BIOINFORMATICS

Vol. 18 no. 10 2002
Pages 1382–1390

Subtle motifs: defining the limits of motif finding algorithms

U. Keich and P. A. Pevzner*

Department of Computer Science and Engineering, University of California San Diego, La Jolla, CA 92093, USA

Received on February 1, 2002; revised on March 25, 2002; accepted on April 25, 2002

ABSTRACT

Motivation: What constitutes a subtle motif? Intuitively, it is a motif that is almost indistinguishable, in the statistical sense, from random motifs. This question has important practical consequences: consider, for example, a biologist that is generating a sample of upstream regulatory sequences with the goal of finding a regulatory pattern that is shared by these sequences. If the sequences are too short then one risks losing some of the regulatory patterns that are located further upstream. Conversely, if the sequences are too long, the motif becomes too subtle and one is then likely to encounter random motifs which are at least as significant statistically as the regulatory pattern itself. In practical terms one would like to recognize the sequence length threshold, or the twilight zone, beyond which the motifs are in some sense too subtle.

Results: The paper defines the motif twilight zone where every motif finding algorithm would be exposed to random motifs which are as significant as the one which is sought. We also propose an objective tool for evaluating the performance of subtle motif finding algorithms. Finally we apply these tools to evaluate the success of our MULTIPROFILER algorithm to detect subtle motifs.

Contact: keich@cs.ucsd.edu

INTRODUCTION

A biologists looking for (unknown) regulatory signals faces a number of choices. One of the problems is how to generate samples of upstream regions for co-regulated genes of interest. For example, a wrong choice of the length of the upstream regions in the sample (for a set of co-regulated genes) may render even the best motif finding algorithm useless. While generating samples of upstream regions biologists often tend to choose longer (up to 2000 bp in expression analysis studies) rather than shorter regions in order to avoid missing some distant regulatory elements. However, some regulatory elements are so subtle (e.g. *E. coli* promoters) that increasing the length of the sequences in the sample to over 100 bp is bound to introduce random motifs which are at least as significant as the regulatory element itself (see Vanet *et al.* (2000) and Eskin *et al.* (2002) on difficulties with finding *E. coli* promoters in large samples). On the other hand, reducing the length of the sample may increase the corruption of the sample (i.e. the number of sequences without signals) and, once again, may lead to losing the signal. This dilemma combined with recent algorithmic advances in finding subtle motifs (Pevzner and Sze, 2000; Buhler and Tompa, 2001; Keich and Pevzner, 2002) raise the following question: what is the maximal length of sample that allows successful detection of a signal of a given strength, or more generally, what is a subtle motif?

This paper answers this question by defining the twilight zone in motif finding where every motif finding algorithm would have difficulties due to the fact that random motifs start competing with biological motifs. Our new MULTIPROFILER algorithm (Keich and Pevzner, 2002) was designed to find particularly subtle motifs even in the cases when a real motif may be blurred by random motifs. We analyze the performance of MULTIPROFILER in a proposed general framework of evaluating the reliability of subtle motif finders and demonstrate that it is able to detect motifs that are in the twilight zone and beyond.

We assume that relying on a scoring function, the motif finder tries to find a motif that is implanted in a sample generated according to the *sample model*. In many motif finding studies the sample, $\mathcal{S} = \{S_1, \ldots, S_n\}$, is formed of n randomly generated sequences of length N. The motif itself is generated and implanted in accordance with the *motif model*. Although much of the discussion in this section holds more generally, we concentrate on two consensus based motif models, where the motif consists of *instances* which are mutated images of the 'backbone' pattern P (Stormo, 2000). The first motif model we consider is the FM model (Pevzner and Sze, 2000) where each of the n sequences contains one instance of an (l, k)*-motif, i.e. each instance of the pattern of length l contains k positions which are randomly mutated. The second model is the VM model (Pevzner and Sze, 2000) where again each sequence contains exactly one instance,

*To whom correspondence should be addressed.

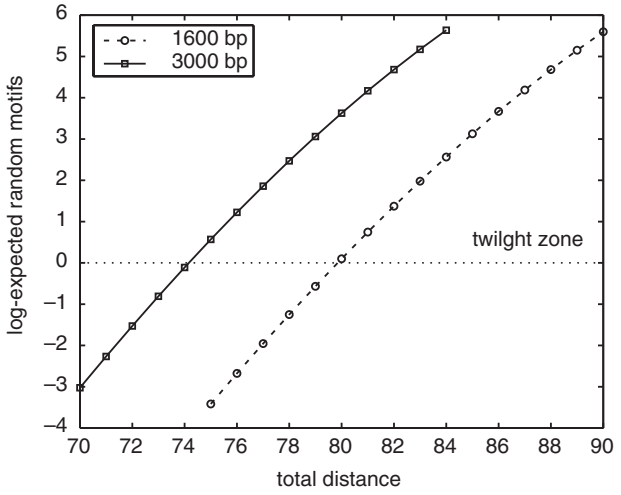

Fig. 1. log expected number of random motifs versus the total distance score.

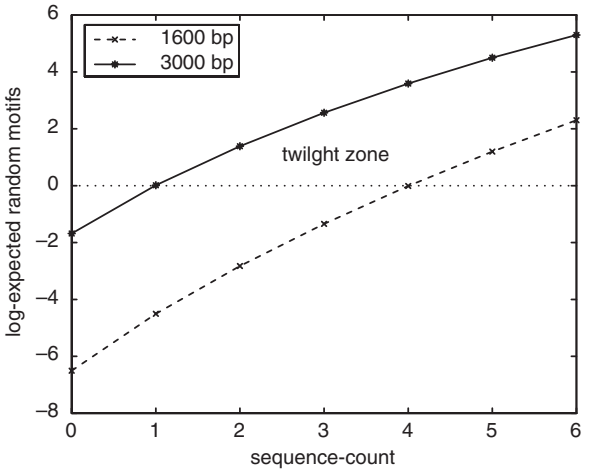

Fig. 2. log expected number of random motifs versus the sequence count score. The point (x, y) is on the graph if the expected number of random motifs with a score of x or better (i.e. x or smaller for both the total distance and sequence count scores) is 10^y. The two curves in each of the two figures correspond to sample models of 20 sequences with 1600 and 3000 base pairs (bp) respectively. The points are connected by lines merely to facilitate visualization. Note how for the total distance score the onset of the twilight zone shifts from a score of 80 to 74 when we move from 1600 to 3000 bp sequences. The analogous shift for the sequence count score is from 4 to 1.

only now each position of the instance is mutated, independently of all other positions, with probability ρ. We mainly discuss the total-distance scoring function[†], but in the context of the FM model we also consider the 'sequence-count' scoring function used by Buhler and Tompa (2001), where we count the number of sequences in the sample whose distance to the putative pattern is greater than k. Note that the total distance of the implanted motif in the FM model is $\leq nk$, while its sequence-count score is always 0.

