Lab 1 – Protein Domain, Motif and Profile Analysis

[Software need: web access]

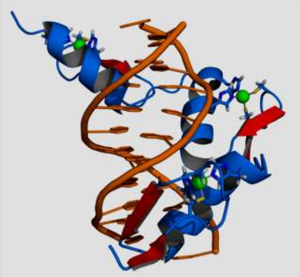
In this lab we will examine several online databases of protein domains, protein motif and profile information, as well as methods for scanning sequences for novel profiles using web-based tools.

From a string of amino acids, the folding of a protein into its secondary and tertiary conformations produces a complex and highly specialized piece of biological machinery. The contiguous regions of a completely folded protein that possess some form of critical functional role are referred to as protein domains – these are often functional when detached from a given protein, and can be exemplified by the DNA binding domain of a transcription factor. The amino acid primary sequence that constitutes a domain often contains smaller protein or sequence motifs that exhibit conservation, not only within a set of conserved domains, but also in other parts of other proteins, for instance a C2H2 Zinc Finger motif. Protein domain functions can range from protein-protein interactions, protein-nucleic acid interactions, catalytic activities or structural roles. Operating in concert with other domains, proteins, nucleic acids, or metabolites, protein domains ultimately function in a modular fashion to produce the variety of protein functional roles seen in nature.

Since the primary sequence of amino acids that make up any single protein domain can be fairly flexible in its composition, methods for the representation of site degeneracy or gaps in sequence have been devised. Sequence motifs are represented in the form of Regular Expressions, Position Specific Scoring Matrices (PSSMs) or an extension of the PSSM, the profile Hidden Markov Model (pHMM). pHMMs allow for the flexible representation of site-specific variability, the presence of gaps in the sequence, and site-to-site probabilistic tendencies in residue content along the length of the motif. Since thousands of protein motifs have already been characterized *in vivo* and *in silico*, there exist expansive databases of motifs and pHMMs to use in scanning a novel sequence whose function cannot be inferred by sequence homology alone.

Box 1. Representing protein motifs with regular expressions

A “classic” motif is the C2H2 zinc finger motif, 3 copies of which are present in Egr1 (early growth response protein 1), a human transcription factor. At right is a cartoon of the DNA binding domain of this transcription factor bound to DNA in orange, showing the 3 zinc fingers in blue and red complexing 3 zinc atoms in green. Two histidines and two cysteines at proper spacing are required for the proper coordination of the Zn atoms. How can we represent such requirements?



Egr1 image courtesy of Thomas Splettstoesser: Creative Commons Attribution ShareAlike 2.5 licence.

Protein motifs can be represented using “regular expressions”. Such expressions were initially developed by computer scientists in order to search text for certain patterns. The C2H2 Zinc Finger motif may be described using the following regular expression:

C-x(2,4)-**C**-x(3)-[LIVMFYWC]-x(8)-**H**-x(3,5)-**H**.
→ *the two* Cs *and two* Hs *are the zinc ligands*

This motif is interpreted as a Cysteine, followed by 2-4 of any amino acid, then another cysteine, followed by any 3 amino acids, any one of L, I, V, M, F, Y, W or C, then 8 of any amino acid, a histidine, 3-5 of any amino acid and finally another histidine.

Rules for describing motifs using regular expressions are as follows:

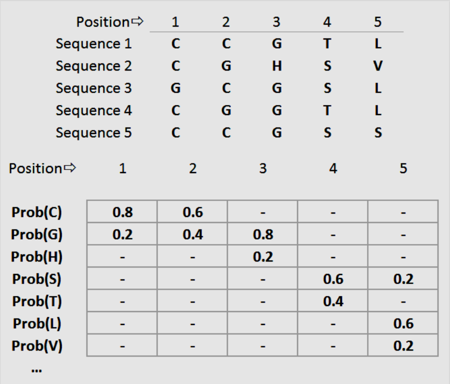
- The standard IUPAC one-letter codes are used.
- x is used for a position where any amino acid is accepted.
- Ambiguities are within [square parentheses]. [ALT] means Ala or Leu or Thr.
- more general ambiguities use a {pair of curly braces} to indicate *disallowed* residues. {AM} is any amino acid *except* Ala or Met.
- Each element in a pattern is separated using a dash -.
- Repetition is denoted using a numerical value or a numerical range between parentheses. x(3) means x-x-x, x(2,4) means x-x or x-x-x or x-x-x-x.
- Patterns at the N- or C-terminal of a sequence start with a < or end with a > symbol, respectively.
- A period ends the pattern.

