Computational Analysis of Cysteine Proteases (Clan CA, Family C1) of *Leishmania major* to Find Potential Epitopic Regions

Babak Saffari and Hassan Mohabatkar*

Department of Biology, College of Sciences, Shiraz University, Shiraz, Iran.

*Corresponding author. E-mail: mohabatkar@susc.ac.ir

Leishmania is associated with a broad spectrum of diseases, ranging from simple cutaneous to invasive visceral leishmaniasis. Here, the sequences of ten cysteine proteases of types A, B and C of *Leishmania major* were obtained from GeneDB database. Prediction of MHC class I epitopes of these cysteine proteases was performed by NetCTL program version 1.2. In addition, by using BcePred server, different structural properties of the proteins were predicted to find out their potential B cell epitopes. According to this computational analysis, nine regions were predicted as B cell epitopes. The results provide useful information for designing peptide-based vaccines.

Key words: bioinformatics, cysteine protease, T cell epitope, B cell epitope, *Leishmania major*

Introduction

Leishmania (Order: Kinetoplastida, Family: Trypanasomatidae) is an obligate intracellular parasite responsible for a broad spectrum of diseases, ranging from simple cutaneous to invasive visceral leishmaniasis (*1*). Protozoan parasites of the genus *Leishmania* present two forms in their life cycle: promastigote, which multiplies in the mid gut of the sand fly vector, and amastigote, the obligate intracellular form that lives within phagolysosomes of the vertebrate host (*2*, *3*). Three major types of leishmaniasis, namely cutaneous, mucocutaneous and visceral, occur in humans depending on the species of *Leishmania*. Infection by species such as *L. major*, *L. tropica* and *L. mexicana* may cause localized cutaneous lesions, resulting in lifelong immunity. Infection by *L. braziliensis* and *L. panamensis* initially presents as cutaneous lesions that may then spread or metastasize causing mucocutaneous lesions. Infection by *L. donovani*, *L. infantum* and *L. chagasi* may result in a chronic disseminating visceral disease in the liver and spleen (*4*).

It has been recognized for many years that proteases of pathogenic organisms may modulate the host's defense mechanisms (*5*). Proteases are grouped into clans and families on the basis of the architecture of their catalytic dyad or triad (*6*). Each clan is identified with two letters, the first representing the catalytic type of the families included in the clan. *L. major* has cysteine proteases (CPs) of eight families within clan CA. Family C1 contains CPA and CPB,

which are both cathepsin L-like in terms of primary amino acid sequence, and CPC, which is cathepsin B-like. CPB is unusual in that it has a 100-amino acid C-terminal extension in comparison with most CPs of the group, and exists as multiple isoenzymes, which are encoded by a tandem array of similar CPB genes located in a single locus (the arrays comprise eight genes in *L. major*). The CPBs of *L. mexicana* are stage-regulated and the isoforms present differences in their substrate specificity and catalytic properties (*7*).

Although the exact roles of CPs in *Leishmania* pathogenesis are unclear, it has been demonstrated that *Leishmania* cannot grow within macrophages in the presence of CP inhibitors (*8*). These observations provide evidence of the importance of these molecules in the survival of both promastigote and amastigote forms of these parasites (*9*). Despite trypanosomatid CPs may be instrumental in modulating the host's immune response to favor parasite survival and proliferation, they are themselves immunogenic. *L. mexicana* CP is a T cell immunogen, resulting in the development of potentially protective Thl cell lines (*10*). This finding suggests that the CP itself is a vaccine candidate and that homologous enzymes in other species may also be so. A CP of *L. pifanoi*, however, provided rather little protection for the host against infection with the parasite (*11*), although more recently a similar *L. amazonensis*

CP provided some protection against subsequent challenge, apparently through inducing a Th1-associated response (*12*). These differing results presumably reflect the complexity of the immune response to the active parasite enzymes and how the response may be determined by the precise immunization conditions. It is therefore encouraging that a CP-rich fraction of *L. major* was shown to be a strong inducer of a primed human immune response and may have protective function (*13*). These observations suggest that trypanosomatid CPs have potential as vaccines although attempts to exploit them are really just beginning (*14*, *15*). In addition, it has also been proven that infected dendritic cells are the critical antigen-presenting cells responsible for T cell priming in *Leishmania* infections. Amastigotes, but not the infectious promastigotes, are the main targets for phagocytosis activity of dendritic cells (*16*, *17*).

A key step in the design of subunit vaccines is the identification of epitopes from overlapping synthetic peptides. This method decreases the possibility of missed epitopes, but lots of peptides need to be synthesized at a high cost. New developments in immunoinformatics and other computational methodologies, combined with the broad versatility in the design and synthesis of genetic (DNA) vaccines, underlay new strategies for the novel design of antigen-specific, epitope-based vaccines against many pathogens that currently have proven refractive to conventional vaccine therapy (*18*). Epitopes are selected by prediction with software, which saves the expense of synthetic peptides and working time (*19*, *20*).

Basically, the recognition of antigenic epitopes by the immune system, either small discrete T cell epitopes or large conformational epitopes recognized by B cells and soluble antibodies, is the key molecular event at the heart of the immune response to pathogens (*21*). The objective of this bioinformatics-based study is to enhance the optimal selection of epitopic regions of clan CA, C1 family of cysteine proteases as potential targets of immune response. Consensus sequence methodology was used to identify sequences of 9 amino acids or longer with complete conservation in 80% or more of C1 families of cysteine proteases. These conserved sequences were further analyzed to identify targets for candidate epitope-based T cell vaccine formulations against *L. major*. Furthermore, concerning the activation of human humoral immunity by leishmaniasis, B cell epitopes were also predicted based on propensity scales for each of

the 20 amino acids. Utilizing bioinformatics servers for vaccine candidates is a time-saving approach that could significantly help to increase our information about various aspects of pathogens in molecular biology. These theoretical predictions can then be tested by using experimental and complementary methods.

