Web-based Tools for Vaccine Design

Ole Lund[1], Morten Nielsen[1], Can Kesmir[1,2], Jens K. Christensen[1], Claus Lundegaard[1], Peder Worning[1], and Søren Brunak[1]

[1]*Center for Biological Sequence Analysis, BioCentrum, Technical University of Denmark.*
[2]*Theoretical Biology/Bioinformatics, Utrecht University, The Netherlands.*

Introduction

Computational methods used in vaccine design have been changing drastically in recent years. In classical immunological research results could be recorded by pen and pencil or in a spreadsheet, but new experimental high-throughput methods such as sequencing, DNA arrays, and proteomics have generated a wealth of data that are not efficiently handled and mined by these approaches. This has fueled the rapid growth of the field of Immunological Bioinformatics (or Immuno-informatics) that addresses how to handle these large amounts of data in the field of immunology and vaccine design. Many of the methods have been made available on the Internet and can be used by experimental researchers without expert knowledge of bioinformatics. This review attempts to give an overview over the methods currently available and to point out the strengths and weaknesses of the different methods.

Immunological processes described by prediction servers

Only a small fraction of the possible peptides that can be generated from proteins of pathogenic organisms actually generate an immune response. In order to be presented to CD8+ T cells a precursor peptide must be generated by the proteasome. This peptide may be trimmed at the N-terminal by other peptidases in the cytosol (Levy et al., 2002). It must then bind to the transporter associated with antigen processing (TAP) in order to be translocated to the endoplasmatic reticulum (ER). Here it can be trimmed N-terminally by the aminopeptidase associated with antigen processing (ERAAP) while it binds to the major histocompatibility complex class I (MHC I) molecule (Serwold, 2002). Hereafter it is transported to the cell surface. Only half the peptides presented on the cell surface are immunogenic probably due to the limited size of the T cell receptor (TCR) repertoire. The most selective step is binding to the MHC I molecule, since only 1/200 binds with an affinity strong enough to generate an immune response (Yewdell, 1999). For comparison the selectivity of TAP binding is reported to be 1/7 (Uebel *et al*., 1997). This all happens in competition with other peptides so in order for a peptide to be immunogenic (immunodominant) it must go through the above described process more efficiently than other peptides produced in a given cell (Reviewed by Yewdell, 1999).

Whereas the MHC I molecule mainly samples peptides from the cytosol, the MHC II molecule presents peptides from endocytosed proteins. Unfolded polypeptides bind to MHC II in the endocytic organelles (Reviewed by Castllino, 1997). Both MHC I and MHC II are highly polymorphic, and the specificity of the alleles are often very different. Different individuals will thus typically react to a different set of peptides from a pathogen.

The specificity of some of the processes involved in antigen presentation can be predicted from the amino acid sequence. This can for example be used to select epitopes for use in a vaccine, and help to understand the role of the immune system in infectious diseases, autoimmune diseases and cancers. Below we describe a number of resources available on the web that can perform such predictions.

Databases of MHC binding peptides

Several databases of MHC binding peptides now exist on the web (Table 1).

SYFPEITHI: The SYFPEITHI database contains information on peptide sequences, anchor positions, MHC specificity, source proteins, source organisms, and publication references. The database comprise approximately 3500 peptide sequences known to bind class I and class II MHC molecules and is based on previous publications on T-cell epitopes and MHC ligands from many species (Rammensee, 1999).

MHCPEP: The other major database of MHC binding peptides, MHCPEP, (Brusic, 1997) comprises over 13,000 peptide sequences known to bind MHC molecules. Entries were compiled from published reports as well as from direct submissions of experimental data. Each entry contains the peptide sequence, its MHC specificity and, when available, experimental method, observed activity, binding affinity, source protein, anchor positions, and publication references. Unfortunately the database has since June 1998 been static. The database can be downloaded as an ASCII file.

JenPep: The JenPep database is a newer database that contains quantitative binding data of peptides to MHC and TAP, as well as T cell epitopes (Blythe, 2001). The database contains more than 8000 entries .

