Se

Mark Voorhies

3/28/2011

Outline

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

Re

Router:

- SSID: BMS270
- password: deoxyribose

Re

Router:

- SSID: BMS270
- password: deoxyribose

Course website:

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

Re

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)

Re

Router:

- SSID: BMS270
- password: deoxyribose

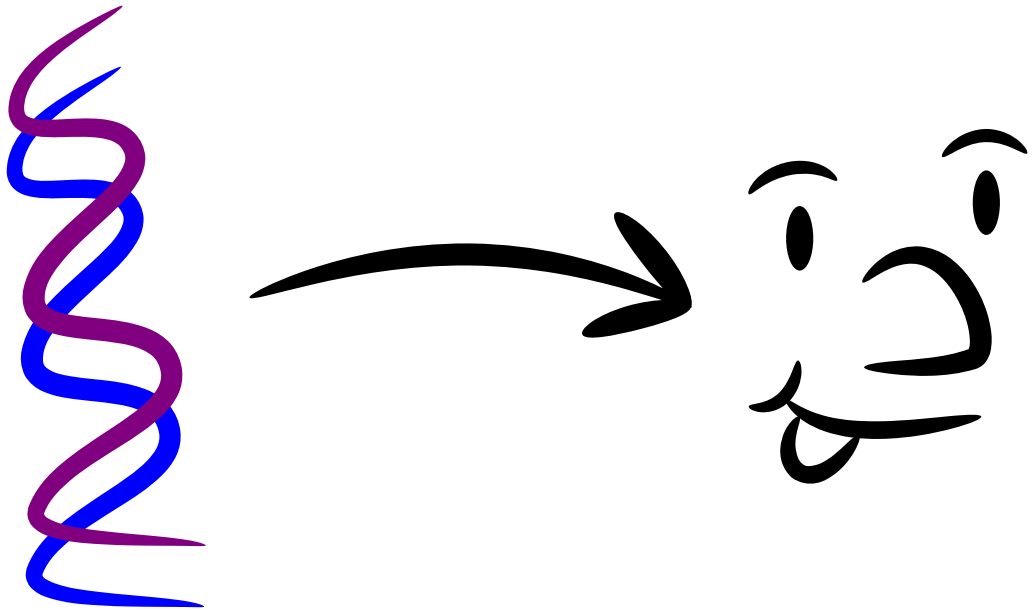
Course website:

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

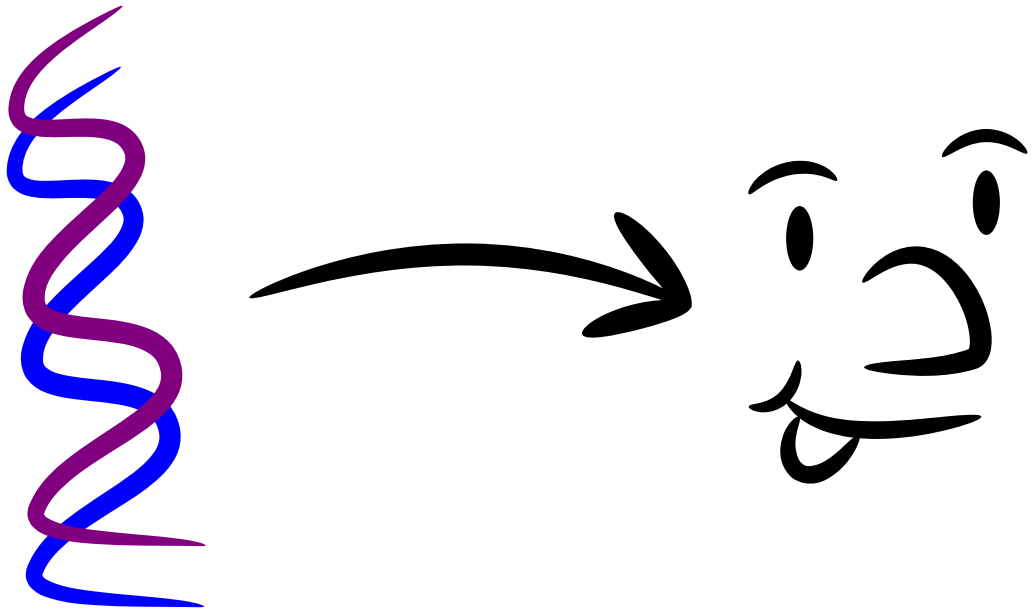
Resources on the course website:

- 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 (CLUSTALX, JalView, MeV, ...)

Course

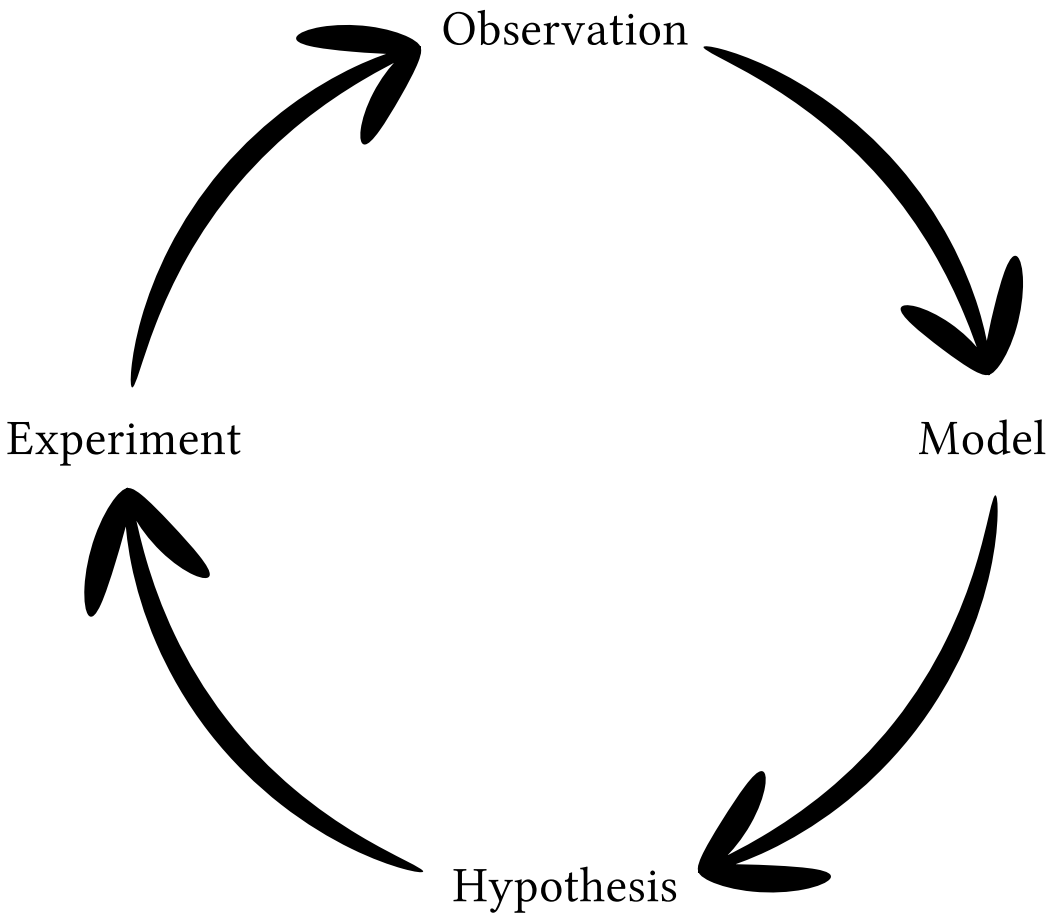


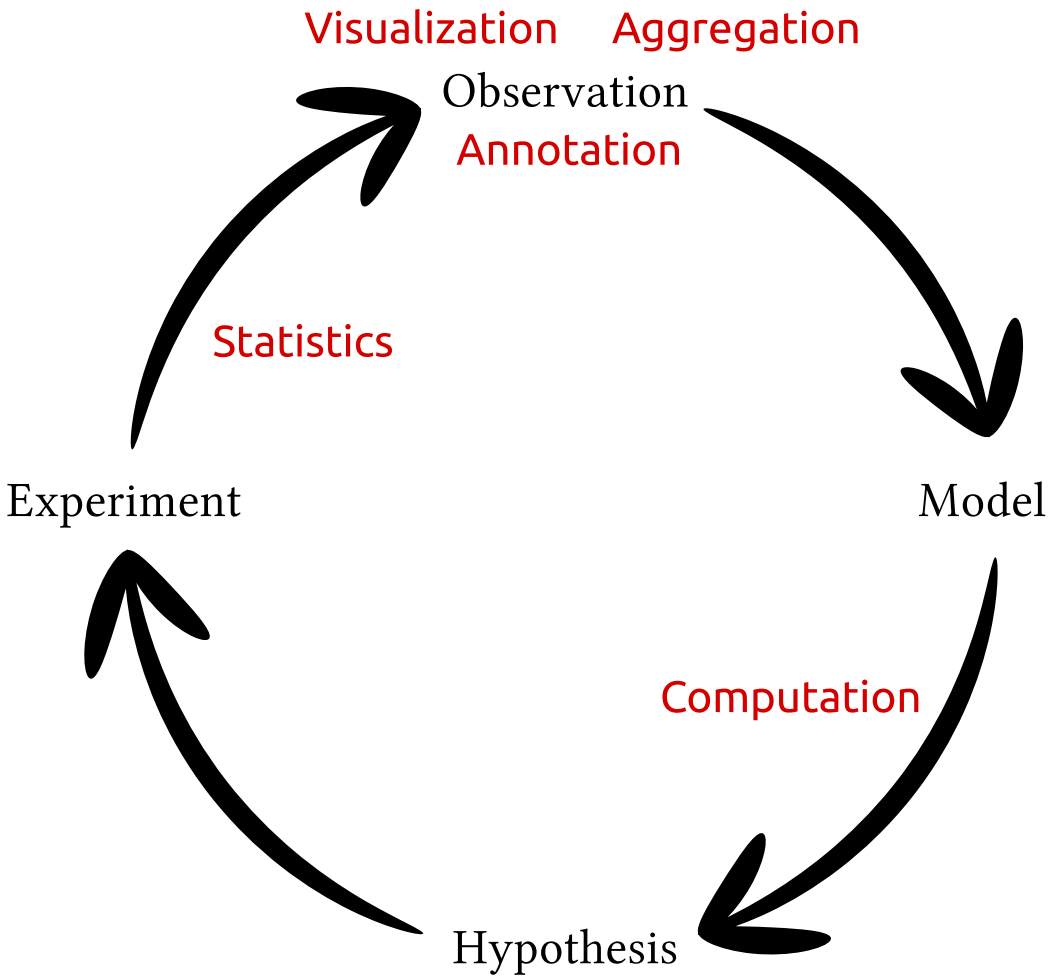
Course



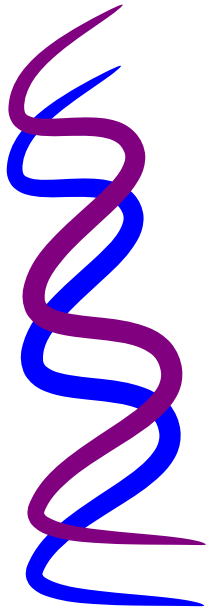
Week 1: Genotype
(Sequence analysis)

Week 2: Phenotype
(Expression pro ling)

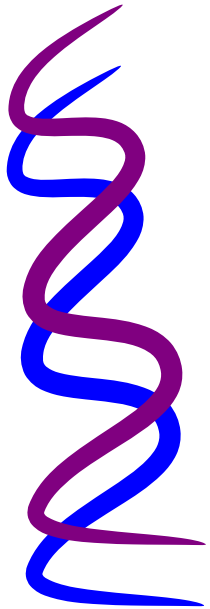

Observation
Model
Hypothesis
Experiment


Visualization
Aggregation
Observation
Annotation
Statistics
Experiment
Model
Computation
Hypothesis

