基于符号化进化动力学的基因组数据采掘①

刘健勤
(中南工业大学信息工程学院, 长沙 410083)

摘 要 提出了一种新的基因组数据模型和模式发现算法。该模型由人工基因组、人工蛋白、进化操作、进化控制、模式匹配、终止判断 6 个环节组成, 其中抽象代数结构由格集合构形和相应有限状态机操作来动态描述, 候选符号序列由符号动力学引导的进化算法所生成, 进化程度由粗糙集所刻划的元进化机制所控制, 模式匹配由句法模式识别器和文法推断过程所完成, 终止判断依具体问题求解的约束条件而定。相应的算法为循环性的群体隐式并行搜索, 数据结构以答号化粗粒度的处理为主, 并与面向语义的模块化程序设计相配合。在该人工生命技术的应用中, 由计算机自动生成了候选符号序列, 从中获得了“真实”的氨基酸序列。实验结果表明, 所提出并实现的计算方法有助于基因组学层次下的生物信息学的统一计算理论的建立和应用系统开发。

关键词 基因组学 生物信息学 进化计算
中图法分类号 TP11

人工生命是通过人工方式和相应技术手段去认识、建模、分析、构造具有抽象意义下生命形式化特征的机器系统的科学, 属于一门新兴的交叉性前沿学科。从计算机科学和技术的角度看, 它具有 NP 问题求解的功能; 就生命科学而言, 它又是一种有力的工具。基因组数据采掘就是一个重要的方面, 在信息、生物、材料这三个领域的交叉点上, 运用现代分子生物学的技术开发新材料的工作就有赖于基因组学的手段 (例如基因调控)。Elghanian 等人采用 DNA 的手段设计纳米级材料结构的工作就是引人瞩目的最新进展之一 (其成果发表在 1997 年 8 月的《Science》上[1, 2])。基因组数据采掘由于其分子层次上的信息特点, 在包含有色金属在内的生物冶金、生物材料等行业中具有较为广阔的应用前景; 同时根据生物基因在同源性、保守性以及模式生物特性方面的联系, 基因组数据采掘具有模式发现的机制, 有助于蛋白质序列的计算机辅助分析与预测。

作为非结构化、非确定性信息处理的计算手段, 数据采掘已被应用于相当广泛的领域。其应用类型主要分为下列 3 个方面。

(1) 面向“科学发现”的计算机系统: 其核心就是 Langley 等所强调的“创造性过程的计算探索”[3], 例如用于基因识别的 GRAIL 系统[4]。

(2) CSDE (计算支持下的发现环境) 工具: 这是一种功能强、效率高的智能信息处理工具, 例如 Jong 和 Rip 所开发的用于大肠杆菌 E. coli 遗传调节机制分析的 CSDE 系统[5]。

(3) 基于代理体 (agent) 结构的过程: 这种计算过程是与具体应用对象相联系的, 如 Anand 等关于数据采掘的内核设计技术[6]就是设计多代理体系统的有效方式之一。

数据采掘是指在缺乏先验知识的情况下, 从包含大量数据的集合中“挖掘”出有用的数据, 从而获得有用的信息, 这种信息往往是传统技术难以获得的。经过数据采掘所求得的信

① 前国家教委留学回国人员科研启动基金、前国家教委国家信息处理与智能控制开放实验室基金、湖南省自然科学基金、前中国有色金属工业总公司“十百千人才”基金资助项目 收稿日期: 1998-04-27; 修回日期: 1998-07-13
刘健勤, 男, 34 岁, 博士, 副教授

息包括模式、信号、知识、符号等形式。数据采掘主要采取两种形式[7]。

(1) 校验驱动的数据采掘: 在用户所给的假设基础上, 通过非确定信息处理而完成有关语义的验证, 获取正确信息, 消除噪声或有错误的假设。这里需涉及随机过程建模理论和统计信号分析技术。

