Data Mining for the Americas: Biowebs

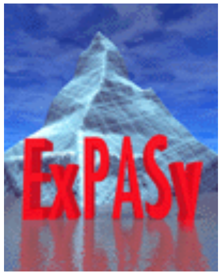
January 29, 2003
AMPATH: Pathway of the Americas

Robert Grossman
University of Illinois at Chicago (grossman at uic.edu)
Open Data (rlg at opendata.biz)

Part 1. Three Trends

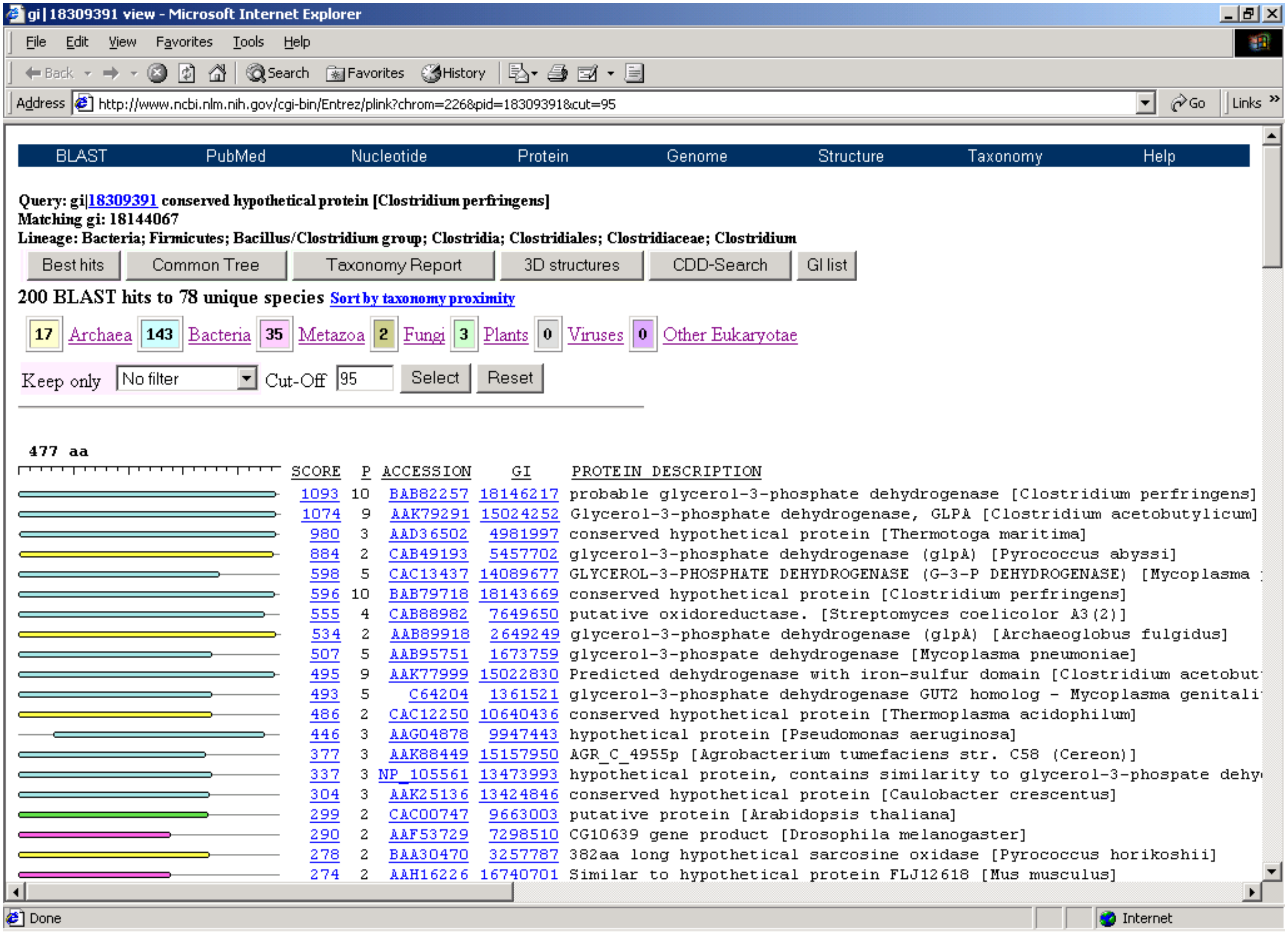
Three trends driving the emergence of biowebs.

Trend 1. Proliferation of Biological Databases



from dozens to
hundreds of databases

…Usually in the Wrong Format


gi|18309391 view - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History
Address http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/plink?chrom=226&pid=18309391&cut=95
Go Links
BLAST PubMed Nucleotide Protein Genome Structure Taxonomy Help
Best hits Common Tree Taxonomy Report 3D structures CDD-Search GI list
Select Reset
Done
Internet

Query: gi|18309391 conserved hypothetical protein [Clostridium perfringens]
Matching gi: 18144067
Lineage: Bacteria; Firmicutes; Bacillus/Clostridium group; Clostridia; Clostridiales; Clostridiaceae; Clostridium

200 BLAST hits to 78 unique species Sort by taxonomy proximity

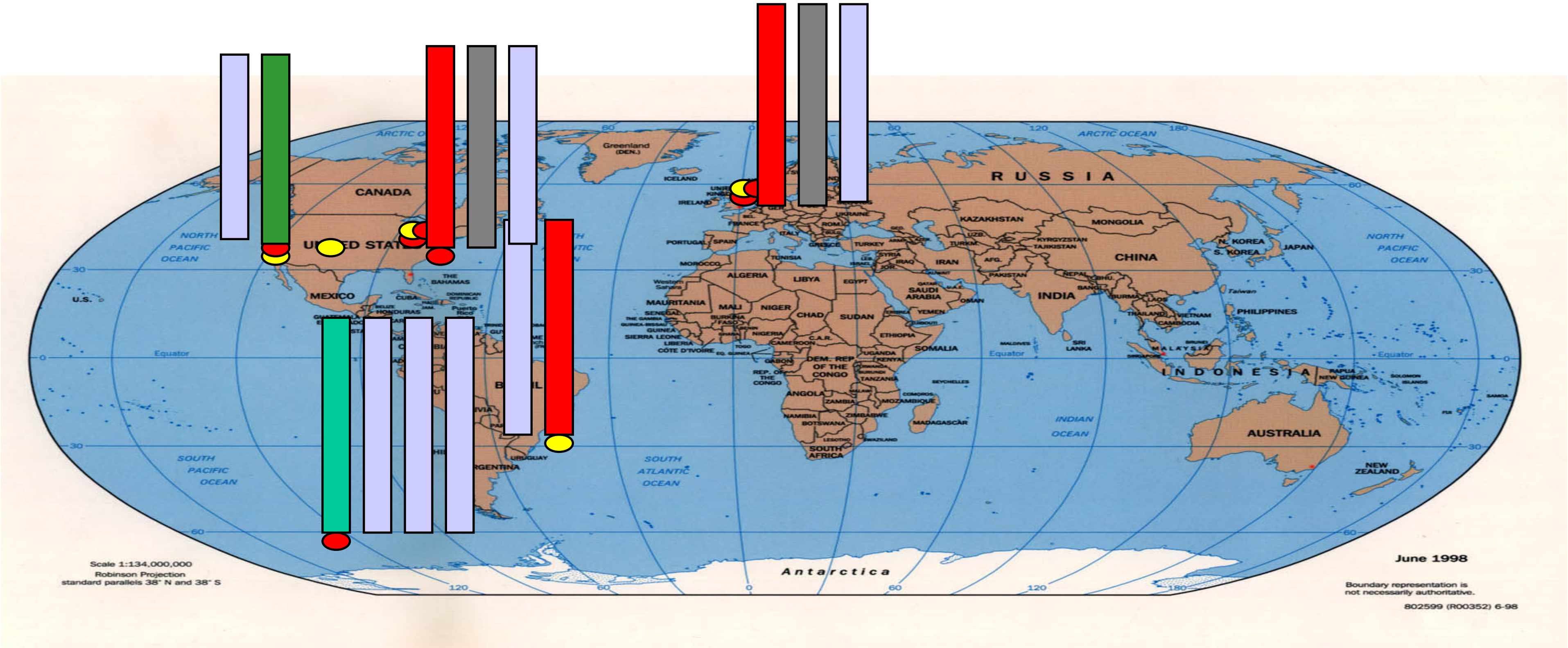
17 Archaea 143 Bacteria 35 Metazoa 2 Fungi 3 Plants 0 Viruses 0 Other Eukaryotae

