PROTEINS: Structure, Function, and Genetics 52:44–55 (2003)

Evolutionary Analysis Reveals Collective Properties and Specificity in the C-Type Lectin and Lectin-Like Domain Superfamily

Sharon Ebner,[1] Nathan Sharon,[2] and Nir Ben-Tal[1]*
[1]*Department of Biochemistry, The George S. Wise Faculty of Life Sciences, Tel Aviv University, Ramat Aviv, 69978, Israel*
[2]*Department of Biological Chemistry, The Weizmann Institute of Science, Rehovot 76100, Israel*

ABSTRACT **Members of the C-type lectin/C-type lectin-like domain (CTL/CTLD) superfamily share a common fold and are involved in a variety of functions, such as generalized defense mechanisms against foreign agents, discrimination between healthy and pathogen-infected cells, and endocytosis and blood coagulation. In this work we used ConSurf, a computer program recently developed in our lab, to perform an evolutionary analysis of this superfamily in order to further identify characteristics of all or part of its members. Given a set of homologous proteins in the form of multiple sequence alignment (MSA) and an inferred phylogenetic tree, ConSurf calculates the conservation score in every alignment position, taking into account the relationships between the sequences and the physicochemical similarity between the amino acids. The scores are then color-coded onto the three-dimensional structure of one of the homologous proteins. We provide here and at http://ashtoret.tau.ac.il/~sharon a detailed analysis of the conservation pattern obtained for the entire superfamily and for two subgroups of proteins: (a) 21 CTLs and (b) 11 heterodimeric CTLD toxins. We show that, in general, proteins of the superfamily have one face that is constructed mostly of conserved residues and another that is not, and we suggest that the former face is involved in binding to other proteins or domains. In the CTLs examined we detected a region of highly conserved residues, corresponding to the known calcium- and carbohydrate-binding site of the family, which is not conserved throughout the entire superfamily, and in the CTLD toxins we found a patch of highly conserved residues, corresponding to the known dimerization region of these proteins. Our analysis also detected patches of conserved residues with yet unknown function(s). Proteins 2003;52:44–55.** © 2003 Wiley-Liss, Inc.

Key words: molecular recognition; protein-protein interactions; protein modeling; phylogenetic trees

INTRODUCTION

The C-type lectins (CTLs) constitute a large and highly diverse protein family, the common feature of which is the presence of a carbohydrate recognition domain (CRD) that requires Ca^{2+} for sugar binding.[1] Included in this family are the asialoglycoprotein receptors and the mannose binding receptor, located on the membranes of hepatocytes and macrophages, respectively,[2] all of which serve as mediators of endocytosis of glycoproteins. Other members of the family are globular proteins, such as mannose binding proteins (MBPs), which play a role in innate immunity against microbial pathogens.[3] A related family that is characterized by C-type lectin-like domains (CTLDs) consists of proteins that do not bind carbohydrates or calcium.[4] Members of this family also have diverse functions, primarily in the immune system. Prominent examples are the receptors located on the membrane of natural killer cells (e.g., CD94, LY49, and NKR-P1), the CTLDs of which serve in the discrimination between healthy and pathogen-infected cells.[5] The family also includes a variety of heterodimeric snake venoms that bind blood coagulation factors and receptors and either inhibit or induce coagulation.[6–8]

A computer program, ConSurf, for analyzing evolutionary characteristics of protein families has recently been developed in our laboratory.[9] Given a multiple sequence alignment (MSA) of homologous proteins and a phylogenetic tree inferred from it, ConSurf calculates the conservation score in every alignment position, taking into account the physicochemical similarity between the amino acids. The higher the similarity between the amino acids of different proteins, residing in an alignment position, the higher the conservation score for that position. Using a color code, the program then maps the conservation scores onto the molecular surface of an arbitrarily chosen protein in the family, the 3D structure of which is known, thus facilitating evaluation of the conservation pattern. The

Abbreviations: CRD, carbohydrate recognition domain; CTL, C-type lectin; CTLD, C-type-lectin-like domain; hMBP-C, human mannose-binding protein C; IX/Xbp, factor IX/X binding protein; MSA, multiple sequence alignment (see also Tables I and II).

*Correspondence to: Nir Ben-Tal, Department of Biochemistry, The George S. Wise Faculty of Life Sciences, Tel Aviv University, Ramat Aviv, 69978, Israel. E-mail: bental@ashtoret.tau.ac.il

Received 20 September 2002; Accepted 30 January 2003

underlying assumption is that patches of highly conserved residues can serve as markers of functionally important regions of the protein.[9–13] In the following we present the results of ConSurf analysis of the combined CTL/CTLD superfamily and of the two selected clades of each of the individual families.

METHODS

We used the ConSurf program that was recently described in detail.[9] In the following we provide a brief outline, with emphasis on the adaptation of ConSurf to the present study and on various aspects that may have been left unclear in our earlier publication.

Constructing the Conservation Pattern of the Entire Protein Family

Multiple sequence alignment and phylogenetic tree

We obtained 238 CTL/CTLD sequences from the SMART web site[14] and multiply-aligned them with the CLUSTAL W program,[15] which has been integrated into the ConSurf package of programs. We then constructed a maximum parsimony evolutionary tree consistent with it, using the PROTPARS program from the PHYLIP package,[16] again with the use of the ConSurf package.

Calculation and presentation of the conservation grades

ConSurf[9] produces a general conservation score for each position in the alignment. Each exchange between any 2 of the 20 amino acids (and gaps) is multiplied by a weight factor according to the physicochemical difference between the amino acids, as determined by the replacement matrix of Miyata.[17] Thus, the physicochemical variability grade, P_k, at position k in the alignment ($k = 1,\ldots L$, where L is the total number of amino acid positions), is calculated as follows:

$$P_k = \sum_{m=1}^{N} (A_{ij}^{m}(k) M_{ij}) \qquad (1)$$

where A_{ij}^{m} is a matrix of elements 0 and 1, describing an exchange between amino acids i and j at any node m in the phylogenetic tree, M_{ij} is the replacement value obtained from the Miyata matrix, and N is the number of nodes in the phylogenetic tree.

After grading each position in the alignment, the grades are averaged. The average variability ($\langle P \rangle$) and standard deviation (σ) are calculated only for ungapped positions in the query protein (the protein with known 3D structure). ConSurf then normalizes each variability grade as follows:

$$W_k = (\langle P \rangle - P_k)/\sigma \qquad (2)$$

where W_k is, by definition, the set of conservation grades for every ungapped alignment position in the query protein. For technical reasons, associated with the use of some of the available molecular graphics programs, the normalized conservation grades are rescaled so that the maximal grade (corresponding to the most highly conserved position) is assigned the value of 1, and the minimum grade (corresponding to the most variable position) is assigned the value of 0. These normalized conservation grades, replacing the B (or temperature) factors in the PDB file of the protein, are now ready to be mapped on the 3D structure of the protein. We used GRASP[18] for mapping the conservation grades onto the molecular surface of the protein.

Calculating the Conservation Pattern of the Query Clade

The conservation grades for the same two proteins (either hMBP-C or the α subunit of IX/Xbp) were calculated again, using the method described above. This time, however, the MSA and phylogenetic tree that were constructed contained only the proteins residing in the specific clade (the CTL clade for hMBP-C and the CTLD toxin clade for the α subunit of IX/Xbp).

Homology Modeling

3D models of proteins from the CTLD toxin clade were constructed using the Modeller program[19] as implemented in the InsightII package (MSI, San Diego, CA). The α subunit of IX/Xbp was used as a template to which sequences were aligned using CLUSTAL X.[20] The models were built using the standard procedure with the default parameters. They are available upon request.

Electrostatic Calculations

The Poisson-Boltzmann equation was solved using GRASP[18] with the default parameters. Values of 2 and 80 were assigned to the dielectric constant inside and outside the protein, respectively. The salt concentration was set to its physiological value of 150 mM. The protein atoms were assigned atomic radii and partial charges using the PARSE set.[21]

RESULTS AND DISCUSSION

The MSA of the CTL/CTLD superfamily in the SMART database[14] contains 238 homologous sequences. Following the SWISS-PROT database annotations, 47 of them are CTLs and 35 are CTLDs, but the rest have not yet been annotated. (Among the latter are, e.g., more than 50 sequences from the *Caenorhabditis elegans* genome.) The MSA and the resultant phylogenetic tree are presented at the web site.[22] The resultant conservation pattern will be referred to as the *conservation pattern of the entire superfamily*. We considered separately two clades of the phylogenetic tree: (a) A CTL clade (Table I) that includes 15 proteins currently annotated as CTLs in the SWISS-PROT database and 6 putative proteins very similar to known CTLs. These 6 putative proteins are not well characterized experimentally, and the fact that our phylogenetic analysis placed them in the CTL clade is a strong indication that they are indeed CTLs. (The rest of the proteins in the CTL family are distributed between other clades.) (b) A toxin clade composed of the 11 heterodimeric snake venom CTLD toxins (Table II).

