Discovering Matrix Attachment Regions (MARs) in Genomic Databases

Gautam B. Singh
Department of Computer Science & Engineering
Oakland University
Rochester, MI 48309
singh@oakland.edu

ABSTRACT

Lately, there has been considerable interest in applying Data Mining techniques to scientific and data analysis problems in bioinformatics. Data mining research is being fueled by novel application areas that are helping the development of newer applied algorithms in the field of bioinformatics, an emerging discipline representing the integration of biological and information sciences. This is a shift in paradigm from the earlier and the continuing data mining efforts in marketing research and support for business intelligence. The problem described in this paper is along a new dimension in DNA sequence analysis research and supplements the previously studied stochastic models for evolution and variability. The discovery of novel patterns from genetic databases as described is quite significant because biological pattern play an important role in a large variety of cellular processes and constitute the basis for gene therapy. Biological databases containing the genetic codes from a wide variety of organisms, including humans, have continued their exponential growth over the last decade. At the time of this writing, the `GenBank` database contains over 300 million sequences and over 2.5 billion characters of sequenced nucleotides. The focus of this paper is on developing a general data mining algorithm for discovering regions of locus control, i.e. those regions that are instrumental for activating genes. One type of such elements of locus control are the MARs or the Matrix Association Regions. Our limited knowledge about MARs has hampered their detection using classical pattern recognition techniques. Consequently, their detection is formulated by utilizing a statistical *interestingness* measure derived from a set of empirical features that are known to be associated with MARs. This paper presents a systematic approach for finding associations between such empirical features in genomic sequences, and for utilizing this knowledge in detecting biologically interesting control signals, such as MARs. This computational MAR discovery tool is implemented as a web-based software called *MAR-Wiz* and is available for public access. As our knowledge about the living system continues to evolve, and as the biological databases continue to grow, a pattern learning methodology similar to that described in this paper will be significant for the detection of regulatory signals embedded in genomic sequences.

General Terms

Medical Data Mining, Bioinformatics

Keywords

Data Mining, Gene Therapy, Matrix Attachment Regions, MARs, DNA Sequence Analysis

1. BACKGROUND

Data mining is a multidisciplinary field spanning several disciplines including statistics, computer science, pattern recognition, artificial intelligence, machine learning, and others. This enables the data mining research to draw from a plethora of existing research in a manner that represents an integration of multiple perspectives. Rapid advances in statistical tools have made many aspects of data analysis routine – as a result our focus is shifting towards higher level issues such as the type of questions we should focus on answering, the type of analysis that would be most suited to the task at hand, and how the knowledge discovered would be assimilated to possibly revise any of our existing beliefs. Our focus is also shifting towards incorporating the analyst's knowledge and strategies within the data mining tools. This in a way is providing the requisite knowledge to the system so that an appropriate criteria for measuring *interestingness* is possible to establish. The problem described in the paper is a problem where a rich set of domain knowledge must be integrated into the knowledge discovery algorithm. Specifically, we are interested in discovering Matrix Attachment Regions or MARs from the raw DNA sequence data available in genomic databases.

The genetic program that forms the basis of life is stored inside the nucleus in eukaryotes (multi-cellular organisms, such as humans, primates, etc.) as a linear macromolecule comprised of nitrogenous *bases*. This large macromolecule is also known as DNA or deoxy-ribonucleic acid. A DNA molecule is comprised of four *bases*, namely, Adenosine (A), Cytosine (C), Thymidine (T) and Guanine (G). The DNA bases occur in pairs, with the base A always pairing with T, and the base C pairing with G. Accordingly, the length of a DNA macromolecule, collectively referred to as the *genome*, is measured as *base-pairs* or *bp*. An extensive multidisciplinary endeavor to identify each base in the 3×10^9 bp long human genome has been ongoing for over a decade. The genome stores the blueprints for the synthesis of a variety of proteins – the macromolecules that enable an organism to be structurally and functionally viable. The blueprint or the program for the synthesis of a single protein is called a *gene*, a unit of the DNA sequence that is generally between 1×10^3–1×10^6 *bp* in length based upon the complexity of the protein that it *codes* for. The process of synthesizing a protein using its genetically coded blueprint is known as *gene expression*. A higher level eukaryote may contain as many as 30,000-40,000 genes[1]. It has also been estimated that only about 1000-10,000 genes may be expressed by a given cell. Although each cell in an organism contains the same genetic program, the subset of genes expressed in one cell type is different from another and is determined by the functional role of that cell. Apart from the gene *coding* regions that account for 10-20% of the genome, the majority of nuclear DNA is *non-coding*. However, it appears to be important as researchers believe that the genetic program is able to regulate the expression of genes based upon the biologically significant DNA sequence patterns that are primarily observed in the non-coding regions within the neighborhood of a gene [13].

