New Distance Measure for Sequence Comparison using Cumulative Frequency Distribution

Meera.A
BMS College of Engineering
Bangalore
Bull temple Road
Basavanagudi

Dr Lalitha Rangarajan
Department of Computer Science
manasagangotri
University of Mysore, Mysore

Shilpa .N
Department of Computer Science
manasagangotri
University of Mysore, Mysore

ABSTRACT

Comparison of two promoter sequences is proposed in this paper. Motifs are extracted from promoter sequences using available software tool 'TF SEARCH'. The promoter sequences are compared using cumulative frequency distribution of motifs. For experimental study, promoter sequences of different mammals of the enzyme Citrate synthase of TCA (kreb) cycle in CMP (Central Metabolic Pathway) are considered. Results reveal high similarity in motif sequences of different organisms in the same chromosome. Also some amount of similarity is present among motif sequences of different chromosomes of the same organism.

General Terms

Pattern Recognition, bioinformatics

Keywords

Cumulative frequency distribution, Distance measure, Pattern matching, Promoter sequence, Regression line, Transcription Factors (TFs), Transcription factor binding sites (TFBS)

1. INTRODUCTION

Sequence comparison is a useful computational technique in molecular biology. Biologists rely heavily on comparison of DNA sequences for understanding of many biological problems. The question that has still remained as an open issue is how to extract information and knowledge from the genetic data available. The evolutionary information in organisms is carried in its genes. Genes are sequences of the polymer DNA which, for our purposes, can be viewed as strings over the alphabet {A,C,G,T}, where each of the four characters corresponds to one of the nucleotide bases that makes up DNA. The expression of coding region is exclusively dependent on promoter region. Conserved region in non-coding sequence are called motifs [5]. These conserved sequences in the region upstream of a gene are as important as coding region. Comparison of promoter sequences may give insight into finding evolutionary distance and gene order [12].

In promoters, primary sequence comparisons, however, have limitations. In the process of aligning nucleotides, structure of motifs may be disturbed. Although similar sequences do tend to play similar functions, the functionality is determined by regulated region. Often similar functions are encoded in higher order sequence elements such as, structural motifs in amino acid sequences and the relation between these and the underlying primary sequence may not be univocal. As a result, similar functions are frequently encoded by diverse sequences [4].

The information for the control of the initiation of the RNA synthesis by the RNA polymerase II is mostly contained in the gene promoter, a region usually 200 to 2,000 nucleotides long upstream of the transcription start site (TSS) of the gene. Transcription factors (TFs) interact in these regions with sequence-specific elements or motifs (the TF binding sites (TFBSs)). TFBSs are typically 5–8 nucleotides long, and one promoter region usually contains many of them to harbor different TFs [10]. However, TFBSs associated to the same TF are known to tolerate sequence substitutions without losing functionality, and are often not conserved. Consequently, promoter regions of genes with similar expression patterns may not show sequence similarity, even though they may be regulated by similar configurations of TFs.

A large amount of work has been carried out in aligning coding regions of DNA sequences for finding homology between different species. Local pairwise alignment methods such as Smith- Waterman [16], BLAST [2], BLASTZ [15], SSAHA [14], and BLAT [11] are able to pinpoint locations of rearrangements between two sequences, and are suitable for aligning nucleotide sequences.

Some useful tools that align promoter sequences are CONREAL [8], Monkey [1], AVID [3]. All these methods aim at identifying TFBS. Despite the recent progress due to the development of techniques based on phylogenetic footprinting [13], lack of nucleotide sequence conservation between functionally related promoter regions may partially explain the still limited success of current available computational methods for promoter characterization [9] and [10]. One of the major challenges facing biologists is to understand the varied and complex mechanisms governing the regulation of gene expression. Sequence conservation across different species is an important indicator of functionality.

Pattern matching has been applied with varying degrees of success on areas as diverse as voice, image and optical character recognition. Nucleotide sequence alignment is also pattern matching and is the basis for DNA sequence analysis that leads to some important bioinformatics activities such as identification of homologous gene, data mining etc. Data mining relies on heuristic pattern matching to locate patterns using a variety of technologies, from simple keyword matching to rule-based expert systems and artificial neural networks. Pattern sequences may be of varying lengths (as small as 5000 in length to as big as 13 x 10^{10} in length).