Eve

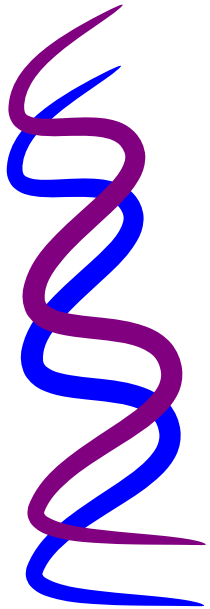


Eve



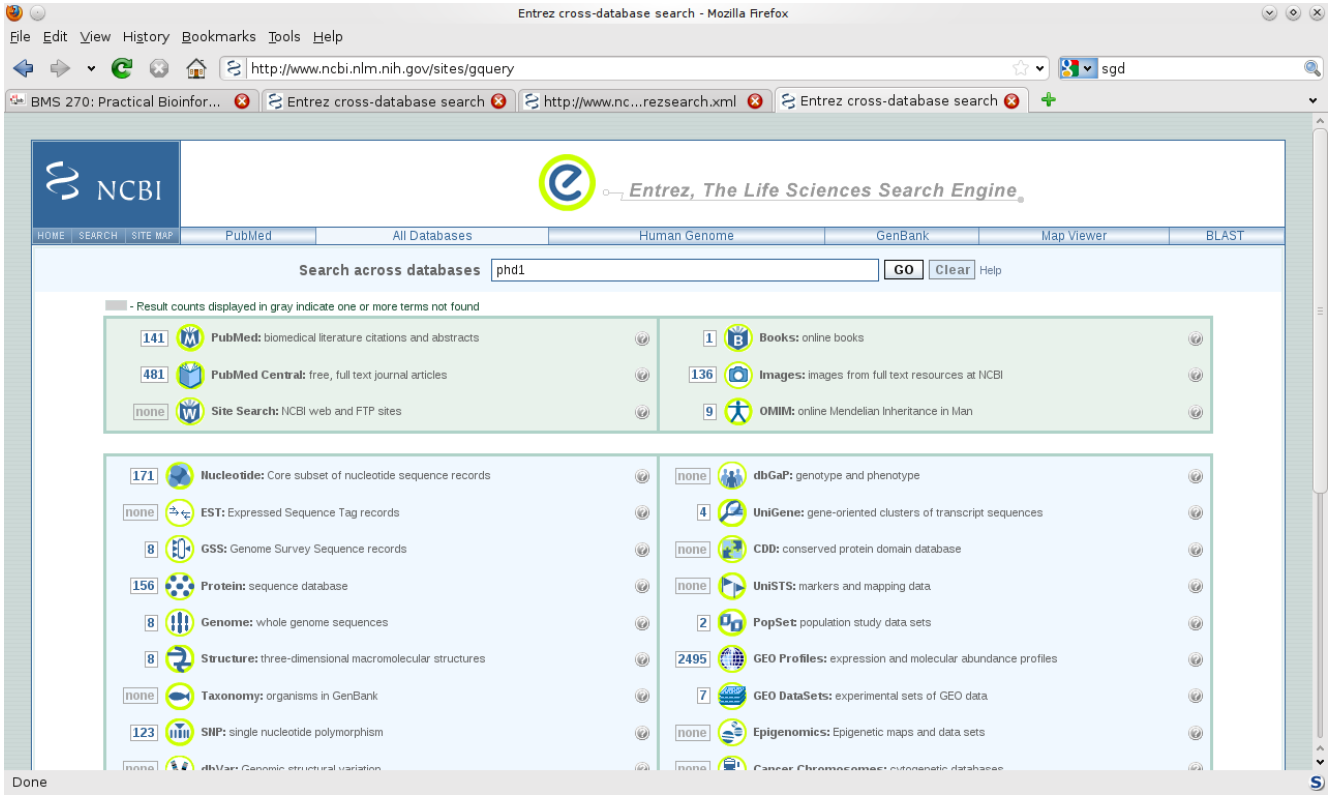
- Export, audit, edit, and import *independent* of a given program.
- Standard le formats for portability.

Eve

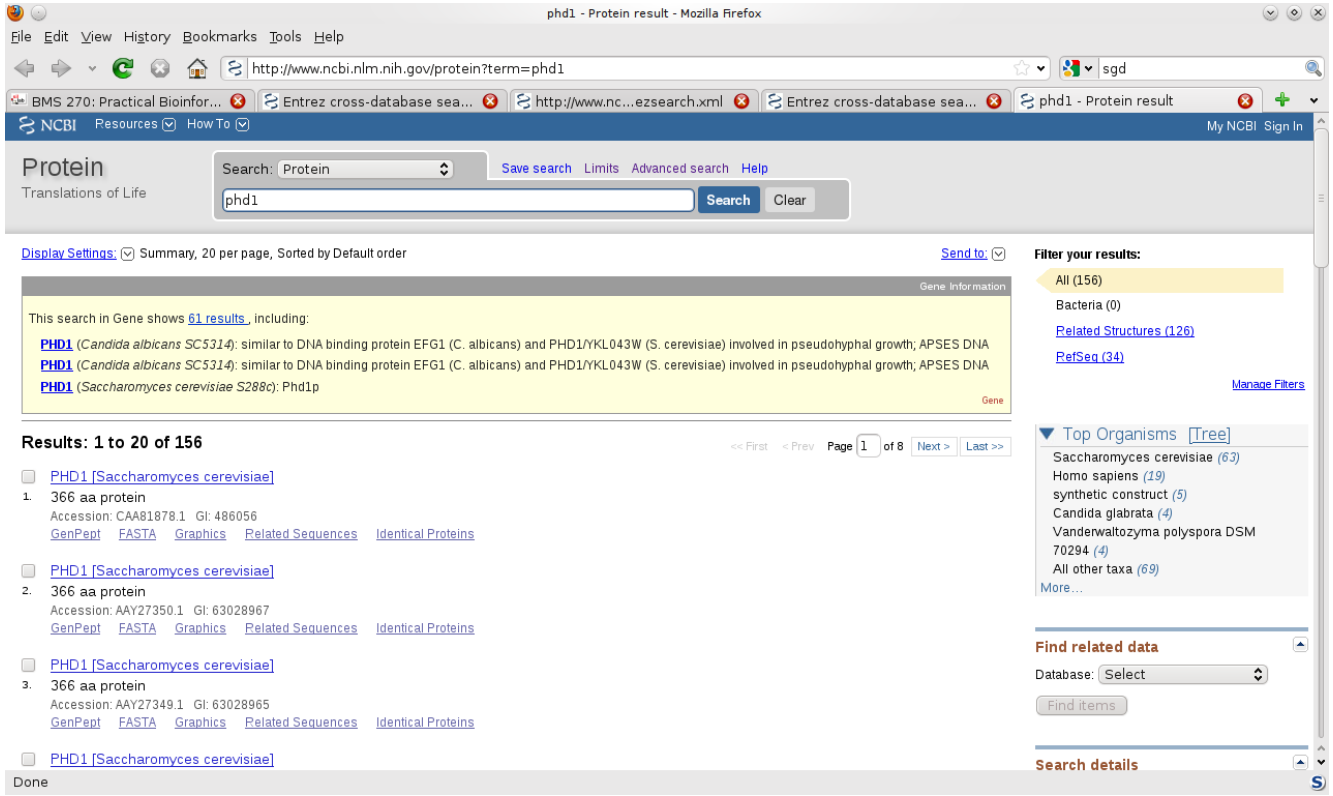


- Export, audit, edit, and import *independent* of a given program.
- Standard le formats for portability.
- Human readable text formats for audit and longevity.

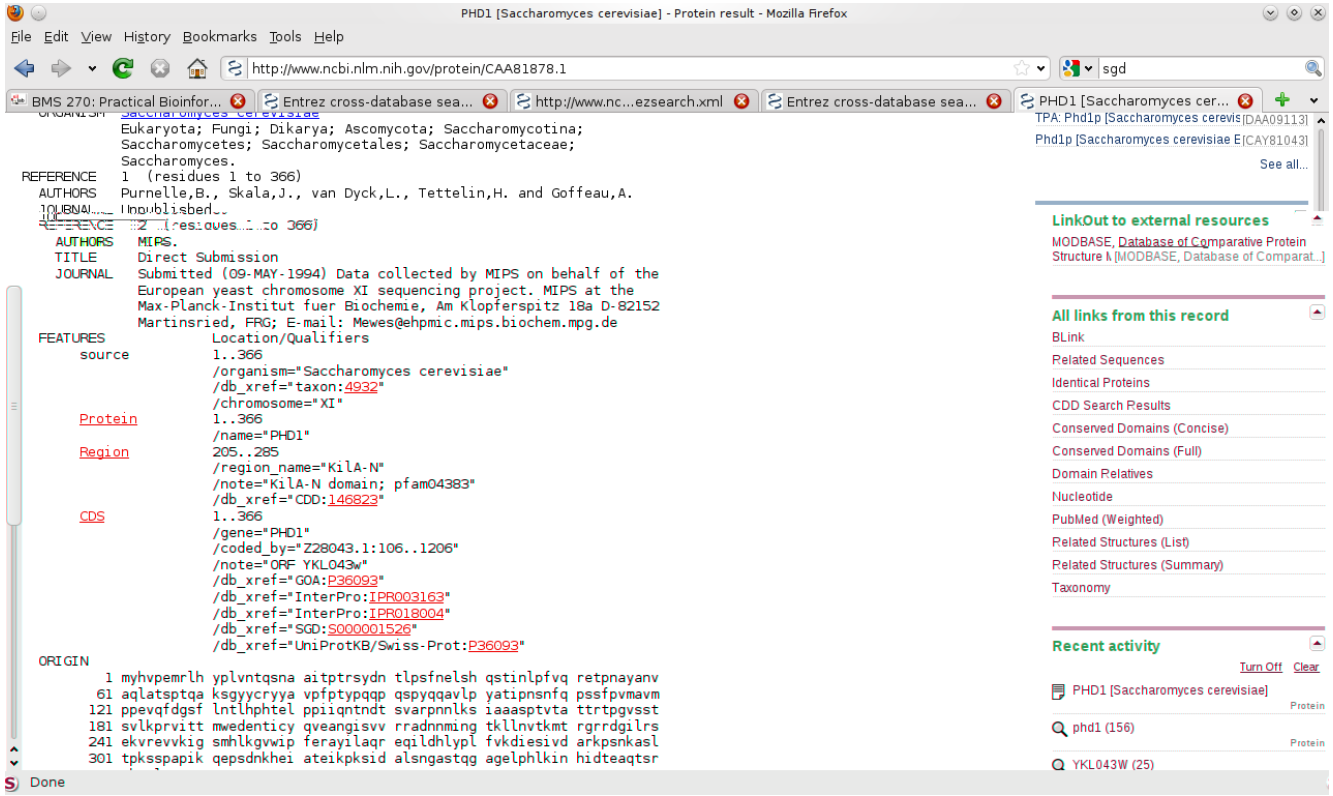
Entre


Entrez cross-database search - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://www.ncbi.nlm.nih.gov/sites/gquery
sgd
BMS 270: Practical Bioinfor...
Entrez cross-database search
http://www.nc...rezsearch.xml
Entrez cross-database search
NCBI
Entrez, The Life Sciences Search Engine
HOME SEARCH SITE MAP
PubMed
All Databases
Human Genome
GenBank
Map Viewer
BLAST
Search across databases
phd1
GO
Clear
Help
- Result counts displayed in gray indicate one or more terms not found
141 PubMed: biomedical literature citations and abstracts
481 PubMed Central: free, full text journal articles
none Site Search: NCBI web and FTP sites
1 Books: online books
136 Images: images from full text resources at NCBI
9 OMIM: online Mendelian Inheritance in Man
171 Nucleotide: Core subset of nucleotide sequence records
none EST: Expressed Sequence Tag records
8 GSS: Genome Survey Sequence records
156 Protein: sequence database
8 Genome: whole genome sequences
8 Structure: three-dimensional macromolecular structures
none Taxonomy: organisms in GenBank
123 SNP: single nucleotide polymorphism
none dbGaP: genotype and phenotype
4 UniGene: gene-oriented clusters of transcript sequences
none CDD: conserved protein domain database
none UniSTS: markers and mapping data
2 PopSet: population study data sets
2495 GEO Profiles: expression and molecular abundance profiles
7 GEO DataSets: experimental sets of GEO data
none Epigenomics: Epigenetic maps and data sets
Done

