

蛋白质表面氨基酸特性研究进展*

荆 艳 宋晓峰[△]

(南京航空航天大学生物医学工程系 江苏 南京 210016)

摘要: 分布在蛋白质分子表面的暴露于溶剂的氨基酸所具有的一些特性对蛋白质的折叠和聚合过程、蛋白质-蛋白质相互作用以及蛋白质的功能具有重要影响。本文分析了蛋白质表面氨基酸在疏水性、保守性、静电势及结构方面的一些特性, 阐述了近年来国际上利用这些特性对蛋白质-蛋白质相互作用界面进行预测的方法, 最后介绍了几款预测蛋白质表面氨基酸的软件。

关键词: 蛋白质表面特性; 蛋白质相互作用界面; 预测方法

中图分类号: Q71 Q517 文献标识码: A 文章编号: 1673-6273(2011)10-1991-04

Progress in Properties of Amino Acids on Surface of Protein Molecules*

JING Yan, SONG Xiao-feng[△]

(Department of Biomedical Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing, 210016, China)

ABSTRACT: The properties of the surface amino acids that exposed to the solvent can influence the fold and polymerization processes, protein-protein interaction and the function of proteins. This paper analyzes the properties of surface amino acids such as hydrophobicity, conservation, electrostatic potential and structural properties. Also, the methods using these properties to predict the interfaces between two interacting proteins in the international filed were reviewed. Finally we introduced several softwares of predicting surface amino acids.

Key words: The properties of protein surface; Protein-protein interface; Prediction methods

Chinese Library Classification(CLC): Q71 Q517 **Document code:** A

Article ID: 1673-6273(2011)10-1991-04

前言

多数蛋白质在细胞内以三维空间结构的形式存在, 分布在蛋白质表面的氨基酸的一些特性(如疏水性)对蛋白质折叠和聚合过程具有重要影响, 蛋白质分子之间的相互作用也受到蛋白质表面氨基酸性质的影响。研究表明蛋白质的催化作用、分子识别作用等功能主要发生在蛋白质分子的表面^[1]。因此, 对蛋白质表面氨基酸特性的研究有利于深入探究蛋白质的功能。

随着结构基因组学的不断发展, 生物信息的不断积累, 出现了越来越多的已知结构未知功能的蛋白质, 准确、快速地对这些蛋白质的功能进行注释显得尤为重要。通过分析蛋白质分子表面氨基酸的特性, 利用生物信息学的方法来识别蛋白质的一些相互作用域及功能位点在后基因组时代是非常有效的方法。

1 概述

蛋白质分子的表面形状很不规则, 对蛋白质表面特性进行分析的第一步就是要定义蛋白质表面。

1.1 溶剂可及表面

蛋白质折叠过程中疏水的氨基酸倾向于埋藏在分子内部, 为了定量描述蛋白质分子疏水性氨基酸的埋藏情况, BK Lee等引进了溶剂可及表面积的概念^[2]。溶剂可及表面是一个“假想曲面”, 是指探针分子在蛋白质表面滚动时探针分子球心所经

过的轨迹(见图1)。图2为一个蛋白质分子的溶剂可及表面示意图(蛋白质来自PDB数据库, ID:1L1O)。

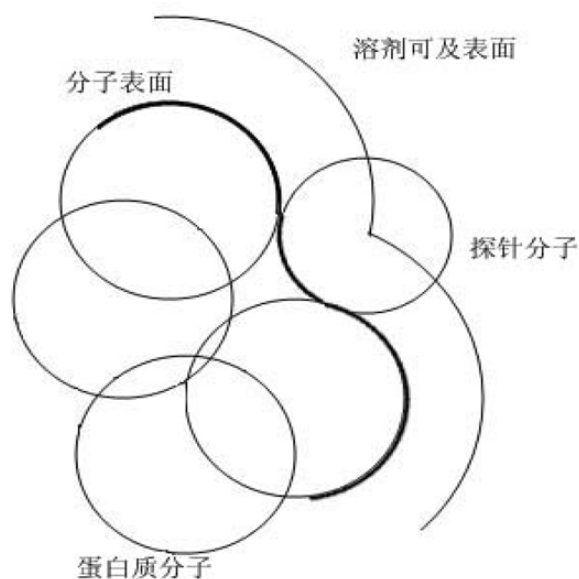


图1 蛋白质分子溶剂可及表面和分子表面示意图

Fig.1 Solvent accessible surface and molecular surface

* 基金项目: 江苏省自然科学基金资助(BK2010500)

