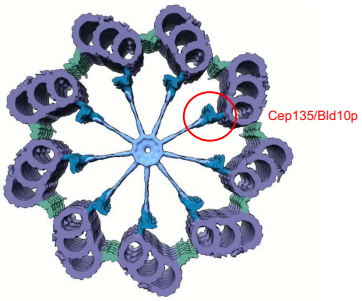
Report bachelor project 2021: Database for Cep135/Bld10p protein sequences

Author: Zied Mustapha (sciper: 282388)
Supervisor: Arpat Alaaddin Bulak

Introduction

The main purpose of this project is to create a database containing as much Cep135 protein sequences as possible and to make it organised and accessible. The Centrosomal protein (Cep135) acts as a scaffolding protein during early centriole biogenesis. The work is divided into two parts: Collecting the data, which took most of the time, and organising this data so that it can be used efficiently.



I- Collect the data

The ultimate goal here would be to find every Cep135 protein sequences available on the internet. It is obviously not feasible because of the decentralised amount of data. We restricted ourselves to one of the biggest bio-informatics database: Uniprot. To gather all the data we need, we first started with simple and naive approaches until some more complex solutions.

Search by name:

First approach was to simply take all protein sequences that are annotated “Cep135” or “Bld10p” in UniProt. The problem with this technique is that it will give us just a small fraction of the data we need. In fact, most of the cep135 sequences that we may find on UniProt are not annotated and referred as “Uncharacterised protein”. We also find a lot of redundancy in the result. The result of this search can be found in: result/search_by_labels/uniprot-cep135.fasta

Search by identity:

Another idea is to use one sequence that we know for sure is Cep135 and check its percentage of similarities with all sequences in UniProt. We can then use a threshold (let’s say 80%) and take all sequences above this threshold. Two problems emerged:

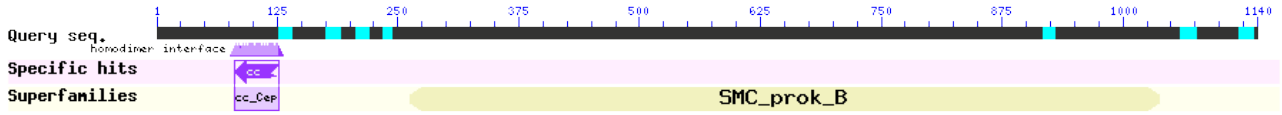
1. There is about 220 millions of sequences, it is computationally unfeasible.
2. What threshold should we use ? Some Cep135 sequences may be below the threshold.

To try to solve this, we used a known bio-informatics algorithm for protein sequences search: PSI-BLAST.

Search with PSI-BLAST:

