Exploring Long DNA Sequences by Information Content

Trevor I. Dix[1,2], David R. Powell[1,2], Lloyd Allison[1], Samira Jaeger[1], Julie Bernal[1], and Linda Stern[3]

[1] Faculty of I.T., Monash University,
[2] Victorian Bioinformatics Consortium,
[3] Computer Science and Software Engineering, Melbourne University

1 Introduction

The paper explores long DNA sequences, of the order of millions of bases, by means of their information content. Compression is used to find the features of a sequence and common features that relate one sequence to another.

The compression problem is to calculate the information content per base, producing an "information sequence". Information is relative, i.e. it depends on the context. The context can include one or more other sequences; this lets you "relate" two or more sequences. Note that an information sequence is 1-dimensional, operations such as difference, zoom, smooth and threshold are efficient, taking linear time and space. This is in contrast to the traditional 2-dimensional dot plots that have to be stored at low resolution for long sequences.

Any compression model can be used to create an information sequence. Here we use our approximate repeats model (ARM) [1, 2, 8]. We present the ARM, introduce our tool to manipulate information sequences, and explore its use for the red alga *Cyanidioschyzon merolae* and the malaria strain *Plasmodium falciparum*.

2 Methods

2.1 DNA Sequence Compression

We wish to examine the information content of sequences. Information content and compressibility are inherently related: low information content implies highly compressible and high information content implies poorly compressible. So, if one has an efficient encoding of a sequence, then it can be argued that one has a good model of that sequence. From Shannon [6] we know that an efficient encoding is related to its probability by the log likelihood. That is, information $I(m) = -logP(m)$, where $P(m)$ is the probability of m occurring.

When trying to make an inference from some data using a Bayesian technique, we attempt to maximize the posterior probability, $P(H|D) = P(D|H) \times P(H)/P(D)$ for hypothesis H and data D. If our model (or hypothesis) has a nuisance parameter about which we do not care to make an inference, we should

sum over all possible values for this parameter. This is necessary when using sequence alignment to infer how related two sequences are. If we are only interested in whether the sequences are related or not we should sum over all possible alignments [5].

2.2 Approximate Repeats Model

The Approximate Repeats Model (ARM) [1] is designed to compress DNA sequences well. DNA sequences often have regions that are highly similar, with only a few changes. Given the double-stranded nature of DNA, it is common for DNA to contain reverse-complement repeats. The ARM compresses a sequence by finding each region that is similar to a previously encountered region and encoding each such region as "similar to this other region, but with these changes". It also looks at the reverse-complement of the sequence so far to find similarities. (For an implementation of the model see `http://www.csse.monash.edu.au/~powell/fuzzyLZ/`.)

The ARM considers a DNA sequence a base-pair (bp) at a time from the left. The model may encode a bp using two possibilities. (1), it may be encoded using some other *base model*. This base model can be any sequence model. We have typically used a small order Markov Model. (2), the bp may be encoded as part of a repeated region. A repeated region is encoded by first encoding the position in the sequence where this region is repeated from. A uniform distribution is used to encode this position. The description to this point is quite similar to the Ziv-Lempel [9] algorithm. The difference is in how a repeated region is treated, each bp from a repeated region may be copied, deleted, changed or a bp inserted. The length of a repeat is encoded using a geometric distribution; while this may not ideal, it allows for a more efficient algorithm.

Notice that this method of treating repeated regions is very similar to the way sequence local-alignment algorithms [7] are used to model sequence variations. This is quite deliberate, the ARM is in effect aligning a sequence against itself, and achieving good compression in regions that would have a good alignment score. The implementation of the ARM supports either simple gap costs or affine gap costs. It is possible to view a two-dimensional plot of the self-alignment used in the ARM, such an image is a very coarse way to look at the compression results. For example, for a sequence of roughly a million bases, each pixel in the image would represent one thousand bases. Thus it is necessary to find a better way to deal with the compression results, we suggest using a 1-dimensional plot of the compression.

Often there are many competing sequence alignments that are almost equally good. This also happens with the ARM, a region may be quite similar to a number of earlier regions and we do not want to pick just one of them to copy from. These repeated regions may be treated as mutually-exclusive hypotheses, and since we do not care to make an inference about which is the best, we may sum over all of them, illustrating a nuisance parameter.

