论 文

一种新型智能预测系统模型 —— 复合金字塔模型

杨炳儒*, 曲武, 王莉军, 周颖

北京科技大学信息工程学院, 北京 100083

* 通信作者. E-mail: bryang_kd@yahoo.com.cn

收稿日期: 2010-07-22; 接受日期: 2011-04-15

国家自然科学基金 (批准号: 61175048) 和科技部创新方法专项 (批准号: 2010IM020900) 资助项目

摘要 将一般性智能预测系统构造向结合具体领域的个性化智能预测系统构造的发展, 表征了智能系统当今发展的重要趋向. 蛋白质结构预测是生物信息学领域的国际性难题与主要挑战性问题之一. 在本文中, 我们提出了一种新型的智能预测系统模型来预测蛋白质的二级结构, 它是一个多层递阶的复合金字塔模型 (CPM). 该模型包含 4 个独立的智能接口, 并且综合涵盖几种相关的知识发现方法. 领域知识贯穿该整个模型, 并通过因果细胞自动机有效的选择属性, 以及高纯度结构数据库的构造. 在 RS126 的数据集上的预测准确性达到 84.31%; 在 CB513 数据集, 预测准确性达到 86.78% (居国际已知之优先地位). 同时对 CASP8 序列预测的结果发现是优于其他方法的, 如 PSIPRED, Jpred, APSSP2, BehairPred 等. 大量实验结果表明: 该智能预测系统模型具有较强的泛化能力, 并提供了其他类型智能系统模型构造的方法论示范.

关键词 知识发现 智能预测 系统模型 复合金字塔模型

1 新型智能预测系统模型构造的背景

1.1 一般性智能预测系统

生物学是研究生命系统以及它们如何工作的学科. 利用智能系统去解决生物学中的难题已被广泛应用, 从生物分子的原子细节描述到在生态系统中的物种的相互作用. 这个领域已经备受关注, 同时, 生物学家继续提出了新的产生大量数据的高效的实验方法, 产生了大量的数据, 就必须借助智能系统来完成对这些数据转化成知识的过程.

如今蛋白质序列和结构的预测中产生的数据量超过预期, 因此需要智能预测系统来解决这个难题. 利用智能系统、动态规划、计算机辅助设计、图像分析和机器学习等来预测和分析蛋白质分子的物理和空间特性是非常好的方法. 二级结构在生物化学及结构生物学中, 是指一个生物巨分子, 如蛋白质及核酸 (DNA 或 RNA), 局部区段的三维通式. 它并不描述任何特定的原子位置, 而这会在三级结构中处理.

二级结构预测的难点和重点, 已引起许多研究人员从事该项工作, 主要的工作在两个方面: 第一, 预测模型和预测方法的研究, 通过不断的努力, 已经开发出数百种预测模型和方法, 并不断涌现新的预测模型和方法; 第二, 是研究预测系统, 即将预测、计算机和人工智能技术结合起来, 帮助用户开发

引用格式: 杨炳儒, 曲武, 王莉军, 等. 一种新型智能预测系统模型 —— 复合金字塔模型. 中国科学: 信息科学, 2013, 43: 555–570

一个用于预测分析的计算机支持系统. 预测支持系统不仅拥有多种预测模型和方法, 而且还成为预测和知识经验领域的专家, 通过用户友好的界面相互作用, 是一种有效的预测工具, 是解决“预测难题”的一个重要途径.

蛋白质二级结构: 指蛋白质多肽链本身的折叠和盘绕方式. 二级结构主要有 α -螺旋、 β -折叠、 γ -转角. 常见的二级结构有 α -螺旋和 β -折叠. 二级结构是通过骨架上的羰基和酰胺基团之间形成的氢键维持的, 氢键是稳定二级结构的主要作用力.

仅从其预测氨基酸序列的问题可看出蛋白质三级结构的预测是一个非常具有挑战性的问题. 只有以它的二级结构序列为基础, 使用相对简单的二级结构去预测蛋白质的三级结构是通常的办法, 很长一段时间也一直是研究的重点.

1.2 新型智能预测系统模型构造背景

回顾蛋白质结构预测中的发展历史, 里程碑式的方法有多种, 如支持向量机^[1]、隐 Markov 模型^[2]、数据挖掘^[3,4]等等. 目前也有: PSIPRED, SAM, PORTER, PROF, SABLE 等方法得到较为准确的预测. 但是, 我们可以发现, 这些方法在一定程度上, 都存在准确度等问题, 例如尽管人们已提出几十种预测方法, 但其预测的准确度大多都不超过 80%^[5], 目前的预测方法的实验色彩较浓, 缺乏坚实的理论基础; 除同源性信息外, 绝大多数领域知识没有得到充分的利用. 迫切需要提出一种新的、精度更高的智能预测系统. 由于数据挖掘 (或知识发现) 在处理海量数据方面具有得天独厚的优势, 越来越多学者逐渐开始利用数据挖掘的方法研究蛋白质的结构预测问题, 并取得了一定成果. 混合方法是未来发展的主流, 最近的研究结果表明, 混合预测模型与集成预测模型通常可以获得更优的预测准确度, 因此吸引了一些生物信息学家的注意, 并建立了一些混合预测模型, 如 HYPROSP II 模型^[6,7]等. 这些模型在一定程度上提高了预测精度, 一般较难超过 80%, 仍然存在稳定性差等不足.

针对这些问题, 本文提出一种新颖的、逐步求精、多层递阶的预测系统模型 — 复合金字塔模型 (compound pyramid model, CPM), 该模型融合了基于原创性理论 KDTICM^[8] 的 KDD* 过程模型^[9,10]以及 SAC^[11], AAC^[11]等多种技法. 结果表明: 复合金字塔模型能够很好地对偏 α/β beta 型蛋白质的二级结构进行预测, 利用层次间的智能接口将各种方法有机融合, 从而充分发挥各个预测模型与方法的优点, 使整体效果达到最优. 实验表明, CPM 能够在 Q_3 ^[12] 评分标准下, 在 RS126^[12], CB513^[13] 和 CASP8^[14] 得到比其他方法都高的精度 (见第 4 节实验与分析).

