基于结构比对的蛋白结合位点预测方法①

刘广钟, 朱佳莉

(上海海事大学 信息工程学院, 上海 201306)

摘 要: 蛋白质通过结合位点与其他分子产生相互作用, 所以对蛋白结合位点的预测具有重要的意义. 现有许多不同的预测方法, 但是这些方法存在命中率低或计算量大的问题, 本文引入了一种基于结构比对的蛋白质位点预测方法, 同时在结构比对过程中引入同源索引, 找出相应的同源模版, 并与之进行结构比对, 然后将结构相似的模版中的配体映射到目标蛋白质中, 采用聚类方法对位点进行分析. 结果表明, 与其他预测方法相比, 本文的方法降低了计算量, 并提高了预测精度.

关键词: 结合位点; 同源索引; 结构比对; 聚类

Binding Sites Prediction Method Based on Structural Alignment

LIU Guang-Zhong, ZHU Jia-Li

(Shanghai Martime University, Information Engineering College, Shanghai 201306, China)

Abstract: Proteins interact with other molecules through binding sites, so it is significant to identify protein binding sites. Although there are different computational methods for the identification of binding sites, the existing prediction methods have problems of low hit rate or large computation. In this paper, a binding sites prediction method based on structural alignment is introduced. In the process of structural alignment, homologous index is applied to screening out homologous templates, with which query chains are aligned, and then the ligands in similar structure templates are mapped onto the query chains. Clustering method is used for analysis of sites. The result indicates that reduced computation and improved prediction accuracy, compared with other prediction methods, can be obtained through our method.

Key words: binding sites; homologous index; structural alignment; clustering

蛋白质是生命物质的基础, 是构成生命体的基本成分. 蛋白质也是生命功能活动的执行者, 主要通过与其他小分子之间的相互作用来实现相应的生物功能, 这些相互作用包括蛋白–配体小分子之间、抗原与抗体之间、蛋白–DNA 之间和蛋白–蛋白之间的结合作用[1], 其中蛋白–配体之间结合作用更能体现蛋白质的生物功能. 因此, 对蛋白质结合位点预测的研究具有重要的意义.

目前, 已经有许多计算方法用于预测和分析结合位点, 可将这些方法分成三类:

第一类是基于几何的计算方法, 研究蛋白质表面的凹槽, 把最大凹槽处的残基作为蛋白质的结合位点[2]. POCKET[3]、LIGSITE[4]、PASS[5]、SURFNET[6]、CAST[7]和 PocketPicker[8]等算法都利用几何的方法来研究蛋白质结合位点. 此类方法主要研究蛋白质表面的凹槽, 认为蛋白质表面最大的凹槽就是蛋白质的结合位点, 所以此类计算方法复杂性较低, 但这类方法并没有考虑氨基酸残基和配体的生化性质, 相对其他方法来说, 命中率并不高.

第二类是基于能量的计算方法, 例如, Q-siteFinder[9]、Grid[10]等, 此类方法将探针小球放在蛋白质表面, 简单的计算两者之间的范德华力或其他作

① 基金项目:国家自然科学基金(61303099)

收稿时间:2015–01–25;收到修改稿时间:2015–03–18

用力, 然后根据空间距离对容易结合的原子进行聚类, 最后根据簇的总结和作用力进行排序, 从而得到最有可能的配体结合位点. 在计算作用力的时候花费的时间较长, 所以虽然这类方法要比一般的几何算法的命中率要高很多, 但其计算量很大, 导致预测效率不高.

