International Journal of Technology (2010) 1: 1-10
ISSN 2086-9614

AN INTEGRATED APPROACH FOR STATISTICAL GENOME SEQUENCE ANALYSIS BETWEEN GENETIC DATASETS

Hassan Mathkour[1*], Muneer Ahmad[1], Hassan Mahmood khan[1]

[1]*Department of Computer Science*
College of Computer & Information sciences
King Saud University, P.O. Box 51178, Riyadh 11543
Kingdom of Saudi Arabia

(Received: May 2010 / Revised: June 2010 / Accepted: July 2010)

ABSTRACT
Genome Sequence Analysis for genetic datasets by using ORF (Open Reading Frames) techniques is an interesting area of research for bioinformatics researchers nowadays. There is a strong research focus on comparative analysis between genetic behaviors and diversity of different species. Contrary to whole genome sequence analysis, scientists are now trying to concentrate specifically on layered analysis to get a better insight of relevancy among genetic datasets. This phenomenon will help to better understand species. An ORF statistical analysis for genetic data-sets of species Chimera Monstrosa and Poly Odontidae is presented. For completion of this analysis, we use a hybrid approach that combines a generic mechanism for statistical analysis with specific approach designed for out performance. At first instance, genetic datasets are refined for better usage at next level. These sets are then passed through layers of filters that perform DNA to Protein translation. Statistical comparison is performed during this translation. This layered architecture helps in better understanding of the degree of similarity and differences in genomic sequences.

Keywords: Amino acid; Codon count; Distributed generation; Open Reading Frame; Pre-processing filter

1. INTRODUCTION
Due to existing and continuously growing bulk of biological data coming from genome projects and experiments nowadays, protein structure prediction and its systematic translation need an efficient and effective way to sequence, analyze and compare coded biological DNA sequence information. The genome sequence analysis is directly related to the sequence comparison and alignment. Sequence similarity is a way to predict the functional similarity among genes and this has been used as a tool for functional prediction. Analysis and Comparison of DNA sequences and genes is useful for finding the facts about how these genes are organized and what are the similarities and differences (Gupta, et al., 2007). These fundamental problems are NP hard (Weng, et al., 2006; Kumar, et al., 2007) and need optimal solutions that can be achieved by improving algorithms and computing architecture. (Ma & Chan 2003). A little work has been done in hybrid statistical analysis of genomic data against exponentially increasing problem size. Usage of Computer aided techniques are not the solution.

* Corresponding author's email: binmathkour@yahoo.com, Tel. +966-4676605, Fax. +966-4675423

There is need to work in computational molecular biological experiments by means of DNA sequence analysis. Finding a unique sequence on the entire target genome is one of the most important problems in molecular biology (Gowda, et al., 2007).

The overall goal of this paper is to present an integrated approach that performs comparative analysis between same species revealing that peptide translation in both has a degree of differences. This task is accomplished by using ORF with statistical analysis. The method used for this purpose is a composite technique that consists of a series of filter from preprocessing level to final analysis.

The human genome project has built rich databases which attracted research interests from biologists and computer scientists to explore and mine these precious data-sets. The computer aided applications now can reveal the hidden information in complex helix DNA structure. They also made it possible to perform fast and accurate analysis. This has been made effective with the availability of cost effective and handy analysis tools. Scientists have developed novel ideas, implemented and resolved complex situations in computational biology whose direct feasible solutions were not possible in yielding optimal solutions in some cases for sequence analysis, an NP hard problem (Kurata, et al., 2003; Weng, et al., 2006; Kumar, et al., 2007; Miranker, 2008).

This paper is organized as follows: Section 1 is the Introduction. Section 2 highlights some related work. Section 3 describes the proposed technique (elaborated in subsections). Section 4 contains fundamentals concluding remarks for this comparative analysis. Section 5 represents an acknowledgement and Section 6 contains references.

