文章编号：1004-0374(2010)03-0248-04

基于人类蛋白质图集的蛋白质表达的总论

Mathias Uhlén
（瑞典皇家技术研究所）

摘 要：新的人类蛋白质图集4.0版本上已经含有了对应5 000个人类基因的6 000多种抗体。这个版本里已经拥有500多万张高分辨率的免疫组化和激光共聚焦图片。每张图片都是经过优秀的病理学家的注释，从而为功能研究提供知识储备，也可以进行正常和病理组织中蛋白质表达谱的查询和文献检索。一个新的结构实现了，它包括了所有预测的基因(大约20 400个)，并且带有可视化的所有编码蛋白质基因的特征。一个新的搜索工具也已经启动了，它可以执行高级检索功能，包括染色体定位、蛋白质分级和(或)组织特异性的检索。蛋白质图集作为一种搜索工具可以发现癌症诊断学的潜在生物标志物。

关键词：组学；蛋白质表达谱；人类蛋白质组资源；生物标志物；抗体百科全书；依赖抗体的蛋白质组学
中图分类号： **文献标识码**：A

A global view of protein expression based on the Human Protein Atlas

Mathias Uhlén
(AlbaNova University Center, Royal Institute of Technology, Stockholm, Sweden)

Abstract: The new version 4.0 of the Human Protein Atlas (www.proteinatlas.org) has been generated with more than 6 000 validated antibodies corresponding to 5 000 human genes. The portal contains more than 5 million high resolution images generated by immunohistochemistry and confocal microscopy. Each image has been manually annotated and curated by a certified pathologist to provide a knowledge base for fuctional studies and to allow searches and queries about protein profiles in normal and disease tissue. A new structure has been implemented with the inclusion of all predicted genes (approximately 20,400) with a visualization of the encoded protein characteristics for all genes. A new search tool is also launched in which advance queries can be performed, including searches for chromosome location, protein class and/or tissue specificity. Protein atlas as a discovery tool can find potential biomarkers for cancer diagnostics.

Key words: mics; protein profile ; HPR; biomarker; antibodypedia; antibody-based proteomics

现在我们处于一个激动人心的生物医学时代，早在18世纪，生物学研究就已经完成了大量系统生物学的工作。19世纪被称作化学的世纪，其中有1/3的元素是在瑞典发现的，最终门捷列夫发现了元素周期表。20世纪则是物理学家的世纪，他们详细研究了各种物质的组成单元。21世纪则是医学的时代，我们通过系统生物学来研究生物体的组成单元，这将是一个非常令人向往的旅程，每个人都可以去享受个过程。

1 生物体的基本组成

我们都知道生物体是由基本单元——核酸和蛋白质组成：

核酸，显而易见，目前我们在各种基因组测序方面取得了很大的成就，中国也积极参与其中。自1965年以来，公共数据库中的DNA序列每年都在成对数增长，目前仍在继续增长。这种增长很大程度上取决于新一代测序技术的出现，1998发表在Science上的一篇文章采用了基于合成法的两种测序

方法[Uhlen & Nyren (1998) 281, 363-365,称为焦磷酸测序]，2005年美国454公司发明了令人惊叹的454测序仪。紧接着一些科学家的工作使得测序可以在更大的规模上进行，2006年推出了新一代Solexa测序仪器，该仪器将“合成测序”化学法(sequencing-by-synthesis chemistry)与“DNA簇技术”(DNA cluster technology)相结合。2007年SOLiD (Sequencing by Oligonucleotide Ligation and Detection)测序出现。最新的PacBio技术使得测序速度达到了每小时100 Gb，这意味着全基因组测序将会在两到三天内完成，这就开启了个性化基因组学的大门。2007年，诺贝尔奖获得者Watson率先解开了他自己的基因组，很快，任何人只要对自己的基因组感兴趣也可以效仿。

