doi:10.3969/j.issn.1672-5565.2013.01.03

Profile-profile 比对方法用于发现远距离同源模板

邢龙生[1], 李 娟[2*], 方慧生[1*], 陈凯先[3,4★]

(1. 中国药科大学生命科学与技术学院,南京市 210009;2. 南京市鼓楼医院血液科, 南京市 210008;
3. 中国科学院上海药物研究所药物发现与设计中心,上海市 201203;4. 上海中医药大学,上海市 201203)

摘 要:基于模板的模建方法是蛋白质结构预测领域中最为准确有效的方法,该类方法的成功与否对模板质量的要求较高。为待预测序列找寻合适的模板,本文提出了一种 profile-profile 比对的方法将查询序列同模板库中的已知结构蛋白进行比对,然后根据比对结果的 Z-score 得分高低顺序挑选出合适的模板。结果表明:本文的 profile-profile 比对方法在测试集上的性能明显优于 PSI-BLAST,相比 PSI-BLAST 在测试集上的准确度提高了约 14.3%,配对 t 检验的结果表明准确度的提高具有统计显著性。从而得出如下结论:本文的 profile-profile 比对方法可以用于为序列相似性较低的待预测序列搜索远距离同源模板,并用于指导后续的三级结构预测。

关键词:profile-profile 比对,结构预测,远距离同源模板,PSI-BLAST

中图分类号:Q343.1+5 **文献标识码**:A **文章编号**:1672-5565(2013)-01-016-06

Profile-profile alignment method for detection of distantly homologous template

XING Long-sheng[1], LI Juan[2*], FANG Hui-sheng[1*], CHEN Kai-xian[3,4]

(1. *School of life science and technology, China Pharmaceutical University, Nanjing* 210009, *China*; 2. *Hematology Department, Nanjing Drum Tower Hospital, Nanjing* 210008, *China*; 3. *Drug Discovery and Design Center of Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai* 201203, *China*; 4. *Shanghai University of Traditional Chinese Medicine, Shanghai* 201203, *China*)

Abstract: Template-based modeling was the most accurate and efficient method in the field of protein tertiary structure prediction, however, this Kind of method was highly dependent on the template quality. This article was aimed at establishing a kind of alignment method, which was used for detection of suitable templates for query sequence. A kind of profile-profile alignment method was proposed in this article, which firstly make alignments of query sequences and proteins with known structure in template library, followed the templates being selected out according to Z-score ranking of alignment results. The proposed profile-profile alignment method obviously outperformed PSI-BLAST on the testing sets. The accuracy was increased by 14.3% in comparison to PSI-BLAST on testing sets, which was statistically significant on a paired Students' t test. The profile-profile alignment method in this article could be employed to identify distantly-related templates for query sequences with low sequence similarity, furthermore, the templates obtained from this method could be used to guide tertiary structure prediction of query sequence subsequently.

Key words: Profile-profile Alignment, Structure Prediction, Distantly Homologous Template; PSI-BLAST

蛋白质三级结构预测是计算生物学领域的最具挑战性的课题之一。蛋白质结构预测的意义和用途

收稿日期:2012-10-26;**修回日期**:2012-11-02.

作者简介:邢龙生,男,安徽,在读硕士研究生,研究方向:蛋白质结构预测。

*通讯作者:方慧生, 男, 浙江, 汉族, 教授, 研究方向:生命科学,虚拟生命科学,Email:hsfang889@cpu.edu.cn;
李娟,女,云南,白族,副教授,副主任医师,研究方向:虚拟生命科学;

