Cadherin-7

Introduction

The Cadherin family is made up of transmembrane proteins which are designed to provide cell-to-cell adhesion, tumour suppression, tissue morphogenesis and cell recognition [10]. They have three main functions to accomplish this. These relate to decreasing interfacial tension upon contact to promote contact expansion, and stabilising contact from mechanical forces that might pull on the contact [1]. Cadherins contain five Cadherin repeats and a cytoplasmic tail [2], both of which are visible in Figure 1. Cadherin behaviour is used to position cells properly in development and help separate tissue layers [4]. Cadherin absence has also been correlated with tumour growth. In particular, E-Cadherin loss is usually a cause of carcinomas and other epithelial types of cancer.

Cadherin-7 (CDH7) is a gene that encodes a type-II classical Cadherin protein [3]. These are characterised by a lack of a histidine-alanine-valine cell adhesion recognition sequence. Mutations of this gene have been linked with bipolar disease in humans. Further diseases that have been associated with the gene are Craniofacial-Deafness-Hand Syndrome and Choanal Atresia. There is a family of three Cadherin 7-like genes which includes CDH19 and CDH20. The CDH7 gene codes for a precursor protein and a preproprotein which in turn undergoes proteolytic processing to form the mature glycoprotein [11].

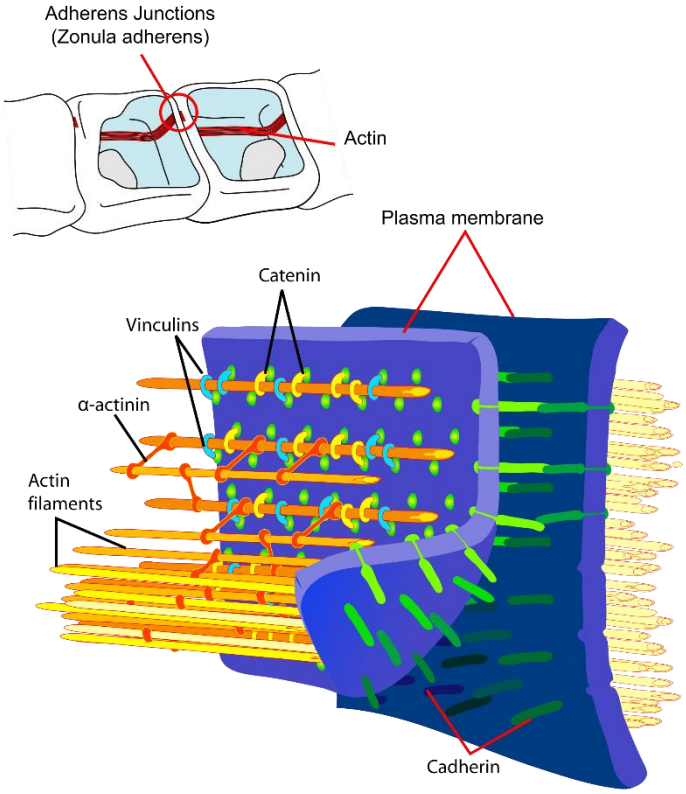

Figure 1 Demonstration of Cadherin's role in cell-to-cell adhesion with clear Cadherin regions (between cells), transmembrane regions (through the purple membranes) and cytoplasmic regions (just within the cells).
https://en.wikipedia.org/wiki/Cadherin#/media/File:Adherens_Junctions_structural_proteins.svg

Part 1

Methods

The Tables generated in this section used the NCBI GenBank (V 256) database to find the alternative transcripts, their respective protein isoforms and the accession IDs. From there I used Biopython to fetch the accession IDs and perform frequency analysis on nucleotide and amino acids present in each gene and protein respectively. In the case of nucleotide frequency, I summed the occurrences of each one and divided them by the total to get a percentage.

The abbreviation aa refers to amino acids.

