ISSN 1007-7626
CN 11-3870/Q
中国生物化学与分子生物学报 http://cjbmb.bjmu.edu.cn
Chinese Journal of Biochemistry and Molecular Biology
2017 年 8 月
33(8):750~755

·生物数据库专栏·

DOI: 10.13865/j.cnki.cjbmb.2017.08.01

抗体信息数据库

周婷婷， 冯健男*
(军事医学科学院 基础医学研究所 免疫学研究室，北京 100850)

摘要 抗体是介导体液免疫的重要效应分子。近年来，随着治疗性抗体药物不断上市及其临床应用范围不断拓宽，治疗性抗体已成为生物制药的产业支柱。另一方面，随着抗体相关的基础研究日益深入和抗体序列、结构、功能表位等相关数据大量涌现，作为抗体数据管理、搜索与利用的重要工具，抗体资源库也层出不穷并得到长足发展，在相关的研究、开发、生产与销售中发挥日益重要的作用。本文对包括 Kabat、IMGT、abYsis 等在内抗体信息数据库进行介绍。

关键词 抗体；数据库；Kabat；IMGT；abYsis
中图分类号 S852.4+3

Antibody Information Databases

ZHOU Ting-Ting, FENG Jian-Nan*
(*Laboratory of Immunology, Institute of Basic Medical Sciences, AMMS, Beijing* 100850, *China*)

Abstract The antibody is an important type of immunoglobulin that mediates humoral immunity. In recent years antibody drugs have emerged in the worldwide pharmaceutical market. With more therapeutic antibody entering the market and with their widening clinical applications, they have become the pillar of bio-pharmaceutical industry. The antibody information database is an important tool to manage, search and use antibodies. With the deepening of antibody research and the emergence of antibody sequences, structures, functional epitopes and other related data, antibody information databases are successively developed and implemented, and continuously playing an important role in the antibody research, development, production and sale. Here we introduce several antibody information databases, such as Kabat, IMGT, abYsis and hope to promote innovation and development for antibody resources.

Key words antibody; database; Kabat; IMGT; abYsis

抗体(antibody, Ab)是免疫细胞分泌的重要的免疫球蛋白(immunoglobulin, Ig)，能特异性结合抗原并激发机体的免疫应答，是机体调节性免疫系统最重要的组成部分。抗体大多以 1 个或多个 Y 字形单体存在，每个单体都是由 2 条相同的重链和 2 条相同的轻链组成的对称结构。Y 结构顶端为抗原结合部位，称为可变区；Y 结构底部为 Fc 区，与诸如 Fc 受体、C1q 等效应分子作用，激发免疫应答。抗体轻链的 V/J 基因和重链的 V/D/J 基因存在 DNA 水平上的可变剪接，使抗体呈现巨大的多样性，这也是机体存在适应性免疫应答及进行大规模不同抗原识别的重要生物学基础之一。

生物技术的飞速发展与信息技术的进步带来了海量的免疫相关生物学数据，基于这些数据的存储、管理、检索、研究和使用来研究免疫学相关规律的免疫信息学也应运而生。抗体信息学作为免疫信息学的一个重要组成部分，其研究对象包括免疫球蛋白或抗体的相关概念以及用来存储分析免疫球蛋白或抗体的数据和其特性的所有数据库与工具[1]。抗

收稿日期：2017-03-06；修回日期：2017-04-24；接受日期：2017-05-09
国家科技重大新药创制专项课题(No. 2014ZX09304311)和北京市自然科学基金(No. 5152022)资助

* 通讯作者 Tel：010-66931325-2；E-mail：fengjiannan1970@qq.com

Received: March 6, 2017; Revised: April 24, 2017; Accepted: May 9, 2017
Supported by National Science and Technology Major Projects of China for "Major New Drugs Innovation and Development" (No. 2014ZX09304311) and Beijing Municipal Natural Science Foundation (No. 5152022)