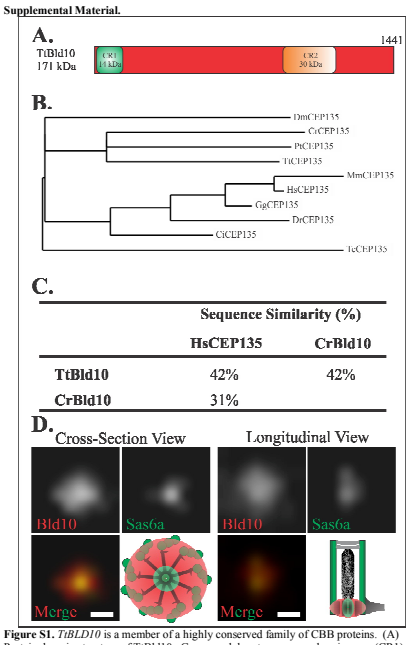
PSI-blast is an iterated profile search method that takes as an input a single protein sequence and compares it to a protein database, using the gapped BLAST (Basic Local Alignment Search Tool) algorithm. BLAST use a rolling window to break down a query sequence into words and word synonyms that form a search set. At least two words or synonyms in the search set must match a target sequence in the database, for that sequence to be reported in the results. This algorithm is then repeated a certain number of time using as initial sequences the result of the precedent execution, until it seems to converge. This algorithm use the concept of “E-value” which works as a threshold and give us an idea of the similarity of the sequences found. We tried this method using as initial sequence the Cep135 sequence of humans using this website: https://blast.ncbi.nlm.nih.gov (the link of the search is not available anymore) and with the default parameters. After 3 iterations the algorithm seems to converge and the result (over 2000 sequences) seem to be Cep135 sequences but some false positive can still be detected (result at result/search_with_PSI-BLAST/PSI-BLAST_full_lengths). To have a better accuracy, we tried to avoid the coiled coil structure. The coiled coil is a part of the protein structure that behave “randomly”, its sequence is variable and is not representative of the Cep135. The website that we used has a tool to find the exact position of coiled coil structure (figure 2). We then did the same PSI-BLAST search but by changing the parameter “position” to only take none coiled coil part (result at result/search_with_PSI-BLAST/PSI-BLAST_N_lengths). There is still one problem: All the sequences we found come from species that are very close with each other in the tree of taxonomy. The variety of species in the resulting sequences is small and we know that cep135 can be found in a wider variety of species. This problem comes from the fact that we begin the algorithm with one sequence: CEP135 from humans. So the taxonomy of the result is very close to the human species. We need to have a wider taxonomy tree. To try to solve this issue, we could

start the search algorithm with more initial sequences that come from very different species, and maybe then we will have a wider taxonomy tree.
Figure 2:



Search with a wider initial sequences:

The strategy here is to find sequences that are very different from each other and far way from each other in the taxonomy tree. Then we do a search with these sequences as initial sequences, then find more sequences like the ones described before and add them to the initial sequences, and repeat this process.
We found this research paper: (result/research_paper.pdf) where very different Cep135 sequences are presented with sometimes a similarity as small as 31% (see image right) (C.). This can be used as our initial input sequences. The problem with this is that if we put these very different sequences in any search algorithm (except HMM which we will talk about later) as input, it will find too much sequences that are not Cep135. The solution would be to find regions of these sequences that are conserved and do the search only on these regions. From the paper, it seems that there is two of such regions (A.). To find these regions, we did a PSI-Coffee alignment which aligns distantly related proteins using homology extension (http://tcoffee.crg.cat/apps/tcoffee/do:psicoffee) and the result can be found at result/search_wwis/alignment_initial_sequences.fasta_aln).
We then found these two regions: one in the beginning of the sequence (position 8 until 116 in human sequence) and one near the end of the sequences (position 722 until 996 in human sequence). We then did a HMM search on these regions of these Cep135 sequences. We didn't use PSI-BLAST this time because it only takes one initial sequence as first input.


Supplemental Material.

A.
TtBld10
171 kDa
1441

B.
DmCEP135
CrCEP135
PtCEP135
TtCEP135
MmCEP135
HsCEP135
GgCEP135
DrCEP135
CiCEP135
TcCEP135

C.

	Sequence Similarity (%)	
	HsCEP135	CrBld10
TtBld10	42%	42%
CrBld10	31%	

D.
Cross-Section View
Longitudinal View
Bld10
Sas6a
Bld10
Sas6a
Merge
Merge

Figure S1. *TtBLD10* is a member of a highly conserved family of CBB proteins. (A) Protein domain structure of TtBld10. Green oval denotes conserved region one (CR1)

MSA

The multiple sequence alignment result as produced by T-coffee.

```
T-COFFEE, Version_11.00 (Version_11.00)
Cedric_Notredame
SCORE=789
*
 BAD AVG GOOD
*
CP135_HUMAN          :  84
M9MSL4_DROME         :  80
A0A481MTH8_SCHM      :  63
XP_032228756.1       :  84
OAJ37363.1           :  81
A8ID55_CHLRE         :  82
TtBld10_TTHERM       :  83
tr|D2UXT3|D2UXT      :  80
tr|A0A1J4KE06|A      :  79
tr|F2U807|F2U80      :  82
tr|L1JRZ9|L1JRZ      :  81
tr|A0A2R5G0S2|A      :  82
tr|D2UXT3|D2UXT      :  80
cons                 :  78
```