Results

Local ID	Accession IDs - NCBI	A %	T %	C %	G %	Isoform	Length (nt)
1	NM_004361.5	33.0	30.8	17.5	18.7	1 preproprotein	12136
2	NM_033646.4	33.2	30.9	17.3	18.5	1 preproprotein	12126
3	NM_001317214.3	28.5	27.4	22.1	22.0	2 precursor	3407
4	NM_001362438.2	32.1	30.4	18.0	19.5	1 preproprotein	12938

Table 1 https://www.ncbi.nlm.nih.gov/datasets/gene/id/1005/products/ The alternate transcripts of the Human Cadherin-7 gene and its composition by each of the four base percentages (3sf)

Local ID	Nucleotide Accession ID - NCBI	Protein Accession ID - NCBI	Length (aa)	Top 3 Modal Amino Acids	Presence of HAV Sequence
1	NM_004361.5	NP_004352.2	785	Serine (66) Aspartic Acid(65) Leucine(59)	No
2	NM_033646.4	NP_387450.1	785	Serine(66) Aspartic Acid(65) Leucine(59)	No
3	NM_001317214.3	NP_001304143.1	630	Serine(52) Aspartic Acid(48) Valine(48)	No
4	NM_001362438.2	NP_001349367.1	785	Serine(66) Aspartic Acid(65) Leucine(59)	No

Table 2 Analysis of the amino acid frequencies in the protein encodings of the different Cadherin-7 (N-Cadherin) transcripts

Discussion and Extension – Background on Cadherins and Comparison with Paralog E-Cadherin

Protein Accession ID - NCBI	Isoform	Length (aa)	Top 3 Modal Amino Acids	Presence of HAV Sequence
NP_004351.1	1 preproprotein	882	Threonine (80) Leucine (76) Valine (70)	Yes
NP_001304113.1	2 precursor	821	Leucine (73) Threonine (72) Aspartic Acid (63)	Yes
NP_001304114.1	3	366	Leucine (41) Aspartic Acid (36) Alanine (28)	No
NP_001304115.1	4	227	Leucine (31) Aspartic Acid (26) Glutamic Acid (18)	No

Table 3 Analysis of the amino acid frequencies in the protein encodings of the different Cadherin-1 (E-Cadherin) transcripts

E-cadherin and N-cadherin are paralogs of the Cadherin cell adhesion protein family. E-Cadherin conversely is used for the establishment and maintenance of Adherens junctions in epithelial tissues which contributes to tissue structural integrity and regulation of some cellular processes. Despite being paralogs there is a clear distinction between the classifications of both transcripts, for example, Serine is the most prevalent amino acid in CDH7 alternatives but is far lower in frequency in CDH1. In its place, others such as Threonine become far more prevalent.

Please find more information in the introduction section.

Part 2

Methods

For pairwise alignment of nucleotides, I used NCBI's BLASTN on the nucleotide collection (2023/10/24) with the standard penalties of 2 for a gap and a score of 1 for a match. For the protein alignment, I used NCBIs BLASTP with a gap opening score of 11, an extension penalty of 1 and the BLOSUM62 matrix.

To find the exon lengths of each transcript I used the NCBI genome data viewer (V 5.3.13) [9] where there is a convenient way to view each transcript exon by exon with their respective lengths attached. From there I manually compared which exons were which and as such which were missing by their respective starting locations.

To compare the features present in proteins I used the NCBI protein database (2023/10/24) and manually compared the noted sites and regions. [7,8]

In the extension section, I used the Alphafold structure prediction website (V2.3.0) to generate images of what each isoform of Cadherin 7 (probably) looks like.

Results

1. The shortest and longest transcripts of CDH7 had a pairwise score of 4072 and query coverage of 17%. There was a 100% identity match meaning the smallest was fully contained within the largest.
2. The shortest and longest protein isoform sequences had a pairwise score of 1278, query coverage of 79% and a 100% identity match again.
3.

