UNIVERSIDAD COMPLUTENSE DE MADRID

FACULTAD DE MEDICINA

Departamento de Inmunología

TESIS DOCTORAL

Modelado computacional del procesamiento y presentación de antígenos

MEMORIA PARA OPTAR AL GRADO DE DOCTOR

PRESENTADA POR

Carmen María Díez Rivero

Director

Pedro Antonio Reche Gallardo

Madrid, 2012

Universidad Complutense de Madrid

Inmunología

Facultad de Medicina

Modelado computacional del procesamiento y presentación de antígenos

Carmen María Diez Rivero

Director de Tesis

Dr. Pedro Antonio Reche Gallardo

Madrid, Marzo 2012

El trabajo presentado en esta memoria ha sido realizado en el Departamento de Mircorbiología I, área de Inmunología, de la Facultad de Medicina de la Universidad Complutense de Madrid (UCM), bajo la dirección del Doctor Pedro Antonio Reche Gallardo

Vº Bº director de tesis

Dr. Pedro Antonio Reche Gallardo

Carmen María Diez Rivero
Madrid, 2012

A mis padres, Juan José y María Jesús

Llega el fin de una etapa en la que he aprendido, compartido y vivido muchas cosas, y este es el momento de agradecer a todos los que han estado presentes y me han ayudado a llegar hasta aquí.

En primer lugar quiero dar las gracias a Pedro A. Reche, quien confió en mi y me dio la oportunidad de incorporarme a su grupo y me ha enseñado y ayudado durante estos años y gracias al cual he podido realizar esta Tesis.

También quiero dar las gracias a toda la gente del departamento de Inmunología de la UCM, Antonio Arnaiz, Edgar Fernández, M. Esther Lafuente, José Manuel Martín Villa, Eduardo Martínez Naves, María José Recio, José Ramón Regueiro, José Manuel Quijano, Juani y Rosi.

A todos mis compañeros y amigos becarios porque han compartido conmigo cada momento, tanto profesional como personal y han estado dispuestos a ayudarme y apoyarme, porque aunque la bioinformática no les resultase muy interesante, han aguantado cada una de mis charlas y porque desde el primer día que me vieron con unos guantes todos se ofrecieron a ayudarme y enseñarme. Sobretodo, muchas gracias a Bea G., con quien he compartido toda esta etapa, por ser mi amiga, por estar siempre ahí para hablar de cualquier cosa y ayudarme. A Miguel, el que cada día me sorprendía más con su "cocina de autor". A Iria, compañera de laboratorio en la última fase, y sobretodo compañera de estrés en los últimos meses, a la que no le gusta discutir, sólo dejar claro su punto de vista, de la que he aprendido mucho; y a sus becarios Dani, Javi y Miriam (¡qué paciencia han tenido!). A Juan, en quien encontré un amigo con quien hablar, contarle mis cosas y con el que me he reído mucho descubriendo el mundo que está ahí fuera. A Vanesa y sus grandes frases, se puede decir más alto, pero no más claro. Ana V. y Marina, que están empezando, mucho ánimo, los nuevos becarios queda en vuestras manos. Y a los becarios que ya se han ido, Alberto y Vero, tardé mucho en conoceros, pero me alegro de haberlo hecho porque sois geniales; Bruno, Dani, Sabela y Ana M. cada vez que uno de vosotros se iba se quedaba un hueco vacío en el departamento; Ana A., Bea A., Jesús, Elena M., Noelia. Y muy especialmente a los *frikimáticos* Santi, Sandra, María, que me ayudó en todo lo que pudo y no sólo explicándome la diferencia entre $, @ y %, Berni, que algunas veces me sacaba de mis casillas pero es un primor, Laura, mi rubia preferida, la que mejor cuenta los chistes y Diego porque, entre otras cosas, nos enseñó que un cubo de Rubik tiene más combinaciones que segundos tiene el Universo.

Quiero dar las gracias a mis amigas de siempre, las que han estado conmigo desde que llevábamos nuestro uniforme azul, no importan los kilómetros que nos separen porque siempre están cuando se las necesita, gracias por vuestro apoyo, por darme fuerzas y por se mis amigas y

aguantarme todos estos años. Coral, gracias por estar ahí para pelearte conmigo y aguantarme cuando tengo un día cruzado. Miriam, creo que todo el mundo debería tener una amiga como tu, eres la persona más buena que conozco. Lucía, que mal me caíste tu primer día de clase y cuando aquella monja me obligó a ir contigo para que no te perdieses, pero la verdad es que me alegro mucho de que lo hiciese porque gracias a eso descubrí lo que significa tener una verdadera amiga. A mis amigos de la facultad, con los que empecé este camino, sobretodo a Criscris, Nacho y Sirgo, que después de tanto tiempo se han convertido en personas imprescindibles para mi. Y a los *nuevos* amigos, Alba, Arturo, Cris, Elena,… sin los que esta etapa en tierras madrileñas no habría sido igual.

He dejado para el final a las personas más importantes para mi, quiero dar las gracias a mi familia, a la que no veo tanto como me gustaría, sobretodo a mis padres, que me han apoyado y ayudado todos estos años y sin los que no habría podido llegar aquí, porque ellos están orgullosos de su hija y así me lo hacen sentir, pero yo estoy aún mas orgullosa de ellos. Os quiero.

I. ÍNDICE

I. ÍNDICE

I. ÍNDICE I

II. ABREVIATURAS VII

III. ABSTRACT IX

1. INTRODUCCIÓN GENERAL 1

1.1 PROCESAMIENTO Y PRESENTACIÓN DE ANTÍGENOS POR MOLÉCULAS DEL MHC DE CLASE I 5
1.1.1 Degradación de proteínas vía ubiquitina-proteasoma 7
1.1.2 Transporte y procesamiento de péptidos en el retículo endoplasmático 10
1.1.3 Carga de péptidos en las moléculas del MHC I y presentación en la superficie celular 13
1.1.4 Reconocimiento antigénico por el TCR de linfocitos T CD8 16
1.2 RELEVANCIA DE LA IDENTIFICACIÓN DE EPÍTOPOS T CD8 17

2. OBJETIVOS 19

3. MATERIAL Y MÉTODOS 23

3.1 BASES DE DATOS 25
3.2 MÉTODOS COMPUTACIONALES 25
3.2.1 N-grams 26
3.2.2 Máquinas de vectors de soporte (SVMs; Support vector machines) 27
3.2.3 Matrices de puntuación específica (PSSMs; Position specific scoring matrices) 29
3.3 DESARROLLO Y EVALUACIÓN DE LOS MODELOS PREDICTIVOS 30
3.3.1 Validación cruzada 30
3.3.2 Test independiente 32
3.3.3 Medidas de rendimiento predictivo 32

4. CAPÍTULO I: Análisis de la distribución de epítopos T CD8 en proteínas virales 37

4.1 JUSTIFICACIÓN Y OBJETIVOS 39
4.2 CONCLUSIONES 39

5. CAPÍTULO II: Modelado computacional de la especificidad de corte del proteasoma 67

5.1 JUSTIFICACIÓN Y OBJETIVOS 69
5.2 CONCLUSIONES 70

6. CAPÍTULO III: Modelado de la afinidad de unión de péptidos a TAP 86

6.1 JUSTIFICACIÓN Y OBJETIVOS 87
6.2 CONCLUSIONES 87

7. CAPÍTULO IV: Desarrollo de otras herramientas computacionales 99

7.1 PVS 101
7.1.1 Justificación y Objetivos 101
7.1.2 Conclusiones 101
7.2 TEPIDAS 111
7.2.1 Justificación y Objetivos 111
7.2.2 Conclusiones 111

8. DISCUSIÓN 117

8.1 SUMARIO 119
8.2 DISTRIBUCIÓN DE EPÍTOPOS T CD8 119
8.3 MODELADO DEL PROTEASOMA 122

8.4 PREDICCIÓN DE PÍTOPOS T CD8 COMBINANDO LAS PREDICCIONES DE UNIÓN A MOLÉCULAS DEL MHC I Y DE CORTE POR EL PROTEASOMA Y/O INMUNOPROTEASOMA 126
8.5 MODELADO DE LA UNIÓN A TAP 127
8.6 PVS 129
8.7 TEPIDAS 131
9. CONCLUSIONES 133
10. REFERENCIAS 137
11. ANEXO I: Otras publicaciones 149

II. ABREVIATURAS

II. ABREVIATURAS

ABC	ATP *binding casette*
ANN	*Artificial neural network*
APC	Célula presentadora de antígenos
BTR	*Better than random*
CDR	Región determinante de complementariedad
CTL	Linfocito T citotóxico
C-terminal	Carboxilo terminal
DRiPs	Productos defectuosos de la síntesis proteica
DAS	*Distributed Annotation Systems*
ER	Retículo endoplásmico
ERAAP	Aminopeptidasa del retículo endoplasmático
FN	Falso negativo
FP	Falso positivo
HCV	Virus de la Hepatitis C
HIV	Virus de la Inmunodeficiencia Humana
HLA	Antígeno leucocitario humano
IAV	Virus Influenza A
IFN	Interferón
ITAM	*Immunoreceptor Tyrosine-based Activation Motifs*
MCC	Coeficiente de correlación de Matthews
MHC	Complejo principal de histocompatibilidad
NK	*Natural Killer*
N-terminal	Amino terminal
ORF	*Open reading frame*
PLC	Complejo de carga peptídica
P_r	Coeficiente de correlación de Pearson
PSSM	*Position specific scoring matrices*
ROC	*Receiver operating characteristic*
SE	Sensibilidad
SP	Especificidad

SVM	*Support vector machine*
TAP	Transportador asociado con el procesamiento de antígenos
TCR	Receptor de célula T
TN	Verdadero negativo
TP	Verdadero positivo
TPPII	Tripeptidil peptidasa II
WEKA	*Waikato Environment for Knowledge Analysis*
β_2m	β_2-microglobulina

III. ABSTRACT

The identification of T cell epitopes is crucial for understanding diseases pathogenesis and aetiology. Moreover, it is also crucial for the development of epitope-based vaccines against infectious agents and treatments for allergic, autoimmune diseases and cancer. CD8 T cell epitopes are peptides presented on the surface of infected or damaged cells bound to MHC I molecules that are recognized by the T cell receptor (TCR). These peptides derive from foreign protein antigens that are degraded in the cytosol by the proteolitic activity of the proteasome. Some of the peptides are translocated by TAP into the endoplasmic reticulum where they can bind to nascent MHC I molecules. Subsequently, peptide loaded MHC I molecules are then displayed on cell surface for recognition by T cells. Traditionally, the identification of T cell epitopes requires the synthesis of overlapping peptides spanning the entire length of a protein, followed by experimental assays over each peptide. This method is expensive and time consuming. Therefore, it is key to develop alternative computational approaches for the prediction of T cell epitopes to decrease the experimental burden associated with epitope identification.

In this Thesis, we have modeled the classical processing pathway of MHC I antigens. We have analyzed the location of 190, 249 and 78 CD8 T cell epitopes of Hepatitis C Virus, Human Immunodeficiency Virus and Influenza A Virus, respectively, in the viral proteins. We found that capsid and matrix proteins encompass, significantly, more epitopes than the expected by their size. We have also modeled the specificity of the cleavage site of the proteasome, using N-grams. We have developed two different models for the proteasome and the immunoproteasome from two distinct sets of MHC I-restricted peptides. The proteasome model was developed using a sets of peptides eluted from human MHC I molecules, whereas the immunoproteasome model was trained using CD8 T cell epitopes naturally processed. In addition, we have also studied the peptide affinity to TAP using support vector machines trained on single residue positions and residue combinations drawn from a large dataset consisting of 613 nonamer peptides of known affinity to TAP. Finally, we have developed two different web tools that are instrumental for epitope selection: PVS and TEPIDAS. PVS (*Protein Variability Server*) is a useful tool for the identification of conserved T and B cell epitopes through the sequence variability analysis. This sever estimates the variability using different methods, like the Shannon entropy, the Simpson diversity index and the Wu-Kabat variability coefficient. PVS can also plot the variability in the MSA and display it in a relevant 3D-structure. TEPIDAS is a DAS Annotation Server that includes CD8 T cell epitopes specific of human pathogenic organisms, the MHC I restriction elements (experimentally determined or predicted) and the associated cumulative phenotypic frequency.

1. INTRODUCCIÓN GENERAL

El sistema inmunitario consiste en un conjunto de células, tejidos y órganos que protegen frente a agentes extraños como bacterias, virus, protozoos, hongos y parásitos. En los vertebrados superiores, y en concreto en el hombre, se pueden distinguir dos tipos de inmunidad: la inmunidad innata y la inmunidad adaptativa.

La inmunidad innata, la más antigua evolutivamente, incluye aquellos componentes del sistema inmunitario que presentan mecanismos para el reconocimiento de estructuras comunes, presentes en diversos microorganismos. Está preparada para iniciar una respuesta rápida frente a estos microorganismos, pero no es capaz de generar memoria inmunológica. La inmunidad innata engloba barreras físicas y químicas (Janeway and Medzhitov, 2002), componentes humorales, como el sistema del complemento (Jules Bordet; 1870 – 1961), el sistema de coagulación, o algunas citocinas, y componentes celulares: macrófagos, células dendríticas, mastocitos, neutrófilos, eosinófilos, células NK (*natural killer*) y células NKT. La inmunidad innata es la primera línea de defensa frente a la infección. La rapidez de respuesta es posible debido a que los mecanismos efectores de la inmunidad innata ya están presentes antes de producirse la infección. La inmunidad innata también es importante para la detección de células tumorales y la inducción de la respuesta inflamatoria. Esta última es esencial para el confinamiento de los agentes infecciosos y la activación de la respuesta inmune adaptativa (Medzhitov and Janeway, 1997; Turvey and Broide, 2010).

Existen otros mecanismos más evolucionados que son estimulados tras la exposición a agentes infecciosos y que, a diferencia de la inmunidad innata, aumentan su intensidad y capacidad defensiva tras una primera exposición a un determinado microorganismo, es la denominada inmunidad adaptativa o adquirida. Esta forma de inmunidad, presente sólo en los vertebrados superiores, se caracteriza por su alta especificidad -que hace que el sistema inmunitario responda de forma singular a variantes de un mismo microorganismo-, y su capacidad para "recordar" y responder a un patógeno frente al que se ha estado expuesto

anteriormente (Pancer and Cooper, 2006). Los componentes de la inmunidad adaptativa son los linfocitos T y B y sus productos, entre ellos los anticuerpos (Elie Metchnikoff; 1845 – 1916). Las sustancias extrañas que inducen respuestas inmunitarias adaptativas y/o son dianas de tales respuestas se denominan antígenos. La respuesta inmunitaria adaptativa requiere del sistema inmunitario innato y de las células presentadoras de antígenos (APC) (Silva, 2010). Una vez que se inicia la respuesta inmunitaria, los componentes de la inmunidad innata y de la inmunidad adquirida no actúan de manera independiente, sino que establecen entre ellos una compleja red de interconexiones cuya finalidad es la de proteger al individuo frente a la infección (Bonilla and Oettgen, 2010; Palm and Medzhitov, 2009). Las respuestas inmunitarias adquiridas se clasifican en dos tipos según el componente del sistema inmunitario que participa en la respuesta:

a. Inmunidad humoral, mediada por los anticuerpos.

b. Inmunidad celular, mediada por células, en la que participan los linfocitos T.

Los linfocitos T se desarrollan en el timo, a partir de progenitores linfoides procedentes de la médula ósea (Hedrick, 2008; Takahama, 2006), donde se diferencian en linfocitos T CD8 y linfocitos T CD4 que migran a la periferia. Los linfocitos T se activan tras la interacción de su receptor TCR (*T cell receptor*) con péptidos antigénicos presentados por moléculas del complejo principal de histocompatibilidad (MHC) de clase I ó II, respectivamente. Los linfocitos T CD4 regulan la respuesta inmunológica mediante la producción de distintas citocinas. Por otro lado, los linfocitos T CD8 se encargan de la eliminación de células infectadas con patógenos intracelulares o que expresan neoantígenos, como los marcadores tumorales, por ello también se les llama linfocitos T citotóxicos (CTLs).

1.1 Procesamiento y presentación de antígenos por moléculas del MHC de clase I

La vía de presentación antigénica por moléculas del MHC I está presente en casi todos los tipos celulares y es el mecanismo por el que se presenta en la superficie celular una muestra de péptidos derivados de proteínas endógenas, haciendo así accesible su proteoma a los linfocitos T CD8. La presentación antigéncia está acoplada con la propia biosíntesis de las moléculas del MHC I. Las moléculas del MHC I presentan fundamentalmente péptidos procedentes de proteínas sintetizadas en el citosol. Estos péptidos provienen mayoritariamente de productos defectivos de la síntesis proteica (DRiPs) (Qian, et al., 2006; Yewdell, et al., 1996), o de productos de la degradación metabólica de proteínas maduras al final de su vida útil. La principal vía de procesamiento antigénico es el sistema ubiquitina-proteasoma (Ciechanover, et al., 2000; Glickman and Ciechanover, 2002) (Fig. 1). Sin embargo, existen otras enzimas que pueden generar epítopos de una manera independiente del proteasoma, como la tripeptidil peptidasa II (TPPII) (Geier, et al., 1999), la furina (Del-Val and Lopez, 2002; Gil-Torregrosa, et al., 1998), la catepsina S (Shen, et al., 2004) o cisteín-proteasas (Lopez and Del Val, 1997). No obstante, los péptidos antigénicos procedentes de estas vías alternativas son, desde un punto de vista cuantitativo, menos importantes.

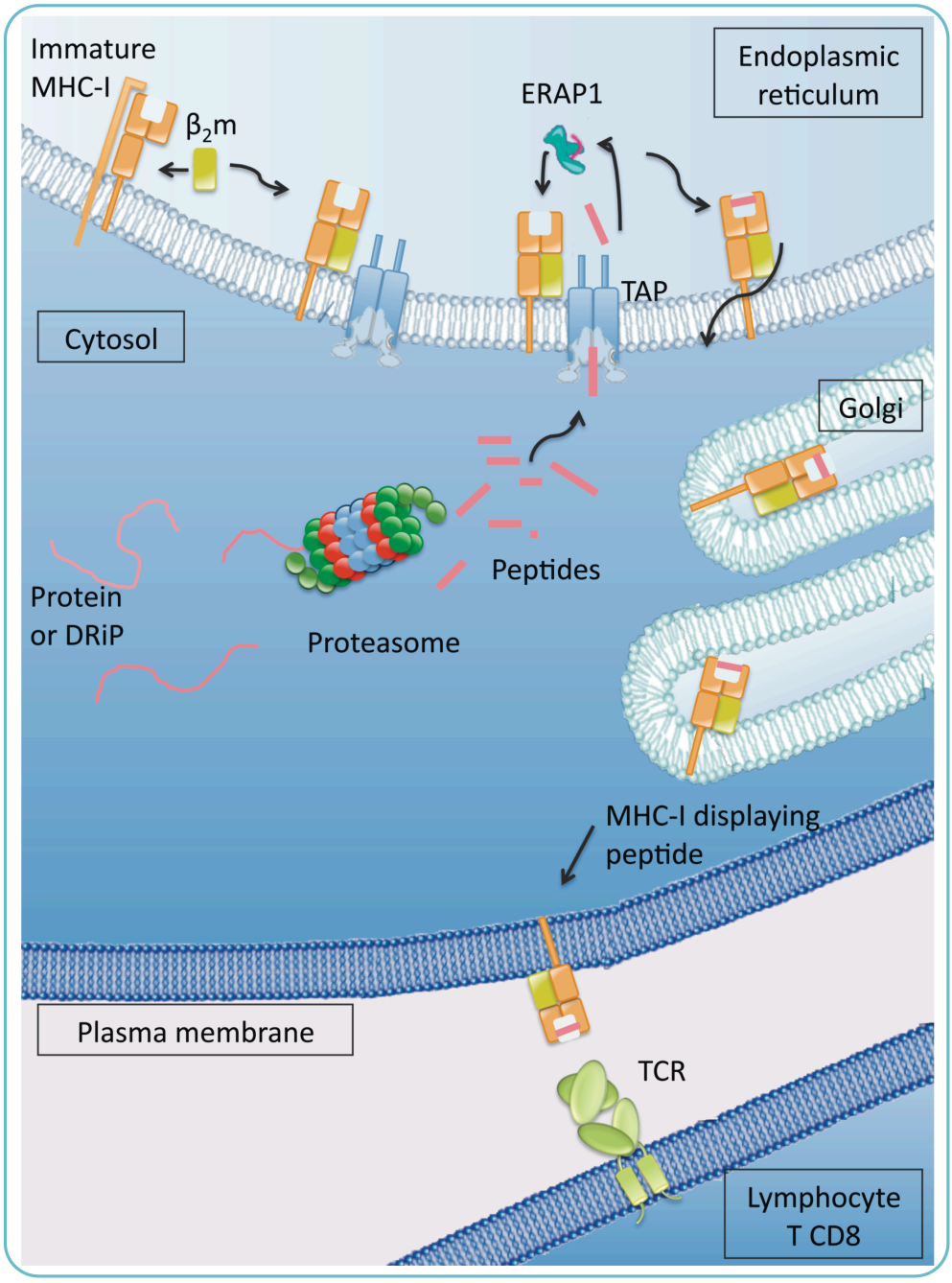


Figure 1. Overview of peptide transport and loading onto MHC I molecules. Peptides generated from cytosolic proteins and DRiPs by the proteasome are translocated into the ER by TAP. If necessary, ERAP1 further trims the N-terminal of the transported peptides to the canonical 9-11 residues required form binding to MHC I molecules. Mature MHC I heterotrimers, consisting of the MHC I heavy chain, the β_2m and the peptide, migrate via the Golgi apparatus to the cell surface where they can be recognized by the TCR of CD8 T cells.

1.1.1 Degradación de proteínas vía ubiquitina-proteasoma

Las proteínas celulares que van a ser destruidas son "marcadas" mediante la unión covalente con el cofactor ubiquitina (Ciechanover, et al., 2000). La ubiquitina es una pequeña proteína presente en todas las células nucleadas y cuya estructura está altamente conservada en la evolución (Finley and Chau, 1991; Schlesinger, et al., 1975). El proteasoma 26S, la maquinaria central de proteolisis de la célula (Pamer and Cresswell, 1998; Rock and Goldberg, 1999), es una proteasa ATP-dependiente formada por un complejo catalítico *core* 20S rodeado en ambos extremos por un complejo regulador 19S (Voges, et al., 1999). Una de las principales funciones de la unidad 19S es el reconocimiento de la "etiqueta" de ubiquitina de las proteínas que van a ser degradadas, evitando así que el proteasoma destruya las proteínas intracelulares de manera indiscriminada (Kloetzel, 2001; Peters, 1994). La unidad *core* 20S está formada por cuatro anillos que forman una cámara interna donde tiene lugar la degradación proteolítica (Groll, et al., 1997; Voges, et al., 1999). Cada uno de los anillos exteriores está formado por 7 subunidades α (α1 - α7) mientras que los anillos internos contienen 7 subunidades β (β1- β7). La actividad proteolítica del proteasoma está mediada por las subunidades β5 (X, LMP7), β2 (Z, MECL-1) y β1 (Y, LMP2), que cortan preferentemente después de residuos hidrofóbicos, básicos y ácidos, respectivamente. Estas tres actividades han sido definidas frecuentemente como actividades tipo quimiotríptica, tríptica y caspasa, respectivamente (Baumeister, et al., 1998; Nussbaum, et al., 1998; Unno, et al., 2002) (Fig. 2). No obstante, estas preferencias no son absolutas y la especificidad también depende de los residuos que rodean el sitio de corte (Eisenlohr, et al., 1992; Nussbaum, et al., 1998). Es importante destacar la relevancia que tiene el lugar de corte del proteasoma en la generación de péptidos presentados por moléculas del MHC I, ya que el extremo C-terminal de estos péptidos corresponde con el residuo P1 del sitio de corte del proteasoma (Rock and Goldberg, 1999; Rock, et al., 1994).

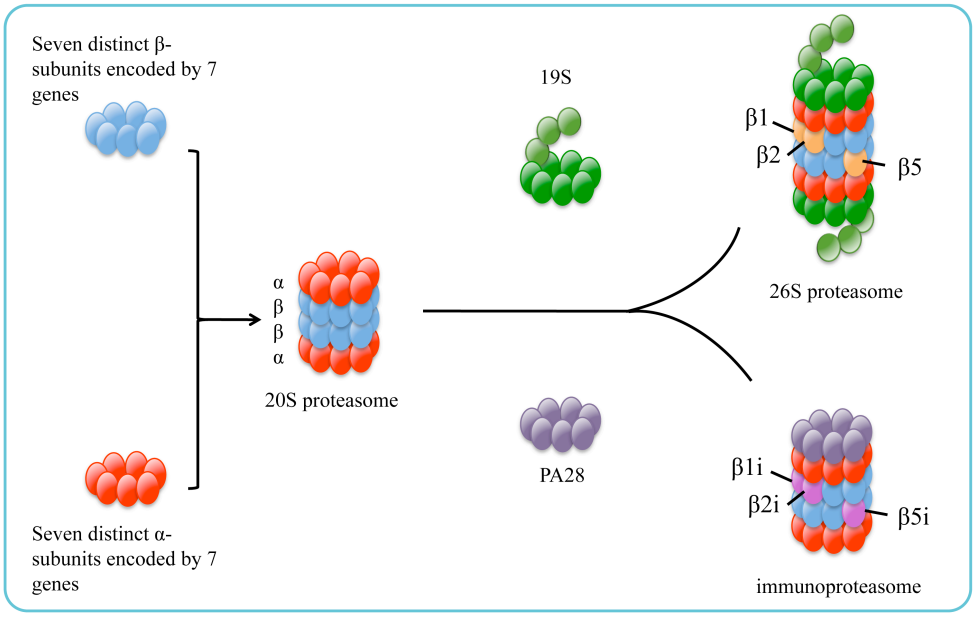


Figure 2. The proteasome. The 26S proteasome is a multicatalytic protease that consists of two α and two β rings forming a hollow cylindrical structure in which proteolysis occurs (20S proteasome). Each of the two inner rings of the 20S proteasome is composed of seven different β-subunits, which host the three different catalytic sites on the inner surface of the 20S proteasome complex, β5, β2 and β1. These proteolytically active sites mediate the hydrolysis of proteins at the C-terminus of hydrophobic, basic and acidic residues, and are referred as the chymotrypsin-like, trypsin-like and peptidylglutamyl-peptide hydrolytic activities, respectively. Upon IFN-γ exposure, the three catalytic subunits of the constitutive 20 S core can be replaced by three new catalytic subunits, β5i, β2i, and β1i. This new form of proteasome is called immunoproteasome. The regulatory complex 19S opens the channel through the 20 S core and unfolds ubiquitinated proteins to allow their entry to the catalytic core; both processes require ATP. The PA28 (11S) regulatory complex can also bind to the 20S core, it has been identified as an activator of the 20S proteasome activities, assessed with small synthetic peptides, and this occurs in a ATP- and ubiquitin-independent manner.

El proteasoma se encuentra presente de manera constitutiva en todas las células, pero existe otra forma, denominada inmunoproteasoma, que se expresa de forma constitutiva en las células dendríticas (Morel, et al., 2000). El inmunoproteasoma también puede encontrarse en otros tipos celulares al ser inducido en presencia de citoquinas proinflamatorias, principalmente

interferón-γ (INFγ) (Griffin, et al., 1998). En el inmunoproteasoma las tres subunidades catalíticas son sustituidas por inmunosubunidades homólogas denominadas β5i (LMP2), β2i (MECL-1), y β1i (LMP2) (Groettrup, et al., 1997) (Fig. 2). Debido a este cambio, disminuye la capacidad de cortar después de residuos ácidos y aumenta la actividad tríptica y quimiotríptica (Toes, et al., 2001). Así, el inmunoproteasoma genera, principalmente, péptidos con residuos C-terminales básicos e hidrofóbicos, razón por la que es más eficiente produciendo ligandos de unión a moléculas del MHC I que el proteasoma constitutivo.

El tamaño de los productos generados por ambos proteasomas es muy semejante, ambos originan fragmentos de entre 3 y 25 aminoácidos, independientemente de la proteína sustrato (Kisselev, et al., 1999). La mayoría de los fragmentos tienen menos de 8 residuos y son, por tanto, demasiado cortos para ser presentados por moléculas del MHC I. Las proteínas contienen numerosos sitios de corte potenciales, de modo que el sitio por el que finalmente el proteasoma o el inmunoproteasoma corta depende de un gran número de factores, incluyendo la especificidad de sus tres actividades catalíticas, y quizás también la distancia entre los sitios de corte, los residuos que rodean el sitio de corte o la concentración del sustrato en el proteasoma. La especificidad mejor caracterizada es la del aminoácido que corresponde al extremo C-terminal del fragmento generado (residuo P1), aunque existen varios estudios donde se demuestra la importancia que el aminoácido flanqueante al extremo C-terminal (residuo P1') tiene en la especificidad del sitio de corte (Altuvia and Margalit, 2000; Beekman, et al., 2000; Mo, et al., 2000). Aunque están menos caracterizadas, también existen otras posiciones que pueden influenciar la especificidad del sitio de corte (Nazif and Bogyo, 2001).

Como se ha demostrado en numerosos trabajos realizados con inhibidores del proteasoma, éste es el responsable de la generación de la mayor parte de los péptidos presentados por moléculas del MHC I (Goldberg, et al., 2002; Mo, et al., 1999; Pamer and Cresswell, 1998; Rock, et al., 1994). Sin embargo, también existen de otras vías de procesamiento de antígenos independientes de proteasoma (Luckey, et al., 2001; Rock, et al., 1994).

El extremo C-terminal de los fragmentos conserva el sitio de corte del proteasoma y/o el inmunoproteasoma. No ocurre así con el extremo N-terminal que sufre la actividad de las aminopeptidasas (Craiu, et al., 1997; Mo, et al., 1999; Serwold, et al., 2002). Este recorte es llevado a cabo por principalmente, en el interior del retículo endoplasmático donde los fragmentos peptídicos son transportados a través de TAP.

1.1.2 Transporte y procesamiento de péptidos en el retículo endoplasmático

Los péptidos generados en el citosol que poseen un tamaño óptimo, entre 8 y 16 aminoácidos, son transportados a través del complejo TAP al interior del RE, donde se unen a las moléculas del MHC I. TAP es un heterodímero compuesto por dos subunidades, TAP1 y TAP2 (Abele and Tampe, 2004) (Fig. 3), pertenecientes a la gran familia de transportadores ABC (ATP *binding casette*). TAP une e hidroliza ATP para la translocación de péptidos a través de la membrana del ER. La expresión de ambas subunidades TAP1 y TAP2 es necesaria para conseguir una translocación eficiente, si bien, ciertos péptidos se unen preferentemente a TAP1, mientras que otros se unen a TAP2 (Androlewicz and Cresswell, 1994; Androlewicz, et al., 1994; Nijenhuis and Hammerling, 1996). TAP está codificado por genes localizados en la región de clase II del MHC estrechamente vinculados a los genes LMP2 y LMP7 que codifican las subunidades del proteasoma inducibles por INFγ (Lankat-Buttgereit and Tampe, 2002). Cada subunidad TAP tiene una región N-terminal hidrofóbica transmembrana, y un dominio de unión a ATP en el extremo C-terminal citosólico (Schrodt, et al., 2006; Vos, et al., 1999).

El transporte mediado por TAP tiene dos pasos secuenciales, primero el péptido se une a TAP y posteriormente es translocado al interior del ER consumiendo ATP (Neefjes, et al., 1993; Shepherd, et al., 1993; van Endert, et al., 1994). Siendo la afinidad de los péptidos a TAP la que controla la cinética de translocación. Además de las preferencias de longitud (8 – 16 aminoácidos) (Androlewicz and Cresswell, 1994; Momburg, et al., 1994), también se ha visto que

los primeros tres residuos del extremo N-terminal y el residuo del extremo C-terminal del péptidos son importantes para la unión a TAP. Estos péptidos presentan generalmente un extremo C-terminal básico o hidrofóbico, los residuos preferidos por las moléculas del MHC I. Igualmente, TAP tiene preferncia por los péptidos con residuos hidrofóbicos en la posición 3 (residuo P3) y residuos cargados o hidrofóbicos en la posición 2 (residuo P2). Por otro lado, residuos aromáticos o ácidos en P1 y prolinas en P1 y P2 tienen efectos perjudiciales (Momburg, et al., 1994; Uebel, et al., 1997;van Endert, et al., 1995).

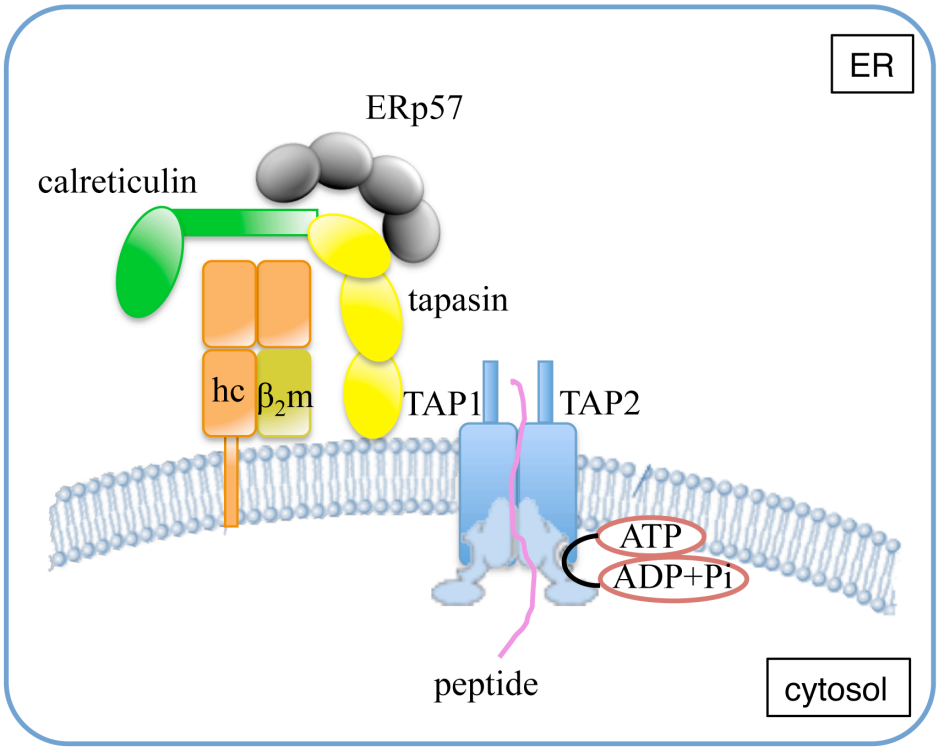


Figure 3. TAP and peptide loading complex. TAP is a heterodimeric protein, made up of the subunits TAP1 and TAP2, that is localized in the ER. TAP is a member of the ABC transporter family of proteins. These proteins consume ATP in order to translocate peptides from the cytosol into the ER. Within the ER, nascent MHC I molecules associate with calreticulin, tapasin and ERp57 to form the peptide-loading complex (PLC), which facilitates the loading of peptides into the MHC I peptide-binding groove.

Muchos de los péptidos transportados por TAP no tienen una longitud óptima para unirse a las moléculas del MHC I, son demasiado largos. ERAAP tiene actividad aminopeptidasa y es esencial para el recorte del extremo N-terminal de los péptidos haciendo que estos alcancen el tamaño óptimo de unión a las moléculas del MHC I (9 – 11 aminoácidos) (Saric, et al., 2002; Serwold, et al., 2002).

ERAAP actúa de manera sucesiva sobre el sustrato, y presenta una preferencia por sustratos con una longitud de entre 8 y 16 aminoácidos, precursores eficientemente trasportados por TAP. Se ha visto que esta preferencia parece estar basada en su capacidad para monitorizar la naturaleza del residuo C-terminal, así como otros residuos en la secuencia del péptido (Evnouchidou, et al., 2008) y el número de residuos hasta el extremo N-terminal. Este mecanismo mediantes el cual ERAAP "conoce" el número de aminoácidos del péptido se llama "regla molecular", y le permite disminuir su actividad cuando el péptido alcanza una longitud de 8 ó 9 aminoácidos (Chang, et al., 2005). ERAAP tiene preferencia por residuos C-terminales hidrofóbicos, presentes en muchos ligandos de moléculas del MHC I. Se ha visto que un residuo básico en esta misma posición disminuye drásticamente el recorte peptídico. De forma similar, se ha comprobado un efecto claro, pero menor, con los residuos ácidos o hidrofóbicos (Chang, et al., 2005). ERAAP degrada completamente los precursores antigéncos en el ER cuando éstos no tienen un motivo de unión a las moléculas del MHC I o en ausencia de moléculas del MHC I (Kanaseki, et al., 2006).

Recientemente se ha visto que existen dos formas de ERAAP, ERAP1 y ERAP2, ambas involucradas en los recortes peptídicos. ERAP1 y ERAP2 se localizan juntas en el ER y asociadas formando complejos, que normalmente son heterodímeros. Aunque se ha visto que existen algunas diferencias en la preferencia de sustratos tienen funciones complementarias (Saveanu, et al., 2005).

1.1.3 Carga de péptidos en las moléculas del MHC I y presentación en la superficie celular

Existen dos tipos diferentes de moléculas del MHC, las moléculas de clase I (MHC I) y las de clase II (MHC II), que se diferencian tanto en su estructura como en la función y el origen de los péptidos que presentan. Las moléculas del MHC II se cagan en vesículas especializadas con péptidos resultantes de la degradación de proteínas fagocitadas y son reconocidos por los linfocitos T CD4. Por el contrario, los péptidos generados en el citosol (generalmente procedentes de proteínas sintetizadas endógenamente), siguiendo la vía del proteasoma aquí descrita, se asocian a moléculas de clase I y son presentados a los linfocitos T CD8. En humanos, las moléculas del MHC se conocen como moléculas HLA (antígenos leucocitarios humanos) y están codificadas en la región p21 del cromosoma 6. Hay tres tipos de moléculas HLA de clase I (HLA I) presentadoras de antígenos clásicas: HLA-A, HLA-B y HLA-C. En la población existen cientos de alelos MHC I diferentes, haciendo que el locus del MHC sea el sistema génico más polimórfico que se conoce (Terasaki, 2007). La expresión de genes del MHC I es codominante y en un solo individuo se expresan hasta 6 moléculas diferentes del MHC. El polimorfismo del MHC I determina la especificidad de unión de péptidos (Reche and Reinherz, 2003), es decir, las variantes alélicas del MHC I unen, generalmente, péptidos distintos. Cada molécula del MHC se une a un gran repertorio de péptidos que se estima entre 1000 y 10000 ligandos diferentes (Hillen and Stevanovic, 2006).

Las moléculas del MHC I son glicoproteínas de membrana formadas por una cadena α polimórfica de 45 kDa unida no covalentemente a una cadena ligera β no polimórfica denominada β_2-microglobulina (β_2m) y un péptido de 8-11 aminoácidos (Fig. 4). La cadena α consta de una región extracelular de 274 aminoácidos, formada por tres dominios (α_1, α_2 y α_3), una región transmembrana de 25 aminoácidos y una región intracelular o citoplásmica de 30 aminoácidos. El domino α_3 presenta una estructura típica de inmunoglobulina y es donde se

concentran los polimorfismos. La β_2m presenta una estructura muy parecida al dominio α_3 (Orr, et al., 1979; Orr, et al., 1979) e interacciona con los tres dominios α_1, α_2 y α_3 contribuyendo significativamente a la estabilidad de la molécula (Madden, 1995; Madden, et al., 1993).

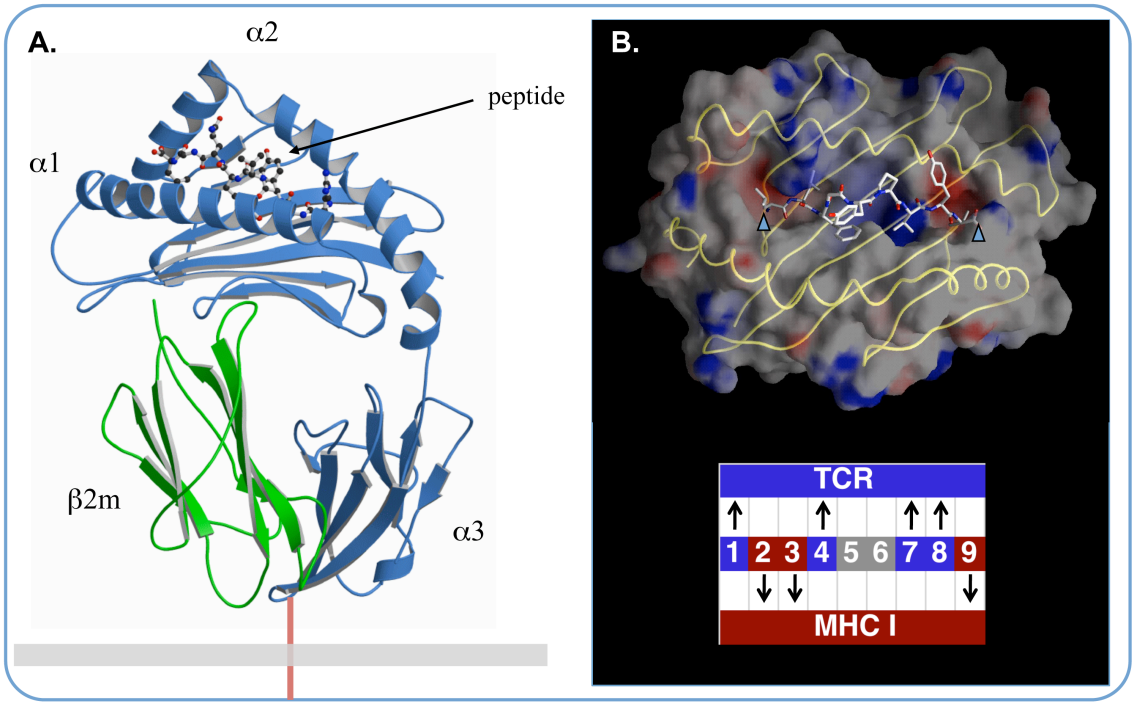


Figure 4. MHC I. A) Ribbon representation of the MHC I 3D structure. MHC I molecules are composed of two chains. The α chain encompasses three α domains, α_1, α_2 and α_3. Domains α_1 and α_2 form a groove where the peptide is located. The β_2-microglobulin (β_2m) is a small globular protein, which associates primarily with the α_3 domain and is necessary for MHC stability. **B)** Peptide binding groove of MHC I molecules. The figure illustrates a view of the α_1 and α_2 domains of HLA-A*0201 with the bounded peptide LLFGYPVYV from HTLV-1 TAX protein (PDB: 1HHK) as seen by the T-cell receptor (TCR). The peptide bound to this molecule is represented by sticks to highlight the contours of the binding groove. The peptide binding groove of the MHC I molecule is closed, and peptides bind in such a manner that both, the N-terminus and C-terminus of the peptide (indicated by arrow heads) are nested into the MHC I binding groove, restricting their length to 8–11 residues. The general binding mode of peptides to MHC I is shown at the bottom of the panel. In this representation, peptide positions contacting the TCR and MHC are shaded in red and blue, respectively, and are also indicated with opposing arrows. Positions shaded in grey can be anchor or TCR contact positions, depending of the specific MHC I molecule.

Las moléculas del MHC I unen péptidos de pequeño tamaño (8-11 aminoácidos), quedando sus extremos N- y C-terminales conectados a los residuos conservados de la molécula del MHC I mediante una red de puentes de hidrógenos (Madden, 1995; Matsumura, et al., 1992; Zhang, et al., 1998). El péptido se une en un surco formado por los dominios α_1 y α_2. Se han definido 6 subcavidades o "*pockets*" denominados A-F, cuya forma, tamaño y polaridad están determinados por las cadenas laterales de los residuos que los conforman, que con frecuencia son altamente polimórficos y proporcionan propiedades únicas en la unión de péptidos a cada moléculas del MHC I (Reche and Reinherz, 2003). En estas subcavidades encajan residuos específicos denominados residuos de anclaje. Generalmente las posiciones 2 y 3, y el extremo C-terminal de los péptidos que se unen a las moléculas del MHC I son los residuos de anclaje (Fig. 4), aunque las preferencias por residuos específicos en las posiciones de unión cambia entre las moléculas del MHC I. Cabe destacar que los péptidos que tienen tamaños diferentes y se unen a las mismas moléculas del MHC I, normalmente, utilizan subcavidades de unión alternativas (Madden, et al., 1993; Reche, et al., 2006).

La carga de péptidos en las moléculas del MHC I está asociada con su biosíntesis e intervienen distintas proteínas dando lugar al PLC (complejo de carga peptídica). El PLC está formado por la molécula del MHC I, TAP, tapasina, calreticulina y ERp57 (Peaper and Cresswell, 2008) (Fig. 3). La tapasina es una chaperona importante para la optimización de la carga peptídica y la eficiencia de la presentación de péptidos por las moléculas del MHC I (Garbi, et al., 2001; Momburg and Tan, 2002; Tan, et al., 2002). Entre sus funciones está la de estabilizar a TAP mediante su asociación a través del dominio transmembrana (Ortmann, et al., 1997), aumentar su expresión e incrementar el aporte de péptidos al lumen del ER (Bangia, et al., 1999; Garbi, et al., 2003) y servir de puente entre TAP y los antígenos MHC I para facilitar la carga peptídica (Lehner, et al., 1998; Li, et al., 2000; Sadasivan, et al., 1996; Tan, et al., 2002). Este

proceso tiene una gran relevancia inmunológica ya que permite seleccionar péptidos de alta afinidad (Elliott and Williams, 2005; Momburg and Tan, 2002; Sijts and Pamer, 1997).

1.1.4 Reconocimiento antigénico por el TCR de linfocitos T CD8

Cada linfocito T CD8 presenta un único TCR que reconoce de manera específica péptidos presentados por una molécula del MHC I determinada. El reconocimiento es específico tanto de la molécula del MHC I como del péptido presentado por ella. Este fenómeno se denomina restricción por moléculas del MHC I (Solheim, 1999). El TCR es un complejo formado por un heterodímero compuestos por las cadenas α y β, generalmente, que se asocia a una serie de cadenas invariantes: CD3 (γ, δ y ε) y CD247 (ζ) (Fig. 5). El reconocimiento del antígeno se produce gracias a las cadenas α y β. Estas cadenas tienen una región variable y otra constante. Las regiones variables están formadas por regiones determinantes de complementariedad (CDRs), que definen la especificidad de unión al complejo péptido-MHC I. Antes de alcanzar la membrana, el heterodímeros del TCR se asocia de manera ordenada a las cadenas CD3. Este hecho tiene lugar en el retículo endoplásmico, donde las cadenas CD3 se unen entre si a través de residuos de carga opuesta en la región transmembrana, justo antes de la unión por puentes disulfuro de las cadenas TCR entre si (Call and Wucherpfennig, 2005; Call and Wucherpfennig, 2007). Las cadenas CD3 son invariantes y se agrupan de manera no covalente en dímeros (γε y δε) (Fig. 5. Estas cadenas ayudan, por una parte, a expresar las cadenas del TCRαβ y, por otra, a transmitir la señal de reconocimiento al interior celular (Dave, 2011). La capacidad señalizadora del complejo TCR/CD3 reside en una secuencia altamente conservada presente en la parte citoplasmática de todas las cadenas CD3 denominada ITAM (*Immunoreceptor Tyrosine-based Activation Motifs*).

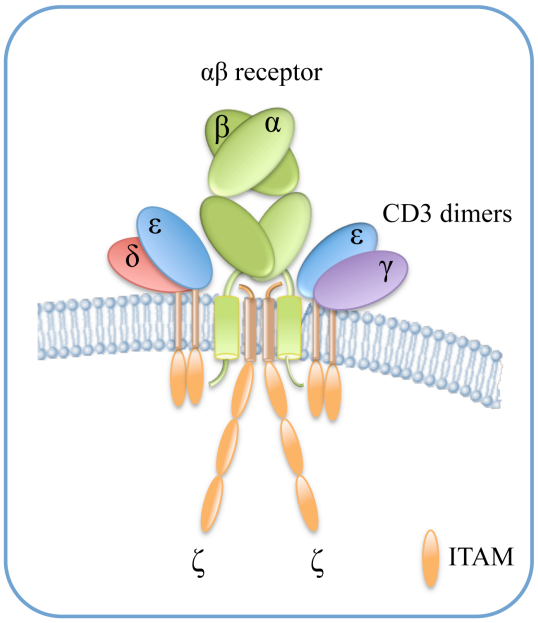


Figure 5. The TCR is a heterodimer composed of a α and a β chain. Both chains have a constant region and a variable region. The variable region of both chains contains hypervariable regions that determine the antigen specificity. Each T cell bears a TCR of only one specificity. The TCR is closely associated to the CD3 chains (γ, δ and ε). The CD3 chains are invariant and they do not contribute to the specificity in any way. The CD3 chains are necessary for cell surface expression of the TCR during T cell development. In addition, the CD3 chains transduce activation signals to the cell following antigen interaction with the TCR.

Sólo cuando el complejo TCR/CD3 reconoce su combinación antígeno-MHC particular, se produce la activación de los linfocitos T CD8 y estos adquieren sus funciones efectoras de manera específica (Alarcon, et al., 2003). La mayor parte de las interacciones del TCR son con la molécula del MHC I, y tan sólo contacta con algunos residuos del péptido, cuyas cadenas son accesibles al TCR (Garcia, et al., 1999; Rudolph and Wilson, 2002).

1.2 Relevancia de la identificación de epítopos T CD8

La identificación de epítopos de células T CD8 es importante para comprender la patogénesis de las enfermedades (Tchernev and Orfanos, 2006). Además, es la base para el desarrollo de vacunas basadas en epítopos contra agentes infecciosos (Reche, et al., 2006) y tratamientos para alergias (Akdis, et al., 1996), enfermedades autoinmunes (Stienekemeier, et al., 2001) y enfermedades neoplásicas (Lazoura and Apostolopoulos, 2005). Tradicionalmente, la identificación de epítopos de células T requería la síntesis de péptidos solapantes que abarcasen la secuencia completa de la proteína, seguida de ensayos experimentales para cada péptido para

determinar la activación de las células T, como por ejemplo, la producción de citoquinas (Draenert, et al., 2003). Este método es costoso y laborioso, por lo que es viable sólo para proteínas únicas o patógenos que consistan en pocas proteínas. Por ello, en esta Tesis doctoral se han desarrollado distintas herramientas computacionales que tratan de facilitar la identificación y selección de epítopos T CD8.