* Corresponding author Tel: 010-66931325-2;
E-mail: fengjiannan1970@qq.com

体数据库主要收集和管理免疫球蛋白或抗体的基因和蛋白质序列以及结构等信息,并从抗体序列和结构信息出发,将抗体的结构与功能信息相对应。

自从 1970 年第一个抗体数据库 Kabat 诞生以来,免疫遗传特别是抗体相关的各类数据库相继涌现,其中规模较大且较为常用的包括服务于免疫信息学的、被称为"免疫遗传门户数据库"的国际免疫遗传信息系统 IMGT 和抗体在线分析系统 abYsis,以及服务于免疫相关实验的抗体在线目录 Abcam 和抗体资源指南页 The Antibody Resource Page 等。本文将围绕包括 Kabat 在内的常用资源库进行详细综述,同时列举一些在抗体相关研究中也会涉及的特色数据库,并在结尾简单提及这些数据库使用时应注意的一些问题,也对抗体相关信息库的发展趋势进行展望。

1　常用数据库

1.1　Kabat 数据库

为了确定抗体轻、重链上的抗原结合部位,著名免疫学家 Elvin A Kabat(1914-2000)和他的研究小组于 1970 年创立了 Kabat 数据库[2-3]。Kabat 数据库是世界上第一个免疫学数据库,多年来一直被免疫学领域的相关研究者奉为宝典。Kabat 数据库于 1976 年在 Kabat 的著作《Sequences of Proteins of Immunological Interest》中公布于众,至 1991 年共经历 5 次大的扩充和修订。随着免疫球蛋白、T 细胞受体、MHC-I 类和-II 类分子以及其他免疫相关蛋白质的核酸及蛋白质序列信息的大量涌现并被加入 Kabat 数据库,1991 年 Kabat 数据库通过互联网正式上线,并于 1993 年提供邮箱搜索接口(seqhunt2@immuno.bme.nwu.edu),于 1995 年提供网页搜索接口(http://immuno/bme/nwu/edu)。

随着新序列的不断加入,Kabat 数据库规模越来越大,提供的信息分析工具也越来越多。到 2000 年 7 月,Kabat 数据库共含有 70 个不同物种的各类序列共 19 382 条,其中有 7 989 条经过详细注释[4547 V_H、3442 V_L(包括 κ 和 λ 链)][4]。Kabat 数据库提供的工具涵盖对数据库的序列高级搜索(seqhunt II),针对抗体和其他免疫相关蛋白质的序列比对(Align-A-Sequence)、分组分析(Subgrouping)、组内计数(Current Counts)、家族分析(Find Your Families)和变异性分析(Variability)等。这些数据和工具最初可从其网页接口免费获得使用,现在需要付费注册使用。2003 年后,Kabat 数据库停止更新,其邮箱与网页接口也久已失效。

1.2　国际免疫遗传信息系统 IMGT

IMGT (the international ImMunoGeneTics information system) 由 Marie-Paule Lefranc(Université Montpellier II, CNRS) 于 1989 年建立,是国际知名的免疫相关基因和蛋白质信息系统,目前已发展成为全球免疫遗传信息的门户数据库。IMGT 提供免疫相关的基因组学、蛋白质组学、遗传学及蛋白质二维和三维结构等数据,其数据的精准性和一致性由 IMGT-ONTOLOGY 来保证。IMGT-ONTOLOGY[5-6] 是第一套也是目前唯一一套针对免疫遗传学和免疫信息学所发展和提出的本体论,由 IMGT 归纳和提出,不仅是 IMGT 数据筛选和添加、内部交换和管理的金标准,也是该资源库构建的基石和最大特色所在。

