A Unified Statistical Framework for Sequence Comparison and Structure Comparison

Michael Levitt[1] and Mark Gerstein[2]

[1]Department of Structural Biology, Stanford University, Stanford, CA 94305, USA

[2]Molecular Biophysics & Biochemistry Department, PO Box 208114, Yale University, New Haven, CT 06520-8114 USA

Classification: Biophysics

Keywords: Sequence Analysis, Structure Analysis, Fold Family, Databank Statistics, Protein Evolution, P-value

Proceedings of the National Academy of Sciences USA (in press)

Page 1 of 33 pages of text and 8 pages of figures

ABSTRACT We present an approach for assessing the significance of sequence and structure comparisons using the same statistical formalism for both sequence and structure. This involves doing an all-vs-all comparison of domains from the scop database, and then fitting a simple distribution function to the observed scores. Using the distribution, we attach a statistical significance to each comparison score in the form of a P-value, the probability that a better score would occur by chance. We find that the scores for sequence matching follow an extreme-value distribution as is expected The agreement between the statistics used by standard programs (e.g., BLAST and FASTA) and our differently derived P-values validates our approach. We find structure comparison scores also follow an extreme-value distribution, when the statistics are expressed in terms of a structural alignment score (essentially the sum of reciprocated distances between aligned atoms less gap penalties). The traditional metric of structural similarity, the RMS deviation in atom positions after fitting aligned atoms, performs less well than structural alignment score. Comparison of the sequence and structure statistics for pairs of proteins known to be distant homologues shows that structural comparison is able to detect about twice as many distant evolutionary relationships as sequence comparison (at the same rate or error). It also indicates that there are very few pairs with significant similarity in terms of sequence but not structure, whereas many pairs have significant similarity in terms of structure but not sequence.

Introduction

Comparison is a most fundamental operation in biology. Measuring the similarities between "things" enables us to group them in families, cluster them in trees, and infer common ancestors and evolutionary progression. Biological comparisons can take place at many levels, from that of whole organisms to that of individual molecules (for an example of systematic comparison applied to organisms see (1) and (2). We are concerned here with the comparison on the latter level, specifically with comparisons of individual protein sequences and structures.

Our overall aim is to describe these two types of comparisons in a self-consistent, unified framework. For sequence or structure comparison, each act of comparing one “entity” to another (i.e. either comparing two sequences or two structures) involves two steps. First, the two objects are optimally aligned through the introduction of gaps in such a way as to maximize their residue-by-residue similarity. This operation generates some form of total similarity score for the number of residues matched -- traditionally, a percent identity for sequences or an “RMS” for structures though we will use other measures here. Second, one has to assess the significance of this score in context of what is known about the proteins currently in the database.

In an earlier paper we tackled the first of these two parts. Gerstein & Levitt (3) extended the work of Subbiah et al. (4) and Laurents et al. (5) and described an approach for structural alignment in an analogous fashion

to the traditional approach for sequence alignment (6 - 9). Like sequence alignment, this method involves applying dynamic programming to a matrix of similarities between individual residues to optimize their overall correspondence through the introduction of gaps. However, the dynamic programming does not necessarily arrive at the best alignment on the first go for protein structures, as it does for sequences. To overcome this, we iterate the procedure until it converges.

In this paper we tackle the second of the two steps in protein comparison, assessing significance. We develop a simple empirical approach for calculating the significance of an alignment score based on doing an all-vs-all comparison of the database and then curve fitting to the distribution of scores of true negatives. This allows us to express the significance of a given alignment score in terms of a P-value, the chance that an alignment between two randomly selected proteins would obtain this score. We apply our approach consistently to both sequences and structures. For sequence we can compare our fit-based P-values with the differently derived statistical score from commonly used programs such as BLAST and FASTA (10-13). The agreement we find validates our approach. For structure alignment, we follow a parallel route to derive an expression for the P-value of a given alignment in terms of the structural alignment score.

Our work follows on much work that has recently been done assessing the significance of sequence and structure comparison. One of

the major developments in the past few years has been the implementation of probabilistic scoring schemes (13, 14, 15, 16). These give the significance of a match in terms of a P-value rather than an absolute, "raw" score (such as percent identity). This places scores from very different programs in a common framework and provides an obvious way to set a significance cutoff (i.e. at P =< 0.0001 or 0.01%). P-values were first used in the BLAST family of sequence searching programs, where they are derived from an analytic model for the chance of an arbitrary ungapped alignment (10, 17). P-values have subsequently been implemented in other programs such as FASTA and gapped BLAST using a somewhat different formalism (13, 18, 19).

There are currently many methods for structural alignment (20 - 31). Some of these have associated with them probabilistic scoring schemes. In particular, one method (VAST) computes a P-value for an alignment based on measuring how many secondary structure elements are aligned, as compared to the chance of aligning this many elements randomly (28). Another method (27, 32) expresses the significance of an alignment in terms of the number of standard deviations it scores above the mean alignment score in an all-vs-all comparison (i.e., a Z-score).

However, in none of the current structure comparison methods is significance derived in the same fashion as it is for the sequence comparison algorithms. Our contribution here is to do just this: to derive

significance P-values in a consistent fashion for comparison of both sequences and structures.

Data set used for Testing

One of the most important aspects of our analysis is that we carefully tested it against the known structural relationships. This allowed us to unambiguously decide whether a given comparison resulted in a true or false positive and to objectively decide between different statistical schemes. In particular, structures were taken from the Protein Data Bank (33 - 34), and domain definitions, definitions of structural class, and known structural similarities were taken from the scop database (version 1.32, May 1996, refs. 35 - 37). The creators of scop have clustered the domains in the PDB on the basis of sequence identity (38, 39). At a sequence identity level of 40%, this procedure results in 941 unique sequences corresponding to the known structural domains. These 941 sequences were what we used as test data for both the sequence and structure comparison. They contain 390 different superfamilies and 281 different folds. Here we concentrate on superfamily pairs, which involve 2107 nontrivial pairwise relationships between the domains, as they have a considerably closer and more certain relationship than fold pairs.

Sequence Comparison Statistics

Sequence matching was done with standard approaches: in particular, we used the FASTA program's (12, 40) version 3.0 SSEARCH implementation of the Smith-Waterman algorithm (Smith & Waterman, 1981), with a gap opening penalty of -12, a gap extension penalty of -2, and the BLOSUM50 substitution matrix (which has a maximal match score of 13 (for C to C) and an expected average match score of -0.36, ref. 40).

A Probability Density Function for Sequence-Comparison Scores

Each pairwise sequence comparison is best quantified by three numbers, S_{seq}, n and m, where S_{seq} is the raw sequence alignment score and n and m are the lengths of the two sequences compared. Comparing all possible pairs of sequences (i.e. 941 x 940) allows us to calculate an observed probability density, ρ^{o}_{seq}, the chance of finding a pair of sequences with particular values for S_{seq} and ln(nm). Figure 1(a) shows the density for pairs between all sequences. This includes the scores for ~300 sequence pairs that are closely related, which clearly show up as "spots" on right side of the plot. These high-scoring, "true positives" are removed in Fig. 1(b), which shows the density for just the sequence pairs in different structural classes (42), i.e. sequences pairs that are definitely unrelated.

Figure 2(a) shows the density distribution as a function of S_{seq} for sections at constant ln(nm). The clear linear relationship between

log(ρ^o_{seq}) and S_{seq} at high values of S_{seq} is indicative of an extreme-value distribution. Thus, we attempt to fit the calculated density using the function

$$\rho^c_{seq}(Z) = \exp(-Z - \exp(-Z))$$

The variable Z is defined in terms of S_{seq} and ln(nm) using a "Z-score-like" expression:

$$Z = (S_{seq} - \mu_{seq}) / \sigma_{seq} \qquad \text{(eq. 2)}$$

where μ_{seq} = a ln(nm) + b and σ_{seq}= a are the most likely sequence score and width parameter for the distribution (using the same parameter *a* for both μ_{seq} and σ_{seq} is done to fit theory as shown below). The two adjustable parameters, a and b, are obtained by fitting the calculated density $\rho^c_{seq}(Z)$ to the observed density $\rho^o_{seq}(Z)$ for all values of S_{seq} and ln(nm). Substituting for μ_{seq} and σ_{seq} in equation (2) gives:

$$Z = (S_{seq} - a \ln(nm) - b)/a = S_{seq}/a - \ln(nm) - b/a$$

To derive specific values for the a and b parameters, we fit the above formulas to the observed density distribution obtained by comparing pairs in different scop classes, getting a = 5.84 and b = -26.3. The fit was done by least-squares optimization using the simplex minimizer in MatLab (44). It has a residual of 0.084, which was calculated using the following standard relation:

$$R = \Sigma\, w_i(O_i - C_i)^2 / \Sigma\, w_i(O_i)^2, \qquad \text{(eq. 3)}$$

where i indexes “bins” with particular S_{seq} and ln(nm) values, O_i = log ($\rho^o_{seq}(Z_i)$) is the observed density, C_i= log ($\rho^c_{seq}(Z_i)$) is the calculated

density, $w_i = 1/N_i$ is a weighting factor, N_i is the number of sequence pairs in a bin, and the summation is over all bins i with ln(nm) between 5.9 and 13.5.

We have also use a four-parameter model: $\mu_{seq} = a \ln(nm) + b$ and $\sigma_{seq} = c \ln(nm) + d$, where a, b, c and d are the four adjustable parameters. The least-squares fit gives a = 6.40, b = -31.9, c = 0.00272 and d = 5.67 with a residual of 0.073, slightly lower than for two parameters. In this study we use the two parameter model as it works almost as well and is closer to the theoretical distribution for ungapped alignment (10).