The success of the Human Genome Project (HGP) is dependent upon a continued effort to develop databases and tools for easy access and comparison of genomic information. Several molecular biology databases have been created that contain diverse information on biological data such as the annotated DNA and protein sequences, 3D-structures, genetic and physical maps [11]. As the first phase of the Human Genome Project aimed at determining the identity of the 3×10^9 characters of the human genome is nearing completion by year 2003 A.D., the bioinformatics focus is shifting toward developing computational tools and algorithms that can assist in the analysis, interpretation and discovery of knowledge contained within this data. At the time of this writing, the main genomic database, `GenBank`, contains about 2.5 billion characters in DNA sequence data (excluding the annotations), in approximately 300 million DNA sequence records. The next phase is expected to focus on completing the transcript map and understanding the functional significance of the genes sequenced during Phase I. During this phase the elements of locus control, such as the MARs or the Matrix Association Regions, will be sought and their localization in genetic sequences will be established in order to facilitate the understanding of genetic processes. This paper describes a data-mining approach based on the information theoretical measure of *interestingness* to discover regions of matrix association from the DNA sequences in genomic databases.

The DNA is *not* merely a homogeneous string of characters {`A`,`C`,`T`,`G`}, but is really comprised of a mosaic of sequence level motifs that come together in a synergistic manner to define the state of that organism. Special sequences of regulatory importance such as introns, promoters, enhancers, and repeats are found on the DNA, and often contain patterns that represent functional control points within the genome [10]. Some *repetitive* DNA patterns serve as a biological clock and bring about the *apoptosis* or programmed cell death. Other examples of these patterns include the $A + T$ or $G + C$ rich regions, telomeric repeats of sequence AGGGTT in human DNA, rare occurrence or absence of dinucleotides TA and GC, and tetranucleotide CTAG, and the GNN periodicity in the gene coding regions. There is evidence that suggests that some deviations from patterns are deleterious to the viability of the organism. These and numerous other examples indicate that the *patterns* embedded in the eukaryote DNA may play a vital role in its viability.

[1]Human genome is estimated to contain about 80,000-100,000 genes

2. BIOLOGICAL FEATURES

The Matrix Attachment Regions are relatively short (100-1000 bp long) sequences of functional importance and anchor the chromatin loops to the nuclear matrix. MARs have been shown to be responsible for gene expression and serve as the origins of replication (ORI) for cell division [2]. MARs have been observed to flank the ends of genic domains encompassing various transcriptional units. It has also been shown that MARs bring together the transcriptionally active regions of chromatin such that transcription is initiated in the region of the chromosome that coincides with the surface of nuclear matrix. Approximately 100,000 matrix attachment sites are believed to exist in the human nucleus, a number that is approximately equal to the expected number of genes [1].

The classical pattern recognition approach for detection utilizes the characteristic properties or features of the object being sought for detection. For example, characters are often recognized using their area and perimeter, by their compactness (i.e. the ratio of their area to the square of their perimeter), or by the degrees of symmetry about the horizontal and vertical axes. Clearly, the features needed depend on the specific problem domain and the design of an appropriate set of features is more of an art and is crucial to the overall success of the classification problem. Often we can think of an unknown observation as being a point in a *k-dimensional* feature space, and develop a feature extractor for the observation $\mathbf{X}$. In an ideal case, the feature extractor would generate the same feature vector $\mathbf{X}$ for all observations in the same class, and distinct feature vectors for observations in different classes. In classical pattern recognition the goal is to design the classifier, given that feature set extracted from the various classes. However, in data mining, our goal to classify the observations in such a manner that the features extracted from observations that are co-classified have similar feature vectors. As described below, the feature vectors utilized for MARs are based upon a vast body of biological research.

MARs have been experimentally defined for several gene loci. However, researchers have not identified a single DNA level pattern that characterizes a MAR. The investigation of MARs in the human interferon-β gene, human β-globin gene[9], human p53, the human protamine gene cluster, and the chicken α-globin and lysozyme genes, have helped in characterizing a *set of patterns* that are found in the vicinity of MARs, although a consensus sequence has not been apparent.

The following set of patterns have been known to be associated with the presence of Matrix Association Regions:

- *The Origin of Replication (ORI) Rule:* It has been established that replication is associated with the nuclear matrix, and the origins of replication share the `ATTA`, `ATTTA` and `ATTTTA` motifs.
- *Curved DNA:* Curved DNA has been identified at or near several matrix attachment sites and has been involved with DNA-protein interaction, such as recombination, replication and transcription [2; 23]. Opti-

mal curvature is expected for sequences with repeats of the motif, $AAAAn_7AAAn_7AAAA$ as well as the motif $TTTAAA$. The term n_7 denotes the occurrence of seven nucleotides of unspecified identity.

