ANTHONY ARTALE
MED LIB NATHAN CUMMINGS CTR (S-46)
MEMORIAL SLOAN KETTERING CANCER CTR
1275 YORK AVENUE
NEW YORK, NY 10021
UNITED STATES

Telephone: 212/639-7441
Fax: 646/422-2316

Protein engineering

10409592

Vol./Issue: 2
Date: 1989
Pages: 589-596
Article Title: RESIDUE CONTACT CONVERSATION IN A FAMILY OF CA-BINDING PROTEINS
Article Author: GODZIK A
Title: **PROTEIN ENGINEERING**
ISSN: ISSN02692139
Report Number: IRN10409592
Client Number: DDS36741/GANGI-DINO, RITA

PATRON REQUESTS COLOUR IF AVAILABLE. THANK YOU.

Protein Engineering vol.2 no.8 pp.589–596, 1989

Conservation of residue interactions in a family of Ca-binding proteins

Adam Godzik[1,2] and Chris Sander[1]

[1]BIOcomputing Programme, EMBL, D-6900 Heidelberg, FRG and
[2]Department of Biophysics, University of Warsaw, Warsaw, Poland

In the TNC family of Ca-binding proteins (calmodulin, parvalbumin, intestinal calcium binding protein and troponin C) ~70 well-conserved amino acid sequences and six crystal structures are known. We find a clear correlation between residue contacts in the structures and residue conservation in the sequences: residues with strong sidechain–sidechain contacts in the three-dimensional structure tend to be the more conserved in the sequence. This is one way to quantify the intuitive notion of the importance of sidechain interactions for maintaining protein three-dimensional structure in evolution and may usefully be taken into account in planning point mutations in protein engineering.

Key words: evolution/protein structure/mutability/conservation/contact maps

Introduction

Among the many known sequences and structures of proteins from diverse species there are families with the same or similar biochemical function, essentially similar three-dimensional fold and some similarity between sequences: a living molecular record of protein evolution. The analysis of such families is an inexpensive substitute for point mutation experiments. Correlating sequence and structure information, one can study the question of how changes in the amino acid sequence of a protein affect its function and three-dimensional structure (Lesk and Chothia, 1980, 1982, 1986), or conversely, how much variation is allowed in the sequence within the constraint of maintaining the same structure (Ponder and Richards, 1987) and function. We focus on the latter question, using the troponin C family as an example.

Proteins in the TNC family participate in various biological processes regulated by calcium. For instance, muscle contraction is triggered by a conformational change in troponin C induced by calcium binding. A number of enzyme complexes are activated by calmodulin. The function of parvalbumins and intestinal calcium binding proteins is less clear. The properties and functions of calcium binding proteins have been extensively reviewed (Kretsinger, 1980). This family provides an excellent basis for systematic studies of activity–structure–sequence relationships. There are >100 known protein sequences, from the work of many authors, recently analyzed by Boguta and Bierzynski (1988) and Godzik and Boguta (1988). Six three-dimensional crystal structures are known, with protein data bank (Koetzle *et al.*) codes CPV, TNC, ICB, CLN: carp muscle calcium binding parvalbumin B (3CPV, Moews and Kretsinger, 1975), turkey skeletal muscle troponin C (2TNC, 5TNC, Herzberg and James, 1985a,b), bovine vitamin D-dependent intestinal calcium-binding protein (3ICB, Szebenyi and Moffat, 1986), rat testis calmodulin (3CLN, Babu *et al.*, 1988), chicken skeletal muscle troponin C (4TNC, Satyshur *et al.*, 1988), pike parvalbumin (Declerc *et al.*, 1988). The calcium binding affinity ranges over four orders of magnitude in the family, and although sequences of the basic calcium binding domain have sequence similarity as low as 40–50% (counting identical amino acids), the three-dimensional structures are remarkably similar in detail.

Here, we first describe two methods of analyzing protein sequence and structure: quantification of sequence mutability (residue conservation) based on family alignment and quantification of residue–residue interactions (sidechain contacts) based on three-dimensional structure. We apply these to the TNC family and derive a hypothesis about the influence of residue contacts on mutability.

Methods

Quantitative measure of residue conservation in a protein family

Inspection of the TNC family multiple sequence alignment (Figure 1a) shows that residues at some positions are strongly conserved while more variation is allowed at others. We need to define a quantitative measure of residue conservation at a given position in the family alignment. Several measures of amino acid similarity exist for pairs of single amino acids (see recent book by Bishop and Rawlings, 1987), quantifying, say, $V \sim I$ or $F \neq D$, but how can this be generalized to describe variation at one sequence position in an entire family? For example, Kabat and Wu (1970, reviewed by Kabat, 1978) have defined *variability* as the number of different amino acids at a given family position divided by the frequency of the most common amino acid at that position (scale: 1–400) while Padlan (1977) has defined *structural variability* as the average pair dissimilarity relative to the most commonly occurring residue at that position. Here, we define *sequence conservation* by averaging pair similarity over all possible exchange pairs at one position—rather than only over pairs involving the most commonly occurring residue, as no one residue can be uniquely traced to the evolutionary ancestor.

In detail: the MDM78 (Mutation Data Matrix) mutational similarity matrix of Schwartz and Dayhoff (1979) describes the probability of exchanging one amino acid for another as based on the analysis of protein families (1572 mutations in 793 protein sequences known in 1978, such as cytochromes, globins, immunogloblins), but it also serves as a general indication of amino acid similarity (e.g. Sander and Schulz, 1979). We use the Dayhoff matrix as distributed with the UWGCG sequence analysis software (Devereux *et al.*, 1984) and as used by Gribskov and Burgess (1986), except that we rescale exchange values (by a factor of 1/1.5) such that for each amino acid its conservation (exchange for itself) has a value equal to 1.0 (lowest value for W↔C −0.8). Then, for each sequence position in a multiple alignment we calculate its conservation value by averaging the similarity value for all pairs of residues from the appropriate column. For every position there are $n(n-1)/2$ such pairs, where n is the number of sequences (rows). If r and s are

protein (column) indices, R the amino acids and *sim* the similarity values, then residue conservation $c(i)$ at position i is:

$$c(i) = \frac{1}{n(n-1)/2} \sum_{r<s} sim(R_{ir}, R_{is})$$

If an amino acid is strictly conserved at position i, $c(i) = 1.0$; $c(i)$ values <0.3 indicate a large mutability. Note that the precise numerical values of residue conservation depend on the somewhat arbitrary sample of proteins in the database. Ideally, the sequences used span a wide range of the naturally occurring sequences in

