BIOINFORMATICS

Vol. 18 no. 3 2002
Pages 434–439

JenPep: a database of quantitative functional peptide data for immunology

*Martin J. Blythe, Irini A. Doytchinova and Darren R. Flower**

Edward Jenner Institute for Vaccine Research, Compton, Berkshire RG0 7NN, UK

Received on August 9, 2001; revised on October 2, 2001; accepted on October 18, 2001

ABSTRACT

Motivation: The compilation of quantitative binding data underlies attempts to derive tools for the accurate prediction of epitopes in cellular immunology and is part of our concerted goal to develop practical computational vaccinology.

Results: JenPep is a family of relational databases supporting the growing community of immunoinformaticians. It contains quantitative data on peptide binding to Major Histocompatibility Complexes (MHCs) and to Transmembrane Peptide Transporter (TAP), as well as an annotated list of T-cell epitopes.

Availability: The database is available via the Internet. An HTML interface allowing searching of the database can be found at the following address: http://www.jenner.ac.uk/JenPep.

Contact: darren.flower@jenner.ac.uk

INTRODUCTION

As the field of Bioinformatics has grown and matured into a new branch of science, new sub-disciplines have emerged within it. Immunoinformatics, the application of informatics and modelling techniques to molecules of the immune system, is one of the most exciting of these newly emergent sub-disciplines. One of the principal goals of immunoinformatics is to develop computer aided vaccine design, or computational vaccinology, and apply it to the quest for new vaccines. At the heart of computational vaccinology is the problem of epitope prediction. The focus of our present work is the development of a new database system in cellular, or T-cell, immunology.

A specialized type of immune cell mediates cellular immunity: the T-cell. These cells constantly patrol the body hunting for foreign proteins originating from pathogenic organisms such as viruses or bacteria. T-cells express a particular kind of receptor: the T-Cell Receptor (TCR), which exhibits a wide range of selectivities and affinities. TCRs bind to Major Histocompatibility Complex (MHC) proteins presented on the surfaces of other cells. These proteins bind small peptide fragments, or epitopes, derived from both host and pathogen proteins. It is recognition of such complexes that lies at the heart of both the adaptive, and memory, cellular immune response.

The overall process leading to the cell-surface presentation of epitopes, derived from antigenic protein, is complex and not yet fully understood. There are two main antigen presentation pathways: classes I and II. Class I MHCs are expressed by most nucleated cells, albeit with some exceptions. T-cells, whose surfaces are rich in CD8 co-receptor protein, recognize class I MHCs. Class II MHCs are only expressed on so-called 'professional antigen presenting cells' and are recognized by T-cells whose surfaces are rich in CD4 co-receptors. Class I peptides are typically, but not exclusively, derived from intracellular proteins, such as viruses. These proteins are targeted to the proteasome, which cleaves them into short peptides of 8–11 amino acids in length. These peptides are bound by the Transmembrane Peptide Transporter (TAP), which translocates them from the cell cytoplasm into the Endoplasmic Reticulum (ER), where they are in turn bound by MHC protein. For class II, receptor mediated ingestion of extracellular protein derived from a pathogen is targeted to an endosomal compartment where the proteins are cleaved by cathepsins, to produce peptides of 15–20 amino acids. Class II MHCs then bind these peptides. Peptide bound MHCs are presented on the surface of the cell where they are recognized, as T-cell epitopes, by T-cells. MHC proteins are polymorphic, each exhibiting slightly different peptide selectivities. The combination of MHC and TCR selectivities determines the power and scope of peptide recognition in the immune system and thus the recognition of foreign and self-antigenic peptides.

Experimental work has established that only peptides that bind with high affinity to MHC molecules are recognized as T-cell epitopes by TCRs (Sette *et al.*, 1994a,b). Weaker or non-binding peptides are simply not recognized. Expressed in terms of a competition assay, the IC_{50} must be less than 500 nM. IC_{50} values are binding affinities measured using a radioisotope-labeled reference peptide. Prediction of MHC binding is thus a pre-requisite to the prediction of T-cell epitopes. Most attempts to predict binding peptides have attempted

*To whom correspondence should be addressed.

