Biological Motif Discovery Algorithm based on Mining Tree Structure

Lounnas Bilal
Department of Computer Science.
Laboratory of Pure and Applied Mathematics LPMA
University of M'sila Algeria

Bouderah Brahim
Department of Computer Science.
Laboratory of Pure and Applied Mathematics LPMA
University of M'sila Algeria

Moussaoui Abdelouahab
Department of Computer Science
UniversityFerhat Abbas of Sétif
Algeria

ABSTRACT
The nucleic acid and protein sequences contain different types of information (genes, RNA structural, active sites, regulatory structure ...), these information can lead to discover many useful knowledge on biology like the functionality of a given protein sequence, another example is toclassifying proteins on different families based on these information. In this paper we focus on the existed motif in the nucleic acid sequences. Before going further it is useful to review the concepts and terminology associated with this study.

The motif is a structural short element that could be found in all members of a family of protein. It contains essential residues for function conserved, not necessarily consecutive, but rather closes to the 3D structure, be-cause they involve the same function (active site, binding site ...). While the pattern or profile is a degenerate sequence and/or composed of different motif that can be separated by variable regions.

In fact, the objective is to develop a new algorithm based on mining tree structure in order to highlight segments of DNA, RNA, or amino acids, which are likely to have a biological role

Keywords
Motif matching, profiles, tree automata, pushdown automata, tree structure, tree mining.

1. INTRODUCTION
In the last decade the field of biology has become in the midst of "Data explosion", due to the huge data recording by biologist using many different techniques [1].

This evolution of biological data as a point of size or nature has caused an appearance of a new concept on the research area, the concept is bioinformatics, so Bioinformatics is conceptualizing biology in term of molecules (in the sense of physical-chemistry) and then applying informatics techniques to understand and organize the information associated with these molecules on large scale [2]. It also can be seen as all the computer techniques and computational statistics for analyzing biological data [1]. There are a few other definitions on the concept of bioinformatics in many different resources in the litterateurs, so in the big picture, bioinformatics is the task that provide necessary search, score, and analyze biological information by algorithms and specific tools from data sets that are accumulated and curated by experts in order to provide a powerful opportunity to improve human health [3].

Many challenges on biology has been determined by biologists, that cause launch of several bioinformatics tasks, but one of the most important task on those challenges is the discovery of motif from biological sequence in order to define the function or the family of biochemical molecular (DNA, RNA, and Protein).

A large number of motif discovery algorithm have been propose and implemented, each algorithm have different proprieties of others, some focus to improve the data structure that represent biochemical molecular, other use combinations between technologies to enhance the matching process. Furthermore technical details the paper show a small survey of motif discovery algorithm that have remarkably used on this bioinformatics task, and for our work we going to present a brief concept of a new algorithm based on mining tree structure.

The paper is dived into four parts, for the first part talk about some motif discovery algorithm by showing different characteristics that discriminate each one from others, in other word the first part is the survey as we mentioned before, for the second part we are going to describe a new concept algorithm for the motif discovery process, the fourth and the fifth part is related work and conclusion that include future work.

2. MOTIF EXTRACTION ALGORITHMS
As there are a big growing interest on regulatory element that can lead to understand some virus function, detect new drug, classify spices, or to get many other helpful new knowledge of biology. The researchers have developed many algorithms in order to discover or predict this small part of biochemical molecular, each of these algorithms have different concept, in the way of data representation to the process of discovering to the results.

2.1 STEPS OF BUILDING MOTIF DISCOVERY ALGORITHM
These steps are based on Brazma et al research [4].

Training set.
There are two different training set types to choose, depending on what king of result that the algorithm need to delivery.

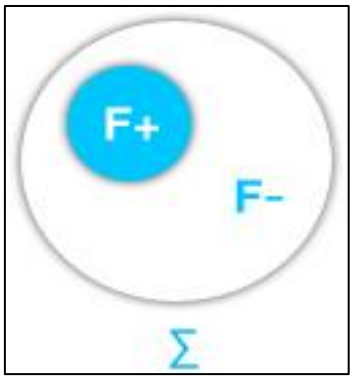

Fig1:Set of training divided into two groups

We assume

$$\Sigma = F^{+} + F^{-}$$

Where Σ is a set of all possible sequence, and F+ is a subset of Σ, all F+ sequence is shared the same characteristics (ex: protein family), we call F+ positive example and F- negative example.