Pattern matching can be applied to the full set of upstream sequences in a genome, in order to predict genes possibly regulated by a given transcription factor. It should be noted that the simple presence of a motif in a given upstream region is generally not sufficient to predict regulation. Indeed, given the short size of the motifs and the large size of the genomes, hundreds, or even thousands of matches could be returned by chance alone. Predictions can be improved by detecting multiple binding sites, either for the same transcription factor, or for combinations of several different transcription factors [7].

In the present work, we take into account the frequency of occurrences of motifs while comparing promoter sequences. Frequency comparison can be misleading. Hence we have extended the work to compare the respective CDFs (Cumulative Distribution Frequency). A new distance measure, to measure the dissimilarity between CDFs is devised.

2. METHODOLOGY

Motifs are extracted using 'TF SEARCH' tool. Extracted motifs are counted and histogram computed separately for the promoter sequences.

The figures 1 and 2 show the histogram of motifs in promoter 1 and 2. The frequencies of motif M33 in promoters 1 & 2 are 5 & 3. Similarly the figures 1 & 2 give the occurrences of all motifs in the motif sequences of promoters 1& 2.

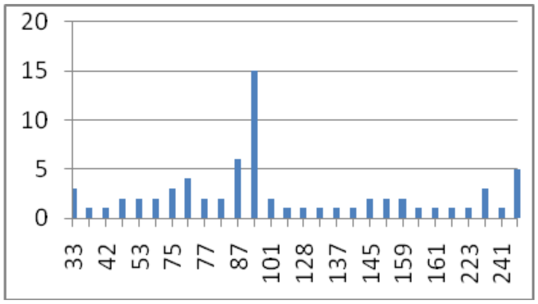

Fig 1: Histogram of motifs in promoter 1

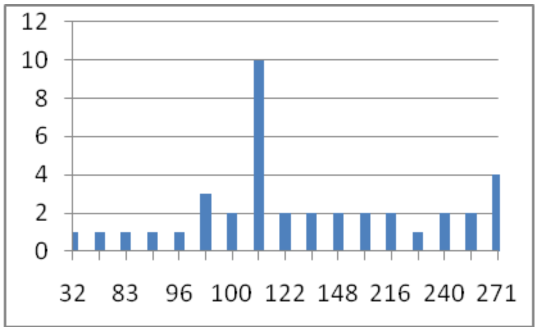

Fig 2: Histogram of motifs in promoter 2

The dissimilarity is computed using the formula

$$D = \sum_{i}[fS_1(i) – fS_2(i)]^2 \text{ ....} \qquad (1)$$

where $fS_1(i)$ and $fS_2(i)$ are the frequencies of motif i in motif sequences S_1 & S_2.For example, the dissimilarity computed from the histograms in figures 1 & 2 is given by $D = (0-1)^2 + (3-0)^2 + \ldots\ldots\ldots\ldots\ldots\ldots + (5-4)^2 = 136$ The dissimilarity measure above depends on number of motifs in the sequences considered. This is then normalized as

$$d = [\, D / \sum_{i} fS_1(i) + \sum_{i} fS_2(i)] * 100 \text{ ....} \qquad (2)$$

For the above example $\sum fS_1(i) = 69$ and $\sum FS_2(i) = 39$

Therefore $d = 136/(69+39)$

$d = 1.26$

Dissimilarity measure= (1/ 1.26) =0.78

The previous dissimilarity measure has some drawbacks. The calculated distance doesn't reflect the differences present in each motif count. For example consider the hypothetical distribution of three promoter sequences:

Table 1: frequency of motifs in 3 promoter sequences

frequency	motif1	motif 2	motif 3	motif 4
Promoter1	10	15	5	6
Promoter2	13	13	9	16
Promoter3	0	18	7	11

The distance when computed will give the following dissimilarity matrix:

$$\begin{array}{c|ccc} & P1 & P2 & P3 \\ \hline P1 & 0 & 806.25 & 862.5 \\ P2 & 0 & 0 & 806.25 \\ P3 & 0 & 0 & 0 \end{array}$$

Observe that distances between P1, P2 and P1, P3 are identical (806.25). There is a large difference in the occurrences of motif 1 in sequences 1 and 3, where as there is a large difference in frequencies of motif 4 in sequence 1and 2.

