The Combinatorics and Extreme Value Statistics of Protein Threading

John L. Spouge, Aron Marchler-Bauer, and Stephen Bryant

National Center for Biotechnology Information
National Library of Medicine
Bethesda MD 20894

Phone: (301) 435-5912
Fax: (301) 435-2433
Email: spouge@nih.gov

For the Los Alamos Conference on Combinatorics In Biology, 1998

Running Title: The Combinatorics of Protein Threading

Version Date: January 7, 2003

Abstract

In protein threading, one is given a protein sequence, together with a database of protein core structures that may contain the natural structure of the sequence. The object of protein threading is to identify correctly the structure(s) corresponding to the sequence. Since the core structures are already associated with specific biological functions, threading has the potential to provide biologists with useful insights about the function of a newly discovered protein sequence. Statistical tests for threading results based on the theory of extreme values suggest several combinatorial problems. For example, what is the number of ways $m' = \#_t \{L_i > x_i\}_{i=0}^{n}$ of choosing a sequence $\{X_i\}_{i=1}^{n}$ from the set $\{1,2,\ldots,t\}$, subject to the difference constraints $\{L_i = X_{i+1} - X_i > x_i\}_{i=0}^{n}$, where $X_0 = 0$, $X_{n+1} = t+1$, and $\{x_i\}_{i=0}^{n}$ is an arbitrary sequence of integers? The quantity m' has many attractive combinatorial interpretations and reduces in special continuous limits to a probabilistic formula discovered by de Finetti. Just as many important probabilities can be derived from de Finetti's formula, many interesting combinatorial quantities can be derived from m'. Empirical results presented here show that the combinatorial approach to threading statistics appears promising, but that structural periodicities in proteins and energetically unimportant structure elements probably introduce statistical correlations that must be better understood.

Keywords: Protein Threading, Extreme Value Statistics, Poisson Clumping Heuristic, Probabilities related to Uniform Distributions

1 Introduction

In recent years, laboratory automation has led to an explosive acquisition of biological information like DNA sequences [10], protein sequences [8, 9], and protein structures [1, 2, 19]. As the corresponding databases have grown, sifting them for biological relationships has become increasingly central to cell biology, biochemistry, and molecular biology, and ultimately to therapeutic drug design. Because the search for relationship is akin to looking for needles in a haystack, or more formally, finding unusual events among many alternatives, extreme value statistics are now an implicit [15, 20, 21] but essential tool of the bench biologist [4, 5].

Let $\{Y_i\}_{i=1}^m$ be a sequence of random variables, and let $Y_1^* \le Y_2^* \le \ldots \le Y_m^*$ represent the corresponding order statistics, i.e., the same values $\{Y_i\}_{i=1}^m$ but in increasing order. The theory of extreme values seeks as its grail the limiting distribution of $Y_m^* = \max_{1 \le i \le m} Y_i$ and the other order statistics as $m \to \infty$.

Extreme value theory has essentially two forms, classical and modern. The prototype for a classical result states that if $\{Y_i\}_{i=1}^m$ are independent and identically distributed, after a linear scaling $\xi_m = s_m^{-1}(Y_m^* - c_m)$, only three non-degenerate limiting distributions for Y_m^* can occur as $m \to \infty$ [17]. If ξ_m does not degenerate to a constant, the limiting distribution of ξ_m (if it exists) is one of only three types of extreme value distribution

$$\left.\begin{array}{l} P(\xi^{(1)} \le x) \\ P(\xi^{(2)} \le x) \\ P(\xi^{(3)} \le x) \end{array}\right\} = \left\{\begin{array}{l} \exp[-(-x)^{\alpha}] \\ \exp(-x^{-\alpha}) \\ \exp(-e^{-x}) \end{array}\right. , \qquad (1)$$

where $\alpha > 0$ is arbitrary. The values of $\xi^{(1)}$, $\xi^{(2)}$, and $\xi^{(3)}$ are restricted to $(-\infty,0]$, $(0,\infty)$, and $(-\infty,\infty)$, respectively. The right-hand tail of the Y distribution determines which of the three extreme value distributions pertains.

Even if the members of $\{Y_i\}_{i=1}^{m}$ are correlated, their maximum Y_m^* often approaches an extreme value distribution anyway. Empirical fits to the scaling parameters s_m and c_m can then yield appropriate thresholds for statistical significance. In this way, the classical theory provides practical statistical methods for many important database searches [6].

On the other hand, the modern theory emphasizes the rich combinatoric structure of extreme values [3]. An extreme value $[Y_m^* \geq y]$ is a rare exceedance $[Y_i \geq y]$ among m trials $\{Y_i\}_{i=1}^{m}$. If $\{Y_i\}_{i=1}^{m}$ are independent and identically distributed, the number N of exceedances is approximately Poisson distributed, where the Poisson parameter λ equals the expected number of exceedances $mP(Y_1 \geq y)$ (consider m Bernoulli trials, each with a vanishing probability of success $P(Y_1 \geq y) = \lambda m^{-1}$, as $m \to \infty$). The events $[Y_m^* < y]$ and $[N = 0]$ are identical, so $P(Y_m^* < y) = P(N = 0) \approx e^{-\lambda} = \exp[-mP(Y_1 \geq y)]$. Likewise, the events $[Y_{m-i}^* < y, Y_m^* \geq y, \ldots, Y_{m-i+1}^* \geq y]$ and $[N = i]$ are identical, since the i largest Y's exceed y if and only if i of the Y's exceed y. Thus, $P(Y_{m-i}^* < y, Y_m^* \geq y, \ldots, Y_{m-i+1}^* \geq y) = P(N = i) \approx e^{-\lambda}\lambda^i / i!$ (an approximation that can appear somewhat mysterious when the Poisson context is not emphasized [17]).

