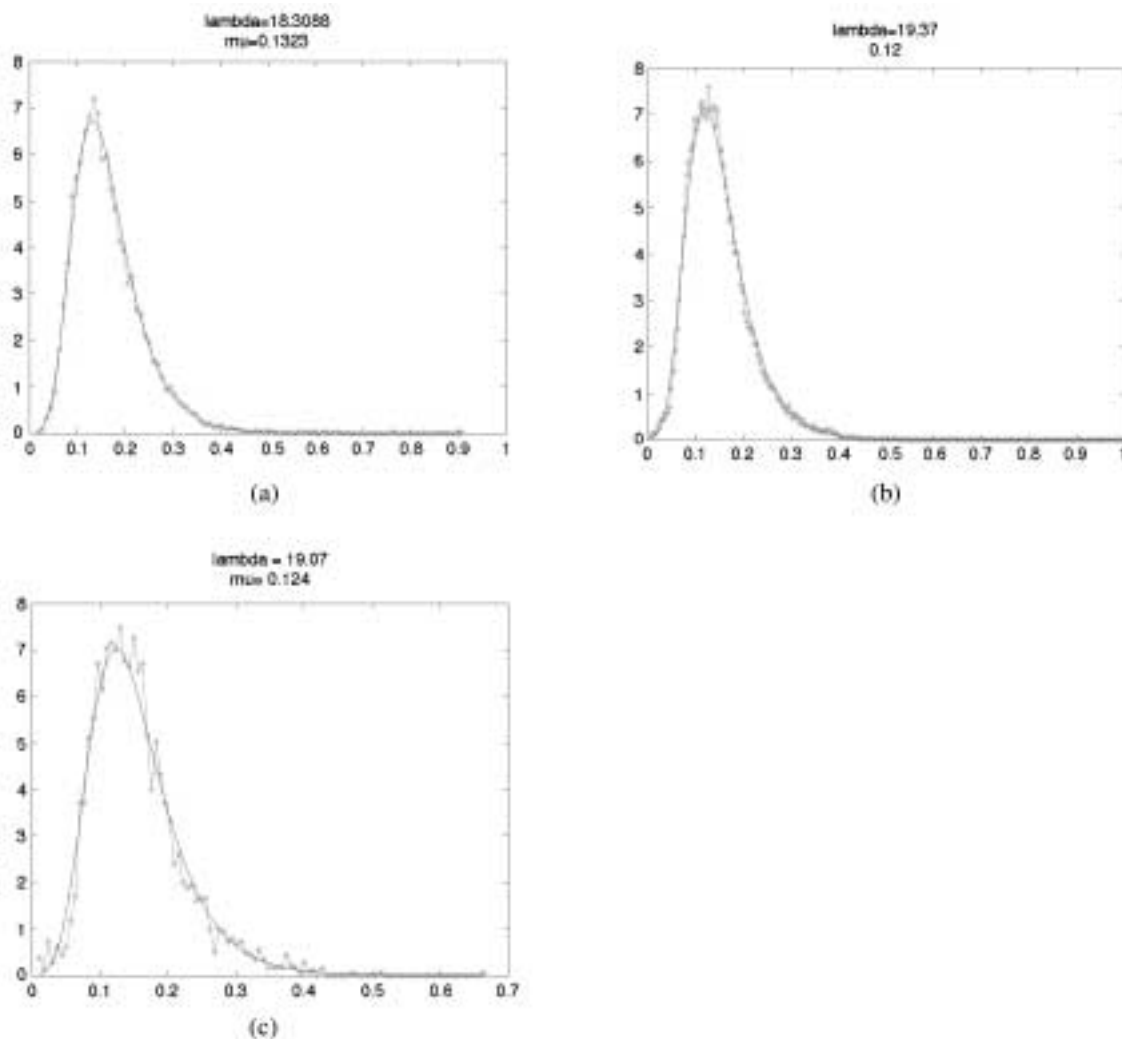
FIG. 3. Mean number of gaps versus aligned length.

find that a  $p$ -value = 5% corresponds to a  $ce\_score$  = 0.0665 and a  $p$ -value = 1% corresponds to a  $ce\_score$  = 0.0439.

