The First Gene, **David L. Abel**, Editor 2011, pp 147-169 ISBN: 978-0-9657988-9-1

5. Functional Sequence Complexity in Biopolymers

Kirk K. Durston & David K.Y. Chiu

Department of computer science, Bioinformatics
University of Guelph
50 Stone Road East, Guelph, ON, Canada, N1G 2W1

ABSTRACT. It is generally recognized that biopolymers such as DNA, RNA and proteins demonstrate a form of sequence complexity. Recent work has provided a more detailed insight into biopolymeric complexity by introducing three types of sequence complexity, Random Sequence Complexity (RSC), Ordered Sequence Complexity (OSC) and Functional Sequence Complexity (FSC). The primary feature of FSC that distinguishes it from RSC and OSC, is the imposition of functional controls upon the sequence. In this paper, we propose that it can be measured using an extended form of Shannon uncertainty that includes a variable of functionality. Clearly, FSC can be found in human languages and carefully designed computer code, but the measure we propose in this paper reveals that it is also found in biopolymers. In the case of proteins, the measure of FSC provides an estimate for the target size of a protein family in the amino acid sequence space, revealing that functional sequences occupy an extremely small fraction of sequence space. Due to the miniscule size of functional sequence space for a given protein family, as mutations accumulate there will be an increasing likelihood of moving the mutated sequence outside that space, with a corresponding deleterious effect on FSC.

Correspondence/Reprint request: Dr. Kirk Durston, Dept. of Computer Science, University of Guelph, 50 Stone Road East, Guelph, Ontario, Canada N1G 2W1 E-mail: kirkdurston@gmail.com

Introduction: sequence complexity in biopolymers

It has recently been pointed out that traditional notions of complexity are inadequate when applied to biosequences [1, 2]. For example, characterizing biosequence complexity in terms of algorithmic complexity fails to account for the redundancy found in numerous different sequences even when they have the same function [1]. Functional controls imposed upon a biological sequence are critical for maintaining specific functions of the sequence within the cell and, ultimately, for the existence of life. A more rigorous formulation for complexity in biosequences that incorporates functionality is therefore required. Abel and Trevors have defined three types of sequence complexity, only one of which accounts for functional controls imposed upon biosequences such as DNA, RNA and proteins. We will discuss these three types of complexity within the context of biopolymers, with a special focus on that form of sequence complexity that incorporates functionality.

1. Random sequence complexity

Abel and Trevors have defined Random Sequence Complexity (RSC) as *a linear string of stochastically linked units, the sequencing of which is dynamically inert, statistically unweighted, and is unchosen by agents; a random sequence of independent and equiprobable unit occurrence* [3]. Implicitly, four components contribute to RSC. First, the sequence is composed of sites, or loci. Second, there is the importance of the symbols that could occupy each site in the sequence. Third, there is a complete absence of

constraints and controls on these symbols, statistically making all options equiprobable. Finally, the value of the symbol at each site must be independent of the values at any other site, such that no site is constrained by any other site in the sequence. An example of RSC can be found in atactic polystyrene, where the orientation of the side chains at each site appears to be completely unconstrained. In summary, if no agent or law of nature controls or constrains the outcomes of any site in a sequence, then they are presumed to be equiprobable, and the complexity of the sequence is characterized as RSC.

2. Ordered sequence complexity

Ordered Sequence Complexity (OSC) is defined as *a linear string of linked units, the sequencing of which is patterned either by the natural regularities described by physical laws (necessity) or by statistically weighted means* (*e.g., unequal availability of units*)*, but which is not patterned by deliberate choice contingency* (*agency*) [3]. Examples of OSC are repeating patterns arising out of chaotic interactions or a string of repeating alphabet characters such as TGTGTGTGTGTG ... In nature, OSC is presumed to occur when laws of nature impose such tight constraints that there is no possibility of variation. In this case, repeatable, highly constrained sequences are produced that cannot, therefore, incorporate new functional inputs as functional information. An example of OSC is the highly ordered and repeating sequence obtained through the formation of polyadenosine absorbed onto the surface of montmorillonite clay [4].

