An Information Theoretic Approach to Macromolecular Modeling: I. Sequence Alignments

Tiba Aynechi* and Irwin D. Kuntz†
*Graduate Group in Biophysics, and †Department of Pharmaceutical Chemistry, University of California, San Francisco, California

ABSTRACT We are interested in applying the principles of information theory to structural biology calculations. In this article, we explore the information content of an important computational procedure: sequence alignment. Using a reference state developed from exhaustive sequences, we measure alignment statistics and evaluate gap penalties based on first-principle considerations and gap distributions. We show that there are different gap penalties for different alphabet sizes and that the gap penalties can depend on the length of the sequences being aligned. In a companion article, we examine the information content of molecular force fields.

INTRODUCTION

Structural biology now has the challenge of providing structural and functional information on a genomic scale. Current methods combine different experimental and computational procedures to deduce the structure and function of biomacromolecules. A partial list includes sequence analysis, crystallography, magnetic resonance, spectroscopy, homology modeling, and molecular dynamics. However, despite the quantitative nature of such undertakings, there is no unifying model of information content and error analysis for the field as a whole. Although there have been important specialized forays (1–7), there is a need to seek a broader approach that would permit the evaluation and comparison of such methods. A further related concern is the additivity of information when different techniques are combined. Previously, we demonstrated (8) that the basic tenets of information theory (9) can be used to quantify the information content of distance constraints. In this and a companion article (10), we apply the same general principles to simple exact models (SEMs) to draw inferences about two important tools of computational biology, sequence analysis, and force fields.

Sequence alignment is an integral part of comparative modeling protocols. Aside from ab initio methods (11), theoretical structure prediction is generally approached in two steps:

1. Given an amino acid sequence, find an appropriate structural template (using homology modeling and/or threading).
2. Refine the structural model to produce an energetically minimized or best-scoring conformation.

The first step requires sequence-alignment algorithms, which rely heavily on the use of empirical parameters such as gap penalties and scoring matrices (12). Because sequence space is poorly characterized, it is difficult to either evaluate or improve overall performance except in the context of specific training sets. What is needed is a unified picture of the fundamental issues.

A standard way to gain insight into complex problems is through SEMs. In the protein-folding field, these models use simplified representations of sequences and structures to mimic sequence and structure interactions in real systems. Thus, self-avoiding two-dimensional lattice walks and simplified alphabets have long been used to evaluate and understand the principles of protein folding (13,14). The ability to exhaustively enumerate all states of the system affords the opportunity to describe the system's behavior unambiguously, and it can provide a clear path relationship between assumptions and consequences. Thus, SEMs are well suited for formulating and evaluating general concepts: a task that may be much more difficult with real-world examples because of their heavy parameter dependence and need for approximations (15).

In this article, we combine the use of simplified systems with information theory to derive the costs of alignment procedures, scoring matrices, and gap penalties of idealized models. We then consider the applicability of the insights gained to the current approaches to sequence alignment. As noted above, our modeling choices are chosen to illuminate the underlying issues. We consider force fields in the companion article (10).

METHODS

Overview

Our basic approach is to explore a simple exact model where it is possible to write out all occurrences of the set of interest (i.e., all possible sequences) and ask what the informational consequences are of performing an operation that combines some of the objects. The information required to select one object from a set of W objects is $\log_2 W$ (9). The normal use of sequence-alignment procedures is considered to increase the information associated with a given probe sequence. That is, one queries a database of sequences

Submitted October 12, 2004, and accepted for publication August 15, 2005.

Address reprint requests to Dr. Irwin D. Kuntz, Dept. of Pharmaceutical Chemistry, University of California at San Francisco, San Francisco, CA 94143-0446. Tel.: 415-476-1937; Fax: 415-502-1411; E-mail: kuntz@cgl.ucsf.edu.

0006-3495/05/11/2998/10 $2.00

doi: 10.1529/biophysj.104.054072

and assigns properties (structure, function) to the probe sequence based on the statistically valid matches that are found. This information increase arises because a single sequence is placed into a cluster smaller than the full set of sequences from which it was indistinguishable before the alignment procedure. However, we can equally well treat alignment as a clustering procedure in which a number of sequences are grouped together as indistinguishable. Such clustering reduces the effective number of distinguishable objects compared to the full set of unique sequences. From this point of view, information is lost because a number of sequences that were treated as distinct from one another are now considered the same (i.e., similar sequence, function, or structure). In this context, gap penalties are a direct reflection of the price that must be paid for the information loss.

Of course, it is not feasible to write out all sequences for all proteins and nucleic acids. Nor can we advance a comprehensive model for the evolutionary and structural constraints that give rise to the sequences that form the current pan-genomic database. Rather, our strategy will be to uncover general properties by making use of model systems and simplified alphabets (14,16). However, we are also interested in exploring the implications of such models for the real world of sequences and conformations. This relationship is not a formal part of information theory, and will involve additional assumptions or hypotheses, the truth of which must be established by other methods. For example, it is straightforward to evaluate the informational consequences of the proposition that the known sequences are a random subset of all possible sequences; this proposal can be directly tested statistically, but information theory, alone, cannot determine its validity.

Shannon information of a set *W*

The information required to select an individual entity from a set of W objects is defined as

$$I^{\mathrm{S}} = \log_2(W). \tag{1}$$

I^{S} is referred to as the source information (9). Given a metric set, M, that partitions the objects into subsets, the information content can be measured in bits using Shannon's formulation (9),

$$I^{\mathrm{M}} = -\Sigma[p_{\mathrm{k}} \log_2(p_{\mathrm{k}})], \tag{2}$$

in which p_{k} is the population of cluster k expressed as a fraction of the ensemble, summed over all clusters. These clusters are subsets of the population that are indistinguishable under particular assumptions or constraints.

As mentioned previously, clustering can lead to a change in information. We relate Eqs. 1 and 2 to yield the information gain/loss of a clustering procedure as

$$I^{\mathrm{gain/loss}} = I^{\mathrm{S}} - I^{\mathrm{M}}. \tag{3}$$

Sequence alignment

Overview

Sequences of proteins or nucleic acids of unknown structure and function are sources of information through association with other sequences whose functions/structures are already known. The most widely used associative process is *alignment*. Alignment algorithms can be divided into two categories, global and local. A global alignment (17) looks for the best overall similarity among sequences, whereas a local alignment (18) searches for similar sub-sequences between two proteins. Both of these algorithms make use of a variety of scoring matrices and gap penalties (19–24). Sequence-alignment problems are underdetermined, having multiple optimal solutions depending on the parameters used. Thus far, there has not been a quantitative analysis of the parameter dependence, one reason being the absence of a standard comparison metric. With an information theoretic approach, we are able to formalize the effects of parameters such as sequence length, alphabet size, etc. We consider the sets of sequences of length N, drawn from an alphabet of A characters. Assume that the characters are used with equal frequency (effects of character correlation can be readily included at a later stage). With this simplifying assumption, each sequence has equal weight and there are A^{N} unique sequences. The information content of the set is simply $N\log_2 A$. Alignment procedures require the definition of a template of length $T < N$. The template may contain gaps—that is, the string for the template may contain one or more positions that match any character. Alternatively, the template may be considered continuous and the sequences with which it is being compared can contain gaps. We ask how many sequences of an exhaustive list match a specific template. Most generally, because there is nothing of special interest for any given template, we are interested in the information content averaged over all templates of a certain type.