FIMM: The database by Schoenbach & Brusic is a functional database of molecular immunology. The database contains 571 antigens and 1591 peptides (Schonbach et al., 2002)

MHCBN: (Bhasin, 2002) is a database of MHC binding and non-binding peptides containing 14,816 binders, 1,782 non-binders and 5,456 T-cell epitope entries.

HLA Ligand/Motif database: This site's database can be searched by defining allele and specificity, amino acid pattern, ligand/motif in sequence of amino acids, author's last name, or advanced search with more criteria.

HIV Molecular Immunology database: The HIV Molecular Immunology Database is an annotated, searchable collection of HIV-1 cytotoxic and helper T-cell epitopes and antibody binding sites. The goal of the database is to provide a comprehensive listing of defined HIV epitopes (Korber *et al*., 2001).

EPIMHC: MHC ligand database that can be searched based on sequence, length, class, species, and on whether a ligand is an epitope or not.

NIH will over the next five to seven years fund an "Immune Epitope Database and Analysis Program" (www.niaid.nih.gov/contract/archive/rfp0331.pdf) to design, develop, populate, and maintain a publicly accessible, comprehensive Immune Epitope Database containing linear and conformational antibody epitopes and T cell epitopes. This database may eventually incorporate most of the data from the above described databases.

Prediction of MHC binding

Several peptide-MHC binding prediction servers exist on the web (Table 2). As indicated in the table some of the web based methods also allow prediction of binding to Class II molecules. Most methods available on the web for predicting MHC-peptide binding are matrix methods. Parameters are often derived from pool sequencing of ligands. Matrices or hidden Markov models may however also be derived from a set of ligand sequences. In these methods the amino acid on each position in the motif gives an independent contribution to the prediction score. Neural networks are able to make more accurate predictions if correlations between positions exist, and there are enough data to model them. This has the potential advantage that it can take correlations between different positions in the binding motif into account.

BIMAS: The BIMAS method was developed by Parker *et al*., (1994). The method is based on coefficient tables deduced from the published literature. For HLA-A2, peptide binding data were combined together to generate a table containing 180 coefficients (20 amino acids x 9 positions), each of which represents the contribution of one particular amino acid residue at a specified position within the peptide (Parker *et al*., 1994).

SYFPEITHI: The SYFPEITHI prediction is based on published motifs (pool sequencing, natural ligands) and takes into consideration the amino acids in the anchor and auxiliary anchor positions, as well as other frequent amino acids. The score is calculated according to the following rules: The amino acids of a certain peptide are given a specific value depending on whether they are anchor, auxiliary anchor or preferred residue. Ideal anchors will be given 10 points, unusual anchors 6–8 points, auxiliary anchors 4–6 and preferred residues 1–4 points. Amino acids that are regarded as having a negative effect on the binding ability are given values between –1 and –3 (Rammensee, 1997; 1999). On the SYFPEITHI web site predictions can be made for 5 different MHC II alleles in addition to a number of Class I alleles.

PREDEPP: In this method the peptide structure in the MHC groove is used as a template upon which peptide candidates are threaded, and their compatibility to bind is evaluated by statistical pairwise potentials. This method has the advantage that it does not require experimental testing of peptide binding, and can thus be used for alleles where only limited data are available (Schueler-Furman *et al*., 2000).

Epipredict: Method using synthetic combinatorial peptide libraries to describe peptide-HLA class II interaction in a quantitative way. The binding contribution of every amino acid side chain in a class II-ligand is described by allele-specific two-dimensional databases (Jung *et al*., 2001).

Predict: The Predict method use neural networks to predict Class I, II and TAP binding (Yu *et al*., 2002).

Propred: The Propred method (Singh, 2001) is based on the matrices published by Sturniolo (1999), and is an implementation and extension of the TEPITOPE program. (Hammer, 1995; Raddrizzani, 2000)). Besides differences that can be attributed to round off errors we have in our tests not seen any differences between the two implementations.

MHCPred: Prediction of binding to 11 different HLA class I alleles using a three-dimensional quantitative structure-activity relationship method (Doytchinova *et al*., 2002).

