Journal of Structural and Functional Genomics **2:** 135–143, 2002.

The cadherin superfamily database

Kevin Truong & Mitsuhiko Ikura*
*Division of Molecular and Structural Biology, Ontario Cancer Institute and Department of Medical Biophysics, University of Toronto, Toronto, Ontario, Canada; *Author for correspondence (e-mail: mikura@uhnres.utoronto.ca; fax: (416) 946-2055 or 6529)*

Received 22 January 2002; accepted in revised form 23 April 2002

Key words: cadherins; classification; hidden Markov model; multi-domain architecture; protein families

Abstract

The cadherin superfamily is a large protein family with diverse structures and functions. Because of this diversity and the growing biological interest in cell adhesion and signaling processes, in which many members of the cadherin superfamily play a crucial role, it is becoming increasingly important to develop tools to manage, distribute and analyze sequences in this protein family. Current profile and motif databases classify protein sequences into a broad spectrum of protein superfamilies, however to provide a more specific functional annotation, the next step should include classification of subfamilies of these protein superfamilies. Here, we present a tool that classified greater than 90% of the proteins belonging to the cadherin superfamily found in the SWISS PROT database. Therefore, for most members of the cadherin superfamily, this tool can assist in adding more specific functional annotations than can be achieved with current profile and motif databases. Finally, the classification tool and the results of our analysis were integrated into a web-accessible database (http://calcium.uhnres.utoronto.ca/cadherin).

Introduction

Proteins in the cadherin superfamily are transmembrane glycoproteins that are involved in many biological functions such as cell–cell adhesion, morphogenesis, synapse formation, cell polarization, cell sorting, cell migration, and cell rearrangements [1–8]. Some members of the cadherin superfamily have even been implicated as proto-oncogenes or tumor suppressors [9, 10]. All these members of the cadherin superfamily share an extracellular cadherin repeat (CR), an approximately 110 amino acid peptide that mediates Ca^{2+}-dependent homophilic interactions between cadherin molecules. CRs assume an immunoglobulin-like β-sandwich fold (Figure 1), which usually occur in tandem and are separated by a linker region that binds three Ca^{2+} ions [11–15]. From the raw sequence data of the various genome projects, it is clear that many sequences have CRs; however, the annotation of the specific biological function is unclear as cadherins are involved in many diverse functions.

Typically, the biological function can be inferred from its similarity to sequences of known function in sequence databases using single-sequence similarity algorithms such as BLAST [16] and FASTA [17]. Such algorithms are suitable for determining highly similar sequences, but are not sensitive enough to capture highly divergent sequences. Therefore, many members of an evolutionarily diverse family of proteins may be overlooked. Within the last decade, the sensitivity of sequence searching techniques has been improved by profile- or motif-based analysis, which uses information derived from multiple sequence alignments (MSAs) to construct and search for sequence patterns [18–20]. Unlike single-sequence similarity, a profile or motif can exploit additional information, such as the position and identity of residues that are conserved throughout the family, as well as variable insertion and deletion probabilities. The hidden Markov model is one powerful way to express a profile or motif because it provides a solid statistical foundation to model information in an MSA [20].

