Dr. L. Segel, Editor
Bulletin of Mathematical Biology
Department of Applied Mathematics
and Computer Science
The Weizmann Institute of Science
Rehovot IL-76100
Israel

23 June, 1998

Dear Dr. Segel,

Thank you for your conditional acceptance of our manuscript BMB 97-51 and 97-52, "A Bayes-Optimal Probability Theory...". We have carefully reviewed the reviewer's comments and substantially agree with all but one, which is beyond the scope of the paper. The manuscript has been revised as suggested by the reviewer, and now is shorter and more tightly focused. The suggestion to add an example helped considerably. Specific changes are summarized below. I would be grateful to learn if the changes made are acceptable to you.

Enclosed is a floppy disk with the text of the paper in Latex and the figures as postscript files. Please let me know if I can supply anything else. Thank you very much for your time and kind attention.

Respectfully,

Richard H. Lathrop

Response to reviewer's comments:

- *the presentation is very detailed, perhaps too detailed:*

  The presentation has been shortened by removing detail.

- *Examples should be given:*

  We have added examples.

- *a strange paragraph on implementation is included in the second paper:*

  We have removed the strange paragraph.

- *the very important observation that simultaneous optimization of structure and alignment is crucial, should be illustrated with an example:*

  It is now illustrated with an example.

- *I don't see any reason to split this work into two papers ... the second paper can be an appendix to the paper:*

  The work is no longer split. The second paper is now an appendix to the paper.

- *I think the authors should address questions related to the size of the data base and the problem of low counts:*

  This is beyond the stated scope of the paper. We have added a sentence pointing out that these problems must be addressed by the particular theory of protein structure adopted, which the paper considers to be arbitrary and fixed in advance.

- *the two lines of "NP-hard Markov, Bayes normalization", and of "NP-hard Partition Function Normalizations":*

  The unexplained passages have been removed.

- *Second paper. ... explicit analysis (or even mentioning) of running time in the paper:*

  We have added sentences describing the complexity analysis of the running times.

- *Second paper. Section 3, results. (page 15) This is a strange paragraph:*

  The strange paragraph has been removed.

A Bayes-Optimal Probability Theory That Unifies Protein Sequence-Structure Recognition and Alignment

Richard H. Lathrop(1,4), Robert G. Rogers Jr.(2),
Temple F. Smith(2), James V. White(2,3)

(1) Department of Information and Computer Science,
University of California, Irvine, CA 92717 USA
(2) BioMolecular Engineering Research Center, Boston University,
36 Cummington Street, Boston, Massachusetts, 02215 USA
(3) TASC, 55 Walkers Brook Drive, Reading, Massachusetts, 01867 USA
(4) Corresponding author.

Contact information:
Prof. Richard H. Lathrop
Dept. of Information and Computer Science
444 Computer Science Bldg. #3425
University of California, Irvine
Irvine, CA 92697-3425
(949) 824-4021 office
(949) 824-4056 fax
rickl@uci.edu email

Running title:
A Bayes-Optimal Sequence-Structure Theory

Keywords:
protein threading; inverse folding; fold recognition; sequence; structure; alignment; pair potentials; contact potentials; knowledge-based potentials.

Abstract

A rigorous Bayesian analysis is presented that unifies protein sequence-structure alignment and recognition. Given a sequence, explicit formulae are derived to select (1) its globally most probable core structure from a structure library; (2) its globally most probable alignment to a given core structure; (3) its most probable joint core structure and alignment chosen globally across the entire library; and (4) its most probable individual segments, secondary structure, and super-secondary structures across the entire library. The computations involved are NP-hard in the general case (3D-3D). Fast exact recursions for the restricted sequence singleton-only (1D-3D) case are given. Conclusions include: (a) the most probable joint core structure and alignment is not necessarily the most probable alignment of the most probable core structure, but rather maximizes the product of core and alignment probabilities; (b) use of a sequence-independent linear or affine gap penalty may result in the highest-probability threading not having the lowest score; (c) selecting the most probable core structure from the library (core structure selection or fold recognition only) involves comparing probabilities summed over all possible alignments of the sequence to the core, and not comparing individual optimal (or near-optimal) sequence-structure alignments; and (d) assuming uninformative priors, core structure selection is equivalent to comparing the ratio of two global means.

1 Introduction — Protein Threading

Protein structure prediction is one of the great unsolved challenges of modern molecular biology. The direct approach, based on modeled atomic force fields [Weiner et al., 1984], [Brooks et al., 1990], as yet faces stiff challenges [Novotny et al., 1988], [Moult et al., 1995], though recent versions using cruder force fields are promising [Srinivasan & Rose, 1995] [Skolnick et al., 1997]. The most successful current method, where applicable, is homology modeling (homological extension) based on primary sequence similarity to another protein of known structure [Sankof & Kruskal, 1983], [Greer, 1990]. This approach is of limited generality because most novel protein sequences have insufficient primary sequence similarity to any known structure for the modeling step to be carried out.

Recently, however, there has been interest in aligning a protein sequence directly to a known structure (a process sometimes called "protein threading" or "inverse structure prediction"). The threading literature is extensive; for reviews see [Bowie & Eisenberg, 1993], [Bryant & Altschul, 1995], [Fetrow & Bryant, 1993], [Jernigan & Bahar, 1996], [Jones & Thornton, 1993], [Jones & Thornton, 1996], [Lemer et al., 1995], [Sippl, 1995], [Wodak & Rooman, 1993], while for cautionary notes see [Crippen, 1996], [Lathrop & Smith, 1996], [Moult et al., 1995], [Ouzounis et al., 1993], [Russell & Barton, 1994], [Smith et al., 1997b], [Thomas & Dill, 1996].

To predict accurately the structure of a novel protein sequence using the threading approach, it is necessary both to select the proper core structure from a library of known examples ("fold recognition"), and to align the sequence to it correctly ("sequence-structure alignment"). The selected core structure provides a discrete set of possible amino acid residue positions in three-dimensional space, e.g., perhaps defined by C_α or C_β locations. The sequence is aligned to the core based on an objective function (alignment score, pseudo-energy, potential function), and is given a similar three-dimensional fold by placing its amino acid residues into the spatial positions implied by the alignment. Non-conserved loop regions are usually too variable to correspond directly, and side-chains are usually abstracted away, so steps of loop placement [Greer, 1990], [Zheng et al., 1993] and side-chain packing [Desmet et al., 1992], [Mandal & Linthicum, 1993] followed by energy minimization [Weiner et al., 1984], [Brooks et al., 1990] still would remain following the protein threading process considered here. A major limitation of the threading approach is that if an appropriate core is not already present in the structure library, correct prediction is obviously impossible. Some attempts have been made to assemble structure fragments into a novel core [Sippl et al., 1992], [Kolinski et al., 1996], and [Simons et al., 1997].

A fundamental algorithmic complexity barrier is crossed if both variable alignments and pair interactions between sequence residue positions are modeled. With the inclusion of pair interactions (the 3D-3D case [Bowie et al., 1991] [Lüthy et al., 1992]), the general problems of protein folding [Ngo & Marks, 1992], [Fraenkel, 1993], [Unger & Moult, 1993], protein threading [Lathrop, 1994], [Akutsu & Miyano, 1997], and protein structure comparison [Holm & Sander, 1996], all are known to be NP-hard [Garey & Johnson, 1979].

With no pair interactions (the singleton-only or 1D-3D case), array-based or dynamic programming methods provide fast exact recursive solutions [Sankof & Kruskal, 1983]. The body of this paper analyzes the pair interaction (3D-3D) case. Appendix A shows that the computations involved are NP-hard. Appendix B provides fast exact recursions for the case of no pair interactions (1D-3D).

This paper provides a mathematical analysis of the general problem of selecting the proper core from a structure library and aligning the sequence to it. Given a sequence, explicit formulae are derived to select (1) its globally most probable core structure from a structure library; (2) its globally most probable alignment to a given core structure; (3) its most probable joint core structure and alignment chosen globally across the entire library; and (4) its most probable individual segments, secondary structure, and super-secondary structures across the entire library. Conclusions include: (a) the most probable joint core structure and alignment is not necessarily the most probable alignment of the most probable core structure, but rather maximizes the product of core and alignment probabilities; (b) use of a sequence-independent linear or affine gap penalty may result in the highest-probability threading not having the lowest score; (c) selecting the most probable core structure from the library (core structure selection or fold recognition) involves comparing probabilities summed over all possible alignments of the sequence to the core, and not comparing individual optimal (or near-optimal) sequence-structure alignments; and (d) assuming uninformative priors, core structure selection is equivalent to comparing the ratio of two global means.

1.1 Protein Threading Bayesian Analysis

Protein threading approaches usually require:

(i) a library of known core structures,

(ii) an objective function for evaluating a given alignment of a sequence to a core structure,

(iii) a method for selecting the best alignment of a sequence to a core structure (see figure 1), and

(iv) a method for selecting the best core structure from the library (see figure 2).

To succeed at the predictive task, the predictions of both core and alignment must be correct simultaneously (see figure 3).

Figures 1, 2, 3, about here.

This paper focuses on a probabilistic Bayesian theory that unifies core recognition and sequence-structure alignment (requirements iii and iv). It analyzes closely the consequences of choosing "best = globally highest conditional probability." For this case, it provides a probabilistic Bayesian theory that unifies core recognition and sequence-structure alignment (requirements iii and iv). The theory is Bayes-optimal because it indicates cores and alignments that are globally most probable.

The particular forms of the core structures and the objective function (requirements i and ii) are determined by the particular theory of protein structure adopted. As always, the theory of protein structure adopted must address the limited size of the data base, the problem of low counts in rare environments, and so on. These issues are beyond the scope of this paper. This paper assumes only that the objective function may be interpreted as encoding the probability of observing a given sequence in a given alignment to a given core structure. Otherwise, requirements (i) and (ii) are considered to be arbitrary and fixed in advance.

Bayes [Bayes, 1763] provided the first exact treatment of inference based on inverting conditional probabilities. His interpretation of the formula, $P(A|B) = P(B|A)P(A)/P(B)$, is well known. Today the mathematics of Bayesian methods is a central component of optimal statistical inference [Box & Tiao, 1973], [Hartigan, 1983]. Conditional probability and Bayesian methods have been applied to protein threading [Bryant & Lawrence, 1993], [Bryant & Altschul, 1995], [Madej et al., 1995], protein secondary structure [Arnold et al., 1992], [Stultz et al., 1995], side-chain packing [Dunbrack & Cohen, 1997], fragment assembly [Simons et al., 1997], solvent exposure prediction [Thompson & Goldstein, 1996], motif discovery [Lawrence et al., 1993], and structure classification [Stultz et al., 1995], [Hunter & States, 1992], [White et al., 1994b], all with good results. In this paper, the Bayesian analysis provides a compact account of the globally most probable cores and alignments.

2 Methods — Problem Formalization

This paper uses the gapped block approach to protein threading [Greer, 1990], [Jones et al., 1992], [Bryant & Lawrence, 1993], [Bryant & Altschul, 1995], [Madej et al., 1995], [Xu & Uberbacher, 1996], [Akutsu & Miyano, 1997], [Akutsu & Tashimo, 1998], [Xu et al., 1998]; see Figure 1. The protein core is modeled as a set of core segments ("blocks") which may represent segments of conserved secondary structure. Each segment is composed of a set of contiguous primitive core elements which are occupied by residues from the sequence to be threaded. Variable-length loop regions separate the core segments and absorb any alignment gaps. Pair interactions occur only between core elements in core segments.

Additional background on the problem formalization and notational conventions may

be found in [Lathrop, 1994], [White et al., 1994a], [Stultz et al., 1995], [Lathrop & Smith, 1996], [Smith et al., 1997a], which this paper follows. Appendix B discusses the formalism and notation in detail. Table 1 summarizes the notation used.

Table 1 about here.

This paper assumes the availability of an objective function f, called the alignment score, that satisfies

$$P(\mathbf{a}|n, C, \mathbf{t}) \propto \exp\left(-f(\mathbf{a}, C, \mathbf{t})\right) \tag{1}$$

where $\mathbf{a}$ is a sequence of length n; C is a core structure; $\mathbf{t}$ is a vector that specifies a sequence-structure alignment (a threading) and whose i^{th} component t_i specifies the alignment of core segment i; and $P(A|B)$ is the conditional probability of A given B. That is, f is the negative logarithm of an unnormalized conditional probability. It encodes the probability of observing sequence $\mathbf{a}$ aligned by $\mathbf{t}$ to core C. For example, White et al. [White et al., 1994a] and Stultz et al. [Stultz et al., 1995] describe how to construct such an objective function based on Markov Random Field (MRF) theory.

Many published threading approaches are grounded in an underlying probabilistic objective function of this general nature. In practice, they may convert the underlying probabilistic objective function from a strict conditional probability to an odds-ratio relative to some assumed reference state, say $P(\mathbf{a}|n, C, \mathbf{t})/P_{ref}(\mathbf{a}|n, C, \mathbf{t})$. The general approach in this paper is directly equivalent to setting $P_{ref}(\mathbf{a}|n, C, \mathbf{t}) = \textit{constant}$ in some odds-ratio approaches. The necessary reference corrections, here derived from first principles of probability theory, play the role of an assumed reference state.

The body of this paper assumes that core segment length is fixed, even though [Lathrop & Smith, 1996] show empirically how this can lead to threading errors. Similarly, here we assume fixed core topology (i.e., segment rank order and direction). Some important approaches [Finkelstein & Reva, 1991], [Madej et al., 1995] treat core segment length as variable by adding residue positions to, or deleting them from, core segment endpoints. This would be modeled using Appendix C and $2m$ additional integer parameters, each specifying one segment endpoint adjustment relative to the model. Variable topology, e.g. alternate arrangements of β-strands in a β-sheet, arises easily from core segment permutations or reversals. This would be modeled using Appendix C and m additional integer parameters, each specifying the rank order and direction of one segment.

3 Results — Selection Criteria

For a given sequence, this section develops formulae for selecting

1. the most probable alignment $\mathbf{t}$ to a given core structure (see figure 1), which maximizes $P(\mathbf{t}|\mathbf{a}, n, C)$;

2. the most probable core structure C across the entire library (see figure 2), which maximizes $P(C|\mathbf{a}, n)$;

3. the most probable joint core structure and alignment $\langle C, \mathbf{t}\rangle$ across the entire library (see figure 3), which maximizes $P(C, \mathbf{t}|\mathbf{a}, n)$, and which need not be the most probable alignment of the most probable core structure; and

4. the most probable core structure segment alignments across the entire library, which maximize $P(C, i, t_i|\mathbf{a}, n)$, and which may potentially allow for the construction of a structural model for a sequence whose core structure is not in the library by selecting piecewise the most probable segment alignments from different core structures.