CDH7 Accession - NCBI	Exon Lengths (aa)
NM_004361.5	138, 406, 295, 120, 168, 188, 254, 137, 122, 118, 252, 9937
NM_033646.4	129, 210, 295, 120, 168, 188, 254, 137, 122, 118, 252, 9937
NM_001317214.3	139, 210, 295, 120, 168, 188, 254, 137, 122, 118, 1460
NM_001362438.2	941, 210, 295, 120, 168, 188, 254, 137, 122, 118, 252, 9937

4. The exon starting at location 65862666 is only of length 252 on the longest transcript but 1460 nucleotides long on the shortest transcript. The final exon of length 9937 is missing on the shortest transcript.
5. Both proteins have an initial signal peptide chain but the preproprotein has a mature peptide chain and a couple of extra regions. The regions that appear in the longer protein but not the shorter one are the Cadherin Cytoplasmic region and a transmembrane region. The transmembrane region goes from site 608-628 (where the numbers refer to positions of amino acids along the peptide chain) and the cytoplasmic region goes from 631 – 777.

Discussion and Extension – Protein Region Analysis

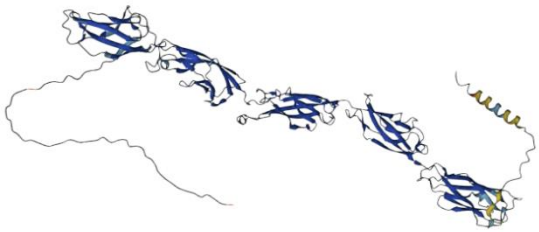

Figure 1 The alphafold structural prediction of the precursor isoform of CDH7

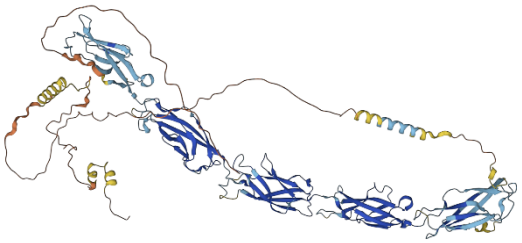

Figure 2 The alphafold structural prediction of the preproprotein isoform of CDH7

There is a clear distinction in structure between the two isoforms of the gene which reflects both the increased length and the additional features. One of these extra features is the cytoplasmic region [6] which has been found to regulate the cell binding function of the domain of the protein. The exclusion of this from the precursor is due to the lack of requirement for regulation as precursors by definition are inactive building blocks which undergo further modification after translation to become the final fully mature proteins. The other region included in the preproprotein is the transmembrane domain. This is the section that passes through the cell membrane, allowing the Cadherin cytoplasmic region to be within the cell's cytoplasmic side of the cell membrane. Once again the precursor protein doesn't have to contain anything that allows the Cadherin to enter the cell when it is not fit for use which is why this section is missing and as such is added when becoming ready to be activated.

Part 3

Methods

From filtering the Pfam database (V 36.0) in Interpro by proteins containing the Cadherin Domain, Homo Sapiens and "reviewed", there are 113 proteins that fit the criteria. Table 4 lists the first few results.

The pairwise sequence comparison in task two was scored using a BLOSUM62 matrix, Gap existence penalty of 11, extension penalty of 1 and a conditional compositional score matrix adjustment, on the NCBI BLASTP tool using only the Swissport (v.2023_04) database option.

I looked at the respective genes on the NCBI transcript table to find alternative transcripts and what proteins they formed.

From the multiple seemingly identically named protein isoforms I found when searching the different transcripts, I verified that they were identical chains with a Biopython script pairwise alignment algorithm and their respective accession IDs.

In the extension section, I searched for Cadherin 7 in the NCBI database and searched by taxonomy to manually select a variety of species that do not appear to have much in common at first glance to then compare proteins formed with each other where possible.

Results

1.

Accession ID - InterPro	Name
O60330	Protocadherin gamma-A12
O14917	Protocadherin-17
O602245	Protocadherin-7
O75309	Cadherin-16
A6H8M9	Cadherin-related family member 4

Table 4 First few results in the Pfam database when searching for human Cadherin domain-containing proteins

2.

Protein Name	Length (aa)	Identity %	Query Coverage %	Accession ID - Swissport
CDH20 (CDH7L3)	801	63.19	98	Q9HBT6.2
CDH6	790	62.89	96	P55285.1
CDH18	790	62.47	96	Q13634.1

Table 5 This table was adapted from the NCBI BLASTP results given when used on the Swissport database.

