Task 1: Motif Discovery using Prosite

Luis Garreta
Electiva de Bioinformática
MAESTRÍA EN INFORMÁTICA BIOMÉDICA
Universidad del Bosque
Bogotá-Colombia

April 14, 2021

1 Goal

The aims of this lab are to enable you to gain experience with sequence pattern discovery and pattern searching.

Databases and search tools:

- PROSITE is a database of protein families and domains. It consists of biologically significant sites, patterns and profiles that help to reliably identify to which known protein family (if any) a new sequence belongs.
- The ScanProsite tool allows users to scan protein sequence(s) (either from UniProt Knowledgebase (Swiss-Prot/TrEMBL) or PDB or provided by the user) for the occurrence of patterns, profiles and rules (motifs) stored in the PROSITE database, or to search protein database(s) for hits by specific motif(s)
- PRATT at EBI, http://ca.expasy.org/tools/pratt/. A quick user guide is here. Note that pratt works by repeated pattern extension, generates regular expression patterns in a Prosite syntax, and can find several patterns which are descriptive of the input sequences. Moreover it can also refine (improve) the pattern set which it finds.

2 Problem

A medical researcher has aligned a set of sequences that possibly contain an important protein motif (amino acids rather than DNA). The researcher needs the help of bioinformaticians to discover the unknown motif. The aligned sequences are:

RHSHYLPFRGGARNCIGTLLRFELLPDPTR
DPFSAYHFDGGARNCIGKQARLVVDTLREL
FDPGRARFFGGARNCIGQFAMKVLTLVRFE
DPSRFAPFDGGARNCIGFAMLKVGLRFELL
DSFADPRFSGGARNCIGQAMMIVLLRFELL
RAPSRSHFRGRARNRIGQFMEKASTLRELP
RFPSHHAFRGRARNPIGKQLTLLRFELLPD
PAPSRSHFPGRARNAIGKQFMEKAATLRDD
AGSSHSHFAGRARNFIGFAMNKVALTLRFE

If there is a significant motif, the researcher wants to confirm and know what protein motif it is and what its biological function is by using the PROSITE database, which is an annotated collection of motif descriptors dedicated to the identification of protein families and domains.

In this realistic exercise, we are going to:

- Construct a motif from aligned sequences
- Search and confirm the motif in Prosite's motif database.
- Retrieve the sequences where the motif is found.
- Use any hits that we find to make some other patterns characteristic of the retrieved set. These patterns will be regular expressions, generated using pratt.

3 Protocol

1. To discover the motif, you can modify the R script for discovering ADN motifs to work with protein motifs (aminoacids). Follow the next steps:

   (a) Create a string vector from the sequences

```
sequences = c(
"RHSHYLPFRGGARNCIGTLLRFELLPDPTR",
"DPFSAYHFDGGARNCIGKQARLVVDTLREL",
"FDPGRARFFGGARNCIGQFAMKVLTLVRFE",
"DPSRFAPFDGGARNCIGFAMLKVGLRFELL",
"DSFADPRFSGGARNCIGQAMMIVLLRFELL",
"RAPSRSHFRGRARNRIGQFMEKASTLRELP",
"RFPSHHAFRGRARNPIGKQLTLLRFELLPD",
"PAPSRSHFPGRARNAIGKQFMEKAATLRDD",
"AGSSHSHFAGRARNFIGFAMNKVALTLRFE")
```

