

Chapter 7

On the Development of Vaccine Antigen Databases: Progress, Opportunity, and Challenge

Hifzur Rahman Ansari, Darren R. Flower, and Gajendra P.S. Raghava

Abstract The accumulation of relevant and appropriate data is the essential preliminary to any successful informatics-based exercise in prediction. Without quality data, meaningful prediction is impossible. This is as true in immunobiology as it is in any other branch of the natural sciences. Within the context of vaccine discovery, the accumulation, storage, and retrieval of immunological data within publically accessible repositories, typically web-based databases, is of overwhelming operational importance. Specifically, and with the special reference to the discovery of subunit vaccines, this chapter explores the current state and status of immunological databases focussed on immunogenic proteins, primarily pathogen antigens and environmental allergens. It sets this exploration firmly into context by simultaneously scoping out the rather more mature backdrop provided by epitope-orientated database systems.

7.1 Introduction

Vaccine antigen databases, which are typically compiled from numerous different varieties of resources, can lead, either directly or indirectly, to the facilitation of vaccine development. Vaccines are substances that train the host immune system to protect itself against the direct effects of pathogenic microorganisms mediating infectious disease, as well as symptoms of the disease itself. A vaccine is a molecular or composite supramolecular agent eliciting an enhanced adaptive immune response to subsequent infection or reinfection, primarily by potentiating immune memory. Vaccines evoke complex, system-wide immunity of various durations, engendering protection lasting year-long to life-long. The remarkable flexibility inherent in the concept of vaccination has prompted science to use it to

D.R. Flower (✉)

School of Life and Health Sciences, Aston University, Birmingham, UK

e-mail: d.r.flower@aston.ac.uk

attack infectious and contagious disease, allergy, addiction, and cancer, among many other targets.

Population-wide vaccination, which also subsumes the beneficial effects of herd immunity, is now commonly agreed to be the most effective prophylactic treatment for infectious disease. Such vaccination also now offers similar hope for many chronic diseases, including among others allergy and cancer. If you find such language dry, then we can further conceptualise how vaccines work in what follows, expressing things in more anthropomorphic terms. Vaccines offer the immune system the opportunity to encounter and respond to diverse structures on the molecular and mesoscopic scales, which are then memorised so that future immune responses are faster and more effective, and without themselves causing illness or disease. A variety of antigenic molecular structures, such as epitopes and PAMPs, can activate different arms of immune system including B-cells, T-cells, and various innate immune responses through the activation of pattern recognition receptors (PRRs), including the current pre-eminent PRR: Toll-like receptors.

Hitherto, vaccine development has been largely categorised by its doggedly empirical nature; as our understanding of immunology has increased, attempts have been made to give vaccine discovery a more explicable, mechanistic basis. With this sea change, has come the opportunity to introduce rational design supported by computational methods for analysis and prediction. Underlying attempts to predict key components of immunological systems are data capable of being understood, manipulated, and utilised, that is: quantitative molecular data and meaningful functional data. Immunoinformatics may reveal or elucidate recondite components of immunological or signalling pathways, or define protein function by establishing structural or sequence similarity to known proteins. Thus, what computational support of vaccinology can achieve is limited. Yet, there are also immense potential opportunities to exploit this technology for the benefit of vaccinology and thus for the benefit of human medicine and the prevention of infectious disease. It is of the greatest importance to understand fully what is useful and what is not.

By offering diverse tools and methods, immunoinformatics is able to support and facilitate experimental and clinical endeavour [1]. To be of value, information that we intend to model and predict must be properly accumulated, collated, annotated, and archived, before being analysed and dissected. This is surely the purpose of a database. Within immunology, such databases have a relatively extended history. Examples such as Kabat [2] and IMGT [3–5] focussed on compiling and annotating host-side sequences and structures. Indeed, for some time, the Kabat database was the only database containing protein sequence data.

In this chapter, we shall offer a succinct overview of immunological databases relevant to the design and discovery of vaccines, offering a concise yet synoptic outline which is comprised of several brief yet insightful descriptions of various publically accessible database systems, each of which impinges, directly or indirectly, upon vaccine design. See also Table 7.1.

