BioInformatics coursework 1 - Sequence Retreival & Analysis of Human Cadherin-7

1. Introduction

The cadherin-7 gene (CDH7) is a member of the cadherin superfamily of calcium-dependent cell-cell adhesion molecules. This calcium dependent cell-cell adhesion molecule is comprised of five extracellular cadherin repeats, a transmembrane region and a highly conserved cytoplasmic tail [1]. The CDH7 gene encodes a type II classical cadherin, meaning that it lacks a histidine-alanine-valine (HAV) cell adhesion recognition sequence specific to type I cadherins [2]. Type II cadherins mediate cell-cell binding in a homophilic manner, contributing to the sorting of heterogeneous cell types.

Cadherin-7 is often associated with neural tissue development, in particular the formation and maintenance of specific neural circuits. One such example is constituted by its use in mossy fiber pontine nucleus (PN) neurons in the human cerebellum [3]. Given its heavy involvement in brain activity, mutations in this gene can have quite drastic consequences: in human patients, it is reported that such mutations may be associated with bipolar disorder.

In this report, we seek to explore how alternate splicing of the cadherin-7 gene can produce different transcripts (mRNAs) and protein isoforms.

2. Data & Methods

2.1 Data sources used

The main data source utilised in this report is the National Center for Biotechnology Information (NCBI). The NCBI is a database of biomedical and genetic information [4]. Within NCBI, three main databases were used to retrieve all the necessary information.

The first of these databases is the gene database. This database contains records for known or predicted genes, which have been curated over several revision cycles. This database contains a unique GeneID for each gene record in the database [5]; in our case, the record for the cadherin-7 gene in Homo sapiens was 1005.

The second database used is the nucleotide database, which contains a collection of sequences such as genome, gene, and transcript sequence data [6]. We used this database to retrieve the nucleotide sequence for each cadherin-7 gene transcript we obtained from the gene database.

The third database used is the SWISS-PROT database. SWISS-PROT is a curated protein sequence database which aims to provide high level of integration with other databases [7]. This database is integrated with the NCBI Basic Local Alignment Search Tool (BLAST), which can be used to find regions of similarity between biological sequences [8]. In our case, we used this database as part of a BLAST query (blastp to be more precise) in order to find all human proteins that have a Cadherin domain.

2.2 Methods used

For all three parts outlined below, python's Jupyter notebooks interface was the primary method used in this research for data analysis. However, below we are going to explore the individual techniques used for each of the individual parts.

Part 1

For this part, we followed the following procedure in python Jupyter notebooks:

1. Using the GeneID previously discussed (1005), we used the Entrez module of the biopython library in order to fetch all transcripts for that GeneID.
2. After all transcripts were obtained, the total number retrieved, along with their individual accession IDs, were outputted.
3. For each transcript, we retrieved the nucleotide sequence (again using the Entrez module).
4. Once we got the nucleotide sequences, we simply computed their lengths and the percentage composition of A, C, G and T for each sequence.
5. After computing the percentage composition, we translated each sequence to a protein sequence, computed the length, and identified the most common amino acid in each of the newly acquired protein sequences.

All the results obtained using these steps were cross-referenced with the NCBI website to ensure correctness and consistency.

Part 2

In this section, we used two approaches to the sequence alignment tasks for both nucleotide and protein sequences (and again cross-referenced the results for consistency). The first approach was using the biopython pairwise2 module, where we performed two alignments, both using a local alignment algorithm [9]; however, we used two different scoring matrices, namely PAM30 [10] and BLOSUM62 [11], to check if any differences arose. The second approach was using BLAST on the NCBI website. Since the observed results were identical, we decided to display the results of the BLAST search in the Results and analysis section, as they are more easily interpretable.

For identifying the missing exons, we compared the sequences using the NCBI Genome Data Viewer [12].

Part 3

For this section, we decided to once again use two distinct approaches and compare the results for consistency. The first approach relied on using the biopython Entrez module to retrieve all human proteins with a Cadherin domain. Once these were obtained, we performed both local and global alignment [13] to determine the top 3 closest proteins to cadherin-7. The second approach was using performing a BLAST search and ordering the outputs based on the computed e-value [8] to retrieve the top 3 closest proteins. Here, we will present all the results obtained (in the Results and analysis section) since we observed some differences, and we will also discuss the possible reasons for these differences.