- *Kinked DNA:* Kinked DNA is typified by the presence of copies of the dinucleotide TG, CA or TA that are separated by 2–4 or 9–12 nucleotides. For example, kinked DNA is recognized by the motif TAn_3TGn_3CA, with TA, TG and CA occurring in any order.

- *Topoisomerase II sites:* It has been shown that Topoisomerase II binding and cleavage sites are also present near the sites of nuclear attachment. Vertebrate and Drosophila topoisomerase II consensus sequence motifs can be used to identify regions of matrix attachment.

- *AT-Rich Sequences:* Typically many MARs contain stretches of regularly spaced AT-rich sequences in a periodic manner.

- *TG-Rich Sequences:* Some T-G rich spans are indicative of MARs. These regions are abundant in the 3'UTR of a number of genes, and may act as recombination signals [2].

- *Consensus Motif:* The sequence defined as the nuclear matrix STAB-1 binding motif: `TCTTTAATTTCTAATATATTTAGAA`, and

- *ATC Sequences:* ATC rule (a stretch of 20 or more occurrences of H, i.e. A or T or C). This rule is an effective indicator of regions with marked helix destabilization potential often associated with MARs.

The process of discovering MARs in genomic database is described below. In the first step, each DNA pattern known to be associated with the MARs is represented using logical disjunction of individual pattern conjunction. With each such `AND-OR` rule, a statistical relevance score is next associated based upon the probability that the pattern might be observed at random. The individual pattern scores are added to generate the overall relevance of a region and used to assess its possibility for being a MAR site.

3. PATTERN DEFINITION AND RELEVANCE SCORES

In such a general framework, a pattern description language is defined that has sufficient power to represent the variety of patterns that characterize the MARs. It is possible to employ a general set of DNA patterns using the `AND-OR` methodology. In such an `AND-OR` pattern specification methodology a disjunction (`OR`) of the conjunctions (`AND`) of the motifs detected in the sequence is used as the definition of the pattern being sought. The sequence level motifs serve as the lowest level *predicates* used to detect the presence of a higher level pattern. In general the following operations may be applied to the lower level motifs:

- Pattern motif sequence, m, represented as a regular expression, or

- The logical `OR` of two motifs m_i and m_j, represented as $m_i \vee m_j$, or

Motif Index	*Motif Name*	*Sequence Predicate*
m_1	ORI Signal	$ATTA$
m_2	ORI Signal	$ATTTA$
m_3	ORI Signal	$ATTTTA$
m_4	TG-Rich Signal	$TGTTTTG$
m_5	TG-Rich Signal	$TGTTTTTTG$
m_6	TG-Rich Signal	$TTTTGGGG$
m_7	Curved DNA Signal	$AAAAn_7AAAAn_7AAAA$
m_8	Curved DNA Signal	$TTTTn_7TTTTn_7TTTT$
m_9	Curved DNA Signal	$TTTAAA$
m_{10}	Kinked DNA Signal	TAn_3TGn_3CA
m_{11}	Kinked DNA Signal	TAn_3CAn_3TG
m_{12}	Kinked DNA Signal	TGn_3TAn_3CA
m_{13}	Kinked DNA Signal	TGn_3CAn_3TA
m_{14}	Kinked DNA Signal	CAn_3TAn_3TG
m_{15}	Kinked DNA Signal	CAn_3TGn_3TA
m_{16}	mtopo-II Signal	$RnYnnCnnGYnGKTnYnY$
m_{17}	dtopo-II Signal	$GTnWAYATTnATnnR$
m_{18}	AT-Rich Signal	$WWWWWW$
m_{19}	Consensus Motif	$TCTTTAATTT-$ $CTAATATATTTAGAA$
m_{20}	ATC Rule	H_{20}

Figure 1: Table of sequence level motifs: The set of motifs characterizing MARs constitute DNA-sequence signals or predicates upon which the rules defining higher level patterns are constructed. Note that the IUPAC characters R,Y, W and K are defined as: R=A `or` G, Y=T `or` C, W = A `or` T, and K=G `or` T.

- The augmented logical `AND` of two motifs m_i and m_j, represented as a $m_i \wedge_a^b m_j$[2], or

- The logical negation of a motif, m, represented as $\overline{m}$, specifying the absence of a given motif.