The first type of training set is when the algorithm is designed to be a classifier method, in this case both positive and negative example (all the training set) are given.

The second type of training set is used the positive example only, this chose when the algorithm is designed to discover the common characteristics on this training set (ex: a protein family data set as an example).

Pattern model.

The input data or the training set is the most influential factor on how pattern model how going to be. A simple pattern model can be define as a Boolean type (return true or false) for a given sequence if it include on a training set or not, this model is called exact pattern matching.

The pattern model can work on approximate way when some measures technics are involve, for example when the algorithm use amino acid substitution matrix.

More complicated pattern model can be found such as pattern model that use union of other model, an example for that when positive example F+ contain a family of protein and other subfamilies. There are also some pattern uses some intelligence techniques for the matching process such as decision tree.

Ranking process.

Ranking process is the way to measure the performance of pattern model over the matching process. There are different ranking methods depend on the pattern model and the type of training set, the quality of the training set may cause some influence on ranking process because in reality sequence come from biological experiment and may contain errors [4].

Developing the algorithm.

Last step according to Brazma et al. research [4] is design the algorithm, he mentioned that are motif discovery algorithm approach can be in two ways. The first is pattern driven, this approach find the highest pattern candidate rating, on a given training set. The second is sequences driven approach, on this one the algorithm try to find pattern candidate by comparing the given sequence and looking for similarity between all sequence on the training set, this approach could base on constructing a local multiple alignment on given sequence and all sequences on the training set.

2.2 CATEGORIES OF MOTIF DISCOVERY ALGORITHM

Many criteria factor can interfere on define the categories of motif discovery algorithm, especially those that mentioned before on steps of building algorithm. On Modan et al [5], classify the available motif discovery algorithms into three major classes as those based on promoter sequences of coregulated genes, and other based on phylogenetic footprinting, and for the third class is the algorithms are based on merge between the first two classes, algorithms based on promoter sequences of coregulated genes and phylogenetic footprinting.

The three categories are resulting on the type of training set that are used on motif discovery algorithm.

Another categorization approach depended on the functionality of the algorithm itself, this divide motif discovery algorithm into two major classes [5]:

Word-based algorithms.

Word-based (WB) relies on regular expression (RE) because RE is a powerful notational algebra describing strings and sequences [6]. WB preferred for discovery short motif especially when implemented with suitable data structure, but still can be a bad choice for extraction transcription factor motif that often has several weakly constrained position [7].

Probabilistic algorithms.

As the probabilistic motif is hard to understand and explain, PA describes this type of motif with a position weight matrix (PWM) [8]. This type of algorithms is preferred to find a longer motif compared with word-based algorithms.

Two categorization approaches are describe, and many other categorization are proposed on the literature of motif discovery algorithms. But still the part of categorize the algorithm depend on the researcher, and each one can see his approach. A proposed approach of categorization can be based on the originality of the algorithm, and this can lead to define many factor criteria for that, such as, if the algorithm has a process of cleaning noise from training set, and if it implement a new data structure approach for handle biological data, another critter factor can distinguish between hybrid and native algorithm.

2.3 SOME WIDELY ALGORITHMS

Motif discovery now is a very active field on bioinformatics, consequences to the availability and the quality of biological data, a wide range of algorithms has been proposed on this field, Table 1 is a list of motif discovery algorithm presented on a chronological way [5].

Table 1. chronologically presentation of motif discovery algorithms.

PSM refer to probabilistic sequence model and PF refer to phylogenetic footprinting.

