A novel approach to compress dna repetative sequences in bio-informatics

S M B Chowdary1, Samparthi V S Kumar2, Dr Deepak Nedunuri3 And Vmnssvkr Gupta4

1Sr. Assistant Professor, Department of CSE, Sir C R Reddy College of Engineering, Eluru.
2Assistant Professor, Department of CSE, Sir C R Reddy College of Engineering, Eluru.
3Associate Professor, Department of CSE, Sir C R Reddy College of Engineering, Eluru.
4Associate Professor, Department of CSE, S R K R Engineering College, Bhimavaramluru.

E-Mail: mohanbabus2009@gmail.com

Abstract: In recent days numbers of gigabyte sequences of nucleotides are stored in a common database Genbank. All the victimization Deoxyribonucleic acid sequences for biological functions are to store the large number of Genomes in a compressed type in economically. Despite the fact that Deoxyribonucleic corrosive arrangements are put away in a packed kind, the information on Deoxyribonucleic corrosive groupings square measure hang on in science databases. For a four-letter alphabet in DNA (Adenine(A), Cytosine(C), Guanine(G) and Thymine(T)), an average description length of 2 bits per base is that the max length required to encode DNA. To reexamine the previous art of compression techniques and its merits and de merits, a novel attempt is initiated. Based on the comparative study of existing algorithms a new method proposed for DNA compression without depending on statistics of sequence set.

Key Words: DNA, GenBank, Phylogenetic Tree, Genomes

1. Introduction
There is large number of databases available for human genomic data. Resulting to the challenging environment on the changes of genomic data (DNA or Protein). The four classifications of DNA were adenine (A), cytosine (C), Guanine (G) and Thymine (T). Without compression two bits required to encode each base by information theory [2]. Even in existing general compression tool like gzip are used [3]. Thus, it became an essential need to compress DNA sequences by developing specific compression algorithms.

1.1 DNA
The genetic information can be passes from one bread to another which is incorporated by DNA is used in the enhancement and functioning of living organisms. Nucleotides and Phosphate, both the teams joined by organic compound were the two long polymers of DNA (Fig.1). In each cell, the organization of DNA is formed into long structures called chromosomes, for ex the human genome contains 23 chromosome pairs. In DNA replication chromosomes are duplicated before cell division.
1.2 Characteristics of DNA sequences
Genetic information is carried through DNA for all the generations. Four types of Nucleotides [4], as shown in Table 1.

| Bases | Nucleotides | Complement |
|-------|-------------|------------|
| Adenine | A           | T          |
| Cytosine | C           | G          |
| Guanine   | G           | C          |
| Thymine   | T           | A          |

Figure 2: DNA chain with complement pairs A<>T and C<>G

1.3 Applications of bioinformatics
The following section describes some of the available Bioinformatics applications.

- **3-D protein prediction:** A microscopic structure known as “Secondary Structure”, determines a detailed analysis and regular patterns that is expressed by 3-D structure of proteins.

- **Phylogenetics:** The study of progressive enhancement and history of species is known as phylogenetics. The outcome of Phylogenetics is an evolutionary tree which explains the generative distance between the set of sequences.
2. Literature review

2.1 Introduction to data compression
Globally there is a wide variety of change for the evolution of computers which can be carried out personally. Radical change in the growth of technology, this change has certainly happened as a part of great contributions around the world. In this connection a remarkable role has been played in this aspect by the advent of data compression. The gigantic world of internet is considerably utilizing data compression techniques in many ways, where it never has been possible without the dReaNRe’s technology booms.

2.2 Similarities of sub sequences among chromosomes
DNA pressure approaches depend on looking redundancies, better pressure proportions can be accomplished by comparative sub successions along the present DNA grouping. Examinations of sub successions could exist among different species which are closer as far as dynamic separation [5] or distinctive chromosome types of one animal categories [6].

2.3 Existing compression methods for dna sequences
Many algorithms failed to compress DNA sequences due to the encoded specialty of text [7]. DNA compression includes many technologies, metrics and computational algorithms, presents the following list.

- Probabilistic Markov models[8]
- Burrows Wheeler Transform[9]
- Greedy algorithms[10]
- Dynamic programming approaches
- Based on the Normalized maximum Likelihood[11]

2.4 s.cerevisiae data set
S.cerevisiae consists of sixteen sequences which are denoted as chromosome I to chromosome XVI shown in

| Chr | Reference   | No. of bases |
|-----|-------------|--------------|
| I   | GI:5059312  | 144157       |
| II  | GI:5059311  | 605184       |
| III | GI:4275985  | 217332       |
| IV  | GI:5059323  | 1229605      |
| V   | GI:7276232  | 391086       |
| VI  | GI:4274217  | 183702       |
| VII | GI:5059321  | 784707       |
| VIII| GI:5088258  | 402792       |
3. Materials and methods

3.1 Data extraction
The open source data sets S.cerevisiae and S.pombe can be used as an input data for the dReaNRe algorithm. The Saccharomyces genome data base (SGD) of Yeast (bacteria) directory contains sequence files of the sixteen nuclear chromosomes and the mitochondrial order of S. cerevisiae strain S288C, in FASTA format. Individual chromosome sequence files are updated once there's an update to the systematic reference sequence stored in SGD. The longest and the shortest chromosomes are ChrIV and ChrI according to their sizes.

