ALIGNMENT-FREE PHYLOGENETIC OUTLINE OF A RANDOM-SEQUENCE LIBRARY OF NON-BIOLOGICAL PROTEINS

Miguel A. Jiménez-Montaño[1] and Matthew He[2]

[1]Facultad de Física e Inteligencia Artificial
Universidad Veracruzana, Xalapa, 91000 Veracruz, México
Email: ajimenez@uv.mx

[2]Division of Math, Science, and Technology
Nova Southeastern University, Ft. Lauderdale, USA
Email: hem@nova.edu

"It seems as though biologists are extraordinarily fond of randomness. A population is defined as one, randomly mating, interbreeding unit, although truly random mating would hardly be practicable in a reasonably large population. Similarly, spontaneous mutations are viewed as randomly sustained base substitutions, in spite of our knowledge of mutational hot spots. I suspect that this extraordinarily strong belief in randomness stems from our too strong faith in the power of natural selection."

S. Ohno, [24]

Abstract - To asses the degree of randomness and complexity of randomly generated sequences, in an *in vitro* selection experiment by Keefe and Szostack [1], we calculated the Kolmogorov complexity, the algorithmic redundancy, and the Shannon entropy of the sequences. We built an alignment-free phylogenetic tree, employing the algorithmic information distance between each pair of sequences to construct the distance-matrix. The tree represents the history of the set of molecular sequences, and allows us to follow in more detail how chemical function improves with respect to the original sequence. We remark the fact that in directed evolution, the highly predominant changes are between neighboring codons. Thus, the amino acid changes in the protein are not arbitrary, but dictated by the amino acid assignments in the code.

Keywords: Kolmogorov complexity, Shannon entropy of the sequences, algorithmic redundancy, phylogenetic tree, non-biological proteins.

1. Introduction

The frequency of occurrence of functional proteins in collections of randomly generated sequences is an important constraint on models of the evolution of biological proteins. Therefore, the experimental determination of this frequency, by isolating proteins with a specific function from a large random-sequence library of known size, is a relevant endeavor in this field. In an effort to substantiate the hypothesis that primordial functional proteins originated from random sequences, Keefe and Szostak [1] used *in vitro* selection of messenger RNA displayed proteins to sample a large population of distinct randomly generated sequences.

Starting from a library of 6×10^{12} polypeptides, each containing 81 contiguous *randomly chosen* amino acids, they selected functional proteins by enriching for those that bind to ATP. As a result, following eight rounds of selection, they obtained four new ATP-binding protein families, designated A, B, C, D (Fig. 3a of their paper), that appear to be unrelated to each other or to anything found in

the current databases of biological proteins. One of these proteins (Family B) was optimized by directed evolution for improved binding affinity. DNA sequencing of the output from this selection revealed a distant clone (clone 18-19) that differed from the consensus sequence at 15 out of 80 positions, and bound ATP with far greater affinity and specificity than all other clones from that round of selection. From this experiment, Keefe and Szostak [1] estimate that roughly 1 in 10^{11} of all random-sequence proteins have ATP-binding activity.

The X-ray crystal structure of the nucleotide binding domain for protein 18-19 was originally solved by Lo Surdo et al. [2] and found to adopt a novel zinc-nucleated a/b-fold not yet observed by nature. As described in detail in [3], the structural comparison of protein 18-19 with the databank of biological protein folds revealed that the *de novo* evolved protein shared certain structural features with some proteins found in nature. However, unlike many naturally occurring proteins, protein 18-19 requires high concentrations of free ligand in order to remain stably folded and soluble.

In two recent publications, Szostak´s group examined the extent to which a *de novo* evolved protein, originally selected on the basis of ligand binding affinity, could be evolved to remain stably folded in the absence of exogenous ligand [3].These authors designed an *in vitro* selection experiment using mRNA display to isolate variants of protein 18- 19 that remained bound to an ATP agarose affinity resin in the presence of increasing concentrations of chemical denaturant. In the second publication [4], they used structural and functional studies to investigate the *in vitro* evolutionary processes in greater detail. We refer the reader to the original papers for further details.