3. Functional sequence complexity

Given the limitations discussed above, neither RSC, OSC, nor a combination of the two, are capable of producing significant levels of FSC since neither, by definition, are controlled by functionality [5]. Szostak [1] has further pointed out that, traditionally, neither algorithmic complexity [6] nor Shannon's measure of uncertainty [7] is adequate for biopolymers. Functional Sequence Complexity (FSC) is therefore defined as *a linear, digital, cybernetic string of symbols representing syntactic, semantic and pragmatic prescription; each successive symbol in the string is a representation of a decision-node configurable switch-setting---a specific selection for function* [3]. Volitional agency (control) is implicitly required to properly set each configurable-switch-position symbol to achieve functionality. Examples of FSC are said to occur in well-designed computer code and, naturally, in human languages. For biopolymers, functionality can be a result of structural requirements of protein families [17], cellular processes, or specific biochemical reactions [8]. Furthermore, biological functions can be nested in a hierarchical manner from the sub-molecular domain structure necessary for the 3D structure of an enzyme, all the way up to the global function of entire species of organisms. Comparing the differences between OSC and RSC on the one hand, and FSC on the other, it is the requirement of functionality that is the distinguishing feature between them.

Recent advances in the synthesis of RNA chains in water are encouraging so far as providing a storage medium for prescriptive information

and FSC [9]. However, the much greater challenge of encoding FSC within RNA remains. If a RNA sequence is highly ordered, it will tend toward OSC. If the highly ordered sequence can mutate, it will tend toward RSC over time. To become functional, controls will be required to properly configure each switch-setting (nucleotide) to select for function.

4. Measuring FSC

The definition of FSC supplied above is essentially a definition of functional information. Shannon uncertainty is well known as a method to measure variability in data and as a measurement of information. Unfortunately, Shannon uncertainty makes no distinction between functional and nonfunctional variability and complexity, as Szostak has pointed out [1]. Hazen, Szostak *et al.* have advanced an equation for the measurement of functional information as follows:

$$I(E_x) = - \log_2[M(E_x)/W] \qquad (1)$$

where E_x is the degree of function x (a measure of a sequence's functionality with regard to function x), $M(E_x)$ is the number of different sequences or configurations that meet or exceed E_x, and W is the total number of possible sequences or configurations. (Note that Hazen *et al*. use the notation N instead of W, but W is used here to be consistent with the notations and equations that follow.) Here we present an alternate method to measure functional information where an estimate of the probability distribution may be required. For example, in the case of a protein family the data may provide only the

probability of each amino acid at each site, but not the total number of functional sequences. Our proposed method allows uncertainty to be managed when sequences with functionality are not exactly known. It also provides further analyses making use of the sequence distribution obtained [10].

Shannon uncertainty can be modified as a joint measure to analyze available data when the data is known to represent a particular function and is entered as input. This modified form of Shannon uncertainty we call *functional uncertainty* (H_f) [11] and is defined as:

$$H(X_f(t)) = - \sum P(X_f(t)) \log P(X_f(t)) \qquad (2)$$

where X_f denotes the data specified with known functionality and t represents time. In the case of a protein family, the data X_f is in the form of a multiple sequence alignment specified by their family label when downloaded from Pfam [12]. When the dataset, composed of a multiple sequence alignment, is corrupted either by irrelevant sequences or irrelevant amino acids within an included sequence, there are methods such as 'noise cleaning' to address that problem. $P(X_f)$ is the *a posteriori* probability of the data with the given functionality F = f, or $P(X_f) = P(X \mid f)$. An explanation as to how this is calculated for proteins is given in section (5).

It may be useful to measure the change in $H(X_f)$ if certain mutations, insertions or deletions occur between time i and time j resulting in a loss, gain, or change in function. For this reason the time variable t is included in Eqn (2). For example, for a protein family that shares a common 3D structure that performs a known, specific functionality task f, X_f represents the dataset X of known sequences that satisfy the functionality f. Changes in sequence due to

mutations may introduce a change in the specified functionality between time *i* and time *j*. (We are currently developing methods to address the dataset when sub-functionality is considered, such as a portion of the dataset coding structural domains within the larger structure of a protein. The sub-functionality in this case could be the function of contributing a critical structural component within the larger 3D structure that defines the larger biological function. In this view, functionality can form a nested hierarchy, composed of lower levels of different functionalities contributing to higher levels of global functionality.

There are three states of uncertainty to consider when measuring functional complexity. The *ground state g* is the state of greatest uncertainty permitted by the constraints imposed by the physical system when no biological function is required or present. Since the physical system may impose constraints on what type of sequences are permitted, it may be the case that not all sequences are equally probable. A special case of the ground state occurs when the physical system imposes no constraints on sequencing whatsoever, with the result that all possible sequences are equiprobable. This special case is classified as the *null state ø*. The third state is that which produces the function under investigation, denoted as the *functional state f*.