Keep only No filter Cut-Off 95

477 aa

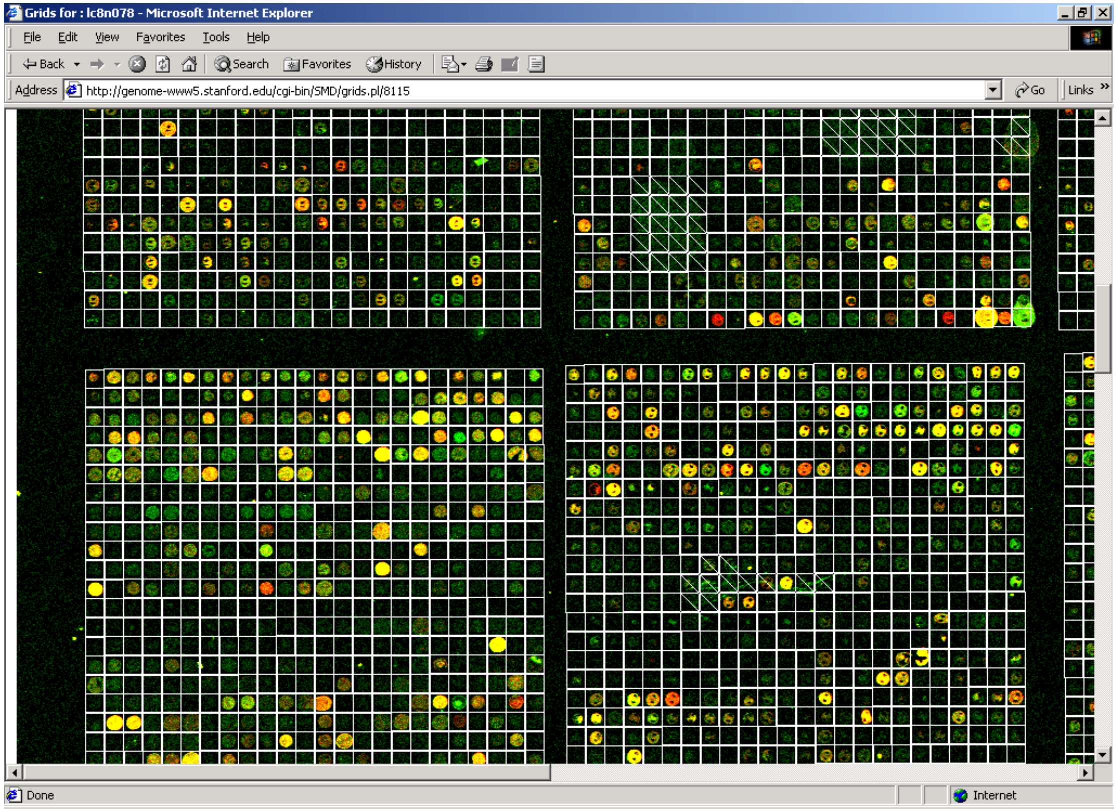
SCORE	P	ACCESSION	GI	PROTEIN DESCRIPTION
1093	10	BAB82257	18146217	probable glycerol-3-phosphate dehydrogenase [Clostridium perfringens]
1074	9	AAK79291	15024252	Glycerol-3-phosphate dehydrogenase, GLPA [Clostridium acetobutylicum]
980	3	AAD36502	4981997	conserved hypothetical protein [Thermotoga maritima]
884	2	CAB49193	5457702	glycerol-3-phosphate dehydrogenase (glpA) [Pyrococcus abyssi]
598	5	CAC13437	14089677	GLYCEROL-3-PHOSPHATE DEHYDROGENASE (G-3-P DEHYDROGENASE) [Mycoplasma
596	10	BAB79718	18143669	conserved hypothetical protein [Clostridium perfringens]
555	4	CAB88982	7649650	putative oxidoreductase. [Streptomyces coelicolor A3(2)]
534	2	AAB89918	2649249	glycerol-3-phosphate dehydrogenase (glpA) [Archaeoglobus fulgidus]
507	5	AAB95751	1673759	glycerol-3-phospate dehydrogenase [Mycoplasma pneumoniae]
495	9	AAK77999	15022830	Predicted dehydrogenase with iron-sulfur domain [Clostridium acetobut
493	5	C64204	1361521	glycerol-3-phosphate dehydrogenase GUT2 homolog - Mycoplasma genitali
486	2	CAC12250	10640436	conserved hypothetical protein [Thermoplasma acidophilum]
446	3	AAG04878	9947443	hypothetical protein [Pseudomonas aeruginosa]
377	3	AAK88449	15157950	AGR_C_4955p [Agrobacterium tumefaciens str. C58 (Cereon)]
337	3	NP_105561	13473993	hypothetical protein, contains similarity to glycerol-3-phospate dehy
304	3	AAK25136	13424846	conserved hypothetical protein [Caulobacter crescentus]
299	2	CAC00747	9663003	putative protein [Arabidopsis thaliana]
290	2	AAF53729	7298510	CG10639 gene product [Drosophila melanogaster]
278	2	BAA30470	3257787	382aa long hypothetical sarcosine oxidase [Pyrococcus horikoshii]
274	2	AAH16226	16740701	Similar to hypothetical protein FLJ12618 [Mus musculus]

Trend 2: More and More Discoveries will be Across Databases



- Pearson’s Law: The usefulness of a column of data varies as the square of the number of columns it is compared to.