In this method, c.d.f is identified and dissimilarity between c.d.f's identified. The c.d.f curve is examined to measure dissimilarity.

For the same example above the c.d.f of promoter sequences 1, 2 and 3 are shown in the below graph. The differences that are not evident with just histogram are apparent with their c.d.fs.

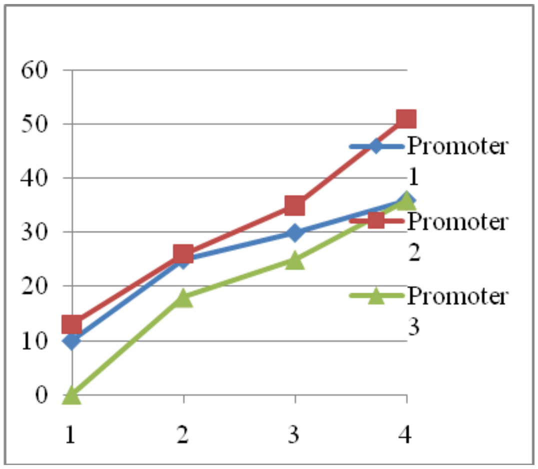

Fig 3: c.d.f of motifs in three promoters.

.

The c.d.f curve is difficult to analyze. So an approximation of lines is done. The approximation will introduce a lot of error if all points in the distribution function are regressed into a single line. So lines are constructed for every successive ten points in the c.d.f graphs. So each c.d.f is reduced to set of lines.

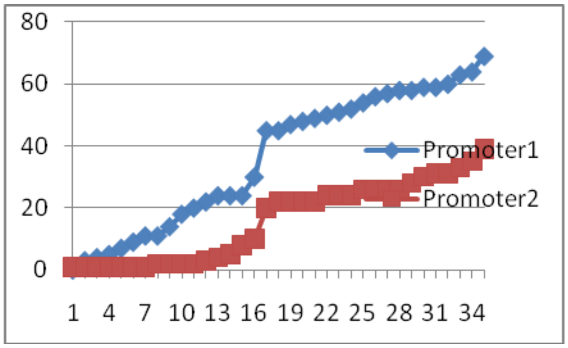

Fig 4: c.d.f of motifs of figures 1and 2.

Figure 4 shows the c.d.f of the histograms (corresponding to two motif sequences) in figures 1& 2. It may be observed that motifs are relabeled in the process of computing c.d.fs

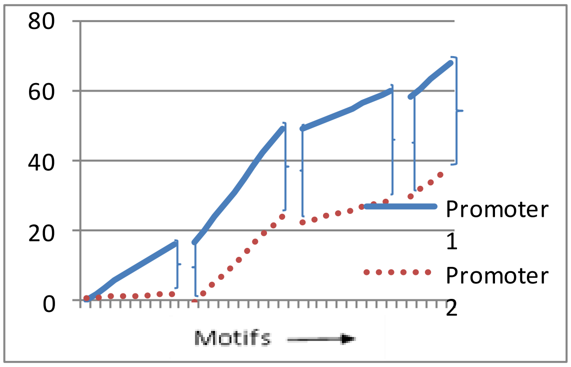

Fig 5: Regression line segments of c.d.fs in figure 4

Figure 5 shows the regression lines fitted for every 10 points in c.d.f curves and also the computation of dissimilarity. The dissimilarity between promoters 1 and 2 is computed as maximum of the difference in y-values at the extreme x-values of each line segment. In figure 5 we have highlighted the difference in y-values and the maximum of these is indicated using bracket.

