The sequence determinants of cadherin molecules

ALEXANDER E. KISTER,[1] MICHAEL A ROYTBERG,[2] CYRUS CHOTHIA,[3]
JURII M. VASILIEV,[4] AND ISRAEL M. GELFAND[1]

[1]Department of Mathematics, Rutgers University, Piscataway, New Jersey 08854, USA
[2]Institute of Mathematical Problems of Biology, RAS, Pushchino, Moscow Region 142292, Russia
[3]MRC Laboratory of Molecular Biology, Hills Road, Cambridge, CB2 2QH, United Kingdom
[4]Oncological Scientific Center of Russia, Moscow 115478, Russia

(RECEIVED August 29, 2000; FINAL REVISION January 23, 2001; ACCEPTED June 11, 2001)

Abstract

The sequence and structural analysis of cadherins allow us to find sequence determinants—a few positions in sequences whose residues are characteristic and specific for the structures of a given family. Comparison of the five extracellular domains of classic cadherins showed that they share the same sequence determinants despite only a nonsignificant sequence similarity between the N-terminal domain and other extracellular domains. This allowed us to predict secondary structures and propose three-dimensional structures for these domains that have not been structurally analyzed previously. A new method of assigning a sequence to its proper protein family is suggested: analysis of sequence determinants. The main advantage of this method is that it is not necessary to know all or almost all residues in a sequence as required for other traditional classification tools such as BLAST, FASTA, and HMM. Using the key positions only, that is, residues that serve as the sequence determinants, we found that all members of the classic cadherin family were unequivocally selected from among 80,000 examined proteins. In addition, we proposed a model for the secondary structure of the cytoplasmic domain of cadherins based on the principal relations between sequences and secondary structure multialignments. The patterns of the secondary structure of this domain can serve as the distinguishing characteristics of cadherins.

Keywords: Classic cadherins; cell adhesion molecules; method for protein family recognition; sequence comparison/classification

In the previous communications (Gelfand and Kister 1995, 1997; Chothia et al. 1998; Galitsky et al. 1998, 1999), we described a new method of sequence-structural analysis of protein families. This method permitted us to find the set of a few key residues in a sequence that will constitute an amino acid pattern of a given family. In this article, we apply this approach to determine defining characteristics of the cadherin family.

Cadherins are a group of proteins essential for the formation of stable specialized cell–cell contacts, that is, adherent contacts in various tissues, and therefore for organization of these tissues and organs. Cadherins are found in many types of animals ranging from nematodes to humans. Humans and other vertebrate animals have several classes of cadherins, each class being characteristic for a group of tissues (Takeichi 1991, 1995; Gumbliner 1996; Suzuki 1996; Gallin 1998; Shapiro and Colman 1999). For example, E-cadherins are specific for epithelial tissues, P-cadherins are found in placenta and other tissues, and N-cadherins are typical of neural and mesenchymal tissues.

The cadherin-like family comprises five subfamilies: classic cadherins types I and II, desmosomal cadherins, and protocadherins and cadherin-related proteins (Koch et al. 1999). In this work, we focus on the classic cadherins. The classic cadherins are transmembrane glycoproteins with five extracellular domains, a single membrane-spanning domain and a single cytoplasmic domain, which are linked to act in microfilaments via several linker proteins such as β-catenin and α-catenin. Cell–cell contacts are formed by homophilic

Reprint requests to: Alexander E. Kister, Department of Mathematics, Rutgers University, Piscataway, NJ 08854, USA; e-mail: akister@math.rutgers.edu; fax: 732-445-55-30.

Article and publication are at http://www.proteinscience.org/cgi/doi/10.1101/ps.37001.

adhesion of external N-terminal domains of cadherin molecules on the surface of one cell with the corresponding domains of cadherin molecules on another cell. Cadherin adhesion is calcium dependent. Within the extracellular region of cadherins, Ca^{2+} ions bind between domains to produce a rigid link part. In the absence of calcium, these domains display excessive motions relative to one another and stable adhesions cannot be formed.

The goal of this work to find the sequence determinants: the residues that occupy the conserved positions in classic cadherins. To describe the sequence determinants, we extend here the methods of sequence and structural analysis that were developed in our previous works (Gelfand and Kister 1995; Chothia et al. 1998). We show here that the sequence determinants can serve as patterns of the classic cadherins. A new method of identification of proteins that is based on the pattern recognition in sequences was suggested. Using this method, we were able to distinguish sequences of the classic cadherins in the SWISS-PROT database.

The currently known structures for the first and the second domains show that they have the same overall immunoglobulin-like fold (Shapiro et al. 1995; Overduin et al. 1995; Nagar et al. 1996; Pertz et al.1999). However, three-dimensional structures of the third, fourth, and fifth domains are unknown. The multialignment of the sequences of all five domains revealed the common conserved positions for extracellular part of the classic cadherins. Discovering the common sequence determinants supports the idea that the all extracellular domains share the immunoglobulin-like structure with the N-terminal domain.

In the second part of this work, we show the possibility of predicting the secondary structure of proteins based on the results of the sequence multialignment. We focus on the analysis of cytoplasmic part of cadherins whose X-ray structures are unknown. We based our research on the results of the sequence multialignment of these sequences. In fact, the multialignment of sequences of a protein family that have no strong homology forces one to make insertion and deletions to make sequences align. As a rule, these gaps in sequences correspond to a beginning or end of the secondary structural units: strands, helices, or loops. On the basis of this observation and of the results of sequence multialignment of the cytoplasmic part, we propose a model for the secondary structures of the cytoplasmic domains of cadherins.

Methods and Results

Classic cadherins: Extracellular domains

Secondary structural analysis of the first two domains

Three-dimensional structures have been determined for the N domains of murine neural cadherins (PDB files: 1NCG, 1NCH, 1NCI, 1NCJ, 2NCM; Shapiro et al.1995; Pertz et al. 1999) and for two domains of murine epithelial cadherins (PDB files: 1EDH, 1SUH, 3NCM; Overduin et al. 1995; Nagar et al. 1996; Jensen et al. 1999). Structural analysis revealed that sequences of these domains form sandwich-like structures with an immunoglobulin-like fold. Each domain consists of ~90–100 amino acids, which form seven β-strands. According to the accepted classification of the immunoglobulin fold, the seven successive strands are termed A′, B, C, D, E, F, and G, and the loops between them are named, respectively, A′B, BC, CD, DE, EF, EF′, and FG (Chothia and Jones 1997). Strands B, E, and D make up one sheet, and strands A′, C, E, and G make up another (Fig. 1).

On the basis of sequence alignments against the known structures, we determined secondary structures in the 37 classic cadherins from the SWISS-PROT database. They include the sequences of E−, N−, P−, R−, and other cadherins of various tissues and species, altogether 19 types of cadherins. The sequences of the N-terminal domains were divided into 15 fragments corresponding to the strands and loops (the loop between E and F strand is divided into two parts: EF and EF′) and a linker that connects the domains (Table 1).

