Analog-Cell 电子细胞模型中基于并发约束的随机模型构架*)

卢欣华　孙吉贵　行　荣　韩霄松

（吉林大学计算机科学与技术学院　符号计算与知识工程教育部重点实验室　长春 130012）

摘　要　提出了一种在电子细胞模型中基于并发约束的随机模型架构方法，应用该架构建立了模拟基因表达过程的电子细胞模型 Analog-Cell。模拟结果表明 Analog-Cell 这种基于并发约束的随机模型构架能更准确地反映出生物系统的真实性，相比其他电子细胞模型含有更丰富的图像信息，能更清晰地观察细胞内基因表达的全过程，具有良好的应用前景。

关键词　电子细胞，并发约束，随机模型，生物信息学

Architecture of Stochastic Model Based on Concurrent Constraint in Analog-Cell

LU Xin-hua　SUN Ji-gui　XING Rong　HAN Xiao-song

(College of Computer Science and Technology, Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun 130012, China)

Abstract　This paper proposed a new architecture of stochastic model based on concurrent constraint in electronic cell (E-Cell), and built an E-Cell model to simulate gene expression by applying this architecture, Analog-Cell. The simulative results indicate that Analog-Cell can reflect the facility of biologic systems more accurately. Compared with other E-Cell models, Analog-Cell has the more abundant picture information, observes the whole process of gene expression in the cell more clearly and has a good prospect of application.

Keywords　Eelectronic cell (E-Cell), Concurrent constraint, Stochastic model, Bioinformatics

1　引言

生命是由地球上与我们共存的无数生物的基因组决定的，每一种生物的基因组都包含着相应的生物学信息。细胞是所有生物的基本结构单位，自然界的生物体中，尽管构成细胞的化学成分相同，但物质构成并不是生命的本质，它们只是生命的载体。生命体是一个具有自复制和自组织特性的开放式系统，生物的多样性和生物体的适应性都是通过生物系统的精密调控来实现。然而，生物体细胞中化学分子间的作用规则并不能完全准确地诠释出生命的真谛，基因组尽管自身存储了生物的信息，却不能将这些信息释放给细胞。这些生物信息的运用，需要酶和其他蛋白协调作用，参与一系列生化反应，称为基因组表达[1]。

电子细胞（E-Cell, Electronic Cell）[2-4]亦称虚拟细胞，它在计算机上模拟真实细胞的结构、物质组成、生命活动的动力学行为和生命现象，通过数学计算，用虚拟现实的方式实现友好人机交互，以便研究者构造细胞结构和其内外部环境物质组成，考察、记录细胞实验现象和功能，将生物反应过程实现数字化模拟，再现细胞生命活动和发现新的生物学现象规律[5]。

在计算机上模拟基因表达的生物化学反应过程，需要构建一个理想的电子细胞模型，从该模型得出的理论预测能够尽可能准确地反映出生物系统的真实性，并发约束（concurrent constraint）[6]建模思想与生物系统之间有很多结构和功能上的共性。我们建立的 Analog-Cell[7]就是一个以真核细胞为研究对象，生动形象地模拟 DNA 序列转录得到 mRNA，mRNA 翻译得到相应多肽链这一基因表达过程的电子细胞模型。该多肽链是各种酶或其他调控因子的前体蛋白质，它将会最终影响整个细胞的活动。Analog-Cell 基于并发约束建立的随机模型，把电子细胞系统中现有的各分子数量作为系统状态列表，其状态转换真实地模拟了细胞中的生命活动，如酶促反应、基因表达调控等，使得对生物系统建模这一复杂过程大大简化。与其他电子细胞模型[8-11]相比，Analog-Cell 能够更准确地观察细胞内的基因表达过程，这为模拟细胞内其他生命活动、总结和发现生物学的新现象和规律提供了一定的可能性。

