Statistical and Alignment Based Methods for Comparison of Non-Coding DNA Sequences

Kouser and Lalitha Rangarajan*

DoS in Computer Science, University of Mysore, Mysore, Karnataka, India.
*lalithar@compsci.uni-mysore.ac.in

Abstract. Increase in the amount of biological sequence data being generated is due to the advancement in the sequencing technologies. This has led to an upsurge in the need for methods/techniques that analyze these sequences to make biological inferences. Promoter sequences are one such biological sequences that play an important role in the process of gene regulation. These promoter sequences are made up of motifs arranged/configured in a particular order which is very important for its functionality. In this work, we propose a few techniques that capture the information regarding the arrangement of these motifs. Further this captured information help in the analysis of the promoter sequences. The methods use the motif features obtained from the Position Specific Motif Matrices (PSMMs) of the promoter sequences and then give a dissimilarity measure. Alignment-free as well as alignment based methods are proposed. The experimental results show that these methods are quite successful in bringing out the similarity/dissimilarity existing between given promoters/promoter sets. The alignment based method also helps in the process of inferring homology and in the phylogenetic analysis. In general, all the methods might help biologists in applications like drug target identification, biological pathway analysis and gene expression analysis.

Keywords: Alignment free, Alignment score, Multiple Sequence Alignment, Promoter sequences, Dissimilarity matrix, Sequence comparison.

1 Introduction

With the rapid growth in the generation of large wealth of biological sequence data, making sense of this huge data is a challenging task. There is an immediate need for good analyzing methodologies. These methods/techniques help in the process of extracting meaningful information from the sequences for better understanding of biomedical mechanisms [1] and to attempt to solve some specific biological problems. Some recent efforts are focused to identify functions of these non-coding DNA sequences which have some important function in gene expression analysis, tissue development, phylogenetic analysis which serves as the motivation to this kind of research work [2].

The analysis of these biological sequences consists of two major steps. Initial step is alignment which is later followed by analysis. Most of the existing multiple sequence alignment (MSA) works well on coding regions of the genes or the proteins.

The MSA algorithms are built using the principles of dynamic programming (Smith-Waterman and Needleman-Wunsch algorithms) [3], hidden markov model [4] and scoring matrices. There are several tools available that work on these concepts to name a few SW-align, Clustal W [3], MUSCLE [5], K align [6], D align [7], T-Coffee [8], M-Coffee [9]. There is a plenty of work on alignment free sequence analysis methods such as L-words frequency [10], spaced-word frequency methods [11], relative frequencies of dual nucleotides [12] etc. Many methods relating to k-mer/word frequency, substring, information theory and graphical representation are also available. Some softwares available for the methods described above include d2Tools [13], AGP [14], Alfy [15] and WNV typer [16].

In this work, we propose methods to analyze promoter sequences which are made up of repetitive patterns called transcription factor binding sites (TFBS)/motifs. The promoter sequences are a part of the non-coding DNA anywhere before the start codon of a gene and the stop codon of the previous gene, which plays an important role in the process of gene expression. The data set is obtained from the NCBI database and the PSMM are created using the TFSEARCH tool [17].

2 Materials and Methods

In this section, we describe the working details of our model and the proposed algorithms. Five frequency based methods proposed and one lacunarity algorithm which are all alignment free sequence comparison methods. Also, we propose one alignment based sequence analysis method which makes use of the motif pair feature. There are two datasets used, dataset 1 consisting of promoters of the Pyruvate kinase gene of different organisms and dataset 2 consisting of all the enzymes of the Glycolysis pathway of the organism human. The overall schema and flow of the method is as described in Figure 1. The detailed description of obtaining the PSMM for a promoter is presented in [18], [19]. Once the PSMM is obtained, they are given as input to the various methods briefly described below.

2.1 Frequency Based Alignment Free Methods

The frequency distribution gives a measure considering the occurrences of motifs in promoters in various ways. The detailed description of these frequency based methods is presented [19], [20]. The output of these methods is a dissimilarity matrix.

2.2 Lacunarity Based Method

The lacunarity algorithm is a multi scaled method that uses box counting across different scales to obtain the lacunarity values across different scales and then the Euclidian distance is used to get the dissimilarity between two promoters. This value then goes as an entry into the dissimilarity matrix. The algorithm can be found in detail in [18].

2.3 Alignment Based Method

This method extracts the motif pair based features from the binarized PSMM [21]. Later, we compare the count of motif pairs between the promoter sequences to find the similarity. The output of alignment is an alignment score.