Since proteins acquire functionality (meaning) throughout evolution, to complement the mentioned works, we consider the construction of a phylogenic tree (Fig. 1) for the evolved proteins in the earliest experiment [1]. The tree represents the history of the set of molecular sequences, and allows us to follow in more detail how chemical function improves with respect to the original sequence. It is commonly believed that to infer such a tree one must first arrange the sequences relative to each other in a way that presents the best available hypothesis of homology at each and every position in those molecules; i.e., an optimal multiple sequence alignment (MSA). There are nonetheless alternative approaches to molecular phylogenetic inference that do not involve prior MSA (reviewed in [5]). These involve two steps: the calculation of a matrix of pairwise distances among unaligned molecular sequences, followed by generation of a tree using a distance-based method such as neighbor joining [6]. The fundamental difference from alignment-based methods lies mainly in the first step; i.e., how pair wise distances in the underlying distance matrix are defined. The majority of alignment-independent approaches involve information theory and the Kullback-Liebler discrepancy or relative entropy; they are based on the statistical properties of of *n*-grams. Or in compression methods, employing the algorithmic information (also called Kolmogorov complexity) shared by two sequences (see Discussion). A notable example of this last approach is the paper by Li et al. [7], who employed the *algorithmic information distance* between a pair of sequences [8,9], to construct a distance-matrix for building a whole mitochondria genome phylogeny without first aligning the sequences. Our approach is closely related to theirs, differing mainly in the software employed to estimate the algorithmic distance.

The simplest way to describe our methodology is in the context of the following linguistically motivated question: Is it possible to identify the subject treated in a text in a way that permits its automatic classification among many other texts in a given corpus? As shown by Benedetto et al. [10] among others, the answer is positive. For DNA sequences, a solution to this kind of problem was delineated by Loewenstern et al. [11] as follows:

"If we took a corpus of DNA sequences, we could gain insight into the degree of similarity between a test sequence and the corpus by compressing the corpus with the test sequence appended, and subtracting the size of this compressed file from the size of the compressed corpus alone. We could classify a test sequence by following the above procedure with two different sample populations of text, assigning the test sequence to the label of the population with which it compressed best"

Here, we follow this idea to classify pseudorandom amino acid sequences.

2. Materials and Methods

Alignment-free Sequence Comparison Algorithms:

In a former publication [12] we introduced the WinGramm Suite [13]. It consists of a set of programs aimed to calculate informational and algorithmic quantities, such as n-gram entropies, context-free grammatical complexity, and algorithmic distance, as well as surrogate statistics, in order to reveal the information content, the complexity or the redundancy embodied in symbol sequences [14, 15, 16, 17, 18].

Here, we have employed the WinGramm Suite to obtain the phylogenetic classification of non-biological amino acid sequences. For this end, we applied our programs to:

1) Calculate the context-free grammatical complexity, algorithmic distance and redundancy, Shannon entropy and surrogate statistics of the protein sequences.

2) Build a phylogenetic tree to classify these sequences, taken from different clones in the directed evolution experiment.

3. Results

Classification of Pseudorandom Proteins:

Globular proteins have amino acid sequences which are highly complex, indistinguishable from pseudorandom sequences [19]. In that paper the authors estimated the Shannon entropy and applied two compression algorithms (one of them is included in the WinGrammm Suite) to estimate the algorithmic complexity of a large, non-redundant, set of protein sequences finding that proteins are fairly close to pseudorandom sequences. They found an entropy reduction due to correlations of about 1 %, corroborated with compression algorithms, which indicates that proteins have approximately 99 % of the complexity of random polypeptides with the same amino acid composition. These results give support to the conclusion of Pande et al. [20], White and Jacobs [21], and others that protein sequences are "slightly edited random sequences".

