Christian Schönbach
Shoba Ranganathan
Vladimir Brusic
Editors

Immunoinformatics

Springer

Immunoinformatics

Immunoinformatics

Edited by

Christian Schönbach
School of Biological Sciences
Nanyang Technological University
Singapore

Shoba Ranganathan
Biotechnology Research Institute
Macquarie University
Sydney, Australia

and

Vladimir Brusic
Cancer Vaccine Center
Dana-Farber Cancer Institute
Boston, Massachusetts, USA

Springer

Christian Schönbach
Division of Genomics and Genetics
School of Biological Sciences
Nanyang Technological University
60 Nanyang Drive, Singapore 637551
Singapore
schoen@ntu.edu.sg

Shoba Ranganathan
Department of Chemistry
and Biomolecular Sciences
Biotechnology Research Institute
Macquarie University
Sydney, NSW 2109
Australia
shoba@els.mq.edu.au

Vladimir Brusic
Director of Bioinformatics
Cancer Vaccine Center
Dana-Farber Cancer Institute
77 Avenue Louis Pasteur, HIM 418
Boston, Massachusetts 02115
USA
vladimir_brusic@dfci.harvard.edu

ISBN 978-0-387-72967-1 e-ISBN 978-0-387-72968-8

Library of Congress Control Number: 2007937448

Printed on acid-free paper.

9 8 7 6 5 4 3 2 1

springer.com

Foreword

Christian Schönbach[1,2]

[1] Immunoinformatics Team, Advanced Genome Information Technology Group, RIKEN Genomic Sciences Center, 1-7-22 Suehiro-cho, Tsurumi, Yokohama, Kanagawa 230-0045, Japan

[2] Present address: Division of Genomics and Genetics, School of Biological Sciences, Nanyang Technological University, 60 Nanyang Drive, Singapore 637551, Singapore, schoen@ntu.edu.sg

Immunoinformatics is an emerging subdiscipline of bioinformatics. It utilizes mathematics, information science, computer engineering, genomics, proteomics, and immunological methods to bridge immunology and informatics (Petrovsky, Schönbach, and Brusic 2003; Brusic and Petrovsky 2003). Similar to bioinformatics which became a driving force in genome research, immunoinformatics enables data-driven research strategies (Van Regenmortel 2006; Moise and De Groot 2006; Korber, LaBute, and Yusim 2006) and systems approaches (Kitano and Oda 2006; Gilchrist, Thorsson, Li, Rust, Korb, Kennedy, Hai, Bolouri, and Aderem 2006) that aim at understanding the networks regulating the immune system. Considering the breadth of topics the volume *Immunoinformatics* was composed to provide a cross section of research ranging from data integration, epitope predictions to systems-level applications. In ten chapters experts introduce and discuss research strategies for immunologists and bioinformaticians who wish to explore existing and new approaches to gain insight into the workings of the immune system.

Data analysis and formulation of new hypotheses concerning the immune system is aided by standardized, integrated databases and tools. In Chapter 1 Lefranc reviews a powerful resource, the international ImMunoGeneTics information system® (http://imgt.cines.fr) IMGT®. Since its beginning in 1989, IMGT® has become an indispensible reference in immunogenetics and immunoinformatics with more than 140,000 accesses per month. The Web resource provides not only standardized data of nucleotide and protein sequences, oligonucleotide primers, gene maps, genetic polymorphisms, specificities, 2D and 3D structures but also analysis tools for immunoglobulins, T-cell receptors, major histocompatibility complex (MHC) molecules, and related proteins of the immune system.

In Chapter 2 Kaas, Duprat, Tourneur, and Lefranc present IMGT® 3D structures and tools that can improve our understanding of the mechanisms governing T-cell-receptor–peptide–MHC recognition. Details of structural analyses and predictions, including docking techniques to predict potential T-cell epitopes, are discussed by Ranganathan, Tong, and Tan in Chapter 3. Hattotuwagama, Doytchinova, Guan, and Flower are taking structural immunoinformatics of peptide-MHC class I and class II

allele complexes to the next level of high-throughput Quantitative Structure-Activity Relationship (QSAR) technologies. The authors discuss in Chapter 4 2D and 3D QSAR methods for the quantitative prediction of peptide-MHC affinity. The latter is an important factor that influences immunogenicity and therefore the identification of T-cell epitopes.

While structural informatics and epitope mapping methods have enhanced vaccine design, the use of computational methods is less prominent in research that aims to counter unfavorable immune responses in transplantation, autoimmunity and allergies. Lee and Brusic give in Chapter 5 an overview of allergy informatics. In the first part, the authors discuss various specialized databases and point out the absence of an integrated database similar to IMGT, that contains a comparable superset of all allergens. In the second part we are introduced to a number of powerful allergenicity prediction methods. However, the current lack of a suffciently large standardized dataset for training and testing reduces the applicability of the prediction methods. Next, De Groot, Knopf, Rivera, and Martin describe in Chapter 6 an intriguing combination of recombinant protein expression and epitope mapping to eliminate antitherapeutic antibody and autoimmune reactions associated with therapeutic proteins. The approach promises to increase the number protein therapeutics and their safety.

The previous chapters cover individual molecules and a limited number of interactions involved in immune response with little consideration of their regulation, molecular pathways, and functions in cells, organs, or whole organism context. To understand the mechanism of a favorable or unfavorable immune response we need to examine, for example, the networks that regulate immune cell phenotypes or natural lymphocyte homeostasis.

The next four chapters introduce strategies suitable to investigate the immune system on a network level. Kellam and Kwan discuss in Chapter 7 host-pathogen interactions focusing on gene expression programs in dendritic cells and plasticity of pathogen-sensing functional states. The next two chapters describe mathematical or computer models of the immune response to HIV infection. In Chapter 8, Bernaschi and Castiglione apply mathematical modeling almost to the level of a virtual patient. Using the modeler called C-ImmSim (http://www.iac.rm.cnr.it/~filippo/C-ImmSim.html) they make predictions on population dynamics, phenotype and specificity of lymphocytes (anergic, proliferating, etc.), viremia and proviral HIV, concentration of anti-HIV antibodies and strenght of cytotoxic response during progression toward AIDS. Another interesting application is the modeling of immune response behavior during an opportunistic infection with *M. tuberculosis*.

Da Silva characterizes in Chapter 9 the adaptation of HIV to immune surveillance at the molecular-genetic level considering both the dynamics of the humoral response to HIV and the viral fitness. The simulations elegantly demonstrate relationships between coreceptor selection, antibody selection, and viral adaptation.

Finally, Gondo presents in Chapter 10 a sophisticated experimental and computational system with mouse as *bona fide* "simulator" (http://www.gsc.riken.go.jp/Mouse/). The ethylnitrosourea (ENU) mutagenesis-based system is one large-scale effort that will bring us closer to understanding how genome sequences affect immune response and immune cell behavior.

Acknowledgements

I wish to thank Dr. Akiyoshi Wada, former RIKEN GSC director, for his encouragement and support for the First International Immunoinformatics Symposium which was held at RIKEN GSC on February 26 and 27, 2004. I am also grateful to all the authors who updated their symposium presentations and contributed them as chapters to the volume *Immunoinformatics*.

References

Brusic, V., and Petrovsky, N. (2003) Immunoinformatics─the new kid in town. Novartis Found. Symp. 254:3–13.

Gilchrist, M., Thorsson, V., Li, B., Rust, A.G., Korb, M., Kennedy, K., Hai, T., Bolouri, H., and Aderem A. (2006) Systems biology approaches identify ATF3 as a negative regulator of Toll-like receptor 4. Nature 441:173–178.

Kitano, H., and Oda, K. (2006) Robustness trade-offs and host-microbial symbiosis in the immune system. Mol. Syst. Biol. 2:2006.0022.

Korber, B., LaBute, M., and Yusim, K. (2006) Immunoinformatics comes of age. PLoS Comput. Biol. 2:e71.

Moise, L., and De Groot, A.S. (2006) Putting immunoinformatics to the test. Nat. Biotechnol. 24:791–792.

Petrovsky, N., Schönbach, C., and Brusic, V. (2003) Bioinformatic strategies for better understanding of immune function. In Silico Biol. 3:411–416.

Van Regenmortel, M.H. (2006) Immunoinformatics may lead to a reappraisal of the nature of B cell epitopes and of the feasibility of synthetic peptide vaccines. J. Mol. Recognit. 19:183–187.

Contents

Contributors

Massimo Bernaschi
Institute for Computing Applications (IAC), National Research Council (CNR), Viale del Policlinico 137, 00161 Rome, Italy

Vladimir Brusic
Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613, Present address: Cancer Vaccine Center, Dana-Farber Cancer Institute, 77 Avenue Louis Pasteur, HIM 418, Boston, MA 02115, USA

Filippo Castiglione
Institute for Computing Applications (IAC), National Research Council (CNR), Viale del Policlinico 137, 00161 Rome, Italy

Jack da Silva
School of Molecular and Biomedical Science, The University of Adelaide, Adelaide, SA 5005, Australia

Irini A. Doytchinova
Edward Jenner Institute for Vaccine Research, Compton, Berkshire RG20 7NN, UK

Elodie Duprat
IMGT®, the International ImMunoGeneTics Information System®, Université Montpellier II, Laboratoire d'ImmunoGénétique Moléculaire LIGM, UPR CNRS 1142, Institut de Génétique Humaine IGH, 141 rue de la Cardonille, 34396 Montpellier Cedex 5, France

Darren R. Flower
Edward Jenner Institute for Vaccine Research, Compton, Berkshire RG20 7NN, UK

Yoichi Gondo
RIKEN Genomic Sciences Center, Functional Genomics Research Group, Population and Quantitative Genomics Team, 1-7-22 Suehiro-cho, Yokohama, Kanagawa 230-0045, Japan

Anne S. De Groot
EpiVax, Inc., 146 Clifford Street, Providence, RI 02903, USA
Brown University, Department of Medicine, Providence, RI 02912, USA

Pingping Guan
Edward Jenner Institute for Vaccine Research, Compton, Berkshire RG20 7NN, UK

Channa K. Hattotuwagama
Edward Jenner Institute for Vaccine Research, Compton, Berkshire RG20 7NN, UK

Quentin Kaas
IMGT®, the International ImMunoGeneTics Information System®, Université Montpellier II, Laboratoire d'ImmunoGénétique Moléculaire LIGM, UPR CNRS 1142, Institut de Génétique Humaine IGH, 141 rue de la Cardonille, 34396 Montpellier Cedex 5, France
Present address: Institute for Molecular Bioscience, The University of Queensland, St Lucia Q 4072, Australia

Paul Kellam
Centre for Virology, University College London, 46 Cleveland Street, London W1T 4JF, UK

Paul M. Knopf
EpiVax, Inc., 146 Clifford Street, Providence, RI 02903, USA

Antonia Kwan
Centre for Virology, University College London, 46 Cleveland Street, London W1T 4JF, UK

Bernett T.K. Lee
Department of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, 8 Medical Drive, MD7 #02- 03, Singapore 117597,
Present address: Bioinformatics Institute, 30 Biopolis Street, #07-01 Matrix, Singapore 138671

Marie-Paule Lefranc
IMGT®, the International ImMunoGeneTics Information System®, Université Montpellier II, Institut Universitaire de France, Laboratoire d'ImmunoGénétique Moléculaire LIGM, UPR CNRS 1142, Institut de Génétique Humaine IGH, 141 rue de la Cardonille, 34396 Montpellier Cedex 5, France

William Martin
EpiVax, Inc., 146 Clifford Street, Providence, RI 02903, USA

Shoba Ranganathan
Department of Chemistry and Biomolecular Sciences & Biotechnology Research Institute, Macquarie University, Sydney, NSW 2109, Australia and Department of Biochemistry, National University of Singapore, 8 Medical Drive, Singapore 117597

Daniel Rivera
EpiVax, Inc., 146 Clifford Street, Providence, RI 02903, USA

Christian Schönbach
RIKEN Genomic Sciences Center, 1-7-22 Suehiro-cho, Tsurumi, Yokohama, Kanagawa 230-0045, Japan
Present address: School of Biological Sciences, Nanyang Technological University, 60 Nanyang Drive, Singapore 637551, Singapore

Tin Wee Tan
Department of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, 8 Medical Drive, MD7 #02- 03, Singapore 117597

Joo Chuan Tong
Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613

Guillaume Tourneur
IMGT®, the International ImMunoGeneTics Information System®, Université Montpellier II, Laboratoire d'ImmunoGénétique Moléculaire LIGM, UPR CNRS 1142, Institut de Génétique Humaine IGH, 141 rue de la Cardonille, 34396 Montpellier Cedex 5, France

Chapter 1

IMGT-ONTOLOGY, IMGT® Databases, Tools, and Web Resources for Immunoinformatics

Marie-Paule Lefranc

IMGT®, the International ImMunoGeneTics Information System®, Université Montpellier II, Institut Universitaire de France, Laboratoire d'ImmunoGénétique Moléculaire LIGM, UPR CNRS 1142, Institut de Génétique Humaine IGH, 141 rue de la Cardonille, 34396 Montpellier Cedex 5, France, Marie-Paule.Lefranc@igh.cnrs.fr

Abstract. IMGT®, the international ImMunoGeneTics information system®, was created in 1989 as a high-quality integrated knowledge resource specialized in immunoglobulins (IG), T cell receptors (TR), major histocompatibility complexes (MHC) of human and other vertebrates, and related proteins of the immune system (RPI) which belong to the immunoglobulin superfamily (IgSF) and to the MHC superfamily (MhcSF). IMGT® is the international reference in immunogenetics and immunoinformatics. IMGT® combines sequence databases (IMGT/LIGM-DB, IMGT/PRIMER-DB, IMGT/PROTEIN-DB, IMGT/MHC-DB), a genome database (IMGT/ GENE-DB), and a three-dimensional (3D) structure database (IMGT/ 3Dstructure-DB) with interactive analysis tools (IMGT/V-QUEST, IMGT/JunctionAnalysis) and Web resources comprising 8000 HTML pages (IMGT Repertoire). The accuracy and consistency of IMGT data are based on IMGT-ONTOLOGY, available for biologists and IMGT users in the IMGT Scientific chart and for computer scientists in IMGT-ML, in XML format. IMGT® components (databases, tools, and Web resources) have been developed according to three main biological approaches: the genomic approach that is gene centered, the genetic approach that refers to genes in relation to their polymorphisms, expression, specificity, and evolution, and the structural approach that analyses 3D structures in relation to protein function and recognition sites. We are implementing Web services for the IMGT databases and tools. This is the first step toward IMGT-Choreography that will trigger and coordinate dynamic interactions between IMGT Web services in order to process complex significant biological and clinical requests. IMGT® is widely used in fundamental and medical research (repertoire analysis of the IG antibody sites and of the TR recognition sites in autoimmune diseases, infectious diseases, AIDS, leukemias, lymphomas, myelomas), veterinary research situations (IG and TR repertoires in farm and wildlife species), genome diversity and genome evolution studies of the adaptive immune responses, biotechnology related to antibody engineering (single chain Fragment variable (scFv), phage displays, combinatorial libraries, chimeric, humanized, and human antibodies), diagnostics (clonalities, detection and follow-up of residual diseases), and therapeutical approaches (graft, immunotherapy, vaccinology). IMGT® is freely available at http://imgt.cines.fr.

1.1 Introduction

Genome and proteome analysis interpretation represents the current great challenge, as a huge quantity of data is produced by many scientific fields, including fundamental, clinical, veterinary, and pharmaceutical research. In particular, the number of sequences

and related data published in the immunogenetics fields is growing exponentially. The number of potential protein forms of the antigen receptors, immunoglobulins (IG), and T cell receptors (TR) is almost unlimited. The potential repertoire of each individual is estimated to comprise about 10^{12} different IG (or antibodies) and 10^{12} different TR, and the limiting factor is only the number of B and T cells that an organism is genetically programmed to produce. This huge diversity is inherent to the particularly complex and unique molecular synthesis and genetics of the antigen receptor chains. This includes biological mechanisms such as DNA molecular rearrangements in multiple loci (three for IG and four for TR in humans) located on different chromosomes (four in humans), nucleotide deletions and insertions at the rearrangement junctions (or N-diversity), and somatic hypermutations in the IG loci (Lefranc and Lefranc 2001a; Lefranc and Lefranc 2001b).

IMGT® (http://imgt.cines.fr), the international ImMunoGeneTics information system® (Lefranc, Giudicelli, Kaas, Duprat, Jabado-Michaloud, Scaviner, Ginestoux, Clément, Chaume, and Lefranc 2005a), was created in 1989, by the Laboratoire d'ImmunoGénétique Moléculaire (LIGM) (Université Montpellier II and CNRS) at Montpellier, France, in order to standardize and manage the complexity of the immunogenetics data. IMGT® is the international reference in immunogenetics and immunoinformatics, and represents a high-quality integrated knowledge resource, specialized in the IG, TR, major histocompatibility complex (MHC) of human and other vertebrates, and related proteins of the immune systems (RPI) of any species which belong to the immunoglobulin superfamily (IgSF) and to the MHC superfamily (MhcSF). As such, IMGT® provides a common access to standardized data from genome, proteome, genetics, and three-dimensional (3D) structures.

1.2 The IMGT® Information System

The IMGT® information system consists of databases, tools, and Web resources (Lefranc, Clément, Kaas, Duprat, Chastellan, Coelho, Combres, Ginestoux, Giudicelli, Chaume, and Lefranc 2004a; Lefranc, Giudicelli, Ginestoux, Bosc, Folch, Guiraudou, Jabado-Michaloud, Magris, Scaviner, Thouvenin, Combres, Girod, Jeanjean, Protat, Monod, Duprat, Kaas, Pommié, Chaume, and Lefranc 2004b; Lefranc et al. 2005a). Databases and tools are summarized in Fig. 1.

Databases include several sequence databases (IMGT/LIGM-DB, IMGT/MHC-DB, IMGT/PRIMER-DB, IMGT/PROTEIN-DB), a genome database (IMGT/ GENE-DB), and a 3D structure database (IMGT/3Dstructure-DB). Interactive tools are provided for nucleotide and amino acid sequence analysis (IMGT/V-QUEST, IMGT/JunctionAnalysis, IMGT/Allele-Align, IMGT/PhyloGene, IMGT/DomainDisplay), genome analysis (IMGT/LocusView, IMGT/ GeneView, IMGT/GeneSearch, IMGT/CloneSearch, IMGT/GeneInfo, IMGT/GeneFrequency), and 3D structure analysis (IMGT/StructuralQuery, IMGT/DomainGapAlign, IMGT/Collier-de-Perles, IMGT/DomainSuperimpose). Web resources (IMGT Marie-Paule page) comprise 8000 HTML pages of synthesis [IMGT Repertoire (for IG and TR, MHC, RPI)], of knowledge [IMGT Scientific chart, IMGT Education] (Aide-mémoire, Tutorials, Questions and answers), IMGT Lexique, The IMGT Medical page, The IMGT

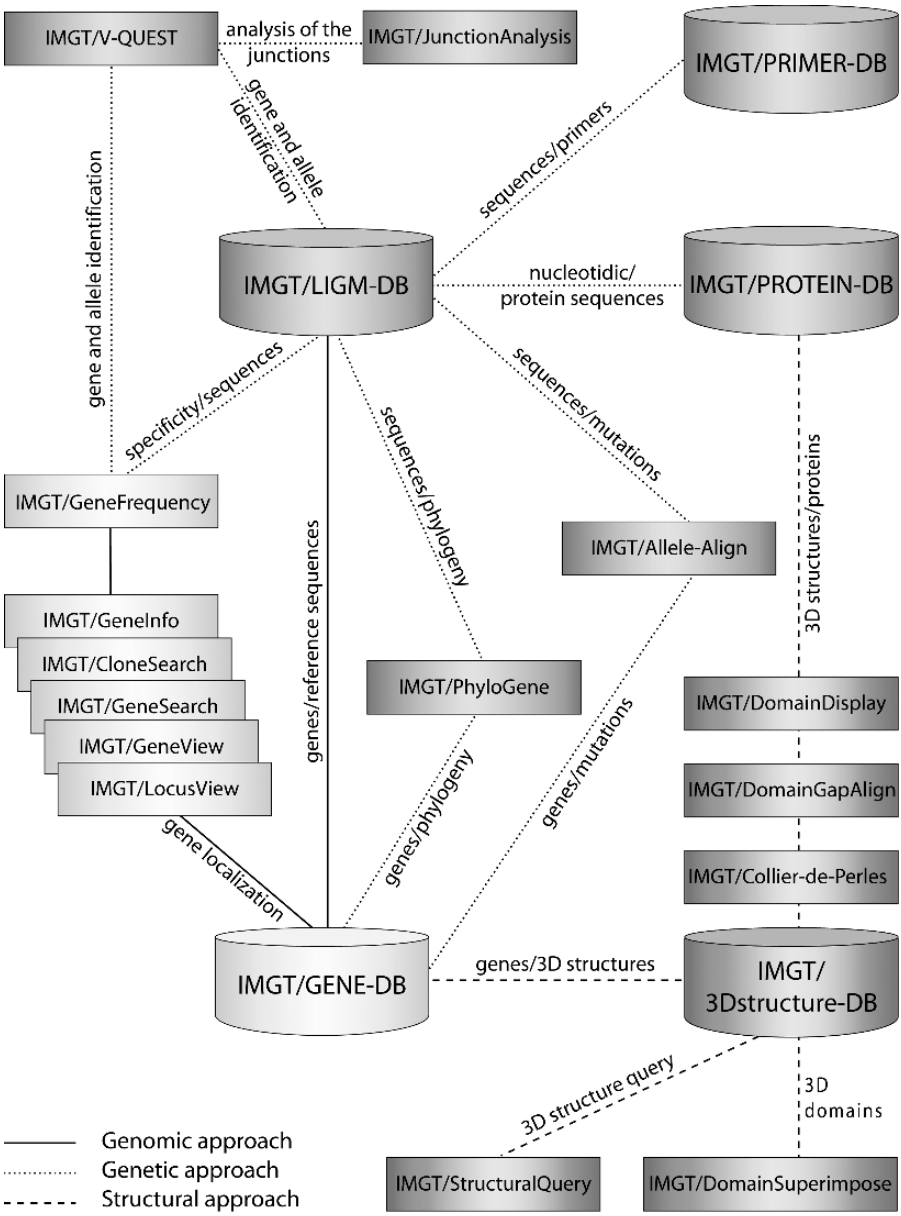


Fig. 1. IMGT®, the international ImMunoGeneTics information system® (http://imgt.cines.fr) databases and tools. The IMGT Repertoire and other IMGT Web resources are not shown. Examples of interactions between the databases (cylinders) and tools (rectangles) in the genomic, genetic and structural approaches are represented respectively by continuous, dotted and broken lines. (A color version of this figure appears between pages 76 and 77.)

Veterinary page, The IMGT Biotechnology page, IMGT Index], and external links [IMGT Bloc-notes (The IMGT Immunoinformatics page, Interesting links), and Other accesses (SRS, BLAST)]. Despite the heterogeneity of these different components, all data in the IMGT® information system are expertly annotated. The accuracy, the consistency, and the integration of the IMGT® data, as well as the coherence between the different IMGT® components (databases, tools, and Web resources) are based on IMGT-ONTOLOGY (Giudicelli and Lefranc 1999), the first ontology in the domain, which provides a semantic specification of the terms to be used in immunogenetics and immunoinformatics, and thus allows the management of immunogenetic knowledge for all vertebrate species. IMGT-ONTOLOGY comprises seven main concepts: IDENTIFICATION, CLASSIFICATION, DESCRIPTION, NUMEROTATION, LOCALIZATION, ORIENTATION, and OBTENTION (Giudicelli and Lefranc 1999; Lefranc et al. 2004a; Lefranc et al. 2004b; Lefranc et al. 2005a). Standardized keywords, standardized IG and TR gene nomenclature, standardized labels, the IMGT unique numbering, annotation rules, and standardized origin/methodology were defined, respectively, based on these seven main concepts.

IMGT-ONTOLOGY concepts are available for biologists and IMGT® users in the IMGT Scientific chart and formalized for computer scientists in IMGT-ML using XML (Extensible Markup Language) Schema. The IMGT Scientific chart (Lefranc, Giudicelli, Ginestoux, Bodmer, Müller, Bontrop, Lemaitre, Malik, Barbié, and Chaume 1999) comprises controlled vocabulary and annotation rules for data and knowledge management of the IG, TR, and MHC of vertebrate species, and of the RPI of any species, that belong to IgSF and MhcSF. All IMGT® data are expertly annotated according to the IMGT Scientific chart rules. The IMGT Scientific chart is available as a section of the IMGT Web resources (IMGT Marie-Paule page). These HTML pages are devoted to biologists, IMGT users, and IMGT annotators. Examples of IMGT expert data concepts, derived from the IMGT Scientific chart rules, correspond to section titles and subtitles in IMGT Repertoire (Lefranc et al. 2004a; Lefranc et al. 2004b).

IMGT-ML (Chaume, Giudicelli, and Lefranc 2001; Chaume, Giudicelli, Combres, and Lefranc 2003; Chaume, Giudicelli, Combres, Ginestoux, and Lefranc 2005) is the formalization of IMGT-ONTOLOGY using XML Schema for interoperability with other information systems. IMGT® components (databases, tools, and IMGT Repertoire Web resources) have been developed according to three main biological approaches. The IMGT® genomic approach is gene-centered focusing on the study of the genes within their loci and on the chromosomes. The IMGT® genetic approach refers to the study of the genes in relation to sequence polymorphisms and mutations and their expression, specificity, and evolution. The IMGT® structural approach refers to the study of 2D and 3D structures of the IG, TR, MHC, and RPI, and to the antigen or ligand binding characteristics in relation to protein functions, polymorphisms, and evolution. IMGT-Choreography, based on the Web service architecture paradigm, will enable significant biological and clinical requests addressing the entire IMGT® information system.

1.3 IMGT-ONTOLOGY Concepts and IMGT® Components for Genomics

1.3.1 IMGT® Genome Database

The IMGT® genomic approach refers to the study of the genes within their loci and on the chromosomes. Genomic data are managed in IMGT/GENE-DB, which is the comprehensive IMGT® genome database, created by LIGM, Montpellier, France, on the Web since January 2003 (Giudicelli, Chaume, and Lefranc 2005).

In March 2007, IMGT/GENE-DB contained 1512 genes and 2461 alleles (673 IG and TR genes and 1215 alleles from *Homo sapiens*, and 839 IG and TR genes and 1,246 alleles from *Mus musculus, Mus cookii, Mus pahari, Mus spretus, Mus saxicola, Mus minutoïdes*). All human and mouse IG and TR genes are available in IMGT/GENE-DB. Based on the IMGT® CLASSIFICATION concept, all the human IMGT® gene names (Lefranc and Lefranc 2001a; Lefranc and Lefranc 2001b; Lefranc 2000a; Lefranc 2000b; Lefranc 2000c; Lefranc 2000d) were approved by the HUman Genome Organisation (HUGO) Nomenclature Committee HGNC in 1999 (Wain, Bruford, Lovering, Lush, Wright, and Povey 2002), and entered in IMGT/GENE-DB (Giudicelli et al. 2005), Genome DataBase GDB, Canada (Letovsky, Cottingham, Porter, and Li 1998), LocusLink at NCBI, USA (Pruitt and Maglott 2001), and GeneCards (Safran, Chalifa-Caspi, Shmueli, Olender, Lapidot, Rosen, Shmoish, Peter, Glusman, Feldmesser, Adato, Peter, Khen, Atarot, Groner, and Lancet 2003). Reciprocal links exist between IMGT/GENE-DB and the generalist nomenclature (HGNC) and genome databases (GDB, Entrez Gene at NCBI, and GeneCards). The mouse IG and TR gene names (Martinez and Lefranc 1998; Bosc and Lefranc 2000; Bosc, Contet, and Lefranc 2001; Martinez, Folch, and Lefranc 2001; Bosc and Lefranc 2003) with IMGT reference sequences were provided by IMGT® to HGNC and to the Mouse Genome Database MGD (Blake, Richardson, Bult, Kadin, Eppig, and Mouse Genome Database Group 2003) in July 2002. Queries in IMGT/GENE-DB can be performed according to IG and TR gene classification criteria, and IMGT reference sequences have been defined for each allele of each gene based on one or, whenever possible, several of the following criteria: germline sequence, first sequence published, longest sequence, mapped sequence (Lefranc et al. 1999). IMGT/GENE-DB interacts dynamically with IMGT/LIGM-DB (Giudicelli, Ginestoux, Folch, Jabado-Michaloud, Chaume, and Lefranc 2006) to download and display gene-related sequence data. This is the first example of an interaction between IMGT® databases using the CLASSIFICATION concept.

1.3.2 IMGT® Genome Analysis Tools

The IMGT® genome analysis on-line tools comprise IMGT/LocusView, IMGT/GeneView, IMGT/GeneSearch, IMGT/CloneSearch, IMGT/GeneInfo, and IMGT/GeneFrequency. IMGT/LocusView and IMGT/GeneView manage the locus organization and the gene location and provide display of physical maps for the human IG, TR, and MHC loci and for the mouse TRA/TRD locus. IMGT/LocusView allows

users to view genes at their loci and then zoom in on a selected area. IMGT/GeneView allows users to directly view a given gene at a locus. IMGT/GeneSearch and IMGT/CloneSearch allow retrieval of information concerning genes and clones, respectively, analysed in IMGT/LocusView. IMGT/GeneSearch allows searching for genes at a locus, based on IMGT gene names, functionality, or chromosomal localization. IMGT/CloneSearch provides information on the clones that were used to build the locus contigs displayed in IMGT/LocusView (accession numbers are from IMGT/LIGM-DB, gene names from IMGT/GENE-DB, and clone position and orientation, and overlapping clones from IMGT/LocusView). IMGT/GeneInfo provides information on potential human and mouse TR rearrangements (Baum, Hierle, Pascal, Bellahcene, Chaume, Lefranc, Jouvin-Marche, Marche, and Demongeot 2006). IMGT/GeneFrequency is an IMGT interactive tool that dynamically computes histograms which represent the contribution of individual V, D, and J genes in sets of expressed IG and TR rearranged V-(D)-J sequences in IMGT/LIGM-DB. IMGT/GeneFrequency results are obtained by querying IMGT/LIGM-DB for sequences which are selected, for example, on the specificity criteria.

1.3.3 IMGT® Genome Web Resources

The IMGT® genomic Web resources are compiled in the IMGT Repertoire "Locus and genes" section that includes Chromosomal localizations, Locus representations, Locus description, Gene exon/intron organization, Gene exon/intron splicing sites, Gene tables, Potential germline repertoires, the complete lists of human and mouse IG and TR genes, and the correspondences between nomenclatures (Lefranc and Lefranc 2001a; Lefranc and Lefranc 2001b). The IMGT Repertoire "Probes and RFLP" section provides data on probes used in Southern analysis and on gene insertion/deletion polymorphisms (Osipova, Posukh, Wiebe, Miyazaki, Matsumoto, Lefranc, and Lefranc 1999; Dard, Lefranc, Osipova, and Sanchez-Mazas 2001; Elemento, Gascuel, and Lefranc 2002; Lefranc and Lefranc 2004).

1.4 IMGT-ONTOLOGY Concepts and IMGT® Components for Genetics

1.4.1 IMGT® Sequence Databases

The IMGT® genetic approach refers to the study of genes in relation to their polymorphisms, mutations, expression, specificity, and evolution. The IMGT® genetics approach heavily relies on the DESCRIPTION concept (and particularly on the V-REGION, D-REGION, J-REGION, and C-REGION core concepts for the IG and TR), on the CLASSIFICATION concept (gene and allele concepts), and on the NUMEROTATION concept (IMGT unique numbering) (Lefranc 1997; Lefranc 1999; Ruiz and Lefranc 2002; Duprat and Lefranc 2003; Lefranc, Pommié, Ruiz, Giudicelli, Foulquier, Truong, Thouvenin-Contet, and Lefranc 2003; Lefranc, Pommié, Kaas,

Duprat, Bosc, Guiraudou, Jean, Ruiz, Da Piedade, Rouard, Foulquier, Thouvenin, and Lefranc 2005b; Lefranc, Duprat, Kaas, Tranne, Thiriot, and Lefranc 2005c).

1.4.1.1 IMGT/LIGM-DB

IMGT/LIGM-DB is the comprehensive IMGT® database of IG and TR nucleotide sequences from human and other vertebrate species, with translation for fully annotated sequences, created in 1989 by LIGM, Montpellier, France, and available on the Web since July 1995 (Lefranc, Giudicelli, Busin, Malik, Mougenot, Déhais, and Chaume 1995; Giudicelli et al. 2006). In March 2007, IMGT/LIGM-DB contained 107,737 sequences of 150 vertebrate species. The unique source of data for IMGT/LIGM-DB is EMBL (Kulikova, Aldebert, Althorpe, Baker, Bates, Browne, van den Broek, Cochrane, Duggan, Eberhardt, Faruque, Garcia-Pastor, Harte, Kanz, Leinonen, Lin, Lombard, Lopez, Mancuso, McHale, Nardone, Silventoinen, Stoehr, Stoesser, Tuli, Tzouvara, Vaughan, Wu, Zhu, and Apweiler 2004) which shares data with the other two generalist databases GenBank and DNA Data Bank of Japan (DDBJ). Based on expert analysis, specific detailed annotations are added to IMGT® flat files. The Web interface allows searches according to specific immunogenetic criteria and is easy to use without any programming language knowledge. Selection is displayed at the top of the resulting sequences pages, so the users can check their own queries. Users have the possibility to modify their request or consult the results with a choice of nine possibilities (Lefranc 2003; Lefranc et al. 2004b). IMGT/LIGM-DB gene and allele name assignment and sequence annotations are performed according to the IMGT Scientific chart rules. These annotations allow retrieval of data from IMGT/LIGM-DB for queries in other IMGT® databases or tools. As an example, the IMGT/LIGM-DB accession numbers of the cDNA expressed sequences for each human and mouse IG and TR gene are available, with direct links to IMGT/LIGM-DB, in the IMGT/GENE-DB entries. IMGT/LIGM-DB data are also distributed by anonymous FTP servers at CINES (ftp://ftp.cines.fr/IMGT/) and EBI (ftp://ftp.ebi.ac.uk/pub/databases/imgt/) and from many Sequence Retrieval System (SRS) sites (EBI Hinxton UK, Institut Pasteur Paris, DKFZ Heidelberg Germany, Columbia University New York USA, IUBio Indiana University USA, DDBJ Japan, etc.). IMGT/LIGM-DB can be searched by BLAST or FASTA on different servers (EBI Hinxton UK, Institut Pasteur Paris).

1.4.1.2 IMGT/Automat for IMGT/LIGM-DB Annotations

IMGT/Automat (Giudicelli, Protat, and Lefranc 2003) is an integrated internal IMGT® Java tool which automatically performs the annotation of rearranged cDNA sequences that represent half of the IMGT/LIGM-DB content. The annotation procedure includes the IDENTIFICATION of the sequences, the CLASSIFICATION of the IG and TR genes and alleles, and the DESCRIPTION of all IG and TR specific and constitutive motifs within the nucleotide sequences. Accuracy and reliability of the annotation are mainly estimated by the program itself with the evaluation of the alignment scores, the deduced sequence functionality, and the coherence of the characterized and delimited IG and TR motifs. So far 9890 human and mouse IG and TR cDNA sequences have

been automatically annotated by the IMGT/Automat tool, with annotations being as reliable and accurate as those provided by a human annotator.

1.4.1.3 Other IMGT® IG and TR Sequence Databases

IMGT/PRIMER-DB (Folch, Bertrand, Lemaitre, and Lefranc 2004) is the IMGT® oligonucleotide primer database for IG and TR, created by LIGM, Montpellier in collaboration with EUROGENTEC S.A., Belgium, on the Web since February 2002 (http://www3.oup.co.uk/nar/database/summary/505). In March 2007, IMGT/PRIMER-DB contained 1864 entries and provides standardized information on oligonucleotides (or Primers) and combinations of primers (Sets, Couples) for IG and TR. These primers are useful for combinatorial library constructions, scFv, phage display, or microarray technologies. The IMGT Primer cards are linked to the IMGT/LIGM-DB flat files, and to the IMGT Repertoire (IMGT Colliers de Perles and Alignments of alleles of the IMGT/LIGM-DB reference sequence used for the primer description). IMGT/PROTEIN-DB is a new IMGT® database related to immunoglobulin and T-cell receptor amino acid sequences. The database will be available on the IMGT® Web site in 2007.

1.4.1.4 IMGT® MHC Sequence Databases

IMGT/MHC-DB comprises databases hosted at EBI and includes a database of human MHC allele sequences, IMGT/MHC-HLA (or IMGT/HLA), developed by Cancer Research, UK and maintained by the Anthony Nolan Research Institute ANRI, London, UK, on the Web since December 1998, and a database of MHC sequences from nonhuman primates IMGT/MHC-NHP, curated by the Biomedical Primate Research Centre BPRC, The Netherlands, on the Web since April 2002 (Robinson, Waller, Parham, de Groot, Bontrop, Kennedy, Stoehr, and Marsh 2003).

1.4.2 IMGT® Sequence Analysis Tools

The IMGT® tools for the genetics approach comprise IMGT/V-QUEST (Lefranc 2003; Giudicelli, Chaume, and Lefranc 2004), for the identification of the V, D, and J genes and of their mutations, IMGT/JunctionAnalysis (Yousfi Monod, Giudicelli, Chaume, and Lefranc 2004) for the analysis of the V-J and V-D-J junctions which confer the antigen receptor specificity, IMGT/Allele-Align for the detection of polymorphisms, IMGT/Phylogene (Elemento and Lefranc 2003) for gene evolution analyses, and IMGT/DomainDisplay for amino acid sequences.

1.4.2.1 IMGT/V-QUEST

IMGT/V-QUEST (V-QUEry and STandardization) is an integrated software for IG and TR (Lefranc 2003, Giudicelli et al. 2004), used for the identification of the V, D, and J genes and of their mutations. This tool is easy to use for the analysis of input IG or TR germline or rearranged variable nucleotide sequences. IMGT/V-QUEST results comprise the identification of the V, D, and J genes and alleles and the nucleotide

alignments by comparison with sequences from the IMGT reference directory, the FR-IMGT and CDR-IMGT delimitations based on the IMGT unique numbering, the translation of the input sequences, the display of nucleotide and amino acid mutations compared to the closest IMGT reference sequences, the identification of the JUNCTION and results from IMGT/JunctionAnalysis (default option), and the V-REGION IMGT Colliers de Perles. IMGT/V-QUEST provides a synthetic view of the results when several sequences (up to 50) are analysed in the same run.

1.4.2.2 IMGT/JunctionAnalysis

IMGT/JunctionAnalysis (Yousfi Monod et al. 2004) is a tool, complementary to IMGT/V-QUEST, which provides a thorough analysis of the V-J and V-D-J junctions which confer the antigen receptor specificity to IG and TR rearranged genes. IMGT/JunctionAnalysis identifies the D-GENEs and alleles involved in the IGH, TRB, and TRD V-D-J rearrangements by comparison with the IMGT reference directory, and delimits precisely the P, N, and D regions. Several hundred junction sequences can be analysed simultaneously.

1.4.2.3 IMGT/Allele-Align

IMGT/Allele-Align is used for the detection of polymorphisms. It allows the comparison of two alleles highlighting the nucleotide and amino acid differences.

1.4.2.4 IMGT/PhyloGene

IMGT/PhyloGene (Elemento and Lefranc 2003) is an easy tool for phylogenetic analysis of IG and TR variable region (V-REGION) and constant domain (C-DOMAIN) sequences. This tool is particularly useful in developmental and comparative immunology. The users can analyse their own sequences by comparison with the IMGT standardized reference sequences for human and mouse IG and TR.

1.4.2.5 IMGT/DomainDisplay

IMGT/DomainDisplay provides a display of amino acid sequences per domain (V, C, or G type domain) for IG, TR, MHC and for RPI (that include IgSF proteins other than IG and TR, and MhcSF proteins other than MHC), based on the IMGT unique numbering (Lefranc et al. 2003; Lefranc et al. 2005b; Lefranc et al. 2005c).

1.4.3 IMGT® Genetics Web Resources

The IMGT® genetic Web resources are compiled in the IMGT Repertoire "Proteins and alleles" section which includes Protein displays, Alignments of alleles, Tables of alleles, Allotypes, and Isotypes (Osipova et al. 1999; Dard et al. 2001; Lefranc and Lefranc 2004).

1.5 IMGT-ONTOLOGY Concepts and IMGT® Components for 2D and 3D Structures

1.5.1 IMGT® Structural Database

The IMGT® structural approach refers to the study of the 2D and 3D structures of the IG, TR, MHC, and RPI, and to the antigen or ligand binding characteristics in relation to the protein functions, polymorphisms, and evolution. The structural approach relies on the CLASSIFICATION concept (IMGT gene and allele names), DESCRIPTION concept (receptor and chain description, domain delimitations), and NUMEROTATION concept (amino acid positions according to the IMGT unique numbering) (Lefranc et al. 2003; Lefranc et al. 2005b; Lefranc et al. 2005c). Structural and functional domains of the IG and TR chains comprise the variable domain or V-DOMAIN (nine-strand β-sandwich) that corresponds to the V-J-REGION or V-D-J-REGION and is encoded by two or three genes (Lefranc and Lefranc 2001a; Lefranc and Lefranc 2001b; Lefranc et al. 2003), the constant domain or C-DOMAIN (seven-strand β-sandwich) (Lefranc et al. 2005b), and, for the MHC chains, the groove domain or G-DOMAIN (four β-strand and one α-helix) (Lefranc et al. 2005c). The IMGT unique numbering has been extended to the V-LIKE-DOMAINs (Lefranc et al. 2003) and C-LIKE-DOMAINs (Lefranc et al. 2005b) of IgSF proteins other than IG and TR, and to the G-LIKE-DOMAINs (Lefranc et al. 2005c) of MhcSF proteins other than MHC.

Structural data are compiled and annotated in IMGT/3Dstructure-DB. IMGT/3Dstructure-DB is the IMGT® 3D structure database for IG, TR, MHC, and RPI, created by LIGM, on the Web since November 2001 (Kaas, Ruiz, and Lefranc 2004). In March 2007, IMGT/3Dstructure-DB contained 1221 atomic coordinate files. IMGT/3Dstructure-DB comprises IG, TR, MHC, and RPI with known 3D structures. Coordinate files extracted from the Protein Data Bank PDB (Berman, Westbrook, Feng, Gilliland, Bhat, Weissig, Shindyalov, and Bourne 2000) (http://www.rcsb.org/pdb/) are renumbered according to the standardized IMGT unique numbering (Lefranc et al. 2003; Lefranc et al. 2005b; Lefranc et al. 2005c). The IMGT/3Dstructure-DB cards provide IMGT annotations (assignment of IMGT genes and alleles, IMGT chain and domain labels, IMGT Colliers de Perles for V, C, and G type domains (Ruiz and Lefranc 2002; Kaas and Lefranc 2005; Kaas and Lefranc 2007), downloadable renumbered IMGT/3Dstructure-DB flat files, visualization tools, and external links. The IMGT/3Dstructure-DB residue cards provide detailed information on the inter- and intra-domain contacts of each residue position. An IMGT/3Dstructure-DB card provides receptor and chain description, IMGT gene and allele names, domain delimitations, and amino acid positions according to the IMGT unique numbering. Standardized IMGT pMHC contact sites have been defined for peptide/MHC complexes (Kaas and Lefranc 2005).

1.5.2 IMGT® Structure Analysis Tools

Several on-line IMGT® structure analysis tools are available for the analysis of 2D and 3D structures, and particularly for the comparison of V, C, and G domains. The IMGT/StructuralQuery tool (Kaas et al. 2004) analyses the interactions of the residues of the antigen receptors (IG and TR), MHC, RPI, antigens and ligands. The contacts are described per domain (intra- and inter-domain contacts) and annotated in term of IMGT labels (chains, domains), positions (IMGT unique numbering) with backbone or side-chain implication. IMGT/StructuralQuery allows users to retrieve the IMGT/3Dstructure-DB entries, based on specific structural characteristics: ϕ (phi) and ψ (psi) angles, accessible surface area (ASA), amino acid type, distance in angstroms between amino acids, CDR-IMGT lengths. IMGT/StructuralQuery is currently available for the V-DOMAINs. IMGT/DomainGapAlign aligns users' amino acid sequences against the closest reference sequences from the IMGT domain sequence directory. IMGT/DomainGapAlign also provides the IMGT gaps and thus allows users to graphically represent their domain sequences with the IMGT/Collier-de-Perles tool. IMGT/DomainSuperimpose allows superimposing two 3D structures of domains from IMGT/3Dstructure-DB.

1.5.3 IMGT® Structural Web Resources

The IMGT® structural Web resources are compiled in the IMGT Repertoire "2D and 3D structures" section which includes 2D representations or IMGT Colliers de Perles (Lefranc et al. 2003; Lefranc et al. 2005b; Lefranc et al. 2005c; Ruiz and Lefranc 2002; Kaas and Lefranc 2007), 3D representations, FR-IMGT and CDR-IMGT lengths, and amino acid physico-chemical characteristic profiles (Pommié, Sabatier, Lefranc, and Lefranc 2004).

In order to appropriately analyse the amino acid resemblances and differences between IG, TR, MHC, and RPI chains, 11 IMGT classes were defined for the "chemical characteristics" amino acid properties and used to set up IMGT Colliers de Perles reference profiles (Pommié et al. 2004). The IMGT Colliers de Perles reference profiles allow one to easily compare amino acid properties at each position whatever the domain, the chain, the receptor, or the species. The IG and TR variable and constant domains represent a privileged situation for the analysis of amino acid properties in relation to 3D structures, by the conservation of their 3D structure despite divergent amino acid sequences, and by the considerable amount of genomic (IMGT Repertoire), structural (IMGT/3Dstructure-DB), and functional data available. These data are not only useful to study mutations and allele polymorphisms, but are also needed to establish correlations between amino acids in the protein sequences and 3D structures and to determine amino acids potentially involved in the immunogenicity.

1.6 IMGT-Choreography

The goal of IMGT-Choreography is to orchestrate dynamic procedure calls between IMGT® databases querying and analysis tools, using IMGT's biological approaches (Lefranc et al. 2004a). Major existing or potential "conversation nodes" can be identified between IMGT® tools, by an analysis of their profiles (IMGT tool diamonds; Lefranc et al. 2004a). IMGT-Choreography is based on the Web service architecture paradigm (see W3C; http://www.w3.org/). Conversations between Web services are expressed using the sole IMGT-ML language both for queries and for result fetches.

1.6.1 IMGT Tool Diamonds

In order to enhance the interoperability between the IMGT® components, IMGT® tools were analysed for input and output parameters, performed tasks, and accompanying databases (IMGT reference directories). Graphical diamond-shaped representations, designated as "IMGT tool diamonds" (Lefranc et al. 2004a), were developed to obtain tool profiles and to compare the state of the art of each tool in relation to the IMGT ontological concepts. Each IMGT tool diamond is composed of modules that correspond to different IMGT-ONTOLOGY concepts and each module comprises four facets: input parameters, task, IMGT reference directory, and output parameters (Lefranc et al. 2004a).

1.6.2 IMGT-ML

IMGT-ML (Chaume et al. 2001; Chaume et al. 2003) (available at IMGT Index>IMGT-ML, http://imgt.cines.fr) represents the specification of the main IMGT-ONTOLOGY concepts (Giudicelli and Lefranc 1999), formalized through a markup language defined in-house, based on Extensible Markup Language (XML) (http://www.w3.org/XML/) and constrained through XML Schema (http://www.w3.org/XML/Schema). Messages that are exchanged between service providers and consumers are encoded using valid IMGT-ML streams. IMGT-ML can be seen as a kind of Rosetta stone since it extends the ease of interconnection between IMGT Web services and is the unique language used for both services inputs and outputs. This ensures semantic consistency between exchanged messages as IMGT-ML is an XML schema formalization of the IMGT-ONTOLOGY concepts (Chaume et al. 2003).

1.6.3 IMGT® Web Services

Web services have been chosen as the means to create dynamic interactions between IMGT® databases and tools. Clients and providers for these services can be written using any SOAP-capable programming language such as SOAP::lite (http://www.soaplite.com/), development library for Perl or webMethods Glue for JAVA, thus facilitating the conversion of legacy applications to services. IMGT Web services are developed using the JAVA programming language and deployed using the Apache Axis (http://ws.apache.org/axis/) Web services development framework.

The IMGT/LIGM-DB Web service is the first Web service currently developed and implemented with Axis (Lefranc et al. 2004a). It includes the "queryKnowledge" and "querySeqData" services. The queryKnowledge service provides the lists of instances for the IMGT-ONTOLOGY concepts, for example the list of chain types, functionalities, specificities defined in the IDENTIFICATION concept, the lists of groups and subgroups defined in the CLASSIFICATION concept, or the list of labels defined in the DESCRIPTION concept. The querySeqData service allows the retrieval of any sequence-related data that are identified, classified, and described in IMGT/LIGM-DB according to the IMGT concepts. The querySeqData input has the form of an incomplete IMGT-ML data entry in which the given values are used as criteria to query IMGT/LIGM-DB. The result is a list of data entries, in IMGT-ML format, sharing these given values. Other Web services are developed to automatically query IMGT® databases and tools.

1.6.4 Perspectives

Composition and chaining of IMGT® Web services through IMGT-Choreography will enable processing of complex significant biological and clinical requests involving every part of the IMGT® information system. IMGT-Choreography has for goal to combine and join the IMGT® database queries and analysis tools.

In order to keep only significant approaches, a rigorous analysis of the scientific standards of the biologist research (Giudicelli and Lefranc 1999; Lefranc and Lefranc 2001a; Lefranc and Lefranc 2001b; Osipova et al. 1999; Dard et al. 2001; Chardes, Chapal, Bresson, Bes, Giudicelli, Lefranc, and Peraldi-Roux 2002;; Chassagne, Laffly, Drouet, Herodin, Lefranc, and Thullier 2004; Bertrand, Duprat, Lefranc, Marti, and Coste 2004) and of the clinician's needs (Ghia, Stamatopoulos, Belessi, Moreno, Stella, Giuda, Michel, Crespo, Laoutaris, Montserrat, Anagnostopoulos, Dighiero, Fassas, Caligaris-Cappio, and Davi 2005; Stamatopoulos, Belessi, Papadaki, Kalagiakou, Stavroyianni, Douka, Afendaki, Saloum, Parasi, Anagnostou, Laoutaris, Fassas, and Anagnostopoulos 2004) has been undertaken in parallel with the modelling of interactions between the IMGT® components (databases, tools, and Web resources). To increase interoperability with other biological information systems and ontologies, IMGT-ONTOLOGY is currently being implemented with Protégé (http://protege.stanford.edu/) (Noy, Fergerson, and Musen 2000).

1.7 Conclusions

Since July 1995, IMGT® has been available on the Web at the IMGT® Home page http://imgt.cines.fr (Montpellier, France). IMGT® has an exceptional response with more than 140,000 requests a month. IMGT® is the international reference in immunogenetics and immunoinformatics and provides a common access to standardized data which include nucleotide and protein sequences, oligonucleotide primers, gene maps, genetic polymorphisms, specificities, 2D and 3D structures, based on IMGT-ONTOLOGY. Although the IMGT® genome, sequence, and 3D structure databases, IMGT® analysis tools, and IMGT Repertoire Web resources, were initially imple-

mented for the IG, TR, and MHC of human and other vertebrates, data and knowledge management standardization has now been extended to the IgSF proteins other than IG or TR (Williams and Barclay 1988) and to the MhcSF proteins other than MHC (Maenaka and Jones 1999), of any species (IMGT Repertoire (RPI)). Thus, standardization in IMGT® contributed to data enhancement of the system and new expertised data concepts were readily incorporated.

The IMGT® information is of much value to clinicians and biological scientists in general. IMGT® databases and tools are extensively queried and used by scientists, from both academic and industrial laboratories, who are equally distributed between the United States, Europe, and the rest of the world. IMGT® is used in very diverse domains: (i) fundamental research and medical research (repertoire analysis of the IG antibody sites and of the TR recognition sites in normal and pathological situations such as autoimmune diseases, infectious diseases, AIDS, leukemias, lymphomas, myelomas), (ii) veterinary research (IG and TR repertoires in farm and wildlife species), (iii) genome diversity and genome evolution studies of the adaptive immune responses, (iv) structural evolution of the IgSF and MhcSF proteins, (v) biotechnology related to antibody engineering (single chain Fragment variable (scFv), phage displays, combinatorial libraries, chimeric, humanized, and human antibodies), (vi) diagnostics (clonalities, detection and follow-up of residual diseases), and (vii) therapeutical approaches (grafts, immunotherapy, vaccinology). By its high quality and its data distribution based on IMGT-ONTOLOGY, IMGT® has an important role to play in the development of immunogenetics Web services. The design of IMGT-Choreography and the creation of dynamic interactions between the IMGT® databases and tools, using Web services and IMGT-ML, represent novel and major developments of IMGT®, the international reference in immunogenetics and immunoinformatics.

1.8 Citing IMGT®

Users are requested to cite this article, and to quote the IMGT® home page URL, http://imgt.cines.fr. Individual IMGT® databases, tools, and Web resources should also be quoted where relevant: IMGT/GENE-DB (Giudicelli et al. 2005), IMGT/GeneInfo (Baum et al. 2006), IMGT/LIGM-DB (Giudicelli et al. 2006), IMGT/MHC-DB (Robinson et al. 2003), IMGT/PRIMER-DB (Folch et al. 2004), IMGT/V-QUEST (Giudicelli et al. 2004), IMGT/JunctionAnalysis (Yousfi Monod et al. 2004), IMGT/PhyloGene (Elemento and Lefranc 2003), IMGT/3Dstructure-DB (Kaas et al. 2004), IMGT/StructuralQuery (Kaas et al. 2004), and IMGT/Collier-de-Perles (Kaas and Lefranc 2005; Kaas and Lefranc 2007).

Acknowledgements

I am very grateful to the IMGT® team for its expertise and its constant motivation. I thank Véronique Giudicelli, Joumana Jabado-Michaloud, Chantal Ginestoux, Géraldine Folch, Patrice Duroux, François Ehrenmann, Xavier Brochet, and Gérard Lefranc for helpful discussion. IMGT® is a registered Centre National de la Recher-

che Scientifique (CNRS) mark. IMGT® has been a National Bioinformatics Platform RIO (CNRS, INSERM, CEA, INRA) since 2001. IMGT® was funded in part by the BIOMED1 (BIOCT930038), Biotechnology BIOTECH2 (BIO4CT960037), and 5th PCRDT Quality of Life and Management of Living Resources (QLG2-2000-01287) programs of the European Union and received subventions from the Association pour la Recherche sur le Cancer (ARC) and from the Génopole-Montpellier-Languedoc-Roussillon. IMGT® is currently supported by the CNRS, the Ministère de l'Education Nationale, de l'Enseignement Supérieur et de la Recherche (MENESR), Université Montpellier II Plan Pluri-Formation, the Réseau National des Génopoles (RNG), BIOSTIC-LR2004, ACI-IMPBIO IMP82-2004, GIS AGENAE, Région Languedoc-Roussillon, Agence Nationale de la Recherche ANR (BIOSYS06_ 135457), and the ImmunoGrid project (IST-2004-028069) of the 6th framework program of the European Union.

References

Baum, T.P., Hierle, V., Pascal, N., Bellahcene, F., Chaume, D., Lefranc, M.-P., Jouvin-Marche, E., Marche, P.N., and Demongeot, J. (2006) IMGT/GeneInfo: T cell receptor gamma TRG and delta TRD genes in database give access to all TR potential V(D)J recombinations. BMC Bioinformatics 7:224.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000) The Protein Data Bank. Nucleic Acids Res. 28:235-242.

Bertrand, G., Duprat, E., Lefranc, M.-P., Marti, J., and Coste, J. (2004) Characterization of human FCGR3B*02 (HNA-1b, NA2) cDNAs and IMGT standardized description of FCGR3B alleles. Tissue Antigens 64:119-131.

Blake, J.A., Richardson, J.E., Bult, C.J., Kadin, J.A., Eppig, J.T., and Mouse Genome Database Group (2003) MGD: The Mouse Genome Database. Nucleic Acids Res. 31:193-195.

Bosc, N., and Lefranc, M.-P. (2000) The Mouse (*Mus musculus*) T cell receptor β variable (TRBV), diversity (TRBD) and joining (TRBJ) genes. Exp. Clin. Immunogenet. 17:216-228.

Bosc, N., Contet, V. and Lefranc, M.-P. (2001) The mouse (*Mus musculus*) T cell receptor delta variable (TRDV), diversity (TRDD) and joining (TRDJ) genes. Exp. Clin. Immunogenet. 18:51-58.

Bosc, N., and Lefranc, M.-P. (2003) IMGT Locus in focus: The mouse (*Mus musculus*) T cell receptor α (TRA) and delta (TRD) variable genes. Dev. Comp. Immunol. 27:465-497.

Chardes, T., Chapal, N., Bresson, D., Bes, C., Giudicelli, V., Lefranc, M.-P., and Peraldi-Roux, S. (2002) The human anti-thyroid peroxidase autoantibody repertoire in Graves' and Hashimoto's autoimmune thyroid diseases. Immunogenetics 54:141-157.

Chassagne, S., Laffly, E., Drouet, E., Herodin, F., Lefranc, M.-P., and Thullier, P. (2004) A high affinity macaque antibody Fab with human-like framework regions obtained from a small phage display immune library. Mol. Immunol. 41:539-546.

Chaume, D., Giudicelli, V., and Lefranc, M.-P. (2001) IMGT-ML a language for IMGT-ONTOLOGY and IMGT/LIGM-DB data. In: CORBA and XML: Towards a bioinformatics integrated network environment. Proceedings of NETTAB 2001, Network Tools and Applications in Biology, pp. 71-75.

Chaume, D., Giudicelli, V., Combres, K., and Lefranc, M.-P. (2003) IMGT-ONTOLOGY and IMGT-ML for Immunogenetics and immunoinformatics. In: *Abstract book of the*

Sequence databases and Ontologies satellite event. European Congress in Computational Biology ECCB'2003, pp. 22-23.

Chaume, D., Giudicelli, V., Combres, K., Ginestoux, C., and Lefranc, M.-P. (2005) IMGT-Choreography: Processing of complex immunogenetics knowledge. In: *Computational Methods in Systems Biology CMSB 2004. Lecture Notes in Bioinformatics LNBI,* Springer, pp. 73-84.

Dard, P., Lefranc, M.-P., Osipova, L., and Sanchez-Mazas, A. (2001) DNA sequence variability of IGHG3 genes associated to the main G3m haplotypes in human populations. Eur. J. Hum. Genet. 9:765-772.

Duprat, E., and Lefranc, M.-P. (2003) IMGT standardization and analysis of V-LIKE-, C-LIKE- and G-LIKE-DOMAINs. In: *Proceedings of the European Conference on Computational Biology ECCB''2003,* PS-32, pp. 223-224.

Elemento, O., Gascuel, O., and Lefranc, M.-P. (2002) Reconstructing the duplication history of tandemly repeated genes. Mol. Biol. Evol. 19:278-288.

Elemento, O., and Lefranc, M.-P. (2003) IMGT/PhyloGene: An on-line tool for comparative analysis of immunoglobulin and T cell receptor genes. Dev. Comp. Immunol. 27 :763-779.

Folch, G., Bertrand, J., Lemaitre, M., and Lefranc, M.-P. (2004) IMGT/PRIMER-DB. In: M.Y. Galperin (Ed.), The Molecular Biology Database Collection: 2004 update. Nucleic Acids Res. 32:3-22.

Ghia, P., Stamatopoulos, K., Belessi, C., Moreno, C., Stella, S., Giuda, G., Michel, A., Crespo, M., Laoutaris, N., Montserrat, E., Anagnostopoulos, A., Dighiero, G., Fassas, A., Caligaris-Cappio, F., and Davi, F. (2005) Geographical patterns and pathogenetic implications of IGHV gene usage in chronic lymphocytic leukemia: The lesson of the IGHV3-21 gene. Blood 105:1678-1685.

Giudicelli, V., and Lefranc, M.-P. (1999) Ontology for Immunogenetics: The IMGT-ONTOLOGY. Bioinformatics 12:1047-1054.

Giudicelli, V., Protat, C., and Lefranc, M.-P. (2003) The IMGT strategy for the automatic annotation of IG and TR cDNA sequences: IMGT/Automat. In: *Proceedings of the European Conference on Computational Biology ECCB'2003,* DKB-31, pp. 103-104.

Giudicelli, V., Chaume, D., and Lefranc, M.-P. (2004) IMGT/V-QUEST, an integrated software program for immunoglobulin and T cell receptor V-J and V-D-J rearrangement analysis. Nucleic Acids Res. 32:W435-W440.

Giudicelli, V., Chaume, D., and Lefranc, M.-P. (2005) IMGT/GENE-DB: A comprehensive database for human and mouse immunoglobulin and T cell receptor genes. Nucleic Acids Res. 33:D256–D261.

Giudicelli, V., Ginestoux, C., Folch, G., Jabado-Michaloud, J., Chaume, D., and Lefranc, M.-P. (2006) IMGT/LIGM-DB, the IMGT® comprehensive database of immunoglobulin and T cell receptor nucleotide sequences. Nucleic Acids Res. 34**:**D781-D784.

Kaas, Q., Ruiz, M., and Lefranc, M.-P. (2004) IMGT/3Dstructure-DB and IMGT/StructuralQuery, a database and a tool for immunoglobulin, T cell receptor and MHC structural data. Nucleic Acids Res. 32:D208-D210.

Kaas, Q., and Lefranc, M.-P. (2005) T cell receptor/peptide/MHC molecular characterization and standardized pMHC contact sites in IMGT/3Dstructure-DB. *In Silico* Biol. 5:505-528.

Kaas, Q., and Lefranc, M.-P. (2007) IMGT Colliers de Perles: Standardized sequence-structure representations of the IgSF and MhcSF superfamily domains. Curr. Bioinformatics 2:21-30.

Kulikova, T., Aldebert, P., Althorpe, N., Baker, W., Bates, K., Browne, P., van den Broek, A., Cochrane, G., Duggan, K., Eberhardt, R., Faruque, N., Garcia-Pastor, M., Harte, N., Kanz, C., Leinonen, R., Lin, Q., Lombard, V., Lopez, R., Mancuso, R., McHale, M., Nardone, F., Silventoinen, V., Stoehr, P., Stoesser, G., Tuli, M.A., Tzouvara, K., Vaughan, R., Wu, D., Zhu, W., and Apweiler, R. (2004) The EMBL Nucleotide Sequence Database. Nucleic Acids Res. 32:D27-D30.

Lefranc, M.-P., Giudicelli, V., Busin, C., Malik, A., Mougenot, I., Déhais, P., and Chaume, D. (1995) LIGM-DB/IMGT: An integrated database of Ig and TcR, part of the Immunogenetics database. Ann. N. Y. Acad. Sci. 764:47-49.

Lefranc, M.-P. (1997) Unique database numbering system for immunogenetic analysis. Immunol. Today 18:509.

Lefranc, M.-P. (1999) The IMGT unique numbering for immunoglobulins, T cell receptors and Ig-like domains. The Immunologist 7:132-136.

Lefranc, M.-P., Giudicelli, V., Ginestoux, C., Bodmer, J., Müller, W., Bontrop, R., Lemaitre, M., Malik, A., Barbié, V., and Chaume, D. (1999) IMGT, the international ImMunoGeneTics database. Nucleic Acids Res. 27:209-212.

Lefranc, M.-P. (2000a) Nomenclature of the human T cell receptor genes. In: *Current Protocols in Immunology,* John Wiley & Sons, New York, Supplement 40, A.1O.1-A.1O.23.

Lefranc, M.-P. (2000b) Nomenclature of the human immunoglobulin genes. In: *Current Protocols in Immunology,* John Wiley & Sons, New York, Supplement 40, A.1P.1-A.1P.37.

Lefranc, M.-P. (2000c) Locus maps and genomic repertoire of the human T cell receptor genes. The Immunologist 8:72-79.

Lefranc, M.-P. (2000d) Locus maps and genomic repertoire of the human immunoglobulin genes. The Immunologist 8:80-88.

Lefranc, M.-P., and Lefranc, G. (2001a) *The Immunoglobulin FactsBook.* Academic Press, London, 458.

Lefranc, M.-P., and Lefranc, G. (2001b) *The T cell receptor FactsBook.* Academic Press, London, 398.

Lefranc, M.-P. (2003) IMGT, the international ImMunoGeneTics information system® (http://imgt.cines.fr). In: B.K.C. Lo (Ed.), *Antibody Engineering: Methods and Protocols.* 2nd edition. *Methods in Molecular Biology.* Humana Press, USA, 248, pp. 27-49.

Lefranc, M.-P., Pommié, C., Ruiz, M., Giudicelli, V., Foulquier, E., Truong, L., Thouvenin-Contet, V., and Lefranc, G. (2003) IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. Dev. Comp. Immunol. 27:55-77.

Lefranc, M.-P., and Lefranc, G. (2004) Immunoglobulin lambda (IGL) genes of human and mouse. In: T. Honjo, F.W. Alt, and M.S. Neuberger (Eds.), *Molecular Biology of B Cells.* Academic Press, Elsevier Science, pp. 37-59.

Lefranc, M.-P., Clément, O., Kaas, Q., Duprat, E., Chastellan, P., Coelho, I., Combres, K., Ginestoux, C., Giudicelli, V., Chaume, D., and Lefranc, G. (2004a) IMGT-Choreography for Immunogenetics and Immunoinformatics. *In Silico* Biol. 5:6.

Lefranc, M.-P., Giudicelli, V., Ginestoux, C., Bosc, N., Folch, G., Guiraudou, D., Jabado-Michaloud, J., Magris, S., Scaviner, D., Thouvenin, V., Combres, K., Girod, D., Jeanjean, S., Protat, C., Monod, M.Y., Duprat, E., Kaas, Q., Pommié, C., Chaume, D., and Lefranc, G. (2004b) IMGT-ONTOLOGY for Immunogenetics and Immunoinformatics (http://imgt.cines.fr). *In Silico* Biol. 4:17-29.

Lefranc, M.-P., Giudicelli, V., Kaas, Q., Duprat, E., Jabado-Michaloud, J., Scaviner, D., Ginestoux, C., Clément, O., Chaume, D., and Lefranc, G. (2005a) IMGT, the international ImMunoGeneTics information system®. Nucleic Acids Res. 33**:**D593-D597.

Lefranc, M.-P., Pommié, C., Kaas, Q., Duprat, E., Bosc, N., Guiraudou, D., Jean, C., Ruiz, M., Da Piedade, I., Rouard, M., Foulquier, E., Thouvenin, V., and Lefranc, G. (2005b) IMGT unique numbering for immunoglobulin and T cell receptor constant domains and Ig superfamily C-like domains. Dev. Comp. Immunol. 29:185-203.

Lefranc, M.-P., Duprat, E., Kaas, Q., Tranne, M., Thiriot, A., and Lefranc, G. (2005c) IMGT unique numbering for MHC groove G-DOMAIN and MHC superfamily (MhcSF) G-LIKE-DOMAIN. Dev. Comp. Immunol. 29:917-938.

Letovsky, S.I., Cottingham, R.W., Porter, C.J., and Li, P.W. (1998) GDB: The Human Genome Database. Nucleic Acids Res. 26:94-99.

Maenaka, K., and Jones, E.Y. (1999) MHC superfamily structure and the immune system. Curr. Opin. Struct. Biol. 9:745-753.

Martinez, C., and Lefranc, M.-P. (1998) The mouse (*Mus musculus*) immunoglobulin kappa variable (IGKV) genes and joining (IGKJ) segments. Exp. Clin. Immunogenet. 15: 184-193.

Martinez, C., Folch, G., and Lefranc, M.-P. (2001) Nomenclature and overview of the mouse (*Mus musculus* and *Mus sp.*) immunoglobulin kappa (IGK) genes. Exp. Clin. Immunogenet. 18:255-279.

Noy, N.F., Fergerson, R.W., and Musen, M.A. (2000) The knowledge model of Protege-2000: Combining interoperability and flexibility. *2th International Conference on Knowledge Engineering and Knowledge Management (EKAW'2000), Juan-les-Pins, France.*

Osipova, L.P., Posukh, O.L., Wiebe, V.P., Miyazaki, T., Matsumoto, H., Lefranc, G., and Lefranc, M.-P. (1999) *Bam*HI-*Sac*I RFLP and Gm analysis of the immunoglobulin IGHG genes in Northern Selkups (West Siberia): New haplotypes with deletion, duplication and triplication. Hum. Genet. 105:530-541.

Pommié, C., Sabatier, S., Lefranc, G., and Lefranc, M.-P. (2004) IMGT standardized criteria for statistical analysis of immunoglobulin V-REGION amino acid properties. J. Mol. Recognit. 17:17-32.

Pruitt, K.D., and Maglott, D.R. (2001) RefSeq and LocusLink: NCBI gene-centered resources. Nucleic Acids Res. 29:137-140.

Robinson, J., Waller, M.J., Parham, P., de Groot, N., Bontrop, R., Kennedy, L.J., Stoehr, P., and Marsh, S.G. (2003) IMGT/HLA and IMGT/MHC sequence databases for the study of the major histocompatibility complex. Nucleic Acids Res. 31:311-314.

Ruiz, M., and Lefranc, M.-P. (2002) IMGT gene identification and Colliers de Perles of human immunoglobulins with known 3D structures. Immunogenetics 53:857-883.

Safran, M., Chalifa-Caspi, V., Shmueli, O., Olender, T., Lapidot, M., Rosen, N., Shmoish, M., Peter, Y., Glusman, G., Feldmesser, E., Adato, A., Peter, I., Khen, M., Atarot, T., Groner, Y., and Lancet, D. (2003) Human Gene-Centric Databases at the Weizmann Institute of Science: GeneCards, UDB, CroW 21 and HORDE. Nucleic Acids Res. 31:142-146.

Stamatopoulos, K., Belessi, C., Papadaki, T., Kalagiakou, E., Stavroyianni, N., Douka, V., Afendaki, S., Saloum, R., Parasi, A., Anagnostou, D., Laoutaris, N. Fassas, A., and Anagnostopoulos, A. (2004) Immunoglobulin heavy and light chain repertoire in splenic marginal zone lymphoma. Mol Med. 10:89-95.

Wain, H.M., Bruford, E.A., Lovering, R.C., Lush, M.J., Wright, M.W., and Povey, S. (2002) Guidelines for human gene nomenclature. Genomics 79:464-470.

Williams, A.F., and Barclay, A.N. (1988) The immunoglobulin family: Domains for cell surface recognition. Annu. Rev. Immunol. 6:381-440.

Yousfi Monod, M., Giudicelli, V., Chaume, D., and Lefranc, M.-P. (2004) IMGT/JunctionAnalysis: The first tool for the analysis of the immunoglobulin and T cell receptor complex V-J and V-D-J JUNCTIONs. Bioinformatics 20:I379-I385.

