Association of nucleotide patterns with gene function classes: application to human 3′ untranslated sequences

Darrell Conklin1,*, Inge Jonassen2, Rein Aasland3 and William R. Taylor4

1ZymoGenetics Inc., 1201 Eastlake Avenue East, Seattle, WA 98102, USA, 2Department of Informatics, 3Department of Molecular Biology, University of Bergen, HIB, N5020 Bergen, Norway and 4Division of Mathematical Biology, National Institute for Medical Research, The Ridgeway, Mill Hill, London NW7 1AA, UK

Received on March 15, 2001; revised on July 24, 2001; accepted on September 21, 2001

ABSTRACT

Motivation: Gene expression is dependent on two main types of signals; one involving transcription factors which initiates gene transcription, and another which regulates the translation of a nascent mRNA. These post-transcriptional events play an important yet incompletely understood role in regulating gene expression and cellular behavior. Many of the identified cis acting elements for translational regulation occur within the 3′ untranslated region (3′ UTR), and some have been observed to occur with surprising regularity within certain protein function classes.

Results: In this study, we present a new association rule mining method for discovering nucleotide sequence patterns that appear in more sequences than expected within protein function classes. The method is applied to a database of human 3′ UTR sequences, and some significant associations between nucleotide patterns and protein function classes are discovered. Among previously identified patterns, the AU-Rich Element (ARE) is found here to occur within the 3′ UTR of cytokines, providing statistical validation of an association often reported in the literature. The method has also identified some GC-rich patterns, found to occur within the 3′ UTR of homeodomain transcription factors and nuclear proteins. The method should be applicable to many types of regulatory element discovery.

Contact: conklin@zgi.com

INTRODUCTION

The interpretation of genomic sequence data requires methods for accurate prediction of exons and introns of genes, and the discovery and identification of their transcription factor binding sites. Many molecular processes involve specific sequences, either at the DNA or RNA level; for instance, in the various steps leading to transcription initiation, many different transcription factors bind to DNA with various degrees of sequence specificity. Although we know fairly well how transcription factors interact with DNA, we are far from being able to predict, with reasonable accuracy, which sites in DNA are actually being read by such proteins. An additional level of complexity arises from the fact that one gene may be regulated by different factors in different types of cells and at different time points in the life cycle of the organism.

Similarly, many proteins involved in post-transcriptional regulation interact with pre-mRNA and/or mature mRNA. One example is the initiation of protein synthesis, which can be influenced by sequence elements in both the 5′ and 3′ Untranslated Regions (UTRs). The process leading to alternative splicing is another example.

Gene transcription of eukaryotic protein-coding genes begins with the binding of RNA polymerase to DNA, followed by elongation, termination, and production of a pre-mRNA by 3′ end cleavage and polyadenylation. mRNA formed from spliced pre-mRNA is exported to the cytoplasm for protein translation. At this point the translation and stability of mRNA is subject to different post-transcriptional regulatory processes. Post-transcriptional events play an important role in regulating gene expression and cellular behavior. Translation of a nascent mRNA is more rapid than de novo transcription and translation, so a cell can respond flexibly and rapidly to environmental cues by employing translational versus transcriptional regulation. Several molecular mechanisms for translational regulation have been identified; many of these involve cis acting elements within the 3′ UTR of an mRNA. Examples of translational regulation include: translational masking of a transcript by mRNP...
particles (Curtis et al., 1995); control of cytoplasmic polyadenylation in the development of oocytes (Richter, 1999); stabilization and circularization of mRNA involving poly(A) binding proteins and translation initiation factors (Wells et al., 1998), nuclear-cytoplasmic shuttling of transcripts (Fan and Steitz, 1998); and targeting of discrete subcellular location of mRNA (St Johnston, 1995). Though many of these processes involve proteins containing conserved RNA-binding domains (Burd and Dreyfuss, 1994), other trans acting mechanisms include the binding of homeodomain transcription factors (Niessing et al., 2000), and RNA–RNA interactions (Ferrandon et al., 1997).

