文章编号: 1006–6330(2011)02–0235–10

利用提升小波的蛋白质相互作用特征提取

冯铁男[1], 答 亮[2], 金鼎立[1], 王翼飞[1]

(1. 上海大学理学院, 上海 200444;
2. 中国科学院上海生命科学研究院上海生物化学与细胞生物学研究所, 上海 200031)

摘要 利用提升小波从蛋白质序列中提取出它们相互作用的频谱特征, 经支持向量机训练学习后, 用于预测蛋白质间的相互作用. 模拟计算结果表明, 在阳性数据和阴性数据平衡的前提下, 利用提升小波获取的低维蛋白质相互作用特征向量可以得到较高预测精度. 进一步阐述了不同物种的蛋白质相互作用网络有着不同特征, 为了得到更准确的预测结果, 需要利用不同的方法提取蛋白质相互作用的特征.

关键词 提升小波; 相互作用特征; 数据平衡; 蛋白质相互作用
2010 数学分类号 92-08
中图分类号 O242.1 **文献标志码** A

Feature-mined of protein-protein interactions using lifting wavelet

FENG Tie-nan[1], DA Liang[2], JIN Ding-li[1], WANG Yi-fei[1]

(1. College of Sciences, Shanghai University, Shanghai 200444, China;
2. Chinese Academy of Sciences, Shanghai Institutes for Biological Sciences, Institute of Biochemistry and Cell Biology, Shanghai 200031, China)

Abstract The protein-protein interacting features only from the protein's sequence are calculated by using the lifting wavelet. These features then are learned by the support vector machine to train a model by which the protein-protein interactions are predicted. Numerical results report that, on the principle of balance between positive dataset and negative dataset, the low-dimensional vector of features has gained a better performance. Results also report that features are different among the local protein-protein interaction network of different species. For making a more accuracy prediction, it is essential to use several methods.

Key words lifting wavelet; interacting feature; dataset balance; protein-protein interaction
2010 Mathematics Subject Classification 92-08
Chinese Library Classification O242.1

收稿日期: 2011–06–24; **修订日期**: 2011–09–05
基金项目: 科技部重大科技专项资助项目 (2009ZX09103-686); 国家自然科学基金资助项目 (30971480); 上海大学研究生创新基金资助项目 (SHUCX101072)
通信作者: 王翼飞, 研究方向为计算分子生物学. E-mail: yifei_wang@staff.shu.edu.cn

蛋白质相互作用网络构成了生命体内活动的基础，深入了解它们的结构可以进一步理解生命的内在机理. 实验或计算方法都可以得到或是推断出相互作用的蛋白质[1−6], 尽管实验可以检测出可靠性强的相互作用蛋白质, 但它有自身局限[7], 而且生命体这个复杂的细胞系统涉及的蛋白质相互作用网络巨大而复杂，目前仅靠实验的手段是无法完成这样宏大的工程，因此低成本且高效的计算预测方法无疑成为了一个重要的技术. 蛋白质一级结构 (序列) 包含着蛋白质高级结构信息[8], 在一级结构的基础上研究蛋白质相互作用的方法多种多样[9−12]. 共鸣识别模型 (RRM) 是一种从数值角度分析蛋白质序列的方法，它可以从一组具有相同生物学功能的蛋白质中提取出对应于此种生物学功能的特征频率[13], 而相互作用蛋白质则往往具有大小相似、相位角方向相反的特征频率，但是获取特征频率需要一定数量的同功能蛋白质，因此很难仅从一对蛋白质中直接确定它们的特征频率. 小波变换可以得到不同分辨率下蛋白质的频谱信息，通过分析相互作用蛋白质对在不同分辨率下的频率强度便可获取它们的特征频率信息，从而直接地预测蛋白质间的相互作用[14]. 本文将提升小波[15−16] 应用于蛋白质相互作用的特征提取. 为了得到同一频率下不同蛋白质序列的谱信息，人们一般在长度较短蛋白质序列的尾端添加零元素使得蛋白质序列长度一致，但是这会使得添加零元素后序列的谱信息和添加零元素前序列的谱信息产生差异. 提升小波可以较好地处理同一蛋白质序列因补零法所造成的差异，使相互作用蛋白质的特征更加准确，之后再利用支持向量机 (SVM) 学习此特征，训练出预测模型，直接预测蛋白质间的相互作用.