基于 IMGT-ONTOLOGY 所提出的公设(axioms),IMGT 中 Ig 相关数据资源可分为 4 类,分别由下列 7 个子数据库存储和管理: 序列数据库(IMGT/LIGM-DB、IMGT/PRIMER-DB、IMGT/CLL-DB)、基因数据库(IMGT/GENE-DB)、结构数据库(IMGT-3Dstructure-DB、IMGT/2Dstructure-DB)和单克隆抗体数据库(IMGT/mAb-DB)。子数据库及其配套的工具软件之间的逻辑联系如 Fig. 1 所示。IMGT/LIGM-DB[10] 是经详细注释的人类和其他脊椎动物免疫球蛋白 Ig 与 T 细胞受体(T cell receptor, TCR)序列数据库,其核酸序列全部来自于 EMBL(http://www.ebi.ac.cn),目前共收录来自 351 个物种的将近 18 万条序列。IMGT/PRIMER-DB 提供来自 11 个物种的 Ig 和 TCR 的寡核苷酸(引物)标准化数据,目前共有 1 864 个条目,对于正常和病理状态下 Ig 和 TCR 的蛋白质表达、组合抗体库构建、单链抗体设计、噬菌体展示和基因芯片技术等方面的研究十分有用。IMGT/CLL-DB 主要收录慢性淋巴性白血病患者的 Ig 序列,由 IMGT/CLL-DB 小组管理,需要向其提交注册后才能使用。IMGT/GENE-DB[11] 是 IMGT 的基因组数据库,提供人类、小鼠和其他脊椎动物 Ig 和 TCR 基因及其等位基因的序列、分类、国际命名、染色体定位等信息。目前最新版本为 v. 3.1.16 版,共收录 Ig 和 TCR 基因共 4 134 个、等位基因 5 845 个。IMGT/2Dstructure-DB 和 IMGT/3Dstructure-DB 是 IMGT 的结构数据库,IMGT/2Dstructure-DB 是 IMGT/3Dstructure-DB 的组成部分,管理并向 3Dstructure-DB 提供收录和整合自 INN/WHO 和 Kabat 数据库的序列信息,其序列数据

以 IMGT 独有编码方式进行编码，并以其独有的珍珠图（IMGT Collier de Peries）形式展现；IMGT/3Dstructure-DB[12-14] 管理和共享 Ig、TCR、MHC 和 RPI 的整合自 PDB 数据库（http://www.rcsb.org/pdb/）的三维结构和经分析的结构域内部、结构域和侧链以及重要残基之间的相互作用信息，目前共有条目 4 779 个。IMGT/mAb-DB 是 IMGT 的单抗数据库，收录诊断或治疗用单克隆抗体、免疫用融合蛋白、临床用复合蛋白及临床适应症中免疫系统相关蛋白质等，截至 2014 年 4 月发布的最新版（201415-1），共收录单克隆抗体 301 个、抗体-受体融合蛋白 20 个，总条目共计 456 个。