2　并发约束

约束是说明性的描述，可以有效地描述目标问题的不完全信息。满足约束条件的变量赋值表达了系统状态或者直接作为实际问题的解，问题求解的过程中可以动态添加约束。约束对于应用领域问题的求解可以主要集中在问题的描述和建模上。所谓约束中的问题建模是指将应用问题的参数用变量表示，问题中各个对象之间较为复杂的相互关系、作用规则用较为基本、抽象的约束来表示。

并发约束是一个多代理的计算模型，彼此独立的计算过程可以通过不同的代理（agent）同时运行。并发约束系统由多个代理和一个共享的约束存储器（constraints store）组成。

*）基金项目：国家自然科学基金重大项目（No. 60496321），国家自然科学基金（No. 60473003），教育部高等学校博士学科点专项科研基金（No. 20050183065），吉林省科技发展计划重大基金项目（No. 20040526），吉林省杰出青年基金资助项目（No. 20030107）。卢欣华　讲师，博士研究生，主要研究人工智能、人工生命、电子细胞；孙吉贵　教授，博士生导师，CCF 理事，研究方向为人工智能、人工生命。

代理之间不能直接进行交互，它们只能和约束存储器之间进行读写交互。约束存储器中存放着目前系统已知的所有约束，也就是对于系统格局的描述。计算过程中可以向存储器中添加一个新出现的约束，也可以通过访问存储器得知某约束是否包含于当前存储器，从而判断当前的系统状态是否满足该约束中所描述的条件。同时，代理对于存储器中信息的更新是单调不可逆的，即：一旦将某相容的约束加入了共享约束存储器中，该约束将始终存在其中。在此之后，其它代理通过访问约束存储器就可以获得该信息。显然，该机制下各代理之间的通信是异步的，通过访问共享约束存储器来实现。代理和约束存储器的分离可以看作计算和逻辑交互功能的分割[6]。

3 Analog-Cell 中基于并发约束的随机模型

并发约束的建模思想与生物系统内复杂而繁多的生物化学反应有很多结构和功能上的共性，随机模型中用系统现有分子的实际数目来描述当前系统的状态，其状态转换真实地反映了细胞内生物化学反应的发生情况，即细胞内的生命活动。Analog-Cell 电子细胞模型模拟了 DNA 序列转录得到 mRNA，mRNA 翻译得到相应多肽链这一基因表达过程，它所采用的基于并发约束的随机模型化方法是将基因表达过程中每一个独立的反应过程作为并发约束中的多个代理；反应过程应遵循的反应规则作为约束，约束存储器中存放着的所有约束即基因表达过程中所有的反应规则。Analog-Cell 中基于并发约束的随机模型整体构架如图 1。

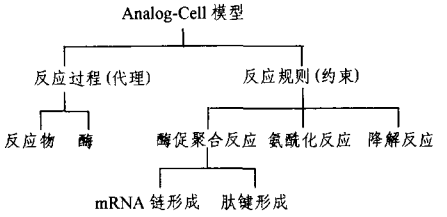

图 1 Analog-Cell 模型架构图

3.1 模型状态定义

模型状态包含用于描述模型在某时刻及未来行为的所有信息。用微分方程描述的生物化学反应，状态是一个包含细胞模型中所有反应物浓度的列表；而随机模型用来描述模型状态的是细胞反应环境中某一时刻各反应物的分子数量列表。分子数量是与反应物浓度相关的，即用离散的方法描述细胞的现有状态，该状态列表的数值变化即状态转换反映了细胞内生物化学反应的进行情况。表 1 列出了我们的模型 Analog-Cell 中假定转录产物 mRNA 序列长度为 180 个核糖核苷酸，几种典型的反应物或酶在基因表达翻译过程完成前后的状态转换。