   (b) Convert the string to a character matrix

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]	[,11]	[,12]	[,13]	[,14]	[,15]	[,16]	[,17]	[,18]	[,19]	[,20]	[,21]	[,22]	[,23]	[,24]	[,25]	[,26]	[,27]	[,28]	[,29]	[,30]
[1,]	"R"	"H"	"S"	"H"	"Y"	"L"	"P"	"F"	"R"	"G"	"G"	"A"	"R"	"N"	"C"	"I"	"G"	"T"	"L"	"L"	"R"	"F"	"E"	"L"	"L"	"P"	"D"	"P"	"T"	"R"
[2,]	"D"	"P"	"F"	"S"	"A"	"Y"	"H"	"F"	"D"	"G"	"G"	"A"	"R"	"N"	"C"	"I"	"G"	"K"	"Q"	"A"	"R"	"L"	"V"	"V"	"D"	"T"	"L"	"R"	"E"	"L"
[3,]	"F"	"D"	"P"	"G"	"R"	"A"	"R"	"F"	"F"	"G"	"G"	"A"	"R"	"N"	"C"	"I"	"G"	"Q"	"F"	"A"	"M"	"K"	"V"	"L"	"T"	"L"	"V"	"R"	"F"	"E"
[4,]	"D"	"P"	"S"	"R"	"F"	"A"	"P"	"F"	"D"	"G"	"G"	"A"	"R"	"N"	"C"	"I"	"G"	"F"	"A"	"M"	"L"	"K"	"V"	"G"	"L"	"R"	"F"	"E"	"L"	"L"
[5,]	"D"	"S"	"F"	"A"	"D"	"P"	"R"	"F"	"S"	"G"	"G"	"A"	"R"	"N"	"C"	"I"	"G"	"Q"	"A"	"M"	"M"	"I"	"V"	"L"	"L"	"R"	"F"	"E"	"L"	"L"
[6,]	"R"	"A"	"P"	"S"	"R"	"S"	"H"	"F"	"R"	"G"	"R"	"A"	"R"	"N"	"R"	"I"	"G"	"Q"	"F"	"M"	"E"	"K"	"A"	"S"	"T"	"L"	"R"	"E"	"L"	"P"
[7,]	"R"	"F"	"P"	"S"	"H"	"H"	"A"	"F"	"R"	"G"	"R"	"A"	"R"	"N"	"P"	"I"	"G"	"K"	"Q"	"L"	"T"	"L"	"L"	"R"	"F"	"E"	"L"	"L"	"P"	"D"
[8,]	"P"	"A"	"P"	"S"	"R"	"S"	"H"	"F"	"P"	"G"	"R"	"A"	"R"	"N"	"A"	"I"	"G"	"K"	"Q"	"F"	"M"	"E"	"K"	"A"	"A"	"T"	"L"	"R"	"D"	"D"
[9,]	"A"	"G"	"S"	"S"	"H"	"S"	"H"	"F"	"A"	"G"	"R"	"A"	"R"	"N"	"F"	"I"	"G"	"F"	"A"	"M"	"N"	"K"	"V"	"A"	"L"	"T"	"L"	"R"	"F"	"E"

   (c) Calculate the position frequency matrix:

   - Take into account that you are working with aminoacids instead of nucleotides.

```
aminos = c ("A","R","N","D","C","Q","E","G","H","I","L",
            "K","M","F","P","S","T", "W","Y","V","B","Z","X")
instead of
bases = c ("A","C","G","T")
```

   - Add a low pseudocount to the frecuency matrix (0.01).

```
freqMatrix = freqMatrix + 0.01
```

	p1	p2	p3	p4	p5	p6	p7	p8	p9	p10	p11	p12	p13	p14	p15	p16	p17	p18	p19	p20	p21	p22	p23	p24	p25	p26	p27	p28	p29	p30
A	1.01	2.01	0.01	1.01	1.01	2.01	1.01	0.01	1.01	0.01	0.01	9.01	0.01	0.01	1.01	0.01	0.01	0.01	3.01	2.01	0.01	0.01	1.01	2.01	1.01	0.01	0.01	0.01	0.01	0.01
R	3.01	0.01	0.01	1.01	3.01	0.01	2.01	0.01	3.01	0.01	4.01	0.01	9.01	0.01	1.01	0.01	0.01	0.01	0.01	0.01	2.01	0.01	0.01	1.01	0.01	2.01	1.01	4.01	0.01	1.01
N	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	9.01	0.01	0.01	0.01	0.01	0.01	0.01	1.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
D	3.01	1.01	0.01	0.01	1.01	0.01	0.01	0.01	2.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	1.01	0.01	1.01	0.01	1.01	2.01
C	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	5.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
Q	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	3.01	3.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
E	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	1.01	1.01	1.01	0.01	0.01	1.01	0.01	3.01	1.01	2.01
G	0.01	1.01	0.01	1.01	0.01	0.01	0.01	0.01	0.01	9.01	5.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	9.01	0.01	0.01	0.01	0.01	0.01	1.01	0.01	0.01	0.01	0.01	0.01
H	0.01	1.01	0.01	1.01	2.01	1.01	4.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
I	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	9.01	0.01	0.01	0.01	0.01	1.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
L	0.01	0.01	0.01	0.01	0.01	1.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	2.01	1.01	2.01	1.01	3.01	4.01	2.01	4.01	1.01	3.01	3.01
K	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	3.01	0.01	0.01	0.01	4.01	1.01	0.01	0.01	0.01	0.01	0.01	0.01
M	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	4.01	3.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
F	1.01	1.01	2.01	0.01	1.01	0.01	0.01	9.01	1.01	0.01	0.01	0.01	0.01	0.01	1.01	0.01	0.01	2.01	2.01	1.01	0.01	1.01	0.01	0.01	1.01	0.01	2.01	0.01	2.01	0.01
P	1.01	2.01	4.01	0.01	0.01	1.01	2.01	0.01	1.01	0.01	0.01	0.01	0.01	0.01	0.01	1.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	1.01	0.01	1.01	1.01	1.01
S	0.01	1.01	3.01	5.01	0.01	3.01	0.01	0.01	1.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	1.01	0.01	0.01	0.01	0.01	0.01	0.01
T	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	1.01	0.01	0.01	1.01	0.01	0.01	0.01	2.01	3.01	0.01	0.01	1.01	0.01
W	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
Y	0.01	0.01	0.01	0.01	1.01	1.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
V	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	5.01	1.01	0.01	0.01	1.01	0.01	0.01
B	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
Z	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
X	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01

