BIOINFORMATICS

Vol. 14 no. 6 1998
Pages 498–507

Detecting periodic patterns in biological sequences

Eivind Coward[1,3] and Finn Drabløs[2]

[1]Department of Mathematical Sciences, Norwegian University of Science and Technology and [2]SINTEF UNIMED MR-Centre, N-7034 Trondheim, Norway

Received on January 19, 1998; revised on April 1, 1998; accepted on April 6, 1998

Abstract

Motivation: *The search for repeated patterns in DNA and protein sequences is important in sequence analysis. The rapid increase in available sequences, in particular from large-scale genome sequencing projects, makes it relevant to develop sensitive automatic methods for the identification of repeats.*

Results: *A new method for finding periodic patterns in biological sequences is presented. The method is based on evolutionary distance and 'phase shifts' corresponding to insertions and deletions. A given sequence is aligned to itself in a certain sense, trying to minimize a distance to periodicity. Relationships between different such periodicity measures are discussed. An iterative algorithm is used, and the running time is nearly proportional to the sequence length. The alignment produces a periodic consensus pattern. A 'phase score' is used to indicate a statistical significance of the periodicity. Three examples using both DNA and protein sequences illustrate how the method can be used to find patterns.*

Availability: *On request from the authors.*

Contact: *eivindc@math.ntnu.no; finn.drablos@unimed.sintef.no*

Introduction

The search for repeated patterns in DNA and protein sequences is an important problem in biology. When the patterns to look for are unknown and the repeats are not exact, this can be difficult.

We present a method for searching for such repeats. It works best when they are approximate tandem repeats, i.e. the repeated units occur one after another, or with few letters between. This is a common situation in practice. The method is pragmatic in the sense that it is reasonably fast, but does not guarantee to find all repeats with up to a given number of mismatches.

Another algorithm for finding approximate tandem repeats has been presented previously (Benson and Waterman, 1994). It involves a rigorous alignment by dynamic programming, but certain heuristics have to be made to decide where and what to align. A simple but crucial step is the creation of a consensus pattern from an alignment, and we use essentially the same principle for creating our consensus pattern.

Phase alignment and distance to periodicity

Our goal is to identify periodic patterns that may be partially hidden by evolution. We need a measure of the degree of periodicity of a given DNA or protein sequence, and we will formulate this in terms of distance to the nearest periodic sequence.

A DNA sequence is considered here as a sequence from the four-letter alphabet {A, C, G, T}. Similarly, a protein sequence can be considered as a sequence from a 20-letter alphabet. Other alphabets can also be used, e.g. the two-letter purine–pyrimidine alphabet. We therefore consider the general case of sequences from an arbitrary finite alphabet $\mathcal{A}$ of N letters, and denote by $\mathcal{F}$ the set of all finite sequences of letters from $\mathcal{A}$. A sequence in $\mathcal{F}$ is typically written $\mathbf{a} = a_1 a_2 \ldots a_n$, where $a_i \in \mathcal{A}$, and $n = |\mathbf{a}|$ is the length of the sequence.

For comparing biological sequences, it is natural to use as our distance function a metric based on sequence alignment, first introduced by Sellers (1974). Smith and Waterman (1981) have shown that the Needleman and Wunsch maximum similarity alignment is equivalent to the minimization of the metric of Sellers. For the algorithm presented here, we could also have used an equivalent similarity score.

The sequence metric depends on a choice of a basic metric d on the alphabet $\mathcal{A}$ and a gap penalty. A dynamic programming technique is then used to minimize the sum of pairwise distances and gap penalties in the two sequences. This minimum defines the distance D, and it can be shown to be a metric (see Waterman, 1989).

For each period length $p \in \mathbb{N}$, let $\mathcal{F}_p$ be the set of all p-periodic sequences in $\mathcal{F}$:

$$\mathcal{F}_p = \{\mathbf{a} \in \mathcal{F} \mid a_{i+p} = a_i, i = 1, 2, \ldots, |\mathbf{a}| - p\}$$

[3]To whom correspondence should be addressed

A natural measure of 'nearness to p-periodicity' for a sequence $a \in \mathcal{F}$ is the aligned distance to the nearest p-periodic sequence:

$$D_p(\mathbf{a}) = \min_{b \in \mathcal{F}_p} D(\mathbf{a}, \mathbf{b})$$

Unfortunately, this is very time consuming to compute. For a linear gap penalty function, the best alignment algorithms are $O(n^2)$ (Waterman, 1989) (the length of the sequences b under consideration will be near n). Since we do not know the minimizing b in advance, we will have to test a lot of such sequences. The straightforward approach is to test all N^p sequences, which is a huge number, even for modest p. It is possible to avoid testing all sequences by using cleverer algorithms, but it is unlikely that a very efficient algorithm can be found.

Instead, we use a different kind of alignment especially tailored for this problem. It is not equivalent to the Needleman–Wunsch alignment, but it is based on the same ideas of an initial metric d on the alphabet and a mechanism similar to gaps.

We first extend the domain of the alphabet metric d to sequences of the same length by:

$$d(\mathbf{a}, \mathbf{b}) = \sum_i d(a_i, b_i)$$

This is the same as the metric D when there are no gaps.

Consider a sequence $\mathbf{a}$ and a fixed period length p. Assume that the length of $\mathbf{a}$ is a multiple of p, $|\mathbf{a}| = n = rp$, deleting the last few elements of $\mathbf{a}$ if necessary. We divide $\mathbf{a}$ into consecutive subsequences $\mathbf{a}^1, \mathbf{a}^2, \ldots, \mathbf{a}^r$ of length p each and use the notation:

$$\mathbf{a}^i = a_1^i a_2^i \ldots a_p^i = a_{(i-1)p+1} a_{(i-1)p+2} \ldots a_{ip-1}$$