Algorithm	Operating principle	Classification
by Galas *et al.*	Enumeration	Word-based
by Mengeritsky and Smith	Enumeration	Word-based
by Staden	Enumeration	Word-based
EM	Expectation maximization	PSM
WordUP	Enumeration	Word-based
Gibbs sampler	Gibbs sampling	PSM
MACAW	Gibbs sampling	PSM
MEME	Expectation maximization	PSM
AlignACE	Gibbs sampling	PSM
Oligo-Analysis	Enumeration	Word-based
Consensus	Weight matrix	PSM
Dyad-Analysis	Enumeration	Word-based
WINNOWER	Graph	Other
ANN-Spec	Gibbs sampling	PSM
SMILE	Suffix tree	Word-based
Verbumculus	Suffix tree	Word-based
MobyDick	Dictionary	Word-based
YMF	Enumeration	Word-based
Bioprospector	Gibbs sampling	PSM
Co-Bind	Gibbs sampling	PSM
ITB	Enumeration	Word-based
Weeder	Enumeration	Word-based
MotifSampler	Gibbs sampling	PSM
MITRA	Prefix tree/Graph	Word-based
MDScan	Greedy algorithm	Other
Projection	Hashing	Other
Footprinter	Dynamic programming	Other
MOPAC	Enumeration	Word-based
DMotif	Enumeration	Word-based
PhyloCon	Consensus	PF
LOGOS	Expectation maximization	PSM
EC	Genetic algorithm	Other
GLAM	Gibbs sampling	PSM
Improbizer	Expectation maximization	PSM

QuickScore	Consensus	PSM
SeSiMCMC	Gibbs sampling	PSM
PhyME	Expectation maximization	PSM
OrthoMEME	Expectation maximization	PSM
FMGA	Genetic algorithm	Other
PHYLONET	Sequence alignment	PF
PhyloGibbs	Gibbs sampling	PSM
GIMF	Expectation maximization	PSM
WordSpy	Dictionary	Word-based
MaMF	Enumeration	Word-based
EMD	Clustering-based ensemble	Other
GibbsST	Gibbs sampling	PSM
MUSA	Biclustering	Other
GAME	Genetic algorithm	Other
ALSE	Expectation maximization	PSM
MotifSeeker	Data fusion and ranking	Other
PhyloScan	Scanning	PF

3. THE PROPOSED ALGORITHM CONCEPT

Data structure can be seen as a simple representation such as string sequences, but sometimes an ordinary structure it's not enough for the need of the purpose of the algorithm, a need of new data structure is necessary.

Graph structure an example of sophisticated way to represent complex data such as biological component interaction [1], and the mining of such data representation also provides a huge knowledge that can never be seeing in mining simple data structure.

3.1 TREE AUTOMATA

A tree automaton has been designed a long time ago in the context of circuit verification [9], and has since been helpful in a broad range of domains [10].

Tree automata can be defined as a set of stats (initial and final) and a set of rules, a mathematic equation of it [9]:

$$\mathcal{A} = (\mathcal{Q}, \mathcal{F}, \mathcal{Q}_f, \Delta)$$

Where. $\mathcal{Q}$ is a set of stats, $\mathcal{Q}_f \subseteq \mathcal{Q}$ is a set of final stats and Δ is a set of transition rules.

A given tree is accepted on a tree language if it can be produced from q (initial stat) using rules of this language.

3.2 PUSH-DOWN TREE AUTOMATA

The notion of push-down automata has been introduced by guessarian [10], a PDA is a tree automata [11] in which has a stack, a kind of simple memory in which it can store information in a last-in-first-out fashion [12].

A formal definition of PDA can be seen as a seven-tuple [13]:

$$M = (\mathcal{Q}, \Sigma, \Gamma, \delta, q_0, Z_0, F)$$ Where.

- $\mathcal{Q}$: finite set of states
- Σ : finite input alphabet
- Γ : finite alphabet of pushdown symbols
- δ : mapping $\mathcal{Q} \times (\Sigma \cup \{\varepsilon\}) \times \Gamma \to 2^{\mathcal{Q} \times \Gamma^*}$ transition function
- $q_0 \in \mathcal{Q}$: starting/initial state
- $Z_0 \in \Gamma$: start symbol on the pushdown
- $F \subseteq \mathcal{Q}$: set of final states.

