JenPep: A Novel Computational Information Resource for Immunobiology and Vaccinology

Helen McSparron, Martin J. Blythe, Christianna Zygouri, Irini A. Doytchinova, and Darren R. Flower*

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

Received January 7, 2003

JenPep is a relational database containing a compendium of thermodynamic binding data for the interaction of peptides with a range of important immunological molecules: the major histocompatibility complex, TAP transporter, and T cell receptor. The database also includes annotated lists of B cell and T cell epitopes. Version 2.0 of the database is implemented in a bespoke postgreSQL database system and is fully searchable online via a perl/HTML interface (URL: http://www.jenner.ac.uk/JenPep).

INTRODUCTION

From the perspective of human disease, a proper understanding of the immune system is vital. Indeed, the immune system has evolved to combat the threat posed by pathogen mediated infection as well as playing an equally pivotal role in other human pathologies, such as allergy, autoimmunity, and cancer. We can do battle with the menace of disease by challenging our immune systems in an appropriate way: this is the concept that underlies the action of vaccines. One of the principal goals of Immunoinformatics,[1] the application of molecular informatic techniques to the immune system, is to develop computational vaccinology, or computer aided vaccine design (CAVD), and apply it in the search for new vaccines. At the heart of computational vaccinology lies the problem of characterizing and predicting the immunogenicity of proteins, as mediated through antigenic determinants or epitopes.

Epitope, at least as it is understood within the immunological and general bioscience literature, is a broad and ill-defined term. Put at its most simplistic, an epitope is any molecular structure that can be recognized by the immune system. Epitopes can be protein, carbohydrate, lipid, or nucleotide. It is through recognition of foreign, or nonself, epitopes that the immune system can identify and, hopefully, destroy pathogens. Hitherto, peptide epitopes have been the best studied and have, traditionally, been categorized as either T cell or B cell epitopes. T cell epitopes are peptides presented to the cellular arm of the immune system. A conformational B cell epitope is composed of one, or more, regions of whole, folded proteins recognized by soluble or membrane bound antibody molecules. Linear B cell epitopes are short peptides that are cross-reactive with conformational epitopes. Putting aside issues such as delivery mechanism or the choice of adjuvants, broadly speaking, vaccines can be grouped together as attenuated pathogens (whole microorganisms which have lost virulence but retained immunogenicity), subunit vaccines (whole protein immunogens), or polyepitope (one or more antigenic epitopes linked together). Historically, attenuated vaccines have been the most successful, but modern immunovaccinology is increasingly turning its attention toward epitope vaccines, which can be designed rationally and offer potential improvements in specificity and safety.[2]

We have developed the JenPep database as an aid to rational vaccine design. Here we outline the continuing development of this system. JenPep is a quantitative database characterizing the thermodynamics of peptide binding as well as focusing on the amino acid identity of epitopes. Experimental studies indicate that only peptides that bind with high affinity to MHC molecules are recognized as T cell epitopes,[3] with weaker or nonbinding peptides seldom being recognized. It would be foolish to ignore the beneficial insights that molecular and physicochemical analysis may give into immunological mechanisms. A significant key to this would be the ability to access a database of immunologically relevant quantitative thermodynamic and kinetic binding data. From such a database, one could, for example, build statistically accurate models for predicting MHC-peptide binding as well as modeling other immunological molecular recognition events. As no such compilation is currently available, we have previously constructed and have now greatly expanded such a database.[4] In this paper we describe version 2.0 of JenPep.

Immunological databases are not, however, without precedent. Several concentrating on the exhaustive, in-depth sequence analysis of particular types of important immunological biomacromolecule have existed for some time.[5] A few other databases focus on themes similar to our own. Arguably, the closest is the now defunct MHCPEP database developed by Brusic.[6] This combines both T cell epitope and MHC binding data. MHCPEP employs a widely used conceptual simplification, operating as a de facto data fusion device, which combines sets of distinct binding measures. It classifies peptides as High, Medium, Low, or Nonbinders, using the following schema: Nonbinders > 10 μM, 10 μM > Low Binders > 100 nM, 100 nM > Medium Binders > 1 nM, High Binders < 1 nM. Although such broad classifications may take account of experimental inaccuracy, they are intrinsically subjective. Subsequently, Brusic and co-workers have developed FIMM.[7] This is a much more complex and sophisticated database, but retains the same

* Corresponding author phone: 44 (0) 1635 577954; fax: 44 (0) 1635 577901/577908; e-mail: darren.flower@jenner.ac.uk.

10.1021/ci030461e CCC: $25.00 © 2003 American Chemical Society
Published on Web 06/14/2003

Table 1. Publically Accessible Functional Immunology Databases[a]

database	URL	reference
FIMM	sdmc.krdl.org.sg:8080/fimm/	7
SYFPEITHI	syfpeithi.bmi-heidelberg.com	9
JENPEP	www.jenner.ac.uk/JenPep	4
MHCBN	www.imtech.res.in/raghava/mhcbn/	8
MHCPEP	wehih.wehi.edu.au/mhcpep/	6
HIV Database	hiv-web.lanl.gov/content/immunology/ index.html/index.html	10

[a] A set of immunology databases focusing on peptide orientated functional data. Each database is available free via the Internet from the indicated URL. The reference shown is to the citation given in the main text.

subjective classification of binding. MHCBN[8] is a database system similar in concept to MHCPEP and FIMM, again focusing primarily on T cell statistics. It does include some quantitative data, although this is primarily parenthetical, it being present as comments rather than as explicitly searchable data. The SYFPEITHI database is an up-to-date and useful compendium of T cell epitopes and MHC peptide ligands.[9] The HIV Molecular Immunology database[10] is a high quality database with a scope, at least in terms of data archived, broader than others, containing lists of MHC binding motifs and ligands as well as B and T cell epitopes. Albeit, in the limited, if very extremely important, context of a single viral species. However, SYFPEITHI and the HIV Molecular Immunology database have no quantitative dimension to their classification of MHC binding: it lists only peptides that bind without record of measured binding affinity. One should note that many of the systems described above are available via the World Wide Web (see Table 1).

As a preliminary to further, and more complete, discussion of JenPep, we present a concise primer to mechanisms involved in the presentation and recognition of antigen within the immune system. We will follow this exordium with an exploration of the computational and conceptual structure of JenPep and then discuss the next steps in the continuing development of the database. Our emphasis in this paper is on the data content, that is the underlying chemical biology, rather than a lengthy discussion of the database structure orientated toward computer scientists. In so doing we seek to make our paper relevant and accessible to chemists, biologists, and immunologists rather than computer scientists.

JENPEP IN CONTEXT: A BRIEF PRIMER ON ANTIGEN PRESENTATION AND RECOGNITION

At the heart of our attempts to design vaccines rationally is the need for a fundamental understanding of immunobiological mechanisms. The manifestation of immunology at the level of the whole animal is, however, an exceedingly complex and hierarchical phenomenon, exhibiting much emergent behavior. Historically, and operationally, the immune system has been thought to divide into two distinct responses, one mediated by cells—cellular immunity—and one by soluble factors, the so-called humoral immunity.

We shall begin by focusing on that aspect of the adaptive immune response that is mediated by cells. A specialized type of immune cell mediates cellular immunity: the T cell, which patrols the body searching out proteins that have a viral, bacterial, fungal, or parasite origin. The cell surface membrane of T cells is enriched in the T cell receptor (TCR), which functions by binding to major histocompatibility complex proteins (MHCs) expressed on the surfaces of other cells. These proteins, in turn, bind small peptide fragments derived from both host and pathogen proteins. It is the recognition of such complexes that lies at the heart of the cellular immune response. Immunologists refer to short peptides such as these as epitopes. The overall process leading to the presentation of antigen-derived epitopes on the surface of cells is a long, complicated, and not yet fully understood story. MHCs fall into two structurally distinct groups, each associated with a distinct presentation process, Class I and Class II. Class I MHCs are expressed by almost all cells in the body. They are recognized by T cells whose surfaces are enriched in CD8 coreceptor protein, so-called CD8+ T cells. Class II MHCs are only expressed on a special subset of cells—professional antigen presenting cells (APCs)—and are recognized by CD4+ T cells. The two classes of MHC molecule are distinct, presenting antigen from different sources. Class I molecules present endogenously synthesized or intracellular protein and class II presents exogenously derived or extracellular proteins. The cell biology and expression of each type of MHC is tailored to address these different functions.

