Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Hypothesis**

Cross chromosomal similarity for DNA sequence compression

Choi-Ping Paula Wu[1, *], Ngai-Fong Law[1] and Wan-Chi Siu[1]

[1]Centre for Signal Processing, Department of Electronic and Information Engineering, The Hong Kong Polytechnic University;
Paula Wu* - E-mail: paul.a@polyu.edu.hk; * Corresponding author

received May 30, 2008; revised June 27, 2008; accepted July 06, 2008; published July 14, 2008

Abstract:
Current DNA compression algorithms work by finding similar repeated regions within the DNA sequence and then encoding these regions together to achieve compression. Our study on chromosome sequence similarity reveals that the length of similar repeated regions within one chromosome is about 4.5% of the total sequence length. The compression gain is often not high because of these short lengths. It is well known that similarity exist among different regions of chromosome sequences. This implies that similar repeated sequences are found among different regions of chromosome sequences. Here, we study cross-chromosomal similarity for DNA sequence compression. The length and location of similar repeated regions among the sixteen chromosomes of *S. cerevisiae* are studied. It is found that the average percentage of similar subsequences found between two chromosome sequences is about 10% in which 8% comes from cross-chromosomal prediction and 2% from self-chromosomal prediction. The percentage of similar subsequences is about 18% in which only 1.2% comes from self-chromosomal prediction while the rest is from cross-chromosomal prediction among the 16 chromosomes studied. This suggests the importance of cross-chromosomal similarities in addition to self-chromosomal similarities in DNA sequence compression. An additional 23% of storage space could be reduced on average using self-chromosomal and cross-chromosomal predictions in compressing the 16 chromosomes of *S. cerevisiae*.

Keywords: DNA; sequence; chromosome; prediction; *S. cerevisiae*

Background:

A DNA sequence is a long stretch consisting of four types of nucleotides: Adenine (A), Cytosine (C), Guanine (G) and Thymine (T). The lengths of the 24 chromosomes in human range from 50 to 250 million base pairs **[1]**. Compression is desirable to uncover similarities among sequences, and provide a means to understand their properties in addition to reduce storage requirement **[2, 3]**. State-of-the-art compression algorithms work by finding approximate repeats and approximate reverse complement repeats in the current DNA sequence. The approximate repeats refer to those repeats that contain errors, i.e., with certain unmatched nucleotides between two subsequences. The reverse complement means nucleotides in a sequence is the reverse ordering of nucleotides in another sequence, but with each nucleotide replaced with its complement. For example, the subsequences AAACGT and ACGTTT are reverse complement repeats as (A, T) and (G, C) are complement bases.

Most DNA-based compression algorithms rely on encoding together similar repeated regions found within the sequence. Biocompress is the first algorithm designed specifically for compressing DNA sequences **[4]**. Both Biocompress and its second version Biocompress-2 are based on a sliding window algorithm known as LZ77 **[4-6]**. Exact matches and complementary palindromes are found so that the matched subsequences can be encoded with respect to identical subsequences occurred in the past. The matched sequences are replaced by two parameters: the start position of the previously occurred subsequence and the repeat length in the analysis. An order-2 arithmetic coding (arith-2) is used for or non-repeated regions.

Cfact **[7]** utilizes a two pass algorithm. A suffix tree is used for finding exact matches in the first pass. The matched subsequences are encoded using previous references if there is a compression gain. Otherwise, they are kept uncompressed in the second pass. Unlike Biocompress and Cfact, GenCompress **[2, 8]** used approximate matches in addition to exact matches. GenCompress-1 uses substitutions while GenCompress-2 uses deletions, insertions and substitutions to encode repeats. CTW+LZ method is based on the context tree weighting method (CTW) and LZ based compression **[3]**. Long exact/approximate repeats and complementary palindromes repeats are encoded by the LZ-based algorithm, whereas short subsequences are compressed using CTW. Execution time is too high for long sequences despite obtaining good compression.

DNACompress **[9]** consists of two phases. The first phase finds all approximate repeats including complementary palindromes by a separate software tool called PatternHunter **[10]**. The second phase encodes those approximate repeats and non-repeating regions. DNACompress not only provides good compression, but is also significantly faster than GenCompress. DNAC **[11]** consists of four phases. The suffix tree is built in the first phase to locate exact matches. In the second phase, all the exact repeats are extended to approximate repeats by dynamic programming. In the third phase, the optimal non-overlapping repeats are extracted from the overlapping regions. All the repeats are then encoded in the last phase. Similar to DNAC, DNAPack uses dynamic programming approach for identification and encoding of repeats **[12]**.