As a measure of the 'mutual agreement' between the subsequences, we define:

$$M_p(\mathbf{a}) = \sum_{j<i} d(\mathbf{a}^i, \mathbf{a}^j)$$

If $\mathbf{a}$ is p-periodic, $\mathbf{a}^1 = \mathbf{a}^2 = \ldots = \mathbf{a}^n$, so that $M_p(\mathbf{a}) = 0$. If $\mathbf{a}$ is periodic except for a few substitutions, $M_p(\mathbf{a})$ will still be small. However, an insertion or a deletion in $\mathbf{a}$ will influence $M_p(\mathbf{a})$ greatly. We will therefore allow for a phase shift ϕ_k of each $\mathbf{a}^i$. For an arbitrary sequence $\mathbf{b} = b_1 b_2 \ldots b_p$, we define:

$$\phi_k \mathbf{b} = b_{p-k+1} b_{p-k+2} \ldots b_p b_1 b_2 \ldots b_{p-k}, \quad k = 0, 1, \ldots, p-1$$

In other words, a circular right shift of k positions (where k is always taken modulo p, to allow for, for example, negative shifts). When a phase shift k_i is applied to each $\mathbf{a}_i$, we get a phase alignment of $\mathbf{a}$, denoted by:

$$\Phi_{\mathbf{k}} \mathbf{a} = \phi_{k_1} \mathbf{a}^1 \ldots \phi_{k_r} \mathbf{a}^r, \qquad \mathbf{k} = [k_1, k_2, \ldots, k_r]$$

The effect of an insertion or a deletion in a periodic sequence can be compensated in the succeeding subsequences by a right or left shift. Motivated by this, we define:

$$M_p^*(\mathbf{a}) = \min_{\mathbf{k}} M_p(\Phi_{\mathbf{k}} \mathbf{a}) = \min_{k_1, \ldots, k_r} \sum_{j<i} d(\phi_{k_i} \mathbf{a}^i, \phi_{k_j} \mathbf{a}^j)$$

Since each of the r subsequences can have p different phases, there are p^r different phase alignments, and again it is normally too time consuming to test all of them. The following iterative method is therefore used.

Let $k_1 = 0$. For each $m = 2, \ldots, r$, choose k_m to minimize $\sum_{j<i\le m} d(\phi_{k_j} \mathbf{a}^j, \phi_{k_i} \mathbf{a}^i)$ with $k_1, \ldots, k_{m-1}$ fixed. For the following iterations, minimize the sum for all $j < i \le r$. Note that only $\sum_j d(\phi_{k_j} \mathbf{a}^j, \phi_{k_m} \mathbf{a}^m)$ needs to be calculated in each step (and this can be done quickly, see the section on implementation and complexity).

We continue to iterate until the sum does not decrease (or until a maximum number of iterations have been performed). The iteration will always converge, since M_p decreases for each phase change. However, it is not guaranteed to converge to $M_p^*(\mathbf{a})$ corresponding to an optimal phase alignment. When the optimal phase alignment is known (by testing all possibilities, necessarily for very modest r), preliminary studies indicate that in most cases an optimal or near-optimal alignment is found.

Determining a consensus pattern

When we have found our phase alignment of $\mathbf{a}$, we want to determine a candidate for the underlying periodic pattern (assuming it exists). A natural choice is the sequence $\mathbf{p}$ which minimizes $\sum_i d(\phi_{k_i} \mathbf{a}_i, \mathbf{p})$. This is simple and fast, since each of the p letters in $\mathbf{p}$ can be found independently of the others (by trial and error).

The minimized sum can now be interpreted as another measure of the degree of periodicity of the sequence. We define:

$$C_p(\Phi_{\mathbf{k}} \mathbf{a}) = \min_{\mathbf{p}} \sum_i d(\phi_{k_i} \mathbf{a}^i, \mathbf{p})$$

and

$$C_p^*(\mathbf{a}) = \min_{\mathbf{k}} C_p(\Phi_{\mathbf{k}} \mathbf{a})$$

Even if we knew the Φ that minimizes M_p, this would not necessarily minimize C_p. If we want to improve our phase alignment after determining our consensus pattern, we can go back and align our sequence to the fixed pattern $\mathbf{p}$, possibly reducing C_p (but M_p might increase). A new, possibly different, consensus can be determined, and the whole process may be iterated. Again, it will always converge, since C_p decreases if a change is made.

The connection between two periodicity distances

We have the following connection between M_p and C_p.

Proposition 1 For every phase alignment $\Phi_{\mathbf{k}}$ of a sequence **a** of length $n = pr$ we have

$$\frac{r}{2} C_p(\Phi_{\mathbf{k}}\mathbf{a}) \leq M_p(\Phi_{\mathbf{k}}\mathbf{a}) \leq (r-1) C_p(\Phi_{\mathbf{k}}\mathbf{a}) \quad (1)$$

The inequalities cannot be improved, except that for odd r, we can in some cases change $r/2$ on the left-hand side to at most $(r + 1)/2$.

Proof: Let $\mathbf{b} = \Phi_{\mathbf{k}}\mathbf{a}, \mathbf{b}^i = \phi_{k_i}\mathbf{a}^i$. By the triangle inequality for the metric d we have:

$$\begin{aligned} M_p(\mathbf{b}) &= \sum_{j<i} d(\mathbf{b}^i, \mathbf{b}^j) \leq \frac{1}{2} \sum_{j \neq i} (d(\mathbf{b}^i, \mathbf{p}) + d(\mathbf{p}, \mathbf{b}^j)) \\ &= (r-1) \sum_i d(\mathbf{b}^i, \mathbf{p}) = (r-1) C_p(\mathbf{b}) \end{aligned}$$

which proves the right inequality. Since **p** is chosen to minimize $\Sigma\, d(\mathbf{b}^i, \mathbf{p})$, we get

$$\begin{aligned} C_p(\mathbf{b}) &= \sum_i d(\mathbf{b}^i, \mathbf{p}) \leq \min_j \sum_i d(\mathbf{b}^i, \mathbf{b}^j) \\ &\leq \frac{1}{r} \sum_{i,j} d(\mathbf{b}^i, \mathbf{b}^j) = \frac{2}{r} M_p(\mathbf{b}) \end{aligned}$$

using the fact that the minimum is less than or equal to the average. This proves the left inequality.

The following example demonstrates that the right inequality cannot be improved. Let $\mathbf{b}^1 = \mathbf{b}^2 = \ldots = \mathbf{b}^{r-1} =$ AAA … A, and $\mathbf{b}^r =$ CCC … C, to be specific. Then **p** = AAA … A (for $r \geq 3$). Normalizing d(A, C) to one, we find that $C_p(\mathbf{b}) = p$ and $M_p(\mathbf{b}) = (r-1)p$, giving $M_p(\mathbf{b}) = (r-1)C_p(\mathbf{b})$.