There are several problems with representing motifs using regular expressions – they are inflexible in that if one position in the regular expression does not match to a search sequence, that sequence is discarded as not containing that motif. Another problem is that no information as to the frequency of ambiguous amino acids is stored in the regular expression.

Box 2. Representing protein motifs using Position-specific Scoring Matrices

Position-specific scoring matrices (PSSMs) provide an improved way to capture the information present in a sequence alignment. They also allow the possibility of statistically evaluating how well any given sequence matches the information represented in the PSSM.

An example alignment and its corresponding PSSM are shown to the right. In the first column of the alignment, 4 of 5 amino acids are cysteines, so a probability of 80% or 0.8 is entered into the PSSM for identifying a cysteine at that position.

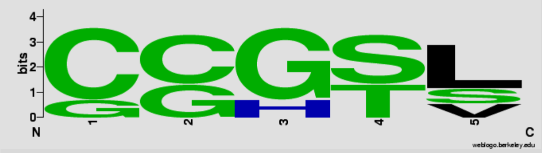


Position⇨	1	2	3	4	5
Sequence 1	C	C	G	T	L
Sequence 2	C	G	H	S	V
Sequence 3	G	C	G	S	L
Sequence 4	C	G	G	T	L
Sequence 5	C	C	G	S	S

Position⇨	1	2	3	4	5
Prob(C)	0.8	0.6	-	-	-
Prob(G)	0.2	0.4	0.8	-	-
Prob(H)	-	-	0.2	-	-
Prob(S)	-	-	-	0.6	0.2
Prob(T)	-	-	-	0.4	-
Prob(L)	-	-	-	-	0.6
Prob(V)	-	-	-	-	0.2
...					

Sequence alignment and corresponding PSSM.

We can use the PSSM to calculate the probability of any given sequence matching the information stored in the PSSM. Consider the sequence CGGSV: the probability of seeing this sequence given the PSSM is simply the product of the probabilities of seeing these amino acids at each position, in this example 0.8 × 0.4 × 0.8 × 0.6 × 0.2 = 0.031.



SeqLogo of the above alignment.

As an aside, sequence logos provide a convenient way to represent the information in a sequence alignment. The height of the block of letters at a given position represents the "information content" (roughly, the degree of conservation) of the alignment column, while the relative letter size within a column represents its proportional abundance in that position. The height of column 5 for a seqlogo representation of the alignment above is lower because there is more variation at that position.

While PSSMs are valuable for representing information present in a sequence alignment, they are not able to represent one biologically important aspect, namely insertions/deletions within individual sequences comprising the alignment – see **Box 3** for a solution to this problem!

Protein Domain Information online

CDD – Conserved Domain Database

NCBI's Conserved Domain Database provides an expansive collection of protein domain architecture information. Connect to http://www.ncbi.nlm.nih.gov and select 'Domains & Structures' in the Resources menu then on 'Conserved Domain Database (CDD)' link. Click on 'HELP' in the top right. Scroll down to read the definition of a conserved protein domain and the CDD content information.

a. What source motif databases does the CDD incorporate?

Let's examine the structure of the Breast Cancer Type 2 susceptibility protein BRCA2. BRCA2 interacts with RAD51 in the DNA damage and repair response pathway. Both proteins are critical to the proper function of the DNA damage response and mutations in each have been linked to multiple forms of cancer including breast cancer. **BRCA2 interacts with RAD51 via the presence of BRC repeat domains in BRCA2.**

Open a separate browser window and connect to NCBI http://www.ncbi.nlm.nih.gov. Enter BRCA2 into the search box and search within the 'Protein' database at NCBI. Select the BRCA2 protein for *Homo sapiens* and retrieve the protein sequence by selecting FASTA from the link just under the accession number. Copy the fasta sequence of BRCA2 to your clipboard by highlighting it and hitting Ctrl (Windows) or ⌘ (Mac) then C, or by using mouse commands.

b. Comment on the size of the BRCA2 protein.

