2005-07-01

PREDBALB/c: a system for the prediction of peptide binding to H2d molecules, a haplotype of ...

W180–W183 *Nucleic Acids Research, 2005, Vol. 33, Web Server issue*
doi:10.1093/nar/gki479

PRED$^{BALB/c}$: a system for the prediction of peptide binding to H2^{d} molecules, a haplotype of the BALB/c mouse

Guang Lan Zhang[1,2], Kellathur N. Srinivasan[3,4], Anitha Veeramani[1], J. Thomas August[3] and Vladimir Brusic[1,5,*]

[1]Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613, [2]School of Computer Engineering, Nanyang Technological University, Singapore 6397984, [3]Department of Pharmacology and Molecular Sciences, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA, [4]Division of Biomedical Sciences, Johns Hopkins in Singapore, #02–01 The Nanos, 31 Biopolis Way, Singapore 138669 and [5]School of Land and Food Sciences and the Institute for Molecular Bioscience, University of Queensland, Brisbane QLD 4072, Australia

Received February 14, 2005; Revised February 15, 2005; Accepted April 15, 2005

ABSTRACT

PRED$^{BALB/c}$ is a computational system that predicts peptides binding to the major histocompatibility complex-2 (H2^{d}) of the BALB/c mouse, an important laboratory model organism. The predictions include the complete set of H2^{d} class I (H2-K^{d}, H2-L^{d} and H2-D^{d}) and class II (I-E^{d} and I-A^{d}) molecules. The prediction system utilizes quantitative matrices, which were rigorously validated using experimentally determined binders and non-binders and also by *in vivo* studies using viral proteins. The prediction performance of PRED$^{BALB/c}$ is of very high accuracy. To our knowledge, this is the first online server for the prediction of peptides binding to a complete set of major histocompatibility complex molecules in a model organism (H2^{d} haplotype). PRED$^{BALB/c}$ is available at http://antigen.i2r.a-star.edu.sg/predBalbc/.

INTRODUCTION

The T cells of the immune system recognize antigens as short peptide fragments (T-cell epitopes) derived from self or foreign proteins. Self proteins include all proteins produced by the cells of the host. Foreign peptides are derived from pathogens, environmental antigens, tumor cells and transplanted tissue. Immune recognition of both self and foreign antigens involves proteolytic processing of antigens, binding of the peptide epitopes by major histocompatibility complex (MHC) molecules and presentation of selected peptide epitopes on the cell surface to activate of T cells (1–4). Cytotoxic T cells (CD8^{+}) recognize peptides bound to MHC class I molecules and helper T cells (CD4^{+}) recognize antigen in the context of MHC class II molecules. MHC class I molecules are present in all cells and bind mainly endogenous peptides (those produced within the presenting cell), whereas class II molecules are present mainly in cells that recognize foreign proteins, such as macrophages and dendritic cells. T-cell epitopes are critical for the immune response to infectious, autoimmune, allergic and neoplastic disease. They have been studied for the development of peptide-based vaccines (5) and may also be important in the diagnosis of pathogens. It is estimated that between 1 and 5% of all peptides can bind a particular MHC molecule (6). Traditional approaches to the identification of T-cell epitopes that involve various biochemical and functional assays of overlapping peptides derived from proteins of interest are costly and not applicable to large-scale studies. Accurate predictions using computer models help speed up the identification of T-cell epitopes (7), minimize the number of experiments necessary and enable systematic scanning for candidate T-cell epitopes from larger sets of protein antigens, such as those encoded by complete viral genomes (8).

The BALB/c inbred laboratory mouse strain is one of the most commonly used animal models in immunological studies and has been used extensively in vaccine research (9,10). BALB/c mice express three class I (H2-K^{d}, H2-L^{d} and H2-D^{d}) and two class II (I-A^{d} and I-E^{d}) molecules. Several publicly available prediction systems for MHC class I and class II binding peptides provide the prediction models for (histocompatibility complex-2) H2^{d} alleles. SYFPEITHI (11) has H2-K^{d} and H2-L^{d} models, BIMAS (12) has H2-K^{d}, H2-D^{d} and H2-L^{d} models, and RANKPEP (13) has models for

*To whom correspondence should be addressed: Tel: +65 96 212 415; Fax: +65 6774 8056; Email: vladimir@i2r.a-star.edu.sg

