Sequence Similarity

Mark Voorhies

3/26/2012

Logistics

Friday (3/30/2012) is a UC holiday.
$\rightarrow$ Monday 4/2/2012?

Outline

Resources

Router:

- SSID: BMS270
- password: deoxyribose

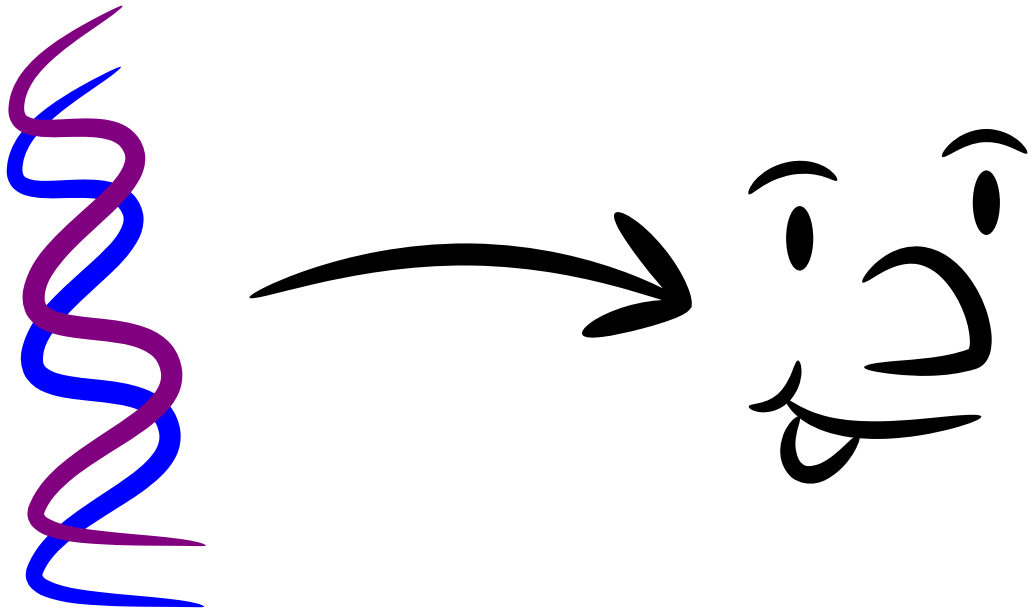
Course website:

- http://histo.ucsf.edu/BMS270/

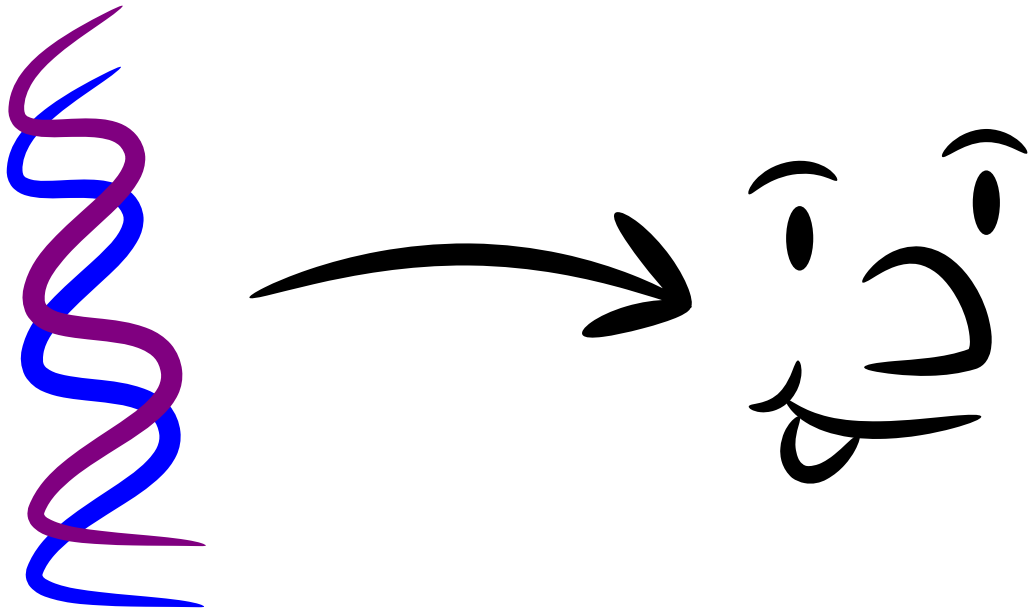
Resources on the course website:

- Syllabus
  - Papers and data sets (for downloading *before* class)
  - Slides (available *during* class)
- On-line textbooks (Safari Bookshelf, the BLAST book, ...)
- Programs for this course (DOTTER, CLUSTALX, JalView, ...)