作者简介: 荆艳(1985-), 女, 硕士研究生, 研究方向: 生物信息学。E-mail: fffjy@126.com

[△]通讯作者: 宋晓峰 Email: xfsong@nuaa.edu.cn

(收稿日期: 2010-12-07 接受日期: 2010-12-31)

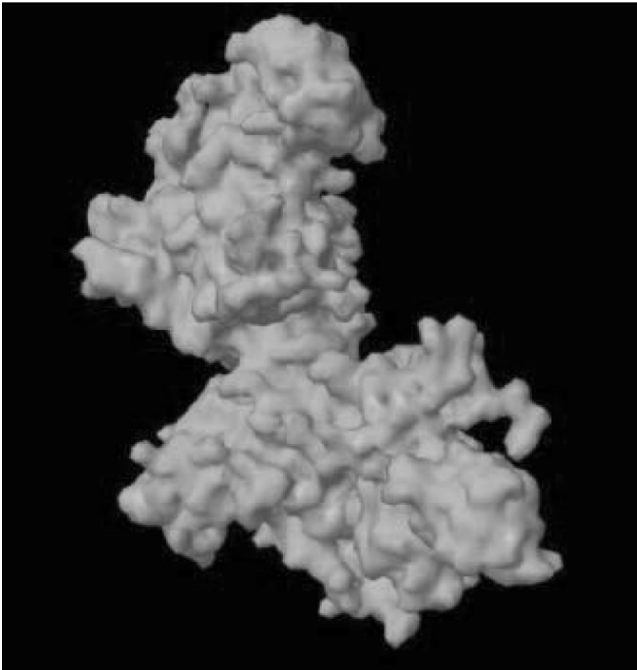


图2 溶剂可及表面示意图

Fig.2 Solvent accessible surface

1.2 蛋白质分子表面

蛋白质分子表面是指探针分子在蛋白质表面滚动时探针分子的内凹面所经过的轨迹(见图1)。相对来说,溶剂可及表面要比分子表面大一些。图3为一个蛋白质的分子表面三维图。(该蛋白质来自PDB数据库, ID:1L1O)。

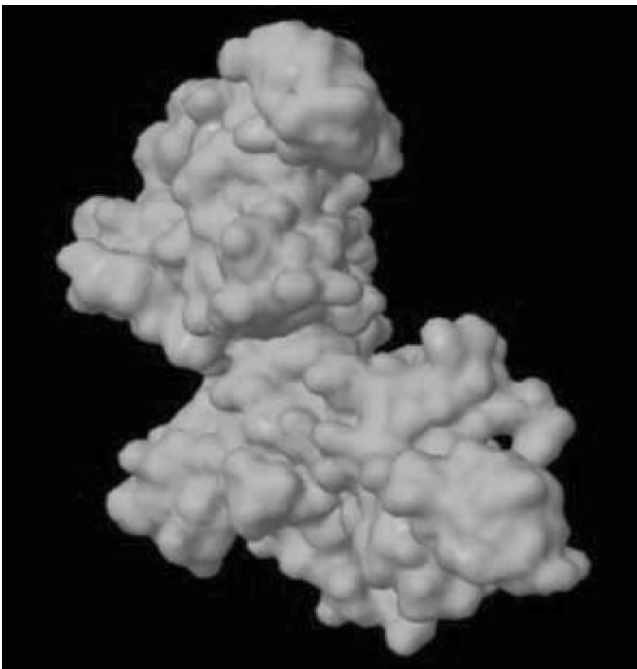


图3 分子表面示意图

Fig.3 Molecular surface

2 蛋白质相互作用域内氨基酸的特性

PDB数据库中约有结构基因组学解析的8000种蛋白质结构,其中约有2000种蛋白质功能未知^[5,28]。蛋白质的相互作用位点与蛋白质的功能有关,识别蛋白质相互作用位点对理解蛋