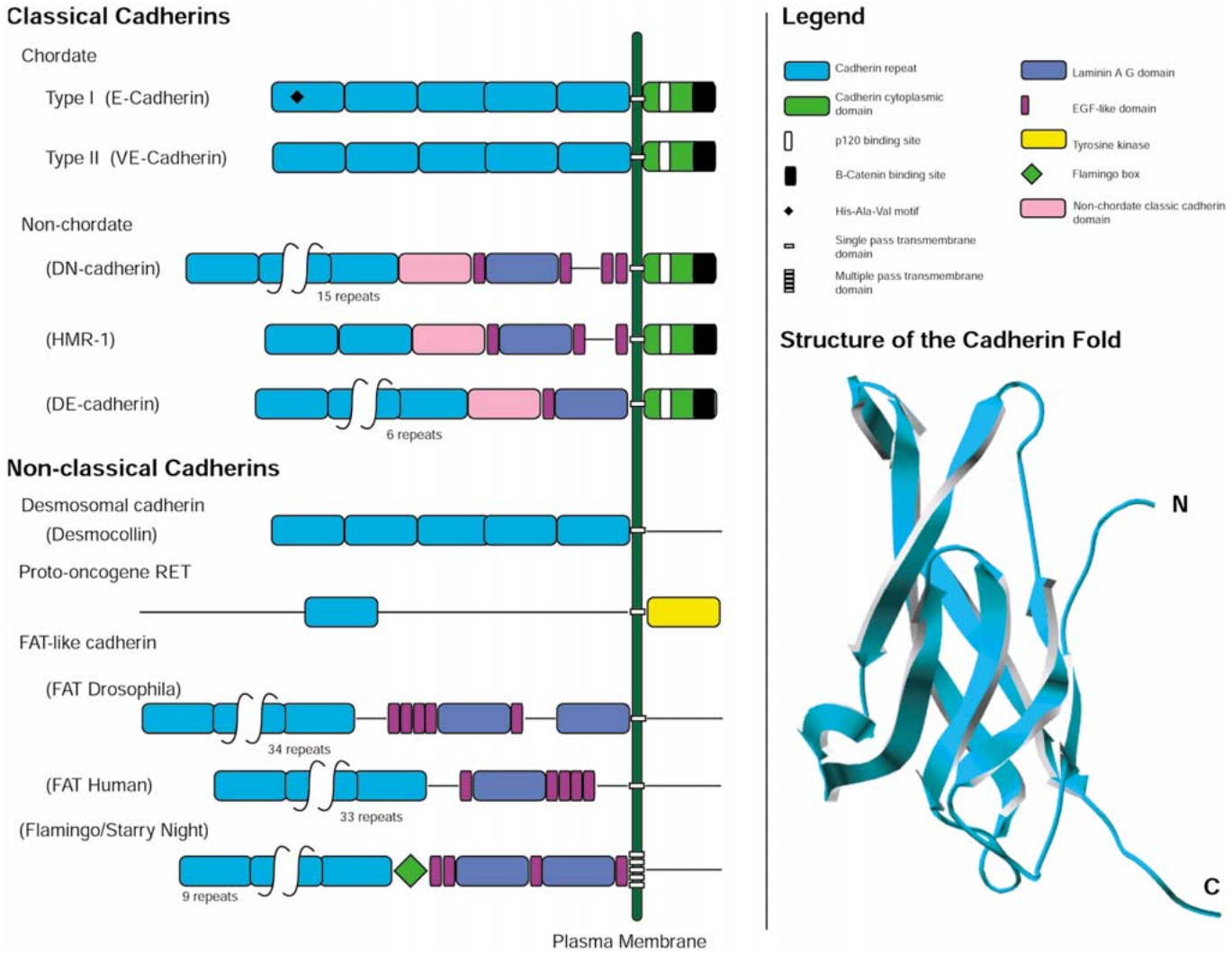

Figure 1 Domain organization of representative members of the cadherin superfamily. Classical cadherins have a well-conserved cytoplasmic domain and can be grouped into chordate and non-chordate classic cadherins. Non-chordate cadherin can be further subdivided into type I and type II, with type I cadherins having a His–Ala–Val sequence in the N-terminal CR [8]. The desmosomal cadherins contain five extracellular CRs and a cytoplasmic domain that interacts with plakoglobin, desmoplakin and the plakophillins [34]. Proto-oncogene RET has one CR in the extracellular domain and a tyrosine kinase domain in its cytoplasmic domain [35]. Fat-like cadherins have a very large extracellular region with 19–34 CRs. Flamingo cadherins have 8–9 CRs, 2 LG domains, and 4 EGF domains and a domain called the Flamingo box [36], located between the last CR and the first EGF domain, which is highly conserved across species in this subfamily. On the bottom right of the figure, the ribbon diagram of the N-terminal cadherin repeat of epithelial cadherin is shown (PDB code 1EDH).

Most profile and motif databases, such as BLOCKS [19], PROSITE [18], Pfam [21], SMART [22], PRINTS [23] and InterPro [24], have patterns that define the CR which can be used to classify protein sequences into the cadherin superfamily. The next step of such database studies should include the development of classification systems capable of distinguishing between subfamilies within a structurally and functionally diverse superfamily, like cadherins. This would be helpful in elucidating sequence–structure–function relationships of proteins as specific classification results in more specific functional annotations (Figure 2). Here, we report a web-accessible classification tool and database that provides a specific classification of known subfamilies of the cadherin superfamily using two methods: multidomain architecture analysis and HMM signatures [25]. This work represents the first classification tool and database focused on a specific protein superfamily.