★总设计师:陈凯先,男,上海中医药大学校长,中科院院士,研究方向:计算机辅助药物设计,化学信息学。

文献[1]中已有报道，此处不再赘述。Anfinsen[2] 及其同事等人发现决定一个蛋白质结构的全部信息都包含在它自身的序列中，蛋白质的天然结构对应其构象的自由能最低点，这一热力学假说为计算预测蛋白质结构奠定了理论基础。在过去的几十年里，国内外众多研究人员一直为解决蛋白质的折叠问题开展了大量的研究工作，并取得了不少的成果。针对蛋白质三级结构问题，人们提出了很多种预测方法。自从 1994 年开始，国际上每两年举办一次的 CASP(Critical Assessment of Protein Structure Prediction)会议，就是专门用来对当前的蛋白质结构预测方法的性能进行评估从而了解该领域的研究进展情况，并根据其预测结果对各种预测方法进行排名。CASP 比赛中将蛋白质结构预测方法主要分为两大类：基于模板的结构预测方法(Template - based modeling)，从头预测(Ab initio)方法。尽管在从头预测方法上已经有不少成功的例子出现，其中最为著名的是 David Baker 实验室的 ROSETTA[3] 方法，以及 I - TASSER[4]、TOUCHSTONE[5] 等方法，然而，到目前为止，基于模板的蛋白质结构预测方法仍然是最为准确有效的方法[6]。基于模板的预测方法根据待预测序列同模板结构之间的序列相似性又分为两种：同源模建(Homology modeling)或称比较模建(Comparative modeling)，该方法需要目标序列同模板间的序列相似性不低于 30%，这样才能保证预测结果的较高的准确性；穿线法[7-8](threading)或称折叠识别法(fold recognition)，该方法是在目标序列同模板序列间相似性较低的情况下(低于 30%)从已知结构中找到同目标序列具有相似折叠方式的模板蛋白。

传统方法主要依靠序列间的配对序列比对用于发现序列之间的相似性，并根据相似性的高低来挑选出合适的模板。David Eisenberg[9] 等人最先报道采用序列 profile 分析的方法用于发现远距离相关的蛋白质，他们首先根据多重序列比对构建了目标序列的 profile，然后利用此 profile 搜索序列库从而发现了属于同一家族的蛋白质。最常用的同源性搜索工具 PSI - BLAST[10] 程序采用查询序列的 profile 作为序列比对的输入，从而提高了远距离同源蛋白的识别能力。这些结果表明了基于 profile 的比对方法用于发现远距离同源模板的正确性和可靠性，因此人们提出了很多基于 profile - profile 比对的方法。如 Marc A. Marti - renom[11] 等人比较了几种基于序列 profile 比对方法的性能，他们认为同单纯的序列相比，profile 中包含了更多结构相关的信息，因而可以显著提高折叠识别的性能并改善序列比对结果。目前文献中已有基于多种方法的穿线算法报道：比如，profile - profile 比对方法，结构 profile 比对法，隐马尔科夫模型(HMM)法，机器学习法等。在多次穿线法盲测(blind test)中，普通的 profile - profile 比对方法都表现出了显著性的优势。例如，在近期 CASP 比赛的服务器组中，几种基于序列 profile 的方法都名列单个穿线法服务器之首。本文即将讨论的内容就是：如何为序列相似性较低的目标序列寻找合适的远距离同源模板，从而能够保证结构预测结果较为准确可靠。本文将结合序列 profile 以及预测的二级结构信息用于查询序列同模板结构的 profile - profile 比对，从而发现远距离同源模板以进一步用于查询序列的三级结构预测。

1　方法

1.1　打分函数

本文中采取的打分函数[4] 如下所示：

$$S(i,j)=\sum_{k=1}^{20}(Pc_q(i,k)+Pd_q(i,k))Lt(j,k)/2+c_1\delta(s_q(i),s_t(j))+c_2 \quad (1)$$