### 3. VALIDATION OF THE NEW SCORING SYSTEM BY SCOP

Here, a quantitative evaluation of the performance of the new scoring system is described. A challenge in performing such an evaluation is how to select a standard dataset. Because of the inherent difficulty of identifying and aligning the key structural motifs that characterize a family of proteins, it is often unclear which proteins from the database should be classified as true members of a family. The manually constructed SCOP (Structural Classification of Proteins) Database (Murzin *et al.*, 1995) has often been cited as the possible “gold standard.” SCOP has been constructed by visual inspection of all PDB entries, classified in a hierarchy (“class,” “fold,” “superfamily,” and “family”) indicating different levels of structural and evolutionary relationship between domains. In the current study, we use ASTRAL SCOP 1.63 protein domain sequence subsets, based on PDB SEQRES records, with less than 90% identity to each other. This subset contains 8,042 domains. There are 5,327 protein chains belonging to 630 different folds after we remove the chains that contain different domains belonging to different folds. There are 5,389 protein chains belong to 971 different superfamilies after we remove the chains that contain different domains belonging to different superfamilies. We performed alignment using CE on randomly chosen pairs of chains from “same folds,” “different folds,” “same superfamilies,” and “different superfamilies” and calculated the  $ce\_score$  for each pair. Then we plot the pdfs of the  $ce\_score$  and  $z$ -scores from “same” and “different” folds (see Fig. 6(a) and (b)) and from “same” and “different” superfamilies (see Fig. 6(c) and (d)). The performances of CE and the new scoring system can be evaluated by their *sensitivity* and *specificity*:

$$sensitivity = \frac{TP}{TP + FN}, \quad specificity = \frac{TN}{FP + TN} \quad (3.1)$$



**FIG. 4.** The pdf of the *ce\_score* and the EVD fittings: (a) PDB\_SELECT 25%; (b) PDB\_SELECT 90%; (c) SCOP 90%.

where  $TP$  = true positive,  $FN$  = false negative,  $TN$  = true negative, and  $FP$  = false positive. Based on the values of sensitivity and specificity of each scoring function, a cutoff score can be chosen to distinguish structures from different fold or superfamilies or from the same folds or superfamily. The cutoffs we chose should minimize the value of  $(1 - sensitivity) + (1 - specificity)$  (i.e., the summation of false positive rate and false negative rate). From plots in Fig. 6, the thresholds for the *score* and *z-score* can be chosen as follows:

$$\begin{cases} S_F = 0.08 \\ Z_F = 3.60 \end{cases} \quad \text{and} \quad \begin{cases} S_S = 0.073 \\ Z_S = 3.79 \end{cases} \quad (3.2)$$

i.e., for a test, if the *score*  $> S_F$  (*score*  $< S_F$ ), we will classify the pair as coming from “different” (“same”) folds. If we use *z-score*, we will classify the pair from “different” or “same” folds depending on the *z-score*  $< Z_F$  or not. A similar classification can be done at the level of the superfamily.

To evaluate our new scoring system and the performance of CE, we summarize the test results in Tables 1 and 2 and conclude that 1) the performance of CE is reliable and 2) both the *z-score* and *ce\_score* show good sensitivity and specificity, while the *ce\_score* is better than the *z-score*. To further compare the *z-score* with the *ce\_score*, we also plot the *receiver operating characteristic curve* (ROC curve) (see Fig. 7).