Chapter 2

IMGT Standardization for Molecular Characterization of the T Cell Receptor/Peptide/MHC Complexes

Quentin Kaas,[1] Elodie Duprat,[1] Guillaume Tourneur,[1] and Marie-Paule Lefranc[1,2]

[1] IMGT®, the International ImMunoGeneTics Information System®, Université Montpellier II, Laboratoire d'ImmunoGénétique Moléculaire LIGM, UPR CNRS 1142, Institut de Génétique Humaine IGH, 141 rue de la Cardonille, 34396 Montpellier Cedex 5, France

[2] Institut Universitaire de France, 103 Boulevard Saint-Michel, 75005 Paris, France, Marie-Paule.Lefranc@igh.cnrs.fr

Abstract. One of the key elements in the adaptive immune response is the presentation of peptides by the major histocompatibility complex (MHC) to the T cell receptors (TR) at the surface of T cells. The characterization of the TR/peptide/MHC trimolecular complexes (TR/pMHC) is crucial to the fields of immunology, vaccination, and immunotherapy. In order to facilitate data comparison and cross-referencing between experiments from different laboratories whatever the receptor, the chain type, the domain, or the species, IMGT®, the international ImMunoGeneTics information system (http://imgt.cines.fr), has developed IMGT-ONTOLOGY, the first ontology in immunogenetics and immunoinformatics. In IMGT/3Dstructure-DB, the IMGT three-dimensional structure database, the molecular characterization of the TR/pMHC is made according to the IMGT Scientific chart rules that are based on the IMGT-ONTOLOGY concepts. IMGT/3Dstructure-DB provides the standardized IMGT gene and allele names (CLASSIFICATION), the standardized IMGT labels (DESCRIPTION), and the IMGT unique numbering (NUMEROTATION). As the IMGT structural unit is the domain, amino acids at conserved positions always have the same number in the IMGT® databases, tools, and Web resources. For the TR α and β chains, the amino acids in contact with the peptide/MHC (pMHC) are defined according to the IMGT unique numbering for V-DOMAIN. The MHC chain cleft that binds the peptide is formed by two groove domains (G-DOMAIN), each one comprising four antiparallel β strands and one α helix. The IMGT unique numbering for G-DOMAIN applies both to the first two domains (G-ALPHA1 and G-ALPHA2) of the MHC class I α chain, and to the first domain (G-ALPHA and G-BETA) of the MHC class II α chain and β chain, respectively. Based on the IMGT unique numbering, we defined 11 contact sites for the analysis of the pMHC contacts. The TR/pMHC contact description, based on the IMGT numbering, can be queried in the IMGT/StucturalQuery tool, at http://imgt.cines.fr.

2.1 Introduction

T cells are implicated in the specific immune response against a stress of viral, bacterial, fungal, or tumoral origin. They identify antigenic peptides presented by the major histocompatibility complex (MHC) cell surface glycoproteins. The recognition is

carried out by the T cell receptor complex (TcR), a multisubunit transmembrane surface complex made up of a T cell receptor (TR) and of the CD3 chains, that is associated, in the immunological synapse, to the CD4 or CD8 coreceptors, to the CD28 and CTLA-4 costimulatory proteins, to the CD2 adhesion molecule, and to intracellular kinases (Lefranc and Lefranc 2001). The TR directly binds the peptide/MHC complex (pMHC), and activates the T cell through interactions with the CD3 and other components of the TcR (Vasmatzis, Cornette, Sezerman, and DeLisi 1996a; Sim, Zerva, Greene, and Gascoigne 1996; Kjer-Nielsen, Clements, Purcell, Brooks, Whisstock, Burrows, McCluskey, and Rossjohn 2003). Three-dimensional (3D) structures of the TR, pMHC, and TR/pMHC complexes provide an atomic description of their interactions (Kaas, Ruiz, and Lefranc 2004; Kaas and Lefranc 2005).

Since 1989, IMGT®, the international ImMunoGeneTics information system® (Lefranc, Giudicelli, Kaas, Duprat, Jabado-Michaloud, Scaviner, Ginestoux, Clément, Chaume, and Lefranc 2005c), http://imgt.cines.fr, has offered standardized genetic and structural data on immunoglobulins (IG), TR, and MHC, and on related proteins of the immune system (RPI) that belong to the immunoglobulin superfamily (IgSF) and to the MHC superfamily (MhcSF). In order to facilitate data comparison and cross-referencing between experiments from different laboratories whatever the receptor, the chain type, the domain, or the species, IMGT® has developed IMGT-ONTOLOGY (Giudicelli and Lefranc 1999), the first ontology in immunogenetics and immunoinformatics.

Based on the IMGT-ONTOLOGY concepts, the IMGT Scientific chart provides the controlled vocabulary and the annotation rules necessary for the identification, the description, the classification, and the numbering of the IG, TR, MHC, and RPI (Lefranc 2004a; Lefranc, Giudicelli, Ginestoux, Bosc, Folch, Guiraudou, Jabado-Michaloud, Magris, Scaviner, Thouvenin, Combres, Girod, Jeanjean, Protat, Monod, Duprat, Kaas, Pommié, Chaume, and Lefranc 2004b; Lefranc, Clément, Kaas, Duprat, Chastellan, Coelho, Combres, Ginestoux, Giudicelli, Chaume, and Lefranc 2005a). The IDENTIFICATION concept refers to the IMGT standardized keywords that are essential for the sequence and 3D structure assignments. The DESCRIPTION concept provides the IMGT standardized labels used to describe structural and functional regions that compose IG, TR, MHC, and RPI sequences and 3D structures. Standardized labels have also been defined to characterize the three-dimensional assembly of domains and chains. The CLASSIFICATION concept provides immunologists and geneticists with a standardized nomenclature per locus and per species. The human IG and TR gene nomenclature elaborated by IMGT was approved by the Human Genome Organisation (HUGO) Nomenclature Committee, HGNC (Wain, Bruford, Lovering, Lush, Wright, and Povey 2002), in 1999. The mouse IG and TR gene names with IMGT reference sequences were provided by IMGT to HGNC and to the Mouse Genome Database (MGD; Blake, Richardson, Bult, Kadin, and Eppig 2003) in July 2002. The NUMEROTATION concept provides the IMGT unique numbering for the IG and TR V-DOMAIN and the V-LIKE-DOMAIN of the IgSF proteins other than IG or TR (Lefranc, Pommié, Ruiz, Giudicelli, Foulquier, Truong, Thouvenin-Contet, and Lefranc 2003b), and for the IG and TR C-DOMAIN and the C-LIKE-DOMAIN of the IgSF proteins other than IG or TR (Lefranc, Pommié, Kaas, Duprat, Bosc,

Guiraudou, Jean, Ruiz, Da Piedade, Rouard, Foulquier, Thouvenin, and Lefranc 2005d). An IMGT unique numbering has also been set up for the MHC G-DOMAIN and the G-LIKE-DOMAIN of the MhcSF proteins other than MHC (Lefranc, Duprat, Kaas, Tranne, Thiriot, and Lefranc 2005b).

The IMGT standardization has allowed the construction of a unique frame for the comparison of the TR, peptide, and MHC interactions in the different resources provided by the IMGT® information system. IMGT/3Dstructure-DB (Kaas et al. 2004), the IMGT structural database, is used with the IMGT sequence databases, IMGT/LIGM-DB (Lefranc 2003a; Giudicelli, Ginestoux, Folch, Jabado-Michaloud, Chaume, and Lefranc 2006) and IMGT/MHC-DB (Robinson, Waller, Parham, de Groot, Bontrop, Kennedy, Stoehr, and Marsh 2003); the IMGT gene database, IMGT/GENE-DB (Giudicelli, Chaume, and Lefranc 2005); the IMGT tools for sequence analysis, IMGT/V-QUEST (Giudicelli, Chaume, and Lefranc 2004), IMGT/JunctionAnalysis (Yousfi Monod, Giudicelli, Chaume, and Lefranc 2004); and the IMGT tool for 3D structure analysis, IMGT/StructuralQuery (Kaas et al. 2004), to explore the TR and MHC conserved structural features. In this paper, we describe the IMGT standardized rules that have been set up for the molecular characterization of the TR/pMHC complexes. Coordinate files are from IMGT/3Dstructure-DB (Kaas et al. 2004), http://imgt.cines.fr, with original crystallographic data from the Protein Data Bank, PDB (Berman, Westbrook, Feng, Gilliland, Bhat, Weissig, Shindyalov, and Bourne 2000). Eleven IMGT pMHC contact sites were defined (C1 to C11) which can be used to compare pMHC interactions (Kaas and Lefranc 2005).

2.2 T Cell Receptor/Peptide/MHC 3D Structures and IMGT Standardization

IMGT/3Dstructure-DB (Table 1) contains 18 TR/pMHC structures: 14 (12 TR/pMHC-I and 2 TR/pMHC-II) with complete extracellular regions of the α-β TR (TR-ALPHA_BETA) and 4 structures with an Fv variable fragment (FV-ALPHA_BETA). The α-β TR chains, TR-ALPHA and TR-BETA, are described with standardized IMGT labels in Fig. 1.

The references for the 18 TR/pMHC 3D structures are: 1ao7 (Garboczi, Ghosh, Utz, Fan, Biddison, and Wiley 1996), 1qrn, 1qse, 1qsf (Ding, Baker, Garboczi, Biddison and Wiley 1999), 1bd2 (Ding, Smith, Garboczi, Utz, Biddison, and Wiley 1998), 1oga (Stewart-Jones, McMichael, Bell, Stuart, and Jones 2003), 1mi5 (Kjer-Nielsen et al. 2003), 1lp9 (Buslepp, Wang, Biddison, Appella, and Collins 2003), 1g6r (Degano, Garcia, Apostolopoulos, Rudolph, Teyton, and Wilson 2000), 1jtr, 1mwa (Luz, Huang, Garcia, Rudolph, Apostolopoulos, Teyton, and Wilson 2002), 2ckb (Garcia, Degano, Pease, Huang, Peterson, Teyton, and Wilson 1998), 1fo0 (Reiser, Darnault, Guimezanes, Gregoire, Mosser, Schmitt-Verhulst, Fontecilla-Camps, Malissen, Housset, and Mazza 2000), 1nam (Reiser, Darnault, Gregoire, Mosser, Mazza, Kearney, van der Merwe, Fontecilla-Camps, Housset, and Malissen 2003), 1kj2 (Reiser, Gregoire, Darnault, Mosser, Guimezanes, Schmitt-Verhulst, Fontecilla-Camps, Mazza, Malissen, and Housset 2002), 1fyt (Hennecke, Carfi, and

Wiley 2000), 1j8h (Hennecke and Wiley 2002), 1d9k (Reinherz, Tan, Tang, Kern, Liu, Xiong, Hussey, Smolyar, Hare, Zhang, Joachimiak, Chang, Wagner, and Wang 1999).

Table 1. TR/peptide/MHC complexes in IMGT/3Dstructure-DB (Kaas et al. 2004), http://imgt.cines.fr. Sp, species; Hs, *Homo sapiens*; Mm, *Mus musculus*; L, length in amino acids. Fourteen 3D structures (12 TR/pMHC-I and 2 TR/pMHC-II) correspond to TR receptors (TR-ALPHA_BETA). Four 3D structures (1d9k, 1fo0, 1kj2, and 1nam) correspond to an Fv variable fragment (FV-ALPHA_BETA). Gene and allele names are according to IMGT/GENE-DB (Giudicelli et al. 2005) for human and mouse TR, to IMGT/HLA-DB (Robinson et al. 2003) for human MHC, and to IMGT for mouse MHC. Amino acid sequences of the TR V-DOMAINs and MHC G-DOMAINs are reported in Figs. 3 and 4, respectively. H2-K1*01 encodes H2-K1b, H2-AB*02 and H2-AA*02 encode I-Abk and I-Aak, respectively. Lengths of the CDR-IMGT are according to Lefranc et al. (2003b).

(A) TR/pMHC-I

	T cell receptor				Peptide		MHC
3D	Name	Sp.	V-DOMAIN genes	CDR-IMGT	Sequence	L	Gene and allele
1ao7	A6	Hs	TRAV12-2-TRAJ24	[6.6.11]	LLFGYPVYV	9	HLA-A*0201
			TRBV6-5-TRBD2-TRBJ2-7	[5.6.14]			
1qrn	A6				LLFGY<u>A</u>VYV	9	
1qse	A6				LLFGYP<u>R</u>YV	9	
1qsf	A6				LLFGYPV<u>A</u>V	9	
1bd2	B7	Hs	TRAV29/DV5-TRAJ54	[6.6.10]	LLFGYPVYV	9	HLA-A*0201
			TRBV6-5-(TRBD2)-TRBJ2-7	[5.6.13]			
1oga	JM22	Hs	TRAV27-TRAJ42	[5.6.10]	GILGFVFTL	9	HLA-A*0201
			TRBV19-(TRBD2)-TRBJ2-7	[5.6.11]			
1mi5	LC13	Hs	TRAV26-2-TRAJ52	[7.4.14]	FLRGRAYGL	9	HLA-B*0801
			TRBV7-8-(TRBD2)-TRBJ2-7	[5.6.11]			
1lp9	12.2	Mm	TRAV12D-2-TRAJ50	[6.6.13]	ALWGFFPVL	9	HLA-A*0201
			TRBV13-3-(TRBD2)-TRBJ2-7	[5.6.11]			
1g6r	2C	Mm	TRAV9-4-TRAJ35	[6.7.10]	SIYRYYGL	8	H2-K1*01
			TRBV13-2-(TRBD2)-TRBJ2-4	[5.6.9]			
1jtr	2C				EQYKFYSV	8	
2dkb	2C				EQYKFYSV	8	
1mwa	2C				EQYKFYSV	8	

	T cell receptor				Peptide		MHC
3D	Name	Sp.	V-DOMAIN genes	CDR-IMGT	Sequence	L	Gene and allele
1fo0	BM3.3	Mm	TRAV16-TRAJ32	[7.7.14]	INFDFNTI	8	H2-K1*01
			TRBV1-TRBD1-TRBJ1-3	[6.6.12]			
1nam	BM3.3				RGYVYQGL	8	
1kj2	KB5-C20	Mm	TRAV14-1-TRAJ15	[6.6.11]	KVITFIDL	8	H2-K1*01
			TRBV1-TRBD2-TRBJ2-3	[6.6.16]			

(B) TR/pMHC-II

	T cell receptor				Peptide		MHC
3D	Name	Sp.	V-DOMAIN genes	CDR-IMGT	Sequence	L	Gene and allele
1fyt	HA1.7	Hs	TRAV8-4-TRAJ48	[6.7.13]	PKYVKQNT LKLAT	13	HLA-DR A *0101
		Hs	TRBV28-TRBD1-TRBJ1-2	[5.6.12]			HLA-DR B1*0101
1j8h	HA1.7				PKYVKQNTL KLAT	13	HLA-DR A*0101
							HLA-DR B1*0401
1d9k	D10	Mm	TRAV14D-2-TRAJ4	[6.6.10]	GNSHRGAIE WEGIESG	16	H2-AA *02
			TRBV13-2-TRBD2-TRBJ2-1	[5.6.11]			H2-AB *02

Each complete TR chain comprises an extracellular region made up of a variable domain and a constant domain (V-ALPHA and C-ALPHA for the α chain, V-BETA and C-BETA for the β chain) (Fig. 1), a connecting region, a transmembrane region, and a very short intracytoplasmic region. The MHC-I is formed by the association of a heavy chain (I-ALPHA) and a light chain (β-2-microglobulin B2M, Fig. 1). The MHC-II is a heterodimer formed by the association of an α chain (II-ALPHA) and a β chain (II-BETA). The I-ALPHA chain of the MHC-I, and the II-ALPHA and II-BETA chains of the MHC-II comprise an extracellular region made of three domains for the MHC-I and of two domains for the MHC-II, a connecting region, a transmembrane region, and an intracytoplasmic region. The I-ALPHA chain comprises two groove domains (G-DOMAIN), G-ALPHA1 [D1] and G-ALPHA2 [D2], and a C-LIKE domain [D3]. The B2M corresponds to a single C-LIKE domain. The II-ALPHA chain and the II-BETA chain each comprise two domains, G-ALPHA [D1] and C-LIKE [D2],

and G-BETA [D1] and C-LIKE [D2]. Only the extracellular region that corresponds to these domains has been crystallized (Fig. 1). The TR V-DOMAINs and MHC G-DOMAINs that are directly involved in TR/pMHC interactions are described in the next sections.

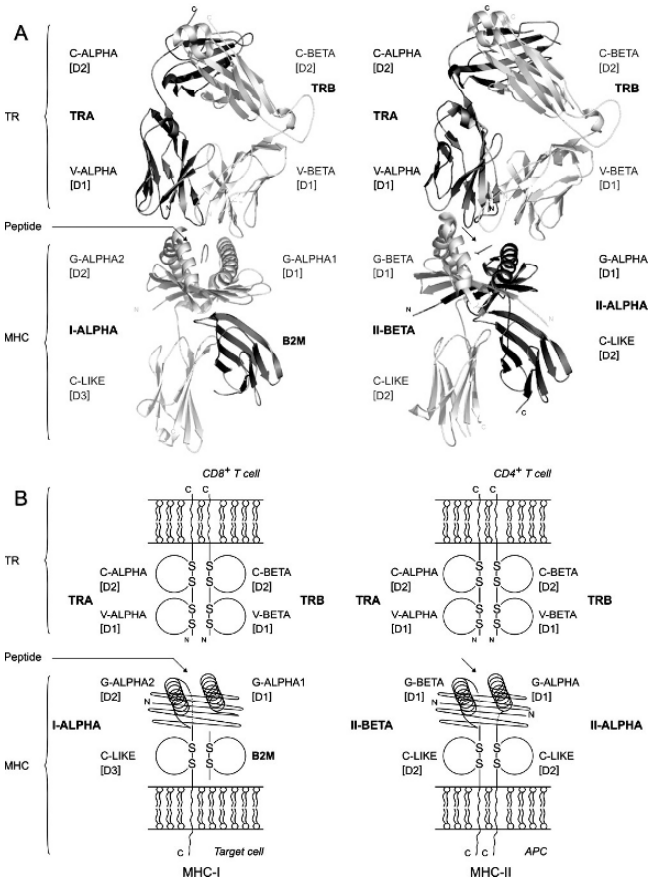


Fig. 1. T cell receptor/peptide/MHC complexes with MHC class I (TR/pMHC-I) and MHC class II (TR/pMHC-II). [D1], [D2] and [D3] indicate the domains. (A) 3D structures of TR/pMHC-I (1oga) and TR/pMHC-II (1j8h). (B) Schematic representation of TR/pMHC-I and TR/pMHC-II. The TR (TR-ALPHA and TR-BETA) chains, the MHC-I (I-ALPHA and β-2-microglobulin B2M) chains and the MHC-II (II-ALPHA and II-BETA) chains are shown with the extracellular domains (V-ALPHA and C-ALPHA for the TR-ALPHA chain; V-BETA and C-BETA for the TR-BETA chain; G-ALPHA1, G-ALPHA2 and C-LIKE for the I-ALPHA chain; C-LIKE for B2M; G-ALPHA and C-LIKE for the II-ALPHA chain; II-BETA and C-LIKE for the II-BETA chain), and the connecting, transmembrane and cytoplasmic regions. Arrows indicate the peptide localization in the G-DOMAIN groove. The MHC G-DOMAINs and TR V-DOMAINs are likely to be in a diagonal rather than in a vertical position relative to the cell surface (Wang, Meijers, Xiong, Liu, Sakihama, Zhang, Joachimiak and Reinherz 2001; Wang and Reinherz 2002).

2.2.1 TR V-DOMAINs

The V-DOMAINs have an immunoglobulin fold, that is an antiparallel β sheet sandwich structure with nine strands (Lefranc et al. 2003b; Lesk and Chothia 1982), the A, B, E, and D strands being on one sheet, and the G, F, C, C', and C" strands on the other sheet. These strands are indicated in the IMGT Colliers de Perles (Fig. 2) and in the IMGT Protein displays (Fig. 3).

IMGT Colliers de Perles are IMGT 2D graphical representations based on the IMGT unique numbering. The IMGT Colliers de Perles of TR V-DOMAINs are based on the IMGT unique numbering for V-DOMAIN and V-LIKE-DOMAIN (Lefranc et al. 2003b) and can be displayed on one layer or on two layers. IMGT Colliers de Perles of the V-ALPHA and V-BETA domains from 1ao7 (Garboczi et al. 1996) are shown as examples in Fig. 2. The IMGT Protein display (Fig. 3) shows the amino acid sequences of the different V-ALPHA and V-BETA domains found in TR/pMHC (Table 1).

The V-ALPHA and V-BETA domains share main conserved characteristics of the V-DOMAIN which are the disulfide bridge between cysteine 23 (1st-CYS) and cysteine 104 (2nd-CYS), and the other hydrophobic core residues tryptophan 41 (CONSERVED-TRP) and leucine (or hydrophobic) 89 (Lefranc et al. 2003b) (Figs. 2 and 3). The A strand comprises positions 1 to 15, B strand positions 16 to 26, C strand positions 39 to 46, C' strand positions 47 to 55, C" strand positions 66 to 74, D strand positions 75 to 84, E strand positions 85 to 96, F strand positions 97 to 104, and G strand positions 118 to 128 (Lefranc et al. 2003b). Compared to the general V-DOMAIN 3D structure, the V-ALPHA domains have shorter C" and D strands at the C" D turn.

The three hypervariable loops or complementarity determining regions (CDR) of each V-DOMAIN are involved in the pMHC recognition. The CDR1-IMGT comprises positions 27 to 38, the CDR2-IMGT positions 56 to 65, and the CDR3-IMGT positions 105 to 117 (Lefranc et al. 2003b). The CDR3-IMGT corresponds to the junction resulting from the V-J and V-D-J rearrangement, and is more variable in sequence and length than the CDR1-IMGT and CDR2-IMGT that are encoded by the V-REGION only (Lefranc and Lefranc 2001). Lengths of the CDR-IMGT are shown separated by dots between brackets (Lefranc et al. 2003b). Lengths of the CDR-IMGT from available TR/pMHC 3D structures are reported in Table 1, together with the names of the V, D, and J genes (Lefranc and Lefranc 2001).

For example, 1ao7 [6.6.11] V-ALPHA means that in the V-ALPHA domain of 1ao7, CDR1-IMGT has a length of 6 amino acids, CDR2-IMGT a length of amino acids, and CDR3-IMGT a length of 11 amino acids. The V-ALPHA CDR3-IMGT results from the TRAV12-2–TRAJ24 rearrangement (Table 1, Fig. 3). In the same way, 1ao7 [5.6.14] V-BETA means that in the V-BETA domain of 1ao7, CDR1-IMGT, CDR2-IMGT, and CDR3-IMGT have a length of 5, 6, and 14 amino acids, respectively (Lefranc et al. 2003b). The V-BETA CDR3-IMGT results from the TRBV6-5–TRBD2–TRBJ2-7 rearrangement (Table 1, Fig. 3).

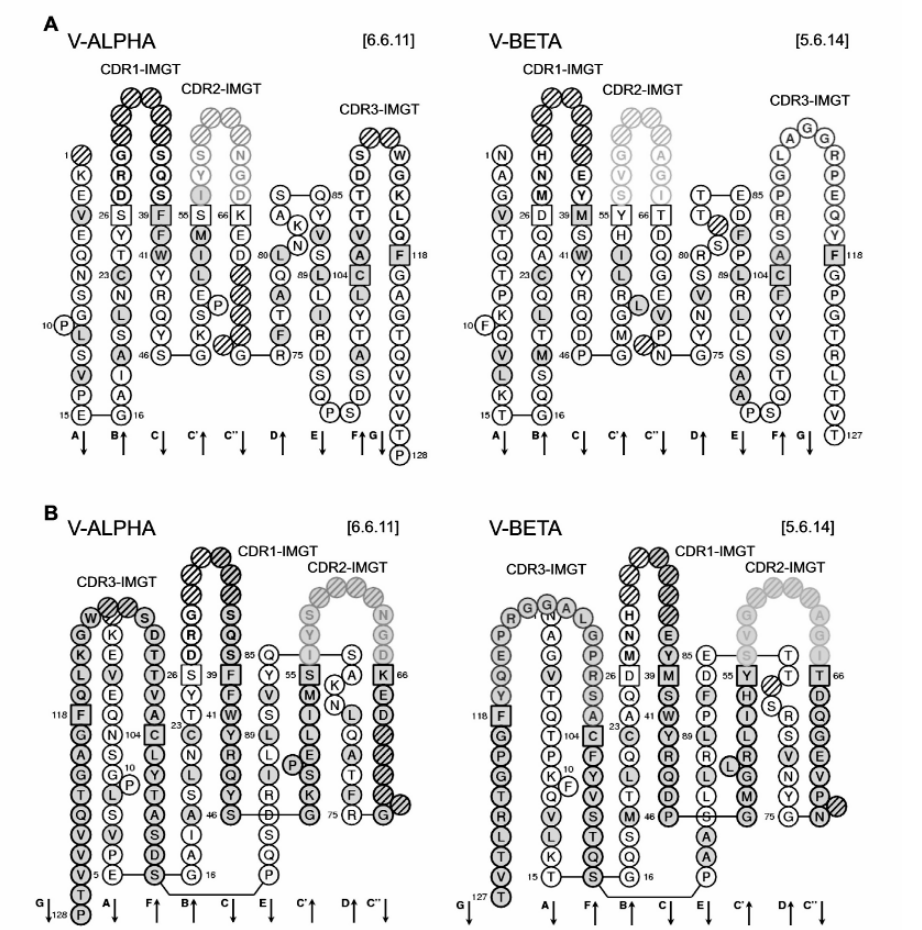


Fig. 2. IMGT Colliers de Perles of the V-ALPHA and V-BETA domains from 1ao7 (IMGT/ 3Dstructure-DB, http://imgt.cines.fr) (A) on one layer (B) on two layers. Amino acids are shown in the one-letter abbreviation. Hydrophobic amino acids (hydropathy index with positive value) and tryptophan (W) found at a given position in more than 50% of analysed IG and TR sequences are shown. The CDR-IMGTs are limited by amino acids shown in squares, which belong to the neighbouring FR-IMGT and represent anchor positions. Hatched circles correspond to missing positions according to the IMGT unique numbering (Lefranc et al. 2003b). Arrows indicate the direction of the β sheets.

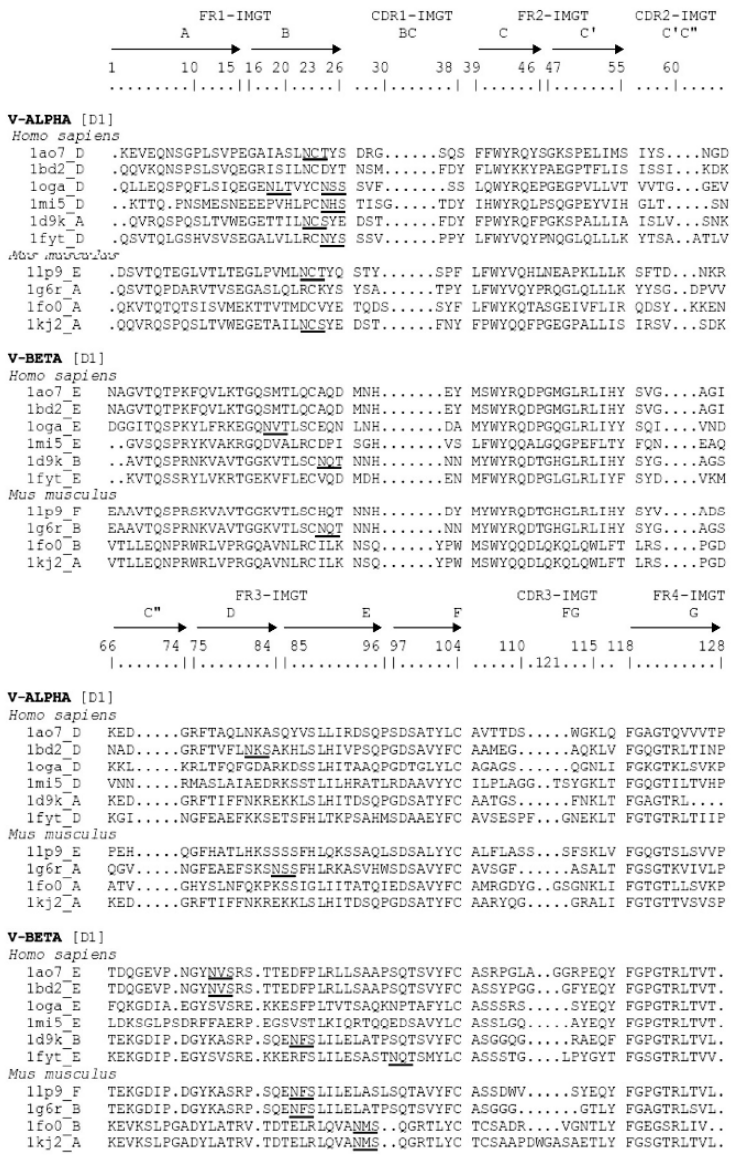


Fig. 3. Protein display of the TR V-ALPHA and V-BETA domains found in the TR/pMHC complexes in IMGT/3Dstructure-DB (Kaas et al. 2004), http://imgt.cines.fr. Amino acid sequences and gaps (shown by dots) are according to the IMGT unique numbering for V-DOMAIN (Lefranc et al. 2003b). The three additional positions in the CDR3-IMGT are 111.1, 112.2 and 112.1. Potential N-glycosylation sites are underlined. Assignments of the V, D and J genes are shown in Table 1.

2.2.2 MHC G-DOMAINs

The four G-DOMAINs, G-ALPHA1 and G-ALPHA2 of the MHC-I, and G-ALPHA and G-BETA of the MHC-II (Figs. 1, 4, and 5), have a similar groove 3D structure that consists of one sheet of four antiparallel β strands ("floor" of the groove or platform) and one long helical region ("wall" of the groove) (Lefranc et al. 2005b). For each G-DOMAIN (Figs. 4 and 5), the A strand comprises positions 1 to 14, B strand positions 18 to 28, C strand positions 31 to 38, and D strand positions 42 to 49 (Lefranc et al. 2005b). The helix (positions 50 to 92) seats on the β sheet and its axis forms an angle of about 40 degrees with the β strands. The helix is split into two parts separated by a kink, positions 58 of G-ALPHA1, 61 of G-ALPHA2, 63 of G-ALPHA, and 62 of G-BETA being the "highest" points on the floor groove. The G-ALPHA2 and G-BETA domains have a disulfide bridge between positions 11 and 74. The G-ALPHA1 and G-ALPHA domains have a conserved N-glycosylation site at position 86 (N-X-S/T, where N is asparagine, X any amino acid except proline, S is serine, and T is threonine), except for HLA-DMB and H2-DMB1. Asparagine 15 of the G-BETA domains also belongs to a conserved N-glycosylation site (Lefranc et al. 2005b).

2.3 TR/pMHC Contact Analysis

2.3.1 Peptide/MHC

The 3D structure of the MHC main chain is well conserved and the peptide binding groove specificity is due to side chain physicochemical characteristics (Reinherz et al. 1999). Both MHC-I and MHC-II grooves have pockets where side chains of bound peptides may anchor (Falk, Rotzschke, Stevanovic, Jung, and Rammensee 1991), the specificity of a peptide to a given MHC being controlled by the physicochemical properties of the pockets. Conversely, comparisons of peptide sequence alignments and pMHC 3D structures have revealed that some anchored peptide positions with conserved properties were needed to bind a peculiar MHC allele. Several databases, SYFPEITHI (Rammensee, Bachmann, Emmerich, Bachor, and Stevanovie 1999), JenPep (Blythe, Doytchinova, and Flower 2002), and MHCpep (Brusic, Rudy, and Harrison 1998), provide peptide sequences associated with MHC alleles together with anchor positions and experimental data on affinity. These observations have extensively been used in peptide/MHC binding prediction (Singh and Raghava 2003; Adams and Koziol 1995; Vasmatzis, Zhang, Cornette, and DeLisi 1996b). A list of prediction programs and servers is available at "The IMGT Immunoinformatics page" (http://imgt.cines.fr). Nevertheless, exceptions have been found (Mandelboim, Bar-Haim, Vadai, Fridkin, and Eisenbach 1997; Apostolopoulos, Yu, Corper, Teyton, Pieters, McKenzie, and Wilson 2002; Scott, Peterson, Teyton, and Wilson 1998) and it was noted that while only 30% of peptides with the expected pattern really bind, peptides without the expected pattern also bind (Gulukota, Sidney, Sette, and

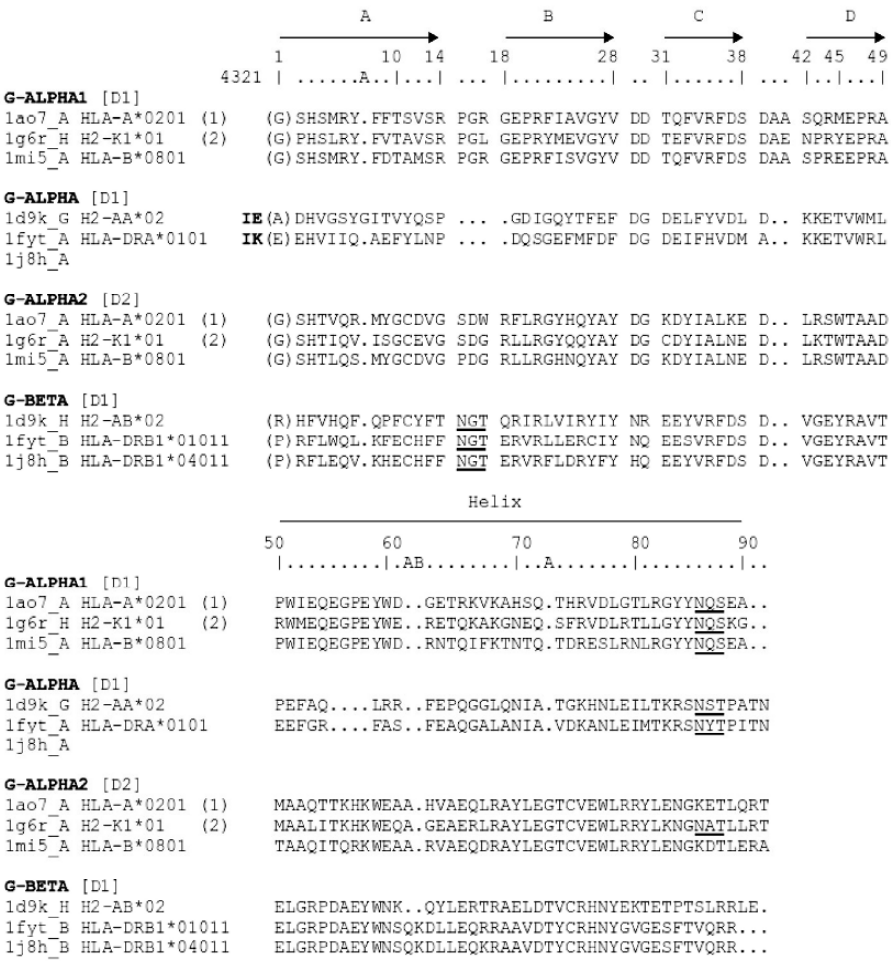


Fig. 4. Protein display of the G-DOMAINs found in the TR/pMHC complexes in IMGT/3Dstructure-DB (Kaas et al. 2004), http://imgt.cines.fr. Amino acid sequences and gaps (shown by dots) are according to the IMGT unique numbering for G-DOMAIN (Lefranc et al. 2005b). Amino acids resulting from the splicing with the preceding exon are shown within parentheses. Potential N-glycosylation sites are underlined. Positions 61A, 61B and 72A are characteristic of the G-ALPHA2 and G-BETA domains. The corresponding gaps in G-ALPHA1 and G-ALPHA shown in this IMGT Protein display are not reported in the IMGT Colliers de Perles as these gaps are shared by those two domains. H2-K1*01 encodes H2-K1b, H2-AB*02 and H2-AA*02 encode I-Abk and I-Aak, respectively.

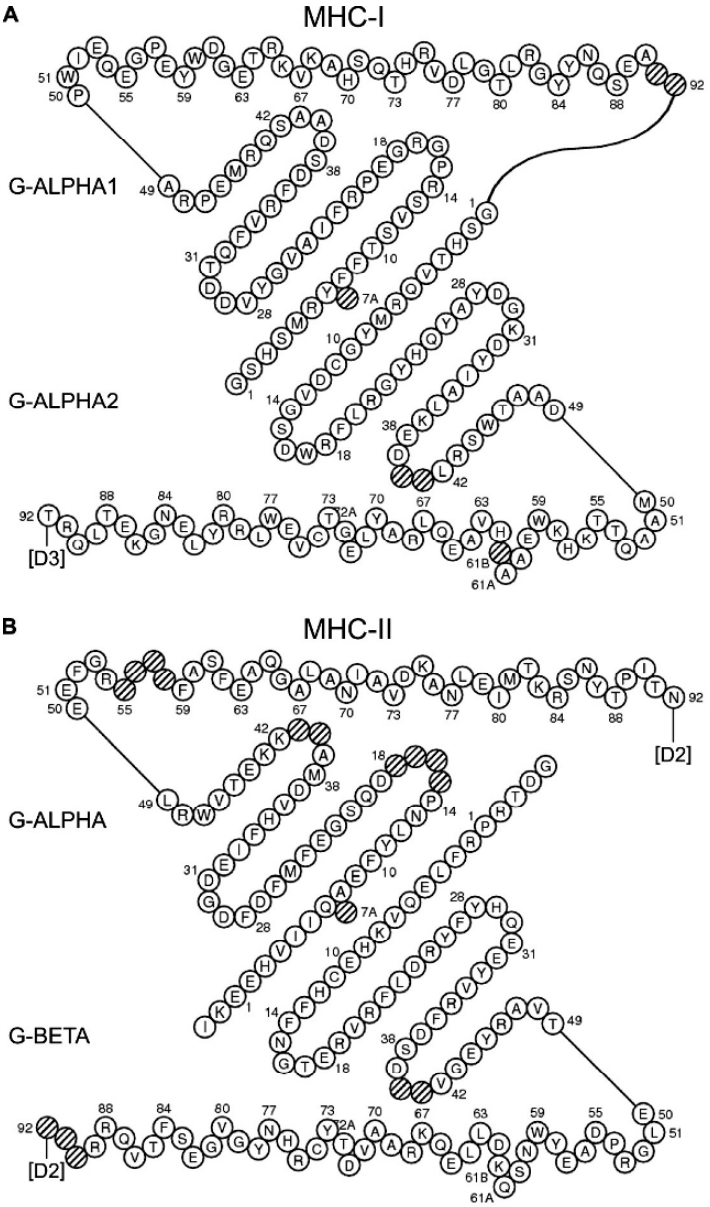


Fig. 5. IMGT Colliers de Perles of MHC G-DOMAINs. (A) MHC-I G-ALPHA1 and G-ALPHA2 domains from 1ao7 (B) MHC-II G-ALPHA and G-BETA domains from 1j8h (IMGT/3Dstructure-DB (Kaas et al. 2004), http://imgt.cines). Amino acids positions are according to the IMGT unique numbering for G-DOMAIN (Lefranc et al. 2005b). Positions 61A, 61B and 72A are characteristic of the G-ALPHA2 and G-BETA domains (and are not reported in the G-ALPHA1 and G-ALPHA IMGT Colliers de Perles).

DeLisi 1997). Peptide/MHC binding prediction and epitope prediction remain a big challenge. In order to compare data from different MHC sequences and 3D structures, the IMGT unique numbering for G-DOMAIN has been set up (Lefranc et al. 2005b) (Figs. 4 and 5). This has allowed to graphically represent, in the IMGT Colliers de Perles for G-DOMAIN (Fig. 5), the MHC amino acid positions that have contacts with the peptide side chains. Eleven IMGT pMHC contact sites were defined (C1 to C11, in Figs. 6–8) which can be used to compare pMHC interactions (Kaas and Lefranc 2005). Examples of contact sites for an MHC-I binding an 8-mer peptide (1jtr), for an MHC-I binding a 9-mer peptide (1ao7), and for an MHC-II binding the nine amino acids of a peptide (1j8h) are shown in Figs. 6, 7, and 8, respectively.

In contrast to previous attempts to define pockets (Zhang, Anderson, and DeLisi 1998), structural data for defining the IMGT pMHC contact sites take into account the length of the peptides and are considered independently of the MHC class and sequence polymorphisms. The interactions between the peptide amino acid side chains and MHC amino acids were computed using an interaction scoring scheme based on true mean energy ratio (Kaas and Lefranc 2005). All direct contacts (defined with a cutoff equal to the sum of the atom van der Waals radii and of the diameter of a water molecule) and water-mediated hydrogen bonds were taken into account for the definition of the IMGT pMHC contact sites (Kaas and Lefranc 2005). The analysis was carried out for the pMHC available in IMGT/3Dstructure-DB (Kaas et al. 2004), http://imgt.cines.fr. One hundred fourteen 3D structures with peptides of 8, 9, and 10 amino acids bound to MHC-I and forty-four 3D structures of pMHC-II were identified. The contact analysis was performed for the peptide amino acid side chains of the 9 amino acids located in the groove. Results for MHC-I with 8-amino acid peptides (30 pMHC-I 3D structures), MHC-I with 9-amino acid peptides (74 pMHC-I 3D structures), and MHC-II for the 9 amino acids located in the groove (44 pMHC-II 3D structures) are reported in Table 2 (the results for the 10 pMHC-I with 10-amino acid peptides are not shown). These "IMGT reference pMHC contact sites" are also available as IMGT Colliers de Perles. They will be updated as the number of 3D structures increases. IMGT Colliers de Perles for IMGT pMHC contact sites are provided for each individual pMHC and TR/pMHC entry in IMGT/3Dstructure-DB. They allow easy identification of the amino acid contacts between the MHC and the peptide amino acid side chains and comparison of them with the "IMGT reference pMHC contact sites".

C1 to C11 refer to the 11 IMGT pMHC contact sites (Kaas and Lefranc 2005). 1 to 9 refer to the numbering of the peptide amino acids in the groove. The peptide binding mode to MHC-I is characterized by the N and C peptide ends docked deeply with C1 and C11 contact sites that correspond to the two conserved pockets A and F, and by the peptide length that mechanically constrains the peptide conformation in the groove. There are no C2, C7, and C8 contact sites for MHC-I with 8-amino acid peptides and no C2 and C7 contact sites for MHC-I with 9-amino acid peptides. In contrast, for MHC-II, C2 is present but there are no C7 and C8. Whereas C1 and C11 correspond to the conserved pockets A and F, respectively, the correspondence between the other

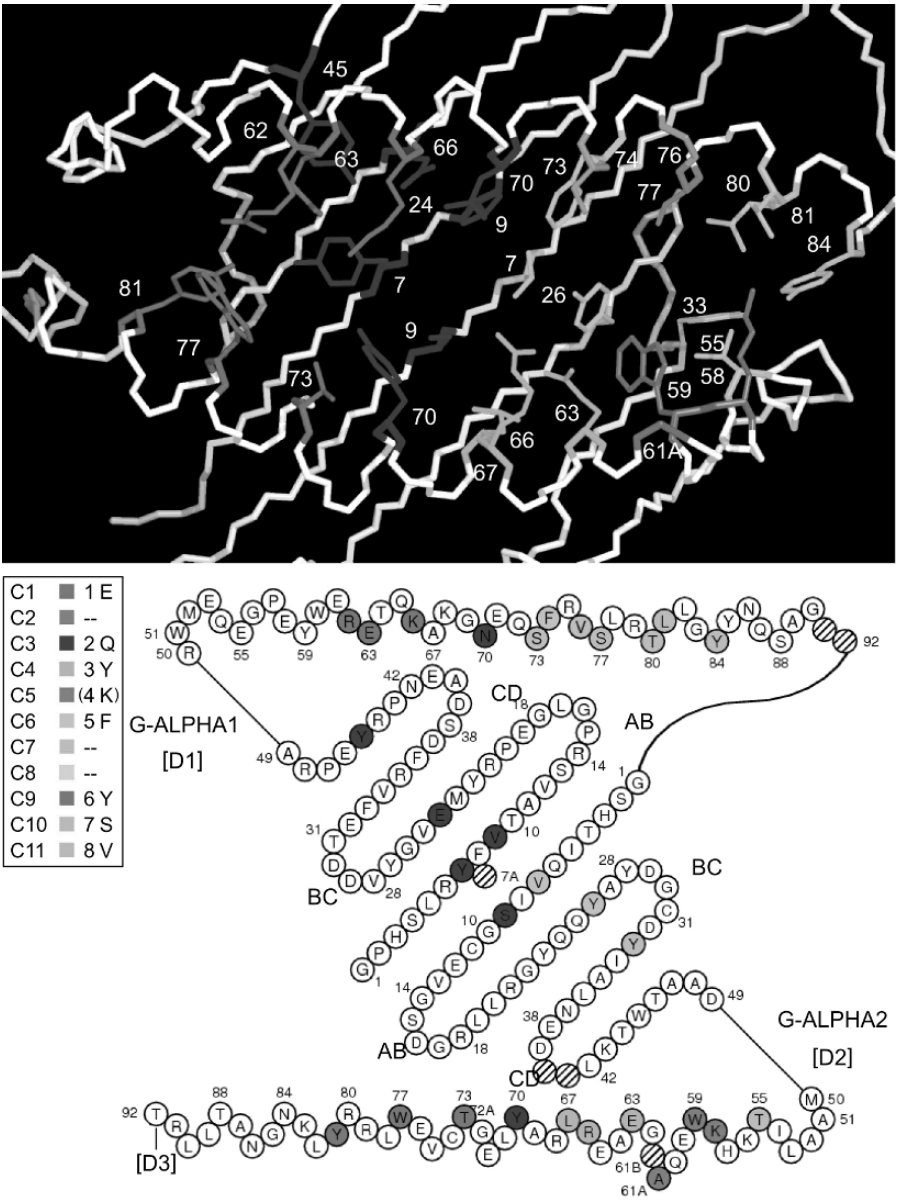


Fig. 6. IMGT pMHC contact sites of mouse H2-K1 MHC-I and a 8-amino acid peptide (1jtr). (A) 3D structure of the mouse H2-K1*01 groove. (B) IMGT pMHC contact sites IMGT Colliers de Perles. Both views are from above the cleft with G-ALPHA1 on top and G-ALPHA2 on bottom. In the box, C1 to C11 refer to contact sites (Kaas and Lefranc 2005), 1 to 8 refer to the numbering of the peptide amino acids P1 to P8. There are no C2, C7 and C8 in MHC-I 3D structures with 8-amino acid peptides. There is no C5 in this 3D structure as P4 does not contact MHC amino acids (4K is shown between parentheses in the box). (A color version of this figure appears between pages 76 and 77.)

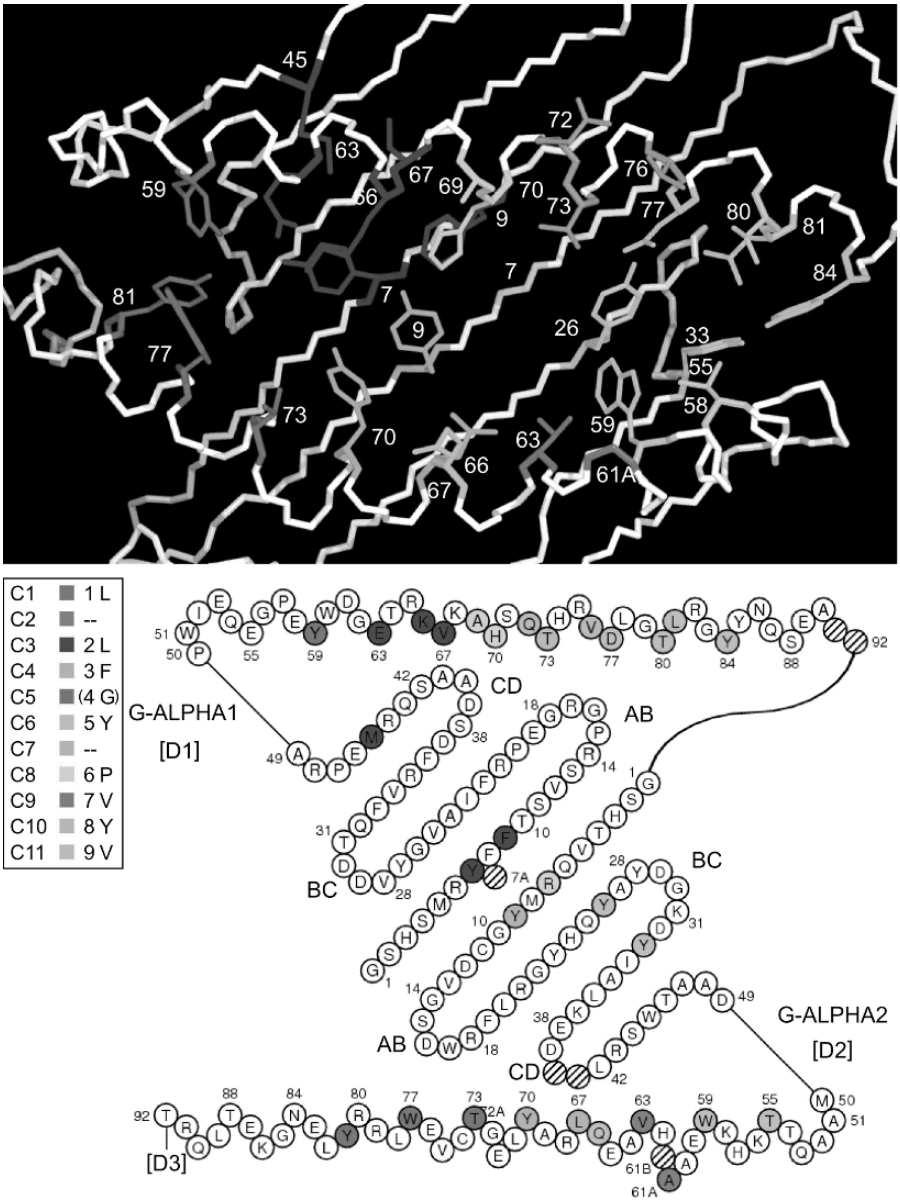


Fig. 7. IMGT pMHC contact sites of human HLA-A*0201 MHC-I and a 9-amino acid peptide (1ao7). (A) 3D structure of the human HLA-A*0201 groove. (B) IMGT pMHC contact sites IMGT Colliers de Perles. Both views are from above the cleft with G-ALPHA1 on top and G-ALPHA2 on bottom. In the box, C1 to C11 refer to contact sites (Kaas and Lefranc 2005). 1 to 9 refer to the numbering of the peptide amino acids P1 to P9. There are no C2 and C7 in MHC-I 3D structures with 9-amino acid peptides. There is no C5 in this 3D structure as P4 does not contact MHC amino acids (4G is shown between parentheses in the box). (A color version of this figure appears between pages 76 and 77.)

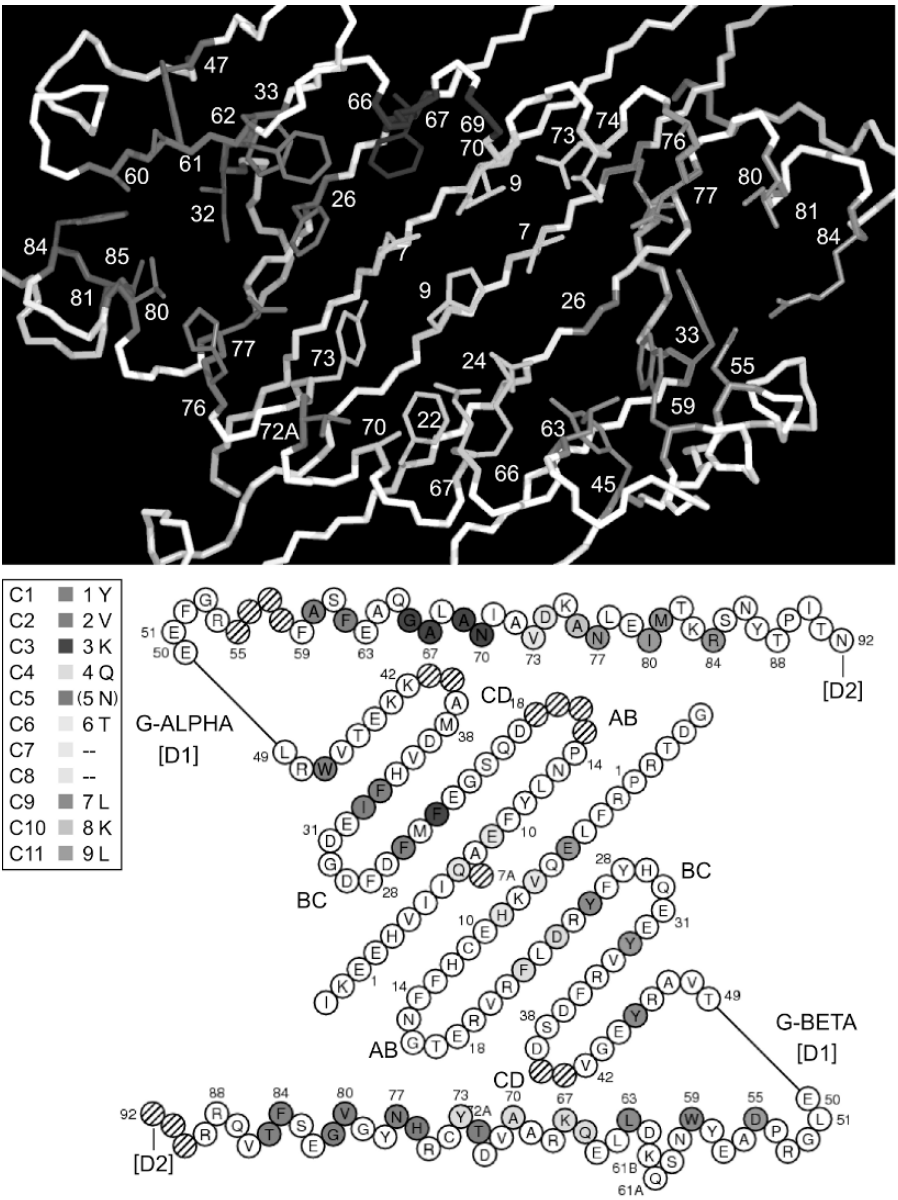


Fig. 8. IMGT pMHC contact sites of the human HLA-DRA*0101 and HLA-DRB1*0401 MHC-II and the peptide side chains (9-amino acids located in the groove). (A) 3D structure of the human HLA-DRA*0101 and HLA-DRB1*0401 groove (1j8h). (B) IMGT pMHC contact sites IMGT Colliers de Perles. Both views are from above the cleft with G-ALPHA on top and G-BETA on bottom. In the box, C1 to C11 refer to contact sites. 1 to 9 refer to the numbering of the peptide amino acids 1 to 9 located in the groove. There is no C5 and C7 in MHC-I 3D structures with 9-amino acid peptides. There is no C5 in this 3D structure as 5 does not contact MHC amino acids (5N is shown between parentheses in the box). (A color version of this figure appears between pages 76 and 77.)

contact sites and the previously defined pockets is more approximative. For MHC-I with a peptide of 8-amino acids, C3, C4, C6, and C9 correspond roughly to the B, D, C, and E pockets, and for MHC-I with a peptide of 9-amino acids C3, C4, and C9 correspond to the B, D, and E pockets.

Table 2. IMGT reference pMHC contact sites. (A) MHC-I. Results for 104 pMHC-I 3D structures (30 with 8-amino acid peptides and 74 with 9-amino acid peptides). (B) MHC-II. Results from 44 pMHC-II 3D structures with 9 amino acids in the groove.

(A) MHC-I			
8-amino acid peptides			
		G-ALPHA1	G-ALPHA2
C1	1	59 62 63 66	73 77 81
C3	2	7 24 45	9
C4	3		9 24 63 66 67 70
C5	4		
C6	5	7 9 22 70 74	7 9 24 26
C9	6		59 61A 63 66
C10	7	77 73 76	
C11	8	77 80 81 84	5 26 33 34 55 59
9-amino acid peptides			
		G-ALPHA1	G-ALPHA2
C1	1	5 59 62 63 66	73 77 81
C3	2	7 9 22 24 34 45 63 66 67 70	
C4	3		7 9 24 66 67 70
C5	4	65 66	66
C6	5	70 73 74	7 26 66 67
C8	6	66 69 70 73 74	7 24 62 66
C9	7		7 24 59 61A 63 66
C10	8	72 73 76 80	58
C11	9	77 80 81 84	5 26 33 34 55 59
(B) MHC-II			
		G-ALPHA	G-BETA
C1	1	26 33 34 47 60 61 62	77 80 81 82 84 85
C2	2		72A 73 76
C3	3	7 24 62 63 66 67 69	
C4	4	7	9 11 22 24 66 67 70 73 74
C5	5		66
C6	6	9 69 70 73 74	7 26
C9	7		24 26 45 59 63 66
C10	8	73 76	
C11	9	77 80 81 84	5 33 55

2.3.2 TR/pMHC

The analysis of the pairwise contacts that occur at the TR/MHC and TR/peptide interfaces was carried out using the IMGT unique numbering for V-DOMAINs (Lefranc et al. 2003b) for the TR, and the IMGT unique numbering for G-DOMAINs for the MHC (Lefranc et al. 2005b). Table 3 shows the interactions of the TR V-ALPHA and TR V-BETA with MHC-I and the peptide, in nine TR/pMHC-I 3D structures. Table 4 shows the interactions of the TR V-ALPHA and TR V-BETA with MHC-II and the peptide, in two TR/pMHC-II 3D structures. The results show that positions implicated in the binding are well conserved but not the pairwise interactions. The MHC contact positions belong to the G-DOMAIN helices. The TR positions that are involved in the contacts belong mostly to the CDR-IMGT and to anchor positions (shown by squares in Fig. 2). The FR-IMGT positions involved in the contacts are positions 84 and 84A that are located at the DE turn (designated as "hypervariable 4" or HV4). The contact analysis confirms that the V-ALPHA CDR2-IMGT seats on top of the G-ALPHA2 (MHC-I) or G-BETA (MHC-II) helices and that the V-BETA CDR2-IMGT seats on top of the G-ALPHA1 (MHC-I) or G-ALPHA (MHC-II) helices (Tables 3 and 4). This agrees with data (Sim et al. 1996) which showed that most of the TR/MHC specificity comes from the CDR1 and CDR2 because mutations in these CDRs are able to change specificity between MHC-I and MHC-II. V-ALPHA and V-BETA CDR3-IMGT usually follow the same G-DOMAIN contact preference as the CDR2-IMGT but they can also have contacts with the other G-DOMAINs. For example, in the 1oga 3D structure, position 66 of G-ALPHA2 is contacted by the V-ALPHA CDR3-IMGT but also by the V-BETA CDR3-IMGT.

The diagonal orientation of the TR/pMHC complex puts the TR in a globally conserved position for a peptide "read-out" (Buslepp et al. 2003). V-ALPHA is on top of the peptide N terminus while V-BETA is on top of the peptide C terminus. TR positions implicated in the peptide recognition are in CDR3-IMGT and generally to a lesser extent in V-ALPHA CDR1-IMGT (Tables 3 and 4). Nearly every 3D structure shows different CDR3 conformations and binding mode. In the JM22/peptide/HLA-A complex (1oga) (Stewart-Jones et al. 2003), the V-BETA CDR3-IMGT extensively contacts the peptide and G-ALPHA2 through hydrogen bonds (Table 3), by inserting itself between the peptide and the G-ALPHA2. In contrast, the 2C/peptide/H2-K1complex (1jtr) (Degano et al. 2000) has comparatively fewer contacts between the V-BETA CDR3-IMGT and the peptide and the MHC; however the V-BETA CDR1-IMGT has more contacts and hydrogen bonds with the peptide and G-ALPHA2.

The TR LC13 and 2C were crystallized both alone and in complex with a pMHC. The structural superimposition of both V-DOMAIN scaffold α carbons reveals large movements of the CDR3 and of the CDR1, respectively. The V-ALPHA domains of LC13, in the 1mi5 and 1kgc 3D structures, have 3.5 Å root mean square (RMS) between their CDR3. The V-ALPHA domains of 2C, in the 2ckb and 1tcr 3D structures, have 2.3 Å RMS between their CDR1. The TR A6 was crystallized in complex with the same MHC but with different peptides. In these structures, the V-BETA CDR3 adopt different conformations to adapt to the different peptides (Rudolph, Luz, and Wilson 2002). The CDR3 conformational change does not increase the binding surface but gives a better shape complementarity to the interface (Lawrence and Colman 1993).

Table 3. TR V-ALPHA and V-BETA CDR interactions with pMHC-I. TR positions in bold indicate hydrogen bonds. 3D structures are from IMGT/3Dstructure-DB (Kaas et al. 2004), http://imgt.cines.fr. Lengths of the CDR-IMGT are shown within brackets. Amino acids are shown in the one-letter code. Sequences of the peptides are reported in Table 1, sequences of the TR V-ALPHA and V-BETA domains in Fig. 3 and sequences of the MHC-I G-ALPHA1 and G-ALPHA2 in Fig. 4. (A) V-ALPHA CDR-IMGT interactions. (B) V-BETA CDR-IMGT interactions. (C) V-ALPHA and V-BETA FR-IMGT interactions.

(A) V-ALPHA CDR-IMGT interactions

V-ALPHA CDR1-IMGT				
PDB	CDR1	G-ALPHA1	Peptide	G-ALPHA2
1ao7 [6.]	27_D	**58_E**		
	28_R	58_E		77_W 80_R
	29_G		1_L	77_W
	37_Q	66_K	1_L **2_L** 3_F 4_G 5_Y	70_Y 73_T
	38_S		5_Y	
1bd2 [6.]	28_S		1_L	76_E **77_W**
	29_M	58_E 59_Y 62_G 63_E 66_K	1_L	77_W
	37_D	66_K	4_G **5_Y**	66_Q 73_T
	38_Y		5_Y	66_Q
1oga [5.]	37_S			65_E 66_Q
1mi5 [7.]	29_S	62_R		
	30_G			69_A
	36_T		4_G	**66_Q** 70_Y 73_T
	38_Y		7_Y	$61A_A$ 62_R 63_V 64_A 65_E 66_Q
1lp9 [6.]	28_T			76_E
	29_Y			69_A $72A_G$ 73_T 76_E 77_W
	36_S			69_A
	38_F			65_E 66_Q 69_A
1g6r [6.]	27_Y	62_R		
	28_S	58_E **62_R**		
	29_A	62_R		
	36_T			76_E
	38_Y		3_Y 4_R	66_R
1jtr [6.]	27_Y	62_R		
	28_S	58_E 59_Y 62_R		
	29_A	62_R	1_E	77_W
	36_T		1_E	
	38_Y		3_Y 4_K	66_R
1fo0 [7.]	28_Q	58_E 62_R		
	29_D	62_R		
	30_S			**73_T**
	36_S			69_A
	38_F			66_R
1kj2 [6.]	27_D	58_E **62_R**		
	29_T	62_R	1_K	77_W
	37_N			73_T

(continued)

Table 3. (continued) V-ALPHA CDR2-IMGT

PDB	CDR2	G-ALPHA1	Peptide	G-ALPHA2
1ao7 [.6.]	57_Y			65_E 66_Q 69_A
	58_S			69_A
	63_N			76_E
1bd2 [.6.]	57_S			65_E 66_Q 69_A
	58_S			69_A 70_Y 73_T
	59_I			68_R 69_A 72_E $72A_G$
1oga [.6.]	57_V			62_H 65_E
1mi5 [.4.]	56_G			62_R
	57_L			65_E 66_Q 69_A
	58_T			65_E
	64_S			65_E
1lp9 [.6.]	57_F			$61A_A$ 62_H 65_E 66_Q
	58_T			62_H 65_E
	64_K			65_E
1g6r [.7.]	57_Y			66_R 69_A
	58_S			69_A $72A_G$ 73_T 76_E
1jtr [.7.]	57_Y			65_E 66_R 69_A
	58_S			69_A 76_E
1fo0 [.7.]	59_Y			62_G 65_E 66_R 69_A
	62_K			65_E
1kj2 [.6.]	57_R			69_A 72_E
	58_S			76_E
	59_V			72_E $72A_G$ 76_E

V-ALPHA CDR3-IMGT

PDB	CDR3	G-ALPHA1	Peptide	G-ALPHA2
1ao7 [.11]	108_T	$\mathbf{65}_R$ 66_K	4_G 5_Y	
	109_D	62_G $\mathbf{65}_R$ 66_K	4_G 5_Y	
	110_S		$\mathbf{4}_G$ 5_Y 6_P	
	113_W	65_R 68_K 69_A 72_Q		
	114_G	65_R		
1bd2 [.10]	107_M		5_Y	
	108_E	58_E 62_G $\mathbf{65}_R$ 66_K		
	109_G	65_R 66_K	4_G 5_Y	
	113_A		4_G 5_Y	
	114_Q	65_R 69_A		
	115_K	65_R		
1oga [.10]	107_A			66_Q
	108_G		5_F	66_Q
	109_S		4_G 5_F	66_Q
	113_Q	66_K	4_G 5_F	
	114_G		4_G 5_F	
1mi5 [.14]	108_L		6_A 7_Y	66_Q
	109_A	62_R		
	110_G	62_R 66_I		
	111_G	65_Q 66_I $\mathbf{69}_T$	4_G	

PDB	CDR3	G-ALPHA1	Peptide	G-ALPHA2
	112.1_T	62_R 65_Q 66_I 69_T		
	112_S	69_T	6_A	
	113_Y	69_T 72_Q	6_A	
1lp9 [.13]	107_F		5_F	66_Q
	109_A		3_W 4_G 5_F	66_Q
	110_S		2_L 3_W 4_G	66_Q 69_A 70_Y 73_T
	111_S	63_E $\mathbf{66}_K$	2_L 4_G	73_T 77_W
	112_S	66_K	4_G 5_F	
	113_F	65_R 66_K 69_A	4_G 6_F	
	114_S		$\mathbf{4}_G$ 5_F 6_F	
1g6r [.10]	107_S		$\mathbf{4}_R$	
	108_G		$\mathbf{4}_R$	
	109_F	62_R 65_Q 66_K	4_R	
	113_A		4_R	
	114_S		4_R	
1jtr [.10]	107_S		4_K	
	108_G		4_K	
	109_F	62_R 65_Q 66_K 69_G	4_K	
	113_A		$\mathbf{4}_K$	
	114_S		4_K	
1fo0 [.14]	110_Y	65_Q		
	111_G	65_Q		
	112.1_G	65_Q		
1kj2 [.11]	108_Y	62_R		
	109_Q	63_E $\mathbf{66}_K$	1_K $\mathbf{2}_V$ 3_I 4_T	70_Y 73_T
	110_G	66_K	4_T	
	114_R	$\mathbf{65}_Q$ 68_K 69_G 72_Q		

(B) V-BETA CDR-IMGT interactions

V-BETA CDR1-IMGT				
PDB	CDR1	G-ALPHA1	Peptide	G-ALPHA2
1ao7 [5.]	37_E		$\mathbf{8}_Y$	
1oga [5.]	37_D		$\mathbf{8}_T$	58_K
1mi5 [5.]	37_V	76_E 80_N		
	38_S	76_E		
1lp9 [5.]	37_D	$\mathbf{72}_Q$ 76_V		
	38_Y	69_A 73_T	6_F	
1g6r [5.]	28_N		6_Y	58_K
	29_H		6_Y	61_Q $61A_A$
	37_N		$\mathbf{6}_Y$ 7_G 8_L	58_K
	38_N		6_Y	
1jtr [5.]	27_N			61_Q
	28_N		6_Y	$\mathbf{58}_K$ 61_Q
	29_H		6_Y	58_K $61A_A$
	37_N		$\mathbf{6}_Y$ $\mathbf{7}_S$ 8_V	58_K
	38_N		6_Y	

(continued)

Table 3. (continued) V-BETA CDR1-IMGT

PDB	CDR1	G-ALPHA1	Peptide	G-ALPHA2
1fo0 [6.]	29_{2}	61_{2}		
	38_{W}	76_{V}	7_{T}	
1kj2 [6.]	29_{Q}			58_{K} 59_{W} 61_{Q} $61A_{A}$
	36_{Y}			$61A_{A}$
	37_{P}		7_{D}	
	38_{W}	69_{G} 72_{Q}		

V-BETA CDR2-IMGT

PDB	CDR2	G-ALPHA1	Peptide	G-ALPHA2
1bd2 [.6.]	65_{I}	72_{Q}		
1oga [.6.]	57_{Q}	69_{A}	$\mathbf{4}_{G}$ 5_{F} $\mathbf{6}_{V}$	
	58_{I}	69_{A} 72_{Q} 73_{T} 76_{V}	6_{V} 8_{T}	
	63_{V}	72_{Q} 76_{V}		
	64_{N}	72_{Q} 75_{R}		
	65_{D}	65_{R} 68_{K} 69_{A} 72_{Q}		
1mi5 [.6.]	57_{Q}	72_{Q} 75_{R} 76_{E}		
	58_{N}	$\mathbf{79}_{R}$		
	63_{E}	$\mathbf{79}_{R}$		
1lp9 [.6.]	57_{Y}	65_{R} 68_{K} 69_{A} 72_{Q}		
	58_{V}	72_{Q}		
	65_{S}	68_{K}		
1g6r [.6.]	57_{Y}	69_{G} 70_{N} 72_{Q} 73_{S} 76_{V}		
	58_{G}	76_{V}		
	63_{A}	76_{V} 79_{R}		
	64_{G}	79_{R}		
	65_{S}	76_{V}		
1jtr [.6.]	57_{Y}	69_{G} 72_{Q} 73_{S} 76_{V}	$\mathbf{7}_{S}$	
	58_{G}	76_{V}		
	63_{A}	76_{V} $\mathbf{79}_{R}$ 80_{T}		
	64_{G}	79_{R}		
	65_{S}	72_{Q} 76_{V} 79_{R}		
1fo0 [.6.]	57_{R}	76_{V} 79_{R} 80_{T}		
	58_{S}	76_{V} $\mathbf{79}_{R}$		
	63_{P}	79_{R}		
1kj2 [.6.]	57_{R}	72_{Q} 73_{S} 76_{V}	$\mathbf{7}_{\mathbf{D}}$	
	58_{S}	72_{Q}		
	65_{D}	72_{Q}		

PDB	CDR3	G-ALPHA1	Peptide	G-ALPHA2
1ao7 [.14]	107_R		5_Y	
	109_G		6_P	
	110_L	69_A 72_Q 73_T	6_P 7_V **8_Y**	
	111_A		7_V 8_Y	$61A_A$
	112_G		5_Y 7_V	**$61A_A$** 62_H 63_V **66_Q**
	112.1_G		5_Y 7_V	$61A_A$
	113_R		5_Y	**61_A** $61A_A$ 62_H 66_Q
	114_P		5_Y	66_Q
1bd2 [.13]	108_Y		8_Y	
	109_P		6_P 7_V	
	110_G		6_P 7_V **8_Y**	
	111_G		7_V 8_Y	$61A_A$
	112_G		7_V	$61A_A$
	114_Y		5_Y 7_V	$61A_A$ 63_V 66_Q
1oga [.11]	108_S			$61A_A$
	109_R		5_F 6_V 7_F	**$61A_A$** **62_H** 63_V **66_Q**
	110_S		5_F 6_V	66_Q
	113_S		5_F	66_Q
	114_Y			**61_A** $61A_A$ 62_H
1mi5 [.11]	108_L	76_E		58_K
	109_G	76_E		
	110_Q	69_T **72_Q** **73_T** **76_E**	5_R **6_A**	
	113_A		6_A 7_Y	
	114_Y	76_E	7_Y **8_G**	58_K 59_W $61A_A$
1lp9 [.11]	109_W		5_F 6_F 7_P 8_V	58_K 59_W $61A_A$ 63_V
	110_V		5_F	$61A_A$
	113_S		5_F	
	114_Y		5_F	61_A $61A_A$ 62_H 66_Q
1g6r [.9]	107_G		6_Y	
	108_G		6_Y	$61A_A$ 63_E
	109_G		4_R 6_Y	$61A_A$ 66_R
	114_G		4_R	66_R
	115_T			$61A_A$
1jtr [.9]	107_G		6_Y	
	108_G		6_Y	**63_E** 66_R
	109_G		4_K 6_Y	66_R
	115_T			61_A
1fo0 [.12]	108_A			58_K
	109_D		6_N **7_T**	58_K 59_W
	110_R	**69_G** 70_N 72_Q **73_S**	4_D **5_F** 6_N	
	112_V		4_D 5_F **6_N**	66_R
	113_G		6_N	
	114_N		6_N	$61A_A$
1kj2 [.16]	108_A		6_I	66_R
	109_A		4_T 6_I	66_R
	110_P		4_T	

(continued)

Table 3. (continued) V-BETA CDR3-IMGT

PDB	CDR3	G-ALPHA1	Peptide	G-ALPHA2
	111_{D}		4_{T}	66_{R}
	111.1_{W}			62_{G} 65_{E} 66_{R} 69_{A}
	112.1_{A}			$61A_{A}$
	112_{S}			61_{Q} $61A_{A}$
	114_{E}			69_{A}

(C) V-ALPHA and V-BETA FR-IMGT interactions

V-ALPHA FR-IMGT

PDB	Position	G-ALPHA1	Peptide	G-ALPHA2
1ao7	2_{K}	58_{E}		
	26_{S}	58_{E}		
	82_{K}			**73_{T}** **76_{E}**
1bd2	2_{Q}	58_{E} 65_{R}		
	82_{K}			$72A_{G}$ 73_{T}
1oga	84_{R}			65_{E}
1mi5	40_{H}		7_{Y}	
	52_{Y}			62_{R}
	55_{H}		7_{Y}	**$61A_{A}$** **62_{R}**
	66_{V}			62_{R}
1lp9	82_{K}			65_{E}
1g6r	2_{Q}		4_{R}	
	55_{K}			65_{E}
1kj2	82_{K}			76_{E}

V-BETA FR-IMGT

PDB	Position	G-ALPHA1	Peptide	G-ALPHA2
1bd2	55_{Y}	65_{R}		
	67_{D}	68_{K}		
1oga	67_{Q}	65_{R}		
1mi5	55_{Y}	72_{Q} 76_{E}		
	66_{L}	72_{Q} 75_{R}		
1lp9	55_{Y}	65_{R}		
	67_{E}	**65_{R}** 68_{K}		
1g6r	67_{E}	72_{Q}		
	84_{Q}			58_{K}
1jtr	67_{E}	72_{Q}		
	84_{Q}			**58_{K}**

Table 4. V-ALPHA and V-BETA CDR interactions with MHC-II. TR positions in bold indicate hydrogen bonds. Three dimensional (3D) structures are from IMGT/3Dstructure-DB (Kaas et al. 2004), http://imgt.cines.fr. Lengths of the CDR-IMGT are shown within brackets. Amino acids are shown in the one-letter code. Sequences of the peptides are reported in Table 1, sequences of the TR V-ALPHA and V-BETA domains in Fig. 3, and sequences of the MHC-II G-ALPHA and G-BETA in Fig. 4. (A) V-ALPHA CDR-IMGT interactions. (B) V-BETA CDR-IMGT interactions. (C) V-ALPHA and V-BETA FR-IMGT interactions.