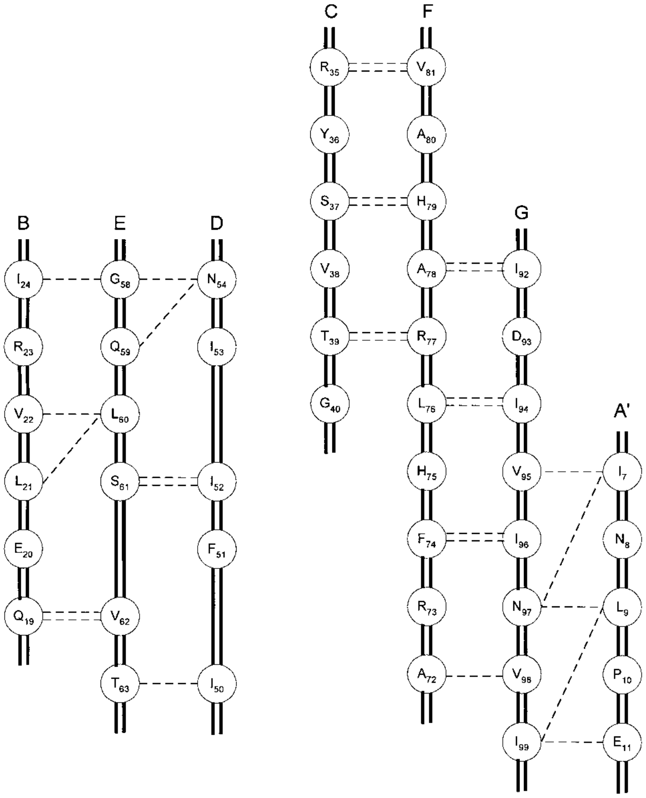


Fig. 1. Schematic representation of the strands in the N-terminal domain of 1NCI structure. A′, B, C, D, E, F, and G strands form two β-sheets (see text). Residues in the circles are shown with their number in the sequence. The dotted lines represent the hydrogen bonds between the main chain atoms.

Because the definition of secondary structure units usually is not very accurate, we strove to improve accuracy by performing a comprehensive multistep multialignment procedure that involved multialignment of structure superposition, as well as multialignment of residue–residue contacts, Cα coordinates, H bonds, and accessibility values (for details, see Gelfand and Kister 1995). As long as multialignments performed in several ways gives the same results, we can be retroactively assured that the division of sequences into secondary structure units was essentially accurate. Nonetheless, it is clear that one cannot be absolutely sure where the border between two secondary structure units lies. We therefore separately studied such borderline regions for the presence of conserved positions. Our analysis shows that conserved positions rarely, if ever, are to be found at the very periphery of strands or loops (Gelfand and Kister 1997). It appears therefore that lack of absolute precision in secondary structure definition has very little effect on the final result.

To classify the conservation of residues, we collected from the various structures all the amino acid fragments that correspond to each of the strands or loops. Alignment was conducted separately for each set of amino acid fragments that describe a particular strand or loop. In our approach, the amino acid sequences of the aligned fragments are given the term “word” (Gelfand and Kister 1995). From this align-

