Journal of Computational Intelligence in Bioinformatics
ISSN: 0973-385X Volume 3 Number 1 (2010) pp. 49–61

http://www.ripublication.com/jcib.htm

GENETAL: A New Technique for Feature Extraction from Large Set of Biological Sequences and Its Use in Classification

Ulavappa. B. Angadi* and M. Venkatesulu

Department of Computer Applications, Kalasalingam University
Krishnankoil, Srivilliputtur (via), Tamil Nadu, India, 626 190.
*Corresponding author *E-mail: angadiub@gmail.com,*
E-mail: venkatesulu_m2000@yahoo.com

Abstract

In bioinformatics, enormous biological data is being accumulated due to genome sequencing projects all over the globe. Compelling need to transform biological data into useful information and knowledge is become an important and challenging task to both computer scientists and biologists. One of the problems arising in the analysis of biological sequences is the discovery of similar motifs/features from set of sequences. Such motifs usually corresponds to residues conserved during evolution due to an important structural or functional rule. In this paper, we develop a new algorithm GENETAL based on genetic theory for discovery of motifs/features in biological sequences and text documents. Our algorithm is able to produce all motifs appearing at least a minimum number of sequences (user defined). It is very efficient compared to other existing algorithms for large set of data, with respect to space and time complexity. Also, we demonstrate clustering of DNA/Protein sequences and text document data using GENETAL as feature extraction algorithm with simple incremental clustering technique and Jaccard coefficient dissimilarity measures.

Key words: Motifs discovery, Clustering, DNA/Protein sequences, Pattern recognition.

Motivation: Discovery of motifs in biological sequences is an important problem and has many applications. However, the existing methods are not able to do exhaustive search with complex motifs representations and each is applicable to only a certain class of problems.

Introduction

In bioinformatics, enormous biological data is being accumulated due to genome-sequencing projects all over the glob. Compelling need to transform biological data into useful information and knowledge is becomes an important and challenging task to both computer and biological professionals. Discovering of motifs and classification of sequences are major directions in predicting the function or the structure of new sequence. Motifs are used for phylogenetic analysis to determine how the families have been derived and to construct evolution tree. The assumption behind motifs discovery approaches is that motifs that appears often enough in a set of biological sequences is expected to play a major role in defining the respective sequences' functional behavior and evolutionary relationships. One of the problems arising in the analysis of biological sequences is the discovery of similar motifs/features from set of sequences. Such motifs usually corresponds to residues conserved during evolution [1], [2], [3], [4], [5], [6], [7], [8].

Motifs/patterns discovery encompasses a wide variety of methods used to find recurrent trends in data. In bioinformatics, the two predominant applications of motifs/patterns discovery are sequence analysis and micro array data analysis. Motifs discovery in sequence analysis typically involves the discovery of binding sites, conserved domain or other discriminatory subsequences.

Lot of research has been done on computational techniques for discovering motif or motifs in small set of data. Discovering motifs in large number of sequences still remains a challenge. In general, motifs discovery uses many types of approaches such as pair wise alignment, combinatorial, moving window, Gibbs sampling, HMM and etc. However, these approaches become less effective, as the number of sequences increases and more complex, especially when a set of sequences includes multiple families.

In this paper, features/words/motifs are considered to be synonymous.

Related work

Several methods have been proposed for discovering motifs and classification of sequences. One widely used class of algorithms [9], [10], [11], [12] employs global string alignment, in this context, edit operations (mutations, insertions, deletions) are used along with their associated costs. Alignment algorithms suffer from several inherent drawbacks. Firstly the task of optimally aligning a set of strings is computationally very expensive. Secondly, alignment of entire sequence can reveal only global similarity [9]. If the sequences under comparison are distantly related or if the relative order of their similar regions varies among sequences, then, in such situations, traces of evolutionary relationships might only be detectable as similarities over short stretches of residues.

Secondly, BLOCK database approach is to look for small numbers (e.g. 3) of exactly or highly conserved positions separated by short fixed spacing and local alignment [9] and were used to provide the initial conserved segments for BLOCK database of conserved sequence blocks [14], [15].