The measure of FSC, denoted as ζ, is the change in functional uncertainty between the ground state *g* and the functional state *f*, or

$$\zeta = \Delta H\,(X_g(t_i), X_f(t_j))\ . \qquad (3)$$

For proteins, the data suggests that actual dipeptide frequencies and single nucleotide frequencies for proteins are closer to random than ordered [13]. For this reason, the ground state *g* for biopolymers can be approximated by the null state *ø*. If we let the number of all possible sequences be represented by *W* and the length of each sequence by *N* and the number of options at each site in the sequence be denoted by *m,* then $W = m^N$. For example, for a 300 amino acid (aa) protein, if we assume 20 aa options per site, then $W = 20^{300}$. If the FSC if a single column within a multiple sequence alignment is being measured, then $N = 1$ and $W = m$. If the FSC of an interdependent cluster of sites is being measured, then N = the number of sites in the cluster. Since for the null state, all options are equally probable, $P(X_ø(t_i)) = 1/W$ and

$$H(X_ø(t_i)) = - \sum (1/W) \log (1/W) = \log W. \qquad (4)$$

The measure of FSC, therefore, reduces to

$$\zeta = \log (W) - H(X_f(t_i)). \qquad (5)$$

If one wishes to take into account the effect of the genetic code on the various *a priori* probabilities of generating the amino acids, then the probability of producing each amino acid given the genetic code can be used to compute a ground state that will be different from the null state, since all amino acids are no longer equiprobable.

With the exception discussed shortly, it is usually the case, in measuring FSC, that the variable *t* is constant, in which case

$$\zeta = \log (W) - H(X_f). \quad (6)$$

The value ζ is a measure of the FSC, or functional information, of any sequence, including biopolymers. As shown above, it is a measure of the change in uncertainty between the ground state and the functional state. This difference in uncertainty is closely related, as is Eqn. (2), to Shannon information and Shannon uncertainty respectively [14]. However, as previously noted, Shannon information is not concerned with function directly. FSC, on the other hand, is inseparable from function and can be regarded, therefore, as a measure of *functional information*, a necessary concept in biology [1, 2]. Since ζ is a measure of functional information, once ζ is known, it can be substituted for $I(E_x)$ in Eqn. (1) and an estimate for the total number of functional sequences $M(E_x)$ can be calculated. Also, the probability of finding a functional sequence in a single search can be estimated by solving Eqn (1) for $M(E_x)/W$.

Change in FSC can be used as a method to quantify evolutionary distance. The change can be between an existing or non-existing function f_a to a modified function f_b between time t_i and t_j described by

$$\Delta\zeta = \Delta H (X_{fa}(t_i), X_{fb}(t_j)). \quad (7)$$

The sequences corresponding to X_{fa} with initial function f_a have two components relative to that of X_{fb} (with resulting mutated function f_b). The *static component* is that portion of the sequence that must remain within the

permitted sequence variation of the original biosequence with function f_a. The *mutating component* is the portion of X_{fa} that must change to achieve either the new function f_b, where the new function is to be understood as either a new level of efficiency for the existing function, or a novel function different from f_a. The mutating component can be assumed to be in the null state relative to the modified function f_b. To clarify, the mutating component, at the outset, is non-functional with respect to the novel function, so the probability of any particular amino acid at a site can be assumed to be equal to the probability of any other amino acid at that site. Since the mutating component is the only part that must change, the static component can be ignored *provided the probability of it remaining static* is included between t_i and t_j. The static probability would be assessed on the basis of the total number of mutations required for the mutating component to achieve functionality and the probability that none of those mutations occur in the static portion. There may be other factors as well in the computation of the static probability, which may also require inclusion in the calculation of the static probability.

5. Application of FSC to protein sequences.

One application of FSC is to protein families and protein structural domains. Measuring the FSC of a protein family can quantify the target size in sequence space for that family or structural domain which, itself, quantifies the degree of difficulty in locating any sequence at all that falls within that target area defined by the same 3D structure or function.

