广西科学 Guangxi Sciences 2017,24(3):286～291

网络优先数字出版时间:2017-06-01 DOI:10.13656/j.cnki.gxkx.20170601.002

网络优先数字出版地址:http://kns.cnki.net/kcms/detail/45.1206.G3.20170601.1109.004.html

基于序列和结构特征的蛋白质自由能预测*

Protein Free Energy Prediction based on Sequence and Structure Features

鲁帮力[1],陈庆锋[1,2**],江家文[1],罗海琼[3]

LU Bangli[1],CHEN Qingfeng[1,2],JIANG Jiawen[1],LUO Haiqiong[3]

(1.广西大学计算机与电子信息学院,广西南宁 530004;2.广西大学亚热带农业生物资源保护与利用国家重点实验室,广西南宁 530004;3.广西医科大学信息与管理学院,广西南宁 530021)

(1.School of Computer,Electronics and Information in Guangxi University,Nanning,Guangxi,530004,China;2.State Key Laboratory for Conservation and Utilization of Subtropical Agro-bioresources,Guangxi University,Nanning,Guangxi,530004,China;3.School of Information and Management,Guangxi Medical University,Nanning,Guangxi,530021,China)

摘要:【目的】蛋白质自由能不仅能准确地反应蛋白质的交互,而且对药物设计有巨大帮助。因此,选择建立精确的蛋白质自由能回归模型是非常有必要的。**【方法】**收集135对蛋白质复合物并计算600个特征,通过最小冗余最大相关(mRMR)选择与蛋白质自由能显著相关的特征并去除冗余特征,从而得到最小冗余最大相关的特征集,用筛选后的特征建立6种回归模型,并对选择后的特征进行移除对比分析特征的重要性;最后通过10折交叉验证对比得到最佳模型,预测蛋白质自由能。**【结果】**相对于其它方法,本研究所建立的模型在预测135对蛋白质复合物的性能,相对于其它方法有着较高的相关系数和较低平均绝对误差。**【结论】**本实验所用方法比其他方法选出的模型有更好的预测精度。

关键词:蛋白质交互 自由能 特征选择 回归模型

中图分类号:TP399 **文献标识码:**A **文章编号:**1005-9164(2017)03-0286-06

Abstract:【Objective】Protein free energy not only can accurately reflect the protein interaction, but also can be a great help to drug design and disease treatment. Therefore,it is necessary to establish an accurate regression model of protein free energy.**【Methods】**In this article,135 proteins complexes were collected and 600 features were calculated. Minimum redundancy maximum relevance algorithm was used to select features which were significantly related to protein free energy and removed redundant features. This was able to obtain the minimum redundancy maximum relevance feature sets. The importance of features was further analyzed by comparing the performance change by removing features. The best model was chosen to predict protein free energy by comparing the result of 10-fold cross validation.**【Results】**The model had a higher correlation coefficient and lower average absolute error in predicting the performance of 135 pairs of protein complexes compared with other methods.**【Conclusion】**The experimental results show that our method has better prediction accuracy than other methods.

