

基于蛋白质进化配对的残基间距离约束挖掘方法

王彩霞¹ 吕 强² 李海鸥¹ 罗 升¹

(苏州大学计算机科学与技术学院 苏州 215006)¹ (江苏省计算机信息处理技术重点实验室 苏州 215006)²

摘 要 蛋白质的进化配对是指在进化过程中残基对之间形成的相对稳定的相互作用。基于已被发现的进化配对,采用机器学习的分类技术,将其转换成残基对之间的距离约束,从而将一种定性的残基对之间的相互作用挖掘为定量的距离约束,这为蛋白质的结构预测提供了新的指导。

关键词 进化配对,距离约束,支持向量机

中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.7.056

Mining Residues Distance Constraints from Protein Evolution Couplings by Classification

WANG Cai-xia¹ LV Qiang² LI Hai-ou¹ LUO Sheng¹

(School of Computer Science and Technology, Soochow University, Suzhou 215006, China)¹

(Provincial Key Laboratory for Computer Information Processing Technology of Jiangsu, Suzhou 215006, China)²

Abstract Protein evolution couplings refer to the relatively stable interactions between residues formed in the process of evolution. Based on the known evolution couplings, we adopted machine learning classification technique to convert evolution couplings into distance constraints, thus quantitative residues distance constraints are extracted from qualitative protein evolution couplings between residues, which can be used as a new guidance for the prediction of protein structure.

Keywords Evolution coupling, Distance constraint, SVM

1 引言

作为生物大分子,蛋白质的空间结构不是由独立的残基位置决定的,残基间的相互关系对结构组成具有重大影响。这种关系会在序列中以某种形式留下一些痕迹,一般这些痕迹在经过长时间的进化、变异后,单一蛋白质上的真实序列信息会被一些噪声掩盖,但是对同一蛋白质家族中序列进行多重序列比对后,能够很好地剔除存在于单一序列的噪声,从而获得同一家族中的蛋白质共同的属性。2011 年, Hopf 等人提出一个大胆的实验,即用 EVfold^[1] 计算出残基间的变异关系,然后应用到蛋白质结构预测中,该方法通过分析蛋白质家族序列比对文件来预测蛋白质空间结构上残基与残基之间存在的联系,进而预测出蛋白质的三维结构^[2]。这种联系被 Hopf 等人定义为进化配对,简称为 EC (Evolution Couplings)。尽管对于残基对间相互联系的研究^[3] 不断扩展,但是从序列信息来预测残基约束仍是一个难题,在如何提高对残基对间联系的预测准确度也一直是一个在不断探究的课题。图 1 是 EC 作用的示意图,一对在序列上相距 195 个残基的残基对由于 EC 的作用在空间中仅仅相距 4 Å。

目前对于蛋白质进化约束的预测方法可粗略分为两类^[4]: 纯序列法和监督学习法。本文中用到的 EC 都是从第

一种方法(即 EVfold 方法)获取的,这种方法完全依赖于蛋白质序列信息,通过序列比对获取同家族中的共性。序列比对^[5] 是一种相似度比较高的比对,但并不是完全比对,这就表明在比对过程中会产生细微误差;在获取 EC 的同时,得到的是整个家族大致符合的残基对,在这一过程中保证了整体的高符合度,但是并不能确保个体高符合度。所以在 EC 的获取过程中会存在误差,同时也会得到大量冗余数据,这些误差和冗余数据将导致所获取的 EC 并不完全准确,因此需要在具有较多误差的数据中挑选出相对准确的进化配对。本文使用 SVM 来进行分类,并得到相对准确的残基对。如何将残基间的这种相互关系转化为计算机能够理解的定量的距离约束,将影响到结构预测的准确性,这也是本文的主要内容。

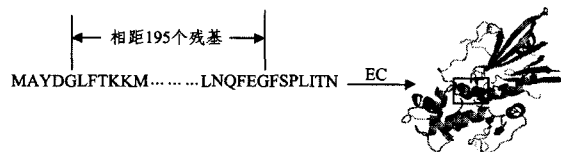


图 1 案例 3DOA(序列长度 $L=279$) 中在序列上相距 195 个残基,由于进化配对关系,形成三维结构反映到空间距离上为图中方框部分,两者空间距离仅仅为 4 Å