NetMHC: Prediction of HLA-A2 binding using neural networks. This method predicts quantitatively the binding affinity, and is different from methods performing classification only (binding versus non-binding according to a threshold). The method has been trained using quantitative binding data generated

by the same assay (Buus *et al*., 2003), and some predicted binders have been tested for their ability to induce a CTL response in mice and be recognized by CD8+ T-cells from HLA-A2 HIV-1 positive patients (Corbet *et al*., 2003). Two well-known prediction methods, TEPITOPE and EpiMatrix (Meister 1995; De Groot, 1997) that are not available through the web are listed in Table 3. TEPITOPE is popular since it allows prediction of peptides to many different Class II molecules.

Prediction of proteasomal cleavage sites

The C terminal of MHC class I ligands must most likely be cleaved by the proteasome. The proteasome usually generates precursors of MHC ligands with an extension at the N-termini. These precursors can be trimmed at the N-terminal in the ER. The existence of proteasome cleavage sites within epitopes need not abrogate the immune response for such epitopes. They may, however, reduce the availability, and thereby the immunogenecity of a given peptide (Yewdell, 1999). The proteasome thus plays an important role in selecting which peptides are presented to CD8+ T cells. In vertebrates stimulation with IFN-γ leads to the replacement of three subunits of the constitutive proteasome to form the so-called immunoproteasome which has a different specificity (reviewed by Uebel, 1999). Different methods for predicting proteasomal cleavage sites exist on the web (Table 4).

PAProC: Prediction Algorithm for Proteasomal Cleavages is a prediction tool for cleavages by human and yeast proteasomes, based on experimental cleavage data. (Kuttler, 2000; Nussbaum, 2001). An updated version of the PAProC program based on *in vitro* immunoproteasome cleavage data (Toes, 2001) is also in the making according to the PAProC homepage.

FRAGPREDICT comprises two different algorithms. One that aims at predicting potential proteasomal cleavage, based on a statistical analysis of cleavage-determining amino acid motifs present around the scissile bond (Holzhütter *et al*., 1999, 2000). The second algorithm, which uses the results of the cleavage site analysis as an input, provides predictions of major proteolytic fragments.

NetChop: (Kesmir, 2002) is a method based on neural networks that have been trained on different data sets. C Kesmir suggests to use the C-term 2.0 network which was trained on C-terminal cleavage sites of 1,110 publicly available MHC class I ligands for predicting the boundaries of CTL. The specificity of this network may resemble the specificity of the immunoproteasome.

Margalits group have also recently made their proteasomal cleavage site propensities (Altuvia and Margalit, 2000) available on the net (bioinfo.md.huji.ac.il/marg/cleavage/index.html).

Combined predictions

A number of sites providing combined predictions have been developed recently. The MAPPP server (Table 2) allows the user to make an open reading frame (ORF) search combined with MHC binding and proteasomal cleavage site predictions, and Raghava have a prediction server (www.imtech.res.in/raghava/propred1/index.html) which implements matrices for 47 MHC Class-I alleles and proteasomal and immunoproteasomal models. The NetMHC server allows combination of HLA-A2 and NetChop predictions.

MHC sequence databases

A number of databases containing sequences of proteins of immunological interest exist on the web (Table 5).

HIG: The HLA Sequence Database currently contains 1,596 allele sequences. To date (October 2002), some 263 HLA-A, 501 HLA-B, 125 HLA-C, 6 HLA-E, 1 HLA-F and 15 HLA-G class I alleles have been named. A total of 3 HLA-DRA, 397 HLA-DRB, 22 HLA-DQA1, 53 HLA-DQB1, 20 HLA-DPA1, 100 HLA-DPB1, 4 HLA-DMA, 6 HLA-DMB, 8 HLA-DOA and 8 HLA-DOB class II sequences have also been assigned. There are also 6 TAP1, 4 TAP2 and 54 MICA sequences. The HLA Sequence Database also contains the comprehensive nomenclature for factors of the HLA system (listings for HLA class I and class II allele names) which is very helpful since the HLA nomenclature is very complicated and cumbersome.

IMGT: IMGT, the international ImMunoGeneTics project, is a collection of databases specializing in Immunoglobulins, T cell receptors and the Major Histocompatibility Complex (MHC) of all vertebrate species. The IMGT project was established in 1989 by the Université Montpellier II and the CNRS (Montpellier, France) and works in close collaboration with the EBI.