Return to the CDD help window and click on 'SEARCH' in the top right navigation menu. Paste the BRCA2 sequence into the CDD search box and click 'Submit Query'. (The search function is also accessible from the main CDD page, under the CD-Search link under CDD Tools)

c. How many distinct protein domains does the BRCA2 protein possess (this is a bit tricky to assess, but go by the number of unique hits in List of Domain Hits from the Specific Hits part of the graphic – mousing over the Specific Hits symbols will helpfully highlight entries in the List of Domain Hits)?

d. How many BRCA2 repeat domains (also called BRC repeats) does BRCA2 possess?

CDART

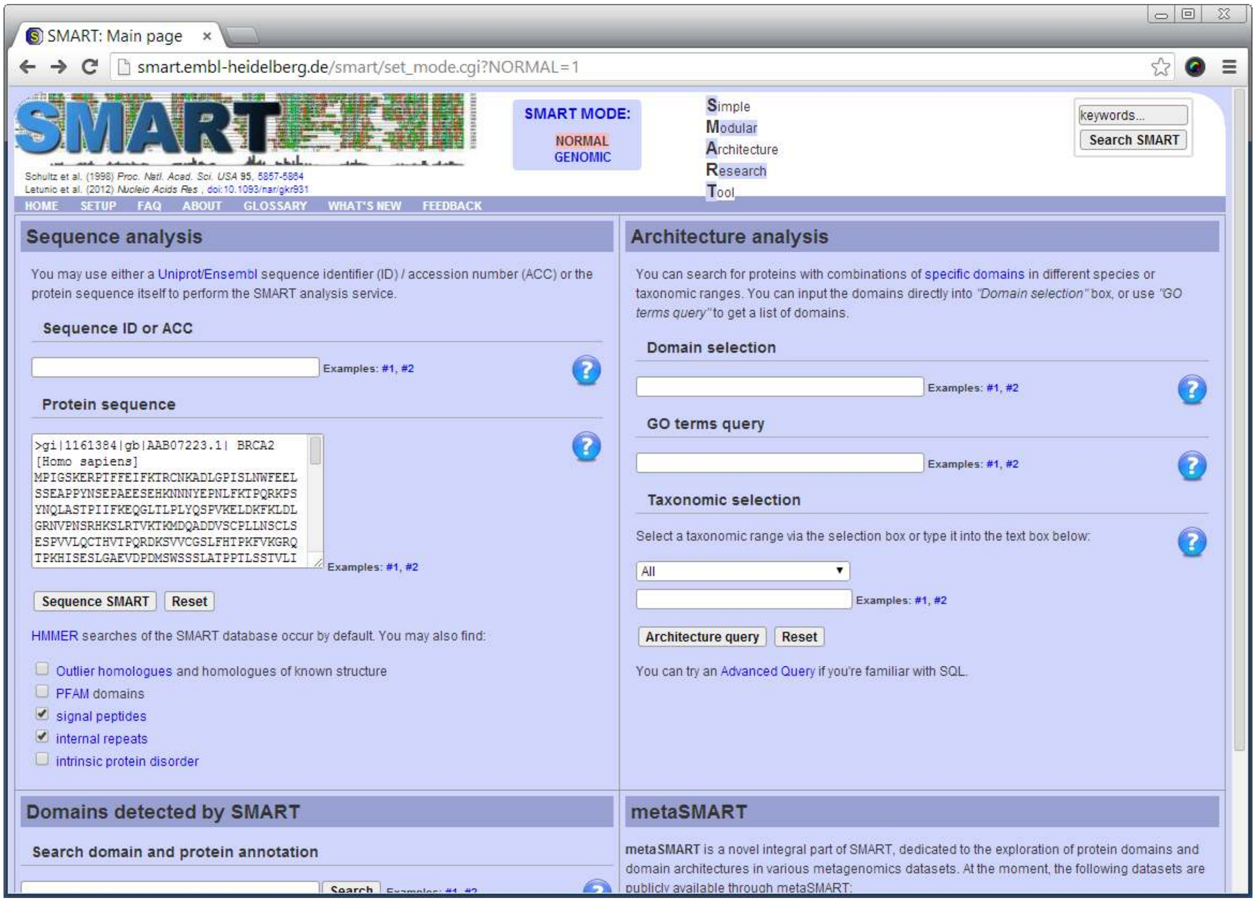
Click on 'Search for similar domain architectures' at the bottom of the 'Graphical Summary' box. This will return the CDART (Conserved Domain Architecture Retrieval Tool) information for BRCA2. You can apply filters to answer the following question – select the small drop down beside the Filter Your Results box, and exclude the BRC repeat by ticking it in the Superfamily list, then click the Boolean operator AND, and finally include the RPA_2b region by ticking it in the list. Click Apply. If you enter *EXCLDOM[2912] AND INCLDOM[9930]* into the Filter box and click Apply you will achieve the same results.