Table 1. *The secondary structures of N-terminal domains of cadherins*

Name	A'	A'B	B	BC	C	CD	D	DE	E	EF	EF'	F	FG	G	LINKER
	12345	1234567	123456	123456789	1234567	123456789	12345	1234	123456	12	12345678901	12345678901	123456789	12345678	123456789
E-CAD_C	ISCLE	NHRGPYP	MRLVQI	KSNKDKESK	VYYSITG	QGADSPPVG	IFIIE	RET	GWLEVT	EQ	EQLDREKI	DRYTLLSHAVS	ASGQPVEDP	MEIIITVM	DQNDNKPVF
E-CAD_H	ISCPE	NEKGPFP	KNLVQI	KSNKDKEGK	VFYSITG	QGADTPPVG	VFIIE	RET	GWLKVT	EP	EPLDRERI	ATYTLFSHAVS	SNGNAVEDP	MEILITVT	DQNDNKPEF
E-CAD_M	ISCPE	NEKGEFP	KNLVQI	KSNRDKETK	VFYSITG	QGADKPPVG	VFIIE	RET	GWLKVT	QP	QPLDREAI	AKYILYSHAVS	SNGEAVEDP	MEIVITVT	DQNDNRPEF
E-CAD_X	IIVSE	NEKGPFP	KRIVQI	KSSYAKEVK	VYYSITG	QGADTPPEG	VFAIG	RED	GWLNVT	RP	RPLDREAI	DNYVLFSHAVS	SNGANVEDP	MEIIIKVQ	DQNDNDPVF
EP-CAD_X	IKVSE	NERGPFP	KRLVQI	KSNKDRFNK	VYYSITG	QGADNPPQG	VFRIE	WET	GWMLVT	RP	RPLDREEY	DKYVLSSHAVS	ENGSPVEEP	MEITINVI	DQNDNRPKF
N-CAD_B	INLPE	NSRGPFP	QELVRI	RSDRDKNLS	LRYSVTG	PGADQPPTG	IFIIN	PIS	GQLSVT	KP	KPLDRELI	ARFHLRAHAVD	INGNQVENP	IDIVINVI	DMNDNRPEF
N-CAD_C	INLPE	NSRGPFP	QELVRI	RSDRDKSLS	LRYSVTG	PGADQPPTG	IFIIN	PIS	GQLSVT	KP	KPLDREQI	ASFHLRAHAVD	VNGNQVENP	IDIVINVI	DMNDNRPEF
N-CAD_H	INLPE	NSRGPFP	QELVRI	RSDRDKNLS	LRYSVTG	PGADQPPTG	IFIIN	PIS	GQLSVT	KP	KPLDREQI	ARFHLRAHAVD	INGNQVENP	IDIVINVI	DMNDNRPEF
N-CAD_M	INLPE	NSRGPFP	QELVRI	RSDRDKNLS	LRYSVTG	PGADQPPTG	IFIIN	PIS	GQLSVT	KP	KPLDRELI	ARFHLRAHAVD	INGNQVENP	IDIVINVI	DMNDNRPEF
N1-CAD_X	INVPE	NARGTFP	QELVRI	RSDRDKNLS	LRYSVTG	PGADQPPIG	VFIIN	PIG	GQLSVT	KP	KPLDREQI	ANFHLRAHAVD	VNGNQVENP	IDIVINVI	DMNDNRPEF
N2-CAD_X	INVPE	NARGTFP	QELVGI	RSDRDKSLS	LRYSVTG	PGADQPPLG	VFIIN	PIS	GQLSVT	KP	KPLDREQI	ATFHLRAHAVD	VNGNQVENP	IDIVINVI	DMNDNRPEF
P-CAD_H	ISVPE	NGKGPFP	QRLNQL	KSNKDRDTK	IFYSITG	PGADSPPEG	VFAVE	KET	GWLLLN	KP	KPLDREEI	AKYELFGHAVS	ENGASVEDP	MNISIIVT	DQNDHKPKF
P-CAD_M	IFVPE	NGKGPFP	QRLNQL	KSNKDRGTK	IFYSITG	PGADSPPEG	VFTIE	KES	GWLLLH	MP	MPLDREKI	VKYELYGHAVS	ENGASVEEP	MNISIIVT	DQNDNKPKF
P-CAD_B	IEIQE	GISTGEP	ICAYTA	RDPDKGSQK	ISYHILR	DPAGWLAM	GWLAM	PDS	GQVTAA	GV	GVLDREDEGFVRN	NIYEVMVLATD	DGSPPTTGT	GTLLLTLM	DINDHGPVP
R-CAD_C	INVPE	NSRGPFP	QQLVRI	RSDKDKEIH	IRYSITG	VGADQPPME	VFSID	PVS	GRMYVT	RP	RPMDREER	ASYHLRAHAVD	MNGNKVENP	IDLYIYVI	DMNDNRPEF
R-CAD_H	INVPE	NSRGPFP	QQLVRI	RSDKDNDIP	IRYSITG	VGADQPPME	VFSIN	SMS	GRMYVT	RP	MRPDREEH	ASYHLRAHAVD	MNGNKVENP	IDLYIYVI	DMNDNHPEF
R-CAD_M	INVPE	NSRGPFP	QQLVRI	RSDKDNDIP	IRYSITG	VGADQPPME	VFNID	SMS	GRMYVT	RP	RPMDREER	ASYHLRAHAVD	MNGNKVENP	IDLYIYVI	DMNDNRPEF
VE-CAD_H	MHIDE	EKNTSLP	HHVGKI	KSSVSRKN	AKYLLKG	EYVGK	VFRVD	AET	GDVFAI	ER	ERLDRENI	SEYHLTAVIVD	KDTGENLETP	SSFTIKVH	DVNDNWPVF
VE-CAD_M	MHIDE	EKNESLP	HYVKDQ	SNVNRQN	AKYVLQG	EFAGK	IFGVD	ANT	GNVLAY	ER	ERLDREKV	SEYFLTALIVD	KNTNKNLEQP	SSFTVKVH	DINDNWPVF
VE-CAD_P	MHIDE	EKNGSLP	HYVGKI	KSSVNHKN	TKYQLKG	ESAGK	VFRVD	ENT	GDVYAF	ER	ERLDREKI	PEYQLVALVVD	KNTEKNLESP	SSFTIKVH	DINDNWPVF
K-CAD_H	FLLEE	YTGSDY	QYVGKL	HSDQDRGDGS	LKYILSG	DGAGD	LFIIN	ENT	GDIQAT	KR	KRLDREEK	PVYILRAQAIN	RRTGRPVEPE	SEFIIKIH	DINDNEPIF
K_CAD_M	FFLLE	GYTGSDY	QYVGKL	HSDQDRGDGS	LKYILSG	DGAGD	LFIIN	EHT	GDIQAT	KR	KRLDREEK	PVYILRAQAVN	RRTGRPVEPE	SEFIIKIH	DINDNEPIF
K-CAD_R	FLLEE	YTGSDY	QYVGKL	HSDQDRGDGS	LKYILSG	DGAGD	LFIIN	ENT	GDIQAT	KR	KRLDREEK	PVYILRAQAIN	RRTGRPVEPE	SEFIIKIH	DINDNEPIF
10-CAD_C	FLLEE	YTGSDY	QYVGKL	HSDQDKGDGS	LKYILSG	DGAGT	LFIID	EKT	GDIHAT	RR	RRIDREEK	AFYTLRAQAIN	RRTLRPVEPE	SEFVIKIH	DINDNEPTF
N8-CAD_H	FVLEE	FSGPEP	ILVGRL	HTDLDPGSKK	IKYILSG	DGAGT	IFQIN	DVT	GDIHAI	KR	KRLDREEK	AEYTLTAQAVD	WETSKPLEPP	SEFIIKVQ	DINDNAPEF
N8-CAD_M	FVLEE	FSGPEP	ILVGRL	HTDLDPGSKK	IKYILSG	DGAGT	IFQIN	DIT	GDIHAI	KR	KRLDREEK	AEYTLTAQAVD	FETNKPLEPP	SEFIIKVQ	DINDNAPEF
OB-CAD_H	FVIEE	YTGPDP	VLVGRL	HSDIDSGDGN	IKYILSG	EGAGT	IFVID	DKS	GNIHAT	KT	KTLDREER	AQYTLMAQAVD	RDTNRPLEPP	SEFIVKVQ	DINDNPPEF
OB-CAD_M	FVIEE	YTGPDP	VLVGRL	HSDIDSGDGN	IKYILSG	EGAGT	IFVID	DKS	GNIHAT	KT	KTLDREER	AQYTLMAQAVD	RDTNRPLEPP	SEFIVKVQ	DINDNPPEF
T-CAD_C	ILIPE	NQRPPFP	RSVGKV	IRSEGTEG	AKFRLSG	KGVDQDPKG	IFRIN	EIS	GDVSVT	RP	RPLDREAI	ANYELEVEVTD	LSGKIIDGP	VRLDISVI	DQNDNRPMF
T-CAD_H	ILIPE	NQRQPFP	RDVGKV	VDSDRPER	SKFRLTG	KGVDQEPKG	IFRIN	ENT	GSVSVT	RT	RTLDREVI	AVYQLFVETTD	VNGKTLEGP	VPLEVIVI	DQNDNRPIF
M-CAD_H	ISVSE	NHKRLP	YPLVQI	KSDKQQLGS	VIYSIQG	PGVDEEPRG	VFSID	KFT	GKVFLN	AM	AMLDREKT	DRFRLRAFALD	LGGSTLEDP	TDLEIVVV	DQNDNRPAF
M-CAD_M	ISVSE	NHKRLP	YPLVQI	KSDKQQLGS	VIYSIQG	PGVDEEPRN	VFSID	KFT	GRVYLN	AT	LDREKT	DRFRLRAFALD	LGGSTLEDP	TDLEIVVV	DQNDNRPAF
B-CAD_C	VPENE	RGPFP	KNLVQI	KSNRDREAK	IFYSITG	QGADAPPEG	IFTIE	KET	GWMKVT	QP	QPLDREHI	NKYHLYSHAVS	ENGKPVEEP	MEIIVTVT	DQNDNKPQF
B_CAD_X	VSENE	RGPFP	KRLVQI	KSNKEKLSK	VFYSITG	QGADTPPEG	IFRIE	KET	GWMQVT	RP	RPLDREEY	EKYVLLSHAVS	ENGASVEEP	MEITVTVI	DQNDNRPKF
LI-CAD_R	FSIFE	GQEPS	QIIFQF	KANPPA	VTFELTG	ETDG	IFKIE	KD	GLLYHT	RV	RVLDRETR	AVHHLQLAALD	SQGAIVDGP	VPIIIEVK	DINDNRPTF
BR-CAD_H	FVLEE	YVGSEP	QYVGKL	HSDLDKGEGT	VKYTLSG	DGAGT	VFTID	ETT	GDIHAI	RS	LDREEK	PFYTLRAQAVD	IETRKPLEPE	SEFIIKVQ	DINDNEPKF
14-CAD_H	FVLEE	HMGPDP	QYVGKL	HSNSDKGDGS	VKYILTG	EGAGT	IFIID	DTT	GDIHST	KS	LDREQK	THYVLHAQAID	RRTNKPLEPE	SEFIIKVQ	DINDNAPKF

Database name of each sequence is given in the first column (E_CAD_C, etc.), while secondary structure units are referenced by letters in topmost row (A′, A′B, etc.). Amino acid sequences of cadherins are given in rows. Secondary structure units are separated by one or more spaces. Numbers in second row (1,2,3, . . .) refer to position number of amino acid within strand or loop.

The sequences of all cadherins are extracted from SWISS-PROT database.

