Chapter 2

Sequence variability, long-range dependence and parametric entropy measures

2.1 Background

The statistical analysis of DNA sequence, which can be regarded as a symbolic strings of four nucleotides, is required for understanding the structure and function of genomes. The characteristic of DNA sequences is that they are not statistically homogenous and depict inherent variability. Biological features such as $G+C$ content, $CpG$ islands, the three base periodicity, $A+T$ strand asymmetry, protein binding motifs, origin of replication are known to exhibit variability in the form of statistical fluctuations along the sequence. For gaining insight into the issues of intrinsic variability, researchers have widely adopted the information theoretic framework (Gatlin, 1972; Schneider et al., 1986). Information theory though developed in the context of communication by Shannon (1948), finds wide applications in a variety of disciplines. This is due to the fact that information theoretic entropy provides a
quantitative measure for uncertainty. Thus entropy measure and related concepts are a natural choice for dealing with symbolic sequences like DNA (Rolda et al., 1996).

Information theoretic framework has been widely adopted to study statistical dependence between nucleotides in genomic sequences. One salient feature which has drawn the attention of several researchers is in connection with the existence of short and long range correlation in DNA (Peng et al., 1992; Doukhaan et al., 2002; Herzel and GroBe, 1995; Li et al., 1994; Li, 1997; Voss, 1992). This feature is not surprising as genomic sequence comprises both information (coding) and regulatory (non-coding) segments. Numerous studies have been carried out to understand and model the underlying mechanisms for generation of long-range correlation structure. It needs to be underlined that the term "long-range correlation" has been used differently in the literature dealing with DNA sequences e.g (i) longer than 3-6 bases (ii) more than 800 bases (iii) 1-10kb (Li, 1997). This long-range correlation is generally believed to persist on account of mixture of several length scales in DNA sequences which leads to 1/f noise phenomenon (Voss, 1992). Long-range dependence leads to power-law behaviour affecting the correlations to decay like a power-law instead of short-range correlations that decay exponentially.

Attempts have been made to generate long-range correlations in terms of stationary and non-stationary fractional Brownian motion process (Allergrini et al., 1998). Another promising approach which can be useful for capturing power law behaviour is based on maximum entropy framework due to Jaynes (Kapur and Kesavan, 1992). Parametric entropy measures such as Rényi, Havrada-Charvat-Tsallis, (Kapur and Kesavan, 1992; Karmeshu and Pal, 2003; Tsallis, 1999) are found to be useful in mimicking power law behaviour.

Tsallis entropy though similar to Havrada-Charvat measure, brought to focus the
applicability of parametric entropy measures to a wide range of problems across disciplines. The importance of parametric entropy measures to sequence variability and long-range dependence in DNA is an active area of investigation. To gain better insight, we also give examples of few problems in computational biology where Shannon’s entropic measure has been employed.

2.2 Rènyi and Tsallis entropy measures

Using different set of postulates, further generalizations of Shannon’s measure have been obtained by changing the requirements of information measure leading to some well known Rènyi and Havrada-Charvat parametric entropies. Tsallis rediscovered Havrada-Charvat entropy in a different context seeking probabilistic description of multi fractals geometries (Tsallis, 1999). This led Tsallis to develop non-extensive thermodynamics applicable to a wide ranging problems (Tsallis, 1999).

Rènyi’s entropy (Renyi, 1961) for a probability distribution $p_1, p_2, \ldots, p_n$, $p_i \geq 0$, $\sum p_i = 1$ is defined as

$$H_{R,\alpha}(P) = \frac{1}{1 - \alpha} \log \sum_i p_i^\alpha \quad \alpha > 0, \quad \alpha \neq 1 \quad (2.2.1)$$

It may be noted that one recovers Shannon entropy as the parameter $\alpha$ tends to 1.

Rènyi’s measure of directed divergence is given by

$$D_{R,\alpha}(P||Q) = \frac{1}{\alpha - 1} \log(\sum p_i^\alpha q_i^{1-\alpha}) \quad \alpha > 0, \alpha \neq 1 \quad (2.2.2)$$

Rènyi’s entropy has found wide applications in coding theory, image reconstruction, feature selection (Smolokova et al., 2004).

For the symbolic sequence of DNA, Information content ($I_\alpha$) is then given by the reduction in uncertainty i.e

$$I_\alpha = 2 - \frac{1}{1 - \alpha} \log_2 \sum_i p_i^\alpha \quad (2.2.3)$$
It is seen from Eq. 2.2.3 that Maximum entropy for DNA is equal to 2.