(A) V-ALPHA CDR-IMGT interactions

V-ALPHA CDR1-IMGT

PDB	Position	G-ALPHA	Peptide	G-BETA
1j8h [6.]	28_S		2_K	76_H
	29_V		2_K 4_V	76_H
	36_P		4_V	$72A_T$ 76_H
	38_Y			$72A_T$
1d9k [6.]	27_D		3_S	
	28_S			$72A_T$ 76_H
	29_T		3_S 4_H $\mathbf{5}_R$	$72A_T$ 76_H
	36_F		5_R	$72A_T$
	37_D		$\mathbf{5}_R$ 8_I	$\mathbf{66}_R$ 69_A $72A_T$
	38_Y			66_R

V-ALPHA CDR2-IMGT

PDB	Position	G-ALPHA	Peptide	G-BETA
1j8h [.7.]	57_T			$\mathbf{65}_E$
	58_S			69_A $72A_T$
	59_A			65_E
1d9k [.6.]	57_S			65_E 66_R 69_A
	58_L			69_A 72_D $72A_T$
	59_V			65_E 66_R 68_R 69_A
	63_S			65_E

V-ALPHA CDR3-IMGT

PDB	Position	G-ALPHA	Peptide	G-BETA
1j8h [.13]	108_E	63_E	$\mathbf{2}_K$ 4_V	
	110_P		7_N	66_Q
	111_F		7_N 9_L	62_D 63_L 66_Q
	114_E	66_G 69_A 70_N	5_K	
1d9k [.10]	107_T			66_R
	108_G		5_R 8_I	66_R
	109_S	69_Q	8_I	66_R
	113_F	$\mathbf{69}_Q$ 73_T	8_I 9_E 10_W 11_E	63_Y 66_R
	114_N	69_Q		66_R
	115_K	65_Q		

(continued)

Table 4. (continued)

(B) V-BETA CDR-IMGT interactions

V-BETA CDR1-IMGT

PDB	Position	G-ALPHA	Peptide	G-BETA
1j8h [5.]	27_M		10_K	
	28_D	76_A	**10_K**	
	29_H		10_K	
	37_E	72_A 73_V 76_A	**10_K**	
	38_N	69_A		
1d9k [5.]	37_N	76_H		
	38_N	**69_Q**		

V-BETA CDR2-IMGT

PDB	Position	G-ALPHA	Peptide	G-BETA
1j8h [.6.]	57_Y	65_Q 66_G 68_L 69_A 72_A		
	58_D	68_L 72_A **75_K**		
	65_M	43_K 68_L		
1d9k [.6.]	57_Y	**65_Q** 66_G 68_L 69_Q 72_A		

V-BETA CDR3-IMGT

PDB	Position	G-ALPHA	Peptide	G-BETA
1j8h [.12]	108_S	73_V	10_K	
	109_T	69_A 70_N 73_V	5_K 7_N 8_T	
	110_G	73_V	8_T 9_L **10_K**	
	112_L		10_K	58_Y
	113_P			$61A_Q$ 62_D 63_L
1d9k [.11]	108_G		11_E	
	109_Q		11_E	58_Y 63_Y
	110_G		10_W 11_E	**63_Y** 66_R
	113_R			61_K **62_Q** 63_Y 65_E 66_R
	114_A			66_R

(C) V-ALPHA and V-BETA FR-IMGT interactions

V-ALPHA FR-IMGT

PDB	Position	G-ALPHA	Peptide	G-BETA
1j8h	55_K			62_D
1d9k	82_K			**72_D**

V-BETA FR-IMGT

PDB	Position	G-ALPHA	Peptide	G-BETA
1j8h	55_F	65_Q		
	66_K	43_K		
	67_E	**43_K** **65_Q**		
	84_K	72_A 76_A	10_K	
1d9k	55_Y	**65_Q**		
	66_T	**43_K**		
	67_E	**43_K** 65_Q 68_L		
	68_K	65_Q		

2.4 Conclusions

With only 18 TR/pMHC 3D structures, the atomic details of TR/pMHC interactions already show a great deal of variability. IMGT standardization is a step toward a better understanding of the mechanisms ruling TR/pMHC recognition. It will help comparing new experimentally resolved 3D structures with published data. However, the TR/pMHC interactions are far from being unravelled and the study of the TR/pMHC interactions with the other proteins of the immunological synapse will be crucial. For example, the interaction between an MHC and the CD4 considerably enhances the pMHC/TR sensibility (Irvine, Purbhoo, Krosgaard, and Davis 2002; Davis 2002). The understanding of the T cell triggering early events is subject to active studies.

Although the TR/pMHC binding represents a necessary step for the TR recognition, many factors, the TR affinity for the pMHC, the relocation of surface proteins such as CD4 or CD8 in the immunological synapse are necessary for generating the T cell activation signal. Each of these steps needs to be described and characterized so that data from different experiments can be integrated. IMGT standardization will be further extended on the IMGT Web site at http://imgt.cines.fr as new parameters become available.

2.5 Citing IMGT/3Dstructure-DB

Users are requested to cite IMGT/3Dstructure-DB (Kaas et al. 2004) and this article, and to quote the IMGT home page URL, http://imgt.cines.fr.

Acknowledgements

We are grateful to Vijay Garapati for his contribution to Tables 3 and 4 and to the IMGT® team for helpful discussion. E.D. was the holder of a doctoral grant from the Ministère de l'Education Nationale, de l'Enseignement Supérieur et de la Recherche (MENESR). K.Q. was the recipient of a doctoral grant from the MENESR and was supported for one year by a grant from the Association pour la Recherche sur le Cancer (ARC). IMGT® is a registered Centre National de la Recherche Scientifique (CNRS) mark. IMGT® has been a National RIO Bioinformatics Platform since 2001 (CNRS, INSERM, CEA, INRA). IMGT® was funded in part by the BIOMED1 (BIOCT930038), Biotechnology BIOTECH2 (BIO4CT960037), and 5th PCRDT Quality of Life and Management of Living Resources (QLG2-2000-01287) programs of the European Union and received subventions from ARC and from the Génopole-Montpellier-Languedoc-Roussillon. IMGT® is currently supported by the CNRS, the MENESR (Université Montpellier II Plan Pluri-Formation), BIOSTIC-LR2004, Région Languedoc-Roussillon, ACI-IMPBIO IMP82-2004, the Réseau National des Génopoles RNG, GIS-AGENAE, Agence Nationale de la Recherche ANR (BIOSYS06_135457), and the European ImmunoGrid project (IST-2004-0280069).

Part of this work was carried out in the frame of the European Science Foundation Scientific Network Myelin Structure and its role in autoimmunity (MARIE).

References

Adams, H.P., and Koziol, J.A. (1995) Prediction of binding to MHC class I molecules. J. Immunol. Methods 185:181-190.

Apostolopoulos, V., Yu, M., Corper, A.L., Teyton, L., Pieters, G.A., McKenzie, I.F.C., and Wilson, I.A. (2002) Crystal structure of a non-canonical low-affinity peptide complexed with MHC class I: A new approach for vaccine design. J. Mol. Biol. 318:1293-1305.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000) The Protein Data Bank. Nucleic Acids Res. 28:235-242.

Blake, J.A., Richardson, J.E., Bult, C.J., Kadin, J.A., and Eppig, J.T. (2003) MGD: The Mouse Genome Database. Nucleic Acids Res. 31:193-195.

Blythe, M.J., Doytchinova, I.A., and Flower, D.R. (2002) JenPep: A database of quantitative functional peptide data for immunology. Bioinformatics 18:434-439.

Brusic, V., Rudy, G., and Harrison, L.C. (1998) MHCPEP, a database of MHC-binding peptides: Update 1997. Nucleic Acids Res. 26:368-371.

Buslepp, J., Wang, H., Biddison, W.E., Appella, E., and Collins, E.J. (2003) A correlation between TCR Vα docking on MHC and CD8 dependence: Implications for T cell selection. Immunity 19:595-606.

Davis, M.M. (2002) A new trigger for T cells. Cell 110:285-287.

Degano, M., Garcia, K.C., Apostolopoulos, V., Rudolph, M.G., Teyton, L., and Wilson, I.A. (2000) A functional hot spot for antigen recognition in a superagonist TCR/MHC complex. Immunity 12:251-261.

Ding, Y.H., Smith, K.J., Garboczi, D.N., Utz, U., Biddison, W.E., and Wiley, D.C. (1998) Two human T cell receptors bind in a similar diagonal mode to the HLA-A2/Tax peptide complex using different TCR amino acids. Immunity 8:403-411.

Ding, Y.H., Baker, B.M., Garboczi, D.N., Biddison, W.E., and Wiley, D.C. (1999) Four A6-TCR/peptide/HLA-A2 structures that generate very different T cell signals are nearly identical. Immunity 11:45-56.

Falk, K., Rotzschke, O., Stevanovic, S., Jung, G., and Rammensee, H.G. (1991) Allele-specific motifs revealed by sequencing of self-peptides eluted from MHC molecules. Nature 351:290-296.

Garboczi, D.N., Ghosh, P., Utz, U., Fan, Q.R., Biddison, W.E., and Wiley, D.C. (1996) Structure of the complex between human T-cell receptor, viral peptide and HLA-A2. Nature 384:134-141.

Garcia, K.C., Degano, M., Pease, L.R., Huang, M., Peterson, P.A., Teyton, L., and Wilson, I.A. (1998) Structural basis of plasticity in T cell receptor recognition of a self peptide-MHC. Antigen Sci. 279:1166-1172.

Giudicelli, V., and Lefranc, M.-P. (1999) Ontology for immunogenetics: The IMGT-ONTOLOGY. Bioinformatics 15:1047-1054.

Giudicelli, V., Chaume, D., and Lefranc, M.-P. (2004) IMGT/V-QUEST, an integrated software program for immunoglobulin and T cell receptor V-J and V-D-J rearrangement analysis. Nucleic Acids Res. 32:435-440.

Giudicelli, V., Chaume, D., and Lefranc, M.-P. (2005) IMGT/GENE-DB: A comprehensive database for human and mouse immunoglobulin and T cell receptor genes. Nucleic Acids Res. 33:256-261.

Giudicelli, V., Ginestoux, C., Folch, G., Jabado-Michaloud, J., Chaume, D., and Lefranc, M.-P. (2006) IMGT/LIGM-DB, the IMGT® comprehensive database of immunoglobulin and T cell receptor nucleotide sequences. Nucleic Acids Res. 34:D781-D784.

Gulukota, K., Sidney, J., Sette, A., and DeLisi, C. (1997) Two complementary methods for predicting peptides binding major histocompatibility complex molecules. J. Mol. Biol. 267:1258-1267.

Hennecke, J., Carfi, A., and Wiley, D.C. (2000) Structure of a covalently stabilized complex of a human αβ T-cell receptor, influenza HA peptide and MHC class II molecule, HLA-DR1. EMBO J. 19:5611-5624.

Hennecke, J., and Wiley, D.C. (2002) Structure of a complex of the human alpha/beta T cell receptor (TCR) HA1.7, influenza hemagglutinin peptide, and major histocompatibility complex class II molecule, HLA-DR4 (DRA*0101 and DRB1*0401): Insight into TCR cross-restriction and alloreactivity. J. Exp. Med. 195:571-581.

Irvine, D.J., Purbhoo, M.A., Krosgaard, M., and Davis, M.M. (2002) Direct observation of ligand recognition by T cells. Nature 419:845-849.

Kaas, Q., Ruiz, M., and Lefranc, M.-P. (2004) IMGT/3Dstructure-DB and IMGT/StructuralQuery, a database and a tool for immunoglobulin, T cell receptor and MHC structural data. Nucleic Acids Res. 32:208-210.

Kaas, Q., and Lefranc, M.-P. (2005) T cell receptor/peptide/MHC molecular characterization and standardized pMHC contact sites in IMGT/3Dstructure-DB. In Silico Biol. 5:505-528.

Kjer-Nielsen, L., Clements, C.S., Purcell, A.W., Brooks, A.G., Whisstock, J.C., Burrows, S.R., McCluskey, J., and Rossjohn, J. (2003) A structural basis for the selection of dominant αβ T cell receptors in antiviral immunity. Immunity 18:53-64.

Lawrence, M.C., and Colman, P.M. (1993) Shape complementarity at protein/protein interfaces. J. Mol. Biol. 234:946-950.

Lefranc, M.-P., and Lefranc, G. (2001) *The T cell receptor FactsBook*. Academic Press, London, 398.

Lefranc, M.-P. (2003a) IMGT, the international ImMunoGeneTics information system® (http://imgt.cines.fr). In: B.K.C. Lo (Ed.), *Antibody Engineering: Methods and Protocols,* 2nd edition. Methods in Molecular Biology. Humana Press, Totowa, NJ, 248, pp. 27-49.

Lefranc, M.-P., Pommié, C., Ruiz, M., Giudicelli, V., Foulquier, E., Truong, L., Thouvenin-Contet, V., and Lefranc, G. (2003b) IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. Dev. Comp. Immunol. 27:55-77.

Lefranc, M.-P. (2004a) IMGT-ONTOLOGY and IMGT databases, tools and Web resources for immunogenetics and immunoinformatics. Mol. Immunol. 40:647-660.

Lefranc, M.-P., Giudicelli, V., Ginestoux, C., Bosc, N., Folch, G., Guiraudou, D., Jabado-Michaloud, J., Magris, S., Scaviner, D., Thouvenin, V., Combres, K., Girod, D., Jeanjean, S., Protat, C., Monod, Y.M., Duprat, E., Kaas, Q., Pommié, C., Chaume, D., and Lefranc, G. (2004b) IMGT-ONTOLOGY for immunogenetics and immunoinformatics. In Silico Biol. 4:17-29.

Lefranc, M.-P., Clément, O., Kaas, Q., Duprat, E., Chastellan, P., Coelho, I., Combres, K., Ginestoux, C., Giudicelli, V., Chaume, D., and Lefranc, G. (2005a) IMGT-Choreography for Immunogenetics and Immunoinformatics. In Silico Biol. 5:6.

Lefranc, M.-P., Duprat, E., Kaas, Q., Tranne, M., Thiriot, A., and Lefranc, G. (2005b) IMGT unique numbering for MHC groove G-DOMAIN and MHC superfamily (MhcSF) G-LIKE-DOMAIN. Dev. Comp. Immunol. 29:917-938.

Lefranc, M.-P., Giudicelli, V., Kaas Q., Duprat, E., Jabado-Michaloud, J., Scaviner, D., Ginestoux, C., Clément, O., Chaume, D., and Lefranc, G. (2005c) IMGT, the international ImMunoGeneTics information system. Nucleic Acids Res. 33:D593-D597.

Lefranc, M.-P., Pommié, C., Kaas, Q., Duprat, E., Bosc, N., Guiraudou, D., Jean, C., Ruiz, M., Da Piedade, L., Rouard, M., Foulquier, E., Thouvenin, V., and Lefranc, G. (2005d) IMGT unique numbering for immunoglobulin and T cell receptor constant domains and Ig superfamily C-like domains. Dev. Comp. Immunol. 29:185-203.

Lesk, A.M., and Chothia, C. (1982) Evolution of proteins formed by β-sheets. II. The core of the immunoglobulin domains. J. Mol. Biol. 160:325-342.

Luz, J.G., Huang, M., Garcia, K.C., Rudolph, M.G., Apostolopoulos, V., Teyton, L., and Wilson, I.A. (2002) Structural comparison of allogeneic and syngeneic T cell receptor-peptide-major histocompatibility complex complexes: A buried alloreactive mutation subtly alters peptide presentation substantially increasing V(β) interactions. J. Exp. Med. 195:1175-1186.

Mandelboim, O., Bar-Haim, E., Vadai, E., Fridkin, M., and Eisenbach, L. (1997) Identification of shared tumor-associated antigen peptides between two spontaneous lung carcinomas. J. Immunol. 159:6030-6036.

Rammensee, H.G., Bachmann, J., Emmerich, N.P.N., Bachor, O.A., and Stevanovie, S. (1999) SYFPEITHI: Database for MHC ligands and peptide motifs. Immunogenetics 50:213-219.

Reinherz, E.L., Tan, K., Tang, L., Kern, P., Liu, J., Xiong, Y., Hussey, R.E., Smolyar, A., Hare, B., Zhang, R., Joachimiak, A., Chang, H.C., Wagner, G., and Wang, J. (1999) The crystal structure of a T cell receptor in complex with peptide and MHC class II. Science 286:1913-1921.

Reiser, J.B., Darnault, C., Guimezanes, A., Gregoire, C., Mosser, T., Schmitt-Verhulst, A.M., Fontecilla-Camps, J.C., Malissen, B., Housset, D., and Mazza, G. (2000) Crystal structure of a T cell receptor bound to an allogeneic MHC molecule. Nat. Immunol. 1:291-297.

Reiser, J.B., Gregoire, C., Darnault, C., Mosser, T., Guimezanes, A., Schmitt-Verhulst, A.M., Fontecilla-Camps, J.C., Mazza, G., Malissen, B., and Housset, D. (2002) A T cell receptor CDR3β loop undergoes conformational changes of unprecedented magnitude upon binding to a peptide/MHC class I complex. Immunity 16:345-354.

Reiser, J.B., Darnault, C., Gregoire, C., Mosser, T., Mazza, G., Kearney, A., van der Merwe, P.A., Fontecilla-Camps, J.C., Housset, D., and Malissen, B. (2003) CDR3 loop flexibility contributes to the degeneracy of TCR recognition. Nat. Immunol. 4:241-247.

Robinson, J., Waller, M.J., Parham, P., de Groot, N., Bontrop, R., Kennedy, L.J., Stoehr, P., and Marsh, S.G. (2003) IMGT/HLA and IMGT/MHC: Sequence databases for the study of the major histocompatibility complex. Nucleic Acids Res. 31:311-314.

Rudolph, M.G., Luz, J.G., and Wilson, I.A. (2002) Structural and thermodynamic correlates of T cell signaling. Annu. Rev. Biophys. Biomol. Struct. 31:121-149.

Scott, C.A., Peterson, P.A., Teyton, L., and Wilson, I.A. (1998) Crystal structures of two I-Ad-peptide complexes reveal that high affinity can be achieved without large anchor residues. Immunity 8:319-329.

Sim, B.C., Zerva, L., Greene, M.I., and Gascoigne, N.R. (1996) Control of MHC restriction by TCR Vα CDR1 and CDR2. Science 273:963-964.

Singh, H., and Raghava, G.P. (2003) ProPred1: Prediction of promiscuous MHC class-I binding sites. Bioinformatics 19:1009-1014.

Stewart-Jones, G.B.E., McMichael, A.J., Bell, J.I., Stuart, D.I., and Jones, E.Y. (2003) A structural basis for immunodominant human T cell receptor recognition. Nat. Immunol. 4:657-663.

Vasmatzis, G., Cornette, J., Sezerman, U., and DeLisi, C. (1996a) TcR recognition of the MHC-peptide dimer: Structural properties of a ternary complex. J. Mol. Biol. 261:72-89.

Vasmatzis, G., Zhang, C., Cornette, J.L., and DeLisi, C. (1996b) Computational determination of side chain specificity for pockets in class I MHC molecules. Mol. Immunol. 33: 1231-1239.

Wain, H.M., Bruford, E.A., Lovering, R.C., Lush, M.J., Wright, M.W., and Povey, S. (2002) Guidelines for human gene nomenclature. Genomics 79:464-470.

Wang, J.H., Meijers, R., Xiong, Y., Liu, J.H., Sakihama, T., Zhang, R., Joachimiak, A., and Reinherz, E.L. (2001) Crystal structure of the human CD4 N-terminal two-domain fragment complexed to a class II MHC molecule. Proc. Natl. Acad. Sci. USA 98:10799-10804.

Wang, J.H., and Reinherz, E.L. (2002) Structural basis of T cell recognition of peptides bound to MHC molecules. Mol. Immunol. 38:1039-1049.

Yousfi Monod, M., Giudicelli, V., Chaume, D., and Lefranc, M.-P. (2004) IMGT/JunctionAnalysis: The first tool for the analysis of the immunoglobulin and T cell receptor complex V-J and V-D-J junctions. Bioinformatics 20:I379-I385.

Zhang, C., Anderson, A., and DeLisi, C. (1998) Structural principles that govern the peptide-binding motifs of class I MHC molecules. J. Mol. Biol. 281:929-947.

Chapter 3

Structural Immunoinformatics

Shoba Ranganathan,[1,3] Joo Chuan Tong,[2] and Tin Wee Tan[3]

[1] Department of Chemistry and Biomolecular Sciences & Biotechnology Research Institute, Macquarie University, Sydney, NSW 2109, Australia, shoba.ranganathan@mq.edu.au
[2] Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613, jctong@i2r.a-star.edu.sg
[3] Department of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, 8 Medical Drive, MD7 #02-03, Singapore 117597, tinwee@bic.nus.edu.sg

Abstract. Normal adaptive immune responses operate under major histocompatibility complex (MHC) restriction by binding to specific short antigenic peptides. Sequence-structure-function information is critical in facilitating the understanding of principles governing MHC-specific peptide recognition and binding. Three-dimensional structures of bound peptide ligands to MHC receptors are today characterized in great number using X-ray crystallography, offering a rich source of information for structural analysis. By utilizing information derived from available experimental structures, it is possible to predict binders for alleles that have not been studied extensively and offers an alternative to sequence-based approaches that require a large dataset for training. This chapter will introduce the use of structural descriptors, as well as comparative modeling and docking techniques for predicting whether a peptide sequence can bind to a specific MHC allele.

3.1 Introduction

The binding of peptide ligands to MHC molecules plays a key role in the activation of normal adaptive immune responses and an intricate theoretical problem that remains unsolved. For an MHC molecule to recognize antigenic peptides, geometric and electrostatic complementarity between the receptor and ligand is essential for the formation of chemical bonds between their functional groups, which in turn determines the stability of the complex. In this context, the introduction of structural information can greatly facilitate our understanding of how well a peptide ligand can associate with an MHC molecule.

In recent years, an increasing number of protein structures have been experimentally determined and deposited in the Protein Data Bank (PDB; Bernstein, Koetzle, Williams, Meyer, Brice, Rodgers, Kennard, Shimanouchi, and Tasumi 1977), providing a wealth of information for structural analysis and prediction. Together with the development of new structural modeling and docking techniques, the use of structure-based approaches to predict potential T-cell epitopes is increasingly successful, often producing modeled structures accurate to within 2.00Å

root-mean-square deviation RMSD (Tong, Tan, and Ranganathan 2004) from the experimental crystal structure. This chapter introduces the use of three-dimensional experimental and modeled structures for peptide-MHC (pMHC) prediction. First, we provide a brief description of the structural characteristics of pMHC complex followed by an introduction on the use of structural descriptors to characterize pMHC interfaces. Following this, we introduce the use of protein structure prediction techniques to generate models of pMHC complexes. Finally, we cover some available pMHC structural resources from the Internet.

3.2 Structural Features of MHC Peptides

Two main classes of MHC molecules, class I and class II, are identified that are responsible for presenting epitopes to cytotoxic T cells (T_c) and helper T cells (T_h), respectively. Each class has different binding characteristics: (i) class I ligands with a typical length of between 8 and 12 amino acids are enclosed within the MHC binding groove with both termini tethered toward the base of the binding groove and the center loop bulges out to interact with T_c (Fig. 1A); (ii) class II ligands (Fig. 1B) are more variable with both termini extending out of the groove thus permitting T_h to bind to ligands of a longer length (between 9 and 25 amino acids).

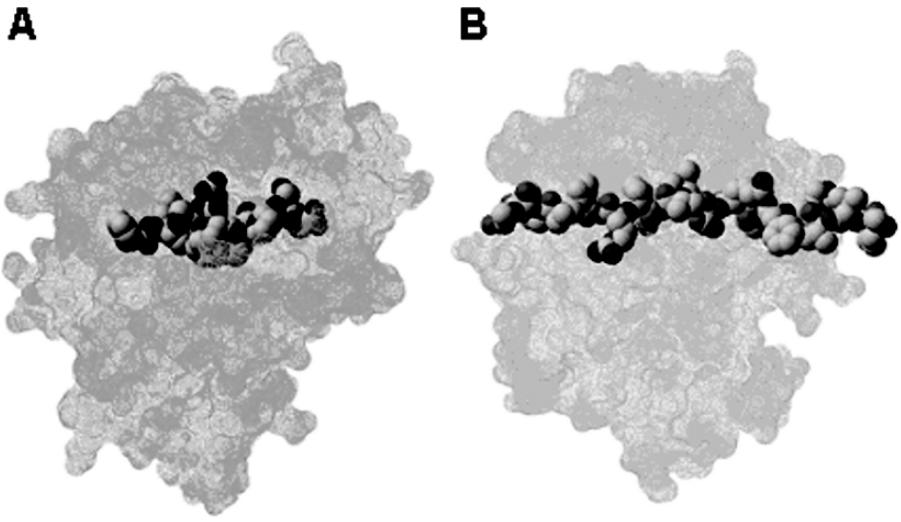


Fig. 1. (A) Class I MHC-peptide complex. (B) Class II MHC-peptide complex.

3.3 MHC-Peptide Interaction Parameters

The forces involved in protein-protein interactions are non-covalent and therefore reversible, and are generally effective over short distances. These include: (i) van der Waals (attractive and repulsive) interaction, (ii) electrostatic interactions (including hydrogen bonds), and (iii) hydrophobic interactions. It is a balance of these interactions against interactions with solvent that determines the stability of a protein. Some interaction parameters have been identified as being significant for the characterization of pMHC interface (Kangueane, Sakharkar, Kolatkar, and Ren 2001; Tong, Kong, Tan, and Ranganathan 2006) and can be calculated from the three-dimensional coordinates of a complex.

3.3.1 Interface Area between Peptide and MHC

The hydrophobic effect is considered to be one of the most significant forces that drive protein folding. A linear correlation exists between the hydrophobic free energy of transfer from polar to hydrophobic environment and the change in solvent accessible surface area (ΔASA) upon complexation (Chothia and Janin 1975). Thus, knowledge of the surface area of a complex interface in direct contact with solvent may provide an indication of the binding strength. The accessible surface area can be measured by tracing out the maximum permitted van der Waals contact that is covered by the center of a water molecule as it rolls over the surface of the protein. Interface area for class I pMHC complexes was defined as the mean ΔASA on complexation when going from a monomeric MHC molecule to a dimeric pMHC complex state and calculated as half the sum of the total ΔASA for both molecules for each type of complex. The mean ΔASA for class I pMHC complexes is 903.30 ± 260.90 $Å^2$. Similarly, the interface area for class II pMHC complexes was defined as the ΔASA when going from a dimeric MHC molecule to a trimeric state. The corresponding ΔASA in class II complexes is 894.40 ± 364.00 $Å^2$.

3.3.2 Intermolecular Hydrogen Bonds

Hydrogen bonds are major contributors to the selectivity and stability of protein-protein complexes. It involves three atoms, a donor electronegative atom to which the hydrogen is bound, and an acceptor electronegative atom in close proximity. The typical observed hydrogen bond distance is approximately 2.60 to 3.10 Å (1.00 to 1.20 Å between donor and hydrogen and 1.60 to 2.00 Å between acceptor and hydrogen). For such bonding to be significant, both electronegative atoms must be derived from the group: F, N, and O (Morrison and Boyd 1992). Only hydrogen bonded to any of these three elements is sufficiently positive, and only these elements are sufficiently negative for the required attraction to exist due to the high concentration of negative charge on their small atoms. Hydrogen bonds are directional and can control and restrict the geometry of the interactions between side-chains. In general, the strength of hydrogen bonds increases with decreasing bond length.

3.3.3 Complementarity between Surfaces

3.3.3.1 Gap Index

One essential feature in receptor-ligand binding is the electrostatic and geometric complementarity observed between associating molecules. Here, we introduce the use of gap index (Jones and Thornton 1996) as a means to evaluate complementarity of interacting interfaces:

$$\text{Gap index (Å)} = \frac{\text{gap volume between pMHC (Å}^3\text{)}}{\text{interface ASA (Å}^2\text{) (per complex)}} \quad (1)$$

The mean gap indices for class I and class II pMHC complexes are 0.95 ± 0.24 Å and 1.12 ± 0.20 Å, respectively (Kangueane et al. 2001). The results indicate that the interacting surfaces in pMHC complexes are significantly complementary. On an average, the gap index is higher in class II complexes than in class I complexes. This implies that the interface area of class I complexes is greater than its corresponding gap volume. On the contrary, the mean gap volume is greater than the interface area in class II complexes. Not much difference can be identified in the gap index between complexes of different alleles in both class I and class II complexes.

3.3.3.2 Gap Volume

The gap volume between the MHC and the peptide in each complex can be computed using the SURFNET program (Laskowski 1991), which provides an estimate of the volume enclosed by the two interacting molecular subunits. The algorithm places a series of spheres (maximum radius 5.00 Å) midway between the surfaces of each pair of subunit atoms, such that its surface is in contact with the surfaces of the atoms in the pair. The size of each sphere is reduced accordingly whenever it is intercepted by other atoms and subsequently discarded if it falls below a minimum allowed radius (1.00 Å). The sizes of all the remaining allowable gap-spheres are subsequently used to compute the gap volume between the two subunits.

3.4 Structural Prediction Techniques

3.4.1 Homology Modeling

The use of known homologous protein structure(s) to predict the unknown structure of a related amino acid sequence represents one of the most reliable strategies for model building of proteins (Swindells and Thornton 1991), often producing model structures with accuracy to within 2.00 Å RMSD from the actual crystal structure. Homology modeling involves a series of steps, with each step depending on the success of the preceding one. A comprehensive coverage of the homology modeling

procedure is described in an earlier study (Sali and Blundell 1993). A brief outline of the comparative modeling process is as follows:

i. Searching for a suitable template structure from the PDB using the target sequence as a query. The template structure is a sequence with known structure that is significantly similar to the target sequence.
ii. Align the target sequence with the template sequence to maximize the structural similarity using either a local-similarity dynamic programming approach (Smith and Waterman 1981) or a global-similarity approach (Needleman and Wunsch 1970).
iii. Substitute amino acid side chains in the template with the corresponding ones from the target sequence.
iv. Model weakly conserved regions such as insertions/deletions and loops between the target and template sequences.
v. Perform energy minimization to improve the stereochemistry of the modeled structure.

Generally, modeled structures are as close to the target structure as their templates (Sanchez and Sali 1997). This is a nontrivial achievement due to the existence of many residue substitutions, deletions, and insertions between the target and template sequences that must be taken into account during comparative modeling. When several templates are selected for modeling, it is possible to generate a model that is significantly closer to the target structure than any of the templates as the model tends to inherit the most conserved regions from each template (Sanchez and Sali 1997).

3.4.2 Docking Algorithm

Computer-simulated ligand binding or docking is a powerful technique for investigating intermolecular interactions. In general, the purpose of docking simulation is twofold: (i) to find the most probable translational, rotational, and conformational juxtaposition of a given ligand-receptor pair and (ii) to evaluate the relative goodness-of-fit or how well a ligand can bind to the receptor. Here, we introduce a highly accurate docking protocol for the modeling of bound peptide ligands to the MHC receptor. The methodology presented here is applicable to the design of both subtype-specific vaccines as well as promiscuous peptide epitopes.

3.4.3 The Peptide Docking Procedure

Beginning with the sequence of the ligand for which the structure is to be generated (herein referred to as the target peptide), and the availability of the target MHC receptor structure, our docking protocol consists of three essential steps: (i) rigid docking of residues at the ends of binding groove; (ii) loop closure of central residues by satisfaction of spatial constraints; (iii) followed by *ab initio* refinements of backbone and ligand interacting side chain. The general flow of the docking protocol is illustrated in Fig. 2.

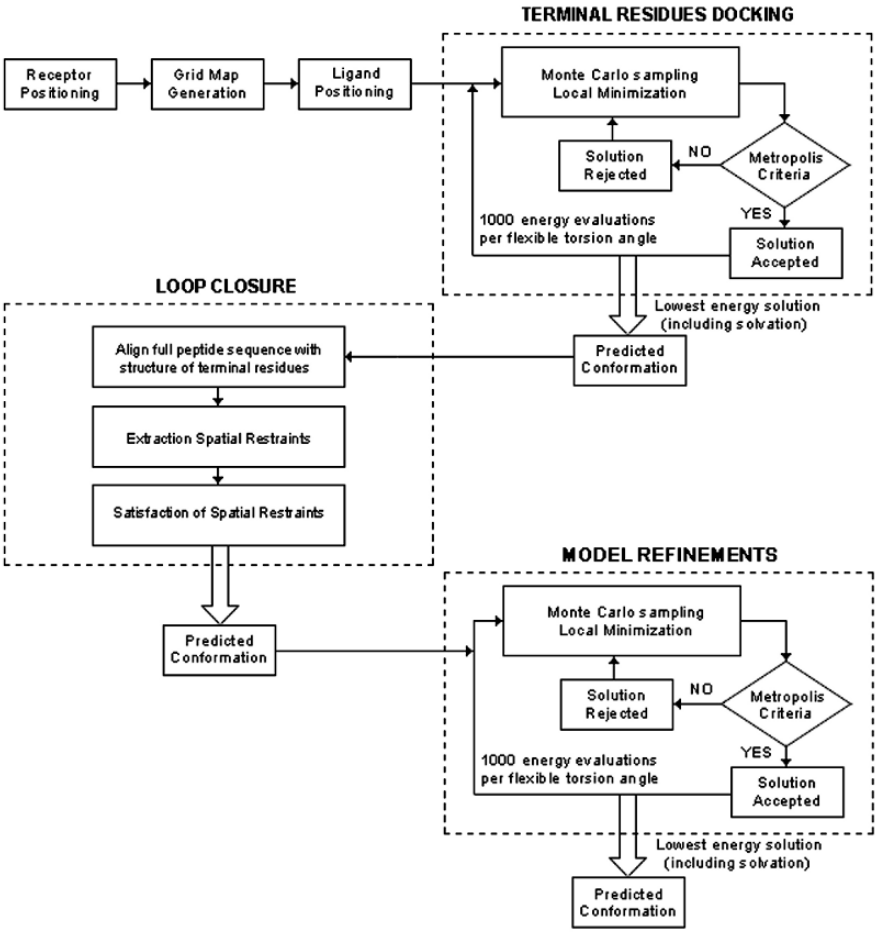


Fig. 2. Flowchart of the docking procedure used in this work.

3.4.3.1 Rigid Docking of Residues at the Ends of Binding Groove

The main problem in docking simulation is to enumerate the number of combinations for two molecules within an enclosed sampling space. There are six degrees of global-rotational and translational freedom of one molecule relative to the other, as well as one internal dihedral rotation per rotational bond. A full search on the conformational space increases exponentially with increasing molecule size and sampling space. As such, a key challenge in pMHC docking simulation is to minimize the conformational search space of ligand within the large sampling space enclosed by the MHC binding groove. A possible approach is to identify suitable base or anchor fragments (herein referred to as probes) for initiating docking simulations. A probe must satisfy two criteria: (i) the anchor must have sufficient contact with the receptor and (ii) the

structure of the anchor must be highly conserved. Peptide fragments at the end of MHC binding groove with mean backbone Cα RMSD within 0.15 ± 0.14 Å (Tong et al. 2004) are ideal for such purpose.

A fast soft-interaction energy function (Fernández-Recio, Totrov, and Abagyan 2002) is adopted to model each probe to the receptor. This is performed using an Internal Coordinate Mechanics (ICM; Abagyan and Totrov 1999) global optimization algorithm; with flexible ligand interface side chains and a grid map representation of the receptor energy localized to small cubic regions of 1.00-Å radius from the backbone of each probe. Each probe performs a random walk within their respective grid map. At each random step, the side-chain torsions were changed using a Biased Monte Carlo procedure, which begins by pseudo-randomly selecting a set of torsion angles in the probe and subsequently finding the local energy minimum about those angles. New conformations are adopted upon satisfaction of the Metropolis criteria with probability min(1, exp[-Δ*G/RT*]), where R is the universal gas constant and *T* is the absolute temperature of the simulation. Loose restraints were imposed on the positional variables of the ligand molecule to keep it close to the starting conformation. The stimulation temperature was set to 300 K. The optimal energy function used during stimulations consisted of the internal energy of the probe and the intermolecular energy based on the same optimized potential maps used in the docking step:

$$E = E_{\mathrm{Hvw}} + E_{\mathrm{Cvw}} + 2.16E_{\mathrm{el}}^{\mathrm{solv}} + 2.53E_{\mathrm{hb}} + 4.35E_{\mathrm{hp}} + 0.20E_{\mathrm{solv}} \quad (2)$$

The internal energy included internal van der Waals interactions, hydrogen bonding and torsion energy calculated with ECEPP/3 parameters, and the Coulomb electrostatic energy with a distance-dependent dielectric constant (e=4r). The configurational entropy of sidechains and the surface-based solvation energy were included in the final energy to select the best-refined solutions.

3.4.3.2 Loop Closure of Center Residues

In this stage, an initial conformation of the central loop is generated by satisfaction of spatial constraints (Sali and Blundell 1993) based on the allowed subspace for backbone dihedrals in accordance with the conformations of peptides docked into the ends of the binding groove. This is performed in three steps: (i) Distance and dihedral angle restraints on the entire peptide sequence are derived from its alignment with the sequences of probes docked into the binding groove. (ii) The restraints on spatial features of the unknown center residues are derived by extrapolation from the known 3D structures of probes in the alignment, expressed as probability density functions. Stereochemical restraints include bond distances, bond angles, planarity of peptide groups and side-chain rings, chiralities of Cα atoms and sidechains, van der Waals contact distances and the bond lengths, bond angles, and dihedral angles of cysteine disulfide bridges. (iii) Spatial restraints on the unknown center residues are satisfied by optimization of the molecular probability density function using variable target function technique that applies the conjugate gradients algorithm to positions of all nonhydrogen atoms.

3.4.3.3 Refinements of Ligand Backbone and Interacting Side Chain

To improve the accuracy of the initial model, partial refinement was performed for both the ligand backbone and side chain, using ICM Biased Monte Carlo procedure (Abagyan and Maxim 1999). Initial stages of refinements attempt to overcome the penalty derived from the initial rigid docking of terminal residues by introducing partial flexibility to the ligand backbone. Restraints were imposed upon the positional variables of the Cα atoms of probes to keep it close to the starting conformation. The energy function adopted for this refinement step is:

$$E = E_{\text{vw}} + E_{\text{hbonds}} + E_{\text{torsions}} + E_{\text{electr}} + E_{\text{solv}} + E_{\text{entropy}} \quad (3)$$

Refinements of ligand and receptor side-chain torsions in the vicinity of 4.00 Å from the receptor were performed upon the final backbone structure.

3.5 Application of Docking Protocol

We will now illustrate the application of our protocol for the discrimination of binders/nonbinders from MHC class II alleles. Concretely, we discuss the docking of a 15-residue peptide ERVICPISSVPGNLA into the binding groove of associated and nonassociated MHC class II alleles DRB1*0402 and DRB1*0406 respectively (Tong, Bramson, Kanduc, Sinha, and Ranganathan 2006).

The core residues of each 15-residue peptide were generated using a sliding window (size 9) to eliminate any bias in selecting core peptides based on sequence patterns before the start of docking simulations. Docking of core peptide residues is performed using the docking procedure earlier described followed by *ab initio* modeling of flanking residues. Our models present the best fit of each peptide into the binding cleft of each disease-associated and non-disease-associated allele based on the following criteria: (i) pattern of hydrogen bonding to the MHC molecule, (ii) pattern of hydrophobic burial of peptide side chains, and (iii) the absence of atomic clashes or repulsive contacts.

In this example, pMHC residues were considered to be in contact if at least one pair of their nonhydrogen ("heavy") atoms was found to be within 4.00-Å radius (Fischer and Marquesee 2000). Intrapeptide interactions and intra-MHC interactions were not considered as they have minor influence on backbone structure. Any atom in the peptide and any atom in the MHC were considered to be experiencing atomic clash if their separation is below 2.00 Å (Samudrala and Moult 1997) for non-hydrogen atoms and below 1.60 Å for atoms participating in hydrogen bonds (Samanta, Bahadur, and Chakrabarti 2002; Wallace, Laskowski, and Thornton 1995). Out of seven possible combinations of core peptide residues for the peptide, only one conformation successfully docked into the binding cleft of DRB1*0402 without any atomic clashes or repulsive interactions. In contrast, atomic clashes are experienced in nonassociated allele DRB1*0406 (Fig. 3).

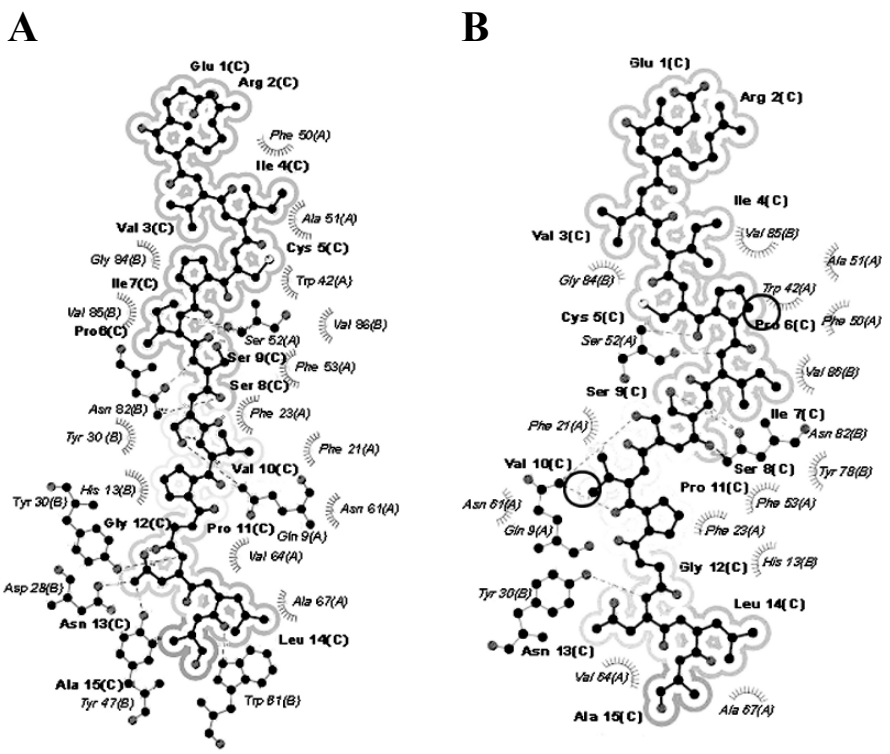


Fig. 3. Comparison of Dsg3 963-977 (ERVICPISSVPGNLA) peptide within the binding grooves of associated and non-associated alleles at 4.00Å. (A) Strongly associated allele *0402. No atomic clash is detected in the modeled peptide-MHC complex. (B) Non-associated allele *0406. Buried peptide residues are shaded grey (in increasing density) and regions of atomic clash occurring between peptide and pocket residues are circled in black.

3.6 Available Resources

A comprehensive dataset to facilitate the sequence-structure-function mapping in peptide binding by MHC receptors is essential for structural analysis and development of predictive algorithms in computational immunology. Listed in Table 1 below are some pMHC structural databases that are freely available for use or download.

Table 1. Some existing pMHC structure databases freely available for use or download.

Name	Description	URL
PDB	Worldwide repository for the processing and distribution of 3D structure data of large molecules of proteins and nucleic acids.	http://www.rcsb.org/pdb/
MPID	A manually curated database on pMHC interactions containing computed interaction parameters relevant to each complex.	http://surya.bic.nus.edu.sg/mpid
MPID-T	A manually curated database containing computed atomic interaction information on TCR-pMHC and pMHC complexes obtained from PDB.	http://surya.bic.nus.edu.sg/mpidt
FIMM	Database containing HLA 3D experimental structures obtained from PDB and models generated by homology modeling.	http://sdmc.lit.org.sg:8080/fimm/
IMGT/3D structure-DB	Database containing annotated information on the sequences, 2D structures and 3D structures of TR and pMHC from human and other vertebrate species according to the IMGT Scientific chart.	http://imgt3d.igh.cnrs.fr/

3.7 Conclusions

In order to fully understand the phenomenon of pMHC interactions, it is necessary to introduce structural information. To date, structure-based techniques are poorly developed and are lagging far behind sequence-based prediction techniques due to higher complexity in development and computational costs. Despite their slow progress, structure-based techniques are highly promising as a predictive tool for vaccine design for its strength in predicting potential T-cell epitopes for MHC subtypes where insufficient data are available for training sequence-based procedures.

References

Abagyan, R., and Totrov, M. (1999) Ab initio folding of peptides by the optimal-bias Monte Carlo minimization procedure. J. Comput. Phys. 151:402-421.

Bernstein, F.C., Koetzle, T.F., Williams, G.J., Meyer, E.E., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., and Tasumi, M. (1977) The Protein Data Bank: A computer-based archival file for macromolecular structures. J. Mol. Biol. 112:535-542.

Chothia, C., and Janin, J. (1975) Principles of protein-protein recognition. Nature 28:705-708.

Fernández-Recio, J., Totrov, M., and Abagyan, R. (2002) Soft protein-protein docking in internal coordinates. Protein Sci. 11:280-291.

Fischer, K.F., and Marquesee, S. (2000) A rapid test for identification of autonomous folding units in proteins. J. Mol. Biol. 302:701-712.

Jones, S., and Thornton, J.M. (1996) Principles of protein-protein interactions. Proc. Natl. Acad. Sci. USA 93:13-20.

Kangueane, P., Sakharkar, M.K., Kolatkar, P.R., and Ren, E.C. (2001) Towards the MHC-peptide combinatorics. Hum. Immunol. 62:539-556.

Laskowski, R.A. (1991) SURFNET computer program (Department of Biochemistry and Molecular Biology, University College, London, England).

Morrison, R.T., and Boyd, R.N. (1992) Organic Chemistry, Sixth Edition. Prentice ─Hall.

Needleman, S.B., and Wunsch, C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. 48:443-453.

Rost, B., and Sander, C. (1994) Conservation and prediction of solvent accessibility in protein families. Proteins 20:216-226.

Sali, A., and Blundell, T.L. (1993) Comparative protein modeling by satisfaction of spatial restraints. J. Mol. Biol. 234:774-815.

Samanta, U., Bahadur, R.P., and Chakrabarti, P. (2002) Quantifying the accessible surface area of protein residues in their local environment. Protein Eng. 15:659-667.

Samudrala, R., and Moult, J. (1997) Handling context-sensitivity in protein structures using graph theory: Bona fide prediction. Proteins Suppl. 1:43-49.

Sanchez, R., and Sali, S. (1997) Evaluation of comparative protein structure modeling by MODELLER-3. Proteins Suppl. 1:50-58.

Smith, T.F., and Waterman, M.S. (1981) Identification of common molecular subsequences. J. Mol. Biol. 147:195-197.

Swindells, M.B., and Thornton, J.M. (1991) Modelling by homology. Curr. Opin. Struct. Biol. 1:219-223.

Tong, J.C., Tan, T.W., and Ranganathan, S. (2004) Modeling the structure of bound peptide ligands to major histocompatibility complex. Protein Sci. 13:2523-2532.

Tong, J.C., Kong, L., Tan, T.W., and Ranganathan, S. (2006) MPID-T: Database for sequence-structure-function information on TCR/peptide/MHC interactions. Appl. Bioinformatics 5:111-114.

Tong, J.C., Bramson, J., Kanduc, D., Sinha, A.A.,and Ranganathan, S. (2006) Modeling the bound conformation of pemphigus vulgaris-associated peptides to MHC class II DR and DQ alleles. Immunome Res. 2:1.

Wallace, A.C., Laskowski, R.A., and Thornton, J.M. (1995) LIGPLOT: A program to generate schematic diagrams of protein-ligand interactions. Protein Eng. 8:127-134.

Chapter 4

In Silico QSAR-Based Predictions of Class I and Class II MHC Epitopes

Channa K. Hattotuwagama, Irini A. Doytchinova, Pingping Guan, and Darren R. Flower

Edward Jenner Institute for Vaccine Research, Compton, Berkshire RG20 7NN, UK, darren.flower@jenner.ac.uk

Abstract. Quantitative Structure-Activity Relationship (QSAR) analysis is a cornerstone of modern informatics. Predictive computational models of peptide-Major Histocompatibility Complex (MHC) binding affinity based on QSAR technology have now become important components of modern computational immunovaccinology. Historically, such approaches were built around semiqualitative, classification methods, but these are now giving way to quantitative regression methods. We review two methods – a 2D-QSAR Additive-Partial Least Squares (PLS) and a 3D-QSAR Comparative Molecular Similarity Index Analysis (CoMISA) method – which can identify the sequence dependence of peptide binding specificity for various class I MHC alleles from the reported binding affinities (IC50) of peptide sets. The Iterative Self-Consistent (ISC) PLS-based Additive Method is a recently developed extension to the Additive method for the affinity prediction of class II peptides. The QSAR methods presented here have established themselves as immunoinformatic techniques complementary to existing methodology, useful in the quantitative prediction of binding affinity: current methods for the *in silico* identification of T-cell epitopes (which form the basis of many vaccines, diagnostics and reagents) rely on the accurate computational prediction of peptide-MHC affinity.

We review a variety of human and mouse class I and class II allele models. Studied alleles comprise HLA-A*0101, HLA-A*0201, HLA-A*0202, HLA-A*0203, HLA-A*0206, HLA-A*0301, HLA-A*1101, HLA-A*3101, HLA-A*6801, HLA-A*6802, HLA-B*3501, H2-Kk, H2-Kb, and H2-Db HLA-DRB1*0101, HLA-DRB1*0401, and HLA-DRB1*0701, I-Ab, I-Ad, I-Ak, I-As, I-Ed, and I-Ek.

In terms of reliability the resulting models represent an advance on existing methods. The peptides used in this study are available from the AntiJen database (http://www.jenner.ac.uk/AntiJen). The PLS method is available commercially in the SYBYL molecular modeling software package. The resulting models, which can be used for accurate T-cell epitope prediction, are freely available online at: http://www.jenner.ac.uk/ MHCPred.

4.1 Introduction

Quantitative Structure-Activity Relationship (QSAR) analysis, as a predictive tool of wide applicability, is one of the main cornerstones of modern cheminformatics and, increasingly, bioinformatics. Immunoinformatics, a newly emergent subdiscipline of bioinformatics, which addresses informatic problems within immunology, uses QSAR technology to tackle the crucial issue of epitope prediction. As high-throughput biology reveals the genomic sequences of pathogenic bacteria, viruses, and parasites, such prediction will become increasingly important in the post-genomic discovery of novel vaccines, reagents, and diagnostics. In order to better understand the sequence dependence of peptide-MHC (Major Histocompatibility Complex) binding of the mouse MHC, we have now used our approach to explore the amino acid preferences of various human and mouse alleles.

The products of MHC play a fundamental role in regulating immune responses. T cells recognize peptide fragments complexed with MHC molecules as antigens, a process requiring antigen degradation through complex proteolytic digestion prior to complexation. The biological role of MHC proteins is thus to bind peptides and "present" these at the cell surface for inspection by T-cell antigen receptors (TCRs). The MHC genes are grouped into two classes on the basis of their chemical structure and biological properties. The two types of MHC protein have related secondary and tertiary structure but with important functional differences. Class I molecules are composed of a heavy chain complexed to β2-microglobulin, while class II molecules consist of two chains (α and β) of similar size. Both classes of MHC molecule have similar 3D structures composed of two domains. The MHC peptide-binding site consists of a β-sheet, forming the base, flanked by two α-helices, which together form a narrow cleft or groove accommodating bound peptides. The principal differences between the two classes are the dimensions of the peptide-binding groove, which is constrained to bind 8- to 11-amino acid peptides in class I, but is open at both ends in class II, allowing much larger peptides of varying length to be bound.

Class II MHC molecules are non-covalently bonded heterodimers, called HLA-DP, HLA-DQ, and HLA-DR in humans and I-A and I-E in mice. Peptides binding to class II MHC molecules are usually 10-25 residues long, with lengths of 13-16 amino acids being the most frequently observed (Rudensky, Preston-Hurlburt, Hong, Buus, and Tschinke 1991; Hunt, Michel, Dickinson, Shabanowitz, Cox, Sakaguchi, and Appella 1992; Chicz, Urban, Lane, Gorga, Stern, Vignali, and Strominger 1992; Chicz, Urban, Gorga, Vignali, Lane, and Strominger 1993). From X-ray crystallographic data of MHC class II and TCR-peptide-MHC class II complexes (Dessen, Lawrence, Cupo, Zaller, and Wiley 1997; Hennecke and Wiley 2002), it is clear that 9 amino acids are bound in an extended conformation within the class II binding site. They are not anchored at their amino and carboxyl termini, but stretch along the binding groove, with residues accommodated by binding pockets along the cleft. Previous interpretations, reported in the literature, suggest that class II peptides have a small number of anchor residues upon which binding depends. These anchors are residues of an appropriate type, which must sit at particular spacings along the peptide in order for allele-restricted binding to occur; residues at other peptide positions

are less constrained. The side chain at peptide position P1 binds into a deep pocket while four shallow pockets bind side chains at peptide positions 4, 6, 7, and 9. The side chains at positions 2, 3, 5, and 8 point toward the T-cell receptor.

We have recently developed an immunoinformatic technique for the prediction of peptide-MHC affinities, known as the Additive Method, a 2D-QSAR technique which is based on the Free-Wilson principle (Kubinyi and Kehrhahn 1976), whereby the presence or absence of groups is correlated with biological activity. For a peptide, the binding affinity is thus represented as the sum of amino acid contributions at each position. We have extended the classical Free-Wilson model with terms which account for interactions between amino acid side chains. An Iterative Self-Consistent (ISC) Partial Least Squares (PLS)-based extension (Doytchinova and Flower 2003) of the Additive Method (Doytchinova, Blythe, and Flower 2002c; Guan, Doytchinova, Zygouri, and Flower 2003a) has also been developed for prediction of class II peptide-binding affinity and applied to human class II alleles. We now address binding to class II human and mouse alleles for peptides of up to 25 amino acids in length. The ISC additive method assumes that the binding affinity of a large peptide is principally derived from the interaction, with an MHC molecule, of a continuous subsequence of amino acids within it. The ISC is able to factor out the contribution of individual amino acids within the subsequence, which is initially identified in an iterative manner. Using literature data, we have applied the Additive Method to peptides binding to several human class I (Doytchinova et al. 2002c; Guan et al. 2003a; Hattotuwagama, Guan, Doytchinova, Zygouri, and Flower 2004) and class II alleles (Doytchinova and Flower 2003).

Three-dimensional QSARs are a technique of significant value in identifying correlations between ligand structure and binding affinity. This value is often enhanced greatly when analysed in the context of high-resolution ligand-receptor structures. In such cases, enthalpic changes – van der Waals and electrostatic interactions – and entropic changes – conformational and solvent-mediated interactions – in ligand binding can be compared with structural changes in both ligand and macromolecule, providing insight into the binding mechanism (Klebe, Abraham, and Mietzner 1994; Klebe and Abraham 1999). Although there are many molecular descriptors that account for free energy changes, 3D-QSAR techniques which use multivariate statistics to relate molecular descriptors in the space around the ligands, to binding affinities, have become preeminent because of their robustness and interpretability (Bohm, Sturzebecher, and Klebe 1999). In the case of CoMSIA (Comparative Molecular Similarity Index Analysis), a Gaussian-type functional form is used so that no arbitrary definition of cutoff threshold is required and interactions can be calculated at all grid points. The obtained values are evaluated using PLS analysis (Stahle and Wold 1988). CoMSIA allows each physicochemical descriptor to be visualized in 3D using a map, which denotes binding positions that are either "favored" or "disfavored".

Recently, CoMSIA has been used to produce predictive models for peptide binding to human MHCs: HLA-A*0201 (Doytchinova and Flower 2002a) and the HLA-A2 and HLA-A3 supertypes (Doytchinova and Flower 2002b; Guan, Doytchinova, and Flower 2003b). We show how CoMSIA has been applied to certain class I MHC alleles. These models were used both to evaluate physicochemical requirements for

binding, and to explore and define preferred amino acids within each pocket. The explanatory power of such a 3D-QSAR method is considerable, not only in its direct prediction accuracy but also in its ability to map advantageous and disadvantageous interaction potentials onto the structures of the peptides being studied. The data are highly complementary to the detailed information obtained from crystal structures of individual peptide-MHC complexes.

4.2 Methodology

4.2.1 Peptide Database

The information and data based on the peptide sequences and their binding affinities were obtained from the AntiJen database, a development of JenPep (Blythe, Doytchinova, and Flower 2002; McSparron, Blythe, Zygouri, Doytchinova and Flower 2003) [URL: http://www.jenner.ac.uk/AntiJen]. Compilations of quantitative affinity measures for peptides binding to class I and class II MHCs were carried out with known binding affinities (IC_{50}). These include human class I (HLA-A*0101, HLA-A*0201, HLA-A*0202, HLA-A*0203, HLA-A*0206, HLA-A*0301, HLA-A*1101, HLA-A*3101, HLA-A*6801, HLA-A*6802, HLA-B*3501), mouse class I (H2-K^k, H2-K^b, and H2-D^b), human class II (HLA-DRB1*0101, HLA-DRB1*0401, and HLA-DRB1*0701), and mouse class II (I-A^b, I-A^d, I-A^k, I-A^s, I-E^d and I-E^k). For class I, only nonameric peptides were included, with the exception of H2-K^b and H2-K^k, where octameric peptides were also examined. For each set of class II alleles, peptide lengths of 10 to 25 were obtained from the AntiJen database. Several QSAR methodologies have been applied to both the class I and class II alleles and their procedures are described as follows. All QSAR and molecular modeling calculations were carried out on a Silicon Graphics octane workstation using the SYBYL 6.9 molecular modeling package (Tripos Inc., USA).

4.2.2 Additive Method – Class I and Class II Alleles

Extracted IC_{50} values were first converted to $\log[1/IC_{50}]$ values (or $-\log_{10}[IC_{50}]$ or pIC_{50}) and used as the dependent variables in a QSAR regression. pIC_{50} can be related to changes in the free energy of binding: $\Delta G_{bind} = - RT \ln IC_{50}$. The values were predicted from a combination of the contributions (p) of individual amino acids at each position of the peptide and used as the dependent variables in a QSAR. The binding affinities were originally assessed by a competition assay based on the inhibition of binding of the radiolabeled standard peptide to detergent-solubilized MHC molecule (Ruppert, Sidney, Celis, Kubo, Grey, and Sette 1993; Sette, Sidney, del Guercio, Southwood, Ruppert, Dalberg, Grey, and Kubo 1994a).

We developed a program to transform the nine-amino-acid (aa) peptide sequences into a matrix with elements 1 and 0. An element is 1 when a certain amino acid at a certain position or a certain interaction between two side chains

exists and 0 when they are absent. For example, 180 columns account for the amino acid contributions (20 aa × 9 positions) while 3200 columns account for the adjacent side chains, or 1-2 interactions (20 × 20 × 8). As these two models were roughly equivalent in terms of statistical quality, we applied the principle of Occam's razor and selected the simplest case, with the amino acids only model, for discussion in this study.

The matrix was assessed using PLS (Sette et al. 1994a), an extension of Multiple Linear Regression (MLR). The method works by producing an equation or QSAR, which relates one or more dependent variables to the values of descriptors and uses them as predictors of the dependent variables (or biological activity) (Wold 1995). The IC_{50} values (the dependent variable y) were represented as negative logarithms (pIC_{50}). The predictive ability of the model was validated using "Leave-One-Out" Cross-Validation (LOO-CV) method.

4.2.3 Cross-Validation Using the "Leave-One-Out" (LOO-CV) Method

Cross-Validation (CV) is a reliable technique for testing the predictivity of models. With QSAR analysis in general and PLS methods in particular, CV is a standard approach to validation. CV works by dividing the dataset into a set of groups, developing several parallel models from the reduced data with one or more of the groups excluded, and then predicting the activities of the excluded peptides. When the number of excluded groups is the same as the number in the set, the technique is called Leave-One-Out Cross-Validation (LOO-CV). The predictive power of the model is assessed using the following parameters: cross-validated coefficient (q^2) and the Standard Error of Prediction (SEP), which are defined in Eqs. (1) and (2).

$$q^2 = 1.0 - \frac{\sum_{i=1} (pIC_{50_{(\exp)}} - pIC_{50_{(pred)}})^2}{\sum_{i=1} (pIC_{50_{(\exp)}} - pIC_{50_{(mean)}})^2} \text{ or simply } q^2 = 1.0 - \frac{PRESS}{SSQ} \quad (1)$$

where $pIC_{50\ (pred)}$ is a predicted value, $pIC_{50\ (exp)}$ is an actual or experimental value, and the summations are over the same set of pIC_{50} values. PRESS is the PRedictive Error Sum of Squares and SSQ is the Sum of Squares of $pIC_{50\ (exp)}$ corrected for the mean.

$$SEP = \sqrt{\frac{PRESS}{p-1}} \quad (2)$$

where p is the number of the peptides omitted from the dataset. The optimal number of components (NC) resulting from the LOO-CV is then used in the non-cross-validated model which was assessed using standard MLR validation terms, explained by variance r^2 and Standard Error of Estimate (SEE), defined in Eqs. (3) and (4).

$$r^2 = 1 - \frac{\sum_{i=1} (pIC_{50(\exp)} - pIC_{50(est)})^2}{\sum_{i=1} (pIC_{50(\exp)} - pIC_{50(mean)})^2} \tag{3}$$

$$SEE = \frac{\sum_{i=1} (pIC_{50(\exp)} - pIC_{50(est)})^2}{n - c - 1} \tag{4}$$

where n is the number of peptides and c is the number of components. In the present case, a component in PLS is an independent trend relating measured biological activity to the underlying pattern of amino acids within a set of peptide sequences. Increasing the number of components improves the fit between target and explanatory properties; the optimal number of components corresponds to the best q^2. Both SEP and SEE are standard errors of prediction or estimation and assess the distribution of errors between the observed and predicted (estimated) values in the regression models.

4.2.4 Iterative Self-Consistent Algorithm – Class II Alleles

An ISC-PLS-based additive method was applied to the set of class II alleles. The ISC PLS-based algorithm (Doytchinova and Flower 2003) works by generating a set of nonameric subsequences extracted from the parent peptide. Values for pIC_{50} corresponding to this set of peptides were predicted using PLS and compared to the experimental pIC_{50} value for each parent peptide. The best predicted nonamer was selected for each peptide, i.e., those with the lowest residual between the experimental and predicted pIC_{50}. LOO-CV was then employed to extract the optimal number of components, which was then used to generate the non-cross-validated model. Each new model is built from the chosen set of optimally scored nonamers. The method works by comparing the new set of peptide sequences with the old set and if the new set is different, the next iteration is begun. The process is repeated until the set of extracted nonameric peptide sequences identified by the procedure have converged. The resulting coefficients of the final non-cross-validated model describe the quantitative contributions of each amino acid at each of the nine positions. An example coefficient matrix for the I-A^b allele is shown in Table 1.

4.2.5 Comparative Molecular Similarity Index Analysis (CoMSIA)

4.2.5.1 Molecular Modeling

Wherever possible an X-ray crystallographic structure for the nonameric/octameric peptide binding to the various class I alleles was chosen as a starting conformation. Using the crystallographic peptide as a template, all the studied peptides were built, and then subjected to an initial geometry optimization using the Tripos molecular force field and charges derived using the MOPAC AM1 Hamiltonian semiempirical method (Dewar, Zoebisch, Healy, and Stewart 1985). Molecular alignment was based on the backbone atoms of the peptides, which was defined as an aggregate during optimization.

Table 1. Additive model* for the binding affinity prediction to the I-A^b allele.

	P1	P2	P3	P4	P5	P6	P7	P8	P9
A	-0.016	-0.008	0.265	-0.115	0.066	-0.442	0.050	0.447	-0.034
C	0.000	0.083	0.037	-0.051	0.090	0.050	0.216	0.079	-0.139
D	-0.065	0.000	-0.067	0.000	0.000	0.107	-0.077	-0.041	-0.203
E	-0.028	-0.129	0.000	0.000	0.000	0.000	0.000	-0.048	0.000
F	0.000	0.000	0.000	-0.283	0.000	0.000	0.000	0.000	0.000
G	-0.286	-0.039	0.050	-0.011	0.000	0.000	-0.003	0.000	-0.067
H	-0.003	-0.013	0.000	0.000	0.000	0.000	0.000	0.213	0.000
I	-0.043	0.090	-0.364	-0.090	0.000	-0.244	-0.351	0.000	-0.069
K	0.094	0.000	0.000	0.000	-0.069	0.000	0.000	0.000	0.000
L	0.000	-0.215	-0.110	0.094	-0.162	0.000	-0.003	-0.242	0.066
M	0.008	-0.067	0.000	0.258	0.223	0.154	0.017	-0.027	0.082
N	0.000	0.298	0.000	0.042	-0.003	-0.069	0.064	-0.097	-0.455
P	0.100	0.000	0.032	0.090	0.030	0.201	0.080	0.000	0.280
Q	-0.013	0.000	-0.235	0.000	0.000	0.000	0.000	-0.067	-0.051
R	0.164	-0.286	0.066	0.122	-0.233	0.120	0.213	-0.229	0.216
S	-0.051	0.090	0.161	0.036	-0.078	0.041	-0.125	0.000	0.213
T	0.054	0.151	0.079	-0.060	0.233	0.000	-0.079	0.012	0.161
V	-0.069	-0.048	0.000	0.064	0.000	0.000	0.000	0.000	0.000
W	0.000	0.000	-0.029	0.000	-0.097	-0.003	0.000	0.000	0.000
Y	0.155	0.092	0.116	-0.097	0.000	0.085	0.000	0.000	0.000

*Constant = 6.044; 0.000 represents positions where amino acids are absent.

4.2.5.2 CoMSIA Method

Five physicochemical descriptors (Steric, Electrostatic, Hydrophobic, Hydrogen Bond Donor and Acceptor) were evaluated using a probe atom placed within a 3D grid. The atom had a radius of 1 Å and charge, hydrophobic interaction, hydrogen bond donor and acceptor properties all equal to +1. The grid was extended beyond the molecular dimensions by 4.0 Å in the X, Y, and Z directions. The spacing between probe points within the grid was set at 2.0 Å and was increased in steps of 0.5 Å. CoMSIA analysis for each allele was carried out using PLS (Young 2001) and models were then validated via the LOO-CV method as previously described.