to simplify the task by using a classification scheme, dividing peptides into non-binders, low affinity binders, medium affinity binders, or high affinity binders. Again, in terms of IC_{50} values: non-binders show no affinity, low binders > 500 nm, 500 nm > medium binders > 50 nm, and high binders < 50 nm. However, more recent work has turned to the development of fully quantitative models (Rognan *et al.*, 1999; Doytchiniva and Flower, 2001, 2002). To achieve this we must have access to a database of allele-specific quantitative binding data. It is only from data of this type that we can build statistically accurate models for the prediction of binding. To accurately model the process we need to focus on well characterized data for the binding of peptides to TAP and to MHCs, and their subsequent functioning as T-cell epitopes. Certain groups have access to some of these data, but currently there is no publicly available database or compilation. As part of our attempts to develop computational vaccinology, we have set about constructing such a database, which we have called JenPep. The following paper describes version 1.0 of this database.

SYSTEMS AND METHODS

Database size and structure

Version 1.0 of JenPep is composed of three component sub-databases: a compilation of quantitative measures of binding for peptides to classes I and II MHCs; a compendium of dominant and subdominant T-cell epitopes, and a similar set of quantitative data for peptide binding to TAP peptide transporter. This compilation was derived through exhaustive, semi-manual searching of the primary literature. We have used extensive searching of available literature databases, using keyword and author searches, retrospective searching, citation matching of key authors (particularly those describing the development of an assay system), to identify new papers detailing experimental quantitative measured values.

Data

The database is organized on the basis of peptides, which are defined by their sequence and length. A schematic of the database structure is included in Figure 1.

Peptide origin. Information on the origin of the peptide is taken from the reference paper and, failing that, from results obtained using BLAST (Altschul *et al.*, 1997). A hypertext link is made to the corresponding SWISS-PROT entry. The reference sequence is taken from that most closely matching the peptide as published.

Restriction allele. Information on the MHC restriction allele is given for all entries except those in the TAP database. MHC nomenclature has been standardized to the best of our ability. Sources of information involving various aspects of the HLA nomenclature can be found at WMDA [http://www.worldmarrow.org/dic99tab.html], the HLA Informatics Group [http://www.anthonynolan.org.uk/HIG/], the IMGT [http://imgt.cines.fr], and at the EBI [http://www.ebi.ac.uk/imgt/index.html]. A list of MHC restriction alleles, as contained within JenPep, is available on the JenPep website.

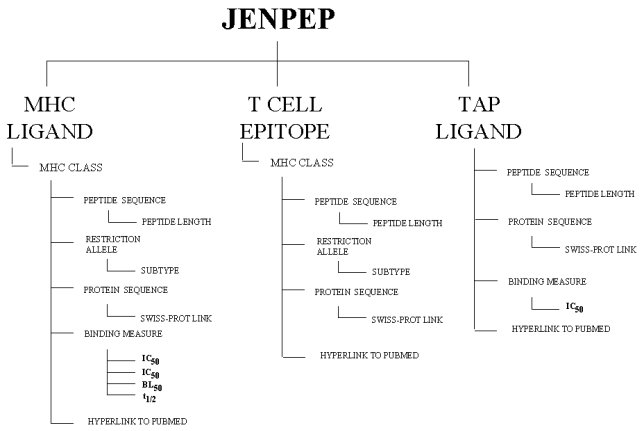


Fig. 1. Schematic of the database structure.

Peptide binding data. We have included measures of binding affinity calculated using a number of different methods. JenPep currently includes published data from the most commonly used methods for evaluating binding. IC_{50} values are binding affinity measures calculated from a competitive binding assay (Ruppert *et al.*, 1994; Sette *et al.*, 1994b; Sidney *et al.*, 1994). The value given is the concentration required for 50% inhibition of a standard labelled peptide by the test peptide. Therefore binding affinity is inversely proportional to the IC_{50} value. Reference peptides can be labelled fluorescently or with a radioisotope (see f and r in Table 1). Values calculated with these two methods are significantly different, making their direct comparison problematic, and are therefore presented separately. BL_{50} values are calculated in a peptide binding stabilization assay (Marshall *et al.*, 1994; van der Burg *et al.*, 1996; ten Bosch *et al.*, 1999). It is the half maximal binding level calculated from a Mean Fluorescence Intensity (MFI) of MHC expressing RMA-S cells. These cells are incubated with the test peptide and then labelled with a fluorescent monoclonal antibody. The nominal binding strength is again inversely proportional to the BL_{50} value. $t_{1/2}$ is the half-life for radioisotope labeled β_2-microglobulin disassociation from an MHC class I complex at 37 °C (Parker *et al.*, 1992, 1994; DiBrino *et al.*, 1994). The greater the half-life the stronger the peptide–MHC complex.