Modern theory can often put bounds on the error in a Poisson approximation, even when the members of $\{Y_i\}_{i=1}^{m}$ are correlated [7]. Although correlations reduce the effective number of independent trials, the corresponding Poisson parameter sometimes can be interpreted directly as $\lambda = m_{eff} P(Y_1 \geq y)$ (an effective number of independent trials m_{eff} times the probability of an exceedance $P(Y_1 \geq y)$, where m_{eff} is called "the extremal index" [25]).

This paper pursues the extreme value statistics for the biological problem of protein threading [23, 24, 27]. Section 2 gives a simplified description of protein threading, its biological significance, the problem of finding appropriate statistical thresholds, and the combinatorial quantity m associated with the corresponding extreme value statistics. Section 3 defines a combinatorial quantity m' related to m, derives a simple formula for m', and then reinterprets m' in several different ways. Section 3 also shows how m' relates in the continuous limit to an interesting probabilistic formula due to de Finetti [14]. Section 4 then elaborates on the formula for m' until it derives m, the combinatorial quantity associated with the extreme value statistics of protein threading. Although the derivation of m is mostly a combinatorial etude, it shows implicitly how several important probabilities can be derived from de Finetti's formula. The Results section then determines m_{eff} empirically and compares it to m for several different protein sequences and core structures. Finally, the Discussion summarizes the main results and examines the prospects for furthering a combinatorial approach to threading statistics.

2 Protein Threading

Figure 1 near here

In protein threading, one is given a protein sequence, together with a database of protein core structures that may contain the sequence's natural structure [27]. The object of protein threading is to identify correctly the structure(s) corresponding to the sequence [11, 28, 29]. Since the core structures are already associated with specific biological functions, threading has the potential to provide biologists with useful insights about the function of a newly discovered protein sequence.

To develop the intuition underlying a mathematical statement of protein threading (see Figure 1), think of the protein sequence as a sequence of differently colored spherical beads (different amino acids). The beads all have the same radius. (Amino acids have different sizes but in the present context, this is irrelevant.) A flexible string (the protein backbone) goes through the center of each bead. Two knots, one just before the first bead and a second just after the last bead, keep the beads on the string in constant contact.

Next, think of a protein core as a set of tubes that are twisted around each other and then fixed rigidly together in space. (The individual tubes correspond to structure elements like alpha helices and beta sheets that are tightly packed together in a protein core.) The tubes' inner diameters are constant and equal the bead diameter, but their lengths vary and are integer multiples of it. The tubes are numbered sequentially, and one end of each tube is labeled "in"; the other, "out".

The second knot, the one at the end of the string of beads, is then grasped and used to draw the beads through all the tubes sequentially, in the correct order. Once the beads have been threaded through all the tubes, the knots at both ends are elaborated and enlarged, to prevent them from entering the tubes again. Loops of string, along with the beads overlying them, now hang out from between successive tubes. We impose length restrictions on each loop (corresponding to the loop lengths found in actual proteins), and we insist that each bead be either entirely in a loop or entirely in a tube. Otherwise, the loop lengths may be adjusted by pulling on the loops as we please to produce different threading configurations. (The Discussion tinkers somewhat with these restrictions, but they provide a useful preliminary framework for threading statistics.)

Finally, threading forces the beads within the adjacent tightly packed tubes into proximity. For each pair of beads, their position within the tubes and their color determine a pairwise energy of interaction (the attraction or repulsion of the corresponding amino acids [13]). In this model, any bead outside the tubes is irrelevant, and the energy calculations ignore it. The sum of the pairwise interaction energies gives the total energy of each threading configuration. Given a particular string of colored beads (a protein sequence) and set of rigidly fixed tubes (a protein core), the total energy can be minimized over all threading configurations. If this minimum energy is statistically significant as an extreme value, and thus unlikely by chance alone, the protein sequence probably corresponds to the core structure in nature [12].

In mathematical terms, let $A_1, A_2, \ldots, A_t$ represent a given sequence of t bead colors. The i^{th} color in the sequence is A_i $(i = 1,2,\ldots,t)$, where $A_i \in C = \{a,b,c,\ldots\}$, the set of all possible colors. Consider a set of n rigidly fixed tubes and a particular threading configuration through it, and let $X_{n_1} \in \{1,2,\ldots,t\}$ be the sequence number of the first bead just inside the "in" end of the n_1 th tube $(n_1 = 1,2,\ldots,n)$. The color of the bead in position i_1 counted from the "in" end of tube n_1 is therefore $A_{X_{n_1}+i_1-1}$, and each threading configuration corresponds to a set of n sequence numbers $\{X_i\}_{i=1}^n$.

Let the function $E_{n_1 i_1, n_2 i_2} : C \times C \mapsto R$ give the interaction energy between position i_1 in tube n_1 and position i_2 in tube n_2. For beads of color a and b, e.g., let the interaction energy be $-E_{n_1 i_1, n_2 i_2}(a,b)$, where we have introduced a minus so that the extreme values are maxima as in the Introduction, and not minima. The (minus) total interaction energy of the threading configuration $\{X_i\}_{i=1}^n$ is the sum over all the pairwise energies:

$$E(\mathbf{X}) = E(\{X_i\}_{i=1}^n) = \tfrac{1}{2}\sum_{n_1=1}^{n}\sum_{i_1=1}^{l_{n_1}}\sum_{n_2=1}^{n}\sum_{i_2=1}^{l_{n_2}} E_{n_1 i_1, n_2 i_2}(A_{X_{n_1}+i_1-1}, A_{X_{n_2}+i_2-1}), \tag{2}$$

where l_i denotes the length of tube i $(i = 1,2,\ldots,n)$, and "self-energy" terms like $\tfrac{1}{2}E_{n_1 i_1, n_1 i_1}(A_{X_{n_1}+i_1-1}, A_{X_{n_1}+i_1-1})$ represent, e.g., interactions with the surrounding solvent.

Eq (2) elides the spatial nature of the energies, but otherwise it has the same content as the prior intuitive verbiage about beads and tubes.