1 材料与方法

1.1 数据集

RRM 是基于相互作用的生物分子 (如蛋白质) 频率互补，而分子频率则是由分子自身主链上的极性所决定，因此本文所用到的数据集均为物理接触的蛋白质相互作用数据. 数据来源为：http: //thebiogrid.org/downloads/archives/Published%20Datasets /HC-BIOGRID-2.0.31.tab, 文件中蛋白质都属于酵母，命名方式为基因命名法. 去除了冗余及 Uniprot 数据库中信息缺失的蛋白质后，最终一共得到了 8 224 对物理上相互作用的蛋白质，涉及到 2 330 个蛋白质，以此作为本研究的一个分析数据集. 为了验证所用方法的有效性，本研究还提取了 IntAct 数据库中物理上相互作用的小鼠蛋白质对作为另一个数据集合，在删除冗余及信息缺失的数据后，得到 4 811 对相互作用的蛋白质，一共涉及 2 324 个蛋白质. 对于酵母，选取其数据中 5 000 对作为训练数据，其余为测试数据 (3 224 对); 对于小鼠，则选取它前 3 000 对作为训练数据，其余为测试数据 (1 811 对).

为了说明计算结果的可信性, 需选取与阳性数据性质相异的数据作为比对数据 (阴性数据). 由于在实验中得到的蛋白质相互作用数据集中，有一定数量高连接度 (连接度表示与此蛋白质相互作用蛋白质的数目) 蛋白质 (也称为枢纽蛋白质), 这些枢纽蛋白质会多次出现在相互作用数据集合中, 导致它们的特征会被多次学习，最终影响预测结果. 为了平衡两类数据中各个蛋白质的连接度， Yu 等[17] 提出了数据平衡方法，其思想是，取原

蛋白质相互作用网络图的补图，从补图中选取蛋白质对来生成阴性数据集合，在阴性数据集合中的每个蛋白质对都不在相互作用数据库中，并且它们的连接度和阳性数据集合中此蛋白质的连接度相同．在本文中，选取两类数据作为比对数据．一类是随机数据：在蛋白质集合中，取随机配对的蛋白质；一类是平衡数据：基于数据平衡思想生成的数据集合．

1.2　小波变换和提升小波变换

小波变换是一种处理信号的数学工具．本质上，可以认为它们是处理信号数据的模块，通过将序列同小波模块的作用，就可以在不同频率 —— 高频 (对应经过高频滤波器分解后的信号) 和低频 (对应经过低频滤波器分解后的信号) 下显示信号信息 (图 1)．图 1 中 c^j 表示原始信号，c^{j-1} 和 d^{j-1} 分别表示经过一次分解后，信号的低频信息和高频信息，$\downarrow 2$ 和 $\uparrow 2$ 分别表示 “下抽样” 和 “上抽样”, H 表示小波的低频滤波器，G 表示小波的高频滤波器．

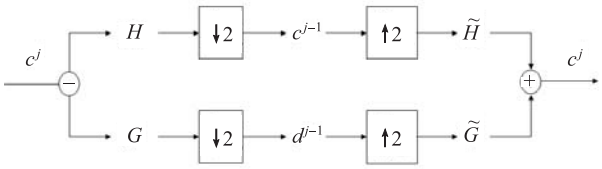

图 1　利用小波的信号分解与重构

可以将图 1 中高低频滤波器的表达方式记为矩阵多项式的形式：

$$P(z)=\begin{pmatrix} H(z) \\ G(z) \end{pmatrix}=\begin{pmatrix} h_{\mathrm{e}}(z) & h_{\mathrm{o}}(z) \\ g_{\mathrm{e}}(z) & g_{\mathrm{o}}(z) \end{pmatrix}, \tag{1.1}$$

$$\widetilde{P}(z)=\begin{pmatrix} \widetilde{H}(z) & \widetilde{G}(z) \end{pmatrix}=\begin{pmatrix} \widetilde{h}_{\mathrm{e}}(z) & \widetilde{g}_{\mathrm{e}}(z) \\ \widetilde{h}_{\mathrm{o}}(z) & \widetilde{g}_{\mathrm{o}}(z) \end{pmatrix}. \tag{1.2}$$