**T1 & T2**

**F7**

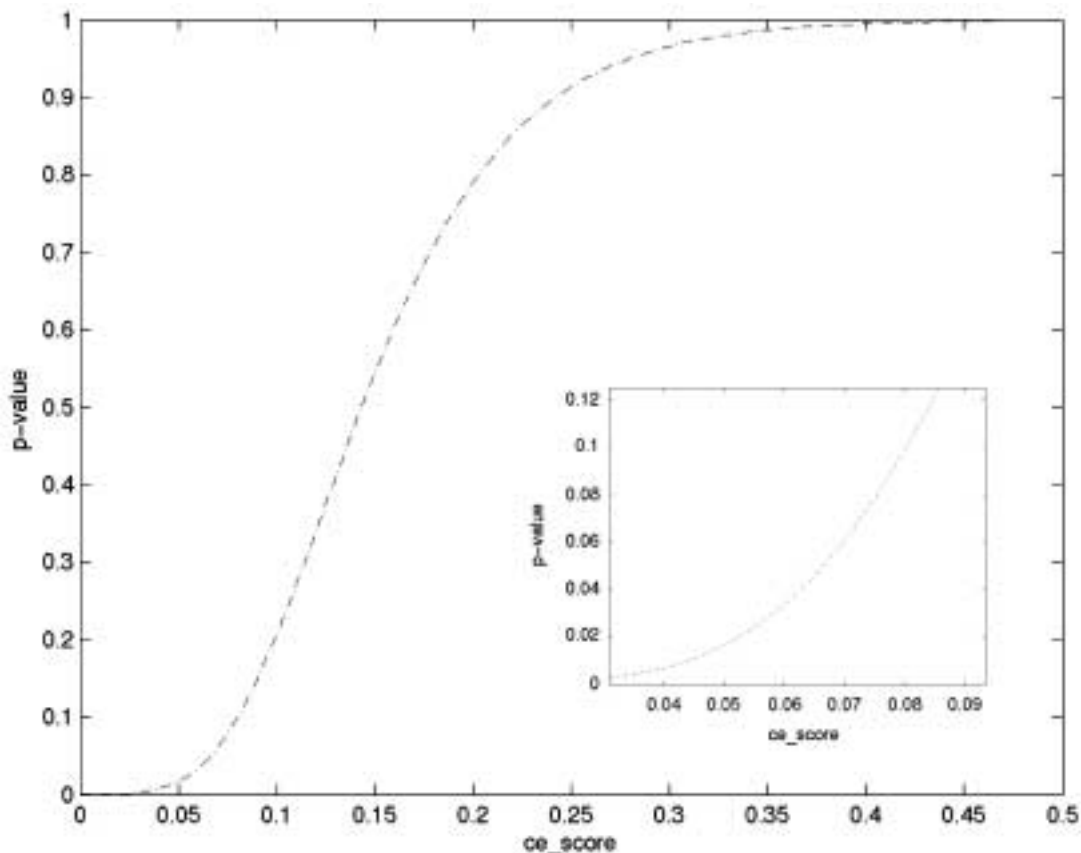


FIG. 5. The  $p$ -value versus  $ce\_score$ .