Methods

Multidomain architecture analysis

All HMMs for commonly occurring domains were created from MSAs using the hmmbuild program in the HMMER software package [32]. The hmmpfam program in HMMER was used to predict the domain layout of a query protein sequence. Finally, Perl regular expressions were used to describe the domain pattern of each subfamily.

PROPERTY	VALUE
Region Name	tcr_TruncatedCadherin
Classification	Cadherin;Truncated_Cadherin
Type	Family
Description	Found in neurons and muscle cells, T-cadherin is an atypical member of the cadherin superfamily. Unlike other cadherins, T (truncated)-cadherin has neither a cytoplasmic domain nor a transmembrane region. Instead, T-cadherin is localized to the plasma membrane by a glycophosphoinositol-lipid anchor. T-cadherin is present in two forms, one pretruncated and one truncated (120/100 kDa in sheep, 130/105 kDa in humans). The two forms are thought to have a precursor-product relationship. Although it is not yet known whether both isoforms are functional, both are expressed on the cell surface. T-cadherin can act as an adhesion molecule even though it does not interact as strongly as other cadherins. The most recent evidence suggests that T-cadherin negatively regulates the cell growth.
HMM Histogram	Click to view
Literature	Doyle DD, *et al*, J Biol Chem 1998 Mar 20;273(12):6937-43
Literature	Takeuchi T, *et al*, J Neurochem 2000 Apr;74(4):1489-97
Literature	Kuzmenko YS, *et al*, FEBS Lett 1998 Aug 28;434(1-2):183-7
Sequence	1381790 (U59288) H-cadherin
Sequence	1381792 (U59289) H-cadherin
Sequence	212709 (M81779) T-cadherin
Sequence	3434957 (AB001103) H-cadherin

Figure 2 Subfamily annotation. Detailed annotations of subfamilies were created. They include information about the function of the subfamily, their relative classification hierarchy, useful literature links, the HMM histogram if applicable and sequences in our database that belong to this subfamily.

HMM signatures

MSAs of subfamilies were created using CLUSTALW [33]. Only the MSAs sharing a 40% sequence identity were used to construct an HMM histogram [25]. From the HMM histogram, HMM signatures were extracted. The hmmpfam program in HMMER was also used to predict HMM signatures of a query protein sequence.

Databases

A relational database was designed to store our subfamily annotations and the results from our analysis of a number of genomes and general genomic databases. The database was implemented in an Oracle 8 relational database management system running on a computer machine with a dual 750-MHz UltraSPARC-III processor and 4G of RAM running SunOS 5.8. It consisted of three main tables: a sequence table, region table and region information table (Figure 3). The sequence table stores general information about the sequence: database source, organism species, descriptions, primary sequence, etc. The region (or domain) information table stores general information about regions: descriptions, literature references, etc. The region table stores the domain layout: start position, ending position, e-value, etc.

Web interface

For wide access to the classification tool and database, a web interface was created (http://calcium.uhnres.utoronto.ca/cadherin). The purpose of the web site is to manage, distribute and analyze information on the cadherin superfamily. The web site has three distinct sections: general, search, and classify. The general section contains a recent compilation of literature and structural information about cadherins as well as a synopsis of methods used in the data analy-

