Assignment 1 : Aligning sequences and detecting motifs

Supporting files can be found on :
http://www.ulb.ac.be/di/map/tlenaert/Home_Tom_Lenaerts/INFO-F-439.html

Part 1; implementing the sequence alignment algorithms

Implement the Needleman-Wunsch (global) and Smith-Waterman (local) alignment algorithms in a Jupyter (python) notebook and report in the notebook on each part of your code with text and images[1]:

- Explain your implementation; for example, the dynamic programming algorithm and the back-tracking, the parser
- Explain your design choices; for example, detail the Abstract Data Types (ADTs) you are using
- Detail intermediate tests you made of each part of your code

Information on Jupyter notebooks (installing, use, etc.) can be found on http://jupyter.org. See also the slides `presentation-jupyter.pdf`, which are available on the website.

The global aligner needs to return the *k* best alignments, with *k* being a parameter in the function call. The local aligner needs to have the capacity to look for novel *l* sub-alignments: this means that once one sub-sequence is found, you need to set the values on the paths equal to zero and recalculate the matrix to find the next sub-alignment (see course discussion). Note that multiple paths (limited again by *k*) are again possible for each sub-alignment.

Make it possible to work with both linear and affine gap penalty.

You can use the sequences in `WW-sequences.fasta` to test the global aligner and the sequences in `protein-sequences.fasta` to check the local alignment algorithm.

The substation matrices (PAM120, PAM250 and BLOSUM matrices) are available online. You will need to write parser to read the files and use them in your code.

Compare the alignments that you make with your software to those made by the online tool LALIGN (http://www.ch.embnet.org/software/LALIGN_form.html). Explain similarities and differences also in the notebook.

When aligning the WW sequences, which ones are more similar? Are they coming from the same protein or different proteins? Check this via http://uniprot.org and report this in your notebook.

[1] Screen dumps are not acceptable. We need to be able to run each part in the Jupyter notebook.

Part 2; detecting motifs by aligning sequences to PSSMs

Now that you have a tool to make alignments we are going to expand it to a tool that can align sequences to motifs, represented by position-specific scoring matrices (PSSM). Such a tool will allow you to identify whether a particular domain family, represented by the PSSM, is present within a given protein sequence.

In a first step, you need to develop software that can construct the PSSM from a set of sequences. In this project the focus will be on WW domains, which are simple domains as visualized by the image in Figure 1.

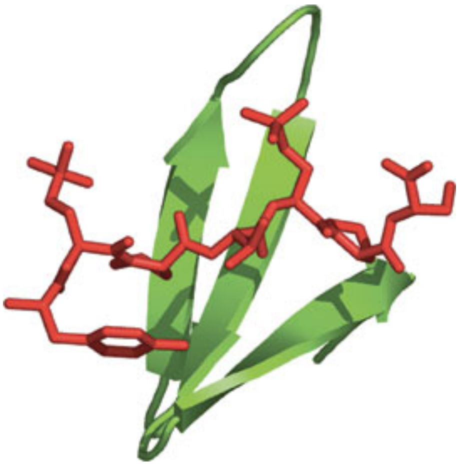
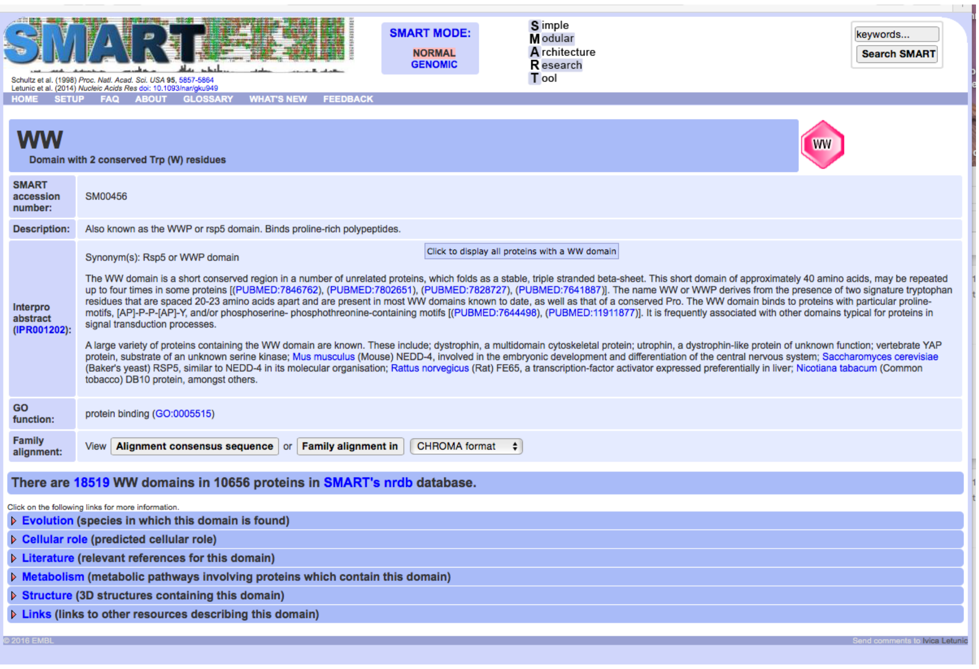