One way to over come the difficulty that alignment algorithms have in identifying local similarities is to focus on the discovery of patterns shared by s set of sequences. Third approach is a practical use of a pattern [16], [17], [18], [19], [20], [21] [22], [23] is to find diagnostic signatures for families. This is well illustrated by the PROSITE data base of proteins patterns [24]. Here groups of functionally and evolutionarily related proteins are listed along with their patterns, which can be used to distinguish each family from all other sequences in the SWISS_PROT protein sequences database. These patterns are extracted semi-manually.

Fourthly, unless the nature of motifs sought is extremely simple, the problem of detecting all existing motifs is NP_hard[25]; typically, a reduction from longest common subsequences problem [26] can be used to prove this. Possible remedies include the use of heuristics [27], [28], [29], [30], [31], [32], [33], [34], which offer enhanced performance, but at the expense of sacrificing the completeness of the results and/or the structural restriction of the patterns. (Maximum length)

In the above techniques, each is applicable to only a certain set of sequences and it could be difficult to know a priori which technique will be most appropriate. Hence, we present GENETAL, a generic efficient novel algorithm for the extraction of motifs/features from biological sequences and text documents without using pair wise alignment and not comparing with other motifs.

Terminology

Take a set of N no. of unaligned DNA/Protein sequences/text documents S_1, S_2, …….S_N of varying lengths L_1, L_2, …… L_N. Each of these sequences can be presented as $Si=a_{i1}, a_{i2}, \ldots. a_{iLi}$ where $a_{iLi} \in \alpha$ i=1 to N and α is the set of nucleic, amino acids and words for DNA, amino acids sequence and text documents, respectively.

Motif data structure

m (L,Q, f) Є M where M is set of motifs, m is motif with L as motif label, Q as parents of the motif; f is number occurrence in the set.

Fitness/Probability function

Probability of survival of new gene m in next generation based on parents' population m1 and m2.

$$P(m) = \frac{\text{Min (Min (freq (m1), freq (m2)), required no of genes)}}{\text{Required number of genes}} \quad (1)$$

Mutation

m1 and m2 are consecutive motifs and new motif m defined as

$$m = m1 \oplus m2 \begin{cases} m1m2` \text{ if suffix (m1) =prefix (m2)} \\ Ø \qquad \text{otherwise} \end{cases} \quad (2)$$

Methodology

The algorithm operates in three phases; initialization, generation, and merging and filtering phases (Fig1). Initialization phase is process of scanning the input set by successive pairs of motifs/genes of given length. The result of scanning process is a set of sequences containing labels of all elementary motifs/genes of the given starting length and list of motifs m (L, Q, f). The resultant set of sequences is input to the generation phase. Generation phase similar to initialization is the process of scanning the resultant set of initialization phase by successive pairs of motifs/genes and generates new motif/gene by mutation based on probability of survival (Eqtn.1.) of the motif/gene in next generation. The result of generation phase is a list of motifs, and a set of sequences containing labels of motifs, which is input set to the next generation. All successive generations are carried out until no the further mutation or new motif/gene takes place. The task, then, of the generation phase is to extract motifs from list of motifs labels by recursive merging process (Eqtn.1), which satisfies the minimum required number of occurrences in the set.