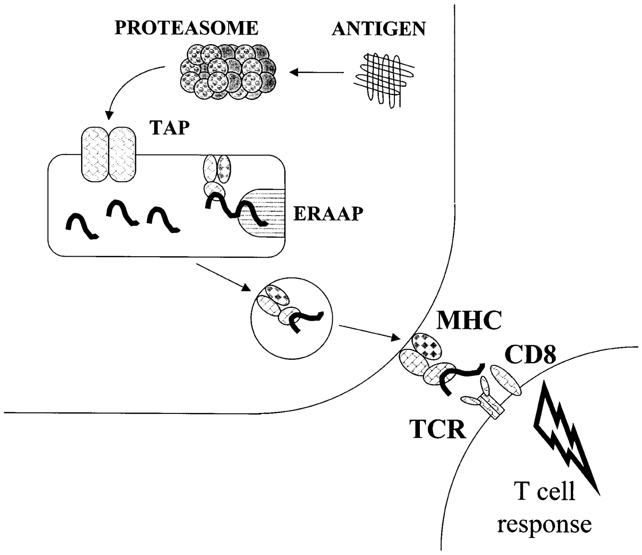

Figure 1. Principal Class I presentation pathway. Class I peptides are primarily derived from viral and other cytosolic proteins, including host, or self, proteins. Ubiquitinylation targets these proteins to the proteasome, which cleaves them proteolytically into short peptides of 8 to 18 amino acids in length. These peptides are then bound by TAP (transporter associated with antigen presentation), which translocates them from the cytoplasm to the endoplasmic reticulum (ER). In the ER, peptides are bound by MHCs, which then passed through the golgi and traffic to the cell surface via exocytic vacuoles. Cell surface MHCs are recognized by TCRs on CD8+ T cells, evoking a subsequent activation of the T cell.

Class I presentation is characterized by a variety of degenerate pathways, but we will consider here only the most important and, consequently, the best understood.[11,12] See Figure 1. Class I peptides are primarily derived from viral and other cytosolic proteins, including host, or self, proteins. Ubiquitinylation targets these proteins to the proteasome, which cleaves them proteolytically into short peptides of 8 to 18 amino acids in length.[13] These peptides are then bound by TAP (Transporter associated with antigen presentation), a transmembrane ATP-binding cassette transporter, which

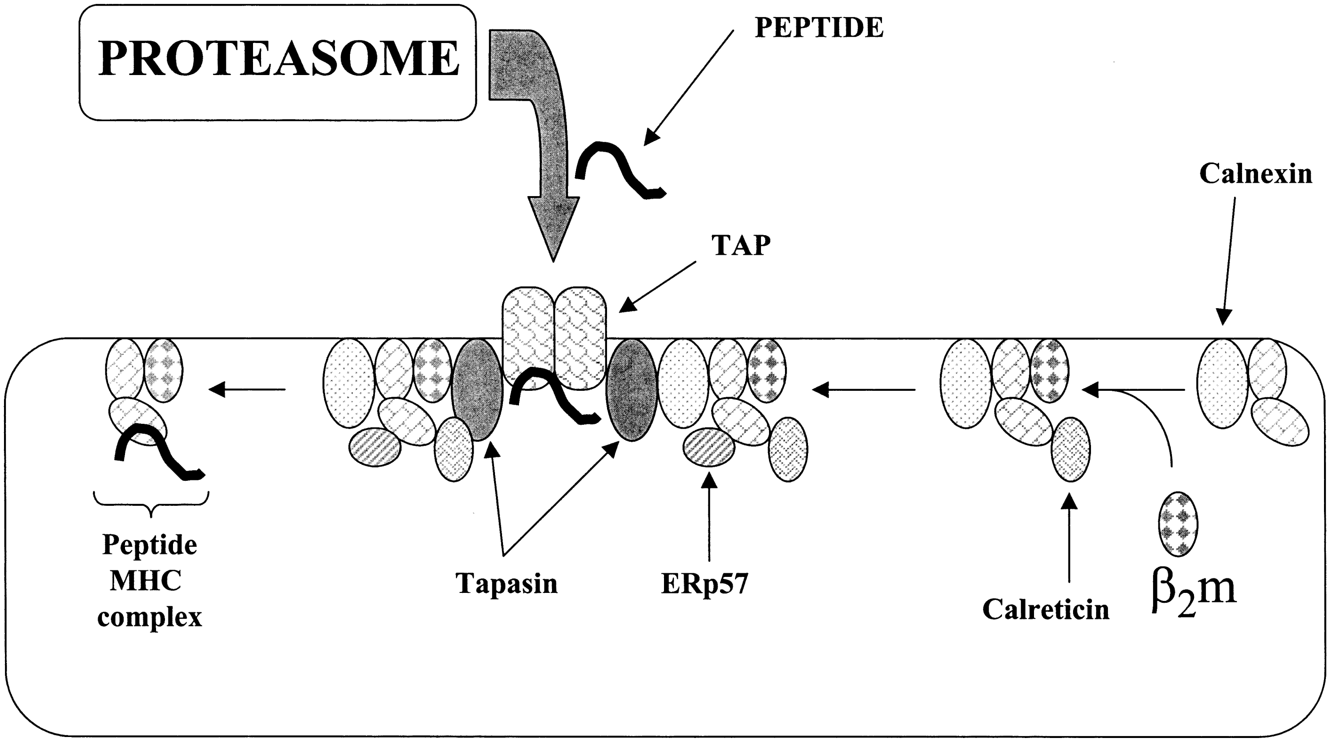

Figure 2. Peptide translocation by TAP. Peptides produced by the proteasome are bound by TAP (transporter associated with antigen presentation), a heterodimer of TAP1 and TAP2, a member of the transmembrane ATP-binding cassette transporters. This protein pump then translocates the peptides to the endoplasmic reticulum (ER) where they are bound by MHCs. Assembly of the MHC-peptide complex is facilitated by a set of chaperones. Free MHC proteins are initially bound by calnexin, which exchanges with other proteins, including β_2 microglobulin, to form a supramolecular complex of MHC, ERp57, and calreticulin. This complex then associates, together with tapasin, to the TAP dimer, prior to nucleotide-driven unidirectional peptide transport.

translocates them from the cytoplasm to the endoplasmic reticulum (ER), where they are bound by MHCs.[14] See Figure 2. It is now increasingly clear however that several other processing routes, including TAP-independent Trojan antigen presentation,[15] act to complicate this simple picture. For example, ERAAP (ER associated aminopeptidase associated with antigen processing), a member of the family of zinc metalloproteases, has been shown to trim peptides cleaved by the proteosome within the ER,[16] as has various other proteases, such as furin.[17] Thus, ultimately the accurate prediction of class I processing will need to rely on a much more comprehensive and integrated modeling of the entire multistep process rather than on individual models of one or more subsidiary stages.

For Class II, receptor mediated internalization of extracellular protein derived from a pathogen is targeted to an acidic endosomal compartment, where proteins are cleaved by cathepsins, a particular variety of protease, to produce longer peptides of 15−20 amino acids. See Figure 3. Unlike Class I MHCs (whose binding sites are closed at both ends and therefore bind a repertoire of peptides restricted in length), Class II MHCs binding sites are open at both ends, and they are thus also able to bind longer peptides.

Peptide bound MHCs (or peptide−MHC (pMHC) complexes) are recognized by receptors on the surface of T cells, so-called TCRs. Many other coreceptors and accessory molecules, in addition to CD4 and CD8 molecules, are also involved in T cell recognition. The recognition process is by no means simple and remains poorly understood. Nonetheless, it has emerged that the process involves the formation of the so-called immunological synapse,[18] a highly organized, spatio-temporal arrangement of receptors and accessory molecules of many types.

T cells that can kill other cells are called cytotoxic T lymphocytes or CTL. Most CTL are MHC class I-restricted CD8+ T cells, but CD4+ T cells can also kill cells under certain conditions. So-called helper CD4+ T cells assist B cells to make antibody in response to antigenic challenge and exist as a set of subtypes. T_H1 cells are a subset of T cells that are characterized by their cytokine expression profile and are mainly involved in activating macrophages. Another subset of helper T cells is the T_H3 cells, which produce transforming growth factor-beta in response to antigen. The most efficient subset of helper T cells is, however, T_H2 cells. They produce cytokines, primarily interleukins 4 and 5, which stimulate B cells to produce antibody.

As we have said, pMHCs are recognized by TCRs on the surface of T cells. In order for this to occur, the antigen must have two distinct interaction sites: one, the epitope, interacts with the TCR, and the other, called the agretope, must interact with an MHC molecule. The formation of such ternary complexes is the molecular recognition event at the heart of the adaptive and memory cellular immune responses. MHCs exhibit extreme polymorphism. Within the human population there are, at each genetic locus, a great number of genetic variants—currently in excess of 1200—known as allelic products or alleles, many represented at high frequency (>1%).[19] MHC alleles may differ by as many as 30 amino acid substitutions. Such a remarkable degree of polymorphism implies a selective pressure to establish and maintain it. Different polymorphic MHC alleles, of both Class I and