residue position>		....;....1....;....2....;....3....;....4....;....5....;....6....;....7..	
most conserved>		* * * * * ** * * * ** * ** * * ** *	
consensus sequence>		F D D G I EL M E D DG G I EF M	
sequence code	from		to
Cabi$Pig	5	SPAELKSIFEKYAAKEGDPNQLSKEELKQLIQAE.FPSLLKGPRTLDDLFQELDKNGNGEVSFEEFQVLVKK	75
Cabi$Rat	5	SPEEMKSIFQKYAAKEGDPNQLSKEELKLLIQSE.FPSLLKASSTLDNLFKELDKDGDGEVSYEEFEVFFKK	75
Calm$Chlre	11	QIAEFKEAFALFD.KDGD.GTITTKELGTVMRSL...GQNPTEAELQDMISEVDADGNGTIDFPEFLMLMAR	77
Calm$Dicdi	10	QIAEFKEAFSLFD.KDGD.GSITTKELGTVMRSL...GQNPTEAELQDMINEVDADGNGNIDFPEFLTMMAR	76
Calm$Drome	8	QIAEFKEAFSLFD.KDGD.GTITTKELGTVMRSL...GQNPTEAELQDMINEVDADGNGTIDFPEFLTMMAR	74
Calm$Eleel	8	QIAEFKEAFSLFD.KDGD.GTITTKELGTVMRSL...GQNPTEAELQDMINEVDADGNGTIDFPEFLTMMAR	74
Calm$Human	8	QIAEFKEAFSLFD.KDGD.GTITTKELGTVMRSL...GQNPTEAELQDMINEVDADGNGTIDFPEFLTMMAR	74
Calm$Metse	8	QIAEFKEAFSLFD.KDGD.GTITTKELGTVMRSL...GQNPTEAELQDMINEVDADGDGTIDFPEFLTMMAR	74
Calm$Parte	8	QIAEFKEAFALFD.KDGD.GTITTKELGTVMRSL...GQNPTEAELQDMINEVDADGNGTIDFPEFLSLMAR	74
Calm$Patsp	8	QIAEFKEAFSLFD.KDGD.GTITTKELGTVMRSL...GQNPTEAELQDMINEVDADGDGTIDFPEFLTMMAR	74
Calm$Schpo	10	QIAEFREAFSLFD.RDQD.GNITSNELGVVMRSL...GQSPTAAELQDMINEVDADGNGTIDFTEFLTMMAR	76
Calm$Spiol	8	QIAEFKEAFSLFD.KDGD.GCITTKELGTVMRSL...GQNPTEAELQDMINEVDADGNGTIDFPEFLNLMAR	74
Calm$Tetpy	8	QIAEFKEAFSLFD.KDGD.GTITTKELGTVMRSL...GQNPTEAELQDMINEVDADGDGTIDFPEFLSLMAR	74
Calm$Trybr	8	QISEFKEAFSLFD.KDGD.GTITTKELGTVMRSL...GQNPTEAELQDMINEVDQDGSGTIDFPEFLTLMAR	74
Calm$Wheat	9	NIAEFKEAFSLFD.KDGD.GCITTKELGTVMRSL...GQNPTEAELQDMINEVDADGNGTIDFPEFLNLMAR	75
Calm$Yeast	9	QIAEFKEAFALFD.KDNN.GSISSSELATVMRSL...GLSPSEAEVNDLMNEIDVDGNHQIEFSEFLALMSR	75
Cals$Chick	8	QIAEFKEAFSLFD.RDGD.GCITTMELGTVMRSL...GQNPTEAELQDMVGEVDADGSGTIDFPEFLSLMAR	74
Calm$Arbpy	80	SEEEIREAFRVFD.KDGN.GFISAAELRHVMTNL...GEKLTDEEVDEMIREADIDGDGQVNYEEFVAMMTS	137
Calm$Chlre	84	HEDELREAFKVFD.KDGN.GFISAAELRHVMTNL...GEKLSEEVDEMIREADVDGDGQVNYEEFVRMMTS	150
Calm$Dicdi	83	TEEEIREAFKVFD.KDGN.GYISAAELRHVMTSL...GEKLTNEEVDEMIREADLDGDGQVNYDEFVKMMIV	147
Calm$Drome	81	SEEEIREAFRVFD.KDGN.GFISAAELRHVMTNL...GEKLTDEEVDEMIREADIDGDGQVNYEEFVTMMTS	147
Calm$Eleel	81	SEEEIREAFRVFD.KDGN.GYISAAELRHVMTNL...GEKLTDEEVDEMIREADIDGDGQVNYEEFVQMMTA	147
Calm$Human	81	SEEEIREAFRVFD.KDGN.GYISAAELRHVMTNL...GEKLTDEEVDEMIREADIDGDGQVNYEEFVQMMTA	147
Calm$Metse	81	SEEEIREAFRVFD.KDGD.GFISAAELRHVMTNL...GEKLTDEEVDEMIREADIDGDGQVNYEEFVKMMTS	147
Calm$Parte	81	SEEELIEAFKVFD.RDGN.GLISAAELRHVMTNL...GEKLTDDEVDEMIREADIDGDGHINYEEFVRMMVS	147
Calm$Patsp	81	SEEEIREAFRVFD.KDGD.GFISAAELRHVMTNL...GEKLTDEEVDEMIREADIDGDGQVNYEEFVTMMTS	147
Calm$Schpo	83	NEEEVREAFKVFD.KDGN.GYITVEELTHVLTSL...GERLSQEEVADMIREADTDGDGVINYEEFSRVISS	149
Calm$Spiol	81	SEEELKEAFRVFD.KDQN.GFISAAELRHVMTNL...GEKLTDEEVDEMIREADVDGDGQINYEEFVKVMMA	147
Calm$Strpu	13	SEEEIREAFRVFD.KDGN.GFISAAELRHVMTNL...GEKLTDEEVDEMIREADIDGDGQVNYEEFVAMMTS	79
Calm$Tetpy	81	SEEELIEAFKVFD.RDGD.GLITAAELRHVMTNL...GEKLTDEEVDEMIREADIDGDGHINYEEFVRMMAK	147
Calm$Trybr	81	SEEEIKEAFRVFD.KDGN.GFISAAELRHIMTNL...GEKLTDEEVDEMIREADVDGDGQINYEEFVKMMMS	147
Calm$Wheat	82	SEEELKEAFRVFD.KDQD.GFISAAELRHVMTNL...GEKLTDEEVDEMIREADVDGDGQINYEEFVKVMMA	148