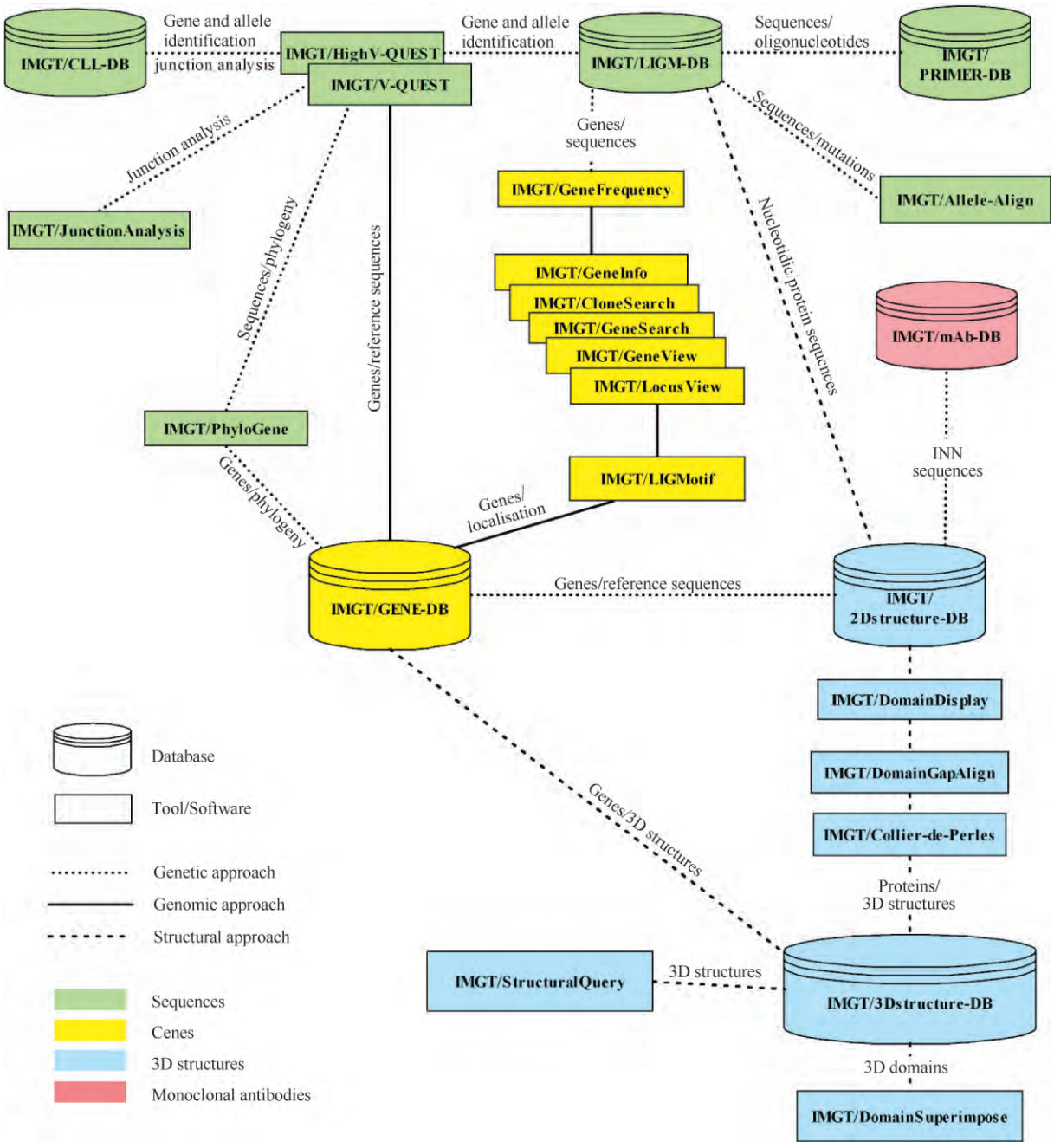

Fig. 1 The major database and tool/software composition of IMGT[9]

IMGT 门户网站还管理超过 15 000 个资源页，并提供与数据库配套的各类分析工具共 17 个。IMGT 提供的工具按照功能大致可分为序列分析、基因分析、结构展示/分析等 3 大类。具体到抗体分析方面，序列分析工具包括专用于 Ig 和 TCR 的 V-Domain 核酸序列比对工具 IMGT/V-QUEST、用于下一代测序结果的高通量分析工具 IMGT/HighV-QUEST、用于 HighV-QUEST 结果的统计分析工具

IMGT/StatClonotype、用于人类和小鼠 IG 与 TCR 的分析工具 IMGT/JunctionAnalysis、用于相关等位基因的比对工具 IMGT/Allele-Align 和进化分析工具 IMGT/PhyloGene 等。基因分析工具包括用于某些人类 Ig、TCR 和 MHC 以及小鼠 TCR 亚类的可视化和搜索工具 IMGT/LocusView、IMGT/GeneView、IMGT/ GeneSearch 和 IMGT/CloneSearch,用于人和小鼠 TCR 基因座中 V-J 和 V-D-J 基因重排的分析工具 IMGT/ GeneInfo 和用于 cDNA 关于 V/D/J/C 基因重排的频度统计工具 IMGT/GeneFrequency 等。结构展示和分析工具包括蛋白质结构域展示工具 IMGT/DomainDisplay、分析工具 IMGT/DomainGapAlign[15]、二维结构珍珠图展示工具 IMGT/Collier-de-Peries[16,17]、结构域比对叠合工具 IMGT/DomainSuperimpose、结构查询工具 IMGT/StructureQuery 等。

