

Statistical Significance of Patterns in Biosequences

Gustavo Stolovitzky

Andrea Califano

*IBM Computational Biology Center
TJ Watson Research Center
PO Box 704, Yorktown Heights, NY 10598*

Abstract

Several algorithms for pattern discovery in sequences have recently emerged, such as Pratt, Teiresias, Splash, and Meme. These are proving useful in biological sequences analysis, as the presence of statistically or biologically significant patterns may indicate subtle relationships that complement those discovered through local sequence alignment tools such as BLAST or FASTA [1, 10].

Statistical tools to assess the statistical relevance of maximal, non-contiguous data patterns are still largely undeveloped [9]. Rather, most recent efforts [8] have concentrated on the analysis of the statistical relevance of contiguous k-tuple repeats.

The goal of this paper, then, is twofold. First statistical analysis tools are developed to determine how likely arbitrary, non-contiguous data patterns are to occur in random proteins and genomes. This is accomplished by modeling the various properties of pattern discovery algorithms such as pattern maximality and pattern density constraints.

Second, a framework is proposed to pinpoint the truly relevant patterns, among the very large number that occur even in small biological datasets. By relevant, in this context, we mean those patterns that reflect a relevant underlying stochastic or deterministic process that deviates from totally random behavior.

Some biological examples are studied. In the first one, patterns are extracted from two proteins (HUMPCRA and MUSHEPGFA). By comparing the results with the theoretical analysis, it is determined that all the patterns discovered in HUMPCRA are statistically irrelevant while a few patterns in MUSHEPGFA are significant. In particular, one of these is likely to occur only once every 10^{40} comparable sequences. An indication that this pattern may be biologically relevant is given by the fact that it includes a known motif, reported in the PROSITE database. In the second example, we analyze a region of DNA of 1000 nucleotides from yeast chromosome 1. Only 6 of the discovered patterns are statistically significant and they all occur in a tandem repeat region.

1. Introduction

Whenever Nature finds a “recipe” to accomplish a task that gives a differential fitness to an organism, chances are that such recipe will be conserved through evolution. At the molecular level, this means that biological sequences, belonging sometimes to widely distant species, will share common motifs. Examples of these motifs are the consensus sequences for transcription promoters (e.g. TATA boxes), sets of a small number of amino acid that determine the active site of families of enzymes, etc.

These motifs need not be the result of evolutionary processes. Some molecular mechanisms that are neutral from an evolutionary perspective may be at work, producing patterns that repeat themselves many times within a given genome. This is manifested, for instance, by Tandem Repeats in many eukaryotic species or by repetitive elements, such as Alu sequences or other types of retroposons.

Thus, the identification of patterns in biological databases is becoming a very transited venue within the bioinformatics community. Pattern databases such as PROSITE [3] are examples of this trend.

In recent years a number of very interesting pattern discovery tools have emerged (Teiresias [11], Pratt [7], Meme [2], Splash [4]), that find all the maximal patterns of constrained local density within a sequence or set of sequences in an unsupervised way. These algorithms are valuable in that they can discover weak motifs, hidden in the biosequences. One difficulty, is that biological relevant patterns discovered in this way will be hidden into a much larger number of meaningless patterns. Thus the need arises to assess at least the statistical relevance of the discovered patterns.

One example will help us bring home our main point: if we run one of these algorithms on a random sequence, the number of discovered patterns will depend, obviously, on the size of the sequence and the letter distribution within the sequence. Surprisingly, however, even for very short sequences the number is very large, as we shall later show. Therefore, if a random sequence hosts a great number of patterns which are irrelevant by construction, we can surmise that a biological sequence will also host, along with its relevant ones, a myriad of other patterns that are bound to be there by mere chance aggregation of letters.

There are two main goals that we wish to accomplish in this paper. The first is to characterize the statistics of maximal patterns of constrained local density in random sequences. The second is to devise a statistical test that allows us to pinpoint, given a biosequence, which patterns are statistically relevant. By the latter we mean, which are the patterns that are not expected to occur in a random sequence of length and composition similar to that given biosequence.

Some of the math involved in the calculations are rather classical. However, the need to account for some important properties of the patterns such as maximality and local density constraint make some of the mathematics a little more involved, and worth presenting in detail.

2. Definitions and Notation

Let us denote by $s = \{v_\omega | \omega = 0, \dots, L-1\}$ a string of L values. The values v_ω (that we shall also call tokens) are taken from a given alphabet Σ , which in concrete examples can be the four bases A,C,G,T or the twenty amino acids. More generally, Σ could be a continuous set, such as an interval on the real line. We define a pattern π as a set of (value, relative offset) pairs $\{(v_i, \delta_i) | \delta_0 = 0; i = 0, \dots, k-1\}$. A pattern that contains exactly k values will be called a k -pattern. A pattern that contains exactly k values, and whose span is l (i.e., the last relative offset is $\delta_{k-1} = l-1$) will be called a kl -pattern.

A convenient notation for patterns is that in which the values are put in a one dimensional string, at the locations given by their relative offsets. For the relative offsets whose values are not specified in the pattern we use the wild character “.”. For instance, for the alphabet $\Sigma = \{A, C, G, T\}$, the pattern $\pi = \{(A, 0), (C, 3), (G, 7)\}$ is a kl -pattern, with $k = 3$ and $l = 8$ which can be written as $\pi = A..C...G$.

Given a distance metric¹ $f(x, y)$, and a distance threshold D , we shall say that a k -pattern matches at offset ω in s , if for each $0 \leq i \leq k-1$, the relation $f(v_i, v_{\omega+\delta_i}) \leq D$ holds true. The locus $\lambda_\pi = \{\omega_m\}$ of a given pattern π is then the set of absolute offsets where the pattern π matches in s . A kl -pattern that matches at j different offsets on s will be called a jkl -pattern. Then, j is the cardinality of the locus.

A jkl -pattern is said to be *maximal in composition* if it cannot be extended to a $j(k+1)l$ -pattern, by adding an extra token within its span, without simultaneously decreasing the cardinality j of its locus. A jkl -pattern is said to be *maximal in length* if it cannot be extended to a $j(k+1)l'$ -pattern (with $l' > l$), by adding an extra token outside its original span l , without simultaneously decreasing the cardinality j of its locus. A *maximal pattern* is a pattern that is both maximal in composition and in length. Maximality is an essential property of pattern discovery algorithms. It avoids reporting a combinatorial number of subpatterns of each maximal pattern in the sequence.

We shall refer to the set of k relative offsets in a k -pattern $\psi = \{\delta_i | i = 0, \dots, k-1\}$ as a k -comb. If the span of a k -comb is l (with $l = \delta_{k-1} + 1$), the comb will be called a kl -comb. It is sometimes convenient to represent a kl -comb with a string of 0s and 1s that has a 1 at each relative offset δ_i and a 0 elsewhere.

1. For an example of a distance metric relevant in computational biology, see Appendix A.

Then, for instance, the comb $\psi_5 = \{0, 3, 4, 7, 11\}$ can be represented by the string 100110010001. By π_ψ we shall indicate a pattern supported by the comb ψ . By $\pi_{\psi\omega}$ we shall indicate the pattern identified by placing the comb ψ at the absolute offset ω in s .

