Comparative Analysis of Structural Properties of the C-Type-Lectin-like Domain (CTLD)

Alex N. Zelensky and Jill E. Gready*

Computational Proteomics and Therapy Design Group, John Curtin School of Medical Research, Australian National University, Canberra, Australia

ABSTRACT **The superfamily of proteins containing the C-type-lectin-like domain (CTLD) is a group of abundant extracellular metazoan proteins characterized by evolutionary flexibility and functional versatility. Several CTLDs are also found in parasitic prokaryotes and viruses. The 37 distinct currently available CTLD structures demonstrate significant structural conservation despite low or undetectable sequence similarity. Our aim in this study was to perform an extensive comparative analysis of all available CTLD structures to establish the most conserved structural features of the fold, and to test and extend the early analysis of Drickamer. By implication, these features should be those critical for maintenance of integrity of the fold. By analyzing CTLD structures superimposed by several methods, we have established groups of conserved structural positions involved in fold maintenance but not in ligand binding; these are consistent with the fold's known functional flexibility. In addition to the well-recognized disulfide bridges, groups of conserved residues are involved in hydrophobic interactions stabilizing the core of the fold and the long loop region, and in an α2-β1–β5 polar interaction. Evaluation of the conclusions of the structure comparison study compared with alignments of all available human, mouse and *Caenorhabditis elegans* CTLD sequences showed that conservation patterns are preserved throughout the whole CTLD sequence space. Our observations provide an improved understanding of CTLD structure, and will help in identification of new CTLDs and the mechanisms that drive and constrain the coevolution of the structure and function of the fold. Proteins 2003;52:466–477.** **© 2003 Wiley-Liss, Inc.**

Key words: C-type lectin; fold evolution; hydrophobic core; structurally conserved residues; domain stability

INTRODUCTION

C-type-lectin-like domains (CTLDs) were first identified as 110–140-residue-long carbohydrate recognition domains (CRDs) of group C of animal lectins, which bound carbohydrates in a Ca-dependent manner.[1] By comparing sequences of C-type lectins, a set of conserved positions was identified, which included residues involved in Ca- and carbohydrate-binding, and fold integrity maintenance.[1,2] A distinct sequence signature defined by this set of residues allowed identification of many new CTLDs by sequence analysis, whereas other examples of the CTLD fold were detected only when structures were solved. Although the CTLD was first identified as a carbohydrate-binding domain, only about half of currently characterized CTLDs have been shown to bind carbohydrates in a Ca-dependent manner.

Here, we use the term "C-type-lectin-like domain," as defined by Drickamer,[3] to describe a superfamily of protein domains that have a structure similar to that of rat mannose-binding protein A (MBP-A), the fold representative whose structure was solved first.[4,5] Structurally, we can divide CTLDs into two groups: *canonical* CTLDs having a long loop region, and *compact* CTLDs that lack it. The second group includes link or protein tandem repeat (PTR) domains[6,7] and bacterial CTLDs.[8–10] It is not clear whether the CTLD superfamily is monophyletic, because the evolutionary relationship of the bacterial CTLDs to the animal C-type lectin family is uncertain.[8] These domains could either have been acquired by horizontal transfer or could have arisen by convergent evolution as mimicry of host proteins. Another family usually included in the CTLD superfamily is that of endostatin.[3,11,12] However, in the comparative structure analysis studies presented below, no substantial evidence of CTLD and endostatin fold similarity was found. Sequence similarity between endostatin and CTLDs is also absent. Therefore, in this article, the endostatin fold was not considered an example of a CTLD and was excluded from the analysis.

The CTLD fold (Fig. 1) characteristically has two antiparallel β-sheets and two α-helices. The β2 strand divides the structure into two lobes, the upper and the lower. The vertical β-sheet, formed by strands β1, β5, and β1′, and both helices (α1 and α2) form the lower lobe of the structure, whereas the upper lobe consists of the long loop region (LLR) (about 40 residues) and the second antiparallel β-sheet formed by strands β2, β3, and β4. The LLR is involved in Ca-mediated carbohydrate binding and in domain-swapping dimerization of some CTLDs,[13–16] which occurs via a unique mechanism.[17] Although the LLR is

*Correspondence to: Jill E. Gready, Computational Proteomics and Therapy Design Group, John Curtin School of Medical Research, Australian National University, P.O. Box 334, Canberra ACT 2601, Australia. E-mail: jill.gready@anu.edu.au

Received 1 November 2002; Accepted 21 January 2003

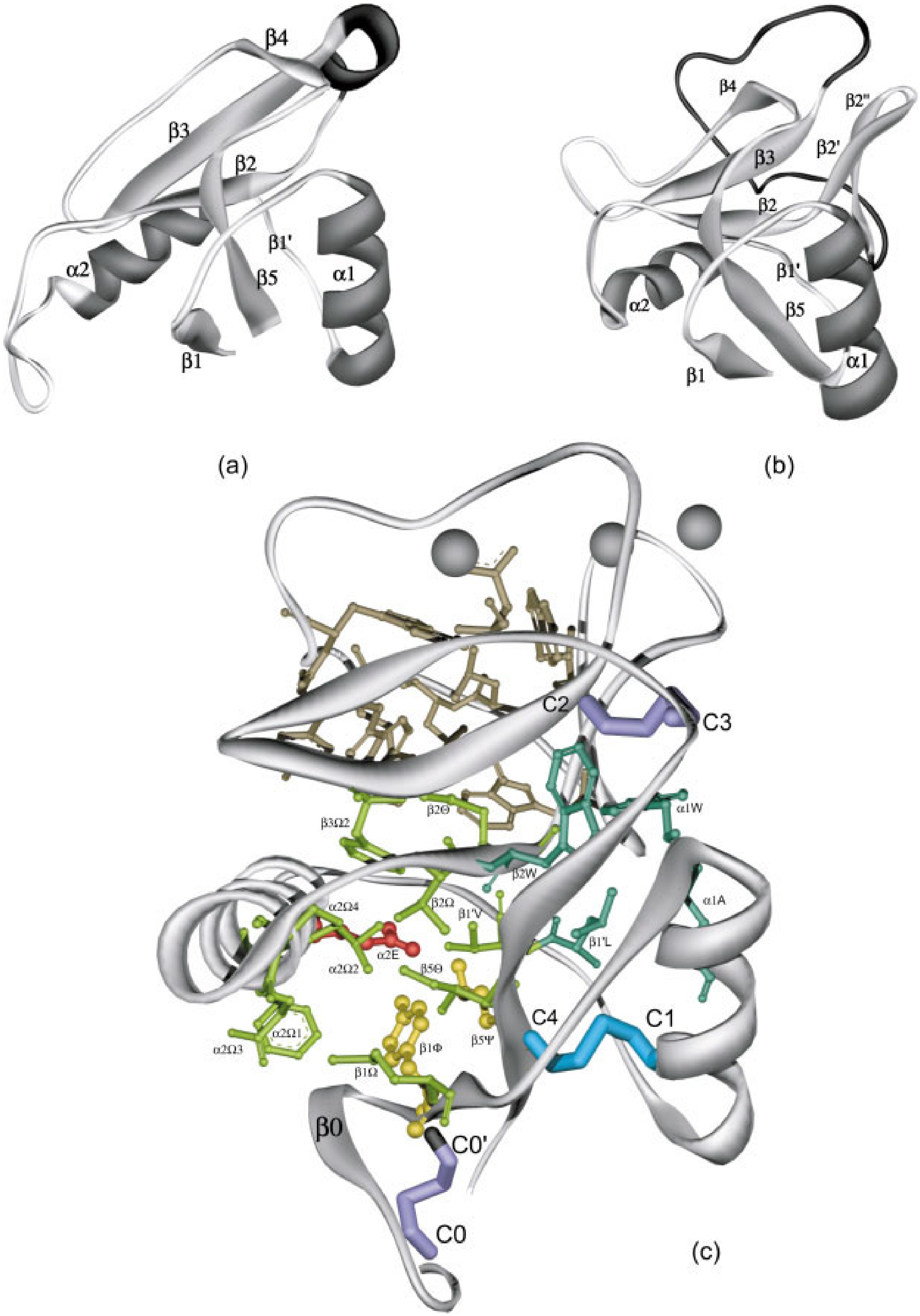

Fig. 1. Ribbon diagrams of (**a**) compact (intimin, 1f00) and (**b**) canonical (E-selectin, 1g1t) CTLD structures. The long loop region in E-selectin, and the short α-helix that replaces the LLR in compact CTLDs, are shown in black. Secondary structure labels are according to the secondary structure element (SSE) naming scheme used in this study. An example of a long-form CTLD (DC-SIGN, 1k9i) is shown in (**c**). Residues discussed in the text are displayed in ball-and-stick, thick-wire, or thick-tube mode and colored as in the alignment in Figure 2(a).

absent in PTR/link domains, the fold still binds saccharides (hyaluronan) by an alternative Ca-independent mechanism. The overall domain has its C- and N-termini coming close together at the bottom of the structure. This end–beginning interaction is stabilized by an α1–β5 disulfide bridge, which is the most conserved feature of the domain. Another highly conserved disulfide bridge links the LLR to the β4–β5 turn.