2. RELATED WORK

Kumar, et al., (2007) gives an approach for a distributed bioinformatics computing system. It was designed for disease detection, criminal forensic and protein analysis. It is a combination of different distributed algorithms that are used to search and identify a triplet repeat pattern in a DNA sequence. It consists of a search algorithm that computes the number of occurrences of a given pattern in a genetic sequence. The distributed subsequence identification algorithm was to detect repeating patterns with sequential and distributed implementation of algorithms relevant to different triplet repeat search patterns and genetic sequences. The result of this system shows that as complexity of the algorithm increases, the response time also increases. There is space to make this work better for more DNA sequences of various lengths.

Kurata, et al., (2003) presents a technique to find unique genome sequences from distributed environment databases. Kurata used implementation of the method upon the European Data Grid and showed its results. The author worked on the unique sequences of E. Cole 0157 (12 genome). The genome is divided into smaller pieces being processed individually. In an example quoted by author, the total file size is 256 MB when it is hashed to 7. It is possible to divide the genomic files into at most 47 = 16384 pieces of 15 KB each. This method results in memory consumption and increases file size. This data grid method is not useful for parallelizing biologically important data.

Li, et al., (2003) proposes a genome sequence learning method by simplifying Bayesian network. The nodes in Bayesian networks are selected as features. A feature selection algorithm is used for structure learning. This algorithm is based on genetic algorithm. The researcher used dataset of 570 vertebrate sequences, including 2079 true donor sites. This approach is limited to the donor site prediction and also confirms that the nucleotides closer to donor site are the key elements in gene expression. There is need to improve the structure learning method, valuable features and analysis etc.

DNA chips (Garbarine & Rosen 2008) have a main role in disease diagnosis, drug discovery and gene identification. They used an approach to detect unique gene regions of particular species. This technique named as an information theoretic method exploits genome vocabularies to distinguish between pathogens. This approach is useful only for finding the gene sequences and most distinguished similarities between two organisms. Oligo probes were used to distinguish between two genes. Experiments were conducted to data from Sanger Institute. Currently 32 out of 92 bacterial pathogen sequencing projects are completed. The author selected a pair of genomes to test the algorithm. Results were shown for a 12-mer and 25-mer Oligo pathogen probe set and confirmed the Garbarine method is less likely to cross-hybridize.

Lousado and Moura (2008) developed a software application for large-scale analysis of codon-triplet associations to shed new light onto this problem. This algorithm describes codon-triplet context biases, codon-triplet analysis and identification of alterations to standard genetic code. The method presents an evolutionary understanding of codons within open reading frames (ORF).

Gene-Split (Chang, et al., 2004) is an application that shows codon triplet patterns in genomes and complete sets of ORFs. Generally this application gives an opportunity to study the characteristics of codon and amino acids triplets in any genome for extraction of hidden patterns.

Zheng, et al., (2006) present a technique that integrates the low pass filter and wavelength de-noising method. Conventional techniques use the low pass filter with cheap hardware resulting in degraded de-noising quality. By properly choosing the cut-off frequency and wavelength de-noising frequency, some enhancement can be made for signal to noise ratio and processed signals can be made for requirement of single base pair resolution in DNA sequencing and vector of targeting signal can be decomposed into the orthogonal matrix of wavelength functions. This is an iterative method with levels n and can be conventionally reconstructed by inverse DWT.

Weng, et al., (2006) apply wavelength transform to extract features from the original measurements. They partition the data in subsequent partitions by a hierarchal clustering method, the terahertz spectroscopy of different DNA samples show the wavelength domain analysis aids the clustering process, authors have clustered six DNA samples into two groups, the data has been cleansed before processing, wavelet function utilized the Haar wavelet methods. The signal trend is separated from the original records. The size of clusters may be calculated by the maximum distance between two points within cluster. Another preprocessing step is balancing the data which can achieve normalization of data.

Bilu, et al., (2006) propose an alignment algorithm for NP hard alignment problem of sequences, the authors outperform an alignment procedure by sufficing optimal alignment of predefined sequence segments. They concentrate on a whole sequence rather than letters and estimate running time by restricting the search space of dynamic programming algorithm. Authors take aid from the observation that encoding sequences used in NP hard problems are not necessarily depictions of protein and DNA sequences. Time expedition is calculated by taking advantage of the biological nature of sequences contrary to traditional approaches that offer good computation leading to optimal alignment; more stress is given to the structure of input sequences.