It is seen that all DNA-based compression algorithms find repetitions within the DNA sequence. Longer repetitive length implies higher compression gain. The compression ratio attained is high if highly similar subsequences are found. It is well known that there are similarities among different chromosome sequences. However, cross-chromosomal similarities are seldom exploited in DNA sequence compression. The objective of this paper is to study self-chromosomal and cross-chromosomal similarities; to investigate use of cross-chromosomal similarity for compression; and to demonstrate the advantage of cross-chromosomal similarity in multiple sequence compression. It should be noted that similar subsequences located within the chromosome sequence are called self-(chromosomal) similarity/ self-reference while those located in other chromosome sequence are called cross-(chromosomal) similarity/ cross-reference in this analysis.

Methodology:

Dataset

The sixteen chromosomes of *S. cerevisiae* are used to investigate chromosome similarities. They are downloaded from elsewhere **[13]**. The search engine PatternHunter is employed to search for repeats. All repeats are ranked by a score. It defines similarity between two subsequences. A large score means that the two subsequences are similar to each other.

Similarity between two chromosome sequences

Repetitions between two chromosome sequences are first investigated. We found that the repetitive lengths found within a chromosome sequence are not necessarily longer than that found in another chromosome. In an example, the top three longest repetitive regions found within Chr I are about 15000, 2600 and 2300 bases long. However, the lengths of the repetitive regions found between Chr I and Chr VIII are 17000, 14000 and 6800. This example shows that the lengths of the repetitive regions found between Chr I and Chr VIII are always larger than those found within Chr I alone. The lengths of the repetitive regions found between Chr I and other chromosomes including Chr II, Chr IV, Chr VII, Chr X, Chr XII, Chr XIII, Chr XV and Chr XVI are also significant. Similar observations are found for other chromosomes. This shows that cross-chromosomal similarities between two sequences are often significant. They are exploited beneficially for compression purposes.

Cross-chromosomal predictions

The potential gain in cross-chromosomal compression is obtained by finding the total lengths of subsequence in the current chromosome sequence that is predicted from another chromosome sequence. The lengths of these cross-reference subsequences determine the potential compression gain in multiple DNA sequences compression. Long length implies a high compression ratio.

Chromosomes are classified into three classes as shown in Column 2 of Table 1 (supplementary material) using comparison of self-chromosomal and cross-chromosomal prediction length. The first class, consisting of Chr III, Chr V, Chr VIII, Chr XI and Chr XIV, has high similarities with chromosomes other than itself. More than 8 chromosomes show cross-repetitive lengths longer than the self-repetitive length. This implies that a potentially high compression gain can be obtained if these sequences employ cross-referencing strategy with subsequences predicted from other chromosomes.

The second class consists of Chr VII, Chr XIII, Chr XV and Chr XVI. Although just 3 to 7 chromosomes show cross-repetitive lengths longer than the self-repetitive length, a potential compression gain is also expected since the cross-repetitive lengths are still large. The last class consists of Chr I, Chr IV and Chr XII. There are less than 3 chromosomes having cross-repetitive lengths longer than the self-repetitive length. In Chr I, the number of nucleotides predicted from Chr VIII is almost doubled of the self-repetitive length within Chr I. In addition, in Chr XII and Chr IV, self-referencing and cross-referencing are indeed significant since the number of nucleotides respectively predicted from Chr IV and Chr XII is comparable to the self-repetitive length. Thus, the combination of self-repetitive and cross-repetitive lengths would still contribute to better compression.

Discussion:

Besides considering the total length of subsequences within a chromosome that can be referenced from other chromosomes, their distribution within the sequence are also important. Let the subsequence in a sequence S that is similar to a subsequence in sequence i be S(i) and the subsequence in S that is similar to a subsequence in sequence j be S(j), the total length of subsequences within S that can be referenced from i and j is given by $T=|S(i)|+|S(j)|-|S(i)\cap S(j)|$. Obviously if these subsequences are well spread out such that $|S(i)\cap S(j)|$ is zero, i.e., they do not overlap in position, T is maximized. This implies that a high proportion of the nucleotides within S can be predicted by cross-referencing among chromosomes, resulting in a high compression gain.

Locations of similar subsequences among chromosomes

Figure 1 shows locations of similar subsequences found among chromosomes. The five chromosomes in Figure 1a prove that the portions of self-repetitive regions are very small, as compared to that of cross-repetitive regions with other chromosomes. In the case of Chr XI, Chr XIV, Chr VIII and Chr V, the self-repetitive subsequence is not seen. Similar subsequences predicted from other chromosomes appear in different locations. For example, in Chr XI, the four similar subsequences appear in four different regions. Similar observations are also seen from Figure1b for second class.

