Dynamics of Continuous, Discrete & Impulsive Systems: Series B Supplement, 2005

A Functional Entropy Model for Biological Sequences

Kirk K. Durston

Department of Biophysics
University of Guelph
Guelph, ON, N1G 2W1
kdurston@uoguelph.ca

David K.Y. Chiu

Department of Computing & Information Science
University of Guelph
Guelph, ON, N1G 2W1
dchiu@snowhite.cis.uoguelph.ca

Abstract-- This paper introduces functional entropy as a measure of entropy that incorporates functional interpretations corresponding to certain biological functions. A measure of change of functional entropy is defined to measure entropy change between two functional states. We show here two biosequence analysis experiments based on the ankyrin repeat and the Ubx hox gene. They show how two related biomolecules with different biological functions can be compared and analyzed. Furthermore, with a given limit on entropy change, intermediaries between states can also be estimated and evaluated.

I. Introduction

Recently, it has become increasingly clear that it is important to incorporate functional interpretations into the measure of information, especially in the analysis of biomolecules. For examples, Winnie suggests that it is the structural organization and the functionality of a bit-string generating program that makes the bit string interesting [1]. Therefore, he argued for a notion of genetic information that is a function of the structural complexity of the genome itself, even though his notion of structure is directly derived from the event sequence and not an interpretation of it. Szostak introduces the term functional information to mean a type of molecularly coded information relating to the biochemical function of biopolymer sequences [2]. MacKay recognizes that it is appropriate to measure information in brain function, using a measure of redundancy, a derivation of entropy [3]. Motivated by this, Chiu has introduced a measure of functional information [7] based on a reduction of surprise between jointly observed data and functional patterns. The purpose of this paper is to show that Shannon entropy can also be redefined as a function of the joint patterns between data and functionality, thus incorporating a functional interpretation into the measure.

Shannon entropy is a measure of variability of data patterns [4]. When the measure is defined on the joint variable (X, F) where X is the variable of data and F is the variable representing a functional pattern, then a measure of Shannon entropy incorporating a functional interpretation can be defined. We call this measure *functional entropy*. For convenience, we denote data corresponding to a defined functional interpretation as X_f.

The difference in functional entropy at different states can be used to provide an information measure in comparing data patterns with different functional characteristics.

In 1951, Brillouin expanded upon Shannon's work and showed that entropy decreases "whenever we happen to have some special information about the structure of our physical system" [5]. Thus, our knowledge of the functional interpretation of the data can be evaluated by the difference or change in a measure of functional entropy. These are evaluated by the different measures of entropy at different states [6].

The measure of functional entropy can be seen as a weighted sum of functional information as proposed by Chiu [7] when each of the joint patterns of a data pattern and a functional pattern are statistically significant. Functional information is then interpreted as the amount of reduction of surprise or uncertainty for the joint pattern.

In the following sections, we shall introduce functional entropy formally. We then define the measure in evaluating change between functional states. It is then applied to examples in biosequence analysis using two biomolecules known as the ankyrin repeat and the Ubx hox protein.

II. Functional Entropy in Biomolecular Analysis

Expanding upon Shannon's definition of entropy, we can define functional entropy as

$$H(X_f(t)) = -\sum P(X_f(t)) \log P(X_f(t)) \qquad (1)$$

where t is the state variable such that the outcome can be fixed, discrete or continuous. When $t = t_i$, we assume the time is fixed. When $t = t_0, t_1, t_2, \ldots, t_m$, then the number of states are defined as discrete time states. The state variable can also be defined with continuous time outcomes, or $t \in \Re$. X_f is the variable of the joint pattern of data and functional pattern. It can be simple, represented as a single variable, or complex, represented by a random n-tuple:

$$X_f(t) = (X_{f1}(t), X_{f2}(t), \ldots, X_{fn}(t)).$$

Functional entropy can be used to compare different sequences at the same state, as $H(X_f(t))$ at a fixed time $t = t_i$. That is, if we let $X_f(t_i) = Y$ and $X_f(t_i) = Z$, we can determine if $H(Y) \approx H(Z)$.

Since biosequences are composed of either nucleotides in the case of DNA or RNA, or amino acids in the case of proteins, discrete changes in those sequences can be evaluated in terms of discrete time states, as $t = t_0, t_1, t_2, \ldots, t_m$. A simple application of this occurs in the case of mutation or mutagenesis at a specified point on the molecule, where the change of functional entropy of the site can be calculated before and after the mutation and evaluated. For larger changes, such as sequence reversal, multiple mutations, a gene split or the lateral transfer of genes or gene segments, the change in functional entropy can be calculated between any

two states t_i, t_j ($0 \le i, j \le m$). We can also evaluate the limits of entropy change between the two states at $t = t_i$ and $t = t_{i+1}$. There may also be cases where functional entropy can be used to evaluate the possibility of missing states or unknown factors between the two states.

