

Discovering patterns in *Plasmodium falciparum* genomic DNA

Linda Stern ^{a,*}, Lloyd Allison ^b, Ross L. Coppel ^c, Trevor I. Dix ^b

^a Department of Computer Science and Software Engineering, The University of Melbourne, Melbourne, Victoria 3010, Australia

^b School of Computer Science and Software Engineering, Monash University, Clayton, Victoria 3800, Australia

^c Department of Microbiology, Monash University, Clayton, Victoria 3800, Australia

Abstract

A method has been developed for discovering patterns in DNA sequences. Loosely based on the well-known Lempel Ziv model for text compression, the model detects repeated sequences in DNA. The repeats can be forward or inverted, and they need not be exact. The method is particularly useful for detecting distantly related sequences, and for finding patterns in sequences of biased nucleotide composition, where spurious patterns are often observed because the bias leads to coincidental nucleotide matches. We show here the utility of the method by applying it to genomic sequences of *Plasmodium falciparum*. A single scan of chromosomes 2 and 3 of *P. falciparum*, using our method and no other a priori information about the sequences, reveals regions of low complexity in both telomeric and central regions, long repeats in the subtelomeric regions, and shorter repeat areas in dense coding regions. Application of the method to a recently sequenced contig of chromosome 10 that has a particularly biased base composition detects a long internal repeat more readily than does the conventional dot matrix plot. Space requirements are linear, so the method can be used on large sequences. The observed repeat patterns may be related to large-scale chromosomal organization and control of gene expression. The method has general application in detecting patterns of potential interest in newly sequenced genomic material. © 2001 Elsevier Science B.V. All rights reserved.

Keywords: Compression; Information theory; Repeated sequences; Pattern discovery

1. Introduction

During the last few years, the amount of DNA sequence material available has been increasing exponentially. A major challenge facing computational molecular biology is the post-sequencing analysis of patterns and motifs in genomic DNA sequences. Since patterns and motifs appear as related sequences at more than one point in the genome, one pattern-finding strategy is to look for repetition. Repetitions can have many different biological meanings. For example, regulatory elements, structural features of the DNA, and low complexity regions all show up as repetition [1]. Since repetition is rarely exact, the problem becomes one of finding related patterns whose location and degree of similarity are not known in advance.

The more general the repeat-finding strategy, the greater the possibility of finding new regions of biological interest. Specific motif searching engines, such as those described in [2], rely on prior biological knowledge, and are therefore limited to finding known patterns in new sequences. In some cases restrictive assumptions must be made, or the range of organisms restricted. The well-known dot matrix plot is a general method for looking at relatedness. In this method, subsequences are scored as ‘matches’ when they show at least as many matches, within a specified window, as a user-determined stringency level.

A particular difficulty of searching for relatedness within sequences with biased composition is that the nucleotide bias may obscure a real relationship between sequences. In dot matrix plots of biased sequences, chance ‘matches’ due to increased coincidental matching of bases appear as a noisy background, and can make it difficult to detect real repeats, particularly when the repeats are only approximate. Scientists working with *P. falciparum*, with an AT content of ~80% [3], experience a similar problem when doing BLAST

Abbreviations: DPA, dynamic programming algorithm; HMM, hidden Markov model; LZ, Lempel Ziv; PFSA, probabilistic finite state automaton.

* Corresponding author.

E-mail address: stern@unimelb.edu.au (L. Stern).

searches, as many ‘matches’ reflect only this compositional bias. Pre-filtering programs such as SEG [4] can remove such areas of low compositional complexity from the query sequence, but one consequence is that information that may be present in these sequences is ignored.

Information theory gives us a way out of this dilemma, by linking probability, complexity, and compression. Shannon showed that, in an optimal code, the length of a code word for an event of probability $p(E)$ is $-\log_2(p(E))$ bits [5]. When sending a message containing a DNA sequence, if the four bases were equally probable, the most efficient code, call it *code 1*, would be to assign a 2-bit code word to each base, e.g. 00, 01, 10 and 11. If, however, the probabilities of the bases were $\frac{1}{2}$, $\frac{1}{4}$, $\frac{1}{8}$ and $\frac{1}{8}$, giving them code words of length 1, 2, 3 and 3 bits, e.g. 0, 10, 110 and 111 (*code 2*) would result in an average message length of $1\frac{3}{4}$ bits per base, which is better than *code 1*, providing that these probabilities apply throughout the sequence. But if the data are really uniform, with all four bases equally probable, then *code 2* requires $2\frac{1}{4}$ bits per base on average, which is worse than using *code 1*. We can use message lengths to see which is the shorter code, and therefore a better model for a sequence.

The complexity of a sequence is defined to be its message length under an optimal code. Shannon shows how to calculate this length without actually encoding the sequence [5]. We do not usually encode sequences, but rather just calculate what the message length would be, using Shannon’s result. But for simplicity we often write of codes and message lengths as though the sequences were being encoded. In many cases it is not possible to calculate probabilities exactly or to compute an optimal code. However, a message length in any comprehensible code, i.e. a code that is decodable without hidden knowledge, is a valid upper bound on the complexity of a sequence, and there are many practical techniques for forming very good codes. It should also be mentioned that the technique of arithmetic coding [6] can in effect allocate a code word of non-integer length to an event, E , so it is not necessary to round $-\log_2(p(E))$ up. In most of the more complicated models and codes, the probabilities of the characters, (e.g. DNA bases), vary from position to position, for example being conditional on the previous k bases in the case of a k th-order Markov model for DNA.

Such models and codes depend on multi-state distributions, e.g. four-state for DNA and RNA, and 20-state for protein sequences. Wootton and Federhen [4,7] defined the notion of local compositional complexity, which is calculated from the multi-state distribution in sliding windows. The original application was to mask out low complexity regions which cause false-positives in searches of sequence databases, with a later application to modelling the domain structure of

proteins [1,8]. Pizzi and Frontali [9] have applied Wootton and Federhen’s SEG algorithm [4] to *P. falciparum* proteins, finding non-globular domains of low complexity that appear as insertions of material not found in homologous proteins in other species. Wan and Wootton [10] have also described a measure for global compositional complexity and used it to explore coding regions of DNA in a number of species. They found the median global compositional complexity of *P. falciparum* coding regions to be considerably lower than that of a number of other eukaryotic species.

Compression is a well-known computer science technique that takes advantage of repetition to reduce the size of a file. The theoretical upper limit of achievable compression, or entropy, is closely related to message length. Our interest in compression here is not for saving file space or communication bandwidth, but in measuring the fit between a model and a sequence. Agarawal and States [11] and Grumbach and Tahi [12] recognised the relevance of compression to pattern discovery in biological sequences. Loewenstern and Yianilos [13] modified a popular file compression algorithm for use with DNA sequences by allowing a certain number of mismatches against past ‘contexts’. Unfortunately their algorithm has several dozen parameters which do not have obvious biological interpretations. Rivals and Dauchet used a compression algorithm that joins nearby exact repeats, allowing some mismatch in the join, thereby making some allowance for mutations within approximate repeats [14]. The method we use in this paper explicitly models repeated subsequences and mutation in DNA.