3. Results and analysis

Part One - Retrieving and Analyzing mRNA and Protein Sequences of Human Cadherin-7

Task 1

Gene ID	Gene symbol	Accession ID	Nucleotide length
1005	CDH7	NM_033646.4	12126
1005	CDH7	NM_004361.5	12126
1005	CDH7	NM_001317214.3	3407
1005	CDH7	NM_001362438.2	12938

Task 2

Accession ID	Percentage composition for each nucleotide			
	Percentage composition of A	Percentage composition of C	Percentage composition of G	Percentage composition of T
NM_033646.4	33.24	17.33	18.51	30.92
NM_004361.5	33.03	17.52	18.68	30.77
NM_001317214.3	28.53	22.07	22.04	27.36
NM_001362438.2	32.13	18.00	19.49	30.38

Task 3

Nucleotide Accession ID	Protein Accession ID	Length of resulting protein (amino acids)	Most frequent amino acid(s)
NM_033646.4	NP_387450.1	785	Serine (S)
NM_004361.5	NP_004352.2	785	Serine (S)
NM_001317214.3	NP_001304143.1	630	Serine (S)
NM_001362438.2	NP_001349367.1	785	Serine (S)

Analysis of tasks 1-3 results

From the results presented above, we can see that the transcript with accession ID NM_001362438.2 has the longest nucleotide sequence (12938 nucleotides) as well as the longest protein sequence (785 amino acids). We also note that the transcript with accession ID NM_001317214.3 has the shortest nucleotide sequence (3407 nucleotides) and the shortest protein sequence (630 amino acids). These results will be used in ***Part 2***, where we are going to perform a pairwise sequence alignment in order to compare these two sequences.

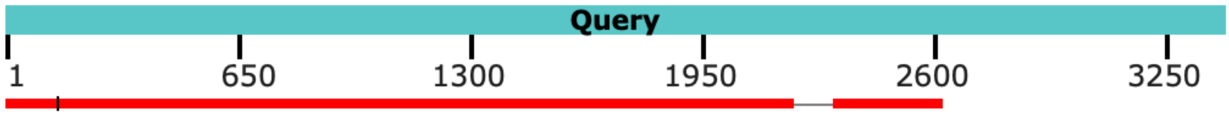
Extension

In this section we compared the human Cadherin-7 to orthologs in other species. Orthologs are genes in different species that evolved from a common ancestral gene by speciation. In order to determine which orthologs had the highest similarity to CDH7, we used two approaches and cross-referenced them to ensure they provide the same results. The first approach was using the OrthoDB database, which provides a database dedicated to the classification of orthologs among various species [14]. One of the main advantages of OrthoDB is that it bases its classification of orthologs on both sequence similarity and phylogenetic relationships, unlike for instance BLAST which only relies on sequence similarity. Nonetheless, we also used BLAST for consistency, and we performed a megablast search which returned similar results as the OrthoDB for the top three orthologs. We show the top five results of the BLAST search (which contain three species), as well as the search parameters used, in the ***Appendix Part One Extension*** section of this report.

Part Two – Comparing mRNA transcripts & protein isoforms of Human Cadherin-7 and interpreting the consequences of alternate transcription

Task 1

Below we show the graphic summary of a BLASTN pairwise sequence alignment between the shortest (NM_001317214.3) and longest (NM_001362438.2) human Cadherin-7 alternate transcript sequences.

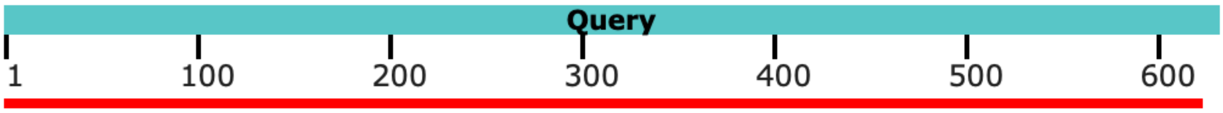


Analysis of task 1 results

We report an e-value of 0.0 and a percentage identity of 100%. We also note that, for the red regions, the alignment score was above 200, indicating significant similarity. However, this is to be expected since both represent nucleotide sequences for the same gene (CDH7). Nonetheless, we do note some gaps in the alignment, such as after 2600, as well as regions with lower alignment scores, such as the small black section spotted between 1950 and 2600, which maps to an alignment score below 40.