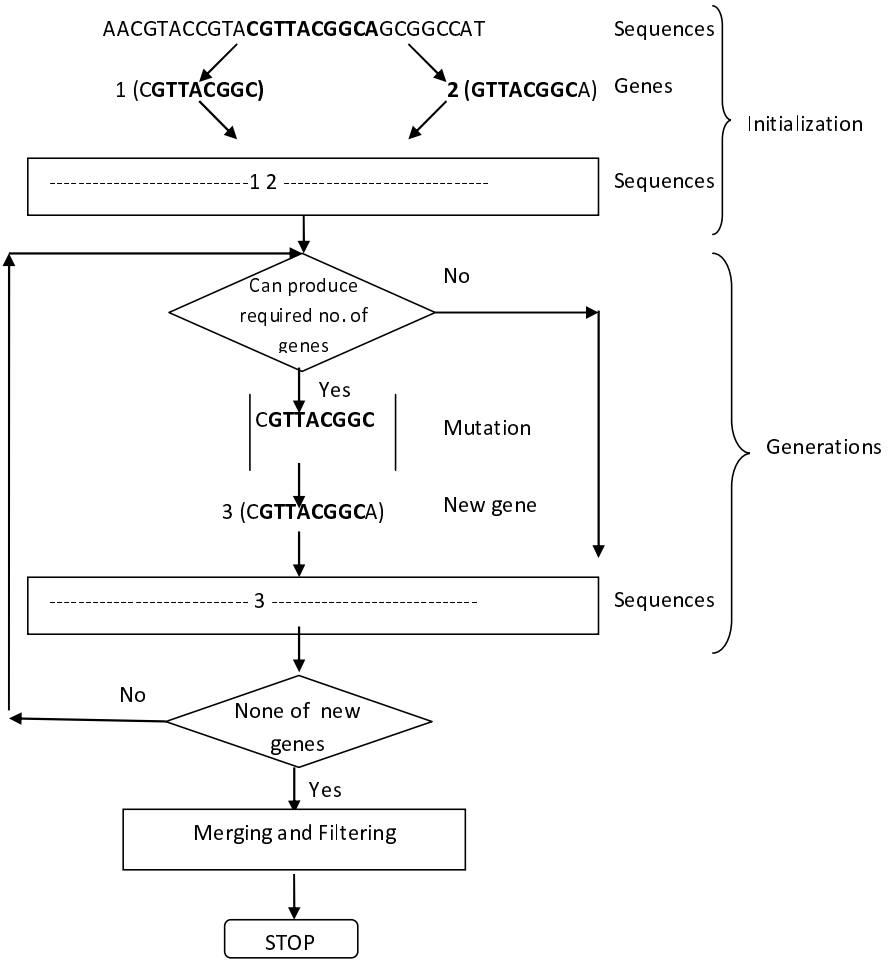

Figure 1: Flow Diagram of GENETAL.

Algorithm

The algorithm has three phases: Initialization, Generation and Extracting motifs from motifs labels. Initialization phase initialize motif/gene by combining two successive characters from given a set of sequences. During the generation phase, new motifs/genes are produced by two successive genes/motifs based on probability of survival. In Extracting, recursive merging of two genes (Eqtn.2) happens.

Algorithm-1: GENETAL Algorithm for discovering features

Input:

S a set of sequences
n minimum number of occurrence in the set
l initial size of motifs

Output:

M set of motifs
M' set of motifs after filtering with redundancy and substring

Initialization:

```
Do for all S
  If m is not existed in M
    // m+ is indicates new motif, frequency=1
     m(S:i,i+l) → m+ Є M for i=1 to size of sequence - 1
  Else
     frequency ++
  End if
  Replacing pairs of alphabets with motif label in the set S
```

Generation:

```
Do for all generations
Do for all S
   If (m is not existed in current generation)
            If (P (m1, m2)>1) →m+ Є M // new motif with frequency=1
   Else
            frequency++
   End if
     Replacing m1, m2 with new motif m+ label
End do
Until ( no new motif)
```

Filtering:

M→M`  m Є M` where frequency (m)> n and m1& m2 are not mutual substring with same frequency

Extracting:

```
Do for all M`
        Merge(m)
```

Algorithm for mutation of pairs of genes

```
Merge(p){
    If (is( alphabets))
       Return elementary motif
    else
       p1=parent1(p)
       p2=parent2(p)
       m1=merge(p1)
       m2=merge(p2)
       p=m1⊕ m2 // suffix(m1)==prefix(m2)
    end if }
```

Clustering

The main purpose of the paper is to present feature extraction technique, which can be used in wide variety of applications. But here we demonstrate application of GENETAL for clustering DNA/Protein sequences and text documents with simple incremental clustering (leader without centroid update) algorithm [35],[36] (Algorithm-2) and Jaccard Co-efficient dissimilarity metric [37].

Let N be the number of given sequences; let n be the number of motifs generated by GENETAL such as m1, m2, ……..mn . Each sequence Si is represented as a vector of length m. Si = [C_{1i}, C_{2i}, …C_{ij}.. C_{mi}.] where C_{ij} is one if motif j appeared in sequence i and is zero if motif j does not appear in sequence i. The entire set of sequences is represented as table (Table 1). An incremental clustering algorithm is used to classify the given data using Jaccard co-efficient dissimilarity.

Jaccard Co-efficient Dissimilarity

Table 1: Motifs appearance table for the sample.