4.2.5.3 CoMSIA Maps

The results of the non-cross-validated CoMSIA models were displayed as contour maps, with each physicochemical descriptor highlighted in different colors, reflecting favorable or disfavorable changes in the peptide structure and its influence on MHC binding. These maps were created using the standard deviation coefficient option based on actual values. The CoMSIA steric bulk map is shown using green (more bulk is favored) and yellow (less bulk is disfavored) contours. The electrostatic potential map is shown with blue (negative potential is disfavored) and red (negative potential is favored) contours. CoMSIA hydrophobic interaction fields are colored yellow (where hydrophobic interaction enhances affinity) and white (where

hydrophilic interactions enhance affinity). The hydrogen-bond donor map is shown in cyan (donors on the ligand are preferred) and purple (donors are disfavored) contours. Finally, in the hydrogen-bond acceptor map favored areas are in magenta and disfavored in yellow.

4.3 Results

The Additive, ISC-PLS, and CoMSIA models were generated for 23 human and mouse class I and class II alleles.

4.3.1 Additive Method – Class I Alleles

The generated models (n=30-335), as shown in Table 2, have an acceptable level of predictive power: LOO-CV statistical terms, SEP and q^2, ranged between 0.565 and 0.907, and 0.317 and 0.531 respectively. The non-cross-validated statistical terms NC, SEE, and r^2 ranged between 2 and 9, 0.085 and 0.456, and 0.731 and 0.997, respectively. An extended motif, as defined by the class I models, is summarized in Table 3. It shows anchor and nonanchor residues relating to strong and weak binding residues. For simplicity, the quantitative contributions of amino acids at each position for the class I mouse alleles are shown in Fig. 1.

Table 2. Class I Additive-PLS Method results.

	Epitope	n[a]	LOO		Non-cross Validation		
			SEP[b]	q^2[c]	NC[d]	SEE[e]	r^2
Human	A*0101	95	0.907	0.420	4	0.146	0.997
	A*0201	335	0.694	0.377	6	0.456	0.731
	A*0202	69	0.606	0.317	9	0.193	0.943
	A*0203	62	0.841	0.327	6	0.197	0.963
	A*0206	57	0.576	0.475	6	0.085	0.989
	A*0301	72	0.680	0.436	6	0.181	0.959
	A*1101	62	0.572	0.458	2	0.321	0.829
	A*3101	30	0.710	0.482	3	0.325	0.892
	A*6801	38	0.594	0.531	4	0.175	0.959
	A*6802	46	0.647	0.500	7	0.119	0.983
	B*3501	52	0.710	0.435	6	0.118	0.984
Mouse	H2-K^k	154	0.565	0.456	6	0.198	0.933
	H2-K^b	62	0.894	0.454	6	0.128	0.989
	H2-D^b	65	0.837	0.493	5	0.268	0.948

[a] Number of epitopes; [b] Standard Error of Prediction; [c] from Leave-One-Out Cross-Validation; [d] Number of components; [e] Standard Error of Estimate.

Table 3. Class I Additive-PLS Method: nonanchor residues related with strong and weak binding for amino acids only.

Allele	P1	P2	P3	P4	P5	P6	P7	P8	P9
Favored binding*									
A*0101	*Y*	*T*	*D*	*H*	*Y*	*P*	*N*	*V*	*AY*
A*0201	FY	LM	FLM WY	T	FY	I	HP	PQ	V
A*0202	K	L	V	M	K	MP	FN	FY	LV
A*0203	DKW	LM	ANSV	PT	LRT	I	QV	T	L
A*0206	AK	*I*	*L*	L	L	F	F	T	V
A*0301	G	IT	F	RT	Y	G	IF	*M*	*K*
A*1101	*S*	*V*	*M*	*N*	*V*	*S*	*F*	*L*	*K*
A*3101	*M*	*L*	*G*	P	R	R	P	E	R
A*6801	Q	A	F	F	M	L	A	L	R
A*6802	F	V	IM	*CG*	P	I	PV	R	V
B*3501	F	*P*	*I*	*H*	T	F	L	AK	M
H2-D^b	AQF	QS	ILP	CT	N	*Q*	DE	*Y*	L
H2-K^b	KS	GISY	RFY	A	F	AG	GPQ	V	
H2-K^k	*F*	ADEP GLSTV	L	*P*	*P*	*F*	RLFW	ANILM FSTWV	
Disfavored binding*									
A*0101	*F*	*N*	*K*	*G*	*D*	*K*	*A*	*E*	*D*
A*0201	T	T	CEHS	AFI	R	R	GNQ	DI	S
A*0202	GI	AV	C	N	N	QS	ST	DE	AR
A*0203	DLR	TV	DFM	ADLN	SW	D	FLT	DLQ	A
A*0206	G	*M*	N	E	PY	P	V	R	A
A*0301	*L*	N	L	*E*	*S*	K	H	E	AQ
A*1101	L	*L*	A	*C*	*R*	*R*	*L*	*A*	Y
A*3101	*Q*	*A*	*F*	*T*	*S*	*F*	*A*	*R*	*K*
A*6801	A	N	G	*V*	*V*	*G*	*R*	*S*	*Y*
A*6802	K	M	STV	*S*	*V*	G	FQ	*D*	L
B*3501	T	A	QK	*S*	*E*	V	QS	T	TW
H2-D^b	LV	E	GS	AIF	G	P	GIMV	CG	Y
H2-K^b	DLY	DQL	EHS	GI	LS	FSV	IFY	M	
H2-K^k	A	NHIK FWY	KS	S	K	G	DK	RDYQ GHKP	

* A cut-off value of > +/- 0.3 is applied, with residues outside the limit in italics.

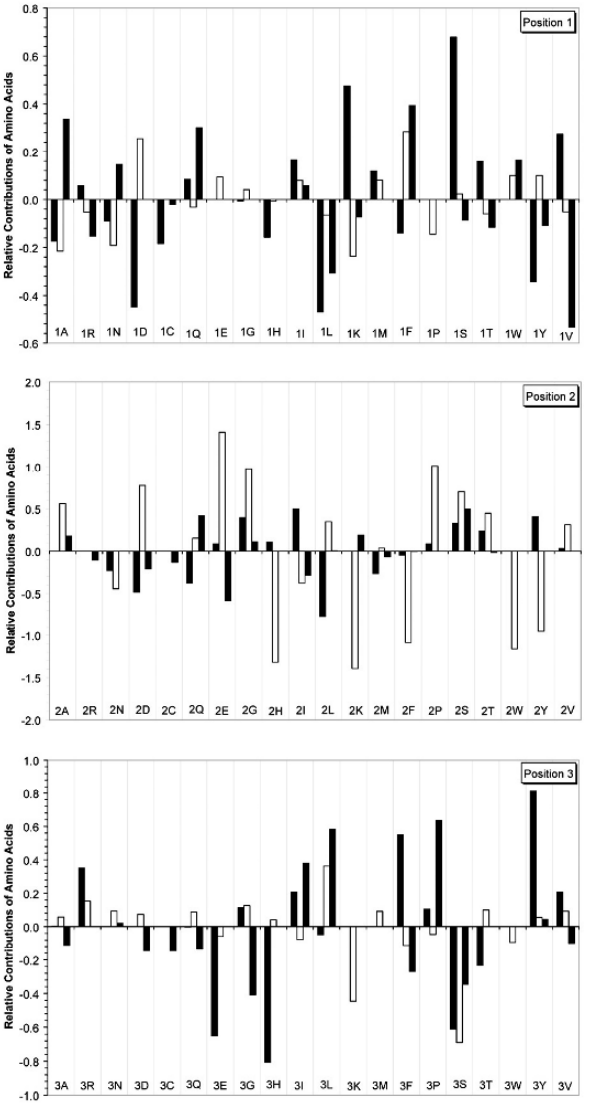


Fig. 1. (Continued)

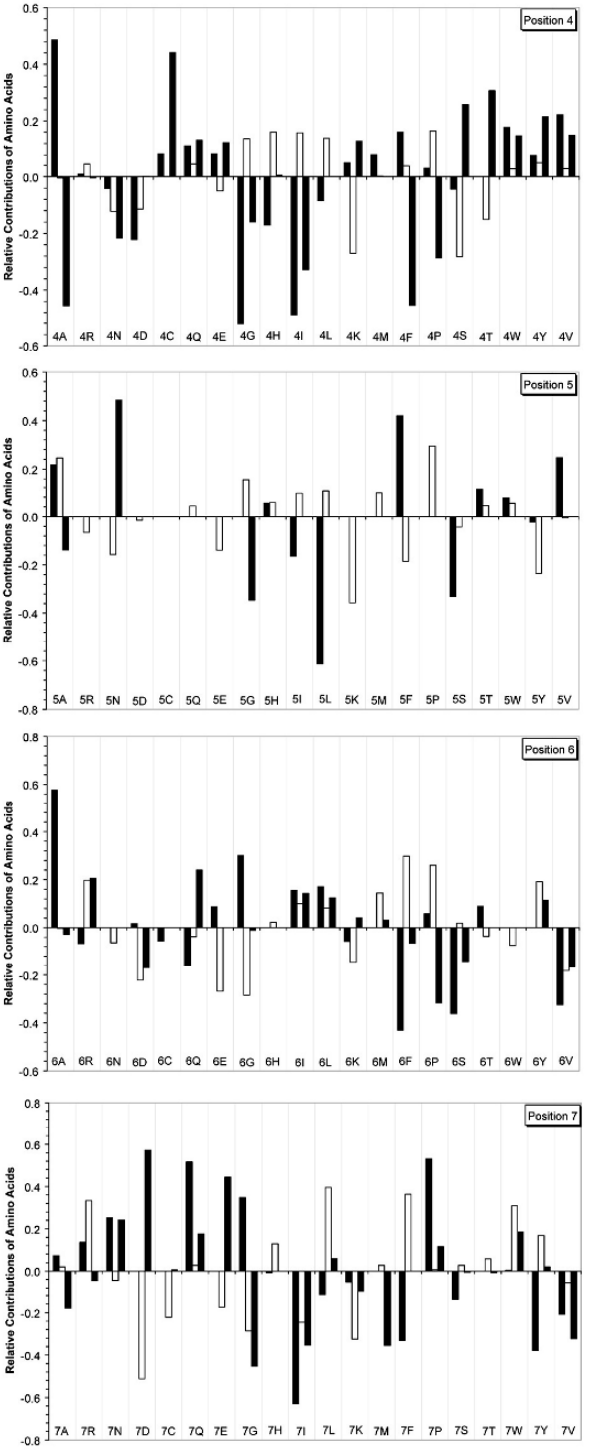


Fig. 1. (Continued)

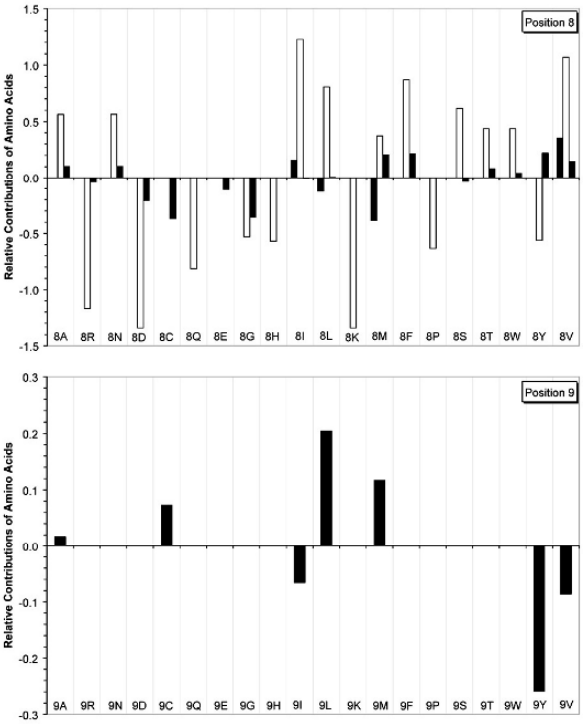


Fig. 1. Relative contributions of positionwise amino acids at each binding position 1 to 9 for the H2-D^b(black bars), H2-K^b (white bars) and H2-K^k (striped bars) alleles. The contribution made by different individual amino acids at each position of the 9mer H2-D^b, H2-K^b and H2-K^k binding peptide. The contribution is equivalent to a positionwise amino acid regression coefficient obtained by PLS regression (as described in the text).

4.3.2 Iterative Self-Consistent (ISC) Algorithm – Class II Alleles

The generated models (n=44-185), as shown in Table 4, have an acceptable level of predictivity: LOO-CV statistical terms, SEP and q^2, ranged between 0.418 and 0.816 and 0.649 and 0.925, respectively. The non-cross-validated statistical terms NC, SEE, and r^2 ranged between 4 and 8, 0.051 and 0.180, and 0.967-0.999, respectively. Convergence ranged between the 4th and 17th iteration. An extended motif, as defined by the class I models, is summarized in Table 5 showing anchor and nonanchor residues related to strong and weak binding residues.

Table 4. Class II ISC Method results.

	Epitope	n[a]	No. of iterations	LOO		Non-cross Validation		
				SEP[b]	q^2[c]	NC[d]	SEE[e]	r^2
Human	DRB1*0101	90	13	0.567	0.808	8	0.075	0.994
	DRB1*0401	185	7	0.701	0.716	4	0.174	0.967
	DRB1*0701	84	11	0.562	0.649	7	0.051	0.999
Mouse	I-A^b	44	7	0.459	0.850	6	0.089	0.994
	I-A^d	145	14	0.534	0.898	6	0.136	0.993
	I-A^k	55	4	0.816	0.790	6	0.180	0.990
	I-A^s	81	17	0.588	0.783	6	0.177	0.980
	I-E^d	69	8	0.557	0.732	6	0.096	0.992
	I-E^k	52	8	0.418	0.925	6	0.106	0.995

[a] Number of epitopes; [b] Standard Error of Prediction; [c] Obtained after Leave-One-Out Cross-Validation; [d] Number of components; [e] Standard Error of Estimate.

Table 5. Class II ISC-Additive Method: nonanchor residues related with strong and weak binding for amino acids only.

Allele	P1	P2	P3	P4	P5	P6	P7	P8	P9
Favored binding*									
DRB1*0101	YF	IY	*T*	*P*	*L*	A	N	*P*	*S*
DRB1*0401	*VW*	*VY*	*L*	*A*	*A*	*S*	*PV*	*L*	*N*
DRB1*0701	Y	P	V	L	T	A	*V*	*N*	*SV*
I-A^b	*P*	*K*	*A*	*L*	*LT*	*M*	*R*	*A*	*M*
I-A^d	CTW	AGM	*T*	CMSW	QLV	L	AIY	GIL	AC FT
I-A^k	T	T	*G*	*C*	GS	*F*	EY	*Q*	*C*
I-A^s	F	C	NL	*F*	*A*	*I*	*I*	G	*H*
I-E^d	M	Q	W	*W*	S	A	*W*	*R*	*C*
I-E^k	*I*	*A*	*Y*	*R*	*R*	Q	L	T	A
Disfavored binding*									
DRB1*0101	*F*	*N*	*K*	*G*	*D*	*K*	*A*	*E*	*D*
DRB1*0401	T	T	CEHS	AFI	R	R	GNQ	DI	S
DRB1*0701	GI	AV	C	N	N	QS	ST	DE	AR
I-A^b	DLR	TV	DFM	ADLN	SW	D	FLT	DLQ	A
I-A^d	G	*M*	N	E	PY	P	V	R	A
I-A^k	*L*	N	L	*E*	*S*	K	H	E	AQ
I-A^s	L	*L*	A	*C*	*R*	*R*	*L*	*A*	Y
I-E^d	*Q*	*A*	*F*	*T*	*S*	*F*	*A*	*R*	*K*
I-E^k	A	N	G	*V*	*V*	*G*	*R*	*S*	*Y*

* A cut-off value of > +/- 0.4 is applied to favored and disfavored binding amino acids, with residues outside the limit in italics.

4.3.3 Comparative Molecular Similarity Index Analysis (CoMSIA)

For each of the 12 alleles, all peptides were built and aligned in three dimensions (Fig. 2), their geometry was optimized, and AM1 (Dewar et al. 1985) calculations performed within SYBYL 6.9. The peptides were placed within individual 3D grids (Fig. 3). The final settings for the three models are shown in Table 6. The generated models (n=30-236) have an acceptable level of predictivity: LOO-CV statistical terms, SEP and q^2, ranged between 0.443 and 0.889 and 0.385 and 0.700, respectively. The non-cross-validated statistical terms NC, SEE, and r^2 ranged between 4 and 12, 0.071 and 0.411, and 0.867 and 0.991, respectively.

To generate CoMSIA coefficient contour maps for each allele, which describes the relationship between the binding affinity and each physicochemical descriptor, three non-cross-validated "all fields" models were created based on the five physicochemical descriptors (steric, electrostatic, hydrophobic, hydrogen bond donor and acceptor). The descriptors involved in the interaction between the peptide and the MHC molecules are presented in the coefficient contour maps as shown in Fig. 4 for the H2-D^b allele. For simplicity, the interaction between only one peptide and its respective contour map is shown with the N-terminus to the left and the C-terminus to the right. Table 7 shows a summary of the position specificities between the physicochemical descriptors and peptide positions for the A2 supermotif and class I mouse alleles.

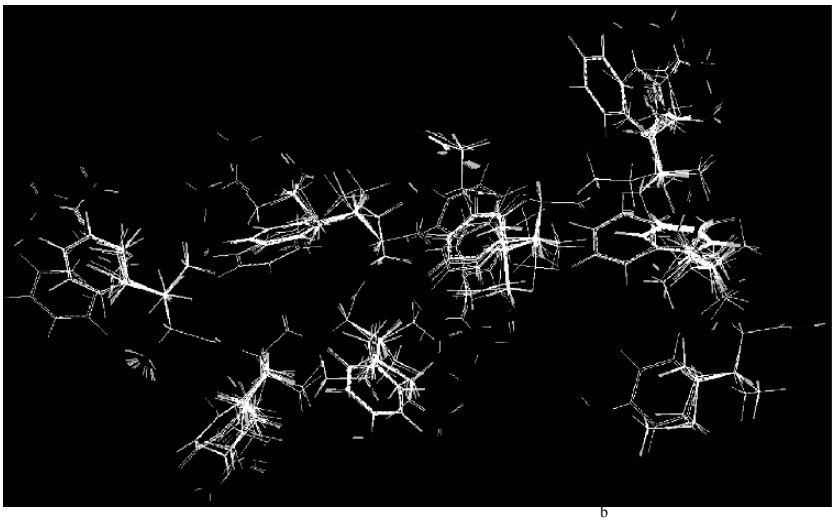


Fig. 2. Superimposed alignment of peptide molecules for the H2-D^b allele.

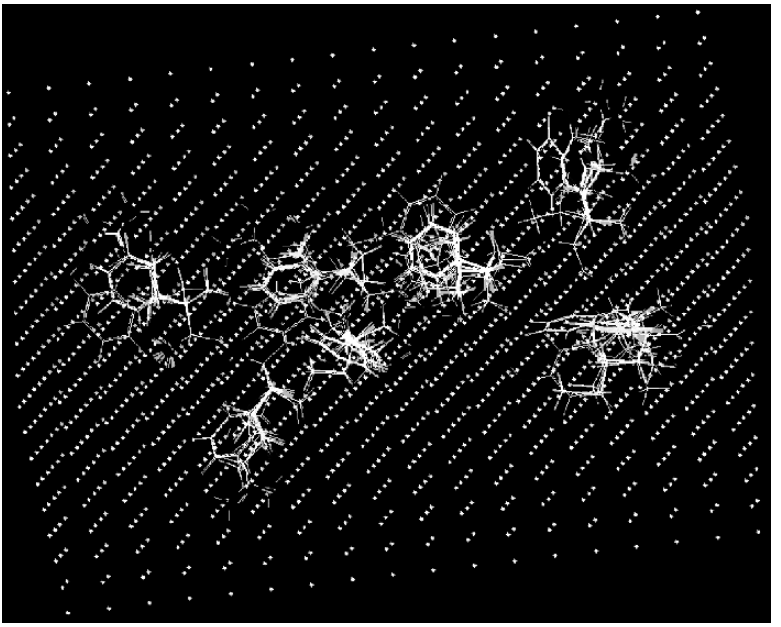


Fig. 3. Superimposed H2-D^{b} peptide molecules placed within 3D grid lattice.

Table 6. Class I Additive-PLS Method results for CoMSIA method.

			LOO		Non-cross Validation			Grid dimensions (Å)	
	Epitope	n	SEP	q^2	NC	SEE	r^2	Size	Spacing
Human	A*0201	236	0.443	0.683	7	0.260	0.891	22x15x15	2.0
	A*0202	63	0.509	0.534	8	0.190	0.935	22x15x15	3.0
	A*0203	60	0.595	0.621	6	0.179	0.966	22x15x15	3.0
	A*0206	54	0.505	0.523	12	0.071	0.991	22x15x15	2.0
	A*0301	69	0.629	0.486	6	0.177	0.959	22x15x15	2.0
	A*1101	59	0.588	0.496	8	0.141	0.972	22x15x15	2.0
	A*3101	30	0.551	0.700	4	0.282	0.921	22x15x15	1.5
	A*6801	39	0.674	0.430	5	0.119	0.950	22x15x15	2.0
	A*6802	45	0.652	0.385	4	0.197	0.944	22x15x15	2.0
Mouse	H2-K^{k}	154	0.525	0.611	6	0.248	0.913	18x13x12	2.0
	H2-K^{b}	62	0.889	0.490	6	0.244	0.962	19x13x11	2.0
	H2-D^{b}	65	0.783	0.518	4	0.411	0.867	18x14x11	2.0

[a] Number of epitopes; [b] Standard Error of Prediction; [c] Obtained after Leave-One-Out Cross-Validation; [d] Number of components; [e] Standard Error of Estimate; [f] Grid steps of 0.5Å.

Table 7. Summary of CoMSIA position specificities* for the A2 supermotif (Doytchinova et al. 2002b) (Class I HLA-A*0201A*0202A*0203A*0206 and A*6802) and Class I Mouse (H2-K^b, H2-K^k and H2-D^b).

Position of side chain	Steric bulk	Electron density	Hydrop-hobicity	H-Bond donor	H-Bond acceptor
A2 Supermotif (Class I HLA-A*0201A*0202A*0203A*0206 and A*6802)					
P1: pocket A		F - Aromatic amino acids preferred.	F		
P2: pocket B	F		F		D
P3: pocket D				D	D
P4: exposed to solvent and T-cell	D	D - Aliphatic amino acids preferred.		F	
P5		D - Aliphatic amino acids preferred.			
P6: pocket C	F		F		D
P7: pocket E	D.	D - Aliphatic amino acids preferred.			D
P8: exposed to solvent and T-cell	D	D - Aliphatic amino acids preferred.	D	F	F
P9: pocket F	F		F		
Class I mouse (H2-K^b, H2-K^k and H2-D^b).					
P1: pocket A	F	D	D	F	F
P2: pocket B	D	D	F		F
P3: pocket D	F		F	F	
P4: exposed to solvent and T-cell			D	F	F
P5			F		F
P6: pocket C	F				
P7: pocket E		D	D		
P8: exposed to solvent and T-cell	F	F	F	F	F
P9: pocket F					

* F, favored; D, disfavored.

4.4 Discussion

Herein we report the development of quantitative, systematic models, based on literature IC_{50} values, for human class I (HLA-A*0101, HLA-A*0201, HLA-A*0202, HLA-A*0203, HLA-A*0206, HLA-A*0301, HLA-A*1101, HLA-A*3101, HLA-A*6801, HLA-A*6802, HLA-B*3501), mouse class I (H2-K^k, H2-K^b, and H2-D^b), human class II (HLA-DRB1*0101, HLA-DRB1*0401, and HLA-DRB1*0701), and mouse class II (I-A^b, I-A^d, I-A^k, I-A^s, I-E^d, and I-E^k).

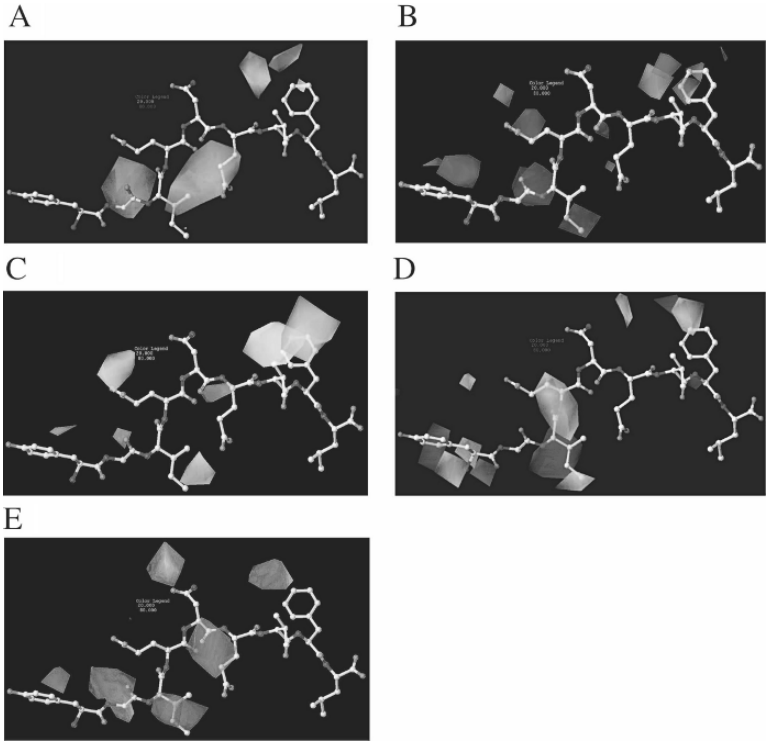


Fig. 4. H2-D^b allele: Steric bulk maps (A), electrostatic potentials maps (B), hydrophobic interaction maps (C), H-bond donor maps (D), H-bond acceptor maps (E). (A color version of this figure appears between pages 76 and 77.)

4.4.1 Additive Method – Class I Alleles

The 2D-QSAR additive method has been applied to the peptide binding specificities of the A3 superfamily human class I alleles: A*1101, A*0301, A*3101, and A*6801. Sequence analysis showed that only 11 of the residues inside the binding pockets are polymorphic. A good, if incomplete, consensus was found in the preferences at the primary anchor positions 2 and 9. Thr and short hydrophobic residues such as Ala and Ile were favored at P2 and nearly all the peptides bound to A3 alleles had positively charged residues Arg or Lys at the C-terminus. The amino acids involved in peptide binding are similar in HLA-A2 and the A3 family. Pocket B interacts with the side chain of the residue at position 2, which was one of the anchor positions in nearly all the MHC class I alleles. Most of the amino acids in pocket B are conserved in the A2 and A3 families; both families accept hydrophobic residues. The amino acid at sequence position 9 of the MHC protein is important in peptide binding in the two families. Alleles (A*3101, A*0301, and A*0201) with small to medium-sized residues (Phe9 or Thr9) were able to accept residues with long side chains such as Leu. On the other hand, only small residues such as Ala and Val could bind to A*6801, A*1101, and A*0206, all of which had the larger residue Tyr9. The

five residues that directly interacted with the peptide in the F pocket are identical in both the A3 family and HLA-B27 (Leu81, Asp116, Tyr123, Thr143, and Trp147). Arg and Lys bound to pocket F and interacted with negatively charged residues Asp116 or Asp77 in both the A3 family and HLA-B27. B27 had been shown to accept hydrophobic residues such as Leu, Ala, and Tyr because of their interaction with Leu81, Tyr123, Thr143, and Trp147 in the binding pocket (Jardetzky, Lane, Robinson, Madden, and Wiley 1991). In the present study, the specificity at position 9 was restricted to Arg and Lys only; both Ala and Tyr had deleterious effects on peptide binding. This suggests a possible difference in the conformation of the binding pocket in spite of sequence similarity. Also, this may be the result of a change in conformation after the binding of other amino acids in the peptide. A peptide-binding motif for the HLA-A3 superfamily has been defined previously (Sidney, Grey, Southwood, Celis, Wentworth, del Guercol, Kubo, Chestnut, and Sette 1996; Rammensee, Friede, and Stevanovic 1995). Some useful similarities can be found on comparing the present motif with those defined by the above two groups. The amino acid preferences for the primary anchor residues are similar. All the motifs show preference for Arg and Lys at position 9 and have a preference for various hydrophobic residues at position 2, such as Ile and Thr. The preferences for secondary anchor residue positions 3 and 7 in the three motifs are hydrophobic amino acids such as Phe.

The amino acid contributions to the affinity of peptides binding to the A2 family: A*0201, A*0202, A*0203, A*0206, and A*6802 alleles using the Additive-PLS Method have also been analysed quantitatively. Certain discrepancies between A*6802 and A*02 molecules concerning the amino acid preferences at P1–P9 were seen in the present study. These discrepancies throw doubt on whether the A*6802 allele belongs to the A2 supertype. The sequence comparison showed that there are only one or two differences in the residues forming the six pockets of A*0201, A*0202, A*0203, and A*0206 molecules. The number of these differences between A*6802 and A*02 molecules is seven residues. Five of them concern pockets A, B, and C and are so substantial that they alter the amino acid preferences at the primary anchor P2 and the secondary anchors P1 and P6. The preferred Val and Thr for P2 brings the A*6802 allele closer to the A3 supertype (Sidney et al. 1996) rather than to the A2 one. But the A3 supermotif requires positively charged residues, such as Arg and Lys, at the C-terminus (Sidney et al. 1996), which is not true in the case of A*6802. Obviously, A*6802 is an intermediate allele standing between A2 and A3 supertypes: in anchor position 2 it is closer to A3 and in anchor position 9 it is nearer to A2. Residues identified as preferred for two or more A*02 molecules, without being deleterious for any molecule, are considered as preferred. Residues identified as deleterious for two or more molecules are considered as deleterious in the common motif. The expansion concerns all positions and especially the anchor P2.

The Additive-PLS results for the mouse alleles are in good agreement with previous studies of the preferred primary anchor positions: 5 and 9 (nonamers); 2, 5, and 8 (octamers – H2-K^k and H2-K^b, respectively). All three models also agree with previous analyses of the preferred residue type at the anchor positions. For H2-D^b: Asn at position 5 and Leu at position 9; for H2-K^b: Phe at position 5 and Val at position 8; and for H2-K^k: Glu, Pro, Gly (best three favored residues) at position 2

and Ile, Val, Phe (best three favored residues) at position 8. The nonameric and octameric alleles show both similarities and differences in amino acids preferred at various binding positions. Preferences for primary anchors show certain similarities: all models exhibit some preference for small amino acids (H2-D^b (Asn), H2-K^b (Val). and H2-K^k (Pro, Ala)), while C-terminal amino acids are strongly hydrophobic: H2-D^b (Leu), H2-K^b (Val), and H2-K^k (Ile, Val). The most noticeable difference between the nonameric and octameric alleles is at position 5, where H2-D^b exhibits a preference for polar Asn, while H2-K^b shows a preference for Phe (aromatic hydrophobic residue) and H2-K^k for Pro (small amino acid residue).

As well as refining and confirming our understanding of sequence dependence at anchor positions, our results throw new light on all other positions within the peptide. Although this study supports the importance of both primary and secondary anchor residues, it is clear that other positions also play a key role in peptide-binding (Hudrisier, Mazarguil, Laval, Oldstone, and Gairin 1996). Table 3 shows a summary of residues associated with both favored and disfavored binding to all three alleles. Looking at Table 3, for weak binding peptides, hydrophobic residues are present at position 1 (Phe) and position 3 (Leu, Ile, Tyr, Phe) in abundance, and there is a probable electrostatic repulsion of both negatively charged polar side chains (Asp and Glu) and positively charged polar side chains (Lys, Arg, and His).

Each class I mouse MHC allele binds a mixture of structurally diverse peptides, typically 8-10 amino acids in length, with each allele exhibiting defined peptide specificity. From our work (Doytchinova and Flower 2002a; Doytchinova and Flower 2002b; Doytchinova et al. 2002c; Doytchinova and Flower 2003; Guan et al. 2003a; Guan et al. 2003b; Hattotuwagama et al. 2004), previous peptide binding experiments, and X-ray crystallographic studies of human class I MHC molecules, it is clear that the molecule binds short peptides, most of which are nonamers (Bjorkman, Saper, Samraoui, Bennett, Strominger, and Wiley 1987). Topologically position 1 corresponds to pocket A of the cleft of the peptide-binding site on HLA-A*0201 (Saper, Bjorkman, and Wiley 1991). Anchor residues at position 2 and at the C-terminus (position 9) are seen to be of primary importance for binding, where pocket B interacts with the side chain of the residue at position 2. The structure of pocket A is mainly polar residues and consists of a network of hydrogen bonding residues. A hydrophobic ridge cuts through the binding cleft forcing the peptide to arch between position 5 and the carboxyl-terminal residue (position 9) which are anchored into the D and F pockets in the floor of the cleft (Fremont, Matsumura, Stura, Peterson, and Wilson 1992). Equivalent data for mice show clear differences and significant similarities. The crystal structure of several mouse class I molecules has revealed that the peptide binding cleft is also closed at both ends, that the length of the cleft is similar for all class I molecules (Fremont, Stura, Matsumara, Peterson, and Wilson 1995; Zhang, Young, Imarai, Nathenson, and Sacchettini 1992; Young, Zhang, Sacchettini, and Nathenson 1994; Smith, Reid, Harlos, McMichael, Stuart, Bell, and Jones 1996a; Smith, Reid, Stuart, McMichael, Jones, and Bell 1996b), and that the carboxyl-terminal peptide position is an anchor residue deeply buried in the F pocket. Analysis of the structure and binding results of the H2-K^b and H2-K^k octameric complex reveals that there is a strong preference for an aromatic and hydrophobic residues Tyr and Phe (H2-K^b) and Leu (H2-K^k) at positions 3 and 5

and for a strong hydrophobic residue Val (H2-K^b) and Ile, Val, and Phe (H2-K^k) at position 8, which is in accordance with the studies of Falk (Falk, Rötzschke, Stevanovic, Jung, and Rammensee 1991). It is found that in H2-K^b the B pocket is large enough to accommodate a bulky Ile residue at position 2, which is in accordance with the crystal structure of the antigenic peptide from the ovalbumin complex OVA-8 (SIINFEKL). In H2-K^b and H2-K^k alleles, the results showed that Tyr, Phe, and Leu are all favored in position 3 (Fremont et al. 1992), which is situated in part of pocket D and would significantly deepen the depth and volume of the D pocket and is complementary to the pocket. The anchor carboxyl-terminal (position 8) prefers hydrophobic residues, which fall into pocket F. Such results show that the peptide binding cleft is closed at both ends, that the cleft has the same length in all class I molecules, that the carboxyl-terminal peptide position is deeply buried in the F pocket, and that there is little restriction on amino acids bound by pocket A (Doytchinova et al. 2002c; Doytchinova and Flower 2003; Guan et al. 2003a; Hattotuwagama et al. 2004; Saper et al. 1991). Our study has identified favored and disfavored regions which are consistent with both the properties of peptide positions and those of pockets, designated by A to F, within the MHC binding groove. It is well known that each class I mouse MHC allele binds a mixture of structurally diverse peptides, typically 8-10 amino acids in length, and that each allele possesses defined peptide specificity. The crystal structure of several mouse class I molecules (Fremont et al. 1995; Zhang et al. 1992; Young et al. 1994; Smith et al. 1996a; Smith et al. 1996b) has helped to rationalize observed peptide binding.

4.4.2 Comparative Molecular Similarity Index Analysis (CoMSIA)

The motif of HLA-A3 superfamily includes main anchor positions 2 and 9 (Zhang, Gavioli, Klein, and Masucci 1993). Peptides bound to members of the A3 family usually had a positively charged residue—arginine or lysine—at the C-terminus, and a variety of hydrophobic residues at position 2. It was found that steric bulk was favored at position 2 for A*0301 and A*3101 but disfavored in A*1101 and A*6801 models. The study of crystal structures of MHC molecules showed that the residue at peptide position 2 bound in pocket B (Saper et al. 1991; Madden, Gorga, Strominger, and Wiley 1991). There are different residues lining pocket B in the different MHC-A3 molecules: Tyr9 in A*1101 and A*6801, Phe9 in A*0301, and Thr9 A*3101 (Schönbach, Koh, Sheng, Wong, and Brusic 2000). This means more space in pocket B for A*0301 and A*3101, allowing them to accommodate larger side chains. Electrostatic potential, hydrophobicity, and hydrogen bond acceptance maps were very varied at this position. This was in good agreement with the broad spectrum of amino acids observed at this position, from the bulky hydrophobic Leu to the small polar Thr. The most important property for the amino acid at position 9 was hydrogen-bond donor ability. It was favored by A*6801 and A*3101, and was disfavoured by A*1101. For A*0301 were found areas of favored and disfavored hydrogen bond donor groups at this position. In some cases, the change of Lys to the larger residue Arg could affect the expression of the molecule (Zhang et al. 1993). Results from the present study suggested the interaction between the residue at peptide position 9 and the MHC molecule may play an important role. The side chain of larger basic residue

Arg could extend to the bottom of pocket F of A*6801 and A*3101, forming complex stabilizing hydrogen bonds with residues at the bottom of the pocket. Among the secondary anchors, positions 1, 3, 5, 6, and 7 were of great importance. The common favored property for position 1 was hydrogen-bond donor/acceptor ability. Hydrogen-bond donor groups with negative electrostatic potential were preferred at position 3 for three of the alleles. Sidney and co-workers (Sidney et al. 1996) found that peptides with an aromatic residue, like Tyr, Phe, and Trp, had a 31-fold increase in binding affinity to A*0301. Bulky side chains with negative electrostatic potential were preferred at position 5. Hydrogen-bond donors and acceptors were disfavored here. Hydrophilic amino acids capable of forming hydrogen bonds were well accommodated at position 6. The only common favored property for position 7 was hydrophobicity. Positions 4 and 8 face the T-cell receptor (Silver, Guo, Strominger, and Wiley 1992), but can still contribute to the affinity. Hydrogen-bond donor ability was important for position 4. Steric bulk and negative electrostatic potential were favored at position 8.

Looking at the CoMSIA results for the mouse alleles, we see that with the H2-D^b allele, steric bulk is favored with the side chains of positions 3 and 6 falling into pockets D and C, respectively. For the electrostatic potential field, the alkyl side chain of position 1 falls into pocket A which consists of Val and Ser residues (Saper et al. 1991). At position 2, where the side chain falls into pocket B, electrostatic potential interaction is favored (Saper et al. 1991). In the remaining positions there are no favorable electrostatic potential interactions. There is a strongly favored hydrophobic interaction at position 8 where the side chain is solvent exposed and contacts the T-cell. The major favored interactions of the hydrogen bond donor fields are found at position 1 and across the peptide backbone between positions 3 and 4. The hydrogen bond acceptor map shows position 2 to be favored and, to a lesser extent, at positions 5 and 7.

For the H2-K^b allele, steric bulk is favored at positions 1, 3, 4, and 5. The side chain at position 1 makes a weak electrostatic interaction; while at position 2 the electrostatic potential map indicates that aromatic-type residues, such as Tyr or Phe, are well tolerated. This is in good agreement with experimental data (Ruppert et al. 1993; Parker, Bednarek, and Coligan 1994). There is no major interaction between side chains at position 3 and pocket D indicated by our model, and in the remaining positions there are no clear favorable electrostatic interactions. The hydrophobic interaction field identifies a favorable interaction at positions 3 and 5. Pocket D is a hydrophobic cavity and amino acids such as Tyr and Ile are well tolerated here which would significantly deepen the depth and volume of pocket D (Fremont et al. 1992). The major favored interactions of the hydrogen bond donor fields are found at positions 1, 3, and 4 (pockets A, D and the "flag" pocket, respectively) (Saper et al. 1991), with a major disfavored interaction found at position 6 (pocket C). The hydrogen bond acceptor map has favoured interactions at positions 1 and 4, pocket A and the "flag" pocket, respectively, but major disfavored interactions between the side chain positions 3 and 5.

For the H2-K^k allele, steric bulk field is favored at positions 1, 7, and 8. There is no favorable electrostatic interaction at position 1, while at position 2 electrostatic potential is favored. Position 3 falls into pocket D but makes little interaction with

the H2-K^k allele. In the remaining positions there seem to be no discernibly favored interactions. Hydrophobic interaction shows a major disfavored interaction at position 2 covering the whole side chain. The only favored interaction in the hydrogen bond donor map in the H2-K^k allele lies between positions 7 and 8. The main disfavored interaction is found at position 2. Within the hydrogen bond acceptor map, there is a strong disfavored interaction between the side chains at positions 2 and 3.

4.4.3 Iterative Self-Consistent (ISC) Algorithm – Class II Alleles

We have examined a recently developed bioinformatics method: the Iterative Self-Consistent (ISC) Partial Least Squares (PLS)-based Additive Method, which was applied to the prediction of class II Major Histocompatibility Complex (MHC)-peptide binding affinity. We have shown previously that ISC is a reliable, quantitative method for binding affinity prediction (Doytchinova and Flower 2003) developing a series of quantitative, systematic models, based on literature IC_{50} values.

Experimental studies of T-cell epitope analogue binding and data from X-ray crystallography, show that peptides bind to MHC molecules through the interaction of side chains of certain peptide residues with pockets situated in the MHC class II peptide-binding site: these side-chains extend into discrete pockets within the binding groove (Hennecke and Wiley 2002; Fremont, Monnaie, Nelson, Hendrickson, and Unanue 1998; Corper, Stratmann, Apostolopoulos, Scott, Garcia, Kang, Wilson, and Teyton 2000). Peptide side chains form favorable interactions with MHC side chains within these pockets (Corper et al. 2000); the most critical determinant of binding, other than the presence of appropriate types of side chain, is their relative spacing. It has been suggested before that different MHC class II molecules can bind the same peptide in several, alternative binding registers, whereby the peptide moves sideways in the binding groove with side chains being bound by different pockets (McFarland, Sant, Lybrand, and Beeson 1999; Li, Li, Martin, and Mariuzza 2000; Vidal, Daniel, Vidavsky, Nelson, and Allen 2000). Reviewing this concept (Bankovich, Girvin, Moesta, and Garcia 2004), identify two main alternative scenarios: binding of the same peptide in different registers by the same or different alleles. The more common second alternative is well demonstrated (Li et al. 2000; Vidal et al. 2000) and results from minor polymorphic differences in the amino acid residue composition of the binding groove. In the DRB5 complex, the large P1 pocket accommodates Phe from the peptide and Ile occupies the shallow pocket at P4. However, in the DRB1 allele, the small pocket at P1 is occupied by Val shifting the peptide to the right, while Phe occupies a deeper pocket at P4. This also causes certain peptide side chains, which are orientated toward the TCR, to change (Li et al. 2000). Unequivocal evidence supporting the former alternative is somewhat scarce: there are few, if any, proper examples of exactly the *same* peptide binding in different registers to exactly the *same* MHC molecule.

Our results are consistent with the view that MHC binding motifs are a less-than-adequate representation of the underlying mechanism of binding. As we have shown elsewhere (Doytchinova, Walshe, Jones, Gloster, Borrow, and Flower 2004; Flower 2003), the whole of a peptide contributes to binding, albeit weighted differently at different positions. At least for class I, it is even possible to generate high-affinity

peptides without using canonical anchors, with extra affinity arising from other interactions made by the rest of the peptide. This is also likely to be a feature of class II binding. For example, Liu, Dai, Crawford, Fruge, Marrack, and Kappler (2002) showed that for I-A^b it was possible for a peptide bearing alanines to bind to its four main pockets – which correspond to positions P1, P4, P6, and P9 and which usually bind larger peptide side chains – with compensatory interactions made by residues at other positions in order to maintain overall affinity. Our class II models suggest that the relative contributions, of particular residues, to binding are spread more evenly through the peptide than is generally supposed, rather than being concentrated solely in so-called anchor positions.

The ISC algorithm described above combines an iterative approach to selecting the best predicted binders with PLS, a robust multivariate statistical tool for model generation. The ISC method is universal in that it can be used for any peptide-protein binding interaction where the peptide length is unrestricted but the binding is limited to a fixed, if unknown, part of the peptide. Implementation of the method is straightforward, it is fast to use, and its interpretation is straightforward. The final models derived from these calculations will be included in an updated version of MHCPred (Doytchinova and Flower 2003; Guan et al. 2003a; Hattotuwagama et al. 2004; Guan, Doytchinova, and Flower 2003c; Guan, Doytchinova, Zygouri, and Flower 2003d).

4.5 Conclusions

From our studies, we find that distinct MHC alleles, both class I and class II, exhibit different peptide specificities: peptides are bound with particular sequence patterns, leading to the development of so-called motifs (Takamiya, Schönbach, Nokihara, Yamaguchi, Ferrone, Kano, Egawa, and Takiguchi 1994). Motifs are usually expressed in terms of anchor residues: the presence of certain amino acids at particular positions that are thought to be essential for binding. Taking human class I allele HLA-B*3501 as our example, previous studies have indicated the need for anchor residues at positions 2 (Pro) and 9 (hydrophobic or aromatic residues, such as Phe, Met, Leu, Ile, and especially Tyr). Primary anchor residues, although generally deemed to be necessary, are not sufficient for peptide binding, and secondary anchors, residues that are favorable, but not essential, for binding, may also be required; other positions show positional preferences for particular amino acids. Moreover, the presence of certain residues at specific positions of a peptide can have a negative effect on binding (Amaro, Houbiers, Drijfhout, Brandt, Schipper, Bavinck, Melief, and Kast 1995; Sidney, del Guercio, Southwood, Engelhard, Appella, Rammensee, Falk, Rötzschke, Takiguchi, and Kubo 1995; Smith et al. 1996b). Although motif methods are admirably simple – easy to implement either by eye or more systematically scanning protein sequences computationally – there remain many problems with the motif approach.

Although it is possible to score the relative contributions of primary and secondary anchors to produce a rough-and-ready measure of binding affinity (Amaro et al. 1995; Sette, Vitiello, Reherman, Fowler, Nayersina, Kast, Melief, Oseroff, Yuan, and Ruppert 1994b), the most significant problem with the motif approach is that it is, fundamentally,

a deterministic method. A peptide is either a binder or is not a binder. A brief reading of the literature shows that motif matches produce many false positives, and are, in all probability, producing an equal number of false negatives. Indeed there are many examples where peptides without both dominant anchors still bind with high affinity. A more accurate description of this phenomenon is to say that MHCs bind peptides with an equilibrium binding constant dependent on the nature of the bound peptide's sequence. The driving forces behind this binding are precisely the same as those driving drug binding. Within the human population there are an enormous number of different, variant genes coding for MHC proteins, each exhibiting a different peptide-binding sequence selectivity. T-cell receptors, in their turn, also exhibit different affinities for pMHC. The combined selectivity of both MHCs and TCRs determines the power of peptide recognition within the immune system and through this phenomenon the recognition of foreign pathogens. Experimentally, there are many ways to measure binding affinity. IC_{50} values are the most widely quoted binding affinity measures and are calculated from a competitive binding assay (Ruppert et al. 1993). Once a peptide has bound to an MHC to be recognized by the immune system, the pMHC complex has to be recognized by one of the TCRs of the T-cell repertoire. It is generally accepted that a peptide binding to an MHC may be recognized, by a TCR, if it binds with a pIC_{50} greater than a value of 6.3.

There is some evidence suggesting that as the MHC binding affinity of a peptide rises, the greater is probability that it will be a T-cell epitope. The prediction, then, of MHC binding is both the best understood and, probably, the most discriminating step in the presentation-recognition pathway. A pragmatic solution to the as yet unsolved problem of what will be recognized by the TCR, and thus activate the T-cell, is to greatly reduce the number of possible epitopes using MHC binding prediction, and then test the remaining candidates using some measure of T-cell activation, such as T-cell killing or thymidine incorporation.

References

Amaro, J.D'., Houbiers, J.G., Drijfhout, J.W., Brandt, R.M., Schipper, R., Bavinck, J.N., Melief, C.J., and Kast, W.M. (1995) A computer program for predicting possible cytotoxic T lymphocyte epitopes based on HLA class I peptide-binding motifs. Hum. Immunol. 43:13-18.

Bankovich, A.J., Girvin, A.T., Moesta, A.K., and Garcia, K.C. (2004) Peptide register shifting within the MHC groove, theory becomes reality. Mol. Immunol. 40:1033-1039.

Bjorkman, P.J., Saper, M.A., Samraoui, B., Bennett, W.S., Strominger, J.L., and Wiley, D.C. (1987) Structure of the human class I histocompatibility antigen, HLA-A2. Nature 329:506-512.

Blythe, M., Doytchiniva, I.A., and Flower, D.R. (2002) JenPep, a database of quantitative functional peptide data for immunology. Bioinformatics 18:434-439.

Bohm, M., Sturzebecher, J., and Klebe, G. (1999) Three-dimensional quantitative structure-activity relationship analyses using comparative molecular field analysis and comparative molecular similarity indices analysis to elucidate selectivity differences of inhibitors binding to trypsin, thrombin and factor Xa. J. Med. Chem. 42:458-477.

Chicz, R.M., Urban, R.G., Lane, W.S., Gorga, J.C., Stern, L.J., Vignali, D.A.A., and Strominger, J.L. (1992) Predominant naturally processed peptides bound to HLA-DR1 are derived from MHC-related molecules and are heterogeneous in size. Nature 358:764-768.

Chicz, R.M., Urban, R.G., Gorga, J.C., Vignali, D.A.A., Lane, W.S., and Strominger, J.L. (1993) Specificity and promiscuity among naturally processed peptides bound to HLA-DR alleles. J. Exp. Med. 178:27-47.

Corper, A.L., Stratmann, T., Apostolopoulos, V., Scott, C.A., Garcia, K.C., Kang, A.S., Wilson, I.A., and Teyton, L. (2000) A structural framework for deciphering the link between I-Ag7 and autoimmune diabetes. Science 288:505-511.

Dessen, A., Lawrence, C.M., Cupo, S., Zaller, D.M., and Wiley, D.C. (1997) X-ray crystal structure of HLA-DR4 (DRA*0101, DRB*0401) complexed with a peptide from human collagen II. Immunity 7:473-481.

Dewar, M.J.S., Zoebisch, E.G., Healy, E.F., and Stewart, J.J.P. (1985) AM1, a new general purpose quantum mechanical molecular model. J. Am. Chem. Soc. 107:3902-3909.

Doytchinova, I.A., and Flower, D.R. (2002a) Physicochemical explanation of peptide binding to HLA-A*0201 major histocompatibility complex, a three-dimensional quantitative structure-activity relationship study. Proteins 48:505-518.

Doytchinova, I.A., and Flower, D.R. (2002b) A comparative molecular similarity index analysis (CoMSIA) study identifies an HLA-A2 binding supermotif. J. Comput. Aided Mol. Des. 16:535-544.

Doytchinova, I.A., Blythe, M.J., and Flower, D.R. (2002c) Additive method for the prediction of protein-peptide binding affinity. Application to the MHC class I molecule HLA-A*0201. J. Proteome Res. 1:263-272.

Doytchinova, I.A., and Flower, D.R. (2003) Towards the *in silico* identification of class II restricted T-cell epitopes, a partial least squares iterative self-consistent algorithm for affinity prediction. Bioinformatics 19:2263-2270.

Doytchinova, I.A., Walshe, V., Jones, N., Gloster, S., Borrow, P., and Flower, D.R. (2004) Coupling *in silico* and *in vitro* analysis of peptide-MHC binding, a bioinformatic approach enabling prediction of superbinding peptides and anchorless epitopes. J. Immunol. 172:7495-7502.

Falk, K., Rötzschke, O., Stevanovic, S., Jung, G., and Rammensee, H.G. (1991) Allele-specific motifs revealed by sequencing of self-peptides eluted from MHC molecules. Nature 351:290-296.Flower, D.R. (2003) Towards *in silico* prediction of immunogenic epitopes. Trends Immunol. 24:667-674.

Fremont, D.H., Matsumura, M., Stura, E.A., Peterson, P.A., and Wilson, I.A. (1992) Crystal structures of two viral peptides in complex with murine MHC class I H-2Kb. Science 257:919-927.

Fremont, D.H., Stura, E.A., Matsumura, M., Peterson, P.A., and Wilson, I.A. (1995) Crystal structure of an H-2Kb-ovalbumin peptide complex reveals the interplay of primary and secondary anchor positions in the major histocompatibility complex binding groove. Proc. Natl. Acad. Sci. USA 92:2479-2483.

Fremont, D.H., Monnaie, D., Nelson, C.A., Hendrickson, W.A., and Unanue, E.R. (1998) Crystal structure of I-Ak in complex with a dominant epitope of lysozyme. Immunity 8:305-317.

Guan, P., Doytchinova, I.A., Zygouri, C., and Flower, D.R. (2003a) MHCPred, bringing a quantitative dimension to the online prediction of MHC binding. Appl. Bioinformatics 2:63-66.

Guan, P., Doytchinova, I.A., and Flower, D.R. (2003b) HLA-A3 supermotif defined by quantitative structure-activity relationship analysis. Protein Eng. 16:11-18.

Guan, P., Doytchinova, I.A., and Flower, D.R. (2003c) A comparative molecular similarity indices (CoMSIA) study of peptide binding to the HLA-A3 superfamily. Bioorg. Med. Chem. 11:2307-2311.

Guan, P., Doytchinova, I.A., Zygouri, C., and Flower, D.R. (2003d) MHCPred, a server for quantitative prediction of peptide-MHC binding. Nucleic Acids Res. 31:3621-3624.

Hattotuwagama, C.K., Guan, P., Doytchinova, I.A., Zygouri, C., and Flower, D.R. (2004) Quantitative online prediction of peptide binding to the major histocompatibility complex. J. Mol. Graph. Model. 22:195-207.

Hennecke, J., and Wiley, D.C. (2002) Structure of a complex of the human α/β T-cell receptor (TCR) HA1.7, influenza hemagglutinin peptide and major histocompatibility complex class II molecule, HLA-DR4 (DRA*0101 and DRB1*0401), insight into TCR cross-restriction and alloreactivity. J. Exp. Med. 195:571-581.

Hudrisier, D., Mazarguil, H., Laval, F., Oldstone, M.B.A., and Gairin, J.E. (1996) Binding of viral antigens to major histocompatibility complex class I H-2Db molecules is controlled by dominant negative elements at peptide non-anchor residues. Implications for peptide selection and presentation. J. Biol. Chem. 271:17829-17836.

Hunt, D.F., Michel, H., Dickinson, T.A., Shabanowitz, J., Cox, A.L., Sakaguchi, K., and Appella, E. (1992) Peptides presented to the immune system by the murine class II major histocompatibility complex molecule I-Ad. Science 256:1817-1820.

Jardetzky, T.S., Lane, W.S., Robinson, R.A., Madden, D.R., and Wiley, D.C. (1991) Identification of self peptides bound to purified HLA-B27. Nature 353:326-329.

Klebe, G., Abraham, U., and Mietzner, T. (1994) Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. J. Med. Chem. 37:4130-4146.

Klebe, G., and Abraham, U. (1999) Comparative molecular similarity index analysis (CoMSIA) to study hydrogen-bonding properties and to score combinatorial libraries. J. Comput. Aided Mol. Des. 13:1-10.

Kubinyi, H., and Kehrhahn, O.H. (1976) Quantitative structure-activity relationships. 3.1 A comparison of different Free-Wilson models. J. Med. Chem. 19:1040-1049.

Li, Y., Li, H., Martin, R., and Mariuzza, R.A. (2000) Structural basis for the binding of an immunodominant peptide from myelin basic protein in different registers by two HLA-DR2 proteins. J. Mol. Biol. 304:177-188.

Liu, X., Dai, S., Crawford, F., Fruge, R., Marrack, P., and Kappler, J. (2002) Alternate interactions define the binding of peptides to the MHC molecule IAb. Proc. Natl. Acad. Sci. USA 99:8820-8825.

Madden, D.R., Gorga, J.C., Strominger, J.L., and Wiley, D.C. (1991) The structure of HLA-B27 reveals nonamer self-peptides bound in an extended conformation. Nature 353: 321-325.

McFarland, B.J., Sant, A.J., Lybrand, T.P., and Beeson, C. (1999) Ovalbumin (323-339) peptide binds to the major histocompatibility complex class II, I-A(d) protein using two functionally distinct registers. Biochemistry 38:16663-16670.

McSparron, H., Blythe, M.J., Zygouri, C., Doytchinova, I.A., and Flower, D.R. (2003) JenPep, a novel computational information resource for immunology and vaccinology. J. Chem. Inf. Comput. Sci. 43:1276-1287.

Parker, K.C., Bednarek, M.A., and Coligan, J.E. (1994) Scheme for ranking potential HLA-A2 binding peptides based on independent binding of individual peptide side-chains. J. Immunol. 152:163-175.

Rammensee, H.G., Friede, T., and Stevanovic, S. (1995) MHC ligands and peptide motifs, first listing. Immunogenetics 41:178–228.

Rudensky, A.Y., Preston-Hurlburt, P., Hong, S.-C., Buus, S., and Tschinke, V. (1991) Sequence analysis of peptides bound to MHC class II molecules. Nature 353:622-627.

Ruppert, J., Sidney, J., Celis, E., Kubo, R.T., Grey, H.M., and Sette, A. (1993) Prominent role of secondary anchor residues in peptide binding to HLA-A*0201 molecules. Cell 74:929-937.

Saper, M.A., Bjorkman, P.J., and Wiley, D.C. (1991) Refined structure of the human histocompatibility antigen HLA-A2 at 2.6A resolution. J. Mol. Biol. 219:277-319.

Schönbach, C., Koh, J.L.Y., Sheng, X., Wong, L., and Brusic, V. (2000) FIMM, a database of functional molecular immunology. Nucleic Acids Res. 28:222-224.

Sette, A., Sidney, J., del Guercio, M.-F., Southwood, S., Ruppert, J., Dalberg, C., Grey, H.M., and Kubo, R.T. (1994a) Peptide binding to the most frequent HLA-A class I alleles measured by quantitative molecular binding assays. Mol. Immunol. 31:813-822.

Sette, A., Vitiello, A., Reherman, B., Fowler, P., Nayersina, R., Kast, W.M., Melief, C.J., Oseroff, C., Yuan, L., and Ruppert, J. (1994b) The relationship between class I binding affinity and immunogenicity of potential cytotoxic T cell epitopes. J. Immunol. 153:5586–5592.

Sidney, J., del Guercio, M.F., Southwood, S., Engelhard, V.H., Appella, E., Rammensee, H.G., Falk, K., Rötzschke, O., Takiguchi, M., and Kubo, R.T. (1995) Several HLA alleles share overlapping peptide specificities. J. Immunol. 154:247–259.

Sidney, J., Grey, H.M., Southwood, S., Celis, E., Wentworth, P.A., del Guercol, M.F., Kubo, R.T., Chestnut, R.W., and Sette, A. (1996) Definition of an HLA-A3-like supermotif demonstrates the overlapping peptide-binding repertoires of common HLA molecules. Hum. Immunol. 45:79–93.

Silver, M.L., Guo, H.C., Strominger, J.L., and Wiley, D.C. (1992) Atomic structure of a human MHC molecule presenting an influenza virus peptide. Nature 360:367-369. Smith, K.J., Reid, S.W., Harlos, K., McMichael, A.J., Stuart, D.I., Bell, J.I., and Jones, E.Y. (1996a) Bound water structure and polymorphic amino acids act together to allow the binding of different peptides to MHC class I HLA-B53. Immunity 4:215-228.

Smith, K.J., Reid, S.W., Stuart, D.I., McMichael, A.J., Jones, E.Y., and Bell, J.I. (1996b) An altered position of the alpha 2 helix of MHC class I is revealed by the crystal structure of HLA-B*3501. Immunity 4:203-213.

Stahle, L., and Wold, S. (1988) Multivariate data analysis and experimental design in biomedical research. Prog. Med. Chem. 25:291-338.

Sybyl 6.9, Tripos Inc., 1699 Hanley Road, St. Louis, MO 63144, USA.

Takamiya, Y., Schönbach, C., Nokihara, K., Yamaguchi, M., Ferrone, S., Kano, K., Egawa, K., and Takiguchi, M. (1994) HLA-B*3501-peptide interactions, role of anchor residues of peptides in their binding to HLA-B*3501 molecules. Int. Immunol. 6:255–261.

Vidal, K., Daniel, C., Vidavsky, I., Nelson, C.A., and Allen, P.M. (2000) Hb (64-76) epitope binds in different registers and lengths to I-Ek and I-Ak. Mol. Immunol. 37:203-212.

Wold, S. (1995) PLS for multivariate linear modelling. In: H. van de Waterbeemd (Ed.), *Chemometric Methods in Molecular Design*. VCH, Weinheim, Germany, pp. 195-218.

Young, A.C., Zhang, W., Sacchettini, J.C., and Nathenson, S.G. (1994) The three-dimensional structure of H-2Db at 2.4 A resolution, implications for antigen-determinant selection. Cell 76:39-50.

Young, D. (2001) *Computational Chemistry, A Practical Guide for Applying Techniques to Real World Problems.* Wiley InterScience, New York.

Zhang, Q.J., Gavioli, R., Klein, G., and Masucci, M.G. (1993) An HLA-A11-specific motif in nonamer peptides derived from viral and cellular proteins. Proc. Natl. Acad. Sci. USA 90:2217-2221.

Zhang, W., Young, A.C., Imarai, M., Nathenson, S.G., and Sacchettini, J.L. (1992). Crystal structure of the major histocompatibility complex class I H-2Kb molecule containing a single viral peptide, implications for peptide binding and T-cell receptor recognition. Proc. Natl. Acad. Sci. USA 89:8403-8407.

Chapter 5

Allergen Bioinformatics

Bernett T.K. Lee[1] and Vladimir Brusic[2,3]

[1] Department of Biochemistry, Yoon Loo Lin School of Medicine, National University of Singapore, 8 Medical Drive, MD7 #02-03, Singapore 117597, bernettl@bii.a-star.edu.sg
[2] Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613
[3] Present address: Cancer Vaccine Center, Dana-Farber Cancer Institute, 77 Avenue Louis Pasteur, HIM 418, Boston, MA 02115, USA, Vladimir_Brusic@dfci.harvard.edu

Abstract. Allergies are a growing health problem in developed and developing countries that result in increased healthcare expenditures. This problem is further compounded by increasing number of allergens found in genetically modified (GM) food and allergens found in unexpected sources (hidden allergens). The importance of allergies has prompted the use of new methods like genomics, proteomics, and microarray in understanding the nature of allergies. These methods have generated large amounts of data that have to be stored, retrieved, and analysed using bioinformatics approaches. Several specialized public databases have been created in response to increasing allergen data. These specialized databases integrate the various information found in general databases into a coherent set of data and provide bioinformatics tools suitable for further analysis. The resources provided by these databases have paved the way for the creation of specialized bioinformatics tools that allow for the prediction of allergenicity. These prediction tools are crucial in view of the new sources of allergens, namely, hidden allergens and potential allergens in the form of recombinant proteins in GM food. Here we review the bioinformatics resources and tools available for the study of allergenicity.

5.1 Introduction

Allergy is a condition where the immune system responds adversely to certain substances that are commonly considered harmless. In recent years, studies have indicated that allergy has become a serious problem in industrial nations, affecting a significant portion of the population (Jansen, Kardinaal, Huijbers, Vlieg-Boerstra, Martens, and Ockhuizen 1994; Malone, Lawson, Smith, Arrighi, and Battista 1997; Larche 2000; Kanny, Moneret-Vautrin, Flabbee, Beaudouin, Morisset, and Thevenin 2001). In fact, allergies are now the most common cause of chronic illness in industrial countries (Larche 2000). Food allergies alone have been found to affect 2.4% (Jansen et al. 1994) and 3.24% (Kanny et al. 2001) of the Dutch and French populations, respectively. This has increased

healthcare costs involved in the detection and treatment of allergies. Thirty-nine million people in the United States suffer from allergic rhinitis but only 12.3% seek medical attention. Nevertheless, this has led to $1.23 billion in healthcare costs (Malone et al. 1997). The percentages of people who seek medical attention are likely to rise leading to even higher healthcare costs.

The problem is further compounded by the fact that in addition to the natural sources of allergens like house dust mites and pollen, introduction of recombinant proteins made possible by molecular genetics into food, medicine, and other products is increasing the number of potential allergens in our environment. The allergenicity of these new recombinant proteins is unknown and this has made safety issues about products containing these proteins paramount. In addition to recombinant proteins, hidden allergens are also found in unexpected sources that people typically do not guard against. For example, milk proteins in processed food are a source of hidden potential allergens that most people would not suspect (Cantani 1999).

In view of these safety concerns, both the FAO (Food and Agriculture Organization) and WHO (World Health Organization) have jointly produced a procedure for evaluating potential allergenicity for any novel protein (FAO/WHO 2001; FAO/WHO 2003). This scheme involves the use of bioinformatics as an initial step to determine whether the protein in question has any allergenicity potential. This is accomplished by determining whether the primary sequence of the novel protein bears significant sequence similarity to another known allergen. The significance is measured as either a greater than 35% similarity over a window of 80 amino acids or a stretch of 6 to 8 identity amino acids to any known allergens.

The increasing importance of allergy has also fueled extensive research in this field and generated large amounts of data. This has been reflected by the number of research articles appearing in the literature. Data contained in PubMed indicate that in the period from 1993 to 2003, the number of allergen articles per year has doubled to 1023 in the year 2003. The rapid growth of sequence information in major public databases like GenBank (Benson, Karsch-Mizrachi, Lipman, Ostell and Wheeler 2003) and Swiss-Prot (O'Donovan, Martin, Gattiker, Gasteiger, Bairoch, and Apweiler 2002) has also contributed significant amounts of allergen-related sequence information. There is also growth in the number of allergen 3D structures although the growth is not as spectacular as that of the sequence databases.

Traditionally, bioinformatics applications have been used in the analysis of individual allergens (Izumi, Sugiyama, Matsuda, and Nakamura 1999; Mills, Hart, Lynch, Thomas, and Smith 1999; Ichikawa, Vailes, Pomes, and Chapman 2001; Iyer, Koonin, and Aravind 2001). In this aspect, bioinformatics applications like sequence similarity searches (Mills et al. 1999; Ichikawa et al. 2001), protein structure comparison (Iyer et al. 2001), sequence profile searches (Iyer et al. 2001), multiple sequence alignments (Iyer et al. 2001), secondary structure prediction (Izumi et al. 1999), protein sequence analysis (Izumi et al. 1999), and homology modeling (Ichikawa et al. 2001) have greatly aided the study of allergens by providing further insights to the workings of allergens. The applications of these methods are similar to those used in other fields and we will not go into details. Instead, we will discuss the various specific issues involved in the management of allergen data as well as some specific recent

implementations of allergen databases. In addition, we will also touch on one of the main bioinformatics applications of these data, namely, allergenicity prediction.

5.2 Allergen Databases

The primary databases like GenBank/EMBL/DDBJ (O'Donovan et al. 2002; Kulikova, Aldebert, Althorpe, Baker, Bates, Browne, van den Broek, Cochrane, Duggan, Eberhardt, Faruque, Garcia-Pastor, Harte, Kanz, Leinonen, Lin, Lombard, Lopez, Mancuso, McHale, Nardone, Silventoinen, Stoehr, Stoesser, Tuli, Tzouvara, Vaughan, Wu, Zhu, and Apweiler 2004; Miyazaki, Sugawara, Ikeo, Gojobori, and Tateno 2004), Swiss-Prot, Protein Data Bank (PDB) (Bourne, Addess, Bluhm, Chen, Deshpande, Feng, Fleri, Green, Merino-Ott, Townsend-Merino, Weissig, Westbrook, and Berman 2004), and PubMed now provide large amounts of publicly available data of various types. Primary databases are the first-stop depositories for biological data and as such are more comprehensive and well maintained. GenBank/EMBL/DDBJ are the major providers of nucleotide sequences. Most nucleotide sequences described in research articles are required to be deposited in any one of these databases. As the data in these three databases are synchronized, they contain virtually the same data. In addition to the requirement by journals on the deposit of nucleotide sequences into these databases, the rapid advances in sequencing technology have tremendously increased the amount of information present in these databases. From 1982 to 2004, the amount of bases in GenBank has doubled every 14 months. This is also reflected in the translated protein sequences derived from the nucleotide sequences available as GenPept, TrEMBL, and DAD. Swiss-Prot, a primary protein sequence database, has experienced lower growth rates due to its manually curated nature. However, its size is still growing at a rapid rate. Release 52.3 (Apr 2007) contains 264,492 protein sequence entries. The manually curated nature of Swiss-Prot provides for quality and rich annotations that have made it popular for specialized allergen databases. The complexity involved in 3D protein structure determination means that there is far less 3D structure information contained in PDB. However, the data contained in PDB, like the rest of the primary databases, is growing and this has placed more 3D structure information on allergens in the hands of researchers. PubMed is a large store of literature information. Sequence and 3D structure information are usually deposited into the previously mentioned primary databases, while the literature contains other types of data that are not found in these primary databases. Some of the information that is of interest in the field of allergens are cross-reactivity data, clinical relevance, and antigen epitopes. Together, these four types of primary databases serve as the primary data source for most if not all specialized allergen databases.

5.2.1 Need for Specialized Databases

Most specialized allergen databases derive their information from the primary databases and provide additional features dedicated to the allergen research community. Since primary databases are meant to be central depositories for

biological data, they do not provide specific subsets of the data. Although the search and retrieval tools of general databases allow to some extent the extraction of allergen data, it takes multiple steps and the results may contain records irrelevant for allergen research. For example, GenBank keyword searches are not sufficiently specific and result in large numbers of false positives (Malandain 2004).

Specialized databases collect allergen-specific data from primary databases and validate the records to ensure they refer to genuine allergens. Some of the primary databases do not perform quality assurance of their data. GenBank only requires that submitters check their records prior to submission. This means that the data may be of low quality, requiring additional validation by the specialized databases. This is one of the reasons why manually curated databases with high-quality data, like Swiss-Prot, are popular with developers of specialized databases.

Most primary databases cover only a certain type of biological data. For example, GenBank is focusing on nucleotide sequences. This presents a problem, as contemporary research is multifaceted and requires different types of biological data. This need is fulfilled by allergen databases that collect allergen-specific data from multiple sources and aggregate them for the benefit of the researcher. Thus, specialized allergen databases serve as a one-stop shop for researchers.

A relatively large amount of allergen data that are required by researchers, such as information on epitopes, cross-reactivity, and clinical phenotypes, are only present in the literature. Although PubMed provides text search and retrieval functions, the information contained in the literature is unstructured, making automated extraction difficult. In addition, PubMed is limited to abstracts only. While PubMed may serve as an initial resource for locating allergen-related literature, expert annotation using full-text literature is often required to extract allergen-specific information. This is a time-consuming process but provides invaluable information that would otherwise be unavailable.

