Protein domain annotation: application on genomic (and meta-genomic) datasets

Juliana Bernardes **Riccardo Vicedomini**

Laboratoire de Biologie Computationnelle et Quatitative
Université Pierre et Marie Curie

25 July, 2016

Outline

1. The Universal Protein Resource

2. The Pfam database and profile Hidden Markov Models

3. Building a custom profile Hidden Markov Model

4. Metagenomics

The Universal Protein Resource (UniProt)

UniProt is comprised of three components, where each one is optimized for different purposes:

1. **UniProt KnowledgeBase (UniProtKB)**
2. UniProt Archive (UniParc)
3. UniProt Reference Clusters (UniRef) databases

The UniProt KnowledgeBase

- The UniProtKB is the central hub for the collection of functional information on proteins, with accurate, consistent and rich annotation.

- As much annotation information as possible is added:
  - biological ontologies
  - classifications
  - cross-references
  - quality of annotation

- The UniProtKB consists of two sections:
  - **UniProtKB/Swiss-Prot**: man all manually-annotated records supported by literature and curator-evaluated computational analysis
  - **UniProtKB/TrEMBL**: computationally analyzed records that await full manual annotation

The UniProt Archive

- The UniProt Archive is a comprehensive and non-redundant database that contains most of the publicly available protein sequences in the world.
- Each unique sequence is stored once and has a stable identifier.
- Such an identifier is never removed, changed or reassigned.

The UniRef databases

- The UniRef databases provide clustered sets of sequences from UniProtKB and selected UniParc entries, in order to obtain complete coverage of sequence space at different resolutions while hiding redundant sequences.
- The UniRef100 database combines identical sequences and sub-fragments at least 11 residues into a single entry
- UniRef90 and UniRef50 are clusters of UniRef100 sequences which have at least 90% or 50% sequence identity, respectively.

UniProt: a simple exercise

1. Access the information about the gene sequence *lacZ* related to the organism *Escherichia coli* (strain K12).

2. What is its function?

3. What is its domain architecture?

4. Download the sequence in FASTA format.

The Pfam database

- One of the largest source of multiple alignments for thousands of protein domain families (16 712 in its current release).
- Tight relationship with HMMer: tool for building profile HMMs.

Pfam family organization

Each Pfam family (or Pfam-A entry) consists of:

- a curated **seed** alignment: a set of representative sequences
- a **profile hidden Markov model** built from the seed subset
- an automatically generated **full** alignment

Pfam's gathering threshold (GA)

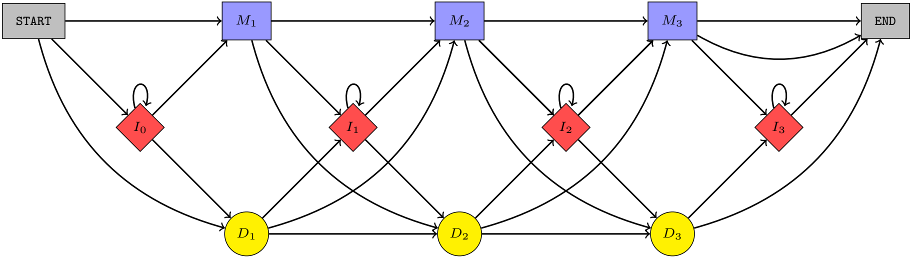
- It is associated to each Pfam domain
- All sequences that have a score are included in the full alignment
- GA values are manually curated and family-specific

Profile Hidden Markov Models

- Classification of protein domains from a multiple sequence alignment (MSA)
- A model which describes a pattern/motif is build according to a given MSA
- Profile HMMs have a formal probabilistic basis
- In contrast to other profile methods which use heuristics, they have a consistent theory behind gap and insertion scores.

Profile Hidden Markov Models

```
AG---C
AA-ACG
CG--A-
A-AG-C
AG---C
```



Build a profile HMM from a MSA

```
$ hmmbuild -h

Usage: hmmbuild [-options] <hmmfile_out> <msafile>

where basic options are:
  -h     : show brief help on version and usage
  -n <s> : name the HMM <s>
  -o <f> : direct summary output to file <f>, not stdout
  -O <f> : resave annotated, possibly modified MSA to file <f>
```

Build a profile HMM from a MSA

```
HMMER3/f [3.1b2 | February 2015]
NAME  zf-C2H2
ACC   PF00096.25
DESC  Zinc finger, C2H2 type
LENG  23
ALPH  amino
RF    no
MM    no
CONS  yes
CS    yes
MAP   yes
DATE  Sat Jul 22 16:51:08 2017
NSEQ  159
EFFN  159.000000
CKSUM 863016284
GA    25.20 15.20
TC    25.20 15.20
NC    25.10 15.10
STATS LOCAL MSV       -6.8107  0.73108
STATS LOCAL VITERBI   -7.1671  0.73108
STATS LOCAL FORWARD   -3.4176  0.73108
```

