Bulgarian Chemical Communications, Volume 42, Number 4 (335- 337) 2010

On analyzing DNA sequences

Parisa Farhami and Ali Reza Ashrafi*

Institute for Nanoscience and Nanotechnology, University of Kashan, Kashan 87317-51167, I. R. of Iran

Received January 5, 2010; accepted May 21, 2010

A DNA sequence can be identified with a word over the alphabet W = {A, C, G, T}. He and Wang presented an algebraic method for analyzing the DNA sequences. We discuss in this paper the model of Randić, Vračko, Nandy and Basak to extend some results of He and Wang.

Keywords: DNA sequence, DNA sequence matrix

INTRODUCTION

The extremely long DNA molecule is actually made of a long string of chemical building blocks called "nucleotides." There are four different nucleotides, which are labeled: adenine (A), thymine (T), guanine (G), and cytosine (C). A DNA sequence is a succession of the letters A, C, G, and T, representing these four nucleotide bases of the DNA strand.

DNA computing is a form of computing which uses DNA, biochemistry and molecular biology instead of the traditional silicon-based computer technologies. DNA computing is fundamentally similar to the parallel computing in that it takes advantage of many different molecules of the DNA to try many different possibilities at once. Generally, in DNA computing, the DNA sequences used for the computation should be critically designed in order to reduce error that could occur during computation.

Since 2000, Randic's research group had proposed several visualization schemes for the DNA sequences [1–8]. In one of his methods, the four vertices, associated with a regular tetrahedron, are assigned to four nucleotides. The mapping between four nucleotides and the corresponding 3–D coordinates is shown below:

$(1–1, –1) \rightarrow$ A, $(–1,1, –1) \rightarrow$ G, $(–1, –1,1) \rightarrow$ C, $(1,1,1) \rightarrow$ T.

For a positive integer n, the $\mathcal{D}_n$ denotes the set of all DNA sequences of length n and $\mathcal{D} = \cup_{n\geq 1} \mathcal{D}_n$. Suppose Σ = {A, C, G, T}, $w = x_1x_2...x_n$ is a DNA sequence of length n and $L_i(w) = |\{j \mid 1 \leq j \leq i \ \& \ x_j = L\}|$, where $L \in \Sigma$. Our other notations are standard and taken mainly from [9].

MAIN RESULTS AND DISCUSSION

He and Wang [10] considered the coordinates of a graphical representation of the DNA sequence, introduced by Zhang [11]. They presented an action of the symmetric group S_4 on Z–curves and the DNA matrices, and proved that the information entropy is invariant under the action of S_4. In this section, we consider another model, presented by Randić and his group [8], and extend the results given by He and Wang, mentioned above. We name the model, presented by Randic et al, RVNB Model. We first consider the following matrix equations:

$$\begin{pmatrix} A_i \\ G_i \\ C_i \\ T_i \end{pmatrix} = \frac{i}{4}\begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} + \frac{1}{4}\begin{pmatrix} 1 & -1 & -1 \\ -1 & 1 & -1 \\ -1 & -1 & 1 \\ 1 & 1 & 1 \end{pmatrix}\begin{pmatrix} x_i \\ y_i \\ z_i \end{pmatrix} \qquad (1)$$

By simplifying Eq. (1), one can see that:

$$\begin{cases} x_i = 2(T_i + A_i) - i \\ y_i = 2(T_i + G_i) - i \\ z_i = 2(T_i + C_i) - i \end{cases}. \qquad (2)$$

For a DNA sequence w, we define its DNA sequence matrix, $\Pi(w)$, to be defined as follows:

$$\Pi(w) = \begin{pmatrix} x_1 & x_2 & \ldots & x_n \\ y_1 & y_2 & \ldots & y_n \\ z_1 & z_2 & \ldots & z_n \end{pmatrix}.$$

* To whom all correspondence should be sent:
E-mail: E-mail: : Ashrafi@kashanu.ac.ir

Let M_n denote the set of all DNA sequence matrix and $\mathbb{M} = \bigcup_{n\geq 1} M_n$. Clearly, Π is defined as a one-to-one correspondence between $\mathcal{D}$ and $\mathcal{M}$. For every $v \in \mathcal{D}_n$ and $w \in \mathcal{D}_m$, one can see that $vw \in \mathcal{D}_{m+n}$, where vw is a DNA sequence, constructed by v and w by juxtaposition. Suppose $A \in M_n$ and $B \in M_m$. Define:

$$\begin{cases} x_i' = x_i + x_A \\ y_i' = y_i + y_A \\ z_i' = z_i + z_A \end{cases} \quad (3)$$

and A*B = (A,B'), where (A,B') is defined by concatenating matrix A and matrix B' is obtained from B by applying Eq. (3). Here (x_i,y_i,z_i) and (x_i',y_i',z_i') are the ith columns of the matrix B and B', respectively, and (x_A,y_A,z_A) is the end column of A.