CTLDs selectively bind a wide variety of ligands. As the superfamily name suggests, carbohydrates (in various contexts) are primary ligands for CTLDs, and the binding is Ca-dependent.[18] However, the fold has been shown specifically to bind proteins,[19] lipids,[20] and inorganic compounds, including $CaCO_3$ and ice.[21–24] In several cases, the domain is multivalent and may bind both protein and sugar.[25–27]

Our purpose in this study was to extend knowledge of the properties of the CTLD fold by comparing all CTLD structures currently available in the Protein Data Bank (PDB)[28] to identify the most conserved positions in the fold. The study was structure-focused, because the CTLDs compared have a variety of biological origins and perform different functions. Hence, our interest in ligand binding was limited to constraints that function imposes on the structure and on its variation. Thus, it is expected that conservation patterns found indicate the most general principles of CTLD fold organization. It has been demonstrated by similar studies carried out for SH3,[29] globin,[30] immunoglobulin,[31] legume lectin[32] folds, the alkaline phosphatase superfamily,[33] and the chymotrypsin family of serine proteases,[34] that comparison of many homologous structures combined with sequence analysis can reveal much more definitive information about fold organization than analysis of individual structures. Though limited structural comparisons of the CTLD fold have been reported previously,[3,35,36] they included only the small number of CTLD structures available then, none of which was beyond the twilight zone (<30% identity)[37] of sequence homology. It is particularly interesting to compare the most deviant examples of the CTLD fold, such as bacterial CTLDs, with the canonical ones, because features common to these two groups were not revealed by sequence alignment.[8–10] Indeed, we find that neither simple pairwise nor more sophisticated profile-based alignments can correctly identify pairs of structurally equivalent residues in such evolutionarily remote (or unrelated) CTLD species as eukaryotic and bacterial CTLDs.

Careful comparison of CTLD structures also allows building of a common scheme for residue and secondary structure element numbering. This was done for other abundant domains, such as the Ig-like domain[38] or SH3 domain,[29] and is an important tool for describing domain variation.

Structural alignment, though hard to define unambiguously,[39] is considered to be a gold standard for assessing protein similarity. Recent comparison of the quality of sequence versus structure alignments demonstrated that the latter detect homologous sites more accurately than the former.[40] In addition to discovery of important structural features of the fold, structure alignment allows creation of a structure-verified sequence alignment, which can be used to create valid position-specific scoring matrices or hidden Markov model (HMM) profiles for the domain.

METHODS

Structure Alignment

Pairwise structure superpositions were generated using the Combinatorial Extension (CE) algorithm,[41] which is available from the San Diego Supercomputer Center's website (ftp://ftp.sdsc.edu/pub/sdsc/biology/CE/src/), with the hierarchical protein structure superposition program LOCK,[42] kindly provided by the authors, and with the DALI program[43] (http://www.ebi.ac.uk/dali/). We used pairs of structurally equivalent residues as established by the structure alignment algorithm or, in order to overcome some discrepancies in structural correspondence assignments that we noticed, generated alignments from the superimposed structures by a dynamic programming algorithm in the following manner: A matrix of inter-residue distances between two superimposed structures was used to generate a position-specific substitution matrix, with distances below a selected threshold, producing an arbitrary positive score (+5), or negative score (−5). We used a fixed-gap penalty of −1 to calculate gap costs.

The alignment description formalism used below was adopted from Godzik at al.[44] Position i in sequence A was considered structurally conserved in a given set of structures if it satisfied two simple criteria:

1. Alignment at this position was consistent regardless of which of the two structures was used as a query, that is, $i = BA[AB(i)]$ is true for all possible B's, where $AB(i)$ is the position in structure B, which corresponds to position i in structure A in a pairwise structural alignment of query A to target B.
2. Alignment at this position is stable, that is, i is aligned to a nongap in more than a specific percentage of possible alignments that involve A.

Additional distance thresholds (3–5 Å) were applied to the definition of a conserved position as described later in this article.

Hydrophobicity and Solvent Accessibility

A hydrophobic core was defined as a cluster of at least four residues, sidechains of which were not more than 10% solvent exposed, and with each sidechain having at least one atom within a distance of 5 Å from at least one other sidechain of a residue belonging to the same cluster. Similar parameters and algorithm were used in one of the previous studies.[32] Although parameters for the hydrophobic core (HFC) detection algorithm have not been comprehensively tested, we attempted to optimize them to avoid obvious false positives and false negatives. Solvent-accessible area was determined with the surfv program,[45] which was downloaded from http://trantor.bioc.columbia.edu/programs/surf.html.

Sequence Entropy

For each alignment position containing less than 20% gaps, Shannon entropy[46] was calculated according to the formula

