coursework1_instructions

October 9, 2023

Introduction

Sequence Retreival & Analysis of Human Cadherin-7

In this coursework we are going to put into practice what you've been learning about biological sequences and the ability to search and retrive them from the NCBI database systems. We are going to look at a very interesting protein called Cadherin-7 which is a member of a large family of proteins involved in cell-cell adhesion and signalling. We will look at the gene and protein sequences to see how alternate splicing of the gene can produce different transcripts (mRNAs) and protein isoforms.

In **Part One** we focus on sequence retreival and understanding the gene and its structure. In **Part Two** we use pairwise sequence alignment (Needleman-Wunsch (global) or Smith-Waterman (local) alignment or BLAST at NCBI) for both transcripts and protein isoforms to find which parts differ and where on the full length protein sequence these differences are located. By viewing where these are on the 3D protein structure and referring to reference material describing Cadherin structure and function we predict what the changes are likely to do. In **Part Three** we query the SwissProt subset of the NCBI protein database to look for the three most similar human Cadherin proteins to Cadherin-7 and look to see whether they also use alternate transcription to produce differnt protein isoforms.

All that you need to successfully complete this project has been covered in the course using web-tools and/or programming approaches. It is fine for you to use either approach or a hybrid of both, but however you aim to undertake the work everyone needs to write up a report of the research. Throughout the instructions, below, we guide you through the project and provide tips, especially where there is an opportunity to do something a little more complex if you choose. We hope you enjoy the project and exploring in a real-world example how you can bring together different sources of information to anayse and gain insight into a research problem.

Useful Sources of Information

Background about Cadherins

- Wikipedia entry for Cadherins - https://en.wikipedia.org/wiki/Cadherin
- Shapiro L, Weis WI. Structure and biochemistry of cadherins and catenins. Cold Spring Harb Perspect Biol. 2009 Sep;1(3):a003053. https://doi.org/10.1101/cshperspect.a003053. PubMed ID: 20066110
- Prosite entry for the Cadherin doman containing useful information and references - https://prosite.expasy.org/PDOC00205
- Cadherin-7 gene entry at NCBI - https://www.ncbi.nlm.nih.gov/datasets/gene/id/1005/products/
- Cadherin-7 protein entry at UniProt - https://www.uniprot.org/uniprotkb/Q9ULB5/entry

Searching

- Search field descriptions for sequence databases - https://www.ncbi.nlm.nih.gov/books/NBK49540/

Hints!

- *Xing Y, Xu Q, Lee C. FEBS Lett. 2003 Dec 18;555(3):572-8. https://doi.org/10.1016/s0014-5793(03)01354-1. PubMed ID: 14675776*
- *Kawano R, Matsuo N, Tanaka H, Nasu M, Yoshioka H, Shirabe K. Journal of Biological Chemistry. 2002 277:49,47679-47685. https://doi.org/10.1074/jbc.M205328200. PubMed ID: 12364338*
- https://en.wikipedia.org/wiki/Single-pass_membrane_protein
- If you perform a protein BLAST search at the NCBI website you can click on an alignment result row it will take you to the details of the alignment and you will oftern see an "AlphaFold Structure" link on the right hand side where you can view your query aligned to the 3D structure of the protein target that it has hit in the database

Research Instructions

Part One - *Retreiving and Analysing mRNA and Protein Sequences of Human Cadherin-7 (25 marks)*

- **Task 1** - Search NCBI for the human Cadherin-7 gene entry, how many alternate transcripts does it have, what are their official accession ids, and how long are they in nucleotides (**5 marks**)
- **Task 2** - Calculate the percentage composition of A, C, G, and T for each transcript and report the results in a table with the transcripts as rows and the percentages for each nucleotide as columns (**5 marks**)
- **Task 3** - Translate the transcript sequences into protein sequences, report the length of the resulting proteins and state the most frequent amino acid in each one, if there are joint frequent ones name them all (**5 marks**)
- up to **6 marks** for including one additional piece of analysis.
- up to **4 marks** for exceptionally well organised and executed approach.

