第 34 卷 第 3 期 计 算 机 工 程 2008 年 2 月
Vol.34 No.3 Computer Engineering February 2008
·博士论文· 文章编号：1000—3428(2008)03—0006—03 文献标识码：A 中图分类号：TN919.81

# 一种新颖的蛋白质序列可视化模型

肖 绚[1]，邵世煌[2]
（1. 景德镇陶瓷学院机电学院，景德镇 333001；2. 东华大学信息学院，上海 200051）

**摘 要：**利用相似规则、互补规则和分子识别理论建立一种氨基酸数字编码模型用于研究序列特征、功能预测。给出一种新的基于元胞自动机的蛋白质序列图像生成方法，其优点是考虑了氨基酸前后的相互作用，生成的图像与基因序列一一对应，许多隐藏在蛋白质序列中的重要特性通过元胞自动机图可以表现出来。基于蛋白质元胞自动机图所得到的蛋白质伪氨基酸成分，蛋白质亚细胞定位预测成功率可以达到 86.4%。

**关键词：**蛋白质序列；可视化；元胞自动机

# New Visualization Model of Protein Sequence

XIAO Xuan[1], SHAO Shi-huang[2]
(1. School of Mechanical & Electronic Engineering, Jingdezhen Ceramic Institute, Jingdezhen 333001;
2. Institute of Information, Donghua University, Shanghai 200051)

**【Abstract】**This paper uses similarity rules, complementary rule and molecular recognition theory to set up a digital coding model of amino acid for investigation into sequence features and their function identification. Cellular automata is used to generate image representation for protein sequences. A protein sequence can be represented by a unique image, and the image considers the interactional actions between amino acids. Many important features hidden in protein sequence can be revealed through its cellular automata image. The protein subcellular location prediction rate reaches 86.4% based on the visualization model.

**【Key words】** protein sequence; visualization; cellular automata

## 1 概述

蛋白质序列是由 20 种氨基酸组成的一维字符序列，要得出更多隐含在其中的生物特性非常困难，为此设计了许多方法，把基因序列转换为数字信号、曲线等，再利用信号处理方法和分形理论等进行研究。其中，可视化技术将符号转换成几何处理，变不可见为可见，使研究者对其研究工作有直观的了解，给予人们新的启示，促进了基因序列的研究。

基因序列可视化研究始于 20 世纪 80 年代，Gates在 1985 年提出了一种DNA二维曲线表示法。4 个单元向量(0,1), (1,0), (0,-1), (-1,0)分别表示碱基T, C, G, A。这种表示法比H curve简单，但它有一个非常大的缺点，就是有高的退化性，序列与二维曲线不是一对一的关系。文献[1]提出另一种基因二维曲线的表示法，它与Gates相比退化性降低了。2003 年Yau进一步提出了没有退化性的表示法[2]。Z曲线是我国张春霆院士提出的，它将DNA序列映射成几何空间的三维曲线图形。Z曲线已被用于真核和原核基因组中若干重要问题研究，包括人与高等真核生物基因组的Isochore结构、微生物基因组的基因水平转移等。

除了 DNA 序列可视化方法，各种蛋白质序列可视化模型也得到发展。例如，文献[3]利用五维空间向量来表示氨基酸，这些向量与氨基酸的化学性质有关，因此，蛋白质的一些结构特点可以通过该方法表现出来。

上述序列可视化方法虽然得出的曲线有二维空间也有多维空间，但其共同点是序列中的某一碱基(或氨基酸)在空间曲线上所对应的点是由这个碱基(或氨基酸)和它之前所有碱基(或氨基酸)共同决定的，而与这个碱基之后的碱基无关。这与基因表达调控的实际情况不符：基因的作用是复杂的，不仅与前面的序列有关，还与后面的序列有关。因此，本文给出了一种新的蛋白质序列可视化方法，利用元胞自动机局部规则，把氨基酸的前后相互作用演化成二维的时空图像，不仅清楚地把不同种类的序列区分开来，且与序列一一对应，为基于氨基酸序列分析的图像处理技术提供了一种途径。

## 2 氨基酸数字编码模型

利用元胞自动机对蛋白质序列进行可视化研究首要解决的问题是建立氨基酸数字编码模型，因为所有分子间的相互作用(如分子识别、分子装配或者任何类型化学键的形成)和分子内部的相互作用(如蛋白质折叠)都受相似规则、互补规则或同时受这两者控制[4]，所以在设计氨基酸数字编码过程中，必须考虑这些规则的影响。