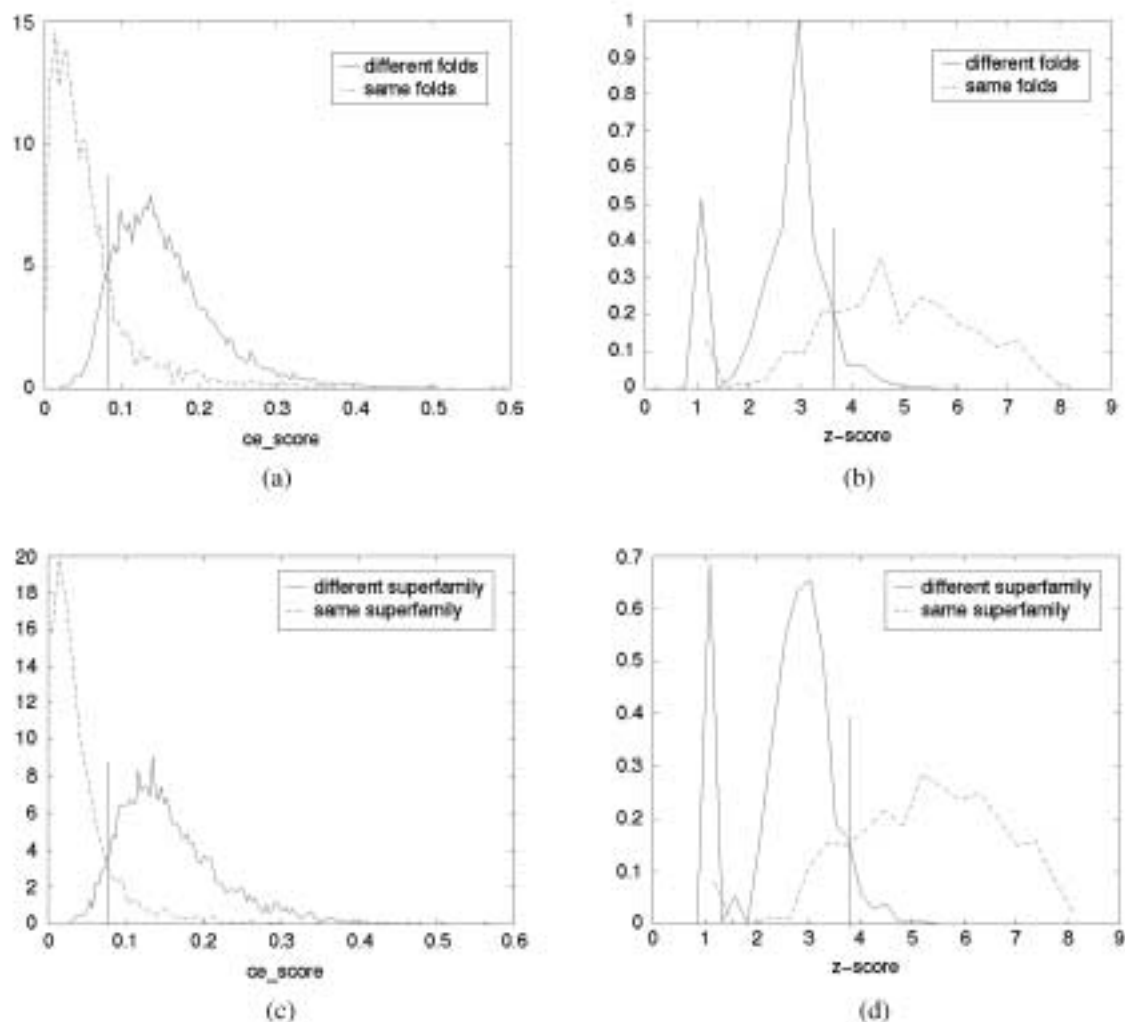
It is a plot of the true positive rate against the false positive rate for the different possible cutpoints of our tests. The results show that the  $ce\_score$  system is of higher sensitivity for almost all possible cutpoints. Hence we believe that our new scoring function is well defined.

We also plot the ROC curves for different values of  $\alpha$  in Fig. 8. We find that for both the fold level and superfamily level,  $\alpha = 1$  and  $\beta = 1$  in the definition of  $ce\_score$  (see Equation (2.6)) is a good choice.

There are a few cases where the two chains were originally classified in the same family by SCOP, but not identified to be even in the same superfamily by our  $ce\_score$ , especially for short chains. For example, 1KBH:B and 1JJS:A were classified in the same family, a.153.1.1. Both chains are short chains (length <60). Alignment results are  $rmsd = 7.045 \text{ \AA}$ ,  $z\_score = 2.3$ ,  $aligned\_length = 40$ ,  $num\_gap = 8$ ,  $ce\_score = 0.21$ , and  $p\_value = 0.83$ . Results for chain length  $\geq 60$  indicated the sensitivity of the  $ce\_score$  is much higher for the long chains than short chains (Table 2 and Fig. 6). Gerstein and Levitt (1998) found that some structures were considered similar in SCOP because they shared a special structural feature rather than an overall structural similarity. An example of this can be found in 1FT5:A (a.138.1.3) aligned to 3CYR (a.138.1.1),  $rmsd = 4.986 \text{ \AA}$ ,  $z\_score = 2.83$ ,  $aligned\_length = 48$ ,  $num\_gap = 17$ ,  $ce\_score = 0.14$ , and  $p\_value = 0.48$ . Interestingly, 1GD1:O aligned to 1DPG:A, one of Gerstein and Levitt's examples, does show statistically significant structure similarity based on our new scoring function ( $p\_value = 0.009$ ).