2. OBJETIVOS

La presente Tesis se centra en descifrar las fuentes que determinan la inmunogenicidad de los epítopos T CD8. Para ello realizamos el modelado *in silico* de la vía clásica del procesamiento de antígenos T CD8. Además se han desarrollado otras herramientas computacionales relacionadas con la predicción y elección de epítopos T CD8. Los objetivos concretos que se han abordado son:

1. Análisis de la distribución de epítopos T CD8 en las proteínas de varios virus.
2. Desarrollo de un método para la predicción de los sitios de corte por el proteasoma constitutivo y el inmunoproteasoma.
3. Desarrollo de un método para predecir y analizar la capacidad de unión de péptidos a TAP.
4. Desarrollo de una herramienta computacional que permita calcular la variabilidad de las secuencias para facilitar la identificación de epítopos conservados.
5. Integración de las anotaciones de epítopos T CD8 con la información funcional y estructural de los antígenos fuente.

3. MATERIAL Y MÉTODOS

3.1 Bases de datos

Los distintos modelos predictivos desarrollados en esta Tesis doctoral están basados en secuencias de epítopos T CD8 y ligandos eluidos de moléculas del MHC I obtenidos de distintas bases de datos (Tabla I).

Table I. *DataBases*.

BASE DE DATOS	TIPO DE DATOS	WEB	REFERENCIA
EPIMHC	Péptidos eluidos de moléculas del MHC y epítopos T que han sido observados en proteínas reales.	http://imed.med.ucm.es/epimhc/	(Reche, et al., 2005)
IMMUNEEPITOPE (IEDB)	Anticuerpos y epítopos de células T de humanos, primates no humanos, roedores y otras especies animales. También incluye información sobre la unión de péptidos a moléculas del MHC de una gran cantidad de antígenos.	http://www.immuneepitope.org/	(Vita, et al. 2010)
Los Alamos DB	Secuencias genéticas de HIV, epítopos inmunogénicos, mutaciones asociadas a resistencia, y ensayos clínicos sobre vacunas.	http://www.hiv.lanl.gov/	
AntiJen	Información cuantitativa de péptidos que se unen a moléculas del MHC, TAP, complejos TCR-MHC, epítopos T, epítopos B, e interacciones inmunológicas proteína-proteína	http://www.ddg-pharmfac.net/antijen/AntiJen/antijenhomepage.htm	(Toseland, et al., 2005)

3.2 Métodos computacionales

En esta sección se describen los algoritmos de inteligencia artificial y otros métodos computacionales en los que hemos basado los modelos predictivos.

3.2.1 *N-grams*

Los *N-grams* son modelos probabilísticos que se utilizan fundamentalmente en el estudio de reglas gramaticales y en programas de reconocimiento de voz (Rosenfeld, 2000). También se han utilizado en bioinformática, para el análisis de secuencias de DNA y proteínas (Jimenez-Montano, et al., 2002; Reche, et al., 2004; Wu and Shivakumar, 1994; Wu, et al., 1996). En esencia, los *Ngrams* consisten en *Hidden Markov Models* de *n* elementos (letras, sílabas, palabras) conectados por probabilidades.

El problema de la predicción de sitios de corte del proteasoma recuerda al de predicción de signos gramaticales de puntuación, por ello, aplicamos *N-grams* para el desarrollo de los modelos de predicción de los sitios de corte del proteasoma y del inmunoproteasoma. En concreto, usamos el paquete SRLIM (Stolcke, 2002). En este caso, los *N-grams* se entrenan para reconocer el sitio de corte del proteasoma, definido por los péptidos presentados por moléculas del MHC I y los aminoácidos flanqueantes de su extremo C-terminal obtenidos de la proteína fuente. Para ello, se generaron distintos grupos de datos con fragmentos de distintos tamaños como aparece en la figura 6.

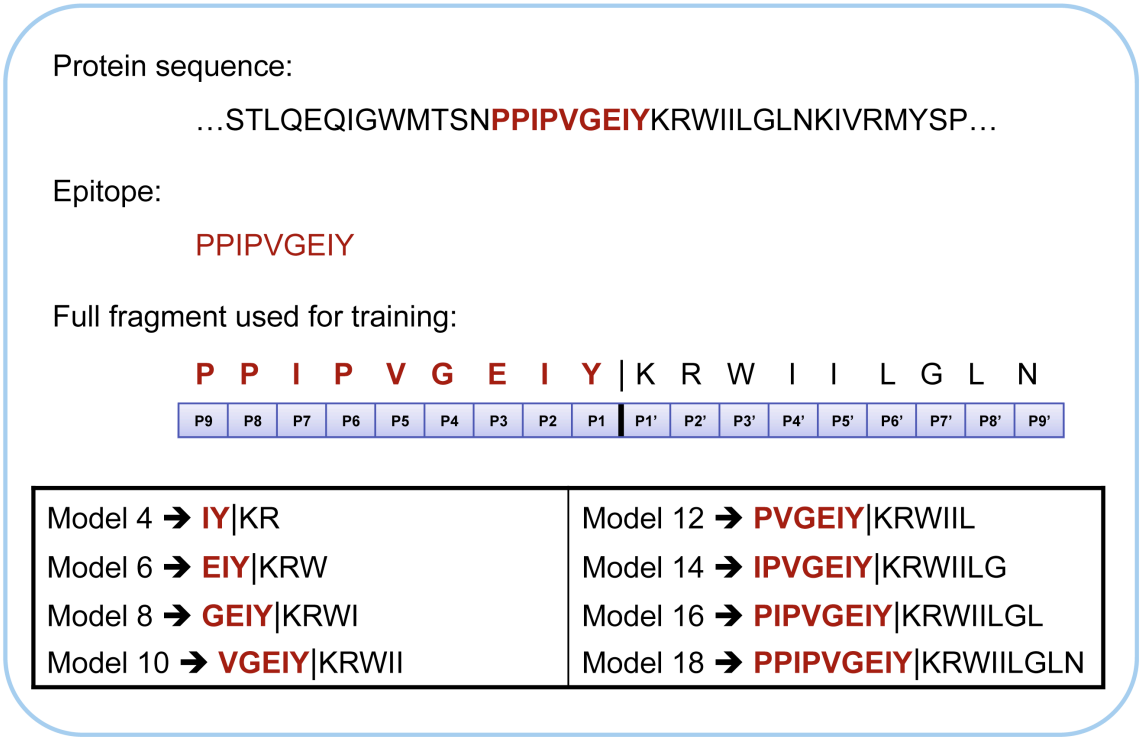
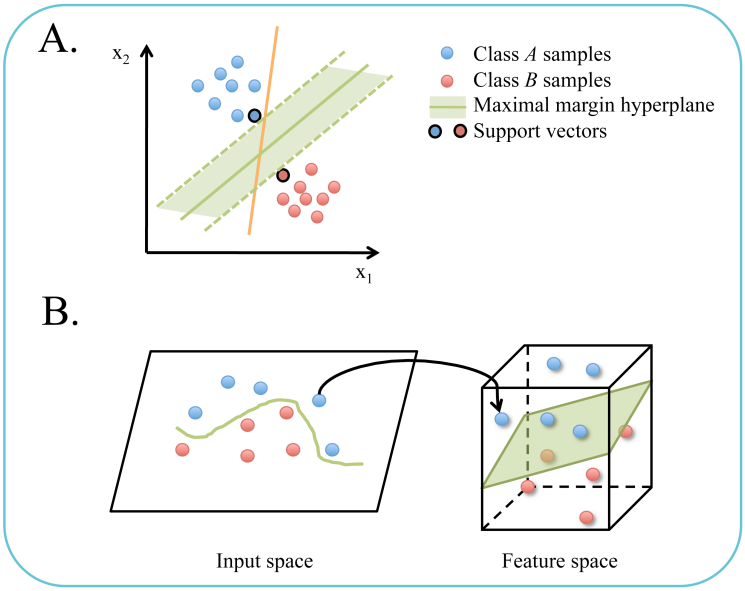


Figure 6. Proteasome and immunoproteasome *N-gram* cleavage models development. Cleavage models were trained and evaluated on datasets consisting of peptide fragments of the same length derived from MHCI-eluted peptides (proteasome models) and CD8 T cell epitopes (immunoproteasome models) and their C-terminal flanking regions, using *N-grams*. Peptide fragments encompassed two portions with the same number of residues, one fraction consisting of the C-terminal end of the peptide, and the other one of their C-terminal flanking region. Cleavage sites −defined between the C-terminus of MHCI-restricted peptides (P1 residue of cleavage site) and the most proximal C-terminal flanking residue (P1' residue)− were indicated by a "|" symbol.

3.2.2 Máquinas de vectores de soporte (SVMs; Support vector machines)

Los SVMs consisten en una técnica de aprendizaje supervisado capaces de resolver problemas de clasificación y regresión (Vapnik, 1998). Los SVM fueron introducidos por Vapnik y colaboradores en 1963 para la clasificación binaria de datos lineales (Burges, 1998; Vapnik, 1995). La principal idea de las clasificaciones de SVM es encontrar plano que maximizar el margen de separación entre las clases (Cristianini and Shawe.Taylor, 2000). Tal y

como se representa en la figura 7A, el plano de separación es aquel que separa objetos de distinta clase, y los puntos que definen dicho plano son los vectores de soporte (*support vectors*).



Figrue 7. SVM classification approach. SVM are used for classification and regression analysis. **A.** A simple 2D example for the decision algorithm in a SVM. **B.** The kernel function may transform the data into a higher dimensional space to make it possible to perform the separation.

Pero la mayoría de las clasificaciones no son tan simples, y generalmente necesitan estructuras complejas que permitan una separación óptima de los datos. Para la clasificación de datos no lineales mediante SVM, Bernhard Boser, Isabelle Guyon y Vapnik introdujeron el uso de las funciones kernels, que permiten distribuir los datos iniciales en un espacio de más dimensiones donde si es posible clasificar los datos utilizando planos de decisión lineales (Fig. 7B) (Bernhard E. Boser, et al., 1992). Esto ajustar el margen máximo del hiperplano en un espacio transformado. Matemáticamente, la funcion de decisión de SVM se escribe de acuerdo a la eq. 1:

$$f(x) = \sum_{i=1}^{m} y_i \alpha_i k(x_i, x) + b \qquad f(x) \begin{cases} 1 \\ -1 \end{cases} \qquad \text{Equation 1}$$

donde $x_i \in R^n$, i = 1, 2,…,n son los elementos de entrenamiento, m es el número de elementos de entrada cuyo valor es distinto de cero dentro de α_i (*Language multipliers,* normalmente, un subconjunto de *n support vectors* conocidos), b es el termino del sesgo, y $k(x_i, x)$ denota la función kernel.

Existen numerosos tipos de kernel, en concreto, aquí aplicamos un kernel Gaussiano (Eq. 2). Este kernel transforma cada elemento de los datos de entrenamiento en un punto en un espacio n-dimensional, donde x_i y x_j son los datos de entrada y γ define la anchura del kernel.

$$k(x_i, x_j) = \exp\left(-\gamma \left\| x_i - x_j \right\|^2\right) \qquad \text{Equation 2}$$

En esta Tesis, aplicamos SVMs de regresión para predecir la afinidad de unión de péptidos a TAP. Los SVMs de regresión funcionan igual que los de clasificación, pero se entrenan con datos cuantitativos. Los SVMs se desarrollaron utilizando el paquete WEKA (*Waikato Environment for Knowledge Analysis*) (Frank, et al., 2004).

3.2.3 Matrices de puntuación específica (PSSMs; *Position specific scoring matrices*)

Las PSSMs o *profiles* se utilizan para representar patrones o motivos (*motifs*) presentes en alineamientos de múltiples secuencias biológicas. Las PSSMs fueron introducidas en 1987 por Gribskov et *al.* (Gribskov, et al., 1987) para detectar relaciones de homología. En esencia, estas matrices consistían en ratios logarítmicos de frecuencias de aminoácidos observadas en alineamientos de secuencias con respecto a frecuencias de fondo de los mismos aminoácidos. Las frecuencias de fondo se estiman a partir de bases de datos. Los *profiles* también corrigen problemas de redundancia (a través del peso de las secuencias) e información no disponible.

En esta Tesis doctoral se emplean las PSSMs desarrolladas por Reche *et al.* (Reche, et al., 2002; Reche, et al., 2004; Reche and Reinherz, 2007) a partir de péptidos alineados cuya unión a moléculas del MHC es conocida. Estas matrices están disponibles para uso público en la herramienta RANKPEP para la predicción de epítopos (Reche, et al., 2004).

3.3 Desarrollo y evaluación de los modelos predictivos

El desarrollo de modelos predictivos basados en datos obtenidos de bases de datos y/o de la literatura requiere del seguimiento de una metodología que evite el sobreentrenamiento y la obtención de sobreestimaciones del rendimiento. En esencia, esta metodología se basa en no evaluar los modelos predictivos con datos que se han utilizado para el entrenamiento del modelo (Fig. 8). En esta Tesis hemos hecho validaciones cruzadas y con datos independientes para evitar estos problemas.

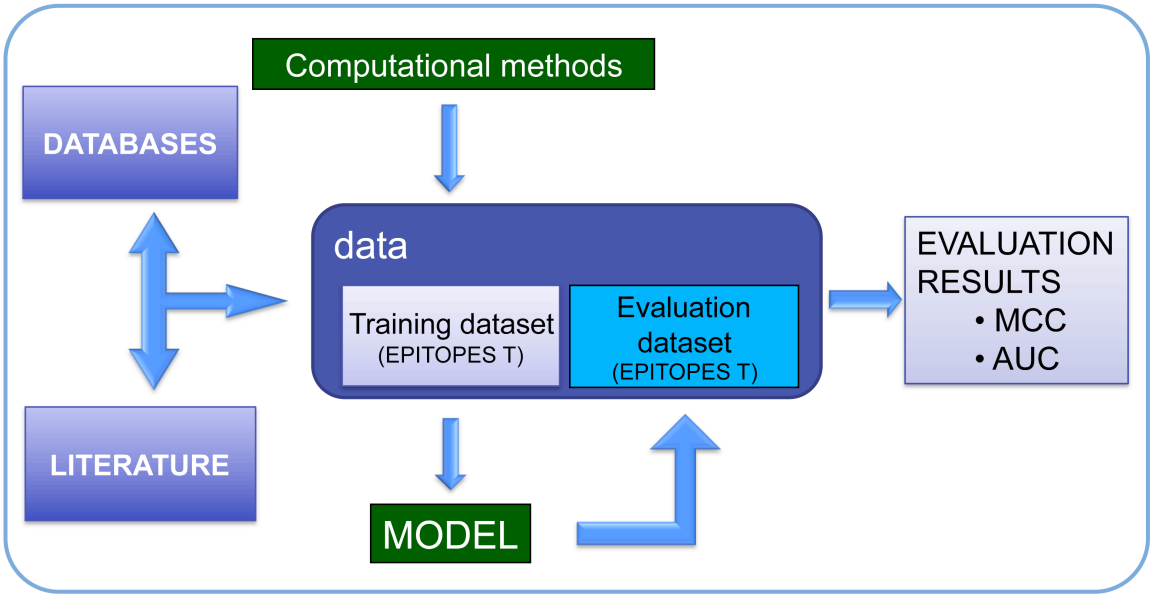


Figure 8. Workflow of the development of a predictive model. First of all, data (epitopes T) are collected from the literature and different databases. Subsequently, the original data is divided in different sets, training datasets and evaluation datasets. The training dataset is used to generate different models based in artificial intelligent algorithms that will be evaluated on the evaluation dataset to calculate different performance parameters (MCC, AUC).

3.3.1 Validación cruzada

La validación cruzada es una técnica empleada para seleccionar los algoritmos y parámetros que maximizan la capacidad predictiva del modelo, así como para evaluar dicha capacidad en datos distintos de los usados para su entrenamiento. Uno de los principales

problemas que presentan los modelos predictivos es el de sobreentrenamiento. Estos problemas de sobreentrenamiento u *overfitting* son más probables cuando el tamaño del grupo de entrenamiento es pequeño, o cuando el modelo tiene muchos parámetros que ajustar, pero la técnica de la validación cruzada de n-campos se puede evitar este problema.

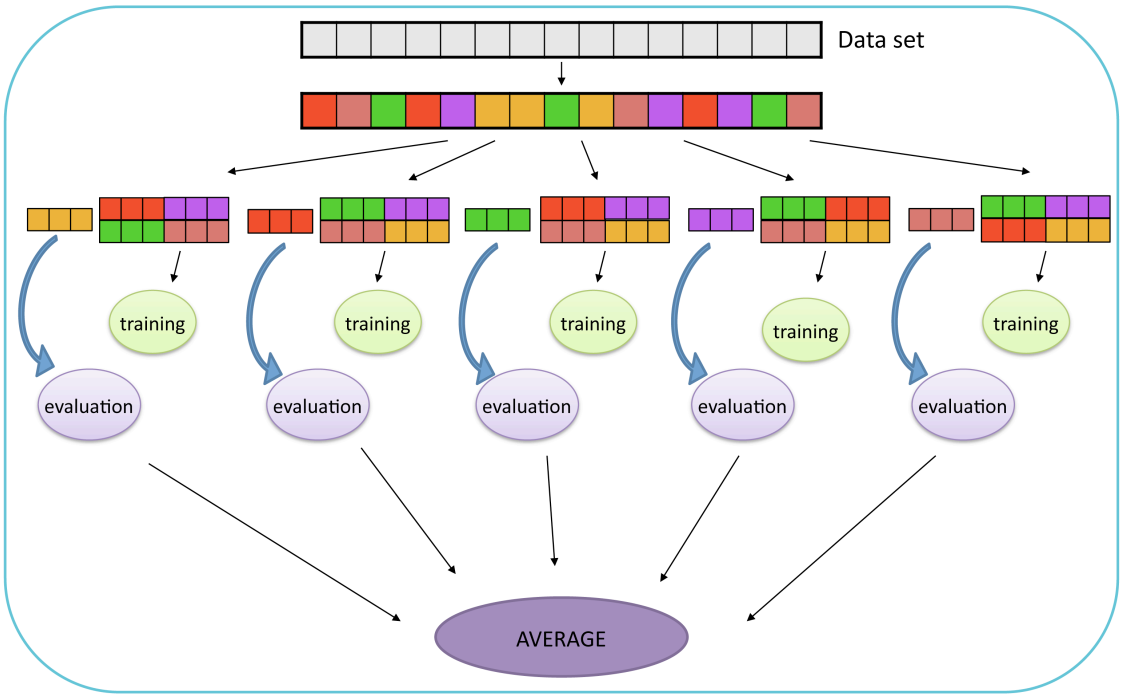


Figure 9. 5-fold cross-validation. The original dataset is randomly divided into 5 sets of the same size. Four of the sets are summed to generate the training dataset that is used to train the model while the other one is used to evaluate such model. The cross-validation process is then repeated 5 times, with each of the five sets used exactly once as the evaluation dataset. The 5 results from the folds give the average value and its standard deviation.

En la validación cruzada de n-campos (*n-fold cross-validation*) (Fig. 9), los datos originales se dividen aleatoriamente en *n* subgrupos del mismo tamaño. De los *n* subgrupos, uno es utilizado para evaluar el modelo generado con el grupo de datos resultado de la suma del resto de subgrupos. Este proceso se repite *n* veces, y cada uno de los *n* subgrupos se emplea únicamente una vez para validar el modelo. Finalmente, la validación cruzada de *n*-campos se repine *k* veces, generalmente 10 ó 100 veces, generando distintos *n* subgrupos cada vez. Los

resultados obtenidos dan finalmente la media y desviación estándar de los parámetros de evaluación.

3.3.2 Test independiente

Cuando es posible, empleamos un grupo de datos independiente para evaluar los modelos predictivos. Este grupo de datos no está presente en los datos de entrenamiento, pero presentan las mismas características que éstos. Por ejemplo, estos datos independientes pueden pertenecer a un organismo distinto a los incluidos en el grupo de entrenamiento. En concreto, los modelos desarrollados en esta Tesis para la predicción de los sitios de corte por el proteasoma y el inmunoproteasoma han sido evaluados en un test independiente. Estos modelos fueron entrenados con datos de epítopos T y ligandos eluidos de moléculas del MHC I de distintos virus, pero no del VIH, utilizándose los epítopos T del VIH con un conjunto de datos independiente para la evaluación de los modelos.

3.3.3 Medidas de rendimiento predictivo

Las medidas de rendimiento predictivo que se han utilizado en esta Tesis son de dos tipos: dependientes e independientes de umbral. Las dependientes de un umbral son la sensibilidad (SE), la especificidad (SP), el coeficiente de correlación de Mathews (MCC) y el parámetro *better than random* (BTR). La medida independiente de umbral es el área bajo la curva ROC (AUC). La capacidad predictiva de los modelos de regresión se ha evaluado determinando los coeficientes de correlación de Pearson o el de Spearman.

La SE es la fracción de casos positivos que son clasificados correctamente como positivos (Eq. 3), mientras que la SP es la fracción de casos negativos que son clasificados correctamente como negativos (Eq. 4).

$$SE = \frac{TP}{TP + FN}$$ Equation 3

$$SP = \frac{TN}{TN + FP}$$ Equation 4

El MCC se emplea como medida de la asociación entre dos variables binarias. Tiene en cuenta los verdaderos y faltos positivos y negativos. Los valores de MCC van entre 1 y -1, donde un coeficiente de 1 indica una correlación, o predicción, perfecta, 0 una correlación aleatoria, y -1 una correlación perfecta pero inversa. El MCC se puede calcular directamente a partir de la tabla de contingencia (Fig 10A) atendiendo a la fórmula:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$ Equation 5

El índice BTR se empleó para comparar la SE de los modelos predictivos del proteasoma constitutivo y del inmunoproteasoma con un modelo aleatorio que predijese el mismo número de sitios de corte. Este parámetro viene dado por la ecuación 6, donde el parámetro ECS (*expected cleavage sites*) representa la proporción de sitios de corte predichos correctamente por un modelo aleatorio.

$$BTR = SE - ECS$$ Equation 6

El análisis mediante las curvas ROC (*Receiver operating charasteristic*) permite evaluar la habilidad de un modelo predictivo a la hora de distinguir entre dos clases. La Curva ROC consiste en la representación gráfica de la sensibilidad frente a (1 – especificidad) a distintos umbrales de decisión (Fig. 10C).

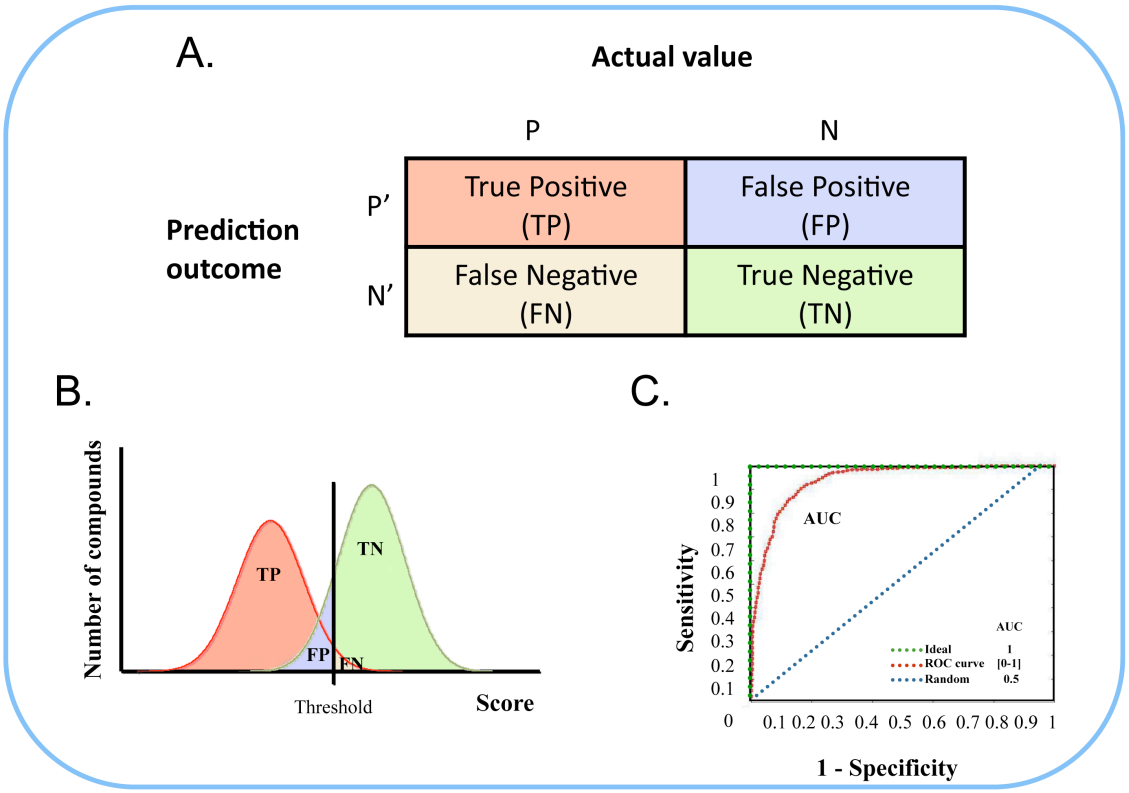


Figure 10. ROC curve. A. Confusion matrix. Upon threshold modifications data is classified as true positive, false positive, true negative or false negative **B.** Theoretical distributions of scores are obtained for both true peptides (red) and false peptides (green) after processing the sample by a suitable computer test. Generally, these distributions overlap, leading to false predictions (FP and FN). **C.** For all possible score thresholds, the evolution of the deduced sensitivity (SE) and specificity (SP) is reported on a ROC graph. Calculating the area under the ROC curve is a practical way to quantify the overall performance of the computer test.

El área que queda bajo la curva ROC es el AUC. El AUC es la medida de exactitud más utilizado en muchos contextos. Este área posee un valor comprendido entre 0,5 y 1, según el cual los modelos se clasifican como:

- AUC > 0,9 = excelente
- 0,8 > AUC > 0,9 = muy bueno
- 0,7 > AUC > 0,8 = bueno
- 0,6 > AUC > 0,7 = malo

- 0,5 > AUC > 0,6 = muy malo

El coeficiente de correlación de Pearson (P_r) y el coeficiente de correlación de Spearman (S_r) miden la existencia de una relación lineal entre dos variables. El coeficiente de correlación de Pearson se calcula dividiendo la covariancia por el producto de las desviaciones estándar de ambas variables (Eq 7).

$$r_{xy} = \frac{S_{xy}}{\sqrt{S_x S_y}} \qquad \text{Equation 7}$$

El coeficiente de Spearman sigue las mismas reglas que la correlación de Pearson, pero se aplica cuando las mediciones se clasifican por rangos (Eq. 8).

$$r_s = 1 - \frac{6\sum d^2}{N3 - N} \qquad \text{Equation 8}$$

Los valores de la correlación en ambos casos van de + 1 a - 1, pasando por el cero, el cual corresponde a ausencia de correlación, mientras que los valores de 1 y -1 indican la existencia de una correlación directa o indirectamente proporcional, respectivamente.

4. CAPÍTULO I

Análisis de la distribución de epítopos T CD8 en proteínas virales.

4.1 Justificación y Objetivos

El desarrollo de vacunas basadas en epítopos es costoso y difícil debido, en gran parte, al trabajo que supone la identificación de epítopos T. Por lo que la planificación de estrategias que ayuden al desarrollo de estas vacunas es fundamental. En este capítulo tratamos de definir un sistema de priorización de epítopos basándonos en su distribución en las secuencias proteicas del Virus de la Hepatitis C (HCV), el Virus de la Inmunodeficiencia Humana (HIV), y el Virus Influenza A (IAV).

Los objetivos propuestos son:

- Visualizar la localización de los epítopos T CD8 en las proteínas de HCV, HIV e IAV.
- Analizar si los epítopos T CD8 de HCV, HIV e IAV están distribuidos homogéneamente en las proteínas virales atendiendo a la longitud de éstas utilizando un test χ^2.
- Ver si la distribución del los epítopos refleja la localización de los sitios de unión a moléculas del MHC I predichos utilizando PSSMs (*Position Specific Scoring Matrices*).

4.2 Conclusiones

- La simple visualización de la distribución de los epítopos T CD8 no permite ver diferencias en la distribución, siendo necesario un análisis estadístico.
- El análisis de la distribución de los epítopos de HCV, HIV e IAV utilizando un test χ^2, revela que éstos no están distribuidos de acuerdo a la longitud de las proteínas virales, localizándose principalmente en las proteínas estructurales que forman la

cápside o la matriz de los virus, proteínas Core, Gag y M1 de HCV, HIV e IAV, respectivamente.

- Los sitios de unión a las moléculas MHC I A*0201, A*0301 y B*0702 están distribuidos, en general, homogéneamente atendiendo a la longitud de las proteínas. Sólo aquellos péptidos de HIV que se unen a moléculas A*0201 no están distribuidos homogéneamente, pero en ningún caso reflejan la distribución de los epítopos T CD8 específicos de estos virus.

Analysis of T cell epitope distribution in hepatitis C, human immunodeficiency and influenza A viruses

Carmen M. Diez-Rivero[1], Pedro A. Reche[1]†

[1]Laboratory of Immunomedicine,
Department of Microbiology I, Facultad de Medicina,
Universidad Complutense de Madrid,
Ave Complutense S/N, Madrid 28040, SPAIN.

† Corresponding author

E-mail addresses:

CMDR: cmdiezri@med.ucm.es

PAR: parecheg@med.ucm.es

Keywords: CD8 T cell epitope, HCV, IAV, HIV, epitope distribution, MHC I-binding peptides

Abstract

Background

Development of T cell epitope vaccines is handicapped by the cost and difficulty associated with T cell epitope identification. Therefore, there is need for defining strategies that can speed translational vaccine research. Here, we tried to define a system for prioritizing protein antigens for vaccine design by investigating epitope distribution patterns.

Methods

We used χ^2 -statistics to analyze whether known CD8 T cell epitopes of Hepatitis C Virus, Human Immunodeficiency Virus-1 and Influenza A virus are distributed in the viral proteomes according to the size/length of the source proteins. We also analyzed the distribution of peptides predicted to bind to several human MHC I molecules using χ^2 -statistics. Finally, we investigated the correlation between epitope distribution and sequence conservation.

Results

We found that epitopes are not distributed homogeneously by the size of the source proteins in any of the viruses. Specifically, structural proteins pack significantly more epitopes than those expected by their size. Moreover, we showed that such non-homogeneous distribution cannot be accounted by underlying MHC I-peptide binding preferences nor it is related to sequence conservation.

Conclusions

Our results provide support for focusing T cell epitope identification efforts on structure-building proteins.

Background

CD8 cytotoxic T cells play a key role in the defense against intracellular pathogens and tumor cells. CD8 T cell immune responses are driven by the recognition of foreign peptides that are presented by major histocompatibility complex class I (MHC I) molecules at the cell surface [1-3]. The identification of these peptides (CD8 T cell epitopes) is therefore important for understanding disease pathogenesis and etiology as well as for vaccine design.

Purely experimental identification of T cell epitopes is costly and time consuming: it requires the synthesis of overlapping peptides spanning the entire length of the protein, followed by complicated *in vitro* cellular assays on each synthesized peptide [4]. Therefore, we, and others, have developed computational approaches to predict T cell epitopes that reduce the experimental load involved in epitope identification. The main basis for anticipating CD8 T cell epitopes is the prediction of MHC I-binding peptides [5]. This approach can also be combined with methods that model other relevant step of the MHC class I antigen processing pathway, such as cleavage by the proteasome [6] and TAP mediated transport [7]. Such combination lead to a refinement of epitope predictions obtained by just considering MHC binding [8, 9]. However, epitope prediction tools are yet far from perfect and generally only 10% of the predicted epitopes are immunogenic (able to elicit a T-cell response) [10, 11]. Therefore, in order to accelerate epitope identification and translational vaccine research, we must improve epitope prediction methods. Additionally, it is key to define rationals for prioritizing protein antigens for epitope prediction and vaccine design [12]. To that end, we analyzed the distribution of known CD8 T cell epitopes.

We focused on three viruses of great clinical relevance: Hepatitis C Virus (HCV), Human Immunodeficiency Virus-1 (HIV) and Influenza A Virus (IAV). Briefly, HCV is a member of the flaviviridae family, which often produces a chronic infection that can lead to cirrhosis and hepatocellular carcinoma. It has a small RNA genome encoding a single polyprotein that is processed into 10 proteins [13], consisting of three structural proteins (core or nucleocapsid, E1 and E2) and seven nonstructural proteins (NS1, NS2, NS3, NS4a, NS4b, NS5a and NS5b). HIV-1 is a lentivirus that causes acquired

immunodeficiency syndrome (AIDS) [14, 15]. HIV is composed of two copies of single-stranded RNA, encompassing 9 gene products (Gag, Pol, Vif, Vpr, Tat, Rev, Vpu, Env and Nef), which, after processing, produce one of more viral proteins. For example, p17 (MA, matrix protein), p24 (CA, capsid protein), p7 (nucleocapsid protein) and p6 are all produced after the Gag polyprotein. Finally, IAV is a member of the Orthomyxoviridae family [16] with eight single (non-paired) RNA strands encoding of a total of eleven proteins (PB2 , PB1, PB1-F2, PA, HA, NP, NA, M1, M2, NS1 and NS2) [17, 18]. Each RNA encodes one or more protein products. For example, the RNA segment 7 encodes M1, the matrix protein that forms the viral envelope, and M2, an integral membrane protein. Using reference strains of these three viruses, we mapped and analyzed the location of the HCV-, HIV- and IAV–specific CD8 T cell epitopes onto the viral proteomes, concluding that CD8 T cell epitopes are not evenly distributed. Notoriously, we found that structural proteins Core (HCV), Gag (HIV) and M1 (IAV) pack significantly more peptides than those expected by their size. Here, we will interpret and discuss the significance of these results.

Methods

CD8 T cell epitopes

We used three datasets of CD8 T cell epitopes specific of HIV, HCV and IAV, encompassing 190, 249 and 78 epitopes, respectively. The datasets consisted of unique peptides of 9 or 10 residues that were collected from EPIMHC [19], Immuneepitope [20] and Los Alamos HIV databases (www.hiv.lanl.gov/). We only selected epitopes that were reported to be restricted by human MHC I molecules and able to elicit immune responses in the course of a natural infection. The corresponding author will provide these datasets upon written request.

Reference sequences and epitope mapping

We applied a fuzzy pattern-matching algorithm based on the *String::Aprox* - Perl extension, allowing a maximum of 3 substitutions for mapping CD8 T epitopes in representative reference amino acid sequences of the viral proteins of HCV, HIV, IAV. Reference sequences were isolated from the genomic sequences specified by the following GenBank accession number: NC_009827.1 for HCV (genotype 6), NC_001802.1 for HIV-1 strain HXB2. For IAV, we used the sequences given by the accessions NC_002016 to NC_002023, specific for the 8-genomic segments of the /Puerto Rico/8/1934(H1N1) strain.

Protein conservation factor

We computed a protein conservation factor (*CF*) for each of the proteins encoded by HCV, HIV and IAV using equation 1:

$$CF = \frac{N_c}{N_t} \qquad \text{Eq. 1}$$

where N_c is the number of non-variable residues and N_t the total number of amino acids of the protein. *CF* ranges between 0 and 1, taking the value of 1 when the protein has no variable residues. Non-variable residues were defined from the relevant multiple protein sequence alignments as those with a Shannon entropy (H) ≤ 1 [21]. Shannon entropy per site was computed as indicated in previous works [22, 23] using equation 2.

$$H = -\sum_{i=1}^{i=20} p_i \log_2 p_i \qquad \text{Eq. 2}$$

where *Pi* is the fraction of residues of amino acid type *i*. *H* ranges from 0 (total conservation, only one amino-acid type is present at that position) to 4.322 (all 20 amino acids are equally represented in that position).

Multiple sequence alignments (MSAs) required for computing sequence variability were obtained as follows. For the IAV, we used the reference genome NC_002016 - NC_002023 and BLAST each of the encoded proteins against a BLAST database built upon all IAV proteins (Taxonomy id: 11320). Subsequently, we realigned the sequences resulting of the BLAST searches using TCOFEE [24]. For HCV and HIV-1, we retrieved the relevant alignments from Los Alamos database and realigned them using TCOFEE.

Statistical analyses

We use χ^2 goodness of fit test to assess whether the distribution of the epitopes in the proteins of HCV, HIV and IAV was uniform –proportional to the size of the proteins– or not. The χ^2 statistics is given by equation 3.

$$\chi^2 = \sum_{i=1}^{k} \frac{\left(O_i - E_i\right)^2}{E_i} \qquad \text{Eq. 3}$$

where O_i is the number of observed epitopes in protein i, and E_i is the number of expected epitopes in the protein i calculated if they were homogeneously distributed according to the size of the proteins. The H_0 hypothesis (epitopes are homogeneously distributed) is rejected if the computed χ^2 statistics exceeds the χ^2 distribution value at $k - 1$ degrees of freedom and a given α value.

We used permutation tests to assess whether Spearman's rank correlation coefficients (R_s), obtained upon correlating protein sequence conservation and epitope distribution, were significantly different from zero.

Prediction of peptide-MHCI binding

We used position Specific Scoring Matrices (PSSMs) [5, 25-27], also known as profiles, to predict peptide binding to the human MHC I molecules HLA-A*0201, HLA-A*0301, HLA-B*0702. We only considered peptide binders of 9 residues in length (9mers). We applied PSSMs to the entire viral proteomes –upon combining all the viral proteins– and assessed the binding of each peptide to the relevant MHC I molecule by comparing its binding score to those of 10000 reference peptides (9-mers randomly obtained from SwissProt) obtained using the same PSSM. Specifically, a given peptide was considered to bind to the MHCI when its binding score was within the top 2% binding scores.

Results

Distribution of CD8 T cell epitopes

T cell epitopes are small peptide fragments obeying to rules for processing and MHC presentation that are not conceived to be highly specific. Hence, the bigger the protein the larger the number of epitopes that one can expect. Here, we used a χ^2 test to examine whether CD8 T cell epitopes specific of HCV, HIV and IAV follow a homogeneous protein-size wise distribution. We proceeded as follows. We first mapped the collected epitopes of HCV (190), VIH-1 (249) and IAV (78) onto their relevant proteins (Figure 1), tallying up the number of epitopes that falls within each viral protein (observed epitopes) (Table 1). Next, we distributed the total number of observed epitopes, into the viral proteins proportionally to their length/size, thus getting the number of expected epitopes (Table 1).

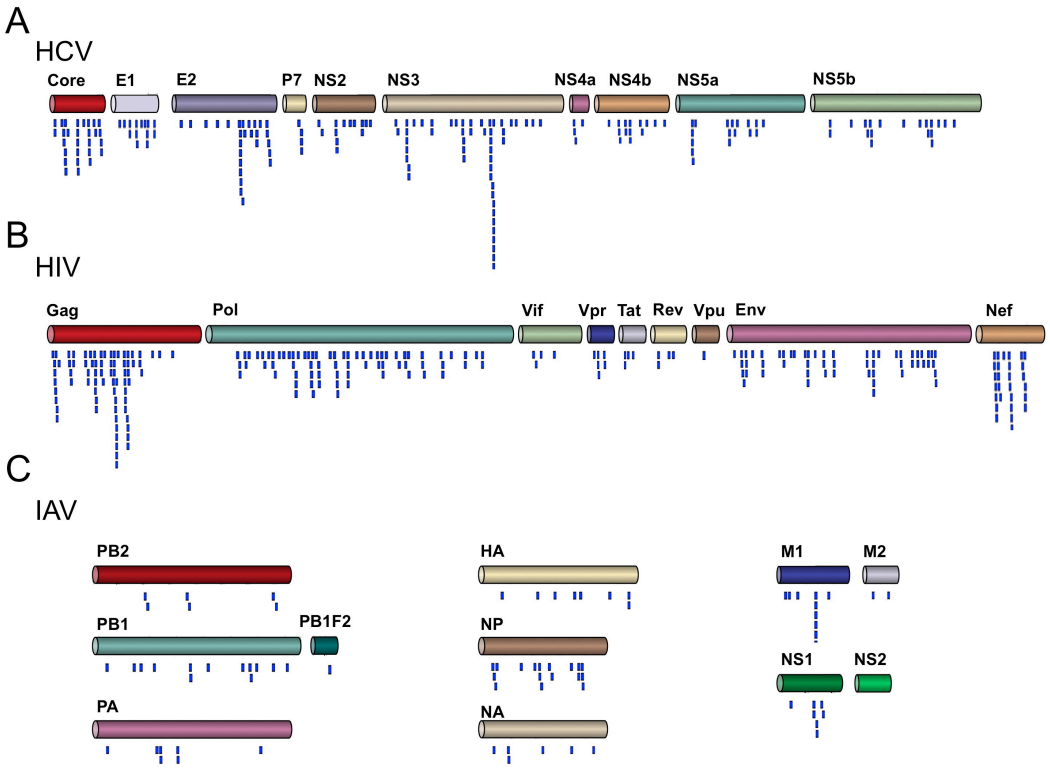


Figure 1. Epitope mapping. The figure shows the localization of CD8 T cell epitopes specific of three viruses: HCV (Panel A), HIV (Panel B) and IAV (Panel C). Epitopes are shown as blue segments underneath of the relevant proteins. IAV proteins that are encoded by the same RNA segment are shown in near proximity.

Table 1. Protein-size distribution analysis of CD8 T cell epitopes in HCV, HIV and IAV

HCV					
Protein	Protein length	*CF*	Observed epitopes	Expected epitopes	χ^2
Core	191	0,95	28	11.99	21.36
E1	192	0,58	14	12.12	0.29
E2	364	0,71	27	22.91	0.73
p7	64	0,68	4	3.98	0.001
NS2	218	0,59	13	13.70	0.04
NS3	632	0,89	50	39.83	2.59
NS4a	55	0,83	4	3.41	0.10
NS4b	262	0,84	14	16.47	0.37
NS5a	449	0,75	17	28.28	4.50
NS5b	592	0,81	19	37.31	8.98
Total	3019		190	190	38.97
HIV					
Protein	Protein length	*CF*	Observed epitopes	Expected epitopes	χ^2
Gag	500	0,68	75	39.73	31.32
Pol	1001	0,84	72	79.53	0.07
Vif	192	0,75	4	15.25	8.30
Vpr	96	0,74	6	7.63	0.35
Tat	86	0,63	4	6.75	1.12
Rev	116	0,57	4	9.22	2.95
Vpu	82	0,45	1	6.51	4.67
Env	856	0,54	55	68.01	2.49
Nef	206	0,62	28	16.37	8.27
Total	3135		249	249	60.19
IAV					
Protein	Protein length	*CF*	Observed epitopes	Expected epitopes	χ^2
PB2	759	0,98	6	13.05	3.81
PB1	757	0,1	13	13.01	0.00
PB1F2	87	0,84	1	1.49	0.16
PA	716	0,98	7	12.31	2.29
HA	566	0,88	8	9.73	0.31
NP	498	0,99	17	8.56	8.32
NA	452	0,92	6	7.81	0.42
M1	252	0,99	11	4.33	10.26
M2	97	0,89	2	1.67	0.07
NS1	230	0,83	7	3.95	2.35
NS2	121	0,92	0	2.08	2.08
Total	4537	10,22	78	78	30.06

The expected epitopes in a given protein are those resulting after distributing all the epitopes present in a virus proportionally to the length of that protein with regard to the total viral proteome.

The results of the χ^2 test showed that the distribution of the CD8 T cell epitopes is not homogeneous ($\alpha = 0.001$) in any of the viral proteomes studied here (HCV: $\chi^2 = 38.97 > \chi^2 H_o = 27.88$; HIV: $\chi^2 = 60.19 > \chi^2 H_o = 26.12$; IAV: $\chi^2 = 30.06 > \chi^2 H_o = 29.59$). To better visualize such uneven distribution, we represented the contribution of each protein, in percentage, to the χ^2 statistics (Figure 2A), and the ratio between observed and expected epitopes in each protein (Figure 2B). In Figure 2B, a ratio > 1 indicates more observed epitopes than expected, whereas a ratio < 1 indicate the opposite (less epitopes than expected). The most significant differences were found in non-enzymatic structural proteins of the viruses; their contribution to the χ^2 statistics is nearly enough to reject the null hypothesis (Figure 2A). These proteins carry more epitopes than the expected by their size. Thus, HCV Core protein encompasses 2.3-fold more epitopes than expected (Figure 2B) and Gag protein, which includes several non-enzymatic HIV-1 structural proteins, has 1.9-times more epitopes than expected (Figure 2B). Finally, the matrix M1 protein of IAV also encompasses 2.5-times more epitopes than the expected by their size (Figure 2B). Likewise, NP encompasses 2-times more epitopes than expected (Figure 2B).

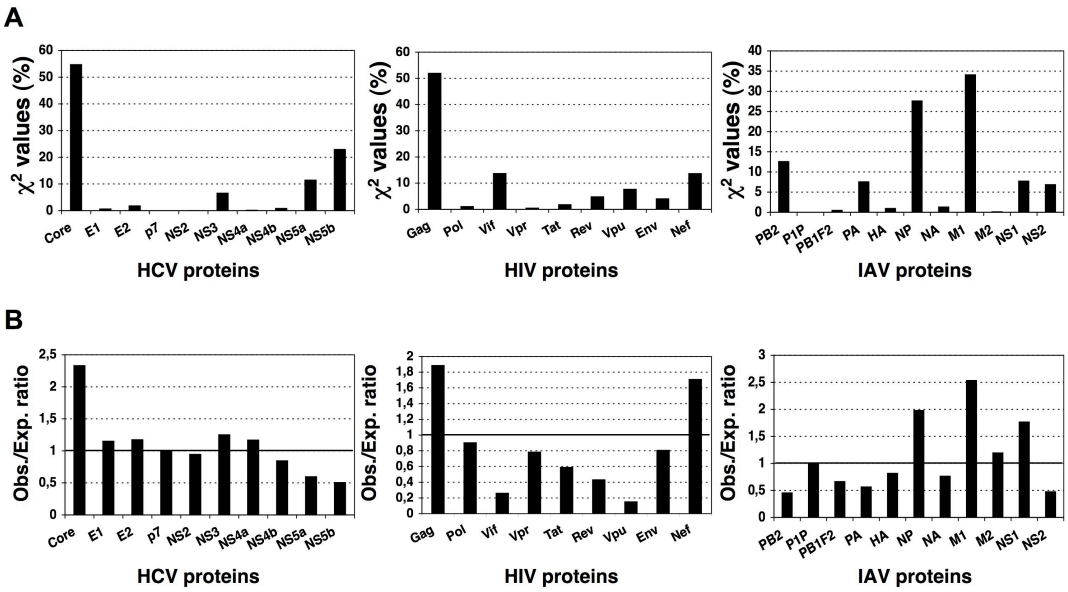


Figure 2. Protein-size distribution of virus-specific CD8 T cell epitopes. Using a χ^2 test, we concluded that CD8 T epitopes specific of HCV, HIV and IAV are not distributed homogeneously throughout their proteomes according to protein size. In this figure, we depicted for each virus, the contribution (in percentage) of the corresponding viral proteins to the χ^2 statistics (Panel **A**) and the ratio

between observed and expected epitopes in each of the proteins (Panel **B**). The ratio is relative to the expected epitopes; a value greater than 1 indicates more observed epitopes than expected, while a value lower than 1 reflects less epitopes than expected.

Other proteins also contributed significantly to the χ^2 statistic (Figure 2A). In HCV, NS5a and NS5b bear 1.6- and 1.9-times, respectively, less epitopes than expected (Figure 2B). In HIV, Vif and Rev encompass 3.8-times and 2.3-times less epitopes than expected (Figure 2B). An interesting case to comment is that of HIV-1 Vpu protein. As shown in Figure 2B, Vpu exhibits 6.5-times less epitopes than expected, the largest difference observed. Nonetheless, this difference does not have a major contribution to the χ^2 statistic (Figure 2A) as Vpu only bears a minor proportion of all HIV epitopes.

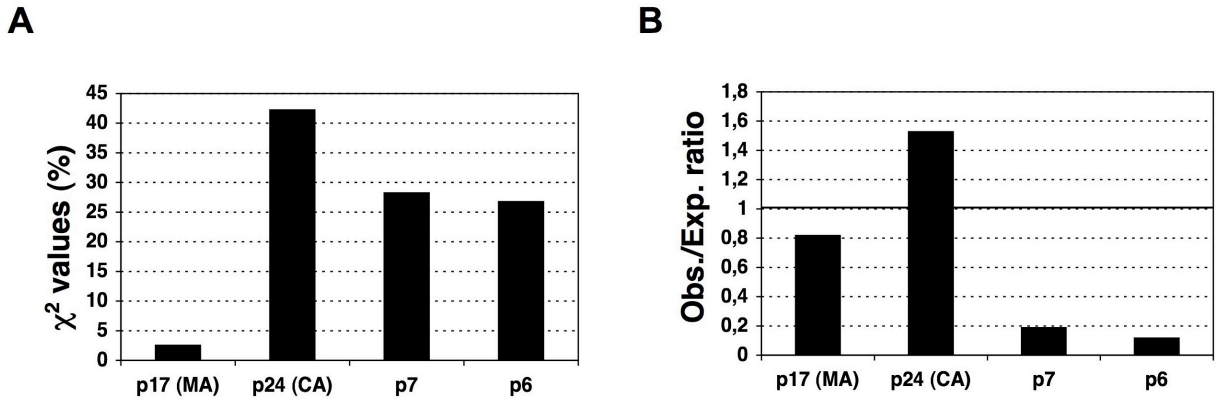


Figure 3. Protein-size distribution of Gag-specific CD8 T cell epitopes. Gag-specific CD8 T epitopes are not distributed homogeneously throughout its four integral proteins (p24, p17, p7 and p6) using χ^2 statistics. In panel **A**, we show the contribution of p17, p24, p7 and p6 to the Gag χ^2 statistics and in panel **B**, the ratio between observed and expected epitopes.

In HCV and IVA, the structure-building proteins Core and M1, respectively, that pack more epitopes than expected by their size are present in the mature viruses. In IAV, M1 is translated after one of the two alternative mRNAs that are produced after the M RNA segment 7 [18]. In HCV, Core is located at the beginning of a single translated open reading frame (ORF). In HIV-1, the Gag protein, in which we also found more epitopes than expected, is actually processed during maturation to produce

four different viral proteins: p17 (MA, matrix), p24 (CA, capsid), p7 and p6 (from the N-terminus to the C-terminus). Therefore, we also used the described χ^2 test to analyze the distribution of the 75 Gag-specific CD8 T-cell eptiopes within the relevant proteins. The results clearly show that Gag-specific epitopes are not distributed homogenously according to protein size/length (Gag: $\chi^2 = 20.31 > \chi^2 H_o = 16.27$) ($\alpha < 0.001$). The most relevant contributions to the χ^2 statistic are observed in protein p24 (CA) and p6 (Figure 3A). Protein p24 encompasses 1.5-times more epitopes than the expected while p6 bears 8.4-times less epitopes than expected (Figure 3B).