	Motif-1	Motif-2	Motif-2	Motif-3	Motif-4		- - -	Motif-n
S1	0	1	1	1	0		0	0
S2	1	0	1	1	1		0	1

Jaccard Co-efficient Dissimilarity (i, j) =(r+s)/ (q+r+s) (3)

where q is the number of features equal to 1 for both sequences, r is number of features that is equal to 1 for sequence-1 and 0 for sequence-2, s is number of features is that equal to 0 for sequence-1 and 1 for sequence-2.

Algorithm-2: Leader algorithm for clustering

Input:

An appearance table of size N X n

Threshold t

Output:

Ci clusters i=1 to k

Algorithm:
L-1= A1j j=1 to n // first sequence
C -1 = first sequence
do for all sequences i=2 to N
finding nearest cluster in existing clusters based on threshold t and Jaccard coefficient
if nearest is available
add sequence to existing cluster
else
create a new cluster i^{th} sequence as leader
end do

Implementation
The algorithm is implemented in C. In the program temp files are used to storing motifs details while testing the algorithm with large data set. User-supplied parameters: The input to GENETAL is a set of sequences in FASTA format. Minimum length of motifs. Minimum number of occurrences of a motif in the set. Availability: The C++ open source single program is available on request angadiub@gmail.com.

The implementation is in two variant, one is for nucleic and amino acids sequences in FASTA format. Another is for text mining. In this variant, during initialization, the pairs of successive words are considered as elementary motifs/genes and similarly in merging process.

Results and Discussion

In this section, we demonstrate GENETAL's capability by presenting clustering of large set of DNA/Protein sequences and text documents. We also evaluate the performance of GENETAL through series of experiments. Specifically, we address features extraction from amino acids, nucleic acids sequences and from text documents. For testing the performance of the algorithm, the following experiments have been done. We consider, two data set for each application, such as DNA, protein and text documents classification. In the two data sets of each application, one is small data set, which is used to analyze the correct classification of known classes, using precision rate. From second data set, we analyze time complexity and classification accuracy based on training class. All our experiments are done on Intel Celeron M processor 370 (1.5GHz, 400MHz FSB, 1MB l2 cache) and 256 MB DDR2 RAM.

Motifs discovery in nucleic acids sequences and using them for classification
In the first data set, fifteen mammals unaligned mitochondrial genomes were collected from Gene Bank; 1. Human (Homo sapiens [V00662]), 2. Chimpanzee (pan troglodytes [D38116]), 3. Pigmy chimpanzee (pan paniscus [D38113]), 4. Gorilla (gorilla [D38114]), 5. Orangutan (pango pygmaeus [D38115]), 6. Gibbon (hylobates

lar [X99256]), 7. Sumatran orangutan (pongo pygmaeus abelii [X97797]), 8. Horse (Equus caballus [X79547], 9. White rhino (ceratotheium simum [U07726]), 10. Harbor seal (Phoca vitulina[X63726]), 11. Gray seal (Haichoerus grupus[X72004]), 12. Cat (felis catus [U20753]), 13. Finback whale[X61145], 14. Blue whale[X72204], 15. Rat(rattus norvegicus[X14848], 16. House mouse(mus musculus[V00711] and 17. Donkey (Equus asinus [X97337]).

Table 1: Results of 17 mammals DNA sequences with classification threshold =0.38.

Total letters/ BPs	**Min. No. of Occurrence**	**Motifs size range**	**Total no. of motifs**	**Motifs after refined**	**Time for motif generation (Secs.)**	**Time for classification (Secs.)**	**Evolution**
2,83,290	**6**	**05 to 76**	**133402**	**32980**	**4577**	**260**	**0.94**
	8	**06 to 74**	**83782**	**21106**	**3826**	**260**	**0.94**
	10	**10 to 73**	**60837**	**13632**	**3497**	**258**	**0.94**
	12	**10 to 71**	**45458**	**8460**	**2492**	**160**	**0.82**

Precision rates (PR) = No. of correctly classified/Total no. of sequences

In order to evaluate the relationship between the performance and input compositions, we carry out four experiments on first dataset of DNA sequences. The input parameter, number of minimum occurrence of a motif in the set is inversely proportion to the size of motif, number of motifs, time, classification accuracy. This experiment yields the result which are same as groups suggested by biologists. Phylogenetic tree can be constructed using neighbor-join algorithm based on Jaccard coefficient dissimilarity.