In biosequence analysis, the functional variable X_f has outcomes that are functional sequences denoted by f. Each functional sequence selected must be identified as related with a specified functional pattern f. If all outcomes of X_f are functional with respect to f, then

$$\sum P(X_f) = 1.$$

If only some, but not all, outcomes are functional, then

$$\sum P(X_f) < 1.$$

The outcome set of X_f can be selected deterministically or statistically.

The change in functional entropy between two states can be defined as

$$\Delta H\,(X_f(t_i), X_g(t_j)) = H(X_g(t_j)) - H(X_f(t_i)).$$

This can be applied to different sequences at the same time t_i or to the same sequence at two different time states. The change in functional entropy can increase, decrease, or stabilize deterministically or statistically. A specified functional pattern for sequence X can be assumed to be fixed (i.e., f = g) when calculating the functional entropy of a particular sequence. From $\Delta(X_f(t_i), X_g(t_j))$, we can also hypothesize a change of specified functional characteristics. The change can be measured when f = g, or when a novel functionality is considered where $f \neq g$. We can also impose a limit on $\Delta(X_f(t_i), X_g(t_j))$ in some domains such that any change must be below a certain threshold and then evaluate what changes to $X_f(t_i)$ might be possible within that limit.

In order to calculate the functional entropy of a sequence related to a particular function, we can estimate the change in functional entropy from the completely random state at $t = t_i$. This completely random state we refer to as the *null state*. The change in functional entropy from the null state is defined as

$$\Delta^{(0)} H\,(X_f(t_i), X_g(t_j)) = H(X_g(t_i)) - H(X_f(t_i)) \qquad (2)$$

where $X_g(t_i)$ is the variable for all possible sequences, or the complete sequence space and g is the completely random state or the null state. From Eqn. (1) we can define the functional entropy of the null state as

$$H(X_g(t_i)) = -\sum P(X_g(t_i)) \log P(X_g(t_i))$$

where g represents the null, or completely random state. The functional entropy for the null state is associated with an outcome space that contains N possible sequences. If we assume that each possible sequence is equally probable, $P(X_g(t_i)) = 1/N$ and

$$H(X_g(t_i)) = -\sum (1/N) \log (1/N) = \log N. \qquad (3)$$

The change in functional entropy from the null state is, therefore,

$$\Delta H(H_g(t_i), H_f(t_j)) = \log (N) - H(X_f(t_i)). \qquad (4)$$

The rate of change of functional entropy over time is equal to the ratio of change of functional entropy at consecutive time states. For continuous time, it is defined as

$$\Delta^2 H(X_f(t_i), X_g(t_j)) = \Delta H(X_f(t_{i+\varepsilon}), X_g(t_{j+\varepsilon})) \,/\, \Delta H(X_f(t_i), X_g(t_j)).$$

And for discrete times, it is defined as:

$$\Delta^2 H\,(X_f(t_i), X_g(t_j)) = \Delta H(X_f(t_{i+1}), X_g(t_{j+1}))/\, \Delta H(X_f(t_i), X_g(t_j)).$$

If $\Delta^2 H(X_f(t_i), X_g(t_j)) = 1$, then a constant rate of change is indicated. The rate of change is decreasing if the value is less than 1 and increasing if the value is greater than 1.

The estimation of probability $P(X_f(t))$ in $H(X_f(t))$ can be based on a sample population or theoretically formulated. Estimation can be based on variable independence for a complex variable of X_f, or a high-order estimator [8]. The sample population may also be pre-screened in some applications.

There are different methods to estimate ΔH based on different probability estimations. A method can assume equal probability for which amino acids are permitted at each site and can also assume independence between sites. Another method can use *a posteriori* probabilities for amino acids permitted at each site in the sample. Then the *a posteriori* probability estimation can be derived from the observed sample data. Other methods based on high-order estimators can also be used.

III. Experimental Evaluation of Ankyrin

A. *Experiments using ankyrin repeat sequences*

A common protein repeat is the 33-residue ankyrin repeat found in a variety of proteins associated with regulation of cell cycles, ion transport and the interconnection of integral proteins with the spectrin-based membrane skeleton [9]. Its deficiency can cause genetic disorders such as anemia and jaundice. The objective of this experiment was to calculate the functional entropy of the 33-residue ankyrin repeat from the observed data of the biosequence.

B. *Functional entropy of ankyrin repeats*

The first 44, aligned 33-residue ankyrin repeat sequences were taken from the PFAM database [10]. As consistent with most biological studies of sequences, statistical independence between sites was assumed. The number of different amino acids permitted at each of the 33 sites was recorded, ignoring any amino acid that appeared less than 5% of the time for any particular site. The estimated number of sequences with the property of ankyrin is thus calculated by taking the product of the number of amino acids observed at each of the 33 sites as permitted and is denoted as Nf.

