Compare tools for protein domain identification

MMG1001 Assignment 1 – Group Heather

Corresponding Reading:

- Sonnhammer, Eddy and Durbin. 1997. Pfam: A comprehensive database of protein domain families based on seed alignments. Proteins 28:405–420.

Additional files needed

- unknown-seq.fa (download from Quercus)

Overview

There are many databases available to search through for information about protein sequences, structures, and domain contents. We will explore Pfam in depth and compare our results to others.

Databases

Pfam	https://pfam.xfam.org/
InterPro	http://www.ebi.ac.uk/interpro/
SMART	http://smart.embl-heidelberg.de/

Useful terminology

- **Clan:** A collection of related Pfam entries. The relationship may be defined by similarity of sequence, structure or profile-HMM
- **Family:** A collection of related protein regions
- **Domain:** A structural unit
- **Motif:** A short unit found outside globular domains
- **Architecture:** The collection of domains that are present on a protein

Learning more about a known protein: PRDM9

1. Copy the amino acid fasta sequence for human PRDM9 into the **Pfam** sequence search box (select the Search tab).
    a. What domains are significant hits?
    b. What is the top hit under insignificant hits (lowest E value)?
    c. How many times does this domain appear in the insignificant list? Why do you think that is?
    d. For one of the significant hits (your choice!) and the top insignificant hit, click on the family name to go to the main page for each domain. Search through the pages to fill out the following table:

Family		
Pfam accession number		
# architectures that contain the domain		
# sequences in full family alignment		

2. Now search the amino acid fasta sequence in **SMART**.
    a. Which domains/motifs are listed and what are their accession numbers?
    b. Does the diagram support your answer for question 1c?

3. Now search the amino acid fasta sequence in **InterPro**.
    a. Under Domain and Conserved Site, hover over the different colored blocks. What other databases are cross referenced?
    b. Do the multiple databases have the same results?
    c. Click on the InterPro accession number (IPR##### for the repeated domain. Do the cross-referenced accession numbers for the domain in Pfam (PF#####) and SMART (SM#####) match those you obtained from searching the sequence in these databases (you should have written them in questions 1 and 2)? If not, why not?
    d. For both Pfam and SMART, click on the accession numbers, then the external links to view the repeated domain in the corresponding databases. Complete the following table comparing information for the repeated domain, this time comparing the three databases:

Database	**Pfam**	**SMART**	**InterPro**
Family/domain name			
Accession number			
# architectures that contain the domain			
# sequences/proteins in family			

    e. Take a look at the HMM sequence logo for this domain on Pfam. What do you notice?

Identifying protein domains of an unknown sequence

Imagine you've obtained the partial sequence of an unknown gene from an unlabeled RNA-seq experiment and you've translated it into an amino acid sequence.

1. How would you try to find out what protein family it belongs to?
2. Copy/paste or upload the contents of unknown-seq.fa to the three search tools. What are the top hits?
    SMART:
    Pfam:
    InterPro:
3. What are some other domains might you expect to be observed in the full protein sequence? How do you find this out?
4. Where did this sample likely come from?

Additional questions to ponder

1. Why is it useful to search through multiple databases to learn about a protein (or anything, really)?
2. What is the difference between a profileHMM consensus sequence and a profile HMM sequence logo?
3. What are the benefits of doing a profile HMM search over a BLAST search for finding related proteins?

Related tools

HMMer	HMM tool used by the above databases http://hmmer.org/ https://www.ebi.ac.uk/Tools/hmmer/search/phmmer
UniProt	Protein database without a sequence search function (for looking up known proteins) https://www.uniprot.org/