Constraints on the loop lengths limit the numbers $\{X_i\}_{i=1}^n$ in Eq (2). These constraints could include the number of beads before the first tube or after the last tube, so let $X_0 = 0$, $X_{n+1} = t+1$, and $l_0 = 0$. The constraint $x_i < X_{i+1} - (X_i + l_i) \le x_i'$ permits the i th loop to have any length between $x_i + 1$ and x_i' inclusive $(i = 0,1,\ldots,n)$. For a bead sequence of length t, these constraints permit some number $m = \#_t \{x_i < X_{i+1} - (X_i + l_i) \le x_i'\}_{i=0}^{n}$ of distinct threading configurations.

Let the extreme value statistic corresponding to a particular bead sequence and set of rigidly fixed tubes be $E_0 = \max_{\mathbf{X}} E(\mathbf{X})$, where the maximum is taken over all threading configurations, subject to the loop constraints. To determine the statistical significance of E_0, we need to compare E_0 to extreme value statistics from "random" bead sequences, which we generate as follows [12].

The given bead sequence has $t!$ permutations, all of which match the original sequence for length and color composition. Pick a random sequence uniformly from the permutations, and calculate its extreme value statistic $\max_{\mathbf{X}} E(\mathbf{X})$. Assume that for the random sequence, the energies $E(\mathbf{X})$ from different threading configurations are independent and identically distributed (which they clearly are not). Then according to the Introduction, $P\{\max_{\mathbf{X}} E(\mathbf{X}) < E_0\} \approx e^{-\lambda}$, where $\lambda = mP\{E(\mathbf{X}) \ge E_0\}$. In both of these probabilities, the bead sequence is chosen uniformly from the $t!$ permutations. In the first, however, $E(\mathbf{X})$ is maximized over the m constrained threading configurations, whereas in the second, $\mathbf{X}$ is chosen uniformly from among them.

If the database of protein cores contains N independent structures ($10^2 \le N \le 10^3$ at present), a p-value correction is necessary for testing a protein sequence against so many cores. Because of the multiple testing, if an overall statistical significance of p is desired, the p-value against each individual protein core must be pN^{-1}. Thus, $pN^{-1} = P\{\max_{\mathbf{X}} E(\mathbf{X}) \ge E_0\} \approx 1 - e^{-\lambda} \approx \lambda$, because pN^{-1} (and therefore λ) is small.

With modern computing, Monte Carlo estimation of $P\{\max_{\mathbf{X}} E(\mathbf{X}) \ge E_0\}$ is time-consuming but feasible [12]. The Monte Carlo estimation of $P\{E(\mathbf{X}) \ge E_0\}$, which omits the maximization, is much faster [30]. If the actual number m of threading configurations were known, it could be compared to the effective number $m_{eff} = P\{\max_{\mathbf{X}} E(\mathbf{X}) \ge E_0\} / P\{E(\mathbf{X}) \ge E_0\}$ of independent threading configurations. Regularities between m_{eff} and m could then speed the computation of the statistical significance $P\{\max_{\mathbf{X}} E(\mathbf{X}) \ge E_0\}$ as follows. If m were to determine m_{eff}, the rapid computation of $P\{E(\mathbf{X}) \ge E_0\}$ would then yield the statistical significance $P\{\max_{\mathbf{X}} E(\mathbf{X}) \ge E_0\} = m_{eff} P\{E(\mathbf{X}) \ge E_0\}$. Thus, we are motivated to calculate $m = \#_t \{x_i < X_{i+1} - (X_i + l_i) \le x_i'\}_{i=0}^{n}$ and to compare it to m_{eff}.

Section 3 following gives an explicit formula for a simple combinatorial quantity m' related to m, reinterprets m' in several different ways, and then in the continuous limit derives from m' a probabilistic formula due to de Finetti. Section 4 elaborates on the formula for m' and determines m in terms of m'.

3 The Basic Formula and Its Interpretations

Consider an arbitrary sequence $\{X_i\}_{i=1}^n$ of n integers (in no particular order, with repetitions permitted), with the sentinels $X_0 = 0$ and $X_{n+1} = t+1$ before and after, with t an arbitrary integer. Let $\{L_i = X_{i+1} - X_i\}_{i=0}^n$ be the sequence of differences, some of which may be negative. Given a second arbitrary sequence of integers $\{x_i\}_{i=0}^n$, let $m' = \#_t \{L_i > x_i\}_{i=0}^n$ be the number of distinct sequences $\{X_i\}_{i=1}^n$ satisfying the constraints $L_i > x_i$ $(i = 0,1,\ldots,n)$ (cf. [18, p. 3]). Then

$$m' = \#_t \{L_i > x_i\}_{i=0}^n = \binom{t - x_0 - x_1 - \ldots - x_n}{n} = \frac{(t - x_0 - x_1 - \ldots - x_n)_+^{(n)}}{n!}, \qquad (3)$$

where $x_+^{(n)} = x(x-1)\ldots(x-n+1)$ for $x \geq 0$, and 0 otherwise. This basic formula is somewhat surprising: it is detailed, simple, and general, all at once.

Before proving Eq (3), let us survey some special cases. The first (and thoroughly elementary) case is $x_0 = x_1 = \ldots = x_n = 0$, corresponding to choosing n distinct elements of the set $\{1,2,\ldots,t\}$ in ascending order. Because each choice corresponds to a subset of n elements from $\{1,2,\ldots,t\}$, Eq (3) is correct. The second case is $x_0 = x_n = 0$ and $x_1 = \ldots = x_{n-1} = -1$, choosing n elements of the set $\{1,2,\ldots,t\}$ in ascending order, but with repetitions permitted. The second case is a combinatorial proto-chestnut due to Euler, who showed that the choice can be made in $(t+n-1)_+^{(n)} / n!$ ways [26].