本文的结构如下: 第 2 节描述了新型智能预测系统模型构造的理论基础; 第 3 节详细介绍了复合金字塔模型; 第 4 节分析了实验结果; 在第 5 和第 6 节分别给出了有关创新性的结论和对未来研究工作的展望.

2 新型智能预测系统模型构造的理论基础 — KDTICM

笔者 1 提出了一个新的研究方向, 研究基于知识发现本体的内在认知机理; 核心思想是: 将知识发现过程视为一个认知过程, 将知识发现系统视为一个认知系统; 从认知科学出发, 相继独立发现了 4 条潜在的规律, 即双库协同机制 (揭示了知识库与数据库间的内在联系); 双基融合机制 (揭示了两个知识发现过程间的内在联系); 信息扩张机制 (揭示了动态挖掘进程的规律); 免疫进化机制 (揭示了动态挖掘系统的规律性); 并且在国际上率先构建了基于内在认知机理的知识发现理论体系 KDTICM^[8].

图 1 知识库中知识素结点与数据子类结构中层之间的对应

Figure 1 KDD* corresponding structure of five algebra systems

2.1 双库协同机制

基于认知心理学的“创见意象”与“心理信息修复”原理,笔者发现了在 KDD 过程中及特定的构造下,数据库与知识库间的对应关系,论证了结构对应定理,设计了启发型协调器与维护型协调器,解决了“定向搜索”、“定向挖掘”、自主发现、实时维护等难题.

1) 结构对应定理. 论域 X 的推理范畴 $Cr(N)$ 与完全数据子类结构可达范畴等价 (建立了 2 个证明路径: (1) 利用范畴论; (2) 利用我们提出的连续映射的同伦理论的拓广——泛同论理论).

2) 通过结构对应定理,建立数据库中数据子类结构的“层”与知识库中知识“素结点”的一一对应关系 (见图 1),实现定向搜索与定向挖掘. 提出并实现了 2 个协调算法: (1) 对领域固有的知识库的实时维护 (通过维护型协调算法与构件); (2) 自主发现知识短缺产生创见意向 (通过启发型协调算法与构件).

3) 关于双库协同机制具体实现的进一步讨论,如可达关系的概率估计定理等 (略).

2.2 由双库协同机制诱导出的新过程模型 KDD*

将双库协同机制及其支持下的两个协调器的构造,融入经典的 KDD 过程中,形成我们独立提出的 KDD*[9,10] 新过程模型,从而从根本上改变了原有的知识发现进程与运行机制. KDD* 过程模型如图 2 所示.

原有知识发现过程模型 KDD 在如下技术与功能方面存在着不足:

- 1) 领域知识不能实质性地介入到数据挖掘 (知识发现) 过程中;
- 2) 系统不能自主地实现对短缺知识发现和挖掘;

图 2 KDD* 过程模型

Figure 2 KDD* process model

3) 仅根据用户的兴趣度产生聚焦, 确立挖掘方向, 会导致大量重复、冗余规则的产生; 与系统自身挖掘的短缺知识不能较好地吻合;

4) 不能对知识库进行动态实时维护;

5) 模型的实现是基于语义层面的.

KDD* 过程模型针对上述各种不足给出了具体的创新方法和实现技术, 具体为:

1) 在挖掘的过程中, 领域知识通过两个协调器直接地、具体地介入到挖掘过程中, 主要借用同步进化和协同计算的思想;

2) 系统能通过有向超图的邻接矩阵产生定向聚焦, 自主地实现对短缺知识需求和挖掘;

3) 聚焦问题: 定向挖掘的方向与进程当且仅当在用户“感兴趣点”与系统自主发现的“短缺知识”相吻合的情况下才能产生. 这样, 不至于挖掘出大量重复、冗余的知识, 大大减少规则评价量. 这样做的主要目的是减少搜索空间、提高算法效率, 为算法能够只通过处理少量数据而达到挖掘效果, 提供了必要的技术支持;

4) 随着知识的积累, 知识库的知识也会越来越多, 为了能快速地对应用问题做出反应, 在新模型中加入了维护协调器, 有效地、动态地、实时地处理了知识的重复、冗余、矛盾、循环与从属;

5) 新模型是基于认识心理学的“创建意象”与“心理信息修复”两个认知特征的, 故新的模型有坚实的理论基础; 模型的实现是基于理论层面的.

2.3 Maradbcm 算法 (简称为 M 算法)

基于双库协同机制和 KDD* 模型, 我们提出了一种全新的挖掘关联规则的 Maradbcm 算法 (以下

图 3 Maradbcm 算法流程图

Figure 3 Maradbcm algorithm

简称 M 算法)^[15]. 输入: 规则强度阈值 Min_Intensity, 支持度阈值 Min_Sup, 可信度阈值 Min_Con; 其算法流程图如图 3 所示 (其中, K_m 表示规则长度为 m 的短缺知识集, 即 $K_m = \{r | \text{Len}(r) = m\}$).

2.4 基于 KD 的因果自动机理论

由于所构建的基于 KD 的因果自动机 (causality automaton based KD, CAKD)^[8], 应涵盖于有限自动机之中, 但它同时又具有其自身的特性. 因此, 将从有限自动机入手, 根据有限自动机的定义方式, 给出一种 CAKD 的概念上的描述.

定义 1 因果状态是因 (果) 事件、因果关系和因 (果) 事件值的三元组, 记为因果状态 $u = (x, r, a)$, 其中 x 为因 (果) 事件, r 为因果关系, a 为因 (果) 事件值. 有时又称 u 为由 x 构建的因果状态. 设 X 为因 (果) 事件集, R 为因果关系集, A 为因 (果) 事件值集. 如果 x 是一个因 (果) 事件, 则记 $x \in X$; 如果 r 是因 (果) 事件间的一个因果关系, 则记 $r \in R$; 如果 a 是因 (果) 事件的值, 则记 $a \in A$, 一个因 (果) 事件 x , 可有多值, 记为 $a_x = a_{1x}, a_{2x}, \dots, a_{kx}$.