Why do false positive occur? Consider the following examples to address this question: (i) 1M1J:B (d.171) aligned to 1FBV:A (g.44),  $rmsd = 2.573 \text{ \AA}$ ,  $z\_score = 4.42$ ,  $aligned\_length = 32$ ,  $num\_gap = 0$ ,  $ce\_score = 0.0804$ , and  $p\_value = 0.100$ ; (ii) 1EO8:H (b.1) aligned to 1JHJ:A (b.18),  $rmsd = 6.219 \text{ \AA}$ ,  $z\_score = 4.25$ ,  $aligned\_length = 99$ ,  $num\_gap = 28$ ,  $ce\_score = 0.0806$ , and  $p\_value = 0.1014$ ; (iii) 1F1O:A (a.127) aligned to 1VHB:A (a.1),  $rmsd = 1.788 \text{ \AA}$ ,  $z\_score = 4.42$ ,  $aligned\_length = 32$ ,  $num\_gap = 0$ ,  $ce\_score = 0.0559$ , and  $p\_value = 0.0256$ ; (iv) 1XIM:A (c.1) aligned to 4TMK:A (c.37),  $rmsd = 5.985 \text{ \AA}$ ,  $z\_score = 4.42$ ,  $aligned\_length = 140$ ,  $num\_gap = 53$ ,  $ce\_score = 0.0589$ , and

**F8**



**FIG. 6.** (a) The pdfs of  $ce\_score$  from different and same folds; threshold  $S_0 = 0.080$ . (b) The pdfs of  $z$ -score from different and same folds; threshold  $Z_0 = 3.60$ . (c) The pdfs of  $ce\_score$  from different and same superfamilies; threshold  $S_0 = 0.073$ . (d) The pdfs of  $z$ -score from different and same superfamilies; threshold  $Z_0 = 3.79$ .

TABLE 1. SENSITIVITY AND SPECIFICITY FOR  $z$ -SCORE AND  $ce\_score$

		$z$ -score		$ce\_score$	
		Same folds 5950	Different folds 11657	Same folds 5950	Different folds 11657
Fold level	Test same folds	$TP = 4677$	$FP = 971$	$TP = 4734$	$FP = 945$
$Z_F = 3.60$	Test different folds	$FN = 1277$	$TN = 10686$	$FN = 1216$	$TN = 10712$
$S_F = 0.080$	Sensitivity	78.6%		79.6%	
$p$ -value = 9.9%	Specificity		91.7%		91.9%
		Same superfamilies 1029	Different superfamilies 3334	Same superfamilies 1029	Different superfamilies 3334
Superfamily level	Test same superfamilies	$TP = 869$	$FP = 160$	$TP = 884$	$FP = 157$
$Z_S = 3.79$	Test different superfamilies	$FN = 160$	$TN = 3174$	$FN = 145$	$TN = 3177$
$S_S = 0.073$	Sensitivity	84.5%		85.9%	
$p$ -value = 7.1%	Specificity		95.2%		95.3%

TABLE 2. SENSITIVITY AND SPECIFICITY FOR  $z$ -SCORE AND  $ce\_score$  FOR CHAIN LENGTH  $>59$

		$z$ -score		$ce\_score$	
		Same folds 5296	Different folds 9858	Same folds 5296	Different folds 9858
Sample size	Fold level	$TP = 4497$	$FP = 809$	$TP = 4501$	$FP = 945$
	Test same folds				
	Test different folds	$FN = 799$	$TN = 9049$	$FN = 795$	$TN = 10712$
	$Z_F = 3.60$				
Sample size	Sensitivity	84.9%		85.0%	
	Specificity		91.8%		93.0%
	$S_F = 0.076$				
	$p$ -value = 8.2%				
		Same superfamilies 941	Different superfamilies 2818	Same superfamilies 941	Different superfamilies 2818
Sample size	Superfamily level	$TP = 835$	$FP = 135$	$TP = 834$	$FP = 133$
	Test same superfamilies				
	Test different superfamilies	$FN = 106$	$TN = 2683$	$FN = 107$	$TN = 2685$
	$Z_S = 3.88$				
Sample size	Sensitivity	88.7%		88.6%	
	Specificity		95.2%		95.2%
	$S_S = 0.070$				
	$p$ -value = 6.1%				

