

Analog-Cell: 一种新的电子细胞图形模型

卢欣华^{1,2}, 孙吉贵^{1,2}

(1. 吉林大学计算机科学与技术学院, 吉林长春 130012; 2. 符号计算与知识工程教育部重点实验室, 吉林长春 130012)

摘要: 本文提出了一种新的模拟基因表达过程的电子细胞图形模型 Analog Cell, 给出了实现该模型设计的主要算法和模拟结果. 与其它电子细胞模型相比, Analog Cell 特点在于含有丰富的图像信息, 并在分子水平上模拟了基因突变.

关键词: 生物信息学; 电子细胞; 人工生命

中图分类号: Q811.4 **文献标识码:** A **文章编号:** 0372-2112(2007)01-0049-05

Analog-Cell: A New Pattern Model of E-Cell

LU Xinhua^{1,2}, SUN Jigui^{1,2}

(1. College of Computer Science and Technology, Jilin University, Changchun, Jilin 130012, China;

2. Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun, Jilin 130012, China)

Abstract: This paper proposes a new pattern model of electronic cell (E-Cell) to simulate gene expression, Analog-Cell, provides major algorithms to implement the design and shows the simulative results of the model. Compared with other E-Cell models, the characteristic of Analog-Cell lies in having abundant picture information and simulating the gene mutation on the molecule level.

Key words: bioinformatics; electronic cell(E-Cell); artificial life

1 引言

电子细胞(E-Cell, Electronic Cell)^[1-3]亦称虚拟细胞(Virtual Cell), 是人工生命的重要研究内容之一. 它在计算机上构建一个虚拟的细胞体系, 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, 以便研究者考察、记录细胞实验现象及功能, 再现细胞生命活动和发现新的生物学现象规律^[4].

目前有三种较成熟的电子细胞模型^[4-5]. 最早的是日本Keio大学Tomita教授领导的研究小组所建立的原核细胞能量代谢、细胞膜物质运输模型E-CELL^[3]. 另一个较成熟的模型是美国康涅狄格州州立大学Leslie M Loew和James C Schaff等学者所建立的真核细胞钙转运模型Virtual Cell^[6], 通过简单的图形模拟钙离子流的移动过程. 第三个较成熟的模型是由美国印第安那大学化学系的Peter J Ortoleva教授等人开发的模拟细胞对外界刺激反应、物质运输及基因组变化的虚拟细胞模型CyberCell^[7]. 本文对这三种模型不作过多介绍, 可参阅文献^[3-8].

这三种电子细胞模型存在一个不足: 图形显示方面较为简单、不能满足充分观察细胞内生物化学反应过程的需要. 本文提出一种新的电子细胞图形模型Analog-Cell, 它以真核细

胞为研究对象, 以生动、形象的图形模拟细胞内基因表达^[9]的生物化学过程. Analog-Cell的特点一方面在于含有更丰富信息的细胞活动影像, 另一方面在于模拟了其它模型所没有的基因突变.

2 Analog-Cell的细胞建模设计

Analog-Cell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, 该多肽链是各种酶或其它调控因子的前体蛋白质. 根据基因表达的生物学原理^[9, 10],

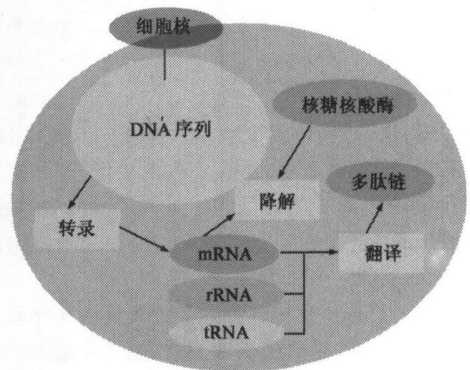


图1 Analog-Cell的三个反应过程

收稿日期: 2006-05-30; 修回日期: 2006-10-31

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(No. 20040526)

我们定义 Analog Cell 模拟的基因表达主要由顺序执行的三个过程组成: 转录、翻译、降解(图 1)。图中椭圆形表示细胞中存在或产生的物质, 方形代表三个反应过程, 箭头指明参与反应的物质及反应的方向。