Proof of Eq (3): Consider the 1-1 correspondence between $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ and $\mathbf{X}' = (X_1 - x_0, X_2 - x_0 - x_1, \ldots, X_n - x_0 - x_1 - \ldots - x_{n-1})$. Every integer n-tuple $\mathbf{X}$ satisfying $L_i > x_i \ (i = 0,1,\ldots,n)$ corresponds to an integer n-tuple $\mathbf{X}'$ satisfying $0 < X_1' < X_2' < \ldots < X_n' \le t - x_0 - x_1 - \ldots - x_n$. Since the latter set has cardinality $(t - x_0 - x_1 - \ldots - x_n)_+^{(n)} / n!$, this establishes Eq (3).

Other proofs of Eq (3) are possible. For example, the 1-1 correspondence between $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ and $\mathbf{X}' = (X_1, X_2, \ldots, X_{j-1}, X_j - 1, X_{j+1} - 1, \ldots, X_n - 1)$ shows that $\#_t \{L_i > x_i\}_{i=0}^n = \#_{t-1} \{L_i > x_i - \delta_{ij}\}_{i=0}^n$, where δ_{ij} is Kronecker's delta ($\delta_{ij} = 1$ if $i = j$, and 0 otherwise). This relation gives $\#_t \{L_i > x_i\}_{i=0}^n = \#_{t - x_0 - x_1 - \ldots - x_n} \{L_i > 0\}_{i=0}^n$ after repeated application. Eq (3), however, has already been established when $x_0 = x_1 = \ldots = x_n = 0$.

Similarly, conditioning on X_n gives

$$\#_t \{L_i > x_i\}_{i=0}^n = \sum_{X_n = x_0 + x_1 + \ldots + x_{n-1} + n}^{t - x_n} \#_{X_n - 1} \{L_i > x_i\}_{i=0}^{n-1}, \tag{4}$$

whereas conditioning on X_1 gives

$$\#_t \{L_i > x_i\}_{i=0}^n = \sum_{X_1 = x_0 + 1}^{t - x_1 - \ldots - x_n - n + 1} \#_{t - X_1} \{L_i > x_{i+1}\}_{i=0}^{n-1}. \tag{5}$$

Both recursions lead to an inductive proof of Eq (3), because the formula $\#_t \{L_i > x_i\}_{i=0}^1 = (t - x_0 - x_1)_+^{(1)} / 1!$ for $n = 1$ is easy to establish.

Eq (3) can be reinterpreted in many interesting ways. Choose a sequence of n distinct elements $\{Y_i\}_{i=1}^n$ (in no particular order) from the set $\{1,2,\ldots,t\}$, and let $X_1 < X_2 <\ldots< X_n$ be the corresponding order statistics. Add to them the sentinels $X_0 = 0$ and $X_{n+1} = t+1$, and define the difference sequence $\{L_i = X_{i+1} - X_i\}_{i=0}^n$ as before. Given an arbitrary sequence of integers $\{x_i \ge 0\}_{i=0}^n$, the number of sequences $\{Y_i\}_{i=1}^n$ whose order statistics satisfy the constraints $L_i > x_i$ $(i = 0,1,\ldots,n)$ is $n! \cdot \#_t \{L_i > x_i\}_{i=0}^n = (t - x_0 - x_1 -\ldots- x_n)_+^{(n)}$. (The $n!$-1 correspondence between $\{Y_i\}_{i=1}^n$ and $\{X_i\}_{i=1}^n$ requires distinct $\{Y_i\}_{i=1}^n$, so the restriction $\{x_i \ge 0\}_{i=0}^n$ really is necessary.)

A circular lattice of length $t+1$ yields another interpretation. Consider $t+1$ equally spaced points around a circle, and label them $0,1,\ldots,t$ in counterclockwise order. Starting with $X_0 = 0$ and ending with $X_{n+1} = 0$, let the sequence $\{X_i\}_{i=0}^{n+1}$ progress in "increasing order" (e.g., counterclockwise) around the circular lattice, with repetitions permitted. Let $\{L_i = X_{i+1} - X_i\}_{i=0}^{n-1}$ as usual, but define $L_n = t+1-X_n$ (not $L_n = X_{n+1} - X_n$), so that that the differences are all nonnegative. Given any sequence of integers $\{x_i \ge -1\}_{i=0}^n$ except the sequence $x_0 = x_1 =\ldots= x_n = -1$, the number of distinct sequences $\{X_i\}_{i=1}^n$ satisfying the constraints $L_i > x_i$ $(i = 0,1,\ldots,n)$ is $\#_t \{L_i > x_i\}_{i=0}^n$ from Eq (3). This can be proved by a standard 1-1 correspondence: break the circular lattice open at $X_0 = X_{n+1}$, and place $X_0 = X_{n+1}$ at <u>both</u> ends of the resulting linear lattice. For any sequence $\{X_i\}_{i=1}^n$ except $\{X_i = 0\}_{i=1}^n$ (hence the exclusion of $x_0 = x_1 =\ldots= x_n = -1$), the circular order and placement corresponds uniquely to a linear placement of $\{X_i\}_{i=1}^n$. Moreover, if

$\{x_i \ge 0\}_{i=0}^{n}$, by analogy with the previous paragraph, the number of circular sequences $\{Y_i\}_{i=1}^{n}$ whose "order statistics" $\{X_i\}_{i=1}^{n}$ satisfy the constraints $L_i > x_i$ $(i = 0,1,\ldots,n)$ is again $n! \cdot \#_t \{L_i > x_i\}_{i=0}^{n} = (t - x_0 - x_1 - \ldots - x_n)_{+}^{(n)}$ (circular "order statistics" being taken counterclockwise on the circular lattice, starting from 0). Circular symmetry adds another factor to give $(t - x_0 - x_1 - \ldots - x_n)_{+}^{(n)}(t+1)$ if an element $Y_0 = X_0$ is prefixed to the sequence, placed arbitrarily on the circular lattice, and taken as 0 for ordering the circular order statistics.