A measure of the lower bound for the FSC of a protein is to assume that each site is independent of all other sites in the sequence. This will yield an

artificially low estimate as discussed shortly, and therefore gives a lower bound. First, a sequence alignment for the protein family or domain being investigated can be downloaded from a web database such as Pfam [12]. It is assumed that the data contains functional sequences, including neutral but functional mutations, with the non-functional sequences filtered out by natural selection. The next step is to compute the functional uncertainty of each site in the sequence. This is done by first calculating the probability of each amino acid occurring at each site. For example, if there are 1000 sequences in the alignment, and proline occurs a total of 235 times at a particular site, then the probability of proline occurring at that site is .235. This is done for each of the 20 commonly occurring amino acids. The functional uncertainty of that site is then computed using Eqn. (2) inputting the 20 amino acid probabilities for that site (ignoring chirality and non-biological amino acids). The functional uncertainty of the entire sequence is obtained by [15] summing all the values obtained for the functional uncertainty of each site in the sequence. The FSC of the protein is then computed using Eqn. (6).

It is much more likely to be the case, for most proteins, that certain sites within the amino acid sequence are associated with other sites in the same sequence, forming 2^{nd}, 3^{rd} and 4^{th} order associations containing one or more amino acid patterns [16, 17], where a 2^{nd} order association is an association between two sites, a 3^{rd} order cluster is an association between three sites, and so on. These associations can be detected through various pattern discovery methods [18, 19, 20, 21]. Measure of FSC becomes more accurate when the sequence of individual sites is transformed into a sequence of individual site clusters. Within each site cluster, there may be one or more amino acid

patterns. The transformation consists of replacing the sequence of sites with a series of non-overlapping site clusters. The functional uncertainty of each site cluster is obtained by observing the *a posteriori* probability from the data of each amino acid pattern within the site cluster. To clarify, Eqn. (2) is applied to each amino acid at one site with respect to what amino acids are associated with it at the other sites in the cluster. The analysis, therefore, runs horizontal across a two dimensional array containing the multiple sequence alignment, where each row represents a different functional sequence and each column represents an aligned site in the sequences representing a protein family. The functional uncertainty of that site cluster is then computed using Eqn. (2), inputting the probabilities of each of the observed amino acid patterns for each site cluster. Next, the functional complexity of each cluster must be computed, where the null state permits any possible amino acid pattern. For example, for a 4^{th} order site cluster, there are a total of $W = 20^4$ possible patterns of amino acids. The functional uncertainty of the null state will depend upon the order of the cluster. The FSC of the site cluster is computed using Eqn. (6), but the variable W represents the total number of possible amino acid patterns, rather than the total number of possible sequences. The total FSC of the protein is then the sum of the individual FSC values for each site cluster within the sequence of sites. In summary, the primary difference between assuming site independence and site inter-dependence is that FSC is computed using probabilities of individual amino acids at individual sites in the case of site independence, and using probabilities of individual patterns of amino acids within clusters of interdependent sites in the case of site inter-dependence. In both cases, equation (2) is used but the unit of data X_f changes depending upon whether individual amino acids at individual sites are the focus (assuming site

independence) or individual amino acid patterns within individual site clusters are the focus (assuming site inter-dependencies).

To illustrate the improvement in the accuracy of measuring FSC when site associations are taken into account, and to contrast the difference in measured FSC between site independence and site inter-dependence, consider a hypothetical 3rd order site cluster. Assume each site in the cluster contains all 20 amino acids and each amino acid is observed to appear an equal number of times in the 1000 sequence alignment. However, each amino acid in the first site in the cluster is uniquely associated with a specific amino acid in the other two sites. If we assume site independence, then since all 20 amino acids appear an equal number of times in all three sites, the site cluster appears to be in the null state and the FSC of the site cluster is 0, since there is no difference between the null state and the functional state in this particular case. The observed amino acid patterns, however, indicate that there are only a total of 20 aa patterns in the site cluster out of a total possible 20^3 patterns. Since each pattern occurs an equal number of times within the 1000-sequence alignment, the probability of each pattern is .05. Using Eqn. (2), the functional uncertainty of the site cluster is 1.30. The functional uncertainty of the null state is log (20^3) or 3.90. The FSC of the cluster, therefore, is 2.60, significantly higher than the lower bound of 0 in this hypothetical case. In reality, there may be fewer patterns per site cluster, some patterns may not be visible due to incomplete data, and the patterns are unlikely to occur with equal probability. Nevertheless, it should be clear as to the importance of considering interdependencies between sites when computing FSC. For the purpose of this paper, however, we shall assume the simplest case of site independence.

Using the site independent assumption and the method described above, the lower bound for the FSC of various proteins can be obtained, with results shown in Table 1. The proteins in Table 1 were chosen because all of them are universal proteins found throughout biological life. Additional results have been published by Durston *et al.* [22]. Results in Table (1) are slightly different from those published earlier due, primarily, to using the genetic code constraints as the ground state, rather than the null state as published earlier.