The allergen information in the primary databases is also void of any form of classifications. Classifications are useful to researchers because they partition the data into meaningful subsets that can be independently analysed and used for deriving generalizations or improving database search functions. The most common form of allergen classification is based on the allergen source, for example, food allergen.

Search tools of allergen databases should be better than the standard tools of primary databases. Often, allergen databases have search tools that use fields relevant to allergies. The adaptation of meaningful search fields and terms allows researchers to quickly and accurately extract the desired information. In addition, allergen databases integrate allergen-specific bioinformatics applications to aid researchers in the analysis of allergens. Often, primary databases do not provide ready-to-use tools but require researchers to use their own computational tools. This can be a lengthy process involving the creation of specific allergen datasets followed by the computation itself. In contrast, allergen databases already contain the allergen datasets and can easily integrate existing bioinformatics applications to provide user-friendly analysis tools to the research community.

5.2.2 Desired Features of Allergen Databases

One of the main desired features of an allergen database is the aggregation of all publicly available allergen-specific information into a comprehensive resource. This aggregation activity should take note of the following points:

1. The database should aim to be as comprehensive as possible. In practice, the creation of a one-stop resource for all allergen information is a nontrivial task. There are already allergen databases that cater to specific needs. Besides, it would require huge efforts and resources to create and maintain a comprehensive database that only a few groups could afford.
2. The records contained in the database should be nonredundant and steps should be taken to ensure this. Redundancy is leading to over- and underrepresentation of data that can cause errors in the allergen analyses. This is particularly important if the records are used as training sets for allergenicity prediction. Moreover, redundancy leads to false estimates of true known allergens. Sequence similarity methods like BLAST (Altschul, Gish, Miller, Myers, and Lipman 1990) can be effectively used to reduce sequence redundancy by searching for similar sequence records.
3. Each source database contains different types of biological data necessitating the design of a common data format that can encompass all available information.
4. The fields contained in the records should be useful for allergen researchers. Therefore, the design of the record format should take into account the expected usage. Some of the common fields required include nucleotide sequence, protein sequence, literature references, and 3D protein structure.
5. As far as possible the allergen names should comply with the nomenclature (King, Hoffman, Lowenstein, Marsh, Platts-Mills, and Thomas 1994) set out by the Allergen Nomenclature subcommittee of the IUIS (International Union of Immunological Societies). Allergens contained in the IUIS allergen list should be used with its official names to prevent naming conflicts.
6. The use of multiple source databases may lead to conflicting data. Manual curation would then be required to resolve these conflicts.
7. There is a need to update the allergen database whenever there are changes or updates in the source databases. The propagation of information from the source databases to the specialized allergen databases ensures that the database is current.
8. Some allergen information is only present in the literature and the lack of a structured form of literature data necessitates the manual extraction of this information. This requires large amounts of time and effort.
9. The source databases may contain errors that have to be validated. In most cases, the validation has to be done manually. Again, like information extraction from the literature, this requires both time and effort.

In view of these factors, the aggregation process should be performed as a two-step process. The first step would be to aggregate the information present in the source databases to a format that encompasses all the required and useful fields. As far as

possible, this should be done computationally to ensure synchronization of the data with the source databases. This is not always possible, especially when information is extracted from the literature. The second step would involve curators who manually validate the data and resolve any arising conflicts. The main advantage of this approach is that the majority of the work can be done computationally, thereby allowing the curators to focus on the validation. This strategy reduces the overall amount of effort while maintaining a high level of quality.

Access to the allergen databases is another required feature. Different users have different types of access requirements and the allergen databases should aim to satisfy all the various needs. The wet lab biologists generally require a Web interface access to the individual records. Relevant search engine facilities are required to enable the quick location of records of interest. The records should be presented in a manner that is easy to interpret. Suitable data visualization methods should be used for data that is difficult to represent textually. An example would be 3D protein structure information that are usually presented in a protein structure viewer.

Bioinformaticians studying allergens would need a different type of database access. Bioinformatics analysis typically requires large sets of data rather than individual records in order to extract meaningful results. Moreover the information contained in these records must be in a computer-readable form. Therefore, the format of the records is far more important to the bioinformatician than to the wet lab biologist. At the very least, the records have to be presented in some structured form. A structured record would allow for the efficient parsing of the information into a computer-readable form for further computational analysis. The extensible markup language XML is ideal for this purpose because most biological data have few issues being represented in this form. Furthermore, the provision of an XML scheme would permit rapid parsing and validation of the records. For efficient linking of database records to other resources, access to the individual records in the database should also be available as hyperlinks.

An allergen database should also provide analysis tools capitalizing on the underlying data that it contains to service the research community. The reasons for this have been discussed in the previous section.

5.2.3 Existing Allergen Databases

An excellent review of existing databases was published in 2003 (Brusic, Millot, Petrovsky, Gendel, Gigonzac, and Stelman 2003). Here, we exclude the reviewed databases except for the IUIS list and Swiss-Prot list of allergens and highlight recent additions to the growing list of allergen databases (Table 1).

Most of the databases covered in the review article are lacking one or more desired features of an allergen database. Only a few databases provide bioinformatics tools and permit the downloading of data. Furthermore, many of the databases described are not actively maintained and lag behind in recording new allergens or changes to existing allergen information.

Table 1. List of allergen databases and their URLs.

Name	URL
Allallergy	http://allallergy.net/
Allergome (Mari and Riccioli 2004)	http://www.allergome.org/
BIFS (Gendel 1998) (Biotechnology Information for Food Safety)	http://www.iit.edu/~sgendel/fa.htm
CSL (Central Science Laboratory) allergen database	http://allergen.csl.gov.uk/
FARRP (Food Allergen Resarch and Resource Program) allergen database	http://www.allergenonline.com/
IUIS List	http://www.allergen.org/
Protall	http://www.ifr.bbsrc.ac.uk/protall/
SDAP (Ivanciuc, Schein and Braun 2003)	http://fermi.utmb.edu/SDAP/index.html
Swiss-Prot allergen list	http://www.expasy.org/cgi-bin/lists?allergen.txt

5.2.3.1 IUIS

The Allergen Nomenclature subcommittee of the IUIS maintains an official list of allergens and isoallergens that conforms to the allergen nomenclature. The nomenclature specifies that the first three characters of the allergen name, for example Bet v1, are derived from the genus name (Bet = Betula). The next character denotes the species name (v = verrucosa). The number at the end of the allergen name indicates the order in which the allergen was identified. Isoallergens (i.e., Bet v 1.0101) have an additional dot followed by four additional numbers. The first two numbers refer to the isoallergen. The third and fourth numbers indicate the particular variant of the isoallergen. The list is available on the Internet (http://www.allergen.org/Allergen.aspx) and is updated periodically. Researchers can submit new allergens for inclusion into the list but the allergens must satisfy a prevalence of IgE reactivity of at least 5% or a minimum of five patients showing IgE reactivity. This ensures that the allergens contained in the list are clinically relevant. As of March 2007, the list contained 574 allergens and 869 isoallergens.

The allergens are classified according to the allergen source and each record contains the species name, allergen name, protein name, molecular weight, type of sequence, database accession, and literature references. Most other allergen databases require access to the IUIS list. Nevertheless, the list is maintained as an HTML file with an Excel-readable download version. The lack of a structured format (i.e., XML) makes parsing of the information inconvenient and error-prone.

5.2.3.2 Swiss-Prot

Swiss-Prot maintains a list of allergens that currently numbers 347 entries (Release 52.3 of Apr 2007). Each allergen contains a link to the Swiss-Prot record, therefore the contents of each record are the same as those of the original Swiss-Prot record. The names of the allergens are in accordance with the nomenclature set out by IUIS.

5.2.3.3 SDAP

SDAP (Structural Database of Allergenic Proteins) is a specialized allergen database that incorporates information obtained from the IUIS list of allergens, Swiss-Prot, PIR (Protein Information Resource), GenBank, GenPept, and the literature. However, the entries in SDAP are guided by the IUIS list of allergens and isoallergens. SDAP which has been static since January 2005 contains 737 allergens and isoallergens, 829 protein sequences, and 22 IgE and IgG epitopes.

Each record in SDAP contains the name of the allergen, the species that the allergen originates from, protein sequences, nucleotide sequences, Pfam (Bateman, Coin, Durbin, Finn, Hollich, Griffiths-Jones, Khanna, Marshall, Moxon, Sonnhammer, Studholme, Yeats, and Eddy 2004), protein domains, 3D protein structure, and IgE epitopes. The IgE epitope information was extracted from the literature. The epitope information makes SDAP rather unique and useful, as it is one of the few databases that contain this information. Records in SDAP can be searched by their names, the allergen source, the description, and the allergen type. In addition to the search facilities, there are compilations of allergens according to alphabetical order, allergens containing PDB structures, allergens containing 3D models, allergens containing epitopes, and various classes of allergens.

Other than containing a comprehensive set of allergen data, SDAP includes computational tools for FASTA (Pearson 1994), sequence similarity search, allergen analysis, and allergenicity prediction. The FAO/WHO allergenicity test evaluates the allergenicity of a given protein sequence against the dataset present in SDAP. In addition, SDAP includes two unique data searching tools. The first is an exact matching tool for searching a query protein sequence against SDAP. This method is useful if the query protein sequence is an epitope sequence. Any SDAP allergens having the same subsequence as the query epitope sequence will be retrieved. The result may then be used as evidence for cross-reactivity between the query sequence and the matched SDAP allergen. If the SDAP allergen has a defined epitope, the search result also provides a link to it.

However, the exact matching method is limited to detecting allergens with identical epitopes. This is hardly practical as the same IgE molecule may bind to two similar but not identical epitopes as seen in cross-reactivity. The second method employs a property distance function (Venkatarajan and Braun 2001) to score the similarity of two peptides. This is a far better solution as it takes into account the level of degeneracy in the epitope sequence. The property distance function employs five descriptors E_1–E_5 that were derived from 237 amino acid properties. The property distance (PD) function for two amino acid sequences A and B of length N would be:

$$PD(A,B) = \frac{1}{N}\sum_{i=1}^{N}\left[\sum_{j=1}^{5}\lambda_j\left(E_j(A_i) - E_j(B_i)\right)^2\right]^{1/2} \quad (1)$$

where λ_j is the eigenvalue of the jth E component, $E_j(A_i)$ is the E_j value for the amino acid in the ith position from sequence A, and $E_j(B_i)$ is the E_j value for the amino acid in the ith position from sequence B.

For any given novel protein sequence, the PD function is used to determine the similarity measure of the novel protein sequence against all same length subsequences in SDAP. The results are then ranked and displayed together with a histogram to the user. The histogram aids the user in determining the significance of the results. If a match is detected that has a much lower similarity measure than the rest of the matches, the match may be significant and should be analysed further. The PD method has been successfully used to detect cross-reactivity among allergens on the basis that they share similar epitopes. Similar to other allergen databases, downloading of the data is not supported. This is a major problem for bioinformaticians who require sets of data for further analysis.

5.2.3.4 Allergome

Allergome (Mari and Riccioli 2004) was started in 2000 and released in February 2003. All the records are manually curated by experts. The primary data source for Allergome is literature published since the early 1960s. As of February 2007, Allergome contains more than 9000 references categorized by topics and allergens (Mari, Scala, Palazzo, Ridolfi, Zennaro , and Carabella 2007).

Not all the allergens in Allergome are found in the IUIS list of allergens and records in Allergome clearly demarcate which records are in the IUIS list. Allergens that are not in the IUIS list are carefully checked to ensure that they are valid allergens prior to the addition in Allergome.

Allergome records integrate literature data and are therefore very information-rich. Each record may contain the allergen name, common names, biological functions, links to primary sequence information, links to PDB structures, sequence motifs, source of allergen, tissue source of allergen, route of exposure, allergen isoforms, prevalence of allergy, references, molecular weight, sequence homologues, posttranslational modifications, test of allergenicity both *in viv*o and *in vitro*, cross-reactivity, recombinant forms of the allergen, and literature references.

Allergome has user-friendly search facilities. A quick search using keywords enables users to filter the results in several ways to display, for example, only IUIS-listed allergens. The advanced search allows users to search specific fields using Boolean modifiers. Similar to the quick search, users can filter their results. In addition to the search facilities, Allergome provides lists of allergens sorted by categories.

Download facilities are also lacking in Allergome. This is particularly dire in the case of Allergome because the richness of the data cannot be easily exploited for large-scale bioinformatics analysis. This is further compounded by the fact that Allergome does not contain any computational tools that could take advantage of these data.

5.2.4 Pitfalls of Current Allergen Databases

Current allergen databases fulfill most of the desired features, except for the availability of a download feature. Although it is possible to parse HTML pages or Excel files to extract the information, it is rather error-prone. HTML pages are structured, but most of the structures are used for describing the appearance rather than the type of content. As a result, bioinformaticians have to re-create the allergen datasets from scratch rather than using existing datasets. Not only is it time consuming and repetitive, it also precludes the creation of a standard set of data for the development of new bioinformatics methods and analysis.
The lack of a standard set of data means that developed methods and analysis results cannot be easily compared to one other, thus hindering the overall progress of development. In contrast, downloadable dataset formats like that provided by GenBank efficiently support bioinformaticians in developing new analysis methods. The adoption of such features by the allergen databases would enhance the bioinformatic tool development for allergen research. The allergen databases could then serve as a platform for the development of new methods and large-scale analysis.

5.3 Allergenicity Prediction

The holy grail of applying bioinformatics to allergen research is the prediction of allergenicity. Accurate prediction of allergenicity is likely to improve the allergenicity assessment of recombinant proteins, thereby lowering the allergenicity testing cost of recombinant proteins. Considering the spread of recombinant protein use in food, medications, and everyday items, the impact of predictive methods is expected to be huge.

Predictive methods are often compared on the basis of their precision and recall. Precision is the ability of the method to correctly predict true allergens among the predicted allergens. Precision is usually expressed as a percentage of the correctly predicted allergens over all the predicted allergens. Recall, on the other hand, is the ability of the method to detect for allergens in the test set. Recall is expressed as the percentage of correct predicted allergens over all the allergens in the test set. The equations for precision and recall are provided below. A high precision would mean that any predicted allergen is likely to be a true allergen while a high recall means that the method is able to correctly predict a large portion of the allergens in the test set. In practice, a trade-off is usually required as it is not possible for one to get both high precision and high recall.

$$precision = \frac{tp}{tp + fp}, \quad recall = \frac{tp}{tp + fn} \tag{2}$$

where *tp* is true positive (a correctly predicted allergen), *fp* is false positive (a non-allergen that has been wrongly predicted to be an allergen) and *fn* is false negative (an allergen wrongly predicted to be a nonallergen).

Allergenicity prediction has been attempted by several groups. The methods are explained and discussed in the next section.

5.3.1 Sequence Similarity Searches

Sequence similarity search methods are an obvious approach to predicting allergenicity. If two proteins are highly similar and one of the proteins is an allergen, the likelihood of the second protein being an allergen is high. Sequence similarity search methods are very mature and easy to implement. However, allergenicity is determined by the binding of epitopes. Since epitopes are in general subsequences of the entire protein sequence, local alignments like BLAST and FASTA (Pearson 1994) tend to perform better. In fact, the FAO/WHO guidelines implement local alignment methods. Sequence similarity search methods are implemented in SDAP and FARRP as means to query the content.

Even for identifying cross-reacting allergens, local alignment methods are useful because cross-reacting allergens are generally more than 70% identical in their sequences (Aalberse 2000). With a few exceptions, cross-reacting allergens share similar protein structures (Aalberse and van Ree 1996). The main drawback of sequence similarity searches is that their performance is limited to linear epitopes. Conformational epitopes, unlike linear epitopes, are not composed of consecutive amino acids. Therefore, sequence similarity search methods are not applicable to conformational epitopes. The similarity in 3D structure often translates to similar primary protein sequences.

Another drawback of sequence similarity search is its dependence on the coverage of the dataset the query sequence is searched against. This makes the detection of novel allergens difficult and requires comprehensive and constantly updated allergen databases. Nevertheless, sequence similarity search methods have also been implemented as a last-resort method for the prediction of allergenicity (Stadler and Stadler 2003; Li, Issac, and Krishnan 2004).

5.3.2 FAO/WHO Guidelines

The FAO/WHO guidelines for allergenicity predictions involve a bioinformatics component. This serves as an initial screening process prior to the use of any laboratory test. Compared to laboratory testing, allergenicity prediction using bioinformatics approaches is comparatively fast and simple. Implementation of the guidelines is relatively simple and the computational complexity is low. However, the method due to its reliance on primary protein sequences similarity is only as good as the underlying dataset used. This means that allergen databases that contain the allergen dataset are in a good position to implement these guidelines as a service. In practice, only SDAP and Allermatch (Fiers, Kleter, Nijland, Peijnenburg, Nap, and van Ham 2004) implemented the guidelines.

The FAO/WHO guidelines produce results that tend to have very high recall and very low precision (Hileman, Silvanovich, Goodman, Rice, Holleschak, Astwood, and Hefle 2002; Kleter and Peijnenburg 2002; Stadler and Stadler 2003; Soeria-Atmadja, Zorzet, Gustafsson, and Hammerling 2004). Consequently, the

method is unlikely to miss any true allergens (a required feature), but is likely to generate large amounts of false positives that may very well overwhelm the laboratory testing capabilities.

5.3.3 Supervised Classification Approaches

Recently, supervised classification approaches have been adopted for allergenicity prediction (Soeria-Atmadja et al. 2004). The study employed three different supervised algorithms, namely, the kNN classifier, the Bayesian linear Gaussian classifier, and the Bayesian quadratic Gaussian classifier. The methods were trained on a set of local alignments produced by FASTA. The feature vector consists of the alignment length and score extracted from the best alignment obtained by FASTA. Training data for the study included both positive and negative datasets.

The results of the study indicate that the Bayesian linear Gaussian classifier was the best algorithm, being able to detect 77% of the allergens with a false positive rate of 10%. This was followed by the Bayesian quadratic Gaussian classifier (77% of allergens detected with a false positive rate of 11%) and the kNN classifier (78% of allergens detected with a false positive rate of 13%). The algorithms may be tuned for either high precision or high recall. Tuning the algorithm for high recall would be critical in a screening procedure as false negatives are far less desirable. By combining feature vectors obtained using different scoring matrices, better results were obtained for the Bayesian linear Gaussian classifier allowing it to detect 77% of the allergens with a false positive rate of 8%.

The results obtained look promising, as they allow for much lower false positive rates than those possible with the FAO/WHO guidelines. However, as the method relies on local alignments, conformational epitopes may still present a challenge.

5.3.4 Expectation Maximization

Allergenicity predictions have also been attempted using MEME (Bailey and Elkan 1994), a motif discovery system employing expectation maximization (Stadler and Stadler 2003). The study attempts to locate common motifs among allergens and then to utilize these motifs for allergenicity predictions. The underlying basis is that these identified motifs are indicators of allergenicity.

The method employs MEME in an iterative manner. First, a dataset of 779 non-redundant allergens was created from public databases. Then MEME was applied to this dataset and the most significant motif extracted and converted into a profile. This profile was then used to search the dataset for any matching allergens, which are then removed from the dataset. The remaining allergens are submitted to the next round of motif discovery and removal. In total, 52 motifs were discovered and 644 allergens in the dataset contain one or more of the 52 motifs. Incomplete sequence information for 78 allergens is the main reason why 135 allergens did not yield any motifs. The remaining 57 allergens are thought to be unique allergens. The 52 discovered motifs can be applied to any novel protein sequence to determine the significance of match. Typically, an *E* value of

10^{-8} is used as an indicator of allergenicity. Should the 52 motifs fail to match the protein sequence, a BLAST search is carried out against the 135 allergens without motifs.

The results of a synthetic dataset indicate that the method had a very high recall (100%) and precision (95.5%). This contrasts with the FAO/WHO guidelines, which scored 98.6% for recall and 36.5% for precision. In a more true-to-life scenario, the method has a recall of 100% and precision of 8.6% for the entire Swiss-Prot dataset. The FAO/WHO guidelines, on the other hand, had 100% recall and 0.5% precision. Therefore, although the recall of both methods was comparable, the method employed here showed improved precision.

The ability to retain the high recall while achieving higher precision (>17 times higher) is particularly useful in a screening procedure. A high recall is important as it prevents any potential allergens from slipping through. Moreover, the increased precision reduces the number of false positives and therefore the number of time-consuming laboratory screenings.

5.3.5 Wavelet Transform

Wavelet transform (Krishnan, Li, and Issac 2004) has also been used instead of MEME to extract motifs from allergens for allergenicity predictions (Li et al., 2004). Wavelet transform is used to convert the aligned amino acid sequences into signals where conserved motifs may be detected on different scales.

The study used a set of 664 allergens collected from the IUIS list of allergens, Swiss-Prot allergen list, BIFS, and FARRP. As the wavelet transform method requires a set of aligned sequences sharing a common motif, the method first clusters the allergen sequences into groups. Clustering into groups was achieved by computing the distance between every pair of allergens using ClustalW (Thompson, Higgins, and Gibson 1994). Allergens were then clustered into groups using the "partitioning around mediods" method (Kaufman and Rousseeuw 1990). Within each group of allergens, ClustalW or T-Coffee (Notredame, Higgins, and Heringa 2000) programs were used to generate multiple aligned amino acid sequences. Wavelet transform was applied to the multiple sequence alignment to extract conserved motifs. Then, HMM (Hidden Markov Model) profiles were created from these motifs using the HMMER package. The HMM profiles are used for searching and predicting the allergenicity of novel proteins. About 20% of the allergens in the dataset did not contain any of the motifs. These allergens were subjected to another round of clustering, wavelet transform, and motif extraction. Any remaining allergens that did not contain motifs were stored separately for sequence similarity search using BLAST.

The allergenicity prediction proceeds with the novel protein sequence being subjected to a search using hmmpfam against all discovered motifs. Should the protein sequence contain any of the discovered motifs, it is predicted to be an allergen. If not, a BLAST search is carried out against the allergens in the dataset that do not contain any discovered motifs. Should the BLAST search result in a good match, then the protein sequence is predicted to be an allergen, otherwise the protein sequence is predicted to be a nonallergen. The threshold values for both the motif search using hmmpfam and the BLAST search were set at an *E* value of 0.001.

The results of a 10-fold cross-validation test indicate that the method performs well with a high precision (99.77%) and reasonable recall (70.61%). The inclusion of the BLAST search does little to improve the precision although it does improve the recall (an increase of 7%). Real-life testing using the entire Swiss-Prot database predicted 2042 allergens of which 295 are known to be allergens (recall of 86.5%). The inclusion of the second-stage BLAST search increased the number of predicted allergens to 4768 and the number of correct allergens to 319 (recall of 93.6%). Comparison with the expectation maximization method is difficult due to the differences in the release of Swiss-Prot used (135,850 proteins in this study as compared to 101,602 proteins in the expectation maximization study). However, this method clearly performs better than the FAO/WHO guidelines as far as precision is concerned. The FAO/WHO guidelines predicted 67% of the Swiss-Prot entries to be allergens.

5.3.6 Current Status of Allergenicity Predictions

Current allergenicity prediction methods have improved on the FAO/WHO guidelines. The methods have retained the high recall required in the screening procedure while increasing the precision of the predictive test.

The main problem facing the development of allergenicity prediction methods is the lack of a standard dataset for training and testing. The lack of a standard dataset has made comparison between the various methods difficult and is hindering progress. Another problem is the lack of a good negative dataset comprising nonallergens. This is critical in the assessment of both the precision and recall measure as both measures are influenced by either the number of false positives or false negatives.

5.4 Conclusion

At present a number of specialized allergen databases offer the allergen research community valuable and information-rich resources. However, none of the allergen databases is comprehensive to the point that it is a superset of all the allergen databases. As such, there is still the need for the continued existence of several allergen databases, each providing some unique information. The biggest drawback of the databases is the lack of download facilities with appropriately formatted data suitable for the development of analysis methods and large-scale analysis. This problem is recurring in the allergenicity prediction methods. Although the predictions have improved on the FAO/WHO guidelines, they are still difficult to compare due to the uniqueness of each dataset used for development.

Acknowledgements

We wish to thank Dr. C. Schönbach for the opportunity to contribute this chapter, and our colleagues at the Institute of Infocomm Research for support.

References

Aalberse, R. C., and van Ree, R. (1996) Cross-reactive carbohydrate determinants. Monogr. Allergy 32:78-83.

Aalberse, R. C. (2000) Structural biology of allergens. J. Allergy Clin. Immunol. 106:228-238. Altschul, S. F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990) Basic local alignment search tool. J. Mol. Biol. 215:403-410.

Bailey, T. L., and Elkan, C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc. Int. Conf. Intell. Syst. Mol. Biol. 2:28-36.

Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E.L., Studholme, D.J., Yeats, C., and Eddy, S.R. (2004) The Pfam protein families database. Nucleic Acids Res. 32: D138-D141.

Benson, D. A., Karsch-Mizrachi, I., et al. (2003) GenBank. Nucleic Acids Res. 31:23-27.

Bourne, P. E., Addess, K. J., Bluhm, W.F., Chen, L., Deshpande, N., Feng, Z., Fleri, W., Green, R., Merino-Ott, J.C., Townsend-Merino, W., Weissig, H., Westbrook, J., and Berman, H.M. (2004) The distribution and query systems of the RCSB Protein Data Bank. Nucleic Acids Res. 32:D223-D225.

Brusic, V., Millot, M., Petrovsky, N., Gendel, S.M., Gigonzac, O., and Stelman, S.J. (2003) Allergen databases. Allergy 58:1093-1100.

Cantani, A. (1999) Hidden presence of cow's milk proteins in foods. J. Invest. Allergol. Clin. Immunol. 9:141-145.

FAO/WHO (2001) Allergenicity of Genetically Modified Foods. Food and Agriculture Organization of the United Nations, Rome, Italy, http://www.who.int/foodsafety/ publications/biotech/ec_jan2001/en/.

FAO/WHO (2003) Codex Principles and Guidelines on Foods Derived from Biotechnology. Food and Agriculture Organization of the United Nations, Rome, Italy, http://www. codexalimentarius .net web/more_info.jsp?id_sta=10007.

Fiers, M. W., Kleter, G.A., Nijland, Peijnenburg, Nap, and van Ham (2004) Allermatch, a webtool for the prediction of potential allergenicity according to current FAO/WHO Codex alimentarius guidelines. BMC Bioinformatics 5:133.

Gendel, S. M. (1998) Sequence databases for assessing the potential allergenicity of proteins used in transgenic foods. Adv. Food Nutr. Res. 42:63-92.

Hileman, R. E., Silvanovich, A., Goodman, R.E., Rice, E.A., Holleschak, G., Astwood, J.D., and Hefle, S.L. (2002) Bioinformatic methods for allergenicity assessment using a comprehensive allergen database. Int. Arch. Allergy Immunol. 128:280-291.

Ichikawa, K., Vailes, L. D., Pomes, A., and Chapman, M.D. (2001) Identification of a novel cat allergen─cystatin. Int. Arch. Allergy Immunol. 124:55-56.

Ivanciuc, O., Schein, C. H., and Braun, W. (2003) SDAP: Database and computational tools for allergenic proteins. Nucleic Acids Res. 31:359-362.

Iyer, L. M., Koonin, E. V., and Aravind, L. (2001) Adaptations of the helix-grip fold for ligand binding and catalysis in the START domain superfamily. Proteins 43:134-144.

Izumi, H., Sugiyama, M., Matsuda, T., and Nakamura, R. (1999) Structural characterization of the 16-kDa allergen, RA17, in rice seeds. Prediction of the secondary structure and identification of intramolecular disulfide bridges. Biosci. Biotechnol. Biochem. 63:2059-2063.

Jansen, J. J., Kardinaal, A. F., Huijbers, G., Vlieg-Boerstra, B.J., Martens, B.P., and Ockhuizen, T. (1994) Prevalence of food allergy and intolerance in the adult Dutch population. J. Allergy Clin. Immunol. 93:446-456.

Kanny, G., Moneret-Vautrin, D. A., Flabbee, J., Beaudouin, E., Morisset, M., and Thevenin, F. (2001) Population study of food allergy in France. J. Allergy Clin. Immunol. 108: 133-140.

Kaufman, L., and Rousseeuw, P. J. (1990) *Finding Groups in Data: An Introduction to Cluster Analysis*, John Wiley & Sons, Brussels, Belgium.

King, T. P., Hoffman, D., Lowenstein, H., Marsh, D.G., Platts-Mills, T.A., and Thomas, W. (1994) Allergen nomenclature. WHO/IUIS Allergen Nomenclature Subcommittee. Int. Arch. Allergy Immunol. 105:224-233.

Kleter, G. A., and Peijnenburg, A. A. (2002) Screening of transgenic proteins expressed in transgenic food crops for the presence of short amino acid sequences identical to potential, IgE -binding linear epitopes of allergens. BMC Struct. Biol. 2:8.

Krishnan, A., Li, K. B., and Issac, P. (2004) Rapid detection of conserved regions in protein sequences using wavelets. In Silico Biol. 4:0013.

Kulikova, T., Aldebert, P., Althorpe, N., Baker, W., Bates, K., Browne, P., van den Broek, A., Cochrane, G., Duggan, K., Eberhardt, R., Faruque, N., Garcia-Pastor, M., Harte, N., Kanz, C., Leinonen, R., Lin, Q., Lombard, V., Lopez, R., Mancuso, R., McHale, M., Nardone, F., Silventoinen, V., Stoehr, P., Stoesser, G., Tuli, M.A., Tzouvara, K., Vaughan, R., Wu, D., Zhu, W., and Apweiler, R. (2004) The EMBL Nucleotide Sequence Database. Nucleic Acids Res. 32:D27-D30.

Larche, M. (2000) Specific immunotherapy. Br. Med. Bull. 56:1019-1036.

Li, K. B., Issac, P., and Krishnan, A. (2004) Predicting allergenic proteins using wavelet transform. Bioinformatics 20:2572-2578.

Malandain, H. (2004) Basic immunology, allergen prediction, and bioinformatics. Allergy 59:1011-1012.

Malone, D. C., Lawson, K. A., Smith, D.H., Arrighi, H.M., and Battista, C. (1997) A cost of illness study of allergic rhinitis in the United States. J. Allergy Clin. Immunol. 99:22-27.

Mari, A., and Riccioli, D. (2004) The Allergome Web site ─ A database of allergenic molecules. Aim, structure, and data of a Web-based resource. J. Allergy Clin. Immunol. 113:S301.

Mari, A., Scala, E., Palazzo, P., Ridolfi, S., Zennaro, D., and Carabella, G. (2007) Bioinformatics applied to allergy: Allergen databases, from collecting sequence information to data integration. The Allergome platform as a model. Cell. Immunol Apr 13; [Epub ahead of print].

Mills, K. L., Hart, B. J., Lynch, N.R., Thomas, W.R., and Smith, W. (1999) Molecular characterization of the group 4 house dust mite allergen from *Dermatophagoides pteronyssinus* and its amylase homologue from *Euroglyphus maynei.* Int. Arch. Allergy Immunol. 120:100-107.

Miyazaki, S., Sugawara, H., Ikeo, K., Gojobori, T., and Tateno, Y. (2004) DDBJ in the stream of various biological data. Nucleic Acids Res. 32:D31-D34.

Notredame, C., Higgins, D. G., and Heringa, J. (2000) T-Coffee: A novel method for fast and accurate multiple sequence alignment. J. Mol. Biol. 302:205-217.

O'Donovan, C., Martin, M. J., Gattiker, A., Gasteiger, E., Bairoch, A., and Apweiler, R. (2002) High-quality protein knowledge resource: SWISS-PROT and TrEMBL. Brief Bioinform. 3:275-284.

Pearson, W. R. (1994) Using the FASTA program to search protein and DNA sequence databases. Methods Mol. Biol. 24:307-331.

Soeria-Atmadja, D., Zorzet, A., Gustafsson, M.G., and Hammerling, U. (2004) Statistical evaluation of local alignment features predicting allergenicity using supervised classification algorithms. Int. Arch. Allergy Immunol. 133:101-112.

Stadler, M. B., and Stadler, B. M. (2003) Allergenicity prediction by protein sequence. FASEBJ. 17:1141-1143.

Thompson, J. D., Higgins, D. G., and Gibson, T.J. (1994) CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22:4673-4680.

Venkatarajan, M. S., and Braun, W. (2001) New quantitative descriptors of amino acids based on multidimensional scaling of a large number of physical-chemical properties. J. Mol. Model. 7:445-453.

Chapter 6

Immunoinformatics Applied to Modifying and Improving Biological Therapeutics

Anne S. De Groot,[1,2] Paul M. Knopf,[1,2] Daniel Rivera,[2] and William Martin[2]

[1] Brown University, Department of Medicine, Providence, RI 02912, USA
[2] EpiVax, Inc., 146, Clifford Street, Providence, RI 02903, USA, AnnieD@EpiVax.com

Abstract. Protein therapeutics have recently emerged as a viable means of treating chronic diseases and are beginning to rival small-molecule drugs in market share. Although their promise of targeted therapy is a major medical advance, repeated administrations in many cases lead to development of antitherapeutic antibodies that compromise treatment. Multiple sources of immunogenicity are considered in this chapter with a focus on the T-cell-dependent immune response. Development of high-affinity antibodies depends on activation of T helper cells by antigen presentation. Disruption of antigen presentation in an antigen-specific manner would be a rational solution to this problem. Here we present the powerful combination of recombinant protein expression and immunoinformatic and molecular modeling tools as a means of reducing immunogenicity by modification of T-cell epitopes. This approach promises to bring to the clinic safer protein therapeutics both as first- and second-generation products.

6.1 Introduction

6.1.1 Deimmunization Defined

The number of therapeutic protein products available for use in clinical settings has dramatically increased in recent years. This category of biomedical products, also known as biological therapeutics, includes neuromuscular antagonists such as botulinum toxin, cytokines such as alpha interferon, growth factors, hormones, and monoclonal antibodies. Biological therapeutics are generally considered to be safe and non-toxic. The production and purification of therapeutic proteins has now become extremely efficient, and a number of such products are commonly used in today's standard medical practice.

Unfortunately, the use of these products in clinical practice is often associated with the development of antibodies directed against the therapeutic proteins. These anti-therapeutic protein antibodies may neutralize or otherwise compromise the clinical effect of the biologics and can also be associated with serious adverse events

such as autoimmunity. Examples of anti-biologic therapeutic antibodies include antibodies to botulinum toxin and antibodies to erythropoietin (Vuong and Jankovic 2005; Haselbeck 2003).

Deimmunization, the subject of this chapter, is the process of modifying biological therapeutics so as to diminish the development of anti-biological therapeutic antibodies. Until recently, reduced immunogenicity could only be achieved by careful formulation of the biologic therapeutic, and in some cases, pegylation of the protein (see below). One of the more recent means of deimmunizing proteins is to apply immunoinformatics tools to identify T-cell epitopes, and then to use related tools to select and modify amino acids contained in those T-cell epitopes. These modifications, achieved using immunoinformatics tools followed by *in vitro* and *in vivo* confirmation, may lead to improved biological therapeutics characterized by significantly diminished immunogenicity profiles.

6.1.1.1 Sources of Biologic Therapeutic Immunogenicity

Several mechanisms for the induction of antibodies against therapeutic proteins have been described. These have included (1) inadvertent formulation of the biologic with proinflammatory contaminants or adjuvants (Haselbeck 2003); (2) formation of immunogenic aggregates; and (3) T-cell-dependent antibody formation. Drug manufacturers have developed methods for addressing the first two mechanisms, including improvements in the purity and formulation of recombinant protein products. The third mechanism for anti-therapeutic antibody formulation involves the adaptive arm of the cellular immune system and the presentation of epitopes derived from the therapeutic proteins to T cells in the context of human leukocyte antigen (HLA) molecules.

HLA class II molecules are genetically encoded on chromosome 6 and expressed as cell-surface proteins on antigen-presenting cells (APC). Class II HLA molecules are involved with the presentation of peptide epitopes derived from therapeutic proteins to T cells, engendering a T-cell-dependent immune response. For more information on HLA-epitope binding see Finkelman, Lees, and Morris (1992).

6.1.1.2 Epitope-Directed Deimmunization

This chapter will discuss a new immunoinformatics-driven strategy for reducing the immunogenicity of biological therapeutics, which is to eliminate HLA class II peptide epitopes by minimal amino acid replacement. This application of immunoinformatics to the modification of protein therapeutics is termed “epitope-directed deimmunization”.

Without such epitopes, helper T-cell (T_H) stimulation is avoided; induction of both humoral and cell-mediated adaptive immunity is significantly diminished, along with the potential for adverse clinical events. T_H-mediated induction of immunological memory is also compromised, reducing affinity maturation of antibodies and encouraging apoptosis (death) of the antigen-specific T and B lymphocytes.

Immunoinformatics is but one of the tools that must be applied in the process of epitope-directed deimmunization. Each step in this new approach to therapeutics is

described in this chapter, including methods for identifying the amino acids that need to be changed using immunoinformatics; measurement of binding of the modified peptide to HLA and measurement of T-cell response by ELISpot. Changes to amino acid sequences, however minimal, can have serious consequences on protein function. This chapter will also describe the need for careful structural modeling and confirmation of activity and decreased immunogenicity *in vitro* and *in vivo*. Epitope-directed deimmunization will also be contrasted with other existing techniques such as pegylation.

6.1.2 Dimensions of the Problem

A number of therapeutic proteins have been shown to induce T-cell-dependent IgG antibody responses when used as biologics (Koren, Zuckerman, and Mire-Suis 2002; Stein 2002; Ryff and Schellekens 2002). The prevalence of antidrug antibodies ranges from less than 1% for drugs such as tissue plasminogen activator (Nilsson, Nilsson, Jansson, Boman, Soderberg, and Naslund 2002) to over 70% for drugs such as OKT3, a murine monoclonal antibody (Jensen, Birkeland, Rohrp, Elbirk, and Jorgensen 1996; Chatenoud 1993). The clinical impact of these antibodies can range from no effect (nonneutralizing or binding antibodies) or some loss of efficacy (neutralizing antibodies) to severe reactions such as anaphylaxis.

Antitherapeutic antibodies can also develop to recombinant human protein or a humanized monoclonal antibody, products that should not, in theory, breach tolerance (Diamond 2003; Wadhwa, Mellstedt, Small, and Thorpe 2003). The basis for the development of T-cell responses to autologous proteins is not well understood. However, it is well known that autologous proteins, including many proteins used in therapy, are not devoid of immunogenic potential. T-cell responses to a number of autologous proteins have been associated with autoimmunity, as is clearly the case with diabetes (Fowell and Mason 1993; Reijonen, Novak, Kochik, Heninger, Liu, Kwok, and Nepom 2002) and multiple sclerosis (MS) (Forsthuber, Shive, Wienhold, de Graaf, Spack, Sublett, Melms, Kort, Racke, and Weissert 2001; Keech, Farris, Beroukas, Gordon, and McCluskey 2001).

6.2 Components of the Immune Response to Biologicals

6.2.1 Types of Antibodies to Biological Therapeutics

6.2.1.1 Cross-Reactive Antibodies

One of the most significant safety concerns related to the clinical use of therapeutic proteins is the formation of neutralizing antibodies that cross-react with or neutralize autologous counterparts. For example, cross-reactive antibodies have been associated with the development of aplastic anemia following treatment with recombinant erythropoietin (rEPO). This severe form of anemia was shown to be due to the development of antibodies cross-reactive with endogenous erythropoietin (Casadevall, Nataf, Viron, Kolta, Kiladjian, Martin-Dupont, Michaud, Papo, Ugo, Teyssandier, Varet, and Mayeux 2002). These cases of

aplastic anemia were associated with exposure to just one formulation of rEPO, suggesting that factors extrinsic to the protein itself were associated with the adverse outcome (Haselbeck 2003). Currently, evidence suggests that this adverse effect was associated with the development of aggregates due to the reformulation of the therapeutic product (Prabhakar and Muhlfelder 1997; Rosenberg 2003).

6.2.1.2 Neutralizing Antibodies

Antibodies that interfere with the function of a therapeutic product may not have as severe consequences for the patient as cross-reactive antibodies, but they can have a dramatic effect on the efficacy of the therapy. Neutralizing antibodies have been observed to common biological therapeutics such as insulin, factor VIII, and beta interferon. In some cases, where different forms of the product are available (e.g., insulin, factor VIII), changing treatment to an alternative has allowed for continued use of the therapy. In addition, as is the case with clotting factors, modifying the dose of the therapy regimen has been shown to induce tolerance. However, this approach is clinically intensive and success is not assured.

6.2.1.3 Nonneutralizing Antibodies

Nonneutralizing antibodies, which do not interfere with the function of the biologic, also known as binding antibodies, are the most common form of anti-therapeutic protein antibody. In some cases, nonneutralizing antibodies to therapeutic compounds have been noted in patients who have never been exposed to the compounds, suggesting that some anti-human protein antibodies are naturally occurring.

6.2.2 Factors Contributing to the Development of Antibodies

6.2.2.1 Extrinsic Factors Contributing to Antibody Formation

Route, dose, and formulation are among a set of extrinsic factors that can influence the immunogenicity of therapeutic proteins (Table 1) (Rosenberg 2003). Contamination of the product with proinflammatory or nonspecific mitogenic compounds such as LPS and the development of product aggregates can provide the critical “second signal” (Signal 2) to the T-cell that is required for induction of T-cell help. In addition, proteins that are denatured during formulation may be more immunogenic than their native counterparts (Braun, Kwee, Labow, and Alsenz 1997), as these products may present new T- and B-cell epitopes that were not present in the parent molecule, leading to the stimulation of an immune response (Josic, Buchacher, Kannicht, Lim, Loster, Pock, Robinson, Schwinn, and Stadler 1999). In contrast, glycosylation and pegylation reduce the immunogenicity of therapeutic proteins. Modification of extrinsic factors such as route, dose, purity, formation of aggregates, and formulation has been used to reduce immunogenicity.

Table 1. Factors associated with induction of immunogenicity.

Extrinsic factors	Intrinsic factors
Route (oral/subcutaneous/IM/IV)	Autologous vs. foreign
Dose (small/large)	T_H epitope content**
Formulation (adjuvant effect)	
Aggregates (crosslinking Ab)	
Contaminants (adjuvant effect)	
Glycosylation/pegylation*	

*Decreases immunogenicity;
**Theoretical – See section 6.2.

6.2.2.2 Intrinsic Factors Contributing to Antibody Formation

Factors that are intrinsic to the therapeutic protein itself can also potentially contribute to immune response. For example, therapeutic proteins that tend to look more like "self," or with few mutations compared to wild type, are considerably less immunogenic than proteins that look very unique to the immune system; for example, mutated or fused proteins that may contain novel antigenic epitopes and/or therapeutic proteins that are derived from pathogens such as streptokinase (Miller, Korn, Stevens, Janik, Gause, Kopp, Holmlund, Curti, Sznol, Smith, Urba, Donegan, Watson, and Longo 1999; Kontsek, Liptakova, and Kontsekova 1999).

Modification of the therapeutic protein sequence from foreign (a murine monoclonal) to a less foreign (a humanized monoclonal) has resulted in reduced immunogenicity. For example, this approach reduced the immunogenicity of the Campath therapeutic (a monoclonal anti-CD52 antibody used to treat cancer patients). Figure 1 illustrates the concept of "humanizing" monoclonal antibodies.

6.2.3 T-Independent and T-Dependent Immune Response

The pathways that lead to B-cell activation, and the resultant production of antibodies, can be divided into T-cell-independent (Ti) and T-cell-dependent (Td) categories (Table 2). Ti activation of B cells occurs when structural features of certain molecules, such as polymeric repeats, induce the "signals" required to stimulate activation of a B-cell subset. Ti activation of B cells results in a weaker immune response than Td activation due to a lack of affinity maturation or of development of B-cell memory. Conversely, Td activation of B cells results in a robust and long-lived antibody response. Most high-affinity IgG responses to therapeutic antibodies are T-dependent.

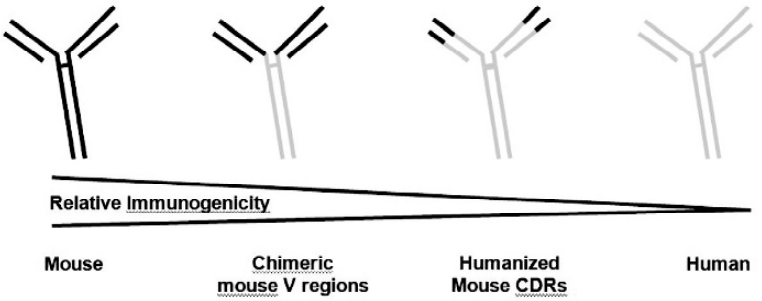


Fig. 1. Minimal effect of "humanization" on immunogenicity of mAbs. Administration of murine antibodies such as Campath 1G is associated with the development of antidrug antibodies in as many as 78% of immunocompetent subjects. Minimalization of the murine component of the antibody by substituting human constant domains (chimeric and humanized antibodies) may result in significant reductions in immunogenicity. However, administration of chimeric antibodies such as Remicade and humanized antibodies such as Campath IH to immunocompetent subjects has been associated with induction of antibodies in as many as 70% and 63% of subjects, respectively. Administration of fully human antibodies can also induce immune responses. For example, Humira has been associated with induction of antidrug antibodies in 18% of subjects in some studies.

6.2.3.1 Ti B-Cell Activation

Activation of naïve B cells to produce specific antibodies (Abs) via the Ti mechanism involves immune responses to polymeric antigens (Ags) with repeating subunit structures (e.g., polysaccharides) that are generally not relevant to the use of soluble, mono-disperse therapeutic proteins; hence, this chapter will not consider them further. The Td mechanism is of greater concern as it generates immunological memory, and also promotes affinity maturation of the Abs and immunoglobulin (Ig) class switching. In fact, the presence of IgG class antibodies implies that a therapeutic protein is a *T-cell-dependent (Td) antigen*, i.e., isotype switching has occurred.

6.2.3.2 Td B-Cell Activation

In order to induce naïve B cells to react to a Td protein antigen and initiate an adaptive immune response, several events must be coordinated, usually within specialized regions of secondary lymphoid organs (e.g., lymph nodes, spleen). An interaction between clonally expressed B-cell transmembrane antigen receptor molecules (IgM

Table 2. T-independent (Ti) and T-dependent (Td) immune response.

T independent	T dependent
No isotype switching	Isotype switching
Low affinity	High affinity
No or low memory	T and B memory

or IgD) and an epitope on the protein antigen is required to initiate activation of a naïve B cell. B-cell epitopes are usually composed of the side chains of nonadjacent amino acids that have been assembled into a three-dimensional conformation on the protein's surface by the folding of the native polypeptide chain. The size and complexity of a protein and its phylogenetic distance from the host contribute to the potential of the protein to be recognized by antibody. Antibodies may also interact with sequential epitopes, segments of consecutive amino acids in the polypeptide chain of a native protein [N- or C- terminals; intrachain loops], or with an unfolded, denatured protein.

In addition to the antigen receptor Ig molecule, the B cell possesses a coreceptor complex that includes CD21. CD21 binds activated components of the innate-immunity complement system. Optimal naïve B-cell activation occurs against target protein antigens that activate the alternative complement pathway, creating a covalent complex between the protein and a cleavage product of complement component C3, called C3d (Carroll 2004). The interaction between the antigen receptor Ig (which is associated with a signal-generating transmembrane heterodimer Igα/β) and the CD21 coreceptor, mediated by the protein: C3d complex, generates "Signal 1" to the B-cell nucleus. Signal 1 must be followed by Signal 2 (T-cell response, see below) for B-cell expansion to occur.

The next step of the process takes place inside the B cell. The target protein: C3d complex is internalized and then it is degraded to peptides within the endocytic compartment and some of the peptides that are generated bind to Class II major histocompatability complex (MHC) molecules that migrate to the B-cell surface. B cells also express the CD40 molecule on their surface and upregulate expression of other adhesion molecules in anticipation of interacting with an activated $CD4^+$ helper T-cell.

Naïve T cells reactive to the same protein antigen that engaged the naïve B cells are likewise stimulated by antigen-processing cells presenting the cognate T-cell epitopes in the context of MHC class II on their surface. Thus, "Signal 2" for activation of a B cell is dependent on (1) proper presentation of the protein antigen by an antigen-presenting cell and (2) prior activation of a T-cell. This process is described in the next paragraph, and then the description of B-cell activation is resumed.

In order for APCs to effectively engage T cells, the APC must become activated, or "mature." APCs that have ingested a protein mature in response to receiving a signal at the cell surface, such as engagement of a Toll receptor or the delivery of cytokines like interleukin 2 (IL-2) to the APC. Contamination of a therapeutic product with proinflammatory or nonspecific mitogenic compounds such as lipopolysaccharide (LPS) could, by binding to a Toll receptor, provide this critical second signal required for the development of T-cell help.

The activation of a T-cell is also a two-step process. The first step involves antigen-presenting cells (APC). These cells internalize, process, and present peptide epitopes derived from the protein antigen, in the context of HLA molecules, to T cells. These T-cell epitopes are derived from therapeutic proteins in the APC or in the B cell as follows: the proteins are taken up into endocytic vesicles, and then digested in a special proteolytic compartment called MIIC. Peptides generated in this process then compete for binding in the binding cleft of HLA Class II molecules,

which are subsequently transported to the surface, where they engage T cells. T cells require the presentation of the MHC:peptide complex and engagement of CD28:CD80/CD86 (Signal 1 and 2) to become activated.

Returning to the case of the B cell, following ingestion of a protein therapeutic by a B cell, the protein is also processed as described above. HLA Class II-peptide complexes are then transported to the surface of the B cell, where they are exposed to interrogation by passing $CD4^+$ T cells (Signal 1), causing the epitope-specific $CD4^+$ to secrete selected cytokines such as gamma-interferon, IL-2, IL-4, and IL-10. Signal 2 is provided by the engagement of the CD40:CD154 proteins on the B cell and T-cell, respectively. These two signals are followed by Signal 3, the release of cytokines from the T cells, initiating a cascade of further immunostimulatory events. The cytokines cause B cells to expand and undergo phenotypic changes resulting in the establishment of memory B cells.

6.2.3.3 Absence of T Help Abrogates Ab Formation

The binding of peptide epitopes (derived from internal processing of proteins by B cells) to HLA Class II molecules and the recognition of the epitope-HLA complex by activated helper T cells are necessary components of any T-cell-dependent antibody response. Without Signal 2 provided by the cytokines released as a result of T-cell interaction, the naïve B-cell response does not mature. Without T-cell help, activation of B cells to antibody-secreting plasma cells can only occur in the presence of aggregates or polymeric proteins (T-cell-independent activation). Attenuation of the helper T-cell response to immunogenic peptides derived from therapeutic proteins has therefore become the focus of considerable research effort.

Since the T-cell epitope plays a critical role in the development of T-cell-dependent antibody responses, it stands to reason that protein sequence modifications that result in the removal of potential T-cell epitopes from autologous (recombinant therapeutic) proteins could indeed reduce the potential for induction of an immune response against the protein. Loss of T-cell help removes Signal 2 for B cells expressing receptor specificity for a therapeutic protein. In theory, loss of Signal 2 due to epitope modification could actually induce B-cell apoptosis and/or tolerance (Kappler, Roehm, and Marrack 1987).

6.2.3.4 Effect of Pegylation and Glycosylation

Pegylation and glycosylation are two commonly used approaches to solving the immunogenicity problem. Polyethylene glycol (PEG) conjugates of protein antigens appear to induce antigen-specific immune tolerance either by masking B-cell epitopes (B-cell Signal 1, structural interference), by stabilizing the protein antigen and increasing the duration of exposure to the antigen (thereby inducing tolerance), or by directly interfering with the processing and presentation of T-cell epitopes to T_H cells (B-cell Signal 1) (So, Ito, Koga, Watanabe, Ueda, and Imoto 1997).

Like pegylation, glycoslyation decreases plasma clearance of therapeutic and increases the half-life of the protein molecules. The primary impact of glycosylation is to interfere with antibody affinity (B-cell Signal 1). No information is available on

the impact of glycosylation on processing and presentation of peptides derived from the therapeutic protein in the context of MHC to T cells.

6.2.3.5 Deimmunization by T-Cell Epitope Modification

While deimmunization by T-cell epitope modification is a relatively recent concept in the field of therapeutics, extensive evidence for the attenuating effect of epitope sequence modification on T-cell response exists in the context of infectious disease, particularly with reference to immune escape from Class I and Class II restricted immune response in viral infections (Vossen, Westerhout, Soderberg-Naucler, and Wiertz 2002). Thus, the idea that therapeutic proteins can be modified so that they may also "escape" immune response, thereby reducing their immunogenicity, has been emerging in recent years, particularly since new tools for mapping T-cell epitopes have become widely available. A number of research teams are actively modifying therapeutic proteins by mutating their amino sequences so as to eliminate the formation of MHC Class II epitopes during intracellular processing. Ideally, these modifications would have little effect on the therapeutic activity of the protein but would ablate MHC Class II presentation of both "self" and "nonself" epitopes.

In the remaining sections of this chapter, we provide a rationale for the deimmunization of therapeutic biologics by T-cell epitope modification. One retrospective example supporting the concept is provided, and examples of possible applications of the deimmunization approach to the design or redesign of therapeutic proteins are discussed.

6.3 A New Concept: Deimmunization by T-Cell Epitope Modification

T helper (T_H) epitope content may explain differences in observed antibody responses to slightly different versions of the same recombinant human protein. For example, "humanizing" chimeric antibodies so that they contain fewer T_H epitopes, as described above, is carried out by swapping "foreign" regions of the antibody for regions that are more like self and has been shown to reduce immunogenicity. The postive effect of humanization may be due at least in part to a reduction in the total Th epitope content of the modified sequence. "Epitope content" may also influence the propensity of any given protein to induce an immune response.

We observed, as illustrated in Fig. 2, that the immunogenicity scores of abundant serum proteins (those contributing almost 90% of serum protein content) were lower than expected. Antigenic proteins from pathogens such as influenza, tetanus and, allergens, in contrast, generally scored much higher. Based on this analysis, it would appear that some common human proteins contain significantly fewer T-cell epitopes, when compared to random proteins and common antigens. This preliminary discovery, which needs to be substantiated using *in vitro* assays, may change existing ideas about the nature of tolerance.

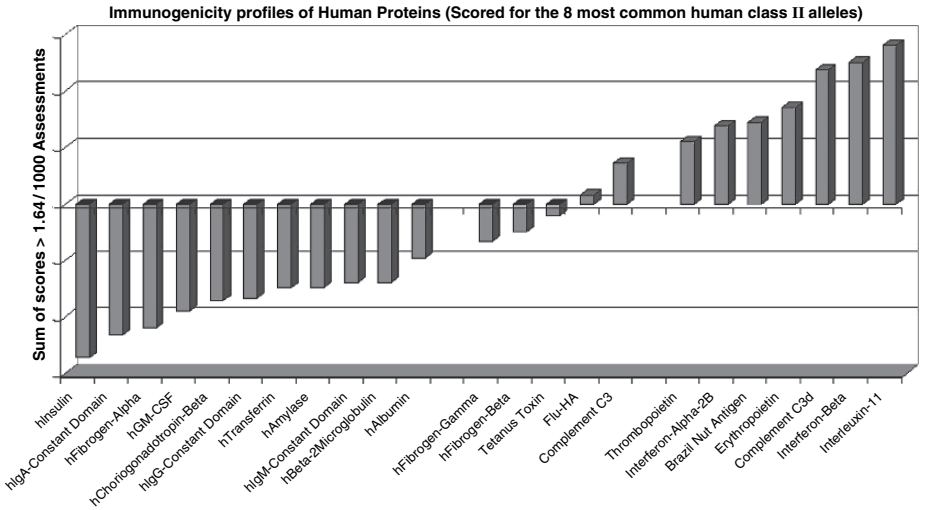


Fig. 2. Inherent immunogenicity of human proteins. To perform this analysis, we used the EpiMatrix tool, which is a matrix-based algorithm for T-cell epitope mapping. This tool is standardized so that comparisons can be made across predictions for different HLA alleles. In this case we used the EpiMatrix prediction matrices for eight class II alleles that are representative of more than 98% of human populations. First, we measured the number of potential T helper epitopes that would occur in random-sequence pseudo proteins composed of amino acids at their naturally occurring frequencies and computed the mean "epitope score" per 1000 assessments of 0.5, with a standard deviation of ±7.9. We then compared proteins by summing the total number of EpiMatrix scores for each frame (9 amino acids) that was above an accepted cutoff for immunogenicity (>1.67), and measured the difference between the scores for "random" proteins and test proteins.

This type of immunogenicity analysis can be used to evaluate and compare protein therapeutics – to other protein therapeutics and to known antigens. Potential T-cell epitopes are not randomly distributed throughout protein sequences but instead tend to be clustered. These regions of unusually high binding potential are called "T-cell epitope clusters," "promiscuous epitopes," or just "clusters" for short (Panina-Bordignon, Tan, Termijtelen, Demotz, Corradin, and Lanzavecchia 1989). Epitope clusters range from 9 to roughly 25 amino acids in length and, considering their affinity to multiple alleles and across multiple frames, can contain anywhere from 4 to 40 binding motifs. ClustiMer, which is an optional feature of EpiMatrix, measures and stores the MHC binding potential for 9 or 10 amino acid sequences to a number of human HLA and then searches for extended regions containing high concentrations of these high-scoring 9-mers.

A protein which scores "average" or even below average on our overall immunogenicity scale may still be immunogenic if it contains one or more clustered regions. For example the well-known antigen tetanus toxin scores just +4 on the overall scale but also contains several local regions with very high potentials (Fig. 3).

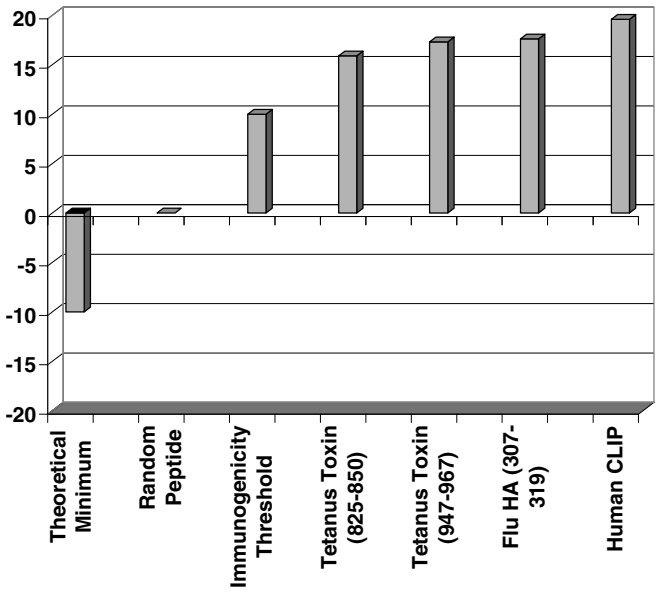


Fig. 3. Predicted potential for immunogenicity of selected T-cell epitope clusters. This analysis was performed as previously described; however, the target proteins in this example are short peptides that are known to be highly immunogenic epitopes. We used the EpiMatrix prediction matrices for eight class II alleles that are representative of more than 98% of human populations. We compared the epitopes to our random peptide standard by summing the total number of EpiMatrix scores for each frame (9 amino acids) that was above an accepted cutoff for immunogenicity (>1.67), and measured the difference between the scores for "random" proteins and test proteins.

Well-described "promiscuous epitopes" tend to score above 10 on our immunogenicity scale. The promiscuous tetanus toxin epitope 825–850 scores +15 and another well-known epitope from tetanus (peptide 947–967) scores +17.3. Influenza HA peptide 307–319 scores +17.6 on the immunogenicity scale (Fig. 3). These peptides are so promiscuous that they are frequently used as positive controls in T-cell activity assays.

After initial exposure to an immunogen, an expanded population of memory T cells is established that is able to respond more rapidly, efficiently, and in greater numbers on subsequent exposure. Since the presence of the epitope bound in the MHC cleft is the trigger for a protective immune response, epitope-driven approaches to reducing the immunogenicity of therapeutic proteins have focused on modifying this response by reducing the ability of peptides from therapeutic proteins to bind to MHC.

6.3.1 Available Epitope Mapping Tools

Alteration of T-cell epitopes is known to result in reduced binding to the MHC and/or altered binding to the T-cell receptor (TCR). This effect has been observed both for Class I and Class II MHC ligands in the context of both tumor cell (Scanlan and Jager 2001) and pathogen (Mullbacher 1992; Hill, Jepson, Plebanski, and Gilbert 1997) escape from immune response. Mapping, confirming, and modifying T-cell epitopes may reduce the immunogenicity of therapeutics.

Based on the hypothesis that the immunogenicity of therapeutic proteins is probably linked to both (1) the presence of T helper epitopes and (2) an event that triggers immune response (the "danger signal"), it follows that removal of T-cell epitopes from a protein that is intended to be used as a therapeutic may reduce the protein's overall potential to stimulate T cells. In order to modify T-cell epitopes, it is important to first identify those epitopes that are responsible for stimulating an immune response and then determine their amino acid sequences. Currently, a number of epitope-mapping tools are available for discovering T-cell epitopes contained within protein sequences. Reviews of these tools have been published (De Groot and Martin 2003a). The following paragraphs provide background on mapping tools developed by the authors of this article.

Prior to the development of tools for T-cell epitope selection, the cost and effort required to identify T-cell epitopes from protein sequences was a significant barrier to the deimmunization of therapeutic proteins. Computational immunology (immunoinformatics) methods dramatically reduce the time and effort involved in screening proteins for potential epitopes, ranging from a reduction of 10- to 20-fold (Kast, Brandt, Sidney, Drijfhout, Kubo, Grey, Melief, and Sette 1994; Schafer, Jesdale, George, Kouttab, and De Groot 1998) to a 95% reduction (De Groot, Bosma, Chinai, Frost, Jesdale, Gonzalez, Martin, and Saint-Aubin 2001a; De Groot, Saint Aubin, Rayner, and Martin 2001b).

EpiMatrix, an algorithm developed by the Brown University TB/HIV Research Lab and licensed to EpiVax, ranks 9- to 10- amino-acid-long segments overlapping by 8 to 9 amino acids derived from any protein sequence by estimated probability of binding to a selected MHC molecule. The EpiMatrix method for ranking prospective epitopes has been published (Schafer et al. 1998; De Groot, Jesdale, Szu, and Schafer 1997). Matrix motifs for 24 HLA Class I alleles are available for use with EpiMatrix. EpiVax used the pocket profile approach first described by Sturniolo and Hammer (Sturniolo, Bono, Ding, Raddrizzani, Tuereci, Sahin, Braxenthaler, Gallazzi, Protti, Sinigaglia, and Hammer 1999) to generate predictive matrices for 74 Class II alleles (De Groot et al. 1997) These new Class II matrices are now included in the EpiMatrix repertoire at EpiVax. Previous studies have demonstrated that EpiMatrix accurately predicts published MHC ligands and T-cell epitopes (De Groot et al. 2001a; De Groot et al.1997; De Groot, Jesdale, Martin, Saint-Aubin, Sbai, Bosma, Lieber-man, Skowron, Mansourati, and Mayer 2003b).

ClustiMer, which is an optional feature of EpiMatrix, can measure and store the MHC binding potential for a 9- or 10- amino-acid sequence to a number of human HLAs. ClustiMer can therefore be used to identify clustered or "promiscuous"

epitopes, which can be presented in the context of more than one HLA; such epitopes are broadly recognized in human populations.

6.3.2 Decreasing Immunogenicity (Case Study)

Evidence that EpiMatrix can evaluate and identify epitopes that can be modified to eliminate immunogenicity is provided by an analysis of SakSTAR and its modified (deimmunized) derivative. SakSTAR is a natural variant of staphylokinase, which is used as a potent thrombolytic agent in the setting of acute myocardial infarction. Administration of SakSTAR results in the generation of IgG antibodies, which limits the efficacy of the therapy. Warmerdam, Plaisance, Vanderlick, Vandervoort, Brepoels, Collen, and De Maeyer (1998) identified the "C3" region of SakSTAR (residues 71–87) as being highly reactive in patients expressing either the DRB1*0301 or DRB1*0701 alleles. EpiMatrix parsed the C3 region into overlapping 9-mer frames and scored each frame for MHC binding potential with respect to eight common alleles. Peptides contained between residues 71 and 87 received the highest scores for the alleles (DRB1*0301 and DRB1*0701) using the EpiMatrix algorithm, consistent with Warmerdam and colleagues' observation that this region was highly immunogenic. The EpiMatrix percentile ranking is shown in Table 3a.

Alanine substitutions to the MHC anchoring residues Y73, K74, R77, E80, and D82, alone or in combination, were subsequently shown to reduce or eliminate T-cell response. These modifications also resulted in dramatically lower EpiMatrix scores for the protein, below the usual cutoff for immunogenicity (i.e., the 95th percentile). An analysis of EpiMatrix scores for these peptides shows that Y73, K74, R77, E80, and D82 all function as anchoring residues in our models of either DRB1*0701, DRB1*0301, or both. Alanine substitutions at these positions had a neutral or negative effect on predicted binding affinity (EpiMatrix percentile ranking shown in Table 3b).

In a separate study by Collen et al. (1996) the SakSTAR variant K74A, E75A, R77A, E80A, D82A was shown to have "induced significantly fewer circulating neutralizing antibodies" in human subjects. Again, this is consistent with the theory that reduction of epitope content also reduces immunogenicity.

In summary, epitope mapping can uncover regions of therapeutic proteins involved in the generation of antibody responses to such proteins. These regions can

Table 3a. SakSTAR epitopes.

Original SakSTAR epitope			
73	YKEFRVVEL	81	DRB1*0701 score: 99th percentile
76	FRVVELDPS	84	DRB1*0301 score: 98th percentile
79	VELDPSAKI	87	DRB1*0301 score: 99th percentile

Table 3b. SakSTAR epitopes (alanine-modified sequence)

Alanine Substituted version			
73	AAEFAVVAL	81	DRB1*0701 score: 82nd percentile
76	FAVVALAPS	84	DRB1*0301 score: 88th percentile
79	VALAPSAKI	87	DRB1*0301 score: 90th percentile

then be targeted for modification. Modifications of these regions may potentially reduce the immunogenicity of therapeutic proteins, as described for the humanized monoclonal antibody and the staphylokinase protein examples. The next section describes the process of deimmunization by epitope mapping and modification of protein sequences.

6.4 A Step-by-Step Approach to Deimmunization

The approach to deimmunization of functional therapeutics described in the next paragraphs is a multistep process involving (1) analysis of the therapeutic protein for the presence of MHC binding motifs; (2) synthesis and testing of the target peptides for MHC Class II binding and immunogenicity *in vitro;* (3) development of "de-immunized" versions of the regions where the MHC binding motifs have been modified; (4) synthesis and testing of the deimmunized counterparts *in vitro*; and (5) testing of the recombinant, deimmunized protein *in vivo* for immunogenicity. Evaluation of the effect of protein modification on protein function is characterized by structural modeling (*in silico)*, at step (3) following the resynthesis of the protein.

6.4.1 Initial Screen for Class II Epitopes and Epitope Clusters

The first step in the deimmunization analysis involves screening the target protein for Class II epitopes. We have used EpiMatrix matrices corresponding to eight common HLA alleles (DRB1*0101, DRB1*0301, DRB1*0401, DRB1*0701, DRB1*0801, DRB1*1101, DRB1*1301, DRB1*1501), which are well represented in most human populations. EpiMatrix identifies clusters of putative epitopes restricted by at least one of these eight common alleles in any given protein. For example, EpiMatrix (Class II, eight alleles) epitope mapping of human beta-interferon demonstrates seven potential epitope cluster (Fig. 4). Collectively these clusters contain almost all of the predicted epitopes. This type of clustering of MHC binding motifs is typical of promiscuous epitopes (Meister, Roberts, Berzofsky, and De Groot 1995).

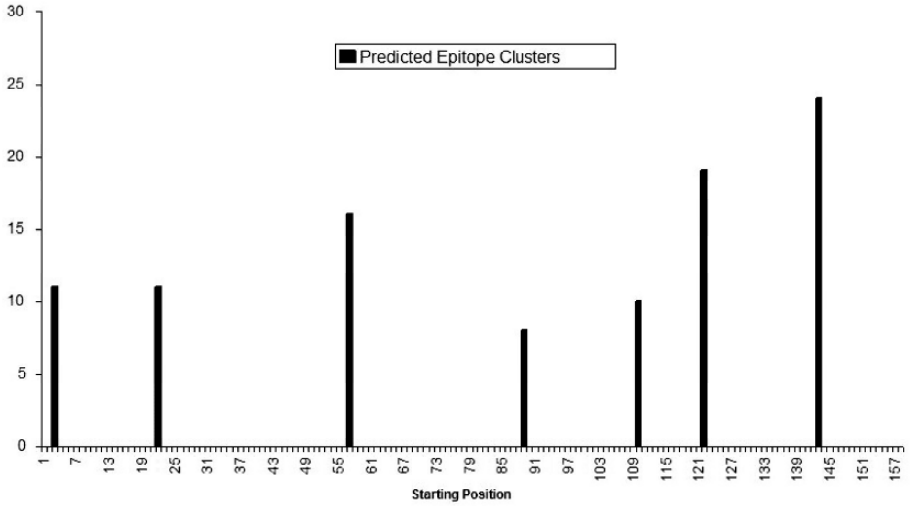


Fig. 4. EpiMatrix analysis of human beta-interferon predicted epitope clusters. For this analysis, the complete sequence of human beta-interferon was downloaded from Genbank (accession 1AU1_A). The sequence was parsed into 158 overlapping 9-mer frames. Each frame was then evaluated for its potential to bind to 8 common Class II HLAs (DRB1*0101, DRB1*0301, DRB1*0401, DRB1*0701, DRB1*0801, DRB1*1101, DRB1*1301, and DRB1*1501). The entire dataset was then scanned for frames that contain more predicted ligands than would be expected by chance alone. In this case any 9-mer frame that is predicted to react to 3 or more alleles defines a cluster. The binomial probability of observing 3 positive signals (a positive signal is defined as an EpiMatrix Z-score in excess of 1.64) in 8 trials is 0.0058. Contiguous and nearly clusters were then combined to yield the 7 clusters displayed in the graph.

6.4.2 Modifying Epitopes

Regions of epitope clustering are often highly immunogenic (Pevear, Luo, and Lipton 1988), thus the first step in deimmunizing a protein is to address the clustered regions defined by EpiMatrix. For example, Tables 4a and 4b show the last 20 overlapping frames of an analysis for human CLIP (an HLA binding portion of Invariant chain) and a modified version of CLIP. Putative epitopes (HLA Class II binding regions) located in the last cluster of epitopes are shown. Each overlapping frame in the entire amino acid sequence is scored against each of the eight common HLA alleles. The total number of epitopes in the natural and modified versions and the highest score per allele are shown at the bottom of each table. For example, by substituting alanine for critical residues the total number of DRB1*0101 epitopes is reduced from three to none and no increase in the number of putative HLA binding regions develops. Confirmation of *in silico* immunogenicity prediction is subsequently performed *in vitro*.

