

基于支持向量机的蛋白质相互作用位点预测

王菲露^{1,2}, 宋杰¹, 王池社¹, 杜秀全¹

(1. 安徽大学 计算智能与信号处理教育部重点实验室, 安徽 合肥 230039;

2. 安徽建筑工业学院 电子与信息工程学院, 安徽 合肥 230601)

摘要:蛋白质的功能常体现在生物大分子的相互作用中, 识别蛋白质相互作用位点对于研究蛋白质功能发挥着重要作用。蛋白质间主要通过表面残基发生相互作用, 蛋白质相互作用形成复合体时, 只有部分表面残基参与了该过程。基于序列谱信息, 提取序列上相邻残基的序列谱作为输入特征向量, 对大小为3和7的残基信息窗(win3, win7), 分别采用支持向量机(SVM)分类器对蛋白质相互作用位点进行预测、比较和分析。最终实验结果为: win3的平均正确率为69.31%, win7的平均正确率为69.68%。

关键词:蛋白质相互作用; 序列谱; 残基信息窗; 支持向量机

中图分类号: TP183

文献标识码: A

文章编号: 1673-629X(2008)10-0127-03

Prediction of Protein-Protein Interaction Sites with SVM

WANG Fei-lu^{1,2}, SONG Jie¹, WANG Chi-she¹, DU Xiu-quan¹

(1. Ministry of Education, Key Lab. of Computing Intelligence and Signal Processing, Anhui

University, Hefei 230039, China;

2. School of Electronics & Information Engineering, Anhui University of Architecture, Hefei 230601, China)

Abstract: Protein usually represents its function through interactions among biological molecules. Identifying protein-protein interaction sites plays an important role in protein's function. The interactions among proteins are produced by surface residues mainly, and only part of surface residues participate in this produce. Adjacent residue sequences profile are as input vectors of two information windows of residue (win3, win7) and interaction sites are classified by support vector machine (SVM). The result shows that the average accuracy of win3 is 69.31%, and the average accuracy of win7 is 69.68%.

Key words: protein-protein interaction; sequence profile; information window of residue; SVM

0 引言

蛋白质相互作用在蛋白质结构和功能预测过程中发挥着越来越重要的作用, 蛋白质相互作用位点的预测是当前生物信息学研究的热点之一。识别蛋白质相互作用位点, 以及检测相互作用氨基酸残基之间的特异性, 是一个具有重要应用前景的课题。然而, 目前蛋白质相互作用预测都通过生化实验确定是不现实的。随着生物信息学和计算生物学的发展, 通过研究已知蛋白质相互作用位点的不同特征, 出现了利用序

列谱与结构等信息作为输入数据来预测蛋白质相互作用位点的计算方法。

在蛋白质相互作用位点预测研究中采用的机器学习算法也越来越多, 如人工神经网络、贝叶斯网、多分类器组合等。文中用支持向量机(SVM)分类器对界面残基(相互作用位点)和非界面残基进行分类, 对不同大小的残基信息窗分类结果进行比较分析。

1 支持向量机(SVM)

支持向量机是 Vapnik 等根据统计学习理论提出的一种机器学习方法, 可用于解决模式分类等问题。它的主要思想是^[1]建立一个最优决策超平面, 使得该平面两侧距平面最近的两类样本之间的距离最大化, 从而为分类问题提供良好的泛化能力。支持向量机建立的分类超平面能够在保证分类精度的同时, 使超平面两侧的空白区域最大化, 从而实现对线性可分问题的最优分类。支持向量机能够较好地避免传统机器学

收稿日期: 2008-01-15

基金项目: 安徽省自然科学基金(KJ2007B239); 安徽建筑工业学院青年科研基金(200510304); 安徽省高校青年教师科研资助计划(2007jql140)

作者简介: 王菲露(1981-), 女, 安徽合肥人, 硕士研究生, 研究方向为智能计算、生物信息学等; 宋杰, 博士, 副教授, 硕士生导师, 研究方向为智能计算、生物信息学、嵌入式系统等。