Regulatory elements for 3′ UTRs can be divided into two groups. One comprises long segments from specific genes, where deletion or replacement of the segment has an observed phenotype. The other group, of interest in this paper, comprises general patterns contained within several genes. An example of a general pattern is the AU-Rich Element (ARE; Chen and Shyu, 1995) which is represented by multiple, overlapping copies of the core pentanucleotide AUUUA. It targets mRNAs for rapid and selective degradation, and is often found in the 3′ UTRs of potent and transiently expressed signaling molecules such as cytokines and chemotactic factors (Kruys and Huez, 1994). Several proteins that bind AREs have been identified; the most extensively characterized being the RNA-binding protein AU/F1.

The availability of genomic sequence and full-length mRNAs for many human genes is promoting the exploration of statistical methods for the automated discovery of recurrent and potentially important cis acting elements (Vanet et al., 1999). Pattern discovery techniques have been applied to the problem of regulatory element discovery in upstream regulatory regions of yeast (Wolfsberg et al., 1999) and bacteria (McGuire et al., 2000) in genes within a common functional pathway. Other studies have looked at yeast downstream regions (van Helden et al., 2000) and human 3′ UTRs (Pesole et al., 1994). In these studies, a few well-known cis acting elements could be identified: for example, 3′ UTR polyadenylation sites and 3′ UTR yeast enhancer elements. Pesole et al. (1994) further demonstrated that the ARE can be discovered by consideration of statistically over-represented patterns in a set of human 3′ UTR sequences. It has been proposed that there should be candidate signals for post-transcriptional regulation between alternative polyadenylation signals (Gautheret et al., 1998). Such signals could allow for differential regulation of messages with different 3′ UTR length and content.

Some sequence patterns, for example the ARE, occur with some specificity in certain protein function classes. It is reasonable to hypothesize the existence of other protein class-specific patterns associated with message degradation or perhaps with other post-transcriptional mechanisms. In this paper we describe a technique that produces associations of sequence patterns with protein function classes. The method can be applied to all kinds of problems where short sequence patterns are associated with function class information. In this study we use a collection of human 3′ UTR sequences to explore the hypothesis that protein class-specific 3′ UTR regulatory elements can be discovered using rigorous, fully automated methods.

Recognition of a particular nucleotide pattern by a trans acting factor is a complex process, involving both sequence specific interactions, the appropriate structural context (e.g. a stem-loop conformation) and the correct intracellular localization of the target mRNA and binding factors. It follows that the presence of an oligonucleotide pattern within a 3′ UTR is not a sufficient condition for a functional cis acting element; the associated patterns will occur within sequences where they are not functional. Furthermore, we cannot expect to find functional patterns that are present in all or even a large proportion of 3′ UTR sequences from a particular protein class. This is because the annotation of genes is generally based on broad biological function, rather than on sharp delineation into proteins whose 3′ UTR sequences are bound by particular trans acting factors.

The data analysis method we have developed is a derivation of association rule mining (Agarwal et al., 1996). The method automatically and efficiently discovers patterns that occur in more sequences than expected within the 3′ UTR of genes grouped into function classes. Important features of the method include the association of sequence patterns with terms from a protein function term vocabulary; the weighting of sequences to correct distorted pattern counts associated with sequence database redundancy; and the accurate computation of the number of sequences within a protein function class expected to contain a pattern. The algorithm for enumerating associations between patterns and protein function terms is simple yet powerful in that all possible associations can be explored in reasonable time.

**METHODS**

**Database of human 3′ UTR regions**

For the discovery of regulatory elements in 3′ UTR sequences, the 3′ UTR section of UTRdb (Release 13.0; 13 680 human sequences; Pesole et al., 2000) is used. Each UTRdb entry contains an associated EMBL accession number that is used to join UTRdb identifiers to Swiss-Prot (Release 39; Bairoch and Apweiler, 2000) identifiers by the following procedure. If an EMBL accession number appeared within more than one Swiss-Prot entry the associated UTRdb entry was not joined. If more
than one UTRdb identifier could be joined to the same Swiss-Prot identifier, only the longest UTRdb sequence was joined.