Item 1 above corresponds to requirement (iii) in section 1.1, item 2 corresponds to requirement (iv), item 3 corresponds to (iii) and (iv) simultaneously, and item 4 corresponds to (iii) and (iv) for individual core segments.

Section 4, below, works a simple example in subsections that parallel the analysis here. The reader is encouraged to read sections 3 and 4 in parallel.

3.1 Selecting an Alignment Given a Core Structure

For fixed sequence and core structure, the task of sequence-structure alignment is to select an alignment of the sequence to the core structure (see figure 1). White et al. [White et al., 1994a] show that

$$P(\mathbf{a}|n, C, \mathbf{t}) = \frac{\exp(-f(\mathbf{a}, C, \mathbf{t}))}{Z_{\mathbf{a}}} \tag{2}$$

$$Z_{\mathbf{a}} = \sum_{\mathbf{b}\in A^n} \exp(-f(\mathbf{b}, C, \mathbf{t})) \tag{3}$$

$$P(\mathbf{t}|\mathbf{a}, n, C) = P(\mathbf{a}|n, C, \mathbf{t})\frac{P(\mathbf{t}|n, C)}{P(\mathbf{a}|n, C)} \tag{4}$$

$$= \frac{P(\mathbf{t}|n, C)}{P(\mathbf{a}|n, C)}\frac{\exp(-f(\mathbf{a}, C, \mathbf{t}))}{Z_{\mathbf{a}}} \tag{5}$$

where the subscript $\mathbf{a}$ in $Z_{\mathbf{a}}$ indicates the summation and the parameters n, C, and $\mathbf{t}$ are inferred from context or postfixed. Note that $P(\mathbf{a}|n, C)$ is constant for fixed $\mathbf{a}$ and C and may be ignored for this task.

When f is modeled as a Markov Random Field, as in White et al. [White et al., 1994a] and Stultz et al. [Stultz et al., 1995], $Z_{\mathbf{a}}$ is the same for every $\mathbf{t} \in \mathcal{T}[C, n]$. This case is treated in the body of the paper. In this case, assuming uninformative priors, the globally most probable alignment has the globally lowest alignment score.

3.1.1 Variable $Z_{\mathbf{a}}$

Appendix C generalizes the equations to the case where $Z_{\mathbf{a}}$ is allowed to vary with $\mathbf{t}$. In this case, even assuming uninformative priors, the globally most probable alignment may not have the globally lowest alignment score. The variability of $Z_{\mathbf{a}}$ also must be accounted for, as the simple example in section 4.1.1 below shows.

One example of this is the common use of an alignment method based on sequence-independent gap penalties, e.g., a gap penalty that is linear or affine in the length of an insertion or deletion but independent of the sequence. In this case some threadings may delete portions of the sequence or the structure. In practical terms, this causes the global sum in equation 3 to be over different effective structures and different effective sequence lengths. In probabilistic terms, the usual linear or affine gap penalty forces loops to become exponentially unlikely in the length of the insertion or deletion. See Flöckner et al. [Flöckner et al., 1995] or Maiorov and Crippen [Maiorov & Crippen, 1994] for cogent criticism of allowing parts of the sequence or structure to "vanish" in this way; see Benner et al. [Benner et al., 1993] for empirical data showing that an exponential distribution does not provide an adequate fit to observed gap lengths; and see Lemer et al. [Lemer et al., 1995] for a discussion of inappropriate gap penalties, leading to gaps that are obviously far too small, as one aspect of threading algorithms that contributes to error.

3.2 Selecting a Core Structure

For a fixed sequence, the task of fold recognition is to select a core structure from the structure library (see figure 2). There is general agreement that one would like to select the core that has the highest conditional probability given the sequence.

$$P(C|\mathbf{a},n) = \sum_{\mathbf{x}\in\mathcal{T}[C,n]} \frac{P(\mathbf{a},n,C,\mathbf{x})}{P(\mathbf{a},n)} \tag{6}$$

$$= \sum_{\mathbf{x}\in\mathcal{T}[C,n]} \frac{P(\mathbf{a}|n,C,\mathbf{x})P(\mathbf{x}|n,C)P(C|n)P(n)}{P(\mathbf{a}|n)P(n)} \tag{7}$$

$$= \frac{P(C|n)}{P(\mathbf{a}|n)} \sum_{\mathbf{x}\in\mathcal{T}[C,n]} \frac{\exp(-f(\mathbf{a},C,\mathbf{x}))P(\mathbf{x}|n,C)}{Z_{\mathbf{a}}} \tag{8}$$

$$= \frac{P(C|n)}{P(\mathbf{a}|n)} \frac{\mu_{\mathbf{t}}}{Z_{\mathbf{a}}} \tag{9}$$

$$\mu_{\mathbf{t}} = \sum_{\mathbf{x}\in\mathcal{T}[C,n]} \exp(-f(\mathbf{a},C,\mathbf{x}))P(\mathbf{x}|n,C) \tag{10}$$

$$\mu_{\mathbf{a}} = \sum_{\mathbf{b}\in A^n} \exp(-f(\mathbf{b},C,\mathbf{t}))P(\mathbf{b}|n,C) \tag{11}$$

where $\mathcal{T}[C,n]$ is the set of all threadings of a sequence of length n onto C. Equation 9 orders all cores by conditional probability across the library.

Normalizing by $\sum_{C' \in \mathcal{L}} P(C'|\mathbf{a}, n)$ imposes the fundamental threading assumption that the proper core is indeed in the library, and converts the ordering into a conditional probability reflecting that assumption. Similar normalizations reflecting the fundamental threading assumption apply throughout. Probabilities shown generally are unnormalized.

The assumption of uninformative priors implies that $P(C|n)$ is constant in equation 9 and $P(\mathbf{b}|n, C)$ is constant in equation 11. In this case $\mu_{\mathbf{a}} = P(\mathbf{a}|n)Z_{\mathbf{a}}$ and so the most probable core structure maximizes the ratio $\mu_{\mathbf{t}}/\mu_{\mathbf{a}}$. This ratio is the mean probability across all possible alignments holding the sequence fixed, divided by the mean probability across all possible sequences holding the alignment fixed.

3.3 Selecting Core Structure and Alignment Jointly

The central problem of inverse structure prediction is to select simultaneously both a core structure and an alignment, given a sequence (see figure 3). To predict accurately, both the core structure and the alignment must be selected correctly. However, it is evident by comparing equations 5 and 9 to equations 12 and 14 that the most probable joint core structure and alignment is not necessarily the most probable alignment of the most probable core structure, but rather maximizes the product of core and alignment probabilities.

$$P(C, \mathbf{t}|\mathbf{a}, n) = P(\mathbf{t}|\mathbf{a}, n, C)P(C|\mathbf{a}, n) \tag{12}$$

$$= \frac{P(\mathbf{a}|n, C, \mathbf{t})P(\mathbf{t}|n, C)}{P(\mathbf{a}|n, C)} \frac{P(\mathbf{a}|n, C)P(C|n)}{P(\mathbf{a}|n)} \tag{13}$$

$$= \frac{P(C|n)}{P(\mathbf{a}|n)} \frac{P(\mathbf{t}|n, C) \exp(-f(\mathbf{a}, C, \mathbf{t}))}{Z_{\mathbf{a}}} \tag{14}$$

This orders all $\langle$structure, alignment$\rangle$ pairs by conditional probability jointly across the entire structure library.

3.4 Selecting Individual Core Segment Alignments

By selecting alignments to the most probable segments across the entire library, it might in principle be possible to construct a new core structure piecewise out of the selected segments even though the constructed core structure does not yet appear in the library. In this way it might in principle be possible to work around a current limitation of protein threading, namely that only known core structures may be predicted.

$$P(C, i, t_i|\mathbf{a}, n) = \sum_{\{\mathbf{x} \in \mathcal{T}[C,n] | x_i = t_i\}} P(C, \mathbf{x}|\mathbf{a}, n) \tag{15}$$

$$= \frac{P(C|n)}{P(\mathbf{a}|n)} \frac{\mu_{i,t_i}}{Z_{\mathbf{a}}} \tag{16}$$

$$\mu_{i,t_i} = \sum_{\{\mathbf{x}\in\mathcal{T}[C,n]|x_i=t_i\}} \exp(-f(\mathbf{a},C,\mathbf{x}))P(\mathbf{x}|n,C) \tag{17}$$

where $\{a|b\}$ is the set of a such that b. This orders all $\langle$structure, segment number, sequence index$\rangle$ triples by conditional probability across the entire library. The triples so generated (a) potentially arise from multiple different cores in the library, (b) have no overlap or ordering constraints, and (c) are selected from the set of legal threadings for each core, and so reflect its mean-field intra-model preferences and constraints. The problem of actually assembling such triples into a novel "meta-core" is left open.

3.4.1 Super-Secondary Structures, or Core Structure Subsets

In many cases a core structure may fit only partially to a core structure. Some secondary structure segments may correspond, while others may not. This might be the case, for example, when a common super-secondary structure motif is shared but the rest of the protein diverges; or when part of the core superposes but another part does not. Suppose that k of the m segments correspond, that the corresponding segments are $I = \{i_1, i_2, \ldots, i_k\}$, and that the corresponding indices are $T = \{t_{i_1}, t_{i_2}, \ldots, t_{i_k}\}$. The previous section gave the special case when $k = 1$.

$$P(C,I,T|\mathbf{a},n) = \sum_{\{\mathbf{x}\in\mathcal{T}[C,n]|j\in I\Rightarrow x_j=t_j\}} P(C,\mathbf{x}|\mathbf{a}) \tag{18}$$

$$= \frac{P(C|n)}{P(\mathbf{a}|n)}\frac{\mu_{I,T}}{Z_{\mathbf{a}}} \tag{19}$$

$$\mu_{I,T} = \sum_{\{\mathbf{x}\in\mathcal{T}[C,n]|j\in I\Rightarrow x_j=t_j\}} \exp(-f(\mathbf{a},C,\mathbf{x}))P(\mathbf{x}|n,C) \tag{20}$$

This orders, by conditional probability across the entire library, all super-secondary structures or core structure subsets that consist of k segments all taken from the same core structure.

3.4.2 Secondary Structure Prediction

Let helix(j) denote the event that the j^{th} sequence residue $\mathbf{a}[j]$ is found in a helical conformation, and let $\Phi_{helix}(\mathbf{a},j,C,i) = \{t_i|\ t_i$ places C_i over $\mathbf{a}[j]$ and C_i is helix$\}$. Then

$$P(\text{helix}(j)|\mathbf{a},n)$$

$$= \sum_{C\in\mathcal{L}} \sum_{\mathbf{x}\in\mathcal{T}[C,n]} P(\text{helix}(j)|\mathbf{a},n,C,\mathbf{x})P(C,\mathbf{x}|\mathbf{a},n,C) \tag{21}$$

$$= \sum_{C\in\mathcal{L}} \sum_{i=1}^{|C|} \sum_{t_i\in\Phi_{helix}(\mathbf{a},j,C,i)} \sum_{\{\mathbf{x}\in\mathcal{T}[C,n]|x_i=t_i\}} P(C,\mathbf{x}|\mathbf{a},n,C) \tag{22}$$

$$= \sum_{C\in\mathcal{L}} \sum_{i=1}^{|C|} \sum_{t_i\in\Phi_{helix}(\mathbf{a},j,C,i)} P(C,i,t_i|\mathbf{a},n) \tag{23}$$

The equation follows because for physical reasons $\mathbf{a}[j]$ cannot simultaneously be in two different helices or two different positions of the same helix. $P(\text{extended}(j)|\mathbf{a}, n)$ is defined similarly (extended = β-sheet). For 3-state prediction, coil is defined as anything that is not helix or extended. Let $\Phi_{coil}(\mathbf{a}, j, C, i) = \{\langle t_i, t_{i+1}\rangle |\ t_i$ places C_i before $\mathbf{a}[j]$ and t_{i+1} places C_{i+1} after it$\}$, where by convention $t_0 =$ the beginning and $t_{m+1} =$ the end of the sequence accounts for the leader and trailer loop regions. Then

$$P(\text{coil}(j)|\mathbf{a}, n) = \sum_{C \in \mathcal{L}} \sum_{i=0}^{|C|} \sum_{\langle t_i, t_{i+1}\rangle \in \Phi_{coil}(\mathbf{a}, j, C, i)} P(C, \{i, i+1\}, \{t_i, t_{i+1}\} |\mathbf{a}, n) \tag{24}$$

The terms are given by equations 9, 16, and 19 with $k = 2$, adjusted for boundary cases at sequence endpoints. As elsewhere, the values correspond to unnormalized probabilities.

3.5 Prior Probabilities

Three prior probabilities are necessary: $P(\mathbf{a}|n)$, $P(C|n)$, and $P(\mathbf{t}|n, C)$. $P(\mathbf{a}|n)$ is constant for a given sequence and may be ignored. $P(C|n)$ corresponds to the sequence-independent part of the core structure probability. It reflects at least two influences: the relative frequencies of different core structures, and the way these shift with sequence length. $P(\mathbf{t}|n, C)$ corresponds to the sequence-independent part of the loop probability. It reflects the loop length probability distribution for C and n, independent of the specific amino acid residue types that actually occupy the loops.

Assuming uninformative priors and fixed n implies that $P(C|n) = |\mathcal{L}|^{-1}$ and $P(\mathbf{t}|n, C) = |\mathcal{T}[C, n]|^{-1}$. There are a number of plausible biological reasons why the assumption of uninformative priors might be relaxed. For example, $P(C|n)$ might instead reflect the observation that some folds are more probable than others [Orengo et al., 1994], [Holm & Sander, 1994], [Murzin et al., 1995]; or that fold-space attractors have unequal population densities [Holm & Sander, 1996]; or that proteins are roughly half secondary structure and half coil; or that longer sequences are more likely to fold into larger structures. $P(\mathbf{t}|n, C)$ might instead reflect an empirical loop length distribution [Benner et al., 1993] constructed by tabulating the loop lengths observed to connect loop endpoints in various geometries across a structural database; or a linear or affine gap penalty, in which case it becomes exponentially improbable in the length of insertions and deletions. This is not to argue for or against any particular set of priors. Rather, different informative priors might be plausible under particular assumptions.