Table 7.1 Immunoinformatic databases and servers for immunomics and vaccinomics

<i>Host databases</i>	
IMGT/HLA	http://www.ebi.ac.uk/imgt/hla/allele.html
IMGT/TR	http://imgt.cines.fr/textes/IMGTrepertoire
IPD Database	http://www.ebi.ac.uk/ipd/index.html
Kabat	http://www.kabatdatabase.com/
VBASE	http://www.vbase2.org/
ABG	http://www.ibt.unam.mx/vir/
V BASE	http://vbase.mrc-cpe.cam.ac.uk/
<i>Pathogen databases</i>	
APB	http://www.engr.psu.edu/ae/iec/abe/database.asp
APDD	http://psychro.bioinformatics.unsw.edu.au/pathogen/index.php
ARS	http://www.ars.usda.gov/research/projects/projects.htm? accn_no=406518
BROP	http://www.brop.org/
EDWIP	http://cricket.inhs.uiuc.edu/edwipweb/edwipabout.htm
FPPD	http://fppd.cbio.psu.edu/
LEGER	http://leger2.gbf.de/cgi-bin/expLeger.pl
ORALGEN	http://www.oralgen.lanl.gov/
Pathema	http://pathema.jcvi.org/Pathema/
ShiBASE	http://www.mgc.ac.cn/ShiBASE/
STDGen	http://www.stdgen.lanl.gov/
VBI	http://phytophthora.vbi.vt.edu/
VIDIL	http://insectweb.inhs.uiuc.edu/Pathogens/VIDIL/index.html
VFDB	http://zdsys.chgb.org.cn/VFs/main.htm
CandiVF	http://research.i2r.a-star.edu.sg/Templar/DB/CandiVF/
TVfac	http://www.tvfac.lanl.gov/
PRINTS	http://www.jenner.ac.uk/BacBix3/PPprints.htm
ClinMalDB-USP	http://malariadb.ime.usp.br/malaria/us/bioinformaticResearch.jsp
Fish Pathogen database	http://dbstdb.nus.edu.sg/fpdb/about.html
PHI-BASE	http://www.phi-base.org/
<i>T-cell databases</i>	
AntiJen	http://www.ddg-pharmfac.net/antijen/AntiJen/aj_tcell.htm
EPIMHC	http://bio.dfci.harvard.edu/epimhc/
FIMM	http://research.i2r.a-star.edu.sg/fimm/
HLA Ligand Database	http://hlaligand.ouhsc.edu/index_2.html
HIV Immunology	http://www.hiv.lanl.gov/immunology
HCV Immunology	http://hcv.lanl.gov/content/immuno/immuno-main.html
IEDB	http://epitope2.immuneepitope.org/home.do
JenPep	http://www.jenner.ac.uk/jenpep2/
MHCBN	http://www.imtech.res.in/raghava/mhcbn
MHCPEP	http://wehih.wehi.edu.au/mhcpep
MPID-T	http://surya.bic.nus.edu.sg/mpidt/
SYFPEITHI	http://www.syfpeithi.de
<i>B-cell databases</i>	
AntiJen	http://www.ddg-pharmfac.net/antijen/AntiJen/aj_bcell.htm
BCIPEP	http://www.imtech.res.in/raghava/bcipep

(continued)

Table 7.1 (continued)

CED	http://web.kuicr.kyoto-u.ac.jp/~ced/
EPITOME	http://www.rostlab.org/services/epitome/
IEDB	http://epitope2.immuneepitope.org/home.do
HaptenDB	http://www.imtech.res.in/raghava/haptendb/
HIV Immunology	http://www.hiv.lanl.gov/immunology
HCV Immunology	http://hcv.lanl.gov/immuno/
<i>Innate immunity and evolution databases</i>	
InnateDB	http://www.innatedb.ca/
ImmTree	http://bioinf.uta.fi/ImmTree
ImmunomeBase	http://bioinf.uta.fi/ImmunomeBase
Immunome Knowledge Base	http://bioinf.uta.fi/IKB/
PRRDB	http://www.imtech.res.in/raghava/prrdb

7.2 Immunoinformatics Databases

For a very long time, indeed rather longer than there has been such a discipline, a principal bed rock of immunoinformatics has been that collection of focussed molecular biology databases, christened IMGT [3–5]. IMGT, standing for the international ImMunoGeneTics information system (URL: <http://www.imgt.org>), describes itself as the world reference in immunogenetics and immunoinformatics, and was created in 1989 at the Laboratoire d'ImmunoGénétique Moléculaire LIGM (Université Montpellier 2 and CNRS, Montpellier, France). Rather than being a monolithic database system, IMGT instead comprises databases, interactive online tools, and Web resources; this database collective concentrates on the compilation and annotation of host-side nucleic acid and protein sequences and corresponding 3D structures. IMGT is an integrated resource encompassing the principal molecules of the adaptive immune system of humans and other vertebrate species: immunoglobulins (IGs) or antibodies, T-cell receptors (TR), and major histocompatibility (MH) complexes. IMGT also covers proteins of the immunoglobulin superfamily, proteins of the MH superfamily, and related proteins of the immune systems (RPI) from both vertebrates and invertebrates, together with therapeutic monoclonal antibodies and fusion proteins for immune applications.

The IMGT/HLA database (URL: <http://www.ebi.ac.uk/imgt/hla/>) is an independent and important self-contained component within the larger edifice of IMGT [6, 7]. The initial release of IMGT/HLA occurred in December 1998, being subsequently updated every 3 months, and includes all publicly available human major histocompatibility complex (MHC) sequences named by the WHO Nomenclature Committee. IMGT/HLA is a locus-specific database for human allelic sequences of the genes of the MHC or HLA system, which is found in the 6p21.3 region of chromosome 6, containing over 220 genes. More widely, the MHC is generally regarded as the most polymorphic of all proteins from higher vertebrates. Growth in the database reflects clinical interest in the MHC as a key mediator of host responses to infection and of tissue rejection in organ and stem cell transplants. The first version of IMGT/HLA included 964 alleles for 24

genes, while the April 2011 release contained in excess of 6,400 alleles for 34 genes. Another important database in the host area is VBASE2, which stores germ-line sequences of human and mouse IG variable (V) genes [8].

The Immuno Polymorphism Database (IPD) system (URL: <http://www.ebi.ac.uk/ipd/>), which covers data from important nonhuman species, including primates, as well as cattle, sheep, and other commercially important farm livestock, has emerged recently from IMGT's long shadow [9–12]. IPD comprises four specialist database systems. IPD-ESTDAB: a database of melanoma cell lines. IPD-KIR: a database of alleles of killer-cell IG-like receptors. PD-HPA: a database of human platelet alloantigens. IPD-MHC: a database of MHC sequences from different species.

