doi:10.1006/jmbi.2000.4361 available online at http://www.idealibrary.com on IDEAL® *J. Mol. Biol.* (2001) **305**, 1011–1024

JMB

Cadherin Superfamily Proteins in *Caenorhabditis elegans* and *Drosophila melanogaster*

Emma Hill[1]*, Ian D. Broadbent[2], Cyrus Chothia[1] and Jonathan Pettitt[2]

[1]*MRC, Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK*

[2]*Department of Molecular and Cell Biology, University of Aberdeen, Institute of Medical Science, Foresterhill, Aberdeen AB25 2ZD, UK*

The ability to form selective cell-cell adhesions is an essential property of metazoan cells. Members of the cadherin superfamily are important regulators of this process in both vertebrates and invertebrates. With the advent of genome sequencing projects, determination of the full repertoire of cadherins available to an organism is possible and here we present the identification and analysis of the cadherin repertoires in the genomes of *Caenorhabditis elegans* and *Drosophila melanogaster*. Hidden Markov models of cadherin domains were matched to the protein sequences obtained from the translation of the predicted gene sequences. Matches were made to 21 *C. elegans* and 18 *D. melanogaster* sequences. Experimental and theoretical work on *C. elegans* sequences, and data from ESTs, show that three pairs of genes, and two triplets, should be merged to form five single genes. It also produced sequence changes at one or both of the 5′ and 3′ termini of half the sequences. In *D. melanogaster* it is probable that two of the cadherin genes should also be merged together and that three cadherin genes should be merged with other neighbouring genes.

Of the 15 cadherin proteins found in *C. elegans*, 13 have the features of cell surface proteins, signal sequences and transmembrane helices; the other two have only signal sequences. Of the 17 in *D. melanogaster*, 11 at present have both features and another five have transmembrane helices. The evidence currently available suggests about one-third of the cadherins in the two organisms can be grouped into subfamilies in which all, or parts of, the molecules are conserved. Each organism also has a ~980 residue protein (CDH-11 and CG11059) with two cadherin domains and whose sequences match well over their entire length two proteins from human brain. Two proteins in *C. elegans*, HMR-1A and HMR-1B, and three in *D. melanogaster*, CadN, Shg and CG7527, have cytoplasmic domains homologous to those of the classical cadherin genes of chordates but their extracellular regions have different domain structures. Other common subclasses include the seven-helix membrane cadherins, Fat-like protocadherins and the Ret-like cadherins. At present, the remaining cadherins have no obvious similarities in their extracellular domain architecture or homologies to their cytoplasmic domains and may, therefore, represent species-specific or phylum-specific molecules.

**Corresponding author*

Keywords: cell adhesion; hidden Markov models; evolution; genomics

Introduction

The cadherin superfamily of cell adhesion molecules is involved in multiple morphogenetic events in animal development, such as the patterning of the central nervous system, and stable tissue formation (Takeichi, 1995; Gumbiner, 1996). Cadherin superfamily genes encode variable numbers of a unique, approximately 110 residue, extracellular domain termed the cadherin domain. These domains mediate intermolecular interactions and are dependent on calcium ions, which bind at sites between adjacent cadherin domains to produce a rigid structure (Figure 1(a)-(c)). The extracellular

Abbreviations used: HMM, hidden Markov model; NC, non-chordate; PCCD, primitive classic cadherin domain.

E-mail address of the corresponding author: eeh@mrc-lmb.cam.ac.uk

0022-2836/01/051011–14 $35.00/0

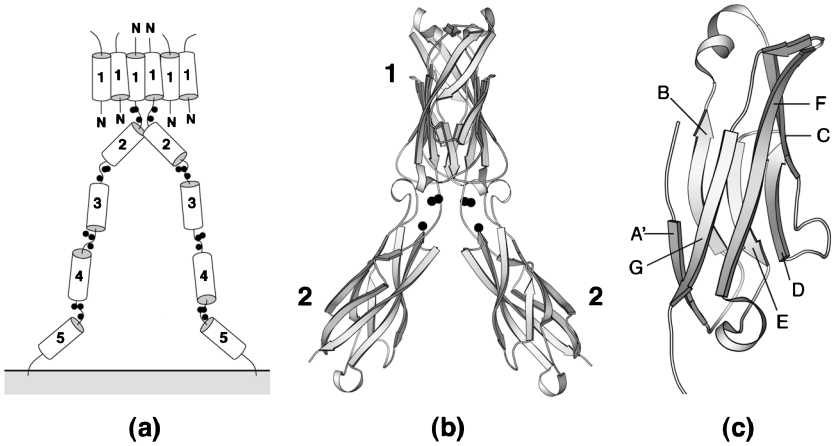

Figure 1. The structure and interactions of classic cadherins (Overduin *et al*., 1995; Shapiro *et al*., 1995; Nagar *et al*., 1996). (a) A view of the structure of the dimer formed by domains 1 and 2 of E-cadherin. Three calcium ions bound at the interface between domains 1 and 2 are shown as filled circles. (b) A model of the association of two classic cadherin molecules on the surface of one cell and the contacts that these two make to other cadherin dimers on other cells. The cadherin domains are represented by small cylinders and calcium ions by small filled circles. (c) The β-sheet structure of a cadherin domain. Strands are shown as ribbons and are labelled A′, B, C, D, E, F and G. Domain 1 of N-cadherin and domain 2 of E-cadherin also have a small A strand.

domains are linked *via* a transmembrane helix to a cytoplasmic domain that is known in some cases to interact with certain classes of intracellular proteins.

The availability of the genome sequences of the nematode *Caenorhabditis elegans* (The *C. elegans* Sequencing Consortium, 1998) and the fruitfly *Drosophila melanogaster* (Celera Genomics and The Berkeley *Drosophila* Genome Project 2000) means we can now begin to define, by a combination of sequence analysis and experiment, the cadherin superfamily proteins in these two organisms. The definition of the cadherin repertoire in these organisms will provide a basis for the experimental determination of their function. It also helps us to identify conserved members of the superfamily, as well as organism or phylum-specific cadherins, and thus contributes to our understanding of the role of this family in the evolution of development.

The Cadherin superfamily

The first cadherins to be identified form a subfamily subsequently termed the ''classic'' cadherins (Takeichi, 1995). In chordates, these cadherins share the same basic structure consisting of an extracellular region composed of five tandem cadherin repeats (Figure 1(a)). Pairs of the rigid form of classic cadherins that are situated on the surface of the same cell form dimers through homophilic interactions between the N-terminal region of each molecule. These dimers can then adhere to dimers on the surface of other cells, thus producing cell adhesion (Figure 1(b); Nose *et al*., 1990; Shapiro *et al*., 1995; Nagar *et al*., 1996). The extracellular domain is linked *via* a transmembrane helix to a highly conserved classic cytoplasmic domain that contains binding sites for a set of cytoplasmic proteins, the catenins (Ranscht, 1994). Catenins regulate the biological function of the classic cadherins through their association with the actin cytoskeleton and other molecules (Grunwald, 1993).

Related groups of cadherins, the desmogleins and desmocollins, display the same extracellular architecture as the classic cadherins, but have a different cytoplasmic domain that interacts with a different set of cytoskeletal components (Koch & Franke, 1994).

C. elegans, *D. melanogaster* and the sea urchin *Lytechinus variegatus* contain cadherins with classic cytoplasmic domains which have been shown experimentally to associate with catenins. Unlike the chordate classic cadherins, the structures of their extracellular domains are more heterogeneous in terms of size and domain composition (reviewed by Tepass, 1999).

There are other members of the cadherin superfamily that do not contain obvious catenin-binding sites, but, at least in some cases, are also able to mediate cell adhesion. These molecules can also be grouped into subfamilies, such as the protocadherins, the Fat-like cadherins and the seven-pass transmembrane cadherins (Sano *et al*., 1993; Suzuki, 1996; Yagi & Takeichi, 2000). At present the molecular functions of these cadherins are poorly understood, particularly in terms of how, or indeed whether, they interact with components of the cytoskeleton.