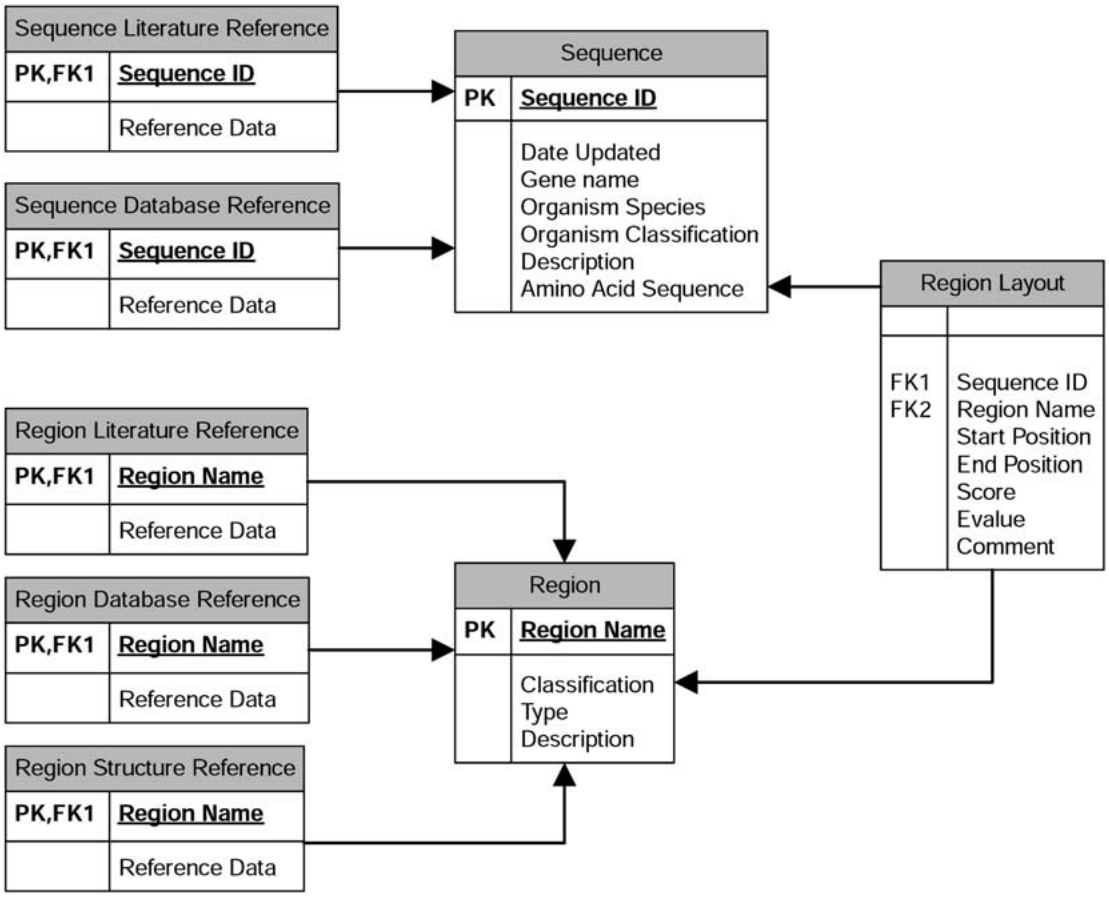

Figure 3. Design of the relational database. The shaded area of each table contains the table name. The black arrows indicate the relationships between tables in the database.

sis. The search section allows the user to submit a single protein sequence for classification. The result of the analysis is a prediction of the protein subfamily based on HMM signatures and multidomain architecture analysis (Figure 4) and a domain layout table with a corresponding alignment to the HMM. The classify section is the conduit to the underlying relational database which supports keyword queries to all tables in the database.

Results and discussion

HMMs for the cadherin superfamily

In our analysis, a protein sequence was considered a member of the cadherin superfamily if it contained a CR. An HMM defining the CR can be found on the Pfam or SMART databases, however they are based on MSAs between CRs found in various species and cadherin subfamilies. Recent biophysical studies showed a cell–cell adhesion interface with variable degrees of antiparallel overlap between multiple CRs [26, 27], which suggests that the position with respect to the N-terminus of the five CRs of chordate classical cadherins is important to its specific role in cell–cell adhesion. To capture the significance of the CR position in chordate classic cadherins, five HMMs were created from MSAs of each repeat. Additionally, an HMM was created for non-chordate CRs. If a query protein sequence had a match for any of the six HMMs for CRs, it was assigned to the cadherin superfamily.

Using the above method for finding proteins in the cadherin superfamily, a number of genomes and general genomic databases were searched, including all available bacterial genomes, the yeast genome [28], the fly genome [29], the worm genome [30], the human genome [31], and the SWISS PROT database (Release 39) (Table 1). No cadherins were found in any of the bacterial genomes or in the yeast genome. This observation strongly supports the hypothesis that transmembrane proteins of the cadherin superfamily have evolved to meet the need for complex cell interactions required for the multicellular organization of metazoans.

*Table 1.*Tabulation of cadherins in various genomes and databases.

