生物化学与生物物理进展 *Progress in Biochemistry and Biophysics*, 2006, 33(8): 800~803
www.pibb.ac.cn

PIBB

科技消息

中国专利基因数据库的创建 *

杨 仑[1,2] 徐朗莱[1)**]

([1]南京农业大学生命科学学院，南京 210095；[2]南京沿溯生物工程有限公司，南京 210095)

摘要 创建了中国专利基因数据库 NASDAP (http://nasdap.generank.org/). 整合了专利序列、专利微阵列、专利基序和专利单核苷酸多态性 (single nucleotide polymorphism，SNP) 等专利对象，并实现对上述对象的 BLAST 检索或基序扫描服务. 这为相关研究的立项、基因研发状态追踪以及基因专利申请和审批等工作提供了生物信息平台，并可为药物开发、疾病诊断和农业等生命科学相关研究的思路启发及知识产权战略制定等方面工作提供参考.

关键词 基因专利，专利基因，数据库，BLAST
学科分类号 Q5，Q6

专利基因是在专利中被权利要求所涵盖的基因，基因专利则体现为包含专利基因的申请公开书或审定授权书. 有关基因专利的知识产权保护问题很早就受到关注[1]. 基因专利处理不当会引发一系列的法律问题[2]或伦理问题[3]. 而即使对于已授权的基因专利，也会因当初申请和查新工作上的缺漏而被判予无效[4]. 基因专利的检索和考察对于生命科学研究战略制定、立项、审批以及执行等均具有重要意义. 这些检索和考察内容包括哪些基因已被专利覆盖，哪些尚未被覆盖，某基因的哪些技术或功能在已覆盖范围之外还有创新余地等相关问题.

随着知识产权逐渐成为科研立项的核心问题之一，人们对专利基因的检索需求也日渐精确化和专业化. 除需进行普通文本检索外，研究者还需进行 BLAST[5]检索、专利微阵列序列检索、专利基序扫描和专利 SNP 扫描，并对它们的法律状态进行考察. 然而专利全文多为非电子化形式的文本，这又成为实现这种检索的瓶颈. 与此同时，国际上生物信息学技术用于专利基因的收集和整理工作已经开始，并将引起科技界极大的关注[6]. 面对上述现状，尽管对我国基因专利实现全文电子化的工作量极大，但我们发现仅对其中的专利基因 (蛋白质) 序列进行电子化便可满足对专利基因的精确化和专业化的检索. 为此构建了中国专利基因数据库 (National Bio-Sequence Database of Chinese Patent，NASDAP)，并免费提供检索服务，以期能够使基因专利文献在较低的电子化程度上为生命科学研究提供尽可能多的信息服务.

1 数据库内容与服务

NASDAP主要收录 1999 年至今，我国申请公开或审定授权的基因专利文本数据及相应的专利基因(蛋白质) 序列. 数据库内容定期更新. 1999 年之前的基因专利因多数未采用知识产权行业标准 (ZC 0003-2001) 且大多已失效而未被收录. 所收录的文本数据包括申请人、标题、摘要以及权利要求等常规信息. 法律状态为审定授权的基因专利在记录中已进行了标注. NASDAP 收录专利序列的具体形式包括普通序列、核酸微阵列序列、专利基序以及 SNP 序列等 4 类 (表 1). 其中，普通序列包括核酸、蛋白质、引物、探针、多肽核酸 (PNA) 和 RNA 序列等. 专利基序是指型如 “一种肽，具有 ‘X_1CYDX_2A’ 的通式，其中 X_1 是 L 或 I，X_2 是 E 或 Q 或缺失” 所描述的序列，在 NASDAP 中以正则表达式[7]的形式存储，专利 SNP 则采用 40 nt 的侧翼序列连同突变位点一同储存为正则表达式的形式.

*国家理科基础科研与教学人才培养基地教育改革研究项目基金资助项目(JD200501).