Course outline

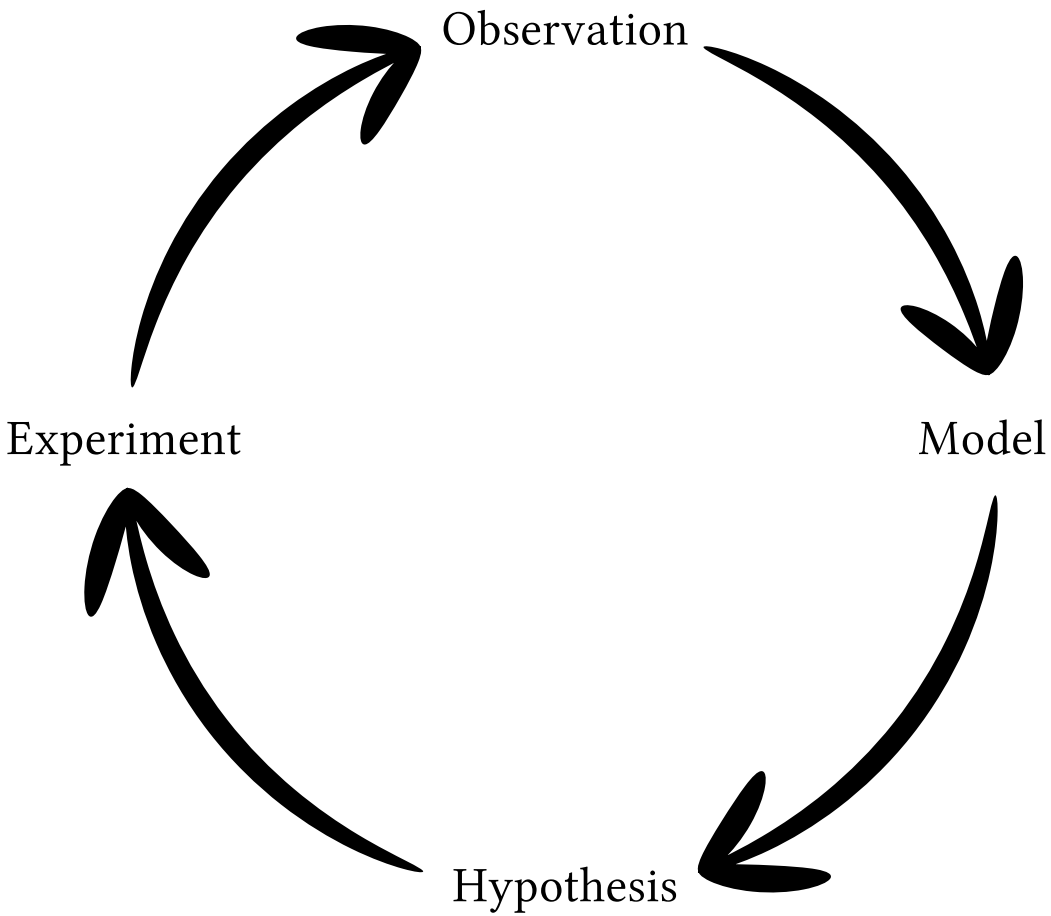


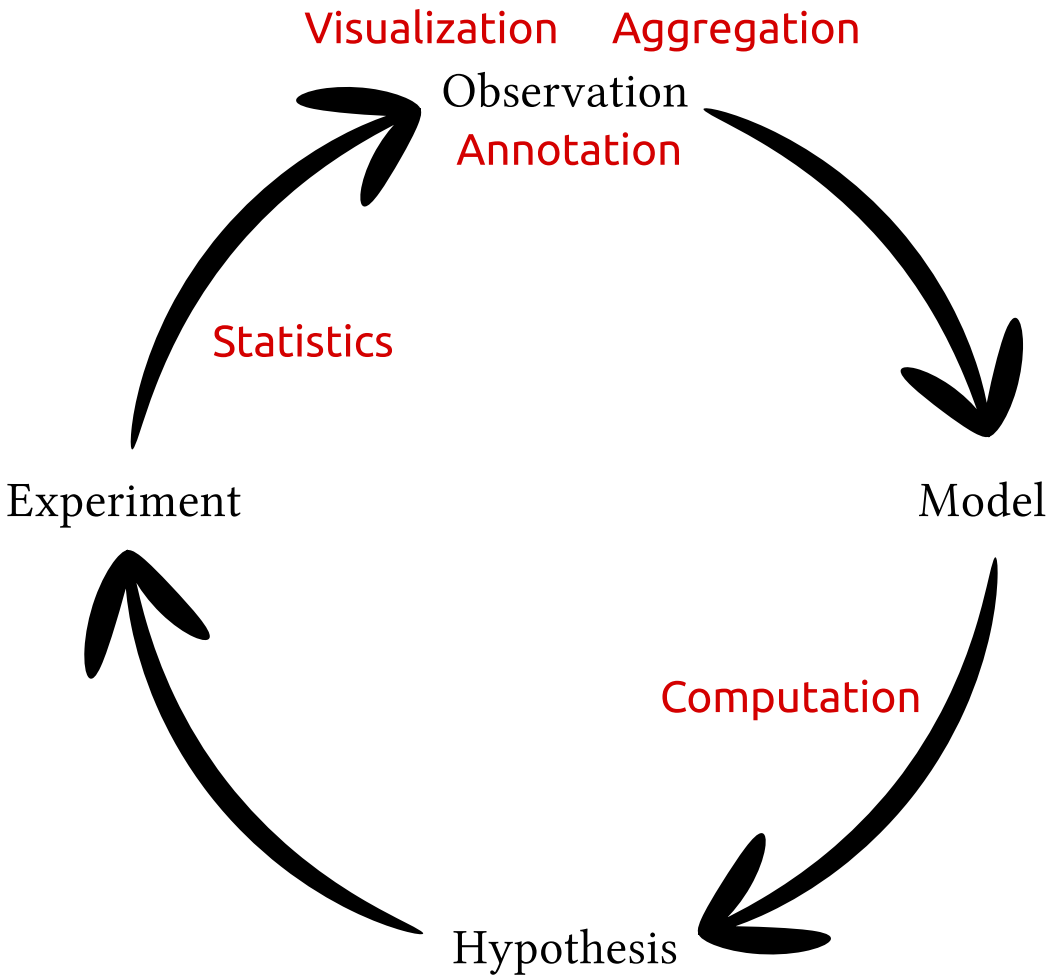
Course outline



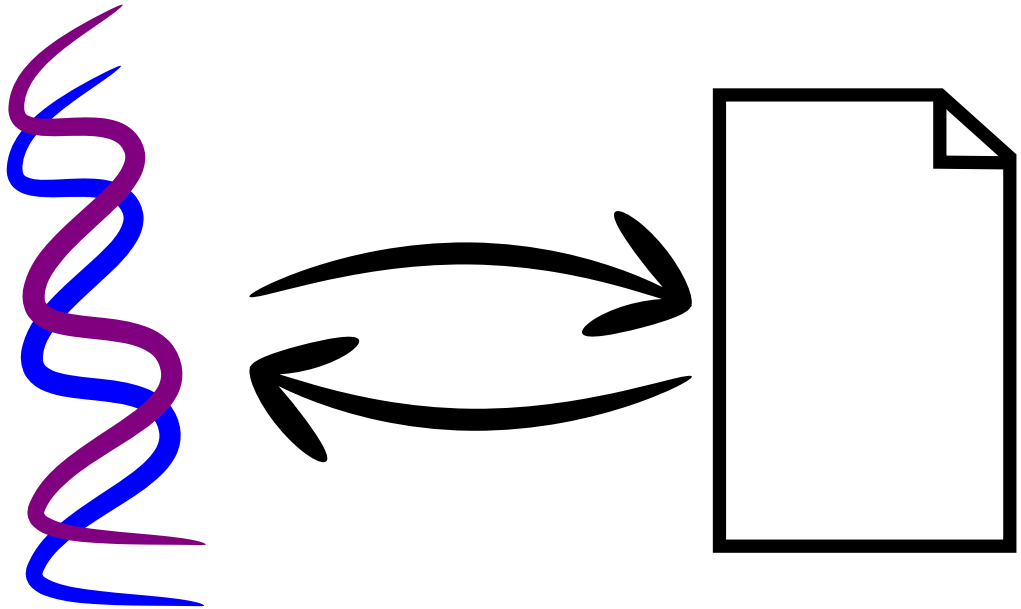
Week 1: Genotype
(Sequence analysis)

Week 2: Phenotype
(Expression profiling)

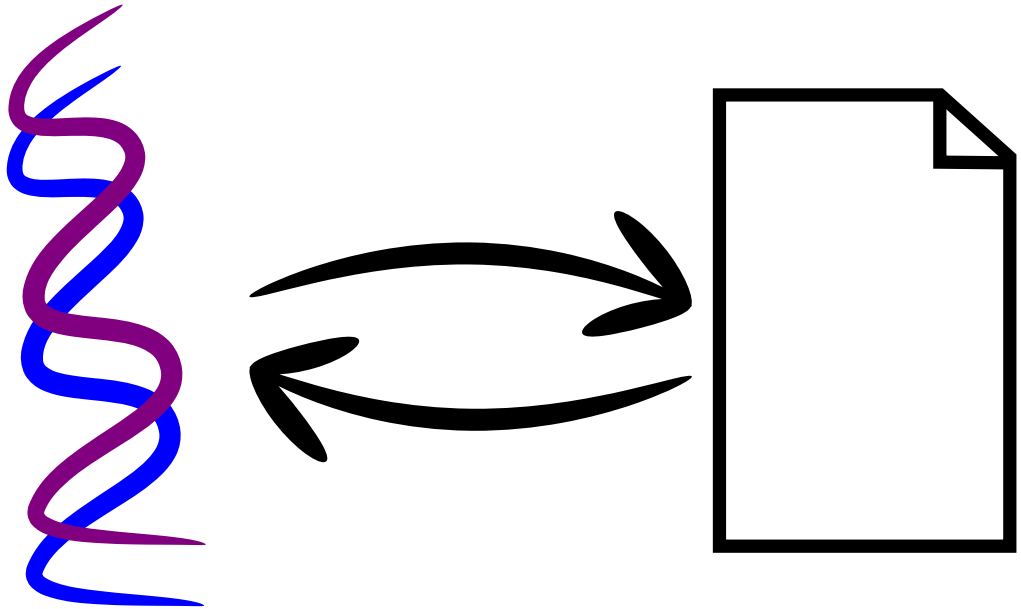

Observation
Model
Hypothesis
Experiment


Visualization
Aggregation
Observation
Annotation
Statistics
Experiment
Model
Computation
Hypothesis

