

Sequence Similarity

Mark Voorhies

3/26/2012

Logistics

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

Outline

- 1 Course Overview
- 2 Sequence File Formats
- 3 Dotplots

Resources

Router:

- SSID: BMS270
- password: deoxyribose

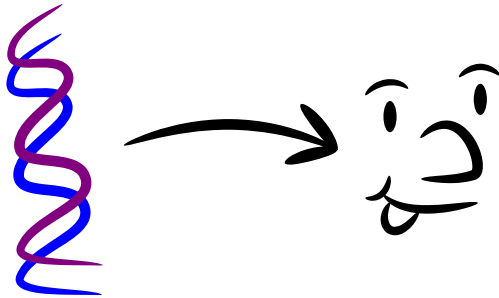
Course website:

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

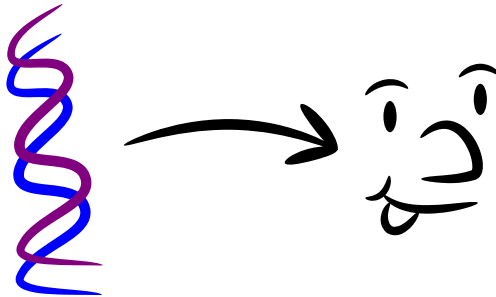
Resources on the course website:

- 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

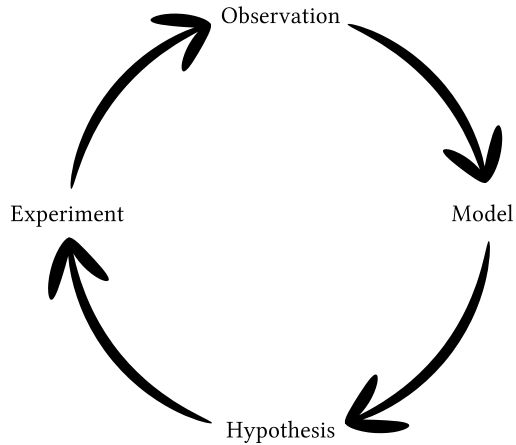


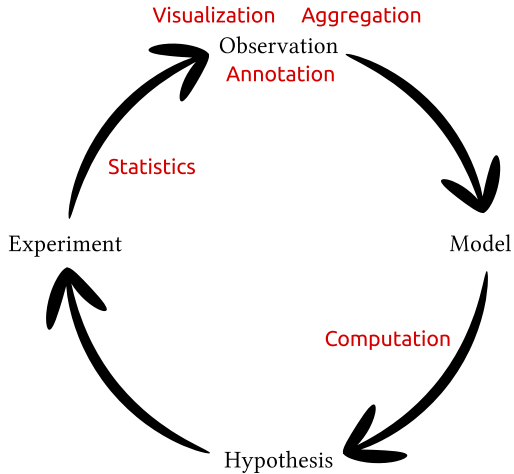
Course outline



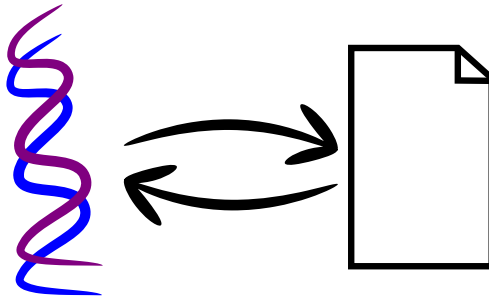
Week 1: Genotype
(Sequence analysis)

Week 2: Phenotype
(Expression profiling)