白质的功能非常必要。蛋白质相互作用位点主要位于蛋白质分子的表面,了解蛋白质分子表面的一些特性,对于识别相互作用域、解释蛋白质功能具有重要的意义。

关于蛋白质相互作用域内氨基酸的保守性,前人从多方面进行了研究。Grishin, Phillips等研究小组对不同蛋白质家族相互作用域内氨基酸的保守性与蛋白质表面其他区域内氨基酸的保守性进行了比较分析,得到蛋白质相互作用界面内的氨基酸比蛋白质表面上其他区域内的氨基酸的保守性高^[3-6]。Thornton等人分析了六类同源二聚物的相互作用,发现相互作用界面内氨基酸的保守性比表面其他区域内氨基酸的保守性高得多^[7,26]。Caffrey对一个更大的数据集进行分析得到了相同的结论^[6,12]。Tseng运用贝叶斯方法估计氨基酸的保守性,发现在蛋白质核上氨基酸的替代率与暴露在蛋白质表面的氨基酸的替代率有很大的差异^[22]。结合前人的很多研究结果,能够判定蛋白质表面上不同区域内氨基酸的保守性存在差异,但是仅仅蛋白质相互作用域内氨基酸的保守性并不能够准确地识别蛋白质的相互结合位点,因此需要将蛋白质表面氨基酸保守性的差异与表面其他的特性综合考虑来预测蛋白质的相互作用域。

蛋白质在卷曲折叠过程中,疏水性氨基酸倾向于远离溶剂而埋藏在蛋白质分子内部,极性氨基酸倾向于暴露在蛋白质分子表面与溶剂接触。但是对一些已知结构的蛋白质表面性质的分析表明,一些疏水基团也会出现在蛋白质表面,使蛋白质表面具有一定的疏水性^[29]。Jan Gruber等人研究过程中,发现蛋白质相互作用界面内氨基酸的疏水性远远高于蛋白质表面其他区域内氨基酸的疏水性^[8-11]。与暴露在表面其他区域内的极性氨基酸相比,蛋白质相互作用域内氨基酸的疏水性也可以作为识别蛋白质相互作用域的一个重要特征。

蛋白质分子的静电势与分子的性质密切相关。静电能为正的区域易发生亲电型反应,静电能为负的区域易发生亲核型反应,静电势能的大小反映了分子反应活性的差异,静电势能绝对值越大,反应活性越强^[30]。蛋白质-蛋白质相互作用域具有静电势互补的特点,可以通过计算蛋白质表面上划分的不同区域内静电势的线性相关系数来估计相互作用界面的静电势互补性^[8,18,20]。蛋白质表面不同区域的静电势也可以作为区分相互作用域和非相互作用域的一个指标。

Ofran等人将蛋白质的相互作用界面分成六类:同源二聚物相互作用界面、异源二聚物相互作用界面、暂时相互作用界面、永久相互作用界面、相同位点相互作用界面和不同位点相互作用界面,并对这六类蛋白质相互作用界面内氨基酸倾向性进行了分析,发现仅仅用氨基酸的组成和氨基酸相互接触的偏好性就可以识别蛋白质-蛋白质相互作用界面,且准确率可以达到63%-100%^[21]。可见氨基酸的倾向性是识别相互作用界面的一个重要特征。

蛋白质表面的结构特征对相互作用位点的识别也有一定影响。Jones, Thornton在分析蛋白质相互作用界面的过程中发现相互作用域表面的形状和氨基酸的溶剂可及性可以用来区分蛋白质相互作用域和蛋白质表面其他区域,相互作用域趋向存在于蛋白质分子表面相对平整、溶剂可及性高的区域^[13]。平均来说,相互作用界面的溶剂可及表面积是1600埃^[14]。Darby Tien-Hao Chang等人分析蛋白质表面氨基酸的溶剂可及表