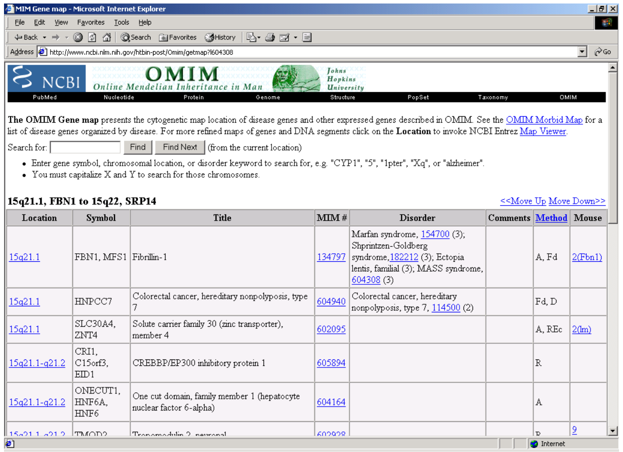
Example:
Microarray Data & Clinical Data



Address http://genome-www5.stanford.edu/cgi-bin/SMD/grids.pl/8115

<publication>
!Citation=Alizadeh AA et al.(2000) Nature 403:503-11
!Title=Distinct types of diffuse large B-cell lymphoma (DLBCL) identified by gene expression profiling.
!PubMedID=10676951

Trend 3. Near a Trifurcation Point



MIM Gene map - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History
Address http://www.ncbi.nlm.nih.gov/htbin-post/Omim/getmap?l604308
Go
NCBI
OMIM
Online Mendelian Inheritance in Man
Johns Hopkins University
PubMed Nucleotide Protein Genome Structure PopSet Taxonomy OMIM
Internet

The OMIM Gene map presents the cytogenetic map location of disease genes and other expressed genes described in OMIM. See the OMIM Morbid Map for a list of disease genes organized by disease. For more refined maps of genes and DNA segments click on the **Location** to invoke NCBI Entrez Map Viewer.

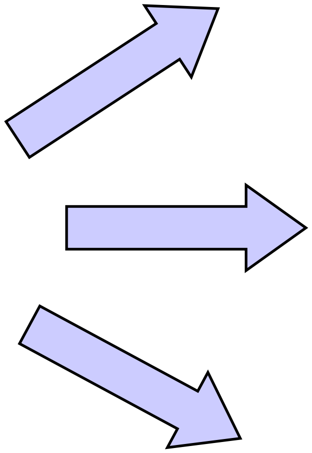
Search for: [ ] Find Find Next (from the current location)

- Enter gene symbol, chromosomal location, or disorder keyword to search for, e.g. "CYP1", "5", "1pter", "Xq", or "alzheimer".
- You must capitalize X and Y to search for those chromosomes.

15q21.1, FBN1 to 15q22, SRP14 <<Move Up Move Down>>

Location	Symbol	Title	MIM #	Disorder	Comments	Method	Mouse
15q21.1	FBN1, MFS1	Fibrillin-1	134797	Marfan syndrome, 154700 (3); Shprintzen-Goldberg syndrome,182212 (3); Ectopia lentis, familial (3); MASS syndrome, 604308 (3)		A, Fd	2(Fbn1)
15q21.1	HNPCC7	Colorectal cancer, hereditary nonpolyposis, type 7	604940	Colorectal cancer, hereditary nonpolyposis, type 7, 114500 (2)		Fd, D	
15q21.1	SLC30A4, ZNT4	Solute carrier family 30 (zinc transporter), member 4	602095			A, REc	2(lm)
15q21.1-q21.2	CRI1, C15orf3, EID1	CREBBP/EP300 inhibitory protein 1	605894			R	
15q21.1-q21.2	ONECUT1, HNF6A, HNF6	One cut domain, family member 1 (hepatocyte nuclear factor 6-alpha)	604164			A	
15q21.1-q21.2	TMOD2	Tropomodulin 2, neuronal	602928			R	9

Biological Databases



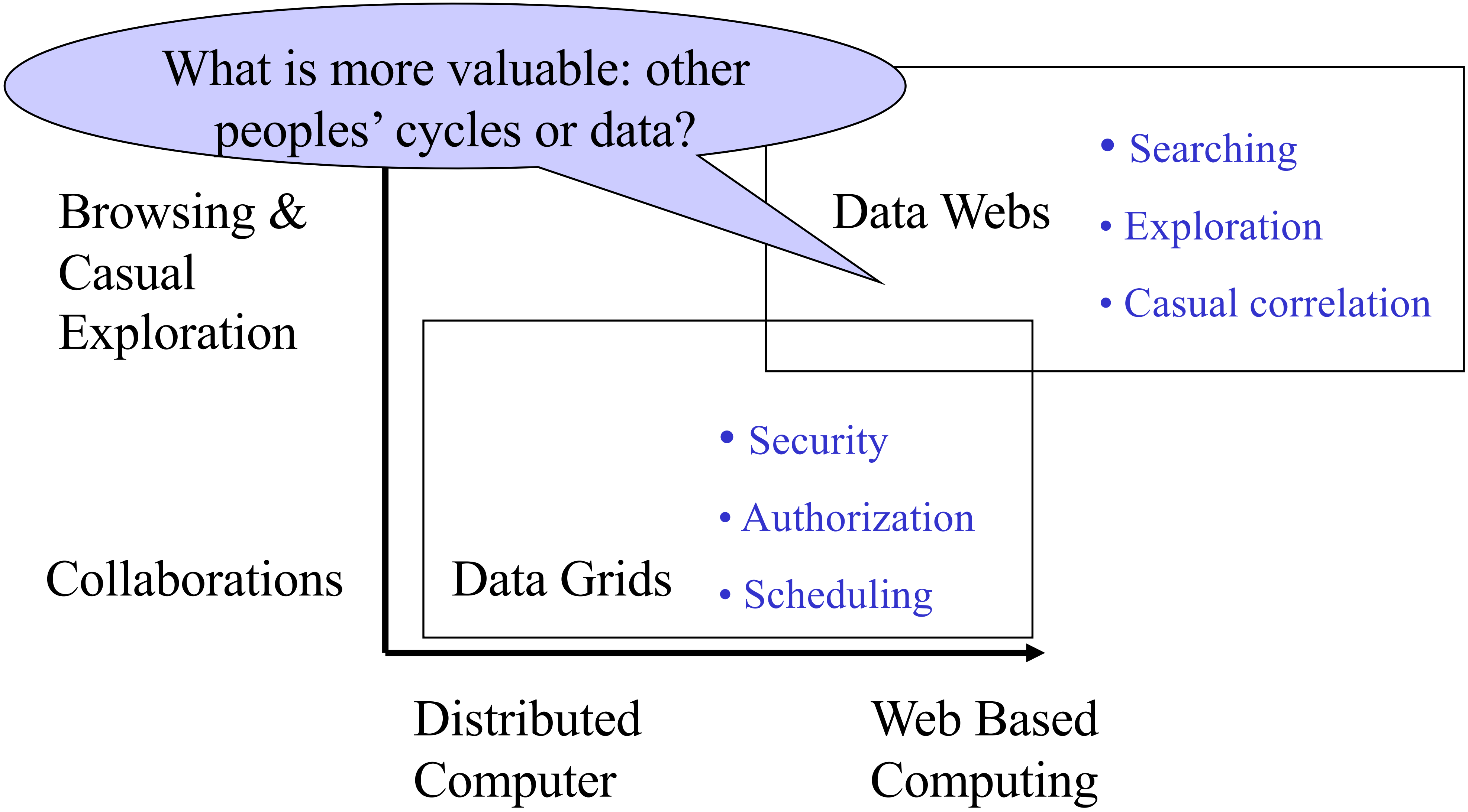
Biowebs – remote data analysis and distributed mining

Biogrids – transparent high end computing

Biological semantic webs – sem. webs for biological knowledge

2003-2008

Data Grids vs. Data Webs



What is more valuable: other peoples' cycles or data?