我们选择了细胞内与基因表达相关的物质如 DNA、核苷酸、mRNA、核糖体、tRNA、氨基酸、多肽链及核糖核酸酶等建立了模型, 采用的模型化方法是将细胞内基因表达涉及的物质分为三类: 反应物、基因序列、反应规则。其中反应规则是把反应物、基因序列整合到一起的模型的核心。模型的整体架构如图 2。

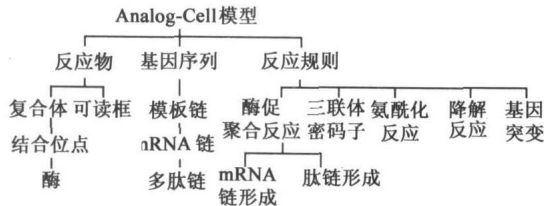


图 2 Analog-Cell 模型总体结构图

2.1 反应物

为了清楚地区分参与不同反应过程的不同反应物, 我们定义了类型标识、颜色、显示半径等属性。例如: 丙氨酸独立存在时颜色为淡紫色 (RGB: 207, 136, 255), 显示半径为 1 像素; 翻译中与其它氨基酸连接形成多肽链后颜色变为紫色 (RGB: 153, 0, 255), 显示半径变为 6 像素。而核糖体、tRNA 等反应物因为只参与一个反应过程, 所以它们的颜色、显示半径等属性不发生变化。类型标识是反应物之间加以区别的标识, 取名多数来自于生物学中物质名称的缩写, 如丙氨酸的类型标识为 'A'。

反应物有几个重要元素: 可读框^[9]、复合体、结合位点、酶。可读框是编码蛋白的基因中, 从三联体密码子的起始密码子开始到终止密码子为止的连续核苷酸密码序列。参与反应的各种反应物有时需要组合成复合体。例如: 转录过程以 DNA 为模板, 按照碱基互补配对原则^[9], 核糖核苷酸逐渐连接到 mRNA 上, 生成 DNA-RNA 复合体。复合体在结合位点由反应物合成, 且一般应在特定酶或因子的催化作用下。例如: DNA-RNA 复合体以基因上游作为结合位点, 由 RNA 聚合酶催化。

2.2 基因序列

基因序列是模型中的关键物质, 它决定了 mRNA 序列, 而多肽链的氨基酸序列又为 mRNA 序列所决定, 多肽链根据自身的氨基酸序列折叠成三维结构才有了蛋白质的活性。基因序列通过 mRNA 把遗传信息传给蛋白质, 从而决定影响细胞活动的蛋白质的结构和功能。

Analog Cell 选用染色体内若干个脱氧核糖核苷酸组成的一个基因序列作为转录的模板链。真核生物细胞内一个基因序列要比原核细胞中的基因序列长得多, 这里只是为了模拟演示而假拟了一条较短的 DNA 序列, 但已经包含了起始密码子和终止密码子, 保证了即使较长的基因序列模型也能够正确进行模拟, 只是随着基因序列的延长, 模拟的时间也会相应增加。Analog Cell 模拟一次基因表达的时间根据外界环境是否

满足反应发生的条件, 大约在 10~30 分钟左右, 比起真实细胞相同基因序列完成一次基因表达的时间要长得多^[9]。实际上, 模型的执行速度作为一个参数是完全可调的, 这里为了方便研究者对基因表达过程进行细致全面地观察而调慢了反应速度。

2.3 反应规则

反应规则是 Analog-Cell 建模设计的核心, 分为两部分: 反应过程与该反应过程中进行的状态控制。

状态控制通过对反应过程中反应物状态值的改变, 来控制反应过程应在何时进行、识别何种反应物或复合体来参与何种反应。基因表达的三个过程必须是按照先转录、再翻译、最后降解的顺序进行; 有些反应过程又必须是某种特定反应物参与该反应。例如: 降解反应要降解已翻译完毕的 mRNA, 那么核糖核酸酶必须识别该 mRNA, 避免把刚转录产生的、还未翻译的 mRNA 降解。表 1 列出了模型中的主要反应过程、反应物的状态及反应发生的条件。