Distribution of MHC I binding sites

We wished to examine whether the noted non-homogeneous distribution of T cell epitopes in the viral proteomes mirrored underlying MHCI binding preferences. To that end, we targeted for peptide binding predictions three human MHCI molecules, HLA-A*0201, HLA-A*0301 and HLA-B*0702 (details in Methods). A*0201, A*0301, B*0702 belong to the A2, A3 and B7 HLA I supertypes, respectively. These HLA I supertypes are expressed in about 90% of population and have peptide binding repertoires that are not overlapping [28]. Then, we used the χ^2 test, as described earlier, to analyze the distribution of the predicted binding peptides to A*0201, A*0301 and B*0702, both, individually to each MHC I and in combination.

Table 2. χ^2–statistics resulting of analyzing the protein-size distribution of MHC I-binding peptides in HCV, HIV and IAV

	χ^2		
MHC I-Binding peptides to:	HCV	HIV	IAV
A*0201	22.56*	27.59**	18.1
A*0301	16.96	4.45	20.12
B*0702	16.48	2.48	12.2
A*0201 + A*0301 + B*0702	11.2	13.4	11.15

Statistically significant deviations are indicated with "*" symbol, where "**" is significant with an α-value < 0.001, and "*" with an α-value < 0.01. MHC-I binding peptides were predicted as indicated in Methods.

Unlike CD8 T cell epitopes, we found that the predicted MHCI-binding peptides are distributed homogeneously with regard to the length of the proteins (Table 2). This result is the expected: the larger

the protein the larger the number of potential peptide binders to MHC I. In fact, at an α-value of 0.001 (the same used in the CD8 T cell epitope analysis), only A*0201 binding peptides in HIV are not distributed homogeneously with regard to protein size (χ^2 = 27.59; χ^2 H_o= 26.12). However, the distribution of HIV-specific A*0201 binding peptides does not match the epitope distribution. For instances, the major contribution to the non-homogeneous distribution of the A*0201-binding peptides lies in Vpu which encompasses 3.6-fold more binding peptides than expected (Figure 4B), whereas Vpu carries less epitopes than expected (Figure 2B). Moreover, the most important contribution to the non-homogenous distribution of the observed epitopes lies in Gag, in which the number of A*0201-binding peptides does not differ from the expected. At a more permissive α-value of 0.01, we find that peptides binding to A*0201 in HCV are neither distributed homogeneously (χ^2 = 22.56 > χ^2 H_o= 21.67). In this case, the most notorious influence to the statistic is seen in NS4a, in which the number of predicted A*0201-binding peptides exceed the number of expected binders (see additional file 1, supplementary Table S1B), again the opposite to that seen with the epitopes (Figure 2B). The combination of the peptides predicted to bind to A*0201, A*0301, B*0702 always followed a homogenous distribution proportional to the size of the source proteins.

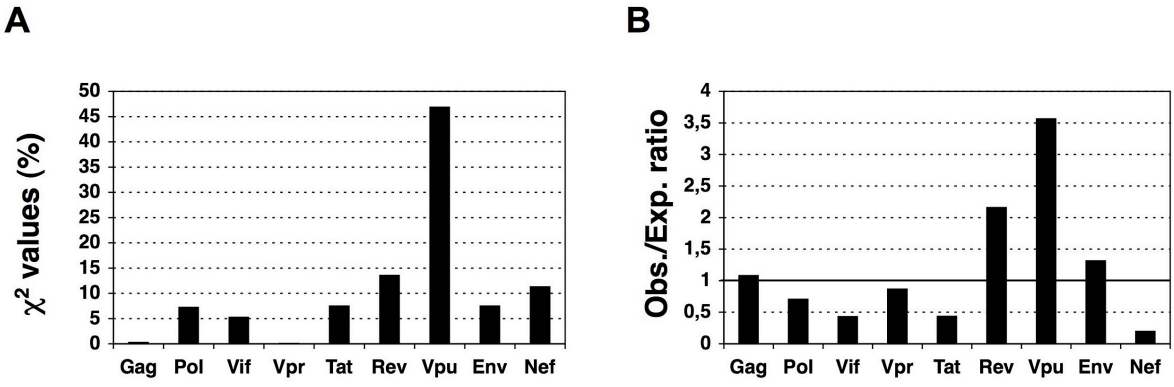


Figure 4. Distribution of predicted A*0201-binding peptides in HVI. We used χ^2 statistics to analyze the distribution of MHC I binding peptides from HCV, VIH and IVA. Only HIV peptides predicted to bind to A*0201 were not distributed homogeneously according to protein size at the same α value than that used in the epitope analysis (0.001). In panel **A,** we show the contribution (in percentage) to the χ^2 statistics of each HIV protein and in panel **B,** the ratio between observed and expected epitopes.

Epitope distribution and sequence conservation

Variable proteins likely bear multiple epitope variants that have not been identified. As result, the epitope distribution that we can obtain using a set of known CD8 T cell epitopes may be conditioned by protein sequence variability. Therefore, we examined the correlation between sequence conservation and epitope distribution. To that end, we computed a protein conservation factor (CF)(details in Methods) for each of the viral proteins and studied their correlation with the corresponding ratio between observed and expected epitopes, using Spearman's rank correlation (R_s)(Figure 5). The largest correlation was found in HCV (R_s = 0.345), followed by HIV (R_s = 0.333) and IAV (R_s = 0.127). However, all of the correlation values were very small and in fact none of the then was statistically different from zero.

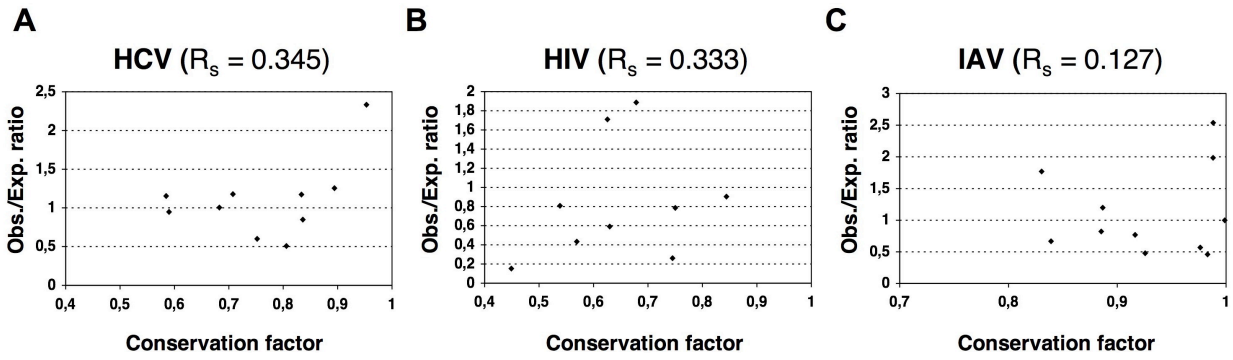


Figure 5. Correlation between epitope distribution and sequence conservation. Graphs depicting the ratio between observed and expected epitopes (Y-axis) in the proteomes of HCV (panel **A**), HIV (panel **B**) and IAV (panel **C**) plotted against the corresponding conservation factors (*CF*)(X-axis).

Discussion

Distinguishing T cell epitope distribution patterns is relevant for epitope-vaccine design. However, to the best of our knowledge, there is little or no evidence on whether T cell epitopes are distributed in any preferential manner onto pathogens' proteomes. Therefore, we investigated this matter in three human viruses, HCV, HIV and IAV, encompassing the largest known collections of CD8 T cell epitopes. Mapping of CD8 T cell epitopes onto the relevant viral proteomes did not reveal any obvious pattern and, in general, the larger the proteins the more epitopes they carry (Figure 1). However, using a χ^2 test we found that CD8 T cell epitopes are not distributed homogeneously according to the size of the proteins. Specifically, structural proteins assembling the viral capsid such as Core in HCV and Gag p24 in HIV display more epitopes than the expected for their size (Figure 2 and Figure 3). Likewise, matrix proteins including M1 of IAV also bear more epitopes than expected (Figure 2). At the other end, there are viral proteins such as NS5a and NS5b in HCV, Vif and Vpu in HIV and PB2 in IAV that display less epitopes than the expected by their size (Figure 2). T cell epitopes consist of peptides that need to bind and be presented by MHC I molecules prior to T cell recognition. However, in contrast to the analyzed epitopes, we found that MHC I-binding peptides are largely distributed proportionally to the size of the source of viral proteins (Table 2). Therefore, the observed epitope distribution does not appear to obey to any underlying MHCI binding preferences.

Another factor that can shape epitope distribution patterns is sequence variability. Experimental verification of epitopes (as those used here) requires determining T cell responses against synthetic peptides and responses elicited against variant epitopes will pass undetected [29]. Therefore, there could be a bias in known CD8 T cell epitopes towards conservation that could lead to observe less epitopes than expected in variable proteins and more than expected in conserved proteins. However, we did not find any significant correlation between the epitope

distributions described here and sequence conservation (Figure 5). Therefore, sequence conservation/variability does not explain the noted epitope distribution.

Although we cannot discard that our results might reflect bias of researchers towards studying specific viral proteins, we also find other possible explanations for the noted epitope distribution. The viral proteins Core in HCV, Gag p24 in HIV that encompass more epitopes are always located near or at the beginning of translated open reading frames (ORF) often encompassing other proteins. Conversely, those located at the end of translated ORF bear less epitopes than expected (*e.g.* p6 from HIV Gag). The extreme paradigm is HCV, whose entire genome is translated into a single polyprotein, in which the structural protein Core is located at the N-terminus and N5Sb at the C-terminus. It has been shown that protein translation often results in incomplete protein products [30-32]. Consequently, that we find more peptides in proteins located near the beginning of translated ORFs is perhaps pointing to the fact that they get translated predominantly and MHC I antigen presentation is linked to protein biosynthesis [30, 31, 33].

Placing the structural proteins at the beginning of translated ORF is likely a strategy used by viruses to guarantee the expression of proteins in high copy numbers. To our knowledge, this simple position-based translational control of protein expression has not been described before and it will require experimental confirmation. A similar mechanism but acting at the transcriptional level has been described in negative-strand RNA viruses. In these viruses, levels of gene expression are primarily regulated by the position of each gene relative to the single promoter and also by cis-acting sequences located at the beginning and end of each gene and at the intergenic junctions [34].

In the case of HIV and HCV, the epitope distribution might appear somewhat paradoxical. After all, we see plenty of epitopes, more than expected, in structural proteins that are made in high copy numbers and can be quite conserved, and yet the immune system is not always capable of clearing up these viruses: they linger causing chronic infections [35, 36].

There is not a simple explanation to this observation. First, it is important to highlight that the CD8 T cell responses, although essential in containing viral infections, do not work alone and may not be sufficient to clear these viruses. On the other hand, the number of epitopes does not say anything about their immunogenicity and T cells might target immunodominant epitopes that are variable [37]. In any case, since immunodominance can be reverted through vaccination [38, 39] one should not underestimate the relevance of subdominant epitopes for vaccine design

Conclusions

CD8 T cell epitopes are preferentially located in viral structural proteins, which, incidentally, are often conserved and get transcribed and/or translated in first place to guarantee the high copy numbers required to ensemble the virus. Altogether, these results support that structure building protein antigens ought to be prioritized for T cell epitope prediction/identification. Experimental identification of CD8 T cell epitopes requires complicated and costly *in vitro* cellular assays and such prioritization ought to save time and resources and speed translational vaccine research.

Competing interests

The authors have no competing interests

Author's contributions

CMDR did the work, prepared figures and helped writing the manuscript. PAR designed the work, interpreted the results and wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We wish to tank Dr. Esther M. Lafuente for helpful comments. This work was supported by Grant SAF2009-08103 to PAR from the Ministerio de Ciencia e Innovación of Spain.

References

1. Garcia KC, Degano M, Pease L, Huang M, Peterson PA, Teyton L, Wilson IA: **Structural basis of plasticity in T cell receptor recognition of a self peptide-MHC antigen**. *Science* 1998, **279**:1166-1172.
2. Margulies DH: **Interactions of TCRs with MHC-peptide complexes: a quantitative basis for mechanistic models**. *Curr Opin Immunol* 1997, **9**(3):390-395.
3. Wang J-H, Reinherz E: **Structural basis of T cell recognition of peptides bound to MHC molecules**. *Molecular Immunology* 2001, **38**:1039-1049.
4. Draenert R, Altfeld M, Brander C, Basgoz N, Corcoran C, Wurcel AG, Stone DR, Kalams SA, Trocha A, Addo MM *et al*: **Comparison of overlapping peptide sets for detection of antiviral CD8 and CD4 T cell responses**. *J Immunol Methods* 2003, **275**(1-2):19-29.
5. Lafuente EM, Reche PA: **Prediction of MHC-peptide binding: a systematic and comprehensive overview**. *Curr Pharm Des* 2009, **15**(28):3209-3220.
6. Diez-Rivero CM, Lafuente EM, Reche PA: **Computational analysis and modeling of cleavage by the immunoproteasome and the constitutive proteasome**. *BMC Bioinformatics* 2010, **11**:479.
7. Diez-Rivero CM, Chenlo B, Zuluaga P, Reche PA: **Quantitative modeling of peptide binding to TAP using support vector machine**. *Proteins* 2010, **78**(1):63-72.
8. Donnes P, Kohlbacher O: **Integrated modeling of the major events in the MHC class I antigen processing pathway**. *Protein Sci* 2005, **14**(8):2132-2140.
9. Tenzer S, Peters B, Bulik S, Schoor O, Lemmel C, Schatz MM, Kloetzel PM, Rammensee HG, Schild H, Holzhutter HG: **Modeling the MHC class I pathway by combining predictions of proteasomal cleavage, TAP transport and MHC class I binding**. *Cell Mol Life Sci* 2005, **62**(9):1025-1037.
10. Wang M, Lamberth K, Harndahl M, Roder G, Stryhn A, Larsen MV, Nielsen M, Lundegaard C, Tang ST, Dziegiel MH *et al*: **CTL epitopes for influenza A including the H5N1 bird flu; genome-, pathogen-, and HLA-wide screening**. *Vaccine* 2007, **25**(15):2823-2831.
11. Zhong W, Reche PA, Lai CC, Reinhold B, Reinherz EL: **Genome-wide characterization of a viral cytotoxic T lymphocyte epitope repertoire**. *J Biol Chem* 2003, **278**(46):45135-45144.
12. Flower DR, Macdonald IK, Ramakrishnan K, Davies MN, Doytchinova IA: **Computer aided selection of candidate vaccine antigens**. *Immunome Res* 2010, **6 Suppl 2**(6):S1.
13. Moradpour D, Penin F, Rice CM: **Replication of hepatitis C virus**. *Nat Rev Microbiol* 2007, **5**(6):453-463.
14. Weiss RA: **How does HIV cause AIDS?** *Science* 1993, **260**(5112):1273-1279.
15. Douek DC, Roederer M, Koup RA: **Emerging concepts in the immunopathogenesis of AIDS**. *Annu Rev Med* 2009, **60**:471-484.
16. Tscherne DM, Garcia-Sastre A: **Virulence determinants of pandemic influenza viruses**. *J Clin Invest* 2011, **121**(1):6-13.
17. Kaverin NV: **[Genome of influenza virus: organization, function, evolution]**. *Mol Biol (Mosk)* 1980, **14**(2):245-260.
18. Cheung TK, Poon LL: **Biology of influenza a virus**. *Ann N Y Acad Sci* 2007, **1102**:1-25.
19. Reche PA, Zhang H, Glutting JP, Reinherz EL: **EPIMHC: a curated database of MHC-binding peptides for customized computational vaccinology**. *Bioinformatics* 2005, **21**(9):2140-2141. Epub 2005 Jan 2118.

20. Peters B, Sidney J, Bourne P, Bui H, Buus S, Doh G, Fleri W, Kronenberg M, Kubo R, Lund O *et al*: **The immune epitope database and analysis resource: from vision to blueprint.** *PLoS Biol* 2005, **3**(3):e91.
21. Shannon CE: **The mathematical theory of communication**. *The Bell System Technical Journal* 1948, **27**:379-423, 623-656.
22. Reche PA, Reinherz EL: **Sequence variability analysis of human class I and class II MHC molecules: functional and structural correlates of amino acid polymorphisms**. *J Mol Biol* 2003, **331**(3):623-641.
23. Reche PA, Keskin DB, Hussey RE, Ancuta P, Gabuzda D, Reinherz EL: **Elicitation from virus-naive individuals of cytotoxic T lymphocytes directed against conserved HIV-1 epitopes**. *Med Immunol* 2006, **5**:1.
24. Notredame C, Higgins DG, Heringa J: **T-Coffee: A novel method for fast and accurate multiple sequence alignment**. *J Mol Biol* 2000, **302**(1):205-217.
25. Reche PA, Reinherz EL: **Prediction of peptide-MHC binding using profiles.** *Methods Mol Biol* 2007, **409**:185-200.
26. Reche PA, Glutting JP, Reinherz EL: **Prediction of MHC class I binding peptides using profile motifs**. *Hum Immunol* 2002, **63**(9):701-709.
27. Reche PA, Glutting J-P, Reinherz EL: **Enhancement to the RANKPEP resource for the prediction of peptide binding to MHC molecules using profiles.** *Immunogenetics* 2004, **56**:405-419.
28. Reche PA, Reinherz EL: **PEPVAC: a web server for multi-epitope vaccine development based on the prediction of supertypic MHC ligands**. *Nucleic Acids Res* 2005, **33**(Web Server issue):W138-142.
29. Chang CX, Dai L, Tan ZW, Choo JA, Bertoletti A, Grotenbreg GM: **Sources of diversity in T cell epitope discovery**. *Front Biosci* 2011, **17**:3014-3035.
30. Princiotta MF, Finzi D, Qian SB, Gibbs J, Schuchmann S, Buttgereit F, Bennink JR, Yewdell JW: **Quantitating protein synthesis, degradation, and endogenous antigen processing**. *Immunity* 2003, **18**(3):343-354.
31. Schubert U, Anton LC, Gibbs J, Norbury CC, Yewdell JW, Bennink JR: **Rapid degradation of a large fraction of newly synthesized proteins by proteasomes**. *Nature* 2000, **404**(6779):770-774.
32. Yewdell JW, Anton LC, Bennink JR: **Defective ribosomal products (DRiPs): a major source of antigenic peptides for MHC class I molecules?** *J Immunol* 1996, **157**(5):1823-1826.
33. Reits E, A. J., Vos J, C., Grommé M, Neefjes J: **The major substrates for TAP in vivo are derived from newly synthesized proteins.** *Nature* 2000(404):774-778.
34. Villarreal LP, Breindl M, Holland JJ: **Determination of molar ratios of vesicular stomatitis virus induced RNA species in BHK21 cells**. *Biochemistry* 1976, **15**(8):1663-1667.
35. Bowen DG, Walker CM: **Adaptive immune responses in acute and chronic hepatitis C virus infection**. *Nature* 2005, **436**(7053):946-952.
36. Sagar M: **HIV-1 transmission biology: selection and characteristics of infecting viruses**. *J Infect Dis* 2010, **202 Suppl 2**(202):S289-296.
37. Yewdell JW: **Confronting complexity: real-world immunodominance in antiviral CD8+ T cell responses**. *Immunity* 2006, **25**(4):533-543.
38. Sandberg JK, Grufman P, Wolpert EZ, Franksson L, Chambers BJ, Karre K: **Superdominance among immunodominant H-2Kb-restricted epitopes and reversal by dendritic cell-mediated antigen delivery**. *J Immunol* 1998, **160**(7):3163-3169.

39. Eberl G, Kessler B, Eberl LP, Brunda MJ, Valmori D, Corradin G: **Immunodominance of cytotoxic T lymphocyte epitopes co-injected in vivo and modulation by interleukin-12**. *Eur J Immunol* 1996, **26**(11):2709-2716.

Additional file 1.

Table IS.A. Protein-size distribution analysis of MHC I-binding peptides from HCV

A*0201-binding peptides				
Protein	Protein length	N. predicted peptides	N. expected peptides	χ^2
Core	191	4	5.24	0.29
E1	192	6	5.29	0.09
E2	364	10	10.01	0
p7	64	3	1.74	0.92
NS2	218	11	5.98	4.2
NS3	632	10	13.4	3.14
NS4a	55	5	1.49	8.27
NS4b	262	12	7.78	3.2
NS5a	449	7	12.35	2.32
NS5b	592	15	16.29	0.1
Total	3019	83	83	22.56
A*0301-binding peptides				
Protein	Protein length	N. predicted peptides	N. expected peptides	χ^2
Core	191	1	2.71	1.08
E1	192	4	2.74	0.58
E2	364	0	5.18	5.19
p7	64	0	0.9	0.9
NS2	218	3	3.1	0.003
NS3	632	7	9.01	0.45
NS4a	55	0	0.77	0.77
NS4b	262	7	3.73	2.87
NS5a	449	6	6.4	0.02
NS5b	592	15	8.44	5.06
Total	3019	43	43	16.96
B*0702-binding peptides				
Protein	Protein length	Predicted peptides	Expected peptides	χ^2
Core	191	6	5.74	0.01
E1	192	4	5.8	0.56
E2	364	7	10.97	1.44
p7	64	3	1.91	0.63
NS2	218	1	6.56	4.71
NS3	632	21	19.08	0.19
NS4a	55	1	1.63	0.24
NS4b	262	9	7.89	0.15
NS5a	449	24	13.54	8.07
NS5b	592	15	17.87	0.46
Total	3019	91	91	16.48
Combination: A*0201-, A*0301- and B*0702-binding peptides				
Protein	Protein length	Predicted peptides	Expected peptides	χ^2

Core	191	11	13.51	0.46
E1	192	14	13.65	0.01
E2	364	17	25.81	3.01
p7	64	6	4.48	0.52
NS2	218	15	15.43	0.01
NS3	632	37	44.86	1.38
NS4a	55	6	3.84	1.22
NS4b	262	27	18.55	3.84
NS5a	449	36	31.85	0.54
NS5b	592	45	42.02	0.21
Total	3019	214	241	11.19

Table 1S.B. Protein-size distribution analysis of MHC I-binding peptides from HIV

A*0201-binding peptides

Protein	Protein length	Predicted peptides	Expected peptides	χ^2
Gag	500	13	11.96	0.09
Pol	1001	17	23.94	2.01
Vif	192	2	4.59	1.46
Vpr	96	2	2.29	0.04
Tat	86	0	2.08	2.08
Rev	116	6	2.77	3.75
Vpu	82	7	1.96	12.95
Env	856	27	20.47	2.08
Nef	206	1	4.93	3.13
Total	3135	75	75	27.59

A*0301-binding peptides

Protein	Protein length	Predicted peptides	Expected peptides	χ^2
Gag	500	5	5.58	0.06
Pol	1001	12	11.17	0.06
Vif	192	3	2.14	0.34
Vpr	96	0	1.07	1.07
Tat	86	1	0.97	0.001
Rev	116	2	1.29	0.38
Vpu	82	1	0.91	0.01
Env	856	11	9.55	0.22
Nef	206	0	2.29	2.3
Total	3135	35	35	4.45

B*0702-binding peptides

Protein	Protein length	Predicted peptides	Expected peptides	χ^2
Gag	500	12	11	0.09
Pol	1001	22	22.02	0
Vif	192	6	4.22	0.75
Vpr	96	1	2.11	0.58
Tat	86	2	1.91	0.003

Rev	116	1	2.55	0.94
Vpu	82	2	1.8	0.02
Env	856	18	18.83	0.04
Nef	206	5	4.53	0.04
Total	3135	69	69	2.48

Combination: A*0201-, A*0301- and B*0702-binding peptides

Protein	Protein length	Predicted peptides	Expected peptides	χ^2
Gag	500	30	28.54	0.07
Pol	1001	51	57.14	0.66
Vif	192	11	10.96	0
Vpr	96	3	5.48	1.12
Tat	86	3	4.96	0.78
Rev	116	9	6.21	0.85
Vpu	82	10	4.68	6.04
Env	856	56	48.86	1.04
Nef	206	6	11.76	2.82
Total	3135	179	179	13.39

Table 1S.C. Protein-size distribution analysis of MHC I-binding peptides from IAV

A*0201-binding peptides

Protein	Protein length	Predicted peptides	Expected peptides	χ^2
PB2	759	14	19.4	1.5
PB1	87	0	2.22	2.22
PB1F2	757	20	19.35	0.02
PA	716	19	18.3	0.03
HA	566	17	14.46	0.44
NP	498	8	12.73	1.75
NA	454	8	11.6	1.12
M1	252	14	6.49	8.65
M2	97	4	2.48	0.93
NS1	230	7	5.88	0.21
NS2	121	5	3.09	1.18
Total	4537	116	116	18.1

A*0301-binding peptides

Protein	Protein length	Predicted peptides	Expected peptides	χ^2
PB2	759	18	14.38	0.91
PB1	87	5	1.65	6.81
PB1F2	757	19	14.34	1.51
PA	716	16	13.56	0.43
HA	566	9	10.72	0.28
NP	498	6	9.43	1.25
NA	454	1	8.6	6.72
M1	252	6	4.81	0.29
M2	97	0	1.84	1.83

NS1	230	4	4.36	0.03
NS2	121	2	2.29	0.04
Total	4537	86	86	20.12

B*0702-binding peptides

Protein	Protein length	Predicted peptides	Expected peptides	χ^2
PB2	759	23	14.54	4.91
PB1	87	1	1.67	0.27
PB1F2	757	16	14.51	0.15
PA	716	9	13.72	1.62
HA	566	8	10.85	0.75
NP	498	7	9.54	0.68
NA	454	12	8.7	1.25
M1	252	4	4.87	0.15
M2	97	2	1.86	0.01
NS1	230	5	4.41	0.08
NS2	121	0	2.23	2.32
Total	4537	87	87	12.19

Combination: A*0201-, A*0301-, B*0702-binding peptides

Protein	Protein length	Predicted peptides	Expected peptides	χ^2
PB2	759	55	47.99	1.02
PB1	87	6	5.5	0.04
PB1F2	757	55	47.86	1.06
PA	716	44	45.27	0.0
HA	566	33	35.79	0.21
NP	498	21	31.49	3.49
NA	454	21	28.71	2.07
M1	252	23	16.06	2.99
M2	97	6	6.13	0.003
NS1	230	16	14.54	0.15
NS2	121	7	7.65	0.05
Total	4537	287	287	11.15

5. CAPÍTULO II

Modelado computacional de la especificidad de corte del proteasoma constitutivo y del inmunoproteasoma

5.1 Justificación y Objetivos

El proteasoma es el responsable de la degradación de las proteínas en el citosol, generando el extremo C-terminal de los péptidos presentados por las moléculas del MHC I (residuo P1). Por lo tanto, la actividad catalítica del proteasoma es uno de los principales pasos del procesamiento de antígenos. Existen dos formas activas del proteasoma, el constitutivo y el inmunoproteasoma, y por tanto es importante ver qué sitios de corte tienen en común para identificar aquellos epítopos que sean protectivos. Por ello, aquí se pretende desarrollar dos modelos para la predicción de los sitios de corte, uno para el proteasoma constitutivo y otro para el inmunoproteasoma.

Los objetivos son:

- Usando *N-grams*, desarrollamos varios modelos distintos del proteasoma constututivo, entrenando con péptidos eluidos de moléculas del MHC I, y del inmunoproteasoma, a partir de un conjunto de epítopos T CD8 que son capaces de activar la respuesta inmunitaria y encontrados en humanos durante el transcurso de la infección. Los modelos se entrenan con fragmentos de péptidos de distintos tamaños, todos ellos incluyendo el mismo número de aminoácidos a cada lado del punto de corte (residuo P1).
- Comparación de los modelos aquí desarrollados con otros modelos existentes.
- Combinación de los modelos predictivos del proteasoma y/o del inmunoproteasoma con la predicción unión de péptidos a las moléculas del MHC I.
- Desarrollo de una herramienta web.

5.2 Conclusiones

- Se han podido desarrollar modelos del proteasoma constitutivo y del inmunoproteasoma que, al ser evaluados mediante un test independiente, obtienen valores de MCC de 0.19 y 0.20, respectivamente. Estos modelos son tan buenos o incluso algo mejores que los desarrollados hasta la fecha (NetChop, MCC = 0.18).
- Tanto en el caso del proteasoma constitutivo como en el del inmunoproteasoma la capacidad predictiva de los modelos mejora al aumentar el tamaño de los fragmentos con los que se entrena el modelo, siendo los mejores modelos, a juzgar por su MCC, aquellos que se entrenaron con fragmentos de 12 residuos, 6 a cada lado del punto de corte.
- Los modelos desarrollados para el proteasoma constitutivo y el inmunoproteasoma, a reproducen los datos experimentales que reflejan patrones de corte distintos aunque solapantes.
- Los modelos del proteasoma parecen ser mejores que los del inmunoproteasoma cuando se evalúan mediante validación cruzada, pero al realizar un test independiente con epítopos T CD8 específicos de HIV, los modelos del inmunoproteasoma presentan mejores resultados.
- La combinación de la predicción de péptidos que se unen a moléculas del MHC I y de la predicción de sitios de corte, empleando los modelos aquí desarrollados de corte por el inmunoproteasoma y el proteasoma constitutivo, mejora de manera significativa el descubrimiento de epítopos de células T CD8 restringidos por distintas moléculas del MHC I.
- Estos modelos aquí desarrollados están disponibles para su libre uso en la página http://imed.med.ucm.es/Tools/PCPS/.

BMC Bioinformatics

RESEARCH ARTICLE Open Access

Computational analysis and modeling of cleavage by the immunoproteasome and the constitutive proteasome

Carmen M Diez-Rivero[1,2], Esther M Lafuente[2], Pedro A Reche[1,2*]

Abstract

Background: Proteasomes play a central role in the major histocompatibility class I (MHCI) antigen processing pathway. They conduct the proteolytic degradation of proteins in the cytosol, generating the C-terminus of CD8 T cell epitopes and MHCI-peptide ligands (*P1* residue of cleavage site). There are two types of proteasomes, the constitutive form, expressed in most cell types, and the immunoproteasome, which is constitutively expressed in mature dendritic cells. Protective CD8 T cell epitopes are likely generated by the immunoproteasome and the constitutive proteasome, and here we have modeled and analyzed the cleavage by these two proteases.

Results: We have modeled the immunoproteasome and proteasome cleavage sites upon two non-overlapping sets of peptides consisting of 553 CD8 T cell epitopes, naturally processed and restricted by human MHCI molecules, and 382 peptides eluted from human MHCI molecules, respectively, using *N-grams*. Cleavage models were generated considering different epitope and MHCI-eluted fragment lengths and the same number of C-terminal flanking residues. Models were evaluated in 5-fold cross-validation. Judging by the Mathew's Correlation Coefficient (*MCC*), optimal cleavage models for the proteasome (*MCC* = 0.43 ± 0.07) and the immunoproteasome (*MCC* = 0.36 ± 0.06) were obtained from 12-residue peptide fragments. Using an independent dataset consisting of 137 HIV1-specific CD8 T cell epitopes, the immunoproteasome and proteasome cleavage models achieved *MCC* values of 0.30 and 0.18, respectively, comparatively better than those achieved by related methods. Using ROC analyses, we have also shown that, combined with MHCI-peptide binding predictions, cleavage predictions by the immunoproteasome and proteasome models significantly increase the discovery rate of CD8 T cell epitopes restricted by different MHCI molecules, including A*0201, A*0301, A*2402, B*0702, B*2705.

Conclusions: We have developed models that are specific to predict cleavage by the proteasome and the immunoproteasome. These models ought to be instrumental to identify protective CD8 T cell epitopes and are readily available for free public use at http://imed.med.ucm.es/Tools/PCPS/.

Background

CD8 cytotoxic T cells play a key role fighting intracellular pathogens, eliminating infected cells that display on their cell surface foreign peptides bound to major histocompatibility complex class I (MHCI) molecules [1-3]. CD8 T cell epitopes and, in general, peptides presented by MHCI molecules, derive from protein fragments produced in the cytosol by the proteolytic action of the proteasome [4,5]. Briefly, the proteasome generates protein fragments between 7 and 15 amino acids. Some of these peptides can be transported from the cytosol into the endoplasmic reticulum (ER) by the transporter associated with antigen processing (TAP), where they can be loaded onto nascent MHCI molecules. Interestingly, whereas different peptidases and proteases in the cytosol and the endoplasmic reticulum shape the N-terminus of the peptides presented by MHCI molecules [6], their C-terminus generally corresponds to the *P1* residue of the proteasome cleavage site [7,8].

The proteasome is a multisubunit ATP-dependent protease and it is primarily responsible for the degradation

* Correspondence: parecheg@med.ucm.es
[1]Laboratory of Immunomedicine, Department of Microbiology I-Immunology, Facultad de Medicina, Universidad Complutense de Madrid, Ave Complutense S/N, Madrid 28040, Spain
Full list of author information is available at the end of the article

of cytosolic proteins [9]. The most common form of the proteasome is known as the 26 S proteasome, which is composed by a catalytic core (20S) and two regulatory complexes (19S), located one at each side of the core [5]. The catalytic activity of the proteasome is located at the subunits β5 (X, LMP7), β2 (Z, MECL-1) and β1 (Y, LMP2) of the 20 S core, which cut after the C-terminus of hydrophobic (chymotrypsin-like activity), basic (trypsin-like activity) or acidic (caspase-like activity) amino acids, respectively [10]. Upon IFN-γ exposure, the three catalytic subunits of the constitutive 20 S core can be replaced by three new catalytic subunits: β5i (LMP2), β2i (MECL-1), and β1i (LMP2) [11]. This new form of proteasome is called immunoproteasome, as opposed to the constitutively expressed proteasome. The immunoproteasome is the constitutive form of proteasome presented in dendritic cells [12]. The immunoproteasome produces different but overlapping cleavage patterns with regard to those of the proteasome [13]; chiefly, the immunoproteasome does not cut after acidic residues [13,14]. Because the antigen-specific cytotoxic function of CD8 T cells is generally acquired upon the recognition of MHCI-bound peptide antigens displayed on the cell surface of dendritic cells (priming), it is likely that protective epitopes are those generated by the proteasome and the immunoproteasome [15].

Prediction of proteasome cleavage sites is relevant for CD8 T cell epitope identification and, subsequently, for the design of epitope-based vaccines eliciting CD8 T cell responses. Therefore, different methods to predict proteasome cleavage sites have been reported. Proteasome cleavage prediction methods were first developed using enolase and β-casein protein fragments generated *in vitro* by human constitutive proteasomes [16-18]. Likewise, a kinetic model of the proteasome proteolytic activity was also developed using peptide fragments from *in vitro* digestions [19,20]. Those models are specific for the constitutive 20 S proteasome that was used to generate the peptide fragments. Proteasome cleavages take place between the C-terminus of MHCI-restricted peptides (*P1* residue of cleavage site) and their most proximal C-terminal flanking residue (*P1'* residue of cleavage site). Therefore, proteasome cleavage prediction methods have also been developed using MHCI-restricted peptide ligands and their C-terminal flanking regions [21-23]. These latter methods appear to outcompete the former methods that were trained on actual proteolytic digestion data on the task of predicting cleavage sites defined by MHC I restricted peptides [24]. However, methods trained on experimental cleavage data can be more suitable for identifying protein fragments produced by the proteasome [18].

The problem of predicting proteasome cleavage sites resembles that of modeling grammatical rules. Therefore, in this manuscript, we have applied statistical language models [25] to analyze and model the cleavage sites of the constitutive proteasome and the immunoproteasome. Proteasome cleavage sites were obtained from MHCI-eluted peptides and their C-terminal flanking regions, whereas immunoproteasome cleavage sites were rendered from naturally processed CD8 T cell epitopes and their C-terminal flanking regions. In cross-validation, optimal proteasome and immunoproteasome cleavage models achieved an *MCC* of 0.43 ± 0.07 and 0.36 ± 0.06, respectively. These models were trained using 12-residue fragments, consisting of the C-terminal end of MHCI-restricted peptides (*P6 - P1* residues of cleavage site) followed by the 6 most-proximal C-terminal flanking residues (*P1' - P6'* residues of cleavage site). The fact that optimal models were trained using peptide fragments consisting of 6 amino acids at each side of the cleavage site is consistent with the activity exhibited by the proteasome [26]. Here, we have also shown that combining cleavage predictions by the constitutive and the immunoproteasome with MHCI-binding predictions serve to improve the prediction rate of CD8 T cell epitopes. Cleavage predictions using our models are available at http://imed.med.ucm.es/Tools/PCPS/.

Methods

Datasets and sequences

We assembled three non-overlapping datasets consisting of distinct MHCI-restricted peptides and their protein sources. The peptide content in these datasets was as follows. The first dataset encompassed 553 CD8 T cell epitopes from different sources but from Human Immunodeficiency Virus (HIV1) and were all restricted by human MHCI molecules. Immune responses against these epitopes have been verified experimentally using T cells from infected humans. Because CD8 T cell immune responses against these epitopes are elicited in the course of an infection, we assume that they are naturally processed. The second dataset included 382 peptides that were eluted from human MHCI molecules, and the third dataset encompassed 137 HIV1-specific CD8 T cell epitopes restricted by human MHCI molecules and naturally processed. MHCI-restricted peptides in these datasets were collected from the EPIMHC [27], Immuneepitope [28] and Los Alamos databases [29], and consisted of unique nonapeptides (9-mers) that were subjected to a sequence similarity reduction schema using the *purge* utility implemented in the Gibbs Sampler [30]. As a result, peptides in these three datasets do not share more than 4 identical residues (global sequence similarity in the first, second and third datasets is 3.1 ± 11.7, 3.9 ± 12.8, and 3.5 ± 11.7, respectively). Moreover, in all datasets the same MHCI molecule restricts less than 18% of all peptides. In additional file 1, we show the

distribution of commonly expressed MHCI alleles in each of the three datasets. The corresponding author will also provide these datasets upon written request.

Model building and evaluation

Cleavage models were trained and evaluated on datasets consisting of peptide fragments of the same length derived from MHCI-eluted peptides (proteasome models) and CD8 T cell epitopes (immunoproteasome models) and their C-terminal flanking regions, using the NGRAM-COUNT utility implemented by the SRILM package [25]. Peptide fragments encompassed two portions with the same number of residues, one fraction consisting of the C-terminal end of MHCI-eluted peptides or CD8 T cell epitopes, and the other one of their C-terminal flanking region. Cleavage sites -defined between the C-terminus of MHCI-restricted peptides (*P1* residue of cleavage site) and the most proximal C-terminal flanking residue (*P1'* residue)- were indicated by a "|" symbol. Cleavage models were generated considering peptide fragments ranging from 4 to 18 residues. Representative peptide fragments of 6 and 12 amino acids are C T L | T I G and P S C C T L | T I G V S S, respectively, where C T L and P S C C T L are two C-terminal portions of the peptide and T I G and T I G V S S are C-terminal flanking residues drawn from the protein source. Cleavage models were tested and evaluated at different thresholds using the SRLIM HIDDEN-NGRAM utility. HIDDEN-NGRAM is a word boundary program that uses *N-gram* models [25] produced by NGRAM-COUNT to predict the probability of hidden tags -cleavage sites- in any peptide fragment. The evaluation of the models was carried out through 5-fold cross-validation experiments that were repeated 5 times, obtaining mean estimations and standard deviations of the measures of performance indicated below.

Measures of performance

Cleavage predictions were examined in each residue at different probability thresholds (*th)* and were judged following the schema proposed in previous works [22,24]. It is assumed that cleavage sites should preferentially occur after the C-terminus of MHCI-restricted peptides (*P1* residue of cleavage site) than over any other position within the peptide. Under such schema, any given test peptide was classified as follows:

- TP (True positive): Cleavage score at the C-terminus (*P1* residue of cleavage site) is above the *th*.
- FN (False negative): Cleavage score of *P1* residue is below the *th*.
- TN (True negative): All the residues within the test fragment have a cleavage score bellow the *th*. Alternatively, if there are residues with cleavage scores above the *th*, but smaller than that of the *P1* residue.
- FP (False positive): There is at least one residue within the peptide with a cleavage score that is both, above the *th* and above that of the *P1* residue.

Upon this classification approach, we computed the Sensitivity (*SE*), Specificity (*SP*) and Matthews correlation coefficient (*MCC*) [31] of the predictions using Equations 1, 2 and 3, respectively,

$$SE = \frac{TP}{TP+FN} \tag{1}$$

$$SP = \frac{TN}{TN+FP} \tag{2}$$

$$MCC = \frac{(TP*TN)-(FN*FP)}{\sqrt{(TN+FN)(TP+FN)(TN+FP)(TP+FP)}} \tag{3}$$

In addition, we also computed the parameter *BTR* (Better Than Random) which was first introduced by Reche *et al.* [32] to compare the *SE* of a given model and that of a random model producing the same number of cleavage sites (Equation 4).

$$BTR = SE - ECS \tag{4}$$

ECS (Expected Cleavage Sites) represents the ratio of cleavage sites correctly predicted by a model that distributes cleavage sites randomly and is given by Equation 5.

$$ECS = \frac{C}{F*N} \tag{5}$$

Where *C* is the total number of cleavage sites (above the *th*) predicted by a given cleavage model in a test set of peptide fragments -specifically, within the MHCI-restricted peptide portion of the peptide fragment-; *F* is the number of MHCI-restricted peptide residues included in the peptide fragments used for training and testing; and *N* is the total number of peptide fragments in the dataset. Note that peptide fragments used for model building and evaluation encompassed two portions with the same number of residues, one consisting of the C-terminal end of MHCI-restricted peptides and the other of their C-terminal flanking region (details elsewhere in Methods). *ECS* is somewhat equivalent to the *SE* of a model that distributes all the cleavage sites randomly. Thus, the bigger the difference between *SE* and *ECS* the better the predictions produced by the model.

Prediction of peptide binding to MHCI

We used Position Specific Scoring Matrices (PSSMs) to compute binding scores of peptides to the relevant

MHCI molecules [33]. Actual binding of peptides to a particular MHCI molecule was assessed relating its binding score to those of 10000 reference peptides, 9-mers randomly obtained from SwissProt, computed using the same relevant PSSM. Thus, a given peptide was considered to bind a specific MHCI molecule when its binding score ranked among the *X* percentile (threshold) of top binding scores. The same peptide was considered not to bind to that MHCI if it ranked below the *X* percentile of top binding scores. PSSMs are derived from alignments of peptides of the same size known to bind to a given MHCI molecule [32,34,35]. Given that MHCI-bound peptides are usually of 9 residues of length, in this study we used PSSMs specific for the prediction of peptide binders of that length (9mers).

ROC analysis

We used 5 different sets of CD8 T cell epitopes consisting of 316, 50, 70, 47 and 30 peptides restricted by A*0201, A*0301, A*2402, B*0702, and B*2705, respectively, to evaluate the discovery rate of CD8 T cell epitopes using MHCI peptide-binding predictions alone, or in combination with proteasome cleavage predictions. Receiver operating characteristic (*ROC*) curves [36] were used to analyze the predictions. In the ROC analysis, we represented the *SE* (Equation 1) *versus 1-SP* (Equation 2) of the T cell epitope predictions obtained over a continuous range of percentile thresholds of MHCI binding (detail elsewhere in Methods). Non-T cell epitopes, required to compute the *SP* of the predictions, consisted of peptides of 9 residues randomly selected from the SwissProt database. A 1:3 ratio of T cell epitopes to non-T cell epitopes data was used. When evaluating the combination of MHCI binding and proteasome cleavage predictions, we applied a filtering approach such as that used by Dönnes and Kohlbacher [37]. Under this approach, peptides that are not predicted to be cleaved by the proteasome are discarded prior to the ROC analysis.

The area under ROC curves (*AUC*) was used as a global threshold-independent measure of performance. The maximum accuracy corresponds to an $AUC = 1$ while an $AUC = 0.5$ is indicative of a random prediction. Predictions are poor for values of $AUC > 0.7$, good for values of $AUC > 0.8$ and excellent for values of $AUC > 0.9$. ROC analyses were repeated 10 times, using the same T cell epitopes but different non-T cell epitopes. Thus, we obtained confident values of *AUC* (mean and standard deviation). Statistical significance of the differences between *AUC* values was evaluated using standard one-side two sample Student *t*- tests ($p <= 0.05$).

Web implementation

Immunoproteasome and proteasome cleavage models were implemented for free public use on the Web using a PERL CGI (Common Gateway Interface) script that executes the predictions on user-provided input data and returns the results to the browser. In addition, we used JavaScript for handling and verification of the input data before submission. Proteasome and immunoproteasome cleavage models exhibited optimal predictions at different model-specific cleavage scores. Therefore, cleavage scores by the different models were normalized and standardized so that cleavage sites are predicted at a score ≥ 0.5.

Results

Proteasome and immunoproteasome cleavage models

Cleavage models were generated from two types of MHCI-restricted peptides and their flanking regions using *N-grams*. *N-gram* models are frequently applied to speech recognition and natural language tagging [38], but they have also been applied to sequence analysis and motif identification [32,39-41]. We built two types of cleavage models. Immunoproteasome cleavage models were built upon a dataset encompassing 553 CD8 T cell epitopes that have been reported to be recognized by humans during the course of an infection. Epitope-specific CD8 T cell responses are generally primed by dendritic cells which express the immunoproteasome. Therefore, naturally processed CD8 T cell epitopes can be used to reproduced the cleavage by the immunoproteasome. In contrast, proteasome cleavage models were based on a set of 382 peptides that were eluted from human MHCI molecules. Peptide elution experiments are generally carried out using various types of cells (virtually never dendritic cells) and under conditions that do not induce the expression of the immunoproteasome. Therefore, we considered that MHCI-eluted peptides are produced by the proteasome. A detailed description of these datasets is elsewhere in Methods.

Numerous immunoproteasome and proteasome cleavage models were obtained from different training sets consisting of peptide fragments varying from 4 to 18 residues -in a given training set, all the peptides have the same size. Peptide fragments used for training included the C-terminus (*P1* residue of cleavage site) of MHCI-restricted peptides (CD8 T cell epitopes and MHCI-eluted peptides) and comprised two distinct portions with the same number of residues: one consisting of the C-terminal end of MHCI-restricted peptides and the other one of their C-terminal flanking region (see Methods section for more details). Cleavage models were evaluated in 5-fold cross-validation experiments, considering a continuous range of cleavage thresholds. As measures of performance we computed *SE*, *SP*, *MCC* and *BTR* (see Methods section for details), but trusted *BTR* as the key measure of the goodness of the predictions. In Figure 1A we show the optimal *BTR* achieved by the cleavage models with regard to the size of the peptide fragments used

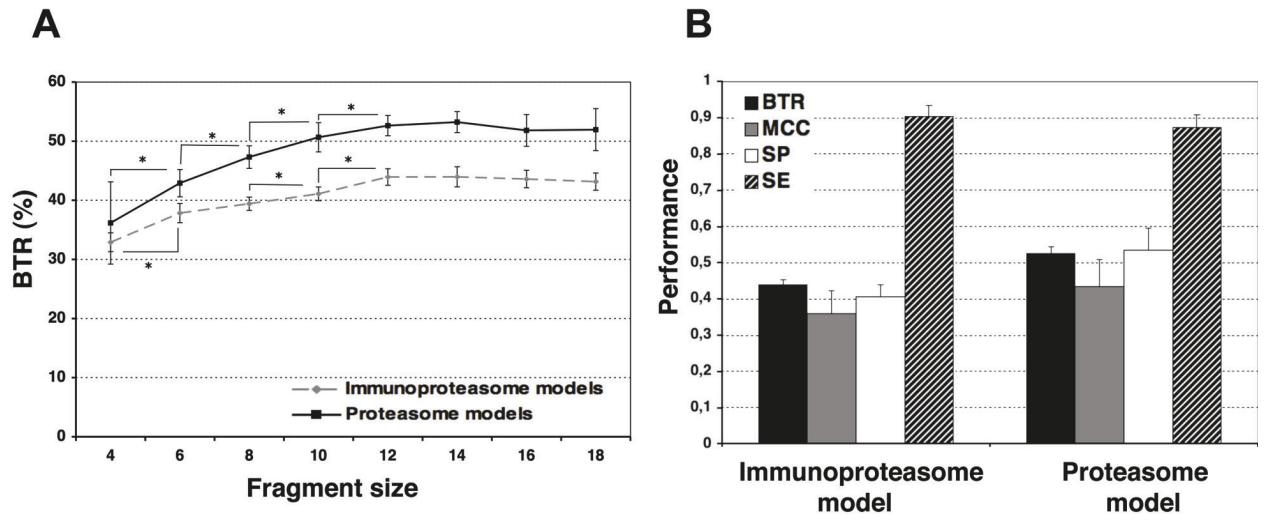


Figure 1 Evaluation of immunoproteasome and proteasome prediction models. The predictive performance of proteasome models was evaluated in 5 fold cross-validation experiments using *MCC*, *BTR*, *SP* and *SE* as measures of performance. Proteasome models were built and tested using MHCI-eluted peptide ligands whereas immunoproteasome models were built and tested using MHCI-restricted CD8 T cell epitopes. A) Predictive performance (*BTR*) achieved by immunoproteasome (*grey line*) and proteasome (*black line*) models trained and tested on peptides of different fragment lengths (abscissa). Statistically significant increments in *BTR* are indicated with "*" symbols. B) Predictive performance (*BTR*, *MCC*, *SP*, *SE*) achieved by the selected proteasome and immunoproteasome models built on peptide fragments of 12 residues.

for training. A complete summary of the performance of the cleavage models, which also includes the *MCC*, *SE*, *SP* of the predictions, is shown in Table 1.

The predictive performance of the cleavage models significantly increased ($p < 0.05$) with the length of the peptide fragments used for training, picking at a fragment size of 12-14 residues (Figure 1A); $BTR = 0.44 \pm 0.02$ for the immunoproteasome model and $BTR = 0.53 \pm 0.02$ for the proteasome model. In general, the predictive performance of proteasome cleavage models built upon MHCI-eluted peptides was higher than that achieved by immunoproteasome cleavage models, regardless of the length the peptides fragments used for training (Figure 1A). Increasing the size of the peptide fragments beyond 14 residues did not improve the predictive performance of the cleavage models (Figure 1A). Judging the predictions by the *MCC*, the immunoproteasome and proteasome models that were built on peptide fragments of 12 residues (Table 1) achieved the best results. Because no statistical difference was observed between the *BTR* achieved by the models trained on 12 and 14 residues, for further analysis, we used the models trained on 12-residue peptide fragments. The performance of the selected proteasome and immunoproteasome models is summarized in Figure 1B.

