PEPVAC: A web server for multi-epitope vaccine development based on the prediction of supertypic MHC ligands

Pedro A. Reche¶†§ and Ellis L. Reinherz¶†

¶Laboratory of Immunobiology and Department of Medical Oncology, Dana-Farber Cancer Institute and †Department of Medicine, Harvard Medical School, 44 Binney Street, Boston, MA 02115, USA

TEL: +1-617-632-3412

FAX: +1-617-632-3351

§To whom correspondence should be addressed

Running title: Prediction of promiscuous MHC I restricted peptides

Abbreviations used: CTL, cytotoxic T cells; MHC, major histocompatibility complex; MHCI, MHC class I; MHCII, MHC class II; HLA, human leukocyte antigens; HLA I, HLA class I; HLA II, HLA class II, PSSM, position specific scoring matrix.

ABSTRACT

Prediction of peptide binding to major histocompatibility complex (MHC) molecules is a basis for anticipating T cell epitopes, as well as epitope discovery-driven vaccine development. In the human, MHC molecules are known as human leukocyte antigens (HLAs), and are extremely polymorphic. HLA polymorphism is the basis of differential peptide binding until now limiting the practical use of current epitope-prediction tools for vaccine development. Here, we describe a web server (PEPVAC) optimized for the formulation of multi-epitope vaccines with broad population coverage. This optimization is accomplished through the prediction of peptides that are to bind to several HLA molecules with similar peptide binding specificity (supertypes). Specifically, we offer the possibility of identifying promiscuous peptide binders to five distinct HLA class I supertypes (A2, A3, B7, A24, B15). We estimated the phenotypic population frequency of these supertypes to be 95%, regardless of ethnicity. Targeting these supertypes for promiscuous peptide binding predictions results in a limited number of potential epitopes without compromising the population coverage required for practical vaccine design considerations. PEPVAC can also identify conserved MHC ligands, as well as those with a C-terminus resulting from proteasomal cleavage. The combination of these features with the prediction of promiscuous HLA class I ligands, further limits the number of potential epitopes.

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

INTRODUCTION

T cells are key component of the adaptive immune system, playing a pivotal role fighting both infectious agents and cancer cells (1). T cell-based immune responses are driven by antigenic peptides (epitopes), presented in the context of major histocompatibility complex (MHC) molecules (2). Therefore, prediction of peptides that can bind to MHC molecules has become the basis for the anticipation of T cell epitopes (3). MHC molecules fall into two major classes, namely MHC class I (MHCI) and MHC class II (MHCII). Antigens presented by MHCI and MHCII are recognized by two distinct sets of T cells, $CD8^+$ T and $CD4^+$ T cells, respectively. Identification of T cell epitopes is important for both understanding disease pathogenesis and vaccine design. Thus, the availability of computational methods that can readily identify potential epitopes from primary protein sequences, have fueled a new paradigm in vaccine develpoment that is driven by this epitope-discovery.

A major complication to this vaccine development approach is the extreme polymorphism of the MHC molecules. In the human, MHC molecules are known as human leukocytes antigens (HLAs), and there are hundreds of allelic variants of the class I (HLA I) and the class II (HLA II) molecules. These HLA allelic variants bind distinct sets of peptides as MHC polymorphism is the basis for distinct peptide binding specificity (4), and are expressed at vastly variable frequencies in different ethnic groups (5). This complexity suggests that a large number of HLA molecules will have to be targeted for peptide-binding predictions, requiring so many peptides to elicit a broadly protective multi-epitope vaccine as to be impractical. Interestingly, groups of several HLA molecules (supertypes) can bind largely overlapping sets of peptides (6,7). Identification of these HLA supertypes facilitates epitope-based vaccine development for the following reasons: first, targeting of representative HLA aleles from distinct supertypes allow the immune response to be stimulated in a variety of genetic backgrounds; second, selection of promiscuous peptide-binders to those alleles included within a given supertype limits the number of peptides to be considered without decreasing the spectrum of the immune response.

In this paper we describe a web server, PEPVAC, that allows the prediction of promiscuous epitopes to 5 HLA I supertypes: A2 (A*0201-07, A*0209, A*6802), A3 (A*0301, A*1101, A*3101, A*3301, A*6801, A*6601), A24 (A*2402, B*3801), B7 (B*0702, B*3501, B*5101-02, B*5301, B*5401), B15 (A*0101, B*1501_B62, B1502). These supertypes were defined using a method based on the clustering of the predicted peptide-binding repertoire of MHC molecules (8). The combined phenotypic frequency of these supertypes is greater than 95% for five major American ethnicities (Black, Caucasian, Hispanic, Native American, and Asian). Thus, targeting these supertypes with epitope predictions would potentially provide a population coverage ≥ 95%, regardless of ethnicity.