Table 1. Summary of MHC binding data and T-cell epitope data in JenPep version 1.0

	Class I	Class II	TAP
Number of peptides	4462	3447	432
IC_{50}[r]	2392	1753	432
IC_{50}[f]	188	789	
BL_{50}	342	110	
$t_{1/2}$	274	0	
T-cell epitopes	1266	795	
Alleles			
MHC	34	37	
T-cell	63	45	
Peptide lengths	7–16	9–35	7–15

A list of peptide data contained within JenPep version 1.0. Data for MHC binding, T-cell epitopes, and TAP binidng are given. Summaries are also provided for each type of quantitative binding data measured (i.e. radiolabeled or fluorescent IC_{50}, BL_{50}, or half-life ($t_{1/2}$)).

T-cell data. Data on T-cell epitopes within JenPep is currently limited to a list of binders. As with MHC binding, there are many different assays used to identify T-cell epitopes. These include T-cell killing, proliferation assays such as thymidine uptake, etc. The quantitative data produced by such assays, while interpretable, ie a peptide either is or is not a T-cell epitope, is not consistent enough to be used outside of the limited criteria for particular experimental conditions. Since there is no real meaning to the idea of a partial T-cell epitope we have decided to rely on the judgement of immunologists to define, as accurately as possible, what are, or are not, T-cell epitopes.

TAP data. Data on peptide binding to TAP, currently the smallest of our compendia is limited to radiolabelled IC_{50} data. As yet, TAP binding has not been studied as deeply as other areas of quantitative immunology.

Database system and graphical user interface

JenPep is a relational database system. It is constructed using MicroSoft ACCESS and is searchable through a Graphical User Interface (GUI) built using the Active Server Pages protocol. The size of the current database system is such that it is easily contained within such a system.

Each query entered must have the minimum number of amino acids as stated on each search page. The entered sequence will match any database entry that contains that sequence as a substring: class I binder query: LPSDYFP will match all longer peptides (FLPSDYFPSP, FLPSDYFPST, etc.). Results are presented as one database entry per page. Action buttons can be used to move from one entry to another. The number and variety of database fields presented vary upon which search type is used. All contain an entry number, restriction information, data on peptide origin, a journal reference, and a hyperlink to the corresponding abstract in PubMed. The frontpage of the web interface to JenPep is shown in Figure 2a, and a representative result page is given in Figure 2b.

IMPLEMENTATION

The current size of the JenPep database is relatively modest, in size, compared to the large, publicly curated, post-genomic sequence databases. It compares more favourably, perhaps, with the large number of disparate databases that have emerged in recent years (Baxevanis, 2001), which are often far smaller and more highly focussed. We are now at a stage, in the lifecycle of our product, similar to that reached by the SWISS-PROT or PIR databases several decades ago. Much useful data is still locked into the written, or hard-copy, literature, presented as tabulated values or in a graphical form. It is an on-going challenge to find and extract these data into a machine-readable format. A significant proportion of quantitative binding data remains as unpublished results written in laboratory notebooks or on conference posters. Nonetheless, we felt that our primary objective was to develop a working database schema and populate it with sufficient data for the database to be useful. Future work, as we probe further into the literature, will, we hope, see the database grow considerably in size and scope.