A Cumulative Sequence Distribution Function, giving the P-value

To estimate the statistical significance of a particular comparison in terms of particular S_{seq}, n, and m values, we need the cumulative distribution function, $P_{seq}(z > Z)$, which is defined as the probability that matching two random sequences will give a z value greater than, or equal to, Z. This is just the integral of $\rho^c_{seq}(z)$ from z = Z to z = infinity.

$$P_{seq}(z > Z) = \int_Z^\infty \exp(-z - \exp(-z))\, dz = \int_Z^\infty \exp(-z)\exp(-\exp(-z))\, dz$$

$$= 1 - \exp(-\exp(-Z)) \qquad \text{(eq. 4)}$$

Note that the cumulative distribution function is just the probability that a value of z greater than the given Z occurs by chance. Writing Z in terms of S_{seq}, n and m gives

$$P_{seq}(s > S_{seq}) = 1 - \exp(-\exp(-S_{seq}/a + \ln(nm) + b/a)) \qquad \text{(eq. 5)}$$

where the parameters a and b are given above.

For large scores and, consequently, Z-values this expression can be considerably simplified. Specifically, for $Z >> 0$, $\exp(-Z)$ is small (i.e. equal to a small ε) so that P_{seq} approximates $1 - \exp(-\varepsilon) = 1 - (1 - \varepsilon) = \varepsilon$. This means that for large Z:

$$P_{seq}(z > Z) = \exp(-Z)$$

$$P_{seq}(s > S_{seq}) = \exp(-S_{seq}/a + \ln(nm) + b/a)$$

$$= \exp(b/a)\, nm \exp(-S_{seq}/a)$$

This makes sense in that the chance of getting a random score greater than S_{seq} depends on the product of the lengths of the sequences and decreases exponentially with increasing S_{seq}.

Relation to BLAST P-value (Karlin & Altschul Parameters)

For sequence comparison without gaps, Karlin & Altschul (10, 11) derived the following cumulative distribution function:

$$P_{K\&A}(s > S_{seq}) = 1 - \exp(-\exp(-\lambda(S_{seq} - \ln(Kmn) / \lambda)))$$

$$= 1 - \exp(-\exp(-\lambda S_{seq} + \ln(Kmn)))$$

where λ and K are calculated analytically based on the sequence composition and amino-acid scoring matrix. Comparison of their analytical form with our P-value expression (equation 4) shows that $\lambda = 1/a$ and $K = \exp(b/a)$. The simple relationship between a and b parameters and Karlin & Altschul's λ and K is one of the reasons we did a two-parameter fit (above). Substituting the specific values for a and b we calculated from the fit, we find that $\lambda = 0.171$ and $K = 0.011$. For the particular database sequences and amino-acid scoring matrix used here (941 scop domains and BLOSUM50), the values for λ analytically calculated by the Karlin & Altschul's formula range from 0.217 to 0.259 with a mean of 0.232. This is significantly larger than our best fit value, of 0.171.

Relation to FASTA E-value

In the FASTA sequence comparison programs (12, 13, 18), the significance of a given alignment score S_{fa}, given by $P_{fa}(s>S_{fa})$, is estimated by fitting an extreme-value distribution to scores resulting from comparison of a given query sequence to each sequence in the database. The distribution is recomputed for each new query so that, unlike our approach, each query sequence is associated with a different distribution function. This has the advantage that it allows any peculiarities of the query sequence to be explicitly taken into account. However, it also means that one can not readily compute the significance for a single pairwise comparison (whereas we express the significance as a simple formula) and

that the results of a given sequence comparison are not symmetrical. By non-symmetrical, we mean that comparison score for matching sequence A to B, where A is the query and B is in the database, is not the same as for matching B to A, where B is the query and A is in the database.

The value commonly used by FASTA and BLAST in judging the significance of a sequence similarity is known as the expectation value or E-value (known here as E_{fa}), which is the number of errors expected when a single query sequence is compared to the entire database. The P-value, defined above, gives the statistical significance of a single comparison, whereas the E-value is an estimate of the number of false positives, or non-similar matches with a score that is judged to be significant, for a search of the entire data base. If there are N_{db} entries in the data base, the E-value is just N_{db} times the P-value (here $N_{db} = 940$). An E-value can be calculated from our $P_{seq}(s > S_{seq})$ using $E_{seq} = N_{db} P_{seq}$. The E-values we obtain (expressed as $\log(E_{seq})$ to allow for the wide range of values), are very similar to those found by FASTA (expressed as $\log(E_{fa})$) over a very wide range of values (Fig. 3). When one considers that our closed-form E_{seq} depends on only two parameters for all pairs, whereas E_{fa} is optimized separately for each query sequence (941 times 2, or 1882 parameters in all), this agreement is astonishing and is likely to be useful for pairwise comparisons that do not involve searches of the entire data base. It also confirms that we have been able to extract the correct underlying distribution by fitting the to observed density of true negative pairs. Thus,

this result validates our approach to some degree and helps us to approach the somewhat more complicated structure comparison situation.

Measuring Coverage vs. Error Rate to Compare Different Formalisms for Significance Statistics

We have presented two forms of E-value statistics for sequence comparison: our method, E_{seq}, which is based on fitting a two parameter model to the observed distribution of alignment scores, and the FASTA method, E_{fa}, which is based on fitting different distributions for each query. Now we are naturally led to ask if there is an objective way to decide which formalism performs the best on some representative test data.

The seminal work of Brenner et al. (39) and Brenner (43) provides a framework for such an assessment, using the known true-positives in the scop dataset and a coverage-vs-error (CVE) plot. To compare any two significance statistics formalisms, we proceed as follows:

(1) For each of the pairs in the all-vs-all comparison (941 x 940 pairs), we determine an E-value, based on the two approaches we are comparing, and a notation of whether or not the pair is a true positive or true negative (for true positives, both sequence must be in the same superfamily in the scop classification). (2) For each E-value measure, we sort the pairs by increasing E-value. (3) We count down the list from best to worst until the number of false positives is 1% of the total number of database entries (here this would be 9, which is about 1% of 941). (4) We look at the threshold E-value at this point. It should ideally be close to 0.01, so as to

correspond to the 1% error rate (per query). (5) We also look at the number of entries which are more significant than the threshold E-value. These define the coverage and it should be as large are possible. (6) After repeating the proceeding five steps with E-value based on the second approach we compare the coverage and also how closely the E-value threshold corresponds to the actual error rate.

Here, we compare the coverage and error rate of our sequence score statistics with those of FASTA (E_{seq} vs. E_{fa}). At the threshold E-value, our sequence statistics have log E_{seq} = -1.98 and a coverage of 328 and the FASTA statistics have a log E_{fa} of -1.68 and a coverage of 379. The FASTA statistics have better coverage but our statistics have an almost perfect threshold value.

Structure Comparison Statistics

Our Basic Pairwise Structural Comparison Procedure

The procedure we use for pairwise structural alignment is described in detail in Gerstein & Levitt (3) and only summarized briefly here. Our core method is based on iterative application of dynamic programming. As such it is a simple application of the Needleman-Wunsch sequence alignment (6). It was originally derived from the ALIGN program of G. Cohen (21, 31) with many subsequent elaborations. One starts with two structures in an arbitrary orientation. Then one computes all pairwise distances between each atom in the first structure and every atom in the

second structure. This results in an inter-protein distance matrix where each entry d_{ij} corresponds to the distance between residue i in the first structure and residue j in the second (inter-residue distances are usually expressed as alpha-carbon distances). This distance matrix can be converted into a similarity matrix S_{ij}, analogous to the one used in sequence alignment, through the relationship

$$S_{ij} = M / (1 + (d_{ij}/d_o)^2),$$

where, somewhat arbitrarily, M = 20 and d_o = 5 Å.

One applies dynamic programming to the similarity matrix to get equivalences (using a gap opening penalty of M/2 = 20 and no gap extension penalty). If this were normal sequence alignment, one would be finished at this point since dynamic programming followed by trace back gives the optimal set of equivalences. However, this is not the case for structural alignment. So one takes these equivalences and uses them to least-squares fit the first structure onto the second one (45). Then one repeats the procedure, finding all pairwise distances and doing dynamic programming to get new equivalences, until convergence. In practice, the iteration is tried from a number of different starting points, and the one that gives the best score is taken.

After determining an alignment, it can be "refined" by eliminating the worst fitting pairs of equivalenced residues and then refitting to get a new RMS, in a similar fashion to the core-finding procedure in Gerstein & Altman (46, 47). This refinement is necessary as the dynamic

programming tries to match as many residues as possible (i.e., it is a global as opposed to local method).

The Structural Comparison Score and the RMS

At the end of the procedure, we are left a number of scores characterizing our final alignment. The most basic is the total number of equivalenced atoms N and the sum of similarity matrix scores S_{ij} for the optimum alignment less the total penalty for opening gaps. We refer to this sum here as S_{str}. Explicitly, it is computed by the following formula:

$$S_{str} = M \left(\left(\sum 1 / (1 + (d_{ij}/d_o)^2) \right) - N_{gap}/2 \right) \qquad \text{(eq. 6)}$$

where N_{gap} is the total number of gaps (not including gaps at the end of a chain) and the summation is carried out over all pairs ij of equivalenced residues. It is important to realize that while the S_{str} is naturally produced by our specific alignment method, it can be calculated from any structural alignment (by substituting the distances between equivalenced alpha-carbons into equation 6). Thus, all the significance statistics that follow could be computed from the results of any structural alignment program, not just our own (and, consequently, provide a uniform basis for comparing the various programs).

However, structural alignments have been traditionally characterized by another quantity, the RMS deviation in alpha-carbon positions after doing a least-squares fit on the positions of equivalenced atoms (the "RMS"). RMS-based statistics were used in our earlier work (e.g. ref. 3-5) and almost all other work in structural-alignment (e.g. the SAS-score in the

original work of ref. 22 is essentially RMS-based). Consequently, we initially felt obligated to try to phrase our structural comparison statistics in this more conventional language.

We describe in detail below two separate statistical treatments of structural comparison, one based on structural alignment score and the other based on the RMS. After doing this, we provide a comparison of the two treatments and show why the one based on structural alignment score is clearly superior.