Tuqan and Rushdi (2008) propose an approach for finding the complete periodicity in DNA sequences. The approach is spliced in three channels: firstly, the authors explain the underlying mechanism for period 3 components; secondly, they directly relate the identification of these components for finding nucleotide bias in codon spectrum; thirdly, they completely characterize

the DNA spectrum by a set of numerical sequences. The authors relate the signal processing problem with genomic one through their proposed multi-rate DSP model. The model identifies the essential components involved in the codon biased, distilling the dual nature of the problem. This phenomenon can further help in understanding the biological significance of codon bias. The period 3 component detection works for a kind of genes and may not be suitable for all genetic datasets.

Ma, et al., (2006) has shown the functionality of popular clustering algorithms for analysis of microarray data and concluded that performance of these algorithms can be further increased. Authors are also proposing an evolutionary algorithm for microarray data analysis in which there is no need for calculation of number of clusters in advance. The algorithm was tested with simulation and different datasets. The noise and missing values are a big issue in this regard. The phenomenon is depicted by encoding the entire cluster grouping in a chromosome so that each gene encodes one cluster and each cluster contains the labels of data used in it. Cross over and mutations are performed suitably. The proposed algorithm has been observed to be slow as compared to other prevailing algorithms.

3. THE PROPOSED TECHNIQUE

Our interest mainly lies in finding genome regions that are responsible for protein translation. We have developed a layered architecture shown in the Figure 1 for this analysis that starts from pre-processing of raw data to final translation analysis. For the sake we have used genetic datasets of Chimaera Monstrosa (rabbit fish, NC_003136) and Poly Odontidae (paddle fish, NC_004419) (Anonym).

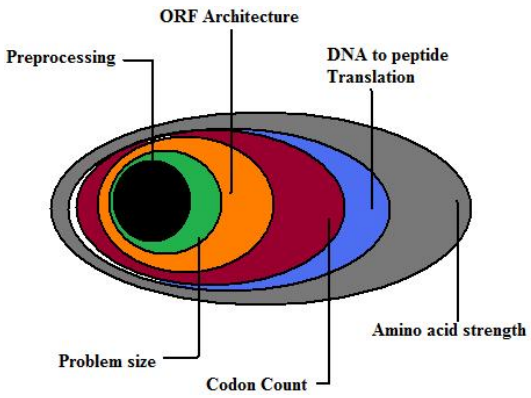

Figure 1 A layered architecture

At pre-processing stage raw data sets are passed through a filter that outputs a more refined form of data which can be further used for actual comparative analysis between species. It is evident from Figure 2 that dataset contains characters other than pure nucleotide bases. These illegal characters are removed by application of a cleansing filter. At first instance it is worth noting that analysis should be made with original data values, any garbage collection may lead to a detrius of results.

Figure 3 depicts that preprocessed data contains only pure nucleotide base pairs without any anomalies. This refined data is later fed into next layer for actual analysis. First we display the ORF in a nucleotide sequence and find the start and stop codon. By using the sequence indice for start and stop, we can extract the subsequences and can determine the codon distribution effectively. The most informative and interesting phenomenon, that whole process is broken into steps and each step fully performs the comparative analysis relevant to DNA to protein

translation.

Figure 2 Dataset before filter application

Figure 3 Pre-processed dataset

3.1. Size of datasets

1. Chimaera Monstrosa contains 18580 nucleotides of Adenine, Guanine, Thymine and Cytosine. Cumulative size of data becomes 37160 bytes arranged in the form of a uni-vector.
2. Poly Odontidae contains 16512 nucleotides of Adenine, Guanine, Thymine and Cytosine. Cumulative size of data becomes 33024 bytes arranged in the form of a uni-vector.