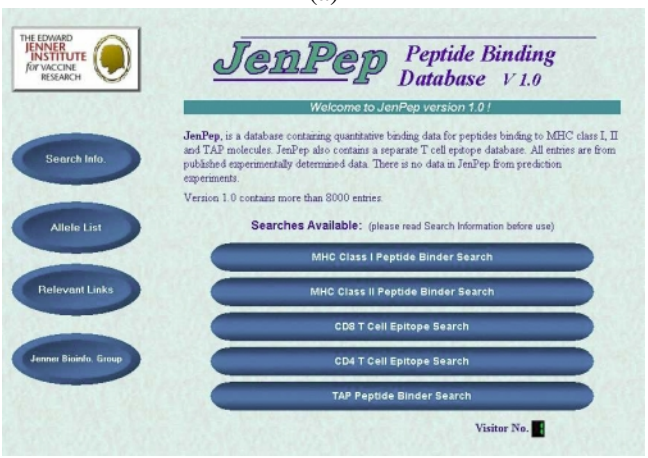
Version 1.0 of JenPep is composed of a compilation of quantitative measures of binding for 6000 peptides to classes I and II MHCs; a compendium of 2300 dominant and subdominant T-cell epitopes, and a set of over 400 quantitative data for peptide binding to TAP peptide transporter. JenPep contains binding data on a wide variety of different MHC alleles: for MHC class I, JenPep has data for 68 different restriction alleles with over 50 genotype variations. For class II MHC molecules there are over 40 restriction alleles with 52 genotype designations. Figures for the number of peptides binding to classes I and II MHCs are given in Table 1, which also lists the number of MHC restricted T-cell epitopes contained in the database. Peptide lengths for class I are in the range of 7–16 residues and for class II are in the range of 9–35 residues.

DISCUSSION

JenPep is a family of relational databases designed to support the growing community of immunoinformaticians. Version 1.0 contains quantitative data on the binding of peptide to MHC, and to TAP peptide transporter, and an annotated list of T-cell epitopes.

Immunological databases are not, however, novel. A number, concentrating on the exhaustive, in-depth sequence analysis of particular types of important im-

(a)



(b)

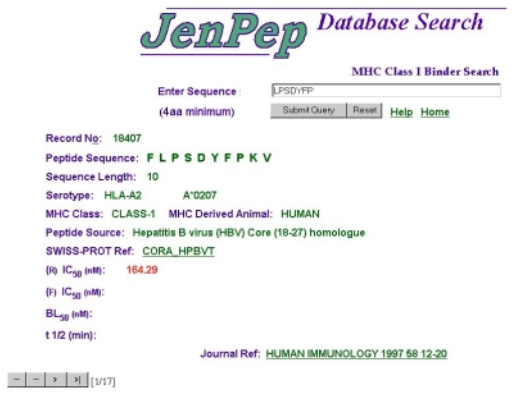


Fig. 2. The JenPep Web Interface. The front page (a), and a sample results page (b), for the JenPep web interface. The server is available over the INTERNET at the URL (http://www.jenner.ac.uk/JenPep).

munological biomacromolecule, have appeared over the last three decades (Brusic *et al.*, 2000). A few other databases focus on similar themes to our own. Probably the closest is the MHCPEP database developed by Brusic *et al.* (1998), which combines data on both T-cell epitopes and MHC binding, and lists about 13 000 peptides. In this database, and in contrast to our work, peptides are classified, in what appears to be a subjective manner, as high, medium, and low affinity binders and as non-binders. This database is now defunct. Subsequently, Brusic and co-workers have developed FIMM (Schonbach *et al.*, 2000). This is a much more complex database. It presents peptide data with a similar subjective classification of binding to that offered by MHCPEP. The SYFPEITHI database is a relatively up-to-date compendium of T-cell epitopes and MHC peptide ligands (Rammensee *et al.*, 1999). Like FIMM [http://sdmc.krdl.org.sg:8080/fimm/], SYFPEITHI is available via the World Wide Web [http://syfpeithi.bmi-heidelberg.com/]. The classification of MHC binding peptides offered by SYFPEITHI has, however, no quantitative dimension: only peptides that bind are listed and these are without any indication of binding strength.

Clearly, there is considerable overlap between these databases and ours. However, these databases do cover significantly different areas. Unlike all other immunological database systems we are aware of, JenPep contains quantitative binding data rather than subjective classifications, such as that used by MHCPEP. JenPep is also more up-to-date than MHCPEP, is contemporary with SYFPEITHI, but with a greater proportional coverage of recent years. JenPep also allows more direct access than databases such as FIMM and is more complete in the types of data given than SYFPEITHI. However, we are not trying to suggest that these databases are not, in their way, excellent. Indeed they are most worthy, but each has its own weaknesses, as well as its own strengths. We view our databases as complementing existing database systems. Together, the ability of all these databases to service the growing needs of the immunoinformatic community is greater with our database than without it.