Like many other combinatorial results, Eq (3) can be paraphrased probabilistically. The constraints $L_i > x_i$ $(i = 0,1,\ldots,n)$ imply $X_0 + \sum_{j=0}^{i-1} x_i + i \le X_i \le X_{n+1} - \sum_{j=i}^{n} x_i - \left(n - i + 1\right)$ $(i = 1,2,\ldots,n)$, yielding $a' \le X_i \le b'$ $(i = 1,2,\ldots,n)$, where $a' = \min_{i=1,\ldots,n}\left(\sum_{j=0}^{i-1} x_j + i\right)$ and $b' = t - \min_{i=1,\ldots,n}\left(\sum_{j=i}^{n} x_j + n - i\right)$. The set $\{a', a'+1, \ldots, b'\}$ therefore provides a universe for probabilists (i.e., their denominator). The "continuum limit" $X_i = s\Xi_i$, $L_i = s\Lambda_i$, $x_i = s\xi_i$, and $t = s\tau$ with $s \to \infty$ puts the resulting probabilities into a continuous setting. The continuum limit and Eq (3) generalize a beautiful result [14] that de Finetti originally proved by geometric considerations (a combinatorial proof may be more appealing to those who mistrust their geometric intuition). De Finetti's result deserves to be better known, even beyond its appearance in Feller [16, p. 42], because it provides a unified derivation of many important probabilities.

Consider an arbitrary real number τ together with an arbitrary sequence of real numbers $\{\xi_i\}_{i=0}^n$, and let $\alpha \le \alpha' = \min_{i=1,\ldots,n} \sum_{j=0}^{i-1} \xi_j$ and $\tau - \min_{i=1,\ldots,n} \sum_{j=i}^{n} \xi_j = \beta' \le \beta$. Consider a sequence $\{\Xi_i\}_{i=1}^n$ of n independent random variables chosen uniformly from $[\alpha,\beta]$, and add to it the sentinels $\Xi_0 = 0$ and $\Xi_{n+1} = \tau$. Define the difference sequence $\{\Lambda_i = \Xi_{i+1} \text{-} \Xi_i\}_{i=0}^n$, and let $P_\tau\{\Lambda_i > \xi_i\}_{i=0}^n$ be the probability that the sequence $\{\Xi_i\}_{i=1}^n$ satisfies the constraints $\Lambda_i > \xi_i \ (i = 0,1,\ldots,n)$. Note that $\Xi_i \in [\alpha',\beta'] \subseteq [\alpha,\beta]$ $(i = 1,2,\ldots,n)$. With the obvious correspondence, the continuum limit of Eq (3) gives

$$n!\cdot P_\tau\{\Lambda_i > \xi_i\}_{i=0}^n = \lim_{s\to\infty} \frac{(t - x_0 - x_1 - \ldots - x_n)_+^{(n)}}{[s(\beta-\alpha)]^n} = \frac{(\tau - \xi_0 - \xi_1 - \ldots - \xi_n)_+^n}{(\beta-\alpha)^n}, \qquad (6)$$

where $\xi_+^n = \xi^n$ for $\xi \ge 0$, and 0 otherwise.

Eq (6) generalizes de Finetti's result, which he subjected to the restriction $\{\xi_i \ge 0\}_{i=0}^n$. Like Eq (3), the common value in Eq (6) can be reinterpreted in many attractive ways under de Finetti's tight restriction $\{\xi_i \ge 0\}_{i=0}^n$. In de Finetti's interpretation for $[\alpha,\beta] = [0,\tau]$ and $\{\xi_i \ge 0\}_{i=0}^n$, it is the probability that the order statistics $\{\Xi_i\}_{i=1}^n$ for a sequence $\{\mathrm{H}_i\}_{i=1}^n$ chosen uniformly from $[0,\tau]$ satisfy the constraints $\Lambda_i > \xi_i \ (i = 0,1,\ldots,n)$. With the same restrictions, it is the probability that the circular order statistics $\{\Xi_i\}_{i=1}^n$ for a sequence $\{\mathrm{H}_i\}_{i=1}^n$ chosen uniformly from a circle of length τ satisfy the constraints $\Lambda_i > \xi_i \ (i = 0,1,\ldots,n)$ (where $\mathrm{H}_0 = \Xi_0 = 0$, and "circular order statistics" progress counterclockwise on the circle starting from 0). Even if Ξ_0 is permitted to vary around the circle, this probability remains the same because of circular symmetry. In all of these cases, $\{\mathrm{H}_i\}_{i=1}^n$ and the corresponding order statistics $\{\Xi_i\}_{i=1}^n$ are in $n!$-1 correspondence, providing the $n!$ factor in Eq (6).

The next section shows implicitly how several continuous probabilities can be derived methodically from de Finetti's Eq (6). This is accomplished by deriving the corresponding combinatorial results from Eq (3), the combinatorial analog of de Finetti's Eq (6). The next section also derives some combinatorial results related to threading statistics.

4 Consequences of the Basic Formula

Eq (3) yields combinatorial analogs of several important probabilities. Many of these analogs require the generalized inclusion-exclusion formula [18], given here for future reference. Let $\{A_i\}_{i=1}^n$ be a sequence of subsets of some set Ω, and let $\#\{i_1,...,i_k\}$ denote the number of elements in $\bigcap_{j=1}^k A_{i_j}$. The number of elements belonging to exactly k of the sets $\{A_i\}_{i=1}^n$ is

$$\#[k] = \sum_{(j \ge k)} (-1)^{j-k} \begin{pmatrix} j \\ k \end{pmatrix} \sum_{i_1 < ... < i_j} \#\{i_1,...,i_j\}. \tag{7}$$

If $j = 0$ (no restricting condition), $\#\{\;\}$ is defined to be the number of elements of Ω.