Loosely speaking, we consider a motif as subtle if its score is unremarkable when compared with the scores of some random motifs present in the same sample. This raises the question of what is the *twilight zone* of scores beyond which we can expect to start seeing random motifs. Clearly, this threshold depends on the sample model and on the choice of the motif scoring function. The sample model determines the distribution of the scores of random motifs and in particular it determines the expected number of random motifs at any given score. Generally, we expect the twilight zone threshold to lie around the score, for which the expected number of random motifs that exceed this score, is about 1[‡]. Figures 1 and 2 demonstrate how this threshold can vary with the size of the sample and with the choice of the scoring function.

In order to determine the 'subtlety' of the motif we need some idea about how the score of the implanted motifs is distributed (Figure 3 demonstrates how this distribution varies with the model). Note that in general the score of the implanted motif varies not only with the particular motif, generated according to the motif model, but also with the sample. For example, the total distance score will improve with good random matches of the pattern.

We call a motif subtle or *dim* if its median score lies beyond the twilight zone[§]. It is interesting to note the impact of the scoring function on this definition. Consider for example our FM 1600 challenge problem: an FM model of, a $(15, 4)^*$-motif implanted in twenty 1600 bp sequences. This motif is dim when viewed with the total distance score, however, the exact same motif 'shines' quite brightly through the sequence-count binocular. Indeed, using this score we can extend the sequence length to over 3000 before this motif becomes dim (Figure 4).

A more refined gauge of the subtlety of a motif can be obtained from what we call the *subtlety graph* (Figure 4). This universally scaled graph is generated by plotting the distribution of the implanted motif score (Figure 3) directly against the expected number of random motifs (Figures 1 and 2). For example, the point $(81, 10^{3.1})$ which lies on the subtlety graph of the total distance - FM 3000

[†] The distance between a word W, of length l, and a sequence S is: $d(W, S) = \min_{B \in S} d(W, B)$, where the minimum is taken over all words of length l in S, and $d(W, B)$ is the Hamming distance between W and B. The *total distance* of W, is $d(W) = \sum_i d(W, S_i)$, where the summation extends over all the sequences in the sample.

[‡] Clearly, the choice of 1 is somewhat arbitrary, 2 or 1/2 are equally plausible however 10^6 or 10^{-6}, for example, are dubious choices for defining this threshold.

[§] The choice of median is, again, somewhat arbitrary

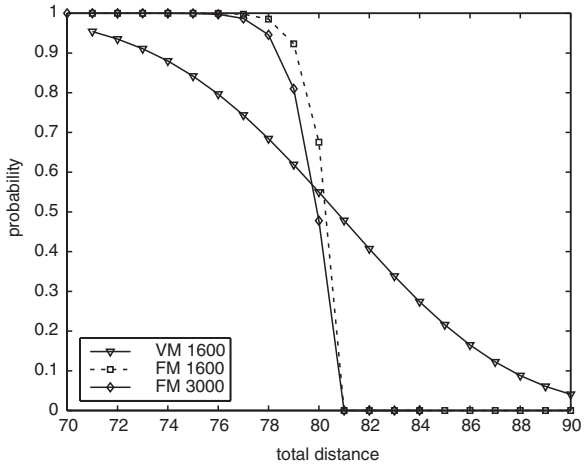

Fig. 3. The distribution of the score of implanted motifs. The point (x, y) is on the graph if y is the probability that the score of the implanted motif will be x or worse (i.e. x or bigger for the total distance). The points are connected by lines merely to facilitate visualization. The two FM models consists of a $(15, 4)^*$-motif implanted once in each of the 20 sequences with 1600 and 3000 bp. In the VM model the positions of the 15 bp pattern are mutated independently in each of the 20 instances (implanted in 1600 bp sequences) with probability 0.3. The location of the points was calculated as described below and agreed well with empirical distributions generated from 10^6 random samples for each of the three motif models: there were no mismatches, up to the third decimal digit, among the 1600 long models, while the few mismatches between the 3000 models were no more than 10^{-3} apart.

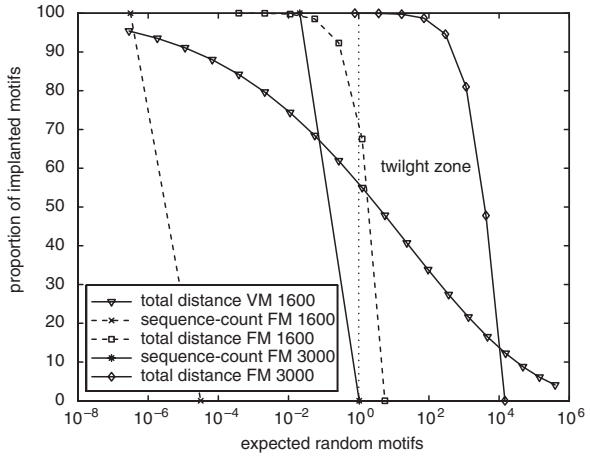

Fig. 4. Subtlety graphs. The point (x, y) is on the graph if the score of y percent of the implanted motifs is (on the average) no better than the score of at least x random motifs. The points are connected by lines merely to facilitate visualization. The points on the two sequence-count based graphs are located either at $y = 0$ or at $y = 1$ as the score of an FM implanted motif is constant at 0. Since the medians of all three total distance based examples are in the twilight zone, all three correspond to dim motifs. Note that the same motif implanted according to an FM 1600 model is dim under the total distance score but shines brightly with the sequence-count score. When the sequences are extended to 3000 bp, the motif is still not dim with the sequence-count score but is 'practically lost' on the total distance score: about 81% of the implanted motifs are buried by an average of over 1259 random motifs whose total distance score is at least as significant. Nevertheless, MULTIPROFILER is equally likely to detect this motif (over 98% of the time) using either score.

problem (a $(15, 4)^*$-motif implanted in twenty 3000 bp sequences) signifies that 81% of the implanted motifs are cluttered by (roughly) an average of at least 1259 random motifs whose score is at least as impressive. Note that the subtlety graph immediately yields whether or not the motif is dim.