e. How many eukaryotic species (that is, the taxonomy span should be Cellular Organisms, or at least Eukaryota) possess a BRCA2-like region containing OB1, OB2, OB3 (denoted as RPA 2b in CDART) but lack BRC repeats (denoted in CDART as BRC2)?

f. What does this suggest about these domains? Could they function independently of one another?

SMART

Go to the SMART search program at http://smart.embl-heidelberg.de/ for protein domain searching. Choose 'Normal Mode' and paste the BRCA2 sequence into the search box (Ctrl or ⌘ then V; it should still be on your clipboard). Select searches for 'signal peptides' and 'internal repeats'. Click on 'Sequence SMART'.



Sequence analysis

You may use either a Uniprot/Ensembl sequence identifier (ID) / accession number (ACC) or the protein sequence itself to perform the SMART analysis service.

Sequence ID or ACC

Examples: #1, #2

Protein sequence

```
>gi|1161384|gb|AAB07223.1| BRCA2
[Homo sapiens]
MPIGSKERPTFFEIFKTRCNKADLGPISLNWFEEL
SSEAPPYNSEPAEESEHKNNNYEPNLFKTPQRKPS
YNQLASTPIIFKEQGLTLPLYQSPVKELDKFKLDL
GRNVPNSRHKSLRTVKTKMDQADDVSCPLLNSCLS
ESPVVLQCTHVTPQRDKSVVCGSLFHTPKFVKGRQ
TPKHISESLGAEVDPDMSWSSSLATPPTLSSTVLI
```

Examples: #1, #2

HMMER searches of the SMART database occur by default. You may also find:

- ☐ Outlier homologues and homologues of known structure
- ☐ PFAM domains
- ☑ signal peptides
- ☑ internal repeats
- ☐ intrinsic protein disorder

Architecture analysis

You can search for proteins with combinations of specific domains in different species or taxonomic ranges. You can input the domains directly into *"Domain selection"* box, or use *"GO terms query"* to get a list of domains.

Domain selection

Examples: #1, #2

GO terms query

Examples: #1, #2

Taxonomic selection

Select a taxonomic range via the selection box or type it into the text box below:

All ▾

Examples: #1, #2

You can try an Advanced Query if you're familiar with SQL.

Domains detected by SMART

Search domain and protein annotation

Examples: #1, #2

metaSMART

metaSMART is a novel integral part of SMART, dedicated to the exploration of protein domains and domain architectures in various metagenomics datasets. At the moment, the following datasets are publicly available through metaSMART:

Figure 1. SMART input page with signal peptide and internal repeat search flags set.

g. Does BRCA2 possess any signal peptides or transmembrane domains?

h. How many regions of low complexity does BRCA2 possess (you may need to scroll to the right so see the entire protein)?

Lab Quiz Question 1

Pfam

Let’s now look at the primary source for pHMM information, Pfam. Go to http://pfam.sanger.ac.uk/ (there are also several ‘mirror’ sites around the world). Click on ‘SEARCH’ and enter your gene sequences in the user box. Click on ‘Submit’ to scan your sequence for Pfam-A domains (use the default search parameters).

i. How many distinct protein domains did Pfam identify?