目前,IMGT 广泛用于抗体设计、分析、优化、改造及相关生物技术的研究和开发[1],所有资源都可以通过其主页 http://www.imgt.org 免费使用。

1.3　抗体研究系统 abYsis

abYsis[18] 由 Andrew C. R. Martin 研究小组开发,是整合抗体序列和结构信息的在线抗体研究系统,序列信息整合自 EMBL-ENA 和 Kabat 数据库,结构信息整合自 PDB 数据库。当前版本(v2.7.4)共收录 299 367 个条目,其中注释过的 EMBL-IG 条目共 135 871 条,Kabat 条目 45 129 条,PDB 条目 117 270 条,NCBI Germline 条目 459 条,V-Base 条目 638 条,共涉及包括人、小鼠、大鼠、食蟹猴、羊驼等在内的 15 个物种。整个数据库基于 PostgreSQL 关系数据库建立,支持最基本的关键词搜索、基于结构搜索、基于序列搜索和基于氨基酸频度或区域频度(指 CDR 或 FR 区)的频度分布搜索等 4 类,能够满足绝大部分使用者的抗体数据搜索需求,数据的管理和使用非常灵活。

abYsis 以流水线的方式智能读取来自 EMBL、PDB 以及 FASTA 编码格式的抗体数据,利用基因建模工具实现 DNA 和蛋白质序列的联配,并提供 Germline 视图,协助用户在基因组范围内研究序列。使用 Kabat、Chothia 和 Chothia + 等经典位置编码方式对抗体序列自动进行编码注释,并对诸如 CDR 区划分及翻译后修饰位点等关键信息进行注释,并基于所发表的方法对抗体 CDR 区进行正则类(Canonical Class)预测。abYsis 提供氨基酸位置和物种特异的残基分布情况,以及抗体序列各个位置残基出现的频率表,并着重显示在特定位置很少出现的“非常见残基”(unusual residue),以此协助研究者确定可行的氨基酸突变方案。这一功能在比较不同物种来源的抗体时十分有用。

abYsis 还提供抗体数据管理、分析和预测的全套工具。这些工具涵盖针对 Kabat 编码序列数据搜索(KabatMan[19])、抗体序列编码(Abnum[20])、人源化程度评测(G-score[21] 和 H-score[22])、轻重链夹角预测(PAPS[23])、Chothia 正则类预测(Chothia canonicals)及非常见残基预测(SeqTest[19])等。

abYsis 的入口为 http://www.bioinf.org.uk/abysis2.7/,允许用户上传自己的数据进行分析并提供多种数据格式支持,其分析工具也提供单机版供下载使用。

1.4　抗体在线目录 Abcam

抗体在线目录 Abcam(http://www.abcam.com)是抗体信息在线查询系统。该系统不仅提供该公司超过 7 000 兔源单抗(RabMab®)的一抗和超过 2 800 条二抗的详细信息,同时提供相关制剂的研发和使用介绍。作为最大最全面的抗体在线目录之一,也一直为抗体与免疫学相关研究者提供详尽的抗体相关信息,以及使用和操作技术指南的查询和技术支持,在大量研究中都得到广泛的应用。

1.5　抗体资源页 ARP

抗体资源页(The Antibody Resource Page, ARP)(http://www.antibodyresource.com/)被认为是最重要的抗体互联网资源汇总。ARP 不仅分门别类地提供全世界超过 180 家抗体供应商及其产品的详细资料,同时也提供抗体研究相关的各类资源,内容涵盖抗体制备与检测等相关生物学实验指南、抗体相关数据库与分析工具等。此外,网站还提供大量抗体相关的图片和一些专题介绍,以及大量抗体研究相关的教育和出版资源等,兼顾到从事抗体生产、销售、购买、使用与研发的所有供应商、研究者与用户的不同需要。

2　其他资源

除上述资源数据库外,抗体研究有时也会用到小众的特色数据库,如全世界最大的抗体搜索引擎 Citeab(https://www.citeab.com),其特色是搜索结果按照引用数进行排序。根据其网站统计,到投稿时为止共收录来自 145 个供应商的 3 535 379 条各类抗体,受到 1 028 160 次引用;抗体注册器 AntibodyRegistry[24] (http://www.antibodyregistry.