Mathematically, the subtlety graph can be defined as follows. Let S be the score of the randomly implanted pattern and let M be the number of random motifs whose score is S or better (both S and M are random variables). Then, the subtlety graph essentially yields the distribution of $\mathrm{E}[M|S]$. Note that while the distribution of M is an even more refined measurement of motif subtlety, it is typically much harder to compute than the distribution of $\mathrm{E}[M|S]$.

The subtlety graph is closely related to the existing notion of an ROC (receiver operating characteristic) curve (e.g. Swets (1988)). ROC graph evaluates a diagnostic system by plotting the percentage of 'true positives' vs. the percentage of 'false positives' as they vary together according to an adjustable parameter. Let us consider a perfect algorithm that finds all motifs above score s, such as an exhaustive pattern driven search (Brazma *et al.*, 1998), and let us replace the percentage of false positive in its ROC curve with the expected number of those. The resulting 'E-ROC' curve holds exactly the same information as does the subtlety graph and there is a trivial equivalent transformation between the two. The emphasis, though, is slightly different: the E-ROC graph readily yields the expected number of random motifs we will encounter if we want to ensure that x percent of the implanted motifs will be found. The emphasis, though, is slightly different: the E-ROC graph readily yields the expected number of random motifs we will encounter if we want to ensure that x percent of the implanted motifs will be found. Note that using the previous definitions of M and S, the ROC curve essentially corresponds to the distribution of $\mathrm{P}(M > 0 \mid S)$.

The performance of a motif finder is determined by its reliability (how well does it find motifs) and its complexity (at what cost). Clearly, how well an algorithm detects a motif might vary with the motif-sample model. However, it also depends on how one defines what constitutes a detected motif. In particular, when looking for dim motifs one quickly realizes that perfectly recovering the set of

instances of a dim motif is a hopeless task due to good random matches of the pattern (for example, see the poor 'average performance coefficient', or apc, reported in Buhler (2001)).

Moreover, when we seek dim motifs we are bound to encounter a non-negligible number of 'false positives', or random motifs which are at least as significant as the ones we seek. Thus we contend that when evaluating the reliability of a dim motif finder one should ignore these false positives. Indeed, the more reliable the finder is, the more false positives it will pick up. We therefore argue that the algorithm only fails to detect the motif if it either completely misses the motif, or if it fails to compute its score correctly. More succinctly we say the algorithm *detects* the motif if it correctly computes its score. For example, regardless of whether the sequence count or total distance score is used, MULTIPROFILER (Keich and Pevzner, 2002) detects the motif in the FM 1600 problem over 99.4% of the time.

A curious implication of this definition is that a dim motif is not necessarily one that is difficult to detect; for example, pattern-driven algorithms that test all 4^l l-letter patterns are 100% reliable according to our definition, and for a small l the cost is not too bad. Of course, the end user might not be able to benefit from this kind of detection; nevertheless, the motif finder is not the weak link in this chain.

In the remainder of this paper we provide a case study of the subtlety analysis of the FM and VM models and continue to study the costs and reliability analysis of applying MULTIPROFILER to resolve these models. The latter part is clearly intended for people who read the account of MULTIPROFILER provided in the companion paper (Keich and Pevzner, 2002).

THE SUBTLETY GRAPH

Defining the twilight zone

Let $\mathcal{S} = \{S_1, \ldots, S_n\}$ be a random sample of n sequences each of which has N independent and uniformly distributed letters. We first look for the expected number of words W of length l whose score is m or better (since our scoring function is the total distance, this is equivalent to $d(W) \leq m$). Clearly, this expectation increases with m.

Let $X_i = d(W, S_i)$, then X_i are independent multinomial random variables. The distribution function of X_i depends on the overlap structure of W (Guibas and Odlyzko, 1981). Indeed, let B_i^j denote the word (of length l) starting at position j of the ith sequence, and let $Y_i^j = d(W, B_i^j)$. Then, since a mismatch between each of the l positions of B_i^j and W occurs independently of the others with probability 3/4, Y_i^j are binomial $b(l, 3/4)$ random variables for *any* W. Had Y_i^j been independent random variables, then the distribution function of $X_i = \min_j Y_i^j$ would have been readily available:

$$\begin{aligned} \mathrm{P}(X_i \geq r) &= \mathrm{P}\left(\cap_j \{Y_i^j \geq r\}\right) = \Pi_j \, \mathrm{P}(Y_i^j \geq r) \\ &= \left[\bar{F}_b[l, 3/4](r)\right]^{N-l+1}, \end{aligned}$$

where $\bar{F}_b[n, p](r) = \sum_r^n \binom{n}{k} p^k (1-p)^{n-k}$ is the probability that a binomial $b[n, p]$ random variable will be r or bigger. Of course, Y_i^j and Y_i^k are not independent for $|j - k| < l$ and the dependency varies with W. Having said that, since the dependency is rather local we can still obtain a decent approximation of the distribution function of X_i using the independence assumption (at least for a 'generic' W, cf. Table 1). Since, $d(W) = \sum_i X_i$ we obtain a readily computable estimate of the distribution of $d(\tilde{W})$ for a 'generic' word $\tilde{W}$ (Table 2). Assuming that the overwhelming number of words are generic ones, we can estimate the expected number of random motifs whose total distance is not bigger than m by $4^l \, \mathrm{P}(d(\tilde{W}) \leq m)$. Figure 1 shows results for sequences of length 1600 and 3000 and the corresponding locations of the twilight zone threshold. Table 2 demonstrates that the various estimates of the distribution of $d(\tilde{W})$ for a generic $\tilde{W}$ are in essential agreement, thus we are fairly confident in the resulting estimates of the expectations, albeit more can be done to quantify this confidence.