一般情况下，h 表示小波的低频滤波系数序列，g 表示小波的高频滤波系数序列，它们都由具体的小波给出，$h(z)$ 和 $g(z)$ 分别表示系数对应生成的劳伦斯多项式 [18], 下标 e 和 o 分别表示序列中的偶数项和奇数项，$P(z)$ 和 $\widetilde{P}(z)$ 中的多项式有如下关系：

$$\begin{cases} h_{\mathrm{e}}(z)=\widetilde{g}_{\mathrm{o}}(z^{-1}), \\ h_{\mathrm{o}}(z)=-\widetilde{g}_{\mathrm{e}}(z^{-1}), \\ g_{\mathrm{e}}(z)=-\widetilde{h}_{\mathrm{o}}(z^{-1}), \\ g_{\mathrm{o}}(z)=\widetilde{h}_{\mathrm{e}}(z^{-1}). \end{cases} \tag{1.3}$$

Daubechies 和 Sweldens[18] 证明了任何离散小波都可以用提升方案来实现，在提升方案中：

$$h_{\mathrm{L}}=h(z)+g(z)s(z), \tag{1.4}$$

$$\widetilde{g}_{\mathrm{L}} = \widetilde{g}(z) + \widetilde{h}(z)\,\widetilde{s}(z), \tag{1.5}$$

所以，提升小波的高低频滤波器可以由如下的方法[18] 得到：

$$P_{\mathrm{L}}(z) = \begin{pmatrix} 1 & s(z) \\ 0 & 1 \end{pmatrix} P(z), \tag{1.6}$$

$$\widetilde{P}_{\mathrm{L}}(z) = \widetilde{P}(z) \begin{pmatrix} 1 & \widetilde{s}(z) \\ 0 & 1 \end{pmatrix}, \tag{1.7}$$

式中， $s(z)$ 为一个劳伦斯多项式.

提升小波不仅继承了小波的优点，还具有如下性质[15−16]:

i) 由于利用了高低频滤通器间的相似性，运算速度更快;

ii) 因为可以直接用原始信号的小波变换系数替换原信号, 所以不需要额外储存空间;

iii) 提升小波正逆变换更加简单，只需要简单交换操作顺序即可;

iv) 小波变换可以提取出相互作用蛋白质对的数据特征，而提升小波变换则能减小因补零法带来的误差，更准确地保留原始信号信息.

1.3 基于提升小波的特征提取流程

i) 分别对两条蛋白质序列赋予与氨基酸残基对应的理化参数，得到两个数值序列 A 和 B, 并统一两条序列长度 (补零法);

ii) 分别对它们进行离散提升小波变换, 将两条序列由高频到低频分解为 3 层, 取 1 、2 层的高频信息和低频信息;

iii) 分别对各层数据序列作离散傅里叶变换，得到傅里叶变换系数序列与其对应的相位序列;

iv) 带入双交叉谱函数中，得到交叉谱系数序列;

v) 计算出每一层满足指定条件的信噪比值，公式如下：

$$S_k = \frac{\max\limits_{i\in n}\left(C_k^{(i)}, a < \left|P_{Ak}^{(i)} - P_{Bk}^{(i)}\right| < b\right)}{\overline{C}_k}, \tag{1.8}$$

式中， S_k 为对应分层的信噪比， $C_k^{(i)}$ 为第 k 层第 i 个交叉谱系数， $\overline{C}_k$ 为 k 层交叉谱系数组的平均数， $P_{Ak}^{(i)}, P_{Bk}^{(i)}$ 分别为蛋白质 A 和蛋白质 B 的相位值， n 为 C_k 中元素的总数， a, b 表示相位差的上下限，本文中分别设定为 4 和 2.

1.4 评价指标参数

一般所用到的评价指标参数包括: 特异性、敏感性、准确率、AUC 及 $\mathrm{AUC}_{0.5}$ 得分. 首先定义：P_t 为将阳性数据预测为相互作用蛋白质对的数量，N_t 为将阴性数据预测为非相互作用蛋白质对的数量，P_f 为将阳性数据预测为非相互作用蛋白质对的数量，N_f 为将阴性数据预测为相互作用蛋白质对的数量，N 表示所有样本的总数，$N = P_t + P_f + N_t + N_f$. 此时，特异度 $S_p = \frac{N_t}{N_t+N_f}$, 表示将阴性数据预测为阴性的百分比，敏感度 $E_s = \frac{P_t}{P_t+P_f}$, 表示将阳性数据预测为阳性的百分比，准确率为 $N = \frac{P_t+N_t}{P_t+P_f+N_t+N_f}$. 为了对已建立的

预测模型进行进一步的评估，分析预测结果的稳定性，引入了 ROC (receiving operator characteristic) 曲线分析，由此定义了 AUC (area under curve), 即 ROC 曲线下的面积大小， AUC 越大，模型的预测效果就越好， $AUC_{0.5}$ 则是当特异度小于等于 0.5% 时 AUC 的大小． $AUC_{0.5}$ 越大，预测结果可信度越高．本文使用准确率、 AUC 和 $AUC_{0.5}$ 进行预测效果的评估指标.