Task 2

Below we show the graphic summary of a BLASTP pairwise sequence alignment between the shortest (NP_001304143.1) and longest (NP_001349367.1) human Cadherin-7 protein isoform sequences.



Analysis of task 2 results

We report an e-value of 0.0 and a percentage identity of 100%. We also note that, for the red regions, the alignment score was above 200, indicating significant similarity. However, this is to be expected since both represent protein isoforms for the same gene (CDH7). However, it is worth noting that, whilst this graph doesn't show it, these are in fact two different isoforms (one is a preprotein and the other is a precursor) and they have different lengths (the precursor has length 630 amino acids, and the preprotein has length 785 amino acids). Nonetheless, we can see that for all 630 amino acids in the precursor isoform, the two protein sequences align really closely.

Task 3

Using the NCBI Genome Data Viewer, we found the following number of exons for the four transcripts available:

Accession ID	Number of exons
NM_033646.4	12
NM_004361.5	12
NM_001317214.3	11
NM_001362438.2	12

We also cross-referenced the exon lengths reported by the Genome Data Viewer with those available on e!Ensembl [15] for consistency, and we report the following exon lengths for each of the twelve exons:

Exon number	Exon length (nucleotides)
Exon 1	139
Exon 2	406
Exon 3	295
Exon 4	120
Exon 5	168
Exon 6	188
Exon 7	254
Exon 8	137
Exon 9	122
Exon 10	118
Exon 11	252
Exon 12	9937

Task 4

By comparing the two transcripts using the NCBI Genome Data Viewer, we found that transcript NM_001317214.3 (representing the shortest sequence) was missing ***Exon 12*** when compared to transcript NM_001362438.2 (representing the longest sequence).

Task 5

The missing Exon 12 appears to have caused a difference in the protein's structure: transcript NM_001362438.2 (representing the longest sequence) forms a preprotein isoform, whereas transcript NM_001317214.3 (representing the shortest sequence) forms a precursor isoform. The difference between the two isoforms is that a signal peptide is present in the preprotein isoform but absent in the precursor isoform. This signal peptide in a preprotein is critical for ensuring that the protein is directed to the appropriate cellular location, which in turn is crucial for a protein's correct functionality. To ensure the protein will still be directed to the appropriate cellular location, the precursor isoform requires additional modifications (such as peptide sequence removal, phosphorylation, glycosylation, etc.) before its target location can be identified and its functional form reached; by contrast, preprotein isoforms are directed to their target location by the signal peptide, and then simply transformed into their functional form by removing the signal peptide.

Extension

In this subsection, we are going to describe the protein domains of Cadherin-7. As with most classical cadherins, Cadherin-7 contains the following three protein domains: **extracellular cadherin domains (ECs)**, **transmembrane domain**, and **cytoplasmic domain**.

Extracellular cadherin domains (ECs) are involved in mediating cell-cell adhesion through homophilic interactions (cadherins binding to the same type of cadherin molecule on another cell). Cadherins are highly dependent on calcium ions to function as expected, and it is the presence of these ions that stabilizes the conformation of these domains and is essential for their adhesive function.

The transmembrane domain's main functionality is to anchor the cadherin molecule in the cell membrane.

The cytoplasmic domain mediates the interactions of cadherins with another protein family called **catenins**. These interactions are key for linking cadherins to the actin cytoskeleton of the cell, thus providing structural support to the adhesive junctions.

Part Three - Do other similar human Cadherin genes use alternate splicing in a similar way to Cadherin-7?

Task 1

For this task, we performed a BLASTP search to obtain all human proteins with a Cadherin domain. We also used the biopython Entrez module; however, since the BLAST search returned a more comprehensive set of results, we have chosen to include those in this report. For the BLAST search, we used the parameters outlined in the table below:

Search type	Sequence type	Organism	Database
BLASTP	Protein	Homo sapiens	UniProtKB/Swiss-prot (swiss-prot)

We display the first ten results of this BLAST search in the table below, and we will include the full table in the ***Appendix Part Three Task 1*** section:

Accession ID	Protein name
Q9ULB5.2	Cadherin-7
Q9HBT6.2	Cadherin-20
P55285.1	Cadherin-6
P55286.2	Cadherin-18
Q9Y6N8.2	Cadherin-10
P55289.2	Cadherin-12
P55286.2	Cadherin-8
Q9ULB4.2	Cadherin-9
P55287.2	Cadherin-11
Q9UJ99.2	Cadherin-22