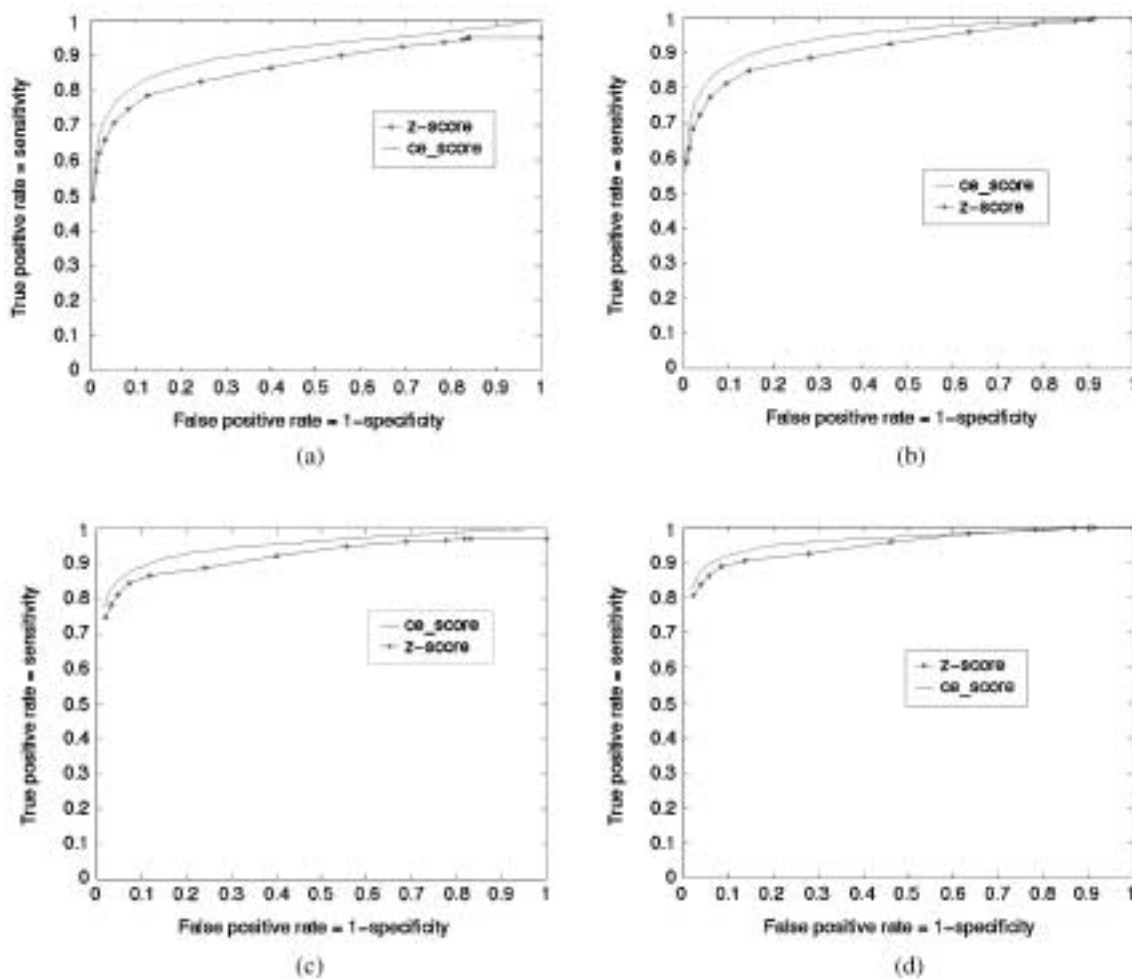
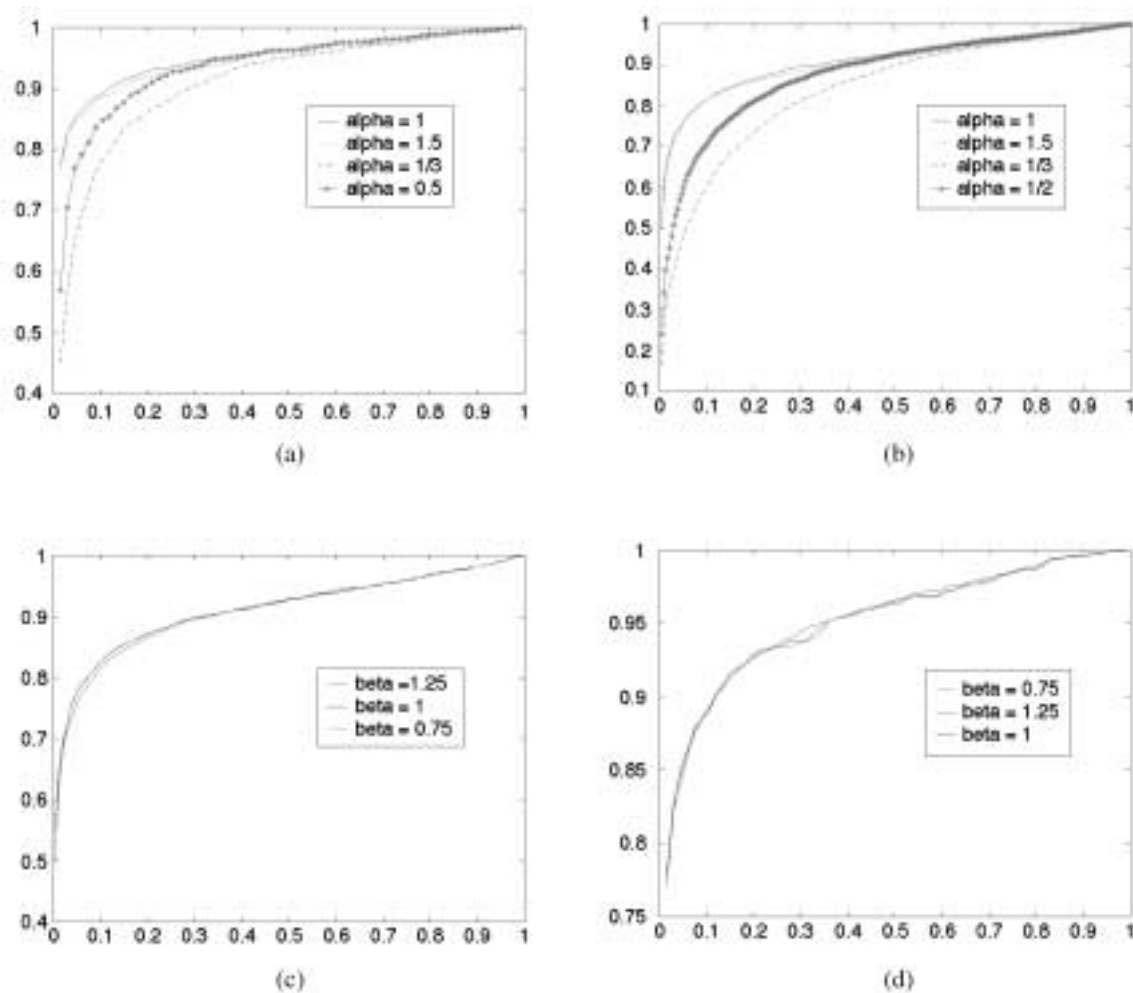