Possible extensions here might be to:

- look at other human Cadherin genes to see whether they also use alternate splicing to create different isoforms
- compare human Cadherin-7 to orthologs in other species
- write a short summary of what Cadherins do
- look at the structure of the Cadherin protein family
- comment on the difference in structure between the different Cadherin-7 isoforms

Part Two - *Comparing mRNA transcripts & protein isoforms of Human Cadherin-7 and interpretting the consequences of alternate transcription* (25 marks)*

- **Task 1** - Perform pairwise sequence comparison between the longest and shortest human Cadherin-7 alternate transcript sequences. Report your alignment results and explain what you think they show (**3 marks**)

- **Task 2** - Perform pairwise sequence comparison between the longest and shortest human Cadherin-7 protein isoform sequences. Report your alignment results and explain what you think they show (**3 marks**)
- **Task 3** - How many exons does the human Cadherin-7 gene have and what are their lengths (**3 marks**)
- **Task 4** - Which exon(s) are missing from the shortest Cadherin-7 transcript compared to the longest one? (**2 marks**)
- **Task 5** - What difference do the missing parts of the protein from the shortest Cadherin-7 isoform compared to the longest make? (HINT you need to find out what the missing pieces do) (**4 marks**)
- up to **6 marks** for including one additional piece of analysis.
- up to **4 marks** for exceptionally well organised and executed approach.

Possible extensions here might be to:

- describe the the protein domains of Cadherin-7
- relate the changes in sequence between the shortest and longest isoforms to their protein structures, what's different?
- in addition to **describing** the difference the missing protein sequence makes also explain **why** it makes the difference (HINT - look up single pass transmembrane proteins).

Part Three - *Do other similar human Cadherin genes use alternate splicing in a similar way to Cadherin-7? (25 marks)*

- **Task 1** - Find all human proteins that have a Cadherin domain (**5 marks**)
  - (HINT - add the swissprot[filter] term to your search)
  - (HINT - NB that the Cadherin-7 official HGNC gene symbol is CDH7)
- **Task 2** - Retrive the sequences and find the three closest proteins to Cadherin-7 (**5 marks**) (HINT - perform a pairwise sequence comparison of all the sequences to Cadherin-7)
- **Task 3** - Do any of these 3 closest Cadherins have short protein isoforms similar to Cadherin-7? (**5 marks**)
- up to **6 marks** for including one additional piece of analysis.
- up to **4 marks** for exceptionally well organised and executed approach.

Possible extensions here might be to:

- look for orthologues of Cadherin-7 in a selection of divergent orgnisms and compare their protein sequences
- count the number of paralogues of Cadherin-7 in humans, mice, and rats
- look for reported mutations in the Cadherin-7 gene and see where they are located in the protein sequence and 3D-structure

Structure of the Report

Write a report of your project including figures, tables and results from your research in Parts One-Three above in the relevant sections. The report should follow the structure described below.

- Introduction
- Data & Methods
- Results
  - Part One - Retreiving and Analysing mRNA and Protein Sequences of Human Cadherin-7.
  - Part Two – Comparing mRNA transcripts & protein isoforms of Human Cadherin-7 and interpretting the consequences of alternate transcription.
  - Part Three –
- Discussion
- References
- Appendices (optional)

Your report should use an 11-point font for the main text and be no more than 8 pages in total (including everything except the references and optional appendices). You should not include code in the main text of the report, but can include them in the appendices if you choose. Please do not paste screenshots of results into the report, but instead generate appropriate tables and figures that are properly labelled and formatted with sufficiently detailed legends to understand their content and key interpretation. Make sure that you document sequences you are working with correctly by using their official accession ids, refrence the databases you use, detail the parameters of the alignments and any BLAST searches that you use. When referring to information you have found either online or in publications site them appropriately in the report text and include them in a consistently formatted bibliography at the end of the report. You may use whichever referencing style you prefer.

Marking Scheme

Your mark for this coursework assignment will contribute 50% to your mark for the Bioinformatics 1 course.

The work will be marked out of 100 with the following scheme:

Overall Report [25 marks]

- Data, Methods, Presentation of Results and their interpretation clearly described paying close attention to keeping track of versions and dates. Think of how able another researcher would be to look at your report and repeat what you have done - reproducibility [10 marks]
- Correct overall structure of the report [3 marks]
- References and any Appendices in good order [2 marks]
- Clear, well-labelled graphs throughout [10 marks]

Part One [25 marks]

- 15 marks for successfully completing the tasks.
- up to 6 marks for including one additional piece of analysis.
- up to 4 marks for exceptionally well organised and executed approach.

Part Two [25 marks]

- 15 marks for successfully completing the tasks.
- up to 6 marks for including one additional piece of analysis.
- up to 4 marks for exceptionally well organised and executed approach.

Part Three [25 marks]

- 10 marks for successfully completing the tasks.
- up to 6 marks for including one additional piece of analysis.
- up to 4 marks for exceptionally well organised and executed approach.