org)，其特色是为来自多个资源的同一抗体提供唯一的注册标识以便于检索和引用；抗体验证数据库 AntibodyValidationDatabase[25] (http://compbio.med.harvard.edu/antibodies)，其特色是收集商用抗体的实验室结果作为其质量和效果的佐证；抗体晶体结构总汇 SACS[26]，其特色是实现抗体晶体结构数据的程序化自管理，数据来自于结构数据库 PDB；抗体结构数据库 SAbDab[27] (http://opig.stats.ox.ac.uk/webapps/sabdab)，其特色是自动收集已知抗体结构信息，并对其进行综合和注释，提供包括实验、基因、抗原信息以及抗原抗体亲和力等。这些数据库同上述常用抗体资源库一起，是抗体研发、改造和生产的重要参考资源。

3 问题与展望

随着抗体研究的日益深入，各类抗体资源信息库也正以不同方式全方位地服务于抗体研发、生产和使用的整个流程。以 IMGT 和 abYsis 为代表的信息综合库提供抗体序列、结构和功能表位等信息的查询、检索与分析服务，借助计算免疫学的分析手段支持抗体的设计、研发和验证；而以 Abcam 和 ARP 等为代表的抗体产品目录和抗体资源平台，则通过提供抗体试剂或产品的详细信息和使用指南来确保抗体在实验过程中的正确选择和使用。

然而，在使用 IMGT 和 abYsis 等数据库提供的抗体序列和结构等资源时，研究者需要注意不同数据库对于抗体序列的编码方式不尽相同。例如 IMGT 采用其特有的编码方式，abYsis 则提供 Kabat、Chothia 和 Chothia + 三种不同的编码方式。这使得各数据库中所提供的对同一抗体可变区中 CDR 区和 FR 区起止位置的定义可能不同。

从世界上第 1 个抗体数据库 Kabat 面世至今 40 年了，抗体资源库得到长足发展。随着抗体研究本身的日益深入以及计算和实验方面大量数据的涌现，抗体相关各类免疫信息数据库也呈现出规模更大、功能定位更加细化的趋势。例如，随着所解析的抗原结构、抗原-抗体复合物结构的不断增加，围绕抗原-抗体相互识别模式的分析数据库、抗体结构的精细分析数据库以及表位数据库等也将进一步发展，为抗体的进一步优化、改进提供基础。可以预见，作为抗体数据管理、搜索与利用的重要工具，抗体数据库或者说抗体信息资源库在抗体及抗体工程的相关研究、开发、生产与销售中的地位和作用也将越来越重要。

参考文献(References)

[1] Lefranc MP. Antibody informatics: IMGT, the international ImMunoGeneTics information system [J]. Microbiol Spectr, 2014, 2(2), doi: 10.1128/microbiolspec.AID-0001-2012

[2] Hood LE. Wu and Kabat 1970: a transforming view of antibody diversity [J]. J Immunol, 2008, **180**(11): 7055-7056

[3] Johnson G, Wu TT. The Kabat database and a bioinformatics example [J]. Methods Mol Biol, 2004, **248**: 11-25

[4] Johnson G, Wu TT. Kabat Database and its applications: future directions [J]. Nucleic Acids Res, 2001, **29**(1): 205-206

[5] Johnson G, Wu TT. Kabat database and its applications: 30 years after the first variability plot [J]. Nucleic Acids Res, 2000, **28**(1): 214-218

[6] Ramirez-Benitez Mdel C, Moreno-Hagelsieb G, Almagro JC. VIR. II: a new interface with the antibody sequences in the Kabat database [J]. Biosystems, 2001, **61**(2-3): 125-131

[7] Giudicelli V, Lefranc MP. IMGT-ONTOLOGY 2012 [J]. Front Genet, 2012, **3**: 79

[8] Lefranc MP. From IMGT-ONTOLOGY IDENTIFICATION axiom to IMGT standardized keywords: for immunoglobulins (IG), T cell receptors (TR), and conventional genes [J]. Cold Spring Harb Protoc, 2011, **2011**(6): 604-613

[9] Giudicelli V, Chaume D, Jabado-Michaloud J, *et al.* Immunogenetics sequence annotation: the strategy of IMGT based on IMGT-ONTOLOGY [J]. Stud Health Technol Inform, 2005, **116**: 3-8