To set up our problem, we consider a sample of 17 sequences from the set generated by Keefe and Szostack [1], in their original *in vitro* selection experiment (appearing in the supplementary information file of the paper). All of the sequences have the following structure:

MDYKDDDDKKT
$(\text{Random})_{81}$WSASCHHHHHHMGMSGS.

From each of these sequences, we dropped the short invariant segments encoding affinity tags for purification, at the beginning and end, retaining the 81 amino acid random segment. The first 13 sequences were obtained from round 8, belonging to families A, B, and C, which have 4 sequences each. The thirteenth sequence constitutes the single representative of family D. The last 4 sequences were acquired from round 18 (Table 1). With the help of the WinGramm Suite [13] we calculated the algorithmic distance between each sequence pair, and obtained the distance matrix (supplementary information Table 2). From this matrix we built the phylogenetic tree (Fig. 1). Comparing this tree with the information in Fig. 3a of [1], we noticed a mistake in their figure: Family A should read Family C and *vice versa*. Professor Szostak acknowledged the misprint (personal communication). The tree displays the right assignment of sequences to families and, correctly, allocates the sequences of generation 18^{th} with family B (see above).

To asses the degree of randomness and the complexity of the experimentally evolved sequences, we calculated the grammatical complexity, the corresponding S-measure (also called *Z*-score), the algorithmic redundancy and the Shannon entropy of the random segments (Table 1). From the S-measure of the complexity, S(K), defined by the difference between the original value of K and its mean surrogate value, divided by the SD of the standard surrogate values:

$$S = \frac{\left| K_{orig} - \langle K_{surr} \rangle \right|}{\sigma_{surr}},$$

it is clear that the evolved sequences are as random as their surrogates. S_{aver} = 1.6191 SD. For the families with more than one member (A,

B and C), we concatenated the strings in each group and compared the resulting string with a sequence, of the same length, constructed from concatenated random surrogates. For example, for Family A, we constructed the sequence F_A concatenating the strings in the family: (08-05), (08-07), (08-09), (08-48) (Table 1). We compared the grammatical complexity of F_A , $K(F_A)$, with the complexity of the string S_A , $K(S_A)$, which was constructed from the concatenation of standard-random surrogates of each sequence in the family. Although, both F_A and S_A were built from pseudorandom sequences, the complexity of F_A is much lower than the complexity of S_A because the sub words of F_A are very similar among themselves, and the sub words of S_A are independent pseudorandom sequences. Thus, the complexity of F_A is a good deal lower than the average complexity of its surrogates (Table 1). The sequence F_A can be considered to be the "corpus" of family A. Thus, an unknown sequence may be identified as belonging or not to family A, after compressing it with this "corpus". While the average algorithmic redundancy of the 17 sequences is very low, 1.4 %, the same quantity of the concatenated sequences is high: 42.2 %, 45.6 %, and 44. 4 % for F_A, F_B, and F_C , respectively (Table 1). However, the average Shannon entropy (H_{aver}) of the evolved sequences and of the concatenated sequences is almost the same (Table 1). H_{aver} differs from its maximum value, H_{max}, only in 0.18398 bits. This is due to the fact that, contrary to the algorithmic quantities, H depends only on the composition of the sequence, except for finite size effects [22, 23], and not on the order of the symbols.

As we mentioned above, the experiment shows that starting from random amino acid sequences, after a few rounds of Darwinian evolution *in vitro*, it is possible to select a functional protein. Nonetheless, the final protein which carries a biochemical function (in a suitable environment), not only *looks as random as the starting polypeptide* without function, from which it was generated, but has informational parameters that confirm this fact (Table 1).

.

4. Discussion and Conclusions

Biological sequences encode information, and the occurrence of evolutionary events separating two sequences sharing a common ancestor will result in the loss of the shared information. Sequences which do not share common ancestor will not share more information than would be expected at random. Therefore, we consider that the appropriate distance matrix was the one defined by be the *algorithmic information distance* between a pair of sequences. Because this distance is based on Kolmogorov complexity (estimated by the grammar complexity), that was designed to measure the information content of individual objects. Here, we made a new application of this concept, since concatenating the sequences of a family we measure the information content of the family. Then we compute the shared information between the new sequence and the family.