到稿日期:2014-08-25 返修日期:2014-12-08 本文受国家自然科学基金(61170125)资助。

王彩霞(1987—),女,硕士生,主要研究方向为生物信息计算,E-mail:20124227016@suda.edu.cn;吕 强 男,教授,博士生导师,主要研究方向为生物信息计算、元启发搜索、并行计算;李海鸥(1988—),女,博士生,主要研究方向为生物信息计算;罗 升(1991—),男,硕士生,主要研究方向为生物信息计算。

由于蛋白质是生物大分子,结构非常复杂,而且蛋白质结构中存在大量的自由度,导致构象的搜索空间巨大,对从头预测产生很大阻力^[6]。如果能够提前加入一些约束,将会大大缩小搜索空间,并能确定蛋白质的一些局部结构,这对蛋白质结构的预测具有重要影响。因此如果能将相对准确的 EC 转化为距离约束,在将距离约束应用到蛋白质结构预测中时,预测的结果将更加接近蛋白质的天然结构^[7]。本文的主要任务是根据已得到的 EC,辅以序列属性,挖掘出残基对间的远距离(相距 12 个残基以上)约束。在蛋白质中按残基间隔大小可以将距离约束分为 3 类^[4]:短距离约束(short-range),残基间隔比较短的距离约束,一般来说距离范围在[6,12]之间,此类约束对蛋白质结构预测影响不大;中等距离约束(medium-range),此类约束的残基序列距离在(12,24]之内;长距离约束(long-range),一般相隔 24 个残基以上,在本文中体现在距离为 8Å–10Å 的 3 部分。

2 材料与方法

2.1 数据准备

蛋白质结构预测评估大会(Critical Assessment of protein Structure Prediction,CASP)^[8]在世界范围内对蛋白质结构预测技术进行评估,评价当前蛋白质结构建模方法的能力及局限性。对于比较难的自由建模(FM)中的蛋白质目标,在 CASP 比赛中基于序列的从头预测模型预测 contact 的准确度要高于模板建模精度^[9]。本文是从 CASP 8、CASP 9 和 CASP 10 中预测相对较难的 FM 中选取了 27 条蛋白质作为数据集。由于要预测 EC 的准确性,首先要保证每条蛋白质的 EC 是可以获取的,因此选取的蛋白质都是已被 PDB 网站收录的。EC 的获取,是采用 Hopf 等提供的源代码^[2]在 Matlab 平台下得到的,一般认为两个残基在序列上距离越近,那么 EC 值就越大,空间距离也较小,从而残基间的联系也就越密切。但也有些特殊的情况,即在序列上距离比较远,但是 EC 值却相对较高,这类情形有一定的研究意义,因为如果这个 EC 是正确的,就表明序列相距较远的这对残基在空间上可能相距较近,如果预测准确,就可以指导蛋白质的折叠方式,从而使蛋白质空间结构的预测更加准确。

2.2 特征提取

本文从蛋白质的一级序列获取三级结构信息,提取的特征更加侧重来自于氨基酸序列导出的特征。对于二级结构^[10]这类需要进一步预测的属性,因为在取得这些属性的过程中就会有误差存在,可能导致后面的预测更加不准确,因此选择抛弃这样的特征属性。将 CASP 上 FM 对象中的 27 条蛋白质作为数据集,提取它们的序列特征。氨基酸序列包含蛋白质的整体信息,在自然情况下氨基酸序列所包含的残基数量是不同的,因此可以把不同长度的氨基酸序列转换成等长的残基频数,即统计每个残基在该蛋白质序列中的频数作为前 20 个属性,这 20 个属性值的总和即这条氨基酸序列的长度。其次是 EC 所在残基对以及 EC 值作为后面的特征值。然而,如果仅仅计算残基的频数而没有可以组成或者跨越至少一个片段的相邻残基信息,就不能明确残基对的周围环境情况,导致无法很好地反映序列隐含的结构信息。为了使得到的残基属性既能反映序列的残基成分又能反映残基间的相邻信息,本文的属性采用了全局属性和局部属性两部分。由