Protein function terms
The objective of this study is to associate patterns with a classification of protein function. Therefore it is necessary to derive an ontology of protein function. The approach used here is to map Swiss-Prot identifiers to one or more of the following protein function terms: a function category, given by a Swiss-Prot keyword; a protein family, given by a Pfam (Bateman et al., 2000) or Prosite (Hofmann et al., 1999) identifier; or a tissue source, given by a TISSUE line annotation in Swiss-Prot. Different vocabularies for protein function could be used; for example, Medline keywords (Masys et al., 2001) that are used to index scientific citations associated with a gene; or gene ontology molecular function concepts (Ashburner et al., 2000). Rison et al. (2000) provide a thorough survey of current functional annotation schemes for genomes. Due to our interest in cytokines and the ARE, we have in addition annotated 43 human Swiss-Prot entries as helical cytokines. A total of 2472 possible different protein function terms was used to describe the UTRdb sequences.

Sequence pattern terms
A sequence pattern term is a string of nucleotides over the set \{A, C, U, G\}. The length of a sequence or sequence pattern term \(P\) is denoted by \(l(P)\).

Sequence weighting
A well-known problem in biological sequence analysis is that closely related sequences may present an undesirable sample bias (Thompson et al., 1994). For pattern discovery, two nearly identical sequences will contribute two counts rather than one count for nearly any pattern occurring within them; these occurrences are due to database redundancy or recent gene divergence and not necessarily due to functional pattern preservation. The most straightforward way to handle this effect is to discard all but one of a group of sequences with high similarity. This has the disadvantage that potentially valuable data might be discarded. The technique used here instead retains all of the data and gives each sequence a numerical weight inversely proportional to its similarity to other sequences within the analysis set. For example, each sequence in a group of \(n\) nearly identical sequences should receive a weight of about \(1/n\). To compute the weight \(w(s)\) for a sequence \(s\) in a database, it is aligned with all other similar sequences using an ungapped blastn multiple alignment (blastn parameters \(m = 6, e = 1e-15\); Altschul et al., 1997); position based sequence weights (Henikoff and Henikoff, 1994) are computed for all columns of this multiple alignment; and the sum of all column weights is computed and normalized by the sequence length. This procedure is repeated for every sequence in the database.

Associations
To represent and discover associations between 3' UTR sequences and protein function, we use ideas derived from the field of association rule mining (Agarwal et al., 1996). In the context of 3' UTR pattern discovery, an association is a conjunction \(P \land C\), where \(P\) is a sequence pattern term and \(C\) is a protein function term. The set of UTRdb identifiers for which the conjunction is true (called the instances of the association) is defined as follows. The instances \(i(P)\) of a pattern term \(P\) is the set of identifiers for 3' UTR sequences containing \(P\) at least once; for a protein function term \(C\), \(i(C)\) is the set of sequence identifiers linked with a Swiss-Prot entry of class \(C\). The instances of an association \(P \land C\) is simply the intersection of their instance sets \(i(P)\) and \(i(C)\). A term or association \(t\) has a support \(N(t)\) which is the sum of the weights of sequences in the set \(i(t)\):

\[
N(t) = \sum_{s \in i(t)} w(s). \tag{1}
\]

The empirical probability \(p(t)\) of a term \(t\) is \(N(t)/n\), where \(n\) is the total (weighted) number of sequences in the analysis set.

Association rules
Associations can be oriented into directed rules, where the right-hand side can be inferred from the left hand side of the rule. The confidence of a rule \(P \Rightarrow C\) is the conditional probability \(p(C|P)\) of the class \(C\) given the pattern \(P\). The lift of a rule is the ratio \(p(C|P)/p(C)\), which indicates the increase in the probability of \(C\) with the condition \(P\). Association rules with high confidence and high lift are of the most interest as they have predictive power.