A Probability Density Function for Structural Alignment Scores

To derive significance statistics for the structural alignment score S_{str}, we proceed exactly as we did for sequence comparison. Structural alignment of all pairs in the database (941 x 940, excluding the protein to itself) gives us an observed probability distribution for comparison scores ρ^c_{str}, which is a function of the number of residues matched N and the comparison score S_{str}. This is shown in Fig. 4. Part (a) shows the data for all pairs. It contains the many pairs of structures that are similar, and these pairs stand out with high values of the structural alignment score S_{str}. Part (b) shows data for pairs that are in different scop structural classes and, therefore, should not show structural similarity. It is much "cleaner" than part (a) and shows the underlying distribution expected for the comparison of structures that are not similar. It is this observed density distribution

function that we need to fit in a similar manner as done for sequences above.

Figure 2(b) shows the density distribution as a function of S_{str} for sections at constant N. There is a close parallel between the structural alignment score S_{str} and the sequence alignment score S_{seq} in Fig. 2(a), and both can be fit modeled by an extreme-value distribution. Thus we fit the calculated structure density by

$$\rho^{c}_{str}(Z) = \exp(-Z - \exp(-Z))$$

where variable, Z, is defined in terms of S_{str} and N using:

$$Z = (S_{str} - \mu_{str}) / \sigma_{str} \qquad \text{(eq. 7)}$$

The most likely structure score μ_{str} and the width parameter σ_{str} have a more complicated dependence on sequence length (N) than was the case for sequences, viz.:

$$\mu_{str}(N) = c \ln(N)^2 + d \ln(N) + e \qquad \text{for } N < 120$$

$$\mu_{str}(N) = a \ln(N) + b \qquad \text{for } N >= 120$$

and

$$\sigma_{str}(N) = f \ln(N) + g \qquad \text{for } N < 120$$

$$\sigma_{str}(N) = f \ln(120) + g \qquad \text{for } N >= 120$$

Continuity of function values and slopes allows a and b to be written in terms of c, d and e. More specifically at N = 120, $a \ln(N) + b = c \ln(N)^2 + d \ln(N) + e$ and $a = 2c \ln(N) + d$. This functional form with the break at N = 120 is the simplest approximation to the observed dependence on N of

the observed density maximum and standard deviation (Fig, 2(b)); it fits well at small N and ensures that the distribution behaves well at high N.

Thus, the expressions for $\mu_{str}(N)$ and $\sigma_{str}(N)$ involve 5 independent parameters c, d, e, f and g. We determined these five parameters via least-squares optimization using the Simplex minimizer in MatLab (44). This yields c = 18.4, d = -4.50, e = 2.64, f = 21.4 and g = -37.5 (a = 419.3 & b = 171.8 are derived as described above). The residual in the fit 0.288. (It is given by the same formula as was used for residual in the sequence statistics fit (equation 5) with $O_i=\rho^o_{str}(Z_i)$, $C_i=\rho^c_{str}(Z_i)$ & $w_i = 1$ (unit weights, w_i, worked better in this case) and the summation is over bins with any value of S_{str} and N between 30 and 170 residues.) The resulting fit of the observed and calculated distribution is good for all values of N and S_{str}, as is apparent in Fig. 2(b).

A Cumulative Structure Distribution Function, giving the P-value

To estimate the statistical significance of a particular structure comparison in terms of its S_{str} and N values, we proceed as we did for sequence comparison. We integrate the score distribution to determine a cumulative distribution function P_{str}, defined as the probability that matching two random structures will give a z value greater than, or equal to, $Z(S_{str},N)$. As the structure score distribution has same extreme-value form as the sequence score distribution, the derivation of P_{str} has the same form as P_{seq}, and we only quote the final result below:

$$P_{str}(z > Z) = 1 - \exp(-\exp(-Z))$$

where Z is expressed in terms of S_{str} and N using

$$Z = (S_{str} - (c \ln(N)^2 + d \ln(N) + e) / (f \ln(N) + g), \quad N < 120$$

$$Z = (S_{str} - (a \ln(N) + b) / (f \ln(120) + g), \quad N >= 120$$

and the seven parameters a, b, c, d, e, f, and g are given above. Finally, we can calculate an E-value for structure comparison E_{str} in the same fashion as sequence comparison: $E_{str}=N_{db} P_{str} = 941 P_{str}$.

Structural Comparison Statistics based on RMS

The traditional characterization of a structural alignment is in terms of the number of residues matched N and the RMS deviation in alpha-carbon positions from least-squares fitting these matched residues, R. Here we derive statistical significance statistics for structural comparison based on these quantities. It is convenient to focus on ln(R) rather than simply R. This ensures that there is good separation of values for small R, which is where the significant pairs occur. (R can never be smaller than zero but ln(R) approaches minus infinity as R tends to zero.) Using the basic quantities N and R we calculate a probability distribution $\rho^o_{rms}(\ln(R),N)$ for the observed RMS values in the true-negative pairs in the same fashion as we did earlier for the observed distribution of structural alignment scores, $\rho^o_{str}(S_{str},N)$. However, this probability distribution has a distinctly different appearance from the earlier one. Figure 5 shows that the profiles of ρ^o_{rms} at constant N are more symmetrical than analogous profiles of ρ^o_{str} shown earlier in figure 2(b). This indicates that an extreme-value distribution would not be an appropriate fitting function.

The fact that log (ρ^{o}_{rms}) varies very slowly near the maximum suggests that we attempt to fit the calculated density using:

$$\rho^{c}_{rms}(Z) = \exp(-Z^4),$$

where the variable Z is a "Z-score-like" expression, defined in terms of ln(R) and N as

$$Z = (\ln(R) - \mu_{rms}(N)) / \sigma_{rms}(N)$$

with

$$\mu_{rms}(N) = c \ln(N)^2 + d \ln(N) + e \qquad \text{for } N < 60$$

$$\mu_{rms}(N) = a \ln(N) + b \qquad \text{for } N >= 60$$

and

$$\sigma_{rms}(N) = f \ln(N) + g \qquad \text{for } N < 60$$

$$\sigma_{rms}(N) = f \ln(60) + g \qquad \text{for } N >= 60$$

The functional form of these expressions for $\mu_{rms}(N)$ and $\sigma_{rms}(N)$ and the choice of the break at N = 60, was determined from the dependence on N of the observed density maximum and standard deviation (Fig. 5). The values of the five independent parameters c, d, e, f and g are determined by least-squares optimization using the Simplex minimizer in MatLab. This yields c = 0.155, d = -0.619, e = 1.73, f = 0.0922 and g = 0.212 (a = 0.872 & b = 0.650, which are determined as before to ensure continuity) with residual of 0.073. (It is given by the same formula as was used for residual in the sequence statistics fit (equation 5) with $O_i=\rho^{o}_{rms}(Z_i)$, $C_i=\rho^{c}_{rms}(Z_i)$, $w_i = 1/N_i$, N_i is the number of pairs in a bin, and the summation over all bins with any value of S_{rms} and N between 30

and 170 residues.) The resulting fit of the observed and calculated distribution is good for all values of N and R, as is apparent in Fig. 5.

To estimate the statistical significance of a particular comparison in terms of its R and N values, we proceed as we did before, deriving a cumulative distribution function $P_{rms}(z > Z)$, defined as the probability that any z will be less than, or equal to, a given Z. This is just the integral of $\rho^c_{rms}(z)$ from z = -infinity to z = Z.

$$P_{rms}(z > Z) = \int_{-\infty}^{Z} \exp(-z^4)\, dz$$

The function, $\exp(-z^4)$ cannot be integrated analytically. Instead we integrate $\exp(-z^4)$ numerically for z from -5 to Z and tabulate its value for 10,000 different Z values from -5 to 5.

It is also convenient to tabulate the limiting values of RMS, R_{lim}, that correspond to particular significance levels (i.e. P_{rms} values) for different values of N. Plots R_{lim} values against N for $P_{rms}(z > Z) = 10^{-7}$, 10^{-6}, 10^{-5} and 10^{-4} show that R_{lim} depends approximately linearly on N and can be written as:

$$R_{lim} = A\ N + B,$$

where the values of A and B depend on log (P_{rms}). For example, to achieve a significance better than 10^{-7}, the RMS must be smaller than 0.0119 N + 1.55, which is 2.74 Å for a 100 residue match. In contrast, for a P-value threshold of 1%, the RMS must be smaller than 0.0172 N + 2.37,

which is ~4 Å for 100 residues. This is approximately the significance cutoff used in Gerstein & Levitt (3).

Comparing Structure Comparison Statistics: Alignment Score S_{str} vs. RMS

Now that we have derived structure comparison statistics based on structural alignment score S_{str} and RMS, we can compare them. The same coverage-vs-error scheme used above to compare the two formulae for sequence significance can be used here. When assessed in terms of coverage (number of true positives found) at a given error rate on our test data, the E-value statistics based on S_{str} give much better performance (i.e. have a larger coverage) than those based on RMS. Specifically, we compare the two approaches (E_{str} vs. E_{rms}) in exactly the same way that we previously compared our sequence E-value to that produced by FASTA (E_{seq} vs. E_{fa}). We find that at the 1% error threshold, the RMS-based statistics have $\log(E_{rms}) = -32.8$ and a coverage of 202, while the structural alignment score (S_{str}) statistics have $\log(E_{str})=-1.58$ and a coverage of 627. Clearly, the statistics based on S_{str} perform much better as the threshold is much more reliable (i.e. closer to the error rate of -2) and the true positive coverage is more than three times higher. The difference between E_{str} and E_{rms} is striking and confirms that the structure score is much better than the RMS score. The coverage (at 1% error per query) obtained with the structural score (627 pairs) is much higher than the coverage obtained with the best sequence score (379 pairs, see above).

There are other reasons why the structural alignment score is a more reliable indicator of structural similarity that the commonly accepted RMS deviation. (1) S_{str} depends most strongly on the best fitting atoms while RMS depends most on the worst fitting atoms. (2) S_{str} penalizes gaps, while RMS does not. (3) S_{str} is formally analogous to the score one gets from a standard sequence comparison, S_{seq}, as both quantities can be derived from a "dynamic-programming" similarity matrix. As such, both S_{str} and S_{seq} directly reflects what the alignment procedure optimizes and are therefore expected to have extreme-value distributions.