Browsing & Casual Exploration

Collaborations

Distributed Computer

Web Based Computing

Data Webs

- Searching
- Exploration
- Casual correlation

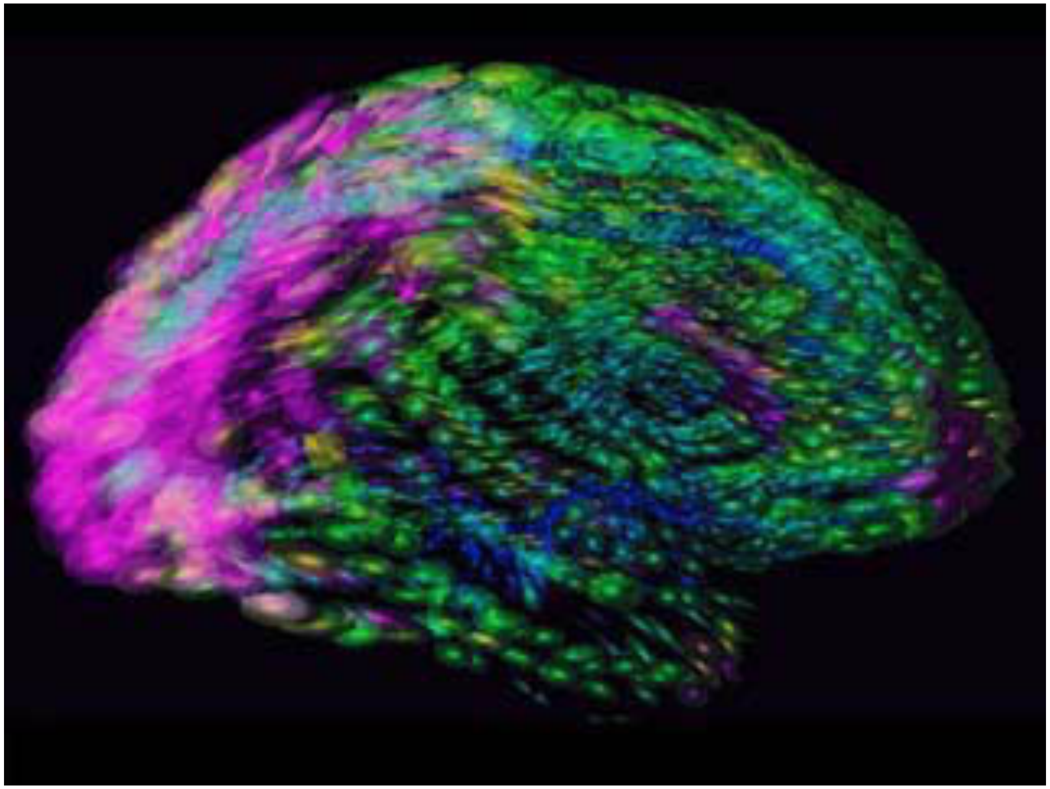
Data Grids

- Security
- Authorization
- Scheduling

Part 2. Examples

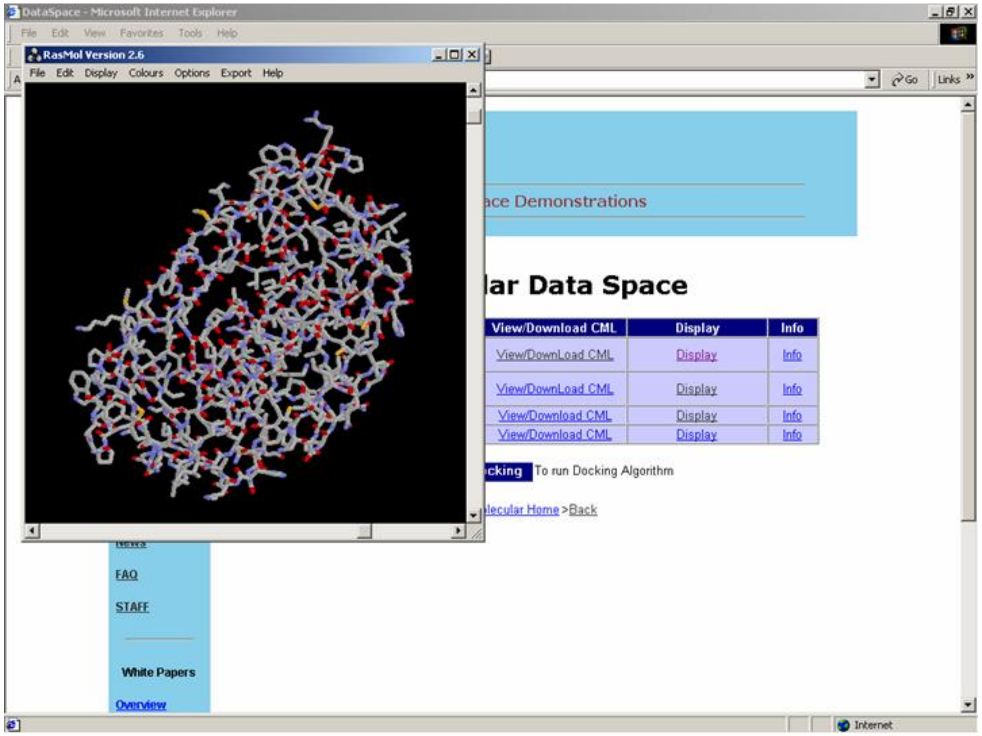
Biogrids, biowebs, and all that.

