Compressing the human genome using exclusively Markov models

Diogo Pratas and Armando J. Pinho

Abstract Models that rely exclusively on the Markov property, usually known as finite-context models, can model DNA sequences without considering mechanisms that take direct advantage of exact and approximate repeats. These models provide probability estimates that depend on the recent past of the sequence and have been used for data compression. In this paper, we investigate some properties of the finite-context models and we use these properties in order to improve the compression. The results are presented using the human genome as example.

Key words: Markov models, DNA signature, DNA entropy, Data compression

1 Introduction

The study of genetics, of which genetic diseases are an important particular case, has been growing during the last decades. Making the genome data easier to transfer over the Internet, as well as reducing its storage size, is a key step to facilitate these studies. Also, the study of data compression algorithms, besides the immediate aim of obtaining data reduction, provides a means for discovering the structure of the data. In fact, in order to compress data, the compression methods have underlying models that represent the data more efficiently. Hence, the better the compression, the better these models describe the information source associated to the data.

Essentially, DNA sequences have been modeled using a combination of two paradigms, one relying on the Lempel-Ziv substitutional scheme, the other one based on the Markov property. This approach is justified by the

Signal Processing Lab, DETI / IEETA, University of Aveiro, 3810–193 Aveiro, Portugal
e-mail: {pratas,ap}@ua.pt

non-stationary nature of the DNA sequence data, which is characterized by an alternation between regions of relatively high and low entropy. Usually, the low entropy regions are modeled by the substitutional methods, whereas those of higher entropy are better described by low-order Markov models.

In this paper, we address the problem of representating the human genome exclusively by a combination of Markov models. To investigate this matter, we used a method based on multiple competing finite-context models [5]. We studied the implications of representing the data with different finite-context models and we discovered some characteristics that allowed us to introduce some techniques to improve the compression. Also, we compared the ability of the approach based on multiple competing finite-context models with that provided by the current state-of-the-art DNA coding method, XM [3], showing comparable results, but at the cost of much less computation time. The XM method also uses finite-context modelling. The algorithm comprises three types of experts: (1) order-2 Markov models; (2) order-1 context Markov models (typically using information from the 512 previous symbols); (3) the copy expert, that considers the next symbol as part of a copied region from a particular offset.

2 Materials and methods

2.1 DNA sequences

In this study, we used the complete DNA sequence of the human genome. The genome was obtained from the following source: *Home sapiens*, Build 33, from the National Center for Biotechnology Information (NCBI) (`ftp://ftp.ncbi.nlm.nih.gov/genomes/H_sapiens/April_14_2003`);

2.2 Single finite-context models

We consider five pseudo-random i.i.d. sequences uniformly distributed over the alphabet {A, C, G, T}, with sizes: Sequence 1, 10^6 symbols; Sequence 2, 10^7 symbols; Sequence 3, 10^8 symbols; Sequence 4, 10^9 symbols; Sequence 5, 10^{10} symbols.

These sequences have been compressed using a DNA compressor based on finite-context models, described in [5], with sixteen different orders (context depths), using single models (no competitive models). The final entropy values have been plotted in Fig 1.

Observing Fig 1, it is possible to identify a property of the finite-context models, characterized by a peak on the average number of bits per base (bpb)

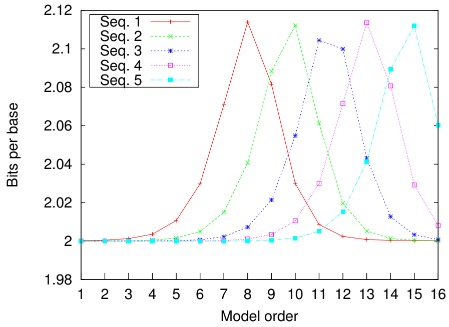

Fig. 1 The entropy of random sequences with different sizes, obtained using finite-context models of several orders (depths).

curve for an order that depends on the size of the sequence. For comparing the behaviour of these entropy curves that we have obtained for random sequences with those generated with real DNA sequences, we have ran the same procedure for all human chromosomes.