A major omission to all extant immunoinformatics and immunogenetics database systems, including IMGT and IPD, is the requirement for a proper database of mouse MHC allele sequences equipped with and based upon a sound and solid consolidated system of nomenclature. The mouse MHC region is located on Chromosome 17, and for historical reasons is referred to *H-2*, for histocompatibility 2, with the different genes denoted by additional letters: D, L, K for Class I and A and E for Class II. Thus, a Class I mouse MHC allele will bear the somewhat inscrutable labels: H-2 Kb or H-2 Db. Within the context of understanding immunological mechanisms underlying vaccine development, the mouse is unarguably the most convenient and thus the most widely used model organism. In bred strains of mice thus provide the principal tool of empirical vaccinology. The albino laboratory strain Balb/c has the *d* haplotype comprising H2-Kd, H2-Dd H2-Ld, H2-IAd, and H2-IEd; the agouti lab strain C3H/He has the *k* haplotype comprising H2-Kk, H2-Dk, H2-IAk, H2-IEk; and the black mouse C57BL/6 has the *b* haplotype composed of H2-Kb, H2-Db, and H2-Iab. The agouti strain CBA also has the *k* haplotype. Outbred mice will obviously contain a greater variety of alleles and thus the need for a more extensible and flexible naming system. However, compilation of such a murine-orientated database, or even the formulation of a more overtly systematic and lucid nomenclature, has progressed little in the last 30 years.

For a long time now, there have been databases that focus upon properties of cellular immunology and look primarily at data pertinent to the discussion of the complexities of MHC processing, presentation, and subsequent T-cell recognition. Perhaps, the earliest such database [13, 14], and probably the first to gain any prominence, is SYFPEITHI (URL: <http://www.syfpeithi.de/>). This is still widely used database system is a high quality development, comprising a useful compendium of known T-cell epitopes. SYFPEITHI also comprises information on MHC ligands: peptides extracted ex vivo from MHCs on the cell surface. It does not include data on synthetic peptide “binders.” “Binders” are peptides with experimentally measured affinity for MHCs. There are several well-used means to measure affinities: IC₅₀ values [15–17]; BL₅₀ values; association and dissociation equilibrium constants, and the half-life of radioisotope-labelled β_2 -microglobulin–MHC complex [18–21]. No consensus has ever emerged on the most appropriate type of affinity measure.

MHCPEP [22–25], defunct database of some vintage, brought together epitope data from T-cells and peptide-MHC (p-MHC) binding data, subsequently making it available as a web-based database system [26]. Subsequently, MHCPEP was developed into a more sophisticated and ambitious database system, called FIMM [27–29]. This integrates data on MHC–peptide interactions: in addition to MHC–peptide binding data and T-cell epitopes, FIMM also includes sequence data on MHC proteins themselves and also information on the disease associations of MHC alleles. Other examples include EPIMHC (URL: <http://imed.med.ucm.es/epimhc/>), a database of T-cell epitopes and MHC-binding peptides [30]. At the time of writing, EPIMHC included 4,867 immunogenic sequences from a plethora of sources. The HLA Ligand Database is another, now defunct databank, with data from the area of T-cell epitology, including information on binding motifs and HLA ligands [31].

Another database of interest is MPID, and its successors: MPID-T and MPID-T2 [32–34]. Focussing primarily on TCR/pMHC interactions, MPID-T is a curated MySQL database which focusses on experimental structures of pMHC and TcR/pMHC complexes (URL: <http://biolinf.org/mpid-t2>). Other related databases have emerged, each addressing data on different aspects of molecular immunology. Middleton and co-workers [35, 36] have described the Allele Frequency Database (URL: <http://www.allelefrequencies.net/>) which lists population frequencies of individual MHC alleles and haplotypes.

Let us now turn to databases that address the humoral, rather than the cellular, immune response. BciPep [37, 38] (URL: <http://www.imtech.res.in/raghava/bcipep/>): sourced from the primary scientific literature, it comprises 3,031 linear B-cell epitopes sequences, within a PostgreSQL database. Cross-referenced to the SWISS-PROT and PDB databases, epitopes are shown as sequences annotated with the host protein, the binding antibody, and the experimental method of immunogenicity determination. The Epitome database (<http://www.rostlab.org/services/epitome/>) consists of structurally inferred antigenic regions, as identified from X-ray protein structures. Together with a host of related information, including experimental methods, antibody, and antigen, the Discontinuous Epitope Database (CED) adumbrates experimentally determined discontinuous epitopes excised from the scientific literature (URL: <http://immunet.cn/ced/index.php>) [39].

Another interesting database, also targeting humoral immunology, is HaptenDB [40]. It presently comprises 2,021 entries for 1,087 haptens and 25 carrier proteins (URL: <http://www.imtech.res.in/raghava/haptendb/>). Entry details the nature of the hapten, as well as its 2D and 3D structures, and its associated carrier proteins. Data is also at present available for anti-hapten antibody production method, the coupling method, and the specificities of antibodies.

There also exist a number of combined databases which seek to cover both cellular and humoral immunology. The first such, and also arguably the best and most detailed and exhaustive, if not necessarily the broadest and most general, is the HIV Molecular Immunology Database (URL: <http://www.hiv.lanl.gov/content/immunology/>) [41]. It archives CD8+ and CD4+ T-cell epitopes, and also B-cell epitopes, derived from various strains of the HIV virus. The database features

sequence alignments, viral protein epitope maps, drug-resistant viral protein sequences, and vaccine-trial data. At the time of writing, T-cell epitope data numbers 3,150 entries, describing 1,600 distinct MHC class I-epitopes. The HCV database comprises 510 entries describing 250 distinct MHC class I-epitope combinations [42]. Perhaps, in the future, all databases will be like the HIV Molecular Immunology Database.