TABLE I. The Proteins Constituting the CTL Clade

Protein name	SWISS-PROT/ TREMBLE code	Identifier in the MSA and tree of Figure 3
Bovine conglutinin	CONG_BOVIN	CONG_BOVIN
Unknown source*	—	G2736145
Polyandrocarpa misakienesis lectin	LECC_POLMI	LECC_POLMI
Bovine L-selectin	LEM1_BOVIN	LEM1_BOVIN
Bovine E-selectin	LEM2_BOVIN	LEM2_BOVIN
Bovine P-selectin	LEM3_BOVIN	LEM3_BOVIN
Human E-selectin	LEM2_HUMAN	LEM2_HUMAN
American cockroach hemolymph lipopolysaccharide-binding protein	LPSB_PERAM	LPSB_PERAM
Rat mannose-binding protein A	MABA_RAT	MABA_RAT
Human mannose-binding protein C	MABC_HUMAN	MABC_HUMAN
Mouse mannose-binding protein C	MABC_MOUSE	MABC_MOUSE
American cockroach lectin-related protein*	P92047	P92047
American cockroach lectin-related protein*	P92048	P92048
American cockroach lectin-related protein*	P92049	P92049
American cockroach lectin-related protein*	P92050	P92050
American cockroach lectin-related protein*	P92051	P92051
Human pulmonary surfactant-associated protein A	PSPA_HUMAN	PSPA_HUMAN
Bovine pulmonary surfactant-associated protein D	PSPD_BOVIN	PSPD_BOVIN
Human pulmonary surfactant-associated protein D	PSPD_HUMAN	PSPD_HUMAN
Beef shark tetranectin-like protein	TETN_CARSP	TETN_CARSP
Human tetranectin	TETN_HUMAN	TETN_HUMAN

Fifteen of these proteins have been designated as CTLs by SWISS-PROT and the other 6, marked with asterisks, are included here since our phylogenetic analysis has suggested that they are CTLs due to their high sequence identity to known CTLs. The proteins are sorted alphabetically by their tree-identifier (the right-most column).

Analyzing the entire superfamily revealed patches of conserved residues shared by all (or most) of the proteins in the superfamily. These are probably related to functions that are typical of the entire superfamily, such as binding to other proteins or protein domains. Similarly, from analyses of the conservation pattern obtained for the clades, patches of conserved residues could be detected that are related to functions unique to the proteins constituting each clade.

TABLE II. The Proteins Constituting the CTLD Toxin Clade

Protein name	SWISSPROT/ TREMBLE code	Identifier in the MSA and tree of Figure 6
Alboaggregin A α subunit from white-lipped pit viper	ABA1_TRIAB	ABA1_TRIAB
AlboaggreginA β subunit from white-lipped pit viper	ABA3_TRIAB	ABA3_TRIAB
Bitiscetin α subunit from *Bitis arietans* snake[30]	—	bitisceA
Bitiscetin β subunit *Bitis arietans* snake[30]	—	bitisceB
Botrocetin α chain from *Bothrop jararaca* snake	BOTA_BOTJA	BOTA_BOTJA
Bothrocetin β subunit from *Bothrop jararaca* snake	BOTB_BOTJA	BOTB_BOTJA
Bothrojaracin α subunit from *Bothrops jararaca* snake[31]	—	bothrojaA
Bothrojaracin β subunit *Bothrops jararaca* snake[31]	—	bothrojaB
CHHB α subunit from timber rattlesnake[28]	—	CHHBa
CHHB β subunit from timber rattlesnake[28]	—	CHHBb
Echicetin α subunit from saw-scaled viper	ECHA_ECHCA	ECHA_ECHCA
Echicetin β subunit from saw-scaled viper	ECHB_ECHCA	ECHB_ECHCA
Coagulation factor IX/factor X-binding protein α subunit from habu snake	IXA_TRIFL	IXA_TRIFL
Coagulation factor IX/factor X-binding protein β subunit from habu snake	IXB_TRIFL	IXB_TRIFL
Convulxin α subunit from snake south American rattlesnake	O93426	O93426
Convulxin α subunit from snake south American rattlesnake	O93427	O93427
Platelet glycoprotein IB - binding protein α subunit from jararaca snake	Q9PSM6	Q9PSM6
Platelet glycoprotein IB - binding protein β subunit from jararaca snake	Q9PSM5	Q9PSM5
Mamushigin α subunit from gloydiusblomhoffii snake	Q9YGG9	Q9YGG9
Mamushigin β subunit from glyodiusblomhoffii snake	Q9YI92	Q9YI92
Rhodocetin α subunit from Malayan pit viper	RHCA_AGKRH	RHCA_AGKRH
Rhodocetin β subunit from Malayan pit viper	RHCB_AGKRH	RHCB_AGKRH

The proteins are sorted alphabetically by their tree-identifier (the right-most column).

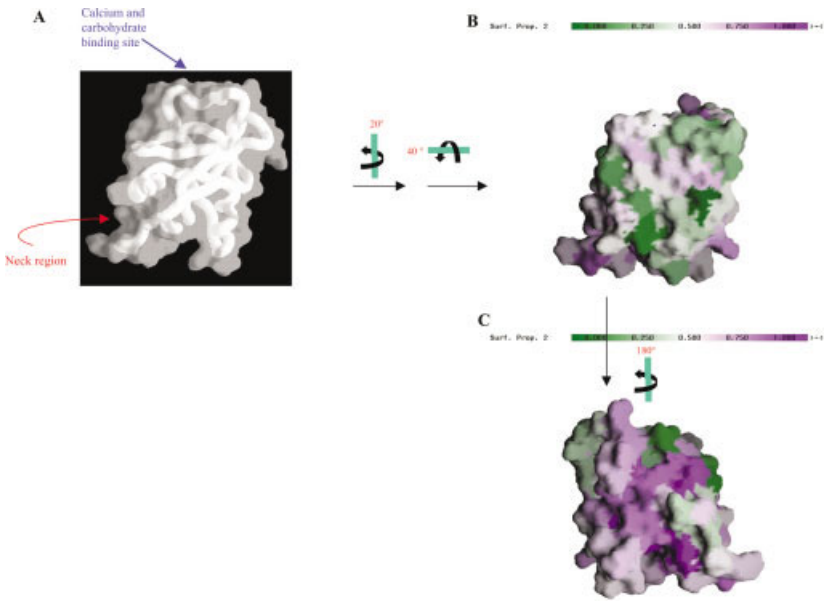

Figure 1.

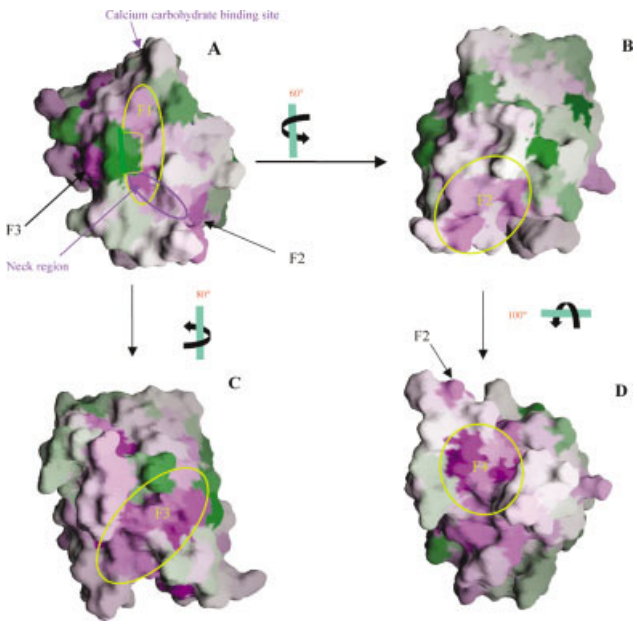

Figure 2.

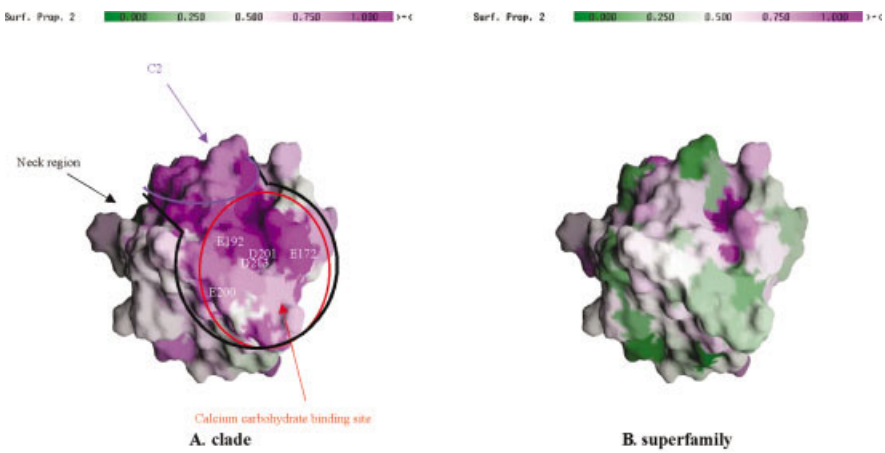

Figure 4.

