BIOINFORMATICS APPLICATIONS NOTE Vol. 21 no. 9 2005, pages 2140–2141
doi:10.1093/bioinformatics/bti269

Databases and ontologies

EPIMHC: a curated database of MHC-binding peptides for customized computational vaccinology

Pedro A. Reche[1,2,*], Hong Zhang[1], John-Paul Glutting[1] and Ellis L. Reinherz[1,2]

[1]Laboratory of Immunobiology and Department of Medical Oncology, Dana-Farber Cancer Institute and [2]Department of Medicine, Harvard Medical School, 77 Avenue Louis Pasteur, Boston, MA 02115, USA

Received on December 9, 2004; revised on January 7, 2005; accepted on January 8, 2005
Advance Access publication January 18, 2005

ABSTRACT

Summary: EPIMHC is a relational database of MHC-binding peptides and T cell epitopes that are observed in real proteins. Currently, the database contains 4867 distinct peptide sequences from various sources, including 84 tumor-associated antigens. The EPIMHC database is accessible through a web server that has been designed to facilitate research in computational vaccinology. Importantly, peptides resulting from a query can be selected to derive specific motif-matrices. Subsequently, these motif-matrices can be used in combination with a dynamic algorithm for predicting MHC-binding peptides from user-provided protein queries.

Availability: The EPIMHC database server is hosted by the Dana-Farber Cancer Institute at the site http://immunax.dfci.harvard.edu/bioinformatics/epimhc/

Contact: reche@research.dfci.harvard.edu

INTRODUCTION

T cell immune responses are driven by antigenic peptides (T cell epitopes) in the context of MHC molecules (Paul, 1998). Therefore, comprehensive databases of MHC-binding peptides are important tools for the analysis of binding to MHC molecules and the development of peptide-based immunotherapies. Current examples of databases of MHC-binding peptides include MHCPEP (Brusic *et al*., 1998), SYFPEITHI (Rammensee *et al*., 1999), JenPep (Blythe *et al*., 2002), MHCBN (Bhasin *et al*., 2003) and FIMM (Schonbach *et al*., 2002). MHCPEP is the oldest database, and it has served as the largest source of data for the other databases. Existing resources have their limitations. In particular, MHCPEP has not been updated since 1998. Peptide annotations in recent databases have not been enhanced with regard to those in the MHCPEP database, and the choices for extraction and analysis of the data are quite limited.

In response to these limitations, we have created the EPIMHC database. The database was compiled from peptides and annotations collected from the above resources and the literature. MHC-binding peptides obtained from SYFPEITHI were all considered to be high binders. Peptide annotations in EPIMHC follow the basic scheme of the MHCPEP database. However, EPIMHC only contains MHC-binding peptides that occur in actual proteins, and it is structured as a relational database of unique MHC<=>peptide-sequence pairs. Also in EPIMHC, peptide annotations have been enhanced with regard to related resources to include new information additional to the usual MHC-binding specificity and T cell activity of peptides. Most importantly, the processing of the peptide and its source (organism and protein sequence) are also annotated in EPIMHC. The processing field is to indicate whether MHC-binding peptides are processed and presented from their protein sources *in vivo* (annotated as natural). EPIMHC also provides links to relevant databases such as PUBMED, IMGT/HLA and GenBank. The database contains 4875 distinct MHC-binding peptides, of which 2224 are T cell epitopes (1267 MHCI-restricted and 957 MHCII-restricted). Peptides in the database target a total of 378 MHC specificities (226 MHCI and 152 MHCII), the majority of which are human (176 MHCI and 119 MHCII). A functionally important subset of epitopes in the database consists of 67 $CD8^+$ and 17 $CD4^+$ T cell epitopes derived from tumor-associated antigens (TAAs).

EPIMHC can be accessed through a flexible web interface that allows retrieval and display of data according to multiple criteria (Fig. 1A). Result queries (Fig. 1B) can be saved in a variety of TEXT formats. Unlike related databases, EPIMHC allows computational experimentation. Specifically, peptides selected from result queries (Fig. 1B) can be used to generate motif-matrices. Subsequently, these motif-matrices can be used to predict related sequences from a protein query using a dynamic search algorithm (Fig. 1C and D). Motif-matrices are generated as position-specific scoring matrices (PSSMs). PSSMs have previously been shown to be adequate tools for prediction of peptide–MHC binding (Nielsen *et al*., 2004; Reche *et al*., 2002, 2004). However, the prediction results employing PSSMs are linked to the specific peptides used to build the PSSM. Thus, the 'create-matrix' feature in EPIMHC empowers users to carry out tailored prediction of MHC-binding peptides. At the moment, PSSM can only be derived from peptides of the same length. This size limitation together with structural features of the peptide binding to MHC molecules (Reche *et al*., 2002, 2004) recommends applying the 'create-matrix' feature only to MHC class I binding peptides.

In sum, databases of MHC-binding peptides are important tools for studying T cell based immunorecognition. This utility has been improved in EPIMHC by providing curated data, enhanced annotations and a design that facilitates the extraction and analysis of data. Furthermore, unlike any related resource, EPIMHC empowers the users to derive their own predictors of MHC-binding, providing a framework for tailored prediction of MHC-binding ligands.

*To whom correspondence should be addressed.