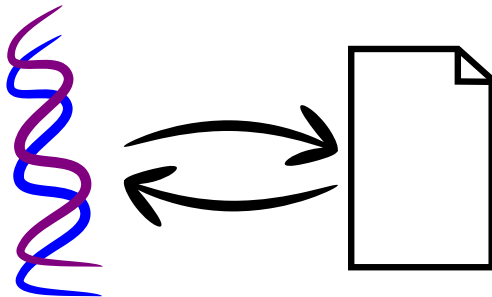




Every object should have an isomorphism to a file

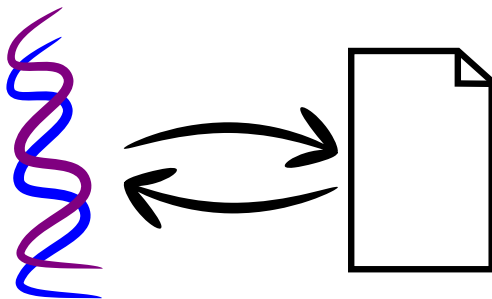


Every object should have an isomorphism to a file



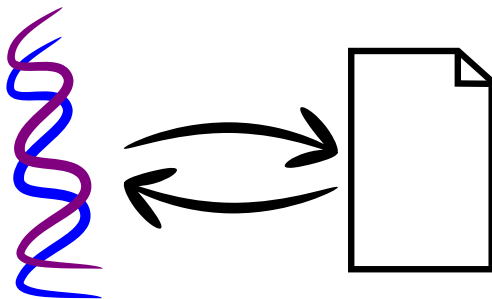
- Export, audit, edit, and import *independent* of a given program.

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

The screenshot shows a web browser window titled "Entrez cross-database search - Mozilla Firefox" with the URL "http://www.ncbi.nlm.nih.gov/sites/eqquery". The search results are displayed on the NCBI Entrez website. The search query is "phtl".

Search across databases: [help](#)

Result counts displayed in gray indicate one or more terms not found

Count	Database	Description
141	PubMed	PubMed: biomedical literature citations and abstracts
481	PubMed Central	PubMed Central: free, full-text journal articles
10000	Site Search	Site Search: NCBI web and FTP sites
171	Nucleotide	Nucleotide: Core subset of nucleotide sequence records
10000	EST	EST: Expressed Sequence Tag records
8	GSS	GSS: Genome Survey Sequence records
156	Protein	Protein: protein sequence database
8	Genome	Genome: whole genome sequences
8	Structure	Structure: three-dimensional macromolecular structures
10000	Taxonomy	Taxonomy: organisms in GenBank
123	SNP	SNP: single nucleotide polymorphism
10000	3D	3D: 3D Protein Structure Structural Variations
1	Books	Books: online books
136	Images	Images: images from full-text resources at NCBI
0	OMIM	OMIM: online Mendelian Inheritance in Man
10000	IBGAP	IBGAP: genotype and phenotype
4	lncRNA	lncRNA: gene-oriented clusters of transcript sequences
10000	CDD	CDD: conserved protein domain database
10000	dbSTS	dbSTS: markers and mapping data
2	PopSet	PopSet: population study data sets
2495	CEO Profiles	CEO Profiles: expression and molecular abundance profiles
7	CEO DataSets	CEO DataSets: experimental sets of CEO data
10000	Epigenetics	Epigenetics: Epigenetic maps and data sets
10000	Protein	Protein: Proteome/Proteomics/Proteomic Database

Entrez: Single Database

The screenshot shows a web browser window displaying the NCBI Entrez Protein search results for the query 'phd1'. The browser's address bar shows the URL 'http://www.ncbi.nlm.nih.gov/protein/?term=phd1'. The search results are displayed in a table with columns for 'Gene information' and 'Accession'. The first three results are listed below:

Accession	Gene information
1. 366 aa protein Accession: CA81878.1 GI: 486056 GenBank FASTA Graphics Related Sequences Identical Proteins	PHD1 [<i>Saccharomyces cerevisiae</i>]
2. 366 aa protein Accession: AY27350.1 GI: 8302997 GenBank FASTA Graphics Related Sequences Identical Proteins	PHD1 [<i>Saccharomyces cerevisiae</i>]
3. 366 aa protein Accession: AY27349.1 GI: 8302995 GenBank FASTA Graphics Related Sequences Identical Proteins	PHD1 [<i>Saccharomyces cerevisiae</i>]

Additional features visible in the screenshot include a search bar at the top with the query 'phd1', a 'Filter your results' sidebar on the right showing 'All (156)' results, and a 'Top Organisms' list including *Saccharomyces cerevisiae* (9/2), *Homo sapiens* (2/8), *synthetic construct* (1), *Candida glabrata* (4), *Vanderwalleya polyspora* DSM 70294 (4), and *All other taxa* (9/3).

Entrez: GenPept (Feature Table) Format

PHD1 [Saccharomyces cerevisiae] - Protein result - Mozilla Firefox

http://www.ncbi.nlm.nih.gov/protein/CAAB1878.1

Entrez cross-database search

BMS 270: Practical Bioinformatics

Entrez cross-database search

http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=protein

Entrez cross-database search

PHD1 [Saccharomyces cerevisiae] [CAAB1878.1]

Phd1p [Saccharomyces cerevisiae] [CAAB1878.1]

See all...

LinkOut to external resources

MDBASE, Database of Comparative Protein Structure [MDBASE, Database of Comparat...]