Task 2

Using the data from the previous task, we then performed a series of pairwise sequence alignments in order to retrieve the three closest proteins to Cadherin-7. We once again used both the Entrez module as well as BLAST; however, we used two approaches as part of the Entrez module: we used both global and local alignment for the pairwise sequence alignments. Below we report the top 3 results of the Entrez with global alignment approach:

Rank	Protein name	Accession ID
1	Protocadherin Fat 4	Q6V0I7.2
2	Cadherin-6	P55285.1
3	Cadherin-12	P55289.2

The results for the Entrez with local alignment approach are displayed in the table below:

Rank	Protein name	Accession ID
1	Cadherin-6	P55285.1
2	Cadherin-12	P55289.2
3	Cadherin-11	P55287.2

Lastly, we show the results of the BLAST search (using the same search parameters as in task 1) in the table below (note that we excluded the result entry for the Cadherin-7 gene itself which had 100% percentage identity):

Percentage Identity	E-value	Protein name	Accession ID
63.19	0.0	Cadherin-20	Q9HBT6.2
62.89	0.0	Cadherin-6	P55285.1
62.47	0.0	Cadherin-18	P55286.2

Analysis of the results

As can be seen, the three methods used to obtain the three closest proteins to Cadherin-7 return overall different results (despite some values matching). A good question to ask is: how do we know which results are correct (or is there such a thing as a correct answer)?

One thing we can conclude with certainty is that Cadherin-6 is one of the three closest proteins, as it is returned by all three methods. One might be inclined to consider Cadherin-12 as another close protein, since the only method not returning that protein in the top 3 was BLAST. However, BLAST provides several additional improvements to the simple pairwise sequence alignment algorithms used for local and global alignment as part of the pairwise2 module.

Another thing worth noting is that when comparing sequences that are part of the same superfamily, local alignment algorithms may give more accurate information regarding the similarity between sequences as opposed to global alignment algorithms. This is because global alignment aims to maximize the overall alignment score for the entire sequence, whereas local alignment focuses more on specific subsections of the sequences. Given this, one could argue that Protocadherin Fat 4 may not in fact be one of the three closest proteins to Cadherin-7.

Overall, we decided to use all proteins returned for future analysis tasks, as these further insights may help guide us to more concrete conclusions.

Task 3

Here we investigated if any of the proteins listed in the ***Task 2*** section formed similar protein isoforms to Cadherin-7. We used the NCBI Gene database to look for all transcripts for each individual protein, and check if they report either the “1 preprotein” or “2 precursor” isoforms (which are reported by CDH7). Below is a table that summarizes our findings:

Protein name	Number of “1 preprotein” isoforms	Number of “2 precursor” isoforms	Accession ID(s)
Cadherin-20	0	0	N/A
Cadherin-6	1	0	NP_004923.1
Cadherin-18	5	2	preprotein: NP_001278885.1, NP_004925.1, NP_001336485.1, NP_001336487.1, NP_001336488.1 precursor: NP_001161139.1, NP_001336490.1
Cadherin-12	7	1	preprotein: NP_004052.2, NP_001304156.1, NP_001351033.1, NP_001351034.1, NP_001351035.1, NP_001351036.1, NP_001351037.1 precursor: NP_001304157.1
Cadherin-11	1	1	preprotein: NP_001788.2 precursor: NP_001295321.1
Protocadherin Fat 4	0	0	N/A

Extension

For this subsection, we are going to investigate paralogues of Cadherin-7 in three organisms: humans, mice, and rats. To do this, we performed three BLAST searches, one for each organism, and counted the number of unique genes that are similar to Cadherin-7 (excluding Cadherin-7 entries). We summarize the findings of these searches by providing the search parameters used, as well as the top 3 paralogues for each species (confirmed by cross-referencing with other databases and current literature) in the ***Appendix Part Three Extension*** section of this report.