Association scoring
The potential biological significance of an association \(P \land C\) is evaluated by measuring statistical correlation. The terms \(P\) and \(C\) are correlated if they occur together more frequently than expected; that is, if the observed support \(N(P \land C)\) is substantially greater than the expected support \(E(P \land C)\). The magnitude of this difference is evaluated using a \(\chi^2\) score:

\[
\frac{(N(P \land C) - E(P \land C))^2}{E(P \land C)^2}. \tag{2}
\]

The score for an association will increase with the difference between its observed and expected support.

Association expectation
A simple definition of the expected support \(E(P \land C)\) of an association is \(p(P) \times p(C) \times n\), where \(n\) is the
total (weighted) number of sequences in the analysis set. However, this definition is acceptable only if all sequences are identical in length. Otherwise, associations with protein classes with long 5′ UTR sequences will have inflated scores because while their observed support will tend to increase for any pattern, their expected support will remain constant. A solution to this problem was provided by Pesole et al. (1992), where sequence length is taken into account in the calculation of pattern expectation. Our modification of their approach to handle sequence weighting and protein function classes is described here.

Consider a pattern P and a sequence s with lengths l(s) and l(P). There are (l(s) − l(P) + 1) sites in s where P can possibly match. Assuming site independence and the expected number of matching sites in the sequence s is therefore

\[ \lambda(P, s) = \hat{p}(P) \times (l(s) - l(P) + 1) \]  

where \( \hat{p}(P) \) is the probability of finding the pattern P at a site (see Section Pattern prior probabilities). Assuming site independence and using the Poisson distribution, the probability that the pattern P does not occur in the sequence s is \( e^{-\lambda(P, s)} \), and the probability of finding one or more occurrences is therefore \( 1 - e^{-\lambda(P, s)} \). The expected support of an association \( P \land C \) is the sequence-weighted sum of this quantity for all sequences in \( i(C) \):

\[ E(P \land C) = \sum_{s \in i(C)} (w(s) \times (1 - e^{-\lambda(P, s)})) \]  

**Pattern prior probabilities**

Equation (3) incorporates a term of the form \( \hat{p}(P) \). This is the probability of finding a pattern P in a segment of l(P) nucleotides. Due to sparse data, this term is not computed empirically, but rather is estimated using a first-order Markov model (van Helden et al., 2000; Pesole et al., 1992). The model is parameterized using sequence-weighted nucleotide and dinucleotide frequencies in the analysis set.

**Statistical significance**

For each association \( P \land C \), it is useful to report a p-value; the probability that its score could have arisen by chance. To explore the form of the score distribution for associations, we applied \( l(s) \log l(s) \) random base swaps for each sequence s in the UTRdb. This produced a sequence database with an identical base composition and length as the original, but destroyed any patterns present in the data.

The distribution of scores for all associations with patterns of length 5–12, with supports of at least 10, on this synthetic database were plotted. The \( \chi^2 \) distribution, which could normally be used to model scores of the form in equation (2), substantially underestimates the synthetic data distribution. This is because there is obviously very high correlation between different overlapping patterns. However, we found that the distribution of scores in the synthetic database is approximated by the exponential distribution

\[ p(X \geq x) \approx e^{-\sqrt{T}} \]  

which is an upper bound on the synthetic data distribution for all pattern lengths tested, for p-values lower than 0.05. Therefore, using this approximation it is possible that we will reject significant patterns, but unlikely that we will accept an insignificant pattern.

The raw p-value of an association must be adjusted to reflect that fact that we are evaluating its significance not in isolation but within a large ensemble of associations found in a data set. Given a particular score, an adjusted p-value is computed by multiplying equation (5) by an adjustment factor which is the total number of associations evaluated for significance.