蛋白质是生物体功能的体现者。95%以上的药物是以蛋白质作为靶点，可见蛋白质对医药工业的重要性。而组学是用高通量的技术系统地寻找生物靶标分子。人类蛋白质组有多大呢？蛋白质总共分为五类：第一类为最主要的非冗余蛋白质，大约有20 500种，每个基因对应有一个典型的蛋白质；第二类是蛋白突变体，有大于200 000种，为冗余的蛋白片段(剪切突变体或者是蛋白降解片段)；第三类是蛋白质异形体，约有100 000种以上，这些蛋白质的区别在于它们的翻译后修饰；第四类是各种蛋白质组合的突变体，约有超过1 000万种，这些蛋白是由体细胞DNA重组而产生的；第五类是等位蛋白，有大于75 000种，这些蛋白质的差异由基因突变造成(可编码的单核苷酸多样性)。那么人的细胞膜蛋白质组又有多大呢？人类有5 514种膜蛋白，占了所有表达蛋白质基因的26%。展望蛋白质组学，我们相信基因组学仍然将会是一个基本的信息来源，它不仅为蛋白质的分析提供依据，而更重要的是为生物学和生物医学的研究提供一幅人类蛋白质组的图谱。现在我们已经从大规模分析中得到一些经验教训：大部分蛋白质在人的所有细胞组织器官表达；少量细胞特异性蛋白(<1%)和特定组蛋白(<10%)在特定部位表达；大规模表达图谱与目前胚胎学、组织学的理念比较一致；组织特异性是通过蛋白质在时空上的精确调控来获得的。

2 HPA计划的由来

人类基因组测序表明，大约有20 500个蛋白质编码基因，这就为探索人体组织和细胞的表达图谱提供了可能性。但是，蛋白质组不足之处在于，大部分人源蛋白质由于缺乏特异的亲和试剂从而不能够进行大规模的研究，这就显现了对高质量、验证过的抗体的迫切需要。

以抗体为基础的蛋白质组学就是系统地创造和利用特异性抗体进行蛋白质组研究。利用特异的抗体我们就可以衍生出很多技术，如 染色体免疫共沉淀(CHIp)、免疫印迹(Western bolt)、免疫沉淀、免疫荧光(immunofluorescence)、免疫组化(immunohistochemistry)、流式分选、ELISA等蛋白质分析技术。在这样一个战略基础上，人类蛋白质图集Human Protein Atlas(HPA)的主要目标是构建出一幅包括大量正常和癌症组织，以及人类细胞株的蛋白质表达图谱。HPA是个多学科交叉的研究计划，2003年7月已经启动。对于一些非冗余人源蛋白，通过采用高通量生产特异性蛋白抗体和自动化免疫组化、组织芯片技术把人体组织和细胞的蛋白质图谱结合起来。根据人类基因组序列，部分编码序列被称为蛋白质抗原表位签名标签(PrESTs)，相当于50~150个氨基酸。基于在整个蛋白质组中的相对特殊性，它们被选作每个基因的抗原(图1)。首先通过引物合成、RT-PCR，然后克隆到大肠杆菌内，就可以得到纯化的PrEST片段。PrEST 蛋白质片段紧接着就可以作为抗原去获得多克隆抗体，然后再通过亲和纯化得到打靶每个蛋白质的多个线性表位的特异单抗。用这种方式生产的单抗有许多优势，它们在蛋白质处于不同状态比如变性、线性化、非变性条件下都可以进行各种分析。抗原设计采用PRESTIGE原理，这是一种生物信息学方法，通过PrESTs来筛选抗原，目前已经启动了19 832个基因，还有一个自动化的细胞和组织注释系统。

对于HPA计划来说，抗体验证是非常重要的，验证策略在抗体高通量生产策略的不同层次实行。PrEST克隆和PrEST蛋白质片段的序列分别通过测序和质谱学的方法进行验证。抗体的特异性则是通过免疫印迹和蛋白质芯片的方法检测，主要检测目的抗体是否结合特定大小的目的片段。但是抗体验证的最终结果是根据免疫组化的结果来推断的。最终目的是用免疫组化的手段来描绘出蛋白质的表达模式图。染色的模式图结果需要通过文献和生物信息学来评估，从而对抗体的可信程度进行打分。