The CTL/CTLD Superfamily

We mapped the conservation grades of the entire superfamily (all 238 CTL and CTLD domains) onto the molecular surface of human MBP-C (hMBP-C), a member of the CTL family. In this protein the CRD, the structure of which is presented, is attached to an α-helical coiled-coil neck region, followed by a collagen-like domain and a cysteine-rich domain [Fig. 1(A)]. Analysis of the conservation pattern obtained revealed that overall, the protein has two faces: one is predominantly composed of evolutionarily divergent residues [Fig. 1(B)], and the other mainly of conserved residues [Fig. 1(C)]. Similar results were obtained when the superfamily conservation grades were mapped onto other of its members, for example, human surfactant protein D,[23] human tetranectin,[24] and IX/Xbp[6] (data not shown). These results suggest that, on average, the conserved face, shared by the proteins of the superfamily, is involved in functions that are common to all its members, while the opposite, variable face is either less important functionally or alternatively is involved in functions that are unique to different families or individual proteins. This issue is discussed in "Limitations of the model" below.

A more detailed analysis revealed four patches of conserved residues (marked as F1–F4 in Fig. 2), three of which (F1, F3, and F4) are located on the conserved face, whereas the fourth (F2) is on the border between the conserved and divergent faces. The various patches are close to one another and can be viewed as a continuous super-patch, which might be involved in a common function. Patch F1 corresponds to the dimerization region of IX/Xbp; its role, however, may be distinct from that in other members of the superfamily. This is discussed in "The CTLD toxin clade" below.

The CTL Clade

The clade sequence alignment is presented in Figure 3(A), and the resultant phylogenetic reconstruction is presented in Figure 3(B). We compared the conservation pattern calculated for the CTL/CTLD superfamily with that calculated for the clade. Both patterns were separately mapped onto the molecular surface of hMBP-C. The comparison revealed two regions of highly conserved residues (C1 and C2 in Figs. 4 and 5), which are unique to the clade. For example, one of those [C2 in Fig. 4(A), surrounded by a black line] is essentially missing in the conservation pattern calculated for the entire superfamily [Fig. 4(B)]. This patch includes amino acids that are involved in sugar and calcium binding in hMBP-C, as seen in the X-ray crystal structure of the protein[25] [Fig. 4(A), circled in red]. A similar patch was detected in other proteins of the clade, such as human surfactant protein D[23] and human tetranectin,[24] and it also consists of amino acids that are involved in sugar-calcium binding. The fact that this region was not conserved throughout the entire superfamily [Fig. 4(B)] is in accord with the fact that many of its members either do not bind sugar and calcium or if they do so, such binding is not in the same region.

The two patches of highly conserved amino acids, C1 and C2 [Fig. 5(A,B), black circles], are located on opposite sides from one another and on the divergent face of the protein close to the border with the conserved face. The surface electrostatic potential of the protein, calculated using GRASP [Fig. 5(C,D)], indicates that these patches are the least polar in hMBP-C.

The CTLD Toxin Clade

These toxins affect blood coagulation markedly, some acting as inhibitors, other as activators of this process. They do not bind carbohydrates or calcium[7,8] (with the exception of IX/Xbp, which does bind calcium but in a different location from that of the CTLs). Indeed, the ConSurf analysis showed that the five residues that coordinate calcium and are essential for carbohydrate recognition in many CTLs are not conserved in the clade.

The clade sequence alignment is presented in Figure 6(A) and the corresponding phylogenetic reconstruction is provided in Figure 6(B). As is evident from the figure, this clade consists of two subclades, one of the α subunits of the toxins together with the β subunit of rhodocetin, and the other of the remaining β subunits.

Fig. 1. The 3D structure of the CTL of human mannose-binding protein-C (hMBP-C, pdb entry: 1hup[25]). The pictures were produced using GRASP.[18] (A) A worm presentation of the backbone and a transparent presentation of the molecular surface. The calcium and carbohydrate binding region and the connection to the α-helical coiled-coil neck region are marked. (B,C) ConSurf mapping of evolutionary conservation onto the molecular surface of the domain is color-coded such that dark green responds to maximal variability, white corresponds to average conservation level, and dark purple to maximal conservation. (B,C) Opposite views of the domain: (B) is the mostly divergent side and (C) is the predominantly conserved side.

Fig. 2. ConSurf mapping of evolutionary conservation in the CTL/CTLD superfamily onto the molecular surface of the CTL domain of hMBP-C. The pictures were produced using GRASP,[18] and the residue conservation is color-coded as in Figure 1. Four patches, of highly conserved residues, were detected, each of which is highlighted by a yellow circle in (A–D). The protein orientation in (A) was rotated 120° to the left, on the *y*-axis, relative to its orientation in Figure 1(A). Patch F1 corresponds to the dimerization region in IX/Xbp (Fig. 7). The amino acids comprising the patches in hMBP-C are as follows. (A) Patch F1: Phe119, Asn151, Gly152, Ala153, Gln155, Tyr185, and Asn187. (B) Patch F2: Ser109, Leu110, Gly111, Lys112, Thr121, and Leu157. (C) Patch F3: Ala140, Ser141, Arg146, Asn147, Glu150, Asp177, Leu178, and Glu225. (D) Patch F4: Phe118, Leu120, Phe138, Gln139, Ala222, and Cys224.

Fig. 4. Conservation pattern of the CTL clade (A) and of the superfamily (B) mapped onto the molecular surface of hMBP-C: the calcium- and carbohydrate-binding site. The structure of hMBP-C has been rotated downward, 90° around the *x*-axis, relative to the orientation of Figure 1(A) so that the neck region is located in the opposite side of the protein. (A) Phylogenetic conservation pattern obtained for the clade of Figure 3. A large patch of conserved residues is observed (marked by the black line). It can be decomposed into patch C2 (blue circle; see Fig. 5 for more detail) and a smaller patch, which corresponds to the calcium-binding site (red circle). The latter is composed of Glu 172, Gly 173, Phe 175, Leu 183, Trp 188, Pro 193, Glu 192, Asn 195, Glu 200, Asp201, and Asp 213. The amino acids that are involved in the coordination of the calcium ion are marked. (B) Conservation pattern of the entire superfamily. Evidently, evolutionary conservation in the calcium-binding region is limited to the clade rather than being shared by all the proteins in the superfamily. The pictures were produced using GRASP[18] and the residue conservation is color-coded as in Figure 1.