**Clustering of associations**

The association finding method can produce a large number of associations including many that are partially redundant. Redundancy may be present in associations with overlapping sequence pattern terms, correlated protein function terms, or a combination of both. To address the problem of redundancy, we note that similar associations have similar instances, and we cluster all discovered associations using a simple method that makes use of this fact. The similarity of two associations \( A \) and \( A' \) is measured using the Dice similarity coefficient (Sneath and Sokal, 1973):

\[ s(A, A') = \frac{2 \times N(A \land A')}{N(A) + N(A')} \]  

That is, the similarity of two associations \( A \) and \( A' \) is twice the support of their conjunction, divided by the sum of their individual supports. Similarity is therefore a measure in the interval \([0, 1]\) that increases as the associations \( A \) and \( A' \) have more instances in common, because the quantity \( N(A \land A') \) will increase. To cluster a set of associations, we use an average linkage hierarchical clustering algorithm to iteratively form clusters whose centroids are similar above a specified threshold. Though we have chosen the Dice similarity coefficient, clustering results will be similar using other methods, because we are liberal in the choice of a threshold for clustering of associations. Furthermore, we have calibrated the choice of this threshold to admit only associations with obvious overlapping patterns or with obviously redundant function terms.

**Algorithm**

Our algorithm finds all significant associations with a specified minimum support. The algorithm performs an
enumeration of the entire space of possible associations. 
Bit vectors are used throughout to efficiently represent instance lists of patterns and classes, and the bit-and operation is used to rapidly determine the instances of an association. Like all association rule mining algorithms, we make use of the fact that $N(P) \geq N(P \land C)$ and $N(C) \geq N(P \land C)$ for any pattern $P$ and function term $C$. Therefore the space of associations can be pruned, at the outset, of pattern or function terms having less than the minimum specified support. The total number of possible associations considered is simply the product of the number of remaining pattern terms and function terms. This is the adjustment factor used to compute adjusted $p$-values for associations (see Section Statistical significance).

**RESULTS**

With the procedure described in Section Methods, 3833 entries in UTRdb were joined with a Swiss-Prot entry, producing a dataset of 2.8 million nucleotides. Applying sequence weighting reduces the total to 2.2 million nucleotides, indicating a redundancy of about 20%. The sequence-weighted dinucleotide frequency composition of the joined sequences from UTRdb closely matches that of the full UTRdb, and also agrees closely with the frequency analysis of Pesole et al. (1994). The least frequent dinucleotides are CG (odds ratio 0.29) and TA (0.71). The most frequent dinucleotides are CC (1.35) and GG (1.25).

The association finding algorithm was directed to find associations with a minimum support of 10 and a maximum adjusted $p$-value (see Section Statistical significance) of 0.01. If a pattern is more significant within the complete set of 3’ UTR sequences than within a specific class, it is not retained. Sequence pattern terms of length 5–12 were used. The similarity threshold for clustering of associations was set to 0.5. Associations were sorted by their lift, and clusters containing an association with a lift of at least 2 were retained. For each such cluster, the association with the highest lift is presented in Table 1.

Some interesting results can be highlighted, for patterns of various lengths.

**Pentamers, hexamers**

No significant associations were found. This is because the base expectation of these patterns is already high, meaning that it is difficult for any class to contain more sequences with the pattern than expected.

**Heptamers**

One significant heptamer cluster was discovered. It is an association with a GC-rich pattern and homeobox genes, including the Pfam homeobox domain identifier (PF00046). The confidence (0.24) and lift (13.7) of this association are relatively high. Table 2 presents the alignment of the instances of this association, showing no sequence conservation outside of the core pattern, but a slight GC-rich context.