因此,我们开发出antibodypedia (抗体百科全书)，www.antibodypedia.org，这是一个抗体验证

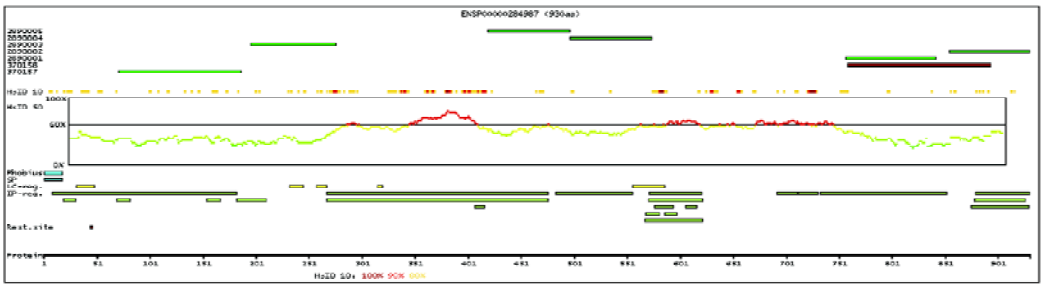

图1 蛋白质抗原表位序列标签(PrESTs)的定位

注：绿色线条表示相对于总蛋白(黑色线条表示的，图的下部)蛋白质抗原表位序列标签(PrESTs)的定位，黄色和绿色曲线表示与其他所有蛋白质的序列相似程度。绿色表示与其他蛋白的同源性最低

的门户网站，是一个通过共享验证好的抗体抗原信息为基础的社群。如果有两种抗体，就可以比较各种检测方法的结果，以“三明治”为基础的分析法可以用来检测其特异性。为了达到这个目标，就需要每个目标蛋白质有与它配对的抗体。实现这个目标还需要很长一段时间。此外，我们还需要确定所有抗体的表位，抗体表位的定位可以通过细菌的表面展示技术、FACS 筛选以及焦磷酸测序来确定。因为抗体的表位分析对于药物靶标的发现非常重要，所以对药物靶标要进行新表位的探索：搜索新的表位以及靶标蛋白质表面上所有的表位。

3 HPR 计划简介

瑞典人类蛋白质资源计划是由Knut 和Alice Wallenberg基金会资助的。它的成立使得大家可以通过以抗体为基础的蛋白质组学对人类蛋白质组进行系统的探索。通过把高通量亲和纯化单克隆抗体和蛋白质表达谱结合起来，目前我们已经得到了许多组织和细胞的蛋白质芯片。共聚焦显微镜对于人细胞系的分析又会进一步得到详细的蛋白质定位的信息。HPR 计划包含了人类蛋白质在各个组织和细胞的表达谱。

该计划的主要站点位于瑞典斯德哥尔摩皇家技术研究所AlbaNova大学中心、乌普萨拉大学的鲁德贝克实验室以及印度孟买的Surgpath实验室。资源中心的主要目的是用高通量的方法(包括基因的克隆和PrESTs 的表达)生产针对人体靶蛋白质的特异性抗体，经纯化后，抗体用来研究蛋白质在细胞和组织的表达谱，以及在各个检测分析平台中进行相应蛋白质的功能研究。斯德哥尔摩站负责生产高质量单克隆抗体和免疫荧光分析，乌普萨拉站负责用免疫组化方法对蛋白质在各种组织和细胞中进行大规模的蛋白质表达谱研究，孟买站则负责免疫组化图的注释。

