Journal of Mathematical Chemistry, Vol. 39, Nos. 3/4, May 2006 (© 2006)
DOI: 10.1007/s10910-005-9053-x

RNA secondary structure 3D graphical representation without degeneracy

Jiawei Luo, Bo Liao,* Renfa Li, and Wen Zhu

School of Computer and Communication, Hunan University, Changsha, Hunan 410082, China
E-mail: dragonbw@163.com

Received 2 June 2005; revised 30 November 2005 / Published online: 12 January 2006

A 3D graphical representation of RNA secondary structures(3DGRR) has been derived for mathematical denotation of RNA structure. The three-dimensional graphical representation avoids loss of information and resolves structures' degeneracy. The RNA pseudoknots also can be represented as three-dimensional graphical representations.

KEY WORDS: RNA secondary structure, pseudoknot, graphical representation, 3DGRR

1. Introduction

Mathematical analysis of the large volume genomic sequence or structure data is one of the challenges for bio-scientists. Graphical representation of DNA sequence provides a simple way of viewing, sorting and comparing various gene structures [1–15].

Ribonucleic acid (RNA) is an important molecule which performs a wide range of functions in the biological system. In particular, it is RNA (not DNA) that contains genetic information of virus such as HIV and therefore regulates the functions of such virus. RNA has recently become the center of much attention because of its catalytic properties, leading to an increased interest in obtaining structural information. Using the similar with the graphical representations of DNA sequences, we also can outline several graphical representations of RNA primary sequences based on 2-D and 3-D to compute the similarity of RNA primary sequences. Current RNA secondary structure comparison algorithms have focused exclusively on tree structures owing to their relative simplicity for quantitative analysis [16–18]. But tree structures refer to mathematical constructs for RNA secondary structures without pseudoknots. So we should present a new representation to analyze and to compare RNA secondary structures with pseudo-

*Corresponding author.

knots. Recently, we have proposed 3D, 6D and 7D graphical representation of RNA secondary structures [19–21], but the representation is not unique.

Here, we present a three-dimensional graphical representation of RNA secondary structures, which has no circuit or degeneracy, so that the correspondence between RNA secondary structures and RNA graphs is one to one.

2. 3-D graphical representation of RNA secondary structures (3DGRR)

The secondary structure of an RNA is a set of free bases and base pairs forming hydrogen bonds between A–U and G–C. Let A', U', G', C' denote A, U, G, C in the base pair A–U and G–C, respectively. Then we can obtain a special sequence representation of the secondary structure. We call it characteristic sequence of the secondary structure. For example, pseudoknot B corresponds the characteristic sequence $C'U'G'G'C'G'AUUGCG'A'G'A'C'C'A'UGUC'G'C'C'A'G'CUCU'G'G'U'C'U'C'CA$ (from 3′ to 5′)(see figure 1).

We will illustrate the three-dimensional characterization of RNA secondary structure. We construct a map between the bases of characteristic sequences and plots in 3-D space, then we will obtain a 3-D representation of the corresponding RNA secondary structure. In 3-D space points, vectors and directions have two components, so we will assign the following basic elementary directions to the four free bases and two base pairs.

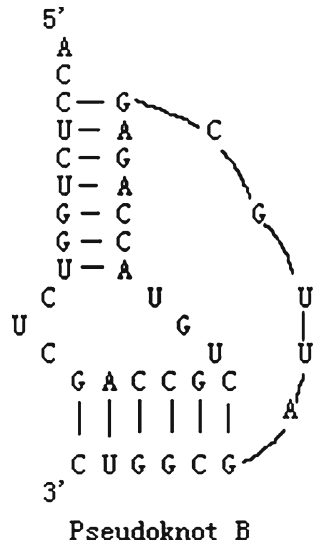

Figure 1. Pseudoknot.

$$(m, m, m) \longrightarrow A, (\sqrt{n}, \sqrt{n}, m) \longrightarrow G, (\sqrt{n}, m, \sqrt{n}) \longrightarrow C, (m, \sqrt{n}, \sqrt{n})$$
$$\longrightarrow U, (m\sqrt{u}, m\sqrt{u}, m\sqrt{u}) \longrightarrow A', (\sqrt{nu}, \sqrt{nu}, m\sqrt{u})$$
$$\longrightarrow G', (\sqrt{nu}, m\sqrt{u}, \sqrt{nu})$$
$$\longrightarrow C', (m\sqrt{u}, \sqrt{nu}, \sqrt{nu}) \longrightarrow U',$$