Figure 1: a typical WW domain structure. For details see the article WW and SH3 domains: two different scaffolds to recognize proline-rich ligands (2002) by Macias, Wiener and Sudol.

You will compare the PSSM to the motif that is available on PFAM for the same domain family. Make sure to explain, illustrate and test all aspects of your code in the notebook. Below I explain in detail how to get the WW domain information



Schultz et al. (1998) *Proc. Natl. Acad. Sci. USA* **95**, 5857-5864
Letunic et al. (2014) *Nucleic Acids Res* doi: 10.1093/nar/gku949

WW

Domain with 2 conserved Trp (W) residues

SMART accession number:	SM00456
Description:	Also known as the WWP or rsp5 domain. Binds proline-rich polypeptides.
Interpro abstract (IPR001202):	Synonym(s): Rsp5 or WWP domain Click to display all proteins with a WW domain The WW domain is a short conserved region in a number of unrelated proteins, which folds as a stable, triple stranded beta-sheet. This short domain of approximately 40 amino acids, may be repeated up to four times in some proteins [(PUBMED:7846762), (PUBMED:7802651), (PUBMED:7828727), (PUBMED:7641887)]. The name WW or WWP derives from the presence of two signature tryptophan residues that are spaced 20-23 amino acids apart and are present in most WW domains known to date, as well as that of a conserved Pro. The WW domain binds to proteins with particular proline-motifs, [AP]-P-P-[AP]-Y, and/or phosphoserine- phosphothreonine-containing motifs [(PUBMED:7644498), (PUBMED:11911877)]. It is frequently associated with other domains typical for proteins in signal transduction processes. A large variety of proteins containing the WW domain are known. These include; dystrophin, a multidomain cytoskeletal protein; utrophin, a dystrophin-like protein of unknown function; vertebrate YAP protein, substrate of an unknown serine kinase; Mus musculus (Mouse) NEDD-4, involved in the embryonic development and differentiation of the central nervous system; Saccharomyces cerevisiae (Baker's yeast) RSP5, similar to NEDD-4 in its molecular organisation; Rattus norvegicus (Rat) FE65, a transcription-factor activator expressed preferentially in liver; Nicotiana tabacum (Common tobacco) DB10 protein, amongst others.
GO function:	protein binding (GO:0005515)
Family alignment:	View **Alignment consensus sequence** or **Family alignment in** CHROMA format

There are 18519 WW domains in 10656 proteins in SMART's nrdb database.

Click on the following links for more information.

