利用一种新的网上工具 Scansite 进行蛋白质磷酸化预测

唐 颖 张令强 贺福初*

(军事医学科学院放射医学研究所,北京 100850)

摘 要 Scansite 分析软件是近两年建立的一种新的利用因特网,基于蛋白质分子中较短的模序进行蛋白质磷酸化和蛋白质-蛋白质相互作用预测的工具。这里综述了 Scansite 的使用方法、功能介绍及与其他磷酸化分析软件的比较,并展望了 Scansite 在进行磷酸化预测中面临的问题和应用前景。

关键词 Scansite, 磷酸化, 模序

中图分类号 Q51 **文献标识码** A **文章编号** 1000-3061(2004)04-0623-04

人类基因组草图已经绘制完成,生命科学全面进入以蛋白质功能注释为主要任务的功能基因组学时代[1]。基因组信息的急速增加势必要求发展新的生物信息学技术来预测蛋白质功能和蛋白质-蛋白质相互作用,为实验体系的验证提供可靠的研究线索。核酸和蛋白质的序列分析已经成为当前生命科学工作者的一个不可或缺的基本技能。利用生物信息学手段确定蛋白质序列所包含的结构域(domain)种类与数量已经具有相当的准确度。根据德国海德堡生物信息中心发布的信息,目前已经鉴定并命名的蛋白质分子中的结构域共有 680 种,其中信号传导相关的 165 种,核内分布的 176 种,胞外分布的 225 种,其他类别 114 种(http://smart.embl-heidelberg.de)。但是,对蛋白质序列中较短的模序(motif)进行预测和功能线索提示较少,Pfam 软件也只提供了 18 种模序,而这些模序对于蛋白质分子之间的相互作用尤其是蛋白质的磷酸化至关重要。Michael B. Yaffe 等建立了一种新的分析软件,基于蛋白质中的较短模序进行搜索[2,3],对解决如上问题提供了一条便捷途径,本文拟对这一网上工具进行介绍。

1 蛋白质模序对蛋白质磷酸化和蛋白质-蛋白质作用的重要性

从蛋白质结构的角度,可以把蛋白质序列分为两类基本组分:一种是结构域,如催化结构域、DNA 结合结构域、蛋白质-蛋白质相互作用结构域等,结构域的基本特征是具有较为完整的二级结构,自身具有功能;另一种是模序,其长度比结构域短,一般在 10 个氨基酸左右(而结构域通常为 60~300 个氨基酸),其特点是可以结合特定的结构域或经修饰后调节蛋白质分子的定位和/或活性[2]。模序按功能可分为信号组件结构域识别模序、激酶识别模序(或称激酶底物模序)、特定蛋白或磷脂识别模序等。参与信号级联传递的蛋白质一般含有一个或多个这样的基本组分。例如,许多受体酪氨酸激酶(receptor tyrosine kinase, RTK)可以发生自身酪氨酸磷酸化,含有磷酸酪氨酸 pY 的模序可以特异地与 SH2 或 PTB 结构域结合,后者经常发现于信号传导的接头蛋白(adaptor protein)中,这种接头蛋白或自身被磷酸化而激活,或将其他的蛋白分子募集到 RTK 附近,把信号传递下去。这种基于模序的结合可以保证信号传递的高保真性和级联放大。当前许多软件可以精确地分析蛋白质含有的结构域,如 Pfam 和 SMART 等,但基于模序这种重要的"蛋白质识别密码"对氨基酸序列进行分析和功能预测的软件还很少。Michael B Yaffe 等建立的 Scansite 新算法既可对单个蛋白的序列进行模序分析,也可根据模序搜索蛋白数据库以寻找含有此模序的所有蛋白。

2 基于模序的蛋白质磷酸化位点和相互作用位点分析软件 Scansite

2.1 Scansite 概述

Scansite 包括 3 种功能:模序搜索(motif scan)、数据库搜索(database scan)和序列匹配(sequence match)[3]。其中,模序搜索是其主要功能。此程序建立在如下的基本原则上:所有的丝/苏氨酸都认为是丝/苏氨酸激酶的可能底物位点,所有的酪氨酸都认为是 SH2 或 PTB 结构域的可能结合位点。每个位点根据其前后各 7 个氨基酸组成的长度为 15 个氨基酸的模序与算法设定的最佳结合模序所匹配的程度打分,计算最终所得的值,以两个参数作为主要衡量标准:分值(score)和百分率(percentile)[2]。后者还是界定搜索严谨度的参数,