Using information theory, we have developed a method for detecting similarities in DNA sequences [15,16]. Our method does not require any a priori knowledge about the motifs to be detected or about the sequences under observation, but can utilise such information where the intent is to search for more specific patterns. Importantly, our method does not require the user to guess at the level of similarity or to set parameters in advance. The method is inspired by the well-known Lempel Ziv (LZ) model used for data compression [17], where matches to previous subsequences are sought. Unlike LZ, the model we use for DNA sequences allows both inexact and reverse-complementary repeats within a sequence, as well as the forward and exact repeats used in data compression. Particular advantages of the method include its ability to simultaneously detect different kinds of repeating regions within a single scan of the same genome, the ability to differentiate and quantitate the degree of similarity between different related regions in the sequence, and the ability of the method to detect long, significant approximate repeats over a background of smaller repeat units. We demonstrate these advantages by showing the application of this compression tech-

nique to the entirety of chromosomes 2 and 3 of *P. falciparum* and to a fragment of chromosome 10. The genome of *P. falciparum* presents special challenges because of its biased nucleotide composition of 80% A + T [3].

2. Methods

2.1. A statistical model for DNA sequences

Our model of DNA [15,16] is based on very general biological knowledge as follows: DNA subsequences can be duplicated and, once there are two or more copies of a subsequence, the copies can individually accumulate mutations and hence diverge. In addition, the general composition of DNA in a particular organism may be biased and can be modelled, in part, by a simple ‘base model’. In our implementation we have used a low-order Markov model as the base model, but other models are possible. No additional biological information is inherent in our model, which considers a DNA sequence to be a mixture of (i) possibly biased DNA generated by a base model; and (ii) approximate repeats in either the forward or reverse-complementary senses. Each repeat is a copy of some earlier subsequence but can differ from the original by mutations (changes, insertions and deletions), as in sequence alignment. The model has a small number of parameters governing the base sub-model, the rate and length of repeats, and the rates of mutations within repeats. The parameters are estimated from the sequence itself by an expectation maximisation process [18–20].

Fig. 1 shows our DNA model as a probabilistic finite state automaton (PFSA), equivalently a hidden Markov model (HMM). The automaton is probabilistic in that the transitions out of each state have probabilities (not shown) associated with them. The base model is labelled ‘B’ and is treated as a ‘black box’ in the diagram. The current implementation allows the base model to be either a zero-order or a first-order Markov model; the work described here employs the latter option. A given sequence may be generated entirely by the base model. However, if there is an approximate repeat, its second occurrence can be generated by making a transition to the *R1* (repeat) state and stating the start position of the original occurrence. From there, state *R2* and associated transitions allow the subsequence to be copied, possibly with mutations, until eventually the model returns to the base model B. If the repeat unit is long enough and of sufficient fidelity, the explanation involving the repeat mechanism will be more probable than that using the base model only. A further set of states, *R1'* and *R2'* (not shown), allow for (approximate) reverse-complementary repeats. Some alternative architectures for the model have been examined and are continuing to be investigated.

The states associated with repeats amount to an alignment sub-model; here one sequence is being aligned with itself and it makes sense to align later parts of the sequence with earlier parts of the sequence. If a sequence is explained by paths through the repeat states, those parts of the explanation amount to a local alignment of the sequence against itself. The local alignments can be combined, not necessarily in order, and are shown in the two-dimensional plots described in the next section.

Each transition of the model has a probability associated with it; these probabilities are the model’s parameters. The probability of a path of transitions which generates a given DNA sequence is the product of the probabilities of the individual transitions making up the path. A dynamic programming algorithm (DPA) [15,16] can be used to find an optimal path. Given a path, the model’s parameters can be estimated from the transition frequencies in the path. The probabilities of the transitions out of a state correspond to a multinomial distribution and Boulton and Wallace [21] give the required theory. Changing the model’s parameters may cause a new path to become optimal which may lead to further changes of parameters and so on. This expectation maximisation process must converge and it does so quickly. In theory it could converge to only a local optimum, but this is not a problem in practice. The probability of a DNA sequence, and hence its compressed message length, can be calculated given a fully parameterised model. One could seek a single optimal

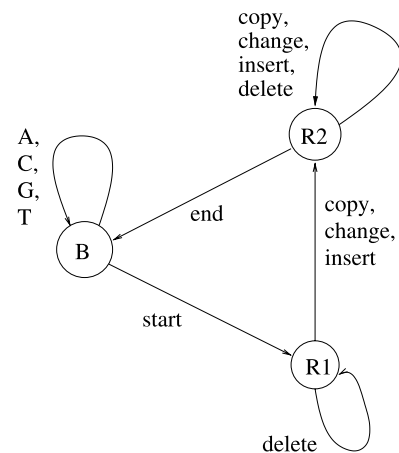


Fig. 1. Finite state machine for generating strings. From the base state, B, the machine can generate ‘random’ characters, returning to the base state. It can also start a repeat, moving to state *R1*, then to *R2*. From state *R2*, characters can be copied from the source substring, but characters can also be changed, inserted or deleted. The auxiliary state *R1* is simply there to ensure that invisible events are prohibited, i.e. at least one character must be output before returning to B. The repeat ends with a return to the base state. The base state is also the start and end state of the machine. Many variations on the ‘architecture’ of the machine are possible to incorporate prior knowledge while staying within the general framework.

path through the model to generate a given DNA sequence. However, it has been shown that comparison based on a single optimal alignment gives biased estimates of parameters and underestimates probabilities [21]. Two different paths are different hypotheses about how the model could create the data, so their probabilities can, and should, be added as has been done for pair-wise sequence alignment under PFSA [22], giving a less-biased estimate of probabilities and parameters [21]. Each state in a PFSA has only a finite number of transitions leading into it and path probabilities are combined where paths meet at a state; the computer implementation actually works with $-\log$ probabilities, for numerical reasons.

In reality, repeats accumulate in DNA over time, and can overlap in complex ways. Fully unravelling this history is a difficult combinatorial problem; the simplifications of our model form an acceptable approximation to reality, while allowing a reasonably efficient inference algorithm. For example, it operates ‘left to right’, not distinguishing between the situation where the subsequence α' is a copy of α , and vice versa.

2.2. Implementation of the statistical model

Overall compression gives a general picture of the presence of repeats in a sequence. We calculate the compressibility under an optimal code without actually encoding the data. A localised picture is often more useful in terms of biological relevance. We have found it useful to produce four distinct kinds of output from an input sequence of nucleotides. The outputs are: (1) a model, with probabilities for starting, continuing, and ending repeats, estimated using an expectation maximisation process [18,19]; (2) a single number measuring the compressibility of the entire sequence, expressed in bits per nucleotide, i.e. information content per nucleotide; (3) a plot showing how compressibility varies along the length of the DNA sequence, calculated in average bits per nucleotide in a local region; and (4) a two-dimensional plot showing how previous subsequences have contributed to compression of the sequence, and showing their location. In the two-dimensional plot, the probability that one subsequence contributed to another subsequence is shown by brightness. The plot is superficially similar to the familiar dot matrix plot, with grey-scale level indicate probability contributions within the statistical model. Two-dimensional plots, and also conventional dot-plots, are necessarily limited in resolution because of computer memory and disk sizes, with each dot representing hundreds or even thousands of bases for long sequences. In contrast, one-dimensional plots, and their file representations, are compact enough to be

kept at full resolution, allowing features to be located precisely, even in very long sequences.