面积的基础上对特征进行整合,通过分类的方法实现了对蛋白质—蛋白质相互作用域的预测^[27]。

3 基于蛋白质表面氨基酸特性对相互作用界面的预测

对于结合位点的识别并没有严格的规则,因此蛋白质表面的很多特性如:疏水性、保守性、静电势、氨基酸的倾向性及表面的结构特征等对氨基酸相互作用位点的识别具有重要作用。运用生物信息学的方法对这些特性进行综合考虑,识别蛋白质

相互作用界面是可行的,这也将使快速、准确地对已知结构未知功能的蛋白质进行功能注释成为可能。这些特征可以用于机器学习方法和经验得分函数等多种方法中。多数方法是用已知的结合位点训练模型进行预测,这样预测的准确率也取决于训练数据集的选取,一个可信度高、信息丰富、非冗余的数据集对模型的训练是非常关键的。表 1 总结了近年来几种采用机器学习方法利用蛋白质表面特性对相互作用域进行预测的方法。

表 1 几种预测蛋白质相互作用域的方法及主要特征^[20]
Table 1 Several methods for predicting PPIs and major features^[20]

Web sever	Method	Major feature
Cons-PPISP ^[15] http://pipe.scs.fsu.edu/ppisp.html	Neural Network	Solvent accessibility and sequence information
InterProSurf ^[16] http://curie.utmb.edu/ http://curie.utmb.edu/	Scoring Function	Solvent accessible surface area, propensity of interface residues
Patch Finder Plus ^[17] http://pfp.technion.ac.il/	Neural Network	Surface concavity, surface area, amino acid hydrophobicity, propensity of interface residues,
PPI-Pred ^[18] http://bioinformatics.leeds.ac.uk/ppi-pred	SVM	surface shape and solvent accessible surface area Solvent accessibility
ProMate ^[23] http://biportal.weizmann.ac.il/promate/	Scoring Function	Solvent accessible surface area, propensity of interface residues

PPI-Pred 和 ProMate 的预测准确率相对较高、应用较广泛,且采用了两种不同的方法。下面对这两种方法做简单介绍:

Bradford 等人通过对蛋白质表面氨基酸簇进行分析,运用支持向量机(SVM)方法实现对蛋白质—蛋白质结合位点的预测,利用 SVM 对蛋白质结合位点的特性(保守性、静电势、疏水性、作用域内氨基酸的倾向性、表面形状和溶剂可及表面积)进行训练,得到一个能够将相互作用域和非相互作用域区分开来的模型。PPI-Pred 对数据集中 180 条蛋白质的预测准确率为 76 %^[18]。ProMate 是基于得分函数的方法来识别蛋白质—蛋白质相互作用域的,该算法通过对异源二聚物的蛋白质复合物进行训练,提取出表面氨基酸的倾向性、残基的保守性、表面的几何特性等特征并对它们进行优化组合,得到一个最适合预测该类蛋白质相互作用位点的模型。ProMate 预测蛋白质相互作用界面的正确率为 70 %^[23,24]。

4 蛋白质表面氨基酸预测软件分析

在蛋白质空间结构上位于表面的氨基酸对于蛋白质—蛋白质相互作用域的预测及蛋白质功能方面的研究具有决定性的作用,准确确定哪些氨基酸位于蛋白质空间结构的表面则显得非常重要。因此国际上许多研究人员开发了一些基于蛋白质序列和结构预测表面氨基酸的软件,这些软件能够较准确、快速的预测出哪些氨基酸在蛋白质分子的表面,对蛋白质功能及药物的研究做出了一定贡献。