The following example demonstrates that the left inequality cannot be improved if we assume that r is even. Let $\mathbf{b}^1 = \mathbf{b}^2 = \ldots = \mathbf{b}^{r/2} =$ AAA … A, and $\mathbf{b}^{r/2+1} = \ldots = \mathbf{b}^r =$ CCC … C. Normalizing as before, we get $M_p(\mathbf{b}) = r^2p/4$. The consensus **p** can be chosen as some combination of these two letters (the triangle inequality implies that other letters are no better). In any case, we get $C_p(\mathbf{b}) = rp/2$ and $M_p(\mathbf{b}) = \frac{r}{2} C_p(\mathbf{b})$.

For odd r, the subsequences AAA … A and CCC … C split into uneven parts of $(r + 1)/2$ and $(r - 1)/2$, yielding $M_p(\mathbf{b}) = (r + 1)\,(r - 1)p/4$ and $C_p(\mathbf{b}) = (r - 1)p/2$; hence, $M_p(\mathbf{b}) = \frac{r+1}{2} C_p(\mathbf{b})$. This is the 'worst case' for a two-letter alphabet. However, if r is divisible by, say, three, and there are three letters in the alphabet with mutual distances one, an equal division into three groups will give $M_p(\mathbf{b}) = \frac{r}{2} C_p(\mathbf{b})$ again. The same is true for any number, so the exact left inequality will depend on the alphabet and the metric in addition to r. □

In particular, the proposition is valid for the phase alignment $\Phi_{\mathbf{k}_M}$ minimizing M_p, and for the phase alignment $\Phi_{\mathbf{k}_C}$ minimizing C_p, but these two alignments may also be compared.

Corollary 1 For a sequence **a** of length $n = pr$ we have

$$\frac{r}{2} C_p^*(\mathbf{a}) \leq M_p^*(\mathbf{a}) \leq (r-1) C_p^*(\mathbf{a})$$

The inequalities cannot be improved, except that for odd r, we can in some cases change $r/2$ on the left-hand side to at most $(r + 1)/2$.

Proof: The inequalities follow from equation (1), using the minimum properties $M_p^*(\mathbf{a}) \leq M_p(\Phi_{\mathbf{k}_C}\mathbf{a})$ and $C_p^*(\mathbf{a}) \leq C_p(\Phi_{\mathbf{k}_M}\mathbf{a})$. To show that they cannot be improved, we can use the same examples as in the preceding proof, since all the subsequences are invariant to phase shifts. □

Phase coincidence as a measure of periodicity

Testing of the described method on both simulated and real sequences shows that the quantities M_p and C_p work well for finding the underlying periodicity when p is known. However, comparing values for different p is difficult and does not always give a clear conclusion, and deciding whether the result is significant is even harder. For this purpose, we use another quantity, which measures to what extent the phases agree. This quantity has the advantage of being easier to analyze statistically.

Suppose we have found a candidate $\Phi_{\mathbf{k}}$ for the best alignment to period p, where $\mathbf{k} = [k_1, k_2, \ldots, k_r]$ If the sequence is exactly p-periodic, we have $k_1 = k_2 = \ldots = k_r$. If the sequence is almost p-periodic, we expect that many of the k_i agree. Let the indicators $I_{i,j}$ be defined by:

$$I_{i,j} = \begin{cases} 1 \text{ if } k_i = j, \\ 0 \text{ if } k_i \neq j, \end{cases} \quad 1 \leq i \leq r, 0 \leq j \leq p-1$$

Furthermore, we define the phase score as:

$$Y = \sum_{j=0}^{p-1} X_j^2, \qquad X_j = \sum_{i=1}^{r} I_{i,j}$$

While $\Sigma X_j = r$ always, Y indicates whether or not the sum is evenly distributed among the X_j. A large phase score means that the alignment is concentrated on a few phases, which we will interpret as a sign of periodicity. An underlying assumption is that insertions and deletions are relatively rare.

When the sequence is random and uncorrelated (but not necessarily with a uniform letter distribution), the X_j have a multinomial distribution, and each of the p phases occurs with the same probability $1/p$, independently of the phase of other subsequences. With some modification (see below), this is also used as a model for other non-periodic sequences.

Under these assumptions, straightforward but tedious calculations show that:

$$\mathrm{E}(Y) = r\left(1 + \frac{r-1}{p}\right)$$
$$\mathrm{Var}(Y) = \frac{2r(r-1)}{p}\left(1 - \frac{1}{p}\right)$$

We want to compute p-values of different scores, so we need the cumulative distribution function. Finding a closed formula is difficult or impossible, but with some combinatorial reasoning one can group the p^r different phase alignments into a computer-manageable number of groups with the same score and thus tabulate an exact distribution function for each occurring (r,p) pair (see the Appendix).

There may be many such pairs, but when $r >> p$ we need not calculate the exact distribution. It then follows from standard asymptotic theory that the X_j tend to a multi-normal distribution. Moreover, if

$$\mathbf{Z} = \frac{1}{\sqrt{r}}\left[X_0 - \frac{r}{p}, X_1 - \frac{r}{p}, \ldots, X_{p-1} - \frac{r}{p}\right]$$

then $p\mathbf{Z}^{\mathrm{T}}\mathbf{Z}$ tends to be χ^2 distributed with $p - 1$ degrees of freedom. But

$$p\mathbf{Z}^{\mathrm{T}}\mathbf{Z} = \frac{p}{r}\sum_{j=0}^{p-1}\left(X_j - \frac{r}{p}\right)^2 = \frac{p}{r}\left(\sum X_j^2 - \frac{2r}{p}\cdot r + \frac{r^2}{p^2}\cdot p\right) = \frac{p}{r}Y - r$$

which is obtained by using $\Sigma X_j = r$.