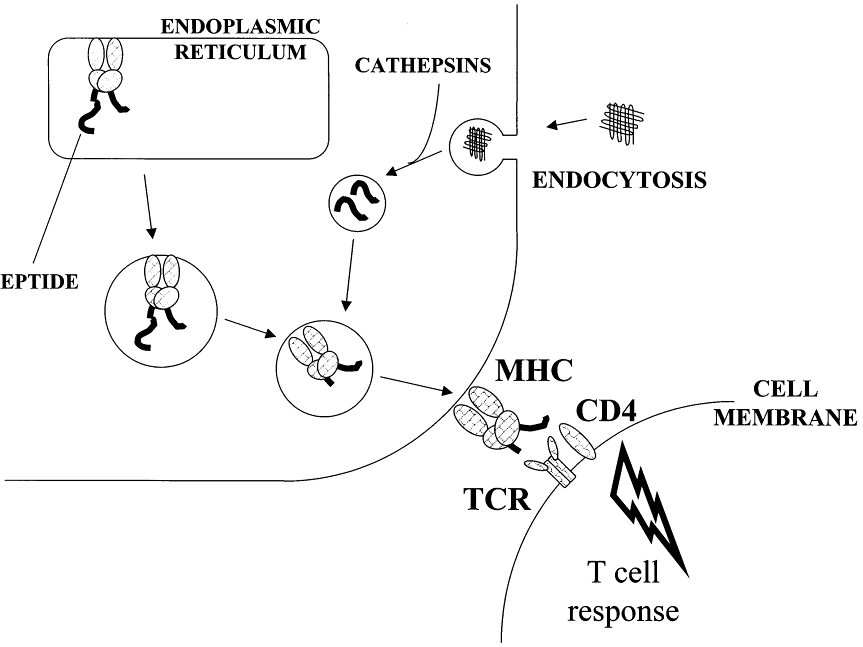

Figure 3. Class II presentation pathway. Receptor mediated internalization of pathogen derived extracellular protein is targeted to an acidic endosomal compartment, where they are cleaved by cathepsins, a particular kind of protease, to produce peptides of 15−20 residues. In the ER, newly synthesized class II MHCs bind a polypeptide called invariant chain or Ii. This serves a dual purpose, first by blocking peptide binding to class II molecules and also by targeting the MHC to a specialized endosomal compartment (MIIC). Membrane vesicle bound external antigen enters the cell by either receptor-mediated endocytosis or fluid phase pinocytosis and is targeted to the MIIC. Here both antigen and li are degraded, the former generating large numbers of antigenic peptides, which are then bound by the MHC prior to trafficking to the cell surface.

Class II, have different peptide specificities: that is to say that they bind peptides with particular sequence patterns. This has led to the development of so-called motifs. Each peptide motif is characterized by the position and occupancy of anchor residues whose side chains protrude into complimentary pockets within the peptide binding groove on the respective MHC molecule.[20] Polymorphic MHC residues cluster in these pockets and in other secondary pockets which are also important in determining binding specificity. A more accurate description of the phenomenon of specificity, which escapes the limitation of the peptide motif so-beloved of immunologists, is to say that MHCs bind peptides with an affinity dependent on the nature of the bound peptide's sequence.

Within a population there are an enormous range of different, variant genes coding for MHC proteins. In man and mouse, as in most species, each MHC class is represented by more than one locus. The class I loci are HLA-A, HLA-B, and HLA-C and the class II loci HLA-DR, HLA-DQ, and HLA-DP. All MHC loci are codominant: both maternally and paternally inherited sets of alleles are expressed. The set of linked MHC alleles found on the same chromosome is known as a haplotype. T cell receptors, in their turn, also exhibit different affinities for pMHC. The combination of MHC and TCR selectivities thus determines the power of peptide recognition in the immune system and thus the recognition of foreign proteins and pathogens.

The accurate prediction of T cell epitopes remains problematic, and because of this, recent theoretical work has largely focused on the prediction of peptide binding to MHC molecules.[21] The prediction of MHC binding is probably both the best understood and the most discriminating step in the presentation-recognition pathway within cellular immunity. The stability, and thus longevity, of MHC-peptide complexes is of general interest because of its fundamental role in the regulation of T cell activation.

Turning now to the humoral response, antibodies are immunological globular proteins secreted by B lymphocyte cells. They are found on serum, lymph, and mucosal surfaces and are a key component in the adaptive immune response of higher vertebrates. B cell-antigen binding occurs via a B cell receptor (BCR)-epitope complex. When accompanied by an appropriate helper T cell response it results in B cell differentiation into an antibody secreting plasma cell. This ultimately results in a B cell-mediated immune response against the pathogen. For B cells the antigen receptor is a cell surface antibody rather than a TCR. Each B cell expresses a single type of BCR. Antibodies can mediate protection from pathogens in several ways. Possibly the most direct route involves forming a high affinity interaction with a pathogen or its products, thus preventing their access to cells. This is called neutralization and is particularly important in protecting against viruses or soluble bacterial toxins. Other important antibody mediated processes include opsonization (the coating of pathogenic cells leading to enhanced recognition and ingestion by phagocytes) and the activation of complement.

In an analogous manner to T cell mediated immunity, the ability to identify linear and conformational B cell epitopes on proteins is of fundamental importance in the development of synthetic peptide vaccines. A B cell epitope (also known as an antigenic site or determinant) may be formed from RNA, DNA, protein, polysaccharides, or glycoproteins. Epitopes recognized in an in vivo immune response must be accessible to antibody binding. An antibody raised against a pathogen tends to bind to epitopes located on the protein surface. However, it is now generally accepted, at least conceptually, that the entire surface of a pathogen acts as one all-encompassing antigenic site. Virtually any part of a globular protein is capable of stimulating the production of particular antibodies under appropriate conditions.

However, the frequency of protein-antibody recognition is not uniform across its surface. Regions that bind a greater proportion of antibodies than other surface areas during a normal in vivo immune response are termed major antigenic sites. These sites may be considered immunodominant, since antibody-producing B-cell clones specific for these sites are stimulated to proliferate more readily. Globular proteins may possess more than one immunodominant site, up to two to three. Predicting the amino acid sequence of protein regions most likely to be recognized by antibodies would enable corresponding peptides to be synthesized facilitating antibody recognition of the naïve epitope.

DATABASE IMPLEMENTATION

Because of its modest size, Version 1.0 of JenPep was constructed using MicroSoft ACCESS as the database engine and could be searched via a graphical user interface (GUI) built using the Active Server Pages protocol (ASP). The growth in the size and scope of the database has required that version 2.0 of JenPep be implemented in a bespoke system using open source postgreSQL[22] as the database engine and a GUI written in perl/HTML. Within postgreSQL, a series of tables was created to accommodate the stored data. The B cell Epitope data are stored in three separate tables (ab_detail, ab_notes, ab_papers), related by the id number, which serves as the primary key. The T Cell Epitope, MHC Ligand, and TAP Ligand data are stored in a larger, generic table: tc_detail. Finally, the TCR−MHC protein complex data are stored in a separate table: mhccmplx_detail. The structure of these tables is shown in Table 2. The GUI is less restrictive than in version 1.0, including a sequence search that allows substring searches and the use of wild card characters.

Data collation has continued to involve exhaustive, semimanual searching of the primary literature. In the development of version 2.0 we have incorporated data from new papers as they are published. Moreover, we have supplemented this by continued searching of available databases containing immunological literature. This has involved the use of author name and keyword searches, prospective searching of cited papers, and citation matching of authors from papers describing novel assay development. The range of information incorporated into JenPep has been expanded to include new data types, such as pMHC−TCR interactions and B cell epitopes, and we have also addressed the issue of extracting data presented in a graphical form using the utility Ungraph,[23] a tool for converting images of graphs into numerical values. We have used this to extract binding measures from bar charts and other nontabulated graphs.

It remains the case however that we cannot be certain how much data remain to be collated. A considerable amount of useful data are still locked into the hardcopy literature, and it is an ongoing challenge to find and extract these data into a machine-readable format. A significant proportion of quantitative binding data remains unpublished or as conference posters or in laboratory notebooks. We can look forward to the day when immunologists submit their experimental binding data to an online archive much as today those involved in genomic sequencing are obliged to submit their data to GenBank or other online archive. In the meantime, we have added a deposition form to our interface allowing experimentalists to add data directly into JenPep. We should like to extend an invitation to all experimental immunologists to begin this process, submitting their raw binding data, either before or after publication, to us for incorporation into JenPep.

Table 2. Internal Structure of the JenPep Database[a]

Data Category: The B Cell Epitope

ab_detail	ab_notes	ab_papers
id	id	id
location	note	author_ref
weau_loc		url
mab		author
neutralizing		journal
epitope		title
immunogen		volume
species_iso		pages
donar		year
		booktitle
		publisher
		address
		abstract

Data Categories: T Cell Epitope, MHC Ligand, and TAP Ligand

tc_detail	tc_detail	tc_detail
entry	swiss_db_ref	c50s
peptide_cat	sp_hyperlink	tm
epitope	nonbinder	ka
seq_length	ungraphed	kd
allele	ic50	ric
subtype	ic50f	ec50
mhc_class	bl50	comments
mhc_species	thalf	journal_ref
pep_desc	sc50	pub_med
category		

Data Category: TCR−MHC Protein Complexes

mhccmplx_detail	mhccmplx_detail	mhccmplx_detail
ENTRY_NO	MHC_DERIVED	K_on
PEP_CAT	PEP_DESC	K_off
EPITOPE	CATEGORY	KA
SEQ_LENGTH	SWISS_DB_REF	THALF
ALLELE	SP_HYPERLINK	TEMP
SUBTYPE	DISSOC_CONS	COMMENTS
MHC_CLASS	AFFINITY	REFERENCE
TCR_REL_INFO	KD	PUBMED
STRUC_SUM	EC50	

[a] The names of tables used within the five subdatabases within JenPep: T cell epitope, MHC binding, TAP binding, pMHC−TCR database, and B cell epitope. The columns within the separate sections are also given.