Scores of implanted motifs

Let $F_{d(P)}$ denote the distribution function of the total distance of the implanted motif P. We begin with the FM model and assume once again that P is 'generic' in the sense that we ignore its overlap structure. Clearly, $d(P) \leq nk$ (n sequences and exactly k mutations per instance) but the inequality could be strict due to good random matches of P. Let $Y_i^j = d(P, B_i^j)$, where B_i^j is the word that starts at position j of the ith sequence, S_i. Let P_i denote the instance of P that is implanted in the ith sequence, and let η_i denote the starting position of that implant. Then, conditional on $j = \eta_i$ (i.e. B_i^j is the ith instance, P_i), the random variable $Y_i^j \equiv k$, while its distribution conditional on $j \neq \eta_i$ is, as in the random case, binomial $b[l, 3/4]$. As in the random sample case, we are interested in the distribution of $X_i = d(P, S_i) = \min_j Y_i^j$, and we assume that for any fixed i, conditional on η_i, Y_i^j are roughly independent (ignoring the overlap). Under this assumption,

$$\mathrm{P}(X_i \leq r) \approx \begin{cases} 1 - \left[\bar{F}_b[l, 3/4](r+1)\right]^{N-l} & r < k \\ 1 & r \geq k \end{cases},$$

and as before we can use this estimated distribution function of X_i to approximate the distribution function of

Table 1. Distribution of the distance to one sequence

Pattern	Minimal distance to a sequence of length 1600 2	3	4	5	6	7	8
aaaaaaaaaaaaaaa	0.001	0.009	0.066	0.288	0.492	0.143	0.002
aaaaaaaccccccc	0.001	0.014	0.109	0.429	0.421	0.026	0.000
gcacggtttcataat	0.001	0.018	0.148	0.552	0.280	0.001	0.000
Randomized pattern	0.001	0.018	0.148	0.549	0.282	0.001	0.000
Theoretical estimate	0.001	0.018	0.148	0.550	0.282	0.001	0.000

The table provides a few examples of the empirical distribution of the minimal distance of a pattern to a random sequence of 1600 bp and contrasts it with the theoretical estimate obtained as explained in the text. In the 'randomized pattern' case we randomly generated a pattern for each randomly generated sequence. The entries in each line were summarized from 10^8 randomly generated 1600 bp sequences.

Table 2. The distribution function of the score of random motifs

Pattern	Total distance 80	82	84	86	88	90
aaaaaaaaaaaaaaa	$4.6 \cdot 10^{-15}$	$1.1 \cdot 10^{-13}$	$2.2 \cdot 10^{-12}$	$3.9 \cdot 10^{-11}$	$6.1 \cdot 10^{-10}$	$8.3 \cdot 10^{-9}$
aaaaaaaccccccc	$1.4 \cdot 10^{-11}$	$2.7 \cdot 10^{-10}$	$4.6 \cdot 10^{-9}$	$6.6 \cdot 10^{-8}$	$8.0 \cdot 10^{-7}$	$8.0 \cdot 10^{-6}$
gcacggtttcataat	$1.2 \cdot 10^{-9}$	$2.2 \cdot 10^{-8}$	$3.5 \cdot 10^{-7}$	$4.5 \cdot 10^{-6}$	$4.6 \cdot 10^{-5}$	$3.8 \cdot 10^{-4}$
Randomized pattern	$1.2 \cdot 10^{-9}$	$2.2 \cdot 10^{-8}$	$3.4 \cdot 10^{-7}$	$4.3 \cdot 10^{-6}$	$4.5 \cdot 10^{-5}$	$3.7 \cdot 10^{-4}$
Theoretical estimate	$1.2 \cdot 10^{-9}$	$2.2 \cdot 10^{-8}$	$3.4 \cdot 10^{-7}$	$4.3 \cdot 10^{-6}$	$4.5 \cdot 10^{-5}$	$3.7 \cdot 10^{-4}$
Monte Carlo	$1.2 \cdot 10^{-9}$	$2.2 \cdot 10^{-8}$	$3.5 \cdot 10^{-7}$	$4.4 \cdot 10^{-6}$	$4.5 \cdot 10^{-5}$	$3.7 \cdot 10^{-4}$

The table provides a sample of empirical values of $F_{d(W)}$, the distribution function of $d(W)$: the total distance of the word W to a sample of twenty 1600 bp sequences. In the first three rows W is specific, while the last four rows provide various estimates for the score of a generic $\tilde{W}$. The first five rows were computed from the corresponding entries in Table 1 using $d(W) = \sum_1^n X_i$. Notice that a direct Monte Carlo estimate of $F_{d(\tilde{W})}$ is not easy to obtain as the probabilities we are trying to estimate are as small as $5 \cdot 10^{-10}$. Thus to obtain a more direct Monte Carlo estimate of $F_{d(\tilde{W})}$ than the one derived from Table 1 (row 5), we estimated $F_{d(W_j)}$ for 10^4 randomly generated words W_j, and reported $F_{d(\tilde{W})}(m) = \sum_j F_{d(W_j)}/10^4$ in row 6. Each such $F_{d(W_j)}$ was computed from the estimated distribution function of $X_i = d(W_j, S_i)$ which in turn was evaluated using $4 \cdot 10^4$ random sequences S.

$d(P) = \sum_i X_i$. Note, however, that unlike in the previous discussion, the sequences now contain the P_is, and since there is an overlap between those words and some of the random words, the X_i are no longer independent. Nevertheless, typically $N \gg l$ and the effects of these weak dependencies should not be of any significance. This assertion is supported by empirical data as explained in Figure 3.

With the same definitions for the VM model, conditional on $j = \eta_i$, Y_i^j is binomial $b[l, \rho]$, while its distribution conditional on $j \neq \eta_i$ is binomial $b[l, 3/4]$. Ignoring the overlap issue we find that $\mathrm{P}(X_i \geq r) \approx \left[\bar{F}_b[l, 3/4](r)\right]^{N-l} \cdot \bar{F}_b[l, \rho](r)$, which combined with $d(P) = \sum_i X_i$ yields an estimate of $F_{d(P)}$ (see Figure 3 for an example).

PERFORMANCE ANALYSIS OF MULTIPROFILER (FM MODEL)

In analyzing the complexity of MULTIPROFILER we assume the input to the algorithm consists of a random sample (generated according to the sample model) *without* any motif implanted in it. This simplifies the cost analysis and moreover, since the fraction of the time the algorithm spends on the implanted motif itself is typically marginal, it introduces only a negligible error into the computed complexity.