It is also useful to control the minimum number of tokens allowed in any given window within a pattern. This can be used to eliminate growing any pattern to arbitrary lengths by means of a few sparse random matches. We shall refer to such constraint as the *density constraint*. Given k_0 and l_0 , we shall say that a comb is a $\langle k_0, l_0 \rangle$ -valid comb (valid in the sense of satisfying the constraint in density), if for all $0 \leq i \leq k - k_0$, the relation $\delta_{i+k_0-1} - \delta_i \leq l_0$ is valid. A pattern π_ψ , supported by a $\langle k_0, l_0 \rangle$ -valid comb is a $\langle k_0, l_0 \rangle$ -valid pattern.

3. Pattern Statistics

3.1. The Basic ingredients

When dealing with long sequences, it is important to have a statistical model to understand the relevance of the observed patterns. For instance, given a sequence of 100 English-alphabet characters chosen from a uniform distribution, is a pattern of 5 characters that repeats 4 times more or less probable than one of 4 characters that repeats 5 times? Or, How does the expected number of $ijkl$ -maximal $\langle k_0, l_0 \rangle$ -valid patterns behave as a function of the sequence length L in random sequences? In this paper we shall construct the theory that allows us to answer this basic kind of questions.

We shall be dealing with random sequences whose values at different offsets are independent random variables, and whose value at each absolute offset is chosen from a discrete alphabet $\Sigma = \{v_i\}$ of σ elements. The probability that the value v_i be chosen is $p(v_i)$. If all the letters in Σ are equiprobable, then clearly $p(v_i) = 1/\sigma$. The first step is to compute the probability that a given value v_j matches at a particular offset on s . The probability of such match is

$$\wp(v_j) = \sum_{i=1}^{\sigma} p(v_i) \alpha_{ij}, \quad (3.1)$$

where α_{ij} is 1 if $f(v_i, v_j) \leq D$ and 0 otherwise. If the values are from a continuous range, then the summation becomes an integral. If D is sufficiently small $\wp(v_j) = p(v_j)$. For later reference we shall denote by $\langle \wp^t \rangle$ the average value of the function $\wp(v)^t$, that is,

$$\langle \wp^t \rangle = \sum_{j=1}^{\sigma} p(v_j) [\wp(v_j)]^t. \quad (3.2)$$

Only when all the $\wp(v_j)$ are the same, i.e. $\wp = \wp(v_j)$, will $\langle \wp^t \rangle = \wp^t$.

Let us now choose a specific kl -comb, ψ and an absolute offset ω in s . We wish to compute the probability that the pattern $\pi_{\psi\omega}$, starting at offset ω in s , matches at exactly j other offsets after ω in s . The probability of such pattern matching at any given offset in s is

$$\rho_{\pi_{\psi\omega}} = \prod_{i=0}^{k-1} \wp(v_i), \quad (3.3)$$

where the v_i are the values that comprise $\pi_{\psi\omega}$. If all the $\wp(v_i) = \wp$ are the same, then $\rho_{\pi_{\psi\omega}} = \wp^k$. Assuming that potential matches of $\pi_{\psi\omega}$ in s are uncorrelated and given that for each kl -comb there are $L - \omega - l$ locations where a match can occur after offset ω , the probability that the kl -pattern $\pi_{\psi\omega}$ will match at exactly j offsets greater than ω is given by the binomial distribution.

$$\mathfrak{S}(j, \pi_{\psi\omega}) = \binom{L - \omega - l}{j} \rho_{\pi_{\psi\omega}}^j [1 - \rho_{\pi_{\psi\omega}}]^{L - \omega - l - j}. \quad (3.4)$$

3.2. The Probability that $\pi_{\psi\omega}$ is a $(j+1)kl$ -pattern

In order to find the probability that the pattern $\pi_{\psi\omega}$ has occurred exactly $j+1$ times in s (i.e., exactly j times after its first occurrence), we must multiply the previous expression by the probability that $\pi_{\psi\omega}$ does not match at any offsets smaller than ω in s . This is simply:

$$p_b(\pi_{\psi\omega}) = [1 - \rho_{\pi_{\psi\omega}}]^\omega. \quad (3.5)$$

3.3. The Maximality in Composition Constraint

As discussed, it is important that a pattern discovery algorithm guarantees the maximality of the patterns detected, see definition in Section 2. If we try to model this property for a pattern $\pi_{\psi\omega}$, then for each relative offset δ that is *not* part of ψ , there must be at least an absolute offset ω' ($\omega' > \omega$) in λ_π such that $f((v_{\omega+\delta}, v_{\omega'+\delta}) > D)$. This probability can be computed as one minus the probability that the value $v_\delta = v_{\omega+\delta}$ matches at each offset $\omega' + \delta$ in s , with ω' in λ_π . The latter is $\wp(v_{\omega+\delta})^j$. Thus the probability $p_{in}(j, \pi_{\psi\omega})$, that the pattern $\pi_{\psi\omega}$ is maximal in composition, is the product of the probabilities that it cannot be extended at any of the $l-k$ relative offsets δ ($0 < \delta < l-1$) which are not part of the kl -comb ψ . This is expressed as:

$$p_{in}(j, \pi_{\psi\omega}) = \prod_{\delta \notin \psi} [1 - \wp(v_\delta)^j], \text{ with } 0 < \delta < l-1. \quad (3.6)$$

3.4. Density-Constrained Combs

Given a span l and a size k , let us determine how many combs, $N_0(k, l)$, can be formed that satisfy the $\langle k_0, l_0 \rangle$ constraint, that is, $N_0(k, l)$ is the number of $\langle k_0, l_0 \rangle$ -valid kl -combs. If we relate to the binary string representation for combs, this can be rephrased as: how many binary strings of length l with exactly k 1s, and starting and ending in a 1 can be formed such that no k_0 consecutive tokens span a window larger than l_0 .

For $k_0 = 2$, it can be shown that this can be computed exactly as the number of $(k-1)$ -parts compositions of the number, $l-k$, of empty offsets in the pattern into integers smaller than $k_0 - 1$. The general solution for arbitrary k_0 and l_0 is significantly more complex. In any case, a fairly accurate approximation can be obtained by the following recursive expression:

$$N_0(k, l) = \sum_{i=k_0}^{l_0} \binom{i-2}{k_0-2} N_0(k-k_0+1, l-i+1), \quad (3.7)$$

with the following set of boundary conditions:

$$N_0(k, l) = 0, \text{ if } (l < k) \vee (k \leq 0) \vee (l \leq 0) \vee (l > l_{max}) \vee (k < k_0) \quad (3.8)$$

$$N_0(k, l) = 1, \text{ if } l = l_{max}$$

$$N_0(k, l) = \binom{l-2}{k-2}, \text{ otherwise}$$

$$\text{where } l_{max} = (\sigma + 1)(l_0 - 1) - (k_0 - 2) + \tau + 1; \sigma = \text{Int}\left(\frac{k-2}{k_0-1}\right); \tau = \text{Mod}\left(\frac{k-2}{k_0-1}\right). \quad (3.9)$$

l_{max} , in this case, is the maximum possible length l that a $\langle k_0, l_0 \rangle$ -valid kl -comb can attain.

For $k_0 \geq 3$ the recursion for $N_0(k, l)$ is only approximated because its iterative formulation only imposes the density constraint starting at tokens that are a multiple of $k_0 - 1$. For $k_0 = 3$ and $k = 8$, for instance, the constraint is only applied to token groups $\{1, 2, 3\}$, $\{3, 4, 5\}$, $\{5, 6, 7\}$, $\{7, 8\}$, while it should be applied to all groups $\{1, 2, 3\}$, $\{2, 3, 4\}$, ..., $\{6, 7, 8\}$. The approximation, however, is quite good.