Peptides binding to HLA I molecules are potential $CD8^+$ T cell epitopes. *In vivo*, the C-terminus of these antigenic epitopes result from the selective proteolysis of cytosolic proteins mediated by the proteasome (9). The proteasome is thus important for determining these epitopes. Therefore, PEPVAC has also been implemented with an algorithm for the identification of those peptides containing a C-terminus that is likely to be the result of proteasomal cleavage. Finally, PEPVAC also allows the prediction of conserved epitopes from sequences with variability masked. The combination of these two features serve to both, refining the predictions of T cell epitopes and limiting the number of potential epitopes.

PREDICTION OF PEPTIDE-MHCI BINDING.

The peptide binding mode of MHC I molecules differs from that of MHC II (10-12), and as result the prediction of peptide-MHCII binding is less reliable than that of peptide-MHCI binding. Thereby, we have focused here in the prediction of MHCI ligands, a class that is specifically recognized by $CD8^+$ cytotoxic T lymphocytes (CTLs). Peptides binding to a specific MHCI molecule are related by sequence similarity, and thus we use position specific scoring matrix (PSSMs) from aligned MHCI ligands as the predictors of peptide-MHCI binding in combination with a dynamic algorithm. PSSMs are also known as profiles and weight-matrices and have previously been shown to be adequate tools for prediction of peptide-MHC binding (13-16). PSSMs are derived from block alignment of MHCI ligands which are of the same length. Such a restriction guarantees proper structural alignment of ligands, and subsequent accuracy of the peptide binding predictions (13,14). Given that MHCI-ligands are usually of 9 residues in length, PSSMs used in this study are for the prediction of ligands of that same size (9 residues). Accuracy of prediction of peptide-MHCI binding using PSSMs varies depending on threshold and the targeted MHCI molecule. On average, however, ROC analyses of the predictions at different thresholds result in *AUC* values (Area Under Roc curve) above 0.8, indicating that these PSSMs are very for good predictors of peptide-MHCI binding. Furthermore, >80% of known $CD8^+$ T cell epitopes can be predicted at a 2% threshold from their protein sources

SUPERTYPES: IDENTIFICATION AND POPULATION COVERAGE ANALYSIS

We defined HLA I supertypes through clustering of predicted MHC peptide binding repertoires (8). In brief, the core of the method consists of the generation of a distance matrix whose coefficients are inversely proportional to the peptide binders shared by any two HLA molecules (Figure 1). Subsequently, this distance matrix is fed to a phylogenic clustering algorithm to establish the kinship among the distinct HLA peptide binding repertoires. Figure 2 shows a phylogenic tree built upon the peptide binding repertoire of 55 HLA I molecules, using a Fitch and Margoliash clustering algorithm (17). We defined supertypes (Figure 2) as groups of HLA I alleles with ≥ 20% peptide binding overlap (pairwise between any pair of alleles). The supertypes identified in this study include the A2, A3, B7, B27 and B44 supertypes previously identified by Sidney-Sette et al. Furthermore, we have also identified two new supertypes, BX and B57 (Figure 2). The cumulative phenotypic frequency (CPF) of these supertypes is shown in Table I. CPF was calculated using the gene and haplotype frequencies reported for 5 distinct American ethic groups including Blacks, Caucasians, Hispanic, North America Natives and Asians (18). CPF represents the population coverage that would be provided by a vaccine composed of epitopes restricted by the alleles included in the supertype. The A2, A3 and B7 supertypes have the largest CPF in the 5 studied ethnic groups, close to 90%, irrespective of ethnicity. To increase the population coverage to ≥ 95%, regardless of ethnicity, it is necessary to include at least two more supertypes. Specifically, the supertypes A2, A3, B7, B15 and A24/B44 represent the minimal supertypic combinations with the indicated population coverage. Alleles belonging to each of these supertypes are shown in Figure 2 and Table I.