Tsallis introduced non-extensive entropy which is defined as

$$H_q(P) = k \frac{1 - \sum p_i^q}{q - 1}; \quad \sum p_i = 1 \quad (2.2.4)$$

where $k$ is a positive constant and $q$ is referred to as non-extensivity parameter. Shannon's entropy is recovered when $q \to 1$. It may be pointed out that Tsallis and Rényi entropies are closely related. One salient property of Tsallis entropy is that it is not additive. Thus for two independent systems one finds

$$H_{Tq}^A(A, B) = H_{Tq}^A(A) + H_{Tq}^B(B) + (1 - q)H_{Tq}^A(A)H_{Tq}^B(B) \quad (2.2.5)$$

The information content $I_q$ for a typical DNA sequence is then

$$I_q = \frac{1 - 4^{1-q}}{q - 1} - \frac{1 - \sum p_i^q}{q - 1} \quad (2.2.6)$$

The attractive feature of Tsallis entropy (Abe and Rajagopal, 2001; Buiatti et al., 1999) is that it can be uniquely identified by the principles of thermodynamics (Beck and Schlogl, 1993). As observed by Abe and Rajagopal (2001) the thermodynamically exotic complex systems occupy non-equilibrium states for significantly long periods with preserving scale invariant and hierarchical structures (Abe and Rajagopal, 2001). He further observes that "the phase spaces are generically homogenous and accordingly the naive additivity requirement may not be satisfied any more" (Abe and Rajagopal, 2001). Based on this, it can be argued that non-extensive statistical mechanisms become a relevant framework for building statistical mechanics aspects in DNA sequence analysis. (Buiatti et al., 1999; Tsallis, 1999)

### 2.3 Properties of Entropy Measures

Several generalized entropy measures have been proposed by changing or replacing some of the requirements of an information measure. Notable among them are the
parametric entropy measures due to Rényi and Tsallis. Some of the important and useful properties are provided below in brief. For clarity the state space is defined as

\[ \Delta_n = \left\{ (p_1, \ldots, p_n | p_i \geq 0), \sum_{i=1}^{n} p_i = 1 \right\} \] (2.3.1)

We outline the properties and axioms in relation to some entropy measure (Karmeshu and Pal, 2003).

### 2.3.1 Shannon’s Entropy

We are providing below some of the properties of the Shannon’s entropy.

1. Shannon’s entropy possesses symmetry property. This property implies that it does not depend on the order of \( p_i \)'s and for any arrangement \( \alpha_1, \alpha_2, \ldots, \alpha_n \), we have,

   \[ H_n(p_1, p_2, \ldots, p_n) = H_n(p_{\alpha_1}, p_{\alpha_2}, \ldots, p_{\alpha_n}) \]

2. Entropy is a monotonically increasing function of \( n \).

3. The maximum entropy corresponds to equally likely case, i.e., when all probabilities are equal.

4. Entropy function is concave function which means the local maximum is same as global maximum.

5. Shannon’s entropy possesses recursivity property, i.e.

\[ H_n(p_1, p_2, \ldots, p_n) = H_{n-1}(p_1, p_2, \ldots, p_n) + (p_1 + p_2) H_2 \left( \frac{p_1}{p_1 + p_2}, \frac{p_2}{p_1 + p_2} \right), \quad p_1 + p_2 > 0, \quad n \geq 3. \]

6. The entropy \( H_n() \) possesses additivity property i.e., for the independent schemes \( P = p_1, \ldots, p_n \) and \( Q = q_1, \ldots, q_m \), we have,

\[ H_{nm}(p_1 q_1, \ldots, p_n q_m) = H_n(p_1, \ldots, p_n) + H_m(q_1, \ldots, q_m). \]
2.3.2 Renyi entropy

Rényi's entropy (Renyi, 1961) for a probability distribution \( p_1, p_2, \ldots, p_n \), \( p_i \geq 0 \), \( \sum p_i = 1 \) is defined as

\[
H_{R,\alpha}(P) = \frac{1}{1 - \alpha} \log \sum p_i^\alpha \quad \alpha > 0, \quad \alpha \neq 1 \tag{2.3.2}
\]

The above defined entity satisfies the following properties

1. It has additive property i.e For two independent probability distributions, \( P = (p_1, \ldots, p_n) \) and \( Q = (q_1, \ldots, q_n) \), the entropy of the combined distribution is equal to the sum of entropies of the individual distributions. This is mathematically expressed as

\[
H_{R,\alpha}(P \ast Q) = H_{R,\alpha}(P) + H_{R,\alpha}(Q) \tag{2.3.3}
\]

2. \( H_{R,\alpha}(P) \) is a symmetric function of its variables for \( n = 2, 3, \ldots \)

3. \( H_{R,\alpha}(P) \) is a continuous function for \( p \) for \( 0 \leq p \leq 1 \).