The names of cadherins are given according to SWISS-PROT identification: E-cad_C, E-cad_H, E-cad_M, E-cad_X are E-cadherins of the chicken, human, mouse, xenla, respectively; E-cad_H; CadF-X, EP cadherin xenla; N-cad_B, N-cad_C, N-cad_H, N-cad_M are N-cadherins from cow, chicken, human and mouse, respectively; N1-cad_X is N-cadherin 1 of xenla and N2-cad_X is N-cadherin 1 of the xenla; Pcdα2, Pcdα2, and Pcdα2, human neural cadherins; P-cad_H, P-cad_M, P-cad_B, P-cadherins from human, mouse, and bovine species, respectively; R-cad_C, R-cad_H, R-cad_M, R-cadherins (retinal) of the chicken, human, and mouse, respectively; VE-cad_H, VE-cad_M, VE-cad_P, vascular endothelial–cadherins of human, mouse, and pig, respectively; K-cad_H, K-cad_M, K-cad_R, cadherin-6, kidney cadherin of the human, mouse, and rat, respectively; 10-cad_C, cadherin-10 of the chicken; N8-cad_H, N8-cad_M, cadherin 8 of the human and mouse; OB-Cad_H, OB-Cad_M, osteoblast cadherin of the human and mouse; T-cad_C, T-cad_C, cadherin 13 of the chicken and human; M-Cad_H, muscle-cadherin of the human; M-cad_M, muscle cadherin of the mouse; B-cad_C and Bcad_X, blastomere-cadherin of the chicken and xenla; LI-cad_R, liver intestine cadherin of the ratl; 14-cad_H, cadherin 14 of the human; BR-cad_H, brain cadherin of the human.

ment, each residue in a sequence is assigned to a position in a word. Residues in sequences are referred to by an index that contains the letter code of the word and its position therein. For example, A′1 is the address of the first residue in the A′ word. Describing residues with the two-part index gives us a common system of numbering for various cadherin sequences. It allows us to compare residue occupation in each position for various sequences and determine residue conservation at all positions.

Residue conservation: Patterns of strands and loops of the N-terminal domain

The first step toward defining characteristic patterns of cadherin strands and loops consists of the analysis of residue frequencies at all positions of words. This analysis reveals the nature and extent of residue conservation at each position. After the classification of residue conservation in immunoglobulins suggested in our previous article (Chothia et al. 1998), we divided residues into three groups: (1) V, L, I, M, A, F, W, and C; (2) R, K, E, D, Q, and N; and (3) P, H, Y, G, S, and T. This classification is based on two properties: hydrophobicity and the tendency to be on the surface or in the interior of a protein.

Inspection of residue frequencies showed that six positions are occupied by a single residue in almost all sequences, and 23 have only a few chemically similar residues from the same group (Table 2). For example, E residue is found at the position A′5 in all known cadherin sequences (Table 1). These 29 positions are considered to be the conserved positions. The other ~66 positions in sequences are variable. They can be occupied by residues from various groups.

These data show that all words that describe the strands and EF′ loop have several conserved positions. Residues at these positions constitute a pattern of the word. Analysis of, for example, the set of B words in the first domain (Table 1) shows that in all sequences position 3 and position 6 are occupied by hydrophobic and aromatic residues, which are assigned to group 1, and the polar and charged residues from group 2 were found at position 5 (Table 2). Thus, the residues at the conserved positions B3, B5, and B6 constitute the pattern of the word B in the Domain I (Table 2). As shown below, the patterns of words can serve as a useful tool for identifying cadherin sequences and for their structural predictions.

Secondary and three-dimensional structure prediction for five extracellular domains

For most molecules in the cadherin family, the three-dimensional structure is unknown. However, for these proteins it is possible to make secondary structure predictions for all extracellular domains based on the knowledge of the patterns of words in the first two domains. To determine secondary structures of cadherin chains in all domains, we have matched the patterns of the domains I and II with the sequences of the domains III, IV, and V. The result of this analysis showed that the patterns of the N-terminal domains fit with the sequences of all domains. It allowed us to divide the sequences of the domains III, IV, and V into the words. Because words describe secondary structural units, dividing a sequence of amino acids into words permits us to predict the secondary structure of a protein.

Because the alignment of cadherins was based on both sequence and structural information, it follows that residues at the identical positions of the words have the same structural role in various molecules. Analysis of the structural role of residues involves determining residue–residue interactions, residue exposure on the surface, and their coordinates in the system of coordinates unified for a given protein family. We can use for this preferred coordinate system, for example, the coordinate system of any of known structure of the cadherin molecule. Thus, it is possible to identify coordinates of residues for extracellular domains of all analyzed cadherins. We suppose that the Cα atoms of the residues at the same positions of the words in various domains can be superimposed on each other.

Conserved positions in the strands and loops in five extracellular domains

Inspection of the sequences of different cadherins shows that the nature of residues and extent of conservation varies greatly at various positions. For example, comparison of the sequences of human E- and K-cadherins shows that in domain I ~32% of the residues are identical. Domains I and II of E-cadherins share only 25% identity. In comparison to the sequences of domains I and II the sequences of domains III, IV, and V show no significant similarity (<20%).

The alignment of the words allowed us to calculate the frequency of residues at every position in the words. Analysis of the residue frequency in the various domains showed that there are no positions that are occupied by a single type of residue in all domains. However, there are many positions where residue conservation was found in one or several domains but not in all five domains. For example, position A′5 is occupied by Glu in all sequences of domains I, II, and III, whereas in the sequences of domains IV and V Glu shares this position with Gln and Asp residues. Residues at the A′1 position are hydrophobic in all sequences of the first domain whereas in the second domain Gly and Ala are the most common residues. The D1 position can be considered as a conserved hydrophobic position in the first domain and conserved hydrophobic and aromatic position in domains II and IV, but a variable position in domains III and V. The residue conservation in the fifth domain differs in many cases from residue variations in the other domains.

The residues at the conserved positions for all strands and EF′ loops in five extracellular domains are presented in Table 2. The comparison of the conserved positions in vari-

Table 2. *Patterns of the strands and EF′ loop of the five extracellular domains*

Units	Positions	DOMAINS					Common Patterns
		I	II	III	IV	V	
A'	1	I F V M Y	G A	*	*	*	*
	2	*		*	*	*	*
	3	L V I C	I V	L V I	L V I	L V I C F	L V I C F
	4	*	*	*	*	*	*
	5	E	E	E	E Q	E D Q	E D Q
B	1	*	*	*	*	*	*
	2	*	*	*	*	*	*
	3	L V I	V L	L V I	L V I	L V I	L V I
	4	*	*	*	*	*	*
	5	R K Q D	*	D N R K S	*	*	*
	6	I L V F	V M L A	L V I	L V I F	L V I	I L V F M A
C	1	V I L A	L V I	A I V M	L V I	F Y I V	V I L A M F Y
	2	*	*	*	*	*	*
	3	Y F L	Y F	Y F	Y F	F Y	Y F L
	4	*	*	*	*	*	*
	5	Y V I	I L	I M L	L V I	L I V A	Y V I L A
	6	*	*	*	*	*	*
	7	*	*	*	*	*	*
D	1	L V I	M L Y	*	*	*	*
	2	F	F	F	L F Y V	F V W	F V F W Y
	3	*	*	*	*	*	*
	4	I V	I V	L V I M	L V I M	I V A L	I V M L A
	5	*	*	*	*	T E Q R H	*
E	1	G	G A	*	G S	*	*
	2	*	*	*	*	*	*
	3	L I V M	I	L V I	I V	L I V M	L I V M
	4	*	*	*	*	*	*
	5	V A L	*	*	*	*	*
	6	*	L V I A	*	*	*	*
EF'	1	L M I	L M	L V I M	L	L F Y	L M F Y I V
	2	D	D	D	D	*	*
	3	R	R	Y F	R	*	*
	4	E	E	E	E	*	*
	5	*	*	*	*	*	*
	6	*	*	*	*	*	*
F	1	*	*	*	*	*	*
	2	*	*	*	*	*	*
	3	Y F C	Y	Y F L	Y I	Y F L	Y F I L C
	4	*	*	*	*	*	*
	5	L I	L V	L V I	A L V I	L V I F	L V I F A
	6	*	*	*	*	*	*
	7	*	V I L	V I	I F V A L	V L I A	*
	8	*	Q E	*	*	*	*
	9	A V I	A V	*	*	*	*
	10	V I L	*	*	*	*	*
	11	*	D	*	D E K N	*	*
G	1	S I M V T F	*	*	*	*	*
	2	*	*	*	*	*	*
	3	I F L V	A V L	V I	V L	L V I	I F L V A
	4	*	*	*	*	*	*
	5	I V	I	L V I	L V I	L V I A M	I L V A M
	6	*	*	*	*	*	*
	7	V I	V L	L V I	L V I	*	*
	8	*	*	*	*	*	*

