Why do we need to develop skills?

Science moves fast

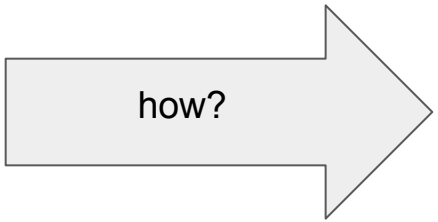
- New information and resources
- Better algorithms
- No more support
- New area

It is important to know how to choose tools and methodologies

Determine the problem

Many functional annotation tools require that the input data are protein sequences, and some tools, which can accept either nucleotide or protein sequence, show superior results when protein sequences are submitted. (Bolger, 2018)

RNA sequence
format: fasta (n)



Protein sequence
format: fasta (aa)

Additional factors: frame, codon table, reverse (EMBOSS Transeq)

How to choose a tool or methodology?

0. Determine the problem

1. Search
2. List of candidates
3. Try and verify
4. Document

This is an iterative process

Search

The first is not always the best

Ideally we want to find a list of candidates with a comparison

- Competition
- Review
- Database of resources
- Ask an expert

Integrated tools for the functional analysis of plant genomes

Resource	Time taken	Annotation rate (%)	Comments
Reference	—	51	At least one GO term assigned including cellular component
Blast2GO	8 h 23 min	78	BLAST is performed locally or as WebBLAST via NCBI; InterProScan is performed as a Web service at the European Bioinformatics Institute (EBI)
KAAS	10 min (only single- directional best hit (SBH) was used as a survey sample of sequence)	29	Runs as a Web service, no user resources needed
GhostKOALA	28 min	26	Runs as a Web service, no user resources needed
Mercator	5 min	56	Runs as a Web service, no user resources needed
TRAPID	5 min	56	Runs as a Web service, no user resources needed

Note. For the analysis, the first 1476 proteins from the Brassica proteome version 5 were downloaded from http://www.genoscope.cns.fr/brassicanapus/data/ alongside their GO annotations, representing exactly 10 000 lines of text and submitted to the various services, where available searches were limited to plant data sets. In the case of Blast2GO, WebBLAST was used. We have rounded the values, as annotations are subjected to updates, and time taken will depend on server loads. Therefore, these values should be seen as a general orientation.

Marie E Bolger, Borjana Arsova, Björn Usadel; Plant genome and transcriptome annotations: from misconceptions to simple solutions, Briefings in Bioinformatics, Volume 19, Issue 3, 1 May 2018, Pages 437–449, https://doi.org/10.1093/bib/bbw135

Search

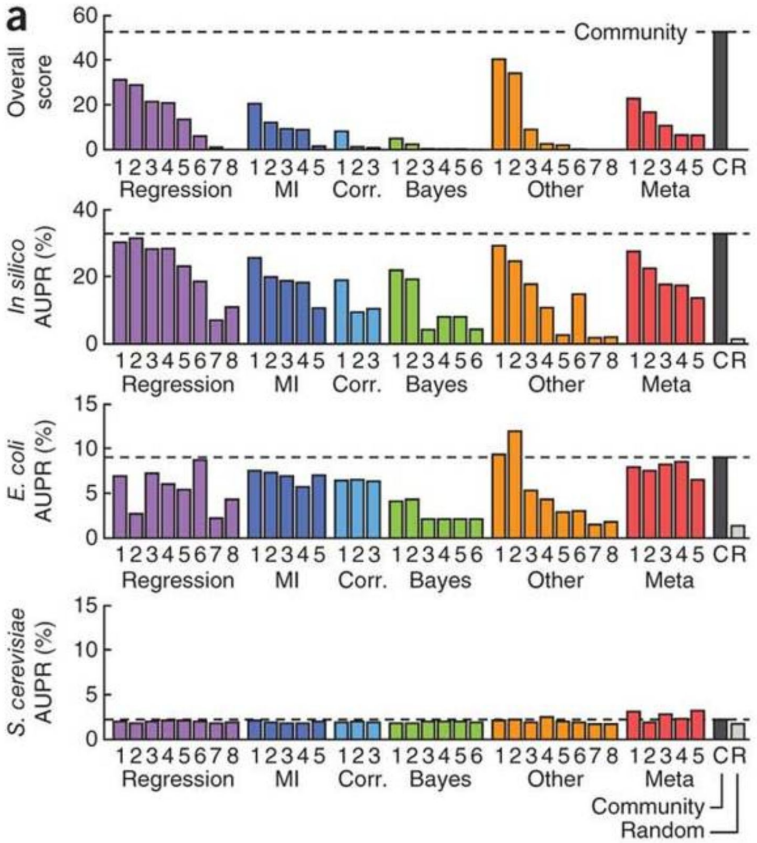
- Google
  - Use search terms https://support.google.com/websearch/answer/2466433?hl=en
- Academic resources
  - google scholar, research gate, biorxiv, NCBI, EBI, databases
- Community
  - stack overflow, ubuntu forums, R-help, local expert
  - Every forum has a format for asking questions https://stackoverflow.com/help/how-to-ask

List of candidates

- All methods depend on the data
- Wisdom of crowds

- Ease of use (good documentation)
- Technical considerations (resources, time, algorithm)
- Biological considerations (organism, question)
- How recent or well maintained

From: Wisdom of crowds for robust gene network



Try and verify

- Use an example you know the answer
  - From your own data (sample and calculate)
  - Previous work (tutorials, articles, etc)
- Compare two (or more) tools
- Check edge cases

What would be a good example in this case?

Exercise

- Get fasta amino acid sequences of example data
- Teams of 3
- Document
  - Program used (links to documentation)
  - Parameters and options used
  - Verification
  - Other relevant details