Choose n distinct elements $\{Y_i\}_{i=1}^n$ from the set $\{1,2,...,t\}$, and let $X_1 < X_2 <...< X_n$ be the corresponding order statistics. The number of sequences $\{Y_i\}_{i=1}^n$ such that every difference between the consecutive order statistics exceeds $x \ge 0$ is

$$n! \cdot \#_t \{L_0 > 0, L_1 > x,..., L_{n-1} > x, L_n > 0\} = [t-(n-1)x]_+^{(n)}, \tag{8}$$

because Eq (3) with $x_0 = x_n = 0$ and $x_1 =...= x_{n-1} = x$ counts the required number of ascending n-tuples satisfying the constraint. In the continuum limit (with $x = s\xi$), Eq (8) yields the probability that when n points are chosen uniformly on a line segment of length τ, no two are closer than ξ [22, p. 132].

The recursion Eq (4) applied repeatedly to itself gives

$$\#_t \{L_i > x_i\}_{i=0}^n = \sum_{X_n = x_0 + x_1 + ... + x_{n-1} + n}^{t - x_n} \quad \sum_{X_{n-1} = x_0 + x_1 + ... + x_{n-2} + n - 1}^{X_n - 1 - x_{n-1}} \cdots \sum_{X_1 = x_0 + 1}^{X_2 - 1 - x_1} 1 \,. \tag{9}$$

This repeated summation, when carried through explicitly, gives another proof of Eq (3). In the continuum limit, a similar but simpler repeated integral proves the continuous analog of Eq (8) directly [22, p. 132]. Similar comments apply to the recursion Eq (5), which when applied repeatedly to itself gives

$$\#_t \{L_i > x_i\}_{i=0}^n = \sum_{X_1 = x_0 + 1}^{t - x_1 - ... - x_n - n + 1} \quad \sum_{X_2 = x_1 + 1}^{t - X_1 - x_2 ... - x_n - n + 2} \cdots \sum_{X_n = x_{n-1} + 1}^{t - X_1 - ... - X_{n-1} - x_n} 1 \quad . \tag{10}$$

Now choose $n+1$ distinct points $\{Y_i\}_{i=0}^n$ from $t+1$ points on a circular lattice, and after starting at Y_0, progressing counterclockwise, and ending at Y_0, specify j of the distances between consecutive circular order statistics of $\{Y_i\}_{i=0}^n$. The number of distinct choices making those j distances greater than $x \ge 0$ is

$$n! \cdot \#_t \{L_i > x \; (i = i_1, ..., i = i_j); \; L_i > 0 \text{ otherwise}\} \cdot (t+1) = (t - jx)_+^{(n)} (t+1) \,. \tag{11}$$

In the continuum limit above, Eq (11) yields a related probability [22, p. 132].

In the context of Eq (11) and the generalized inclusion-exclusion formula of Eq (7), let $\Omega = \{L_i > 0\}_{i=0}^n$ and $A_i = \Omega \cap \{L_i > x\}$. Above, the number of distinct choices making exactly k of the consecutive distances greater than $x \ge 0$ is

$$\#[k] = \sum_{j=k}^{n+1} (-1)^{j-k} \binom{j}{k} \binom{n+1}{j} (t - jx)_+^{(n)} (t+1) \,. \tag{12}$$

The specialization $k = 0$ gives the number of distinct choices making all the consecutive distances less than or equal to x. Eq (12) and the specialization $k = 0$ both yield standard probability results in the continuum limit [16, p. 28, 22, p. 132].

Results on the line resemble those on the circle. We now confine the discourse to choosing n distinct elements $\{Y_i\}_{i=1}^{n}$ from $\{1,2,...,t\}$, as in protein threading. Let $X_1 < X_2 <...< X_n$ be the corresponding order statistics, with sentinels $X_0 = 0$ and $X_{n+1} = t + 1$, and difference sequence $\{L_i = X_{i+1} - X_i\}_{i=0}^{n}$, as usual. From the inclusion-exclusion formula and Eq (3), the number of sequences $\{Y_i\}_{i=1}^{n}$ satisfying the constraints $\{0 < L_i \le x_i\}_{i=0}^{n}$ is

$$n! \ \cdot \ \#_t \ \{0 < L_i \le x_i\}_{i=0}^{n} = \#[0] = \sum_{j=0}^{n+1} (-1)^j \sum_{(0 \le i_1 <...< i_j \le n)} (t - x_{i_1} - ... - x_{i_j})_+^{(n)} . \qquad (13)$$

Since we have confined the discourse to <u>distinct</u> elements $\{Y_i\}_{i=1}^{n}$ from $\{1,2,...,t\}$, our notation will now omit vacuous constraints like $0 < L_i \le t$. If $x_1 =...= x_{n-1} = x$ with $x_0 = x_n = t$ in Eq (13), i.e.,

$$n! \ \cdot \ \#_t \ \{0 < L_i \le x\}_{i=1}^{n-1} = \#[0] = \sum_{j=0}^{n-1} (-1)^j \binom{n-1}{j} (t - jx)_+^{(n)} . \qquad (14)$$

The corresponding continuum probability is a standard result [16, p. 28].

As in Eqs (4) and (5), recursions determine the common value in Eq (13):

$$\begin{aligned}\#_t\,\{0<L_i\le x_i\}_{i=0}^{n} &= \sum_{X_n=t+1-x_n}^{t}\#_{X_n-1}\,\{0<L_i\le x_i\}_{i=0}^{n-1} \\ &= \sum_{X_n=t+1-x_n}^{t}\ \sum_{X_{n-1}=X_n-x_{n-1}}^{X_n-1}\cdots\sum_{X_1=X_2-x_1}^{X_2-1}1\end{aligned}\quad . \tag{15}$$

and

$$\begin{aligned}\#_t\,\{0<L_i\le x_i\}_{i=0}^{n} &= \sum_{X_1=1}^{x_0}\#_{t-X_1}\,\{0<L_i\le x_{i+1}\}_{i=0}^{n-1} \\ &= \sum_{X_1=1}^{x_0}\ \sum_{X_2=X_1+1}^{X_1+x_1}\cdots\sum_{X_n=X_{n-1}+1}^{X_{n-1}+x_{n-1}}1\end{aligned}\quad . \tag{16}$$