References

1. Yeh-Shiu Chu, Olivier Eder, William A Thomas, Inbal Simcha, Frederic Pincet, Avri Ben-Ze'ev, Eric Perez, Jean Paul Thiery, and Sylvie Dufour. Prototypical type I E-cadherin and type II cadherin-7 mediate very distinct adhesiveness through their extracellular domains. Journal of Biological Chemistry, 281(5):2901–2910, 2006.
2. Linn Fagerberg, Bj¨orn M Hallstro¨m, Per Oksvold, Caroline Kampf, Dijana Djureinovic, Jacob Odeberg, Masato Habuka, Simin Tahmasebpoor, Angelika Daniels- son, Karolina Edlund, Anna Asplund, Evelina Sj¨ostedt, Emma Lundberg, Cristina Al-Khalili Szigyarto, Marie Skogs, Jenny Ottosson Takanen, Holger Berling, Hanna Tegel, Jan Mulder, Peter Nilsson, Jochen M Schwenk, Cecilia Lindskog, Frida Daniels- son, Adil Mardinoglu, Asa Sivertsson, Kalle von Feilitzen, Mattias Forsberg, Martin Zwahlen, Ingmarie Olsson, Sanjay Navani, Mikael Huss, Jens Nielsen, Fredrik Ponten, and Mathias Uhĺ en. Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. Mol. Cell. Proteomics, 13(2):397–406, February 2014.
3. Ken-ichiro Kuwako, Yoshinori Nishimoto, Satoshi Kawase, Hirotaka James Okano, and Hideyuki Okano. Cadherin-7 regulates mossy fiber connectivity in the cerebellum. Cell Reports, 9(1):311–323, 2014.
4. National Center for Biotechnology Information. https://www.ncbi.nlm.nih.gov/, Ac- cessed October 24, 2023.
5. Gene Database. National Center for Biotechnology Information. https://www.ncbi.nlm.nih.gov/gene/, Accessed October 24, 2023.
6. Nucleotide Database. National Center for Biotechnology Information. https://www.ncbi.nlm.nih.gov/nucleotide/, Accessed October 24, 2023.
7. Amos Bairoch and Rolf Apweiler. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic acids research, 28(1):45–48, 2000.
8. Mark Johnson, Irena Zaretskaya, Yan Raytselis, Yuri Merezhuk, Scott McGinnis, and Thomas L Madden. NCBI BLAST: a better web interface. Nucleic acids research, 36(suppl 2):W5–W9, 2008.
9. William R Pearson. Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms. Genomics, 11(3):635– 650, 1991.
10. Stephen F Altschul. A protein alignment scoring system sensitive at all evolutionary distances. Journal of molecular evolution, 36:290–300, 1993.
11. David W Mount. Using BLOSUM in sequence alignments. Cold Spring Harbor Pro- tocols, 2008(6):pdb–top39, 2008.
12. Genome Data Viewer. National Center for Biotechnology Information. https://www.ncbi.nlm.nih.gov/genome/gdv/browser/gene/?id=1005, Accessed October 25, 2023.
13. Saul B Needleman and Christian D Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of molecular biology, 48(3):443–453, 1970.

14. Evgeny M Zdobnov, Dmitry Kuznetsov, Fredrik Tegenfeldt, Mos`e Manni, Matthew Berkeley, and Evgenia V Kriventseva. OrthoDB in 2020: evolutionary and functional annotations of orthologs. Nucleic Acids Research, 49(D1):D389–D393, 11 2020.
15. Fergal J Martin, M Ridwan Amode, Alisha Aneja, Olanrewaju Austine-Orimoloye, Andrey G Azov, If Barnes, Arne Becker, Ruth Bennett, Andrew Berry, Jyothish Bhai, Simarpreet Kaur Bhurji, Alexandra Bignell, Sanjay Boddu, Paulo R Branco Lins, Lucy Brooks, Shashank Budhanuru Ramaraju, Mehrnaz Charkhchi, Alexander Cockburn, Luca Da Rin Fiorretto, Claire Davidson, Kamalkumar Dodiya, Sarah Donaldson, Bilal El Houdaigui, Tamara El Naboulsi, Reham Fatima, Carlos Garcia Giron, Thiago Genez, Gurpreet S Ghattaoraya, Jose Gonzalez Martinez, Cristi Guijarro, Matthew Hardy, Zoe Hollis, Thibaut Hourlier, Toby Hunt, Mike Kay, Vinay Kaykala, Tuan Le, Diana Lemos, Diego Marques-Coelho, Jos´e Carlos Marug´an, Gabriela Alejandra Merino, Louisse Paola Mirabueno, Aleena Mushtaq, Syed Nakib Hossain, Denye N Ogeh, Manoj Pandian Sakthivel, Anne Parker, Malcolm Perry, Ivana Piliˇzota, Irina Prosovetskaia, Jos´e G P´erez-Silva, Ahamed Imran Abdul Salam, Nuno Saraiva-Agostinho, Helen Schuilenburg, Dan Sheppard, Swati Sinha, Botond Sipos, William Stark, Emily Steed, Ranjit Sukumaran, Dulika Sumathipala, Marie-Marthe Suner, Likhitha Surapaneni, Ky¨osti Sutinen, Michal Szpak, Francesca Floriana Tricomi, David Urbina-G´omez, Andres Veidenberg, Thomas A Walsh, Brandon Walts, Elizabeth Wass, Natalie Willhoft, Jamie Allen, Jorge Alvarez-Jarreta, Marc Chakiachvili, Bethany Flint, Stefano Giorgetti, Leanne Haggerty, Garth R Ilsley, Jane E Loveland, Benjamin Moore, Jonathan M Mudge, John Tate, David Thybert, Stephen J Trevanion, Andrea Winterbottom, Adam Frankish, Sarah E Hunt, Magali Ruffier, Fiona Cunningham, Sarah Dyer, Robert D Finn, Kevin L Howe, Peter W Harrison, Andrew D Yates, and Paul Flicek. Ensembl 2023. Nucleic Acids Research, 51(D1):D933–D941, November 2022.