Results

Prediction of MHC class I binding peptides

For the prediction of major histocompatibility complex (MHC) class I binding peptides of C1 family of cysteine proteases, one sequence for each of the cysteine proteases A, B and C was selected. In the case of CPB, due to the high similarity between these cysteine proteases, only one out of eight sequences was chosen (LmjF08.1050). This sequence is a good candidate of CPB because of the complete identity of its sequence with the consensus sequence of CPB. All overlapping nonamer peptides were generated from this dataset and were screened for potential T cell antigens using the NetCTL algorithm, from which 716 peptides were short-listed. Most of the peptides were found to exhibit mono-supertype specificity, meaning that they bind to a single supertype. Some of them, however, appeared to bind to multiple supertypes; the highest number of supertypes a given nonamer could bind is 5 out of the 12 supertypes tested. In fact, out of the 716 human leukocyte antigen (HLA)-binding nonamers, one peptide binds to 5 supertypes, 4 bind to 4 supertypes, 9 bind to 3 supertypes and 44 bind to 2 supertypes. Sequences of binding peptides to the 4 and 5 supertypes and the name of proteins they belong to are summarized in **Table 1**.

Knowing the number of binding peptides of each of the analyzed proteins is important, considering the polymorphic nature of HLA and its diversity in populations of different geographical regions. Therefore,

Table 1 Number of supertypes that a given nonamer could bind

Peptide sequence	Protein	No. of binding to the supertypes
339-HVSQSPTPG-347	CPB	5
9-FAIVVTILF-17	CPA	4
270-FMSYHSGVL-278	CPB	4
42-AEVNSKAKG-50	CPC	4
258-LTMQVYSDF-266	CPC	4

a good T cell antigen should have peptides recognized by many HLA alleles. The analysis revealed that CPB has the maximum number of binding peptides, followed by CPC and CPA, respectively (**Table 2**).

For short-listing potential vaccine candidates, it is important to analyze the binding profiles from a supertype perspective. Of the 12 supertypes studied, the largest number of nonamers was found to be recognized by the allele B62 (53), followed by B58 (38), A2 (36), A24 (24), A1, B8 and B39 (21), B44 (20), A3 and B7 (19), B27 (17) and A26 (15), as illustrated in **Figure 1**.

The score with which a peptide binds to HLA ranges from 0.75 to 3.1949. In general, the binding score of CPC peptides to A1 locus supertypes is higher compared with CPB, CPA and other supertypes (**Table 3**).

Putative promiscuous T cell epitopes may be localized in clusters, as reported in studies of HIV-1 (*22–25*), the outer membrane of *Chlamydia trachomatis*

Table 2 Generated and binding peptides from selected cysteine proteases

Systematic name in GeneDB	Protein	Features	No. of amino acid	No. of generated peptides	No. of binding peptides	Percent of binding peptides	No. of binding peptides with score >1.5
LmjF19.1420	CPA	cathepsin L-like	354	346	91	26.3%	8
LmjF08.1050	CPB	cathepsin L-like	348	340	103	30.3%	9
LmjF29.0820	CPC	cathepsin B-like	340	332	96	28.9%	15

Table 3 Features of binding peptides with high scores

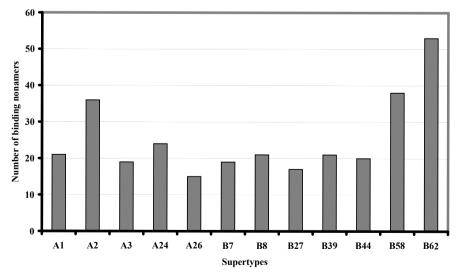| No.	Protein	Supertype	Peptide	Score
1	CPC	A1	52-WTASADNGY-60	3.1949
2	CPC	A1	308-NTDWGDKGY-316	2.6687
3	CPC	A1	19-LATTVSGLY-27	2.3211
4	CPA	A1	77-NTHNPHAHY-85	2.0979
5	CPC	A1	131-AVEAISDRY-139	1.9722
6	CPC	A1	76-VTDMSTEAV-84	1.9262
7	CPC	A24	204-KYPPCPSTI-212	1.9244
8	CPC	B44	279-GEFLGGHAV-287	1.9117
9	CPC	A26	52-WTASADNGY-60	1.8531
10	CPC	A1	224-RSEMDLVKY-232	1.8397

Figure 1 Observed variation in the binding of predicted peptides to the 12 different HLA class I supertypes studied. The figure depicts the proportion of binding peptides to individual supertypes.

(*26*), and among others (*27, 28*). The clusters are also ideal for developing epitope-based vaccines because they contain multiple promiscuous epitopes. The number of immunogenic hotspots for CPA, CPB and CPC is 4, 5 and 0, respectively, as shown in **Table 4**.

The identification of conserved sequences is very important to design peptide vaccines, because vaccines that are developed on the basis of the conserved segments among candidate proteins can be used against a large majority of pathogen's variants. In **Figure 2**, three segments (I, II and III) with identity ≥90% and have ≥9 amino acids in length are shown as conserved regions. Obviously the epitopes predicted in these regions are very significant. For immunological applications, a minimum conserved sequence length of 9 amino acids is required because this represents the typical length of peptides that bind to HLA molecules (*29*). The features of potential epitopes located in conserved regions with maximum scores are summarized in **Table 5**.