DATABASE CONTENT

Version 1.0 of JenPep was composed of three compendia: compilations of T cell epitopes and quantitative measures of peptide binding to TAP and to Class I and Class II MHCs. Version 2.0 has expanded this to include pMHC−TCR interactions and linear B cell epitopes. The database is organized on the basis of peptides defined by their sequence and length. The number and size range of peptides in each of these five categories are listed in Table 3.

The content of the database is either extracted directly from the literature or is generated by searching, or by making reference to, other online data repositories. A considerable

Table 3. Summary of the Characteristics of Epitope Data Contained with JenPep[c]

peptide class	total no. of peptides	length[a] distribution	class i[b]	length[a] distribution	class ii[b]	length[a] distribution
Tap transporter	441	7−15				
MHC binding	12336	4−28	6411	4−23	5925	7−28
TCR−pMHC	49	8−20				
T cell epitope	3218	7−35	2060	7−24	1158	8−35
B cell epitope	816	3−47				

[a] Range in amino acids. [b] Number of peptides. [c] The number and class and length distributions for the five classes of epitode and binding data contained within JenPep.

quantity of the information we have collated in JenPep is, essentially, generic data: it is the same irrespective of data category. For each entry, for example, we record the peptide sequence (e.g. *YLDDPDLKY*) of the epitope using the standard one-letter code, its length (9 in this case), and, through a link to the sequence database SWISS-PROT, the antigen to which the peptide sequence most closely matches (in the case of *YLDDPDLKY − DNA (cytosine-5)-methyltransferase 1, SWISS-PROT code P26358*). The description of the antigen from which the peptide is derived is, wherever possible, obtained directly from the literature. There are occasions, for example, when the peptide is synthetic, when this information cannot be provided. Sequence searching is used to identify an appropriate database link. The principal drawback here is that, due to their short length, the same peptide sequences can be found in a number of different potential antigens—orthologues, paralogues, or even in completely unrelated proteins. This is not a significant problem for MHC binding but does pose a dilemma for epitopes. Binders are essentially independent of sequence context: a peptide either binds or its does not. Epitopes are, however, processed from whole proteins via a complex processing pathway, as described in earlier sections. It is possible to use the sequence context of a particular epitope to deduce preferred cleavage patterns of the proteasome or endoplasmic protease, but only if this context is correctly defined. Likewise, wrongly identifying particular proteins as antigens can lead to the percolation of annotation errors, assuming that JenPep is used subsequently to assign the antigenic status of proteins. As JenPep is not currently used for either purpose, this is not an issue of current concern, but we will have to remain aware of this possibility.

We have also implemented a classification scheme, categorizing peptides into simple class (self-peptides, viral, bacterial, cancer, etc.) related to the purpose of the original experiment or the origin of the antigen. JenPep also links to the PUBMED citation of the paper from which the recorded data were derived (for YLDDPDLKY: *J. Immunol. 1994, 152, 3913−3924, PUBMED ID 8144960*). For the T cell epitope, MHC ligand, and TCR−pMHC complex categories, we also record, for each peptide, the MHC restriction in terms of the host species, class (class I vs class II), and allele. For YLDDPDLKY, these data would be *human*, *class I*, and *HLA-A*0101*. As far as the variable ways of naming alleles inherent in the primary literature permit us, we present MHC nomenclature standardized to the best of our ability. The primary resource for HLA nomenclature is the HLA Informatics Group [http://www.anthonynolan.org.uk/HIG/]. The naming of an allele follows a defined pattern:[19] for HLA-A*0101, the HLA-A refers to the HLA locus, the initial 01 to the group of alleles which encode the serologically recognized A1 antigen, and the final 01 to the individual HLA allele protein sequence. The nomenclature has recently been extended to include null sequences, synonymous mutations, and mutations outside the coding region. In JenPep we store the antigen classification (i.e. HLA-A1) and, where available, the specific allele. Data on null sequences and synonymous mutations do not affect peptide binding and we omit them. We often encounter problems with nonstandardization in the reporting of alleles. While a four-digit HLA name implies the two-digit antigen type, a two-digit classification clearly does not imply a specific allele. Other rich sources of information regarding HLA nomenclature are available at the IMGT [http://imgt.cines.fr; http://www.ebi.ac.uk/imgt/index.html] and at WMDA [http://www.worldmarrow.org/ dic99tab.html]. JenPep contains data on a wide variety of MHC alleles: for MHC class I, JenPep contains data for over 70 class I alleles and for over 40 class II alleles.

Compared with version 1.0, JenPep now incorporates more forms of binding measurement. These include equilibrium constants, which cover true association (K_A) and dissociation constants (K_D), as well as radiolabeled and fluorescent IC_{50} values that approximate equilibrium binding constants under suitable conditions. Other types of measurement include BL_{50} values, together with closely related SC_{50}, EC_{50}, and C_{50} values, as calculated in a peptide binding stabilization assay,[24] and T_m values (the temperature at which 50% of MHC protein is denatured). We also record β_2-microglobulin dissociation half-life,[25] which is strictly a kinetic measurement, but one believed, at least by immunologists, to correlate well with binding affinity.

These different measures form a hierarchy, with equilibrium constants, when calculated correctly, being the most reliable and accurate. Peptide binding to MHC molecules can be quantified as one would quantify any other biomolecular receptor−ligand interaction

$$\mathrm{R} + \mathrm{L} \leftrightarrow \mathrm{RL}$$

where R is the receptor, L the ligand, and RL the receptor−ligand complex. Such interactions frequently obey the law of mass action, which states that the rate of reaction is proportional to the concentration of reactants. The rate of the forward reaction is proportional to [L][R]. The rate of the reverse reaction is proportional to [RL], since there is no other species involved in the dissociation. At equilibrium, the rate of the forward reaction is equal to the rate of the reverse reactions, and so (using k_1 and k_{-1} as the respective proportionality constants)

$$k_1[\mathrm{R}][\mathrm{L}] = k_{-1}[\mathrm{RL}]$$

Rearranging

$$\frac{[\mathrm{R}][\mathrm{L}]}{[\mathrm{RL}]} = \frac{k_{-1}}{k_1} = K_\mathrm{D}$$

where K_D is the equilibrium dissociation constant, which also represents the concentration of ligand which occupies 50% of the receptor population at equilibrium. Experimentally, the measurement of equilibrium dissociation constants has most often been addressed using radioligand binding assays. There are many other ways to determine equilibrium constants more exactly (BIAcore technology[26] and Isothermal titration calorimetry[27] are two well-known examples) yet most have not been applied to the study of MHC peptide interactions. Saturation analysis measures equilibrium binding at various radioligand concentrations to determine receptor number (usually denoted B_max) and affinity (K_D). Competitive binding experiments measure binding at single concentration of labeled ligand in the presence of various concentrations of unlabeled ligand. Competition experiments can be either homologous (where the labeled and unlabeled peptides are the same) or, more commonly, heterologous (where labeled and unlabeled peptides are different) inhibition assays. Homologous inhibition experiments perform a similar function to saturation analysis.

IC_{50} values, obtained from a competitive radioligand or fluorescence binding assay,[28] are the most frequently reported affinity measures. The value given is the concentration required for 50% inhibition of a standard labeled peptide by the test peptide. Therefore nominal binding affinity is inversely proportional to the IC_{50} value. Values obtained from radioligand or fluorescence methods may be significantly different. IC_{50} values for a peptide may vary between experiments depending on the intrinsic affinity and concentration of the standard radiolabeled reference peptide as well as the intrinsic affinity of the test peptide.

The K_D of the test peptide can be obtained from the IC_{50} value using the relationship derived by Cheng and Prussoff[29]

$$K_\mathrm{D}{}^\mathrm{i} = \frac{\mathrm{IC}_{50}}{\left(1 + \dfrac{[L_\mathrm{tot}{}^\mathrm{S}]}{K_\mathrm{D}{}^\mathrm{S}}\right)}$$

where $K_\mathrm{D}{}^\mathrm{i}$ is the dissociation constant for the inhibitor or test peptide, $K_\mathrm{D}{}^\mathrm{S}$ is the dissociation constant for the standard radiolabeled peptide, and $[L_\mathrm{tot}{}^\mathrm{s}]$ is the total concentration of the radiolabel. This relation holds at the midpoint of the inhibition curve under two principal constraints: the total amount of radiolabel is much greater than the concentration of bound radiolabel and that the concentration of bound test peptide is much less than the IC_{50}. This relation, although an approximation, holds well under typical assay conditions.

For competition assays, it can be shown that IC_{50} values are defined by

$$K_\mathrm{D}{}^\mathrm{i} = \frac{[\mathrm{R}_\mathrm{free}](\mathrm{IC}_{50} - [\mathrm{RL}^\mathrm{i}])}{[\mathrm{RL}^\mathrm{i}]}$$

where $[\mathrm{RL}^\mathrm{i}]$ is the concentration of test peptide bound to MHC and R_free is the concentration of free MHC. Both $[\mathrm{R}_\mathrm{free}]$ and $[\mathrm{RL}^\mathrm{i}]$ are independent of the test IC_{50} value. It is clear, then, that the measured IC_{50} value varies with the equilibrium dissociation constant, at least within a single experiment. In practice, the variation in IC_{50} is often sufficiently small that values can be compared between experiments. For the peptide discussed above, YLDDPDLKY, the radiolabeled IC_{50} value recorded in JenPep is 2.8 nM. RIC^{-1}, which are calculated in a relative binding assay, is the amount of the test peptide required to inhibit 50% of a radiolabeled reference peptide's binding. The value is then normalized to the concentration of unlabeled reference peptide required to achieve 50% of the labeled reference peptide's binding. Data on peptide binding to TAP, currently the smallest of our compendia, are limited to radiolabeled IC_{50} data. As yet, TAP binding has not been studied as deeply as other areas of quantitative immunology.