4.1 基于蛋白质序列的表面氨基酸预测软件

NetSurfP 和 Protein Solvent Accessible Surface Area Predictor (ASAP)是两款基于蛋白质序列的预测软件,均可以实现在线预测。NetSurfP 预测软件的准确率为 79 %,主要采用双层神经网络实现对蛋白质可及表面积的预测,该软件可以输入单条序列也可以大批量的输入多条序列^[25]。该软件首先对每条蛋白质序列进行二级结构的预测,将二级结构信息和该序列对应的 PSSM 矩阵作为第一层神经网络的输入,第一层的输出和 PSSM 矩阵作为第二层神经网络的输入,第二层神经网络的输出作为最终结果。该软件的输出结果主要包括了以下参数:氨基酸分类(分为表面氨基酸和内部氨基酸)、序列名称、氨基酸序号、相对可及性表面积、绝对可及表面积、每个氨基酸以 α 螺旋、 β 折叠及不规则卷曲结构存在的可能性等。ASAP(<http://ccb.imb.uq.edu.au/ASAP/>)预测软件一次只能接收一条氨基酸序列,输出结果包括 ASA 谱图、绝对可及表面积、相对可及表面积等参数。

4.2 基于蛋白质结构的表面氨基酸预测软件

CCP4 6.1.13 (<http://www.ccp4.ac.uk/>) 和 NACCESS V2.1.1 (<http://www.bioinf.manchester.ac.uk/naccess/>) 是两款基于蛋白质结构的计算蛋白质溶剂可及性的软件。PDB 格式的数据给出了蛋白质中每个原子的坐标以及温度因子等参数,这两款软件可以根据 PDB 文件提供的信息比较准确的计算出蛋白质中各个氨基酸的溶剂可及性,以溶剂可及表面积的形式体现出来。这两款软件可应用于多个操作系统中,比如:Linux、Unix、Windows。通常情况下,在对溶剂可及表面积进行计算时将探针分子的半径取为 1.4 埃,不过这两款软件均允许根据实际情

况修改此参数。

5 小结

蛋白质—蛋白质相互作用域内的氨基酸与蛋白质表面其他区域内的氨基酸在物理、化学性质及结构等方面存在的差异,使得利用统计学的方法预测蛋白质—蛋白质相互作用域成为可能。蛋白质相互作用域的识别有利于深入挖掘蛋白质的功能,有利于药物的研制。目前疫苗的研究在预防各种疾病的过程中发挥了重要的作用。识别 B、T 细胞上抗原决定簇对疫苗的开发具有重要意义^[31]。由于抗原决定簇是决定抗原性的特殊化学集团,大多存在于抗原物质的表面,因此利用蛋白质表面氨基酸的一些特性应用生物信息学的方法来识别 B、T 细胞上的抗原决定簇是可行的。相信随着进一步对蛋白质表面氨基酸特性的研究,在药物的研制、疾病的治疗方面都会有一定的应用前景。

参考文献(References)