Each cadherin molecule is composed of five domains. Patterns of strands or loops of the first domain can be read off in top-to-bottom direction in the first of the five broad columns (Roman numeral). The letters in the Units column (A′, B . . .) refer to names of the words that make up each domain; numbers in Positions column refer to position within the words. Corresponding row contains amino acids that are commonly found at particular position within the words. Thus, position A′1 of the I domain is commonly occupied by residues I, F, V, M, and Y, and the positions A′2 and A′4 are the variable positions (marked off by *). Patterns of the words of the II, III, IV, and V domains can be read off in an analogous fashion from I, II, etc., broad columns. (E.g., residues E, D, and Q are common at position A′5 of the V domain.) Positions occupied by residues from same amino acid group (described in text) in all five domains constitute the Common Patterns of the extracellular domains.

ous domains revealed 15 extracellular conserved positions. All positions except one are occupied by hydrophobic residues in all five domains. The polar and charged residues are found at A′5 position.

Buried and surface positions in cadherins

The role of residues at each position was determined from the examination of their accessible surface areas. To give an overview of the positions of residues, we calculated the accessible surface area (ASA) of residues in three structures: domain 1 of N-cadherins and domains 1 and 2 of E-cadherins (Table 3). ASA are divided into 0, 1, 2, . . ., 9 groups, where 0 indicates ASA in the range 0–9 Å^2, 1 indicates 10–19 Å^2, etc. Residues at 12 positions in all structures are buried in the protein interior (ASA are calculated in the range 0–2). Eight of these positions (A′3, B6, C3, D4, E3, EF′′1, F3, F5) are hydrophobic and aromatic conserved positions at the center of the structure.

Method of attributing a protein to a protein family by using patterns

Discovering a small set of key residues that furnishes us with the amino acid patterns for the structural units in a the protein family allows us to develop a computer algorithm for classification of proteins.

To assign a query sequence to its proper protein family, we need to find a match between residues at positions in the query sequences and the residues in the patterns of the words of family members. In fact, we need not know residues at all positions in the query sequence. The advantage of our approach is that it allows one to find a few of the class-determining positions that uniquely determine a family. We developed a new approach for assigning a protein to a protein family, which we applied for identification of classic cadherins.

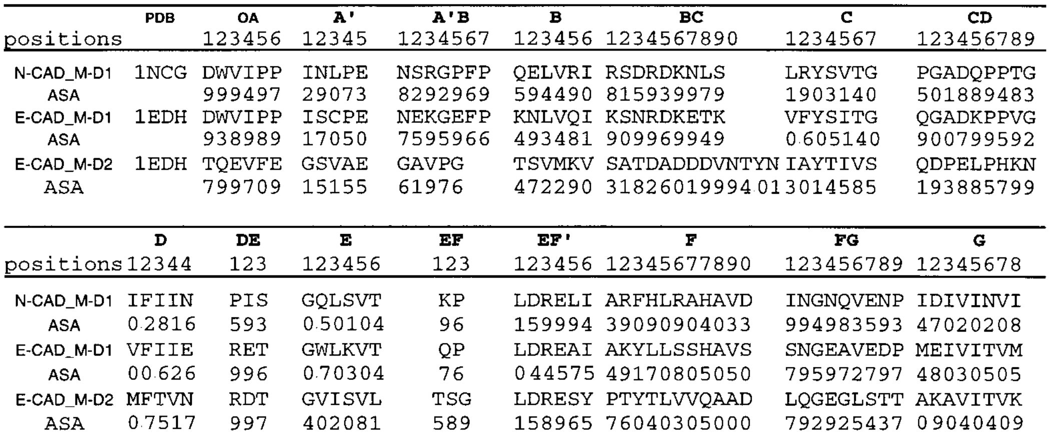
Algorithm

A sequence in a protein family can be defined in terms of an ordered set of patterns of words. For each pattern of a word the following are determined: (1) number of positions in a given word; (2) conserved positions and the various sets of residues that can occupy these positions; (3) interval (a possible range of residues) to the next word in the sequence.

In the search procedure, we matched the patterns of words with a query sequence. To check it, we implemented an algorithm based on appropriate modification of the dynamic programming. The algorithm of the method is the following: patterns of all or several secondary structural units are matched with a query sequence in consecutive order, starting from the first pattern. First, we pick out those sequences of the database that contain a fragment that fits one of the known basic patterns describing the first (A′) fragment of cadherins. Then, we again search out the entire database, this time using patterns for B fragment as our query patterns, and selecting sequences containing one of the B patterns. We continue this procedure with patterns of other words.

Results of the analysis are formulated in the following way: how many words (more precisely: fragments describable by cadherin patterns) are found in a given sequence. If in a sequence in question fragments are found that match with patterns of all, or almost every, cadherin word, then that sequence is considered to belong to the cadherin family.

Table 3. *Structural alignments of the sequences and the residue accessible surface areas of the cadherin domains*



	PDB	OA	A'	A'B	B	BC	C	CD
positions		123456	12345	1234567	123456	1234567890	1234567	123456789
N-CAD_M-D1	1NCG	DWVIPP	INLPE	NSRGPFP	QELVRI	RSDRDKNLS	LRYSVTG	PGADQPPTG
ASA		999497	29073	8292969	594490	815939979	1903140	501889483
E-CAD_M-D1	1EDH	DWVIPP	ISCPE	NEKGEFP	KNLVQI	KSNRDKETK	VFYSITG	QGADKPPVG
ASA		938989	17050	7595966	493481	909969949	0.605140	900799592
E-CAD_M-D2	1EDH	TQEVFE	GSVAE	GAVPG	TSVMKV	SATDADDDVNTYN	IAYTIVS	QDPELPHKN
ASA		799709	15155	61976	472290	31826019994.01	3014585	193885799