where m is a real number and $m \neq \sqrt{n}, m \neq \sqrt{u}, m \neq \sqrt{un}, m\sqrt{u} \neq \sqrt{n}$, n, u, nu are positive real number but not perfect square number. So that we will reduce a RNA secondary structure into a series of nodes $P_0, P_1, P_2, \ldots, P_N$, whose coordinates $(x_i, y_i)(i = 0, 1, 2, \ldots, N$, where N is the length of the RNA secondary structure being studied) satisfy

$$\begin{cases} x_i = (a_i + \sqrt{u}a'_i) \times m + \sqrt{u}g'_i) \times \sqrt{n} + (c_i + \sqrt{u}c'_i) \times \sqrt{n} + (u_i + \sqrt{u}u'_i) \times m \\ y_i = (a_i + \sqrt{u}a'_i) \times m + (g_i + \sqrt{u}g'_i) \times \sqrt{n} + (c_i + \sqrt{u}c'_i) \times m + (u_i + \sqrt{u}u'_i) \times \sqrt{n}, \\ z_i = (a_i + \sqrt{u}a'_i) \times m + (g_i + \sqrt{u}g'_i) \times m + (c_i + \sqrt{u}c'_i) \times \sqrt{n} + (u_i + \sqrt{u}u'_i) \times \sqrt{n} \end{cases} \quad (1)$$

where $a_i, c_i, g_i, u_i, a'_i, c'_i, g'_i$ and u'_i are the cumulative occurrence numbers of A, C, G, U, A', C', G', and U', respectively, in the subsequence from the 1st base to the i-th base in the sequence. We define $a_0 = c_0 = g_0 = u_0 = a'_0 = c'_0 = g'_0 = u'_0 = 0$.

We call the corresponding plot set be characteristic plot set. The curve connecting all plots of the characteristic plot set in turn is called 3D characteristic curve (3D-curve), which is determined by m, n and u that satisfy the above mentioned condition.

3. Properties

Property 1. For a given RNA secondary structure structure there is a unique 3D representation corresponding to it.

Proof. Let (x_i, y_i, z_i) be the coordinates of the i-th base of RNA secondary structure, then we have $a_i(m, m, m) + g_i(\sqrt{n}, \sqrt{n}, m) + c_i(\sqrt{n}, m, \sqrt{n}) + u_i(m, \sqrt{n}, \sqrt{n}) + a'_i(m\sqrt{u}, m\sqrt{u}, m\sqrt{u}) + g'_i(\sqrt{nu}, \sqrt{nu}, m\sqrt{u}) + c'_i(\sqrt{nu}, m\sqrt{u}, \sqrt{nu})$ $(\sqrt{nu}, m\sqrt{u}, \sqrt{nu}) + u'_i(m\sqrt{u}, \sqrt{nu}, \sqrt{nu}) = (x_i, y_i, z_i)$ i.e.

$$\begin{cases} (a_i + \sqrt{u}a'_i) \times m + (g_i + \sqrt{u}g'_i) \times \sqrt{n} + (c_i + \sqrt{u}c'_i) \times \sqrt{n} + (u_i + \sqrt{u}u'_i) \times m = x_i \\ (a_i + \sqrt{u}a'_i) \times m + (g_i + \sqrt{u}g'_i) \times \sqrt{n} + (c_i + \sqrt{u}c'_i) \times m + (u_i + \sqrt{u}u'_i) \times \sqrt{n} = y_i. \\ (a_i + \sqrt{u}a'_i) \times m + (g_i + \sqrt{u}g'_i) \times m + (c_i + \sqrt{u}c'_i) \times \sqrt{n} + (u_i + \sqrt{u}u'_i) \times \sqrt{n} = z_i \end{cases} \quad (2)$$

Obviously, x_i, y_i and z_i are irrational numbers of form $jm + k\sqrt{n} + qm\sqrt{u} + s\sqrt{nu}$, where j, k, q and s are integers. We suppose

$$x_i = j_x m + k_x\sqrt{n} + q_x m\sqrt{u} + s_x\sqrt{un}$$
$$y_i = j_y m + k_y\sqrt{n} + q_y m\sqrt{u} + s_y\sqrt{un}$$
$$z_i = j_z m + k_z\sqrt{n} + q_z m\sqrt{u} + s_z\sqrt{un}$$

then we have

$$\begin{cases} a_i + u_i = j_x \\ g_i + c_i = k_x \\ a'_i + u'_i = q_x \\ g'_i + c'_i = s_x \\ a_i + c_i = j_y \\ g_i + u_i = k_y. \\ a'_i + c'_i = q_y \\ g'_i + u'_i = s_y \\ a_i + g_i = j_z \\ c_i + u_i = k_z \\ a'_i + g'_i = q_z \\ c'_i + u'_i = s_z \end{cases} \tag{3}$$