(2) 发现驱动的数据采掘: 在已有的数据基础上进行有效的操作, 以自动地抽取出所需的信息。为了实现这个过程可采用粗糙逻辑、非单调推理等途径, 关键是需要模拟人脑思维的创造性机理, 这样发现驱动的数据采掘就与机器学习和计算机辅助软件工程相联系。

综上所述, 校验驱动属于自上而下的概念驱动行为, 发现驱动则可归入自底向上的数据驱动行为, 在设计实际应用系统时将这两者加以集成, 可以提高系统效率, 同时在进行复杂模式的非线性信息处理时, 该集成方式下的数据采掘技术迫切需要构造新的形式化理论规范。数据采掘具有多学科交叉的突出特点, 其内容涉及数学、理论计算机科学、数据库技术、计算机系统开发、认知科学, 其核心是粒度化信息处理, 并且多与自适应知识处理相关。

基因组数据采掘是针对基因组相关信息的处理, 以获取必要的知识性结论的过程。人类基因组计划 (HGP) 等重大国际合作科研项目的开展, 使得认识基因组学层次的生物信息学工具也得以进一步开发。基因组和蛋白组之间的联系、基因组序列测定和功能分析是涉及到高维多变量非线性情况下的复杂系统的问题。目前量子水平的分子生物学在生物物理、生物化学方面着眼于微观的定量描述, 在基因组、蛋白组的尺度上, 如何探讨生命现象中的结构与功能间关系、作用以及有关因素, 还是国际学术界很前沿的课题。本文着重从复杂性科学的角度提出一种创发性机制引导的人工生命模型和相应算法, 以由符号化进化动力学生成满足一定约束条件的蛋白质符号序列, 由计算机的模式发现过程确认其中有意义的蛋白质序列。该工作的意义在于由粗粒度的计算过程自动地演化出自组织的模式数据, 经确认其获得的“真实”数据后, 以达到蛋白质序列的计算机辅助分析的结果 (这里的“真实”是指该数据与生物模式数据相匹配)。

1 模型的形式化表示

令概率测度空间为 (S, M, P), 其中 S 表示状态的总体、M 表示 Borel 测度定义下的 σ 一域、P 表示概率度量。本文提出一个基于符号化进化动力学的基因组数据采掘模型, 首先用下列六元组将其加以定义:

$$\langle U_g,\ U_p,\ Q_E,\ V_d,\ Z_c,\ W_T \rangle$$

其中 U_g 和 U_p 表示人工生命意义下的人工基因组和人工蛋白组, Q_E 表示进化操作, V_d 表示采用粗糙集形式的进化控制机制(属于元进化作用), Z_c 表示模式匹配, W_T 表示计算终止的判据。

U_g 为抽象代数系统, 用来表达动态的符号描述对象, 它主要由有限状态机的操作和格集合构造来完成。令 L_J 表示满足下列性质的格、L_Q 表示相应的有限状态机:

(1) L_J 包含交集、并集操作;

(2) L_J 满足交换律、结合律、吸收律;

(3) L_J 中间可定义“$\rightarrow$”(推出) 关系, 即有: $u \rightarrow v \Leftrightarrow u \cup v = u$, 其中 u 和 v 属于 L_J 中的元素;

(4) L_Q 的状态集合, 也包含(3) 中“$\rightarrow$”关系所形成的半序集;

(5) L_Q 相关的半序集, 在一定条件下在逻辑上具有与 L_J 的同一性。

在这里, 格这种抽象代数系统的结构为有限状态自动机的构成提供了柔性的合理约束, 有限状态自动机的框架则是格集合及其符合化算子发生作用的论域, 这是由符号化粗粒度信息处理的形式化体系理论所要求的。在演化的基因组动力学所决定下, 这里的格系统可衍生出若干种不同形态的格: 子格、商格、对偶格、