The complexity of counting wordlets

Consider first the case where s, the syllable size, equals k, the wordlet size. In other words, we only look for complete wordlets, not parts of them. If the distance between the word B and the reference word A is strictly less than k, then B does not contain any wordlet that is disjoint (totally different) from A. Thus, we can restrict our attention to a shortened list of neighbors, namely to $\mathcal{J}'_i = \{B : k \leq d(A, B) \leq \alpha\}$ (instead of considering $\mathcal{J}_i = \{B : d(A, B) \leq \alpha\}$). Recall that α is typically set to $2k$ for the FM model.

Let $Z_i^j = d(A, B_i^j)$, then B_i^j (the word that starts at position j of S_i) contains exactly $\binom{Z_i^j}{k}$ k-wordlets which are disjoint from A. Thus, the number of (disjoint)

k-wordlets in the sequence S_i that MULTIPROFILER counts is $\sum_{B_i^j \in \mathcal{J}'_i} \binom{Z_i^j}{k}$. The expectation of this random variable is

$$\mathrm{E} \sum_j \binom{Z_i^j}{k} 1_{\{k \le Z_i^j \le \alpha\}} = (N - l + 1) \sum_{m=k}^{\alpha} \binom{m}{k} P_b[l, 3/4](m),$$

where $P_b[n, p](m)$ is the binomial probability, $\binom{n}{m} p^m (1-p)^{n-m}$, and 1_X is the indicator function of the set X.

Since one can show that generating the lists $\mathcal{J}'_i(A)$ for all $N-l+1$ reference words can be done in $\mathrm{O}(N^2 n)$ time, and since the counting itself takes O(1) per wordlet, the expected complexity of the counting task is given by

$$(n-1)(N-l+1)^2 \left[\sum_{m=k}^{\alpha} \binom{m}{k} P_b[l, 3/4](m) \right] \cdot \mathrm{O}(1) + \mathrm{O}(N^2 n).$$

In the case of our challenge problem ($k = 4, \alpha = 8$) the first term is dominant and it amounts to roughly $1.6 \cdot 10^8$. Note that explicit costs should be calibrated against the particular machine the algorithm is running on. For the case $s < k$ we redefine the neighbors' list to be $\mathcal{J}'_i = \{B \in S_i : s \le d(A, B) \le \alpha\}$, and with $Z_i^j = d(A, B_i^j)$, we note that B_i^j has $\binom{Z_i^j}{s}$ (s-)syllables that are disjoint from (the corresponding syllables of) A. Each such syllable of B_i^j appears in $\binom{l-s}{k-s} 3^{k-s}$ wordlets that are disjoint from A. Thus, B_i^j contributes 1 to the count of $\binom{Z_i^j}{s}\binom{l-s}{k-s}3^{k-s}$ wordlets that are disjoint from A (note that the same wordlet can receive up to $\binom{k}{s}$ contributions). The expected overall complexity of the counting process in this case is therefore:

$$nN^2 \left[\sum_{m=s}^{\alpha} \binom{m}{s} \binom{l-s}{k-s} 3^{k-s} P_b[l, 3/4](m) \right] \cdot \mathrm{O}(1)$$

This amounts to $4.8 \cdot 10^{10}$ in the case of our challenge problem when we choose $s = 2$. This is evidently much higher than in the case $s = k = 4$. It is exactly this difference in the 'fixed costs' that makes the case $s = k$ a better choice for our challenge problem when we move from using one reference sequence to using n reference sequences.

Computing the score of the modified words

Recall that we compute the score (total distance) of A_γ, the reference word A modified by the disjoint wordlet γ, provided $C(\gamma)$, the count of γ, is β or higher. The cost of scoring these putative patterns is therefore proportional to $\mathrm{E}\left[\, |\{\gamma : C(\gamma) \ge \beta\}| \,\right]$, the expected number of wordlets whose count reaches β. For the case $s = k$, which we begin with, $C(\gamma)$ counts the number of sequences S_i for which there exists a word $B \in \mathcal{J}'_i$ with $\gamma \subset B$ (i.e. B contains the wordlet γ).

Fix a reference word A, and let γ be a disjoint k-wordlet. Then, the probability that γ will be part of the word B_i^j and that $B_i^j \in \mathcal{J}'_i$, i.e. of the event $E_i^j = \{\gamma \subset B_i^j \in \mathcal{J}'_i\}$, is

$$\mathrm{P}(E_i^j) = 4^{-k} \cdot F_b[l-k, 3/4](\alpha - k), \qquad (1)$$

where $F_b[n, p](m)$ is the binomial distribution with parameters $[n, p]$ evaluated at m. Assuming that the events E_i^j are 'roughly independent', we can estimate, p_r, the probability that γ is present in $\mathcal{J}'_i$ as

$$p_r = \mathrm{P}(\cup_j E_i^j) \approx 1 - \left[1 - \mathrm{P}(E_i^j)\right]^{N-l+1}. \qquad (2)$$

Clearly, $C(\gamma)$ is a binomial $b[n-1, p_r]$ random variable. In particular, $\mathrm{P}(C(\gamma) \ge \beta) = \bar{F}_b[n-1, p_r](\beta)$, and since A has $3^k\binom{l}{k}$ disjoint wordlets of size k, and there are $N-l+1$ different reference words,

$$\mathrm{E}\left[|\{\gamma : C(\gamma) \ge \beta\}|\right] = (N-l+1) \cdot 3^k \binom{l}{k} \cdot \bar{F}_b[n-1, p_r](\beta). \qquad (3)$$

Now that we know the expected number of words A_γ whose score will be computed, we need to determine the 'typical' cost of actually computing $d(A_\gamma)$[¶]. A naive implementation would require $O(nNl)$, however we can do significantly better than that. Setting[‖] $\alpha \ge 2k$, $d(A_\gamma)$ can be estimated from $\cup_i \mathcal{J}_i$ in only $O(k|\cup_i \mathcal{J}_i|)$. There is a price to be paid for this significant saving: $\widehat{d(A_\gamma)}$, this estimator of $d(A_\gamma)$ might be strictly bigger than its estimated target. However, it can easily be demonstrated that if $A_\gamma = P$ (which is the major case of concern) then $\widehat{d(A_\gamma)} = d(A_\gamma)$. In other words, the aforementioned 'shortcut' does not affect the reliability of the algorithm for the FM model discussed here.