A

```
TETN_CARSP/11      --CLKGTKIHKKCYLASRGSKSYHAANEDCIAQGGTLSIPRSSDEGNSLRSYAKKSL---
TETN_HUMAN/11      --CLKGTKVHMKCFLAFTQTKTFHEASEDCISRGGTLSTPQTGSENDALYEYLRQSV---
PSPD_BOVIN/11      --FPNGRSVGEKIFKTVGSEKTFQDAQQICTQAGGQLPSPRSGAENEALTQLATA-----
HSP_D              ---------GEKIFKTAGFVKPFTEAQLLCTQAGGQLASPRSAAENAALQQLVVA-----
CONG_BOVIN/11      -LFPDGQAVGEKIFKTAGAVKSYSDAEQLCREAKGQLASPRSSAENEAVTQMVRA-----
MABC_HUMAN/11      -TFSLGKQVGNKFFLTNGEIMTFEKVKALCVKFQASVATPRNAAENGAIQNLIK------
MABA_RAT/1122      HAFSMGKKSGKKFFVTNHERMPFSKVKALCSELRGTVAIPRNAEENKAIQEVAK------
MABC_MOUSE/11      VLFSLSEKVGKKYFVSSVKKMSLDRVKALCSEFQGSVATPRNAEENSAIQKVAK------
G2736145/1121      -SLKDVVNIGKKMFVSTGKKYNFEKGKSLCAKAGSVLASPRNEAENTALKDLIDP-----
PSPA_HUMAN/11      SLQGSIMTVGEKVFSSNGQSITFDAIQEACARAGGRIAVPRNPEENEAIASFVKK-----
P92047/1131        --LGYDLVPGFGYYKLHTDVKTWHEALRACEQEGAHLAIINSEAEAKSLTPFWDMNP--K
P92048/1127        ASLGYNLVAGLGYYKFHTDPKTWHEAQNICEKEGTHLAVINSEYEAKVLT----------
P92050/1134        -APDYELVPGLGYYKLHTDVNTWHNAKMVCEEEGAHLVVINSEKEAQVLKNLWNKTPPAK
P92049/1135        PSAGYELLPGLGYYKFHTDYKNWYDARKTCIQEGAHLAVINSETESKALLKLWLPHP--K
LPSB_PERAM/11      IPPGYELSAVLGYYKFHKTPKTWDEARIICQQEGGHLVIINSEDESKVLQNLFSKVT--K
P92051/1134        PRPGYVLFPKLGYFKLYNIGKTWNEAKLICEAEGAHLGIVNSKKEVEIVQELRARLP--K
LEM2_BOVIN/11      --LLLFKESRTWSYHASTEMMTFEEARDYCQKTYTALVAIQNQEEIEYLNSTFS------
Eselect            -----------WSYNTSTEAMTYDEASAYCQQRYTHLVAIQNKEEIEYLNSILS------
LEM1_BOVIN/11      CDFFAHHGTDCWTYHYSKRPMPWEKARAFCRENYTDLVAIQNKGEIEYLNKTLP------
LEM3_BOVIN/11      FEVINQKEVSAWTYHYSNKTYSWNYSRAFCQKYYTDLVAIQNKNEIAYLNETIP------
polyan             ---------MDYEILFSDETMNYADAGTYCQSRGMALVSSAMRDSTMVKAILAFT-----

TETN_CARSP/11      ----VGARDFWIGVNDMTTEGKFVDVNGL-PITYFNWD---RSKPVG---GTRENCVAAS
TETN_HUMAN/11      ----GNEAEIWLGLNDMAAEGTWVDMTGA-RIAYKNWETEITAQPDG---GKTENCAVLS
PSPD_BOVIN/11      -----QNKAAFLSMSDTRKEGTFIYPTGE-PLVYSNWA---PQEPNN--DGGSENCVEIF
HSP_D              -----KNEAAFLSMTDSKTEGKFTYPTGE-SLVYSNWA---PGEPND--DGGSEDCVEIF
CONG_BOVIN/11      -----QEKNAYLSMNDISTEGRFTYPTGE-ILVYSNWA---DGEPNNSDEGQPENCVEIF
MABC_HUMAN/11      -------EEAFLGITDEKTEGQFVDLTGN-RLTYTNWN---EGEPNN--AGSDEDCVLLL
MABA_RAT/1122      -------TSAFLGITDEVTEGQFMYVTGG-RLTYSNWK---KDEPND--HGSGEDCVTIV
MABC_MOUSE/11      -------DIAYLGITDVRVEGSFEDLTGN-RVRYTNWN---DGEPNN--TGDGEDCVVIL
G2736145/1121      ------SSQAYIGISDAQTEGRFMYLSGG-PLTYSNWK---PGEPNN--HKN-EDCAVIE
PSPA_HUMAN/11      -----YNTYAYVGLTEGPSPGDFRYSDGT-PVNYTNWY---RGEPAG---RGKEQCVEMY
P92047/1131        ILDGGSNNWAHVGFHDQYKEGQYLTIFNQ-SLVAAGYIKWNPGEPH----GVGANCGCVI
P92048/1127        -N-MGDFTWAYVGFHDNYVDGQYVTIFNQ-SLYDAGYAKWHRGEPVP---GSEYNCGCIS
P92050/1134        ITGGTHTDWAWIGFHDLYQEGEFVTIFNE-TLKSAGYSKFHPSDGKG---GTSQNCGLID
P92049/1135        MFNDWRNDWAHIGFHDHYTEGQFVTIFDT-PLNEAGFSKWQPPNPDG---GDKDDCGVFR
LPSB_PERAM/11      TEGATNNDYIFIGIHDRFVEGEFITIFGK-PLATTGFTRWVDSIQPDN-AGGNENCGSMH
P92051/1134        LFGNWVDDHIYTGVNDLQHANTWVTIFEQ-PLSATGFSVWDNGEPD---VGANEHCVALH
LEM2_BOVIN/11      ------YSPSYYWIGIRKINGTWTWIGTN-KSLTKEATNWAPGEPNN--KQSDEDCVEIY
Eselect            ------YSPSYYWIGIRKVNNVWVWVGTQ-KPLTEEAKNWAPGEPNN--RQKDEDCVEIY
LEM1_BOVIN/11      ------FSRTYYWIGIRKVEGVWTWVGTN-KSLTEEAKNWGAGEPNN--RKSKEDCVEIY
LEM3_BOVIN/11      ------YYNSYYWIGIRKINNKWTWVGTK-KTLTEEAENWADNEPNN--KRNNQDCVEIY
polyan             ----EVKGHDYWVGADNLQDGAYNFLWNDGVSLPTDSDLWSPNEPSN--PQSWQLCVQIW

TETN_CARSP/11      TSG---QGKWSDDVCRSE-KRYICE-----
TETN_HUMAN/11      GAA---NGKWFDKRCRDQ-LPYICQ-----
PSPD_BOVIN/11      P-----NGKWNDKVCGEQ-RLVICE-----
HSP_D              T-----NGKWNDRACGEK-RLVVCEF----
CONG_BOVIN/11      P-----DGKWNDVPCSKQ-LLVICE-----
MABC_HUMAN/11      K-----NGQWNDVPCSTS-HLAVCE-----
MABA_RAT/1122      D-----NGLWNDISCQAS-HTAVCE-----
MABC_MOUSE/11      G-----NGKWNDVPCSDS-FLAICE-----
G2736145/1121      D-----SGKWNDLDCSNSNIFIICE-----
PSPA_HUMAN/11      T-----DGQWNDRNCLYS-RLTICE-----
P92047/1131        RR----ENLLADIICTAK-QPFFCE-----
P92048/1127        KE----DSSLGVAKCRDK-IEFFCE-----
P92050/1134        R-----AVQLGDHSCEDK-DPFFCE-----
P92049/1135        RN----FGTLGDIPCSAK-LAFICE-----
LPSB_PERAM/11      P-----NGGLNDIPCPWK-LPFVCE-----
P92051/1134        T----------------------------
LEM2_BOVIN/11      IKREKDSGKWNDEKCTKQ-KLALCY-----
Eselect            IKREKDVGMWNDERCSKK-KLALC------
LEM1_BOVIN/11      IKRNKDSGKWNDDACHKA-KTALCY-----
LEM3_BOVIN/11      IKSLSAPGKWNDEPCWKR-KRALCY-----
polyan             S----KYNLLDDVGCGGA-RRVICEKELDD
```

Fig. 3. The CTL clade of Table I. (A) MSA of the sequences comprising the clade. (B) The phylogenetic relationships in the clade. The black arrow indicates the location of hMBP-C.

Again, we compared the conservation pattern calculated for the CTL/CTLD superfamily with that calculated for the toxin clade, by mapping the patterns onto the molecular surface of the α subunit of IX/Xbp. The high-resolution structure of this subunit isolated from the habu snake[6] is shown in Figure 7(A). The general fold is similar to that of the CTL domain of hMBP-C [Fig. 1(A)], except for a large open loop that extends away