Appendix

Part One Extension

Search parameters

Search type	Sequence type	Organism	Database
MEGABLAST	Nucleotide	(exclude) Homo sapiens (include) mammals	Non-redundant protein sequences (nr)

Results

Accession ID	Scientific species name	Percentage Identity
XM_034943952.2	Pan paniscus	99.40
XM_001149161.5	Pan troglodytes	99.33
XM_009434118.4	Pan troglodytes	99.29
XM_031003575.2	Gorilla gorilla gorilla	98.64
XM_004059526.4	Gorilla gorilla gorilla	98.63

Part Three Task 1

Accession ID	Protein name
Q9ULB5.2	Cadherin-7
Q9HBT6.2	Cadherin-20
P55285.1	Cadherin-6
P55286.2	Cadherin-18
Q9Y6N8.2	Cadherin-10
P55289.2	Cadherin-12
P55286.2	Cadherin-8
Q9ULB4.2	Cadherin-9
P55287.2	Cadherin-11
Q9UJ99.2	Cadherin-22
Q9H159.1	Cadherin-19
Q86UP0.1	Cadherin-24
P33151.5	Cadherin-5
P55283.2	Cadherin-4
P19022.4	Cadherin-2
P55291.1	Cadherin-15
P12830.3	Cadherin-1
P22223.2	Cadherin-3
P55290.1	Cadherin-13
Q08554.2	Desmocollin-1
Q02487.1	Desmocollin-2
Q14574.3	Desmocollin-3
Q14517.2	Protocadherin Fat 1
Q6V0I7.2	Protocadherin Fat 4