完备格、有余格、模格。

令 $\delta_J = \{L_J^{(0)}, L_J^{(1)}, L_J^{(2)}, L_J^{(3)}, L_J^{(4)}, L_J^{(5)}, L_J^{(6)}\}$。其中 $L_J^{(0)} \triangleq L_J$; $L_J^{(1)}$, $L_J^{(2)}$, $L_J^{(3)}$, $L_J^{(4)}$, $L_J^{(5)}$, $L_J^{(6)}$ 分别表示子格、商格、对偶格、完备格、有余格、模格, 并满足下列性质:

(1) $L_J^{(1)} \subseteq L_J^{(0)}$, 并且满足 $L_J^{(0)}$ 相关的交集、并集操作的封闭性;

(2) $L_J^{(2)}$ 为 $L_J^{(0)}$ 的一个闭区间;

(3) $L_J^{(3)} \subseteq L_J^{(0)}$, 并满足关于交集、并集操作的对偶性;

(4) $L_J^{(4)} \subseteq L_J^{(0)}$, 并存在最小上界和最大下界;

(5) $L_J^{(5)} \subseteq L_J^{(0)}$, 令 L_{Max} 和 L_{Min} 分别表示 $L_J^{(5)}$ 的极大元素和极小元素, 关于任意元素 $l \in L_J^{(5)}$, $l^* \in l_J^{(5)}$, 且 l 与 l^* 不同, 则有

$$l \cup l^* = l_{\mathrm{Max}}$$

$$l \cap l^* = l_{\mathrm{Min}}$$

(6) $L_J^{(6)} \subseteq L_J^{(0)}$, 并满足于 $\alpha \rightarrow \beta$ 的下列性质:

$$\alpha \cap (\beta \cup \gamma) = \beta \cup (\alpha \cap \gamma)$$

其中 α, β, γ 均为 $L_J^{(6)}$ 中的元素。

这样, 该模型环节中偏序关系"$\rightarrow$"和集合元素间的拓扑联系决定了所需构造的基因形态。并与符号化非线性耦合过程相关联。

U_p 是符号序列, 其字符集为蛋白质编码的有限集, 这是在多维空间中由符号动力学所产生的, 相互间的耦合关系可通过非线性映射来描述。每维的基本符号序列由进化操作 Q_E 生成, 其定量过程为

$$\begin{cases} G_L(n+1) = h(x_n) \\ \quad = \{\theta_i \mid C \cdot x_n \in T_i,\ \theta_i \in S_L\} \\ x_n = a \cdot x_{n-1}(1 - x_{n-1}),\ a \in A_s \end{cases}$$

其中 θ_i 为符号, T_i 表示闭区间 $\{M_T^{(i)}, N_T^{(i)}\}$, $M_T^{(i)}$ 和 $N_T^{(i)}$ 为正整数, C 为实系数(一般可取为正值), a 为正数, A_s 表示 Logistic 方程的奇怪吸引子域, $n \geqslant 1$, $x_0 = 0.05$。

各维之间的耦合关系可采用下列矩阵来表示:

$$W = \begin{bmatrix} \omega_{00} & \omega_{01} & \cdots & \omega_{0d} \\ \omega_{10} & \omega_{11} & \cdots & \omega_{1d} \\ \vdots & \vdots & \vdots & \vdots \\ \omega_{j0} & \omega_{j1} & \cdots & \omega_{jd} \end{bmatrix}$$

其中 d 和 j 分别为某两维的符号序列个数, 整个空间中的非线性关系以两两相关的矩阵为基础, 在对称、自反、传递的性质下, 又构成一个代数系统, 显然也可通过定义算子而建立包含环模在内的抽象代数结构。由此可知, 该环节的计算效率可通过该算法的计算过程来获得。从生命科学中与生命起源相关的 DNA 及 RNA 共生关系学说来看, 这种仿真的建模与计算也是可取的。模型中 V_d, Z_c, W_T 这三个环节则通过数据结构的设计在具体算法中来体现。