ASHI: The American Society for Histocompatibility and Immunogenetics (ASHI) hosts databases of gene and allele frequencies (www.ashi-hla.org/).

MHCDB: "Registered users only" database of MHC sequences. This is an ACeDB-style database holding the Human Major Histocompatibility Database. It is largely superseded by 6ace which is ACeDB-style database of human chromosome 6 from the Sanger Centre.

Other sites

A number of other databases relevant to immunology and vaccine design are listed in Table 6. Table 7 contains a compilation of lists of links. As stated

in Table 7 we will also make an HTML version of this article available on the net.

References

Altuvia Y, Margalit H. Sequence signals for generation of antigenic peptides by the proteasome: implications for proteasomal cleavage mechanism. *J Mol Biol*. 2000 **295**:879–90.

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

Blythe MJ, Doytchinova IA, Flower DR. JenPep: a database of quantitative functional peptide data for immunology. *Bioinformatics*. 2002 **18**:434–9.

Brusic V, Rudy G, Harrison LC. MHCPEP, a database of MHC-binding peptides: update 1997. *Nucleic Acids Res*. 1998 **26**:368–71.

Buus S, Lauemøller SL, Worning P, Kesmir C, Frimurer T, Corbet S, Fomsgaard A, Hilden J, Holm A, and Brunak S. Sensitive quantitative predictions of peptide-MHC binding by a "Query by Committee" artificial neural network approach. Accepted for publication in *Tissue Antigens*, 2003.

Castellino F, Zhong G, Germain RN. Antigen presentation by MHC class II molecules: invariant chain function, protein trafficking, and the molecular basis of diverse determinant capture. *Hum Immunol*. 1997 **54**:159–69.

Corbet S, Nielsen HV, Vinner L, Lauemøller SL, Therrien D, Tang S, Kronborg G, Mathiesen L, Chaplin P, Brunak S, Buus S, and Fomsgaard A. Optimisation and immune recognition of multiple novel conserved HLA-A2, HIV-1-specific CTL epitopes. Accepted for publication in *General Virology*.

De Groot AS, Jesdale BM, Szu E, Schafer JR, Chicz RM, Deocampo G. An interactive Web site providing major histocompatibility ligand predictions: application to HIV research. *AIDS Res. Hum. Retroviruses* 1997 **13**:529–31.

Doytchinova IA, Flower DR. Physicochemical explanation of peptide binding to HLA-A*0201 major histocompatibility complex: a three-dimensional quantitative structure-activity relationship study. *Proteins*. 2002 **48**:505–18.

Hammer J. New methods to predict MHC-binding sequences within protein antigens. *Curr Opin Immunol*. 1995 **7**:263–9.

Holzhutter HG, Frommel C, Kloetzel PM. A theoretical approach towards the identification of cleavage-determining amino acid motifs of the 20 S proteasome. *J Mol Biol*. 1999 **286**:1251–65.

Holzhutter HG, Kloetzel PM. A kinetic model of vertebrate 20S proteasome accounting for the generation of major proteolytic fragments from oligomeric peptide substrates. *Biophys J*. 2000 **79**:1196–205.

Johnson G, Wu TT. Kabat Database and its applications: future directions. *Nucleic Acids Res*. 2001 **29**:205–6.

Jung G, Fleckenstein B, von der Mulbe F, Wessels J, Niethammer D, Wiesmuller KH. From combinatorial libraries to MHC ligand motifs, T-cell superagonists and antagonists. *Biologicals*. 2001 **29**:179–81

HIV Molecular Immunology 2001, Editors: Bette T. M. Korber, Christian Brander, Barton F. Haynes, Richard Koup, Carla Kuiken, John P. Moore, Bruce D. Walker, and David I. Watkins. Publisher: Los Alamos National Laboratory, Theoretical Biology and Biophysics, Los Alamos, New Mexico. LA-UR 02-4663.

Kuttler C, Nussbaum AK, Dick TP, Rammensee HG, Schild H, Hadeler KP. An algorithm for the prediction of proteasomal cleavages. *J Mol Biol*. 2000 **298**:417–29.