表 1 反应规则

反应过程	反应物或复合体	反应物所需状态值	反应发生条件
转录—mRNA 合成	DNA 序列	1	核内存在 DNA 序列;
	mRNA	12	基因上游与核膜结合;
	核糖核苷酸 细胞核核膜	0 -4	核内存在游离的核糖核苷酸。
翻译—多肽链合成	氨基酸	0	核外存在 mRNA;
	氨酰 tRNA	-2	核糖体与 mRNA 5' 末端结合;
	核糖体	-5	核外存在氨酰 tRNA。
	多肽链 mRNA	22 16	
氨酰化反应	tRNA	-3	核外存在游离的 tRNA;
	氨基酸	0	核外存在游离的氨基酸。
三联体密码子的识别	mRNA	18, 31~38	核外存在氨酰 tRNA;
	氨酰 tRNA	-2	核外存在 mRNA;
	核糖体	-5, -6, -7	核糖体已移动三个位置。
降解反应	核糖核酸酶	-1	核外存在核糖核酸酶;
	mRNA	20	核外存在翻译完毕的 mRNA。
基因突变	DNA 序列	1	核内存在 DNA 序列;
	脱氧核糖核苷酸	0	核内存在游离的脱氧核糖核苷酸。

反应物状态用整数表示, 反应发生条件通过反应物所需状态值进行控制, 同一种反应物可能有多种不同的状态值。例如: 转录中正在合成的 mRNA 状态为 12; 转录完毕状态为 16; 翻译完毕被降解时状态为 20。多数酶或因子的催化作用也由状态控制来完成。例如: 氨酰化反应在氨酰 tRNA 合成酶的催化下把核外游离的 tRNA 和氨基酸连接在一起, 模型定义游离的 tRNA 状态为 -3, 连接氨基酸后 tRNA 状态改为 -2。

3 反应过程和状态控制机制

3.1 转录——mRNA 的合成

mRNA 实际上是核糖核苷酸以 3', 5' - 磷酸二酯键连接起来的线性多聚体, 由 RNA 聚合酶催化合成^[10]。我们针对 mRNA 合成反应定义的状态控制见表 2。

表 2 转录过程状态定义

反应物或复合体	状态值	状态值含义
脱氧核糖核苷酸	1	未转录的核苷酸.
	2~5	转录中所需的中间状态.
	6, 7	转录完毕, DNA 与 mRNA 双链分开时的中间状态.
	8, 14	使 DNA 最后一个核苷酸正常转录的中间状态.
	9	DNA 第一个核苷酸结合核膜前.
	10	DNA 第一个核苷酸结合核膜后.
	13	DNA 最后一个核苷酸的初始状态.
核糖核苷酸	15	DNA 最后一个核苷酸转录完毕, 转录结束标识.
	0	游离的核糖核苷酸.
细胞核核膜	12	新合成的 mRNA 上核苷酸.
	-4	不参与反应的核膜.
	10	mRNA 结合核膜后, 核膜的初始状态.
mRNA	11	核膜结合 mRNA 所需的中间状态.
	16	转录完毕、翻译未开始时的 mRNA.

参照此状态定义表, 利用状态控制并依据转录的生物学原理, 我们设计了模拟转录过程的算法如下:

算法 1 转录过程及状态控制算法

1. 初始化 DNA 序列的核苷酸、核膜以及游离的核苷酸的状态;
2. 转录的起始:
 - (1) 若反应范围内存在 RNA 聚合酶:

则 DNA 第一个核苷酸状态更新为结合核膜前的状态;
 - (2) 若 DNA 第一个核苷酸类型为 't' (起始密码子第一位), 且状态为结合核膜前的状态:

则该核苷酸与核膜上任意一点结合, 且状态更新为结合核膜后的状态;
3. DNA 第一个核苷酸转录:

结合核膜后的 't' 与游离的 'a' 连接('a' 与 't' 配对), 且 't' 的状态更新为转录后中间状态, 'a' 的状态更新为新合成的核糖核苷酸的状态;
4. 循环, 直到转录到 DNA 最后一个核苷酸:

把游离的核苷酸按照碱基互补配对原则与新合成的核糖核苷酸连接;
5. DNA 最后一个核苷酸转录:

DNA 最后一个核苷酸 't' (终止密码子最后一位) 与游离的核苷酸 'a' 连接, 且 't' 的状态更新为转录结束的状态, 'a' 的状态更新为 DNA 与 mRNA 分开的状态;
6. mRNA 与 DNA 的分开:
 - (1) 转录结束状态的核苷酸与核膜连接, 状态更新为分开状态;
 - (2) 循环, 直到不存在分开状态的核苷酸:

断开分开状态两点之间的连接;
7. mRNA 所有核苷酸的状态更新为转录完毕的状态;

以下的模拟示例解释算法中的状态控制机制. 例如: 第 3 步

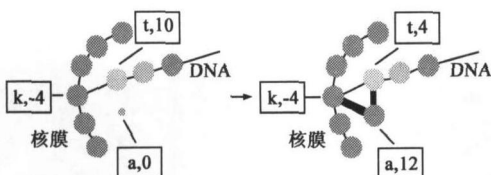


图 3 DNA 第一个核苷酸的转录

DNA 序列第一个核苷酸的转录如图 3. 游离的核苷酸与 DNA 第一个核苷酸、核膜结合后, 两个核苷酸的状态都发生了改变.

3.2 翻译——多肽链的合成

多肽链实际上是 20 种氨基酸以肽键连接起来的线性多聚体, 由肽基转移酶催化合成^[10]. 我们针对多肽链合成反应定义的状态控制见表 3.

表 3 翻译过程状态定义

反应物或复合体	状态值	状态值含义	
tRNA	-2	已携带氨基酸.	
	-3	未携带氨基酸.	
核糖体	-5	游离的核糖体;	
	-6, -7	结合 mRNA 后, 不需要移动的核糖体.	
	16	还未识别的三联体密码子.	
	17	起始密码子第一位标识, 核糖体识别并结合于该位点.	
	18	与核糖体结合后的三联体密码子第一位.	
	19	多肽链沿着 mRNA 移动的标识.	
	21	识别后的密码子第一位, 还未形成多肽链.	
	23	识别后的密码子第二位, 还未形成多肽链.	
	26	识别后的密码子第二位, 肽链已形成, 核糖体开始移动三位.	
	27, 28	核糖体移动三位时, 核苷酸的中间状态.	
核糖核苷酸	29	终止密码子的状态标识.	
	31~38	依赖第三位决定的密码子识别时的中间状态.	
	mRNA	20	mRNA 的最终状态;
		30	核糖核酸酶识别并降解 mRNA 的标识.
氨基酸	22	翻译完毕时 mRNA 的中间状态, 多肽链继续移动的标识.	
	0	已连接 tRNA.	
	22	已连接 mRNA, 还未形成多肽链.	
多肽链	24	刚刚与 tRNA 脱离.	
	25	多肽链的最终状态.	
	39	肽链合成所需的中间状态.	

参照此状态定义表, 利用状态控制并依据翻译的生物学原理, 我们设计了模拟翻译过程的算法如下:

算法 2 翻译过程及状态控制算法

1. 若反应范围内存在起始因子 eIF-3:

则 (1) mRNA 起始密码子第一位 'A' 的状态更新为供核糖体识别的状态;

(2) 核糖体 'r' 与识别出来的 'A' 连接, 且 'A' 的状态更新为与核糖体结合后的状态;
2. 循环, 直到翻译到终止密码子:
 - (1) 调用三联体密码子识别算法, 找到密码子对应的氨基酸, 并连接到 mRNA 上;
 - (2) 若存在两个未连接在一起的氨基酸:

则两个氨基酸之间形成肽键;
 - (3) 先进入 'r' 的氨基酸与连接的 tRNA 'n' 断开连接, 且状态更新为与 'n' 脱离后的状态;
 - (4) 该氨基酸再与 mRNA 断开连接, 且状态更新为翻译完毕多肽链的最终状态;

- (5) r' 沿着 mRNA 向 3' 末端移动三个位置, 读取下一个密码子;
- (6) 多肽链随着 r' 移动三位;
- 3. 若 r' 移动到终止密码子处:
则 r' 与 mRNA 断开连接;
- 4. 若多肽链移动到终止密码子处:
则多肽链与 mRNA 断开连接;
- 5. 更新 mRNA 所有核苷酸状态为翻译完毕 mRNA 的最终状态;

以下的模拟示例解释算法中的状态控制机制. 例如: 第 1 步核糖体与 mRNA 上起始密码子 AUG 的结合如图 4. mRNA 上起始密码子的第一个核苷酸状态首先改变供核糖体识别, 与核糖体结合后状态又发生了改变.

