Chapter 6

Dependence within motif in multiple sequence alignment: Entropy based approach

6.1 Introduction

Gene and genome sequencing projects have generated vast amount of data which require use of computational tools and techniques. The control and regulation of gene expression is carried out through regulatory proteins and transcription factors. Their sequence specific binding to DNA influences gene expression. The DNA binding motif is generally very short in length spanning 10-12 bases and may exhibit variability across species. It is a challenging task to delineate and specify these motifs by computational means (Hertz and Stormo, 1999).

Several efforts have been made to predict and identify motifs in the regulatory DNA sequences (Bailey and Gribskov, 1997; Hertz and Stormo, 1999; Fickett and Wasserman, 2000; Benos, Bulyk and Stormo, 2002; Barash et al., 2003; King and Roth, 2003; Zhou and Liu, 2004). Multiple sequence alignment (MSA) is one of the important steps for finding conserved patterns in a given family of DNA sequences. One can find biologically interesting motifs by identifying over represented n-mers in the upstream sequences (Apostolico et al., 2005; Sinha and Tompa, 2002) and also
by phylogenetic footprinting methods (Blanchette and Tompa, 2002).

It is possible to characterize the given multiple sequence alignment either in terms of a consensus sequence or through a position weight matrix (PWM). In the consensus model, only the dominant base in each column gets represented whereas in PWM contribution from all bases are included as elements of the matrix. It is assumed that each column (or position) in the multiple alignment is independently distributed. The assumption of independence may not be a valid one, particularly the order of nucleotides in a given sequence is dictated by biological constraints (Bulyk et al., 2002; Barash et al., 2003; Zhou and Liu, 2004). These constraints in fact introduce dependence which could either be weak or strong. A strong dependence may provide biologically meaningful patterns. Accordingly, one would be interested to identify an appropriate measure of dependence. The usual measure of dependence based on correlation would not serve the purpose as it deals with two variates at a time. In DNA sequence context, one may have to deal with a combination of nucleotides resulting in several variates.

The objective of our study is to explore a measure of dependence which should incorporate all positional information. It is natural to expect that information theory may be appropriate framework to study dependence (Shannon, 1948; Shultzabeger et al., 2001; Gadiraju and et al, 2003; Krishnamachari et al., 2004).

6.2 Material and methods

6.2.1 Data

The first data set contained 471 pre-aligned E. coli promoter sequences and were obtained from the PromeEC (Hersberg et al., 2001) database which can be downloaded from http://bioinfo.md.huji.ac.il/marg/promec/promec/. This data set contains the
range of nucleotide positions from -75 to +25, where +1 denotes the transcription start site.

The second data set corresponds to 1055 E.coli Ribosome binding site sequences and can be obtained from http://www.ccrnp.ncifcrf.gov/toms/.

The third data set made up of 4737 of putative human promoter sequences and is published by Ramirez and et al (2004). We have analyzed sequences from -100 to +50 with respect to the transcription start site.

The fourth data set is generated randomly and the size is same as E. coli promoter sequences.

6.2.2 The model

A DNA binding pattern or motif such as "TACTTG," for example, can be treated as the joint occurrence of six bases (word size) with alphabets from $S = \{A, T, G, C\}$. In this case, the number of possible patterns for word of size six ($n = 6$) is $4^6$. A typical multiple sequence alignment may have $N$ sequences (rows) and $m$ positions (columns) and one can search for occurrence of $n$-mers in them. The relative frequency of these words can be used to estimate information content.

Shannon’s measure of average uncertainty has been used to compute information content from MSA assuming the positions to be independent. The focus of this study is to examine the statistical dependence within motif. To address this issue, we need a measure of dependence when length of the motif varies.