Key words: protein interaction, free energy, feature selection, regression model

收稿日期:2017-03-25

修回日期:2017-05-24

作者简介:鲁帮力(1991—),男,硕士研究生,主要从事生物信息学和数据挖掘研究。

*国家自然科学基金项目(61363025)和广西自然科学基金重点项目(2013GXNSFDA019029)资助。

**通信作者:陈庆锋(1972—),男,教授,主要从事数据挖掘和生物信息学研究,E-mail:qingfeng@gxu.edu.cn。

0 引言

【研究意义】蛋白质是组成生物的基础物质之一，蛋白质与蛋白质间相互作用所产生的功能对生命有重要意义，如 DNA 的合成、基因转录、蛋白质运输、信号转导等，但并非所有的蛋白质之间都能结合发生相互作用。研究蛋白质的交互作用对药物设计有很大的帮助。药物设计方法比较多样，比如计算诱变、分子对接以及从头设计等等，但这些方法都很难预测药物相互作用的过程，如果选择的两种蛋白质之间结合自由能很低的话，它们之间的相互作用就越低。此外，蛋白质交互[1]的细微变化都有可能改变复合物的功能，甚至给病人带来危害[2]。因此，通过计算预测蛋白质之间的自由能，对提高药物设计的效率以及理解蛋白质交互功能都有重大作用。**【前人研究进展】**近些年来，计算自由能的方法都是以提高预测精度和减少计算开销为目的。尽管一些方法能够比较准确地预测自由能，但是往往他们都有巨大的空间和时间开销，并且使用有局限的实验集。这些方法大致分为 3 大类：精确方法、终点法、经验公式。典型的精确方法有自由能微扰[3](Free Energy Perturbation，FEP)和热力学积分(Thermodynamic Integration，TI)[4]。FEP 和 TI 都是典型的计算自由能的精确方法，但是它们都需要一定的计算时间，并且只能使用在结合和未结合状态[5]上平均信息熵变化小的蛋白质上，有很大的局限性。基于经验力场公式的终点法计算速度快，例如线性交互能方法(Linear Interaction Energy，LIE)[6]和 Molecular Mechanics Poisson-Boltzmann Surface Area 方法(MM-PBSA)[3]。LIE 能计算蛋白质的结合和非结合状态下的平均静电和范德尔瓦斯交互自由能，它使用参数 α 和 β 来确定溶剂状态和蛋白质状态的内部势能的变化，但是参数值极易被不同的训练集影响。MM-PBSA 分解结合自由能为 $\triangle E_{MM}$、$\triangle G_{sol}$、$T\triangle S$，然后计算不同状态下蛋白质的不同自由能。但是，MM-PBSA 不适合在那些构象熵变化特别大的蛋白质上，因此它们的误差很大。由于上述方法的局限性，现在很多研究基于蛋白质特征建立模型来预测自由能。此类方法的模型计算精度有很大的提升，而且计算时间大大减少。但是预测结果依然不完美，而且许多与自由能相关的特征都被忽略。有的方法是计算 200 个蛋白质结构特征[7]构建回归模型，从而预测蛋白质自由能。但是这个回归模型中的许多蛋白质结构特征以不同的方式被重复计算，而且有些特征计算开销大、时间长，无关特征和冗余特征会影响预测模型的精度。因此，需要对特征集合使用最小冗余最大相关算法[8]进行筛选。除结构特征，也有些研究使用序列特征来预测自由能[9]。尽管序列特征能够提高模型的表现，但是结构特征却使用很少；另外他们没有使用信息增益或其他可靠的方法对特征进行选择。同时，因为使用的数据集较小，很多算法的模型并没有对特征重要性进行评估，这些训练样本过小以及未经过特征筛选的模型很容易过拟合。因此在模型构建上，有效的特征选择和特征重要性评估对蛋白质之间自由能预测特别重要。**【本研究切入点】**收集 135 对标有结合前后蛋白质变化的复合物，同时计算和选择那些与蛋白质自由能相关的结构特征和序列特征[10]，并对特征进行选择和重要性分析。**【拟解决的关键问题】**首先使用 Minimum Redundancy Maximum Relevance(mRMR) 特征选择方法除去不相关特征以及冗余特征，从而获得最好的特征集；然后建立 6 种回归模型并通过 10 折交叉验证进行对比来选择最佳的模型；最后对选择后的结构特征进行移除特征分析，判断和确定结构特征对模型性能的重要性，从而证明 mRMR 特征选择的意义；并用该模型训练和预测训练集 135 对蛋白质的自由能，然后与他人所建立的模型进行对比。

1 材料与方法

1.1 数据预处理

从文献[7]中收集 135 对蛋白质复合物，用于训练回归模型。每一对复合物都由 2 个蛋白质组成，每个蛋白质的氨基酸链都已经标出。根据功能分为 9 类：13 个抗体-抗原类(A)，5 个抗原-结合抗体类(AB)，34 个酶-抑制剂类(EI)，10 个酶-底物类(ES)，11 个有调节和附件链的酶(ER)，16 个 G-蛋白质(OG)，12 个受体(OR)，34 个混杂蛋白质(OX)。这些复合物都含有真实的来自实验结果的自由能(表 1)。

表 1　135 对蛋白质复合物的蛋白质编号、功能类型和自由能

Table 1　Protein ID, functional class and free energy value of 135 protein complexes