The second data set is having 4627 DNA sequences of Bermuda grass, switch grass, switch grass, sorghum propinquum, sorghum halepense, rice panicle and etc. Sequences in the set were well shuffled before the experiments and experimental results of clustering on the dataset are show in table-2. The classification accuracy (CA) is not much affected among the various combination of training and testing data. It is evident that the classification with features is having more consistency.

Table 2: Results of 4627 fodder related DNA sequences.

Data set	**Total letters/ BPs**	**Motifs range**	**No. of motifs (total/ selected)**	**Motifs generating time (Secs.)**	**Appearance tabling time(Secs.)**	**Evolution**
Second data set	**3093090**	**4 to 15**	**2061/625**	**4219**	**96 Secs.**	**PR=0.87**
	Threshold	**Training data**	**Testing data**	**Training time**	**Testing Time**	**Classification accuracy**
	0.38	**1000**	**3627**	**4**	**20**	**99.47**
		2000	**2627**	**10**	**14**	**99.54**
		3000	**1627**	**14**	**10**	**99.75**

CA= Total no. of testing sequences assigned into training clusters X 100 / Total no. of testing sequences.

Features extraction from protein sequences and using them for classification
First data set consist of 497 sequence belongs to 6 superfamily, namely Globin-like (105), EF-hand (74), Cupredoxins(24), (trans) glycosidases(99), Thioredixin-like (100), Membrane all-alpha(95). Second data set consist of 5917 sequences. Results of time complexity and classification accuracy obtained on various training and testing data sets are shown in the table 3.

Table 3: Results of 525 and 5917 of protein sequences.

Data set	**Total characters**	**Motifs range**	**No of motifs**	**Time for motif generation**	**Time for appearance table and classification**	**Evolution**
First data set	248653	3 to 5	3186	892 Secs.	22 Secs.	
	Threshold	Training data	Testing data	Training time	Testing Time	Classification accuracy
	.984	200	325	1	1	98.37
		300	225	1	1	99.65
		400	125	1	2	99.76
Second data set	1795387	3 to 7	2780	7350 Secs.	124 Secs.	
	Threshold	Training data	Testing data	Training time (Secs.)	Testing Time (Secs.)	Classification accuracy
	.95	2000	3917	42	121	98.34
		3000	2917	69	94	98.56
		4000	1917	100	63	98.80

Features extraction from text documents and using them for classification
In this experiments, we have selected 13 articles from references of this paper (9, 10, 12, 14, 15, 16, 20, 21, 27, 29, 31, 33, 34), which consists 306292 characters. Total 32326 motifs are generated by GENETAL in 1067 secs. After filtering, finally 204 motifs (Table. 4) are selected for preparing appearance table. Motifs range(in words) 2 to 5. Incremental clustering has grouped all the articles in 3 classes viz. Class-1:9, 16, 20, 27, 33 ; Class-2: 12, 21, 34, ; and Class-3: 10, 14, 15, 29.

Table 4: Some patterns from that are presented 14 reference articles.

Motif Label	No. of Occurrence In set	Motifs
32321	2	based on also single
32320	2	if and only if
32318	3	the size of the
32314	2	of pattern discovery in
32273	4	in the original
32271	4	in the same
32270	3	from an singleton
32269	2	sequences and tools

32267	2	patterns of the
32266	3	in the Creative
32265	5	based on the
32263	5	in any singleton
32261	2	of the original
32255	4	notion also singleton
32254	5	based on also
32229	2	motifs can be
32214	2	of different lengths.
32212	2	pattern may be
32209	3	each sequence of
32208	2	that each sequence
32202	4	have been identified
318	2	biological sequences.
314	2	to compress
312	2	used advantageously
307	11	based on
300	4	classification and
295	3	and classification,
292	3	for structural
195	2	the discovery
175	2	discovery techniques
173	2	sequences. Motif

Conclusion

The overall computational time greatly depends on the number of iterations, here the number of iteration is equal to maximum size of motifs. The overall computational time complexity is O(N x l), where N is total number of BPs/characters in the data set and l is maximum size of motifs. The space and time usage of the implementation of the algorithm is very reasonable. During merging process, if motif size increases then the time complexity also increases. In case of space complexity, there is no constraint, it can support any number of sequences and size, because it is simple iterative solution and disk space can be used. The algorithm works better in non-redundant data set.