The change in functional entropy from the null state was estimated using Eqn. (4). Since each possible ankyrin sequence

was assumed to be equally probably at a fixed time, $P(X_f(t))= 1/Nf$ which resulted in $H(X_f(t)) = \log N_f$. Assuming site independence and ignoring amino acids that occurred less than 5% at a site, the value for $H(X_f(t))$ was calculated. Using Eqn. (4), the functional entropy required to construct a functional 33-residue ankyrin repeat was estimated by converting the change in functional entropy from the null state into a base 2 notation in units of 'bits'.

C. *Calculated results*

From the data sequences, the product of the number of amino acids permitted at each site yielded an estimate of 10^{15} for N_f. The total number of possible 33-site sequences where 20 amino acids were equally probable at each site was $N = 20^{33} \approx 10^{43}$. The probability of finding a functional ankyrin sequence within sequence space was therefore $N_f/N \approx 10^{-28}$. The functional entropy of the 33-residue ankyrin repeat was found to be $H(X_f(t)) = \log 10^{15} \approx 16$. The functional entropy of the null state was $H(X_g(t_i)) = \log 10^{43} \approx 44$. The change in functional entropy from the null state to functional 33-residue ankyrin was, therefore, $H(X_g(t_i)) - H(X_f(t_i)) = 28$. Converting to base 2, the information required to construct a functional 33-residue ankyrin repeat was estimated to be ≈ 93 bits.

IV. Experimental evaluation of Ubx evolution

A. *Simulation experiments*

The Ultrabithorax (Ubx) hox protein can suppress limb development in arthropods. The wild type 276-amino acid *Artemia* Ubx protein was found to have little effect on the suppression of limb development in *Drosophila*. However, mutations in the first 6 C-terminal Ser/Thr residues to Alanine (*Art* Ubx S/T to A 1-5 and 7) can result in strong repression of embryonic limbs [11]. The objective of this computer simulation was to calculate the number of expected trials required to change the wild type Ubx protein sequence (denoted as S_w) to the functionally different, mutated Ubx protein sequence (denoted as S_m). The functional entropy change calculated between S_w and S_m can be interpreted as a quantifier of evolutionary change.

B. *Experimental design*

In this simulation, it was assumed that the change from sequences S_w to S_m occurred through a series of single-step, random mutations at the amino acid level. Since the novel function as expressed did not come into effect until all 6 mutations were in place, the evolutionary path was modeled as a random walk. Each mutational event had to satisfy two criteria. First, the site where each mutation occurred was selected at random and could occur with equal probability anywhere along the 276-amino acid sequence. Second, each mutation selected 1 out of 20 possible amino acids with equal probability. In addition, it was assumed that the target sequence S_m was unchanged from S_w except for the 6 sites that were changed to Alanine. The specific sites that conferred the new function were the 6 mutated sites. (In Ubx hox protein, *Artemia* Ubx S/T are mutated to Alanine at sites 1-5 and 7.) The mutations are assumed to be statistically independent from the remaining 270 unmutated sites.

Using a random walk, the number of possible evolutionary pathways can be very large. Considering the shortest possible path where S_w changed to S_m in 6 steps of mutated amino acids, the number of sites requiring mutation was 6 (denoted here as b). The number of possible amino acids that would satisfy each target site was 1 (or denoted as $a = 1$). The total sequence length remained a constant at $S = 276$. The probability (P_s) that a mutation would occur at one of the sites requiring Alanine decreased with each successful mutation and was equal to

$$P_s = [b - (k - 1)]/S,$$

where k was the step number and $1 \le k \le b$. The probability (P_a) that a given mutation satisfied the target amino acid requirement of Alanine was a divided by the number of possible amino acids, or 20. Since statistical independence was assumed in each of the selections, the probability of one successful mutation step was $P_s * P_m$, or

$$P_s * P_m = a/20*[b - (k - 1)]/S.$$

The probability of obtaining S_m by the shortest path based on mutated amino acids, was therefore

$$P_r = (a/20)b/S * (a/20)(b\text{-}1)/S * \ldots * (a/20)/S$$

$$= b![a / (20S)]^b.$$

Since the evolutionary process was assumed to be a random walk, the expected number of trials T to find the shortest path to S_m was expected to be inversely proportional to the probability P_r or,

$$T \approx 1/P_r$$

$$\approx 1/(b![a / (20S)])^b.$$

Since the novel function did not arise until all 6 mutations had occurred [11], it was assumed that there was no change in functional entropy at intermediate steps unless a successful mutation occurred at one of the functional sites. Since the original sequence S_w did not have the novel function, the functional entropy of the 6 sites in S_w was assumed to be in the null state, relative to the novel function in S_m. Since the changes between S_w and S_m were only in the 6 specified sites, the change in functional entropy between the two sequences was assumed to be the changes between the 6 sites.