All links from this record

BLINK

Related Sequences

Identical Proteins

CDD Search Results

Conserved Domains (ConStr)

Conserved Domains (Full)

Domain Relatives

Nucleotide

Pubmed (Weighted)

Related Structures (List)

Related Structures (Summary)

Taxonomy

Recent activity

PHD1 [Saccharomyces cerevisiae]

phd1 (156)

YKL043W (25)

Done

REFERENCE

1 (residues 1 to 366)

AUTHORS Purnelle,B., Skala,J., van Dyck,L., Tettelin,H. and Goffeau,A.

JOURNAL Unpublished

REFERENCE

2 (residues 1 to 366)

AUTHORS MIPS.

TITLE Direct Submission

JOURNAL Submitted (09-MAY-1994) Data collected by MIPS on behalf of the European yeast chromosome XI sequencing project. MIPS at the Max-Planck-Institut fuer Biochemie, Am Klopferspitze 18a D-82152 Martinsried, FRG; E-mail: News@mpbc.mps.biochem.mpg.de

FEATURES

source

1..366

/organism="Saccharomyces cerevisiae"

/db_xref="taxon:4932"

/chromosome="XI"

1..366

/name="PHD1"

Region

205..285

/region_name="Kila-N"

/note="Kila-N domain; pfam04383"

/db_xref="CDD:146823"

CDS

1..366

/gene="PHD1"

/coded_by="Z28043.1:1106..1206"

/note="GFP: YKL043W"

/db_xref="GSI:P36093"

/db_xref="InterPro:IP0001163"

/db_xref="InterPro:IP0118004"

/db_xref="GDI:P000001528"

/db_xref="UniProtKB/Swiss-Prot:P36093"

ORIGIN

1 mylpeerlh ylnltqena aiptraydn lpfafnelh qatinlpfv retgnayan

61 aqlateptqe ksgycryya wptpypqp qsyqqavlv yatipenfq psfpvawv

121 ppevfdgfv lntlqhtel psliantndt svarpnniks iaasatvta tirtvgusst

181 svlkrvritt mdeadentcy qvawgsivv rradmnnq skllntvkn rprdgdlra

241 kvrvewkig mnlkqavp feryylyqr mqlidltlyl fvdktstsvl arlknkvel

301 tpksepapik qepodnkhei ateklksid alsngastq agelphlkin hidteatqr

Entrez: FASTA Format

PHD1 [Saccharomyces cerevisiae] - Protein result - Mozilla Firefox

http://www.ncbi.nlm.nih.gov/protein/486056?report=fasta

NCBI Resources (2) How To (2) My NCBI Sign In

Protein
 Translations of Life

Search: Protein Limits Advanced search Help

Display Settings (2) FASTA

PHD1 [Saccharomyces cerevisiae]

GenBank: CAA81878.1
 GenPept Statistics

```

>g|486056|emb|CAA81878.1| PHD1 [Saccharomyces cerevisiae]
MTHVREMLHYPLVNTQSNAAIYPTRSYDNTLPSFNELSHQSTINLRFVQNETPNAYANVAQLATSPTQA
KSGYFYQYAVRPTYPQSQSPYQQAULPYATIPKGNFQSPSPFPMVMPREYVQDSFLNLIHPHTEL
RPIIQNTKOTDVARPKNALSIAAASPTFYATIPFQVGSSTSLKPRVITTIMCEDNTLYQVEAMQSVV
PFAADNHEINGTKLLNVTNMTKPPFDGLRSEKVRVVKI GSMILKGVGIPFERATYLAQREQLLDHLVPL
FKVDIESTIQAARKSPNKASLTPKSSPAPKIQEPPDKKHETATEIKPKSI DALSGNASTQGAELPHLKI N
HETDEAGTSPWANELS
    
```

Change region shown

Analyze this sequence
 Run BLAST
 Identify Conserved Domains
 Find in this Sequence

Identical proteins for CAA81878.1
 Ph1p [Saccharomyces cerevisiae] FEIG481434
 TPA: Ph1p [Saccharomyces cerevisiae] [DA04911.3]
 Ph1p [Saccharomyces cerevisiae] EICAY81043
 See all...

LinkOut to external resources
 MIM: Database of Comparative Protein Structure | MIM: Database of Comparat...