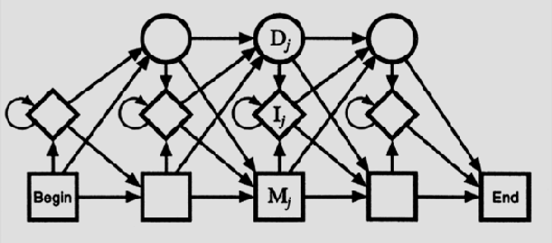
Click on the ‘BRCA2 repeat’ domain link in table below the sequence cartoon and select ‘Domain Organisation’ tab in the left navigation pane.

j. Would you say the ‘BRC repeat’ domain occurs in non-BRCA2 orthologous proteins?

k. If so, can you say anything interesting about the species that possess strictly BRC repeat possessing proteins and no other BRCA2-type domains?

Box 3. Representing protein motifs using profile Hidden Markov Models

While PSSMs offer a probabilistic way of representing relative amino acid abundance at a given alignment position, the representation of insertions and deletions is problematic. Profile HMMs offer a framework to represent both these and the probabilities of the individual amino acids in each column of the alignment.



Profile HMM representation for a sequence alignment of 3 columns.

In the figure in this box, the squares are the "match" states, and these correspond to the columns of an amino acid alignment. For each of the 20 amino acids, there is an "emission" probability associated with each match state. These emission probabilities correspond to the frequency of occurrence of the amino acids in a given column of the PSSM.

The diamonds in the figure correspond to "insert" states. An insert state has associated with it probabilities of "emitting" each of the amino acids as an inserted amino acid. These probabilities would come from sequence alignments containing insertions in some of the sequences, but to maintain flexibility in the model, insertions can be allowed with a very small probability anywhere in the sequence.

The circles in the figure correspond to "delete" states. By definition, deletions are the absence of amino acids at that position and are thus not associated with any amino acid emission probabilities.

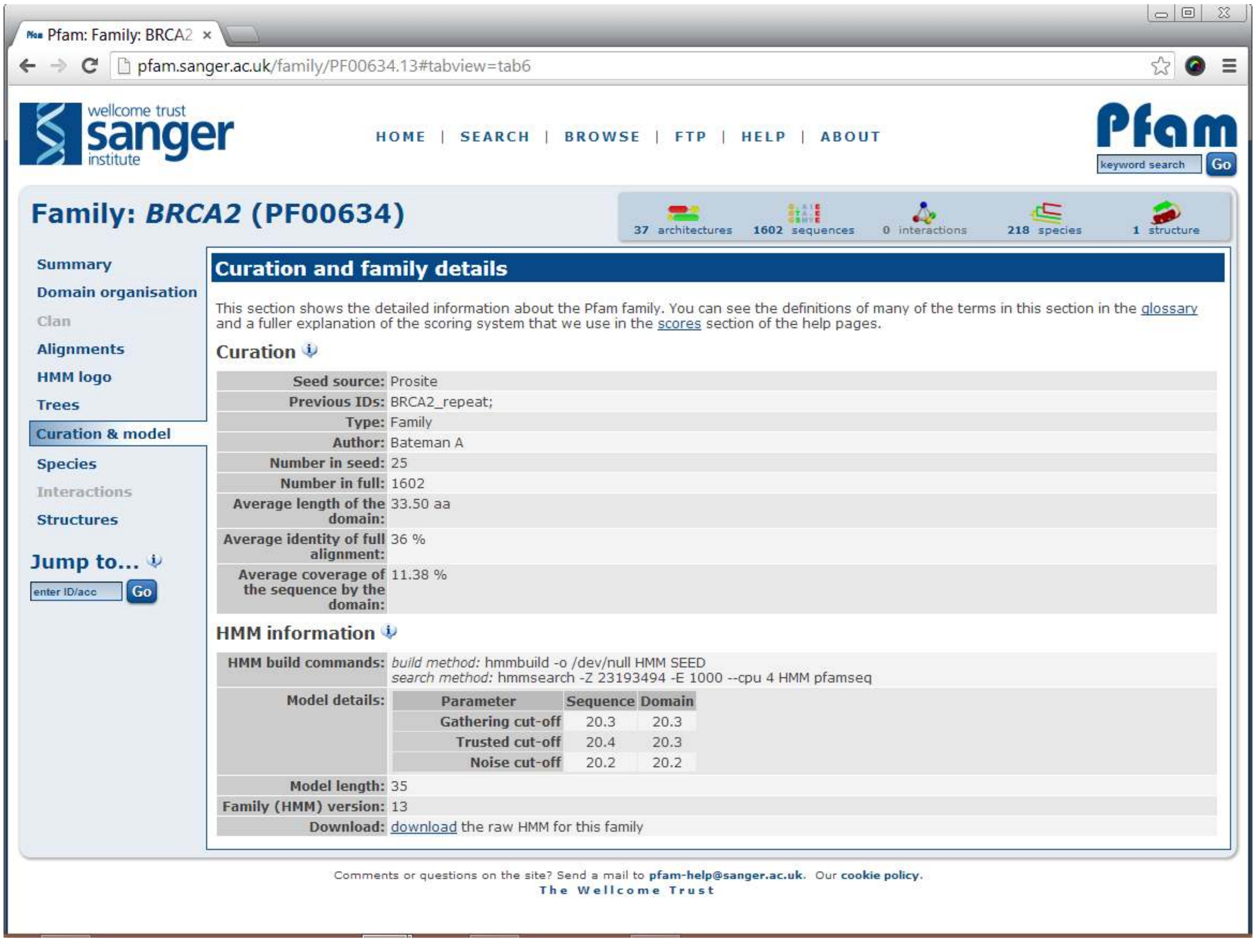
Finally, the arrows represent the "transitions" between states. Each arrow has associated with it a "transition" probability. In the above example, we start at the beginning, and can then move to either an insert state, delete state or the 1st match state. The probability of any of these happening is the transition probability, and depends on the sequence alignment used to generate the HMM. We then move to the next state and can emit an amino acid at the next position. For any given amino acid sequence produced by the HMM "automaton" we can calculate the probability of its being generated by simply multiplying the transition and emission probabilities along the path through the HMM that was used to generate the sequence.