We begin with the case of gapless pairwise alignments and then move to multigapped alignments. We will use both exhaustive and stochastic data sets, along with simple alignment models, to provide insight into the informational issues associated with sequence comparisons.

Statistics from alignments are gathered under two scenarios:

1. For every sequence in the data set, a single (first) occurrence of the template, T, is sufficient for assignment.
2. All possible occurrences of a template are sought in each sequence of the data set (multiple-occurrence model).

Gapless alignments

For an A-letter alphabet, the total number of possible N-letter sequences is A^{N},

$$W = A^{\mathrm{N}}. \tag{4}$$

In the simplest case, we consider those template sequences of length T whose elements are found in contiguous positions in probe sequences of length N. Defining $K = N-T$, the templates can be anchored in $K+1$ positions each with A^{K} possible matches, leading to an estimate of $(K+1)A^{\mathrm{K}}$ sequences if there are no duplicate sequences. Consequently, the information required to distinguish among the ungapped matches in an exhaustive set is

$$\begin{aligned} W &= (K+1)A^{\mathrm{K}} \\ I^{\mathrm{M}} &= K\log_2 A + \log_2(K+1). \end{aligned} \tag{5}$$

This formula counts exactly all occurrences of the template in complete (multiple occurrence) alignments. For a two-letter alphabet consisting of zeroes and ones, the templates are of the form 01, 001, 0001, ... Templates of higher symmetry, e.g., 000 or 010, have somewhat fewer hits (data not shown). Using the asymmetric templates gives the maximum number of hits, which also corresponds to the numerical results from the formulas.

For single-occurrence, ungapped alignments, we have an alternative approach using 1), the contiguous string; and 2), the standard formula for the probability of failure to match, p_{F}. Given the probability of occurrence of the template in a single sequence, $p_{\mathrm{T}} = (1/A)^{\mathrm{T}}$, and the number of independent attempts $(K+1)$, $p_{\mathrm{F}} = (1-p_{\mathrm{T}})^{\mathrm{K}+1}$. The probability of a hit for a sequence, p_{H}, is then $(1-p_{\mathrm{F}})$ and the total number of hits for the set is $A^{\mathrm{N}} \times p_{\mathrm{H}}$,

$$\begin{aligned} W &= A^{\mathrm{N}} \times [1-(1-p_{\mathrm{T}})^{\mathrm{K}+1}], \\ I^{\mathrm{M}} &= N\log_2 A + \log_2[1-(1-p_{\mathrm{T}})^{\mathrm{K}+1}]. \end{aligned} \tag{6}$$

This formula assumes that multiple occurrences of a template occur with an equal probability, p_{T}, in a given sequence. However, once a template appears once in a sequence, it fixes the positions it occupies and the probabilities of any subsequent overlapping templates will no longer be independent. As a result, this formula may either underestimate or overestimate the hit count depending on the template type. In the case of the templates used here (01,

0001,...), the formula underestimates. This effect is lessened as the template/sequence length ratio becomes larger, because of the diminishing possibility of overlaps.

However, Eq. 6 provides useful values for I for the full range of K for single occurrence of templates (see Results and Discussion, below).

Gapped alignments

For more general gap distributions, in which all templates of length $T = N - K$ are aligned against a probe of length N, we need to consider the combinatorial arrangement of gaps of varying length. For gaps of minimum-length one, there will be $C(N, K) = N!/K!(N–K)!$ ways to arrange the gaps in an N-long sequence. However, if we require the minimum-gap size, $G_{\min}$, to be greater than one, then the effective length of the sequence is reduced to $N_{\text{effective}} = N - K \times G_{\min} + K$. There are A^K sequences for each arrangement:

$$W = C(N_e, K)A^K$$
$$I^M = K\log_2 A + \log_2 C(N_e, K).$$

Results for $G_{\min} = 1$ are exact for complete alignments (see Results and Discussion).

We have also found a formulation leading to an exact solution for the number of gapped matches for single occurrences of the template. The number of hits to match a given template of length T, where $K = N - T$, against an exhaustive set of sequences becomes

$$W = \sum_1^K [C(N, K) \times (A-1)^K]. \quad (8)$$

This equation (discovered empirically from the counting data) provides exact counts over the complete range of N, T. When converting to bits of information, the right-hand side of Eq. 8 generally cannot be reduced to a simpler form; however, when $A > 2$ and $T < 95\%$ of N, the highest order term sufficiently dominates so that the summation is no longer needed. Under these circumstances, the information is, to a good approximation,

$$I^M = K\log_2(A-1) + \log_2 C(N, K). \quad (9)$$

Gap penalties

The formulas above quantify the amount of information associated with successful alignments when an exhaustive basis set of all possible sequences is available. They also can be used to set bounds on gap penalty values (see Results and Discussion).

Gap penalties can be derived by examining the length distributions of gaps in systems where structural alignment is possible. This approach is based on a general affine model of gap penalties (25) and uses a geometric distribution to assess the probability distribution of gaps, yielding, in the Qian and Goldstein treatment (26), the formula for gap initiation (γ_I) and gap extension (γ_E) penalties of

$$\gamma_I = \log_2\left[\frac{P_g}{1-\exp(-1/\lambda)}\right] - 2/\lambda$$
$$\gamma_E = -1/\lambda. \quad (10)$$