Levy F, Burri L, Morel S, Peitrequin AL, Levy N, Bachi A, Hellman U, Van den Eynde BJ, Servis C. The final N-terminal trimming of a subaminoterminal proline-containing HLA class I-restricted antigenic peptide in the cytosol is mediated by two peptidases. *J Immunol* 2002 **169**:4161–71.

Meister GE, Roberts CG, Berzofsky JA, De Groot AS. Two novel T cell epitope prediction algorithms based on MHC-binding motifs; comparison of predicted and published epitopes from Mycobacterium tuberculosis and HIV protein sequences. *Vaccine*. 1995 **13**:581–91.

Nussbaum AK, Kuttler C, Hadeler KP, Rammensee HG, Schild H. PAProC: a prediction algorithm for proteasomal cleavages available on the WWW. *Immunogenetics*. 2001 **53**:87–94.

Parker KC, Bednarek MA, Coligan JE. Scheme for ranking potential HLA-A2 binding peptides based on independent binding of individual peptide side-chains. *J Immunol*. 1994 **152**:163–75.

Raddrizzani L, Hammer J. Epitope scanning using virtual matrix-based algorithms. *Brief Bioinform*. 2000 **1**:179–89.

Rammensee H, Bachmann J, Emmerich NP, Bachor OA, Stevanovic S. SYFPEITHI: database for MHC ligands and peptide motifs. *Immunogenetics*. 1999 **50**:213–9.

Rammensee H-G, Bachmann J, Stevanovic S. MHC ligands and peptide motifs. *Landes Bioscience*, 1997.

Serwold T, Gonzalez F, Kim J, Jacob R, Shastri N. ERAAP customizes peptides for MHC class I molecules in the endoplasmic reticulum. *Nature*. 2002 **419**:480–3.

Schonbach C, Koh JL, Flower DR, Wong L, Brusic V. FIMM, a database of functional molecular immunology: update 2002. *Nucleic Acids Res*. 2002 **30**:226–9.

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**:1838–46.

Singh H, Raghava GP. ProPred: prediction of HLA-DR binding sites. *Bioinformatics*. 2001 **17**:1236–7.

Sturniolo T, Bono E, Ding J, Raddrizzani L, Tuereci O, Sahin U, Braxenthaler M, Gallazzi F, Protti MP, Sinigaglia F, Hammer J. Generation of tissue-specific and promiscuous HLA ligand databases using DNA microarrays and virtual HLA class II matrices. *Nat Biotechnol*. 1999 **17**:555–61.

Toes RE, Nussbaum AK, Degermann S, Schirle M, Emmerich NP, Kraft M, Laplace C, Zwinderman A, Dick TP, Muller J, Schonfisch B, Schmid C, Fehling HJ, Stevanovic S, Rammensee HG, Schild H. Discrete cleavage motifs of constitutive and immunoproteasomes revealed by quantitative analysis of cleavage products. *J Exp Med*. 2001 **194**:1–12.

Uebel S, Kraas W, Kienle S, Wiesmuller KH, Jung G, Tampe R. Recognition principle of the TAP transporter disclosed by combinatorial peptide libraries. *Proc Natl Acad Sci U S A* 1997 **94**:8976–81

Uebel S, Tampe R. Specificity of the proteasome and the TAP transporter. *Curr Opin Immunol*. 1999 **11**:203–8.

Yewdell JW, Bennink JR. Immunodominance in major histocompatibility complex class I-restricted T lymphocyte responses. *Annu Rev Immunol*. 1999 **17**:51–88.

Yu K, Petrovsky N, Schonbach C, Koh JY, Brusic V. Methods for prediction of peptide binding to MHC molecules: a comparative study. *Mol Med*. 2002 **8**:137–48