Figure 1c shows locations of similar subsequences for the third class. In Chr I, we can see that the portions of cross-repetitive regions with either Chr VIII or Chr XV are much larger than that of self-repetitive region. In Chr XII, the portions of cross-repetitive regions with Chr XIII or Chr IV are comparable to that of self-repetitive region. In Chr IV, the portions of cross-repetitive regions with Chr XII are also comparable to that of self-repetitive region. Figure 1 illustrates that cross-repetitive regions are often significant when compared with self-repetitive regions. Furthermore, subsequences that are cross-referenced from other chromosomes appear in different locations within the chromosome.

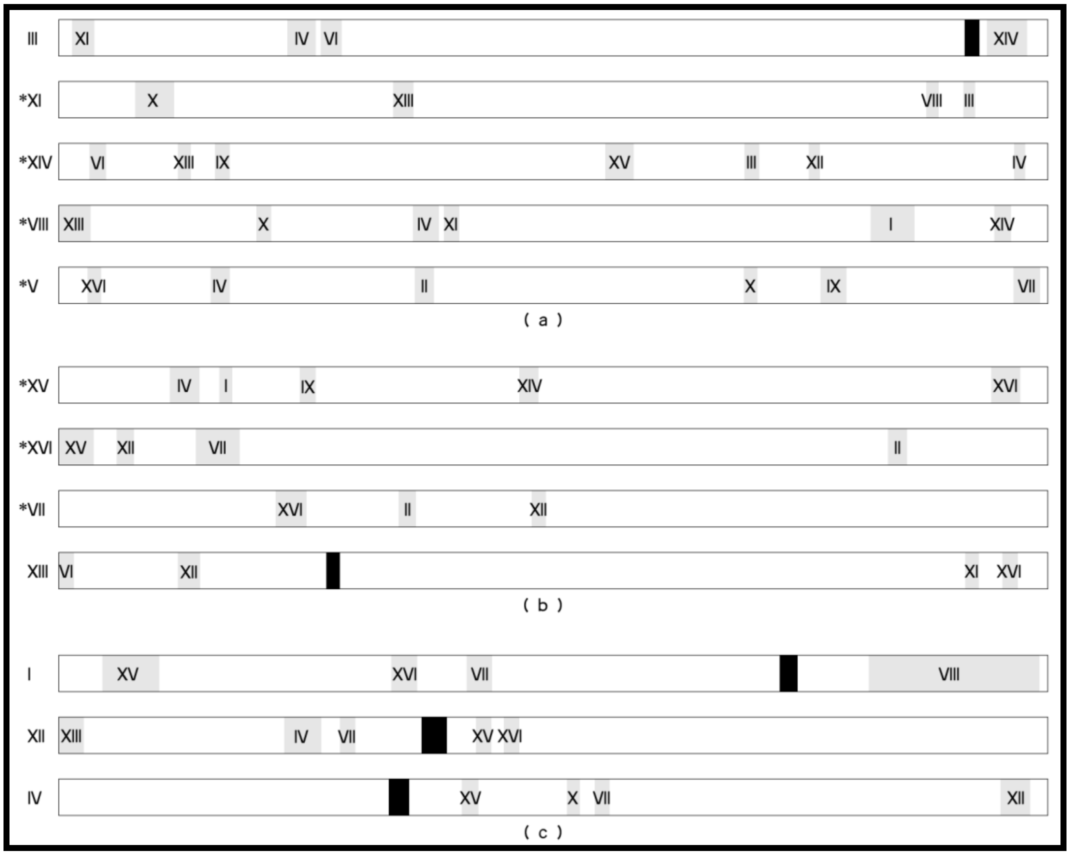

Figure 1: Locations and lengths of similar subsequences. Locations of similar subsequences for (a) the first class, (b) the second class and (c) the third class are shown. The colored region indicates the length and the approximated location of a repetitive subsequence that can be found in a particular chromosome. Self-sequence repetitions are shown in black color while cross-sequence repetitions with other chromosomes are in light grey color. The sequence number of the chromosome is marked inside the colored region. Only significant regions are presented (i.e. with score larger than 100 in the PatternHunter software tool) and are drawn on scale for each chromosome. Note that the * next to the chromosomes represent those chromosomes without significant self-sequence repetitions.

Cross-chromosomal predictions

We considered two cases for cross-chromosomal prediction. In the first case named prediction-2, the prediction is restricted to only two chromosome sequences including the current chromosome sequence. In the second case named prediction-16, the prediction is from the current chromosome and the other 15 chromosome sequences. The self-prediction and cross-predictions are examined to remove all those overlapping regions and are sorted to produce a combined list. This combined list is then used to show all the repetitive regions including both self-chromosomal and cross-chromosomal repetitions. Table 1 (supplementary material) shows the results of the analysis.