When analyzing protein-coding DNA sequences, it is no surprise to find that there is almost always a significant periodicity three, and this also affects the scores for all period lengths divisible by three. We are not normally interested in this periodicity, and we want to filter out this effect in order to be able to detect codon periodicities of six, nine and so on. Our method is to divide the aligned rows into three groups. Group 0 consists of rows with phase 0, 3, 6, etc., group 1 contains the rows with phase 1, 4, 7, etc., and group 2 the rows with phase 2, 5, 9, etc. Owing to the underlying periodicity three, most of the rows are usually in one of the three groups. We choose the biggest group and discard the other rows. The phase score is now computed using the $p/3$ different phases of the remaining group, thus replacing p by $p/3$ in the calculation of the p-value.

Of course, this method can also be used to filter out the effect of a dominant periodicity of length other than three.

Phase alignment with gap penalty

Sequence alignment involves a gap penalty to balance the search for good letter agreement against the avoidance of large and numerous gaps. The gap penalty for each gap is added to the letter distances. It is an increasing function of the gap length, often of the form $g(k) = a + bk$ for a gap of length k.

In the same manner, it is possible to introduce a gap penalty in the phase alignment, where a gap corresponds to the phase difference between two consecutive subsequences. This will disfavor frequent phase changes. For the distance to consensus, this can be introduced as:

$$\hat{C}_p(\mathbf{\Phi_k a}) = C_p(\mathbf{\Phi_k a}) + G(\mathbf{k})$$
$$G(\mathbf{k}) = \sum_{i=1}^{r} g(|k_i - k_{i-1}|_p)$$

where g is the gap penalty function and the phase difference $|k_i - k_{i-1}|p$ is the shortest distance between k_i and k_{i-1}, when counting modulo p (such that $|1 - 8|_9 = |8 - 1|_9 = 2$, for example). For the mutual distance we define:

$$\hat{M}_p(\mathbf{\Phi_k a}) = M_p(\mathbf{\Phi_k a}) + crG(\mathbf{k})$$

where c is a constant. If we choose $\frac{1}{2} \le c \le 1 - \frac{1}{r}$, equation (1) implies that

$$\frac{r}{2}\hat{C}_p(\mathbf{\Phi_k a}) \le \hat{M}_p(\mathbf{\Phi_k a}) \le (r-1)\hat{C}_p(\mathbf{\Phi_k a})$$

The inequalities are still sharp in the same sense as in the Proposition, because $G(\mathbf{k})$ may be zero.

If we compare this with the Smith–Waterman alignment distance D_p to the periodic sequence with the same alphabet metric and the same gap penalty, it is not difficult to see that D_p defines a lower limit for the phase alignment distance $\hat{C}_p$ (because every phase alignment can be transformed into an ordinary sequence alignment in a natural way). On the other hand, if the periodicity is distinctive and the gaps are not too frequent (rarely more than one in each period), the difference between D_p and $\hat{C}_p$ tends to be small.

While introducing a gap penalty might look like a good idea for obtaining a better phase alignment, it has serious drawbacks. It is difficult for the iterative algorithm to find the optimal alignment because it would usually have to 'climb a hill' (make several expensive phase shifts) before the reward is paid. Moreover, the alignment to consensus afterwards has the same problem and is no longer a straightforward one-sweep procedure. In addition, computing p-values from phase scores would be much more complicated because the phases are dependent on each other.

Implementation and complexity

We have implemented the phase alignment, the consensus search and the phase coincidence score in a C program running on a UNIX machine. The program can analyze either one specific sequence or a collection of sequences in a data-

base file. Since a periodic pattern is often present only locally in a sequence, every sequence can be split into subsequences of a given length, partially overlapping if desired. The user can specify the alphabet metric, which period lengths to investigate and several other parameters. The amount of output can be regulated, ranging from detailed phase alignment information for each iteration for a single sequence and a few period lengths, to a one-line summary for each sequence when investigating a large database.

The following main steps are carried out for each period length p given, and for each (sub)sequence:

- phase alignment;
- determination of the consensus pattern;
- adjustment of phase alignment according to the consensus pattern (optional);
- computation of phase score with corresponding p-value.

Most of the consumed time is spent on the phase alignment. As above, the sequence length is $n = rp$, deleting leftovers if p does not divide n evenly. For one iteration of the phase alignment procedure, each of the r subsequences is aligned by minimizing the sum of the distances to the rest of them. The minimization implies trying all p phases, but the sum of distances can in fact be calculated in $O(p)$ time if the distances are tabulated for each of the N letters in the alphabet. This gives a complexity of each iteration of $O(rp^2 + rpN)$, or $O(np + nN)$, where the last term is for making the distance table. The number of iterations seems to be slowly growing with n. An upper limit may of course be specified. Experience shows that DNA sequences of 10 000 bases may need up to 30–40 iterations, 100 base sequences less than five.

The consensus pattern is determined by trying each of the N bases for each of the p positions. Again the distance sum is calculated by using the table, so the complexity is just $O(pN)$, independent of sequence length (because the table is made during the alignment).

The optional adjustment to the consensus pattern can be done if the alignment itself is of interest (e.g. for calculating the phase score) and one considers the consensus to be a better criterion than the alignment algorithm. The phase of each subsequence is chosen to minimize the distance to the consensus, this is done in $O(np)$ time.

After computing the phase score, we want a p-value to indicate its significance. If $r \geq 5p$, we use the χ^2 approximation. The χ^2 distribution function can be evaluated efficiently (see Press *et al*., 1992). If $r < 5p$, the approximation is not so good, and we need the exact distribution. We store the distributions in tabular form in a separate file for each (r,p) pair. If the requested table does not exist, it is created once and for all (see the Appendix).

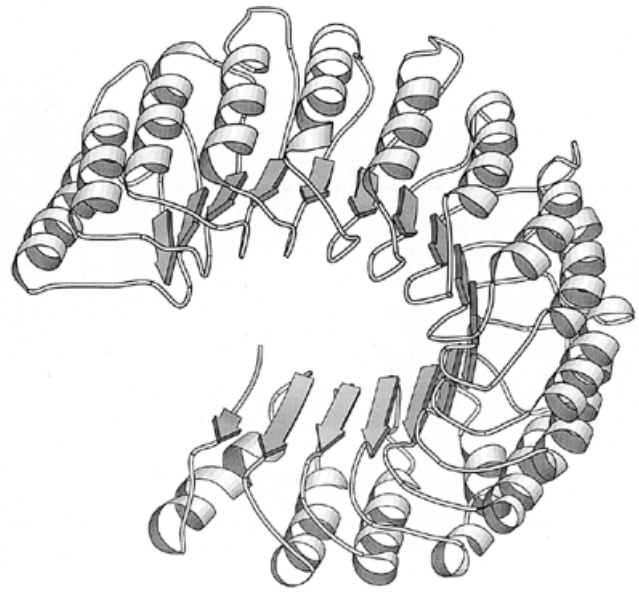

Fig. 1. Structure of the ribonuclease inhibitor. The structure of the ribonuclease inhibitor with secondary structure elements highlighted. The drawing was generated using Molscript (Kraulis, 1991).