Table 1. Databases of MHC binding peptides

Name	Principal Investigator	URL	Description
SYFPEITHI	Rammensee	syfpeithi.bmi-heidelberg.com/scripts/MHCServer.dll/home.htm	Database and prediction server for peptides that bind MHC molecules.
MHCPEP	Brusic, Harrison	wehih.wehi.edu.au/mhcpep	Database of MHC binding peptides
JenPep	Flower	www.jenner.ac.uk/JenPep	Database of MHC and TAP binding peptides
FIMM	Schoenbach & Brusic	sdmc.krdl.org.sg:8080/fimm	Database of functional molecular immunology/binding prediction
MHCBN	Raghava	www.imtech.res.in/raghava/mhcbn	Tools for subunit vaccine design
HLA Ligand/Motif Database	Hildebrand	hlaligand.ouhsc.edu	Ligand database/prediction
HIV Molecular Immunology	Korber	hiv-web.lanl.gov/content/immunology/	HIV CTL epitopes
EPIMHC	Reinherz	mif.dfci.harvard.edu/Tools/db_query_epimhc.html	Peptides that bind to MHC molecules

Table 2. HLA Peptide Binding Predictions

Name	URL	Description
BIMAS	bimas.dcrt.nih.gov/molbio/hla_bind	Prediction of MHC class I binding using matrices
SYFPEITHI	syfpeithi.bmi-heidelberg.com/Scripts/MHCServer.dll/EpPredict.htm	Prediction of Class I and II binding
PREDEPP	bioinfo.md.huji.ac.il/marg/Teppred/mhc-bind	MHC Class I epitope prediction
Epipredict	www.epipredict.de/index.html	Prediction of HLA class II restricted binding
Predict	http://sdmc.krdl.org.sg:8080/predict-demo	Prediction of Class I, II and TAP binding
Propred	www.imtech.res.in/raghava/propred	MHC class II prediction
MHCPred	www.jenner.ac.uk/MHCPred	HLA class I predictions
NetMHC	www.cbs.dtu.dk/services/NetMHC	Prediction of HLA-A2 binding using Neural networks
MAPPP	www.mpiib-berlin.mpg.de/MAPPP/expertquery.html	Combined ORF, MHC binding and proteasomal cleavage Registration needed for expert mode

Table 3. Non web MHC binding predictions

Name	URL	Description
TEPITOPE	www.vaccinome.com	PC Program for Class II predictions can be downloaded
EpiMatrix	epivax.com/epimatrix.html	Commercial epitope prediction

Table 4. Prediction of proteasomal cleavage sites

Name	URL	Description
Paproc	paproc.de	A matrix based method for prediction of prota-somal cleavage
FRAGPREDICT	www.mpiib-berlin.mpg.de/MAPPP/cleavage.html	Proteolytic fragment predicter
NetChop	www.cbs.dtu.dk/services/NetChop	A neural network based method for prediction of proteasomal cleavage

Table 5. MHC sequence databases

Name	URL	Description
HIG	www.anthonynolan.org.uk/HIG	HLA sequence database
IMGT	www.ebi.ac.uk/imgt	Sequences of MHC, TCR and immunoglobulin molecules
ASHI	www.ashi-hla.org	Sequences and Gene and Haplotype frequencies
MHCDB	www.hgmp.mrc.ac.uk/Registered/Option/mhcdb.html	Registered users only database of MHC sequences

Table 6. Other sites

Name	URL	Description
HIV Molecular Immunology database	hiv-web.lanl.gov/content/immunology	HIV immunology
School of Crystallography, Birkbeck College, University of London	www.cryst.bbk.ac.uk/pps97/assignments/projects/coadwell/MHCSTFU1.HTM	Structure and Function of the Major Histocompatibility Complex (MHC) Proteins
MHC-Peptide Interaction Database (MPID)	surya.bic.nus.edu.sg/mpid/	Structural information and characterization of MHC peptide interaction
ELF	hiv-web.lanl.gov/content/hiv-db/ALABAMA/epitope_analyzer.html	Epitope Location Finder
ASHI	www.ashi-hla.org	The American Society for Histocompatibility and Immunogenetics

Table 7. Links to lists of links

Name	URL	Description
Syfpeithi	http://syfpeithi.bmi-heidelberg.com/Scripts/MHCServer.dll/Info.htm	Rammensees links
FIMM	http://sdmc.krdl.org.sg:8080/fimm	Brusics links
CBS	www.cbs.dtu.dk/courses/27485.imm/links.html	Our links
HLA-RELATED LINKS	home.att.net/ dorak/hla/linkhla.html	Doraks links
This article	www.cbs.dtu.dk/researchgroups/immunology/webreview.html	The present article in HTML format