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

D.R. Flower et al. (eds.), *Bioinformatics for Immunomics*, Immunomics Reviews 3,
DOI 10.1007/978-1-4419-0540-6_8, © Springer Science+Business Media, LLC 2010

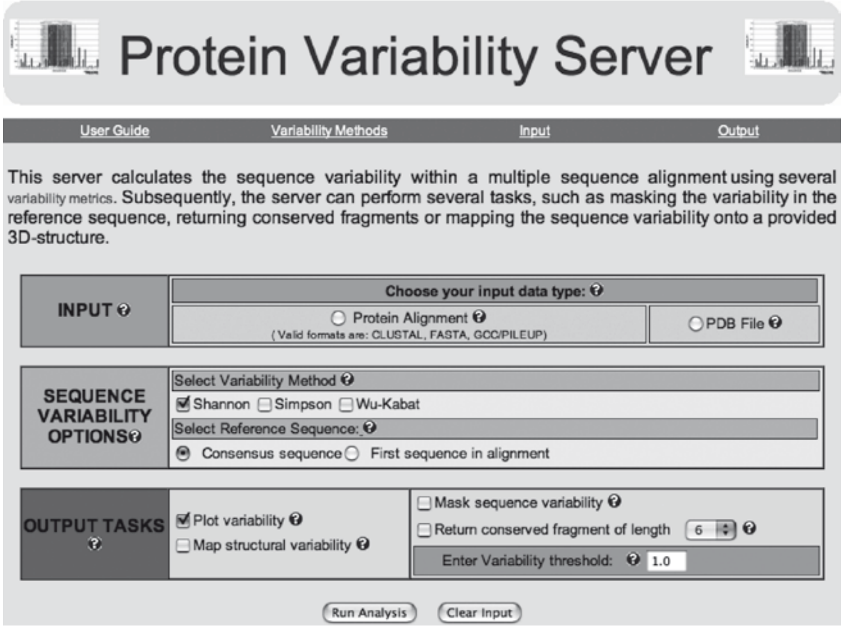
Materials and Methods

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-



This server calculates the sequence variability within a multiple sequence alignment using several variability metrics. Subsequently, the server can perform several tasks, such as masking the variability in the reference sequence, returning conserved fragments or mapping the sequence variability onto a provided 3D-structure.

INPUT
Choose your input data type:
○ Protein Alignment (Valid formats are: CLUSTAL, FASTA, GCC/PILEUP)
○ PDB File

SEQUENCE VARIABILITY OPTIONS
Select Variability Method
☑ Shannon □ Simpson □ Wu-Kabat
Select Reference Sequence:
● Consensus sequence ○ First sequence in alignment

OUTPUT TASKS
☑ Plot variability
□ Map structural variability
□ Mask sequence variability
□ Return conserved fragment of length 6
Enter Variability threshold: 1.0

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

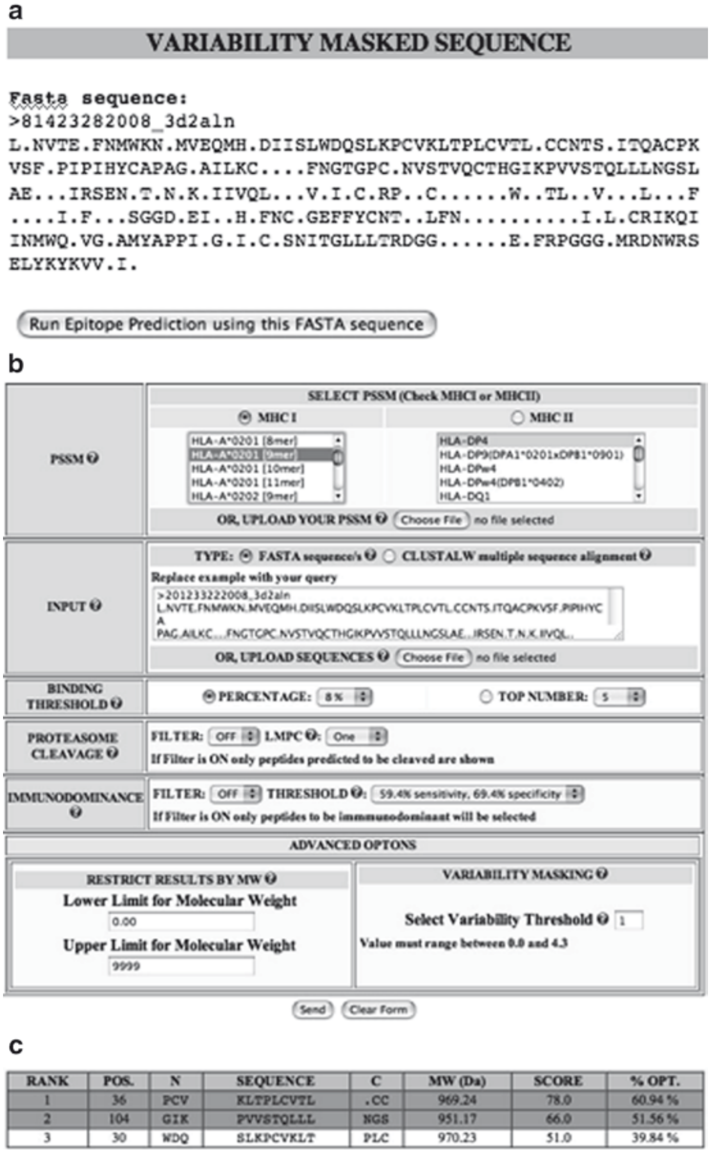
function studies and vaccine design. PVS main input is an MSA provided by the user, but it can also take a PDB file as main input, generating an MSA from it (for details see García-Boronat et al. 2008) The sequence variability in the MSA is computed *per site* using three different metrics: The Shannon 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,



