CFW 的 CBR 动态预测

杨振刚,刘伟章,方永美
YANG Zhen-gang,LIU Wei-zhang,FANG Yong-mei

华南农业大学 信息学院,广州 510642
College of Informatics,South China Agricultural University,Guangzhou 510642,China

YANG Zhen-gang,LIU Wei-zhang,FANG Yong-mei.CBR dynamic forecast for CFW.Computer Engineering and Applications, 2009,45(6):236-239.

Abstract: This research presents a method for Cucumber Fusarium Wilt(CFW) dynamic forecast combining Case-Based Reasoning(CBR) methodology.A case indexing mechanism guided by idea of vantage case is developed to rapidly generate a similar case set for a new case and speed up case retrieval process,and an idea of sensitivity analysis for determining optimal similar case is used to construct the reasoning algorithm in this CBR forecast system.By analyzing performances of the exhaustive search and the search using proposed indexing mechanism,the range of optimal cluster number for this application is inferred.The precision and recall evaluations are conducted for each testing case using the X fold cross-validation approach,the reasoning effectiveness employing different thresholds of dissimilarity distance (R) is figured out and the optimal R for this CBR system is determined.

Key words: Case-Based Reasoning(CBR);Cucumber Fusarium Wilt(CFW);dynamic forecast

摘 要: 结合 CBR(Case-Based Reasoning,基于案例推理)方法学,探索了 CFW(Cucumber Fusarium Wilt,黄瓜枯萎病)动态预测技术。提出一种优势案例机制辅助案例检索,快速定位相似案例集以提高检索效率,并借助灵敏度分析思想确定最优相似案例。对遍历检索及基于优势案例机制的检索进行了对比分析,确定了系统案例库的最优分类数范围。利用交叉验证方法,对每个测试案例进行准确度及联想特性值评价,得出不同相异阈值下推理算法的推理有效性,并依此确定了系统案例检索的最优相异阈值。

关键词: 基于案例推理;黄瓜枯萎病;动态预测
DOI: 10.3778/j.issn.1002-8331.2009.06.068 **文章编号:** 1002-8331(2009)06-0236-04 **文献标识码:** A **中图分类号:** TP18

传统的黄瓜病害预测是依靠经验进行人工预测或根据预设模型利用软件工具进行预测,其缺点是不能根据实时气候及生态条件动态地对作物病害进行智能预测。因此,提出集成 CBR(Case-Based Reasoning,基于案例推理)方法来研究 CFW(Cucumber Fusarium Wilt,黄瓜枯萎病)动态预测技术。CBR 方法是通过适应(修订)先前相似案例来解决新面对的问题[1-5],它不需要建立非常明确的作物病害知识模型,而收集病害预测历史案例及开发有效的推理算法就变得相对重要。CBR 系统要通过大量的案例来获取新知识[1,3,6-10],这是它适合于需要经验知识导向的作物病害预测领域的原因。本研究提出一种基于优势案例(Vantage Case)的索引机制来提高案例检索效率,借助灵敏度分析思想确定最优相似案例,并利用交叉验证方法分析了系统推理有效性。在 CBR 预测系统的协助下,CFW 可能发生的情况可以有效预知,从而增加防治胜算并能辅助生产决策的动态修订。

1 CBR 预测系统结构

CBR 预测系统结构如图 1 所示。系统根据作物生态特征及环境信息定时决策作物病害预测方案,或根据用户请求动态返回作物病害预测结果。CBR 执行过程如图 2 所示,具体包

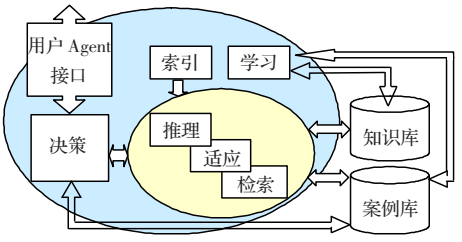

图 1 CBR 系统结构图

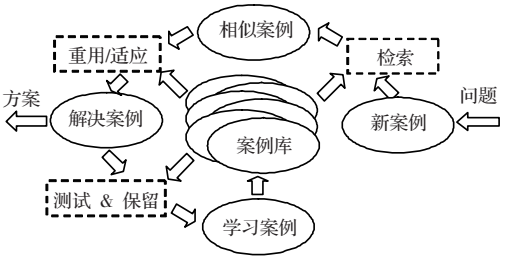

图 2 CBR 执行过程

基金项目: 2007 年广东省农业标准化项目;华南农业大学新学科扶持基金项目(No.2008X032)。
作者简介: 杨振刚(1976-),男,工学博士,讲师,主研领域:信息系统工程、人工智能。
收稿日期: 2008-01-14 **修回日期:** 2008-03-28