B cell epitope prediction

Before the prediction of B cell epitope of CPB (LmjF08.1050), signal peptide of this protein predicted by SignalP 3.0 hidden Markov model (HMM) (signal peptide probability 0.999, signal anchor probability 0.001, with cleavage site probability 0.760 between residues 27 and 28) was excluded. Hydrophilicity, flexibility, accessibility, turns, exposed surface, polarity and antigenic propensity scales were applied to predict B cell epitopes. These parameters were correlated with the location of continuous epitopes. As a result, 9 regions were predicted to be B cell epitopes (Figure 2).

Discussion

The aim of this investigation was to apply bioinformatics methods to study the B and T cell epitopic sites of C1 family cysteine proteases of *L. major*. To help the development of vaccines, understanding the structural basis for the cell-mediated immune response is necessary (*30*). The perfect bioinformatics prediction of T cell epitopes can to a great extent reduce the experimental cost in candidate epitope identification (*31*).

In the present study, NetCTL program has been used to predict MHC class I of cysteine proteases A, B and C of *L. major* (*32*). CP proteins are immunogenic and are potential vaccine candidates. Efficient processing and presentation of vaccine antigens by class I and/or class II MHC are essential for a good T cell response. Since humans carry only a limited number of co-dominant HLA alleles in their genome (2 each for A, B and C loci), out of hundreds of polymorphic alleles present in the population, it becomes important that a candidate vaccine must generate peptides that bind to a wide range of HLA molecules to provide good population coverage.

Table 4 Features of immunogenic hotspots of CPA and CPB

CP	Supertype	Position	Score	Threshold	Length
CPA	A2	1–39	57.38	55	39
	A2	174–224	56.20	55	51
	A3	61–93	45.46	45	33
	A3	230–279	45.21	45	50
CPB	A2	1–38	55.65	55	38
	A2	188–221	55.04	55	34
	A3	18–138	46.51	45	121
	A3	151–193	45.19	45	43
	A3	228–270	45.42	45	43