Since $\mathrm{E}\left[|\cup_i \mathcal{J}_i|\right] = (n-1)(N-l+1) \cdot F_b[l, 3/4](\alpha)$ we are tempted to say that the overall complexity of computing the scores is of the order of:

$$\left[N \cdot 3^k \binom{l}{k} \cdot \bar{F}_b[n-1, p_r](\beta)\right] \cdot [knN \cdot F_b[l, 3/4](\alpha)]. \qquad (4)$$

This is essentially correct though some explanation is in order. The problem is that $|\{\gamma : C(\gamma) \ge \beta\}|$ is positively correlated with $|\cup_i \mathcal{J}_i|$, so we cannot simply multiply the two expectations. However, a large deviation argument for example can show that the probability that $|\cup_i \mathcal{J}_i|$

[¶] More generally, if ψ is our adopted scoring function then we should evaluate the cost of computing $\psi(A_\gamma)$.

[‖] typically $\alpha = 2k$ is the only reasonable choice for this neighborhood defining distance

deviates outside of the interval $\left((1-\varepsilon)\,\mathrm{E}\left[|\cup_i \mathcal{J}_i|\right], (1+\varepsilon)\,\mathrm{E}\left[|\cup_i \mathcal{J}_i|\right]\right)$ decays exponentially fast (with nN), so the estimate (4) is indeed valid to first order. We omit the work spent on ranking the total distance of the patterns which pass the threshold as this is typically much smaller than computing the total distance itself.

In order to avoid a notational nightmare we study the case $s < k$ by way of example with $s = 2 < k = 4$. That is, we seek to modify 4-wordlets based on conserved pairs. For a word $B \in \mathcal{J}'_i$ let $C(\gamma, B)$ denote the number of syllables of size s (pairs) of the wordlet γ that are present in B. For example, if m of γ's letters are preserved in B, then $C(\gamma, B) = \binom{m}{2}$. Let γ be a disjoint wordlet of the reference word A, and for $m = 1, 2, 3, 4$ let $p_{c_m} = \mathrm{P}\left(C(\gamma, B) = \binom{m}{2} \text{ and } B \in \mathcal{J}'_i\right)$. Then, for $m = 2, 3, 4$:

$$p_{c_m} = \sum_{j=0}^{k-m} P_t[k, 1/4, 1/4, 1/2](m, j, k-m-j) \\ F_b[l-k, 3/4](\alpha - (k-j)),$$

where $P_t[M, p_1, p_2, p_3](i_1, i_2, i_2) = \frac{M!}{i_1! i_2! i_3!} p_1^{i_1} p_2^{i_2} p_3^{i_3}$ is the trinomial probability function, and $p_{c_1} = 1 - p_{c_2} - p_{c_3} - p_{c_4}$.

Recall that $C(\gamma, S_i)$, the count of γ in the sequence S_i, is defined as $\max_{B\in\mathcal{J}'_i} C(\gamma, B)$. For $m = 1, 2, 3, 4$, let $q_m = \mathrm{P}(C(\gamma, S_i) = \binom{m}{2})$. Then, using our standard approximate independence assumption,

$$q_m = \Big[\sum_{j=1}^{m} p_{c_j}\Big]^{N-l+1} - \Big[\sum_{j=1}^{m-1} p_{c_j}\Big]^{N-l+1}. \qquad (5)$$

For $m = 1, 2, 3, 4$, let $X_m = |\{i : C(\gamma, S_i) = \binom{m}{2}\}|$. Then, the random vector (X_1, X_2, X_3, X_4) has a multinomial distribution with parameters $[n-1; q_1, q_2, q_3, q_4]$, and since $C(\gamma) = 6X_4 + 3X_3 + X_2$, finding $\mathrm{P}(C(\gamma) \geq \beta)$ is now an elementary computation. The rest follows as in (3) and (4) with $\bar{F}_b[n-1, p_r](\beta)$ replaced by the appropriate $\mathrm{P}(C(\gamma) \geq \beta)$.

Motif detection rate

We next study the reliability, or the motif detection rate starting again with the case $s = k$. We stress that the results in this section are independent of the particular scoring function used. In particular they hold for the sequence-count scoring function (Buhler and Tompa, 2001).

Assume that the reference word A satisfies $d(A, P) = k$. This will be the case, for example, if A coincides with P_1. Let $\gamma = \gamma(A)$, be the correct modification of the 'mutated' wordlet of A (Keich and Pevzner, 2002). Let p_w be the probability that γ is preserved in P_i ($i \geq 2$):

$$p_w = \mathrm{P}(\gamma \subset P_i \in \mathcal{J}'_i | d(A, P) = k) = \binom{l-k}{k} / \binom{l}{k}.$$

Even if one or more of γ's positions are mutated in P_i, γ can still be present in $\mathcal{J}'_i$ provided there exists a (random word) $B \in \mathcal{J}'_i$ which 'preserves' γ. The probability of such an event was essentially computed in (2) (see also (1)). Thus,

$$p_r = \mathrm{P}(C(\gamma, S_i \setminus \{P_i\}) = 1) \\ \approx 1 - \left[1 - 4^{-k} \cdot F_b[l-k, 3/4](\alpha - k)\right]^{N-l}.$$

Since the random words are essentially independent of the P_is, the probability that γ will be counted in S_i is:

$$p_{wr} = \mathrm{P}(C(\gamma, S_i) \geq 1 | d(A, P) = k) \\ \approx p_w + (1 - p_w) p_r.$$

In the case of our challenge problem, $p_w \approx 0.24$, $p_r \approx 0.046$ and $p_{wr} \approx 0.276$. The total number of sequences in which γ is counted, $C(\gamma)$, is therefore essentially a binomial $b[n-1, p_{wr}]$ random variable, so

$$p_d \stackrel{d}{=} \mathrm{P}(C(\gamma(A)) \geq \beta | d(A, P) = k) \\ \approx \bar{F}_b[n-1, p_{wr}](\beta).$$

The reason we say essentially binomial, is that there are some marginal interactions between random words which overlap the P_is. If $N \gg l$ these dependencies should be negligible.