Relationship Between Sequence and Structure Comparison

Having derived sequence and structure significance scores using all-vs-all comparisons on the same data base of 941 sequences and structures, we are now in a position to directly compare structure and sequence significance scores. Fig. 6. shows such a comparison for the 2107 pairs of proteins in our data set that are considered to be evolutionarily related according to scop (i.e. they are the true positives in the same superfamily). The lines at $\log(E_{seq}) = -2$ and at $\log(E_{str}) = -2$ divide the 2107 true-positive pairs amongst four quadrants, depending on whether or not their sequence or structure matches are significant, as follows:

1204 pairs in TOP-RIGHT (non-significant sequence match, non-significant structure match). Over half (1204 out of 2107) of the pairs of domains

thought to be evolutionary related by scop fall into this category of having no significant match, indicating that the combination of manual measures used in scop is more sensitive than either automatic sequence or structure comparison.

244 pairs in LOWER-LEFT (significant sequence match, significant structure match). These pairs are evenly distributed in the lower left quadrant, indicating that both sequence and structure significance scores are on the same scale.

576 pairs in LOWER-RIGHT (non-significant sequence match, significant structure match). There are many more pairs with good structure matches but without sequence matches than with the converse (sequence match but no structure match). This objectively shows how much more structure is conserved in evolution than is sequence. These 576 pairs are very good test cases for threading algorithms that match a sequence to a structure, and we are currently testing them in this way.

83 pairs in TOP-LEFT (significant sequence match, non-significant structure match). Almost all the pairs (70 out of 83) in this category involve matches with a small number of residues ($N < 70$). For such short matches, the structures may be deformed and not match well. There are seven labeled pairs that are exceptions as the match is extensive ($N>70$) but the pairs are structurally less similar than would be expected from the strong sequence match. There are 11 coordinate sets involved in these 7 exceptions. Three of these sets were solved by X-ray crystallography to

only medium resolution (>2.9 Å, 1mys, 1scm and 1tlk), five were solved by NMR (1prr, 1ntr, 2pld, 2pna and 1tnm) and three are high-resolution X-ray structures (better than 1.7 Å for 1osa, 3chy and 1sha). None of the seven exceptional pairs involves two high-resolution structures and it seems likely that some of the seven exceptions would have had a more significant structural match if both structures in the pair were determined to high-resolution. Furthermore, as determined from consultation of a Database of Macromolecular Movements (48), some of the seven exceptions involve proteins with extensive conformational changes that have been solved in different conformational states. In particular 1osa, 1mys, and 1scm involve proteins with the highly flexible calmodulin fold. These are clearly examples where one would expect sequence similarity and structural differences.

Discussion and Conclusion

Summary

We have presented an approach for assessing the significance of a given sequence or structure comparison in a unified statistical framework. For either sequence or structure we fit an extreme-value distribution to the observed distribution obtained from the all-vs-all comparison of pairs of domains in different classes in the scop databank. For sequence comparison this result is as expected. Thus, we empirically observe for

gapped alignments what Karlin & Altschul (11) derived for ungapped alignments.

For structure comparison, we find that the score distribution follows an extreme-value distribution when expressed in terms of the structural alignment score S_{str}. Using this measure, expressions for statistical significance can be formulated in an almost identical way for structures as they are for sequences. In contrast, when the score distribution is expressed in the more traditional RMS terms, it is more complicated and the resulting very steep dependence of the probability on the Z-value makes it much less useful for significance statistics.

In connection with this, it is interesting that recent work (39, 43) indicates find that the significance statistics based the optimized "sum" scores from dynamic programming (i.e. Smith-Waterman scores, which are essentially sums of BLOSUM matrix values, less gap penalties) perform much better than those based on the traditional measure of sequence similarity, percentage identity. This parallels the poor performance of our statistics based on the traditional measure of structural similarity, RMS. It is disconcerting that such well-established and intuitive measures like percentage identity or RMS work so much worse than the statistical measures based on the sequence or structure alignment scores.

It is surprising that over half of the relationships between distant homologues in the scop data base are not statistically significant (at the rate of 1% error per query) using either pure sequence comparison or pure

structure comparison. Almost all the pairs found by sequence comparison are also found by structure comparison, but there are many pairs found by structure comparison that are not found by sequence comparison. Overall, structural comparison is able to detect about twice as many of the scop distant homology superfamily pairs as sequence comparison (at the same rate or error)

Future Directions

The approach we have used to derive statistical significance could easily be generalized to other contexts. In particular it can be adapted to provide significance statistics for threading.

We have not presented a detailed examination of the significance values for specific pairs of sequences or structures. This could prove to be a useful endeavor in the future, particularly focusing on pairs of proteins with the same fold but insignificant E-values and those with different folds but significant E-values. These two classes of pairs characterize the twilight zone for structure, which has yet to be fully described.

Acknowledgments

We thank S. E. Brenner for carefully reading the manuscript and S. E. Brenner and T. Hubbard for providing the pdb40d-1.32 dataset. MG acknowledges the NSF for support (Grant DBI-9723182) and ML, the DOE (Grant DE-FG03-95ER62135).

References

1. Rohlf, F. & Slice, D. (1990), *Sys. Zoology* **39**, 40-59.
2. Bookstein, F. L., *Morphometric tools for landmark data* (Cambridge UP, Cambridge, 1991).
3. Gerstein, M. & Levitt, M. (1998), *Protein Science* (in press).
4. Subbiah, S., Laurents, D. V. & Levitt, M. (1993) *Current Biol.* **3**, 141-148.
5. Laurents, D.V., S. Subbiah & Levitt, M. (1994). *Protein Science* **3** 1938-1944.
6. Needleman, S. B. & Wunsch, C. D. (1971), *J. Mol. Biol.* **48**, 443-453.
7. Smith, T. F. & Waterman, M. S. (1981), *J. Mol. Biol.* **147**, 195-197.
8. Doolittle, R. F., *Of Urfs and Orfs* (University Science Books, Mill Valley, CA, 1987).
9. Gribskov, M. & Devereux, J., *Sequence Analysis Primer* (Oxford University Press, New York, 1992).
10. Karlin, S. & Altschul, S. F. (1990), *Proc Natl Acad Sci U S A* **87**, 2264-8.
11. Karlin, S. & Altschul, S. F. (1993), *Proceedings of the National Academy of Sciences of the United States of America* **90**, 5873-7.
12. Lipman, D. J. & Pearson, W. R. (1985), *Science* **227**, 1435-1441.
13. Pearson, W. R. (1996), *Meth. Enz.* **266**, 227-259.
14. Karlin, S., Bucher, P., Brendel, V. & Altschul, S. F. (1991), *Annu Rev Biophys Biophys Chem* **20**, 175-203.
15. Altschul, S. F., Boguski, M. S., Gish, W. & Wootton, J. C. (1994), *Nature Genetics* **6**, 119-29.
16. Bryant, S. H. & Altschul, S. F. (1995), *Curr Opin Struct Biol* **5**, 236-44.
17. Altschul, S. F. & Gish, W. (1996), *Methods in Enzymology* **266**, 460-80.
18. Pearson, W. R. (1997), *Comput Appl Biosci* **13**, 325-32.
19. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997), *Nucleic Acids Res* **25**, 3389-402.
20. Remington, S. J., Matthews, B. W. (1980), *J. Mol. Biol.* **140**, 77-99.
21. Satow, Y., Cohen, G. H., Padlan, E. A., Davies, D. R. (1987), *J. Mol. Biol.* **190**, 593-604.
22. Taylor, W. R., Orengo, C. A. (1989), *J. Mol. Biol.* **208**, 1-22.
23. Artymiuk, P. J., Mitchell, E. M., Rice, D. W., Willett, P. (1989), *J. Inform. Sci.* **15**, 287-298.
24. Sali, A., Blundell, T. L. (1990), *J. Mol. Biol.* **212**, 403-428.
25. Vriend, G., Sander, C. (1991), *Proteins* **11**, 52-8.
26. Russell, R. B., Barton, G. B. (1992), *Proteins* **14**, 309-323.
27. Holm, L., Sander, C. (1993), *J. Mol. Biol.* **233**, 123-128.
28. Gibrat, J. F., Madej, T., Bryant, S. H. (1996), *Curr. Opin. Str. Biol.* **6**, 377-385.
29. Falicov, A., Cohen, F. E. (1996), *Journal Of Molecular Biology* **258**, 871-892.

30. Gerstein, M., Levitt, M., in *Proc. Fourth Int. Conf. on Intell. Sys. Mol. Biol.* (AAAI Press, Menlo Park, CA, 1996), pp. 59-67.
31. Cohen, G. H. (1997), *J. Appl. Cryst.* (in press).
32. Holm, L., Sander, C. (1996), *Science* **273**, 595-602.
33. Bernstein, F. C.*, et al.* (1977), *J. Mol. Biol.* **112**, 535-542.
34. Abola S.J., Prilusky J, Manning N.O. (1997), *Meth. Enz.* **277**, 556-571.
35. Murzin, A., Brenner, S. E., Hubbard, T., Chothia, C. (1995), *J. Mol. Biol.* **247**, 536-540.
36. Brenner, S., Chothia, C., Hubbard, T. J. P., Murzin, A. G. (1996), *Meth. Enz.* **266**, 635-642.
37. Hubbard, T. J. P., Murzin, A. G., Brenner, S. E., Chothia, C. (1997), *Nucleic Acids Res* **25**, 236-9.
38. Brenner, S., Hubbard, T., Murzin, A., Chothia, C. (1995), *Nature* **378**, 140.
39. Brenner, S., Chothia, C., Hubbard, T. (1998), *Proc. Natl. Acad. Sci. USA* (submitted).
40. Pearson, W. R., Lipman, D. J. (1988), *Proc. Natl. Acad. Sci. USA* **85**, 2444-2448.
41. Henikoff, S., Henikoff, J. G. (1993), *Proc. Natl. Acad. Sci.* **19**, 6565-6572.
42. Levitt, M., Chothia, C. (1976), *Nature* **261**, 552-558.
43. Brenner, S. E., (PhD Thesis, Cambridge University, 1996)
44. MatLab Version 5. The Math Works Inc., Natick, MA 01760, USA. http://www.mathworks.com.
45. Kabsch, W. (1976), *Acta Cryst.* **A32**, 922-923.
46. Gerstein, M., Altman, R. (1995a), *CABIOS* **11**, 633-644.
47. Gerstein, M., Altman, R. (1995b), *J. Mol. Biol.* **251**, 161-175.
48. Gerstein, M., Lesk, A. M., Chothia, C. (1994), *Biochemistry* **33**, 6739-6749. See database at http://bioinfo.mbb.yale.edu/MolMovDB.