So, for given x-projection, y-projection and z-projection of any point $P = (x, y, z)$ on the structure, after uniquely determining $j_x, k_x, j_y, k_y, j_z, k_z, q_x, q_y, q_z, s_x, s_y, s_z$ from x, y and z, the number $a_p, g_p, c_p, u_p, a'_p, g'_p, c'_p$ and u'_p of A, G, C, U, A', G', C' and U' from the beginning of the sequence to the point P can be found by solving linear system (3). By successive x-projection, y-projection and z-projection of points on the sequence, we can recover the original RNA secondary structure uniquely from the RNA graph.

The vector pointing to the point P_i from the origin O is denoted by r_i. The component of r_i, i.e. x_i, y_i and z_i are calculated by equation (1). Let $\Delta r_i = r_i - r_{i-1}$, then we have property 2.

Property 2. For any $i=1, 2, \ldots, N$, where N is the length of the studied RNA secondary structure, the vector Δr_i has only four possible direction. Furthermore, the length of Δr_i, i.e., $|\Delta r_i|$, is always equal to $\sqrt{3m^2}$, $\sqrt{m^2 + 2n}$ or $\sqrt{(m^2 + 2n)u}$, for any $i = 1, 2, \ldots, N$.

Proof. Actually, the components of Δr_i, i.e., Δx_i, Δy_i and Δz_i can be calculated for each possible residue (A,G,C, U, A', G', C' and U') at the i th position of the RNA secondary structure by using eq.(1). For example, when the i th residue is A, we find $\Delta x_i = m, \Delta y_i = m, \Delta z_i = m$. This result is independent of the conformation state of the $(i-1)$ th residue. The two numbers (m, m, m) are called the direction of Δr_i. The direction number and the length of Δr_i for each possible residue type at the i th position are summarized as follows (Table 1).

Property 3. There is no circuit or degeneracy in our three-dimensional graphical representation.

Table 1
Eight possible direction.

	Δx_i	Δy_i	Δz_i	$\lvert\Delta r_i\rvert$
A	m	m	m	$\sqrt{3m^2}$
G	$\sqrt{n}$	$\sqrt{n}$	m	$\sqrt{m^2+2n}$
C	$\sqrt{n}$	m	$\sqrt{n}$	$\sqrt{m^2+2n}$
U	m	$\sqrt{n}$	$\sqrt{n}$	$\sqrt{m^2+2n}$
A'	$m\sqrt{u}$	$m\sqrt{u}$	$m\sqrt{u}$	$\sqrt{3m^2u}$
G'	$\sqrt{nu}$	$\sqrt{nu}$	$m\sqrt{u}$	$\sqrt{(m^2+2n)u}$
C'	$\sqrt{nu}$	$m\sqrt{u}$	$\sqrt{nu}$	$\sqrt{(m^2+2n)u}$
T'	$m\sqrt{u}$	$\sqrt{nu}$	$\sqrt{nu}$	$\sqrt{(m^2+2n)u}$

Proof. We assume that: (1) the number of nucleotide forming a circuit is e; (2) the number of A,G,C, U, A', G', C' and U' in a circuit is $a_e, g_e, c_e, u_e, a'_e, g'_e, c'_e$ and u'_e, respectively. So $a_e + g_e + c_e + u_e + a'_e + g'_e + c'_e + u'_e = e$. Because $a_eA, g_eG, c_eC, u_eU, a'_eA', g'_eG', c'_eC'$ and u'_eU form a circuit, the following equation holds: $a_i(m,m,m)+g_i(\sqrt{n},\sqrt{n},m)+c_i(\sqrt{n},m,\sqrt{n})+u_i(m,\sqrt{n},\sqrt{n})+a'_i(m\sqrt{u},m\sqrt{u},m\sqrt{u})+g'_i(\sqrt{un},\sqrt{nu},m\sqrt{u})c'_i(\sqrt{nu},m\sqrt{u},\sqrt{nu})+u'_i(m\sqrt{u},\sqrt{nu},\sqrt{nu})=(0,0,0)$ i.e.