Once we have a profile HMM, we can also search (and assign a probability score to) any arbitrary sequence for how well it matches the information captured in the pHMM. This is how the PFAM database search function identifies whether or not a sequence contains a match.

Let's view the sequences used to generate the BRC repeat HMM in a sequence logo-like format by clicking on the 'HMM logo' tab on the BRCA2 Family page from the Pfam database from the previous step. You can see that not all positions are equally well conserved.

l. Which position is best conserved in the BRC repeat (PF000634)?

Lab Quiz Question 2



Family: *BRCA2* (PF00634)

Curation and family details

This section shows the detailed information about the Pfam family. You can see the definitions of many of the terms in this section in the glossary and a fuller explanation of the scoring system that we use in the scores section of the help pages.

Curation

Seed source:	Prosite
Previous IDs:	BRCA2_repeat;
Type:	Family
Author:	Bateman A
Number in seed:	25
Number in full:	1602
Average length of the domain:	33.50 aa
Average identity of full alignment:	36 %
Average coverage of the sequence by the domain:	11.38 %

HMM information

HMM build commands: *build method:* hmmbuild -o /dev/null HMM SEED
search method: hmmsearch -Z 23193494 -E 1000 --cpu 4 HMM pfamseq

Model details:

Parameter	Sequence	Domain
Gathering cut-off	20.3	20.3
Trusted cut-off	20.4	20.3
Noise cut-off	20.2	20.2

Model length: 35

Family (HMM) version: 13

Download: download the raw HMM for this family

Comments or questions on the site? Send a mail to pfam-help@sanger.ac.uk. Our cookie policy.
The Wellcome Trust

Figure 2. Curation and family details for the BRCA2 repeat in Pfam. The HMM model may be downloaded by clicking on the appropriate download link.

Now click on 'Curation & Models' in the left navigation pane of the BRC repeat domain screen from Pfam.