AntiJen v2.0 [43] is another database system, somewhat more ambitious in its scope. It retains information on B- and T-cell epitopes, but also archives quantitative data on a variety of immunological interactions, thus making it both unique and useful. The database (URL: <http://www.ddg-pharmfac.net/antijen/>) contains more than 31,000 entries with 11 sub-databases. These interactions include peptide binding to transport-associated antigen processing (TAP), peptide binding to MHC, p-MHC binding to TCR, and protein–protein interactions in terms of molecular thermodynamic, kinetic, and biophysical data such as receptor copy number and diffusion coefficient

Updated version 4.0 of the MHCBN [44, 45] contains more than 25,000 experimentally characterised T-cell and MHC binding peptides with affinity and activity data compiled from 1,500 papers, as well as from diverse public databases (URL: <http://www.imtech.res.in/raghava/mhcbn/>). Coverage of 1,000 TAP and 4,000 MHC non-binder peptides put MHCBN at a position distinct from other similar immunological databases. Database covers all major hosts like human, mouse, rat, monkey, etc., and provide information of 20 major MHC-linked diseases. The database provides a user-friendly search interface which can be used for simple and customised queries. This also provides several tools for peptide searching, mapping, and BLAST search against MHCBN data.

However, the pre-eminent database of our age is doubtlessly the Immune Epitope Database (IEDB) [46–49]. It is the most comprehensive and progressive database yet available in the area of immunoinformatics. IEDB contains more than 75,000 experimentally characterised B- and T-cell epitopes, MHC binding and elution experiment data compiled from 10,500 literature sources, and data directly submitted by a series of labs (URL: <http://www.immuneepitope.org/>). They have covered 2,500 organisms based on the priority set by National Institute of Allergy and Infectious Diseases which includes allergens and auto-antigens. IEDB also provides useful complementary tools based on the latest algorithms for the analysis and prediction of epitopes and MHC binding. It provides improved user interface for general and advanced searching or browsing by source organism or MHC allele type.

7.3 Antigen and Virulence Factor Databases

Arguably, the clearest and least ambiguous example of an antigen is the so-called virulence factor (VF). Such antigenic proteins induce disease and/or undertake host colonisation. Analysis of known pathogenic species reveals recurrent “systems” of toxins and VFs. VFs have been categorised into several thematic groups: exotoxins,

adherence/colonisation factors, transporters, invasins, iron-binding siderophores, and miscellaneous cell surface factors. A simpler yet more synoptic definition places VFs into three groups: (1) “true” virulence factors; (2) VFs associated with the expression and regulation of class 1 VF genes; and (3) VFs required for host colonisation [50].

Obviously, there is no absolute requirement for antigens to be VFs. Antigens do not need to be directly or indirectly implicated in mechanisms of infectivity; instead, they need only be available for immune surveillance. Thus, databases focussing solely on VFs are not enough and so other kinds of database are needed, capable of capturing and containing more and different relevant data. In the recent past, such a database has been created: AntigenDB [51]; it contains a compilation of in excess of 500 antigens drawn from other immunological resources, as well as the primary scientific literature. AntigenDB is a new stage in the development of immunoinformatics, signalling a switch away from the peptide epitope and towards the whole protein antigen. The AntigenDB database project commenced with the key objective of compiling all immunogenic antigens from all major pathogens for which epitope information is known. AntigenDB (URL: <http://www.imtech.res.in/raghava/antigenadb/>) is intended to help researchers working on the development of antigen prediction algorithms; and also to aid vaccinologists who would like to use these antigens as a tool in the discovery of subunit vaccine candidates.

In AntigenDB, a database entry summarises a large amount of relevant information regarding antigens. It lists origin, structure, sequence, etc., with additional information such as MHC binding, T-cell and B-cell epitopes, gene expression, function, and post-translational modifications. Currently, AntigenDB contains more than 500 antigens coming from important pathogenic species. These antigens originate from 44 key pathogenic species. Entries comprise information pertaining to structure, origin, sequence, etc. They also hold information such as MHC binding data, B- and T-cell epitope, function, and gene expression information if available. A database can be searched by using simple keywords or browsing using pathogen names, after which data can be easily exported to an Excel file. AntigenDB also provides useful tools like BLAST for analysis of the antigenic similarity, epitope search to verify if similar epitopes exist in the database, and peptide mapping to determine the presence of previously known epitopes. AntigenDB also provides links to major internal and external databases. Data submission tool is also provided to submit data online which is included in the database after validation. AntigenDB will be updated on a continuous and continuing basis, with the regular, if infrequent, addition of new antigens and antigens from new organisms. AntigenDB should act as the kernel for future attempts to predict antigens accurately using sequence similarity or via more complex and subtle analysis.

A related database system has been christened VIOLIN (vaccine investigation and online information network) [52]: it allows straightforward curation, and the analysis and comparison of research data across diverse pathogens in the context of human medicine, animal models, laboratory model systems, and natural hosts. It contains 2,759 vaccines or candidates for 173 pathogens culled from 1,631 peer-reviewed papers, as well as 24,345 vaccine-related abstracts and 10,317 full-text

articles. VIOLIN is a web-accessible resource (URL: <http://www.violinet.org>). Peer-reviewed vaccine-related literature has been downloaded into VIOLIN; and the database is searchable in various ways. Archived information includes vaccine preparation and characteristics, microbial pathogenesis, host-protective immunity, stimulated host responses post vaccination, and protection efficacy after challenge.