上式表示查询序列的第 i 个残基同模板的第 j 个残基的比对得分，其中，“q”表示查询序列，“t”表示模板蛋白。打分函数中每一项的具体含义如下所述：

两种序列 profile

公式(1)中的第一项表示由查询序列得到的两种 profile。利用 PSI - BLAST 为查询序列搜索非冗余序列数据库，根据比对结果构建出多重序列比对，Pcq(i,k)表示该多重序列比对中第 i 个位点处氨基酸 k(k 表示 20 种氨基酸的编号)的出现频率，PSI - BLAST 搜索时的 E 值设为 0.001。这一项称为近距离同源产生的频率 profile [12]。另一项远距离的频率 profile Pdq(i,k)用相同的方法产生，只是 E 值有所不同为 1.0。此处之所以联合使用近距离和远距离的序列 profile，是由于 Skolnick[12] 等人认为这样可以在不同同源性区域提高比对的敏感性。在计算这两项频率 profile 时，我们采用 Henikoff 和 Henikoff[13] 提出的基于位点的序列权重法以降低比对序列的冗余性。此外，我们参照文献[12]中的方法：为了强调显著性高的 PSI - BLAST 命中序列，E 值较小的序列赋予较高权重，E 值较大的序列赋予较低权重。分配权重的具体方案为：E 值小于 10 - 10 的序列赋权重 1.0，其余序列权重随相应 E 值的对数值线性下降直至 E 值为 1.0 的序列赋权重 0.5。公式(1)中的 Lt(j,k)表示模板序列的第 j 个位点处氨基酸 k 的对数几率(log - odds)profile 值。模板的

对数几率值 profile 是由 E 值为 0.001 的 PSI - BLAST 搜索非冗余序列库得到的。

公式(1)中的第二项将查询序列第 i 个残基的预测二级结构 sq(i)同模板的第 j 个残基的真实二级结构 st(j)进行比较。如果 sq(i)和 st(j)的二级结构类型相同,δ(sq(i), st(j))值等于 1,否则等于 0。二级结构类型分为螺旋,片层和无规则卷曲三种,查询序列的二级结构是由 PSIPRED[14] 软件预测得到,模板的二级结构由 STRIDE 程序包计算得到。

公式(1)中的第三项 c_2 为漂移常数,引入这一项是用来避免局部区域内的不相关残基的比对。

1.2 动态规划算法

我们采用全局动态规划算法[15] 用于发现查询序列同模板序列间的最佳比对。采用了位点依赖型的空位罚分方法,即模板的 α 螺旋以及 β 片层二级结构区域内不允许出现空位,其他区域采用空位起始罚分 g_o 以及空位延伸罚分 g_e,忽略末端空位罚分。

1.3 模板排名方案

文献中[12,16] 多有报道按照 Z - score 值由大到小的顺序对搜索结果进行排序,从而挑选出可能的最佳模板以及同目标序列的比对结果。本方法亦采用同样的方法将初始比对得分转化为 Z - score 值,然后按照 Z - score 值大小顺序挑选合适的模板。其中,Z - score 值的计算公式如下:

$$Z-score=\frac{(R'_{score}-\langle R'_{score}\rangle)}{\sqrt{\langle R^{2\prime}{}_{score}\rangle-\langle R'_{score}\rangle^2}} \quad (2)$$

其中,Rscore 表示初始比对得分,R'_{score} 表示归一化以后的比对得分 R_{score}/L_{full} 或者 $R_{score}/L_{partial}$,L_{full} 表示比对的全长(包括查询序列和模板的末端空位),而 $L_{partial}$ 表示比对的部分长度(不包括查询序列的末端空位),如图 1 所示,; < … > 表示所有模板的平均值。采用 R_{score}/L_{full} 还是 $R_{score}/L_{partial}$ 对初始比对得分进行归一化由两种方法挑选出的第一个模板同查询序列的序列相似性高低所决定,R_{score}/L_{full} 得到的第一个模板同查询序列相似度较高就选择 R_{score}/L_{full} 作为归一化方法,否则选择部分比对长度进行比对得分的归一化。