Here, P_g is the probability of opening a gap, and λ is the half-decay length of the gap length distribution. The values for P_g and λ are determined in a similar way to Qian and Goldstein (26). For a given sequence-length and template, we plot the distribution of gap lengths versus the probability for the observed hits (as an example, see Fig. 1). We then fit the data to an exponential of the form

$$p(n) = B \times \exp(-n/\lambda), \quad (11)$$

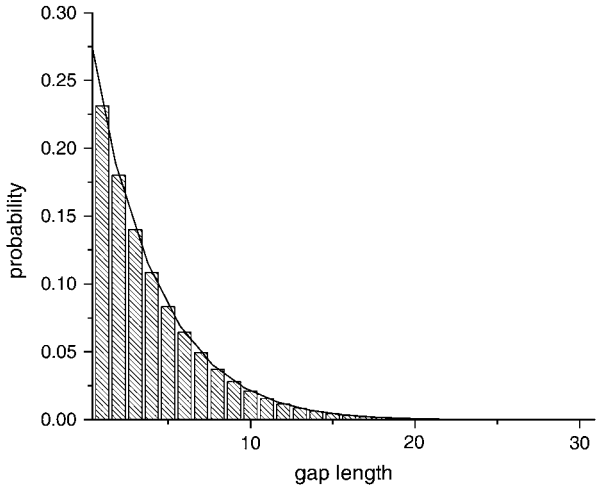

FIGURE 1 Calculation of λ and P_g from gap distributions. Sample distribution of gaps for the first-occurrence model, using a stochastic set of 50-mers with a six-letter alphabet, using a 10-mer template. The data is fit (*solid line*) to the form $p(n) = B \times \exp(-n/\lambda)$, for this case $\lambda = 3.833$, $B = 0.30298$. According to the formulation of Qian and Goldstein (20), $B = P_g \times \exp(-1/\lambda)/[1-\exp(-1/\lambda)]$. We calculate P_g by substituting λ back into the expression for B to obtain gap initiation and extension penalties according to Eq. 10.

where n is length of the gap and B is defined as $B = P_g \times \exp(-1/\lambda)/[1-\exp(-1/\lambda)]$.

To compare our exhaustive reference-state distributions to previously determined values, we use our counting experiments to measure the distribution of gap lengths for sets of sequences and templates of varying length. We then use the Qian and Goldstein equation (10) to calculate gap initiation (γ_I) and extension(γ_E) penalties.

Search algorithm methods

First-occurrence alignments

For every sequence, S, in the set, given a template T, we look for the first occurrence of the symbol in the first position of the template. Looking forward, we search for the first instance of the symbol in the second position of the template and so forth, until the last position in T. If a symbol is not observed in S in order of appearance in T, the search is terminated. Indices of each hit in the sequence are tabulated to determine the length of the gap among instances of each symbol present in the template.

Multiple-occurrence alignments

Fig. 2 shows how the occurrences of a template T are sought in a sequence, S, consisting of an alphabet of size A. The final list contains all the occurrences of T in S by specifying the indices of the symbols in S. The positional indices for each occurrence are used to determine the distribution of gap lengths.

Exhaustive versus incomplete sequence sets

Mapping

We turn to the question of how to compare results from the exhaustive list of sequences with those generated from a (sub)set of observed sequences. There are several issues. First, the set of observed sequences is not fixed but is continually updated with new sequences being added and old sequences being

```
for each α, in the alphabet A
        for (i = 0 to length(Sequence);i++)
                if (Sequence[i] == α, add i to T_ α)

// initialize the match matrix
t = template(0)
Match = T_t
//populate the match matrix
for (j=1 to length(Template); j ++)
        t = Template(j)
        w = width(Match)
        r = 0
        for (h =0 to height(Match); h++)
                for (k = 0 to length(T_t); k++)
                        if (Match[h,w] > T_t[k]) continue
                        M[r] = {Match[h], T_t[k]
                        r = r + 1
        Match = M
```

FIGURE 2 Search algorithm for finding all occurrences of a template, T, in a sequence S, using alphabet A.

modified or even deleted. For our purposes, we will ignore such issues and simply take a snapshot of the existing data. A second concern is that the observed sequences show unequal utilization of the characters. Such variable weightings were part of Shannon's initial formulation and Eq. 2 yields a single correction term equal to 0.12 bits/amino acid, based on the nonuniform composition of amino acids in real proteins (27). Higher-order terms dealing with joint probability of multiple characters can also be considered as corrections to the simple assumption of equal frequencies (28). In the same way, scoring matrices with partial weights for alternative characters reduce the effective alphabet below the limit set by equal utilization of all characters. A correction term can be generated for any scoring matrix of interest.

A harder question is the relationship of the observed sequences to the exhaustive set. Many hypotheses can be put forward about the mechanisms of evolution and the types of structural constraints imposed upon both nucleic acids and proteins. We do not propose in this article to select among them. Instead, we provide simple examples to illustrate how the mapping from exhaustive sequences to sequence subsets changes the information content of alignment operations and hence changes the values of gap penalties. These simple hypotheses are:

1. The observed sequences are a random subset of the exhaustive sequences.
2. The observed sequences are a particular evolutionary subset of the exhaustive sequences.

Again, our purpose is not to espouse these models, but to show how Eqs. 5–9 are modified in each case.

To examine the information content of a random subset of the exhaustive sequences, we generated sets of 10,000–100,000 random sequences of lengths $N = 10$, 20, 30 for $A = 2$. These were scanned with templates of various lengths and gap lengths. The number of hits was recorded with each of the sequences as a starting point and the probabilities of clustering were calculated. The information content for each alignment procedure was tabulated.

To generate a simple evolutionary model, we used the constraint that L of the N positions would not vary. This assumption produces a subset of sequences that are in exact correspondence to sequences from the exhaustive list for N' where $N' = N - L$. Equations 5–9 can then be applied directly to this subset.

Correlation of sequence alignment and conformational resolution

Of course, one random model and one simple evolutionary model just begin to explore the sequence constraints operating on the natural sequences. Presumably, one of the critical limits is that most of the experimental sequences arise from sequence subsets that provide stable three-dimensional (tertiary) structures for some range of physical variables. We have not attempted to construct such a model in this article, but others have approached the problem (29,30).

In a seminal article, Chothia and Lesk (31) provided the first quantitative relationship between sequence identity and structural similarity. In recent work, this relationship has been revisited in great detail (32). For our purposes, we use the methods above and those of Sullivan and Kuntz (33) to compare the information content of sequence and structural alignments, as follows. As a model of real proteins, we choose the backbone conformations for 100-mer compact polyalanine chains, a 20-letter alphabet, and a multiple-occurrence gapped alignment model. For a given homology level, we then compare the information of sequence and structural alignments. For our example, we select a similarity level of 90%. We use Eqs. 2, 3, and 7 to determine the information from sequence alignments. From Chothia and Lesk, a 90% identity yields a root-mean-square deviation (RMSD) of ~0.5 Å. Therefore, we ask: What is the conformational probability of 100-mer compact polyalanine models falling within 0.5 Å RMSD, as derived by Sullivan and Kuntz, to determine the information required for a corresponding structural alignment?

RESULTS AND DISCUSSION

We have studied two alignment models, each treating ungapped and gapped templates. We have both analytic formulas and statistical results for the information. In addition, we have collected statistics on gap frequencies and the probability distribution of gap lengths. As noted earlier, Eqs. 5 and 6 describe the ungapped data exactly for multiple hits (Table 1) and within an average of 3% for single-occurrence hits (Table 2), respectively. Tables 3 and 4 show that Eqs. 7 and 8 provide an exact numerical result for gapped alignments for multiple and first-occurrence models.

