Hindawi Publishing Corporation
Journal of Theoretical Chemistry
Volume 2013, Article ID 961378, 5 pages
http://dx.doi.org/10.1155/2013/961378

Research Article

Studying the Polypeptide Sequence (α-Code) of *Escherichia coli*

Vladimir R. Rosenfeld

Mathematical Chemistry Group, Noble Institution for Environmental Peace, P.O. Box 163, 720 Spadina Avenue, Toronto, ON, Canada M5S 2T9

Correspondence should be addressed to Vladimir R. Rosenfeld; vladimir_rosenfeld@yahoo.com

Received 19 March 2013; Accepted 21 November 2013

Academic Editors: A. M. Lamsabhi, A. Stavrakoudis, and B. M. Wong

This paper is devoted to algebraically simulating the α-code of bacterium *Escherichia coli* and studying contrast factors (words) in its polypeptide sequence. We utilize the methods of spectral theory of graphs which were previously employed by us for enumerating De Bruijn and Kautz sequences. The empirical material is borrowed from the computer investigation of contrast factors in the polypeptide sequences of prokaryotes.

1. Introduction

It was proposed [1, 2] to divide 19 out of all the 20 amino acids into two subgroups, an *alanine* subgroup (of fatty amino acids: alanine, phenylamine, isoleucine, leucine, methionine, proline, threonine, and valine) and a *glycine* subgroup (of more polar amino acids: cysteine, aspartic acid, glutamic acid, glycine, histidine, lysine, asparagine, glutamine, arginine, tryptophan, and tyrosine), while serine remains a spare element in the full classification thereof. In a shorthand notation, this gives **a**, **f**, **i**, **l**, **m**, **p**, **t**, and **v** (an alanine subgroup); **c**, **d**, **e**, **g**, **h**, **k**, **n**, **q**, **r**, **w**, and **y** (a glycine subgroup); and **s** (a free character). The three numbers 1, 2, and 3 were picked to represent the main two subgroups and the character **s**, respectively.

Brute statistics, under the natural ratio of 1s to 2s being 0.526: 0.474, had predicted an almost regular distribution (alternation) of the two ciphers. To check such a hypothesis, there were found the frequencies of all 2^l ($1 \le l \le 11$) possible substrings of the length l in the genomic sequence of *E. coli*. The results at once showed that the respective perfectly alternating substrings are in fact the least frequent ones, in the entire genomic sequence. However, visual observation allowed suspecting that the main condition for near-to-statistical distribution of the two subgroups of amino acids may be disguised in grouping equal ciphers (either 1 or 2) representing the respective subgroups in adjacent pairs thereof. That is, in lieu of the code $121212\cdots12$ it should be $112211\cdots22$, where general ratio of ciphers stays thus unaltered.

The situation with pairing of equal ciphers reminded us of the phenomenon of the so-called α-code in polypeptides. According to it, one turn of polypeptide spiral involves 3.5 amino acids on an average. Since the nearest to 3.5 multiple integer equals 7, it was of interest to interpret the α-code as a one in which all structural features are due to conditionally grouping amino acids into consecutive sevens thereof.

Merging the ideas of sevens (which suit well for interpreting the α-code) and pairs (which better obey the natural proportion of amino acids and follow experimental observation) allows us to set forward a universal model of the α-code. In this, the distribution of amino acids along the entire polypeptide sequence should maximize a mean number of pairs of equal integers (1 or 2) that are contained in consecutive sevens of it. Thus, the optimal case is that one has 3 pairs (11 and/or 22) and one unpaired number (1 or 2) which can either be between any two pairs out of the three or be outside of these (e.g., it may comprise a pair with an equal number of the adjacent seven). However, it is more interesting to consider the case when the window of length 7 cuts out a substring without 3 such pairs; this is also a necessary pattern in the sequence.

Logically, we could readily establish that the optimal α-code should have *not less than* 2 pairs of adjacent equal numbers in every window of length 7. Another criterion that seemed to be of use was that the α-code should reject the subsequences 1212 and 2121 since these allow no more than

one pair 11 or 22 in most seven-cipher substrings that contain them. Experimentally, there was the third (and strongest of all we know) criterion; the α-code should avoid the inclusion of the palindrome of the type 1211121; however, here, very serious reservations must be made concerning the fact that substituting 2 for 1 (and conversely) in this palindrome produces (longer) substrings that are, on the contrary, not very frequent in the natural sequence of *E. coli*. In case of longer palindromes including 1211121 (or 2122212), the situation may seem rather ambiguous.