Genomes	
Data source	Matches
Worm	11
Human	328
Fly	15
General sequence database	
Species	SWISS PROT
Bombyx mori	1
Bos taurus	5
Botryllus schlosseri	
Brachydanio rerio	5
Caenorhabditis elegans	10
Canis familiaris	1
Ciona intestinalis	
Cricetulus griseus	
Danio rerio	
Drosophila melanogaster	4
Gallus gallus	12
Homo sapiens	108
Lytechinus variegatus	1
Mus musculus	30
Mus sp.	
Oryctolagus cuniculus	1
Rattus norvegicus	13
Rattus rattus	
Rattus sp.	
Sus scrofa	2
Synechocystis sp.	1
Tetraodon fluviatilis	1
Xenopus laevis	8
Matches	203

Classification using multidomain architecture analysis

One method used for the classification of cadherin protein sequences into subfamilies was multidomain architecture analysis. Based on our study of the primary structure, there are several key domains in the cadherin superfamily which were predicted by carefully constructed HMMs (Table 2). Certain subfamilies of cadherins had a characteristic arrangement of these domains within the primary sequence (Figure 1). For example, classical cadherins were identified by their well-conserved cytoplasmic domain which interact with the catenins, β-catenin/Armadillo

Table 2. Tabulation of commonly occurring domains in the cadherin superfamily.

Domain name	Number of sequences	Length[a]
Cadherin repeat 1	38	111
Cadherin repeat 2	44	115
Cadherin repeat 3	41	118
Cadherin repeat 4	40	109
Cadherin repeat 5	43	120
Non-chordate cadherin repeat	55	97
Classical cytoplasmic domain	44	168
Desmosomal cytoplasmic domain	10	183
Desmoglein repeat	18	29
Epidermal growth factor (EGF) domain	87	45
Flamingo box	6	211
Laminin B domain	9	148
Laminin EGF domain	72	59
Laminin G domain	22	161
Non-chordate classical cadherin domain (NCCD)	4	184
Protein kinase domain	6	276

[a]the number of amino acid residues needed to define the domain

and p120^{ctn}/δ-catenin, and proto-oncogene RET were identified by a single CR in the extracellular domain and a tyrosine kinase domain in its cytoplasmic domain [8].

Many cadherin subfamilies were identified, such as the ones above, by their multidomain architecture. Using multidomain architecture analysis, 73% of cadherin sequences in the SWISS PROT database were classified into subfamilies. Within genomes, 18% were classified in the worm, 12% in the human and 40% in the fly. Although the classification percentages in the worm, fly and human genomes are small, they are optimal because many cadherins in these species are novel and do not fall under any well-defined subfamily. As more research is done in this area, we expect a clearer classification to emerge.

Classification by HMM signatures

The second method involved using HMMs to find the smallest windows of residues in an MSA of a subfamily that is necessary to significantly differentiate it from the other subfamilies (called signatures). Once signatures were identified, HMMs were built from the corresponding MSA segments and used to find the signatures in query protein sequences. The detection

PREDICTED FAMILY BY HMM: EpithelialCadherin
PREDICTED FAMILY BY LAYOUT: TypeIClassicCadherin
SEQUENCE LENGTH: 887

DOMAIN	FROM	TO	SCORE	EVALUE
dom_cad_repeat_1	161	267	236.6	3.5e-67
dom_cad_repeat_2	269	380	238.8	7.5e-68
dom_cad_repeat_3	382	492	218.4	1.1e-61
dom_cad_repeat_4	494	598	212.5	6.5e-60
dom_cad_repeat_5	600	704	148.2	1.5e-40
dom_classic_cytoplasmic	735	886	352.6	4.4e-102

Figure 4. Results from a sequence search. This figure shows the results of a sequence search of an epithelial cadherin protein sequence. The database returns the predicted subfamily based on HMM signatures (denoted HMM) and also multidomain architecture analysis (denoted by layout). It outputs a table of the domain layout, which includes the name of the domain, the starting position, ending position, score and e-value. Finally, it outputs the alignment between the domains and the query sequence.