表 1 Analog-Cell 模型翻译过程完成前后的状态比较

分子数量	氨基酸	ATP	GTP	核糖体	游离 tRNA	延伸因子
翻译前	200	200	200	10	0	10
翻译后	140	39	78	10	60	10

3.2 微分方程描述的生物化学反应

细胞的生命活动通过其内繁多而独立的生物化学反应过程体现，借助光学显微镜我们可以清晰地观察到细胞内的生命活动，电子细胞模型要求可以对细胞内生物化学反应过程实现精确的模拟。生物化学反应过程 $A \xrightarrow{k} A'$ 中，反应物 A 以某个与常数 k 相关的速率转化成生成物 A'。反应过程中各个分子的浓度变化是连续变量，可以用微分方程来描述化学反应，dt 表示时间的微分单元，$[A]$ 表示浓度，则反应中浓度随时间的变化符合如下规律：

$$\frac{d[A]}{dt}=-k[A] \quad \frac{d[A']}{dt}=k[A]$$

类似地，在反应 $A+B \xrightarrow{k} AB$ 中，满足如下规律：

$$\frac{d[AB]}{dt}=k[A][B] \frac{d[A]}{dt}=-k[A][B] \frac{d[B]}{dt}=-k[A][B]$$

如果将时间划分成固定长度的时间间隔，将连续描述离散化，将分子浓度描述为反应环境中分子的实际数目，将浓度变化描述为分子数量列表的变化，就可以得到下面的随机模型。

3.3 随机模型描述的一般生物化学反应

细胞中生物化学反应的一般形式：$n_aA+n_bB \xrightarrow{k} n_cC+n_dD$，表示 n_a 个 A 分子和 n_b 个 B 分子作为底物发生反应，产物为 n_c 个 C 分子和 n_d 个 D 分子。k 是一个与细胞环境温度、反应发生所占用时间相关的参数，表示单位时间内某离散事件发生的概率。化学反应中，反应物和生成物的浓度持续变化从而导致了细胞环境的状态改变，当某些分子浓度变化达到一定的阈值，就可能触发其它化学反应的发生或者终止。上述化学反应式 dt 时间内两种反应物 A，B 作为底物，反应物 A 的分子数目表示为 $\{\#A\}$，反应物 B 的分子数目表示为 $\{\#B\}$，则反应发生即状态转换的概率为 $k\{\#A\}\{\#B\}dt$。

以酶促聚合反应为例，酶促聚合反应是指在酶的催化作用下，多种物质聚合在一起转变为另一种物质，而酶本身没有任何变化[12]。Analog-Cell 电子细胞模型模拟的基因表达过程中，酶促聚合反应主要指转录过程中 mRNA 的合成与翻译过程中多肽链的合成。mRNA 实际上是 4 种核糖核苷酸的线性多聚体，反应由 RNA 聚合酶Ⅱ催化，以 4 种核糖核苷酸作为底物，以 DNA 为模板，每加入一个核苷酸脱去一个焦磷酸 PPi。生物化学反应式可表示为：

$$n_1\mathrm{ATP}+n_2\mathrm{GTP}+n_3\mathrm{CTP}+n_4\mathrm{UTP} \xrightarrow{k} \mathrm{mRNA}+(n_1+n_2+n_3+n_4)\mathrm{PPi}$$

则该反应发生的概率为 $k\{\#\mathrm{ATP}\}\{\#\mathrm{GTP}\}\{\#\mathrm{CTP}\}\{\#\mathrm{UTP}\}dt$。

3.4 随机模型描述的典型反应过程：DNA 与蛋白质的结合

在基因表达过程中普遍认为转录起始是最重要的环节，是前转录复合物(多个蛋白质的复合物)识别并结合 DNA 序列上的核心启动子启动转录开始的过程[1]，它决定特定细胞在特定时间表达哪些基因。因此 DNA 与蛋白质的结合是基因表达中起核心作用的反应过程，是细胞中具有普遍意义的生物化学反应，蛋白质和 DNA 的结合与分离以不同的速率在同一环境下同时发生。Analog-Cell 电子细胞模型模拟的转录起始过程中以 RNA 聚合酶Ⅱ为主体的前转录复合物与 DNA 的结合如图 2 所示。

图 2 Analog-Cell 模拟的前转录复合物与 DNA 的结合

用随机模型描述的 DNA 与蛋白质的结合反应过程：$X+\mathrm{DNA} \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} X\cdot\mathrm{DNA}$，即细胞环境中含有蛋白质分子 X，DNA