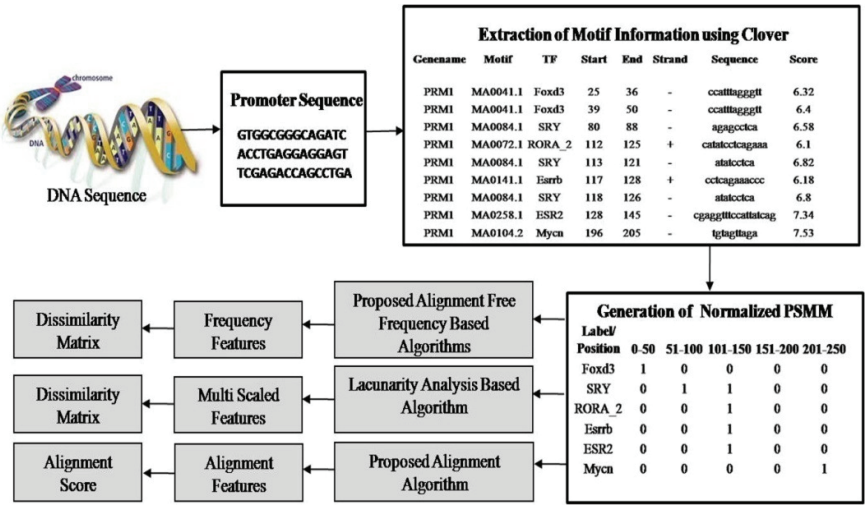

Fig. 1. Overall schema and flow of the methods.

3 Results and Discussion

All the methods described are quite successful in bringing out the similarity between organisms that are known to belong to the same family. The dissimilarity is very low for similar organisms and high otherwise. The results of the frequency based method and the alignment based methods on dataset 1 are presented in Tables 1 and 2. Also, the result of lacunarity analysis is given in Table 3. Results of the frequency based method and the alignment based methods on dataset 2 are presented in Tables 4 and 5. However some cases where some unusual results were obtained are highlighted in the respective dissimilarity matrices.

Table 1. Dissimilarity matrix of frequency based method on dataset 1. (Organisms: 1. Homo sapien (human), 2. Gorilla gorilla (gorilla), 3.Macaca mulatta (rhesus monkey), 4. Bos taurus (cattle), 5.Felis catus (cat), 6.Pan troglodytes (chimpanzee) 7. Canis lupus (dog), 8.Rattus norvegicus (rat), 9.Drosophila melanogaster (fruit fly) and 10. Pseudomonas aeruginosa (bacterium)).

	Human	Gorilla	Monkey	Cattle	Cat	Chimpanzee	Dog	Rat	Fruit Fly	Bacteria
Human	0	299	1629	903	1411	1378	1439	451	1531	1346
Gorilla	299	0	890	256	568	581	604	266	796	587
Monkey	1629	890	0	400	500	283	488	1608	306	629
Cattle	903	256	400	0	300	213	256	820	256	369
Cat	1411	568	500	300	0	267	198	1302	480	125
Chimpanzee	1378	581	283	213	267	0	169	1263	101	298
Dog	1439	604	488	256	198	169	0	1310	262	173
Rat	451	266	1608	820	1302	1263	1310	0	1606	1275
Fruit fly	1531	796	306	256	480	101	262	1606	0	497
Bacteria	1346	587	629	369	125	298	173	1275	497	0

Table 2. Alignment scores of enzyme pyruvate kinase in different organisms. (Organisms: 1. Homo sapien (human), 2. Gorilla gorilla (gorilla), 3.Macaca mulatta (rhesus monkey), 4. Bos taurus (cattle), 5.Felis catus (cat), 6. Drosophila melanogaster (fruit fly), 7. Pseudomonas aeruginosa (bacterium), 8. Pan troglodytes (chimpanzee), 9. Canis lupus (dog) and 10.Rattus norvegicus (rat)).

	Human	Gorilla	Monkey	Cattle	Cat	Fruit fly	Bacteria	Chimpanzee	Dog	Rat
Human	125	117	1	7	0	2	1	32	5	4
Gorilla	117	173	1	11	0	2	1	45	5	4
Monkey	1	1	111	10	1	0	3	0	9	8
Cattle	7	11	10	146	5	2	2	3	6	4
Cat	0	0	1	5	63	2	2	0	6	2
Fruit fly	2	2	0	2	2	201	1	1	1	7
Bacteria	1	1	3	2	2	1	67	1	1	0
Chimpanzee	32	45	0	3	0	1	1	118	3	1
Dog	5	5	9	6	6	1	1	3	142	1
Rat	4	4	8	4	2	7	0	1	1	119

Table 3. Lacunarity scores of enzyme pyruvate kinase in different organisms. (Organisms: 1. Homo sapien (human) , 2. Gorilla gorilla (gorilla) , 3.Macaca mulatta (rhesus monkey), 4. Bos taurus (cattle), 5.Pan troglodytes (chimpanzee), 6. Canis lupus (dog), 7.Rattus norvegicus (rat), 8.Drosophila melanogaster (fruit fly) and 9. Pseudomonas aeruginosa (bacterium)).