To this end, consider the joint occurrence of $n$ bases $(X_1, X_2, \ldots, X_n)$ with the probability distribution $P(X_1 = x_{i_1}, X_2 = x_{i_2}, \ldots, X_n = x_{i_n}) = p_{i_1 i_2 \ldots i_n}$, where $(x_{i_1}, x_{i_2}, \ldots, x_{i_n}) \in S^n$, since in the case of DNA $S_i \equiv S$. Thus, the average uncertainty in this case takes
the form
\[ H(X_1, X_2, \cdots, X_n) = - \sum_{i_1 i_2 \cdots i_n} p_{i_1 i_2 \cdots i_n} \log_2 p_{i_1 i_2 \cdots i_n}. \] (6.2.1)

In particular, Shannon's measure (Applebaum, 1996) of average uncertainty for a single random variate \( X \) has the form
\[ H(X) = - \sum_i p_i \log_2 p_i. \] (6.2.2)

**Measure of Dependence**

It would be instructive to introduce the notion of dependence of random variables in terms of entropy of joint distribution and entropies of individual random variables. Such a measure gains more relevance as the information about \( \frac{1}{2}m(m-1) \) correlation coefficients of \( m \) random variables do not yield a single well-defined measure of stochastic dependence (Kapur and Kesavan, 1992). It may be noted that if the positions in the alignment are statistically independent then \( P(x_{i_1}, x_{i_2}, \cdots, x_{i_n}) = p_{i_1} p_{i_2} \cdots p_{i_n} \) and Eq.(6.2.1) simplifies to
\[ H(X_1, X_2, \cdots, X_n) = \sum_{i=1}^{n} H(X_i). \] (6.2.3)

In case the variables are statistically *not independent*, the measure of stochastic dependence \( D() \), termed *cross-entropy*, is given as
\[ D(X_1, X_2, \cdots, X_n) = \sum_{i=1}^{n} H(X_i) - H(X_1, X_2, \cdots, X_n). \] (6.2.4)

It is easy to check that \( D \geq 0 \), with the equality holding for the case of statistically independent variables and its magnitude depicting the degree of dependence among the variates (Kapur and Kesavan, 1992). The independence in a MSA will imply that the occurrence of the nucleotide at a particular position will not depend on the preceding positions.
It is possible to write Eq.(6.2.4) in the form

\[
\mathcal{D}(X_1, X_2, \ldots, X_n) = - \sum_{i=1}^{n} p_i \log_2 p_i - \sum_{i=1}^{n} p_i \log_2 p_i - \cdots \\
- \sum_{i=1}^{n} p_i \log_2 p_i + \sum_{i=1, i \neq k}^{n} p_i \log_2 p_i - \sum_{i=1}^{n} p_i \log_2 p_i
\]

\[
= \sum_{i=1, i \neq k}^{n} p_i \log_2 \left( \frac{p_i}{p_1 p_2 \ldots p_n} \right). \tag{6.2.5}
\]

This compares to the well-known expression which is referred to as K-L measure.

**Normalised measure of dependence**: Noting that the dependence measure \(\mathcal{D}\) is a function of motif length, comparison among them for different length would not be meaningful. In this regard, a normalised measure of dependence (Kapur and Kesavan, 1992) would be useful. The measure \(\mathcal{\hat{D}}\) is defined as

\[
\mathcal{\hat{D}} = \frac{\sum_{i=1}^{n} H(X_i) - H(X_1, X_2, \ldots, X_n)}{(n-1) H(X_1, X_2, \ldots, X_n)}, \tag{6.2.6}
\]

with \(\mathcal{\hat{D}} \in [0, 1]\).

**The Statistic**: The estimated value of the dependence measure for \(n\)-mer sequence starting at the position \(k\) corresponding to Eq.(6.2.4) is denoted as \(d^n_k\). The following two key issues arise, viz. (i) the comparison of \(d^n_k\) for different word (motif) size and (ii) the sampling distribution of the statistic \(d\).

On similar lines, another statistic \(d^n_k\) based on Eq.(6.2.6) can be constructed. In contrast to \(d^n_k\), this statistic along with standardized \(d^n_k\) denoted as \(d^n_k\), will enable comparison for different values of \(n\).