Comparison of the immunoproteasome and proteasome cleavage models

For further comparisons, we evaluated the immunoproteasome and proteasome cleavage models in an independent test set built from 137 HIV1-specific CD8 T cell epitopes and their flanking regions (Figure 2). The immunoproteasome model achieved better results than the proteasome model, as judged by both, the *BTR* (0.45 for the immunoproteasome model and 0.39 for the proteasome model) and the *MCC* (0.30 for the immunoproteasome model and 0.18 for the proteasome model). These results indicate that the immunoproteasome model appears to be more suitable than the proteasome model to predict the cleavage sites defined by CD8 T cell epitopes.

Using the immunoproteasome and proteasome cleavage models, we analyzed the fragmentation patterns resulted from 100 proteins randomly selected from the SwissProt database (Figure 3). The immunoproteasome cleavage model generated fragments with a mean size of 2.23 ± 1.61 residues, whereas the proteasome cleavage model generated fragments with a mean size of 3.02 ± 2.33 residues. Using a Wilcoxon test, we observed no significant difference between the sizes of the fragments generated with the proteasome and immunoproteasome models (Figure 3A). This analysis also revealed that 36% of the peptide fragments generated by the proteasome and immunoproteasome are identical, and 67% of the cleavage sites are shared (Figure 3B).

Comparison with NetChop

We also used the 137 HIV1-specific CD8 T cell epitopes and their flanking regions to compare the cleavage predictions obtained with our *N-gram* cleavage models and those obtained using the NetChop web sever. The NetChop system uses an artificial neural-network model

Table 1 Predictive performance of immunoproteasome and proteasome cleavage models

Immunoproteasome					
Size	***SE***	***SP***	***ECS***	***MCC***	***BTR***
4	0.807 ± 0.030	0.851 ± 0.039	47.828 ± 2.001	0.660 ± 0.038	0.329 ± 0.016
6	0.763 ± 0.036	0.708 ± 0.042	38.495 ± 0.614	0.472 ± 0.069	0.378 ± 0.016
8	0.906 ± 0.023	0.545 ± 0.038	51.219 ± 1.008	0.484 ± 0.059	0.394 ± 0.011
10	0.802 ± 0.024	0.462 ± 0.019	39.083 ± 1.003	0.281 ± 0.045	0.411 ± 0.012
12	0.903 ± 0.031	0.407 ± 0.031	46.339 ± 0.481	0.357 ± 0.062	0.439 ± 0.014
14	0.872 ± 0.035	0.374 ± 0.023	43.190 ± 1.498	0.284 ± 0.056	0.434 ± 0.017
16	0.855 ± 0.030	0.306 ± 0.041	41.908 ± 1.406	0.193 ± 0.047	0.436 ± 0.015
18	0.857 ± 0.031	0.290 ± 0.028	42.536 ± 1.081	0.179 ± 0.039	0.432 ± 0.015
Proteasome					
Size	***SE***	***SP***	***ECS***	***MCC***	***BTR***
4	0.803 ± 0.125	0.871 ± 0.052	44.110 ± 9.249	0.681 ± 0.089	0.362 ± 0.069
6	0.792 ± 0.048	0.723 ± 0.037	36.274 ± 1.943	0.516 ± 0.082	0.429 ± 0.023
8	0.855 ± 0.037	0.603 ± 0.047	38.160 ± 2.112	0.473 ± 0.072	0.473 ± 0.019
10	0.885 ± 0.050	0.537 ± 0.046	37.839 ± 2.355	0.452 ± 0.069	0.506 ± 0.025
12	0.874 ± 0.034	0.534 ± 0.062	34.970 ± 1.704	0.434 ± 0.075	0.526 ± 0.017
14	0.871 ± 0.037	0.468 ± 0.065	33.699 ± 1.432	0.371 ± 0.085	0.532 ± 0.018
16	0.844 ± 0.058	0.403 ± 0.065	32.657 ± 1.692	0.276 ± 0.096	0.518 ± 0.027
18	0.794 ± 0.077	0.392 ± 0.060	27.510 ± 1.978	0.206 ± 0.126	0.519 ± 0.035

Cleavage models were built on peptide fragments of a given size encompassing the C-terminal end of MHCI-eluted ligands (proteasome model) or CD8 T cell epitopes (immunoproteasome) and the corresponding C-terminal flanking residues. Predictive performance was evaluated in 5-fold cross-validation experiments. *SE*: sensitivity; *SP*: specificity; *MCC*: Matthew's correlation coefficient; *BTR*: better than random (Eq. 4).

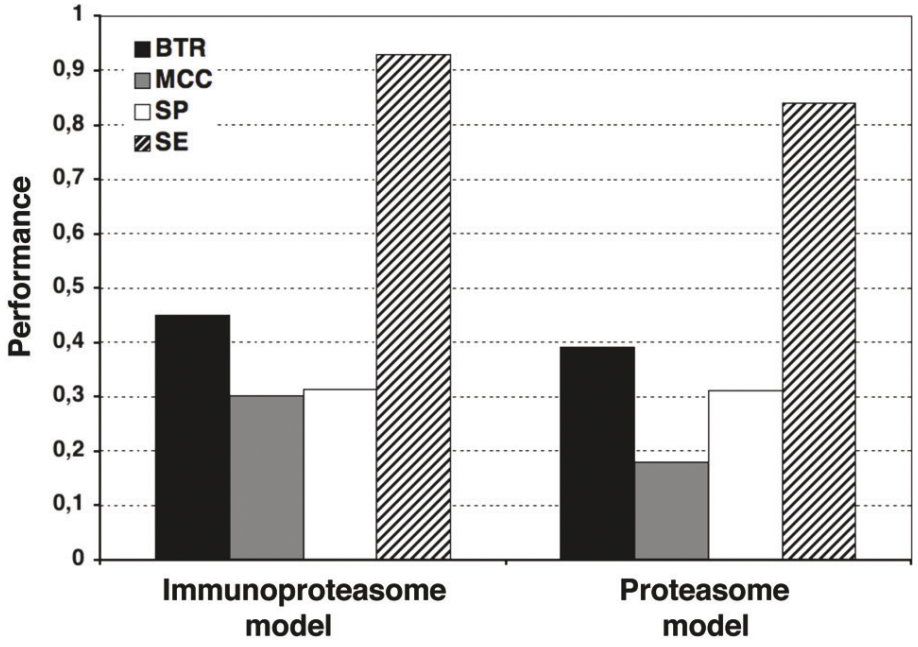


Figure 2 Model evaluation using an independent test dataset. The proteasome and immunoproteasome models were evaluated using an independent test consisting of HIV1-specific CD8 T cell epitopes. The predictive performance was evaluated using *BTR* (*black bars*), *MCC* (*grey bars*), *SP* (*white bars*) and *SE* (*pattern bars*).

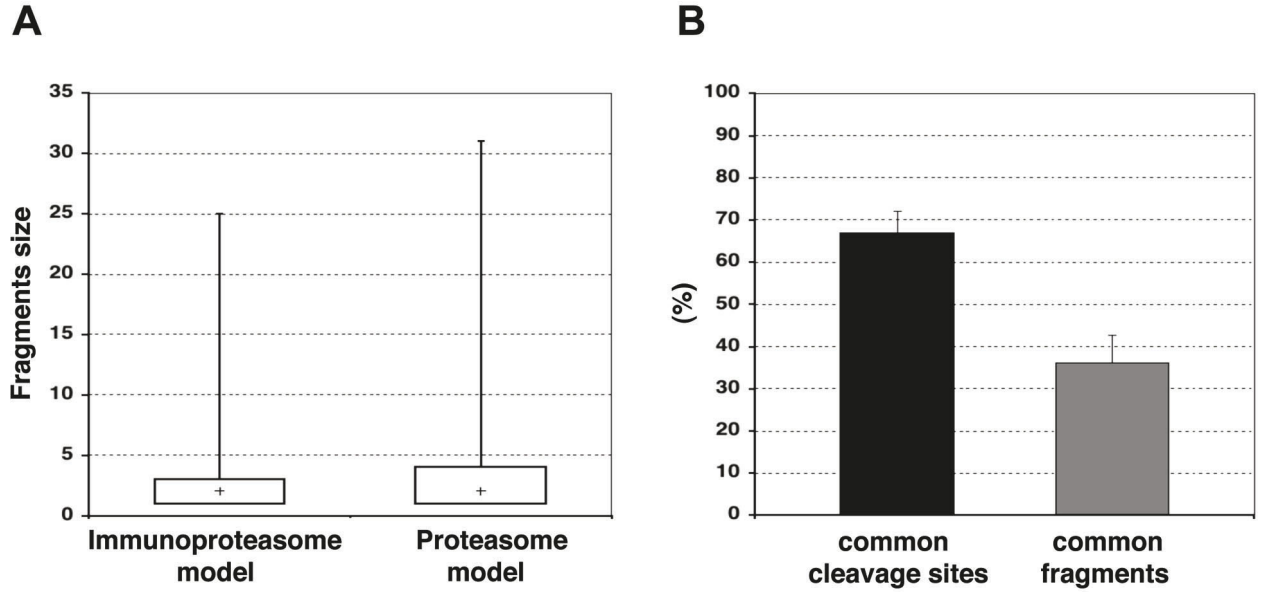


Figure 3 Analysis of fragmentation patterns produced by the immunoproteasome and proteasome models. A) BoxPlot of the fragment size distribution obtained with the immunoproteasome and the proteasome models. Not significant difference was observed, using a Wilcoxon test, between the sizes of the fragments generated by each model B) Overlap between the fragments generated by the immunoproteasome and proteasome models represented as the percentage of common sites (*black bar*) and the percentage of common fragments (*grey bar*). Fragments were obtained from 100 proteins randomly selected from the Swissprot database. We used the optimal proteasome and immunoproteasome models, which were built on 12-residue peptide fragments.

that was built upon MHCI-restricted peptides [21]. For this comparison, we used NetChop default settings (cleavage sites occur after residues having a probability of 0.5 or higher) in computing the *SE*, *SP*, and *MCC* of the predictions following the same schema reported by the NetChop developers [22] (see Methods section for details). In addition, we computed the *BTR* parameter defined in this study. Because NetChop models were trained on 18-residue peptide fragments consisting of full-length MHCI-restricted peptides (9 residues) and the most proximal 9 residues flanking the C-terminus, in this comparison we evaluated *SE*, *SP*, *MCC* and *BTR* on peptide fragments consisting of the full-length HIV1-specific CD8 T cell epitopes. Note that in previous analyses these parameters were evaluated on the portion of the peptide fragments corresponding to the MHCI-restricted peptides. The results of this analysis are depicted in Figure 4. The immunoproteasome and proteasome *N-gram* models achieved *MCC* values (0.20 and 0.19, respectively) similar to those obtained using NetChop (0.18). Likewise, NetChop and our *N-gram* models achieved similar *BTR* values around 0.44 (Figure 4).

Combination of MHCI-peptide binding and cleavage predictions

We also evaluated the impact of combining cleavage and MHCI-peptide binding predictions on T cell epitope identification. Specifically, using a ROC analysis (see Methods section for details), we analyzed the result of such combination to discriminate CD8 T cell epitopes restricted by 5 different MHCI molecules (A*0201, A*0301, A*2402, B*0702 and B*2705) from random peptides. We combined MHCI-peptide binding predictions with cleavage predictions by the immunoproteasome and proteasome models, individually or together, and used *AUC* values (computed after the ROC analyses, see Methods for details) as a measure of the goodness of the predictions (Figure 5).

MHCI-peptide binding predictions alone achieved high *AUC* values above 0.9 -regardless of the MHCI molecule-, that did not leave much margin to observe any large improvements on CD8 T cell epitope predictions. Nevertheless, combining the proteasome and immunoproteasome models separately or together with MHCI-peptide binding predictions resulted in increased *AUC* values (Figure 5B). Moreover, such increases were statistically significant ($p < 0.05$) in all cases. The major increment in *AUC* was observed for A*0301-restricted epitopes. Alone, MHCI-peptide binding predictions reached an $AUC = 0.9063 \pm 0.0141$ for A*0301, whereas in combination with the immunoproteasome and proteasome cleavage predictions achieved *AUC* values of 0.9416 ± 0.017, and 0.9411 ± 0.0095, respectively.

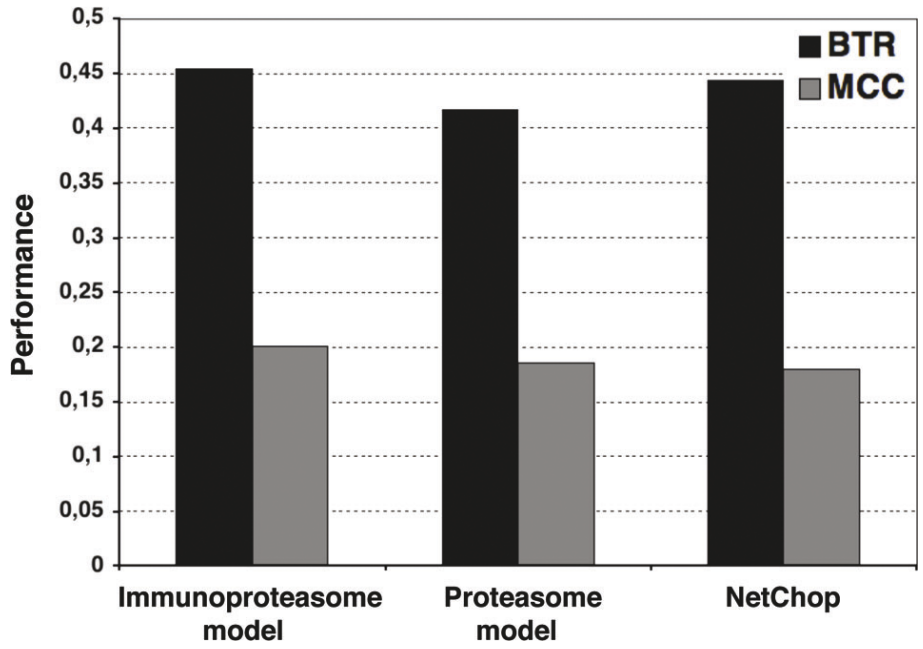


Figure 4 Comparative analysis of cleavage predictions. The figure depicts the *MCC* (*black bars*) and the *BTR* (*grey bars*) achieved by our immunoproteasome and proteasome models and NetChop on an independent test set of 137 HIV1-specific CD8 T cell epitopes. Because NetChop was built using complete nonameric MHCI-restricted peptides, in this analysis we have evaluated the cleavage predictions by the three models over the entire length of the T cell epitopes being tested.

There were no differences between the results obtained combining MHCI-peptide binding and the cleavage predictions by the immunoproteasome model or the proteasome model, but the joint combination of both cleavage models (immunoproteasome and proteasome) with the MHCI-peptide binding resulted in *AUC* values larger than those obtained using single cleavage models (Figure 5B). Nevertheless, with the exception of A*0201 ($p < 0.05$), these increases in *AUC* were not statistically significant with regard to those *AUC* obtained using solely either cleavage model (Figure 5B).

Enhanced *AUC* values obtained upon combining the cleavage models with MHCI-peptide binding predictions are due to the reduction of the number of false positives detected with regard to the MHCI-peptide binding predictions alone (Figure 5C). Taking MHCI-peptide binding predictions alone as reference, we observed a ~56% decrease of false positives (computed over the entire range of thresholds used in the *ROC* analysis) when using the immunoproteasome model. The reduction of false positives was even larger (68%) when using the proteasome model and increased slightly when both models were combined (70%).

Proteasome Cleavage Prediction Server (PCPS)

We developed PCPS (Proteasome Cleavage Prediction Server) to allow the prediction of proteasome and immunoproteasome cleavage through our *N-gram* models. PCPS is available for free public use at http://imed.med.ucm.es/Tools/PCPS/. PCPS was designed to be intuitive and user friendly (Figure 6A). The main input data for PCPS is one or several protein sequences that can be pasted or uploaded to the server in multiple formats, including FASTA, IG, GenBank, EMBL, Phylip, NBRF, GCG, DNAStrider, PIR, MSF, ASN and PAUP. The sequences provided to the server are subjected to a cleavage analysis using *N-gram* models that are selected by the user from the CLEAVAGE MODELS section. There are several models available for both proteasomes, constitutive and immunoproteasome, which differ in sensitivity and specificity, and users can combine different proteasome and immunoproteasome models. Cleavage models in PCPS were trained on peptide fragments of 12 (*models 1*), 8 (*models 2*) and 6 (*models 3*) residues. The models trained on 12 residues exhibited the best performance ($MCC = 0.43 \pm 0.07$ for the proteasome cleavage model and $MCC = 0.36 \pm 0.06$ for the immunoproteasome cleavage model) (Table 1). The output of PCPS consists of a table indicating the cleavage score of each residue in the protein queries (Figure 6B). Computed scores reflect the likelihood that the proteasome/immunoproteasome would cleave the protein after such residue (*P1* residue of cleavage site). Whenever the cleavage score is higher than 0.5, a tick marks the corresponding residue. The different models actually differ in the sensitivity, specificity, and BTR of the predictions.

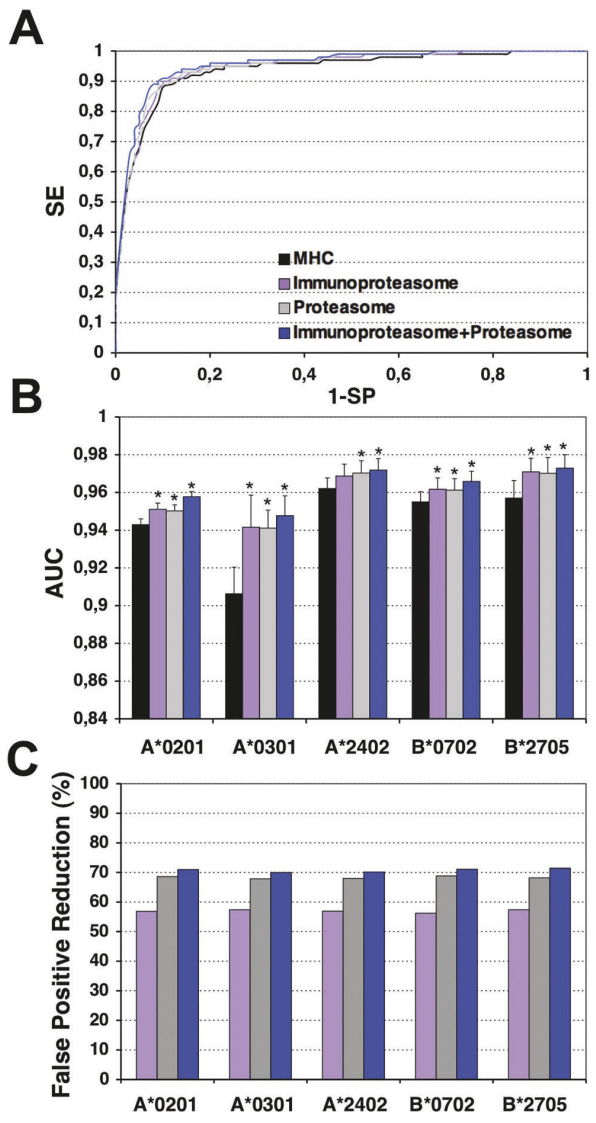


Figure 5 Prediction of T cell epitopes using MHCI-peptide binding and cleavage models. A) ROC curves depicting the prediction of T cell epitopes restricted by A*0201 using MHCI-peptide binding prediction alone (*black line*) or in combination with cleavage predictions by the immunoproteasome (*purple line*), the proteasome (*grey line*) and both cleavage models together (*blue line*). B) *AUC* values obtained for the prediction of T cell epitopes restricted by 5 different HLA I molecules (A*0201, A*0301, A*2402, B*0702, B*2705) using MHCI-peptide binding predictions alone (*black bars*) and in combination with cleavage predictions by the immunoproteasome (*purple bars*), proteasome (*grey bars*) and both cleavage models together (*blue bars*). Significant increases ($p < 0.05$) in *AUC* values with regard to MHCI-peptide binding predictions alone (*black bars*) are indicated with the symbol "*". C) Reduction of false positives. The figure represents the decrease in percentage of false positives after introducing cleavage predictions by the immunoproteasome (*purple bars*), the proteasome (*grey bars*) and both together (*blue bars*). False positive reduction was computed over the entire ROC analysis.

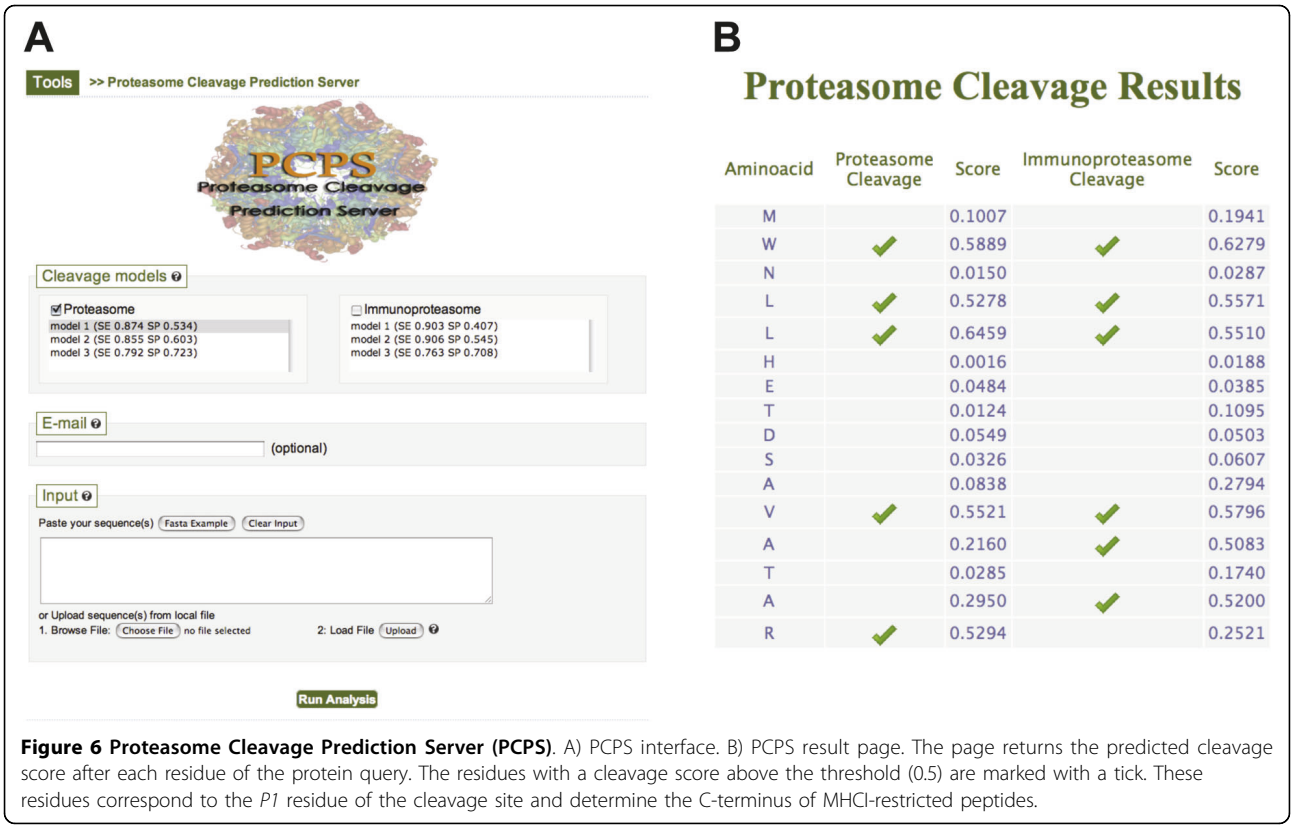
In PCPS, the indicated specificity and sensitivity of the models were achieved at cleavage thresholds of 0.5, but users can experiment with the server and decide different cleavage thresholds.

Discussion

It is generally believed that the C-terminus of most CD8 T cell epitopes, and in general that of most MHCI-restricted peptides, results from the proteolytic cleavage by the proteasome [4,9]. Some other proteases, chiefly tripeptidyl peptidase II (TPP II), also play some role generating the C-terminus of some MHCI-restricted peptides[42-44], specifically through the degradation of some proteolytic products generated by the proteasome that are longer than 15 residues [43]. However, because the majority of the peptide fragments generated by the proteasome are shorter than 15 residues [13], the proteasome is still the principal source of the C-terminus of peptides that are bound to MHCI molecules. As a result, proteasome cleavage models can be derived using cleavage sites recreated from MHCI-restricted peptides and their C-terminal flanking regions [21-23].

There are two types of proteasomes, the immunoproteasome and the constitutive proteasome, which differ in their cleavage patterns [14]. The constitutive proteasome is the form expressed in most nucleated cells, whereas the immunoproteasome is constitutively expressed in mature dendritic cells. Antigen presentation by dendritic cells is generally required to prime and instruct naïve CD8 T cells in an antigen specific manner. Subsequently, the effector function of CD8 T cells is executed upon recognizing the same antigenic peptides on target cells [45]. Consequently, the immunoproteasome is responsible for the generation of the C-terminus of the peptides that elicit the CD8 T cell response, whereas the constitutive proteasome determines the C-terminus of the MHCI-peptide ligands that can be the targets of such response. Protective CD8 T cell epitopes are likely those generated by both, the constitutive proteasome and the immunoproteasome [15].

In this work, we have assumed that MHCI-eluted peptides reflect protein degradation by the proteasome, whereas bona fide identified CD8 T cell epitopes elicited in patients during the course of an infection reflect protein degradation by the immunoproteasome, but not necessarily by the proteasome. The latter is due to the fact that epitope verification is generally carried out by measuring the response of T cells to synthetic peptides loaded onto antigen presenting cells, which are seldom dendritic, thus bypassing antigen processing by the proteasome in the test target cells. Subsequently, using *N-grams*, we have modeled the proteasome and immunoproteasome cleavage



Aminoacid	Proteasome Cleavage	Score	Immunoproteasome Cleavage	Score
M		0.1007		0.1941
W	✓	0.5889	✓	0.6279
N		0.0150		0.0287
L	✓	0.5278	✓	0.5571
L	✓	0.6459	✓	0.5510
H		0.0016		0.0188
E		0.0484		0.0385
T		0.0124		0.1095
D		0.0549		0.0503
S		0.0326		0.0607
A		0.0838		0.2794
V	✓	0.5521	✓	0.5796
A		0.2160	✓	0.5083
T		0.0285		0.1740
A		0.2950	✓	0.5200
R	✓	0.5294		0.2521

Figure 6 Proteasome Cleavage Prediction Server (PCPS). A) PCPS interface. B) PCPS result page. The page returns the predicted cleavage score after each residue of the protein query. The residues with a cleavage score above the threshold (0.5) are marked with a tick. These residues correspond to the *P1* residue of the cleavage site and determine the C-terminus of MHCI-restricted peptides.

from datasets of peptide fragments of different length built upon MHCI-eluted peptides (proteasome model) and CD8 T cell epitopes (immunoproteasome model) and their C-terminal flanking regions. These models predict whether the C-terminus of a given peptide, in the context of its flaking residues, is likely to result from the proteolytic activity of the proteasome and/or the immunoproteasome (*P1* residue of cleavage site).

The best cleavage predictions for both proteasomes, constitutive and immunoproteasome, were obtained using *N-grams* trained on 12-residue peptide fragments, encompassing the 6 most proximal flanking residues to the C-terminus of the MHCI-restricted peptides preceded by 6 residues from the C-terminal end of the MHCI-restricted peptides (Figure 1). These results are consistent with reports indicating that proteasomes and immunoproteasomes scrutinize between 10 and 12 residues [10,26]. In contrast, related methods for the prediction of proteasome cleavage that are based on MHCI-restricted peptides have been trained using 18 to 20 residue peptide fragments [19,21,23], which, makes these models, regardless of the results, somewhat artificial.

In cross-validation, the predictive performance of proteasome models exceeded that of immunoproteasome models; the best proteasome cleavage model achieved a $BTR = 0.53 \pm 0.02$ and an $MCC = 0.43 \pm 0.07$, whereas the best immunoproteasome model achieved a $BTR = 0.44 \pm 0.01$ and an $MCC = 0.36 \pm 0.06$. Despite that both sets of peptides were subjected to the same sequence reduction procedure (See Methods), these results likely reflect that the set of CD8 T cell epitopes is more numerous and arguably more diverse than the set of MHCI-eluted peptide (see Results). Dendritic cells exhibit non-classical pathways on antigen presentation and some can be immunoproteasome independent [45,46], which could actually account for a higher diversity in the epitope dataset. Nonetheless, the best immunoproteasome model achieved better results than the corresponding proteasome model when predicting the cleavage sites encompassed by an independent set consisting of HIV1-specific CD8 T cell epitopes (Figure 2). Taking into account all the above, the immunoproteasome model appears to be the most suitable to predict the C-terminus of CD8 T cell epitopes.

Our constitutive proteasome and immunoproteasome models produced different but overlapping fragmentation patters that mirror those observed experimentally [13]; 68% of the cleavage sites (*P1* residues) and 36% of the fragments generated were identical (Figure 3B). However, the fragments yielded by the immunoproteasome and proteasome models were much smaller (2-3 residues)

than those determined experimentally (7-9 residues) [13,47]. The smaller fragment sizes produced by our models may reproduce the clustering and overlapping of epitopes found in protein regions [48]. On the other hand, it is important to note that our models are not meant, and are not suitable, to predict proteolytic fragments, but to indicate whether the C-terminus of a peptide can result from the cleavage produced by the proteasome and/or the immunoproteasome. Proteasome fragmentation patterns (the size of fragments) may be better reproduced by methods trained on actual cleavage data such as that by Tenzer *et al* [18].

Using a test set of HIV1-specific CD8 T cell epitopes, we found that the predictive performance of our optimal proteasome and immunoproteasome cleavage models was comparable to that of NetChop [22]; a reference method to predict proteasome cleavage sites [24] that it was also developed from MHCI-restricted peptides. The immunoproteasome and proteasome cleavage models achieved *MCC* values of 0.20 and 0.19, respectively, while NetChop achieved an *MCC* = 0.18. It is worth nothing that these results were obtained under conditions that were optimal for NetChop. First, NetChop was trained on peptide fragments encompassing full length MHCI-restricted peptides [22], and here we have evaluated and compared the cleavage predictions over the entire epitope sequences. Note that we only used a portion of the MHCI-restricted peptides for training (6 residues). Second, the HIV1-specific CD8 T cell epitopes used for testing were not used for training our *N-gram* models but were likely included in the NetChop training dataset. It is also important to mention that NetChop has been described as an immunoproteasome cleavage prediction method, but in fact it was trained on a dataset consisting of both, MHCI-eluted ligands and CD8 T cell epitopes. As we have discussed here, CD8 T cell epitopes can be considered as generated by the immunoproteasome. However, it is more appropriated to consider MHCI-eluted peptides as generated by the constitutive proteasome because they are obtained from different type of cells but seldom from dendritic cells. In sum, we have dealt with the prediction of proteasome and immproteasome cleavage sites from MHC-restricted peptides in a manner that is consistent with the mechanism of antigen presentation and recognition, and achieved a notorious performance.

Prediction of proteasome and immunoproteasome cleavage sites using our models is available at http://imed.med.ucm.es/Tools/PCPS/. In addition, there are several other online servers to predict proteasome cleavage, which differ in the data and approach used for generating the models [16,22,49]. Nonetheless, the problem of identifying proteasome cleavage sites with high precision is still far from being solved. A simple manner to improve the prediction of proteasome cleavage sites could likely be achieved trough a meta-server that would arrive to a consensus prediction from the available proteasome cleavage predictors. Such a consensus approach has resulted successful in the also difficult task of predicting peptide binding to MHC class II molecules [50].

It has been reported that proteasome prediction models can improve T cell epitope identification when combined with MHCI-peptide binding predictions [18,22,37,51,52]. Likewise, our proteasome and immunoproteasome models, separately or together, also served to improve CD8 T cell epitope discrimination when combined with MHCI-binding predictions (Figure 5). The improvements, judged by increases in *AUC*, could appear minor but were statistically significant (Figure 5), and were linked to a large reduction of the number of false positives detected (up to 70%). Therefore, combining proteasome cleavage and MHCI-peptide binding predictions would serve to decrease the experimental toll involved in epitope identification; there will be less peptides to be tested. The proteasome cleavage model alone or juxtaposed with the immunoproteasome model resulted in a significant loss of true positives (up to 20%). Therefore, the proteasome cleavage model will be more useful on large-scale epitope identification scenarios (e.g. predicting CD8 T cell epitopes from a large number of antigens). Finally, combining cleavage predictions by both proteasomes, constitutive and immunoproteasome, with MHCI-binding predictions ought to help defining protective CD8 T cell epitopes. Overall, these results call for the integration of our proteasome models with others taking into account TAP transport and MHC binding, as already pioneered by other authors [18,22,37,51,52].

Conclusion

We have derived *N-gram* models specific for the proteasome and the immunoproteasome that are consistent with the known biology of antigen presentation. The proteasome models were built upon MHCI-eluted peptides whereas the immunoproteasome models were built upon CD8 T cell epitopes. The *N-gram* models that exhibited the best performance were trained on 12-residue peptides, 6 residues at each side of the cleavage site, defined by the C-terminus of MHCI-restricted peptides and the most proximal C-terminal flanking residue. Finally, we have shown that combining cleavage predictions by the proteasome and immunoproteasome models with MHCI-binding predictions improves CD8 T cell epitope prediction. Cleavage predictions using our *N-gram* models are available for free public use at the PCPS site http://imed.med.ucm.es/Tools/PCPS/.

Additional material

Adittional file 1: MHCI allele distribution in peptide datasets. The figure depicts the percentage of peptides restricted by 7 commonly expressed human MHCI alleles (A*0201, A*0301, A*1101, A*2402, B*0702, B*0801, B*2705) in the three datasets used in this study.

Abbreviations used

MHC: I molecules, major histocompatibility class I molecules; N-terminus: amino-terminus; C-terminus: carboxy-terminus.

Authors' contributions

CMDR did the work and wrote paper. EML interpreted results and wrote paper. PAR designed the work, interpreted results and rendered the final paper. All authors read and approved the final manuscript.

Acknowledgements

We wish to tank Dr Elena Rodriguez-Garcia for corrections and thoughtful comments. This work was supported by Grants SAF2006-07879 and SAF2009-08103 from Ministerio de Ciencia e Innovación of Spain, and by Grant CCG08-UCM/BIO-3769 from Comunidad Autonoma de Madrid to PAR.

Author details

[1]Laboratory of Immunomedicine, Department of Microbiology I-Immunology, Facultad de Medicina, Universidad Complutense de Madrid, Ave Complutense S/N, Madrid 28040, Spain. [2]Department of Microbiology I-Immunology, Facultad de Medicina, Universidad Complutense de Madrid, Ave Complutense S/N, Madrid 28040, Spain.

Received: 6 May 2010 Accepted: 23 September 2010
Published: 23 September 2010

References

1. Garcia KC, Teyton L, Wilson IA: **Structural basis of T cell recognition.** *Annu Rev Immunol* 1999, **17**:369-397.
2. Margulies DH: **Interactions of TCRs with MHC-peptide complexes: a quantitative basis for mechanistic models.** *Curr Opin Immunol* 1997, **9(3)**:390-395.
3. Wang J-H, Reinherz E: **Structural basis of T cell recognition of peptides bound to MHC molecules.** *Molecular Immunology* 2001, **38**:1039-1049.
4. Pamer E, Cresswell P: **Mechanisms of MHC class I–restricted antigen processing.** *Annu Rev Immunol* 1998, **16**:323-358.
5. Kloetzel PM: **Antigen processing by the proteasome.** *Nat Rev Mol Cell Biol* 2001, **2(3)**:179-187.
6. Serwold T, Gonzalez F, Kim J, Jacob , Shastri N: **ERAAP customizes peptides for MHC class I molecules in the endoplasmic reticulum.** *Nature* 2002, **419(6906)**:480-483.
7. Craiu A, Akopian T, Goldberg A, Rock KL: **Two distinct proteolytic processes in the generation of a major histocompatibility complex class I-presented peptide.** *Proc Natl Acad Sci USA* 1997, **94(20)**:10850-10855.
8. Rock KL, Gramm C, Rothstein L, Clark K, Stein R, Dick L, Hwang D, Goldberg AL: **Inhibitors of the proteasome block the degradation of most cell proteins and the generation of peptides presented on MHC class I molecules.** *Cell* 1994, **78(5)**:761-771.
9. Rock KL, Goldberg AL: **Degradation of cell proteins and the generation of MHC class I-presented peptides.** *Annu Rev Immunol* 1999, **17**:739-779.
10. Nussbaum AK, Dick TP, Keilholz W, Schirle M, Stevanović S, Dietz K, Heinemeyer W, Groll M, Wolf DH, Huber R, *et al*: **Cleavage motifs of the yeast 20 S proteasome β subunits deduced from digests of enolase 1.** *Proc Nat Acad Sci* 1998, **95**:12504-12509.
11. Groettrup M, Standera S, Stohwasser R, Kloetzel PM: **The subunits MECL-1 and LMP2 are mutually required for incorporation into the 20 S proteasome.** *Proc Natl Acad Sci USA* 1997, **94(17)**:8970-8975.
12. Morel S, Levy F, Burlet-Schiltz O, Brasseur F, Probst-Kepper M, Peitrequin AL, Monsarrat B, Van Velthoven , Cerottini JC, Boon T, *et al*: **Processing of some antigens by the standard proteasome but not by the immunoproteasome results in poor presentation by dendritic cells.** *Immunity* 2000, **12(1)**:107-117.
13. Toes E, Nussbaum AK, Degermann S, Schirle M, Emmerich NP, Kraft M, Laplace C, Zwinderman A, Dick TP, Muller J, *et al*: **Discrete cleavage motifs of constitutive and immunoproteasomes revealed by quantitative analysis of cleavage products.** *J Exp Med* 2001, **194(1)**:1-12.
14. Gaczynska M, Rock K, Spies T, Goldberg A: **Peptidase activities of proteasomes are differentially regulated by the major histocompatibility complex-encoded genes for LMP2 and LMP7.** *Proc Natl Acad Sci USA* 1994, **91(20)**:9213-9217.
15. Chapiro J, Claverol S, Piette F, Ma W, Stroobant V, Guillaume B, Gairin JE, Morel S, Burlet-Schiltz O, Monsarrat B, *et al*: **Destructive cleavage of antigenic peptides either by the immunoproteasome or by the standard proteasome results in differential antigen presentation.** *J Immunol* 2006, **176(2)**:1053-1061.
16. Nussbaum A, Kuttler C, Hadeler K, Rammensee H, Schild H: **PAProC: a prediction algorithm for proteasomal cleavages available on the WWW.** *Immunogenetics* 2001, **53(2)**:87-94.
17. 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(3)**:417-429.
18. Tenzer S, Peters B, Bulik S, Schoor O, Lemmel C, Schatz MM, Kloetzel PM, Rammensee HG, Schild H, Holzhutter HG: **Modeling the MHC class I pathway by combining predictions of proteasomal cleavage, TAP transport and MHC class I binding.** *Cell Mol Life Sci* 2005, **62(9)**:1025-1037.
19. Holzhütter H, Frömmel C, Kloetzel P: **A theoretical approach towards the identification of cleavage-determining amino acid motifs of the 20 S proteasome.** *J Mol Biol* 1999, **286(4)**:1251-1265.
20. Holzhutter HG, Kloetzel PM: **A kinetic model of vertebrate 20 S proteasome accounting for the generation of major proteolytic fragments from oligomeric peptide substrates.** *Biophys J* 2000, **79(3)**:1196-1205.
21. Kesmir C, Nussbaum AK, Schild H, Detours V, Brunak S: **Prediction of proteasome cleavage motifs by neural networks.** *Protein Eng* 2002, **15(4)**:287-296.
22. Nielsen M, Lundegaard C, Lund O, Kesmir C: **The role of the proteasome in generating cytotoxic T-cell epitopes: insights obtained from improved predictions of proteasomal cleavage.** *Immunogenetics* 2005, **57(1-2)**:33-41.
23. Bhasin M, Raghava GPS: **Pcleavage: an SVM based method for prediction of constitutive proteasome and immunoproteasome cleavage sites in antigenic sequences.** *Nucleic Acids Research* 2005, **33**:W202-W207.
24. Saxová P, Buus S, Brunak S, Kesmir C: **Predicting proteasomal cleavage sites: a comparison of available methods.** *International Immunology* 2003, **15(7)**:781-787.
25. Stolcke A: **SRILM – An Extensible Language Modeling Toolkit.** In *Proceedings of the International Conference of Spoken Language Processing.* Edited by: JJ Ohala TMN, BL Derwing M, Hodge M, Wiebe GE. Boulder, CO: Center for Spoken Language Research; 2002:**2**:901-904.
26. Altuvia Y, Margalit H: **Sequence signals for generation of antigenic peptides by the proteasome: implications for proteasomal cleavage mechanism.** *J Mol Biol* 2000, **295(4)**:879-890.
27. Reche PA, Zhang H, Glutting JP, Reinherz EL: **EPIMHC: a curated database of MHC-binding peptides for customized computational vaccinology.** *Bioinformatics* 2005, **21(9)**:2140-2141, Epub 2005 Jan 2118.
28. Peters B, Sidney J, Bourne P, Bui H, Buus S, Doh G, Fleri W, Kronenberg M, Kubo R, Lund O, *et al*: **The immune epitope database and analysis resource: from vision to blueprint.** *PLoS Biol* 2005, **3(3)**:e91.
29. **HIV Molecular Immunology 2006/2007.** Los Alamos, New Mexico: Los Alamos National Laboratory, Theoretical Biology and Biophysics 2007.
30. Neuwald AF, Liu JS, Lawrence CE: **Gibbs motif sampling detection of bacterial outer membrane protein repeats.** *Prot Sci* 1995, **4**:1618-1632.
31. Matthews B: **Comparison of the predicted and observed secondary structure of T4 phage lysozyme.** *Biochim Biophys Acta* 1975, **405**:442-451.
32. Reche PA, Glutting J-P, Reinherz EL: **Enhancement to the RANKPEP resource for the prediction of peptide binding to MHC molecules using profiles.** *Immunogenetics* 2004, **56**:405-419.
33. Lafuente EM, Reche PA: **Prediction of MHC-peptide binding: a systematic and comprehensive overview.** *Curr Pharm Des* 2009, **15(28)**:3209-3220.
34. Reche PA, Glutting JP, Reinherz EL: **Prediction of MHC class I binding peptides using profile motifs.** *Hum Immunol* 2002, **63(9)**:701-709.

35. Reche PA, Reinherz EL: **Prediction of peptide-MHC binding using profiles.** *Methods Mol Biol* 2007, **409**:185-200.
36. Swets JA: **Measuring the accuracy of diagnostic systems.** *Science* 1988, **240(4857)**:1285-1293.
37. Dönnes P, Kohlbacher O: **Integrated modeling of the major events in the MHC class I antigen processing pathway.** *Protein Science* 2005, **14(8)**:2132-2140.
38. Rosenfeld : **Two decades of statistical language modeling: Where do we go from here?** *Proceedings of the IEEE* 2000, **88(8)**:1-11.
39. Jimenez-Montano MA, Ebeling W, Pohl T, Rapp PE: **Entropy and complexity of finite sequences as fluctuating quantities.** *Biosystems* 2002, **64(1-3)**:23-32.
40. Wu C, Shivakumar S: **Back-propagation and counter-propagation neural networks for phylogenetic classification of ribosomal RNA sequences.** *Nucleic Acids Res* 1994, **22(20)**:4291-4299.
41. Wu CH, Zhao S, Chen HL, Lo CJ, McLarty J: **Motif identification neural design for rapid and sensitive protein family search.** *Comput Appl Biosci* 1996, **12(2)**:109-118.
42. Kloetzel PM: **Generation of major histocompatibility complex class I antigens: functional interplay between proteasomes and TPPII.** *Nat Immunol* 2004, **5(7)**:661-669.
43. Reits E, Neijssen J, Herberts C, Benckhuijsen W, Janssen L, Drijfhout JW, Neefjes J: **A major role for TPPII in trimming proteasomal degradation products for MHC class I antigen presentation.** *Immunity* 2004, **20(4)**:495-506.
44. Yewdell JW, Princiotta MF: **Proteasomes get by with lots of help from their friends.** *Immunity* 2004, **20(4)**:362-363.
45. Heath W, Belz T, Behrens G, Smith C, Forehan S, Parish I, Davey G, Wilson N, Carbone F, Villadangos J: **Cross-presentation, dendritic cell subsets, and the generation of immunity to cellular antigens.** *Immunological Reviews* 2004, **199(1)**:9-26.
46. Banchereau J, Briere F, Caux C, Davoust J, Lebecque S, Liu Y, Pulendran B, Palucka K: **Immunobiology of dendritic cells.** *Annu Rev Immunol* 2000, **18**:767-811.
47. Kisselev AF, Akopian TN, Woo KM, Goldberg AL: **The sizes of peptides generated from protein by mammalian 26 and 20 S proteasomes. Implications for understanding the degradative mechanism and antigen presentation.** *J Biol Chem* 1999, **274(6)**:3363-3371.
48. 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(6)**:581-591.
49. Ginodi I, Vider-Shalit T, Tsaban L, Louzoun Y: **Precise score for the prediction of peptides cleaved by the proteasome.** *Bioinformatics* 2008, **24(4)**:477-483, Epub 2008 Jan 2023.
50. Wang P, Sidney J, Dow C, Mothe B, Sette A, Peters B: **A systematic assessment of MHC class II peptide binding predictions and evaluation of a consensus approach.** *PLoS Comput Biol* 2008, **4(4)**:e1000048.
51. Doytchinova IA, Flower CR: **Class I T-cell epitope prediction: Improvements using a combination of proteasome cleavage, TAP affinity, and MHC binding.** *Molecular Immunology* 2006, **43(13)**:2037-2044.
52. Larsen MV, Lundegaard C, Lamberth K, Buus S, Brunak S, Lund O, Nielsen M: **An integrative approach to CTL epitope prediction: a combined algorithm integrating MHC class I binding, TAP transport efficiency, and proteasomal cleavage predictions.** *Eur J Immunol* 2005, **35(8)**:2295-2303.

doi:10.1186/1471-2105-11-479
Cite this article as: Diez-Rivero *et al.*: Computational analysis and modeling of cleavage by the immunoproteasome and the constitutive proteasome. *BMC Bioinformatics* 2010 **11**:479.

6. CAPÍTULO III

Modelado de la afinidad de unión de péptidos a TAP

6.1 Justificación y Objetivos

El transporte de los péptidos al interior del retículo endoplasmático por el transportador asociado al procesamiento de antígenos (TAP) es otro de los pasos esenciales del procesamiento de epítopos de células T CD8. En este capítulo se realiza un estudio de la capacidad para predecir péptidos que sean transportados por TAP de distintos modelos basados todos ellos en "*support vector machine*" (SVM) y entrenados con cada una de las posiciones de los residuos del péptido solas o combinadas.

Los objetivos de este capítulo son:

- Definir cuantitativamente las posiciones que contribuyen a la unión del péptido a TAP.
- Partiendo de un conjunto de péptidos cuya unión a TAP es conocida, desarrollamos varios modelos para predecir la afinidad de unión de péptidos a TAP, entrenando dichos modelos con cada residuo individualmente, con la combinación de distintas posiciones y utilizando la secuencia del péptido completa.
- Desarrollo de una herramienta web.

6.2 Conclusiones

- La evaluación de los modelos desarrollados con cada una de las posiciones del péptido individualmente muestra que todas ellas (P1-P9) contribuyen a la unión a TAP, a juzgar por el coeficiente de correlación de Pearson (R_p), aunque la principal contribución para la unión del péptido a TAP corresponde al residuo C-terminal.
- Cuando se entrenan los modelos con combinaciones de residuos del péptido, generalmente, se mejora la capacidad predictiva, alcanzando una correlación máxima cuando el modelo se entrena con la secuencia completa del péptido o con una

selección de residuos consistente en los primeros 5N- y los últimos 3 C-terminales de los péptidos.

- De la combinación de residuos del péptido podemos ver que la mitad del extremo N-terminal del péptido contribuye más a la unión a TAP que la mitad del extremo C-terminal del péptido.
- El modelo para la predicción de la afinidad de unión de los péptidos a TAP usando los modelos desarrollados está implementado en la herramienta TAPREG, disponible para uso público en http://imed.med.ucm.es/Tools/tapreg/.

PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Quantitative modeling of peptide binding to TAP using support vector machine

Carmen M. Diez-Rivero,[1] Bernardo Chenlo,[1] Pilar Zuluaga,[2] and Pedro A. Reche[1]*

[1] Laboratorio de InmunoMedicina, Departamento de Microbiología I-Immunología, Facultad de Medicina, Universidad Complutense, Madrid 28040, Spain

[2] Departamento de Estadística e Investigación Operativa, Facultad de Medicina, Universidad Complutense, Madrid 28040, Spain

ABSTRACT

The transport of peptides to the endoplasmic reticulum by the transporter associated with antigen processing (TAP) is a necessary step towards determining CD8 T cell epitopes. In this work, we have studied the predictive performance of support vector machine models trained on single residue positions and residue combinations drawn from a large dataset consisting of 613 nonamer peptides of known affinity to TAP. Predictive performance of these TAP affinity models was evaluated under 10-fold cross-validation experiments and measured using Pearson's correlation coefficients (R_p). Our results show that every peptide position (P1–P9) contributes to TAP binding (minimum R_p of 0.26 ± 0.11 was achieved by a model trained on the P6 residue), although the largest contributions to binding correspond to the C-terminal end (R_p = 0.68 ± 0.06) and the P1 (R_p = 0.51 ± 0.09) and P2 (0.57 ± 0.08) residues of the peptide. Training the models on additional peptide residues generally improved their predictive performance and a maximum correlation (R_p = 0.89 ± 0.03) was achieved by a model trained on the full-length sequences or a residue selection consisting of the first 5 N- and last 3 C-terminal residues of the peptides included in the training set. A system for predicting the binding affinity of peptides to TAP using the methods described here is readily available for free public use at http://imed.med.ucm.es/Tools/tapreg/.

Proteins 2009; 00:000–000.

Key words: antigen processing; peptide; TAP; prediction; WEKA; SVM.

INTRODUCTION

CD8 T cells play a key role in tumor immunosurveillance and clearing of intracellular infectious agents, and a subset of them known as cytotoxic T lymphocytes (CTLs) are capable of directly killing infected and tumor cells.[1] CTLs discriminate between normal and damaged cells using their T cell receptor (TCR) to monitor the peptides presented by major histocompatibility class I (MHCI) molecules on the cell surface. T cells recognizing self-peptides are eliminated during the process of thymic selection, and, thereby, T cell immune responses are triggered by the recognition of MHC molecules incorporating foreign or antigenic peptides (T cell epitopes).[2] T cell epitopes result from the degradation of proteins through pathways that determine the repertoire of peptides that are available for binding to MHC and recognition by T cells. The dominant pathway for class I antigen processing is reviewed next.