收稿日期:2003-11-18,修回日期:2004-04-19。

基金项目:国家高技术研究发展计划(863 计划)课题(No.2001AA221111)。

* 通讯作者。Tel:86-10-66931246;Fax:86-10-68214653;E-mail:hefc@nic.bmi.ac.cn

当其小于0.2%时,为高严谨度搜索;当其介于1%～5%之间时,为中严谨度搜索;当其大于5%时,为低严谨度搜索。搜索结果的百分率越低,表明结果的可靠性越高[2]。

当前的Scansite软件为2.0版本,包含62种模序(图1),分为磷酸化丝/苏氨酸结合模序、磷酸化酪氨酸结合模序、酪氨酸激酶底物模序、丝/苏氨酸激酶底物模序、DNA损伤激酶底物模序、SH2结构域结合模序、SH3结构域结合模序、磷脂结合模序等多种类型,因此可以进行Abl等26种激酶底物的预测,及与多种重要信号蛋白的SH2、SH3、PTB结构域以及14-3-3蛋白的结合特性的分析。

磷酸化丝/苏氨酸结合模序	14-3-3 mode 1	PDZ结合模序	PDZ class 1, PDZ class 2, PDZ (nNOS) class 1, PDZ (nNOS) class 3
磷酸化酪氨酸结合模序	Shc	磷脂结合模序	PIP3-binding PH
酪氨酸激酶底物模序	Abl, EGFR, FGFR, insulin receptor, Itk,, Lck, PDGFR, Src	激酶结合模序	Erk1, PDK1
丝/苏氨酸激酶底物模序	Akt, Calmodulin- dependent kinase 2, Casein kinase 1, Casein kinase 2, Cdc2, Cdk5, Clk2, Erk1, GSK3, p38 MAP kinase, Protein kinase A, PKC α/β/γ, PKC δ, PKCε, PKCζ, PKCμ	SH2结构域结合模序	Abl, Crk, FGFR, Fyn, Grb2, Itk, Lck, Nck, p85, PLC γ, Shc, SHIP, Src
DNA损伤激酶底物模序	ATM, DNA protein kinase	SH3结构域结合模序	Abl, Amphiphysin, Cbl-associated protein, Cortactin, Crk, Grb2, Intersectin, Itk, Nck, p85, PLC γ, Src

图1 Scansite包含的模序分类表

Fig.1 Current list of motifs included in Scansite

2.2 使用方法

登录网址http://scansite.mit.edu进入Scansite程序主页面,选择Motif Scan程序,其中又可以选择输入待分析蛋白的注册号或其氨基酸序列。以原癌蛋白RET为例,选择输入蛋白序列,在DATA entry输入其氨基酸序列。原癌基因*c-ret*编码的蛋白为受体酪氨酸激酶,与肾脏发育、神经系统发育和辐射致甲状腺癌等密切相关[4]。将其全长1114个氨基酸输入,选择分析所有模序。由于高严谨度搜索得到的模序在这个例子中较少(3个),为了便于说明问题,选择中严谨度搜索,得到结果如图2所示。