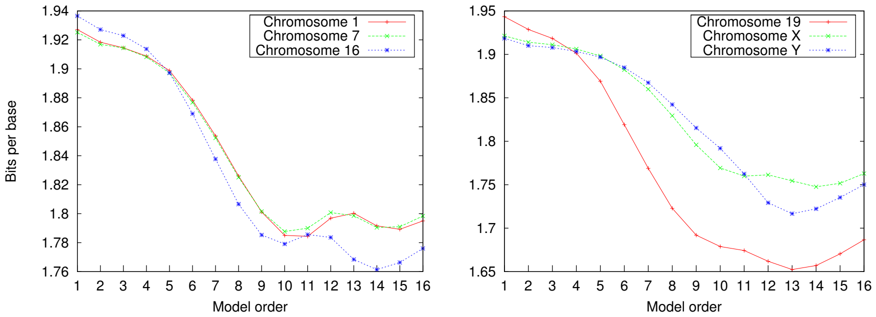

Fig. 2 Entropy curve for chromosomes 1, 7, 16, 19, X and Y using sixteen single models.

We have observed that almost all chromosomes have an identical pattern, although significantly different from that obtained with random sequences. This pattern can be observed for three examples of chromosomes in Fig. 2. In this case, the curves show a peak, although much less evident than it is for the random sequences. Based on this observation, we conclude that there are parts in these chromosomes that seem to be random. Furthermore, we also observe that the horizontal position of the peak is correlated with the sizes of the chromosomes (sizes of the samples): ≈ 219 million bases for chromosome 1, ≈ 155 million bases for chromosome 7 and ≈ 80 million bases for chromosome 16.

However, chromosomes 19 (≈ 56 million bases), X (≈ 148 million bases) and Y (≈ 23 million bases) do not show an entropy pattern similar to the others, specially chromosome Y. Chromosomes 19 and X are generally better compressed than the others (apart from chromosome Y), revealing that there are more repetitive zones and less random parts in these chromosomes. Chromosome 19 is the one containing the largest number of small repeats [1], thus justifying the shape of the peak in the corresponding entropy curve.

Chromosome Y showed the most different behaviour in this process, lacking the peak in the entropy curve. As reported before [4], the Y chromosome is highly repetitive, a property that agrees with the observations and that strengthens the conclusion that the main reason for peak absence is the existence of extensive repetitive zones in this chromosome.

2.3 Competitive finite-context models

It is known that DNA data is better represented by multiple finite-context models [5], because the data are non-stationary. Having observed the presence of a peak in the entropy curve when using single models, we now address the case of using multiple competing models to investigate if this property still holds. Therefore, we compressed the random sequences using the sixteen models, one by one, but now competing with a fixed order-1 model, using a block size of one base. In this evaluation, we did not include the additional information needed to describe which of the two models is used for each base and, therefore, the presented values do not correspond to real compression values (which, of course, cannot be lower than two bits per base for random sequences). Here, we just wanted to assess the peak property with the competitive models.

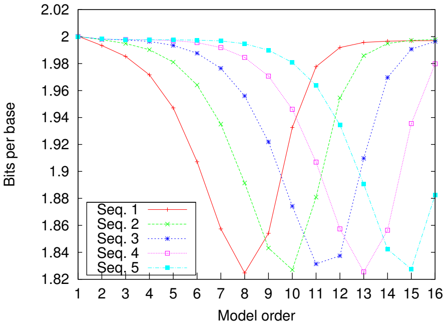

Fig. 3 Entropy curve for random sequences, using two competing models: one fixed with order-1, the other varying from order one to sixteen.

Apparently, in Fig 3 the peak has been inverted, if we compare with the previous results (Fig 1). Moreover, we tested the same process with more competitive models and the peak remained inverted. Using this property, it is possible to know where the best theoretical compression model order is. However, random sequences and DNA sequences are different, so we used the same method to test if a similar behaviour would also appear in the DNA sequences. Fig 4 is an example of that test (using a block size of ten bases). It shows an inverted peak in all chromosomes, as occurred with the random sequences, apparently revealing the best compression model orders for the corresponding block size and models usage.