MHCI molecules preferably bind peptides nine residues long that generally originate from endogenous proteins that are degraded in the cytosol of the cell by the proteolytic activity of the proteasome.[3,4] Peptide fragments cleaved by proteasomes are shuttled to the lumen of the endoplasmic reticulum (ER) by the transporter associated with antigen processing (TAP), where they can bind to newly assembling MHCI molecules.[5,6] Before MHCI binding, peptides can also undergo an optional N-terminal trimming by ER-associated amino peptidases (ERAAP).[7] Finally, peptide-MHCI complexes are exported to the cell surface for presentation to the CD8 T cells.[5,6] There is evidence supporting that these processing steps limit/shape the peptides that can be presented by MHCI molecules *in vivo*,[7-9] thus explaining the numerous observations of high affinity MHCI binding peptides that are unable to elicit CTL responses.[10,11] Nonetheless, peptide transport by TAP represents the single most selective step in T cell epitope processing.[12] In addition, TAP is also important for presentation of epitopes derived from exogenous antigens.[13]

Additional Supporting Information may be found in the online version of this article.
The authors state no conflict of interest.
Carmen M. Diez-Rivero and Bernardo Chenlo contributed equally to this work.
Grant sponsor: Ministerio de Ciencia e Innovación (MICINN) of Spain; Grant number: SAF2006-07879; Grant sponsor: Universidad Complutense de Madrid (U.C.M); Grant number: CCG08-UCM/BIO-3769.
*Correspondence to: Pedro A. Reche, Laboratorio de InmunoMedicina, Departamento de Microbiología I-Immunología, Facultad de Medicina, Universidad Complutense, de Madrid, Ave. Complutense s/n, Madrid 28040, Spain. E-mail: parecheg@med.ucm.es
Received 8 April 2009; Revised 2 July 2009; Accepted 7 July 2009
Published online in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/prot.22535

TAP belongs to the ATP-dependent binding cassette (ABC) transporter superfamily, and it is expressed as a heterodimer consisting of the TAP1 and TAP2 proteins subunits.[14,15] Both TAP1 and TAP2 proteins encode one hydrophobic transmembrane domain and one ATP binding domain. Transport of peptides by TAP proceeds in two sequential steps, where peptide binding to TAP occurs first followed by a translocation step consuming ATP.[16-18] Peptide transport rate by TAP is governed by the initial binding step.[19,20] Likewise, TAP preselection of peptides available for MHCI presentation is also controlled by their affinity to TAP. Selectivity of TAP has been studied from data generated using assays that determine peptide binding to TAP or peptide accumulation in the ER.[17,18] TAP preferentially transports peptides with a length of 8–16 residues,[14,21] whereas longer peptides may be transported but with much lower efficiency. Besides peptide length preferences, the first three N-terminal residues and the C-terminal end of the peptides have also been shown to be important for binding to TAP.[12,22] Furthermore, a peptide-binding motif for TAP has been defined by van Endert et al.,[22] which indicates a TAP preference for hydrophobic aromatic residues at the C-terminus, hydrophobic residues at position 3 (P3), and charged and hydrophobic residues at position 2 (P2). On the other end, aromatic or acidic residues at P1 and prolines at P1 and P2 have strong deleterious effects.

A number of methods have also been applied for predicting and analyzing the binding affinity of peptides to TAP, such as artificial neural networks,[23-25] support vector machines (SVMs),[26,27] and matrices generated using the Stabilized Matrix Method[28] and the additive method.[29,30] The majority of these methods were trained on the same training set of ~435 nonamer (9-mer) peptides of known affinity to TAP made available by Dr. van Endert, and until now their performance has not been compared in an independent testing set. In contrast, here we have used a much larger training set, encompassing 178 new peptides, to analyze TAP binding preferences using SVMs. Interestingly, our results indicate that each peptide residue has a significant contribution to TAP binding. Moreover, we have generated TAP binding affinity models that in cross-validation experiments achieved a correlation between experimental and predicted values of 0.89 ± 0.03, which is stronger than that of related methods. Based on these results, we have implemented a system, TAPREG, for predicting affinity of peptides to TAP that is available for free public use at http://imed.med.ucm.es/Tools/tapreg/.

MATERIAL AND METHODS

Peptide datasets

The main dataset used in this study to analyze the peptide selectivity of TAP consisted of 613 unique nonamer (9-mer) peptides of known binding affinity to human TAP relative to the reference peptide RRYNASTEL ($IC_{50relative}$). The lower the $IC_{50relative}$, the stronger the peptide binds to TAP. This dataset encompasses 435 peptides, kindly provided by Dr. Peter van Endert[23] (INSERM U580, Paris Descartes University, Paris, France)—$IC_{50relative}$ already referenced to RRYNASTEL—plus 178 peptides parsed from the TAP binding affinity peptide collection of the Antijen Database,[31] kindly provided by Dr. Darren Flower (The Jenner Institute, Compton, UK). To combine the peptides into a single dataset, the TAP binding affinity (IC_{50}) of peptides collected from the Antigen Database was also referenced to the peptide RRYNASTEL. For peptides obtained from the Antigen Database that were identical in sequence but had different TAP binding affinities, median values were considered before referencing. This dataset is provided as Supporting Information in Table 1S. We thank to Dr. Peter van Endert and Dr. Darren Flower for showing no inconvenience in that we provided Table 1S as Supporting Information.

Peptide datasets with reduced sequence similarity were generated from the 613-peptide dataset using the purge utility of the Gibbs Sampler[32] with an exhaustive method and maximum blosum 62 relatedness scores of 25, 30, 35, and 37. The resulting datasets had 293, 332, 465, and 530 peptides and are provided as Supporting Information (Table 2S, Table 3S, Table 4S, and Table 5S, respectively).

To compare TAP affinity scores predicted by available methods, we used a set of 723 unique 9-mer CD8 T cell epitopes obtained from the IMMUNEEPITOPE[33] and EPIMHC[34] databases (provided as Supporting Information in Table 6S).

Model building and evaluation

Predictive models of TAP affinity were trained and evaluated under the EXPERIMETER application of the Waikato Environment for Knowledge Analysis (WEKA) package.[35] WEKA provides a framework for data classification, clustering, and feature selection using a large collection of machine-learning algorithms. In this study, we have selected kernel-based SVMs. Specifically, we used a radial basis function (RBF) as the kernel in combination with Alex Smola and Bernhard Scholkopf's sequential minimal optimization algorithm for training SVMs (SMOreg algorithm in WEKA).[36,37] Model refinement was achieved by varying the C (0.2, 0.4, 0.8, 1, 2, 4, 8, 10) and gamma (0.001, 0.0025, 0.005, 0.01, 0.025, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5) values of the RBF kernel. Predictive models were generated from distinct training sets, consisting of different residue selections drawn from the peptide sequences of the training set and encoded using sparse and blosum representations. In the sparse encoding, each amino acid is coded by the relevant amino acid symbol, whereas in the blosum encoding, it is represented by 20 digits corresponding to the relevant amino

acid substitution scores given by the BLOSUM62 substitution matrix.[38] TAP affinity ($IC_{50relative}$) values of the training sets were provided to WEKA as $logIC_{50relative}$ values. Pearson's correlation coefficient (R_p) was used to measure the performance of SVMs to fit the experimental data. Since SVM models were built and evaluated using 10-fold cross-validation experiments that were repeated 10 times, R_p mean values and standard deviations were computed from 100 different values. Predicted peptide affinity scores yielded by the models generated with WEKA were transformed to IC_{50} values by considering an IC_{50} for the reference peptide RRYNASTEL of 400 n*M*.

Sequence similarity analyses

Sequence similarity in peptide datasets was analyzed from pairwise sequence alignments between all peptides in the dataset. Sequence alignments were obtained using the Needleman-Wunsch global alignment algorithm implemented with the needle application that is included in the EMBOSS package.[39] Alignments with peptide positions shifted were not evaluated (e.g., residues 1–4 of a peptide aligned with residues 3–7 of another peptide). Generally, for any given peptide (query) in the dataset, one could find several peptides that shared sequence similarity with it (hits), but the majority of the peptides in the dataset had no similarity with the query. In this study, we have computed average sequence similarities in the peptide datasets in two ways: globally, considering all possible pairwise comparisons between the peptide sequences but those with themselves (for a dataset with *N* peptides there will be $N \times N\text{-}1$ comparisons), and using only the hits.

For a given query peptide in the dataset, the relationship between sequence similarity and binding affinity was studied by correlating sequence similarity with hits and differences in binding affinity ($logIC_{50relative}$) using Spearman's rank correlation (R_s). For instance, let us consider the peptide PLAKAAAAV ($logIC_{50relative} = 8.370$) had the following hits:

Hit:ALAKAAAAV; Identity:88.9%; Similarity:88.9%; $logIC_{50relative}$:3.984; Dif:4.386

Hit:ALAKAAAAL; Identity:77.8%; Similarity:88.9%; $logIC_{50relative}$:0.688; Dif:7.682

Hit:AAASAAAAF; Identity:66.7%; Similarity:77.8%; $logIC_{50relative}$:−0.734; Dif:9.104

Hit:ALAKAAAAF; Identity:55.6%; Similarity:66.7%; $logIC_{50relative}$:0.332; Dif:8.038

Hit:GRQKGAGSV; Identity:33.3%; Similarity:44.4%; $logIC_{50relative}$:6.215; Dif:2.155

Then, for peptide PLAKAAAAV, an R_s value was computed by correlating the similarity/identity with its peptide hits (88.9, 77.8, 66.7, 55.6, 33.3) and the differences in $logIC_{50relative}$ values (4.386, 7.682, 9.104, 8.038, 2.155). R_s values were thus computed for each peptide in the dataset. Peptides with less than five hits were discarded from this analysis. These peptide-specific R_s values were determined considering all peptide hits and only those with an identity $\geq$ 50%.

Statistical analyses

To assess whether the correlation achieved by a given SVM model, *i*, during training was stronger than that of another SVM model, *j*, we used one-sided two-sample *t*-test to examine if the differences of the relevant R_p mean values were significantly above 0 (Ho: $R_pi - R_pj = 0$; $P \leq 0.05$). To evaluate if R_p values were statistically significant (H_0: $R_p = 0$), we computed the statistics given by Eq. (1), which follows a *t*-Student distribution with $N - 2$ degrees of freedom, and tested subsequently ($P < 0.05$).

$$t = \frac{R_p}{\sqrt{\frac{1-Rp^2}{N-2}}} \tag{1}$$

To evaluate the correlation coefficients obtaining by comparing the TAP affinity scores predicted by different methods with each other or with experimental data, we applied the test for comparing overlapping correlation coefficients described by Meng et al.,[40] as implemented in the *R* package *compOverlapCorr* by Ka-Lon Li (http://cran.us.r-project.org/web/packages/compOverlapCorr/index.html). Briefly, Fisher's Z-transform is applied first to the relevant correlation coefficients (R_i) using Eq. (2).

$$Z_i = \frac{1}{2}\ln\left(\frac{1+R_i}{1-R_i}\right) \tag{2}$$

Next, a statistics Z, which follows a normal distribution is computed using Eq. (3), and tested subsequently ($P < 0.05$).

$$Z = (z_i - z_j)\sqrt{\frac{N-3}{2(1-R_{ij})h}} \tag{3}$$

In Eq. (3), R_{ij} is the correlation between the predicted values by the methods *i* and *j* being compared, and $h = (1 - f\overline{R}^2)/(1 - \overline{R}^2)$, with $\overline{R}^2 = (R^2_i + R^2_j)/2$ and $f = (1 - R_{ij})/2(1 - \overline{R}^2)$.

Web server implementation

The TAPREG Web server for predicting the binding affinity of peptides to TAP was implemented on an Apache Web server under the Mac OSX operating system. The TAPREG core consists of a PERL CGI (Common Gateway Interface) script that executes the predictions on

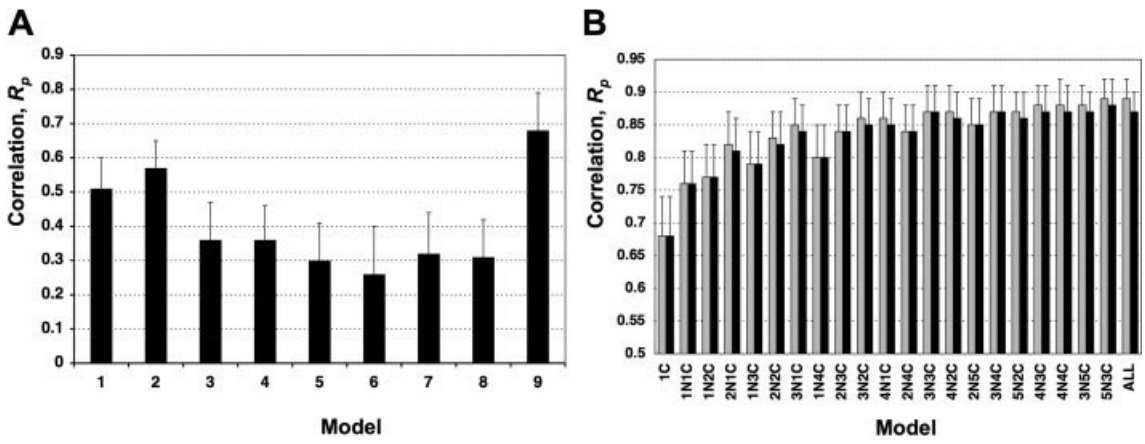


Figure 1

Performance of TAP-affinity prediction models. Models were trained using SVM and their performance was measured using R_p values between predictions and experimental values determined under 10-fold cross-validation experiments that were repeated 10 times. Thus, R_p mean values and standard deviations obtained over 100 measures are represented in the figure. Moreover, plotted R_p values were those achieved by SVMs after parameter optimization. (**A**) Performance of models trained on individual residues of the 9-mer peptides (1–9) included in the training set. (**B**) Performance of models trained on different peptide fragments consisting of the first *i* N-terminal and the last *j* C-terminal residues of the peptides in the training set. Residue selections, *iNjC* are indicated in the abscissa. Grey bars are for SVM models trained on sparse sequence representations and black bars for models trained using blosum sequence representations. There was no difference between sparse and blosum trained models on single peptide residues. Data for making these representations—including the relevant RBF parameters of SVMs—are provided as Supporting Informtion in Table 7S.

user-provided input data and returns the results to the browser. In addition, the TAPREG web interface uses JavaScript for handling and verification of input data before submission.

RESULTS

Quantitative analysis of TAP selectivity using TAP affinity models

We have approached the study of TAP selectivity using a large dataset consisting of 613 9-mer peptides (DS_{613}) of known affinity to TAP ($logIC_{50relative}$) and SVMs under a regression schema. SVMs are among the most widely used methods for solving common data mining problems in bioinformatics[41-43] and were chosen because of their solid theoretical foundations and proven generalization ability.[44] A key feature of SVMs is the use of nonlinear functions (kernels) to map the input onto a higher dimensional space in which an optimal separation is achieved—in the regression task—using a linear regression conducted with an ϵ-insensitive loss function for error minimization.[44] In this study, we have selected RBF kernels (Material and Methods) because in preliminary training experiments they outperformed the alternative linear and polynomial kernels (data not shown). Moreover, we have chosen two peptide sequence representations, sparse and blosum (Material and Methods), as input for SVMs. The evolutionary relationships between amino acids are taken into consideration with blosum representations of peptide sequences, which may enhance the generalization power of the resulting models. Using WEKA as the framework for model building and parameter optimization (Material and Methods), we first evaluated the ability of SVM models to predict TAP affinity data when trained on individual peptide residues (P1–P9), judging from the relevant Pearson's correlation coefficient (R_p). No differences were observed for models generated on blosum or sparse encoded sequences. Interestingly, for each peptide residue position, it was possible to generate SVM models that fitted the data with R_p values [Fig. 1(A)] that are significant for a linear correlation ($P \leq 0.05$, Material and Methods). The lowest correlation was obtained with a model trained on the P6 residue (R_p of 0.26 ± 0.11), whereas the largest correlation corresponded to a model trained on the C-terminal end of the peptide ($R_p = 0.68 \pm 0.06$) followed by the models trained on the P2 (0.56 ± 0.08) and the P1 ($R_p = 0.51 \pm 0.09$) residues of the peptide. Systematic pairwise comparisons between the predictive performance of the different position-specific TAP affinity models using one-side *t*-tests over the relevant R_p means (Material and Methods) showed the following peptide residue position relevance to TAP binding: $(P6 = P5) < (P8 = P7) \leq (P3 = P4) \leq P1 \leq P2 \leq P9$ (C-terminal end).

To evaluate the contribution of several peptide residues to TAP binding and to improve the correlation results, SVMs were trained on peptide fragments consisting of residue combinations drawn from the peptides of the training set. A total of 20 SVM models were generated

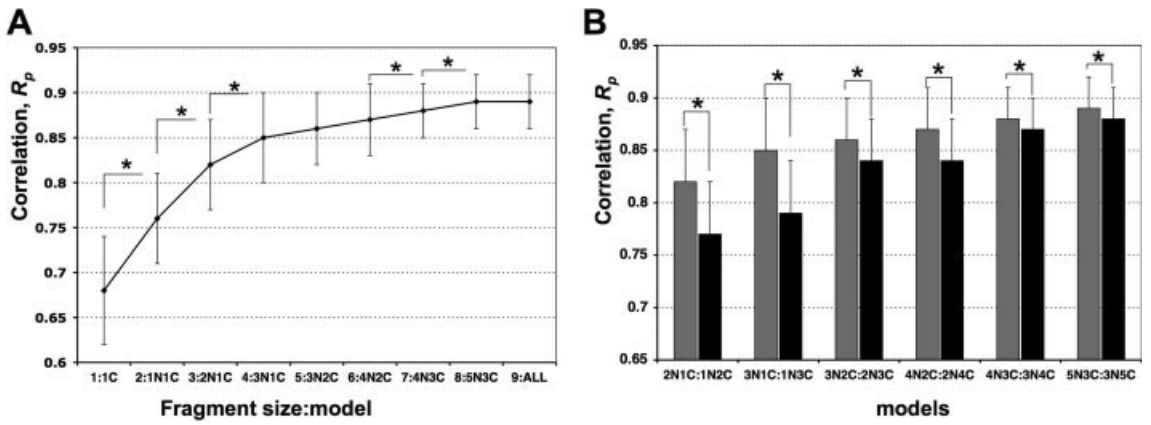


Figure 2

Analysis of TAP selectivity using TAP-affinity prediction models. SVM-Models trained using sparse sequence representation were selected. (**A**) Predictive performance (R_p) of SVM-models with regard to the fragment size used for training (1–9). Only the largest R_p value achieved by a specific model (indicated in the abscissa) at each fragment size is represented. Statistically significant increments between R_p values of neighboring models are indicated with a "*" symbol. (**B**) Predictive performance of the best SVM-models generated upon optimal first *i* N- and last *j* C-terminal residue selections (gray bars) compared with those generated from suboptimal first *j* N- and last *i* C-terminal residue selections (black bars). Statistically significant differences were found between R_p values in all cases (indicated with a "*" symbol). Statistical significance was assessed using *t*-tests (Material and Methods).

and named after the specific peptide residue selection used for training (model *iNjC* was generated from a fragment of $i + j$ residues, consisting of the first *i* N-terminal and last *j* C-terminal residues of the peptides of the training set). R_p values achieved by these models on the training set together with those achieved by the models trained on just the C-terminus and the full-length peptide sequences (9-mers) are shown in Figure 1(B). Few or no differences were observed between SVMs trained using different sequence representations: sparse [gray bars in Fig. 1(B)] and blosum [black bars in Fig. 1(B)]. However, when differences were found, correlations obtained with the models trained on sparse encoded sequences were always larger than their blosum counterparts and were significantly stronger ($P \leq 0.05$) for models *3N2C, 4N1C, 4N2C, 5N2C, 4N3C, 4N4C, 3N5C, 5N3C,* and *ALL* (trained on the full-length sequences). Several other general features emerged upon a detailed analysis of these results. Increasing the number of selected residues in the training sets (drawn from the peptides of known affinity to TAP) significantly improved the correlations achieved by the models [Fig. 2(A)], which went from an R_p value of 0.68 ± 0.06 for a model trained on just the C-terminal end of the peptides of the training set to an R_p of 0.89 ± 0.03 for the model trained on the full-length sequences (nonamers). Interestingly, a model trained on just eight residues (*5N3C*) achieved the same or better correlation (for blosum encoding) than models trained on the full-length peptide sequences [Figs. 1(B) and 2]. Nevertheless, for each fragment size, the best correlations were obtained with models trained on fragments encompassing more N-terminal than C-terminal peptide residue selections (*2N1C, 3N1C, 4N2C, 4N3C,* and *5N3C)* [Fig. 2(A)], and these correlations were significantly stronger ($P \leq 0.05$) than those obtained with models with reversed N-terminal and C-terminal residue selections (*1N2C, 1N3C, 2N4C, 3N4C,* and *3N5C*) [Fig. 2(B)]. This observation supports a larger contribution of the N-terminal half of the peptide to TAP binding when compared with its C-terminal half.

Sequence similarity in peptide datasets and predictive performance of SVM models

To explore the predictive performance of SVM models in relation to the sequence similarity between testing and training sets, we generated four peptide datasets of 293, 332, 465, and 530 peptides (DS_{293}, DS_{332}, DS_{465}, DS_{530}, respectively) by discarding similar sequences from the original DS_{613} dataset (Material and Methods). The global sequence identity in percentage in these datasets varied from 1 ± 6% in the DS_{293} dataset to 9 ± 23% in the DS_{530} dataset, whereas in the DS_{613} dataset it was 10 ± 25% (Table I). In the 435-peptide dataset provided by Peter van Endert (PVE_{435}) the global identity is 5 ± 16%. The overall low sequence similarity in the datasets reflects that the peptides do not belong to a single class or group related by a given property. On the contrary, each peptide is linked to a different numeric value ($logIC_{50relative}$). The average number of similarity hits per peptide in the datasets varied from nine peptides in the DS_{293} dataset to 110 hits in the DS_{613} dataset (Table I). Sequence identity between hits was considerably larger

Table I
Predictive Performance of SVMs Trained on Datasets with Different Sequence Similarity

Dataset	R_p	Identity (%)[a]	Similarity (%)[a]	Identity (%)[b]	Similarity (%)[b]	Hits[c]
DS_{293}	0.71 ± 0.1	1 ± 6	2 ± 10	23 ± 11	43 ± 11	9 ± 7
DS_{332}	0.76 ± 0.09	2 ± 8	3 ± 11	28 ± 18	46 ± 14	14 ± 12
DS_{465}	0.85 ± 0.05	7 ± 19	8 ± 21	52 ± 25	60 ± 19	59 ± 45
DS_{530}	0.87 ± 0.03	9 ± 23	10 ± 25	57 ± 24	62 ± 26	86 ± 62
DS_{613}	0.89 ± 0.03	10 ± 25	11 ± 26	59 ± 23	66 ± 18	110 ± 77
PVE_{435}	0.83 ± 0.05	5 ± 16	6 ± 18	45 ± 26	56 ± 19	40 ± 33

[a]Identity and similarity computed considering all possible pairwise comparisons between the peptides in the datasets.
[b]Identity and similarity computed considering only hits (Material and Methods).
[c]Average number of similarity hits per peptide in the dataset.

and ranged from 23% in the DS_{293} dataset to 59% in the DS_{613} dataset (Table I).

Because we train and evaluate the predictive performance of SMVs using 10-fold cross-validation experiments, and we repeat these experiments 10 times. we can assume that sequence similarity between testing and training sets to be comparable to that in the entire datasets. The correlation between predictions and experimental $\log IC_{50relative}$ values achieved by SVMs trained and evaluated on the datasets of reduced sequence similarity (DS_{293}, DS_{332}, DS_{465}, DS_{530}, and PVE_{435}) was significantly lower ($P \leq 0.05$; one-sided *t*-tests) than that obtained in the DS_{613} dataset (Table I). The smallest R_p was achieved in the DS_{293} dataset (0.71 ± 0.1), and these values increased significantly ($P \leq 0.05$) as the number of peptides in the datasets (Table I). Thus, $DS_{613}R_p > DS_{530}R_p > DS_{465}R_p > PVE_{435}R_p > DS_{332}R_p > DS_{293}R_p$.

These results may apparently suggest that prediction rates by our SVM models became inflated as sequence similarity in the datasets increased. However, this is an unlikely scenario because R_p values were computed in cross-validation, and the differences in R_p that we observed were statistically significant. For sequence similarity to be responsible for inflating prediction rates, the larger the sequence similarity between peptides in the datasets the closer their binding affinity must be. As a result, for any given peptide in the dataset one would expect to find a negative correlation between the similarity to its peptide hits and the differences in binding affinity (Material and Methods for details). However, we have not found such a negative correlation for the vast majority of the peptides in any of the datasets, as shown in the boxplot depicted in Figure 3. On the contrary, we have found these correlations to be shifted toward positives values; correlation medians in the DS_{293}, DS_{332}, DS_{465}, DS_{613}, and PVE_{435} datasets were 0.083, 0.109, 0.102, 0.139, 0.1945, and 0.114, respectively. Notably, the median of the correlation values in the DS_{613} dataset is significantly larger than those of the remaining datasets ($P \leq 0.05$), as judged from Wilcoxon-Mann-Whitney tests. Virtually identical results were obtained when only hits with ≥ 50% identity were considered (data not shown).

These results indicate that sequence similarity between peptides in the datasets does not correlate with proximity in binding affinity—in fact the opposite would appear to be the case. Therefore, the prediction rates obtained with SVMs trained on DS_{613} dataset are not inflated due to sequence similarity redundancy. Furthermore, similar sequences in the DS_{613} dataset are not redundant and contribute to the appropriated modeling of TAP binding affinity by SVMs; hence, the enhanced prediction rates achieved by models trained on the DS_{613} dataset.

Comparison of methods for predicting binding affinity of peptides to TAP

We have compared our SVM model trained on 9-mer peptide sequences that achieved an $R_p = 0.89 \pm 0.03$ (hereafter TAP_{613}) with four alternative predictive

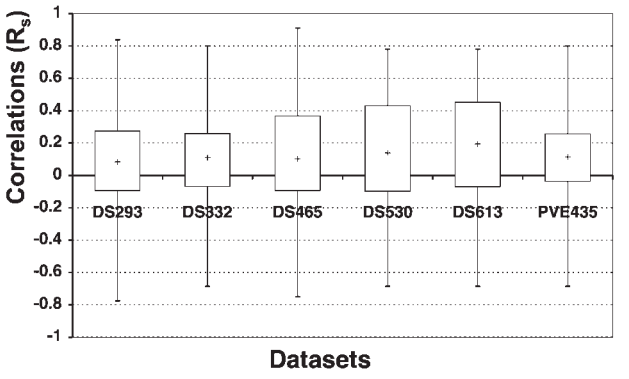


Figure 3
Relationship between sequence similarity in peptide datasets and binding affinity proximity. This figure depicts a boxplot of R_s values computed for each peptide in a dataset by correlating their identity with its hits and the difference in $\log IC_{50relative}$ values (Material and Methods). Boxplot were generated for peptides in DS_{293}, DS_{332}, DS_{465}, DS_{530}, DS_{613}, and PVE_{435} datasets. Median R_s values in peptide datasets are indicated with a cross. A negative R_s will indicate that the larger the sequence similarity between peptides the closer their binding affinity. Conversely, a positive correlation will reflect that the larger the sequence similarity between peptides the larger the difference in their binding affinity.

Table II
Correlation Between Experimental TAP Binding Affinities and Predicted Values Using Different Methods

Method	R_s	Reference
TAP_{613}	0.89 ± 0.03	This study
SMM	0.87 (0.82)	28
ADM	0.74 (0.72–0.83)	29
TAPPRED	0.67 (0.88)	26
SVMTAP	0.61 (0.82)	27

R_s were computed using a testing set of 178 peptides of known affinity to TAP. For the TAP_{613} model, R_s shown in the table is that achieved in cross-validation. Correlations reported in the literature for the different methods are shown in parentheses.

methods of peptide binding affinity to TAP, which are readily available from the relevant publications (those by Peters et al.[28] and Doytchinova et al.[29]) or from dedicated Web services (TAPPRED[26] and SVMTAP[27]). The method developed by Doytchinova et al.[29] consists of a matrix generated from 163 poly-Alanine 9-mer peptides of known affinity to TAP using an additive method[30]; hence, we will refer to this method as ADM. The ADM method achieved a reported R_p between 0.72 and 0.83, depending of the testing set.[29] The remaining methods have been trained on the PVE_{435} dataset.[28] Briefly, Peters' et al.[28] method is based on a consensus matrix (CM) that was obtained from three scoring matrices, which included a poly-Alanine derived matrix and a SMM-matrix (generated using the Stabilized Matrix Method) trained on the PVE_{435} dataset. The CM method achieved a reported R_p of 0.782 on the PVE_{435} dataset. The TAPPRED[26] and SVMTAP[27] methods are based on SVMs trained solely on the PVE_{435} dataset and achieved reported R_p of 0.82 and 0.88, respectively. The TAPPRED method is based on two layers of SVMs, whereas SVMTAP consists of a single SVM model, similar to those trained in this study. We have evaluated all these methods in a testing set consisting of the 178 peptides of known affinity to TAP collected in this study (DS_{178}), using Spearman's correlation coefficients (R_s) (Table II). Interestingly, the lowest R_s values were achieved by TAPPRED and SVMTAP (0.67 and 0.61), the methods with the largest reported correlations. On the other hand, CM achieved an R_s (0.87) comparable to the value achieved by our TAP_{613} model in cross-validation (0.89), and AMD achieved an intermediate R_s value of 0.74. Statistical comparison of these R_s values (Material and Methods) indicated that the correlations obtained with the CM and TAP_{613} methods were significantly stronger than those obtained with the remaining methods. However, TAP_{613} was also trained on the DS_{178} testing set used for the comparisons, as surely were both the CM and ADM methods (DS_{178} contains binding affinity data of poly-Alanine peptides).

To further compare these methods, we have used a reference set of 723 MHCI-restricted T cell epitopes and correlated the scores predicted by the different methods (Table III). Interestingly, TAP_{613} predictions were significantly closer to the predictions by CM (R_s = 0.86), a matrix-based method, than to those by TAPRED (0.29) and SVMTAP (0.76), which are based on SVM. Likewise, ADM predictions also correlated better with TAP_{613} predictions (0.59) than with those by TAPPRED (0.17) and SVMTAP (0.51). The extreme disparity of TAPPRED predictions with regard to the remaining methods was already noted by Zhang et al.[25] Overall, these results support the view that existing SVM-based methods (TAPPRED and SVM) have suffered to some extent from data over-fitting, particularly TAPPRED, while we do not expect such a problem with our TAP_{613} model, as it was trained on a much larger dataset.

The TAPREG server

We have implemented a Web tool, TAPREG, for predicting the binding affinity of peptides to TAP, which is available for free public use at http://imed.med.ucm.es/Tools/tapreg/ [Fig. 4(A)]. There are two models available at the TAPREG site that were trained both on the DS_{613} dataset using the entire peptide sequences; one was generated from a sparse representation of peptide sequences and the other from a blosum representation. The model trained on blosum-encoded sequences displayed a somewhat lower predictive performance (R_p = 0.87 ± 0.03) than the sparse counterpart (R_p = 0.89 ± 0.03), but nonetheless, it is included in the TAPREG server because blosum representation of sequences can often increase the generalization power of predictive models. The input data for TAPREG can consist of either protein sequences or multiple peptide sequences. For the protein sequence, TAPREG returns all 9-mer peptides encompassed by the protein, ranked by their affinity to TAP (IC_{50}). The number of peptides listed in the output can also be limited using a user-defined threshold of binding affinity [Fig. 4(B)]. For the peptide input, the server returns the affinity of each individual peptide [Fig. 4(C)]. As TAP can bind and transport peptides of arbitrary length ranging from eight to 16 residues,[14,21] TAPREG will predict the affinity of any peptide within that length range as described below.

Table III
Correlation Between TAP Binding Affinity Predictions by Different Methods

	CM	TAP_{613}	TAPPRED	ADM	SVMTAP
CM	1	0.86	0.26	0.84	0.68
TAP_{613}	0.86	1	0.29	0.59	0.76
ADM	0.84	0.59	0.17	1	0.51
TAPPRED	0.26	0.29	1	0.17	0.34
SVMTAP	0.68	0.76	0.34	0.51	1

Table shows R_s values that were obtained by correlating the TAP binding affinity scores of 723 MHCI-restricted T cell epitopes predicted with the different methods.

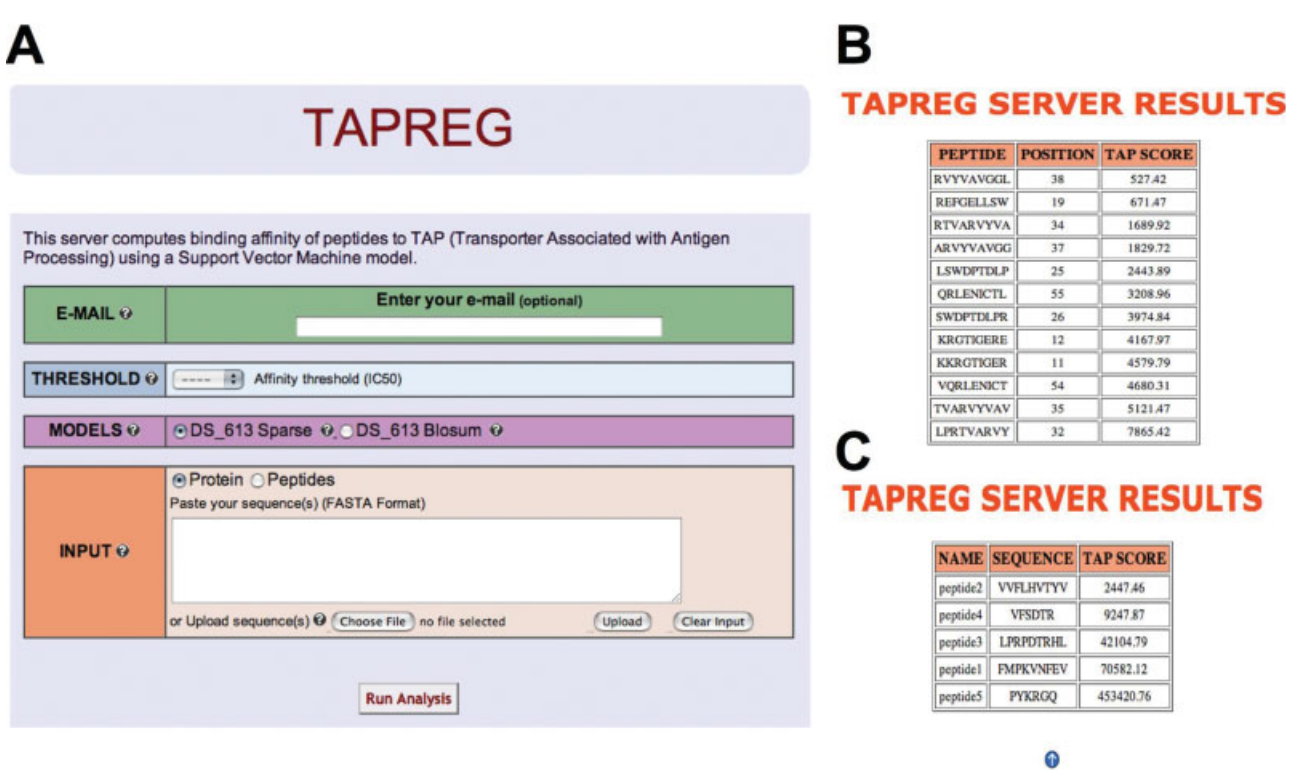


Figure 4

TAPREG server for predicting peptide binding affinity to TAP. (**A**) TAPREG Web interface. TAPREG can take two types of input data consisting of either multiple peptides in FASTA format (size 8 to 16 allowed) or a protein sequence in FASTA format. For protein sequences, TAPREG computes the TAP affinity of all 9-mer peptides in the protein and returns the peptides sorted by their affinity (IC_{50}) (Panel **B**). When multiple peptides are submitted, the program returns the binding affinity to TAP (IC_{50}) of each peptide (Panel **C**).

In general, models generated using machine-learning algorithms require input data of the same format as the data used for training. Therefore, in TAPREG, we have implemented a system to predict the TAP binding affinity of any peptide longer than nine residues, for example, ALRQFDSMERDNAVFL, by applying the model to a peptide fragment encompassing the first five N-terminal and last four C-terminal residues of the longer peptide; in this example, ALRQFAVFL. For peptides of eight residues, for example AVDFSDRS, we simply insert an Alanine at P6, AVDFSADRS, and then predict the binding affinity. Note that the P6 residue had the lower contribution to TAP binding [Fig. 1(A)]. Using the *5N3C* model, which achieved the same correlation as the TAP_{613} model that was trained on the entire 9-mer peptides (Fig. 2), the binding of any peptide longer than eight residues could be predicted by applying the model to a derivative fragment consisting of the first 5 N-terminal and last 3-C terminal residues.

DISCUSSION

The majority of TAP binding models have been derived from the same dataset consisting of ~435 9-mer peptides of known affinity which was made available by Dr. Peter van Endert[28] (PVE_{435}). In contrast, in this work, we have used a larger dataset of 613 peptides (DS_{613})—encompassing 178 new extra peptides—to study TAP selectivity quantitatively, using SVM regression models that were trained on single residue and residue combinations drawn from the peptides in the dataset. Thus, we have been able to recognize that each peptide position has a significant contribution to TAP binding, and that the contribution of the P4 residue is equivalent to that of the P3 residue [Fig. 1(A)]. Previously, only the positions P1, P2, P3, and the C-terminal end of the peptide were thought to be clearly relevant for binding to TAP.[12,22,26,28,29] We have confirmed that the C-terminal end of the peptide has the largest quantitative input to TAP binding; a model trained on this residue alone reached an $R_p = 0.68 \pm 0.06$. Nonetheless, we have shown that the N-terminal half of the peptide has a larger contribution to TAP binding than the C-terminal half of the peptide, as judged by the predictive performance of SMVs trained on peptide fragments encompassing a varying number of N-terminal and C-terminal residues of the peptides in the DS_{613} dataset (Fig. 2).

Optimal modeling of the binding affinity of peptides in the DS_{613} dataset was achieved by SVM models trained on the full-length peptide sequences (TAP_{613}) or on 8-residue fragments consisting of the first five N-terminal and last three C-terminal residues (*5N3C*) of the peptides ($R_p = 0.89 \pm 0.03$) [Figs. 1(B) and 2]. These results may reflect the observation that TAP can transport peptides of eight and nine residues with comparable efficiency.[14,21] Overall, that optimal fitting of TAP binding affinity data required training on multiple peptide residues also implies that all peptide residues—perhaps with the exception of the P6 residue—have a relevant contribution to TAP binding.

The correlation between predictions and experimental binding affinity values achieved by models TAP_{613} and *5N3C*, both trained on the DS_{613} dataset, is larger (0.89 ± 0.03) than that reported for any predictive model of TAP binding affinity.[26-29] It is worth noting that, unlike any of the related studies, we have not only evaluated the predictive performance of our models in cross-validation experiments but have also repeated the experiments 10 times and provided confidence values (standard deviations). Moreover, we have also shown that the enhanced predictive performance obtained with the model trained on the DS_{613} dataset is not related to sequence similarity redundancy (Fig. 3). In fact, we have found that peptides with high sequence similarity generally differ in their binding affinity (Fig. 3). Therefore, similar sequences are not redundant, and instead of inflating prediction rates, have a genuine contribution to model TAP binding affinity appropriately; hence, the enhanced prediction rates that we have obtained with the model trained in the DS_{613} dataset (Table I).

Using the new 178 peptides of known affinity to TAP collected in this study as a testing set (DS_{178} dataset), we have proved that two previous SVM-based methods (TAPPRED[26] and SMVTAP[27]) for predicting binding affinity of peptides to TAP, which were trained on the PVE_{435} dataset, appear to have suffered to some extent from data overfit; they achieved much lower correlation coefficients in the testing DS_{178} dataset than those reported on the PVE_{435} dataset (Table II). We have also evaluated two matrix-based methods, ADM[29] and CM,[28] on the same DS_{178} dataset, and they achieved correlations (0.87 and 0.74, respectively) that were similar to those originally reported by the authors (Table II). However, it is likely that these two matrix-based methods were trained on some of the peptides included in the DS_{178} dataset, because they were developed using binding affinity data of poly-Alanine peptides, such as those included in the DS_{178} dataset. In any case, TAP binding affinity predicted by our SVM models correlated more closely with those predicted by CM than with those predicted by related SVM-based methods (Table III). Overall, these results highlight the relevance of identifying and including new data points for training predictive models.

In this study, we have also developed a Web-based tool, TAPREG, to predict the binding affinity of peptides to TAP, which is available for free public use at http://imed.med.ucm.es/Tools/tapreg/. Currently, there are two dedicated web-based tools to predict the binding affinity of peptides to TAP: SMVTAP[27] (http://www-bs.informatik.uni-tuebingen.de/Services/SVMTAP/) and TAPPRED[26] (http://www.imtech.res.in/raghava/tappred/), both of them based on SVMs. These two resources use a protein sequence as input and report the 9-mer peptides encompassed by the protein, ranked by their predicted binding affinity to TAP. In addition to this task, TAPREG can be used to predict the binding affinity to TAP of multiple peptides with a length ranging from eight to 16 residues,[14,21] which is consistent with the transport activity displayed by TAP.

Until now TAP binding affinity of peptides longer than nine residues could only be achieved using quantitative matrices, and only the 3 N-terminal residues and the C-terminus of the peptide were considered to matter for TAP binding.[28] In contrast, in TAPREG, we compute the TAP affinity using nine residues selected from the larger peptides—those equivalent to the 9-mer peptides used for training—as we have shown that all residues in a 9-mer peptide contribute to binding. To our knowledge, this is the first machine-learning based approach that can predict the binding affinity to TAP of peptides longer than nine residues.

CONCLUSIONS

We have used a large dataset of 9-mer peptides of known affinity to TAP to dissect the TAP binding preferences, concluding that each peptide position has a quantitative contribution to TAP binding. Moreover, we have been able to generate SVM models with enhanced predictive performance as a result of including new peptide binding data. Because accurate modeling of TAP activity is relevant for T cell epitope selection,[12,13] we have implemented the Web-based tool TAPREG (http://imed.med.ucm.es/Tools/tapreg/). Unlike any related resource, TAPREG can be used to predict the binding affinity of peptides ranging from eight to 16 residues, in a manner that is consistent with the activity exhibited by TAP.

REFERENCES

1. Paul WE. Fundamental immunology. Philadelphia, PA: Lippincott, Williams & Wilkins; 1998.
2. Von Boehmer H. Positive and negative selection of the ab T cell repertoire in vivo. Curr Opin Immunol 1991;3:210–215.
3. Craiu A, Akopian T, Goldberg A, Rock KL. Two distinct proteolytic processes in the generation of a major histocompatibility complex class I-presented peptide. Proc Natl Acad Sci USA 1997;94:10850–10855.
4. Yewdell JW, Haeryfar SM. Understanding presentation of viral antigens to CD8+ T cells in vivo: the key to rational vaccine design. Annu Rev Immunol 2005;26:651–682.

5. Pamer E, Cresswell P. Mechanisms of MHC class I--restricted antigen processing. Annu Rev Immunol 1998;16:323–358.
6. York IA, Goldberg AL, Mo XY, Rock KL. Proteolysis and class I major histocompatibility complex antigen presentation. Immunol Rev 1999;172:49–66.
7. Serwold T, Gonzalez F, Kim J, Jacob N. ERAAP customizes peptides for MHC Class I molecules in the endoplasmic reticulum. Nature 2002;419:480–483.
8. Beekman NJ, Van Veelen PA, Van Hall T, Neisig A, Sijts A, Camps M, Kloetzel PM, Neefjes JJ, Melief CJ, Ossendorp F. Abrogation of CTL epitope processing by single amino acid substitution flanking the C-terminal proteasome cleavage site. J Immunol 2000;164:1898–1905.
9. Smith KD, Lutz CT. Peptide-dependent expression of HLA-B7 on antigen processing-deficient T2 cells. J Immunol 1996;156:3755–3764.
10. Zhong W, Reche PA, Lai CC, Reinhold B, Reinherz EL. Genome-wide characterization of a viral cytotoxic T lymphocyte epitope repertoire. J Biol Chem 2003;278:45135–45144.
11. Wang M, Lamberth K, Harndahl M, Røder G, Stryhn A, Larsen MV, Nielsen M, Lundegaard C, Tang ST, Dziegiel MH, Rosenkvist J, Pedersen AE, Buus S, Claesson MH, Lund O. CTL epitopes for influenza A including the H5N1 bird flu; genome-, pathogen-, and HLA-wide screening. Vaccine 2007;25:2823–2831.
12. Uebel S, Krass P, Kienle S, Wiesmuller KH, Jung G, Tampe R. Recognition principle of the TAP transporter disclosed by combinatorial peptide libraries. Proc Natl Acad Sci USA 1997;94:8976–8981.
13. Ackerman AL, Cresswell P. Cellular mechanisms governing cross-presentation of exogenous antigens. Nat Immunol 2004;5:678–684.
14. Androlewicz MJ, Cresswell P. Human transporters associated with antigen processing possess a promiscuous peptide-binding site. Immunity 1994;1:7–14.
15. Abele R, Tampe R. The ABCs of immunology: structure and function of TAP, the transporter associated with antigen processing. Physiology (Bethesda) 2004;19:216–224.
16. Shepherd JC, Schumacher TN, Ashton-Rickardt PG, Imaeda S, Ploegh HL, Janeway CA, Tonegawa S. TAP1-dependent peptide translocation in vitro is ATP dependent and peptide selective. Cell 1993;74:577–584.
17. Neefjes JJ, Momburg F, Hammerling GJ. Selective and ATP-dependent translocation of peptides by the MHC-encoded transporter. Science 1993;261:769–771.
18. Van Endert PM, Tampe R, Meyer TH, Tisch R, Bach JF, Mcdevitt HO. A sequential model for peptide binding and transport by the transporters associated with antigen processing. Immunity 1994;1:491–500.
19. Gubler B, Daniel S, Armandola EA, Hammer J, Caillat-Zucman S, Van Endert PM. Substrate selection by transporters associated with antigen processing occurs during peptide binding to TAP. Mol Immunol 1998;35:427–433.
20. Armandola EA, Momburg F, Nijenhuis M, Bulbuc N, Fruh K, Hammerling GJ. A point mutation in the human transporter associated with antigen processing (TAP2) alters the peptide transport specificity. Eur J Immunol 1996;26:1748–1755.
21. Momburg F, Roelse J, Howard JC, Butcher GW, Hammerling GJ, Neefjes JJ. Selectivity of MHC-encoded peptide transporters from human, mouse and rat. Nature 1994;367:648–651.
22. Van Endert PM, Riganelli D, Greco G, Fleischhauer K, Sidney J, Sette A, Bach JF. The peptide-binding motif for the human transporter associated with antigen processing. J Exp Med 1995;182:1883–1895.
23. Daniel S, Brusic V, Caillat-Zucman S, Petrovsky N, Harrison L, Riganelli D, Sinigaglia F, Gallazzi F, Hammer J, Van Endert PM. Relationship between peptide selectivities of human transporters associated with antigen processing and HLA class I molecules. J Immunol 1998;161:617–624.
24. 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.
25. Zhang GL, Petrovsky N, Kwoh CK, August JT, Brusic V. PRED(TAP): a system for prediction of peptide binding to the human transporter associated with antigen processing. Immunome Res 2006;2:3.
26. Bhasin M, Raghava G. Analysis and prediction of affinity of TAP binding peptides using cascade SVM. Protein Sci 2004;13:596–607.
27. Donnes P, Kohlbacher O. Integrated modeling of the major events in the MHC class I antigen processing pathway. Protein Sci 2005;14:2132–2140.
28. Peters B, Bulik S, Tampe R, Van Endert PM, Holzhutter HG. Identifying MHC class I epitopes by predicting the TAP transport efficiency of epitope precursors. J Immunol 2003;171:1741–1749.
29. Doytchinova I, Hemsley S, Flower DR. Transporter associated with antigen processing preselection of peptides binding to the MHC: a bioinformatic evaluation. J Immunol 2004;173:6813–6819.
30. Doytchinova IA, Blythe MJ, Flower DR. Additive method for the prediction of protein-peptide binding affinity. Application to the MHC class I molecule HLA-A*0201. J Proteome Res 2002;1:263–272.
31. Toseland CP, Clayton DJ, Mcsparron H, Hemsley SL, Blythe MJ, Paine K, Doytchinova IA, Guan P, Hattotuwagama CK, Flower DR. AntiJen: a quantitative immunology database integrating functional, thermodynamic, kinetic, biophysical, and cellular data. Immunome Res 2005;1:4.
32. Neuwald AF, Liu JS, Lawrence CE. Gibbs motif sampling detection of bacterial outer membrane protein repeats. Protein Sci 1995;4:1618–1632.
33. Sathiamurthy M, Peters B, Bui HH, Sidney J, Mokili J, Wilson SS, Fleri W, Mcguinness DL, Bourne PE, Sette A. An ontology for immune epitopes: Application to the design of a broad scope database of immune reactivities. Immunome Res 2005;1:2.
34. Reche PA, Zhang H, Glutting JP, Reinherz EL. EPIMHC: a curated database of MHC-binding peptides for customized computational vaccinology. Bioinformatics 2005;21:2140–2141.
35. Frank E, Hall M, Trigg L, Holmes G, Witten IH. Data mining in bioinformatics using Weka. Bioinformatics 2004;20:2479–2481.
36. Smola AJ, Scholkopf B. A Tutorial on support vector regression. NC2-TR-1998–030. NTRS. Berlin, Germany: Springer; 1998.
37. Shevade SK, Keerthi SS, Bhattacharyya C, Murthy KRK. Improvements to SMO Algorithm for SVM Regression, Technical Report CD-99–16, 2000.
38. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA 1992;89:10915–10919.
39. Rice P, Longden I, Bleasby A. EMBOSS: the European molecular biology open software suite. Trends Genet 2000;16:276–277.
40. Meng X-L, Rosenthal R, Rubin DB. Comparing correlated correlation coefficients. Psychol Bull 1992;111:172–175.
41. Yang ZR. Biological applications of support vector machines. Brief Bioinform 2004;5:328–338.
42. Bhasin M, Reinherz EL, Reche PA. Recognition and classification of histones using support vector machine. J Comput Biol 2006;13:102–112.
43. Bhasin M, Zhang H, Reinherz EL, Reche PA. Prediction of methylated CpGs in DNA sequences using a support vector machine. FEBS Lett 2005;579:4302–4308.
44. Vapnik V. The nature of statistical learning theory. New York: Springger-Verlag; 1995.