Methods for the Identification of Cadherin Domains in the Predicted Protein Sequences of *C. elegans* and *D. melanogaster*

The complete set of predicted protein sequences of *C. elegans* and *D. melanogaster* were obtained by ftp from:

ftp://ftp.sanger.ac.uk/pub/databases/C. elegans_sequences
and
ftp://ncbi.nlm.nih.gov/genbank/genomes/D_melanogaster/
respectively.

These predicted protein databases were searched for sequences that contain cadherin domains using two methods: hidden Markov models (HMMs) (Krogh *et al*., 1994; Eddy, 1996). Currently, HMMs are probably the most sensitive automatic sequence comparison method available (Park *et al*., 1998). The HMM package used here was the iterative procedure SAM-T98 (Karplus *et al*., 1998). The set of models that were most effective for detecting cadherin domains were the three created using the sequences of the cadherin domains of known structure, domains 1 and 2 of murine epithelial (E) cadherin and domain 1 of murine neural (N) cadherin (Overduin *et al*., 1995; Shapiro *et al*., 1995; Nagar *et al*., 1996). These models were run against the *C. elegans* and *D. melanogaster* predicted protein sets. A cut-off score of −15 was taken to indicate significant matches based on the assessment of SAM-T98 (Park *et al*., 1998). The cadherin HMMs made significant matches to (i) 141 regions in 21 *C. elegans* sequences and (ii) 178 regions in 18 *D. melanogaster* sequences. Key residue inspection (Chothia *et al*., 1988). The unmatched regions of the 21 *C. elegans* and 18 *D. melanogaster* sequences found by the HMMs were examined by eye for the pattern of key residues that is characteristic of cadherin structures. An alignment with these key residues highlighted is available on the accompanying website. A total of 25 complete domains and four partial domains were found using this procedure within the unmatched regions of the *C. elegans* sequences; 16 complete and three partial domains were identified within the unmatched regions of the *D. melanogaster* sequences. The extent of divergence in these sequences prevented their detection by the HMMs.

Cadherin domains found by the HMMs and the key residue analysis

From the sequence analyses, we found in 21 *C. elegans* sequences a total of 166 complete and four partial cadherin domains (Table 1A). In 18 *D. melanogaster* sequences we found 194 complete and three partial cadherin domains (Table 1B). The sequence for the *D. melanogaster* gene *ds* given by the *Drosophila* Genome Project (2000) is much shorter than sequences found for the cloned gene (Clark *et al*., 1995). The latter has 22 additional cadherin domains and we use this description of *ds* in this work. These numbers for cadherin domains in *C. elegans* were modified further by the subsequent experimental work described below.

Methods Used to Identify Non-Cadherin Regions in the Cadherin Sequences

The following resources were used to further define the predictions obtained from the HMM analysis.

SignalP server (Nielsen *et al*., 1997). For the detection of signal peptide sequences using the program options 25 3.0 and 3,3 & 45 3.4 and 3.75.

SMART HMM server (Schultz *et al*., 1998). To detect matches to HMMs for extracellular domains of cell surface and matrix proteins.

TMHMM server (Sonnhammer *et al*., 1998). For the detection of transmembrane, intracellular and extracellular regions.

FASTA (Pearson, 1998). Regions that were not assigned a domain using the three servers described above were searched against the NRDB90 (Holm & Sander, 1998) and Swissprot (Bairoch & Apweiler, 1999) databases using FASTA with an expectation value threshold of 0.001. The complete sequences of all 21 *C. elegans* and 18 *D. melanogaster* sequences were also searched against NRDB90 to look for whole-protein homologues.

Refining Cadherin Gene Predictions

In *C. elegans* five sets of predicted cadherin proteins were identified whose gene sequences are adjacent on their respective chromosomes (Table 1A). In these cases, the translations of the individual Genefinder predicted genes were missing either their signal peptides or their transmembrane helices, or both of these. Merging the adjacent predicted genes resulted in a gene product that possesses one or both motifs. Three mergers, one comprising C45G7.6 and C45G7.5, one comprising F18F11.3 and Y66H1B.1 and the other comprising R10F2.2 and R10F2.1 bring together both of the motifs. Two of these mergers are also supported by experimental work described below.

There are also two sets of three genes that appeared as though they should be merged. The first of these, comprising Y71D11A.1, Y92C3A and Y119D3B.L, was, at the time of this work, three regions of unfinished sequence. The gene resulting from this merger comprises a signal sequence, 21 cadherin domains, a transmembrane helix and a cytoplasmic domain. The predicted merger of the second set, comprising ZK39.1, Y52B11B.2 and W02B9.1, has been investigated experimentally and is described below and will be elsewhere (I.D.B. & J.P., unpublished results).

The genome sequence in the regions around suspected split genes for *C. elegans* was also examined by Daniel Lawson (The Sanger Centre, UK), the current curator of the *C. elegans* sequence database. In each case he showed that there are reasonable alternative gene predictions consistent with the merging of the suspected genes. These mergers reduce the number of cadherin protiens within *C. elegans* from 21 to 15.

Examination of the positions and predicted sequences of the *D. melanogaster* genes (Table 1B) suggests that one adjacent pair should be merged. The first sequence (CG15511) codes for

Table 1. Cadherin superfamily genes in the *C. elegans* and *D. melanogaster* genomes and the number of cadherin domains they encode

a. *C. elegans*

Chrom-osome	Gene prediction	Gene name	Protein	Position of the gene on the chromosome			Number of cadherin domains identified		
				Start	Stop	Strand	SAM T-98	K.R (k) EXP (e)	Total
I	ZK39.1			10502017	10498071	-	1	2k	3
	Y52B11B.2			10493668	10482772	-	12	0.5k	12.5
	W02B9.1			10464478	10446729	-	3	0.5k	3.5
		hmr-1	HMR-1A	10464478	10446729	-	3	0	3
		hmr-1	HMR-1B	10504917	10446729	-	16	3k	19
II	B0034.3	*cdh-11*	CDH-11	5986437	5975648	-	2	0	2
	R05H10.6			14874013	14877475	+	4	2k, 2e	8
		cdh-7	CDH-7	14870514	14877863	+	4	2k, 2e	8
III	Y71D11A.1			1098613	1101261	§	1	1k, 2e	4
	Y92C3A			1102837	1123741	§	2	2k	4
	Y119D3B.L			1123542	1126797	§	10	3k	13
		cdh-12	CDH-12	1098613	1126797	+	13	8	21
	R10F2.1			2364410	2375476	+	5	0	5
	R10F2.2			2353264	2358952	+	16	4k	20
		cdh-1	CDH-1	2365910	2374216	+	21	4	25
	F25F2.2	*cdh-4*	CDH-4	3956081	3972283	+	32	0	32
	ZK112.7	*cdh-3*	CDH-3	7202427	7189377	-	15	4k	19
IV	F18F11.3			346134	348552	+	2	1k	3
	Y66H1B.1			351636	360087	+	9	1k	10
		cdh-8	CDH-8	346134	360087	+	11	2	13
	C45G7.6			2431421	2422367	-	4	3.5k	7.5
	C45G7.5			2412396	2404624	-	1	1.5k	2.5
		cdh-10	CDH-10	2431421	2404624	-	5	5	10
	F08B4.2	*cdh-5*	CDH-5	8309517	8318349	+	8	0	8
	Y37E11A.94.a	-	-	3403688	3471814	+	1	0	1
V	F15B9.7	*cdh-6*	CDH-6	13256576	13270637	+	8	0	8
	T01D3.1	-	-	13885906	13899175	+	1	0	1
X	F59C12.1	*cdh-9*	CDH-9	16285942	16291128	+	4	1k	5