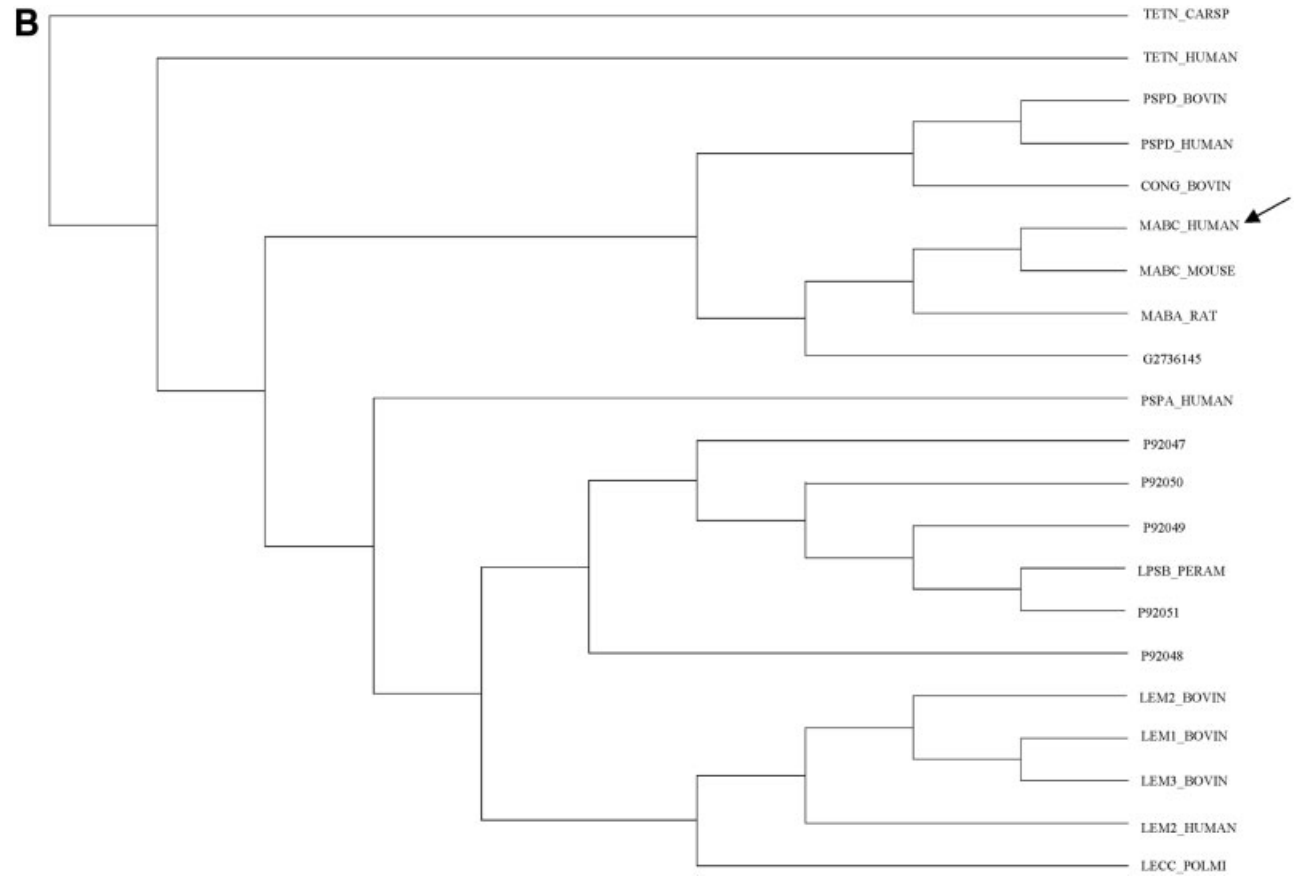

Figure 3. (Continued.)

from the protein, projecting into the adjoining subunit and generating a loop exchange heterodimer; in hMBP-C the corresponding chain is folded back onto the monomer.[6]

Analysis of the conservation pattern calculated for the toxin clade and mapped on the α subunit of IX/Xbp revealed a region of conserved residues at the interface with the β subunit of this toxin [Fig. 7(B)]. A similar conserved region was detected in the pattern calculated for the β subunit and for the entire superfamily (data not shown). The location of this region corresponds to patch F1 detected in the conservation pattern of the entire superfamily (Fig. 2). It contains residues that are highly conserved in the toxin clade and to a lesser extent throughout the entire superfamily as well, indicating that they have an important role not just in the toxin α subunit, but in other proteins of the superfamily too. As suggested for the corresponding patch F1, these amino acids, which are involved in the binding to the other subunit, may play a similar role in other proteins or function in binding to other targets.

Another conserved patch found in the toxin clade is T1 [Fig. 8(A), blue circle], which is missing in the conservation pattern of the entire superfamily [Fig. 8(B)]. Part of patch T1, marked as T1a in Figure 8(A) (white circle), which overlaps to ~80% with patch C1 in the CTL clade (Fig. 5), is not a part of the dimerization interface. In contrast to patch C1, characterized by a weak electrostatic potential, patch T1a has a strong positive potential, as can be seen in the surface electrostatic potentials calculated using GRASP [Fig. 8(C)]. It was of interest to test whether the electrostatic potential just mentioned is characteristic of all the toxins in the α sub-clade. To this end, we built 3D models of the α subunit of several other toxins in the clade (botrocetin, mamushigin, GPIB-bp, CHH-Ba, rhodocetin, convulxin, alboaggregin, and echicetin), using the IX/Xbp α subunit as a template, and calculated the electrostatic potential around each of them. Most of these toxins do indeed have a positive surface potential in a region corresponding to patch T1a (results not shown). The two exceptions are GPIB-bp, which has a positive patch located in a slightly different region, and convulxin, which does not have a positive potential in this region at all.

Possible role for patch T1a

Most of the CTLD snake toxins bind to negatively charged receptors or other proteins.[6,26–28] We speculate that patch T1a, which is characterized by a positive electrostatic potential, plays a role in the binding of the toxins to their receptors or other proteins. For example,

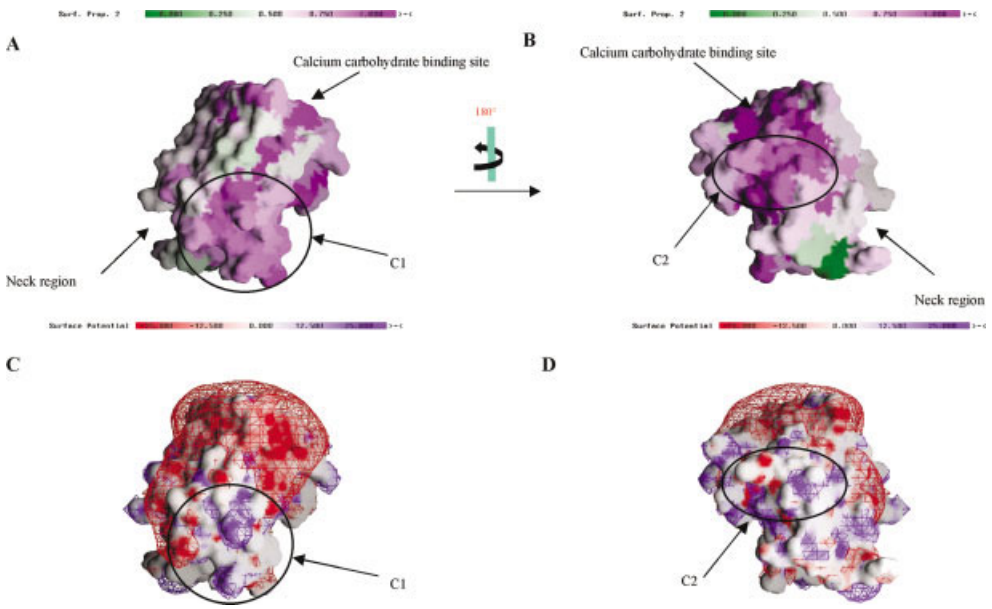

Figure 5.

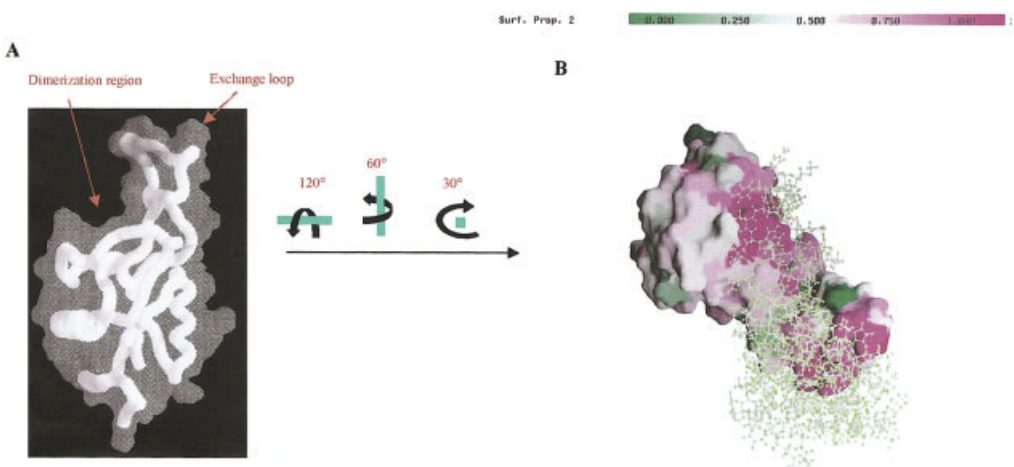

Figure 7.

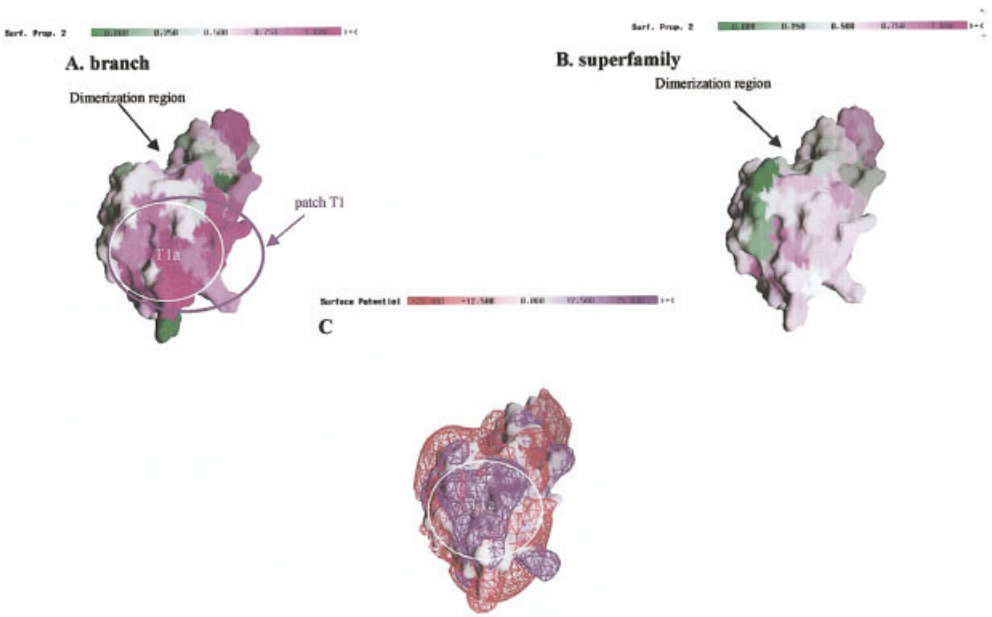

Figure 8.