- **Evolution (species in which this domain is found)**
- **Cellular role (predicted cellular role)**
- **Literature (relevant references for this domain)**
- **Metabolism (metabolic pathways involving proteins which contain this domain)**
- **Structure (3D structures containing this domain)**
- **Links (links to other resources describing this domain)**

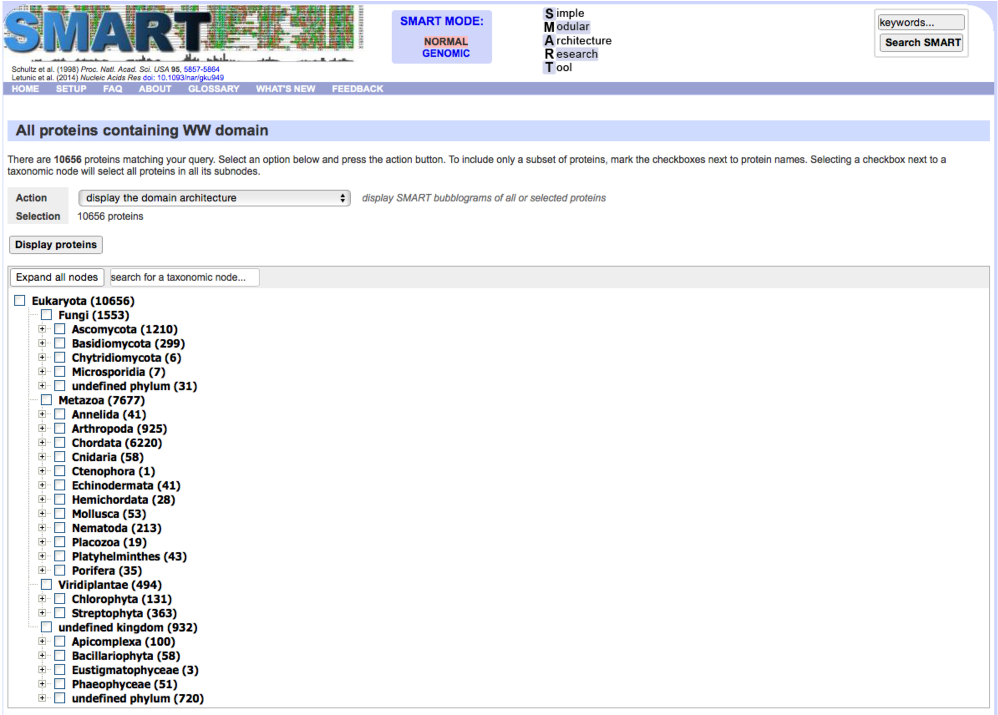
© 2016 EMBL Send comments to Ivica Letunic

Figure 2. SMART page for WW domains

The data

The set of sequences representing the entire family of WW domains is available in the database SMART (http://smart.embl.de), which needs to be used in normal mode (see home page of the smart website). When you select normal mode, you will move to another page that consists of 4 parts. In the part with the title « *Domains detected by SMART* », you need to type the word "WW" and click search. This will provide the page as visualized in Figure 2.

The SMART page for WW domains provides all information that is relevant for WW domains. It reports that there are 18519 instances of WW domains available. When you click the number 18519, the system searches for all proteins that contain WW domains. You should see a page equivalent to the page in Figure 3.



Schultz et al. (1998) *Proc. Natl. Acad. Sci. USA* **95**, 5857-5864
Letunic et al. (2014) *Nucleic Acids Res* doi: 10.1093/nar/gku949

All proteins containing WW domain

There are **10656** proteins matching your query. Select an option below and press the action button. To include only a subset of proteins, mark the checkboxes next to protein names. Selecting a checkbox next to a taxonomic node will select all proteins in all its subnodes.

Action: display the domain architecture — *display SMART bubblograms of all or selected proteins*
Selection: 10656 proteins