Cals$Chick	81	SEEEIREAFRVFD.KDGN.GYISAAELRHVMTNL...GEKLTDEEVDEMIKEADCNNDGQVNYEEFVRMMTE	147
Prva$Cypca	38	TPDDIKKAFAVID.QDKS.GFIEEDELKLFLQNFSAGARALTDAETKAFLKAGDSDGDGKIGVDEFAALVKA	108
Prva$Ampme	39	SPADVTKAFHILD.KDRS.GYIEEEELQLILKGFSKEGRELTDKETKDLLIKGDKDGDGKIGVDEFTSLVAE	108
Prva$Esolu	38	SANDVKKVFKAID.ADAS.GFIEEEELKFVLKSFAADGRDLTDAETKAFLKAADKDGDGKIGIDEFETLVHE	107
Prva$Latch	41	PDDTLKEVFGILD.QDKS.GYIEEEELKFVLKGFAAGGRELTANETKALLKAGDQDGDDKIGVDEFTNLVKA	110
Prva$Rabit	39	STEDVKKVFHILD.KDKS.GFIEEEELGFILKGFSPDARDLSVKETKTLMAAGDKDGDGKIGADEFSTLVSE	108
Prva$Rajcl	39	SDAELAEIFNVLD.GDQS.GYIEVEELKNFLKCFSDGARVLNDKETSNFLAAGDSDGDHKIGVDEFKSMAKM	107
Prva$Ranes	39	SKEIMQKVFHVLD.QDQS.GFIEKEELCLILKGFTPEGRSLSDKETTALLAAGDKDGDGKIGVDEFVTLVSE	108
Prva$Rat	39	SADDVKKVFHILD.KDKS.GFIEEDELGSILKGFSSDARDLSAKETKTLMAAGDKDGDGKIGVDEFQALVRS	108
Prvb$Ampme	39	PTDQVKKVFDILD.QDKS.GYIEEDELQLFLKNFCSSARSLSNAETKAFLFAGDSDGDGKIGVDEFSTLVAE	108
Prvb$Boaco	39	SKDQLTKVFGVID.RDKS.GYIEEDELKKFLQNFDGKARDLTDKETAEFLKEGDTDGDGKIGVEEFVVLVTK	108
Prvb$Cypca	39	SADDVKKAFAIID.QDKS.GFIEEDELKLFLQNFKADARALTDGETKTFLKAGDSDGDGKIGVDEFTALVKA	108
Prvb$Esolu	38	SLDDVKKAFYVID.QDKS.GFIEEDELKLFLQNFSPSARALTDAETKAFLADGDKDGDGMIGVDEFAAMIKA	107
Prvb$Gadca	39	SADELKKLFKIAD.EDKE.GFIEEDELKLFLIAFAADLRALTDAETKAFLKAGDSDGDGKIGVDEFGALVDK	108
Prvb$Latch	39	SNEELEAIFKILD.QDKS.GFIEDEELELFLQNFSAGARTLTKTETETFLKAGDSDGDGKIGVDEFQKLVKA	108
Prvb$Leuce	37	SAGDVKKAFEIID.EDKS.GFIEEEELKLFLQNFKAGARALTDAETKIFLKAGDADGDGKIGIDEFAALVKA	106
Prvb$Merme	39	SAADIKKVFGIID.QDKS.DFVEEDELKLFLQNFSAGARALTDAETATFLKAGDSDGDGKIGVEEFAAMVKG	108
Prvb$Mermr	39	SADDIKKAFVFID.QDKS.GFIEEDELKLFLQVFKAGARALTDAETKAFLKAGDSDGDGAIGVEEWVALVKA	108
Prvb$Opsta	36	TPNVIKKAFYVID.QDKS.GFIEEDELKLFLQNFASSARALTDKETETFLKAGDSDGDGKIGIDEFADLVKE	108
Prvb$Ranes	39	SAADAKKVFEILD.RDKS.GFIEQDELGLFLQNFRASARVLSDAETSAFLKAGDSDGDGKIGVEEFQALVKA	108
Prvb$Xenla	42	SADDVKNVFAILD.QDRS.GFIEEEELKLFLQNFSASARALTDAETKAFLAAGDSDGDGKIGVEEFQSLVKP	112
Tpcs$Chick	18	MIAEFKAAFDMFD.ADGG.GDISTKELGTVMRML...GQNPTKEELDAIIEEVDEDGSGTIDFEEFLVMMVR	84
Tpcs$Human	15	MIAEFKAAFDMFD.ADGG.GDISVKELGTVMRML...GQTPTKEELDAIIEEVDEDGSGTIDFEEFLVMMVR	81
Tpcs$Pig	15	MIAEFKAAFDMFD.ADGG.GDISVKELGTVMRML...GQTPTKEELDAIIEEVDEDGSGTIDFEEFLVMMVR	81
Tpcs$Rabit	15	MIAEFKAAFDMFD.ADGG.GDISVKELGTVMRML...GQTPTKEELDAIIEEVDEDGSGTIDFEEFLVMMVR	81
Tpcs$Ranes	18	MIAEFKAAFDMFD.TDGG.GDISTKELGTVMRML...GQTPTKEELDAIIEEVDEDGSGTIDFEEFLVMMVR	84
Tpcc$Human	93	SEEELSDLFRMFD.KNAD.GYIDLEELKIMLQAT...GETITEDDIEELMKDGDKNNDGRIDYDEFLEFMKG	159
Tpcc$Rabit	93	SEEELSDLFRMFD.KNAD.GYIDLDELKIMLQAT...GETITEDDIEELMKDGDKNNDGRIDYDEFLEFMKG	159
Tpcc$Cotjy	93	TEEELSDLFRMFD.KNAD.GYIDLEELKIMLQAT...GETITEDDIEELMKDGNKNNDGRIDYDEFLQFMKG	159
Tpcs$Chick	94	SEEELANCFRIFD.KNAD.GFIDIEELGEILRAT...GEHVIEEDIEDLMKDSDKNNDGRIDFDEFLKMMEG	160
Tpcs$Human	91	SEEELAECFRIFD.RNAD.GYIDPEELAEIFRAS...GEHVTDEEIESLMKDGDKNNDGRIDFDEFLKMMEG	157
Tpcs$Pig	91	SEEELAECFRIFD.RNMD.GYIDAEELAEIFRAS...GEHVTDEEIESIMKDGDKNNDGRIDFDEFLKMMEG	157
Tpcs$Rabit	91	SEEELAECFRIFD.RNAD.GYIDAEELAEIFRAS...GEHVTDEEIESLMKDGDKNNDGRIDFDEFLKMMEG	157
Tpcs$Ranes	94	SEEELAECFRIFD.KNAD.GYIDSEELGEILRSS...GESITDEEIEELMKDGDKNNDGKIDFDEFLKMMEG	160
secondary structure		HHHHHHHHHHHH HHHHHHHHHH HHHHHHHHHHHH HHHHHHHHHH	

