368–371 *Nucleic Acids Research, 1998, Vol. 26, No. 1*

MHCPEP, a database of MHC-binding peptides: update 1997

Vladimir Brusic*, George Rudy and Leonard C. Harrison

The Walter and Eliza Hall Institute, Parkville, Victoria 3050, Australia

Received October 3, 1997; Accepted October 6, 1997

ABSTRACT

MHCPEP (http://wehih.wehi.edu.au/mhcpep/) is a curated database comprising over 13 000 peptide sequences known to bind MHC molecules. Entries are compiled from published reports as well as from direct submissions of experimental data. Each entry contains the peptide sequence, its MHC specificity and where available, experimental method, observed activity, binding affinity, source protein and anchor positions, as well as publication references. The present format of the database allows text string matching searches but can easily be converted for use in conjunction with sequence analysis packages. The database can be accessed via Internet using WWW or FTP.

INTRODUCTION

MHCPEP, a database of peptides that bind to MHC class I or II molecules was established in 1994 (1). Peptide binding to an MHC binding site is a prerequisite for T-cell recognition of peptide, although not all binding peptides function as T-cell epitopes. Allele-specific motifs for peptides binding MHC class I are now well documented (2). Although motifs for peptides binding MHC class II have been reported (2–5), they are generally less well defined. The comprehensive compilation of binding peptides in MHCPEP enables the analysis of sequence properties governing binding to MHC molecules and the identification of T-cell epitopes. It should also facilitate research on antigen processing and transport, the mechanisms of T-cell receptor activation and the development of specific approaches to immunotherapy. More than 3000 new entries have been added since the last report (6).

DESCRIPTION

Version 1.3 of MHCPEP has 13 423 entries (as of September 1997) compiled from published sources or directly submitted experimental data. A few peptides binding non-classical MHC molecules (e.g., mouse Qa-2a) are included. The majority of entries contain human or mouse MHC binding peptides. There are also a small number of entries containing peptides that bind rat, rhesus macaque, chimpanzee or goat MHC. Each report of a peptide sequence is assigned to a separate entry identified by a unique entry ID. Entries consist of 12 fields composed of the field name followed by a colon (:) delimiter and the field value. Field values may be textual, numeric or empty. Fields are written one to a row and delimited by a hash (#). Entries are delimited by ellipses (...). Representative entries are shown in Figure 1. A description of each entry field follows.

Entry ID

A unique identifier exists for each entry. The format is: [>organism][class][xxxx]. The ‘organism’ is designated by a three letter code (e.g. HUM for human, MUS for mouse); ‘class’ is a single digit number; the right side is an unique four digit hexadecimal number.

MHC molecule

The designation of the MHC molecule, according to the nomenclature of Klein *et al*. (7), is followed by the specific allele (where known) within brackets. Human alleles are designated according to the DNA sequence nomenclature (8). This field also shows MHC class and host organism. The format is: [MHC molecule], [MHC class], [(host)].

Method

Peptide binding to MHC molecules can be determined indirectly by T-cell activity based assays or directly by biochemical methods. T-cell recognition of MHC class I-bound peptides is usually detected by cytotoxicity assay, and of MHC class II-bound peptides by proliferation assay. Biochemical methods include stabilization assays, competitive inhibition assays to purified MHC molecules or cells bearing MHC, or elution followed by sequencing.

Activity

The ‘activity’ of a peptide is a semi-quantitative measure of its immunogenic ‘potency’. For an MHC class I-bound peptide, ‘activity’ is a measure of the extent of lysis by cytotoxic T-cells of target cells displaying the MHC class I–peptide complexes. A peptide is considered immunogenic if it mediates killing of at least 15% of the cells that display it. The ‘activity’ is expressed as the PD_{50}, (PD, peptide dose), the concentration of peptide giving 50% of maximum specific lysis, and is given a descriptive value of none, little, moderate, high, immunogenic-not-quantified or unknown. A $PD_{50} > 10$ μM is considered non-immunogenic and assigned ‘none’