In prediction-2, the cross-predictions come from another chromosome that gives the longest repetitive regions. In column 5(a) and (b), it is clear that the cross-predictions are always significant, as compared to the self-predictions. In particular, the cross-predictions are in the range of 5% to 22%. In contrast, the self-predictions are always less than

3.5%. In prediction 16, the cross-predictions from the other 15 chromosomes are listed in column 5(c). The cross-predictions are in the range of 12.5% to 32%, whereas the self-predictions are always less than 3%. As a result, our study indicates that it would be advantageous to compress different chromosomes together to take into account both self-similarity and cross-similarities.

Self-chromosomal and cross-chromosomal similarities for compression

Two chromosome sequences are compressed by considering self-chromosomal similarities or by both self-chromosomal and cross-chromosomal similarities using GenCompress.

Column 6(b) shows the number of bits used if two chromosomes are compressed separately (consider only self-chromosomal similarities). Column 6(c) shows the number of bits if the two chromosomes are concatenated and compressed together. The savings are shown in Column 7. Column 7(a) is the savings resulting from self-chromosomal predictions as compared to the no compression case. Column 7(b) is the savings from cross-chromosomal predictions as compared to the self-chromosomal predictions. We can see that there is always extra savings by considering cross-chromosomal predictions in addition to self-chromosomal predictions. Since the cross-prediction found between Chr I and Chr VIII is the highest as shown in Column 5(a), the saving from cross-chromosome predictions is the largest. While the size of repetitive regions in cross-predictions ranged from 5% to 22%, their savings are between 9% and 60%.

Conclusion:

The state-of-the-art DNA compression algorithms consider repetitions within the current sequence. However, similarities exist across different chromosome sequences. Here, we described cross-chromosomal similarities in *S. cerevisiae*.

We find that cross-chromosomal similarities are always significant as compared to self-chromosomal similarities. For example, the average percentage of similar subsequences between two chromosome sequences is about 10% in which 8% comes from cross-chromosomal prediction and 2% from self-chromosomal prediction. For 16 chromosome sequences of *S. cerevisiae*, the average percentage is about 18% in which 16.8% comes from cross-chromosomal prediction and 1.2% from self-chromosomal prediction. Therefore, it would be advantages to compress different chromosome sequences together to take advantage of cross-chromosomal similarities.

Our experimental results demonstrate that on average an additional 23% of storage is reduced in cross-chromosomal predictions as compared to self-chromosomal predictions. Therefore, a high compression ratio is obtained by considering both self-prediction and cross-predictions for the entire set of chromosomes. Our future work is to extend this analysis to cross-similarities between species and to develop a systematic approach for incorporating both self-chromosomal and cross-chromosomal predictions into DNA sequence compression.

Acknowledgement:

This work is supported by the Centre for Signal Processing (1BB9F), the Hong Kong Polytechnic University. Paula Wu acknowledges the research studentships provided by the University.

References:

[01] http://www.ornl.gov/sci/techresources/Human_Genome/project/info.shtml
[02] M. Li, *et al., Bioinformatics,* 17: 149 (2001) [PMID: 11238070]
[03] T. Matsumoto, *et al., Genome Inform Ser Workshop Genome Inform.,* 11: 43 (2000) [PMID: 11700586]
[04] O. Delgrange, *et al., Pac Symp Biocomput.,* 254 (1999) [PMID: 10380202]
[05] V. D. Gusev, *et al., Bioinformatics,* 15: 994 (1999) [PMID: 10745989]
[06] X. Chen, *et al., IEEE Eng Med Biol Mag.,* 20: 61 (2001) [PMID: 11494771]
[07] É. Rivals, *et al., Biochimie.,* 78: 315 (1996) [PMID: 8905150]
[08] X. Chen, *et al., Genome Inform Ser Workshop Genome Inform.,* 10: 51 (1999) [PMID: 11072342]
[09] X. Chen, *et al., Bioinformatics,* 18: 1696 (2002) [PMID: 12490460]
[10] B. Ma, *et al., Bioinformatics,* 18: 440 (2002) [PMID: 11934743]
[11] http://www.im.ntu.edu.tw/theses/r92/R91725026.pdf
[12] A. J. Pinho, *et al., IEEE Trans Biomed Eng.,* 53: 2148 (2006) [PMID: 17073319]
[13] ftp://ftp.ncbi.nlm.nih.gov/genomes/Saccharomyces_cerevisiae/