Push-down automata can handle strings in which there is some relationship be-tween two symbols, some bioinformatics phenomena, such as looped RNA structures [1].

3.3 AUTOMATA AND BIOLOGICAL DATA

In nature DNA, RNA, and Protein are formed as a chemical compounds that can be represented on a sophisticated structure such as graph, tree, petrinet….

We propose to use pushdown automata as a data (tree) structure for the new algorithm, because pushdown automata is a context free language that can allow us to create a suitable grammar for biological data, and with a stack memory of it, many features of biological data can be represented such as loops of RNA structure.

3.4 MINING PUSH-DOWN AUTOMATA AS A GRAPH STRUCTURE

Mining pushdown automata as a tree structure for biological data can lead to characterize tree sets, discriminating different groups of trees, classifying and clustering trees. Those trees are valuable information of the field of biology.

First of all the algorithm define a grammar for creating a pushdown automata notation that had to be suitable for biological data (DNA, RNA, Protein).

The second step is to transfer a set of motif sequences from a biology bank to a set of pushdown automata based on the grammar on the first step.

The third step, for a given sequence (an input data that will be analyzed) we construct correspond pushdown automata.

Last step is the use of graph mining techniques such as mining frequent sub-tree, or AppriorіGraph [14] to extract a sub trees from the pushdown automata generated from step three and compare all resulted trees with a set of trees generated from the step two in order to highlight the motifs for the given input data.

3.5 ARCHITECTURE

As we mentioned on the beginning of this paper, this is a brief description of a new algorithm, so the architecture in the figure 2 describe a scenario of how this motif discovery algorithm extract motifs from biological sequences.

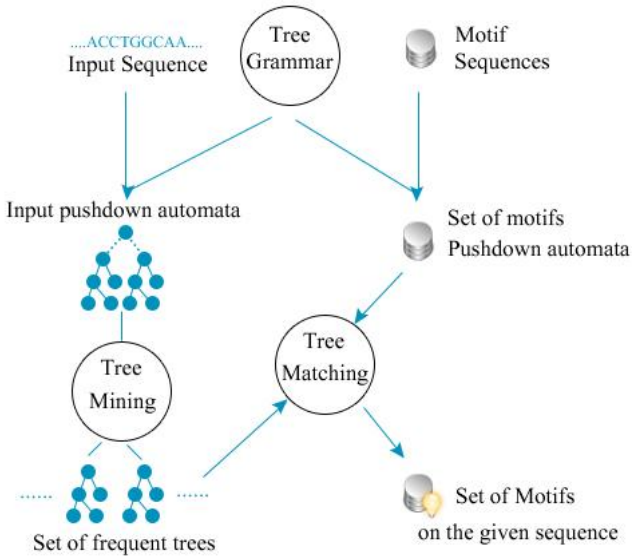

Fig 7: Proposed algorithm architecture.

4 RELATED WORK

The evolution of the field of bioinformatics has increased in observed way, resulting a huge number of techniques and application for the purpose of solving biological problems. Even that many researchers believe that the field of bioinformatics has reached the first phase which is gathering and analyzing data using computer science simple techniques and computational statistic [1]. And it's time to move on to the second phase when the bioinformatics, instead of being informed by just computer science and computational statistics, is also informed by artificial intelligence techniques [1].

One of the most influence factors of motif discovery algorithms is the concept of data structure that can handle a biomolecule data. Graph structure has become increasingly important in modeling sophisticated data especially for bioinformatics [14], and the mining of a structure like that bring a meaningful knowledge to biology. A survey of such kind of mining has been published by Lin et. Al [17]. In the past few years many researches have been proposed to mine this type of data in a wide range of application. Chi et. Al [15] propose an algorithm CmTreeMiner to mine closed subtree. De Amoet. Al [16] is another research for mine tree structure, in this research a new algorithm CoBoMiner for mining tree structure using tree automata as a mechanism to specify user constraints over tree pattern has been proposed, Xeuyiet. Al [18] present a mining RNA tertiary motifs with structure graphs. And still there are a few others proposed algorithms for the context of mining tree or graph structure, also there are a different class of research that describe a way to represent a biological data on tree automata and pushdown automata, Krasiński et. Al [12] has introduced a new model for representing a biomolecule data on pushdown automata.