蛋白质 Protein	类型 Type	自由能 Free energy	蛋白质 Protein	类型 Type	自由能 Free energy	蛋白质 Protein	类型 Type	自由能 Free energy
1AHW_AB:C	A	11.55	3SGB_E:I	EI	14.51	2AJF_A:E	OR	10.63
1BJ1_HL:VW	AB	11.55	1E6E_A:B	ES	8.28	2HLE_A:B	OR	10.09
1BVK_DE:F	A	10.53	1EWY_A:C	ES	7.43	2I9B_E:A	OR	12.93
1DQJ_AB:C	A	11.67	1F6M_A:C	ES	7.6	2NYZ_AB:D	OR	12.69
1E6J_HL:P	A	10.28	1GLA_G:F	ER	6.76	1MLC_AB:E	A	9.61
1FSK_BC:A	AB	13.12	1IJK_A:BC	ER	10.42	2VIS_AB:C	A	7.36
1JPS_HL:T	A	13.64	1JMO_A:HL	ER	9.47	1AY7_A:B	A	13.23
1KXQ_H:A	AB	11.54	1JWH_CD:A	ER	11.14	1CBW_ABC:D	EI	10.75
1NCA_HL:N	AB	11.02	1KKL_ABC:H	ES	10.02	2PCB_A:B	ES	6.82
1P2C_AB:C	A	13.63	1M10_A:B	ER	11.24	2TGP_Z:I	EI	7.54
1VFB_AB:C	A	11.46	1NW9_B:A	ER	11.19	2WPT_A:B	EI	10.67
1WEJ_HL:F	A	12.48	1OC0_A:B	ER	12.28	1AVZ_B:C	OX	6.55
2I25_N:L	A	12.28	1R6Q_A:C	ER	8.84	2AQ3_A:B	OX	6.71
2JEL_HL:P	AB	11.59	1US7_A:B	ER	8.09	1ATN_A:D	OX	12.07
2VIR_AB:C	A	12.28	1WDW_BD:A	ER	12.72	1DE4_AB:CF	OX	9.78
1ACB_E:I	EI	13.05	1ZM4_A:B	ES	8.03	1FC2_C:D	OX	10.43
1AVX_A:B	EI	12.5	2A9K_A:B	ES	10.25	1H1V_A:G	OX	10.2
1BRS_A:D	EI	17.32	2MTA_HL:A	ES	7.42	1IB1_AB:E	OX	9.76
1BUH_A:B	EI	9.7	2OOB_A:B	ES	5.66	1KLU_AB:D	OX	7.28
1BVN_P:T	EI	15.06	2OOR_AB:C	ER	10.65	1KXP_A:D	OX	12.34
1DFJ_E:I	EI	18.05	2PCC_A:B	ES	7.91	1XQS_A:C	OX	7.08
1EAW_A:B	EI	14.06	1A2K_C:AB	OG	9.31	2B4J_AB:C	OX	10.86
1EMV_A:B	EI	18.58	1E96_A:B	OG	7.42	2C0L_A:B	OX	9.82
1EZU_C:AB	EI	13.77	1FQJ_A:B	OG	9.79	2VDB_A:B	OX	13.4
1F34_A:B	EI	14.19	1GRN_A:B	OG	9.03	1AKJ_AB:DE	OX	5.32
1FLE_E:I	EI	12.28	1HE8_B:A	OG	7.37	1MQ8_A:B	OX	7.53
1GXD_A:C	EI	11.3	1I2M_A:B	OG	15.83	1RLB_ABCD:E	OX	8.18
1JIW_P:I	EI	15.55	1I4D_D:AB	OG	7.46	1XD3_A:B	OX	8.9
1JTG_B:A	EI	12.82	1IBR_A:B	OG	12.07	1ZHI_A:B	OX	9.08
1MAH_A:F	EI	14.51	1K5D_AB:C	OG	12.77	2BTF_A:P	OX	7.69
1NB5_AP:I	EI	13.86	1LFD_B:A	OG	7.79	2HQS_A:H	OX	10.15
1OPH_A:B	EI	11.32	1NVU_Q:S	OG	7.43	2HRK_A:B	OX	10.98
1PXV_A:C	EI	12.97	1WQ1_R:G	OG	6.62	2OZA_B:A	OX	11.73
1R0R_E:I	EI	14.17	1Z0K_A:B	OG	6.98	3BP8_AB:C	OX	11.44
1YVB_A:I	EI	11.17	2FJU_B:A	OG	7.2	1AK4_A:D	OX	6.43
1ZLI_A:B	EI	12.04	3CPH_G:A	OG	8.84	1B6C_A:B	OX	8.94
2ABZ_B:E	EI	11.67	1NVU_R:S	OG	7.8	1EFN_B:A	OX	10.12
2B42_A:B	EI	12.11	1E4K_AB:C	OR	7.87	1FFW_A:B	OX	8.09
2J0T_A:D	EI	13.34	1EER_A:BC	OR	15.59	1GCQ_B:C	OX	6.51
2O3B_A:B	EI	15.68	1HCF_AB:X	OR	13.08	1GPW_A:B	OX	11.32
2OUL_A:B	EI	11.96	1KAC_A:B	OR	10.68	1H9D_A:B	OX	9.18
2PTC_E:I	EI	18.04	1KTZ_A:B	OR	8.92	1QA9_A:B	OX	7.16
2SIC_E:I	EI	13.84	1PVH_A:B	OR	9.52	1S1Q_A:B	OX	4.29
2SNI_E:I	EI	15.96	1RV6_VW:X	OR	13.86	2GOX_A:B	OX	12.08
2UUY_A:B	EI	11.26	1T6B_X:Y	OR	13.1	3BZD_A:B	OX	9.57