7. CAPÍTULO IV

Desarrollo de otras herramientas computacionales:

7.1 PVS: *Protein Variability Server*

7.2 TEPIDAS: *Integrating T-cell epitope annotations with sequence and structural information using DAS*

7.1 PVS

7.1.1 Justificación y Objetivos

El análisis de la variabilidad de secuencias permite deducir información funcional y evolutiva de las proteínas. En este trabajo hemos desarrollado la herramienta PVS, que permite:

- Calcular la variabilidad de las secuencias de un alineamiento, según la entropía de Shannon, el índice de diversidad de Simpson y el coeficiente de variabilidad de Wu-Kabat.
- Visualización de dicha variabilidad en la estructura 3D relevante.
- Obtención de una secuencia de referencia con las posiciones variables codificadas, determinadas por un umbral de variabilidad seleccionado por el usuario.
- Utilizar la secuencia de referencia con las posiciones enmascaradas en el servidor RANKPEP para la predicción de epítopos conservados.

7.1.2 Conclusiones

- Los análisis de variabilidad y conservación de secuencias, especialmente cuando se combinan con la visualización de la variabilidad en la estructura 3D relevante, son útiles para estudiar las relaciones entre estructura y función, así como para revelar los residuos importantes.
- El análisis de la variabilidad de secuencias llevado a cabo con la herramienta PVS facilita el descubrimiento de epítopos T y B conservados , facilitando el diseño de vacunas basadas en epítopos.

Published online 27 April 2008 *Nucleic Acids Research, 2008, Vol. 36, Web Server issue* ***W35–W41***
doi:10.1093/nar/gkn211

PVS: a web server for protein sequence variability analysis tuned to facilitate conserved epitope discovery

Maria Garcia-Boronat[1], Carmen M. Diez-Rivero[1], Ellis L. Reinherz[2,3] and Pedro A. Reche[1,*]

[1]Immunomedicine Group, Department of Microbiology I, Division of Immunology, Facultad de Medicina, Universidad Complutense de Madrid, Ave Complutense s/n, Madrid 28040, Spain, [2]Laboratory of Immunobiology and Department of Medical Oncology, Dana-Farber Cancer Institute and Department of Medicine and [3]Harvard Medical School, 77 Avenue Louis Pasteur, Boston, MA 02115, USA

Received January 20, 2008; Revised April 3, 2008; Accepted April 9, 2008

ABSTRACT

We have developed PVS (Protein Variability Server), a web-based tool that uses several variability metrics to compute the absolute site variability in multiple protein-sequence alignments (MSAs). The variability is then assigned to a user-selected reference sequence consisting of either the first sequence in the alignment or a consensus sequence. Subsequently, PVS performs tasks that are relevant for structure-function studies, such as plotting and visualizing the variability in a relevant 3D-structure. Neatly, PVS also implements some other tasks that are thought to facilitate the design of epitope discovery-driven vaccines against pathogens where sequence variability largely contributes to immune evasion. Thus, PVS can return the conserved fragments in the MSA—as defined by a user-provided variability threshold—and locate them in a relevant 3D-structure. Furthermore, PVS can return a variability-masked sequence, which can be directly submitted to the RANKPEP server for the prediction of conserved T-cell epitopes. PVS is freely available at: http://imed.med.ucm.es/PVS/.

INTRODUCTION

Multiple sequence alignments (MSAs) of homologous proteins encompass unique patterns of conserved and variable residues. The functional relevance of conserved residues is widely acknowledged. Indeed, functionally important residues such as those defining interacting sites, substrate binding sites or simply relevant to protein-structure integrity, display a low rate of substitution. This observation is predicted by the neutral evolution model (1), which also indicates that variable residues are somehow less important. Consequently, many methods have been developed to look for general and subfamily conservation patterns (2–8) as a key to identify functionally important residues. Moreover, some of these approaches are available for public use through the web (9–11). While these methods and related servers are very useful to identify functionally relevant residues, they generally underestimate the variability in the MSAs and certainly dismiss the significance of variable sites.

Variable residues in proteins can however be functionally relevant. Indeed, sequence variability is widely used by biological systems to generate functional heterogeneity. Thus, the hypervariable residues in the T-cell receptors (TCR) and Immunoglobulins match the antigen-binding residues (12). Likewise, the most polymorphic (variable) residues in the human leukocyte antigens (HLAs) are located on their binding groove, explaining the distinct peptide-binding specificities of the HLA allelic variants (13,14). Therefore, having a direct estimate of the sequence variability in an MSA is important to fill gaps in structural knowledge and to offer insight for function-structure studies. Indeed, long before the first antigen-bound immunoglobulin crystal structures were solved (15–17), Kabat (18) was able to anticipate that highly variable segments in immunoglobulin molecules match the antigen contact sites. Importantly, the estimation of sequence variability in rapidly evolving protein antigens from pathogens that use sequence variation for immune evasion (19–21) provides a mean to identify conserved antigenic determinant targets (epitopes), and consequently it is useful for epitope-vaccine design.

For all the above, we have developed PVS, a web server that provides absolute sequence variability estimates 'per site' in an MSA as determined by the Shannon Entropy

*To whom correspondence should be addressed. Tel: +34 91 394 7229; Fax: +34 91 394 1641; Email: parecheg@med.ucm.es

(22), the Simpson Diversity Index (23) and the Wu-Kabat Variability Coefficient (18). The Wu-Kabat's coefficient, perhaps the most popular sequence variability metric, is effective in resolving the highest diversity positions, but as it has been noted, underestimates the diversity in the MSA (24). In comparison, Shannon and Simpson methods are statistically more sound for quantifying a system diversity, and are widely used in ecology and sequence analyses (25). Following the variability computations, PVS can plot the variability in the MSA and display it in a relevant 3D-structure. PVS can also return the selected reference sequence with the variable positions masked, as well as the sequence fragments (minimum length selected by the user) containing only nonvariable residues, as determined by a user-provided variability threshold. Within the PVS output page, the user can also locate the conserved fragments in the provided 3D-structure, and submit the variability-masked sequence to the RANKPEP server (26,27) for the prediction of conserved T-cell epitopes. Here we will show that these features are particularly relevant for epitope discovery-driven design of vaccines against pathogens displaying large sequence variability.

SYSTEMS AND METHODS

Automated generation of MSAs

Automated MSAs are obtained from the protein sequence of a Protein Data Bank (PDB) file following a BLAST (28) search against the SWISSPROT database. The BLAST search is performed using an E value of $1e^{-20}$ and a maximum of 250 hits are considered. Subsequently, the relevant sequence hits are aligned using MUSCLE (29).

Computation of sequence variability

The Shannon Diversity Index (Shannon Entropy) (22), the Simpson Diversity Index (23) and the Wu-Kabat Variability Coefficient (30) are used to estimate the sequence variability 'per site' (V) in MSAs.

The Shannon Diversity Index (H) is given by

$$H = -\sum_{i=1}^{M} p_i \log_2 p_i \quad \mathbf{1}$$

where, p_i is the fraction of residues of amino acid type i, and M represents the total number of amino acid types in a given site. H ranges from 0 (only one amino acid type is present at that position) to 4.322 (all 20 amino acids are equally represented in that position). Note, that for a site including gaps the maximum value of H will be 4.39.

We estimate the Simpson Diversity Index (D) using the following equation:

$$D = 1 - \sum_{i=1}^{S} \frac{n_i(n_i - 1)}{N(N-1)} \quad \mathbf{2}$$

where, n_i is the number of residues of type i, N is the total number of residues and S is the number of different symbols 'per site'. From Equation (2) it follows that $0 \leq D \geq 1$. Those sites with D values near 1 are highly variable and those with D values near 0 are almost constant.

The Wu-Kabat Variability Coefficient (W) is given by:

$$W = \frac{Nk}{n} \quad \mathbf{3}$$

Here, N is the number of sequences in the MSA, k is the number of different amino acids at a given position and n is the frequency of the most common amino acid at that position. The minimum value of W is 1. Unlike for H and D, W maximum value increases with the number of sequences in the MSA.

Mapping sequence variability onto a 3D-structure

Given a relevant PDB file with the coordinates of a 3D-structure, the V in an MSA is mapped onto the 3D-structure by simply replacing the B-factor of the relevant residues in the PDB with the computed V values.

Implementation

PVS is implemented on an Apache Web server running under the Mac OSX operating system. The PVS functional core consists of a PERL CGI (Common Gateway Interface) script that handles the input, executes several subroutines implementing the above outlined methods, and then assembles and displays the results. PVS uses GNUPLOT (http://www.gnuplot.info) to plot the variability and the Bioperl Bio::Graphics module (http://www.bioperl.org) to generate sequence graphs with features. For displaying 3D-structures, PVS uses Jmol, an open-source Java molecular viewer for three-dimensional chemical structures (http://www.jmol.net).

DESCRIPTION AND USAGE OF THE SERVER

Web interface

The PVS web interface will dynamically change to present only those fields that apply to the user made selections. This is done using JavaScript. Moreover, the web interface is divided into the INPUT, SEQUENCE VARIABILITY OPTIONS and OUTPUT TASKS sections which overall facilitate an intuitive use of the server. The web interface also provides links to help pages, and specific information regarding the elements featured by the server can be obtained from the question mark icons. A description of the server usage, including the input and output follows here.

Input and variability options

The main input data for PVS can either be (i) an MSA or (ii) a PDB and users have to select one type or another from the INPUT section. Once a selection is made, the PVS web interface will show only the fields relevant to the selected input type. Thus, for the MSA option, the user can either paste or upload the alignment, which can be in CLUSTALW, GCG or FASTA formats. For the PDB input option, the user can either upload a PDB file or supply a PDB code and PVS will retrieve the corresponding PDB file from the Brookhaven database (http://www.rcsb.org/). Next, an MSA will be built from the sequence of the PDB chain—specified by the user—as

detailed in 'Systems and methods' section. If no chain is provided, the first chain in the PDB file will be taken by default. Currently, PVS will only process MSAs with less than 400 sequences and 250 000 symbols. Also, automated MSAs will only be generated from PDB protein sequences shorter than 400 residues. If such limits are exceeded, the server will return an error.

Subsequently, PVS will subject the MSA to a sequence variability analysis using several methods that can be selected by the user from the 'Sequence variability options' section. The default method, 'Shannon', uses the Shannon Diversity Index as the variability metric [Systems and methods section, (Equation (1)]. Additionally, users can also select the 'Wu-Kabat' Variability Coefficient [Systems and methods, Equation (2)] and the 'Simpson' Diversity Index [Systems and methods, Equation (3)].

Output

The output for PVS will be determined by the user-selected options in the 'Output tasks' section. By default, PVS will 'plot the variability' in the MSA—computed for each selected variability method—against a reference sequence selected by the user (Figure 1A). The reference sequence can either be a consensus sequence (default) or the first sequence in the MSA. Additionally, the following tasks can be performed by PVS: (i) 'Mask sequence variability'; (ii) 'Return conserved fragments' and (iii) 'Map structural variability'. The outputs and restrictions resulting from selecting these tasks are discussed below.

Mask sequence variability. This option returns the selected reference sequence so that those residues with *V*

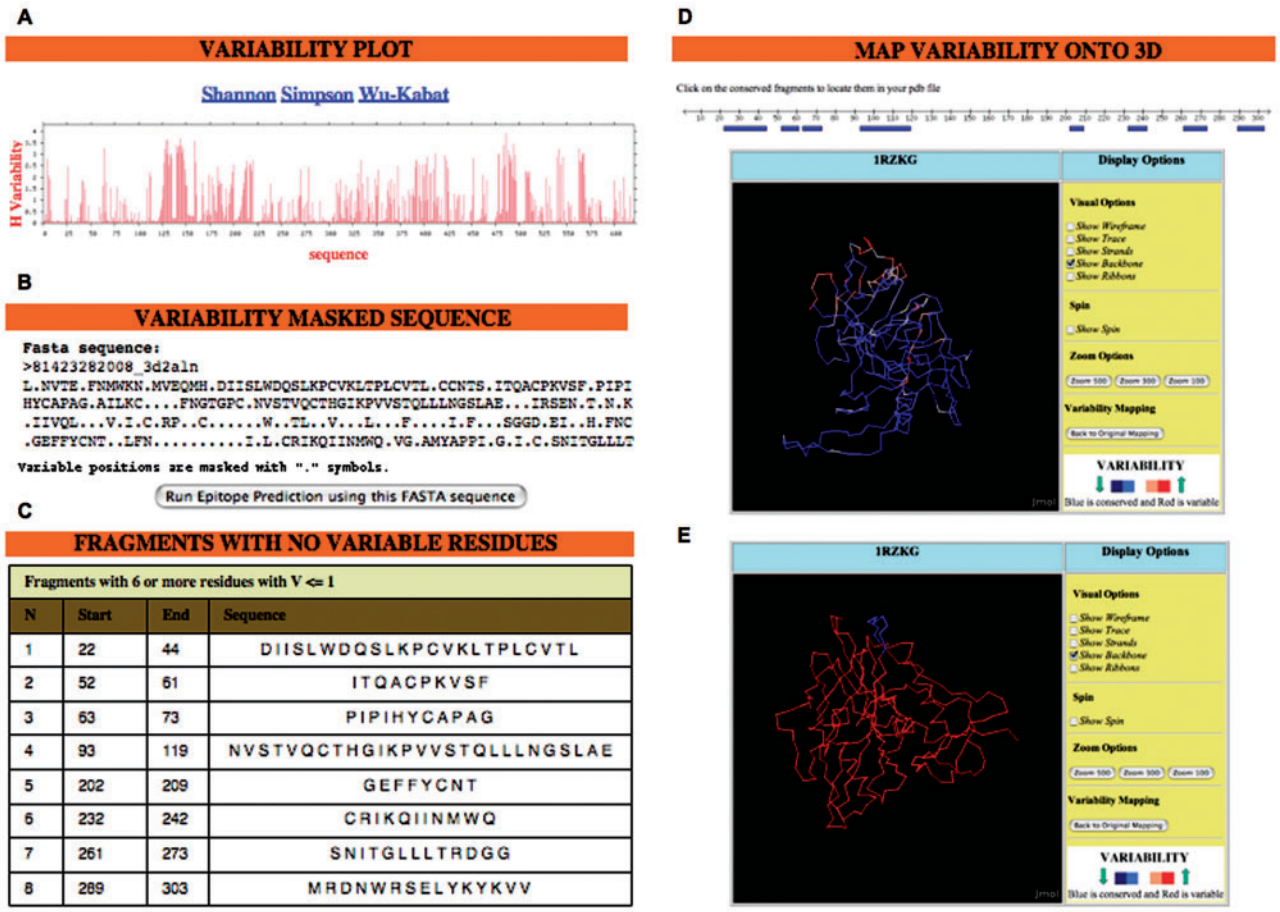


Figure 1. PVS output. The figure shows a composition with the possible outputs of PVS. Results were obtained using an MSA corresponding to the HIV1 glycoprotein gp120 (residues 31–183 in gp160 from HIV-1 strain H2XB2). The MSA was generated from 359 representative sequences of the HIV-1 clades A (73), B (85), C (85), D (51) and 01_AE (65) using the program MUSCLE (29). The MSA is available at http://imed.med.ucm.es/PVS/supplemental/gp120_aln.html. The sequence variability was computed using the 'Shannon', 'Simpson' and 'Wu-Kabat' methods, and from the 'sequence variability options', a reference 'consensus sequence' and the default 'variability threshold of 1.0' were selected. (**A**) 'Variability plot'. Users can change the variability metric ('Shannon', 'Simpson' and 'Wu-Kabat') by clicking on the relevant links. (**B**) 'Variability masked sequence'. The sequence is returned in FASTA and T-cell epitope predictions can be obtained by clicking on the 'Run Epitope Prediction' bottom. (**C**) 'Conserved fragments with no variable residues'. In this example, a 'minimal fragment length' of eight was selected. (**D**) 'Structural variability mapping'. Sequence variability in the alignment was mapped onto the 3D-coordinates of gp120 (chain G of PDB 1RZK). The output allows the visualization of the variability in several user-selected renderings of the 3D structure. PVS can also display a graph of the protein sequence with the conserved fragments shown in blue. By clicking on a fragment, the user will locate it on the 3D-structure as shown in (**E**) with fragment 2. The output used to make this figure is available at: http://imed.med.ucm.es/PVS/supplemental/gp120_pvs.html.

greater or equal than the selected variability threshold are masked using a '.' symbol. The variability-masked sequence is returned in FASTA format (Figure 1B), and it can be submitted to RANKPEP (26,27), the only T-cell epitope prediction tool that can anticipate conserved T-cell epitopes from a variability-masked sequence.

Return conserved fragments. This option identifies those fragments (minimum length selected by user) in the selected reference sequence consisting only of consecutive residues with V below the set variability threshold (Figure 1C). These fragments are returned, sorted in a table by their position in the MSA. For options (i) and (ii), the variability threshold must be between 0 and 4.3 in the case of the Shannon Entropy and between 0 and 1 for the Simpson Diversity Index (See Systems and methods section), otherwise PVS will return an error message. The default 'variability threshold' is 1.0 for the 'Shannon' Entropy method and 0.46 for the 'Simpson' Diversity Index, values which are regarded as indicatives of low variability (24). If the Shannon and Simpson methods were selected, PVS will proceed considering the variability threshold as for Shannon. Note that unlike the Shannon and Simpson Diversity Index, the upper value of the Wu-Kabat Variability Coefficient increases with the number of sequences in the MSA (see Systems and methods section). Therefore, since the 'variability threshold' must be entered prior to submitting the job, the options of masking the variability and returning conserved fragments are not available if the Wu-Kabat Variability Coefficient is the only variability metric selected.

Map structural variability. The sequence variability in the MSA is mapped onto a 3D-structure through a B-factor (see Systems and methods section). If an MSA was entered in PVS, the user must upload a relevant PDB to map the sequence variability onto it. Obviously, if the input was a PDB, PVS will map the sequence variability onto that same 3D structure. Note that when the 'Map structural variability' option is selected the variability is only computed for the positions in the MSA that map with the PDB. The resulting 3D structure is displayed using an interactive Jmol applet (JavaScript must be enabled in the browser) that allows the user to visualize the variability over several structural renderings, in a color scale that goes from blue for constant residues to red for highly variable residues (Figure 1D). In addition, if the 'Return Conserved fragments' task had also been selected, PVS will display a graph of the protein sequence with the conserved fragments shown in blue. By clicking on a fragment, the user will locate it on the 3D structure (Figure 1E).

Limitations

Proper computation of sequence variability from MSAs is contingent on the quality of the alignments. Therefore, we suggest evaluating the reliability of MSAs using the corresponding applications implemented in the TCOFFEE web server (http://www.igs.cnrs-mrs.fr/Tcoffee/) (31). This evaluation is particularly relevant when working with MSAs of distantly related proteins. However, the users should not have problems with the quality of MSAs built from very similar sequences (e.g. allelic and antigen variants). Likewise, we do not anticipate quality problems on the automated MSAs generated by the server because they are built considering only highly similar protein sequences. Finally, while the methods implemented in PVS are for computing sequence variability from MSAs, other methods do exist that can estimate sequence variability without the need of an MSA (32–34).

COMPARISON WITH AVAILABLE SERVERS

Sequence variability or conservation analyses, particularly when combined with mapping the variability onto a relevant 3D-structure, are useful to explore structure–function relationships and to reveal functionally relevant residues. Not surprisingly, some servers are already available (summarized in Table 1) that given an MSA can perform related tasks, such as providing a consensus sequence as 'Consensus', or plotting the relative sequence variability as in 'WebVar' (35). Other servers such as 'Conseq' and 'TreeDet' (20) carry out sophisticated conservation analyses to identify functionally relevant residues, and 'Consurf' (20), using the same phylogeny-dependent algorithms as 'Conseq' (9), maps the conservation scores onto a relevant 3D-structure. The 'Conservancy' (36) server is another related tool that from a set of user-provided predefined epitopes, identifies their conservation as determined by a percentage of identity. In comparison, PVS can handle more input types (PDBs or MSAs) and formats (MSAs can be in FASTA, CLUSTAW and GCG) that most of the related servers, and offers the largest set of functional tasks (Table 1). In any case, despite all these servers being related to some extent, they differ with regard to their methods and specific objectives, and indeed PVS is unique for using sequence variability analyses to help with epitope-vaccine design.

PVS RELEVANCE FOR EPITOPE DISCOVERY: WORKED EXAMPLES

Sequence variability analyses are commonly applied to infer evolutive and functional information in systems where functional diversity is achieved through sequence variation. For example, we previously applied a sequence variability analysis to human class I and class II MHC molecules (13), which, when correlated with the available structural information, clearly showed that the majority of the polymorphisms exhibited by these molecules are related with their differential peptide-binding specificity. In addition, we could also identify some other polymorphisms that could determine the restriction by their cognate T-cell receptors. While these classic structure–function studies can be carried in PVS, we will focus here on illustrating the use of PVS in the context of epitope-vaccine design.

PVS results are in fact tuned to facilitate the design of vaccines driven by epitope discovery against pathogenic organisms such as HIV-1, where sequence variation largely contributes to immune evasion, and sequence

Table 1. Web servers related to PVS

Web server	Input: formats	Output and tasks	Ref
• PVS http://imed.med.ucm.es/PVS/	• MSA: CLUSTAL, FASTA, GCG/MSF • PDB: Uploaded or retrieved • MSA and PDB	1. Compute sequence variability 2. Plot sequence variability 3. Map and display variability in 3D structures 4. Mask sequence variability 5. T-cell epitope prediction 6. Return conserved fragments 7. Locate conserved fragments into 3D structures/B-cell epitope prediction	
• SVS* http://bio.dfci.harvard.edu/Tools/svs.html	• MSA: CLUSTAL	1. Compute sequence variability as given by Shannon Entropy 2. Plot sequence variability 3. Return conserved fragments	
• SiteVarProt http://159.149.109.16/Tools/SiteVarProt.php	• MSA: FASTA	1. Compute relative sequence variability 2. Plot sequence variability	(35)
• Consensus http://coot.embl.de/Alignment//consensus.html	• MSA: CLUSTAL and GCG/MSF	1. Consensus sequence at various thresholds with amino acid groupings	
• Conseq http://conseq.bioinfo.tau.ac.il/	• SEQUENCE: FASTA • MSA: NBRF/PIR, EMB, FASTA, GDE, CLUSTAL, GCG/MSF and RSF	1. Compute conservation scores 2. Compute solvent accessibility 3. Return color-coded sequence with calculations	(9)
• Consurf http://consurf.tau.ac.il/	• PDB: Uploaded or retrieved • MSA and PDB	1. Compute conservation scores 2. Map and display conservation scores in 3D structures	(11)
• TreeDet http://www.pdg.cnb.uam.es/Servers/treedet/	• MSA: CLUSTAL, FASTA, MSF and PIR	• Predicts and display functionally relevant residues	(10)
• Conservancy http://tools.immuneepitope.org/tools/conservancy	• SEQUENCES: FASTA	• Computes *per site* sequence identity of epitopes in protein sources	(36)

PVS is an enhanced version of SVS, a server previously developed by Dr Reche. SVS has >85 000 hits since it started running in 2002.

variability analyses are needed to identify conserved epitopes (37). The discovery of conserved T-cell epitopes (antigenic peptides recognized by the T cells when bound and displayed by MHC molecules in the cell surface of target cells) is facilitated in PVS by providing variability-masked sequences that can be submitted directly to the RANKPEP web server. Subsequently, RANKPEP will only return predicted conserved T-cell epitopes, thus also reducing the number of T-cell epitopes that have to be considered for experimental epitope confirmation. For example, from the gp120 variability masked sequence shown in Figure 1, RANKPEP will return two conserved T-cell epitopes restricted by the HLA I molecule A*0201 (KLTPLCVTL and PVVSTQLLL) as judged by their above-threshold binding score to A*0201 and by the predicted proteasomal cleavage. These predictions can be obtained from the gp120 PVS result page at: http://imed.med.ucm.es/PVS/supplemental/gp120_pvs.html. In comparison, the corresponding gp120 sequence of HIV-1 H2XB2 strain will yield 10 epitopes, a 5-fold increase in the epitope number (data not shown). Therefore, regardless of the predictive power of RANKPEP, this strategy saves the time, effort and resources one would need to consume confirming nonconserved T-cell epitopes that are not as suitable for vaccine design.

PVS results can also be helpful for the identification of conserved B-cell epitopes, the antigenic determinants of antibodies (Abs). As an example, we were able to detect seven highly conserved fragments of six or more residues (Table 2) from an MSA of the ectodomain of HIV-1 gp41 (details in Table 2 legend), which is the target of various broadly neutralizing Abs (38). Interestingly, fragments 5 and 7 encompass the antigenic determinants (B-cell epitopes) of the monoclonal antibodies CL3 and ZE10, respectively, both broadly neutralizing (38). Abs, however, only recognize solvent-exposed epitopes and most of them are conformational but can also be linear. Consequently, when used as immunogens, the majority of these conserved fragments will fail to yield Abs cross-reacting with the native antigen. However, one can also use PVS to locate the conserved fragments in the 3D-structure (when available), and select those that are surface exposed. Under such scenario, the chance of producing Abs that are cross-reactive with the native antigen and broadly neutralizing will be greatly increased. For example, in Figure 1E we have chosen to display the conserved fragment 2 (ITQACPKVSF) from HIV-1 gp120, which is readily accessible to the solvent. Moreover, from the PVS results obtained from the gp120 MSA (http://imed.med.ucm.es/PVS/supplemental/gp120_pvs.html) one could

Table 2. Conserved fragments in ectodomain of HIV-1 gp41

N	Start	End	Sequence
1	1	7	S T M G A A S
2	9	25	T L T V Q A R Q L L S G I V Q Q Q
3	27	55	N L L R A I E A Q Q H L L Q L T V W G I K Q L Q A R V L A
4	62	67	D Q Q L L G
5	69	74	W G C S G K
6	87	92	S W S N K S
7	153	158	W L W Y I K

Fragments were selected to have six or more consecutive residues with $H \leq 1$, and were obtained from an MSA of the HIV-1 gp41 ectodomain (residues 528–674 in gp160 from HIV-1 strain H2XB2). The MSA includes 359 representative sequences of HIV-1 clades A (73), B (85), C (85), D (51) and 01_AE (65) that were aligned using MUSCLE (29). The MSA is available at http://imed.med.ucm.es/PVS/supplemental/gp41_ecto_aln.html

also see that fragment 3 and significant portions of fragments 1, 4 and 6 are also accessible to the solvent.

CONCLUSIONS AND FUTURE DIRECTIONS

PVS is a user-friendly and versatile web server where sequence variability computations are exploited to facilitate structure-function studies and, unlike any other related server, *de novo* epitope discovery. In the future, we plan to include additional variability and conservation scores. Moreover, we will implement solvent accessibility calculations, which should enhance the potential of PVS in structure–function studies and B-cell epitope discovery.

ACKNOWLEDGEMENTS

This work was supported by a Ramón y Cajal Grant ('convocatoria 2005') and by grant SAF2006-07879 from the 'Ministerio de Educación y Ciencia' (M.E.C) of Spain, both to P.A.R. The authors wish to thank Dr Jose R. Regueiro for corrections and thoughtful comments. Funding to pay the Open Access publication charges for this article was provided by M.E.C of Spain (SAF2006-07879).

Conflict of interest statement. None declared.

REFERENCES

1. Kimura,M. (1983) *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge, pp. 34–55.
2. del Sol Mesa,A., Pazos,F. and Valencia,A. (2003) Automatic methods for predicting functionally important residues. *J. Mol. Biol.*, **326**, 1289–1302.
3. Hannenhalli,S.S. and Russell,R.B. (2000) Analysis and prediction of functional sub-types from protein sequence alignments. *J. Mol. Biol.*, **303**, 61–76.
4. Lichtarge,O., Bourne,H.R. and Cohen,F.E. (1996) An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.*, **257**, 342–358.
5. Madabushi,S., Yao,H., Marsh,M., Kristensen,D.M., Philippi,A., Sowa,M.E. and Lichtarge,O. (2002) Structural clusters of evolutionary trace residues are statistically significant and common in proteins. *J. Mol. Biol.*, **316**, 139–154.
6. Mihalek,I., Res,I. and Lichtarge,O. (2004) A family of evolution-entropy hybrid methods for ranking protein residues by importance. *J. Mol. Biol.*, **336**, 1265–1282.
7. Pupko,T., Bell,R.E., Mayrose,I., Glaser,F. and Ben-Tal,N. (2002) Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics*, **18**, S71–S77.
8. Thibert,B., Bredesen,D.E. and del Rio,G. (2005) Improved prediction of critical residues for protein function based on network and phylogenetic analyses. *BMC Bioinformatics*, **6**, 213.
9. Berezin,C., Glaser,F., Rosenberg,J., Paz,I., Pupko,T., Fariselli,P., Casadio,R. and Ben-Tal,N. (2004) ConSeq: the identification of functionally and structurally important residues in protein sequences. *Bioinformatics*, **20**, 1322–1324.
10. Carro,A., Tress,M., de Juan,D., Pazos,F., Lopez-Romero,P., del Sol,A., Valencia,A. and Rojas,A.M. (2006) TreeDet: a web server to explore sequence space. *Nucleic Acids Res.*, **34**, 115.
11. Landau,M., Mayrose,I., Rosenberg,Y., Glaser,F., Martz,E., Pupko,T. and Ben-Tal,N. (2005) ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures. *Nucleic Acids Res.*, **33**, 302.
12. Paul,W.E. (1998) *Fundamental Immunology*. 5th edn. Lippincott Williams & Wilkins, Philadelphia, pp. 47–59, pp. 227–259.
13. Reche,P.A. and Reinherz,E.L. (2003) Sequence variability analysis of human class I and class II MHC molecules: functional and structural correlates of amino acid polymorphisms. *J. Mol. Biol.*, **331**, 623–641.
14. Stern,L.J. and Wiley,D.C. (1994) Antigen peptide binding by class I and class II histocompatibility proteins. *Structure*, **2**, 245–251.
15. Padlan,E.A., Silverton,E.W., Sheriff,S., Cohen,G.H., Smith-Gill,S.J. and Davies,D. (1989) Structure of an antibody-antigen complex: crystal structure of the HyHEL-10 Fab-lysozyme complex. *Proc. Natl Acad. Sci. USA*, **86**, 5938–5942.
16. Rose,D.R., Strong,R.K., Margolies,M.N., Gefter,M.L. and Petsko,G.A. (1990) Crystal structure of the antigen-binding fragment of the murine anti-arsonate monoclonal antibody 36-71 at 2.9-A resolution. *Proc. Natl Acad. Sci. USA*, **87**, 338–342.
17. Stanfield,R.L., Fieser,T.M., Lerner,R.A. and Wilson,I.A. (1990) Crystal structures of an antibody to a peptide and its complex with peptide antigen at 2.8 A. *Science*, **248**, 712–719.
18. Kabat,E.A. (1970) Antigenic determinants and antibody complementarity. *Folia Allergol.*, **17**, 425.
19. Mendis,K.N., David,P.H. and Carter,R. (1991) Antigenic polymorphism in malaria: is it an important mechanism for immune evasion? *Immunol. Today*, **12**, A34–A37.
20. Phillips,R.E., Rowland-Jones,S., Nixon,D.F., Gotch,F.M., Edwards,J.P., Ogunlesi,A.O., Elvin,J.G., Rothbard,J.A., Bangham,C.R., Rizza,C.R. *et al.* (1991) Human immunodeficiency virus genetic variation that can escape cytotoxic T cell recognition. *Nature*, **354**, 453–459.
21. Weber,F. and Elliott,R.M. (2002) Antigenic drift, antigenic shift and interferon antagonists: how bunyaviruses counteract the immune system. *Virus Res.*, **88**, 129–136.
22. Shannon,C.E. (1948) The mathematical theory of communication. *Bell Sys. Tech. J.*, **27**, 379–423, 623–656.
23. Simpson,E.H. (1949) Measurement of diversity. *Nature*, **163**, 688.
24. Stewart,J.J., Lee,C.Y., Ibrahim,S., Watts,P., Shlomchik,M., Weigert,M. and Litwin,S. (1997) A Shannon entropy analysis of immunoglobulin and T cell receptor. *Mol. Immunol.*, **34**, 1067–1082.

25. Baczkowski,A.J., Joanes,D.N. and Shamia,G.M. (1998) Range of validity of alpha and beta for a generalized diversity index H (alpha, beta) due to Good. *Math. Biosci.*, **148**, 115–128.
26. Reche,P.A., Glutting,J.-P. and Reinherz,E.L. (2004) Enhancement to the RANKPEP resource for the prediction of peptide binding to MHC molecules using profiles. *Immunogenetics*, **56**, 405–419.
27. Reche,P.A., Glutting,J.P. and Reinherz,E.L. (2002) Prediction of MHC class I binding peptides using profile motifs. *Hum. Immunol.*, **63**, 701–709.
28. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
29. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797. Print 2004.
30. Wu,T.T. and Kabat,E.A. (1970) An analysis of the sequences of the variable regions of Bence Jones proteins and myeloma light chains and their implications for antibody complementarity. *J. Exp. Med.*, **132**, 211–250.
31. Poirot,O., O'Toole,E. and Notredame,C. (2003) Tcoffee@igs: a web server for computing, evaluating and combining multiple sequence alignments. *Nucleic Acids Res.*, **31**, 3503–3506.
32. Calhoun,J.R., Kono,H., Lahr,S., Wang,W., DeGrado,W.F. and Saven,J.G. (2003) Computational design and characterization of a monomeric helical dinuclear metalloprotein. *J. Mol. Biol.*, **334**, 1101–1115.
33. Dahiyat,B.I. and Mayo,S.L. (1997) De novo protein design: fully automated sequence selection. *Science*, **278**, 82–87.
34. Kuhlman,B., Dantas,G., Ireton,G.C., Varani,G., Stoddard,B.L. and Baker,D. (2003) Design of a novel globular protein fold with atomic-level accuracy. *Science*, **302**, 1364–1368.
35. Mignone,F., Horner,D.S. and Pesole,G. (2004) WebVar: A resource for the rapid estimation of relative site variability from multiple sequence alignments. *Bioinformatics*, **20**, 1331–1333.
36. Bui,H.H., Sidney,J., Li,W., Fusseder,N. and Sette,A. (2007) Development of an epitope conservancy analysis tool to facilitate the design of epitope-based diagnostics and vaccines. *BMC Bioinformatics*, **8**, 361.
37. Reche,P.A., Keskin,D.B., Hussey,R.E., Ancuta,P., Gabuzda,D. and Reinherz,E.L. (2006) Elicitation from virus-naive individuals of cytotoxic T lymphocytes directed against conserved HIV-1 epitopes. *Med. Immunol.*, **5**, 1.
38. Zolla-Pazner,S. (2004) Identifying epitopes of HIV-1 that induce protective antibodies. *Nat. Rev. Immunol.*, **4**, 199–210.

7.2 TEPIDAS

7.2.1 Justificación y Objetivos

Es necesario que todos los datos anotados en las bases de datos sean almacenados de manera eficiente y accesible, permitiendo la integración de la información. En este capítulo mostramos TEPIDAS, una base de datos de epitopos que se integra dentro de otras bases de datos utilizando DAS (*Distributed Annotation Systems*) cuyo objetivo es:

- Incluir la información disponible acerca de los epítopos T CD8 junto con la información disponible de sus proteínas fuente en otras bases de datos.

7.2.2 Conclusiones

- TEPIDAS es una herramienta útil que permite acceder a la información de los epítopos T CD8 de manera integrada con otras informaciones y compartirla con otros investigadores, usando DAS.

Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Database**

Integrating T-cell epitope annotations with sequence and structural information using DAS

Carmen M. Diez-Rivero[1], María García-Boronat[1] and Pedro A Reche[1,] *

[1]ImmuneMedicine Group, Department of Microbiology, Division of Immunology, Facultad de Medicina, Universidad Complutense de Madrid, Ave Complutense, s/n. Madrid 28040, Spain;
Pedro A Reche* - Email: parecheg@med.ucm.es; Phone: 34 91 394 7229; Fax: 34 91 394 1641; * Corresponding author

received November 09, 2008; accepted November 23, 2008; published December 06, 2008

Abstract:
Immunoinformatics is an emerging new field that benefits from computational analyses and tools that facilitate the understanding of the immune system. A large number of immunoinformatics resources such as immune-related databases and analysis software are available through the World Wide Web for the benefit of the research community. However, immunoinformatics developments have sometimes remained isolated from mainstream bioinformatics. Therefore, there is clearly a need for integration, which will empower the exchange of data and annotations within the scientific community in a quick and efficient fashion. Here, we have chosen the Distributed Annotation System (DAS), for integrating in house annotations on experimental and predicted HLA I-restriction elements of CD8 T-cell epitopes with sequence and structural information.

Keywords: DAS; annotation; epitope; HLA I

Abbreviations: CMV - Cumulative Phenotypic Frequency; DAS - Distributed Annotation System; HLA I - Human Leukocyte Antigen class I; PSSM - Position Specific Scoring Matrix

Background:
Recent years have witnessed the birth of Immunoinformatics, an emerging subdiscipline of Bioinformatics. With the burgeoning explosion of immunological data, computational analysis has become an essential element of immunology research, facilitating the understanding of the immune function by modeling the interactions among immunological components **[1]**. Another major role in Immunoinformatics is the efficient management, storage, and annotation of such data. Following those principles, a large number of immunoinformatics resources including immune-related databases and sophisticated analysis software, are available through the World Wide Web. Collectively, these resources contribute to the advances made in immunological research. Yet, there is still a major step to be taken towards the integration of all these resources, as ideally, multiple research groups should be able to exchange and compare their data, in a quick and efficient fashion.

The distributed annotation system (DAS) defines a communication protocol used to exchange biological annotations from a number of heterogeneous distributed databases **[2]**. The key idea behind the DAS concept is that annotations should not be provided by single centralized databases but instead be spread over multiple sites. DAS follows a simple http-based client-server protocol, where clients make requests in the form of a URL to the servers, and receive simple XML responses. The basic system is composed of a reference server, one or more annotation servers, and an annotation viewer. The reference server is responsible for serving genome maps, sequences and information related to the sequencing process. Annotation servers are responsible for returning the annotations on a defined region (given a start and stop position coordinates) of the genome or proteome. The annotation viewer can either be a simple web browser, which will visualize the raw XML data provided by the server, or a graphical client which translates the XML annotations such as the Center for Biological Sequence Analysis (CBS) DAS viewer **[3]** accessible at http://www.cbs.dtu.dk/cgi-gin/das.

In this article, we will show how an epitope database can be integrated to other database resources using DAS. For that we will describe TEPIDAS, a DAS Annotation Server of HLA I-restricted CD8 T-cell epitopes specific of human pathogenic organisms. TEPIDAS falls into the category of annotation servers and is registered at the DAS registry since February of 2008, and has the unique id DS_545.

Description:
Overview
TEPIDAS is a DAS annotation server that follows the UniProt coordinates system to annotate the experimental and potential HLA I-restriction elements of a set of CD8 T-cell epitopes. TEPIDAS is implemented using ProServer **[4]**, a lightweight Perl-based DAS server. When a client makes a query to the

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 3(4): 156-158 (2008)

TEPIDAS server, ProServer simply retrieves the relevant information from the relational database and composes the XML response. The annotations in TEPIDAS are pre-calculated and stored in a relational database. The coordinate system defined for TEPIDAS is Uniprot **[5]**, as the "authority", and Protein Sequence, as the "type". As for TEPIDAS capabilities, our server implements the "types" and "features" queries.

Annotations served by TEPIDAS

TEPIDAS annotates the HLA I molecules that can restrict a set of 3250 CD8 T-cell epitopes. Epitopes were obtained from the EPIMHC **[6]** and IMMUNEEPITOPE (http://www.immuneepitope.org/) databases, and were selected to be experimentally defined in humans infected with the pathogen or immunized with the relevant source antigen. HLA I-restriction annotations can be classified as experimental, when determined experimentally, or predicted. Predictions of the epitopes binding HLA I molecules, were obtained using a set of 72 position-specific scoring matrices (PSSMs), also known as weight matrices of profiles, which are obtained from aligned peptides known to bind to the relevant HLA I molecules. This predictive method is described in full detail at **[7]**. In addition to the experimental and predicted data, the cumulative phenotypic frequency (CMV) of the T-cell epitope HLA I restriction is also provided for five ethnic groups (Black, Caucasian, Hispanic, North American natives and Asian). CMV was computed using the gene and haplotype frequencies of the relevant HLA I alleles **[8]**. The potential population protection coverage of a T cell epitope-based vaccine is determined by the percentage of the population that could elicit a T cell response to the epitopes, which in turn is given by the CMV of HLA I molecules restricting these epitopes.

Accessing TEPIDAS from the SPICE graphical client

SPICE **[9]** is a Java program that can be used to visualize annotations of protein sequences and protein structures. It is available at: http://www.efamily.org.uk/software/dasclients/spice. SPICE accepts either a PDB or a UniProt accession code, and integrates information from four different types of DAS servers: 1) a protein sequence server that provides the sequence (typically UniProt), 2) an alignment server that provides the alignment between the protein sequence and its structure, 3) a structure server that serves the 3D coordinates displayed, and 4) several feature servers that provide pre-calculated annotations, as for example TEPIDAS among others.

SPICE retrieves the protein sequence pertaining to the selected UniProt accession number, and displays it as a ruler with relative position numbers. Annotations, such as TEPIDAS annotation features, are listed below the sequence in that figure. On the left of the panel, below the 'tepidas' descriptor, appears the type of HLA I molecule of the corresponding feature shown as a colored rectangle on the right. When the user clicks on a feature, a pop-up window appears, containing all the information of the feature, including the explanatory note. In addition, the PDB coordinates of the selected feature, if available, will be highlighted at the left panel, enabling the location of the epitope at the 3D structure whenever there is a match between sequence and structure (Figure 1).

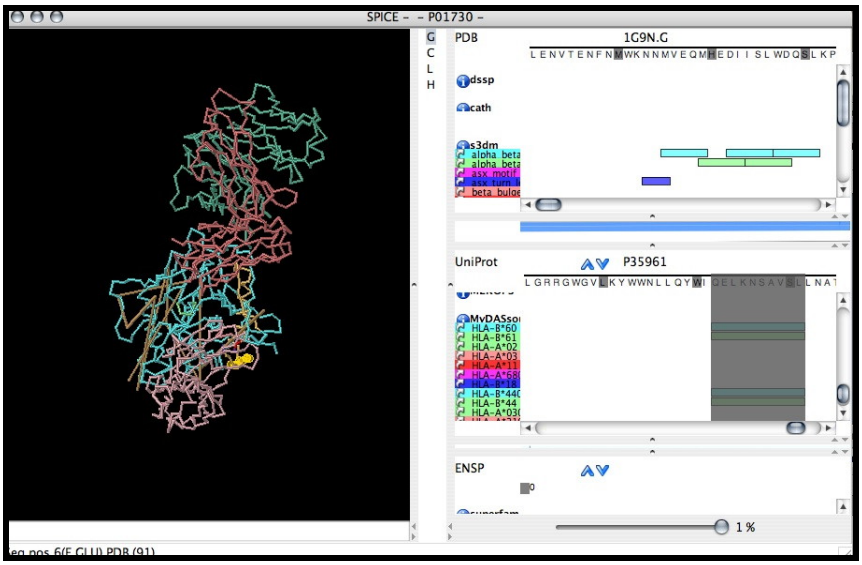


Figure 1: SPICE viewer window. Left panel provides a 3D visualization of the molecule. Right panel displays the annotations provided by the distributed serves. This figure was generated using the UniProt code P35961 as the reference sequence. SPICE's alignment server automatically maps the protein sequence to a 3D structure (1G9N in this example). Feature annotations from TEPIDAS are displayed in the right center panel as rectangular tracks colored as the HLA I molecules on their left under the te -pidas source descriptor.

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 3(4): 156-158 (2008)

Conclusion:

DAS is an important, simple and yet powerful system for exchanging and viewing biological data that is already being used in real-world bioinformatics applications. The TEPIDAS annotation server described in this chapter is a clear example of how epitope data can be integrated and shared by the research community using the DAS architecture. The complexity of immune interactions and the data intensive nature of immune research make Immunoinformatics a suitable area that could greatly benefit from the advantages of using such a powerful integration and annotation system, allowing to gain a more insightful understanding of the complexities of the immune system.

Acknowledgment:

We would like to thank Alfonso Valencia, Osvaldo Graña, and Jaime Fernandez Vera from the Spanish National Cancer Research Center (CNIO) for their helpful advice on DAS and ProServer. Work and authors were supported by grant SAF2006-07879 from the "Ministerio de Educación y Ciencia" of Spain, granted to PAR.

References:

[01] M. N. Davies and D. R. Flower, *Drug Discov. Today*, 12: 389 (2007) [PMID: 17467575]

[02] R. D. Dowell, R.M. *et al*., *BMC Bioinformatics*, 2: 7 (2001) [PMID: 11667947]

[03] P. I. Olason, *Nucleic Acids Res*., 33: W468 (2005) [PMID: 15980514]

[04] R. D. Finn *et al., Bioinformatics* 23: 1568 (2007) [PMID: 17850653]

[05] C. H. Wu *et al*., *Nucleic Acids Res.*, 34: D187 (2006) [PMID: 16381842]

[06] P. A Reche *et al., Bioinformatics*, 21: 2140 (2005) [PMID: 15657103]

[07] P. A Reche and E. L. Reinherz, *Methods Mol. Biol.*, 409:185 (2007) [PMID: 18450001]

[08] P. A Reche *et al., Med. Immunol.*, 5: 1 (2006) [PMID: 16674822]

[09] Prlic *et al., Bioinformatics*, 21: ii40 (2005) [PMID: 16204122]

Edited by P. Kangueane

Citation: Diez-Rivero *et al.*, Bioinformation 3(4): 156-158 (2008)

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 3(4): 156-158 (2008)

8. DISCUSIÓN

8.1 Sumario

Los linfocitos T CD8 son las células del sistema inmunitario adaptativo que se encargan de eliminar las células infectadas o dañadas. Su función depende del reconocimiento de péptidos antigénicos presentados por las moléculas del MHC I en la superficie de dichas células. Estos péptidos proceden de proteínas sintetizadas en el citosol que son degradadas por el proteasoma. El proteasoma genera fragmentos proteicos de distintos tamaños, algunos de los cuales son transportados al interior del ER a través de TAP. Una vez en el interior del ER pueden unirse a las moléculas del MHC I nacientes.

La identificación de aquellos péptidos que son capaces de inducir la respuesta de las células T CD8 tiene gran interés para comprender la patogénesis de las enfermedades y el desarrollo de vacunas e inmunoterapias (Purcell, et al., 2007; Vivona, et al., 2008). En la última década, se han desarrollado modelos *in silico* que facilitan la identificación de epítopos T CD8. Originalmente, la identificación de epítopos recaía exclusivamente en la predicción de péptidos que se unen a las distintas moléculas del MHC I, ya que es el paso más restrictivo de la presentación de antígenos. No obstante, el lugar de corte del proteasoma, así como el transporte de los péptidos mediado por TAP, también son críticos en la presentación antigénica. Por este motivo, en esta Tesis doctoral nos hemos centrado en el desarrollo de métodos capaces de modelar estos otros procesos, integrándolos en la predicción de epítopos T CD8. También hemos desarrollado herramientas que ayudan a seleccionar epítopos T CD8 conservados y recursos que permiten integrar nuestras anotaciones sobre los epítopos T CD8 en otros recursos con información funcional y estructural de los antígenos fuente.

8.2 Distribución de epítopos T CD8

Una primera aproximación que facilite el desarrollo de vacunas de epítopos sería definir si los epítopos T CD8 se distribuyen de manera preferencial en los antígenos de los patógenos.

Por ello, aquí hemos analizando la distribución de los epítopos T CD8, que están anotados en distintas bases de datos, en las proteínas de HCV, HIV e IAV. En principio, cuanto más larga sea la secuencia de una proteína más epítopos T CD8 podrá contener. Pero, ¿es realmente así?

A simple vista, la representación gráfica de la distribución de los epítopos en sus proteínas fuente no parece indicar ninguna istribución preferencial. En general, las proteínas más largas contienen más epítopos (Fig. 1; capítulo I). Sin embargo, al analizar su distribución mediante el test χ^2, observamos que los epítopos T CD8 no estaban distribuidos homogéneamente de acuerdo al tamaño de las proteínas. De hecho, el número de epítopos que se localizan en las proteínas estructurales, Core de HCV, Gag de HIV y M1 de IAV, es mucho mayor del esperado, mientras que el número de epítopos en las proteínas no estructurales es mucho menor que el que cabría esperar (Fig. 2; capítulo I). Además, al analizar, mediante el test χ^2, la distribución de los péptidos predichos presentados por distintas moléculas del MHC I, observamos que éstos se distribuyen homogéneamente de acuerdo a la longitud de las proteínas (Tabla 2; capítulo I). De modo que la esta distribución no homogénea de los epítopos no parece reflejar ningún patrón de preferencia de unión a moléculas del MHC I.

El número de epítopos que se puede observar en una determinada proteína está, de algún modo, condicionado por la variabilidad de dicha proteína. La identificación experimental de epítopos T requiere la activación de la respuesta de células T por parte de péptidos sintéticos y por lo tanto es muy probable que las respuestas frente a epítopos variables no sean detectadas (Chang, et al., 2011). Por ello, los epítopos en proteínas variables están infrarrepresentados y podría dar lugar a la distribución de peítopos que observamos. No es el caso, ya que no observamos ninguna correlación entre la distribución de los epítopos y la conservación de las secuencias (Fig. 5; capítulo I).

Dado que los datos de epítopos T CD8 de los que partimos son epítopos que han sido verificados de manera experimental, no podemos descartar que nuestros resultados reflejen un sesgo marcado por el interés de los investigadores. Las distintas líneas de investigación pueden

estar centradas en el estudio de proteínas concretas, haciendo que en dichas proteínas el número de epítopos observados sea mucho mayor que en otras.