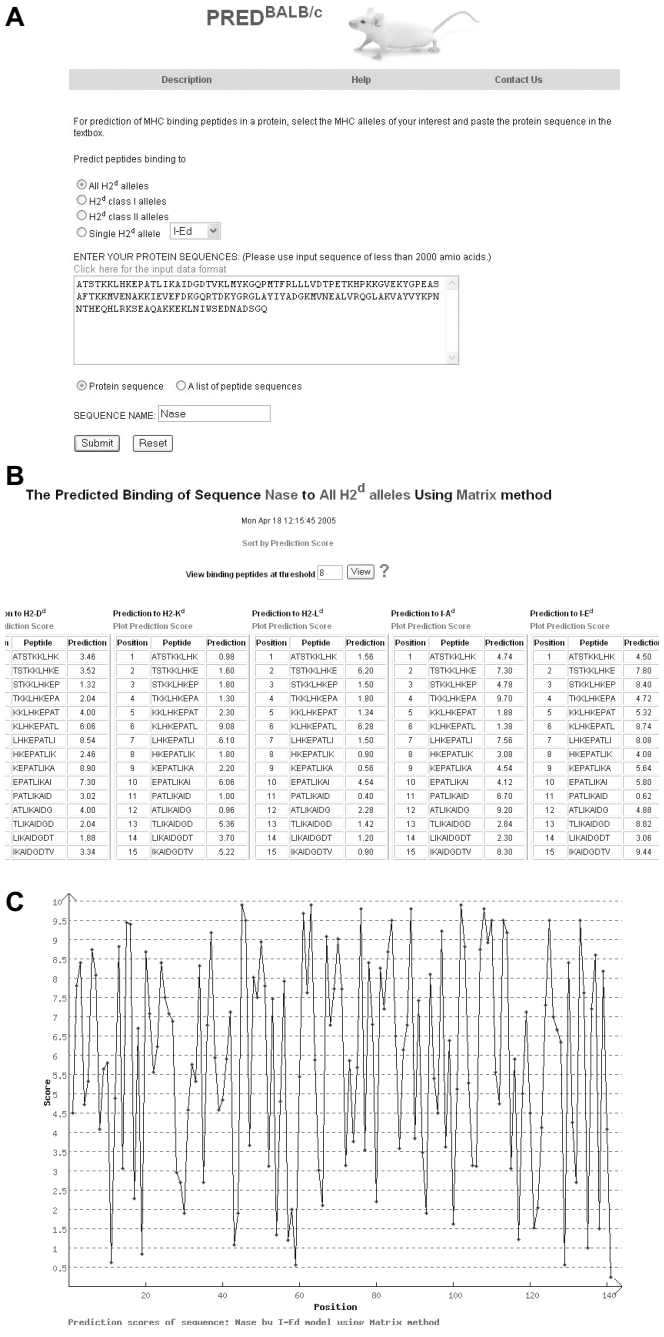

Figure 1. An example of the output pages of PRED$^{BALB/c}$ when the input is a single protein. The input protein sequence is Staphylococcal nuclease (Nase), the sequence type chosen is 'protein sequence' and the H2^d alleles of interest are all H2^d alleles. (**A**) The input page. (**B**) The main result page. The input sequence is decomposed into overlapping 9mers for the prediction of binding scores to each allele. (**C**) Graphical view of the predicted binding scores to the I-E^d molecule.

Table 1. Number of peptides in the training sets for H2^d matrices

	Class I			Class II	
	D^d	K^d	L^d	A^d	E^d
Binders (B)	93	139	73	127	191
Non-binders (NB)	331	105	91	963	627
Ratio B/NB	28%	132%	87%	13%	30%

all H2^d molecules. SYFPEITHI uses binding motifs, whereas BIMAS and RANKPEP use binding matrices. These are general servers that contain prediction models for a range of MHC molecules in human and mouse, and several other mammalian organisms, but the accuracies of individual models have not been determined. On the other hand, quantitative matrices for H2-K^b, H2-D^b, H2-L^d and H2-K^k (14,15) have been developed and validated.