$$\begin{cases} a_em + g_e\sqrt{n} + c_e\sqrt{n} + u_em + a'_em\sqrt{u} + g'_e\sqrt{nu} + c'_e\sqrt{nu} + u'_em\sqrt{u} = 0 \\ a_em + g_e\sqrt{n} + c_em + u_e\sqrt{n} + a'_em\sqrt{u} + g'_e\sqrt{nu} + c'_em\sqrt{u} + u'_e\sqrt{nu} = 0. \\ a_em + g_em + c_e\sqrt{n} + u_e\sqrt{u} + a'_em\sqrt{u} + g'_em\sqrt{u} + c'_e\sqrt{nu} + u'_e\sqrt{nu} = 0 \end{cases} \quad (4)$$

Clearly eq.(4) hold if, and only if $a_e = g_e = c_e = u_e = a'_e = g'_e = c'_e = u'_e = 0$. Therefore, $e = 0$, which means no circuit exists in this graphical representation.

Property 4. The 3D representation possesses the reflection symmetry.

Proof. usually the sequence is expressed in the order from 5′ to 3′. Suppose that the 3D representation for RNA secondary structure is described by $(x_i, y_i, z_i), i = 0, 1, 2, \ldots, N$. Suppose again that the 3D representation for the reverse structure, i.e, the same sequence but from 3′ to 5′ is described by $(\hat{x_i}, \hat{y_i}, \hat{z_i})$, we find

$$\begin{cases} \hat{x_i} = x_N - x_{N-i} \\ \hat{y_i} = y_N - y_{N-i}. \\ \hat{z_i} = z_N - z_{N-i} \end{cases} \quad (5)$$

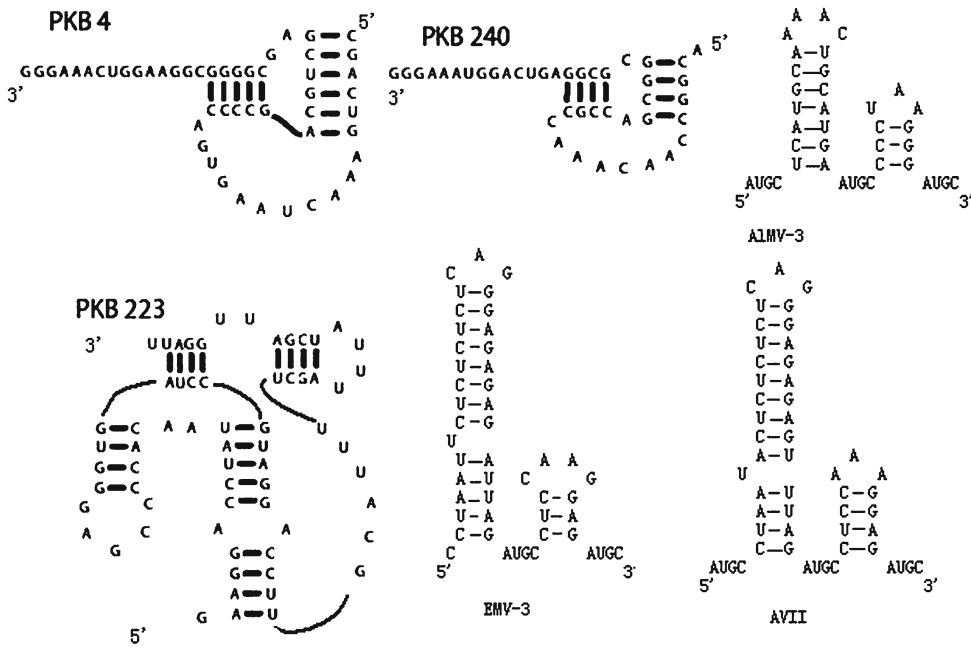

Figure 2. RNA secondary structures.

4. Similarities/Dissimilarities

In this section, we will make a comparison for the secondary structures belonging to six different species (see figure 2) based on our 3DGRR.