Since running time is almost linear with respect to sequence length (only a slight increase for longer sequences due to more iterations), the length of split sequences is not very important for the time consumption. The choice should be based on the expected length of periodic patterns.

Examples

Three examples from using this approach on real DNA and protein sequences are shown. The first example is the porcine ribonuclease inhibitor. The protein sequence of this molecule [SwissProt (Bairoch and Boeckmann, 1992) entry RINI_PIG] is 456 residues long, and the coding region from the corresponding DNA entry (EMBL entry SSRI) was used for this analysis. The ribonuclease inhibitor is one of several proteins with leucine-rich repeats. In the ribonuclease inhibitor, there is an alternating pattern of 'A' and 'B' repeats, and the repeats are 29 and 28 residues long, respectively. This makes an effective repeat length of 57 residues, and there are 7.5 such repeats. The three-dimensional (3D) structure of this molecule is known (Kobe and Deisenhofer, 1993), and shows that the protein is horseshoe shaped with the interior face formed by a parallel β sheet of 17 individual β strands, and the exterior from 16 α helices (Figure 1). Therefore, each 'AB' repeat corresponds to a strand–helix–strand–helix motif.

The sequence was analyzed as described in this paper. Periods from nine to 300 in steps of three were tested, in

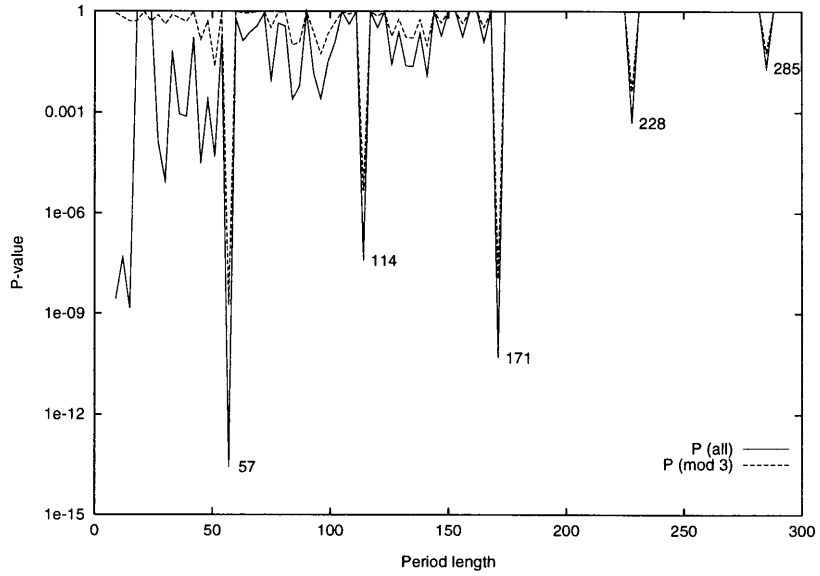

Fig. 2. The p-values for phase scores of the ribonuclease inhibitor. The p-values of all phase scores, P(all), and of phase scores fro m phase shifts corrected for a periodicity of three, P(mod 3), as described in the text. The p-values are plotted as a function of the assumed period.

```
Period length 171:
  Iteration  1:   Dm = 1746
  Iteration  2:
Converged during iteration 1
   1: :a.cc...a.....c...c........tc........ag..g......gg..t. - ...cag.....  0
 172: C.......a...gca.t.....t.....c.......g......a.....g..cc. - t.t.c......  0
 343: :.............................t..c..gc.a...aa......a... - ....c..c..g  0
 514: C..ca....c..gt....ctg.g.g..t.a.a.c......agt.......ga... - ...........  0
 685: :.......t.c..gt..g..a..tc....a....c..g..g...t.....c..c.. - ..g...ac.ac  0
 856: A...cg.gt....t.t..t.g.g....t.agg.....aa.ag........gcag. - ..g......gc  0
1027: t.ga..a.t.....cC......c.gctg.t....ta..a.a..ta...g.c...g - ...a....tac156
Cons: GCTGGAGACGCTCAGGCTGGAGAACTGCGGCCTGACCGCCGCCGGCTGCAAGGAC - CCCGGCTGCCA
Mutual distance: Dm = 1746. Distance to consensus: Dc = 381
Phase score with n=7 rows and p=171 phases: 37 (normalized score: 896.857)
Calculating distribution (n=7, p=171)
P-value from exact distribution: 4.8e-11
Distribution of phases modulo 3: 7, 0, 0.
Analysis of rows with phase 0 modulo 3 only:
Phase score with n=7 rows and p=57 phases: 37 (normalized score: 294.286)
Calculating distribution (n=7, p=57)
P-value from exact distribution: 1.1e-08
```

Fig. 3. Alignment of repeats in the ribonuclease inhibitor. The size of the alignment has been reduced by cutting out part of the sequence (indicated with a '–'). For each line of the alignment, the start of the section in the full sequence and the phase shift are shown. Sequence positions identical to the consensus sequence are shown as a dot, for non-conserved positions the base code is given. An uppercase base code or a ':' indicates the real start of the sequence, this becomes relevant if circular shifts are used in the alignment. The zero phase shifts in this example indicate simple repeats of constant length.

order to focus on properties related to the protein structure. The p-values corresponding to phase scores for different periodicities are plotted (Figure 2), showing very small p-values for periodicities that are multiples of 57, representing different integer fractions and combinations of the 'true' repeat, which is found at periodicity 171 (corresponding to a protein sequence of 57 residues) (Figure 3). For the true periodicity, all phase shifts in the repeat region are zero (because the repeat length is constant), for the other periodicities there are non-zero phase shifts. This makes the correct repeat length easy to identify.