That is why we first tried to attest the simplest (for implementation) mathematical criterion of the above; the prohibition of 1212 and 2121, in a model α-code. Here, we shall turn to some rigorous mathematical matters.

2. Preliminaries

First of all, we must introduce an ancillary digraph D' which is constructed as follows. The vertices of D' are all the eight ordered triples of 1s and 2s (i.e., 111, 112, 121, 211, 122, 202, 220, and 222) and an arc (a self-loop) goes out of one vertex to another (this same vertex) if the last two ciphers (on the right) in the first triple coincide with the first two ciphers in the second triple. Say there is an arc from 111 to 112 (with two common ciphers: 11) and also an oriented selfloop attached to the vertex 111 (i.e., an arc from 111 into 111 itself).

Using the connectivity of D', one can mentally travel in it from any vertex v, consecutively traversing arcs and visiting adjacent vertices. So, the passage to any adjacent vertex u along the respective arc (or returning to the original vertex v through the selfloop attached to it, if any) means that one has done a walk of length 1 (*et seq.*). However, what is very essential is that every walk of the length l, in D', factually covers $l+3$ ciphers since every next step increases the number of visited vertices by one but the very first vertex already contained three ciphers, by definition. It is very important to note that the auxiliary digraph D' allows one to reproduce every sequence of 1s and 2s of length $l + 3$ ($l \geq 0$), which also includes sequences containing 1212 and 2121.

In order to rule out all forbidden sequences with 1212 and/or 2121, we shall derive from D' a "better" working digraph D. Specifically, D is the above digraph D' less two opposite arcs going from 121 to 212, and vice versa, from 212 to 121. Since now, the problem on enumerating all (model) α-code sequences of length $l + 3$ ($l \geq 3$) can be reduced to another one enumerating all walks of length l in the working digraph D.

It is not difficult to form the adjacency matrix $A = A(D) = [a_{ij}]_{i,j=1}^{8}$ of D, where an entry $a_{ij} = 1$ if and only if there is an arc (a selfloop) from the ith vertex to the jth vertex and $a_{ij} = 0$, otherwise, namely,

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix}. \tag{1}$$

The characteristic polynomial $P(D; x)$ of the digraph D (which is by definition the characteristic polynomial $P(A; x)$ of its adjacency matrix A [3]) is

$$\begin{aligned} P(D; x) &= x^8 - 2x^7 + x^6 - 2x^5 + x^4 + x^2 \\ &= x^2(x-1)\left(x^2+1\right)\left(x^3 - x^2 - x - 1\right). \end{aligned} \tag{2}$$

The roots of $P(D; x)$ are $\lambda_1 = 1.83929$, $\lambda_2 = 1$, $\lambda_{3,4} = 0$, $\lambda_{5,6} = \pm i$, and $\lambda_{7,8} = 0.419645 \pm 0.6063i$, where $i = \sqrt{-1}$. These roots comprise the spectrum (of eigenvalues) of the digraph D [3].

Also, we need to construct a derivative matrix $\overline{A} = J - I - A$, where J is the matrix of all 1s while I is a diagonal identity matrix; namely,

$$\overline{A} = \begin{bmatrix} -1 & 1 & 1 & 1 & 0 & 1 & 1 & 1 \\ 1 & -1 & 1 & 1 & 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 \end{bmatrix}. \tag{3}$$

The matrix $\overline{A}$ is the matrix of the complementary digraph $\overline{D}$. The corresponding characteristic polynomial is

$$\begin{aligned} P\left(\overline{D}; x\right) &= x^8 + 2x^7 - 15x^6 - 80x^5 - 180x^4 \\ &\quad - 232x^3 - 180x^2 - 80x - 16. \end{aligned} \tag{4}$$

Let $H_D(t) = \sum_{l=0}^{\infty} N_l t^l$ be the generating function of the number of walks in the digraph D, wherein the coefficient N_l of t^l is the number of walks of length l, in D. Taking into account that a walk of length l ($l \geq 0$) corresponds to a substring of length $l+3$ of the α-code, one can also write down the generating function $S(t)$ for the number of substrings (of the α-code) of length l as follows:

$$S(t) = \sum_{l=0}^{\infty} N_l t^{l+3} = t^3 H_D(t). \tag{5}$$

In the next subsection, we shall demonstrate that analytically calculating $S(t)$ can readily be done with the aid of spectral theory of graphs.

3. Main Part

We shall start this subsection with adapting adapted in our notation the following fundamental result (see Theorem 1.11 in [3]).