第三类是基于模版(序列比对或结构比对)的方法, 通过与同源模版进行序列或者结构比对, 来确定蛋白质相应的结合位点, 为结合位点的预测打开了一条新思路. Capra 等[11]认为配体的结合位点在蛋白质家族进化过程中是高度保守的, 基于这个假说, 采用序列比对的方式, 研究分析蛋白质进化过程中的保守氨基酸, 达到预测蛋白质结合位点的目的. 在蛋白质相似性比对方面, 由于蛋白质的结构比序列更为保守, 因此, 基于结构的比对方法与单纯的序列比对方法相比, 更有利于找出结构和功能相似的蛋白质, 能更加准确的预测结合位点残基[12]. 2009 年, Jeffrey Skolnick 和 Michal Brylinski 提出了一种基于蛋白质结构比对的预测配体结合位点和功能注释的算法 FINDSITE[13], 该算法基于折叠识别确定的结果模版叠加处的配体所具有的相似性结合位点来预测结合位点和配体[14], 使用 TM-align[15]蛋白质结构比对算法把同源模版叠加到目标蛋白质上, 通过配体聚类中心来预测假定的结合位点, 最后根据聚类中同源模版的数量对预测所得的结合位点进行排序. 2010 年 Wass 等人提出的 3DLigandSite 方法利用 MAMMOTH[16]工具来识别与目标蛋白质结构相似的蛋白结构, 把这些蛋白质上配体叠加到目标蛋白质上, 聚类后, 选择配体数量最多的簇, 作为预测的依据[17]. 2012 年 Roy 等人研究的 COFACTOR 算法在位点预测方面, 采用全局和局部结构比对的方式, 从一个完整的结合位点模版数据库中寻找结构相似的模版[18]. 2013 年 Yang 等人也提出了类似的预测方法 TM-SITE, 首先生成一个模版库 SSFL, 然后只用模版中的位点残基与目标蛋白进行比较, 这样弥补了全局和局部比对的缺陷, 在排列聚类时, TM-SITE 方法计算了每个簇的置信分数 CS_t, 它与簇中的模版配体数量有关, 最后根据 CS_t 的大小对簇进行排序[19], 从而选择预测结合位点的簇. 2011 年 CASP(Critical Assessment of protein Structure Prediction, 蛋白结构预测方法评估)证实, 用这类方法预测结合位点的精确度是最高的[20].

综上所述, 基于几何的计算方法存在命中率较低的问题, 而基于能量的方法命中率虽然比较高, 但是其计算量很大. 文本采用基于结构比对的方法对结合位点进行预测, 引入同源索引减少计算量, 构建一个模版数据库用于结构比对, 最后表明本文的计算方法在保证计算量适当的情况下能得到较好的计算结果.

1 方法

1.1 数据准备

1.1.1 模版库

对于本身结构中带配体的蛋白质的预测方法比较简单, 而对于结构中不带配体的蛋白质, 需要通过同源模版比对的方式来预测结合位点. 所以, 需要获取模版. 首先, 从蛋白结构数据库(Protein Data Bank)中以一定的过滤规则查询并下载所有符合条件的蛋白质. 其中, 过滤规则设置为: (1)大分子的类型为蛋白质, 不包含 DNA 和 RNA; (2)测定蛋白质的实验方法为 X 射线衍射法; (3)蛋白质本身必须包含配体; (4)序列长度大于 50. 共有 66903 个蛋白质符合所有条件. 然后, 对每个 PDB 文件进行分析, 以链为单位保存, 并只保存包含配体的情况, 以此作为模板库, 其中共包含 113102 个蛋白结构.

1.1.2 测试集

为了便于与其他方法进行比较, 采用了两组基准数据集作为测试集, 第一组包含 210 个带配体蛋白质结构, 在 LIGSITE、LIGSITEcs、LIGSITEcsc[21]、PASS、SURFNET 和 Q-siteFinder 等方法中也采用了这组数据集作为测试集; 另一组使用广泛的数据集是 48 对蛋白质结构, 包含了 48 个不带配体蛋白质和 48 个与之结构相似的带配体蛋白质, 使用不带配体的蛋白进行计算, 而使用带配体的蛋白进行检验.

1.2 流程及说明

整个算法的流程如图 1 所示.