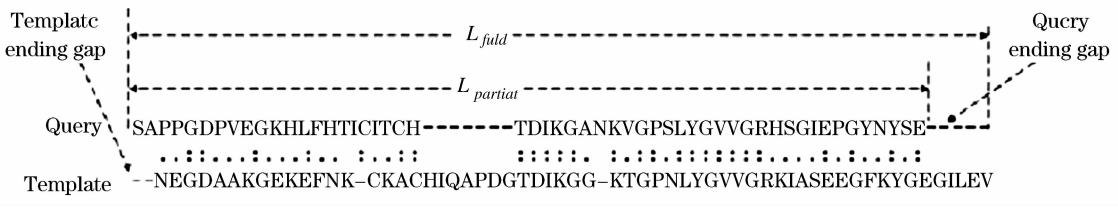

图 1 用于归一化比对初始得分的比对全长和比对部分长度的示意图[12]

符号“-”,“.”和“:”分别表示未比对的空位,比对的不相同残基对以及比对的相同残基对

Fig. 1 Schematic of alignment full length and partial length used for normalization of alignment raw scores

Symbol “-”, “.” and “:” denote unaligned gaps, aligned nonidentical residue pairs and aligned identical residue pairs, respectively

1.4 参数训练

在我们的 profile - profile 比对算法中一共有 4 个参数需要进行合理的调试,它们分别是 c_1、c_2、g_o 以及 g_e。文献[12]报道的参数训练方法很多是基于 PROSUP 结构比对数据库的,该数据库包含了由结构比对程序 PROSUP 产生的 127 对非同源蛋白质的最佳比对结果。可以通过调节程序中的参数以使得比对的残基对同数据库 PROSUP 中一致的数量最大化,从而获得较优的参数组合。本论文中我们选取了来自 PROSUP 数据库的 50 个最佳比对作为训练集。采用文献中报道的网格搜索的方法来进行参数的训练,具体的做法就是让训练集中的每个蛋白遍历所有的网格点,然后统计得到同最佳比对中相同的残基对数目最多的那组参数,即为我们需要的最佳参数。

本论文中采用的蛋白质结构模板全部来自于 Yang Zhang[12] 等人报道中所采用的蛋白质结构模板库。

2 结果与讨论

2.1 训练集蛋白

为了对本文的 profile - profile 比对算法的参数进行训练,我们从 PROSUP 结构比对数据库中选取了 50 对非同源蛋白质的结构比对作为本方法的训练集,表 1 中列出了训练集中所有的蛋白质对。

表 1　训练集中包含的结构比对的蛋白质 PDB 号

Table 1　PDB codes of proteins for structural alignment contained in training sets

蛋白质对		蛋白质对		蛋白质对		蛋白质对		蛋白质对	
1cnv_	1nar_	1hce_	4fgf_	1aizB	1rcy_	1ecmB	1csmB	1sacA	1y4wA1
1nulA	1hgxA	2cpgA	1bazA	1dcfA	1fyeA	1ujpA	1i1wA	1t4hA	1nw1A
1acf_	1pne_	1vhwA	1vheA2	1elg_	2alp_	1h80A	1k5cA	1snyA	1vjpA1
1gtqA	1gtpA	1pgs_	1phm_	1bbzA	1jb0E	1dlwA	1h97A	1fjjA	1qouA
1h05A	1ydgA	1senA	1a81_1	1pjaA	1lzlA	2pth_	1vheA2	2bltB	3pte_
1ajsA	1ohwA	1oeyA	1c9fA	1ttzA	1t1vA	1wfzA	1mzgA	1uwdA	1josA
1n1jB	1tzyB	1wf6A	1cdzA	1den_	1tcp_	2gmfB	1rcb_	1m5wA	1ur4A
1keaA	1ornA	1afi_	1aps_	1rhcA	1eokA	1vhnA	1qnrA	1elg_	1havA
1vk4A	1vi9A	1dlwA	1it2A	1hce_	1i1b_	1ptvA	1ytn_	1agjA	1elg_
1s1qA	1jatA	1dhkA	1bag_	1cvl_	1mnaA	1rliA	1g2iA	1m8zA	1jdhA