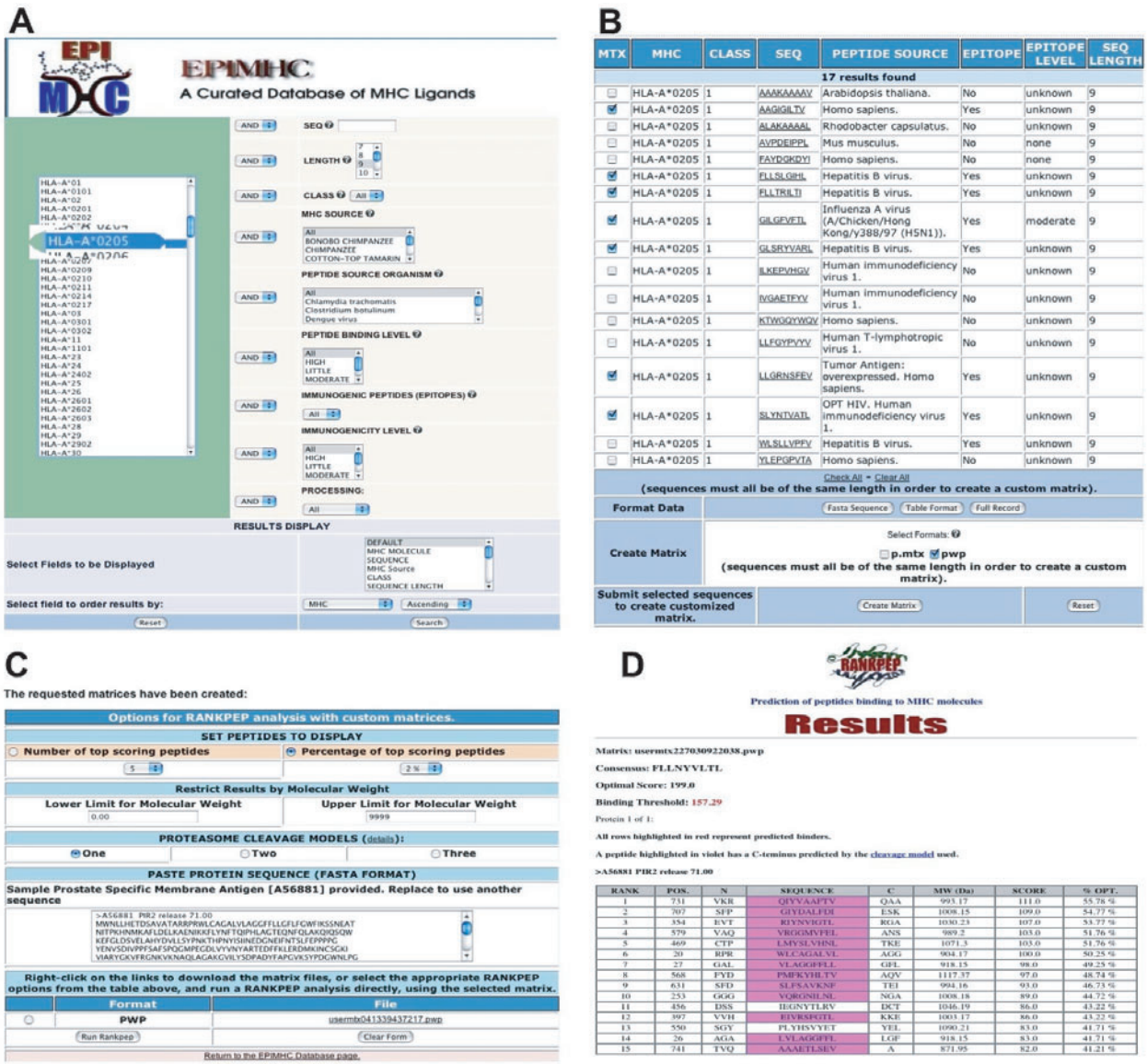

Fig. 1. (**A**) EPIMHC web interface. (**B**) Output from an example query (peptide binders to HLA-A*0205 of nine residues in length). Data can be saved in various text formats. In this example, peptides eliciting T cell activity (epitopes) have been selected to create a motif-matrix. (**C**) Page resulting from selecting the create matrix option. This page allows the prediction of MHC-binding peptides from a user-provided protein using the previously created PSSM. (**D**) Predicted MHC-binding peptides at a 2% threshold from a prostate-specific membrane antigen (GenBank: A56881). Peptides highlighted in violet contain a C-terminus predicted to result from immunoproteasomal cleavage (Reche *et al*., 2004).

ACKNOWLEDGEMENTS

This manuscript was supported by NIH grants AI50900 and AI43649, and the Molecular Immunology Foundation.

REFERENCES

Bhasin,M., Singh,H. and Raghava,G.P. (2003) MHCBN: a comprehensive database of MHC binding and non-binding peptides. *Bioinformatics*, **19**, 665–666.

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

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

Nielsen,M., Lundegaard,C., Worning,P., Hvid,C.S., Lamberth,K., Buus,S., Brunak,S. and Lund,O. (2004) Improved prediction of MHC class I and class II epitopes using a novel Gibbs sampling approach. *Bioinformatics*, **20**, 1388–1397.

Paul,W.E. (1998) *Fundamental Immunology*. 4th edn. Raven Press, NY.

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

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. 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.

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