13

A New Definition and Look at DNA Motif

Ashkan Sami[1,2] and Ryoichi Nagatomi[2]
[1]*Department of Computer Science and Engineering; Shiraz University; Shiraz 71348;*
[2]*Graduate School of Biomedical Engineering; Tohoku University; Sendai 980-8575;*
[1]*Iran*
[2]*Japan*

1. Introduction

Genetics is the main source of life. The more insights added to the knowledge of genetics, the more accurate prediction and even diagnosis of diseases may become. In genetics, a **sequence motif** is a nucleotide or amino-acid sequence pattern that is widespread and has, or is conjectured to have, a biological significance. Sequence motifs are short, recurring patterns in DNA that are presumed to have a biological function. Often they indicate sequence-specific binding sites for proteins such as nucleases and transcription factors (TF). Others are involved in important processes at the RNA level, including ribosome binding, mRNA processing (splicing, editing, polyadenylation) and transcription termination.

In this study we first concentrate on promoter motifs; however the discussions can easily be extended to other type of sequences. Promoter is a fragment of DNA sequence that is responsible for the transcription from DNA to RNA. Through the study on promoter, it can be found out which DNA sequence will be transcribed into RNA, and even transcription of any DNA sequence which is intended to study into RNA. In bacteria, the target sequence for RNA polymerase attachment is called the promoter. However, in eukaryotes, the term 'promoter' is used to describe all the sequences that are important in initiation of transcription of a gene.

Although no clear definition of motif exist (Pisanti et al., 2005), some define motifs based on statistical representations in the form of PSSM (Position Specific Scoring Matrix) (Gribskov et al., 1987; Hertz & Stormo, 1996; Lawrence & Reilly, 1990; Lawrence et al., 1993). Another school of thought defines a motif as a consensus (Brazma et al., 1998; Vanet et al., 1999). A motif is therefore a pattern that appears repeatedly in a sequence or set of sequences of interest.

The sequences that make up the E-coli promoter were first identified by comparing the regions upstream of over 100 genes. It was assumed that promoter sequences would be very similar for all genes and so should be recognizable when the upstream regions are compared. These analyses showed that the E. coli promoter consists of two motifs described as -10 box - TATA-box or 5′-TATAAT-3′and -35 box - TTG-Box or 5′-TTGACA-3′ (Brown, 2002).

Most of the patterns known to biologist are contingent. In other words, nucleotides that these motifs have are located in consecutive positions. These patterns give more insight to understanding of DNA.

Source: Data Mining in Medical and Biological Research, Book edited by: Eugenia G. Giannopoulou,
ISBN 978-953-7619-30-5, pp. 320, December 2008, I-Tech, Vienna, Austria

In this chapter, a need to a new definition for motif will be provided due to two kind of mentality. First of all by use of a very standard dataset and known concept of independence in statistics, it will be shown why TTG, a very well-known pattern in promoter, is not actually a “valuable pattern”. This illustration leads us to a new direction of defining what actually a motif or pattern is in a DNA sequence like promoter. A measure to find significance based on shape distribution for two-item and multi-item patterns are presented. Later on limitation of the motif evaluation measure to patterns that lead to classification capabilities will be presented. The chapter concludes with summary and future research.

2. Why TTG is not a valuable pattern

Before entering the technical details of the chapter some very simple definitions just to avoid further confusions will be provided.

Frequency: The number of sequences having a specific pattern or property divided by number of all the sequences. Support and frequency are used interchangeably and have the same meaning in this chapter. ‘F’ is used as symbol for frequency. Its subscript presents the type or item. As an example, F_A means frequency of item A and F_{Exc} means ‘excess’ frequency (explained later in this chapter).

Position: position is presented by letter ‘p’ and afterwards an integer follows. The number represents the position with respect to a specific place. Positive integers present the position after the specific place and negative numbers present number of nucleotides prior. For example, in case of the promoters p-36 means the 36th nucleotide prior to Transcriptional Start Site. When p-36 = T is stated, it means at the mentioned position a Thymine exists.

E. coli promoter sequences in UC Irvine – Machine Learning Repository (UC-Irwin MLR) are used. The reason to use the dataset was familiarity of data mining community with the data and ease of access. There are 106 instances of 57 sequential DNA nucleotides (strings consisting of Adenine, Guanine, Cytosine and Thymine) that half of them are sample promoter sequences and the rest are non-promoter sequences. The range of a promoter sequence is starting at p-50 and ending at p7 relative to the Transcriptional Start Site (TSS). It is important to note that position zero does not exist. In other words, the position after position p-1 is p1.