The time complexity of our basic algorithm is quadratic in the length of the sequence, resulting in relatively slow performance on very long sequences. However, we have implemented a heuristic that speeds up processing significantly. Space requirements are linear in the length of the sequence, so the use of the heuristic makes it feasible to work on sequences of millions of nucleotides, i.e. whole chromosomes. We start with the assumption that most important approximate repeats will contain some small exact repeats, and use this assumption to determine where in the sequence to concentrate our search for approximate repeats. We construct a hash-table which contains k -tuples and their locations in the sequence, where k is a constant, typically in the range 6–16. A match in the sequence with a k -tuple ‘turns on’ a region of ± 5 nucleotides around the match. A region is turned off when its paths are making only a negligible contribution to the probability of the sequence [15]. In contrast to the k -mismatch problem, which has been used to find approximate repeats in DNA sequences [23], no assumptions are made about the overall number of mismatches. Regions can grow, shrink, merge, or be turned off. Adjusting the value of k allows the algorithm to process long sequences quickly, at some loss of accuracy. Note that k is a parameter of the speed-up heuristic; it is not a model parameter, as such. Ideally one would use $k = 1$, given a fast enough computer. A useful technique, particularly for newly sequenced material, is to perform a rapid scan of a large sequence, such as an entire chromosome, in an approximate manner, to give a general idea of where the interesting regions are located, and then to ‘zoom in’ on these areas, detecting more subtle patterns by using less restrictive assumptions.

3. Results and discussion

3.1. Compressibility across *P. falciparum* chromosomes

The DNA sequences for chromosomes 2 and 3 of *P. falciparum* have been determined [24,25]. Chromosome 2 is 947,103 base pairs in length, with a biased base composition of 81% A + T content overall [24]; chromosome 3 is 1,060,106 nucleotides in length, with 80% A + T content [25]. The biased base composition of these chromosomes implies that the sequences will be compressible, and poses challenges in looking for additional patterns over and above this base level of compressibility.

Using our model to compress the sequences of chromosomes 2 and 3, we achieved compression of 1.556 bits per nucleotide for chromosome 2 and 1.586 bits

per nucleotide for chromosome 3, setting the heuristic parameter $k = 16$ nucleotides, i.e. this is an upper bound. The values obtained are better than the theoretical maximum compression achievable without invoking repeats or context, which is $\sum_i -p_i \log_2 p_i$, or 2.0 bits per nucleotide for random nucleotide sequences, and ~ 1.7 bits per nucleotide for sequences with the biased A + T composition of the *P. falciparum* chromosomes. The compression observed using our method represents the combined effects of the biased base composition, regions of low complexity, regions containing multiple tandem short repeat elements, and approximate repeats of long subsequences. Sequences that are associated with lower information content are more compressible, indicating relatedness to other sequences by some form of repetition.

Compressibility is not uniform across the chromosome, as seen when the average information content across a small window is plotted for the length of the chromosome. It is clear from the plots of compressibility along both chromosomes 2 and 3 (Fig. 2a and b) that the ends are more compressible than the middle, and that there are small, clearly defined regions of compressibility in the middle.

We have also explored chromosome 3 in the context of chromosome 2, allowing sequences in chromosome 3 to be repeats from either chromosome 2 or 3 (or both), by concatenating the two chromosomes and running the algorithm. The overall compression of the concatenated chromosomes is 1.538 bits per nucleotide, less than the figure obtained for either chromosome alone. The additional compression must be interpreted to mean that there are regions that show similarity across both chromosomes.

The compressibility plot of chromosome 3, computed in the context of chromosome 2 (Fig. 2c), shows additional areas of low information content, over and above those seen on the plot of chromosome 3 without this context (Fig. 2b). The additional areas include sizeable regions at the ends of the chromosome and one particularly striking spike of low information content that can be seen ~ 100 kb from the 5' end of the chromosome. Further analysis shows that the central spike represents the close relationship between the region 116–121 kb on chromosome 3, which includes the *clag 3.1* gene PFC0110w that encodes a cytoadherence-linked asexual protein [26], and the region 839–844 kb on chromosome 2 that contains *clag 2*, PFB0935w, PFB0940w, and PFB0945w fragments and the regions between them [24,27]. While the subsequence is relatively short (~ 5 kb) relative to the length of the sequences being searched (~ 1 Mb), the method detects the repeat quite readily, since its length and fidelity allow it to stand out above the background.

3.2. Different classes of repeat patterns

Regions with low information content point to areas that may contain patterns of interest. Starting at the 5' end of chromosome 2, a number of low information content regions seen in Fig. 2a were examined further. The two-dimensional plot shows the location of subsequences that are likely to have contributed to the compression of the sequence. 'Zooming in' on chromosome 2, the telomeric and subtelomeric region at 0–100 kb (overall compressibility ~ 1.4 bits per nucleotide) was examined further. Fig. 3 shows a two-dimensional plot of this region, with the sequence running from left to right, 5'–3', on the x -axis, and from top to bottom (5'–3') on the y -axis. The main diagonal shows the contribution of the base model (here, first order Markov model). Lines off the main diagonal show approximate repeats, parallel to the main diagonal when in the same orientation, perpendicular to the main diagonal when reverse complementary. The positions of repeats can be found by extending a horizontal line through the internal diagonal, dropping perpendiculars to the x -axis from the line's intersection with any internal diagonals and with the main diagonal, and reading the positions on the x -axis. Horizontal and vertical discontinuities within a diagonal indicate deletions in the repeat region. Discontinuities that do not alter the alignment of the diagonal represent regions within a repeat that are either unrelated or more weakly related to each other than the rest of the repeat. Numerous short diagonals close together indicate a number of short approximate repeats close together in the sequence.

Several regions of low information content at the 5' end of chromosome 2 are outlined in Fig. 3. The low information areas fall into two distinct classes. Dark spots and areas with many short diagonals, e.g. in the region 0–23 kb (upper left quadrant), indicate areas of low complexity, in this case correlating with known repetitive areas in the telomeric region of chromosome 2, or 'telomere associated repetitive elements (TAREs)' [28]. Another region of low complexity, at 90–95 kb, is located in the gene PFB0095c, which encodes the erythrocyte membrane protein PfEMP3, and can be ascribed to the 15 amino acid repeat in the gene product [26]. In between these two areas of low complexity, at 30–70 kb, is a relatively large subtelomeric region marked by numerous long repeats that appear as diagonal lines, which we examined in more detail.