以及 X 和 DNA 的复合物 $X\cdot$DNA。用{＃X}表示细胞中蛋白质 X 的分子数目，{＃DNA}表示细胞中 DNA 的分子数目，{＃$X\cdot$DNA}表示 DNA 与蛋白质复合物的分子数目，初始状态表示为({＃X}，{＃DNA}，{＃$X\cdot$DNA})。则后继时刻的状态有 3 种可能：(({＃X}，{＃DNA}，{＃$X\cdot$DNA})，({＃X}－1，{＃DNA}－1，{＃$X\cdot$DNA}＋1)，({＃X}＋1，{＃DNA}＋1，{＃$X\cdot$DNA}－1)。状态转换及反应发生的概率如图 3 所示。

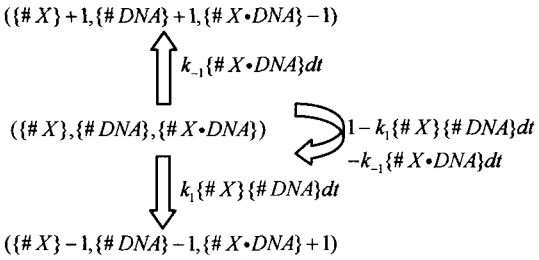

图 3　DNA 与蛋白质结合反应的状态转换及发生概率

用随机模型描述的多种蛋白质分子如两种蛋白质分子和一个 DNA 的结合反应过程：

$$N+\mathrm{DNA}\underset{k_{10}}{\overset{k_{01}}{\rightleftharpoons}}N\cdot\mathrm{DNA};M+\mathrm{DNA}\underset{k_{20}}{\overset{k_{02}}{\rightleftharpoons}}M\cdot\mathrm{DNA}$$

$$M+N\cdot\mathrm{DNA}\underset{k_{31}}{\overset{k_{13}}{\rightleftharpoons}}M\cdot N\cdot\mathrm{DNA}$$

$$\mathrm{A}=\begin{pmatrix}1-nk_{01}\Delta t-mk_{02}\Delta t & k_{10}\Delta t & k_{20}\Delta t & 0\\ nk_{01}\Delta t & 1-k_{10}\Delta t-mk_{13}\Delta t & 0 & k_{31}\Delta t\\ mk_{02}\Delta t & 0 & 1-k_{20}\Delta t-nk_{23}\Delta t & k_{32}\Delta t\\ 0 & mk_{13}\Delta t & nk_{23}\Delta t & 1-k_{31}\Delta t-k_{32}\Delta t\end{pmatrix}$$

Pi 表示系统转换至 i 状态的概率，Δt 是单位时间间隔，下一时刻系统状态转换至 i 的可能性为 $Pi(t+\Delta t)$。

3.5　Analog-Cell 的模拟结果

Analog-Cell 是一个在分子水平上模拟细胞内基因表达过程的电子细胞模型。模型运行以后，用户会看到一条 DNA 模板链在具有棕色核膜的细胞核内游动，细胞中还有游离的核苷酸、核糖体、已连接氨基酸的氨酰 tRNA、各种调控因子[14]等物质。在以 RNA 聚合酶Ⅱ为主体的前转录复合物的作用下，转录过程开始，产物为以 DNA 序列为模板，依据碱基配对原则[1]生成的一条 mRNA 链，如图 5。

图 5　转录过程：mRNA 链正在产生

转录过程完毕后，DNA 和 mRNA 形成的双链分开。mRNA 则游动出细胞核，在细胞核外即胞浆中准备合成多肽链，进行下一步的翻译过程(图 6)。

mRNA 与核糖体(蛋白质的复合体)结合引发翻译过程开始。在起始因子的作用下，核糖体结合在 mRNA 的 5'端并扫描到起始密码子 AUG 处(图 7)。

$$N+M\cdot\mathrm{DNA}\underset{k_{32}}{\overset{k_{23}}{\rightleftharpoons}}N\cdot M\cdot\mathrm{DNA}$$

