Digital Coding for Amino Acid Based on Cellular Automata

Xuan Xiao[1, 2], Shihuang Shao[1], Yongsheng Ding[1], and Xiaojing Chen[1]

[1] College of Information Sciences and Technology, Donghua University, Shanghai 200051, P. R. China
[2] Department of Computer, Jingdezhen Ceramic Institute, Jingdezhen, Jiangxi 333000, P. R. China
Email: xioaxuan@mail.dhu.edu.cn, shshao@dhu.edu.cn, ysding@dhu.edu.cn

Abstract *- With the development of bioinformatics, it is a valuable subject to study the biological sequence on the protein aspect with the computer based on the threshold of encoding 20 kinds of amino acids, because many biological phenotype characters, gene regulating control network are determined by amino acids sequence. It is feasible to analyze genetic signals in amino acid sequences because there are already a lot of mature information process systems. This paper makes use of similarity rule, complementaity rule, molecular recognition theory, and information theory to set up a model of digital coding for amino acids. The model reflects better amino acid chemical physical properties and degeneracy. It transforms the symbolic DNA sequences into digital genetic signals of amino acids and offers the possibility to utilize Cellular Automata. Correspondingly, it opens the possibility to apply a whole range of powerful signal processing methods for analysis of amino acids.*

Keywords: Digital coding, amino acid, genetic information, cellular automata

1. Introduction

In the last few decades, advances in molecular biology and the equipments available for research in this field have allowed the increasing rapid in sequencing of large portions of the genomes of several species. In fact, the human genome project, designed to sequence all of the 24 human chromosomes, has been achieved [1-3]. Popular sequence databases, such as GenBank and EMBL, have been growing at exponential rates. This deluge of information has offered tremendous opportunities to explore in depth its content. Information science applied to biology initiates a promising field called Bioinformatics**.** To understand the function of genome and proteome is the main task of post-genome era. Genome has contained all the genetic information needed, inc[1]luding its products and regulation information. In nature, genome is a complicated information system, which obeys the same regulation law as a general information system does. Consequently, we get a good way to understand the function of genome on the perspective of the information system**,** which makes it feasible to deal with the biological information for there are already a lot of mature information process systems. Till now, the more important thing is how to change biological information into the digital signal.

Biological information can be analyzed on several levels, for instance nucleotide sequence, protein, genome, etc. Among them, proteins are the molecules which play the largest role in the functioning of the body. Some kinds of protein molecules buildup the structures in body, and other kind of proteins are involved in carrying energy around the body and in communicating chemical signals between organs. In conclusion, it is a valuable subject to study the biological sequence on the protein sequences because many biological phenotype characters, gene regulating, and control network are determined by amino acids sequence.

Many kinds of models on amino acid digital encoding have been built. P. Cristea proposed a representation of genetic code, which converts the DNA sequences into digital signals and uses a base four representation of the nucleotides. It leads to the conversion of the codons into numbers in the range 0-63 and of the amino acids, together with the terminator, into numbers in the range 0-20[4]. In this model, amino acids are coded as: F=0, L=1, S=2, Y=3, end=4, C=5, W=6, P=7, H=8, Q=9, R=10, I=11, M=12, T=13, N=14, K=15, V=16, A=17, D=18, E=19, G=20. This model reflects better amino acid structure and degeneracy, and the genetic signals built from genes on this model show low auto-correlation.

Pan et al. also proposed a kind of amino acid coding when they predicted protein sub-cellular location based on stochastic signal processing approach [5]. For simplicity, the model is: A=10, C=20, D=30, E=40, F=50, G=60, H=70, I=80, K=90, L=100, M=110, N=120, P=130, Q=140, R=150, S=160, T=170, V= 180, W=190, Y=200.

[1] * **0-7803-8566-7/04/$20.00** Ó **2004 IEEE.**

Though two different procedures above can encode a protein sequence to a serial of digital signals, they only distinguish each amino acid in the process of encoding amino acid, however the physics chemical properties of the amino acid were neglected.

When Sofer predicted secondary structure of proteins using genetic algorithms [6], he assigned one or two five-digit codes to each amino acid because the rules of genetic algorithms are often encoded as binary strings. Amino acid with similar properties has similar code words. But a shortcoming of this model is that the amino acid and its digital coding are not one-one corresponding. According to this rule, there are 12 amino acids own two possibility digital coding.