Based on the definition of -35 box motif it is a six nucleotide motif TTGACA. An exact match for the pattern may not exist at position -35 of TSS. However here it will be shown that even for highly observed TTG pattern the pattern is not a valuable pattern. In other words, it is not TTG pattern that presents functionality.

By observing the promoter sequences of the dataset, we will notice that TTG pattern at position -35 occurs with %49.1 frequency. A closer look, although painstaking, reveals patterns with their respected frequencies at Table 1.

Calculation of the previous frequencies by simply counting the occurrence is very difficult. One of the methods to calculate the frequencies of different patterns in data is to convert a sequence to a graph and use of graph-data mining algorithms (Matsuda, et al., 2002). Matsuda et al. use BGI which is a greedy algorithm. Due to its greedy nature some of the pattern may be missed. Use of complete graph data mining algorithms (Yan & Han, 2002; Kuramochi & Karypis, 2001) solves the problem. However, due to NP-completeness of graph-isomorphism checking, the computational complexities of complete graph data mining algorithm are high. In the following section, a much simpler method of finding and calculating the patterns will be presented.

2.1 A simple method of finding patterns and their frequencies

A very simple and effective method of finding all the patterns and their corresponding support in DNA sequences is to use FAF (Sami, 2006; Sami & Takahashi, 2005a). FAF (Finding All Features) uses a special mapping that allows regular Frequent Itemset Mining or Apriori type algorithm (Hipp et al., 2000) be applied to Genetic sequences.

Position -36	*Position -35*	*Position -34*	*Frequency*
T			%81.1
	T		%81.1
T	**T**		%66.0
		G	%79.2
	T	**G**	%64.2
T		**G**	%60.4
T	**T**	**G**	%49.1

Table 1. The frequency of patterns in the promoter sequence

Mapping or pre-processing is one of the main issues that can be treated based on number of main factors. The main purpose of mapping is to come up with a number for each nucleotide in the record that can uniquely represent all the information regarding that main factors of the sequence. The FAF mapping is performed in 5 stages. However before the definition of the mapping, some formal definitions are presented.

A gene in the data set is a sequence R_m, an ordered collection of nucleotides and is represented as $R_m=\{x_1, x_2, \dots ,x_q\}$, where indexes are arranged with regarding to a specific position like transcriptional start site. The alphabet Alpha={A, C, G, T} is used for symbols (x). Each sequence in the data set can be treated as a string. The index of x is of high importance. Records have a class label C is also fixed and known in advance. The class labels are $C = \{C_1, C_2, \dots ,C_t\}$. |C| presents the cardinality of set C. Even though here treated cases had |C| equal to two, the formula is given for generalization purposes. Patterns like $P_i = \{p_1,p_2, ... ,p_n\}$ are desired, where each p_i represents a specific alphabet and i is the index of x that belongs to a unique C_k within the same sequences with a frequency above a given threshold. Now the mapping is as follows:

1. First |R|, |Alpha| and |C| should be considered. In other words, to decide a mapping first the number of outcomes, types, positions, and etc must be calculated.
2. Secondly for |R|, |Alpha| and |C|, k's should be obtained through calculations
   - $k_R = 10^n$ such that $10^{n-1} \leq |R| < 10^n$
   - $k_A = 10^m$ such that $10^{m-1} \leq |Alpha| < 10^m$
   - $k_C = 10^p$ such that $10^{p-1} \leq |C| < 10^p$
3. After calculating the k's, the results based on k_i and |i| where i can be R, A or C must be sorted. As an example, we assume that $k_R > k_A = k_C$,and |R| > |Alpha| > |C| regardless of the fact that the change in order can be easily generalized.
4. Now each value of records, x_i being a_j and belonging to C_t class the mapping is calculated based on Equation 1.

$$p_i = i + j * k_R + t * k_A * k_R \tag{1}$$

5. For each record in the database a unique number will be assigned that can be its order in the database.

The mapping should be done in a sense that each mapped member represents the type, position and class of the nucleotide in the sequence.