3.3. Subtelomeric repeats

In order to examine further the repeats in both the 5' and 3' subtelomeric regions of chromosome 2, 40 kb from each end of the chromosome was extracted, and the subsequences were concatenated. A plot obtained

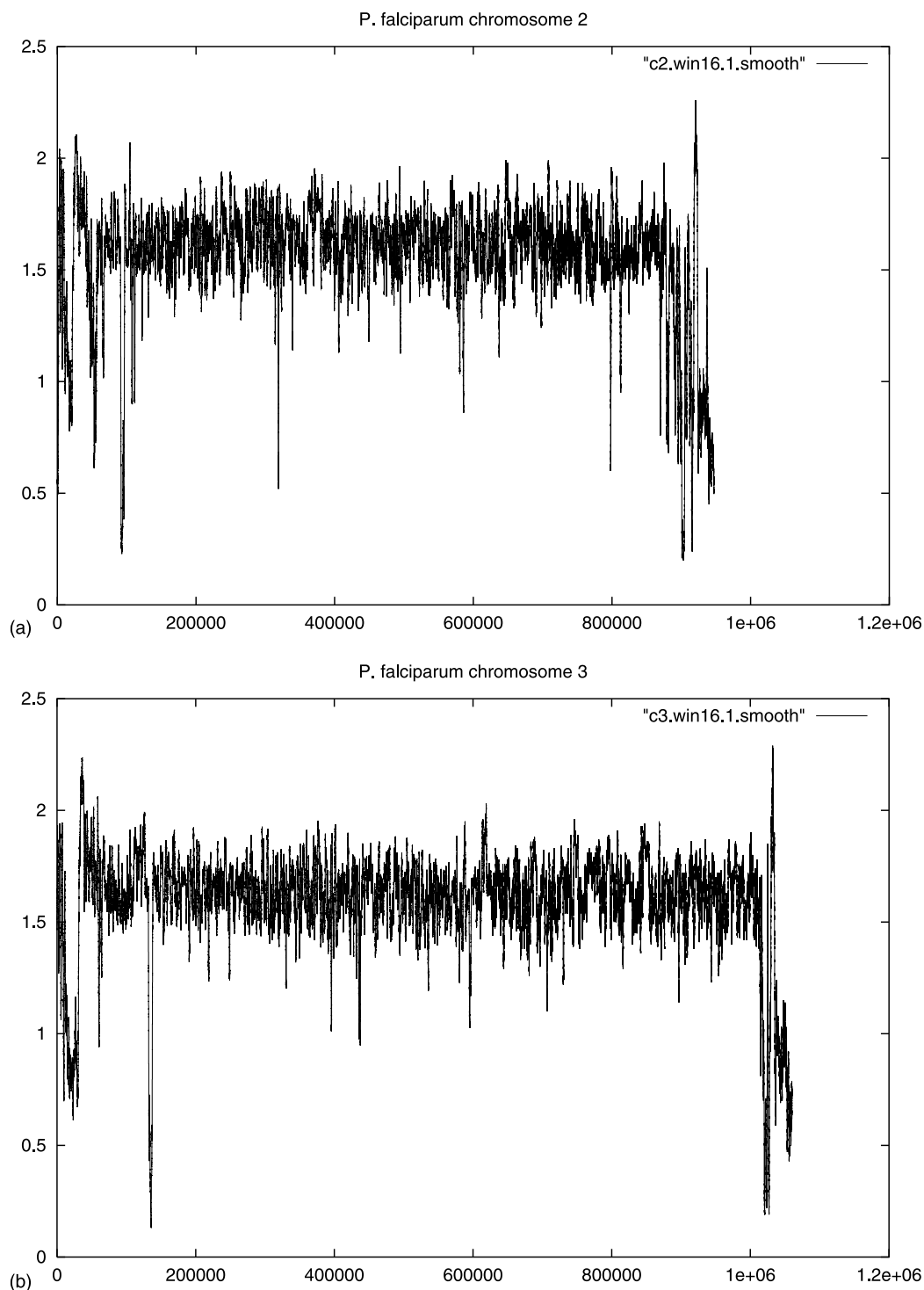


Fig. 2. Plots of information content (bits per nucleotide) across *P. falciparum* chromosomes: (a) chromosome 2; (b) chromosome 3; and (c) chromosome 3 in the context of chromosome 2. Nucleotide position is shown on the *x*-axis, average information content across a window of 1000 nucleotides centered on this nucleotide on the *y*-axis. The minimum hash-hit word length was 16 nucleotides.

from the concatenated sequences (Fig. 4) shows repeats in the 5' and in the 3' region in the context of the 5' region. A horizontal line separates the 5' (30–70 kb) and 3' (880–920 kb) regions. The abundance of internal diagonals in the figure indicate that the two regions contain numerous approximate repeats. The degree of

relatedness of various subsequences within this subtelomeric region is indicated by the intensity and continuity of the diagonals. The positions of repeat units, as determined by more detailed analysis, are shown along the main diagonal by half boxes. The lines in the lower left quadrant of the figure, representing subsequences in

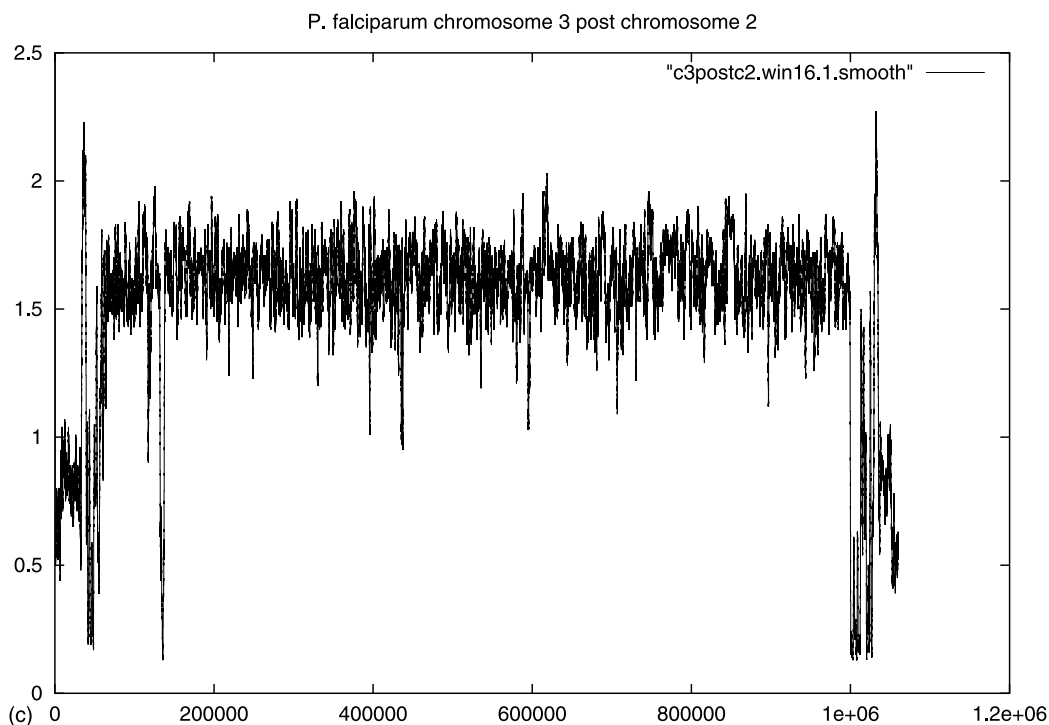


Fig. 2. (Continued)

the 5' region that are repeated in the 3' region, are perpendicular to the main diagonal, indicating that the repeats at the two ends of the chromosome are in reverse complementary orientation to each other.

Some of the repeats observed in Fig. 4 can be accounted for by known genes in the subtelomeric regions, such as repeated copies of the ~1.5 kb *rif* gene [25,24], and the longer *var* gene which encodes PfEMP1 [29,30]. The *var* genes PFB0010w and PFB1055c make only a small contribution to the observed repeats, because they are mostly outside of the region under observation. The relationship between the last exons of these two *var* genes is shown by the short diagonal line in the lower left quadrant of the Fig. 4, representing a repeat near the two ends of the concatenated sequence, while the rest of these genes lie outside the region.