定义 2 因果状态空间为所有因果状态的集合, 记为 $U = (X, R, A) = \{u | u = (x, r, a)\}$. CAKD 作为有限自动机的一种特殊形式, 它既有有限自动机的通用性质, 同时还具有其自身的特性, 因此可

参照有限自动机的定义方式, 将 CAKD 表示成一个六元组, 其形式化的定义如下.

定义 3 基于知识发现的因果自动机是一个六元组: $M = (U, \Sigma, \Phi, \mu_0, F, T)$. 其中: U 为因果状态空间, 是状态的非空有穷集合, $u \in U$, u 称为 U 的一个状态; Σ 为输入字母表, 它的每一个因 (果) 事件值 a 可当作一个输入字符, 输入字符串都是 Σ 上的字符串; u_0 为 M 的开始状态, 也可叫初始状态或者启动状态, 因 (果) 状态 $u_0 \in U$; F 为 M 的终止状态集合 (可空), 也可叫接受状态, $F \subseteq U, \forall u \in F$, 则果 (因) 状态 u 称为 M 的终止状态; T 为时态评价算子; Φ 为状态转换函数 (因果关系), $\Phi: U \times \Sigma \rightarrow U$, 从因 (果) 状态空间到果 (因) 状态空间产生的映射. 在上述六元组中, 其中 4 个 (U, Σ, u_0, F) 是有限自动机中固有的通用性元组, 在此不再赘述, 只对另外两个可反映其自身特性的元组 T 和 Φ 加以解释.

因果关系的本质特性是: 原因在时间上不晚于结果. 要全面地反映因果关系的特性, 就必须考虑时间因素. 这是因果关系与一般条件关系的本质区别. 为全面刻画因果关系的特性, 下面将引入算子 T 的概念.

定义 4 时态评价算子 T 是一个映射, 它所对应的是一个评价函数值, 可具体表示为 $T: XYP \in [0, 1]$. 其中 X 为因状态域, Y 为果状态域, P 是指因状态 x 与果状态 y 在特定的时间 t 里的结合强度 (即指在因状态出现时果状态出现的必然结果, 如必出现、频繁出现、时隐时现、出现少、不出现等情况). 时态评价算子 T 主要包括 $\Box c, \Diamond c, \circ c, \triangleright c, \neg c$.

- 1) $A\Box cB$ — 表示“有 A 则 B 必出现”;
- 2) $A\Diamond cB$ — 表示“有 A 则 B 频繁出现”;
- 3) $A\circ cB$ — 表示“有 A 则 B 时隐时现”;
- 4) $A\triangleright cB$ — 表示“有 A 则 B 出现少”;
- 5) $A\neg cB$ — 表示“有 A 则 B 不出现”.

时态评价算子是二元算子, 它是以时间做为参数, 针对原因和结果的不同搭配 (组态), 通过因、果出现的结合强度的强弱, 可表现出因状态和果状态这一对序偶 (因, 果) 的模糊关系. 其中序偶 (因, 果) 可以是一对一、一对多、多对一等形式.

状态转换函数 Φ 是在因果状态空间内, 整个因果变化的链条中所映射的转换. (因果关系) 状态转换函数 $\Phi: U \times \Sigma \rightarrow U$, 是指从因 (果) 状态空间到果 (因) 状态空间产生的映射.

在 CAKD 的定义中, 因果状态空间 U 中可包括因状态空间和果状态空间, 根据 M 的开始状态及终止状态中所取的因、果状态不同, 可将 CAKD 划分为判定型、诊断型两种形式.

值得注意的是, 复合金字塔模型利用了基于知识发现的因果细胞自动机理论, 即在氢键形成概率、碳环、疏水性、带电量、残基大小、电荷密度、分子量等多种物化属性中, 利用因果细胞自动机建立窗口中间氨基酸二级结构类型与临近氨基酸物化属性间的因果关系. 从图 4 中我们可以发现, 对二级结构影响最大的是氢键形成概率、疏水性及带电量 3 个物化属性, 残基大小也有一定的影响. 以此作为后继分析的基础, 进而精简了物化属性的数目, 降低了模型整体的复杂性.

3 新型智能预测系统模型 — 复合金字塔模型

现代研究表明: 分层递阶结构是降低系统复杂度的最有效的处理手段, 而有序的粒度空间理论是建立复杂系统的分层递阶结构最有效的手段之一. 复合金字塔模型采取了逐步求精、多层递阶的架构; 各层的粒度空间由粗到细; 各层功能相对独立且通过智能接口紧密衔接. 其总体架构如图 5 所示.

图 4 基于知识发现的因果细胞自动机所得因果关系图

Figure 4 Causal cellular automata based on knowledge discover

图 5 复合金字塔模型 (综合分析层综合了 SVM 多分类方法与同源性分析方法, 其结果分为两个集合, 其中集合 (1) 基本确定为 H, E, C 结构, 直接输送结果库; (2) 为尚待 SAC, AAC 模块进一步判定的部分)

Figure 5 Compound pyramid model. Comprehensive analysis layer integrates the ISVM and IHA methods, then generates two kinds of results: one of which, denoted as (1), is sent directly to the results database; while the other, denoted as (2), needs further determination by the SAC and AAC modules

3.1 综合分析层

本层综合了改进的同源性分析方法与优化的 SVM 类化方法, 是整个模型的基础层, 可以完成 50% 以上的特征明显的待测氨基酸二级构象的预测. 同源序列方法^[16]是蛋白质二级结构预测领域广泛采用的一种方法. 复合金字塔模型中, 采用 APSSP2 方法, 该方法是一种较为成熟的基于多序列匹配的蛋白质二级结构预测方法, 其基本步骤如下.