BL_{50} values are also obtained from a peptide binding assay.[24] They are the half-maximal binding levels calculated from mean fluorescence intensities (M.F.I.) of MHC expression by RMA-S or T2 cells. Cells are incubated with the test peptide and then labeled with a fluorescent monoclonal antibody. The nominal binding strength is again inversely proportional to the BL_{50} value. These assays are often termed stabilization assays, as it is presumed that cell surface MHCs are only stable when they have bound peptide. Given that peptides are typically administered extracellularly, there remain questions about the mechanism of peptide induced MHC stabilization. Moreover, the measured BL_{50} values also represent an approximate overall value from a complex multicomponent equilibrium. The interaction between peptide and MHC, as reflected in complex stability, is measured by binding to it either an allele- or class I-specific antibody, which is then bound by a flourescently labeled antibody specific for the first antibody. The resulting complex is then assayed spectrophotometrically using FACS or an equivalent technique. Affinity measures very similar to BL_{50}s, known by various similar names such as SC_{50}, C_{50}s, etc., are also found in the literature. SC_{50} is the binding affinity calculated from a stabilization assay. It is the 50% maximal stabilization concentration inducing half of the maximal up regulating effect. The binding strength is inversely proportional to the SC_{50} value. C_{50} is, similarly, the molar concentration of the peptide at 50% of the maximum fluorescence obtained with that peptide. EC_{50} expresses relative binding. When multiple peptides are compared the C50 values of a reference peptide is obtained. Binding of the other peptides is expressed as EC50, which is the molar concentration of a given peptide required to obtain the fluorescence value at the C50 of the reference peptide.

The half-life for radioisotope labeled β_2-microglobulin dissociation from an MHC class I complex, as measured at 37 °C, is a commonly reported alternative binding measure.[25] This is a kinetic measurement rather than a thermodynamic one, although it is often assumed that the greater the half-life the stronger the peptide−MHC complex. The half-life ($t_{1/2}$) equals

$$t_{1/2} = \frac{\ln 2}{k_{-1}} \sim \frac{0.693}{k_\mathrm{off}}$$

Here the $t_{1/2}$ corresponds to the dissociation of the MHC-β_2 microglobulin complex rather than the kinetics of the protein−ligand interaction. One would anticipate that the peptide dissociation would be related to the overall dissociation of the complex, but quite what this relationship is has not been characterized.

T_m is the temperature at which 50% of the MHC protein is denatured as measured by circular dichroism.[30] Jenpep also records so-called Weak/Nonbinder as a category. This indicates that the peptide has been tested in an MHC restriction assay and has been found to exhibit a binding affinity so low that it can be categorized as inactive. The database also records whether a recorded value has been obtained using Ungraph.[23] In JenPep, Ungraph is used to extract binding values from graphs where exact numerical values are not given. Numerical data gathered using this method are operationally derived approximations to the experimentally verifiable measured values but still retain a sufficiently high degree of accuracy to be worth recording.

We may wish to ask the question: which of these measures is best? Unfortunately, there is no simple answer. The prima facie response might be "equilibrium constant" but in what context? K_D and IC_{50} values are probably the most accurately measured constants, as they are usually assayed using soluble protein. However, as MHCs are membrane bound in their functional context, a value, such as a BL_{50}, might be more relevant to processes in vivo. However, BL_{50} values are typically measured using a cascade of antibodies and the multiple equilibrium that results may obscure salient experimental details. Many other measures, different to those adumbrated above, have also been reported in the context of MHC-peptide interaction. For example, Mean Fluorescence Intensities, or MFIs, typically measured at a single peptide concentration, are widely reported. However, as yet, no clear consensus has emerged on the most appropriate type of affinity measurement or assay strategy. A step forward will come when we are able to deploy methods that return richer affinity measurements. ITC is an example, returning not only highly accurate dissociation constants but also corresponding entropy, enthalpy, and heat capacity values. However, routine ITC again only operates on soluble molecules and is very time-consuming, labor intensive, and relatively expensive to perform. No method readily addresses the joint goals of effectively mimicking the in vivo, membrane bound nature of the interaction and the need for accuracy. It is a prime concern to establish a correlation between these different binding measures, so that information-rich measures can work synergistically with those that are facile to perform. Moreover, it reinforces the need to establish effective predictive methodology that can substitute for experimental assays.

Once a peptide has bound to a MHC, then in order for it to be "recognized" by the immune system, the pMHC complex has to be "recognized", at the molecular level, by one of the TCRs of the T cell repertoire. As part of our database expansion, we have added data on the thermodynamic stability of the ternary complex of TCR, MHC, and peptide, as expressed as the affinity of the TCR for the peptide−MHC complex. We have gathered together information related to this process including data on affinities (essentially a subset of MHC ligand measures) and kinetics (on rates and off rates). It is generally accepted that a peptide binding to an MHC may be recognized by a TCR if it binds with a pIC50 $>$ 6.3, or a half-life $>$ 5 min, or some similar figure for other binding measures.[3] Some peptides binding at these affinities will become immunodominant epitopes, others weaker epitopes, and still others will show no T cell activity. Data on B-cell and T-cell epitopes are currently limited to a list of "binders". There are many different assays used to identify B cell and T cell epitopes. These include, for cellular immunity, T cell killing, proliferation assays such as thymidine uptake, etc. and for humoral immunity, techniques such as enzyme-linked immunosorbent assay (ELISA) or competitive inhibition assays yield values for the Antibody Titer. The quantitative data produced by such assays, while interpretable (a peptide is or is not an epitope), are not consistent enough to be used outside of the limited criteria appropriate for a particular set of experimental conditions. Instead, we have decided to rely on the judgment of immunologists to define, as accurately as possible, what are, or are not, T cell epitopes. It should be noted that for some peptides it is quite common for them to be as classified both as T cell epitopes and MHC ligands, verifying the general rule that correlates affinity with immunogenicity. Thus a T cell epitope is always a MHC ligand, but not all MHC ligands are T cell epitopes.

DISCUSSION

In this paper we have described a significantly updated, and much expanded, version of our JenPep database. Since the first report of Jenpep,[4] there has been a rapid increase in the number and range of data that it contains and in the sophistication of its underlying database architecture. Jenpep version 2.0 has increased significantly in size, scope, and searchability and now contains five categories of immunological peptide binding data: TAP transporter ligands, MHC binding ligands, and peptide−MHC−TCR complexes as well as B cell and T cell epitopes. For the first three categories, JenPep records a variety of quantitative binding measures, and for the last two categories the database comprises annotated lists of epitopes. As well as the addition of two extra categories, this represents an increase from version 1.0 of more than 100% for MHC ligands, 50% for T cell epitopes, but only a nugatory increase for TAP transporter ligands due to the paucity of newly available data.

JenPep shares characteristics with a number of recently emerged databases: functional immunological databases,[7,9,10] thermodynamic binding databases, such as ProTherm[31] and BindingDB,[32] and a variety of other databases, of which BIND[33] and Brenda[34] are good exemplars, whose similarity the Jenpep system is less clearly defined. Databases containing experimentally measured binding affinities are a relatively recent development. The focus of these databases is rigorously measured thermodynamic properties derived from experimental protocols such as isothermal titration calorimetry, which can return not only free energies of binding but also equivalent enthalpies, entropies, and heat capacities. Moreover, because these protocols are well standardized, such databases are able to record easily precise information on experimental conditions.

Although JenPep also focuses, in part, on thermodynamic properties, the extreme diversity of experimental measurements we record currently prevents us from matching the rigor promulgated by databases such as BindingDB. Data standardization remains a significant issue, and the problems we face are not trivial. This greatly affects the degree to which we can automate this process of mining the bioscience literature. Literature or text mining is the unsupervised extraction of data and information directly from machine-readable text. While it may be easy to identify data—the sequences of peptides or numerical values such as an IC_{50}—on the basis of case, or the unequivocal association of an unambiguous symbol with a fixed format number, the identification of information is intrinsically harder, being highly context dependent. Thus far, relatively little has been published that focuses on the immunological literature. This is partly due to its scale—perhaps a tenth of PUBMED is immunological—and partly the confusing nature of written language, which confounds all attempts at text mining. Whatever people may say, capturing and representing complex knowledge is time-consuming, expensive, and surprisingly difficult. Literature mining, useful though it may be, is simply not enough; we still need people to manage the process. For example, it is remarkable how diverse the explanations of a similar piece of work can be. To illustrate this, let us look at the determination of radiolabeled IC_{50}s from a competitive peptide assay binding to HLA-A*0201. The HLA allele could be recorded as the serotype A2 or the allele A*0201. If the serotype is given, then we cannot be certain which allele is implied (0201, 0202, 0203, etc). We also require the sequence of the peptide tested and the IC50 value measured. The peptide sequence is usually given but sometimes as only as a subsequence of a specific protein (i.e. residues 189−198). If the identification of the protein is vague or equivocal, this again proves problematic. The IC50 value is fairly standard but can be given in several units: grams per mililiter or molarity. We might also like to record standard experimental details, such as pH, temperature, or the concentration range over which the experiment was conducted. Also the sequence and concentration of the reference radiolabeled peptide competed against, so that we may be able to calculated the dissociation constant (K_D). Few, if any, papers contain all these details, and it requires human interaction to read between the lines in order to extract as much data as are available.