Build a profile HMM from a MSA

```
HMMER3/f [3.1b2 | February 2015]
NAME  zf-C2H2
ACC   PF00096.25
DESC  Zinc finger, C2H2 type
LENG  23
ALPH  amino
RF    no
MM    no
CONS  yes
CS    yes
MAP   yes
DATE  Sat Jul 22 16:51:08 2017
NSEQ  159
EFFN  159.000000
CKSUM 863016284
GA    25.20 15.20
TC    25.20 15.20
NC    25.10 15.10
STATS LOCAL MSV       -6.8107  0.73108
STATS LOCAL VITERBI   -7.1671  0.73108
STATS LOCAL FORWARD   -3.4176  0.73108
```

Search a Pfam domain in a sequence using HMMer: `hmmsearch`

```
Usage: hmmsearch [options] <hmmfile> <seqdb>

  -h               : show brief help on version and usage
  --domtblout <f> : save parseable table of per-domain hits to file <f>
  --domE <x>       : report domains <= this E-value threshold in output
  --cut_ga         : use profile's GA gathering cutoffs to set all
                     thresholding
  --cpu <n>        : number of parallel CPU workers to use for
                     multithreads
```

Example: `hmmsearch --domtblout output.domtlbout PF00096.hmm X6NX52.fasta`

Exercise: build/search a profile HMM

Consider the Pfam *Zinc finger* domain (zf-C2H2) `PF00096`:

1. Download the SEED alignment in Stockholm format from Pfam
2. Build a profile HMM from the retrieved file using HMMer with default parameters and save it as `PF00069.hmm`
3. Retrieve the sequence `X6NX52` from UniProtKB and save it as `X6NX52.fasta`
4. Search the model in `X6NX52_RETFI` using the command `hmmsearch` with default parameters.
    - How many hits do you find?
    - Compare the results with Pfam annotation.

Exercise: build/search a profile HMM

Consider now the Pfam *ubiquitin* domain `PF00240`:

1. Construct the HMMer model for this domain from the SEED alignment (as done in the previous exercise).
2. Retrieve all FULL alignment sequences of the same domain in FASTA format (without gaps) and save it as `PF00240_full.fasta`
3. Retrieve all the ubiquitin domain occurrences. How many domain hits do you observe? Compare the number of detected domains with the expected one. What do you notice?
4. Write a script which returns only those domain hits with e-value less than `1e-10`. How many domains do you observe now?
5. Write a script which reports the best (and the worst) hits according to the `e-value`

Comparison of different annotation methods

In this exercise we want to compare the performance of several methods to possibly identify the SH2 domain in a couple of human proteins which might contain it. More precisely, we will consider:

- BLAST
- PSI-BLAST with 2 iterations
- HMMer

The SH2 domain can be found with the Pfam accession number `PF00017`.

The SH2 domain

We will focus only on five specific proteins:

- `SRC_HUMAN` whose UniProt ID is `P12931`
- `SHC2_HUMAN` whose UniProt ID is `P98077`
- `STAT1_HUMAN` whose UniProt ID si `P42224`
- `SPT6H_HUMAN` whose UniProt ID is `Q7KZ85`
- `RINL_HUMAN` whose UniProt ID is `Q6ZS11`

The SH2 domain

1. **Retrieve the domain sequence**

Access to SRC_HUMAN in the UniProtKB database by using its accession number (P12931). The protein has an SH2 domain in the range 151-248. Extract this region and save it in a file called SRC_HUMAN_SH2.fasta

2. **Run BLAST**

Search the extracted sequence using BLAST against the UniProtKB database choosing human (taxid:9606) as organism. Then, download the output file choosing the Hit table(text) format. Using the grep command, verify whether the four proteins were detected by BLAST or not.

The SH2 domain

3. **PSI-BLAST with 3 iterations:** as done with BLAST, search the same sequence using PSI-BLAST but download the results after the second iteration. Verify whether the four proteins were detected.

4. **HMMer:** run `hmmsearch` command with the GA cutoff

5. Is there a method that performed better than the others? Why?

Building a *custom* profile HMM

- The HH-suite is a quite recent iterative method which allows to efficiently find homologous protein sequences.
- Compared to PSI-BLAST it has 50–100% higher sensitivity and generates more accurate alignments.

```
$ source $CONDA2/activate hhsuite-2.0.16
$ hhblits -d /db/off_biomaj/hhsuite/dbname
    -i sequence.fasta
    -o output.hhr
    -oa3m output_msa.a3m
    -ohhm output_model.hhm
    -M first -e 1e-8 -n 3

$ reformat.pl a3m a2m output_msa.a3m output_msa.a2m
$ hmmbuild model_name.hmm output_msa.a2m
```