Vertebrates have developed an efficient immune system and pathogens need to adapt to this very hostile environment; therefore, they devise different counter strategies like antigenic drift and variation to evade a host immune system. Some pathogens mutate their antigens so that antibodies are no longer able to recognise them, while some hide themselves from the immune cells like malaria-causing pathogen plasmodium to reside in red blood cells. Genome-sequencing efforts revealed the existence of a large number of multi-copy gene families with suspected involvement in antigenic variation and virulence. The genomics and proteomics projects release a huge amount of sequence and phenotypic data. The varDB project [53–56] aims to produce a dedicated platform for the antigen variation, phenotypes, and clinical manifestations (URL: <http://www.vardb.org/vardb/>). So far, the varDB database has covered 42 gene families, accommodating 67,000 sequences from 27 species of organisms causing 19 different diseases. varDB has created a framework for automatic data download, processing, and annotation. Entries can be queried and sorted based on the UniProt ID, gene products, isolation source, host, locus tag, etc. varDB has implemented a user-friendly tool, a shopping cart where sequences can be added and analysed, and a correlation with disease outcome can be revealed, leading to the identification of functionally important residues and drug targets.

Several other databases targeting VFs have also been reported. They include TVFac (Los Alamos National Laboratory Toxin and Virulence Factor database; URL: <http://www.tvfac.lanl.gov/>). This has within it information on in excess of 250 organisms, with records for thousands of virulence genes. The *Candida albicans* virulence factor (CandiVF) is a fully searchable species-specific database containing diverse VFs [57]. The Virulence Factor Database (VFDB; URL: <http://www.mgc.ac.cn/VFs/>) targets 16 bacterial genomes; it focusses on functional and structural biology and is fully searchable in several ways [58, 59]. PHI-base integrates VFs from a veritable gallimaufry of plant and animal pathogens [60]. The Fish Pathogen Database (URL: <http://www.fishpathogens.eu/vhsv/index.php>), from the Bacteriology and Fish Diseases Laboratory, archives 500 VFs from fish.

7.4 Allergen Databases

Allergy is a complex pathological condition caused by allergenic proteins resulting from uncontrolled IgE antibody-mediated immune responses. Allergies of all types are escalating as the world moves inexorably towards a western lifestyle, with all its benefits and dangers. This gives rise to a reduction in the depth and breadth of immunity to microbial pathogens. As a consequence, the prevalence of atopy rises. Making inappropriate and immediate hypersensitivity reactions to harmless substances is the mark of atopy. Thus, significant exposure to pathogens, both

viral and bacterial, during early life is vital to the proper regulation of allergen-specific immune responses. If such inappropriate and potentially catastrophic reactions are not the ransom to be delivered by certain populations for their freedom from microbial diseases, we must be prepared to continually train our immune systems, especially during infancy, in order to prevent allergic conditions. Allergy affects over 30 % of the population in developed countries, with a concomitant major economic health burden [61].

The aforementioned hypersensitivity reactions occur when allergens are bound by IgE tethered to the surface of basophils or mast cells. Allergenic proteins form a distinct class of antigens; they are able to induce specific production of IgE, which in turn leads to the secretion of inflammatory mediators. An individual sensitised to, say, pollen-derived allergens might additionally react to allergens also present in plant-based foods as a result of the intrinsic commonality of sequence and structure. This cross-reactive of distinct allergenic proteins from different sources can be problematic. Owing to the importance of allergens, a number of databases and analysis tools have been created. Fortuitously, a standard for the naming of allergens, maintained by Allergen Nomenclature Sub-committee of the World Health Organization and International Union of Immunological Societies (WHO/IUIS), has now been adopted, making construction of databases much more facile than hitherto. Major active databases in this area are adumbrated, limned, and otherwise outlined below.

The Structural Database of Allergenic Proteins, or SDAP, was released in 2001 and has been continuously updated [62, 63]. Currently, SDAP contains 1,396 allergens, 1,221 protein sequences and 28 IgE and IgG epitopes, 70 allergens with PDB structures, and 582 3D models. SDAP also provides information on a wide variety of relevant information, including: allergen name, IUIS nomenclature, type, source species, sequence, IgE epitopes, literature, family domain, and structure information when available (URL: <http://fermi.utmb.edu/SDAP/>). Database can be searched through a user-friendly interface. SDAP also hosts an array of computational tools for the prediction of allergenicity and cross-reactivity.

The Allergen Database for Food Safety [64, 65], or ADFS, was originally an initiative of the Japanese National Institute of Health Sciences (URL: <http://allergen.nihs.go.jp/ADFS/>). It was updated in February 2010, mainly deals with food safety, principally addressing Novel Foods and related immunochemistry. The database comprised 1,285 sequences, 77 molecular structures, and 91 epitopes. It also contains 88 sugar-attached allergens derived from air, pathogens, food, and venom/salivary allergens. ADFS provides domain information, links to other external databases, and tools for the allergen analysis and prediction.