表 1　案例库

案例 ID	问题描述部分(ATR)				…	方案部分 1(PDT)		…	方案部分 2(STM)			…
	温度/(℃)	土壤 pH 值	平均菌量	土壤湿度		病株率/(%)	病叶率/(%)		药剂	药含量/(%)	倍液数	
1	25.5	6.8	107.19	82.5	…	23	18	…	多菌灵可湿性粉剂	50	500	…
2	27.6	6.9	78.56	86.3	…	19	15	…	多菌灵可湿性粉剂	50	500	…
3	24.3	6.9	50.34	90.2	…	17	10	…	多菌灵胶悬剂	40	400	…
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
432	23.5	7.1	19.63	80.6	…	11	4	…	双效灵水剂	10	200	…

括:从案例库中检索相似案例、重用相似案例并推断新情况下系统解决方案、修订解决方案、保存新案例以备后用[11-15]。

2　基于 CBR 的 CFW 预测原理

2.1　案例表达

案例表达是 CBR 系统最为关键的部分之一,它需要正确描述案例问题部分及方案部分的独特特征。设系统案例库 $CB=\{c_a|a=1,2,\cdots,n\}$,案例 c_a 问题描述部分属性集合 $ATR_a=\{atr_{a1},\cdots,atr_{al}\}$,其中 atr_{al} 为第 l 个问题描述属性。案例 c_a 的方案部分 $SLT_a=\{PDT_a,STM_a\}=\{\{pdt_{a1},\cdots,pdt_{aj}\},\{stm_{a1},\cdots,stm_{ah}\}\}$,其中 PDT_a 为预测属性集合,STM_a 为病害处理措施及其它案例描述内容集合。因此案例可以表示为 $c_a=\{ATR_a,PDT_a,STM_a\}$。

CFW 是世界性病害,是在中国各地都有发生的黄瓜重要病害之一。黄瓜产地土壤中的枯萎病病菌量会由于连年耕作而逐年积累增多,发病概率增大。CFW 的发生同时受自然环境、生态条件等因素影响,如气温暖和、土壤偏酸、土质粘重等都会助长病菌的生长。因此 CFW 案例的问题部分为影响病害发生的自然条件及生态条件,如温度、土壤 PH 值、土壤湿度、平均菌量、作物生长期、土壤线虫密度、品种特性和栽培方式(l=8)。方案部分的预测属性集合 PDT 包括病株率和病叶率(j=2),而处理措施及其它描述内容集合 STM 包括处理药剂、药含量、倍液数、添加剂、添加剂含量、添加剂倍液数、添加比例、剂量、使用方式、使用频数、使用次数、其它措施和备注(h=13)。案例库如表 1 所示。

2.2　案例索引

当案例较多时,CBR 系统检索过程非常耗时。为解决这个问题,建立一种基于优势案例的案例索引机制,以帮助相似案例的快速查找。优势案例代表一组相似案例,用它与新案例进行匹配,可减少检索时间。优势案例可用 K-means[16]、DBSCAN[17]、CURE[18]等方法来确定。索引机制根据作物的生态特征、环境条件利用 CURE 聚类算法把案例库分为 k 类案例 $CS_t(t=1,\cdots,k)$。定义每个 CS_t 的优势案例 c_t^* 为该子集中与其余案例相异距离和最小的案例,即:

$$c_t^*=\{c_a|\min[\sum_{c_b\in CS_t}dt(c_a,c_b)]\} \tag{1}$$

定义案例相异距离(dissimilarity distance)函数为:

$$dt(c_a,c_b)=\sqrt{\sum_{s=1}^{l}[w_s(atr_{as}-atr_{bs})^2]} \tag{2}$$

其中 w_s 为 ATR 中第 s 个属性的重要性权值,总体上温度、土壤 PH 值、平均菌量、土壤湿度等属性权值相对较大,因为它们对病害的发生有较大影响。由于 ATR 中属性的单位不同,在计算相异距离前要采用 Z-score 方法对属性进行标准化。其中描述性属性,如作物生长期、品种特性和栽培方式等,用预处理算法将其映射到[0,1]区间,再进行标准化。