PEPVAC WEB SERVER

Following the HLA I supertypic analysis as discussed, we have implemented a tool for the prediction of promiscuous peptide binders to a set of supertypes with a CPF greater than 95%, irrespective of ethnicity. We named this tool PEPVAC (Promiscuous Epitopes based VACcines), and it is online at the site http://immunax.dfci.harvard.edu/PEPVAC/ hosted by the Molecular Immunology Foundation/Dana-Farber Cancer Institute. The web interface to PEPVAC is divided in several sections that facilitate intuitive use (Fig. 3A). Main features of the web server are discussed bellow.

Input and limitations

In PEPVAC, input query to carry epitope predictions is entered in the GENOME section (Fig 3A). Input consists of a single or various protein sequences in FASTA format. Only the standard 20 amino acid residues are considered. There are several translated genomes from pathogenic organisms that can be selected as inputs. More useful, a user-provided local file containing a set of protein sequences can be uploaded to the server using the choose/browse bottom. PEPVAC can also process files with protein sequences in which the variable sites have been masked with a dot "." symbol. In that case, peptide-binding predictions will be carried out only over consecutive stretches of 9 or more residues. Sequences with variable positions masked according to the Shannon Entropy variability metric (4,19) can be obtained at the site http://immunax.dfci.harvard.edu/bioinformatics/Tools/sva.html. Currently, there is limit of 200 sequences and 50,000 symbols that can be processed per request. If such limits are exceed the server will return an error.

Supertypes and thresholds

The A2, A3, B7, B15 and A24 (Figure 2 and Table I) supertypes have been chosen for promiscuous peptide binding predictions in PEPVAC. Only those peptides that are predicted to bind to all the alleles included in the supertypes are returned in the output (Figure 3B). Threshold for the prediction of promiscuous peptide binders in PEPVAC have been fixed to provide a reduced and manageable set of promiscuous peptide binders to each supertype. As an example, predicted promiscuous peptides to the above 5 supertypes from a genome such as that of *Influenza virus A* (4160 amino acids distributed in 10 distinct open reading frames (ORF), represents only 5.51% (254 9mer peptides) of all possible peptides (4617 9mer peptides).

Proteasome cleavage.

In PEPVAC, predictions of supertypic peptide binders are combined with prediction of proteasomal cleavage using probabilistic language models derived from HLA I-restricted epitopes (14). Currently, there are three optional models for proteasomal cleavage that differ in their sensitivity/specificity ratio of the predictions as discussed elsewhere (14). These models are selected within the PROTEASOME CLEAVAGE section. Model 1 has the highest sensitivity (~95%) and the lower specificity (~60 %). Conversely, Model 3 has the lowest sensitivity (65%) with the largest specificity (80%). Model two has a sensitivity and specificity around 70%. Promiscuous peptide binders containing a C-terminal end predicted to be the result of proteasomal cleavage are shown in violet in the result page (Figure 3B). In the previous example with the *Influenza* virus A, the list of promiscuous peptide binders to the 5 selected supertypes decreases from 254 down to 170 peptides (3.7% of all 9mer peptides from *Influenza virus A* genome) after considering proteasomal cleavage using Model 1. Furthermore, a combination of the predictions of peptide-MHCI binding and proteasomal cleavage increases the specificity of the epitope predictions by discarding predicted peptide-MHCI binders that experimentally are unable to elicit $CD8^+$ T cell responses (20).

Output

The results page returned by PEPVAC is shown in Figure 3B. This page first displays a summary of the predictions, including the chosen selections, the number of predicted peptides and the minimum population coverage provided by the supertypic selection, followed by the predicted peptide binders to each of the selected supertypes (only A3 in the shown example). Peptides are predicted to bind to all alleles included in the supertype, and appear ranked with regard to the PSSMs of the first allele included in the supertype. Relevant information about each sorted peptide includes its protein source as well as its molecular weight.

ACKNOWLEDGMENTS

This manuscript was supported by NIH grant AI50900 and the Molecular Immunology Foundation. We wish to acknowledge John-Paul Glutting for programming assistance.