Let A^j be the jth reference word (assuming S_1 is our reference sequence, $A^j = B_1^j$), and let η denote the starting position of P_1. Let $E_j = \{d(A^j, P) = k, C(\gamma(A^j)) \geq \beta\}$, i.e. E_j is the event 'the motif will be picked up using the reference word A^j'. Then,

$$\begin{aligned} \mathrm{P}(E_j | \eta = m) &= \mathrm{P}(E_j | d(A^j, P) = k, \eta = m) \\ &\quad \cdot \mathrm{P}(d(A^j, P) = k | \eta = m) \\ &= \mathrm{P}(C(\gamma(A^j)) \geq \beta | d(A^j, P) = k) \\ &\quad \cdot \mathrm{P}(d(A^j, P) = k | \eta = m) \\ &= p_d \cdot \begin{cases} P_b[l, 3/4](k) & j \neq m \\ 1 & j = m \end{cases} \end{aligned} \qquad (6)$$

As usual, we assume that given η, $d(A^j, P)$ are essentially independent random variables. Moreover, we assume that conditional on $d(A^j, P) = k$, $C(\gamma(A^j))$ are iid random variables**, and therefore the events E_j are roughly independent conditional on $\eta = m$. Thus, the probability that using one reference sequence, the algorithm will detect the pattern P is:

$$\begin{aligned} \mathrm{P}\left(\cup_j E_j\right) &= \mathrm{P}\left(\cup_j E_j | \eta = m\right) \approx 1 - \Pi_j \mathrm{P}(E_j^c | \eta = m) \\ &= 1 - [1 - P_b[l, 3/4](k) \cdot p_d]^{N-l} \cdot (1 - p_d), \end{aligned} \qquad (7)$$

**For a typical application this is a fairly harmless assumption.

Table 3. Detection rates: theoretical estimates contrasted with simulations ($s = 4$)

	β - the score computing threshold										
	0	1	2	3	4	5	6	7	8	9	10
Estimate (1)	1.000	0.998	0.985	0.940	0.839	0.677	0.478	0.288	0.146	0.061	0.021
Observed (1)	1.000	0.998	0.985	0.939	0.836	0.672	0.472	0.283	0.144	0.061	0.021
Estimate (20)	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.957	0.721	0.356
Observed (20)	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.994	0.927	0.676	0.334

The Table contrasts the theoretical estimates of MULTIPROFILER's motif detection rate (as a function of β) with empirically obtained data. The motif model is our FM challenge problem and MULTIPROFILER is set with syllable size $s = 4$. The empirical distributions were obtained from 10^6 randomly generated samples of the model. In parentheses are the numbers of reference sequences used.

Table 4. Detection rates: theoretical estimates contrasted with simulations ($s = 2$)

	β - the score computing threshold										
	48	51	54	57	60	63	66	69	72	75	78
Estimate (1)	1.000	0.999	0.996	0.984	0.952	0.883	0.768	0.613	0.440	0.280	0.156
Observed (1)	1.000	0.999	0.996	0.984	0.950	0.880	0.763	0.608	0.435	0.275	0.152
Estimate (20)	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.966
Observed (20)	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.995	0.945

Same caption as for Table 3 only the syllable size MULTIPROFILER uses is $s = 2$.

where E_j^c is the complement of the event E_j. Table 3 contrasts these estimates with empirical results obtained by Monte Carlo methods.

We now return to the case $s = 2 < k = 4$ to find its detection rate. Suppose that the reference word A satisfies $d(A, P) = k$, and let $\gamma = \gamma(A)$ (the correct modification of A). Let

$$p_{w_m} \stackrel{d}{=} \mathrm{P}(C(\gamma, P_i) = \tbinom{m}{2}) = \frac{\binom{k}{m}\binom{l-k}{k-m}}{\binom{l}{k}}.$$

As before, the count of (the syllables of) γ might increase due to random words. The probabilities of these events were essentially computed in (5), with the original q_m replaced with $q_m = \mathrm{P}(C(\gamma, S_i \setminus \{P_i\}) = \binom{m}{2})$ and $N-l+1$ replaced by $N - l$. Thus,

$$p_{wr_m} = \mathrm{P}(C(\gamma, S_i) = \tbinom{m}{2}) = p_{w_m} \sum_1^m q_j + q_m \sum_1^{m-1} p_{w_j}.$$

Let $Y_m = |\{i : C(\gamma, S_i) = \binom{m}{2}\}|$, then (Y_1, Y_2, Y_3, Y_4) is essentially a multinomial random vector with parameters $[n - 1; p_{wr_1}, p_{wr_2}, p_{wr_3}, p_{wr_4}]$, and since $C(\gamma) = 6Y_4 + 3Y_3 + Y_2$, this yields an effective way to compute $p_d \stackrel{d}{=} \mathrm{P}(C(\gamma) \geq \beta | d(A, P) = k)$. The rest of the computation is exactly as in (6) and (7) (p_d is different though). Table 4 contrasts these estimates with their Monte Carlo obtained analogs.

Using multiple reference sequences

Our implementation of MULTIPROFILER using, say, n_r reference sequences is a naive one. Thus, for the *same* β, the complexity increases by a factor of n_r. The exact detection rate has so far eluded the authors. We do however have a 'ballpark' estimate for that which is more of an empirical result. The difficulty in finding the overall detection rate is that the events E_i = {signal was detected using the ith sequence as reference} are not independent events and the correlations are not at all clear. However, as extensive Monte Carlo tests indicate, using the independence assumption, we can obtain somewhat reasonable estimates of the detection rates we seek. Tables 4 and 3 provide the evidence, and Figure 1 in the companion paper (Keich and Pevzner, 2002) demonstrates in this context the advantage of the $s = 4$ variant over the $s = 2$ variant due to the cheaper 'fixed costs'.

MULTIPROFILER AND THE VM MODEL

Recall that in the VM model each instance of P is generated by randomly mutating each position at a rate ρ. Based on our experience with the FM model, we chose to use n reference sequences and $s = k$ for the VM model. Our implementation is a naive one, essentially enumerating over 'reasonable' values of k ($k = 0, 1, \ldots, K$). For each such value of k we have to set α_k, the analogue of α in the FM model, and β_k, the analogue of β in the FM model. Note that in dealing with the FM

model it was reasonable to set $\alpha = 2k$, however this value will not necessarily be optimal for the VM case. As of writing this paper given the parameters of the problem we pick the 'right' thresholds by a trial and error process based on the analysis that follows.