*To whom correspondence should be addressed. Tel: +61 3 9345 2588; Fax: +61 3 9347 0852; Email: vladimir@wehi.edu.au

```
>HUM10075#
MHC MOLECULE: HLA-B27, CLASS-1, (HUMAN)#
METHOD: competitive inhibition/reference#
ACTIVITY: yes, ?#
BINDING: yes, ?#
SOURCE: HIV GAG p24 protein (265-276)#
DB REFERENCE: SWISS:(GAG_HV1A2,GAG_HV1B1,GAG_HV1B5,GAG_HV1BR,GAG_HV1C4,#
&                   GAG_HV1H2,GAG_HV1J3,GAG_HV1JR,GAG_HV1MA,GAG_HV1MN,#
&                   GAG_HV1N5,GAG_HV1ND,GAG_HV1OY,GAG_HV1PV,GAG_HV1RH,#
&                   GAG_HV1W2,GAG_HV1Y2,GAG_HV1Z2)#
&             PIR1:(FOVWH3,FOVWLV,A44001,FOLJND,FOVWH4,FOVWA2,FOVWVL,A38068)#
&             PIR2:(S60704,S60702,S60699,S54377,S60708,S60698,S60703,S60697)#

&             PIR3:(S19598,S33979)#
ANCHOR POSITIONS: 2#
REFERENCES: rammensee93a,carreno92a,jardetzky91a,huet90a,parker92b,buseyne93a,#
&           brander95b,hivdb97a#
COMMENT:#
SUMMARY: HLA-B27,actyesu,bindyesu,KRWIILGLNK#
SEQUENCE: KRWIILGLNK*#
...#
>HUM20076#
MHC MOLECULE: HLA-DR1(HLA-DRB1*0101), CLASS-2, (HUMAN)#
METHOD: binding assays and proliferation assays#
ACTIVITY: yes, moderate#
BINDING: yes, ?#
SOURCE: FLU M1 (17-29)#
DB REFERENCE: SWISS:(VMT1_IAANN,VMT1_IABAN,VMT1_IACKB,VMT1_IAFOW,VMT1_IAFPR,#
&                   VMT1_IAFPW,VMT1_IALE1,VMT1_IALE2,VMT1_IAMAN,VMT1_IAPOC,#
&                   VMT1_IAPUE,VMT1_IAUDO,VMT1_IAUSS,VMT1_IAWIL)#
&             PIR1:(MFIV,MFIVC,MFIV61,B45539,MFIVWS,MFIV1K,MFIV1F,MFIV1M,#
&                   JN0392)#
&             PIR2:(S04056,S04054,S04052,S07429,S07945,S04058,S04050,S04060#
&                   S14616)#
ANCHOR POSITIONS:#
REFERENCES: hill91a,rothbard88a#
COMMENT:#
SUMMARY: HLA-DR1,actyesm,bindyesu,SGPLKAEIAQRLE#
SEQUENCE: SGPLKAEIAQRLE*#
...#

.
.
.
```

Figure 1. Representative MHCPEP entries.

as the value of the field 'ACTIVITY'. For an MHC class II-bound peptide, 'activity' is a measure of the extent of T-cell proliferation induced by cells displaying the MHC class II–peptide complexes. Again 'activity' is expressed as PD_{50}, now defined as the concentration of peptide giving 50% of maximum proliferation. The range of values of the field 'ACTIVITY' is given in Table 1.

Table 1. Range of values assigned to the field 'ACTIVITY'

PD_{50}	Value
> 10 μM	none
10 μM–100 nM	yes, little
100 nM–1 nM	yes, moderate
< 1 nM	yes, high
Immunogenic but unknown	yes, ?
Immunogenicity unknown	?

Binding

It is assumed that all 'active' peptides also bind; if no measure of binding is reported for an 'active' peptide, the value 'yes, ?' is assigned to the field 'BINDING'. As several different methods exist for determining binding affinity, only a descriptive value is assigned to the data; the user should consult the original source for more specific details. The same scale as for 'activity' is used (none, little, moderate, high, unknown).

Source

MHC binding peptides are fragments of larger proteins. This field indicates the parent protein with the start and end positions of the fragment. Synthetic peptides (e.g., those generated by mutations of a naturally occurring sequence) are designated by the word 'homologue'.

DB reference

This field specifies the name of the source protein(s) as it appears in the major protein databases: SWISS-PROT (9) and PIR (10), versions 34.0 and 53.0, respectively. These databases have been searched for sequences that match the MHC peptide entry sequence. This field may spread over several lines of text. The continuation of the field is designated by an ampersand (&) as the first character in the line.

Anchor positions

Presumed anchor residues are numbered relative to the N-terminus of the peptide sequence. The main criterion for determining the values for this field is conformance to proposed binding motifs.

References

A separate list of references to the published sources of entry sequences is supplied with the database. The value of the 'REFERENCES' field is a set of reference words, each of which consists of the first author's surname, year of publication or submission (two digit number) and an identifier (single letter). Each reference word uniquely designates a single reference. The format of the 'REFERENCES' field is: [author][year][x]. This field may also spread over several lines of text.

Comment

This field is reserved for any relevant comments or observations.

Summary

The summary field is a one-line description of the main fields of an entry (MHC molecule, activity, binding and peptide sequence), which is useful for rapid indexing of the database.