Example 1. BIRN



- NIH Sponsored project developing collaborative infrastructure for studying brains in humans and animals

Example 2. OptIPuter


DataSpace - Microsoft Internet Explorer
File Edit View Favorites Tools Help
RasMol Version 2.6
File Edit Display Colours Options Export Help
Go
Links
Internet

ace Demonstrations

lar Data Space

View/Download CML	Display	Info
View/DownLoad CML	Display	Info
View/DownLoad CML	Display	Info
View/Download CML	Display	Info
View/Download CML	Display	Info

cking To run Docking Algorithm

lecular Home >Back

FAQ

STAFF

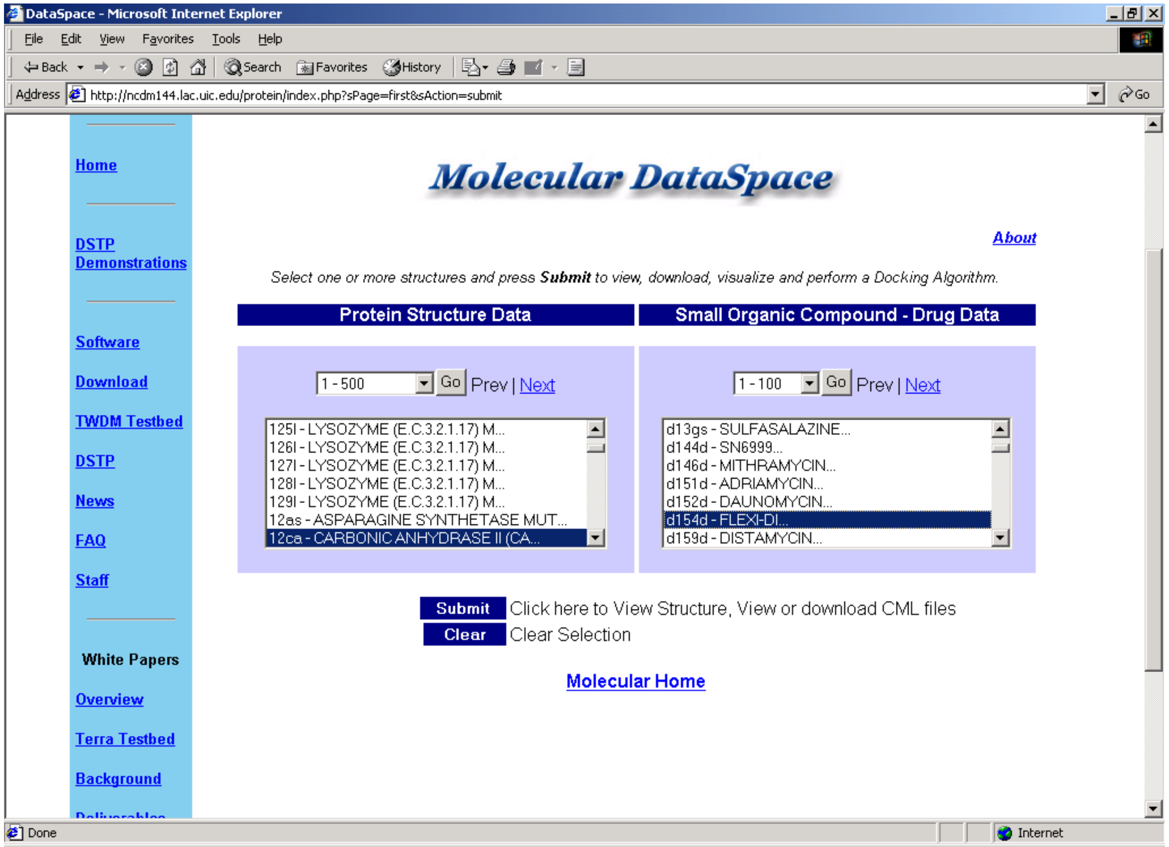
White Papers

Overview

Photonic DataSpace

- ❑ Data intensive computing over photonic networks
- ❑ Replication of the protein data bank (PDB).
- ❑ Linked with a chemical library of small organics molecules.
- ❑ Distributed docking algorithms

Example 3. Molecular DataSpace



http://ncdm144.lac.uic.edu/protein/index.php?sPage=first&sAction=submit

Molecular DataSpace

*Select one or more structures and press **Submit** to view, download, visualize and perform a Docking Algorithm.*

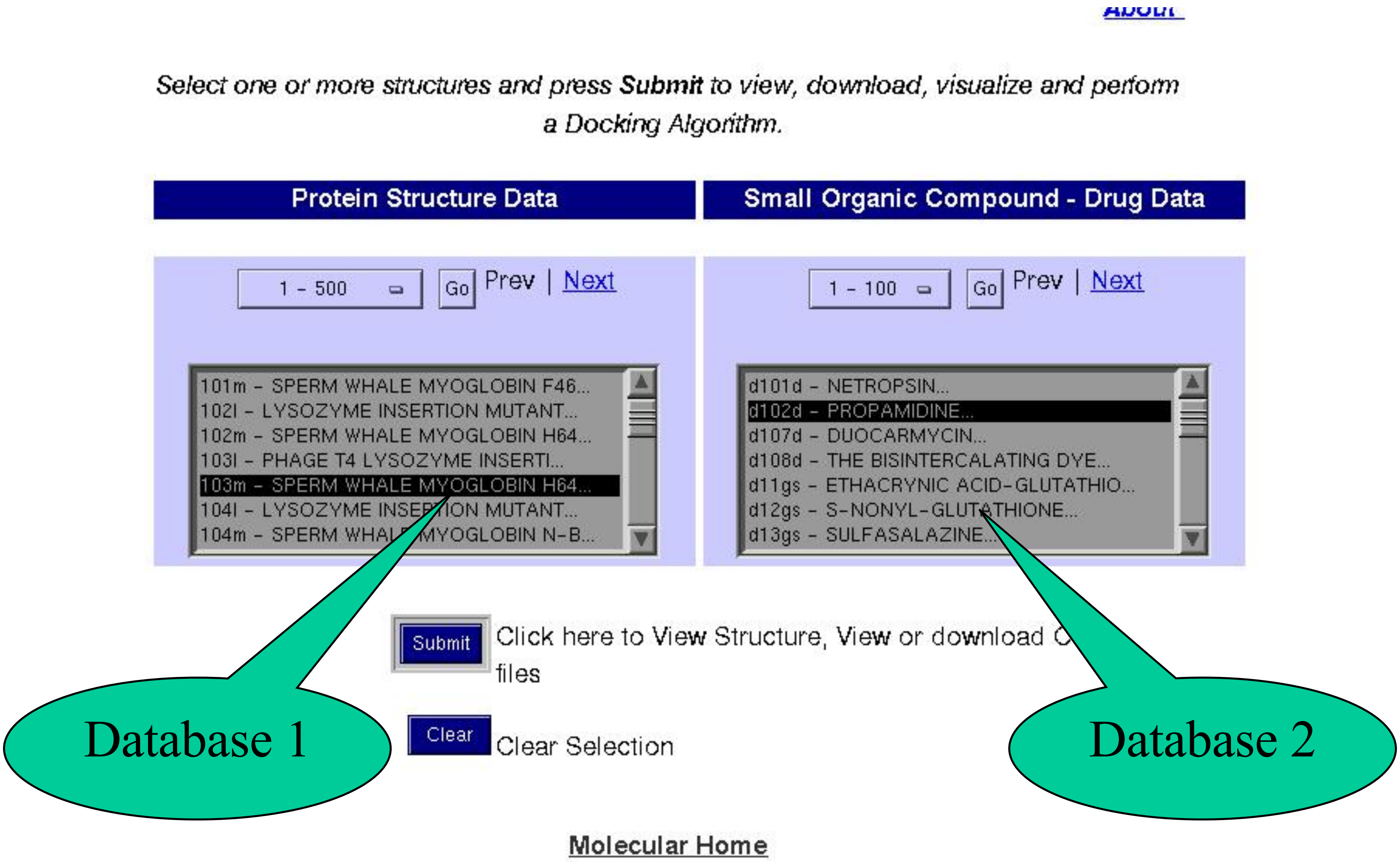
Protein Structure Data	Small Organic Compound - Drug Data
1 - 500 Go Prev | Next	1 - 100 Go Prev | Next
125l - LYSOZYME (E.C.3.2.1.17) M... 126l - LYSOZYME (E.C.3.2.1.17) M... 127l - LYSOZYME (E.C.3.2.1.17) M... 128l - LYSOZYME (E.C.3.2.1.17) M... 129l - LYSOZYME (E.C.3.2.1.17) M... 12as - ASPARAGINE SYNTHETASE MUT... 12ca - CARBONIC ANHYDRASE II (CA...	d13gs - SULFASALAZINE... d144d - SN6999... d146d - MITHRAMYCIN... d151d - ADRIAMYCIN... d152d - DAUNOMYCIN... d154d - FLEXI-DI... d159d - DISTAMYCIN...