Fig. 1 (a)

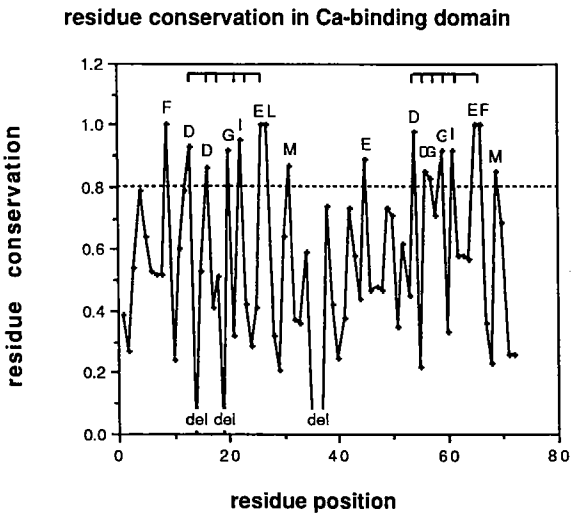

Fig. 1. (**a**) Aligned sequences of 69 Ca-binding domains. Each domain contains a pair of helix−loop−helix structures. Only sequences with >75% homology to one of seven crystallographically known proteins are included. Positions with residue conservation value >0.8 (above the dotted horizontal line in **b**) are denoted by stars in the top row. Sequence fragments are identified by their EMBL/SWISS-PROT entry identifier (A.Bairoch, G.Cameron, D.Hazledine, G.Stoesser, P.Kahn, B.Boeckmann, release 9.0, 1988) and first (from) and last (to) residue in the sequence entry. The first four letters identify the family: Calm, calmodulins; Tpcs(c), troponins C; Prva(b), parvalbumins; Cabi, intestinal calcium binding proteins. Sequences closest to (but not necessarily identical to) the protein of known three-dimensional structure are: 5TNC/Tpcs$Chick, 3CLN/Cln$Human, 3ICB/Cabi$Pig, 3CPV/Prvb$Cypca, Pike pavalbumin/Prva$Esolu (Protein Data Bank, EMBL/SWISSPROT Release 8 identifier). The sequence numbering of sea urchin calmodulin Calm$Strpu is from a fragment. Secondary structure (bottom row) is an approximate consensus based on the known three-dimensional structures: H, helix; blank, loop. (**b**) Not all residues in a protein sequence are essential for its structure and function and consequently well conserved in the course of evolution. From a mutiple sequence alignment (**a**) residue conservation at a given sequence position can be calculated using a pair exchange matrix (Schwartz and Dayhoff, 1979). Conservation statistics are based on the 69 sequences in **a**. Residue conservation varies strongly along the chain, values near 1.0 mean very little change at a given position, random amino acid content gives values close to 0.2. Peaks above 0.8 (dotted line) are marked by the name of the amino acid prevalent at this position: F, Phe; D, Asp; G, Gly; I, Ile; E, Glu; L, Leu; M, Met; del marks deletions/insertions. The two Ca-binding loops are indicated by horizontal lines, the attached short vertical lines mark the position of calcium ligands (sidechain or backbone C=O; six per loop). Of the 17 most conserved residues (conservation value >0.8, marked 'most conserved' in **a**), six are negatively charged calcium ligands: Asp 13, Asp 16, Glu 26 from the first calcium-binding loop in the domain and Asp 54, Asp 56 and Glu 65 from the second; three are Gly 20, Gly 57 and Gly 59 in the Ca-binding loops; seven are large hydrophobic residues: Phe 9, Ile 22, Leu 27, Met 31, Ile 61, Phe 66 and Met 70, five of these in helices; and one is Glu 45, also on a helix.

the family. We call the resulting plot of conservation value against position number, the 'conservation profile' (Figure 1b); it is similar in spirit to Kabat and Wu's (1970) variability plot.

If conserved positions are important for maintaining the biochemical activity of the protein and/or for determining its three-dimensional structure, then analysis of these positions may lead to discovery of the molecular reasons for residue conservation.

Contact maps

The specific fold of protein three-dimensional structure appears to be primarily determined by interactions of residues that are nearest neighbors in space, although such residues may be far apart in the sequence. A two-dimensional contact map, although a reduction from full three-dimensional information, is a powerful tool for displaying and analyzing such contacts.

In the contact maps used here (Figure 2), a marked square at i,j represents a pair of side chains in spatial contact, where i and j are the sequence positions of the residues. A close contact (black square) is detected in the crystal structure if at least one sidechain atom of residue i is within 3.75 Å of a sidechain atom of residue j; more distant contacts have decreasing grey values; contacts beyond 5 Å are disregarded (white squares). In the statistical average over several TNC structures (Figure 3) a cutoff of 3.75 Å is used, i.e. only black squares are counted. Our definition is similar to previous forms of distance or contact plots, but different in detail (Phillips, 1970; Rossman and Liljas, 1974; Nishikawa and Ooi, 1974; Kuntz, 1975; Liebman *et al.*, 1985; Kraulis and Jones, 1987; Richards and Kundrot, 1988); contact maps were used for parvalbumin by Moews and Kretsinger (1975), for ICaBP by Szebeni and Moffat (1986) and for TNC by Satyshur *et al.* (1988) to describe the pattern of hydrophobic contacts. For this analysis, two aspects of contact maps are particularly useful. (i) Spatial proximity between residues far apart in the sequence is seen as far off-diagonal clusters, e.g. helix−helix contacts. (ii) As contact maps are based on interatomic distances and not on absolute spatial positions, their inspection and analysis may reveal common structural patterns in spite of conformational changes. Here we are especially interested in detecting tertiary structure contact patterns common to all known structures in the TNC family.

Sequences and structures in the TNC family

For the present analysis we have selected 69 domains containing two neighboring calcium-binding units from the >100 known sequences of parvalbumins, calmodulins, troponins C and intestinal calcium-binding proteins. Related proteins that have 'lost' some of the Ca-binding properties, in particular myosin light chains, have been omitted. For the selected domains the family alignment is well determined and the sequence similarity to one or more of the crystallographically solved proteins is sufficiently strong to guarantee that the three-dimensional structure is similar for all domains. The multiple sequence alignment is the basis for quantification and detailed analysis of residue conservation (Figure 1b).

From the six crystal structures we know that the calcium binding sites for proteins from the TNC family have almost identical structures. Each binding site has a 32- to 34-residue-long helix−loop−helix pattern, also called EF hand (Moews and