3.1.1 改进的同源性分析方法

同源序列分析方法是蛋白质二级结构预测领域广泛采用的一种方法. 在我们的实验中, 采用加以改进的 SNNS 神经网络模块 (包括多序列比对功能块和人工神经网络功能块).

1) 使用 SNNS 神经网络技术和 PSIBLAST^[17,18] 算法生成多序列比对结果. 我们首先应用由长度为 13 的滑动窗口内的氨基酸残基生成的 PSSM(position specific scoring matrix) 矩阵, 并加入氨基酸的三类物化属性 (即: 疏水性, 氢键和带电性) 作为输入节点; 该网络中包含了另外 9 个隐藏节点和 3 个输出节点; 标准神经网络的输入是一个窗口, 输出的是 H, E, C (3 类二级结构). 然后使用 PSIBLAST 生成的 PSSM 矩阵作为神经网络第二层的输入, 网络的每一层是相互结合和均衡考虑预测的结构状态, 即 α -螺旋 (α -helix, H)、 β -折叠 (β -sheet, E)、无规则卷曲等 (C-coil, C).

2) 根据神经网络的前两层对蛋白质的二级结构进行预测: 如果预测的残基一致时, 对蛋白质的结构分类为相同位置; 否则将其分类为不同位置. 使用不同的神经网络方法在预测相同位置的残基结构时, 能得出一致的结果; 而那些不同位置的残基结构, 需要训练分类神经网络 (如第二层 BP 神经网络), 通过该层网络的同源性不同的位置将替代最后的预测结构.

3.1.2 优化的 SVM 多分类方法

此处优化的 SVM 多分类 (SVM multiple classification) 方法是采用了结构与属性相结合的“轮换式”分类方法.

1) 我们的大多数测试都是采用径向基核函数 (RBF), 其公式定义为 $K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2)$. 在最优化过程中, 为使得 RBF 处理复杂分类问题更加有效, 可根据每一步的计算结果和其中的二元分类器, 来确定合适的参数 γ 和 C . 最优化参数的选取是数据集 RS126 在 $\gamma = 0.05$ 和 $C = 2.0$ 时; 数据集 CB513 在 $\gamma = 0.05$ 和 $C = 1.0$ 时; 数据集 CASP8 在 $\gamma = 0.05$ 和 $C = 1.0$ 时预测准确率最优, 为此我们分别采用上述参数构建 SVM 分类器.

2) 在构建三元分类器时, 我们首先需要建立多个二元分类器对样本进行分类, 应包括 $H/\tilde{H}$, $E/\tilde{E}$ 和 $C/\tilde{C}$ 这 3 个二元分类器 (其中 “one” 表示正例, “rest” 表示反例), 并分别命名为 H/E , E/C 和 C/H , 这样可以区分任意两种状态的不同样本. 在 CPM 中对文献 [19] 提出的方法做了相应的改进, 即基于三个正反例二元分类器 ($H/\tilde{H}$, $E/\tilde{E}$, $C/\tilde{C}$) 和 3 个正例二元分类器 (E/C , C/H , H/E), 生成一系列的三元分类器为 SVM-TREE1($H/\tilde{H}$, E/C), SVM-TREE2($E/\tilde{E}$, C/H) 和 SVM-TREE3 ($C/\tilde{C}$, H/E). 在 SVM-MAX-D 三元分类器中, 对于测试集的分类根据它相应的最大正距离优化区分 SVM-TREE1, SVM-TREE2 和 SVM-TREE3 三元分类器. SVM VOTE 分类器采用简单的投票机制综合了 6 个二元分类器对样本进行分类, 这 6 个二元分类器能够预测测试集中蛋白质二级结构的状态. SVM 利用评审技术再结合三元分类器对蛋白质进行结构分类.

3) 尽管 SVM 已被证明是机器学习中有用的方法之一, 但是对于处理大样本数据集时, 特别是支持向量、优化速度和分类速度一直不尽人意. 关键在于在不降低分类准确率的前提下, 如何缩小大样本训练集的样本空间. 我们采用了新的学习策略: 首先从训练集中选取一个小规模样本集, 用这个小样本集训练得到一个初始的分类器, 然后用这个分类器对大规模训练集进行修剪, 修剪后得到一个小规模的约减集, 再用这个约减集训练得到最终的分类器.

3.2 核心判定层

本层采用以 KDD* 过程模型和 M 算法对 RS126 的二级结构库挖掘所得的关联规则集为基础的

图 6 SAC 示意图

Figure 6 SAC method

关联分类算法; 构建的 SAC(structural association classifier) 模块主要承担综合分析层中难以判定的数据分类. RS126 由 Rost 等人构建, 该数据集由非冗余的 126 个蛋白质组成 (即两两同源小于 25%), 因此更符合现实情况. 我们采用交叉验证的方式, 将 RS126 分为 10 个子集, 分别学习分类规则, 最后以分类准确率最高的一组规则作为分类规则库. 其原理是基于二级结构间的关联影响, 也即二级结构之间构象影响信息; 其核心理论基础为前述基于内在认知机理的知识发现理论 KDTICM; 工具为该理论下的 KDD* 过程模型及其 Maradbcm 算法. 通过与相关关联规则方法的比较, 可以发现 Maradbcm 算法在相同的支持度与可信度下, 通常可以获得更多的规则, 这些规则具有低支持度, 同时具有很高的可信度, 我们称之为“意外规则”. 大量研究结果表明, 正是由于意外规则的引入, 使得 SAC 的分类更为准确.

研究表明: 目标氨基酸紧邻的 12 个氨基酸对其的二级结构形成具有较大影响^[19]. 因此, 在 KDD* 挖掘过程中, 训练数据首先基于滑动窗口理论, 将序列处理为长度为 13 的窗口, 其中位于中间位置的为目标氨基酸, 其他为与其相邻的氨基酸二级结构. 对于序列的头部及尾部, 采取镜像预处理法, 即使用对于中间位置的相反位置结构填充缺失部分.