** 通讯联系人. Tel: 025-84395773, E-mail: xulanglai@njau.edu.cn

收稿日期：2006-02-23，接受日期：2006-04-27

NASDAP收录的 8 563 件基因专利 (123 218 条序列) 由 2 278 个法人或个人申请 (截至 2006 年 2 月 1 日). 公司申请 6 465 件，科研院所申请 2 067 件 (其中公司与科研院所共同申请 282 件)，个人申请 313 件. 申请数最多的五个申请人分别是：上海博德基因开发有限公司、上海博道基因技术有限公司、国家人类基因组南方研究中心、复旦大学和杭州华大基因研发中心. 但目前获得授权最多的 5 个专利权人是：深圳市康哲药业股份有限公司、上海市肿瘤研究所、上海中信国健药业有限公司、上海新世界基因技术开发有限公司和清华大学. NASDAP 所收录的中国基因专利中，国外申请件累计 4 606 件，占总数的 53.8%.

用户从 http://nasdap.generank.org/ 可获得 NASDAP 所提供的检索服务，具体的检索方式和检索实例可从在线帮助中获得. 这些服务除了常规针对标题或摘要的文本检索外，还包括对用户提交的待检序列进行 BLAST 序列同源比对、微阵列序列 BLAST 检索、专利基序扫描以及专利 SNP 扫描等(表 1)，从而使用户了解其待检序列在中国基因专利中的存在形式和法律状态.

表 1 专利基因在 NASDAP 中的序列类型及检索方法

Table 1 Sequence types and relative search methods of patented genes in NASDAP

Sequence type	Count	Search method	Form of storage	Reference(application No.)
Sequence	116 681	BLAST	CKGKGAKCSRLAYDCCTGSCRSGKC	CN00109828.4
Microarray sequence	5 985	BLAST	accatatggtatttgttgcctgtaggtatc	CN200510085480.7
Motif	128	Regular expression matching	WKYM(F|W)M	CN03800414.3
SNP	424	Regular expression matching	tcacctggctctgtcccatc(c|t)agagcctgatgctggccaag	CN200410013383.2

NASDAP中的序列存储为 4 种类型，数据库分别采用 BLAST[5]和正则表达式匹配的方法[7]对其进行检索. 表中第 4 列为上述 4 种序列存储类型的实例，它们的出处列于第 5 列.

2 数据库特点

为说明 NASDAP 区别于普通核酸数据库或专利数据库的特点，现从防止重复研究、挖掘隐藏信息和启迪研发思路三方面举例说明.

2.1 防止重复研究

某研究组在癌旁组织与癌组织差异表达的文库中获得一全长序列，欲了解其功能，将其 ORF 对应的蛋白质序列提交针对 GenBank 的 nr 库 (2006 年 1 月 12 日版) 的 BLASTP 服务. 在所有 BLAST 参数均为默认值的情况下可得到数条 (第一条序列 gi 号为 10732642) 功能未知序列，这表明该基因的功能可能尚未鉴定. 此时研究者希望追加投资对此基因进行下一步研究. 然而，如将此序列对 NASDAP 执行 BLASTP，可得到一条存在于申请号为“CN00111997.4”的中国专利中的序列. 检索该专利摘要得知，该基因早在 2000 年就已由上海市肿瘤研究所申请了较大覆盖范围的物质专利，权利要求涵盖此基因在肿瘤治疗中的应用，并已于 2004 年被授权.

2.2 挖掘隐藏信息

为了解美国 FDA 批准药物“ω 芋螺毒素 MVⅡA”及其类似物在中国的专利申请和授权状况，首先在国家知识产权局网站的专利名称检索栏中输入“芋螺毒素”，结果返回 5 条记录，其中 MVⅡA 的相关专利仅 2 条. 然而用 MVⅡA 的氨基酸序列对 NASDAP 执行 BLASTP 后发现，除检出上述 2 条专利外，还有 5 条 *E* 值最高为 5×10^{-11} 的序列出现在已授权专利“CN00109828.4”中. 进一步检索发现该专利全文中包含 MVⅡA 序列. 由此可见，采用传统文本检索可能造成重要信息的漏检，这是由于作者在专利申请中对“ω 芋螺毒素”采用了“欧米加 - 海螺毒素”的非标准提法而导致的. 然而，通过 NASDAP 对专利基因进行的检索则排除了提法不标准和同义词等原因对检索的影响，因此可以挖掘出此类隐藏在基因专利文献中的重要信息，从而确保了检索结果的全面性.