3.5. The Maximality in Length Constraint

The length maximality constraint strongly depends on the density constraint. In fact, a jkl -pattern can also be extended if k_0 tokens, including at least one outside the comb (before its first or after its last 1), in a window of size l_0 , match simultaneously at all j locations. This probability can be approximated as:

$$p_{out}(j, \pi_{\psi\omega}) \cong \sum_{n=0}^{k_0-2} \left\{ \sum_{\substack{1 \leq i_1 < i_2 < \dots \\ \dots < i_n \leq l_0-1}} \left[\prod_{\substack{\vartheta = i_1 \\ \dots, i_n}} \wp(v_{l-1+\vartheta})^j \right] \left[\prod_{\substack{1 \leq \iota \leq l_0-1 \\ \iota \neq i_1, \dots, i_n}} [1 - \wp(v_{l-1+\iota})^j] \right] \right. \\ \left. \times \sum_{m=d(l_0-1-i_n)}^{k_0-2-n} \Phi(m|l_0-i_n-1) \right\} \quad (3.10)$$

where $d = (k-2)/(l-2)$ is the mean token density inside the pattern, and

$$\Phi(m|w) = \begin{cases} \binom{w}{m} d^m (1-d)^{w-m} & \text{if } w < l-2 \\ \delta_{m, l-2} & \text{if } w = l-2 \\ \delta_{m, l-1} & \text{if } w = l-1 \end{cases} \quad (3.11)$$

is the probability of having exactly m tokens in a window of size w within the pattern. For further details on Equation (3.10), which will get considerably simplified in the next sections, see Appendix B.

4. Average number of $\langle k_0, l_0 \rangle$ -valid jkl -maximal patterns in a string

In this Section we shall compute the average number of different jkl -maximal patterns that satisfy the density constraint in a string s of length L . That number is the average of a random variable. For each realization of the string s , that is, the number of patterns will be different and distributed with a probability function that we shall estimate in Section 6., on page 7.

Let us first write the number v_{jkl} of jkl -maximal patterns as the sum over all $\langle k_0, l_0 \rangle$ -valid kl -combs ψ and absolute offsets ω , of the random variable $B(j, \pi_{\psi\omega})$ which is 1 if $\pi_{\psi\omega}$ is a jkl -maximal pattern not matching at any offsets smaller than ω , and 0 otherwise. That is,

$$v_{jkl} = \sum_{\substack{\langle k_0, l_0 \rangle\text{-valid} \\ kl\text{-combs } \psi}} \sum_{\omega=0}^{L-l+1} B(j, \pi_{\psi\omega}). \quad (4.1)$$

This analysis is carried out in details in Appendix C, where we obtain the following closed formula for the average number of different, density constrained, jkl -maximal patterns:

$$\langle v_{jkl} \rangle = N_0(k, l) \binom{L-l+1}{j} \left(\sum_{i=0}^{L-l-1-j} \binom{L-l-1-j}{i} (-1)^i \langle \wp^{j-1+i} \rangle^k \right) \langle p_{in} \rangle \langle p_{out} \rangle^2. \quad (4.2)$$

$$\text{where } \langle p_{in} \rangle = [1 - \langle \wp^{j-1} \rangle]^{l-k}, \quad (4.3)$$

$$\text{and } \langle p_{out} \rangle = \sum_{n=0}^{k_0-2} \langle \wp^{j-1} \rangle^n [1 - \langle \wp^{j-1} \rangle]^{l_0-n-1} \sum_{i=n}^{l_0-1} \binom{i-1}{n-1} \sum_{m=d(l_0-i-1)}^{k_0-n-2} \Phi(m|l_0-i-1) \quad (4.4)$$

Finally, the average number of maximal k -patterns that appear j times in s , irrespective of their span l , is obtained by summing $\langle v_{jkl} \rangle$ over all the allowable values for l between the minimum $l_{min} = k_0$ and the maximum, l_{max} , defined in Equation (3.9). The results of this calculation for different values of the parameters are shown in Fig. 1. It can be observed that for the chosen values of parameters, even short random

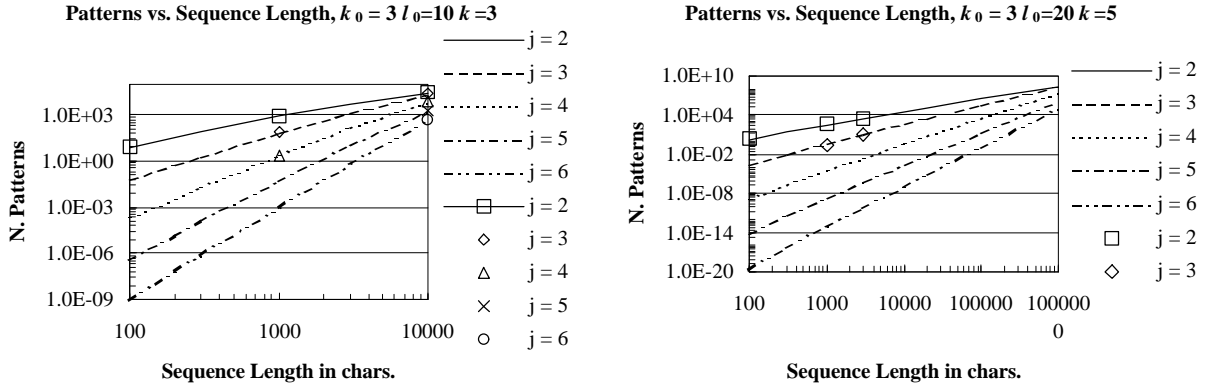


Fig. 1: Number of discovered patterns, theoretical and numerical results.

sequences can exhibit a rather surprising number of patterns, which arise out of chance collisions in the alphabet space. For example, for $k_0 = 3$, $l_0 = 20$, and $k = 5$ we have, in the average, about 3,000 different patterns that appeared twice in a string of length 3,000. Let us give another example. For $k_0 = 3$, $l_0 = 10$, and $k = 3$, we have that a sequence of length 10,000 will host in the average around 30,000 different maximal patterns with $j = 2$, 25,000 patterns with $j = 3$, 10,000 patterns with $j = 4$, 2,000 patterns with $j = 5$, and 500 patterns with $j = 6$.

Except for the approximations done in the computation of p_{out} and $N_0(k, l)$, the formulae presented so far are exact. The complexity in their actual calculation can be reduced considerably if we further assume a “mean-field” effect, by approximating $\langle \wp^j \rangle \approx \langle \wp \rangle^j$. This approximation gives reasonable results if the ratio $\langle (\wp - \langle \wp \rangle)^2 \rangle^{1/2} / \langle \wp \rangle$ is small enough. If such is the case, Equation (4.2) simplifies to

$$\langle v_{jkl} \rangle \approx N_0(k, l) \binom{L-l+2}{j} \langle \wp \rangle^{k(j-1)} [1 - \langle \wp \rangle^k]^{L-l-2-j} \langle p_{in} \rangle \langle p_{out} \rangle^2. \quad (4.5)$$

We shall call this, the mean-field approximation.

5. Two Case Studies

The understanding of $\langle v_{jkl} \rangle$ already allows us to make some inferences from the analysis of real biological sequences. In this Section we shall consider two examples. We have run Teiresias with parameters $k_0 = 3$ and $l_0 = 15$ on the aminoacid sequence of HUMPRCA, a human serine protease (GenBank accession # M11228) with 461 amino acids. We tabulated the number of patterns so obtained in Table 1 of Appendix D. In the same Table we report the expected number of patterns computed using the mean-field approximation for random sequences with the same $\langle \wp \rangle = 0.059$ as HUMPRCA ($\langle \wp \rangle$ was obtained with $\alpha_{ij} = \delta_{ij}$; see Equation (3.1)). It can be seen that the number of patterns occurring in the real protein are very close to those expected in a random sequence of similar composition.