1.5　蛋白质相互作用的预测

利用机器学习方法来预测蛋白质相互作用一般分为 3 个步骤 (图 2):

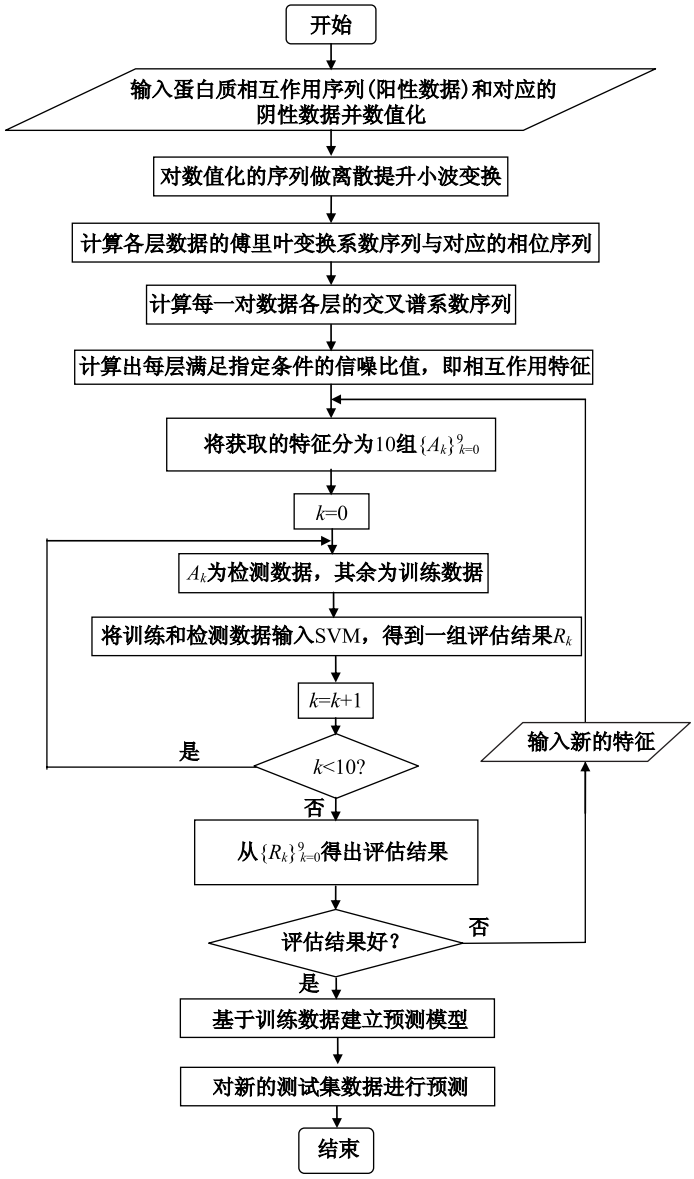

图 2　基于提升小波特征提取的蛋白质相互作用预测流程图

步骤一　选取特征, 计算出训练数据 (训练数据包括已知相互作用的蛋白质数据——阳性数据和对应的阴性数据) 的相互作用特征，目前最常用的特征是 3-频肽特征[6], 本文则是利用了提升小波得到的频谱特征.

步骤二　确立 SVM 的核函数和核函数的参数，将计算得到的训练数据特征代入到 SVM 中，为了确定训练后的预测模型稳定且可信，加入十次交叉验证的方法，即将训练数据特征随机地分为数据量都相同的 10 组数据，且 10 组数据中阳性数据和阴性数据的数量相等，并依次将它们记为第一组、第二组、……、第十组，按顺序分别将其中一组用作检测数据，其它 9 组作为训练数据，经过 10 次训练 – 检测的过程，可以得到 10 次计算后的 10 组评估参数，以平均值 ± 范围的方式来表示最终的评估结果.