Allergome [66, 67] is arguably the most complete resource for allergens produced to date and is updated on a daily basis; up-to-date statistics can be found at the Allergome statistics page (URL: <http://www.allergome.org/script/statistic.php>). The database comprises 173 biological functions, 1,759 allergen sources, 53 tissues, over 700 structures, 17,000 bibliographies, totalling 5,276 entries. Allergome entries contain information on allergen name, route of exposure, taxonomy, source tissue and organism, and cross-reactivity; as well as data on protein

primary sequence, motifs, and post-translational modifications. Allergome introduced the concept of allergenicity score as a real-time characterisation status for a given allergen. This score can be visualised using a dedicated page using a 3-colour, 11-dot sequence. Dots represent data generation, sequence availability, and different functional/non-functional tests. Users can search the database for allergens matching these scores. The database also includes a user-friendly reference archive and an “allergome aligner” tool for the analysis of allergen sequences.

7.5 Discussion and Conclusion

The principal means of vaccine development remains largely empirical, although targets have moved away from an exclusive focus on live virulent or attenuated pathogens as candidate vaccines. Safety concerns have led to such entities being viewed as poor vaccine candidates despite their potent immunogenicity, although seasonal flu vaccines are still of this type. They are more likely to cause disease in children or immune-compromised individuals. Due to safety and availability of powerful adjuvants, the trend is shifting towards antigen, acellular, or subunit- and epitope-based vaccines. Approaches to antigen identification can include cumbersome assays for epitopes, antibodies, cytokine levels, as well as testing for extended survival within animal test groups. Most of the databases described in this chapter catalogue antigens for which B- and T-cell epitopes are known, other antigens, despite high intrinsic immunogenicity, are often lost in the literature. Yet, such antigens could be useful in the context of vaccine design and discovery, hence, the recent development of databases specifically target antigens as opposed to epitopes.

There remain many outstanding or unresolved challenges in the development of antigen database. We need immunology to assess antigens with the same rigour that it applies to the characterizing of epitopes, providing quantitative data on protein affinity measures for binding to B-cell receptors or soluble antibodies. All antigen databases should coordinate and define a universal format such as has been done for other protein sequence and structure databases. Terminology should be properly defined much as it has been done for vaccine ontology by the developers of IEDB. Lastly, and perhaps most importantly, updating should be a constant and consistent process, never stopping always progressing: this remains a remorseless, and thus challenging, task.

References

1. Flower DR (2007) Immunoinformatics and the in silico prediction of immunogenicity. An introduction. *Methods Mol Biol* 409:1–15
2. Wu TT, Kabat EA (1970) An analysis of the sequences of the variable regions of Bence Jones proteins and myeloma light chains and their implications for antibody complementarity. *J Exp Med* 132(2):211–250
3. Lefranc MP (2011) IMGT, the international ImMunoGeneTics information system. *Cold Spring Harb Protoc* 2011(6):595–603

4. Lefranc MP, Giudicelli V, Ginestoux C, Jabado-Michaloud J, Folch G, Bellahcene F, Wu Y, Gemrot E, Brochet X, Lane J et al (2009) IMGT, the international ImMunoGeneTics information system. *Nucleic Acids Res* 37:D1006–D1012, Database issue
5. Lefranc MP (2008) IMGT, the International ImMunoGeneTics Information System for Immunoinformatics: methods for querying IMGT databases, tools, and web resources in the context of immunoinformatics. *Mol Biotechnol* 40(1):101–111
6. Robinson J, Mistry K, McWilliam H, Lopez R, Parham P, Marsh SG (2011) The IMGT/HLA database. *Nucleic Acids Res* 39:D1171–D1176, Database issue
7. Robinson J, Waller MJ, Fail SC, McWilliam H, Lopez R, Parham P, Marsh SG (2009) The IMGT/HLA database. *Nucleic Acids Res* 37:D1013–D1017, Database issue
8. Retter I, Althaus HH, Munch R, Muller W (2005) VBASE2, an integrative V gene database. *Nucleic Acids Res* 33:D671–D674, Database issue
9. Robinson J, Marsh SG (2007) IPD: the immuno polymorphism database. *Methods Mol Biol* 409:61–74
10. Robinson J, Mistry K, McWilliam H, Lopez R, Marsh SG (2010) IPD—the immuno polymorphism database. *Nucleic Acids Res* 38:D863–D869, Database issue
11. Robinson J, Waller MJ, Fail SC, Marsh SG (2006) The IMGT/HLA and IPD databases. *Hum Mutat* 27(12):1192–1199
12. Robinson J, Waller MJ, Stoeckl P, Marsh SG (2005) IPD—the immuno polymorphism database. *Nucleic Acids Res* 33:D523–D526, Database issue
13. Schuler MM, Nastke MD, Stevanovik S (2007) SYFPEITHI: database for searching and T-cell epitope prediction. *Methods Mol Biol* 409:75–93
14. Rammensee H, Bachmann J, Emmerich NP, Bachor OA, Stevanovic S (1999) SYFPEITHI: database for MHC ligands and peptide motifs. *Immunogenetics* 50(3–4):213–219
15. Sette A, Sidney J, del Guercio MF, Southwood S, Ruppert J, Dahlberg C, Grey HM, Kubo RT (1994) Peptide binding to the most frequent HLA-A class I alleles measured by quantitative molecular binding assays. *Mol Immunol* 31(11):813–822
16. Sidney J, Oseroff C, del Guercio MF, Southwood S, Krieger JI, Ishioka GY, Sakaguchi K, Appella E, Sette A (1994) Definition of a DQ3.1-specific binding motif. *J Immunol* 152(9):4516–4525
17. Kubo RT, Sette A, Grey HM, Appella E, Sakaguchi K, Zhu NZ, Arnott D, Sherman N, Shabanowitz J, Michel H et al (1994) Definition of specific peptide motifs for four major HLA-A alleles. *J Immunol* 152(8):3913–3924
18. DiBrino M, Parker KC, Margulies DH, Shiloach J, Turner RV, Biddison WE, Coligan JE (1994) The HLA-B14 peptide binding site can accommodate peptides with different combinations of anchor residues. *J Biol Chem* 269(51):32426–32434
19. Parker KC, Biddison WE, Coligan JE (1994) Pocket mutations of HLA-B27 show that anchor residues act cumulatively to stabilize peptide binding. *Biochemistry* 33(24):7736–7743
20. DiBrino M, Parker KC, Shiloach J, Turner RV, Tsuchida T, Garfield M, Biddison WE, Coligan JE (1994) Endogenous peptides with distinct amino acid anchor residue motifs bind to HLA-A1 and HLA-B8. *J Immunol* 152(2):620–631
21. Parker KC, Bednarek MA, Coligan JE (1994) Scheme for ranking potential HLA-A2 binding peptides based on independent binding of individual peptide side-chains. *J Immunol* 152(1):163–175
22. Brusic V, Rudy G, Harrison LC (1998) MHCPEP, a database of MHC-binding peptides: update 1997. *Nucleic Acids Res* 26(1):368–371
23. Brusic V, Rudy G, Kyne AP, Harrison LC (1997) MHCPEP, a database of MHC-binding peptides: update 1996. *Nucleic Acids Res* 25(1):269–271
24. Brusic V, Rudy G, Kyne AP, Harrison LC (1996) MHCPEP—a database of MHC-binding peptides: update 1995. *Nucleic Acids Res* 24(1):242–244
25. Brusic V, Rudy G, Harrison LC (1994) MHCPEP: a database of MHC-binding peptides. *Nucleic Acids Res* 22(17):3663–3665