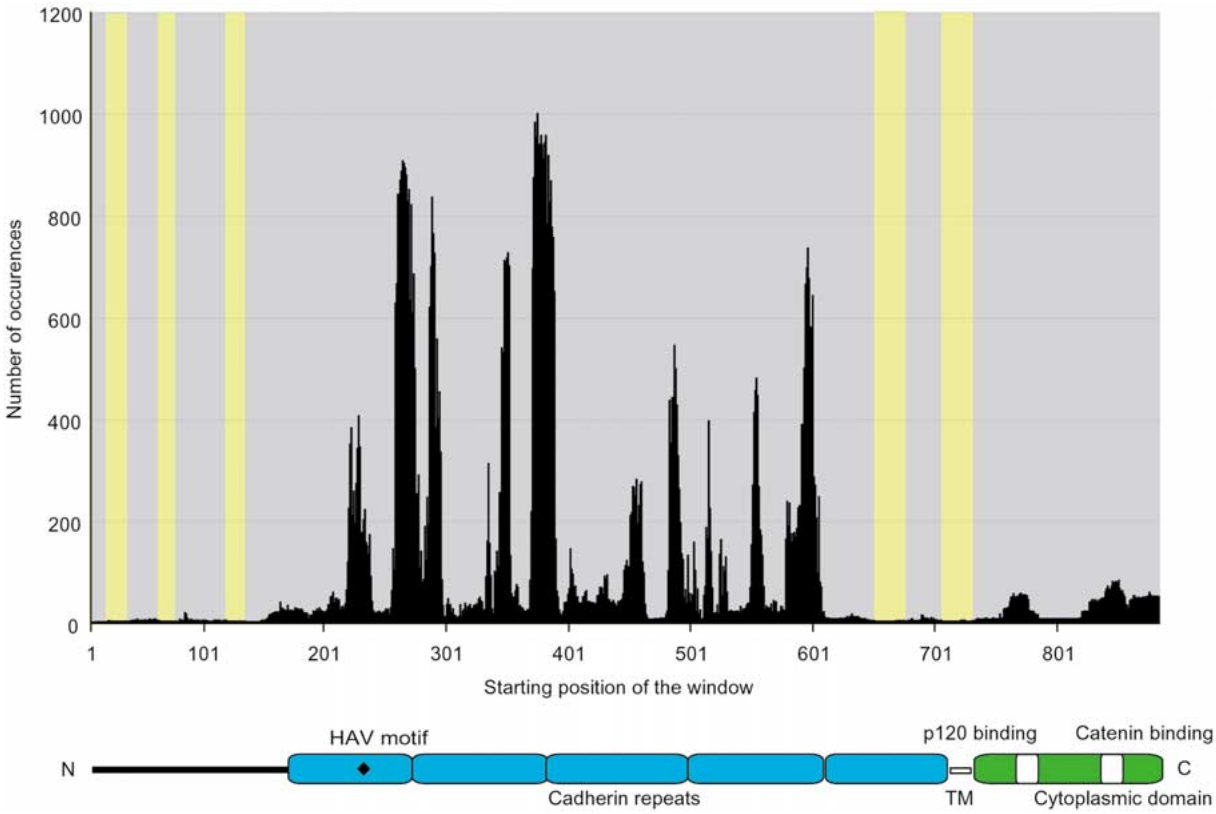

Figure 5. HMM histogram of the epithelial cadherin subfamily with a window size of 20. The MSA of the epithelial cadherin subfamily is 893 residues wide and therefore there are 874 possible windows of 20 residues. The *x*-axis plots the starting position of the window while the y-axis plots the number of occurrences. The yellow regions (15–33, 58–73, 113–127, 642–668, 697–723) are the HMM signatures of the subfamily because they include only epithelial cadherins. Below the HMM histogram, the layout of the epithelial cadherin relative to the starting position of the windows is shown. There are a large number of matches in the important sites in cadherins, such as the HAV motif, p120 binding site and catenin binding sites because these regions are well conserved throughout the superfamily. This suggests that HMM histogram is not only useful for finding signature regions but also regions of high conservation which may have functional significance.

of a signature in a query protein sequence implied its membership to the corresponding subfamily. In short, the method to find signatures started with an MSA of the subfamily that was used to build an HMM database representing all sliding windows of the MSA of a fixed size. Then, an HMM histogram was built from the number of matches from HMM searches of each sliding window to the protein sequence database of the superfamily. From an HMM histogram, subfamily signatures were identified because they had an equal number of matches to the number of sequences in the subfamily (Figure 5). The method to find signatures is described in detail elsewhere [25].