The pattern specification methodology must account for the motif variability in its representation. As an example, consider the rule to define the Origin of DNA Replication. This can be based on an `OR` or the $\vee$ operator applied to the three motifs m_1 =`ATTA`, m_2 =`ATTTA`, and m_3 =`ATTTTA`.

$$R_1 = m_1 \vee m_2 \vee m_3 \tag{1}$$

Similarly, the requirement for multiple motif occurrences can be specified using the `AND` or the $\wedge$ operator. An additional parameter is incorporated when using the `AND` rule to constrain the allowable gap between the two co-occurring motifs. For example, the AT-Richness rule, R_6, can be formulated as the occurrence of two hexanucleotide strings, m_{18}=`WWWWWW`[3], that are separated by distance of 8–12 nt, using the augmented `AND` operator using $\bigwedge_{low}^{high}$ to define the acceptable distance between the two motifs:

$$R_6 = m_{18} \bigwedge_{8}^{12} m_{18} \tag{2}$$

[2]In this augmented `AND` operator, the parameters a and b specify the acceptable separation between the co-occurrence of the two motifs

[3]Note: The IUPAC code `W` denotes the base `A` or `T`

The **relevance score** of a pattern is defined as $\log_2(p_i)$, where p_i is the probability that the pattern defined by the `AND-OR` rule R_i will occur at random. The base composition of the sequences being analyzed is considered for this purpose [21]. For an `OR` type of pattern, the composite probability of the pattern is derived by addition of the probabilities of the underlying patterns. This is possible since the occurrences of the underlying patterns are mutually exclusive by definition for a given start position on the DNA sequence. For an `AND` type of pattern, this value is derived using that the occurrence of *at least one motif* within an acceptable distance from the reference motif will satisfy the pattern occurrence constraint. In general for a pattern rule of the type, $m_x \bigwedge_u^v m_y$, this probability will be equal to $P(m_x) \cdot (1.0 - e^{-(v-u+1)P(m_y)})$. This is derived by assuming the occurrence of m_y as a Poisson's process, and deriving the probability of least one occurrence of m_y by subtracting from 1.0 the probability that no occurrences of m_y will be observed in the interval $(v - u + 1)$.

The significance of the occurrence of a *pattern* in a DNA sequence is inversely related to the probability that the pattern will occur purely by chance. This is captured by the *unexpectedness* term [22] that is based on Shannon's information theory. The unexpectedness for a rule R_i is equal to $\log_2(Pr(R_i))$. In a certain sense, this information content for a pattern defined by a rule is equal to the amount of uncertainty that the occurrence of that rule removes about the identify of a specific region as being MAR. (In general, the regions are considered to be random samples from the DNA). Naturally, the patterns that occur less frequently at random have a larger contribution in removing this uncertainty. The unexpectedness values for the various rules are defined in Table 2.

Rule	*Name*	*Definition*	*Unexpectedness*
R_1	ORI Rule	$m_1 \vee m_2 \vee m_3$	$p_1 = \sum_{i=1}^{3} Pr(m_i)$
R_2	TG-Richness Rule	$m_4 \vee m_5 \vee m_6$	$p_2 = \sum_{i=4}^{6} Pr(m_i)$
R_3	Curved DNA Rule	$m_7 \vee m_8 \vee m_9$	$p_3 = \sum_{i=7}^{9} Pr(m_i)$
R_4	Kinked DNA Rule	$m_{10} \vee m_{11} \vee m_{12} \vee$ $m_{13} \vee m_{14} \vee m_{15}$	$p_4 = \sum_{i=10}^{15} Pr(m_i)$
R_5	Topoisomerase Rule	$m_{16} \vee m_{17}$	$p_5 = \sum_{i=16}^{17} Pr(m_i)$
R_6	AT-Richness Rule	$m_{18} \ \wedge_8^{12} \ m_{18}$	$p_6 = Pr(m_{18}) \cdot$ $(1 - \exp(-5 \cdot Pr(m_{18})))$
R_7	Consensus Rule	m_{19}	$-\log_2(Pr(m_{19}))$
R_8	ATC Rule	m_{20}	$-\log_2(Pr(m_{20}))$

Figure 2: The set of biological rules defining patterns that were used for detecting structural MARs. The table also specifies the unexpectedness associated with each MAR rule defined using the combination of lower level patterns. The unexpectedness values form the basis for mining MARs from DNA sequences.

4. DATA MINING THE MATRIX ATTACHMENT REGIONS

Data mining efforts aim at detecting statistically significant patterns, that are useful to the user as they are not redundant, novel in regards to user's previous knowledge, simple for the user to understand, and sufficiently general to the referred population [5; 15; 3; 14; 19; 18; 4; 6]. When searching for patterns, one must strive for a balance between the *specificity* and their *generality*. Correspondingly, a distinction is often drawn between the tasks of finding patterns and that of finding models. Generally, a model is a global representation that summarizes the phenomenon underlying the data and aims at describing how the data may have arisen. The methods aimed at building global models fall within the category of statistical exploratory analysis [8].

In contrast, *patterns* are local representations that characterize a smaller number of cases using a subset of variables that might be utilized for global models. As an example, local patterns are often sought in the time-series data analysis [12], as in the case of analysis of daily stock prices with the objective of detecting unusual market behaviour. These methods seek patterns by sifting through the data, measuring co-occurrences of specific values of specific variables, and seek *nuggets* of information amongst the mass of otherwise voluminous data. The problem of detecting MARs from anonymous DNA sequence data falls within this category, and relies on statistically measuring the interestingness of a region from the standpoint of characterizing it as a MAR.