Lemma 1. *Let D be a finite connected (di)graph. Then the generating function $H_D(t)$ of the number of walks in D is*

$$H_D(t) = \frac{1}{t}\left\{(-1)^n \frac{P_{\overline{D}}[-(t+1)/t]}{P_D(1/t)} - 1\right\}, \tag{6}$$

where n is the number of vertices in D.

For practically using (6), one should substitute to the R.H.S. of it the polynomials $P(D;x)$ and $P(\overline{D};x)$ obtained above. After elementary but tedious manipulations (which will be omitted herein) we have arrived at the following formula for our specific digraph D:

$$H_D(t) = \sum_{l=0}^{\infty} N_l t^l = -2\left(\frac{2t^4+3t^3+6t^2+3t+4}{t^5+t^4+2t^3+t-1}\right). \quad (7)$$

It should also be noted that along with the exact result given by Theorem 1.11 from [3], there exists a remarkable asymptotic evaluation for N_l, based on Theorem 1.12 in [3]. In our specific case, it gives the following simple formula:

$$\lim_{l\to\infty} N_l = 8 \times 1.83929^l, \quad (8)$$

where 8 is the number of vertices in the digraph D while 1.83929 is its eigenvalue λ_1 with the maximum modulus ($|\lambda_1| > |\lambda_s|$; $2 \le s \le 8$). Thus, there can be deduced the following approximate formula for the entire series $H_D(t)$:

$$H_D(x) \approx \frac{8}{1-1.83929t}. \quad (9)$$

From the obtained spectral results for the number of walks, one can immediately derive the respective working formulae for the number of substrings of length l ($l \ge 3$) of the α-code. So, we arrive at the following final expression:

$$\begin{aligned}
S(t) &= \sum_{l=0}^{\infty} N_l t^{l+3} = -2t^3\left(\frac{2t^4+3t^3+6t^2+3t+4}{t^5+t^4+2t^3+t-1}\right) \\
&\approx \frac{8t^3}{1-1.83929t}, \\
S(t) &= 8t^3 + 14t^4 + 26t^5 + 48t^6 + 88t^7 \\
&\quad + 162t^8 + 298t^9 + 548t^{10} + 1008t^{11} \\
&\quad + 1854t^{12} + 3410t^{13} + 6272t^{14} \\
&\quad + 11536t^{15} + 21218t^{16} + 39026t^{17} \\
&\quad + 71780t^{18} + 132024t^{19} + 242830t^{20} \\
&\quad + 446634t^{21} + 821488t^{22} + 1510952t^{23} \\
&\quad + 2779074t^{24} + 5111514t^{25} + 9401540t^{26} \\
&\quad + 17292128t^{27} + 31805182t^{28} + 58498850t^{29} \\
&\quad + 107596160t^{30} + 197900192t^{31} + 363995202t^{32} \\
&\quad + 669491554t^{33} + 1231386948t^{34} \\
&\quad + 2264873704t^{35} + 4165752206t^{36} \\
&\quad + 7662012858t^{37} + 14092638768t^{38} \\
&\quad + 25920403832t^{39} + 47675055458t^{40} \\
&\quad + 87688098058t^{41} + 161283557348t^{42} \\
&\quad + 296646710864t^{43} + 545618366270t^{44} \\
&\quad + 1003548634482t^{45} + 1845813711616t^{46} \\
&\quad + 3394980712368t^{47} + 6244343058466t^{48} \\
&\quad + 11485137482450t^{49} + 21124461253284t^{50} \\
&\quad + 38853941794200t^{51} + 71463540529934t^{52} + \cdots, \\
S(t) &\approx 8t^3 + 14.714t^4 + 27.064t^5 \\
&\quad + 49.778t^6 + 91.557t^7 + 168.4t^8 \\
&\quad + 309.74t^9 + 569.69t^{10} + 1047.8t^{11} \\
&\quad + 1927t^{12} + 3544.8t^{13} + 6519.9t^{14} \\
&\quad + 11992t^{15} + 22057t^{16} + 40569t^{17} \\
&\quad + 74618t^{18} + 137240t^{19} + 252430t^{20} \\
&\quad + 464290t^{21} + 853970t^{22} + 1570700t^{23} \\
&\quad + 2889000t^{24} + 5313700t^{25} + 9773400t^{26} \\
&\quad + 17976000t^{27} + 33063000t^{28} + 60813000t^{29} \\
&\quad + 111850000t^{30} + 205730000t^{31} + 378400000t^{32} \\
&\quad + 695980000t^{33} + 1280100000t^{34} \\
&\quad + 235450000t^{35} + 4330600000t^{36} \\
&\quad + 7965200000t^{37} + 14650000000t^{38} \\
&\quad + 26946000000t^{39} + 49562000000t^{40} \\
&\quad + 91159000000t^{41} + 167670000000t^{42} \\
&\quad + 308390000000t^{43} + 567220000000t^{44} \\
&\quad + 1043300000000t^{45} + 19189000000t^{46} \\
&\quad + 3529400000000t^{47} + 6491600000000t^{48} \\
&\quad + 11940000000000t^{49} + 21961000000000t^{50} \\
&\quad + 40392000000000t^{51} + 74293000000000t^{52} \\
&\quad + 136650000000000t^{53} + \cdots.
\end{aligned} \quad (10)$$