$$H(X) = -\sum_{i=A}^{Y} P(x_i)\ln P(x_i), \quad (1)$$

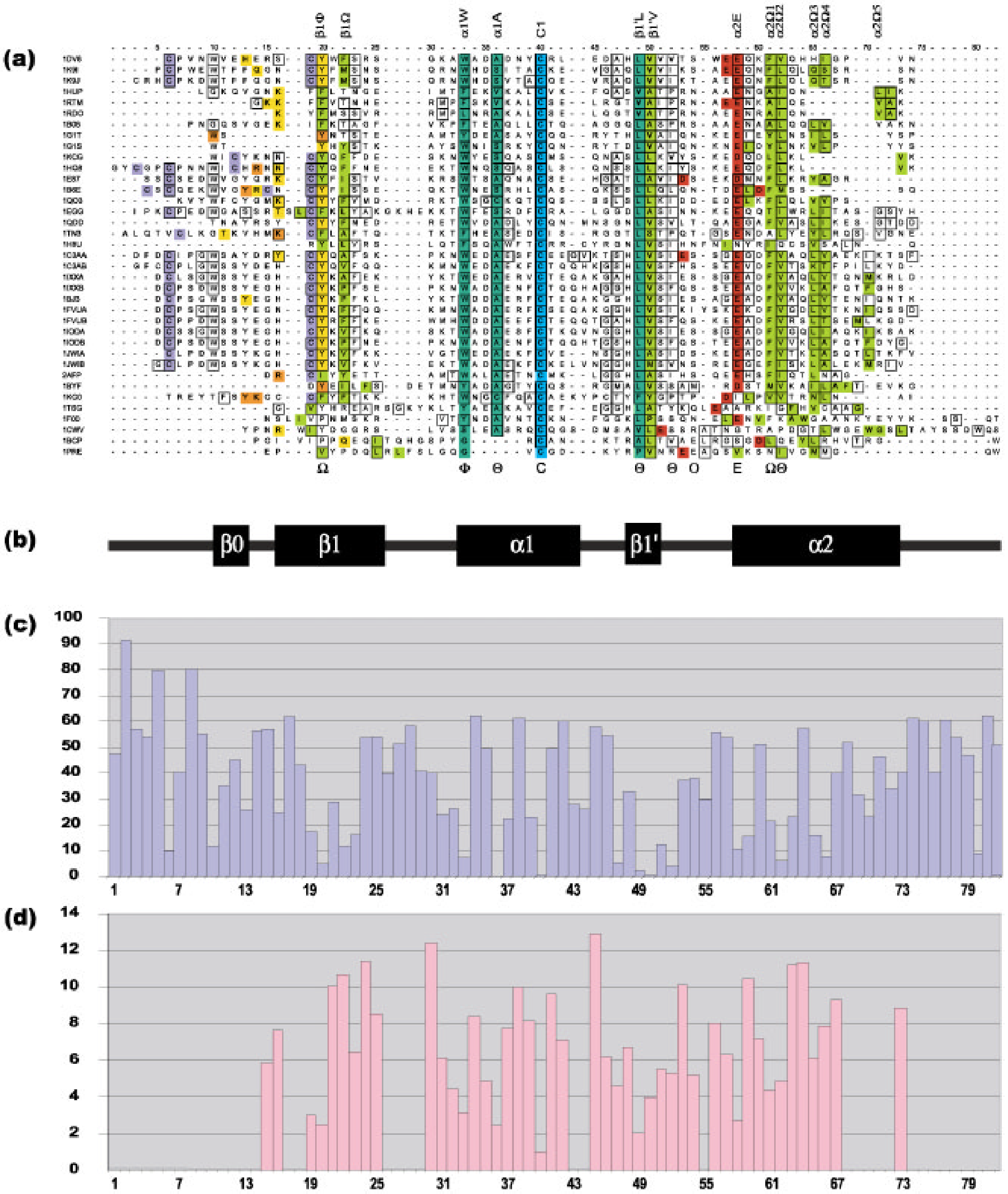

Fig. 2. (**a**) Structure-verified alignment of the sequences of 37 CTLDs, whose structures have been solved (order as in Table I). Sequences were first aligned with ClustalW, followed by manual correction of misaligned structurally equivalent positions. Corrections were not made when structural feature was conserved but transferred to a different position in the structure. Residues are colored according to their presumed role in fold maintenance. Absolutely conserved Cys are in blue. Cys specific for canonical CTLDs are lavender. Glu at α2 N-terminus is red, and its partners in β1 and β5 are yellow or orange, depending on whether the interaction is observed or presumed, respectively. Primary, small, and LLR hydrophobic core residues are light-green, green, and brown, respectively. Conserved positions are labeled above the alignment. Letters below the alignment indicate residues identified as the sequence motif characteristic of the CTLD by Drickamer;[2] red letters are for Ca- and carbohydrate-binding residues. Boxed residues are those included into hydrophobic cores by the computer algorithm described in the Methods section. (**b**) Consensus diagram of secondary structure element positions. (**c**) Average sidechain solvent accessibility of the aligned residues. (**d**) Positional entropy plot. Entropy was calculated only for positions containing less than 20% gaps.

where i is each of 20 amino acids (A,C,D…Y) and $P(x_i)$ is the frequency of residue i in position x. Positional entropy N is plotted in Figure 2 in exponential form:

$$N(X) = e^{H(X)} \qquad (2)$$

Sequence Analysis

We built a database of the CTLD-containing protein sequences by searching the GenPept[47] database with PSI-BLAST[48] (up to 20 iterations, default settings), using sequences of CTLD domains whose structures were solved

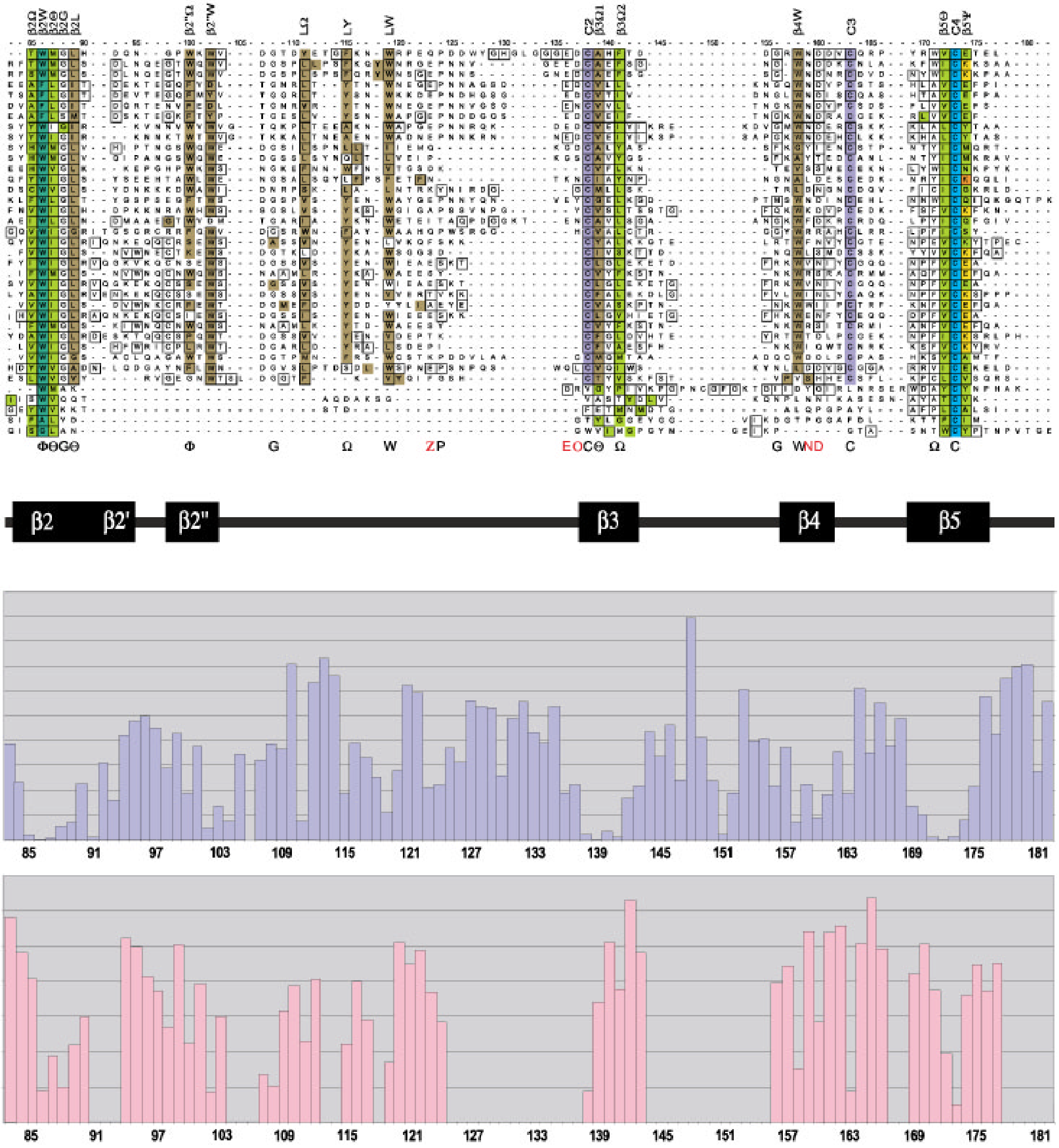

Figure 2. (Continued.)

as seed queries. The resulting database (*cdb*) contained about 1300 sequences. Nonredundant (95%) derivatives of the database were created (*cdb95*) with the *cd-hi* sequence-clustering algorithm.[49] To build a structure-verified sequence alignment, we aligned in Clustal W sequences of CTLDs used in the structure comparison analysis.[50] The alignment was then corrected manually according to the structure comparison results. We built a HMM profile on the basis of the alignment with the *hmmbuild* program from the HMMER package.[51] This profile was used for identification of CTLD boundaries within *cdb95* sequences. Regions corresponding to CTLDs were extracted from the sequences (*cdb95ctld*) and aligned with the *hmmalign*[51] or Clustal W programs.