The discovery of MARs is based on the observation that a group of patterns are bonded together by the virtue of their similar function. After such a grouping, a search for the patterns in a given group can be performed to identify these regions in the query DNA sequence. If a large subset of members of a functionally related group of patterns is found in a given region of the DNA sequence, one can justifiably classify it as a MAR. The detection problem is to identify these regions by estimating the statistical significance of the observation within a given region. The discovery of these regions of high pattern density is achieved by sliding a window along the DNA sequences and measuring the statistical *unexpectedness* values for the patterns observed within it. In the context of our problem domain of detecting MARs, this unexpectedness value is referred to as the *MAR-Potential* or ρ. The pattern-density is measured within a window of size W centered at location x along the sequence. The value of ρ, although a function of x, is mathematically independent of the window size, W. Successive window measurements are carried out by sliding this window in the increments of δ nucleotides. If δ is small, linear interpolation can be used to join the individual window estimates that are gathered at x, $x+\delta, \ldots\ x+k\delta$. In this manner, a continuous distribution of the pattern-density is obtained as a function of x [20].

The value for the pattern density is evaluated using probabilistic methods. The density of patterns observed in each window is defined as the inverse of the probability of observing the patterns in a random window of W nucleotides. The inverse function chosen as, $\rho = -\log(\alpha)$, where the parameter α is the probability of observing frequencies equal or higher than those observed in the window. From a standpoint of hypothesis testing, this would correspond to erroneously rejecting null hypothesis that states that the patterns observed in the region of investigation are no different from those expected from a random sample of the DNA sequence. The value of ρ is computed for both the forward and the reverse strands of the double helix and the average value taken to be the true density estimate for a given location.

The computation of ρ proceeds as follows. Assume that we are searching for k distinct types of patterns within a given window of the sequence. In general, these patterns are defined as rules $R_1, R_2, \ldots, R_k$. The probability of random occurrence of the various k patterns is calculated using the `AND-OR` relationships between the individual motifs. Assume that these probabilities for k patterns are $p_1, p_2, \ldots, p_k$. Next, a random vector of pattern frequencies, F, is constructed. F is a *k-dimensional* vector with components, $F = \{x_1, x_2, \ldots, x_k\}$, where each component x_i is a random variable representing the frequency of the pattern R_i in the W base-pair window. The component random variables x_i are assumed to be *independently* distributed Poisson processes, each with the parameter $\lambda_i = p_i \cdot W$. Thus, the joint probability of observing a frequency vector $F_{obs} = \{f_1, f_2, \ldots, f_k\}$ purely by chance is given by:

$$P(F_{obs}) = \prod_{i=1}^{k} \frac{e^{-\lambda_i} \lambda^{f_i}}{f_i!} \quad where\ \lambda_i = p_i \cdot W \tag{3}$$

The steps required for computation of α, the cumulative probability that pattern frequencies equal to or greater than the vector F_{obs} occurs purely by chance is given by Eq. 4 below. This corresponds to the one-sided integral of the multi-variate Poisson distribution.

$$\begin{aligned}
\alpha &= Pr(x_1 \geq f_1, x_2 \geq f_2, \ldots, x_k \geq f_k) \\
&= Pr(x_1 \geq f_1) \cdot Pr(x_2 \geq f_2) \cdot \ldots \cdot Pr(x_k \geq f_k) \\
&= \sum_{x_1=f_1}^{\infty} \frac{\exp^{-\lambda_1} \lambda_1^{x_1}}{x_1!} \cdot \sum_{x_2=f_2}^{\infty} \frac{\exp^{-\lambda_2} \lambda_2^{x_2}}{x_2!} \cdot \\
&\quad \ldots \cdot \sum_{x_k=f_k}^{\infty} \frac{\exp^{-\lambda_k} \lambda_k^{x_k}}{x_k!}
\end{aligned} \tag{4}$$

The p-value, α, in Eq. 4 is utilized to compute the value of ρ or the *pattern-density* as specified in Eq. 5 below. The pattern density is the unexpectedness $(-\log_2 \alpha)$ associated with the tail probability α. The infinite summation term in Eq. 5 constitutes a convergent series that may be calculated to the precision desired.

$$\begin{aligned}
\rho &= -\log_2(\alpha) \\
&= \sum_{i=1}^{k} \log_2 e \cdot \lambda_i + \sum_{i=1}^{k} \log_2 f_i! - \sum_{i=1}^{k} f_i \log_2 \lambda_i \\
&\quad - \sum_{i=1}^{k} \log_2 (1 + \frac{\lambda_i}{f_i + 1} \ldots \frac{\lambda_i^t}{(f_i+1)\ldots(f_i+t)})
\end{aligned} \tag{5}$$