Table 4a. Z score indicates the potential of a 9-mer frame to bind to a given HLA allele; the strength of the score is indicated by the shading: top 10% ;top 5% ; top 1% . All scores in the top 5% are considered "hits." *Frame 107–115 (sequence highlighted) contains four or more alleles scoring above 1.64 and as such is referred to as an EpiBar. EpiBars have an increased likelihood of binding to multiple HLAs.

Frame Start	AA Sequence	Frame Stop	DRB1*0101 Z-Score	DRB1*0301 Z-Score	DRB1*0401 Z-Score	DRB1*0701 Z-Score	DRB1*0801 Z-Score	DRB1*1101 Z-Score	DRB1*1301 Z-Score	DRB1*1501 Z-Score	Hits
1	PVSKMRMAT	9					1.91	1.55			1
2	VSKMRMATP	10									0
3	SKMRMATPL	11	1.46			1.87				1.90	2
4	KMRMATPLL	12	2.16		1.42			1.45			1
5	MRMATPLLM	13	2.81	2.80	2.54	2.46	1.65	1.62	2.76	2.95	7*
6	RMATPLLMQ	14									0
7	MATPLLMQA	15								1.44	0

Summarized Results	DRB1*0101	DRB1*0301	DRB1*0401	DRB1*0701	DRB1*0801	DRB1*1101	DRB1*1301	DRB1*1501	Total
Maximum Single Z score	2.81	2.80	2.54	2.46	1.91	1.82	2.76	2.95	--

Table 4b. Z score indicates the potential of a 9-mer frame to bind to a given HLA allele; the strength of the score is indicated by the shading: top 10% ;top 5% ; top 1% . All scores in the top 5% are considered "hits." Scores below the top 10% are omitted for simplicity. *Frame 5 (sequence highlighted) contains high scores for 2 of the 8 alleles tested as compared to 7 of the 8 alleles in the parent sequence. The modified CLIP contains no EpiBars. EpiBars (frames with high scores across at least 4 alleles) have an increased likelihood of binding to multiple HLAs.

Frame Start	AA Sequence	Frame Stop	DRB1*0101 Z-Score	DRB1*0301 Z-Score	DRB1*0401 Z-Score	DRB1*0701 Z-Score	DRB1*0801 Z-Score	DRB1*1101 Z-Score	DRB1*1301 Z-Score	DRB1*1501 Z-Score	Hit
2	VSKAKAATP	10	1.42								0
3	SKAKAATPL	11	1.29								0
4	KAKAATPLL	12	1.60								0
5	AKAATPLLM	13	1.78	1.52	1.49	1.43			1.52	1.78	2*
6	KAATPLLMQ	14									0
7	AATPLLMQA	15									0

Summarized Results	DRB1*0101	DRB1*0301	DRB1*0401	DRB1*0701	DRB1*0801	DRB1*1101	DRB1*1301	DRB1*1501	Total
Maximum Single Z score	1.78	1.52	1.49	1.43	1.76		1.52	1.78	--

6.4.3 Confirming the Potential of Epitope Clusters *in Vitro*

MHC binding assays can be used as the first screen for immunogenicity *in vitro* (Elvin, Potter, Elliott, Cerundolo, and Townsend 1993). These assays confirm the binding affinity of predicted epitopes for various MHC alleles and are based on competition between the test peptide and a known binder to recombinant soluble MHC molecules (Tompkins, Rota, Moore, and Jensen 1993). Since MHC binding is the first event involved in the stimulation of a T-cell response and is one of the more critical criteria determining potential immunogenicity, a peptide's ability to bind to a given MHC can indicate its potential to generate a T-cell response. These binding assays can also be used to determine whether suggested modifications would disrupt the binding and presentation of predicted epitope to T cells. However, processing of the protein antigen is not measured by the binding assay. Thus,

the assays cannot predict which peptides are actually expressed on the surface following natural processing of a protein antigen *in vivo*.

6.4.4 Confirming the Potential of Epitope Clusters Using T Cells from Donors

The only way to accurately evaluate whether a peptide epitope is processed and presented, or not (following modification of key amino acid residues) is to measure T-cell responses to the peptides, using peripheral blood T cells from exposed subjects or splenocytes from exposed mice (see next section). Alternatively, blood from unexposed donors can be primed *in vitro*. T-cell lines can be developed by incubating blood samples drawn from unexposed volunteers with the candidate protein and selected stimulatory molecules such as GMCSF, CD80, or CD86.

While there are many assays for evaluating T-cell response, one of the most accurate means of measuring the response of the individual T-cell is to perform an ELISpot assay. In most cases, the number of T cells responding to a given protein is extremely few; therefore, an *in vitro* T-cell restimulation assay is used to increase the assay's sensitivity by expanding the number of T cells present that will respond to a given peptide. Restimulation assays are usually performed using blood from exposed subjects (i.e., persons who have been treated with the therapeutic protein and/or may have participated in clinical trials.)

No matter which cells are used, the first task is to establish an overall level of response by performing an ELISpot T-cell assay using the candidate protein as the antigen. Peptides representing the predicted epitopes can also be assessed at this time. Peptide epitopes that are observed to be associated with a strong T-cell response can then be modified at critical amino acid residues. Next, in an iterative approach, these modified peptides can be reevaluated for both binding and immunogenicity.

The goal of these *in vitro* experiments is to demonstrate that the target protein and peptides representing the clustered epitope regions do engender immune responses from exposed donors i*n vitro*; that the derivative clustered epitope region peptides account for the majority of the immunogenicity engendered by the target protein *in vitro*; and that the deimmunized proteins do not engender immune responses *in vitro* and *in vivo*.

6.4.5 Confirming Epitope Clusters *in Vivo*

Another means of evaluating the impact of epitope modifications on *de novo* T-cell response is to measure the immunogenicity of the original epitopes and the modified epitopes in HLA-transgenic mice, either in the context of the whole protein (the modifications are engineered in) or as peptides.

A number of transgenic mouse strains that express the most common HLA A, HLA B, and HLA DR molecules have been developed. These mice process and present epitopes in the context of human HLA. In order to compare the immunogenicity of wild-type and modified epitopes, immunizations are carried out with preparations of the peptide epitopes in adjuvant. The standard regimen is to immunize

the mice subcutaneously or intradermally with 100 micrograms of peptides in adjuvant three times, at 2-week intervals. The animals are sacrificed at 42 days, and the splenocytes are obtained for use in T-cell (ELISpot) assays.

One can expect to find that these assays confirm that the original epitopes (wild type) are more immunogenic than the modified epitopes. Measurable T-cell responses seen in mice given the wild-type epitopes, and not seen in modified peptide or placebo-vaccinated mice, validate the effect of the modifications.

6.4.6 Evaluating the Effect on Protein Structure and Function

Some amino acid substitutions may have a profound effect on the functionality of the protein. Several methods of assessing the effect of deimmunizing changes on the function target protein exist. A comparison of the sequence containing the proposed changes with the sequences of other similar proteins can indicate areas of variability that may be tolerant of change. Molecular modeling can also help elucidate the impact of the proposed changes on the structure of the protein. These two approaches are explained in more detail in the next paragraphs.

6.4.6.1 Evaluate by Comparison with Other Similar Proteins

Within a particular protein, amino acid sequences that are conserved in the course of evolution can be presumed to be required for the proper function of the protein. For example, an analysis of 92 variants of cytochrome c derived from eukaryotic species shows that 21 of 104 residues are exactly conserved and another 14 are mainly restricted to amino acids that share common properties such as I, L, V or F, and Y. Thus, modifications of the cytochrome c protein aiming to de-immunize the protein should not be performed at any of these 35 residues, as they could eliminate monomers that are critical to the function and/or structure of the protein. This approach was recently described in a paper by Kersh, Miley, Nelson, Grakoui, Horvath, Donermeyer, Kappler, Allen, and Fremont (2001).

6.4.6.2 Evaluate the Effect on Structure Using Modeling

Structural analyses comparing wild-type and modified protein sequences can also be performed. In general these methods start by modeling the observed structure of the wild-type protein. Interactions between the individual atoms that make up the structure are measured and an overall energy value is calculated. Modified proteins are then overlaid on top of the wild-type structure and reiterative modeling of the protein is performed until a minimum energy configuration is achieved. The difference in overall energy between the wild-type protein and its homologue, as measured by calculating the root mean square (RMS) deviation of the positions of the alpha-carbons in the protein chains, can be used as a predictor of the structural deviation between the two proteins.

Several *in silico* methods are available for evaluating the effect of amino acid substitution on protein structure. A number of computational chemistry software programs such as InsightII (Accelrys, Inc.), Cerius2 (Accelrys, Inc.), Sybyl (Tripos, Inc.),

MacroModel (Schrodinger, Inc.), and MOE (Chemical Computing Group, Inc.) are commercially available. Additional modeling programs are freely available on the Internet; though not as complete as the commercially available products, they perform many of the same functions. Such programs include pymol (http://www.pymol.org) for visualization and in silico mutagenesis of proteins, namd (http://www.ks.uiuc.edu) for molecular dynamics simulations, vmd (http://www. ks. uiuc.edu) for more advanced visualization, and grace (http://plasma-gate.weizmann.ac.il/Grace/) for data analysis.

Using any of these programs, differences in structure between wild-type and amino-acid-substituted protein sequences can be measured by calculating the RMS deviation of the positions of the alpha-carbons in the protein chains after identifying the lowest energy state of each sequence. A relatively low RMS deviation suggests that the mutations do not cause broad structural changes in the protein, while a relatively large RMS difference suggests that the substitutions do affect tertiary structure. Close examination of the results permits the exact identification of the amino acids responsible for the greatest energy difference; thus, resubstitution and reiterative analysis of the effect of substitutions on protein structure may be performed.

6.5. When Can Deimmunization Be Useful?

Deimmunization may be extremely useful at certain stages of protein therapeutic development. Three scenarios are illustrated in the next few paragraphs.

6.5.1 Prioritizing in the Preclinical Stage of Development

Given several similar candidates that may have all demonstrated a reasonable level of efficacy in preclinical evaluations, drug developers need a means for selecting the one or two that are most likely to succeed. One means of reducing the list of candidates to evaluate is to score each therapeutic protein on a "potential immunogenicity scale" such as the one described in Section 2.3.3. This scale allows for comparisons between proteins that are known to be nonimmunogenic, such as the constant regions of human antibodies, and viral or bacterial proteins known to be highly immunogenic such as tetanus toxin, ESAT6 derived from TB, or haemagglutinin derived from influenza. Candidate therapeutic proteins displaying low or limited potential for immunogenicity may be prioritized for clinical trials using this scale. Candidate therapeutics displaying high potential immunogenicity might be set aside or returned to the developmental pipeline for reengineering.

6.5.2 Modifying a Lead Candidate Following Immunogenicity Testing

In some cases, a good candidate has been identified but is predicted to contain an unacceptable amount of immunogenic potential. In this case, regions of the therapeutic protein that account for at least 50% of the total potential for immunogenicity contained within the candidate protein can be identified. These regions of high potential can then be evaluated in order to identify individual amino acids that are

primarily responsible for the peptide's binding affinity with respect to multiple alleles. Substitutions at these key amino acids may then be considered and re-evaluated in a reiterative manner.

When modifying a protein sequence, it is important to consider the effect of the substitutions on the function of the protein. This process may involve the analysis of multiple substitutions and comparative *in silico* and *in vitro* analyses of the substitutions' effects on protein function (see Structural Modeling, above).

6.5.3 Reducing Immunogenicity Following Clinical Trials

The final scenario involves the evaluation and deimmunization of protein therapeutics that have already been tested in human subjects. Blood samples from exposed subjects are obtained, restimulation is performed in order to activate and expand the relevant T memory cells, and T-cell response in ELISpot assays is measured. T-cell response to the native protein and to immunogenic peptides can then be contrasted with response to peptides that have been modified to reduce immunogenicity. The protein can then be modified to reduce immunogenicity while carefully preserving function.

6.6 Conclusions

The approach to deimmunization of functional therapeutics described in this article is a multistep process involving (1) analysis of the therapeutic protein for presence of MHC binding motifs; (2) synthesis and testing of the target peptides in the MHC Class II binding and immunogenicity *in vitro*; (3) development of "deimmunized" versions of these regions for which the MHC binding motifs have been modified; (4) synthesis and testing of the deimmunized counterparts *in vitro*; and (5) testing of the recombinant, deimmunized protein *in vivo* for immunogenicity and function following natural translation.

In summary, epitope-mapping tools can be used to predict Class II restricted T-cell epitopes contained in therapeutic protein sequences. These tools can also be used to accurately discriminate between immunogenic and nonimmunogenic peptides. The demand for preclinical methods for evaluating the immunogenic potential of therapeutic proteins is expected to increase as the number of therapeutic proteins and monoclonal antibodies entering the preclinical pipeline increases. While more extensive validation is needed, this chapter provides a road map for deimmunization that may be worth pursuing as it may accelerate the development of improved therapeutic proteins.

References

Braun, A., Kwee, L., Labow, M.A., and Alsenz, (1997) Protein aggregates seem to play a key role among the parameters influencing the antigenicity of interferon alpha (IFN-alpha) in normal and transgenic mice. J. Pharm. Res. 14:1472-1478.

Carroll, M.C. (2004) The complement system in B cell regulation. Mol. Immunol. 41:141-146.

Casadevall, N., Nataf, J., Viron, B., Kolta, A., Kiladjian, J.J., Martin-Dupont, P., Michaud, P., Papo, T., Ugo, V., Teyssandier, I., Varet, B., and Mayeux, P. (2002) Pure red-cell aplasia and antierythropoietin antibodies in patients treated with recombinant erythropoietin. N. Engl. J. Med. 346:469-475.

Chatenoud, L. (1993) Immunologic monitoring during OKT3 therapy. Clin. Transplant. 7(4 Pt 2):422-430.

Collen, D. (1996) Fibrin-selective thrombolytic therapy for acute myocardial infarction. Circulation 93:857-865.

De Groot, A.S., Jesdale, B.M., Szu, E., and Schafer, J.R. (1997) An interactive Web site providing MHC ligand predictions: Application to HIV research. AIDS Res. Hum. Retroviruses 13:539-541.

De Groot, A.S., Bosma, A., Chinai, N., Frost, J., Jesdale, B.M., Gonzalez, M.A., Martin, W., and Saint-Aubin, C. (2001a) From genome to vaccine: In silico predictions, ex vivo verification. Vaccine 19:4385-4395.

De Groot, A.S., Saint Aubin, C.S., Rayner, J., and Martin, W. (2001b) Rapid determination of HLA B*07- ligands from the West Nile Virus genome. Emerg. Infect. Dis. 7:706-713.

De Groot, A.S., and Martin, W. (2003a) From immunome to vaccine: Epitope mapping and vaccine design tools. Novartis Found. Symp. 254:57-72.

De Groot, A.S., Jesdale, B., Martin, W., Saint-Aubin, C., Sbai, H., Bosma, A., Lieberman, J., Skowron, G., Mansourati, F., and Mayer, K.H. (2003b) Mapping cross-clade HIV-1 vaccine epitopes using a bioinformatics approach. Vaccine 21:4486-4504.

Diamond, B. (2003) Speculations on the immunogenicity of self-proteins. Dev. Biol. (Basel)112:29-34.

Elvin, J., Potter, C., Elliott, T., Cerundolo, V., and Townsend, A. (1993) A method to quantify binding of unlabeled peptides to class I MHC molecules and detect their allele specificity. J. Immunol. Methods 158:161-171.

Finkelman, F.D., Lees, A., and Morris, S.C. (1992) Antigen presentation by B lymphocytes to CD4+ T lymphocytes in vivo: Importance for B lymphocyte and T lymphocyte activation. Semin. Immunol. 4:247-255.

Forsthuber, T.G., Shive, C.L., Wienhold, W., de Graaf, K., Spack, E.G., Sublett, R., Melms, A., Kort, J., Racke, M.K., and Weissert, R. (2001) T-cell epitopes of human myelin oligodendrocyte glycoprotein identified in HLA-DR4 (DRB1*0401) transgenic mice are encephalitogenic and are presented by human B cells. J. Immunol. 167:7119-7125.

Fowell, D., and Mason, D. (1993) Evidence that the T-cell repertoire of normal rats contains cells with the potential to cause diabetes. Characterization of the CD4+ T-cell subset that inhibits this autoimmune potential. J. Exp. Med. 177:627-636.

Haselbeck, A. (2003) Epoetins: Differences and their relevance to immunogenicity. Curr. Med. Res. Opin. 19:430-432.

Hill, A.V., Jepson, A., Plebanski, M., and Gilbert, S.C. (1997) Genetic analysis of host-parasite coevolution in human malaria. Philos. Trans. R. Soc. London Ser. B 352: 1317-1325.

Jensen, P.B., Birkeland, S.A., Rohrp, N., Elbirk, A., and Jorgensen, K.A. (1996) Development of anti-OKT3 antibodies after OKT3 treatment. Scand. J. Urol. Nephrol. 30:227-230.

Josic, D., Buchacher, A., Kannicht, C., Lim, Y.P., Loster, K., Pock, K., Robinson, S., Schwinn, H., and Stadler, M. (1999) Degradation products of factor VIII which can lead to increased immunogenicity. Vox Sang. 77(Suppl. 1):90.

Kappler, J.W., Roehm, N., and Marrack, P. (1987) T-cell tolerance by clonal elimination in the thymus. Cell 49:273-280.

Kast, W.M., Brandt, R.M., Sidney, J., Drijfhout, J.W., Kubo, R.T., Grey, H.M., Melief, C.J., and Sette, A. (1994) Role of HLA-A motifs in identification of potential CTL epitopes in human papillomavirus type 16 E6 and E7 proteins. J. Immunol. 152:3904-3912.

Keech, C.L., Farris, A.D., Beroukas, D., Gordon, T.P., and McCluskey, J. (2001) Cognate T-cell help is sufficient to trigger anti-nuclear autoantibodies in naive mice. J. Immunol. 166:5826-5834.

Kersh, G.J., Miley, M.J., Nelson, C.A., Grakoui, A., Horvath, S., Donermeyer, D.L., Kappler, J., Allen, P.M., and Fremont, D.H. (2001) Structural and functional consequences of altering a peptide MHC anchor residue. J. Immunol.166:3345-3354.

Kontsek, P., Liptakova, H., and Kontsekova, E. (1999) Immunogenicity of interferon-alpha 2 in therapy: Structural and physiological aspects. Acta Virol. 43:63-70.

Koren, E., Zuckerman, L.A., and Mire-Suis, A.R. (2002) Immune responses to therapeutic proteins in humans ─ Clinical significance, assessment and prediction. Curr. Pharm. Biotechnol. 3:349-360.

Meister, G.E., Roberts, C.G.P., Berzofsky, J.A., and De Groot, A.S. (1995) Two novel T cell epitope prediction algorithms based on MHC-binding motifs; comparison of predicted and published epitopes from Mycobacterium tuberculosis and HIV protein sequences. Vaccine 13:581-591.

Miller, L.L., Korn, E.L., Stevens, D.S., Janik, J.E., Gause, B.L., Kopp, W.C., Holmlund, J.T., Curti, B.D., Sznol, M., Smith, J.W., Urba, W.J., Donegan, S.E., Watson, T.M., and Longo, D.L. (1999) Abrogation of the hematological and biological activities of the interleukin-3/granulocyte-macrophage colony-stimulating factor fusion protein PIXY321 by neutralizing anti-PIXY321 antibodies in cancer patients receiving high-dose carboplatin. Blood 93:3250-3258.

Mullbacher, A. (1992) Viral escape from immune recognition: Multiple strategies of adenoviruses. Immunol. Cell Biol. 70(Part 1):59-63.

Nilsson, J.B., Nilsson, T.K., Jansson, J.H., Boman, K., Soderberg, S., and Naslund, U. (2002) The effect of streptokinase neutralizing antibodies on fibrinolytic activity and reperfusion following streptokinase treatment in acute myocardial infarction. J. Intern. Med. 252: 405-411.

Nissenson, A.R. (2001) Novel erythropoiesis stimulating protein for managing the anemia of chronic kidney disease. Am. J. Kidney Dis. 38:1390-1397.

Panina-Bordignon, P., Tan, A., Termijtelen, A., Demotz, S., Corradin, G., and Lanzavecchia, A. (1989) Universally immunogenic T cell epitopes: Promiscuous binding to human MHC class II and promiscuous recognition by T cells. Eur. J. Immunol. 19:2237-2242.

Pevear, D.C., Luo, M., and Lipton, H.L. (1988) Three-dimensional model of the capsid proteins of two biologically different Theiler virus strains: Clustering of amino acid difference identifies possible locations of immunogenic sites on the virion. Proc. Natl. Acad. Sci. USA 85:4496-4500.

Prabhakar, S.S., and Muhlfelder, T. (1997) Antibodies to recombinant human erythropoietin causing pure red cell aplasia. Clin. Nephrol. 47:331-335.

Reijonen, H., Novak, E.J., Kochik, S., Heninger, A., Liu, A.W., Kwok, W.W., and Nepom, G.T. (2002) Detection of GAD65-specific T-cells by major histocompatibility complex class II tetramers in type 1 diabetic patients and at-risk subjects. Diabetes 51:1375-1382.

Rosenberg, A.S. (2003) Immunogenicity of biological therapeutics: A hierarchy of concerns. Dev. Biol. (Basel) 112:15-21.

Ryff, J.C., and Schellekens, H. (2002) Immunogenicity of rDNA-derived pharmaceuticals. Trends Pharmacol. Sci. 23:254-256.

Scanlan, M.J., and Jager, D. (2001) Challenges to the development of antigen-specific breast cancer vaccines. Breast Cancer Res. 3:95-98.

Schafer, J.A., Jesdale, B.M., George, J.A., Kouttab, N.M., and De Groot, A.S. (1998) Prediction of well conserved HIV-1 ligands using a matrix-based algorithm, EpiMatrix. Vaccine 16:1880-1884.

So, T., Ito, H.-O., Koga, T., Watanabe, S., Ueda, T., and Imoto, T. (1997) Depression of T-cell epitope generation by stabilizing hen lysozyme. J. Biol. Chem. 272:32136-32140.

Stein, K.E. (2002) Immunogenicity: Concepts/issues/concerns. Dev. Biol. (Basel) 109:15-23.

Sturniolo, T., Bono, E., Ding, J., Raddrizzani, L., Tuereci, O., Sahin, U., Braxenthaler, M., Gallazzi, F., Protti, M.P., Sinigaglia, F., and Hammer, J. (1999) Generation of tissue-specific and promiscuous HLA ligand databases using DNA microarrays and virtual HLA class II matrices. Nature Biotech. 17:555-561.

Tompkins, S.M., Rota, P.A., Moore, J.C., and Jensen, P.E. (1993) A europium fluoroimmunoassay for measuring binding of antigen to class II MHC glycoproteins. J. Immunol. Methods 163:209-216.

Vossen, M.T., Westerhout, E.M., Soderberg-Naucler, C., and Wiertz, E.J. (2002) Viral immune evasion: A masterpiece of evolution. Immunogenetics 54:527-542.

Vuong, K.D., and Jankovic, J. (2005) Long-term botulinum toxin efficacy, safety, and immunogenicity. Mov. Disord. 20:592-597.

Wadhwa, M., Mellstedt, H., Small, E., and Thorpe, R. (2003) Immunogenicity of GM-CSF products in cancer patients following immunostimulatory therapy with GM-CSF. Dev. Biol. (Basel) 12:61-67.

Warmerdam, P.A., Plaisance, S., Vanderlick, K., Vandervoort, P., Brepoels, K., Collen, D., and De Maeyer, M. (2002) Elimination of a human T-cell region in staphylokinase by T-cell screening and computer modeling. Thromb. Haemost. 87:666-673.

Chapter 7

Plasticity of Dendritic Cell Transcriptional Responses to Antigen: Functional States of Dendritic Cells

Paul Kellam and Antonia Kwan

Department of Infection, University College London, 46, Cleveland Street, London W1T 4JF, UK, p.kellam@ucl.ac.uk

Abstract. The vertebrate immune system protects the host from harmful encounters with pathogenic microorganisms and other dangerous components of the environment. To do this the immune system must gather information about the 'nonself' pathogens and by processing this information, initiate an appropriate immunological response. The immune system represents an example of emergent behavior from a complex, multifactorial, adaptive system. To understand emergent behavior we need to know the systems components and the rules that govern the system. To identify the components, immunology research has catalogued and characterized probably all major cell types involved in the innate and adaptive immune response. In the postgenomic world, we are now able to further characterize global changes in cellular gene expression and thereby identify and infer changes in the functional state of the cells. Together this should allow modeling of immune system function. Dendritic cells orchestrate the host immune response. By identifying a pathogen and processing this information through a coordinated differentiation program, phenotypic changes effected in dentritic cells allow appropriate information to be conveyed to the adaptive arm of the immune system, thereby shaping downstream immunological responses. Transcriptional profiling of human and mouse dendritic cell responses to different antigens have demonstrated this functional plasticity. Understanding the regulation of these dendritic cell differentiation states will contribute to computational models of the immune system, and our understanding of the parameters that affect the immune system response to infection.

7.1 Introduction

The vertebrate immune system is remarkably versatile, functioning to protect animals from pathogenic microorganisms and environmental toxins. The immune system is also able to recognize and attack host cells that have become tumors. To help understand many of the diverse functions required to fight pathogens the immune system has been partitioned conceptually into a number of functionally distinct components: recognition and response, innate and adaptive or effector and memory. While theoretically simplifying the immune system, these divisions also

highlight the complexity of relationships that occur. It is clear that the nonlinear integration of cell surface and cytokine signals from pathogen recognition, through innate to adaptive immune responses bestows incredible complexity and flexibility on and leads to the renewable recall capacity of the immune system.

Complex, dynamic nonlinear systems commonly produce unusual, non-intuitive properties, so-called emergent properties. Emergence in this sense applies where the activity of the parts does not simply sum to give the activity of the whole. Therefore, ultimately, mathematical models will be required to help understand immune function (Callard, George, and Stark 1999). Formulating such models requires knowledge of the systems components and the rules that govern the system. Some of the component details are known. For example, the vertebrate immune system has evolved to comprise many specialist cell types, which themselves have complex ontogenesis. These cell types reflect functional partitioning, with cells of the myeloid lineage (monocyte/macrophage, neutrophils, eosinophils, and basophils) functioning primarily in the innate, antigen-nonspecific immune response and cells of the lymphoid lineage (T cells and B cells) functioning primarily in the adaptive antigen-specific response. Other components are less well understood. For example, gene expression programs and functional gene networks that manifest as different immune cell phenotypes have not been defined. In this review we will discuss aspects of how the host immune system functions in the recognition of pathogens and begin to illustrate how plasticity in gene expression programs creates different functional states in one of the pathogen-sensing components, the dendritic cell.

7.2 Dendritic Cells

The recognition and initial innate response to diverse pathogens is linked to an effective adaptive immune response, with dendritic cells (DCs) providing the bridge (Hoebe, Janssen, and Beutler 2004). Bone marrow-derived DCs in their “immature” form are distributed at anatomical sites most likely to be breached by microbes. Here they continuously sample their environment. When microbial antigens are encountered, along with the presence of “danger signals” from locally infected cells, DCs undergo a complex “maturation” process resulting in their migration out of peripheral tissues and transit to secondary lymphoid tissues. Where mature, DCs present processed microbial peptides on Major Histocompatibility complex (MHC) molecules to T cells. Importantly, DCs express costimulatory molecules allowing them to interact and prime antigen naïve T cells to proliferate. DCs are therefore able to recognize and interpret the presence of different antigens, process the information facilitating their own maturation, and present the processed information to T cells to shape the adaptive immune response.

7.2.1 Recognizing Pathogens

7.2.1.1 Dendritic Cell Subsets

At least two, probably nonexclusive mechanisms can exist for differential recognition of pathogens by DCs. First, distinct cell subsets specialize to recognize particular pathogens. Second, cells can express different receptor molecule repertoires that recognize specific pathogens. Evidence exists for both these

scenarios in DCs. Because DCs interact with all major lymphocyte types (B cells, T cells and natural Killer cells) and because DCs have mutually exclusive functions (antigen uptake compared to antigen presentation), it is perhaps not surprising that different DC subsets/subtypes can be found. What is not clear, however, is whether such subsets are the product of different ontogenesis with particular pathogen recognition and effector functions (specialized lineage model) or if they represent different functional states of a single lineage that itself can detect and respond differently to distinct pathogens (functional plasticity model). Recent evidence suggests that the simple ontogeny distinction of myeloid (derived from a common myeloid progenitor cell) and lymphoid (derived from a common lymphoid progenitor cell) DCs is not true.

DC subsets can be identified in the mouse and humans on the basis of the expression of different cell surface markers (proteins and glycoproteins) (Shortman and Liu 2002) in support of the specialized lineage model. Myeloid DCs (mDCs) express the myeloid cell surface marker CD11c whereas plasmacytoid DCs (pDCs) (proposed to be of the lymphoid lineage and so named because of their ultrastructural resemblance to antibody-secreting plasma cells) express the CD45 isoform (B220) normally expressed on B cells. Exposure of murine bone marrow pDCs to the mouse virus lymphocytic choriomeningitis virus (LCMV), however, induced these pDCs to differentiate into authentic mDCs by undergoing profound phenotypic and functional changes, highlighting the functional plasticity of bone marrow pDCs (Zuniga, McGavern, Pruneda-Paz, Teng, and Oldstone 2004).

The situation becomes more complicated with the evidence that pDCs can develop efficiently from myeloid and lymphoid committed progenitors suggesting the existence of a common DC progenitor (Shigematsu, Reizis, Iwasaki, Mizuno, Hu, Traver, Leder, Sakaguchi, and Akashi 2004). In addition, hematopoietic cell lineage commitment may not be an irreversible transition from progenitor to terminally differentiated cell, as experiments show that enforced expression of the transcription factor C/EBP α and/or β in mature B cells efficiently reprograms them into macrophages (Xie, Ye, Feng, and Graf 2004). Complex cell ontogeny is certainly a feature of the immune system. Lineage plasticity may also be a common but hidden feature of hematopoietic cell types that are influential in the immune response, with the extracellular signaling context maintaining an apparently fixed phenotypic steady state, masking the capacity of changing to an alternative lineage phenotypic state if different cellular input signals were supplied (Xie et al. 2004).

7.2.1.2 Toll-like Receptors

Regardless of whether DC subsets represent stages in a functional continuum or distinct functional types, the problem of recognizing diverse pathogens remains. The strategy of pattern recognition has evolved to deal with the problem of detecting microbes in the context of microbial heterogeneity and rapid evolution. Pattern recognition detects a limited set of conserved molecular patterns that are associated with microbes, often called pathogen-associated molecular patterns (PAMPs). These are recognized by pattern recognition receptors (PRRs) of which the Toll-like receptor (TLR) family is the best characterized (Akira and Takeda 2004). DCs and

other cell types such as epithelial cells and B cells express a range of the ten known human TLRs. In addition, DC subsets express different combinations of TLRs (Table 1). This implies a given DC population or subset will only respond to the pathogens for which they have the appropriate TLRs.

Table 1. TLRs expressed on human dendritic cells (adapted from Iwasaki and Medzhitov 2004).

	Monocytes[1]	mDCs[2]	pDCs[3]	*In vitro* differentiated DCs[4]
TLR1	+	+	+	+
TLR2	+	+	–	+
TLR3	–	+	–	+
TLR4	+	+	–	+
TLR5	+	+	–	+/–
TLR6	+	+	+	+
TLR7	+/–[5]	+/–	+	–
TLR8	+	+	–	+
TLR9	–	–	+	–
TLR10	–	+	+	?

[1] human CD14 positive monocytes.
[2] human myeloid dendritic cells.
[3] human plasmacytoid dendritic cells.
[4] Dendritic cells generated *in vitro* from $CD14^+$ monocytes by culture in the presence of GM-CSF and IL4.
[5] +/– indicates that the results of different studies do not concur.

TLRs differ in their ligand specificity and the signal transduction pathways they activate, thereby triggering antimicrobial and inflammatory responses and DC maturation. This comprises both conserved signaling pathways such as MyD88-dependent activation of the NFkB pathway, and receptor-specific signaling such as the MyD88-independent activation of the IRF3 transcription factor by TLR4 (Iwasaki and Medzhitov 2004). How these signals combine to influence the various facets of DC function is not known in detail. However, it is reasonable to assume that the pathogen recognition and receptor signaling logic must influence the transcriptional state of the DC and is therefore crucial for DC function.

7.2.2 Differential Outcomes; T_H1, T_H2, and Tolerogenic T-cell Responses

DCs are essential for activation and differentiation of naive T cells into T helper type 1 (T_H1) cells, T helper type 2 (T_H2) cells, and cytotoxic T lymphocytes. They are also important for stimulating mature B cells to become antibody-producing plasma cells in the absence of T-cell help (Jego, Palucka, Blanck, Chalouni, Pascual, and Banchereau 2003; Poeck, Wagner, Battiany, Rothenfusser, Wellisch, Hornung, Jahrsdorfer, Giese, Endres, and Hartmann 2004), and for inducing peripheral tolerance to self antigens. These processes, however, normally occur in secondary lymphoid organs. Therefore, following antigen exposure DCs must migrate to local lymphoid organs where they can interact with T and B cells. This involves functional reprogramming of DCs including the upregulation of costimulatory and MHC

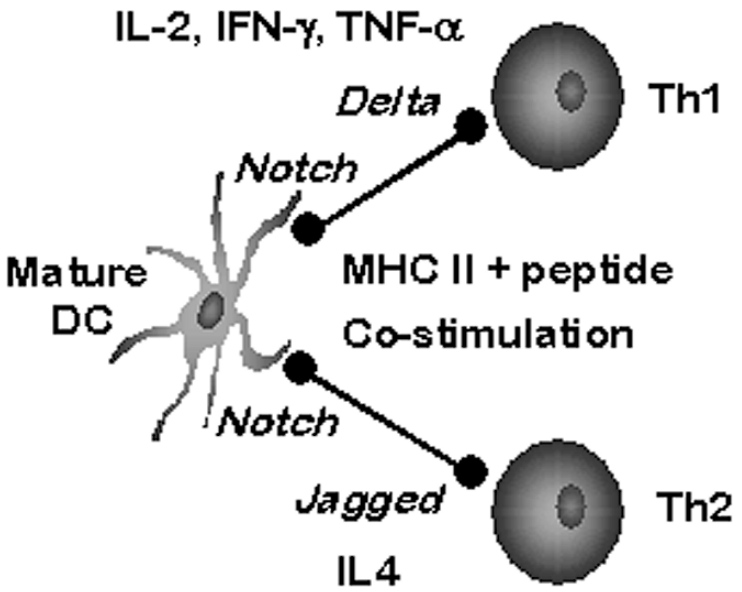


Fig. 1. Signals provided by mature dendritic cells during the polarization of naive T cells.

molecules, changes in chemokine receptor expression to allow trafficking to secondary lymphoid organs, and changes in cytokine and chemokine production to influence T and B cells. Once in the local lymphoid tissue, DC and lymphocyte interaction is far from simple.

Elegant experiments using two-photon microscopy of labeled DCs and T cells in murine lymph nodes *in vivo* show that during the first 8 hours of antigen-matured DC entry into the lymph node, mobile DCs and highly mobile T cells undergo short encounters but with a progressive decrease in cell motility of both. During the next 12 hours T cells form long-lasting stable conjugates with the DCs and begin to secrete specific cytokines. After 24 hours the T cells begin proliferating and resume rapid migration and short DC contacts, although DCs do not resume motility (Mempel, Henrickson, and Von Andrian 2004). The signals for prolonged DC/T-cell interaction are not known but demonstrate that the function of DCs may also change within lymphoid tissue. During the phase of prolonged interactions, the DC is likely to be providing the signals required for MHC-restricted T-cell proliferation. These signals include antigenic peptide loaded in MHC class I or II, costimulatory signals, cytokine signals (O'Garra 1998), and signals through the Notch/Delta/Jagged pathways (Maekawa, Tsukumo, Chiba, Hirai, Hayashi, Okada, Kishihara, and Yasutomo 2003; Amsen, Blander, Lee, Tanigaki, Honjo, and Flavell 2004). The strength, duration (Langenkamp, Messi, Lanzavecchia, and Sallusto 2000), and combinatorial logic of these signals provide the input for the T-cell to proliferate as a T_H1 or T_H2 polarized cell, thereby shaping the type of immune response to the pathogen (Fig. 1). DCs could also provide the signals for T-cell anergy or apoptosis in the absence of T-cell proliferation.

7.3 Gene Expression Programs in Antigen-Presenting Cells

Antigen presenting cells (APCs) must exhibit considerable functional plasticity to initiate the complex series of events that leads to these different adaptive immune responses being appropriate to the initial antigenic trigger. APCs have therefore been the subject of considerable interest in determining global gene expression programs using postgenomic tools such as DNA microarrays.

7.3.1 Common Transcriptional Reprogramming of Dendritic Cells by Pathogens

Before antigen encounter, it seems that APCs that are similar in ontogeny (such as macrophages and immature monocyte-derived DCs) are similar in gene expression profile, sharing 96% of their expressed gene networks (Chaussabel, Semnani, McDowell, Sacks, Sher, and Nutman 2003). Following exposure to the same pathogen the two cell types respond quite differently at the transcriptional level, with only 40% of the regulated genes being shared by both (Chaussabel et al. 2003). This is perhaps not too surprising as macrophages and DCs have different roles in the immune response. Macrophages and monocytes mainly sustain and control the local inflammatory process as well as phagocytosing and destroying microbes. The distinct functional roles of DCs, participating in the inflammatory process at the site of infection, while being able to transit to local lymphoid tissue to prime T cells, suggest a number of different functional and therefore transcriptional states for DCs.

Time course studies of murine and human DCs exposed to whole pathogens and PAMPs such as lipopolysaccharide (LPS) and double-stranded RNA (dsRNA), show DCs have at least three distinct transcriptional states (Huang, Liu, Majewski, Schulte, Korn, Young, Lander, and Hacohen 2001; Granucci, Vizzardelli, Virzi, Rescigno, and Ricciardi-Castagnoli 2001a; Granucci, Vizzardelli, Pavelka, Feau, Persico, Virzi, Rescigno, Moro, and Ricciardi-Castagnoli 2001b; Kwan & Kellam, unpublished observations). These "core" transcriptional responses occur independent of the type of antigen and reflect the essential functions of a maturing DC (Fig. 2).

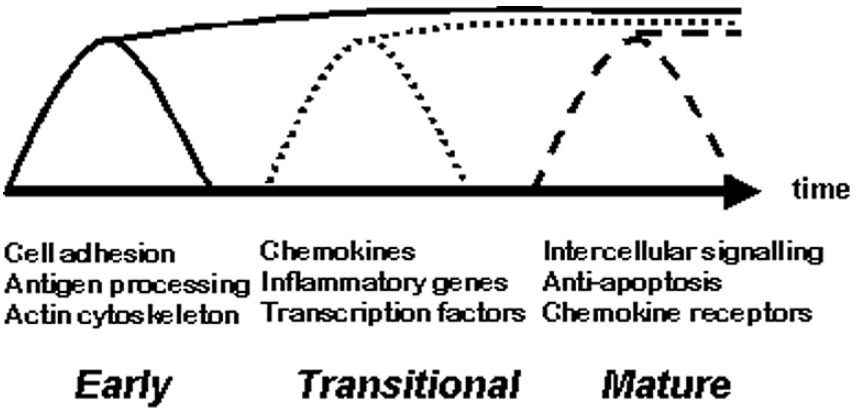


Fig. 2. Gene expression programs at different transcriptional stages of dendritic cell maturation. The solid and dashed lines represent transient and sustained upregulation of different functional groups of genes.

As immature DCs progress to an early activated state, gene expression is characterized by induction of genes involved in antigen processing and cytoskeletal rearrangements and downregulation of genes involved in pathogen recognition and phagocytosis. In addition, the transcription of proinflammatory genes such as TNFα and IL-1 is induced. Within a few hours the early "activated" DC progresses to a second "transitional" gene expression program consisting of the induction of many chemokines, transcription factors, and components of intracellular signaling pathways. Finally, by 18 hours following antigen exposure, DCs display a "mature" transcriptional program characterized by the induction of genes involved in apoptosis regulation, intercellular signaling and T-cell stimulation. These time-dependent classes of gene expression correspond broadly to known DC functions over time. Initially DCs are involved in pathogen phagocytosis. Following pathogen exposure, however, DCs rapidly influence the local innate immune response, producing cytokines that affect macrophage, natural killer cell, and neutrophil function. Finally, as discussed, DCs in local lymphoid tissue interact and stimulate T and B cells requiring the transcriptional induction of specific functions.

7.3.2 Plasticity in Dendritic Cell Transcriptional Programs

Given that dendritic cells can direct different outcomes at the level of T-cell polarization they must therefore interpret and translate different antigenic input stimuli to instruct these outcomes. As we have seen, two nonexclusive theories predict that different DC subsets produce differential outcomes to a given pathogen (2.1.1), or a given DC is flexible in its response. Transcriptional profiling of DCs exposed to different pathogens and PAMPs clearly shows the induction of different gene sets superimposed over the "core" DC maturation program (Huang et al. 2001; Granucci et al. 2001a; Granucci et al. 2001b; Kwan and Kellam, unpublished observations). The first study comparing bacterial, viral, and yeast exposure to DCs (Huang et al. 2001) showed that the bacterium, *E. coli* specifically induced the expression of 118 genes. These included the rapid activation of a potent inflammatory response and the later induction of T-cell stimulating genes including a subset of chemokines thought to attract naïve T_H2 T helper cells. In contrast, influenza virus strongly induced antiviral genes including type 1 interferons and interferon-inducible chemokines, although the induction of these genes may simply reflect the ability of influenza to replicate in DCs (Huang et al. 2001). A related study comparing LPS and TNFα stimulation of DCs showed very different patterns of DC gene expression but importantly identified the TNFα stimulus as providing only a "mild alert" effect compared to LPS which stimulates full DC maturation (Granucci et al. 2001a). These studies highlight the importance of the stimulatory context of different pathogens, but clearly highlight DC transcriptional plasticity to diverse pathogens, although care should be taken in such studies as different LPS preparations can have variable effects on DCs.

Using similar approaches we have determined that DCs can transcriptionally distinguish between two RNA genome viruses, human rhinovirus and influenza virus, which differ by the presence of a viral lipid envelope and the type of RNA genome. In addition, DCs can discriminate between replication-competent and inactivated viruses (Fig. 3).

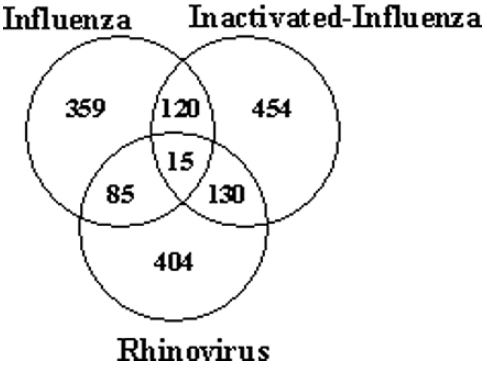


Fig. 3. Distribution of induced and repressed genes in dendritic cells responding to viruses.

The differences between viruses that infect DCs and those that do not, either through lack of DC tropism or because they are inactivated, have important implications for understanding immune control of viral infections. Certain viruses such as human immunodeficiency virus (HIV) (Izmailova, Bertley, Huang, Makori, Miller, Young, and Aldovini 2003) and murine cytomegalovirus (Andrews, Andoniou, Granucci, Ricciardi-Castagnoli, and Degli-Esposti 2001) are known to infect DCs and prevent authentic maturation. How these viruses prevent the maturation program is not known but should allow the identification of crucial functional gene networks that are required for DC maturation. In addition, the differences in DC transcriptional responses to live and inactivated influenza imply that inactivated vaccines may not produce the complete transcriptional repertoire that leads to fully authentic immunization. Understanding such differences should improve both vaccines for infectious agents and vaccines directed against tumors.

7.3.3 Transcriptional Plasticity Toward T_H1, T_H2, and Tolerogenic Immunity

Most of the DC gene expression studies have so far focused on antigens that stimulate a T_H1 response, raising the question as to whether transcriptional differences will define T_H1, T_H2, and tolerogenic DC states. Recently Ryan et al. described the transcriptional changes in monocyte-derived DCs to the contact allergen dinitrobenzenesulfonic acid (DNBS) (Ryan, Gildea, Hulette, Dearman, Kimber, and Gerberick 2004). Many of the core DC maturation program genes were upregulated consistent with observed DC maturation. Indications of DC gene expression differences were identified by comparing DNBS-induced genes with other published DC gene expression data. This suggested that Signaling Lymphocyte Activation Molecule (SLAM), known to be highly upregulated on DCs matured by LPS or dsRNA, was downregulated by DNBS. SLAM activation in T cells results in the production of T_H1 polarizing IFN-γ and in DCs in the production of IL-12 and IL-8 (Bleharski, Niazi, Sieling, Cheng, and Modlin 2001). SLAM activity may therefore be detrimental in a T_H2 polarizing environment perhaps explaining its down-regulation in DCs by DNBS.

Gene expression programs involved in the production of peripheral tolerance have so far only been investigated using DCs derived from a myelomonocytic tumor cell line, but also suggest that DC interaction with regulatory T cells can induce a tolerogenic gene expression program including the induction of antiapoptotic genes (Suciu-Foca Cortesini, Piazza, Ho, Ciubotariu, LeMaoult, Dalla-Favera, and Cortesini 2001). More detailed analysis of the way in which various pathogens and PAMPs affect transcriptional profiles leading to differential DC functions will help uncover the contribution of DC plasticity in shaping the immune response.

7.4 Integrating Genomics Data: A System View of the Immune Response

What becomes immediately clear with genome-scale surveys of transcriptional programs and protein interaction maps is that focusing on single genes and gene families will not provide an understanding of the complexity of the integrated immune response. This understanding requires abstraction of models from existing theories and high-dimensional genomics datasets. Modeling should not be an anathema to biology although models are often treated with suspicion relative to empirical data. This arises mainly through a misunderstanding of the purpose of models. Biologists intuitively construct models all the time by forming hypotheses. Such mental and verbal models concentrate on describing and integrating selected aspects of their research, leaving aside certain facts as irrelevant. Good models make planning the next experiment possible and allow a prediction of the results. If the results differ from predicted, the model is adjusted. Modeling in systems biology is simply a formalization of this process mathematically and extending the model to large datasets.

7.4.1 Models of the Immune Response

To be able to produce such abstracted models from large genomic datasets it is necessary to define the structure of interactions within a network of genes and proteins, determine the dynamic relationships between the gene and protein components, and determine the integrated network behavior. Insights come through either data-driven or model-driven approaches to these aims. Network structures are beginning to be compiled either by physically mapping protein—protein interactions (Bouwmeester, Bauch, Ruffner, Angrand, Bergamini, Croughton, Cruciat, Eberhard, Gagneur, Ghidelli, Hopf, Huhse, Mangano, Michon, Schirle, Schlegl, Schwab, Stein, Bauer, Casari, Drewes, Gavin, Jackson, Joberty, Neubauer, Rick, Kuster, and Superti-Furga 2004; Gavin, Bosche, Krause, Grandi, Marzioch, Bauer, Schultz, Rick, Michon, Cruciat, Remor, Hofert, Brajenovic, Ruffner, Merino, Klein, Hudak, Dickson, Rudi, Gnau, Bauch, Bastuck, Huhse, Leutwein, Heurtier, Copley, Edelmann, Querfurth, Rybin, Drewes, Raida, Bouwmeester, Bork, Seraphin, Kuster, Neubauer, and Superti-Furga 2002), by predicting network interactions from gene expression data (Zhou, Kao, and Wong 2002), and by computing networks from protein identification experiments using databases of known interactions (Yan, Lee,

Yi, Reiss, Shannon, Kwieciszewski, Coito, Li, Keller, Eng, Galitski, Goodlett, Aebersold, and Katze 2004). At present these networks are incomplete, dependent on the context of a given experiment and often lack the detail of knowledge available within the literature, all of which should improve with time.

Determining the dynamic relationships between network components from genomics data is a vibrant area of research but suffers from many problems, both technical and theoretical. Problems with lack of long time courses of geneexpression data, lack of integration of genome-scale transcription rate data, and lack of integration of existing knowledge bases all hinder dynamic modeling. Even with better data, advances in modeling multivariate time series data consisting of a great excess of variables to observations (as is true from gene expression microarrays) are needed (Kellam, Liu, Martin, Orengo, Swift, and Tucker 2001; Bar-Joseph 2004)

To date, the best approaches to modeling immune cell functions have used theoretical models that are tested against known biologically relevant parameters or intuitive estimates of the parameters. In immunology, such models have been assessed at the level of T-cell proliferation following priming by APCs (Allan, Callard, Stark, and Yates 2004). These models show control of $CD4^+$ and $CD8^+$ proliferation is mediated by different mechanisms. To prevent runaway T-cell expansion, the rate of apoptosis must progressively increase over time. Apoptosis mediated by cell-to-cell contacts alone is sufficient to regulate both $CD4^+$ and $CD8^+$ T-cell responses. However, if proliferation is controlled by other mechanisms such as cytokine signaling or by APCs then $CD8^+$ cells must change both apoptosis and cell division rates over time to reduce cell numbers possibly reflecting the need to rapidly reduce $CD8^+$ T-cell numbers after pathogen clearance to prevent immune pathology.

Intracellular signaling has also benefited from a system-theoretic approach for modeling TNFα-mediated NFκB signaling (Cho, Shin, Lee, and Wolkenhauer 2003). When signaling and cell phenotype models are combined, the influence of T-cell stimulating signals on T_H1 and T_H2 polarization can be modeled (Yates, Callard, and Stark 2004). The model shows specific T_H1 and T_H2 polarization signals give rise to rapid but reversible induction of the transcription factors T-bet (T_H1) and GATA-3 (T_H2). The model predicts that T_H differentiation can be reversed at the single cell level, suggesting a possible therapeutic means of manipulating T_H1 and T_H2 responses. Furthermore, such models would be of considerable interest in the context of DCs interacting with more than one T-cell, where potential T-cell interactions could be orchestrated and manipulated by the DC.

7.4.2 Systems Immunology

Despite the promise of models of immune cell function the problem remains of how we can assemble the parts into a more integrated understanding. This moves from modeling in isolation into a more systems biology framework. Theories of systems biology also impact on immunology both at intracellular signaling and at the cell phenotype and interaction level. One feature of complex biological systems is robustness against environmental and genetic changes. Robust systems, however,

also suffer from fragility as a trade-off. This may seem contradictory, but robustness in this sense means stability in the face of reasonable changes. For example, the complex autopilot system of a passenger airplane maintains a given flight path while being robust against changes in atmospheric conditions. The system is inherently fragile, however, to unexpected extreme events such as complete engine failure (Kitano 2004). The same applies to biological systems where robustness allows buffering against environmental changes and perhaps more importantly against small variations in genotype representing interindividual differences. Therefore, inter-individual variations will not necessarily affect the functioning of the system unless they are catastrophic defects, for example deficiency in the enzyme adenosine deaminase leading to Severe Combined Immunodeficiency (SCID) in children, or unless they affect the dynamics of the system, for example susceptibility to mycobacterial and *Salmonella* infections in individuals with defects in the IL-12 receptor. Both types of events can therefore define disease (Kitano 2004).

Systems control consists of a number of principles of which positive and negative feedback mechanism are well known in the immune system. Negative feedback is the main method of control that enables a robust response buffered against noise and perturbations. From this point of view, DCs sensing PAMPs might exhibit negative feedback control to buffer against differences in PAMP signaling intensity and strength allowing maintenance of a fixed time-dependent maturation program. Positive feedback contributes to a robust system by amplifying stimuli and often results in the phenomenon of bistability. This occurs in signal transduction/transcription settings when for a given range of input stimuli only a modest increase in transcription occurs, but for a different, higher range of input positive feedback results in a large increase in transcription. This in effect produces a bistable switch (Fig. 4). The modeling of T_H1 and T_H2 stimulation by Yates et al. clearly shows how the influence of bistable switch controls T_H1 or T_H2 polarization (Yates et al. 2004), making the induction of a state change from naïve to polarized T-cell insensitive to the absolute level of initial input signals provided they are above a switch threshold.

Positive feedback bistability is therefore important for a robust immune response allowing both the amplification of variable input signals as will occur through differences in receptor affinities and signaling intensities in inter-individual variation. Positive feedback bistability should also allow a layer of functional redundancy in combinatorial signaling cascades where the absence of one component of a signal can be compensated for by increasing other components to achieve the desired switch threshold. In both cases a robust and consistent output is conferred from variable level inputs. The system breaks down and disease occurs when a switch threshold is not reached. How to investigate relative inputs and switch thresholds represents a considerable challenge, but it is possible that experiments and models of viral interaction with DCs may prove particularly informative.

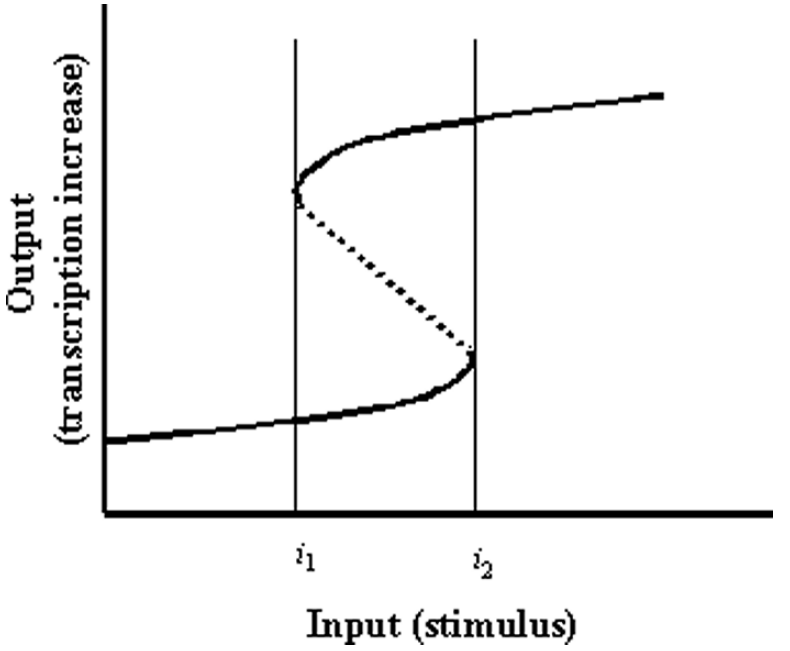


Fig. 4. A schematic representation of bistability. Inputs below i_1 result in only a small increase in transcription (y axis). Above i_1 the system exhibits two stable states (solid line), the higher state being a result of positive feedback driving high-level transcription. Above i_2 the system is insensitive to higher input signal and between i_1 and i_2 the system is insensitive to small reductions in input signal once the higher state is achieved.

7.5 Conclusions

Here we have shown some of the complexity and progress in determining genome-scale views of DC function. Modeling of immune system components, especially models that define input and output parameters, may allow the chaining of model components with one providing outputs that serve as inputs for the next. Modeling DC gene expression programs corresponding to cytokine and cell-surface molecule expression may well provide a range of inputs into the models of T-cell polarization described (Yates et al. 2004). Understanding such integrated processes within the immune system should help improve therapies against infections, produce better vaccines against infectious agents and tumors, and provide means of intervening in autoimmune disease.

Acknowledgements

We would like to acknowledge Wendy Barclay, Dave Rowlands, and Toby Tuthill for providing us with purified viral preparations for our studies. We would also like to acknowledge our colleagues, Benny Chain, David Katz, and Jane Rasaiyaah, for their

helpful comments and critical reading of this chapter. Antonia Kwan is supported by grants from the Triangle Trust and an Overseas Research Student Award.

References

Akira, S., and Takeda, K. (2004) Toll-like receptor signalling. Nat. Rev. Immunol. 4:499-511.

Allan, M.J., Callard, R., Stark, J., and Yates, A. (2004) Comparing antigen-independent mechanisms of T cell regulation. J. Theor. Biol. 228:81-95.

Amsen, D., Blander, J.M., Lee, G.R., Tanigaki, K., Honjo, T., and Flavell, R.A. (2004) Instruction of distinct CD4 T helper cell fates by different notch ligands on antigen-presenting cells. Cell 117:515-526.

Andrews, D.M., Andoniou, C.E., Granucci, F., Ricciardi-Castagnoli, P., and Degli-Esposti, M.A. (2001) Infection of dendritic cells by murine cytomegalovirus induces functional paralysis. Nat. Immunol. 2:1077-1084.

Bar-Joseph, Z. (2004) Analyzing time series gene expression data. Bioinformatics 20:2493-2503.

Bleharski, J.R., Niazi, K.R., Sieling, P.A., Cheng, G., and Modlin, R.L. (2001) Signaling lymphocytic activation molecule is expressed on CD40 ligand-activated dendritic cells and directly augments production of inflammatory cytokines. J. Immunol. 167:3174-3181.

Bouwmeester, T., Bauch, A., Ruffner, H., Angrand, P.O., Bergamini, G., Croughton, K., Cruciat, C., Eberhard, D., Gagneur, J., Ghidelli, S., Hopf, C., Huhse, B., Mangano, R., Michon, A.M., Schirle, M., Schlegl, J., Schwab, M., Stein, M.A., Bauer, A., Casari, G., Drewes, G., Gavin, A.C., Jackson, D.B., Joberty, G., Neubauer, G., Rick, J., Kuster, B., and Superti-Furga, G. (2004) A physical and functional map of the human TNF-alpha/NF-kappa B signal transduction pathway. Nat. Cell Biol. 6:97-105.

Callard, R., George, A.J., and Stark, J. (1999) Cytokines, chaos, and complexity. Immunity 11:507-513.

Chaussabel, D., Semnani, R.T., McDowell, M.A., Sacks, D., Sher, A., and Nutman, T.B. (2003) Unique gene expression profiles of human macrophages and dendritic cells to phylogenetically distinct parasites. Blood 102:672-681.

Cho, K.H., Shin, S.Y., Lee, H.W., and Wolkenhauer, O(2003) Investigations into the analysis and modeling of the TNF alpha-mediated NF-kappa B-signaling pathway. Genome Res. 13: 2413-2422.

Gavin, A.C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J.M., Michon, A.M., Cruciat, C.M., Remor, M., Hofert, C., Brajenovic, M., Ruffner, H., Merino, A., Klein, K., Hudak, M., Dickson, D., Rudi, T., Gnau, V., Bauch, A., Bastuck, S., Huhse, B., Leutwein, C., Heurtier, M.A., Copley, R.R., Edelmann, A., Querfurth, E., Rybin, V., Drewes, G., Raida, M., Bouwmeester, T., Bork, P., Seraphin, B., Kuster, B., Neubauer, G., and Superti-Furga, G. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 415:141-147.

Granucci, F., Vizzardelli, C., Virzi, E., Rescigno, M., and Ricciardi-Castagnoli, P. (2001a) Transcriptional reprogramming of dendritic cells by differentiation stimuli. Eur. J. Immunol. 31:2539-2546.

Granucci, F., Vizzardelli, C., Pavelka, N., Feau, S., Persico, M., Virzi, E., Rescigno, M., Moro, G., and Ricciardi-Castagnoli, P. (2001b) Inducible IL-2 production by dendritic cells revealed by global gene expression analysis. Nat. Immunol. 2:882-888.

Hoebe, K., Janssen, E., and Beutler, B. (2004) The interface between innate and adaptive immunity. Nat. Immunol. 5:971-974.

Huang, Q., Liu, D., Majewski, P., Schulte, L.C., Korn, J.M., Young, R.A., Lander, E.S., and Hacohen, N. (2001) The plasticity of dendritic cell responses to pathogens and their components. Science 294:870-875.

Iwasaki, A., and Medzhitov, R. (2004). Toll-like receptor control of the adaptive immune responses. Nat. Immunol. 5:987-995.

Izmailova, E., Bertley, F.M., Huang, Q., Makori, N., Miller, C.J., Young, R.A., and Aldovini, A. (2003) HIV-1 Tat reprograms immature dendritic cells to express chemoattractants for activated T cells and macrophages. Nat. Med. 9:191-197.

Jego, G., Palucka, A.K., Blanck, J.P., Chalouni, C., Pascual, V., and Banchereau, J. (2003) Plasmacytoid dendritic cells induce plasma cell differentiation through type I interferon and interleukin 6. Immunity 19:225-234.

Kellam, P., Liu, X., Martin, N., Orengo, C., Swift, S., and Tucker, A. (2001) A framework for modelling short, high-dimensional multivariate time series. Preliminary results in virus gene expression data analysis. In: *Advances in Intelligent Data Analysis: 4th International Conference, IDA 2001. Lecture Notes in Computer Science 2189*. Springer, Berlin, pp. 218-227.

Kitano, H. (2004) Biological robustness. Nat. Rev. Genet. 5:826-837.

Langenkamp, A., Messi, M., Lanzavecchia, A., and Sallusto, F. (2000) Kinetics of dendritic cell activation: Impact on priming of TH1, TH2 and nonpolarized T cells. Nat. Immunol. 1:311-316.

Maekawa, Y., Tsukumo, S., Chiba, S., Hirai, H., Hayashi, Y., Okada, H., Kishihara, K., and Yasutomo, K. (2003) Delta1-Notch3 interactions bias the functional differentiation of activated CD4+ T cells. Immunity 19:549-559.

Mempel, T.R., Henrickson, S.E., and Von Andrian, U.H. (2004) T-cell priming by dendritic cells in lymph nodes occurs in three distinct phases. Nature 427:154-159.

O'Garra, A. (1998) Cytokines induce the development of functionally heterogeneous T helper cell subsets. Immunity 8:275-283.

Poeck, H., Wagner, M., Battiany, J., Rothenfusser, S., Wellisch, D., Hornung, V., Jahrsdorfer, B., Giese, T., Endres, S., and Hartmann, G. (2004). Plasmacytoid dendritic cells, antigen, and CpG-C license human B cells for plasma cell differentiation and immunoglobulin production in the absence of T-cell help. Blood 103:3058-3064.

Ryan, C.A., Gildea, L.A., Hulette, B.C., Dearman, R.J., Kimber, I., and Gerberick, G.F. (2004) Gene expression changes in peripheral blood-derived dendritic cells following exposure to a contact allergen. Toxicol. Lett. 150:301-316.

Shigematsu, H., Reizis, B., Iwasaki, H., Mizuno, S., Hu, D., Traver, D., Leder, P., Sakaguchi, N., Akashi, K. (2004). Plasmacytoid dendritic cells activate lymphoid-specific genetic programs irrespective of their cellular origin. Immunity 21:43-53.

Shortman, K., and Liu, Y.J. (2002) Mouse and human dendritic cell subtypes. Nat. Rev. Immunol. 2:151-161.

Suciu-Foca Cortesini, N., Piazza, F., Ho, E., Ciubotariu, R., LeMaoult, J., Dalla-Favera, R., and Cortesini, R. (2001) Distinct mRNA microarray profiles of tolerogenic dendritic cells. Hum. Immunol. 62:1065-1072.

Xie, H., Ye, M., Feng, R., and Graf, T. (2004). Stepwise reprogramming of B cells into macrophages. Cell 117:663-676.

Yan, W., Lee, H., Yi, E.C., Reiss, D., Shannon, P., Kwieciszewski, B.K., Coito, C., Li, X.J., Keller, A., Eng. J., Galitski, T., Goodlett, D.R., Aebersold, R., and Katze, M.G. (2004) System-based proteomic analysis of the interferon response in human liver cells. Genome Biol. 5:R54.

Yates, A., Callard, R., and Stark, J. (2004) Combining cytokine signalling with T-bet and GATA-3 regulation in Th1 and Th2 differentiation: A model for cellular decision-making. J. Theor. Biol. 231:181-196.

Zhou, X., Kao, M.C., and Wong, W.H. (2002) Transitive functional annotation by shortest-path analysis of gene expression data. Proc. Natl. Acad. Sci. USA 99:12783-12788.

Zuniga, E.I., McGavern, D.B., Pruneda-Paz, J.L., Teng, C., and Oldstone, M.B. (2004) Bone marrow plasmacytoid dendritic cells can differentiate into myeloid dendritic cells upon virus infection. Nat. Immunol. 5:1227-1234.

Chapter 8

Understanding the Immune System by Computer-Aided Modeling

Massimo Bernaschi and Filippo Castiglione

Institute for Computing Applications (IAC), National Research Council (CNR), Viale del Policlinico 137, 00161 Rome, Italy, m.bernaschi@iac.cnr.it and f.castiglione@iac.cnr.it

Abstract. We describe some computer models of the immune system and in particular of its response to the HIV infection. Then we introduce our model and show some results of simulations of the AIDS disease progression.

8.1 Introduction

Recently the term "systems biology" started to become popular among scientists working in the interdisciplinary field of theoretical biology. It suggests that to understand biology we should examine the structure and the dynamics of the combined system components rather than the isolated parts of a cell or organism (Kitano 2001). Along this line some general properties of biological systems as adaptation, insensitivity to specific parameters, and graceful degradation have been identified. It is interesting to note that engineers routinely use similar concepts to describe the properties of mechanical or electric systems.

It is a relatively new field of science that involves the application of experimental, theoretical, and modeling techniques to the study of biological organisms at all levels, from the molecular, through the cellular, up to the behavioral (Kitano 2001).

Such long-term goal, though well posed in scientific terms, appears quite ambitious. On a more practical ground, the problems can be formulated in a narrower context. For example, in the design of drugs, the objective is to find a synthetic molecule that binds with high affinity a certain protein (or nucleic acid macromolecule) and either blocks its normal function or mimics another ligand for structures as the receptors in order to induce a normal physiological response. Actually, this task includes two distinct subproblems: the first one is to identify the molecules, the second one is to understand the impact of the drug on the organism. Though not independent, they can be addressed with different mathematical techniques and by different expertise. As another example, consider the question to find the best schedule

for the drug administration to a patient undergoing a certain therapy or, closer to our topic, for an immune stimulator used in immune therapies. To these ends, the mathematics can describe just a part of the whole system and, based on clinical data, can help clinicians to choose among a set of possibilities spelled out on the basis of their experience.

The present chapter aims to describe examples of the theoretical models that scientists employ to address the second of the aforementioned problems. As a matter of fact, in a number of situations, the molecular interactions can be included in very simple and stylized mathematical terms in equations describing the relationships among cells and molecules. The real value of the mathematical models is their capability to describe the complex relationships among the large number of components of the (immune) system. By means of perturbations of the *network* of molecules/cells/organs, it is possible to understand many properties of the whole system. In other words, it is possible to answer questions like "what happens to the immune response if we administer this drug which is able to block the action of this molecule?" or "what is the best administration dose/schedule/route for interleukin-2 in terms of a stimulation of the response?" or "what is the net effect of downregulating the expression of such receptor on the surface of the lymphocytes?"

In the following sections we introduce some mathematical/computer models of the immune response and in particular of the HIV infection. Then, we describe our model (C-ImmSim) and the way we use it to uncover some unknown aspects of the pathology and to make predictions about AIDS disease progression.

8.2 Computational Immunology

In the development of a mathematical model of the immune system of a vertebrate animal, three levels of abstraction are usually considered: the *microscopic* level which is the scale of subcellular activities (e.g., DNA synthesis and degradation, gene expression, alteration mechanisms of the cell cycle, absorption of vital nutrients, activation and inactivation of receptors, transduction of chemical signals within the cell that regulate cellular activities such as duplication, motion, adhesion, or detachment), the *mesoscopic* level (that refers to the cellular level and therefore to the main activities of the cell population, e.g., the statistical description of the progression and activation state, cooperation/competition, aggregation properties, and intra/extravasation processes), and the *macroscopic* level (the tissue level that refers to the typical phenomena of continuum systems, e.g., cell migration, convection and diffusion of nutrients and chemical factors, mechanical responses, interactions with external tissues) (Preziosi 2003).

Besides that, the mathematical models of the immune system can be classified according to: (1) the mechanism of regulation of immune processes (i.e., the clonal selection theory versus. the idiotypic network); (2) the choice of the *mathematical space* (i.e., continuous versus. discrete time and/or space), and (3) the presence of stochastic (i.e., nondeterministic) components.

The implications of the first modeling choice are well known to biologists, whereas the other two require a few more words. The difference between discrete

and continuous models is in the choice of the numerical space chosen to represent, for instance, the time scale. It is pretty different to let time change continuously, and to consider quanta of time in which a number of events may happen simultaneously. The choice of the mathematical space determines the mathematical techniques one is able to use, the kind of solution that can be derived and, in general, the difficulty in carrying on a nontrivial analysis. A deterministic process is completely predictable provided an exact knowledge is available about the initial conditions whereas a stochastic process, by definition, can be described only by means of the methods of statistics and probability theory. Stochastic components are often included to take into account that we do not have perfect knowledge either of the initial conditions or of the process that a system follows (or both). Therefore, stochastic models seem more suitable to describe biology but more difficult to analyse.

All existing models of the immune system derive from either the *clonal selection theory* or the *idiotypic network theory*. Nowadays immunologists consider these as two independent and perhaps complementary theories (Zorzenon dos Santos 1999). However, while clonal selection theory is believed to be the fundamental one in the present knowledge of the immune system, the idiotypic network theory is believed correct as far as the existence of anti-idiotypic reactions but probably not relevant in determining the immune response (Anachini and Mortarini 1999).

Both immunological theories inspired a number of continuous models (Perelson 1988a; Perelson 1988b) whereas most of the discrete models are based on Jerne's (*idiotypic network*) theory (Jerne 1973; Jerne 1974). On the other hand, the Celada–Seiden model, which *may* include both theories, rests its foundation on the clonal selection theory (Celada and Seiden 1992).

Basically, continuous mathematical models try to represent the dynamics of certain quantities, like the number of cells or the concentration of a molecule in a compartment of the immune system, as a function of the birth and death rate and of other variables. A simple example is the so-called AB model (Segel and Perelson 1991) used in theoretical studies of the immune network. It describes the dynamics of the i-th clone (i=1,…,M) of B-lympocytes ($B_i(t)$) and antibodies ($A_i(t)$) by means of the following two equations:

$$dB_i / d_t = m + B_i (pf(h_i) - dB) \tag{1}$$

$$dA_i / d_t = sB_i f(h_i) - d_C h_i A_i - d_A A_i \tag{2}$$

Here m is the rate of maturation of B cells from the bone marrow, p is the division rate of activated lymphocytes, d characterizes the death rate, s is the secretion rate of antibodies, and d_C is the elimination rate of antibody–antigen complexes. Idiotypic interactions are mediated through a field h_i which is determined by the concentration of antibodies A_j and by the affinity J_{ij}:

$$h_i = \sum_{j=1}^{M} J_{ij} A_i \tag{3}$$

The activation function $f(h_i)$ determines the intensity of the stimulation of the clone i by the idiotypic field h_i. These equations are used to study the dynamical regimes: oscillatory, chaotic, etc. Further details are outside the scope of the present work.