我们提出的原创性的 SAC 方法 (基于结构的关联分类算法) 本质为关联规则分类器, 其规则由 KDD* 模型挖掘二级结构互作用信息而得. 根据蛋白质二级结构预测问题的背景需要, 我们在挖掘过程的结果中, 选择了规则后件分别是 α -螺旋 (α -helix, H)、 β -折叠 (β -sheet, E) 与无规则卷曲 (C-coil, C) 组成了 H 规则集、E 规则集与 C 规则集, 并由 H 规则集、E 规则集与 C 规则集组成了蛋白质二级结构知识库. 蛋白质知识库中的这些关联规则归纳了紧邻氨基酸的二级结构对中间氨基酸二级结构的决定关系. SAC 方法依据蛋白质二级结构的临近信息以及 KDD* 挖掘得到的蛋白质知识库, 利用 CMAR(classification based on multiple association rules)^[20] 算法进行蛋白质二级结构的多分类预测. CMAR 作为经典关联分类预测方法 CBA 的改进类型, 突破了使用单一规则进行预测的方法, 而是使用了多条满足条件的关联规则进行联合预测, 如图 6 所示.

表 1 针对数据集 RS126 的 SAC 的部分挖掘结果

Table 1 The SAC module's mining results for RS126

Rule ID	Condition	Result	Support	Confidence
1	Negative6H	CenterStructureH	0.180932	0.580645112
2	Negative5H, Positive4H	CenterStructureH	0.134312	0.682655712
3	Negative4E	CenterStructureE	0.101231	0.994865125
4	Negative3E	CenterStructureE	0.102318	0.987121524
5	Negative2C	CenterStructureC	0.381257	0.890542132
6	Positive2C	CenterStructureC	0.386423	0.998095924
7	Negative1C	CenterStructureC	0.432474	0.783655676
8	Positive1C	CenterStructureC	0.436790	1.000000000
9	Negative2H	CenterStructureH	0.271264	0.990615372
10	Negative2H, Negative1H	CenterStructureH	0.261689	0.980275131

该系统从训练数据中挖掘出精度高、适应度好的关联规则, 这些关联规则描述了临近二级结构间的相互影响, 利用这些规则, 基于改进的 CBA 算法, SAC 模块得以构建. 在实现方面, SAC 同样基于滑动窗口技术, 长度仍为 13. 不同之处在于训练数据采用窗口内的氨基酸二级结构, 挖掘其共生规律, 从而形成关联规则, 并作为预测基础. 针对数据集 RS126 的部分挖掘结果, 如表 1 所示.

3.3 辅助判定层

本层的核心方式是我们提出的原创的 AAC(attribute association classifier) 模块. 通过对氨基酸物化属性的关联分析, 建立精化规则库, 然后利用改进的 CBA 算法, 对下两层无法判断的数据进行预测. AAC 方法 (基于属性的关联分类算法) 主要的工具同样是基于原创性理论 KDTICM 的 KDD* 过程模型及其 Maradbcm 算法以及改进的 CBA 算法. 通过 KDD* 系统生成的 α/β 的规则, 进行一定程度的约简后得到精炼的 α/β 规则库; 使用改进的 CBA 算法, 经过反复实验得到适用于 α/β 的支持度与可信度的阈值, 通过实验验证其结果是可靠的; 对此我们得到的 α/β 规则库与支持度、可信度的阈值是可以作为知识确定下来, 嵌入到整体的复合金字塔模型中.

在预测过程中, 我们仍然采用与 SAC 类似的分类算法. 在产生规则的过程中我们打破常规, 根据蛋白质生物数据的特性放弃了使用通用数据库来生成规则, 而是高起点地使用内容分别相对偏向 α/β 型的蛋白质库. 这样解决了长期以来, 蛋白质关联规则挖掘中存在的生成规则的支持度与可信度较低、规则不可信的问题. 经过大量实验验证与原先的采用蛋白质通用数据库相比, 其判定准确度有了较大的提升. 针对数据集 RS126 的部分挖掘结果, 如表 2 所示. 这不仅使得 CPM 保持较高的预测准确度, 而且还提供了多种可理解性的关联规则 (如表 2 所示), 这对于蛋白质空间折叠结构的研究将是非常有用的.

3.4 优化层

本层主要设计倾向性因子、位能函数及合情推理 3 类方法. 前两类方法属于生物信息学的固有方法, 其主要利用生物信息背景知识进行结构预测. 合情推理方法是建立在各种二级结构具备的不同物化属性规律基础上的. 3 种方法从不同角度对其下 3 层的结果加以优化, 以最大程度提高整体预测精

表 2 针对数据集 RS126 的 AAC 的部分挖掘结果

Table 2 The AAC module's mining results for RS126

RuleID	Condition	Result	Support	Confidence
1	HydrogenGlobal(0.57, 0.60]	CenterStructureH	0.090935	0.995805865
2	HydrophobicLeft3(2,5]	CenterStructureH	0.281751	0.995773060
3	HydrogenGlobal(0.62, 0.65], IsoelectricpointLeftrange(24,36]	CenterStructureH	0.432767	0.988526462
4	HydrogenGlobal(0.62,0.65],HydrogenRight5(0.6,0.7]	CenterStructureH	0.370103	0.989514070
5	HydrophobicLeft4(2,5],HydrogenRight6(0.7,0.7]	CenterStructureH	0.199569	0.989514070
6	HydrophobicLeft4(2,5],IsoelectricpointLeftrange(4,6]	CenterStructureH	0.218082	0.994989380
7	HydrogenTightrange(2,3],HydrogenGloabal(0.62,0.65],...	CenterStructureH	0.076872	0.992814170
8	HydrogenTightrange(3,4.5],HydrogenGloabal(0.62,0.65],...	CenterStructureH	0.394451	0.990146451
9	HydrogenTightrange(3,4.5],HydrogenGloabal(0.62,0.65],...	CenterStructureH	0.147333	0.992268371
10	HydrogenTightrange(3,4.5],HydrogenGloabal(0.62,0.65],...	CenterStructureH	0.082660	0.995399918