3.6 Global Sums

Only four global sums or means are sufficient to accomplish all of the probabilistic selections described above: $Z_{\mathbf{a}}$ of equation 3, $\mu_{\mathbf{t}}$ of equation 10, μ_{i,t_i} of equation 17, and $\mu_{I,T}$ of equation 20. Computing the global sums is NP-hard if specific pair interactions and

gaps both are permitted, and consequently either approximations or long computation must be employed. Appendix A provides brief proof sketches of NP-hardness for the general pair interaction case. Appendix B provides fast exact recursions for the case of no pair interactions.

4 Example — Selection Criteria

This section works a simple example showing the major selection criteria of section 3. The subsections here parallel the subsections in section 3. The reader is reminded that this example was contrived to be as simple as possible while illustrating the points in the paper compactly and transparently, and is not intended to shed light on protein structure.

The example uses the HP model [Dill et al., 1995] in which there are only two amino acid types, H (hydrophobic) and P (polar). Figure 4 shows the components of the example.

Figure 4 about here.

Figure 4a shows the core library $\mathcal{L}$, together with the cores' native sequences for reference. Library member $\mathcal{L}_1$ has two interacting core segments. Library member $\mathcal{L}_2$ has three; C_1 and C_3 both interact with C_2, but not with each other. All core segments are only one core element long, hence correspond to one sequence residue.

Figure 4b shows the three-residue sequence $\mathbf{a} = HHP$ to be threaded. It is not identical to either core's native sequence.

Suppose any threading that fills all core segments is legal, i.e., the minimum loop lengths are all zero. Figure 4c–e shows all three legal threadings of $\mathcal{L}_1$, labeled τ_1–τ_3, and figure 4f shows the only legal threading of $\mathcal{L}_2$, labeled τ_4.

For simplicity, let the singleton scores all be zero. Let the pairwise scores f_p be

$$f_p(H,H) = f_p(P,P) = 0 \tag{25}$$
$$f_p(H,P) = f_p(P,H) = 1 \tag{26}$$

let the loop scores f_l be

$$f_l(P) = 0 \tag{27}$$
$$f_l(H) = 1 \tag{28}$$

let the total score be $f = f_p + f_l$, and let $h(x) = \exp(-f(x))$. This simple score function says that hydrophobe-hydrophobe or polar-polar interactions are more favorable

than hydrophobe-polar, and that a polar residue is more favorable in a loop than is a hydrophobe. The native sequences of $\mathcal{L}_1$ and $\mathcal{L}_2$ (see Figure 4a) both score zero under this score function.

Uninformative priors are assumed throughout the example. For any feasible $\mathbf{a}$, C, or $\mathbf{t}$:

$$P(\mathbf{a}|n=3) = 1/8 \quad (29)$$
$$P(C|n=3) = 1/2 \quad (30)$$
$$P(\mathbf{t}|n=3, C=\mathcal{L}_1) = 1/3 \quad (31)$$
$$P(\mathbf{t}|n=3, C=\mathcal{L}_2) = 1 \quad (32)$$

4.1 Example — Select Alignment Given Core Structure

Assume as in section 3.1 that the sequence and core have been fixed in advance and that the task is to align the sequence $\mathbf{a} = HHP$ of Figure 4b to the two-segment model $\mathcal{L}_1$ of Figure 4a. The three legal threadings are τ_1 through τ_3 of Figure 4c-e.

The scores of the three threadings of HHP onto $\mathcal{L}_1$ are

$$f(HHP, \mathcal{L}_1, \tau_1) = f_p(H,P) + f_l(H) = 1 + 1 = 2 \Rightarrow h(HHP, \mathcal{L}_1, \tau_1) \approx 0.135 \quad (33)$$
$$f(HHP, \mathcal{L}_1, \tau_2) = f_p(H,H) + f_l(P) = 0 + 0 = 0 \Rightarrow h(HHP, \mathcal{L}_1, \tau_2) = 1 \quad (34)$$
$$f(HHP, \mathcal{L}_1, \tau_3) = f_p(H,P) + f_l(H) = 1 + 1 = 2 \Rightarrow h(HHP, \mathcal{L}_1, \tau_3) \approx 0.135 \quad (35)$$

To compute $Z_{\mathbf{a}}$ for these three threadings, sum $h(\mathbf{b}, \mathcal{L}_1, \tau_i) = \exp(-f(\mathbf{b}, \mathcal{L}_1, \tau_i))$ over all sequences $\mathbf{b}$ of length 3 over the alphabet $\{H, P\}$. For τ_1,

term	f_p	$+ f_l$	$= f$	$\approx h$
$f(HHH, \mathcal{L}_1, \tau_1)$	0	1	1	0.368
$f(HHP, \mathcal{L}_1, \tau_1)$	1	1	2	0.135
$f(HPH, \mathcal{L}_1, \tau_1)$	0	0	0	1.000
$f(HPP, \mathcal{L}_1, \tau_1)$	1	0	1	0.368
$f(PHH, \mathcal{L}_1, \tau_1)$	1	1	2	0.135
$f(PHP, \mathcal{L}_1, \tau_1)$	0	1	1	0.368
$f(PPH, \mathcal{L}_1, \tau_1)$	1	0	1	0.368
$f(PPP, \mathcal{L}_1, \tau_1)$	0	0	0	1.00
$Z_{\mathbf{a}}[\mathcal{L}_1, \tau_1] = \sum h \approx$				3.742

It is easy to see that $Z_{\mathbf{a}}[\mathcal{L}_1, \tau_2] = Z_{\mathbf{a}}[\mathcal{L}_1, \tau_3] = Z_{\mathbf{a}}[\mathcal{L}_1, \tau_1]$ because permuting the alignment only permutes the rows of the table, i.e., $Z_{\mathbf{a}}$ is constant for every threading of $\mathcal{L}_1$.

Finally, the unnormalized probabilities are

$$P(\tau_1|\mathbf{a}=HHP, n=3, C=\mathcal{L}_1) = \left(\frac{P(\tau_1|n,\mathcal{L}_1)}{P(\mathbf{a}|n,\mathcal{L}_1)}\right)\left(\frac{\exp(-f(\mathbf{a},\mathcal{L}_1,\tau_1))}{Z_{\mathbf{a}}}\right) \quad (36)$$

$$\approx (0.333/0.125)\,(0.135/3.742) \approx 0.096 \quad (37)$$

$$P(\tau_2|\mathbf{a}=HHP, n=3, C=\mathcal{L}_1) \approx (0.333/0.125)\,(1/3.742) \approx 0.71 \quad (38)$$

$$P(\tau_3|\mathbf{a}=HHP, n=3, C=\mathcal{L}_1) \approx (0.333/0.125)\,(0.135/3.742) \approx 0.096 \quad (39)$$

Consequently the threading of lowest score, τ_2, is also the threading of highest conditional probability, τ_2. This is always the case when $Z_{\mathbf{a}}$ is constant across different threadings of the same core.

4.1.1 Example — Variable $Z_{\mathbf{a}}$.

If $Z_{\mathbf{a}}$ is not constant for each threading of a given core and sequence, then the threading of global maximum probability may not be the threading of global minimum score. In this subsection a superscript "*" indicates this.

Consider the case where loop scores are a sequence-independent gap penalty that is linear in the length of an insertion or deletion but independent of the sequence. Let the score $w(k)$ for an insertion or deletion of length k be $w(k) = ck$ where c is an arbitrary constant. Then threading τ_1 has no insertions or deletions and so has a loop score of 0, while τ_2 and τ_3 each have one interior deletion and one leader or trailer insertion of one residue each (two residues total) and so have loop scores of $2c$.

Then the scores of the three threadings are

$$f^*(\tau_1) = f_p(H,P) = 1 \quad (40)$$

$$f^*(\tau_2) = f_p(H,H) + w(1) + w(1) = 0 + 2c \quad (41)$$

$$f^*(\tau_3) = f_p(H,P) + w(1) + w(1) = 1 + 2c \quad (42)$$

Clearly, τ_2 will always score lower than τ_3, but by adjusting the arbitrary constant c it can be made to score higher or lower than τ_1.

The following table shows the calculation of $Z^*_{\mathbf{a}}$ for τ_2,

term	$f_p^* + w$		$= f^*$	$\approx h^*$
$f(HHH, \mathcal{L}_1, \tau_2)$	0	$2c$	$2c$	$\exp(-2c)$
$f(HHP, \mathcal{L}_1, \tau_2)$	0	$2c$	$2c$	$\exp(-2c)$
$f(HPH, \mathcal{L}_1, \tau_2)$	1	$2c$	$1 + 2c$	$0.368\exp(-2c)$
$f(HPP, \mathcal{L}_1, \tau_2)$	1	$2c$	$1 + 2c$	$0.368\exp(-2c)$
$f(PHH, \mathcal{L}_1, \tau_2)$	1	$2c$	$1 + 2c$	$0.368\exp(-2c)$
$f(PHP, \mathcal{L}_1, \tau_2)$	1	$2c$	$1 + 2c$	$0.368\exp(-2c)$
$f(PPH, \mathcal{L}_1, \tau_2)$	0	$2c$	$2c$	$\exp(-2c)$
$f(PPP, \mathcal{L}_1, \tau_2)$	0	$2c$	$2c$	$\exp(-2c)$
$Z^*_{\mathbf{a}}[\mathcal{L}_1, \tau_2] = \sum h^* \approx$				$5.472\exp(-2c)$

It is easy to see that $Z^*_{\mathbf{a}}[\mathcal{L}_1, \tau_1] \approx 5.472$ and that $Z^*_{\mathbf{a}}[\mathcal{L}_1, \tau_3] = Z^*_{\mathbf{a}}[\mathcal{L}_1, \tau_2] \approx 5.472 \exp(-2c)$.

Finally, the unnormalized probabilities are

$$P(\tau_1|\mathbf{a}, n, C) \approx (0.333/0.125)\,(0.368/5.472) \approx 0.179 \quad (43)$$

$$P(\tau_2|\mathbf{a}, n, C) \approx (0.333/0.125)\,(\exp(-2c)/5.472\exp(-2c)) \approx 0.487 \quad (44)$$

$$P(\tau_3|\mathbf{a}, n, C) \approx (0.333/0.125)\,(0.368\exp(-2c)/5.472\exp(-2c)) \approx 0.179 \quad (45)$$

Threading τ_2 still has the highest conditional probability, but for some choices of the arbitrary constant c threading τ_1 has the lowest score. Consequently, the threading of lowest score is not necessarily the threading of highest conditional probability.

4.2 Example — Selecting a Core Structure

Next turn to selecting the most probable core model, as in section 3.2. The task is to choose between $\mathcal{L}_1$ and $\mathcal{L}_2$, with f_l as in equations 27–28, $f = f_p + f_l$ for $\mathcal{L}_1$ as in section 4.1, and $f = f_p^{1,2} + f_p^{2,3}$ for $\mathcal{L}_2$ where $f_p^{i,j}$ denotes f_p between C_i and C_j.

The threading score of $\mathbf{a}$ on $\mathcal{L}_2$ under its only threading τ_4 is

$$f(HHP, \mathcal{L}_2, \tau_4) = f_p^{1,2}(H, H) + f_p^{2,3}(H, P) = 0 + 1 = 1 \quad (46)$$

$$h(HHP, \mathcal{L}_2, \tau_4) = \exp(-1) \approx 0.368 \quad (47)$$

The following table shows the calculation of $Z_{\mathbf{a}}[\mathcal{L}_2]$,

term	$f_p^{1,2} + f_p^{2,3}$	$= f$	$\approx h$
$f(HHH, \mathcal{L}_2, \tau_4)$	0	0	1.000
$f(HHP, \mathcal{L}_2, \tau_4)$	0	1	0.368
$f(HPH, \mathcal{L}_2, \tau_4)$	1	2	0.135
$f(HPP, \mathcal{L}_2, \tau_4)$	1	1	0.368
$f(PHH, \mathcal{L}_2, \tau_4)$	1	1	0.368
$f(PHP, \mathcal{L}_2, \tau_4)$	1	2	0.135
$f(PPH, \mathcal{L}_2, \tau_4)$	0	1	0.368
$f(PPP, \mathcal{L}_2, \tau_4)$	0	0	1.000
		$Z_{\mathbf{a}}[\mathcal{L}_2] = \sum h \approx$	3.742

The numerical equality of $Z_{\mathbf{a}}[\mathcal{L}_1]$ and $Z_{\mathbf{a}}[\mathcal{L}_2]$ is accidental.

Next it is necessary to compute $Z_{\mathbf{t}}[\mathcal{L}_1]$ and $Z_{\mathbf{t}}[\mathcal{L}_2]$ by summing h over all threadings.

$$\begin{aligned} Z_{\mathbf{t}}[\mathcal{L}_1] &= h(HHP, \mathcal{L}_1, \tau_1) + h(HHP, \mathcal{L}_1, \tau_2) + h(HHP, \mathcal{L}_1, \tau_3) && (48) \\ &= \exp(-2) + \exp(0) + \exp(-2) \approx 1.271 && (49) \\ Z_{\mathbf{t}}[\mathcal{L}_2] &= h(HHP, \mathcal{L}_2, \tau_4) \approx 0.368 && (50) \end{aligned}$$

Assuming uninformative priors, $\mu_{\mathbf{t}}[\mathcal{L}_1] = Z_{\mathbf{t}}[\mathcal{L}_1]/3 \approx 0.424$ and $\mu_{\mathbf{t}}[\mathcal{L}_2] = Z_{\mathbf{t}}[\mathcal{L}_2]/1 \approx 0.368$.

Finally, the unnormalized probabilities are

$$\begin{aligned} P(\mathcal{L}_1 | n = 3, \mathbf{a} = HHP) &= \left(\frac{P(\mathcal{L}_1|n)}{P(\mathbf{a}|n)}\right)\left(\frac{\mu_{\mathbf{t}}[\mathcal{L}_1]}{Z_{\mathbf{a}}[\mathcal{L}_1]}\right) && (51) \\ &\approx (0.5/0.125)\,(0.424/3.742) \approx 0.453 && (52) \\ P(\mathcal{L}_2 | n = 3, \mathbf{a} = HHP) &\approx (0.5/0.125)\,(0.368/3.742) \approx 0.393 && (53) \end{aligned}$$

Thus $\mathcal{L}_1$ is the most probable core model for HHP, but $\mathcal{L}_2$ is a plausible alternative (the normalized probabilities are 0.54 and 0.46 respectively).