As well as maintaining the database, that is keeping the database up to date as new papers are published, it is also our intention to continue the development of JenPep, extending both its size and scope, as we increase its coverage of cellular immunology data. We look forward to the day when immunologists are obliged to submit their experimental binding data to an on-line archive, such as ours, much as today, molecular biologists and protein crystallographers must submit their data to GenBank or the PDB. There is clearly a need to extend our existing databases, identifying more T-cell epitopes, and more data on TAP and MHC binding. It would also be interesting to extend JenPep to include non-amino acid ligands of MHC molecules, such as peptidomimetic compounds (Krebs and Rognan, 1998) and post-translational modifications of peptides (phosphorylated or glycosylated peptides, etc.) (Kastrup *et al.*, 2000; Zarling *et al.*, 2000). We would also like to cover the formation of the ternary complex of TCR, MHC, and peptide, where we would seek to gather together affinity data and kinetic data. We would also like to extend our coverage to peptide-related data from humoral immunology, specifically, the delineation of B cell epitopes. Projections on the foreseeable size of the database, particularly when new data types are added, suggest that we need to move away from ACCESS as our relational database system to a bespoke system capable of dealing with the particular requirements of our data.

CONCLUSION

Databases are the lingua franca of bioinformatics. The creation, manipulation, and use of databases containing

biologically relevant information is, perhaps, the most crucial feature of modern-day bioinformatics, both as a discipline in its own right and through its crucial support of post-genomic biological science. Such data is the currency in the information economy of modern biology. JenPep is the first database in immunology to concentrate on quantitative measurements and represents a complement to existing systems.

One of the primary motivations for constructing JenPep is to simplify the construction of quantitatively predictive QSAR models. The prediction of epitopes is vital to our goal of developing computational vaccinology. The accurate prediction of the immunogenicity of epitope, multi-epitope, or subunit vaccines, whether delivered as peptide or DNA, is a complex and unsolved problem. The prediction of MHC binding is a component of the process leading to T-cell activation, and probably forms the most selective filter along the natural epitope presentation pipeline. Binding data for TAP and MHC molecules will allow us, and other groups, to develop quantitatively predictive models, in our case based on QSAR techniques (Doytchiniva and Flower, 2001, 2002). The ability to predict MHC binding, for example, will enable us to analyze microbial genomes, identifying the most immunogenic proteins and thus selecting a set of favoured putative vaccines. As part of on going efforts, we would ultimately expect the development of computational vaccinology to have a similar effect on the search for new vaccines as informatics techniques have had on the search for new drugs.

ACKNOWLEDGEMENTS

We should like to thank Dr Valdimir Brusic for his helpful and encouraging comments and Mr Andrew Worth for his excellent technical assistance.

REFERENCES

Altschul,S.F., Madden,T.L., Schäffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Baxevanis,A.D. (2001) The molecular biology database collection: an updated compilation of biological database resources. *Nucleic Acids Res.*, **29**, 1–10.

ten Bosch,G.J., Kessler,J.H., Joosten,A.M., Bres-Vloemans,A.A., Geluk,A., Godthelp,B.C., van Bergen,J., Melief,C.J. and Leeksma,O.C. (1999) A BCR-ABL oncoprotein p210b2a2 fusion region sequence is recognized by HLA-DR2a restricted cytotoxic T-lymphocytes and presented by HLA-DR matched cells transfected with an Ii(b2a2) construct. *Blood*, **94**, 10 381–10 387.

van der Burg,S.H., Visseren,M.J., Brandt,R.M., Kast,W.M. and Melief,C.J. (1996) Immunogenicity of peptides bound to MHC class I molecules depends on the MHC-peptide complex stability. *J. Immunol.*, **156**, 3308–3315.

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

Brusic,V., Zeleznikow,J. and Petrovsky,N. (2000) Molecular immunology databases and data repositories. *J. Immunol. Methods*, **238**, 17–28.

DiBrino,M., Parker,K.C., Shiloach,J., Turner,R.V., Tsuchida,T., Garfield,M., Biddison,W.E. and Coligan,J.E. (1994) Endogenous peptides with distinct amino acid anchor residue motifs bind to HLA-A1 and HLA-B8. *J. Immunol.*, **152**, 620–629.