2　算法的计算过程

与前面所阐述的模型相对应的算法为模式发现算法, 其过程中综合了粗糙集、进化计算、符号动力学三个因素, 而这三个因素恰恰是现有数据采掘技术中所没有集成和嵌入的。整个算法的流框图如图 1 所示。进化操作有别于现有技术之处在于, 它分为实数取值和符号取值两种形式; 约束条件作用下的选择过程着重在有限长度、有限字符种类、合理数据结构的复杂度 (空间) 等方面对候选模式 (符号串) 进行限制, 该约束还是比较弱的, 因而使得该算法具有很强的创发性, 并且适于通用型的对象描述。在该算法中 V_d 由对混沌动力系统的信号控制来达到对进化控制的目的, 而 Z_c 模式匹配采用文法推断意义下的句法模式识别形式, W_T 主要包括对匹配完成和最大世代数满足两个因素的判断。

采用该算法所生成的一个候选符号序列示于图 2。在该算法所产生的候选符号串中值得注意的是产生了幽门螺杆菌H. Pylori相关N

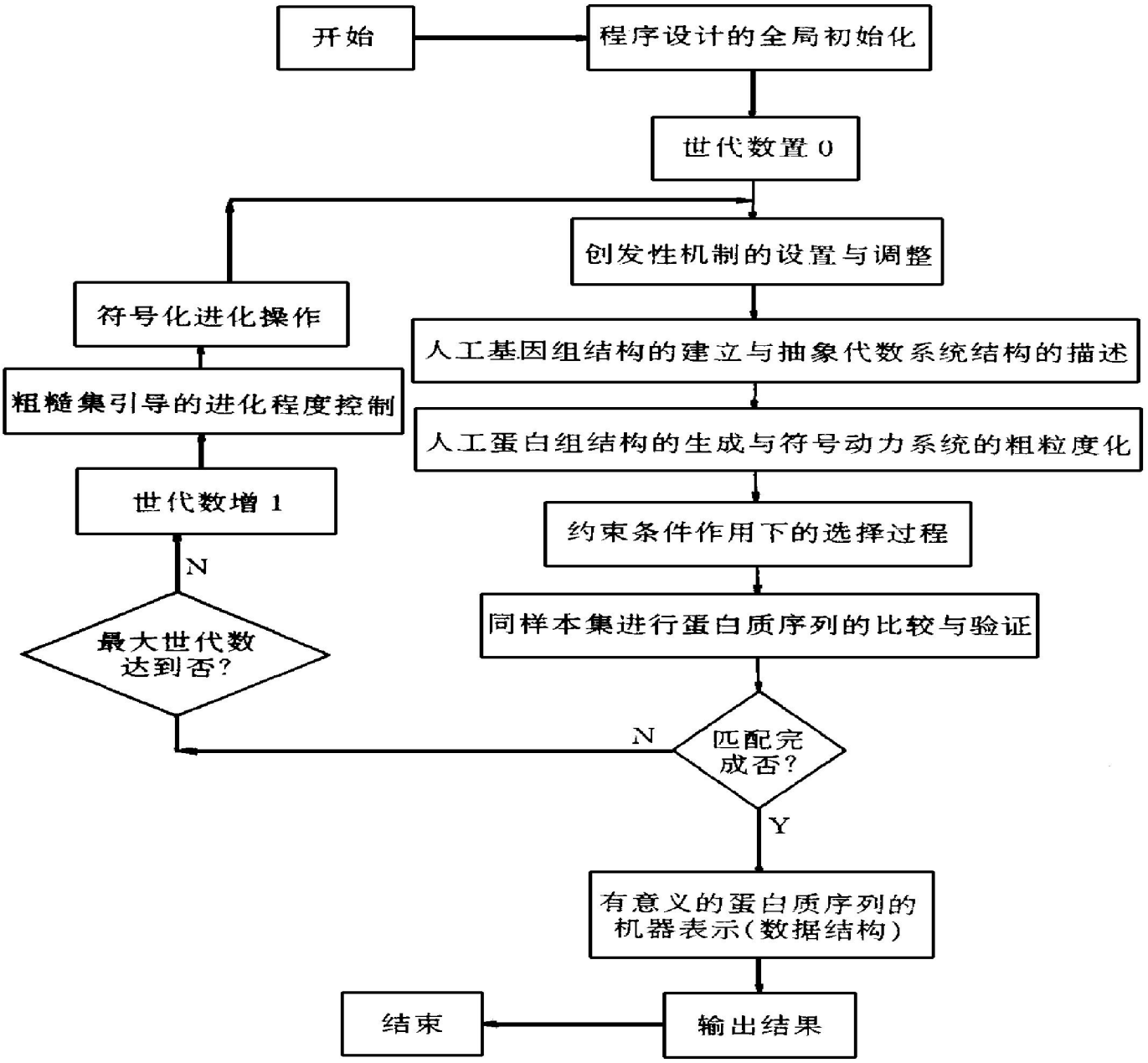

图 1 算法概貌

Fig. 1 Overview of algorithm

末端肽序列 HopC 的符号串, 示于图 3 (该符号串可参看文献 [8]) 。国际上在 1997 年 8 月发表了 H. Pylori 基因组的全序列, 该科学发现的意义在于人类获得了认识与包括胃溃疡在内的胃病原体引发疾病诊断紧密联系的基因组学基础。破译基因组编码是非常复杂的非结构化工作, 现有计算机技术是采用将待分析的序列与库中模式相比较, 遵循有关生物学的规律, 并最终由生物学实验检验, 这种方法可称为正向法。本文中的方法则是由计算机模型产生、演化、最后取得其候选集中存在生物学实验所公认的序列这样的效果, 这是递向法。这里所阐述的工作为建立计算机自动发现新序列和预测其功能的辅助系统进行了必需的探索, 并构造了一定的理论基础。

3 结论

本文工作与现有技术相比的主要区别在于: (1) 将粗糙集、符号动力学、进化计算嵌入到数据采掘过程中, 并将其加以综合; (2) 采用了反向式的模式发现过程, 体现了人工生