b. *D. melanogaster*

Chrom-osome	Gene or gene prediction	Cytological Position	Position of the gene on the chromosome			Number of cadherin domains identified		
			Start	Stop	Strand	SAM T-98	K.R	Total
II	*Shg*	57B19-57B20	15993073	15986515	-	7	1	8
	Ft	24D7-24E1	4135697	4116991	-	31	3	34
	*Ds**	21C7-21D1	691165	632019	-	5	22*	27
	Stan	47B4-47B7	5722663	5736204	+	9	3	12
	CadN	36C8-36D1	17572851	17483186	-	16	3	19
	CG7527	36D1-36D2	17649054	17621637	-	7	0.5	7.5
	[CG14396	39B4	20996609	20995850	-	[1	0	1
	CG1061	39B4	20993719	20990902	-	0	0	0
	merged gene RET	39B4	20996609	20990902	-	1	0	1]
III	CG7749	76E2-76E4	19862473	19878242	+	31	3	34
	CG6445	74B2	17277942	17266885	-	14	0	14
	CG6977	87A4	7694742	7686820	-	12	0.5	12.5
	CG3389	88C10	10390932	10397792	+	14	1	15
	CG14900	89C6-89C7	12251595	12243030	-	12	0	12
	CG10421	96C3	20967534	20965006	-	3	1.5	4.5
	[CG4655	86C7	6630219	6634815	+	0	2	2
	CG4509	86C7	6635616	6641306	+	[3	0	3
	merged gene]	86C7	6630219	6641306	+	3	2	5]
	[CG10244	96C2	20948892	20940196	-	[1	0	1
	HD-14	96C2	20939599	20938602	-	0	0	0
	merged gene]	96C2	20948892	20938602	-	1	0	1]
	[CG15511	99C6	25566718	25568737	+	[1	0	1
	CG7805	99C6	25570161	25575811	+	9	1	10
	merged gene]	99C6	25566718	25575811	+	10	1	11]
IV	CG11059	102F1	829155	841178	+	2	0	2

a signal peptide and a cadherin domain, and the second (CG7805) codes for ten cadherin domains followed by a transmembrane helix and a cytoplasmic domain. The effect of this merger is to reduce the number of *D. melanogaster* sequences from 18 to 17. Analysis of the predicted proteins adjacent to these cadherin proteins in Flybase identified three other genes which should probably be merged with two of those identified previously. The first of these is CG4655, which contains two cadherin domains and lies upstream of CG4509. The second of these was that of CG14396 and CG1061, which together form the Ret protein. The third is that of HD-14, which lies directly downstream of CG10244 and merging the two produces a complete cytoplasmic tyrosine kinase domain. This domain architecture of the merged protein is similar to that of the Ret proto-oncogenes which strengthens the evidence for the merger.

Examination of the Predicted Cadherin Sequences in *C. elegans*

Previous experimental work has defined the structure and function of the *C. elegans* cadherin genes *cdh-3* (Pettitt *et al*., 1996) and *hmr-1* (Costa *et al*., 1998). We have proposed gene names for the other members of the cadherin family in *C. elegans*, and the relation between these gene names and the Genefinder identifiers given to the genome sequences is described in Table 1A. Results from the HMMs and key residue inspection indicated that some of the cadherin sequences produced by the Genefinder predictions were incomplete, therefore, RT-PCR experiments were carried out to check various aspects of the predicted cadherin sequences (see Materials and Methods, and website).

In the case of *cdh-1*, the merger of R10F2.2 and R10F2.1 was supported and the 3′ cytoplasmic domain was completely redefined. We determined the full cDNA sequence of *cdh-1* by RT-PCR. The Genefinder prediction of the 5′ end of *cdh-1* did not encode a putative signal peptide, so the upstream sequence was scanned to identify exons predicted to encode signal peptides by the SignalP server. An exon was found that contained a predicted signal peptide and part of a cadherin repeat and RT-PCR was used to confirm that this first exon can be spliced to the rest of *cdh-1*. A similar RT-PCR approach confirmed the merger between F18F11.3 and Y66H1B.1 (*cdh-8*).

Scanning upstream of sequences for extra exons altered the 5′ ends of three other *C. elegans* genes (*cdh-5, cdh-7* and *cdh-12*), providing in each case sequence found to encode a signal peptide.

RT-PCR analysis of the proposed merger between W02B9.1, Y52B11B.2 and ZK39.1 extends the size of the previously identified *hmr-1* gene (Costa *et al*., 1998), and demonstrates that this gene is capable of producing two transcripts using alternative promoters and alternative splicing (I.D.B. & J.P., unpublished results). We therefore designate the original *hmr-1* gene product HMR-1A and the longer alternative isoform HMR-1B. The full cDNA sequence of the HMR-1B transcript was determined by RT-PCR, and this confirmed the predicted merger between W02B9.1, Y52B11B.2 and ZK39.1 and modified it both 5′ and 3′. The SignalP server was used to verify that the 5′ ends of the HMR-1B and HMR-1A transcripts encoded putative signal peptides.

We have also used information from the Kohara laboratory EST database (http://www.ddbj.nig.ac.jp/htmls/c-elegans/html/CE_INDEX.html) to confirm the structures of regions of the cadherin sequences. Partial cDNA sequences are available that match regions of *cdh-4, cdh-5, cdh-6, cdh-7, cdh-11,* and *cdh-12* (see website for details of these ESTs: http://www.mrc-lmb.cam.ac.uk/genomes/Cadherins/cad_web_pages.html). The information from the ESTs altered the 3′ ends of four of the Genefinder sequences (*cdh-5, cdh-7, cdh-11* and *cdh-12*), and therefore modified their cytoplasmic domains.

Finally, alterations to the Genefinder prediction for both the 5′ and 3′ ends of *cdh-9* are supported by the comparison with the *C. briggsae cdh-9* orthologue, derived from the sequence of fosmid G45J16 (R. Babbar & J.P., unpublished; see website).

Overall these investigations produced changes to the 5′ ends of six *C. elegans* genes and to the 3′ ends of six, and confirmed the mergers of three sets of genes experimentally.

Table 1. (*footnote*).

Sequences, genes and positions: for *C. elegans*, in cases where we believe the Genefinder prediction to be correct, a *cdh* gene and protein designation is given on the same line e.g. B0034.3, *cdh-11* and CDH-11. In those cases where the definition of the coding region has been modified by work reported here the *cdh* designation is placed on a subsequent line, and the revised positions for the gene are given in the appropriate columns. For example the predicted genes R10F2.1 and R10F2.1 have been merged and extended at the 5′ and 3′ ends. The revised gene prediction is given on the next line along with its corresponding gene and protein designation, *cdh-1*.

Note that the gene *hmr-1* has two alternative protein products (HMR-1A and HMR-1B).

For *D. melanogaster*, the entry under Gene is the gene symbol. Those in the form CGXXXX come from the genome projects whilst the other refer to previously identified genes.

The number of cadherin domains: here we list the number of cadherin domains found by (i) hidden Markov models (under SAM T-98); (ii) key residue analysis under (under K.R. and k); and (iii) the RT-PCR experiments (under Exp. and e).

[a] Ds is at present truncated in the gene prediction. Its cloned sequence has been determined and the 22 domains listed here are from Clark *et al*. (1995).

Domains in the Cadherin Superfamily Proteins of *C. elegans* and *D. melanogaster*

The computational and experimental work identified a total of 175 cadherin domains in 15 *C. elegans* proteins The computational work identified 217 complete cadherin domains and three partial domains in 17 *D. melanogaster* proteins. Figure 2 shows the domain architectures of these proteins. Full details of the matches made within the sequences are available on the website:

http://www.mrc-lmb.cam.ac.uk/genomes/Cadherins/cad_web_pages.html

Hutter *et al*. (2000) have also described the domain structure of 12 *C. elegans* cadherin proteins. Their analysis, for the most part, concurs with ours. The differences arise from our refinement of the gene predictions based upon experimental evidence, coupled with the additional cadherin domains identified by key residue analysis.

The number of cadherin domains in the different proteins varies greatly: from one to 32 in *C. elegans* and one to 34 in *D. melanogaster* (Figure 2). This is quite different from the situation for cadherin proteins of higher metazoans where the large majority of proteins that are currently known have five or six domains and only a small proportion have a large number. Note that although the number of cadherin domains in the proteins varies in the two organisms, the extent of the variation is very similar (Table 2 and Figure 2).

Cytoplasmic domains

In *C. elegans* the cytoplasmic domains range in length from 56 to 233 residues. In *D. melanogaster* they tend to be larger, ranging in length from 43 to 968 residues with only five less than 227 residues. This means that, in most cases, the cytoplasmic regions of the *C. elegans* cadherin proteins are much shorter than those in the *D. melanogaster* cadherin proteins.