FIG. 7. ROC curves: true positive rate versus false positive rate. (a) The fold level, (b) fold level with aligned length  $>59$ , (c) superfamily level, and (d) superfamily level with aligned length  $>59$ .



**FIG. 8.** ROC curves for different values of  $\alpha$  and  $\beta$ : (a) superfamily level with different  $\alpha$  and  $\beta = 1$ ; (b) fold level with different  $\alpha$  and  $\beta = 1$ ; (c) superfamily level with different  $\beta$  and  $\alpha = 1$ ; and (d) fold level with different  $\beta$  and  $\alpha = 1$ .

$p$ -value = 0.0315. All four examples are false positive if we use a  $z$ -score ( $z$ -score  $> Z_F$  in Tables 1 and 2), but only (iii) and (iv) are false positives if we consider the  $ce$ -score ( $ce$ -score  $< S_F$  in Tables 1 and 2). The aligned length for (iii) is only 32, so the real false positive is example (iv). In (iv) the similarity is valid, but for only part of the respective chain 1XIM:A. The number of aligned residues is only 140 from 393 in 1XIM:A and 213 in 4TMK:A. From eight beta-strands of the TIM-beta/alpha topology of 1XIM:A, only four are matched to another four from five beta-strands from the P-loop topology of 4TMK:A. A further four alpha-helices are matched between the two proteins.

#### 4. SUMMARY

An approach for assessing the statistical significance of a given comparison of protein structures is presented. A new scoring function is defined and tested using the CE structure comparison and alignment algorithm. An extreme value distribution was fitted to the observed distribution of the new scores obtained from aligning structures randomly chosen from public databases. After one estimates the parameters, the statistical significance of the alignment can be easily inferred. The ability of CE to distinguish similar and nonsimilar structures was examined using a  $z$ -score and the new score,  $ce$ -score, indicating that  $ce$ -score was superior. Also,  $ce$ -score can be applied to any other structure alignment algorithms.