N. Stambuk built an encoding model according to molecular recognition theory [7], As far as physical chemical properties of amino acid are considered in this model, 64 codons are subbed one-to-one correspondence 0-63 6 binary numbers. However, in such a model, if a nucleotide changes and the amino acid doesn't change, the digital code will change as before. As a result, the image may be changed if the digital signals changed, no matter the amino acid sequence has changed or not. So this case is unfavorable for us to utilize cellular automata to study gene expression and regulation mechanism. In some sense, it is essential to build a model for amino acid digital encoding which considers amino acid chemical physical properties and also make each amino acid corresponding to only one digital code.

In this paper, a solution based on molecular recognition theory, similarity rule, complementarity rule, and information theory is put forward.

2. Genetic code

The genetic code is universal as it is used by all known organisms. DNA (deoxyribonucleic acid), RNA (ribonucleic acid), and proteins are macromolecules which are unbranched polymers built up from smaller units. In the case of DNA, these unites are the 4 nucleotide residues A (adenine), C (cytosine), G (guanine), and T (thymine), while for RNA the units are A, C, G, and U (uracil). When coming to proteins, the units are the 20 amino acid residues A (alanine), C(cysteine), D(aspartic acid), E (glutamic acid), F(phenylalanine), G (glycin), H(histidine), I(isoleucine), K(lysine), L(leucine), M(methionine), N(asparagines), P (proline), Q (glutamine), R(arginine), S(serine), T(threonine), V(valine), W(tryptophan), and Y(tyrosine). The genetic code of proteins is given in Table 1. There are 64 codons, out of which 61 codons encode 20 amino acids, while the other three ones correspond to terminators [3]. Consequently, there is a degeneracy of the genetic code, most amino acids being inserted into a growing polypeptide chain in response to two or more different triplets in the mRNA.

The origin and development of the genetic code are not well understood. Each of the 20 a-amino acids found in proteins can be distinguished by the R-group substitution on the a-carbon atom. There are two broad classes of amino acids based upon whether the R-group is hydrophobic or hydrophilic.

Is there a relationship between the pattern of nucleic acids in the codons and the physico-chemical properties of the amino acids? J. C. Biro designed a periodic table of codons where the codons are in regular locations. The resulting nucleic acid periodic table showed perfect axial symmetry for codons [8]. The corresponding amino acid table also displaced periodicity regarding the biochemical properties (charge and hydropathy) of the 20 amino acids and the position of the stop signals. The periodic table shows the importance of the central nucleotide in the codons and predicts the control of charge by purines, while the polarity of the amino acids were determined by pyrimidines

Table 1: Genetic Code

First Position in Codon	Second Position in Codon: T	C	A	G	Third Position in Codon
T	TTT Phe [F]	TCT Ser [S]	TAT Tyr [Y]	TGT Cys [C]	T
	TTC Phe [F]	TCC Ser [S]	TAC Tyr [Y]	TGC Cys [C]	C
	TTA Leu [L]	TCA Ser [S]	TAA Ter [end]	TGA Ter [end]	A
	TTG Leu [L]	TCG Ser [S]	TAG Ter [end]	TGG Trp [W]	G
C	CTT Leu [L]	CCT Pro [P]	CAT His [H]	CGT Arg [R]	T
	CTC Leu [L]	CCC Pro [P]	CAC His [H]	CGC Arg [R]	C
	CTA Leu [L]	CCA Pro [P]	CAA Gln [Q]	CGA Arg [R]	A
	CTG Leu [L]	CCG Pro [P]	CAG Gln [Q]	CGG Arg [R]	G
A	ATT Ile [I]	ACT Thr [T]	AAT Asn [N]	AGT Ser [S]	T
	ATC Ile [I]	ACC Thr [T]	AAC Asn [N]	AGC Ser [S]	C
	ATA Ile [I]	ACA Thr [T]	AAA Lys [K]	AGA Arg [R]	A
	ATG Met [M]	ACG Thr [T]	AAG Lys [K]	AGG Arg [R]	G
G	GTT Val [V]	GCT Ala [A]	GAT Asp [D]	GGT Gly [G]	T
	GTC Val [V]	GCC Ala [A]	GAC Asp [D]	GGC Gly [G]	C
	GTA Val [V]	GCA Ala [A]	GAA Glu [E]	GGA Gly [G]	A
	GTG Val [V]	GCG Ala [A]	GAG Glu [E]	GGG Gly [G]	G

The hydrophobic amino acids tend to repel the aqueous environment, and therefore reside predominantly in the interior of proteins. This class of amino acids does not ionize nor participate in the formation of H-bonds. The hydrophilic amino acids that tend to interact with the aqueous environment are often involved in the formation of H-bonds and are predominantly found on the exterior surfaces proteins or in the reactive centers of enzymes.