Figure Legends

Fig. 1. A probability density distribution for sequence comparison scores, ρ^o_{seq}, contoured against S_{seq}, the sequence alignment score (along the horizontal axis) and ln(nm), where n and m are the lengths of the pair sequences (along the vertical axis). This density is closely related to the raw data (via normalization) obtained by counting the number of pairs with the particular S and ln(nm) values. Due to the wide range of density values, contours of $\log(\rho^o_{seq})$ are drawn with an interval of 1 (a full order of magnitude). When contouring the logarithm of a density function, special attention must be paid to the zero values. Here a zero value is set to 0.001, which effectively lifts the entire surface by 3 log units. The data is then smoothed by averaging with a Gaussian function ($\exp(-s/(\Delta S_{seq}/3)^2)$ over a window 14 units wide along the S_{seq} axis. This smoothing together with the treatment of zero observations serves to emphasize the smallest observed counts (values of 1) by surrounding them with three contour levels. Panel (a) shows the data from all 884,540 pairs between any one of the 941 sequences and any other sequence (pairs A-B and B-A are both included). The significant sequence matches are seen as the isolated spots at high values of the score S_{seq}. Panel (b) shows the data from 352,168 pairs including only those pairs of sequences in different scop classes. We also exclude pairs between an all-alpha or all-beta domain and alpha+beta domain as well as sequences that are not in one of the five main scop classes -- alpha, beta, alpha/beta, alpha+beta and alpha+beta (multidomain). This is done to ensure that no significant matches will be found, and this is indeed seen in the figure by the absence of any outlying spots at high score values. Thus, the density in (b) is free of any significant matches and shows the underlying density distribution expected for comparison of unrelated sequences.

Fig. 2. Cross-sections of the sequence and structure density distribution show they are both extreme-value distributions and that the calculated distribution fits the observed distribution well. Part (a) shows plots of the logarithm of the observed, $\log(\rho^o_{seq})$, and calculated sequence pair density, $\log(\rho^c_{seq})$, against the sequence match score, S_{seq}; it is taken from the data for pairs in different classes (Fig. 1b). Each panel shows the variation of the density with S_{seq} for a particular value of ln(nm), the product of the lengths of the sequences compared; this value is indicated by assuming n =

m and showing the value of n. The observed density is clearly an extreme-value distribution with a linear fall-off of $\log(\rho^o_{seq})$ with S_{seq}. The calculated distribution obtained with a two parameter fit (dashed line, see text) is a good fit for all values of n (or ln(nm)). Part (b) shows plots of the logarithm of the observed, $\log(\rho^o_{str})$, and calculated structure pair density, $\log(\rho^c_{str})$, against the structure match score, S_{str}; it is taken from the data for pairs in different classes (Fig. 4b). Each panel shows the variation of the density with S_{str} for a particular value of the number of aligned residues, N. The observed density is clearly an extreme-value distribution with a linear fall-off of $\log(\rho^o_{str})$ with S_{str}. The calculated distribution obtained with a five parameter fit (dashed line, see text) is a good fit for all values of N. In both parts, there is more noise at high values of ln(nm) or N where there is much less data (see Figs. 1b and 4b).

Fig 3. The statistical significance derived here is shown to be similar to that derived in a completely different way by the sequence comparison program SSEARCH from the FASTA package (13). We plot the expected number of errors per search of the data base obtained by Pearson's method, $\log(E_{fa})$, against the same value calculated here, $\log(E_{seq})$ (which is a function of the sequence match score, S_{seq}, and the length of the two sequences). More specifically, E_{fa} is the E-value output by the FASTA-SSEARCH program, whereas E_{seq} is calculated as $940*P_{seq}(s > S_{seq})$ for score S_{seq} using equation 5. The accuracy of our simple two parameter fit is confirmed by the fact that most pairs of $\log(E_{fa})$ and $\log(E_{seq})$ values are perfectly correlated, lying along the line $\log(E_{fa}) = \log(E_{seq})$ over the entire range.

Fig. 4. The logarithm of the density distribution for structure comparison scores, ρ^o_{str}, is contoured against S_{str}, the structural alignment score (along the horizontal axis), and N, the number of aligned residues (along the vertical axis). Following the protocol used for Fig. 1, the raw data obtained by counting the number of pairs with the particular S_{str} and N values is first 'lifted' by setting 0 values to 0.001, it is then smoothed by the same Gaussian averaging used for S_{seq} (Fig. 1) over a window 90 units wide along the S_{str} axis, and finally the log value is contoured in intervals of 1 log unit. Given the different scales used for S_{seq} and S_{str}, the extent of smoothing is very similar for both. Panel (a) shows the data from all

884,540 pairs between any one of the 941 sequences and any other sequence. Panel (b) shows the data from 352,168 pairs including only those pairs of sequences in different scop classes (described in Figure 1). Comparison of (a) and (b) shows that the true positive structural matches are seen in the contours at the higher values of the alignment score, S_{str}. and also at higher values of the number of matches, N. The density in (b) is free of these significant matches and shows the underlying density distribution expected for comparison of unrelated structures.

Fig. 5. The fit to the structure pair density using the RMS score. The observed, $\log(\rho^{o}_{str})$, and calculated structure pair density distributions, $\log(\rho^{c}_{str})$, are plotted against the RMS score, ln(R), for different numbers of aligned residues, N. The observed structure pair density, which is derived from pairs in different classes, is clearly not an extreme-value distribution as it is symmetrical about the maximum value and falls off faster than a linear function with increasing Z. In fact, it is best fit by $\exp(-Z^4)$. The calculated distribution obtained with a five parameter fit (dashed line) is a good fit when the numbers of aligned residues exceeds 50. For smaller values of N, there is a preponderance of pairs. In particular, for short residues matches (30 or 40) there is a clear spike around ln(R) = 1 (RMS = 2.7 Å); it is not clear what this common sub-structure is.

Fig. 6. Comparison of structure significance with sequence significance. Plots of the structure significance, $\log(E_{str})$, against the sequence significance, $\log(E_{seq})$, for the 2107 pairs of proteins judged to be homologous in the scop database (in the same superfamily). Pairs are distinguished by the extent of the structural match, with solid squares used for pairs with N >= 70 and unfilled diamonds used for N < 70. The horizontal and vertical dashed lines, which divide the figure into four quadrants, are at $\log(E_{str})$ =-2 and at $\log(E_{seq})$ = -2, respectively. Both these thresholds correspond to an E-value of 10^{-2} and P-value of 10^{-2} / 941 = 10^{-5} so that we judge matches with lower values to be significant at the 1% level.

Fig. 1a

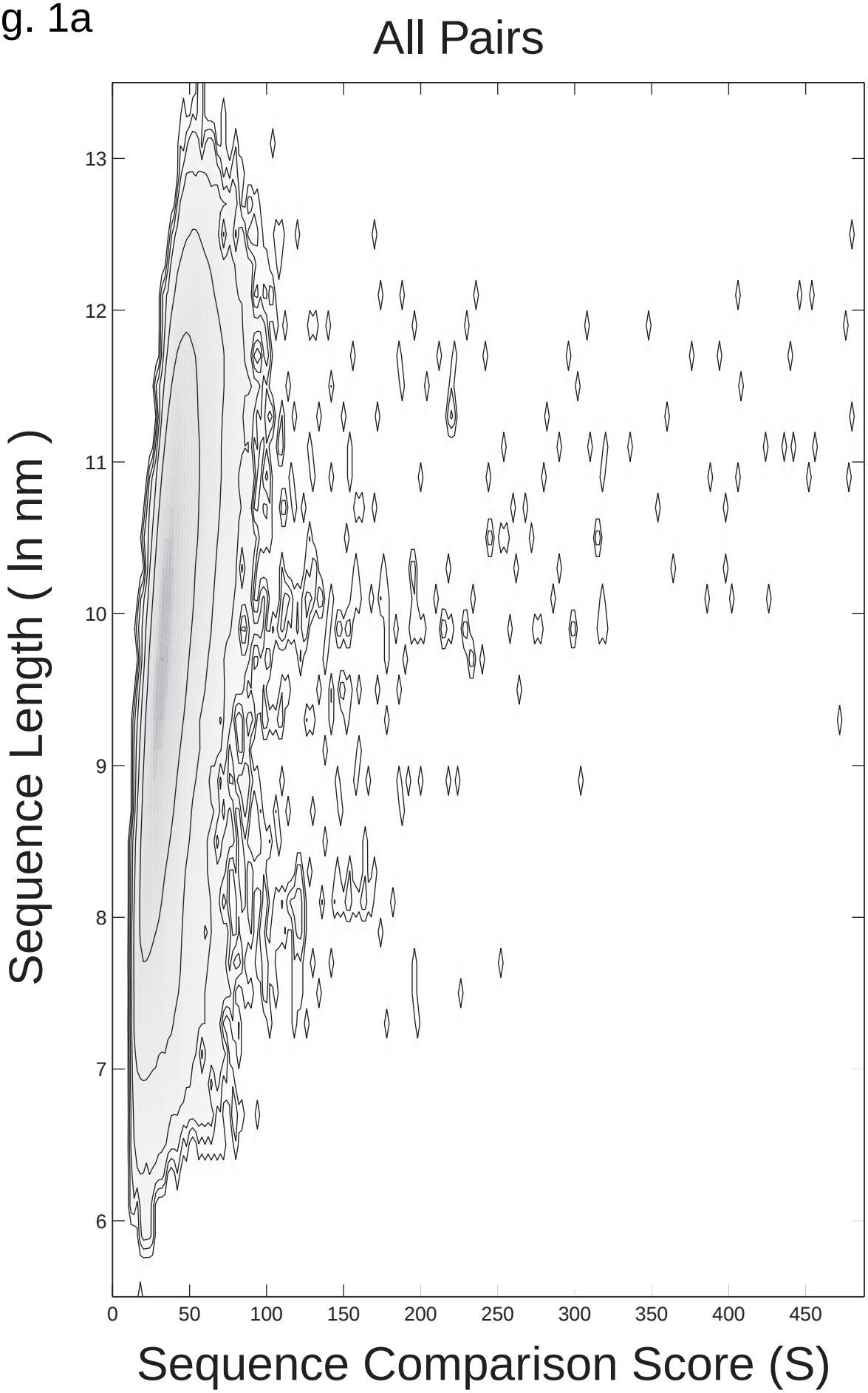

All Pairs
Sequence Length (ln nm)
13
12
11
10
9
8
7
6
0
50
100
150
200
250
300
350
400
450
Sequence Comparison Score (S)