RESULTS AND DISCUSSION

We built a collection of CTLD structures (Table I) available in the Protein Data Bank (PDB) by searching the PDB sequence database with PSI-BLAST[48] and HMMER,[51] and by analyzing automated (FSSP[52]) and manual (SCOP[11]) classifications of protein structures. This approach allowed us to create a comprehensive list of 92 known CTLD structures, including theoretical models: 1afa, 1afb, 1afd, 1b08, 1b6e, 1bch, 1bcj, 1bcp, 1bj3, 1buu,

TABLE I. Representative Structures of C-Type-Lectin-Like Domains Used in the Analysis

Name	PDB ID	Chain	CTLD residues	Origin
Asialoglycoprotein receptor (ASGR)	1dv8	A	153–280	*Homo sapiens*
DC-SIGN	1k9i	A	285–382	*Homo sapiens*
DC-SIGNR	1k9j	A	265–394	*Homo sapiens*
Mannose-binding protein (MBP)	1hup		110–228	*Homo sapiens*
Mannose-binding protein A (MBP-A)	1rtm	1	108–221	*Rattus norvegicus*
Mannose-binding protein C (MBP-C)	1rdo	1	115–225	*Rattus norvegicus*
Pulmonary surfactant protein D (PSP-D)	1b08	A	235–355	*Homo sapiens*
E-selectin	1g1t	A	1–121	*Homo sapiens*
P-selectin	1g1s	A	1–121	*Homo sapiens*
NKG-2D	1kcg	A	103–215	*Homo sapiens*
NKG-2D	1hq8	A	110–232	*Mus musclulus*
CD69	1e87		83–199	*Homo sapiens*
CD94	1b6e		59–179	*Homo sapiens*
Ly49A	1qo3	C	140–258	*Mus musclulus*
Macrophage mannose receptor (MMR)	1egg	B	625–768	*Homo sapiens*
Lithostathine	1qdd	A	18–144	*Homo sapiens*
Tetranectin	1tn3		45–181	*Homo sapiens*
Eosinophil major basic protein (EMBP)	1h8u	A	3–117	*Homo sapiens*
Flavocetin-A	1c3aA	A	1–135	*Trimeresurus flavoviridis*
Flavocetin-A	1c3aB	B	201–325	*Trimeresurus flavoviridis*
IX/X-binding snake venom protein	1ixxA	A	1–129	*Trimeresurus flavoviridis*
IX/X-binding snake venom protein	1ixxB	B	1–123	*Trimeresurus flavoviridis*
IX-binding snake venom protein	1bj3	A	1–129	*Trimeresurus flavoviridis*
Botrocetin	1fvuA	A	1–133	*Bothropos jararaca*
Botrocetin	1fvuB	B	401–525	*Bothropos jararaca*
Coagulation factor X-binding protein	1iodA	A	1–129	*Deinagkistrodon acutus*
Coagulation factor X-binding protein	1iodB	B	201–323	*Deinagkistrodon acutus*
Bitiscetin	1jwiA	A	4–127	*Bitis arientans*
Bitiscetin	1jwiB	B	3–125	*Bitis arientans*
Fish antifreeze protein	2afp	A	16–129	*Hemitripterus americanus*
Polyandrocarpa lectin	1byf	A	2–124	*Polyandrocarpa misakiensis*
GP42	1kg0	C	104–221	Epstein–Barr virus
TSG-6 link module	1tsg		1–98	*Homo sapiens*
Intimin	1f00	I	840–939	*Escherichia coli*
Invasin	1cwv	A	885–986	*Yersinia pseudotuberculosis*
Pertussis toxin	1bcp	B	3–89	*Bordetella pertussis*
Proaerolysin	1pre	A	2–85	*Aeromonas hydrophila*

1bv4, 1byf, 1c3a, 1cwv, 1dv8, 1e5u, 1e87, 1e8i, 1egg, 1egi, 1esl, 1f00, 1f02, 1fif, 1fih, 1fm5, 1fvu, 1g1q, 1g1r, 1g1s, 1g1t, 1gie, 1h8u, 1hli, 1hlj, 1hq8, 1htn, 1hup, 1hyr, 1ijk, 1iod, 1ixx, 1ja3, 1jsk, 1jwi, 1k9i, 1k9j, 1kcg, 1kg0, 1kja, 1kjb, 1kjd, 1kje, 1kmb, 1kwt, 1kwu, 1kwv, 1kww, 1kwx, 1kwy, 1kwz, 1kx0, 1kx1, 1kza, 1kzb, 1kzc, 1kzd, 1kze, 1lit, 1msb, 1pre, 1prt, 1pto, 1qdd, 1qo3, 1rdi, 1rdj, 1rdk, 1rdl, 1rdm, 1rdn, 1rdo, 1rtm, 1tlg, 1tn3, 1tsg, 1ytt, 2afp, 2kmb, 2msb, 3kmb, 4kmb. Each of the structures found has already been attributed to the CTLD superfamily either in the SCOP classification or in the publication describing the structure. For structures solved more than once in mutant or variously complexed forms, the PDB entry for wild-type with the highest resolution was used: this yielded the 37 structures in Table I. Theoretical models were excluded from the analysis. We employed two approaches for structure analysis. The first involved detection of conserved positions and interactions in CTLDs by manual comparison of the CTLD structures. The second was to produce structural alignments with several computer programs (DALI,[43] LOCK,[42] and CE[41]).

Manual Structure Comparison

Secondary structure

There is no general agreement on the numbering of CTLD secondary structure elements (SSEs) in the literature. The first numbering scheme was proposed by Weis et al.[4] on the basis of the MBP-A structure and included 5 strands, 2 helices, and 4 loops. However, this description appeared to be insufficient, because MBP lacks some SSEs that are present in long-form (defined below) CTLD structures, whereas other small strands were not defined. Other studies describing structures of CTLDs that have a different number of SSEs than MBP-A either introduced their own SSE numbering [β strands 1–6 in asialoglycoprotein receptor (ASGR);[53] 6 β strands in link module, with labeling not consistent with ASGR or MBP-A[6]; β1–β7 in NKG2D;[54] β1–β8 in EMBP[55]] or extended the SSE naming scheme used for MBP-A (Ly49A SSE numbering is consistent with that in MBP-A[56]). The latter approach seems to be more universal, because only the most conserved SSEs get individual numbers, whereas others have derived names/numbers. Therefore, we have adopted the MBP-A SSE numbering scheme[4] with the following extensions (Fig. 1): the β strand specific for the long-form CTLD is labeled β0; the short β strand between α1 and α2 is labeled β1′; and the two β strands forming a hairpin C-terminal to β2 are labeled β2′ and β2″.

Nomenclature

Here, we use naming conventions to discuss structurally important positions. A residue identifier is built from three fields: (1) the SSE to which the position belongs; (2) the most frequent residue in the position, or the group name symbol if a group of residues occupies the position with similar frequencies (Θ for aliphatic, Φ for aromatic, Ω for aliphatic or aromatic, Ψ for charged); and (3) a sequential number if combination of fields (1) and (2) does not give a unique identifier. For example, the second of the two conserved hydrophobic positions in β3 is named β3Ω2. The four highly conserved Cys residues are labeled C1–C4. If a reference to a residue number is required to define a position unambiguously, the rat MBP-A (PDB code: 1rtm) structure is used as the template. This is for historical reasons, because the MBP-A structure was the first CTLD structure solved and has often been used as a reference structure in the literature. It is not, however, the best consensus structure from the set, because in some conserved positions it has residues that rarely occur in the same position in most other structures.

Groups of identified structurally equivalent residues are listed in order of decreasing conservation:

- Absolutely conserved Cys residues—“cystine staple” joining the N- and C-termini of the structure (C1, C4).
- A highly conserved pair of consecutive aliphatic residues in β1′ defining the primary and the small hydrophobic cores, and other highly conserved residues that form the primary hydrophobic core.
- In the small hydrophobic core, a pair of orthogonal Trp residues forming a hydrophobic interaction between α1 and β2 (α1W and β2W) and two aliphatic residues.
- A highly conserved Glu in α2.
- A highly conserved Cys pair in the base of the β3–β4 hairpin (C2,C3).
- Residues comprising the hydrophobic core of canonical CTLDs.

Each group is discussed in more detail below [Fig. 2(a,b)].