---

**Table 1. Application of the association finding algorithm to human 3’ UTR sequences, for pattern lengths of 7–12**

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Protein class</th>
<th>Obs</th>
<th>$p$-value</th>
<th>Confidence</th>
<th>Lift</th>
</tr>
</thead>
<tbody>
<tr>
<td>CGCCGCC</td>
<td>pfam:PF00046</td>
<td>10.2</td>
<td>5.1e−05</td>
<td>0.24</td>
<td>13.7</td>
</tr>
<tr>
<td>CGCGGCGG</td>
<td>Transcription regulation</td>
<td>11.8</td>
<td>0.003</td>
<td>0.33</td>
<td>3.9</td>
</tr>
<tr>
<td>ATTTAATT</td>
<td>Helical cytokine</td>
<td>12.2</td>
<td>1.5e−07</td>
<td>0.13</td>
<td>15.1</td>
</tr>
<tr>
<td>AAATTTAAT</td>
<td>Cytokine</td>
<td>11.7</td>
<td>0.0025</td>
<td>0.15</td>
<td>7.0</td>
</tr>
<tr>
<td>TATTTATTA</td>
<td>Helical cytokine</td>
<td>11.2</td>
<td>7.5e−23</td>
<td>0.22</td>
<td>25.6</td>
</tr>
<tr>
<td>TTTATTTTTAT</td>
<td>Transcription regulation</td>
<td>10.2</td>
<td>0.0086</td>
<td>0.37</td>
<td>4.3</td>
</tr>
<tr>
<td>ATATTTATTA</td>
<td>Signal</td>
<td>12.4</td>
<td>6.2e−07</td>
<td>0.67</td>
<td>2.7</td>
</tr>
<tr>
<td>TATTTTTATTT</td>
<td>Nuclear protein</td>
<td>11.9</td>
<td>0.0036</td>
<td>0.43</td>
<td>2.5</td>
</tr>
<tr>
<td>TATATATATTT</td>
<td>Nuclear protein</td>
<td>11.5</td>
<td>1.5e−07</td>
<td>0.41</td>
<td>2.4</td>
</tr>
<tr>
<td>TATTTAATAT</td>
<td>Signal</td>
<td>10.2</td>
<td>0.00019</td>
<td>0.60</td>
<td>2.4</td>
</tr>
<tr>
<td>TTTTATTAAAAT</td>
<td>Nuclear protein</td>
<td>24.4</td>
<td>5.3e−06</td>
<td>0.41</td>
<td>2.3</td>
</tr>
<tr>
<td>TTAATTTTTAA</td>
<td>Glycoprotein</td>
<td>15.4</td>
<td>0.0095</td>
<td>0.62</td>
<td>2.3</td>
</tr>
<tr>
<td>TATTTATTTAT</td>
<td>Signal</td>
<td>15.3</td>
<td>3.6e−09</td>
<td>0.50</td>
<td>2.0</td>
</tr>
<tr>
<td>AAATTTATTTT</td>
<td>Signal</td>
<td>12.2</td>
<td>4.3e−11</td>
<td>0.70</td>
<td>2.8</td>
</tr>
<tr>
<td>TTTTTATTATT</td>
<td>Nuclear protein</td>
<td>14.3</td>
<td>5.2e−06</td>
<td>0.46</td>
<td>2.6</td>
</tr>
</tbody>
</table>

Each entry of the table is the highest lift association within a cluster of similar associations. Indicated are the number of sequences within the protein class observed to contain the pattern, the association $p$-value, confidence, and lift. PF00046 is the Pfam homeobox family identifier. Signal is the Swiss-Prot keyword for proteins containing a signal sequence.
The method described in this paper has verified an association often reported in the literature; that AREs tend to occur within the 3’ UTR of cytokines. The method reported a significant association with patterns containing the ARE pentamer and cytokines; and similarly with the subclass of helical cytokines. It is very encouraging that a known cis acting element and protein class association can be discovered computationally.