It may be interesting to point out that the dominant terms in the formulation of ρ above represents the cumulative unexpectedness observed within a window. If the unexpectedness for rule R_i defined in Table 2 was denoted as σ_i, the value of ρ is proportional to the sum of the total pattern unexpectednesss. The total pattern unexpectedness is equal to a sum of unexpectedness associated with every pattern observed within the given region of the DNA sequence. The value of ρ is therefore approximately equal to this sum as shown in Eq. 6. The constant κ is dependent upon the window size being used for measuring ρ.

$$\rho \quad \approx \quad \kappa \ \times \sum_{i=1}^{k} \sigma_i \ \cdot \ f_i \tag{6}$$

Effective analysis of large data sets requires an integrated support for results visualization within the knowledge discovery environment. For example, competent interactive tools for exploring complex real-world data as well as visualization strategies and summarizing capabilities are needed to supplement the intelligent search methods such as the one described above. The integration of interpretive and analysis techniques is needed to help the scientific users focus more on the *interesting* aspects of their tasks, instead of the routine and the mundane components. Such a design philosophy is adopted by the *MAR-Wiz* tool where an applet provides the visualization abilities for this web-based discovery system. Summarizing statistics are also provided where the users can drill down into the exact matches of a putative region of Matrix Attachment. Statistical significance as well as normalized scoring are both optionally provided for scoring MARs. The *MAR-Wiz* discovery tool is available at the following URL:
http://www.futuresoft.org/MAR-Wiz
Fig. 3 presents the output generated from the analysis of the human β-globin gene sequence produced by this MAR discovery tool called the *MAR-Wiz*. The pattern-density values have been normalized to fall within the range of [0 ... 1]. As indicated in Table 5, the regions of high MAR potential values, ρ, have corresponded well to the regions that have been experimentally established by wet-bench analysis to be MARs [16; 17].
However, there are cases where improvement is required. For example, consider the *hprt* gene, that when impaired results in Lesch-Nyhan syndrome, a devastating self-mutilating disease. In this case a MAR has been biologically defined to be contained within the first intron (positions 5534–6107) and has shown great promise when incorporated as an integral component of gene therapy. However, as shown below in Fig. 4, when the hprt sequence was tested with *MAR-Wiz*, this MAR region was not identified. This may reflect the fact that this MAR functions as an ARS (Autonomously Replicating Sequence) which is yet another type of MAR. The other candidate MAR sequences that have been identified by *MAR-Wiz* are coincident with known human mutations causing this disease. This is significant as MARs have been found to be associated with hot spots recombination events that lead to gene mutation.

5. CONCLUSIONS AND FUTURE RESEARCH

It is estimated that mammalian somatic nucleus contains approximately 100,000 MARs of which 30,000–40,000 serve as origins of replication [1]. This begs the question: what is the function of the remaining MARs? One can begin to answer this question by observing that MARs most often flank the ends of gene units. It is reasonable to propose that the remaining 30,000–35,000 pairs of MARs in each cell, anchor the paired ends of the approximately 12,000–30,000 gene domains to the nuclear matrix. It is likely that it is not a simple coincidence that this is the number of genes expressed in each cell. If MARs act, or participate as regions of locus control, then it is likely that their reiteration throughout the genome provides a means to specifically tag genes. If verified this will help move gene therapy from the bench to the bedside by ensuring that the implanted gene adopts the open chromatin conformation when integrated into the genome. This paper demonstrates that the task of detecting the MARs is effectively possible through data mining techniques. The paper also illustrates the synergies between the use of genomic pattern databases, the development computational model for representing these patterns and the application of data mining algorithms as the basis for observing and understanding higher level biological knowledge.

Our future goals are to extend this work to discovery of MARs in a species specific manner by utilizing the different features that are characteristic of each species. More research is also ongoing to establish the validity of the feature independence assumption as well as discovering any causal relationships amongst the constituent features. Contingency tables [7], a basic form for discovering 2-D regularities, will be utilized for discovering pattern interdependence.

6. ACKNOWLEDGEMENT

The author would like to acknowledge the collaboration with Dr. Stephen A. Krawetz, School of Medicine, Wayne State University, Detroit, MI. Dr. Krawetz's laboratory compiled the known lower level patterns that were utilized for prediction of MAR sites, and provided the wet-bench results for establishing accuracy of the data-mining tasks.

7. REFERENCES

[1] J. Bode, M. Stengert-Iber, V. Kay, T. Schlake, and A. Dietz-Pfeilstetter. Scaffold/Matrix Attchment Regions: Topological switches with multiple regulatory functions. *Crit. Rev. in Eukaryot. Gene Expr.*, 6(2&3):115–138, 1996.