HAIYjMjTAAAADMjSAACKj- - nWACJr- - - - jRACIZjOAK- - - - jSABEQAFSABEQAEQADMjRHBFUYj
AABEPAGWMDLnVCBGUnABFSrBEQRFUIZjAABHjI- nVQABGWBCJjrjr- - - - jTHBFAABHBHAHMHnHXB
GUnABGVrADMjRUBGUAAACK- - - - nUVAABGVBBERrCKjr- - - ZjQFEPBI- nUAABHXFSXBEQnEPGG
VAACIZjNjNjQAEP- - - GWUCI- nUVBAACJ- - - - - - WrDNjQFDNjNjQUFSAADL- - - - XSEOEK- -
- ZjQGDNjPII- nUNAADMnVGACIZjNjPBFUCAABFUDNAABFSjACK- - - ZjQAEP- - - - GWEBGX- - -
DNjOjL- - - - XIERnDMjTDMAABFSIABFTPBFUCAAABEPNJ- nVSADOjMjSQADMjTWAACJr- - - - nU
jAAADLnVDBFSDBEOBJrj- - - - - YjMjSUBFSABEQNFSnADNjPjHXSFSDACKj- - nWJBGVnADMjS
DABGUBAACJ- - - - - - XJEQ- - ERjCKj- - nWSBHXQFSSBFSNACIZjPFGWACI- jSWBEQVEODK- - -
- jRBBHXAFTJ- - AAACJrj- - - - - YjMjSWBFSXBFTSBAAACKj- - - - - YjLnWCCK- - - YjNjQVFS
DACI- nVBADLnVJBFRABHXjEOYK- - - ZjPBIYjMjSSABGUCABFTYjLnWCCAABHXjDOjL- - - - CHXnF
TAABEPJGVRACHYjKj- - - - - - ZGVBACHYjLnVBBFSAABHjHXWFSAACKj- - - - - - XjEOJ- - - -
- - YFKjrj- nVGADNjPYHXjFUZGVAABFRjBGWBDNjQAEPHXI- jRABH- - GVjAEOJL- - - YjL- - - -
jHYVKj- - - - - - WPDMjRFCHYjL- - - - BGWjCIZjPNGWACJ- - - - - - XIZjNjOjLnWVBGWODL- -
- - jGVBFUVKAACHYDJr- - - ZjQYFS- - - - BEOWJ- - - - - - XjFSGACJ- - - - - - YJr- - -
ZjQOFSnABGWWBHX- - - - EPGI- jTUAABFSHABGV- - - BEPYHX- - - - EPOGV- - ADLnVGACK-
- - ZjPJIZjPQHSIZjPWGVHADLnWICK- - - YjMjTSABEPVGWEDL- - - - X- - - - FSOADMjT
LAABEOKKjrjr- - - ZjQQFSSBFTPWGAACIYjNjQKFSjACJ- - - - - - EHYjLnWLBGWFBGWDCK- - - -
jRBBGVKBFTrAAADNjNjQFFSWADL- - - - YGWjCK- - - - jUCJ- - AABHjGVWACKj- - - - - - XWE
QKFSjADNjQKEQTFTAABERNCKjrj- nVLABHWGVKBEOUJ- - - - - YjMnUjAADNjOjL- - - - XW
ERKCK- - - - jTjABEOTJ- - - - - - XBEQNEQjFS- BFTABAACJ- nVEADMjRYCKj- - - - - - AIZjQ
jEQXFSRADL- - - - jHAJ- - - - - - JHXXFRQBGWjCK- - - ZjOjMjTCAADL- - - - XDEPRI- jT- -
AABFUjQjAABEQREPLHXjEQAEP- - - - J- - nWXCJrjrjr- - - ZjQjFTMAABEQDFUXDEAABFTT- -
AAADNjQjEPBGVRADLnWjDMnUPAABGUXABFTrAACKjr- - ZjQAEQEDNjPX
HYjLnWTBGVABFTjAACIZjORKj- - - - - - jHYnKj- - - - - - XGEOAJr- - - ZjOjMjR- - - BGVAADL
- - - - XjFTHABEQWDNjPAGVTBFSCABFTRKj- - AABGVnADNjOjL- - - - XJEPZHYjL- - - - - - -
IZjRACK- - - ZjOTJ- - - - - YjMjRjCJ- - - - - - BHYCJrj- - - - - YjL- - - - BGWnDLnWOBHX
HXPEQYFS- - - - ACK- - - YjMjR- - - CKj- - - - - YjMjTjRjAAADMjUYCAAACL- - - YjL- - - -
XGFTDLAAADLnWHDL- - - - YIYjLnV- - - ACL- - - - Y- - - Jr- - - ZjOjK- - - ZjQjDNjPMGWYCI
- jRLBHLHYGK- - - - jTAAAABFTDLAAACJ- - - - - YjL- - - - LGV- - - ADMj
TOAABEQQEPNGUMAABGW- BGWBCI- jSGABHXTEQAEOFKj- - - - - - YjMjR- - - BH- - - - -
GWGDMjSMADNjOjMjTEPAABGVABEO- K- - - ZjPIIZjNjPHJ- - - - - - XAEOFK- - - YjMjRjDMn
VACHYjL- - - - XMEPNJ- - nVjBFTAABEQAERODMnUjAABGVNBEQJERXCK- - - - jRYCI- jTD
ABEPAGWYDLnWMCJrj- - - - - - XjDNjPAIYjMjSOACIZjPAHVJ- - nWJCKjr- - - - jTRABFTjABEPEI-
jRYCI- jUCJrAAACJrrjrjr- - - - nUSOAACI- jTHVAABEQCDOjL- - - - WTDNjQTEPBI- jR- CI- jS- AD
MjTAAADMjTjAADMjRUBHXAFS- BEQERADL- - - - XDEPLI- jRNCI- nUBAAADMjS- BF