Clearly, then, there is a certain degree of crossover between our database and related systems.[7,9,10,31−34] However, the nature of the data within JenPep sets it apart from other functional immunological databases: it is the first database in immunology to concentrate on quantitative measurements and represents an important complement to existing systems. Most methods for predicting peptides that bind MHCs are predicated on an apparent dichotomy in affinity between epitopes and nonepitopes. They utilize a classification scheme to greatly simplify the great diversity of extant affinity binding measurements.[6] However, recent attempts have turned to the development of more quantitative models.[35−37] The development of JenPep underlies our attempts to generate statistically sound QSAR models for the prediction of epitopes, which is vital to our goal of developing computer aided vaccine design.

First attempts to characterize in silico MHC binding peptides led to the development of motifs which seek to describe the specificity of individual MHC alleles. Motifs currently experience a wide popularity among immunologists and characterize a short peptide in terms of anchor positions with highly restricted preferences for certain amino acids: the presence of certain amino acids at particular positions that are thought to be essential for binding. For example, human Class I allele HLA-A*0201, probably the best studied of all alleles, has anchor residues at two peptide positions: residue two (P2) and residue nine (P9) for a peptide of length 9. At P2, anchor amino acids would be L and M, and at P9: V and L. Secondary anchors, residues that are favorable, but not essential, to peptide binding, may also be present, and other positions can show preferences for particular residue types. The motif approach is commendably uncomplicated: it is simple to implement either by eye or, more systematically, using a computer to scan protein sequences. However, there are many fundamental problems with this approach, the most significant being that it is, at a fundamental level, a deterministic method. A peptide is either a binder or is not a binder. Even a brief reading of the literature in immunology and vaccinology shows that motif matches produce many false positives and, in all likelihood, an equal number of false negatives, though these are seldom screened.

As a consequence of these shortcomings, alternative approaches abound, each exhibiting different strengths and weaknesses. A significant progression from simple motifs came with the work of Kenneth Parker.[38] This method, which is based on regression analysis, gives quantitative predictions in terms of half-lives for the dissociation of β_2-microglobulin from the MHC complex. Another common strategy is to use data from binding experiments to generate matrices able to predict MHC binding. Positional scanning peptide libraries (PSPLs) have, for example, been used to generate such matrices.[39−41] Other empirical methods include EpiMatrix and EpiMer developed by DeGroot and co-workers[42] and TEPITOPE developed by Hammer and colleagues.[43] An alternative strategy has been to use methods from sequence analysis: Reche et al.[44] have recently developed RANKPEP, a program for epitope prediction based on the use of standard sequence profiles.

Several groups have used machine learning techniques, principally artificial neural networks (ANNs) and hidden Markov models (HMMs), to tackle the problem of predicting peptide−MHC affinity. However, ANN development is complicated by several adjustable factors whose optimal values are seldom known initially. These include the initial distribution of weights between neurons, the number of hidden neurons, the gradient of the neuron activation function, and the training tolerance. Other than chance effects, ANN suffer from three limiting factors: interpretation, memorization, and overfitting. As better ANN methods have developed, and rigorous statistics applied to their use, overfitting and overtraining have largely been overcome. Interpretation remains largely intractable: few scientists can readily decipher the complicated weighting schemes used by ANNs. Among the most famous names among those interested in this area has been Vladimir Brusic. His group has developed a range of machine learning techniques, including evolutionary algorithms as well as ANNs and HMMs, aimed at solving MHC peptide binding.[45,46] His work contains models of both Class I and Class II MHC alleles as well as the TAP transporter.[47,48] Other machine learning techniques applied to this problem recently have been support vector machines[49] and HMMs.[50]

A quite different approach to this problem is based on atomistic Molecular Dynamic (MD) simulations, which attempts to calculate the free energy of binding for a given molecular system. In principle, there is no reliance on known binding data, as it attempts the de novo prediction of all relevant parameters: all that is required is the experimental structure, or a convincing homology model, of a MHC peptide complex. Delisi and co-workers were among the first to apply MD in this area and have, recently, developed a series of different methods.[51,52] Rognan has, over a long period, also made important contributions to this area.[53,54] "Virtual Screening" (VS) is a set of techniques closely related to MD simulation and is a term derived from pharmaceutical research: the use of predicted receptor−ligand interactions to rank and/or filter molecules as an alternative to high throughput screening. Both VS and MD are based on the use of pairwise atomistic potential energy functions. There are two main types of virtual screening methodology that have been used to predict MHC binding: one from structural bioinformatics and fold prediction and another from computational chemistry. Bioinformatic approaches to VS include the work of Margalit and Colleagues, who have proposed a number of virtual screening methodologies,[55,56] each of increasing complexity. Rognan has developed a VS approach called FRESNO, which relies on a simple physicochemical model of host−guest interaction, and used it to predict peptide binding to MHCs.[37] Models were trained on HLA-A*0201 and H-2Kk data: lipophilic interactions contributed most to the human allele model, whereas H-bonding predominated in H-2Kk recognition. In a study of peptides binding to HLA-B*2705.[57] Rognan and colleagues found that FRESNO out performed six other VS methods (Chemscore, Dock, FlexX, Gold, Pmf, and Score). Because of the relative celerity of virtual screening methods compared with MD methods and its ability to tackle MHC alleles for which no binding data are available, this method has considerable potential. While both MD and VS methods hold out the greatest hope of true de-novo predictions of MHC binding, their present success rate is not comparable to that of data driven models. In this respect, we have recently applied a number of QSAR based data mining techniques to the problem of T-cell epitope prediction. More specifically, we have developed models using both a 3D QSAR technique called CoMSIA and a 2D QSAR approach, which we have christened the additive method, to determine a number of class I allele specificities.[35,36,58−60]

The processing, presentation, and recognition of peptides by the immune system is a complicated process. Through the integration of data for peptide binding to TAP, MHC, TCR, B cell receptors, and soluble antibodies we will allow the development quantitatively predictive models for the prediction of the immunogenicity of epitope, multiepitope, or subunit vaccine.

FUTURE WORK

As part of a continuing program, we shall seek to expand both the size and scope of JenPep, as we probe more deeply into the immunological literature, hopefully seeing the database grow considerably. There is a clear need to augment our existing set of five databases. Although we cannot guarantee it, we are nonetheless reasonably confident that our data for peptide binding to TAP and MHC approaches completeness. In this regard, our retrospective searching of available data has reached something like saturation. Apart from the incremental increase in binding data, principally as new papers are published, we see the main route to increased peptide−MHC binding data being through our own efforts to generate experimental measurements [Walsh, Doytchinova, Borrow, and Flower, unpublished]. Another option, of course, is to expand the types of data included in the database. We could, for example, conceive of including data on the number of peptide bound MHC complexes expressed on the surface of antigen presenting cells or archive relative or normalized binding data or experiment specific measurements such as MFI values. We do not currently see these as a priority, however.

For T cell epitopes, and especially B cell epitopes, we are, however, only taking early, tentative steps into the literature. There is considerable scope for expanding both the number and nature of epitope data. In particular, and within the context of cellular immunity, we would like to distinguish, where possible, between immunodominant and nondominant epitopes as well as identify agonist and antagonist peptides.

It is our goal to increase the depth as well as the breadth and scope of our treatment of binding data. For example, we intend to supplement the database with important experimental conditions such as temperature, pH, radioligand sequence, radioligand concentrations, etc. used in binding measurements. Moreover, we should also like to extend JenPep to facilitate the analysis of nonnatural mutants of MHC molecules and non-amino acid ligands of MHC molecules, such as post-translationally modified peptides, such as glycosylation or phosphorylation[61,62] and peptidomimetic compounds,[63] and druglike non-peptide small molecules. We could approach this goal from several directions. One would be to implement a GUI driven substructure search algorithm,[64] or we could employ some form of standard encoding, such as CHUCKLES,[65] corresponding to the prevalent one-letter code we use now, or we could attempt to develop our own encoding of nonstandard amino acids. Likewise, to effect completeness, data concerning other types of epitope, primarily carbohydrate and lipid epitopes, must, in time, also be added to the JenPep system.

