蛋白质功能位点预测

李伍举　吴加金

(军事医学科学院基础医学研究所，北京　100850)

提　要

在 IBM-PC 机上开发了蛋白质功能位点预测软件：PROSITE。根据 EMBL 发布于激光光盘上的蛋白质功能位点氨基酸片段的保守模式数据库，对给定的蛋白质序列，可按 19 类 443 个氨基酸保守模式来探测蛋白质的所属家族，各种功能区的位置和活性部位等性质，通过 52 个序列的验证结果和 SWISS 蛋白质数据库相一致。此外该软件还具有操作灵活，多种输入输出方式等特点。

关键词　蛋白质，功能位点预测，计算机软件。

众所周知，蛋白质分子具有某些功能位点、活性部位或功能结构域。如：N糖基化位点，cAMP 和 cGMP 有关的蛋白激酶磷酸化位点；碱性磷酸脂酶的活性部位，EPSP 合酶活性部位，胰蛋白酶族丝氨酸蛋白酶的活性部位；核酸靶序列，细胞吸附序列，ATP/GTP 结合位点等等。对这些位点的探测及定位是蛋白质研究的重要方面。现已积累了许多关于蛋白质功能位点的性质，这些功能位点活性部位的氨基酸构成资料。尤其是对蛋白质的研究进入分子水平以来，发现蛋白质功能位点氨基酸构成具有相当保守性。当把蛋白质的氨基酸序列建立成蛋白质序列数据库，再用电子计算机分析研究蛋白质各种功能位点，就会发现功能位点附近的氨基酸片段有的是对所有蛋白质都保守，有的是在蛋白质某家族内保守。于是许多研究者对蛋白质功能位点的保守的氨基酸序列作统计分析，建立了蛋白质功能位点的氨基酸保守模式。目前已有许多文章报道了各种功能位点的氨基酸保守模式[1-3]。Bairoch 把发表在文献的功能位点的氨基酸保守模式收集、整理并用计算机程序对蛋白质数据库作检索验证，建立了各种蛋白质功能位点的氨基酸保守模式的数据库[4]。这数据库发表在 EMBL (欧洲分子生物学实验室的简称)所发布的激光光盘上。

在研究分析了蛋白质功能位点的氨基酸保守模式数据库的基础上，我们提出这样设想：即然蛋白质功能位点的氨基酸保守模式是来自实验研究结果，也是研究蛋白质所描准的目标。那么能否利用经实验确定，用统计分析方法所建立的蛋白质功能位点的氨基酸保守模式的数据库，通过电子计算机的软件技术，去预测蛋白质氨基酸序列中各种功能位点的位置，达到预测蛋白质的功能呢?

基于上述设想，我们设计了蛋白质功能位点预测软件——PROSITE。下面介绍 PROSITE 软件的设计思想和使用情况。

在激光光盘上关于蛋白质的功能位点保守氨基酸片段模式由二个主要文件组成。一个命名为 PROSITE. DAT，另一个命名为 PROSITE. DOC。在 PROSITE. DAT 的文件中共收集 443 种功能位点氨基酸片段模式，其中每一个氨基酸片段模式的格式如表 1 所示。

表 1 中每行前两个字符代表该行的功用，

收稿日期：1991-11-05　修回日期：1992-03-02

表 1 PROSITE. DAT 文件中一种模式的数据格式

ID	ASN GLYCOSYLATION; PATTERN. (天冬酰胺糖基化位点;模式)
AC	PS00001. (表示天冬酰胺糖基化位点在文件 PROSITE. DAT 的代号)
DT	APR-1990 (CREATED); APR-1990 (DATA UPDATE); APR-1990 (INFO UPDATE). (该模式生成时间)
DE	N-glycosylation site. (表示位点的名称为 N-糖基化位点)
PA	N-{P}-[ST]-{P}. (表示位点的组成情况)
CC	/TAXO-RANGE = ??E?V; (表示该位点的分类范围是与真核生物和真核病毒类的蛋白质有关)
CC	/SITE = 1, carbohydrate; (表示该位点的有意义残基是位于该模式的第一元素位置,其类型为碳水化合物)
DO	PDOC00001. (表示该位点在文件 PROSITE. DOC 中的代号)

例如 ID 表示该行的内容“ASN GLYCOSYLATION”为天冬酰胺糖基化的功能位点的名称，ID 表示位点的识别行．又如 PA 行表示该行为天冬酰胺糖基化位点的氨基酸模式，N-{P}-[ST]-{P}．该模式中的字母 N,P,S,T 为单字母的氨基酸的代号．例如 N 代表天冬酰胺．排列次序代表该位点的保守多肽中氨基酸的位置，“-”为氨基酸之间分隔符，{}，[]表示该位氨基酸占有情况，通过对整个库的分析，氨基酸占有情况分为：