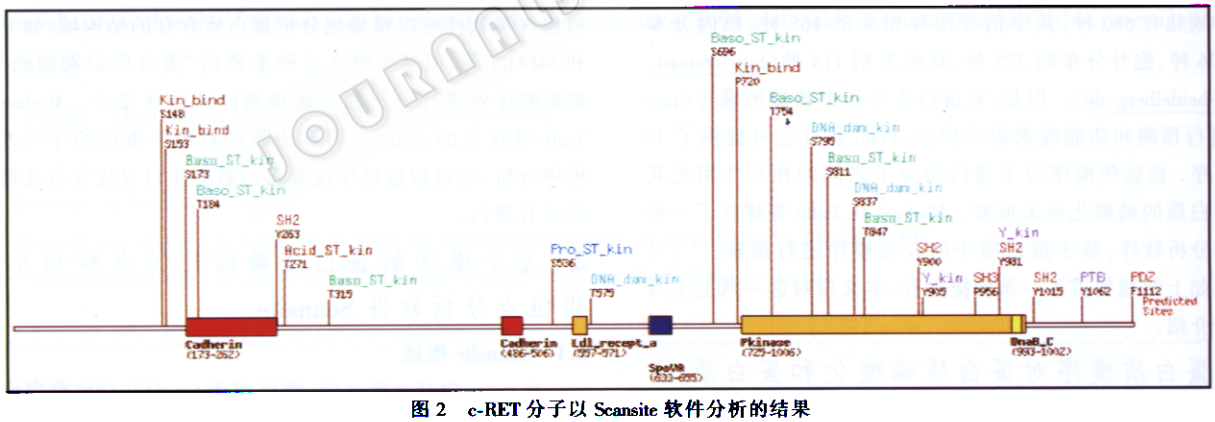

图2 c-RET分子以Scansite软件分析的结果

Fig.2 Analysis result of c-RET by Scansite program

图2以简洁的示意图形式给出了RET所含的结构域:cadherin(173-262)、cadherin(486-506)、ldl-recept-a (557-571)、pkinase (725-1006)等。在示意图的上方以不同颜色标识出可能的磷酸化位点及激酶名称、可能的结合结构域名称等。在此图的下方列出了详细的分析结果,图3列出了部分分析得到的模序。其中有些预测已得到实验验证,如与p85 SH2结构域结合、被PKA磷酸化等[5,6]。

进一步,通过点击Gene Card PRKACG可以链接至PKA的GeneCards。GeneCards是人类基因、基因产物及其在疾病中的作用的数据库,提供所有库中列出基因的功能的准确信息,实质上是生物医学信息的电子百科全书。GeneCards以卡片的形式给出所查询基因的官方名称、GDB同义列表、小鼠中的同源物、细胞遗传学定位、基因产物功能、相关基因家族、相关疾病列表、有关的研究论文及医学应用等。

利用Scansite的模序搜索的功能可以对功能未知的基因进行分析,以获取功能线索。本实验室在对人22周孕龄胎肝EST库进行新基因发掘时,曾发现并克隆了一个新的p53和Rb结合蛋白PACT,利用Scansite程序进行模序分析发现,其N端存在5个Akt激酶的磷酸化位点:RGRHHS406、RSRSYS549、RGRGKS574、RSRSRS583、RSRSRS593 PP,均符合Akt底物一致模序要求RXRXXS/T(X代表任意氨基酸,S/T为磷酸化位点)[7],其中S593旁侧序列还符合14-3-3结合一致

Nck SH2			Gene Card NCK1	
Site	Score	Percentile	Sequence	SA
Y263	0.1707	0.150 %	VPFPVTVYDEDDSAP	1.858
p85 SH2			**Gene Card PIK3R1**	
Site	Score	Percentile	Sequence	SA
Y981	0.1918	0.599 %	DNCSEEMYRLMLQCW	0.984
Src homology 3 group (SH3)				
Intersectin SH3A			**Gene Card ITSN**	
Site	Score	Percentile	Sequence	SA
P956	0.5137	0.765 %	GNPYPGIPPERLFNL	1.289
Basophilic serine/threonine kinase group (Baso_ST_kin)				
Protein Kinase A			**Gene Card PRKACG**	
Site	Score	Percentile	Sequence	SA
S696	0.0669	0.060 %	SSGARRPSLDSMENQ	1.717
Protein Kinase A			**Gene Card PRKACG**	
Site	Score	Percentile	Sequence	SA
T315	0.1216	0.670 %	GELVRRYTSTLLPGD	1.619
Calmodulin dependent Kinase 2			**Gene Card CAMK2G**	
Site	Score	Percentile	Sequence	SA
S696	0.2623	0.756 %	SSGARRPSLDSMENQ	1.717

图 3　c-RET 分子以 Scansite 软件分析的模序详细信息

Fig.3　Detailed information of motifs in c-RET analyzed by Scansite

序列要求 RSXpSXP(其中 pS 代表磷酸化的丝氨酸)[8],因此 S593 很可能首先被 Akt 磷酸化,而后结合支架蛋白(scaffold protein)14-3-3。同时还含有至少 3 个脯氨酸富含区,可能参与同 SH3 结构域的结合:203PPPPPPIPPPRP215、425 PVPPPPLYPPPPHTLPLPPGVPPP448 和 466PPPGFPPAP475。根据 Scansite 的预测,Abl、Amphiphysin、Crk、cortactin、Grb2、Itk、Nck、p85、PLC-gamma、Src 等蛋白的 SH3 结构域可能与 PACT 的上述脯氨酸富含区结合。这些相互作用的揭示有助于研究 PACT 复合体的组成及其可能的信号传导途径。