3. Similarity rule, complementarity rule and molecular recognition theory

All intermolecular processes (molecular recognition and molecular assembling or the formation of any kinds of chemical bond) and intermolecular processes between molecular moieties are governed either by similarity rule or by complementarity rule or both [9-10]. Similarity rule shows that a component in a molecular recognition process loves others of analogical properties, such as hydrophobic

interaction, similarity in softness of the well-known hard-soft-acid-base rules. It predicts the affinity of individuals of similar properties.

On the contrary, complementarity rule predicts the affinity of individuals of certain contrast properties. All kinds of donor-acceptor interaction, such as enzyme and substrate combination, which may involve hydrogen bond, electrostatic interaction, and stereochemical key-and-lock docking, follow the complementarity rule. Both types of rules still remain strictly empirical.

An interesting pattern in the genetic code was reported previously. Whether in the 5'-to-3' or the 3'-to-5' direction, codons for hydrophilic and hydrophobic amino acids are generally complemented by codons for hydrophobic and hydrophilic amino acids respectively. The average tendency of codons for unchanged (slightly hydrophilic) amino acids was to be complemented by codons for unchanged amino acids.

Molecular Recognition Theory (MRT) [11-15] explains the experimental results that peptides specified by the complementary RNAs and DNAs bind to each other with high efficacy and specificity. This is supported by the evidence that peptides deduced from the same reading frame of complementary strands of nucleic acids (antisense peptides) have the ability to bind one and another. Thus, peptides deduced from complementary DNA strands possess a hydropathic complementary which should induce amphiphilic structures and promote binding. Furthermore, observations show that antibodies to peptides encoded by the complementary mRNA seemed to recognize their own peptide receptors and comparison of the mRNA sequences. These sequences are related to interleukin-2, epidermal growth factor, and transferring with their respective receptor. Such an observation result suggests that this binding has important biological and evolutionary significance. According to probabilities of appearance for different amino acid pairs from high to low, they are arranged in an order to be : L-D, A-R, E-L, G-P, S-S, V-H, M-Y, Y-I, V-Q, L-N, C-T, S-R, F-K, W-T. 14 distinct groups of complementary RNA/DNA coding pairs define all 64 codons. The probability of appearance for each amino acid complementary pair (P) within peptide motifs is defined as:

$$P = n / N \tag{1}$$

where, n being the number of detected pairs of the same type and N being the total number of all matching pairs.

4. Optimal model of amino acids digital coding

It is well known that all the proteins occurring in living organisms are composed of a total of just 20 different chemical building blocks (amino acids). Information theory makes it possible to determine the smallest binary number of a word in order to allow unambiguous identification of all amino acids. If words are made up of 4 bits/word, these contain too little information. Six bits/word would be too complex. According to information theory, words having five bits/word are sufficient and are therefore the most economical method of coding. Five binary numbers can mostly present 32 states from which we have to select available 21 states among them. According to combinatorics, this encoding format has C_{32}^{21} kinds. Each amino acid has its possible complementary pair based on MRT, they have some partial symmetrical on the physics chemical property, so their code should be designed symmetrical.

The existence of the four different nitrogenous bases strongly suggests the mapping of the nucleotides to the digits {0, 1, 2, 3} [16]. There are two complementary pairs 0 (00) and 3(11), 1(01), and 2(10) in the four digits, and also C and G, A and U in the four nucleotide base. To preserve the symmetry within complementary and stationary RNA coding strands, the binary notation is given in Table 2.

Table 2. Mapping of Nucleotides to Digits in Four Bases

Pyrimidines
Cytosine =C =00
Uracil =U=01
Purines
Adenine =A =10
Guanine=G=11

The reasons are given as:

(1) In history, there is a precedent to arrange elements with the order on the molecular weight. For example, the periodic table of elements is arranged in the order according to the size of atomic weight at first in chemistry. Four kinds of bases are arranged in an order according to the molecular weight: C=111.1, U=112.1, A=135.13, G=151.13, so the code is 0123/CUAG.