度. 复合金字塔模型中各个层次功能独立, 但同时紧密衔接, 构成了一个逐步求精、多层递阶的预测系统:

1) 以第 1 层 — 综合分析层作为基础, 基本完成对相当一部分特征明显的待测氨基酸二级构象的预测, 其准确率通常可以达到 90% 以上, 该层较高的预测精度是模型整体精度的重要保障. 第 2 层与第 3 层是复合金字塔模型中最为重要的部分, 由于综合分析层只能对部分氨基酸二级结构进行有效判定, 并且该部分氨基酸的结构特征最为明显, 因此留待核心判定层与辅助判定层解决的氨基酸通常不具备明显的结构特征, 进而要求更为复杂的方法.

2) 核心判定层作为复合金字塔模型中的核心部分, 其主要任务是强化综合判定层没有确定的预测结果, 保证最少的转角环形 (coil) 的待预测氨基酸被输送到辅助判定层. 虽然一般来讲这部分氨基酸数量较少, 然而由于其结构特征不明显, 采用序列比对、同源性分析等方法很难获得较好的预测结果, 因此复合金字塔模型另辟蹊径, 利用 KDD* 过程模型对 α/β 库进行挖掘, 由于 α/β 库的构建采用了高纯度的方法, 并且 KDD* 模型本身具有挖掘意外规则的特性, 因此挖掘所获得的关联规则集是精华的, 即其区别于一般关联规则挖掘结果, 该规则集具有较高的支持度与可信度, 为关联规则分类奠定了较为坚实的基础.

3) 辅助判定层经过 AAC 模块及改进的 CBA 方法的双重校验, 使绝大多数待测氨基酸在此得到解决.

4) 优化层在复合金字塔模型起到辅助的作用, 由于其采用了合情推理、位能函数与倾向性因子方法, 虽然这 3 种方法较为传统, 在单独使用时预测准确度较低, 然而我们发现当使用这些方法对预测结果进行修正时, 往往可以获得更高的精度, 这是由于这些方法内聚的知识是其他方法难以获得的, 因此可以与其他 3 层形成有效互补.

4 实验与分析

我们使用了 3 种不同的数据集, 以开发和测试我们的新方法, 分别为 RS126^[12] 数据集、CB513^[13] 数据集和 CASP8^[14] 数据集. 同时采用 Q_3 标准作为评价指标, 其定义为预测正确的氨基酸数与氨基

表 3 CPM 在 RS126 数据集上每层预测精度

Table 3 Prediction accuracy for each layer and scale of CPM on the RS126 dataset

Module	Accuracy		Percent
Comprehensive analysis layer	ISVM	IBP	18851/24806=75.99%
	19997/24806=80.61%	20263/24806=81.69%	
	17020/18851= 90.29%		
Kernel judgment layer	3876/5904= 65.65%		5904/24806= 23.80%
Assistant judgment layer	18/113= 15.93%		113/24806=0.46%
Total	20914/24806 = 84.31%		

表 4 CPM 在 CB513 数据集上每层预测精度

Table 4 Prediction accuracy for each layer and scale of CPM on the CB513 dataset

Module	Accuracy		Percent
Comprehensive analysis layer	ISVM	IBP	
	118862/146233=81.28%	124711/146233=85.28%	114295/146233= 78.16%
	106522/114295= 93.20%		
Kernel judgment layer	20283/31608= 64.71%		31608/146233= 21.61%
Assistant judgment layer	95/467 = 20.34%		467/146233= 0.32%
Total	126900/146233 = 86.78%

图 7 分别在 RS126 和 CB513 数据集上 CPM 与其他研究结果的 Q_3 准确度比较

Figure 7 Q_3 accuracy comparison with other research results for both the RS126 and CB513 datasets

酸总数的比值. 这个评分只依赖 3 个状态 (螺旋/折叠/卷曲), 因此它的名字是取 Q_3 . 在整个的三态预测正确的残基准确度的定义 (Q_3) 是测量预测性能的.

CPM 的每一层的结果如表 3 和 4 所示. 同时, 我们与最好的 9 个二级结构预测的方法 (包括 PSIPRED^[21], SSPRO^[22], SAM-T02^[2], PHD Expert^[23], PROF^[24], JPRED^[25]) 在 RS126 和 CB513 数据集上进行对比实验. 实验结果显示在图 7 所示. 对于数据集 CASP8, 我们选择预测结果最好的几个方法作为对比对象. 单体实验结果如图 8 所示.

图 8 在 CASP8 数据集上 CPM 与其他 4 种方法的个体分布预测结果对比

Figure 8 Comparison with the results of 4 methods using the CASP8 dataset

对典型的研究文献回顾: Hu 使用 SVM 方法^[26]在 RS126 数据集上使准确度达到 78.8%; Xie^[27]使用神经网络分别在 RS126 和 CB513 数据集上得到了 79.65% 和 69.11%; CHEN^[28]使用层次的神经网络在 RS126 数据集上得到 74.38% 的准确度; Chopra^[29]使用细胞自动机方法分别在 RS126 和 CB513 上得到了 58.21% 和 56.51% 的准确度; liu^[30]使用文本分析的方法分别在 RS126 和 CB513 数据集上得到了 69.8% 和 69.6% 的准确度; Guo^[31]使用了双层 SVM 方法在 CB513 数据集上得到了 75.2% 准确度; Wang^[32]使用了优化的 SVM 方法在 CB513 数据集上得到了 78.44% 的准确度. 现在我们可以看到: CPM 的预测精度比其他方法都要高.