Edited by P. Kangueane
Citation: Wu *et al.*, Bioinformation 2(9): 412-416 (2008)

Supplementary material

Chr *b*	Class of Chr *b*	Length of Chr *b*	Chr *a*	Repetitive length in bases (%)				Total no. of bits required for Chr *a* and Chr *b*			Total no. of bits saved (%) from	
				Prediction-2		Prediction-16						
				a. Cross predictions	b. Self predictions	c. Cross predictions	d. Self predictions	a. Without compression	b. Compressing separately	c. Compressing together	a. Self predictions	b. Cross predictions
I	3	230208	VIII	50536 (22.0%)	5526 (2.4%)	74058 (32.2%)	4209 (1.8%)	1585702	1499256	1447264	86446 (5.8%)	51992 (60.1%)
III	1	316617	XIV	28818 (9.1%)	6416 (2.0%)	54714 (17.3%)	4737 (1.5%)	2201900	2112392	2096936	89508 (4.2%)	15456 (17.3%)
IV	3	1531918	XII	79909 (5.2%)	44897 (2.9%)	197093 (12.9%)	31532 (2.1%)	5220186	4855360	4815592	364826 (7.5%)	39768 (11.9%)
V	1	576869	VII	39909 (6.9%)	6859 (1.2%)	94421 (16.4%)	4094 (0.7%)	3335630	3177920	3149392	157710 (5.0%)	28528 (18.1%)
VII	2	1090946	XVI	66619 (6.1%)	17936 (1.6%)	156422 (14.3%)	5812 (0.5%)	4078016	3881368	3841968	196648 (5.1%)	39400 (20.0%)
VIII	1	562643	I	36808 (6.5%)	15086 (2.7%)	104628 (18.6%)	6129 (1.1%)	1585702	1499256	1447432	86446 (5.8%)	51824 (59.9%)
XI	1	666454	X	35013 (5.3%)	3930 (0.6%)	85186 (12.8%)	2655 (0.4%)	2824398	2729104	2720464	95294 (3.5%)	8640 (9.1%)
XII	3	1078175	IV	87678 (8.1%)	36310 (3.4%)	164488 (15.3%)	27744 (2.6%)	5220186	4855360	4816424	364826 (7.5%)	38936 (10.7%)
XIII	2	924429	XII	51845 (5.6%)	17079 (1.9%)	117607 (12.7%)	12670 (1.4%)	4005208	3742920	3713616	262288 (7.0%)	29304 (11.2%)
XIV	1	784333	XV	51084 (6.5%)	8952 (1.1%)	122687 (15.6%)	6396 (0.8%)	3751244	3604120	3566944	147124 (4.1%)	37176 (25.3%)
XV	2	1091289	IV	70056 (6.5%)	14168 (1.3%)	183165 (16.8%)	7434 (0.7%)	5246414	4973664	4931832	272750 (5.5%)	41832 (15.3%)
XVI	2	948062	VII	67662 (7.1%)	8658 (0.91%)	145116 (15.3%)	4860 (0.5%)	4078016	3881368	3845376	196648 (5.1%)	35992 (18.3%)
Average				55495 (7.9%)	15485 (1.8%)	124965 (16.7%)	9856 (1.2%)	3594384	3401007	3366103	193376 (5.7%)	34904 (23.0%)

Table 1: Lengths of cross-chromosomal and self-chromosomal repetitions and the number of bits required/saved in compressing two chromosomes. Column 2 and 3 give the class and the number of bases of Chr b, respectively. Chr a in Column 4 is the most similar chromosome with Chr b in Column 1. In Column 5, the sub-column (a)(b) and (c)(d) provide the length of repetitive regions in cross-chromosomal prediction from one chromosomes - Chr a (i.e. prediction-2) and from the other 15 chromosomes (i.e. prediction-16), respectively. The sub-column (a) (c) and (b) (d) refers to cross-chromosomal and self-chromosomal predictions, respectively. The total number of bits required for storing Chr a and Chr b without any compression is listed in Column 6(a). In considering self-chromosomal repetitions (i.e. compressing Chr a and Chr b separately), the total number of bits required by GenCompress scheme is shown in Column 6(b). In considering, both self-chromosomal and cross-chromosomal repetitions (i.e. Chr a and Chr b are concatenated together before compression), the total number of bits required is shown in Column 6(c). Column 7(a) shows the result of the number of bits saved in self-chromosomal repetitions obtained by 6(a) with 6(b) and column 7(b) shows the additional saving in bits from cross-chromosomal repetitions which is obtained by comparing 6(b) and 6(c).