The ARM has a number of parameters, probabilities for the beginning of a repeat, for the possible mutations and for ending a repeat. The ARM is used

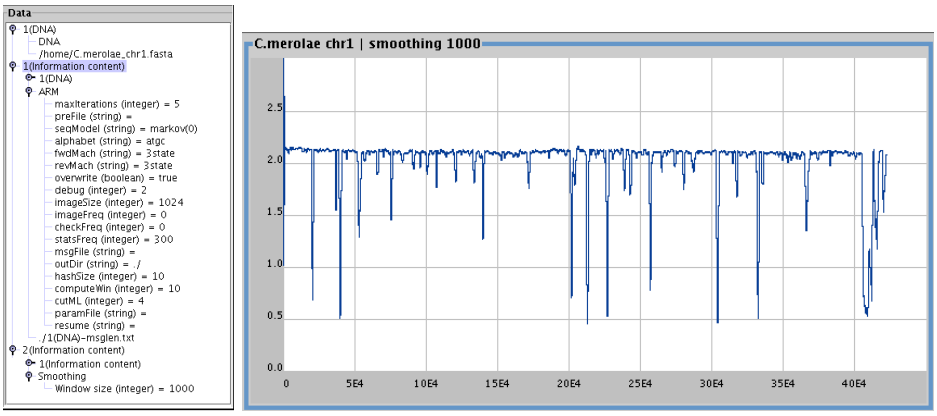

Fig. 1. 1-d plot for *C. merolae* chromosome 1, smoothing window 1000

with some initial values for these parameters, then the results from applying the model are used to choose new values for the parameters, and the ARM is applied again. An EM algorithm iterates until the parameters converge.

2.3 1-D Information Content Viewer

InfoV is a Java platform used to explore the structure of sequences using arbitrary compression models. It provides functionality to import biological sequences such as DNA, use compression models to generate information content sequences and interactively display multiple plots for the analysis. This tool also provides various functions to manipulate sequences such as smooth, cut, append, calculate difference between numeric sequences and find the reverse complement of DNA sequences. Additionally, InfoV annotates how sequences are derived, this includes the storage of the model parameters and functions used to create sequences.

The current implementation of InfoV is focused on DNA sequences and includes the ARM. However, it has a generic, extensible design, which enables the analysis of other type of sequences, such as character and numeric sequences from other compression models.

3 Results

We applied the ARM to find approximate repeats within each of chromosomes 1, 2, 3, 4, 5, 6, 11, 12, 19 and 18 of *C. merolae* and between pairs of chromosomes. The 1-d information content graph, $I(c1)$, is given in figure 1 for chromosome 1. It has been smoothed, displaying the average of a 1000 wide sliding window. We can easily store the whole graph and dynamically explore the low information areas. The window size should be of the order of what we are looking for. The viewer