PRED$^{BALB/c}$ is a computational system for the prediction of peptides binding to all five MHC molecules in BALB/c mice (H2^d) class I (H2-K^d, H2-L^d and H2-D^d) and class II (I-A^d and I-E^d) that allows analysis of proteins for the presence of binding motifs to all five H2^d molecules in parallel. We derived the initial quantitative matrices for PRED$^{BALB/c}$ using logarithmic equations based on the frequency of amino acids at specific positions within the training set of 9mer peptides as described previously (16). The initial matrices were refined by including information on the consensus (11) and other binding motifs, for example, H2-K^d binding peptides that have I, L or M at major anchor position p2 (17). The anchor positions (e.g. positions 2 and 9 in K^d binding peptides) were assigned higher weights than other positions. In addition, the prediction scores were inspected for all permissible amino acids at each of the anchor positions. All amino acids at the anchor positions other than the permissible ones were assigned low scores to exclude peptides with non-permissible amino acids from the list of predicted binders. The final binding scores were normalized to a scale of 1–9 and the final models were tested and validated rigorously. To our knowledge, PRED$^{BALB/c}$ is the first online server for the prediction of peptides binding to a complete set of MHC molecules in a model organism (H2^d haplotype).

SYSTEM DESCRIPTION

The training data containing binding and non-binding peptides were extracted from MHCPEP (18), MHCBN (19), SYFPEITHI (11), JenPep (20) and a set of non-binders (V. Brusic, unpublished data). The 9mer peptides were used for deriving H2^d class I matrices because the majority of peptides that bind these molecules are 9 amino acids long (21). Although the majority of H2^d class II binding peptides are 12–20 amino acids long, their binding cores are 9 amino acids long (22,23). An iterative elimination method starting from I-A^d and I-E^d motifs in SYFPEITHI (11) was used to identify the core 9mer regions from long peptides (K.N. Srinivasan, G.L. Zhang, A. Veeramani, J.T. August and V. Brusic, manuscript in preparation). The number of peptides in the training sets is shown in Table 1. No one method of predicting peptide–MHC binding consistently outperforms the rest and the most appropriate predictive model depends on the amount of data available in Ref. (16). In our previous work, an artificial neural network method and hidden Markov models were applied to the prediction of human leukocyte antigens (HLAs) binding peptides (24,25), where more training data were available. Because relatively small training data sets are available for H2^d, we adopted matrix models as the prediction method.

Five 9 × 20 matrices were built, one for each of the five H2^d alleles, and 10-fold cross-validations were performed to test the accuracy of the prediction models. The results show that PRED$^{BALB/c}$ predicts peptides binding to I-E^d, I-A^d and H2-K^d

with excellent accuracy [area under the receiver operating characteristic (ROC) curve, $A_{ROC} \geq 0.90$], and to H2-D^d and H2-L^d with good accuracy ($A_{ROC} \geq 0.85$). The models were also rigorously tested using experimentally known peptides from viral, prokaryotic and eukaryotic origins (14,26–29) and validated by *in vivo* studies using severe acute respiratory syndrome (SARS) nucleocapsid and HIV GAG proteins. The $H2^d$ models accurately predicted 11 out of 12 ELISPOT positive regions from BALB/c mice splenocytes immunized with SARS nucleocapsid DNA vaccines (data not shown).

The web interface of $PRED^{BALB/c}$ uses a set of graphical user interface forms. The interface was built using a combination of Perl, CGI and C programs. $PRED^{BALB/c}$ has been implemented in a SunOS 5.9 UNIX environment.