Table 5 Epitopes with highest scores located in conserved regions

Conserved region	Epitope sequence	Supertype	Score	CP
I	133-REKGAVTPV-141	B44	1.2763	CPB
II	148-GMCGSCWAF-156	B62	1.2696	CPA
III	283-GEQLNHGVL-291	B44	1.7068	CPB

```
                ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                     10        20        30        40        50        60        70        80        90       100
CPA(LmjF19.1420) MARRNPFLFA IVVTILFVVC YGSALVAQTP LGVDNFIASA HYGRFKERHG KSFGEDADEG HRFNAFKQNM QTAYFLNTHN PHAHYDVSGK FADLTPQEFA

CPC(LmjF29.0820) MALRAKSALC LVAVFALLLA TTVSGLYAKP SDFP-LLGKS FVAEVNSKAK GQWTASADNG --YLVTGKSL GEVRKLMGVT DMSTEAVP-- ---PRNFSVE
CPB(LmjF08.1080) MATSRAALCA VAV-VCVVLA AACAPARAIY VGTP---AAA LFEEFKRTYQ RAYGTLTEEQ QRLANFERNL ELMREHQARN PHARFGIT-K FFDLSEAEFA
CPB(LmjF08.1060) MATSRAALCA VAV-VCVVLA AACAPARAIY VGTP---AAA LFEEFKRTYQ RAYGTLTEEQ QRLANFERNL ELMREHQARN PHARFGIT-K FFDLSEAEFA
CPB(LmjF08.1030) MATSRAALCA VAV-VCVVLA AACAPARAIY VGTP---AAA LFEEFKRTYQ RAYGTLTEEQ QRLANFERNL ELMREHQARN PHARFGIT-K FFDLSEAEFA
CPB(LmjF08.1020) MATSRAALCA VAV-VCVVLA AACAPARAIY VGTP---AAA LFEEFKRTYQ RAYGTLTEEQ QRLANFERNL ELMREHQARN PHARFGIT-K FFDLSEAEFA
CPB(LmjF08.1010) MATSRAALCA VAV-VCVVLA AACAPARAIY VGTP---AAA LFEEFKRTYQ RAYGTLTEEQ QRLANFERNL ELMREHQARN PHARFGIT-K FFDLSEAEFA
CPB(LmjF08.1070) MATSRAALCA VAV-VCVVLA AACAPARAIY VGTP---AAA LFEEFKRTYQ RAYGTLTEEQ QRLANFERNL ELMREHQARN PHARFGIT-K FFDLSEAEFA
CPB(LmjF08.1040) MATSRAALCA VAV-VCVVLA AACAPARAIY VGTP---AAA LFEEFKRTYQ RAYGTLTEEQ QRLANFERNL ELMREHQARN PHARFGIT-K FFDLSEAEFA
CPB(LmjF08.1050) MATSRAALCA VAV-VCVVLA AACAPARAIY VGTP---AAA LFEEFKRTYQ RAYGTLTEEQ QRLANFERNL ELMREHQARN PHARFGIT-K FFDLSEAEFA

Clustal Consensus **    . :.. . ::.  :          .   . :   ..:      :  :::       :.:        .  :  :.         ..
Consensus        MA
```

```
                ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                     110       120       130       140       150       160       170       180       190       200
CPA(LmjF19.1420) KLYLN-PDYY AHRGKDYKEH VHVDDSVLSG A-MSVDWREK GAVTPVKNQG MCGSCWAFSA IGNIESQWAL KN-HSLVSLS EQMLVSCDDI DD-GCNGGLM

CPC(LmjF29.0820) ELQQDLPEFF DAAEH----- ---------- ----WPMC LTISEIRDQS NCGSCWAIAA VEAISDRYCT FGGVPDRRMS TSNLLSCCFI CGLGCHGGIP

CPB(LmjF08.1080) ARYLNGAAYF AAAKQHAGQH YRKARADLSA VPDAVDWREK GAVTPVKNQG ACGSCWAFSA VGNIESQWAV AG-HKLVRLS EQQLVSCDHV DN-GCGGGLM
CPB(LmjF08.1060) ARYLNGAAYF AAAKQHAGQH YRKARADLSA VPDAVDWREK GAVTPVKNQG ACGSCWAFSA VGNIESQWAV AG-HKLVRLS EQQLVSCDHV DN-GCGGGLM
CPB(LmjF08.1030) ARYLNGAAYF AAVKQHAGQH YRKARADLSA VPDAVDWREK GAVTPVKNQG ACGSCWAFSA VGNIESQWAV AG-HKLVRLS EQQLVSCDHV DN-GCGGGLM
CPB(LmjF08.1020) ARYLNGAAYF AAAKQHAGQH YRKARADLSA VPDAVDWREK GAVTPVKNQG ACGSCWAFSA VGNIESQWAV AG-HKLVRLS EQQLVSCDHV DN-GCGGGLM
CPB(LmjF08.1010) ARYLNGAAYF AAAKQHAGQH YRKARADLSA VPDAVDWREK GAVTPVKNQG ACGSCWAFSA VGNIESQWAV AG-HKLVRLS EQQLVSCDHV DN-GCGGGLM
CPB(LmjF08.1070) ARYLNGAAYF AAAKQHAGQH YRKARADLSA VPDAVDWREK GAVTPVKNQG ACGSCWAFSA VGNIESQWAV AG-HKLVRLS EQQLVSCDHV DN-GCGGGLM
CPB(LmjF08.1040) ARYLNGAAYF AAAKQHAGQH YRKARADLSA VPDAVDWREK GAVTPVKNQG ACGSCWAFSA VGNIESQWAV AG-HKLVRLS EQQLVSCDHV DN-GCGGGLM
CPB(LmjF08.1050) ARYLNGAAYF AAAKQHAGQH YRKARADLSA VPDAVDWREK GAVTPVKNQG ACGSCWAFSA VGNIESQWAV AG-HKLVRLS EQQLVSCDHV DN-GCGGGLM

Clustal Consensus : . ::    :                *    ::: :::*. ******::* : *...::.   .    :*   . *:**  :  . ** **:
Consensus            W            Q   CGSCWA  A          I                S     L SC       GC GG
                                           I                  II
```

```
                ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                     210       220       230       240       250       260       270       280       290       300
CPA(LmjF19.1420) DQAMEWIIQH ---------- ----HNGTVP TEKSYPYASA GGTSPPCHDK GEFG--ARIS GYMSLPH--- DEKAIAAYVE KKGPVAVAVD ATT-WQLYFG

CPC(LmjF29.0820) TVAWLWWVWV GIATEDCQPY PFDPCSHHGN SEKYPPCPST IYDTPKCNTT CERSEMDLVK YKGSTSYSVK GEKELMIELM TNGPLELTMQ VYSDFVGYKS

CPB(LmjF08.1080) LQAFEWVLRN ---------- ----MNGTVF TEKSYPYPVSG NGDVPECSNS SELAPGARID GYVSMES--- SERVMAAWLA KNGPISIAVD ASS-FMSYHS
CPB(LmjF08.1060) LQAFEWVLRN ---------- ----MNGTVF TEKSYPYPVSG NGDVPECSNS SELAPGARID GYVSMES--- SERVMAAWLA KNGPISIAVD ASS-FMSYHS