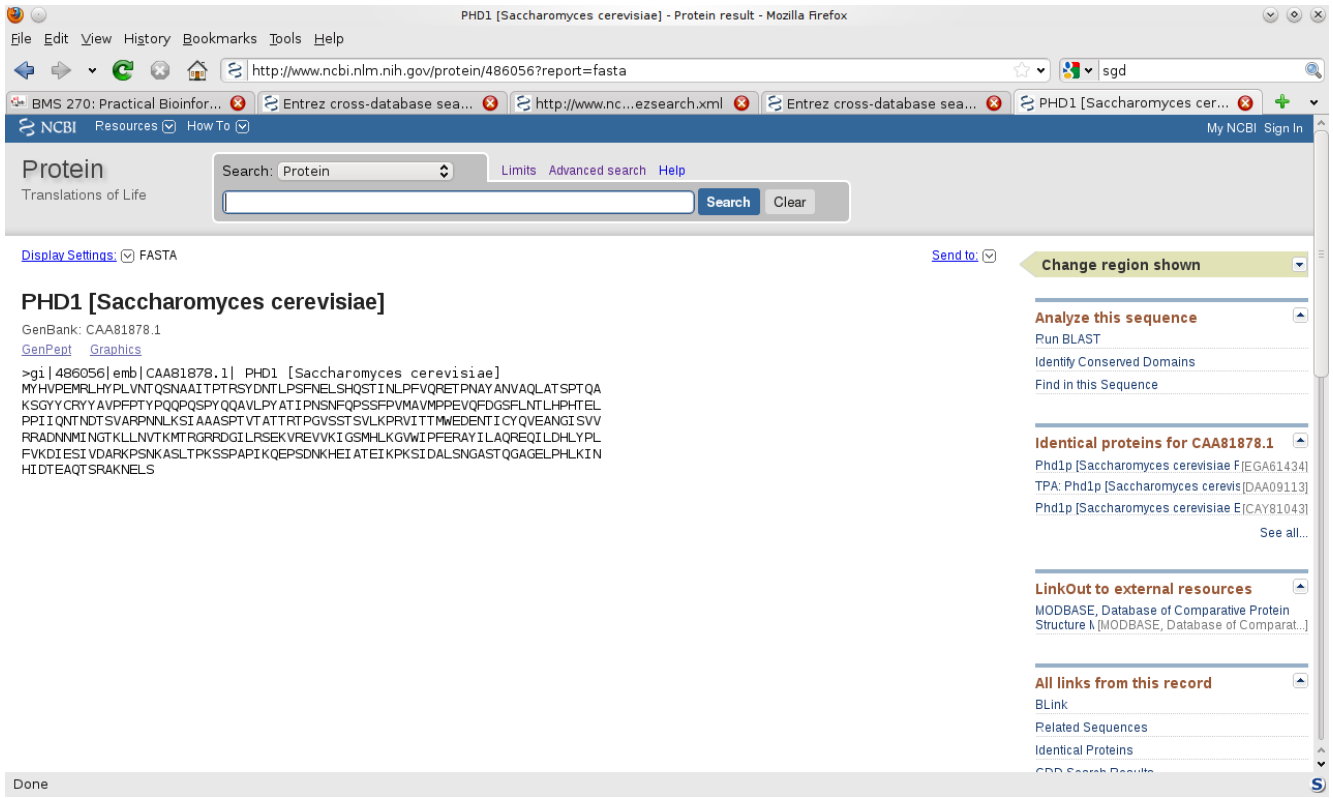
Entre


phd1 - Protein result - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://www.ncbi.nlm.nih.gov/protein?term=phd1
sgd
BMS 270: Practical Bioinfor...
Entrez cross-database sea...
http://www.nc...ezsearch.xml
Entrez cross-database sea...
phd1 - Protein result
NCBI Resources How To
My NCBI Sign In
Protein
Translations of Life
Search: Protein
Save search Limits Advanced search Help
phd1
Search
Clear
Display Settings: Summary, 20 per page, Sorted by Default order
Send to:
Gene Information
This search in Gene shows 61 results , including:
PHD1 (Candida albicans SC5314): similar to DNA binding protein EFG1 (C. albicans) and PHD1/YKL043W (S. cerevisiae) involved in pseudohyphal growth; APSES DNA
PHD1 (Candida albicans SC5314): similar to DNA binding protein EFG1 (C. albicans) and PHD1/YKL043W (S. cerevisiae) involved in pseudohyphal growth; APSES DNA
PHD1 (Saccharomyces cerevisiae S288c): Phd1p
Gene
Results: 1 to 20 of 156
<< First < Prev Page 1 of 8 Next > Last >>
PHD1 [Saccharomyces cerevisiae]
1. 366 aa protein
Accession: CAA81878.1 GI: 486056
GenPept FASTA Graphics Related Sequences Identical Proteins
PHD1 [Saccharomyces cerevisiae]
2. 366 aa protein
Accession: AAY27350.1 GI: 63028967
GenPept FASTA Graphics Related Sequences Identical Proteins
PHD1 [Saccharomyces cerevisiae]
3. 366 aa protein
Accession: AAY27349.1 GI: 63028965
GenPept FASTA Graphics Related Sequences Identical Proteins
PHD1 [Saccharomyces cerevisiae]
Filter your results:
All (156)
Bacteria (0)
Related Structures (126)
RefSeq (34)
Manage Filters
Top Organisms [Tree]
Saccharomyces cerevisiae (63)
Homo sapiens (19)
synthetic construct (5)
Candida glabrata (4)
Vanderwaltozyma polyspora DSM 70294 (4)
All other taxa (69)
More...
Find related data
Database: Select
Find items
Search details
Done

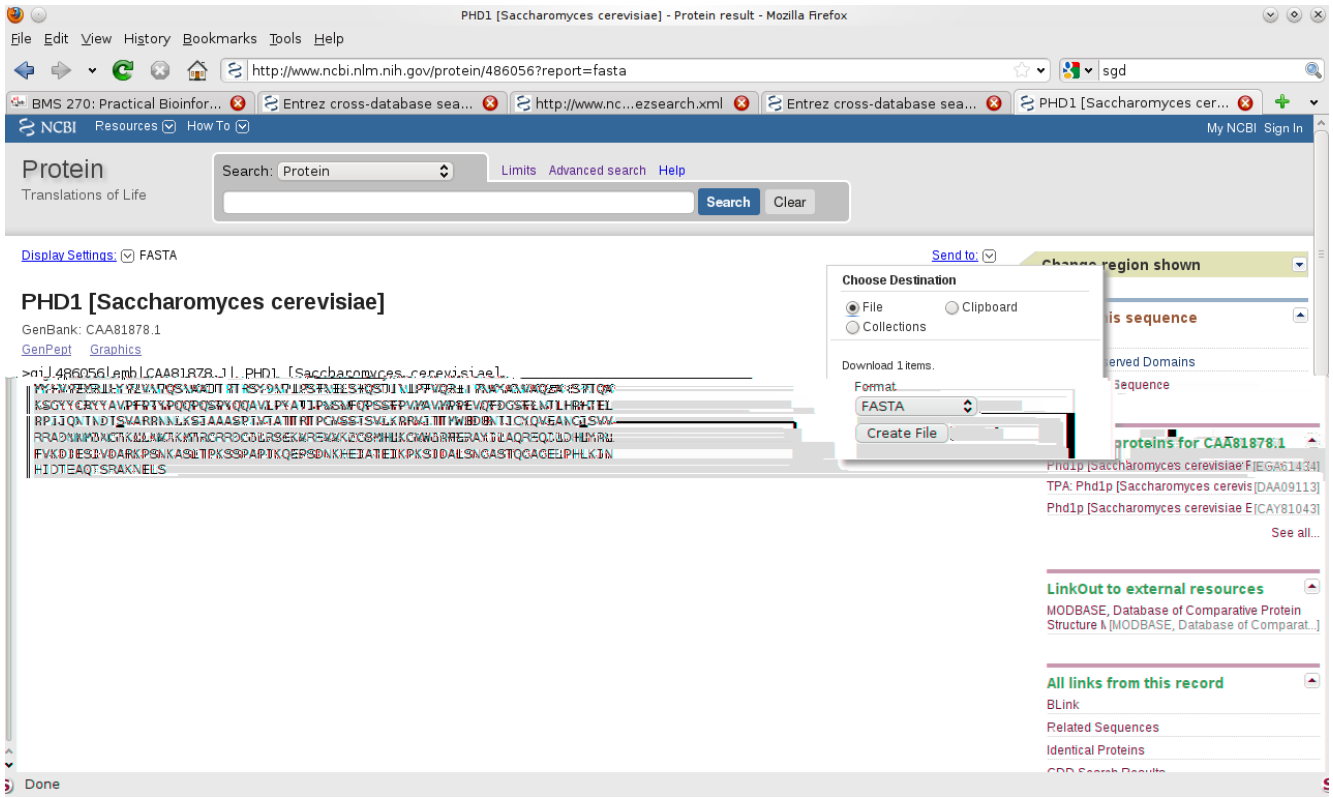
Entre


PHD1 [Saccharomyces cerevisiae] - Protein result - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://www.ncbi.nlm.nih.gov/protein/CAA81878.1
sgd
BMS 270: Practical Bioinfor...
Entrez cross-database sea...
http://www.nc...ezsearch.xml
Entrez cross-database sea...
PHD1 [Saccharomyces cer...
Eukaryota; Fungi; Dikarya; Ascomycota; Saccharomycotina;
Saccharomycetes; Saccharomycetales; Saccharomycetaceae;
Saccharomyces.
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 Klopferspitz 18a D-82152
Martinsried, FRG; E-mail: Mewes@ehpmic.mips.biochem.mpg.de
FEATURES Location/Qualifiers
source 1..366
/organism="Saccharomyces cerevisiae"
/db_xref="taxon:4932"
/chromosome="XI"
Protein 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="Z28043.1:106..1206"
/note="ORF YKL043w"
/db_xref="GOA:P36093"
/db_xref="InterPro:IPR003163"
/db_xref="InterPro:IPR018004"
/db_xref="SGD:S000001526"
/db_xref="UniProtKB/Swiss-Prot:P36093"
ORIGIN
1 myhvpemrlh yplvntqsna aitptrsydn tlpsfnelsh qstinlpfvq retpnayanv
61 aqlatsptqa ksgyycryya vpfptypqqp qspyqqavlp yatipnsnfq pssfpvmavm
121 ppevqfdgsf lntlhphtel ppiiqntndt svarpnnlks iaaasptvta ttrtpgvsst
181 svlkprvitt mwedenticy qveangisvv rradnnming tkllnvtkmt rgrrdgilrs
241 ekvrevvkig smhlkgvwip ferayilaqr eqildhlypl fvkdiesivd arkpsnkasl
301 tpksspapik qepsdnkhei ateikpksid alsngastqg agelphlkin hidteaqtsr
TPA: Phd1p [Saccharomyces cerevis [DAA09113]
Phd1p [Saccharomyces cerevisiae E [CAY81043]
See all...
LinkOut to external resources
MODBASE, Database of Comparative Protein Structure M [MODBASE, Database of Comparat...]
All links from this record
BLink
Related Sequences
Identical Proteins
CDD Search Results
Conserved Domains (Concise)
Conserved Domains (Full)
Domain Relatives
Nucleotide
PubMed (Weighted)
Related Structures (List)
Related Structures (Summary)
Taxonomy
Recent activity
Turn Off Clear
PHD1 [Saccharomyces cerevisiae]
Protein
phd1 (156)
Protein
YKL043W (25)
Done

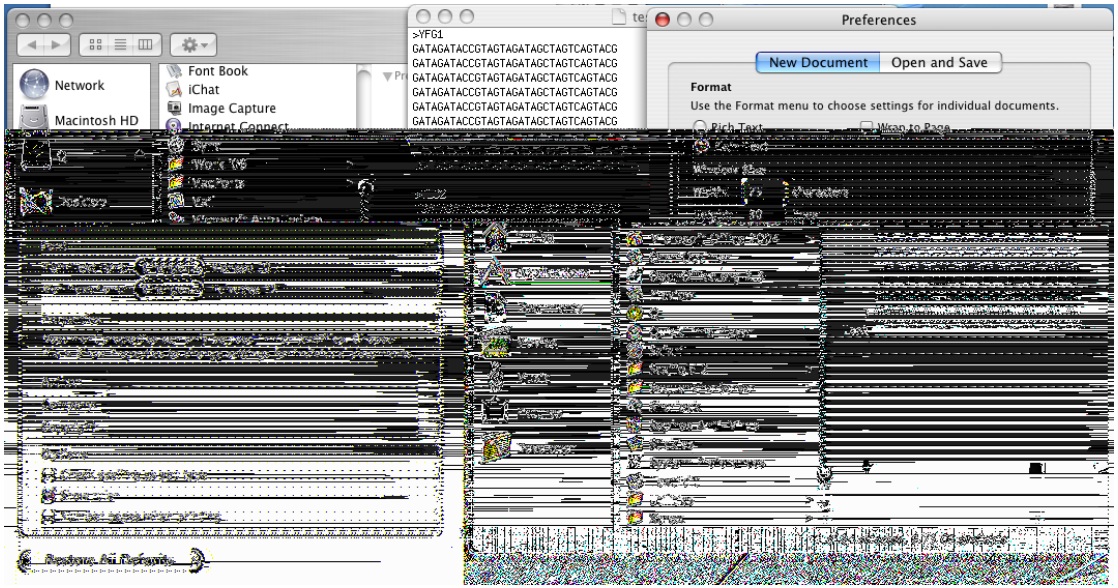
Entre


PHD1 [Saccharomyces cerevisiae] - Protein result - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://www.ncbi.nlm.nih.gov/protein/486056?report=fasta
sgd
BMS 270: Practical Bioinfor...
Entrez cross-database sea...
http://www.nc...ezsearch.xml
Entrez cross-database sea...
PHD1 [Saccharomyces cer...
NCBI Resources How To
My NCBI Sign In
Protein
Translations of Life
Search: Protein
Limits Advanced search Help
Search Clear
Display Settings: FASTA
Send to:
Change region shown
PHD1 [Saccharomyces cerevisiae]
GenBank: CAA81878.1
GenPept Graphics
>gi|486056|emb|CAA81878.1| PHD1 [Saccharomyces cerevisiae]
MYHVPEMRLHYPLVNTQSNAAITPTRSYDNTLPSFNELSHQSTINLPFVQRETPNAYANVAQLATSPTQA
KSGYYCRYYAVPFPTYPQQPQSPYQQAVLPYATIPNSNFQPSSFPVMAVMPPEVQFDGSFLNTLHPHTEL
PPIIQNTNDTSVARPNNLKSIAAASPTVTATTRTPGVSSTSVLKPRVITTMWEDENTICYQVEANGISVV
RRADNNMINGTKLLNVTKMTRGRRDGILRSEKVREVVKIGSMHLKGVWIPFERAYILAQREQILDHLYPL
FVKDIESIVDARKPSNKASLTPKSSPAPIKQEPSDNKHEIATEIKPKSIDALSNGASTQGAGELPHLKIN
HIDTEAQTSRAKNELS
Analyze this sequence
Run BLAST
Identify Conserved Domains
Find in this Sequence
Identical proteins for CAA81878.1
Phd1p [Saccharomyces cerevisiae F [EGA61434]
TPA: Phd1p [Saccharomyces cerevis [DAA09113]
Phd1p [Saccharomyces cerevisiae E [CAY81043]
See all...
LinkOut to external resources
MODBASE, Database of Comparative Protein Structure M [MODBASE, Database of Comparat...]
All links from this record
BLink
Related Sequences
Identical Proteins
Done

Entre


PHD1 [Saccharomyces cerevisiae] - Protein result - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://www.ncbi.nlm.nih.gov/protein/486056?report=fasta
NCBI Resources How To
My NCBI Sign In
Protein
Translations of Life
Search: Protein
Limits Advanced search Help
Search Clear
Display Settings: FASTA
Send to:
PHD1 [Saccharomyces cerevisiae]
GenBank: CAA81878.1
GenPept Graphics
Choose Destination
File Clipboard
Collections
Download 1 items.
Format
FASTA
Create File
LinkOut to external resources
All links from this record
BLink
Related Sequences
Identical Proteins
Done