One of our primary concerns is the implication of these equations for gap penalties. We can get estimates of these penalties by examining the equations directly, or we can calculate the distribution of gap lengths. Equations 5–9 contain terms sensitive to K (the total length of all gaps), as well as terms that depend on the size of the alphabet, A, and the length of the sequence, N. These formulas cannot be separated cleanly into gap initiation terms and gap extension terms. However, they are generally consistent with a gap penalty that costs information at the rate of $\log_2 A$ per unit of gap length (K). The full loss (initiation + extension) at $K = 1$ is $\log_2(A \times N)$. For $N = 100$, such a penalty would be equivalent to -6.0 for $A = 4$ and -7.6 for $A = 20$ in the units normally used for sequence-alignment programs (i.e., ln A). These values are model-dependent. We also note that, for equivalent coding, the nucleic acid model, $A = 4$, would have an N of 300, yielding a penalty term of -7.1. These results are in reasonable agreement with the range of empirical gap initiation penalties reported by Qian and Goldstein (20) (see Fig. 3).

We can also examine the probability distribution of gap lengths (26). We gathered these data either from short exhaustive binary sequences or from stochastic samples of longer sequences with larger alphabets. There are two ways that we can count gap frequencies and gap lengths (see

TABLE 1 Gapless alignments for exhaustive sequence sets—multiple-occurrences as described by Eq. 5, verified numerically from the counting data

Alphabet size (A)	Template length	Sequence length	K	A^K	$(K+1)A^K$	Number of hits (actual count)
2	3	20	17	131,072	2,359,296	2,359,296
2	4	20	16	65,536	1,114,112	1,114,112
2	5	20	15	32,768	524,288	524,288
2	6	20	14	16,384	245,760	245,760
2	7	20	13	8192	114,688	114,688
2	8	20	12	4096	53,248	53,248
2	9	20	11	2048	24,576	24,576
2	10	20	10	1024	11,264	11,264
2	11	20	9	512	5120	5120
2	12	20	8	256	2304	2304
2	13	20	7	128	1024	1024
2	14	20	6	64	448	448
2	15	20	5	32	192	192
2	16	20	4	16	80	80
2	17	20	3	8	32	32
2	18	20	2	4	12	12
2	19	20	1	2	4	4
2	20	20	0	1	1	1
3	3	12	9	19,683	196,830	196,830
3	4	12	8	6561	59,049	59,049
3	5	12	7	2187	17,496	17,496
3	6	12	6	729	5103	5103
3	7	12	5	243	1458	1458
3	8	12	4	81	405	405
3	9	12	3	27	108	108
3	10	12	2	9	27	27
3	11	12	1	3	6	6
3	12	12	0	1	1	1

Methods, above). First, for the equations given above, we have used a first-occurrence model in which the gap-length data are taken from the initial successful match of a template to a sequence. Alternatively, we can identify all matches of a template with a specific sequence, and for each match, tabulate the gap-length information (multiple-occurrence) model. This model appears to be closer to the empirical data reported by Qian and Goldstein (26).

Our basic findings are:

1. Most of the gap-length distributions can be approximated by an exponential, but those arising from larger alphabets and longer templates clearly have a more complex character. The distributions can be numerically fit as multiple exponentials similar to those found by Qian and Goldstein for sequence alignments of proteins of known structures. They can also be fit with polynomial functions. It is not obvious if these expanded functions carry any physical significance.
2. For the first-occurrence model, the exponential decay increases strongly with alphabet size (Table 5). However, for the multiple-occurrence model, the exponential decay is independent of alphabet size, although it increases with N and decreases with T (Table 6). If we use the single-exponential approximation and the treatment of Qian and Goldstein (see Eq. 10), we get the range of gap penalties shown in Fig. 3. Our values are consistently on the low end of the empirical range.

To continue the comparison with the values in the literature (20), we return to the relationship between the set of observed sequences and the exhaustive reference state based on the two scenarios described earlier. *Random* models are easily constructed and tested. The stochastically derived probabilities for the alignments of random sequences show no surprises and are equal to those from the exhaustive set of sequences within the expected statistical variation (Table 7). *Evolutionary* models for the observed sequences can also be constructed. Two explicit models would be: 1), use of a full alphabet for a subset of the sequence positions with the other positions fixed; and 2), restricted alphabets at all sequence positions. In the former case, we would expect Eqs. 5–9 to apply directly, but with a reduced chain length (see Methods); in the latter, Eq. 2 can be used to compensate for the unequal probabilities of each character. To test the first evolutionary model numerically, we took the exhaustive set of fully variable 15-mers embedded in 20-mers with the first five L-positions invariant. The results (Fig. 4) for first-occurrence gapped hits closely correspond to the exhaustive 15-mer data, suggesting that Eqs. 8 and 9 are good approximations for sequences generated by evolutionary relationships. However, when we compute the multiple-occurrence gap-length distributions for the same data set, the situation is more complicated. The distribution functions are intermediate between the 15-mer and 20-mer data, with the distributions closer to the 20-mer results (Fig. 5). Others have also studied the evolutionary relationships among protein and model sequences using simple models. For example, Irback and Sandelin (34) show that real sequences deviate from random sequences in the distribution of hydrophobic residues in the chains, whereas Cui et al. (35) show that crossovers and nonhomologous combinations are favored in the evolution of low energy states. However, neither study explores information content for their systems.

The other question raised above is what type of gap simulation best captures normal alignment procedures, as for example, in the Needleman-Wunsch algorithm. The essential issue is that real-world data are drawn from a highly heterogeneous sequence set. The sequences and templates are of variable lengths, and the alphabets, although nominally of four or 20 letters, have unequal utilization of the letters in a sequence/structure-dependent manner. Furthermore, the results depend on whether first-occurrence or multiple-occurrence statistics are used. Thus, it seems unlikely that there is a *best set* of gap penalties for all alignment problems.