Primary hydrophobic core (PHC)

The PHC of the CTLD fold is formed by interacting residues from five distinct regions of the structure and can be subdivided into five subgroups on this basis. Two positions at the C-terminus of the β1′ strand (β1′L and β1′V; V135 and A136 in MBP-A) are the most conserved structurally and tightly interact with β5. The direction of the sidechain of the first residue of the pair, which is highly conserved in all structures, is parallel to β5 and is a part of the small hydrophobic core. The second position is occupied by a less conserved residue, the sidechain of which projects into the middle of the PHC. There is only one deviation from hydrophobicity of this β1′ pair: in tetranectin the second position is occupied by Ser. At least one β1 residue (β1Φ; F112) is always involved in PHC formation by interacting with hydrophobic residues in α2, whereas many structures have another position C-terminal to β1Φ (β1Ω; T114), which is also involved in PHC formation. In canonical CTLDs, β1Φ is occupied by either Tyr or Phe; in compact CTLDs, the aromatic residue is replaced by an aliphatic one. It does not seem possible to identify structurally equivalent conserved positions in α2 from different structures, although the general conservation pattern is obvious. Three to five hydrophobic residues (α2Ω1–α2Ω5) on the inner side of the amphiphilic α2 helix toward its C terminus interact with both the upper (β2, β3) and lower (β1) part of the PHC. Only the second of the two conserved positions in β2 (β2Ω and β2Θ; A155 and L157) always contains a hydrophobic residue (either aliphatic or methionine), and this “stacks” over the β1′L-β1′V pair. The β2Ω position is in most cases either aromatic or aliphatic, but very weakly conserved, and can also be occupied by a polar or charged residue. At the same time, it is always buried by solvent (~5% average exposure), projects toward the PHC center, and is recognized as part of the hydrophobic core by our algorithm. This promiscuity of β2Ω can be partially explained by its dual structural role, because, apart from PHC formation, it may be involved in stabilizing the loop joining α2 to β2. For example, in 1g1t (E-selectin) β2Ω is occupied by Y49, which forms a hydrogen bond with the S46 backbone carbonyl. A similar interaction is observed in 1kcg and 1hq8 (human and mouse NKG2D), where β2Ω histidine is H-bonded to the backbone carbonyl of the position −3 relative to it. β3 also contains a pair of conserved residues (β3Ω1, β3Ω2; V196, I198), usually aliphatic; the orientation of their sidechains is quite similar to that of the corresponding pair in β2 (β2Θ and β2Ω, respectively). β3Ω2 is involved in PHC formation, whereas β3Ω1 in canonical CTLDs is involved in formation of the LLR hydrophobic core but projects into the PHC in compact CTLDs. In some cases in which the second of the two PHC positions contributed by β5 (β5Θ, β5Ψ; V216, E218) is occupied by a large charged or polar residue (Glu, Lys, Tyr), it may play a dual role, with its apolar portion being a part of the PHC and the charged portion interacting with α2E. Where β5Ψ is E, this interaction is mediated by a Ca^{2+} ion.

Small hydrophobic core

This hydrophobic core is formed by two bulky aromatic residues, one in the initial part of α1 (α1W; F121), and the other in β2 (β2W, F156); their sidechain planes are orthogonal to each other. The small hydrophobic core also includes two positions occupied by aliphatic residues: β1′L and a residue in the +3 position to α1W (α1A; V124). In MBP-A, the α1W and β2W residues are Phe, whereas in most other structures they are Trp. The residue type and orientation in the α1W and β2W positions are highly conserved in canonical CTLDs, but more variability exists in several compact structures both in relative residue orientation (in the link module of TSG-6, the sidechain planes are almost parallel) and composition (α1W mutated to Y or S). Two similar structures—proaerolysin and pertussis toxin—both lack the α1 helix, which is replaced by a loop, and the small hydrophobic core, having Gly or Ala in the α1W and β2W positions. In the pertussis toxin structure, the hydrophobic interaction between β2 and α1 is mediated by F78, which is in the loop between β4 and β5. In proaerolysin, this mediator is absent.

Long loop region hydrophobic core (LLRHC)

The LLRHC is present only in canonical CTLDs and is formed by residues that come from the LLR itself (β2″Ω, β2″W, LΩ, LY, LW) and from 3 β strands: β2 (β2L; L159), β3 (β3Ω1; V196), and β4 (β4W; W204). It is interesting to note that snake venom CTLDs that dimerize by swapping the LLR have most of the LLRHC residues conserved. There is no clear spatial margin between the LLRHC and PHC, but conservation of the latter and absence of the former in compact CTLDs indicates that these two hydrophobic clusters are independent.

We also performed an automated detection of clusters of hydrophobic residues (see Methods section) to confirm that our definition of hydrophobic cores is correct. As can be seen in Figure 2(a,c), most of the positions with conserved low-solvent exposure are covered by the hydrophobic cores described.

Residues from the so-called “WIGL” motif (positions β2W, β2Θ and β2L), which is extremely well conserved in canonical CTLD sequences,[2] are involved in formation of all three hydrophobic cores described. Therefore, the WIGL motif can be considered an integrating component of the structure, which unites clusters of hydrophobic residues that stabilize all parts of the structures. The role of the almost completely conserved Gly residue in WIGL (β2G) is less clear. Dihedral angles for the β2G residue in canonical CTLD structures lie within the “allowed” area of the Ramachandran plot [Phi (−90,−60), Psi (−30,75)], which rules out unusual backbone conformation as a reason for the Gly conservation. Although there is some limitation on the size of the sidechain in this position, simple modeling shows that the sidechains of Ala or Ser would fit without clashing with the backbone or structurally adjacent regions. The fact that it is not conserved in compact CTLDs suggests that it is required for LLR formation, possibly by maintaining the β2-β2′ turn.

Highly conserved Glu residue

A highly conserved Glu residue (α2E; E143) in the beginning of α2 (within 1–3 N-terminal residues) forms in most cases an ionic and/or hydrogen bond with a less conserved partner(s) in β1 (β1Φ and other N-terminal positions) and/or β5 (β5Ψ). Only in EMBP (1h8u) is this position not occupied by a Glu; this may be due to the inherent high basicity of the protein. The importance of the interaction maintained by α2E is emphasized by the manner in which it is conserved in bacterial CTLDs. In invasin, the α2–β1 interaction is mediated by a Glu that is N-terminal to the usual position. Pertussis toxin has an Asp in the + 2 position, which forms an interaction with a β1 partner, similar to canonical CTLDs. Although both proaerolysin and intimin have Glu in appropriate positions and orientations, these residues do not mediate an interaction with β1 or β5 in the X-ray structures. In about half of the structures, a second Glu residue is present upstream to α2E and in a few cases (1dv8, 1k9i, 1rtm) forms an interaction with β1/β0 analogous to α2E. However, this position is much less conserved both conformationally and in the sequence, and in the majority of cases the Glu occupying it just protrudes into the solvent.

β3–β4 disulfide bridge (C2–C3)

Previous descriptions of residues conserved in CTLDs[2,4] usually have included two pairs of cysteines. It can be seen from the alignment that although the outer Cys pair (α1–β5 bridge) is indeed one of the most conserved CTLD features, the inner pair is only present together with the LLR, which suggests that its role is LLR stabilization. Moreover, many examples of both mammalian and *Caenorhabditis elegans* canonical CTLD sequences lack one or both β3 and β4 Cys residues, suggesting that their presence is not essential.

Comparison with previous findings

Results of this manual structure comparison of 37 currently available nonredundant CTLD structures reproduce many of the findings of other authors on the basis of analysis of individual or small groups of CTLDs[2,4,35,36] and the CTLD profile deposited in the curated pattern-based protein domain and family database PROSITE (accession number: PDOC00537),[57] but provide a more general view of the residue positions that determine the CTLD fold. For example, our definition of hydrophobic cores is different from that introduced by Weis et al. in the first article describing the CTLD structure.[4] In particular, their large hydrophobic core[4] is a combination of PHC and LLRHC defined by us. We consider such a separation valid and important, because absence of the LLRHC does not significantly affect the PHC structure, as may be seen in the analysis of the compact CTLD structures. Also, some of the conserved positions we identified were not described in the previous studies (e.g., β1Ω, β1′V, β2Ω, β2″W, β5Θ, β5Ψ). Another set of differences is due to the fact that sequence signatures often used for CTLD analysis, such as the one presented by Drickamer,[2] include positions involved in Ca- and carbohydrate-binding. Because these