习方法中的维数恶化问题,其方法是:将输入向量映射到一个高维特征向量空间,如果选用的映射函数适当且特征空间的维数足够高,则多数非线性可分模式在特征空间中可以转化为线性可分模式。因此可以在该特征空间构造最优超平面进行模式分类,这个构造与内积核有关。

常用的内积核函数有以下几种:

* 多项式核函数: $K(X, X^p) = [(X \cdot X^p) + 1]^q$ (1)

* Gauss 核函数: $K(X, X^p) = \exp(-\frac{|X - X^p|^2}{2\sigma^2})$ (2)

* Sigmoid 核函数: $K(X, X^p) = \tanh(k(X \cdot X^p) + c)$ (3)

* 径向基函数(RBF): $k(X, X^p) = \exp(-\gamma |X - X^p|^2)$ (4)

支持向量机对非线性可分数据,在进行非线性变换后的高维特征空间实现线性分类,其最优分类判别函数为:

$$f(x) = \text{sgn}[\sum_{i=1}^{N_s} a_i d^i K(X^i, X) + b_0] \quad (5)$$

其中 N_s 为支持向量的数量。

支持向量机结构模型如图 1 所示。

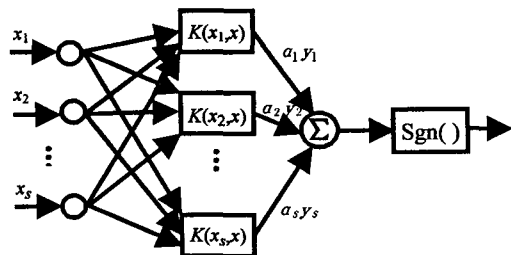


图 1 支持向量机示意图

2 用支持向量机预测蛋白质相互作用位点

2.1 相关定义

相互作用位点通常有两种定义方式:一是用形成复合体时残基的溶剂可及表面积(ASA)的减少来定义;二是用残基之间的原子距离来定义^[2]。文中采用第一种方法定义。

表面残基的定义:一般用形成复合体时可及表面积(ASA)的减小来定义表面残基,不同的人选择的阈值不一样,Asake Koik 和 Toshihisa Takagi^[3]把溶剂的 ASA 对残基的最大面积的比值超过 10% 的残基定义为表面残基,Chanhui Yan, Drena Dobbs 等人^[4]则把在未绑定单体中的 ASA(MASA)超过在复合体中的 ASA(CASA)的 25% 的残基定义为表面残基。文中采用 Yohei Minakuchi 等人^[5]把表面积与最大面积的比值大

于 16% 的残基定义为表面残基。蛋白质通过表面残基发生相互作用,蛋白质链相互作用形成复合体时,只有部分表面残基参与了该过程。蛋白质链相互作用的界面,通常称之为相互作用位点。

界面残基的定义:文中采用 Changhui Yan^[4]等的定义方法,如果一个表面残基的 MASA(残基在单体中的可及表面积)与 CASA(残基在复合体中的可及表面积)的差值大于 1A,则定义为界面残基,否则定义为非界面残基。

2.2 序列谱

蛋白质序列谱是一个二维数组,第一维对应于序列中的位置序号,每一行是一个 20 维的向量(对应 20 种氨基酸),向量中的每个元素(对应于数组的第二维)分别代表 20 种氨基酸在这个位置出现的频率。文中从蛋白质 hssp 文件中提取序列谱(如图 2 所示)。

SeqNo	PDNo	V	L	I	M	F	W	Y	G	A	P	S	T	C	H	R	K	Q	E	N	D
1	1A	14	2	2	54	29	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	2A	2	64	1	1	31	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	3A	0	0	0	0	0	0	0	0	0	0	17	80	0	0	0	0	0	1	0	0
4	4A	0	0	0	0	0	1	0	12	45	12	2	1	0	0	0	1	0	6	1	21
5	5A	1	0	0	0	3	0	0	1	12	1	4	2	0	1	0	1	1	64	0	8
6	6A	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	89	1	10
7	7A	0	1	1	0	0	0	0	0	1	0	0	0	0	0	18	79	0	0	1	0
8	8A	1	0	0	0	0	0	0	4	44	0	21	7	0	1	0	2	11	2	6	1
9	9A	2	17	6	0	1	0	2	0	55	0	1	12	0	1	1	0	1	1	1	1
10	10A	42	0	56	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0