Sequence

The sequence of the peptide is the one actually reported, not the minimum or optimum sequence. Therefore, a given T-cell epitope may be found within several entries representing different sequences which overlap or include it. The value for this field has an asterisk (*) following the C-terminal residue. The letter 'X' is used to represent ambiguity or an unknown residue. In some instances it is not possible to distinguish certain amino acids, e.g., tandem mass spectrometry does not distinguish leucine (L) and isoleucine (I). In such cases, separate entries are created and the ambiguity is noted in the 'COMMENT' field.

ACCURACY AND COMPLETENESS OF THE DATA

MHCPEP is largely compiled from published reports. However, numerous potential sources of error exist. Double-checking, comparison with original papers, comparison with other databases and multiple entry of the same sequence from different sources have been used to minimise errors. Some entries describing the same peptide may have different values in fields other than 'SEQUENCE' when derived from independent sources. Differences between T-cell clones used in experiments have not been considered; in cases where an MHC-bound peptide is recognised by any of the clones it is entered into the database. Observations regarding clones which do or do not recognise the peptide are included in the comment field. The database has a degree of redundancy reflecting the variety of ways of detecting MHC-bound peptides. Earlier reports of MHC binding peptides were less specific: reported peptides were usually longer than the optimum size and the fine specificity of MHC molecules was not determined. The quality of data in recent reports has improved.

DATABASE ACCESS

MHCPEP is accessible via Internet using WWW or FTP to the following respective WEHI addresses: http://wehih.wehi.edu.au/mhcpep/ ; ftp.wehi.edu.au/pub/biology/mhcpep/ ; Gopher access option is not longer available.

MHCPEP has been linked with SWISS-PROT and PIR databases and also with references file via the 'DB REFERENCE' and 'REFERENCES' fields, respectively. On-line text

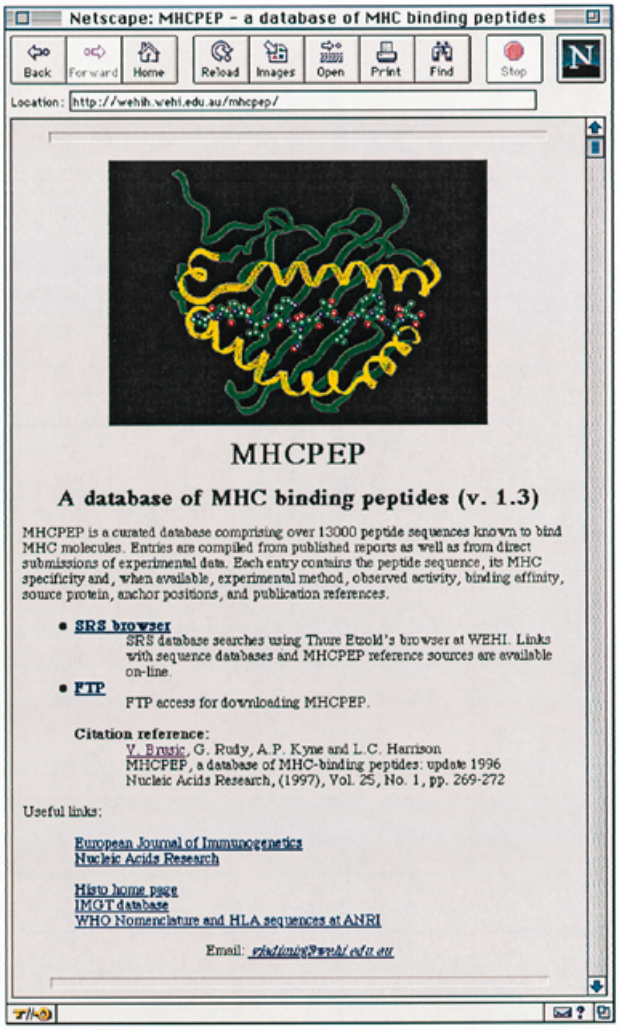

Figure 2. MHCPEP WWW home page.

string searches and retrieval of MHCPEP entries as well as their linked SWISS-PROT, PIR and reference entries are now available using Sequence Retrieval System (SRS) (11), accessible through the WEHI WWW home page (Fig. 2). MHCPEP has been converted for local use with sequence analysis packages (e.g., GCG; 12) which allows more sophisticated sequence analysis.

Authors who wish to cite MHCPEP should quote this paper as the reference.

For queries and comments regarding the MHCPEP database contact Vladimir Brusic (preferably by electronic mail) at the address given at the start of this paper.