	Human	Gorilla	Monkey	Cattle	Chimpanzee	Dog	Rat	Fruit fly	Bacteria
Human	0	0.02	0.37	0.31	0.12	0.47	0.9	0.09	2.27
Gorilla	0.02	0	0.35	0.3	0.1	0.45	0.88	0.07	2.25
Monkey	0.37	0.35	0	0.05	0.25	0.1	0.54	0.28	1.9
Cattle	0.31	0.3	0.05	0	0.19	0.15	0.59	0.23	1.96
Chimpanzee	0.12	0.1	0.25	0.19	0	0.35	0.78	0.03	2.15
Dog	0.47	0.45	0.1	0.15	0.35	0	0.43	0.38	1.8
Rat	0.9	0.88	0.54	0.59	0.78	0.43	0	0.81	1.37
Fruit fly	0.09	0.07	0.28	0.23	0.03	0.38	0.81	0	2.18
Bacteria	2.27	2.25	1.9	1.96	2.15	1.8	1.37	2.18	0

The features obtained from the proposed alignment free methods are not very useful in the study of phylogeny. However, these methods are faster. The lacunarity features obtained at different scales individually or together seem to be more effective in phylogeny analysis when compared to the frequency based methods. They perform poor when used for gene expression differentiation. The motif pair based alignment features seem to be performing best when one is interested in performing phylogenetic analysis. The proposed alignment does not insert gaps preserving the information regarding the position of motifs, which is important in gene expression studies.

Table 4. Dissimilarity matrix of frequency based method on dataset 2. (Enzymes: 1. HK1 hexokinase 1, 2. GPI glucose-6-phosphate isomerase, 3. PFKM phosphofructokinase, 4. ALDOA aldolase A, fructose-bisphosphate, 5. TPI1 triosephosphate isomerase 1, 6. GAPDH glyceraldehyde-3-phosphate dehydrogenase, 7. PGK1 phosphoglycerate kinase 1, 8. PGAM2 phosphoglycerate mutase 2 (muscle), 9. ENO1 enolase 1, (alpha) and 10.PKM pyruvate kinase (muscle)).

	HK1	GPI	PFKM	ALDOA	TPI1	GAPDH	PGK1	PGAM2	ENO1	PKM
HK1	0	449	256	328	499	363	381	420	217	303
GPI	449	0	470	261	564	495	222	212	315	218
PFKM	256	470	0	344	474	343	348	314	151	307
ALDOA	328	261	344	0	196	223	108	138	266	111
TPI1	499	564	474	196	0	340	394	214	446	290
GAPDH	363	495	343	223	340	0	323	339	414	316
PGK1	381	222	348	108	394	323	0	128	242	86
PGAM2	420	212	314	138	214	339	128	0	196	109
ENO1	217	315	151	266	446	414	242	196	0	174
PKM	303	218	307	111	290	316	86	109	174	0

Table 5. Alignment scores of enzymes in glycolysis pathway of Homo sapien. (Enzymes: 1. HK1 hexokinase 1, 2. GPI glucose-6-phosphate isomerase, 3. PFKM phosphofructokinase, 4. ALDOA aldolase A, fructose-bisphosphate, 5. TPI1 triosephosphate isomerase 1, 6.GAPDH glyceraldehyde-3-phosphate dehydrogenase, 7. PGK1 phosphoglycerate kinase 1, 8. PGAM2 phosphoglycerate mutase 2 (muscle), 9. ENO1 enolase 1, (alpha) and 10. PKM pyruvate kinase (muscle)).