While *rif* genes contribute to the repeats observed, the pattern of repeats cannot be totally explained by *rif* and *var* genes in the subtelomeric region. The longer diagonals in Fig. 4 represent sequences of 9–10 kb that are repeated (approximately) as a unit. This unit includes *rif* genes, but is considerably longer. The repeat unit, in fact, is spanned by two *rif* genes, and includes a considerable non-coding region between them, along with *var* fragments. There are three instances of this repeat unit on chromosome 2: (1) 5'–most repeat spanned by *rif* genes PFB0015c and PFB0025, including *var* fragment PFB0020c and non-coding regions between genes; (2) nearby repeat unit spanned by *rif* gene PFG0040c and *rif* pseudogene PFB0050c, includ-

ing *var* fragment PFB0045c and associated non-coding regions (internal diagonal, upper left quadrant); and (3) reverse complement repeat unit at the 3' end of the chromosome spanned by *rif* genes PFB1020w and PFB1035w, including the two *var* fragments PFB1025w and PFB1030w and non-coding sequences (internal diagonal, lower left quadrant). This multi-gene repeat unit corresponds to the canonical repeat unit we reported previously [31]. We can further infer from the

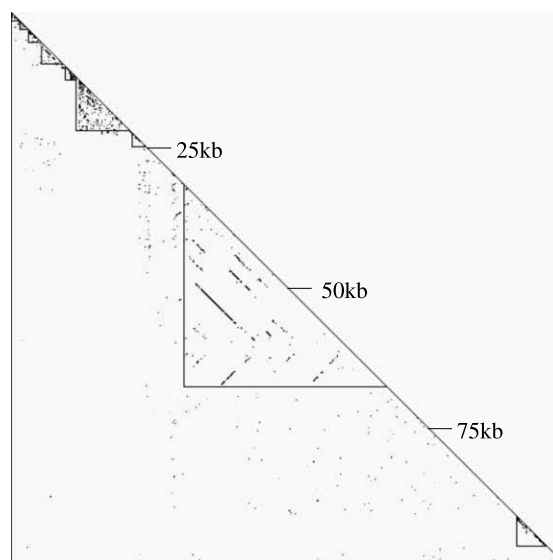


Fig. 3. Two-dimensional plot of 0–100 kb at the 5' end of chromosome 2. Regions of low complexity and repeats are outlined.

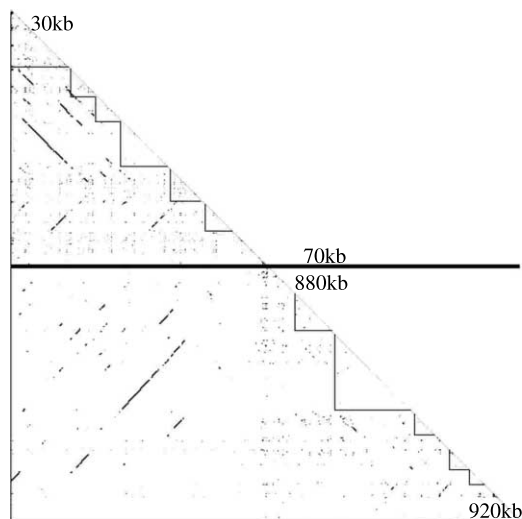


Fig. 4. Two-dimensional plot of subtelomeric regions of chromosome 2. The regions 30–70 and 880–920 kb are shown. Basic repeat units are outlined. The thicker horizontal line delineates between the 5' and the 3' region.

relative brightness of the diagonal lines that repeat (3) is more closely related to repeat (2) than it is to repeat (1).

The pattern seen strongly suggests a single duplication event copied a ~ 10 kb length of DNA, containing multiple genes and substantial amount of non-coding sequence, and that this region was duplicated twice on chromosome 2 (plus one original length of DNA). Shorter internal diagonals that are vertically aligned with the ends of the long diagonals represent the relationships among the *rif* genes within the 10 kb segment and other *rif* genes. The *rif* genes flanking the repeat unit are not as closely related to each other as they are to other *rif* genes, hence overlapping diagonals are not observed within the 10 kb unit.

The analogous subtelomeric region on chromosome 3, 30–70 kb and 1010–1040 kb, also shows repeats and a relationship between the two ends (Fig. 5). In chromosome 3, the long diagonal in the lower left of the figure indicates that a long region ~ 28 kb of the 5' region was copied into the 3' region in a single event, or vice versa, with subsequent mutation. This region includes multiple *rif* genes, which of course would have arisen from previous duplication events, a *var* gene, a *stevor* gene, and substantial non-coding material. Thus, while not originally designed for such purposes, our method has a potential usefulness in unravelling the sequence of events that lead to a region of repeats, over and above simply locating the repeats.

Remembering that intensity on the two-dimensional plot is a measure of the probability that one subsequence has contributed to another, it is also apparent in the two-dimensional plots that some repeat regions are more closely related to each other, shown by stronger

diagonals, than others (see for example Fig. 4 and Fig. 5). The ability to differentiate among repeats based on their length and overall fidelity is another advantage of our method over traditional dot matrix plots, where all subsequences that match above the threshold within a fixed-sized window appear the same; the unique ability of our model to take into account background composition is particularly useful in sequences of low complexity. A more quantitative measure of the degree of similarity between two sequences can be obtained from the normalised compressibility of one sequence in the context of another, relative to the compressibility in the absence of any context. Such analysis can show, for example whether the most closely related telomeric duplication event at the 5' end of a chromosome is to its 3' end or to another chromosome, and may provide a phylogeny of development of repeated genes such as *var* and *rif* sequences. It will be of interest to see if a genomic phylogeny derived using this method matches the relatedness of the coding genes themselves.

3.4. Repeat patterns in coding and non-coding regions

Localised dips in information content were noted at approximately 450 kb in chromosome 2 and approximately 600 kb in chromosome 3 (Fig. 2a and b), and correlate with the putative centromeres reported in [25]. While these centromeres are readily detected on a plot of G + C content, it is notable that they are observed here on the same scan that detects repeat units. When the sequencing of the *P. falciparum* genome is further advanced, it will be interesting to see whether our method suggests a lineage among centromeric regions on different chromosomes.

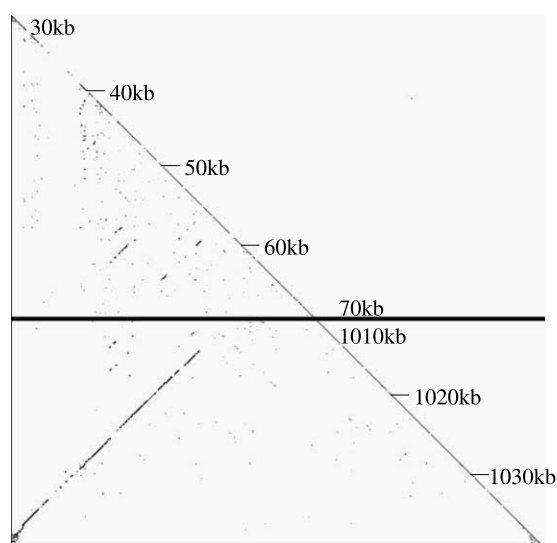


Fig. 5. Two-dimensional plot of subtelomeric regions of chromosome 3, showing the regions 30–70 and 1010–1040 kb. The thicker horizontal line delineates between the 5' and the 3' region.