训练集的结果

运用前面提到的网格搜索法对本方法中所包含的四个待训练参数进行训练,最终得到的最优参数组合为:c_1,c_2,g_o,g_e 的值分别为 0.8,1.8,-7.1,-0.5。表 2 中列出了使用最优参数组合将本方法应用于训练集蛋白质所产生的结果。由表 2 可知,训练集所得比对结果中的比对残基对与结构比对数据库中的残基对相同的数目同结构比对数据库中包含的比对的残基对总数的比值为 27.93%。很明显,本文所采用的 profile-profile 比对方法在训练集上的结果并不出色,准确度相对偏低,分析其中原因可能主要有如下几方面:首先,训练集中的结构比对都是非同源蛋白质间的结构比对,参与比对的蛋白质结构彼此间的同源性都较低;另外,参数范围的选取带有一定的经验性质,参考了相关文献中的取值,可能最终得到的参数组合实际上并非真正的最优参数集。

表 2　profile-profile 比对方法在训练集蛋白上的性能

Table 2　Performance of profile-profile alignment method on training sets

数据	方法	准确度
训练集	profile-profile 比对	27.93%

准确度表示训练集的比对结果中同结构比对库中相同的残基对的数目

2.2　测试集的结果

为了验证本方法的性能,我们选取了 CASP9 比赛中 human 组的 62 个目标蛋白作为本方法的测试集蛋白。表 3 中列出了选取的 CASP9 的目标蛋白编号以及每个目标蛋白利用本方法搜索到的模板个数,其中用红色标注出的数字 0 表示对于这些目标蛋白,本方法未能找到合适的模板(即 CASP 比赛官方公布的最优模板)。从表 3 中我们可以统计得到,采用我们的方法测试集中一共有 33 个目标蛋白可以找到 CASP 官方认可的合适模板,其余 29 个测试集蛋白未能找到合适的模板。根据测试集的结果,本方法在测试集上的准确率大约为 53.23%。

表 3　profile-profile 比对方法在测试集蛋白上的性能

Table 3　Performance of profile-profile alignment method on testing sets

编号	模板数	编号	模板数	编号	模板数
T0515	8	T0564	0	T0604	0
T0516	4	T0568	0	T0605	0
T0517	4	T0569	1	T0606	0
T0518	5	T0571	0	T0608	5
T0520	4	T0574	1	T0610	1
T0521	1	T0576	0	T0612	0
T0523	3	T0578	0	T0614	0
T0526	2	T0579	0	T0616	0
T0529	0	T0580	1	T0618	0
T0531	0	T0581	0	T0619	2
T0534	0	T0582	4	T0621	0
T0537	0	T0584	4	T0622	1
T0540	0	T0586	4	T0624	0
T0543	3	T0588	1	T0625	1
T0544	0	T0590	1	T0627	5
T0547	8	T0592	1	T0628	0
T0550	0	T0594	4	T0629	1
T0553	0	T0596	6	T0630	2
T0558	1	T0598	1	T0637	0
T0561	0	T0600	6	T0643	0
T0562	1	T0602	0		

表格中的数字表示 profile-profile 比对方法搜索到的前 25 个模板同 CASP9 比赛以后官方公布的最优 25 个模板相同的个数。

另外,运用 PSI-BLAST 为同样的测试集蛋白搜索模板,执行 PSI-BLAST 搜索时 E 值采用的是 1.0,同样将搜索到的模板同 CASP9 官方公布的模板相比较。然后将本论文方法在测试集上的结果同 PSI-BLAST 的结果进行了初步的比较,以评价本论文的方法和 PSI-BLAST 在搜索模板性能上的优劣。表 4 中列出了同样的测试集蛋白利用 PSI-BLAST 搜索模板所得的结果。