2.2 Observation of patterns based on process-oriented mentality

Meaningful patterns should present a combination that the combination by itself presents a functionality or identification. To present the mentality a process-oriented methodology is deployed. Considering occurrence of each nucleotide at specific position as a process, significance of co-occurrence of more than one nucleotide simultaneously at different positions will be will be judged based on the notion of independence. In other words, when two processes are independent of each other, their co-occurrence does not show any specific property.

Co-occurrence of p-36=T and p-35=T is not valuable. Based on Table 1, 66% of sequences have p-36=T and p-35=T, so why is this pattern not valuable or significant? At p-36 and p-35 more than 81% of all sequences have T. The occurrence of each T is completely independent of the other. In more details, two processes a and b are independent if p(ab)=p(a)p(b), where p(a) is the probability of occurrence of a. In case of p-36=T or p-35=T taking frequency directly as probability, it can be seen that 0.81*0.81=0.658 which is almost equal to 0.66. Stated differently, even though 66% of the sequences have T at position p-35 and p-36, considering process oriented mentality reveals that this is not a valuable pattern because the co-occurrence of two T's is independent of each other.

Other combinations of two nucleotides like TG at position -35 and -34 lead to same results (0.811*0.792=0.642). The conclusion can be drawn that no two-nucleotide pattern in TTG is a pattern but is occurring statistically due to high frequency of each of its components..

Since each single two-nucleotide motifs are just statistically occurring patterns due to high frequency of T or G, the discussion can further be extended to conclude that TTG is not a pattern by itself but a pattern due to high frequency of each single nucleotide. In other words, the high frequency of T at position -36 and -35 and G at position -34 lead to the observed co-occurrence of TTG. Having T alone at positions -36 or -35 or G at position -34 is a better indicator of a promoter.

It was shown by use of statistical concept of independence TTG (regardless of its high frequency) is not a significant or valuable pattern. There are several other statistical measures to present interestingness. For a review of the measures refer to (McGarry, 2005; Geng and Hamilton, 2006). Some researchers (Ohsaki et al., 2007) showed that these measures can fairly represent expert needs in a specific domain. It can be shown that deployment of some other measures will also lead to insignificance of TTG.

Next section will discuss another model of finding patterns that can be considered motifs. The main mentality of the new measure basically is based on ideas presented in (Sami, 2006). The measure uses the ranges of values that pattern may have and defines significance and value based on how close the actual value is compare to highest and lowest probable frequencies. By this view, it is the shape of distribution that presents knowledge not purely statistical parameters. In other words, instead of the statistical concepts shape distribution is used.

3. The need for evaluation measure of motifs

In the previous section use of process oriented observation of the frequencies of co-occurrence lead to show that TTG is not an actual pattern. Here a method to evaluate valuableness of the motif with the same mentality but different view is presented.

Basic mentality of the proposed method is based on the idea that the patterns that their support or frequency is comparable to support of single constituents are interesting. In other

words, the frequency of each itemset should not be explainable based on the distribution of frequencies of the two categories that constitute the itemset. As an example: if A and B are two items where: F_A=30%, F_B=35% and F_{AB}=25%
Pattern AB is interesting, since 25% distribution is not explainable based on 30 and 35% (support of A and B). Illustration of frequent pattern evaluation measure is shown in Figure 1 (Good Example) and Figure 2 presents a pattern which is not a valuable pattern.

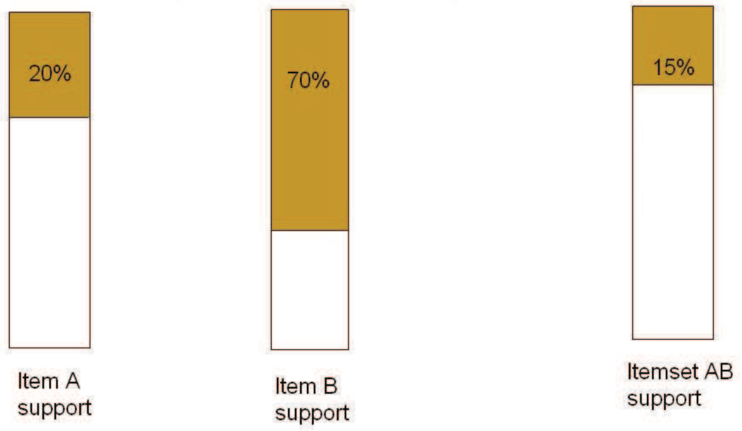

Fig. 1. Illustration of frequent itemset which is valuable

Figure 1 presents a valuable pattern since the frequency of the pattern AB is comparable to the frequency of the lowest frequent item of the pattern. In other words, it could have been possible to have item B's distributed in a sequences that no occurrence of AB existed. However, 15% of 20% of item A co-occur with item B which makes a valuable pattern. In contrast in Figure 2, item B is highly frequent and due to high frequency of item B it is impossible to have items A and B occur together less than 15%.