The second example is the PIR1 gene from the complete genome sequence of yeast (*Saccharomyces cerevisiae*). This gene codes for a heat shock protein, although the exact function of this protein is unknown. The protein is known to contain eight tandem repeats, most of them 19 residues long, although repeat 3 is 24 residues long (corresponding to 57 and 72 bases, respectively). A separate investigation using

```
a)
   1: ...ggt..aact.atT.c...tt....c.a..ag.t.c.taca.g 30
  46: .aa.c...tgcca.t...Tactctt......tcggt....c..tt 27
  91: tc......tggt:.tt.ct.c..gg...agagag.cgc..c.at. 33
 136: ....................c.....c...g..g.tgc..t.:..  3
 181: ........................c.aa:...a.......... 15
 226: ..c.............cG..gctg..gtc....a.......... 27
 271: tcc...:c.a.ga.t..ct.agct.ag...a.cg.tgca.cc.t. 39
 316: ........t.................c.aag..g.t:.......  9
 361: ....................c...G..gtc....a.......... 21
 406: ............A.a...g..gc.g..gtc....a.......... 33
 451: :ccact.cta..a.....g..gctg..gt.....a..........  0
 496: :................c..c..t......gt.g.tcca.tctc.  0
 541: ..c.............a...tt....t..g.a:a...cac...t 12
 586: ..cact:c.attatac.at..cc.g..c.ag........ac.a.t 39
 631: t..a...g.agt.g....ttag..Gacc.agtaa.tgc..aaaca 21
Cons: GGTCAAATCCAAGCCACTACTAAAACTACTTCTCCAATTGGTGAC

b)
   1: ...Tcc..ttt.a..c.at.a.cta..t.c.ag....gtataactgatt.ct.tt.. 54
  58: g..tcct...ag...aagaga.c.:tcggt....c..tt.aa.c...tgcca.t... 33
 115: ..c........t...:.....a.................................... 42
 172: .........a....:........t.................................. 42
 229: ..c..g..t..ct.a:.........................c..............c 42
 286: G....g..t..c.....a..c.....c...................t.........  0
 343: :.........a.............................................  0
 400: :.c.....a..t.....a.....................................  0
 457: :..........t...........t..............................c  0
 514: :.c..t..t..t.t....c.a.....c.....cac...t..c.............a  0
 571: ..c.tt.tac..t..c.a...ccag...c....ac.a.t..cActt.aactt.tg.a 15
 628: :..g.cc.agt.a......aaaca.g.a...gca...g.ac.tt.gaaatgaa.tta  0
Cons: ACTAAAACCACAGCTGCTGCTGTCTCTCAAATTGGTGACGGTCAAATCCAAGCCACT
```

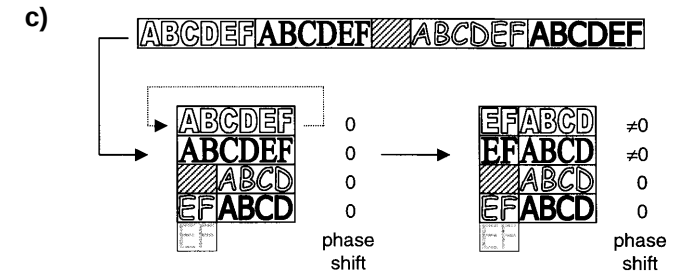

Fig. 4. Identification of repeats with phase shifts (PIR1). Testing for periodicity using different periods in PIR1. In (**a**), the assumed period is 45, and the apparent lack of any pattern in the phase shifts indicates that 45 is not a true periodicity of PIR1. This may be compared to (**b**), where the assumed period is 57. Although the first and the last lines of the alignment seem to be random, the regular pattern of the eight lines in the middle is a strong indication of a repeat. The phase shift in three of the lines is further explained in (**c**), where it is shown how a short insertion in the repeat region is compensated for with circular shifts.

the approach described here in an automatic large-scale screening process for the identification of repeats in genomes identified this as one of a large number of repeats in the yeast genome (Coward, 1998). To use it as an example in this study, a part of the gene sequence known to contain this repeat was analyzed as already described. The output for two different periodicities is shown (Figure 4). For a period of 45, there is no corresponding periodicity in the DNA sequence, and the phases are mainly random. This may be compared to the output for a periodicity of 57, which is equivalent to the dominating repeat length. Here the phases are zero for most of the repeat region. However, the increased length of repeat 3 is compensated for by a phase shift of 42 bases, corresponding to the difference between the dominating repeat length and the 15 extra bases found in repeat 3.

The final example shows how this tool may be used for analyzing protein sequences. An amino acid distance matrix was generated from the Blosum-50 mutation matrix (Henikoff and Henikoff, 1992) by using a normalized negation as described by Taylor and Jones (1993). This distance matrix was used to search the NRL_3D database (Pattabiraman *et al*., 1990) of protein sequences extracted from the PDB database (Bernstein *et al*., 1977; Abola *et al*., 1987). The complete sequences were tested for repeats. Several protein sequences were identified as repetitive, and one of these, the sequence for UDP-*N*-acetylglucosamine acyltransferase (PDB code 1LXA), was selected for further analysis. The plot of p-values versus assumed period length (Figure 5) shows the dominating repeat to be of length 6, with the next possibly significant score at length 18, indicating triplets of the basic repeat length. The listing of the search result for repeat length 18 (Figure 6) shows that the repeat is only weakly conserved with respect to sequence, it is probably found only in the first 180 residues, and the phase shifts indi-