图 2 序列谱示意图(取自 1ABY_A 的序列谱)

2.3 实现过程

2.3.1 数据集的选取

文中从 FariSelli 等^[6]使用的数据集中进行筛选,使用 BLAST 工具排除一致性大于等于 35% 且序列长度在 85% 左右的蛋白质链,最终得到 46 个蛋白质复合物的 74 条蛋白质链作为文中研究的数据集(见表 1)。文中选择的数据集中:残基总数为 18392;表面残基总数为 11235, 占有残基的 61.0863%;界面残基为 3648, 占表面残基数 32.47%, 占总残基数的 19.8347%。

表 1 选取的数据集

1ABY-A	1ABY-B	1ACY-L	1ADO-B	1AGB-A	1AGB-B	1AGR-A	1AGR-E
1AIS-A	1AIS-B	1ALL-A	1AOK-A	1AQD-A	1AQD-B	1ATN-A	1ATN-D
1AUI-A	1AUI-B	1AXI-A	1AXI-B	1BFV-H	1BFV-L	1BPL-A	1BPL-B
1BRL-A	1BRL-B	1CAU-A	1CAU-B	1EBD-A	1EBD-C	1EFU-A	1EFU-B
1EFV-A	1FIN-A	1FTN-B	1FRV-B	1GLA-F	1GLA-G	1GUA-A	1GUA-B
1IBC-A	1IBC-B	1IGT-B	1IHF-A	1IHF-B	1LGB-A	1MEL-L	1MHL-C
1MIO-A	1MIO-B	1NPO-C	1PHN-A	1PHN-B	1RBL-A	1RBL-M	1RLB-E
1SCT-B	1SCU-A	1SCU-B	1SEB-D	1TCR-A	1TCR-B	1TMC-B	1TTP-A
1TTP-B	1VOL-A	1YRN-A	1YRN-B	1YUH-L	2BTF-P	2PCC-A	2PCC-B
2REQ-A	2REQ-B						

2.3.2 用 DSSP 程序生成 dssp 文件,确定输入向量

使用 DSSP 程序生成 dssp 文件,计算每个残基的 CASA 和 MASA。并按 2.1 中的定义计算出界面残基和非界面残基,分别用 +1、-1 标识。

从 HSSP 中下载蛋白质 hssp 文件,并提取出需要的蛋白质链的序列谱。文中采用在序列上相邻残基的序列谱作为输入特征向量,残基信息窗大小分别为 3 和 7。

对于窗口大小为 3 或 7 的输入向量是由目标残基的序列谱和与之在序列上相邻的 2 个或 6 个残基(两侧各 1 个或 3 个残基)的序列谱组合在一起,构成的大小为 20×3 或 20×7 维的输入向量,如式(6)、(7)。

$$X_n = (P_{n-11}, \dots, P_{n-120}, P_{n1}, \dots, P_{n20}, P_{n+11}, \dots, P_{n+120}) \quad (6)$$

$$X_n = (P_{n-31}, \dots, P_{n-320}, \dots, P_{n1}, \dots, P_{n20}, \dots, P_{n+31}, \dots, P_{n+320}) \quad (7)$$

如果该目标残基是界面残基,则在相应的输入向量中标记 +1,若是非界面残基则标记 -1。

2.3.3 预测器的构建

文中使用 SVM 分类器,分别用 3 窗口、7 窗口的目标残基输入向量作为训练数据。采用“74 留一法”进行实验。每次拿出一条单链,对剩下的 73 条蛋白质单链训练 SVM 分类器。最后用训练过的 SVM 分类器,对拿出去的那条单链进行预测,如果输出结果是 +1,则认为目标残基是界面残基,如果输出结果是 -1,则认为该目标残基是非界面残基。实验时使用径向基核函数 RBF(核函数中 $\gamma = 0.0$, 惩罚因子 $c = 1.0$),用 SVM 分类器进行分类预测。