IX/Xbp binds to coagulation factors IX and X through a calcium ion that interacts with their negatively charged γ-carboxyglutamic acid domains.[6] We propose that the positive patch of IX/Xbp is involved in enhancing the interactions between the toxin and the above coagulation factors. Interestingly, Mizuno et al.[6] noted a cluster of basic residues (Lys100, Lys105, Lys107, and Arg109) in subunit α in the X-ray crystal structure of the IX/Xbp in a location distinct from patch T1a. They further speculated that these positively charged residues interact with the negatively charged γ-carboxyglutamic acid residues of factors IX and X. The positively charged region corresponding to this cluster appears to be composed of variable residues, suggesting that it is unique to IX/Xbp. A detailed discussion of other negatively charged binding targets of the other snake toxins located in the clade is presented at the web site.[22]

LIMITATIONS OF THE MODEL

Possible problems related to the phylogenetic reconstruction and the evolutionary analysis performed with the ConSurf program have been discussed in our earlier report.[9] In the following we elaborate on the issues that are of particular importance for the present study.

Fig. 5. Conservation pattern of the CTL clade (A) and of the superfamily (B) mapped onto the molecular surface of hMBP-C: patches C1 and C2. The pictures were produced using GRASP,[18] and the residue conservation is color-coded as in Figure 1. (A) Patch C1 (black circle), conserved only in the clade (Fig. 3) and not in the superfamily. The structure has been rotated 30° to the right, around the *y*-axis, relative to the orientation of Figure 1(A). Patch C1 is constructed of Gly123, Glu124, Ile125, Met126, Thr127, Glu129, Lys130, Ala133, Leu134, Val136, Lys137, and Gln139. (B) Patch C2 (black circle), conserved only in the clade (Fig. 3) and not in the superfamily. The structure has been rotated 150° to the left, around the *y*-axis, relative to the orientation of Figure 1(A). Patch C2 is constructed of Ala149, Ala148, Leu183, Thr184, Tyr185, and Thr186. This patch is located near the calcium- binding region. (C,D) A projection of the electrostatic potential (ϕ) onto the molecular surface of the protein domain (oriented as in A and B, respectively); $\phi > 25$ kT/e is dark red, $\phi = 0$ is white, and $\phi < -25$ kT/e is dark blue. Equipotential contours of $\phi = \pm 2$ kT/e are depicted in blue and red mesh.

Fig. 7. (A) The CTLD of subunit α of IX/Xbp[6] (pdb entry: 1ixx), oriented as the hMBP-C in Figure 1(A). A transparent presentation of the molecular surface is shown, and the dimerization region and loop extending toward the β subunit (not shown) are marked. (B) Evolutionary conservation, calculated according to the CTLD toxin clade (Fig. 6), was mapped onto the molecular surface of the α subunit. Subunit β is displayed as a yellow ball-and-stick model. Most of the surface area involved in the dimerization with this subunit is highly conserved. The pictures were produced using GRASP,[18] and residue conservation grades are color-coded as in Figure 1.

Fig. 8. Patch T1 in the α subunit of IX/Xbp. Evolutionary conservation grades, calculated according to the CTLD toxin clade (Fig. 6), were mapped onto the molecular surface of the α subunit of IX/Xbp. The structure has been oriented to the right, 30° on the *y*-axis and downward, 30° on the *x*-axis relative to the orientation in Figure 7(A). (A,B) Patch T1 (blue circle) is conserved in the clade (A) but not in the superfamily (B). Patch T1 comprised of patch T1a (white circle) composed of amino acids Leu3, Ser4, Trp6, Lys15, Ala16, Phe17, Glu18, Asp25, Val29, Arg28, Glu32, Gln33, Lys35, Lys60, and Arg61. (C) Projection of the electrostatic potential (ϕ) onto the molecular surface of the α subunit of IX/Xbp (oriented as in A and B); $\phi > 25$ kT/e is dark red, $\phi = 0$ is white, and $\phi < -25$ kT/e is dark blue. Equipotential contours of $\Phi = \pm 1$ kT/e are depicted in blue and red mesh. A region of positive potential is observed in the area corresponding to patch T1a.

The separation of the CTL/CTLD superfamily into subfamilies, which is central to our analysis, is based on the topology of the phylogenetic tree. Thus, it is important to verify that the tree topology is correct. To do this, we compared the tree used here to the one constructed by us according to the neighbor-joining (NJ) method,[29] as implemented in CLUSTAL X.[20] The trees created by NJ and PROTPARS are very similar; furthermore, the two clades analyzed here were found in both trees. However, the CTL clade contained 10 more protein sequences according to the NJ method than the equivalent clade according to the protein parsimony method. It is noteworthy that the six putative proteins of Table I were in the CTL clade of the phylogenetic trees generated by both NJ and PROTPARS. It has been suggested, based on sequence alignment, that the six proteins are CTLs and our phylogenetic analysis provides strong evidence in favor of this suggestion.

The outcome of the ConSurf analysis depends crucially on the quality of the MSA; for example, the inclusion of nonhomologous sequences in the MSA may introduce noise that will screen the signals from evolutionarily conserved regions within the true homologous proteins. Thus, we used an MSA of the CTL/CTLD superfamily from the SMART database.[14] SMART presents experts-derived MSAs, in which the relations between sequence, structure and function were carefully studied.

The ConSurf analysis presented in this article is based on eyeballing the amino acid conservation pattern that was color-coded onto the molecular surface of the protein. It is qualitative rather than quantitative, and a measure of the likelihood of each patch of conserved region to be statistically significant is missing. However, recent studies that were carried out with datasets of tens of protein complexes of known 3D structure demonstrated that patches of evolutionarily conserved residues are very often indicative of functionality.[10–13] The current study further underscores this point in that it demonstrated that the known functionally important regions within the CLT and CTLD families are associated with patches of evolutionarily conserved residues. Other patches of putatively functionally important residues were detected, whose conservation scores were equally high, suggesting that they are significantly important too.

One of our main goals in this study was to focus on clades of the phylogenetic tree and to find specific characteristics of the proteins in these clades. To this end, we had to use multiple sequence alignment containing sequences of limited diversity. Thus, the conservation patterns obtained may be a combination of functionally important residues and shortness of divergence time. Reassuringly, at least one of the evolutionarily conserved patches detected in each of the clades corresponds to a well-characterized functionally important region, namely the calcium-binding region of hMBP-C and the dimerization region of IX/Xbp.