We can now calculate the number of constrained threading configurations $m =\#_t\,\{x_i < X_{i+1}-(X_i+l_i)\le x_i'\}_{i=0}^{n} =\#_t\,\{x_i+l_i < L_i \le x_i'+l_i\}_{i=0}^{n}$. The 1-1 correspondence between $\mathbf{X}=(X_1,X_2,\ldots,X_n)$ and $\mathbf{X}'=(X_1-x_0,X_2-x_0-x_1,\ldots,X_n-x_0-x_1-\ldots-x_{n-1})$ (where $L_i'=X_{i+1}'-X_i'$; $i=0,1,\ldots,n$) shows

$$\#_t\,\{x_i<L_i\le x_i'\}_{i=0}^{n} =\#_{t-x_0-x_1-\ldots-x_n}\,\{0<L_i'\le x_i'-x_i\}_{i=0}^{n}, \tag{17}$$

Thus, Eqs (13), (15), and (16) provide practical formulas for computing m rapidly.

5 Results

Figure 2 near here

Figure 2 shows a log-log plot for the loop-constrained threading of several different protein sequences and core structures. The effective number of threading configurations $m_{eff} = P\{\max_{\mathbf{X}} E(\mathbf{X}) > E_0\} / P\{E(\mathbf{X}) > E_0\}$ as estimated by Monte Carlo simulation (Y-axis) is plotted against actual number of threading configurations $m = \#_s \{x_i < X_{i+1} - (X_i + l_i) \leq x_i'\}_{i=0}^{n}$ as calculated by Eqs (13)-(17) (X-axis). There are some outliers, but the empirical relationship $m_{eff} \approx 10^{-2.5} m$ is surprisingly consistent.

For a particular random sequence and rigid core structure, if all the threading configurations were independent, the empirical relationship would be $m_{eff} = m$. The empirical discrepancy indicates that the energies of different threading configurations are correlated, even for random sequences. First, correlations can be caused by periodicities in protein core structures. Periodic interactions abound in nature (e.g., pairs of alpha helices and beta strands, either parallel or anti-parallel). Consider an extreme case (period 1), where two long "rigid tubes" (e.g., beta strands) are fixed parallel and side-to-side along their length. Pulling the beads through them in parallel may generate new threading configurations, but it changes the bead interactions very little. Second, correlations can be caused by structure elements that do not interact strongly with other parts of the protein core. The corresponding rigid tubes can be placed arbitrarily without much influence on the total interaction energy. These two effects conspire to introduce correlations between threading configurations, so they lower m_{eff} relative to m.

6 Discussion

The extreme value statistics of protein threading suggest a combinatorial problem, calculating the number of ways $m' = \#_t \{L_i > x_i\}_{i=0}^{n}$ of choosing a sequence $\{X_i\}_{i=1}^{n}$ from the set $\{1,2,\ldots,t\}$, subject to the difference constraints $\{L_i = X_{i+1} - X_i > x_i\}_{i=0}^{n}$, where $X_0 = 0$ and $X_{n+1} = t+1$, and $\{x_i\}_{i=0}^{n}$ is an arbitrary sequence of integers. The quantity m' has many attractive combinatorial interpretations and is closely related to a probabilistic formula discovered by de Finetti. Just as many important probabilities can be derived from de Finetti's formula, many interesting combinatorial quantities can be derived from m'.

In particular, m' is involved in many combinatorial problems related to protein-threading statistics. These statistics involve constraints $\{x_i < X_{i+1} - (X_i + l_i) \le x_i'\}_{i=0}^{n}$ on the loop lengths, and Section 4 calculates the corresponding combinatorial quantity $m = \#_t \{x_i < X_{i+1} - (X_i + l_i) \le x_i'\}_{i=0}^{n}$ in terms of m'.

These loop constraints are somewhat arbitrary. To increase statistical power, e.g., one might try constraining only the internal loop lengths, along with their sum. This yields the following combinatorial problem: what is the number of ways of choosing a strictly increasing sequence $\{X_i\}_{i=1}^{n}$ from $\{1,2,\ldots,t\}$, subject to the restrictions $X_n - X_1 = s$ and $x_i < L_i \le x_i'$ $(i = 1,\ldots,n-1)$? The answer is

$$\#_t \{X_n - X_1 = s, x_i < L_i \le x_i' \, (i = 1,\ldots,n-1)\} = (t-s) \cdot \#_{s-1} \{x_{i+1} < L_i \le x_{i+1}'\}_{i=0}^{n-2}, (18)$$

where the right side can be found in Section 4. The proof follows: X_1 can be chosen in $t-s$ ways (which then fixes X_n). The sentinels X_1 and X_n that bracket $\{X_i\}_{i=2}^{n-1}$ are s apart, and since the sequence $\{L_i = X_{i+2} - X_{i+1}\}_{i=0}^{n-2}$ satisfies the difference constraints on the right above, this proves Eq (18). Obviously, many other combinatorial problems in protein threading can be solved systematically within the framework of this paper.

The extreme value statistics of protein threading would have been completely solved if the empirical relationship in Figure 2 were $m_{eff} = m$ instead of $m_{eff} \approx 10^{-2.5} m$. Although the relationship $m_{eff} \approx 10^{-2.5} m$ might be consistent enough for some purposes, the factor $10^{-2.5}$ displays a confounding effect from energetic correlations between different threading configurations. These correlations probably reflect natural periodicities found in protein structures and are currently under investigation.

Acknowledgements: The authors would like to thank Dr. W.J. Wilbur for several useful conversations.

References

1. E.E. Abola *et al.*, Protein Data Bank, in: Crystallographic Databases - Information Content, Software Systems, Scientific Applications, F.H. Allen, G. Bergerhoff, and R. Sievers, eds., International Union of Crystallography, Bonn, 1987, pp. 107-132.