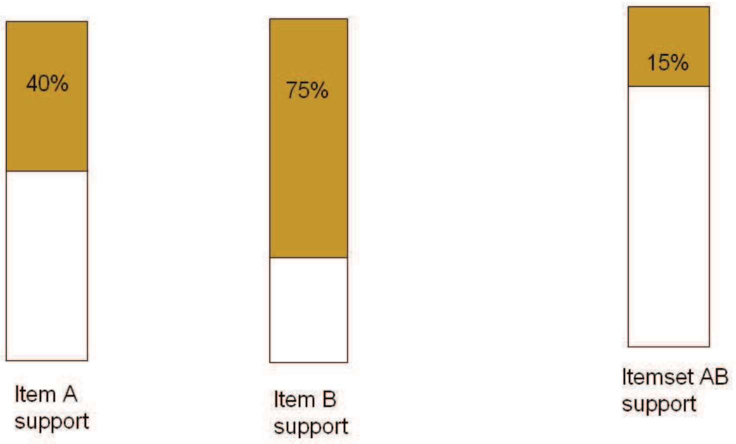

Fig 2. Illustration of frequent pattern which is not significant

In the following section evaluation measure for two-item patterns will be presented.

4. Evaluation measure for two-item patterns

As illustrated in section 3 of this chapter, lowest frequency of occurance of A or B should be comparable to frequency of sequences containing AB. The closer the frequency of sequences containing AB to the minimum frequency of sequences having A or B, the more valuable the pattern.

In case of having two items their frequencies summation can be more than unit, below than one or exactly one. With respect to this summation, a two-item pattern that frequencies of its items add up to be more 100% is defined as having excess frequency. In contrast, two-item itemsets that the summation of frequencies of each item add up to less than one will be defined as a pattern without excess frequency. If the frequencies of the two constituantes of two-item pattern add up to one, the case can be categorized as either of the two. To deeply understand the derivation of the measure based on shape distribution the cases of having excess frequency and not having an excess frequency will be treated separately.

4.1 Evaluation measure for two-item patterns with excess frequency

As a definition for itemset AB, excess frequency exists if summation of frequencies of A and B exceeds 100%. Now a detail discussion on how to reach the measure is presented by an example.

Considering the pattern presented in Figure 2, frequency of A is 40% and B is 75%. Since F_A+F_B = 0.4+0.75 = 1.15 (the summation of the frequencies exceeds 100%), excess frequency exists. The excess frequency is equal to 15%. Due to having excess frequency, at least 15% of the sequences will have pattern AB. It is impossible to have a sequence that has only A or B exclusively. At least 15% of the sequences have AB in their sequences. Stated differently, regardless of shape of distribution of A and B among sequences at least 15% overlap exists. Therefore 15% support is least significant value for the itemset.

In contrast, the pattern is most prominent if AB occurs with its highest possible frequency or near the minimum frequency of its constituants. Since minimum of the frequency of A and B is 40%, pattern AB would have been most meaningful if AB had occured with 40%. It is obvious that frequencies more than 40% is not possible. This is a knowm property in KDD community and was first introduced by Agrawal and Srikant (Agrawal and Srikant, 1994) as “downward closure property”.

Due to downward closure property of co-occuring patterns, maximum frequency would be equal to the lowest frequency of the two items. Minimum of each item in the itemset frequencies is the frequency of A which is 40%. In other words, the maximum frequency of sequences that have both A and B cannot be more than 40%. The closer the F_{AB} to 40%, the more significant the pattern.

Since the relationship is based on shape of distribution of A and B, a linear form can be considered suitable. The evaluation measure would rank pattern with frequency of excess as not significant and with frequency of lowest frequency of the two items as the most valuable. Zero presents not valuable and one most valuable. So by a linear relationship the degree of significance in the relationship is assessed.