最新的研究思路曾吸引了不少研究人员使用混合挖掘算法. CDM^[33], 它结合 GORV 及分类数据挖掘方法. 但它与 CPM 不同:

- 1) CPM 的方法是一种普适的预测系统, 它能使用在不同类型的蛋白质二级结构预测中; 而混合两种方法的 CDM 方法只能预测混合型蛋白质的结构.
- 2) CPM 的方法是一个自动系统, 而 CDM 则需要使用者的主观体验.
- 3) CDM 是平面铺展的, 而 CPM 是逐步精化的立体结构. 需要指出的是, 我们可以进一步优化和改善 CPM, 使得预测结果更准确.

5 结论

综合复合金字塔模型各层功能与方法, 分析可知基于复合金字塔模型的预测系统具有以下创新性:

- 1) 该项研究成果具有完全的自主性, 即所有智能预测系统模型中涉及的模型、算法均为我们原创性或改进型的, 绝无一点利用外界服务器提供的任何输出成果; 并且通用于各典型蛋白质数据库和任何蛋白质个体的测试.
- 2) 该项研究属于蛋白质二级结构的智能预测系统模型的研究, 不是当今国际上主流发展中单纯预测方法的研究. 它涵盖了模型、方法、优化等核心构件. 这是迄今为止, 在解决蛋白质二级结构预测这

一重大难题中, 我们力图实现的学术思想、技术路线和方法论上的突破。

3) 就模型而论: 通过文献跟踪与查询检索, 尚未发现形如“合成金字塔”式的多层递阶、综合集成、逐步求精的模型构造; 也未见过在每一层中, 均采用物化属性判定与结构序列判定相结合的构建线路; 特别是采用了笔者 1 独立提出的基于结构和基于属性的原创性理论 KDTICM, 对构造高准确度的关联分类器形成有效支撑。

4) 就方法而论: 是彼此联系、无缝对接、走向有序、相互融合的若干预测方法构成的方法群, 故而称之为系统方法。所涵盖的方法中, 有的是原创性的 (如: 基于 KDTICM 理论的 KDD* 模型与 M 算法), 有的是对现有方法做出重要改进 (如: 基于动态规划的同源性分析、优化的 SVM 类化分析); 形成了一个高内聚、低耦合、紧密协同的预测方法体系。特别是利用我们独立提出的基于知识发现的因果自动机理论, 寻求与优化影响蛋白质空间构象的物化因素, 这样深入到机理层面为其物化属性预测法奠定了坚实的基础。

5) 就优化而论: 在系统模型自下而上的演绎中, 各层的粒度空间由粗到细。这种分层结构的设置, 体现了优化的迭代、渐近细化的粒度空间构建思想, 也体现了领域知识与背景知识的贯穿性。

6 展望

复合金字塔模型及其以创新性知识发现技术为主线的智能预测系统, 是蛋白质数据库上的蛋白质二级结构预测通用、典型的智能预测系统, 并能取得很好的预测效果。对于这阶段性研究成果, 尚存在可在优化层进一步提高预测准确度的空间。应强调的是, RS126, CB513 数据集准确度都超过 80%, 均是目前国际范围内所知最高。生物信息学领域普遍认为: 当蛋白质二级结构预测准确度达到 80% 以上时, 其三级结构是基本可知的; 因此基于我们所提出的预测系统, 将有望在蛋白质三级结构预测方面取得较大突破; 并且此模型及其系统构造对于类似问题具有普适意义。CPM 服务器较其他蛋白质二级结构预测服务器有更好的性能, 颇具推广价值。为此, 我们开通了 CPM 开放式接入网络自动服务网站。可以随时访问 <http://kdd.ustb.edu.cn/protein-Web/>, 获取在线服务。

参考文献

- 1 Hua S J, Sun Z R. A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach. *J Mol Biol*, 2001, 308: 397–407
- 2 Karplus K, Karchin R, Draper J, et al. Combining local-structure, foldrecognition, and new-fold methods for protein structure prediction. *ProteinsMayo*, 2003, 53: 491–496
- 3 Haoudi A, Bensmail H. Bioinformatics and data mining in proteomics. *Expert Rev Proteom*, 2006, 3: 333–343
- 4 Li J Y, Wong L S, Yang Q. Data mining in bioinformatics. *IEEE Intell Syst*, 2005, 20: 16–18
- 5 Wu X, Jain L, Wang J, et al. *Data Mining in Bioinformatics*. Berlin: Springer, 2005
- 6 Preisach C, Schmidt-Thieme L. Relational ensemble classification. In: *Proc of IEEE International Conference on Data Mining ICDM Optical Fiber Sensors Conference (APOS)*, Hong Kong, 2006
- 7 Lin H N, Chang J M, Wu K P, et al. HYPROSP II—a knowledge-based hybrid method for protein secondary structure prediction based on local prediction confidence. *Bioinformatics*, 2005, 21: 3227–3233
- 8 Yang B R. *Knowledge Discovery Theory Based on Inner Cognitive Mechanism*. Beijing: Electron Industry Publishment, 2004
- 9 Yang B R, Shen J T, Song W. KDK based double-basis fusion mechanism and its process model. *Int J Artif Intell Tool*, 2005, 14: 399–423
- 10 Yang B R, Xiong F. KD(D&K) and double - bases cooperating mechanism. *J Syst Eng Electron*, 1999, 10: 48–54