于 profile 是在无法获取蛋白质三维结构时采用相关序列比对的方式得到远相似性矩阵,因此它是一种能够较好地反应蛋白质序列方面的信息的属性。实际上,profile 就是对每个残基的评估分数,如果一个残基高度保守在某一位置,那么这个分数就是比较大的正数,反之是一个绝对值较大的负数^[11]。本文进行了归一化处理,通过调用 BLAST 程序计算所有残基的 profile,对于每一对残基,取其对应前后 3 个残基的 profile。由于蛋白质主要由 20 种常见残基组成,在计算 profile 过程中,每个残基都会产生与 20 种常见残基相似的比率,即当前基在重组过程中可能会变成其他残基的概率。profile 中包含几乎所有的蛋白质性质,本文采用统计每对残基对前后的 3 个残基 profile 作为局部特征。

大部分蛋白质是由 20 种常用氨基酸组成,根据 SVM 数据集的特性,需要将所有氨基酸字母转换成对应的数字。这样就组成了一个共有 265 个属性的数据集:以 20 种氨基酸的统计频数作为整体特征(20 个特征属性),以残基对每个残基前、后 3 个残基的 profile 作为局部特征(240 个特征属性),还包括每对残基在序列中的序列号及残基的名字和 EC 值(5 个特征属性)。对于每条蛋白质的数据集文件,首先取出 EC 相对准确的记录,EC 的准确性是根据记录中的一对氨基酸的几何距离判断,如果几何距离在 10Å 以内,则认为是相对准确的记录,全部留用,超过 10Å 的部分按 EC 从大到小排序,取前 30%的记录作为参考。

对于不同的蛋白质其序列长度不同,而不同的序列长度产生的 EC 对数目也不同,同样不同蛋白质计算出的几何距离对数目也不同,所以每条蛋白质序列比对得出的记录条数也不同。27 条序列产生的记录总数为 43980 条,抽出任何一条蛋白质的所有记录作为测试集,其他作为训练集,每条记录含有 265 个序列属性和 1 个分类属性,分类属性根据几何距离标为 0、3、4、...、10 共 9 个分类标签。训练数据生成流程如图 2 所示。

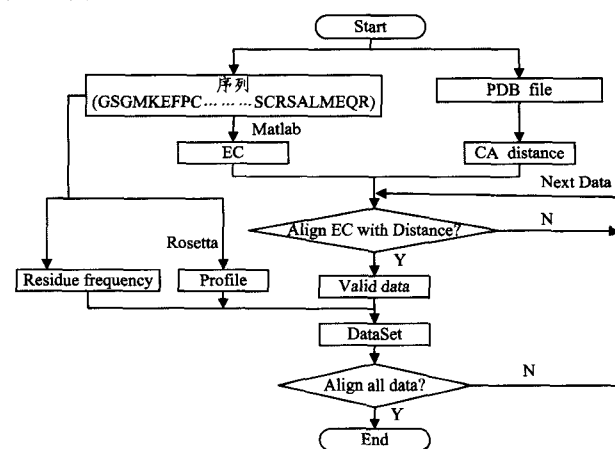


图 2 数据生成流程

2.3 训练过程

对 EC 进行分类,目的在于预测每条蛋白质在不同分类情况下的准确率。由于把分类标签分为了 9 个,因此对 EC 的预测是一个多元分类问题。要确定 SVM,需要选择合适的核函数,现在较常用的核函数主要有以下 4 种:线性核、多项式核、径向基(RBF)核、Sigmoid 核。在这 4 种核函数中,RBF 核函数适用范围最为广泛,并且具有较宽的收敛域,是较为理想的分类依据函数。此外,线性核是 RBF 核的一个特例。

Sigmoid 核不是正定核,而且对于一定的参数取值,Sigmoid 核和 RBF 的性能相当。相对来说,多项式核的参数比较多,而且它在运算时间上可能会溢出。相比较而言,RBF 核函数就具有相对的优势。

本文采用 libsvm 分类器来进行预测,选择了 RBF 核函数来把输入空间的向量映射到高维特征空间中:

$$K(x_i, x_j) = \exp(-\gamma(\|x_i - x_j\|)^2), \gamma > 0 \quad (1)$$

γ 是一个影响高维空间中的特征向量的坐标的参数。同时使用了 C-SVM 样本分类器:

$$\begin{cases} \min \frac{1}{2} \omega^T \omega + C \sum_{i=1}^l \xi_i, & C \geq 0 \\ y_i(\omega^T \phi(x_i) + b) \geq 1 - \xi_i, & \xi_i \geq 0 \end{cases} \quad (2)$$