3.2. ORF in nucleotide sequences

It is worth noting that comparative analysis between both species is being done at translation level, so this level is vital in analysis. We split this layer into three more layers to get a better benefit of this layered analysis. In each phase, our interest lies in determining the accurate start and stop position of codons that perform the relative analysis.

3.2.1. ORF primary Frames

At ORF primary frame level, Figure 4 shows that the start position for the first frame is at 7156 and second at 8761. These start positions represent the major translation regions in entire frames. These regions are pure depiction of tri-nucleotide molecules. This process leads towards the extraction of sub-chains that later will be shifted to peptide regions.

Figure 4 ORF of Chimaera Monstrosa in Frame 1

Figure 5 ORF of Poly Odontidae in Frame 1

Likewise we obtain the ORF in the second data set of Poly Odontidae shown in Figure 5. The by entering the start positions we can get stop codons. The start positions of the second dataset Frame 1 are 10798 to 11395, 14641 to 15559. It is clear that there is an evident difference in codon regions for both frames of these species. The corresponding translated regions are so entirely different that we cannot even guess the idea of sub-channel similarity.

3.2.2. ORF secondary Frames

At second level, we intend to find the codon positions for Frame 2 of both species, Figure 6 describes that major ORF start from 2753, 5426 and 10325, this represents that there are series of other regions occupied between the first and second frames that do not contribute to the peptide translation regions.

Figure 6 Frame 2 (Chimaera Monstrosa)

Figure 7 Frame 2 (Poly Odontidae)

Similarly the frame 2 of Poly Odontidae shown in Figure 7 describes its codon position from 11120 to 11465 and 12464 to 12887. This shows a massive difference in datasets at this level. As we move with increasing nucleotide subsequences, we may get larger differences, but this case does not seem to be true for all genetic datasets. This is the reason that phenomenon has been given importance in selection of these particular sets.

3.2.3. ORF Tertiary Frames

Discussing the last frame set in this sequence, we first find the codon composition for these frames, for instance, when we consider frame 3 of Chimaera Monstrosa. Figure 8 shows that major ORF starts from 4019, 11948 and 14328. This massive difference in codon compositions also provide evidence that first translated region lies in the region of some four thousand while second and third regions have jump gaps. This is the variation in translated regions in species.

Figure 8 Frame 3 (Chimaera Monstrosa)

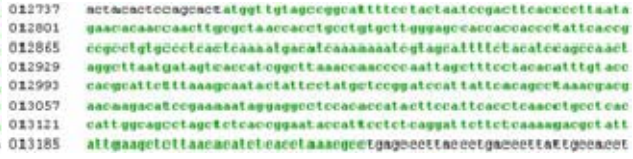

Figure 9 Frame 3 (Poly Odontidae)

In Figure 9, third frame for Poly Odontidae goes from 2796 to 3242, 6315 to 6722 and 12753 to 13217. Figure 8 shows that first 2 codon positions are relative similar while third position again describe a jump gap. Performing comparative analysis at this level, reveals the facts that both genetic data finds a kind of extremity in behavior which makes them relevant at certain codon compositions and different at others.

3.3. Codon Count

The codon count describes the tri-nucleotide behavior of sequences. We need to find the degree of relevancy in terms of strengths of nucleotide bases. For instance, we have selected Frame 1 from codon composition of both species and then we compare the strength.

AAA - 4	AAC - 7	AAG - 0	AAT - 1
ACA - 10	ACC - 2	ACG - 0	ACT - 2
AGA - 1	AGC - 3	AGG - 0	AGT - 1
ATA - 8	ATC - 14	ATG - 2	ATT - 11
CAA - 10	CAC - 7	CAG - 0	CAT - 3
CCA - 3	CCC - 4	CCG - 0	CCT - 6
CGA - 1	CGC - 3	CGG - 0	CGT - 2
CTA - 10	CTC - 8	CTG - 1	CTT - 6
GAA - 13	GAC - 5	GAG - 1	GAT - 6
GCA - 4	GCC - 4	GCG - 0	GCT - 5
GGA - 3	GGC - 1	GGG - 2	GGT - 3
GTA - 7	GTC - 4	GTG - 1	GTT - 3
TAA - 0	TAC - 4	TAG - 0	TAT - 5
TCA - 6	TCC - 3	TCG - 0	TCT - 3
TGA - 4	TGC - 0	TGG - 1	TGT - 2
TTA - 4	TTC - 5	TTG - 1	TTT - 4