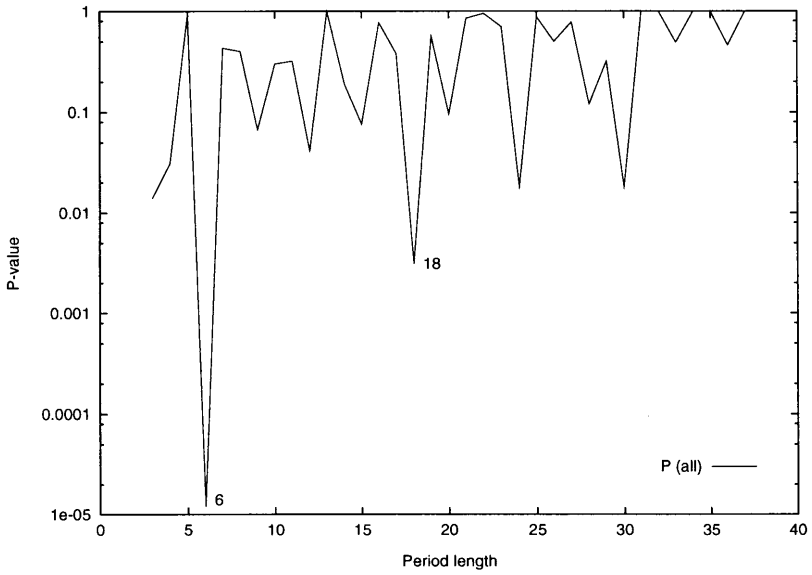

Fig. 5. Plot of phase score p-value versus assumed period length for the sequence of UDP-*N*-acetylglucosamine acyltransferase.

```
Period length 18:
  Iteration  1:   Dm = 5162
  Iteration  2:   Dm = 5117
  Iteration  3:   Dm = 5108
  Iteration  4:
Converged during iteration 3
    1: Midk.af.hptai.eega  0 X
   19: Si.anahigpfci.gp.v  0 X
   37: Ei.egtvlksh...ng..  0 X
   55: Ki.rdn.iyqfasigevn  0 X
   73: ei.drnQdlkyageptrv 12              X
   91: vq.gglRir...tihrg. 12              X
  109: k...dnllminahi.hd:  1  X
  127: ...nrcil.nnatl.g.C  1  X
  145: ii.gmta.hqf:s.ddfa  7        X
  163: m..gc.g..qdCiiga.v  7        X
  181: viaqgnhatpfg.nVppy  4     X
  199: rr.f.r.aitairIeglk  5      X
  217: liyr.gktlde.kpNayk  4     X
  235: Eiaela.type.kaftdf  0 X
Cons.: TVGSSSEVAESVVVASHT
Mutual distance: Dm = 5108. Distance to consensus: Dc = 584
Phase distribution: 5 2 0 0 2 1 0 2 0 0 0 0 2 0 0 0 0 0
Phase score with n=14 rows and p=18 phases: 42 (normalized score: 40)
P-value from exact distribution: 0.0031
```

Fig. 6. Alignment of possible repeats for period length 18 in the sequence of UDP-*N*-acetylglucosamine acyltransferase. The phase shifts indicate insertions after the first four repeats, the relatively random phase shifts for the last four segments (after residue number 180) indicate that this region is not repetitive. Please observe that the consensus sequence is estimated so that the total distance to all subsequences is minimized, therefore it may not correspond exactly to a more traditional consensus sequence based on the most frequent residue at each position. This can be seen in the first position of this alignment, where threonine (T) is used in the consensus, although glutamic acid (E) is the most frequent residue at this position.

cate two insertions after the first four 18 residue repeats. This may be compared to the 3D structure (Raetz and Roderick, 1995) of this protein. A TOPS (Flores *et al.*, 1994) cartoon of the structure (Figure 7) shows that the first domain of the structure is a left-handed parallel β helix, there are nine turns in this helix, and each turn consists of three β strands of six residues each. In the 3D structure itself (not shown), insertions after turn 4 and 5 can easily be seen. This shows that our tool identifies a repeat pattern which is consistent with 3D structure, and it is also consistent with previous studies of this pattern (Vuorio *et al.*, 1994). It is important to realize that this pattern was easily identified by automatic non-supervised screening of a database, despite the facts that the repeat region is only partly conserved, it contains insertions, and it does not cover the entire protein sequence that was tested.

Discussion

The method presented in this paper can be used as a tool to find repeated patterns in DNA and protein sequences. It turns

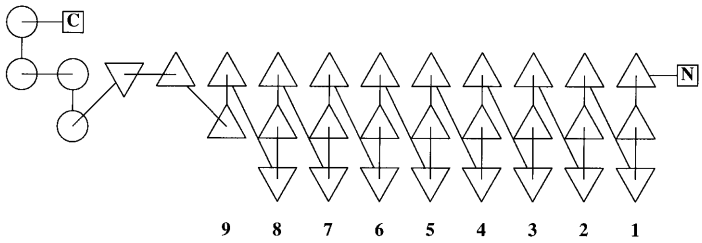

Fig. 7. A TOPS cartoon of the structure of UDP-*N*-acetylglucosamine acyltransferase. The β strands are shown as triangles, α helices as circles. The relative positions of the secondary structure elements correspond to the organization of the 3D structure. The figure shows clearly how the nine turns of the β helix are built from three β strands each.

out to be most effective for tandem repeats of at least 12–15 bases. When there are few insertions and deletions, the phase agreement is a good indicator of the presence of a repeat, and this agreement can be analyzed statistically. The sensitivity of the method is also good when there are many substitutions.

Alternatively, we can use the distance D_c to consensus (or the mutual distance D_m) as our indicator. Its statistical properties are more complicated and are not analyzed here. It is complementary to the phase agreement in the sense that it punishes substitutions, but not phase shifts, and in some situations this is important. An example is shown in Coward (1998), where we have just three long repetitions with a phase shift. Since only two phases agree, the phase agreement would not be very significant. In many cases, however, the phase agreement and the distance to consensus would both indicate the correct periodicity.

The process of phase shifting is useful because it is simple and splits the sequence into independent parts, making optimization as well as statistical analysis easier. It can be argued that phase shifts are artificial and do not correspond to our understanding of biological sequences. This is a price we have to pay for simplicity and efficiency, but for reasonably nice almost periodic sequences (few insertions and deletions compared to the period length) the resulting phase alignment is close to an ordinary alignment.

Once a consensus sequence is found, an ordinary alignment could be produced if desired, using, for example, the wraparound dynamic programming technique (see Benson and Waterman, 1994).

Further work should be carried out in order to create a fully automatic procedure to find and isolate repeated patterns in long sequences. Some attempts in this direction are made in Coward (1998).

The inclusion of gap penalties would also be a useful addition, but the global minimization problem of finding the optimal alignment has to be solved.

Acknowledgement

This work was supported by grant 100548/410 from the Norwegian Research Council.