Configuring Te



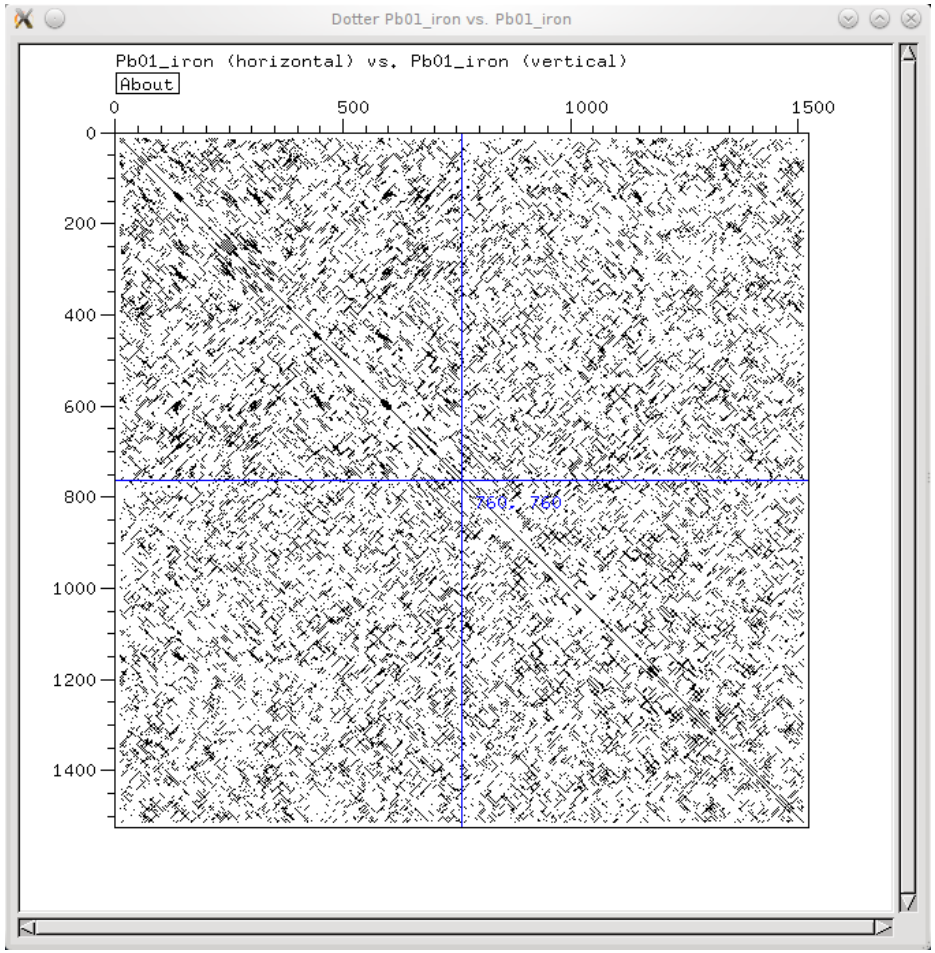
- Format $\rightarrow$ Plain Text
- Uncheck \check spelling as you type"
- In \Open and Save" uncheck \.txt extension"

Te

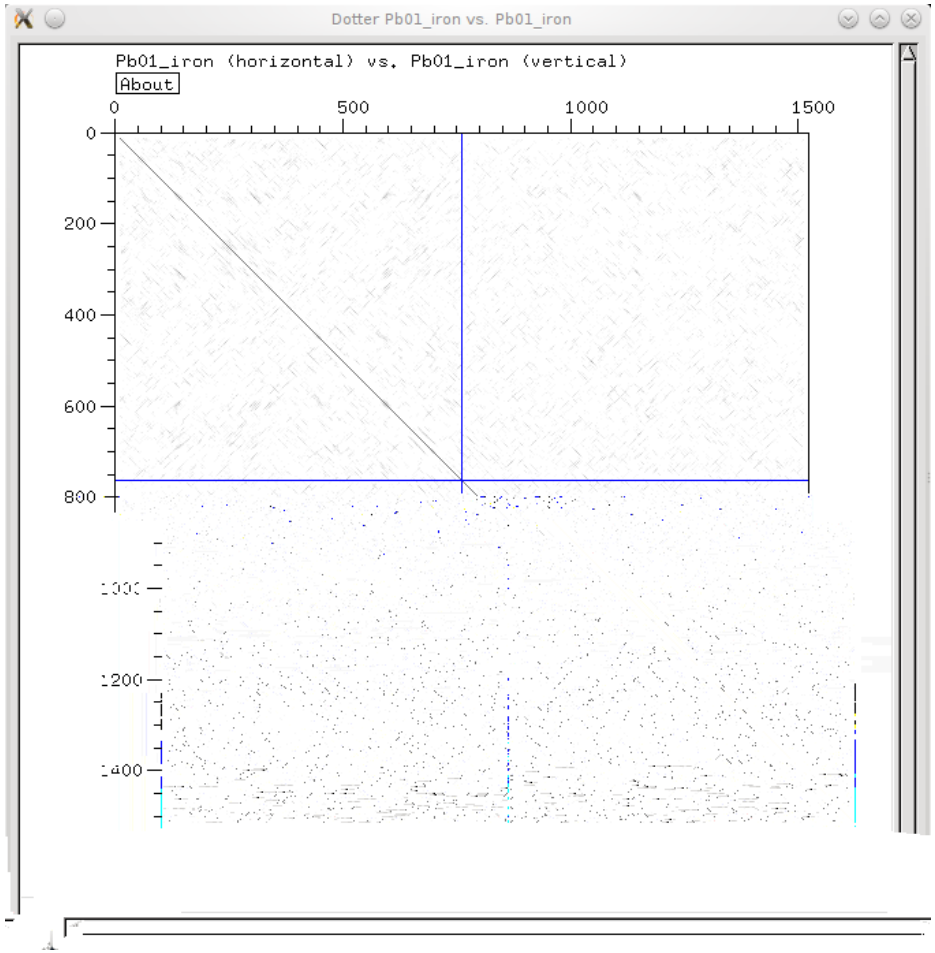
- Line terminators:
  - Unix/Linux: \n (linefeed)
  - MacOS: \r (carriage return)
  - DOS/Windows: \r \n (CRLF)
- Converting from MacOS to Unix on OS X:

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

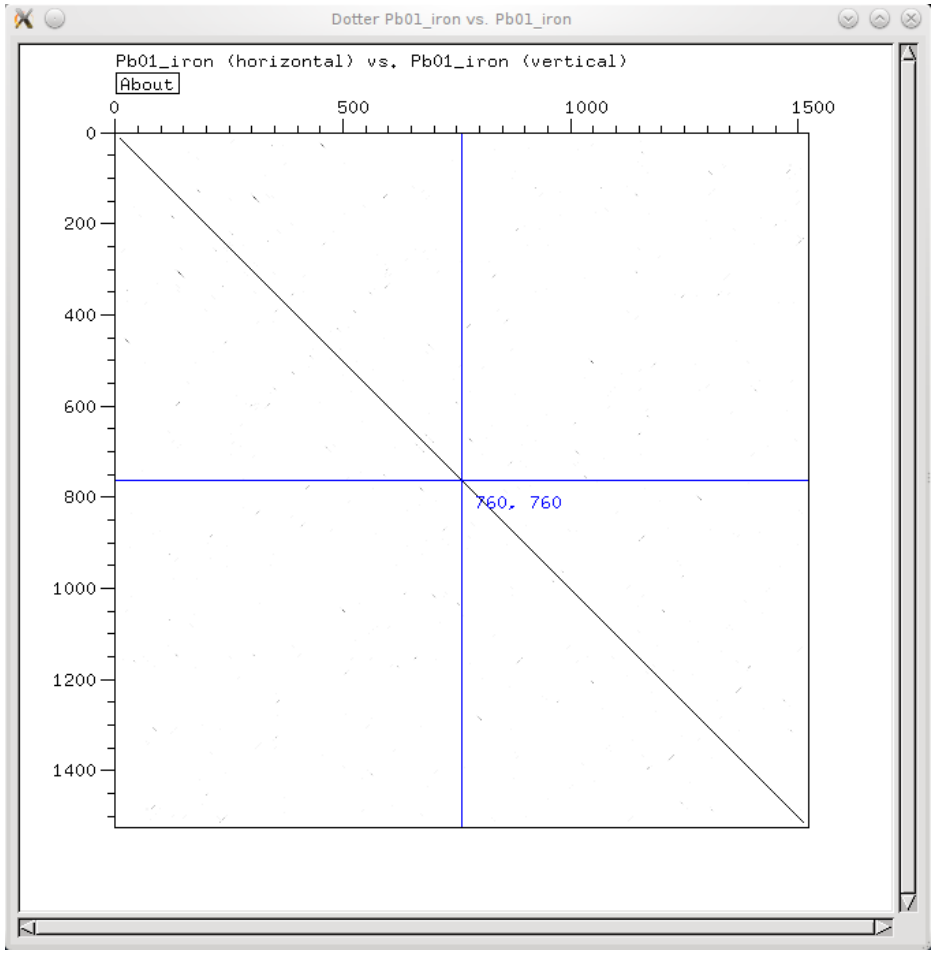
Dotplots: Dot = ide



DOTTER: Windowe



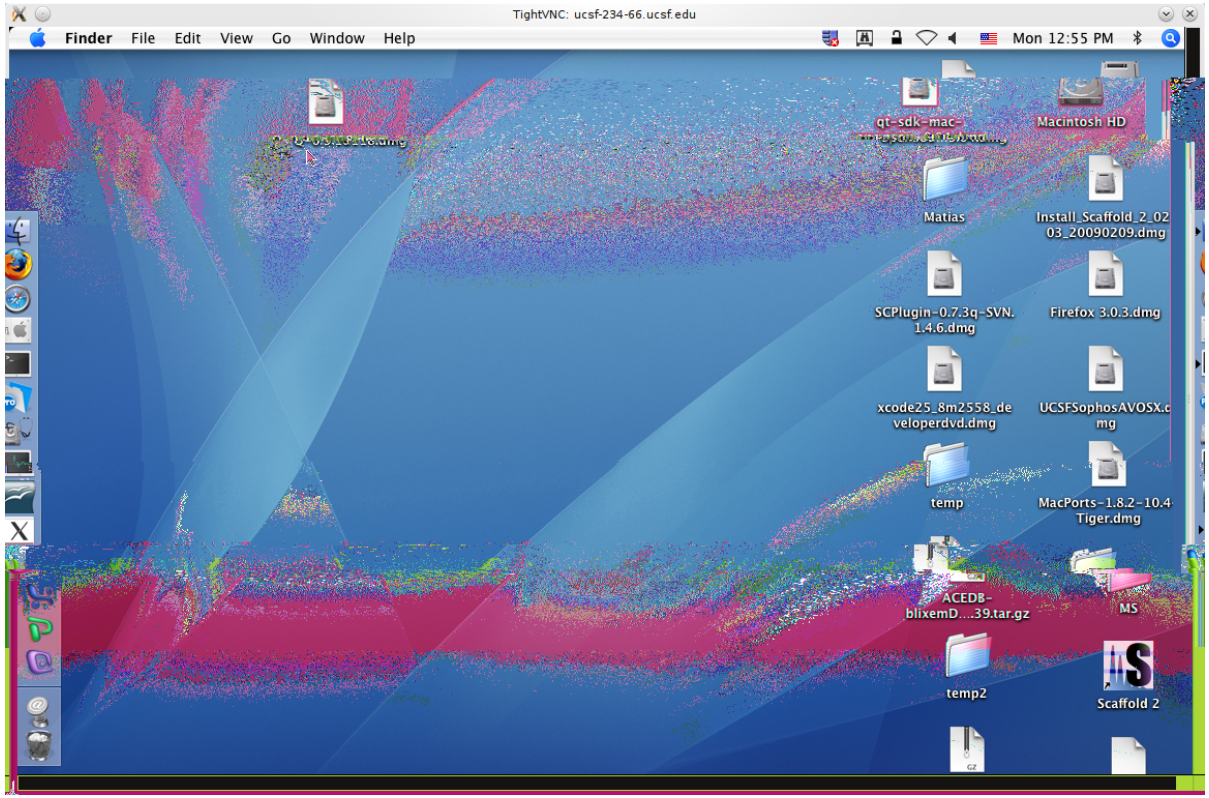
DOTTER: Windowe



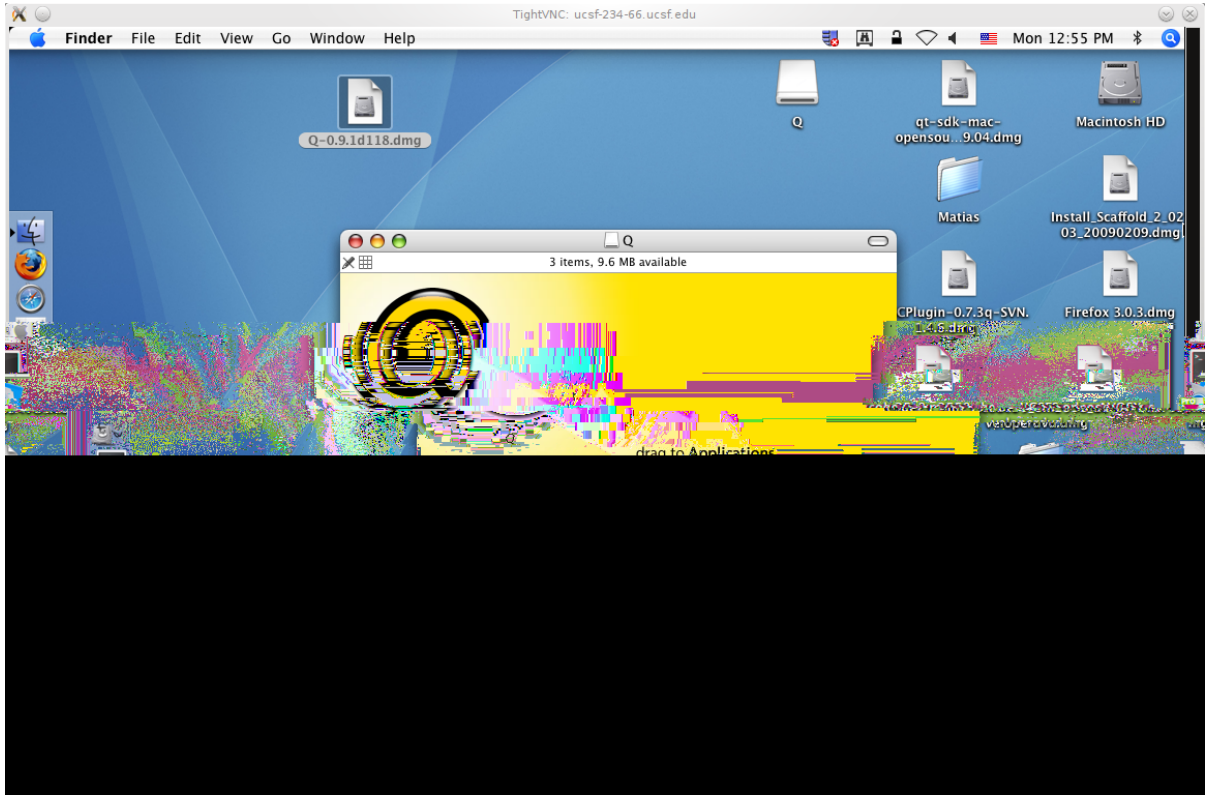
OS X Part 1: Se

1. Open the \Q" disk image
2. Click on the \Q" program
3. Click on \+" to create a new virtual machine
4. Choose the biosqueeze.qcow disk image
5. Allow SSH to pass through the \Q" rewall
6. Connect the virtual machine's SSH port to a local port: -redir tcp:8022::22
7. Start the virtual machine (wait for it to nish booting)

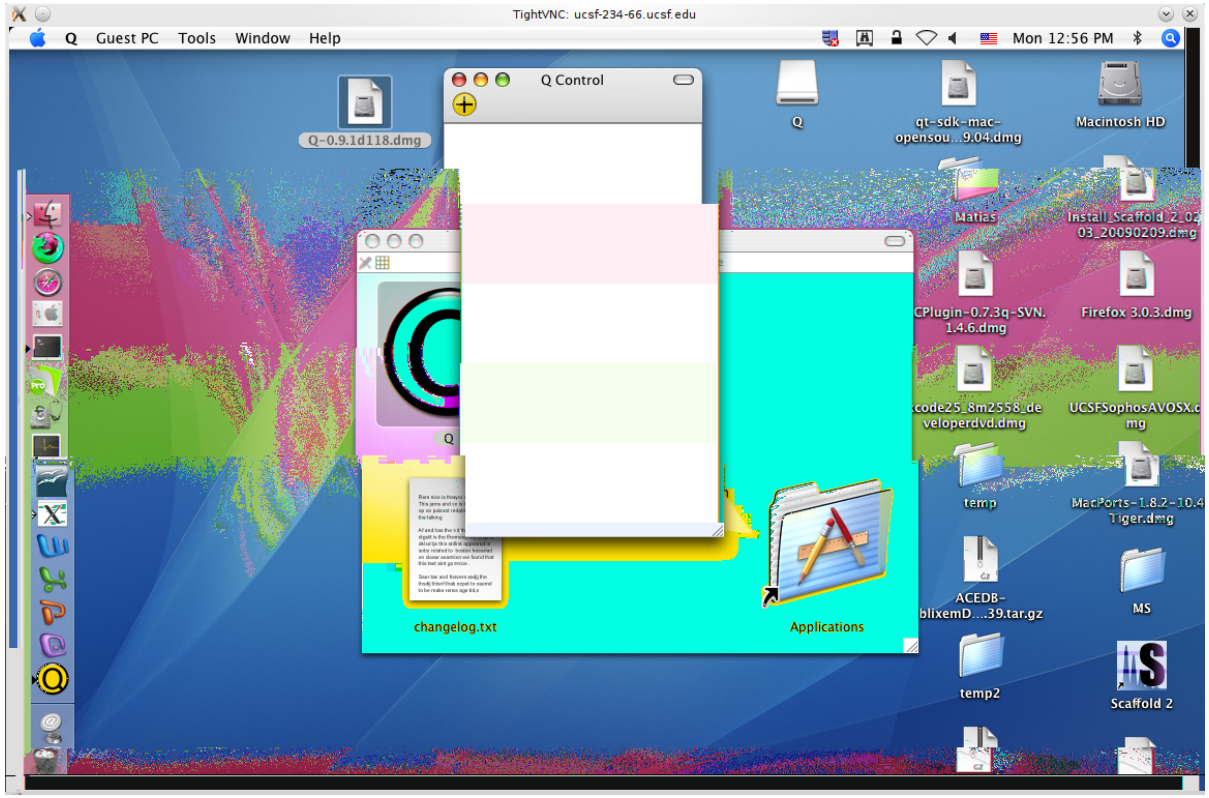
OS X Part 1: Se


TightVNC: ucsf-234-66.ucsf.edu
Finder File Edit View Go Window Help
Mon 12:55 PM
Macintosh HD
Matias
Install_Scaffold_2_02
03_20090209.dmg
SCPlugin-0.7.3q-SVN.
1.4.6.dmg
Firefox 3.0.3.dmg
xcode25_8m2558_de
veloperdvd.dmg
temp
MacPorts-1.8.2-10.4
Tiger.dmg
ACEDB-
blixemD....39.tar.gz
MS
temp2
Scaffold 2

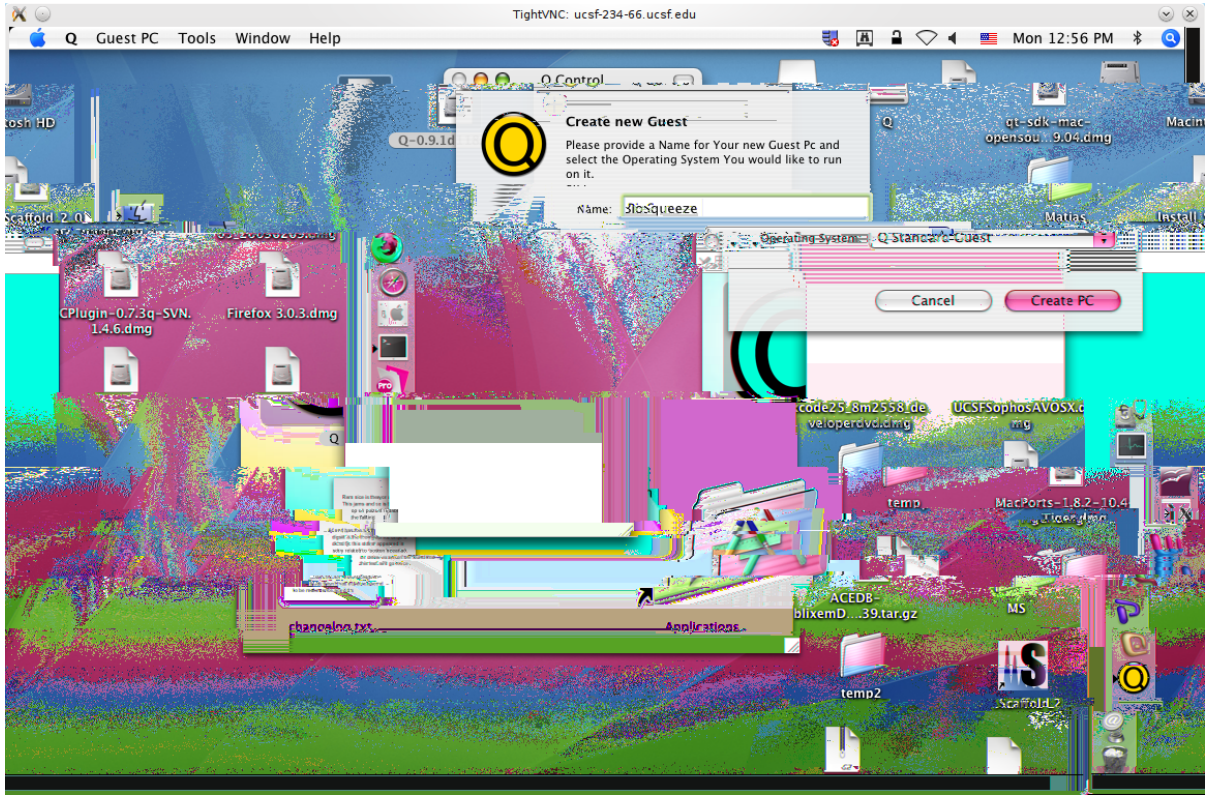
OS X Part 1: Se


TightVNC: ucsf-234-66.ucsf.edu
Finder File Edit View Go Window Help
Mon 12:55 PM
Q-0.9.1d118.dmg
Q
qt-sdk-mac-opensou...9.04.dmg
Macintosh HD
Matias
Install_Scaffold_2_0203_20090209.dmg
Q
3 items, 9.6 MB available
CPlugin-0.7.3q-SVN.
Firefox 3.0.3.dmg
drag to Applications

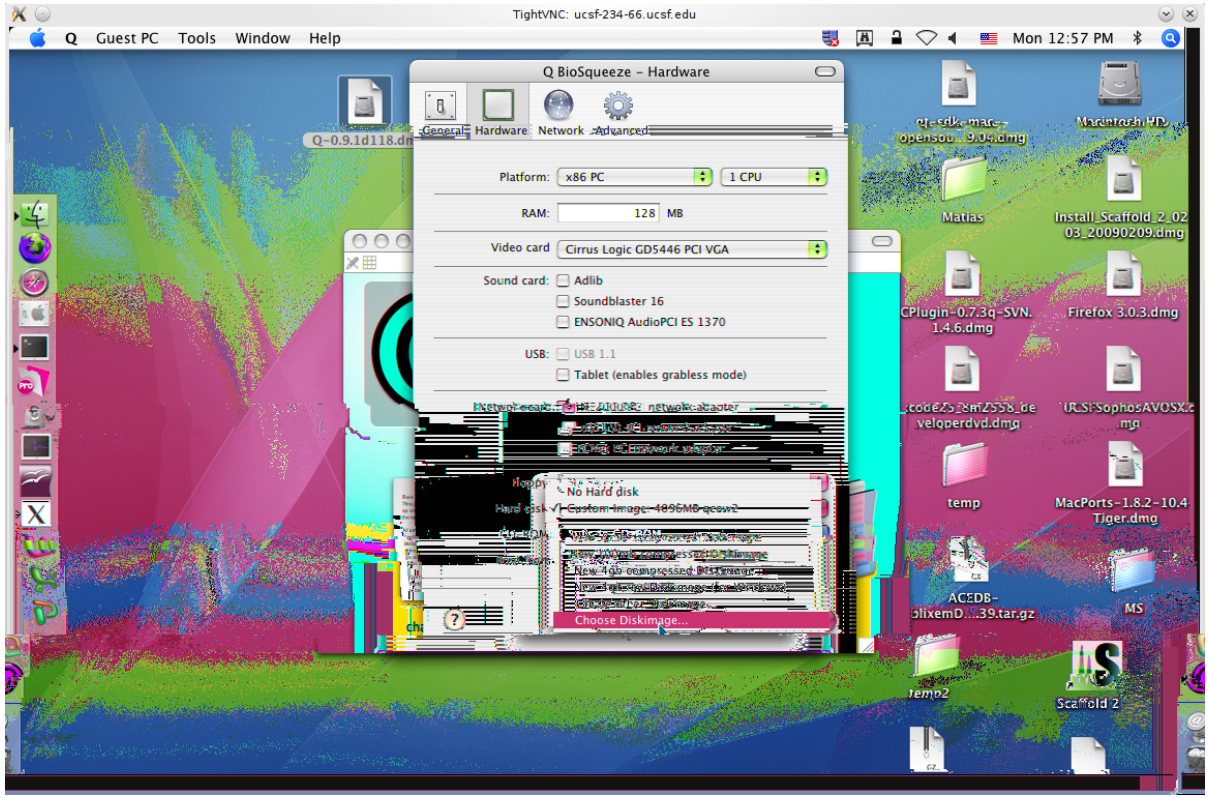
OS X Part 1: Se


TightVNC: ucsf-234-66.ucsf.edu
Guest PC
Tools
Window
Help
Mon 12:56 PM
Q Control
Q-0.9.1d118.dmg
Q
qt-sdk-mac-opensou...9.04.dmg
Macintosh HD
Matias
Firefox 3.0.3.dmg
temp
temp2
MS
Scaffold 2
changelog.txt
Applications