4.2.1 Example — Primary Sequence Similarity

In contrast, note that selection by primary sequence similarity to the cores' native sequences (Figure 4a) would prefer $\mathcal{L}_2$ over $\mathcal{L}_1$. This is because the ungapped alignment of $\mathbf{a} = HHP$ to the native sequence of core $\mathcal{L}_2$, HHH, has two sequence identities and only one mismatch. However, the native sequence of core $\mathcal{L}_1$, HPH, can achieve two sequence identities only by introducing alignment gaps. Consequently the primary sequence alignment of $\mathbf{a} = HHP$ to HHH would score better than the alignment to HPH, and so $\mathcal{L}_2$ would be preferred over $\mathcal{L}_1$.

4.3 Example — Selecting Structure and Alignment Jointly

Next turn to selecting the most probable joint core model and alignment, as in section 3.3. All of the necessary constants already have been calculated.

$$P(\mathcal{L}_1, \tau_1 | \mathbf{a}, n) = \left(\frac{P(\mathcal{L}_1|n)}{P(\mathbf{a}|n)}\right)\left(\frac{P(\tau_1|n,\mathcal{L}_1)\exp(-f(\mathbf{a},\mathcal{L}_1,\tau_1))}{Z_{\mathbf{a}}}\right) \quad (54)$$

$$\approx (0.5/0.125)\,(0.333 \times 0.135/3.742) \approx 0.048 \quad (55)$$

$$P(\mathcal{L}_1, \tau_2 | \mathbf{a}, n) \approx (0.5/0.125)\,(0.333 \times 1.0/3.742) \approx 0.356 \quad (56)$$

$$P(\mathcal{L}_1, \tau_3 | \mathbf{a}, n) \approx (0.5/0.125)\,(0.333 \times 0.135/3.742) \approx 0.048 \quad (57)$$

$$P(\mathcal{L}_2, \tau_4 | \mathbf{a}, n) \approx (0.5/0.125)\,(1.0 \times 0.368/3.742) \approx 0.393 \quad (58)$$

Thus $\langle \mathcal{L}_2, \tau_4 \rangle$ is the most probable joint core and alignment for HHP, but $\langle \mathcal{L}_1, \tau_2 \rangle$ is a plausible alternative (the normalized probabilities are 0.47 and 0.42 respectively). On the other hand, neither $\langle \mathcal{L}_1, \tau_1 \rangle$ nor $\langle \mathcal{L}_1, \tau_3 \rangle$ are rated highly (their normalized probabilities both are 0.057). This is plausible, as $\mathcal{L}_1$ and $\mathcal{L}_2$ are both plausible core models for HHP (shown in section 4.2) and τ_2 and τ_4 are respectively their best threadings (shown in section 4.1).

The reason that $\mathcal{L}_1$ is the most probable core, but not the core of the most probable joint core and alignment, is because the threadings of $\mathcal{L}_1$ have more uncertainty than do the threadings of $\mathcal{L}_2$. Indeed, $\mathcal{L}_2$ has no threading uncertainty. Uncertainty in core selection and in alignment selection both contribute to threading error. This effect is commonly noted in protein structure prediction. For example, uncertainty in the identification of a fold as a β-sandwich differs from uncertainty in the strand alignments.

4.4 Example — Selecting Individual Core Segment Alignments

Next turn to selecting individual core segment alignments, as in section 3.4. This part of the example is vacuous for $\mathcal{L}_2$ as there is only one threading. Secondary and supersecondary structure, also discussed in section 3.4, are omitted because they are not meaningful in this simple example.

First the global sums corresponding to μ_{i,t_i} must be computed. For example, $\mu_{i,t_i}[\mathcal{L}_1, C_1, 1]$ will sum h over τ_1 and τ_2, the two threadings that place C_1 of $\mathcal{L}_1$ at $\mathbf{a}[1]$. Assuming uninformative priors,

$$\mu_{i,t_i}[\mathcal{L}_1, C_1, 1] = (h(HHP, \mathcal{L}_1, \tau_1) + h(HHP, \mathcal{L}_1, \tau_2))/3 \quad (59)$$

$$= (\exp(-2) + \exp(0))/3 \approx 0.378 \quad (60)$$

$$\mu_{i,t_i}[\mathcal{L}_1, C_1, 2] = \exp(-2)/3 \approx 0.0451 \quad (61)$$

$$\mu_{i,t_i}[\mathcal{L}_1, C_2, 2] = \exp(0)/3 \approx 0.333 \quad (62)$$

$$\mu_{i,t_i}[\mathcal{L}_1, C_2, 3] = (\exp(-2) + \exp(-2))/3 \approx 0.0902 \quad (63)$$

$$\mu_{i,t_i}[\mathcal{L}_2, C_1, 1] \;=\; \mu_{i,t_i}[\mathcal{L}_2, C_2, 2] = \mu_{i,t_i}[\mathcal{L}_2, C_3, 3] = \exp(-1)/1 \approx 0.368 \qquad (64)$$

The unnormalized probabilities are

$$\begin{aligned}
P(\mathcal{L}_1, C_1, 1|\mathbf{a}, n) &= \left(\frac{P(\mathcal{L}_1|n)}{P(\mathbf{a}|n)}\right)\left(\frac{\mu_{i,t_i}[\mathcal{L}_1, C_1, 1]}{Z_{\mathbf{a}}[\mathcal{L}_1]}\right) && (65)\\
&\approx (0.5/0.125)\,(0.378/3.742) \approx 0.405 && (66)\\
P(\mathcal{L}_1, C_1, 2|\mathbf{a}, n) &\approx (0.5/0.125)\,(0.0451/3.742) \approx 0.0482 && (67)\\
P(\mathcal{L}_1, C_2, 2|\mathbf{a}, n) &\approx (0.5/0.125)\,(0.333/3.742) \approx 0.356 && (68)\\
P(\mathcal{L}_1, C_2, 3|\mathbf{a}, n) &\approx (0.5/0.125)\,(0.0902/3.742) \approx 0.0964 && (69)\\
P(\mathcal{L}_2, C_1, 1|\mathbf{a}, n) &= P(\mathcal{L}_2, C_2, 2|\mathbf{a}, n) = P(\mathcal{L}_2, C_3, 3|\mathbf{a}, n) && (70)\\
&\approx (0.5/0.125)\,(0.368/3.742) \approx 0.393 && (71)
\end{aligned}$$

Thus $\langle \mathcal{L}_1, C_1, 1\rangle$ is the most probable individual core segment alignment, followed by the $\mathcal{L}_2$ segments and $\langle \mathcal{L}_1, C_2, 2\rangle$. Neither $\langle \mathcal{L}_1, C_1, 2\rangle$ nor $\langle \mathcal{L}_1, C_2, 3\rangle$ are rated highly. The most probable individual core segment alignments are those that tend to participate in the most highly rated threadings.

5 Discussion

We have presented a probabilistic Bayes-optimal analysis which unifies the protein "threading" problems of fold recognition and sequence-structure alignment. The analysis is consistent with probability theory. The theory involved three specific prior probabilities that model background knowledge about protein structure, and four specific global sums. The global sum computations are NP-hard in the pair interaction case, and have fast exact recursions if pair interactions are omitted. Conclusions include: (a) the most probable joint core structure and alignment is not necessarily the most probable alignment of the most probable core structure, but rather maximizes the product of core and alignment probabilities; (b) use of a sequence-independent linear or affine gap penalty may result in the highest-probability threading not having the lowest score; (c) selecting the most probable core structure from the library (core structure selection or fold recognition only) involves comparing probabilities summed over all possible alignments of the sequence to the core, and not comparing individual optimal (or near-optimal) sequence-structure alignments; and (d) assuming uninformative priors, core structure selection is equivalent to comparing the ratio of two global means.

A long-range goal of this work is to integrate structural and functional pattern recognition. The reader will have noticed that the gapped block alignment method discussed here is conceptually similar to block patterns, consensus patterns, weight matrices, profile patterns, and hierarchical patterns, among many other gapped block pattern

methods (reviewed in [Smith et al., 1996]). Combined structural and functional pattern recognition is likely to prove more powerful than either alone.

The probabilistic Bayesian view of protein structure prediction presented here complements other views by which protein structure prediction has been understood, such as Boltzmann's principle [Sippl, 1993], [Finkelstein et al., 1995], [Wilbur et al., 1996], simplified lattice representations [Dill et al., 1995], and spin-glasses [Friedrichs & Wolynes, 1989], [Goldstein et al., 1992]. The underlying problem of predicting protein structure from sequence has proven extremely difficult, and each different perspective has advanced our understanding. Strengths of the Bayesian analysis include a large body of well-understood mathematics; an explicit provision for the use of prior knowledge about protein structure; the ability to ask very precise probabilistic questions and derive rigorous formulae that precisely answer them; and a guarantee that the resulting conclusions are globally correct with respect to the axioms of probability.

Acknowledgments

Ilya Muchnik helped develop the probabilistic framework within which this work is situated. Ljubomir Buturović and Raman Nambudripad contributed greatly to the implementation of the MRF score function, and together with Jadwiga Bienkowska, Chrysanthe Gaitatzes, Loredana Lo Conte, Lisa Tucker-Kellog, and Sophia Zarakhovich contributed greatly to core definition and construction. Comments from Janice Glasgow, Nick Steffen, Jadwiga Bienkowska, and a blind reviewer greatly improved the presentation. Thanks to Barbara Bryant, Tomás Lozano-Pérez and Patrick Winston for discussions of computational protein folding. Special thanks to all crystallographers who deposited their coordinates in the international scientific databases.

This paper describes research performed at TASC; at the Information and Computer Science Department of the University of California, Irvine, sponsored by the National Science Foundation under grant IRI-9624739; and at the Artificial Intelligence Laboratory of the Massachusetts Institute of Technology in consortium with the BioMolecular Engineering Research Center of Boston University, sponsored by the National Science Foundation under grant DIR-9121548. Support for the Information and Computer Science Department's research is provided in part by the Department of Education under grant GAANN-P200A50166. Support for the Artificial Intelligence Laboratory's research is provided in part by the Advanced Research Projects Agency of the Department of Defense under Office of Naval Research contract N00014-91-J-4038. Support for the BioMolecular Engineering Research Center's research is provided in part by the National Institutes of Health under grant RR02275-05.

References

[Akutsu & Miyano, 1997] Akutsu, T., Miyano, S. On the approximation of protein threading. pp. 3–8 in *Proc. Intl. Conf. on Computational Molecular Biology,* (ed. Istrail, S., Karp, R., Lengauer, T., Pevzner, P., Shamir, R., Waterman, M.), ACM Press, New York, 1997.

[Akutsu & Tashimo, 1998] Akutsu, T., Tashimo, H. Linear programming based approach to the derivation of a contact potential for protein threading. pp. 413–424 in *Proc. Pacific Symposium on Biocomputing'98,* (ed. Altman, R.B., Dunker, A.K., Hunter, L., Klein, T.E.), World Scientific, Singapore, 1998.

[Arnold et al., 1992] Arnold, G.E., Dunker, A.K., Johns, S.J., Douthart, R.J. Use of conditional probabilities for determining relationships between amino acid sequence and protein secondary structure. Proteins: Structure, Function, and Genetics, 12:382–399, 1992.

[Bayes, 1763] Bayes, T. An essay toward solving a problem in the doctrine of chances. Philosophical Transactions of the Royal Society of London 53:370–418, 1764. Reprinted pp. 131–153 in "Studies in the History of Statistics and Probability," (ed. Pearson, E.S., Kendall, M.G.), Charles Griffin, London, 1970.

[Benner et al., 1993] Benner, S.A., Cohen, M.A., Gonnet, G.H. Empirical and structural models for insertions and deletions in the divergent evolution of proteins. J. Mol. Biol. 229:1065–1082, 1993.

[Bowie & Eisenberg, 1993] Bowie, J., Eisenberg, D. Inverted protein structure prediction. Current Opinion in Structural Biol. 3:437–444, 1993.

[Bowie et al., 1991] Bowie, F.U., Lüthy, R., Eisenberg, D. A method to identify protein sequences that fold into a known three-dimensional structure,. Science. 253:164–170, 1991.

[Box & Tiao, 1973] Box, G.E., Tiao, G.C. Bayesian inference in statistical analysis. Addison-Wesley, Reading, MA, 1973.

[Brooks et al., 1990] Brooks, C.L., Karplus, M., Pettitt, B.M. Proteins: A theoretical perspective of dynamics, structure, and thermodynamics. John Wiley and Sons, New York. 1990.

[Bryant & Altschul, 1995] Bryant, S.H., Altschul, S.F. Statistics of sequence-structure threading. Current Opinion in Structural Biol. 5:236–244, 1995.

[Bryant & Lawrence, 1993] Bryant, S.H., Lawrence, C.E. An empirical energy function for threading protein sequence through the folding motif. Proteins: Structure, Function, and Genetics 16:92–112, 1993.

[Crippen, 1996] Crippen, G.M. Failures of inverse folding and threading with gapped alignment. Proteins. 26:167-71, 1996.

[Desmet et al., 1992] Desmet, J., De Maeyer, M., Hazes, B., Lasters, I. The dead-end elimination theorem and its use in protein side-chain positioning. Nature (London) 356:539–542, 1992.

[Dill et al., 1995] Dill, K.A., Bromberg, S., Yue, K., Fiebig, K.M., Yee, D.P., Thomas, P.D., Chan, H.S. Principles of protein folding — a perspective from simple exact models. Protein Science. 4:561–602, 1995.

[Dunbrack & Cohen, 1997] Dunbrack, R.L., Jr., Cohen, F.E. Bayesian statistical analysis of protein sidechain rotamer preferences, Protein Science, 6:1661–1681, 1997.

[Fetrow & Bryant, 1993] Fetrow, J.S., Bryant, S.H. New programs for protein tertiary structure prediction. Bio/Technology 11:479–484, 1993.

[Finkelstein et al., 1995] Finkelstein, A.V., Badretdinov, A.Y., Gutin, A.M. Why do proteins have Boltzmann-like statistics? Proteins: Structure, Function, and Genetics 23:142–150, 1995.