FURTHER DEVELOPMENTS

A summary of MHCPEP contents is given in Table 2. The growing numbers of peptides known to bind to a specific MHC molecule facilitates building of predictive models for determination of novel T-cell epitope candidate peptides. Predictive models utilizing MHCPEP have been successfully applied in cancer (13) and autoimmunity (14,15) research.

Table 2. Summary of MHCPEP contents

HOST/allelic region	Number of entries	
	MHC Class I	MHC Class II
Human	4617	5394
/HLA-A	2530	–
/HLA-B	1894	–
/HLA-C	64	–
/HLA-DR	–	4545
/HLA-DP	–	106
/HLA-DQ	–	550
Mouse	1145	2213
/H-2K	682	–
/H-2D	266	–
/H-2L	120	–
/H-2A	–	1252
/H-2E	–	666
Rat	15	8
Chimpanzee	13	2
Rhesus Macaque	9	5
Goat	0	2

Knowledge of MHC–peptide interactions continues to expand rapidly, together with the number of methods for determining binding peptides. Combining data generated by diverse sources imposes additional standardization requirements on the further developments. The main considerations in the improvement of MHCPEP include: (i) retrieval of data, (ii) internal data cleansing, (iii) linkage to other databases containing MHC- or antigen-related data, and (iv) extraction of high-level relationships hidden within data. Optimizing these requires a more complex structure than the current MHCPEP. We are investigating the utility of a structure that integrates a knowledge base (KB), a WWW interface, and a set of computational tools similar to the RIBOWEB system (16). A KB comprising structured hierarchical representations of MHC data, methods and literature sources, is currently being developed. This set of computational tools should facilitate further applications of the database.

ACKNOWLEDGEMENT

We thank Russ Altman of Stanford University for helpful suggestions and guidance about Knowledge Base development.

REFERENCES

1 Brusic, V., Rudy, G. and Harrison, L.C. (1994) *Nucleic Acids Res.*, **22**, 3663–3665.
2 Rammensee, H.G., Friede, T. and Stevanovic, S. (1995) *Immunogenetics*, **41**, 178–228.
3 Hammer, J., Valsasnini, P., Tolba, K., Bolin, D., Higelin, J., Takacs, B. and Sinigaglia, F. (1993) *Cell.*, **74**, 197–203.
4 Chicz, R.M., Urban, R.G., Gorga, J.C., Vignali, D.A., Lane, W.S. and Strominger, J.L. (1993) *J. Exp. Med.*, **178**, 27–47.
5 Harrison, L.C., Honeyman, M.C., Tremblau, S., Gregori, S., Gallazzi, F., Augstein, P., Brusic, V., Hammer, J. and Adorini, L. (1997) *J. Exp. Med.*, **185**, 1013–1021.
6 Brusic, V., Rudy, G., Kyne, A.P. and Harrison, L.C. (1997) *Nucleic Acids Res.*, **25**, 269–271.
7 Klein, J., Bontrop, R.E., Dawkins, R.L., Erlich, H.A., Gyllensten, U.B., Heise, E.R., Jones, P.P., Parham, P., Wakeland, E.K. and Watkins, D.I. (1990) *Immunogenetics*, **31**, 217–219.
8 Bodmer, J.G., Marsh, S.G., Albert, E.D., Bodmer, W.F., Bontrop, R.E., Charron, D., Dupont, B., Erlich, H.A., Mach, B., Mayr, W.R. *et al.* (1995) *Hum. Immunol.*, **43**, 149–164.
9 Bairoch, A. and Apweiler, R. (1997) *Nucleic Acids Res.*, **25**, 21–25. [See also this issue *Nucleic Acids Res.* (1998), **26**, 38–42.]
10 George, D.G., Barker, W.C., Mewes, H.W. and Tsugita A. (1997) *Nucleic Acids Res.*, **25**, 17–20. [See also this issue *Nucleic Acids Res.* (1998), **26**, 27–32.]
11 Etzold, T. and Argos, P. (1993) *Comput. Applic. Biosci.*, **9**, 49–57.
12 Program Manual for the Wisconsin Package, Version 8, September 1994, Genetics Computer Group, Madison, Wisconsin.
13 Ramakrishna, V., Negri, D.R.M., Brusic, V., Fontanelli, R., Canevari, S., Bolis, G., Castelli, C. and Parmiani, G. (1997) *Int. J. Canc.*, **73**, 143–150.
14 Honeyman, M.C., Brusic, V. and Harrison, L.C. (1997) *Ann. Med.*, **29**, 401–404.
15 Brusic, V., Rudy, G., Honeyman, M.C., Hammer, J. and Harrison, L.C. (1998) *Bioinformatics.*, **14**, in press.
16 Chen, R.O., Feliciano, R. and Altman, R.B. (1997) *ISMB-97*, p. 84–87, AAAI.