步骤三　用训练数据生成的预测模型预测待测试的或者是相互作用未知的的蛋白质对，进而可以建立新的蛋白质相互作用网络.

2　结果与讨论

使用光滑性 (smooth) 和消失动量 (vanishing moment) 适中的小波和提升小波，如 db4 小波和 db4 提升小波，处理补零法所造成的误差都很难准确获取相互作用蛋白质的特征，因此本文采用的是光滑度更高的 db8 提升小波，选取理化参数 EIIP 和 IC 值[19]. 对于一个蛋白质作用对，按照本文所述的特征提取流程，最终得到了含有 8 个频率的特征向量，而相应地， 3-频肽特征[6] 向量则有 686 个元素.

经过计算得到，酵母相互作用蛋白质在各层信噪比的平均值都略高于随机配对蛋白质，之后以随机数据作为阴性数据，在支持向量机模型[20] (采用 C-SVM 支持向量机，核函数为 RBF 核函数) 中分析蛋白质相互作用特征的显著性. 首先，分别计算出酵母训练数据 (5 000 对) 的频谱特征和 3-频肽特征以及对应阴性数据的频谱特征和 3-频肽特征，然后计算出十次交叉的各评价指标参数.

表 1 列出了酵母蛋白质对应于频谱特征和 3-频肽特征所获得的各个评价指标参数的大小. 在相同的阳性数据下，以随机数据对作为比对数据时，使用 3-频肽特征可以得到更好的评估结果，这与 YU 等人[17] 的结果是一致；但是如果是以平衡法产生的数据作为比对数据时，提升小波获取的频谱特征有更好的评估结果. 提升小波获取的频谱特征对于经实验直接确认的，在物理上相互作用的蛋白质具有很好的评估结果，对于仅通过基因间关系推断出的相互作用蛋白质则没有很好的评估结果. 这说明 3-频肽特征相对于局部网络外的蛋白质对可以很好地抓取一个局部网络内相互作用蛋白质对的特征，而不能判别都属于一个局部网络中的蛋白质对是否相互作用；频谱特征则可以区分同属于一个局部网络中在物理上相互作用的蛋白质对和不相互作用的蛋白质对.

通过大量的数值实验发现，不同物种的相互作用网络特征不一样. 结果显示对于小鼠的蛋白质相互作用网络而言，小鼠相互作用蛋白质的 3-频肽特征无法被经参数优化过的支持向量机模型检测出来，但是它们的频谱特征却能被很好地识别出来 (表 2). 在利用基于频谱特征的预测模型预测测试数据后得到了 77% 以上的预测准确率 (表 3).

表 1　酵母蛋白质相互作用预测的评估结果

	提升小波		3-频肽	
	随机数据对	平衡阴性数据对	随机数据对	平衡阴性数据对
准确率	0.82±0.02	0.802±0.03	0.84±0.01	0.75±0.02
AUC	0.84±0.02	0.82±0.02	0.92±0.01	0.76±0.01
$AUC_{0.5}$	0.03	0.03	0.02	0.02

表 2　小鼠蛋白质相互作用预测的评估结果

	平衡阴性数据对 2 324 个蛋白质 4 811 对相互作用蛋白质	
	提升小波	3-频肽
准确率	0.83±0.03	0.624±0.03
AUC	0.84±0.02	0.67±0.03
$AUC_{0.5}$	0.01	0.01

表 3　基于特征小波的两类物种预测结果

	准确率	AUC	$AUC_{0.5}$
酵母	77.23%	0.78	0.01
小鼠	78.53%	0.79	0.01

最后，再从 uniprot 数据库中找出了 538 个小鼠蛋白质，蛋白质序列经过两两配对去冗余后作为待预测数据集，一共有 [((537+1)*538)/2]=144 453 对的数据. 用训练得到的频谱特征预测模型预测出这些数据中相互作用的蛋白质，最终得到了 2 045 对相互作用蛋白质. 由于局部网络中高连接度的蛋白质具有十分重要的生物学意义，如果它们自身的性质发生变化，整个网络的拓扑结构也将随之发生改变，从而影响细胞器的生物功能，因此，确定局部网络中的这些关键蛋白质以及与它们关联的蛋白质有着非常重要的意义. 在这个局部网络中，连接度高的蛋白质为 Q01097(74), P35438(59), Q9CQJ4(31), Q60631(30), 括号内的数为此蛋白质连接度. 在预测结果中包含了所有与 Q01097 和 P35438 已经确认的相互作用蛋白质；对于 Q9CQJ4 和 Q60631, 在预测的结果中，本文不仅捕捉到了已经实验确认的与之相互作用的蛋白质，同时还预测出了与它们有着潜在相互作用关系的蛋白质 (表 4).