Fig. 1b

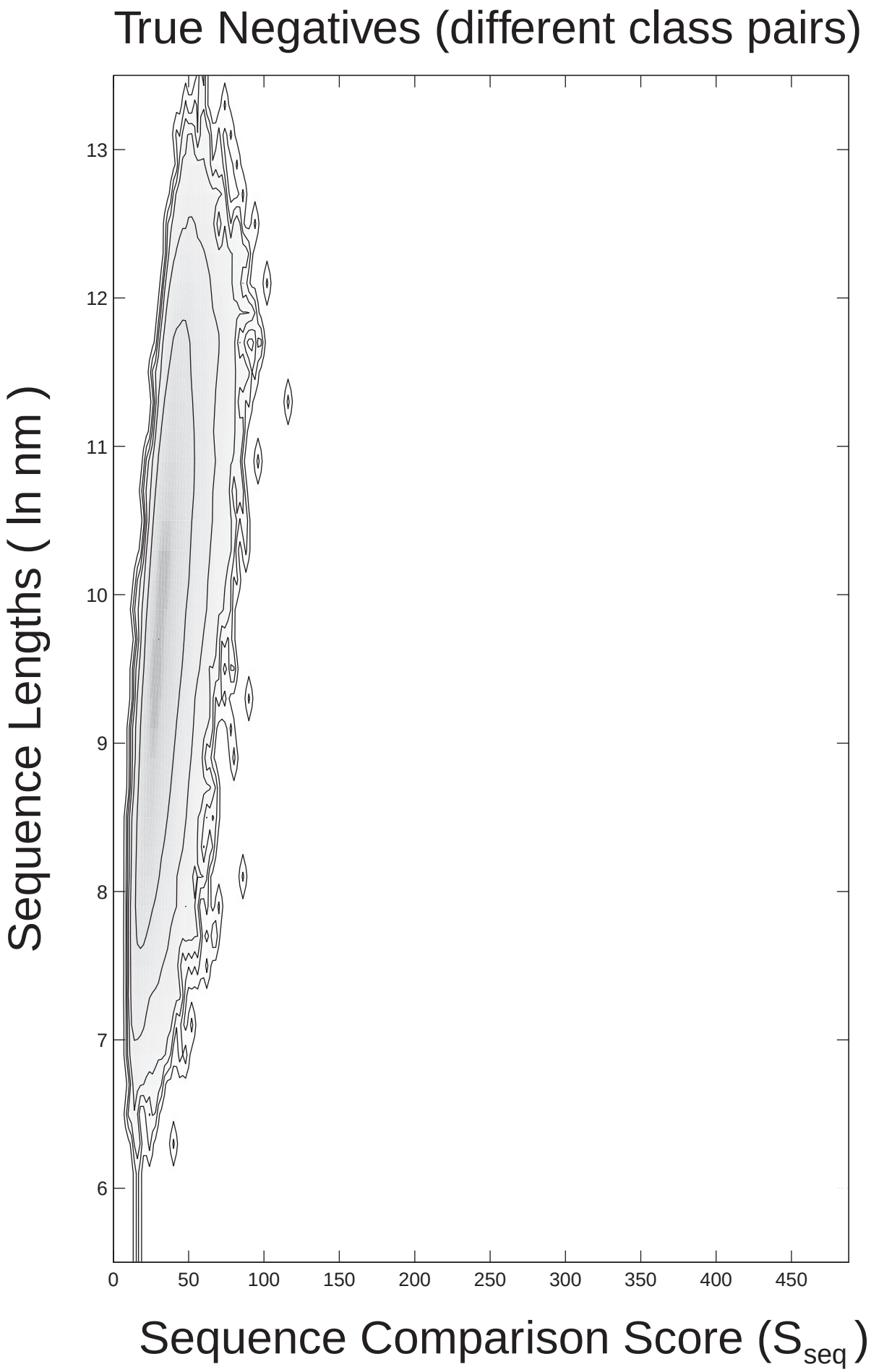
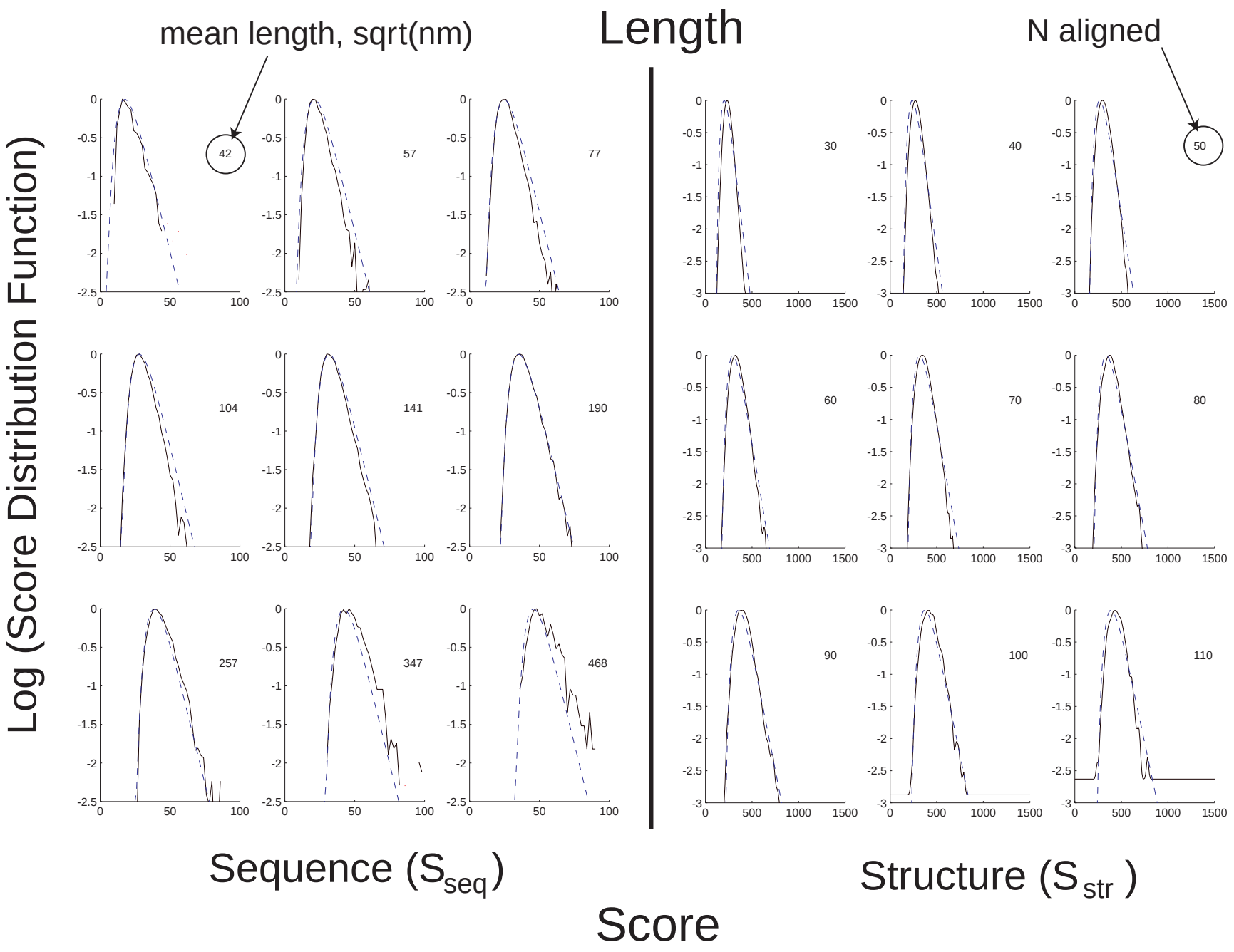

True Negatives (different class pairs)
Sequence Lengths (ln nm)
13
12
11
10
9
8
7
6
0
50
100
150
200
250
300
350
400
450
Sequence Comparison Score (S_{seq})

Fig. 2


Length
mean length, sqrt(nm)
N aligned
Log (Score Distribution Function)
42
57
77
30
40
50
104
141
190
60
70
80
257
347
468
90
100
110
Sequence (S_{seq})
Structure (S_{str})
Score