1.2　方法

1.2.1　总体架构

如图 1 所示，本实验主要分为 5 步：第一步是收集蛋白质复合物和特征数据；第二步是提取并计算序列和结构特征值；第三步是使用 mRMR 进行特征选择，从而得到最有价值的特征集；第四步是建立 6 种回归模型并进行 10 折交叉验证，通过最高相关性和最小 Mean Absolute Error(MAE)来选择最佳模型；第五步通过特征移除分析进一步证明特征价值。

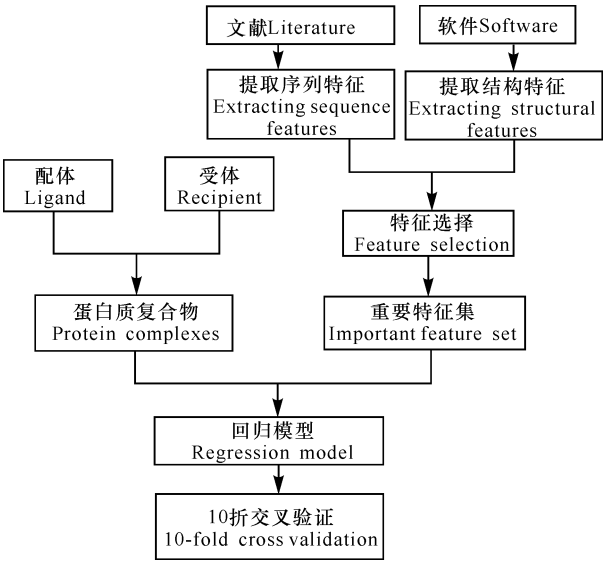

图 1　实验流程图

Fig. 1　The flow chart of experiment

1.2.2　特征计算

(1)序列特征计算

从 2 篇文献中共收集 592 种氨基酸的值(氨基酸的值代表不同环境下测试的氨基酸能量)。其中 544 种氨基酸的值来自 AAindex 数据库[11],剩下的 48 种氨基酸的值来自 Gromiha[12] 的研究。将蛋白质的氨基酸序列进行统计,得到每种氨基酸的数量,再将每种氨基酸的数量乘以对应的氨基酸的值,将相乘后得到的数相加便是蛋白质序列特征值。最终得到 592 个序列特征值。

(2)结构特征计算

原子表面积(Accessible Surface Area,ASA)的计算是基于 NACCESS[13] 软件,通过旋转一个 1.4 Å 半径探针原子在蛋白质表面计算。将一对蛋白质结合前的 ASA 减去结合后形成复合物的 ASA 便得到△ASA。得到的△ASA 的值越大,就越说明减少的表面积越多,相互作用接触的面积越多。