Submit Click here to View Structure, View or download CML files

Clear Clear Selection

- ❑ How do you interactively explore other people’s data?
- ❑ How do you overlay other peoples data on your own?
- ❑ How do you do distributed data mining?

Simplify the Integration of Two or More Distributed Data Sets



*Select one or more structures and press **Submit** to view, download, visualize and perform a Docking Algorithm.*

Protein Structure Data

101m - SPERM WHALE MYOGLOBIN F46...
102l - LYSOZYME INSERTION MUTANT...
102m - SPERM WHALE MYOGLOBIN H64...
103l - PHAGE T4 LYSOZYME INSERTI...
103m - SPERM WHALE MYOGLOBIN H64...
104l - LYSOZYME INSERTION MUTANT...
104m - SPERM WHALE MYOGLOBIN N-B...

Small Organic Compound - Drug Data

d101d - NETROPSIN...
d102d - PROPAMIDINE...
d107d - DUOCARMYCIN...
d108d - THE BISINTERCALATING DYE...
d11gs - ETHACRYNIC ACID-GLUTATHIO...
d12gs - S-NONYL-GLUTATHIONE...
d13gs - SULFASALAZINE...

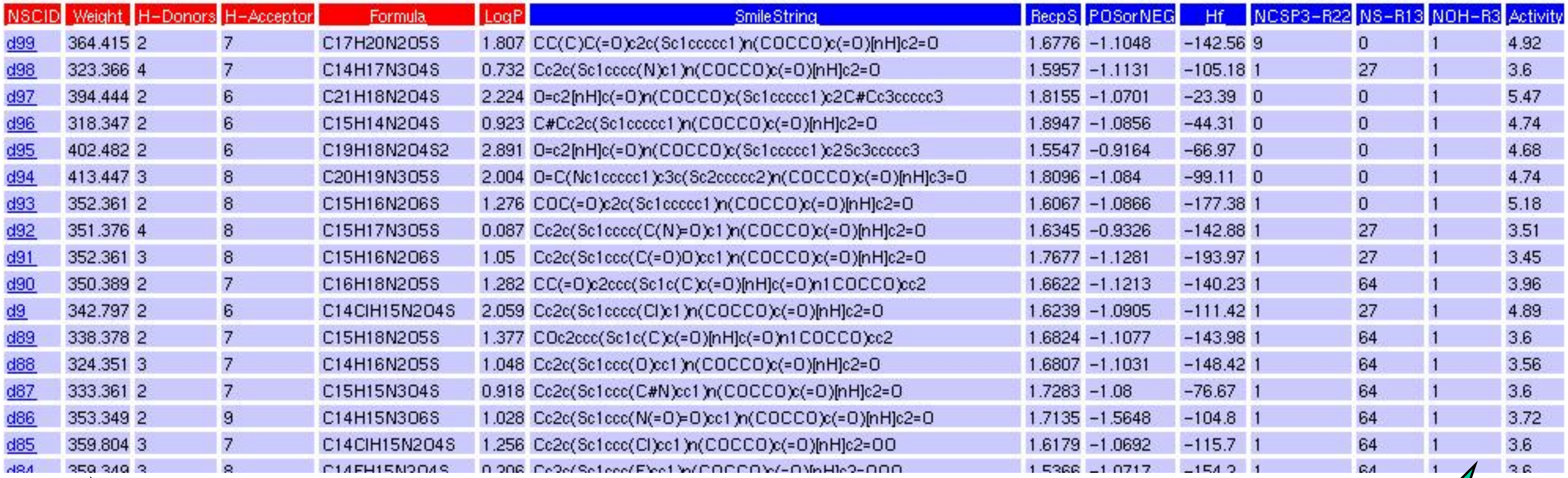
Submit Click here to View Structure, View or download C files