As an example, these are the 4 patterns corresponding to the case $k = 8, j = 2$:

$\pi_1 = \text{H} \dots \text{W} \dots \text{L} \dots \text{E} \dots \text{L} \dots \text{L} \dots \text{G} \dots \text{Y}$
 $\pi_2 = \text{I} \dots \text{G} \dots \text{C} \dots \text{S} \dots \text{G} \dots \text{F} \dots \text{L} \dots \text{G}$
 $\pi_3 = \text{K} \dots \text{D} \dots \text{D} \dots \text{I} \dots \text{S} \dots \text{T} \dots \text{I} \dots \text{L}$
 $\pi_4 = \text{A} \dots \text{T} \dots \text{L} \dots \text{G} \dots \text{L} \dots \text{G} \dots \text{T} \dots \text{V}$

The fact that the expected number of patterns in random sequences with an aminoacid frequency equivalent to HUMPRCA is approximately the same as the number found for HUMPRCA, indicates that these patterns are statistically irrelevant.

We performed the same analysis for the sequence MUSHEPGFA, a mouse hepatocyte growth factor-like protein (GenBank accession # M74180) with 716 amino acid. The results are reported in Table 2 of Appendix D. For MUSHEPGFA, we have $\langle \wp \rangle = 0.058$. This is a much more interesting case. Contrary to what happens for HUMPRCA, except for $j = 2$ (and maybe $j = 3, k = 3$), the number of patterns found in MUSHEPGFA are much larger than the expected number of patterns in random sequences of equivalent composition. As an example, these are the 3 patterns corresponding to the case $j = 3, k = 6$.

$\pi_1 = \text{E} \dots \text{K} \dots \text{N} \dots \text{R} \dots \text{G} \dots \text{T}$
 $\pi_2 = \text{G} \dots \text{N} \dots \text{R} \dots \text{D} \dots \text{G} \dots \text{T}$
 $\pi_3 = \text{P} \dots \text{W} \dots \text{C} \dots \text{Y} \dots \text{T} \dots \text{C}$

The mean number of patterns expected for this value of the parameters is 0.004; the fact that there are 3 patterns (750 times the expected number of patterns), tells us that these patterns are at least statistically significant, and might be relevant from a biological standpoint. Another interesting example is the following pattern, (not listed in Table 2 of Appendix D),

$\pi_4 = \text{C} \dots \text{G} \dots \text{E} \dots \text{Y} \dots \text{R} \dots \text{T} \dots \text{G} \dots \text{C} \dots \text{Q} \dots \text{W} \dots \text{P} \dots \text{H} \dots \text{F} \dots \text{P} \dots \text{L} \dots \text{N} \dots \text{C} \dots \text{R} \dots \text{N} \dots \text{P} \dots \text{D} \dots \text{G}$,

with $j = 3$ and the remarkable $k = 23$. The mean number of times of a pattern like this would be expected in an random sequence with the same composition as MUSHEPGFA is less than 10^{-15} . Finally, another remarkable pattern, that includes π_4 , is:

$\pi_5 = \text{C} \dots \text{G} \dots \text{E} \dots \text{D} \dots \text{Y} \dots \text{R} \dots \text{T} \dots \text{G} \dots \text{C} \dots \text{Q} \dots \text{R} \dots \text{W} \dots \text{D} \dots \text{Q} \dots \text{P} \dots \text{H} \dots \text{H} \dots \text{F} \dots \text{P} \dots \text{E} \dots \text{K} \dots \text{D} \dots \text{L} \dots \text{N} \dots \text{C} \dots \text{R} \dots \text{N} \dots \text{P} \dots \text{D} \dots \text{G} \dots \text{S} \dots \text{E} \dots \text{P} \dots \text{W} \dots \text{C} \dots \text{T} \dots \text{P} \dots \text{F}$,

which appears only twice ($j = 2$), but has $k = 41$. We estimated that such a pattern would only be likely to occur once every 10^{40} sequences of this size! An indication that this pattern may be biologically significant is given by the fact that it contains, towards the end, a known motif [FY]CRNP[DNR], which is reported in the PROSITE database.

6. Statistical test for pattern relevance

In Section 4., on page 5, we have computed the *mean value* of v_{jkl} for random sequences. In this Section we wish to compute the probability of observing a specific value of v_{jkl} . We shall then used that distribution to give a quantitative measure of significance: a score.

Let us call N_ω the actual number of offsets that are going to contribute new patterns that have not occurred upstream of ω . Within the mean field approximation, that number can be estimated as

$$N_{\omega} = \sum_{i=0}^{L-1} (1 - \langle \wp \rangle^k)^i = \frac{1 - (1 - \langle \wp \rangle^k)^L}{\langle \wp \rangle^k}. \quad (6.1)$$

If we express $\langle v_{jkl} \rangle$ as the number of valid combs ($N_0(k, l)$) and offsets (N_{ω}) multiplied by the effective probability ϕ_{jkl} that any comb and offset produce a jkl -pattern, we obtain

$$\phi_{jkl} = \frac{\langle v_{jkl} \rangle}{N_{\omega} N_0(k, l)}. \quad (6.2)$$

We now want to compute the probability $P(v_{jkl})$ that a number v_{jkl} of jkl -patterns occurs in a random sequence. For that it is convenient to interpret $\langle v_{jkl} \rangle$ as the mean value of the number of successes when $N_{\omega} N_0(k, l)$ Bernoulli experiments are performed, with a probability of success equal to ϕ_{jkl} . In such case, we have that

$$P(v_{jkl}) = \binom{N_{\omega} N_0}{v_{jkl}} \phi_{jkl}^{v_{jkl}} (1 - \phi_{jkl})^{N_{\omega} N_0 - v_{jkl}}. \quad (6.3)$$

It should be made clear, at this point, that the above formula is an ansatz. Its validity is discussed in Appendix E, where we show that it approximates very well a set of numerical simulations. However, the actual probability $P(v_{jkl})$ can be approximated to a higher degree, by computing the variance of v_{jkl} in addition to the mean. These results will be reported in a follow-up paper.

Based on the theory presented so far, we would like to study an important problem in computational biology, namely the assignment of a quantitative measure of statistical relevance to patterns arising from biological datasets. To do that, we shall take the following approach: we have seen that one is to expect a *null hypothesis* number of jkl -patterns, merely from a random sequence. However, if for any reason, there are regions that are conserved through evolution or there are molecular mechanisms that replicate specific motifs, then patterns from those regions will tend to occur more often than those from regions affected only by the neutral mutation drift. Indeed, if sequences belonging to the latter category can be effectively modeled as random processes then one expects the corresponding distribution of the number of jkl -patterns to be close to the ones computed with Equation (6.3).

The previous discussion suggests the following scoring system for discovered patterns. Given a test biological sequence of length L , one can run a pattern discovery tool with parameters k_0 and l_0 . From the list of extracted patterns, we count the number v_{jkl} of jkl -patterns. From the given sequence we also compute the corresponding $\wp(v)$ with the actual token frequency of that sequence. Each jkl -pattern will then be assigned the score¹

$$s(j, k, l) = \sum_{v = v_{jkl}}^{\infty} P(v), \quad (6.4)$$

where $P(v)$, the probability of observing v jkl -patterns in a random sequence with the same token composition as that of the test sequence, is computed from Equation (6.3). The interpretation of this scoring system is straightforward: $s(j, k, l)$ is the probability of obtaining a number $v \geq v_{jkl}$ of patterns from an equally composed random sequence.