Figure 10 Codon count
(Chimaera Monstrosa in Frame 1)

AAA - 8	AAC - 2	AAG - 1	AAT - 11
ACA - 5	ACC - 3	ACG - 3	ACT - 9
AGA - 4	AGC - 11	AGG - 7	AGT - 8
ATA - 2	ATC - 6	ATG - 4	ATT - 6
CAA - 3	CAC - 5	CAG - 0	CAT - 7
CCA - 5	CCC - 5	CCG - 0	CCT - 17
CGA - 2	CGC - 3	CGG - 6	CGT - 2
CTA - 3	CTC - 2	CTG - 6	CTT - 4
GAA - 0	GAC - 2	GAG - 0	GAT - 4
GCA - 0	GCC - 1	GCG - 0	GCT - 0
GGA - 0	GGC - 1	GGG - 2	GGT - 0
GTA - 0	GTC - 1	GTG - 0	GTT - 0
TAA - 1	TAC - 0	TAG - 0	TAT - 8
TCA - 0	TCC - 0	TCG - 0	TCT - 3
TGA - 0	TGC - 8	TGG - 1	TGT - 0
TTA - 3	TTC - 2	TTG - 1	TTT - 1

Figure 11 Codon count
(Poly Odontidae in Frame 1)

Figure 10 represents the codon count for Chimera Monstrosa. Our aim focuses on comparative analysis of codon strength at this stage. For this purpose, we need to calculate the codon count for Poly Odontidae. Figure 11 shows the codon count of the first ORF of the Poly Odontidae. The tri-nucleotide composition of these molecules represents the amino acids. By calculating these combinations, we can get the volume of the specific amino acids. Some of the amino acids for these codons ATA, CTA, ACC and ATC are as follows respectively.

Ile: soleucine, Leu: Leucine, Thr: Threonine, and Ile: Isoleucine

3.4. Amino Acid Conversion and composition

In this section, we try to obtain relative amino acid composition that will give us the characteristic profile of the protein. Once we get the Open Reading Frame in a genetic data, we can convert it into an amino sequence to bring in its acid composition. Consider first dataset Chimaera Monstrosa at first frame level for conversion from nucleotide sequence to an amino acid sequence. This conversion is shown in Figure 12.

```
MAHPSQLGFQDAASPVMEELLHFHDHTLMIVFLISTLILYIITV
MVTTKLTNKFILDSQGIEIIWTILPAIILISIALPSLRILYLMD
EIINPHLTIKAIGHQWYWSYEYTDYENLEFDSYMVQTQDLNPGQ
FRLLETDHRMIIPMESPIRILVSADDVLHSWAVPALGVKMDAVP
GRLNQTAFLVTRPGVYYGQCSEICGANHSFMPIVVEAVPLQHFE
NWSLLTLEENSLS
```

Figure 12 Amino Acid sequence
(Chimaera Monstrosa)

```
MKPNRTSQRHLLPILHTSRVPPPTSCPTDPTKNLPINTNYPMHT
TCPPMHLSRNLMSSLLNCLPSKNTPLRSPPLITKSTRSPSCIYS
PSCRTTKTRLWHDTNDYYTACIQKPCMPIYYPGPMHYHDRINLS
TTDSKIPNCLLFSKPHGTSSSGNPHPNPLLHRSYHFNDRTRISI
FRIILPSKHQL
```

Figure 13 Amino Acid sequence
(Poly Odontidae)

From Figures 12 and 13, both genetic datasets have strongly different translated composition even at the primary frame level. This provides us evidence to strengthen the idea that both species will behave differently in other frames too.