The method also discovered two GC-rich patterns associated with human DNA binding proteins. One of these patterns is strongly associated with homeodomain containing genes. Though elucidation of this association needs experimental research, several speculations on its biological meaning can be made. Many homeodomain transcription factors are expressed in a limited number of cells during embryonic development and in adult organs and tissues. Such transcription factors often require exquisitely precise control of activation and inactivation. The discovered pattern may function, like the ARE, to target mRNAs for rapid degradation and poly(A) tail removal. Some maternal mRNAs encoding homeodomain proteins are non-polyadenylated, therefore another possibility is that the discovered pattern represents a type of cytoplasmic polyadenylation signal. Perhaps the GC-rich pattern is contained within CpG islands and represents a DNA methylation site that happens to be within the 3’ UTR of several homeobox genes, overlapping with the promoter region of a neighboring gene. This is a plausible explanation because many homeobox gene families are tightly clustered in the genome.

For the ARE there is a diversity of binding factors, and though functional AREs contain a conserved pentamer they differ in their exact flanking sequences. Therefore, a fully sensitive model for the ARE and other 3’ UTR cis acting elements may require a representation of RNA secondary structure. A powerful method to find conserved stem-loop structural patterns in RNA by sensitive multiple sequence alignment is described by Gorodkin et al. (2001). Their algorithm, while computationally feasible for a few sequences, would be prohibitively expensive to apply to the discovery task described in this paper, where thousands of sequences and protein classes need to be explored and where there is no preconceived notion of where a regulatory element might lie within a 3’ UTR. Furthermore it is unlikely that stem-loop structures will be sufficient to describe 3’ UTR regulatory elements (Fan et al., 1997). An interesting idea is to explore a hybrid method, using our technique to propose conserved nucleotide patterns and subsequently attempting to find a structural pattern within these patterns and their flanking contexts. Alternatively, structural patterns could be verified by our method, which could report whether they were involved in associations with specific protein classes.

A large number of sequence pattern discovery methods, some related to the one described here, have been proposed. A survey is given by Brazma et al. (1998) where a framework is presented describing each algorithm in terms of solution space, scoring scheme, and algorithm-
mic approach. The solution space can consist of a set of substring patterns, regular expression type patterns, weight matrices, or hidden Markov models. Most scoring schemes measure how frequently a pattern occurs in the sequence set in relation to the number of occurrences expected. As for algorithmic approaches, Brazma et al. (1998) distinguish between sequence and pattern driven approaches where the first are based on sequence comparisons while the latter are based on a search through the solution space. Pattern driven algorithms can tabulate counts on all possible patterns, or they can use specialized algorithms and data structures such as suffix trees (Apostolico et al., 2000; Vilo et al., 2000). In terms of applying association rule mining methods to biology, previous work has considered finding associations between Swiss-Prot keywords and enzyme EC class identifiers (Satou et al., 1997). Brazma et al. (1997) have applied association rule mining to find combinations of known yeast transcription factor binding sites. Here we have applied association rule mining ideas to the more difficult problem of nucleotide sequence pattern discovery.

In evaluating associations, we have included an adaptation of a sequence weighting scheme used in profile analysis (Henikoff and Henikoff, 1994; Altschul et al., 1997). This weighting scheme helps to avoid biases in the sequence sets that distort the observed and expected support of associations. A frequently used alternative approach is to reduce the input set by removing all except one from each set of highly similar sequences (Saqi and Sternberg, 1994). Our approach is less sensitive to the choice of similarity threshold and it avoids the wasteful deletion of potentially valuable sequence data.

The pattern discovery method presented in this paper has been used to find nucleotide sequence patterns associated with gene function classes. For this particular variant of the pattern discovery problem it improves upon other approaches (van Helden et al., 2000; Pesole et al., 1992; Vilo et al., 2000) which would need to define separate sequence sets for each protein function term and perform pattern discovery on each such set separately. With a slight modification to the way association expectation is computed, the algorithm generalizes to find associations between multiple patterns and function terms. This method may be applied for identification of putative binding sites in 5' UTRs and upstream regulatory regions within groups of genes exhibiting coordinated expression.

ACKNOWLEDGEMENT

I.J. was supported by grants from the Norwegian Research Council.

REFERENCES


Nucleotide patterns in 3′ UTR sequences