其中, C 是惩罚参数。为解式(2),需要调节式(1)中的参数 γ 和式(2)中的参数 C 。对于参数选择问题,在实际应用中一般采用取试探法、网格法(Grid Search)或启发式的选参方法等。本文在 libsvm 中可以调用命令 grid.py 来训练这两个参数,使用逐格搜索的方法来确定最优 C 和 γ 。libsvm 采用特有的对数取值法,默认参数 C 的取值范围为 $2^{-15} \sim 2^5$, $\log_2 C$ 的步长为 1;默认参数 γ 的取值范围为 $2^{-15} \sim 2^5$, $\log_2 \gamma$ 的步长为 1,交叉验证为 5 重。经过 grid.py 优化后,得到 C 和 γ 的值分别是 512.0 和 3.0517578125e-05。

3 结果与分析

根据训练得到的参数,同时可以训练得到一个 SVM 多类分类模型,然后对测试集中的 EC 进行准确性预测。调用 libsvm 中的可执行程序 svm-train.exe,根据得到的最优参数训练得到文本的分类器。本文采用了封闭式和开放式两种测试方案。封闭式测试随机选取 5 个测试文件,由于封闭式测

试中的测试集包含在训练集里面,因此得到的测试准确率在 98% 以上。但是这种测试对于准确挑选 EC 没有意义,所以尽管它的准确率很高,其也不能作为挑选的凭证。由文献[12]的结果可知,对残基对约束的预测一直以来就是一个比较难的项目,准确率也相对较低,所以本文选取部分准确率。鉴于以上原因,本文采用开放式测试,将 27 条蛋白质产生的 27 个文件中的任意一个作为测试集,其余 26 个作为训练集,循环进行 27 次。每个文件分别作为测试集 1 次,作为训练集 26 次。

表 1 统计了每个目标在几何距离为 8Å-10Å 的个数及比率。本文更侧重于部分准确率,因为在决定蛋白质结构中,只有空间上相对较远的残基对蛋白质结构预测才更具指导意义。表 1 第二列选取的是浮动个数,因为在选取标签时采用的是四舍五入取整,而且蛋白质的结构距离约束不是固定的,在某个范围内就可以接收,所以上、下浮动两个距离值内的都正确;表中第四列体现的是部分准确率,未显示出的标签在 4Å 到 7Å 时,准确率整体比较高。

这种高比例在蛋白质结构中是显而易见的,因为在序列上相距近的残基对空间上必定距离较近,对结构预测作用不是很大(图 3 中能够得到验证)。有相关联系的残基对在空间上相隔 12 个残基以上对结构预测意义是比较大的,这些残基对大多集中在距离为 8Å 到 10Å 的这 3 部分。表 1 第五列即显示了在预测结果中预测出的 8Å 到 10Å 内相隔 11 个残基以上的残基对个数与所有预测出的相隔 11 个以上残基的残基对个数对比,可以看出 8Å 到 10Å 内残基对在总体残基对中所占比例除 3NQK 和 3TD7 外都在 0.6 以上,而且大部分在 0.7 以上,这也是本文选择 8Å 到 10Å 这 3 部分进行研究的重要原因之一。

表 1 实验结果

PDBid	浮动后个数	总数	比率(%)	Rate(8Å-10Å/total)
2KJX	28-54-21	51-103-46	54.9-52.4-45.7	0.94(75/83)
2L06	54-75-46	123-178-134	43.9-42.1-34.3	1.000(8/8)
2L3W	72-110-49	120-192-144	60.0-57.3-34.0	0.697(124/178)
2XQF	201-221-131	719-1114-750	17.9-19.8-17.5	0.680(398/585)
3D3Q	138-213-106	172-359-193	80.2-59.3-54.9	0.866(220/254)
3D4R	73-140-69	128-236-175	57.0-59.3-39.4	0.904(161/178)
3DO9	50-60-30	100-174-108	50.0-34.5-27.8	0.901(91/101)
3DOA	211-271-114	292-470-328	72.3-57.7-34.8	0.725(322/444)
3KLU	23-66-31	84-197-111	27.4-33.5-27.9	1.000(19/19)
3MWT	376-511-215	526-903-552	71.5-56.6-38.9	0.727(769/1058)
3N8U	71-58-44	137-258-162	51.8-22.5-27.2	0.796(90/113)
3N91	33-4-9	63-80-52	52.4-5.0-17.3	0.066(7/106)
3NAT	72-109-59	129-214-166	55.8-50.9-35.5	0.902(120/133)
3NPD	93-153-93	96-164-118	96.9-93.3-78.8	0.853(221/259)
3NQK	34-47-17	44-100-48	77.3-47.0-35.4	0.597(92/154)
3TD7	13-32-18	66-125-98	19.7-25.6-18.4	0.300(3/10)
3UX4	137-232-71	165-306-208	83.0-75.8-34.1	0.743(286/385)
3NYY	26-42-24	79-144-109	32.9-29.2-22.0	0.833(80/96)
4FGM	100-191-114	108-197-120	92.6-97.0-95.0	0.837(221/264)
4FJ6	61-56-26	118-182-110	51.7-30.8-23.6	0.602(115/191)
4FS7	120-192-94	132-217-102	90.9-88.5-92.2	0.817(326/399)
4G2A	52-73-51	55-81-72	94.5-90.1-70.8	0.773(85/110)