表 4 PSI – BLAST 方法在测试集蛋白上的性能

Table 4 Performance of PSI – BLAST method on testing sets

编号	模板数	编号	模板数	编号	模板数
T0515	1	T0564	0	T0604	0
T0516	8	T0568	0	T0605	0
T0517	2	T0569	0	T0606	0
T0518	3	T0571	0	T0608	4
T0520	4	T0574	0	T0610	1
T0521	7	T0576	0	T0612	0
T0523	0	T0578	0	T0614	0
T0526	2	T0579	0	T0616	0
T0529	0	T0580	0	T0618	0
T0531	0	T0581	0	T0619	1
T0534	0	T0582	0	T0621	0
T0537	0	T0584	3	T0622	0
T0540	0	T0586	3	T0624	0
T0543	5	T0588	5	T0625	1
T0544	0	T0590	1	T0627	5
T0547	1	T0592	0	T0628	0
T0550	0	T0594	2	T0629	2
T0553	0	T0596	2	T0630	0
T0558	0	T0598	2	T0637	0
T0561	0	T0600	1	T0643	0
T0562	0	T0602	0		

表格中的数字表示 PSI – BLAST 方法在 E 值为 1.0 时搜索到的模板同 CASP9 赛后官方公布的最优 25 个模板相同的个数

对比表 3 和表 4，我们可以清晰地发现本论文的方法和 PSI – BLAST 在测试集性能上的差异，PSI – BLAST 可以搜索到合适模板的目标蛋白，本文的 profile – profile 比对方法也都能为其搜索到 CASP 官方认可的模板。除此之外，本文的 profile – profile 比对方法还能为额外的 9 个目标蛋白搜索到合适的模板，它们分别是：T0523、T0558、T0562、T0569、T0574、T0580、T0592、T0622、T0630，而 PSI – BLAST 方法未能搜索到合适的模板。在所选取的测试集上，本文的 profile – profile 比对方法相比于 PSI – BLAST 在准确度上提高了约 14.3%。

为了验证本文的 profile – profile 比对方法同 PSI – BLAST 在测试集性能上的差异有无统计学显著性，我们对 profile – profile 方法和 PSI – BLAST 的结果进行了配对 t 检验。在进行配对 t 检验时，凡是能够找到模板的目标蛋白统一记为 1，未能找到模板的目标蛋白记为 0，配对 t 检验的结果见表 5。

表 5 profile – profile 方法和 PSI – BLAST 在测试集上结果的配对 t 检验

Table 5 Paired student's t test of the results of profile – profile method and PSI – BLAST on testing sets

	自由度	t Stat	t 单尾临界	P(T <= t) 单尾	t 双尾临界	P(T <= t) 双尾
profile – profile / PSI – BLAST	61	3.425021	1.670219	0.000552	1.999624	0.001105

显著性水平 α 为 0.05

由表 5 可知，t 统计值为 3.425，大于 t 双尾临界 1.999，因此本论文的 profile – profile 比对方法和 PSI – BLAST 在测试集性能上的差异显著，具有统计学上的意义。另外，需要指明的是，本论文的 profile – profile 搜索到的模板精确到模板蛋白的单个亚基，然而 PSI – BLAST 搜索模板时只到模板蛋白的 PDB 编号。

分析本论文中的 profile – profile 比对方法在测试集上搜索模板的性能优于 PSI – BLAST 的原因，主要包括如下几方面：首先，序列 profile 包含了比蛋白质序列更多的有关结构方面的信息；另外，本方法中引入了二级结构匹配得分，文献[12]中报道打分函数中引入更多有关结构方面的信息，可在一定程度上提高折叠识别的性能；其次，本文中的 profile – profile 比对方法的查询序列和模板的 profile 就是根据 PSI – BLAST 的比对结果构建的，所以性能上优于 PSI – BLAST 也在预料之中。

2.3 结论与展望

本文提出了一种用于搜索远距离同源模板的 profile – profile 比对方法，且其在测试集上的性能明显优于 PSI – BLAST 方法，并具有显著的统计学意义。虽然本文中的 profile – profile 比对方法在测试集上的性能优于 PSI – BLAST，但是从结果部分列出的表格来看，仍有相当一部分的目标蛋白采用我们的方法未能搜索到合适的模板，这说明了本文的方法在远距离同源模板的发现上仍有不足之处，尚有待进一步的改进和完善。不过就文献中报道的用于搜索远距离同源模板的方法来看，目前尚没有一种方法能够保证为每个待测序列都找到合适的模板，因此大家一致认为采用一致性的方法可以取得更好的结果，较为出名的一致性方法有 Pcons[17]，LOMETS[18]等等。

参考文献(References)

[1] Yang Zhang. Protein structure prediction: when is it useful? [J]. Current Opinion in Structural Biology, 2009, 19:145 – 155.