While experimenting with various data sets, we have observe that the most important input parameter is minimum number of appearance of motifs in the dataset, Time and space complexity depends on this, but it does not affect much the final clustering result. In this paper we presented better results using simple clustering technique to demonstrate feature extraction technique and its application. The experimental results demonstrate the potential and consistency of our proposed technique. Classification performance can be improved using other advanced clustering technique based on features discovered by GENETAL with frequencies of features in a sequence.

The algorithm is generic; it is applicable to any variant of sequential data and text data. It returns exhaustive set motifs. Furthermore the GENETAL algorithm could be modified and applied to gapped motifs discovery in DNA/Protein sequences, discovery of structural motifs in protein sequences, discovering pattern in protein as

described in PROSITE, prototype / median sequence selection in multiple sequence alignment, identifying text pattern in text documents, text pattern search in internet.

References

[1] D., Bashfold, C,. Chothia, and A., M., Lesk, 1987, "Determinants of a protein fold unique features of the globin amino acid sequences", J. Mol. Biol., 196, pp. 199-215.

[2] M., Caserta, W., Zacharias, D., Nwankwo, G., G.. Wilson, and R., D., Wells, 1987, "Cloning, sequencing in vivo promoter mapping and expression in Escherichia coli of the gene for Hhal methyltransferase," J. Biol. Chem., 262, pp. 4770-4777.

[3] S., Som, A., S., Bhagwat, and S., Friedman, 1987, "Nucleotide sequence and expression of the gene encoding the ECORII modification enzyme," Nucleic Acids Research, 15, pp. 313-332.

[4] L., A., Sznyter, B., Slatko, L., Moran, K., H., O'Donnell, and J., E., Brooks, 1987, "Nucleotide sequence of the Ddel restriction modification system and characterization of the methylase protein," Nucleic Acids Research, 15, pp. 8249-8266.

[5] Jony Hunter, 1987, "A thousand and one protein kinoses," Cell, vol.50, pp. 823-829.

[6] R., J., Lipton, T., G., Marr, and J., D., Welsh, 1989, "Computational approaches to discovering semantics in molecular biology," Proceedings of the IEEE, 77, pp. 1056-1060.

[7] J. Posfai, A., S., Bahwat, G., Posfai, and R., J., Roberts, 1989, "Predictive motifs derived from cytosine methyltransferases," Nucleic Acids Research 17, pp. 2421-2435.

[8] K., Romisch, J., Webb, J., Herz, S., Prehn, R., Frank, M., Vingron, and B., Dobberstein 1989, "Homology of 54k protein of signal recognition particle, docking protein and two E.coli protein with putative GTP-binding domains," Nature(London), 340, pp. 478-482.

[9] Michael S., Waterman, " Efficient sequence alignment algorithm," 1984, J. Theor. Biol., 108, pp. 333-337.

[10] Eric Sobel and Hugo M., Martinez, 1986, "A multiple sequence alignment program," Nucleic Acids Research, Vol. 14, pp.363-374.

[11] Randall F., Smith and Temple F., Smith, 1990, "Automatic generation of primary sequences patterns from set of related protein sequences," Proc. Natl. Acad. Sci. UAS, Vol. 87, pp.118-122.

[12] Thomas D., Wu and Douglas L., Brutlag, 1995, "Identification of protein motifs using conserved amino acid property and partitionery technique," Proc. of Third International Conference on Intelligent System for Biology (ISMB 95), pp. 402-10.

[13] Michail A., Roytberg, Aleksy Y., Ogurtsov, Svetlana A., Shabalina, and Alexey S., Kondrashov, 2002, "A hierarchical approach to aligning collinear regions of genomes," Bioinformatics, Vol.18, No. 12, pp. 1673-1680.

[14] Steven Hanikoff and Jorja G., Henikoff, 1991, "Automated assembly of protein blocks for database searching," Nucleic Acids Research, Vol. 19, No. 23, pp. 6565-6572.

[15] Steven Henokoff and Jorja G., Henikoff, 1994, "A Protein family classification method for analysis of large DNA Sequences," Proc. of the Twenty-seventh Annual Hawaii International Conference on System Sciences, pp. 265-274.

[16] Graziano Pesole, Nicola Prunella, Sabino Liuni, Marcella Attimonel, and Cecilia Saccone, 1992, "WORDUP: an efficient algorithm for discovering statistically significant patterns in DNA sequences," Nucleic Acids Research, Vol. 20, No.11, pp. 2871-2875.