对于任一蛋白质, 都必须给出它的 PDB 文件作为输入, 所以对于结构未知的蛋白质来说, 该方法并不适用. 本文的方法考虑了两种情况, 一种是目标蛋白质结构中本身带有配体; 另一种是不带配体或者对第一种预测方法不适用的情况. 整个流程主要分为六个步骤: 第一步, 判断输入的蛋白结构是否带配体, 如果发现带配体, 那么获取并保存带配体的链, 直接对该链进行位点定义, 否则继续执行第二步; 第二步, 对不带配体的结构, 构建同源索引; 第三步, 如果目

标蛋白有同源蛋白, 那么针对同源蛋白模版进行结构比对, 否则第四步, 与模板库中的结构进行结构比对; 第五步, 进行配体映射和聚类分析; 第六步, 位点定义, 最后输出预测结果.

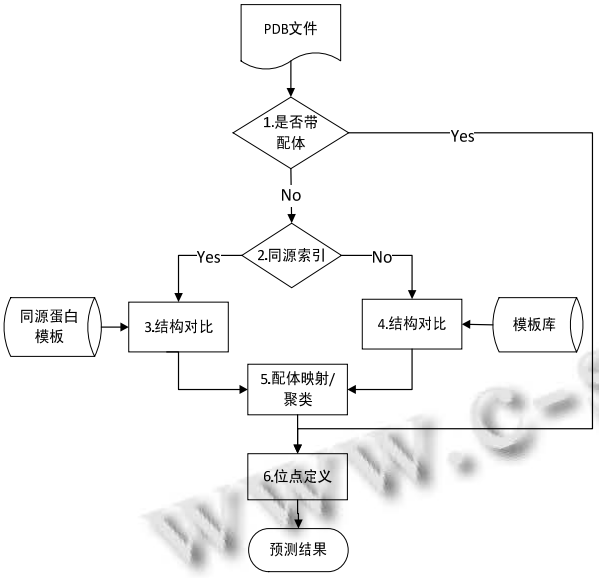

图 1 算法流程图

1.2.1 判断是否带配体

输入一个蛋白质的 PDB 文件后, 首先要做的就是判断蛋白结构中是否带有配体, 从 PDB 文件中分别获取 “HET” 和 “MODRES” 标签下的两个残基组, 从 “HET”标签下的残基组中去除“MODRES”标签下的残基, 得到的就是该蛋白质的配体残基信息. 然后去除蛋白质中的水, 并以链为单位, 保存带配体的蛋白质链和配体信息.

1.2.2 同源索引

为了减少计算量和算法的运行时间, 可以从已有的模版库中筛选出目标蛋白的同源蛋白质作为比对模版. 借助 SCOPe(Structural Classification of Proteins–extended)[22]蛋白质结构分类数据库建立同源索引, 首先从 SCOPe 数据库中识别出目标蛋白质所在的超家族(superFamily), 并获取该超家族中其他所有的蛋白质链, 由于是在同一超家族中的蛋白链, 所以它们与目标蛋白的结构较为相似. 然后判断所获得的这些蛋白质链是否包含在之前所构建的模板库中, 提取所有包含在模板库中的同源蛋白链, 作为同源蛋白模版, 用于之后的结构比对.

1.2.3 结构比对

通过蛋白质两两结构比对的方式, 选择最可能的配体结合位点. 目标蛋白与模版库中的每个同源蛋白模版一一执行基于 CE 算法[23]的结构比对, 至于模板库的选择, 取决于目标蛋白链是否有同源蛋白链, 如果存在同源模版, 则只需要与这些同源模版进行结构比对, 否则需要与之前建立的模板库中的所有模版进行比对, 获得最终的比对结果, 然后去除结果中不理想的情况, 即比对结果中相同残基的数量小于 10 个, 也就是说两者的相似度极低的情况. 最后得到两个比对蛋白质的覆盖度(coverage), 分别记为 coverage1 和 coverage2, 以两者的乘积作为相似度标准, 乘积值越大, 表明相似度就越高, 所以根据所有 $coverage1 \times coverage2$ 值对同源模版进行排序, 选择并保存相似度最高的 30 个蛋白链, 即排列在前 30 的蛋白质结构, 用于之后的配体映射及聚类过程.