The further optimization of sequence 18-19 described in [4] consist of twelve single-base mutations, seven of which are transitions. Therefore, the increased stabilization and solubility of the protein is highly influenced by the structure of the genetic code. In the vicinity of a functional protein, in protein space, it is not very difficult to get improvements by fine-tuning it. This is so because, although DNA base mutations are random, each codon does not have the same probability to mutate to any of the other 61 sense codons. In short-term natural evolution and in directed evolution, the highly predominant changes are between neighboring codons. Thus, the amino acid changes in the protein are not arbitrary, but dictated by the amino acid assignments in the code.

Sequence	K	<K>sd-surr	$<K>_{pair\text{-}Surr}$	S (K)	R %	H bits
A8-05	81	79.9	81	0.9047	0.74600	4.13920994
A8-07	81	80.1	80.5	1.2853	1.12359	4.14729973
A8-09	81	79.6	80.7	0.7905	1.25000	4.13801925
A8-48	81	79.9	80	0.8162	0.99751	4.11634521
B8-01	78	80	81	2.5276	2.62172	4.1396914
B8-04	78	80.7	80	3.9185	3.22580	4.15270775
B8-08	78	80.3	80	3.5941	2.86426	4.13992263
B8-10	79	80.4	81	2.0314	1.61893	4.14920872
C8-06	80	80	81	0.6031	0.49751	4.13510411
C8-11	80	80.3	81	0.0000	0.0000	4.12701735
C8-17	80	80.6	81	0.3331	0.24937	4.13212155
C8 -19	81	80.6	80.5	0.8160	0.49627	4.12905168
D8-20	81	80.6	79	1.0938	0.87172	4.12407239
18-01	78	80.3	81	3.6191	2.98507	4.14924243
18-02	81	80.7	80	0.9047	0.74626	4.17014976
18-03	78	80.4	79	2.2447	2.74313	4.12228268
18-19	79	80.5	81	2.0417	1.37328	4.13367787
Average	79.70	80.28	80.45	1.6191	1.43591	4.13794837
SQR	1.2727	0.3141	0.6643	1.1743	1.01817	0.01276136
$\mathbf{F_A}$	171	296.1	299	31.38664	42.2490	4.13989713
$\mathbf{F_B}$	165	299.5	301.7	28.4291	45.67007	4.1497239
$\mathbf{F_C}$	168	300.8	301	37.8086	44.426	4.13400159

Table 1 Grammatical Complexity, S-values, algorithmic redundancy and entropy for pseudorandom protein sequences[a]

[a] The labels of the first 17 sequences are the same as in the additional information from Keefe and Szostack [1]. The last three sequences were obtained by concatenating the sequences of the corresponding family, as explained in the text.

In the first column, K is the grammar complexity; the 2nd and 3rd columns are average values of K, for standard and pair-conserving surrogates [12, 13]. S (K) is the S-measure of K, R is the algorithmic redundancy in % and H is the entropy in bits.