2.3.3 Tsallis entropy

Tsallis in the context of understanding the dynamical systems exhibiting multi-fractal structures, proposed a measure which is similar to the one Havrda-Charvat (Havrda and Charvat, 1967) measure which was introduced by him in 1967. Tsallis introduced non-extensive entropy which is defined as

\[
H_{T,\gamma}(P) = k \frac{1 - \sum p_i^\gamma}{\gamma - 1}; \quad \sum p_i = 1 \tag{2.3.4}
\]

where \( k \) is a positive constant and \( \gamma \) is referred to as non-extensivity parameter.

Suyari in a recent paper proposed a set of axioms which can result in Tsallis entropy. By introducing function,

\[
H_{q}(p_1, \ldots, p_n) = \frac{1 - \sum_{i=1}^{n} p_i^\gamma}{\phi(q)} \tag{2.3.5}
\]
where \( q \in R^+ \) and \( \phi(q) \) satisfies the following properties

(i) \( \phi(q) \) is continuous, i.e., \( \phi(q)(q - 1) > 0 \quad (q \neq 1) \) (2.3.6)

(ii) \( \lim_{q \to 1} \phi(q) = \phi(1) = 0, \quad \phi(q) \neq 0 \quad (q \neq 1) \)

(iii) in the interval \((a, b) \subset R^+ \) such that \( a < 1 < b \) and \( \phi(q) \) is differentiable on this

\[ (a, 1) \cup (1, b) \quad (2.3.7) \]

(iv) there exists a constant \( K > 0 \) such that

\[ \lim_{q \to 1} \frac{d\phi(q)}{dq} = \frac{1}{K} \quad (2.3.8) \]

In addition 1. \( H_q \) is continuous in \( \Delta_n \) and \( q \in R^+ \)

2. For any \( q \in R^+ \), any \( n \in N \) and any \((p_1, ..., p_n) \in \Delta_n.\)

\[ H_q(p_1, ..., p_n) \leq H_q \left( \frac{1}{n}, ..., \frac{1}{n} \right) \quad (2.3.9) \]

This referred as maximality.

3. Under the normalization constraint of probabilities, the following equality holds

\[ H_q(p_1, ..., p_{nm_n}) = H_q(p_1, ..., p_n) + \sum_{i=1}^{n} p_i^q H_q \left( \frac{p_{i1}}{p_i}, ..., \frac{p_{in}}{p_i} \right) \quad (2.3.10) \]

This is termed as Generalised Shannon additivity.

4. \( H_q \) has Expandability i.e

\[ H_q(p_1, ..., p_n, 0) = H_{1}(p_1, ..., p_n) \quad (2.3.11) \]

2.4 Maximum entropy principle, Tsallis entropy and powerlaw

There are situations where only partial or incomplete information, say in terms of a first few moments is available. In such cases, there could be infinitely large number of
distributions which are consistent with the given information. A pertinent question in this regard is to choose the 'most objective' probability distribution consistent with the given moments. The principle of maximum entropy as enunciated by Jaynes' states that the most objective probability distribution is one which has maximum entropy subject to the given constraints (Kapur and Kesavan, 1992).

For the purpose of illustration, we consider a case where the first moment "A" is known. The most objective probability distribution consistent with the first moment is to be obtained. Employing Tsallis entropic framework, the problem can be stated as

\[ \text{Max} H_f^q(P) = k - \frac{\sum p_i^q}{q - 1} \]  
subject to

\[ \sum_i i p_i = A \quad \text{and} \quad \sum_i p_i = 1 \]  

Maximization of \( H_f^q(p) \) yields

\[ p_i = \frac{[1 + \beta(1 - q)i]^{\frac{1}{q-1}}}{\sum_i [1 + \beta(1 - q)i]^{\frac{1}{q-1}}}, \quad q > 0 \]  

where \( \beta \) is Lagrange parameter. It is straightforward to see that as \( q \to 1 \), one finds

\[ p_i = \frac{e^{-\beta i}}{\sum_i e^{-\beta i}}, \]  

which corresponds to Boltzmann-Gibbs statistics.

The limiting behaviour of \( p_i \) as given in (2.4.3) for large \( i \) yields

\[ p_i \sim \frac{1}{i^{q-1}}, \]  

leading to power law distribution. Several problems in computational biology depict power-law like behaviour.

Information theoretic measures particularly Shannon's entropy has been widely used as it captures the sequence variability in a variety of problems. Very little
attention has been given to study the advantage and suitability of parametric entropic measures for symbolic DNA and protein sequences. These entropic measures are ideal choice to deal with short range and long-range correlations. Tsallis entropy has been found to mimic power law distributions. This aspect has been emphasized as the non-extensive statistical mechanics is likely to play a role in computational molecular biology.