[Finkelstein & Reva, 1991] Finkelstein, A.V. Reva, B. A search for the most stable folds of protein chains. Nature (London) 351:497–499, 1991.

[Flöckner et al., 1995] Flöckner, H., Braxenthaler, M., Lackner, P., Jaritz, M., Ortner, M., Sippl, M.J. Progress in fold recognition. Proteins: Structure, Function, and Genetics, 23:376–386, 1995.

[Fraenkel, 1993] Fraenkel, A.S. Complexity of protein folding. Bull. Math. Biol. 55(6):1199-1210, Nov. 1993.

[Friedrichs & Wolynes, 1989] Friedrichs, M.S., Wolynes, P.G. Toward protein tertiary structure recognition by means of associative memory Hamiltonians. Science, 246:371–373, 1989.

[Garey & Johnson, 1979] Garey, M.R., Johnson, D.S. *Computers and Intractability: A Guide to the Theory of NP-Completeness.* W.H. Freeman and Company, New York. 1979.

[Goldstein et al., 1992] Goldstein, R.A., Luthey-Schulten, Z.A., Wolynes, P.G. Tertiary structure recognition using optimized Hamiltonians with local interactions. Proc. Natl. Acad. Sci., USA 89:9029–9033, 1992.

[Greer, 1990] Greer, J. Comparative modeling methods: application to the family of the mammalian serine proteases. Proteins: Structure, Function, and Genetics 7:317–333, 1990.

[Hartigan, 1983] Hartigan, J.A. (1983) *Bayes Theory.* Springer-Verlag, New York.

[Holm & Sander, 1994] Holm, L., Sander, C. The FSSP database of structurally aligned protein fold families. Nucl. Acids Res. 22:3600–3609, 1994.

[Holm & Sander, 1996] Holm, L., Sander, C. Mapping the protein universe. Science 273:595–602, 1996.

[Hunter & States, 1992] Hunter, L., States, D.J. Bayesian classification of protein structure. IEEE Expert 7(4):67–75, 1992.

[Jernigan & Bahar, 1996] Jernigan, R.L., Bahar, I. Structure-derived potentials and protein simulations. Current Opinion in Structural Biol. 6:195–209, 1996.

[Jones et al., 1992] Jones, D.T., Taylor, W.R., Thornton, J.M. A new approach to protein fold recognition. Nature (London). 358:86–89, 1992.

[Jones & Thornton, 1993] Jones, D. T., Thornton, J. M. Protein fold recognition. J. Computer-Aided Mol. Design. 7:439–456, 1993.

[Jones & Thornton, 1996] Jones, D. T., Thornton, J. M. Potential energy functions for threading. Current Opinion in Structural Biol. 6:210–216, 1996.

[Kolinski et al., 1996] Kolinski, A., Skolnick, J., Godzik, A. An algorithm for prediction of structural elements in small proteins. pp. 446–460 in *Proc. Pacific Symposium on Biocomputing'96,* (ed. Hunter, L., Klein, T.E.), World Scientific, Singapore, 1996.

[Lathrop, 1994] Lathrop, R.H. The protein threading problem with sequence amino acid interaction preferences is NP-complete. Protein Engng. 7:1059–1068, 1994.

[Lathrop & Smith, 1996] Lathrop, R.H., Smith, T.F. Global optimum protein threading with gapped alignment and empirical pair score functions. J. Mol. Biol. 255:641–665, 1996.

[Lawrence et al., 1993] Lawrence, C.E., Altschul, S.F., Boguski, M.S., Liu, J.S., Neuwald, A.F., Wootton, J.C. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. Science, 262:208-214, 1993.

[Lemer et al., 1995] Lemer, C.M.-R., Rooman, M.J., Wodak, S.J. Protein structure prediction by threading methods: Evaluation of current techniques. Proteins: Structure, Function, and Genetics, 23:337–355, 1995.

[Lüthy et al., 1992] Lüthy, R., Bowie, J.U., Eisenberg, D. Assessment of Protein Models with Three-dimensional Profiles. Nature (London), 356:83–85, 1992.

[Madej et al., 1995] Madej, T., Gibrat, J.-F., Bryant, S.H. Threading a database of protein cores. Proteins: structure, Function, and Genetics, 23:356–369, 1995.

[Maiorov & Crippen, 1994] Maiorov, V.N., Crippen, G.M. Learning about protein folding via potential functions. Proteins: Structure, Function, and Genetics, 20:167–173, 1994.

[Mandal & Linthicum, 1993] Mandal, C., Linthicum, D. S. PROGEN: An automated modelling algorithm for the generation of complete protein structures from the α-carbon atomic coordinates. J. Computer-Aided Mol. Design. 7(2):199–224, 1993.

[Moult et al., 1995] Moult, J., Pedersen, J.T., Judson, R., Fidelis, K. A large-scale experiment to assess protein structure prediction methods. Proteins: Structure, Function, and Genetics, 23:ii–iv, 1995.

[Murzin et al., 1995] Murzin, A.G., Brener, S.E., Hubbard, T., Chothia, C. SCOP: A structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol. 247:536–540, 1995.

[Ngo & Marks, 1992] Ngo, J.T., Marks, J. Computational complexity of a problem in molecular structure prediction. Protein Engineering 5(4):313–321, 1992.

[Novotny et al., 1988] Novotný, J., Rashin, A.A., Bruccoleri, R.E. Criteria that discriminate between native proteins and incorrectly folded models. Proteins: Structure, Function, and Genetics 4:19–30, 1988.

[Orengo et al., 1994] Orengo, C.A., Jones, D.T., Thornton, J.M. Protein superfamilies and domain superfolds. Nature (London), 372:631–634, 1994.

[Ouzounis et al., 1993] Ouzounis, C., Sander, C., Scharf, M., Schneider, R. Prediction of protein structure by evaluation of sequence-structure fitness. J. Mol. Biol. 232:805–825, 1993.

[Rabiner, 1989] Rabiner, R.L. A tutorial on hidden Markov models and selected applications in speech recognition. Proc. of the IEEE, 77:257–285, 1989.

[Russell & Barton, 1994] Russell, R.B., Barton, G.J. Structural features can be unconserved in proteins with similar folds. J. Mol. Biol. 244:332–350, 1994.

[Sankof & Kruskal, 1983] Sankof, D., Kruskal, J.B., eds. Time warps, string edits and macromolecules. Addison-Wesley, Reading, MA, USA, 1983.

[Sippl et al., 1992] Sippl, M.J., Hendlich, M., Lackner, P. Assembly of polypeptide and protein backbone conformations from low energy ensembles of short fragments. Protein Sci. 1:625–640, 1992.

[Sippl, 1993] Sippl, M.J. Boltzmann's principle, knowledge-based mean fields and protein folding. J. Computer-Aided Mol. Design 7:473–501, 1993.

[Sippl, 1995] Sippl, M.J. Knowledge-based potentials for proteins. Current Opinion in Structural Biol. 5:229–235, 1995.

[Simons et al., 1997] Simons, K.T., Kooperberg, C., Huang, E., Baker, D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J. Mol. Biol. 268:209–225, 1997.

[Smith et al., 1996] Smith, T.F., Lathrop, R.H. and Cohen, F.E. The Identification of Protein Functional Patterns. pp. 29-61 in *Integrative Approaches to Molecular Biology,* eds. Collado-Vides, J., Magasanik, B., and Smith, T.F., MIT Press, Cambridge, MA, 1996.

[Smith et al., 1997a] Smith, T.F., Lo Conte, L., Bienkowska, J., Rogers Jr., R.G., Gaitatzes, C., Lathrop, R.H. The threading approach to the inverse folding problem. pp. 287–292 in *Proc. Intl. Conf. on Computational Molecular Biology,* (ed. Istrail, S., Karp, R., Lengauer, T., Pevzner, P., Shamir, R., Waterman, M.), ACM Press, New York, 1997.

[Smith et al., 1997b] Smith, T.F., Lo Conte, L., Bienkowska, J. Gaitatzes, C., Rogers Jr., R.G., Lathrop, R.H. Current limitations to protein threading approaches. J. Comp. Biol., 4:217–225, 1997.

[Skolnick et al., 1997] Skolnick, J., Kolinski, A., Ortiz, A.R. MONSSTER: A method for folding globular proteins with a small number of distance restraints. J. Mol. Biol., 265:217–241, 1997.

[Srinivasan & Rose, 1995] Srinivasan, R., Rose, G.D. LINUS: A hierarchic procedure to predict the fold of a protein. Proteins: Structure, Function, and Genetics, 22:81–99, 1995.

[Stultz et al., 1995] Stultz, C.M., Nambudripad, R., Lathrop, R.H., White, J.V. Predicting protein structure with probabilistic models. In *Protein Folding and Stability* (Allewell, N., Woodward, C., eds), JAI Press, Greenwich, in press.

[Thomas & Dill, 1996] Thomas, P.D., Dill, K.A. Statistical potentials extracted from protein structures: How accurate are they? J. Mol. Biol. 257:457–469, 1996.

[Thompson & Goldstein, 1996] Thompson, M.J., Goldstein, R.A. Predicting solvent accessibilities: Higher accuracy using Bayesian statistics and optimized residue substitution classes. Proteins: Structure, Function, and Genetics, 25:38–47, 1996.

[Unger & Moult, 1993] Unger, R., Moult, J. Finding the lowest free energy conformation of a protein is an NP-hard problem: Proof and implications. Bull. Math. Biol. 55(6):1183-1198, Nov. 1993.

[Weiner et al., 1984] Weiner, S.J., Kollman, P.A., Case, D.A., Singh, U.C., Ghio, C., Alagona, G., Profeta, S., Weiner, P. A new force field for molecular mechanical simulation of nucleic acids and proteins J. Am. Chem. Soc., 106:765-784, 1984.

[White et al., 1994a] White, J., Muchnik, I., Smith, T.F. Modeling protein cores with Markov random fields. Mathematical Biosciences 124:149–179, 1994.

[White et al., 1994b] White, J.V., Stultz, C.M., and Smith, T.F., Protein Classification by State-Space Modeling and Optimal Filtering of Amino-Acid Sequences. Mathematical Biosciences, 191:1, 35–75, 1994.

[Wilbur et al., 1996] Wilbur, W.J., Major, F., Spouge, J., Bryant, S. The statistics of unique native states for random peptides. Biopolymers 38:447-459, 1996.

[Wilmanns & Eisenberg, 1993] Wilmanns, M., Eisenberg, D. Three-dimensional profiles from residue-pair preferences: Identification of sequences with β/α-barrel fold. Proc. Natl. Acad. Sci., USA 90:1379–1383, 1993.

[Wodak & Rooman, 1993] Wodak, S. J., Rooman, M. J. Generating and testing protein folds. Current Opinion in Structural Biol. 3:247–259, 1993.

[Xu & Uberbacher, 1996] Xu, Y., Uberbacher, C.E. A polynomial-time algorithm for a class of protein threading problems. CABIOS. 12:511–517, 1996.

[Xu et al., 1998] Xu, Y., Xu, D., Uberbacher, C.E. A new method for modeling and solving the protein fold recognition problem. pp. 285–292 in *Proc. Intl. Conf. on Computational Molecular Biology,* (ed. Istrail, S., Karp, R., Lengauer, T., Pevzner, P., Shamir, R., Waterman, M.), ACM Press, New York, 1998.

[Zheng et al., 1993] Zheng, Q., Rosenfeld, R., Vajda, S., DeLisi, C. Determining protein loop conformation using scaling-relaxation techniques. Protein Sci. 2:1242–1248, 1993.

A Appendix — Pair Interactions

It is easy to show that computing any of the global sums is NP-hard if pair interactions are allowed. This follows by modifications to a previously published proof that computing the global optimum sequence-structure alignment is NP-hard under similar conditions [Lathrop, 1994]. This appendix assumes that the reader is familiar with that proof. Here we sketch only the proof modifications that would be made in each case; formal proofs are left to the reader.

A.1 Sketch of proof for $\mu_{\mathbf{t}}$, equation 10

Combine the edge score functions of a given threading using multiplication instead of addition, and change the edge score function so that threadings that score zero encode failures to the original ONE-IN-THREE-3SAT problem and threadings that score 1 encode solutions. Then $\mu_{\mathbf{t}}$ is greater than zero exactly when a solution exists to the original ONE-IN-THREE-3SAT problem.

A.2 Sketch of proof for $Z_{\mathbf{a}}$, equation 3

Use the same embedding as above. Again, a score of zero corresponds to a failure, and a score of 1 corresponds to a solution, of the original ONE-IN-THREE-3SAT problem. Shorten the sequence so that there is exactly one amino acid per core segment in the encoded problem, hence exactly one threading in the solution search space. Then $Z_{\mathbf{a}}$ is greater than zero exactly when a solution exists to the original ONE-IN-THREE-3SAT problem.

A.3 Sketch of proof for μ_{i,t_i} and $\mu_{I,T}$, equations 17 and 20

Because $\mu_{\mathbf{t}} = \sum_{t_i=1}^{\tilde{n}} \mu_{i,t_i}$, a polynomial-time computation for μ_{i,t_i} would imply a polynomial-time computation for $\mu_{\mathbf{t}}$, which is NP-hard by section A.1. In turn, μ_{i,t_i} is a special case of $\mu_{I,T}$, which therefore cannot be easier.

B Appendix — Sequence singleton-only case

Low-order polynomial time exact recursive relationships are derived for the restricted gapped sequence singleton-only case of a Bayesian analysis that unifies protein sequence-structure alignment and recognition. This allows the Bayes-optimal selection and alignment terms to be computed rapidly and exactly for proteins of realistic size, but at the price of giving up information potentially encoded in pair (or higher-order) interactions between sequence residues.

This appendix gives exact formulae first for arbitrary sequence-specific segment and loop score functions, then specializes them to the per-residue case. The recurrence relations are similar to the low-order polynomial forward-backward and Viterbi procedures of Hidden Markov Models [Rabiner, 1989].

B.1 Additional Notation

B.1.1 Notation for Sequence, Structure, Alignment

The sequence $\mathbf{a}$ is a string of length n over an alphabet A of twenty characters (amino acid residue types). The set A^n consists of all strings over A of length n. The sequence $\mathbf{b}$ is a summation variable over A^n.