Proteins of uncertain status

There are two *C. elegans* sequences of uncertain status, T01D3.1 and Y37E11A.94a. They each contain one cadherin domain and both have a predicted signal peptide and six and one EGF domains, respectively. Other sequences made matches on the margin of significance, these may turn out to be very divergent cadherin proteins.

For *D. melanogaster* there are at least three apparently incomplete genes (CG7527, CG6977 and CG10421) which contain partial cadherin domains. Further sequence information and experimental work is needed to clarify the structures of these genes.

Undefined regions in the cadherin proteins

There are 27 unmatched regions of over 100 residues in 11 of the *C. elegans* cadherin proteins. Within the *D. melanogaster* superfamily of identified cadherins there are 19 such regions in 14 predicted proteins. The positions of the regions in the different sequences are given on the website that accompanies this work.

Comparisons with the Cadherin Proteins Identified by Other Groups

Two *C. elegans* and five *D. melanogaster* cadherin proteins have been characterized experimentally prior to the current study; these are discussed below. Whilst the work described here was in progress, or subsequent to it, three groups made available on the internet assignments for cadherin proteins in *C. elegans* and/or *D. melanogaster*.

Hutter *et al*. (2000) made domain assignments to the protein products of putative cell adhesion and extracellular genes. They identified 19 sequences in *C. elegans* which are the same as those described here except for the two we described as being of uncertain status. In these sequences Pfam HMMs (Bateman *et al*., 2000) detect 137 cadherin domains. Comparison with our results shows that these 137 domains are very largely the same as the 139 we found for these sequences using the SAM HMM procedure but include only a few of the 31 domains found by key residue analysis. The signal sequences, non-cadherin domains, transmembrane helices and cytoplasmic domains found by Hutter *et al*. (2000) are similar to those found by our procedures.

Schultz *et al*. (2000) recently extended the SMART database and made available domain assignments to putative cell surface and matrix proteins from genome sequences. The assignments SMART makes for *D. melanogaster* genome sequences are close to those given by the HMM calculations described here: it assigns 169 cadherin domains to 17 sequences; our HMMs assign 178 cadherin domains to the same 17 sequences plus one additional sequence. The results for *C. elegans* are less close: SMART assigns 82 cadherin domains to 13 of the 21 genome sequences described here as opposed to our HMM assignment of 141 cadherin domains. Given the closeness of the *D. melanogaster* results, the discrepancy of the *C. elegans* results are likely to arise from a programming or data error in SMART, rather than an error in its HMMs.

Hynes & Zhao (2000) found 17 cadherin proteins in *D. melanogaster* which are the same as those described here except one. They also give numbers for the different types of domains that they found in each sequence and these are close to those detected using the Pfam and SAM HMMs. The matching procedures that they used and the arrangement of domains in the different sequences are not described.

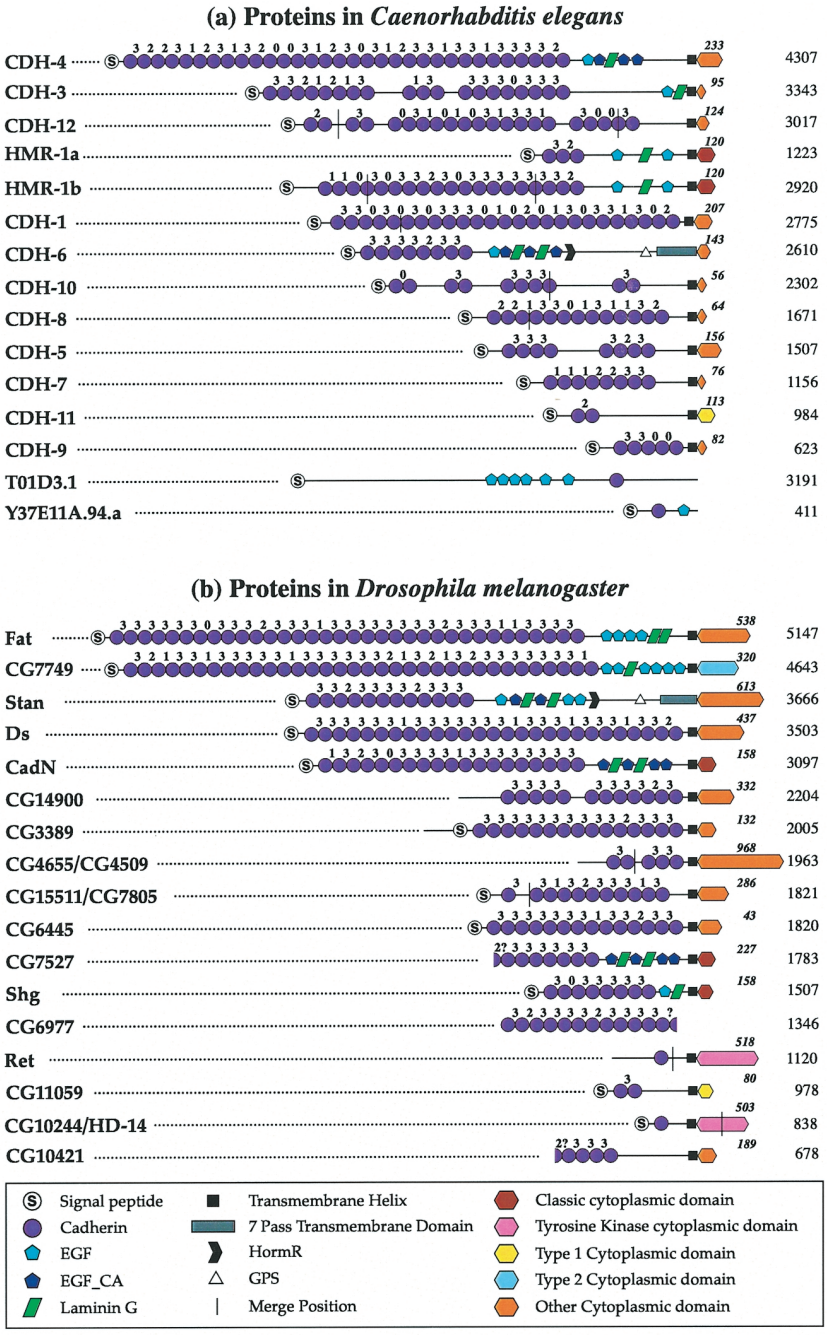

Figure 2. The repertoires of cadherin superfamily proteins within the genomes of (a) *Caenorhabditis elegans* and (b) *Drosophila melanogaster*. EGF domains are epidermal growth factor-like. EGF_CA are calcium-binding EGF domains. HormR is a protein domain found in hormone receptors. GPS is a G-protein coupled receptor proteolytic site domain found in latrophilin/CL-1, sea urchin REJ and polycystin. Type 1 Cytoplasmic domain refers to the homo-logous cytoplasmic domain of *C. elegans* protein CDH-11 and *D. melanogaster* protein CG11059. Type 2 Cytoplasmic domain refers to the homologous cytoplasmic domain of the *D. melanogaster* protein CG7749 with those of human Fat protein and the rat protocadherin Fat (Trembl IDs: Q14517 and Q9WU10, respectively). The numbers shown above tandem cadherin domains represent the number of calcium ions likely to be bound between them.

Therefore the results we obtained using HMMs and other automatic procedures are in good agreement with the work of others who used similar automatic procedures to assign domains to cadherin sequences. However, the work described here extends these types of analyses considerably, firstly by using key residue inspection to extend the detection of cadherin domains, and secondly by using RT-PCR to substantially improve 11 *C. elegans* gene predictions. Indeed, this work implies that, at present, computational analyses of complex predicted protein sequences can only produce

Table 2. Distribution of the number of cadherin domains found in *C. elegans* and *D. melanogaster* proteins

Ranges for no. cadherin domains (*n-m*) within the sequences	Number of sequences with *n* to *m* domains	
	C. elegans	*D. melanogaster*
1-9	10	6
10-19	4	7
20-29	2	1
30+	1	2

initial approximate results (see also Teichmann & Chothia, 1999). We would expect, for example, that RT-PCR experiments on the *D. melanogaster* cadherin sequences, and on the *C. elegans* sequences not included in the current experiments, would extend further the results described here.