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.
```

PSSM	SELECT PSSM (Check MHCI or MHCII) ● MHC I: HLA-A*0201 [8mer], **HLA-A*0201 [9mer]**, HLA-A*0201 [10mer], HLA-A*0201 [11mer], HLA-A*0202 [9mer] ○ MHC II: HLA-DP4, HLA-DP9(DPA1*0201xDPB1*0901), HLA-DPw4, HLA-DPw4(DPB1*0402), HLA-DQ1 OR, UPLOAD YOUR PSSM: no file selected
INPUT	TYPE: ● FASTA sequence/s ○ CLUSTALW multiple sequence alignment Replace example with your query >201233222008_3d2aln L.NVTE.FNMWKN.MVEQMH.DIISLWDQSLKPCVKLTPLCVTL.CCNTS.ITQACPKVSF.PIPIHYC A PAG.AILKC....FNGTGPC.NVSTVQCTHGIKPVVSTQLLLNGSLAE...IRSEN.T.N.K.IIVQL.. OR, UPLOAD SEQUENCES: no file selected
BINDING THRESHOLD	● PERCENTAGE: 8 % ○ TOP NUMBER: 5
PROTEASOME CLEAVAGE	FILTER: OFF LMPC: One If Filter is ON only peptides predicted to be cleaved are shown
IMMUNODOMINANCE	FILTER: OFF THRESHOLD: 59.4% sensitivity, 69.4% specificity If Filter is ON only peptides to be immmunodominant will be selected

ADVANCED OPTONS

RESTRICT RESULTS BY MW	VARIABILITY MASKING
Lower Limit for Molecular Weight: 0.00 Upper Limit for Molecular Weight: 9999	Select Variability Threshold: 1 Value must range between 0.0 and 4.3

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

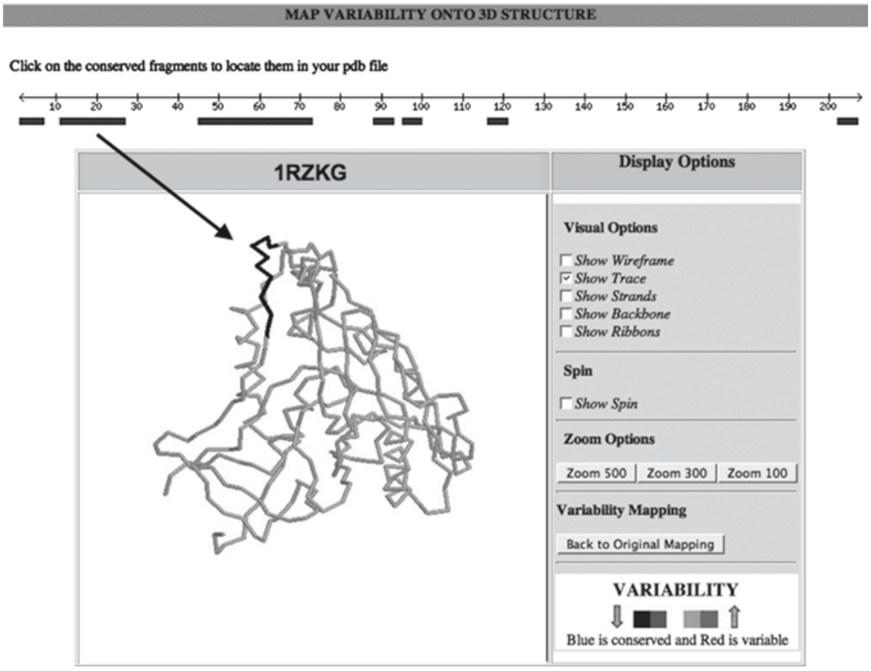
N	Start	End	Sequence
1	1	7	S T M G A A S
2	9	25	T L T V Q A R Q L L S G I V Q Q Q
3	27	55	N L L R A I E A Q Q H L L Q L T V W G I K Q L Q A R V L A
4	62	67	D Q Q L L G
5	69	74	W G C S G K
6	87	92	S W S N K S
7	153	158	W L W Y I K

Fragments were selected to have six or more consecutive residues with H≤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 ≤ 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



MAP VARIABILITY ONTO 3D STRUCTURE

Click on the conserved fragments to locate them in your pdb file

1RZKG

Display Options

Visual Options

□ Show Wireframe
☑ Show Trace
□ Show Strands
□ Show Backbone
□ Show Ribbons

Spin

□ Show Spin

Zoom Options

Variability Mapping

VARIABILITY

Blue is conserved and Red is variable

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