3 结果分析

实验使用 SVM 分类器预测蛋白质相互作用位点,分别选取目标残基和与之在序列上相邻的 2 个、6 个残基,形成残基信息窗大小为 3 和 7 的输入向量,经训练后,再用 SVM 分类器进行预测,分别得出 74 条蛋白质链的正确率。最终计算出 3 窗口(Win3)和 7 窗口(Win7)的平均正确率分别为:69.31%, 69.68%。74 条蛋白质链的预测结果如图 3、图 4 所示。

从实验结果得出:7 窗口的正确率较 3 窗口的正确率略有提高,但效果不是特别明显。甚至有个别目标残基的 7 窗口正确率略低于 3 窗口的正确率。这说明并不一定是窗口越大,实验结果的正确率就越高。目标残基与其相邻残基存在一定的协调问题。在运行过程中,7 窗口的运算复杂度要大于 3 窗口。窗口的增大虽然带来了正确率的提高,但也使运算复杂度增

大了。

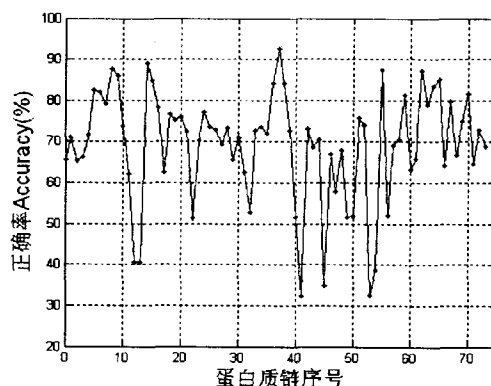


图 3 Win3 实验结果(平均正确率为 69.13%)

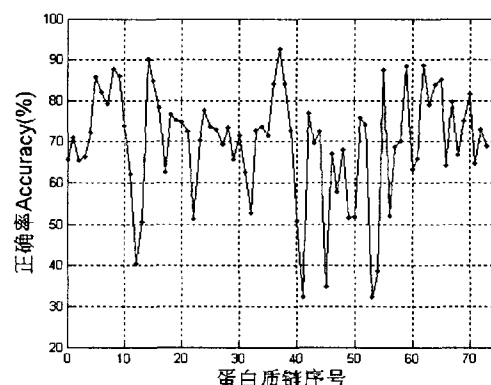


图 4 Win7 实验结果(平均正确率为 69.68%)

4 结束语

蛋白质是通过相互作用发挥其作用的,研究蛋白质相互作用位点有着重要的意义。文中使用 SVM 分类器,对大小不同的残基信息窗分别进行蛋白质相互作用位点预测,得到较好的结果,说明了该方法的有效性。为进一步提高预测的准确度,可选取目标残基在序列上相邻的残基序列谱与可及表面积的结合、目标残基在空间上相邻的残基序列谱、目标残基在空间上相邻的残基序列谱与可及表面积的结合等作为输入特征向量进行研究,来获得更好的预测结果。

参考文献:

- [1] 韩力群. 神经网络教程[M]. 北京:北京邮电大学出版社,2006:185-192.
- [2] 安书君. 改进的蛋白质相互作用位点预测方法研究[D]. 哈尔滨:哈尔滨工业大学,2006.
- [3] Koike A, Takagi T. Prediction of protein-protein interaction sites using support vector machines[J]. Protein Engineering, Design & Selection, 2004, 17(2): 165-173.
- [4] Yan Changhui, Dobbs D, Hoavar V. A two-stage classifier for identification of protein-protein interface residues[J]. Bioinformatics, 2004, 20(s): 371-378.

(下转第 132 页)

步^[6]。具体说来,包括以下几个方面。

1)不一致数据转换:这个过程是一个整合的过程,将不同业务系统的相同类型的数据统一,比如同一个供应商在结算系统的编码是 XX0001,而在 CRM 中编码是 YY0001,这样在抽取过来之后统一转换成一个编码。

2)参照转换:在转换中通常要用数据源的一个或多个字段作为 Key,去一个关联数组中搜索特定值,而且应该只能得到唯一值。这个关联数组使用 Hash 算法实现是比较合适也是最常见的,在整个 ETL 开始之前,它就装入内存,对性能提高的帮助非常大。