CPB(LmjF08.1030) LQAFEWVLRN ---------- ----MNGTVF TEKSYPYPVSG NGDVPECSNS SELAPGARID GYVSMES--- SERVMAAWLA KNGPISIAVD ASS-FMSYHS
CPB(LmjF08.1020) LQAFEWVLRN ---------- ----MNGTVS TEKSYPYPVSG NGDVPECSNS SELAPGARID GYVSMES--- SERVMTAWLA KNGPISIAVD ASS-FMSYHS
CPB(LmjF08.1010) LQAFEWVLRN ---------- ----MNGTVF TEKSYPYTST FGYVPECSNS SELAPGARID GYVSMES--- SERVMAAWLA KNGPISIAVD ASS-FMSYHS
CPB(LmjF08.1070) LQAFEWVLRN ---------- ----MNGTVF TEKSYPYPVSG NGDVPECSNS SELAPGARID GYVSMES--- SERVMTAWLA KNGPISIAVD ASS-FMSYHS
CPB(LmjF08.1040) LQAFEWVLRN ---------- ----MNGTVF TEKSYPYPVSG NGDVPECSNS SELAPGARID GYVSMES--- SERVMAAWLA KNGPISIAVD ASS-FMSYHS
CPB(LmjF08.1050) LQAFEWVLRN ---------- ----MNGTVF TEKSYPYPVSG NGDVPECSNS SELAPGARID GYVSMES--- SERVMAAWLA KNGPISIAVD ASS-FMSYHS

Clustal Consensus *  * :            .       :** * *      ** . *.   :.       *       .*:.: :   : ..:**: :::: . : : * .
Consensus        A  W                  EK  P  S        P C    E             S          E              GP          Y
```

```
                ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                     310       320       330       340       350       360       370       380       390       400
CPA(LmjF19.1420) GVVTLCFGLS LN-HGVLVVG FNKRAKPPYW IVKNSWGTSW GEKGYIRLAM GSNQCLLKNY PVTATVDDSN TS-------- ---------- -----HVPTT

CPC(LmjF29.0820) GVYKHVLGEF LGGHAVKLVG WGTQDGVPYW KVANSWNTDW GDKGYFLIQR GNNECKIESG GVAGIPAQE- ---------- ---------- ----------

CPB(LmjF08.1080) GVLTSCIGEQ LN-HGVLLVG YNMTGEVPYW VIKNSWGEDW GEKGYVRVTM GVNACLLTGY PVSVHVSQSP TPGPNTTTTT HAPKRVTVKQ ITCTDYFCRK
CPB(LmjF08.1060) GVLTSCIGEQ LN-HGVLLVG YNMTGEVPYW VIKNSWGEDW GEKGYVRVTM GVNACLLTGY PVSVHVSQSP TPGPNTTTTT HAPKRVTVKQ ITCTDYFCRK
CPB(LmjF08.1030) GVLTSCIGEQ LN-HGVLLVG YNMTGEVPYW VIKNSWGEDW GEKGYVRVTM GVNACLLTGY PVSVHVSQSP TPGPNTTTTT HAPKRVTVKQ ITCTDYFCRK
CPB(LmjF08.1020) GVLTSCIGEQ LN-HGVLLVG YNMTGEVPYW VIKNSWGEDW GEKGYVRVTM GVNACLLTGY PVSVHVSQSP TP-------- ---------- -----YL---
CPB(LmjF08.1010) GVLTSCIGEQ LN-HGVLLVG YNMTGEVPYW VIKNSWGKDW GEKGYVRVTM GVNACLLTGY PVSVHVSQSP TP-------- ---------- -----YL---
CPB(LmjF08.1070) GVLTSCIGEQ LN-HGVLLVG YNMTGEVPYW VIKNSWGEDW GEKGYVRVTM GVNACLLTGY PVSVHVSQSP TP-------- ---------- -----YL---
CPB(LmjF08.1040) GVLTSCIGEQ LN-HGVLLVG YNMTGEVPYW VIKNSWGKDW GEKGYVRVTM GVNACLLTGY PVSVHVSQSP TP-------- ---------- -----YL---
CPB(LmjF08.1050) GVLTSCIGEQ LN-HGVLLVG YNMTGEVPYW VIKNSWGEDW GEKGYVRVTM GVNACLLTGY PVSVHVSQSP TP-------- ---------- -----YL---

Clustal Consensus **.  :*   *. *.* :** :.      *** : ***. .* *:***. :    * * * :.   *:   :.
Consensus        GV     G   L  H V  VG        PYW      NSW    W G KGY    G N C    V
                           III
```

```
                ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                     410       420       430       440       450       460
CPA(LmjF19.1420) TA-------- ---------- ---------- ---------- ---------- ---------
CPC(LmjF29.0820) ---------- ---------- ---------- ---------- ---------- ---------
CPB(LmjF08.1080) GCKTTVIPTK ECLPNGAGGS FQMECGDHQV LKLTYTSMNC TGEAKYTVTR EGKCGISWSG SSKSICQYV
CPB(LmjF08.1060) GCKTTVIPTK ECLPNGAGGS FQMECGDHQV LKLTYTSMNC TGEAKYTVTR EGKCGISWSG SSKSICQYV
CPB(LmjF08.1030) GCKTTVIPTK ECLPNGAGGS FQMECGDHQV LKLTYTSMNC TGEAKYTVTR EGKCGISWSG SSKSICQYV
CPB(LmjF08.1020) ---------- ---------- ---------- ---------- ---------- ---------
CPB(LmjF08.1010) ---------- ---------- ---------- ---------- ---------- ---------
CPB(LmjF08.1070) ---------- ---------- ---------- ---------- ---------- ---------
CPB(LmjF08.1040) ---------- ---------- ---------- ---------- ---------- ---------
CPB(LmjF08.1050) ---------- ---------- ---------- ---------- ---------- ---------
Clustal Consensus
Consensus
```
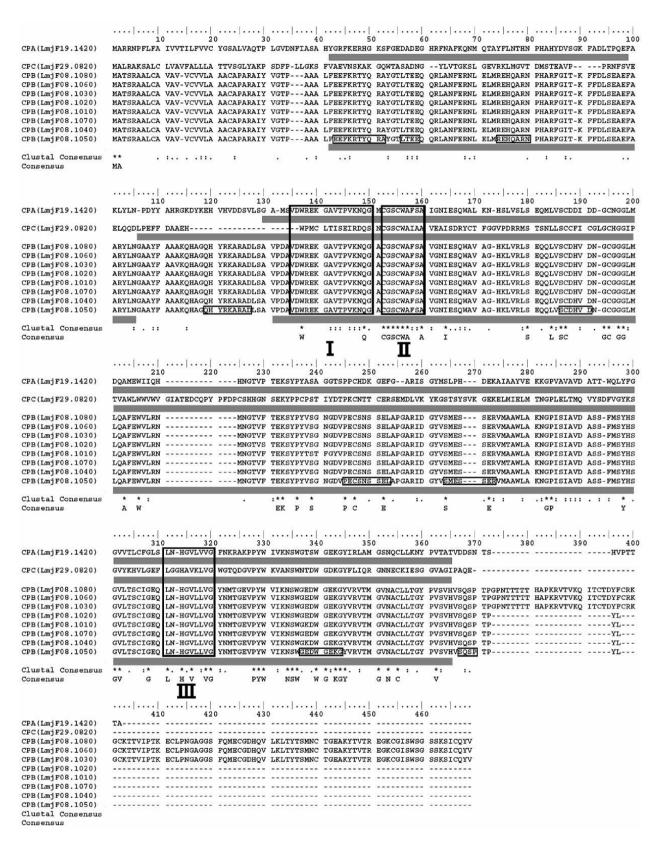
Figure 2 Alignment of C1 family of clan CA of cysteine proteases. Gray rectangular under each sequence is the corresponding regions of that sequence, which is determined with Pfam database as conserved sequences. Sequences shown in boxes are predicted B cell epitopes of CPB (LmjF08.1050). Regions that have identity ≥90% are shown with Roman numerals. The most conserved residues that appear in the consensus sequence are indicated by "*".