REFERENCES

1. Paul, W.E. (1998) *Fundamental Immunology*.
2. Zinkernagel, R.M. and Doherty, P.C. (1974) Restriction of in vitro T cell-mediated cytotoxicity in lymphocytic choriomeningitis within a syngeneic or semiallogeneic system. *Nature*, **248,** 701-702.
3. Flower, D.R. and Doytchinova, I.A. Immunoinformatics and the prediction of immunogenicity. 167-176.
4. 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.
5. David W. Gjertson and Paul I. Terasaki, E. (1998) *HLA 1998.*
6. Sette, A. and Sidney, J. (1999) Nine major HLA class I supertypes account for the vast preponderance of HLA-A and -B polymorphism. *Immunogenetics*, **50,** 201-212.
7. Sette, A. and Sidney, J. (1998) HLA supertypes and supermotifs: a functional perspective on HLA polymorphism. *Curr Opin Immunol*, **10,** 478-482.
8. Reche, P.A. and Reinherz, E.L. (2004). Definition of MHC supertypes through clustering of MHC peptide binding repertoires. *Artificial Immune Systems.* Springer-Verlag Berling Heidelberg., Catania, Italy, Vol. LNCS 3239, pp. 189-196.
9. 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.
10. Madden, D.R. (1995) The three-dimensional structure of peptide-MHC complexes. *Annu. Rev. Immunol.*, **13,** 587-622.
11. 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.
12. Stern, L.J. and Wiley, D.C. (1994) Antigen peptide binding by class I and class II histocompatibility proteins. *Structure*, **2,** 245-251.
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. 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.
15. 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. Epub 2004 Feb 1312.
16. Peters, B., Tong, W., Sidney, J., Sette, A. and Weng, Z. (2003) Examining the independent binding assumption for binding of peptide epitopes to MHC-I molecules. *Bioinformatics*, **19,** 1765-1772.
17. Fitch, W.M. and Margoliash, E. (1967) Construction of phylogenetic trees. *Science*, **155,** 279-284.
18. Cao, K., Hollenbach, J., Shi, X., Shi, W., Chopek, M. and Fernandez-Vina, M.A. (2001) Analysis of the frequencies of HLA-A, B, and C alleles and haplotypes in the five major ethnic groups of the United States reveals high levels of diversity in these loci and contrasting distribution patterns in these populations. *Hum Immunol*, **62,** 1009-1030.
19. 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.

20. 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.
21. Dawson, D.V., Ozgur, M., Sari, K., Ghanayem, M. and Kostyu, D.D. (2001) Ramifications of HLA class I polymorphism and population genetics for vaccine development. *Genet Epidemiol*, **20,** 87-106.

FIGURES LEGENDS

Figure 1. Strategy to define HLA I supertypes. HLA I supertypes are identified by clustering their peptide binding repertoire (8). The method consists of 4 basic steps. (1) Predict the peptide binding repertoire (*i,j* sets in figure) of each HLA I molecule from the same random protein using the relevant PSSMs in combination with the RANKPEP scoring algorithm (13). (2) Compute the number of common peptides between the binding repertoire of any two HLA I molecules. (3) Build a distance matrix whose coefficients are inversely proportional to the peptide binding overlap between any pair of HLA I molecules. (4) Use a phylogenic clustering algorithm to compute and visualize HLA I supertypes (clusters of HLA I molecules with overlapping peptide binding repertoires).

Figure 2. HLA I peptide binding overlap and supertypes. The figure shows an unroot dendrogram built after clustering the overlap between the peptide binding repertoire of the indicated HLA I molecules. Peptide binding repertoires of HLA I molecules were obtained from a random protein (1000 amino acids) using the relevant PSSMs at a 2% peptide-binding threshold. This dendrogram reflects the relationship between the peptide-binding specificities of HLA I molecules. HLA I alleles with similar peptide binding specificities branch together in groups or clusters. The closer HLA I alleles branch, the larger is the overlap between their peptide-binding repertoires. Supertypes (shadowed with different colors) consist of groups HLA I alleles with at least a 20% peptide binding overlap (pairwise between any pair of alleles).

Figure. 3 The PEPVAC web server. A) PEPVAC input page. The page is divided into several sections. E-MAIL, for obtaining the results via e-mail (optional). GENOMES, where a selection of genomes from pathogenic organisms are available, as well as the possibility of uploading a user-provided genome. SUPERTYPES, the supertypes A2, A3, B7, A24, and B15 are available for selection. Alleles targeted for peptide-biding predictions in each supertype are indicated. The minimum population coverage of the selected supertypes is calculated on the fly and shown on the relevant window. PROTEASOMAL

CLEAVAGE, prediction of proteasomal cleavage using three optimal language models are carried out in parallel to the peptide biding predictions. *B*) PEPVAC result page. An example result page where the A3 supertype was selected for peptide-binding predictions from the genome of *Influenza A virus* is shown. The result page first displays a summary of the predictions, followed by the predicted peptide binders to each of the selected 18 774 c342