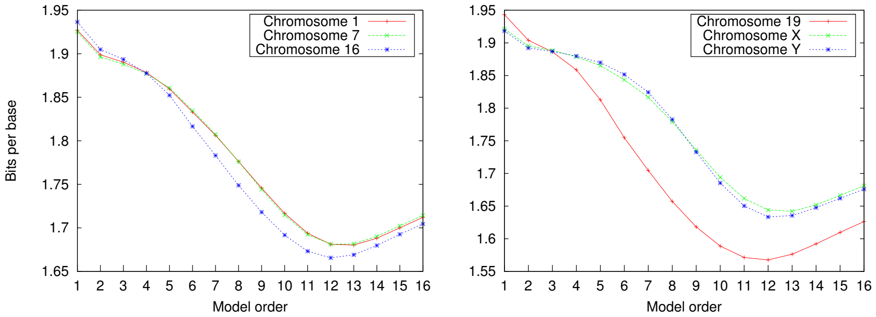

Fig. 4 Entropy curve for chromosome 1, 7, 16, 19, X and Y using one fixed order-1 model competing with sixteen single models.

2.4 Sequence concatenation

Finite-context modelling provides probability estimates that depend on the recent past of the sequences. Generally, bigger sequences provide better statistics, consequently providing more accurate models. Normally, the human genome (like other eukaryote organism) is compressed chromosome by chromosome, which prevents the model from exploring inter-chromosome correlations[2]. In order to explore the advantage of using these models in more than one chromosome at the same time, we compressed chromosome 1 concatenated with chromosome 2 and compared the compression ratio with the average resulting from compressing the chromosomes individually. Table 1 shows the results.

In this case, the average rate without concatenation would stand for 1.7221 bpb. Although, with concatenation, the ratio value is 1.7147 bpb, indicating that there is an advantage of using concatenation in finite-context models. Moreover, we used this method with more sequences and also including more competing models, consistently obtaining better compression results.

Table 1 Compressing results regarding chromosome 1, chromosome 2 and a concatenation of chromosomes 1 and 2. It has been used two competitive models (order-3 and order-16) and a block size of 50 bases.

Sequence	Rate (bpb)	Time (min.)	Length (Mb)
Chromosome 1	1.7121	9.33	218.71
Chromosome 2	1.7314	14.10	237.04
Concatenated	1.7147	20.92	455.75

3 Results and discussion

In the previous section, we presented a property of finite-context models, based on the observation of a peak in the entropy curve obtained as a function of the model order, apparently revealing the most repetitive chromosomes. Consequently, we discovered that the peak observed in the single finite-context models indicated one of the best models to compress the sequences using the competitive finite-context models. On the other hand, we realized that using concatenated sequences we could archive better compression results. Therefore, we concatenated all chromosomes from the human genome and addressed it as a single sequence.

The version of the human genome that we are using has about 2.6×10^9 bases (excluding the unknown symbols). Since the human genome is bigger than Sequence 4 (10^9 symbols) and smaller than Sequence 5 (10^{10} symbols), then its entropy curve should show a peak between order-13 and order-15 (see Fig 3) and, more probably, between order-13 and order-14. For the reason explained below, of uniform model order distribution, we chose order-13.

We recall that, due to the non-stationary characteristic of the DNA data, the multiple finite-context models should combine, at least, a low order model and a high order one. Accordingly, we used an order-4 model, because to compress the human genome with competing models, model order-4 seems to have the best compression ratio in the low order category. For the high order category, an order-16 model seems to be the best. Together with the order-13 indicated by the presence of the peak for this sequence size, we get three models with depths 4, 13 and 16. However, since there is a difference of 3 between model order-13 and order-16, we decided to include an additional model for order 7 and, therefore, to have an uniformly model order distribution: 4, 7, 13, 16.

The size of the data block is also a parameter that needs to be chosen. To assess how this paramenter may affect the performance of the compression algorithm, we performed an exhaustive search in the interval from size 80 to size 120, using the four models mentioned above. Fig 5 shows the compression results obtained, revealing that a block size of 105 bases lead to the best compression. Nevertheless, as can be seen in the graphic, for this range the exact block size does not affect the compression ratio in a significant way.