In this work we found that generated peptides bind in larger numbers to B supertypes. However, almost all of the peptides with the highest binding score belong to CPC. In other words, CPC is the major source of peptides that bind to HLA loci with more affinity. These observations suggest that greater emphasis must be placed on cytotoxic T lymphocyte (CTL) response generated by the presentation of antigen by B alleles and should design epitope-based vaccine directed towards these HLA.

T cell epitopes specific to multiple HLA supertypes are advantageous for vaccine design because they effectively increase the number of epitopes to which an individual can respond, and provide much more extensive coverage of the population (*33*). The peptides binding to more than one HLA are termed promiscuous and such peptides are of prime interest for vaccine design because of their relevance in coverage of higher proportions of human populations. *In silico* approaches would help to predict some of the HLA-binding motifs, which could act as promiscuous epitopes (*34*). Most of the generated nonameric peptides in this work are mono-allelic binders. To cover majority of the population, it is essential to have vaccine candidates that have multi-binding behavior. Consequently, peptides with the binding ability of ≥4 supertypes were taken as promiscuous epitopes. Note that each supertype consists of multiple HLA alleles, and peptides that can bind to ≥4 supertypes have a great potentiality to activate the most proportion of T cell population.

It is generally recognized that conserved protein sequences represent important functional domains (*35*), for which mutations would be detrimental to the survival of the pathogen. The functions of conserved sequences can be elucidated by databases that comprise data on protein families, domains and functional sites, such as the Pfam database (www.sanger.ac.uk/Software/Pfam) (*36*). In Figure 2, in addition to the ClustalW consensus sequence, the results of the Pfam database and the highly conserved regions that have ≥9 amino acids in length have also been shown. It is clear that the predicted epitopes located in the conserved segments have more validity.

Eventually, identification of proteins with peptides binding to larger number of alleles, assessment of alleles or supertypes of MHC that bind large number of peptides than others have great importance in determining epitopes as a candidate vaccine. In addition, allelic variation in binding affinity, immunologi-

cal hotspots, HLA distribution analysis and similarity of epitopes to the self proteins play a key role in identification of these epitopes (*34–36*).

In proteins, turns are located on the surface; these parts are accessible and hydrophilic but the core regions are mostly devoid of water molecules (*37*). Antigenic determinants lie in regions that are hydrophilic, exposed and polar, while accessibility and flexibility of these segments are high. This has led to the rules that would allow the position of B cell epitopes to be predicted from these features of the protein sequence (*37, 38*).

In conclusion, recognizing epitopes on proteins is essential for developing synthetic vaccines and can facilitate immunotherapy of leishmaniasis and many other infectious diseases. In the present work, employing a bioinformatics approach, a set of peptides has been identified, which can be used in either a natural or a synthetic vaccine cocktail. This approach could be extended to the entire proteome of *L. major* to identify newer sets of potentially antigenic proteins and yet reducing the number of T and B cell antigens for experimental verification. These kinds of researches can be applied for omitting non-functional sequences of proteins, which would help in designing new immunological methods.

Materials and Methods

Amino acid sequences

The sequences of ten cysteine proteases (clan CA, family C1) of *L. major* were obtained from GeneDB database (www.genedb.org) (*39*). Sequences included one CPA (Systematic ID: LmjF19.1420), one CPC (LmjF29.0820) and eight CPBs (LmjF08.1010, LmjF08.1020, LmjF08.1030, LmjF08.1040, LmjF08.1050, LmjF08.1060, LmjF08.1070 and LmjF08.1080).

Prediction of MHC class I binding peptides

NetCTL program version 1.2 (http://www.cbs.dtu.dk/services/NetCTL/) (*40*) predicts peptides restricted to 12 HLA class I supertypes (A1, A2, A3, A24, A26, B7, B8, B27, B39, B44, B58 and B62), integrated with predictions of HLA binding, proteasomal C-terminal cleavage and transport efficiency by the transporter associated with antigen processing (TAP) molecules. HLA binding and proteasomal cleavage

predictions were performed by an artificial neural networks (ANN) method and TAP transport efficiency was predicted using a weight matrix method. The parameters used for NetCTL prediction were: 0.15 weight on C terminal cleavage, 0.05 weight on TAP transport efficiency, and 0.75 threshold for HLA supertype binding. The final scores are the predicted MHC class I affinities in form of −logIC50 and IC50 values.