[2] Christian B. Anfinsen. Principles that Govern the Folding of Protein Chains[J]. Science, 1973, 181(4096): 223 – 230.

[3] Kim T. Simons, Charles Kooperberg, Enoch Huang, David Baker. Assembly of Protein Tertiary Structures from Fragments with Similar Local Sequences using Simulated Annealing and Bayesian Scoring Funcitons[J]. Journal of Molecular Biology, 1997, 268:209 – 225.

[4] Sitao Wu, Jeffery Skolnick, Yang Zhang. Ab initio modeling of

small proteins by iterative TASSER simulations[J]. BMC Biology, 2007, 5 – 17.

[5] Yang Zhang, Andrzej Kolinski, Jeffery Skolnick. TOUCHSTONE II: A New Approach to Ab Initio Protein Structure Prediction[J]. Biophysical Journal, 2003, 85:1145 – 1164.

[6] Yang Zhang. Progress and challenges in protein structure prediction [J]. Current Opinion in Structural Biology, 2008, 18:342 – 348.

[7] D. T. Jones, W. R. Taylor, J. M. Thornton. A new approach to protein fold recognition[J]. Nature, 1992,358:86 – 89.

[8] James U. Bowie, Roland Luthy, David Eisenberg. A Method to Identify Protein Sequences that Fold into a Known Three – Dimensional Structure[J]. Science, 1991, 253(5016):164 – 170.

[9] Michael Gribskow, Andrew D. Mclachlan, David Eisenberg. Profile analysis: Detection of distantly related proteins [J]. Proc. Natl. Acad. Sci. USA, 1987, 84:4355 – 4358.

[10] Stephen F. Altschul, Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, David J. Lipman. Gapped BLAST and PSI – BLAST: a new generation of protein database search programs[J]. Nucleic Acids Research, 1997, 25(17):3389 – 3402.

[11] Marc A. Marti – Renom, M. S. Madhusudhan, Andrej Sali. Alignment of protein sequences by their profiles[J]. Protein Science, 2004, 13:1071 – 1087.

[12] Sitao Wu, Yang Zhang. MUSTER: Improving protein sequence profile – profile alignments by using multiple sources of structure information[J]. Proteins, 2008, 72:547 – 556.

[13] Steven Henikoff, Jorja G. Henikoff. Position – based sequence weights[J]. Journal of Molecular Biology, 1994, 243:574 – 578.

[14] Liam J. McGuffin, Kevin Bryson, David T. Jones. The PSIPRED protein structure prediction server[J]. Bioinformatics, 2000, 16(4):404 – 405.

[15] Saul B. Needleman, Christian D. Wunsch. A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins[J]. Journal of Molecular Biology, 1970, 48:443 – 453.

[16] Krzysztof Ginalski, Jakub Pas, Lucjan S. Wyrwicz, Marcin von Grotthuss, Janusz M. Bujnicki, Leszek Rychlewski. ORFeus: detection of distant homology using sequence profiles and predicted secondary structure[J]. Nucleic Acids Research, 2003, 31(13):3804 – 3807.

[17] Jesper Lundstrm, Leszek Rychlewski, Janusz Bujnicki, Arne Elofsson. Pcons: A neural – network – based consensus predictor that improves fold recognition[J]. Protein Science, 2001, 10(11):2354 – 2362.

[18] Sitao Wu, Yang Zhang. LOMETS: A local meta – threading – server for protein structure prediction [J]. Nucleic Acids Research, 2007, 35(10):3375 – 3382.