We are now ready to extend Proposition 2.1 of He and Wang [10] to Randic's et al. model [8].

Lemma 1. If $v, w \in \mathcal{D}$ then $\Pi(vw) = \Pi(v)*\Pi(w)$.

Proof. Suppose $|v| = n_1$, $|w| = n_2$. Then $|vw| = n_1 + n_2$ and we have:

$$\Pi(v) = \begin{pmatrix} \alpha_1(i) \\ \alpha_2(i) \\ \alpha_3(i) \end{pmatrix}_{1\leq i\leq n_1}, \quad \Pi(w) = \begin{pmatrix} \beta_1(i) \\ \beta_2(i) \\ \beta_3(i) \end{pmatrix}_{1\leq i\leq n_2},$$

$$\Pi(vw) = \begin{pmatrix} \gamma_1(i) \\ \gamma_2(i) \\ \gamma_3(i) \end{pmatrix}_{1\leq i\leq n_1+n_2}.$$

If $i \in [1,n_1]$ then

$\gamma_1(i) = 2((A_i(v)+T_i(v)-i = \alpha_1(i)$ and if $i = n_1 + k \in [n_1+1,n_1+n_2]$ then

$\gamma_1(i) = 2(T_i(vw) + A_i(vw)) - i = 2(T_{n1}(v) + A_{n1}(v)) + 2(T_k(vw) + A_k(vw)) - n_1 - k = \beta_1(i) + \alpha_1(n)$. Therefore, we prove that:

$$\gamma_1(i) = \begin{cases} \alpha_1(i) & i \in [1, n_1] \\ \beta_1(k) + \alpha_1(n) & i \in [n_1+1, n_1+n_2] \end{cases}.$$

In a similar way, one can prove:

$$\gamma_2(i) = \begin{cases} \alpha_2(i) & i \in [1, n_1] \\ \beta_2(k) + \alpha_2(n) & i \in [n_1+1, n_1+n_2] \end{cases},$$

$$\gamma_3(i) = \begin{cases} \alpha_3(i) & i \in [1, n_1] \\ \beta_3(k) + \alpha_3(n) & i \in [n_1+1, n_1+n_2] \end{cases}.$$

This completes our proof.

He and Wang [10] also presented an action of the symmetric group S_4 on the DNA sequences. We now present new generators for this group, compatible with the RVNB model. Clearly, $S_4 = \langle (G\ T), (A\ T), (C\ T) \rangle$. Suppose O(3) denotes the set of all 3×3 orthogonal matrices on real numbers. Define:

$$\Phi(G\ T) = \begin{pmatrix} 0 & 0 & -1 \\ 0 & 1 & 0 \\ -1 & 0 & 0 \end{pmatrix},$$

$$\Phi(A\ T) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & -1 & 0 \end{pmatrix},.$$

$$\Phi(C\ T) = \begin{pmatrix} 0 & -1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Extend these values to an injective homomorphism $\Phi: S_4 \rightarrow O(3)$. We now define an action of the symmetric group S_4 on the set of all DNA sequence matrices by $A^{\alpha} = A\Phi(\alpha)$, where $\alpha \in S_4$ and $A \in \mathcal{M}$.

Lemma 2. Suppose $\alpha \in S_4$ and $P, Q \in \mathcal{M}$. Then:

$\Pi\ \alpha = \alpha\Pi$,

$\alpha\ (P * Q) = \alpha\ (P) * \alpha\ (Q)$.