根据式(2)对 CS_t 中的每个案例计算其与 c_t^* 的相异距离,并按相异距离升序用索引码重排列 CS_t 中案例,以有利于案例检索。

2.3　案例检索

当新案例 P 与案例 c_a 之间的相异距离小于等于预定义的相异阈值 R 时,系统认定 P 与 c_a 相似。定义子集半径 R_t 为 CS_t 中案例与其优势案例 c_t^* 的最大相异距离,即:

$$R_t=\max\{dt(c_a,c_t^*)|c_a\in CS_t\} \tag{3}$$

则案例检索算法描述如下:

(1)对于每个 $CS_t(t=1,\cdots,k)$,计算新案例 P 与其优势案例 c_t^* 的相异距离 $dt(c_t^*,P)$,如果 $dt(c_t^*,P)\geqslant R_t+R$,$P$ 与 CS_t 无需满足预定义阈值的相似案例,案例检索程序无需再对 CS_t 进行相似性评价;

(2)如果 $dt(c_t^*,P)<R_t+R$,根据式(4)确定与 CS_t 对应的候选案例集 CS'_t;

$$CS'_t=\{c_a\in CS_t|dt(c_t^*,P)+R\geqslant dt(c_t^*,c_a)\geqslant dt(c_t^*,P)-R\} \tag{4}$$

(3)计算 CS'_t 中每个案例与新案例 P 的相异距离,满足小于预设阈值 R 的案例保留在集合中,否则删除。重复步骤(1)~(3),直至所有 CS_t 都被计算;

(4)计算 P 的相似案例集 $CCS_P=\bigcup_{t=1}^{k}CS'_t$。

2.4　案例推理

设与新案例 P 的相似案例集 CCS_P 的案例个数为 N,即 $CCS_P=\{c_x|x=1,\cdots,N\}$,$P$ 经推理后得到的解决方案表达为 $c_P=\{ATR_P,PDT_P,STM_P\}$,定义最优相似案例为 CCS_P 中与 P 相异距离最小的案例,即

$$c_\#=\{c_x\in CCS_P|\min[dt(c_x,P)]\} \tag{5}$$

则案例推理算法描述如下:

(1)当 CCS_P 为空集时,问题 P 为异常案例,计算获取案例库中与 P 最小相异距离的案例作为对比案例返回结果并显示 CFW 相关描述内容,算法终止;

(2)如果 N=1,令最优相似案例 $c_\#=c_1$,转到步骤(7);

(3)进行最优相似案例灵敏度分析,即通过微调 ATR 中属性权重 w_s,分析最优相似案例的变化情况,为最优相似案例的确定提供决策依据。由于温度、土壤 pH 值、平均菌量以及土壤湿度对 CFW 的影响程度较大,故仅对其权重进行微调分析。设温度、土壤 pH 值、平均菌量以及土壤湿度的初始属性权重分别为 w_T、w_{PH}、w_{MP} 和 w_{SH},则每次权重同时变化为

$$\begin{cases}w_T^i=w_T+i\cdot\varepsilon\\ w_{PH}^i=w_{PH}-i\cdot\varepsilon\\ w_{MP}^i=w_{MP}+i\cdot\varepsilon\\ w_{SH}^i=w_{SH}-i\cdot\varepsilon\end{cases} \tag{6}$$

其中 i 与 f 均为整数,$f>0$,$i\in[-f,f]$,ε 为调节因子,且同时满足

$$\begin{cases}|f\cdot\varepsilon|<w_T/10\\|f\cdot\varepsilon|<w_{MP}/10\\|f\cdot\varepsilon|<w_{PH}/10\\|f\cdot\varepsilon|<w_{SH}/10\end{cases}\tag{7}$$

当 i 变化时，根据式(6)用 w_T^i、w_{PH}^i、w_{MP}^i和 w_{SH}^i 替代原先对应属性的 w_s 来计算 CCS_P 中案例与 P 的相异距离，确定对应于 i 的最优相似案例 $c_i^{\#}$；

(4)统计 CCS_P 中案例 c_x 成为最优相似案例的频度。令频度最高的案例为最优相似案例 $c_{\#}$。当出现多个案例频度最高时，令 min$|i|$对应的 $c_i^{\#}$ 为最优相似案例 $c_{\#}$；