通过文献作者提供的 Vangone[14-15] 软件来计算接触的残基(ICs)的数量和非相互作用的表面(NIS)的比例,根据得到的结果进行分类。对于 ICs 得到带电-带电、带电-非极性、极性-极性、极性-非极性 4 种类型的数量作为特征,对于 NIS 得到带电、非极性、极性 3 种类型的比例作为特征。

1.2.3　特征选择和回归模型

将得到的序列特征和结构特征使用最小冗余最大相关(mRMR)算法进行过滤。mRMR 算法一共分为 2 步:根据相关性系数 r 来计算特征与蛋白质自由能之间的相关性,将非显著相关的特征作为不相关特征除去;再对特征与特征之间相关性进行计算,如果相关性为 1,说明这 2 个特征是同一种特征,因此去掉其中一个冗余特征,从而得到更加完善的特征集。这样得到的特征集是最小冗余最大相关特征。

$$r=\frac{\sum_{i=1}^{n}(Xi-\bar{X})(Yi-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(Xi-\bar{X})^2}\sqrt{\sum_{i=1}^{n}(Yi-\bar{Y})^2}}。$$

使用经过特征选择后的特征集建立回归模型,并预测蛋白质自由能。所建立的 6 种回归模型分别是 Support Vector Regression (SVR), Radial Basis Function Network (RBF network), Linear Regression, Sequential Minimal Optimization Regression (SMOreg), Additive Regression, Regression By Discretization。此外,对每种模型进行 10 折交叉验证来检测性能,然后选取性能最佳的模型来预测蛋白质自由能。对于回归模型性能的判断,选择相关系数 r 和平均绝对误差(MAE)。

$$\mathrm{MAE}=\frac{1}{n}\sum|Xi-\bar{X}|。$$

1.2.4　特征移除分析

特征对模型[16] 的性能有一定的影响,一个好的特征往往对模型性能的提高起到一定的作用。因此,当一个好的特征被移除时,模型的性能可能会降低。通过移除每个结构特征,对比移除该特征前后模型性能的变化来判断该结构特征对模型的价值。每次移除一个结构特征,将剩下的特征作为特征集来建立回归模型,然后对比移除前后性能的变化来判断该特征对模型的价值。对比过后再将特征放回到特征集中,移除另一个结构特征直到所有特征都被移除,从而对结构特征进行移除分析。其结果显示当结构特征被移除后回归模型的相关性有所降低。

2　结果与分析

通过最小冗余最大相关(mRMR)算法对得到的序列特征和结构特征进行过滤,最终得到 45 个序列特征和 5 个结构特征(表 2)。

通过 mRMR 来选择有价值的特征之后,将这些特征投入来建立 6 种回归模型来预测蛋白质自由能,并将这些模型通过 10 折交叉验证,表 3 中显示的是 6 种模型通过 10 折交叉验证后的性能。其中模型性能最好的是 Linear Regression 模型,它的 10 折交叉验证的性能明显比其他算法更加优越,其相关系数达到 0.502 9,比其它算法高;它的 MAE 为 1.973 8 kcal·mol^{-1},比其它算法都低。被选择的 5 个结构特征与蛋白质自由能具有显著性相关,表 4 显示他们

有着很高的相关性和很低的 P 值。

表 2　45 个序列特征和 5 个结构特征

Table 2　45 sequence features and 5 structure features

序列特征 Sequence feature	序列特征 Sequence feature	序列特征 Sequence feature	序列特征 Sequence feature	结构特征 Structure feature
CIDH920101	QIAN880110	QIAN880135	WOLS870103	△ASA
CIDH920105	QIAN880112	QIAN880137	NADH010102	Ics_charg-apolar
HOPT810101	QIAN880114	ROBB760101	NADH010103	Ics_polar-apolar
JANJ790102	QIAN880115	ROBB760108	KOEP990101	NISpol
LEVM760101	QIAN880122	ROBB760110	WOLR790101	NISchar
PRAM900101	QIAN880124	ROBB760112	KIDA850101	
QIAN880101	QIAN880125	ROBB760113	GUYH850104	
QIAN880102	QIAN880126	SNEP660102	CORJ870102	
QIAN880104	QIAN880130	SNEP660104	ENGD860101	
QIAN880105	QIAN880132	SWER830101		
QIAN880106	QIAN880133	VHEG790101		
QIAN880109	QIAN880134	WOLS870102		