(2) 0123/CUAG code can reflect the chemical properties of four kinds of bases. The first digit is called as structural encoding bit in the binary coding of the bases. For the first digit discrimination, pyrimidines (Y) are denoted by 0 and purines (R) by 1. And the end digit is encoding bit for functional gene group, 1 denotes keto group, such as U (01) and G (11); 0 codes amino group, such as C(00) and A(10).

This notation ensures that 0-1 digit replacements define complementary signal changes with respect to the stationary one by means of strong-weak H bonding distinction.

In general, the first two nucleotide bases are the same in the codons what determine the same amino acid while the third nucleotide base is different. This indicates that the first two bases determine the different properties of amino acid, and the third base plays an unimportant role. According to MRT, we know that first two bases of each possible complementary amino acid pair are completed complementary which reminds us when we code the amino acid, the first two bases should be considered especially, the first four digit of five binary coding should be encode by the first two bases of codons. On the other words, the first four 01 digital coding of amino acid is coded at the order of appearance of the first two bases coded amino acid pairs, the fifth digit is decided by other two factors. They are : (1) two amino acid have alike property, their coding should be close too; (2) if the first two bases of the codons of the amino acid are same, the fifth digit is decided by molecular weight of the amino acid, the big one of molecular weight is 1, the small one is 0.

The optimal code model is drawn as Table 3 according to above-mentioned principles. There are only two one codon-one amino acid (non degenerated) mappings for Tryptophan and Methionine, but ten double, three triple, six quadrille, and two sextuple degeneracy. Judging from the frequency of the amino acids in the proteins, it is obviously that the genetic code presents the features of an entropic coding.

Table 3. Binary Notation of Amino Acid Coding Language

Amino acid	codon	binary notation	amino acid	codon	binary notation
ccu ccc cca ccg	P	00001	cuu cuc cua cug uua uug	L	00011
caa cag	Q	00100	cau cac	H	00101
cgu cgc cga cgg aga agg	R	00110	ucu ucc uca ucg agu agg	S	01001
uau uac	Y	01100	uuu ucc	F	01011
ugg	W	01110	ugu ugc	C	01111
acu acc aca acg	T	10000	auu auc aua	I	10010
aug	M	10011	aaa aag	K	10100
aau aac	N	10101	gcu gcc gca gcg	A	11001
guu guc gua gug	V	11010	gau gac	D	11100
gaa gag	E	11101	ggu ggc gga ggg	G	11110
uaa uag uga	end	11111			

If symbol system is composed of n symbols, entropy S of the structure is given by the following familiar Gibbs-Shannon expression:

$$S = -\sum_{i=1}^{n} p_i \ln p_i \qquad (2)$$

where, p_i is the probability of the each symbol to appear, with the understanding that $0\ln 0 = 0$ and $1\ln 1 = 0$. Because $1 \geq p_i \geq 0$ and $\ln p_i < 0$, entropy is nonnegative. According to the rule of information theory, Lin proved that the higher-similarity-higher-entropy relation and the Similarity principle: The higher the similarity among the components is, the higher value of entropy will be and the higher stability will be [9].

The encoding model in Table 3 owns seven completely symmetrical pairs while other three are majority symmetrical too. It accords with physics chemical property of amino acid as well as the code request of information theory.

5. Application: gene visualization based on cellular automata

Cellular automata consist of a regular lattice with a discrete variable at each site. A set of rules specify the time and space evolution of the system, which is discrete in both variables [17]. These systems have attracted much interest in very recent years because even with simple rules cellular automata may show very complex evolution patterns. For example, it can imitate the character of dynamics which describes a complicated immense system existing in the nature [18]. We can also adopt it to study about biological information visualization. For example, visualization of

amino acid sequence.

There is a common characteristic in foregone visual methods of gene sequence, though the curves that are educed have two-dimensional or three-dimensional. The point of the special curve that corresponding to the certain nucleic acids, is jointly determined by the base itself and bases before without considering the bases behind[19-20]. This shortcoming can be solved by use of cellular automata, because the neighbors of lattice in cellular automata contain both fore-and-aft lattices.