即细胞环境中含有一个 DNA，m 个蛋白质分子 M 和 n 个蛋白质分子 N，蛋白质 M 和 N 都可以与 DNA 结合，组成复合物 $M\cdot$DNA，$N\cdot$DNA 或 $M\cdot N\cdot$DNA。该结合反应有 4 种可能存在的状态，其状态转换如图 4[13]所示。

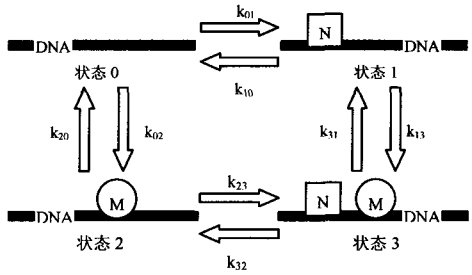

图 4　两种蛋白质与 DNA 结合反应的状态转换

状态转换发生的概率可通过下列公式[13]计算：

$$\begin{pmatrix}P_0(t+\Delta t)\\ P_1(t+\Delta t)\\ P_2(t+\Delta t)\\ P_3(t+\Delta t)\end{pmatrix}=A\cdot\begin{pmatrix}P_0(t)\\ P_1(t)\\ P_2(t)\\ P_3(t)\end{pmatrix}$$

其中，

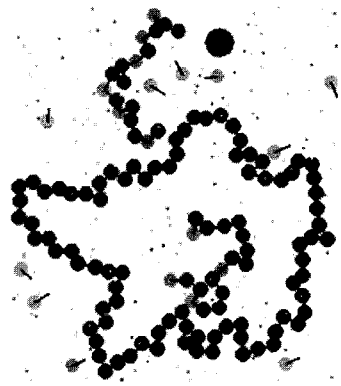

图 6　DNA 与 mRNA 已分开，mRNA 游出细胞核

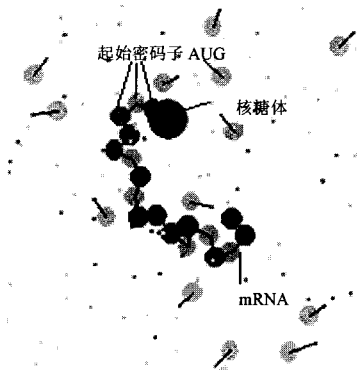

图 7　核糖体与 mRNA 结合

起始密码子 AUG 对应甲硫氨酸 Met，携带着 Met 的 tRNA 与 mRNA 结合，多肽链上的第一个氨基酸产生。接着核

糖体移动三个位置准备读取下一个密码子(图 8)。

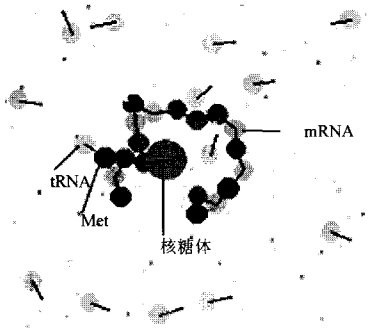

图 8 多肽链上的第一个氨基酸 Met 产生

随着核糖体的移动,密码子依次被读取,氨基酸被氨酰 tRNA 送到 mRNA 链上,这些氨基酸通过肽键连接在一起形成多肽链,同时已连接到多肽链上的氨基酸与 tRNA 脱离。这一延伸过程持续下去,直到核糖体移动到终止密码子处,释放因子发出终止信号,使核糖体、多肽链、tRNA 等物质都与 mRNA 脱离。翻译过程模拟结束(图 9)。