Table 1: FSC Results for Four Universal Protein Families using the Genetic Code frequencies as the ground state

Protein Family	sites	Number of unique sequences in data	FSC (Fits)	Fits per site	Probability of locating a functional sequence in a single search for same-length sequence space	Estimate for upper limit of functional sequences $M(E_x)$
Ribosomal S12	122	1774	346	2.8	10^{-104}	10^{55}
Ribosomal S7	149	535	359	2.4	10^{-108}	10^{93}
Ribosomal S2	211	2469	465	2.2	10^{-140}	10^{135}
RecA	320	4301	976	3.0	10^{-294}	10^{122}

Table (1) shows lower bound of FSC since we are ignoring any additional constraints imposed by other sites in the sequence. The probability of locating a functional sequence in a single search is derived from the FSC of the protein family. Once we have solved for ζ we can then solve for $M(E_x)/W$ using Eqn (1). This will be an upper probability limit due to the fact that we are assuming no interdependencies between sites. Site interdependencies will introduce additional constraints which will reduce the number of possible functional sequences, as illustrated earlier in this section (5). Thus, assuming site independence gives an upper limit for the number of functional sequences $M(E_x)$ and, therefore, an upper limit for the probability $M(E_x)/W$ of locating a functional sequence in a single search. It should also be noted that W is a lower limit since, as noted earlier, $W = m^n$ where, for proteins, $m = 20$ and n = the length of the sequence, the total sequence space is radically reduced to just n-aa sequence space. Realistically, a search of sequence space is not limited to just the length of the sequences in the protein family being analyzed. Therefore, sequence space target size shown in Table 1 is only for the sequence space for the same length protein.

If all of amino acid sequence space is used for even just up to 300-aa sequence space, the probability of locating a functional sequence for a given protein family would be many orders of magnitude smaller, since W would be many orders of magnitude larger. In summary, if site independence is assumed, and given the artificially low value of W, the value of FSC calculated this way is artificially low and can safely be taken to be a lower bound. Similarly, the probability of locating a functional sequence within a protein family in sequence space is likely to be much smaller by numerous orders of magnitude. This, coupled with the results shown in Table 1, underscores the

almost infinitesimal size of functional sequence space relative to the size of the entire sequence space for a given number of sites.

6. Relationship between RSC, OSC and FSC

Preliminary attempts have already been made to model the relationship between RSC, OSC and FSC [3, 22]. The following model improves upon those earlier attempts and is consistent with the method to measure FSC discussed earlier. It may not be the only way to model this relationship, but may provide a helpful model for the comparison of RSC, OSC and FSC. Figure 1 models one approach to describing the relationship between the three types of sequence complexity, portrayed as a three dimensional coordinate system, with the X coordinate representing RSC, the Y coordinate representing OSC, and the Z coordinate representing FSC.

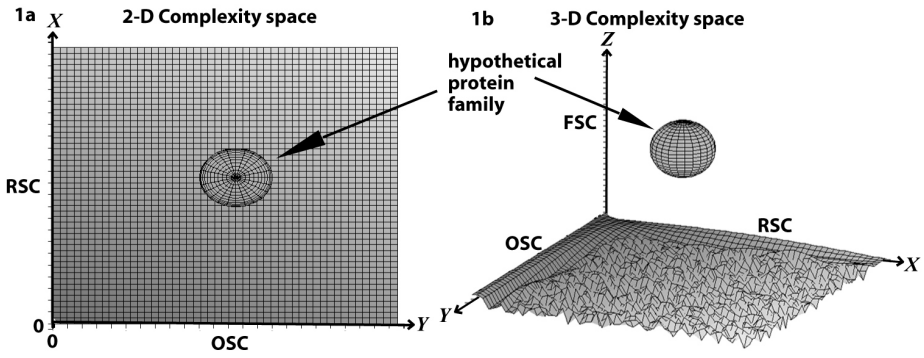

Figure 1: Relationship between RSC, OSC and FSC. In 1a, 2-D complexity space, composed of RSC and OSC, is inadequate to distinguish FSC from RSC and OSC. A third coordinate is necessary, representing the information required to achieve the function, which is a function of probability from Eqn. (1). In 1b, the uneven surface of the XY plane represents low-level, statistically insignificant FSC that can stochastically occur without any controls imposed on the generated sequences. The FSC of a hypothetical protein family can clearly be distinguished from RSC and OSC in this 3-D coordinate model of complexity space.