TABLES

Table 1. *Cumulative phenotype frequency of defined supertypes*

Supertype	Alleles	Blacks	Caucasians	Hispanics	*N.A.Natives	Asians
A2	A*0201-7, A*6802	43.7%	49.9%	51.8%	52.4%	44.7%
A3	A*0301, A*1101, A*3101, A*3301, A*6801, A*6601	35.4%	46.9%	41.5%	40.7%	47.9%
B7	B*0702, B*3501, B*5101-02, B*5301, B*5401	45.9%	42.2%	40.5%	52.0%	31.3%
B15	A*0101, B*1501_B62, B1502	13.06%	37.80%	16.75%	27.26%	21.04%
A24	A*2402, B*3801	15.5%	17.28%	25.85%	41.94%	35.0%
B44	B*4402, B*4403	10.4%	27.7%	17.15%	14.4%	10.1%
B57	B5701-02, B5801, B*1503	19.2%	10.3%	5.9%	5.8%	16.5%
ABX	A*2902, B*4002	7.4%	11.3%	19.1%	16.3%	16.3%
B27	B*2701-06, B*2709, B*3909	2.3%	4.8%	5.1%	16.9%	4.7%
BX	B*1509, B*1510, B*39011	3.1%	0.7%	4.2%	7.8%	4.1%

Cumulative phenotype frequency was obtained using the HLA I gene and haplotype frequencies published by Cao et al (18) corresponding to the indicated 5 American ethnic groups. Method for computing the cumulative phenotype frequency considered the disequilibrium linkage between the HLA-A and -B gene and was based on that reported by Dawson et al (21) *North American Natives.