JenPep would also benefit from a properly annotated list of whole protein antigens, indicating, where available, where such antigens, or derivatives thereof, have been shown to offer protection in vaccination studies. This list would sit "above" the peptide data described above, as a de facto meta-layer, and allow us to overcome one of the few serious shortcomings of our existing system. It is a fair and justifiable criticism that our database is orientated toward peptide data and does not offer other kinds of useful and relevant searches. This is a reflection of the fact that Jenpep originated from our interest in peptide QSAR.[21,35,36,58−60] It would be useful to be able to search, say, for all data associated with a specific protein or organism. This type of search would be best undertaken by querying data associated with particular antigens and linking them to available, and peptide orientated, epitope and binding data. In this way we would quite naturally complement our existing peptide sequence searches with key word and even whole protein sequence searches using, say, BLAST.[66]

We should like to complement our existing thermodynamic data with kinetic rate constants characterizing peptide binding to MHCs, as we have already begun to do with the pMHC−TCR complex data we have added to JenPep. Moreover, another addition to our cellular immunological data would be the addition of data on the thermodynamics and kinetics of other immunological recognition events, such superantigen binding to MHCs and TCRs, the interaction of cell surface coreceptors with the pMHC–TCR complex, or natural killer cell receptor interactions with MHC and MHC homologues.

Additionally, it would also be interesting to complement our existing data on binding to the TAP transporter binding with information on other aspects of the class I and class II presentation pathways, such as proteasomal and cathepsin cleavage patterns. Integration of this kind of data is a prerequisite to the development of sophisticated mathematical models for antigen presentation, ultimately affording the ability to predict mechanistically antigens from sequence.

As we have intimated, the compilation of B cell or antibody epitope data is an area ripe for robust development. The number of both linear and conformational B cell epitopes is large—very much larger than the compilation currently contained in JenPep. The scope exists therefore to greatly increase the number of epitopes recorded and to add a data category for conformational epitopes where we can record both epitope sequences and thermodynamic data for antibody-protein interaction.

Taking a lead from the Interpro Project[67] we can envisage an international collaboration aimed at producing a broadly focused immunogenicity database. In Interpro, existing databases of sequence families, such as PRINTS,[68] have been combined to produce a more comprehensive and complete coverage of known sequence families, combining annotation details from the different component databases. A similar superdatabase, which incorporates, perhaps, inter alia, JenPep, FIMM,[7] SYFPEITHI,[9] and the HIV Molecular Immunology database[10] into a powerful and comprehensive database of immunogenic peptides, would appear the obvious immunological counterpart.

CONCLUSION

Within the context of molecular mechanisms underlying immunovaccinology, JenPep lays the foundation of a proper quantitative physicochemical understanding of peptide presentation and recognition within the immune system. Using the data compiled within our database system, it should be possible to model, at both the molecular and phenomenological level, the complex behavior of immunological systems using mathematical expressions involving the binding constants of ligand–receptors interactions.[69] The database is also a practical tool of utilitarian value in the search for new vaccines, and we have attempted to build the database, and its interface, to meet these joint needs. JenPep itself, unique in the data it archives, is both a high-quality and value-added database, but, useful as it is, the database is clearly less an end and more a beginning. As JenPep grows we will endeavor to increase its usefulness by both deepening our treatment of its thermodynamic aspects and also by broadening the breadth, scope, and quality of the data we cover.

ACKNOWLEDGMENT

We should like to thank Debra Taylor, Prof. Peter Beverley, Dr. Vladimir Brusic, Dr. Persephone Borrow, and Dr. Anne De Groot for their help and encouragement.

REFERENCES AND NOTES

(1) De Groot A. S.; Sbai, H.; Aubin, C. S.; McMurry, J.; Martin, W. Immuno-informatics: Mining genomes for vaccine components. *Immunol. Cell Biol.* **2002**, *80*, 255–269.

(2) Sette, A.; Livingston, B.; McKinney, D.; Appella, E.; Fikes, J.; Sidney, J.; Newman, M.; Chesnut, R. The development of multi-epitope vaccines: epitope identification, vaccine design and clinical evaluation. *Biologicals* **2001**, *29*, 271–276.

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

(4) Blythe, M. J.; Doytchinova, I. A.; Flower, D. R. JenPep, a database of quantitative functional peptide data for immunology. *Bioinformatics* **2002**, 1*8*, 434–439.

(5) Brusic, V.; Zeleznikow, J.; Petrovsky, N. Molecular immunology databases and data repositories. *J. Immunol. Methods* **2002**, *238*, 17–28.

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

(7) Schonbach, C.; Koh, J. L; Flower, D. R.; Wong, L.; Brusic, V. FIMM, a database of functional molecular immunology: update 2002. *Nucleic Acids Res*. **2002**, *30*, 226–229.

(8) MHCBN. Bhasin, M.; Singh, H.; Raghava, G. P. S. MHCBN: A Comprehensive Database of MHC Binding and Non-Binding Peptides. *Nucleic Acids Res.* **2002** (online) (http://www3.oup.co.uk/nar/database/summary/180.

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

(10) Korber, B. T. M.; Brander, C.; Haynes, B. F.; Koup, R.; Kuiken, C.; Moore, J. P.; Walker, B. D.; Watkins, D. *HIV molecular Immunology 2001*; Los Alamos National Laboratory: Theoretical Biology and Biophysics, Los Alamos, NM, 2001.

(11) Williams, A.; Peh, C. A.; Elliott, T. The cell biology of MHC class I antigen presentation. *Tissue Antigens* **2002**, *59*, 3–17.

(12) Gromme, M.; Neefjes, J. Antigen degradation or presentation by MHC class I molecules via classical and nonclassical pathways. *Mol. Immunol.* **2002**, *3*, 181–202.

(13) van Endert, P. M.; Saveanu, L.; Hewitt, E. W.; Lehner, P. Powering the peptide pump: TAP crosstalk with energetic nucleotides. *Trends Biochem. Sci.* **2002**, *27*, 454–461.

(14) Ulrich, H. D. Natural substrates of the proteasome and their recognition by the ubiquitin system. *Curr. Top Microbiol. Immunol*. **2002**, *268*, 137–174.

(15) Lu, J.; Wettstein, P. J.; Higashimoto, Y.; Appella, E.; Celis, E. TAP-independent presentation of CTL epitopes by trojan antigens. *J. Immunol.* **2001**, *166*, 7063–7071.

(16) Serwold, T.; Gonzalez, F.; Kim, J.; Jacob, R.; Shastri, N. ERAAP customizes peptides for MHC class I molecules in the endoplasmic reticulum. *Nature* **2002**, *41*, 480–483.

(17) Gil-Torregrosa, B. C.; Castano, A. R.; Lopez, D.; Del, Val M. Generation of MHC class I peptide antigens by protein processing in the secretory route by furin. *Traffic* **2000**, *1*, 641–651.

(18) Davis, D. M. Assembly of the immunological synapse for T cells and NK cells. *Trends Immunol.* **2002**, *23*, 356–363.

(19) Marsh, S. G.; Albert, E. D.; Bodmer, W. F ; Bontrop, R. E.; Dupont, B.; Erlich, H. A.; Geraghty, D. E.; Hansen, J. A.; Mach, B.; Mayr, W. R.; Parham, P.; Petersdorf, E. W.; Sasazuki, T.; Schreuder, G. M.; Strominger, J. L.; Svejgaard, A.; Terasaki, P. I. Nomenclature for factors of the HLA system, 2002. *Hum. Immunol.* **2002**, *63*, 1213–1268.

(20) Sette, A.; Buus, S.; Appella, E.; Smith, J. A.; Chesnut, R.; Miles, C.; Colon, S. M.; Grey, H. M. Prediction of major histocompatibility complex binding regions of protein antigens by sequence pattern analysis. *Proc. Natl. Acad. Sci. U.S.A.* **1989**, *86*, 3296–3300.

(21) Flower, D. R.; Doytchinova, I. A.; Paine, K.; Taylor, P.; Blythe, M. J.; Lamponi, D.; Zygouri, C.; Guan, P.; McSparron, H.; Kirkbride, H. Computational Vaccine Design. In *Drug Design: Cutting Edge Approaches*; Flower, D. R., Ed.; 2002.

(22) Geschwinde, E; Schonig, H.-J. *Postgresql: Developer's Handbook*, 1st ed.; SAMS: 2001.

(23) UnGraph. Version 4.0. http://www. biosoft.com.

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

(25) Parker, K. C.; DiBrino, M.; Hull, L.; Coligan, J. E. 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*. **1992**, *149*, 1896–1903.

(26) Roos, H.; Karlsson, R.; Nilshans, H.; Persson, A. Thermodynamic analysis of protein interactions with biosensor technology. *J. Mol. Recognit.* **1998**, *11*, 204−210.

(27) Pierce, M. M.; Raman. C. S.; Nall, B. T. Isothermal titration calorimetry of protein−protein interactions. *Methods* **1999**, *19*, 213−221.

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

(29) Cheng, Y.; Prusoff, W. H. Relationship between the inhibition constant (K1) and the concentration of inhibitor which causes 50% inhibition (I50) of an enzymatic reaction. *Biochem. Pharmacol.* **1973**, *22*, 3099−3108.

(30) Sato, A. K.; Zarutskie, J. A.; Rushe, M. M.; Lomakin, A.; Natarajan, S. K.; Sadegh-Nasseri, S.; Benedek, G. B.; Stern, L. J. Determinants of the peptide-induced conformational change in the human class II major histocompatibility complex protein HLA-DR1. *J. Biol. Chem.* **2000**, *275*, 2165−2173.