Calcium Binding Sites Between Cadherin Domains

To form effective adhesion complexes classical cadherins bind three calcium ions at interfaces between domains (Ringwald *et al.*, 1987) (Figure 1(b)). In the two domain fragments of E-cadherin, whose structure is known, the three ions are bound by the side-chains of two Glu and one Asp from the N-terminal domain, two Asp and an Asn from the linker region, and three Asp from the C-terminal domain. Conservation of these residues in the other domains of classic cadherins implies that they bind Ca^{2+} in the same manner (Nagar *et al.*, 1996).

To determine the Ca^{2+} binding capabilities of *C. elegans* and *D. melanogaster* cadherin domains, they were examined for the presence of the same set of side-chains as seen in classical cadherins. We assume that a Ca^{2+}-can be held in some cases when one, or in the case of the third Ca^{2+}, two of the residues involved are absent. The number of ions likely to be fixed between domains is shown in Figure 2.

We predict that between one and three calcium ions can be bound in 84 % and 98 % of tandem cadherin domains in *C. elegans* and *D. melanogaster*, respectively (Table 3). From these numbers it is apparent that the conservation of the calcium binding sites is greater in the *D. melanogaster* proteins than in those of *C. elegans*. It is unclear whether the links that lack calcium-binding residues are flexible joints in their active state, have evolved a rigid structure, or have evolved a different ion-binding site.

An Ancient Association Between Cadherin, EGF and Laminin G Domains

Many of the cadherins that we have identified possess EGF and laminin G domains in the membrane-proximal region of their extracellular domains. Seven *C. elegans* proteins and six *D. melanogaster* proteins encode between one and six EGF domains, and all but two of these EGF domains are coupled with one or two laminin G domains (as is commonly observed). This observation suggests that the association of these three domains is evolutionarily ancient. Their function in the different cadherins is unclear but recent evidence suggests that, at least in the case of the non-chordate classic cadherins, they may form a site at which the cadherins are post-translationally processed (Oda & Tsukita, 1999). This site, found in all non-chordate classic cadherins, has been termed the primitive classic cadherin domain (PCCD), and consists of a non-chordate cadherin domain (NC), an EGF domain and a laminin G domain. We compiled hidden Markov models for the NC sequence from CadN, Shg and HMR-1 which were searched against the *C. elegans, D. melanogaster* and NRDB90 (Holm & Sander, 1998) databases. In addition to the classic cadherins, Shg, HMR-1 (both isoforms), CadN and CG7527, these models matched to the final cadherin domains of CDH-4, CG7749 and Stan. Both the CadN and Shg HMMs also made significant matches to Fat-like cadherins from vertebrates. In each case the match corresponded to the final cadherin domain preceding the EGF and laminin G domains. This cadherin domain is more divergent than most, in that it often lacks many of the residues involved in calcium binding.

Thus PCCD domains are found in non-classic cadherins in both chordates and non-chordates. This raises the possibility that these proteins may also be processed in the same manner as the non-chordate classic cadherins (Oda & Tsukita, 1999).

This simple picture does have a few exceptions; a match is also made to the third cadherin repeat within HMR-1B. This is not the final cadherin domain, but does lack the calcium-binding residues. Also, EGF and laminin G domains are

Table 3. Number of tandem cadherin domains predicted to bind 3, 2, 1 or 0 Ca^{2+} in *C. elegans* and *D. melanogaster* proteins

Organism	Number (%) of tandem cadherin domains that are predicted to bind *m* calcium ions			
	m			
	3	2	1	0
C. elegans	79 (52 %)	23 (15 %)	25 (17 %)	24 (16 %)
D. melanogaster	161 (80 %)	18 (9 %)	18 (9 %)	4 (2 %)

The procedure used to predict the number of bound Ca^{2+} is described in the text.

observed in some cases without an NC domain (CDH-3, CDH-6 and Fat).

Conserved and Unique Cadherins in *C. elegans*, *D. melanogaster* and Vertebrates

Two *C. elegans* and five *D. melanogaster* cadherin genes had been defined experimentally prior to the work presented here. These are *hmr-1* and *cdh-3* in the former (Pettitt *et al*., 1996; Costa *et al*., 1998) and *stan* (also known as flamingo or cad47B), *ds*, *ft*, *shg*, and *CadN* in the latter (Mahoney *et al*., 1991; Oda *et al*., 1994; Iwai *et al*., 1997; Adler *et al*., 1998; Chae *et al*., 1999; Usui *et al*., 1999). The availability of the complete cadherin repertoires in *C. elegans* and *D. melanogaster* allows us to look for cadherin subfamilies conserved in the worm, fly and other organisms.

Amongst these other organisms, the most extensively characterized set of cadherin proteins are those in humans. The currently known members of the human repertoire have been reviewed by Nollet *et al*. (2000). Some 85 proteins are known and most of these belong to one of a small number of subfamilies. Nineteen are in the classic cadherin subfamily, whose members have five or, in two cases, seven cadherin domains, and 52 are in the protocadherin α, β or γ subfamilies and have six cadherin domains. Larger cadherins are human Fat (34 domains) and the human Flamingos 1 and 2 (nine domains). The classic five-domain cadherin is also found in a primitive chordate: the golden star tunicate (Levi *et al*., 1997).

Within human subfamilies sequence identities tend to be high, 40-70 % and, in some cases, this is also true between subfamilies. In addition, intron/exon patterns are conserved in different subsets of human cadherins. This means that sequence identities and intron/exon patterns can be used to classify the human repertoire (Nollet *et al*., 2000). Examination of the intron positions in the *C. elegans* and *D. melanogaster* cadherins shows that there is no, or very little conservation in intron positions (our unpublished data). For example, the sequences CDH-11 and CG11059 match over their entire length but none of their introns occurs at equivalent positions (Figure 3). Similarly the comparison of cadherin domains within and between *C. elegans* and *D. melanogaster* proteins shows that they are much more divergent than in humans. In most cases, sequence identities are less 28 % and in the other few cases it only goes up to 29-33 %.

This means that to identify proteins in *C. elegans*, *D. melanogaster* and humans that might be functionally equivalent, we need to look for similarities beyond those given by simple sequence matches. Two features that are clearly related to function are (i) the nature of their cytoplasmic domains, which determines which intracellular pathways are activated by their interactions, and (ii) the lengths of the proteins, which are a major determinant of the geometry of their interactions. Here, in conjunction with what is known about their function, we discuss these features of the two sets of cadherins.

CDH-11 and CG11059

On the basis of the high level of overall primary sequence similarity, CDH-11 and CG11059 are clear orthologues. CDH-11 is homologous to CG11059 along its entire length, making a FASTA match with an *e*-value of zero and 28 % sequence identity. Both encode single-pass transmembrane proteins of similar sizes that contain two tandem cadherin domains in their extracellular portions. The cytoplasmic domains contain a number of conserved motifs, including a run of acidic residues. We have designated this conserved cytoplasmic domain as type 1 in Figure 2.

FASTA searches with *C. elegans* protein CDH-11 and *D. melanogaster* protein CG11059 show that they both match (with very significant *e*-values) two proteins of unknown function isolated from human brain (Trembl identifiers O94985 and O94831; Nagase *et al*., 1998). Comparison between the vertebrate and invertebrate homologues reveals that they share significant sequence similarity that extends along the entire length of the molecules: see the alignment of their sequences in Figure 3. This is the only known example of a cadherin found in both vertebrates and invertebrates where the homology is not just confined to particular domains. The existence of homologues with such high sequence similarity in vertebrates, *C. elegans* and *D. melanogaster* indicates that the function of this cadherin is likely to be conserved throughout evolution. None of the other cadherins shows the level of similarity shared by CDH-11 and CG11059 and assignments of orthology are less straightforward.