This method was used to find signatures in cadherin subfamilies where an MSA could be constructed with a greater than 40% sequence identity because at that level of sequence identity, the MSA is structurally correlated [25]. Ninety-five HMM signatures were found in 21 cadherin subfamilies as many subfamilies, could be characterized with multiple HMM signatures. A cadherin protein was classified into a particular subfamily if one of the subfamily's HMM signatures was found. Using HMM signatures, 71% of cadherin sequences were classified into subfamilies in the SWISS PROT database. Within genomes, no cadherins were classified in the worm and fly because their cadherin sequences are variable in length and, therefore, do not align well globally, making it difficult to create HMM histograms to find signatures. However, 38% in the human genome were classified. Using HMM signatures, many subfamilies of cadherins were identified with greater specificity than with multidomain architecture analysis. For example, we were able to distinguish between various subfam-

Table 3. Tabulation of results by the combination of classification techniques.

	Genomes			General Database SWISS PROT
	Worm	Human	Fly	
Family				
Arcadlin				3
Classic cadherin		4		3
Desmocollin type I				3
Desmosomal cadherin		1		10
Epithelial cadherin		1		7
FAT-like cadherin	1		1	3
Flamingo cadherin		5	1	3
Kidney cadherin		3		7
Kidney specific cadherin		4		2
Liver intestine cadherin		2		4
Muscle cadherin		2		2
Neural cadherin		2		12
Non-chordate classic cadherin	1		3	4
Osteoblast cadherin		2		5
PB cadherin		2		2
Placental cadherin		1		2
Protocadherin-α		90		28
Protocadherin-β		10		14
Protocadherin-γA		5		26
Protocadherin-γB		1		10
Protocadherin-γC				8
Protocadherins		11		5
Truncated cadherin				4
Type I classic cadherin				1
Type II classic cadherin				6
Tyrosine kinase receptor			1	6
Vascular endothelial cadherin		1		5
Unknown	**9**	**181**	**9**	**18**
Total	**11**	**328**	**15**	**203**

ilies of chordate classical cadherins such as epithelial and neural cadherin.

Combining both methods

By combining both HMM signatures and multidomain architecture analysis, we can exploit the advantages of each method to improve the overall classification. The HMM signatures are very specific and are able to classify subfamilies that are similar to each other. In contrast, multidomain architecture analysis is able to classify proteins with distinct domain layouts. The order of precedence in the combined classification is HMM signatures, and then the multidomain architecture analysis.

Using the combined methods, we achieved a 91% coverage in the classification of cadherin sequences in the SWISS PROT database compared to 73% and 71% using only multidomain architecture analysis or HMM signatures, respectively (Table 3). Within genomes, 18% were classified in the worm, 45% in the human and 40% in the fly. Multidomain architecture analysis was able to classify proteins with varying sequence lengths like FAT and flamingo cadherins that HMM signatures missed, while HMM signatures was able to classify similar subfamilies, like epithelial and neural cadherins that multidomain architecture analysis missed. Thus, the combination of both methods significantly improved the classification.

Conclusions

Over 300 cadherin protein sequences were identified in the human genome, 15 in the fly genome and 11 in the worm genome, many of these sequences were classified into one of 27 subfamilies. By using HMM signatures and multidomain architecture analysis, our classification tool and database classifies query protein sequences into subfamilies that are richly annotated in our web-accessible database. Future work in this area could involve applying the methods to create similar web-accessible databases of other large protein superfamilies. Such highly specific databases could play an important role in the automatic annotation of genomes in complement with other popular databases like BLOCKS, PROSITE, Pfam, SMART, PRINTS and InterPro.

Acknowledgements

We thank Marc Sherman for his help in annotating the cadherin subfamilies and Jean-Rene Alattia and Jane Gooding for critical reading of the manuscript. This work was supported by a grant to M.I. from the National Cancer Institute of Canada. M.I. is a Canadian Institutes of Health Research Scientist.