The main task of the immune system is to recognize the antigen by means of cell receptors. The binding mechanism, whose fine details are mostly unknown, is based on different physical-chemical effects (short-range noncovalent interactions, hydrogen binding, van der Waals interactions, etc.) (Perelson and Weisbuch 1997). The binding of a receptor with a molecule requires that they complement each other over a significant portion of their surface. This *generalized shape* is the constellation of features that determine the binding among molecules (Perelson and Oster 1979). Under the assumption that the shape can be described by K parameters, a point in a K-dimensional space (the *shape space*) specifies the generalized shape of the molecular binding region. Oster and Perelson estimated that in order to be complete the receptor repertoire should fulfill the following conditions: (1) each receptor should recognize a set of related epitopes, each of which differs slightly in shape; (2) the repertoire size should be on the order of 10^{16} or larger; (3) at least a subset of the repertoire should be distributed randomly throughout the shape space (Perelson and Weisbuch 1997).

Later, Farmer, Packard, and Perelson (1986) introduced the idea of using binary strings to represent the shape of a receptor. To determine the degree of affinity between strings it is possible to resort to different string-matching criteria. For instance, by using a "key–lock" analogy, two binary strings have high affinity if they "complement" each other, that is, when the two strings are lined up every 0 in one corresponds to a 1 in the other and conversely.

The following paragraph is an introduction (by no means complete) to mathematical models of the immune system response and in particular to the class of models more close to our own approach that will be discussed later in this chapter. At this point of the discussion we just need to say that we deal with a stochastic and discrete description of the immune system response at the cellular scale.

8.2.1 An Overview of Discrete Models

A number of discrete models of the immune system working at the cellular scale have been proposed in the past by using a variety of different techniques and aims. The first aim in modeling the immune system is to reproduce the (primary and secondary) response. However, many other aspects of its behavior, like autoimmune diseases (Weisbuch and Atlan 1988), selection and hypermutation of antibodies during an immune reaction (Celada and Seiden 1996), autoimmunity and T-lymphocyte selection in thymus (Morpurgo, Serenthà, Seiden, and Celada 1995), the immune response to known virus inducing cancer (Melief, Toes, Medema, Van der Burg, Ossendorp and Offringa 2000), the acquired HIV-related tumors (Carbone and Gaidano 2001; Varthakavi, Smith, Deng, Sun, and Spearman 2002), etc., have been studied and modeled. Moreover, it is worth pointing out that the activation of the immune response against some tumors has been known since the 1950s. Other studies showed that tumor cells resort to escape mechanisms that prevent the activation of the immune system (Anachini and Mortarini 1999). Recent extensive reviews on

discrete models of the immune system can be found in Perelson and Weisbuch (1997), Zorzenon dos Santos (1999), and Forrest and Hofmeyr (2000).

The discrete models of the immune system in use at this time (Zorzenon dos Santos 1999) can be classified, from the technical viewpoint, as Boolean networks, cellular automata, and lattice gases. We now describe briefly some of these models. Kaufman, Urbain, and Thomas (1985) proposed one of the first applications of discrete automata in the study of the adaptive immune response. The original model (*KUT model*) considered five types of cells and molecules but, for the sake of simplicity, we describe here a simplified version (Fig. 1). This "submodel" considers antibodies (Ab), helper cells (T_H), lymphocytes, B cells (B), and antigens (Ag). Each entity is represented by a two-valued variable denoting high/low concentration. The rules governing the network of interdependencies/interactions among these variables are expressed by logical operations. The application of the rules is iterated over discrete steps and the dynamics is observed. The rules are:

$$\begin{aligned} Ab_{t+1} &= Ag_t \ \text{AND} \ B_t \ \text{AND} \ T_{H\,t} \\ B_{t+1} &= T_{H\,t} \ \text{AND} \ (Ag_t \ \text{OR} \ B_t) \\ Ag_{t+1} &= Ag_t \ \text{AND NOT} \ Ab_t \\ T_{H\,t+1} &= T_{H\,t} \ \text{OR} \ Ag_t \end{aligned} \tag{4}$$

where AND, OR, and NOT are the standard logical operators. These rules should be read as follows: Abs are produced at time step $t+1$ by B cells if (at time t) the antigen is present together with stimulating lymphocytes and T_H cells; the B cells grow if T_H cells and either antigens or other B cells are present; the antigen proliferates in the absence of

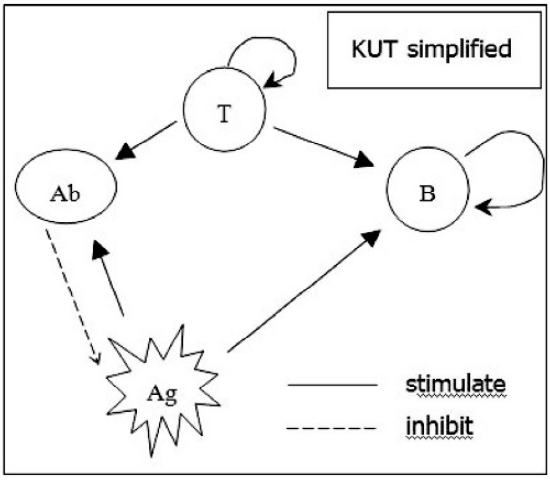


Fig 1. Simplied KUT Model with antibodies (Ab), T helper cells (T), lymphocytes, (B) cells, and antigens (Ag).

antibodies; T_H cells depend on the presence of other helpers (to keep homeostasis) or of the antigens (implicit presentation is assumed). A network like that in the figure can schematically represent this simplified model and its evolution in time can be studied by means of computer simulations. For example, the complete KUT model has five fixed points in the state space composed by 2^5=32 points. Fixed points identify the global state of the immune system: naive, vaccinate, immune, paralyzed, paralyzed and sick.

Many authors followed this simple approach. Weisbuch and Atlan proposed a model (*WA model*) (Weisbuch and Atlan 1988), based on Jerne's theory, to study the special case of autoimmune diseases, like multiple sclerosis, in which the immune system attacks the cells of the nervous system. As in the KUT model, this model uses five binary variables representing killer cells (S_1), activated killers (S_2), suppressor cells (S_3), helper cells (S_4), and activated suppressor cells stimulated by the helpers (S_5). The different types of cells influence each other with a strength that is 1, 0, or –1. The system evolves according to the following rule: at each time step, the concentration of one variable is set to unity if the sum of the interactions with the various cell types is positive, otherwise the concentration is set to zero. This model shows the existence of only two basins of attraction over 2^5=32 possible states: the empty state where all the concentrations are zero and a state where only activated killers disappear whereas the other four concentrations are unity.

These two models have been extensively studied (Stauffer 1989; Pandey and Stauffer 1990; Atlan and Cohen 1989). Moreover, Pandey and Stauffer further extended the KUT model by using a probabilistic generalization of the original deterministic cellular automata. Their model tried to provide a possible explanation of the time delay between HIV infection and the onset of AIDS (Pandey and Stauffer 1990; Pandey and Stauffer 1989). They represented helper cells (H), cytotoxic cells (S), virus (V), and interleukin (I). The interleukin molecules produced by helper cells induce the suppressor cells to kill the virus. The dynamics shows an oscillatory behavior followed by a fixed point where the immune system is totally destroyed, similar to the real onset of the AIDS.

Dayan, Havlin, and Stauffer (1988) studied the WA model on a square lattice in order to take into account spatial fluctuations of cell concentrations. In their model each lattice point influences itself and its nearest neighbors with the same rules of the WA model. Interestingly, this lattice version of the WA model was found to have a different dynamics compared to the original WA model as the number of fixed points is smaller.

Chowdhury and Stauffer (Chowdhury and Stauffer 1992; Chowdhury 1998) proposed a unified model of the immune system that includes, as special cases, the KUT and WA models. The model describes the immune response to HIV and reproduces some features of experimental data. Chowdhury and Stauffer also proposed extensions of the original network approach for modeling HIV and cancer (Chowdhury and Stauffer 1992).

A majority rule cellular automaton was used by Agur (1991) to study the signal processing in a multilayered network. Chowdhury, Deshpande, and Stauffer (1994) proposed a model to describe the interaction between various types of immune components considering intra- and interclonal interactions.

The interaction among different T_H cell subsets (Brass, Bancroft, Clamp, Grencis, and Else 1994) and the HIV interaction with T cells have been modeled as well

(Pandey and Stauffer 1990; Sieburg, McCutchan, Clay, Caballero, Ostlund, and James 1990; Mosier and Sieburg 1994; Zorzenon dos Santos and Coutinho 2001). Other possible approaches to study different aspects of the immune system dynamics in a discrete space/time framework can be found in Atlan and Cohen (1989).

Hereafter, we introduce specific models of the HIV infection. The focus is on C-ImmSim that can be seen as a synthesis of a number of different approaches.

8.3 Discrete Models of HIV Infection

As an extension of the already mentioned work with Stauffer (Pandey and Stauffer 1990; Pandey 1991), Pandey proposed a model that views a whole organism (e.g., a person) as a three-dimensional cellular automaton. The entities represented are the helper T cells, the killer T cells, macrophages, and virions (cells infected by HIV or free virus particles).

The evolution of the automaton is determined by two sets of rules corresponding to different infection modes (fast replication followed by rupture of the cell and slower reproduction). The rules take the form of logical statements using the Boolean operators on the binary codes of the entities. The simulation of the model produced nontrivial results, but it did not show the characteristic "three-phase'' dynamics of HIV (Pantaleo, Graziosi, and Fauci 1993). An interesting variation of that model allowed the authors to study the viral load as a function of viral growth factor and mutation rate (Ruskin, Pandey, and Liu 2002).

Perelson and Nelson presented various models of HIV infection (Perelson and Nelson 1999) based on sets of simple ordinary differential equations (ODEs). By means of clinical data, the value of the parameters is estimated and a classic stability analysis is carried out for the different phases of the HIV infection. The authors described the impact of drug therapies on the HIV dynamics by means of appropriate changes in the equations. In a model extension, they also consider the role of macrophages and write the corresponding ODE. However, the model does not include B cells. This choice is very common (see other models below) and it may appear like a major omission in any description of the immune response.

Hershberg, Louzoun, Atlan, and Solomon (2001) proposed a discrete model that acts in the Perelson "generalized shape space." The mutations of HIV are represented by propagation of the virions in the infinite-dimensional shape space. There is neither a spatial structure nor a distinction of the immune system components. The status of each site of the (discrete) shape space is represented simply by the number of virions existing with that shape and by the number of immune system cells that recognize the same shape.

The virions of any shape proliferate exponentially killing immune system cells at random until an immune system cell with that shape starts to proliferate and kill the virions in turn. In the absence of virus diffusion in shape space (i.e., if there were no mutations), this would stop the dynamics of the system, that is, the disease would be defeated. The diffusion of the virus in the shape space is responsible for the continuation of the infection.

The authors are able to simulate the "three-phase" dynamics of the HIV infection. However, their analysis is pretty qualitative and does not take into account at all the complexity of the real immune system. For instance, there is no distinction between immune cells targeted by HIV and other cells.

The spatial distribution of the virus infection and diffusion in the host's body plays a major role in the work of Zorzenon dos Santos and Coutinho (2001). Their model describes the immune system cells in the lymphoid tissues that can be a target for HIV by means of a two-dimensional cellular automaton. Each lattice site contains a single cell and each time step corresponds to one week of real life.

The values of the parameters required to tune the system have been determined on the basis of experimental data. The results of the density of healthy and infected T cells are in good qualitative agreement with those reported by Pantaleo et al. (1993). However, this model does not describe how the immune system interacts with an antigen like HIV that mutates at an extremely high rate. Many important entities like the macrophages or the B cells are not included in the model.

8.3.1 A Detailed Model of the Immune Reaction

The model we use to study the immune response to HIV branched years ago from the Celada–Seiden model (Celada and Seiden 1992). In the original Celada–Seiden model a single lymph node (or generically a small portion of a secondary lymphoid organ) of a vertebrate animal is mapped onto a two-dimensional hexagonal lattice, with full periodic boundary conditions. The primary lymphoid organs, thymus and bone marrow, are modeled apart: the thymus is implicitly represented by the positive and negative selection of immature thymocytes before they get into the lymphatic system, whereas the bone marrow generates already mature B lymphocytes. Hence, on the lattice there are only immunocompetent lymphocytes.

The Celada–Seiden model belongs to the class of bit string models. Bit strings represent the "binding site" of cells and molecules as, for example, lymphocyte receptors (T lymphocyte receptor, B lymphocyte receptor, Major histocompatibility complex, antigen peptides and epitopes, immunocomplexes, etc.). The model includes the major classes of cells of the lymphoid lineage (T_H, cytotoxic T lymphocytes or CTLs, B lymphocytes, and antibody-producing plasma cells) and some of the myeloid lineage (macrophages and dendritic cells).

The interactions among these cells define their functional behavior. With respect to other immune system models, the Celada–Seiden model has an additional level of description, representing the intracellular processes of antigen digestion and presentation. Both the cytosolic and endocytic pathways are implemented. Usually, each time step of the simulation corresponds to 8 hours of "real life".

8.4 Simulation of HIV-1 Infection

To account for the features of HIV-1, a number of additions are required to the original Celada–Seiden model. First, the T_H cells become, along with the dendritic cells and the macrophages, a possible target of the antigen action. From our simulations, it

appears that the infection of each cell type has different and specific consequences. For instance, if HIV-1 does not infect dendritic cells, the only effect is a reduction of the cytotoxic activity in recognizing new strains of the virus.

On the contrary, extending the infection to macrophages has a more striking implication since it weakens the innate response during the first phase. Moreover, it partially reduces the number of active antigen-presenting cells (B cells are also presenting). In such a way the efficiency in stimulating the growth of helper cells and triggering both cytotoxic activity and humoral response is impaired.

We assume that the infection of a cell is a stochastic event. There is a fixed probability that an HIV-1 infects a target cell. Once inside the target cell, the virus remains silent until, with a certain probability p_w, it starts to transcribe its RNA genome in the host DNA. In such a way, we account for variants of the virus with a low value of p_w, which may be interpreted either as strains having a poor adaptation (those having good chances to become extinct) or as strains that are activated very late. Newly assembled virions in productively infected cells accumulate inside the cell at a rate given by another parameter, p_r. With the same rate p_r, a part of these virions bud from cell membranes. Hence, if p_r is high, the accumulation of virions inside the cells causes cell rupture and consequent release of viral content into extracellular space. Finally, HIV-1 mutates in productively infected cells with a mutation rate given by a third parameter, p_m.

The activation (p_w), replication (p_r), and mutation rate (p_m) are a triplet of numbers between 0 and 1. The virus is represented by two binary strings (each l bitlong), one corresponds to the epitope (i.e., the B-cell-receptor's binding site) and the other to the peptide (i.e., the MHC class I and II binding site). It is possible to specify an arbitrary number of epitopes and peptides. A string of l bits can assume 2^l possible values. However, since the virus is represented by one epitope and one peptide, each viral strain is identified by $2l$ bits. This means that for l equal to 12 the potential number of different virus strains becomes equal to 16,777,216 (2^{24}). Note that if p_m is equal to 1×10^{-2} per bit, the probability of having, at least, one mutation in a 24-bit string is $1-(1-p_m)^{24} \sim 0.22$, in accordance with other studies (Perelson and Nelson 1999).

Recently, we started to associate a "meaning" to parts of the bit strings in order to specify the functional properties of the simulated virus (Castiglione, Poccia, D'Offizi, and Bernaschi 2004). We map the genotype to the phenotype by means of a simple formula that computes the values p_w, p_r, and p_m from different, nonoverlapping, zones of the binary string that describes the epitope of the virus. Since the bit-mutation is completely random, it may flip any of the bits representing the peptide or the epitope. In either case there is a nontrivial outcome: (1) if the peptide is modified, the affinity with the (class I or II) MHC molecules changes. This corresponds to the appearance of variants of the virus that might not undergo the cytotoxic activity of CD8 cells. (2) if the epitope is modified, then one of the three values of the triplet (p_w, p_r, p_m) is modified.

By looking at the results of the simulations, one can observe the population dynamics of the lymphocyte cells classified according to their specificity and state (i.e., duplicating, anergic, and so on), the plasma viremia and proviral HIV, the concentration of anti-HIV antibodies produced, and the magnitude of cytotoxic response. An example

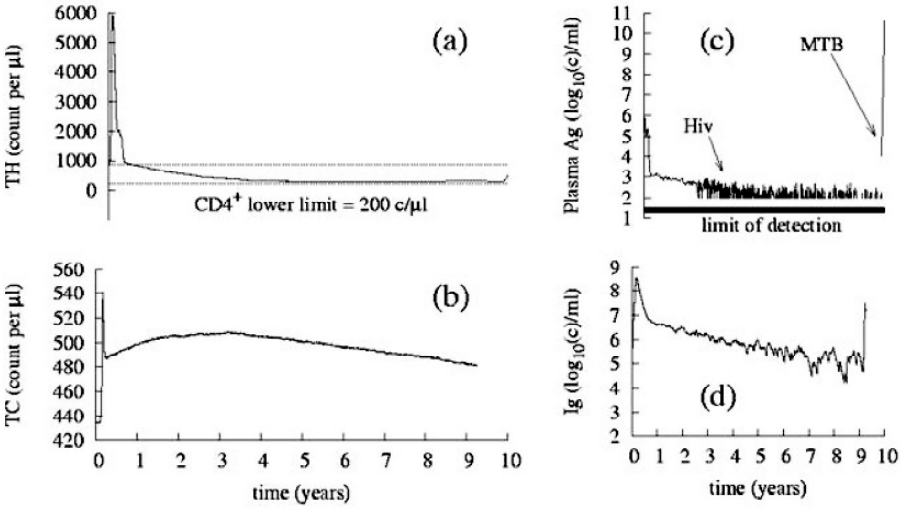


Fig. 2. We simulate the onset of an opportunistic disease by injecting a bacterium (e.g., *Mycobacterium tuberculosis*) during the ninth year after the primal infection with HIV. While the HIV count is still under control by means of a continuous cytotoxic activity (plot b) and by a humoral response (plot d), the system that is weakened in its CD4 count (plot a) cannot cope with the replication rate of the newly injected bacteria (plot c).

of output is shown in Fig. 2 that is produced by the simulation of a number of cells contained in 10 µl of peripheral blood (about 20,000 immune cells with a repertoire of 4096 possible receptors fighting with an equivalent number of potential HIV strains).

Besides that, we can look at the HIV peptides that are expressed on the surface of the antigen-processing cells, to determine if the wild-type virus undergoes mutations able to escape MHC presentation. Likewise we can monitor any viral mutation to check if there is a systematic behavior of the virus that allows it to escape the immune system control (Bernaschi and Castiglione 2006).

An interesting feature of the model is the possibility to schedule the injection of an arbitrary antigen at any time. For instance, at the beginning of a simulation, we can infect the "virtual patient" with a wild-type HIV that weakens the immune system. Then, during the simulation, it is possible to challenge again the immune system by means of a fast-replicating bacterium like, for example, *Mycobacterium tuberculosis* (MTB). What we get, as depicted in Fig. 2, is the appearance of the expected opportunistic disease. The MTB is injected 9 (simulated) years after the primal infection with HIV. At the time of the second antigenic challenge, the HIV is still under control by means of a continuous anti-HIV cytotoxic activity (plot b) and by a specific humoral response (plot d). However, in the latent phase of the AIDS disease, the system is weakened in its CD4 count (plot a) and eventually cannot cope with the replication rate of the newly injected bacteria (plot c).

Note, however, that by injecting the same bacterium at the time the CD4 count is not too low, the immune system is able to defeat the challenge (not shown), as it happens in reality for healthy immune systems. This demonstrates

not only that the model is sensible enough to any kind of validation test but also that it is ready to be used to test HAART therapy as we expect to do in the near future.

8.5 Conclusions

We presented some general concepts about mathematical modeling of the immune system. We focused on those models that resort to the discrete representation of time and space. Then we introduced the model that we currently use to study the HIV infection in order to show how a computational model can both reproduce what is known about the dynamics of a biological phenomenon and then make predictions about other aspects that cannot be easily measured or observed in clinical experiments. In this way we tried to give an idea of how mathematical/computer modeling can help biologists to understand all the aspects of a disease and provide indications to clinicians about a possible therapy. Hopefully, a closer collaboration among mathematicians, computer scientists, and biologists, along with the availability of large clinical datasets, will enhance the understanding of diseases and suggest new drug discoveries and/or therapeutic regimens.

References

Agur, Z. (1991) Fixed points of majority rule cellular automata with application to plasticity and precision of the immune system. Complex Syst. 5:351-357.

Anachini, A., and Mortarini, R. (1999) Il Controllo della Crescita Neoplastica: Ruoli Dell'Immunità specifica nella risposta alle metastasi. In: G. Bevilacqua, R. Gavari, and P.L. Lollini (Eds.), *Invasione e Metastasi*. Pacini Editore, Pisa, pp. 325-340.

Atlan, H., and Cohen, I.R. (1989) *Theories of Immune Networks*. Springer-Verlag, Berlin.

Bernaschi, M. and Castiglione, F. (2002) Selection of escape mutants from immune recognition during HIV infection. Immunol. Cell Biol. 80:307-313.

Bernaschi, M., and Castiglione, F. (2006) HIV-1 strategies of immune evasion. Int. J. Mod. Phys. C 16:1869-1879.

Brass, A., Bancroft, A.J., Clamp, M.E., Grencis, R.K., and Else, K.J. (1994) Dynamical and critical behavior of a simple discrete model of the cellular immune system. Phys. Rev. E 50:1589-1593.

Burnet, F. (1959) *The Clonal Selection Theory of Acquired Immunity.* Vanderbilt University, Nashville.

Carbone, A., and Gaidano, G. (2001) Acquired immunodeficiency syndrome-related cancer. A study model for the mechanisms contributing to the genesis of cancer. Eur. J. Cancer 37:1184-1187.

Castiglione, F., Poccia, F., D'Offizi, G., and Bernaschi, M. (2004) Mutation, fitness, viral diversity and predictive markers of disease progression in a computational model of HIV-1 infection. AIDS Res. Hum. Retrovir. 20:1314-1323.

Celada, F., and Seiden, P.E. (1992) A computer model of cellular interaction in the immune system. Immunol. Today 13(2):56-62.

Celada, F., and Seiden, P.E. (1996) Affinity maturation and hypermutation in a simulation of the humoral immune response. Eur. J. Immunol. 26:1350-1358.

Chowdhury, D., Stauffer, D., and Choudary, P.V. (1990) A unified discrete model of immune response. J. Theor. Biol. 145:207-215.

Chowdhury, D., and Stauffer, D. (1992) Statistical physics of immune networks. Physica A 186:61-81.

Chowdhury, D., Deshpande, V., and Stauffer, D. (1994) Modelling immune network through cellular automata: A unified mechanism of immunological memory. Int. J. Mod. Phys. C 5:1049-1072.

Chowdhury, D. (1998) Immune network: An example of complex adaptive systems. In: D. Dasgupta (Ed.), *Artificial Immune Systems and Their Applications*. Springer-Verlag, Heidelberg, pp. 84-104.Cohen, I.R., and Atlan, H. (1989) Network regulation of autoimmunity: An automation model. J. Autoimmun. 2:613-625.

Dayan, I., Havlin, S., and Stauffer, D. (1988) Cellular automata generalization of the Weisbuch-Atlan model for immune response. J. Phys. A 21:2473-2476.

Farmer, J.D., Packard, N., and Perelson, A.S. (1986) The immune system, adaptation and machine learning. Physica D 22:187-204.

Forrest, S., and Hofmeyr, S.A. (2000) Immunology as information processing. In: L.A. Segel and L. Cohen (Eds.), *Design Principles for the Immune System and Other Distributed Autonomous Systems*. Santa Fe Institute Studies in the Sciences of Complexity. Oxford University Press, New York, pp. 361-387.

Hershberg, U.R., Louzoun, Y., Atlan, H., and Solomon, S. (2001) HIV time hierarchy: Winning the war while losing all the battles. Physica A 289:178-190.

Jerne, N.K. (1973) The immune system. Sci. Am. 229:52-60.

Jerne, N.K. (1974) Towards a network theory of the immune system. Ann. Immunol. 125C:373-389.

Kaufman, M., Urbain, J., and Thomas, R. (1985) Towards a logical analysis of the immune response. J. Theor. Biol. 114:527-561.

Kitano, H. (2001) System biology: Toward system-level understanding of biological system. In: H. Kitano (Ed.), *Foundations of System Biology*. MIT Press, Cambridge, MA, pp. 1-36.

Melief, C.J., Toes, R.E., Medema, J.P., Van der Burg, S.H., Ossendorp, F., and Offringa, R. (2000) Strategies for immunotherapy of cancer. Adv. Immunol. 75:235-282.Morpurgo, D., Serenthà, R., Seiden, P.E., and Celada, F. (1995) Modelling thymic functions in a cellular automaton. Int. Immunol. 7:505-516.

Mosier, D., and Sieburg, H.B. (1994) Macrophage-tropic HIV: Critical for AIDS pathogenesis? Immunol. Today 15:332-339.

Pandey, R.B., and Stauffer, D. (1989) Immune response via interacting three dimensional network of cellular automata. J. Phys. (Paris) 50:1-12.Pandey, R.B., and Stauffer, D. (1990) Metastability with probabilistic cellular automata in an HIV infection. J. Stat. Phys. 61:235-240.

Pandey, R.B. (1991) Cellular automata approach to interacting cellular network models for the dynamics of cell population in an early HIV infection. Physica A 179:442-470.

Pantaleo, G., Graziosi, C., and Fauci, A.S. (1993) New concepts in the immunopathogenesis of human immunodeficiency virus infection. N. Engl. J. Med. 328:327-335.

Perelson, A.S., and Oster, G.F. (1979) Theoretical studies on clonal selection: Minimal antibody repertoire size and reliability of self-nonself discrimination. J. Theor. Biol. 81:645-670.

Perelson, A.S. (Ed.) (1988a) *Theoretical Immunology*, *Part II*. Addison Wesley, Redwood City.

Perelson, A.S. (Ed.) (1988b) *Theoretical Immunology, Part I*. Addison Wesley, Redwood City.

Perelson, A.S., and Weisbuch, G.I. (1997) Immunology for physicists. Rev. Mod.Phys. 69:1219-1267.

Perelson, A.S., and Nelson, P.W. (1999) Mathematical analysis of HIV-1 dynamics in vivo. SIAM Rev. 41:3-44.Preziosi, L. (Ed.) (2003) *Cancer Modelling and Simulation*. CRC Press , London.

Ruskin, H.J., Pandey, R.B., and Liu, Y. (2002) Viral load and stochastic mutation in a Monte Carlo simulation of HIV. Physica A 311:213-220.

Segel, L.A., and Perelson, A.S. (1991) Exploiting the diversity of time scales in the immune system: A B-cell antibody model. J. Stat. Phys. 63:1113-1131.

Sieburg, H.B., McCutchan, J.A., Clay, O., Caballero, L., and Ostlund, J.J. (1990) Stimulation of HIV-infection in artificial immune systems. Physica D 45:208-227.

Stauffer, D. (1989) Immunologically motivated cellular automata. In: A. Pires, D.P. Landau, and H.J. Herrmann (Eds.), *Computational Physics and Cellular Automata*. World Scientific, Singapore, pp. 89-97.

Thomè, T., and Drugowich de Felício, J.R. (1996) Probabilistic cellular automaton describing a biological immune system. Phys. Rev. E 53:3976-3981.

Varthakavi, V., Smith, R.M., Deng, H., Sun, R., and Spearman, P. (2002) Human immunodeficiency virus type-1 activates lytic cycle replication of Kaposi's sarcoma-associated herpesvirus through induction of KSHV Rta. Virology 297:270-280.

Weisbuch, G.I., and Atlan, H. (1988) Control of the immune response J. Phys. A 21:189-192.

Zorzenon dos Santos, R.M. (1999) Immune responses: Getting close to experimental results with cellular automata models. In: D. Stauffer (Ed.), *Annual Review of Computational Physics VI*. World Scientific, Singapore, pp.159-202.

Zorzenon dos Santos, R.M., and Coutinho, S.C. (2001) The dynamics of the HIV infection: Acellular automata approach. Phys. Rev. Lett. 87:168102-168114.

Chapter 9

Simulation of HIV-1 Molecular Evolution in Response to Chemokine Coreceptors and Antibodies

Jack da Silva

School of Molecular and Biomedical Science, The University of Adelaide, Adelaide, SA5005, Australia, jack.dasilva@adelaide.edu.au

Abstract. The form of the neutralizing antibody response to human immunodeficiency virus type 1 (HIV-1) and the evolutionary response by the virus are poorly understood. In order for a virus particle (virion) to infect a cell, exterior envelope glycoprotein (gp120) molecules on the virion's surface must interact with receptors on the cell's surface. Antibodies that bind to gp120 may neutralize a virion by interfering with these interactions. Therefore, gp120 is expected to evolve in response to selection by both cell-surface receptors and antibodies. The rate of such adaptation and the constraints imposed by a response to one selective force on the response to the other are unknown. Here, I describe a simulation modeling approach to these problems. The population of viral genomes infecting a single patient is represented by the intensely studied third variable (V3) region of gp120, the main determinant of which chemokine coreceptor a virion uses to enter a cell, and an important target of neutralizing antibodies. Mutation and recombination are applied by realistically simulating the viral replication cycle. Selection by chemokine coreceptors is simulated by taking advantage of the fact that mean site-specific amino acid frequencies are measures of the site-specific marginal fitnesses of amino acids in relation to coreceptor interactions. Selection by antibodies is imposed by simulating the affinity maturation of B-cell lineages that produce neutralizing antibodies to HIV-1 V3. These simulations make clear predictions about the functional cost of adaptation to antibody surveillance, which may help explain the pattern of chemokine coreceptor usage by HIV-1.

9.1 Introduction

Understanding the immunology and evolution of infectious disease is not only a fundamental requirement of the successful treatment and prevention of disease, it also provides an exceptional opportunity to study adaptive evolution at the molecular level (Frank 2002). Arguably, the main barrier to the development of an effective vaccine against infection by human immunodeficiency virus type 1 (HIV-1) is our lack of detailed understanding of the process by which the virus adapts to immune surveillance. We lack an understanding of how viral mutation, recombination, cell superinfection, and protein structural and functional constraints affect HIV's

adaptation to the cell-mediated and humoral branches of the immune system. We also lack detailed understanding of the dynamics of immune responses to HIV.

Here, I describe a simulation approach to understanding and predicting the adaptation of HIV to immune surveillance at the molecular genetic level. Published estimates of fundamental population genetic parameters for HIV-1 allow the realistic simulation of HIV intrapatient evolution in the absence of selection. Selection is more difficult to simulate realistically at the molecular level, however, because it requires knowledge of the effects of all genetic changes on replication rate within a given environment. For example, to simulate selection at the protein level would require knowledge of the fitness effect of each amino acid at each site of a protein, or protein region, and the effects on fitness of interactions among amino acids at different sites. This information is not available for any protein region. However, in the case of selection by host cell receptors, fitness effects can be modeled easily for HIV-1 because the mean site-specific frequency of an amino acid is positively correlated with its effect on fitness (da Silva 2006a). For the special case of immune selection, in which the environment coevolves antagonistically with the virus, knowledge is required of the dynamics of the immune response as well as the targeted protein region (epitope).

The evolution of a viral population infecting a single patient is examined by focusing on the intensely studied third variable region (V3) of the HIV-1 exterior envelope glycoprotein (gp120). I begin by providing background information on the HIV replication cycle and the special role of V3 in interactions with cell-surface receptors and neutralizing antibodies. Then I briefly review what little is known about the dynamics of the humoral response to HIV and the evolutionary response of the virus. This is followed by a description of the model, focusing on the methods used to estimate viral fitness. Finally, results are presented from simulations that explore the effects of coreceptor selection, antibody selection, and the interaction of these on viral adaptation.

9.2 The HIV Replication Cycle

HIV virus particles (virions) contain two single-stranded copies of their RNA genome, which is approximately 9.5 kilobases long and includes nine protein-coding genes. In order to replicate its genome a virion must infect a cell, reverse transcribe its genome into DNA, and integrate the DNA copy into the host's genome. The first step, infection, requires that gp120, which is on the virion's surface, interact with protein receptors on the cell surface. Typically, gp120 binds to a CD4 receptor molecule, which causes conformational changes to gp120 that allow it to then bind to either of two chemokine receptors: CCR5 or CXCR (Coffin 1999). As a result of these interactions, the primary targets of infection are $CD4^+$ T cells expressing either CCR5 or CXCR4. The second major step in the replication cycle is the reverse transcription of the viral genome by the viral enzyme reverse transcriptase. Reverse transcription of the viral genome is error prone and lacks proofreading, resulting in the high mutation rate characteristic of retroviruses ($\sim 10^{-5}$ mutations per nucleotide per replication cycle (Mansky and Temin 1995)). Reverse transcriptase also jumps

frequently from one RNA template to the other during reverse transcription, producing a DNA copy that is a recombinant of the two RNA copies of the genome. Such recombination has recently been reported to occur at a rate two orders of magnitude higher than the mutation rate ($\sim 10^{-3}$ crossovers per nucleotide per replication cycle (Levy, Aldrovandi, Kutsch, and Shaw 2004)). The end result is a double-stranded DNA copy of the viral genome that is then integrated into the host genome. This provirus, as the integrated viral genome is called, is eventually transcribed to RNA, and the RNA translated to protein, by cellular enzymes. Pairs of newly transcribed strands of viral RNA are packaged into virions formed by newly translated viral proteins, and these are budded from the host cell as free virions. The entire cycle takes about 2 days when an activated T-cell is infected (Perelson, Neuman, Markovitz, Leonard, and Ho 1996).

9.2.1 The V3 Loop

A virion's interaction with cell-surface chemokine coreceptors during the infection process is largely determined by the V3 loop of gp120 on the virion surface (Sharon, Kessler, Levy, Zoller-Pazner, Gorlach, and Anglister 2003). V3 is a loop of 35 amino acids (typically) defined by a disulfide bond between terminal cysteine residues. The amino acid sequence of V3 appears to be the primary determinant of which chemokine coreceptor, CCR5 or CXCR4, is used to enter a cell, or whether both coreceptors may be used. Hence, virions may be classified by their chemokine coreceptor-utilization phenotype as exclusively CCR5-utilizing (R5), exclusively CXCR4-utilizing (X4), or dual tropic (R5X4). This aspect of a virion's phenotype is important because it determines which cells a virion may infect and because the X4 phenotype has been linked to increased disease pathogenesis (Mosier 2000).

V3 is also a target of neutralizing antibodies; antibodies that bind to V3 may interfere with chemokine coreceptor interactions. Twenty-two monoclonal antibodies that target V3 and neutralize primary HIV-1 isolates have been isolated from humans infected with HIV-1 subtype B, the main North American and European subtype. These antibodies are listed in the HIV Molecular Immunology Database. These antibodies recognize linear epitopes from the central region of V3, and amino acid changes in this region have been implicated in escape from neutralization (McKeating, Gow, Goudsmit, Pearl, Mulder, and Weiss 1989; McKnight, Weiss, Shotton, Takeuchi, Hoshino, and Clapham 1995; Yoshida, Nakamura, and Ohno 1997). Other studies suggest that polyclonal antibodies in human sera that are capable of neutralizing HIV-1 target conformational V3 epitopes (Gorny, Williams, Volsky, Revesz, Cohen, Polonis, Honnen, Kayman, Krachmarov, Pinter, and Zolla-Pazner 2002). Unfortunately, these epitopes have not been described.

9.2.2 The Neutralizing Antibody Response and the HIV-1 Adaptive Response

Recent studies have laid to rest any doubt that there is a strong neutralizing antibody response to HIV-1 and that the virus evolves in response to the resulting selection (Wei, Decker, Wang, Hui, Kappes, Wu, Salazar-Gonzalez, Salazar, Kilby, Saag,

Komarova, Nowak, Hahn, Kwong, and Shaw 2003; Richman, Wrin, Little, and Petropoulos 2003). However, HIV-1-specific antibodies are not detected until 20 days after the onset of symptoms (Wei et al. 2003). Assuming that the onset of symptoms coincides with peak viremia, about 6 weeks after infection, then HIV-1-specific antibodies are not detected until approximately 2 to 3 months after infection. Antibodies capable of neutralizing HIV-1 are detected approximately 2.5 months after peak viremia, or 4 months after infection (Wei et al. 2003; Richman et al. 2003). This initial neutralizing response is followed by a turnover of the viral population that results in resistant virus. A new neutralizing response is then stimulated and followed by viral turnover. The viral turnover may occur in as little as 2.5 to 3 months (Wei et al. 2003) and the interval between neutralizing antibody responses varies between 3 and 10 months (Richman et al. 2003).

The mode of neutralization escape by the virus appears to involve amino acid changes both within and outside epitopes. In addition to amino acid changes in the central region of V3, other changes associated with resistance involve glycosylation motifs within or outside V3 (Wei et al. 2003). These motifs are binding sites for glycans that somehow interfere with antibody binding.

9.3 The Model

The basic data structure of the model represents a cell that may be infected by zero or more proviruses, each represented by its V3 DNA sequence. A vector of such cells represents the population of target cells in a patient. The population sizes of target cells and proviruses are held constant for simplicity and to simulate the quasi-stable state during the 5 or so years of the asymptomatic, chronic stage of infection in untreated patients. Simulation flow follows the HIV replication cycle. An infected cell forms virions by pairing its proviruses at random. The fitness of a virion, which depends on its V3 amino acid sequences, determines its probability of escaping neutralization and integrating its genome into that of a new host cell. In reality, a virion has many gp120 molecules on its surface, each with a V3 loop, and these molecules are potentially translated from any of the proviruses sharing a host cell. In the model, it is assumed that a virion possesses all of the possible V3 loops translated from the proviruses sharing its host cell and that its fitness is equal to that of the V3 loop with the highest fitness. The calculation of fitness from a V3 amino acid sequence is described below. A virion infects a randomly chosen cell, which may already be infected, with a probability equal to its fitness (scaled from 0 to 1). During the reverse transcription step, crossovers between the two copies of the virion's genome and mutation of the final recombinant DNA copy occur probabilistically at specified rates.

9.3.1 Fitness

Fitness is determined by the amino acid sequence of a V3 loop and consists of two components. The functional component of fitness reflects a V3 loop's interaction with chemokine coreceptors and is the probability of infection. The neutralization component of fitness reflects a V3 loop's interaction with antibodies and is the probability of escaping neutralization. Total fitness is the product of these probabilities.

9.3.1.1 The Functional Component of Fitness

The functional component of fitness was estimated from V3 amino acid mean site-specific relative frequencies. Site-specific frequencies were calculated for the viral population infecting a single patient and were then averaged across patients. These mean site-specific frequencies are linearly related to the site-specific marginal fitness effects of the amino acids, measured as relative virion infectivity, and are equivalent to the relative marginal fitnesses of the amino acids when scaled from 0 to 1 (da Silva 2006a). The marginal relative fitness of a particular V3 loop is simply the product of the 35 site-specific marginal relative fitnesses of its 35 amino acid sites.

V3 amino acid site-specific frequencies were calculated from sequences associated with a patient and a chemokine coreceptor-usage phenotype in the HIV Sequence Database. Only sequences from subtype B, the most studied and sequenced HIV-1 subtype, were used. Mean site-specific frequencies vary among chemokine coreceptor-usage phenotypes (Fig. 1), as would be expected if these frequencies are measures of site-specific marginal fitnesses with respect to the interaction between V3 and a chemokine coreceptor. Therefore, the fitness of a particular V3 loop was calculated from the site-specific relative fitnesses of each phenotype and the loop was assigned the phenotype that gave the highest fitness.

To explore the effect of the strength of selection by a coreceptor on viral adaptation, amino acid site-specific selection coefficients were multiplied by a scaling constant ranging from 0 to 1. Selection coefficients were calculated by subtracting the relative fitness of the amino acid under consideration from that of the most frequent amino acid at the same site, which is assigned a relative fitness of 1. A higher scaling constant corresponds to stronger selection.

This method of estimating fitness assumes no fitness interactions (epistasis) among amino acids at different sites. However, the amino acids of the V3 loop are known to co-vary between pairs of sites (Korber, Farber, Wolpert, and Lapedes 1993), indicating epistatic interactions between sites. To take this into account, the fitness of a sequence in which only one member of a pair of associated amino acids was present was reduced. Fitness was reduced by a factor equal to the ratio of the expected mean frequency of a covarying pair of amino acids (the product of their individual mean frequencies) to the observed mean frequency of the pair (which is higher).

9.3.1.2 The Neutralization Component of Fitness

To calculate the neutralization component of fitness requires a model of the neutralizing antibody response to HIV-1 that describes how the probability of neutralization of a virion changes over time. In the absence of any directly applicable model, I have made the reasonable assumptions that antibody affinity maturation increases logistially (Rundell, DeCarlo, Hogenesch, and Doerschuk 1998) and that neutralization is

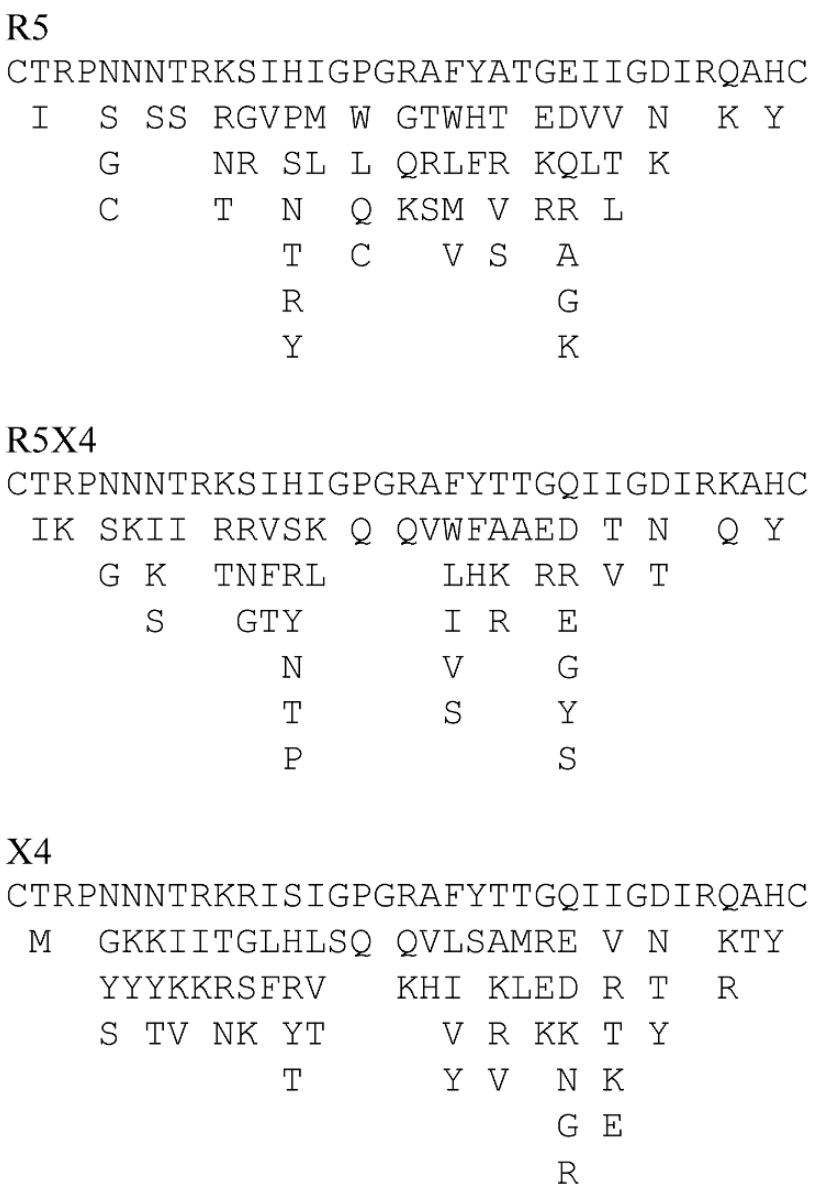


Fig. 1. HIV-1 subtype B V3 amino acids of chemokine coreceptor-usage phenotypes in order of decreasing site-specific frequency. Frequencies were calculated from 181 sequences from 36 patients for R5, 43 sequences from 12 patients for R5X4, and 83 sequences from 10 patients for X4.

limited by antibody affinity. The well-known logistic growth equation is used to model affinity maturation in response to an epitope in the viral population. The affinity of an antibody to an epitope is

$$N_t = K / \left[1 + \left(K / N_0 - 1 \right) e^{-rt} \right] \quad (1)$$

where t is the number of viral generations of stimulation (maturity), K is the maximum affinity (ranging from 0 to 1), N_0 is the affinity at $t = 0$, and r is the intrinsic rate of increase of affinity (rate of increase when N_t is small). Once the production of an antibody has been stimulated (see below), t is incremented for each viral generation in which the targeted epitope is present. The probability of neutralization of a virion that carries the epitope targeted by the antibody is N_t, and the probability of neutralization escape is $1 - N_t$. The neutralization component of fitness for a virion then is the product of the neutralization escape probabilities associated with each antibody that targets epitopes contained in the virion's V3 sequences.

The V3 linear epitopes of known monoclonal neutralizing antibodies produced during HIV-1 subtype B infection of humans are shown in Table 1. Before any of these antibodies can affect virion fitness, their initial production must be stimulated.

Table 1. Monoclonal neutralizing antibodies that target V3, generated during HIV-1 subtype B infection of humans, and their epitopes. A period indicates any amino acid. Data are from the HIV Molecular Immunology Database.

Monoclonal antibody	V3 epitope
412-D	RKRIHIGPGRAFYTT
DO142-10	KRIHIGPGRAFYTT
391/95-D	KRIHIGPGRAFY
41148D	KRIHIGP
311-11-D	KRIHIGP
257-D	KRIHI
908-D	KSITKG
782-D	KSITKG
838-D	KSITK
MN215	RIHIGPGRAFYTTKN
19b	I....G..FY.T
419-D	IHIGPGR
504-D	IHIGPGR
453-D	IHIGPGR
4117C	I.IGPGR
418-D	HIGPGRA
386-D	HIGPGR
268-D	HIGPGR
537-D	IGPGR
447-52D	GP.R
N70-1.9b	PGRAFY
694/98-D	GRAF

In the model, initial production of an antibody is stimulated by the presence of its epitope in the viral population at a frequency above a specified threshold. This frequency threshold determines the narrowness of the antibody response. For example, a high threshold frequency of 0.9 corresponds to a narrow response. Once an antibody's initial production is stimulated, the presence of its epitope at any frequency is sufficient to increase its maturity (t in Eq. (1)). If the individual frequencies of epitopes for two or more antibodies exceed the stimulation threshold in the same viral generation, then only one antibody, chosen at random, has its initial production stimulated. If a V3 sequence carries the epitope for an antibody whose production has been stimulated, then none of the epitopes from that sequence are available to stimulate the initial production of any other antibody.

9.4 Simulations

9.4.1 The Simulation Environment

The simulation program was written in FORTRAN 90 and parallelized using the Message Passing Interface library. Random numbers were generated using the Scalable Parallel Random Number Generators Library (SPRNG). Simulation replicates were run in parallel on an IBM eServer 1350 Linux cluster. Details of the simulation methods used are given in da Silva (2006b).

9.4.2 Adaptation to Coreceptors

Initial simulations were carried out to demonstrate viral adaptation to chemokine coreceptors. To study adaptation to CCR5, the viral population was initialized with a suboptimal R5 V3 sequence and allowed to evolve over several hundred viral generations in the absence of antibody selection. Figure 2 shows the increase in population mean fitness, determined entirely by the functional component of fitness, for various selection coefficient scaling constants. Fitness increased as the suboptimal sequence evolved toward the optimal sequence for the phenotype, that is, as the viral population adapted to the CCR5 chemokine coreceptor.

Figure 3 compares simulated V3 sequences at the end of one simulation replicate with the suboptimal R5 V3 sequence of the initial population and the optimal R5 sequence. The initial sequence differed from the optimal sequence at three sites, and the simulated sequences evolved to match the optimal sequence at two of these sites. At the third, unmatched, site the difference between the initial and optimal sequences involved amino acids with nearly identical site-specific frequencies, thereby imposing only weak selection for amino acid replacement.

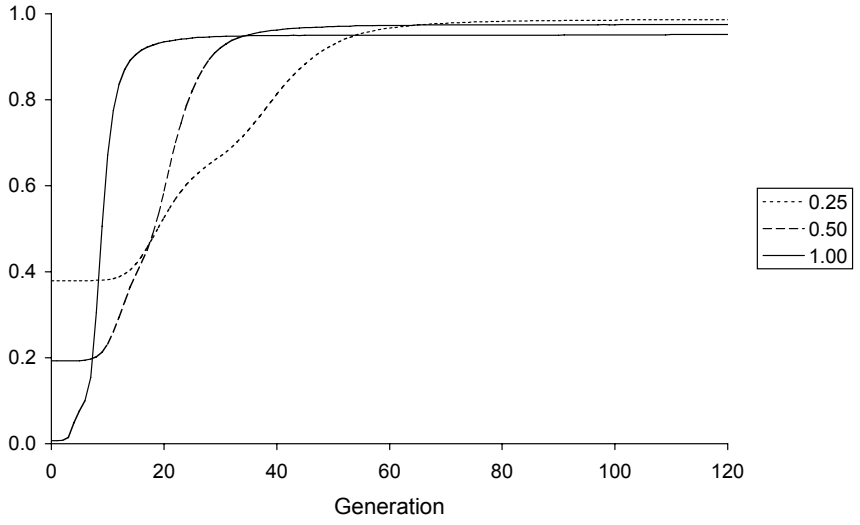


Fig. 2. The change in viral population mean fitness over viral generations in the absence of antibody selection. Populations were initialized with a suboptimal R5 V3 sequence. Plots for different selection coefficient scaling constants (strengths of coreceptor selection) are shown. Lines are means of 20 replicate simulations. The following parameter values were used: 10^5 cells, 3 x 10^5 virions, 10^{-5} mutation per nucleotide per cycle, 10^{-3} crossover per nucleotide per cycle, 100 virions budded per cell, and 0.03 probability of viral genome integration.

Simulations were also run with populations initialized with optimal, or near-optimal, R5, X4R5, or X4 V3 sequences (Fig. 4). The optimal R5 V3 sequence has a fitness of one in the absence of antibody selection and does not change over the several hundred generations simulated. The X4R5 V3 sequence is near-optimal in the sense that it contains the most common amino acid for its phenotype at all sites but one; this was the sequence in the Swiss-Prot protein database most similar to the optimal sequence. The relative fitness of this near-optimal X4R5 sequence is 0.37 because of the difference from the optimal sequence and because it violates several amino acid covariation patterns (See 9.3.1.1). Selection of covarying amino acids appears to have changed the sequence phenotype to R5, followed by evolution to the optimal R5 V3 sequence. The optimal X4 sequence also has a fitness of less than one because it violates several amino acid covariation patterns. However, selection of covarying amino acids, which increased population mean fitness, did not change the phenotype. X4 sequences attained a mean fitness of only 0.64 after 500 generations of selection.

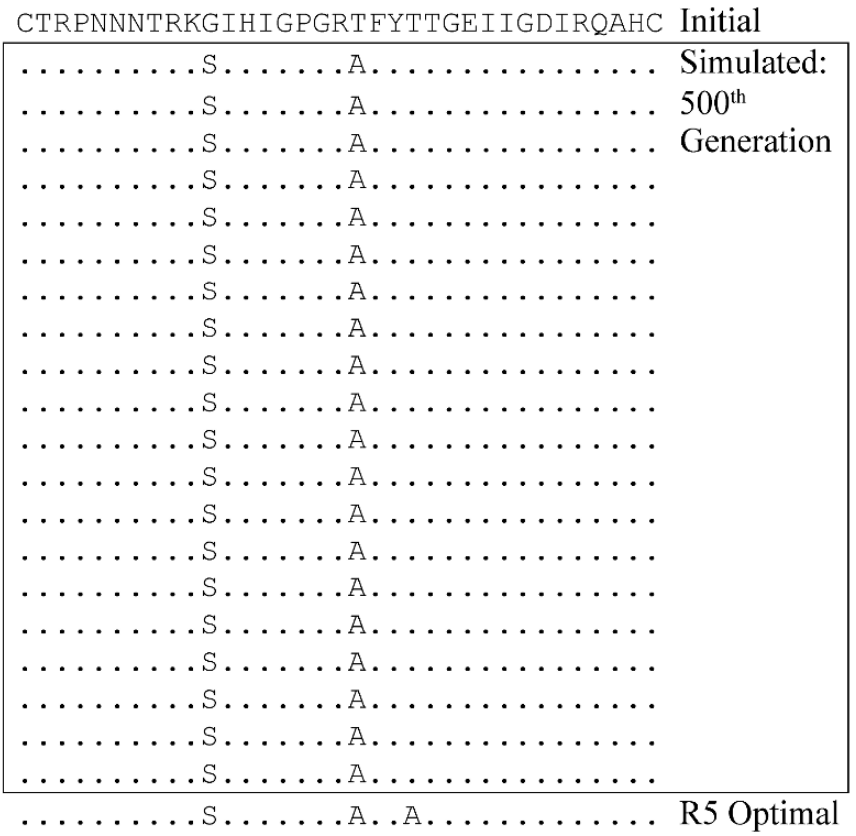


Fig. 3. Twenty randomly chosen V3 sequences from one replicate simulation with a selection coefficient scaling constant of 0.25. Other parameter values are as described in Fig. 2. Periods indicate identity with the initial sequence.

9.4.3 Adaptation to Antibody Surveillance

Parameter values for the logistic growth equation (Eq. (1)), used to model affinity maturation, were chosen to give realistic antibody response dynamics. The affinity of antibody to hapten increases up to 10,000-fold (Wedemayer, Patten, Wang, Schultz, and Stevens 1997), but increases only up to 1000-fold for a protein immunogen (lysozyme) (Cauerhff, Goldbaum, and Braden 2004), and affinity increases of only around 100-fold have been reported in response to HIV-1 gp120 (Toran, Kremer, Sanchez-Pulido, de Alboran, del Real, Llorente, Valencia, de Mon, and Martinez 1999). Therefore, it was assumed that a 1000-fold increase in affinity is the maximum in response to a protein immunogen. Setting the maximum affinity (K) to 1.0

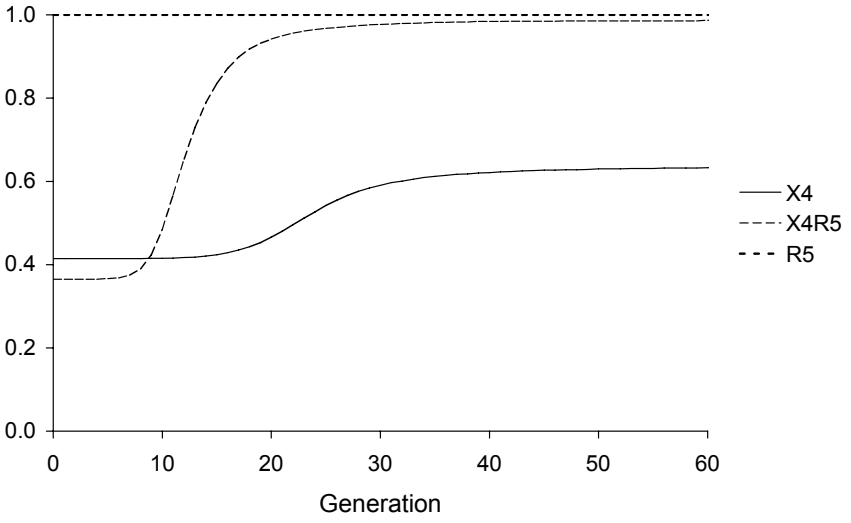


Fig. 4. Changes in viral population mean fitness over viral generations in the absence of antibody selection for populations initialized with optimal (R5) or near-optimal (X4 and X4R5) V3 sequences. Lines are means of 20 replicate simulations. Simulations were run with a selection coefficient scaling factor of 0.25. Other parameter values are as described in Fig. 2.

and the initial affinity (N_0) to 10^{-3} allows a 1000-fold increase. The value for the remaining free parameter, the intrinsic rate of increase of affinity (r), was set to 0.25, which gives viral population turnovers at intervals of about 3 months. The threshold epitope frequency in the viral population at which antibody initial production is stimulated was set to 0.1, 0.5, and 0.9 in separate simulations.

9.4.3.1 Effect of the Stimulation Threshold

Figure 5 shows changes in the components of fitness over time starting with a viral population of optimal R5 V3 sequences and with antibody selection. With antibody selection, the neutralization component of fitness cycled as antibody affinity increased and then new, neutralization-resistant viral variants progressively replaced older, sensitive variants. Increasing the antibody production stimulation threshold increased the period of the neutralization fitness cycles, but in every case the virus escaped humoral control. Note that with each increase in the neutralization component of fitness, which corresponds to viral escape from the circulating antibody, the functional component of fitness decreased. This shows a clear trade-off between fitness components: adaptation to antibody surveillance necessarily reduced adaptation to the chemokine coreceptor. The final value of the functional component of fitness in each simulation was about 0.6 (compared with an initial value of 1.0).

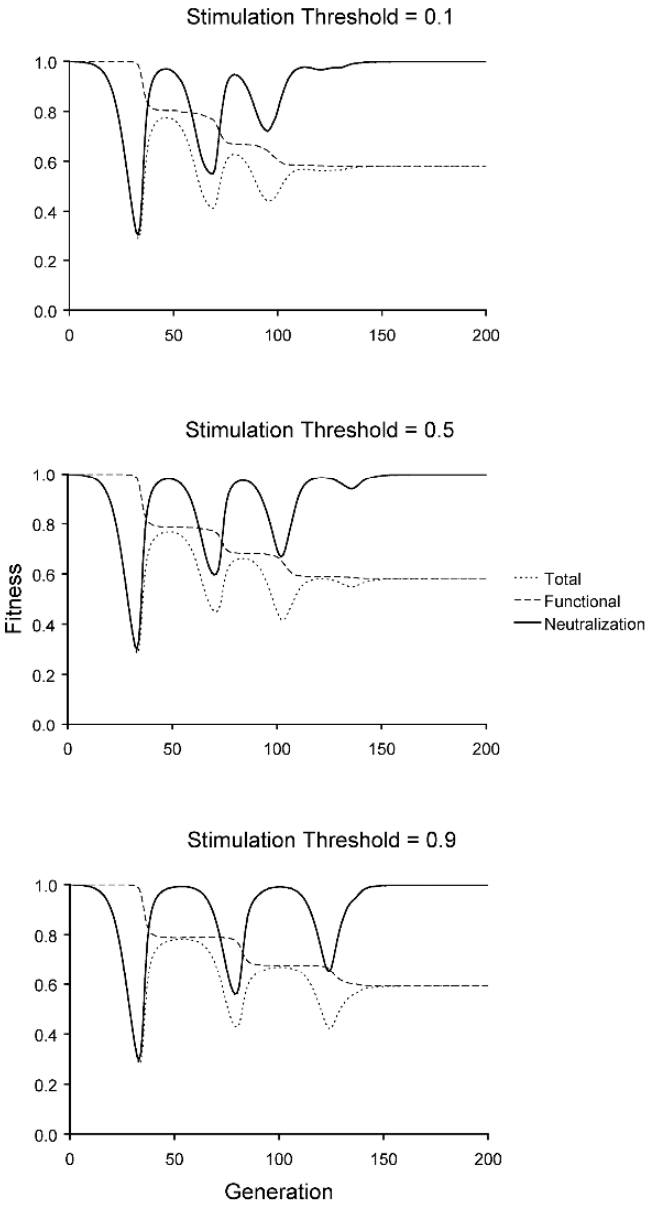


Fig. 5. Viral population mean fitness components plotted over viral generations for different antibody stimulation frequency thresholds. Lines are means of 20 replicate simulations. Simulations were run with antibody affinity maturation parameters $N_0 = 10^{-3}$, $K = 1$, $r = 0.25$, and a coreceptor selection coefficient scaling factor of 0.25. Remaining parameter values are as described in Fig. 2.

Therefore, adaptation to antibody surveillance under these simulation conditions reduced the functional component of fitness by 40%. Also note that a narrower antibody response (higher stimulation threshold) increased the duration of high V3 function.

Figure 6 shows V3 sequences from one replicate simulation after 500 generations with an antibody production stimulation frequency threshold of 0.1. Amino acid replacements, mainly at two sites, produced sequences that are neutralization resistant but suboptimal with regard to their interaction with CCR5.

9.4.3.2 Effect of the Strength of Coreceptor Selection

The effect of the strength of selection by the chemokine coreceptor was investigated by rerunning simulations with different coreceptor selection coefficient scaling factors (Fig.7). These simulations were run with an antibody production stimulation frequency threshold of 0.1 and should be compared with the corresponding plot in Fig. 5. A coreceptor selection coefficient scaling factor of 0.1, instead of 0.25, as was used previously (Fig. 5), did not alter appreciably the dynamics of the neutralization

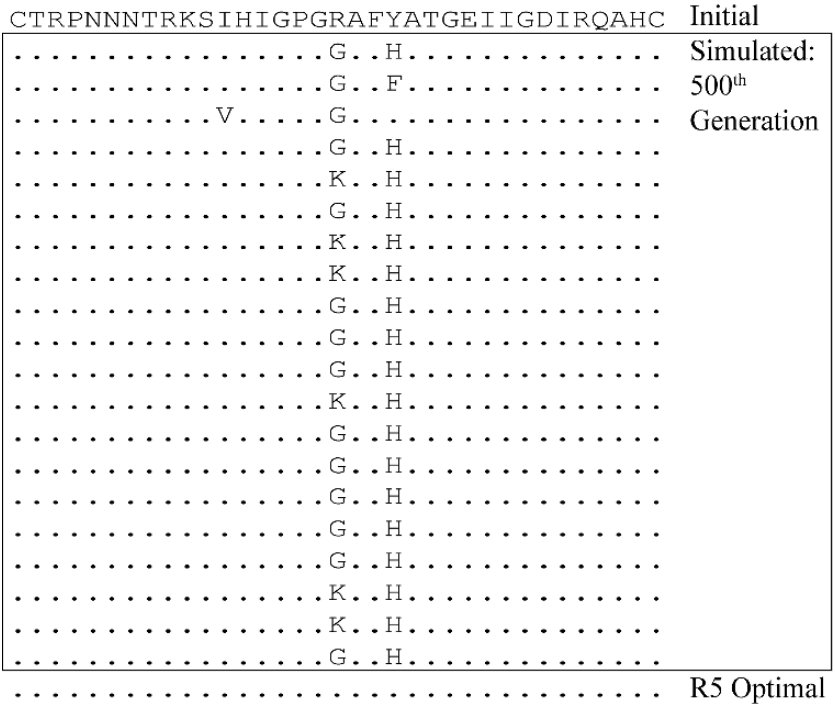


Fig. 6. Twenty randomly chosen V3 sequences from one replicate simulation with an antibody production stimulation threshold of 0.1. Other parameter values are as described in Fig. 5. Periods indicate identity with the initial sequence.

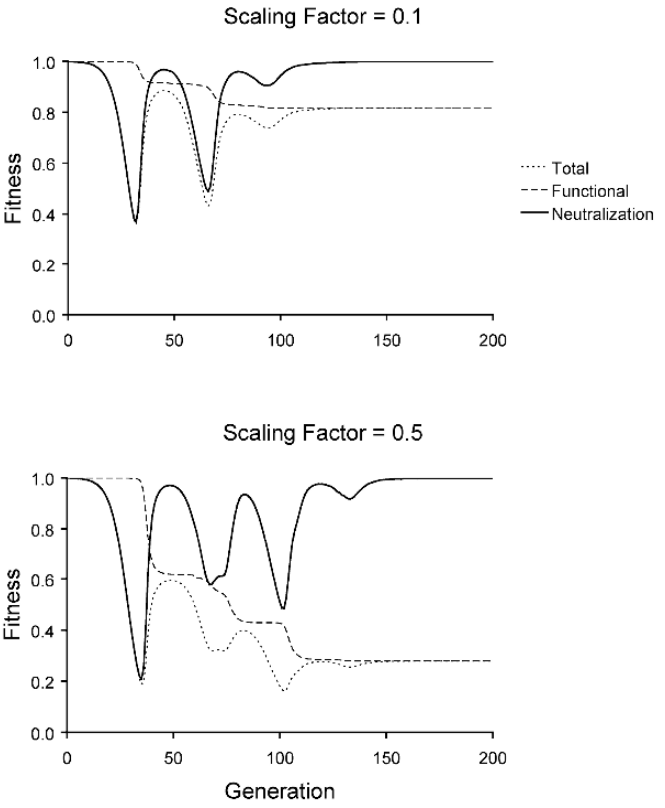


Fig. 7. Viral population mean fitness plotted over viral generations for simulations with different coreceptor selection coefficient scaling factors. Lines are means of 20 replicate simulations. The antibody production stimulation threshold is 0.1. Other parameter values are as described in Fig. 5.

component of fitness, but, as expected, resulted in a final functional component of fitness that is higher than before, around 0.8 (compared to 0.6). Similarly, a scaling factor of 0.5 also did not alter appreciably the dynamics of the neutralization component, but, as expected, resulted in a final functional component of fitness that is lower than before, around 0.3. Therefore, regardless of the strength of coreceptor selection, there was a functional cost of adaptation to antibody surveillance, but the magnitude of the cost depended on the strength of coreceptor selection. Weaker selection by a coreceptor resulted in a lower cost of adaptation to antibody surveillance.

9.4.3.3 Effect of the Coreceptor Utilization Phenotype

The evolutionary response of the virus to antibody selection was also investigated for viral populations initialized with V3 phenotypes other than R5 (Fig. 8). The simulation conditions used were otherwise the same as described in Fig. 5 and with an antibody stimulation threshold frequency of 0.1. The dynamics of the neutralization component of fitness were altered very little, given that the dynamics of the functional component of fitness were substantially different from when the population was initialized with an optimal R5 sequence. As in previous simulations, the functional component of fitness decreased with adaptation to antibody surveillance. For populations initialized with a near-optimal X4R5 V3 sequence, the final value for the functional component of fitness, 0.58, is near that observed for a population initialized with an optimal R5 V3 sequence under otherwise identical simulation conditions. This is because the population evolved to R5 phenotype within the first 30

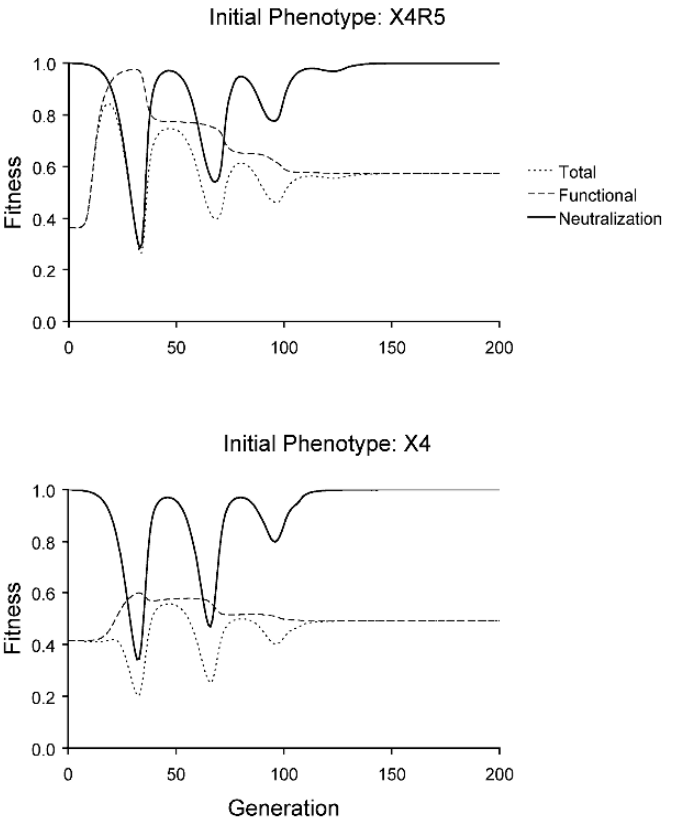


Figure 8. Viral population mean fitness plotted over viral generations for simulations initialized with X4R5 or X4 phenotype V3 sequences. Lines are means of 20 replicate simulations. The antibody production stimulation threshold is 0.1. Other parameter values are as described in Fig. 5.

generations, as described above (Section 9.4.2). For populations initialized with an optimal X4 V3 sequence, the final value of the functional component of fitness was about 0.5. Therefore, after escaping humoral control, X4 virus remained less well adapted to its coreceptor than R5 virus was to its coreceptor (0.6). However, this difference is considerably smaller than the difference between the maximum values of the functional fitness components of the two phenotypes (1.0 and 0.6 for R5 and X4, respectively).

9.5 Conclusions and Future Directions

The main obstacle to simulating selection at the molecular genetic level is a lack of knowledge of the fitness effects of individual nucleotide and amino acid changes. For HIV-1 V3, this obstacle is easily surmounted because, in the context of selection by coreceptors, mean site-specific amino acid frequencies are measures of amino acid site-specific marginal relative fitnesses (da Silva 2006a). Using this approach to model selection by coreceptors on V3 produced plausible evolutionary dynamics.

Selection by neutralizing antibodies may be simulated for defined epitopes by assuming a plausible model of affinity maturation. Affinity maturation was modeled as a logistic increase in affinity. This aspect of the model could be made more realistic by incorporating knowledge of the actual dynamics of affinity maturation for specific antibodies. Inclusion of changes to antibody titer may also add to the realism. This could be accomplished by assuming logistic growth of antibody numbers, scaled from zero to one, and using the product of the antibody titer and affinity maturation as a measure of neutralization capacity. A more difficult problem in simulating the neutralizing antibody response to HIV-1 may be the definition of conformational epitopes. The current model uses linear V3 epitopes of monoclonal antibodies known to neutralize primary isolates. However, it is apparent that the most potent neutralization *in vivo* is by antibodies with conformational V3 epitopes (Gorny et al. 2002). Unfortunately, these epitopes have yet to be described.

With the current model, a maximum of four consecutive antibody responses to newly emerged (or initialized) viral variants were each followed by viral escape before the final escape from humoral control. A cyclical pattern of antibody response and viral escape and final loss of humoral control is also observed in patients (Wei et al. 2003; Richman et al. 2003). In the model, this occurred regardless of the stimulation threshold for initial antibody production, the strength of coreceptor selection, or the coreceptor utilization phenotype of the virus. The other consistent result is a decrease in V3 function associated with escape from neutralization. This decrease in function corresponds to a decrease in viral infectivity. Such a pattern has not been reported from either *in vivo* or *ex vivo* studies of HIV. However, this predicted trade-off between fitness components might help explain the relatively low viral population size during the nonsymptomatic, chronic phase of infection (Coffin 1999). As was expected, the decrease in V3 function was greater with stronger coreceptor selection. Of more interest, however, is that after escape from humoral control, the fitness difference between R5 and X4 virus was diminished from about 60% to about 20%. This predicted reduction in the difference

in fitness between these viral phenotypes may allow other factors to tip the balance in favor of X4 virus and help explain why in about 50% of patients the viral phenotype switches from R5 to X4 late in infection (Mosier 2000).