Nosotros consideramos que existen otras causas que expliquan la distribución observada. Las proteínas Core de HCV, Gag de HIV y M1 de IAV, en las que se localiza un número de epítopos mucho mayor del esperado, están localizadas al comienzo de ORFs (*Opend Reading Frames*) que incluyen varias proteínas junto a las que se traducen. El caso más claro es el de HCV, cuyo genoma se traduce en una única poliproteína en la que la proteína Core se localiza en el extremo N-terminal. Se ha visto que el proceso de traducción de proteínas virales es my ineficiente y que frecuentemente resulta en productos proteicos defectuosos (Khan, et al., 2001; Princiotta, et al., 2003; Princiotta, et al., 2001; Qian, et al., 2006; Schubert, et al., 2000; Yewdell, et al., 1996). Además, se sabe que la presentación de antígenos por las moléculas del MHC I está relacionada con la biosíntesis de proteínas (Princiotta, et al., 2003; Reits, et al., 2000; Schubert, et al., 2000). Por ello nosotros encontramos más epítopos en las proteínas localizadas al comiendo de un ORF, ya que éstas son translocadas de manera prioritaria y la presentación de péptidos por las moléculas del MHC I está estrechamente relacionada con la biosíntesis de proteínas. La localización estratégica de ciertas proteínas al comienzo de un ORF es una manera de garantizar un alto número de copias de estas proteínas. Un mecanismo semejante al propuesto aquí, de regulación a nivel transcripcional ha sido descrito en RNA virus de cadena negativa, pero, hasta donde hemos visto, este control translacional basado en la posición de las proteínas es la primera vez que se describe y necesita una confirmación experimental.

Los resultados obtenidos parecen ser algo paradójicos, ya que el número de epítopos que observamos en las proteínas estructurales Core y Gag de HCV y HIV, respectivamente, es mucho mayor del esperado, y aún así el sistema inmunológico no es capaz de combatir eficazmente estos virus (Bowen and Walker, 2005; Sagar, 2010). Este hecho puede deberse a que aunque las respuestas de los linfocitos T CD8 son esenciales para contener la infección viral, pueden no ser suficientes para eliminar el virus. Por otro lado, el número de epítopos no refleja la

inmunogenicidad de éstos y los linfocitos T pueden tener como dianas otros epítopos T variables. En cualquier caso, no se debe relegar la importancia que los epítopos subdominantes tienen, ya que la inmunodominancia se puede revertir a través de la vacunación (Eberl, et al., 1996; Sandberg, et al., 1998).

8.3 Modelado del proteasoma

El proteasoma tiene un papel fundamental en el procesamiento de antígenos presentados por moléculas del MHC I. Aunque existen algunas peptidasas, como TPPII que pueden tener un papel importante en la generación de algunos péptidos presentados por las moléculas del MHC I (Kloetzel, 2004; Reits, et al., 2004; Yewdell and Princiotta, 2004), se ha visto que el extremo C-terminal de los péptidos que se unen a las moléculas del MHC I es el resultado del corte por el proteasoma (Goldberg, et al., 2002; Pamer and Cresswell, 1998; Rock, et al., 1994).

El modelado de la especificidad proteolítica del proteasoma ya ha sido abordado por varios grupos investigadores. Los primeros métodos que se desarrollaron empleaban fragmentos de enolasa y proteína β-caseína producidos *in vitro* por el proteasoma constitutivo humano (Kuttler, et al., 2000; Nussbaum, et al., 2001). Del mismo modo, se desarrolló un modelo cinético de la actividad proteolítica del proteasoma empleando fragmentos originados en digestiones *in vitro* (Holzhutter, et al., 1999; Holzhutter and Kloetzel, 2000). Sendos modelos son específicos para el proteasoma constitutivo 20S, ya que ésta fue la forma del proteasoma empleada para generar los fragmentos. Posteriormente, se generaron modelos predictivos del corte por el proteasoma que emplean péptidos restringidos por moléculas del MHC I y la región que flanquea al extremo C-termial en la proteína de origen (Bhasin and Raghava, 2005; Kesmir, et al., 2002; Nielsen, et al., 2005). Estos métodos parecen mejorar las predicciones de corte por el proteasoma respecto a los primeros modelos que se entrenaban con fragmentos de digestiones *in vitro* (Saxova, et al., 2003). Sin embargo, los métodos entrenados con los sitios de corte

generados experimentalmente parecen ser más idóneos para la identificación del tamaño de los fragmentos generados por el proteasoma (Tenzer, et al., 2005).

La mayoría de los modelos desarrollados hasta la fecha utilizan un único grupo de péptidos que principalmente han sido generados por el proteasoma constitutivo. Sin embargo, existe otra forma del proteasoma, el inmunoproteasoma, cuyo patrón de corte es distinto al del proteasoma constitutivo pero solapante (Gaczynska, et al., 1994). En general, el inmunoproteasoma se expresa de manera constitutiva en las células dendríticas, mientras que el proteasoma es la forma constitutiva presente en el resto de células. Por todo ello, se podría concluir que los epítopos T CD8 protectivos son aquellos que pueden ser generados por ambos, el inmunoproteasoma y el proteasoma constitutivo (Fig. 11) (Chapiro, et al., 2006).

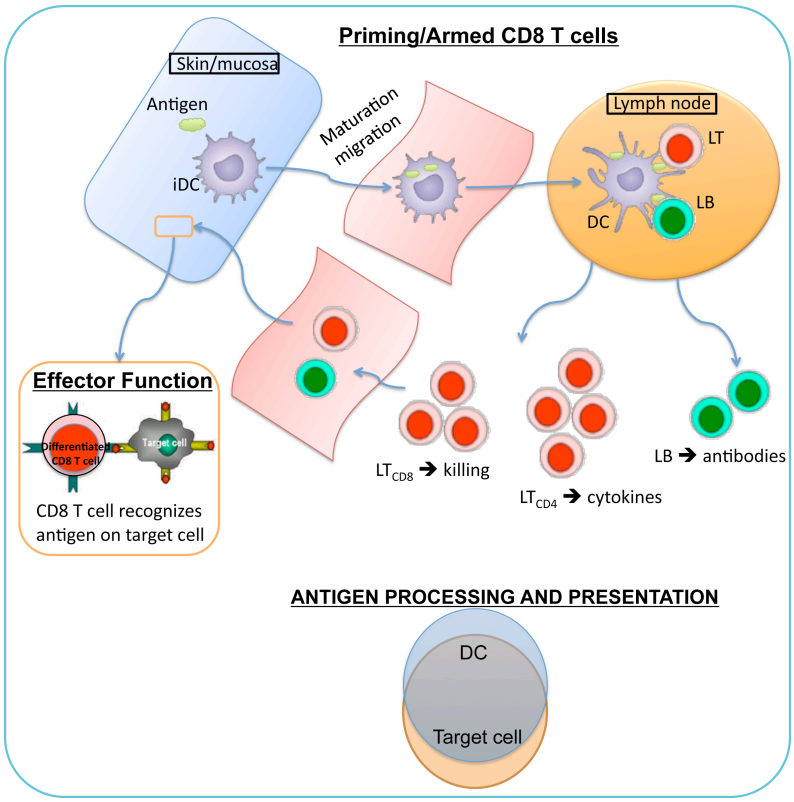


Figure 11. Antigen presentation by MHC I. The immunoproteasome produces different but overlapping cleavage patterns with regard to those of the proteasome. The immunoproteasome is constitutively expressed in dendritic cells, which prime CD8 T cells acquiring their cytotoxic function, whereas, the proteasome is the form expressed in other cells. Therefore, it is likely that protective epitopes are those generated by both, the proteasome and the immunoproteasome.

En esta Tesis, hemos desarrollado dos modelos distintos, uno para predecir los sitios de corte del proteasoma constitutivo y otro para los del inmunoproteasoma. Para ello, asumimos que los péptidos eluidos de las moléculas del MHC I conservan el corte proteolítico realizado por el proteasoma, mientras que los epítopos T CD8 identificados en pacientes durante el transcurso de la infección reflejan el corte realizado por el inmunoproteasoma. Los modelos se desarrollaron usado *N-grams* entrenados con fragmentos del péptidos de distinto tamaño; y se evaluaron mediante validación cruzada (ver Material y Métodos) empleando como parámetros BTR (Better Than Random), que permite ver la mejora del modelo respecto a un modelo aleatorio, y MCC. Los resultados de la validación cruzada muestran que los modelos del proteasoma constitutivo predicen mejor los sitios de corte que los del inmunoproteasoma. El mejor modelo del inmunoproteasoma alcanzó un BTR = 0.53 ± 0.02 y un MCC = 0.43 ± 0.07, mientras que el mejor modelo del inmunoproteasoma obtuvo un BTR = 0.44 ± 0.01 y un MCC = 0.36 ± 0.06. Esto refleja, posiblemente, el hecho de que el grupo de epítopos T CD8 era mucho más numeroso y por tanto más diverso que el de péptidos eluidos de moléculas del MHC I. Además, se ha visto que las células dendríticas presentan vías de procesamiento de antígenos alternativas, y algunas de ellas, incluso independientes del inmunoproteasoma (Banchereau, et al., 2000; Heath, et al., 2004), lo que provocaría una mayor variabilidad entre los epítopos con los que se entrenó el modelo.

A juzgar por los resultados de la validación cruzada, la capacidad de predicción de los modelos aumenta a medida que aumenta el tamaño de los fragmentos con los que se entrenan los modelos, alcanzándose los mejores resultados en aquellos modelos entrenados con fragmentos de 12 residuos (Fig 1A, Tabla I; capítulo II). Al entrenar los modelos con fragmentos de péptidos de mayor tamaño la capacidad predictiva de los modelos no mejoró significativamente, e incluso disminuyó en algunos casos. Por eso, los modelos seleccionados en este trabajo y por tanto a los que haremos referencia de aquí en adelante, tanto para el proteasoma constitutivo como para el inmunoproteasoma, fueron aquellos que se habían entrenado con fragmentos de péptidos de 12

residuos, compuestos por los 6 residuos del extremo C-terminal del péptido (P1-P6) y los 6 más próximos al otro lado de corte (P1'-P6'). Estos resultados son consistentes con los trabajos experimentales en los que se indica que el proteasoma y el inmunoproteasoma examinan entre 5 y 6 residuos a cada lado del sitio de corte (Altuvia and Margalit, 2000; Nussbaum, et al., 1998). En comparación, otros modelos predictivos basados en péptidos presentados por moléculas del MHC I han sido entrenados con fragmentos de mayor tamaño (18 – 20 residuos), que hacen que estos modelos sean un tanto artificiales independientemente de sus resultados.

Los modelos aquí desarrollados para ambos proteasomas producen patrones de fragmentación distintos pero solapantes que reflejan, de algún modo, los patrones descritos experimentalmente (Toes, et al., 2001); el 68% de los sitios de corte (residuos P1) y el 36% de los fragmentos generados eran idénticos (Fig. 3B; capítulo II). La mayoría de los fragmentos generados por ambos modelos son de pequeño tamaño (2 – 3 residuos), también de manera experimental se ha visto que dos tercios de los fragmentos producidos por el proteasoma tienen un tamaño menor del necesario para ser presentados por moléculas del MHC I (Kisselev, et al., 1999). Aún así, es importante destacar que nuestros modelos no tratan de predecir los posibles fragmentos proteicos, sino indicar si el extremo C-terminal de un péptido puede ser el resultado de la actividad catalítica del proteasoma y/o el inmunoproteasoma.

Los modelos también fueron evaluados mediante un test independiente, utilizando un grupo de epítopos T CD8 de HIV-1 que no se empleó para entrenar los modelos. Esta evaluación mostró que el modelo del inmunoproteasoma obtiene mejores resultados que el modelo del proteasoma (Fig. 2; capítulo II). Por ello, y teniendo en cuenta todo lo anterior, el modelo del inmunoproteasoma, entrenado con epítopos T CD8, parece ser más adecuado para la predicción del extremo C-terminal de los epítopos T CD8.

Por otra parte, el test independiente también se utilizó para comparar nuestros modelos con NetChop, el mejor modelo de predicción del corte por el proteasoma descrito hasta la fecha. Atendiendo a los valores de MCC alcanzados por el modelo del inmunoproteasoma, el del

proteasoma y NetChop (MCC = 0.20, MCC = 0.19 y MCC = 0.18, respectivamente) podemos decir que nuestros modelos son tan buenos o mejores que NetChop (Fig. 4; capítulo II). Sin embargo, es importante tener en cuenta que estos resultados se obtuvieron utilizando las condiciones óptimas indicadas para NetChop. Primero, dado que NetChop fue entrenado utilizando fragmentos que contenían la secuencia completa del péptido, aquí hemos evaluado y comparado la capacidad predictiva de los distintos modelos sobre la secuencia completa del péptido aunque nuestros modelos óptimos tan sólo utilizan un fragmento del péptido formado por los 6 residuos del extremo C-terminal. Segundo, es probable que los epítopos TCD8 de HIV-1 que nosotros utilizamos como un conjunto de datos independiente, si hayan sido incluidos en los datos de entrenamiento de NetChop.

Los modelos para la predicción del sitio de corte por el proteasoma constitutivo y el inmunoproteasoma están disponibles en http://imed.med.ucm.es/Tools/pcps/index.html. En esta herramienta no sólo están disponibles los mejores modelos aquí desarrollados, los entrenados con 6 aminoácidos a cada lado del punto de corte, sino que también se pueden elegir otros dos modelos del proteasoma constitutivo y del inmunoproteasoma con distinta sensibilidad y especificidad.

8.4 Predicción de epítopos T CD8 combinando la predicción de sitios de unión a moléculas del MHC I y la predicción de sitios de corte por el proteasoma y/o el inmunoproteasoma.

La base para predecir epítopos T CD8 es la predicción de los sitios de unión a las moléculas del MHC I. Cuando se utilizan solamente las PSSMs (ver Material y Métodos) para ver los sitios de unión a las moléculas del MHC se obtienen predicciones muy precisas, más del 80% de los epítopos reales son identificados entre el 2% de los péptidos con más puntuación (Reche, et al., 2002). En este caso, hemos realizado predicciones de unión a las moléculas del

MHC I A*0201, A*0301, A*2402, B*0702 y B*2705, que por si solas, alcanzan valores de AUC (ver Material y Métodos) mayores de 0.9. Para tratar de mejorar estos resultados hemos combinado la predicción de unión a moléculas del MHC I con la predicción de sitios de sitios de corte por el proteasoma constitutivo y/o el inmunoproteasoma. Para ello realizamos un filtrado en el que aquellos péptidos que no posean un extremo C-terminal predicho como sitio de corte son eliminados.

Con todas las posibles combinaciones se obtiene una mejora en la identificación de epítopos T CD8 estadísticamente significativa ($p < 0.05$) (Fig. 5; capítulo II). Las mejoras obtenidas se deben principalmente a una gran reducción del número de falsos positivos detectados (más del 70%). Por lo tanto, la combinación de los modelos de corte por el proteasoma y/o inmunoproteasoma con las predicciones de unión a moléculas del MHC I permiten disminuir el trabajo experimental que la identificación de epítopos conlleva dado que habrá menos péptidos que necesitan ser probados. Es cierto, que la combinación con el modelo del proteasoma solo o junto con el del inmunoproteasoma suponen una pérdida significativa de verdaderos positivos (más del 20%), por ello, esta combinación de modelos predictivos será más útil cuando se quieran predecir los epítopos de un gran número de antígenos.

8.5 Modelado de la unión de péptidos a TAP

El transporte mediado por TAP tiene dos pasos secuenciales, primero el péptido se une a TAP y posteriormente es translocado al interior del ER consumiendo ATP. Ésta es una de las etapas obligadas de la vía clásica de presentación de antígenos por las moléculas del MHC I. Esto hace que sea una etapa importante, y por ello distintos grupos han desarrollado modelos computacionales para predecir y analizar la afinidad de unión de los péptidos a TAP. Estos métodos están basados en "*artificial neural networks"* (ANNs) (Brusic, et al., 1999; Daniel, et al., 1998; Zhang, et al., 2006), en "*support vetor machines*" (SVMs) (Bhasin and Raghava, 2004; Donnes and Kohlbacher, 2005) y en matrices generadas utilizando el "*Stabilized Matrix Method*"

(Peters, et al., 2003) y el "*additive method*" (Doytchinova, et al., 2004; Doytchinova, et al., 2002). Hasta la fecha, la mayoría de los modelos desarrollados se basan en el mismo grupo de datos consistente en 435 péptidos de 9 aminoácidos cuya afinidad a TAP es conocida y que han sido hechos públicos por el Dr. Peter Van Endert (Peters, et al., 2003). En este trabajo, hemos utilizado un grupo de datos mayor, que incluye 178 péptidos nuevos, para estudiar la selectividad de TAP de manera cuantitativa.

El modelo aquí desarrollado está basado en los modelos de regresión de SVM entrenados con residuos individuales de los péptidos o en combinación, generando numerosos modelos que se evaluaron mediante el coeficiente de correlación de Pearson (R_p) (ver Material y Métodos). Anteriormente, se pensaba que sólo las posiciones P1, P2, P3 y el extremo C-terminal eran relevantes en la unión del péptido a TAP (Bhasin and Raghava, 2004; Doytchinova, et al., 2004; Peters, et al., 2003; Uebel, et al., 1997; van Endert, et al., 1995). Sin embargo, en los modelos que hemos desarrollado aquí utilizando de manera independiente cada una de las posiciones del péptido se ha visto que todos los residuos del péptido contribuyen en la unión a TAP y que la contribución del residuo P4 es equivalente a la del residuo P3. Asimismo, hemos confirmado que el residuo del extremo C-terminal del péptido es el que más contribuye, de manera cuantitativa, a la unión a TAP, ya que el modelo entrenado sólo con este residuo alcanzó un $R_p = 0.68 \pm 0.06$.

Cuando los modelos se entrenan con distintas combinaciones de los residuos del péptido vemos que la mitad del extremo N-terminal del péptido contribuye más a la unión del péptido a TAP que la mitad del extremo C-terminal (Fig. 2; capítulo III). Los mejores resultados se obtienen con los modelos entrenados con fragmentos de 8 residuos consistentes en los primeros 5 residuos del extremo N-terminal y los últimos 3 residuos del extremo C-terminal (5N3C) de los péptidos ($R_p = 0.89 \pm 0.03$). Estos resultados son iguales que cuando el modelo se entrena con la secuencia completa de los péptidos (TAP_{613}) ($R_p = 0.89 \pm 0.03$) (Fig. 1B y Fig. 2; capítulo III). Esto refleja las observaciones de que TAP transporta péptidos de 8 y 9 aminoácidos con una eficiencia comparable (Androlewicz and Cresswell, 1994; Momburg, et al., 1994).

La correlación alcanzada por estos dos modelos, TAP_{613} y 5N3C, es mayor que la del resto de modelos de afinidad de TAP desarrollados (Bhasin and Raghava, 2004; Donnes and Kohlbacher, 2005; Doytchinova, et al., 2004; Peters, et al., 2003). Es importante destacar que los buenos resultados alcanzados por nuestros modelos no se deben a una redundancia de los datos, de hecho hemos visto que péptidos con secuencias altamente similares suelen tener distinta afinidad a TAP (Fig. 3; capítulo III).

En esta Tesis, también hemos desarrollado una herramienta web, TAPREG, para predecir la afinidad de los péptidos a TAP que está disponible para uso público en http://imed.med.ucm.es/Tools/tapreg/. TAPREG puede usarse no sólo para predecir la afinidad de unión a TAP de péptidos de 8 ó 9 aminoácidos, sino también para péptidos de mayor tamaño, hasta 16 residuos, correspondiente con la actividad transportadora de TAP (Androlewicz and Cresswell, 1994; Momburg, et al., 1994). Hasta ahora la afinidad de péptidos de más de 9 aminoácidos sólo podía calcularse utilizando matrices cuantitativas, y se consideraba que sólo los 3 últimos residuos del extremo N-terminal y el residuo del C-terminal eran importantes para la unión a TAP (Peters, et al., 2003). Por el contrario, en TAPREG, se seleccionan 9 residuos de los péptidos de mayor tamaño, ya que se ha visto que todos los residuos de los péptidos de 9 aminoácidos contribuyen en la unión a TAP. Este es el primero modelo que puede predecir la unión de péptidos a TAP de péptidos de más de 9 aminoácidos.

8.6 PVS

Con el objetivo de facilitar la identificación de epítopos, tanto de linfocitos T como B, hemos desarrollado la herramienta PVS (Protein Variabiliby Server) (http://imed.med.ucm.es/PVS/). PVS es un servidor web que permite calcular la variabilidad absoluta de la secuencia estimada para cada posición dentro de un MSA y determinada según la entropía de Shannon (Shannon, 1948; Simpson, 1949), el índice de diversidad de Simpson (Simpson, 1949) y el coeficiente de variabilidad de Wu-Kubat (Kabat, 1970).

La importancia de los residuos variables siempre ha quedado relegada a un segundo plano. Sin embargo, los residuos variables también son funcionalmente importantes dando lugar a la heterogenicidad de las secuencias, por ejemplo, los residuos más variables de las moléculas del MHC I están en las subcavidades donde se unen los péptidos, por eso las distintas moléculas del MHC I presentan especificidades de unión a péptidos distintas (Reche and Reinherz, 2003; Stern and Wiley, 1994). Por todo ello, el análisis de la variabilidad de las secuencias se ha empleado para tratar de dar respuestas a las incógnitas de las estructuras de las proteínas y su función (Reche and Reinherz, 2003; Wu and Kabat, 1970).

PVS, no sólo permite calcular la variabilidad de las secuencias, sino que además sus resultados pueden facilitar el diseño de vacunas basadas en epítopos mediante la identificación de epítopos conservados. Este tipo de aproximación tiene muchas ventajas en aquellos microorganismos que emplean la variabilidad de sus secuencias como método para evadir la respuesta inmunitaria. En este contexto es importante ver dónde se localiza la variabilidad. Por ejemplo, al utilizar la secuencia de la proteína gp120 de HIV-1 con los residuos variables enmascarados obtenida con la herramienta PVS, RANKPEP predice tan sólo dos epítopos T CD8 conservados restringidos por moléculas A*0201 (KLTPLCVTL y PVVSTQLLL), mientras que si se emplea la secuencia H2XB2 de la proteína gp120 de HIV-1, RANKPEP devuelve 10 posibles epítopos. Por ello, independientemente de la capacidad predictiva de RANKPEP, es evidente que esta estrategia ahorra tiempo y trabajo necesario para encontrar epítopos T CD8 conservados.

Por otro lado, PVS también sirve como herramienta para la identificación de epítopos B conservados. Los anticuerpos sólo reconocen epítopos expuestos al exterior, y muchos de ellos son conformacionales, aunque también pueden ser lineales. PVS puede utilizarse para identificar los fragmentos conservados en la estructura en 3D de las proteínas e identificar aquellos que quedan expuestos al exterior. La utilización de fragmentos conservados que estén expuestos al exterior como inmunógenos aumenta de manera considerable la posibilidad de producir anticuerpos que presenten tanto actividad cruzada con el antígeno natural como una amplia

capacidad neutralizadora. Como se explica en el ejemplo del capítulo IVa, la proteína gp41 es la diana de varios anticuerpos neutralizantes. Cuando se hace el análisis con PVS del alineamiento de secuencias de esta proteína se obtienen 7 fragmentos conservados de al menos 6 residuos (Table 2; capítulo IVa). Entre estos fragmentos se encuentran WGCSGK y WLWYIK que son conocidas dianas de estos anticuerpos neutralizantes.

8.7 TEPIDAS

La integración de la información anotada sobre distintas materias científicas es fundamental para poder llevar a cabo un intercambio eficaz de los conocimientos entre los distintos grupos investigadores. Sin embargo, la inmunoinformática ha estado, tradicionalmente, aislada de la bioinformática. De hecho, hasta la fecha, no existen datos de epítopos recogidos junto al resto de anotaciones en las bases de datos.

Recientemente se ha desarrollado un sistema conocido como DAS que permite integrar toda esta información. Utilizando este sistema DAS hemos desarrollado TEPIDAS, que integra las anotaciones de los epítopos T CD8 junto con la información de sus proteínas de origen. TEPIDAS proporciona información sobre epítopos T CD8 de acuerdo a las moléculas del MHC I que los restringen, incluyendo datos sobre moléculas del MHC I predichas, la localización del epítopo dentro de la secuencia de la proteína de procedencia y la frecuencia fenotípica acumulativa de la molécula del MHC I.

9. CONCLUSIONES

1. CD8 T cell epitopes are preferentially located in viral structural proteins making the capsid and matrix of the viruses.

2. Prioritization of structural proteins as a source of CD8 T cell epitopes could save time and resources for experimental identification of epitopes.

3. We have developed two different proteasome cleavage prediction models trained with two different datasets, using *N-grams*. The constitutive proteasome models were built upon MHC I-eluted peptides whereas the immunoproteasome models were trained with a set of CD8 T cell epitopes. These models mirror the different cleavage patterns of both forms of proteasomes.

4. The proteasome and immunoproteasome models that exhibited the best performance were trained, in both cases, on 12-residue peptide fragments, 6 residues at each side of the cleavage site, which is the fragment size reported to be accessible for the proteasome.

5. Combining cleavage predictions by the proteasome and immunoproteasome models with MHCI-binding predictions improves CD8 T cell epitope discovery rate.

6. We have confirmed that the residue at the C-terminal end of the peptide has the largest quantitative input to TAP binding.

7. TAP affinity models trained with each residue individually showed that each epitope position has a quantitative contribution to TAP binding, and the contribution of the P4 residue is equivalent to that of the P3 residue.

8. The N-terminal half of the peptide has a larger contribution to TAP binding that the C-terminal half of the peptide.

9. Optimal modeling of the binding affinity to TAP was achieved by SVM models trained

on the full-length peptide sequences or on 8-residue fragments consisting of the first five N-terminal and the last three C-terminal residues (5N3C) of the peptides.

10. Sequence variability analyses performed by PVS are useful not only to facilitate structure-function studies, but also to facilitate the discovery of conserved B- and T- cell epitopes.

11. TEPIDAS annotation server is a handy tool to access to epitope data integrated with information of their source proteins and to share it to other investigators using DAS.

10. REFERENCIAS

Abele, R. and Tampe, R. (2004) The ABCs of immunology: structure and function of TAP, the transporter associated with antigen processing, *Physiology (Bethesda)*, **19**, 216-224.

Akdis, C.A., Akdis, M., Blesken, T., Wymann, D., Alkan, S.S., Muller, U. and Blaser, K. (1996) Epitope-specific T cell tolerance to phospholipase A2 in bee venom immunotherapy and recovery by IL-2 and IL-15 in vitro, *J Clin Invest*, **98**, 1676-1683.

Alarcon, B., Gil, D., Delgado, P. and Schamel, W.W. (2003) Initiation of TCR signaling: regulation within CD3 dimers, *Immunol Rev*, **191**, 38-46.

Altuvia, Y. and Margalit, H. (2000) Sequence signals for generation of antigenic peptides by the proteasome: implications for proteasomal cleavage mechanism, *J Mol Biol*, **295**, 879-890.

Androlewicz, M.J. and Cresswell, P. (1994) Human transporters associated with antigen processing possess a promiscuous peptide-binding site, *Immunity*, **1**, 7-14.

Androlewicz, M.J., Ortmann, B., van Endert, P.M., Spies, T. and Cresswell, P. (1994) Characteristics of peptide and major histocompatibility complex class I/beta 2-microglobulin binding to the transporters associated with antigen processing (TAP1 and TAP2), *Proc Natl Acad Sci U S A*, **91**, 12716-12720.

Banchereau, J., Briere, F., Caux, C., Davoust, J., Lebecque, S., Liu, Y.J., Pulendran, B. and Palucka, K. (2000) Immunobiology of dendritic cells, *Annu Rev Immunol*, **18**, 767-811.

Bangia, N., Lehner, P.J., Hughes, E.A., Surman, M. and Cresswell, P. (1999) The N-terminal region of tapasin is required to stabilize the MHC class I loading complex, *Eur J Immunol*, **29**, 1858-1870.

Baumeister, W., Walz, J., Zuhl, F. and Seemuller, E. (1998) The proteasome: paradigm of a self-compartmentalizing protease, *Cell*, **92**, 367-380.

Beekman, N.J., van Veelen, P.A., van Hall, T., Neisig, A., Sijts, A., Camps, M., Kloetzel, P.M., Neefjes, J.J., Melief, C.J. and Ossendorp, F. (2000) Abrogation of CTL epitope processing by single amino acid substitution flanking the C-terminal proteasome cleavage site, *J Immunol*, **164**, 1898-1905.

Bernhard E. Boser, Isabelle M. Guyon and Vapnik, V.N. (1992) A training algorithm for optimal margin classifiers., *In Proceedings of the fifth annual workshop on Computational learning theory (COLT '92). ACM, New York, NY, USA*, 144-152.

Bhasin, M. and Raghava, G.P. (2004) Analysis and prediction of affinity of TAP binding peptides using cascade SVM, *Protein Sci*, **13**, 596-607.

Bhasin, M. and Raghava, G.P. (2005) Pcleavage: an SVM based method for prediction of constitutive proteasome and immunoproteasome cleavage sites in antigenic sequences, *Nucleic Acids Res*, **33**, W202-207.

Bonilla, F.A. and Oettgen, H.C. (2010) Adaptive immunity, *J Allergy Clin Immunol*, **125**, S33-40.

Bowen, D.G. and Walker, C.M. (2005) Adaptive immune responses in acute and chronic hepatitis C virus infection, *Nature*, **436**, 946-952.

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

Burges, C.J.C. (1998) A Tutorial on Support Vector Machines for Pattern Recognition, 121-167.

Call, M.E. and Wucherpfennig, K.W. (2005) The T cell receptor: critical role of the membrane environment in receptor assembly and function, *Annu Rev Immunol*, **23**, 101-125.

Call, M.E. and Wucherpfennig, K.W. (2007) Common themes in the assembly and architecture of activating immune receptors, *Nat Rev Immunol*, **7**, 841-850.

Chang, C.X., Dai, L., Tan, Z.W., Choo, J.A., Bertoletti, A. and Grotenbreg, G.M. (2011) Sources of diversity in T cell epitope discovery, *Front Biosci*, **17**, 3014-3035.

Chang, S.C., Momburg, F., Bhutani, N. and Goldberg, A.L. (2005) The ER aminopeptidase, ERAP1, trims precursors to lengths of MHC class I peptides by a "molecular ruler" mechanism, *Proc Natl Acad Sci U S A*, **102**, 17107-17112.

Chapiro, J., Claverol, S., Piette, F., Ma, W., Stroobant, V., Guillaume, B., Gairin, J.E., Morel, S., Burlet-Schiltz, O., Monsarrat, B., Boon, T. and Van den Eynde, B.J. (2006) Destructive cleavage of antigenic peptides either by the immunoproteasome or by the standard proteasome results in differential antigen presentation, *J Immunol*, **176**, 1053-1061.

Ciechanover, A., Orian, A. and Schwartz, A.L. (2000) Ubiquitin-mediated proteolysis: biological regulation via destruction, *Bioessays*, **22**, 442-451.

Craiu, A., Akopian, T., Goldberg, A. and Rock, K.L. (1997) Two distinct proteolytic processes in the generation of a major histocompatibility complex class I-presented peptide, *Proc Natl Acad Sci U S A*, **94**, 10850-10855.

Cristianini, N. and Shawe.Taylor, J. (2000) An introducion to support vector machines and other kernel-based learning methods, *Cambridge University Press, Cambridge*.

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

Dave, V.P. (2011) Role of CD3epsilon-mediated signaling in T-cell development and function, *Crit Rev Immunol*, **31**, 73-84.

Del-Val, M. and Lopez, D. (2002) Multiple proteases process viral antigens for presentation by MHC class I molecules to CD8(+) T lymphocytes, *Mol Immunol*, **39**, 235-247.

Donnes, P. and Kohlbacher, O. (2005) Integrated modeling of the major events in the MHC class I antigen processing pathway, *Protein Sci*, **14**, 2132-2140.

Doytchinova, I., Hemsley, S. and Flower, D.R. (2004) Transporter associated with antigen processing preselection of peptides binding to the MHC: a bioinformatic evaluation, *J Immunol*, **173**, 6813-6819.

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

Draenert, R., Altfeld, M., Brander, C., Basgoz, N., Corcoran, C., Wurcel, A.G., Stone, D.R., Kalams, S.A., Trocha, A., Addo, M.M., Goulder, P.J. and Walker, B.D. (2003) Comparison of overlapping peptide sets for detection of antiviral CD8 and CD4 T cell responses, *J Immunol Methods*, **275**, 19-29.

Eberl, G., Kessler, B., Eberl, L.P., Brunda, M.J., Valmori, D. and Corradin, G. (1996) Immunodominance of cytotoxic T lymphocyte epitopes co-injected in vivo and modulation by interleukin-12, *Eur J Immunol*, **26**, 2709-2716.

Eisenlohr, L.C., Yewdell, J.W. and Bennink, J.R. (1992) Flanking sequences influence the presentation of an endogenously synthesized peptide to cytotoxic T lymphocytes, *J Exp Med*, **175**, 481-487.

Elliott, T. and Williams, A. (2005) The optimization of peptide cargo bound to MHC class I molecules by the peptide-loading complex, *Immunol Rev*, **207**, 89-99.

Evnouchidou, I., Momburg, F., Papakyriakou, A., Chroni, A., Leondiadis, L., Chang, S.C., Goldberg, A.L. and Stratikos, E. (2008) The internal sequence of the peptide-substrate determines its N-terminus trimming by ERAP1, *PLoS One*, **3**, e3658.

Finley, D. and Chau, V. (1991) Ubiquitination, *Annu Rev Cell Biol*, **7**, 25-69.

Frank, E., Hall, M., Trigg, L., Holmes, G. and Witten, I.H. (2004) Data mining in bioinformatics using Weka, *Bioinformatics*, **20**, 2479-2481.

Gaczynska, M., Rock, K.L., Spies, T. and Goldberg, A.L. (1994) Peptidase activities of proteasomes are differentially regulated by the major histocompatibility complex-encoded genes for LMP2 and LMP7, *Proc Natl Acad Sci U S A*, **91**, 9213-9217.

Garbi, N., Tan, P., Momburg, F. and Hammerling, G.J. (2001) Role of tapasin in MHC class I antigen presentation in vivo, *Adv Exp Med Biol*, **495**, 71-78.

Garbi, N., Tiwari, N., Momburg, F. and Hammerling, G.J. (2003) A major role for tapasin as a stabilizer of the TAP peptide transporter and consequences for MHC class I expression, *Eur J Immunol*, **33**, 264-273.

Garcia, K.C., Teyton, L. and Wilson, I.A. (1999) Structural basis of T cell recognition, *Annu Rev Immunol*, **17**, 369-397.

Geier, E., Pfeifer, G., Wilm, M., Lucchiari-Hartz, M., Baumeister, W., Eichmann, K. and Niedermann, G. (1999) A giant protease with potential to substitute for some functions of the proteasome, *Science*, **283**, 978-981.

Gil-Torregrosa, B.C., Raul Castano, A. and Del Val, M. (1998) Major histocompatibility complex class I viral antigen processing in the secretory pathway defined by the trans-Golgi network protease furin, *J Exp Med*, **188**, 1105-1116.

Glickman, M.H. and Ciechanover, A. (2002) The ubiquitin-proteasome proteolytic pathway: destruction for the sake of construction, *Physiol Rev*, **82**, 373-428.

Goldberg, A.L., Cascio, P., Saric, T. and Rock, K.L. (2002) The importance of the proteasome and subsequent proteolytic steps in the generation of antigenic peptides, *Mol Immunol*, **39**, 147-164.

Gribskov, M., McLachlan, A.D. and Eisenberg, D. (1987) Profile analysis: detection of distantly related proteins, *Proc Natl Acad Sci U S A*, **84**, 4355-4358.

Griffin, T.A., Nandi, D., Cruz, M., Fehling, H.J., Kaer, L.V., Monaco, J.J. and Colbert, R.A. (1998) Immunoproteasome assembly: cooperative incorporation of interferon gamma (IFN-gamma)-inducible subunits, *J Exp Med*, **187**, 97-104.

Groettrup, M., Standera, S., Stohwasser, R. and Kloetzel, P.M. (1997) The subunits MECL-1 and LMP2 are mutually required for incorporation into the 20S proteasome, *Proc Natl Acad Sci U S A*, **94**, 8970-8975.

Groll, M., Ditzel, L., Lowe, J., Stock, D., Bochtler, M., Bartunik, H.D. and Huber, R. (1997) Structure of 20S proteasome from yeast at 2.4 A resolution, *Nature*, **386**, 463-471.

Heath, W.R., Belz, G.T., Behrens, G.M., Smith, C.M., Forehan, S.P., Parish, I.A., Davey, G.M., Wilson, N.S., Carbone, F.R. and Villadangos, J.A. (2004) Cross-presentation, dendritic cell subsets, and the generation of immunity to cellular antigens, *Immunol Rev*, **199**, 9-26.

Hedrick, S.M. (2008) Thymus lineage commitment: a single switch, *Immunity*, **28**, 297-299.

Hillen, N. and Stevanovic, S. (2006) Contribution of mass spectrometry-based proteomics to immunology, *Expert Rev Proteomics*, **3**, 653-664.

Holzhutter, H.G., Frommel, C. and Kloetzel, P.M. (1999) A theoretical approach towards the identification of cleavage-determining amino acid motifs of the 20 S proteasome, *J Mol Biol*, **286**, 1251-1265.

Holzhutter, H.G. and Kloetzel, P.M. (2000) A kinetic model of vertebrate 20S proteasome accounting for the generation of major proteolytic fragments from oligomeric peptide substrates, *Biophys J*, **79**, 1196-1205.

Janeway, C.A., Jr. and Medzhitov, R. (2002) Innate immune recognition, *Annu Rev Immunol*, **20**, 197-216.

Jimenez-Montano, M.A., Ebeling, W., Pohl, T. and Rapp, P.E. (2002) Entropy and complexity of finite sequences as fluctuating quantities, *Biosystems*, **64**, 23-32.

Kabat, E.A. (1970) Antigenic determinants and antibody complementarity, *Folia Allergol (Roma)*, **17**, 425.

Kanaseki, T., Blanchard, N., Hammer, G.E., Gonzalez, F. and Shastri, N. (2006) ERAAP synergizes with MHC class I molecules to make the final cut in the antigenic peptide precursors in the endoplasmic reticulum, *Immunity*, **25**, 795-806.

Kesmir, C., Nussbaum, A.K., Schild, H., Detours, V. and Brunak, S. (2002) Prediction of proteasome cleavage motifs by neural networks, *Protein Eng*, **15**, 287-296.

Khan, S., de Giuli, R., Schmidtke, G., Bruns, M., Buchmeier, M., van den Broek, M. and Groettrup, M. (2001) Cutting edge: neosynthesis is required for the presentation of a T cell epitope from a long-lived viral protein, *J Immunol*, **167**, 4801-4804.

Kisselev, A.F., Akopian, T.N., Woo, K.M. and Goldberg, A.L. (1999) The sizes of peptides generated from protein by mammalian 26 and 20 S proteasomes. Implications for understanding the degradative mechanism and antigen presentation, *J Biol Chem*, **274**, 3363-3371.

Kloetzel, P.M. (2001) Antigen processing by the proteasome, *Nat Rev Mol Cell Biol*, **2**, 179-187.

Kloetzel, P.M. (2004) Generation of major histocompatibility complex class I antigens: functional interplay between proteasomes and TPPII, *Nat Immunol*, **5**, 661-669.

Kuttler, C., Nussbaum, A.K., Dick, T.P., Rammensee, H.G., Schild, H. and Hadeler, K.P. (2000) An algorithm for the prediction of proteasomal cleavages, *J Mol Biol*, **298**, 417-429.

Lankat-Buttgereit, B. and Tampe, R. (2002) The transporter associated with antigen processing: function and implications in human diseases, *Physiol Rev*, **82**, 187-204.

Lazoura, E. and Apostolopoulos, V. (2005) Insights into peptide-based vaccine design for cancer immunotherapy, *Curr Med Chem*, **12**, 1481-1494.

Lehner, P.J., Surman, M.J. and Cresswell, P. (1998) Soluble tapasin restores MHC class I expression and function in the tapasin-negative cell line .220, *Immunity*, **8**, 221-231.

Li, S., Paulsson, K.M., Chen, S., Sjogren, H.O. and Wang, P. (2000) Tapasin is required for efficient peptide binding to transporter associated with antigen processing, *J Biol Chem*, **275**, 1581-1586.

Lopez, D. and Del Val, M. (1997) Selective involvement of proteasomes and cysteine proteases in MHC class I antigen presentation, *J Immunol*, **159**, 5769-5772.

Luckey, C.J., Marto, J.A., Partridge, M., Hall, E., White, F.M., Lippolis, J.D., Shabanowitz, J., Hunt, D.F. and Engelhard, V.H. (2001) Differences in the expression of human class I MHC alleles and their associated peptides in the presence of proteasome inhibitors, *J Immunol*, **167**, 1212-1221.

Madden, D.R. (1995) The three-dimensional structure of peptide-MHC complexes, *Annu Rev Immunol*, **13**, 587-622.

Madden, D.R., Garboczi, D.N. and Wiley, D.C. (1993) The antigenic identity of peptide-MHC complexes: a comparison of the conformations of five viral peptides presented by HLA-A2, *Cell*, **75**, 693-708.

Matsumura, M., Fremont, D.H., Peterson, P.A. and Wilson, I.A. (1992) Emerging principles for the recognition of peptide antigens by MHC class I molecules, *Science*, **257**, 927-934.

Medzhitov, R. and Janeway, C.A., Jr. (1997) Innate immunity: impact on the adaptive immune response, *Curr Opin Immunol*, **9**, 4-9.

Metz, C.E. (1978) Basic principles of ROC analysis, *Semin Nucl Med*, **8**, 283-298.

Mo, A.X., van Lelyveld, S.F., Craiu, A. and Rock, K.L. (2000) Sequences that flank subdominant and cryptic epitopes influence the proteolytic generation of MHC class I-presented peptides, *J Immunol*, **164**, 4003-4010.

Mo, X.Y., Cascio, P., Lemerise, K., Goldberg, A.L. and Rock, K. (1999) Distinct proteolytic processes generate the C and N termini of MHC class I-binding peptides, *J Immunol*, **163**, 5851-5859.

Momburg, F., Roelse, J., Howard, J.C., Butcher, G.W., Hammerling, G.J. and Neefjes, J.J. (1994) Selectivity of MHC-encoded peptide transporters from human, mouse and rat, *Nature*, **367**, 648-651.

Momburg, F. and Tan, P. (2002) Tapasin-the keystone of the loading complex optimizing peptide binding by MHC class I molecules in the endoplasmic reticulum, *Mol Immunol*, **39**, 217-233.

Morel, S., Levy, F., Burlet-Schiltz, O., Brasseur, F., Probst-Kepper, M., Peitrequin, A.L., Monsarrat, B., Van Velthoven, R., Cerottini, J.C., Boon, T., Gairin, J.E. and Van den Eynde, B.J. (2000) Processing of some antigens by the standard proteasome but not by the immunoproteasome results in poor presentation by dendritic cells, *Immunity*, **12**, 107-117.

Nazif, T. and Bogyo, M. (2001) Global analysis of proteasomal substrate specificity using positional-scanning libraries of covalent inhibitors, *Proc Natl Acad Sci U S A*, **98**, 2967-2972.

Neefjes, J.J., Momburg, F. and Hammerling, G.J. (1993) Selective and ATP-dependent translocation of peptides by the MHC-encoded transporter, *Science*, **261**, 769-771.

Nielsen, M., Lundegaard, C., Lund, O. and Kesmir, C. (2005) The role of the proteasome in generating cytotoxic T-cell epitopes: insights obtained from improved predictions of proteasomal cleavage, *Immunogenetics*, **57**, 33-41.

Nijenhuis, M. and Hammerling, G.J. (1996) Multiple regions of the transporter associated with antigen processing (TAP) contribute to its peptide binding site, *J Immunol*, **157**, 5467-5477.

Nussbaum, A.K., Dick, T.P., Keilholz, W., Schirle, M., Stevanovic, S., Dietz, K., Heinemeyer, W., Groll, M., Wolf, D.H., Huber, R., Rammensee, H.G. and Schild, H. (1998) Cleavage motifs of the yeast 20S proteasome beta subunits deduced from digests of enolase 1, *Proc Natl Acad Sci U S A*, **95**, 12504-12509.

Nussbaum, A.K., Kuttler, C., Hadeler, K.P., Rammensee, H.G. and Schild, H. (2001) PAProC: a prediction algorithm for proteasomal cleavages available on the WWW, *Immunogenetics*, **53**, 87-94.

Orr, H.T., Lancet, D., Robb, R.J., Lopez de Castro, J.A. and Strominger, J.L. (1979) The heavy chain of human histocompatibility antigen HLA-B7 contains an immunoglobulin-like region, *Nature*, **282**, 266-270.

Orr, H.T., Lopez de Castro, J.A., Lancet, D. and Strominger, J.L. (1979) Complete amino acid sequence of a papain-solubilized human histocompatibility antigen, HLA-B7. 2. Sequence determination and search for homologies, *Biochemistry*, **18**, 5711-5720.

Ortmann, B., Copeman, J., Lehner, P.J., Sadasivan, B., Herberg, J.A., Grandea, A.G., Riddell, S.R., Tampe, R., Spies, T., Trowsdale, J. and Cresswell, P. (1997) A critical role for tapasin in the assembly and function of multimeric MHC class I-TAP complexes, *Science*, **277**, 1306-1309.

Palm, N.W. and Medzhitov, R. (2009) Pattern recognition receptors and control of adaptive immunity, *Immunol Rev*, **227**, 221-233.

Pamer, E. and Cresswell, P. (1998) Mechanisms of MHC class I--restricted antigen processing, *Annu Rev Immunol*, **16**, 323-358.

Pancer, Z. and Cooper, M.D. (2006) The evolution of adaptive immunity, *Annu Rev Immunol*, **24**, 497-518.

Peaper, D.R. and Cresswell, P. (2008) Regulation of MHC class I assembly and peptide binding, *Annu Rev Cell Dev Biol*, **24**, 343-368.

Peters, B., Bulik, S., Tampe, R., Van Endert, P.M. and Holzhutter, H.G. (2003) Identifying MHC class I epitopes by predicting the TAP transport efficiency of epitope precursors, *J Immunol*, **171**, 1741-1749.

Peters, J.M. (1994) Proteasomes: protein degradation machines of the cell, *Trends Biochem Sci*, **19**, 377-382.

Princiotta, M.F., Finzi, D., Qian, S.B., Gibbs, J., Schuchmann, S., Buttgereit, F., Bennink, J.R. and Yewdell, J.W. (2003) Quantitating protein synthesis, degradation, and endogenous antigen processing, *Immunity*, **18**, 343-354.

Princiotta, M.F., Schubert, U., Chen, W., Bennink, J.R., Myung, J., Crews, C.M. and Yewdell, J.W. (2001) Cells adapted to the proteasome inhibitor 4-hydroxy- 5-iodo-3-nitrophenylacetyl-Leu-Leu-leucinal-vinyl sulfone require enzymatically active proteasomes for continued survival, *Proc Natl Acad Sci U S A*, **98**, 513-518.

Purcell, A.W., McCluskey, J. and Rossjohn, J. (2007) More than one reason to rethink the use of peptides in vaccine design, *Nat Rev Drug Discov*, **6**, 404-414.

Qian, S.B., Reits, E., Neefjes, J., Deslich, J.M., Bennink, J.R. and Yewdell, J.W. (2006) Tight linkage between translation and MHC class I peptide ligand generation implies specialized antigen processing for defective ribosomal products, *J Immunol*, **177**, 227-233.

Reche, P.A., Glutting, J.P. and Reinherz, E.L. (2002) Prediction of MHC class I binding peptides using profile motifs, *Hum Immunol*, **63**, 701-709.

Reche, P.A., Glutting, J.P., Zhang, H. and Reinherz, E.L. (2004) Enhancement to the RANKPEP resource for the prediction of peptide binding to MHC molecules using profiles, *Immunogenetics*, **56**, 405-419.

Reche, P.A., Keskin, D.B., Hussey, R.E., Ancuta, P., Gabuzda, D. and Reinherz, E.L. (2006) Elicitation from virus-naive individuals of cytotoxic T lymphocytes directed against conserved HIV-1 epitopes, *Med Immunol*, **5**, 1.

Reche, P.A. and Reinherz, E.L. (2003) Sequence variability analysis of human class I and class II MHC molecules: functional and structural correlates of amino acid polymorphisms, *J Mol Biol*, **331**, 623-641.

Reche, P.A. and Reinherz, E.L. (2007) Prediction of peptide-MHC binding using profiles, *Methods Mol Biol*, **409**, 185-200.

Reche, P.A., Zhang, H., Glutting, J.P. and Reinherz, E.L. (2005) EPIMHC: a curated database of MHC-binding peptides for customized computational vaccinology, *Bioinformatics*, **21**, 2140-2141.

Reits, E., Neijssen, J., Herberts, C., Benckhuijsen, W., Janssen, L., Drijfhout, J.W. and Neefjes, J. (2004) A major role for TPPII in trimming proteasomal degradation products for MHC class I antigen presentation, *Immunity*, **20**, 495-506.

Reits, E.A., Vos, J.C., Gromme, M. and Neefjes, J. (2000) The major substrates for TAP in vivo are derived from newly synthesized proteins, *Nature*, **404**, 774-778.

Rock, K.L. and Goldberg, A.L. (1999) Degradation of cell proteins and the generation of MHC class I-presented peptides, *Annu Rev Immunol*, **17**, 739-779.

Rock, K.L., Gramm, C., Rothstein, L., Clark, K., Stein, R., Dick, L., Hwang, D. and Goldberg, A.L. (1994) Inhibitors of the proteasome block the degradation of most cell proteins and the generation of peptides presented on MHC class I molecules, *Cell*, **78**, 761-771.

Rosenfeld (2000) Two decades of statistical language modeling: Where do we

go from here?, *Proceedings of the IEEE*, **88**, 1-11.

Rudolph, M.G. and Wilson, I.A. (2002) The specificity of TCR/pMHC interaction, *Curr Opin Immunol*, **14**, 52-65.

Sadasivan, B., Lehner, P.J., Ortmann, B., Spies, T. and Cresswell, P. (1996) Roles for calreticulin and a novel glycoprotein, tapasin, in the interaction of MHC class I molecules with TAP, *Immunity*, **5**, 103-114.

Sagar, M. (2010) HIV-1 transmission biology: selection and characteristics of infecting viruses, *J Infect Dis*, **202 Suppl 2**, S289-296.

Sandberg, J.K., Grufman, P., Wolpert, E.Z., Franksson, L., Chambers, B.J. and Karre, K. (1998) Superdominance among immunodominant H-2Kb-restricted epitopes and reversal by dendritic cell-mediated antigen delivery, *J Immunol*, **160**, 3163-3169.

Saric, T., Chang, S.C., Hattori, A., York, I.A., Markant, S., Rock, K.L., Tsujimoto, M. and Goldberg, A.L. (2002) An IFN-gamma-induced aminopeptidase in the ER, ERAP1, trims precursors to MHC class I-presented peptides, *Nat Immunol*, **3**, 1169-1176.

Saveanu, L., Carroll, O., Hassainya, Y. and van Endert, P. (2005) Complexity, contradictions, and conundrums: studying post-proteasomal proteolysis in HLA class I antigen presentation, *Immunol Rev*, **207**, 42-59.