相似规则显示在分子识别过程中，一个已有的成分总是选择与它性质相近的成分结合。它暗示了具有相似性质的个体之间有亲密的关系。相反，互补规则预测具有一些确定的相反性质的个体之间也有亲密关系。从遗传密码中发现，不管是从 5'到 3'方向，还是 3'到 5'方向，具有亲水性和疏水性的氨基酸密码子一般分别被疏水性和亲水性的氨基酸密码子互补对称，而中性(很小的亲水性)的氨基酸的密码子一般被同样中性的氨基酸密码子所对应[5]。

**基金项目：**国家自然科学基金资助项目(60661003)；江西省自然科学基金资助项目(0611060)

**作者简介：**肖 绚(1970－)，男，教授、博士，主研方向：生物信息学；邵世煌，教授、博士生导师

**收稿日期：**2007-05-10 **E-mail：**xiaoxuan0326@yahoo.com.cn

决定同一种氨基酸的各密码子中前 2 个核苷酸往往是相同的，只有第 3 个核苷酸不同，表明密码子的特异性往往由前 2 个核苷酸决定，第 3 个不太重要。而且根据分子识别理论，每对互补对称的氨基酸对前 2 个碱基都是互补的，因此，编码氨基酸时应重点考虑密码子的前 2 位碱基。依据信息理论可以确定，满足氨基酸与编码一一对应的最小二进制数为 5。因此，5 位二进制编码中的前 4 位应由编码氨基酸的密码子的前 2 位碱基决定，第 5 位数由其他 2 个因素决定，它们是：

(1)有相同性质的氨基酸，它们的编码也应相近；

(2)如氨基酸密码子的前 2 位碱基相同，则第 5 位由氨基酸的分子量决定，分子量大的为 1，小的为 0。

因为存在 4 种不同的含氮基强烈暗示核苷酸可以编码成数字{0,1,2,3}，由组合数学可知，这种编码格式有 24 种编码方式的组合。在 4 个数字中，0(00)与 3(11)互补，1(01)与 2(10)互补，而在 4 个碱基中，C 与 G 互补，A 与 U 互补，因此，4 种碱基的数字编码应满足互补法则。本文选择的编码方式为：C=00, U=01, A=10, G=11。原因为：0123∕CUAG 编码可以反映出 4 种碱基的化学性质。在碱基的 2 位二进制数字编码中，首位称为结构编码位。首位为 1 时，编码嘌呤，如 10 为腺嘌呤，11 为鸟嘌呤；当首位为 0 时，编码嘧啶。而末尾数字为功能基团的编码位，当末位为 1 时代表酮基，如 U(01)和 G(11)；而当末位为 0 时则编码氨基基团，如 C(00)和 A(10)。

根据上述的原则得出的编码模型见表 1。数字编码与氨基酸是一一对应的。它既符合氨基酸的物理化学特性，也符合信息理论的要求[6]。

**表 1　氨基酸二进制数字编码模型**

| codon | Amino acid | Binary notation | codon | Amino acid | 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 UUC | 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 | | | |

## 3　基于元胞自动化的蛋白质序列可视化模型

对生物数据进行元胞自动机建模的需求来源于生命系统的复杂性和多样性。生命系统呈现很高的维度，但遗传基因是由 A, C, G, U 这 4 个核苷酸组成的，这些简单分子的高度组合形成了复杂的生命系统，这与元胞自动机有相似之处。本文将一维序列通过元胞自动机合适的演化规则进行时空演化，生成二维蛋白质图像，从而把图像识别技术引入到蛋白质的研究中。

### 3.1　蛋白质序列的二维映射

经过数字编码，氨基酸序列变成了一维的 01 序列。本文采用 Wolfram 元胞自动机，每个格子的状态在某一时刻只能是状态 0, 1 中的一个，演化规则采用简单的 3 点决定 1 点，即元胞下一时刻的状态由本元胞和它最近的 2 个邻居的状态决定。因为许多基因都是环状的，所以采用循环边界条件。

假设蛋白质序列 $S$ 的长度为 $N$，经过数字编码后长度变为 $5N$，将一维序列转化为二维矩阵的迭代公式为

$$\begin{cases} \boldsymbol{D}(i,j)=F(\boldsymbol{D}(i-1,j-1),\boldsymbol{D}(i-1,j),\boldsymbol{D}(i-1,j+1)) & 1<i\le n,\ i<j<5N \\ \boldsymbol{D}(i,1)=F(\boldsymbol{D}(i-1,5N),\boldsymbol{D}(i-1,1),\boldsymbol{D}(i-1,2)) & 1<i\le n \\ \boldsymbol{D}(i,5N)=F(\boldsymbol{D}(i-1,5N-1),\boldsymbol{D}(i-1,5N),\boldsymbol{D}(i-1,1) & i<i\le n \end{cases} \tag{1}$$