- 11 Yang B R, Hou W, Zhou Z. KAAPRO: an approach of protein secondary structure prediction based on KDD* in the compound pyramid prediction model. *Expert Syst Appl*, 2009, 36: 9000–9006
- 12 Rost B, Sander C. Prediction of secondary structure at better than 70% accuracy. *J Mol Biol*, 1993, 232: 5840–5899
- 13 Cuff J A, Barton G J. Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. *Proteins Struct Func Genet*, 1999, 34: 508–519
- 14 Protein Structure Prediction Center. CASP8 (2008). <http://predictioncenter.org/>
- 15 Yang B R, Sun H H, Xiong F L. Mining quantitative association rules with standard SQL queries and its evaluation. *J Comput Res Dev*, 2002, 39: 307–312
- 16 Park J, Karplus K, Barrett C, et al. Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J Mol Biol*, 1998, 284: 1201–1210
- 17 Altschul S F, Gish W, Miller W, et al. Basic local alignment search tool. *J Mol Biol*, 1990, 215: 403–410
- 18 Altschul S F, Madden T L, Schäffer A A, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, 1997, 25: 3389–3402
- 19 Mark K, Ashutosh T, Jifrn M. A review of soft computing applications in supply chain management. *Appl Soft Comput J*, 2010, 10: 661–674
- 20 Teixeira C A, Pereira W C A, Ruano A E, et al. On the possibility of non-invasive multilayer temperature estimation using soft-computing methods. *Ultrasonics*, 2010, 50: 32–43
- 21 Jones D. Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol*, 1999, 292: 195–202
- 22 Baldi P, Brunak S, Frasconi P, et al. Bidirectional dynamics for protein secondary structure prediction. In: *Proceedings of the 16th International Joint Conference on Artificial Intelligence (IJCAI99)*, Stockholm, 1999
- 23 Rost B, Sander C, Schneider R. PHD-an automatic mail server for protein secondary structure prediction. *Comput Appl Biosc*, 1994, 10: 53–60
- 24 Ouali M, King R. Cascaded multiple classifiers for secondary structure prediction. *Protein Sci*, 2000, 9: 1162–1176
- 25 Cuff J, Clamp M, Siddiqui A, et al. JPRED: a consensus secondary structure prediction server. *Bioinformatics*, 1998, 14: 892–893
- 26 Hu H J, Pan Y, Harrison R, et al. Improved protein secondary structure prediction using support vector machine with a new encoding scheme and an advanced tertiary classifier. *IEEE Trans NanoBiosci*, 2004, 3: 265–271
- 27 Xie X, Yang B, Chen Y H. Prediction of secondary structure of protein using neural network. *J Jinan Univ(Sci Technol)*, 2008, 22: 111–115
- 28 Chen J M, Chaudhari N S. Cascaded bidirectional recurrent neural networks for protein secondary structure prediction. *IEEE/ACM Trans Comput Biol Bioinform*, 2007, 4: 572–582
- 29 Chopra P, Bender A. Evolved cellular automata for protein secondary structure prediction imitate the determinants for folding observed in nature. In *Silico Biol*, 2007, 7: 87–93
- 30 Liu Y, Carbonel J, Klein-Seetharaman J, et al. Context sensitive vocabulary and its application in protein secondary structure prediction. In: *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Sheffield: ACM, 2004. 538–539
- 31 Guo J, Chen H, Sun Z R, et al. A novel method for protein secondary structure prediction using dual-layer SVM and profiles. *Proteins*, 2004, 54: 738–743
- 32 Wang L H, Liu J, Li Y F, et al. Predicting protein secondary structure by a support vector machine based on a new coding scheme. *Genome Inform*, 2004, 15: 181–190
- 33 Cheng H T, Sen T Z, Jernigan R L, et al. Consensus data mining (CDM) protein secondary structure prediction server: combining GOR V and fragment database mining (FDM). *Bioinformatics*, 2007, 23: 2628–2630

A new intelligent prediction system model-the compound pyramid model

YANG BingRu*, QU Wu, WANG LiJun & ZHOU Ying

School of Information Engineering, University of Science and Technology Beijing, Beijing 100083, China

*E-mail: bryang_kd@yahoo.com.cn

Abstract A current development trend in research on intelligent systems is to optimize a general intelligent prediction system into an individuation intelligent prediction system that is applied in specialized fields. Protein structure prediction is a challenging international issue. In this paper, we propose a new intelligent prediction system model, designed as a multi-layer compound pyramid model, for predicting secondary protein structure. The model comprises four independent intelligent interfaces and several knowledge discovery methods. The model penetrates throughout the domain knowledge, with the effective attributes chosen by Causal Cellular Automata. Furthermore, a high pure structure database is constructed for training. On the RS126 dataset, the overall state per-residue accuracy, Q_3 , reached 84.31%, while on the CB513 dataset, Q_3 reached 86.78%. Meanwhile, on the CASP8 sequences, the results are superior to those produced by other methods, such as Psipred, Jpred, APSSP2 and BehairPred. These results confirm that our method has a strong generalization ability, and that it provides a model for the construction of other intelligent systems.

Keywords knowledge discovery, intelligent prediction, system model, compound pyramid model

YANG BingRu was born in 1943. He serves in University of Science and Technology Beijing as a Lifetime Chief Professor, Principal Level, Ph.D. Supervisor of School of Computer and Communication Engineering and dean of Institute of Knowledge Engineering. Prof. Yang has always been engaging in teaching and researching in the field of modern mathematics and computer science and technology. He is not only

a specialist who obtained State Council Special Allowance, but also a certified international science and technology specialist, Master Teacher of Beijing, member of expert advisory group of Country Technology Basement Condition Platform Construction, accreditation expert of Ministry of Education Degree Center, member of China Educational Expert Committee, National 863 High-tech Project, National Natural Science Fund Project, accreditation expert of science and technology projects of Ministry of Education.

QU Wu was born in 1981. He received the Ph.D. degree in 2012 from the School of Information Engineering, Beijing University of Science and Technology. Currently, he is a Post Doctor at Knowledge Engineering Group, Tsinghua University. His research interests include data mining, social network analysis, cloud computing, and natural language processing. Dr. Qu is a member of China Computer Federation (CCF).

WANG LiJun was born in 1978. She received the Ph.D. degree in computer application from the Beijing University of Science and Technology, in 2012. Currently, she is a Post Doctor at Institute of Scientific and Technical Information of China. Her research interests include knowledge discovery, data mining, and intelligent retrieval.