[2] T. Boulikas. Nature of DNA sequences at the attachment regions of genes to the nuclear matrix. *J. Cellular Biochemistry*, 52:14–22, 1993.

[3] N. Cercone and M. Tsuchiya. Special issue on learning and discovery in databases. *IEEE Trans. Knowledge & Data Engg.*, 5(6), Dec. 1993.

[4] D. Conklin, S. Fortier, and J. Glasgow. Knowledge discovery in molecular databases. *IEEE Trans. Knowledge & Data Engg.*, 5(6):985–987, 1993.

[5] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From data mining to knnowledge discovery: An overview. In U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthuruswamy, editors, *Advances in Knowledge Discovery and Data Mining*, pages 1–34. AAAI Press, Menlo Park, CA, 1996.

[6] W. Frawley, G. Piatetsky-Shapiro, and C. Matheus. Knowldege discovery in databases: An overview. *AI Magazine*, 14(3):57–70, 1992.

[7] D. Gokhale and S. Kullback. *The Information in Contingency Tables*. New York:M. Dekker, 1978.

[8] D. Hand. Data mining: Statistics and more? *The American Statistician*, 52(2):112–118, 1998.

[9] A. Jarman and D. Higgs. Nuclear scaffold attachment sites in the human globin gene complexes. *EMBO Journal*, 7(11):3337–3344, 1988.

[10] J. Kadonaga. Eukaryotic transcription: An interlaced network of transcription factors and chromatin-modifying machines. *Cell*, 92:307–313, 1998.

[11] G. Keen, G. Redgrave, J. Lawton, M. Cinkowsky, S. Mishra, J. Fickett, and C. Burks. Access to molecular biology databases. *Mathematical Computer Modeling*, 16:93–101, 1992.

[12] E. Keogh and P. Smyth. A probabilistic approach to fast pattern matching in time series databases. In *Proc. of the 3rd. Int'l Conf. on Knowledge Discovery and Data Mining*, pages 24–30, 1997. Menlo Park, CA: AAAI Press.

[13] L. Kliensmith and V. Kish. *Principles of cell and molecular biology*. HarperCollins, 2nd. edition, 1995.

[14] W. Klosgen. Problems for knowledge discovery in databases and their treatment in the statistics interpretor EXPLORA. *Intl. Jou. for Intell. Sys.*, 7(7):649–673, 1992.

[15] W. Klosgen. Efficient discovery of interesting statements in databases. *J. Intell. Info. Sys.*, 4(1):53–69, 1995.

[16] J. Kramer, M. Adams, G. Singh, N. Doggett, and S. Krawetz. Extended analysis of the region encompassing the PRM1→PRM2→TNP2 domain: genomic organization, evolution and gene identification. *Jou. of Exp. Zoology*, 282:245–253, 1998.

[17] J. Kramer and S. Krawetz. PCR-Based assay to determine nuclear matrix association. *Biotechniques*, 22:826–828, 1997.

[18] C. Matheus, P. Chan, and G. Piatetsky-Shapiro. Systems of knowledge discovery. *IEEE Trans. Knowledge & Data Engg.*, 5(6):903–913, 1993.

[19] G. Piatetsky-shapiro. Special issue on knowledge discovery in databases and knowledgebases. *Intl. Jou. for Intell. Sys.*, 7(7), 1992.

[20] G. Singh, J. Kramer, and S. Krawetz. Mathematical model to predict regions of chromatin attachment to the nuclear matrix. *Nucleic Acid Res.*, 25:1419–1425, 1997.

[21] R. Staden. Methods for calculating the probabilities of finding patterns in sequences. *Comput. Applic. Biosci.*, 5(2):89–96, 1988.

[22] M. Tribus. *Thermostatics and Thermodynamics*. D. van Nostrand Company, Inc., Princeton, N.J., 1961.

[23] J. von Kries, L. Phi-Van, S. Diekmann, and W. Strätling. A non-curved chicken lysozyme 5' matrix attachment site is 3' followed by a strongly curved DNA sequence. *Nucleic Acid Res.*, 18:3881–3885, 1990.