	D	DE	E	EF	EF'	F	FG	G
positions	12344	123	123456	123	123456	12345677890	123456789	12345678
N-CAD_M-D1	IFIIN	PIS	GQLSVT	KP	LDRELI	ARFHLRAHAVD	INGNQVENP	IDIVINVI
ASA	0.2816	593	0.50104	96	159994	39090904033	994983593	47020208
E-CAD_M-D1	VFIIE	RET	GWLKVT	QP	LDREAI	AKYLLSSHAVS	SNGEAVEDP	MEIVITVM
ASA	00.626	996	0.70304	76	044575	49170805050	795972797	48030505
E-CAD_M-D2	MFTVN	RDT	GVISVL	TSG	LDRESY	PTYTLVVQAAD	LQGEGLSTT	AKAVITVK
ASA	0.7517	997	402081	589	158965	76040305000	792925437	09040409

In the row ‘positions’ the number of residues at the positions of A′, A′B, B, BC, C, CD, D, DE, E, EF, EF′, F, FG, and G strands and loops are shown, N-cad_M-D1, sequence of the mouse N cadherin in the domain 1; E-cad_M-D1 and E-cad_M-D2, sequence of the E cadherin in the domains 1 and 2, respectively. Accessible surface areas of residues are given as 0, 1, 2, . . . and 9 where 0 indicates accessible surface areas in the range 1–9 Å^2, 1 area in the range 10–19 Å^2, 2 areas in the range 20–29 Å^2, . . ., and 9 areas greater than 90 Å^2.

Results of the analysis of sequences in SWISS-PROT database

We used patterns of eight words (A′, B, C, D, E, EF′, F, and G) of the first domain of classic cadherins in the search procedure (Table 2). These patterns are presented in Table 2. The goal of this test is to show that these patterns are sufficient to identify the classic cadherins. We analyzed the sequences in SWISS-PROT (release 38 with 79,909 entries). The results of the analysis are presented in Table 4. Thirty sequences were found to contain all eight cadherin patterns, that is, there are eight fragments within these sequences that sequentially match with A′, B, C, D, E, EF′ F, and G patterns (the first row in the table). According to the description in SWISS-PROT, all of these proteins are classic cadherins.

Six sequences were found to contain seven cadherin patterns, that is, one of the patterns of words was not found in the sequences (the second row in Table 4). For example, the analysis of the VE-CAD_M sequence (Table 1) showed that the patterns of seven words; all except B word match with the sequence. (No fragment corresponding to B word was observed because the position B6 is occupied by Q residue and does not match with the conserved hydrophobic position in the pattern of B word.) According to the description in SWISS-PROT, five of six found sequences are classic cadherins and one protein is a noncadherin protein. In row 3, it is shown that seven sequences match with the patterns of exactly six words. It was found that two of these sequences are classic cadherins. Analysis of the sequences where 5, 4, 3, 2, 1, and no cadherin words were found showed that all of these proteins are not classic cadherins.

In total, there are 43 sequences (30 + 6 + 7) in which the patterns of at least six words were found. Thirty-seven of these proteins are classic cadherins. Six other proteins in which at least six or seven patterns were found can be called false-positive. These proteins are identified in SWISS-PROT as desmogleins and desmocollins. They are not classic cadherins but belong to cadherin family. These proteins have sequence homology with classic cadherins. However, the patterns of the classic cadherins developed in this work mainly allow us to distinguish the classic cadherins from other cadherin-like proteins.

Table 4. *The numbers of sequences where cadherins' words are found*

Words	The number of sequences	Classical cadherin	Non-cadherin sequences
8	30	30	0
7	6	5	1
6	7	2	5
5	7	0	7
4	88	0	88
3	1375	0	1375
2	36659	0	36659
1	37547	0	37547
0	4190	0	4190

Words, the numbers of the cadherins' words discovered in the sequences. The number of sequences, the number of sequences in SWISS-PROT, where a given number of cadherins' words are found. Classical cadherin, the number of classical cadherins' sequences where a given number of words are found. Non-cadherin sequences, the number of non-cadherin proteins where a given number of words are found (see the text).

Thus, the result of a search of the cadherin sequences shows that patterns at least of six words allow us to find all classic cadherins in the database. It gives us a new tool for identifying of proteins. Thus, if the patterns of eight, seven, or six words are observed in a sequence in question, then there is a great probability that the sequence is a classic cadherin. Because in total there are 27 conserved positions in the patterns of eight words, we can classify a protein sequence if we know residues at no more than 27 conserved positions.

Comparison of secondary structural units with the results of sequence multialignment

The comparison of sequence and structural multialignment shows that the gaps (deletions and insertions) in the sequences are almost never found in the middle of the strands or helices but at the borders. This observation could help us to predict a secondary structure for proteins with unknown three-dimensional structure. Consider, for example, the sequence multialignment. We present the results of the multialignments for seven cadherin sequences of the I domains in Table 5. Sequence multialignment shows the sequences to be divided into 10 ungapped fragments. For example, there are two ungapped fragments at the beginning of E-cadherin of the xenla (E-CAD_X) sequence: VSENE (fragment 1) and KGPFP (fragment 2). In such manner the sequences were divided into 10 fragments (Table 5a).

The comparison of the sequence multialignments with the secondary structures of these proteins obtained from the analysis of three-dimensional structures (Table 5b) shows that most fragments and secondary structural units coincide. In fact, in E-CAD_X sequence the fragment VSENE corresponds to the A′ strand and KGPEP residues correspond to A′B loop (Table 5). This relationship – sequence ungapped fragment and secondary structural units, are observed for all fragments except fragments 7, 8, and 10. Fragment 7 corresponds to strand D and loop DE together and fragment 8 corresponds to strand E and loop EF′, whereas fragment 10 involves FG loop and G strand. Thus, there is a strong relationship between sequence multialignments and the secondary structures of cadherins.

It is obvious that the greater the number of sequences we consider for multialignment, the greater the accuracy in predicting the secondary structure. The classic cadherins give us a good example of this. We have analyzed 37 sequences,

Table 5. *The comparison of the sequence and secondary structural multialignments*

a

Names	fragments 1	2	3	4	5	6	7	8	9	10
E-CAD_M	ISCPE	NEKGEFP	KNLVQI	_KSNRDKETK	VFYSITG	QGADKPPVG	VFIIERET	GWLKVTQPLDREAI______	AKYILYSHAVS	_SNGEAVEDPMEIVITVT
E-CAD_X	VSENE	__KGPFP	KRIVQI	_KSSYAKEVK	VYYSITG	QGADTPPEG	VFAIGRED	GWLNVTRPLDREAI______	DNYVLFSHAVS	_SNGANVEDPMEIIIKVQ
VE-CAD_P	MHIDE	EKNGSLP	HYVGKI	_KSSVNHKN_	TKYQLKG	___ESAGK_	VFRVDENT	GDVYAFERLDREKI______	PEYQLVALVVD	KNTEKNLESPSSFTIKVH
R-CAD_M	INVPE	NSRGPFP	QQLVRI	_RSDKDNDIP	IRYSITG	VGADQPPME	VFNIDSMS	GRMYVTRPMDREER______	ASYHLRAHAVD	_MNGNKVENPIDLYIYVI
T-CAD_H	ILIPE	NQRQPFP	RDVGKV	VDSDRPER__	SKFRLTG	KGVDQEPKG	IFRINENT	GSVSTRTLDREVI______	AVYQLFVETTD	_VNGKTLEGPVPLEVIVI
LI-cad_R	FSIFE	GQEPS__	QIIFQF	_KANPPA___	VTFELTG	_____ETDG	IFKIEKD_	GLLYHTRVLDRETR______	AVHHLQLAALD	_SQGAIVDGPVPIIIEVK
P-CAD_B	IEIQE	GISTGEP	ICAYTA	RDPDKGSQK	ISYHILR	DPAGWLAM	GWLAMPDS	GQVTAAGVLDREDEGFVRN	NIYEVMVLATD	_DGSPPTTGTGTLLLTLM