Acknowledgements

I thank the Discipline of Genetics and the School of Molecular and Biomedical Science, University of Adelaide, for support, and the South Australian Partnership for Advanced Computing (SAPAC) for access to high-performance computing resources.

References

Cauerhff, A., Goldbaum, F.A., and Braden, B.C. (2004) Structural mechanism for affinity maturation of an anti-lysozyme antibody. Proc. Natl. Acad. Sci. USA 101:3539-3544.

Coffin, J.M. (1999) In: K.A. Crandall (Ed.), *The Evolution of HIV*. Johns Hopkins University Press, Baltimore, pp. 3-40.

da Silva, J. (2006a) Site-specific amino acid frequency, fitness and the mutational landscape model of adaptation in human immunodeficiency virus type 1. Genetics 174:1689-1694.

da Silva, J. (2006b) In: A.Y. Zomaya (Ed.), *Parallel Computing for Bioinformatics and Computational Biology*. John Wiley & Sons, New York, pp. 29-57.

Frank, S.A. (2002) *Immunology and Evolution of Infectious Disease*. Princeton University Press, Princeton.

Gorny, M.K., Williams, C., Volsky, B., Revesz, K., Cohen, S., Polonis, V.R., Honnen, W.J., Kayman, S.C., Krachmarov, C., Pinter, A., and Zolla-Pazner, S. (2002) Human monoclonal antibodies specific for conformation-sensitive epitopes of V3 neutralize human immunodeficiency virus type 1 primary isolates from various clades. J. Virol. 76:9035-9045.

Korber, B.T., Farber, R.M., Wolpert, D.H., and Lapedes, A.S. (1993) Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelope protein: An information theoretic analysis. Proc. Natl. Acad. Sci. USA 90:7176-7180.

Levy, D.N., Aldrovandi, G.M., Kutsch, O., and Shaw, G.M. (2004) Dynamics of HIV-1 recombination in its natural target cells. Proc. Natl. Acad. Sci. USA 101:4204-4209.

Mansky, L.M., and Temin, H.M. (1995) Lower in vivo mutation rate of human immunodeficiency virus type 1 than that predicted from the fidelity of purified reverse transcriptase. J. Virol. 69:5087-5094.

McKeating, J.A., Gow, J., Goudsmit, J., Pearl, L.H., Mulder, C., and Weiss, R.A. (1989) Characterization of HIV-1 neutralization escape mutants. AIDS 3:777-784.

McKnight, A., Weiss, R.A., Shotton, C., Takeuchi, Y., Hoshino, H., and Clapham, P.R. (1995) Change in tropism upon immune escape by human immunodeficiency virus. J. Virol. 69:3167-3170.

Mosier, D.E. (2000) Virus and target cell evolution in human immunodeficiency virus type 1 infection. Immunol. Res. 21:253-258.

Perelson, A.S., Neumann, A.U., Markowitz, M., Leonard, J.M., and Ho, D.D. (1996) HIV-1 dynamics in vivo: Virion clearance rate, infected cell life-span, and viral generation time. Science 271:1582-1586.

Richman, D.D., Wrin, T., Little, S.J., and Petropoulos, C.J. (2003) Rapid evolution of the neutralizing antibody response to HIV type 1 infection. Proc. Natl. Acad. Sci. USA 100:4144-4149.

Rundell, A., DeCarlo, R., Hogenesch, H., and Doerschuk, P. (1998) The humoral immune response to *Haemophilus influenzae* type b: A mathematical model based on T-zone and germinal center B-cell dynamics. J. Theor. Biol. 194:341-381.

Sharon, M., Kessler, N., Levy, R., Zolla-Pazner, S., Gorlach, M., and Anglister, J. (2003) Alternative conformations of HIV-1 V3 loops mimic beta hairpins in chemokines, suggesting a mechanism for coreceptor selectivity. Structure (Camb.) 11:225-236.

Toran, J.L., Kremer, L., Sanchez-Pulido, L., de Alboran, I.M., del Real, G., Llorente, M., Valencia, A., de Mon, M.A., and Martinez, A.C. (1999) Molecular analysis of HIV-1 gp120 antibody response using isotype IgM and IgG phage display libraries from a long-term non-progressor HIV-1-infected individual. Eur. J. Immunol. 29:2666-2675.

Wedemayer, G.J., Patten, P.A., Wang, L.H., Schultz, P.G., and Stevens, R.C. (1997) Structural insights into the evolution of an antibody combining site. Science 276:1665-1669.

Wei, X., Decker, J.M., Wang, S., Hui, H., Kappes, J.C., Wu, X., Salazar-Gonzalez, J.F., Salazar, M.G., Kilby, J.M., Saag, M.S., Komarova, N.L., Nowak, M.A., Hahn, B.H., Kwong, P.D., and Shaw, G.M. (2003) Antibody neutralization and escape by HIV-1. Nature 422:307-312.

Yoshida, K., Nakamura, M., and Ohno, T. (1997) Mutations of the HIV type 1 V3 loop under selection pressure with neutralizing monoclonal antibody NM-01. AIDS Res. Hum. Retrovir. 13:1283-1290.

Chapter 10

MUTANT MOUSE: *bona fide* Biosimulator for the Functional Annotation of Gene and Genome Networks

Yoichi Gondo

RIKEN Genomic Sciences Center, Functional Genomics Research Group, Population and Quantitative Genomics Team, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan, gondo@gsc.riken.jp

Abstract. The advancements of genomics and genome projects led to the current paradigm that the blueprint of life is depicted in the genome sequences. To decipher the life system, deductive methods have been applied from genome sequences to genes, transcripts, proteins, organelles, cells, tissues, organs, organisms, and populations. As a result we encountered an astronomical scale of complicated molecular and cellular networks in the life system. There is a way, however, to directly connect the function of a single base pair (bp) in genome sequences to the life system by bypassing all the molecular and cellular labyrinths. "MUTANT" provides the ultimate tool as a *bona fide* biosimulator for the functional annotation of gene and genome networks. Genetics, with mutations and mutants, is revealing the life system. Mendel deduced the concept of "gene" from a large dataset of the pea phenome. Snell discovered the mouse *H2* locus by graft rejection that led to the identification and understanding of the major histocompatibility complex. Many other mouse mutants (i.e., *nu*, *scid*, *lpr*, *gld*, *Sl*, and *W)* provided model systems for the functional characterization of key genes in immunological networks. In this context, "reverse genetics" methods have been developed since the 1980s to systematically produce mutant mice carrying a particular gene of interest, for example, transgenic mice, knockout mice, and gene targeting. Recently, more versatile, large-scale, and high-throughput methods such as ENU mutagenesis and insertional mutagenesis are being used to generate mutant mice. This chapter offers a review of the history and current status of mouse mutagenesis and discusses the value of mouse model systems.

10.1 Introduction

Theoretical modeling and computer simulation are two powerful tools to elucidate the mechanism of complex systems. Testable modeling, parameter setting, and empirical input of initial parameter values are some of the key requirements for simulation. The efficacy of simulations is finally evaluated by how closely the simulation output reflects the complex system. The calculation power of computer hardware is another key factor that determines whether the results are obtained within a reasonable period of time.

Since life is a typical complex adaptive system, a major challenge is to depict its composition and dynamism. Apart from the structural design of whole parts, the concrete working plans and protocols in time and space for life to emerge, are all inscribed in the genomic DNA sequences.

10.1.1 Relevancy of Mouse as Simulator

Given that the entire blueprint of life is depicted in the genome, an ideal life simulator must reconstruct all the biology solely from three billion letters of the genomic DNA sequences. At the time of fertilization, the life of the mouse starts and after 4 months it matures and is able to produce the next generation as shown in Fig. 1.

The 4-month "calculation" period appears to be an acceptable time range because the reconstructed life is valid to start, with nothing left to be proven or evaluated. Hence, it is obvious that all the life phenomena are reflected in the mouse model system. The key question, however, is whether the genomic function would be effectively and appropriately elucidated by using the mouse system as a *bona fide* simulator.

In this chapter we discuss recent developments in using mutant mice to directly decode the genomic DNA sequence toward its biological outcome in a genome-wide manner. This approach is particularly effective in deciphering complicated biological systems like neurological behavior and immunological processes at the organism level.

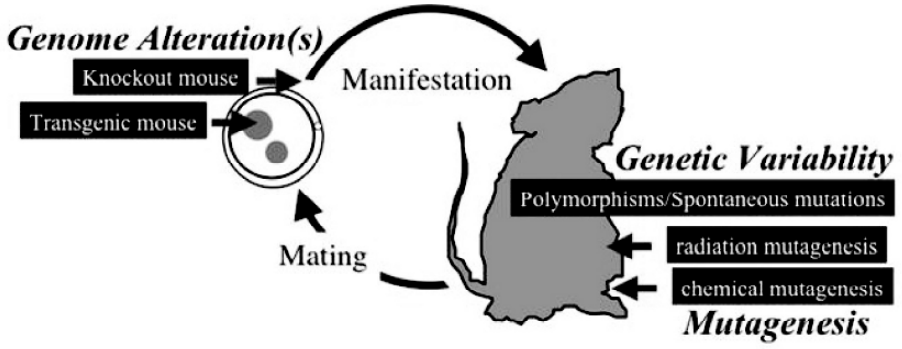


Fig. 1. Life cycle of the mouse with respect to the genomic DNA program and its execution. A part of genomic DNA sequences may be artificially altered by transgenesis or gene-targeting. The differences may have been induced by mutagenesis or they already existed as natural polymorphisms. At the step of mating and fertilization, all the sufficient and necessary information is compiled. All the biological traits are autonomously manifested during the 20-day gestation period and the 3-month of maturation cycle.

10.2 I/O System in Mouse Genetics

The genomic DNA sequences of the mouse genome are products shaped by the course of evolution. To decode the function of genomic DNA sequences, the alteration of the finished product is considered to be INPUT of the *bona fide* biosimulator. There are several ways to alter or introduce new sequences to the genomic sequences (Fig. 1). In natural populations, many polymorphisms accumulated whereas some spontaneous mutations arose recently. Mutations can be artificially induced by genotoxic agents, for instance, X-ray irradiation (Muller 1927) and chemical mutagens (Auerbach and Robson 1946). In classical genetics, polymorphisms and mutations represent a major resource for research studies; however, the approach is usually limited to phenotype-driven analyses of a small number of loci. Even if a genetic alteration is vindicated, it used to be impossible to conclude what kind of genomic sequence change(s) is responsible for the phenotypic consequence. Eventually, in the 1980s with the development of various positional cloning methods and genetic engineering technologies it became feasible to identify the causative genomic sequence change(s).

10.2.1 Reverse Genetics

In the 1980s, new gene-driven approaches were developed to decode the genomic sequence function in the mouse. One method that introduces an artificially designed DNA fragment into the genome of the mouse (Palmiter, Brinster, Hammer, Trumbauer, Rosenfeld, Birnberg, and Evans 1982) enabled the generation of transgenic mice (Fig. 2). Another technology that disrupts or eliminates a specified part of genomic DNA sequences in the mouse (reviewed by Capecchi 1989) is called gene targeting or knockout mouse (Fig. 3). Together, these approaches are defined as "reverse genetics." Both became practicable with the innovations of genetic engineering as well as mouse embryonic technology. Conversely, classical genetics and its phenotype-driven approach is now called "forward genetics".

10.2.1.1 Transgenic Mouse

As shown in Fig. 2, the transgenic mouse allows us to elucidate the gain of function due to a particular DNA sequence that has been designed and constructed *in vitro* by using genetic engineering technology. However, some caution must be taken when interpreting the outcome. For instance, the integration of the exogenous DNA disrupts at least one site in genomic DNA sequences. The expression of endogenous genes at and around the integrated site may be affected by this disruption. In addition, the expression of the integrated DNA sequence often varies depending on the site and mode of integration.

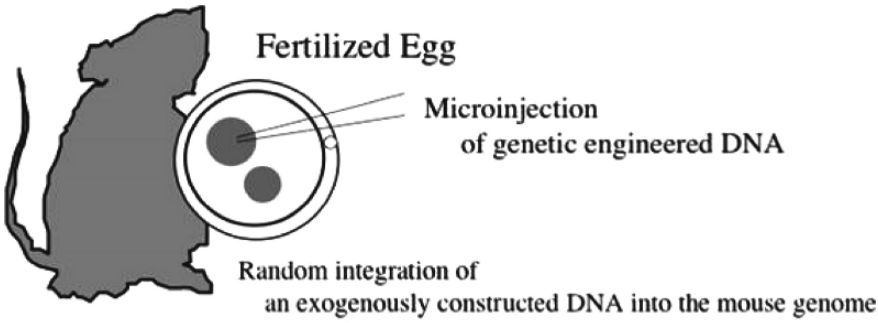


Fig. 2. System of transgenic mouse. An *in vitro* constructed DNA fragment is injected directly into the pronucleus of a fertilized egg. Occasionally the injected DNA is occasionally integrated into the genomic DNA of the fertilized egg. Once integrated, it is delivered to all the cells as if it were a part of the authentic endogenous genome. The integrated DNA sequence is usually inherited stably to the offspring. If any phenotypic changes are observed in the transgenic mouse, it is primarily expected to be a gain of function.

10.2.1.2 Knockout Mouse (Gene Targeting)

The gene targeting method enabled the disruption of any part of genomic sequences with a marker gene in the mouse genome in a site-specific manner (Fig. 3), and the study of the functional loss of the endogenous genomic sequence. Occasionally the disruption might be lethal, thereby hampering the elucidation of the normal function of the sequence.

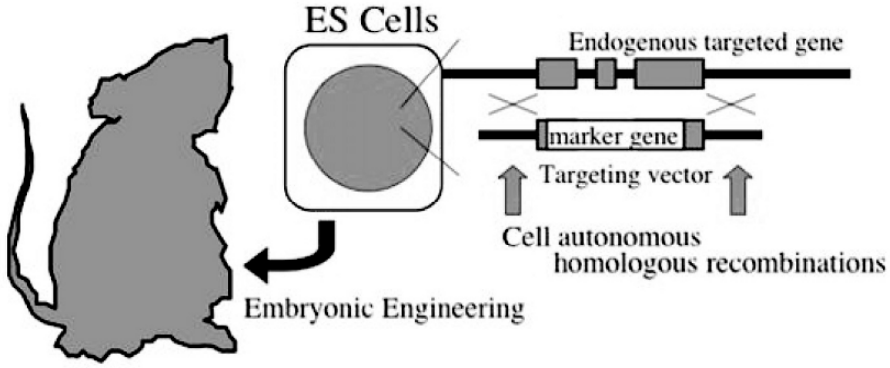


Fig. 3. Knockout mouse system by gene targeting. Homologous recombination occurs even in somatic cells with a low frequency. A targeting vector is constructed with part of the genomic sequences from the gene of the interest and a marker gene: e.g., the neomycin-resistant gene. The targeting vector is introduced into ES cells by electroporation. Homologous recombination gives rise to the disruption of the targeted gene by the marker gene. The embryonic engineering allows the ES cells to become mice.

10.2.2 Mouse Phenotyping as Output

Since the 1990s transgenic mice and knockout mice have been widely used to study particular DNA sequences at the organism level. A typical case is the *p53* gene, officially called *Trp53*. Originally, *Trp53* was identified as a major tumor marker product, but its function was not well understood until results of knockout mouse studies became available. Besides, Trp53 mutant lines allowed the discovery of novel *Trp53* functions that had not been expected from the molecular knowledge of the gene. Several examples of the *bona fide* biosimulator are briefly summarized in the following sections.

10.2.2.1 *Trp53* Function in Adulthood

Donehower, Harvey, Slagle, McArthur, Montgomery, Butel, and Bradley(1992) were the first to use knockout mice for analyzing the nature of the *Trp53* tumor suppressor gene. Almost all of the Trp53 homozygous knockout mice developed malignant thymic lymphoma and died by the age of 30 weeks. The heterozygotes were also highly tumor-prone combined with the loss of heterozygosity. The current understanding of the *Trp53* gene function is that of a gatekeeper for DNA repair. When genomic DNA is damaged during the cell cycle, *Trp53* causes the G1 arrest. If the damage is repaired, the cell cycle resumes, otherwise the cell becomes apoptotic and dies.

10.2.2.2 *Trp53* Function in Embryogenesis

We confirmed similar homozygous and heterozygous effects of *Trp53* by using a different knockout system (Gondo, Nakamura, Nakao, Sasaoka, Ito, Kimura, and Katsuki 1994). In addition, we found that the Trp53 homozygous knockout mice were extremely susceptible to teratogenesis rather than tumorigenesis during embryogenesis (Norimura, Nomoto, Katsuki, Gondo, and Kondo 1996). As summarized in Fig. 4, the frequency of abnormal development without radiation exposure was 30 – 40%. On the other hand, it increased twofold (60–80%) after 2-Gy exposure at the embryonic development stage E9.5 for any genotype. The nature of abnormalities was, however different from genotype to genotype. Three-quarters of the abnormal development in wild-type homozygotes were late deaths. In contrast, almost all of the abnormalities seen in the knockout homozygotes were morphological anomalies (e.g., polydactyly, megadactyly, tail anomaly, dwarf, cleft palate, exencephaly). No tumors were observed in this embryogenesis study.

10.2.2.3 *Trp53* Point Mutation

The introduction of point mutations in a gene can improve the elucidation and dissection of its higher functions. For example, the Cre-loxP system allows the construction of missense mutations. Liu, Parant, Lang, Chau, Chavez-Reyes, El-Naggar, Multani, Chang, and Lozano (2004) used the Cre-loxP system to introduce in *Trp53* a base substitution that causes the arginine-to-proline missense mutation

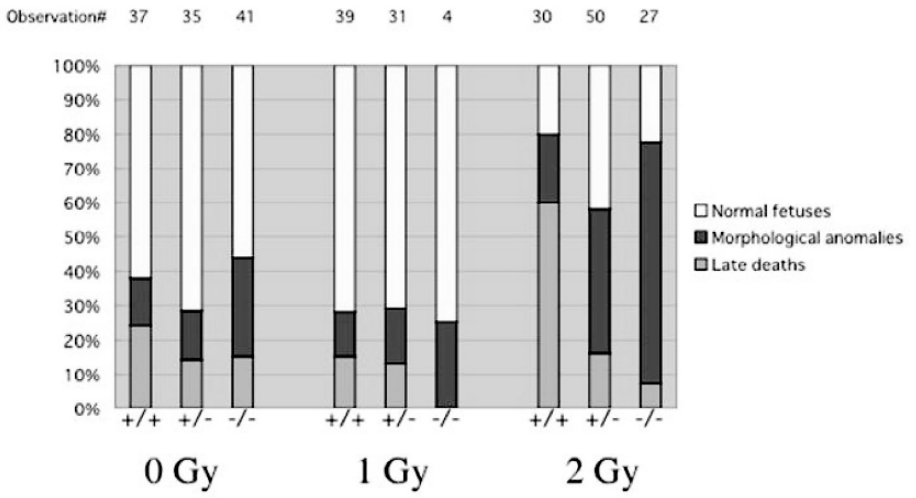


Fig. 4. The effect of γ-ray irradiation at E9.5 day of embryogenesis in p53 knockout mice. The genotypes are indicated below the chart (+, wild-type p53 allele; -, knockout p53 allele). The examined numbers of mice are shown above the chart.

at position 172 (R172P). The homozygotes for the R172P mutation showed signs of cell cycle arrest but no apoptosis. The early onset of tumorigenesis was not observed. These results strongly vindicate the usefulness of point mutation studies in characterizing gene functions.

10.3 Renaissance of Classical Genetics

The *Trp53* gene study using gene targeting demonstrated that the finer structure and functions of a gene are revealed by in-depth phenotyping at various tissues and developmental stages. In addition, many sets of mutant alleles are ideal for revealing the comprehensive function of each base pair in genomic DNA sequences. Since the 1990s these issues have been exploited in another, *N*-ethyl-*N*-nitrosourea-based (ENU) mouse mutagenesis approach for large–scale functional genomics studies (e.g., Brown and Nolan 1998; Hrabé de Angelis and Balling 1998). ENU is a highly potent and extensively studied mutagen that induces point mutations during mouse spermatogenesis (Russell, Kelly, Hunsicker, Bangham, Maddux, and Phipps 1979; Russell, Hunsicker, Raymer, Steele, Stelzner, and Thompson 1982a; Russell, Hunsicker, Carpenter, Cornett, and Guinn 1982b; Hitotsumachi, Carpenter, and Russell 1985; Noveroske, Weber, and Justice 2000).

10.3.1 Chemical Mutagenesis for Genome Wide Studies

The concept of mouse ENU mutagenesis is to induce randomly and genomewide as many point mutations as possible (the method is not able to target specified DNA sequences). While the throughput of transgenic and gene targeting technologies is

low-throughput and knockout mouse methods may confer lethality, ENU mutagenesis results in partial loss of function at much higher levels of throughput.

10.3.2 ENU-Based Phenotype-Driven Mouse Mutagenesis

The primary objective of the ENU mouse mutagenesis project is to identify and construct mutant lines on a scale such that the whole mutant archive encompasses at least one mutant for any gene. Worldwide, more than ten large-scale ENU mouse mutagenesis projects with a genomewide coverage have been initiated.

10.3.2.1 Phase I: Dominant Screens

The identification of useful mutants depends on how meticulously the phenotype assessment is conducted. All the G1 progenies depicted in Fig. 5 are subject to various phenotype screens. The full description of the RIKEN screening platform is available at the URL http://www.gsc.riken.jp/Mouse/. RIKEN started full screening in 2000. So far more than 20,000 (Masuya, Nakai, Motegi, Niinaya, Kida, Kaneko, Aritake, Suzuki, Ishii, Koorikawa, Suzuki, Inoue, Kobayashi, Toki, Wada, Kaneda, Ishijima, Takahashi, Minowa, Noda, Wakana, Gondo, and Shiroishi 2004) G1 mice have been screened. The basic screens are modified SHIRPA (SmithKline Beecham Pharmaceuticals, Harwell MRC Mouse Genome Centre and Mammalian Genetics Unit, Imperial College School of Medicine at St Mary's, Royal London Hospital, St Bartholomew's and the Royal London School of Medicine, Phenotype Assessment) (Rogers, Fisher, Brown, Peters, Hunter, and Martin 1997) which includes morphological and behavioral screens. Other additional phenotype screenings include hematology, urine and serum biochemical analyses.

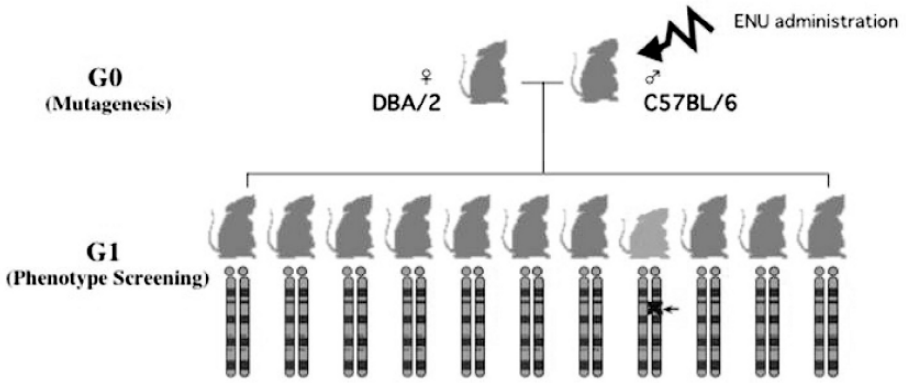


Fig. 5. Overall scheme of the dominant mutant screening at RIKEN. ENU is administered to male C57BL/6 inbred mice (abbreviated G0). The ENU-treated males are mated to another inbred strain DBA/2. All the F1 hybrid offspring, designated as G1, are subjected to exhaustive phenotype screens. In this scheme dominant mutations are collectively identified in a genomewide manner.

Mutant candidates or phenodeviants are identified by comparison with control mice of the same genetic background. It is therefore critical to prepare a reliable and large-scale control phenotype dataset. Phenodeviants are mated to produce the G2 offspring, which is called the inheritance test cross. If a candidate phenotype anomaly occurred due to an ENU-induced mutation, the phenotype should be observed in half of the G2 offspring. Only the phenodeviants that passed the inheritance test are registered as mutants.

Mutants have been identified in approximately two to three percent of G1 mice using the basic screening platforms of RIKEN and other ENU mouse mutagenesis projects. Thus, the expected number of dominant mutant lines that have been newly generated in the past 5 years by large-scale ENU mouse mutagenesis programs is roughly 5000 (= 2 – 3% x 20,000 G1 x 10 projects). About the same number of mutant lines have been identified and established during the 100-year history of mouse genetics. Transgenics and gene targeting have produced an equivalent number of mutants since the late 1980s. Altogether, more than 15,000 mutant lines are available as a resource for studying gene and genome functions.

10.3.2.2 Phase II: Gene Identification and Recessive Screens

Connecting the causative genomic DNA changes to the phenotype outcome demands the identification of the ENU-induced mutation site that is responsible for the established lines. The classical genetics approach of backcrossing, to map the mutation in the mouse chromosomes or genome-based candidate approaches are combined to identify the site of the mutation. Currently, both mapping and positional cloning are utilized to locate the mutations obtained by ENU mouse mutagenesis programs.

Another key objective of Phase II of the large-scale ENU mouse mutagenesis project is to systematically establish recessive mutants. While many dominant mutants have been collected, many human genetic diseases are recessive. Dominant screens have also revealed that some genes exhibit their dominant phenotype due to loss of function. In such cases, other genes that would show the same phenotypes by gain of function can hardly be identified. The mutation rate of dominant mutations due to the loss of function is estimated to be about 10^{-3}/locus/G1. This frequency is probably 1000-fold higher than the frequency for gain of function. In order to establish comprehensive human genetic disease models that cover most of the mouse genes, it is necessary to develop recessive screens. As shown in Fig. 6, recessive screens take much more time, space, and manpower.

10.3.2.3 Informatics Infrastructure for ENU Mouse Mutagenesis

Database System. Large-scale ENU mouse mutagenesis requires the construction and organization of an informatics infrastructure. RIKEN chose Oracle database as the core of the infrastructure (Gondo 2001; Masuya et al. 2004). All manipulations, protocols, data recordings, data analyses, annotations, etc. are stored in and retrieved from the database server by multiple client machines. Some examples are described below.

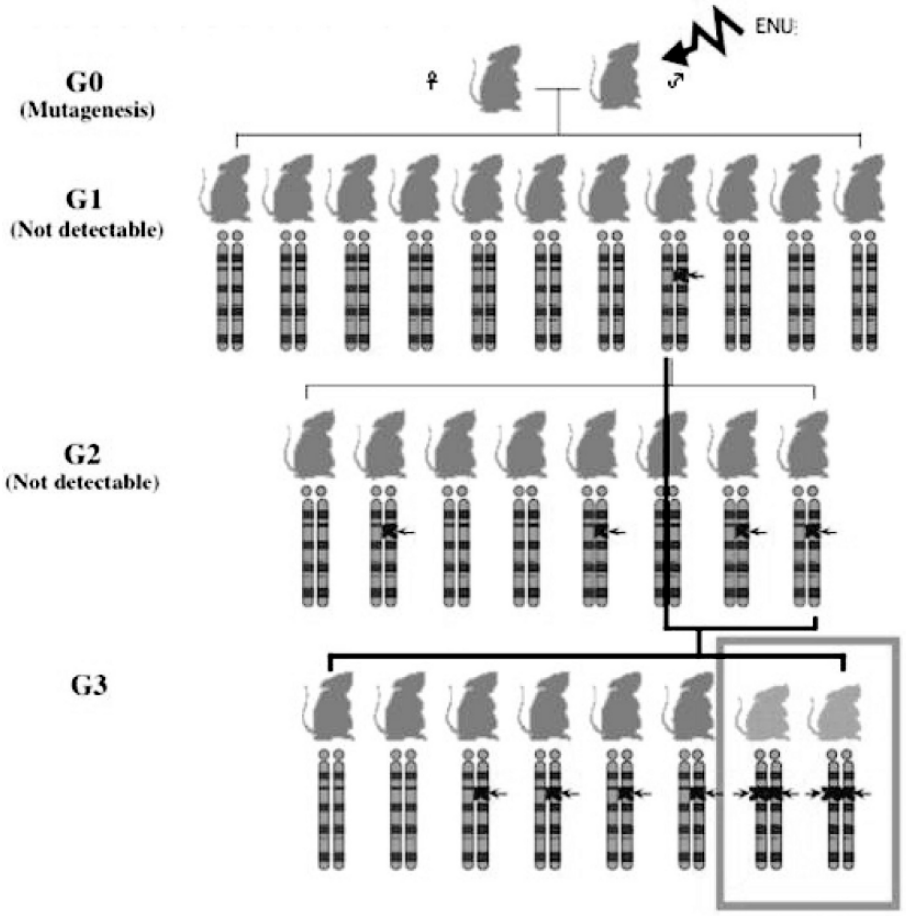


Fig. 6. Overall scheme of recessive mutant screening at RIKEN. G1 and G2 mice are obtained by the same scheme as shown in Fig. 5 for the dominant mutant screening. Recessive traits, however, are concealed in G1 and G2. In order to detect recessive mutations, randomly chosen G1 and its direct progeny of G2 are mated to reproduce G3. In some cases, females and males of the same G2 litter are mated to obtain G3 offspring.

Bar-coding. More than 100 G1 mice are produced every week. They remain under investigation until 78 weeks of age. As a result, at any given time approximately 8000 G1 mice are present in our facility. G0, G2, and G3 mice and control mice are maintained in a specific pathogen-free (SPF) animal facility. A bar-coding system that facilitates the identification of each mouse is crucial for the daily husbandry work.

Local-Area Network (LAN). The mice must be protected from lethal infections by pathogens. Consequently, physical material transfer including the use of pens and papers are severely limited to maintain the SPF grade. Data transfer also requires no or minimal writing and copying errors. To meet these requirements, the entire data

input is conducted only once at the original location, using a client computer. The data are instantaneously transferred to the Oracle database server. Even one minute of LAN interruption can cause serious problems in the data fidelity. Thus, a robust and secure LAN configuration and its maintenance are extremely critical.

Biometrical Devices. Automated biometrical devices for phenotype screening have been installed at various locations. Since stand-alone-type computers usually control the devices it is often necessary to develop device-specific data-transfer protocols to guarantee efficient data transport to the database server. Occasionally, we modified and even built new biometrical devices. This type of transformation of biometrics will also contribute to the development of new diagnostic systems for human use.

10.3.3 ENU-Based Gene-Driven Mouse Mutagenesis

Genetics has a more than 100-year history. During the foundation era, the mouse was an important species that supported the establishment and advancement of genetics. Mouse genetics was also a driving force in the field of mutagenesis. The first report of artificially induced mutations was the X-ray mutagenesis study on *Drosophila* by Muller (1927). However, 4 years earlier Little and Bagg (1923) already pointed out the possibility of X-ray-induced mutations in the mouse. In spite of the long history and extensive studies, mouse mutagenesis studies and mutations are not considered to be a high-throughput INPUT system for the biosimulator. Mutations are neither quick nor efficient enough to alter and operate the *bona fide* life program of the mouse genome.

Until very recently, mutations were detected in natural populations as polymorphisms or randomly induced by mutagens (see Fig. 1). Mutations are rare and random events. It used to be impossible to change the mouse genome by targeting a particular base pair or sequence. Even for an established mutant derived from a natural population or mutagenesis it will take a few years to identify the site of base pair change in the mouse genome. Currently available tools to alter a genomic DNA sequence are the transgenic and knockout mouse systems as described in Section 10.2.1. The transgenic and knockout mouse systems are low-throughput. Moreover both methods modify only a stretch of genomic DNA sequences by insertion and deletion, respectively. An ideal INPUT system to alter the program of the mouse genome is to deliberately change any single base pair versus another.

10.3.3.1 Mutant Mouse Library

There are as yet no efficient methods to alter a target base pair in the mouse genome. The second best approach is to construct a mutant mouse library that covers a sufficiently large number of point mutations. As described in Section 10.3.2, the large-scale ENU mouse mutagenesis to collect and establish mutant lines has started with a phenotype-driven approach.

At RIKEN, we are focusing on the screening of late-onset phenotypes in G1 population (Fig. 5). G1 mutants, carrying a late-onset phenotype (i.e., tumors or diabetes), may be lethal or develop sterility during disease progression. For that

reason we take sperm from all the G1 males when they are 12 weeks old and subject them to cryopreservation before the onset time.

The frozen sperm archive is considered to be the mutant mouse library. A key question is how many mutations are indeed preserved in this archive. We started to estimate the ENU-induced mutation rate per base pair in since 2001 and found roughly one mutation per million base pairs per gamete (Sakuraba, Sezutsu, Takahasi, Tsuchihashi, Ichikawa, Fujimoto, Kaneko, Nakai, Uchiyama, Goda, Motoi, Ikeda, Karashima, Inoue, Kaneda, Masuya, Minowa, Noguchi, Toyoda, Sakaki, Wakana, Noda, Shiroishi, and Gondo 2005). Recently, results of other groups supported our findings (Quwailid, Hugill, Dear, Vizor, Wells, Horner, Fuller, Weedon, McMath, Woodman, Edwards, Campbell, Rodger, Carey, Roberts, Glenister, Lalanne, Parkinson, Coghill, McKeone, Cox, Willan, Greenfield, Keays, Brady, Spurr, Gray, Hunter, Brown, and Cox 2004; Augustin, Sedlmeier, Peters, Huffstadt, Kochmann, Simon, Schöniger, Garke-Mayerthaler, Laufs, Mayhaus, Franke, Klose, Graupner, Kurzmann, Zinser, Wolf, Voelkel, Kellner, Kilian, Seelig, Koppius, Teubner, Korthaus, Nehls, and Wattler 2005; Michaud, Culiat, Klebig, Barker, Cain, Carpenter, Easter, Foster, Gardner, Guo, Houser, Hughes, Kerley, Liu, Olszewski, Pinn, Shaw, Shinpock, Wymore, Rinchik, and Johnson 2005). Based on this mutation rate, each G1 is expected to carry 3000 ENU-induced heterozygous mutations in 3×10^9 bp of the paternally inherited genome. So far we have archived about 10,000 G1 (Sakuraba et al. 2005) male sperm samples at RIKEN. Worldwide there are more than 40,000 G1 sperm archives. Within several years the total number of G1 sperm archives will exceed 100,000. The cumulative number of ENU-induced mutations will be

$$3{,}000 \text{ mutations/G1} \times 100{,}000 \text{ G1} = 3 \times 10^8 \text{ mutations} \quad (1)$$

or one mutation per 10 bp on average. This level is considered to be semisaturated. If the size of the target gene is 10,000 bp, 100 independent mutations in the gene would be available in the frozen sperm archive. This number of stored mutants should be more than sufficient to elucidate the molecular and biological function(s) of a gene in detail.

10.3.3.2 New Point Mutation Discovery System

If we consider for practical reasons the frozen sperm archive equivalent to the mutant mouse library, the next key question is how can we discover effectively ENU-induced point mutations in target genes. The basic scheme to identify ENU-induced mutation in target genes is depicted in Fig. 7. First, PCR primer pairs are designed to amplify the sequences of the target gene. Mutations can be identified in the PCR products by direct sequencing. For practical reasons, it is necessary to prepare either genomic DNA or cDNA archives from all G1 males. If we screen the genomic sequences of the target genes, genomic DNA is simply isolated from any organs or tissues of the G1 males. However, all genomic sequence information including the exon and intron structure must be known to permit the design of PCR primers for the

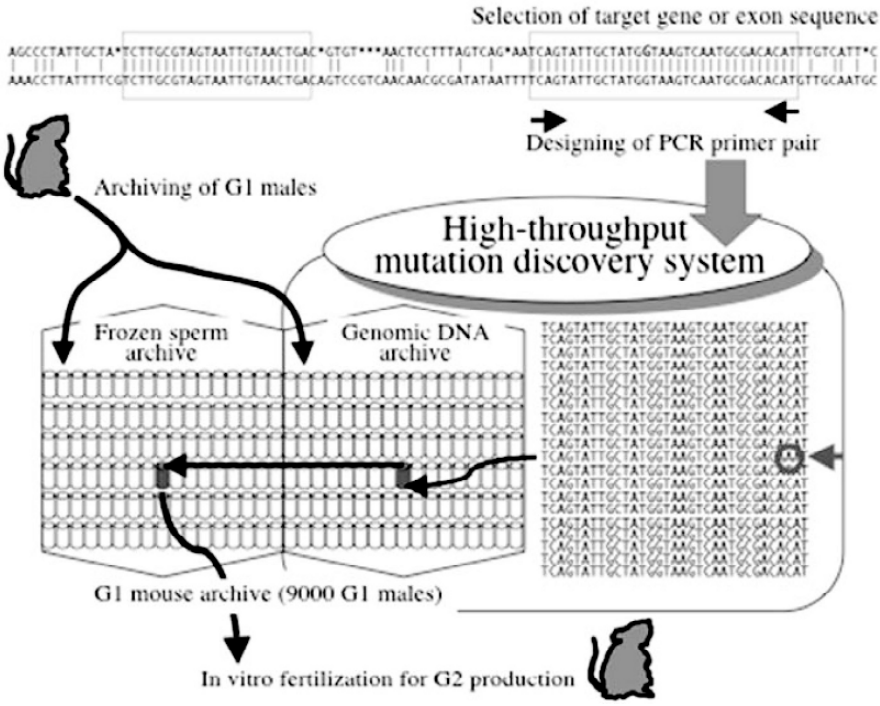


Fig. 7. Overall scheme of the ENU-based gene-driven mutagenesis at RIKEN. All the G1 males (see Fig. 5) are subjected to sperm cryopreservation and genomic DNA extraction to construct "Frozen sperm archive" and "Genomic DNA archive," respectively. To construct mutant mouse carrying a point mutation in the target gene, several appropriate PCR primer pairs are designed. The target gene is amplified and screened for any ENU-induced mutations from the genomic DNA archive. Once a mutation is found, the corresponding sperm sample in the frozen sperm archive is used for *in vitro* fertilization (IVF) to reconstruct the strain as live mice.

target gene. On the other hand, if the focus is on the protein-coding sequences of the target genes, PCR primer design and screening are simpler and faster. Most cDNA sequence information is deposited in public databases. In addition, cDNA sequences are much more compact and shorter than genomic sequences. Nevertheless, the success of cDNA-based screening depends on the careful selection of tissues that express the target mRNA and the sampling time.

At RIKEN, genomic DNA was chosen to construct the mutant DNA archive when we started the system development in 2000. We assumed that the mouse genome project would be completed and all the mouse genomic sequences would become available within a few years. Indeed, 2 years later the Mouse Genome Sequencing Consortium (2002) published the first paper. The updated mouse genomic sequences are available from the Ensembl database (http://www.ensembl.org/).

We started a feasibility study with a new mutation discovery system by using a direct sequencing method, when the number of G1 in the archive exceeded 2,000. Although we

identified mutations, the process was very labor intensive and not cost effective. Hence, we started to investigate other high-throughput mutation discovery systems.

A new mutation discovery system, Temperature Gradient Capillary Electrophoresis (TGCE), became commercially available in 2002. The TGCE system detects the difference of a heteroduplex DNA fragment from the homoduplex counterpart during the capillary electrophoresis (Gao and Yeung 2000; Murphy, Hafez, Philips, Yarnell, Gutshall, and Berg 2003). The formation of heteroduplex fragments by a point mutation is schematically shown in Fig. 8.

By using the TGCE system as the primary screening method for identifying mutations in the target sequences of the G1 genomic DNA archive, the ENU-based gene-driven mutagenesis has become practicable. At present, the mutation discovery rate is about 100 per year per TGCE system (Sakuraba et al. 2005). Each mutation corresponds to one knockout mouse line. Since it takes more than a year to characterize the phenotypes, the mutation discovery rate vastly exceeds the capacity of the biological and functional analyses of the mutant mouse lines. The size of the RIKEN frozen sperm archive is about 10,000 as shown above (Sakuraba et al. 2005). In this archive, 3×10^7 independent point mutations are preserved for the biological analyses (see Eq. (1) in Section 10.3.3.1). This is a tremendous genetic resource for the biological annotation of the genome function.

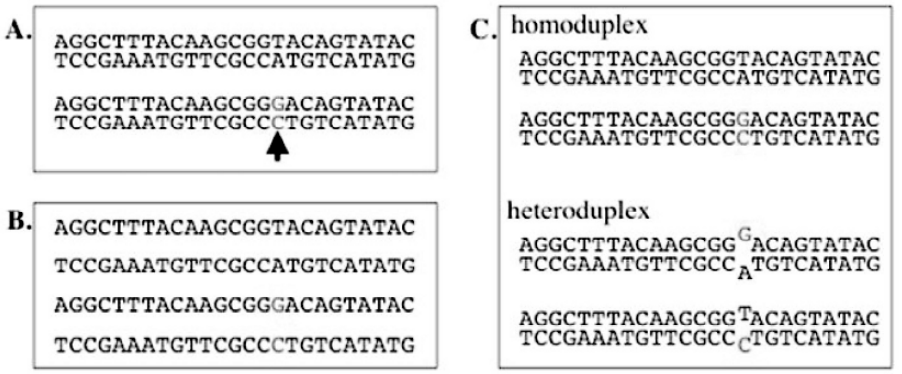


Fig. 8. PCR fragments amplified from G1 mutant carrying an ENU-induced mutation and the heteroduplex formation. (A) Design of PCR primers that amplify a target fragment of the G1 genome. When the target region of the G1 genome carries an ENU-induced mutation (arrow), it appears only in the paternally inherited genome as heterozygote. (B) Heat denaturation of PCR products makes all the fragments single stranded. (C) Renaturation of the heat-denatured PCR products reconstructs the double helix structure and heteroduplex fragments that have a mismatch-pairing at the site of the mutation. Theoretically the molar ratio is equal between homoduplex and heteroduplex fragments.

10.3.3.3 Use of the RIKEN Gene-Driven Mutagenesis

Future progress in mouse genetics and functional studies depends on the opening of ENU-based gene-driven mutagenesis systems to the public and any researchers who have expertise in analyzing mutant mice. At RIKEN we envision the following scenario:

1. USER designs the PCR primers for the target gene and sends them to RIKEN.
2. RIKEN screens the mutagenized genomic DNA archive.
3. RIKEN reports all the identified mutations to USER.
4. USER decides which mutation to analyse.
5. RIKEN retrieves live mice from the corresponding frozen sperm and sends the mice to USER.
6. USER conducts biological and functional studies on the mutant mice as *a bona fide* biosimulator.
7. The established mutant line will be open to public in an appropriate time.

Based on the above plan, we have already started cooperative feasibility studies for a number of genes with many collaborators. All information about the current target genes and their chromosomal locations are open in our website (http://www.gsc.riken.jp/Mouse/, then see "gene-driven mutagenesis"). As of today, 251 genes are listed as targets of which 176 genes or 70% are collaborative targets. For each collaborative gene, we ask USER to be the principal investigator of the study. Using this cooperative framework we have identified to date about 300 point mutations in more than 60 target genes.

10.4 Conclusions

Mouse genetics provides a versatile tool to study gene networks and whole genome function. Here, mutant mice are considered to be *bona fide* biosimulators. Previously, when geneticists discovered a mutant, it became their life work based on mating and extensive phenotype analyses. Now, transgenic and knockout mouse systems have made it possible to alter the mouse genome using DNA technology and embryonic engineering. Furthermore, ENU mouse mutagenesis projects have opened a new platform for the genomewide study of mouse functional genomics. The renaissance of classical genetics gave rise to a large number of mutant mouse lines derived from both phenotype-driven and gene-driven approaches. Prior to this era, more than 5000 mutant mouse lines were available. Comparable numbers of transgenic and knockout mouse lines were generated in the past 15 years. Today, ENU mouse mutagenesis contributes almost 1000 new mutant lines every year. Abundant numbers of mutants as INPUT for the *bona fide* biosimulator are now available for the development and transformation of the reproducible high-throughput OUTPUT system as a phenometrics platform.

Acknowledgements

The author thanks Drs. Y. Sakuraba, H. Sezutsu, K. R. Takahasi, M. Uchiyama, R. Fukumura, and T. Murata for their contribution in the development of the ENU-based gene-driven mutagenesis at RIKEN, Y. Nakai, A. Ikeda, N. Fujimoto, K. Tsuchihashi, Y. Karashima, R. Ichikawa, S. Kaneko, H. Sasaki, K. Takenouchi, N. Goda, R. Motoi, and K. Karouji for their excellent technical assistance and extensive data analysis, and in particular the former director, Dr. Akiyoshi Wada, for his leadership, encouragement, and support. This study was partly supported by Grants-in-Aid for Scientific Research (A) from the Ministry of Education, Culture, Sports, Science and Technology in Japan.

References

Auerbach, C., and Robson, J.M. (1946) Chemical production of mutations. Nature 157:302.

Augustin, M., Sedlmeier, R., Peters, T., Huffstadt, U., Kochmann, E., Simon, D., Schöniger, M., Garke-Mayerthaler, S., Laufs, J., Mayhaus, M., Franke, S., Klose, M., Graupner, A., Kurzmann, M., Zinser, C., Wolf, A., Voelkel, M., Kellner, M., Kilian, M., Seelig, S., Koppius, A., Teubner, A., Korthaus, D., Nehls, M., and Wattler, S. (2005) Efficient and fast targeted production of murine models based on ENU mutagenesis. Mamm. Genome 16: 405-413.

Brown, S.D.M., and Nolan, P.M. (1998) Mouse mutagenesis-systematic studies of mammalian gene function. Hum. Mol. Genet. 7:1627-1633.

Capecchi, M.R. (1989) Altering the genome by homologous recombination. Science 244:1288-1292.

Donehower, L.A., Harvey, M., Slagle, B.L., McArthur, M.J., Montgomery, C.A., Jr., Butel, J.S., and Bradley, A. (1992) Mice deficient for p53 are developmentally normal but susceptible to spontaneous tumours. Nature 356:215-221.

Gao, Q., and Yeung, E.S. (2000) High-throughput detection of unknown mutations by using multiplexed capillary electrophoresis with poly(vinylpyrrolidone) solution. Anal. Chem. 72:2499-2506.

Gondo, Y., Nakamura, K., Nakao, K., Sasaoka, T., Ito, K., Kimura, M., and Katsuki, M. (1994) Gene replacement of the p53 gene with the lacZ gene in mouse embryonic stem cells and mice by using two steps of homologous recombination. Biochem. Biophys. Res. Commun. 202:830-837.

Gondo, Y. (2001) Bioinformatics for the large-scale mouse mutagenesis project. In: N. Baba, L.C. Jain, and R.J. Howlett (Eds.), *Knowledge-Based Intelligent Information Engineering Systems & Allied Technologies KES'2001*. IOS Press, Tokyo, pp. 763-767.

Hitotsumachi, S., Carpenter, D.A., and Russell, W.L. (1985) Dose-repetition increases the mutagenic effectiveness of N-ethyl-N-nitrosourea in mouse spermatogonia. Proc. Natl. Acad. Sci. USA 82:6619-6621.

Hrabé de Angelis, M., and Balling, R. (1998) Large scale ENU screens in the mouse: Genetics meets genomics. Mutat. Res. 400:25-32.

Little, C.C., and Bagg, H.J. (1923) The occurrence of two heritable types of abnormality among the descendants of X-rayed mice. Am. J. Roentgenol. Radiat. Ther. 10:975-989.

Liu, G., Parant, J.M., Lang, G., Chau, P., Chavez-Reyes, A., El-Naggar, A.K., Multani, A., Chang, S., and Lozano, G. (2004) Chromosome stability, in the absence of apoptosis, is critical for suppression of tumorigenesis in Trp53 mutant mice. Nat. Genet. 36:63-68.

Masuya, H., Nakai, Y., Motegi, H., Niinaya, N., Kida, Y., Kaneko, Y., Aritake, H., Suzuki, N., Ishii, J., Koorikawa, K., Suzuki, T., Inoue, M., Kobayashi, K., Toki, H., Wada, Y., Kaneda, H., Ishijima, J., Takahashi, K.R., Minowa, O., Noda, T., Wakana, S., Gondo, Y., and Shiroishi, T. (2004) Development and implementation of a database system to manage a large-scale mouse ENU-mutagenesis program. Mamm. Genome 15:404-411.

Michaud, E.J., Culiat, C.T., Klebig, M.L., Barker, P.E., Cain, K.T., Carpenter, D.J., Easter, L.L., Foster, C.M., Gardner, A.W., Guo, Z.Y., Houser, K.J., Hughes, L.A., Kerley, M.K., Liu, Z., Olszewski, R.E., Pinn, I., Shaw, G.D., Shinpock, S.G., Wymore, A.M., Rinchik, E.M., and Johnson, D.K. (2005) Efficient gene-driven germ-line point mutagenesis of C57BL/6J mice. BMC Genomics 6:164.

Mouse Genome Sequencing Consortium (2002) Initial sequencing and comparative analysis of the mouse genome. Nature 420:520-562.

Muller, H.J. (1927) Artificial transmutation of the gene. Science 66:84-87.

Murphy, K., Hafez, M., Philips, J., Yarnell, K., Gutshall, K., and Berg, K. (2003) Evaluation of temperature gradient capillary electrophoresis for detection of the factor v leiden mutation : Coincident identification of a novel polymorphism in factor v. Mol. Diagn. 7:35-40.

Norimura, T., Nomoto, S., Katsuki, M., Gondo, Y., and Kondo, S. (1996) p53-dependent apoptosis suppresses radiation-induced teratogenesis. Nat. Med. 2:577-580.

Noveroske, J.K., Weber, J.S., and Justice, M.J. (2000) The mutagenic action of N-ethyl-N-nitrosourea in the mouse. Mamm. Genome 11:478-483.

Palmiter, R.D., Brinster, R.L., Hammer, R.E., Trumbauer, M.E., Rosenfeld, M.G., Birnberg, N.C., and Evans, R.M. (1982) Dramatic growth of mice that develop from eggs microinjected with metallothionein-growth hormone fusion genes. Nature 300:611-615.

Quwailid, M.M., Hugill, A., Dear, N., Vizor, L., Wells, S., Horner, E., Fuller, S., Weedon, J., McMath, H., Woodman, P., Edwards, D., Campbell, D., Rodger, S., Carey, J., Roberts, A., Glenister, P., Lalanne, Z., Parkinson, N., Coghill, E.L., McKeone, R., Cox, S., Willan, J., Greenfield, A., Keays, D., Brady, S., Spurr, N., Gray, I., Hunter, J., Brown, S.D., and Cox, R.D. (2004) A gene-driven ENU-based approach to generating an allelic series in any gene. Mamm. Genome 15:585-591.

Rogers, D.C., Fisher, E.M., Brown, S.D., Peters, J., Hunter, A.J., and Martin, J.E. (1997) Behavioral and functional analysis of mouse phenotype: SHIRPA, a proposed protocol for comprehensive phenotype assessment. Mamm. Genome 8:711-713.

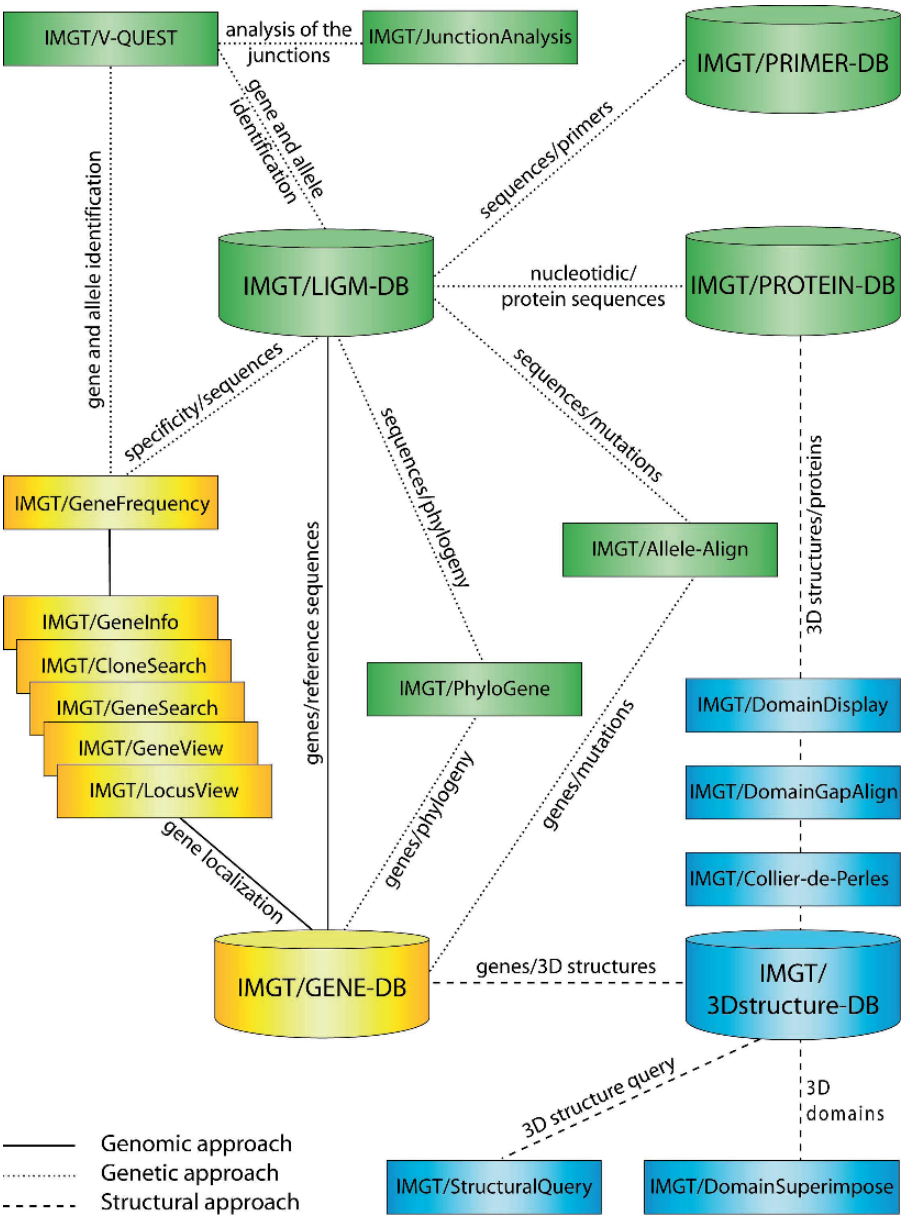
Russell, W. L., Kelly, E.M., Hunsicker, P.R., Bangham, J.W., Maddux, S.C., and Phipps, E.L., (1979) Specific-locus test shows ethylnitrosourea to be the most potent mutagen in the mouse. Proc. Natl. Acad. Sci. USA 76:5818-5819.

Russell, W.L., Hunsicker, P.R., Raymer, G.D., Steele, M.H., Stelzner, K.F., and Thompson, H.M. (1982a) Dose-response curve for ethylnitrosourea-induced specific-locus mutations in mouse spermatogonia. Proc. Natl. Acad. Sci. USA 79:3589-3591.

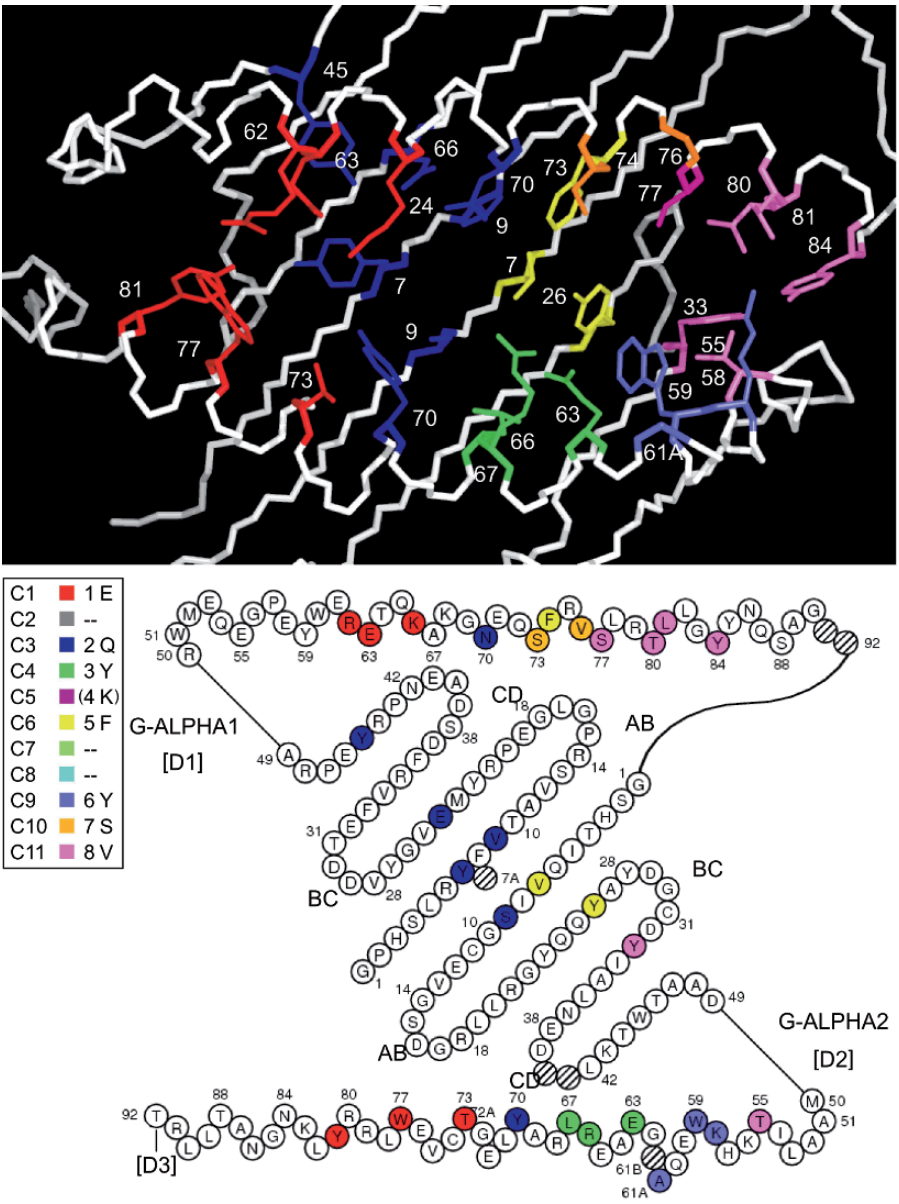
Russell, W.L., Hunsicker, P.R., Carpenter, D.A., Cornett, C.V., and Guinn, G.M. (1982b) Effect of dose fractionation on the ethylnitrosourea induction of specific-locus mutations in mouse spermatogonia. Proc. Natl. Acad. Sci. USA 79:3592-3593.

Sakuraba, Y., Sezutsu, H., Takahasi, K.R., Tsuchihashi, K., Ichikawa, R., Fujimoto, N., Kaneko, S., Nakai, Y., Uchiyama, M., Goda, N., Motoi, R., Ikeda, A., Karashima, Y., Inoue, M., Kaneda, H., Masuya, H., Minowa, O., Noguchi, H., Toyoda, A., Sakaki, Y., Wakana, S., Noda, T., Shiroishi, T., and Gondo, Y. (2005) Molecular characterization of ENU mouse mutagenesis and archives. Biochem. Biophys. Res. Commun. 336:609-616.

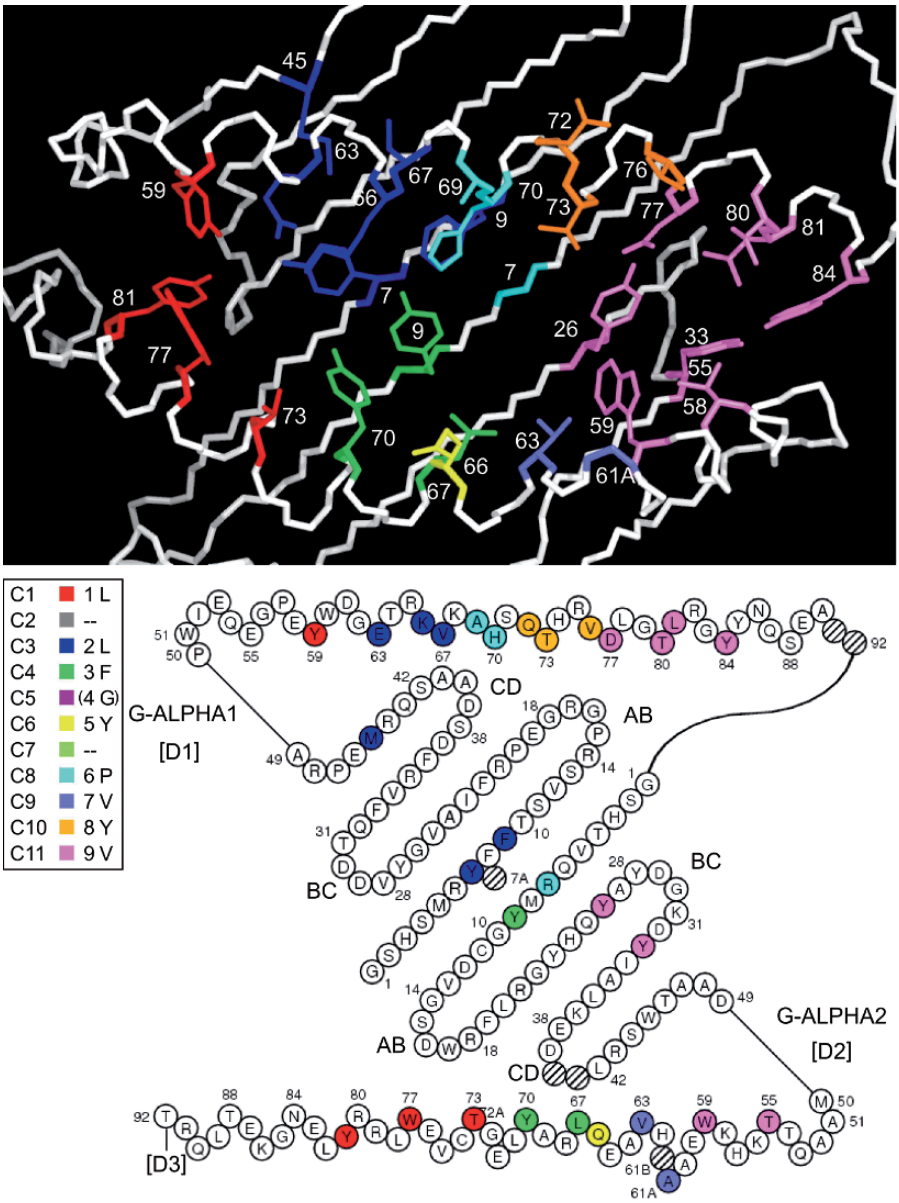
Index



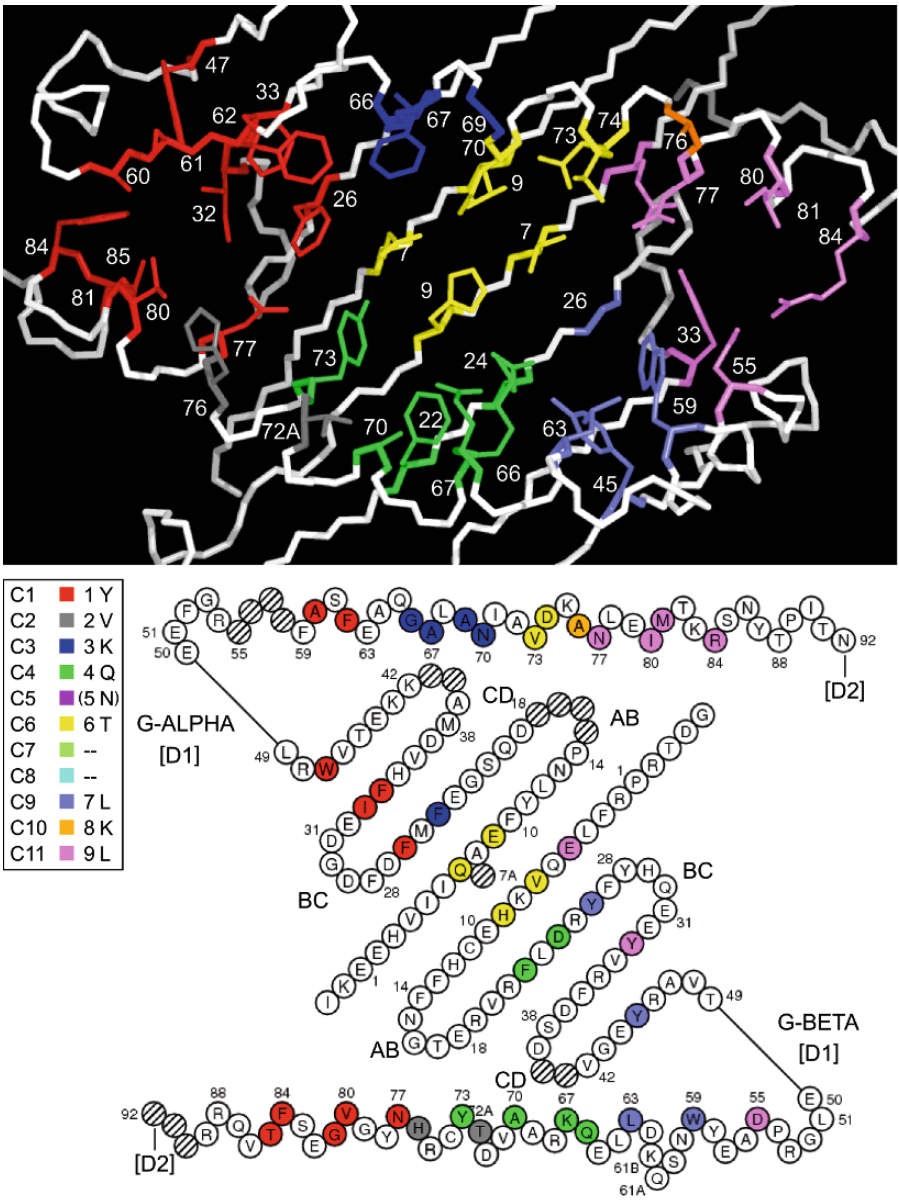
Ch. 1 Fig. 1. IMGT ®, the international ImMunoGeneTics information system® (http://imgt.cines.fr) databases and tools. The IMGT Repertoire and other IMGT Web resources are not shown. Examples of interactions between the databases (cylinders) and tools (rectangles) in the genomic, genetic and structural approaches are represented respectively by continuous, dotted and broken lines.



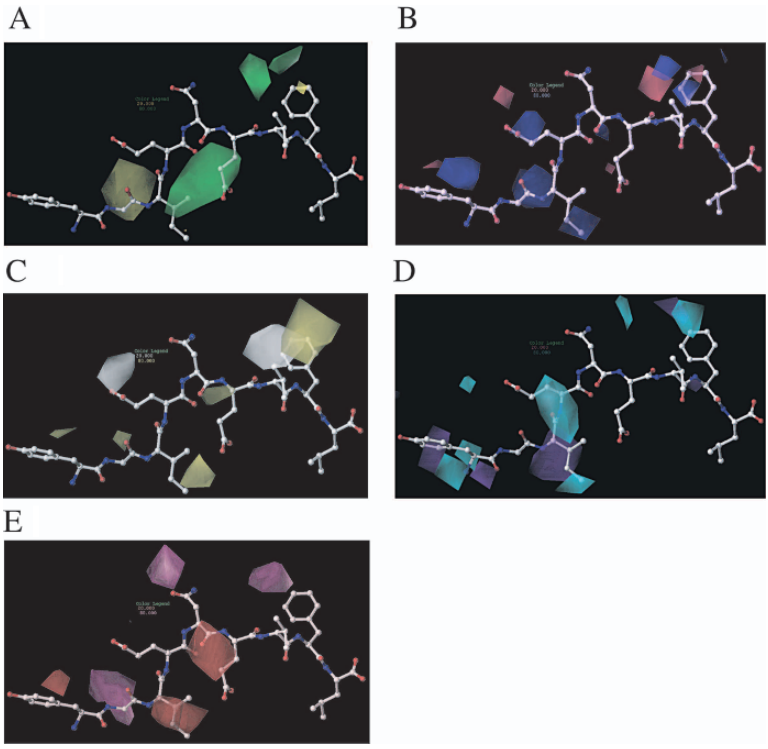
Ch. 2 Fig. 6. IMGT pMHC contact sites of mouse H2-K1 MHC-I and a 8-amino acid peptide (1jtr). (A) 3D structure of the mouse H2-K1*01 groove. (B) IMGT pMHC contact sites IMGT Colliers de Perles. Both views are from above the cleft with G-ALPHA1 on top and G-ALPHA2 on bottom. In the box, C1 to C11 refer to contact sites (Kaas and Lefranc 2005), 1 to 8 refer to the numbering of the peptide amino acids P1 to P8. There are no C2, C7 and C8 in MHC-I 3D structures with 8-amino acid peptides. There is no C5 in this 3D structure as P4 does not contact MHC amino acids (4K is shown between parentheses in the box).



Ch. 2 Fig. 7. IMGT pMHC contact sites of human HLA-A*0201 MHC-I and a 9-amino acid peptide (1ao7). (A) 3D structure of the human HLA-A*0201 groove. (B) IMGT pMHC contact sites IMGT Colliers de Perles. Both views are from above the cleft with G-ALPHA1 on top and G-ALPHA2 on bottom. In the box, C1 to C11 refer to contact sites (Kaas and Lefranc 2005). 1 to 9 refer to the numbering of the peptide amino acids P1 to P9. There are no C2 and C7 in MHC-I 3D structures with 9-amino acid peptides. There is no C5 in this 3D structure as P4 does not contact MHC amino acids (4G is shown between parentheses in the box).



Ch. 2 Fig. 8. IMGT pMHC contact sites of the human HLA-DRA*0101 and HLA-DRB1*0401 MHC-II and the peptide side chains (9-amino acids located in the groove). (A) 3D structure of the human HLA-DRA*0101 and HLA-DRB1*0401 groove (1j8h). (B) IMGT pMHC contact sites IMGT Colliers de Perles. Both views are from above the cleft with G-ALPHA on top and G-BETA on bottom. In the box, C1 to C11 refer to contact sites. 1 to 9 refer to the numbering of the peptide amino acids 1 to 9 located in the groove. There is no C5 and C7 in MHC-I 3D structures with 9-amino acid peptides. There is no C5 in this 3D structure as 5 does not contact MHC amino acids (5N is shown between parentheses in the box).



Ch. 4 Fig. 4. H2-D^{b} allele: Steric bulk maps (A), electrostatic potentials maps (B), hydrophobic interaction maps (C), H-bond donor maps (D), H-bond acceptor maps (E).