2.3 启迪研发思路

通过对 NASDAP 执行 TBLASTX 获得东亚钳蝎 α 毒素基因家族专利群的多条序列. 该家族序列间相似性多在 60%以上，且权利要求多样化，如抗昆虫[8]、抗心律失常[9]、抗肿瘤[10]、抗神经兴奋[11]等. 这启发研究人员对该家族具有不同生物活性根源的探索或引发对该家族其他成员开发潜力的思

索. 对该家族序列执行多重比对发现这些序列间存在高度保守的半胱氨酸残基，进一步获悉这是一种名为“CSαβ”的基序[12]，它存在于多个物种内并承载多样的生物学功能. 昆虫防御素、芋螺毒素、人内皮素、蜂毒和 Brazzein 甜味蛋白等均具有该基序. “CSαβ”基序的形成可能是自然选择的结果，类似地，这种基序在专利基因中多次出现，也可以看作是科学界和人类社会选择的结果. 既然自然界和人类都不约而同地选择了“CSαβ”基序，这提示针对拥有此基序的蛋白质的研究开发可能还存在着很多空间.

3 讨 论

当前，序列的查新工作多限于对 GenBank 等公共数据库执行序列同源比对，多数研究者可能忽视并且也无法对专利基因进行 BLAST 检索. 在 NASDAP 创建之前，检索我国专利序列的唯一方法是通过基因名进行文本检索. 然而专利中存在的用词艰深隐讳及基因名使用不规范等问题均可能造成重要信息漏检. 除“海螺毒素”的案例外，类似漏检还会发生于许多采用“肽”、“新肽”等短标题命名以及低信息量摘要的专利之中. 尽管这些文本处理方法是一种申请策略，然而却极有可能导致纠纷. 因此，基因专利文本检索呈现“阴性”的结果并不能反映真实情况，还需进行更加谨慎而深入的考察.

文本检索对已命名序列的检索具有一些作用，却无法实现对未命名序列的考察. 例如，高通量技术带来的今后大量 SNP 和微阵列的专利序列，均属于无法进行文本索引的序列. 这些序列是基因专利的核心，BLAST 检索正是一种不通过标题或摘要而直接到达基因专利核心的检索方法. 然而据 FIZ Karlsruhe 信息研究所的数据，我国专利序列几乎未被任何公共数据库收录. 以 GenBank 为代表的公共数据库仅收录因论文发表需要而提交的序列，它们对我国专利序列无法做到强制收录. 由此表明，如果某基因仅在公共数据库中不存在功能已知的同源序列，研究者并不能将其判定为新基因，更不能武断地追加投资进而对其展开深入研究和开发.

NASDAP 创建的初衷是为了打破上述专利基因考察中亟待打破的瓶颈，并成为对文本检索和序列检索的重要补充. 联合运用其摘要信息和序列信息对科研的参考作用，可类比于联合 PubMed[13]和 GenBank[14]对科研的参考作用. 除此之外，NASDAP 最大限度地保证了专利基因检索和考察的全面性，并维护了专利的公开原则，其更为重要的应用还体现在防止重复研究、避免人力和资源浪费以及防止知识产权纠纷隐患的产生等问题上.

致谢 感谢国家知识产权局和知识产权出版社提供中国专利说明书及专利法律状态等信息，以及陈剑、于维前和夏振华等在数据整理方面的帮助.