Exercise: Pfam/HH-suite model comparison

We want to annotate the ZNRF2_PLAF7 sequence of *Plasmodium falciparum* (UniProt ID Q8I480). More precisely we want to seek the following two Pfam domains:

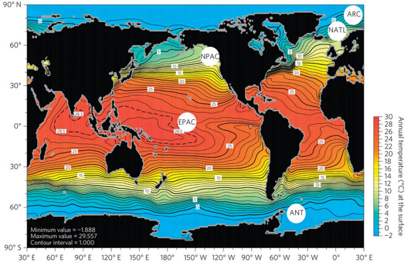
- VPS11_C, accession number PF12451
- Clathrin, accession number PF00637

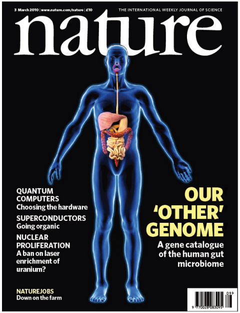
Consider then the UniProt sequence E2B3X3 where the aforementioned two domains can be found in the regions 265-413 and 681-719, respectively. Download these two *fragments* and create two HMMer models using the HH-suite (hhblits)

Search the Pfam models (always considering the GA threshold) and those two you have just built. Compare the results. What do you notice?

What is a microbiome?

The totality of microbes in a specific environment: their genomes and interactions with each other and the surrounding environment.




3 March 2010 | www.nature.com/nature | £10
THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE
nature
QUANTUM COMPUTERS
Choosing the hardware
SUPERCONDUCTORS
Going organic
NUCLEAR PROLIFERATION
A ban on laser enrichment of uranium?
NATUREJOBS
Down on the farm
OUR 'OTHER' GENOME
A gene catalogue of the human gut microbiome

Metagenomics

What is it?

The study of the entire genetic content of the microbiome of an environment of interest, using DNA sequencing techniques.

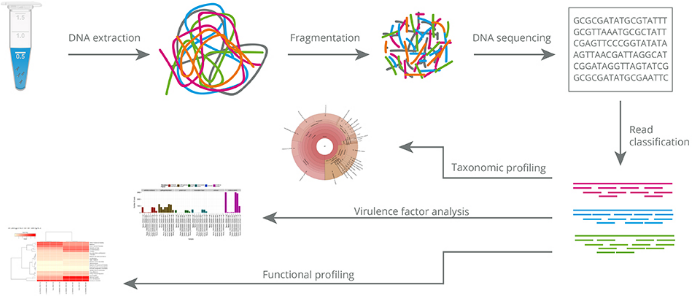
The advantages of metagenome sequencing

- studying environmental microbes without the need to culture them
- close estimations of microbial diversity
- information on composition and functional capabilities of an environment

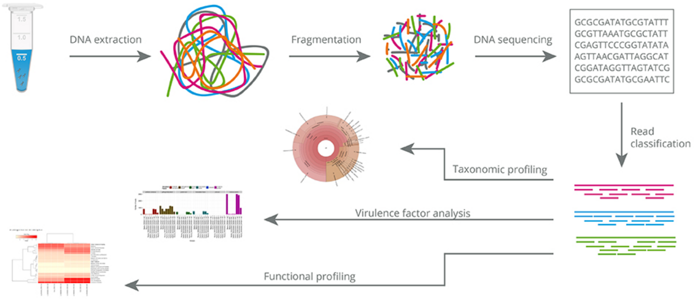
Challenges

- Data analysis can be hard.
- For medium/high diversity communities, assembly can easily produce chimeric sequences.

Metagenomic analysis workflow



Metagenomic analysis workflow



Functional profiling

Protein domain identification of metagenomic (MG) data allow to study and analyze which are the main functions expressed in a specific environment.

Functional profiling of MG/MT datasets

1. Read preprocessing:
    - quality trimming
    - paired-end reads merging
    - assembly?

2. Coding sequences prediction

3. Domain annotation (*e.g.,* with Pfam or other domain databases)

4. Functional/comparative analysis (*e.g.,* using GO terms)

Sequence manipulation exercises

Write a python function which performs the following tasks:

1. Reverse complement of a nucleotide sequence
2. Output all the six reading frames of a nucleotide sequence
   - You can find a template (`translateSequence.py`) at the following repository: `https://github.com/vice87/roscoff2017/`
3. Given a FASTA file, print all six reading frames for each read in input (the output should be in FASTA format as well)

Annotation and analysis of a MG dataset

Download the metatranscriptomic dataset `HumanGut_MG_pCDS_subset.faa.gz` you can find at `https://github.com/vice87/roscoff2017`

1. Annotate the set of translated reads / predicted CDS with the entire Pfam HMM library, available at the path `/db/pfam/current/flat/Pfam-A.hmm`
2. Write a python program which performs the following operations:
   a. considers only the best hit for each read
   b. returns the 10 most abundant domains (along with the number of occurrences)