翻译算法内部调用的三联体密码子识别算法是我们根据遗传密码表^[9, 10]中遗传密码对应氨基酸的规律, 利用状态控制机制设计完成的. 该算法以三位密码子的类型及状态作为识别依据, 根据遗传密码表找到这三位密码子对应的氨基酸, 把识别出来的氨基酸连接到 mRNA 上, 并且改变其状态值以满足肽键形成的条件. 限于篇幅在本文中不给出此算法的具体内容.

3.3 基因突变

基因组的进化是由突变引起的小规模序列改变与重组产生的大规模重排的累积结果. 许多突变是单个核苷酸被替换的点突变, 其它形式包括核苷酸的插入或缺失. 突变发生的几率在细胞中往往是很低的^[9].

Analog Cell 是分子水平上的模拟, 即 DNA 由多个核苷酸组成, 为模拟基因突变提供了条件. 针对基因突变中比较典型的点突变, 我们设计了基因点突变模拟算法. 该算法利用随机数是否能整除一个较大的整数来控制突变发生的几率, 符合生物学上基因突变发生几率很低的客观事实. 图 5 解释了算法的具体执行过程并给出算法的模拟结果. DNA 上的核苷酸 x 被反应范围内的核苷酸 y 替换. Analog Cell 使用类似的方法也模拟了其他类型的基因突变, 如插入和缺失, 限于篇幅在本文中不给出这些算法的具体内容.

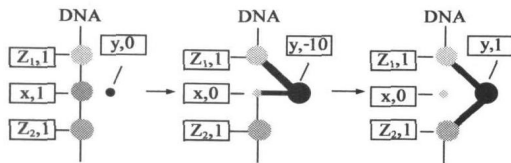


图 5 基因点突变: x 被 y 替换

4 Analog Cell 的模拟结果

Analog Cell 采用 C++ 面向对象的方法实现计算机上对基因表达过程的模拟, 并参照了 Tim J Hutton 的 Squim3 程序

面^[11]. 模型运行以后, 可以看到一条 DNA 模板链在具有核膜的细胞核内游动, 核外还有核糖体、氨酰 tRNA、核糖核酸酶等物质. 当基因上游结合到核膜上时转录开始, 依据碱基互补配对原则, 以 DNA 序列为模板, mRNA 链逐渐产生于细胞核内(图 6).

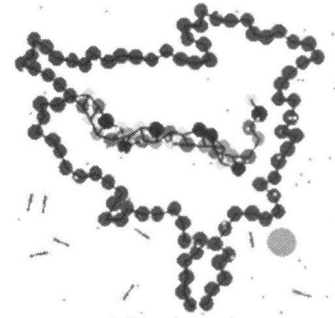


图 6 转录过程: mRNA 链正在产生

转录完毕后, DNA-mRNA 双链分开, mRNA 游动出细胞核, 在胞浆中与核糖体结合引发翻译的产生. 随着核糖体的移动, 依据遗传密码识别出来的氨基酸通过肽键连接成多肽链, 直到核糖体移动到终止密码子处为止(图 7).

核糖体识别出 mRNA 3' 末端的终止密码子, 核糖体、多肽链、tRNA 都与 mRNA 脱离. 这时核外可以看到完整的多肽链、游离的 tRNA、独立的核糖体等物质(图 8). 翻译过程模拟结束.

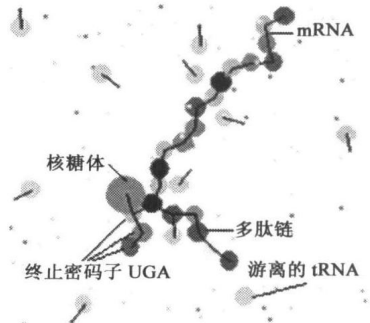


图 7 翻译过程: 核糖体移动到终止密码子处

当核糖核酸酶遇到翻译完毕的 mRNA 时, 会导致 mRNA 的断裂, 断裂片断继续被降解, 直到重新变成游离的核苷酸(图 9). 至此, 一次完整的基因表达过程模拟结束.