其中，$\boldsymbol{D}(i,j)$ 为存放基因序列二维图像数组；$F$ 为演化规则；$n$ 为演化次数；$N$ 为序列长度。

需要说明的是，应用元胞自动机蛋白质序列可视化模型时，演化次数和演化规则可以根据研究问题的不同而改变；演化次数和基因序列的长度成正比，本文要求最小的演化次数应使所产生的基因图像纹理不随演化次数的增加而发生明显改变。演化次数并非越多越好，因为次数增加后，图像也变大了，过大的图像很难寻找特征。一般常用的演化规则有 43, 57, 65, 67, 84, 87, 89, 90, 99, 142, 143, 184, 212, 213, 226 等，如研究乙肝病毒时采用的是 84 号规则，如图 1 所示。

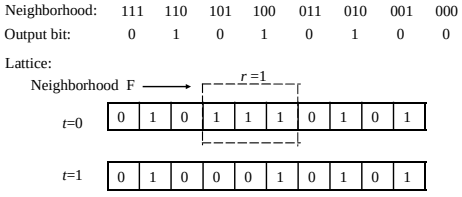


**图 1　元胞自动机 84 号演化规则**

### 3.2　图像生成

为了处理方便，本文采用最基本的位图(BMP)格式。矩阵每个元素的值都是 0 或者 1，因此，图像上的色彩可以用黑白 2 种颜色表示。元素的值为 1，则图像对应该像素点表示黑色，元素的值为 0，该像素点表示白色。

### 3.3　图像压缩

通常由图像缩放而产生的图像中，像素可能在原图像中找不到相应的像素点，这就必须进行近似处理。一般的方法是直接赋值为与它最相近的像素值，也可以通过一些插值算法来计算。后者的效果较好，但是运算量也会相应增加很多。本模型采用直接赋值的方法。

假设图像 $X$ 轴方向缩放比率为 $f_x$，$Y$ 轴方向缩放比率为 $f_y$，那么原图像中的点 $(x_0,y_0)$ 对应于新图形中的点 $(x_1,y_1)$ 的转换矩阵为

$$\begin{bmatrix} x_1 \\ y_1 \end{bmatrix}=\begin{bmatrix} f_x & 0 \\ 0 & f_y \end{bmatrix}\begin{bmatrix} x_0 \\ y_0 \end{bmatrix} \tag{2}$$

## 4　模型特点与应用

本文以一些蛋白质数据图像为例说明元胞自动机图像如何提供有用信息，所用的蛋白质序列都是从美国国家生物技术信息中心(http://www.ncbi.nlm.nih.gov)下载的。对于一序列，演化规则不同，得出的图像也不相同，即基于元胞自动机一个序列可以生成 256 个图像。生成图像所需的演化规则必须能使图像特征很容易地得到，这些图像特征可用于区分蛋白质是否属于同一类的。

本文基因序列可视化模型的特点之一是所生成的图像与蛋白质序列一一对应，前提条件是演化规则和演化次数必须确定不变。这要归功于本文设计的氨基酸编码模型，因为它

保证了不同蛋白质序列的图像至少在第 1 行是不同的。图 2 和图 3 中的序列都是乙肝病毒编号为 ab059661 的 C 基因图，演化规则都为 84 号，演化次数为 300，压缩比为 2:2，编码模型为 Nikola 所设计。2 张图虽然氨基酸序列相同，但核苷酸序列发生了变化，从中可以看出条码的纹理并不相同，得出的图并不能反映两者的蛋白质序列是相同的。

**图 2　乙肝细胞 C 基因图**

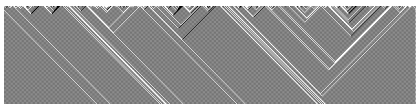

**图 3　乙肝病毒 C 基因图**

但如果用本文的编码系统就不会出现上述情况。图 4 是乙肝病毒序列 2-18 的 P 基因序列比对图。比对图就是对发生突变前后的序列图进行比较，如果相应位上的值相同，比对图中的位置就保持原来的黑白值，如果不同，则用灰色代替。图 4 在第 652 个氨基酸上发生了由丝氨酸(S)到铺氨酸(P)的突变，序列其他位置都保持原样。从图 4 中可以看出，只要有一个位点上发生变化，它的基因图就会改变，图与序列是一一对应的。