Every object should have an isomorphism to a file

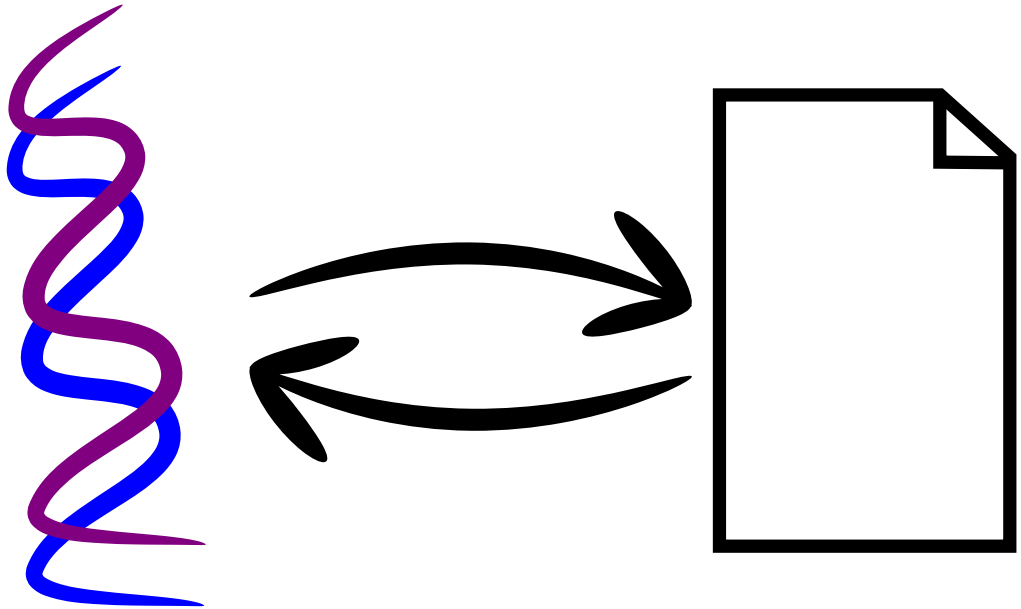


Every object should have an isomorphism to a file



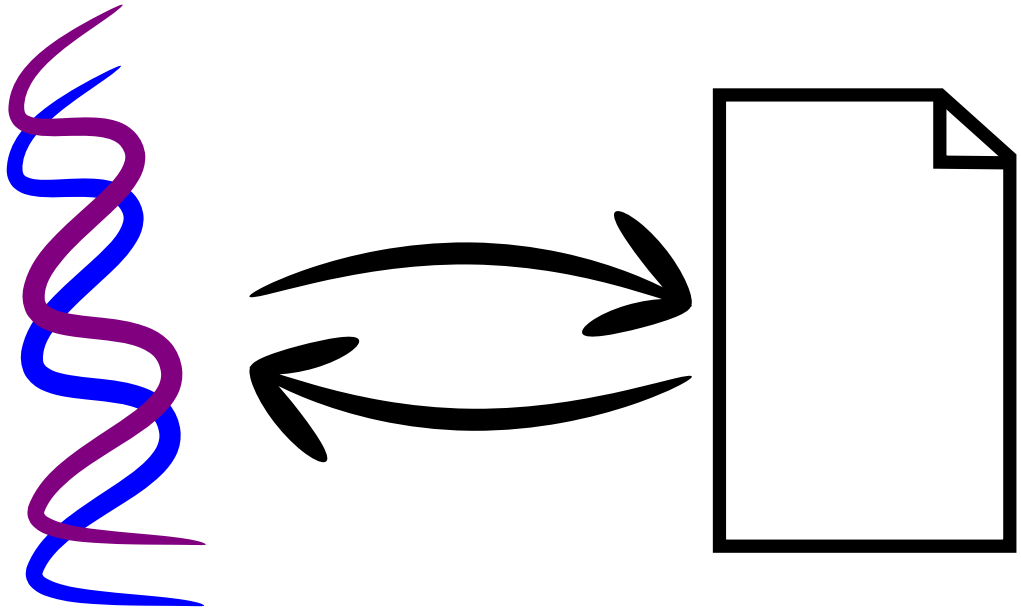
- Export, audit, edit, and import *independent* of a given program.

Every object should have an isomorphism to a file



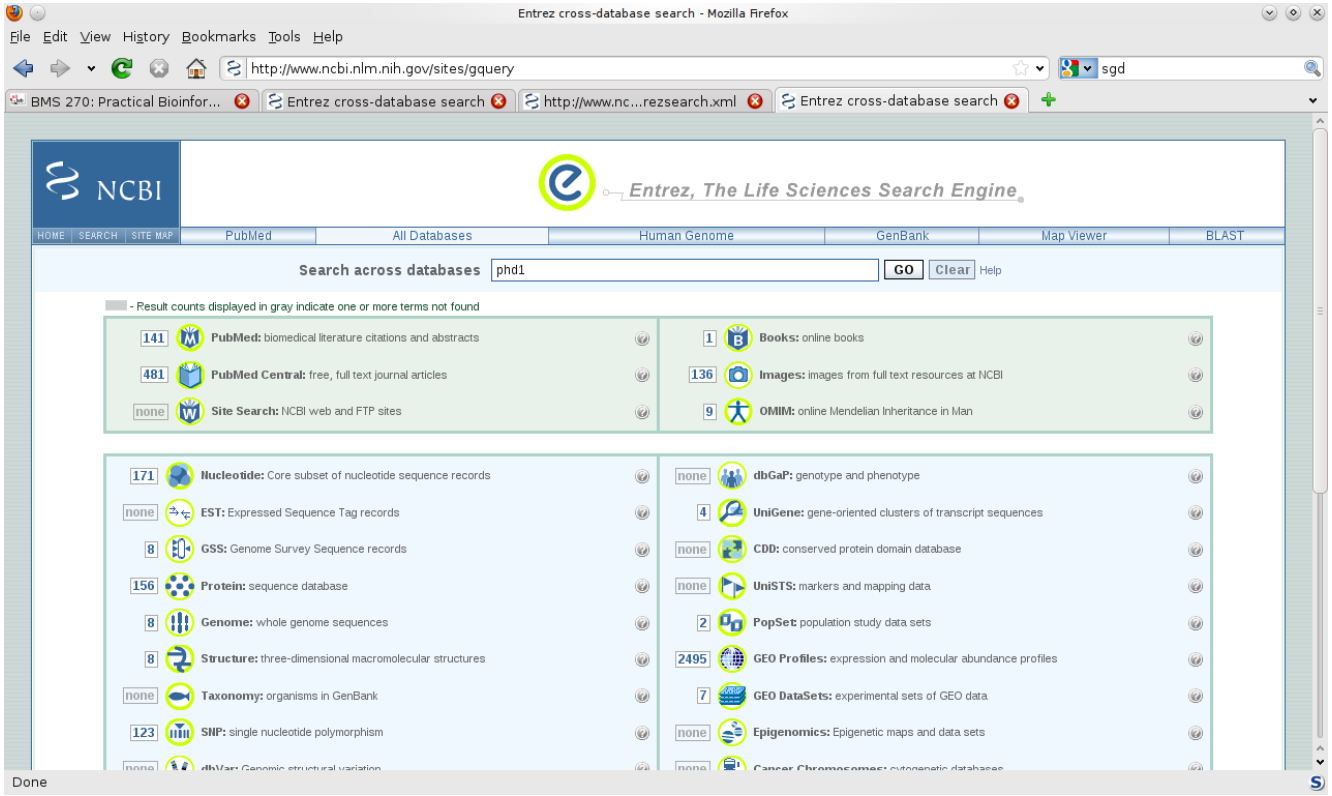
- Export, audit, edit, and import *independent* of a given program.
- Standard file formats for portability.