More specifically, if the excess frequency is presented as F_{Exc}, minimum of frequencies of A and B is presented as F_{Max}, (F_{Max} = min(F_A, F_B)). Minimum of the frequencies is called max frequency since this is the highest possible value for AB support. The probable values of frequency of itemset AB may start from minimum value of F_{Exc} and reaches the maximum of F_{Max}. Therefore interestingness or significance of two-item pattern of AB based on shape distribution is the position of the actual frequency of AB on a simple line connecting (0, F_{Exc}) to (1, F_{Max}) or:

$$S = (F_{AB} - F_{Exc}) / (F_{Max} - F_{Exc}) \qquad (2)$$

where F_{AB} is the actual value of support of AB.
Following the example, S=0 for 15% frequency for AB pattern
As another example, if F_A=50% and F_B=60% and F_{AB}=30%, it is clear 10% excess frequency exists. Thus,
S=(30-10)/(50-10) = 0.5, which makes sense since 30% is in the middle of F_{Max} and F_{Exc}

4.2 Evaluation measure for two-item patterns without excess frequency

In case of frequencies of items of an itemset that do not add up to more than 100%, the maximum will not differ. In other words, in a two-item itemsets the highest frequency will be equal to the frequency of the lowest frequent item, F_{Max}. However the minimum is definately equal to zero.
To develop the measure consider two cases. In case I, F_A = 30% and F_B = 32%. In case II F_A is the same as case I but F_B = 60%. In either case the frequeny of itemset AB ranges from zero to 40%. However it does make sense to consider pattern F_{AB} = 20% in case I more valuable than case II. In other words, higher frequncy of B in case II increases the likelihood of having sequences with AB occuring in them. Therefore a hypothetical chance of having negative frequency to compensate for the effect is considered. Negative frequency is the value need to make the summation of frequencies equal to one or 100%.

$$F_{Neg} = 1 - (F_A + F_B) \tag{3}$$

As an illustration, in case I from the previous paragraphs, F_{Neg} = 100-(30+32) = 38% and in case II, F_{Neg} = 100-(30+60) = 10%. Again a linear relationship is considered between the amount of negative frequency and maximum as significance for negative frequency equal to zero and maximum frequency as one. Significance is defined as the value of F_{AB} on the line connecting (0, F_{Neg}) to (1, F_{Max}). Thus,

$$S = (F_{AB} + F_{Neg}) / (F_{Max} + F_{Neg}) \tag{4}$$

Going back to the case I and II of previouse paragraphs,
Significance for case I: S = (20+38)/(30+38) = 0.853. Where significance for case II: S = (20+10)/(30+10) = 0.75. These numbers somehow present the fact that regardless of same motif frequency the pattern is more valuable in case I than case II.
Close observation of measures derived in section 4 reveals that we can use Equation 3 instead of Equation 4 if F_{Neg} was actually a negative number. Therefore, only one equation exists for two-item patterns.

5. Evaluation measure for patterns having more than two items

Again the mentality behind the measure is same as before. Significance is measured with respect of how unlikely it is to have the pattern. The lower the chance of having the pattern, the higher the significance.
Due to downward closure property, maximum value for frequency of pattern with several constituents is equal to the lowest frequency among all the items. It is intuitive to consider concept of F_{Neg} one more time. F_{Neg} presents the freedom of choices that may lead to no pattern. The same type of concept is extended to consider multiple factors involving a multi-item pattern, F_{NegMul}. Definitely the value that F_{Neg} type of parameter has should increase in magnitude as the number of items increases.

Let's start with an example, assuming A has frequency of 30%, B 40% and C 50%. How can a value be given to the significance of frequency of pattern ABC?
The maximum possible frequency would be the lowest frequency of the items. Thus, F_{Max} = 30%. What number should be assigned to present the lower possibility of having ABC?
First defining $F_{\sim A}$ as the frequency of sequences not having A. If BC occurs in ~A sequences, no ABC pattern would exist. Thus negate of a pattern provides a constraint on construction of the pattern. Obviously the greater the $F_{\sim A}$, the less likely existence of ABC. On the same token, ABC is more valuable when F_A, F_B and F_C are all small. So, ~A, ~B and ~C are profound factors. To present significance value as a linear relationship as before, the lower bound must be calculated in a way that increase of number of items departs it further away from F_{Max}. A simple extension of negative frequency would be based on linear assumption of increasing chance that is presented as follows; if F_{Xi} presents the frequency of x_i, where x_i is *i*-th item of an n-item pattern (*i*=1,…,n) and

$$F_{x_k} = \min\left(F_{x_1},...,F_{x_n}\right) \tag{5}$$

$$F_{NegMul} = \sum_{i=1\&i\neq k}^{n} F_{\sim x_i} - F_{x_k} = \sum_{i=1}^{n} F_{\sim x_i} - 100 \tag{6}$$