3.5. Strength of amino acid in the Protein sequence

At last phase of this comparative analysis, we need to find the relevant strength of peptide pairs in protein sequences (resulted as a translation from DNA to protein). Figure 14 shows the strength of amino acid in Chimera Monstrosa. Now we will determine the atomic decomposition and molecular weight of the protein:

C: 1220, H: 1886, N: 298, O: 341, S: 12, Molecular weight is $2.6569e^{+004}$

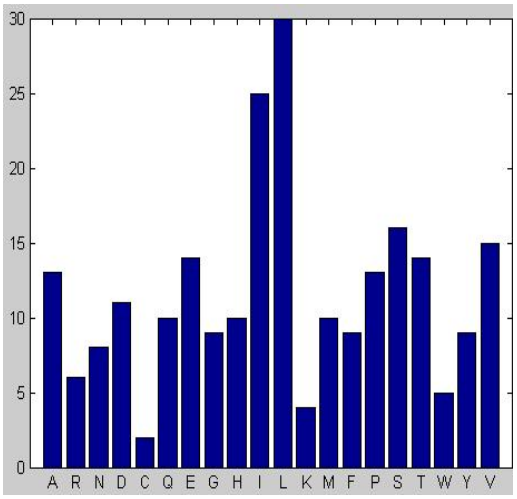

Figure 14 Strength of amino acid (Chimaera Monstrosa)

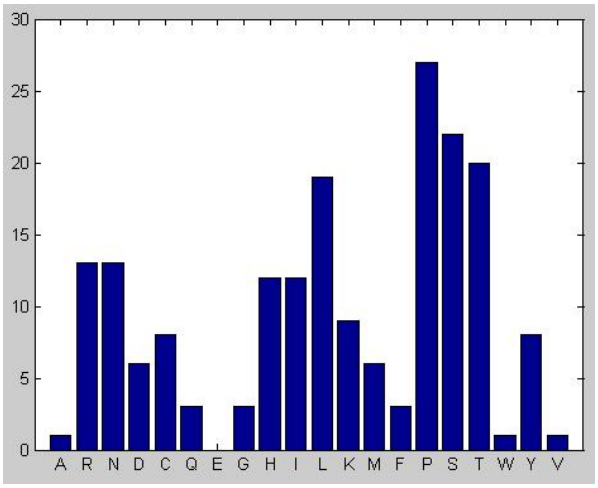

Figure 15 Strength of amino acid (Poly Odontidae)

The strength of amino acid in protein sequence of the Poly Odontidae is depicted in Figure 15. Similarly, the atomic decomposition and molecular weight of the protein are

C: 940, H: 1488, N: 276, O: 266, S: 14, and Molecular weight is $2.1360e^{+004}$

Table 1 Amino acid sequence comparison

Amino acid	Chimaera Monstrosa	Poly Odontidae
C	1220	940
H	1886	1488
N	298	276
O	341	266
S	12	14

The comparison of amino acid sequences of both species obtained from the primary codon translation is shown in Table 1. The corresponding molecular weights are $2.6569e^{+004}$ and $2.1360e^{+004}$ for Chimaera Monstrosa and Poly Odontidae respectively. These results clearly describe the phenomenon that despite both species from same class differ greatly in patterns of ORF. Their codon count and numerical measures of amino acid and molecular weights make them different in behavior, appearance, habits, characters and living.

4. CONCLUSION

An Open Reading Frame (ORF) contains a start codon region. This subsequent region contains pairs of nucleotides in length multiple of 3 and end with a stop codon. This paper describes the phase wise comparative analysis of two genetic data of species Chimaera Monstrosa and Poly Odontidae. It represents an integrated approach composed of step by step processes to elaborate the results effectively. The process gives more stress on peptide translation using Open Reading Frame concept and data refining methodology. At the end we look for all outcomes that make this effort optimal by performing a sensitive analysis of DNA to protein conversion. Variations at each step were observed even the data classes remained same.

5. ACKNOWLEDGEMENTS

This work was partially supported by Research Center, College of Computer and Information Sciences, King Saud University, Riyadh, Kingdom of Saudi Arabia.

6. REFERENCES

Anonym, http://www.ncbi.nlm.nih.gov.