References

1. Takeichi, M. (1991) *Science* **251**, 1451–145.
2. Nollet, F., Kools, P., and van Roy, F. (2000) *J. Mol. Biol.* **299**, 551–572.
3. Yagi, T., and Takeichi, M. (2000) *Genes Dev.* **14**, 1169–1180.
4. Tepass, U. (1999) *Curr. Opin. Cell Biol.* **11**, 540–548.
5. Nose, A., Nagafuchi, A., and Takeichi, M. (1988) *Cell* **54**, 993–1001.
6. Steinberg, M.S., and Takeichi, M. (1994) *Proc. Natl. Acad. Sci. USA* **91**, 206–209.
7. Godt, D., and Tepass, U. (1998) *Nature* **395**, 387–391.
8. Tepass, U., Truong, K., Godt, D., Ikura, M., and Peifer, M. (2000) *Nat. Rev. Mol. Cell Biol.* **1**, 91–100.
9. Mulligan, L.M., Kwok, J.B., Healey, C.S. *et al.* (1993) *Nature* **363**, 458–460.
10. Mahoney, P.A., Weber, U., Onofrechuk, P., Biessmann, H., Bryant, P.J., and Goodman, C.S. (1991) *Cell* **67**, 853–868.
11. Overduin, M., Harvey, T.S., Bagby, S. *et al.* (1995) *Science* **267**, 386–389.
12. Shapiro, L., Fannon, A.M., Kwong, P.D. *et al.* (1995) *Nature* **374**, 327–337.
13. Nagar, B., Overduin, M., Ikura, M., and Rini, J.M. (1996) *Nature* **380**, 360–364.
14. Tamura, K., Shan, W.S., Hendrickson, W.A., Colman, D.R., and Shapiro, L. (1998) *Neuron* **20**, 1153–1163.
15. Pertz, O., Bozic, D., Koch, A.W., Fauser, C., Brancaccio, A., and Engel, J. (1999) *EMBO J.* **18**, 1738–1747.
16. Altschul, S.F., Madden, T.L., Schaffer, A.A. *et al.* (1997) *Nucleic Acids Res.* **25**, 3389–33402.
17. Pearson, W.R. (1994) *Meth. Mol. Biol.* **25**, 365–389.
18. Hofmann, K., Bucher, P., Falquet, L., and Bairoch, A. (1999) *Nucleic Acids Res.* **27**, 215–219.
19. Henikoff, S., Henikoff, J.G., and Pietrokovski, S. (1999) *Bioinformatics* **15**, 471–479.
20. Eddy, S.R. (1998) *Bioinformatics* **14**, 755–763.
21. Bateman, A., Birney, E., Durbin, R., Eddy, S.R., Howe, K.L., and Sonnhammer, E.L. (2000) *Nucleic Acids Res.* **28**, 263–266.
22. Schultz, J., Milpetz, F., Bork, P., and Ponting, C.P. (1998) *Proc. Natl. Acad. Sci. USA* **95**, 5857–5864.
23. Attwood, T.K., Croning, M.D., Flower, D.R. *et al.* (2000) *Nucleic Acids Res.* **28**, 225–227.
24. Apweiler, R., Attwood, T.K., Bairoch, A. *et al.* (2000) *Bioinformatics* **16**, 1145–1150.
25. Truong, K., and Ikura, M. (2002) *BMC Bioinformatics* **3**, 1.
26. Sivasankar, S., Brieher, W., Lavrik, N., Gumbiner, B., and Leckband, D. (1999) *Proc. Natl. Acad Sci. USA* **96**, 11820–11824.
27. Sivasankar, S., Gumbiner, B., and Leckband, D. (2001) *Biophys J.* **80**, 1758–1768.
28. Mewes, H.W., Albermann, K., Bahr, M. *et al.* (1997) *Nature* **387**, 7–65.
29. Adams, M.D., Celniker, S.E., Holt, R.A. *et al.* (2000) *Science* **287**, 2185–2195.
30. Hodgkin, J., Plasterk, R.H., and Waterston, R.H. (1995) *Science* **270**, 410–414.
31. Lander, E.S., Linton, L.M., Birren, B. *et al.* (2001) *Nature* **409**, 860–921.
32. Eddy, S.R. (1995) *The HMMER Package*. Washington University, Washington.
33. Thompson, J.D., Higgins, D.G., and Gibson, T.J. (1994) *Nucleic Acids Res.* **22**, 4673–4680.
34. Kowalczyk, A.P., Bornslaeger, E.A., Norvell, S.M., Palka, H.L., and Green, K.J. (1999) *Int. Rev. Cytol.* **185**, 237–302.
35. Kuma, K., Iwabe, N., and Miyata, T. (1993) *Mol. Biol. Evol.* **10**, 539–551.
36. Usui, T., Shima, Y., Shimada, Y. *et al.* (1999) *Cell* **98**, 585–595.