2.3　Scansite 的其他功能

Scansite 除了可以对一段已知序列进行模序搜索外,还可以对一些已知的模序进行反向搜索,即搜索蛋白质数据库或数据库亚套(subset),以寻找含有此模序的所有可能蛋白,进行蛋白功能分类和模序功能验证[2,3]。比如,经过大规模肽库筛选技术已经发现丝/苏氨酸激酶 Akt/PKB 的底物分子必须含有 RXRXX-(S/T)-X,这段模序中的 S/T 为 Akt 的磷酸化位点,其-3 和-5 位的 R(Arg)必须保守[7]。根据这个原则,对脊椎动物蛋白质组数据库 GenPept 中收录的 76,310 条序列进行检索,发现了大量已经被鉴定和尚未被鉴定的可能的 Akt 底物分子。许多已知的底物如 FKHR、GSK3、IRS-1、BAD、eNOS 等分析得到的百分率都在 0.1% 以下(高严谨度域值设定为 0.2%),因此分析结果的可信度是相当高的。预测的可能性非常大的 Akt 底物分子包括蛋白激酶 CLK2(percentile: 0.0001%)、MEK3(0.025%)、p53 结合蛋白 MDM2(0.0001%)、疾病相关蛋白 Huntingtin(0.046%)和 Tuberin(0.005%)等,其中有些已经在近期的报道中得到实验证实[9-11]。利用这个软件,有助于发现新的激酶底物分子,对于阐明激酶新的生理功能具有重要的实用价值。

3　Scansite 和其他磷酸化分析软件的比较

3.1　Netphos

NetPhos 同样可以通过 Internet 链接进行蛋白的磷酸化分析,当前版本为 NetPhos 2.0(http://www.cbs.dtu.dk/services/NetPhos/)。NetPhos 要求以 FASTA 格式输入一条或多条序列,结果给出分子中所有的 S/T/Y 位点及可能的磷酸化位点,每个位点的可能性同样以打分(score)的高低来衡量,得分在 0~1之间,NetPhos 设定的域值为 0.5,只有高于 0.5,才被认为可能为磷酸化位点。NetPhos 的不足之处有 3 个方面:一是严谨度比较低,实际分析中得到的可能位点数量非常多,存在大量"假阳性"结果,相比之下,Scancite 设定了 3 种严谨度供选择,一般来说,高严谨度情况下分析得到的结果可靠性是比较高的。二是没有给出可能位点的激酶,NetPhos 只列出位点,而没有列出磷酸化此位点的激酶,因此不能给实验验证提供具体的线索提示。三是 NetPhos 只限于分析 S、T 和 Y 三种磷酸化的氨基酸残基,而 Scancite 除了分析以上三种磷酸化氨基酸外,还进一步给出结合这些磷酸化氨基酸的蛋白,如 14-3-3 是特异结合磷酸化丝/苏氨酸的支架蛋白[12],SH2 和 PTB 结构域可特异地结合磷酸化的酪氨酸等[13]。

3.2　ScanProsite

ScanProsite 是欧洲分子生物学实验室(EMBL)建立的针对 Prosite 数据库进行搜索的软件(http://www.expasy.ch/cgi-bin/scanprosite),ScanProsite 分析得到的结果所列出的种类比较繁多,如糖基化修饰、磷酸化修饰、豆蔻酰化、亮氨酸拉链模式、bZIP 转录因子模式以及各种结构域等,因此针对性不强。ScanProsite 和 Scansite 都可以给出磷酸化位点的可能激酶,并进一步直接链接至其他地址了解此激酶的详细信息。同样地,ScanProsite 也没有提供结合这些磷酸化氨基酸的蛋白信息,所以,它与 Scansite 在一定程度上可以相互弥补各自的不足,使得生物信息学分析的结果更加全面。

4　展望

随着人类基因组计划的完成,蛋白质组学(功能基因组学)将成为分子生物学发展的主流。人类基因组计划测得的大量序列信息还有待破译与分析,有关蛋白质功能的研究既是当前国际上研究的热点,又是研究的难点,没有生物信息学的鼎立相助是不可想象的。在几乎所有的真核生物细胞内,蛋白质磷酸化都是最主要的调控与修饰方式之一。在当前最为热点的细胞信号转导、细胞凋亡与基因组研究中,蛋白质磷酸化介导了几乎所有已知的信号途径[14],蛋白激酶与蛋白磷酸酶的研究报道也越来越多,相应地,越来越多的底物分子和调节分子被发现。许多激酶的底物一致序列被总结归纳,与之相关的特异相互作用模序也被揭示。典型的例子是,以前了解得比较清楚的是磷酸化酪氨酸及其结合模序——SH2 结构域和 PTB 结构域,不同类型的分子结合磷酸酪