A very short repeating sequence would be an example of OSC and would be placed closer to the axis than a longer repeating sequence. Similarly, a short random sequence could be an example of RSC and would be placed closer to the axis than a longer random sequence. A sequence consisting of repeating random sequences would have components of both RSC and OSC

and could be placed somewhere on the XY plane, as would any non-functional sequence that contained a mix of RSC and OSC.

For both OSC and RSC, the magnitude of their values is contingent upon the sequence length. This is not the case for FSC. From Eqn. (2), the magnitude of FSC is a function of the probability of finding a functional sequence in a single blind search, also a function of target size in sequence space, as has already been discussed. That probability is determined, in the simplest case, by the ratio of the number of sequences that will produce the function, $M(E_x)$, over the total number of sequence options W, both functional and non-functional, as used in Eqn. (1). This ratio also represents the target size, in sequence space, of the region that produces the function.

The location of FSC relative to the horizontal XY plane is plotted according to the combination of RSC and OSC within the sequences when the sequence, or set of sequences in the case of a protein family, is assumed to be non-functional. The size of the FSC of a protein family, along the Z-coordinate, allows for some variation of efficiency about the optimum value. From the examples in Table 1 and from Eqn. (2) and (6) it can be seen that the greater the FSC, the less probable a functional sequence becomes which, therefore, results in a greater quantum jump from the horizontal X-Y plane. This leads to the conundrum of how functional biopolymeric sequences such as protein families can be discovered in the overall sequence space when, necessarily, the higher the FSC, the less probable it becomes and, from Table 1, those probabilities are quite miniscule.

7. Biological FSC is unique in nature

Although FSC is found within human languages and computer code, the only known occurrence in nature occurs within biopolymers such as DNA, RNA and proteins [3]. To appreciate the oddity of FSC in biopolymers, it is essential to understand the distinction between constraints and controls. *Constraints* are imposed upon any physical system found in nature by the deterministic laws of nature, relevant initial conditions, and probabilistic limitations to possible outcomes [23]. The deterministic outcome of constraints is often referred to as *necessity*. Deterministic necessity steers a process toward repeatability and order. Thus, sequences determined by the constraints thus described will be highly ordered, demonstrating high OSC. The relaxation of constraints, or necessity, permits greater degrees of freedom in the outcome and is often referred to as *chance*. Chance, itself, is not causal but is permitted by the relaxation of the constraints imposed by the laws of nature, initial conditions, and probabilistic boundaries [24]. Chance, or the relaxation of necessity, in sequence formation, permits RSC. Thus, natural processes can be summarized by chance and necessity, the interplay of which produces OSC and RSC. Natural processes, therefore, are known to be limited to producing effects that lay in the XY/OSC-RSC plane in Figure 1. These effects determine the ground state of any sequence, as described earlier. FSC, however, requires a deviation from the ground state to achieve a particular function, as shown in Eqn. (3). To clarify, FSC requires a deviation from the normal results of natural processes operating in the XY plane (Figure 1). Small deviations from the ground state are statistically possible, such that low-level FSC can be achieved for very low-level functions, but the greater the FSC value ζ required by a particular function, the greater the deviation from the ground state. FSC, therefore, represents an *anomaly* within nature, a deviation from mere chance

and necessity. To clarify, FSC represents something that one would not predict given the natural processes of OSC and RSC; it is improbable and the higher the FSC, the more improbable it becomes as shown experimentally (column 6, Table 1). Biological proteins require very high levels of FSC, as illustrated in Table 1. For this reason, chance and necessity define the ground state of a sequence, but are inadequate to produce the kind of cybernetic algorithms found within the genomes of life [25].