Every object should have an isomorphism to a file

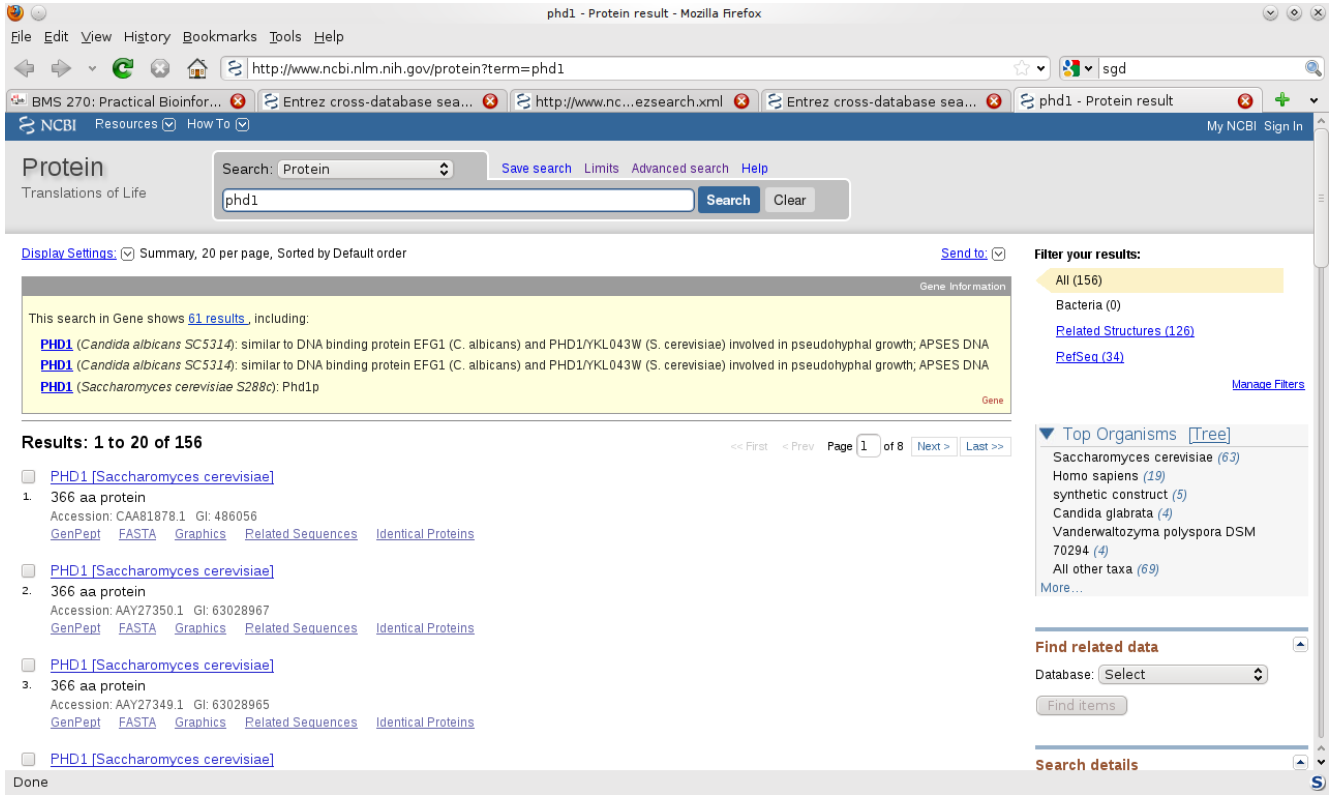


- Export, audit, edit, and import *independent* of a given program.
- Standard file formats for portability.
- Human readable text formats for audit and longevity.