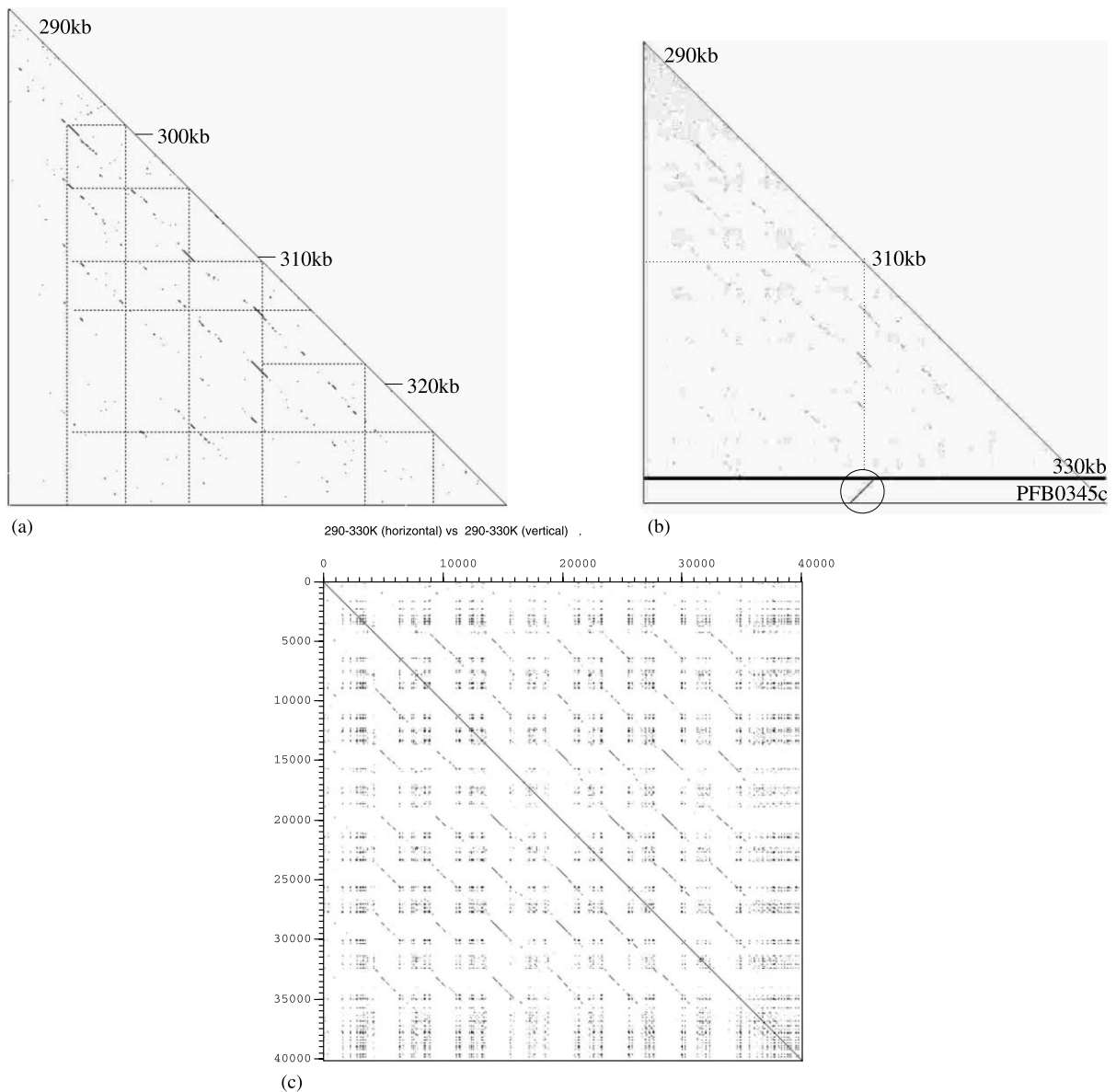


Fig. 6. Serine erythrocyte-binding antigen (SERA) cluster on chromosome 2. The region 290–330 kb, encompassing the SERA genes PFB0325c–PFB0370c is shown: (a) using our statistical model; (b) using our statistical model, with exon 4 of PFB0345c appended; and (c) using the Dotter dot-matrix plotter, with sliding window length 50 and grey ramp range 90–150. The grid in (a) shows similarities between different subsequences; the positions of related subsequences can be read from the *x*-axis. Weak and missing diagonals represent more weakly related sequences. In (b) the circled diagonal shows the last exon of PFB0345c.

A region around 300 kb in chromosome 2 in which the information content dropped to approximately 1.2 bits per nucleotide was also noted (Fig. 2a). This location corresponds to the cluster of serine erythrocyte-binding repeat antigen (SERA) genes (PFB0325c, PFB0330c, PFB0335c, PFB0340c, PFB0345c, PFB0350c, PFB0355c, PFB0360c) which lie in this region of chromosome 2 [24,32]. ‘Zooming in’ on the region, our method showed a pattern in which a ~ 1 kb segment of DNA is repeated, approximately, at regular intervals (Fig. 6a). Variations in the fidelity of the repeat are seen in a qualitative way, with the more

closely related sequences showing as stronger lines and the more weakly related sequences as broken lines. Occasionally the relationship between two repeats is so subtle that it is not detected directly at the degree of resolution achieved in the plot, but is inferred transitively, through a relationship in common with a third subsequence. The strongest relationships among sequences are shown in the grid pattern superimposed on the program output in Fig. 6a; some of the weaker relationships can be deduced from transitivity. Identification of the strong repeat as exon 4 of the SERA gene was confirmed by post-pending the exon 4 of PFB0345c to the region and rescanning (Fig. 6b).

We compared the utility of our method for finding repeats with the utility of two well-established techniques. Using BLASTN [33] to search for matches to last exon of PFB0345c, the regular repeat pattern in the SERA genes was not as readily apparent as it was with our method. The repeats were also evident using the advanced dot matrix plotter Dotter [34]. The Dotter plot of this region (Fig. 6c) also shows considerable background noise due to the biased nucleotide composition of the chromosome; in our method the biased nucleotide composition was correctly handled by the base model, and did not appear as a repeat. In our model the variations in the degree of relatedness among multiple subsequences can be inferred from the relative brightness of the diagonal lines; this takes into account the background characteristics of the sequence.

We also examined preliminary sequence from chromosome 10, using contigs obtained from The Institute for Genomic Research website (www.tigr.org). Contig c10m304 from this source proved to be very compressible using the approximate repeat model, giving an overall compression of 0.607 bits per nucleotide over its length of 24 kb, considerably better than the best compression achieved using Markov models of different orders without modelling the approximate repeats explicitly. The best compression without repeat modelling was achieved using a sixth-order Markov model, and was 0.820 bits per nucleotide. Compression of the biased sequence c10m304 (38.7% A) using a zero-order Markov model gave 1.846 bits per nucleotide.