A direct comparison of these RNA secondary structures using computer codes is somewhat less straightforward due to the fact that the RNA secondary structures have different lengths. We will construct a 4-component vector consisting of the normalized leading eigenvalue of the M/M matrix (proposed by M.Randic [8]) of the 3DGRR with different parameters. As an example, we will use the following four 3DGRR:

Case a. Letting $m = 1/2; n = 3/4;$ and $u = \frac{4}{5}$, then we get

$$\begin{cases} x_i = (a_i + \sqrt{\frac{4}{5}}a_i')/2 + (g_i + \sqrt{\frac{4}{5}}g_i') \times \sqrt{\frac{3}{4}} + (c_i + \sqrt{\frac{4}{5}}c_i') \times \sqrt{\frac{3}{4}} + (u_i + \sqrt{\frac{4}{5}}u_i')/2 \\ y_i = (a_i + \sqrt{\frac{4}{5}}a_i')/2 + (g_i + \sqrt{\frac{4}{5}}g_i') \times \sqrt{\frac{3}{4}} + (c_i + \sqrt{\frac{4}{5}}c_i')/2 + (u_i + \sqrt{\frac{4}{5}}u_i') \times \sqrt{\frac{3}{4}}. \\ z_i = (a_i + \sqrt{\frac{4}{5}}a_i')/2 + (g_i + \sqrt{\frac{4}{5}}g_i')/2 + (c_i + \sqrt{\frac{4}{5}}c_i') \times \sqrt{\frac{3}{4}} + (u_i + \sqrt{\frac{4}{5}}u_i') \times \sqrt{\frac{3}{4}} \end{cases}$$

Case b. Letting $m = 1/2; n = 4/5$ and $\frac{4}{7}$, then we get

$$\begin{cases} x_i = (a_i + \sqrt{\frac{4}{7}}a_i')/2 + (g_i + \sqrt{\frac{4}{7}}g_i') \times \sqrt{\frac{4}{5}} + (c_i + \sqrt{\frac{4}{7}}c_i') \times \sqrt{\frac{4}{5}} + (u_i + \sqrt{\frac{4}{7}}u_i')/2 \\ y_i = (a_i + \sqrt{\frac{4}{7}}a_i')/2 + (g_i + \sqrt{\frac{4}{7}}g_i') \times \sqrt{\frac{4}{5}} + (c_i + \sqrt{\frac{4}{7}}c_i')/2 + (u_i + \sqrt{\frac{4}{7}}u_i') \times \sqrt{\frac{4}{5}}. \\ z_i = (a_i + \sqrt{\frac{4}{7}}a_i')/2 + (g_i + \sqrt{\frac{4}{7}}g_i')/2 + (c_i + \sqrt{\frac{4}{7}}c_i') \times \sqrt{\frac{4}{5}} + (u_i + \sqrt{\frac{4}{7}}u_i') \times \sqrt{\frac{4}{5}} \end{cases}$$

Case c. Letting $m = 1/3; n = 4/7$ and $u = \frac{1}{2}$, then we get

$$\begin{cases} x_i = (a_i + \sqrt{\frac{1}{2}}a_i')/3 + (g_i + \sqrt{\frac{1}{2}}g_i') \times \sqrt{\frac{4}{7}} + (c_i + \sqrt{\frac{1}{2}}c_i') \times \sqrt{\frac{4}{7}} + (u_i + \sqrt{\frac{1}{2}}u_i')/3 \\ y_i = (a_i + \sqrt{\frac{1}{2}}a_i')/3 + (g_i + \sqrt{\frac{1}{2}}g_i') \times \sqrt{\frac{4}{7}} + (c_i + \sqrt{\frac{1}{2}}c_i')/3 + (u_i + \sqrt{\frac{1}{2}}u_i') \times \sqrt{\frac{4}{7}}. \\ z_i = (a_i + \sqrt{\frac{1}{2}}a_i')/3 + (g_i + \sqrt{\frac{1}{2}}g_i')/3 + (c_i + \sqrt{\frac{1}{2}}c_i') \times \sqrt{\frac{4}{7}} + (u_i + \sqrt{\frac{1}{2}}u_i') \times \sqrt{\frac{4}{7}} \end{cases}$$

Case d. Letting $m = 1/2; n = 1/3$ and $u = \frac{4}{5}$, then we get

$$\begin{cases} x_i = (a_i + \sqrt{\frac{4}{5}}a_i')/2 + (g_i + \sqrt{\frac{4}{5}}g_i') \times \sqrt{\frac{1}{3}} + (c_i + \sqrt{\frac{4}{5}}c_i') \times \sqrt{\frac{1}{3}} + (u_i + \sqrt{\frac{4}{5}}u_i')/2 \\ y_i = (a_i + \sqrt{\frac{4}{5}}a_i')/2 + (g_i + \sqrt{\frac{4}{5}}g_i') \times \sqrt{\frac{1}{3}} + (c_i + \sqrt{\frac{4}{5}}c_i')/2 + (u_i + \sqrt{\frac{4}{5}}u_i') \times \sqrt{\frac{1}{3}}. \\ z_i = (a_i + \sqrt{\frac{4}{5}}a_i')/2 + (g_i + \sqrt{\frac{4}{5}}g_i')/2 + (c_i + \sqrt{\frac{4}{5}}c_i') \times \sqrt{\frac{1}{3}} + (u_i + \sqrt{\frac{4}{5}}u_i') \times \sqrt{\frac{1}{3}} \end{cases}$$