氨酸的特异性也比较清楚[13]。近期的研究发现,结合磷酸化的丝/苏氨酸的蛋白分子也有内在的规律性,现在已经发现的有 14-3-3 蛋白、WW 结构域、FHA 结构域等[8]。Scansite 比较全面地总结了介导蛋白质磷酸化和蛋白质-蛋白质作用的特异性模序的特征,是功能基因组学研究者进行相关分析的有力帮手。当前的 Scansite 共收录了 62 种模序,我们应用此软件对一些磷酸化位点已知的蛋白进行分析后,发现其可靠性是现在诸多同类分析工具中较高的,对一些未知功能的新基因进行分析也获得了许多可喜的功能线索,其中一些正在实验验证中。但是,由于激酶的种类远远超出此软件所列出的二十几种,所以,仍难以满足更详尽分析的需要,将来的更新版本相信能够涵盖更多的激酶、更多的模序,为研究提供更准确、更详尽的生物信息服务。

REFERENCES(参考文献)

[1] Venter JC, Mark D, Adam S *et al*. The sequence of the human genome. *Science*, 2001, **291**: 1304 – 1351

[2] Yaffe MB, Leparc GG, Lai J *et al*. A motif-based profile scanning approach for genome-wide prediction of signaling pathways. *Nat Biotechnol*, 2001, **19**: 348 – 353

[3] Obenauer JC, Cantley LC, Yaffe MB. Scansite 2.0: proteome-wide prediction of cell signaling interactions using short sequence motifs. *Nucleic Acids Res*, 2003, **31**: 3635 – 3641

[4] Jhiang SM. The RET proto-oncogene in human cancers. *Oncogene*, 2000, **19**: 5590 – 5597

[5] Besset V, Scott RP, Ibanez CF. Signaling complexes and protein-protein interactions involved in the activation of the Ras and phosphatidylinositol 3-kinase pathways by the c-Ret receptor tyrosine kinase. *J Biol Chem*, 2000, **275**: 39159 – 39166

[6] Fukuda T, Kiuchi K, Takahashi M. Novel mechanism of regulation of Rac activity and lamellipodia formation by RET tyrosine kinase. *J Biol Chem*, 2002, **277**: 19114 – 19121

[7] Obata T, Yaffe MB, Leparc GG *et al*. Peptide and protein library screening defines optimal substrate motifs for AKT/PKB. *J Biol Chem*, 2000, **275**: 36108 – 36115

[8] Yaffe MB, Elia AE. Phosphoserine/threonine-binding domains. *Curr Opin Cell Biol*, 2001, **13**: 131 – 138

[9] Gottlieb TM, Leal JF, Seger R *et al*. Cross-talk between Akt, p53 and Mdm2: possible implications for the regulation of apoptosis. *Oncogene*, 2002, **21**: 1299 – 303

[10] Zhou BP, Liao Y, Xia W *et al*. HER-2/neu induces p53 ubiquitination via Akt-mediated MDM2 phosphorylation. *Nat Cell Biol*, 2001, **3**: 973 – 982

[11] Gratton JP, Morales-Ruiz M, Kureishi Y *et al*. Akt down-regulation of p38 signaling provides a novel mechanism of vascular endothelial growth factor-mediated cytoprotection in endothelial cells. *J Biol Chem*, 2001, **276**: 30359 – 30365

[12] Tzivion G, Avruch J. 14-3-3 proteins: active cofactors in cellular regulation by serine/threonine phosphorylation. *J Biol Chem*, 2002, **277**: 3061 – 3064

[13] Sudol M. From Src Homology domains to other signaling modules: proposal of the 'protein recognition code'. *Oncogene*, 1998, **17**: 1469 – 1474

[14] Graves JD, Krebs EG. Protein phosphorylation and signal transduction. *Pharmacol Ther*, 1999, **82**: 111 – 121

A Motif-based Scanning Approach for Prediction of Protein Phosphorylation

TANG Ying ZHANG Ling-Qiang HE Fu-Chu*

(*Beijing Institute of Radiation Medicine*, *Beijing* 100850, *China*)

Abstract Scansite is a short linear motif-based scanning approach established in the latest two years. It's accessible over the World Wide Web and can be used to identify sequence motifs likely to be phosphorylated by specific protein kinases or likely to bind to specific protein domains such as 14-3-3, SH2 and SH3 domains. The usage and function of the potent approach were reviewed and compared with previously established tools for phosphorylation prediction. The facing problems and application outlook of Scansite in prediction of cell signaling networks within proteomes were also presented.

Key words Scansite, protein phosphorylation, motif

Received: 11-18-2003

This work was supported by the Grant from Chinese National High-tech Program(863) (No.2001AA221111).

* Corresponding author. Tel: 86-10-66931246; Fax: 86-10-68214653; E-mail: hefc@nic.bmi.ac.cn