Chapter 4

Distribution of Nucleotides in the vicinity of DNA binding site: Analysis using Jaynes’ Entropy Concentration Theorem

4.1 Introduction

Bioinformatics based approach mainly focus on utilising vast amount of sequence, structure and its derived data to gain insight into biological process. Deriving inferences from the data is an important and challenging task which may reveal and capture the basic underlying process. Transcription factors and other regulatory proteins control and modulate gene expression by specifically binding to DNA sequences in non-coding segments (Lewin, 1997). Information theoretic framework based on Shannon’s measure of information have been extensively employed for studying the distribution of nucleotides in the binding site regions (O’Neill, 1998; Schneider et al., 1986; Schneider, 1997; Shultzaberger et al., 2001; Krishnamachari et al., 2004). After aligning the binding site sequences for a particular protein or biomolecule, Shannon’s measure of uncertainty is computed from the observed nucleotide frequency at each position. The computation of "uncertainty" at each nucleotide position is carried out under the assumption of independence. Based on these assumptions, position weight matrix and profiles are constructed which are used as a tool to search for
potential binding sites.

The graphical representation of the information content at and in the vicinity of a binding site, generally reveal high values at binding site and low values at other positions [Schneider 1986]. The high information content is due to strong conservation at particular positions and it is occurring because of biological reasons. On the contrary the low values at other positions are generally believed to be not carrying any biologically significant features and nucleotides are randomly distributed as pure noise. The need for study of the information content of the background or flanking region is emphasized by Percus (2002) in the context of evolutionary process. Percus observes "the random or null hypothesis with respect to additional information must be assessed. Ignorance of this background bias not only ignores available information but also poses as noise against which a signal must compete".

The assumption of random noise implies that the distribution of nucleotides follows uniform distribution. Accordingly, binding site region is taken to represent a signal (molecular contact points) and the rest is treated as background noise.

An important assumption that has remained untested is: Whether the distribution of information content near and away from binding site position is actually part of random noise or does it also carry a weak signal, or evolving binding site?. The validity of this hypothesis is examined using the theoretical framework based on Jaynes' concentration theorem (Jaynes, 1982). For investigating this aspect, one has to resort to pooling as the available experimental data sets are scanty. The pooling is permissible as the test statistic has the same distribution for all data sets. Our proposed work aimed at examining the assumption of independence empirically using available data sets and also a novel attempt in this direction.
4.2 Entropy, Maximum Entropy Principles and Jaynes

Entropy concentration theorem

Shannon (1948) laid the foundation of "information theory" and this framework has found extensive application in areas of physical, biological, engineering and even in social sciences (Karmeshu and Pal, 2003). In information theory entropy is a measure of the average amount of information required to describe the distribution of some random variable of interest (Cover and Thomas, 1991). More generally, information received can be thought of as a measure of reduction in uncertainty given a specific message.

4.2.1 Shannon’s Entropy:

The Shannon’s formula which computes the average uncertainty "H" (Applebaum, 1996; Krishnamachari et al., 2004) at a position or column of a given multiple alignment is

\[ H = \sum_{i=1}^{4} q_i \log_2 q_i \]  

(4.2.1)

where index \( i \) stands for the four symbols A,T,G,C and \( q_i \) denote the respective probabilities of observing those symbols at that position.

Defining Information content (I) at a position as "decrease in Uncertainty", we write

\[ I = \text{Maximum uncertainty} - \text{Observed uncertainty} \]

and for DNA sequences in terms of its four alphabets A,T,G,C reduces to

\[ I = 2 - \sum_{i=1}^{4} q_i \log_2 q_i, \]  

(4.2.2)
The graphical representation of the information content at and in the vicinity of the binding site, generally have high values at a binding site and low values at other positions (Schneider et al., 1986). The high information content is due to strong conservation of bases at particular positions.

### 4.2.2 Maximum Entropy Principle

Maximum entropy principle can be regarded as the most powerful quantitative tool for inference in probabilistic systems when only partial information is given. The maximum entropy principle in simplest term, as stated by Jaynes "when we make inferences based on incomplete information, we should draw them from that probability distribution that has the maximum entropy permitted by the information we do have" (Jaynes, 1982). Further maximum entropy probability distribution is maximally noncommittal with regard to the missing information (Kapur, 1989; Karmeshu and Pal, 2003).

In many realistic situations, we may have to construct models from partial or incomplete information. A typical example is modeling protein-DNA binding sites, for which only few experimentally verified datasets are available even for the well studied organism like *E. coli*. Hence maximum entropy framework is an ideal choice for studying problems with incomplete or missing data. When dealing with biological system one has to assign probabilities to the states of the system when partial or incomplete information is given. It is possible to construct number of distribution which are consistent with the prior knowledge about the system but the challenging task is to find the probability distribution which is optimum in some sense.
4.2.3 Jaynes Concentration theorem:

Consider a random experiment which can be repeated N times. Assuming, $i^{th}$ outcome results in $N_i = Nf_i$ ways with frequency $f_i$ ($i= 1,2,3...n$). Out of the $n^N$ conceivable outcomes, the number which yields a particular set of frequencies $f_i$ has the multiplicity factor

$$Wf_1, ... f_n = \frac{N!}{(Nf_1)!(...Nf_n)!}$$

On using Stirlings approximation, one finds for large N

$$\ln W \rightarrow NH(f_1, ... f_n),$$

where $H(f_1, ... f_n)$ corresponds to the entropy function.