26. Hon L, Abernethy NF, Brusic V, Chai J, Altman RB (1998) MHCWeb: converting a WWW database into a knowledge-based collaborative environment. *Proc AMIA Symp* :947–951
27. Schonbach C, Koh JLY, Flower DR, Brusic V (2005) An update on the functional molecular immunology (FIMM) database. *Appl Bioinformatics* 4(1):25–31
28. Schonbach C, Koh JLY, Flower DR, Wong L, Brusic V (2002) FIMM, a database of functional molecular immunology: update 2002. *Nucleic Acids Res* 30(1):226–229
29. Schonbach C, Koh JL, Sheng X, Wong L, Brusic V (2000) FIMM, a database of functional molecular immunology. *Nucleic Acids Res* 28(1):222–224
30. Reche PA, Zhang H, Glutting JP, Reinherz EL (2005) EPIMHC: a curated database of MHC-binding peptides for customized computational vaccinology. *Bioinformatics* 21(9):2140–2141
31. Sathiamurthy M, Hickman HD, Cavett JW, Zahoor A, Prilliman K, Metcalf S, Fernandez Vina M, Hildebrand WH (2003) Population of the HLA ligand database. *Tissue Antigens* 61(1):12–19
32. Khan JM, Cheruku HR, Tong JC, Ranganathan S (2011) MPID-T2: a database for sequence-structure-function analyses of pMHC and TR/pMHC structures. *Bioinformatics* 27(8):1192–1193
33. Tong JC, Kong L, Tan TW, Ranganathan S (2006) MPID-T: database for sequence-structure-function information on T-cell receptor/peptide/MHC interactions. *Appl Bioinformatics* 5(2):111–114
34. Govindarajan KR, Kanguane P, Tan TW, Ranganathan S (2003) MPID: MHC-Peptide Interaction Database for sequence-structure-function information on peptides binding to MHC molecules. *Bioinformatics* 19(2):309–310
35. Gonzalez-Galarza FF, Christmas S, Middleton D, Jones AR (2011) Allele frequency net: a database and online repository for immune gene frequencies in worldwide populations. *Nucleic Acids Res* 39:D913–D919, Database issue
36. Middleton D, Menchaca L, Rood H, Komerofsky R (2003) New allele frequency database: <http://www.allelefrequencies.net>. *Tissue Antigens* 61(5):403–407
37. Saha S, Raghava GP (2007) Searching and mapping of B-cell epitopes in Bcipep database. *Methods Mol Biol* 409:113–124
38. Saha S, Bhasin M, Raghava GP (2005) Bcipep: a database of B-cell epitopes. *BMC Genomics* 6:79
39. Huang J, Honda W (2006) CED: a conformational epitope database. *BMC Immunol* 7:7
40. Singh MK, Srivastava S, Raghava GP, Varshney GC (2006) HaptenDB: a comprehensive database of haptens, carrier proteins and anti-hapten antibodies. *Bioinformatics* 22(2):253–255
41. Kuiken C, Korber B, Shafer RW (2003) HIV sequence databases. *AIDS Rev* 5(1):52–61
42. Yusim K, Richardson R, Tao N, Dalwani A, Agrawal A, Szinger J, Funkhouser R, Korber B, Kuiken C (2005) Los alamos hepatitis C immunology database. *Appl Bioinformatics* 4(4):217–225
43. Toseland CP, Clayton DJ, McSparron H, Hemsley SL, Blythe MJ, Paine K, Doytchinova IA, Guan P, Hattotuwegama CK, Flower DR (2005) AntiJen: a quantitative immunology database integrating functional, thermodynamic, kinetic, biophysical, and cellular data. *Immunome Res* 1(1):4
44. Lata S, Bhasin M, Raghava GP (2009) MHCBN 4.0: a database of MHC/TAP binding peptides and T-cell epitopes. *BMC Res Notes* 2:61
45. Bhasin M, Singh H, Raghava GP (2003) MHCBN: a comprehensive database of MHC binding and non-binding peptides. *Bioinformatics* 19(5):665–666
46. Peters B, Sidney J, Bourne P, Bui HH, Buus S, Doh G, Fleri W, Kronenberg M, Kubo R, Lund O et al (2005) The immune epitope database and analysis resource: from vision to blueprint. *PLoS Biol* 3(3):e91
47. Sette A (2004) The immune epitope database and analysis resource: from vision to blueprint. *Genome Inform* 15(2):299
48. Vita R, Zarebski L, Greenbaum JA, Emami H, Hoof I, Salimi N, Damle R, Sette A, Peters B (2010) The immune epitope database 2.0. *Nucleic Acids Res* 38:D854–D862, Database issue