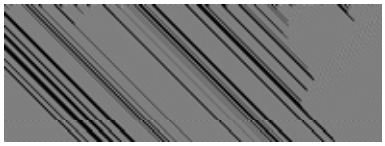

**图 4　乙肝序列 2-18 在 P 基因发生突变的比对图(S652P)**

分子生物中有许多功能相似的基因，表现为同源性。本文对转化生长因子(Transforming Growth Factor-alpha, TGFA)进行了测试。序列包括：人类(AAA61157, AAH05308, AAH05309, CAA49806)，羊(P98135)，Capreolus(AAF73229)，鲔(CAE30382)，恒河猴(P55244)，Mus musculus(AAB50554)，野兔(P98138)，鸡(NP_001001614)，老鼠(NP_036803)和犬类(AAR21186)。图 5 为老鼠 TGFA 基因压缩图，其序列编号为 P0113，长度为 159 氨基酸，压缩比为 2:2，演化次数为 300。图 6 为人类 TGFA 基因压缩图，序列编号为 AAH05308，长度为 159 氨基酸，压缩比为 2:2，演化次数为 300。根据图 5 和图 6，虽然 2 个基因分别来自人和老鼠，但由于功能相似，它们的图像也是非常相似的，即它们在序列上有共同特征，而这很难从它们的字符序列中看出。

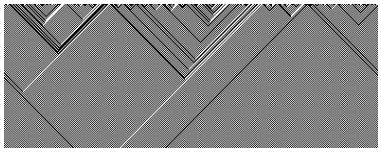

**图 5　老鼠 TGFA 基因压缩图**

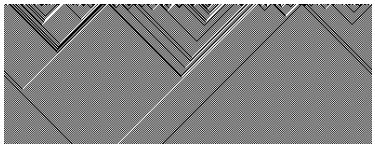

**图 6　人类 TGFA 基因压缩图**

从同一生物体中得到的不同基因也被用于测试本文的模型。TGFA 和 beta-globin major genes 是 2 个具有不同功能的基因。图 7 是老鼠 beta-globin major 基因压缩图，序列编号为 J00413，压缩比为 2:2，演化次数为 300。图 5 和图 7 分别显示了从老鼠身上得到的这 2 个基因的元胞自动机图，比较两者可以看出它们具有明显不同的纹理，从中可以明显区分不同功能的基因。跟据上述分析可以得出，本方法对不同功能基因序列的区分非常有效。

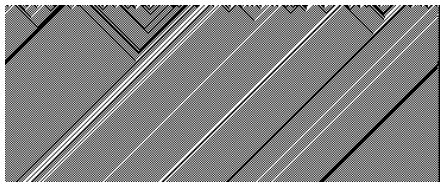

**图 7　老鼠 beta-globin major 基因压缩图**

元胞自动机生成的蛋白质序列图像可用于蛋白质亚细胞定位预测研究。蛋白质亚细胞定位预测是当今分子和细胞生物学的一个研究热点，蛋白质功能与它的亚细胞定位是密切相关的。简单地利用元胞自动机所生成图像中的 01 序列，计算出这些 01 序列的复杂度作为蛋白质伪氨基酸成分，就可以大大提高预测成功率，对同样的数据库进行预测，本方法的预测率在 self-consisency 测试中达到 86.4%，在 jackknife 测试中也达到 72.4%，比基于数字信号处理方法的预测率高出了 5%。

## 5　结束语

可视化是理解复杂现象和大规模数据的重要工具，在生物信息领域中得到了广泛应用。通过可视化技术可以把基因序列变为各种二维和三维的曲线，使各种信号处理技术和图形处理技术得以应用于基因的研究中，为了解基因提供了新方法。本文提出了基于元胞自动机的蛋白质序列图像生成方法，利用元胞自动机处理复杂问题的能力，把前后氨基酸的相互作用用各种演化规则作出不同图像表达，为图像处理技术应用于蛋白质的研究提供了新的途径。

## 参考文献


[1] Guo Xiaofeng, Randic M, Basak S C. A Novel 2-D Graphical Representation of DNA Sequences of Low Degeneracy[J]. Chemical Physics Letters, 2001, 350(1/2): 106-112.

[2] Stephen S, Yau T. DNA Sequence Representation without Degeneracy[J]. Nucleic Acids Research, 2003, 31(12): 3078-3080.

[3] Williams A, Chenault K, Melcher U. Graphic Representations of Amino Acid Sequences[M]//Visualizing Biological Information. River Edge NJ: World Scientific, 1995.

[4] Stambuk N. On the Genetic Origin of Complementary Protein Coding[J]. Croatica Chemica Acta, 1998, 71(3): 573-589.

[5] Bashford J D, Tsohantjis L, Jarvis P D. A Supersymmetric Model for the Evolution of the Genetic Code[C]//Proc. of National Academy of Sciences. Berkeley, CA: [s. n.], 1998.

[6] Xiao Xuan, Shao Shihuang, Ding Yongsheng, et al. Digital Coding for Amino Acid Based on Cellular Automata[C]//Proc. of 2004 IEEE Int'l Conf. on Systems, Man, and Cybernetics. Netherlands: [s. n.], 2004.