FIGURES

Figure 1

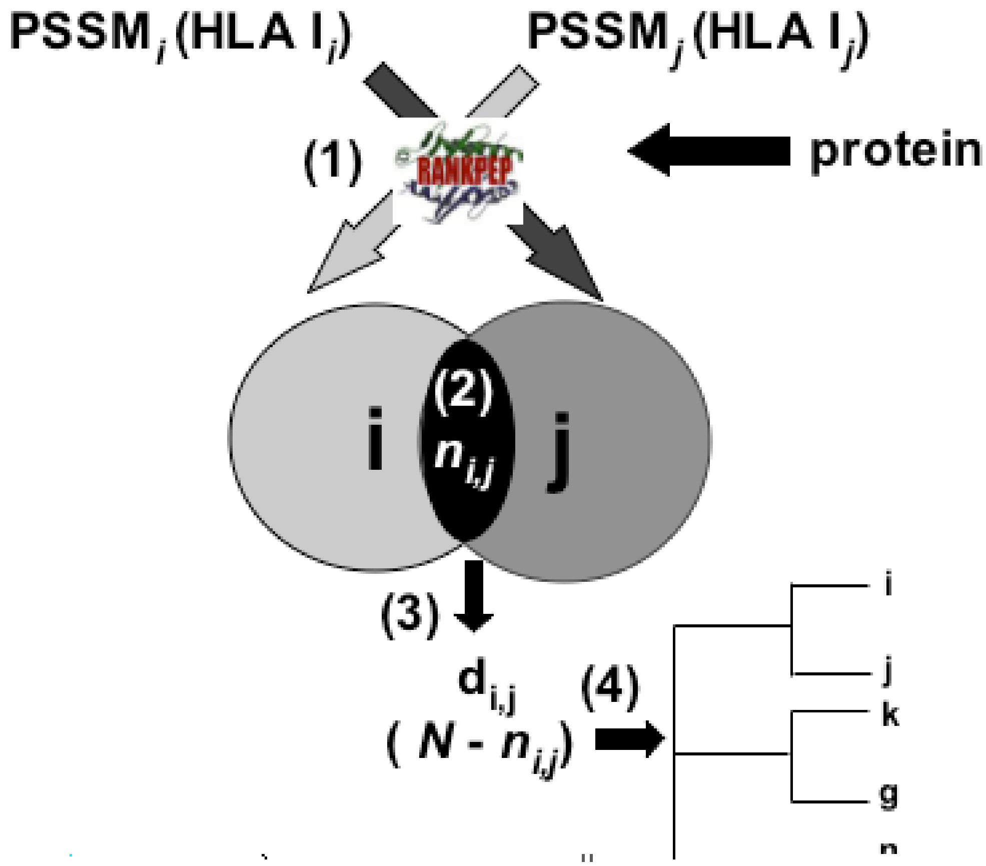

Figure 2

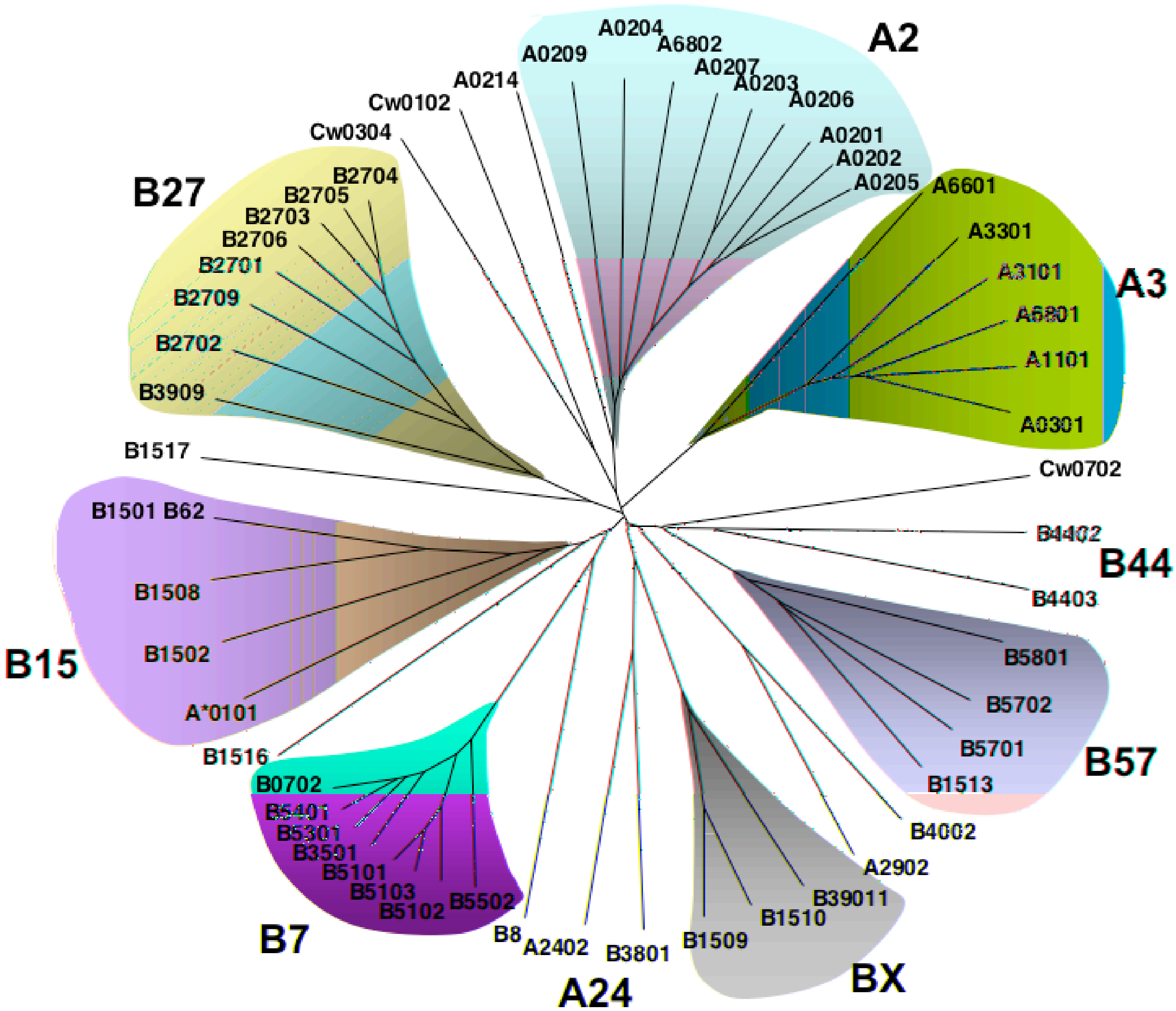
A2
A0204
A6802
A0209
A0207
A0203
A0206
A0201
A0202
A0205
A0214
Cw0102
Cw0304
B27
B2704
B2705
B2703
B2706
B2701
B2709
B2702
B3909
A3
A6601
A3301
A3101
A6801
A1101
A0301
B1517
Cw0702
B44
B4402
B4403
B15
B1501 B62
B1508
B1502
A*0101
B1516
B57
B5801
B5702
B5701
B1513
B4002
A2902
B7
B0702
B5401
B5301
B3501
B5101
B5103
B5102
B5502
B8
A24
A2402
B3801
BX
B1509
B1510
B39011

Figure 3