**Computational procedure**

For the sake of completeness, we denote \(u_{ij}\) as the \(i\)-th \(n\)-mer symbol combination in the \(j\)-th aligned sequence. Thus, for a fixed \(n\)-mer sequence, starting at the position \(k\), there are a total of \(N\) samples from which we can obtain the distribution having
\[ |S^n| = 4^n \text{ classes } |C_i|_1^{4^n}. \]  
The frequency \( f_i \) for the class \( C_i \) is computed for the position \( k \in [1, m - n + 1] \) from
\[
N \hat{p}_i = f_i = \sum_{j=1}^{N} 1_{[u_k \in C_i]}.
\]  
(6.2.7)

For example, in case \( k = 10 \) and \( n = 1 \) (mono nucleotide), probability \( p_i \) for a given data set is estimated using
\[
\hat{p}_i = \frac{f_i}{N},
\]
where \( f_i \) denotes the observed frequency of the \( i \)-th symbol in \( S \) such that \( \sum_{i=1}^{4} f_i = N \). It may be noted that this computation can be carried out for each position (or column) starting at any arbitrary position \( k \in [1, m - n + 1] \).

An immediate task at hand is to validate our model against available data sets. To this end, we need a statistic which would enable us to quantitatively test the hypothesis of independence of \( n \)-mers.

### 6.3 Asymptotic distribution of the statistic \( d \)

In this section we develop a test for the hypothesis of complete independence, which in the MSA context can be stated as
\[
H_0 : p_{i_1 i_2 \ldots i_n} = p_{i_1} p_{i_2} \ldots p_{i_n} \text{ for all } i_1, i_2, \ldots, i_n = 1, 2, 3, 4.
\]

For example, consider the probability of the occurrence of the motif 'ATC' in \( n = 3 \). In this case the assertion under \( H_0 \) in terms of the elements in \( S \) (denoted by 1, 2, 3, 4) can be stated as
\[
p_{214} = p_2 \cdot p_1 \cdot p_4,
\]
where \( p_2 = \sum_{j} \sum_{k} p_{2jk} \), \( p_1 = \sum_{i} \sum_{k} p_{i1k} \) and \( p_4 = \sum_{i} \sum_{j} p_{ij4} \).

The sampling distribution of the statistic \( d \equiv d(X_1, X_2, \ldots, X_n) \), which measures distance of the probability distribution from that of the reference distribution (corresponding to independently distributed random variables) is required. Assuming the
deviations to be small, we write
\[ \hat{p}_{i_1 i_2 \ldots i_n} = \hat{p}_{i_1} \hat{p}_{i_2} \cdots \hat{p}_{i_n} (1 + \epsilon_{i_1 i_2 \ldots i_n}) \text{ for all } i_1, i_2, \ldots, i_n \] (6.3.1)
where \( \epsilon_{i_1 i_2 \ldots i_n} \) corresponds to small deviations. Using the normalization condition
\[ \sum_{i_1 i_2 \ldots i_n} \hat{p}_{i_1 i_2 \ldots i_n} = 1 = \sum_{i_1 i_2 \ldots i_n} \hat{p}_{i_1} \hat{p}_{i_2} \cdots \hat{p}_{i_n} \]
we find
\[ \sum_{i_1 i_2 \ldots i_n} \hat{p}_{i_1} \hat{p}_{i_2} \cdots \hat{p}_{i_n} \epsilon_{i_1 i_2 \ldots i_n} = 0. \] (6.3.2)
Substituting Eq.(6.3.1) into Eq.(6.2.5), one finds
\[ d(X_1, X_2, \ldots, X_n) = \sum_{i_1 i_2 \ldots i_n} \hat{p}_{i_1} \hat{p}_{i_2} \cdots \hat{p}_{i_n} (1 + \epsilon_{i_1 i_2 \ldots i_n}) \log_2 (1 + \epsilon_{i_1 i_2 \ldots i_n}). \] (6.3.3)
After simplification, one may write Eq.(6.3.3) in the form
\[ d(X_1, X_2, \ldots, X_n) = \frac{1}{2} \sum_{i_1 i_2 \ldots i_n} (\hat{p}_{i_1} \hat{p}_{i_2} \cdots \hat{p}_{i_n} \epsilon_{i_1 i_2 \ldots i_n}^2 + o(\epsilon_{i_1 i_2 \ldots i_n})). \] (6.3.4)
Neglecting higher order terms \( o(\epsilon_{i_1 i_2 \ldots i_n}) \), we get
\[ d(X_1, X_2, \ldots, X_n) \approx \frac{1}{2} \sum_{i_1 i_2 \ldots i_n} \frac{(\hat{p}_{i_1} \hat{p}_{i_2} \cdots \hat{p}_{i_n} - \hat{p}_{i_1} \hat{p}_{i_2} \cdots \hat{p}_{i_n})^2}{\hat{p}_{i_1} \hat{p}_{i_2} \cdots \hat{p}_{i_n}}. \] (6.3.5)
In order to get further insight into the form in Eq.(6.3.5), we introduce \( f_{i_1 i_2 \ldots i_n} \) denoting the frequency of the class \( C_i \) corresponding to motif \((i_1 i_2 \ldots i_n)\). The marginal frequencies are denoted by \( f_{i_1} = \sum_{i_2 i_3 \ldots i_n} f_{i_1 i_2 \ldots i_n}, f_{i_2} = \sum_{i_1 i_3 \ldots i_n} f_{i_1 i_2 \ldots i_n}, \ldots \) and \( f_n = \sum_{i_1 \ldots i_{n-1}} f_{i_1 i_2 \ldots i_n} \).
Writing, \( \hat{p}_{i_1 i_2 \ldots i_n} = \frac{1}{N} f_{i_1 i_2 \ldots i_n}, \hat{p}_{ij} = \frac{1}{N} f_{ij} \) for \( j = 1, \ldots, n \); Eq.(6.3.5) reads as
\[ d(X_1, X_2, \ldots, X_n) \approx \frac{1}{2N} \sum_{i_1 i_2 \ldots i_n} \left( \frac{N^p-1 f_{i_1 i_2 \ldots i_n} - f_{i_1} f_{i_2} \cdots f_{i_n}}{N^p-1 f_{i_1} f_{i_2} \cdots f_{i_n}} \right)^2 \] (6.3.6)
\[ = \frac{1}{2N} \chi^2. \] (6.3.7)
Consequently one can say that \( 2Nd(X_1, X_2, \ldots, X_n) \sim \chi^2 \) with \( v = 4^n - 3n - 1 \) df.
CHAPTER 6: DEPENDENCE WITHIN MOTIF IN MULTIPLE SEQUENCE ALIGNMENT...