- [1] David Lee, Oliver Redfern, Christine Orengo. Predicting protein function from sequence and structure [J]. *Nature Reviews Molecular Cell Biology*, 2007, 8(12): 995-1005
- [2] BK Lee, FM Richards. The interpretation of protein structures: estimation of static accessibility [J]. *J Mol Biol*, 1971, 55(3): 379-400
- [3] Grishin NV, Phillips MA. The subunit interfaces of oligomeric enzymes are conserved to a similar extent to the overall protein sequences [J]. *Protein Sci*, 1994, 3(12): 2455-2458
- [4] Ouzounis C, Perez-Irratxe C, Sander C, et al. Are binding residues conserved [J]. *Pac Symp Biocomput*, 1998, 3(1): 401-412
- [5] Bartlett GJ, Porter CT, Borkakoti N, et al. Analysis of catalytic residues in enzyme active sites [J]. *J Mol Biol*, 2002, 324(1): 105-121
- [6] Caffrey DR, Somaroo S, Hughes JD, et al. Are protein protein interfaces more conserved in sequence than the rest of the protein surface [J]. *Protein Sci*, 2004, 13(1): 190-202
- [7] William SJ, Valdar, Janet M Thornton. Protein Protein Interfaces: Analysis of Amino Acid Conservation in Homodimers [J]. *Proteins*, 2001, 42(1): 108-124
- [8] Jan Gruber, Alexander Zawaira, Rhodri Saunders C, et al. Computational analyses of the surface properties of protein—protein interfaces [J]. *Acta Cryst*, 2007, 63(1): 50-57
- [9] Young L, Jernigan RL, Covell DG. A role for surface hydrophobicity in protein protein recognition [J]. *Protein Sci*, 1994, 3(5): 717-729
- [10] Glaser F, Steinberg DM, Vakser IA, et al. Residue frequencies and pairing preferences at protein protein interfaces [J]. *Proteins*, 2001, 43(2): 89-102
- [11] James R Bradford, David R Westhead. Improved prediction of protein protein binding sites using a support vector machines approach [J]. *Bioinformatic*, 2005, 21(8): 1487-1494
- [12] Bradford JR, Westhead DR. Asymmetric mutation rates at enzyme-inhibitor interfaces: implications for the protein-protein docking problem [J]. *Protein Sci*, 2003, 12(9): 2099-2103
- [13] Jones S, Thornton JM. Analysis of protein protein interaction sites using surface patches [J]. *J Mol Biol*, 1997, 272(1): 121-132
- [14] Lo Conte L, Chothia C, Janin J. The atomic structure of protein protein recognition sites [J]. *J Mol Biol*, 1999, 285(5): 2177-2198
- [15] Zhou HX, Shan Y. Prediction of protein interaction sites from sequence profile and residue neighbor list [J]. *Proteins*, 2001, 44 (3): 336-343
- [16] Negi SS, Schein CH, Oezguen N, et al. InterProSurf: a web server for predicting interacting sites on protein surfaces [J]. *Bioinformatics*, 2007, 23(24): 3397-3399
- [17] Stawiski EW, Gregoret LM, Mandel-Gutfreund Y. Annotating nucleic acid-binding function based on protein structure [J]. *J Mol Biol*, 2003, 326(4): 1065-1079
- [18] Bradford JR, Westhead DR. Improved prediction of protein—protein binding sites using a support vector machines approach [J]. *Bioinformatics*, 2005, 21(8): 1487-1494
- [19] Porollo A, Meller J. Prediction-based fingerprints of protein—protein interactions [J]. *Proteins*, 2007, 66(3): 630-645
- [20] Nurcan Tuncbag, Gozde Kar, Ozlem Keskin, et al. A survey of available tools and web servers for analysis of protein-protein interactions and interfaces [J]. *Bioinformaticis*, 2009, 10(3): 217-232
- [21] Ofra Y, Rost B. Analysing six types of protein—protein interfaces [J]. *J Mol Biol*, 2003, 325(2): 377-387
- [22] Tseng YY, Liang J. Estimation of amino acid residue substitution rates at local spatial regions and application in protein function inference: a Bayesian Monte Carlo approach [J]. *Mol Biol Evol*, 2006, 23(2): 421-436
- [23] Neuvirth H, Raz R, Schreiber G. ProMate: a structure based prediction program to identify the location of protein-protein binding sites [J]. *J Mol Biol*, 2004, 338(1): 181-199
- [24] Liang S, Zhang C, Liu S, et al. Protein binding site prediction using an empirical scoring function [J]. *Nucleic Acids Res*, 2006, 34(13): 3698-3707
- [25] Bent Petersen, Thomas Nordahl Petersen, Pernille Andersen, et al. A generic method for assignment of reliability scores applied to solvent accessibility predictions [J]. *BMC Structural Biology*, 2009, Doi: 10.1186/1472-6807-9-51
- [26] Valdar WS, Thornton JM. Protein protein interfaces: analysis of amino acid conservation in homodimers [J]. *Proteins*, 2001, 42(1): 108-124
- [27] Darby Tien-Hao Chang, Yu-Tang Syu, Po-Chang Lin. Predicting the protein-protein interactions using primary structures with predicted protein surface [J]. *Bioinformatics*, 2010, Doi: 10.1186 / 1471-2105-11-S1-S3
- [28] Janez Konc, Dušana Janežić. ProBiS algorithm for detection of structurally similar protein binding sites by local structural alignment [J]. *Bioinformatics*, 2010, 26(9): 1160-1168