(5)判断 CCS_P 中各案例的 stm_1 属性和 stm_4 属性与 $c_{\#}$ 的相应属性是否一致(即对案例解决方案中药剂处理方法的一致性判断)，一致的案例保留在 CCS_P 中，不一致的案例从 CCS_P 中删除；

(6)令此时的 CCS_P 案例个数为 N_2，当 N_2=1 时，取 PDT_P=$PDT_{\#}$(# 表示最优相似案例 $c_{\#}$)，当 N_2>1 时，推理 PDT_P 各属性值时考虑 CCS_P 中所有案例的贡献，采用加权平均法适应。设 CCS_P 中 N_2 个案例的权重为 $w_x^P(x=1,\cdots,N_2)$，定义

$$w_x^P=\frac{dt(c_{\#},P)/dt(c_x,P)}{\sum_{c_a\in CCS_P}[dt(c_{\#},P)/dt(c_a,P)]}\tag{8}$$

则 PDT 属性的适应值

$$pdt_{P_i}=\sum_{x=1}^{N}[w_x^P\cdot pdt_{xi}],i=1,2,\cdots,j\tag{9}$$

(7)推理 STM_P 时，先取 STM_P=$STM_{\#}$，推理器基于知识和 $STM_{\#}$ 的属性值对相应的案例属性进行适应，得到最终的STM_P，决策时除了提供案例处理方案，还可能提供其它 CFW 描述内容，如 CFW 症状表现、病原菌、病害传染途径、发病条件等，具体内容与案例输入特征情况相关。

3　系统应用分析

CBR 系统使用黄瓜产地提供的一年多实际数据(432 个案例，如表 1 所示)来实证其性能及有效性。通常在案例检索时相似性计算评价方面会花费较多时间，本文基于优势案例算法可提高检索效率，因为根据索引机制，部分案例经判断可能不需进行相似性评价。对于检索计算复杂性，假定遍历检索(Exhaustive Search)的计算复杂度为 $CC(n)$(n 为案例总数)，则基于优势案例算法的计算复杂度为 $CC(pn+k)$，其中 k 为案例聚类数，p 为事件 $dt(c_i^*,P)<R_t+R$ 的概率。k 的最优值通常介于[1,n]之间且远小于 n，故 $CC(pn+k)\approx CC(pn)$，从而计算复杂性低于遍历检索算法。为比较遍历检索及所提出的基于优势案例检索算法，在不同案例聚类数 k 下，计算两者检索相似案例的时间，结果如图 3 所示。由结果分析可知，案例聚类数在 9 到 36 之间时检索效率较好。

对于案例推理，其有效性通常用推理准确度及联想特性值来评价。为增强实验可靠性，采用交叉验证方法(X fold Cross-validation)[19]来评价系统推理有效性。在实验中，每个案例库案例被顺序选作测试案例，其余(n-1)案例作为训练案例集。对每个测试案例进行准确度及联想特性值评价，可统计得到平均准确度(AVP)及平均联想特性值(AVR)，而 CBR系统推理有效性定义为 Ef，可用 Ef=(2×AVP×AVR)/(AVP+AVR)[20]计算，Ef 值越大，推理有效性越高。在不同相异阈值 R 下对 AVP、AVR 和 Ef 进行分析，结果如图 4 所示。当 R 增加时，推理有效性 Ef 也将增加，但随之案例检索效率会降低，因而 R 最优值可选取当 Ef 曲线开始趋于平缓增长的 0.3。当 R 为 0.3 时，Ef、AVP 和 AVR 分别为 66.3%，83%和 57%。