In adapting MULTIPROFILER to the VM model we had to adjust the way we estimate the score, $d(A_\gamma)$, as it is no longer the case that $\widehat{d(A_\gamma)} = d(A_\gamma)$ when A_γ coincides with P_i (the section on computing the scone of the modified words). Other than this comment, the complexity of the VM variant is the same as that of the FM variant subject to the obvious necessary summation over $k = 0, \ldots, K$. As for the detection rate, we next evaluate that fairly accurately per one reference sequence. Then, as in the FM model, assuming signal detections in different reference sequences are independent, we obtain cruder estimates of the overall detection rates.

Let A, be a reference word with $d(A, P) = k$, and let $\gamma(A)$ be the correct modification of the 'mutated' wordlet of A. Let p_{w_k} be the conditional probability that $\gamma(A)$ will be counted in P_i, i.e. that $\gamma(A) \subset P_i$ and that $d(P_i, A) \leq \alpha_k$, given that $d(A, P) = k$. Then,

$$\begin{aligned} p_{w_k} &= \mathrm{P}(C(\gamma(A), P_i) = 1) | d(A, P) = k) \\ &= P_b[k, \rho](0) \cdot F_b[l - k, \rho](\alpha_k - k). \end{aligned}$$

Let p_{r_k} be the probability that $\gamma(A)$ will be detected in the random word $B \in \mathcal{J}'_i$, given that $d(A, P) = k$. Then,

$$\begin{aligned} p_{r_k} &= \mathrm{P}(C(\gamma, B) = 1 | B \neq P_i, d(A, P) = k) \\ &= 4^{-k} F_b[l - k, 3/4](\alpha_k - k). \end{aligned}$$

Then,

$$\begin{aligned} p_{wr_k} &\stackrel{d}{=} \mathrm{P}(C(\gamma, S_i) = 1 | d(A, P) = k) \\ &\approx 1 - (1 - p_{r_k})^{N-l}(1 - p_{w_k}). \end{aligned}$$

Let A^j be the jth reference word (say, $A^j = B_1^j$), and let $E_j = \left\{\exists k \in \{0, \ldots, K\} : d(A^j, P) = k, C(\gamma(A^j)) \geq \beta_k\right\}$. Then, for $k \in \{0, 1, \ldots, K\}$,

$$\begin{aligned} \mathrm{P}(E_j | d(A^j, P) = k) &= \mathrm{P}(C(\gamma(A^j)) \geq \beta_k | d(A^j, P) = k) \\ &\approx \bar{F}_b[n - 1, p_{wr_k}](\beta_k). \end{aligned}$$

Thus, with η denoting the starting position of P_1,

$$\begin{aligned} \mathrm{P}(E_j | \eta = m) &= \sum_{k=0}^{K} \mathrm{P}(E_j | d(A^j, P) = k, \eta = m) \\ &\qquad \cdot \mathrm{P}(d(A^j, P) = k | \eta = m) \\ &= \sum_{k=0}^{K} \bar{F}_b[n - 1, p_{wr_k}](\beta_k) \\ &\qquad \cdot \begin{cases} P_b[l, \rho](k) & j = m \\ P_b[l, 3/4](k) & j \neq m \end{cases}. \end{aligned}$$

Assuming that given η the events E_j are 'roughly' independent[††], we have:

$$\begin{aligned} \mathrm{P}\left(\cup_j E_j\right) &= \mathrm{P}\left(\cup_j E_j | \eta = m\right) \approx 1 - \Pi_j \, \mathrm{P}(E_j^c | \eta = m) \\ &= 1 - \left(1 - \mathrm{P}(E_j | \eta \neq j)\right)^{N-l} \\ &\qquad \times (1 - \mathrm{P}(E_j | j = \eta)). \end{aligned}$$

When applied to the VM 1600 problem (20 sequences of 1600 bp and $\rho = 0.3$), MULTIPROFILER has an observed detection rate of 98%[‡‡] at a running time of just over an hour on a 500 MHz G4.

ACKNOWLEDGEMENTS

We would like to thank the anonymous referees for their numerous insightful comments.

REFERENCES

Brazma,A., Jonassen,I., Eidhammer,I. and Gilbert,D. (1998) Approaches to the automatic discovery of patterns in biosequences. *J. Comput. Biol.*, **5**, 279–305.

Buhler,J. (2001) *Search Algorithms for Biosequences Using Random Projection*, Ph.D. Thesis, University of Washington.

Buhler,J. and Tompa,M. (2001) Finding motifs using random projections. In *Proceedings of the Fifth Annual International Conference on Computational Molecular Biology (RECOMB-01)*. ACM Press, Montreal, pp. 69–76.

Eskin,E., Gelfand,M. and Pevzner,P. (2002) Genome-wide analysis of bacterial promoter regions. *submitted*.

Guibas,L. and Odlyzko,A. (1981) String overlaps, pattern matching and nontransitive games. *J. Combin. Theor.* Series A, **30**, 183–208.

Keich,U. and Pevzner,P. (2002) Finding motifs in the twilight zone. *Bioinformatics*, **18**, 1374–1381.

Pevzner,P. and Sze,S. (2000) Combinatorial approaches to finding subtle signals in DNA sequences. In *Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology*. AAAI Press, San Diego, pp. 269–278.

Stormo,G. (2000) DNA binding sites: representation and discovery. *Bioinformatics*, **16**, 16–23.

Swets,J. (1988) Measuring the accuracy of diagnostic systems. *Science*, **240**, 1285–1293.

Vanet,A., Marsan,L., Labigne,A. and Sagot,M. (2000) Inferring regulatory elements from a whole genome. an analysis of *Helicobacter pylori* σ^{80} family of promoter signals. *J. Mol. Biol.*, **297**, 335–353.

[††] This assumption can lead to non-negligible errors when there are many random words A^j with $d(A^j, P) = k$.

[‡‡] Empirical result based on 10^6 randomly generated samples of our challenge problem and a particular choice of MULTIPROFILER's parameter setting.