CDH-6 and Stan

Since both *C. elegans* and *D. melanogaster* each have only one seven-helix transmembrane cadherin, it might be thought that CDH-6 and Stan are orthologues. Both seven-helix membrane proteins are homologous to the members of secretin group in the G-protein coupled receptor family 2. Also, their extracellular regions are of similar lengths and possess similar numbers of cadherin, EGF, laminin G, GPS and HormR domains: they match one another with an *e*-value of zero and 31 % sequence identity. However, in their cytoplasmic domains the proteins are different: their sequences do not match and they are quite different in size.

CDH-6 and Stan make sequence matches with *e*-values of, or close to, zero to the seven-helix membrane proteins rat MEGF2, mouse CELSR1 and the human Flamingos 1 and 2. The extracellular regions of these proteins are all very similar with only small discrepancies of one cadherin domain and/or one or two EGF domains. Again,

Figure 3. An alignment of the sequences of CDH-11 from *C. elegans*, CG11059 from *D. melanogaster* and O94985 and O94832 from humans. These are the only cadherins whose sequences match over their entire length. We also indicate the positions of the 21 introns that occur in the *C. elegans* and *D. melanogaster* genes. A number, given the phase of the intron, is placed over the residue in whose codon it is found. A c precedes this number if the intron is in the *C. elegans* sequence and a d if it is in the *D. melanogaster* sequence. Of the 11 introns in the *C. elegans* sequence and ten in *D. melanogaster*, none occurs at equivalent positions.

neither of the invertebrate cytoplasmic domains matches those in the vertebrates.

Functional information is available for the *D. melanogaster* protein Stan. It is required for the regulation of planar polarity *via* a Frizzled-dependent pathway (Usui *et al*., 1999; Chae *et al*., 1999). No functional information exists for CDH-6 in *C. elegans* at present.

Beyond the two pairs of proteins discussed in the previous paragraphs, the evolutionary relationships between the cadherins in the two organisms are less obvious.

CDH-3, CDH-4, Fat and CG7749

D. melanogaster and *C. elegans* both have two large Fat-like cadherins, however the relationship between these four proteins is unclear. Their extracellular regions are similar in size and structure and have good sequence matches over long regions but their cytoplasmic domains do not have significant similarities.

Fat acts to regulate both the morphogenesis and proliferation of the larval imaginal discs (Mahoney *et al*., 1991), though the mechanism by which it coordinates these processes is at present unclear. Like Fat, CDH-3 is required for the morphogenesis of epithelia, though there is no evidence that it functions to regulate cell proliferation (Pettitt *et al*., 1996; L.A. Hodgson & J.P., unpublished results). The partial overlap in their functions, along with their similar domain architecture, suggest that they may function *via* a related mechanism.

CDH-4 is expressed almost exclusively in neurons, rather than epithelial cells (Birchall *et al*., 1995; I.D.B. & J.P., unpublished results), so it is unlikely to play a similar role to CDH-3 and Fat in epithelial morphogenesis. However, both human and rat Fat-like cadherins are also expressed in the developing nervous system (Dunne *et al*., 1995; Ponassi *et al*., 1999), suggesting that this subfamily includes regulators of both epithelial and neuronal morphogenesis, that may share a common mechanism.

Although there is no evidence for homology of the cytoplasmic domains of the Fat-like cadherins of *C. elegans* and *D. melanogaster*, the intracellular domain of the *D. melanogaster* protein CG7749 is homologous to those in both the human and rat Fat-like proteins. Its cytoplasmic domain of 320 residues makes a good FASTA match to their cytoplasmic domains with *e*-values of 0.00057 and 0.0017, respectively. We refer to this conserved cytoplasmic domain as the type 2 cytoplasmic domain (Figure 2(b)). All three encode either 34 or 35 cadherin repeats, varying numbers of

EGF domains and a laminin G domain. This homology suggests that these three proteins represent orthologues.

CDH-1 and Ds

These two have extracellular regions that are very similar; 25 and 27 cadherin domains, respectively and lack EGF and laminin G domains. However, their cytoplasmic domains show no similarity and their functions appear to be different. Ds functions in the regulation of imaginal disc morphogenesis, and may well interact with Fat in this process (Clark *et al*., 1995). However, it does not appear to play a role in regulation of cell proliferation. A *cdh-1* based GFP fusion construct is expressed largely in the developing nervous system (I.D.B. & J.P., unpublished results), suggesting that it is unlikely to function in the regulation of epithelia.

Ret-like

There are two Ret-like proteins in *D. melanogaster*, both are formed by gene mergers. The first one, CG10244/HD-14 encodes a signal peptide, one cadherin domain, a transmembrane helix and a cytoplasmic domain encoding a tyrosine kinase domain. The second, CG14396/CG1061 is annotated as Ret in Flybase, and encodes one cadherin domain, a transmembrane helix and a cytoplasmic domain encoding a tyrosine kinase domain. There is no *C. elegans* equivalent Ret-like protein. However good matches are made to similarly constructed Ret proteins in human, mouse, chicken, *Brachydanio rerio* and *Tetraodon fluvialitis.* Ret proteins are thought to be proto-oncogene receptors with a tyrosine-protein kinase activity important for development (Takahashi & Cooper, 1987).

HMR-1A, HMR-1B, CadN, SHG and CG7527

The relationship between the cadherins of *C. elegans* and *D. melanogaster* that have classic cytoplasmic domains is complicated by the fact that the two present in *C. elegans* are generated by a single gene (*hmr-1*), whereas in *D. melanogaster* they are encoded by separate genes (*CadN, Shg* and *CG7527*). Our analysis of the sequences upstream of the previously defined *hmr-1* gene has led to evidence for the production of two overlapping gene products from the *hmr-1* gene. The smaller product is that originally defined by Costa *et al*. (1998), and we propose to rename this gene product HMR-1A. For the larger product we propose the designation HMR-1B.

CadN and HMR-1B show significant sequence similarity and both encode the same number of cadherin repeats. Moreover, the similarity between corresponding cadherin repeats in HMR-1B and CadN is significantly higher than between non-corresponding repeats. Importantly both CadN and HMR-1B are expressed almost exclusively in neurons, where they appear to play similar roles in regulating neuronal morphogenesis (Iwai *et al*., 1997; I.D.B. & J.P., unpublished results).

HMR-1A does not closely resemble Shg in terms of structure, but the two proteins are of similar size (1223 and 1507 residues, respectively). Nevertheless both appear to represent the major epithelial classic cadherins in their respective organisms, and given that they both interact with a set of conserved cellular proteins (Oda *et al*., 1994; Tepass *et al*., 1996; Uemura *et al*., 1996), are highly likely to function *via* the same mechanism. Thus, they are clearly functionally equivalent, if not true orthologues.

Thus, both organisms appear to have pairs of classic cadherins that are functionally equivalent. HMR-1B and CadN appear to be orthologues but the phylogenetic relationship between the HMR-1A and Shg is unclear. A single classic cadherin gene could have given rise, *via* partial gene duplication, to the two-gene condition. Alternatively, the partial fusion of two classic cadherin genes could have produced the situation we observe in *C. elegans*. It will be interesting to determine the structure of classic cadherins from other protostomes to determine which arrangement is more likely to be the ancestral condition.

Our analyses show that *D. melanogaster* has the potential to encode a third classic cadherin by the CG7527 gene. The coding region appears to have been produced by a recent duplication of the second half of the CadN protein as their sequences are adjacent and 75 % identical in their overlapping region. No signal peptides can be identified for CG7527, and it is possible that CG7527 is a pseudogene. Further experimental evidence is required to determine the status of this protein.

Other cadherins in *C. elegans* and *D. melanogaster*

Beyond the fact that all have cadherin repeats the remaining cadherins in the two organisms do not share any obvious sequence similarities that would suggest they represent functional homologues; in particular, their cytoplasmic domains have no detectable sequence similarities. In some cases, this may be because they have evolved beyond the point at which homologous relationships can be detected by primary sequence alone. In most cases, however, the *D. melanogaster* domains are also much larger than those in *C. elegans*, which suggests that they have probably been selected for organism or phylum-specific processes.