Beckmann et al. [4] and Brendel et al. [5] studied the measure for evaluating the deviation of the frequency of a string of length n, in a genomic sequence, from its statistically expected magnitude. Here, we shall also adopt their approach [4, 5]. Let $w = a_1 a_2 \cdots a_n$ be a sequence of letters, of length n, encoding amino acids in a genome or random sequence ($a_i \in \mathscr{A}_{20} = \{\mathbf{a}, \mathbf{c}, \mathbf{d}, \mathbf{e}, \mathbf{f}, \mathbf{g}, \mathbf{h}, \mathbf{i}, \mathbf{k}, \mathbf{l}, \mathbf{m}, \mathbf{n}, \mathbf{p}, \mathbf{q}, \mathbf{r}, \mathbf{s}, \mathbf{t}, \mathbf{v}, \mathbf{w}, \mathbf{y}\}$, $1 \le i \le n$). Let further $f(w) = f(a_1 a_2 \cdots a_n)$ denote

Table 1: Contrast 8-letter factors in the polypeptide sequence of *Escherichia coli*.

Number	w	$f(w)$	std(w)	Number	w	$f(w)$	std(w)
1	11111111	10436	4.686098	61	21121112	3489	3.470266
2	11111112	5225	−6.185607	70	21211111	3439	−3.430215
3	21111111	5214	−6.595223	74	21121111	3430	−3.594918
5	12111111	4886	4.464338	78	11211112	3403	−3.546551
6	11112111	4852	3.606632	82	12211222	3378	−3.168103
7	11111121	4835	3.235524	92	21111121	3341	−4.150119
10	11111122	4126	−3.226558	94	22211221	3337	7.776085
13	22111111	4064	−4.725432	110	12111112	3234	−5.338942
16	11221122	3977	5.215973	119	12121111	3193	3.538106
26	21111112	3749	8.531731	132	21211112	3124	4.053076
29	21112111	3728	−3.948552	155	12211212	3018	−3.143041
30	22111112	3726	5.740505	162	21111212	2980	3.490677
33	22112112	3665	−3.097771	165	12112222	2971	−4.237803
34	11112112	3639	−3.956724	177	22122212	2910	3.511413
40	21111122	3618	4.192473	206	22222222	2767	3.071546
41	21112112	3609	4.460223				

the observed frequency of a string w in a genome. Given the observed frequencies of $(n-2)$- and $(n-1)$-mers, one can calculate the expected frequency of n-mers in a genome as follows:

$$\phi(a_1 a_2 \cdots a_n) = \frac{f(a_1 \cdots a_{n-1})\, f(a_2 \cdots a_n)}{f(a_2 \cdots a_{n-1})}. \quad (11)$$

To measure the deviation of the observed frequency of string w from its expected occurrence in a genome sequence, one should further define the standard deviation of w as

$$\mathrm{std}(w) = \frac{f(w) - \phi(w)}{\max\left(\sqrt{\phi(w)}, 1\right)}. \quad (12)$$

A string is called a contrast factor (or word) if the absolute value of std(w), defined as above, is greater or equal to some threshold value (which is set to be 3.0, in [4, 5]). In other words, a contrast word w should obey the claim $|\mathrm{std}(w)| \geq 3$.

Table 1 contains the values std(w) for all 31 8-character contrast factors in the polypeptide sequence of *Escherichia coli*, wherein the numbers are taken from the full inventory of all possible 256 8-character factors over the two-character alphabet $\mathscr{A}_2 = \{1, 2\}$, as these follow in the nonincreasing order of the magnitudes of $f(w)$.

Now we shall consider some conclusions.

4. Conclusions

The following inferences can be drawn.