1.2.4 配体映射及聚类

接下来, 需要通过旋转、平移的方式, 将相似度最高的 30 个蛋白质链上的配体分别映射到目标蛋白结构上. 然后, 对这 30 个配体进行聚类分析, 计算配体几何中心两两之间的距离, 得到一个 30*30 的距离矩阵. 通常配体数量最多的区域被认为是最有可能的结合区域, 所以基于这个理论, 计算配体数量最多的区域就可以预测结合位点. 借助距离矩阵, 计算每个配体周围一定距离范围内的配体数量, 即如果某个配体 A 的几何中心与其他配体的几何中心的距离小于某一阈值, 这里距离阈值取 2Å, 那么把这些配体归于一个簇中. 最后, 从这些簇中选取配体数量最多的三个簇, 分别计算簇的几何中心, 比较簇中的所有配体与该几何中心的距离, 找出离该几何中心最近的配体, 保存下来, 用于位点定义.

1.2.5 位点定义

位点定义时, 比较计算目标蛋白链上所有残基和配体上的原子, 配体的选择也分为两种情况, 一种是本身带有配体的情况, 那么计算时就选择蛋白质结构中的配体, 另一种不带配体的情况, 那么就用之前通过结构比对后得到的模版上的配体. 当目标蛋白某一残基上的任一原子与配体上的任一原子之间的距离小于 8 Å 时, 该原子所属的残基就被定义为结合位点, 否则为非结合位点. 定义完目标蛋白链上的所有残基后, 计算所有位点形成的活性口袋的几何中心与配体

上任一原子之间的距离, 如果小于 4Å, 则预测成功. 若目标蛋白链本身不带配体, 那么完成三组位点定义之后, 需要根据预测位点残基的数量对三组结合位点进行排序, 选择位点残基数量最多的一组作为最后的预测结果.

1.3 评估

采用命中率(Hit Rate)、准确度(Accuracy, ACC)[24]和马修斯相关性系数(Matthews correlation coefficient, MCC)[25]这三大指标来评估预测结果.

针对 1.2 小节中的两种情况, 计算命中率方法也有不同. 对于第一种情况, 用结构中本身带有的配体来预测位点, 只需要测试预测位点的几何中心和配体的最小距离是否小于 4 Å, 如果满足条件, 那么就可以判断为命中, 否则为没有命中. 而对于本身结构中不带配体链来说, 计算命中方法与前者相似, 唯一不同的是引入了相对应的带配体的结构. 在 48 对蛋白结构测试集中, 把 48 对带配体的结构上的配体作为基准测试配体, 分别映射到不带配体的结构上, 然后使用与第一种情况相同的计算方法来判断是否命中.

ACC 主要衡量测试值与真实值之间相异程度, 而 MCC 除了衡量相异程度外, 当样本的阴性和阳性数据规模差异很大时, 用得更为广泛. ACC 与 MCC 的计算方法如下和所示, 其中, 把蛋白质的结合位点残基作为阳性数据, 非结合位点残基作为阴性数据 .

$$ACC = \frac{TP + TN}{TP + TN + FN + FP} \tag{1}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(FN + FP)(TN + FN)}} \tag{2}$$

2 结果与讨论

2.1 预测结果

下面以 210 测试集中蛋白质 7tim.A 链为例, 给出带配体的蛋白质的预测结果. 如下图 2 所示, 蓝色部分为该蛋白质本身所带的配体, 红色部分就是通过预测得到的结合位点.

对于不带配体蛋白质链的预测, 以 48 对测试集中的 3p2p.A 链为例, 图 3 为蛋白质 3p2p 的 A 链原本的结构, 图 4 是把 5p2p.B 链上的配体映射到 3p2p.A 后的结构, 蓝色部分为 5p2p.B 上的配体, 红色部分为根据该配体预测到的结合位点.