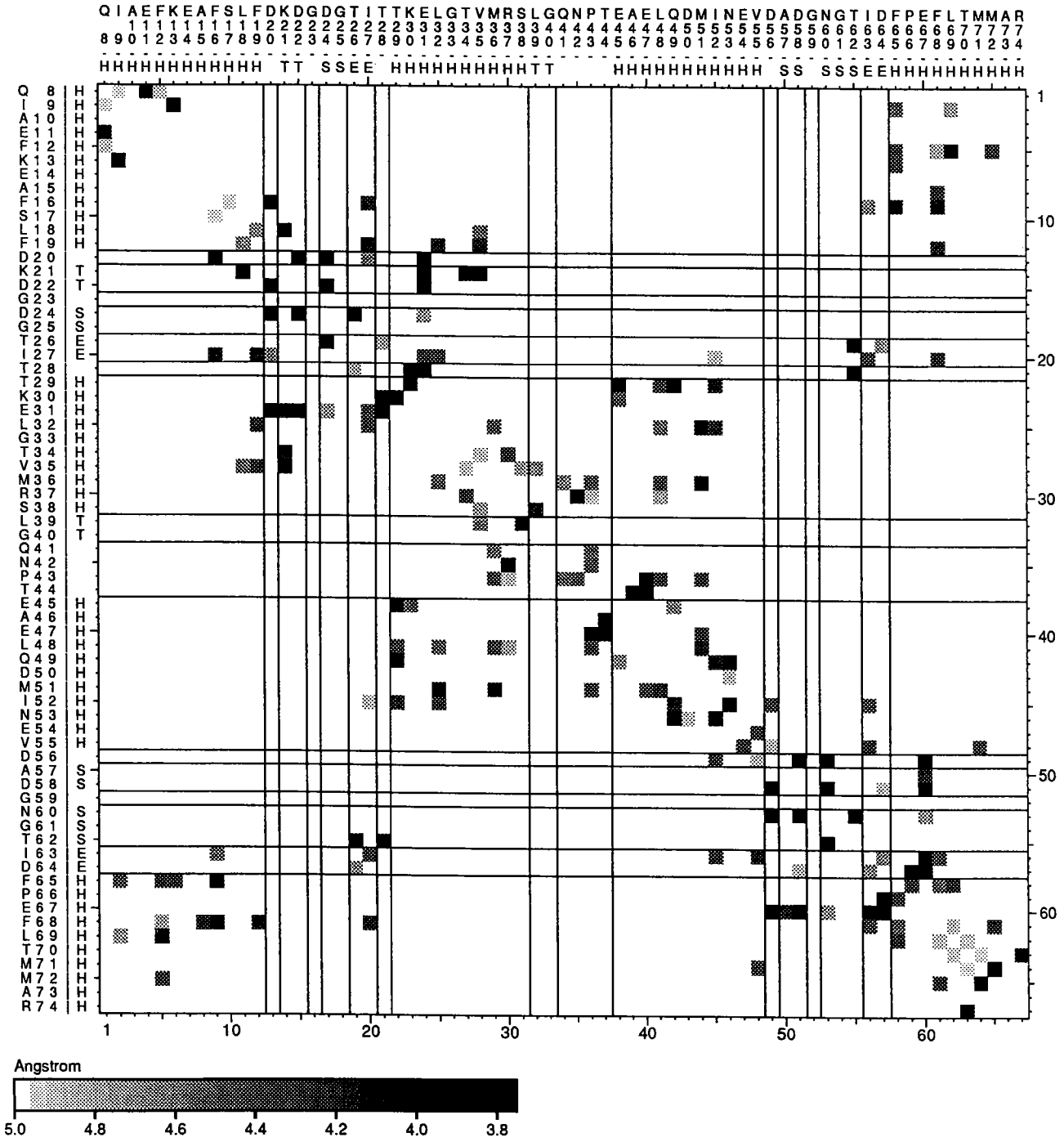

Fig. 2 (a)

Kretsinger, 1975), where a 9- to 11-residue-long central calcium-binding loop connects two roughly antiparallel α-helices. Two such EF hand patterns, linked by a β-bridge and helix−helix contacts, form a single domain that binds two calcium ions.

The structural similarity between them is not confined to the overall fold but extends to the pattern of residue interactions. For example, the sidechain−sidechain contact maps for the first domain of calmodulin and the second domain of troponin C (Figure 2a and b) are strikingly similar. The comparison contact map (color, Figure 2c; C.Sander and M.Scharf, unpublished) highlights contacts which are common to both structures as full squares and stronger contacts by higher color intensity. We are not aware of prior use of color comparison maps in protein contact analysis.

The off-diagonal contact regions are particularly interesting, as they contain the 'tertiary' structure interactions. Starting from the lower left or upper right: region I involves the contact between the first and fourth helix (in 3CLN: FKEAFSLF/12−19 make contacts FPEFLTMM/65−72); region II includes the beta-bridge contact between the central residues of the two calcium binding

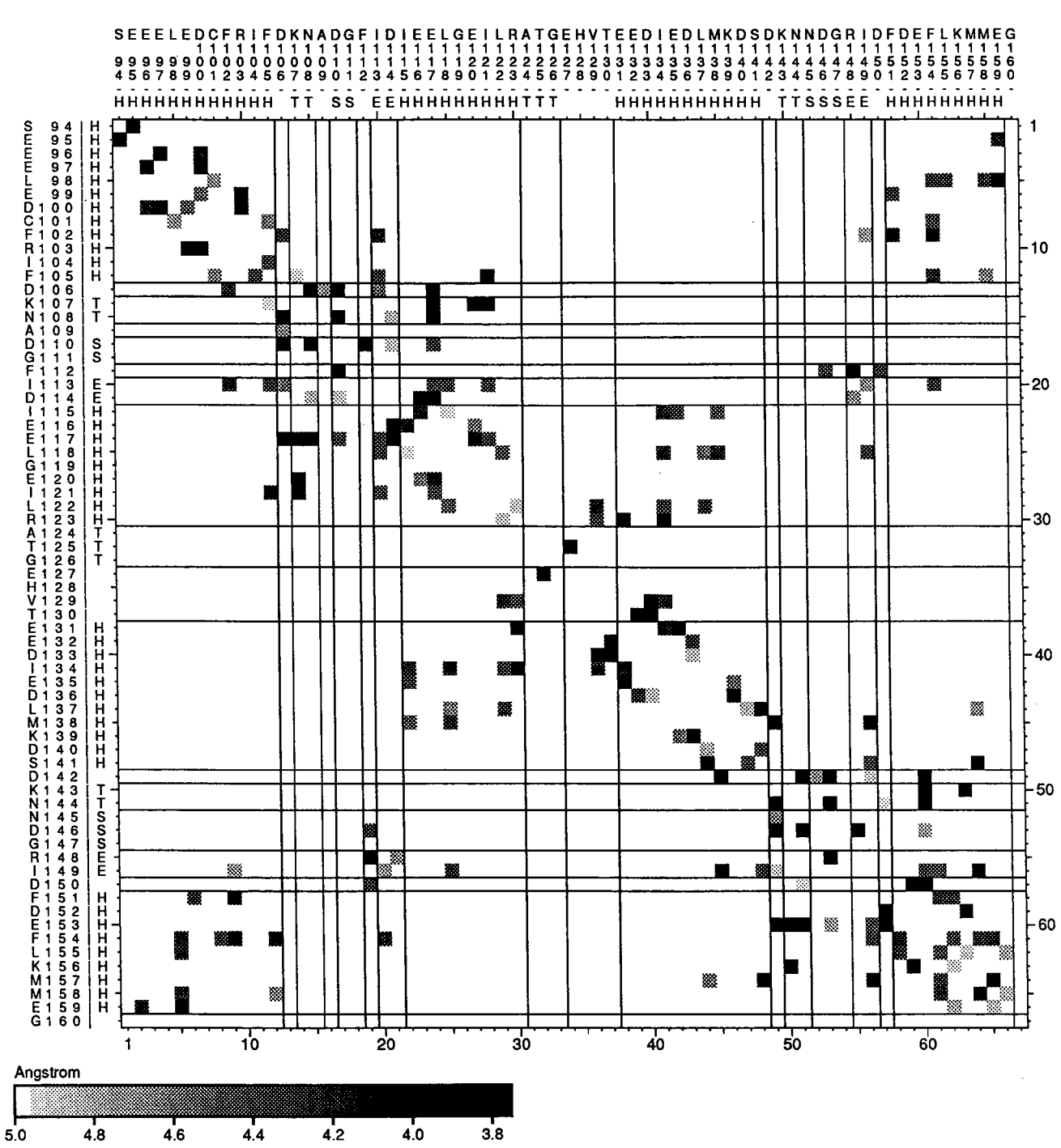

Fig. 2 (b)

loops (in 3CLN: TIT/26–28 make contacts with TID/62–64); and region III involves the contact between the second and third helix (in 3CLN: TKELGTVMR/29–37 make contacts with LQDMI/48–52). Other conserved contacts are in the calcium-binding loops and at the N-termini of the helices (Figure 2c).