(31) Sarai, A.; Gromiha, M. M.; An, J.; Prabakaran, P.; Selvaraj, S.; Kono, H.; Oobatake, M.; Uedaira, H. Thermodynamic databases for proteins and protein-nucleic acid interactions. *Biopolymers* **2002**, *61*, 121−126.

(32) Chen, X.; Lin, Y.; Liu, M.; Gilson, M. K. The Binding Database: data management and interface design. *Bioinformatics* **2002**, *18*, 130−139.

(33) Bader, G. D.; Hogue, C. W. BIND- -a data specification for storing and describing biomolecular interactions, molecular complexes and pathways. *Bioinformatics* **2000**, *16*, 465−477.

(34) Schomburg, I.; Chang, A.; Hofmann, O.; Ebeling, C.; Ehrentreich, F.; Schomburg, D. BRENDA: a resource for enzyme data and metabolic information. *Trends Biochem. Sci.* **2002**, *27*, 54−56.

(35) Doytchiniva, I. A.; Flower, D. R. 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.* **2001**, *44*, 3572−3581.

(36) Doytchiniva, I. A.; Flower, D. R. Quantitative approaches to computational vaccinology. *Immunol. Cell Biol.* **2002**, *80*, 270−279.

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

(38) Parker, K. C.; Shields, M.; DiBrino, M.; Brooks, A.; Coligan, J. E. Peptide binding to MHC class I molecules: implications for antigenic peptide prediction. *Immunol. Res.* **1995**, *14*, 34−57.

(39) Stevens, J.; Wiesmuller, K. H.; Barker, P. J.; Walden, P.; Butcher, G. W.; Joly, E. Efficient generation of major histocompatibility complex class I-peptide complexes using synthetic peptide libraries. *J. Biol. Chem.* **1998**, *273*, 2874−2884.

(40) Stryhn, A.; Pedersen, L. O.; Romme, T.; Holm, C. B.; Holm, A.; Buus, S. Peptide binding specificity of major histocompatibility complex class I resolved into an array of apparently independent subspecificities: quantitation by peptide libraries and improved prediction of binding. *Eur. J. Immunol.* **1996**, *26*, 1911−1818.

(41) Udaka, K.; Wiesmuller, K. H.; Kienle, S.; Jung, G.; Tamamura, H.; Yamagishi, H.; Okumura, K.; Walden, P.; Suto, T.; Kawasaki, T. An automated prediction of MHC class I-binding peptides based on positional scanning with peptide libraries. *Immunogenetics* **2000**, *51*, 816−828.

(42) De Groot, A. S.; Sbai, H.; Aubin, C. S.; McMurry, J.; Martin, W. Immuno-informatics: Mining genomes for vaccine components. *Immunol. Cell Biol.* **2002**, *80*, 255−269.

(43) Kwok, W. W.; Gebe, J. A.; Liu, A.; Agar, S.; Ptacek, N.; Hammer, J.; Koelle, D. M.; Nepom, G. T. Rapid epitope identification from complex class-II-restricted T-cell antigens. *Trends. Immunol.* **2001**, *22*, 583−588.

(44) Reche, P.; Glutting, J.; Reinherz, E. Prediction of MHC class I binding peptides using profile motifs. *Hum. Immunol.* **2002**, *63*, 701−708.

(45) Honeyman, M. C.; Brusic, V.; Stone, N. L.; Harrison, L. C. Neural network-based prediction of candidate T-cell epitopes. *Nat. Biotechnol.* **1998**, *16*, 966−969.

(46) Brusic, V.; Rudy, G.; Honeyman, G.; Hammer, J.; Harrison, L. Prediction of MHC class II-binding peptides using an evolutionary algorithm and artificial neural network. *Bioinformatics* **1998a**, *14*, 121−130.

(47) Daniel, S.; Brusic, V.; Caillat-Zucman, S.; Petrovsky, N.; Harrison, L.; Riganelli, D.; Sinigaglia, F.; Gallazzi, F.; Hammer, J.; van Endert, P. M. Relationship between peptide selectivities of human transporters associated with antigen processing and HLA class I molecules. *Immunol.* **1998**, *161*, 617−624.

(48) Brusic, V.; van Endert, P.; Zeleznikow, J.; Daniel, S.; Hammer, J.; Petrovsky, N. A. Neural network model approach to the study of human TAP transporter. *In Silico Biol.* **1999**, *1*, 109−121.

(49) Donnes, P.; Elofsson, A. Prediction of MHC class I binding peptides, using SVMHC. BMC *Bioinformatics* **2002**, *3*, 25−32.

(50) Udaka, K.; Mamitsuka, H.; Nakaseko, Y.; Abe, N. Prediction of MHC Class I Binding Peptides by a Query Learning Algorithm Based on Hidden Markov Models. *J. Biol. Phys.* **2002**, *28*, 183−194

(51) Sezerman, U.; Vajda, S.; DeLisi, C. Free energy mapping of class I MHC molecules and structural determination of bound peptides. *Protein Sci.* **1996**, *5*, 1272−1281.

(52) Vasmatzis, G.; Zhang, C.; Cornette, J. L.; DeLisi, C. Computational determination of side chain specificity for pockets in class I MHC molecules. *Mol. Immunol.* **1996**, *33*, 1231−1239.

(53) Rognan, D.; Reddehase, M. J.; Koszinowski, U. H.; Folkers, G. Molecular modeling of an antigenic complex between a viral peptide and a class I major histocompatibility glycoprotein. *Proteins* **1992**, *13*, 70−85.

(54) Rognan, D.; Scapozza, L.; Folkers, G.; Daser, A. Molecular dynamics simulation of MHC-peptide complexes as a tool for predicting potential T cell epitopes. *Biochemistry* **1994**, *33*, 11476−11485.

(55) Altuvia, Y.; Sette, A.; Sidney, J.; Southwood, S.; Margalit, H. A structure-based algorithm to predict potential binding peptides to MHC molecules with hydrophobic binding pockets. *Hum. Immunol.* **1997**, *58*, 1-11.

(56) Schueler-Furman, O.; Altuvia, Y.; Sette, A.; Margalit, H. Structure-based prediction of binding peptides to MHC class I molecules: application to a broad range of MHC alleles. *Protein Sci.* **2000**, *9*, 1838−1846.

(57) Logean, A.; Sette, A.; Rognan, D. Customized versus universal scoring functions: application to class I MHC-peptide binding free energy predictions. *Bioorg. Med. Chem. Lett.* **2001**, *11*, 675−679.

(58) Doytchinova, I. A.; Blythe, M. J.; Flower, D. R. An additive method for the prediction of binding affinity. Application MHC Class I Molecule HLA-A*0201. *J. Proteome Res.* **2002**, *1*, 263−272.

(59) Doytchinova, I. A.; Flower, D. R. Physicochemical explanation of peptide binding to HLA-A*0201 major histocompatibility complex. A Three-Dimensional Quantitative Structure − Activity Relationship Study. *Proteins* **2002a**, *48*, 505−518.

(60) Doytchinov, I. A.; Flower, D. R. A Comparative Molecular Similarity Index Analysis (CoMSIA) study identifies an HLA-A2 binding supermotif. *J. Comput.-Aided Mol. Design* **2002**, *16*, 535−544.

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

(62) Zarling, A. L.; Ficarro, S. B.; White, F. M.; Shabanowitz, J.; Hunt, D. F.; Engelhard, V. H. Phosphorylated peptides are naturally processed and presented by major histocompatibility complex class I molecules in vivo. *J. Exp. Med.* **2001**, *192*, 1755−1762.

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

(64) Stobaugh, R. E. Chemical Substructure Searching. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 271−275.

(65) Siani, M. A.; Weininger, D.; Blaney, J. M. CHUCKLES: a method for representing and searching peptide and peptoid sequences on both monomer and atomic levels. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 588−593.

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

(67) Apweiler, R.; Attwood, T. K.; Bairoch, A.; Bateman, A.; Birney, E.; Biswas, M.; Bucher, P.; Cerutti, L.; Corpet, F.; Croning, M. D.; Durbin, R., Falquet, L.; Fleischmann, W.; Gouzy, J.; Hermjakob, H.; Hulo, N.; Jonassen, I.; Kahn, D.; Kanapin, A.; Karavidopoulou, Y.; Lopez, R.; Marx, B.; Mulder, N. J.; Oinn, T. M.; Pagni, M.; Servant, F. The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res.* **2001**, *29*, 37−40.

(68) Attwood, T. K.; Blythe, M. J.; Flower, D. R.; Gaulton, A.; Mabey, J. E.; Maudling, N.; McGregor, L.; Mitchell, A. L.; Moulton, G.; Paine, K.; Scordis, P. PRINTS and PRINTS-S shed light on protein ancestry. *Nucleic Acids Res.* **2002**, *30*, 239−241.

(69) Andersen, P. S.; Menne, C.; Mariuzza, R. A.; Geisler, C.; Karjalainen, K. A response calculus for immobilized T cell receptor ligands. *J. Biol. Chem.* **2001**, *276*, 49125−49132.

CI030461E