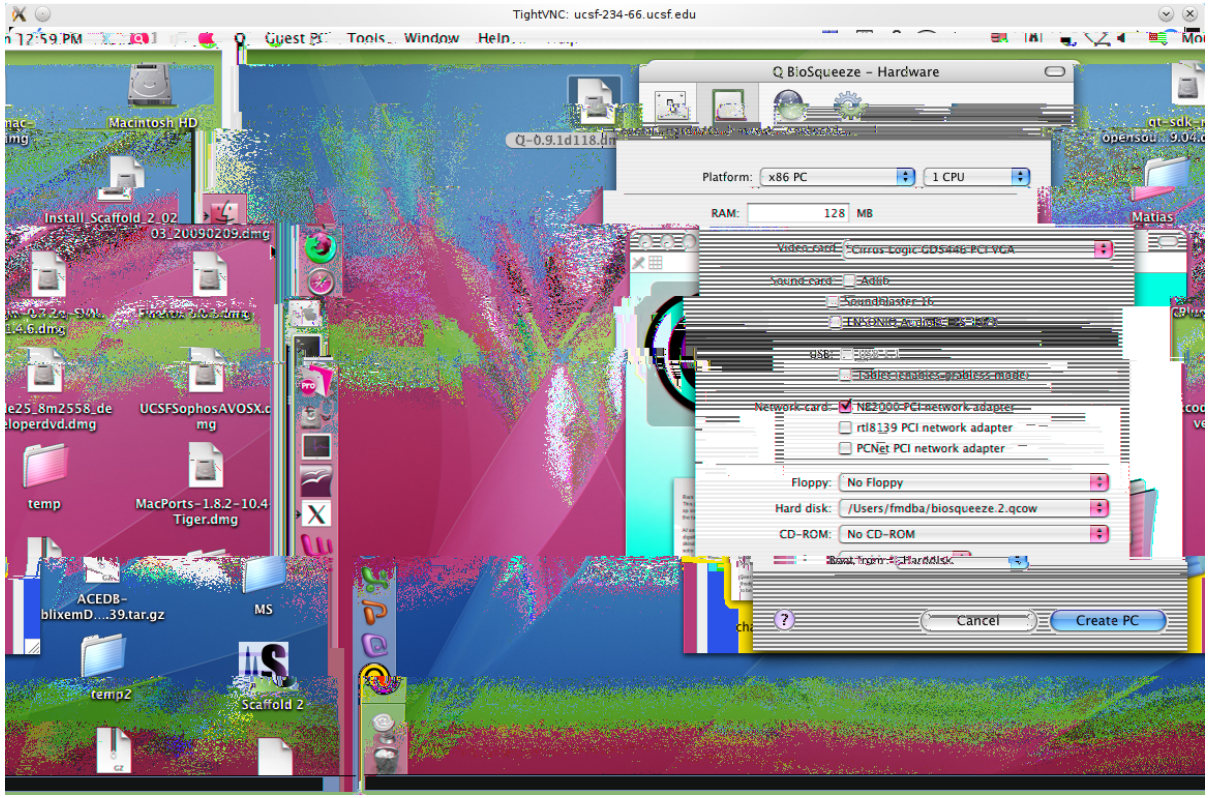
OS X Part 1: Se


TightVNC: ucsf-234-66.ucsf.edu
Guest PC
Tools
Window
Help
Mon 12:56 PM
Create new Guest
Please provide a Name for Your new Guest Pc and select the Operating System You would like to run on it.
Name:
Cancel
Create PC
Firefox 3.0.3.dmg
temp
temp2
MS
Applications

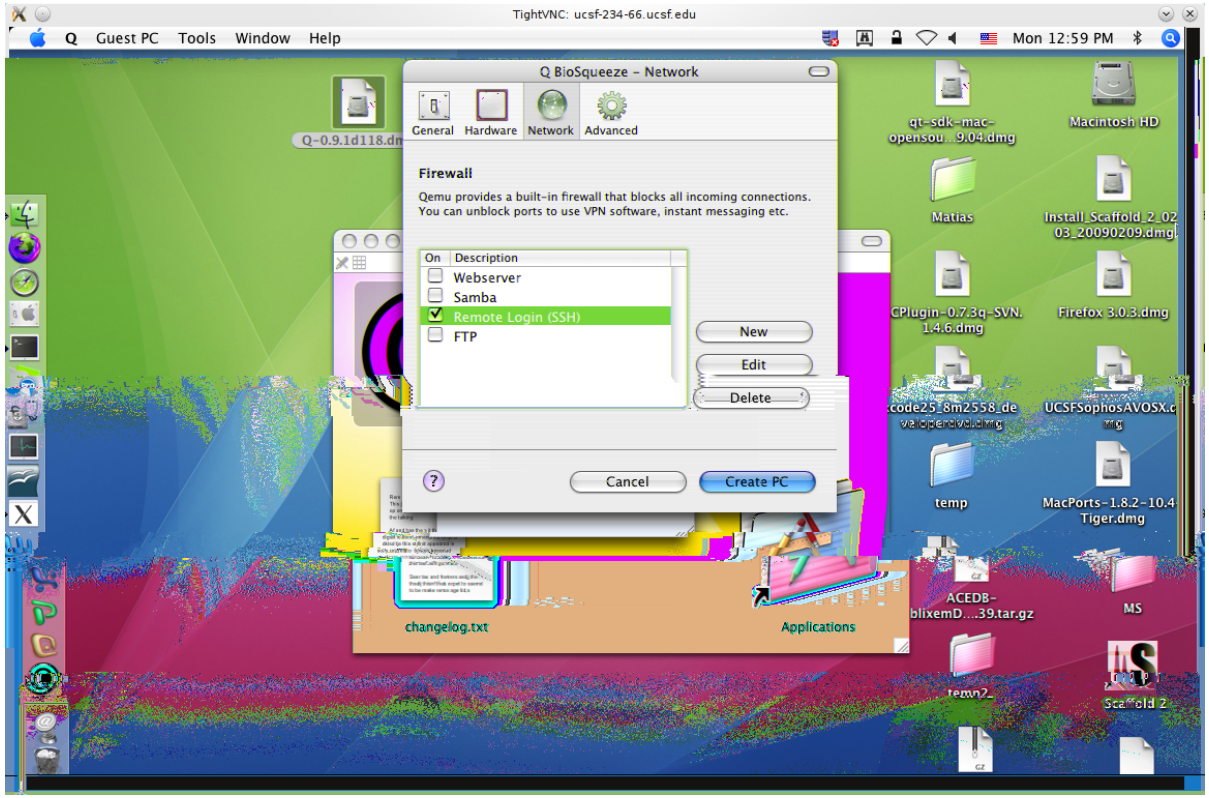
OS X Part 1: Se


TightVNC: ucsf-234-66.ucsf.edu
Guest PC
Tools
Window
Help
Mon 12:57 PM
Q BioSqueeze – Hardware
General
Hardware
Network
Advanced
Platform:
x86 PC
1 CPU
RAM:
128
MB
Video card
Cirrus Logic GD5446 PCI VGA
Sound card:
Adlib
Soundblaster 16
ENSONIQ AudioPCI ES 1370
USB:
USB 1.1
Tablet (enables grabless mode)
No Hard disk
Choose Diskimage...
Matias
Install_Scaffold_2_02_03_20090209.dmg
Firefox 3.0.3.dmg
temp
MacPorts-1.8.2-10.4-Tiger.dmg
temp2
Scaffold 2

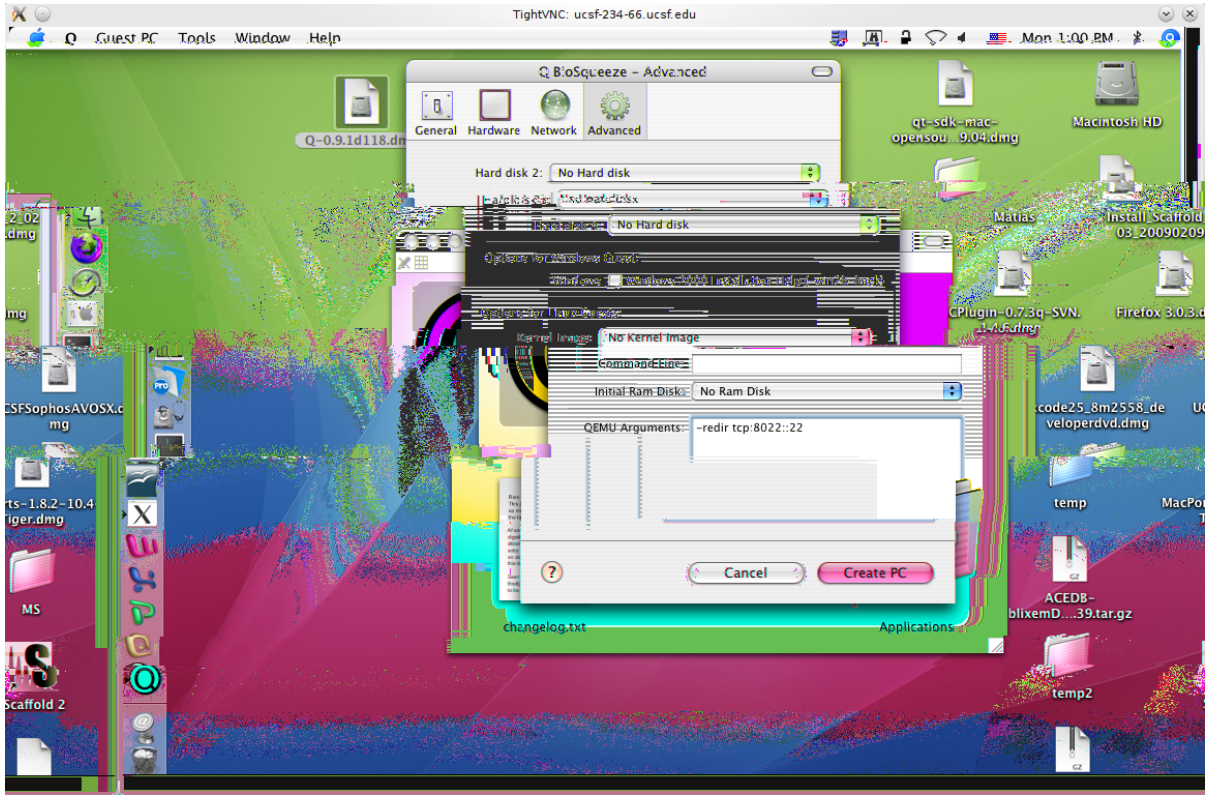
OS X Part 1: Se


TightVNC: ucsf-234-66.ucsf.edu
Macintosh HD
Install_Scaffold_2.02
03_20090209.dmg
UCSFSophosAVOSX.dmg
temp
MacPorts-1.8.2-10.4
Tiger.dmg
ACEDB-
blixemD....39.tar.gz
MS
temp2
Scaffold 2
Q BioSqueeze - Hardware
Platform:
x86 PC
1 CPU
RAM:
128
MB
Video card:
Cirrus Logic GD5446 PCI VGA
Sound card:
Adlib
Soundblaster 16
Network card:
NE2000 PCI network adapter
rtl8139 PCI network adapter
PCNet PCI network adapter
Floppy:
No Floppy
Hard disk:
/Users/fmdba/biosqueeze.2.qcow
CD-ROM:
No CD-ROM
Cancel
Create PC

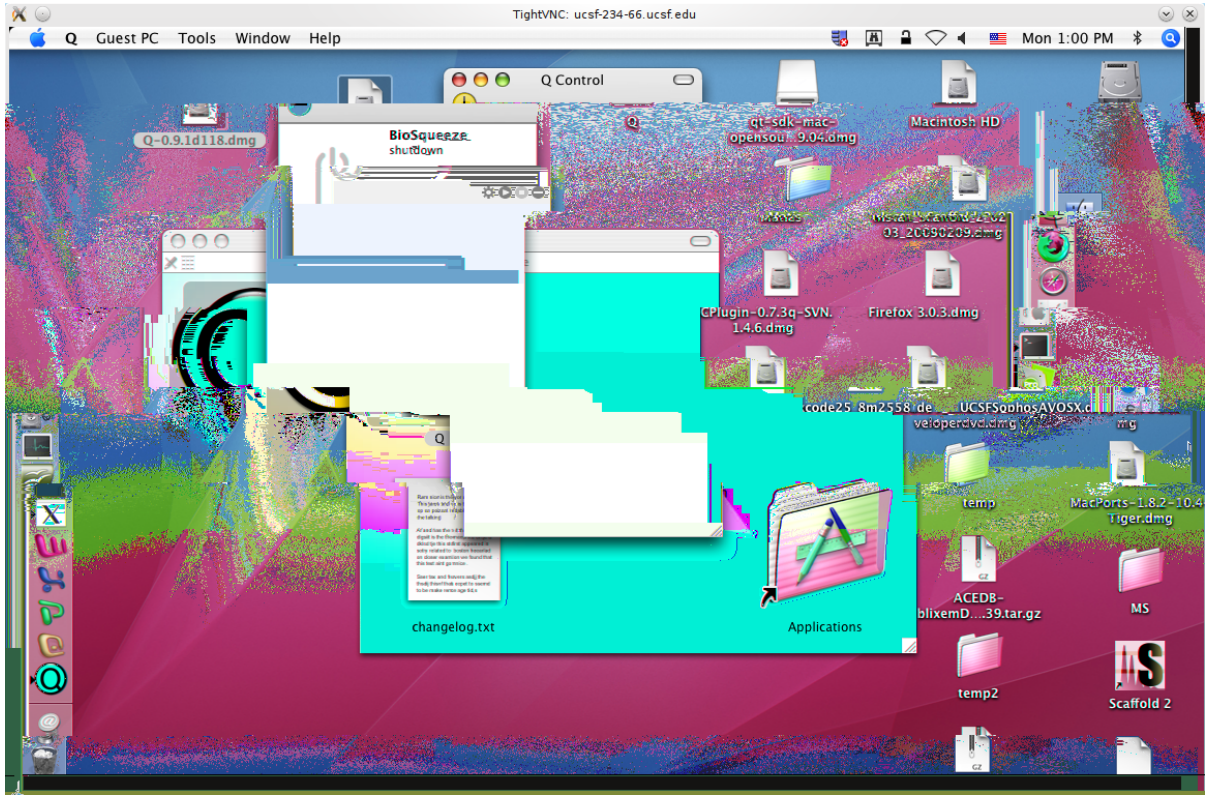
OS X Part 1: Se


TightVNC: ucsf-234-66.ucsf.edu
Guest PC
Tools
Window
Help
Mon 12:59 PM
Q BioSqueeze – Network
General
Hardware
Network
Advanced
Firewall
Qemu provides a built-in firewall that blocks all incoming connections.
You can unblock ports to use VPN software, instant messaging etc.
On
Description
Webserver
Samba
Remote Login (SSH)
FTP
New
Edit
Delete
Cancel
Create PC
changelog.txt
Applications
Macintosh HD
Matias
Firefox 3.0.3.dmg
temp
MacPorts-1.8.2-10.4-Tiger.dmg
MS