[17] I., Jonassen, J., F., Collins, and D., G., Higgins, 1995, "Finding flexible patterns in unaligned protein sequences," Protein Science, 4(8), pp. 1587-1595.

[18] Alvis Brazma, Inge Jonassen, Esko Ukkonen, and Jaak Vilo, 1996, "Discovering Patterns and subfamilies in Biosequences," Proc. Intelligent Systems for Molecular Biology (ISMB-96), pp. 34-43.

[19] M., F., Sagot and A., Viari,1996, "A double combinatorial approach to discovering patterns in biology sequences," in Proc. of Seventh Symposium on Combinatorial Pattern Matching, pp.186-208.

[20] Jens Hanke, George Beckmann, Peer Bork, and Jens G. Reich, 1996, "Self-organizing hierarchic network for pattern recognition in protein sequences," Protein Science, 5, pp. 72-82.

[21] Isidore Rigoutsos and Aris Floratos, 1998, "Combinatorial pattern discovery in biological sequences: the TEIRESIAS algorithm," Bioinformatics, Vol. 14, No.1, pp.55-67.

[22] Gerald Z., Hertz and Gary D., Stormo, 1999, "Identifying DNA and protein patterns with statistically significant alignments of multiple sequences," Bioinformatics Vol.15, Nos. 7/8, pp. 563-577.

[23] Isidore Rigoutsos, Aris Floratos, Laxmi Parida, Yuan Gao, and Daniel Platt, 2000, "The emergence of pattern discovery techniques in computational biology," Metabolic Engineering, Vol.2:3, 159-177.

[24] A., Bairoch, P., Bucher, and K., Hofmann, 1996, "The PROSITE database: its status in 1995," Nucleic Acids Res., 24, pp. 189-196.

[25] M., R., Garey and D., S., Jahnson, 1979, "Computers and intractability: A guide to the theory of NP-Completeness.

[26] D., Maier, "Complexity of some problems on subsequences and super sequences," 1978, J. ACM, pp. 322-336.

[27] T., F., Smith and M., S., Waterman, "Identification of common molecular subsequences," 1981, J. Mol. Biol., 147, pp. 195-197.

[28] Hamilton O., Smith, Thomas M., Annau, and Srinivasan Chandrasegaran, 1990, "Finding sequences motifs in groups of functionally related proteins," Proc. Natl. Acad. Sci. USA, Vol.87, pp. 826-830.

[29] C., E., Lawrence, S., F., Altschul, M., S., Boguski, J., S., Liu, A., F., Neuwald, and J., C., Wootton, 1993, "Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment," Science, 262, pp. 208-214.

[30] Jason T., L., Wang, Thomas G., Marr, Dennis Shasha, Bruce Shapiro, and Gung-Wei Chirn, 1994, "Discovering active motifs in sets of related protein sequences and using them for classification," Nucleic Acids Res., pp. 2769-2775.

[31] Timothy L., Bailey and Charles Elkan, 1995, "The value of prior knowledge in discovering motifs with MEME," Proc. of the Third International Conference on Intelligent Systems for Molecular Biology (ISMB, 95), pp. 21-29.

[32] Jeremy Buhler and Martin Tompa, 2001, "Finding motifs using random projections," Proc. of the Fifth Annual International Conference on Computational Molecular Biology (RECOMB01), pp. 69-76.

[33] Alkes Price, Sriram Ramabhadran, and Pavel A., Pevzner, 2003, "Finding subtle motifs by branching from sample strings," Bioinformatics, vol. 19 Suppl.2, pp. ii149-ii155.

[34] Alberto Apostolico, Matteo Comin, and Laxmi Parida, 2006, "Mining, compressing and classifying with extensible motifs," Algorithms for Molecular Biology, 1:4.

[35] P.,A., Vijaya, M., Narasimha Murty, and D., K., Subramanian, 2004, "Leaders-Subleaders: An efficient hierarchical clustering algorithm for large data sets," Pattern Recognition Letters, 25, pp. 503-511.

[36] Margaret H., Dunhom, 2005, "Data mining introductory and advanced topics," Pearson Education Singapore.

[37] Jiawei Han and Micheline Kamber, 2005, "Data mining concepts and techniques," Elsevier, pp. 341-342.