A

```
ABA1_TRIAB/1-1      -----CPSDWSSYDQYCYRVFKRIQTWED-----AERFCSEQANDGHLVSIESAG-EADF
ix/xA               ----DCLSGWSSYEGHCYKAFEKYKTWED-----AERVCTEQAKGAHLVSIESSG-EADF
convulA             ----HCPSDWYYYDQHCYRIFNEEMNWED-----AEWFCTKQAKGAHLVSIKSAK-EADF
BOTA_BOTJA/1-1      -----CPSGWSSYEGNCYKFFQQKMNWAD-----AERFCSEQAKGGHLVSIKIYSKEKDF
bothrjarA           ----DCPSDWSSHEGHCYKFFQQKMNWAD-----AERFCSEQAKGGHLVSFQSDG-ETDF
GPIB_bpA            DTPFECPSDWSTHRQYCYKFFQQKESWDDRSEYDAERFCSEQAKGGHLVSIESDE-EADF
mamushA2            ----DCPSDWSSNGRFCYKLFQQKMKWAD-----AERFCTEQRTGAHLVSIESNT-EAAF
rhodoceA            ----DCPDGWSSTKSYCYRPFKEKKTWEE-----AERFCTEQEKEAHLVSMENRL-EAVF
rhodoceB            --DFRCPTTWSASKLYCYKPFKEKKTWIE-----AERFCAKQAENGHLVSIGSAA-EADF
CHH-Ba              --DLECPSGWSSYDRYCYKPFKQEMTWAD-----AERFCSEQAKGRHLLSVETAL-EASF
ECHA_ECHCA/1-1      -----CLSGWSFYEGHCYQLFR-LKTWDE-----AEKYCN-QWDGGHLVSIESNA-KAEF
bitiscetA           --DPGCLPDWSSYKGHCYKVFKKVGTWED-----AEKFCVE--NSGHLASIDSKE-EADF
BOTB_BOTJA/1-1      -----CPPDWSSYEGHCYRFFKEWMHWDD-----AEEFCTEQQTGAHLVSFQSKE-EADF
bothrojarB          ----DCPPDWSSYEGSCYRVFEQKMNWED-----AEKFCSQQQTGGHLVSFQSSE-EADF
ABA3_TRIAB/11       -----CPFGWSSYEGYCYKVYNKKMNWED-----AESFCREQHKRSHLVSFHSSG-EVDF
echicB              ----NCLPDWSVYEGYCYKVFKERMNWAD-----AEKFCMKQVKDGHLVSFRNSK-EVDF
mamhshB2            ----DCPSDWSSYEGHCYRVFQKEMTWED-----AEKFCTQQRKESHLVSFHSSE-EVDF
CHH-Bb              ----DCPSDWSSYEGHCYRVFQQEMTWDD-----AEKFCTQQHTGGHLVSFRSSE-EVDF
convulB             --GFCCPSHWSSYDRYCYKVFKQEMTWAD-----AEKFCTQQHTGSHLVSFHSTE-EVDF
GPIB_bpB            ----DCPSDWSPYGGHCYKLFKQRMNWAD-----AENLCAQQRKESHLVSFHSSE-EVDF
IXB_TRIFL/111       -----CPSDWSSYEGHCYKPFSEPKNWAD-----AENFCTQQHAGGHLVSFQSSE-EADF
bitiscetinB         --DEGCLPDWSSYKGHCYKVFKVEKTWAD-----AEKFCKELVNGGHLMSVNSRE-EGEF

ABA1_TRIAB/1-1      VTQLVSENIRSEKH--YVWIGLRVQGKGQQCSSEWSDGSSVHYDNLQENKTRKCYGLEKR
ix/xA               VAQLVTQNMKRLDF--YIWIGLRVQGKVKQCNSEWSDGSSVSYENWIEAESKTCLGLEKE
convulA             VAWMVTQNIEESFS--HVSIGLRVQNKEKQCSTKWSDGSSVSYDNLLDLYITKCSLLKKE
BOTA_BOTJA/1-1      VGDLVTKNIQSSDL--YAWIGLRVENKEKQCSSEWSDGSSVSYENVVERTVKKCFALEKD
bothrjarA           VVNLVTEKIQSTDL--YAWIGLRVQNKEKQCSSKWSDGSSVSYENVVGRTVKKCFALEKE
GPIB_bpA            VAQLVAPNIGKSKY--YVWIGLRIENKKQQCSSKWSDYSSVSYENLVRGNVKKCFALEKK
mamushA2            VNQMISENIKKTD---YVWIGLTVQNEEQQCKSRWSDRSSVSYENLVKPNSKKCFVLKEY
rhodoceA            VDMVMENNFENKIY--RSWIGLKIENKGQRSNLEWSDGSSISYENLYEPYMEKCFLMDHQ
rhodoceB            LDLVIVVNFDKQRY--RAWTGLTERN------LKWTNGASVSYENLYEPYIRKCFVVQPW
CHH-Ba              VDNVLYANKEYLTR--YIWIGLRVQN-------KGQPCSSIYSENLVDPF--ECFMVSDR
ECHA_ECHCA/1-1      VAQLISRKLPKSAIEDRVWIGLRDRSKREQCGHLWTDNSFVHYEHVVPPT--KCFVLERQ
bitiscetA           VTKLASQTLTKFVY--DAWIGLRDESKTQQCSPQWTDGSSVVYENVDEPT--KCFGLDVH
BOTB_BOTJA/1-1      VRSLTSEMLKGDV----VWIGLSDVWN--KCRFEWTDGMEFDYDDYY--LIAEYECVASK
bothrojarB          VVSLTSPILRDSF----VWTGLSDVWK--ECSFEWSDGSDLSYKDNYQFVFSEYECVASK
ABA3_TRIAB/11       VVSKTFPILRYDF----VWMGLSDIWK--ECTKEWSDGARLDYKAWS----GKSYCLVSK
echicB              MISLAFPMLKMEL----VWIGLSDYWR--DCYWEWSDGAQLDYKAWD----NERHCFAAK
mamhshB2            VVSMTWPILKYDF----VWIGLNNIWN--ECMVEWTDGTRLSHNAWI----TESECIAAK
CHH-Bb              LVS----ILKFDL----FWMGWRDIWN--ERRLQWSDGTKVNYKAWS----AEPECIVCR
convulB             VVKMTHQSLKSTF----FWIGANNIWN--KCNWQWSDGTKPEYKEWH----EEFECLISR
GPIB_bpB            LVLLTFPILGPDL----YWTGLSNIWN--GCSFEWSDGTKVNYNAWA----SESECVASK
IXB_TRIFL/111       VVKLAFQTFGHSI----FWMGLSNVWN--QCNWQWSNAAMLRYKAWA----EESYCVYFK
bitiscetinB         ISKLALEKMRIVL----VWIGLSHFWR--ICPLRWTDGARLDYRALS----DEPICFVAE

ABA1_TRIAB/1-1      AEFRTWSNVYCGHEYPFVCK------
ix/xA               TDFRKWVNIYCGQQNPFVCEA-----
convulA             TGFRKWFVASCIGKIPFVCKFPPQC-
BOTA_BOTJA/1-1      LGFVLWINLYCAQKNPFVCK------
bothrjarA           QEFFVWINIYCGQQNPFVCKSPPP--
GPIB_bpA            QGFRKWVNIDCVEGNPFVCKFIRPR-
mamushA2            EGSRKWFNVYCGQKYAFMCKFLRPR-
rhodoceA            SGLPKWHTADCEEKNVFMCKFQLPR-
rhodoceB            EGKSKWYKADCEEKNAFLCKFPKPH-
CHH-Ba              TRLREWFKVDCEQQHSFICKFTRPRR
ECHA_ECHCA/1-1      TEFRKWIAVNCEFKFPFVCK------
bitiscetA           TEYRTWTDLPCGEKNPFICKSRLPH-
BOTB_BOTJA/1-1      PTNNKWWIIPCTRFKNFVCE------
bothrojarB          TKNNKWRIIPCTKLEYFVCEFQA---
ABA3_TRIAB/11       TTNNEWLSMDCSRTRYPVCK------
echicB              TTDNQWMRRKCSGEFYFVCKCPA---
mamhshB2            TTDNQWLSRPCSRTYNVVCKFQE---
CHH-Bb              ATDNQWLSTSCSKTHNVVCKF-----
convulB             TFDNQWLSAPCSDTYSFVCKFEA---
GPIB_bpB            TTDNQWWSFPCTRLQYFVCEFQA---
IXB_TRIFL/111       STNNKWRSRACRMMAQFVCE------
bitiscetinB         SFHNKWIQWTCNRKKSFVCKYRV---
```

Fig. 6. The CTLD toxin clade of Table II. (A) MSA of the proteins comprising the clade. (B) Phylogenetic reconstruction of the clade. The clade is constructed of two sub-clades: the first (green) is composed of α subunits and includes also the β subunit of rhodocein, and the second (blue) is composed of the β subunits. A black arrow indicates the location of the α subunit of IX/Xbp, the 3D structure of which is known.