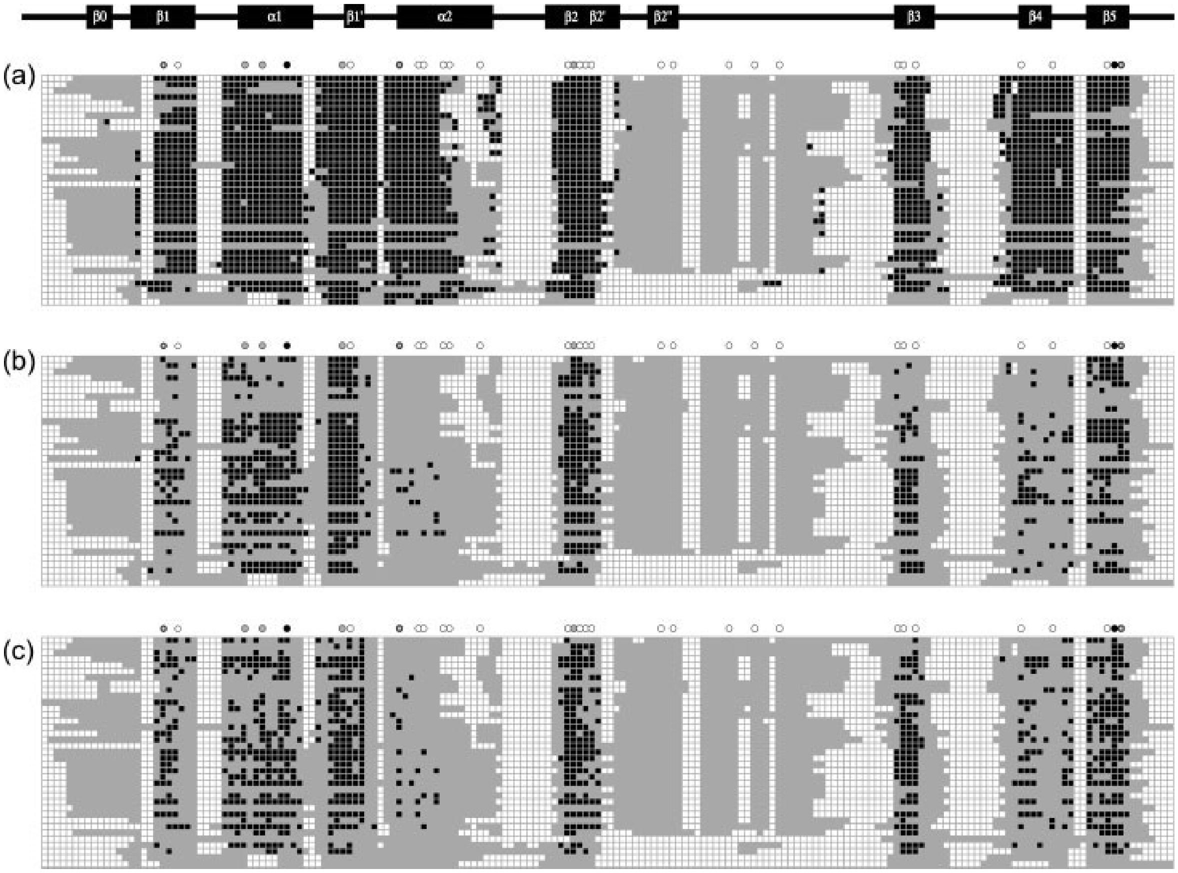

Fig. 3. Structurally conserved positions in CTLDs superimposed by the CE structure alignment program. Sequence alignment is the same as in Figure 2. Nongap conserved and unconserved positions are shown as black and gray squares, respectively, and gaps as white squares. (**a**) Conserved pairs as defined by CE alignment. Residues that satisfy alignment consistency and stability criteria in more than 90% of the structures are displayed (see Methods section for details). No distance threshold. Circles above the alignment indicate conserved positions shown in Figure 2(a). Filled circles indicate C1 and C4, partly filled circles indicate α2E and its partners, gray circles indicate small hydrophobic core residues, and empty circles indicate PHC and LLRHC residues. (**b**) Same as (a), but a 3-Å pairwise distance threshold between Cα atoms was applied. (**c**) Positions conserved in alignments built with a dynamic programming algorithm (described in the Methods section) with the use of distances between sidechain geometric centers as a measure of similarity. Distance cutoff is 3 Å, 95% conservation threshold.

functions are not general for CTLDs, structure comparison fails to detect corresponding positions as conserved.

Automated Structure Alignment

To test whether automated structural alignment methods are able to detect similarities found by manual structure comparison, we applied three different methods (CE,[41] DALI,[43] and LOCK[42]) of structure alignment to determine structurally corresponding pairs of residues. We performed alignments for all structures against all others, followed by analysis of the residue-pairing persistence. A residue in a given structure was considered structurally conserved if it satisfied criteria of consistency (aligned to the same position in another structure regardless of which of the two structures was used as a template) and stability (included in more than 90% of all possible alignments for the structure). We used these two simple constraints to eliminate sporadic matches. As expected, and as demonstrated in Figure 3(a) for the CE method, structure alignment algorithms produce consistent and stable superpositions of canonical CTLDs but noticeably vary in aligning compact and canonical CTLDs. LOCK and DALI produced similar results (not shown). It is also clear that a high level of structural similarity among the available canonical CTLDs prevents discrimination of truly conserved positions from the rest. Nonconserved positions cluster mostly in the LLR, which can be explained by the difference in LLR structure in snake domain–swapped CTLD dimers and mammalian CTLDs. Of the bacterial CTLDs, pertussis toxin and proaerolysin appeared to be the most difficult targets for automated structure alignment.

Analysis of the residue pairs equivalenced by the structure alignment programs suggested two ways to improve automated detection of conserved residue pairs. The first was to introduce a minimal distance threshold in the range of 3–5 Å for inclusion of residue pairs in the alignment, whereas the second was to use points other than Cα centers to measure pairwise distances between residues. The latter correction was inspired by our observation that many of the equivalenced positions had different sidechain orientation, whereas positions involved in fold integrity maintenance did not, and the expectation that functionally

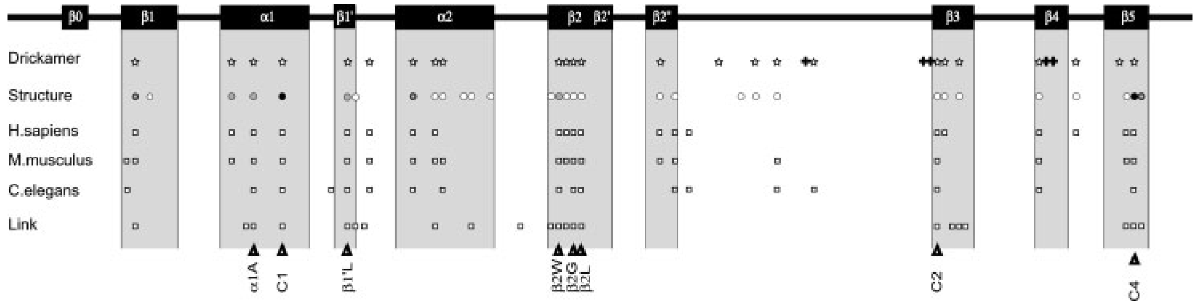

Fig. 4. The most conserved positions in different CTLD groups revealed by sequence alignments. Consensus secondary structure diagram is shown at the top. Positions of conserved residues relative to secondary structure elements are shown in the table below. (**Row 1**) CTLD sequence signature from Drickamer;[2] + indicates residues involved in Ca- and carbohydrate-binding; *, other conserved residues. (**Row 2**) Circles indicate structurally conserved positions detected by manual structure comparison, as in Figure 3. (**Rows 3–6**) Residues with the lowest positional entropy in sequence alignments of several groups of CTLDs: 95% nonredundant set of 111 human CTLD sequences, positions with entropy below 4 are plotted; 95% nonredundant set of 141 mouse CTLD sequences (entropy cutoff 4); 95% nonredundant set of 295 *C. elegans* CTLD sequences (entropy cutoff 6); 10 distinct mammalian link domain sequences (entropy cutoff 2), respectively. Triangles at the bottom of the table indicate positions conserved in all four sets of sequences; labeling as in Figure 2(a).

similar sidechains would project into the same part of the structure and form similar interactions. It has also been reported previously that residue interaction maps can be a good source of structure similarity information.[44] To apply these corrections, we implemented an algorithm (see Methods section) for detection of structurally equivalent positions important for structure maintenance, which allowed filtering of equivalenced pairs based on maximal pairwise distance (measured either between Cα centers, residue geometric centers, or sidechain geometric centers), and optional creation of an alternative sequence alignment by dynamic programming with pairwise distance as a similarity measure.