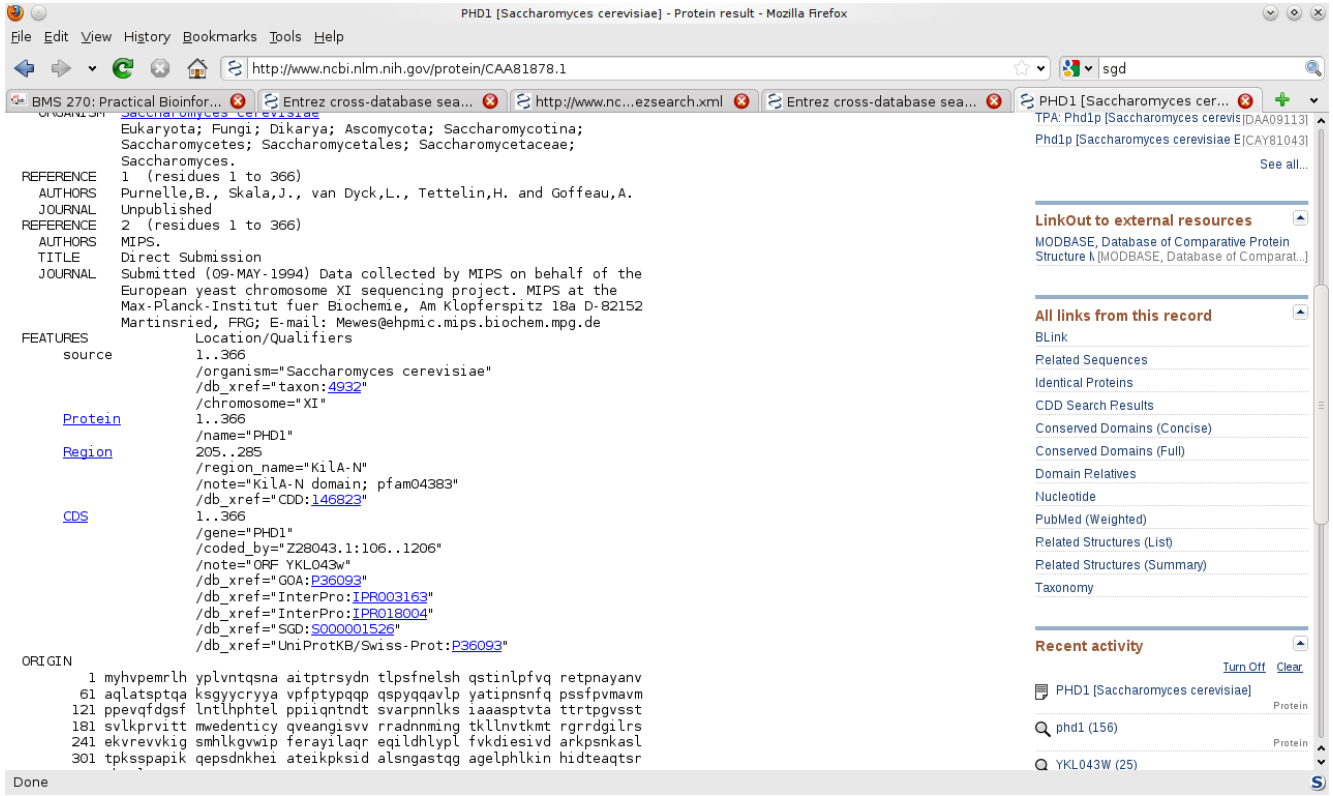
Entrez: Cross-Database Search



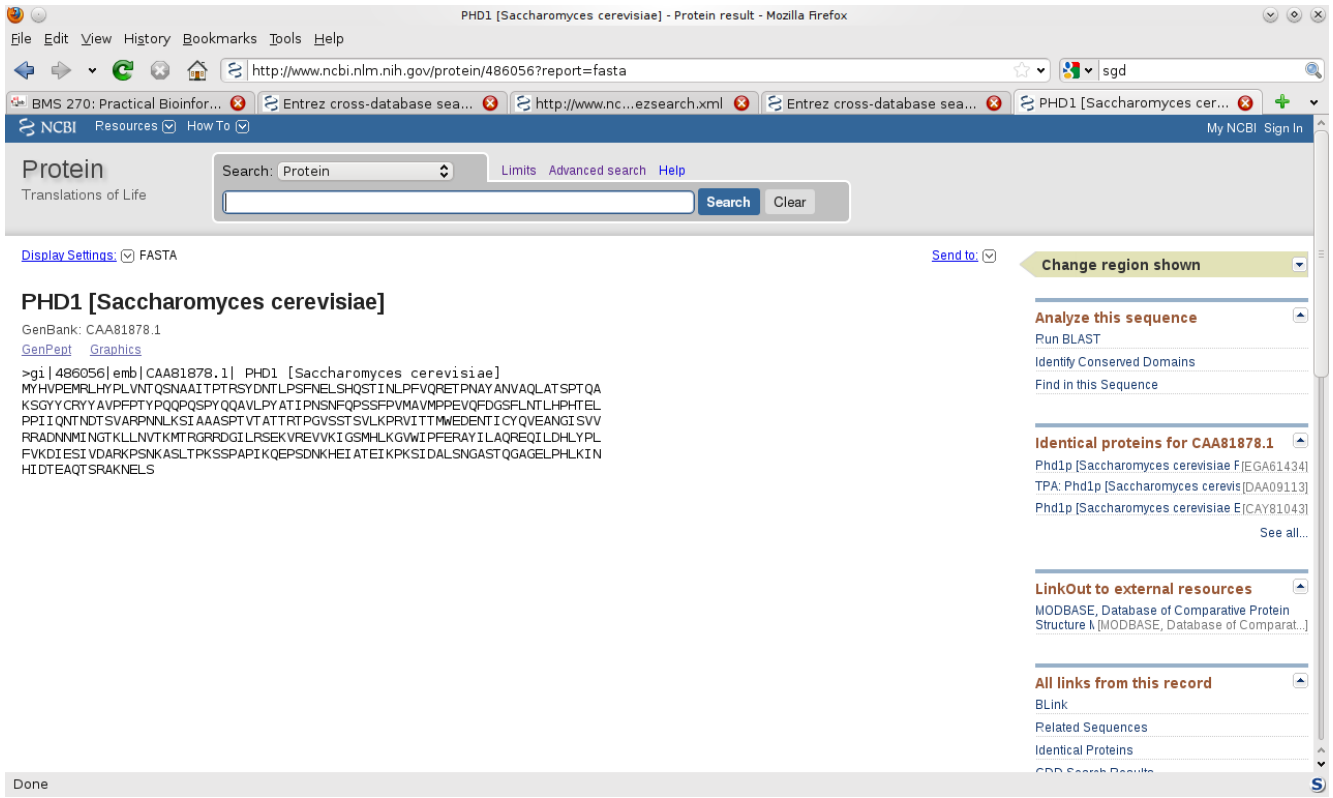
Entrez: Single Database



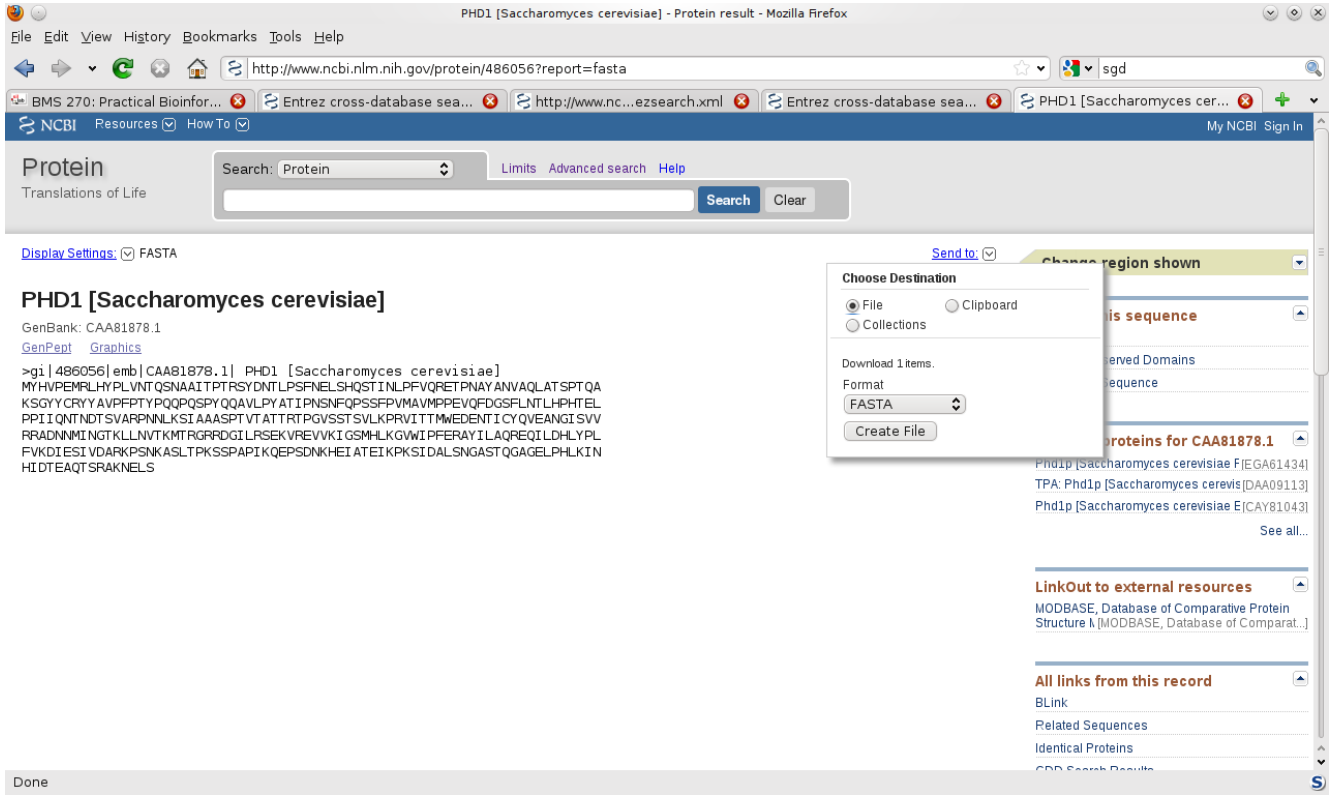
Entrez: GenPept (Feature Table) Format



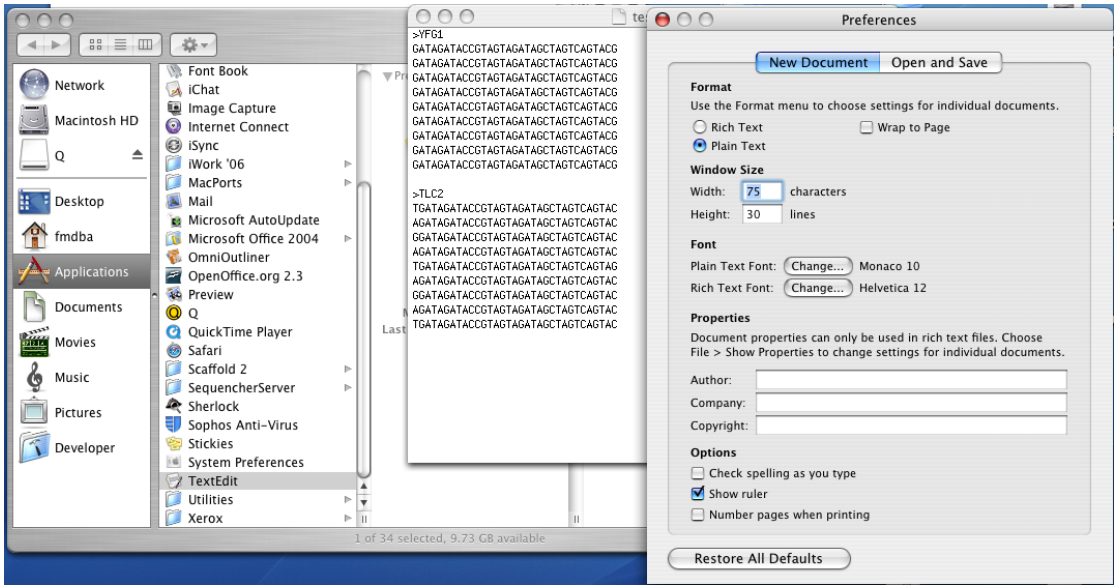
Entrez: FASTA Format



Entrez: Downloading Files



Configuring TextEdit for text files



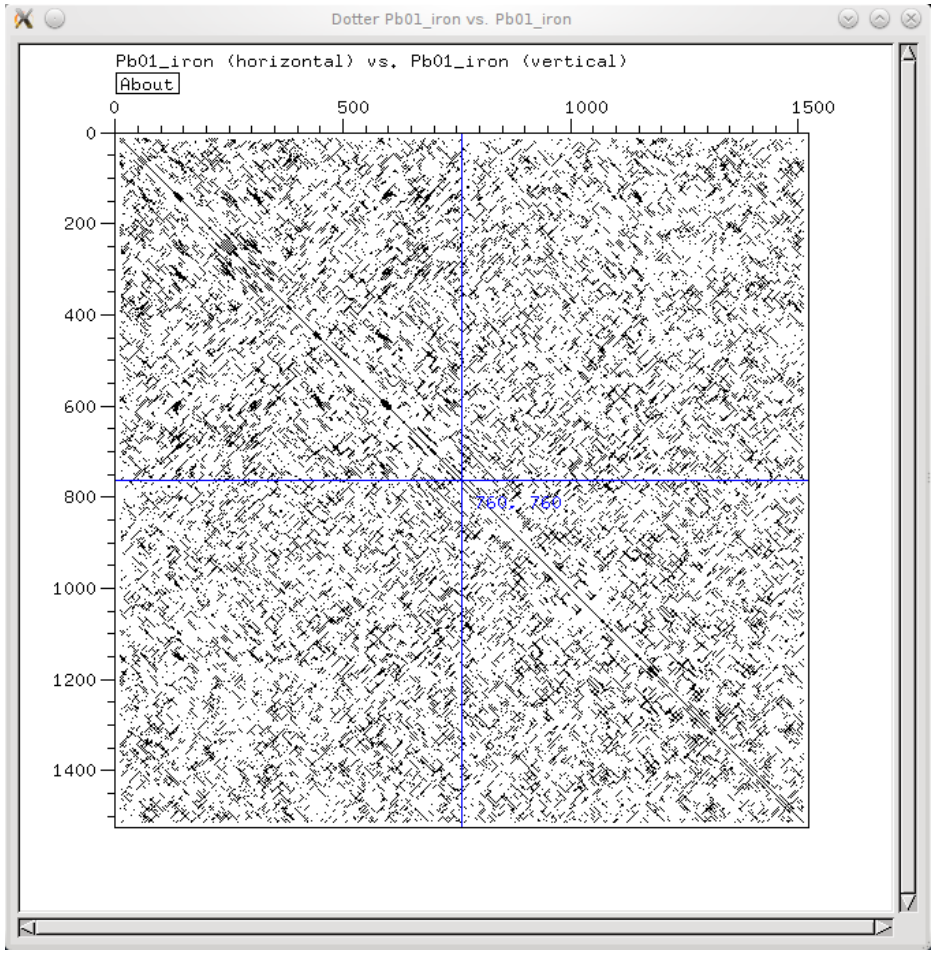