Conclusions

We have described the domain architecture of 15 predicted cadherin proteins in *C. elegans* and of 17 in *D. melanogaster*. The initial assignment of domains to these sequences by HMMs and other

such procedures was supplemented by key residue analysis and, in the case of *C. elegans*, RT-PCR experiments. This supplementary work, particularly the experiments, improved substantially the results obtained from computational procedures. This work implies that, at present, computational analyses of complex genome sequences can only produce initial approximate results.

Though *C. elegans* and *D. melanogaster* differ greatly in size consisting of just under 10^3 and approximately 10^6 cells, respectively, comparison of their cadherin repertoires shows that they have two broad features that are very similar: the number of proteins in the two organisms and the distribution of the lengths of their extracellular regions. On a more detailed level, the similarities are fewer. There are three pairs that are probably orthologues: HMR-1B and CadN, CDH-11 and CG11059, and, CDH-6 and Stan. Another three pairs: CDH-3 and Fat, HMR-1A and Shg, and CDH-1 and Ds, have fairly similar structures in the extracellular regions, and, at least in the cases of CDH-3 and Fat, and HMR-1A and Shg, functions that partly overlap.

The common ancestor of human, *D. melanogaster* and *C. elegans* predates the Protostome-Deuterostome divide. Therefore, proteins common to humans and to one or both of the other two organisms were probably present in the earliest metazoan. The classic cytoplasmic domain is present in all three organisms but is associated with different extracellular domain arrangements. Thus the classic five-domain cadherin of chordates appears to represent a derived class of cadherins that may have evolved to fulfil a role unique to chordate development. The striking genomic arrangement of the vertebrate CNR genes (Wu & Maniatis, 1999; Sugino *et al*., 1999) is also absent from *C. elegans* and *D. melanogaster*, suggesting that this is also a recent invention.

Conversely, the presence of the Fat-like cadherins, seven-pass transmembrane cadherins and the CDH-11/CG11059 cadherins in all three organisms suggests that these represent ancient cadherin classes. In addition, there are three *D. melanogaster* cadherins (the Fat-like cadherin CG7749 and the two Ret-like cadherins) that have homologues in vertebrates, implying an ancient origin, which seem not to be present in the lineage leading to *C. elegans*. The determination of the full repertoire of human cadherins may lead to the identification of additional homologies in the three species.

The remaining cadherin proteins have neither similar sequence features, nor known functional roles that, with the presently available data, would indicate that they are orthologues. Although there may be some cases where the apparent lack of homology is the result of sequences having diverged beyond the point where their relationships can be detected, the current evidence suggests that a significant proportion of the cadherin repertoires in *C. elegans* and *D. melanogaster* have been selected for organism, or phylum-specific processes.

Materials and Methods

Standard molecular biology techniques were used throughout. RT-PCR was performed using Ready-to-Go™ RT-PCR beads (Amersham Life Science Ltd., UK). Total *C. elegans* RNA (1 μg) was added to the bead and first strand cDNA was synthesized using pd(N6) random hexamers at 42 °C for 30 minutes. After denaturation for five minutes at 95 °C, the outer set of nested primers were added to the reaction, and 35 PCR cycles (typically 95 °C 40 seconds, 54 °C 40 seconds, 72 °C 2 minutes 30 seconds) were performed. One microlitre of this reaction was then used as a template in a standard *Taq* PCR reaction for 35 cycles using the inner set of nested primers. The resulting RT-PCR product (typically approximately 1 kb) was cloned into the pGEM-T Easy™ vector (Promega UK Ltd.) for DNA sequencing using the M13 Forward and Reverse Universal primers. Automated DNA sequencing was carried out using the ABI Big Dye™ labeling kit (Perkin Elmer) on an ABI 377 DNA Sequencer by NCIMB Ltd. (Aberdeen, UK). MWG Biotech Ltd. (Germany) synthesized all oligonucleotide primers. The sequences of the primers used in this work are described in the data on the web site.

Acknowledgements

We thank Daniel Lawson (The Sanger Centre, UK) for comments on predicted gene mergers, and Aileen Flett for technical assistance. We thank one of our referees for making us aware of the merger that should take place between HD-14 and CG10244, and also of the presence of CG14396 in a subsequent release of the *D. melanogaster* genome. I.D.B. was supported by a Wellcome Trust Project Grant (no. 050720/Z/97/Z/PMG/LB).

References

Adler, P. N., Charlton, J. & Liu, J. C. (1998). Mutations in the cadherin superfamily member gene *dachous* cause a tissue polarity phenotype by altering frizzled signaling. *Development,* **125**, 959-968.

Bairoch, A. & Apweiler, R. (1999). The Swiss-prot protein sequence data bank and its supplement Trembl in 1999. *Nucl. Acids Res.* **27**, 49-54.

Batemen, A., Birney, E., Durbin, R., Eddy, S. R., Howe, K. L. & Sonnhammer, E. L. L. (2000). The Pfam protein families database. *Nucl. Acids Res.* **28**, 263-266.

Birchall, P. S., Fishpool, R. M. & Albertson, D. G. (1995). Expression patterns of predicted genes from the *C. elegans* genome sequence visualized by FISH in whole organisms. *Nature Genet.* **11**, 314-320.

C. elegans Sequencing Consortium (1998). Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science,* **282**, 2012-2018.

Celera Genomics and The Berkeley *Drosophila* Genome Project (2000). The genome sequence of *Drosophila melanogaster*. *Science,* **287**, 2185-2185.

Chae, J., Kim, M. J., Goo, J. H., Collier, S., Gubb, D., Charlton, J., Adler, P. N. & Park, W. J. (1999). The

Drosophila tissue polarity gene *starry night* encodes a member of the protocadherin family. *Development,* **126**, 5421-5429.

Chothia, C., Boswell, D. R. & Lesk, A. M. (1988). The outline structure of the T-cell αβ-receptor. *EMBO J.* **7**, 3745-3755.

Clark, H. F., Brentrup, D., Schneitz, K., Bieber, A., Goodman, C. & Noll, M. (1995). Dachsous encodes a member of the cadherin superfamily that controls imaginal disc morphogenesis in *Drosophila*. *Genes Dev.* **9**, 1530-1542.

Costa, M., Raich, W., Agbunag, C., Leung, B., Hardin, J. & Priess, J. R. (1998). A putative catenin-cadherin system mediates morphogenesis of the *Caenorhabditis elegans* embryo. *J. Cell Biol.* **141**, 297-308.

Dunne, J., Hanby, A. M., Poulsom, R., Jones, T. A., Sheer, D., Chin, W. G., Da, S. M., Zhao, Q., Beverley, P. C. L. & Owen, M. J. (1995). Molecular-cloning and tissue expression of FAT, the human homolog of the *Drosophila fat* gene that is located on chromosome 4q34-q35 and encodes a putative adhesion molecule. *Genomics,* **30**, 207-223.

Eddy, S. R. (1996). Hidden Markov models. *Curr. Opin. Struct. Biol.* **6**, 361-365.

Grunwald, G. B. (1993). The structural and functional analysis of calcium-dependent cell adhesion molecules. *Curr. Opin. Cell Biol.* **5**, 797-803.

Gumbiner, B. M. (1996). Cell adhesion: the molecular basis of tissue architecture and morphogenesis. *Cell,* **84**, 345-357.

Holm, L. & Sander, C. (1998). Removing near-neighbour redundancy from large protein sequence collections. *Bioinformatics,* **14**, 423-429.

Hutter, H., Vogel, B. E., Plenefisch, J. D., Norris, C. R., Proenca, R. B., Spieth, J., Guo, C. B., Mastwal, S., Zhu, X. P., Scheel, J. & Hedgecock, E. M. (2000). Cell biology: conservation and novelty in the evolution of cell adhesion and extracellular matrix genes. *Science,* **287**, 989-994.

Hynes, R. O. & Zhao, Q. (2000). The evolution of cell adhesion. *J. Cell Biol.* **150**, F89-F96.

Iwai, Y., Usui, T., Hirano, S., Steward, R., Takeichi, M. & Uemura, T. (1997). Axon patterning requires *D*N-cadherin, a novel neuronal adhesion receptor, in the *Drosophila* embryonic CNS. *Neuron,* **19**, 77-89.

Karplus, K., Barrett, C. & Hughey, R. (1998). Hidden Markov models for detecting remote protein homologies. *Bioinformatics,* **14**, 846-856.