Appendix: Computing the exact phase score distribution

Consider a phase alignment $\Phi_{\mathbf{k}}$, where $\mathbf{k} = [k_1, k_2, \ldots, k_r]$, and each $k_i \in \{0, 1, \ldots, p\}$. Assuming that every alignment is equally probable, we want to find the distribution function of the phase score Y, i.e. to count the number of phase alignments whose score is above (or below) a given value. There are p^r different phase alignments, so going through all of them is soon unrealistic (for example, $r = 50$ and $p = 30$ give 7×10^{73} combinations).

Let r_j be the number of rows with phase j, i.e. $r_j = \#\{i : k_i = j\}$, $j = 0, 1, \ldots, p-1$, $\Sigma_j r_j = r$. The phase score is then $y = \Sigma r_j^2$, regardless of the order of the phases k_i. The number of such phase alignments is the number of permutations of the given $k_1, k_2, \ldots k_r$, which is given by the multinomial coefficient:

$$\binom{r}{r_0, \ldots, r_{p-1}} = \frac{r!}{r_0! \ldots r_{p-1}!} = \binom{r}{r_1}\binom{r - r_1}{r_2}\binom{r - r_1 - r_2}{r_3} \ldots \binom{r_{p-1}}{r_{p-1}}$$

Moreover, the order of the r_j (i.e. the labeling of the different phases) does not matter either. Let p_i be the number of row counts r_j that equal i, i.e. $p_i = \#\{j : r_j = i\}$, $i = 0, 1, \ldots, r$, $\Sigma_i p_i = p$. The values $r_0, r_1, \ldots, r_{p-1}$ can be permuted in

$$\binom{p}{p_0, \ldots, p_r}$$

different ways. The total number of phase alignments with score greater than or equal to y is then

$$N(y) = \sum_{\substack{r_0 \le \ldots \le r_{p-1} \\ \sum r_j = r \\ \sum r_j^2 \ge y}} \binom{r}{r, \ldots, r_{p-1}} \binom{p}{p_0, \ldots, p_r}$$

The p-value for this score is $N(y)/p^r$. Note that even after this grouping of phase alignments, several terms in the sum may still correspond to the same score.

To make a table of the distribution, we evaluate the term for each $r_0 \le \ldots \le r_{p-1}$, with $\Sigma r_j = r$ and add it to the table in a position depending on Σr_j^2. The total number of terms equals the number of partitions of r into at most p positive integers or, equivalently, the number of partitions of r into any number of the integers from 1 to p. No closed formula for this number exists, but it is clearly bounded by the number $p(r)$ of all partitions of r into positive integers, which is approximated by the Hardy–Ramanujan formula:

$$p(r) \approx \frac{1}{4\sqrt{3r}} e^{\pi\sqrt{2r/3}}$$

(see Cohen, 1978, pp. 73–79). The example with $r = 50$ and $p = 30$ produced a sum with 202 139 terms, which can easily be handled by a computer. Much larger cases may require other techniques, like simulation, if a χ^2 approximation is not satisfactory.

References

Abola,E.E., Bernstein,F.C., Bryant,S.H., Koetzle,T.F. and Weng,J. (1987) Protein data bank. In Allen,F.H., Bergerhoff,G. and Sievers,R. (eds), *Crystallographic Databases: Information Content, Software Systems, Scientific Applications*. Data Commission of the International Union of Crystallography, Bonn, pp. 107–132.

Bairoch,A. and Boeckmann,B. (1992) The SWISS-PROT protein sequence data bank. *Nucleic Acids Res.*, **20**, 2019–2022.

Benson,G. and Waterman,M.S. (1994) A method for fast database search for all *k*-nucleotide repeats. *Nucleic Acids Res.*, **22**, 4828–4836.

Bernstein,F.C., Koetzle,T.F., Williams,G.J.B., Meyer, Jr., E.F. Brice,M.D., Rodgers,J.R., Kennard,O., Shimanouchi,T. and Tasumi,M. (1977) The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.*, **112**, 535–542.

Cohen,D.A. (1978) *Basic Techniques of Combinatorial Theory.* John Wiley, New York.

Coward,E. (1998) Mathematical Methods for Periodic Patterns in Biological Sequences. Dr. ing. Thesis 1998:13, Norwegian University of Science and Technology, Trondheim, Norway.

Flores,T.P., Moss,D.S. and Thornton,J.M. (1994) An algorithm for automatically generating protein topology cartoons. *Protein Eng.*, **7**, 31–37.

Henikoff,S. and Henikoff,J. (1992) Amino acid substitution matrices from protein blocks. *J. Proc. Natl Acad. Sci. USA*, **89**, 10915–10919.

Kobe,B. and Deisenhofer,J. (1993) Crystal structure of porcine ribonuclease inhibitor, a protein with leucine-rich repeats. *Nature*, **366**, 751–756.

Kraulis,P.J. (1991) MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallogr.*, **24**, 946–950.

Pattabiraman,N., Namboodiri,K., Lowrey,A. and Gaber,B. (1990) NRL_3D: a sequence-structure database derived from the protein data bank (PDB) and searchable within the PIR environment. *Protein Seq. Data Anal.*, **3**, 387–405.

Press,W.H., Teukolsky,S.A., Vetterling,W.T. and Flannery,B.P. (1992) *Numerical Recipes in C*. Cambridge University Press, Cambridge.

Raetz,C.R.H. and Roderick,S.L. (1995) A left-handed parallel 3 helix in the structure of UDP-N-acetylglucosamine acyltransferase. *Science*, **270**, 997–1000.

Sellers,P.H. (1974) On the theory and computation of evolutionary distances. *SIAM J. Appl. Math.*, **26**, 787–793.

Smith,T.F. and Waterman,M.S. (1981) Comparison of biosequences. *Adv. Appl. Math.*, **2**, 482–489.

Taylor,W.R. and Jones,D.T. (1993). Deriving an amino acid distance matrix. *J. Theor. Biol.*, **164**, 65–83.

Vuorio,R., Härkönen,T., Tolvanen,M. and Vaara,M. (1994) The novel hexapeptide motif found in the acyltransferases LpxA and LpxD of lipid A biosynthesis is conserved in various bacteria. *FEBS Lett.*, **337**, 289–292.

Waterman,M.S. (1989) *Sequence Alignments*. CRC Press, Boca Raton, FL, pp. 53–92.