Simple filtering of the equivalenced pairs with a maximal pairwise distance threshold between Cα centers dramatically reduces the number of structurally conserved positions [Fig. 3(b)]. The most invariant parts of the structure we detected with these criteria were β1′ and β2, which, if compared with the results of the manual structure analysis, include the highly conserved pair of aliphatic residues in β1′ and the WIGL motif in β2. Conservation of the α1 and β5 regions containing the absolutely conserved cystine staple is also observed, though less pronounced than in the β1′ and β2 regions. α2 contains effectively no conserved positions, which is consistent with the variability of relative position and orientation of SSE even in closely related mammalian CTLDs. Therefore, although distance filtering allows elimination of most of the false-positive matches, it is not able to detect all positions that we consider to be structurally important (produces false negatives).

Use of sidechain geometric centers as pivot points for pairwise distance measurements followed by realignment of the structures does improve the correlation of automated detection of structurally conserved residues with the results of manual analysis [Fig. 3(c)]. For the most important residues, the change in number of correctly detected conserved cases compared with the Cα-distance filtering approach was +11% for C1 (20/18), +76% C4 (30/17), −26% β1′L (23/29), −8% β1′V (25/27), +8% β2W (27/25), −20% β1W (16/20), +160% α2E (8/3), whereas the total number of conserved positions changed hardly at all (834/808). Therefore, the third approach improves specificity for most of the positions but also produces a significant number of false negatives.

Overall, despite the good superpositions generated by the structure alignment programs, in essence, we failed to reproduce the results of the manual comparison with an automated structure-based sequence alignment algorithm we developed. It might be possible, however, to improve the performance of our algorithm by other choices of parameters, or to explore other approaches to solving this problem that would produce more satisfactory results.

Sequence Comparison

Because the set of CTLDs with solved structure is neither representative nor complete, we derived a database of all CTLD-containing sequences available in the GenPept database. CTLD sequences were extracted from the full-length sequences and aligned with ClustalW or HMMER; results of the sequence and structure conservation analysis are compared in Figure 4. We used Shannon entropy (see Methods section) as a measure of position conservation, selecting the cutoff arbitrarily depending on the number of sequences in the alignment: 2 for alignment of 10 link domains, 4 for 111 human CTLDs, 4 for 141 mouse CTLDs, and 6 for 295 *C. elegans* CTLDs. In general, structurally conserved residues are also significantly conserved at the sequence level, with eight positions being conserved through all groups. These include the absolutely conserved cystine staple (C1–C4), three residues from the small hydrophobic core (β2W, β1′L, and α1A), one of the Cys residues of the inner disulfide (C2, or a position structurally corresponding to it in link domains), and β2G and β2L from the WIGL motif. The latter two positions in the link domain alignment are occupied by residues that are not consistent with the WIGL motif (Ala and Asp, respectively), yet are conserved. Interestingly, the β2W partner in α1 is conserved above threshold only in human and mouse CTLDs, although in the structures, the nature of interaction between the two residues and the correlation of the sidechain size is almost always observed. Also, α2E and its partners are not detectable by sequence comparisons, whereas the second position in the β1′ aliphatic pair has relatively high entropy. This and other discrepancies between sequence and structure conservation analysis results may indicate that conserved positions not detectable by sequence alignment are subject to only partial structural constraints, which permit a broad range of residues. In particular, the β1′V conserved orientation may be determined by the highly conserved orientation of the preceding β1′L, and may allow any small, noncharged sidechain (ALVIPST). Other possibilities are that (as in case of α2E and its partners) a conserved structural feature (i.e., α2-β1-β5) may be maintained by structurally slightly dissimilar positions, or that sequence alignment algorithms fail to align the structurally equivalent positions correctly. Finally, it is possible that the limited number of CTLD structures available for the analysis is biased and cannot reveal all the important information about the CTLD fold organization. For example, a significant number of *C. elegans* putative CTLDs do not contain even the most conserved CTLD fold features, such as the outer cystine staple or key hydrophobic core residues. If these outliers are expressed and functional, they would be good targets for structural genomics initiatives, because they represent new groups of this otherwise structurally well-characterized domain superfamily.

Insights from the Study

The combined results provide several insights into the fold structure. There is both evolutionary (existence of compact CTLDs) and structural (LLRHC separation from the PHC) evidence that the LLR is a structurally independent part of canonical CTLDs. Thus, it is possible that canonical CTLDs might have arisen from compact CTLDs by insertion of an exogenous subdomain. The structural independence of the LLR makes it a very flexible unit for adaptive evolution. Indeed, sequence analysis (Fig. 4) shows that LLRs of CTLDs from several species are largely devoid of consistently conserved positions. In this respect, CTLDs could be regarded as analogous to Ig-like domains, in which hypervariability regions are maintained on a conserved scaffold. The relatively large size of the LLR and its structural independence provide scope for the functional plasticity observed in the CTLD fold, which can bind carbohydrates, proteins, and inorganic substances. The very variable structure, with a low proportion of formal secondary structure in the LLR, is also the reason why the CTLD overall has a relatively low secondary-structure content.

On specific points, it is interesting to observe that structural reasons for the extremely high β2G conservation in canonical CTLDs are lacking, as already noted, suggesting a different role for it than previously proposed,[4] perhaps involvement in the folding process, or maintaining some dynamic properties of the LLR that cannot be understood from the crystal structure. Another interesting point is that in addition to the β1–β5 ribbon formation that joins the N- and C-termini, β5 interacts with the N-terminal helices by a disulfide bridge and salt bridges or H-bonds with the conserved Glu. Thus, there appear to be several mechanisms stabilizing this part of the structure.

Finally, although structure alignment methods produce good superpositions for most of the CTLD structures analyzed, some of the interesting features we detected manually were generally missed by the algorithms we tested, whereas others were difficult to discriminate unambiguously from the noise. Automated structure-based sequence alignment methods would be a useful analytical tool, but our experience is that they are not easy to develop.

CONCLUSIONS

A combination of careful manual structure comparison, automated structure alignment, and analysis of structural properties, together with extensive study of sequence space, has allowed detection and annotation of the key conserved structural features that underlie CTLD fold stability. Verification of the structure comparison data against alignments of all available human, mouse, and *C. elegans* CTLD sequences showed that conservation patterns are preserved throughout the whole of CTLD sequence space. This study also provides an example of performance tests of several automated structure alignment algorithms in detection of structural features, compared with those found by manual comparison. We suggest that our observations provide a better understanding of CTLD structure and help in identification of new CTLDs and the mechanisms that drive and constrain the fold evolution.

REFERENCES

1. Drickamer K. Two distinct classes of carbohydrate-recognition domains in animal lectins. J Biol Chem 1988;263:9557–9560.
2. Drickamer K. Evolution of Ca(2+)-dependent animal lectins. Prog Nucleic Acid Res Mol Biol 1993;45:207–232.
3. Drickamer K. C-type lectin-like domains. Curr Opin Struct Biol 1999;9:585–590.
4. Weis WI, Kahn R, Fourme R, Drickamer K, Hendrickson WA. Structure of the calcium-dependent lectin domain from a rat mannose-binding protein determined by MAD phasing. Science 1991;254:1608–1615.
5. Weis WI, Drickamer K, Hendrickson WA. Structure of a C-type mannose-binding protein complexed with an oligosaccharide. Nature 1992;360:127–134.
6. Kohda D, Morton CJ, Parkar AA, Hatanaka H, Inagaki FM, Campbell ID, Day AJ. Solution structure of the link module: a hyaluronan-binding domain involved in extracellular matrix stability and cell migration. Cell 1996;86:767–775.
7. Brissett NC, Perkins SJ. The protein fold of the hyaluronate-binding proteoglycan tandem repeat domain of link protein, aggrecan and CD44 is similar to that of the C-type lectin superfamily. FEBS Lett 1996;388:211–216.
8. Rossjohn J, Buckley JT, Hazes B, Murzin AG, Read RJ, Parker MW. Aerolysin and pertussis toxin share a common receptor-binding domain. EMBO J 1997;16:3426–3434.
9. Kelly G, Prasannan S, Daniell S, Fleming K, Frankel G, Dougan G, Connerton I, Matthews S. Structure of the cell-adhesion fragment of intimin from enteropathogenic *Escherichia coli*. Nat Struct Biol 1999;6:313–318.
10. Hamburger ZA, Brown MS, Isberg RR, Bjorkman PJ. Crystal structure of invasin: A bacterial integrin-binding protein. Science 1999;286:291–295.
11. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: A structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247:536–540.
12. Hohenester E, Sasaki T, Olsen BR, Timpl R. Crystal structure of

the angiogenesis inhibitor endostatin at 1.5 Å resolution. EMBO J 1998;17:1656–1664.