Saxova, P., Buus, S., Brunak, S. and Kesmir, C. (2003) Predicting proteasomal cleavage sites: a comparison of available methods, *Int Immunol*, **15**, 781-787.

Schlesinger, D.H., Goldstein, G. and Niall, H.D. (1975) The complete amino acid sequence of ubiquitin, an adenylate cyclase stimulating polypeptide probably universal in living cells, *Biochemistry*, **14**, 2214-2218.

Schrodt, S., Koch, J. and Tampe, R. (2006) Membrane topology of the transporter associated with antigen processing (TAP1) within an assembled functional peptide-loading complex, *J Biol Chem*, **281**, 6455-6462.

Schubert, U., Anton, L.C., Gibbs, J., Norbury, C.C., Yewdell, J.W. and Bennink, J.R. (2000) Rapid degradation of a large fraction of newly synthesized proteins by proteasomes, *Nature*, **404**, 770-774.

Serwold, T., Gonzalez, F., Kim, J., Jacob, R. and Shastri, N. (2002) ERAAP customizes peptides for MHC class I molecules in the endoplasmic reticulum, *Nature*, **419**, 480-483.

Shannon, C.E. (1948) The mathematical theory of communication, *Bell Sys. Tech. J.*, **27**, 379-423, 623-656.

Shen, L., Sigal, L.J., Boes, M. and Rock, K.L. (2004) Important role of cathepsin S in generating peptides for TAP-independent MHC class I crosspresentation in vivo, *Immunity*, **21**, 155-165.

Shepherd, J.C., Schumacher, T.N., Ashton-Rickardt, P.G., Imaeda, S., Ploegh, H.L., Janeway, C.A., Jr. and Tonegawa, S. (1993) TAP1-dependent peptide translocation in vitro is ATP dependent and peptide selective, *Cell*, **74**, 577-584.

Sijts, A.J. and Pamer, E.G. (1997) Enhanced intracellular dissociation of major histocompatibility complex class I-associated peptides: a mechanism for optimizing the spectrum of cell surface-presented cytotoxic T lymphocyte epitopes, *J Exp Med*, **185**, 1403-1411.

Silva, M.T. (2010) When two is better than one: macrophages and neutrophils work in concert in innate immunity as complementary and cooperative partners of a myeloid phagocyte system, *J Leukoc Biol*, **87**, 93-106.

Simpson, E.H. (1949) Measurement of diversity, *Nature*, **163**, 688.

Solheim, J.C. (1999) Class I MHC molecules: assembly and antigen presentation, *Immunol Rev*, **172**, 11-19.

Stern, L.J. and Wiley, D.C. (1994) Antigenic peptide binding by class I and class II histocompatibility proteins, *Structure*, **2**, 245-251.

Stienekemeier, M., Falk, K., Rotzschke, O., Weishaupt, A., Schneider, C., Toyka, K.V., Gold, R. and Strominger, J.L. (2001) Vaccination, prevention, and treatment of experimental autoimmune neuritis (EAN) by an oligomerized T cell epitope, *Proc Natl Acad Sci U S A*, **98**, 13872-13877.

Stolcke, A. (2002) SRILM -- An Extensible Language Modeling Toolkit. In J. J. Ohala, T.M.N., B. L. Derwing, M. M. Hodge, and G. E. Wiebe (ed), *Proceedings of the International Conference of Spoken Language Processing*. Center for Spoken Language Research, Boulder, CO, 901-904.

Takahama, Y. (2006) Journey through the thymus: stromal guides for T-cell development and selection, *Nat Rev Immunol*, **6**, 127-135.

Tan, P., Kropshofer, H., Mandelboim, O., Bulbuc, N., Hammerling, G.J. and Momburg, F. (2002) Recruitment of MHC class I molecules by tapasin into the transporter associated with antigen processing-associated complex is essential for optimal peptide loading, *J Immunol*, **168**, 1950-1960.

Tchernev, G. and Orfanos, C.E. (2006) Antigen mimicry, epitope spreading and the pathogenesis of pemphigus, *Tissue Antigens*, **68**, 280-286.

Tenzer, S., Peters, B., Bulik, S., Schoor, O., Lemmel, C., Schatz, M.M., Kloetzel, P.M., Rammensee, H.G., Schild, H. and Holzhutter, H.G. (2005) Modeling the MHC class I pathway by combining predictions of proteasomal cleavage, TAP transport and MHC class I binding, *Cell Mol Life Sci*, **62**, 1025-1037.

Terasaki, P.I. (2007) A brief history of HLA, *Immunol Res*, **38**, 139-148.

Toes, R.E., Nussbaum, A.K., Degermann, S., Schirle, M., Emmerich, N.P., Kraft, M., Laplace, C., Zwinderman, A., Dick, T.P., Muller, J., Schonfisch, B., Schmid, C., Fehling, H.J., Stevanovic, S., Rammensee, H.G. and Schild, H. (2001) Discrete cleavage motifs of constitutive and immunoproteasomes revealed by quantitative analysis of cleavage products, *J Exp Med*, **194**, 1-12.

Toseland, C.P., Clayton, D.J., McSparron, H., Hemsley, S.L., Blythe, M.J., Paine, K., Doytchinova, I.A., Guan, P., Hattotuwagama, C.K. and Flower, D.R. (2005) AntiJen: a quantitative immunology database integrating functional, thermodynamic, kinetic, biophysical, and cellular data, *Immunome Res*, **1**, 4.

Turvey, S.E. and Broide, D.H. (2010) Innate immunity, *J Allergy Clin Immunol*, **125**, S24-32.

Uebel, S., Kraas, W., Kienle, S., Wiesmuller, K.H., Jung, G. and Tampe, R. (1997) Recognition principle of the TAP transporter disclosed by combinatorial peptide libraries, *Proc Natl Acad Sci U S A*, **94**, 8976-8981.

Unno, M., Mizushima, T., Morimoto, Y., Tomisugi, Y., Tanaka, K., Yasuoka, N. and Tsukihara, T. (2002) The structure of the mammalian 20S proteasome at 2.75 A resolution, *Structure*, **10**, 609-618.

van Endert, P.M., Riganelli, D., Greco, G., Fleischhauer, K., Sidney, J., Sette, A. and Bach, J.F. (1995) The peptide-binding motif for the human transporter associated with antigen processing, *J Exp Med*, **182**, 1883-1895.

van Endert, P.M., Tampe, R., Meyer, T.H., Tisch, R., Bach, J.F. and McDevitt, H.O. (1994) A sequential model for peptide binding and transport by the transporters associated with antigen processing, *Immunity*, **1**, 491-500.

Vapnik, V. (1995) The Nature of Statistical Learning Theory, *Springer, New York*.

Vapnik, V. (1998) Statistical Learning Theory, *Wiley-Interscience; New York*.

Vita, R., Zarebski, L., Greenbaum, J.A., Emami, H., Hoof, I., Salimi, N., Damle, R., Sette, A. and Peters, B. (2010) The immune epitope database 2.0, *Nucleic Acids Res*, **38**, D854-862.

Vivona, S., Gardy, J.L., Ramachandran, S., Brinkman, F.S., Raghava, G.P., Flower, D.R. and Filippini, F. (2008) Computer-aided biotechnology: from immuno-informatics to reverse vaccinology, *Trends Biotechnol*, **26**, 190-200.

Voges, D., Zwickl, P. and Baumeister, W. (1999) The 26S proteasome: a molecular machine designed for controlled proteolysis, *Annu Rev Biochem*, **68**, 1015-1068.

Vos, J.C., Spee, P., Momburg, F. and Neefjes, J. (1999) Membrane topology and dimerization of the two subunits of the transporter associated with antigen processing reveal a three-domain structure, *J Immunol*, **163**, 6679-6685.

Wang, M., Lamberth, K., Harndahl, M., Roder, G., Stryhn, A., Larsen, M.V., Nielsen, M., Lundegaard, C., Tang, S.T., Dziegiel, M.H., Rosenkvist, J., Pedersen, A.E., Buus, S., Claesson, M.H. and Lund, O. (2007) CTL epitopes for influenza A including the H5N1 bird flu; genome-, pathogen-, and HLA-wide screening, *Vaccine*, **25**, 2823-2831.

Wu, C. and Shivakumar, S. (1994) Back-propagation and counter-propagation neural networks for phylogenetic classification of ribosomal RNA sequences, *Nucleic Acids Res*, **22**, 4291-4299.

Wu, C.H., Zhao, S., Chen, H.L., Lo, C.J. and McLarty, J. (1996) Motif identification neural design for rapid and sensitive protein family search, *Comput Appl Biosci*, **12**, 109-118.

Wu, T.T. and Kabat, E.A. (1970) An analysis of the sequences of the variable regions of Bence Jones proteins and myeloma light chains and their implications for antibody complementarity, *J Exp Med*, **132**, 211-250.

Yewdell, J.W., Anton, L.C. and Bennink, J.R. (1996) Defective ribosomal products (DRiPs): a major source of antigenic peptides for MHC class I molecules?, *J Immunol*, **157**, 1823-1826.

Yewdell, J.W. and Princiotta, M.F. (2004) Proteasomes get by with lots of help from their friends, *Immunity*, **20**, 362-363.

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

Zhang, G.L., Petrovsky, N., Kwoh, C.K., August, J.T. and Brusic, V. (2006) PRED(TAP): a system for prediction of peptide binding to the human transporter associated with antigen processing, *Immunome Res*, **2**, 3.

Zhong, W., Reche, P.A., Lai, C.C., Reinhold, B. and Reinherz, E.L. (2003) Genome-wide characterization of a viral cytotoxic T lymphocyte epitope repertoire, *J Biol Chem*, **278**, 45135-45144.

11. ANEXO I

Otras publicaciones

TEPIDAS: A DAS Server for Integrating T-Cell Epitope Annotations

M. García-Boronat, C.M. Díez-Rivero, and Pedro Reche

Abbreviations

CMV	Cumulative phenotypic frequency
DAS	Distributed annotation system
HLA I	Human leukocyte antigen class I
PSSM	Position-specific scoring matrix

Introduction

Recent years have witnessed the birth of Immunoinformatics, an emerging subdiscipline of Bioinformatics. With the burgeoning explosion of immunological data, computational analysis has become an essential element of immunology research, facilitating the understanding of the immune function by modeling the interactions among immunological components (Petrovsky and Brusic 2006). Another major role in Immunoinformatics is the efficient management, storage, and annotation of such data. Following those principles, a large number of immunoinformatics resources, including immune-related databases and sophisticated analysis software, are available through the World Wide Web (Davies and Flower 2007). Collectively, these resources contribute to the advances made in immunological research. Yet, there is still a major step to be taken toward the integration of all these resources, as ideally, multiple research groups should be able to exchange and compare their data, in a quick and efficient fashion.

In this chapter, we show an example of how an epitope database can be integrated to other database resources using the Distributed Annotation System (DAS) (Dowell et al. 2001). For that we describe the TEPIDAS server, a DAS Annotation Server of HLA I-restricted CD8 T-cell epitopes specific of human pathogenic organisms.

P. Reche (✉)
Facultad de Medicina, Departamento de Immunología (Microbiología I), Universidad Complutense de Madrid, Pabellón 5º, planta 4ª, 28040, Madrid, Spain
e-mail: parecheg@med.ucm.es

The Distributed Annotation System

Introduction

The distributed annotation system defines a communication protocol used to exchange biological annotations from a number of heterogeneous distributed databases. The key idea behind the DAS concept is that annotations should not be provided by single centralized databases but instead be spread over multiple sites. This distribution of data encourages a divide-and-conquer approach to annotation, where experts provide and maintain their own annotations.

The Protocol

Currently, there are two versions of the DAS protocol. The original DAS protocol (DAS/1) was designed to serve annotation of genomic sequences. That protocol was later extended (DAS/2) to be applicable to alignments and 3D structure information (Prlic et al. 2005). It is very likely that further extensions of the protocol will appear in a near future, such as the new extension for electron microscopy data recently published by Macias et al. (2007).

The DAS protocol is a simple http-based client–server system. DAS clients make requests in the form of a URL to the servers and receive simple XML responses (Crook and Howell 2007). The architecture of the system will be next described in the following subsection.

The Architecture of the System

The basic system is composed of a reference server, one or more annotation servers, and an annotation viewer. The reference server is responsible for serving genome maps, sequences and information related to the sequencing process. Annotation servers are responsible for returning the annotations on a defined region (given a start and stop position coordinates) of the genome. The annotation viewer can either be a simple web browser, which will visualize the raw XML data provided by the server, or a graphical client such as the Center for Biological Sequence Analysis (CBS) DAS viewer (Olason 2005) accessible at http://www.cbs.dtu.dk/cgi-gin/das. This viewer translates the XML annotations to aligned graphical tracks making it easier to visualize the features along the length of the protein. Additional information about the annotations is shown in a pop-up window when the mouse points to an annotation track.

Although the servers are conceptually divided between reference and annotation servers, there is in fact no key difference between them. A single server can provide both reference sequence information and annotation information. The only functional

difference is that the reference sequence server is required to serve the coordinate map and the raw DNA, while annotation servers have no such requirement. Our TEPIDAS server falls into the category of annotation servers.

The DAS Registry

The DAS Registry is a public server (http://www.dasregistry.org) dedicated to the registration, validation, and listing of worldwide DAS servers. One can browse the list of available DAS sources at the Registry, as well as register his own DAS server for public use. The Registry automatically validates the DAS server when it is being registered, ensuring that it returns well-formed XML responses. In addition, it periodically tests DAS sources and notifies their administrators if they are unavailable.

When you register your DAS server, you have to specify the Coordinate System of your source in order to describe the kind of data that are being made available. This information is important for the DAS clients to deal with data correctly, as they often can accept data served in multiple coordinate systems. The Coordinate System is described by the following four fields: "Authority," "(assembly) Version," "Type," and "Organism." The assembly version is important for genome assemblies, but not really applicable for other datasets like UniProt sequences; therefore, this field is optional. The "authority" is the name of an authority/institution that defines the accession codes of a coordinate system or that provides a gene build. In the latter case this field also contains the "version" number of the assembly. The "type" or category of the coordinate system refers to the physical dimension of the annotated data. Some examples include: Chromosome, Clone, Protein Sequence, and Protein Structure. The last field is the "organism" the data refer to. Not every DAS source is organism specific, and therefore this field is optional.

During the registration process, you also have to specify the capabilities of your DAS source, that is the types of queries that your server will be able to serve a response to. Some basic queries that can be used by a client to interrogate a DAS server are: "dna," "features," and "types." The "dna" query can be used to fetch a segment of DNA from a reference server. "features" is the query used to retrieve the actual annotations, and the "types" query returns a summary of the available annotation types. These three are just some examples of DAS queries. Readers can access the full list and specification of query types at the DAS web page (http://www.biodas.org).

The TEPIDAS server has been registered at the DAS registry since February 2008 and has the unique id DS_545. The coordinate system defined for TEPIDAS is Uniprot (Wu et al. 2006), as the "authority," and Protein Sequence, as the "type." As for TEPIDAS capabilities, our server implements the "types" and "features" queries. Note that our server is just an annotation server, and therefore it does not provide the "dna" query, served only by reference servers. A comprehensive description of the TEPIDAS server follows next.

TEPIDAS

TEPIDAS is a DAS annotation server that provides annotations for CD8 T-cell epitopes consisting of the distinct HLA I molecules to which that epitope binds, following the UniProt coordinates system. TEPIDAS is implemented using ProServer (Finn et al. 2007), a lightweight Perl-based DAS server that does not depend on a separate HTTP server. The annotations are precalculated and the results stored in a relational database, allowing for fast retrieval and update of data. When a client makes a query to the TEPIDAS server, ProServer simply retrieves the relevant information from the relational database and composes the XML response.

Annotations Served by TEPIDAS

TEPIDAS annotates CD8 T-cell epitopes according to the HLA I molecules that restrict them. Epitopes were obtained from the EPIMHC (Reche et al. 2005) and IMMUNEEPITOPE (Peters et al. 2005) databases, and were selected to be experimentally defined in humans infected with the pathogen or immunized with the relevant source antigen. HLA I-restriction annotations can be classified as experimental, when determined experimentally, or predicted. Predictions of the epitopes binding HLA I molecules were obtained using a set of 72 position-specific scoring matrices (PSSMs), also known as weight matrices of profiles, which are obtained from aligned peptides known to bind to the relevant HLA I molecules. This predictive method is described in full detail at (Reche et al. 2002, 2004). In addition to the experimental and predicted data, the cumulative phenotypic frequency (CMV) of the T-cell epitope HLA I restriction is also provided for five ethnic groups (Black, Caucasian, Hispanic, North American natives, and Asian). CMV was computed using the gene and haplotype frequencies of the relevant HLA I alleles (Reche et al. 2006). The potential population protection coverage of a T cell epitope-based vaccine is determined by the percentage of the population that could elicit a T cell response to the epitopes, which in turn is given by the CMV of HLA I molecules restricting these epitopes.

TEPIDAS Query Capabilities

As we mentioned before, TEPIDAS capabilities include the “types” and “features” queries. An explanation and an example for each query follow next.

The “types” query returns a list of all the distinct HLA I molecules that are used to annotate the epitopes. A total of 125 different HLA I restriction elements are included in TEPIDAS. To make this query to the server, you simply have to access the following URL through your web browser:

http://imed.med.ucm.es:9000/das/tepidas/types

and the XML response you will get is shown as follows.

```
<?xml version="1.0" standalone="no"?>
<!DOCTYPE DASTYPES SYSTEM "http://www.biodas.org/dtd/dastypes.dtd">
<DASTYPES>
  <GFF version="1.0" href="http://imed.med.ucm.es:9000/das/tepidas/types">
  <SEGMENT version="1.0">
    <TYPE id="HLA-A*02" method="Experimental" category="default"></TYPE>
    <TYPE id="HLA-A*0201" method="Experimental" category="default"></TYPE>
                                .
                                .
                                .
    <TYPE id="HLA-B*02706" method="Predicted" category="default"></TYPE>
    <TYPE id="HLA-B*02709" method="Predicted" category="default"></TYPE>
    <TYPE id="HLA-B*027" method="Predicted" category="default"></TYPE>
  </SEGMENT>
</GFF>
</DASTYPES>
```

Only a part of the XML response file is shown due to length constraints. Each type has an "id" that corresponds to the name of the HLA I molecule. There is also a "method" attribute that distinguishes between experimental and predicted annotations. In addition, a third attribute named "category" can be used to group different types, although we have not used that attribute, and therefore *default* is the "category" shown in the response.

The other type of query supported by TEPIDAS is the "features" query, which returns the actual annotations made on a reference UniProt sequence. An annotation feature includes the following information: the start and end position of the feature annotated, the method used to annotate it (experimental or predicted), the type of the annotation (the HLA I molecule to which it binds), a link to the UniProt page of the reference protein sequence, and a note field with additional complementary information. The information on the note varies depending on the feature's method. Common fields in the note of both methods are: the epitope source species name and taxonomy identifier, the name of the source protein, the cumulative phenotypic frequency (CMV) of the T-cell epitope HLA-I restriction for five ethnic groups (Black, Caucasian, Hispanic, North American natives, and Asian), and the immunogen type. Specific fields for the features with an experimental "method" are: T-cell epitope activity assays, the experimental HLA I restriction element, its binding level (low, moderate, high, or unknown), and the predicted HLA I restriction elements. As for the features with a predicted "method" the note also includes the predicted HLA I restriction element, as well as an extended prediction with additional HLA I restriction elements for that epitope.

The "features" query has several arguments that can be optionally used to restrict the results. For example, the following URL string:

http://imed.med.ucm.es:9000/das/tepidas/features?segment=P26664

will return all the features annotated on the UniProt protein sequence identified with the accession number P26664 (which will also be the features id).

If we want to restrict our query to the annotations on a particular region of the protein sequence, we could use:

http://imed.med.ucm.es:9000/das/tepidas/features?segment=Q9WMX2:885,893

which returns all the features for the protein sequence with accession number Q9WMX2 that lie within the region defined by the start and end positions 885 and 893. The XML response to this query is shown as follows.

```
<?xml version="1.0" standalone="yes"?>
<!DOCTYPE DASGFF SYSTEM "http://www.biodas.org/dtd/dasgff.dtd">
<DASGFF>
  <GFF version="1.01" href="http://imed.med.ucm.es:9000/das/tepidas/features">
    <SEGMENT id="Q9WMX2" version="1.0" start="885" stop="893">
    <FEATURE id="Q9WMX2" label="Q9WMX2">
      <TYPE id="HLA-A*2402" reference="no" subparts="no" superparts="no">
       HLA-A*2402</TYPE>
      <METHOD id="Experimental">Experimental</METHOD>
      <START>885</START>
      <END>893</END>
      <ORIENTATION>0</ORIENTATION>
      <NOTE>
       Epitope Source Species: Hepatitis C virus; TaxID: 11103
       Epitope Source Protein: Genome polyprotein
       T cell Epitope Activity positive on: 51 Chromium Release,
       Cytokine bioassay
       MHCI Restriction Element: HLA-A*2402 (Experimental)
       MHCI Binding level: unknown
       Predicted MHCI Restriction: HLA-A*24,  HLA-A*2402
       Cummulative Phenotypic Frequency of MHCI(%):
       5.5(Black),12.8(Caucasian),22.9(Hispanic),
       40.3(North American Natives),34.3(Asian)
       Immunogen: Infection</NOTE>
     <LINK href="http://www.ebi.uniprot.org/unipro t-
srv/uniProtView.do?proteinAc=Q9WMX2">http://www.ebi.uniprot.org/uniprot-
srv/uniProtView.do?proteinAc=Q9WMX2</LINK>
    </FEATURE>
    <FEATURE id="Q9WMX2" label="Q9WMX2">
      <TYPE id="HLA-A*24" reference="no" subparts="no" superparts="no">
       HLA-A*24</TYPE>
      <METHOD id="Predicted">Predicted</METHOD>
      <START>885</START>
      <END>893</END>
      <ORIENTATION>0</ORIENTATION>
      <NOTE>
       Epitope Source Species: Hepatitis C virus; TaxID: 11103
       Epitope Source Protein: Genome polyprotein
       T cell Epitope Activity: predicted
       MHCI Restriction Element: HLA-A*24 (Predicted)
       MHCI Binding level: unknown
       Extended predicted MHCI Restriction: HLA-A*24, HLA-A*2402
       Cummulative Phenotypic Frequency of MHCI(%):
       5.5(Black),12.8(Caucasian),22.9(Hispanic),
       40.3(North American Natives),34.3(Asian)
       Immunogen: Infection</NOTE>
     <LINK href="http://www.ebi.uniprot.org/uniprot-
srv/uniProtView.do?proteinAc=Q9WMX2">http://www.ebi.uniprot.org/uniprot-
srv/uniProtView.do?proteinAc=Q9WMX2</LINK>
    </FEATURE>
    </SEGMENT>
  </GFF>
   </DASGFF>
```

Example: Access TEPIDAS from the SPICE Graphical Client

In the previous section we have described how to access TEPIDAS annotations using formatted queries from a web browser, and we have also shown examples of the XML responses to the queries. We will now describe a different way of accessing TEPIDAS from a graphical client such as SPICE (Prlic et al. 2005). We hope that this example will illustrate the integration capability of DAS.

SPICE is a Java program that can be used to visualize annotations of protein sequences and protein structures. It is available at: http://www.efamily.org.uk/software/dasclients/spice. SPICE accepts either a PDB (Berman 2008) or a UniProt code, and integrates information from four different types of DAS servers: (1) a protein sequence server that provides the sequence (typically UniProt), (2) an alignment server that provides the alignment between the protein sequence and its structure, (3) a structure server that serves the 3D coordinates displayed, and (4) several feature servers that provide precalculated annotations, as for example TEPIDAS among others.

The SPICE viewer window consists of (1) a left structure panel, which provides a 3D visualization of the molecule using the open source Jmol library (http:/www.jmol.org), and (2) a right 2D feature panel that displays the annotations provided by the distributed servers. This is illustrated in Fig. 1 using the protein sequence with UniProt code P35961 as an example. As we can appreciate in Fig. 1, SPICE has automatically mapped that protein sequence to PBD "19GN" using its default alignment server. Figure 1 clearly shows how different annotations from several DAS servers can be integrated and collectively visualized through a graphical client such as SPICE. Users can choose which DAS annotations servers to use, as well as add new local DAS sources that are still under development or have not been registered with the DAS registry.

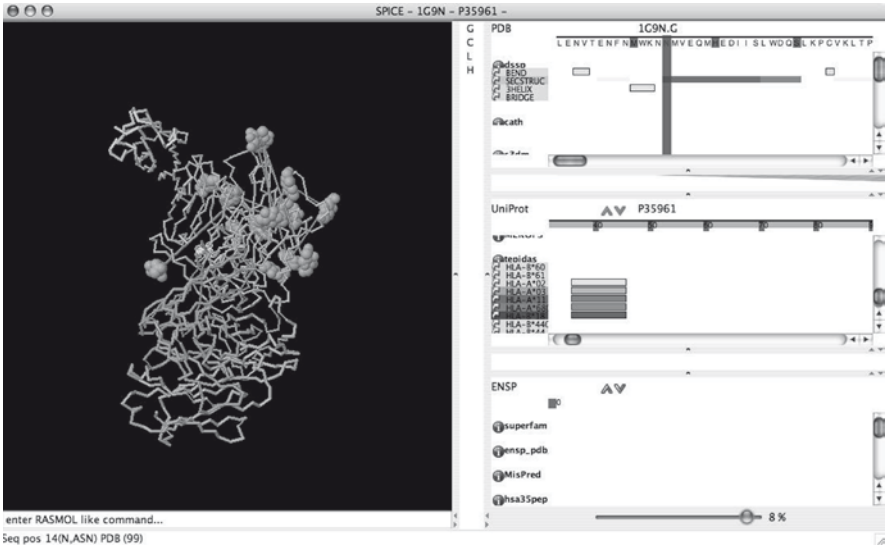


Fig. 1 SPICE viewer window. *Left panel* provides a 3D visualization of the molecule. *Right panel* displays the annotations provided by the distributed serves. This figure was generated using the UniProt code P35961 as the reference sequence. SPICE's alignment server automatically maps the protein sequence to a 3D structure (1G9N in this example). Feature annotations from TEPIDAS are displayed in the *right center panel* as *rectangular* tracks colored as the HLA I molecules on their *left* under the tepidas source descriptor

SPICE retrieves the protein sequence pertaining to the selected UniProt code and displays it as a ruler with relative position numbers, although there is a zoom feature that allows it to be expanded up to amino acid level as shown in Fig. 2 TEPIDAS annotation features are listed below the sequence in that figure. On the left of the panel, below the "tepidas" descriptor, appears the type of HLA I molecule of the corresponding feature shown as a colored rectangle on the right. When the user clicks on a feature, a pop-up window appears, containing all the information of the feature, including the explanatory note. In addition, the PDB coordinates of the selected feature will be highlighted at the left panel, enabling the location of the epitope at the 3D structure. Figure 3 shows an example of a pop-up window with feature information.

Fig. 2 SPICE zooming capability. Protein sequence visualized at amino acid level

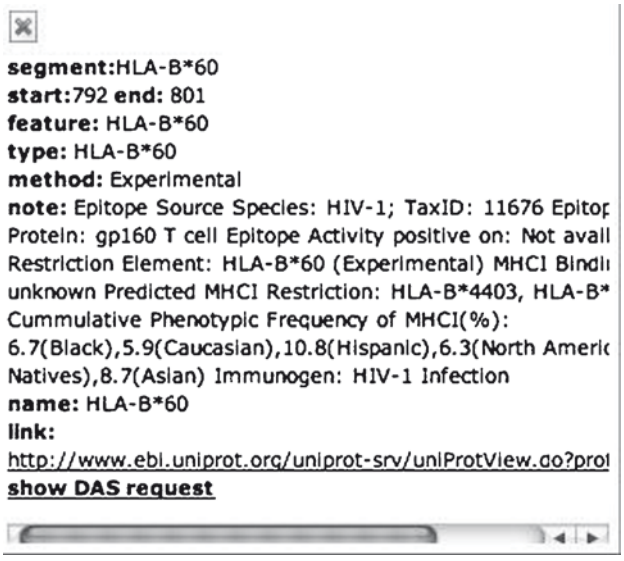


Fig. 3 Pop-up window containing all the information for feature HLA-B*60 annotated for protein sequence referenced by UniProt code P35961

Conclusion

DAS is an important, simple, and yet a powerful system for exchanging and viewing biological data that are already being used in real-world bioinformatics applications. The TEPIDAS annotation server described in this chapter is a clear example of how epitope data can be integrated and shared by the research community using the DAS architecture. The complexity of immune interactions and the data-intensive nature of immune research make Immunoinformatics a suitable area that could greatly benefit from the advantages of using such a powerful integration and annotation system, allowing to gain a more insightful understanding of the complexities of the immune system.

Acknowledgments We would like to thank Alfonso Valencia, Osvaldo Graña, and Jaime Fernandez Vera from the Spanish National Cancer Research Center (CNIO) for their helpful advice on DAS and ProServer. Work and authors were supported by grant SAF2006-07879 from the "Ministerio de Educación y Ciencia" of Spain, granted to PR.

References

Berman HM (2008) The Protein Data Bank: a historical perspective. Acta Crystallogr A 64(Pt 1): 88–95

Crook SM, Howell FW (2007) XML for data representation and model specification in neuroscience. Methods Mol Biol 401:53–66

Davies MN, Flower DR (2007) Harnessing bioinformatics to discover new vaccines. Drug Discov Today 12(9–10):389–395

Dowell RD, Jokerst RM et al (2001) The distributed annotation system. BMC Bioinformatics 2:7

Finn RD, Stalker JW, Jackson DK et al (2007) ProServer: a simple, extensible Perl DAS server. Bioinformatics 23(12):1568–1570

Macias JR, Jimenez-Lozano N, Carazo JM (2007) Integrating electron microscopy information into existing Distributed Annotation Systems. J Struct Biol 158(2):205–213

Olason PI (2005) Integrating protein annotation resources through the Distributed Annotation System. Nucleic Acids Res 33(Web Server issue):W468–W470

Peters B, Sidney J et al (2005) The immune epitope database and analysis resource: from vision to blueprint. PLoS Biol 3(3):e91

Petrovsky N, Brusic V (2006) Bioinformatics for study of autoimmunity. Autoimmunity 39(8):635–643

Prlic A, Down TA, Hubbard TJ (2005) Adding some SPICE to DAS. Bioinformatics 21(Suppl 2): ii40–ii41

Reche PA, Glutting JP, Reinherz EL (2002) Prediction of MHC class I binding peptides using profile motifs. Hum Immunol 63(9):701–709

Reche PA, Glutting JP et al (2004) Enhancement to the RANPEP resource for the prediction of peptide binding to MHC molecules using profiles. Immunogenetics 56(6):405–419

Reche PA, Zhang H et al (2005) EPIMHC: a curated database of MHC-binding peptides for customized computational vaccinology. Bioinformatics 21(9):2140–2141

Reche PA, Keskin DB, Hussey RE et al (2006) Elicitation from virus-naïve individuals of cytotoxic T lymphocytes directed against conserved HIV-1 epitopes. Med Immunol 5:1

Wu CH, Apweiler R, Bairoch A et al (2006) The Universal Protein Resource (UniProt): an expanding universe of protein information. Nucleic Acids Res 34(Database issue):D187–D191

Discovery of Conserved Epitopes Through Sequence Variability Analyses

Carmen M. Díez-Rivero and Pedro Reche

Introduction

Many pathogens exhibit high mutation rates, generating new genetic variants that are resistant to an existing immune response to earlier pathogen subtypes (Mendis et al. 1991; Phillips et al. 1991; Weber and Elliott 2002), difficulting the task of vaccine development. It is therefore important to focus on conserved regions during the process of vaccine design.

Several research groups have tried to develop vaccines based on quimeric consensus sequences (Thomsona et al. 2005). However, these vaccines have a major disadvantage as chimeric consensus proteins still bear nonconserved connecting regions, which might be more inmunogenic than conserved ones and thus truncate the development of a protective immune response. Nonprotective immunodominance can however be overcome using antigenic determinants (epitopes) as vaccines, as one can drive the immune response only towards the conserved epitopes of interest (Sette et al. 2002; Tsuji and Zavala 2001; Disis et al. 2001; Reche et al. 2006).

The estimation of sequence variability from MSAs of protein antigens also provides a means to identify conserved antigenic determinants. In this chapter, we will illustrate the use of PVS (García-Boronat et al. 2008), a Protein Variability Server that has been tuned to facilitate the discovery of conserved epitopes. Specifically, we will use PVS to obtain the conserved regions of the HIV-1 gp120 and gp41 proteins, identifying those that are solvent exposed, and therefore, likely the targets of cross-neutralizing antibodies (Abs). Likewise, we will use PVS to generate a variability-masked sequence of the HIV-1 gp120 protein, which will be targeted for T cell epitope predictions. Epitope-vaccine development requires confirming the immunogenicity of vaccine candidates, which consumes a vast amount of time and resources. Interestingly, sequence variability analyses in PVS dramatically reduce the number of potential epitope-vaccine candidates one would need to consider. PVS is freely available at the site http://imed.med.ucm.es/PVS.

P. Reche (✉)
Facultad de Medicina, Departamento de Immunología (Microbiología I), Universidad Complutense de Madrid, Pabellón 5º, planta 4ª, 28040, Madrid, Spain
e-mail: parecheg@med.umc.es

Materials and Methods

MSAs

For this study two proteins are used: The gp120 (residues 31-183 in gp160) and the gp41 (residues 528–674 in gp160), which are both membrane glycoproteins of HIV-1 (strain H2XB2). Both the gp120 and gp41 MSAs, were generated from 359 representative sequences of the HIV-1 clades A (73), B (85), C (85), D (51) and 01_AE (65) using the program MUSCLE (Edgar 2004). The gp41 and gp120 MSAs are available at http://imed.med.ucm.es/PVS/supplemental/gp120_pvs.html and http://imed.med.ucm.es/PVS/supplemental/gp41_pvs.html, respectively.

PVS Description and Usage

PVS (Protein Variability Server) is a web-based tool (Fig. 1) that following a protein sequence variability analysis performs several tasks that are relevant for structure-

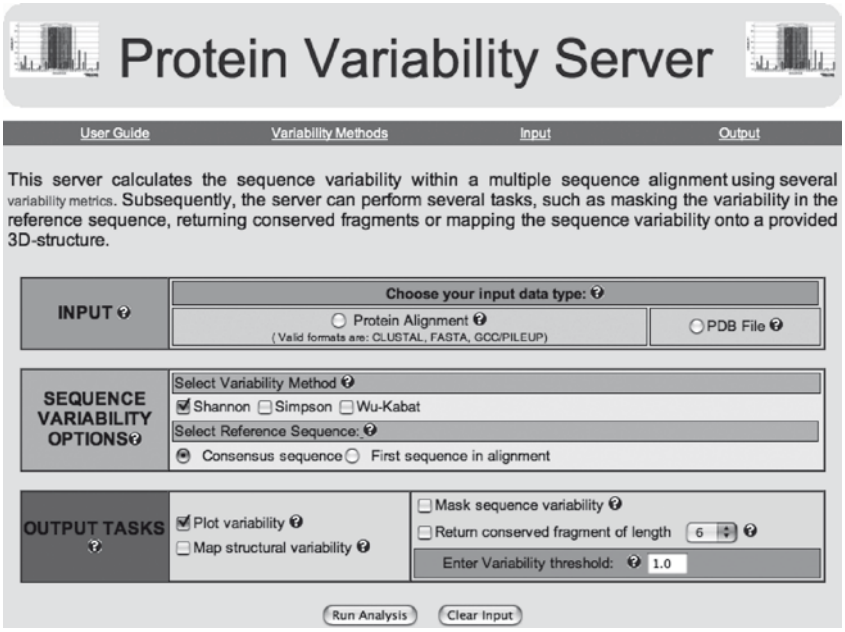


Fig. 1 *PVS web interface.* The web interface is divided into the INPUT, SEQUENCE VARIABILITY OPTIONS and OUTPUT TASKS sections which overall facilitate an intuitive use of the server. The web interface also provides links to help pages and specific information regarding the elements featured by the server accessible from the question mark icons

function studies and vaccine design. PVS main input is an MSA provided by the user, but it can also take a PDB file as main input, generating an MSA from it (for details see García-Boronat et al. 2008) The sequence variability in the MSA is computed *per site* using three different metrics: The Shannon Diversity index (Shannon Entropy) (Shannon 1948), the Simpson Diversity Index (Simpson 1949) and the Wu-Kabat Variability Coefficient (Wu and Kabat 1970). In this study, we have selected the Shannon Diversity Index (H) as the variability metric. H ranges from 0 (only one amino acid type is present at that position) to 4.322 (all 20 amino acids are equally represented in that position). Note, that for a site including gaps the maximum value of H will be 4.39. A site with a value of H under 1.0 is indicative of a site with very low variability (Reche and Reinherz 2003).

PVS optional tasks include that of plotting the variability in MSA – computed for each selected variability method – against a sequence consisting of a consensus sequence or the first sequence in the MSA. If the task "map structure variability" is selected and a PDB with relevant 3D-coordinates is submitted, PVS will map the sequence variability in the MSA onto the provided 3D-structure. Mapping the sequence variability onto the provided PDB is achieved by simply replacing the B-factor of the relevant residues with the variability values. Variability mapped 3D-structures can be visualized and manipulated interactively using JMOL (http://jmol.sourceforge.net/). The variability is shown in the 3D-structrure using a color scale that goes from blue for constant residues to red for highly variable residues. PVS also offers the possibility of returning the "conserved fragments." A variability threshold (*Vt*) and a minimum length of the conserved fragments need to be provided with this option. Under these selections, if a PDB is provided, PVS will also display a graph of the protein sequence with the conserved fragments shown in blue. By clicking on a fragment, one can locate the fragment on the 3D structure.

Finally, PVS can return the selected reference sequence with the variable positions masked. Specifically, those residues with variability greater than a user selected threshold will be shown with a "." symbol. The returned masked sequence is in FASTA format and can be directly submitted to RANKPEP (Reche and Reinherz 2007; Reche et al. 2004; Reche et al. 2002), a T cell epitope prediction tool that can anticipate conserved T-cell epitopes from a variability-masked sequence.

Results and Conclusion

Sequence variability is exploited by biological systems to generate functional heterogeneity (e.g., receptors involved in antigen recognition). Therefore, sequence variability analyses have traditionally been used to fill gaps in structural knowledge (Wu and Kabat 1970; Reche and Reinherz 2003). In addition, sequence variability analyses are also important for vaccine development as they also enable the identification of conserved antigenic determinants (Reche et al. 2006). For that purpose, we recently developed PVS, a web-based tool for protein variability analysis,

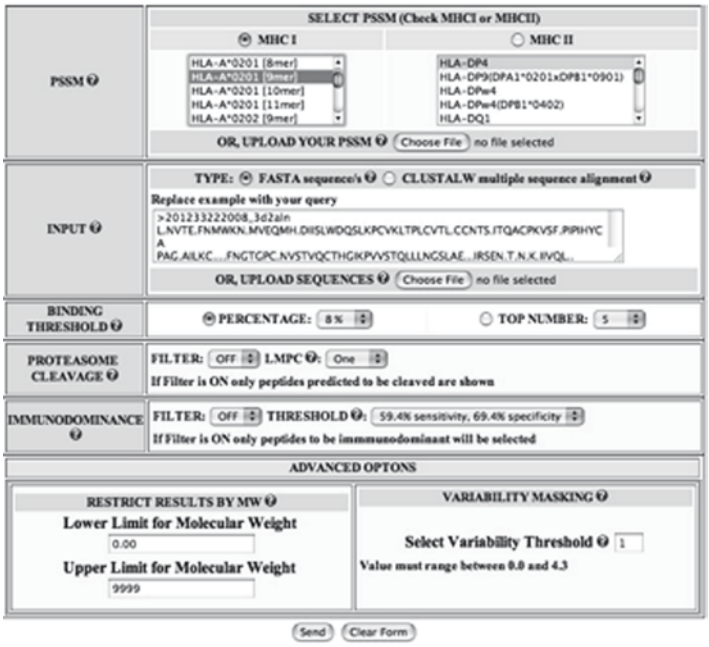
a

VARIABILITY MASKED SEQUENCE

```
Fasta sequence:
>81423282008_3d2aln
L.NVTE.FNMWKN.MVEQMH.DIISLWDQSLKPCVKLTPLCVTL.CCNTS.ITQACPK
VSF.PIPIHYCAPAG.AILKC....FNGTGPC.NVSTVQCTHGIKPVVSTQLLLNGSL
AE...IRSEN.T.N.K.IIVQL...V.I.C.RP..C......W..TL..V...L...F
....I.F...SGGD.EI..H.FNC.GEFFYCNT..LFN..........I.L.CRIKQI
INMWQ.VG.AMYAPPI.G.I.C.SNITGLLLTRDGG......E.FRPGGG.MRDNWRS
ELYKYKVV.I.
```

Run Epitope Prediction using this FASTA sequence

b



c

RANK	POS.	N	SEQUENCE	C	MW (Da)	SCORE	% OPT.
1	36	PCV	KLTPLCVTL	.CC	969.24	78.0	60.94 %
2	104	GIK	PVVSTQLLL	NGS	951.17	66.0	51.56 %
3	30	WDQ	SLKPCVKLT	PLC	970.23	51.0	39.84 %

Fig. 2 *PVS and T cell epitope predictions.* (**a**) *Variability-masked sequence*. The shown sequence obtained from an MSA of HIV-1 gp120 (consensus sequence was selected as the reference sequence). The sequence is in FASTA format and positions indicated by dots, ".", display a variability > 1.0. (**b**) *Rankpep web interface.* By clicking on the button "Run Epitope Predictions" one will directly submit this sequence for conserved T cell epitope predictions *using* the RANKPEP algorithm. (**c**) RANKPEP results for the variability-masked sequence of the gp120. Only fragments KLTPLCUTL and PVVSTQLLL were predicted to have a binding score above the threshold

which implements several features that are thought to facilitate epitope-vaccine design. Next we will discuss such features using HIV-1 as the pathogenic model.

PVS can be used to facilitate the identification of conserved T cell epitopes. As an example we used an MSA from the HIV-1 gp120 protein (see Sect. 1 for details) to first obtain a variability masked sequence (Fig. 2a), which was subsequently targeted for the prediction of CD8+ T cell epitopes restricted by the HLA I molecule A*0201 (Fig. 2b). Interestingly, only two T cell epitopes (KLTPLCVTL and PVVSTQLLL) were predicted to have a binding score above the threshold (Fig. 2c) In comparison, the complete gp120 sequence (strain H2XB2) would yield 10 different epitopes. Thus, regardless of the predictive power of RANKPEP, this strategy saves the time, effort and resources that one will need to confirm non-conserved T cell epitopes that are not as suitable for epitope-vaccine design.

PVS results can also be useful for the identification of conserved B cell epitopes, the antigenic determinants of Abs. For example, the ectodomain of HIV-1 gp41 is known to be the target of various broadly neutralizing Abs (Zolla-Pazner 2004). When PVS is run with an MSA of this protein, 7 highly conserved fragments of 6 of more residues are returned (Table 1). Interestingly, fragments WGCSGK and WLWYIK encompass the antigenic determinants of the monoclonal Abs CL3 and ZE10, both broadly neutralizing. As we can see, the targets of broadly neutralizing Abs lie within conserved fragments.

Abs only recognize solvent-exposed epitopes, and most of them are conformational –although, some can also be linear–. To help identifying solvent-exposed fragments, PVS also allows exploring the location of the conserved fragments in the 3D-structure of the protein (when available). The use of such solvent-exposed conserved fragments as immunogens greatly increases the chance of raising Abs that are both, crossreactive with the native antigen and broadly neutralizing. For example, Table 2 shows that there are only eight highly conserved fragments lying within the reported gp120 structure (PDB 1RZK, chain G).

However, by mapping the conserved gp120 fragments onto the 3D-structure (Fig. 3) one could see that only fragment 2 and fragment 3 and significant portions of fragments 1, 4 and 6 are accessible to the solvent. Therefore, these solvent-exposed

Table 1 Conserved fragments in the ectodomain of HIV-1 gp41 calculated by PVS

N	Start	End	Sequence
1	1	7	S T M G A A S
2	9	25	T L T V Q A R Q L L S G I V Q Q Q
3	27	55	N L L R A I E A Q Q H L L Q L T V W G I K Q L Q A R V L A
4	62	67	D Q Q L L G
5	69	74	W G C S G K
6	87	92	S W S N K S
7	153	158	W L W Y I K

Fragments were selected to have six or more consecutive residues with $H \leq 1$, and were obtained form an MSA of the HIV-1 gp41 ectodomain

Table 2 Conserved fragments of the HIV-1 glycoprotein gp120 calculated by PVS

N	Start	End	Sequence
1	22	44	D I I S L W D Q S L K P C V K L T P L C V T L
2	52	61	I T Q A C P K V S F
3	63	73	P I P I H Y C A P A G
4	93	119	N V S T V Q C T H G I K P V V S T Q L L L N G S L A E
5	202	209	G E F F Y C N T
6	232	242	C R I K Q I I N M W Q
7	261	273	S N I T G L L L T R D G G
8	289	303	M R D N W R S E L Y K Y K V V

Fragments were selected to have eight or more consecutive residues with $H \leq 1$, and were obtained from an MSA of HIV-1 gp120 (See Material and Methods). The "Map structure variability" task was selected and chain G of PDB 1RZK containing the 3D-coordinates of HIV-1 gp120 was entered in the server. Relevant sequence in PDB is considerably shorter than that of MSA, and only those fragments mapping within the PDB sequence are reported by the server

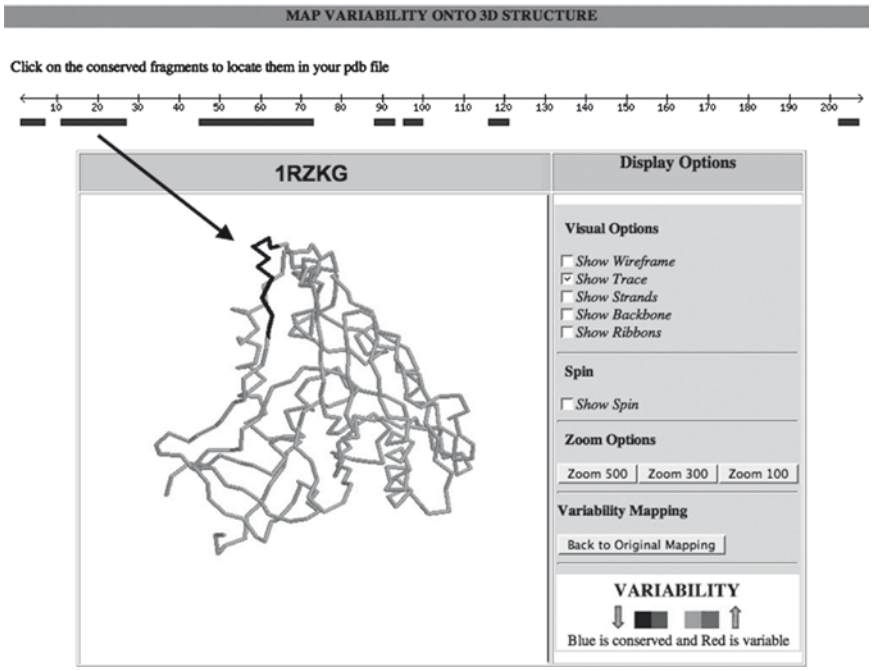


Fig. 3 *Exploring solvent accessibility of conserved fragments.* Arrow shows the location of fragment 2 (ITQACPKVSF) in the 3D-structure of gp120 (chain G of PDB 1RZK). It was located on the 3D-structure by simply clicking on the corresponding fragment shown under the linear representation of gp120

fragments are the only peptides from HIV-1 gp120 that may elicit both cross-neutralizing cross-reactive Abs with the native gp120.

Acknowledgments The work and the authors were supported by grant SAF2006-07879 from the "Ministerio de Educación y Ciencia" of Spain to PR.

References

Disis ML, Knutson KL, McNeel DG et al (2001) Clinical translation of peptide-based vaccine trials: The HER-2/neumodel. Crit Rev Immunol 21:263–274

Edgar RC (2004) MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–1797

García-Boronat M, Diez-Rivero CM, Reinherz EL et al (2008) PVS: A web server for protein sequence variability analysis tuned to facilitate conserved epitope discovery. Nucleic Acids Res 36:W35–W41

Mendis KN, David PH, Carter R (1991) Antigenic polymorphism in malaria: Is it an important mechanism for immune evasion? Immunol Today 12:A34–A37

Phillips RE, Rowland-Jones S et al (1991) Human immunodeficiency virus genetic variation that can escape cytotoxic T cell recognition. Nature 354:453–459

Reche PA, Reinherz EL (2003) Sequence variability analysis of human class I and class II MHC molecules: Functional and structural correlates of amino acid polymorphisms. J Mol Biol 331:623–641

Reche PA, Reinherz EL (2007) Prediction of peptide-MHC binding using profiles. Mol Biol 409:185–200

Reche PA, Glutting JP, Reinherz EL (2002) Prediction of MHC class I binding peptides using profile motifs. Hum Immunol 63:701–709

Reche PA, Glutting J-P, Reinherz EL (2004) Enhancement to the RANKPEP resource for the prediction of peptide binding to MHC molecules using profiles. Immunogenetics 56:405–419

Reche PA, Keskin DB, Hussey RE et al (2006) Elicitation from virus-naive individuals of cytotoxic T lymphocytes directed against conserved HIV-1 epitopes. Med Immunol 5:1

Sette A, Newman M, Livingston B et al (2002) Optimizing vaccine design for cellular processing, MHC binding and TCR recognition. Tissue Antigens 59:443–451

Shannon CE (1948) The mathematical theory of communication. Bell Syst Tech J 27(379–423): 623–656

Simpson EH (1949) Measurement of diversity. Nature 163:688

Thomsona SA, Jaramillo AB, Shoobridge M et al (2005) Development of a synthetic consensus sequence scrambled antigen HIV-1 vaccine designed for global use. Vaccine 23:4647–4657

Tsuji M, Zavala F (2001) Peptide-based subunit vaccines against preerythrocytic stages of malaria parasites. Mol Immunol 38:433–442

Weber F, Elliott RM (2002) Antigenic drift, antigenic shift and interferon antagonists: How bunyaviruses counteract the immune system. Virus Res 88:129–136

Wu TT, Kabat EA (1970) An analysis of the sequences of the variable regions of Bence Jones proteins and myeloma light chains and their implications for antibody complementarity. J Exp Med 132:211–250

Zolla-Pazner S (2004) Identifying epitopes of HIV-1 that induce protective antibodies. Nat Rev Immunol 4:199–210