2.5 Applications of Shannon’s entropy to DNA sequences

Shannon’s entropy measure has been extensively applied to both DNA and protein sequences (Yockey, 1992), but here only few applications dealing with DNA sequences are highlighted. The general framework prescribed for DNA can also be easily be extended to Protein sequences as well.

2.5.1 Sequence variability and pattern OR motif extraction

It is a general practice to perform multiple alignment of closely related sequences and then deduce a consensus pattern for that family of sequences. This is done by considering only the dominant character or base in each column of the multiple alignment and the consensus motif is prescribed. This method has several problems (Schneider, 2002). It is possible to consider the individual contributions of nucleotides and can be quantitatively (Schneider et al., 1986) shown as sequence logos or represent the alignment by a position weight matrix (PWM) using Shannon’s entropy measure (Schneider, 1997).
The weights are proportional to the degree conservancy of each base at each column. Critical positions (i.e., biologically important) residues generally highly conserved and it is absolutely required. Any change in the base composition in these positions may alter the gene expression directly. One may even consider these critical positions as reference points while aligning sites. Near the vicinity of these sites, one may see lowly conserved bases meaning there is a flexibility in base preference. By employing this type of profile or weight matrix, it is also possible to detect weak but related sequence patterns. The weights are generally computed as a log odds ratio, or information content.

This approach is particularly useful for predicting transcription and ribosome binding sites, promoters, and other signals (Pandey and Krishnamachari, 2006). Since the range of the binding sites are relatively short (from 12 to 20 bases approx.) weight matrix methods are simple, an ideal choice, than employing computationally intensive methods such as Markov models, neural networks etc.

2.5.2 Modelling dependency in protein binding sites

Computation of "uncertainty" at each nucleotide position is carried out under the assumption that nucleotides are uniformly and independently distributed as i.i.d variates. Based on these assumptions, position weight matrix and profiles are constructed and are used to search for potential binding sites. The important question is to test this assumption. Markov models which considers positional dependency in DNA sequences have been used extensively.

The simplest model is a classical Markov Chain is a collection of states. Each state correspond to a particular residue e.g, DNA contains four symbols i.e four States. Each position in a typical one dimensional linear DNA sequence can be considered as time points. Hence any given base at a particular position can be expressed in probabilistic
terms as

\[ a_{st} = P(x_t = t \mid x_{t-1} = s) \]  \hspace{1cm} (2.5.1)

Left hand side expression represents the transition probability i.e probability for the transition from the state "s" to the state "t". From right hand side it is clear that both "S" and "t" are neighbouring states and also known as first order ie immediate position.

This probabilistic framework has been successively used for Gene finding and other applications but their applications (Durbin et al., 1998; Ewens and Grant, 2001) to prediction DNA binding sites is still a challenging task. The reason may be due to computational complexity and availability of few experimentally valid sample sequences for learning process. Entropy measures of higher orders also can capture dependency and have been used to study oligomer frequency distribution in genomic sequences. In addition mutual information employs conditional probabilistic terms to capture the dependency factor. Though not much work has been carried out in this regard, still information theoretic measures are powerful to study positional dependency.

### 2.5.3 Predicting gene coding sequences

Many statistical patterns have been found that distinguishes clearly the coding and non-coding DNA. However they differ from species to species and hence specific learning models have to be built.

I Grosse et al (Grosse et al., 2000) proposed a measure called "Average mutual information" (AMI) to identify coding segments. Individual codon positions are considered while computing AMI. The performance of this approach is comparable with
other methods.

2.5.4 Segmenting genomic DNA

Segmentation is a procedure that partitions a given sequence into domains of homoge­
nous composition. It is a known fact that the statistical properties of DNA are not
homogenously distributed along the sequence (Bernaola-Galvan et al., 2000). The
segmentation procedure employing Jensen-Shannon divergence has been used to find
domains, isochores, \(C_p G\) islands, replication origin and terminus, coding and non­
coding sequence borders etc (Azad, Rao, Li and Ramaswamy, 2002; Azad, Bernaola­
Galvan, Ramaswamy and Rao, 2002; Bernaola-Galvan et al., 2000). This simple
entropic based distance method is powerful in detecting many biological features.

2.5.5 Evolutionary analysis

Inferring phylogenetic tree across organisms is important for biologists as it can pro­
vide information and insights about the evolutionary process. Alignments are per­
formed before drawing the trees and proper choice of scoring matrices improves the
alignment. These matrices are evaluated using information theoretic perspective(Altschul,
1991). Recently the relative entropy (or) Kullback-Leibler divergence has been used
to compute evolutionary distances (Robins et al., 2005) and then to build phylogentic
trees.