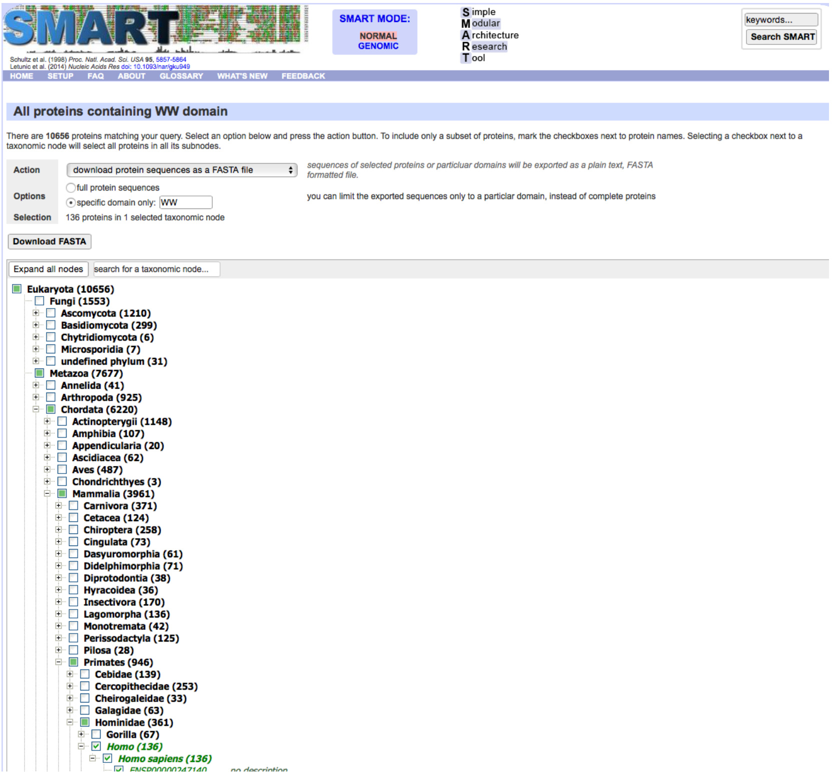
- □ Eukaryota (10656)
  - □ Fungi (1553)
    - ⊞ □ Ascomycota (1210)
    - ⊞ □ Basidiomycota (299)
    - ⊞ □ Chytridiomycota (6)
    - ⊞ □ Microsporidia (7)
    - ⊞ □ undefined phylum (31)
  - □ Metazoa (7677)
    - ⊞ □ Annelida (41)
    - ⊞ □ Arthropoda (925)
    - ⊞ □ Chordata (6220)
    - ⊞ □ Cnidaria (58)
    - ⊞ □ Ctenophora (1)
    - ⊞ □ Echinodermata (41)
    - ⊞ □ Hemichordata (28)
    - ⊞ □ Mollusca (53)
    - ⊞ □ Nematoda (213)
    - ⊞ □ Placozoa (19)
    - ⊞ □ Platyhelminthes (43)
    - ⊞ □ Porifera (35)
  - □ Viridiplantae (494)
    - ⊞ □ Chlorophyta (131)
    - ⊞ □ Streptophyta (363)
  - □ undefined kingdom (932)
    - ⊞ □ Apicomplexa (100)
    - ⊞ □ Bacillariophyta (58)
    - ⊞ □ Eustigmatophyceae (3)
    - ⊞ □ Phaeophyceae (51)
    - ⊞ □ undefined phylum (720)

Figure 3: SMART protein selection page

We will use this page now to look for the 136 WW sequences that can be found in human proteins. To get this data we first need to select the human species in the hierarchy visualized in Figures 3. Figure 4 shows where to find the human species exactly in this hierarchy. By clicking the "+" symbols you can descend in the tree to the correct level. You will see the number 136 next to the species "homo sapiens", which indicates the number of WW sequences found in that species.

Once you checked the box before "homo sapiens" you need to go back to the beginning of the page and select in the dropbox with title « *Action* » the option « *download protein sequences as fasta files* ». You also need to select « *Options* --

specific domain only: » and add the domain name, which is WW. Once you have done this you click « *Download FASTA* ».



All proteins containing WW domain

There are **10656** proteins matching your query. Select an option below and press the action button. To include only a subset of proteins, mark the checkboxes next to protein names. Selecting a checkbox next to a taxonomic node will select all proteins in all its subnodes.

Action: download protein sequences as a FASTA file — *sequences of selected proteins or particluar domains will be exported as a plain text, FASTA formatted file.*

Options: ○ full protein sequences / ● specific domain only: WW — you can limit the exported sequences only to a particlar domain, instead of complete proteins

Selection: 136 proteins in 1 selected taxonomic node

Download FASTA

Expand all nodes | search for a taxonomic node...