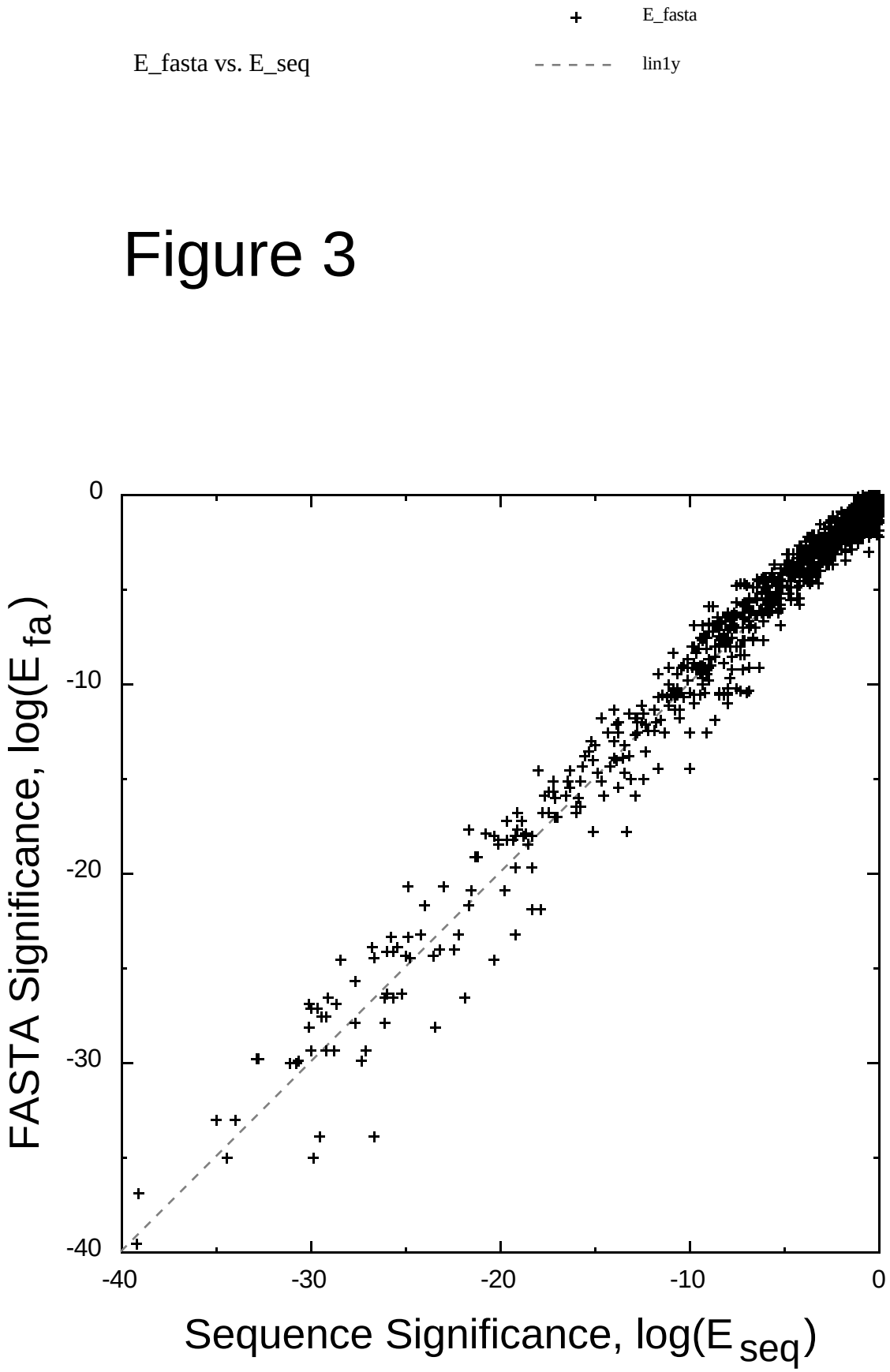

E_fasta
E_fasta vs. E_seq
lin1y
Figure 3
0
-10
-20
-30
-40
-40
-30
-20
-10
0
FASTA Significance, log(E fa)
Sequence Significance, log(E seq)

Fig. 4

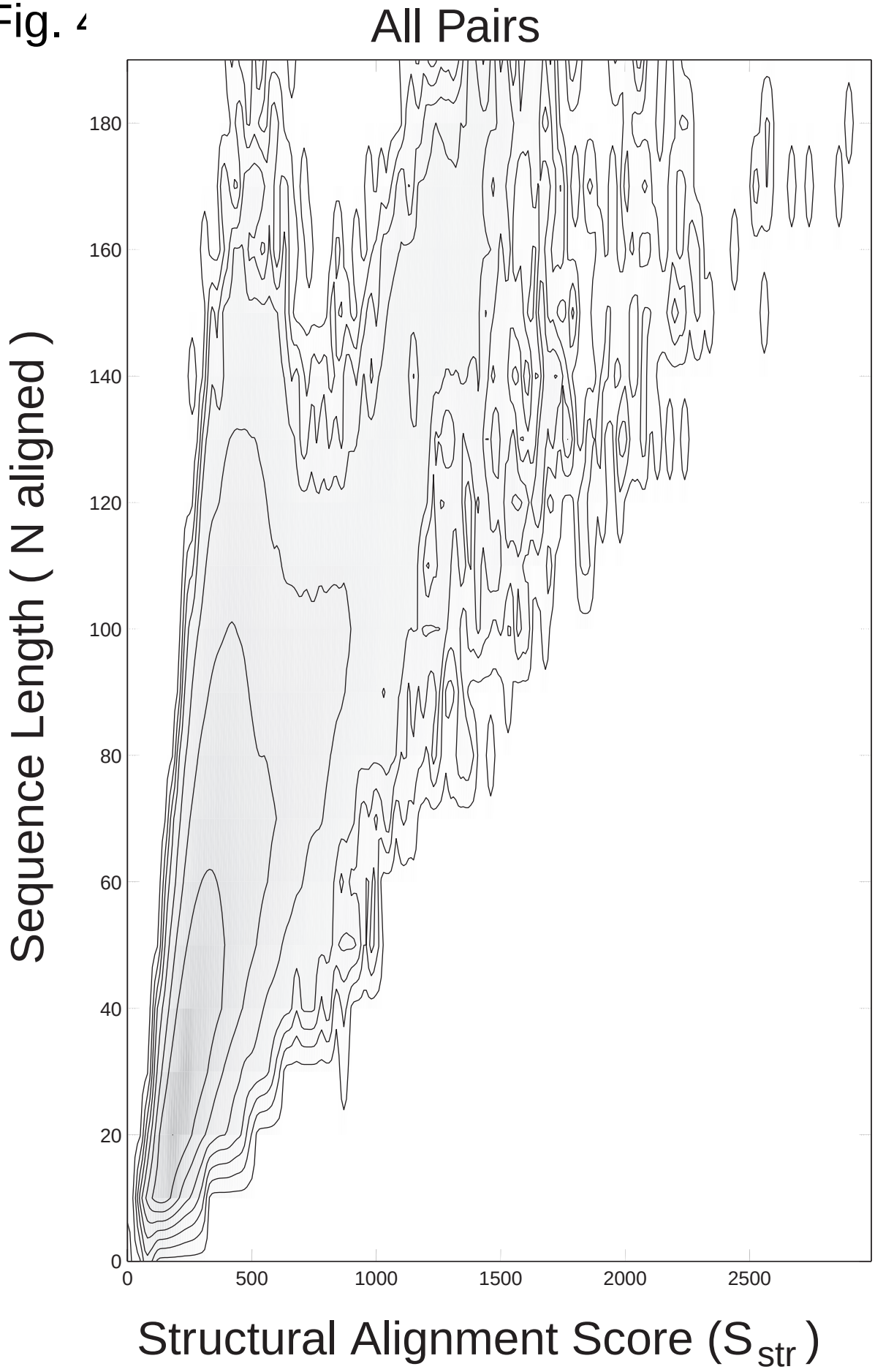

All Pairs
Sequence Length (N aligned)
Structural Alignment Score (S_{str})
0
20
40
60
80
100
120
140
160
180
0
500
1000
1500
2000
2500