3. RESULT AND DISCUSSION

In this paper we have attempted to compare promoter sequences of different mammals. For experimental study, promoter sequences of different mammals of the enzyme Citrate synthase of TCA (kreb) cycle in CMP (Central Metabolic Pathway) are considered. Table below shows the computation of dissimilarity measure between promoters of different mammals of enzyme citrate synthase extracted from NCBI database. The motifs are extracted using ‘TF SEARCH’ tool. cdf for each motif sequence is constructed and compared.

Table 1: Dissimilarity measure between promoters of different mammals.

	Rat7	**Can10**	**Pan12**	**Pan3**	**Hs6**	**Hs12**	**Bos5**	**Sus5**	**Mac11**	**Hs19**	**Pan19**	**Bos10**
Rat7	0	203.5	332.7	342.2	516.5	300.9	363.5	210.2	387.8	299.4	304.1	346.2
Can10	203.5	0	169.2	171.6	355.6	135.7	201.4	30.7	218.5	139.8	143.2	181.8
Pan12	332.7	169.2	0	64.8	194.0	35.7	45.7	131.4	64.9	42.5	37.5	64.5
Pan3	342.2	171.6	64.8	0	181.7	78.3	54.8	143.8	61.7	48.6	49.1	35.0
Hs6	516.5	355.6	194.0	181.7	0	223.4	151.0	317.8	130.3	219.6	215.2	171.4
Hs12	300.9	135.7	35.7	78.3	223.4	0	76.2	99.9	93.4	32.5	33.5	81.6
Bos5	363.5	201.4	45.7	54.8	151.0	76.2	0	169.4	63.5	71.2	71.6	44.4
Sus5	210.2	30.7	131.4	143.8	317.8	99.9	169.4	0	187.0	100.3	104.8	151.5
Mac11	387.8	218.5	64.9	61.7	130.3	93.4	63.5	187.0	0	91.6	87.9	42.4
Hs19	299.4	139.8	42.5	48.6	219.6	32.5	71.2	100.3	91.6	0	5.6	50.6
Pan19	304.1	143.2	37.5	49.1	215.2	33.5	71.6	104.8	87.9	5.6	0	46.3
Bos10	346.2	181.8	64.5	35.0	171.4	81.6	44.4	151.5	42.4	50.6	46.3	0

Method-2:

TFs are able to tolerate some alterations in the sequence of the binding sites without losing functionality. For eg. CdxA can bind to motifs 101 (GTTAATA) and 100 (CATAAAG). In method 1, these motifs are considered be distinct where as in this method they are not differentiated. Each motif is labeled after the corresponding TF that is reported to bind to it. Thus the number of distinct motifs is reduced.

Table 2: Similarity measure between promoters of some mammals of citrate synthase using method 1

	Rat 7	**Can 10**	**Pan12**	**Pan 3**	**Hs 6**	**Hs 12**	**Bos 5**	**Sus 5**	**Mac 11**	**Hs 19**	**Pan 19**	**Bos 10**
Rat 7	100	23.5	24.2	22.6	22.9	26.9	22.7	46.9	24.3	27.5	26.9	23.1
Can 10	23.5	100	61.5	61.2	15.3	67.1	57.1	92.2	29.9	66.2	65.5	59.6
Pan12	24.2	61.5	100	62.7	55.8	91.8	90.3	70.1	85.2	90.3	91.5	85.7
Pan 3	22.6	61.2	62.7	100	58.9	82.3	88.3	67.5	87	89	88.9	92.2
Hs 6	22.9	15.3	55.8	58.9	100	46.8	67.9	24.3	68.9	47.7	48.8	61.9
Hs 12	26.9	67.1	91.8	82.3	46.8	100	83.8	75.8	77.3	92.1	91.9	81.9
Bos 5	22.7	57.1	90.3	88.3	67.9	83.8	100	64	86.5	84.9	84.8	90.6
Sus 5	46.9	92.2	70.1	67.5	24.3	75.8	64	100	52.8	75.7	74.8	66.3
Mac 11	24.3	29.2	85.2	87	68.9	77.3	86.5	52.8	100	77.8	78.9	90.6
Hs 19	27.5	66.2	90.3	89	47.7	92.1	84.9	75.7	77.8	100	98.7	88.8
Pan 19	26.9	65.5	91.5	88.9	48.8	91.9	84.8	74.8	78.9	98.7	100	89.7
Bos 10	23.1	59.6	85.7	92.2	61.9	81.9	90.6	66.3	90.6	88.8	89.7	100