49. Zhang Q, Wang P, Kim Y, Haste-Andersen P, Beaver J, Bourne PE, Bui HH, Buus S, Frankild S, Greenbaum J et al (2008) Immune epitope database analysis resource (IEDB-AR). *Nucleic Acids Res* 36:W513–W518, Web Server issue
50. Wassenaar TM, Gaastra W (2001) Bacterial virulence: can we draw the line? *FEMS Microbiol Lett* 201(1):1–7
51. Ansari HR, Flower DR, Raghava GPS (2010) AntigenDB: an immunoinformatics database of pathogen antigens. *Nucleic Acids Res* 38:D847–D853
52. Xiang Z, Todd T, Ku KP, Kovacic BL, Larson CB, Chen F, Hodges AP, Tian Y, Olenzek EA, Zhao B et al (2008) VIOLIN: vaccine investigation and online information network. *Nucleic Acids Res* 36:D923–D928, Database issue
53. Hayes CN, Diez D, Joannin N, Kanehisa M, Wahlgren M, Wheelock CE, Goto S (2009) Tools for investigating mechanisms of antigenic variation: new extensions to varDB. *Genome Inform* 23(1):46–59
54. Diez D, Hayes N, Joannin N, Normark J, Kanehisa M, Wahlgren M, Wheelock CE, Goto S (2010) varDB: a database of antigenic variant sequences—current status and future prospects. *Acta Trop* 114(3):144–151
55. Allred DR, Barbet AF, Barry JD, Deitsch KW (2009) varDB: common ground for a shifting landscape. *Trends Parasitol* 25(6):249–252
56. Hayes CN, Diez D, Joannin N, Honda W, Kanehisa M, Wahlgren M, Wheelock CE, Goto S (2008) varDB: a pathogen-specific sequence database of protein families involved in antigenic variation. *Bioinformatics* 24(21):2564–2565
57. Tongchusak S, Chaiyaroj SC, Veeramani A, Koh JLY, Brusica V (2005) CandiVF—Candida albicans virulence factor database. *Int J Pep Res Ther* 11(4):271–277
58. Yang J, Chen L, Sun L, Yu J, Jin Q (2008) VFDB 2008 release: an enhanced web-based resource for comparative pathogenomics. *Nucleic Acids Res* 36:D539–D542, Database issue
59. Chen L, Yang J, Yu J, Yao Z, Sun L, Shen Y, Jin Q (2005) VFDB: a reference database for bacterial virulence factors. *Nucleic Acids Res* 33:D325–D328, Database issue
60. Winnenburg R, Baldwin TK, Urban M, Rawlings C, Kohler J, Hammond-Kosack KE (2006) PHI-base: a new database for pathogen host interactions. *Nucleic Acids Res* 34:D459–D464
61. Kamble S, Bharmal M (2009) Incremental direct expenditure of treating asthma in the United States. *J Asthma* 46(1):73–80
62. Ivanciuc O, Mathura V, Midoro-Horiuti T, Braun W, Goldblum RM, Schein CH (2003) Detecting potential IgE-reactive sites on food proteins using a sequence and structure database, SDAP-food. *J Agric Food Chem* 51(16):4830–4837
63. Ivanciuc O, Schein CH, Braun W (2003) SDAP: database and computational tools for allergenic proteins. *Nucleic Acids Res* 31(1):359–362
64. Nakamura R, Teshima R (2009) Major revision of the allergen database for food safety (ADFS) and validation of the motif-based allergenicity prediction tool. *Kokuritsu Iyakuhiin Shokuhin Eisei Kenkyusho Hokoku* 127:44–49
65. Nakamura R, Teshima R, Takagi K, Sawada J (2005) Development of Allergen Database for Food Safety (ADFS): an integrated database to search allergens and predict allergenicity. *Kokuritsu Iyakuhiin Shokuhin Eisei Kenkyusho Hokoku* 123:32–36
66. Mari A, Scala E (2006) Allergome: a unifying platform. *Arb Paul Ehrlich Inst Bundesamt Sera Impfstoffe Frankf AM* 95:29–39, discussion 39–40
67. Chardin H, Peltre G (2005) Allergome: the characterization of allergens based on a 2D gel electrophoresis approach. *Expert Rev Proteomics* 2(5):757–765