- Format → Plain Text
- Uncheck “check spelling as you type”
- In “Open and Save” uncheck “.txt extension”

Text File Tips

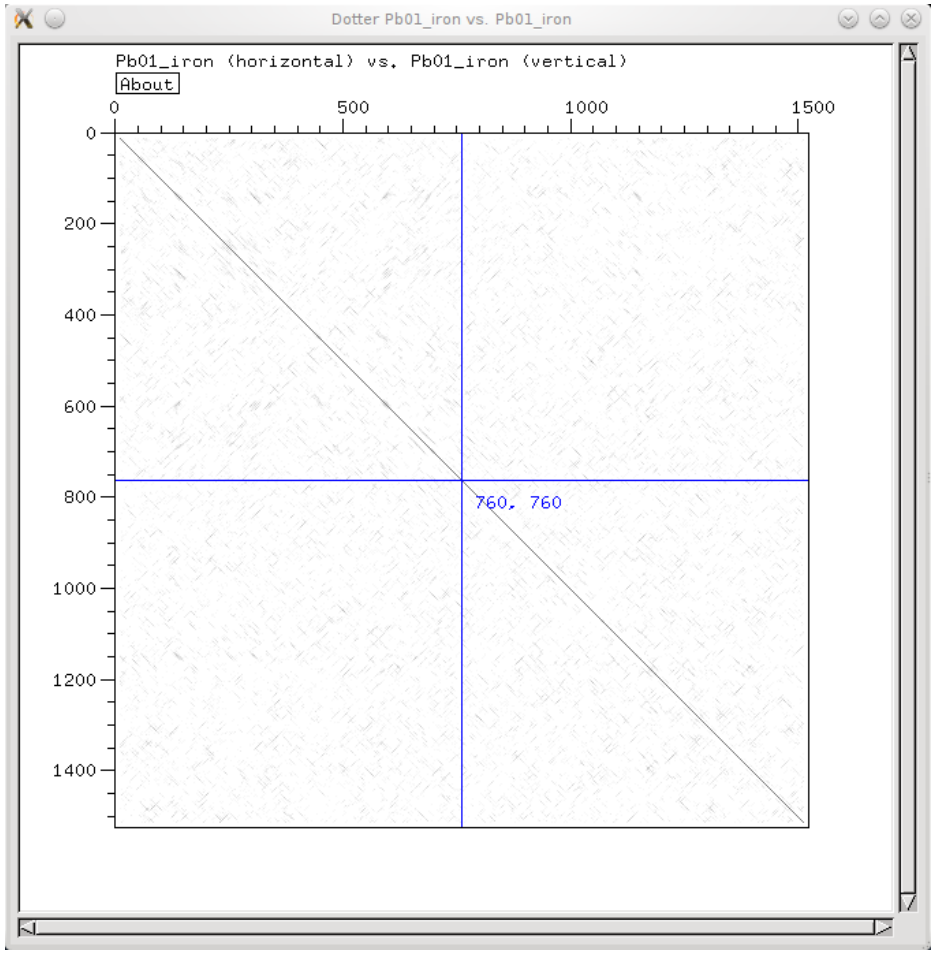
- Line terminators:
  - Unix/Linux: \n (linefeed)
  - MacOS: \r (carriage return)
  - DOS/Windows: \r \n (CRLF)
- Converting from MacOS to Unix on OS X:

```
tr '\r' '\n' < macfile.txt > unixfile.txt
```