The change in functional entropy of the 6 sites during mutation was assumed to be a 2-event process for each site, where one event was the choosing of the proper site and the other event was the choosing of the proper amino acid out of a possible 20 amino acids, all with equal probability. Each selection event was assumed to be independent to each other. The probability of a successful mutation at one of the required sites was estimated as

$$(a/20)[b - (k - 1)]/S; \; 1 \le k \le b.$$

The change in functional entropy between the two sequences was the difference between the functional entropy of the null state for the 6 sites, and the functional entropy of the final state for the 6 sites.

C. Functional entropy from the simulation results

The functional entropy of the null state for the 6 sites was found to be $H(X_g(t_i)) = \log 20^6 \approx 7.8$. The functional entropy of the 6 sites after successful mutations was found to be $H(X_f(t_i)) \approx .007$. The difference in functional entropy between the two states was, therefore, $H(X_g(t_i)) - H(X_f(t_i)) \approx 7.8$. Converting to base 2, the difference was found to be ≈ 26 bits. The number of expected trials required to find the shortest evolutionary path from S_w to S_m was found to be $T \approx 3.9 \times 10^{19}$ trials.

V. Discussion

The first experiment estimates the change in functional entropy from a null state to a functional state using ankyrin repeat sequences. The second experiment estimates the change in functional entropy from one functional state to a functional state different from the original one. The difference in functional entropy between the null state and a functional state is useful in providing an estimate for the amount of information required to construct a sequence with the particular function under investigation. This is not to be confused with the amount of information a sequence actually carries. No correspondence has been made here between the information required to configure a functional sequence, and the functional information the finished sequence actually carries.

The difference in functional entropy between the two different sequences not only provides an estimate for the amount of information required to change the starting sequence into the final sequence, but it also calculates the estimated number of trials to achieve the final sequence in evolution. From the simulation study, one can see that even a small amount of information may require a very large number of trials. If the number of trials becomes too large such that the target sequence is not likely to be obtained, then a functional, intermediate sequence may exist, reducing the number of expected trials. Thus, the method demonstrated in the simulation study may provide a way of predicting the existence of intermediate states that could have some selective value in the process of evolution.

In a similar way, the method used in the simulation study could be adapted for predicting how far a viral strain could mutate over time, giving a search area in sequence space within which possible novel strains could be looked for. A variation of this could also be applied to rapidly mutating viruses to predict new viral genotypes. Data including population size, replication rate, and information about possible other functional states for certain rapidly mutating genes within a viral genome can be analyzed to predict future strains, as well as reconstructing a possible evolutionary history.

VI. Conclusions

We have introduced a measure of functional entropy, as an extension of Shannon entropy, to biosequence analysis. Changes in functional entropy can provide a method to estimate the relationship between two sequences, as well as the number of trials to achieve one sequence using the other sequence as a starting point. Intuitively, in biomolecules, two sequences that produced the same fold and biological function would possibly have little difference in functional entropy. If an upper limit for the estimated number of trials is set due to imposed constraints, then the method can also be used to predict possible intermediate states. Thus this paper has demonstrated the usefulness of functional entropy in biosequence analysis.

References

[1] J.A. Winnie, "Information and Structure in Molecular Biology: Comments on Maynard Smith, *Philosophy of Sciences*," Vol. 67, pp. 517-526, 2000.

[2] J.W. Szostak, "Molecular messages," *Nature* Vol. 423, p. 689, 2003.

[3] D.M. MacKay, "Information Theory and Brain Function," *Encyclopedia of Neuroscience*, Adelman, G. (ed.), second edition, 1999.

[4] C. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, Vol. 27, pp. 379-423; 623-657, 1948.

[5] C.E. Brillouin, "Physical entropy and information II," *Journal of Applied Physics*, Vol. 22, pp. 338-343, 1951.

[6] L.L. Gatlin, *Information Theory and the Living System*, Columbia University Press: New York, 1972.

[7] D.K.Y. Chiu, "An informatics approach to bioinformatics," *Data Complexity in Pattern Recognition*, T. Ho and M. Basu (eds.), Springer Verlag, in press.

[8] C.K. Chow and C.N. Liu, "Approximating discrete probability distributions with dependence trees," *IEEE Trans. on Information Theory*, Vol. IT-14, pp. 462-467, 1968.

[9] L.K. Mosavi, D.L. Minor Jr., Z. Peng, "Consensus-derived structural determinants of the ankyrin repeat motif," *PNAS,* Vol. 99, pp. 16029-16034, 2002.

[10] http://pfam.wustl.edu/

[11] M. Ronshaugen, N. McGinnis, W. McGinnis, "Hox protein mutation and macroevolution of the insect body plan," *Nature*, Vol. 415, 914-917, 2002.