图 2　计算机生成的候选序列示例

Fig. 2　Candidate sequence examples generated by computer

ED- DGGFFTVGYQLGQ- VMQDVQNPG

图 3　Hopc 的符号序列

Fig. 3　Symbolic sequence of HopC

命的创发性特点。基因组学在很大程度上改变了人类认识生命现象的方式，特别是人类基因组计划不仅在探索自然规律方面，而且在医疗、制药、生物技术方面具有广阔的应用前景和社会经济效益[9, 10]。完整基因组的获得、基因组的生物信息学分析等环节均依赖于基因组

功能辨识，连锁图、物理图、转录图、序列图诸层次的有机联系，为工作的开展提供了逻辑基础。就遗传学意义下的分散性而言，基因型与表现型的非线性高维模式关系又是定性分析与定量计算相综合的途径。显然，基因组序列测定与功能分析中，对象信息处理的非线性非确定性又要求相应的定位、克隆等工作具有很大的经验性。基因组功能辨识与分散性描述是基因组学中具有很大难度的工作，但也是不可回避的关键问题。从本质上来看，整个计算机信息处理的核心就是模式识别和系统辨识。为了实现这个功能，基于人工生命的计算机系统的设计和开发就是必要和可行的。

REFERENCES

1 Elghanian R, Storhoff J J, Mucic R C *et al*. Science, 1997, 277 (5329): 1078- 1081.

2 Service R F. Science, 1997, 277 (5329): 1036- 1037.

3 Langley P, Simon H A, Bradshaw G *et al*. Scientific Discovery: Computational Explorations of the Creative Processes. Cambridge, MA: MIT Press, 1987.

4 Xu Y, Mural R, Einstein J *et al*. Proc of the IEEE, 1996, 84 (10): 1544- 1552.

5 Jong H and Rip A. Artificial Intelligence, 1997, 91 (2): 183- 203.

6 Anand S S *et al*. IEEE Expert: Intelligent Systems & Their Applications, 1997, 12 (2): 65- 74.

7 Simoudis E. IEEE Expert: Intelligent Systems & Their Applications, 1996, 11 (5): 26- 33.

8 Tomb J F *et al*. Nature, 1997, 388: 539- 547.

9 Cohen J. Science, 1997, 275 (5301): 767- 772.

10 Weinstein J N *et al*. Science, 1997, 275 (5298): 343- 349.

GENOMIC DATA MINING BASED ON SYMBOLIC EVOLUTIONARY DYNAMICS

Liu Jianqin

College of Information Engineering,

Central South University of Technology, Changsha 410083, P. R. China

ABSTRACT A novel model of genomic data mining and a corresponding algorithm for pattern discovery were proposed. The model consists of six units such as artificial genome, artificial proteome, evolutionary operation, evolutionary control, pattern matching and termination judgement. The abstract algebraic structure is described by lattice set configuration and finite state automata dynamically. The candidate string sequence is generated by evolutionary algorithm with symbolic dynamics. The degree of evolution is controlled by meta-evolution mechanism and expressed by rough sets. The pattern matching procedure is implemented by syntatic pattern recognizer and grammar inference. Termination judgement is dependent on concret problem solving paradigm. The algorithm is with the cycle type of implicit parallation and population searching. The data structure focusses on coarse-grained symbolic information processing and modular programming oriented to semantics. With the application of the above-mentioned artificial life techniques, candidate symbolic sequences have been automatically produced by computer system and "real" amino-acid sequence obtained among them. The experimental result shows that the computational method proposed and implemented here is helpful to the building of unified computational theory of bioinformatics in the genomics level and development of application systems.

Key words genomics bioinformatics evolutionary computation

(编辑 袁赛前)