比较特殊的案例 3NQK 和 3TD7 属于序列比较长、结构比较复杂的案例,其复杂性造成结果相对较低是比较正常的。2L06 和 3KLU 中所有测出的相隔 11 个残基以上的残基对都在 8Å 到 10Å 内,而且个数比较少,这可能是因为在这两个案例中所有在序列上相距较远的残基对在不同的空间距离中分布较为均匀,不能很好地区分所致。对于 3N91,结果完全与预想相反,在 8Å 到 10Å 内个数远远少于在剩余部分的个数,

可能由两个原因导致,一是结构复杂,而且该目标中含有较多的特殊原子;另外一种原因可能是所选的属性特征不适合该目标,无法进行正确的分类。图 3 为选取的 4 个案例结果的散点图。从图中可以看出挖掘准确的大部分集中在对角线上残基相距较近的部分,远离对角线的点即本文挖掘的重点大部分点在我们的挖掘范围内,这进一步验证了本文方法结果

(下转第 299 页)

Chen Xiu-feng, Xu Hong-guo, Ni An-ning. Dynamic Load Balancing Mechanism and Algorithms in Parallel Microscopic Traffic Simulation[J]. Computer Simulation, 2013, 30(8): 164-168

- [5] Leontiadis I, Marfia G, Mack D, et al. On the effectiveness of an opportunistic traffic management system for vehicular networks [J]. IEEE Transactions on Intelligent Transportation Systems, 2011, 12(4): 1537-1548
- [6] Shen Wei, Wynter L. A New One-level Convex Optimization Approach for Estimating Origin-destination Demand [J]. Transportation Research Part B: Methodological, 2012, 46(10): 1535-1555
- [7] Sun Hui-jun, Zhang Hui, Wu Jian-jun. Correlated scale-free network with community: modeling and transportation dynamics [J]. Nonlinear Dynamics, 2012, 69(4): 2097-2104
- [8] 王光浩, 吴越. 一种车载自组织网络路况信息的数据信任模型 [J]. 计算机科学, 2014, 41(6): 89-93
Wang Guang-hao, Wu Yue. Data Trust Model for Road Information in Vehicular Ad hoc Network[J]. Computer Science, 2014,

41(6): 89-93

- [9] 张子龙, 薛静, 乔鸿海, 等. 基于改进 SURF 算法的交通视频车辆检索方法研究[J]. 西北工业大学学报, 2014, 32(2): 297-301
Zhang Zi-long, Xue Jing, Qiao Hong-hai, et al. The Vehicle Retrieval Methods of Traffic Video Based on Improved SURF Algorithm[J]. Journal of Northwestern Polytechnical University, 2014, 32(2): 297-301
- [10] 高颢悦. 一种基于随机网格简化的 Web 可靠性分析方法研究 [J]. 科技通报, 2013, 4(29): 67-69
Gao Jin-yue. A Web reliability analysis method based on random mesh simplification research [J]. Bulletin of Science and Technology, 2013, 4(29): 67-69
- [11] 韩国卫, 彭伟, 唐晋韬. 基于路标的最短路径长度快速估计算法 [J]. 重庆理工大学学报(自然科学版), 2013, 27(7): 96-102, 118
Han Wei-guo, Peng Wei, Tang Jin-tao. A Landmark-Based Fast Shortest-Path Length Estimation Algorithm [J]. Journal of Chongqing University of Technology(Natural Science), 2013, 27(7): 96-102, 118

(上接第 264 页)