Q8IXH8.4	Cadherin-like protein 26
Q9NYQ7.2	Cadherin EGF LAG seven-pass G-type receptor 3
Q14126.2	Desmoglein-2
Q9NYQ8.2	Protocadherin Fat 2
Q8TDW7.3	Protocadherin Fat 3
Q86SJ6.1	Desmoglein-4
Q12864.3	Cadherin-17
Q6V1P9.2	Protocadherin-23
Q9HCU4.1	Cadherin EGF LAG seven-pass G-type receptor 2
Q9H251.2	Cadherin-23
Q9NYQ6.1	Cadherin EGF LAG seven-pass G-type receptor 1
Q9Y5G8.1	Protocadherin gamma-A5
Q9Y5G0.1	Protocadherin gamma-B5
Q9Y5H4.1	Protocadherin gamma-A1
Q9Y5G7.1	Protocadherin gamma-A6
Q9Y5H1.1	Protocadherin gamma-A2
Q96JQ0.1	Protocadherin-16
Q9Y5G3.1	Protocadherin gamma-B1
Q9Y5F9.1	Protocadherin gamma-B6
P32926.2	Desmoglein-3
Q9Y5H3.1	Protocadherin gamma-A10
Q9Y5E2.1	Protocadherin beta-7
Q9Y5G4.1	Protocadherin gamma-A9
Q9HC56.2	Protocadherin-9
O60330.1	Protocadherin gamma-A12
Q8TAB3.3	Protocadherin-19
Q9Y5H0.2	Protocadherin gamma-A3
Q9Y5G1.4	Protocadherin gamma-B3
Q9Y5H2.1	Protocadherin gamma-A11
Q9Y5F8.1	Protocadherin gamma-B7
Q9UN71.1	Protocadherin gamma-B4
O75309.1	Cadherin-16
Q9Y5E8.1	Protocadherin beta-15
Q9Y5G2.1	Protocadherin gamma-B2
Q9P2E7.2	Protocadherin-10
Q9Y5G6.1	Protocadherin gamma-A8
Q9Y5F0.1	Protocadherin beta-13
Q9Y5E7.1	Protocadherin beta-2
Q9Y5G5.1	Protocadherin gamma-A8
Q9HCL0.3	Protocadherin-18
Q9UN70.1	Protocadherin gamma-C3
Q08174.2	Protocadherin-1
Q9Y5I4.1	Protocadherin alpha-C2
Q9UN66.4	Protocadherin beta-8
Q9Y5F6.1	Protocadherin gamma-C5
Q9NRJ7.4	Protocadherin beta-16
Q9BZA7.1	Protocadherin-11 X-linked
Q02413.2	Desmoglein-1
Q9BZA8.1	Protocadherin-11 Y-linked
Q9Y5E1.3	Protocadherin beta-9

Q9Y5F1.1	Protocadherin beta-12
O14917.2	Protocadherin-17
Q96QU1.2	Protocadherin-15
Q9UN74.1	Protocadherin alpha-4
Q9Y5H8.1	Protocadherin alpha-3
Q9UN75.1	Protocadherin alpha-12
O60245.2	Protocadherin-7
Q9Y5H5.1	Protocadherin alpha-9
Q9Y5F3.2	Protocadherin beta-1
Q9Y5G9.2	Protocadherin gamma-A4
Q9Y5I3.1	Protocadherin alpha-1
Q9Y5E4.2	Protocadherin beta-5
Q9Y5F7.1	Protocadherin gamma-C4
Q9UN72.1	Protocadherin alpha-7
Q8N6Y1.2	Protocadherin-20
Q9Y5I2.1	Protocadherin alpha-10
Q9Y5H7.1	Protocadherin alpha-5
Q9Y5E9.1	Protocadherin beta-14
Q96TA0.2	Putative protocadherin beta-18
Q9Y5I0.1	Protocadherin alpha-13
Q9Y5E3.2	Protocadherin beta-6
Q9UN73.1	Protocadherin alpha-6
Q9Y5F2.1	Protocadherin beta-11
Q9Y5H6.1	Protocadherin alpha-8
Q9Y5E5.1	Protocadherin beta-4
Q9UN67.1	Protocadherin beta-10

Part Three Extension

Search 1 parameters

Search type	Sequence type	Organism	Database
BLASTP	Protein	Homo sapiens	Reference Sequences (refseq_protein)

Note: once the results were returned, we applied a filter for the percentage identity in order to remove the Cadherin-7 entries; we limited the percentage identity to the range 30 - 70.

Search 2 parameters

Search type	Sequence type	Organism	Database
BLASTP	Protein	Mus musculus	Reference Sequences (refseq_protein)

Note: once the results were returned, we applied a filter for the percentage identity in order to remove the Cadherin-7 entries; we limited the percentage identity to the range 30 - 70.

Search 3 parameters

Search type	Sequence type	Organism	Database
BLASTP	Protein	Rattus rattus	Reference Sequences (refseq_protein)

Note: once the results were returned, we applied a filter for the percentage identity in order to remove the Cadherin-7 entries; we limited the percentage identity to the range 30 - 70.

Results for search 1 (humans)

Accession ID	Protein name
NP_001278886.1	Cadherin-18
NP_114097.2	Cadherin-20
NP_066976.1	Cadherin-19

Results for search 2 (mice)

Accession ID	Protein name
NP_035930.1	Cadherin-20
NP_001074768.1	Cadherin-18
NP_001074855.1	Cadherin-19

Results for search 3 (rats)

Accession ID	Protein name
XP_032771636.1	Cadherin-20
XP_032771502.1	Cadherin-19
XP_032754637.1	Cadherin-18