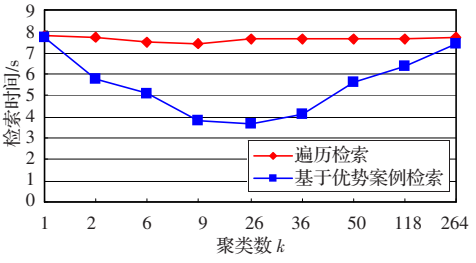

图 3　检索效率比较

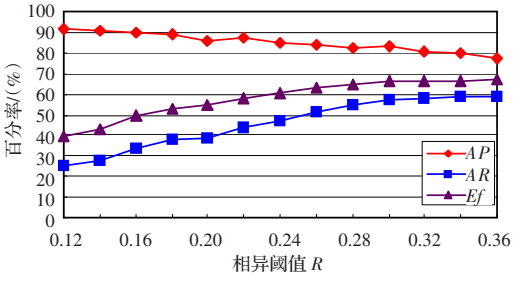

图 4　不同相异阈值下的推理有效性

4　结论

结合 CBR 方法学，探索了 CFW 动态预测技术。在预测过程中提出一种优势案例机制辅助案例检索，快速定位相似案例集以提高了检索效率，并借助灵敏度分析思想构建了推理算法。对遍历检索及基于优势案例机制的检索进行了对比分析，确定了本应用的最优分类数范围为 9 至 36。为增强实验可靠性，利用交叉验证方法，对每个测试案例进行准确度及联想特性值评价，得出不同相异阈值下所提出推理算法的推理有效性，并依此确定了本应用中案例检索的最优相异阈值为 0.3。性能分析及实际应用表明，所提出的基于 CBR 的病害预测方法可为 CFW 防治提供有用信息，并能有效协助生产决策的动态修订。

参考文献：

[1] Aamodt A,Plaza E.Case-based reasoning:foundational issues methodological variations and system approaches[J].Artificial Intelligence Communications,1994,7(1):39-59.

[2] Bichindaritz I,Kansu E,Sullivan K M.Case-based reasoning in CARE-PARTNER:gathering evidence for evidence-based medical practice[C]//Advances in Case-Based Reasoning:Proceedings of the 4th European Workshop on Case-Based Reasoning.Berlin:Springer-Verlag,1998:334-345.

[3] Krampe D,Lusti M.Case-based reasoning for information system design[C]//Proceedings of the 2nd International Conference on Case-Based Reasoning,1997:63-73.

[4] Bradburn C,Zeleznikow J.The application of case-based reasoning to the tasks of health care planning[C]//Topics in Case-Based Reasoning: Proceedings of the 1st European Workshop on Case-Based Reasoning.Berlin:Springer-Verlag,1993:365-378.

[5] Arts R J,Rousu J.Towards CBR for Bioprocess planning[C]//Advances in Case-Based Reasoning:Proceedings of the 3rd European Workshop on Case-Based Reasoning.[S.l.]:Springer-Verlag,1996:16-27.
[6] 徐晓臻,高国安.案例推理在多准则评价智能决策支持系统中的应用研究[J].计算机集成制造系统,2001,7(1):16-18.
[7] 张荣梅,涂序彦.基于 CBR 的交通事故处理智能决策支持系统[J].计算机工程与应用,2002,38(2):247-249.
[8] 黄继鸿,姚武,雷战波.基于案例推理的企业财务危机智能预警支持系统研究[J].系统工程理论与实践,2003(12):46-52.
[9] 汪季玉,王金桃.基于案例推理的应急决策支持系统研究[J].管理科学,2003,16(6):46-51.
[10] 路云,吴应宇,达庆利.基于案例推理技术的企业经营决策支持模型设计[J].中国管理科学,2005,13(2):81-87.
[11] Watson I.Applying case-based reasoning: techniques for enterprise system[M].San Francisco,CA:Morgan Kaufmann,1997.
[12] Leake D B.Case-based reasoning:experiences,lessons,and future directions[M].[S.l.]:AAAI Press/MIT Press,1996.
[13] Reisbeck C K,Schank R C.Inside case based reasoning[M].[S.l.]:Lawrence Erlbaum,1989.
[14] Kolodner J.Case-based reasoning[M].[S.l.]:Morgan Kaufmann,1993.
[15] Yang B S,Han T,Kim Y S.Integration of ART-Kohonen neural network and case-based reasoning for intelligent fault diagnosis[J].Expert Systems with Applications,2004,26:387-395.
[16] MacQueen J.Some methods for classification and analysis of multivariate observations[C]//Proceedings of the 5th Berkeley Symposium on Mathematics,Statistics and Probability,Berkeley,CA,1967:281-297.
[17] Ester M,Kriegel H,Sander J,et al.A density-based algorithm for discovering clusters in large spatial databases[C]//Proceeding of the 2nd International Conference on Discovery and Data mining,Portland,OR,1996:226-231.
[18] Gehrke J,Ramakrishnan R,Rainforest V G.A framework for fast decision tree construction of large datasets[C]//Proceedings of the 24th International Conference on Very Large Databases,New York,1998:416-427.
[19] Witten I H,Frank E.Data mining:practical machine learning tools and techniques with Java implementations[M].San Francisco,CA:Morgan Kaufmann,1999.
[20] Lewis D.Evaluating text categorization[C]//Proceedings of a Workshop on Speech and Natural Language,Pacific Grove,CA,1991:312-318.