的准确率,同时也反映出能够挖掘出的残基对在序列上的分布。

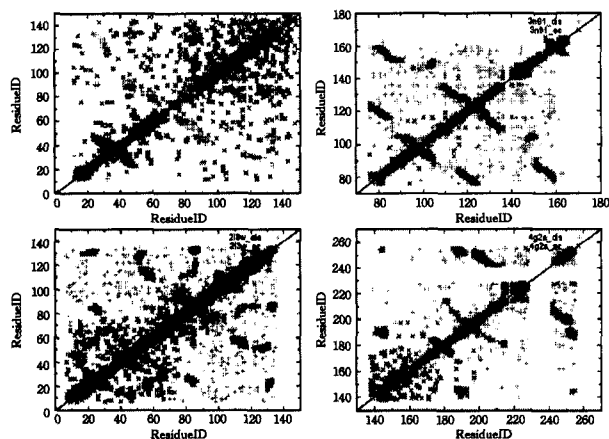


图3 3KLU、3N91、2L3W 和 4G2A 4 个案例的结果图,其中灰色点为准确的残基对,黑色为挖掘出的进化配对,黑色能够覆盖灰色的部分表示挖掘准确。

结束语 本文以从蛋白质残基配对关系中挖掘出的距离约束为目标,将一种定性的生化信息转化为定量的距离约束,使之成为计算机更容易理解的约束项。本文训练属性更多的是来自蛋白质氨基酸序列,避免了复杂属性的不准确性,简化了计算,提高了运算效率。实验表明,EC 的准确性影响到蛋白质三维结构预测的准确性,将 EC 充分运用到预测中将对后面的实验产生积极的影响。因此,尝试找到更有代表性的特征属性来提高预测的准确性是工作的难点,另外如何将挖掘出的距离约束运用到结构预测中是下一步工作的重点。

参考文献

- [1] Marks D S, Colwell L J, Sheridan R, et al. Protein 3D structure computed from evolutionary sequence variation[J]. PloS one, 2011, 6(12): e28766
- [2] Hopf T A, Colwell L J, Sheridan R, et al. Three-dimensional structures of membrane proteins from genomic sequencing[J].

Cell, 2012, 149(7): 1607-1621

- [3] Wang Zhi-yong, Xu Jin-bo. Predicting protein contact map using evolutionary and physical constraints by integer programming [J]. Bioinformatics, 2013, 29(13): i266-i273
- [4] Ma Jian-zhu, Wang Sheng, Xu Jin-bo. Joint Multi-family Evolutionary Coupling Analysis for Protein Contact Prediction[J]. arXiv preprint arXiv:1312.2988, 2013
- [5] Wu Si-tao, Zhang Yang. MUSTER: improving protein sequence profile-profile alignments by using multiple sources of structure information[J]. Proteins: Structure, Function, and Bioinformatics, 2008, 72(2): 547-556
- [6] Lazaridis T, Karplus M. Effective energy functions for protein structure prediction[J]. Current opinion in structural biology, 2000, 10(2): 139-145
- [7] Jones D T, Buchan D W A, Cozzetto D, et al. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments[J]. Bioinformatics, 2012, 28(2): 184-190
- [8] Kryzhtafovich A, Monastyrskyy B, Fidelis K. CASP Prediction Center infrastructure and evaluation measures in CASP10 and CASP ROLL[J]. Proteins: Structure, Function, and Bioinformatics, 2014, 82(S2): 7-13
- [9] Ezkurdia I, Grana O, Izarzugaza J M G, et al. Assessment of domain boundary predictions and the prediction of intramolecular contacts in CASP8[J]. Proteins: Structure, Function, and Bioinformatics, 2009, 77(S9): 196-209
- [10] 郭海娟, 吕强, 吴宏杰, 等. 一个识别蛋白质折叠模式的 SVM 分类器[J]. 生物信息学, 2010(4): 287-290
Guo Hai-juan, Lv Qiang, Wu Hong-jie, et al. A svm classifier for protein fold recognition [J]. China Journal of Bioinformatics, 2010(4): 287-290
- [11] Altschul S F, Koonin E V. Iterated profile searches with PSI-BLAST—a tool for discovery in protein databases[J]. Trends in biochemical sciences, 1998, 23(11): 444-447
- [12] Cheng Jian-lin, Baldi P. Improved residue contact prediction using support vector machines and a large feature set[J]. BMC bioinformatics, 2007, 8(113): 1-9