表 3　6 种模型基于 10 折交叉验证后的性能

Table 3　The performance of 6 regression models are based on 10-fold cross-validation

回归模型 Regression models	相关系数 Correlation coefficient	平均绝对误差 Mean absolute error
SVR	−0.121 7	2.322 5
RBF network	0.306 8	2.127 3
Linear Regression	0.502 9	1.973 8
SMOreg	0.429 5	2.175 3
Additive Regression	0.382 2	2.200 5
Regression By Discretization	0.283 1	2.408 2

表 4　5 个结构特征与蛋白质自由能之间的相关性和 P 值

Table 4　The correlation coefficient and P value between five structure features and protein free energy

结构特征 Structure feature	相关系数 Correlation coefficient	P 值 P value
△ASA	−0.222	$P=0.005$
Ics_charg-apolar	−0.256	$P=0.001$
Ics_polar-apolar	−0.363	$P<0.000\ 1$
NISpol	−0.315	$P<0.000\ 1$
NISchar	0.326	$P<0.000\ 1$

将选择后的 5 个结构特征进行移除分析，观察移除后 Linear Regression 模型的 10 折交叉性能。表 5 显示移除每个结构特征后 Linear Regression 模型的相关性和 MAE 的值，从中可以看到相对于原来 Linear Regression 的相关性 0.502 9，移除后相关性有着明显的降低。

从 DFIRE[17]，PMF[18] 和 ICs/NIS[14] 提供的方法来预测 135 对蛋白质复合物的自由能。如图 2 所示，Linear Regression 模型预测结果相关系数为 0.66，DFIRE 和 PMF 为 0.34，ICs/NIS 为 0.48。从相关系数来看，Linear Regression 模型有着更好的性能。相关性越大，说明模型与蛋白质自由能有着很强的关联。此外，Linear Regression 模型性能在平均绝对误差上也是比较优秀的，该模型的误差为 1.973 8 kcal·mol^{-1}，DFIRE 为 4.82 kcal·mol^{-1}，PMF 为 3.22 kcal·mol^{-1}，ICs/NIS 为 1.893 7 kcal·mol^{-1}。平均绝对误差越小，说明在预测蛋白质自由能的值上，预测的结果可以更加精确，所造成的误差更小。

表 5　结构特征移除后的 Linear Regression 模型基于 10 折交叉验证的性能

Table 5　The performance of 10-fold cross-validation of Linear Regression using feature set by pruning structure features

移除的结构特征 Removed structure features	相关系数 Correlation coefficient	平均绝对误差 Mean absolute error
△ASA	0.518 9	1.923 6
Ics_charg-apolar	0.450 4	2.025 8
Ics_polar-apolar	0.484 4	1.939 7
NISpol	0.482 5	1.983 5
NISchar	0.539 4	1.862 9

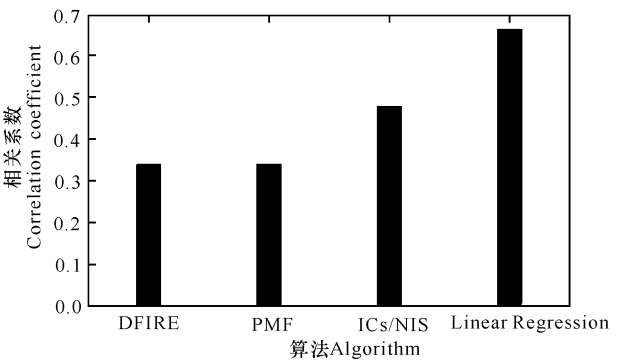

图 2　线性回归算法与其它方法性能的对比

Fig. 2　Performance comparison between Linear Regression algorithm and other methods