The empirical gap penalties currently in use are obtained from training sets on homologous proteins. Our approach in this article has been to explore simple, exhaustive treatments of sequence alignment with a much broader reference state. We show that these efforts lead directly to a priori gap penalties that depend on models of the protein and nucleic-acid

TABLE 2 Gapless alignments for exhaustive sequence sets—first occurrence as described by Eq. 6, verified numerically from the counting data

Alphabet size (A)	Template length	P_T	Seq. length*	K	$P_H^\dagger$	Eq. 6: Number of hits	Number of hits (actual)	ΔI	% Difference, number of hits
2	3	0.125	20	17	9.10E-01	953,790	1,019,920	19.96	6.48
2	4	0.0625	20	16	6.66E-01	698,541	782,497	19.58	10.73
2	5	0.03125	20	15	3.98E-01	417,637	458,495	18.81	8.91
2	6	0.015625	20	14	2.10E-01	220,618	234,280	17.84	5.83
2	7	0.007813	20	13	1.04E-01	109,042	112,896	16.78	3.41
2	8	0.003906	20	12	4.96E-02	52,018	53,008	15.69	1.87
2	9	0.001953	20	11	2.32E-02	24,314	24,552	14.58	0.97
2	10	0.000977	20	10	1.07E-02	11,209	11,263	13.46	0.48
2	11	0.000488	20	9	4.87E-03	5109	5120	12.32	0.22
2	12	0.000244	20	8	2.20E-03	2302	2304	11.17	0.10
2	13	0.000122	20	7	9.76E-04	1024	1024	10.00	0.04
2	14	6.1E-05	20	6	4.27E-04	448	448	8.81	0.02
2	15	3.05E-05	20	5	1.83E-04	192	192	7.58	0.01
2	16	1.53E-05	20	4	7.63E-05	80	80	6.32	0.00
2	17	7.63E-06	20	3	3.05E-05	32	32	5.00	0.00
2	18	3.81E-06	20	2	1.14E-05	12	12	3.58	0.00
2	19	1.91E-06	20	1	3.81E-06	4	4	2.00	0.00
2	20	9.54E-07	20	0	9.54E-07	1	1	0.00	0.00
3	3	0.037037	12	9	3.14E-01	167,064	176,957	17.43	5.59
3	4	0.012346	12	8	1.06E-01	56,215	57,835	15.82	2.80
3	5	0.004115	12	7	3.25E-02	17,246	17,442	14.09	1.12
3	6	0.001372	12	6	9.56E-03	5082	5102	12.32	0.39
3	7	0.000457	12	5	2.74E-03	1456	1458	10.51	0.11
3	8	0.000152	12	4	7.62E-04	405	405	8.66	0.03
3	9	5.08E-05	12	3	2.03E-04	108	108	6.75	0.01
3	10	1.69E-05	12	2	5.08E-05	27	27	4.75	0.00
3	11	5.65E-06	12	1	1.13E-05	6	6	2.58	0.00
3	12	1.88E-06	12	0	1.88E-06	1	1	0.00	0.00

*Sample size $= A^N$.
$^\dagger P_H = 1 - P_F$.

sequence universe. Our formulas suggest alphabet-size and sequence-length dependencies that are not included in current methods. They lead directly to two practical suggestions:

1. Gap penalties should differ for nucleic acid versus amino-acid sequence alignments.
2. It would be useful to generate sequence-length-dependent penalty corrections.

We also have examined gap-occurrence probabilities, finding that they (approximately) follow a geometric distribution.

Most of our results are at the low end of reported range of gap penalties. One reasonable explanation is that the set of known sequences is significantly nonrandom because of some combination of evolutionary and structural constraints; for example, reduced gap probabilities inside secondary structure elements, which would lead to higher-than-random gap penalties for gaps in loops. Although we cannot say that gap penalties based on SEMs will lead to better alignments than current methods, we hope that exploring the underlying relationships in simple systems

TABLE 3 Gapped alignments for exhaustive sequence sets—multiple-occurrences as described by Eq. 7, verified numerically from the counting data

Alphabet size (A)	Template length	Sequence length*	K	Number of hits (actual)	ΔI	$C(N,K)$	$(A)^K$	Eq. 7: $C(N,K)(A)^K$
2	3	20	17	149,422,080	27.15	1140	131,072	149,422,080
2	4	20	16	317,521,920	28.24	4845	65,536	317,521,920
2	5	20	15	508,035,072	28.92	15,504	32,768	508,035,072
3	3	12	9	4,330,260	22.05	220	19,683	4,330,260
3	4	12	8	3,247,695	21.63	495	6561	3,247,695
3	5	12	7	1,732,104	20.72	792	2187	1,732,104

*Sample size $= A^N$.

TABLE 4 Gapped alignments for exhaustive sequence sets—first-occurrence model as described by Eq. 8, verified numerically from the counting data

Alphabet size (A)	Template length	Sequence length*	K	Number of hits (actual)	# Failures	ΔI	$C(N,K)(A^{-1})^K$
2	3	20	17	1,048,365	1140	20.00	1140
2	4	20	16	1,047,225	4845	20.00	4845
2	5	20	15	1,042,380	15,504	19.99	15,504
2	6	20	14	1,026,876	38,760	19.97	38,760
2	7	20	13	988,116	77,520	19.91	77,520
2	8	20	12	910,596	125,970	19.80	125,970
2	9	20	11	784,626	167,960	19.58	167,960
2	10	20	10	616,666	184,756	19.23	184,756
2	11	20	9	431,910	167,960	18.72	167,960
2	12	20	8	263,950	125,970	18.01	125,970
2	13	20	7	137,980	77,520	17.07	77,520
2	14	20	6	60,460	38,760	15.88	38,760
2	15	20	5	21,700	15,504	14.41	15,504
2	16	20	4	6196	4845	12.60	4845
2	17	20	3	1351	1140	10.40	1140
2	18	20	2	211	190	7.72	190
2	19	20	1	21	20	4.39	20
3	3	12	9	435,185	112,640	16.78	112,640
3	4	12	8	322,545	126,720	16.95	126,720
3	5	12	7	195,825	101,376	16.63	101,376
3	6	12	6	94,449	59,136	15.85	59,136
3	7	12	5	35,313	25,344	14.63	25,344
3	8	12	4	9969	7920	12.95	7920
3	9	12	3	2049	1760	10.78	1760
3	10	12	2	289	264	8.04	264
3	11	12	1	25	24	4.58	24

*Sample size $= A^N$.

will lead to improved understanding of the basic principles.

At a higher level, we can use the machinery set up above to ask how the information content of sequence alignment compares to the information content of structural alignments. To do this we draw on the basic formulas derived above. We also use the work of Chothia and Lesk (31), which establishes the relationship between sequence identity and