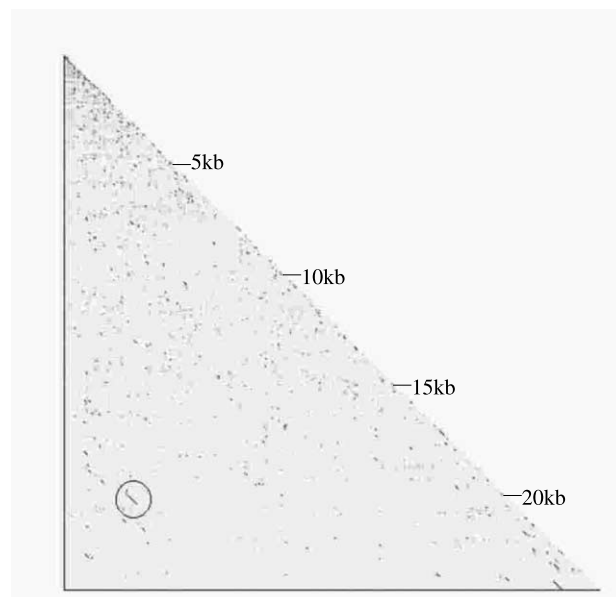
There are numerous short approximate repeats in c10m304 and many strings of A's. Using the approximate repeat model, a subsequence of approximately 600 nucleotides in length stood out as unusual over and above the numerous smaller repeats. This long repeat unit appears once at around 3000 nucleotides into the contig, and again at around 20000 nucleotides (Fig. 7a). This repeat was not as readily detectable using the Dotter dot matrix plotter, due to the noisy background of small-period repeats and the biased nucleotide composition (Fig. 7b).

3.5. Advantages of the statistical model

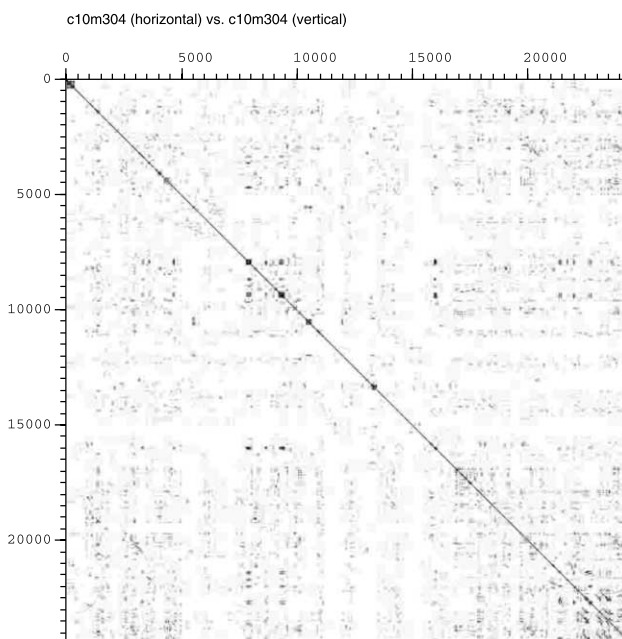
We have applied a statistical model [15,16] to the task of looking for approximate repeats in *P. falciparum* genomic sequences. The model is very general, using only the most minimal biological knowledge, i.e. that regions of DNA can be repeated, in a forward or inverted (reverse complement) direction, and that mutations, deletions, and insertions can occur. Because the model does not rely on prior information about the genome, it can find new motifs, and can find repeat sequences of different classes at the same

time. It can also use information about sequences, for example, by prepending or appending known sequences of interest to the sequence under exploration, and searching for repetitions of the known sequence.

Another strength of the current model is that it can detect weakly related sequence repetitions, and can differentiate repeats of different degrees of fidelity from each other, both visually and numerically. The



(a)



(b)

Fig. 7. Contig c10m304 from *P. falciparum* chromosome 10: (a) Two-dimensional plot using the statistical model. The circle encloses the 600 bp repeat at positions 3000 and 20000 from the 5' end of the contig; (b) Dot-matrix plot generated using the Dotter program, with sliding window length 37 and grey ramp range 150–190.

use of probabilities, rather than a threshold, is a key factor in differentiating among multiple repeats of different strengths, and can allow us to suggest likely gene duplication events. It would be interesting to extend this analysis to examine the degree of relatedness between entire genomes, to examine whether the conventional phylogenetic relationships deduced from examination of the sequences of particular genes holds true for the genome as a whole [35,36].

Other important properties of the method include its separate modelling of the base and repeat states, which allows significant motifs to emerge above a noisy background. Degrees of relationship between the sequences are revealed by examination at various threshold settings, and depending on the specific sequences that are examined. Regions of interest may include entire genes or only portions of genes. For example, when chromosome 2 is compared to itself, in addition to the regions around the telomere, a segment at positions 538000–539000 is highlighted. This corresponds to the exon 2 and surrounding introns of PFB0595c, a protein containing a putative DnaJ domain. Other proteins with DnaJ-like domains include PFB0085c, PFB0090c, PFB0920w, PFB0925w. Some of these also contain RESA-like sequences. The inclusion of the introns within this region suggests the possibility of a modular insertion of a functional region into multiple genes. Similarly, an examination of chromosome 2 against itself after comparison to chromosome 3 identifies the central region of PFB0695c (position 629000–630000), an ATP-dependent acyl-CoA synthetase as having a region of interest. This can be matched to a similar gene on chromosome 3, MAL3P8_AL034560 (position 66265–68586), which was annotated as a hypothetical gene originally, but would now appear to be an acyl CoA synthetase. Interestingly, this protein is quite similar to the octapeptide repeat antigen (ORA), which was originally described as an antigen commonly and strongly recognized by immune sera [37]. Other regions clearly highlighted include the clag genes on chromosome 3 which are related to a gene on chromosome 2 [38].

The model is an information theoretic abstraction of very general biological knowledge and is applicable to organisms other than *P. falciparum*. The generality of the model makes it useful as a pointer to new motifs, as well as a tool for locating known motifs and non-motif repeats.

The current implementation is a prototype lacking a convenient user interface; a new and generalised implementation of the model with an improved interface is under development and will be made available on the web. Requests for a Linux binary of the prototype should be made to the authors.

Acknowledgements

We would like to thank Tim Edgoose for implementing the compression algorithm, partially supported by Australian Research Council grant A9800558. We are grateful to Robert Huestis for turning our attention to interesting sequences in chromosome 10 and for many useful discussions.

Dot matrix plots were generated using the Dotter program of Sonnhammer and Durbin [34]. BLASTN searches were performed using the service provided by the National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health (USA).

We wish to thank the scientists and funding agencies comprising the International Malaria Genome Project for making sequence data from the genome of *P. falciparum* (3D7) public prior to publication of the completed sequence. A consortium composed of The Institute for Genome Research, along with the Naval Medical Research Center (USA), sequenced chromosomes 2, 10, 11 and 14, with support from NIAID/NIH, the Burroughs Wellcome Fund, and the Department of Defense.

Sequences for *P. falciparum* chromosomes 2 and 3 were obtained from the National Center for Biotechnology Information (www.ncbi.nlm.nih.gov), National Library of Medicine, National Institutes of Health (USA). Preliminary sequence data for *P. falciparum* chromosomes 10 was obtained from The Institute for Genomic Research website (www.tigr.org). Sequencing of chromosomes 10 and 11 was part of the International Malaria Genome Sequencing Project and was supported by award from the National Institute of Allergy and Infectious Diseases, National Institutes of Health (USA).