3)字符串处理:从数据源某个字符串字段中经常可以获取特定信息,例如身份证号。而且,经常会有数值型值以字符串形式体现。对字符串的操作通常有类型转换、字符串截取等。但是由于字符类型字段的随意性也造成了脏数据的隐患,所以在处理这种规则的时候,一定要加上异常处理。

4)日期转换:在数据仓库中日期值一般都会有特定的、不同于日期类型值的表示方法,例如使用 8 位整型 20071001 表示日期。而在数据源中,这种字段基本都是日期类型的,所以对于这样的规则,需要一些共通函数来处理将日期转换为 8 位日期值、6 位月份值等。

5)日期运算:基于日期,通常会计算日差、月差、时长等。一般数据库提供的日期运算函数都是基于日期类型的,而在数据仓库中采用特定类型来表示日期的话,必须有一套自己的日期运算函数集。

6)既定取值:这种规则和以上各种类型规则的差别就在于它不依赖于数据源字段,对目标字段取一个固定的或是依赖系统的值。

7)聚集运算:业务系统一般存储非常明细的数据,而数据仓库中数据是用来分析的,不需要非常明细的数据。一般情况下,会将业务系统数据按照数据仓库粒度进行聚合。

8)商务规则计算:不同的企业有不同的业务规则、不同的数据指标,这些指标有的时候不是简单的加加减减就能完成,这个时候需要在 ETL 中将数据指标计算好了之后存储在数据仓库中,以供分析使用。

4 数据加载

数据加载是将转换后的数据加载到数据仓库中。

数据加载策略包括加载周期和数据追加策略,数据加载周期要综合考虑经营分析需求和系统加载的代价,对不同业务系统的数据采用不同的加载周期,但必须保持同一时间业务数据的完整性和一致性。

数据追加策略,根据数据抽取阶段的不同选择,又可以分为两种不同的方式:

一、在数据抽取阶段选择增量抽取,那么在数据转换完成后可以直接将数据加载到数据仓库中;

二、若在抽取阶段选择全量抽取,则可以根据数据的时间标记或源数据的操作日志或采用全表对比的方式选择相应的数据加载,几种方式各有优缺点,应针对不同的情况进行考虑。

5 结束语

ETL 是 BI 项目的关键部分,也是一个长期的过程,同时这部分的工作直接关系数据仓库中数据的质量,从而影响到决策分析的结果的质量。在 ETL 过程中的每一步都会发现大量的问题,有些可以直接解决,有些则需要回溯到前一个甚至几个过程。通常情况下,每次对 ETL 过程的修改都需要重新运行整个 ETL 过程并对结果进行验证。这样一来,开发整个 ETL 过程的所需的时间必然很长。因此,只有认真、仔细地设计 ETL 过程中的每一步,尽量使得 ETL 过程每一步的运行效率更高、结果更准确同时更容易修改,才能有效保证整个 BI 项目的最终成功。

参考文献:

- [1] 程跟上. 基于公共仓库模型的 ETL 系统的研究和应用[D]. 南京:南京航空航天大学,2005.
- [2] Rahmand E, Hong Haido. Data Cleaning, Problems and Current Approaches[J]. IEEE Bulletin of the Tenniel Committee Data Engineering, 2000, 23(4): 3-13.
- [3] 章水鑫,徐宏炳,于立. 增量式 ETL 工具的研究与实现[J]. 现代计算机, 2005(3): 6-10.
- [4] Hernandez M. A Generation of Band Joins and the Merge/Purge Problem[R]. USA: Department of Computer Science, Columbia University, 1995.
- [5] Inmon W H. Building the Data Warehouse[M]. 北京:机械工业出版社, 2007.
- [6] 张宁,贾自艳. 数据仓库中 ETL 技术的研究[J]. 计算机工程与应用, 2002, 38(24): 213-216.

(上接第 129 页)

- [5] Minakuchi Y, Satou K, Konagaya A. Prediction of Protein-Protein Interaction Sites Using Support Vector Machines[J]. Genome Informatics, 2002(13): 322-323.

- [6] Fariselli P, Pazos F, Valencia A, et al. Predication of protein-protein interaction sites in heterocomplexes with neural networks[J]. Eur. J. Biochem, 2002, 269: 1356-1361.