Clear Clear Selection

Database 1

Database 2

Easy to Overlay External, Third Party Data with Local Data

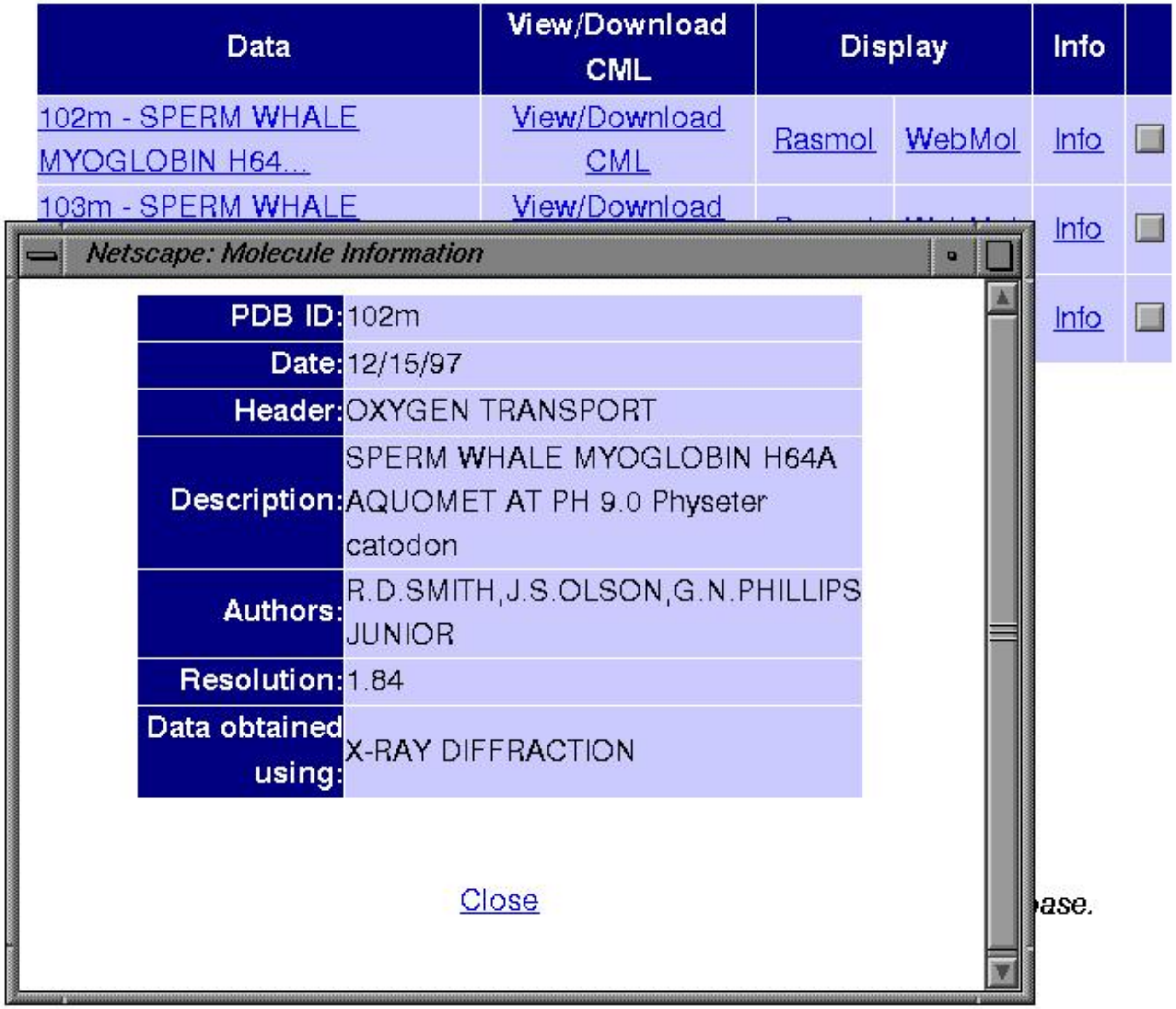


NSCID	Weight	H-Donors	H-Acceptor	Formula	LogP	SmileString	RecpS	POSorNEG	Hf	NCSP3-R22	NS-R13	NOH-R3	Activity
d99	364.415	2	7	C17H20N2O5S	1.807	CC(C)C(=O)c2c(Sc1ccccc1)n(COCCO)c(=O)[nH]c2=O	1.6776	-1.1048	-142.56	9	0	1	4.92
d98	323.366	4	7	C14H17N3O4S	0.732	Cc2c(Sc1cccc(N)c1)n(COCCO)c(=O)[nH]c2=O	1.5957	-1.1131	-105.18	1	27	1	3.6
d97	394.444	2	6	C21H18N2O4S	2.224	O=c2[nH]c(=O)n(COCCO)c(Sc1ccccc1)c2C#Cc3ccccc3	1.8155	-1.0701	-23.39	0	0	1	5.47
d96	318.347	2	6	C15H14N2O4S	0.923	C#Cc2c(Sc1ccccc1)n(COCCO)c(=O)[nH]c2=O	1.8947	-1.0856	-44.31	0	0	1	4.74
d95	402.482	2	6	C19H18N2O4S2	2.891	O=c2[nH]c(=O)n(COCCO)c(Sc1ccccc1)c2Sc3ccccc3	1.5547	-0.9164	-66.97	0	0	1	4.68
d94	413.447	3	8	C20H19N3O5S	2.004	O=C(Nc1ccccc1)c3c(Sc2ccccc2)n(COCCO)c(=O)[nH]c3=O	1.8096	-1.084	-99.11	0	0	1	4.74
d93	352.361	2	8	C15H16N2O6S	1.276	COC(=O)c2c(Sc1ccccc1)n(COCCO)c(=O)[nH]c2=O	1.6067	-1.0866	-177.38	1	0	1	5.18
d92	351.376	4	8	C15H17N3O5S	0.087	Cc2c(Sc1cccc(C(N)=O)c1)n(COCCO)c(=O)[nH]c2=O	1.6345	-0.9326	-142.88	1	27	1	3.51
d91	352.361	3	8	C15H16N2O6S	1.05	Cc2c(Sc1ccc(C(=O)O)cc1)n(COCCO)c(=O)[nH]c2=O	1.7677	-1.1281	-193.97	1	27	1	3.45
d90	350.389	2	7	C16H18N2O5S	1.282	CC(=O)c2ccc(Sc1c(C)c(=O)[nH]c(=O)n1COCCO)cc2	1.6622	-1.1213	-140.23	1	64	1	3.96
d9	342.797	2	6	C14ClH15N2O4S	2.059	Cc2c(Sc1cccc(Cl)c1)n(COCCO)c(=O)[nH]c2=O	1.6239	-1.0905	-111.42	1	27	1	4.89
d89	338.378	2	7	C15H18N2O5S	1.377	COc2ccc(Sc1c(C)c(=O)[nH]c(=O)n1COCCO)cc2	1.6824	-1.1077	-143.98	1	64	1	3.6
d88	324.351	3	7	C14H16N2O5S	1.048	Cc2c(Sc1ccc(O)cc1)n(COCCO)c(=O)[nH]c2=O	1.6807	-1.1031	-148.42	1	64	1	3.56
d87	333.361	2	7	C15H15N3O4S	0.918	Cc2c(Sc1ccc(C#N)cc1)n(COCCO)c(=O)[nH]c2=O	1.7283	-1.08	-76.67	1	64	1	3.6
d86	353.349	2	9	C14H15N3O6S	1.028	Cc2c(Sc1ccc(N(=O)=O)cc1)n(COCCO)c(=O)[nH]c2=O	1.7135	-1.5648	-104.8	1	64	1	3.72
d85	359.804	3	7	C14ClH15N2O4S	1.256	Cc2c(Sc1ccc(Cl)cc1)n(COCCO)c(=O)[nH]c2=OO	1.6179	-1.0692	-115.7	1	64	1	3.6
d84	359.349	3	8	C14FH15N2O4S	0.206	Cc2c(Sc1ccc(F)cc1)n(COCCO)c(=O)[nH]c2=OOO	1.5366	-1.0717	-154.2	1	64	1	3.6