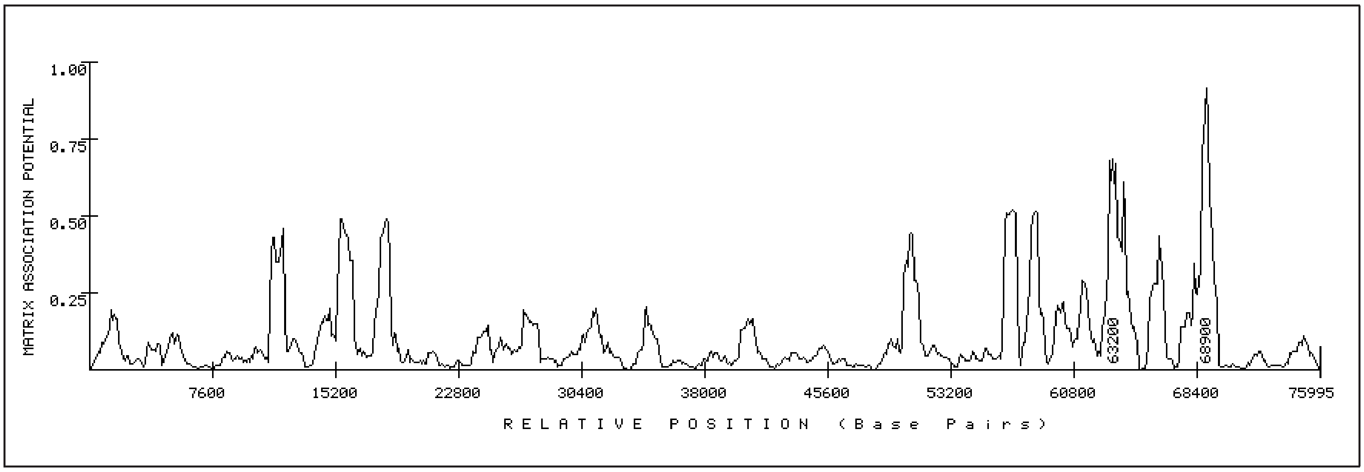

Figure 3: The analysis of human beta-globin gene cluster using the *MAR-Wiz* tool. Default analysis parameters were used for this case.

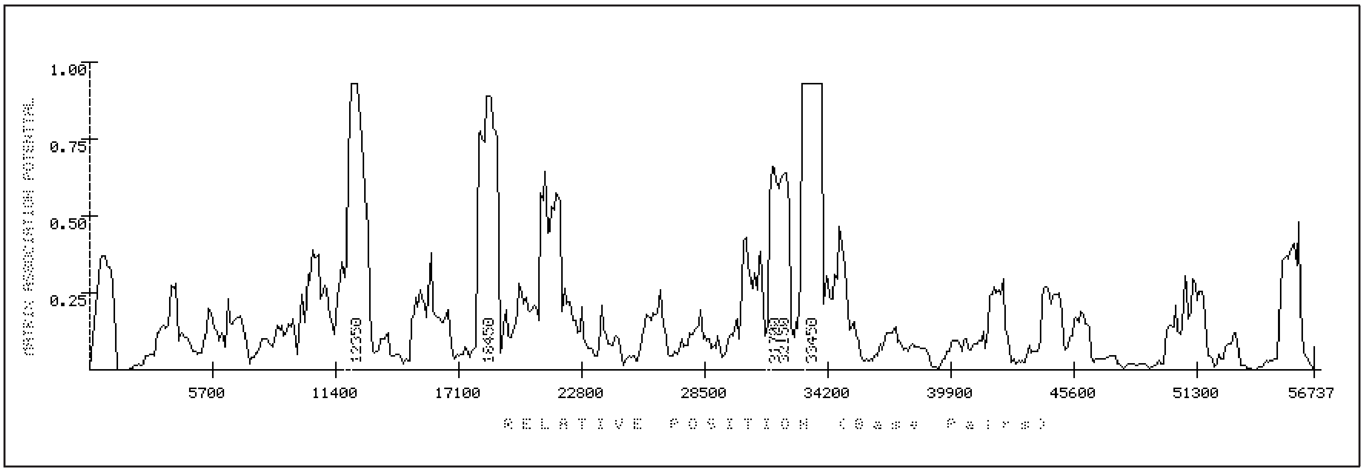

Figure 4: The analysis of the Lesch-Nyhan syndrome HPRT gene shows that additional research is required to identify MARs such as those embedded in introns (positions 5534–6107). Interestingly, most peaks are coincident with known mutations in this human gene. To date there are 104 mutations reported for this gene, described in detail at: `http://www.uwcm.ac.uk/uwcm/mg/search/119317.html`

		Predicted MARs		Experimental MARs	
Sequence	*Length*	*No.*	*Position*	*No.*	*Position*
Human β-globin	75,995	–	–	1	~1.5 kb
		2	~15–18 kb	1	~15–18 kb
		1	50,550	1	~46–54 kb
		4	~56–69 kb	2–4	~58–70 kb
Human Protamine PRM1→PRM2→TNP2	40,573	1	8,700	1	~9.3 kb
		1	34,350	1	~33 kb
		1	38,150	1	~38 kb
Human Apolipoprotein B	7,274	1	225	1	~200 bp
Human Interferon	9,937	2	~3-4 kb	2	~3-4 kb
		1	~ 8 kb	1	~ 8 kb

Figure 5: Predicted and Experimental MARs: Comparison of the MAR locations predicted and those determined experimentally.