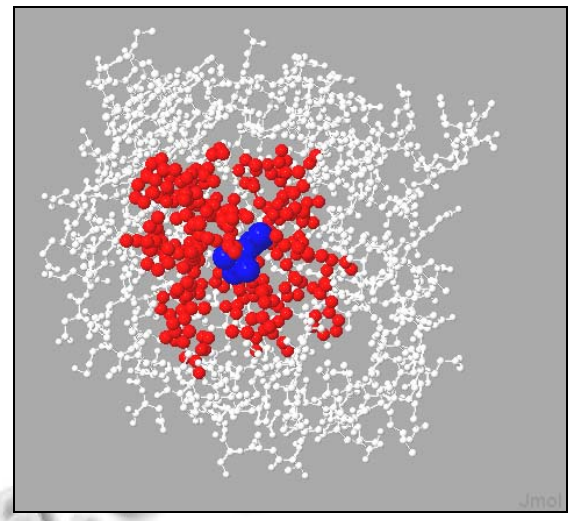

图 2　7tim.A 预测结果

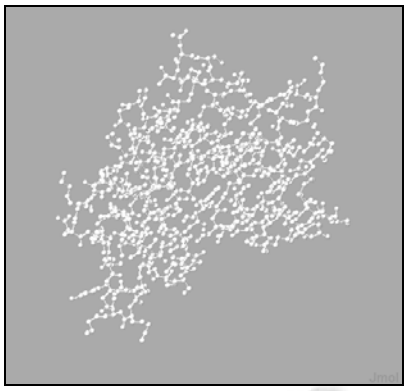

图 3　蛋白质 3p2p 的 A 链原本的结构

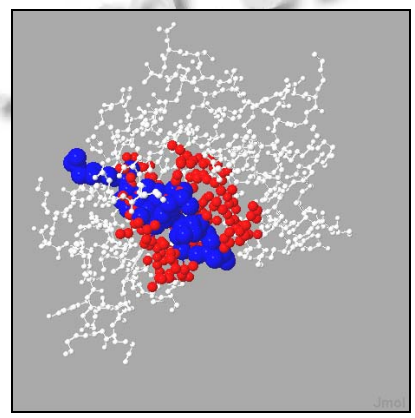

图 4　配体映射后的 3p2p.A 结构和预测结果

下表 1 给出了不同方法采用 210 带配体结构作为测试集的 Top1 和 Top3 命中率, Top1 和 Top3 的命中率均为 90%, 也就是说 210 个蛋白中仅有 21 个没有预测成功. 从表中看到, 基于几何的方法 Dai[26]、PASS、LIGSITE、LIGSITEcs、LIGSITEcsc 的 Top1 的命中率

仅在 42%~75%之间, Top3 的命中率在 57%~87%之间, 不管是 Top1 还是 Top3, 本文方法的命中率都要高于该类方法. 而基于能量的方法 Q–siteFinder 的 Top1 为 70%, 较低, LISE 达到 83%, 考虑 Top3 的命中率, 本方法的 Top3 命中率与 Q–siteFinder 的命中率持平, 而不及 LISE, 但基于能量的方法计算时间较长, 本方法在算法效率方面优势较大. 而 MPK1(MetaPocket1.0)和 MPK2(MetaPocket2.0)方法采用的是同时执行几种不同的类型的方法, 如 MPK2 进行 LIGSITEcs、PASS、Q–SiteFinder 等 8 个算法, 然后选取其中较优的结果, 所以命中率较高, 从 Top1 的命中率来看, 本方法有较大的优势, 而相比 Top3, 本方法稍低, 同样, 由于 MPK 方法需要进行 8 个算法的计算, 所以效率不高, 相对而言, 本方法引入了同源索引, 效率较高. 当然, 针对没有同源模版的蛋白链来说, 计算量比较大, 本方法的优势并不明显. 通过计算, 本方法的平均 ACC 值可达 0.93, MCC 可达 0.80. 基于能量的方法 Q–siteFinder 的 ACC 为 0.68, 而同样基于结构比对的预测算法 TM–align 算法的 ACC 也仅有 0.57, MCC 为 0.48, 3DLigandSite 所得到的 ACC 值为 0.60, MCC 为 0.64. 由此可见, 本文所采用的方法的命中率要比其他方法的命中率高, 预测效果较好.