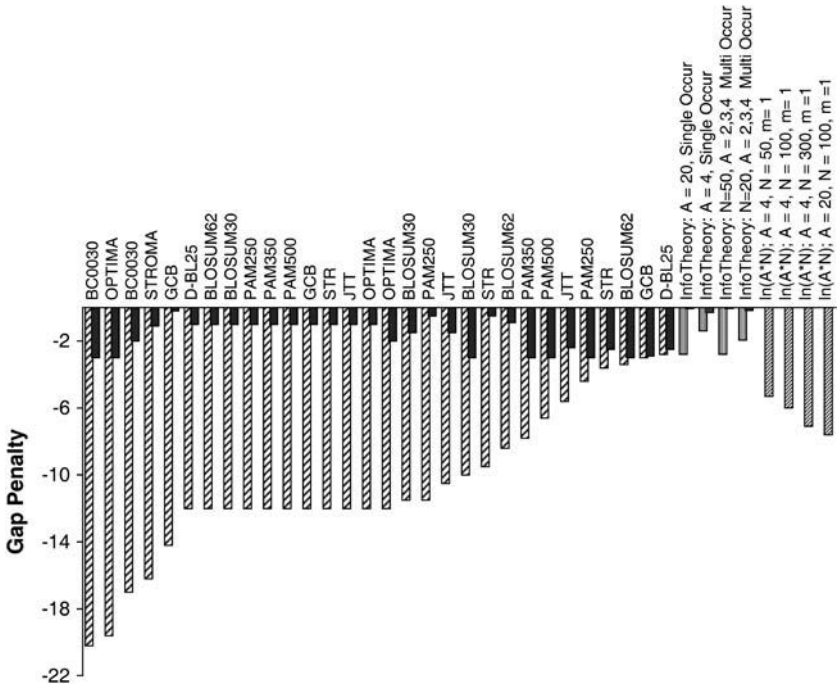

FIGURE 3 Distribution of gap initiation and extension penalties. Medium hashed bars designate gap initiation; solid bars, gap extension. Bars labeled *Info Theory* represent values derived from our gap distributions using Eq. 10. Dense hash bars indicate gap initiation + gap extension approximated using Eqs. 5–9. Data for the following were taken from Qian and Goldstein (20): BLOSUM62, BLOSUM30 (38); PAM250, PAM350, PAM500 (39); GCB (40); STR (41); JTT(42); BC0030 (43); OPTIMA (44); D-BL25(45); and STROMA (20).

TABLE 5 Gap distributions and gap penalties for the first-occurrence model

N	Alphabet size	Template length	Total number of gaps	Total number of hits	P_g	λ	$(-)\ \gamma_{gap-I}$	$(-)\ \gamma_{gap-E}$
10	2	2	1,004	1,013	1.043	1.425	0.676	0.702
10	2	3	1,398	968	1.219	1.352	0.632	0.739
10	2	4	1,528	848	1.695	1.209	0.552	0.827
20	2	2	1,048,536	1,048,555	1.000	1.443	0.693	0.693
20	2	3	1,572,291	1,048,365	1.001	1.442	0.693	0.693
20	2	4	2,092,512	1,047,225	1.004	1.441	0.692	0.694
50	2	2	1,009,981	1,000,000	0.997	1.444	0.694	0.692
50	2	3	1,514,903	1,000,000	0.996	1.445	0.694	0.692
50	2	4	2,020,116	1,000,000	0.999	1.443	0.694	0.693
50	2	5	2,524,154	1,000,000	0.997	1.444	0.694	0.692
100	2	2	999,853	1,000,000	1.000	1.442	0.693	0.693
100	2	3	1,501,252	1,000,000	1.000	1.443	0.693	0.693
100	2	4	2,000,332	1,000,000	0.999	1.443	0.693	0.693
100	2	5	2,501,670	1,000,000	0.998	1.444	0.693	0.693
20	4	4	28,163,173	774,544	0.170	2.928	1.215	0.341
20	4	5	54,233,834	585,323	0.228	2.589	1.112	0.386
20	4	6	108,467,668	383,398	0.316	2.264	1.004	0.442
20	4	7	215,925,355	214,460	0.447	1.972	0.897	0.507
50	4	4	2,998,161	999,517	0.111	3.477	1.386	0.288
50	4	5	3,740,043	997,900	0.112	3.471	1.384	0.288
50	4	6	4,461,340	992,992	0.112	3.460	1.381	0.289
50	4	7	5,126,973	980,293	0.114	3.439	1.374	0.291
100	4	4	2,999,727	1,000,000	0.111	3.472	1.385	0.288
100	4	5	3,750,300	1,000,000	0.111	3.473	1.386	0.288
100	4	6	4,500,088	999,998	0.111	3.472	1.386	0.288
100	4	7	5,248,490	999,998	0.111	3.474	1.386	0.288
100	20	2	1,826,955	963,111	0.003	19.056	2.958	0.052
100	20	3	2,500,674	881,617	0.003	17.958	2.894	0.056
100	20	4	2,792,228	741,973	0.004	16.435	2.804	0.061
100	20	5	2,637,250	564,189	0.005	14.822	2.703	0.067
100	20	6	2,133,962	383,374	0.006	13.287	2.597	0.075

structural variation, and the article of Sullivan and Kuntz (36), which gives log odds probabilities of structural variation. For this purpose, we treat a specific example of a compact 100 polyalanine backbone conformations because the Chothia-Lesk relation applies to native proteins. We ask what is the information content of a 90% identity match using a 20-letter, equal frequency, alphabet. From Eq. 7 we get 87 bits of required information (I^M) for a 90% identity

TABLE 6 Gap distributions and gap penalties for the multiple-occurrence model

N	Alphabet size	Template length	Total number of gaps	Total number of hits	P_g	λ	$(-)\ \gamma_{gap-I}$	$(-)\ \gamma_{gap-E}$
20	2	3	298,844,160	149,422,080	0.031	5.974	1.944	0.167
20	2	4	952,565,760	317,521,920	0.058	4.298	1.737	0.233
20	2	5	2,032,140,288	508,035,072	0.096	3.330	1.592	0.300
20	3	3	84,355,136	42,177,568	0.028	6.301	1.990	0.159
20	3	4	178,967,385	59,655,795	0.052	4.542	1.783	0.220
20	3	5	254,217,392	63,554,348	0.085	3.528	1.637	0.283
20	3	7	212,005,428	35,334,238	0.217	2.222	1.414	0.450
20	4	3	35,598,996	17,799,498	0.031	5.976	1.945	0.167
20	4	4	56,714,217	18,904,739	0.052	4.545	1.784	0.220
20	4	5	60,481,940	15,120,485	0.085	3.529	1.637	0.283
20	4	7	28,234,680	4,705,780	0.187	2.367	1.456	0.423
50	2	3	4,947,211,366	2,473,605,683	0.004	15.899	2.774	0.063
50	2	4	43,593,312,450	14,531,104,150	0.008	11.713	2.529	0.085
50	2	5	264,694,071,300	66,173,517,825	0.012	9.488	2.362	0.105
50	3	3	1,452,6205,02	7,26,310,251	0.004	15.902	2.774	0.063
50	3	4	8,535,094,731	2,845,031,577	0.008	11.717	2.529	0.085
50	3	5	34,896,710,928	8,724,177,732	0.012	9.292	2.342	0.108
50	3	7	319,645,429,159	45,663,632,737	0.023	6.661	2.087	0.150
50	4	3	611,835,994	305,917,997	0.004	16.223	2.794	0.062
50	4	4	2,697,6219,72	8,99,207,324	0.008	11.715	2.529	0.085
50	4	5	8,262,674,472	2,065,668,618	0.012	9.291	2.342	0.108
50	4	7	36,452,069,610	6,075,344,935	0.023	6.659	2.087	0.150