RLC is supported by the Howard Hughes Medical Institute International Scholars in Infectious Diseases and Parasitology Program, the Burroughs Wellcome Fund and the Australian National Health and Medical Research Council.

References

- [1] Wootton JC. Simple sequences of protein and DNA. In: Bishop MJ, Rawlings CJ, editors. DNA and Protein Sequence Analysis, a Practical Approach. Oxford; New York: IRL Press at Oxford University Press, 1997:169–83.
- [2] Fickett JW. Finding genes by computer: the state of the art. Trends Genet 1996;12:316–20.
- [3] Yermanian E. The physics of DNA and the annotation of the *Plasmodium falciparum* genome. Gene 2000;255:151–68.
- [4] Wootton JC, Federhen S. Statistics of local complexity in amino acid sequences and sequence databases. Comput Chem 1993;17:149–63.
- [5] Shannon CE. The Mathematical Theory of Communication. Champaign, Illinois: University of Illinois Press, 1949.

- [6] Langdon GG. An introduction to arithmetic coding. IBM J Res Dev 1984;28:135–49.
- [7] Wootton JC, Federhen S. Analysis of compositionally biased regions in sequence databases. In: Doolittle RF, editor. Methods in Enzymology 1996; 266: 554–71.
- [8] Wootton JC. Non-globular domains in protein sequences: automated segmentation using complexity measure. Comput Chem 1994;18:269–85.
- [9] Pizzi E, Frontali C. Low-complexity regions in *Plasmodium falciparum* proteins. Genome Res 2001;11:218–29.
- [10] Wan H, Wootton JC. A global compositional complexity measure for biological sequences: AT-rich and GC-rich genomes encode less complex proteins. Comput Chem 2000;24:71–94.
- [11] Agarawal P, States DJ. The repeat pattern toolkit (RPT): analyzing the structure and evolution of the *C. elegans* genome. In: Altman R, Brutlag D, Karp P, et al., editors. Proceedings of the Second Conference on Intelligent Systems in Molecular Biology. Menlo Park: AAAI Press, 1994:1–9.
- [12] Grumbach S, Tahi F. A new challenge for compression algorithms: genetic sequences. Inf Processing Manag 1994;30:875–86.
- [13] Loewenstern DM, Yianilos PN. Significantly lower entropy estimates for natural DNA sequences. In: Storer JA, Cohn M, editors. Proceedings of the IEEE Data Compression Conference, DCC97. Piscataway: IEEE Press, 1997:151–60.
- [14] Rivals E, Dauchet M. Fast discerning repeats in DNA sequences with a compression algorithm. In: Proceedings of the Genome Informatics Workshop. Tokyo: Universal Academy Press, 1997:215–26.
- [15] Allison L, Edgoose T, Dix T. Compression of strings with approximate repeats. In: Glasgow J, Littlejohn T, Major F, et al., editors. Proceedings of the Sixth International Conference on Intelligent Systems for Molecular Biology. Menlo Park: AAAI Press, 1998:8–16.
- [16] Allison L, Stern L, Edgoose T, Dix TI. Sequence complexity for biological sequence analysis. Comput Chem 2000;24:43–55.
- [17] Ziv J, Lempel A. A universal algorithm for sequential data compression. IEEE Trans Inf Theory 1977;IT-23:337–43.
- [18] Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. J Roy Stat Soc B 1977;39:1–38.
- [19] Baum LE, Eagon JE. An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model of ecology. Bull AMS 1967;73:360–3.
- [20] Baum LE, Petrie T, Soules G, Weiss N. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. Ann Math Stat 1970;41:164–71.
- [21] Yee CN, Allison L. Reconstruction of strings past. Comp Appl Biosci 1993;9:1–7.
- [22] Allison L, Wallace CS, Yee CN. Finite-state models in the alignment of macro-molecules. J Mol Evol 1992;35:77–89.
- [23] Kurtz S, Ohlebusch E, Schleiermacher C, et al. Computation and visualization of degenerate repeats in complete genomes. In: Bourne P, Gribskov K, Altman R, et al., editors. Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology. Menlo Park: AAAI Press, 2000:228–38.
- [24] Gardner MJ, Tattelin H, Carucci DJ, et al. Chromosome 2 Sequence of the Human Malaria Parasite *Plasmodium falciparum*. Science 1998;282:1126–32.
- [25] Bowman SD, Lawson D, Basham D, et al. The complete nucleotide sequence of chromosome 3 of *Plasmodium falciparum*. Nature 1999;400:532–8.
- [26] Pasloske BL, Baruch CI, van Schravendijk MR, et al. Cloning and characterization of a *Plasmodium falciparum* gene encoding a novel high-molecular weight host membrane-associated protein PfEMP3. Mol Biochem Parasitol 1993;59:59–72.
- [27] Huestis R, Cloonan N, Tchavtchitch M, Saul A. An algorithm to predict 3' intron splice sites in *Plasmodium falciparum* genomic sequences. Mol Biochem Parasitol 2001;112:71–7.
- [28] Figueiredo LM, Pirritt LA, Scherf A. Genomic organisation and chromatin structure of *Plasmodium falciparum* chromosome ends. Mol Biochem Parasitol 2000;106:169–74.
- [29] Smith JD, Chitnis CE, Craig AG, et al. Switches in expression of *Plasmodium falciparum* var genes correlate with changes in antigenic and cytoadherent phenotypes of infested erythrocytes. Cell 1995;82:101–10.
- [30] Baruch DI, Pasloske BL, Singh HB, et al. Cloning the *P. falciparum* gene encoding PfEMP1, a malarial variant antigen and adherence receptor on the surface of parasitized human erythrocytes. Cell 1995;82:77–87.
- [31] Stern L, Allison L, Coppel RL, Dix TI. Information Theoretic Analysis of *Plasmodium falciparum* Genomic DNA. The University of Melbourne Technical Report 1998; 1998/7.
- [32] Bzik DJ, Li WB, Horii T, Inselburg J. Amino acid sequence of the serine-repeat antigen (SERA) of *Plasmodium falciparum* determined from cloned cDNA. Mol Biochem Parasitol 1988;30:279–88.
- [33] Altschul SF, Madden TL, Schaffer AA, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–402.
- [34] Sonnhammer ELL, Durbin R. A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. Gene 1995;167:GC1–GC10.
- [35] Escalante AA, Ayala FJ. Evolutionary origin of *Plasmodium* and other apicomplexa based on rRNA genes. Proc Natl Acad Sci USA 1995;92:5793–7.
- [36] Rich SM, Light MC, Hudson RR, Ayala FJ. Malaria's eve-evidence of a recent population bottleneck throughout the world populations of *Plasmodium falciparum*. Proc Natl Acad Sci USA 1998;95:4425–30.
- [37] Favaloro JM, Marshall VM, Crewther PE, et al. cDNA sequence predicting an octapeptide-repeat antigen of *Plasmodium falciparum*. Mol Biochem Parasitol 1989;32:297–9.
- [38] Holt DC, Gardiner DL, Thomas EA, et al. The cytoadherence linked asexual gene family of *Plasmodium falciparum*: are there roles other than cytoadherence? Int J Parasitol 1999;29:939–44.