b

Names	secondary structural units A'	A'B	B	BC	C	CD	D	DE	E	EF	F	FG	G
E-CAD_M	ISCPE	NEKGEFP	KNLVQI	KSNRDKETK	VFYSITG	QGADKPPVG	VFIIE	RET	GWLKVT	QPLDREAI	AKYILYSHAVS	SNGEAVEDP	MEIVITVT
E-CAD_X	VSENE	KGPFP	KRIVQI	KSSYAKEVK	VYYSITG	QGADTPPEG	VFAIG	RED	GWLNVT	RPLDREAI	DNYVLFSHAVS	SNGANVEDP	MEIIIKVQ
VE-CAD_P	MHIDE	EKNGSLP	HYVGKI	KSSVNHKN	TKYQLKG	ESAGK	VFRVD	ENT	GDVYAF	ERLDREKI	PEYQLVALVVD	KNTEKNLESP	SSFTIKVH
R-CAD_M	INVPE	NSRGPFP	QQLVRI	RSDKDNDIP	IRYSITG	VGADQPPME	VFNID	SMS	GRMYVT	RPMDREER	ASYHLRAHAVD	MNGNKVENP	IDLYIYVI
T-CAD_H	ILIPE	NQRQPFP	RDVGKV	VDSDRPER	SKFRLTG	KGVDQEPKG	IFRIN	ENT	GSVSVT	RTLDREVI	AVYQLFVETTD	VNGKTLEGP	VPLEVIVI
LI-cad_R	FSIFE	GQEPS	QIIFQF	KANPPA	VTFELTG	ETDG	IFKIE	KD_	GLLYHT	RVLDRETR	AVHHLQLAALD	SQGAIVDGP	VPIIIEVK
P-CAD_B	IEIQE	GISTGEP	ICAYTA	RDPDKGSQK	ISYHILR	DPAGWLAM	GWLAM	PDS	GQVTAA	GVLDREDEGFVRN	NIYEVMVLATD	DGSPPTTGT	GTLLLTLM

Names, the names of the sequences (see the footnote of Table 1). The gaps in the sequence multialignments are shown by '_'
(a) Results of the sequence multialignments. The sequences are divided into fragments: 1, 2, 3, 4, 5, 6, 7, 8, 9, and 10.
(b) Results of the secondary structural multialignments. The sequences are divided into secondary structural units: A′, B, C, D, E, F, and G strands, and A′B, BC, CD, DE, EF, and FG loops.

involving 14 types of cadherins (Table 1). Thus, we propose that the results of sequence multialignment gives a reliable basis to predict secondary structure.

Sequence multialignment gives important information about three-dimensional structure as well. Residues of molecules that are aligned with each other have approximately the same structural characteristics, such as H bonds between main chain atoms, approximately the same residue–residue of contacts, or equal values of accessibility. This observation has been made in the analysis of proteins (see, e.g., Lesk et al. 1987)

Classic cadherins: Cytoplasmic part

In this part, we describe the result of our investigation of the cytoplasmic domain of cadherins. Currently, there is no structural information about the intracellular domains. We analyzed amino acid sequences of 36 cytoplasmic domains. They consist of ~120 amino acids. Because we have found the relationship between sequence and structural alignment for the extracellular domains, it is likely that the sequence alignment can give some information about secondary structures of the cytoplasmic domains. The mutialignment of 36 sequences resulted in 14 ungapped fragments. We can speculate that these fragments correspond to some extent to the helices or strands and loops in this part of cadherins.

The residue frequency was calculated at each position of the sequences. It was found that 71 of ~120 positions are occupied by only one residue or very similar residues in all or almost all sequences (Table 6). This observation shows that unlike the extracellular domain, the cytoplasmic part is characterized by a high degree of residue conservation. Twenty-four positions are occupied by hydrophobic and/or aromatic residues. The polar and charged amino acids are found in 26 positions, and hydrophilic and neutral residues are found in 21 positions. The conserved positions are mostly found near the N and C termini in sequences. Fragments 4 and 14 have the most conserved positions (13 and 18 positions, respectively), whereas the ungapped fragments in the middle of the cytoplasmic part (fragments 6, 7, 8, 9, and 10) have one conserved positions in each fragment. (Note that residues in fragment 4 are involved in binding with β-catenin.)

On the basis of the analysis of the extent of conservation, we determined the amino acid patterns for each fragment (Table 6). We expected patterns of several long fragments to be characteristic of the cytoplasmic part. For example, there are 18 conserved positions in fragment 14. To test our suggestion that the pattern of a single fragment is sufficient for cadherin recognition, we used the pattern matching method that we developed for analysis of the extracellular part. The patterns of fragments 4, 5, 11, 12, 13, and 14 were matched separately with the sequences of the SWISS-PROT database. The results of the analysis showed that the pattern of just one fragment, either 4 or 12 or 14, can be used for identification of the cadherins (Table 7).

Discussion

To find reasonable criteria for classification of proteins into families, one needs to find invariant characteristics that are shared by all members of the family. Traditional tools for sequence classification use different methods of align-

Table 6. *The most common residues in the cytoplasmic domain*

```
fragments |  1  |    2    |  3    |            4           |    5
          |     |         |       |          1         2 |
position  |12345| 12345678| 123456| 12345678901234567890123| 1234567
residues  |RRR**| R*******| **L***| **D*RDNII*Y*E*GGGE*D***| YDLS*LH
           K K    K           I       E HEQVL   D            F IT  Q
                                   V                   VG  R
                                   F                    A  N
                                   Y

fragments |   6   |     7      |     8      |     9     |     10
          |       |          1 |          1 |          1|          1
position  |123456 | 12345678901| 12345678901| 1234567890| 1234567890123
residues  |****** | ***********| P**********| P*********| P************

fragments |         11           |        12
          |         1         2 |          1
position  |12345678901234567890 | 12345678901234567
residues  |EI**FI**K****D*D***P | PYDSL**Y*YEG**S*A
           DM   M  R                  TI  F
                   D
                   E

fragments |     13       |             14
          |         1   |          1         2         3
position  |123456789012 | 1234567890123456789012345678901
residues  |*SLSSL*S**S* | D*DYDYL*DWG*RFK*LADMYG*********
                  T  T    E N NF   E         EL S
```

Fragments, the number of the ungapped fragments. Position, the positions of the residues in the fragments. Residues, the most common residues are shown at the conserved positions. The variable positions are marked by *.

ment: BLAST, FASTA, HMM, and others require one to know all or almost all residues in sequences (Smith and Waterman 1981; Eddy 1996; Pearson 1996; Altschul et al. 1997; Gusfield 1997). Another method used for dividing proteins into families in the Prosite database (Hofmann et al. 1999) identified specific sites of conserved regions in protein families.