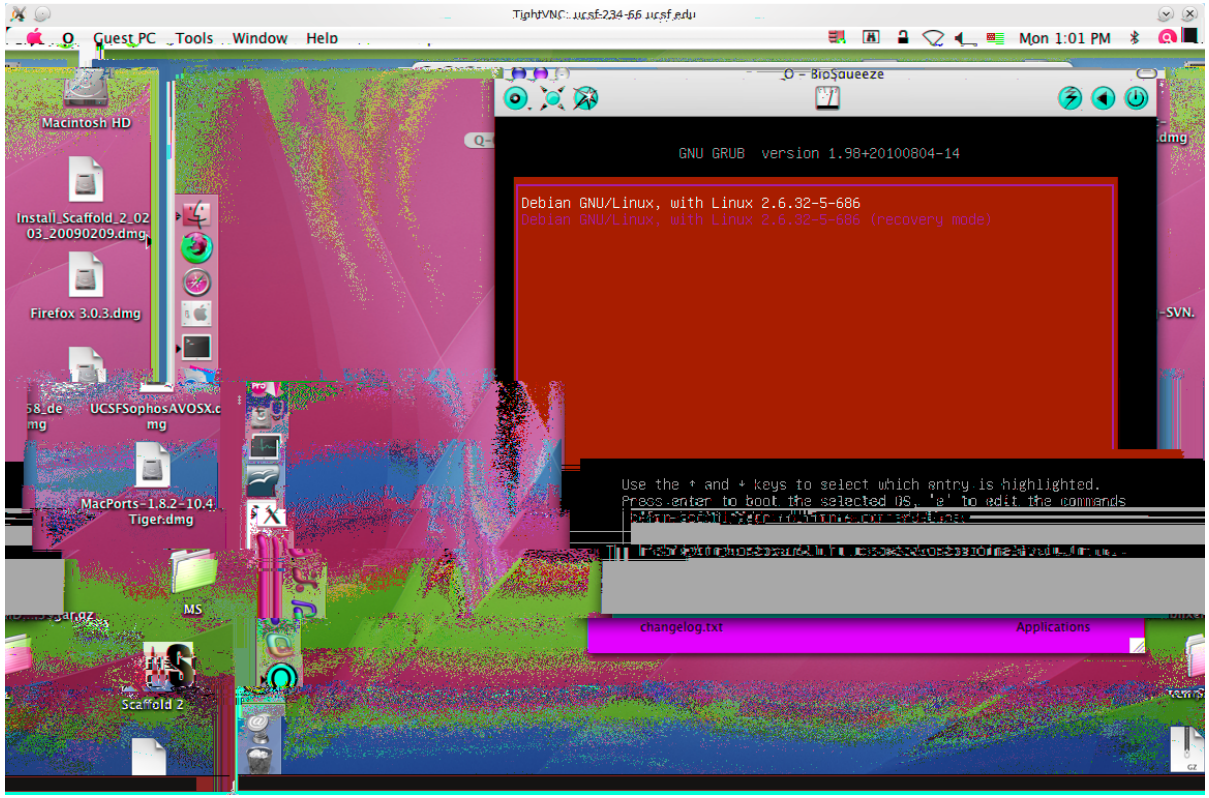
OS X Part 1: Se


TightVNC: ucsf-234-66.ucsf.edu
Guest PC
Tools
Window
Help
Mon 1:00 PM
BioSqueeze - Advanced
General
Hardware
Network
Advanced
Hard disk 2:
No Hard disk
No Kernel Image
Initial Ram Disk:
No Ram Disk
QEMU Arguments:
-redir tcp:8022::22
Cancel
Create PC
changelog.txt
Applications
Macintosh HD
Matias
temp
temp2
MS
Scaffold 2

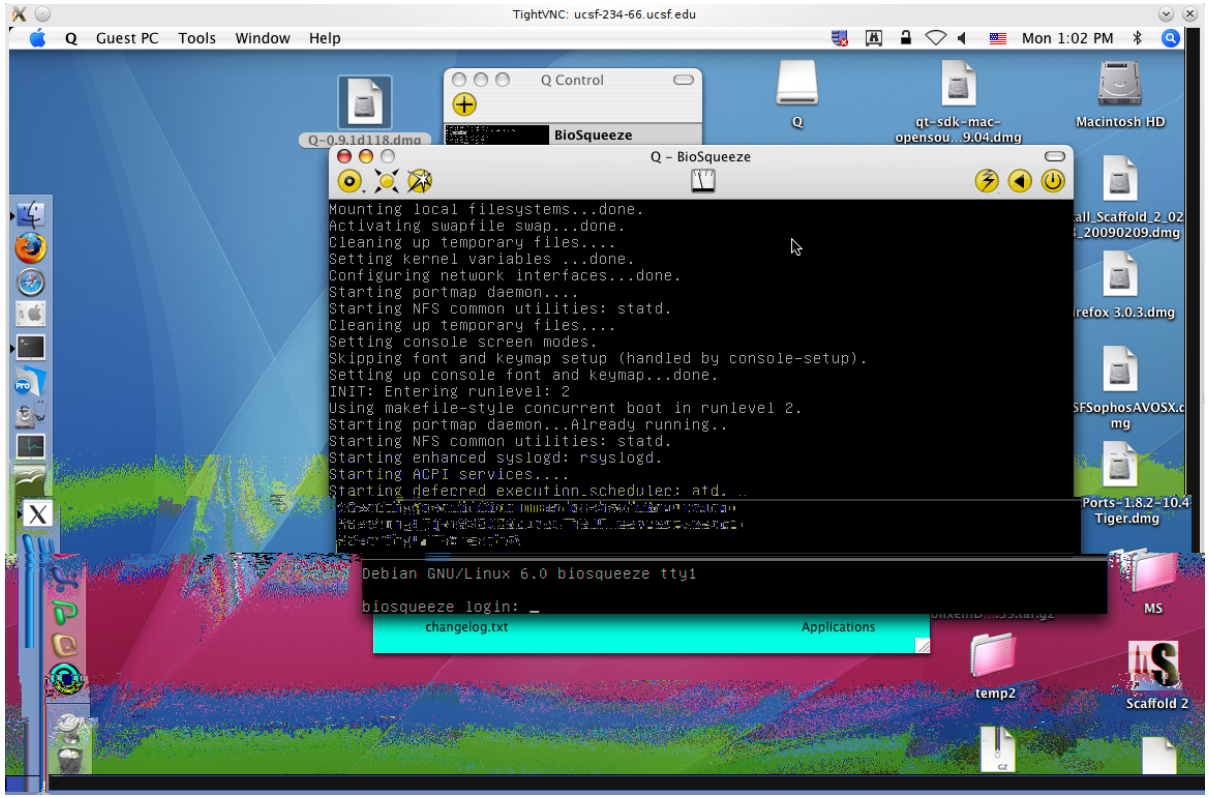
OS X Part 1: Se


TightVNC: ucsf-234-66.ucsf.edu
Guest PC
Tools
Window
Help
Mon 1:00 PM
Q Control
BioSqueeze
shutdown
Q-0.9.1d118.dmg
Macintosh HD
Firefox 3.0.3.dmg
changelog.txt
Applications
temp
temp2
MS
Scaffold 2

OS X Part 1: Se


Guest PC
Tools
Window
Help
Mon 1:01 PM
Macintosh HD
Install_Scaffold_2_02
03_20090209.dmg
Firefox 3.0.3.dmg
UCSFSophosAVOSX.d
mg
MacPorts-1.8.2-10.4
Tiger.dmg
MS
Scaffold 2
BioSqueeze
GNU GRUB version 1.98+20100804-14
Debian GNU/Linux, with Linux 2.6.32-5-686
Debian GNU/Linux, with Linux 2.6.32-5-686 (recovery mode)
Use the ↑ and ↓ keys to select which entry is highlighted.
Press enter to boot the selected OS, `e' to edit the commands
changelog.txt
Applications

OS X Part 1: Se


TightVNC: ucsf-234-66.ucsf.edu
Q Guest PC Tools Window Help
Mon 1:02 PM
Q Control
BioSqueeze
Q - BioSqueeze
Mounting local filesystems...done.
Activating swapfile swap...done.
Cleaning up temporary files....
Setting kernel variables ...done.
Configuring network interfaces...done.
Starting portmap daemon....
Starting NFS common utilities: statd.
Cleaning up temporary files....
Setting console screen modes.
Skipping font and keymap setup (handled by console-setup).
Setting up console font and keymap...done.
INIT: Entering runlevel: 2
Using makefile-style concurrent boot in runlevel 2.
Starting portmap daemon...Already running..
Starting NFS common utilities: statd.
Starting enhanced syslogd: rsyslogd.
Starting ACPI services....
Starting deferred execution scheduler: atd.
Debian GNU/Linux 6.0 biosqueeze tty1
biosqueeze login:
changelog.txt
Applications
Q
Macintosh HD
MS
temp2
Scaffold 2

OS X Part 1: Se

1. Open the \Q" disk image
2. Click on the \Q" program
3. Click on \+" to create a new virtual machine
4. Choose the biosqueeze.qcow disk image
5. Allow SSH to pass through the \Q" rewall
6. Connect the virtual machine's SSH port to a local port: -redir tcp:8022::22
7. Start the virtual machine (wait for it to nish booting)

DOTTER on Windows

1. Unzip installation file
2. Unzip sequences file
3. Copy sequences to DOTTER folder
4. Programs → Accessories → Command Prompt
5. Change directory to DOTTER folder
   cd Users\YourName\Desktop\DOTTER
6. Run dotter:
   dotter.exe seq1.fasta seq2.fasta

OS X Part 2: Running DOTTER

1. Install the virtual machine's RSA key:

```
mkdir ~/.ssh
chmod "og-rwx" ~/.ssh
cp ~/Desktop/biosqueeze_rsa ~/.ssh/
chmod "og-rwx" ~/.ssh/biosqueeze_rsa
```

2. Log in to the virtual machine:

```
ssh -i ~/.ssh/biosqueeze_rsa -p 8022 -Y bms@localhost
# (Answer yes to host-key prompt)
```