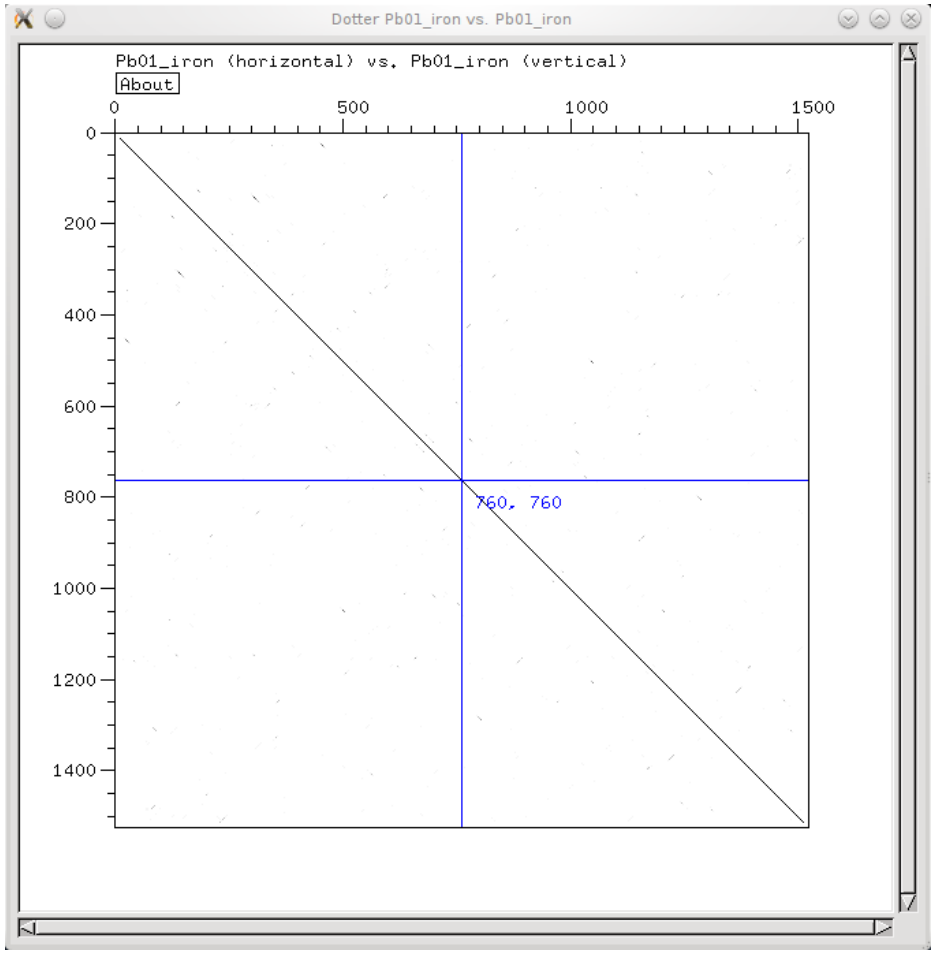
Dotplots: Dot = identity



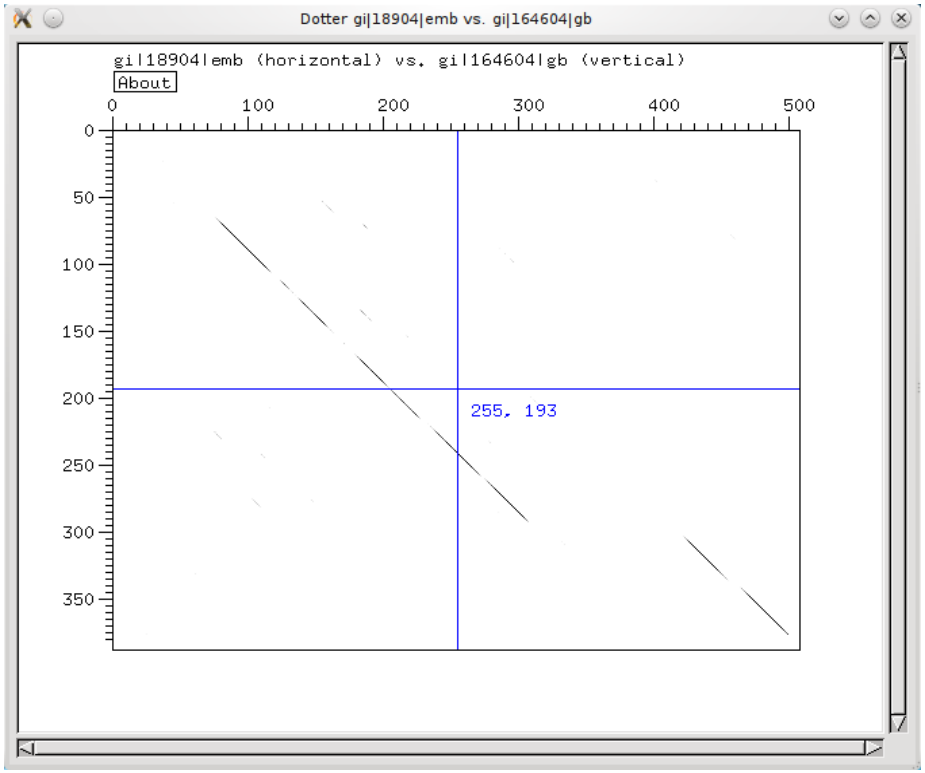
DOTTER: Windowed similarity scores



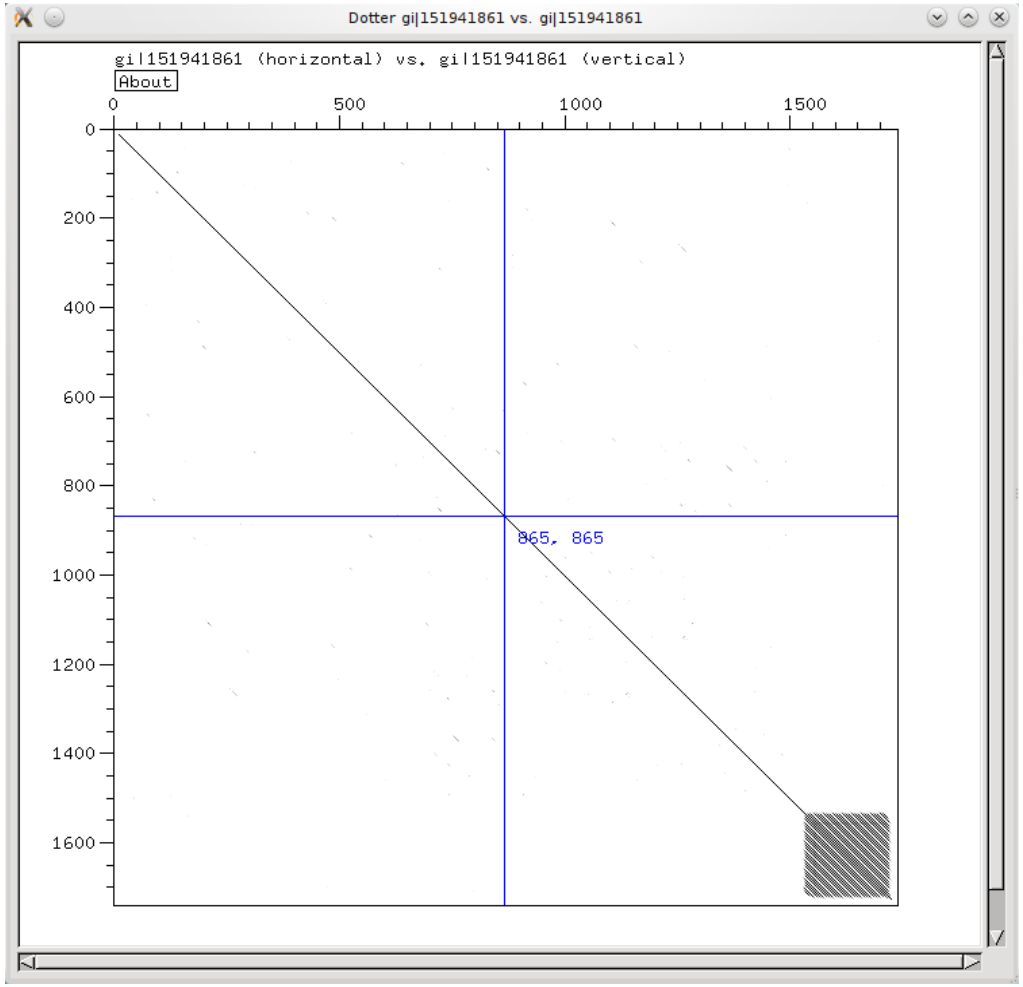
DOTTER: Windowed similarity scores with cutoff



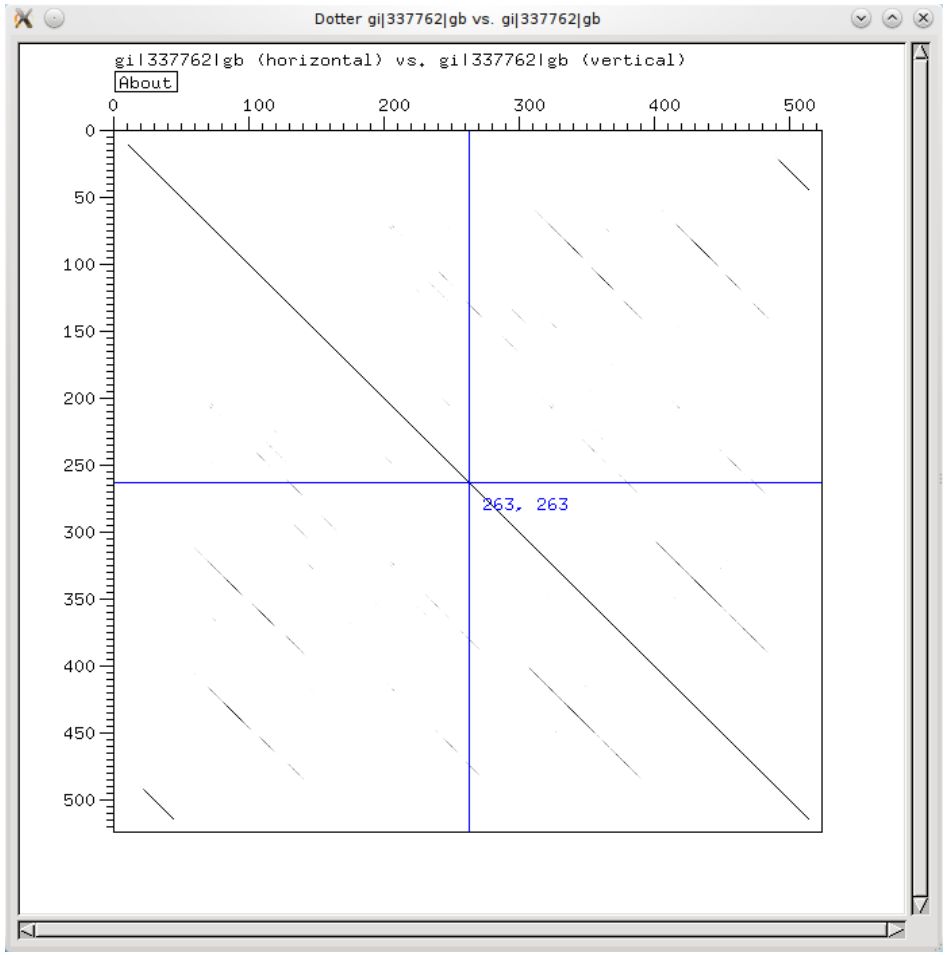
Phytepsin (barley) vs. Pepsinogen (pig)