Proof. (i) Since {(G T), (C T), (A T)} is a generating set for S_4, it is enough to investigate equation (i), for α = (G T), (C T) and (A T). Suppose $v \in \mathcal{D}_n$. Then:

$$\Pi(v) = \begin{pmatrix} x_i \\ y_i \\ z_i \end{pmatrix} = \begin{pmatrix} 2(T_i + A_i) - i \\ 2(T_i + G_i) - i \\ 2(T_i + C_i) - i \end{pmatrix}.$$

Since $A_i + T_i + G_i + C_i = i$, we have:

$$\Phi(G\ T)\Pi(v) = \begin{pmatrix} -z_i \\ y_i \\ -x_i \end{pmatrix} =$$

$$= \begin{pmatrix} i - 2(T_i + C_i) \\ 2(T_i + G_i) - i \\ i - 2(T_i + A_i) \end{pmatrix} = \begin{pmatrix} 2(A_i + G_i) - i \\ 2(T_i + G_i) - i \\ 2(C_i + G_i) - i \end{pmatrix},$$

$$\Phi(A\ T)\Pi(v) = \begin{pmatrix} x_i \\ -z_i \\ -y_i \end{pmatrix} =$$

$$= \begin{pmatrix} 2(T_i + A_i) - i \\ i - 2(T_i + C_i) \\ i - 2(T_i + G_i) \end{pmatrix} = \begin{pmatrix} 2(T_i + A_i) - i \\ 2(G_i + A_i) - i \\ 2(C_i + A_i) - i \end{pmatrix},$$

$$\Phi(T\ C)\Pi(v) = \begin{pmatrix} -y_i \\ -x_i \\ z_i \end{pmatrix} =$$

$$= \begin{pmatrix} i - 2(T_i + G_i) \\ i - 2(T_i + A_i) \\ 2(T_i + C_i) - i \end{pmatrix} = \begin{pmatrix} 2(C_i + A_i) - i \\ 2(G_i + C_i) - i \\ 2(T_i + G_i) - i \end{pmatrix}.$$

Therefore,

$$\Phi(\alpha)\Pi(v) = \begin{pmatrix} 2(\alpha(T_i) + \alpha(A_i)) - i \\ 2(\alpha(T_i) + \alpha(G_i)) - i \\ 2(\alpha(T_i) + \alpha(C_i)) - i \end{pmatrix} = \Pi\alpha(v).$$

To prove (ii), we assume that v, w $\in \mathcal{D}$ such that Π(v) = P and Π(w) = Q. Since *α(vw) = α(v) α(w)*, Π α(vw) = Π(α(v) α(w)) = Π(α(v)) Π(α(w)) = αΠ(v)* αΠ(w) = α(P) * α(Q), which completes the proof.

REFERENCES

1 M. Randić, Chem. Phys. Lett. **317**, 29 (2000).
2 M. Randić, Chem. Phys. Lett. **386**, 468 (2004).
3 M. Randić and A. T. Balaban, J. Chem. Inf. Comput. Sci. **43**, 532 (2003).
4 M. Randić and S. C. Basak, J. Chem. Inf. Comput. Sci. **41**, 561 (2001).
5 M. Randić G. Krilov, Chem. Phys. Lett. **272**, 115 (1997).
6 M. Randić and M. Vracko, J. Chem. Inf. Comput. Sci. **40**, 599 (2000).
7 M. Randić, A. F. Kleiner and L. M. De Alba, J. Chem. Inf. Comput. Sci. **34**, 277 (1994).
8 M. Randić, M. Vračko, A. Nandy and S. C. Basak, J. Chem. Inf. Comput. Sci. **40**, 1235 (2000).
9 H.-H. Hsu, Advanced Data Mining Technologies in Informatics, Idea Group Inc., London, 2006.
10 P. He and J. Wang, J. Phys. A: Math. Gen. **37**, 7135 (2004).
11 R. Zhang and C. T. Zhang, J. Biomol. Struct. Dyn. **11**, 767 (1994).

Върху анализа на ДНК-секвенции

Париса Фархами, Али Реза Ашрафи

Институт за нано-науки и нано-технологии, Университет в Кашан, Кашан 87317-51167,

Ислямска република Иран

Постъпила на 5 януари, 2010 г.; приета на 21 май, 2010 г.

(Резюме)

He и Wang са представили алгебричен метод за анализ на ДНК-секвенциите. ДНК-секвенциите могат да се идентифицират с буква от латиницата W = {A, C, G, T}. В настоящата работа се обсъжда развитието на модела на Randić, Vračko, Nandy and Basak за разширяване на резултатите на He и Wang.