5 CONCLUSION

This paper has present a small survey of motif discovery algorithms, a step of building algorithm and a categorization for the most widely algorithm on this area, it also mentioned a new notion for categorize motif discovery algorithms.

That's the first part, on the second part a proposed concept of a new algorithm in the field of motif discovery algorithm has been describe, by Taking Advantage of representing bio-logical data on pushdown automata and applied mining tree techniques. On the future work we plan to implement the proposed algorithm and evaluate the performance of it using biological datasets.

6. REFERENCES

[1] Edward Keedwell, Ajit Narayanan, Intelligent Bioinformatics: The Application of Artificial Intelligence Techniques to Bioinformatics Problems, 2005.K. Karu, A.K. Jain, ,"Fingerprint Classification, Proceedings of Pattern Recognition", Vol. 29, No. 3, pp.389-404, 1996.

[2] Luscombe NM, Greenbaum D, Gerstein M, What is bioinformatics? A proposed definition and overview of the field.Schattauer GmbH, 2001.

[3] Venkatarajan Mathura, PandjassarameKangueane, Bioinformatics: A Concept-Based Introduction. Springer. 2009.

[4] alvisbrazma, ingejonassen, ingvareidhammer,david gilbert,Approaches to the Automatic Discovery of Patterns in Biosequences. JOURNAL OF COMPUTATIONAL BIOLOGY. Volume 5, Number 2, 1998.

[5] Modan K Das, Ho-Kwok Dai, A survey of DNA motif finding algorithms, BMC Bioinformatics, 2007.

[6] A. Heger, M. Lappe, and L. Holm. Accurate detection of very sparse sequence motifs. In Proceedings of RECOMB 2003, pages 139–147, 2003.

[7] Vilo J, Brazma A, Jonassen I, Robinson A, Ukonnen E: Mining for putative regulatory elements in the yeast genome using gene expression data. In Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology AAAI Press San Diego, CA; 2000.

[8] Bucher P: Weight matrix description for four eukaryotic RNA polymerase II promoter element derived from 502 unrelated promoter sequences. J MolBiol 1990, 212:563-578.

[9] Hubert Comon, Max Dauchet, Remi Gilleron, Florent Jacquemard, Denis Lugiez,ChristofLodingTison, Marc Tommasi. Tree Automata Techniques and Applications. 2007.

[10] IrrneGuessarian, Pushdown Tree Automata. Math. Systems Theory 16, 237-263 (1983).

[11] Irene GUESSARIAN, On pushdown tree automata, Lecture Notes in Computer Science Volume 112, 1981.

[12] TadeuszKrasiński, Sebastian Sakowski, Autonomous Push-down Automaton Built on DNA, Informatica 36 (2012) 263–276.

[13] DušanKolá, Formal Pushdown Automata, Lecture Formal Pushdown Automata on the 2009.

[14] Jiawei Han, MichelineKamber and Jian Pei, Data Mining: Concepts and Techniques, 3rd Edition. The Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann Publishers, July 2011. ISBN 978-0123814791.

[15] Yun Chi, Yirong Yang, Yi Xia, and Richard R. Muntz. CMTreeMiner: Mining Both Closed and Maximal Frequent Subtrees. In The Eighth Pacific Asia Conference on Knowledge Discovery and Data Mining. 2003.

[16] S. de Amo, N.A. Silva, R.P. Silva, F.S.F. Pereira. Tree Pattern Mining with Tree Automata Constraints. Twenty-second Brazilian Symposium on Databases. 2007.

[17] Lin Shi, Nick Rizzolo. Survey of Graph Mining Techniques. 2005.

[18] Xueyi Wang, Jun Huan, Jack S. Snoeyink, Wei Wang. Mining RNA Tertiary Motifs with Structure Graphs. Scientific and Statistical Database Management, 2007.