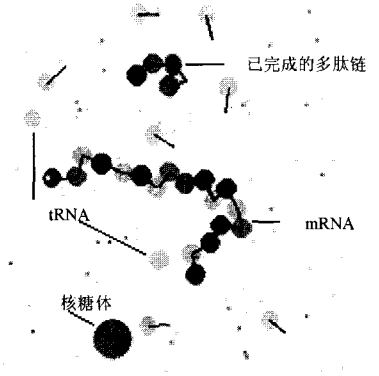

图 9 多肽链形成,翻译过程结束

模拟的最后一步是 mRNA 的降解,即翻译成多肽链的 mRNA 被核糖核酸酶识别并降解,重新变成游离的核苷酸(图 10)。至此,一次完整的基因表达过程模拟结束。

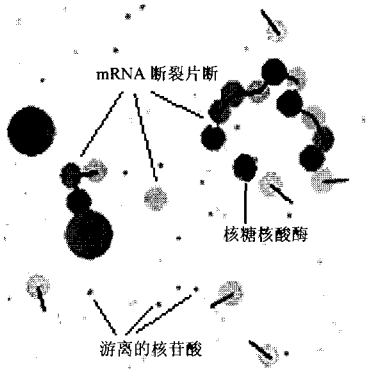

图 10 mRNA 被核糖核酸酶降解

结束语 为了在计算机上模拟基因表达的生物化学反应过程,需要构建能够尽可能准确地反映出生物系统真实性的电子细胞模型。并发约束的建模思想提高了用计算机解决生物系统模拟的能力;随机模型把细胞中各分子数量作为系统状态列表,其状态转换真实地模拟了细胞中的生命活动。Analog-Cell 基于并发约束思想建立了随机模型,用图形显示方式模拟了细胞内与基因表达相关的生物化学过程。相比其他电子细胞模型,Analog-Cell 含有更丰富的图像信息,能够更清晰地观察细胞内基因表达的全过程,并且完全符合基因表达的生物学原理,这为模拟细胞内其他生命活动、总结和发现生物学的新现象和规律提供了一定的可能性。

参 考 文 献

[1] Brown T A. Genomes 2[M]. Oxford, England: Bios Scientific Publishers Ltd, 2002

[2] Stephen H M. Exceeding human limits[J]. Nature, 2006, 440(7083): 409-410

[3] Emonet T, Charles M M, Michael J N, et al. AgentCell: a digital single-cell assay for bacterial chemotaxis[J]. Bioinformatics, 2005, 21(11): 2714-2721

[4] Tomita M. Whole-cell simulation: a grand challenge of the 21st century[J]. TRENDS in Biotechnology, 2001, 19(6): 205-210

[5] 赵明生,尚彤,孙冬泳,等. 电子细胞的研究现状与展望[J]. 电子学报, 2001, 29(12A): 1740-1743

[6] Valencia F. Decidability of Infinite-State Timed CC PProcess and First-Order LTL[J]. Theory Computer Science, 2005, 330(3): 577-607

[7] 卢欣华,孙吉贵. Analog-Cell: 一种新的电子细胞图形模型[J]. 电子学报, 2007, 35(1): 49-53

[8] Ishii N, Robert M, Nakayama Y, et al. Toward large-scale modeling of the microbial cell for computer simulation[J]. Journal of Biotechnology, 2004, 113: 281-294

[9] Slepchenko B M, Schaff J C, Macara I G, et al. Quantitative cell biology with the Virtual Cell[J]. Trends in Cell Biology, 2003, 13(11): 570-576

[10] Eungdamrong N J, Iyengan R. Compartment-specific feedback loo Pand regulated trafficking can result in sustained activation of Ras at the Golgi[J]. Biophys. J, 2007, 92(3): 808-815

[11] Broderick G, Ru'aini M, Chan E, et al. A life-like virtual cell membrane using discrete automata[OL]. http://www.bioinfo.de/isb/2004/05/0016/, 2004-11-27/2006-05-30

[12] 王镜岩,朱圣庚,徐长法. 生物化学(下册)[M]. 第三版. 北京: 高等教育出版社, 2002

[13] Gibson M A, Mjolsness E. Modeling the Activity of Single Genes[M]. California: MIT Press, 1999

[14] Lu Xin-hua, Sun Ji-gui, Ren Ying, et al. The Regulation of Gene Expression in E-Cell[C] // Proceedings of Third International Conference on Natural Computation (ICNC 2007). Volume Ⅲ. Los Alamitos: IEEE Computer Society Press, 2007