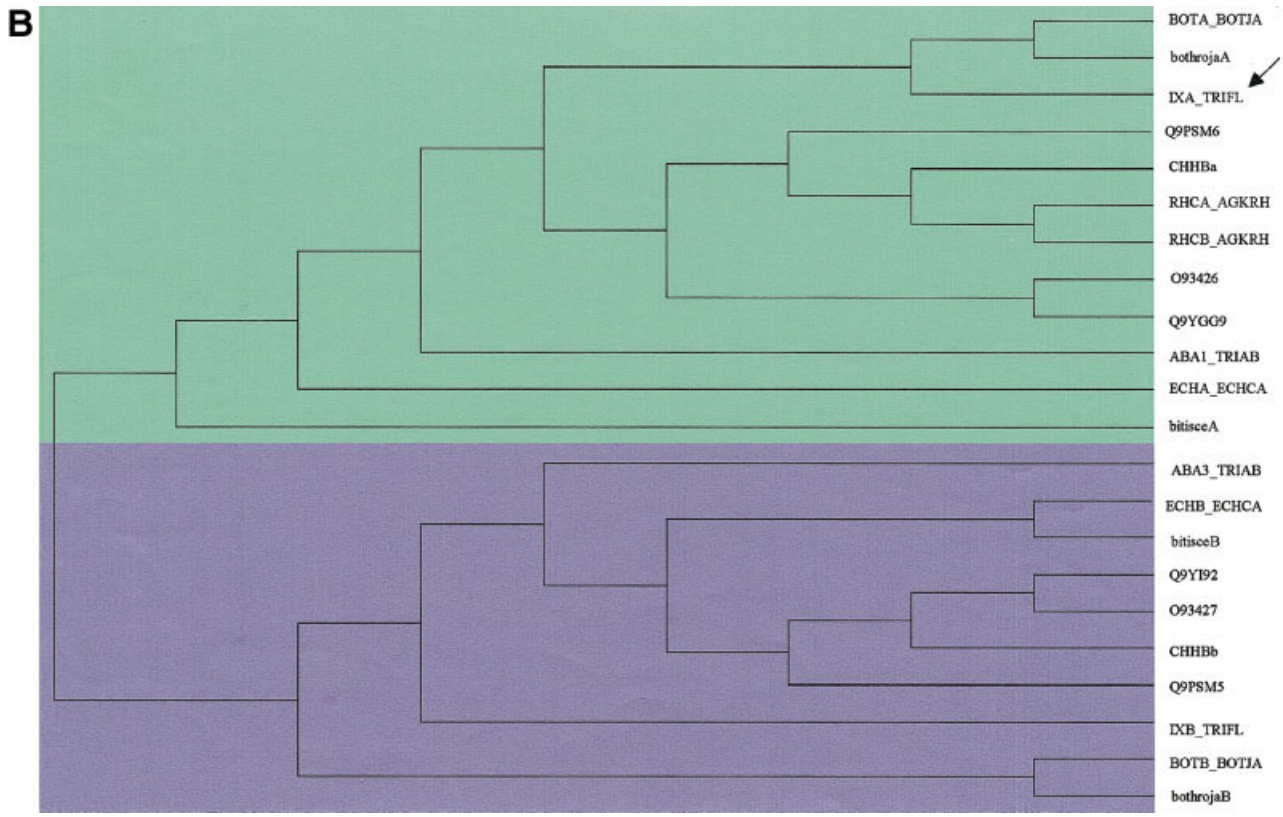

Figure 6. (Continued.)

CONCLUSIONS

We have shown here that the CLT/CTLD superfamily has certain common characteristics that are related to the function of all (or most of) its members and some that are shared only by proteins of a given family. We also found patches of conserved residues to which no function has yet been assigned, such as patch C1 and C2 in the CTL clade (Fig. 5) and patch T1a in the toxin clade (Fig. 8). The latter contains positively charged residues that may be involved in binding to negatively charged targets. Other patches of conserved residues for which no functions have been proposed are discussed in detail in the web site.[22] We want to reiterate that the high conservation grades calculated for the amino acids constructing the patches detected might be due to shortness of divergence time, rather than being indicative of an important function. Site-directed mutation studies are required in order to establish the role of these patches and to characterize them more extensively.

The CTL/CTLD superfamily contains well-studied proteins. Indeed, some of the patches of conserved residues have been documented as functionally important based on experimental data. Interestingly, other patches have not. This suggests that ConSurf can be viewed as a computational tool that enables one to quickly skim through the protein structure for conserved patches. These patches can be further studied using biochemical, biophysical, and molecular biology tools.

ACKNOWLEDGMENTS

This study was supported by a Research Career Development Award from the Israel Cancer Research Fund. We are grateful to the Bioinformatics Unit at the George S. Wise Faculty of Life Sciences at Tel Aviv University for providing technical assistance and computation facilities.

REFERENCES

1. Drickamer K, Taylor ME. Biology of animal lectins. Annu Rev Cell Biol 1993;9:237–264.
2. Hofmann K, Bucher P, Falquet L, Bairoch A. The PROSITE database, its status in 1999. Nucleic Acids Res 1999;27(1):215–219.
3. Drickamer K. C-type lectin-like domains. Curr Opin Struct Biol 1999;9(5):585–590.
4. Dodd RB, Drickamer K. Lectin-like proteins in model organisms: implications for evolution of carbohydrate-binding activity. Glycobiology 2001;11(5):71R–79R.
5. Kogelberg H, Feizi T. New structural insights into lectin-type proteins of the immune system. Curr Opin Struct Biol 2001;11:635–643.
6. Mizuno H, Fujimoto Z, Koizumi M, Kano H, Atoda H, Morita T. Structure of coagulation factors IX/X-binding protein, a heterodimer of C-type lectin domains. Nat Struct Biol 1997;4(6):438–441.
7. Bon C, Leduc M. Cloning of subunits of convulxin, a collagen-like platelet-aggregating protein from *Crotalus durissus terrificus* venom. Biochem J 1998;333(Pt 2):389–393.

8. Morita T, Shin Y. Rhodocytin, a functional novel platelet agonist belonging to the heterodimeric C-type lectin family, induces platelet aggregation independently of glycoprotein Ib. Biochem Biophys Res Commun 1998;245(3):741–745.
9. Armon A, Graur D, Ben-Tal N. ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. J Mol Biol 2001;307(1):447–463.
10. Aloy P, Querol E, Aviles FX, Sternberg MJ. Automated structure-based prediction of functional sites in proteins: applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking. J Mol Biol 2001;311(2):395–408.
11. Landgraf R, Xenarios I, Eisenberg D. Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins. J Mol Biol 2001;307(5):1487–1502.
12. Valdar WS, Thornton JM. Protein-protein interfaces: analysis of amino acid conservation in homodimers. Proteins 2001;42(1):108–124.
13. Lichtarge O, Sowa ME, Evolutionary predictions of binding surfaces and interactions. Curr Opin Struct Biol 2002;12:21–27.
14. Schultz J, Milpetz F, Bork P, Ponting CP. SMART, a simple modular architecture research tool: identification of signaling domains. Proc Natl Acad Sci USA 1998;95(11):5857–5864.
15. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 1994;22(22):4673–4680.
16. Felsenstein J. Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. Methods Enzymol 1996;266:418–427.
17. Miyata T, Miyazawa S, Yasunaga T. Two types of amino acid substitutions in protein evolution. J Mol Evol 1979;12(3):219–236.
18. Nicholls A, Sharp KA, Honig B. Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons. Proteins 1991;11(4):281–296.
19. Sali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 1993;234(3):779–815.
20. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res 1997;25(24):4876–4882.
21. Sitkoff D, Ben-Tal N, Honig B. Calculation of alkane to water solvation free energies using continuum solvent models. J Phys Chem 1996;100:2744–2752.
22. http://ashtoret.tau.ac.il/~sharon
23. Hakansson K, Lim NK, Hoppe HJ, Reid KB. Crystal structure of the trimeric alpha-helical coiled-coil and the three lectin domains of human lung surfactant protein D. Structure Fold Des 1999;7(3):255–264.
24. Kastrup JS, Nielsen BB, Rasmussen H, Holtet TL, Graversen JH, Etzerodt M, Thogersen HC, Larsen IK. Structure of the C-type lectin carbohydrate recognition domain of human tetranectin. Acta Crystallogr D Biol Crystallogr 1998;54(Pt 5):757–766.
25. Sheriff S, Chang CY, Ezekowitz RA. Human mannose-binding protein carbohydrate recognition domain trimerizes through a triple alpha-helical coiled-coil. Nat Struct Biol 1994;1(11):789–794.
26. Sakurai Y, Fujimura Y, Kokubo T, Imamura K, Kawasaki T, Handa M, Suzuki M, Matsui T, Titani K, Yoshioka A. The cDNA cloning and molecular characterization of a snake venom platelet glycoprotein Ib-binding protein, mamushigin, from *Agkistrodon halys blomhoffii* venom. Thromb Haemost 1998;79(6):1199–1207.
27. Matsui T, Kunishima S, Hamako J, Katayama M, Kamiya T, Naoe T, Ozeki Y, Fujimura Y, Titani K. Interaction of von Willebrand factor with the extracellular matrix and glycocalicin under static conditions. J Biochem (Tokyo) 1997;121(2):376–381.
28. Andrews RK, Kroll MH, Ward CM, Rose JW, Scarborough RM, Smith AI, Lopez JA, Berndt MC. Binding of a novel 50-kilodalton alboaggregin from *Trimeresurus albolabris* and related viper venom proteins to the platelet membrane glycoprotein Ib-IX-V complex. Effect on platelet aggregation and glycoprotein Ib-mediated platelet activation. Biochemistry 1996;35(38):12629–12639.
29. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 1987;4(4):406–425.
30. Hamako J, Matsui T, Suzuki M, Ito M, Makita K, Fujimura Y, Ozeki Y, Titani K. Purification and characterization of bitiscetin, a novel von Willebrand factor modulator protein from *Bitis arietans* snake venom. Biochem Biophys Res Commun 1996;226(1):273-279.
31. Arocas V, Castro HC, Zingali RB, Guillin MC, Jandrot-Perrus M, Bon C, Wisner A. Molecular cloning and expression of bothrojaracin, a potent thrombin inhibitor from snake venom. Eur J Biochem 1997;248(2):550–557.