Now we transformed the symbolic DNA sequences into digital genetic signals of amino acids by using two kinds of coding model respectively, the first coding model is Nikola's, 64 codons are subbed one-to-one correspondence 0-63 6 binary numbers, and the second is our model. The gene images can be gained through inputting digital signal into cellular automata and running the suitable evolution rule.

The coding model of Figure 1 to 3 all adopted Nikola's. Figure 1 is the C gene image of Hepatitis B virus (HBV) whose accession is ab059661. The sequence comes from the website of National Center for Biotechnology Information (NCBI) in the United States, www.ncbi.nlm.nih.gov. The evolution rule is 84^{th}, compressed by 2:2 after running 300 times. C gene is between 1814 and 2452 in the sequence. Figure 2 is also the C gene image that the amino acid sequence is the same as Figure 1's, but the nucleotides have been mutated. Though Figure 1 and Figure 2 have the same amino acid sequence, two images have different pattern of white and black thread, the image that created by this coding model cannot reflect the amino acid composition of the gene sequence.

Figure 3 is created by ours coding model. If the amino acid sequence is same, the image will not changed no matter how the nucleotides has been mutated, the image and its sequence are one-one corresponding. From Figure 3 to Figure 5, we can see that the images can easily discriminate the differences and similarities among various gene sequences and potentially embody important parameters concerning the gene expression and regulation because evolving rule considers the interaction of gene signal in the course of image producing. It offers a bridge between image recognition and gene expression.

Figure1. Image of HBV C gene is generated by CA 84^{th} rule, the evolving step is 300, the sequence is obtained from NCBI GenBank (ab059661), the coding model is Nikola's, the compression ratio is 4:4.

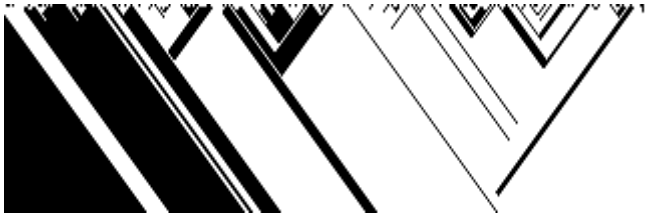

Figure 2. Image of HBV C gene is generated by CA 84^{th} rule, the amino acid sequence is same as Figure 1 but the nucleotides has been mutated. The evolving step is 300, the coding model is Nikola's. the compression ratio is 4:4.

Figure 3. Image of Hepatitis B virus C gene is generated by CA 84^{th} rule, the evolving step is 300, the sequence is obtained from NCBI GenBank (ab059661). The coding model is ours.

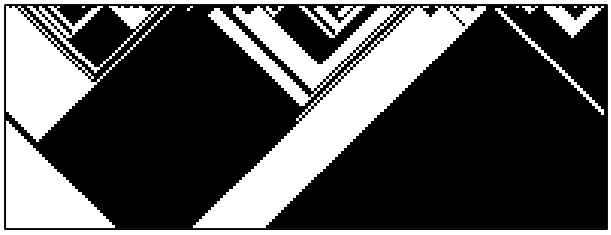

Figure4. Compressed image of the mouse TGFA gene. The sequence was obtained from NCBI GenBank (P 01134), the compression ratio is 4:4, the evolving step is 300. The coding model is ours.

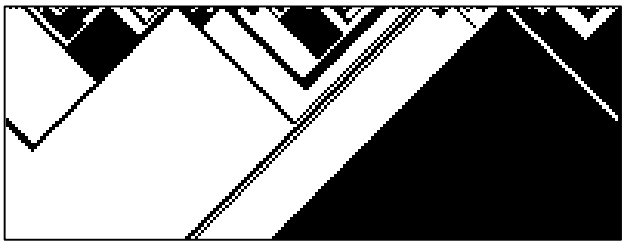

Figure5. Compressed image of the human TGFA gene. The sequence was obtained from NCBI GenBank (AAH05308), the compression ratio is 4:4, the evolving step is 300. The coding model is ours.

6. Conclusions

This paper introduces optimal symbolic-to-digital mappings for amino acids based on the similarity rule,

complementarity rule and molecular recognition theory. This code model reflects better the amino acid physics chemical properties and degeneracy as well, which can not only be applied in gene visualization based on cellular automata but also be used to study the genetic signals by many kinds of powerful signal processing methods.