USING $PRED^{BALB/C}$

Users have the option to predict peptides binding to all $H2^d$ molecules, $H2^d$ class I molecules, $H2^d$ class II molecules or a single $H2^d$ molecule. The default selection on the webpage is 'all $H2^d$' molecules. To perform predictions using $PRED^{BALB/c}$, the user must paste a protein sequence into a textbox and assign a name to the sequence. The sequence must contain between 9 and 2000 amino acids. If the prediction is run with an input sequence containing symbols other than the 20 amino acid codes (spaces and carriage returns are allowed) or the total sequence length is outside the 9–2000 amino acids range, an error message will be displayed. The input can be either a contiguous protein sequence or a list of peptides, one per line. The default selection on the webpage is 'protein sequence' (Figure 1A), which means the input sequence is treated as a contiguous protein sequence (carriage returns and line breaks will be ignored). The $PRED^{BALB/c}$ input processing program decomposes protein sequence (or the list of peptides) into a series of 9mer peptides overlapping by eight amino acids. Individual 9mer peptides are then submitted for prediction. Predicted binding scores for all 9mers are displayed in the result tables (Figure 1B). The 9mer binding scores are within the range 0–10; the higher the score, the higher the probability of the peptide being a binder. $PRED^{BALB/c}$ has the option to plot the binding scores of all the overlapping 9mer peptides as a graph, in which the *x*-axis represents the start position of a 9mer peptide and the *y*-axis represents the binding score of the 9mer peptide (Figure 1C). The user can sort the peptides by their binding scores and choose to view only predicted binders with binding scores above a certain threshold. To assess prediction accuracy, we used measures of sensitivity SE = TP/(TP + FN) and specificity SP = TN/(TN + FP) (TP: true positives; TN: true negatives; FP: false positives; FN: false negatives). The higher the value of SP, the lower is the value of SE, which results in lower number of both TPs and FPs. The lower the value of SP, the higher is the value of SE, which results in higher number of both TPs and FPs. Raw binding scores are mapped to a linear scale that corresponds to SP values, and therefore the prediction thresholds across different models have similar meaning. For example, when a user sets the threshold to 8, the specificity of the predictions to all five alleles is 0.8. The corresponding sensitivities of each model can be viewed at http://antigen.i2r.a-star.edu.sg/predBalbc/HTML/specificity.html.

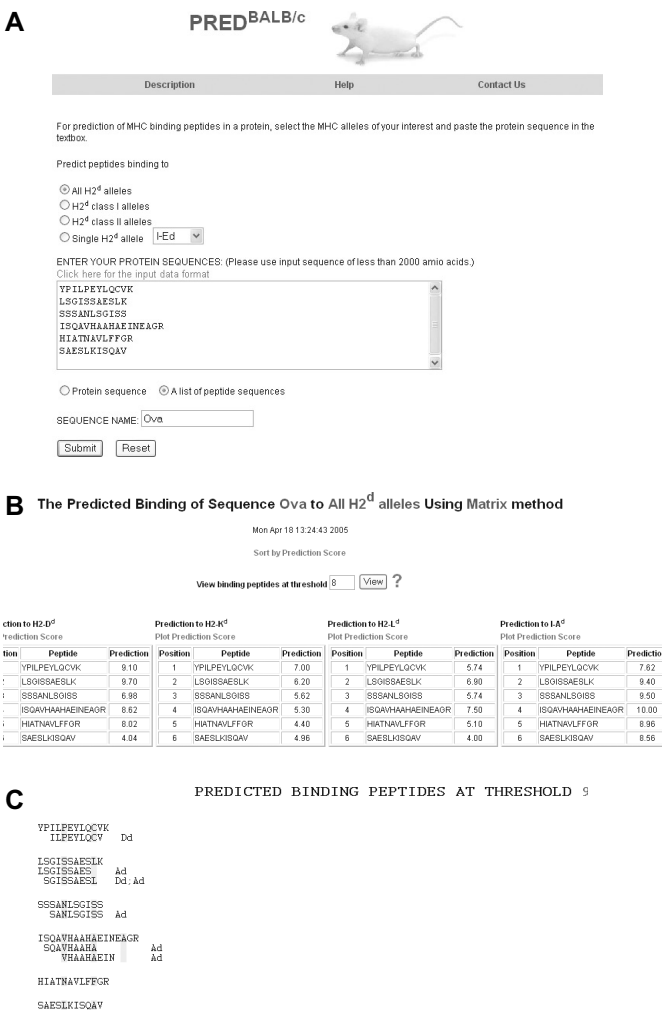

Figure 2. An example of the output pages of $PRED^{BALB/c}$ when the input is a list of peptides. The input is a list of peptides from chicken ovalbumin and the $H2^d$ alleles of interest are all $H2^d$ alleles. (**A**) The input page. (**B**) The main result page. All 9mers in each peptide are submitted for prediction. The predicted binding scores are represented by the highest individual 9mer binding score of each input peptide. (**C**) The alignment view of the predicted binding peptides at threshold 9, which indicates that the specificity of the prediction is 0.9.