## ACKNOWLEDGMENTS

This work was supported by NIH grant GM63208. The authors thank Professor Cheng-Yao Kao for helpful discussions.

## REFERENCES

- Brenner, S.E., Koehl, P., and Levitt, M. 2000. The ASTRAL compendium for sequence and structure analysis. *Nucl. Acids Res.* 28, 254–256.
- Chandonia, J.M., Walker, N.S., Lo Conte, L., Koehl, P., Levitt, M., and Brenner, S.E. 2002. ASTRAL compendium enhancements. *Nucl. Acids Res.* 30, 260–263.
- Dewey, T.G. 1993. Protein structure and polymer collapse. *J. Chem. Phys.* 98(3), 2250–2257.
- Doi, M., and Edwards, S.F. 1986. *The Theory of Polymer Dynamics*, Clarendon Press, Oxford.
- Eddy, S.R. 1997. Maximum likelihood fitting of extreme value distributions. Unpublished, see <ftp://ftp.genetics.wustl.edu/pub/eddy/papers/evd.pdf>.
- Gerstein, M., and Levitt, M. 1996. Using iterative dynamic programming to obtain accurate pairwise and multiple alignments of protein structures. *Proc. 4th Int. Conf. on Intelligent Systems in Molecular Biology*.
- Gerstein, M., and Levitt, M. 1998. Comprehensive assessment of automatic structural alignment against a manual standard, the SCOP classification of proteins. *Protein Sci.* 7(2), 445–456.
- Gibrat, J.F., Madej, T., and Bryant, S.H. 1996. Surprising similarities in structure comparison. *Curr. Opin. Struct. Biol.* 6(3), 377–385.
- Hobohm, U., and Sander, C. 1994. Enlarged representative set of protein structures. *Protein Sci.* 3, 522–524.
- Holm, L., and Sander, C. 1993. Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.* 223, 123–128.
- Jia, Y., and Dewey, T.G., 2003. A random polymer model of the statistical significance of structure alignment. Submitted to *J. Comp. Biol.*
- 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.
- Lessel, U., and Schomburg, D. 1994. Similarities between protein 3-D structures. *Protein Eng.* 7, 1175–1187.
- Levitt, M., and Gerstein, M. 1998. A unified statistical framework for sequence comparison and structure comparison. *Proc. Natl. Acad. Sci. USA* 95, 5913–5920.
- May, A.C.W., and Johnson, M.S. 1994. Protein structure comparison using a combination of a genetic algorithm, dynamic programming and least-squares minimization. *Protein Eng.* 7, 475–485.
- Nussinov, R., and Wolfson, H.J. 1991. Efficient detection of three-dimensional structural motifs in biological macromolecules by computer vision techniques. *Proc. Natl. Acad. Sci. USA* 88(23), 10495–10499.
- Remington, S.J., and Matthews, B.W. 1978. A general method to assess similarity of protein structures, with applications to T4 bacteriophage lysozyme. *Proc. Natl. Acad. Sci. USA* 75, 2180–2184.
- Rose, J., and Eisenmenger, F. 1991. A fast unbiased comparison of protein structures by means of the Needleman–Wunsch algorithm. *J. Mol. Evol.* 32, 340–354.
- Shindyalov, I.N., and Bourne, P.E. 1998. Protein structure alignment by incremental combinatorial extension (CE) at the optimal path. *Protein Eng.* 11, 739–747.
- Taylor, W.R., and Orengo, C.A. 1989. Protein structure alignment. *J. Mol. Biol.* 208, 1–12.

Address correspondence to:

Philip E. Bourne  
Department of Pharmacology  
University of California at San Diego  
9500 Gilman Drive  
La Jolla, CA 92039

E-mail: bourne@sdsc.edu

**Author/Pub**

**All art originals supplied for this article were low-resolution. Unable to differentiate colors and hold lines and shades for Figs. 7 and 8 due to low resolution originals. Okay as is or supply new art?**

**AU1**

**Changes okay: “residues, and then . . . are necessary”?**