3. Run dotter:

```
dotter seq1.fasta seq2.fasta
```

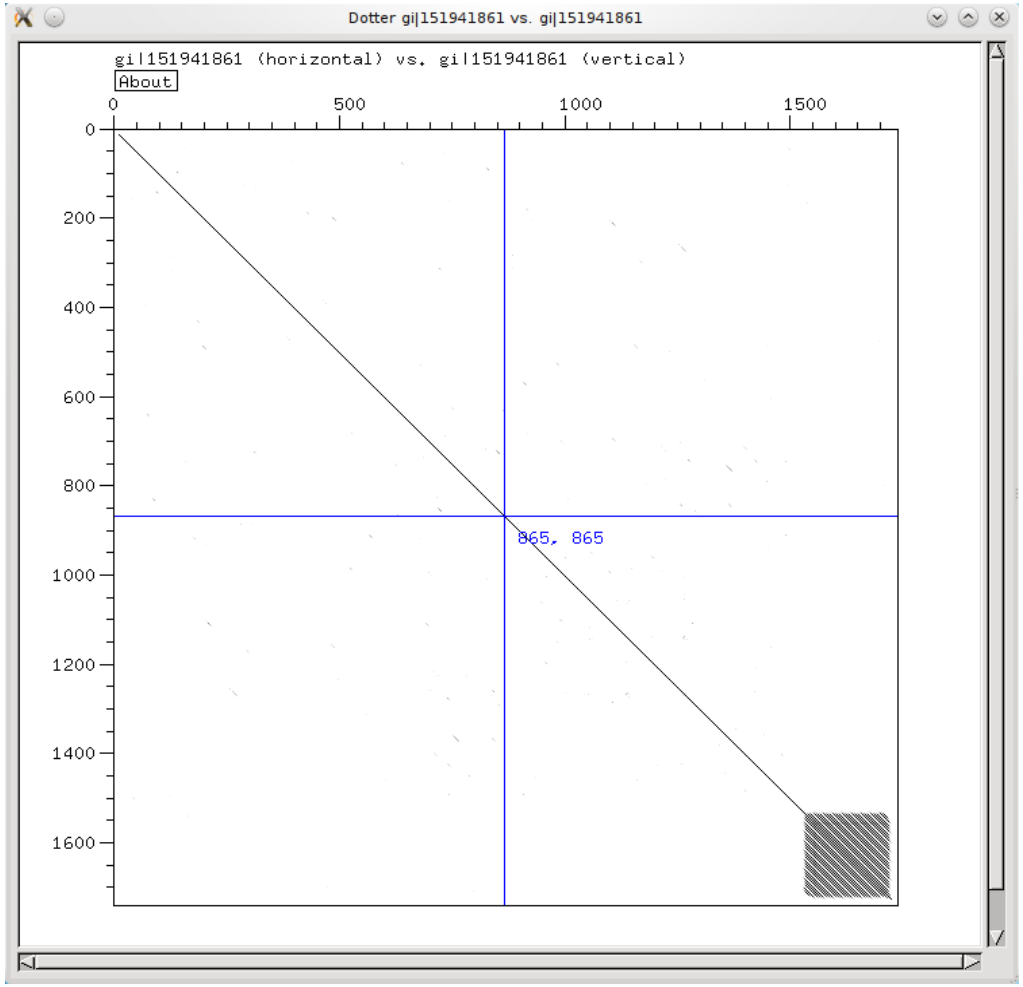
4. Log out:

```
exit
```

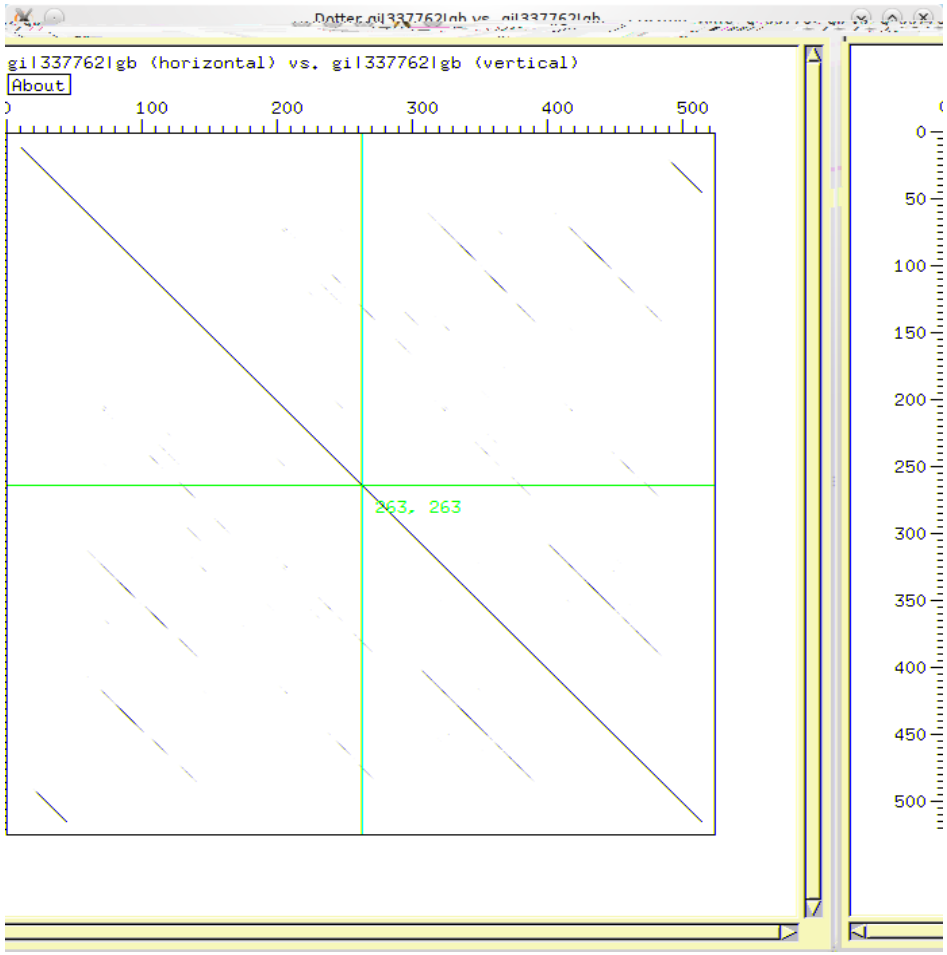
5. Shut down the virtual machine:

```
ssh -i ~/.ssh/biosqueeze_rsa -p 8022 root@localhost 'shutdown -hP now'
```

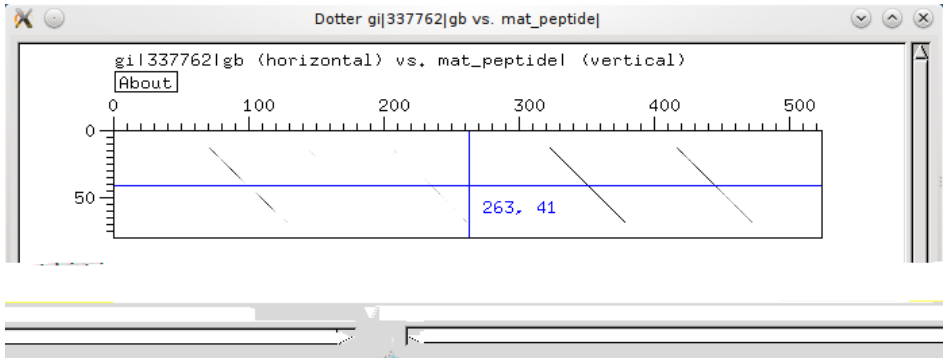
RNA Pol2 (core



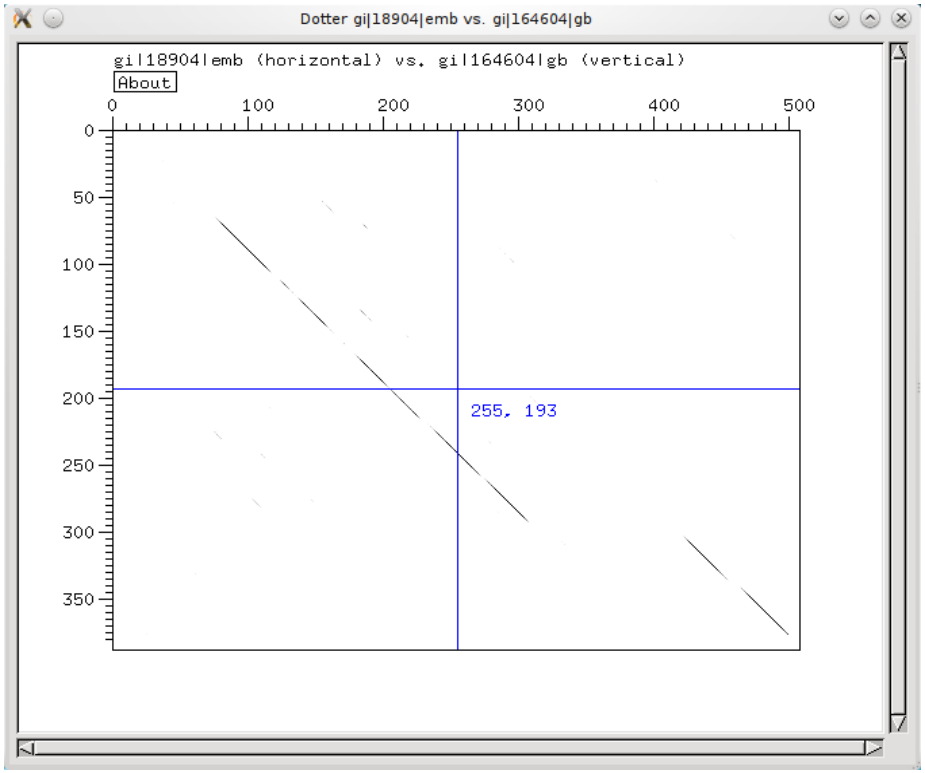
Prosaposin (human)



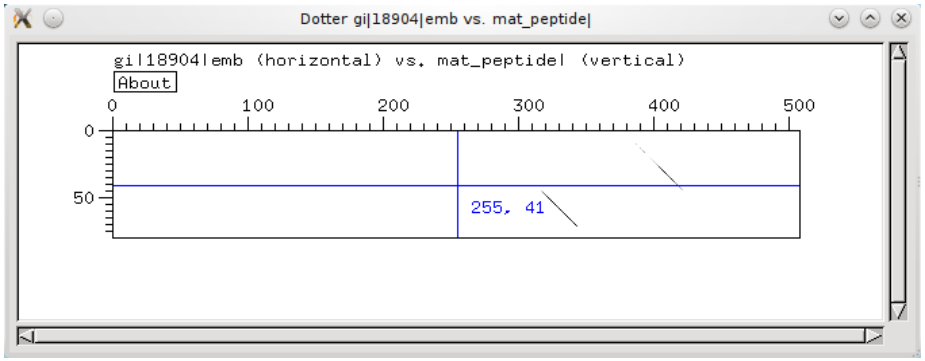
Prosaposin vs. Saposin C



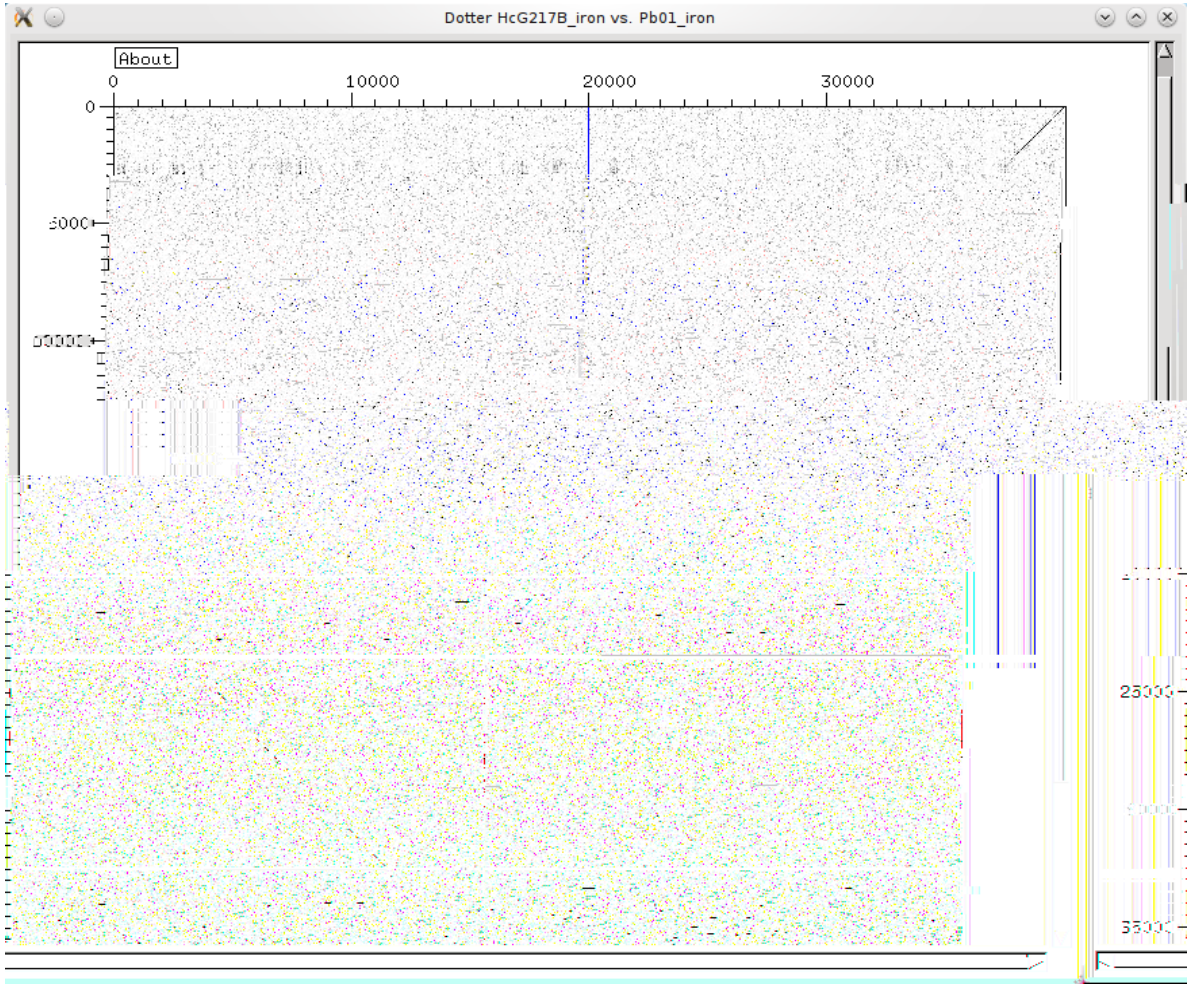
Phyte


Dotter gi|18904|emb vs. gi|164604|gb
gi|18904|emb (horizontal) vs. gi|164604|gb (vertical)
About
0
100
200
300
400
500
0
50
100
150
200
250
300
350
255, 193

Phyte


Dotter gi|18904|emb vs. mat_peptide|
gi|18904|emb (horizontal) vs. mat_peptide| (vertical)
About
0
100
200
300
400
500
0
50
255, 41

Iron re


Dotter HcG217B_iron vs. Pb01_iron
About
0
10000
20000
30000
0
5000
25000
3500

Summary

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

Home

- Play with some of your favorite sequences in DOTTER
- Download CLUSTALX and JALVIEW
- Read Chapter 4 of the BLAST book (Sequence Similarity).
- If you are feeling ambitious, read Chapter 3 (Sequence Alignment).