参 考 文 献

1 Doll J J. The Patenting of DNA. Science, 1998, **280** (5364): 689～690

2 Abbott A. Clinicians win fight to overturn patent for breast-cancer gene. Nature, 2004, **429** (6990): 329～329

3 Abbott A. Europe pares down double patents on breast-cancer gene. Nature, 2005, **433** (7024): 344～344

4 Paradise J, Andrews L, Holbrook T. Patents on human genes: an analysis of scope and claims. Science, 2005, **307**(5715): 1566～1567

5 Schaffer A A, Aravind L, Madden T L, *et al*. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. Nucleic Acids Res, 2001, **29**(14): 2994～3005

6 Jensen K, Murray F. Intellectual property enhanced: intellectual property landscape of the human genome. Science, 2005, **310** (5746): 239～340

7 Stephens S M, Chen J Y, Davidson M G, *et al*. Oracle Database 10g: a platform for BLAST search and regular expression pattern matching in life sciences. Nucleic Acids Res, 2005, **33**(Database issue): D675～D679

8 陈海宝, 黄 昊. 重组东亚马氏钳蝎毒 rBmKaIT1 的基因工程. 中国专利, 02110581.2. 2002-1-18
Chen H B, Huang H. Genetic engineering for recombining scorpion venom rBmKaIT1. Chinese patent, 02110581.2. 2002-1-18

9 李文鑫, 彭 方, 曾宪春, 等. 蝎抗心律失常肽及其制备方法和应用. 中国专利, 02115487.2. 2002-1-30
Li W X, Peng F, Zeng X C, *et al*. Preparation and the application for an anti-arrhythmia peptide of scorpion. Chinese patent, 02115487.2. 2002-1-30

10 张景海, 马润林, 王思玲, 等. 蝎镇痛抗肿瘤缬精甘肽及获得方法. 中国专利, 01128235.5. 2001-9-30
Zhang J H, Ma R L, Wang S L, *et al*. Analgesic and anti-tumor Val-Arg-Gly peptide of scorpion as well as its acquisition method. Chinese patent, 01128235.5. 2001-9-30

11 张景海, 华子春, 朱德煦. 蝎抗神经兴奋肽. 中国专利, 00112016.6. 2000-1-13
Zhang J H, Hua Z C, Zhu D X. Anti-neuroexcitation peptide of scorpion. Chinese patent, 00112016.6. 2000-1-13

12 Cornet B, Bonmatin JM, Hetru C, *et al*. Refined three-dimensional solution structure of insect defensin A. Structure, 1995, **3** (5): 435～448

13 Wheeler D L, Barrett T, Benson D A, *et al*. Database resources of the national center for biotechnology information. Nucleic Acids Res, 2006, **34**(Database issue): D173～D180

14 Benson D A, Karsch-Mizrachi I, Lipman D J, *et al*. GenBank. Nucleic Acids Res, 2006, **34**(Database issue): D16～D20

Construction of The National Bio-Sequence Database of Chinese Patent*

YANG Lun[1,2], XU Lang-Lai[1)**]

([1)]*College of Life Sciences, Nanjing Agricultural University, Nanjing* 210095, China;

[2)]*Nanjing Ye$u Bio-engineering Co. Ltd., Nanjing* 210095, China)

Abstract There were neither available database nor deep investigation on gene patents or patented genes in China before. But the National Bio-Sequence Database of Chinese Patent (NASDAP, http://nasdap.generank.org/), comprising 123 218 patented bio-sequences within 8 563 Chinese patents, has been constructed now. Genes, microarrays, single nucleotide polymorphisms (SNPs), motifs and any other objects of Chinese patent are involved in, and BLAST-based search or motif scan are carried out to navigate these objects, which can serve as an indispensable bioinformatics platform for setting up of research work, tracing of the advancement in patented genes and facilitating the decision of both applicants and assigners of gene patent. Such platform also gives comprehensive consults on drawing intellectual property strategies in the areas of pharmaceutics, diagnostics and agriculture.

Key words gene patent, patented gene, database, BLAST

---

*This work was supported by grants from The Education Innovation Project of National Basic Science Foundation for Research and Education Training (JD200501).

**Corresponding author . Tel: 86-25-84395773, E-mail: xulanglai@njau.edu.cn

Received: February 23, 2006 Accepted: April 27, 2006