Doytchiniva,I.A. and Flower,D.R. (2001) Towards the quantitative prediction of T-cell epitopes: CoMFA and CoMSIA studies of peptides with affinity to class I MHC molecule HLA-A*0201. *J. Med. Chem.*, **44**, 3572–3581.

Doytchiniva,I.A. and Flower,D.R. (2002) Physicochemical explanation of peptide binding to HLA-A*0201 major histocompatibility complex. A three-dimensional quantitative structure–activity relationship study. *Proteins*, submitted.

Kastrup,I.B., Stevanovic,S., Arsequell,G., Valencia,G., Zeuthen,J., Rammensee,H.G., Elliott,T. and Haurum,J.S. (2000) Lectin purified human class I MHC-derived peptides: evidence for presentation of glycopeptides *in vivo*. *Tissue Antigens*, **56**, 129–135.

Krebs,S. and Rognan,D. (1998) From peptides to peptidomimetics: design of nonpeptide ligands for major histocompatibility proteins. *Pharm. Acta Helv.*, **73**, 173–181.

Marshall,K.W., Liu,A.F., Canales,J., Perahia,B., Jorgensen,B., Gantzos,R.D., Aguilar,B., Devaux,B. and Rothbard,J.B. (1994) Role of the polymorphic residues in HLA-DR molecules in allele-specific binding of peptide ligands. *J. Immunol.*, **152**, 4946–4953.

Parker,K.C., DiBrino,M., Hull,L. and Coligan,J.E. (1992) The beta 2-microglobulin dissociation rate is an accurate measure of the stability of MHC class I heterotrimers and depends on which peptide is bound. *J. Immunol.*, **149**, 1896–1903.

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**, 163–170.

Rammensee,H., Bachmann,J., Emmerich,N.P., Bachor,O.A. and Stevanovic,S. (1999) SYFPEITHI: database for MHC ligands and peptide motifs. *Immunogenetics*, **50**, 213–219.

Rognan,D., Lauemoller,S.L., Holm,A., Buus,S. and Tschinke,V. (1999) Predicting binding affinities of protein ligands from three-dimensional models: application to peptide binding to class I major histocompatibility proteins. *J. Med. Chem.*, **42**, 4650–4658.

Ruppert,J., Sidney,J., Celis,E., Kubo,R.T., Grey,H.M. and Sette,A. (1994) Prominent role of secondary anchor residues in peptide binding to HLA-A2.1 molecules. *Cell*, **74**, 929–934.

Sakaguchi,T., Ibe,M., Miwa,K., Yokota,S., Tanaka,K., Schonbach,C. and Takiguchi,M. (1997) Predominant role of N-terminal residue of nonamer peptides in their binding to HLA-B* 5101 molecules. *Immunogenetics*, **46**, 245–255.

Schonbach,C., Koh,J.L., 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., Dahlberg,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**, 813–820.

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.

Shiga,H., Shioda,T., Tomiyama,H., Takamiya,Y., Oka,S., Kimura,S., Yamaguchi,Y., Gojoubori,T., Rammensee,H.G., Miwa,K. and Takiguchi,M. (1996) Identification of multiple HIV-1 cytotoxic T-cell epitopes presented by human leukocyte antigen B35 molecules. *AIDS*, **10**, 1075–1085.

Sidney,J., Oseroff,C., del Guercio,M.F., Southwood,S., Krieger,J.I., Ishioka,G.Y., Sakaguchi,K., Appella,E. and Sette,A. (1994) Definition of a DQ3.1-specific binding motif. *J. Immunol.*, **152**, 4516–4523.

Sobao,Y., Tsuchiya,N., Takiguchi,M. and Tokunaga,K. (1999) Overlapping peptide-binding specificities of HLA-B27 and B39: evidence for a role of peptide supermotif in the pathogenesis of spondylarthropathies. *Arthritis Rheum.*, **42**, 175–182.

Zarling,A.L., Ficarro,S.B., White,F.M., Shabanowitz,J., Hunt,D.F. and Engelhard,V.H. (2000) Phosphorylated peptides are naturally processed and presented by major histocompatibility complex class I molecules *in vivo*. *J. Exp. Med.*, **192**, 1755–1762.