由于蛋白质网络中充斥着大量蛋白质节点，如果在二维平面中直接观察网络，则在平面内某些区域上会出现重叠现象，为了可以更加直观形象地观察网络结构，将此基于枢纽蛋白质 Q01097, P35438, Q9CQJ4, Q60631 的相互作用网络以立体的方式呈现出来 (图 3).

立体图的原理是基于人的两只眼有一定距离，这使得两眼的视角会有所不同，由于视角的不同所看到是影像也会有一些差异，大脑会根据这种差异感觉到立体的景象. 利

用此原理，分别从两个不同的视角生成两个图片，当这两个图片在两只眼中重合时，就看到了立体的图像.

表 4　小鼠蛋白质 Q9CQJ4 及 Q60631 的预测结果

蛋白质	与 Q9CQJ4 相互作用的预测结果	蛋白质	与 Q60631 相互作用的预测结果
Q9WTX5	已确认	Q9Z1S8	已确认
Q9QWH1	已确认	Q9WU78	已确认
Q9JIA7	尚未确认	Q9QYY0	已确认
Q9QZ06	尚未确认	Q9JLQ0	已确认
A2AWL7	已确认	Q9H204	已确认
Q8CCI5	已确认	Q9ES52	已确认
Q6P1G2	已确认	Q9EQ32	已确认
Q64028	已确认	Q99PM9	已确认
O35730	已确认	Q99JZ7	已确认
O54833	已确认	Q8R550	已确认
Q5U4D9	已确认	Q62077	已确认
Q3UK78	已确认	Q60992	已确认
Q08639	已确认	Q60749	已确认
O55187	已确认	Q62083	尚未确认
O88974	已确认	Q6P9R2	尚未确认
P67871	已确认	Q8JZS0	尚未确认
P63017	已确认	Q99ML1	尚未确认
P61965	已确认	Q9WV55	尚未确认
P59178	已确认	Q02384	已确认
P23198	已确认	P98083	已确认
P23798	已确认	P54763	已确认
P25916	已确认	P06800	已确认
P27641	已确认	P35991	已确认
P28574	已确认	P35438	已确认
P30658	已确认	P35329	已确认
O88746	尚未确认	P34152	已确认
Q6PIC6	尚未确认	P51807	尚未确认
Q8R5C8	尚未确认	Q06507	尚未确认
Q922Q8	尚未确认	Q3UK78	尚未确认
O09061	尚未确认	Q5U4D9	尚未确认
Q99PM9	尚未确认		

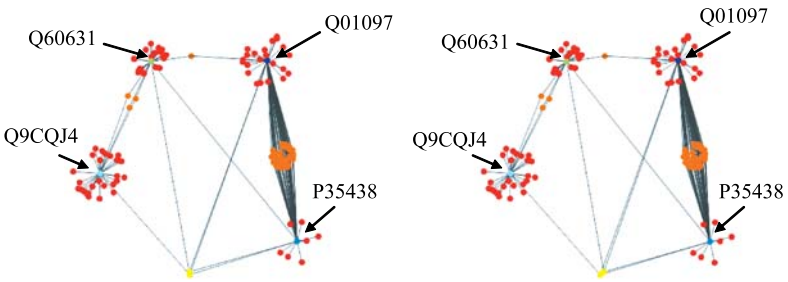

图 3　小鼠局部蛋白质立体相互作用网络图

3　结　语

一般机器学习方法的预测结果，会大大地受到高连接度的枢纽蛋白质的影响，而造成过拟合的问题，因此，需要根据阳性数据的结构来生成对应的阴性数据，从而得到可信的预测结果．在此规则的约束下，本文利用提升小波，从蛋白质的序列信息中提取了蛋白质间相互作用的频谱特征，用于相互作用的预测．经过大量的数值实验发现，蛋白质相互作用网络会随着物种或者来源的差异有着不同的特征．对于物理上相互作用的蛋白质，频谱特征相对于 3-频肽较易被识别，可以得到更高的预测准确率，而且由于频谱特征所使用的维数较小，它的计算时间会大大减少．