(上接 215 页)

参考文献:

[1] 陈彬,洪家荣,王亚东.最优特征子集选择问题[J].计算机学报,1997,20(2):133-138.
[2] Barzilay O,Brailovsky V L.On domain knowledge and feature selection using a support vector machines[J].Pattern Recognition Letters,1999,20(5):475-484.
[3] Cao L J,Tay F E.Feature selection for support vectormachines in financial time series forecasting[C]//Proceedings Second International Conference on Intelligent Data Engineering and Automated Learning:Data Mining,Financial Engineering,and Intelligent Agents.[S.l.]:Springer-Verlag,2000.
[4] 任江涛,赵少东.基于二进制 PSO 算法的特征选择及 SVM 参数同步优化[J].计算机科学,2007,34(6):179-182.
[5] 智强,杨梅.景观设计概论[M].北京:中国轻工业出版社,2006:89-93.
[6] 安秀.公共设施与环境艺术设计[M].北京:中国建筑工业出版社,2007:65-78.
[7] 卢新海,杨祖达.园林规划设计[M].北京:化学工业出版社,2006:45-60.
[8] 李铮生.城市园林绿地规划与设计[M].北京:中国建筑工业出版社,2005:93-110.
[9] Kennedy J,Eberhart R C.Particle swarm optimization[C]//Proc IEEE Int Conf on Neural Networks.Pscataway,NJ:IEEE Service Center,1995:1942-1948.
[10] Kennedy J,Eberhart R C.A discrete binary version of the particle swarm algorithm[C]//Proc 1997 Conf on Systems,Man and Cybernetics.Piscataway,NJ:IEEE Press,1997:4104-4109.
[11] 苑敏,杨奎河.基于支持向量机理论的多类分类算法[J].福建电脑,2007(2):9-10.

(上接 220 页)

[7] 韩敏,林丽玉.基于神经网络集成的蛋白质二级结构预测模型[J].计算机与应用化学,2006,23(10):959-961.
[8] Haykin S.神经网络原理[M].叶世伟,史忠值,译.北京:机械工业出版社,2004:109-178.
[9] 贾光峰,任爱华,吴强,等.基于多表达式编程的汇率预测的研究[J].济南大学学报:自然科学版,2008,22(1):77-80.
[10] Cuff J A,Barton G J.Evaluation and improvement of multiple sequence methods for protein secondary structure prediction[J].PROTEIN:Structure,Function,and Genetics,1999,34:508-519.
[11] Jones D T.Protein secondary structure prediction based on position-specific scoring matrices[J].J Mol Biol,1999,292:195-202.
[12] Pagni M,Cerutti L,Bordoli L.An introduction to patterns,profiles[C]//HMMs and PSI-BLAST,2003.
[13] 丁永生.计算智能-理论,技术与应用[M].北京:科学出版社,2004:64-114.
[14] Lee B C,Kim D.New design of neural network input and output vectors in the protein secondary structure prediction[J].Bioinformatics and Biosystems,2006,1(4):82-90.
[15] Yuksektepe F U,Yilmaz O,Turkay M.Prediction of secondary structures of proteins using a two-stage method[J].Computers & Chemical Engineering,2008,32:78-88.
[16] Rost B.Review:protein secondary structure prediction continues to rise[J].Journal of Structural Biology,2001,134:204-218.
[17] Salamov A A,Soloveyev V V.Protein secondary structure prediction using local alignments[J].Journal of Molecular Biology,1997,268:31-36.
[18] King R D,Sternberg M J E.Identification and application of the concepts important for accurate and reliable protein secondary structure prediction[J].Protein Science,2006,5:2298-2310.
[19] Frishman D,Argos P.Seventy-five percent accuracy in secondary structure prediction[J].Proteins:Structure,Function and Genetics,1997,27:329-335.
[20] Zvelebil M J,Barton G J,Taylor W R,et al.Prediction of protein secondary structure and active sites using the alignment of homologous sequences[J].Journal of Molecular Biology,1987,195:957-961.
[21] Cuff J A,Barton G.Evaluation and improvement of multiple sequence methods for protein secondary structure prediction[J].Proteins,1999,34:508-519.