Following Jaynes (Kapur and Kesavan, 1992), given two sets of frequencies $f_i$ and $f'_i$, the ratio $\frac{W_i}{W'_i}$ of the number of ways $i^{th}$ outcome $f_i$ can be realised to that of $f'_i$ asymptotically yields

$$\frac{W_i}{W'_i} \sim e^{N(H-H')}$$

If H corresponds to maximum entropy $H_{max}$, then it can be easily seen that the number favouring $f_i$ increases exponentially with N. A related question is to have a quantitative demonstration of this entropy concentration theorem. Jaynes (1982) established that asymptotically $2N\Delta H = 2N(H_{max} - H)$ is distributed as chi-square with $k = n - m - 1$ degrees of freedom, where m is the number of linear constraints. This result can be easily understood by resorting to the Kullback-Leibler measure (Kullback and Leibler, 1951; Kapur and Kesavan, 1992).

Noting that $p_i$ corresponds to the maximum distribution and $q_i$ to the observed distribution satisfying the given constraints, then it is known (Kapur and Kesavan, 1992)
\[ \Delta H = H_{\text{max}} - H = \sum q_i \ln \frac{q_i}{p_i} \]  \hspace{1cm} (4.2.3)

Note \[ \sum p_i = 1, \ \sum q_i = 1 \] and \[ H_{\text{max}} = 2 \text{ bits} \]

Writing \[ q_i = p_i (1 + \epsilon_i) \] with \( \epsilon_i \) random errors, one finds \[ \sum p_i \epsilon_i = 0 \]

Neglecting higher order terms \[ \sum p_i \epsilon_i^3, \sum p_i \epsilon_i^4 \text{ etc.} \] one gets

\[ \Delta H = H_{\text{max}} - H \approx \frac{1}{2} \sum \frac{(q_i - p_i)^2}{p_i} \]  \hspace{1cm} (4.2.4)

In terms of total realizations \( N \), Eq.(4.2.4) can be written as

\[ \Delta H = \frac{1}{2} \sum \frac{(Nq_i - Np_i)^2}{Np_i} \]  \hspace{1cm} (4.2.5)

where \( Nq_i \) and \( Np_i \) are the observed and expected frequencies \( O_i \) and \( E_i \) respectively.

This can be seen by observing that

\[ 2N \Delta H = \sum_{i=1}^{4} \frac{(O_i - E_i)^2}{E_i} \]  \hspace{1cm} (4.2.6)

Hence for large \( N \), \( 2N \Delta H \) follows a chi-square distribution with three degrees of freedom. The reduction of the number of degrees of freedom by one is due to the natural constraints.

In essence, for large \( N \), most distributions satisfying given constraints will have entropies close to maximum entropy distribution. In case it is found that the deviation from maximum entropy distribution does not fit the chi-square then there is a possibility that data may be subject to some constraints in addition to the natural one. As observed by Jaynes (1982) "entropy concentration theorem establishes the fundamental status of entropy as a criterion for testing hypotheses about the systematic effects in experiments where frequency data are used".
4.3 Analysis of Binding site data

Three different data sets i.e 1. E.coli promoter data (Hersberg et al., 2001) 2. Data provided in Schneider et al. (1986) and 3. Plant Promoter data (Shahmuradov et al., 2003) were used and all these are pre-aligned sequences.

The first data set contained Four hundred and twenty 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 contained the range of nucleotide positions from -75 to +25 where +1 denotes the transcription start site.

The second dataset pertaining to Schneider et al (Schneider et al., 1986) who investigated several E.coli DNA binding sites using information theoretic principles. The data includes HincII, lexA, ribosome binding site, λ cI/cro and T7 sites. The third set contained three hundred and two Plant promoter sequences were downloaded from the PlantProme database (Shahmuradov et al., 2003) using the URL http://mendel.cs.rhul.ac.uk/ uk/, http://softberry.com. Sequence data ranging from -75 to +25 were extracted where +1 denotes the transcription start site (TSS).

Although each individual data set has scanty observations, the entropy deviation term $z = 2N\Delta H$ from different data sets can be pooled as they all are expected to come from chi-square distribution with three degrees of freedom under the null hypothesis. This is also a convenient way to handle small data.

4.4 Results

For the purpose of illustration, E.coli Hinc II site is shown in Figure 4.1 The positions starting from position -4 to 6 forms the signal and all other neighbourhood regions
are treated as noise. The graph depicts the information content, a measure of uncertainty or variability against binding site positions. This quantity was computed using equation 4.2.2. The peak corresponding to 2 bit conveys that a specific nucleotide out of possible A,T,G,C is completely conserved at this position. Lower Values indicates the richness in nucleotides variability at these positions.