3.2 The dreanre algorithm
In prior, differencing has its inceptions in each longest regular subsequence (LCS) calculations [12] and furthermore the string-to-string amendment disadvantage. The establishment constraints of LCS after they made a brand new file evaluate the program [13]. An improved metric edit distance [14] for distinction of files and techniques supported this technique increased the utility and speed of file differencing. Probabilistic mismatching [15] mostly depend on Compression –based distance measures (CBMs) with neglect the locations of variations, aren't distinct enough among completely different categories. The three patterns are listed below:

P1= ACGTGTAC
P2= CGATGCAT
P3= TCGAAACGT
Table 3: Variance Matrix (R*-Repeats-self similarity-Vice-Versa)

| S.no | P1   | P2 | P3 |
|------|------|----|----|
| P1   | -    | 5  | 6  |
| P2   | R*   | -  | 6  |
| P3   | R*   | R* | -  |

The proposed algorithm dReaNRe (DNA repetitive and non repetitive encoding analysis model) developing of distinction algorithmic rule particularly for DNA information sets. Some sequences are identical to the reference have minimal impact on storage despite their length or depth of sequencing coverage is the key component of the algorithmic instruction. To boot, reference ordering is used strictly as a compression framework and don't need any biological correctness for the reference. From a spread of “identical” sequences, the information is stored as an approach, e.g., those exploitation identical individual or species. Then as an alternate compression framework these sequences function upon that we are able to offer economical storage.

3.3 The DNA coder

The first stage of dReaNRe is DNACoder which is one of the important algorithmic components. The main function of this component is to identify the similarities among DNA sequences and then encode them to achieve better compression ratios. The DNACoder have two sub components Encoder and Decoder for compression and decompression processes.

The compression process as shown in Figure 4.1. This process will include DNACoder and dReaNRe-2 Encoder. The sequence such as NC-001133 (shown in Figure 4.2) is applied as input into the DNACoder and the output is bypassed to dReaNRe-2 encoder to make compressed file. Obviously the compressed one is having lesser in size than that of the original one.

The first sub component of DNACoder is DNAEncoder and it will consist of the following steps.

- Retrieval of identical subsequences among different DNA sequences.
- Identical subsequences can be sorted out according to their priority.
- Removal of redundancy from different DNA sequences.
- Sort out the non-overlapping subsequences according to their index position.
- Residual sequences after removal of redundancy, bypassed to dReaNRe-2 Coder for further compression.

In the first step mentioned above, similar subsequence can be extracted by using pattern Hunter. In addition to that, similarities may found in self and cross chromosomal sequences. The results of two similar chromosome sequences can be stored in variance table which will contains two files i.e. (selfsimilar.aln and cross-reference.aln) shows variation point (which shows variations) among chromosomes. In every .aln file, corresponding subsequences together with the scores, the direction, the beginning and also the ending positions of query are recorded. A precedent record is appeared in Figure 3. The inquiry and the subject successions see the comparative subsequences inside the present grouping on account of self-referencing. The scores identify with the similitude between the inquiry and furthermore the subject successions. High score demonstrates that they're much the same as each other. The heading are frequently either 'in addition to' or 'short'.
Figure 3: NC-001133.aln file

The 'in addition to' heading recommends that an estimated rehash, that the arrangement should be peruse in rising request. The 'short' heading recommends that a turnaround supplement rehash, along these lines the succession should be peruse backward course. The start and closure position of the subsequence denotes its area inside the first arrangement. A posting is made for putting away data concerning monotonous records with score over a limit. The edge is set all together that exclusively fundamental comparative subsequences square measure thought of by our anticipated pressure calculation.

In the third step, each tedious record inside the joined rundown from the second step is analyzed. In particular, covering comparative subsequences from two records are cut. On the off chance that the point groupings in 2 redundant records are covered with each other in position, the covering half will be whole inside the record with the following score and be evacuated inside the diverse record with a lower score. The guideline behind is to remain an extended monotonous length rather than a short one. In the event that the length of the cut dreary record is a littler sum than the edge once evacuating the covering part, the redundant record will be off from the rundown. In the wake of evacuating all the covering components inside the comparative subsequences, DNACoder spares the varieties between the 2 comparative subsequences. This progression is vital as rough rehashes, rather than exact rehashes, are thought of in DNACoder.