Testing procedure

A test of the hypothesis $H_0$ can be carried out using the test statistic $\chi^2 = 2Nd_k^2$ for a given $n$ and various values of $k \in [1, m-n+1]$. However, in practice it can be seen that values of df corresponding to $n = 2, 3, 4, 5, \ldots$ increase rapidly and $v = 243$ for $n = 4$ and $v = 1008$ for $n = 5$. Kendall and Stuart (1977) state that the statistic $\sqrt{\frac{\chi^2}{v}}$ converges to normality faster than $\chi^2$ and the well-known Fisher's approximation $\sqrt{2\kappa^2}$ with $v$ df. Accordingly, the test statistic corresponding to large df is of the form $\sqrt{\frac{2Nd_k^2}{v}}$, which is approximately normal with mean $1 - \frac{2}{9(4^n - 3n - 1)}$ and variance $2/\{9(4^n - 3n - 1)\}$.

6.4 Results

Computations of entropy or information content for a set of DNA is normally carried out after proper alignment. These alignments are ungapped alignment and also with respect to a reference position. While analyzing upstream sequences, the reference sequence position is generally either the translation start site or transcription start site. Given the alignment, the entropy is computed for each position using the formula given in Eq.(6.2.2). These values are generally shown as sequence logos or graphic plots (Schneider et al., 1986). In entropy computation, the positions (columns) in the alignment are assumed to be independent.

In this study, we have identified a measure of dependence $D$ and its normalized version $\overline{D}$ based on entropy framework. The assumption in respect of $d^p_k$ and $\overline{d}^p_k$ are carried out for a window of size $n$. As we move the window by unit distance (in the overlapping manner), the values of $d^p_k$ and $\overline{d}^p_k$ are computed. In this manner, we obtain a set of values for dependence measure for a fixed $n$ and varying values of $k$. While plotting the values for different $n$, the points are shifted so as to bring in frame,
the corresponding positions.