Then the significance based on linear relationship would be evaluated based on where pattern frequency on the line of connecting (F_{NegMul},0) and (F_{Max}, 1) lies or:

$$S = (F_{Pattern} + F_{NegMul}) / (F_{Max} + F_{NegMul}) \tag{7}$$

Going back to our example, instead of x_i's, A, B, and C exist.

$$\begin{aligned} F_{NegMul} &= \sum_{i=1}^{n} F_{\sim x_i} - 100 \\ &= (100-30)+(100-40)+(100-50)-100 = 80 \end{aligned} \tag{8}$$

In case of TTG box;

F_{NegMul} = (100-81.1)+(100-81.1)-79.2 = -41.4
S= (49.1-41.4) / (79.2-41.4) = 0.20

Thus, TTG pattern is not significant even based on a more relax measure of shape distribution in comparison to regular independence in statistics.

6. Motif based on functionality

It is important to note that all that has been said was with respect to a specific sequence namely promoter in E. Coli. The purpose of viewing the motif in other situations can lead to different definitions. In other words, the distinction between sequences was not considered. All the evaluation measures discussed so far are not suitable for classification purposes. Viewing motifs with classification capabilities may lead to different motifs. This issue has been addressed to some extend by some researchers in graph data mining community (Geamsakul et al., 2003a and 2003b). As stated before graph data mining algorithms are either greedy and fast or complete and very slow. Another approach to motif discovery with classification capability is based on the mapping of FAF discussed (Sami & Takahashi, 2005b).

7. Conclusions and future research

In this chapter a close look at genetic motifs especially TTG-box or -35 box was provided. Based on statistical measure of independence it was shown that TTG box is not actually a significant pattern. Based on statistical notion of independence, it was shown that in TTG if occurrence of each nucleotide is considered as a process are completely independent of each other. Afterwards, another view that focused on shape distribution was deployed. Again after developing the model and measure, it was shown that the TTG pattern is not valuable. Even though TTG has near 50% support; it has a low frequency with respect to its constituents' frequencies.

In addition to use of bigger datasets, this research can be extended in two major ways. As suggested in the chapter, deployment of other interestingness measures to reach the same results or similar is one direction. Secondly, devising non-linear measures of significance based on shape distributions that form in high dimensional space of multi-item patterns.

8. References

Agrawal, R. and Srikant, R. (1994). Fast algorithms for mining association rules. In Proc. of *the 20th Int'l Conf. on Very Large Databases (VLDB' 94),* Santiago, Chile.

Brazma, A.; Jonassen, I.; Eidhammer, I. and Gilbert, D. (1998). "Approaches to the Automatic Discovery of Patterns in Biosequences," *Journal of Computational Biology*, vol. 5, pp. 279-305, 1998.

Brwon, T.A., (2006). *Genomes*. Garland Science, Taylor & Francis Group, May 2006, ISBN: 9780815341383

Geamsakul, W.; Matsuda, T.; Yoshida, T.; Motoda, H. and Washio. T. (2003a). Classifier construction by graph-based induction for graph-structured data. *In PAKDD'03: Proc. of 7th Pacific-Asia Conference on Knowledge Discovery and Data Mining,* LNAI2637, pp. 52--62.

Geamsakul, W.; Matsuda, T.; Yoshida, T.; Motoda H. and Washio, T. (2003b). Constructing a Decision Tree for Graph Structured Data, P*roc. of First International Workshop on Mining Graphs, Trees and Sequences (MGTS-2003), 14th European Conference on Machine Learning (ECML'03) and 7th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'03),* pp.1–10.

Geng, L. and Hamilton, H.J. (2006). Interestingness measures for data mining—a survey, *ACM Comput Surveys,* Vol. 38, No. 3, pp. 1-32.