The cybernetic requirements of biological life, encoded within biopolymers such as DNA, require the addition of *controls* which can steer the physical system away from the ground state to produce the desired function [23]. Each step, or site, in a sequence represents a decision node that determines the course of physicodynamic events in such a way that the physical system can be steered in the direction of the desired function. Computer code, for example, is an illustration of FSC in the encoding of prescriptive information, using a physical medium, to steer the physical system (a computer) such that a desired function is achieved. FSC is only known to be achieved by encoding formal choices into a physical information storage medium, such as DNA through a series of volitional decisions at each configurable switch symbol. The large deviation from the ground state determined by constraints is an indicator that it has been achieved through the application of controls that arise out of choice contingency or volition. It is easy to understand why human languages and designed computer code exhibit such a high degree of FSC, since human intelligence provides the necessary volitional agency required to select the proper switch configurations. The high level FSC observed in biopolymers, however, successfully locates miniscule areas of sequence space, as shown in column 6 of Table 1, requiring very tight

controls. The results are high-FSC sequences representing massive deviations from the ground state, deviations from the normal OSC-RSC of natural processes. These examples of FSC, therefore, represent something that is unique in nature. In other words, FSC breaks out of the XY plane in Figure 1 and into the Z dimension, deviating from natural processes of chance and necessity that produce OSC and RSC.

8. The effect of mutations on FSC

The set of all sequences $M(E_x)$ that can perform a given function forms the functional sequence space within the larger sequence space. For example, Table 1 shows that the universal protein RecA has an estimated number of 10^{122} functional sequences within the larger 320-site sequence space, assuming site independence. Site interdependency will substantially reduce this number, as discussed earlier, but site independence will be assumed here for the sake of simplicity. These 10^{122} RecA sequences form the functional sequence space for RecA. Mutations that merely move the sequence from one area of functional sequence space to another area of functional space will have little effect on the FSC of the sequence. As discussed in Section 5, FSC is not measured on the basis of a single sequence, but on a large set of functional sequences. Some sequences, however, may be more efficient than others, as pointed out by Hazen *et al.* [2]. If that is the case, then there can be some variation in the value ζ for the FSC of a protein family. This is represented in Figure 1 as a range of Z-values in the FSC of the hypothetical protein family shown in the model. Therefore, mutations within functional sequence space may reduce FSC as they become functionally less efficient until the mutations reduce the

functionality of the sequence below the threshold required by biological life [2]. At that point, any further change can be measured using Eqn. (7) relative to the original function.

With regard to a protein with function f_a evolving into a completely novel structure with novel function f_b, once a sequence has mutated out of f_a sequence space, the evolutionary path may have to traverse a region of non-folding, non-functional sequence space. As pointed out by Blanco, natural selection is of no use in navigating non-functional sequence space, so further mutations will be unguided and take the form of a random walk [26]. Given the results shown in Table 1, and given experimental evidence [27], the area of functional sequence space for stable, folding functional proteins may be so miniscule, that attempting to locate them without the aid of controls may exceed an objective universal plausibility cut-off [28]. Mutations that occur within functional sequence space can move a functional sequence to an area within functional sequence space that may enhance fitness. However, given the miniscule size of functional sequence space, as suggested by Table 1, an obvious prediction is that an accumulation of mutations will tend to be harmful.

Since natural constraints tend to produce repeatable results, non-random mutations imposed by natural constraints will tend to move the sequence in the direction of order, or OSC. As shown in Figure 1, this constitutes movement in the XY plane which can move the sequence outside the area of functionality shown in Figure 1b. Thus, as can be seen in Figure 1b, both a limited amount of random and non-random mutations can be permitted provided the sequence remains within the functional area. But both random and non-random mutations can render the sequence non-functional, collapsing the Z component

to zero, if the mutations move the sequences outside of the small functional sphere representing a hypothetical protein family.

9. Conclusion

FSC can be measured by extending Shannon uncertainty to include the joint variables of data and function. This measure can provide an estimate of the variability and hence the size of the functional sequence space for a specific functional protein. It also can measure change in a sequence due to mutation relative to the required functionality. The information calculated from an observed sequence ensemble constrained by the specified functionality then reflects the underlying sub-molecular information structure that could be used to reconstruct the structural or functional properties of the molecule. FSC thus provides a foundational measure that can form the basis for more detailed analysis.