2. E.E. Abola *et al.*, Protein Data Bank archives of three-dimensional macromolecular structures, Methods Enzymol, **277 (**1997) 556-71.

3. D. Aldous, Probability Approximations via the Poisson Clumping Heuristic, F. John, J.E. Marsden, and L. Sirovich, eds., Applied Mathematical Sciences, vol. 77, Springer-Verlag, New York, 1989.

4. S.F. Altschul *et al.*, Issues in searching molecular sequence databases, Nat Genet, **6 (**1994) 119-29.

5. S.F. Altschul *et al.*, Basic local alignment search tool, J Mol Biol, **215 (**1990) 403-10.

6. S.F. Altschul *et al.*, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Res, **25 (**1997) 3389-402.

7. R. Arratia, L. Goldstein, and L. Gordon, Two moments suffice for Poisson approximations: The Chen-Stein method, The Annals of Probability, **17 (**1989) 9-25.

8. A. Bairoch and R. Apweiler, The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1998, Nucleic Acids Res, **26 (**1998) 38-42.

9. A. Bairoch, P. Bucher, and K. Hofmann, The PROSITE database, its status in 1997, Nucleic Acids Res, **25 (**1997) 217-21.

10. D.A. Benson *et al.*, GenBank, Nucleic Acids Res, **26 (**1998) 1-7.

11. S.H. Bryant, Evaluation of threading specificity and accuracy, Proteins, **26 (**1996) 172-85.

12. S.H. Bryant and S.F. Altschul, Statistics of sequence-structure threading, Curr Opin Struct Biol, **5 (**1995) 236-44.

13. S.H. Bryant and C.E. Lawrence, An empirical energy function for threading protein sequence through the folding motif, Proteins, **16 (**1993) 92-112.

14. B. de Finetti, Alcune osservazioni in tema di "suddivisione casuale", Giornale Istituto Italiano degli Attuari, **27 (**1964) 151-173.

15. A. Dembo, S. Karlin, and O. Zeitouni, Limit distribution of maximal non-aligned two-sequence segmental score, The Annals of Probability, **22 (**1994) 2022-2039.

16. W. Feller, An Introduction to Probability Theory and its Applications, vol. 2, Wiley, New York, 1971.

17. J. Galombos, The Asymptotic Theory of Extreme Order Statistics, Wiley, New York, 1978.

18. M. Hall, Combinatorial Theory, Blaisdell, Waltham MA, 1967.

19. C.W. Hogue, H. Ohkawa, and S.H. Bryant, A dynamic look at structures: WWW-Entrez and the Molecular Modeling Database, Trends Biochem Sci, **21 (**1996) 226-9.

20. S. Karlin and S.F. Altschul, Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes, Proc Natl Acad Sci U S A, **87 (**1990) 2264-8.

21. S. Karlin and A. Dembo, Limit distributions of maximal segmental score among Markov-dependent partial sums, Adv Appl Prob, **24 (**1992) 113-140.

22. S. Karlin and H.M. Taylor, A Second Course in Stochastic Processes, Academic Press, New York, 1981.

23. R.H. Lathrop, The protein threading problem with sequence amino acid interaction preferences is NP-complete, Protein Eng, **7 (**1994) 1059-68.

24. R.H. Lathrop and T.F. Smith, Global optimum protein threading with gapped alignment and empirical pair score functions, J Mol Biol, **255 (**1996) 641-65.

25. R. Leadbetter and H. Rootzen, Extremal theory for stochastic processes, Annals of Probability, **16 (**1988) 431-478.

26. C.L. Liu, Introduction to Combinatorial Mathematics, McGraw-Hill, New York, 1968.

27. T. Madej, J.F. Gibrat, and S.H. Bryant, Threading a database of protein cores, Proteins, **23 (**1995) 356-69.

28. A. Marchler-Bauer and S.H. Bryant, A measure of success in fold recognition, Trends Biochem Sci, **22 (**1997) 236-40.

29. A. Marchler-Bauer and S.H. Bryant, Measures of threading specificity and accuracy, Proteins, **Suppl (**1997) 74-82.

30. W.J. Wilbur, Accurate Monte Carlo estimation of very small p-values in Markov chains, Computational Statistics, **in press (**1998) .

Captions for the Figures

Figure 1: A Schematic Representation of Protein Threading.

Figure 1 displays a protein sequence as a string of colored beads, here in the two colors, black and white. It also displays a protein core structure as 4 horizontal “tubes”, the rectangles in heavy outline. The beads have been threaded through the tubes. Threading has forced the beads within the adjacent tightly packed tubes into proximity. The resulting energy interactions between pairs of beads are indicated by the vertical rectangles. If the two beads have the same color, the corresponding rectangle is black, indicating one particular strength of interaction; if they have different colors, it is white, indicating another. The arrows indicate the possibility of pulling the beads through the tubes. Note, e.g., if one pulls the middle loop on the left and advances the beads in both middle tubes leftward one position, this changes only two interactions in the middle tubes: the one on the left disappears, and a new one appears on the right. Thus, the bead interactions in different threading configurations may be highly correlated.

Figure 2: A Plot of $m_{eff} = \#\{\text{effective}\}$ against $m = \#\{\text{combinatorial}\}$ (Unavailable electronically).

Figure 2 plots the effective number of threading configurations $m_{eff} = P\{\max_{\mathbf{X}} E(\mathbf{X}) > E_0\} / P\{E(\mathbf{X}) > E_0\}$ as estimated by Monte Carlo simulation (Y-axis) against actual number of threading configurations $m = \#_s \{x_i < X_{i+1} - (X_i + l_i) \le x_i'\}_{i=0}^{n}$ as calculated by Eqs (13)-(17) (X-axis). There are some outliers, but the empirical relationship $m_{eff} \approx 10^{-2.5} m$ is surprisingly consistent.