Analysis of residue conservation

Here we are interested in how much sequence variation a given structure allows. More specifically, can we understand the molecular reasons for sequence conservation in particular positions? Which structural properties of a residue or residue pair influence residue conservation? Can we, by inspection of a single protein structure, predict sequence positions with strong sequence conservation?

There is remarkable variation of conservation of amino acid residues in the protein—plotted as a function of sequence position in Figure 1b and color-coded in three dimensions in Figure 3. Conservation is lowest in the loop between the two helix–loop–helix fragments and highest in or near the

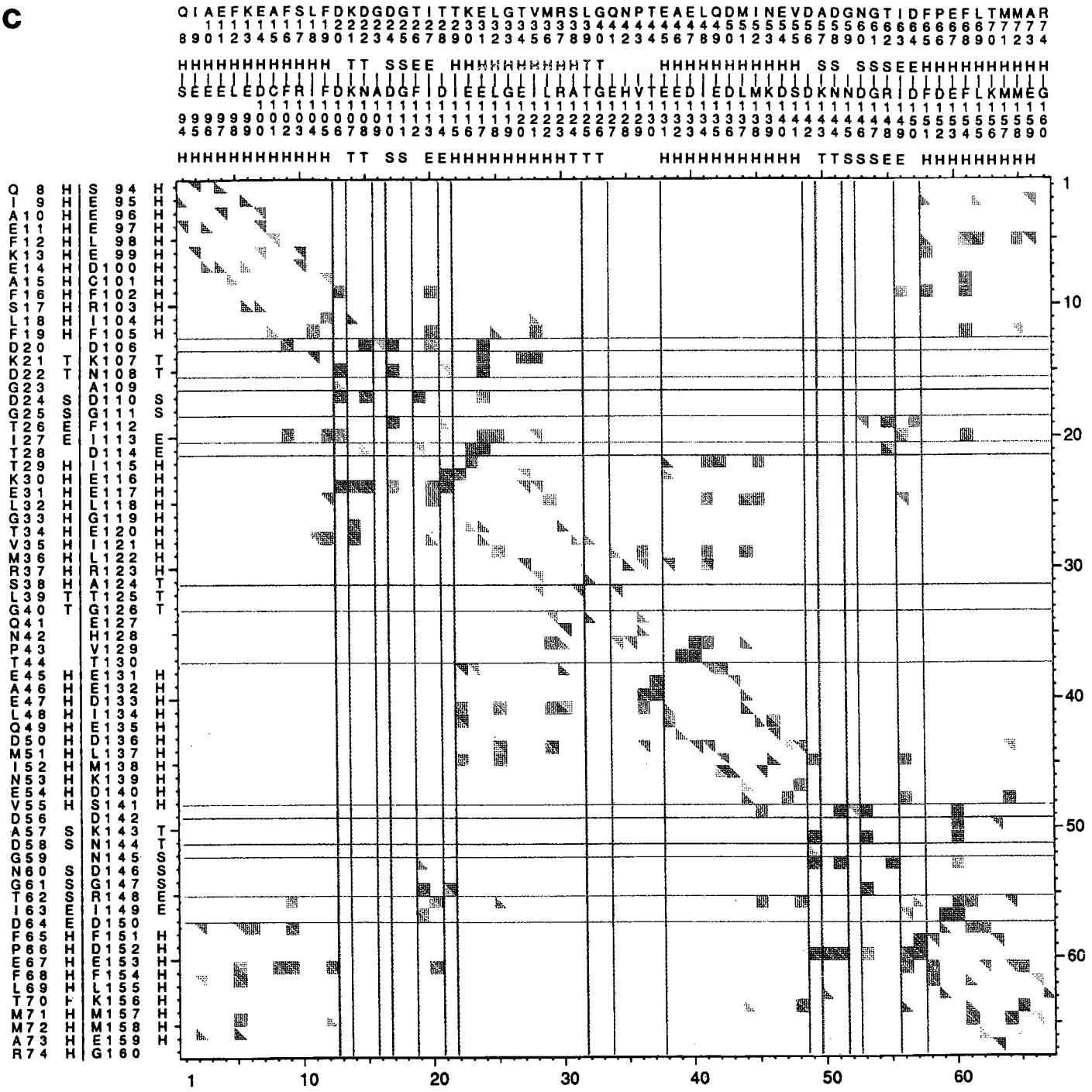

Fig. 2. Which residues are close to each other in the folded state of the protein? Sidechain contact maps are calculated from the three-dimensional structures of (**a**) calmodulin [PDB, 3CLN, Babu *et al.*, 1988] and of (**b**) the second domain of troponin C [PDB, 5TNC, Herzberg and James, 1985a,b] (36% identical amino acids). (**c**) Residue pairs that make structurally conserved sidechain contacts. Top and side legend: sequence in one-letter code, PDB residue number, DSSP secondary structure (Kabsch and Sander, 1983) for one or two proteins. (**a,b**) A sidechain residue−residue contact at map position *i,j* indicates that at least one pair of atoms, one from the sidechain of residue *i* and one from the sidechain of residue *j*, are within 3.75 Å (black squares) up to 5.0 Å (almost white squares) of each other. The three most characteristic off-diagonal contact regions are: (I) loop−loop contact involving β-bridge Ile 22/Ile 61; (II) helix-A−helix-D contact involving Phe 9, Phe 66 and Met 70; (III) helix-B−helix-C contact involving Leu 27 and Met 31 (naming only the most conserved residues in each region by their residue number in the family alignment (Figure 1a). (**c**) The comparison contact map compares the sidechain contacts in the two domains of (**a**) and (**b**). Red triangles are the contacts in the calmodulin (3CLN, res 8−74) domain, green triangles those in the troponin C domain (5TNC, res 94−160). Conserved contacts are visible as full squares (half red, half green). Variable contacts, present only in either protein, are visible as isolated red or green triangles. The three main off-diagonal (chain-distant) contact regions are highly conserved: look for clusters of full squares, starting from lower left. Also conserved are near-diagonal (chain-local) contacts in the Ca-binding loops and at helix N-termini. Color intensity goes from full saturation (minimum distance between sidechains 3.75 Å to zero saturation (minimum distance 5.0 Å). All contacts maps were produced using the software CONAN (M.Scharf, Diplomarbeit, Heidelberg University, 1989) on a SUN-3 workstation.