[10] Giudicelli V, Duroux P, Ginestoux C, *et al.* IMGT/LIGM-DB, the IMGT comprehensive database of immunoglobulin and T cell receptor nucleotide sequences [J]. Nucleic Acids Res, 2006, **34**(Database issue): D781-784

[11] Giudicelli V, Chaume D, Lefranc MP. IMGT/GENE-DB: a comprehensive database for human and mouse immunoglobulin and T cell receptor genes [J]. Nucleic Acids Res, 2005, **33**(Database issue): D256-261

[12] Ehrenmann F, Lefranc MP. IMGT/3Dstructure-DB: querying the IMGT database for 3D structures in immunology and immunoinformatics (IG or antibodies, TR, MH, RPI, and FPIA) [J]. Cold Spring Harb Protoc, 2011, **2011**(6): 750-761

[13] Ehrenmann F, Kaas Q, Lefranc MP. IMGT/3Dstructure-DB and IMGT/DomainGapAlign: a database and a tool for immunoglobulins or antibodies, T cell receptors, MHC, IgSF and MhcSF [J]. Nucleic Acids Res, 2010, **38**(Database issue): D301-307

[14] Kaas Q, Ruiz M, Lefranc MP. IMGT/3Dstructure-DB and IMGT/StructuralQuery, a database and a tool for immunoglobulin, T cell receptor and MHC structural data [J]. Nucleic Acids Res, 2004, **32**(Database issue): D208-210

[15] Ehrenmann F, Lefranc MP. IMGT/DomainGapAlign: the IMGT (R) tool for the analysis of IG, TR, MH, IgSF, and MhSF domain amino acid polymorphism [J]. Methods Mol Biol, 2012, **882**: 605-633

[16] Vlachakis D, Feidakis C, Megalooikonomou V, *et al.* IMGT/Collier-de-Perles: a two-dimensional visualization tool for amino acid domain sequences [J]. Theor Biol Med Model, 2013, **10**: 14

[17] Ehrenmann F, Giudicelli V, Duroux P, *et al.* IMGT/Collier de Perles: IMGT standardized representation of domains (IG, TR, and IgSF variable and constant domains, MH and MhSF groove domains) [J]. Cold Spring Harb Protoc, 2011, **2011**(6): 726-736

[18] Swindells MB, Porter CT, Couch M, *et al.* abYsis: integrated antibody sequence and structure-management, analysis, and Prediction [J]. J Mol Biol, 2017, **429**(3): 356-364

[19] Martin AC. Accessing the Kabat antibody sequence database by computer [J]. Proteins, 1996, **25**(1): 130-133

[20] Abhinandan KR, Martin AC. Analysis and improvements to Kabat and structurally correct numbering of antibody variable

domains [J]. Mol Immunol, 2008, **45**(14): 3832-3839
[21] Thullier P, Huish O, Pelat T, *et al*. The humanness of macaque antibody sequences [J]. J Mol Biol, 2010, **396**(5): 1439-1450
[22] Abhinandan KR, Martin AC. Analyzing the "degree of humanness" of antibody sequences [J]. J Mol Biol, 2007, **369**(3): 852-862
[23] Abhinandan KR, Martin AC. Analysis and prediction of VH/VL packing in antibodies [J]. Protein Eng Des Sel, 2010, **23**(9): 689-697
[24] Vasilevsky NA, Brush MH, Paddock H, *et al*. On the reproducibility of science: unique identification of research resources in the biomedical literature [J]. Peer J, 2013, **1**: e148
[25] Egelhofer TA, Minoda A, Klugman S, *et al*. An assessment of histone-modification antibody quality [J]. Nat Struct Mol Biol, 2011, **18**(1): 91-93
[26] Allcorn LC, Martin AC. SACS—self-maintaining database of antibody crystal structure information [J]. Bioinformatics, 2002, **18**(1): 175-181
[27] Dunbar J, Krawczyk K, Leem J, *et al*. SAbDab: the structural antibody database [J]. Nucleic Acids Res, 2014, **42**(Database issue): D1140-1146