TABLE 7 Data comparing the hit probabilities from exhaustive and stochastic sequence sets for $A = 2$, $N = 20$

	Number of sequences					
	10,000		100,000		1,048,576*	
Template	Hit count	Probability	Hit count	Probability	Hit count	Probability
Gapped						
0.*0.*1	9997	0.1244	99,974	0.1251	1,048,365	0.1250
0.*0.*0.*1	9991	0.1243	99,876	0.1250	1,047,225	0.1248
0.*0.*0.*0.*1	9941	0.1237	99,383	0.1243	1,042,380	0.1243
0.*0.*0.*0.*0.*1	9798	0.1219	97,877	0.1225	1,026,876	0.1224
0.*0.*0.*0.*0.*0.*1	9434	0.1174	94,166	0.1178	988,116	0.1178
Gapless						
001			97,208	0.9721	1,019,920	0.9727
0001			74,514	0.7451	782,497	0.7462
00001			43,569	0.4357	458,495	0.4373
000001			22,198	0.2220	234,280	0.2234
0000001			10,734	0.1073	112,896	0.1077

*Exhaustive set.

alignment. The full protein requires 432 bits (I^S) (Eqs. 1 and 2), so the information gain from a 90% alignment is 345 bits (Eq. 3). A correction for the known frequency of use of the amino acids is ~ -17 bits, so that a 90% homology match using realistic frequencies contains something approaching 3.5 bits/residue of information. From Chothia and Lesk, as well as later work (32), we see that 90% homology implies a structural variance of $\sim$0.5 Å. Using the data from the cumulative distribution function of Sullivan and Kuntz (36) and Eq. 2 of this article, we can estimate how much information is required to achieve an RMSD of 0.5 Å for a stochastic population of compact polyalanine chains. We get $\sim$2.4–2.5 bits/residue required to select this RMSD distribution from a stochastic set of compact chains. This calculation indicates that, roughly speaking, high-end sequence alignment combined with homology modeling could approach the quality of direct structural measurements for determining backbone geometries. It also implies that the designability hypothesis (i.e., many sequences per structure) derived from lattice models (37) can also be supported from information content assessments of off-lattice conformational estimates.

In the companion article following, we extend these calculations to include comparison with force fields as well, showing how the use of information theory allows direct comparison of quite diverse techniques.

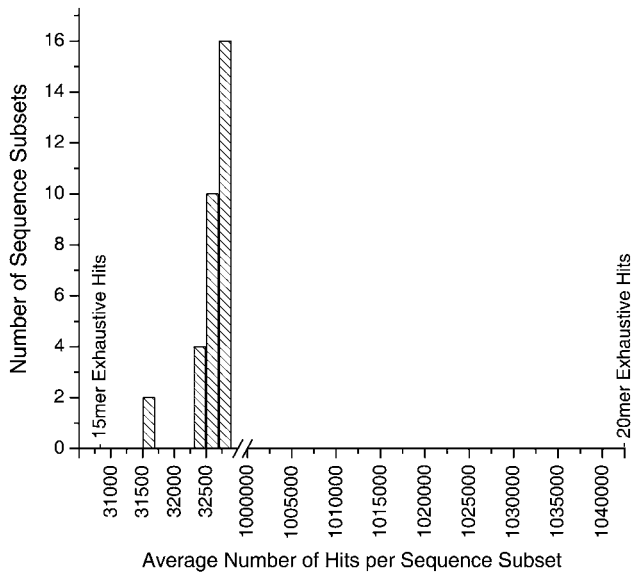

FIGURE 4 Average number of gapped hits for all possible 5-mer templates in 32 related 20-mer evolutionary subsets ($N = 20$, $L = 5$) using the single-occurrence model. Average hits for the 15-mer and 20-mer exhaustive sets are shown on the axis ends.

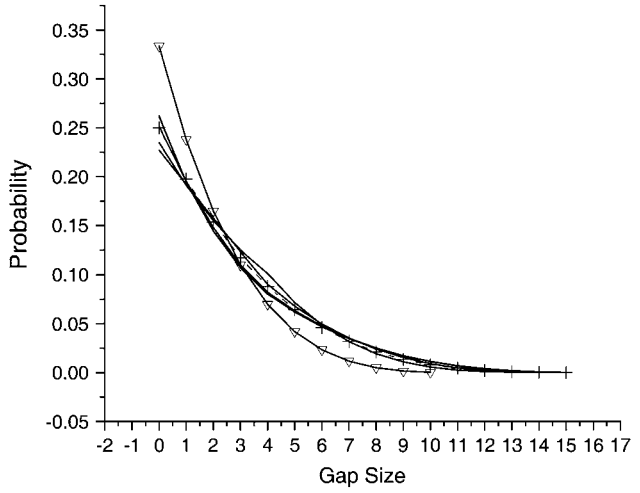

FIGURE 5 Gap-length distributions in multiple-occurrence alignments for a 20-mer evolutionary subset ($N = 20$, $L = 5$). All possible 5-mer binary templates are aligned against the sequence set resulting in 32 different distributions represented with black solid lines. For comparison we also show the behavior of the 5-mer templates in 20-mer ($-+-$) and 15-mer ($-\nabla-$) exhaustive sets. Some of the distributions from the evolutionary set are almost identical to the distributions from the 15-mer and 20-mer exhaustive sets.

We are grateful for helpful discussions with Scott Pegg and Kevin Masukawa.

This work was supported by a grant from the National Science Foundation (No. CHE-0118481), R. Kip Guy, Principal Investigator; a National Institutes of Health grant (No. RR019864), C. Pancerella, Principal Investigator; and a University of California President's Dissertation Year Fellowship (to T.A.).