	HK1	GPI	PFKM	ALDOA	TPI1	GAPDH	PGK1	PGAM2	ENO1	PKM
HK1	313	13	11	9	3	10	3	1	4	2
GPI	13	173	10	8	1	9	5	5	9	2
PFKM	11	10	194	5	1	8	3	9	11	3
ALDOA	9	8	5	122	10	9	1	6	4	2
TPI1	3	1	1	10	61	11	4	5	3	0
GAPDH	10	9	8	9	11	78	3	6	4	2
PGK1	3	5	3	1	4	3	92	1	0	0
PGAM2	1	5	9	6	5	6	1	124	8	5
ENO1	4	9	11	4	3	4	0	8	223	0
PKM	2	2	3	2	0	2	0	5	0	97

4 Conclusion

The results obtained from different proposed methods on the two datasets show their success in bringing out the similarity/dissimilarity existing between given promoters. The alignment based method also helps in the process of inferring homology and in the phylogenetic analysis. We notice that results are similar from both the alignment free and alignment based methods showing that both methods are equally useful. In general, all the methods might aid biologists in the process of drug target identification,

biological pathway analysis and gene expression analysis. As a future work, we are contemplating on selection of subset of important motifs from the promoters. This could result in substantial reduction in PSMM size and will have direct influence on running time of algorithms. Perhaps, dissimilarity measure may also improve since less important motifs are ignored.

References

1. Qin, Y., Yalamanchili, H. K., Qin, J., Yan, B, Wang, J.: The Current Status and Challenges in Computational Analysis of Genomic Big Data. Big Data Research—Special Issue on Computation, Business, and Health Science. Vol.2(1) (2015) 12–18.
2. Ghiurcuta, C.G.: Models and Algorithms for Noncoding Genes. Edic Research Proposal. (2009).
3. Renxiang, Y., Dong, X., Jianyi, Y., Sara, W., Yang, Z.: A comparative assessment and analysis of 20 representative sequence alignment methods for protein structure prediction. Scientific Reports 3. (2013).
4. Siepel, A., Haussler, D.: Combining phylogenetic and hidden Markov models in biosequence analysis. J Comput Biol, Vol.11(2–3) (2004) 413–428.
5. blast.ncbi.nlm.nih.gov
6. Thompson, D., Higgins, D.G., Gibson,T. J.: CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. Vol 22(22) (1994) 4673–4680.
7. Lassmann, T., Sonnhammer, E.L.: Kalign—an accurate and fast multiple sequence alignment algorithm. BMC Bioinformatics. (2005).
8. Edgar, R.C.: MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. Vol 32(5) (2004) 1792–1797.
9. Notredame, C., Higgins, D.G., Heringa, J.: T-Coffee: A novel method for fast and accurate multiple sequence alignment. J Mol Biol. Vol 302(1) (2000) 205–217.
10. Soares, I., Goios, A., Amorim,A.: Sequence Comparison Alignment-Free Approach Based on Suffix Tree and L-Words Frequency. The Scientific World Journal.Vol. 2012 (2012).
11. Leimeister, C.A., Boden, M., Horwege, S., Linder, S., Morgenstern, B.: Fast Alignment-Free Sequence Comparison Using Spaced-Word Frequencies. Bioinformatics. Vol.30(14) (2014) 1991–1999.
12. Luo, J., Li, R., Zeng,Q.: A Novel Method for Sequence Similarity Analysis Based on the Relative Frequency of Dual Nucleotides. MATCH Commun. Math. Comput. Chem. Vol. 59 (2008) 653–659.
13. http://omictools.com/d2-tools-s3305.html
14. Cheng, J., Cao, F., Liu, Z.: AGP: A multimethods web server for alignment-free genome phylogeny. Mol Biol Evol. (2013) 1032–1037.
15. Domazet, L., Haubold, M.: Alignment-free detection of local similarity among viral and bacterial genomes. Bioinformatics. Vol.27(11) (2011) 1466–1472.
16. Pandurang, K., Nilesh, H., Mohan, K., Urmila, K.K.: WNV Typer: A server for genotyping of West Nile viruses using an alignment-free method based on a return time distribution. Journal of Virological Methods. Vol. 198 (2014) 41–55.
17. http://www.cbrc.jp/research/db/TFSEARCH.html

18. Kouser, Lalitha, R.: Similarity Analysis of Position Specific Motif Matrices using Lacunarity for Promoter Sequences. ICONIAAC'14.Article. 37. ACM New York. NY. USA (2014).
19. Kouser, Lalitha, R., Darshan, S.C., Acharya, K.K., Emin, M.A.: Alignment Free Frequency Based Distance Measures for Promoter Sequence Comparison. IWBBIO 2015. Vol.2 (2015) 183–193.
20. Kouser, Lalitha, R., Darshan, S.C., Acharya, K.K.: Frequency Distribution and lacunarity based Promoter Sequence Analysis (FDL-PSA). Communicated.
21. Kouser, Lalitha, R.: Promoter sequence analysis through ungapped multiple sequence alignment of motif pairs. ICACCI 2015. (2015).