When users select the input sequence type to be 'a list of peptide sequences', the input sequences separated by carriage returns or line breaks are treated as different peptides (Figure 2A). All overlapping 9mers in each peptide are submitted for prediction. In the result tables, predicted binding scores are represented by the highest individual 9mer binding score within the input peptide. The predicted binding scores of individual 9mers in each peptide in the list are not shown (Figure 2B). To display the top-scoring 9mer peptides from each input peptide, the user can use the function 'View binding peptides at threshold 9' (Figure 2B). In the result page (Figure 2C), the 9mers with binding scores equal to or above the threshold of 9 are aligned with the input peptides. The predicted 9mers are displayed with the names of the $H2^d$ alleles to which the 9mer binding scores are above the threshold. For example, the first input peptide, YPILPEYLQCVK, has binding scores 9.10, 7.00, 5.74, 7.62 and 8.62 to H2-D^d, H2-K^d, H2-L^d, I-A^d and I-E^d, respectively (Figure 2B).

At threshold 9 (SP level 0.9), there are no 9mer binders to $H2^d$ class II alleles and the 9mer ILPEYLQCV has the highest binding score to H2-D^d, 9.10. Thus, in Figure 2C, this 9mer is aligned with the input peptide and followed by 'D^d'.

CONCLUSION

$PRED^{BALB/c}$ marks a new direction in predictive modeling of MHC-binding peptides and T-cell epitopes. The main advantage is that $PRED^{BALB/c}$ focuses on a complete organism and its predictions represent a complete set of predicted targets of T-cell immune responses. The focus on the complete set of MHC alleles is closer to studies involving laboratory animals. This approach provides a more complete view of the immune responses of an organism. The BALB/c mouse is an important laboratory model and $PRED^{BALB/c}$ is, therefore, useful for the analysis of immunization regimens and deciphering responses to infections. Further development of $PRED^{BALB/c}$ will include addition of matrices for prediction of 8mer and 10mer binders to $H2^d$ class I molecules and further improvement of prediction matrices by cyclical refinement—using newly defined binders and non-binders from experiments.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

This project has been funded in part by US Federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under Grant No. 5 U19 AI56541 and Contract No. HHSN266200400085C. Funding to pay the Open Access publication charges for this article was provided by the Institute for Infocomm Research.

Conflict of interest statement. None declared.