In table 2, we give the similarities and dissimilarities for the six RNA secondary structures based on the Euclidean distances between the end points of the 4-component vectors of the normalized leading eigenvalues of the M/M matrices. We believe that it is not accidental that the smallest entries in table 2 are associated with the pairs (pkb240, pkb4), and (EMV-3, AVII).

Table 2
The similarity/dissimilarity matrix for the six RNA secondary structures based on the Euclidean distances between the end points of the 4-component vectors of the normalized leading eigenvalues of the M/M matrices.

Species	AlMV-3	pkb240	pkb 223	EMV-3	pkb 4	AVII
AlMV-3	0	0.0880	0.0506	0.0272	0.0920	0.0214
pkb240		0	0.0459	0.0637	0.0056	0.0718
pkb 223			0	0.0237	0.0481	0.0302
EMV-3				0	0.0672	0.0084
pkb 4					0	0.0752
AVII						0

5. Conclusion

High complexity and degeneracy are major problems in previous RNA secondary structure representations. Our representation provides a direct plotting method to denote RNA secondary structures without degeneracy. From the RNA graph, the A,U,G,C,A–U and C–G usage as well as the original RNA structure can be recaptured mathematically without loss of textual information. Many properties of visual importance in a RNA secondary structure are preserved in the 3D characteristic curve. It is useful for visualizing the local and global features of big or small RNA secondary structures and can facilitate the visual discovery of interesting features in a RNA secondary structure. The current three-dimensional graphical representation of RNA secondary structures provides different approaches for both computational scientists and molecular biologists to analysis RNA secondary structures efficiently with different parameter n, m and u.

References

[1] B. Liao, Chem. Phys. Lett. 401 (2005) 196.
[2] C. Yuan, B. Liao and T. Wang, Chem. Phy. Lett. 379 (2003) 412.
[3] B. Liao and T. Wang, J. Comput. Chem. 25(11) (2004) 1364.
[4] B. Liao and T. Wang, J. Mol. Struct. *THEOCHEM,* 681(2004) 209.
[5] B. Liao and T. Wang, Chem. Phys. Lett. 388 (2004) 195.
[6] S. S.-T. Yan, J. Wang, A. Niknejad, C. Lu, N. Jin and Y.-K. Ho, Nucleic Acids Res. 31(12) (2003) 3078.
[7] M. Randic, M. Vracko, A. Nandy and S.C. Basak, J. Chem. Inf. Comput. Sci. 40 (2000) 1235.
[8] M. Randic, M. Vracko, N. Lers and D. Plavsic, Chem. Phys. Lett. 368 (2003) 1.
[9] E. Hamori and J. Ruskin, J. Biol. Chem. 258 (1983) 1318.
[10] E. Hamori, Nature 314 (1985) 585.
[11] M.A. Gates, Nature 316 (1985) 219.
[12] A. Nandy, Curr. Sci. 66 (1994) 309.
[13] A. Nandy, Comput. Appl. Biosci. 12 (1996) 55.
[14] B. Liao, M. Tan and K. Ding, Chem. Phys. Lett. 402 (2005) 380.
[15] B. Liao, Y. Zhang, K. Ding and T. Wang, J. Mol. Struct. *THEOCHEM,* 717 (2005) 199.
[16] H.H. Gan, S. Pasquali and T. Schlick, Nucleic Acids Res. 31 (2003) 2926.
[17] B.A. Shapiro and K.Z. Zhang, Comput. Biomed. Res. 6 (1990) 309.
[18] S.Y. Le, R. Nussinov and J.V. Maizel, Comput. Biomed. Res. 22 (1989) 461–473.
[19] B. Liao and T. Wang, J. Biomol. Struc. Dynamics 21 (2004) 827.
[20] B. Liao, K. Ding and T. Wang, J. Biomol. Struct. Dynamics 22 (2005) 455.
[21] B. Liao, T. Wang, and K. Ding, Lecture Ser. Comp. Comput. Sci. 1 (2004) 310.