表 1 210 带配体结构测试集的不同方法命中率(%)比较

方法	Top1	Top3	类型
this work	90	90	比对
Dai	59	75	几何
LIGSITE^{csc}	75	—	几何
LIGSITE^{cs}	67	87	几何
LIGSITE	65	85	几何
PASS	51	80	几何
SURFNET	42	57	几何
Q–siteFinder	70	90	能量
LISE	83	94	能量
MPK1	75	93	混合
MPK2	81	95	混合

对于应用更广泛的 48 对蛋白质结构测试集来说, 它的预测结果更具代表性. 从下表 2 中可以看出, 对于本身带配体的蛋白质链来说, Top1 的命中率为 90%, 比其他预测方法高. 对于结构中不带配体的蛋白链, 本文提出的方法 Top1 命中率为 83%, 也就是说 48 个蛋白质中仅 8 个没有命中, 而基于几何的方法 CAST 的 TOP1 的命中率仅 58%, 最高的命中率 Top1 也只有 71%, 由此可见, 本方法较优, 而本方法的 Top3 命中率为 85%, 相对于其他方法来说, 没有较大提高. 通过计算得出, 平均 ACC 值为 0.93, MCC 值为 0.52. 从得到的三个评估参数值中综合分析得出, 本方法的预测精度较高, 预测效果显著.

表 2 48 对蛋白质结构测试集的不同方法命中率(%)比较

方法	带配体		不带配体		类型
	Top1	Top3	Top1	Top3	
this work	90	90	83	85	比对
CAST	67	83	58	75	几何
PASS	63	81	60	71	几何
PocketPicker	72	85	59	85	几何
FPocket	83	92	69	94	几何
LIGSITE^{cs}	81	92	71	85	几何
LIGSITE^{csc}	79	—	71	—	几何
SURFNET	54	78	54	75	几何

2.2 引入同源索引的影响

下表 3 给出了同源索引引入前后结构比对数量的区别. 建立同源索引之前, 每个不带配体的蛋白结构都需要与模板库中所有的蛋白链(总共 113102 条蛋白链)进行结构比对, 对于 48 个结构, 那就需要进行 113102×48=5428896 次结构比对, 而引入同源索引之后, 48 个结构的比对次数减少到了 27589 次, 只占原来的 0.51%. 由此可见, 建立同源索引可以使得结构比对的次数大大减少, 算法的计算时间被缩短.

表 3 同源索引引入前后结构比对数量的区别

ChainID	引入前	引入后	ChainID	引入前	引入后
3tms.A	113102	323	1cge.A	113102	263
8adh.A	113102	261	1hsi.B	113102	344
1hxf.H	113102	1672	1a4j.B	113102	1342
2fbp.A	113102	136	1ime.A	113102	69
1gcg.A	113102	286	1nna.A	113102	387
1hel.A	113102	483	1ahc.A	113102	209
1npc.A	113102	677	2tga.A	113102	578
1esa.A	113102	1076	4ca2.A	113102	129
1brq.A	113102	322	1pdy.A	113102	547
8rat.A	113102	147	1phc.A	113102	1046
1swb.A	113102	162	1psn.A	113102	160
1ula.A	113102	713	3lck.A	113102	2401

1ifb.A	113102	185	1bbs.A	113102	212
3ptn.A	113102	612	1stn.A	113102	114
1ypi.A	113102	164	1pts.A	113102	160
5dfr.A	113102	574	2ctb.A	113102	485
3phv.A	113102	276	2cba.A	113102	374
2ctv.A	113102	1452	1krn.A	113102	9
5cpa.A	113102	472	2sil.A	113102	239
1a6u.H	113102	749	1l3f.E	113102	528
1qif.A	113102	2718	1chg.A	113102	1196
3app.A	113102	775	6ins.E	113102	23
1djb.A	113102	603	3p2p.A	113102	145
1bya.A	113102	1648	7rat.A	113102	143

3 结语

对蛋白–配体的结合位点的研究具有重要的意义, 虽然已存在很多不同的预测蛋白质结合位点的方法, 但一般基于几何的方法的存在命中率低的问题, 而对于命中率较高的基于的能量的方法来说, 计算量又过大, 影响计算效率, 所以选择采用基于结构比对的方法比较适合. 此外引入同源索引, 大大地减少计算量和计算时间. 本文引入了一种基于结构比对的蛋白结合位点预测方法, 经测试, 本文的方法在计算量适当的情况下, 具有较高的命中率.