Core Structures and Library

The core structure C is drawn from a library $\mathcal{L}$ of cores. Core structure C is composed of m core segments C_i, each of length c_i amino acid residues. The segments C_i may correspond to the backbone trace of conserved secondary structure segments. When $c_i = 1$ the segments may correspond to single amino acid residue positions. Each segment C_i is composed of c_i primitive core elements $C_{i,j}$. Each element $C_{i,j}$ corresponds to a spatial position that may be occupied by a residue from the sequence. No alignment gaps are permitted within segments, so adjacent elements within a segment are aligned to adjacent residues from the sequence.

Core segments are connected by a set λ of loops, with loop λ_i connecting segment C_i to C_{i+1}, N-terminal leader λ_0 preceding C_1, and C-terminal trailer λ_m following C_m. The length of loop λ_i is the variable l_i and its maximum (respectively minimum) length is l_i^{max} (respectively l_i^{min}). Unless stated otherwise, $l_i^{max} = +\infty$ and l_i^{min} = the minimum geometric spanning loop length (i.e., the minimum loop length capable of spanning the distance between the end of C_i and the beginning of C_{i+1}), with $l_0^{min} = l_m^{min} = 0$.

Alignment

The set $\mathcal{T}[C, n]$ consists of all legal alignments of a sequence of length n to the core structure C. The vector $\mathbf{x}$ is a summation variable over $\mathcal{T}[C, n]$.

A sequence-structure alignment ("threading") is specified by a vector of m integers, denoted by $\mathbf{t}^a$ in absolute coordinates and $\mathbf{t}$ in relative coordinates. Each absolute coordinate t_i^a specifies the index in the sequence $\mathbf{a}$ that is aligned to the first element of the i^{th} core segment. That is, t_i^a is the index of the sequence residue that occupies $C_{i,1}$.

For simpler notation [Lathrop & Smith, 1996] we generally replace absolute sequence coordinates $\mathbf{t}^a$ by relative coordinates $\mathbf{t}$, defined by $t_i = t_i^a - \sum_{j<i}(c_j + l_j^{min})$. Let $\tilde{n} = n + 1 - \sum_i(c_i + l_i^{min})$ and $\tilde{l}_i = l_i^{max} - l_i^{min}$. Then $t_i = 1$ corresponds to the lowest legal value of t_i^a and $t_i = \tilde{n}$ to the highest. Below, the absence of the superscript a will indicate relative coordinates.

The i^{th} loop length l_i and segment length c_i are related to $\mathbf{t}^a$ and $\mathbf{t}$ by $l_i = t_{i+1}^a - t_i^a - c_i = t_{i+1} - t_i + l_i^{min}$. Due to the minimum spanning loop length constraints, $1 + \sum_{j<i}(c_j + l_j^{min}) \le t_i^a \le n + 1 - \sum_{j \ge i}(c_j + l_j^{min})$. Due to core segment topological ordering constraints, $t_i^a + c_i + l_i^{min} \le t_{i+1}^a \le t_i^a + c_i + l_i^{max}$. In relative coordinates, the minimum loop length constraints simplify to $1 \le t_i \le \tilde{n}$ and the ordering constraints simplify to $t_i \le t_{i+1} \le t_i + \tilde{l}_i$.

Fictitious segments C_0 (respectively C_{m+1}) are fixed at the beginning (respectively end) of the sequence whenever it is convenient for indicated summations or recurrence limits. By convention, $c_0 = c_{m+1} = 0$, i.e., fictitious segments have zero length; and $t_0 = 1$ and $t_{m+1} = \tilde{n}$, i.e., they are fixed.

B.1.2 Notation for Arbitrary Sequence Singleton-Only Objective Function

In this section we define an objective function that allows an arbitrary sequence-specific score function for each segment or loop, but ignores all pairwise or higher order interactions between non-adjacent segments. We use f^1 to distinguish this singleton-only objective function from the general case, and Z^1 and μ^1 to distinguish corresponding global sums and means. To simplify notation we omit the sequence ($\mathbf{a}$) and core structure (C) arguments when they are clear from context, writing $f_s(i, t_i)$ to abbreviate $f_s(\mathbf{a}, C, i, t_i)$ and so on. Recall throughout that the relative coordinates shown must be converted to absolute coordinates as above to obtain an actual index into $\mathbf{a}$; specifically, add $\sum_{j<i}(c_j + l_j^{min})$ (respectively $\sum_{j<i+1}(c_j + l_j^{min})$) to the second argument t_i (respectively the third argument t_{i+1}) of f_s, f_l, h_s, h_l, h_λ, H_s, and H_l; and add $\sum_{j<i}(c_j + l_j^{min})$ to the relative coordinate x in $\mathbf{a}[x]$.

Let $f_s(i, t_i)$ be the score for occupying segment C_i by the substring of length c_i beginning at $\mathbf{a}[t_i]$, and let $f_l(i, t_i, t_{i+1})$ be the score for occupying loop λ_i by the substring of length $l_i = t_{i+1} - t_i + l_i^{min}$ beginning at $\mathbf{a}[t_i + c_i]$. If desired, pair interactions entirely within segment C_i may be encoded in $f_s(i, t_i)$, and those between segments C_i and C_{i+1} may be encoded in $f_l(i, t_i, t_{i+1})$. This allows an arbitrary sequence-specific score function for each segment or loop, but ignores all pairwise or higher order interactions between non-adjacent segments.

Assume that the threading score f^1 is the sum of the segment f_s and loop f_l scores

separately.

$$f^1(\mathbf{a}, C, \mathbf{t}) = \sum_{i=1}^{m} f_s(i, t_i) + \sum_{i=0}^{m} f_l(i, t_i, t_{i+1}) \tag{72}$$

Functions h_s and h_l are the unnormalized probability functions corresponding to f_s and f_l.

$$h_s(i, t_i) = \exp(-f_s(i, t_i)) \tag{73}$$
$$h_l(i, t_i, t_{i+1}) = \exp(-f_l(i, t_i, t_{i+1})) \tag{74}$$

By convention, all illegal or out-of-range indices imply score $+\infty$ (infinitely bad) and probability zero; and $h_s(0, x) = h_s(m+1, x) = 1$, i.e., fictitious segments have zero score and unit probability; and $h_l(0, x, x) = h_l(m+1, x, x) = 1$, i.e., they have zero length.

Function H_s (respectively H_l) is the sum of h_s (respectively h_l) over all strings over A of length c_i (respectively l_i).

$$H_s(i, t_i) = \sum_{\mathbf{b} \in A^{c_i}} h_s(\mathbf{b}, C, i, t_i) \tag{75}$$
$$H_l(i, t_i, t_{i+1}) = \sum_{\mathbf{b} \in A^{l_i}} h_l(\mathbf{b}, C, i, t_i, t_{i+1}) \tag{76}$$

Function $h_\lambda(i, t_i, t_{i+1})$ is the sequence-independent prior probability of observing loop length $l_i = t_{i+1} - t_i + l_i^{min}$ at loop λ_i. It differs from h_l and H_l in being a prior probability distribution over loop lengths, while h_l and H_l are posterior probabilities derived from the sequence residues to occupy the loop. The assumption that loop lengths are independent yields

$$P(\mathbf{t}|n, C) = \prod_{i=0}^{m} h_\lambda(i, t_i, t_{i+1}) \tag{77}$$

If uninformative priors are used, then $h_\lambda(i, t_i, t_{i+1}) = |\mathcal{T}[C, n]|^{-1/(m+1)}$ and the equation is exact. If an empirical loop length distribution [Benner et al., 1993] is used, then h_λ is taken from empirical tables; in this case the equation is approximate because $\sum_i \tilde{l}_i = \tilde{n} - 1$ so the assumption of loop length independence is violated, but it may yield a biologically more plausible result in some cases.

B.1.3 Notation for Per-Residue Sequence Singleton-only Objective Function

In many current proposals, the sequence singleton-only f_s (respectively f_l) is specialized further to be the sum of the individual sequence residue scores at each element of the segment (respectively loop). Here we give a simple way to derive f_s, f_l, H_s, and H_l, in such proposals.

Let $s(C_{i,j})$ be the structural environment assigned to core element $C_{i,j}$ and $s(\lambda_i)$ be the structural environment assigned to loop λ_i. $s(C_{i,j})$ potentially reflects a different structural environment for each core element, as defined by the theory of protein structure used. The loop structural environment $s(\lambda_i)$ might be used to divide loops into categories, e.g., tight, short, medium, and long; or all loops might be assigned to a single generic loop environment. Let $f_a^A(a', s)$ be the score assigned to amino acid residue type $a' \in A$ in environment s, and let $h_a^A(a', s) = \exp(-f_a^A(a', s))$.

$$f^A(\mathbf{a}, C, \mathbf{t}) = \sum_{i=1}^{m} f_s^A(i, t_i) + \sum_{i=0}^{m} f_l^A(i, t_i, t_{i+1}) \tag{78}$$

$$f_s^A(i, t_i) = \sum_{j=1}^{c_i} f_a^A(\mathbf{a}[t_i + j - 1], s(C_{i,j})) \tag{79}$$

$$f_l^A(i, t_i, t_{i+1}) = \sum_{j=1}^{l_i} f_a^A(\mathbf{a}[t_i + c_i + j - 1], s(\lambda_i)) \tag{80}$$

$$h_s^A(i, t_i) = \exp(-f_s^A(i, t_i)) = \prod_{j=1}^{c_i} h_a^A(\mathbf{a}[t_i + j - 1], s(C_{i,j})) \tag{81}$$

$$h_l^A(i, t_i, t_{i+1}) = \exp(-f_l^A(i, t_i, t_{i+1})) = \prod_{j=1}^{l_i} h_a^A(\mathbf{a}[t_i + c_i + j - 1], s(\lambda_i)) \tag{82}$$

$$H_s^A(i, t_i) = \prod_{j=1}^{c_i} \sum_{a' \in A} h_a^A(a', s(C_{i,j})) \tag{83}$$

$$H_l^A(i, t_i, t_{i+1}) = \prod_{j=1}^{l_i} \sum_{a' \in A} h_a^A(a', s(\lambda_i)) \tag{84}$$

B.2 Results — Sequence Singleton-only Recursions

B.2.1 Arbitrary Sequence Singleton-only Objective Function

This section allows arbitrary sequence-specific score functions for the segment and loops, and an arbitrary length-dependent function for h_λ. The time complexity of recursions in this section is $\mathcal{O}(m\tilde{n}^2)$, where $\tilde{n}$ is the relative or effective sequence length and m is the number of core segments involved.

Recursion for $Z_{\mathbf{a}}^1$

In the sequence singleton-only case, if h_s and h_l are normalized probabilities, then $Z_{\mathbf{a}}^1 = H_s = H_l = 1$; otherwise, $Z_{\mathbf{a}}^1$ corresponds to a normalizing constant for $\exp(-f^1)$.

$$Z_{\mathbf{a}}^1 = \sum_{\mathbf{b} \in A^n} \exp\left(-f^1(\mathbf{b}, C, \mathbf{t})\right) \tag{85}$$

$$= \prod_{i=0}^{m} H_l(i, t_i, t_{i+1}) H_s(i, t_i) \tag{86}$$

Recursion for $\mu^1_{\mathbf{t}}$

$$\mu^1_{\mathbf{t}} = \sum_{\mathbf{x} \in \mathcal{T}[C,n]} \exp(-f^1(\mathbf{a}, C, \mathbf{x})) P(\mathbf{x}|n, C) \tag{87}$$

$$= \sum_{\mathbf{x} \in \mathcal{T}[C,n]} \prod_{i=0}^{m} h_l(i, x_i, x_{i+1}) h_\lambda(i, x_i, x_{i+1}) h_s(i+1, x_i) \tag{88}$$

Define an intermediate function R by the recurrence

$$R(m, x) = h_l(m, x, \tilde{n}) h_\lambda(m, x, \tilde{n}) \tag{89}$$

$$R(i, x) = \sum_{y=x}^{\tilde{n}} h_l(i, x, y) h_\lambda(i, x, y) h_s(i+1, y) R(i+1, y) \quad , \ 0 \le i < m \tag{90}$$

$R(i, x)$ is the unnormalized probability corresponding to placing segment i at relative coordinate x but assigning it zero score, together with all following segments and loops, summed over all possible placements of the following segments. That is, $R(i, x)$ is $\mu^1_{\mathbf{t}}$ restricted to segments $i+1$ and above and the substring $\mathbf{a}[x]$ and beyond. Consequently,

$$\mu^1_{\mathbf{t}} = R(0, 1) \tag{91}$$

Recursion for μ^1_{i,t_i}

$$\mu^1_{i,t_i} = \sum_{\{\mathbf{x} \in \mathcal{T}[C,n] | x_i = t_i\}} \exp(-f^1(\mathbf{a}, C, \mathbf{x})) P(\mathbf{x}|n, C) \tag{92}$$

$$= \sum_{\{\mathbf{x} \in \mathcal{T}[C,n] | x_i = t_i\}} \prod_{i=0}^{m} h_s(i, x_i) h_l(i, x_i, x_{i+1}) h_\lambda(i, x_i, x_{i+1}) \tag{93}$$

where equation 93 follows because $h_s(0, x) = h_s(m+1, x) = 1$.