3.

a. CDH20 only has one verified protein isoform which is marked in Table 4.
b. CDH6 has the 2-preproprotein isoform NP_001349364.1 of length 663.
c. CDH18 has three unique smaller isoforms which are listed in table 5. Multiple transcripts yield identical isoforms.

CDH	Protein Accession	Length (aa)	Isoform	Identity %	Query Coverage %
6	NP_001349364.1	663	2 preproprotein	62.81	89
18	NP_001161139.1	574	2 precursor	62.64	90
18	NP_001278886.1	575	3 precursor	62.62	90

18	NP_001336491.1	567	4	60.74	87

Table 6 Comparison of smaller protein isoforms of other Cadherin gene transcripts with the CDH7 precursor protein

Discussion and Extension – Orthologs of CDH7

For the following sequence comparisons I am using the BLOSUM62 scoring matrix, 11 gap existence cost and a 1 gap extension cost. All of the following proteins are compared with the Homo Sapiens Cadherin 7 Preproprotein 1 Protein.

Scientific Name	Name	Accession ID	Length (aa)	Identity %	Coverage %
Mus Musculus	Mouse	NP_001303672.1	785	97.96	100
Macaca Mulatta (Precursor)	Macaque	NP_001180923.2	785	97.71	100
Rattus Norvegicus	Norway Rat	NP_001012755.1	785	97.45	100
Gallus Gallus (Precursor)	Junglefowl	NP_989518.3	785	93.25	100
Bos Taurus	Cattle	NP_001179833.1	785	98.09	100

Table 7 Comparison of CDH7 orthologs in other species

Despite being a diverse array of species, there is still a large similarity between each of the different species' versions of the same protein. Notably, the precursor in humans is only 630 amino acids in length compared to the full 785-length activated protein yet the precursor in both the Junglefowl and the Macaque is already the same length. The comparison of regions on the protein between the human preproprotein, precursor and the Macaque precursor shows that the Macaque precursor is much closer to the final protein, due to already having the Cadherin cytoplasmic region. The final piece missing in the precursor is the transmembrane region and subsequent activation, meaning it is much more ready to go than the human precursor and is just missing the part that allows it to take its place in the cell.

References

[1] Maître JL, Heisenberg CP. Three functions of cadherins in cell adhesion. Curr Biol. 2013 Jul 22;23(14):R626-33. doi: 10.1016/j.cub.2013.06.019. PMID: 23885883; PMCID: PMC3722483.

[2] https://www.sciencedirect.com/topics/agricultural-and-biological-sciences/cadherin

[3] Gheldof, A., & Berx, G. (2013). Cadherins and Epithelial-to-Mesenchymal Transition. *Progress in Molecular Biology and Translational Science*, *116*, 317-336. https://doi.org/10.1016/B978-0-12-394311-8.00014-5

[4] Gumbiner, B. M. (2005). Regulation of cadherin-mediated adhesion in morphogenesis. *Nature Reviews Molecular Cell Biology*, *6*(8), 622-634. https://doi.org/10.1038/nrm1699

[5] https://www.genecards.org/cgi-bin/carddisp.pl?gene=CDH7

[6] Nagafuchi, A., & Takeichi, M. (1988). The cell binding function of E-cadherin is regulated by the cytoplasmic domain. *The EMBO Journal*, *7*(12), 3679-3684. https://doi.org/10.1002/j.1460-2075.1988.tb03249.x

[7] Homo Sapien Preproprotein CDH7 - https://www.ncbi.nlm.nih.gov/protein/NP_004352.2

[8] Homo Sapien Precursor CDH7 - https://www.ncbi.nlm.nih.gov/protein/NP_001304143.1

[9] https://www.ncbi.nlm.nih.gov/genome/gdv/browser/genome/?id=GCF_000001405.40

[10] Colman DR, Filbin MT. The Cadherin Family. In: Siegel GJ, Agranoff BW, Albers RW, et al., editors. Basic Neurochemistry: Molecular, Cellular and Medical Aspects. 6th edition. Philadelphia: Lippincott-Raven; 1999. Available from: https://www.ncbi.nlm.nih.gov/books/NBK27935/

[11] https://www.ncbi.nlm.nih.gov/gtr/genes/1005/