参考文献

1 Jones S, Thornton JM. Principles of protein-protein interactions. Proc. of the National Academy of Sciences, 1996, 93(1): 13–20.

2 Xie ZR. Ligand-binding site prediction using ligand-interacting and binding site-enriched protein triangles. Bioinformatics, 2012, 28(12): 1579–1585.

3 Levitt DG, Banaszak LJ. POCKET: a computer graphies method for identifying and displaying protein cavities and their surrounding amino acids. Journal of Molecular Graphics, 1992, 10(4): 229–234.

4 Hendlich M, Rippmann F, Barnickel G. LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. Journal of Molecular Graphics and Modelling, 1997, 15(6): 359–363.

5 Brady Jr GP, Stouten PF. Fast prediction and visualization of protein binding pockets with PASS. Journal of Computer-Aided Molecular Design, 2000, 14(4): 383–401.

6 Laskowski RA. SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. Journal of Molecular Graphics, 1995, 13(5): 323–330.

7 Liang J, Woodward C, Edelsbrunner H. Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. Protein Science, 1998. 7(9): 1884–1897.

8 Weisel M, Proschak E, Schneider G. PocketPicker: analysis of ligand binding-sites with shape descriptors. Chem Cent J, 2007, 1(7): 1–17.

9 Laurie AT, Jackson RM. Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites. Bioinformatics, 2005, 21(9): 1908–1916.

10 Goodford PJ. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. Journal of Medicinal Chemistry, 1985, 28(7): 849–857.

11 Capra JA, Singh M. Predicting functionally important residues from sequence conservation. Bioinformatics, 2007. 23(15): 1875–1882.

12 徐占.蛋白质空间结构的相似性比较[学位论文].无锡:江南大学,2009.

13 Skolnick J, Brylinski M. FINDSITE: a combined evolution/structure-based approach to protein function prediction. Briefings in bioinformatics. 2009. bbp017.

14 Jones D, Hadley C. Threading methods for protein structure prediction. Bioinformatics, Sequence, Structure and Databanks. 2000. 1–13.

15 Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Research, 2005, 33(7): 2302–2309.

16 Ortiz AR, Strauss CE, Olmea O. MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. Protein Science, 2002, 11(11): 2606–2621.

17 Wass MN, Kelley LA, Sternberg MJ. 3DLigandSite: predicting ligand-binding sites using similar structures. Nucleic Acids Research. 2010. gkq406.

18 Roy A, Yang J, Zhang Y. COFACTOR: an accurate comparative algorithm for structure-based protein function annotation. Nucleic Acids Research. 2012. gks372.

19 Yang J, Roy A, Zhang Y. Protein-ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment. Bioinformatics, 2013, 29(20): 2588–2595.

20 Schmidt T, et al. Assessment of ligand-binding residue predictions in CASP9. Proteins: Structure, Function, and Bioinformatics, 2011, 79(S10): 126–136.

21 Huang B, Schroeder M. LIGSITEcsc: predicting ligand binding sites using the Connolly surface and degree of conservation. BMC Structural Biology, 2006, 6(1): 19.

22 Fox NK, Brenner SE, Chandonia JM. SCOPe: structural classification of proteins-extended, integrating SCOP and ASTRAL data and classification of new structures. Nucleic Acids Research, 2014, 42(D1): D304–D309.

23 Shindyalov IN, Bourne PE. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Engineering, 1998, 11(9): 739–747.

24 Metz CE. Basic principles of ROC analysis. Semin Nucl Med, 1978 Oct, 8(4): 283–98.

25 Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochimica et Biophysica Acta (BBA) - Protein Structure . 1975, 405(2): 442–451.

26 Dai T, Liu Q, Gao J, et al. A new protein-ligand binding sites prediction method based on the integration of protein sequence conservation information. BMC Bioinformatics, 2011, 12(Suppl 14): S9.