Acknowledgment

The work in this research was supported in part by Doctoral Foundation from National Education Committee (20030255009), P. R. China.

REFERENCES

1. J. C. Venter., H. O. Smith and L. Hood, "A new strategy for genome sequencing", *Nature*, Vol. 381, pp. 364-366, 1996

2. J. C. Venter et al. "Shotgun sequencing of the human genome", *Science*, Vol. 280, pp. 1540-1542,1998.

3. J. W. Meinke, J. M. Cherry, C. Dean, S. D. Rounsley and M. Koornneef, "Arabidopsis thaliana: A moder plant for genome analysis", *Science*, Vol. 282, pp. 662-682, 1998.

4. P. Cristea, "Independent component analysis for genetic signals", *SPIE Conference BIOS 2001-international Biomedical optics Symposium*, San Jose, USA, pp. 20-26 January 2001.

5. Y. X. Pan, Z. Z. Zhang, Z. M. Guo, G. Y. Feng, Z. D. Huang and L. He, "Application of pseudo amino acid composition for predicting protein subcellular location: Stochastic signal processing approach", *Journal of Protein Chemistry*, Vol. 22,No. 4, pp.395-402, 2003.

6. W. H. Sofer, "Predicting secondary structure of proteins using genetic algorithms". http://waksman. Rutgers. Edu/Waks/Sofer/sofer.Html.

7. S. Nikola, "On the genetic origin of complementary protein coding", *Croatica Chemica acta*, Vol. 71, No 3, pp. 573-589, 1998.

8. J. C. Biro, B. benyo, C. Sansom, A. Szlavecz, et al. "A common periodic table of codons and amino acids", *Biochemical and Biophysical research communications*, Vol. 306, pp. l408-1415, 2003.

9. S. K. Lin, "The nature of the chemical process. 1. symmetry evolution-revised information theory, similarity principle and ugly symmetry", *Int. J. Mol. Sci*., Vol. 2, pp. 10-39, 2001.

10. Y. Duan, S. C. Harvey and P. A. Kollman, "A protein folding and beyond", in *Chemistry for the 21st century* ; E. keinan, I. Schechter, Eds, Wiley-VCH: Weinheim, pp. 89-101, 2001.

11. E. Blalock, "Genetic origins of protein shape and interaction rules", *Nature Medicine*, Vol. 1, pp. 876-878, 1995

12 . L. Baranyi, W. Campbell, K. Ohishima, S. Fujimoto, M. Boros and H. Okada, "The antisense homology box: A new motif within proteins that encodes biologically active peptides", *Nature medicine*, Vol. 1, pp. 1894-901, 1995

13. N. Stambuk, "On the optimization of complementary protein coding", in: S. Ohno, K. Aoki, M. Usui and E. Uchio(Eds.), *Uveitits Today*, Elsevier, Amsterdam, pp. 315-318, 1998.

14. J. E. Blalock and K. L. Bost, "Binding of peptides that are specified by complementary RNAs", *Biochem. J*., Vol. 234, pp. 679-683, 1986.

15. G. Fasina, P. P. Roller, A. D. Olson, S. S. Thorgeirsson and J. G. Omichinski, "Recognition properties of peptides hydropathically complementary to residues 356-375 of the c-raf protein", *J. Biological Chemistry*, Vol. 264, No. 19, pp. 11252-11257, 1989.

16. S. C. Li and J. Xu, "Digital coding for RNA based on DNA computing", *Chinese Journal of Computer Engineering and Application*, No. 5, pp. 45-47, 2003

17. S. Wolfram, "Cellular automation fluid: Basic theory". *J. Stat. Phys.*, Vol. 45, pp. 471, 1986.

18.R. M. Z. Dos Santos and S. Coutinho, "Dynamic of HIV infection: A cellular automata approach", *Physical Review Letters*, Vol. 87, pp. 168102, 2001.

19. M. Randic, M. Vracko, A. Nandy and S. C. Basak, "On 3-D graphical representation of DNA primary sequences and their numerical characterization", *J. Chem. Inf. Comput. Sci.*, Vol. 40, pp. 1235-1244, 2000.

20. Y. Liu, X. Guo, J. Xu, L. Pan and S. Wang, " Some notes on 2-D graphical representation of DNA sequence", *J. Chem. Inf. Comput. Sci.*, Vol. 42, pp. 529-533, 2002