All links from this record
 BLAST
 Related Sequences
 Identical Proteins
 3D Structure

Done

Entrez: Downloading Files

The screenshot shows a web browser window displaying the NCBI Entrez Protein page for PHD1 [Saccharomyces cerevisiae]. The browser's address bar shows the URL: <http://www.ncbi.nlm.nih.gov/protein/486056?report=fasta>. The page title is "PHD1 [Saccharomyces cerevisiae] - Protein result - Mozilla Firefox".

The main content area shows the protein name "PHD1 [Saccharomyces cerevisiae]" and its GenBank accession number "CAA81878.1". Below this, the FASTA sequence is displayed:

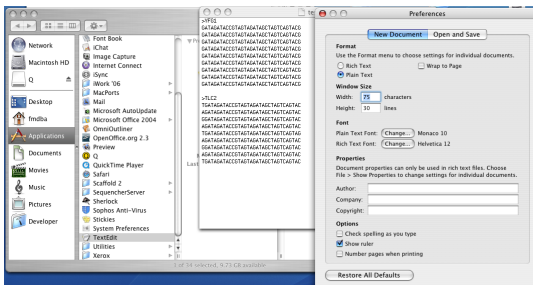
```

>| 486056| emb| CAA81878.1| PHD1 [Saccharomyces cerevisiae]
MTHVREMLJYPLVNTQSNAAIYPTRSYDNTLPSFNEISHQSTINLRFVQNETPNAYANVAQLATSPTQA
KSDQYCHYAVRPTFYQQGQSPYQQAALPYATIPKGNMGQSPSPPMVAWPREVQVQDSFLNLIHPHTEL
RPIIQNTKOTDVARPKNALSIAAASPTFYATIPKQVGSSTSLKSRVLIITMDEKNTLCYQVAMGDSVY
PRADNKHINGTLLNVTMYTPRQPDGLRSEKVRVVKI GSHMLKGVGIPFERATILADREQLLDLHLYPL
FKVDIESTIWAQRKPSNKAASLTPKSSPAPKIQEPPDKKHETATEIKPKSI DALSGNASTQGAELPHLKI N
HETDEAGTSPWANELS
    
```

A "Send to" menu is open over the "FASTA" format, showing options: "File", "Clipboard", "Collections", "Download Items", "Format" (set to "FASTA"), and "Create File".

Below the sequence, there are sections for "proteins for CAA81878.1" (listing EFG481434, TPA, and EIC491043), "LinkOut to external resources" (listing NCBI/NCBI and NCBI/NCBI), and "All links from this record" (listing BLAST, Related Sequences, Identical Proteins, and Protein Families).

Configuring TextEdit for text files



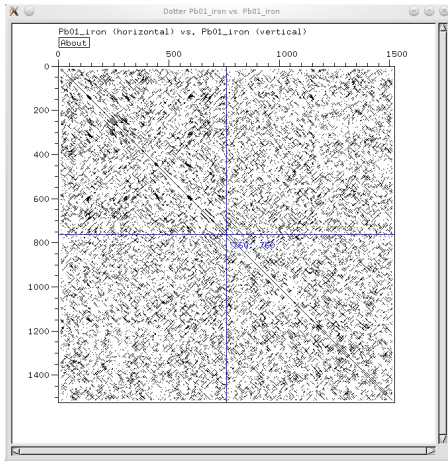
- Format → Plain Text
- Uncheck “check spelling as you type”
- In “Open and Save” uncheck “.txt extension”

Text File Tips

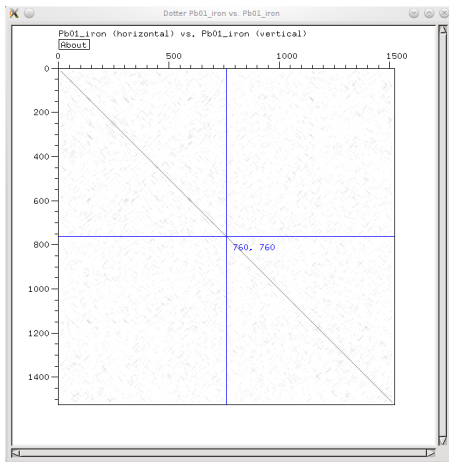
- Line terminators:
 - Unix/Linux: `\n` (linefeed)
 - MacOS: `\r` (carriage return)
 - DOS/Windows: `\r \n` (CRLF)
- Converting from MacOS to Unix on OS X:

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

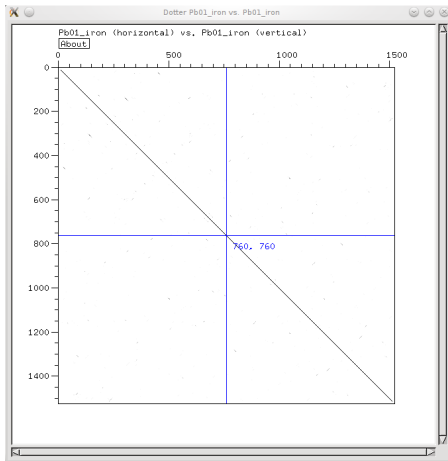
Dotplots: Dot = identity



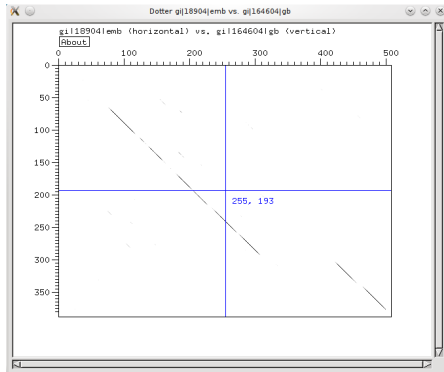
DOTTER: Windowed similarity scores



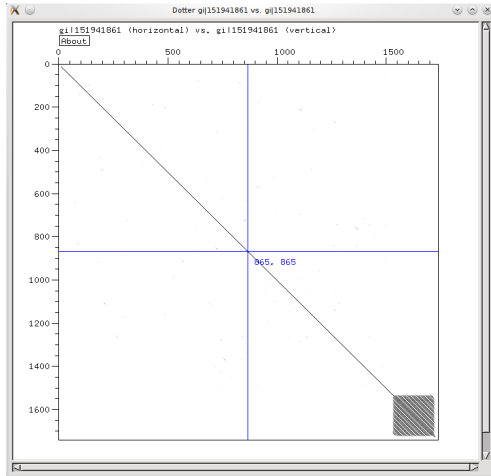
DOTTER: Windowed similarity scores with cutoff



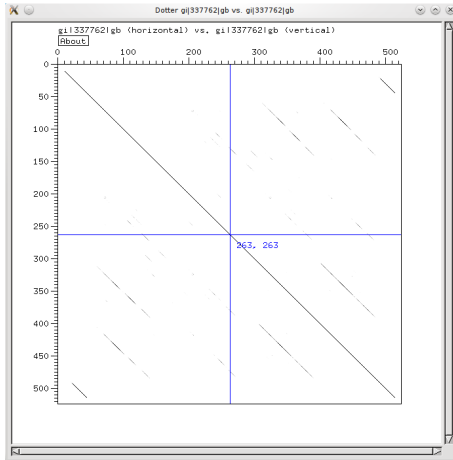
Phytepsin (barley) vs. Pepsinogen (pig)



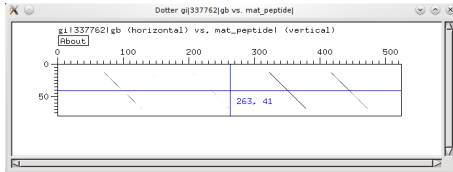
RNA Pol2 (core subunit)



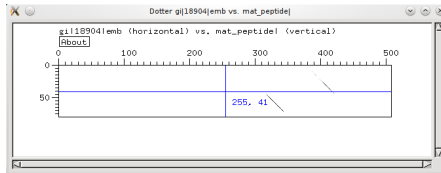
Prosaposin (human)



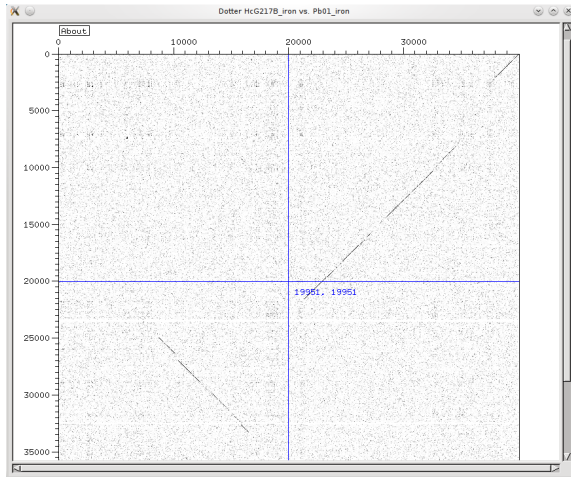
Prosaposin vs. Saposin C



Phytepsin (barley) vs. Saposin C (human)



Iron response locus, *Histoplasma* vs. *Paracoccidioides*



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