对于某些蛋白质相互作用网络，机器学习的方法往往不能通过 3-频肽特征检测出相互作用的蛋白质对，因此，在预测蛋白质相互作用时，不能只是选取单一的相互作用蛋白质数值特征，而需要尽可能地对特定的蛋白质集合进行分析，利用多种方法来选取它们的特征，之后才有可能构造出更加可信的预测模型，从而得到更可靠的预测结果．

最后，本文使用一个置信度高的小鼠预测模型，预测了小鼠体内蛋白质间的相互作用，构建了一个局部相互作用网络，并以立体的方式呈现．

参考文献

[1] Fields S, Song O. A novel genetic system to detect protein-protein interactions [J]. *Nature*, 1989, **340**(6230): 245–246.

[2] Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y. A comprehensive two-hybrid analysis to explore the yeast protein interactome [J]. *Proceedings of the National Academy of Sciences of the United States of America*, 2001, **98**(8): 4569–4574.

[3] Marcotte E M, Pellegrini M, Ng H L, Rice D W, Yeates T O, Eisenberg D. Detecting protein function and protein-protein interactions from genome sequences [J]. *Science*, 1999, **285**(5428): 751–753.

[4] Uetz P, Giot L, Cagney G, et al. A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae [J]. *Nature*, 2000, **403**(6770): 623–627.

[5] Gonzalez A J, Liao L. Predicting domain-domain interaction based on domain profiles with feature selection and support vector machines [J]. *BMC Bioinformatics*, 2010, **11**: 537.

[6] Shen J W, Zhang J, Luo X M, et al. Predictina protein-protein interactions based only on sequences information [J]. *Proceedings of the National Academy of Sciences of the United States of America*, 2007, **104**(11): 4337–4341.
[7] Sprinzak E, Sattath S, Margalit H. How reliable are experimental protein-protein interaction data? [J]. *Journal of Molecular Biology*, 2003, **327**(5): 919–923.
[8] Anfinsen C B. Principles that govern the folding of protein chains [J]. *Science*, 1973, **181**(96): 223–230.
[9] Pazos F, Valencia A. Similarity of phylogenetic trees as indicator of protein-protein interaction [J]. *Protein Engineering*, 2001, **14**(9): 609–614.
[10] Ben-Hur A, Noble W S. Kernel methods for predicting protein-protein interactions [J]. *Bioinformatics*, 2005, **21**: I38–I46.
[11] Valencia A, Pazos F. Computational methods for the prediction of protein interactions [J]. *Current Opinion in Structural Biology*, 2002, **12**(3): 368–373.
[12] Burger L, Nimwegen E V. Accurate prediction of protein-protein interactions from sequence alignments using a Bayesian method [J]. *Molecular Systems Biology*, 2008, **4**: 165.
[13] Cosic I. Macromolecular bioactivity: is it resonant interaction between macromolecules?—theory and applications [J]. *IEEE Transactions on Biomedical Engineering*, 1994, **41**: 1101–1114.
[14] Liu X, Wang Y F. A modified resonant to predict protein-protein interaction [J]. *Frontiers of Biology in China*, 2007, **2**: 268–271.
[15] Sweldens W. The lifting scheme: a construction of second generation wavelets [J]. *SIAM Journal on Mathematical Analysis*, 1998, **29**(2): 511–546.
[16] Sweldens W. The lifting scheme: a custom-design construction of biorthogonal wavelets [J]. *Applied and Computational Harmonic Analysis*, 1996, **3**(2): 186–200.
[17] Yu J T, Guo M Z, Needham C J, Huang Y C, Cai L, Westhead D R. Simple sequence-based kernels do not predict protein-protein interactions [J]. *Bioinformatics*, 2010, **26**: 2610–2614.
[18] Daubechies I, Sweldens W. Factoring wavelet transforms into lifting steps [J]. *Journal of Fourier Analysis and Applications*, 1998, **4**(3): 247–269.
[19] Veljkovic V, Veljkovic N, Este J A, Huther A, Dietrich U. Application of the EIIP/ISM bioinformatics concept in development of new drugs [J]. *Current Medicinal Chemistry*, 2007, **14**(4): 441–453.
[20] Chang C C, Lin C J. LIBSVM: a library for support vector machines [J]. *ACM Transactions on Intelligent Systems and Technology*, 2011, **2**(3): 27.