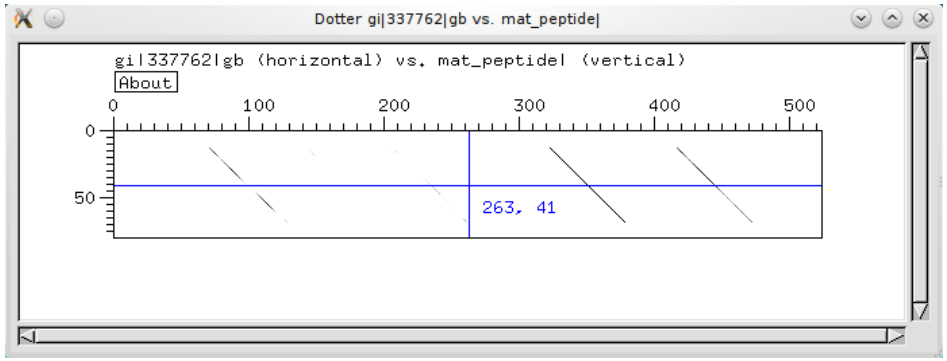
RNA Pol2 (core subunit)



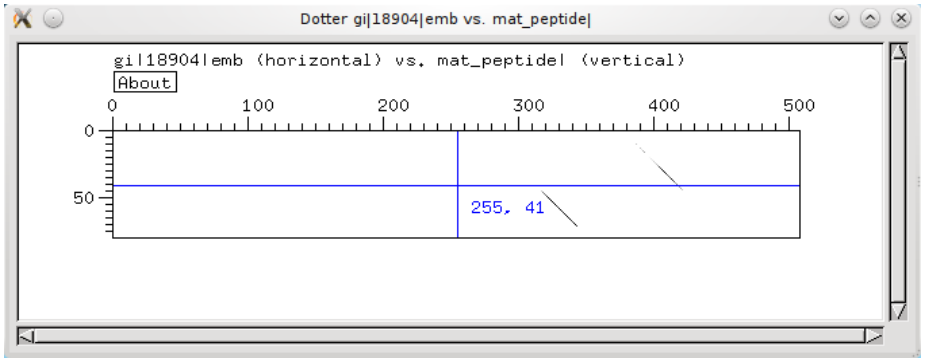
Prosaposin (human)



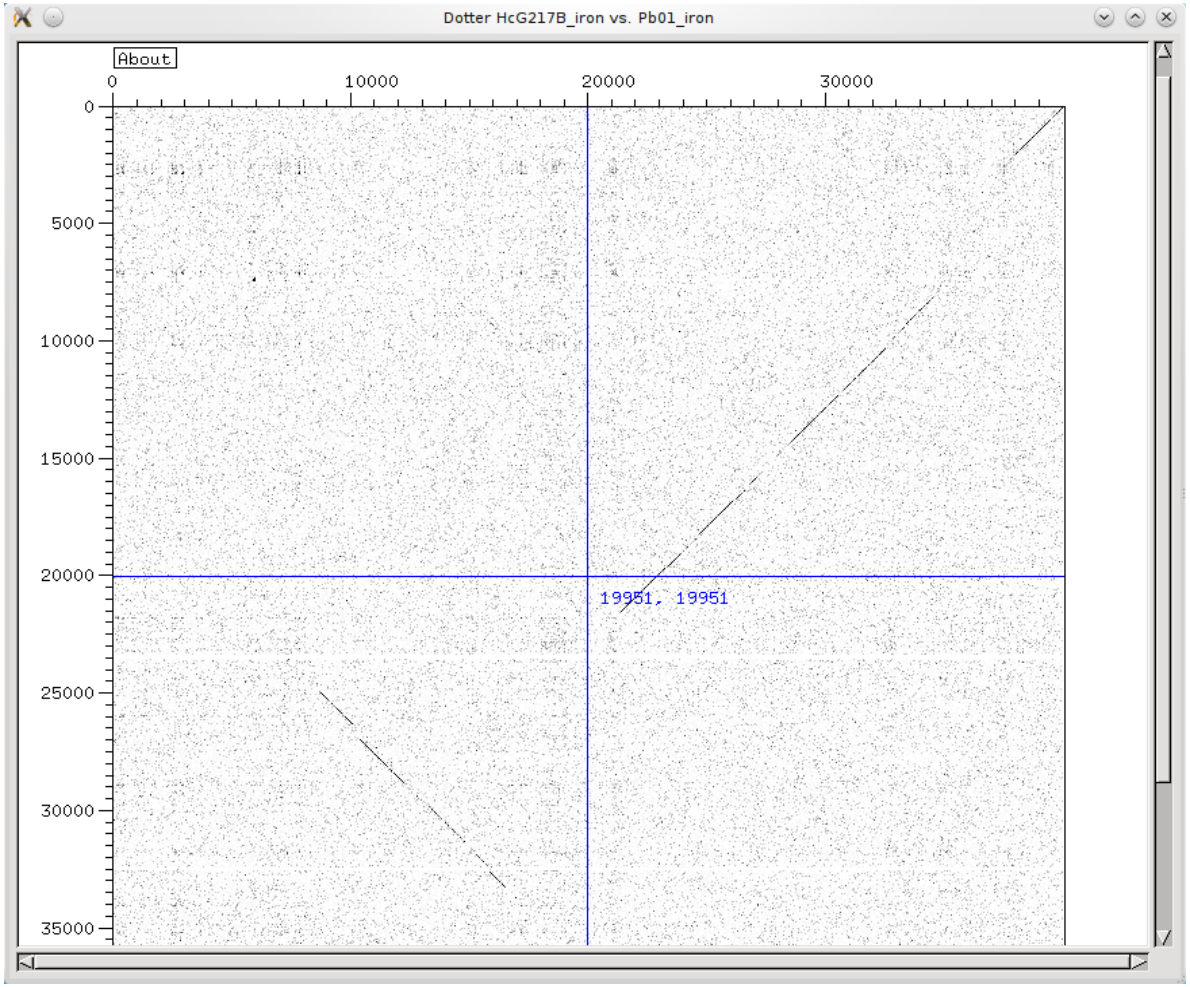
Prosaposin vs. Saposin C



Phytepsin (barley) vs. Saposin C (human)



Iron response locus, Histoplasma vs. Paracoccidiodes



Summary

- Every object has an isomorphism to a file. Don't be afraid to look inside and hack on *your* data files.
- Standard file formats allow different programs to work together.
- Analysis should start from an unbiased visualization of primary data.
- Dotplots provide a good first impression of the similarity between two sequences (or one sequence with itself) and are useful for debugging tricky sequence alignments.

Homework

- Play with some of your favorite sequences in DOTTER
  - Start from sequences with known insert/delete/repeat patterns and see if you can recapitulate them in DOTTER
  - See what you can infer in an unannotated sequence or pair of sequences
- Read Chapter 4 of the BLAST book (Sequence Similarity).
- Download CLUSTALX and JALVIEW for tomorrow