REFERENCES

1. Pamer,E. and Cresswell,P. (1998) Mechanisms of MHC class I–restricted antigen processing. *Annu. Rev. Immunol.*, **16**, 323–358.
2. Villadangos,J.A., Bryant,R.A., Deussing,J. and Driessen,C. (1999) Proteases involved in MHC class II antigen presentation. *Immunol. Rev.*, **172**, 109–120.
3. Yewdell,J.W. and Bennink,J.R. (2001) Cut and trim: generating MHC class I peptide ligands. *Curr. Opin. Immunol.*, **13**, 13–18.
4. Bryant,P. and Ploegh,H. (2004) Class II MHC peptide loading by the professionals. *Curr. Opin. Immunol.*, **16**, 96–102.
5. 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.
6. Brusic,V. and Zeleznikow,J. (1999) Computational binding assays of antigenic peptides. *Lett. Pept. Sci.*, **6**, 313–324.
7. Brusic,V., Bajic,V.B. and Petrovsky,N. (2004) Computational methods for prediction of T-cell epitopes—a framework for modelling, testing, and applications. *Methods*, **34**, 436–443.
8. De Groot,A.S., Saint-Aubin,C., Bosma,A., Sbai,H., Rayner,J. and Martin,W. (2001) Rapid determination of HLA B*07 ligands from the West Nile virus NY99 genome. *Emerg. Infect. Dis.*, **7**, 706–713.
9. Buchwald,U.K., Lees,A., Steinitz,M. and Pirofski,L.A. (2005) A peptide mimotope of type 8 pneumococcal capsular polysaccharide induces a protective immune response in mice. *Infect. Immun.*, **73**, 325–33.
10. Sakai,Y., Morrison,B.J., Burke,J.D., Park,J.M., Terabe,M., Janik,J.E., Forni,G., Berzofsky,J.A. and Morris,J.C. (2004) Vaccination by genetically modified dendritic cells expressing a truncated neu oncogene prevents development of breast cancer in transgenic mice. *Cancer Res.*, **64**, 8022–8028.
11. Rammensee,H.G., Bachmann,J., Emmerich,N.P., Bachor,O.A. and Stevanovic,S. (1999) SYFPEITHI: database for MHC ligands and peptide motifs. *Immunogenetics*, **50**, 213–219.
12. Parker,K.C., Bednarek,M.A. and Coligan,J.E. (1994) Scheme for ranking potential HLA-A2 binding peptides based on independent binding of individual peptide side-chains. *J. Immunol.*, **152**, 163–175.
13. 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.
14. Udaka,K., Wiesmuller,K.H., Kienle,S., Jung,G., Tamamura,H., Yamagishi,H., Okumura,K., Walden,P., Suto,T. and Kawasaki,T. (2000) An automated prediction of MHC class I-binding peptides based on positional scanning with peptide libraries. *Immunogenetics*, **51**, 816–828.
15. Hattotuwagama,C.K., Guan,P., Doytchinova,I.A. and Flower,D.R. (2004) New horizons in mouse immunoinformatics: reliable in silico prediction of mouse class I histocompatibility major complex peptide binding affinity. *Org. Biomol. Chem.*, **2**, 3274–3283.
16. Yu,K., Petrovsky,N., Schönbach,C., Koh,J.Y.L. and Brusic,V. (2002) Methods for prediction of peptide binding to MHC molecules: a comparative study. *Mol. Med.*, **8**, 137–148.
17. Quesnel,A., Hsu,S.C., Delmas,A., Steward,M.W., Trudelle,Y. and Abastado,J.P. (1996) Efficient binding to the MHC class I K(d) molecule of synthetic peptides in which the anchoring position 2 does not fit the consensus motif. *FEBS Lett.*, **387**, 42–46.
18. Brusic,V., Rudy,G. and Harrison,L.C. (1994) MHCPEP, a database of MHC-binding peptides. *Nucleic Acids Res.*, **22**, 3663–3665.
19. Bhasin,M., Singh,H. and Raghava,G.P. (2003) MHCBN: a comprehensive database of MHC binding and non-binding peptides. *Bioinformatics*, **19**, 665–666.
20. McSparron,H., Blythe,M.J., Zygouri,C., Doytchinova,I.A. and Flower,D.R. (2003) JenPep: a novel computational information resource for immunobiology and vaccinology. *J. Chem. Inf. Comput. Sci.*, **43**, 1276–1287.
21. Rammensee,H.G., Falk,K. and Rotzschke,O. (1993) Peptides naturally presented by MHC class I molecules. *Annu. Rev. Immunol.*, **11**, 213–244.
22. Rammensee,H.G. (1995) Chemistry of peptides associated with MHC class I and class II molecules. *Curr. Opin. Immunol.*, **7**, 85–96.
23. Scott,C.A., Peterson,P.A., Teyton,L. and Wilson,I.A. (1998) Crystal structures of two I-A^d-peptide complexes reveal that high affinity can be achieved without large anchor residues. *Immunity*, **8**, 319–329.
24. Khan,A.M., Srinivasan,K.N., August,J.T. and Brusic,V. (2005) Neural Models for predicting viral vaccine targets. *J. Bioinform. Comput. Biol.* (in press).
25. Zhang,G.L., Khan,A.M., Srinivasan,K.N., August,J.T. and Brusic,V. (2005) MULTIPRED: a computational system for prediction of promiscuous HLA binding peptides. *Nucleic Acids Res.* (in press).
26. Sette,A., Buus,S., Appella,E., Smith,J.A., Chesnut,R., Miles,C., Colon,S.M. and Grey,H.M. (1989) Prediction of major histocompatibility complex binding regions of protein antigens by sequence pattern analysis. *Proc. Natl Acad. Sci. USA*, **86**, 3296–3300.
27. Mata,M., Travers,P.J., Liu,Q., Frankel,F.R. and Paterson,Y. (1998) The MHC class I-restricted immune response to HIV-gag in BALB/c mice selects a single epitope that does not have a predictable MHC-binding motif and binds to K^d through interactions between a glutamine at P3 and pocket D1. *J. Immunol.*, **161**, 2985–2993.
28. Saikh,K.U., Martin,J.D., Nishikawa,A.H. and Dillon,S.B. (1995) Influenza A virus-specific H-2^d restricted cross-reactive cytotoxic T lymphocyte epitope(s) detected in the hemagglutinin HA2 subunit of A/Udorn/72. *Virology*, **214**, 445–452.
29. Zatechka,D.S.,Jr, Hegde,N.R., Hariharan,K. and Srikumaran,S. (1999) Identification of murine cytotoxic T-lymphocyte epitopes of bovine herpesvirus 1. *Vaccine*, **17**, 686–69.