人类蛋白质图集网作为一个公共数据库网站有着数以百万计的高分辨率图片，其中包括了48种正常组织和20种癌症类型以及47个人类细胞系的蛋白质空间分布信息。此外，抗体与其验证方式将会共同发布，包括免疫组织化学、免疫印迹分析、大量的蛋白质芯片分析以及以荧光为基础的激光共聚焦显微镜图片。该数据库的发展是以基因为中心的，包括了从基因组中预测到的所有人类基因以及每个蛋白质的可视化特点，比如预测到的细胞膜区域、信号肽、蛋白质结构域以及能显示出每个蛋白质与其他人类蛋白质的独特性(序列的相似性)的图片。搜索功能可以对蛋白质表达谱、蛋白质分类和染色体定位进行复杂的查询。

人类蛋白质表达图集含有蛋白质在正常组织、癌细胞、培养的细胞系中表达与定位的免疫组织化学(IHC)以及免疫荧光(IF)共聚焦显微镜图像。2009-06-16发布第五版，包括8 832种抗体，7 334 244 幅图像。

4 HPA 数据库的组织及其内容

基因的染色体信息，到Uniprot、NCBI、EnsEmble 数据库的链接，表达蛋白质所属的分类、谱图等；抗体的HPA 或者CAB 编号；验证结果以及包括四种方法：蛋白质芯片(Protein array)、免疫印迹、免疫组织化学和免疫荧光。不同方法有不同的评分系统，但结果都是使用统一的颜色编码抗原/抗体信息PrEST 信息、抗体信息、以及四种方法验证的信息。

5 检索方式

通过染色体、基因名称、描述信息、EnsEMBLe

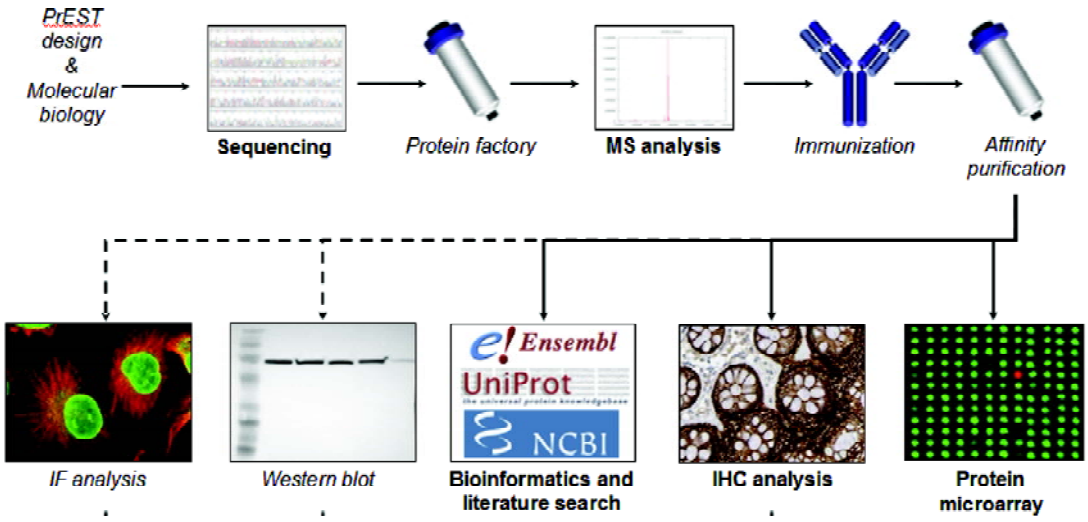

图2　HPA数据库生成流程

基因编号、抗体编号等进行查询，也可以通过蛋白质表达水平、表达组织、蛋白质分类信息进行高级查询，同时也可以通过染色体、蛋白质分类进行浏览。

[参　考　文　献]

[1]　Berglund L, Björling E, Oksvold P, et al. A gene-centric human protein atlas for expression profiles based on antibodies. Mol Cell Proteomics., 2008, 7(10):2019-27

[2]　Uhlén M, Björling E, Agaton C, et al. A Human Protein Atlas for normal and cancer tissues based on antibody proteomics. Mol Cell Proteomics, 2005, 4(12):1920-32

（中国科学院上海生物化学和细胞生物学研究所
曹　灿　编译）