3 结论

最近几年，越来越多与蛋白质自由能相关的特征被计算出来。在此基础上，本研究提出一种基于序列和结构特征的蛋白质自由能预测方法：在特征计算上，计算蛋白质序列特征和结构特征；在特征选择上，使用 mRMR 来选择重要的特征，使得被筛选后的特征达到最小冗余最大相关，从而确保模型更加精确；在特征分析上，通过计算该特征移除前后模型性能的变化来判断该特征的价值，实验结果显示结构特征能提高模型的性能；在模型选择上，本研究建立多种模型基于 10 折交叉验证，通过选择最高相关性和最低平均绝对误差来选择最准确最可靠的模型；最终，将模型的性能与其它方法进行对比，得到基于序列特征和结构特征的 Linear Regression 模型，相对其他模型有着更准确的性能，其相关性为 0.66，MAE 为 1.973 8 kcal·mol^{-1}。

参考文献：

[1] VIDAL M, CUSICK M E, BARABÁSI A L. Interactome networks and human disease[J]. Cell, 2011, 144(6): 986-998.

[2] CHEN Q F, LUO H Q, ZHANG C Q, et al. Bioinformatics in protein kinases regulatory network and drug discovery[J]. Mathematical Biosciences, 2015, 262: 147-156.

[3] GILSON M K, ZHOU H X. Calculation of protein-ligand binding affinities[J]. Annual Review of Biophysics and Biomolecular Structure, 2007, 36(1): 21-42.

[4] BORDNER A J, MITTELMANN H D. Prediction of the binding affinities of peptides to class Ⅱ MHC using a regularized thermodynamic model[J]. BMC Bioinformatics, 2010, 11: 41.

[5] CHEN Q F, CHEN Y P P. Function annotation for pseudoknot using structure similarity[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2011, 8(6): 1535-1544.

[6] HANSSON T, MARELIUS J, ÅQVIST J. Ligand binding affinity prediction by linear interaction energy methods[J]. Journal of Computer-Aided Molecular Design, 1998, 12(1): 27-35.

[7] MOAL I H, AGIUS R, BATES P A. Protein-protein binding affinity prediction on a diverse set of structures [J]. Bioinformatics, 2011, 27(21): 3002-3009.

[8] YU L, LIU H. Feature selection for high-dimensional data: A fast correlation-based filter solution[C]//Proceedings of the 20th International Conference on Machine Learning. Washington, DC: ICML, 2003.

[9] YUGANDHAR K, GROMIHA M M. Protein-protein binding affinity prediction from amino acid sequence [J]. Bioinformatics, 2014, 30(24): 3583-3589.

[10] CHEN Q F, LAN W, WANG J X. Mining featured patterns of miRNA interaction based on sequence and structure similarity[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2013, 10(2): 415-422.

[11] KAWASHIMA S, POKAROWSKI P, POKAROWSKA M, et al. AAindex: Amino acid index database, progress report 2008[J]. Nucleic Acids Research, 2008, 36(S1): D202-D205.

[12] GROMIHA M M. A statistical model for predicting protein folding rates from amino acid sequence with structural class information[J]. Journal of Chemical Information and Modeling, 2005, 45(2): 494-501.

[13] HUBBARD S J, THORNTON J M. NACCESS 2.1.1. [R]. London: Department of biochemistry and molecular biology, University College.

[14] VANGONE A, BONVINA M J J. Contacts-based prediction of binding affinity in protein-protein complexes [J]. eLife, 2015, 4: e07454.

[15] KASTRITISP L, RODRIGUES J P G L M, FOLKERS G E, et al. Bonvin: Proteins feel more than they see: Fine-tuning of binding affinity by properties of the non-interacting surface[J]. Journal of Molecular Biology, 2014, 426(14): 2632-2652.

[16] CHEN Q F, CHEN Y P P. Mining protein kinases regulation using graphical models[J]. IEEE Transactions on Nano Bioscience, 2011, 10(1): 1-8.

[17] LIU S, ZHANG C, ZHOU H Y, et al. A physical reference state unifies the structure-derived potential of mean force for protein folding and binding[J]. Proteins: Structure, Function, and Bioinformatics, 2004, 56(1): 93-101.

[18] SU Y, ZHOU A, XIA X F, et al. Quantitative prediction of protein-protein binding affinity with a potential of mean force considering volume correction[J]. Protein Science, 2009, 18(12): 2550-2558.

（责任编辑：米慧芝）