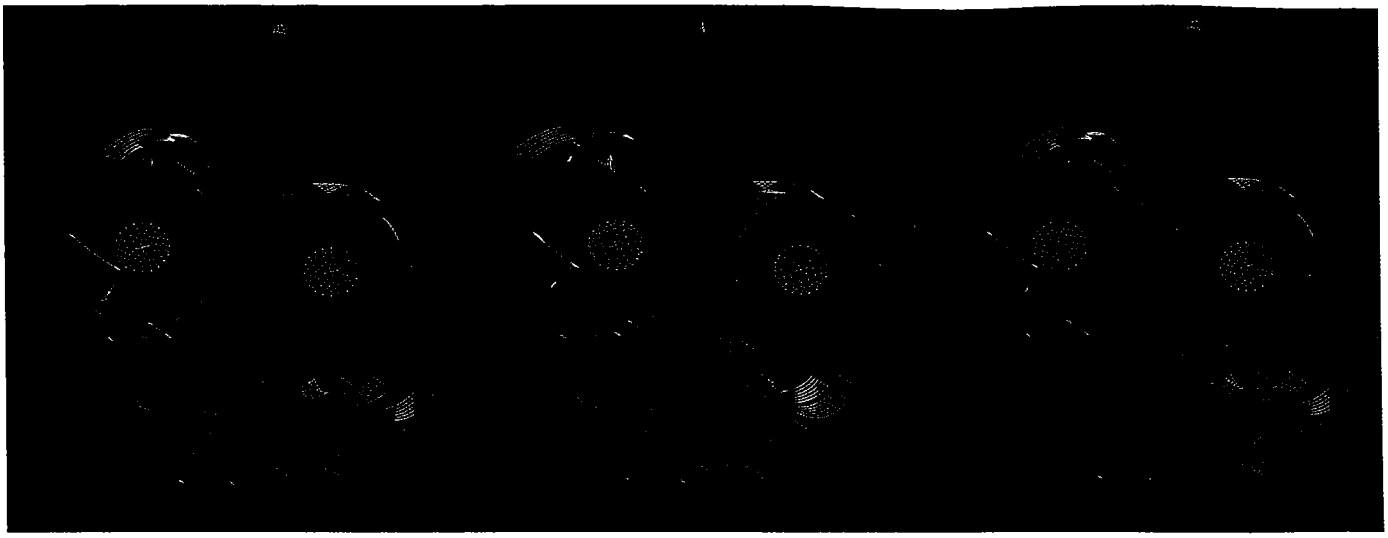

Fig. 3. Stereo view of ribbon cartoon of one member of the family, the first domain of rat testis calmodulin (residues 8–74 of PDB:3CLN; Babu *et al.*, 1988), color-coded by residue conservation. Conservation decreases from red (most conserved) via violet (medium conserved) to blue (least conserved). Calcium ions are dotted orange spheres. Gln8 and helix-A are at top right, Arg74 and helix-D at top left. The view is from the 'stem' side of the first calcium-binding domain. Conserved contacts are between the top two helices (helix-A with helix-D, region I), between the two loops in the back, sandwiched by the two calcium ions (beta bridge, region II) and between the two helices at the bottom (helix B with helix-C, region III). Note that this protein structure cartoon (made using INSIGHT) only represents rough residue location, while sidechains and their contacts are NOT visible. For a representation of sidechain contacts, an important quantity affecting conservation, the contact map view of protein structure (Figure 2a–c) is more appropriate.

Ca-binding loops; conservation is different in detail in the first and the second Ca-binding fragment within the domain. We initially focus on the 17 highly conserved residues—those with a conservation value >0.8 (arbitrary cut-off, Figure 1b). Six of these have negatively charged sidechains directly involved in Ca coordination: D13, D16 and E26 in the first Ca-binding loop, and D54, D56 and E65 in the second Ca-binding loop at completely equivalent positions. Gly 20 and Gly 59 (tight corner, $\Phi \sim +90°$, $\Psi \sim +5°$) as well as Gly 57 appear to have a unique structural role in the Ca-binding loops and are conserved (although their carboxyl groups point away from the Ca ion).

The reason for conservation at the eight other positions is not immediately obvious, but inspection of the contact map identifies seven of them as important contact residues: F9, I22, L27, M31, E45, I61, F66, M70. Many of the contacts made by these residues cluster in the three off-diagonal contact regions (Figure 2) that are similar in all known structures down to atomic detail, suggesting the existence of a core of tertiary structure contacts essential for this structure, analogous to (but different in detail from) the conserved pattern of helix–helix contacts in the globins (Lesk and Chothia, 1980).

Can this observation be generalized, i.e. does higher contact strength imply higher residue conservation not only for residues in the most conserved parts of a globular protein but for all residues?

There is a clear correlation between contact strength and residue conservation of a given residue position (Figure 4). More precisely, note the absence of points in the upper left-hand corner of Figure 4: the higher the contact strength the more pronounced the tendency to be conserved. This is our main result.

The converse, however, is not strictly true: note some points in the lower right-hand corner. Some residues are highly conserved, yet they make few sidechain contacts. The most obvious examples are the conserved glycines, for which sidechain contacts are not defined (counted as zero) and which may be

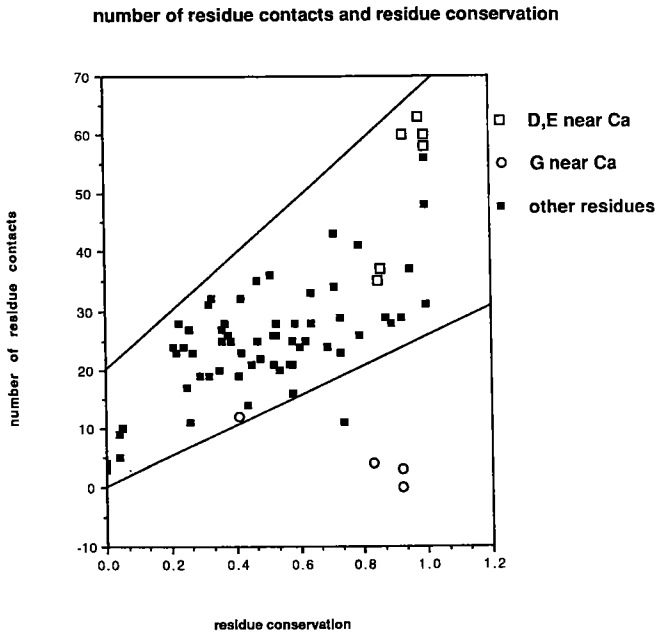

Fig. 4. What are the reasons for different variation allowed at different positions? Each point represents one position in the family alignment. Correlation of the number of inter-residue side chain contacts made by an amino acid at that position with residue conservation shows that if a residue makes many contacts it is more likely to be conserved. Residue conservation is calculated by averaging over all possible exchange pairs at that position (same as in Figure 1b). The number of contacts is defined as the number of different residues that have at least one sidechain atom within 3.75 Å of any sidechain atom of a given residue, summed over the six known structures. The oblique straight lines enclose all sequence positions, except four that have predominantly Gly, for which sidechain contacts are zero. The single black square at the bottom right is Gly 38 in the loop connecting the second and third helix, distinctly away from the Ca-binding loops. The correlation coefficient is 0.56 for all points and 0.77 if Gly 20, 38, 57 and 59 (lower right) are excluded.

required for unusual backbone conformation in tight turns [e.g. backbone angles (Φ,Ψ) = (88,7), (79,15), (90,2) in degrees for Gly 20, Gly 57 and Gly 59 (family numbering) in parvalbumin]. Any non-glycine residues in or near the lower right-hand corner of such a contact/conservation plot are interesting candidates for further analysis: they are conserved for requirements other than the need to maintain side chain interactions.