Define Q by the recurrence

$$Q(1, x) = h_l(0, 1, x) h_\lambda(0, 1, x) \tag{94}$$

$$Q(i, x) = \sum_{y=1}^{x} Q(i-1, y) h_s(i-1, y) h_l(i-1, y, x) h_\lambda(i-1, y, x) \tag{95}$$
$$, \ 1 < i \le m+1$$

$Q(i,x)$ is the unnormalized probability corresponding to placing segment i at relative coordinate x but assigning it zero score, together with all preceding segments and loops, summed over all possible placements of the preceding segments. That is, $Q(i,x)$ is $\mu^1_{\mathbf{t}}$ restricted to segments $i-1$ and below and the substring $\mathbf{a}[x]$ and before. Consequently,

$$\mu^1_{i,t_i} = Q(i,t_i)h_s(i,t_i)R(i,t_i) \tag{96}$$

Recursion for $\mu^1_{I,T}$

$$\mu^1_{I,T} = \sum_{\{\mathbf{x}\in\mathcal{T}[C,n]|j\in I\Rightarrow x_j=t_j\}} \exp(-f^1(\mathbf{a},C,\mathbf{x}))P(\mathbf{x}|n,C) \tag{97}$$

$$= \sum_{\{\mathbf{x}\in\mathcal{T}[C,n]|j\in I\Rightarrow x_j=t_j\}} \prod_{i=0}^{m} h_s(i,x_i)h_l(i,x_i,x_{i+1})h_\lambda(i,x_i,x_{i+1}) \tag{98}$$

Recall that $I=\{i_1,i_2,\ldots,i_k\}$ and $T=\{t_{i_1},t_{i_2},\ldots,t_{i_k}\}$. By convention, let $i_0=0$ and $t_{i_0}=1$. Define Q_j by the recurrence

$$Q_j(i_{j-1}+1,x) = h_l(i_{j-1},t_{i_{j-1}},x)h_\lambda(i_{j-1},t_{i_{j-1}},x) \tag{99}$$

$$Q_j(i,x) = \sum_{y=t_{i_{j-1}}}^{x} Q_j(i-1,y)h_s(i-1,y)h_l(i-1,y,x)h_\lambda(i-1,y,x) \tag{100}$$
$$, \; i_{j-1}+1 < i \le i_j$$

$Q_j(i,x)$ is the unnormalized probability corresponding to placing segment i at relative coordinate x but assigning it zero score, together with all preceding segments and loops back to but excluding placing segment i_{j-1} at $t_{i_{j-1}}$, summed over all possible placements of the intervening segments. Consequently, with $k=|I|$,

$$\mu^1_{I,T} = \Big(\prod_{j=1}^{k} Q_j(i_j,t_{i_j})h_s(i_j,t_{i_j})\Big)R(i_k,t_{i_k}) \tag{101}$$

For use with secondary structure prediction, loop modeling, α-/β-hairpins, etc., observe the special case of

$$\mu_{\langle \mathbf{a},n,C,\{i,i+1\},\{t_i,t_{i+1}\}\rangle} \tag{102}$$

$$= Q(i,t_i)h_s(i,t_i)h_l(i,t_i,t_{i+1})h_\lambda(i,t_i,t_{i+1})h_s(i+1,t_{i+1})R(i+1,t_{i+1}) \tag{103}$$

Invariants

Useful diagnostic invariants include

$$\begin{aligned}
\mu^1_{\mathbf{t}} &= \sum_{\mathbf{t}\in\mathcal{T}[C,n]} \exp\left(-f^1(\mathbf{a},C,\mathbf{t})\right) && (104)\\
&= R(0,1) && (105)\\
&= Q(m+1,\tilde{n}) && (106)\\
&= \sum_{t_i=1}^{\tilde{n}} \mu^1_{i,t_i} && (107)
\end{aligned}$$

B.2.2 Per-Residue Sequence Singleton-only Objective Function

Specializing the sequence singleton-only f_s (respectively f_l) to be the sum of the individual sequence residue scores at each element of the segment (respectively loop) is discussed in section B.1.3. This leads to recurrence relations that are more efficient than section B.2.1 by a factor of $\tilde{n}$, because the loop function can be included in the recursion. The time complexity of recursions in this section is $\mathcal{O}(m\tilde{n})$, where $\tilde{n}$ is the relative or effective sequence length and m is the number of core segments involved.

However, the new recurrences no longer make loop endpoints or lengths explicit, so the uninformative loop prior $P(\mathbf{t}|n,C) = |\mathcal{T}[C,n]|^{-1}$ is assumed, a per-loop structural environment $s(\lambda_i)$ is used, and pair interactions between adjacent segments are not allowed. A superscript A indicates these assumptions.

The new recurrences are

$$\begin{aligned}
R^A(m,x) &= h^A_l(m,x,\tilde{n})|\mathcal{T}[C,n]|^{-1} && (108)\\
Q^A(1,x) &= h^A_l(0,1,x) && (109)\\
Q^A_j(i_{j-1}+1,x) &= h^A_l(i_{j-1},t_{i_{j-1}},x) && (110)\\
R^A(i,x) &= \begin{cases} h^A_l(i,x,x)h^A_s(i+1,x)R^A(i+1,x) \\ \qquad +h^A_a(\mathbf{a}[x+c_i],s(\lambda_i))R^A(i,x+1) \\ \quad , 1\le x\le\tilde{n} \text{ and } 0\le i<m \\ 0 \quad , \text{otherwise} \end{cases} && (111)\\
Q^A(i,x) &= \begin{cases} Q^A(i-1,x)h^A_s(i-1,x)h^A_l(i-1,x,x) \\ \qquad +Q^A(i,x-1)h^A_a(\mathbf{a}[x-1],s(\lambda_{i-1})) \\ \quad , 1\le x\le\tilde{n} \text{ and } 1<i\le m+1 \\ 0 \quad , \text{otherwise} \end{cases} && (112)
\end{aligned}$$

$$Q_j^A(i,x) = \begin{cases} Q_j^A(i-1,x)h_s^A(i-1,x)h_l^A(i-1,x,x) \\ \qquad +Q_j^A(i,x-1)h_a^A(\mathbf{a}[x-1],s(\lambda_{i-1})) \\ \quad , t_{i_{j-1}} \le x \le t_{i_j} \text{ and } i_{j-1}+1 < i \le i_j \\ 0 \quad , \text{otherwise} \end{cases} \tag{113}$$

Consequently,

$$Z_{\mathbf{a}}^A = \prod_{i=0}^{m} H_l^A(i,t_i,t_{i+1})H_s^A(i,t_i) \tag{114}$$

$$\mu_{\mathbf{t}}^A = R^A(0,1) \tag{115}$$

$$\mu_{i,t_i}^A = Q^A(i,t_i)h_s^A(i,t_i)R^A(i,t_i) \tag{116}$$

$$\mu_{I,T}^A = \Big(\prod_{j=1}^{k} Q_j^A(i_j,t_{i_j})h_s^A(i_j,t_{i_j})\Big)R^A(i_k,t_{i_k}) \tag{117}$$

C Appendix — Variable $Z_{\mathbf{a}}$

Equation 3 gave the threading-independent definition of $Z_{\mathbf{a}}$. The body of the paper treated the case where $Z_{\mathbf{a}}$ is the same for every $\mathbf{t} \in \mathcal{T}[C,n]$. Here we treat the case where $Z_{\mathbf{a}}$ varies with $\mathbf{t}$. In this case, f induces a partition on $\mathcal{T}[C,n]$ such that two threadings $\mathbf{t}$ and $\mathbf{u}$ are in the same partition element if and only if $Z_{\mathbf{a}}(\mathbf{t}) = Z_{\mathbf{a}}(\mathbf{u})$. Let the induced partition be $\mathcal{T}^*[C,n] = \{\mathcal{T}_i^*[C,n]\}$ where $\mathcal{T}_i^*[C,n] \subseteq \mathcal{T}[C,n]$ is the i^{th} partition element and the $\mathcal{T}_i^*[C,n]$ are disjoint and cover $\mathcal{T}[C,n]$. Let ${}_iZ$ (respectively ${}_i\mu$) represent global sums (respectively global means) over threadings in partition element $\mathcal{T}_i^*[C,n]$. Define

$$ {}_iZ_{\mathbf{a}} = \sum_{\mathbf{b} \in A^n} \exp(-f(\mathbf{b}, C, \mathbf{t})) \quad \text{, where } \mathbf{t} \in \mathcal{T}_i^*[C,n] \tag{118} $$

$$ {}_i\mu_{\mathbf{t}} = \sum_{\mathbf{x} \in \mathcal{T}_i^*[C,n]} \exp(-f(\mathbf{a}, C, \mathbf{x}))P(\mathbf{x}|n, C) \tag{119} $$

$$ {}_i\mu_{I,T} = \sum_{\{\mathbf{x} \in \mathcal{T}_i^*[C,n] | j \in I \Rightarrow x_j = t_j\}} \exp(-f(\mathbf{a}, C, \mathbf{x}))P(\mathbf{x}|n, C) \tag{120} $$

Equation 9 must be generalized to

$$ P(C|\mathbf{a}, n) = \frac{P(C|n)}{P(\mathbf{a}|n)} \sum_{T_i^*[C,n]} \frac{{}_i\mu_{\mathbf{t}}}{{}_iZ_{\mathbf{a}}} \tag{121} $$

Equation 19 must be generalized to

$$ P(I, T, C|\mathbf{a}, n) = \frac{P(C|n)}{P(\mathbf{a}|n)} \sum_{T_i^*[C,n]} \frac{{}_i\mu_{I,T}}{{}_iZ_{\mathbf{a}}} \tag{122} $$

Recall that equation 16 was the special case of equation 19 when $k = 1$, and must be generalized accordingly. Equations 5 and 10 are unchanged, but must be interpreted with variable $Z_{\mathbf{a}}$.

Figure Legends

Legend for figure 1.

A schematic view of the gapped block alignment approach to protein threading (adapted from [Lathrop & Smith, 1996]).
(a) Conceptual drawing of two structurally similar proteins and a common core of four secondary structure segments (dark lines, I-L). Note that there is no restriction on core segment length, from a single residue position upwards. To form the structural models used here, side-chains were replaced by a methyl group resulting in polyalanine, and loops or variable regions were removed resulting in discrete core segments.
(b) Abstract structural model showing spatial adjacencies (interactions). Small circles represent amino acid positions (core elements), and thin lines connect neighbors that interact in the objective function. The structural environments and interacting positions will be recorded for later use by the objective function.
(c) Illustration of the combinatorically large number of threadings (sequence-structure alignments) possible with a novel sequence. t_x^a indexes the sequence amino acid placed into the first element of segment X. Sequence regions between core segments become connecting turns or loops, which are constrained to be physically realizable. All alignment gaps are confined to turn or loop regions.
(d) A sequence is threaded through the model by placing successive sequence amino acids into adjacent core elements. Alignments are rank-ordered by their probability according to equation 5. The globally most probable alignment is shown selected. It is also possible to enumerate all alignments in order of probability, or to sample the near-optimal alignments.

Legend for figure 2.

Selecting a core from a structural library. Cores in the structural library are rank ordered by their probability according to equation 9. The globally most probable core is shown selected. It is also possible to enumerate all cores in order of probability, or to sample most-probable cores.

Legend for figure 3.

Selecting a core and alignment jointly. Conceptually, every core structure of the structural library (Str. Lib., C_i) is used to generate a pool of all possible sequence-structure alignments (All Aligns., $\cup_i \mathcal{T}[C_i, n]$) with the input sequence. The pooled (core, alignment) pairs are rank ordered by probability according to equation 14. The globally most probable core structure and alignment pair is shown selected. It is also possible to enumerate all (core, alignment) pairs in order of probability, or to sample near-optimal pairs.

Legend for figure 4.

A simple example using the HP model. (a) The core library $\mathcal{L}$, showing the cores' native sequences for reference. (b) The unknown sequence to be threaded, HHP. (c) The first threading of HHP onto $\mathcal{L}_1$, no gaps. (d) The second threading of HHP onto $\mathcal{L}_1$, H–HP. (e) The third threading of HHP onto $\mathcal{L}_1$, HH–P. (f) The only threading of HHP onto $\mathcal{L}_2$, no gaps.

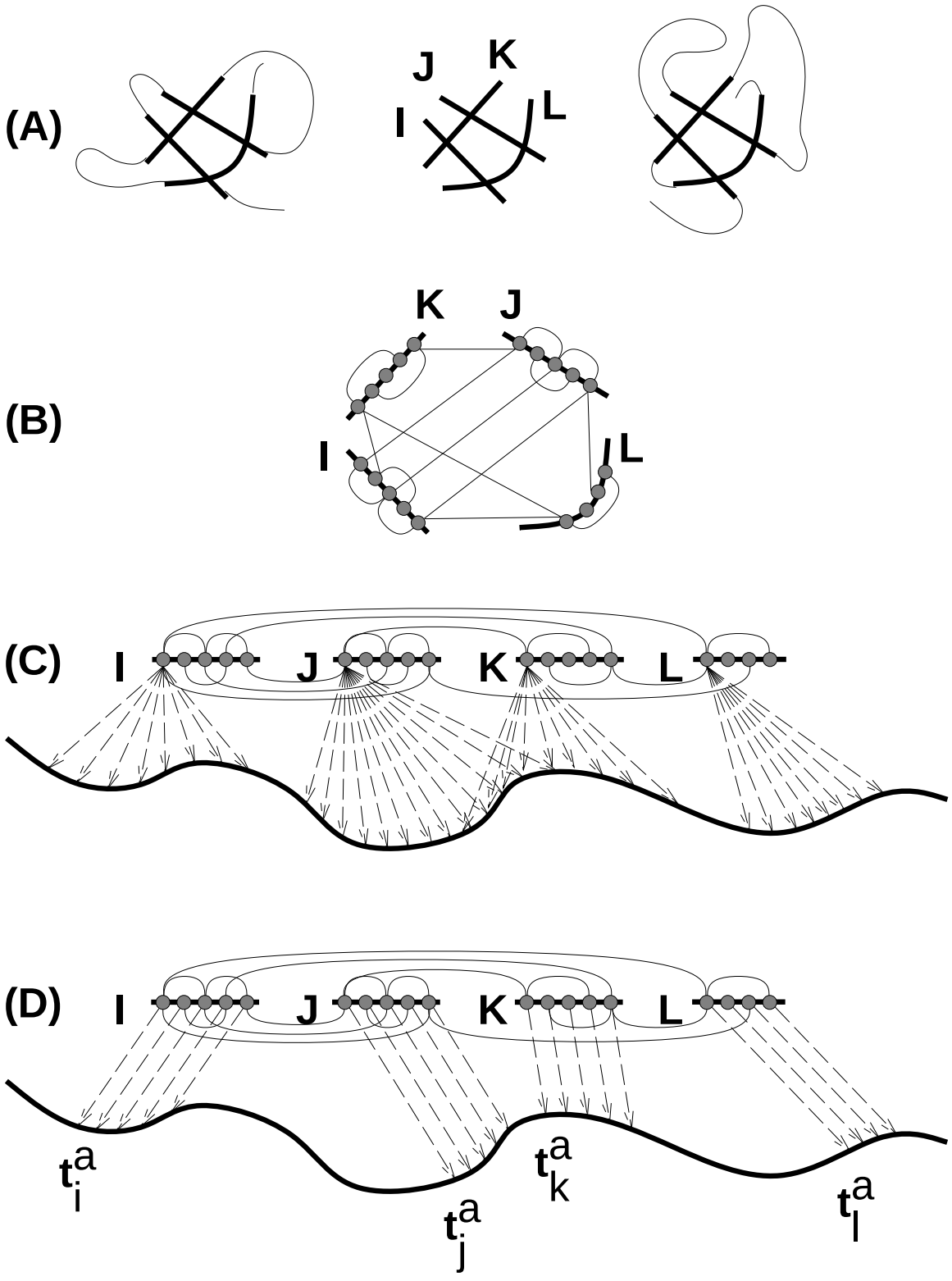

Figure 1:

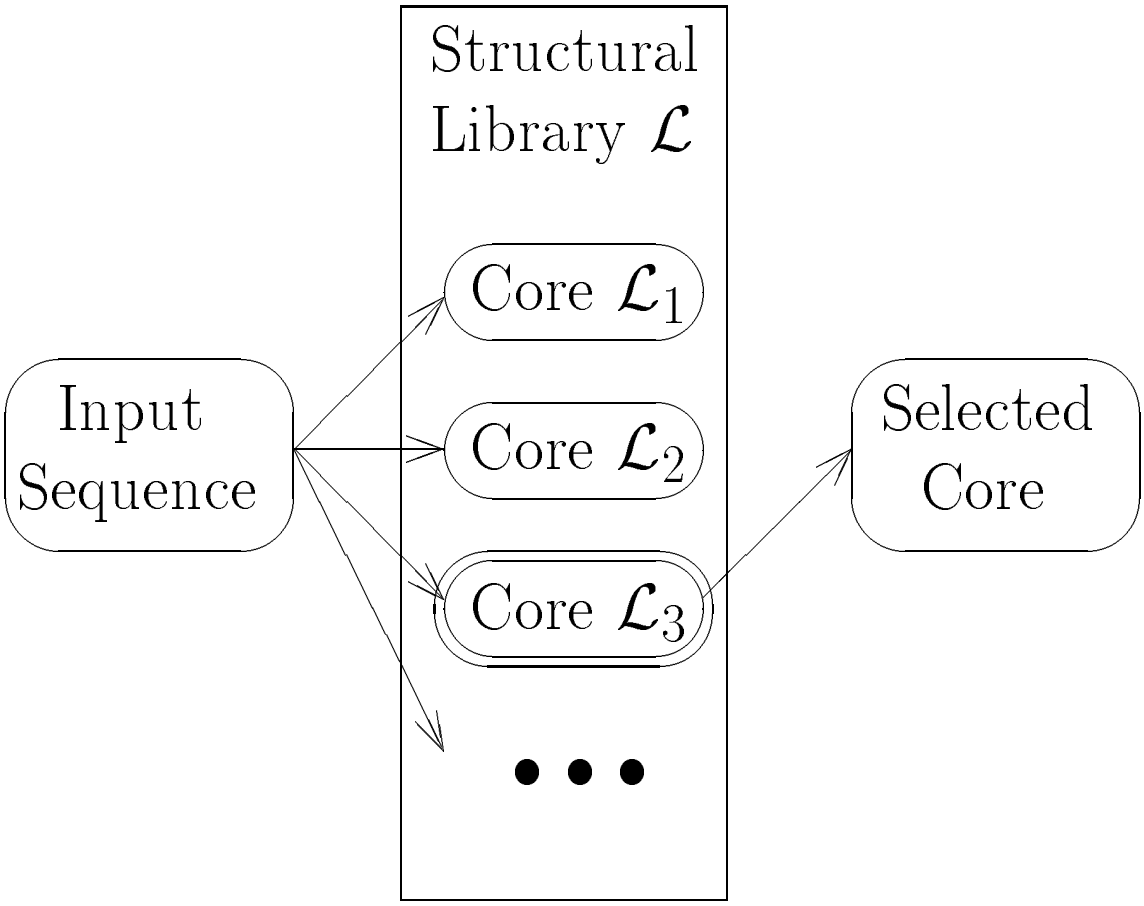
Structural
Library $\mathcal{L}$
Core $\mathcal{L}_1$
Core $\mathcal{L}_2$
Core $\mathcal{L}_3$
Input
Sequence
Selected
Core

Figure 2:

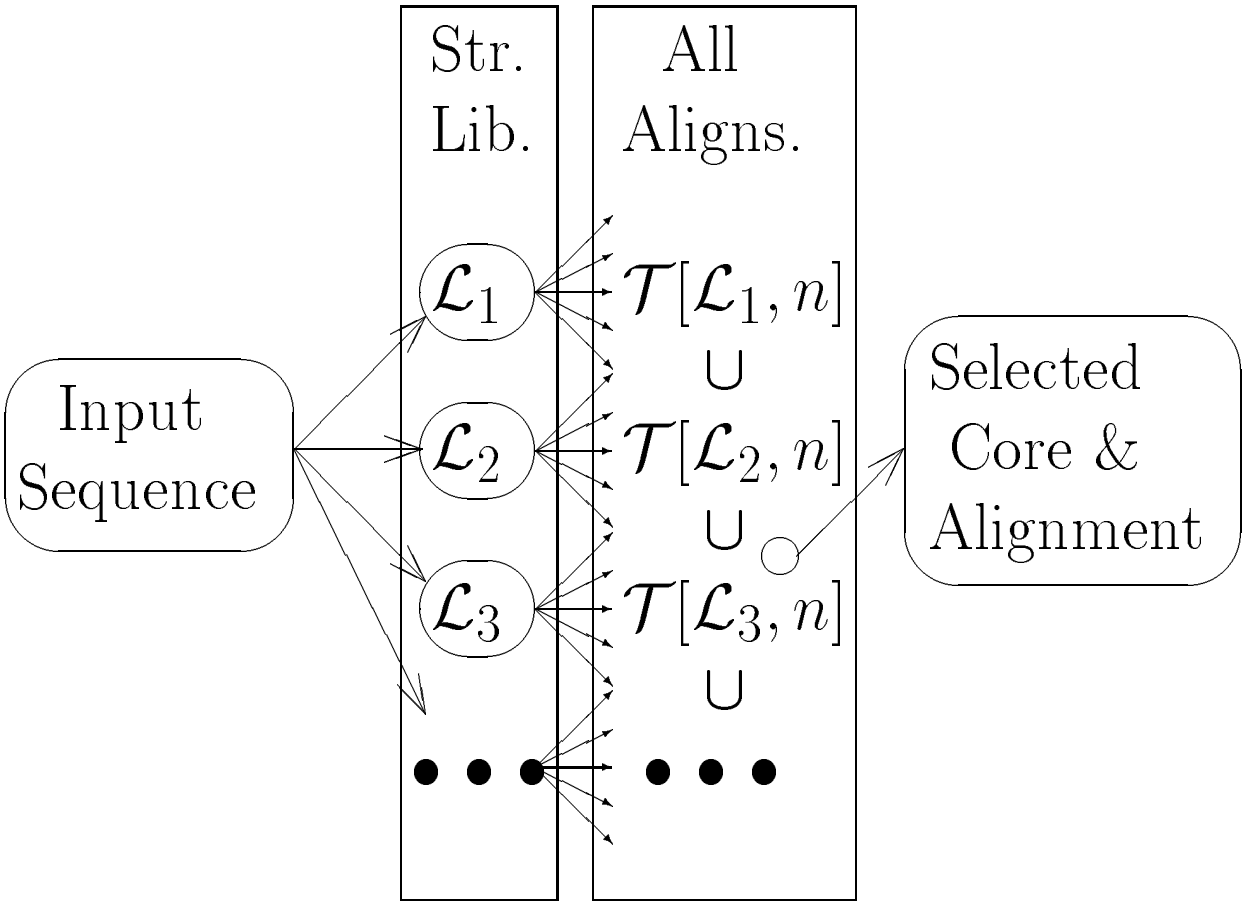
Str.
Lib.
All
Aligns.
$\mathcal{L}_1$
$\mathcal{L}_2$
$\mathcal{L}_3$
$\mathcal{T}[\mathcal{L}_1, n]$
$\cup$
$\mathcal{T}[\mathcal{L}_2, n]$
$\cup$
$\mathcal{T}[\mathcal{L}_3, n]$
$\cup$
Input
Sequence
Selected
Core &
Alignment

Figure 3:

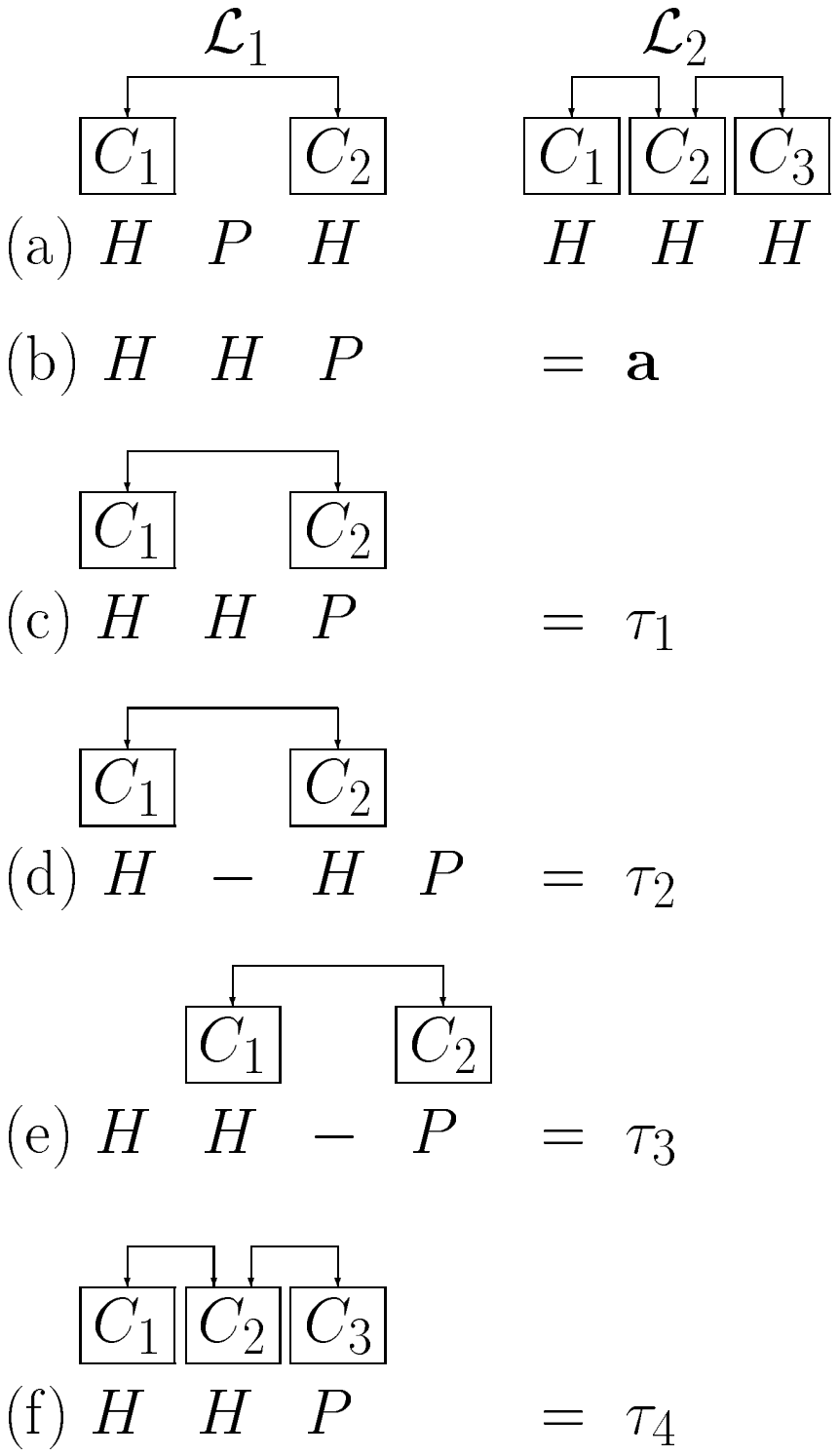
$\mathcal{L}_1$
$\mathcal{L}_2$
C_1 C_2
C_1 C_2 C_3
(a) H P H
H H H
(b) H H P $= \mathbf{a}$
C_1 C_2
(c) H H P $= \tau_1$
C_1 C_2
(d) H $-$ H P $= \tau_2$
C_1 C_2
(e) H H $-$ P $= \tau_3$
C_1 C_2 C_3
(f) H H P $= \tau_4$

Figure 4:

Notation	Usage
$\mathbf{a}$	a sequence or string over A of length n
A	an alphabet of 20 characters (amino acid types)
A^n	the set of all strings over A of length n
$\mathbf{b}$	a summation variable over A^n
C	a core structure; its i^{th} segment is C_i, whose j^{th} element is $C_{i,j}$
c_i	$\|C_i\|$, the length of the i^{th} core segment C_i
f	an objective function (score function)
f^1	a sequence singleton-only version of f
f^A	a per-residue version of f^1
f_a	f restricted to amino acid residue types
f_l	f restricted to loops or variable regions
f_p	f restricted to pair interactions
f_s	f restricted to core segments
h	$\exp(-f)$
H	$\sum_{\mathbf{b}\in A^n} h$
h_λ	the loop length prior probability
l_i	$\|\lambda_i\|$, the variable length of the i^{th} loop λ_i
l_i^{min} (or l_i^{max})	the minimum (or maximum) value of l_i
$\mathcal{L}$	a library of core structures; the i^{th} library member is $\mathcal{L}_i$ or C
m	$\|C\|$, the number of core segments in C
n	$\|\mathbf{a}\|$, the length of the sequence $\mathbf{a}$
$\tilde{n}$	$n + 1 - \sum_i (c_i + l_i^{min})$; the relative sequence length
$P(A\|B)$	the conditional probability of A given B
Q, Q_j, R	recurrence functions
$\mathbf{t}$ (or $\mathbf{t}^a$)	a vector of m integers; t_i (or t_i^a) is the i^{th} relative (or absolute) coordinate
$\mathcal{T}[C, n]$	the set of all alignments given core C and sequence length n
$\mathbf{x}$	a summation variable over $\mathcal{T}[C, n]$
$Z_x[y]$	a global sum specified by x, optionally specified by y
λ	a set of loops; the i^{th} loop is λ_i, whose j^{th} element is $\lambda_{i,j}$
$\mu_x[y]$	a global mean specified by x, optionally specified by y

Table 1: Notational usage of this paper.