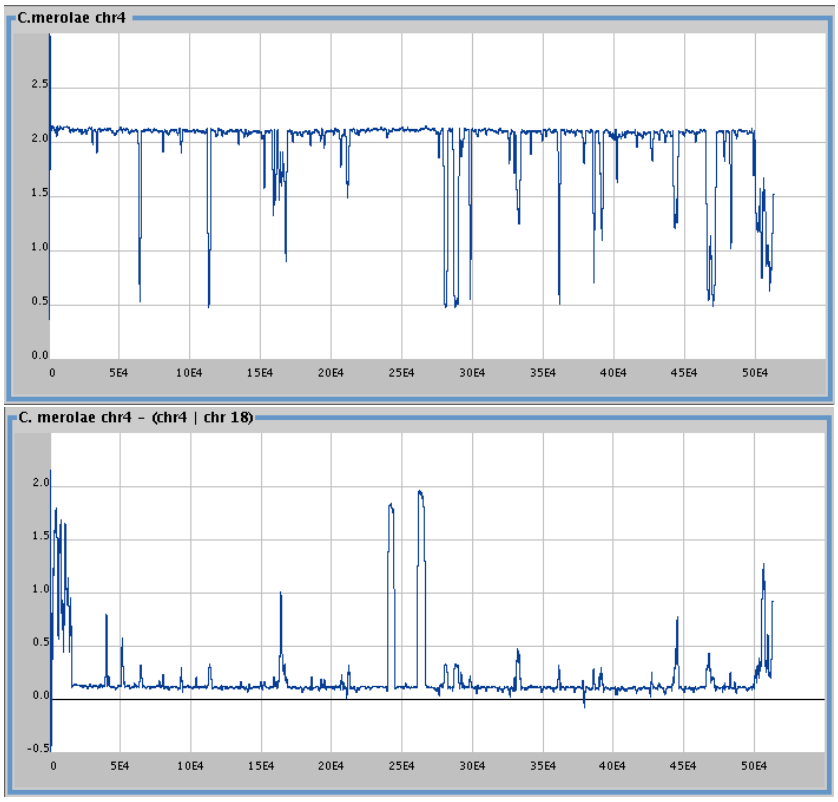

Fig. 2. 1-d plot for *C. merolae* chromosome 4 given 18, smoothing window 100

facilitates zooming-in and re-smoothing, typically with smaller window size, to find interesting regions. These can be copied to file and for further investigation starting say with a Blast search. We also give the history window for the plot.

Figure 2 shows *C. merolae* chromosome 4 compressed independently. The figure also contains a difference plot of information content for chromosome 4 alone minus that for chomosome 4 given 18, i.e. $I(c4) - I(c4|c18)$. To calculate the information sequence $I(c4|c18)$, the ARM prepends chromosome 18 to 4, and thus compresses chromosome 4 in the presence chromosome 18. This shows explicitly what information content chromosome 18 brings to the model for chromosome 4.

In this case, we find repeated regions from 239406 to 244000 corresponding to 974903 to 970308 in chromosome 18, and another from 260529 to 265988 corresponding to 961910 to 967371 in 18. The first region is a probable myo-inositol 2-dehydrogenase and the second contains a hypothetical protein.

Importantly, all of these plots are 1-dimensional. They can be computed at full resolution and stored, even on a small computer. We used the ARM but this can be done for any (your favourite) compression model. Common operations

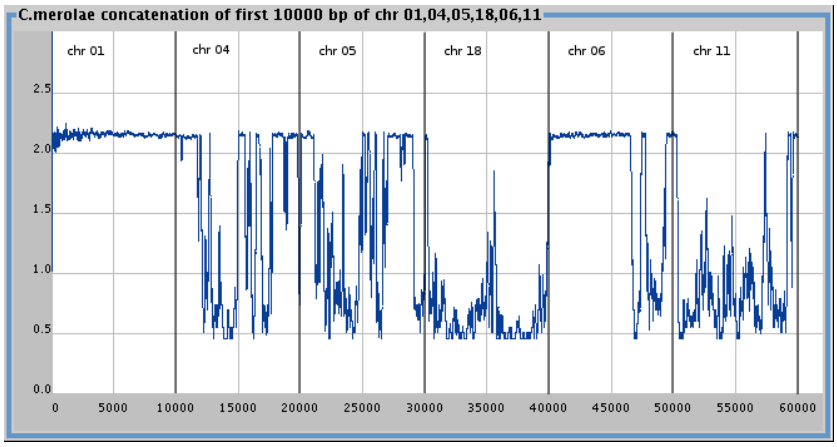

Fig. 3. 1-d plot for *C. merolae* 10000 bp for chromosomes 1, 4, 5, 18, 6 and 11, smoothing window 100

like difference, smooth, zoom and threshold can be performed quickly in linear time. A difference plot shows what *new* information an addition to a context tells us about a sequence; features already revealed by the original context are discounted by the difference.

We also investigated the subtelemeric regions of *C. merolae*. Pairwise comparisons $I(c_i|c_j)$ confirmed known results [4]. We summarize the results in figure 3 showing that the subtelemeric regions for chromosomes 1, 4, 5 and 18 belong to element P and those for chromosomes 6 and 11 belong to element PH.