The correlation does not crucially depend on the precise measure of side chain interactions used. For example, the atomic sidechain contact strength, i.e. the number of sidechain atoms of other residues in contact with a residue, correlates almost as well with conservation as the residue sidechain contact strength used here (the correlation coefficient drops slightly: from 0.56 to 0.49 for all points and from 0.77 to 0.63 if four Glys are excluded). The correlation is clearly less good, however, when the number of contacts is normalized by residue size (correlation coefficient drops from 0.56 to 0.29 for all points and from 0.77 to 0.44 if four Glys are excluded).

It appears that in this context the more relevant quantity is the total contact strength, not the fraction of possible contacts made by a residue. Large residues are more likely to make many contacts and are therefore more likely to be conserved for reasons of contact conservation.

Lesk and Chothia (1980) have made a related observation in the globins: 'For closely related globins, the homology of the buried and surface contact residues is always greater than that of the non-contact residues.' Similarly, in a study of immunoglobulin sequences, Padlan (1970) observed lower sequence variability for interior residues compared to exterior residues.

Conclusions

For the family of calcium binding proteins the set of highly conserved residues includes three subsets: (i) residues in sidechain contact with many other residues; (ii) ligand binding residues; (iii) glycines in ligand binding loops. These three sets overlap and the number of conserved residues outside of these sets is small.

The correlation between contact strength and residue conservation is an interesting observation based on statistical averages. This suggests that an important structural property influencing residue conservation is the strength of residue interactions in the folded structure, quantified, for example, as the number of nearest-neighbor sidechain contacts, and is consistent with the view that the folded structure is optimized with respect to residue–residue potential energy.

Interpreted as a rule, the correlation allows prediction of the minimal conservation (maximal mutability) of a residue, given its sidechain interprotein contact strength in a known structure (but not vice versa). This may be useful in choosing residues to be mutated in protein engineering. However, the ability to predict more precisely which residue exchanges at a particular position are compatible with a given structure is still lacking. We expect that further progress with this problem can be made by focusing on the energetics and structural properties of conserved residues in protein families for which at least one three-dimensional structure and many sequences are known.

Acknowledgements

We thank G.Boguta for suggestions and help with sequence data, Osnat Herzberg, Mike James and J.-P.Declerq for early communication of coordinate datasets, Amos Bairoch for sequence updates, Arthur Lesk, Susan Miller and Martin Vingron for comments and hints, Michael Scharf for contact maps (program CONAN), Anna Tramontano for help with graphics, and CPBR 3.13 and CPBP 1.06, and the BIOcomputing visitor's programme for support (A.G.).

References

Babu,S.Y,. Bugg,C.E. and Cook,W.J. (1988) *J. Mol. Biol.*, **204**, 191–204.
Bernstein,F.C. *et al.* (1977) *J. Mol. Biol.*, **112**, 535–542.
Bishop,M.J. and Rawlings,C.J. (eds) (1987) *Nucleic Acid and Protein Sequence Analysis: a Practical Approach.* IRL Press, Oxford.
Boguta,G. and Bierzynski,A. (1988) *Biophys. Chem.*, **31**, 133–137.
Declerc,J.-P. *et al.* (1988) *J. Mol. Biol.*, **202**, 349–353.
Devereux,J., Haeberli,P. and Smithies,O. (1984) *Nucleic Acids Res.*, **12**, 387–395.
Godzik,A. and Boguta,G. (1988) *Stud. Biophys.*, in press.
Gribskov,M. and Burgess,R.R. (1986) *Nucleic Acids Res.*, **14**, 6745–6763.
Herzberg,O. and James,M.N.G. (1985a) *Nature*, **313**, 653.
Herzberg,O. and James,M.N.G. (1985b) *Biochemistry*, **24**, 5298–5302.
Kabat,E.A. (1978) *Adv. Protein Chem.*, **32**, 1–75.
Kraulis,P.J. and Jones,T.A. (1987) *Proteins*, **2**, 188–201.
Kretsinger,R.H. (1980) *CRC Crit. Rev. Biochem.*, **8**, 119–174.
Kuntz,I.D. (1975) *J. Am. Chem. Soc.*, **97**, 4362–4366.
Lesk,A.M. and Chothia,C. (1980) *J. Mol. Biol.*, **136**, 225–270.
Lesk,A.M. and Chothia,C. (1982) *J. Mol. Biol.*, **160**, 325–342.
Lesk,A.M. and Chothia,C. (1986) *Phil. Trans. R. Soc. Lond.*, **A317**, 345–356.
Liebman,M., Venanzi,C.A. and Weinstein,H. (1985) *Biopolymers*, **24**, 1721–1758.
Nishikawa,K. and Ooi,T. (1974) *J. Theor. Biol.*, **43**, 351–374.
Moews,P.C. and Kretsinger,R.H. (1975) *J. Mol. Biol.*, **91**, 201–228.
Padlan,E.A. (1977) *Proc. Natl. Acad. Sci. USA*, **74**, 2551–2555.
Phillips,D.C. (1970) *Biochem. Soc. Symp.*, **31**, 11–28.
Ponder,J.W. and Richards,F.M .(1987) *J. Mol. Biol.*, **193**, 775–791.
Richards,F.M. and Kundrot,C.E. (1988) *Proteins*, **3**, 71–84.
Rossman,M.G. and Liljas,A. (1974) *J. Mol. Biol.*, **85**, 177.
Sander,C. and Schulz,G.E. (1979) *J. Mol. Evol.*, **13**, 245–252.
Satyshur,K.A. *et al.* (1988) *J. Biol. Chem.*, **263**, 1628–1647.
Schwartz,R.M. and Dayhoff,M.O. (1979) In Schwartz,R.M. and Dayhoff,M.O. (eds), *Atlas of Protein Sequence and Structure.* National Biomedical Research Foundation, Washington, DC, Vol. 5, Suppl. 3, pp. 353–358.
Szebenyi,D.M.E. and Moffat,K. (1986) *J. Biol. Chem.*, **261**, 8761–8777.
Wu,T.T. and Kabat,E.A. (1970) *J. Exp. Med.*, **132**, 211–250.

Received on February 16, 1989; revised on March 10, 1989