A small score for a jkl -pattern (say $s(j, k, l) \leq 0.05$) indicates statistical significance for that pattern and suggests a possible biological *raison d'être* for that motif. A larger score (which should of course be smaller than 1 by definition) means that any random, biologically irrelevant sequence would contain

1. This score tends to be smaller as the patterns become more relevant. If a score that increases with relevance is required, then other definitions could be used, such as the z -score, $-\log(s)$, etc.

approximately the same number of *jkl*-patterns as the test sequence. In that sense, patterns with score above threshold are statistically (and probably biologically) irrelevant.

As proof of concept, we will next discuss a concrete example. A region of DNA of length $L=1,000$ from yeast chromosome 1 [14], starting at location 219,000, was analyzed with Teiresias, with parameters $k_0 = 10$ and $l_0 = 14$. This region belongs to an intergenic portion of chromosome 1. We found a grand total of 99 patterns. The organization of these patterns according to their j , k and l properties, is reported in Table 3 of Appendix F.

With a significance threshold of 5% (i.e., we accept as statistically significant only scores smaller than 0.05), only the first two rows comprise statistically relevant patterns. We shall see next where the relevance of these patterns stems from. The important message to be drawn from this exercise is that of the almost 100 patterns found by Teiresias, only six happen to be statistically relevant. We list them below:

for $j = 2$, $k = 26$, and $l = 34$: P0 = TC.GT..T.TCTTCCTC.GTCAT.TCTTC.TC.G

for $j = 3$, $k = 10$, and $l = 14$: P1 = TCTT..TC.GT.AT; P2 = CAT.TCTT..TC.G; P3 = CTC..TCA..TCTT; P4 = GT..T.TCTTC.TC; P5 = TC.GT..T.TCTTC

The following is the DNA region aligned with the statistically relevant patterns

```

DNA   ACCTCATCCAGTTTGTGCATCATCTTCTTCAGAACAAATCACCAGCTCTATCAGTCTTAGCATCCAATTATTACTCCATTCTATCGCAGCA
P3                                         |CTC..TCA..TCTT|
DNA   ATGGAACTTCTGTAGTTTCTTCCCTCAGTCATGTCTTCCCTCGGTCATTTCTTCTTCTGCAACGACCTCCACTTCTATATTCTCTGAATCATC
P0                                         |TC.GT..T.TCTTCCTC.GTCAT.TCTTC.TC.G|
P0     |TC.GT..T.TCTTCCTC.GTCAT.TCTTC.TC.G|
P1           |TCTT..TC.GT.AT|TCTT..TC.GT.AT|
P2           |CAT.TCTT..TC.G|CAT.TCTT..TC.G|
P3           |CTC..TCA..TCTT|CTC..TCA..TCTT|
P4           |GT..T.TCTTC.TC|GT..T.TCTTC.TC|GT..T.TCTTC.TC|
P5           |TC.GT..T.TCTTC|TC.GT..T.TCTTC|TC.GT..T.TCTTC
DNA   TAAATCATCCGTCATTCAAACCAAGTAGTTCCACCTCTGGTTCTTCTGAGAGCGAAACAATCTTAGTGATTGCTGGCATGTCATTAGCGACA
DNA   AGACGCTTATTACCGTAGTAGCCCCCAAGGCAAACATCTCTTTATCAGTAATATCCAAA
P1           |TCTT..TC.GT.AT|
P2           |CAT.TCTT..TC.G|

```

We can see that all the statistical relevant patterns are all from the same region. Indeed, these patterns correspond to 3.3 copies of the following tandem repeat of period 15:

.....TCTGTAGTTTCTTCC TCAGTCATGTCTTCC TCGGTCATTTCTTCT TCTG.....

It is interesting to note that some of the patterns that correspond to this tandem repeat (specifically P1, P2 and P3) also appear not far upstream and downstream of the tandem repeat. These regions, flanking the tandem repeat itself, might also be biologically relevant in this context

7. Summary and Conclusions

In this paper we have shown that the combinatorial nature of arbitrary patterns can result in a high number of chance coincidences in random sequences with composition similar to actual biosequences. Thus, when detecting patterns in real biosequences, it is likely that biologically relevant patterns, if present at all, will be buried in a sea of irrelevant patterns. Thus we need a statistical tool designed, so to speak, to find a needle in a haystack.

An analytical framework to evaluate the statistical significance of patterns discovered by several algorithms has been proposed. The framework models all major properties of pattern discovery algorithms, such as maximality and density constraints. We show that our formulae can be reliably used to represent the probability functions for the number of *jkl*-patterns for $j \geq 3$. For $j = 2$, such a distribution tends to slightly overestimate (by about a 10%) the number of *jkl*-patterns in random sequences. This will result in

a somewhat higher rate of false negatives, in the assignment of statistical significance to patterns with $j = 2$. Even though this overestimate will be unimportant in the majority of cases, a precise assessment of the rate of increase of false negatives for $j = 2$ is missing, and will be attempted in the future.

We have developed a quantitative scoring measure for the statistical significance of maximal patterns. The interpretation of this scoring measure is as follows: if a jkl -pattern appears v_{jkl} times in a sequence, then its score $s(j, k, l)$ is the probability that the number v of jkl -patterns in a random sequence (of the same base or aminoacid composition as the one given) is bigger or equal than v_{jkl} .

In a biological setting, this scoring approach would assume that irrelevant patterns live in a sea of neutrally drifting sequences, not undergoing any specific natural selection pressure. We only use first order statistics to model our null-hypothesis. Other possibilities might be, to constrain the n -token composition (di-nucleotide, tri-nucleotide, di-amino acid, etc.).

Pattern discovery tools, such as Teiresias or SPLASH, are model-less and produce all the maximal patterns in a given sequence or sets of sequences. In that context, our score can be used to hierarchically order the discovered patterns according to their statistical significance.

It has been pointed out [14] that “statistical significance is neither necessary nor sufficient for biological significance, but it is a good indicator”. Some simple examples, in this paper, show that statistical tools can be useful in highlighting interesting regions in biological sequences. We have applied the significance criterion to single string analysis of a stretch of DNA from yeast chromosome 1 (spanning from location 219,000 to 220,000), and showed that the most significant patterns in the statistical sense, correspond to a tandem repeat structure of 3.3 repetitions (with variation) of a basic unit of size 15. Thus, to the extent that tandem repeats have a biological activity (caused by molecular mechanisms acting at the DNA level), our significance criterion provided hints to the biological relevance of the pattern

Applications of the same statistical techniques to multiple strings are being studied currently and will be reported in a follow-up paper. They are clearly relevant in the field of homology analysis.

Acknowledgments

Very special thanks go to I. Rigoutsos and A. Floratos, who made Teiresias available to us, and provided considerable insights into the substance of this paper. Special thanks go to G. Sorkin for many discussions pertaining the adequate counting of constrained-density tuples. We wish to acknowledge Y. Gao, L. Parida, Barry Robson, D. Silverman, D. Platt, Ajay Royyuru, and Mike Pitman for useful comments.