   (d) Calculate de position weight matrix (PWM):

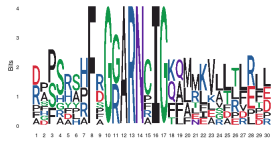
   - Take into account the added pseudocount to normalize the frequencies:

```
probMatrix = freqMatrix/(N+23*0.01)
```

	p1	p2	p3	p4	p5	p6	p7	p8	p9	p10	p11	p12	p13	p14	p15	p16	p17	p18	p19	p20	p21	p22	p23	p24	p25	p26	p27	p28	p29	p30
A	0.11	0.22	0.00	0.11	0.11	0.22	0.11	0.00	0.11	0.00	0.00	0.98	0.00	0.00	0.11	0.00	0.00	0.00	0.33	0.22	0.00	0.00	0.11	0.22	0.11	0.00	0.00	0.00	0.00	0.00
R	0.33	0.00	0.00	0.11	0.33	0.00	0.22	0.00	0.33	0.00	0.43	0.00	0.98	0.00	0.11	0.00	0.00	0.00	0.00	0.00	0.22	0.00	0.00	0.11	0.00	0.22	0.11	0.43	0.00	0.11
N	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.98	0.00	0.00	0.00	0.00	0.00	0.00	0.11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
D	0.33	0.11	0.00	0.00	0.11	0.00	0.00	0.00	0.22	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.11	0.00	0.11	0.00	0.11	0.22
C	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.54	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Q	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.33	0.33	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
E	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.11	0.11	0.11	0.00	0.00	0.11	0.00	0.33	0.11	0.22
G	0.00	0.11	0.00	0.11	0.00	0.00	0.00	0.00	0.98	0.54	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.98	0.00	0.00	0.00	0.00	0.00	0.11	0.00	0.00	0.00	0.00	0.00	0.00	0.00
H	0.00	0.11	0.00	0.11	0.22	0.11	0.43	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
I	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.98	0.00	0.00	0.00	0.00	0.11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
L	0.00	0.00	0.00	0.00	0.00	0.11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.11	0.22	0.11	0.22	0.11	0.33	0.43	0.22	0.43	0.11	0.33	0.33
K	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.33	0.00	0.00	0.43	0.11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
M	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.43	0.33	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
F	0.11	0.11	0.22	0.00	0.11	0.00	0.00	0.98	0.11	0.00	0.00	0.00	0.00	0.00	0.11	0.00	0.00	0.22	0.22	0.11	0.00	0.11	0.00	0.00	0.11	0.00	0.22	0.00	0.22	0.00
P	0.11	0.22	0.43	0.00	0.00	0.11	0.22	0.00	0.11	0.00	0.00	0.00	0.00	0.00	0.11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.11	0.00	0.11	0.11	0.11
S	0.00	0.11	0.33	0.54	0.00	0.33	0.00	0.00	0.11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.11	0.00	0.00	0.00	0.00	0.00	0.00
T	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.11	0.00	0.00	0.11	0.00	0.00	0.00	0.22	0.33	0.00	0.00	0.11	0.00
W	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Y	0.00	0.00	0.00	0.00	0.11	0.11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
V	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.54	0.11	0.00	0.00	0.11	0.00	0.00	0.00
B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Z	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
X	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

(e) Create the logo from the PWM, and from it construct the regular expression pattern. For the regular expresion take into account:

- The motif must start with a completely conserved aminoacid
- The motif must end with a completely conserved aminoacid



2. Now, search the motif in SwissProt using the ScanProsite:

   (`https://prosite.expasy.org/scanprosite/`).

   Select "***Option 2***" in the form and enter the motif to search for sequences matching this motif.

Report the name of the motif and its biological function.

1. Retrieve these hits in fasta format, go to "STEP 3" of the form, select a maximum of 10 sequences and save them all in one file. To do this, you have to select either:

   - "FASTA" output format, then select the sequences, **or**
   - "Graphical view" "output format, then go to the page for each sequence, hit and click on the 'fasta' button, and save them all in one file.

   Report the sequences.

2. Check that these sequences are specific to the search motif using the ScanProsite. Select ***"Option 1"*** in the form, paste the sequences into the ScanProsite input form to search for PROSITE collection of motifs in the sequences.

   Report the graphical view of the first result.

3. Use PRATT (`http://www.ebi.ac.uk/pratt/`) as an automated method to construct regular expressions characterising these sequences. To obtain these regular expressions, use the fasta sequences retrieved in the previous step, copy them into the input form and send them to PRATT. Wait for the results and take from them larger regular expression. Report this regular expression.

4. Finally. Search back into SwissProt with ScanProsite using the largest pattern and see what hits are obtained.

   - Do these searches return all of the original sequences?
   - What other sequences (if any) are identified by these patterns?