It is to be noted that the values of computed statistic $z = 2N\Delta H$ from different data sets can be Pooled because they are merely independent drawings from a chi-squared distribution with 3 degrees of freedom under the null hypothesis of uniform distribution of the four symbols in the vicinity of the binding site. This provides us with a test of the null hypothesis. We have pooled the $z$ values computed from three published data sets which contain information at several positions in all in the vicinity of the respective binding sites. The binding site positions, which are nothing but the signal, may be identified as outliers of the data. See under the section "data" below for a discussion and detailed reference to the three published data sets.

The outliers have been detected using the criteria: an estimated value of $z$ is an outlier if it is greater than 12.84 i.e. $P[\chi^2 > 12.84] = 0.005$. Based on this criterion the
number of outliers have been detected. The data which are not identified as outliers should be coming from a chi-square distribution with three degrees of freedom under the null hypothesis. The frequency distribution of these data with that of the expected one are shown in Figure 2.

![Figure 4.2: Distribution of entropy deviation values](image)

Fig. 4.2a corresponds to E.coli promoter binding sites (Hersberg et al., 2001) and Fig.4.2b is the distribution is for the pooled data from five E.coli Regulatory Protein binding sites [Schneider 1986]. Fig.4.2c also corresponds to the pooled data from E.coli promoters, five E.coli Regulatory protein binding sites and the plant promoters (Shahmuradov et al., 2003). Pooling of the data enable us to have a fairly
large data set to test our hypothesis. It is to be noted that the outliers are not included while studying the distribution. We have also attempted to examine fit of the data to some well-known skewed distributions such as gamma, lognormal, beta, pareto distributions. We find that in all cases the fit is poor (data not shown). The chi-square test of goodness of fit does not support the null hypothesis even at 1% level of significance.

4.5 Discussion and Conclusion

The fact that the data does not support null hypothesis leads us to suspect that the uncertainty values in the vicinity of a binding site may not correspond to random noise. It is possible that they may be carrying weak signals. This observation also support the known fact that array of transcription factors can bind to these positions with some specificity. This phenomenon needs to be examined in depth from larger data sets. The poor fit of the observed data to any well-known skewed distributions may be due to the fact that the observed data is generally multimodal, suggesting that it could be a mixture of random noise and signals (McLachlan and Peel, 2000). This aspect requires further investigation.

Recent works on ribosome binding sites by Shultzaberger et al. (2001) lends support to our present study. In rigid model of ribosome binding sites the "Translation Start Sites [TSS]" and "Shine-Dalgarno (SD)" sequences are treated separately whereas in flexible model these two sites are treated in conjunction with a spacer element in between. Spacer region is assumed to be not having any signal or shows any marked conservancy in bases. Spacer region is assumed as a surprisal element in this model which may not be correct. When mRNA transcript is made, it consists of Shine-Dalgarno sequence, spacer and coding segment and therefore the assumption of
spacer as surprisal element does not hold good. There exist a strong correlation between spacer and Translation start site and dictated by DNA structural constraints. Spacers paly a major role and contribute directly to DNA binding affinity or property of that particular site.

It is generally assumed that distribution of nucleotides at various positions of the aligned sequences is treated as independent. This is implied in the calculation of information content at each individual position of nucleotides. Moreover position weight matrix and profiles are constructed under these assumptions. They are extensively used to scan for potential sites having some biological functions. It is difficult to define the boundary for a binding site when the underlying distribution is not known. The range for a binding site is normally defined arbitrarily. Statistically validated outlier criterion may be the right strategy. Modelling the regulatory regions needs much more in-depth investigation (Fickett and Wasserman, 2000).

One aspect, which needs to be investigated, is the validity of assumption of independence, which becomes even more questionable as one observes a cooperative binding phenomenon in biological systems (Benos, Lapedes and Stormo, 2002) Recent micro array experiments on Zing fingers Transcription factor binding sites by Bulyk et al. (2002) showed that the binding sites cannot be treated as independent. It was further suggested that the current practice of characterizing binding sites by mutating individual positions one base at a time does not provide a true picture of the sequence specificity. Bioinformatics methods which uses either consensus or mono nucleotide frequency weight matrices rely upon the assumption that the nucleotides of binding sites exert independent effects on binding affinity are not accurate (Bulyk et al., 2002). Uncertainties manifest itself in several forms and may arise from random fluctuations, incomplete information, imprecise perception, vagueness etc. Entropy framework are ideally suitable to deal with this type of situations. Noting this fact
we employed entropy framework based methods to investigate DNA binding sites. In addition Maximum entropy principle find widespread application in time series application, image reconstruction, feature extraction in pattern recognition, computerised tomography, nonlinear system, queueing theory etc.

We further conclude that modeling binding site positions as independent may not be valid and few investigations confirms this fact (Zhou and Liu, 2004; Barash et al., 2003; Yeo and Burge, 2004).