References

- [1] Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990) *J.Mol.Bio.*, 215, 403-410
- [2] T.L.Bailey and M.Gribskov, “Methods and statistics for combining motif match scores,” *Journal of Computational Biology*, Vol. 5, pp. 211-221, 1998.
- [3] A. Bairoch. “PROSITE: a dictionary of sites and patterns in proteins.” in *Nucleic Acids Res.* (1991) **19**, 2241-2245
- [4] A.Califano, “SPLASH: Structural Pattern Localization Algorithm by Sequential Histograming,” to be published as IBM RC
- [5] M.O.Dayhoff: “ATLAS of PROTEIN SEQUENCE and STRUCTURE,” Volume 5 SUPPLEMENT 3, 1978.
- [6] S.Henikoff and J.G.Henikoff: “Amino acid substitution matrices from protein blocks.” *Proc. Natl. Acad. Sci. USA.* 89: 10915-10919, Nov., 1992.
- [7] I.Jonassen, J.F.Collins, D.G.Higgins. “Finding flexible patterns in unaligned protein sequences,” *Protein Science* 4, 1587-1595 (1995)
- [8] Karlin, S., and Brendel, V., “Chance and Statistical Significance in Protein and DNA Sequence Analysis,” *Science*, **257**, 39-49 (1992)

- [9] Neuwald, A.F., and Green, P., "Detecting Patterns in Protein Sequences," J. Mol. Bio. **239**, 698-712, (1994).
- [10] Pearson, W.R. and Lipman, D.J. (1988) Proc. Natl. Acad. Sci. USA, **85**, 2444-2448.
- [11] I.Rigoutsos and A.Floratos: "Motif Discovery without Alignment or Enumeration," in S.Istrail, P.Pevzner, and M.S.Waterman editors, Proc. 2nd annual ACM Intl. Conf. on Comp. Mol. Biol, RECOMB98, 221-227, 1998
- [12] R.M.Schwartz and M.O.Dayhoff: "Matrices for Detecting Distant Relationships," Atlas of Protein Sequence and Structure, 1978, ed. Dayhoff, M.O., pp. 353-358.
- [13] M.S.Waterman, "Introduction to Computational Biology," Chapman & Hall, London, 1995
- [14] Data was obtained from *ftp.ebi.ac.uk*, directory *pub/databases/yeast*

Appendix A

The use of a distance metric, allows one to extend the notion of a pattern in useful ways. For instance, using a BLOSUM50 mutation probability matrix [6] as a distance metric¹ and a threshold $D = 2$, the following amino acid equivalences would be allowed in a pattern:

$$\begin{array}{lll} \text{Ala} = [\text{Ala}] & \text{Arg} = [\text{Arg}, \text{Lys}] & \text{Asn} = [\text{Asn}, \text{Asp}] \\ \text{Asp} = [\text{Asp}, \text{Glu}] & \text{Cys} = [\text{Cys}] & \text{Gln} = [\text{Gln}, \text{Glu}, \text{Lys}] \\ \text{Glu} = [\text{Glu}, \text{Asp}, \text{Gln}] & \text{Gly} = [\text{Gly}] & \text{His} = [\text{His}, \text{Tyr}] \\ \text{Ile} = [\text{Ile}, \text{Leu}, \text{Met}, \text{Val}] & \text{Leu} = [\text{Leu}, \text{Ile}, \text{Met}] & \text{Lys} = [\text{Lys}, \text{Arg}, \text{Gln}] \\ \text{Met} = [\text{Met}, \text{Ile}, \text{Leu}] & \text{Phe} = [\text{Phe}, \text{Tyr}] & \text{Pro} = [\text{Pro}] \\ \text{Ser} = [\text{Ser}, \text{Thr}] & \text{Thr} = [\text{Thr}, \text{Ser}] & \text{Trp} = [\text{Trp}, \text{Tyr}] \\ \text{Tyr} = [\text{Tyr}, \text{His}, \text{Phe}, \text{Trp}] & \text{Val} = [\text{Val}, \text{Ile}] & \end{array} \quad (7.1)$$

In this case, with the above notation and the following sequence,

$$\text{Ala Cys Gln Gln Val Trp Ala Gly Ala Phe Ile Tyr Leu His Pro}, \quad (7.2)$$

the pattern $\{(Ala, 0)(Ile, 4)(Tyr, 5)\}$ would have locus $\lambda = \{0, 6, 8\}$. Based on the same matrix, however, the pattern $\{(Ala, 0)(Val, 4)(Trp, 5)\}$, which appears at position 0, would have a different locus since both $f(\text{Val}, \text{Leu}) > D$ and $f(\text{Trp}, \text{His}) > D$.

If the distance metric is such that only identical values are considered equivalent, then there are several algorithms such as Pratt [2], Meme [7], and Teiresias [11], among others, that can automatically discover patterns from data streams. Pratt is also notable because it allows for arbitrary insertions and deletions in the patterns. Splash [4] allows arbitrary metrics to be used such as PAM [12] or BLOSUM [5] matrices.

Appendix B

Finding the probability that a pattern is not extensible beyond its boundaries present some difficulties. These can be overcome by using some heuristics. In the present analysis, for simplicity, we analyze only the right end of the comb. Exactly the same reasoning applies to the left end. The idea of the calculation is that we are allowed to have matching tokens to the right of the comb, as long as they do not meet the density constraint. Those additional tokens could be anywhere between the first site to the right of the rightmost comb token, and the $(l_0 - 1)$ -th site. Let us denote the location of a number n of outside matching tokens by i_1, i_2, \dots, i_n , and the probability of having m tokens strictly within the pattern in a window of size s by $\Phi(m|s)$. With that notation, the probability of non-violation of the maximality in length property having into account the density constraint is:

1. In rigor, a distance metric has to be symmetric. The measure of “relatedness” f is not symmetric (e.g $f(\text{Asp}, \text{Asn}) > 2$, whereas $f(\text{Asn}, \text{Asp}) < 2$). We shall nevertheless abuse notation and still call f a “distance”.

$$p_{out}(j, \pi_{\psi\omega}) \cong \sum_{n=0}^{k_0-2} \left\{ \sum_{\substack{1 \leq i_1 < i_2 < \dots \\ \dots < i_n \leq l_0-1}} \left[\prod_{\substack{\vartheta = i_1 \\ \dots, i_n}} \wp(v_{l-1+\vartheta})^j \right] \left[\prod_{\substack{1 \leq v \leq l_0-1 \\ v \neq i_1, \dots, i_n}} [1 - \wp(v_{l-1+v})^j] \right] \right. \\ \left. \times \sum_{m=m_{min}}^{k_0-2-n} \Phi(m|l_0-i_n-1) \right\} \quad (7.3)$$

The interpretation of this formula is as follows. The probability of a configuration in which exactly n tokens match in a span of $l_0 - 1$ to the right of pattern in *all* the j loci where the pattern occurs is

$$\left[\prod_{\substack{\vartheta = i_1 \\ \dots, i_n}} \wp(v_{l-1+\vartheta})^j \right] \left[\prod_{\substack{1 \leq v \leq l_0-1 \\ v \neq i_1, \dots, i_n}} [1 - \wp(v_{l-1+v})^j] \right] \quad (7.4)$$

(the factors of the form $[1 - \wp(v_{l-1+v})^j]$ take into account the probability that there are $l_0 - n - 1$ sites where at least one token of the j does not match). In order for this configuration not to violate the maximality constraint, we have to make sure that the window spanning a length l_0 starting somewhere inside the pattern and ending in the right-most one of the n matching tokens, contains at most $k_0 - 1$ tokens. Given that we have already n outside the pattern, and the right extreme of the pattern is anchored, that leaves room for at most $k_0 - 2 - n$ tokens inside the pattern. Therefore we have to weigh this configuration of n outside tokens with the probability that the number of tokens in a window of size $l_0 - 1 - i_n$ (where i_n is the location of the last token) strictly inside the pattern contains at most $k_0 - 2 - n$ and not less than m_{min} tokens. m_{min} has to be calculated from considerations involving the density constraint, but we shall proceed in a more intuitive, albeit less rigorous way, and consider m_{min} to be the typical number of tokens in a window of size $l_0 - 1 - i_n$ inside the pattern. Using now that the token density strictly inside the pattern is $d = \frac{(k-2)}{(l-2)}$, we estimate m_{min} to be $m_{min} = d(l_0 - 1 - i_n)$.