	Rat 7	Can 10	Pan12	Pan 3	Hs 6	Hs 12	Bos 5	Sus 5	Mac 11	Hs 19	Pan 19	Bos 10
Rat 7	100	33.4	24.5	23.1	17.7	27	22.7	47.7	24.3	27.5	27.3	23.1
Can 10	33.4	100	61.5	61.6	43.4	67.9	58.2	92.2	30.7	67.1	65.5	61.7
Pan12	24.5	61.5	100	89.3	70.7	99.2	92.3	71.9	87.5	90.8	91.6	92
Pan 3	23.1	61.6	89.3	100	71.3	87.4	89.0	70.1	86.2	89.3	90.5	96
Hs 6	17.7	43.4	70.7	71.3	100	66.7	75.6	50.7	79.3	66.4	67.4	72.5
Hs 12	27	67.9	99.2	87.4	66.7	100	86.7	85.5	77.7	93.2	91.9	88.5
Bos 5	22.7	58.2	92.3	89.0	75.6	86.7	100	64.9	89.0	86.4	87.0	92.0
Sus 5	47.7	92.2	71.9	70.1	50.7	85.5	64.9	100	54.9	75.7	74.8	69.4
Mac 11	24.3	30.7	87.5	86.2	79.3	77.7	89.0	54.9	100	77.8	79.8	90.8
Hs 19	27.5	67.1	90.8	89.3	66.4	93.2	86.4	75.7	77.8	100	98.8	91.3
Pan 19	27.3	65.5	91.6	90.5	67.4	91.9	87.0	74.8	79.8	98.8	100	91.3
Bos 10	23.1	61.7	92	96	72.5	88.5	92.0	69.4	90.8	91.3	91.3	100

Table 3: Similarity measure between promoters of some mammals of citrate synthase using method 2

In the above tables,

Sus 5- Sus scrofa chromosome 5
Hs 12- Homosapiens chromosome 12
Bos- 5 BosTaurus chromosome 5
Can 10- Cannis familiaris chromosome10
Hs 6- Homosapiens chromosome 6
Bos 10- BosTaurus 10

Mac 11- Macaca Mulatta chromosome 11
Pan 19- Pan Trygolodytes chromosome 19
Rat 7- Rattus chromosome 7
Pan12- PanTrygolodytes chromosome12
Hs19- Homosapiens chromosome 19
Pan 3- Pan Trygolodytes chromosome 3

In all the methods, percentage similarity is calculated as [100- (dissimilarity score/maximum of lengths of the two motif sequences * 100)].

When comparison is made between sequences within organisms (table 1), Homosapien chromosomes 12 &19 have more similarity (rows 6 & 10). And Pan trygolodyte chromosomes 3, 12 and 19 (rows 3, 4 & 11) have high similarity.

It is reported that regulatory elements are evolutionarily important and thus conserved across species [17]. Highly conserved non-coding regions between human and mouse sequences are more likely to perform an important function, such as comprising regulatory elements to which TFs can bind, than non-conserved non-coding regions [7]. It is evident from our work, when comparison is made considering chromosomes (table 1). Homosapien 19 and Pan trygolodytes 19 (rows 10 &11) are highly similar and also compare more or less similar with all other mammals. Also, Homo sapien chromosome 12 and Pan Trygolodytes chromosome 12 show high similarity (rows 3 & 6).

Table 2 gives the similarity scores between motif sequences using method 1 and table 3 gives the similarity scores between motif sequences using method 2. It may be observed that the match score has increased in table 3 when compared with Table 2. This is expected because the different combinations of same TF have been grouped and assigned a common TF name. For eg. TF Caudal type homeodomain protein/ cardiac specific homeo box (CdXA) binds to CAATAAAACT, AACACGTTATT, AATAAATG, CATTTAAG, ACTTAAATT, TTGTGCAATA, ACTTAAAT, ACACGTTA. These motifs are not differentiated in method 2. Hence alignment score has increased.