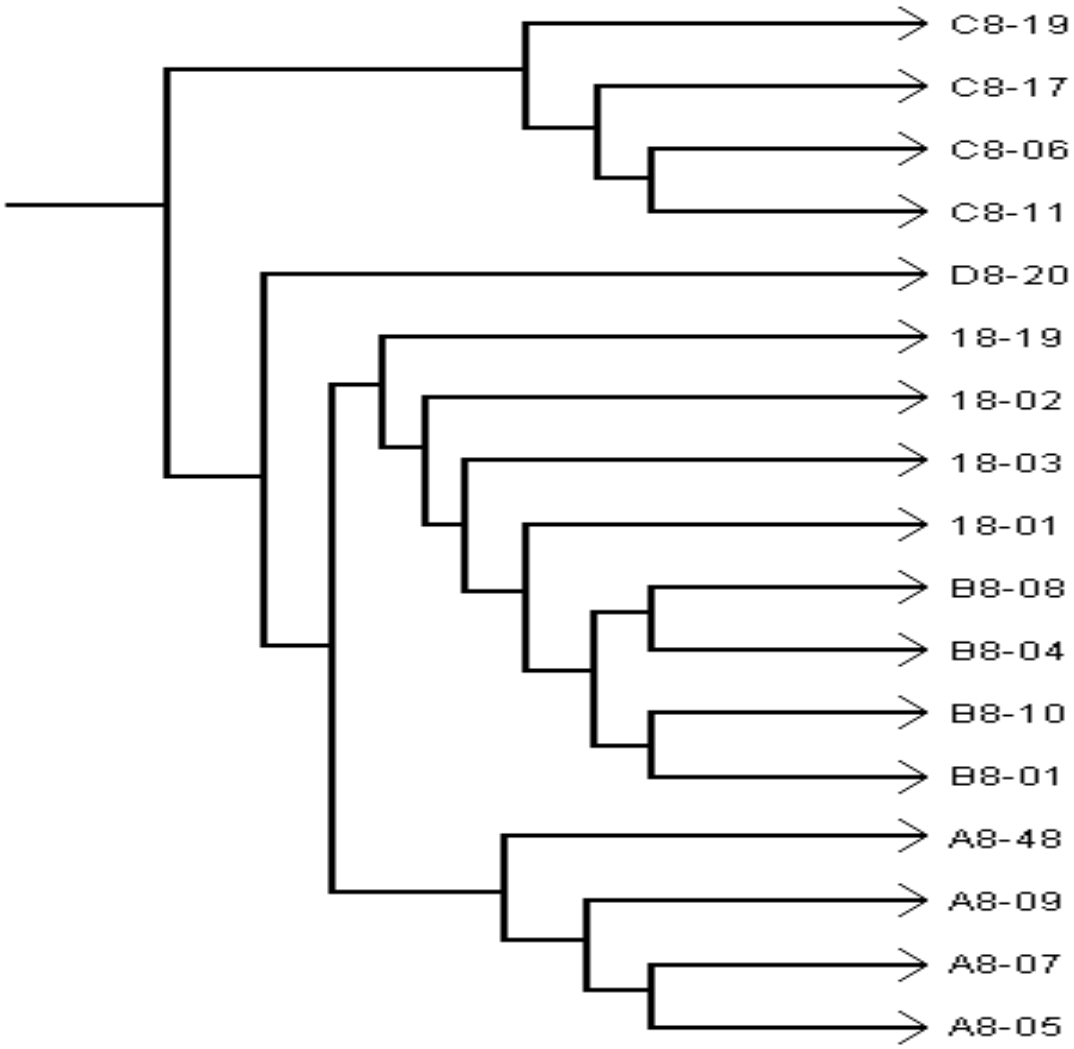

Fig. 1 Phylogenetic tree for the non-biological protein sequences from the experiment performed by Keefe and Szostak (2001).

Acknowledgements

The first author of this paper would like to thank CONACYT, MEXICO Project: 81484; Sistema Nacional de Investigadores; and PROMEP, Project: UV-CA-197, for partial support.

References

1. Keefe, A.D., Szostak, J.W.: Functional proteins from a random-sequence library. Nature 410, 715–718 (2001)
2. Lo Surdo, P., Walsh, M.A., Sollazzo, M.: A novel ADP-and zinc-binding fold from function-directed in vitro evolution. Nat Struct Molec Biol 11, 382–383 (2004)
3. Smith, M.D., Rosenow, M.A., Wang, M., Allen, J.P., Szostak, J.W., Chaput, J.C.: Structural insights into the evolution of a non-biological protein: Importance of surface residues in protein fold optimization. PLoS ONE 2, e467. doi:10.1371/journal.pone.0000467. (2007)
4. Mansy, S.S., Zhang, J.L., Kummerle, R., Nilsson, M., Chou, J.J., Szostak, J.W.,