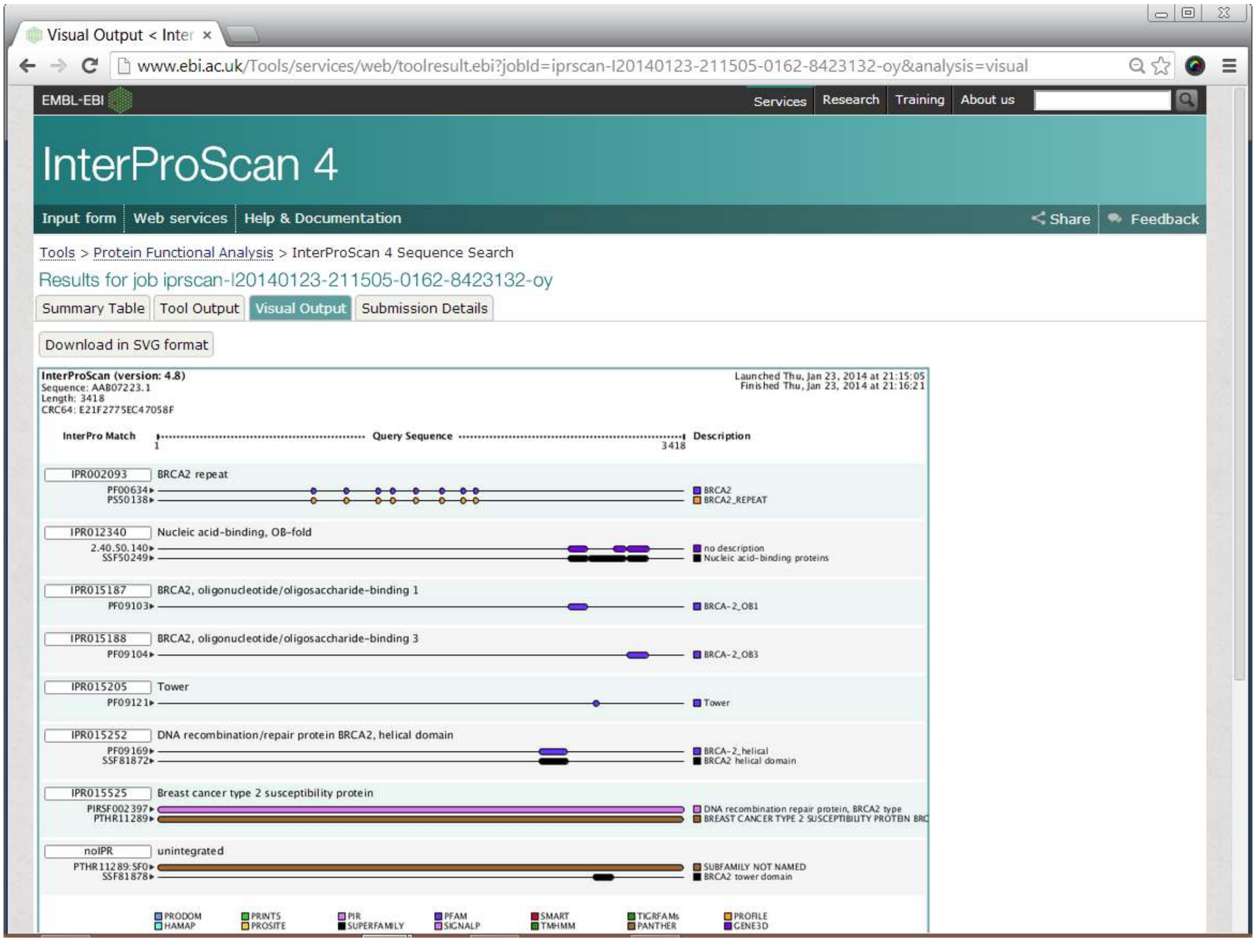
m. Under 'HMM information' what were the HMMer build commands to create the HMM for the BRC repeat? NOTE: 'SEED' refers to the multiple sequence alignment file to build from.

You can download the HMM model for the BRC repeat by clicking on the 'download' link at the bottom of the page and saving to your computer.

InterProScan

Interpro is an overarching collection of many domain, motif, and profile HMM databases. You can search a given sequence against all of the associated Interpro DBs using Interproscan: http://www.ebi.ac.uk/InterProScan/. Paste your sequence into the search box and click Submit. By default, all databases that comprise Interpro will be searched using the appropriate algorithms.

It is useful to compare the results of a couple of different motif scanning programs as they may encompass different databases. See the results of an InterProScan of the human BRCA2 protein in **Figure 3**.