Within the same mean-field scheme, we can estimate the probability of having m tokens strictly inside the pattern in a window w as the $\Phi(m|w)$ given in Equation (3.11). The third sum in the equation for p_{out} above is, therefore, the weight given to an outside configuration of n tokens given by the number of compatible inside configurations.

After averaging p_{out} over all the tokens v_δ , the indices i_1, \dots, i_{n-1} are not used any longer, and they can be summed right away to yield the binomial number $\binom{l_n-1}{n-1}$. Thus we obtain the formula for $\langle p_{out} \rangle$ discussed in the following Appendix.

Appendix C

To compute $\langle v_{jkl} \rangle$ we only need to compute the mean value of $B(j, \pi_{\psi\omega})$. It is convenient to compute the latter in two steps: first as its mean value conditioned on having fixed $\pi_{\psi\omega}$ at offset ω ; second take the resulting conditional expectation and average it over all possible patterns $\pi_{\psi\omega}$. That is, $\langle B(j, \pi_{\psi\omega}) \rangle = \langle \langle B(j, \pi_{\psi\omega}) | \pi_{\psi\omega} \rangle \rangle$.

The ingredients for the conditional expectation $\langle B(j, \pi_{\psi\omega}) | \pi_{\psi\omega} \rangle$ have been worked out in the previous Section. In effect, $\langle B(j, \pi_{\psi\omega}) | \pi_{\psi\omega} \rangle$ is nothing but the probability that $\pi_{\psi\omega}$ is a $\langle k_0, l_0 \rangle$ -maximal jkl -pattern, and thus it is the product of the probability $\mathfrak{S}(j-1, \pi_{\psi\omega})$ that it has appeared $j-1$ times after ω , multiplied by the probability $p_b(\pi_{\psi\omega})$ that it has not appeared at earlier offsets, multiplied by the probability

$p_{in}^L(j-1, \pi_{\psi\omega})$ that it is maximal in length, multiplied by the probabilities $p_{out}^R(j-1, \pi_{\psi\omega})$ and $p_{out}^L(j-1, \pi_{\psi\omega})$ that it cannot be extended neither to the right or to the left of $\pi_{\psi\omega}$, respectively:

$$\langle B(j, \pi_{\psi\omega}) | \pi_{\psi\omega} \rangle = \mathfrak{S}(j-1, \pi_{\psi\omega}) p_b(\pi_{\psi\omega}) p_{in}(j-1, \pi_{\psi\omega}) p_{out}^R(j-1, \pi_{\psi\omega}) p_{out}^L(j-1, \pi_{\psi\omega}). \quad (7.5)$$

The next step is to average the previous equation with respect to the tokens that have been frozen for its computation. Close inspection indicates that the tokens frozen in $\mathfrak{S} p_b$, p_{in} , p_{out}^R and p_{out}^L are all statistically independent, and thus the global average of $\langle B(j, \pi_{\psi\omega}) | \pi_{\psi\omega} \rangle$ can be factorized into the averages of $\mathfrak{S} p_b$, p_{in} , p_{out}^R and p_{out}^L . The expressions for these averages are

$$\langle \mathfrak{S}(j-1, \pi_{\psi\omega}) (p_b(\pi_{\psi\omega})) \rangle = \binom{L-\omega-l}{j-1} \sum_{i=0}^{L-l+1-j} \binom{L-l-1-j}{i} (-1)^i \langle \wp^{j-1+i} \rangle^k \quad (7.6)$$

(where we have used that $\langle \rho_{\pi_{\psi\omega}}^j \rangle = \langle \wp^j \rangle^k$),

$$\langle p_{in}(j-1, \pi_{\psi\omega}) \rangle = [1 - \langle \wp^{j-1} \rangle]^{l-k}, \quad (7.7)$$

and

$$\langle p_{out}^{R,L}(j-1, \pi_{\psi\omega}) \rangle = \sum_{n=0}^{k_0-2} \langle \wp^{j-1} \rangle^n [1 - \langle \wp^{j-1} \rangle]^{l_0-n-1} \sum_{i=n}^{l_0-1} \binom{i-1}{n-1} \sum_{m=d(l_0-i-1)}^{k_0-n-2} \Phi(m | l_0-i-1) \quad (7.8)$$

The only dependence on ω of the previous formulae appears through the binomial in $\langle \mathfrak{S} p_b \rangle$, which can be readily summed:

$$\sum_{\omega=0}^{L-l+1-j} \binom{L-l-\omega}{j-1} = \binom{L-l+1}{j-1}. \quad (7.9)$$

After averaging each $B(j, \pi)$, none of the terms participating in the sums that yield $\langle v_{jkl} \rangle$ depend on ψ . Thus the sum over combs results in the term $N_0(k, l)$ (the number of constrained $\langle k_0, l_0 \rangle$ -valid combs). This produces the result reported in Equation (4.2).

Appendix D

TABLE 1. $\langle k_0 = 3, l_0 = 15 \rangle$ -valid pattern distribution for human serine protease HUMPRCA

| # of times (j) | k=3 | | k=4 | | k=5 | | k=6 | | k=7 | | k=8 | |
|----------------|-----|-------|-----|-------|-----|------|-----|------|-----|------|-----|------|
| | exp | theo | exp | theo | exp | theo | exp | theo | exp | theo | exp | theo |
| 2 | 357 | 349.9 | 166 | 124.9 | 64 | 59.8 | 32 | 23.3 | 4 | 11.3 | 4 | 4.5 |
| 3 | 47 | 45.8 | 1 | 1.41 | 0 | 0.05 | 0 | 0.00 | 0 | 0.00 | 0 | 0.00 |
| 4 | 1 | 1.18 | 0 | 0.00 | 0 | 0.00 | 0 | 0.00 | 0 | 0.00 | 0 | 0.00 |
| 5 | 1 | 0.02 | 0 | 0.00 | 0 | 0.00 | 0 | 0.00 | 0 | 0.00 | 0 | 0.00 |
| 6 | 0 | 0.00 | 0 | 0.00 | 0 | 0.00 | 0 | 0.00 | 0 | 0.00 | 0 | 0.00 |

TABLE 2. $\langle k_0 = 3, l_0 = 15 \rangle$ -valid pattern distribution for mouse hepatocyte growth factor MUSHEPGFA.

| # of times (j) | k=3 | | k=4 | | k=5 | | k=6 | | k=7 | | k=8 | |
|----------------|-----|-------|-----|-------|-----|-------|-----|------|-----|-------|-----|------|
| | exp | theo | exp | theo | exp | theo | exp | theo | exp | theo | exp | theo |
| 2 | 646 | 787.9 | 363 | 288.8 | 127 | 136.9 | 79 | 52.5 | 21 | 25.07 | 11 | 9.7 |
| 3 | 204 | 141.6 | 30 | 4.31 | 7 | 0.15 | 3 | 0.00 | 1 | 0.00 | 0 | 0.00 |
| 4 | 38 | 5.20 | 0 | 0.01 | 1 | 0.00 | 0 | 0.00 | 0 | 0.00 | 1 | 0.00 |
| 5 | 14 | 0.14 | 0 | 0.00 | 0 | 0.00 | 0 | 0.00 | 0 | 0.00 | 0 | 0.00 |
| 5 | 1 | 0.00 | 0 | 0.00 | 0 | 0.00 | 0 | 0.00 | 0 | 0.00 | 0 | 0.00 |