Given a data set in the form of a multiple aligned sequences, we compute the values of $D$ and $\hat{D}$ for various values of word size $n$. Figure 6.1(a) and (b) show the graphs of $D$ and $\hat{D}$ in case of E. Coli promoter for specific word size $n = 6$. A close inspection of the Figure 6.1(a) shows that the measure of dependence continues to be at around 3.1 with some fluctuations as we increase $k$, exhibiting dependence among the nucleotides. This situation continues to exist till we reach the zeroth position—so called transcription start site around which there is a sudden decrease in the measure of dependence falling to a level close to zero, indicating independence among nucleotides. This marks the onset of transition from dependence to independence which may be on account of biological reasons. It may be worth noting that there are minor dips at other positions as well in the upstream region. These are in the vicinity of positions where biologically interesting features like TATA box (at -10) and CAAT box (at -35) are located. It would be instructive to compare the graphs of dependence measure (see Fig. 6.1 with that of purely randomized sequence data as given in Fig. 6.2.

It can be seen from the figure that the dependence measure fluctuates around value zero indicating general independence in the columns of MSA obtained from random sequences. It is easy to see that graphs of dependence for biologically realistic data sets (see Fig. 6.1 defers markedly from the randomized data sets, (See Fig. 6.2).

Fig. 6.3 and 6.4 show the graphic plot of dependence for the data sets (i) E. Coli promoter and (ii) Ribosome binding sites. As it has been mentioned earlier that low values of dependence correspond to biologically interesting features. Further, it can be seen that by varying the word size the pattern shown in Fig. 6.3 remains more or less identical. It will interesting to examine the effect of increase in word size on discovery of biologically relevant features. This raises a fundamental question as to
Figure 6.1: Plots of (a) $d_k^n$, (b) $\hat{d}_k^n$ for E.Coli promoter Data (471x101) for $n = 6$
the determination of optimal word size.

We have also plotted the standardized dependence measure $\tilde{d}_k^n$ for Human promoter data for varying word sizes, i.e. 7 - 20 (See Fig. 6.5). It is easy to notice the dip near the transcription start site. To get an idea about the degree of dependence, we plotted the minimum value of the standardized dependence measure, i.e. $\min_k \{\tilde{d}_k^n\}$ for various word size.( See Fig. 6.6). It is observed that for the word size 5 to 8 there is a dip. This $n$-mer combination may exhibit strong dependence compared to others. This may be a clue to the degree of dependence one can obtain from the MSA. It is to be noted that hexamer motif (word size) is known to be one of the interesting biological features and many studies analyze over represented hexamers in the upstream data.
Figure 6.3: Plots of $\hat{d}_k^n$ for E.Coli promoter Data (471x101) in the range $n = 7$ to 12

Figure 6.4: Plots of $\hat{d}_k^n$ for Ribosome binding sites Data (1055x40), in the range $n = 7$ to 12
Figure 6.5: Plots of $d^n_k$ for Human promoter Data (4737x151), in the range $n = 7$ to 20

Figure 6.6: Plot of the least value i.e., $\min_k \{d^n_k\}$ against word size for the E.Coli promoter Data
6.4.1 Discussion and Conclusion

The use of Shannon's entropy for characterization of binding sites is well known. The basis for characterization involves computation of entropy or information content from multiple aligned sequences treating the columns (positions) as independent. Noting the assumption of independence as a serious limitation, we have attempted to explore information theory based measure of dependence which can quantify the presence of dependence amongst columns. Identifying the dependence measure in terms of difference between joint entropy of random variables and the respective individual entropy for each random variable, we have worked out the sampling distribution of dependence statistic. This will enable us to test the data sets against the null hypothesis corresponding to independent column hypothesis. Based on the measure of dependence, we have analyzed data for each promoter and CAAT box.

Our study clearly brings out a need for dependence measure in studying MSA. This becomes evident when one compares Fig. 6.1 with Fig. 6.2 based on random MSA. The proposed study indicates the usefulness and relevance of columns in MSA has forcibly been evoked. The dependence measure is based to be biologically meaningful as it considers the dependence among the nucleotides. The utility and power of the dependence measure could be gauged by the fact that it automatically reveals the well known biologically interesting binding sites in the upstream region such as TATA, CAAT. This framework can be easily extended to protein sequences as well. It may be noted that one requires massive data sets to realize the full potential of the framework.