Results for job iprscan-I20140123-211505-0162-8423132-oy

```
InterProScan (version: 4.8)                                   Launched Thu, Jan 23, 2014 at 21:15:05
Sequence: AAB07223.1                                           Finished Thu, Jan 23, 2014 at 21:16:21
Length: 3418
CRC64: E21F2775EC47058F

InterPro Match   1 ........ Query Sequence ........ 3418   Description

IPR002093        BRCA2 repeat
      PF00634▸                                             BRCA2
      PS50138▸                                             BRCA2_REPEAT

IPR012340        Nucleic acid-binding, OB-fold
  2.40.50.140▸                                             no description
     SSF50249▸                                             Nucleic acid-binding proteins

IPR015187        BRCA2, oligonucleotide/oligosaccharide-binding 1
      PF09103▸                                             BRCA-2_OB1

IPR015188        BRCA2, oligonucleotide/oligosaccharide-binding 3
      PF09104▸                                             BRCA-2_OB3

IPR015205        Tower
      PF09121▸                                             Tower

IPR015252        DNA recombination/repair protein BRCA2, helical domain
      PF09169▸                                             BRCA-2_helical
     SSF81872▸                                             BRCA2 helical domain

IPR015525        Breast cancer type 2 susceptibility protein
   PIRSF002397▸                                            DNA recombination repair protein, BRCA2 type
     PTHR11289▸                                            BREAST CANCER TYPE 2 SUSCEPTIBILITY PROTEIN BRC

noIPR            unintegrated
 PTHR11289:SF0▸                                            SUBFAMILY NOT NAMED
     SSF81878▸                                             BRCA2 tower domain
```

Figure 3. InterProScan output for the human BRCA2 protein.

n. Are the results of your InterProScan congruent with those of the CDD search?

Lab Quiz Question 3

End of Lab!

Lab 1 Objectives

By the end of Lab 1 (comprising the labs including their boxes, and the lectures), you should:

- know why we are interested in searching for motifs and profiles in sequences;
- know the advantages and disadvantages of representing structural elements in protein sequences as motifs or profiles;
- be able to generate a motif given an alignment;
- understand how to score a given sequence with a PSSM or profile HMM;
- be able to use CDD, CDART, SMART, Pfam, and InterProScan to identify specific functional units within protein sequences;

Do not hestitate to ask the instructor or TA if you do not understand any of the above after reading the relevant material.

Further Reading

Section 6.1 “Profiles and Sequence Logos” in Chapter 6 “Patterns, Profiles, and Multiple Alignments” in *Understanding Bioinformatics* by Marketa Zvelebil and Jeremy Baum, Garland Science, 2008. pp 167-178.

Section 6.2 “Profile Hidden Markov Models” in Chapter 6 “Patterns, Profiles, and Multiple Alignments” in *Understanding Bioinformatics* by Marketa Zvelebil and Jeremy Baum, Garland Science, 2008. pp 179-192.

Sonnhammer ELL, Eddy SR, Durbin R (1997). Pfam: A Comprehensive Database of Protein Families Based on Seed Alignments. Proteins, 28:405-420.

Eddy SR (1998). Profile Hidden Markov Models. Bioinformatics, 14:755-763.

Appendix 1: Common Unix commands

`ls`	list directory contents
`ls -x`	sort by extension
`ls -t`	sort by time stamp
`cd`	change directory
`cd /home/student_name/docs`	absolute path to directory “docs”
`cd docs/`	
`cp`	copy files and/or directories
`cp myfile.txt myfile2.txt`	
`clear`	clear screen
`mkdir`	make new directory
`mkdir downloads/`	
`rm`	remove file or directory
`rm test`	
`rm -r directory_name`	
`head`	output the first lines of files (by default 10 lines)
`head -n 200 myfile.txt`	write the first 200 lines of myfile.txt
`tail`	output the last part of files (by default 10 lines)
`tail -n 20 myfile.txt`	write the last 20 lines of the myfile.txt
`wc`	word count
`wc myletter.txt`	
`cat`	concatenate file
`cat file1.txt file2.txt > merged.txt`	write file1.txt then file2.txt to merged.txt
`chmod`	change file access permissions
`chmod a+x myfile.txt`	give everybody execute permission
`chmod g-w another.txt`	take away write permission (read-only)
`pwd`	display name of current/working directory
`man`	text manual page
`man ls`	show instructions for the ls program
`exit`	close the current shell window and logout

CTRL-C is like CTRL-ALT-DELETE on Windows – stops the current application in the shell