X：表示在该位置上可以是任一种残基．

N：表示在该位置上是一个确定的残基 N．

{P}：表示在该位置上的残基是除去 P 以外的所有残基．

[ST]：表示在该位置上的残基要么是 S，要么是 T．

[ST](2)：表示有两个位置氨基酸占有情况相同，都是 S 或 T．

X(2, 4)：表示有 2，3 或 4 个位置可以是任一种氨基酸．

整个 PROSITE. DAT 文件中的每个功能位点都是由上面的几种或全部符号组成．

表 1 中的 DO 行所给的 PDOC00001 表示该位点的更详细解释资料可从 PROSITE. DOC 文件中编号为 00001 的条款中查出．

通过对 PROSITE. DAT 文件中各功能位点的氨基酸片段模式的分析，产生了如下的软件设计思想：

首先扫描 PROSITE. DAT 文件，从每个功能位点的数据条款中抽出 3 行：DE，PA 和 DO，其中 DE 和 DO 为输出时提供功能位点名称和代号，以便必要时可通过 DO 行的代号进一步查询该功能位点的资料．PA 行中功能位点的氨基酸模式为检索模式．由于各功能位点的氨基酸排列模式很复杂，对蛋白质的氨基酸序列作功能位点查找时，采用逐点匹配的方法，具体讲：

根据模式中每个位置的残基占有情况，分别与序列中的残基作比较，若配上，则考虑模式中的下一个位置．否则从模式中的第一个位置重新开始，继续查找．现以糖基化位点模式 PAT = N-{P}-[ST]-{P} 为例说明逐点匹配算法，对给定的序列，如 SEQ = NNSPN 位点查找按如下步骤进行：PAT(1) = N，SEQ(1) = N，两者一致，判断模式下个位置，PAT(2) = {P}，SEQ(2) = N，SEQ 的第二个位置为 N，非 P 故两者匹配，PAT(3)=[ST] 与 SEQ(3) = S 亦满足条件，可是 PAT(4)={P} 与 SEQ(4) = P 没有匹配，从 SEQ(2) 和 PAT(1) 继续按上述步骤运行．

为使软件运行时提供用户良好的界面条件，软件在会话菜单中尽可能给出有关提示，如程序运行时首先显示如表 2 的 19 类功能位点代号和名称，便于用户选择预测的类别．

按照上面的设计思想，采用 Fortran 语言设计了蛋白质功能位点预测软件，取名为 PROSITE．

为了验证软件 PROSITE 的预测可靠性，从 EMBL 的激光盘的蛋白质数据库 (SWISS) 中取出 52 个含有功能位点信息的序列，经

表 2　屏幕显示的 19 类蛋白质功能位点的列表

1——Post-translational modification (翻译后修饰位点)	11——Electron transport proteins (电子传递蛋白)
2——Domains (功能域)	12——Other transport protein (其它传递蛋白)
3——DNA or RNA associated proteins (DNA 或 RNA 结合蛋白)	13——Structural proteins (结构蛋白)
4——Oxidoreductases (氧化-还原酶)	14——Receptors (受体)
5——Transferases (转移酶)	15——Cytokines and growth factors (细胞因子和生长因子)
6——Hydrolases (水解酶)	16——Hormones and active peptides (激素和活性肽)
7——Lyases (裂解酶)	17——Toxins (毒素)
8——Isomerases (同工酶)	18——Inhibitors (抑制因子)
9——Ligases (连接酶)	19——Other Signature (其它信号)
10——Other enzymes (其它酶)	0——Quit (退出)

PROSITE 软件的检测后，其中 48 个序列和 SWISS 标注的功能位点完全一致。仅少数序列没有检出功能位点。

经对原数据库进一步用人工查证，这些序列确实不存在功能位点，原因为数据库仅收入序列中的部分氨基酸，不是完整的序列。

通过验证说明，PROSITE 能可靠地检测蛋白质序列的功能位点。目前此软件已提供给有关实验组使用，力求从实验结果返回关于功能位点预测的信息，以便进一步改进提高。

参 考 文 献

1 Bause E. Structural requirements of N-glycosylation of proteins. *Biochem J*, 1983; **209**: 331

2 Glass D B, El-Maghrabi M R, Pilkis S J. Synthetic peptides corresponding to the site phosphorylated in 6-phosphofructo-2-kinase/Fructose-2, 6-bisphosphatase as substrates of cyclic nucleotide-dependent protein kinases. *J Biol Chem*, 1986; **261**: 2987

3 Woodget J R, Gould K L, Hunter T. Substrate specificity of protein kinase C. *Eur J Biochem*, 1986; **161**: 177

4 Bairoch A. PROSITE: a dictionary of sites and patterns in protein. *Nucl Acids Res*, 1991; **19**: 2241

生物科学类核心期刊表

序号	刊　名	序号	刊　名	序号	刊　名
1.	生物化学杂志	13.	植物学报	25.	中华微生物学和免疫学杂志
2.	生物化学与生物物理学报	14.	病毒学报	26.	细胞生物学杂志
3.	生物化学与生物物理进展	15.	上海免疫学杂志	27.	植物生理学通讯
4.	微生物学报	16.	生理科学进展	28.	药学学报
5.	微生物学通报	17.	遗传	29.	水生生物学报
6.	中国科学 · B 辑	18.	微体古生物学报	30.	昆虫学报
7.	生物物理学报	19.	生态学报	31.	动物学报
8.	科学通报	20.	生理学报	32.	人类学学报
9.	古生物学报	21.	中国免疫学杂志	33.	动物分类学报
10.	古脊椎动物学报	22.	实验生物学报	34.	植物生理学报
11.	生物工程学报	23.	生态学杂志		
12.	遗传学报	24.	生物学通报		

摘引自[北京高校图书馆期刊工作研究会，北京大学图书馆，中文核心期刊要目总览。北京：北京大学出版社，1992：166]