Fig. 4

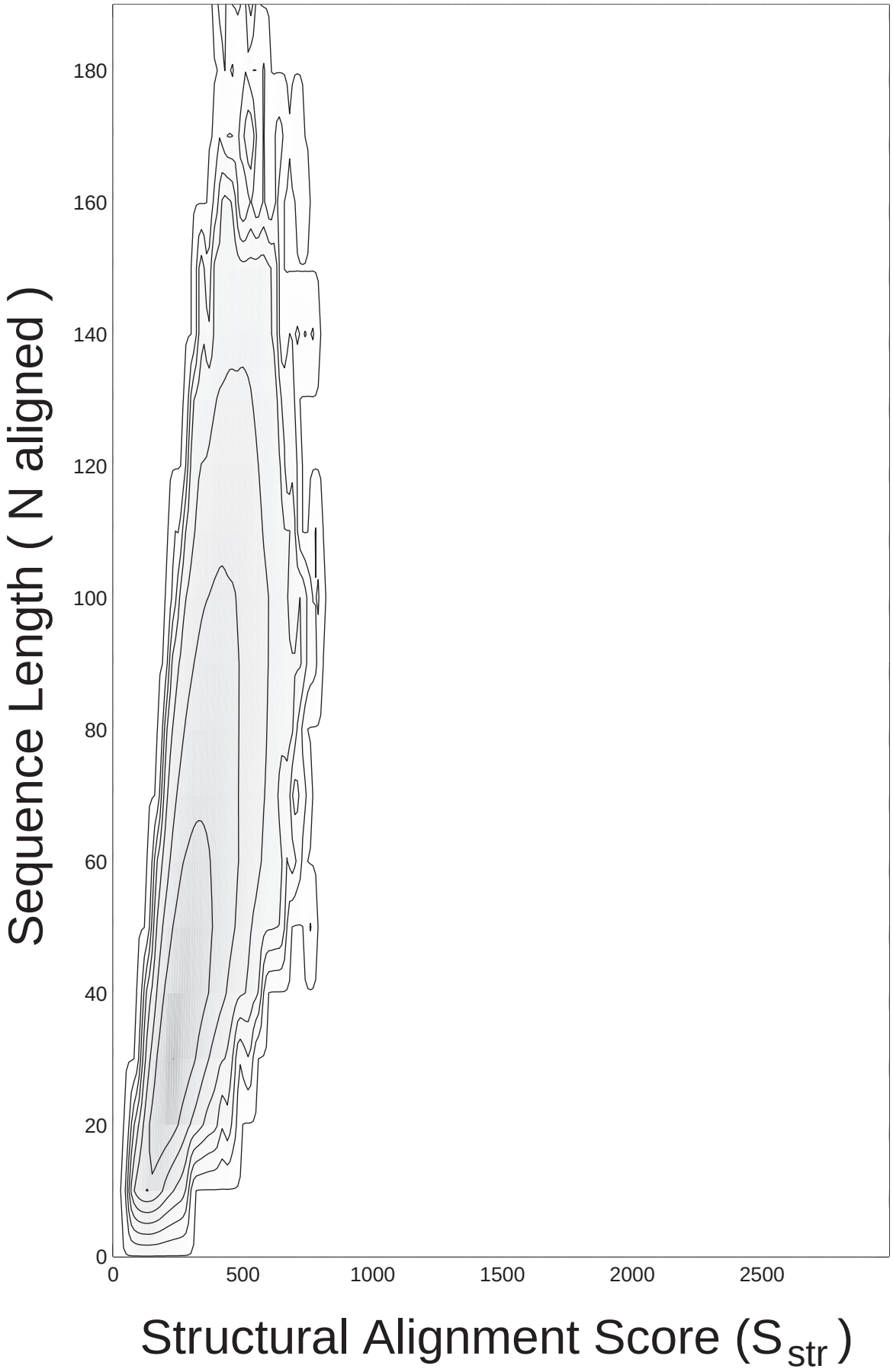
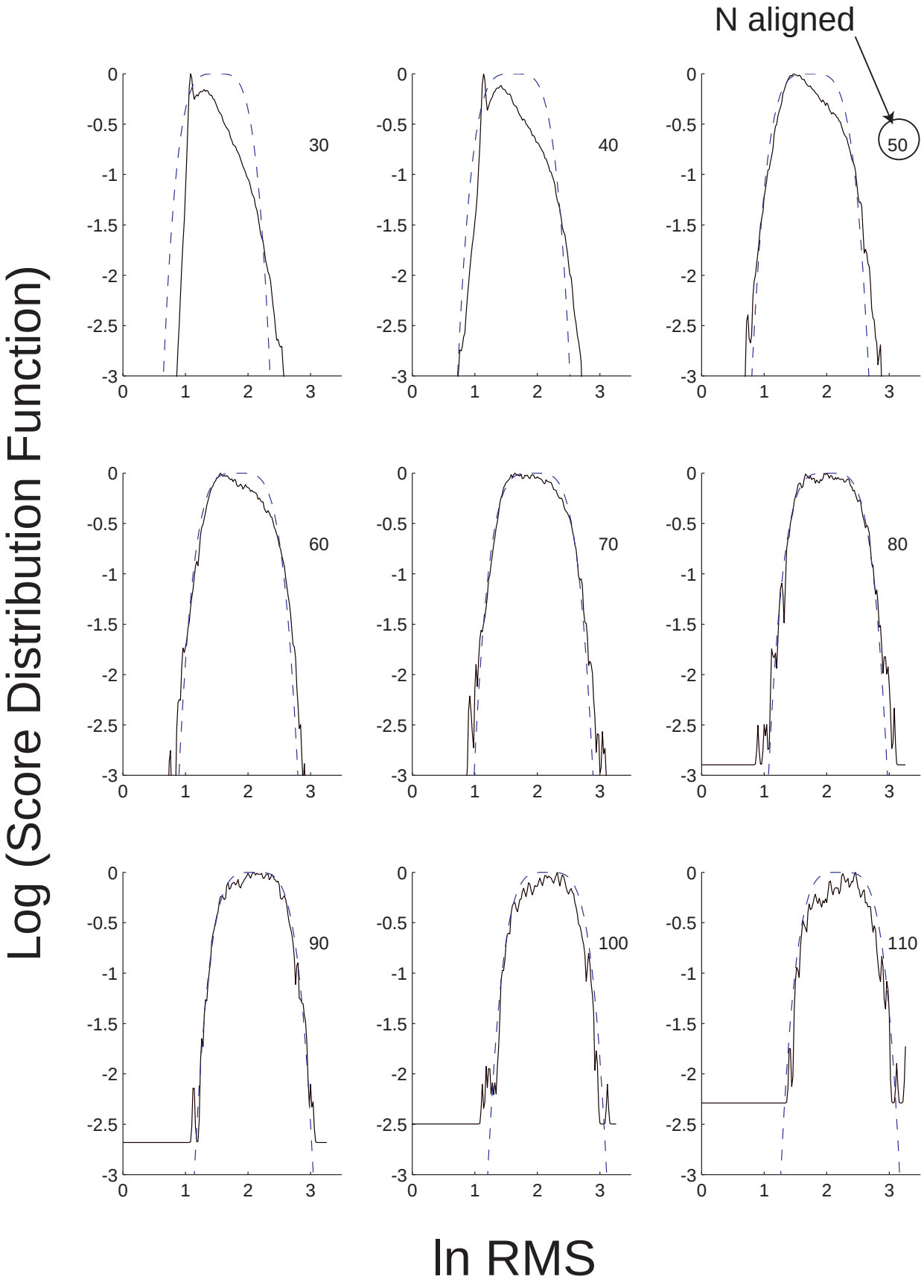
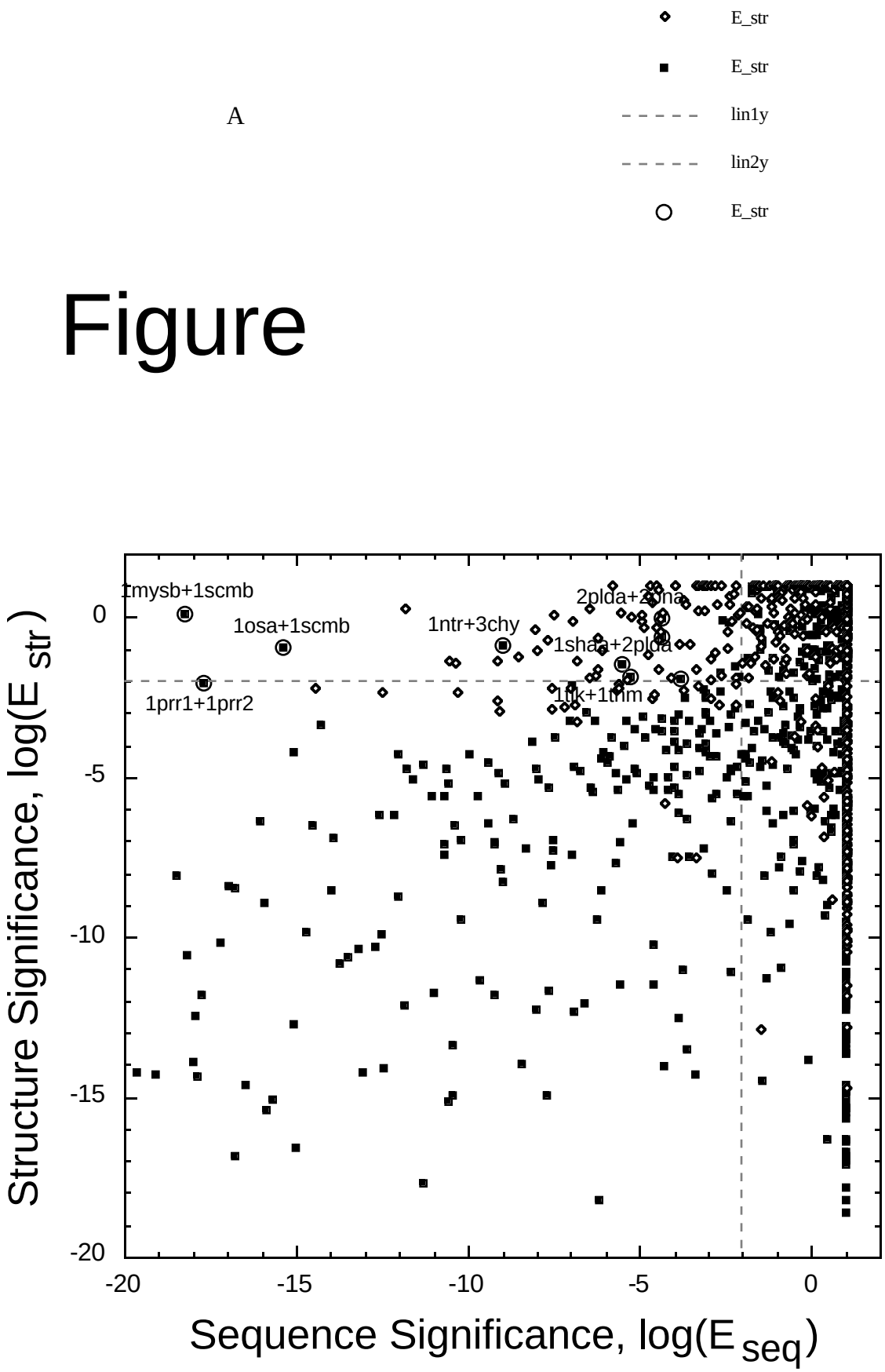

True Negatives (different class pairs)
Sequence Length (N aligned)
0
20
40
60
80
100
120
140
160
180
0
500
1000
1500
2000
2500
Structural Alignment Score (S_{str})

Fig. 5


N aligned
30
40
50
60
70
80
90
100
110
Log (Score Distribution Function)
ln RMS

A

Figure


E_str
E_str
lin1y
lin2y
E_str
1mysb+1scmb
1osa+1scmb
1ntr+3chy
1shaa+2plda
1prr1+1prr2
1tlk+1tlm
0
-5
-10
-15
-20
-20
-15
-10
-5
0
Structure Significance, log(E str)
Sequence Significance, log(E seq)