4. CONCLUSION

The similarity search techniques have variety of applications from Medical Imaging, Molecular Biology, Spatial and Multimedia databases. Thus similarity search techniques should be flexible and adaptable to requirements of the applications or individual user preferences.
Comparison of promoter sequences may give insight into gene therapy and drug design [6]. The results reveal high similarity between related mammals.

5. REFERENCES

[1] Alan M Moses,Derek Y Chiang,Daniel A Pollard,Venky N Iyer & Michael BEisen, 2004 .MONKEY:Identifying conserved transcription factor binding sites in multiple alignments using a binding site-specific evolutionary model **;**Genome biology vol.5, issue 2,article 98,

[2] Altschul S.F., Gish,W., Miller,W., Myers,E.E. and Lipman,D.J. 1990. Basic local alignment search tool, J. Mol. Biol., 21**5,** 403–410. [PubMed]

[3] Nick Bray, Inna Dubchak and Lior Pachter, 2003.

[4] AVID: A Global Alignment Program. Genome Res. 13: 97-102.

[5] Blanco E, Messeguer X, Smith TF, Guigo´ R 2006. Transcription factor map alignment of promoter regions , PLoS Comput Biol 2(5): e49. DOI: 10.1371/journal.pcbi 0020049

[6] Brutlag.D. 2002. Multiple sequence alignment and Motifs, Bioinformatics methods and Techniques. Stanford University, Stanford center for Professional development,

[7] Davidov, E., Holland, J., Marple, E., Naylor, S., 2003. Advancing drug discovery through systems biology. Drug Discov Today. 8: 175-83.

[8] Down, T.A, Hubbard, T.J.P. 2004. What can we learn from non-coding regions of similarity between genomes. BMC Bioinformatics 5, 131-137.

[9] Eugene Berezikov, Victor Guryev and Edwin Cuppen, 2005. CONREAL web server: identification and visualization of conserved transcription factor binding sites**.** Nucleic Acids Research, Vol. 33, Web Server issue W447–W450 doi:10.1093/nar/gki378. Gen. Biol., 5, R98

[10] Ficket, J.W., A.G. Hatzigeorgiou, 1997. Eukaryotic promoter recognition. Genome Res 7, 861–878.

[11] Jacques van Helden, 2003. Regulatory Sequence Analysis Tools. Nucleic Acids Research, Vol. 31, No. 13 3593–3596 DOI : .1093/nar/gkg567.

[12] Kent, W.J. 2002. BLAT—The BLAST-like alignment tool. Genome Res. 12**:** 656–664,

[13] Meera A, Lalitha Rangarajan, Savithri Bhat, 2009.Computational Approach Towards Finding Evolutionary Distance And Gene order Using Promoter Sequences Of Central Metabolic Pathway. Interdisciplinary sciences-computational life sciences DOI: 0.1007/s12539-009-0017-3 [Spriger link],

[14] Mount.D. 2001. Bioinformatics - sequence and Genome analysis. Cold Spring Harbor, NY: Cold spring Harbor Laboratory Press,

[15] Ning, Z., Cox, A. J., and Mullikin, J.C. 2001. SSAHA: A fast search method for large DNA databases. *Genome Res.* 11**:** 1725–1729.

[16] Schwartz, ,S., Kent, W. J., Smith, A., Zhang, Z., Baertsch, R., Hardison, R. C., Haussler, D. and Miller, W., 2003. Human–mouse alignments with BLASTZ. Genome Res., 13, 103–107

[17] Smith, T. F. and Waterman, M. S., 1981. Identification of common molecular subsequences. J. Mol. Biol. 147**:** 195–197

[18] Ureta-Vidal A., Ettwiller, L., Birney, E, 2003. Comparative genomics: genome-wide analysis in metazoan eukaryotes. Nat Rev Genet. Apr; 4(4):251-62.