References

1. Szostak, J.W. 2003, Functional information: Molecular messages, Nature, 423, (6941) 689.
2. Hazen, R.M.; Griffin, P.L.; Carothers, J.M.; Szostak, J.W. 2007, Functional information and the emergence of biocomplexity, Proc Natl Acad Sci U S A, 104 Suppl 1, 8574-81.
3. Abel, D.L.; Trevors, J.T. 2005, Three subsets of sequence complexity and their relevance to biopolymeric information, Theor Biol Med Model, 2, 29.
4. Ferris, J.P. 2002, Montmorillonite catalysis of 30-50 mer oligonucleotides: laboratory demonstration of potential steps in the origin of the RNA world, Orig Life Evol Biosph, 32, (4) 311-32.
5. Abel, D.L.; Trevors, J.T. 2006, Self-Organization vs. Self-Ordering events in life-origin models., Physics of Life Reviews, 3, 211-228.
6. Gammerman, A.; Vovk, V. 1999, Kolmogorov complextiy: sources, theory and applications, The Computer Journal, 42, 252-255.
7. Shannon, C. 1948, Part I and II: A mathematical theory of communication, The Bell System Technical Journal, XXVII, 379-423.
8. Karp, P. 2000, An ontology for biological function based on molecular interactions, Bioinformatics Ontology, 16, (3) 269-285.
9. Costanzo, G.; Pino, S.; Ciciriello, F.; Di Mauro, E. 2009, Generation of long RNA chains in water, J Biol Chem, 284, (48) 33206-16.

10. Durston, K.K. 2010, Statistical analyses of site variability and site inter-dependencies in sub-molecular hierarchical protein structuring. University of Guelph, Guelph.
11. Durston, K.K.; Chiu, D.K.Y. 2005, A functional entropy model for biological sequences, Dynamics of Continuous, Discrete & Impulsive Systems, Series B.
12. Finn, R.D.; Tate, J.; Mistry, J.; Coggill, P.C.; Sammut, S.J.; Hotz, H.R.; Ceric, G.; Forslund, K.; Eddy, S.R.; Sonnhammer, E.L.; Bateman, A. 2008, The Pfam protein families database, Nucleic Acids Res, 36, (Database issue) D281-8.
13. Weiss, O.; Jimenez-Montano, M.A.; Herzel, H. 2000, Information content of protein sequences, J Theor Biol, 206, (3) 379-86.
14. Schneider, T.D. 2006, Claude Shannon: biologist. The founder of information theory used biology to formulate the channel capacity, IEEE Eng Med Biol Mag, 25, (1) 30-3.
15. Wong, A.K.; Liu, T.S.; Wang, C.C. 1976, Statistical analysis of residue variability in cytochrome c, J Mol Biol, 102, (2) 287-95.
16. Lui, T.W.H.; Chiu, D.K.Y. 2009, Multi-value association patterns and data mining. In Foundations of Compuational Intelligence, Abraham, A.; Hassanien, A. E.; de Carvalho, A. P.Snael, V., Eds. Springer-Verlag: Vol. 6: Data Mining.
17. Lui, T.W.H.; Chiu, D.K.Y. 2010, Associative classification using patterns from nested granules, International Journal of Granular Computing, Rough Sets and Intelligent Systems, 1, (4) 393-406.
18. Wong, A.K.C.; Chiu, D.K.Y.; Huang, W. 2001, A discrete-valued clustering algorithm with applications to biomolecular data, Information Sciences, 139, 97-112.
19. Au, W.H.; Chan, K.C.; Wong, A.K.; Wang, Y. 2005, Attribute clustering for grouping, selection, and classification of gene expression data, IEEE/ACM Trans Comput Biol Bioinform, 2, (2) 83-101.
20. Chiu, D.K.; Wang, Y. 2006, Multipattern consensus regions in multiple aligned protein sequences and their segmentation, EURASIP J Bioinform Syst Biol, 35809.
21. Chiu, D.K.Y.; Lui, T.W.H. 2002, Integrated use of multiple interdependent patterns for biomolecular sequence analysis, International Journal of Fuzzy Systems, 4, (3) 766-775.
22. Durston, K.K.; Chiu, D.K.; Abel, D.L.; Trevors, J.T. 2007, Measuring the functional sequence complexity of proteins, Theor Biol Med Model, 4, 47.
23. Abel, D.L. 2010, Constraints vs Controls, The Open Cybernetics & Systemics Journal, 4, 14-17.
24. Pearle, J. 2000, Causation. Cambridge University Press: Cambridge.
25. Trevors, J.T.; Abel, D.L. 2004, Chance and necessity do not explain the origin of life, Cell Biol Int, 28, (11) 729-39.
26. Blanco, F.J.; Angrand, I.; Serrano, L. 1999, Exploring the conformational properties of the sequence space between two proteins with different folds: an experimental study, J Mol Biol, 285, (2) 741-53.
27. Axe, D.D. 2004, Estimating the prevalence of protein sequences adopting functional enzyme folds, J Mol Biol, 341, (5) 1295-315.
28. Abel, D.L. 2009, The Universal Plausibility Metric (UPM) & Principle (UPP), Theor Biol Med Model, 6, 27.