REFERENCES

1. Luzzati, V. 1952. Statistical treatment of errors in the determination of crystalline structures. *Acta Crystallogr.* 5:802–810.
2. Stroud, R. M., and E. B. Fauman. 1995. Significance of structural changes in proteins: expected errors in refined protein structures. *Protein Sci.* 4:2392–2404.
3. Brunger, A. T. 1992. Free *R* value: a novel statistical quantity for assessing the accuracy of crystal structures. *Nature.* 355:472–475.
4. Berger, B., J. Kleinberg, and T. Leighton. 1999. Reconstructing a three-dimensional model with arbitrary errors. *J. ACM.* 46:212–235.
5. Levitt, M., and M. Gerstein. 1998. A unified statistical framework for sequence comparison and structure comparison. *Proc. Natl. Acad. Sci. USA.* 95:5913–5920.
6. Park, B., and M. Levitt. 1996. Energy functions that discriminate x-ray and near-native folds from well-constructed decoys. *J. Mol. Biol.* 258:367–392.
7. Carothers, J. M., S. C. Oestreich, J. H. Davis, and J. W. Szostak. 2004. Informational complexity and functional activity of RNA structures. *J. Am. Chem. Soc.* 126:5130–5137.
8. Sullivan, D. C., T. Aynechi, V. A. Voelz, and I. D. Kuntz. 2003. Information content of molecular structures. *Biophys. J.* 85:174–190.
9. Shannon, C. E. 1948. A mathematical theory of communication. *Bell Sys. Tech. J.* 27:379–423,623–656.
10. Aynechi, T., and I. D. Kuntz. 2005. An information theoretic approach to macromolecular modeling: II. Force fields. *Biophys. J.* 89:3008–3016.
11. Bonneau, R., and D. Baker. 2001. Ab initio protein structure prediction: progress and prospects. *Annu. Rev. Biophys. Biomol. Struct.* 30:173–189.
12. Vingron, M., and M. S. Waterman. 1994. Sequence alignment and penalty choice. Review of concepts, case studies and implications. *J. Mol. Biol.* 235:1–12.
13. Dill, K. A., S. Bromberg, K. Yue, K. M. Fiebig, D. P. Yee, P. D. Thomas, and H. S. Chan. 1995. Principles of protein folding—a perspective from simple exact models. *Protein Sci.* 4:561–602.
14. Wroe, R., E. Bornberg-Bauer, and H. S. Chan. 2005. Comparing folding codes in simple heteropolymer models of protein evolutionary landscape: robustness of the superfunnel paradigm. *Biophys. J.* 88:118–131.
15. Habeck, M., M. Nilges, and W. Rieping. 2005. Replica-exchange Monte Carlo scheme for Bayesian data analysis. *Phys. Rev. Lett.* 94:018105.
16. Solis, A. D., and S. Rackovsky. 2000. Optimized representations and maximal information in proteins. *Proteins.* 38:149–164.
17. Needleman, S. B., and C. D. Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* 48:443–453.
18. Smith, T. F., and M. S. Waterman. 1981. Identification of common molecular subsequences. *J. Mol. Biol.* 147:195–197.
19. Apostolico, A., and R. Giancarlo. 1998. Sequence alignment in molecular biology. *J. Comput. Biol.* 5:173–196.
20. Qian, B., and R. A. Goldstein. 2002. Optimization of a new score function for the generation of accurate alignments. *Proteins.* 48:605–610.
21. Altschul, S. F. 1998. Generalized affine gap costs for protein sequence alignment. *Proteins.* 32:88–96.
22. Koretke, K. K., Z. Luthey-Schulten, and P. G. Wolynes. 1996. Self-consistently optimized statistical mechanical energy functions for sequence structure alignment. *Protein Sci.* 5:1043–1059.
23. Lesk, A. M., M. Levitt, and C. Chothia. 1986. Alignment of the amino acid sequences of distantly related proteins using variable gap penalties. *Protein Eng.* 1:77–78.
24. Benner, S. A., M. A. Cohen, and G. H. Gonnet. 1993. Empirical and structural models for insertions and deletions in the divergent evolution of proteins. *J. Mol. Biol.* 229:1065–1082.
25. Zachariah, M. A., G. E. Crooks, S. R. Holbrook, and S. E. Brenner. 2005. A generalized affine gap model significantly improves protein sequence alignment accuracy. *Proteins.* 58:329–338.
26. Qian, B., and R. A. Goldstein. 2001. Distribution of index lengths. *Proteins.* 45:102–104.
27. Strait, B., and T. Dewey. 1996. The Shannon information entropy of protein sequences. *Biophys. J.* 71:148–155.
28. Cline, M. S., K. Karplus, R. H. Lathrop, T. F. Smith, R. G. Rogers, Jr., and D. Haussler. 2002. Information-theoretic dissection of pairwise contact potentials. *Proteins.* 49:7–14.
29. Helling, R., H. Li, R. Melin, J. Miller, N. Wingreen, C. Zeng, and C. Tang. 2001. The designability of protein structures. *J. Mol. Graph. Model.* 19:157–167.
30. Lau, K. F., and K. A. Dill. 1990. Theory for protein mutability and biogenesis. *Proc. Natl. Acad. Sci. USA.* 87:638–642.
31. Chothia, C., and A. M. Lesk. 1986. The relation between the divergence of sequence and structure in proteins. *EMBO J.* 5:823–826.
32. Gan, H. H., R. A. Perlow, S. Roy, J. Ko, M. Wu, J. Huang, S. Yan, A. Nicoletta, J. Vafai, D. Sun, L. Wang, J. E. Noah, S. Pasquali, and T. Schlick. 2002. Analysis of protein sequence/structure similarity relationships. *Biophys. J.* 83:2781–2791.
33. Sullivan, D. C., and I. D. Kuntz. 2001. Conformation spaces of proteins. *Proteins.* 42:495–511.
34. Irback, A., and E. Sandelin. 2000. On hydrophobicity correlations in protein chains. *Biophys. J.* 79:2252–2258.
35. Cui, Y., W. H. Wong, E. Bornberg-Bauer, and H. S. Chan. 2002. Recombinatoric exploration of novel folded structures: a heteropolymer-based model of protein evolutionary landscapes. *Proc. Natl. Acad. Sci. USA.* 99:809–814.
36. Sullivan, D. C., and I. D. Kuntz. 2004. Distributions in protein conformation space: implications for structure prediction and entropy. *Biophys. J.* 87:113–120.
37. Li, H., C. Tang, and N. S. Wingreen. 2002. Designability of protein structures: a lattice-model study using the Miyazawa-Jernigan matrix. *Proteins.* 49:403–412.
38. Henikoff, S., and J. G. Henikoff. 1992. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA.* 89:10915–10919.
39. Dayhoff, M. O. 1978. A model of evolutionary change in proteins. *In* Atlas of Protein Sequence and Structure. National Biomedical Research Foundation, Silver Spring, MD. 345–352.
40. Gonnet, G. H., M. A. Cohen, and S. A. Benner. 1992. Exhaustive matching of the entire protein sequence database. *Science.* 256:1443–1445.
41. Overington, J., D. Donnelly, M. S. Johnson, A. Sali, and T. L. Blundell. 1992. Environment-specific amino acid substitution tables: tertiary templates and prediction of protein folds. *Protein Sci.* 1:216–226.
42. Jones, D. T., W. R. Taylor, and J. M. Thornton. 1992. The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci.* 8:275–282.
43. Blake, J. D., and F. E. Cohen. 2001. Pairwise sequence alignment below the twilight zone. *J. Mol. Biol.* 307:721–735.
44. Doolittle, R. F. 1986. Of Urfs and Orfs: A Primer on How to Analyze Derived Amino Acid Sequences. University Science Books, Mill Valley, CA.
45. Mallick, P., D. Rice, and D. Eisenberg. DAPS: database of distant aligned protein structures. http://www.doe-mbi.ucla.edu/~parag/DAPS/,2001.