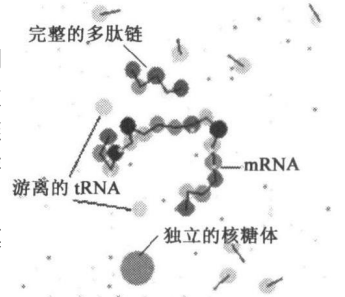


图 8 多肽链形成, 翻译过程结束

5 结束语

Analog Cell 是以真核细胞为研究对象, 用图形显示方式模拟细胞内与基因表达相关的生物化学过程的电子细胞模型. 它具有以下几个特点: (1) Analog Cell 是国内第一个电子细胞模型, 到目前为止还没有查到有电子细胞模型发表; (2) Analog Cell 包含丰富的图像信息, 用户会直观、形象地观察到 DNA 表达为多肽链

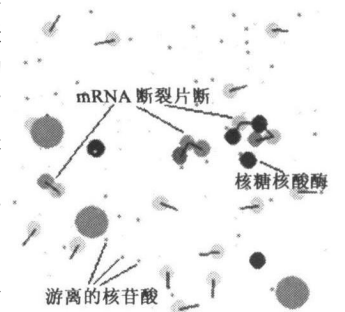


图 9 mRNA 的降解

的基因表达过程; (3) Analog Cell 在分子水平上模拟了基因突变, 而基因突变是生物进化的重要手段之一。但是模型也存在一些不足, 比如: 目前模型的设计要实现三维可视化存在一定难度; 对于基因表达调控模拟得不够全面等等, 这都是我们进一步的研究工作。

参考文献:

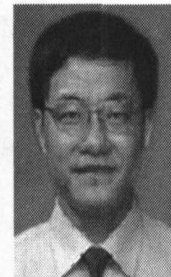
- [1] Stephen H Muggleton. Exceeding human limits[J]. Nature, 2006, 440(7083): 409- 410.
- [2] Thierry Emonet, Charles M Macal, Michael J North, et al. AgentCell: a digital single cell assay for bacterial chemotaxis[J]. Bioinformatics, 2005, 21(11): 2714- 2721.
- [3] Nobuyoshi Ishii, Martin Robert, Yoichi Nakayama, et al. Toward large scale modeling of the microbial cell for computer simulation[J]. Journal of Biotechnology, 2004, 113: 281- 294.
- [4] 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.
Zhao Ming sheng, Shang Tong, Sun Dong yong, et al. The current state and prospect of E-Cell[J]. Acta Electronica Sinica, 2001, 29(12A): 1740- 1743. (in Chinese)
- [5] 陈源, 李朝军. 虚拟细胞[J]. 细胞生物学杂志, 2004, 26(3): 231- 234.
Chen Yuan, Li Chao jun. Virtual cell[J]. Chinese Journal of Cell Biology, 2004, 26(3): 231- 234. (in Chinese)
- [6] Slepchenko BM, Schaff JC, Macara IG, et al. Quantitative cell biology with the Virtual Cell[J]. Trends in Cell Biology, 2003, 13(11): 570- 576.
- [7] Gordon Broderick, Melania Ru' aini, Eugene Chan, 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.

- [8] 卢欣华, 孙吉贵. 模拟细胞生命活动的电子细胞[J]. 计算机科学, 2006, 33(8. 增刊): 315- 318.
Lu Xirr hua, Sun Ji gui. E-Cell to simulate the life activity of cells[J]. Computer Science, 2006, 33(8. Suppl): 315- 318. (in Chinese)
- [9] T A Brown. Genomes[M]. Oxford, England: BIOS Scientific Publishers Ltd, 1999.
- [10] 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.
- [11] Tim J Hutton. 集智俱乐部 Squirml3 源程序[OL]. <http://www.swarmagents.com/thesis/detail.asp?id=143>, 2004 04 16/2006 05 30.

作者简介:



卢欣华 女, 1977年2月生于吉林长春, 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博士研究生。研究方向: 人工智能、人工生命、电子细胞。
E-mail: luxh@jlu.edu.cn



孙吉贵 男, 1962年11月生于辽宁庄河, 教授、博士生导师、吉林大学计算机科学与技术学院副院长、符号计算与知识工程教育部重点实验室主任。研究方向: 人工智能、人工生命。