B cell epitope prediction

All prediction calculations were based on propensity scales for each of the 20 amino acids. Sequence of each protein was read as a moving window. In order to compare the profiles obtained by different methods, various scales were normalized where the original values of each scale were set between +3 and −3. Hydrophilicity (*41*), flexibility (*42*), accessibility (*43*), turns (*44*), exposed surface (*45*), polarity (*46*) and antigenic propensity (*47*) scales were applied to predict B cell epitopes by BcePred server (http://www.imtech.res.in/raghava/bcepred) with default threshold.

Signal peptide prediction

Due to the elimination of signal peptides of CPBs before secretion to the outer space of the infected cells, this region must be excluded from the entire sequence of the protein for exerting the prediction analysis on it. Signal peptide prediction was achieved using SignalP 3.0 HMM (http://www.cbs.dtu.dk/services/SignalP-3.0/) (*48*).

Immunogenic hotspot prediction

Putative promiscuous T cell epitopes may be localized in clusters that are also ideal for developing epitope-based vaccines because they contain multiple promiscuous epitopes. For determining the immunogenic hotspots, MULTIPRED server (http://antigen.i2r.a-star.edu.sg/multipred/) (*49*) was utilized.

Alignment

Sequences were aligned using ClustalW program (*50*) from the BioEdit v5.0.9 package (*51*).

Authors' contributions

BS conceived the study and carried out the computational analysis. HM supervised the study. BS and HM prepared the manuscript. Both authors read and approved the final manuscript.

Competing interests

The authors have declared that no competing interests exist.

References

1. Bates, P.A. 1994. The developmental biology of *Leishmania* promastigotes. *Exp. Parasitol.* 79: 215-218.
2. Chang, K.P., *et al.* 1990. Molecular determinates of *Leishmania* virulance. *Annu. Rev. Microbiol.* 44: 499-529.
3. Berman, J.D. 1997. Human leishmaniasis: clinical, diagnostic, and chemotherapeutic developments in the last 10 years. *Clin. Infect. Dis.* 24: 684-703.
4. Desjeux, P. 1996. Leishmaniasis. Public health aspects and control. *Clin. Dermatol.* 14: 417-423.
5. Brooks, D.R., *et al.* 2001. The stage-regulated expression of *Leishmania mexicana* CPB cysteine proteases is mediated by an intercistronic sequence element. *J. Biol. Chem.* 276: 47061-47069.
6. Barrett, A.J. and Rawlings, N.D. 2001. Evolutionary lines of cysteine peptidases. *Biol. Chem.* 382: 727-733.
7. Travis, J., *et al.* 1995. Are bacterial proteases pathogenic factors? *Trends Microbiol.* 3: 405-407.
8. Mottram, J.C., *et al.* 1996. Evidence from disruption of the lmcpb gene array of *Leishmania mexicana* that cysteine proteases are virulence factors. *Proc. Natl. Acad. Sci. USA* 93: 6008-6013.
9. Alexander, J., *et al.* 1998. *Leishmania mexicana* cysteine protease-deficient mutants have attenuated virulence for mice and potentiate a Th1 response. *J. Immunol.* 161: 6794-6801.
10. Wolfram, M., *et al.* 1995. Antigen presentation by *Leishmania mexicana*-infected macrophages: activation of helper T cells specific for amastigote cysteine proteases requires intracellular killing of the parasites. *Eur. J. Immunol.* 25: 1094-1100.

11. Soong, L., *et al.* 1995. *Leishmania pifanoi* amastigote antigens protect mice against cutaneous leishmaniasis. *Infect. Immun.* 63: 3559-3566.

12. Beyrodt, C.G., *et al.* 1997. Characterization of an antigen from *Leishmania amazonensis* amastigotes able to elicit protective responses in a murine model. *Infect. Immun.* 65: 2052-2059.

13. Rafati, S., *et al.* 1997. Biochemical analysis and immunogenicity of *Leishmania major* amastigote fractions in cutaneous leishmaniasis. *Clin. Exp. Immunol.* 110: 203-211.

14. Handman, E. 1997. *Leishmania* vaccines: old and new. *Parasitol. Today* 13: 236-238.

15. McMahon-Pratt, D., *et al.* 1998. *Leishmania* amastigote target antigens: the challenge of a stealthy intracellular parasite. *Parasitol. Today* 14: 31-34.

16. von Stebut, E., *et al.* 1998. Uptake of *Leishmania major* amastigotes results in activation and interleukin 12 release from murine skin-derived dendritic cells: implications for the initiation of anti-*Leishmania* immunity. *J. Exp. Med.* 188: 1547-1552.

17. von Stebut, E., *et al.* 2000. *Leishmania major*-infected murine Langerhans cell-like dendritic cells from susceptible mice release IL-12 after infection and vaccinate against experimental cutaneous leishmaniasis. *Eur. J. Immunol.* 30: 3498-3506.

18. Khan, A.M., *et al.* 2006. A systematic bioinformatics approach for selection of epitope-based vaccine targets. *Cell. Immunol.* 244: 141-147.

19. Bian, H., *et al.* 2003. The use of bioinformatics for identifying class II-restricted T-cell epitopes. *Methods* 29: 299-309.

20. Li, G.F., *et al.* 2005. Identification of immunodominant Th1-type T cell epitopes from *Schistosoma japonicum* 28 kDa glutathione-S-transferase, a vaccine candidate. *Acta Biochim. Biophys. Sin.* 37: 751-758.

21. Doytchinova, I.A. and Flower, D.R. 2002. Quantitative approaches to computational vaccinology. *Immunol. Cell Biol.* 80: 270-279.