Koch, P. J. & Franke, W. W. (1994). Desmosomal cadherins - another growing multigene family of adhesion molecules. *Curr. Opin. Cell Biol.* **6**, 682-687.

Krogh, A., Brown, M., Mian, I. S., Sjolander, K. & Haussler, D. (1994). Hidden Markov models in computational biology: applications to protein modeling. *J. Mol. Biol.* **235**, 1501-1531.

Levi, L., Douek, J., Osman, M., Bosh, T. C. G. & Rinkevich, B. (1997). Cloning and characterization of BS-cadherin, a novel cadherin from the colonial urochordate *Botryllus schlosseri*. *Gene,* **200**, 117-123.

Mahoney, P. A., Weber, U., Onofrechuk, P., Biessmann, H., Bryant, P. J. & Goodman, C. S. (1991). The *fat* tumor suppressor gene in *Drosophila* encodes a novel member of the cadherin gene superfamily. *Cell,* **67**, 853-868.

Nagar, B., Overduin, M., Ikura, M. & Rini, J. M. (1996). Structural basis of calcium-induced E-cadherin rigidification and dimerization. *Nature,* **380**, 360-364.

Nagase, T., Ishikawa, K., Suyama, M., Kihuno, R., Hirosawa, M., Miyajima, N., Tanaka, A., Kotani, H., Nomura, N. & Ohara, O. (1998). Prediction of the coding sequences of unidentified human genes. XII. The complete sequences of 100 new cDNA clones from brain which code for large proteins *in vitro*. *DNA Res.* **5**, 355-364.

Nielsen, H., Engelbrecht, J., Brunak, S. & von Heijne, G. (1997). Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng.* **10**, 1-6.

Nollet, F., Kools, P. & vanRoy, F. (2000). Phylogenetic analysis of the cadherin superfamily allows identification of six major subfamilies besides several solitary members. *J. Mol. Biol.* **299**, 551-572.

Nose, A., Tsuji, K. & Takeichi, M. (1990). Localization of specificity determining sites in cadherin cell-adhesion molecules. *Cell,* **61**, 147-155.

Oda, H. & Tsukita, S. (1999). Nonchordate classic cadherins have a structurally and functionally unique domain that is absent from chordate classic cadherins. *Dev. Biol.* **216**, 406-422.

Oda, H., Uemura, T., Harada, Y., Iwai, Y. & Takeichi, M. (1994). A *Drosophila* homolog of cadherin associated with armadillo and essential for embryonic cell-cell adhesion. *Dev. Biol.* **165**, 716-726.

Overduin, M., Harvey, T. S., Bagby, S., Tong, K. I., Yau, P., Takeichi, M. & Ikura, M. (1995). Solution structure of the epithelial cadherin domain responsible for selective cell-adhesion. *Science,* **267**, 386-389.

Park, J., Karplus, K., Barrett, C., Hughey, R., Haussler, D., Hubbard, T. & Chothia, C. (1998). Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J. Mol. Biol.* **284**, 1201-1210.

Pearson, W. R. (1998). Empirical statistical estimates for sequence similarity searches. *J. Mol. Biol.* **276**, 71-84.

Pettitt, J., Wood, W. B. & Plasterk, R. H. A. (1996). *cdh-3*, a gene encoding a member of the cadherin superfamily, functions in epithelial cell morphogenesis in *Caenorhabditis elegans*. *Development,* **122**, 4149-4157.

Ponassi, M., Jacques, T. S., Ciani, L. & FfrenchConstant, C. (1999). Expression of the rat homologue of the *Drosophila fat* tumour suppressor gene. *Mech. Devel.* **88**, 127.

Ranscht, B. (1994). Cadherins and catenins-interactions and functions in embryonic-development. *Curr. Opin. Cell Biol.* **6**, 740-746.

Ringwald, M., Schuh, R., Vestweber, D., Eistetter, H., Lottspeich, F., Engel, J., Dolz, R., Jahnig, F., Epplen, J., Mayer, S., Muller, C. & Kemler, R. (1987). The structure of cell-adhesion molecule uvomorulin - insights into the molecular mechanism of Ca^{2+} dependent cell-adhesion. *EMBO J.* **6**, 3647-3653.

Sano, K., Tanihara, H., Heimark, R. L., Obata, S., Davidson, M., Stjohn, T., Taketani, S. & Suzuki, S. (1993). Protocadherins-a large family of cadherin-related molecules in central nervous system. *EMBO J.* **12**, 2249-2256.

Schultz, J., Milpetz, F., Bork, P. & Ponting, C. P. (1998). Smart, a simple modular architecture research tool: Identification of signaling domains. *Proc. Nat. Acad. Sci. USA,* **95**, 5857-5864.

Schultz, J., Copley, R. R., Duerks, T., Ponting, C. P. & Bork, P. (2000). SMART: a web-based tool for the study of genetically mobile domains. *Nucl. Acids Res.,* **28**, 231-234.

Shapiro, L., Fannon, A. M., Kwong, P. D., Thompson, A., Lehmann, M. S., Grubel, G., Legrand, J. F., Alsnielsen, J., Colman, D. R. & Hendrickson, W. A.

(1995). Structural basis of cell-cell adhesion by cadherins. *Nature,* **374**, 327-337.

Sonnhammer, E. L. L., Von Heijne, G. & Krogh, A. (1998). A hidden Markov model for predicting transmembrane helices in protein sequences. In *Proc. of Sixth Conf. I. S. M. B*. (Glasgow, J., Littlejohn, T., Major, F., Lathrop, R., Sankoff, D. & Sensen, C., eds), pp. 175-182, AAAI Press, Mendo Park, CA.

Sugino, H., Hamada, S., Yusuda, R., Tuji, A., Matsuda, Y., Jujita, M. & Yagi, T. (1999). Genomic organization of the family of CNR cadherin genes in mice and humans. *Genomics,* **63**, 75-87.

Suzuki, S. T. (1996). Protocadherins and diversity of the cadherin superfamily. *J. Cell Sci.* **109**, 2609-2611.

Takahashi, M. & Cooper, G. M. (1987). Ret transforming gene encodes a fusion protein homologous to tyrosine kinases. *Mol. Cell. Biol.* **7**, 1378-1385.

Takeichi, M. (1995). Morphogenetic roles of classic cadherins. *Curr. Opin. Cell Biol.* **7**, 619-627.

Teichmann, S. A. & Chothia, C. (2000). Immunoglobulin superfamily proteins in *Caenorhabditis elegans*. *J. Mol. Biol.* **296**, 1367-1383.

Tepass, U. (1999). Genetic analysis of cadherin function in animal morphogenesis. *Curr. Opin. Cell Biol.* **11**, 540-548.

Tepass, U., Gruszynski-DeFeo, E., Haag, T. A., Omatyar, L., Torok, T. & Hartenstein, V. (1996). Shotgun encodes *Drosophila* E-cadherin and is preferentially required during cell rearrangement in the neurectoderm and other morphogenetically active epithelia. *Genes Dev.* **10**, 672-685.

Uemura, T., Oda, H., Kraut, R., Hayashi, S., Kataoka, Y. & Takeichi, M. (1996). Zygotic *Drosophila* E-cadherin expression is required for processes of dynamic epithelial cell rearrangement in the *Drosophila* embryo. *Genes Dev.* **10**, 659-671.

Usui, T., Shima, Y., Shimada, Y., Hirano, S., Burgess, R. W., Schwarz, T. L., Takeichi, M. & Uemura, T. (1999). Flamingo, a seven-pass transmembrane cadherin, regulates planar cell polarity under the control of frizzled. *Cell*, **98**, 585-595.

Wu, Q. & Maniatis, T. (1999). A striking organization of a large family of human neural cadherin-like cell adhesion genes. *Cell*, **97**, 779-790.

Yagi, T. & Takeichi, M. (2000). Cadherin superfamily genes: functions, genomic organisation, and neurological diversity. *Genes Dev.* **14**, 1169-1180.

Edited by G. von Heijne

(Received 17 August 2000; received in revised form 28 November 2000; accepted 29 November 2000)