Appendix E

In this Appendix we test the theoretical distribution of v_{jkl} found in Section 6 against a numerical example. We have produced 1,000 random “protein” sequences of length $L=2,000$, and for each of them, we performed a pattern search with parameters $k_0 = 3$ and $l_0 = 10$, and computed v_{jkl} for $2 \leq j \leq 6$, $k = 3$ and $l = 10$. The results corresponding to these numerical experiments are marked with circles in Fig. 2(a-e). As expected, the experimental distributions peak at smaller values of v_{jkl} as j increases. The theoretical distributions corresponding to Equation (6.3) are plotted as thick solid lines in the same figures. It is clear that except for the slight overestimate in Fig. 2a (for which the mean value for the theoretical distribution is 10% bigger than the experimental one), the plots corresponding to Equation (6.3) are a good approximation of the experimental distributions.

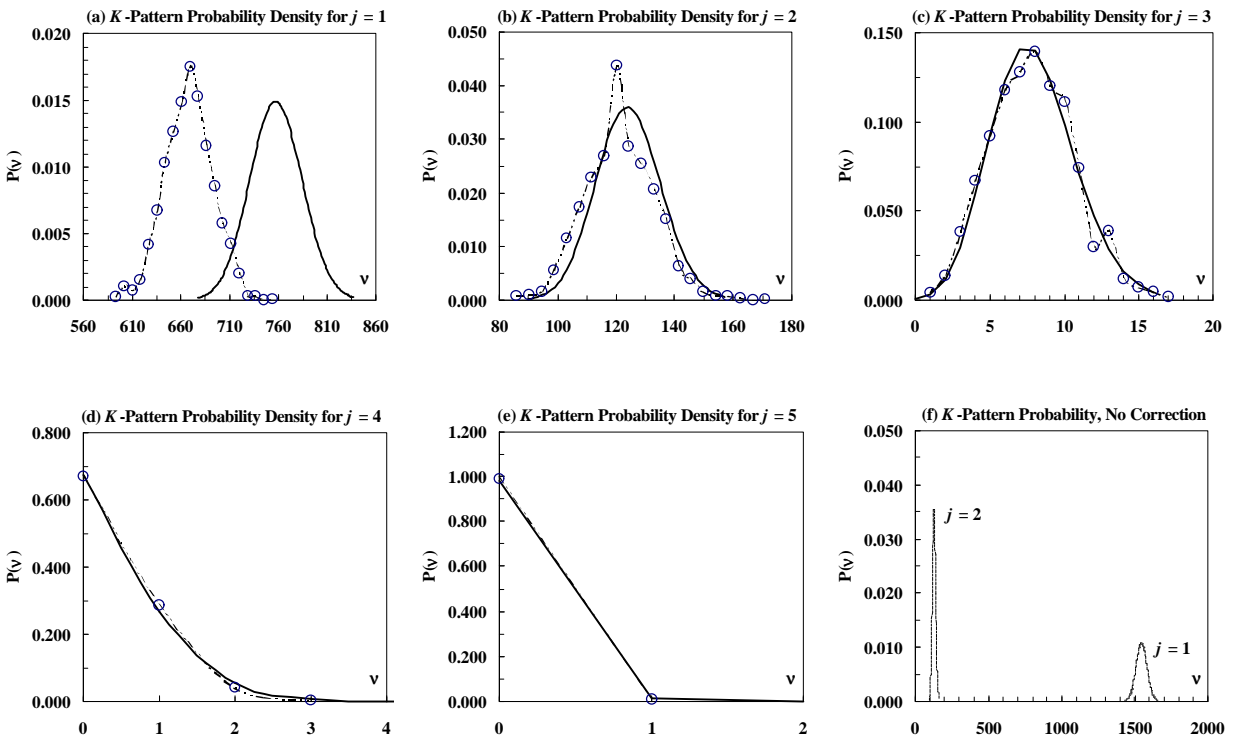


Fig. 2: jkl -Pattern Probability functions (Please note that values of j should be replaced by $j+1$)

It is important, however, to understand the source of the overestimate for $j = 2$. We have modeled in Section 3 the length and composition maximality constraints. The former is taken into account *via* p_{out} (Equation (3.10)), whereas the latter is accounted for *via* p_{in} (Equation (3.6)). The density constraint requires that we estimate the number of valid combs N_0 (Equation (3.7)). Of these three constraints, only p_{in} is computed exactly. The estimates we have got for p_{out} and N_0 are only approximations, and they are the responsible for any discrepancy between theory and experiment.

In the examples of Fig. 2, however, we have chosen $k = k_0$ and for that choice of k , the exact value of N_0 is computed: $N_0(k, l) = \binom{l-2}{k-2}$. Therefore, the only source for the overestimation comes from p_{out} . Our expression for p_{out} tends to slightly overestimate the actual values of v_{jkl} . With respect to N_0 , if higher accuracy is demanded in its calculation, one could easily compute it from Montecarlo simulations in constant time.

If none of these constraints were imposed, the previous calculations would still be valid for the statistics of v_{jkl} with unrestricted k -tuples. In such case, our formulae remain valid if we set $p_{in} = p_{out} = 1$, and $N_0(k, l) = \binom{l-2}{k-2}$. This instance will be called the *no-correction* case.

Fig. 2f shows the no-correction case for $j = 2, 3$ on the same scale. In this case, for $j = 2$, the values of v_{jkl} are seriously overestimated with respect to the correct results in Fig. 2a by a full 100%.

Other analyses (not shown) have been performed, that allow us to conclude that the theory is robust with respect to the various variables involved, including $\langle \wp \rangle$ in the mean field approximation, as long as its standard deviation is a small fraction of its mean.

Appendix F

TABLE 3. Results of Pattern Discovery on Yeast Chromosome 1 region

| j | k | l | v_{jkl} | score (in %) |
|----------|----------|----------|-----------|---------------------|
| 3 | 10 | 14 | 5 | 6.e-07 |
| 2 | 26 | 34 | 1 | 4.e-03 |
| 2 | 16 | 24 | 1 | 6 |
| 2 | 14 | 19 | 3 | 11 |
| 2 | 14 | 17 | 1 | 27 |
| 2 | 12 | 19 | 1 | 28 |
| 2 | 12 | 18 | 2 | 45 |
| 2 | 11 | 17 | 1 | 61 |
| 2 | 11 | 16 | 4 | 80 |
| 2 | 12 | 15 | 2 | 82 |
| 2 | 13 | 18 | 1 | 91 |
| 2 | 12 | 17 | 2 | 92 |
| 2 | 10 | 11 | 1 | 94 |
| 2 | 11 | 13 | 1 | 96 |
| 2 | 11 | 14 | 4 | 98 |
| 2 | 10 | 14 | 44 | 99 |
| 2 | 11 | 15 | 12 | 100 |
| 2 | 12 | 16 | 2 | 100 |
| 2 | 12 | 16 | 2 | 100 |
| 2 | 10 | 13 | 9 | 100 |