22. Shankar, P., *et al.* 1996. Three regions of HIV-1 gp160 contain clusters of immunodominant CTL epitopes. *Immunol. Lett.* 52: 23-30.

23. Surman, S., *et al.* 2001. Localization of CD4+ T cell epitope hotspots to exposed strands of HIV envelope glycoprotein suggests structural influences on antigen processing. *Proc. Natl. Acad. Sci. USA* 98: 4587-4592.

24. Brown, S.A., *et al.* 2003. Clustering of Th cell epitopes on exposed regions of HIV envelope despite defects in antibody activity. *J. Immunol.* 171: 4140-4148.

25. Berzofsky, J.A., *et al.* 1991. Construction of peptides encompassing multideterminant clusters of human immunodeficiency virus envelope to induce *in vitro* T cell responses in mice and humans of multiple MHC types.

J. Clin. Invest. 88: 876-884.

26. Kim, S.K. and DeMars, R. 2001. Epitope clusters in the major outer membrane protein of *Chlamydia trachomatis. Curr. Opin. Immunol.* 13: 429-436.

27. Gupta, V., *et al.* 2006. SARS coronavirus nucleocapsid immunodominant T-cell epitope cluster is common to both exogenous recombinant and endogenous DNA-encoded immunogens. *Virology* 347: 127-139.

28. Srinivasan, K.N., *et al.* 2004. Prediction of class I T-cell epitopes: evidence of presence of immunological hot spots inside antigens. *Bioinformatics* 20: I297-302.

29. Rammensee, H.G. 1995. Chemistry of peptides associated with MHC class I and class II molecules. *Curr. Opin. Immunol.* 7: 85-96.

30. Collins, F.M. 1974. Vaccines and cell-mediated immunity. *Bacteriol. Rev.* 38: 371-402.

31. Mohabatkar, H. 2007. Prediction of epitopes and structural properties of Iranian HPV-16 E6 by bioinformatics methods. *Asian Pac. J. Cancer Prev.* 8: 602-606.

32. Chen, C.C., *et al.* 2006. (PS)^2: protein structure prediction server. *Nucleic Acid Res.* 34: W152-157.

33. Wilson, C.C., *et al.* 2003. Development of a DNA vaccine designed to induce cytotoxic T lymphocyte responses to multiple conserved epitopes in HIV-1. *J. Immunol.* 171: 5611-5623.

34. Vani, J., *et al.* 2006. A combined immuno-informatics and structure-based modeling approach for prediction of T cell epitopes of secretory proteins of *Mycobacterium tuberculosis. Microbes Infect.* 8: 738-746.

35. Valdar, W.S. 2002. Scoring residue conservation. *Proteins* 48: 227-241.

36. Bateman, A., *et al.* 2004. The Pfam protein families database. *Nucleic Acids Res.* 32: D138-141.

37. Pellequer, J.L., *et al.* 1991. Predicting location of continuous epitopes in proteins from their primary structures. *Methods Enzymol.* 203: 176-201.

38. Mohabatkar, H. and Kar, S.K. 2004. Prediction of exposed domains of envelope glycoprotein in Indian HIV-1 isolates and experimental confirmation of their immunogenicity in humans. *Braz. J. Med. Biol. Res.* 37: 675-681.

39. Hertz-Fowler, C., *et al.* 2004. GeneDB: a resource for prokaryotic and eukaryotic organisms. *Nucleic Acids Res.* 32: D339-343.

40. Larsen, M.V., *et al.* 2005. An integrative approach to CTL epitope prediction: a combined algorithm integrating MHC class I binding, TAP transport efficiency, and proteasomal cleavage predictions. *Eur. J. Immunol.* 35: 2295-2303.

41. Parker, J.M., *et al.* 1986. New hydrophilicity scale derived from high-performance liquid chromatography peptide retention data: correlation of predicted surface residues with antigenicity and X-ray-derived ac-

cessible sites. *Biochemistry* 25: 5425-5432.

42. Karplus, P.A. and Schulz, G.E. 1985. Prediction of chain flexibility in proteins: a tool for the selection of peptide antigens. *Naturwissenschaften* 72: 212-213.

43. Emini, E.A., *et al.* 1985. Induction of hepatitis A virus-neutralizing antibody by a virus-specific synthetic peptide. *J. Virol.* 55: 836-839.

44. Pellequer, J.L., *et al.* 1993. Correlation between the location of antigenic sites and the prediction of turns in proteins. *Immunol. Lett.* 36: 83-99.

45. Kolaskar, A.S. and Tongaonkar, P.C. 1990. A semi-empirical method for prediction of antigenic determinants on protein antigens. *FEBS Lett.* 276: 172-174.

46. Janin, J. and Wodak, S. 1978. Conformation of amino acid side-chains in proteins. *J. Mol. Biol.* 125: 357-386.

47. Ponnuswamy, P.K., *et al.* 1980. Hydrophobic packing and spatial arrangement of amino acid residues in globular proteins. *Biochim. Biophys. Acta* 623: 301-316.

48. Bendtsen, J.D., *et al.* 2004. Improved prediction of signal peptides: SignalP 3.0. *J. Mol. Biol.* 340: 783-795.

49. Zhang, G.L., *et al.* 2005. MULTIPRED: a computational system for prediction of promiscuous HLA binding peptides. *Nucleic Acids Res.* 33: W172-179.

50. Thompson, J.D., *et al.* 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22: 4673-4680.

51. Hall, T.A. 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symp. Ser.* 41: 95-98.