We propose another approach for classification of protein families. An essential feature of the method is that it combines sequence and structural data. Putting together the results of the sequence and structural multialignments, we are able to give a description of the major structural units in a protein family. Patterns of strands and loops serve as defining characteristics of a protein family. In this work, we applied this method to one particular protein family, cadherins. The results of this analysis showed that, in fact, on the basis of defining characteristics one could unequivocally select all members of the cadherin family from ~80,000 proteins. Qualitatively specific patterns are characteristics of both the extracellular and the cytoplasmic domains. We can use independently the patterns of any of these parts. Notably, the sequence of the cytoplasmic tail is especially specific: the pattern of one unit is sufficient to determine a family. In contrast, patterns of transmembrane parts cannot assign proteins to a proper family, because they were found in >2000 proteins. These results confirm that defining patterns can be successfully used for reliable assignment of proteins to a proper protein family. We plan to expand the investigation of defining characteristics of protein families of the β fold.

In this work, we found that the gaps in sequences of cadherins obtained as the result of insertions and deletions in the sequence multialignment divide the sequences into the structural units (strands and loops). Thus, sequence mul-

Table 7. *Numbers of sequences where patterns of the fragments of the cytoplasmic part are found*

Fragment	Sequence	Cadherin
4	34	34
5	5565	34
11	52	34
12	34	34
13	363	34
14	34	34

Fragment, the number of the ungapped fragment. Sequence, the number of sequences in the SWISS-PROT database where a given pattern is found. Cadherin, the number of cadherin sequences where a given pattern is found. It is shown, for example, that a fragment that corresponds to the pattern of the sequence fragment #5 in the cytoplasmic domain was found in 5565 sequences; 34 of them are cadherins.

tialignments may give us a clue about secondary structure. The assignment of sequence units to a secondary structure has, however, some limitations. The multialignment of sequences with homology results in long ungapped fragments that include several structural units. To obtain a more reliable secondary structural assignment in the protein family, we need to use as many diverse sequences as possible. In our further analysis of other protein families, we plan to test the hypothesis about relationship between the sequence and structural alignments.

Acknowledgments

Authors are grateful to Dr. P. Ehrlich for critical review of the manuscript. The authors acknowledge the assistance of L. Pogost in performing computer calculations. We acknowledge with deep gratitude the support of the Gabriella and Paul Rosenbaum Foundation and also thank M. Goldman for continuous encouragement. A.E.K. is supported by the Gabriella and Paul Rosenbaum Foundation. M.A.R. was supported by grants from Russian Fund of Basic Research (Grants 00-04-48246 and 00-07-90037), Russian State Scientific Program Human Genome (Grant 1gc/00), INTAS (Grant 99-1476), and Merck Genome Research Institute (Grant 244).

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

References

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25:** 3389–3402.

Chothia, C. and Jones, E.Y. 1997. The molecular structure of cell adhesion molecules. *Annu. Rev. Biochem.* **66:** 823–862.

Chothia, C., Gelfand, I.M., and Kister, A.E 1998. Structural determinants in the sequences of immunoglobulin variable domain. *J. Mol. Biol.* **278:** 457–479.

Galitsky, B., Gelfand, I.M., and Kister, A.E. 1998. Predicting amino acids sequences of antibody human VH chains from its first several residues. *Proc. Natl. Acad. Sci.* **95:** 5193–5198.

———. 1999. Class-defining characteristics in the mouse heavy chains of variable domains. *Protein Eng.* **12:** 101–107.

Gallin, W.J. 1998. Evolution of the classical cadherin family of cell adhesion molecules in vertebrates. *Mol. Biol. Evol.* **15:** 1099–1107.

Gelfand, I.M. and Kister, A.E., 1995. Analysis of the relation between the sequence and secondary and three dimensional structures of immunoglobulin molecules. *Proc. Natl. Acad. Sci.* **92:** 10884–10888.

———. 1997. A very limited number of keywords main patterns) describes all sequences of the human variable heavy (V_H) and κ (V_κ) domains. *Proc. Natl. Acad. Sci.* **94:** 12562–12567.

Gumbliner, B.M. 1996. Cell adhesion: The molecular basis of tissue architecture and morphonegenesis. *Cell* **84:** 345–357.

Gusfield, D. 1997. *Algorithms on strings, trees and sequences: Computer science and computational biology.* Cambridge University Press, New York.

Eddy, S.R. 1996. Hidden Markov models. *Curr. Opin. Struct. Biol.* **6:** 361–365

Hill, E., Broadbent, I., Chothia, C., and Peltitt, J. 2001. Cadherin superfamily proteins in *Caenorhabditis elegans* and *Drosophila melanogaster*. *J. Mol. Biol.* **305:** 1011–1024.

Hofmann, K., Bucher, P., Falquet, L., and Bairoch, A. 1999. The PROSITE database, its status in 1999. *Nucleic Acids Res.* **27:** 215–219.

Jensen, P.H., Soroka, V., Thomsen, N.K., Ralets, I., Berezin, V., Bock, E., and Poulsen, F.M. 1999. Structure and interactions of Ncam modules 1 and 2, basic elements in neural cell adhesion. *Nat. Struct. Biol.* **6:** 486–493.

Koch, A.W., Bozic, D., Pertz, O., and Engel, J. 1999. Homophilic adhesion by cadherins. *Curr. Opin. Struct. Biol.* **9:** 275–281.

Lesk, A.M., Levitt, M., and Chothia, C. 1987. Alignment of the amino acid sequences of distantly related proteins using variable gap penalties. *Protein Eng.* **1:** 77–78.

Nagar, O., M., Ikura, M., and Rinl, J.M. 1996. Structural basis of calcium-induced E-cadherin rigidification and dimerization. *Nature* **380:** 360–364.

Overduin, M., Harvey, T.S., Bagby, S., Tong, K.L., Yau, P., Takeishi, M., and Ikura, M. 1995. Solution structure of the epithelial cadherin domain responsible for selective cell adhesion. *Science* **267:** 386–389.

Pearson, W.R. 1996. Effective protein sequence comparison. *Methods Enzymol.* **266:** 227–258.

Pertz, O., Bozic, D., Koch, A.W., Fauser, C., Brancaccio, A., and Engel, J. 1999. A new crystal structure, Ca^{2+} dependence and mutational analysis reveal molecular details of E-cadherin homoassociation. *EMBO J.* **18:** 1738–1747.

Takeichi, M. 1991. Cadherin cell adhesion receptors as a morphogenetic regulator. *Science* **251:** 1451–1455.

Takeichi, M. 1995. Morphogenetic roles of classic cadherins. *Current Opin. Cell Biol.* **7:** 619–627.

Shapiro, L. and Colman, D.R. 1999. The diversity of cadherins and implications for a synaptic adhesive code in the CNS. *Neuron* **23:** 427–430.

Shapiro, L., Fannon, A.M., Kwong, P.D., Thompson, A., Lehmann, M.S., Grubel, G., Legrand, J-F., Als-Nielsen, J., Colman, D.R., and Hendrickson, W.A. 1995. Structural basis of cell-cell adhesion by cadherins. *Nature* **374:** 327–336.

Smith T.F. and Waterman, M.S. 1981. Identification of common molecular subsequences. *J. Mol. Biol.* **147:** 195–197.

Suzuki, S.T. 1996. Structural and functional diversity of cadherin superfamily: Are new members of cadherin superfamily involved in signal transduction pathway? *J. Cell. Biochem..* **61:** 531–542.