Our final example is for chromosome 2 of *Plasmodium falciparum*. Figure 4 shows a difference plot of $I(c2) - rev(I(revcomp(c2)))$ where $revcomp(c2)$ finds the reverse complement of DNA sequence $c2$ and rev simply reverses the resulting information content sequence. The sequence from the first term is computed left to right; the second is computed right to left and then reversed. Such difference plots highlight first and last of all approximate repeated sequences.

Most of the difference plot gives values close to zero. But at both ends there are large differences from the baseline reflecting the known repetitive structure of chromosome ends for *Plasmodium falciparum*. The differences in sign are a result of asymmetry and subtraction. Telomere-associated repeat elements include Rep20, and the var, rif and stevor genes that are involved in its virulence [3].

4 Conclusion

We have shown how to explore 1-dimensional information sequences derived from long DNA sequences. Recall that information is relative to what is known. A sequence Y can be compressed firstly in a context $ctx1$ and then in a context $ctx2$ where $ctx2$ is $ctx1$ plus a sequence X. The difference between the information sequences for $Y|ctx1$ and for $Y|ctx2$ shows the *new* information that X gives

us about Y. Mere common "statistical" features of Y and X that were already known from $ctx1$ and/or Y itself are discounted. Exploration of full-resolution information sequences is carried out in linear time and space.

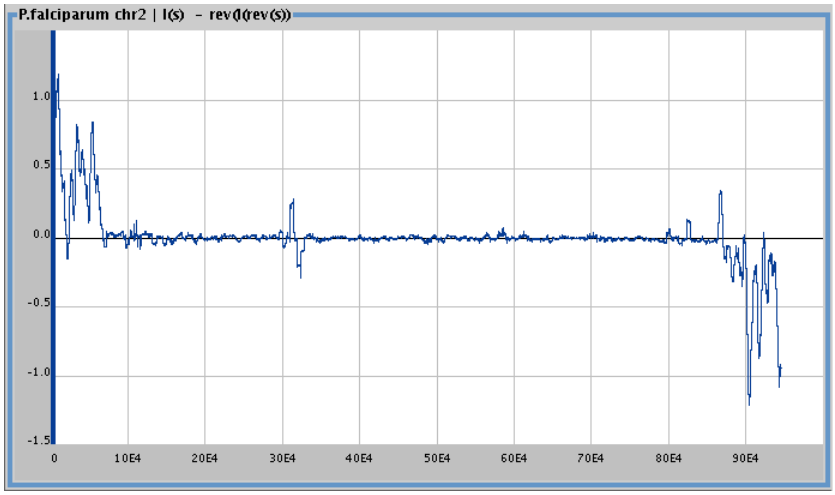

Fig. 4. 1-d plot of $I(c2) - rev(I(revcomp(c2)))$ for chromosome 2 of *P. falciparum*, smoothing window 5000

References

1. L. Allison, T. Edgoose, and T. I. Dix. Compression of strings with approximate repeats. *Intell. Sys. in Mol. Biol. '98*, pages 8–16, 1998.
2. L. Allison, L. Stern, T. Edgoose, and T. I. Dix. Sequence complexity for biological sequence analysis. *Computers and Chemistry*, 24(1):43–55, 2000.
3. B. Crabb and A. Cowman. *Plasmodium falciparum* virulence determinants unveiled. *Genome Biology*, 3(11):103.1–1031.4, 2002.
4. M. Motomichi and *et al.* Genome sequence of the ultrasmall unicellular red alga *Cyanidioschyzon merolae* 10D. *Nature*, 428:653–657, 2004.
5. D. R. Powell, L. Allison, and T. I. Dix. Modelling alignment for non-random sequences. *LNCS/LNAI*, 3339:203–214, 2004.
6. C. E. Shannon. A mathematical theory of communication. *Bell Systems Technical Journal*, 27:379–423,623–656, 1948.
7. T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147:195–197, 1981.
8. L. Stern, L. Allison, R. L. Coppel, and T. I. Dix. Discovering patterns in plasmodium falciparum genomic dna. *Molecular and Biochemical Parasitology*, 118(2):175–186, 2001.
9. J. Ziv and A. Lempel. A universal algorithm for sequential data compression. *IEEE Trans. Inf. Theory*, IT-23:337–343, 1977.