Figure 4: DNACoder.

This instance includes all the operations - substitution, deletion and insertion. For example, the base "C" within the fifth position of the query subsequence is replaced by "A" within the corresponding position of the subject subsequence. There square measure 2 a lot of bases extra within the subject subsequence in between the twelve and also the 13 bases of the question subsequence.

4. Results and discussions
4.1 Simulation results on s.cerevisiae-sgd
The Saccharomyces Cerevisiae Genome Database (S.cerevisiae-SGD) can be considered in this experiment to assess the effectiveness of the compression algorithm. The S.Cerevisiae is an open source repository can be downloadable from the web:- ftp://ftp.ncbi.nlm.nih.gov/inigenomes/ Different representations of SGD can be shown in Figure 5.

![Figure 5. representations of SGD](image)

4.2 Simulation results on s.pombe-pbd
The 3 chromosome arrangements of schizosaccharomyces pombe (S. pombe) are tried. Note that the normal length of S. pombe is 4200k, which is significantly longer than that of S. cerevisiae. In this way, the long length would build the pressure time altogether since it would go searching for the redundant records inside the successions.

![Figure 6: Area Chart Representation of SGD](image)

5. Conclusion
The simulation study has been performed between the chromosome sequences of Saccharomyces Cerevisiae Genome Database (S.cerevisiae-SGD) and schizosaccharomyces pombe (S. pombe). In this thesis, for the first time an attempt is made to recognize self and cross sequence similarities and make it use in different real time applications. A detailed analysis has been performed in between SGD and PBD to find out the variations in between the sequences, length and locations by constructing variance table. From this analysis it is observed that the cross similarities have much impact than that of self similarities in SGD and PBD. To till date most of the methods consider self references only but the present study shows to attain higher compression ratios it would be more advantages if cross similarities can be considered.

References
[1] B.A.Gata, “Database Similarity Searching Using BLAST and FastA”, Australasian Biotechnology, vol. 5, pp.282-290, 1995.
[2] U. Ozkan Nalbantoglu, J. David Russell and Khalid Sayood, "Data Compression Concepts and Algorithms and Their Applications to Bioinformatics", OPEN ACCESS, Entropy, vol.12, pp.3452; 2010.
[3] T. Matsumoto, K. Sadakane, H. Imai and T. Okazaki, “Can General-Purpose Compression Schemes Really Compress DNA Sequences?”, Currents in Computational Molecular Biology, pp.76-77, 2000.

[4] Open Bioinformatics Foundation: BOSC”, Official website. Open Bioinformatics Foundation, Retrieved 10 May 2011.

[5] Sequence Similarity Search – BLAST: http://blast.genome.jp/

[6] Hari Prasad, P.V. Kumar Efficient compression through CBT, vol.4, no.2, 2102 CIIT proceedings.

[7] Hari Prasad, Kumar, A new revisited DNA CRAMP tool based approach of chopping DNA repetitive and non repetitive sequences of genomes in vol.9, issue 6, November 2012 IJCSI proceedings.

[8] P. J. S. G. Ferreira, A. J. R. Neves, V. Afreixo, and A. J. Pinho, “Exploring Three-Base Periodicity for DNA Compression and Modeling,” presented at ICASSP 2006, vol. 5, Toulouse, France, 2006.

[9] D. Adjeroh, Y. Zhang, A. Mukherjee, M. Powell, and T. Bell, “DNA Sequence Compression Using the Burrows-Wheeler Transform,” presented at IEEE Computer Society Bioinformatics Conference (CSB’02), 2002.

[10] A. Apostolico and S. Leonardi, “Compression of biological sequences by greedy off-line textual substitution,” presented at Data Compression Conference, pp. 143-152, 2000.

[11] G. Korodi and I. Tabus, “An efficient normalized maximum likelihood algorithm for DNA sequence co

[12] Stong J.A. Storer, and T.G. Szymanski, “Data compression via textual substitution,” J. ACM, 29, 928–951, 1982.

[13] R. WAGNER, and M. FISCHER, “The string-to-string correction problem”, Journal of the ACM 21, 1, pp.168–173, Jan. 1973.

[14] M. Li, JH. Badger, X. Chen, S. Kwong, P. Kearney, HY. Zhang, “An information-based sequence distance and its application to whole mitochondrial genome phylogeny”, Bioinformatics, 17:149-154, 2001.

[15] Stong J.A. Storer, and T.G. Szymanski, “Data compression via textual substitution,” J. ACM, 29, 928–951, 1982.