Search by HMM:

We used this tool: https://www.ebi.ac.uk/Tools/hmmer/search/hmmsearch to do the hmm search online against UniProt. We did one search using only the first region and another using the second region. The results seemed fine (a lot of cep135 sequences found) but the taxonomy tree is still not wide enough (result at result/search_with_hmm/HMM_cep135). So we decided to get more initial sequences by searching the taxonomy tree for some outsider sequences that seems to be Cep135: 5 more sequences have been added to the initial input (result at result/search_with_hmm/initial_seq_aligned.fasta_aln). We selected these three sequences by looking at their coiled coil structure (to see if it match the typical coiled coil of centrosome) and simply its gene and protein name. Meanwhile the EBI tool to do the hmm search stopped working for Uniprot. We tried using another database (SwissProt) but didn't get expected results so we turned to another tool: https://toolkit.tuebingen.mpg.de/. This time we didn't bother searching only on the first or second region because of the way hmm work: it will automatically detects and favors these two regions that we found by alignment (result at result/search_with_hmm/final_result.fasta). The result is very large (around 9000 hits without redundancy) and the protein sequences seem to be Cep135 for most of them. More details on the result can be found at : https://toolkit.tuebingen.mpg.de/jobs/8803289_4 . We now have to take a look at the taxonomy

tree: result/search_with_hmm/taxnonomy.html. As showed in this html file, the taxonomy tree seems very wide and with more species that are genetically far from each other.

Limitations and further possible modifications:

In this final result, we can certainly find some false positive (i.e sequences that are not really Cep135). To solve this, we could create an algorithm or partially manually clean the data.
Example of algorithm would be a coiled coil detector that predict the coiled coil structure of every sequences and, based on this, classify the sequences as Cep135 or not. Other checks (how it has been aligned, reviews on this sequence etc…) have also to be used to ensure the prediction.

II- Make data organised and accessible

The purpose of this part was to add the data in the database of the laboratory, create some queries to facilitate accessibility and to create a web interface for this database. Unfortunately we did not have enough time to do everything.

Add data in the online database:

Now that we have the data, we have to store it and make it accessible. For this we will use these tools: BioSQL, BioPyton and MySQL. We populated the laboratory database with the data found (the final result) using Arpat Bulak's python algorithm. This algorithm go through every sequence, convert some metadata, find the species it belongs to, add more meta data using the sequence id by correlating it with NCBI database to find more informations. This algorithm allowed us to add every sequence in the online database of the laboratory and gave us extra informations on the data:

2	**Finding proper IDs (from NCBI, UniProt)**	
3	Total num of sequences	8257
4	IDs found	8205
5	IDs from entrez	8192
6	No alternative IDs found	52
7		
8	**Writing to DB**	
9	Total num of sequences	8257
10	Identified as a search results	8163
11	Loaded OK	8163
12	Loading failed	94
13		
14	**taxon_search**	
15	Found from cache	6784
16	Unique match	1461
17	Partial match	22
18	Automatically assigned	2
19	Manual assignment	5
20	Unidentified	5
21		
22	**filtering_recs**	
23	Total input	8257
24	Has 'species'	8257
25	Taxon search OK	8257
26	Unique species	1467
27	Unique species sequences	8257

Queries and web interface:

Because of time restriction, this part has not been done, but here is what could've been done:
Two types of queries: some to classify data using extra informations on sequences, and other to extract particular subset of the data. It will depends on what and how members of the laboratory want to use the data. For the web interface, web technologies like AJAX (Asynchronous Javascript And Xml) would make an intuitive and fast web interface to access data without even recharging the page.