- ■ Eukaryota (10656)
  - □ Fungi (1553)
    - □ Ascomycota (1210)
    - □ Basidiomycota (299)
    - □ Chytridiomycota (6)
    - □ Microsporidia (7)
    - □ undefined phylum (31)
  - ■ Metazoa (7677)
    - □ Annelida (41)
    - □ Arthropoda (925)
    - ■ Chordata (6220)
      - □ Actinopterygii (1148)
      - □ Amphibia (107)
      - □ Appendicularia (20)
      - □ Ascidiacea (62)
      - □ Aves (487)
      - □ Chondrichthyes (3)
      - ■ Mammalia (3961)
        - □ Carnivora (371)
        - □ Cetacea (124)
        - □ Chiroptera (258)
        - □ Cingulata (73)
        - □ Dasyuromorphia (61)
        - □ Didelphimorphia (71)
        - □ Diprotodontia (38)
        - □ Hyracoidea (36)
        - □ Insectivora (170)
        - □ Lagomorpha (136)
        - □ Monotremata (42)
        - □ Perissodactyla (125)
        - □ Pilosa (28)
        - ■ Primates (946)
          - □ Cebidae (139)
          - □ Cercopithecidae (253)
          - □ Cheirogaleidae (33)
          - □ Galagidae (63)
          - ■ Hominidae (361)
            - □ Gorilla (67)
            - ☑ *Homo (136)*
              - ☑ *Homo sapiens (136)*
                - ☑ ENSP00000247140 no description

Figure 4: Where to find the human species in the species tree.

Once you clicked « *Download FASTA* », you will get the 136 domains that you need to produce the PSSM. You need to copy-and-paste the data in this page to a text file that can be named `to-be-aligned.fasta`. The sequences in this file will be used to create a multiple-sequence alignment.

> **IMPORTANT**: When you hand in your project, provide also this file next to the Jupyter notebook.

Making the multiple sequence alignment

Once you have the file `to-be-aligned.fasta` you can now align all sequences at the same time using one of the following tools. Mention clearly in your notebook which tool you used.

1. CLUSTAL Omega : http://www.ebi.ac.uk/Tools/msa/clustalo/
2. TCoffee : http://www.ebi.ac.uk/Tools/msa/tcoffee/
3. MUSCLE : http://www.ebi.ac.uk/Tools/msa/muscle/

Store the alignment in FASTA format in a file called `msaresults-<replace this by name of MSA tool>.fasta`.

IMPORTANT: When you hand in your project, provide also this file next to the Jupyter notebook.

Jupyter implementation

Now that you have the multiple-sequence alignment of the 136 sequences you can implement your code to construct the PSSM. See the slides of the course for the detail. Remember to use pseudo-counts. Explain in the notebook which approach for the PSSM construction and pseudo-counts you used.

Once you have the code to produce the PSSM, you should validate your results with what is known about the WW domains. Provide answers to the following questions and add them to your notebook. Use images to make our answers clear.

1) Construct a Weblogo (http://weblogo.threeplusone.com) for the WW domain family and compare this logo to what you see in your PSSM. Do the conserved positions correspond to what you see in the Weblogo?
2) Compare your results to the HMM-logo that can be found on the PFAM website (http://pfam.xfam.org). Write “WW” in the box next to « view a PFAm entry » on the main page) en click « go ». You will reach the page PF00397 and you can find the HMM logo on that page. What are the similarities and differences with your PSSM?

Expand your alignment code

Take now the code of part 1 and expand it so you can align a sequence to a PSSM.

1) Expand your alignment code (local!!) with linear gap-penalty. The recurrence relation is now (for linear gap penalty);
    a. Initialize the first row and first column as before
    b. Matrix S(i,j) = max {S(i-1,j-1) + PSSM(seq(i), j), S(i-1,j)+PSSM(“-“, j), S(i,j-1)+PSSM(“_”,j-1), 0}

        With j being the column in the PSSM and
        PSSM(“-“,j) is the penalty for the gap in position j as stored in your PSSM.

2) Align the sequences from `protein-sequences.fasta` to your PSSM. Show that you can identify the WW domains inside those sequences.
3) Use the information in Uniprot for these proteins to verify whether you have identified the domains correctly. Do you find for instance the same starting and ending positions?