Bartkowiak, 2008. Nonlinear dimensionality reduction by isomap and MLEdim as applied to Amino-Acid distribution in yeast ORFs, *Computer Information Systems and Industrial Management Application*, 2008, pp.183-188.

Bilu Y., Agarwal P.K. & Kolodny R., 2006. Faster algorithms for optimal multiple sequence alignment based on pairwise comparisons, *IEEE/ACM Transactions on Computational Biology and Bioinformatics 2006*, Volume 3, Issue 4, pp.408-422.

Chang P.H.-M., Soo V.-W., Chen T.-Y., Lai W.-S., Su S.-C. & Huang Y.-L., 2004. Automating the determination of open reading frames in genomic sequences using the Web service techniques - a case study using SARS coronavirus, *Fourth IEEE Symposium on Bioinformatics and Bioengineering 2004*, pp.451-458.

Garbarine E. & Rosen G., 2008. An information theoretic method of microarray probe design for genome classification, *30th Annual International Conference of the Engineering in Medicine and Biology Society 2008*, pp.3779-3782.

Gowda T., Leshner S., Vrudhula S. & Kim S., 2007. Threshold logic gene regulatory networks, *International Workshop on Genomic Signal Processing and Statistics 2007*, pp.1-4, ISBN: 978-1-4244-0998-3.

Gupta R., Mittal A., Singh K., Bajpai P., Suraj, & Prakash, 2007. A time series approach for identification of Exons and Introns, *10th International Conference on Information Technology 2007*, pp.91-93.

Hireche N., Langlois J.M.P. & Nicolescu G., 2006. Survey of biological high performance computing: Algorithms, Implementations and Outlook Research, *Canadian Conference on Electrical and Computer Engineering 2006*, pp.1926-1929.

Kumar R., Kumar A. & Agarwa S., 2007. A distributed bioinformatics computing system for analysis of DNA sequences, *In: IEEE proceedings of Southeast Conference 2007*, pp.358-363.

Kurata K.-i., Breton V. & Nakamura H., 2003. A method to find unique sequences on distributed genomic databases, *IEEE/ACM International Symposium on Cluster Computing and the Grid 2003*, 3rd Volume, pp.62-69.

Li A., Wang T., Zhou Y., Wang M.-h. & Feng H.-q., 2003. An efficient structure learning method in gene prediction, *In: Proceedings of the International Conference on Neural Networks and Signal Processing 2003*, Volume 1, pp.567-570.

Lousado J. & Moura R.G., 2008. Exploiting codon-triplets association for genome primary structure analysis, *International Conference on Bio-computation, Bioinformatics, and Biomedical Technologies 2008*, pp.155-158.

Ma P. & Chan K.,C.C., 2003. Discovering clusters in gene expression data using evolutionary approach, *15th IEEE International Conference on Tools with Artificial Intelligence 2003*, pp.459-466.

Ma P.C.H., Chan K.C.C., Yao X. & Chiu D.K.Y., 2006. An evolutionary clustering algorithm for gene expression microarray data analysis, *IEEE Transactions on Evolutionary Computation 2006*, Volume 10, Issue 3, pp.296-314.

Miranker D., 2008. Evolving models of biological sequence similarity, *First International Workshop on 2008*, pp.3-9.

Tuqan J. & Rushdi A., 2008. A DSP approach for finding the codon bias in DNA sequences, *IEEE Journal of Selected Topics in Signal Processing 2008*, Volume 2, pp.343-356.

Weng B., Xuan G., Kolodzey J. & Barner K.E., 2006. Discriminating DNA sequences from terahertz spectroscopy - A wavelet domain analysis, *In: Proceedings of the IEEE 32nd Annual Northeast Bioengineering Conference 2006*, pp.211-212.

Zheng H., Shi Y., Wang J., Wang L. & Lu Z., 2006. The analysis on the signals denoising and single base pair resolution of DNA sequencing, *International Symposium on Biophotonics, Nanophotonics and Metamaterials 2006*, pp.118-121.