Chaput, J.C.: Structure and evolutionary analysis of a non-biological ATP-binding protein. J. Mol. Biol. 371, 501–513 (2007)
5. Vinga, S., Almeida, J.: Alignment-free sequence comparison-a review. Bioinformatics 19, 513-523 (2003)
6. Saitou, N., Nei, M.: The neighbor-joining method: A new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4, 406-425 (1987)
7. Li, M., Badger, J.H., Chen, X., Kwong, S., Kearney, P., Zhang, H.: An information-based sequence distance and its application to whole mitochondrial genome phylogeny. Bioinformatics 17 (2), 149-154 (2001)
8. Zurek, H.: Thermodynamic cost of computation, algorithmic complexity and the information metric. Nature 341, 119-124 (1989)
9. Li, M., Vitányi, P.: *An introduction to Kolmogorov complexity and its applications.* Berlin: Springer. (1997)
10. Benedetto, D., Caglioti, E., Loreto, V.: Language Trees and Zipping. Physical Review Letters 88 (4), 048702 (2002)
11. Lowenstern, D., Hirsh, H., Yianilos, P., Noordewier, M.: DNA sequence classification using compression-based induction. DIMACS Technical Report 95-04, 1-12 (1995)
12. Jiménez-Montaño, M.A., Feistel, R., Diez-Martínez, O.: Information Hidden in Signals and Macromolecules I. Symbolic Time-series Analysis. Nonlinear Dynamics, Psychology & Life Sciences 8 (4), 445-478 (2004)
13. Jiménez-Montaño, M. A., Feistel, R.: WinGramm: *Grammatical Complexity Analysis of Sequences.* Internal Report. Faculty of Physics & Artificial Intelligence, University of Veracruz, Mexico (2003). The suite of programs and user manual may be downloaded from: http://www.io-warnemuende.de =/homepages/rfeistel/
14. Gatlin, L.L.: *Information Theory and the Living System*. New York: Columbia University Press (1972)
15. Ebeling, W., Jiménez-Montaño, M.A.: On grammars, complexity, and information measures of biological macromolecules. Mathematical Biosciences 52, 53-71 (1980)
16. Ebeling, W., Feistel, R.: *Physics of self-organization and evolution* (in German). Berlin: Akademie-Verlag (1982)
17. Jiménez-Montaño, M.A.: On the syntactic structure of protein sequences and the concept of grammar complexity. Bull. Math. Biol. 46(4), 641–659 (1984)
18. Milosavljevic, A.: Discovering patterns in DNA sequences by the algorithmic significance method. In J.T.L. Wang, B.A. Shapiro, & D. Shasha (Eds.), *Pattern discovery in biomolecular data* (pp. 3-23). Oxford: Oxford University Press (1999)
19. Weiss, O., Jiménez-Montaño, M.A., Herzel, H.: Information content of protein sequences. J. Theor. Biol. 206, 379-386 (2000)
20. Pande, S.V., Grosberg, A.Y., Tanaka, T.: Nonrandomness in protein sequences: Evidence for a physically driven stage of evolution? Proc. Natl. Acad. Sci. U.S.A. 91, 12972-12975 (1994)
21. White, S.H., Jacobs, R.E.: The evolution of proteins from random amino acid sequences. I. Evidence from the lengthwise distribution of amino acids in modern protein sequences. J. Mol. Evol. 36, 79-95 (1993)
22. Schmitt, O., Herzel, H., Ebeling, W.: A New Method to Calculate Higher-Order Entropies from Finite Samples. Europhysics Letters 23, 303 (1993)
23. Herzel, H.: Complexity of symbol sequences. Systems Analysis Modelling Simulation 5, 435-444 (1988)
24. Ohno, S.: Modern Coding Sequences Are in the Periodic-to-Chaotic Transition. Hämatol. Bluttransf. Vol 32 pp 512-519 (1989)