Gribskov, M., McLachlan, A. and Eisenberg, D., (1987). Profile Analysis: Detection of Distantly Related Proteins, *Proc. Nat'l Academy of Sciences*, vol. 84, no. 13, pp. 4355-4358.

Hertz, G.Z. and Stormo, G.D. (1996). "Escherichia Coli Promoter Sequences: Analysis and Prediction," *Methods in Enzymology*, vol. 273, pp. 30-42.

Hipp, J.; Güntzer, U. and Nakhaeizadeh, G. (2000). Algorithms for Association Rule Mining A General Survey and Comparison, *SIGKDD Explorations,* vol. 2, no. 1, July 2000, pp. 58-64.

Kuramochi, M. and Karypis, G. (2001). Frequent Subgraph Discovery, *Proceedings of the 2001 IEEE International Conference on Data Mining*, November 29-December 02, 2001, pp. 313-320.

Lawrence, C.E.; Altschul, S.F.; Boguski, M.S.; Liu, J.S.; Neuwald, A.F. and Wooton, J.C. (1993). "Detecting Subtle Sequence Signals: A Gibbs Sampling Strategy for Multiple Alignment," *Science*, vol. 262, pp. 208-214, 1993.

Lawrence C.E. and Reilly A.A., (1990). "An Expectation Maximization (EM) Algorithm for the Identification and Characterization of Common Sites in Unaligned Biopolymer Sequences," *Proteins: Structure, Function, and Genetics*, vol. 7, pp. 41-51, 1990.

McGarry, K. (2005). A survey of interestingness measures for knowledge discovery, *The Knowledge Engineering Review*, Vol. 20 No. 1, March 2005, pp.39-61.

Matsuda, T. Motoda, H. and Washio. T. (2002). Graph-based induction and its applications. *Advanced Engineering Informatics*, Vol. 16 No. 2, pp:135−143, 2002.

Ohsaki, M., Abe, H., Yokoi, H., Tsumoto, S., Yamaguchi, T. (2007). Evaluation of rule interestingness measures in medical knowledge discovery in databases. *Artificial Intelligence in Medicine*, Vol. 41, No. 3, pp. 177−196.

Pisanti, N.; Crochemore, M.; Grossi, R. and Sagot, M.F. (2005). Bases of Motifs for Generating Repeated Patterns with Wild Cards. *IEEE/ACM Transaction on Computational Biology and Bioinformatics*, Vol. 2, No. 1, JANUARY-MARCH 2005

Sami, A. (2006). *Knowledge Discovery in Biomedical Sciences Based on Shape Distribution Methods*. Ph.D. Thesis; August 2006; Tohoku University; Sendai, Japan.

Sami, A. and Takahashi, M. (2005a). "FAF: Finding All Features Relating to Different Gene Sequences" *Workshop on Knowledge Discovery and Data Management in Biomedical Sciences, (KDDMBS 2005) in conjunction with PAKDD*, pp. 4-13; 18 May 2005; Hanoi, Vietnam.

Sami, A. and Takahashi, M. (2005b). Decision Tree Construction for Genetic Applications based on Association Rules, *IEEE TENCON 2005*, Melbourne, Australia, November 2005, pp.21-25.

UC-Irwin MLR., University of California at Irwin – Machine Learning Repository; http://www.ics.uci.edu/~mlearn/MLRepository.html, Last visited March 2006.

Vanet, A.; Marsan, L. and Sagot, M.F. (1999). "Promoter Sequences and Algorithmical Methods for Identifying Them," *Research in Microbiology*, vol. 150, pp. 779-799, 1999.

Yan X. and Han, J. (2002). "gSpan: Graph-Based Substructure Pattern Mining," *Proc. Int'l Conf. Data Mining,* (ICDM 2002), December 2002, pp. 721-724.

Data Mining in Medical and Biological Research

Edited by Eugenia G. Giannopoulou

ISBN 978-953-7619-30-5

Hard cover, 320 pages

Publisher InTech

Published online 01, November, 2008

Published in print edition November, 2008

This book intends to bring together the most recent advances and applications of data mining research in the promising areas of medicine and biology from around the world. It consists of seventeen chapters, twelve related to medical research and five focused on the biological domain, which describe interesting applications, motivating progress and worthwhile results. We hope that the readers will benefit from this book and consider it as an excellent way to keep pace with the vast and diverse advances of new research efforts.

INTECH

open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821