(1) The distribution the most contrast factors approximately corresponds to that of the most frequent ones. Say the first 16 out of 31 contrast factors fall into the group of the 41 most frequent factors, whereas 50 of the least frequent factors include no contrast factors whatever. The last, 31st, contrast word is 206th in the complete list of all 8-character sequences over $\mathscr{A}_2$.

(2) It is clearly seen that a "nonpolar" alanine subgroup of amino acids, denoted by 1, and a "polar" glycine subgroup, denoted by 2, play asymmetric parts in composing the contrast words. For instance, 1 is the last character of word 12 times while 2 is at the tail 19 times; truly, at the head of word the respective numbers are 14 and 17, whose difference seems to be not so essential. The sums 12 + 14 = 26 and 19 + 17 = 36 also conserve this disbalance of frequencies. Thus, mutually substituting 1s for all 2s, and vice versa, in the words collected in Table 1 is not at all an invariant action on it. Additionally note that out of $8 \times 31 = 248$ characters comprising the 31 contrast words 2 appears only 86 times (or in 34.68% of cases). Separately, the number of occurrences of 2 in Table 1 for the words w with std(w) ≥ 3 is 48 (or 19.35% of cases) and that for the words with std(w) ≤ -3 is 38 (or 15.32% of cases). Since the share of 2 in a natural polypeptide sequence of *Escherichia coli* is 0.474, contrast factors in it can comprise only a minor part of its total length (it is a trivial qualitative fact that is easily deducible from the data in Table 1). But another readily seen fact is of paramount importance. Contrast factors occur chiefly due to the presence of a fattier alanine subgroup of amino acids, denoted by 1.

(3) Since (the most preferable) contrast factors w with std(w) ≥ 3 and (the most avoidable) contrast factors with std(w) ≤ -3 have practically equal frequencies in Table 1 (16 and 15, resp.), both types of contrast factors play a commanding role in forming the polypeptide sequence (in particular of *Escherichia coli*). That is, the synthesis of polypeptides in nature is carried out so as to give preference to the maximum number of admissible "preferable" factors and reject the maximum number of avoidable factors. Though the contrast factors themselves comprise only a minor part

of polypeptide sequence (see above), their existence, with the underlined preference of one of them and avoidance of the others, crucially controls the synthesis of polypeptides. Accordingly, noncontrast factors occur in a quite statistical way and thereby ensure the conservation of the natural ratio of polar and fatty amino acids. The current study just emphasizes the significance of compiling a special dictionary of contrast factors for polypeptides.

(4) Thus, the role of noncontrast factors is to be the main building material for the aminoacid sequence of *Escherichia coli* and, as we may guess, also of such a sequence of any other organism. Since our proposed mathematical approach is applicable to an arbitrary aminoacid sequence, it would be of interest to check it for different organism, as well as to investigate aminoacid factors of different lengths (longer than 7, as in our present work).

The material of this paper has entirely been borrowed from [6]. Additionally, the interested reader may see also unsolved combinatorial problems in our earlier publications [7, 8].

References

[1] B. M. Broome and M. H. Hecht, "Nature disfavors sequences of alternating polar and non-polar amino acids: implications for amyloidogenesis," *Journal of Molecular Biology*, vol. 296, no. 4, pp. 961–968, 2000.

[2] Y. Mandel-Gutfreund and L. M. Gregoret, "On the significance of alternating patterns of polar and non-polar residues in beta-strands," *Journal of Molecular Biology*, vol. 323, no. 3, pp. 453–461, 2002.

[3] D. M. Cvetković, M. Doob, and H. Sachs, *Spectra of Graphs: Theory and Application*, Academic Press, Berlin, Germany, 1980.

[4] J. S. Beckmann, V. Brendel, and E. N. Trifonov, "Intervening sequences exhibit distinct vocabulary," *Journal of Biomolecular Structure and Dynamics*, vol. 4, no. 3, pp. 391–400, 1986.

[5] V. Brendel, J. S. Beckmann, and E. N. Trifonov, "Linguistics of nucleotide sequences: morphology and comparison of vocabularies," *Journal of Biomolecular Structure and Dynamics*, vol. 4, no. 1, pp. 11–21, 1986.

[6] V. R. Rosenfeld, *Function-related linguistics of noncoding sequences [Ph.D. thesis]*, The University of Haifa, 2003.

[7] V. R. Rosenfeld, "Enumerating De Bruijn sequences," *MATCH: Communications in Mathematical and in Computer Chemistry*, no. 45, pp. 71–83, 2002.

[8] V. R. Rosenfeld, "Enumerating Kautz sequences," *Kragujevac Journal of Mathematics*, vol. 24, pp. 19–41, 2002.