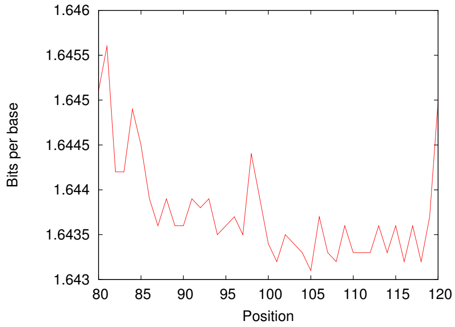

Fig. 5 Entropy curve for the compression of the human genome using four models, as a function of the block size.

Table 2 Compressing results of the human genome with different approaches. The FCM-S, FCM-C and FCM-CA columns contain, respectively, the results provided by the single finite-context models, by the eight competitive finite-context models on the individual chromosomes, and by the four competitive finite-context models on the complete genome sequence. The XM-50 and XM-200 columns show the results obtained with the XM algorithm, using 50 and 200 experts.

Method	FCM-S	FCM-C	FCM-CA	XM-50	XM-200
Rate (bpb)	1.739	1.695	1.643	1.644	1.618
Time (min)	46	323	197	1035	1780

The experimental results included in Table 2 show that previous compression results of the human genome with competitive finite-context models (FCM-C), using eight order models (2, 4, 6, 8, 10, 12, 14 and 16), indicated a ratio of 1.695 bpb. In this work (FCM-CA), we were able to compress the human genome slightly better than the state-of-the-art XM encoder [3] (with 50 experts). Moreover, FCM-CA was 5 times faster than XM-50. Regarding XM-200, also associated to the XM technique but using 200 experts, it has better compression ratio (0.025 bpb) than FCM-CA, but FCM-CA is approximately 9 times faster.

4 Conclusion

We have pointed out a property of finite-context models, characterized by a peak in the entropy curve obtained using different model orders. The amplitude of this peak seems to be related with the amount of repetitiveness of the sequence (the higher the randomness, the more pronounced the peak), whereas the position of the peak depends on the size of the sequence. Using competitive finite-context modelling, the peak is inverted, indicating a model order for which compression is efficient.

We concluded that using finite-context modelling in the concatenated human genome gives better compression results than when using the chromosomes one by one. This means that inter-chromosome information can be used by these models. Using only Markov models, we were able to compress the human genome with values that are competitive with the XM technique and that require much less computation time.

Taking into account the results that we report in this paper, we can say, perhaps somewhat surprisingly, that complete genomes can be quite well described using only discrete Markov models, i.e., by models that rely on short-term knowledge of the past.

5 Acknowledgements

This work was supported in part by the grant with the COMPETE reference FCOMP-01-0124-FEDER-007252 (FCT, Fundação para a Ciência e Tecnologia, reference PTDC/EIA-EIA/103099/2008).

References

1. Berg, I., Bosnacki, D., Hilbers, P.: Large scale analysis of small repeats via mining of the human genome. In: 20th Int. Workshop on Database and Expert Systems Application, DEXA'09, pp. 198–202 (2009). DOI 10.1109/DEXA.2009.78
2. Botta, M., Haider, S., Leung, I., Lio, P., Mozziconacci, J.: Intra- and inter-chromosomal interactions correlate with CTCF binding genome wide. Molecular Systems Biology **6** (2010). DOI 10.1038/msb.2010.79
3. Cao, M.D., Dix, T.I., Allison, L., Mears, C.: A simple statistical algorithm for biological sequence compression. In: Proc. of the Data Compression Conf., DCC-2007, pp. 43–52. Snowbird, Utah (2007)
4. Haubold, B., Wiehe, T.: How repetitive are genomes? BMC Bioinformatics **7**(1), 541 (2006). DOI 10.1186/1471-2105-7-541
5. Pinho, A.J., Neves, A.J.R., Martins, D.A., Bastos, C.A.C., Ferreira, P.J.S.G.: Finite-context models for DNA coding. In: S. Miron (ed.) Signal Processing, pp. 117–130. INTECH (2010)