13. 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;4:438–441.
14. Feinberg H, Park-Snyder S, Kolatkar AR, Heise CT, Taylor ME, Weis WI. Structure of a C-type carbohydrate recognition domain from the macrophage mannose receptor. J Biol Chem 2000;275: 21539–21548.
15. Mizuno H, Fujimoto Z, Atoda H, Morita T. Crystal structure of an anticoagulant protein in complex with the Gla domain of factor X. Proc Natl Acad Sci U S A 2001;98:7230–7234.
16. Hirotsu S, Mizuno H, Fukuda K, Qi MC, Matsui T, Hamako J, Morita T, Titani K. Crystal structure of bitiscetin, a von Willebrand factor-dependent platelet aggregation inducer. Biochemistry 2001;40:13592–13597.
17. Liu Y, Eisenberg D. 3D domain swapping: as domains continue to swap. Protein Sci 2002;11:1285–1299.
18. Weis WI, Drickamer K. Structural basis of lectin-carbohydrate recognition. Annu Rev Biochem 1996;65:441–473.
19. Natarajan K, Dimasi N, Wang J, Mariuzza RA, Margulies DH. Structure and function of natural killer cell receptors: multiple molecular solutions to self, nonself discrimination. Annu Rev Immunol 2002;20:853–885.
20. Sano H, Kuroki Y, Honma T, Ogasawara Y, Sohma H, Voelker DR, Akino T. Analysis of chimeric proteins identifies the regions in the carbohydrate recognition domains of rat lung collectins that are essential for interactions with phospholipids, glycolipids, and alveolar type II cells. J Biol Chem 1998;273:4783–4789.
21. Geider S, Baronnet A, Cerini C, Nitsche S, Astier JP, Michel R, Boistelle R, Berland Y, Dagorn JC, Verdier JM. Pancreatic lithostathine as a calcite habit modifier. J Biol Chem 1996;271: 26302–26306.
22. Ewart KV, Li Z, Yang DS, Fletcher GL, Hew CL. The ice-binding site of Atlantic herring antifreeze protein corresponds to the carbohydrate-binding site of C-type lectins. Biochemistry 1998;37: 4080–4085.
23. Mann K, Siedler F. The amino acid sequence of ovocleidin 17, a major protein of the avian eggshell calcified layer. Biochem Mol Biol Int 1999;47:997–1007.
24. Weiss IM, Kaufmann S, Mann K, Fritz M. Purification and characterization of perlucin and perlustrin, two new proteins from the shell of the mollusc *Haliotis laevigata*. Biochem Biophys Res Commun 2000;267:17–21.
25. Yokoyama WM. Natural killer cell receptors. Curr Opin Immunol 1998;10:298–305.
26. Matsumoto N, Ribaudo RK, Abastado JP, Margulies DH, Yokoyama WM. The lectin-like NK cell receptor Ly-49A recognizes a carbohydrate-independent epitope on its MHC class I ligand. Immunity 1998;8:245–254.
27. Kijimoto-Ochiai S. CD23 (the low-affinity IgE receptor) as a C-type lectin: a multidomain and multifunctional molecule. Cell Mol Life Sci 2002;59:648–664.
28. Berman HM, Battistuz T, Bhat TN, Bluhm WF, Bourne PE, Burkhardt K, Feng Z, Gilliland GL, Iype L, Jain S, Fagan P, Marvin J, Padilla D, Ravichandran V, Schneider B, Thanki N, Weissig H, Westbrook JD, Zardecki C. The Protein Data Bank. Acta Crystallogr D 2002;58:899–907.
29. Larson SM, Davidson AR. The identification of conserved interactions within the SH3 domain by alignment of sequences and structures. Protein Sci 2000;9:2170–2180.
30. Bashford D, Chothia C, Lesk AM. Determinants of a protein fold: Unique features of the globin amino acid sequences. J Mol Biol 1987;196:199–216.
31. Chothia C, Gelfand I, Kister A. Structural determinants in the sequences of immunoglobulin variable domain. J Mol Biol 1998; 278:457–479.
32. Chandra NR, Prabu MM, Suguna K, Vijayan M. Structural similarity and functional diversity in proteins containing the legume lectin fold. Protein Eng 2001;14:857–866.
33. Galperin MY, Jedrzejas MJ. Conserved core structure and active site residues in alkaline phosphatase superfamily enzymes. Proteins 2001;45:318–324.
34. Lesk AM, Fordham WD. Conservation and variability in the structures of serine proteinases of the chymotrypsin family. J Mol Biol 1996;258:501–537.
35. Hakansson K, Reid KB. Collectin structure: A review. Protein Sci 2000;9:1607–1617.
36. Weis WI, Taylor ME, Drickamer K. The C-type lectin superfamily in the immune system. Immunol Rev 1998;163:19–34.
37. Rost B. Twilight zone of protein sequence alignments. Protein Eng 1999;12:85–94.
38. Honegger A, Pluckthun A. Yet another numbering scheme for immunoglobulin variable domains: An automatic modeling and analysis tool. J Mol Biol 2001;309:657–670.
39. Godzik A. The structural alignment between two proteins: is there a unique answer? Protein Sci 1996;5:1325–1338.
40. Marchler-Bauer A, Panchenko AR, Ariel N, Bryant SH. Comparison of sequence and structure alignments for protein domains. Proteins 2002;48:439–446.
41. Shindyalov IN, Bourne PE. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Eng 1998;11:739–747.
42. Singh AP, Brutlag DL. Hierarchical protein structure superposition using both secondary structure and atomic representations. Proc Int Conf Intell Syst Mol Biol 1997;5:284–293.
43. Holm L, Sander C. Protein structure comparison by alignment of distance matrices. J Mol Biol 1993;233:123–138.
44. Godzik A, Skolnick J, Kolinski A. Regularities in interaction patterns of globular proteins. Protein Eng 1993;6:801–810.
45. Nicholls A, Sharp KA, Honig B. Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons. Proteins 1991;11:281–296.
46. Shenkin PS, Erman B, Mastrandrea LD. Information-theoretical entropy as a measure of sequence variability. Proteins 1991;11: 297–313.
47. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Rapp BA, Wheeler DL. GenBank. Nucleic Acids Res 2002;30:17–20.
48. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25: 3389–3402.
49. Li W, Jaroszewski L, Godzik A. Clustering of highly homologous sequences to reduce the size of large protein databases. Bioinformatics 2001;17:282–283.
50. 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;22:4673–4680.
51. Eddy SR. Profile hidden Markov models. Bioinformatics 1998;14: 755–763.
52. Holm L, Sander C. Mapping the protein universe. Science 1996;273: 595–603.
53. Meier M, Bider MD, Malashkevich VN, Spiess M, Burkhard P. Crystal structure of the carbohydrate recognition domain of the H1 subunit of the asialoglycoprotein receptor. J Mol Biol 2000;300: 857–865.
54. Wolan DW, Teyton L, Rudolph MG, Villmow B, Bauer S, Busch DH, Wilson IA. Crystal structure of the murine NK cell-activating receptor NKG2D at 1.95 Å. Nat Immunol 2001;2:248–254.
55. Swaminathan GJ, Weaver AJ, Loegering DA, Checkel JL, Leonidas DD, Gleich GJ, Acharya KR. Crystal structure of the eosinophil major basic protein at 1.8 Å. An atypical lectin with a paradigm shift in specificity. J Biol Chem 2001;23:26197–26203.
56. Tormo J, Natarajan K, Margulies DH, Mariuzza RA. Crystal structure of a lectin-like natural killer cell receptor bound to its MHC class I ligand. Nature 1999;402:623–631.
57. Falquet L, Pagni M, Bucher P, Hulo N, Sigrist CJ, Hofmann K, Bairoch A. The PROSITE database, its status in 2002. Nucleic Acids Res 2002;30:235–238.