Local Database 1

External Database 2

Data and Metadata Separated

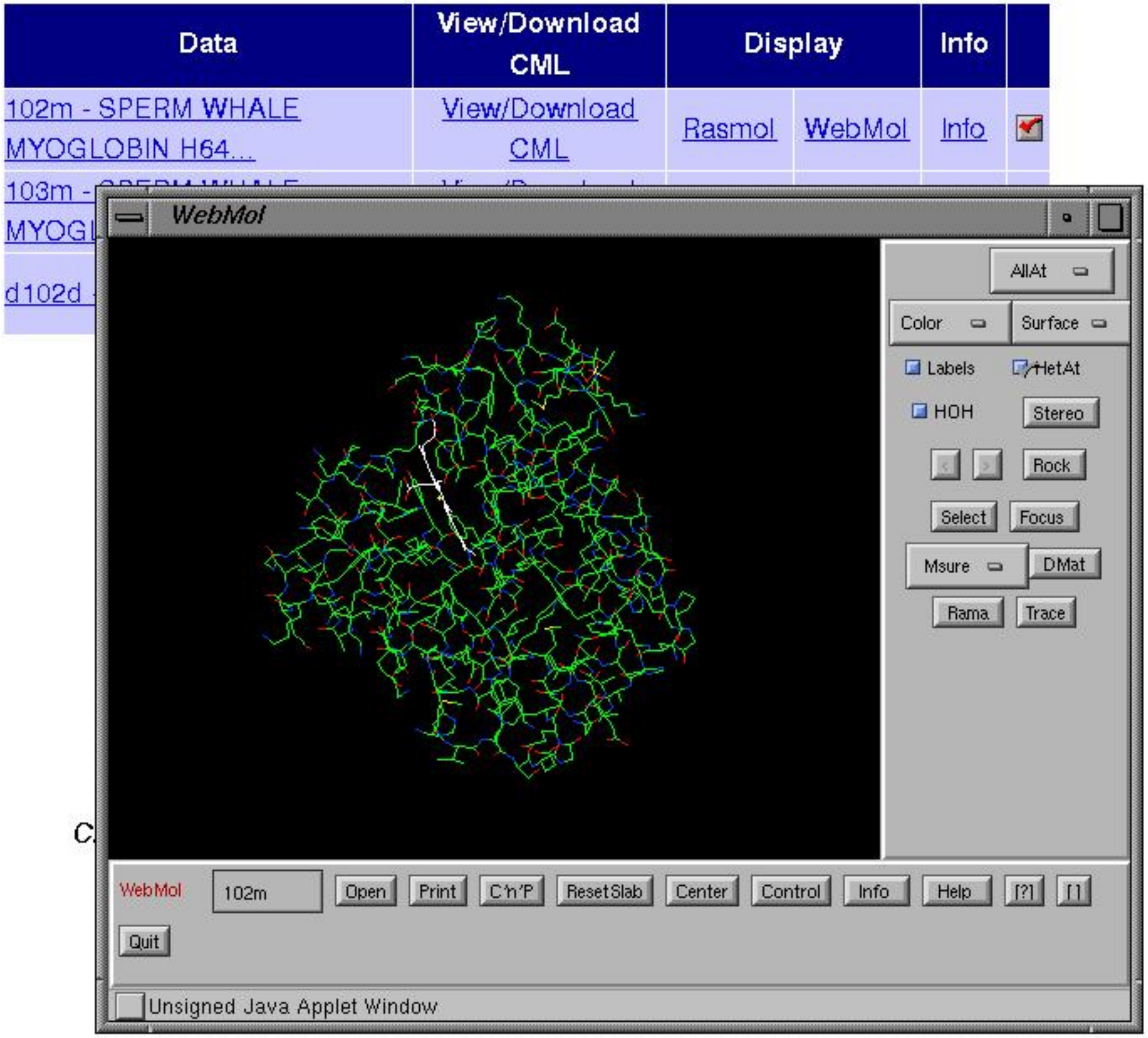


Data	View/Download CML	Display		Info	
102m - SPERM WHALE MYOGLOBIN H64...	View/Download CML	Rasmol	WebMol	Info	□
103m - SPERM WHALE	View/Download			Info	□
				Info	□

ase.

PDB ID:	102m
Date:	12/15/97
Header:	OXYGEN TRANSPORT
Description:	SPERM WHALE MYOGLOBIN H64A AQUOMET AT PH 9.0 Physeter catodon
Authors:	R.D.SMITH,J.S.OLSON,G.N.PHILLIPS JUNIOR
Resolution:	1.84
Data obtained using:	X-RAY DIFFRACTION

Data Can be Streamed…



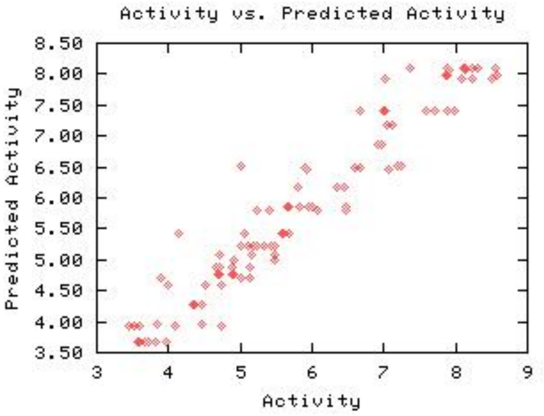
Data	View/Download CML	Display		Info	
102m - SPERM WHALE MYOGLOBIN H64...	View/Download CML	Rasmol	WebMol	Info	✓
103m - MYOGL					
d102d					

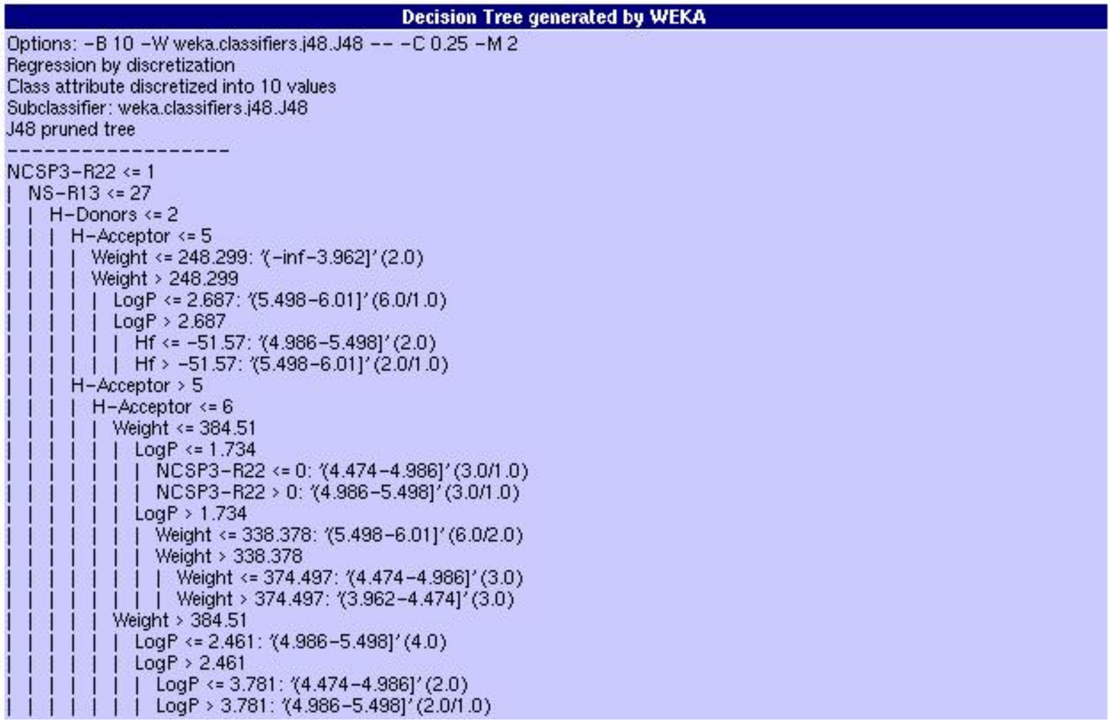
WebMol

102m

Unsigned Java Applet Window

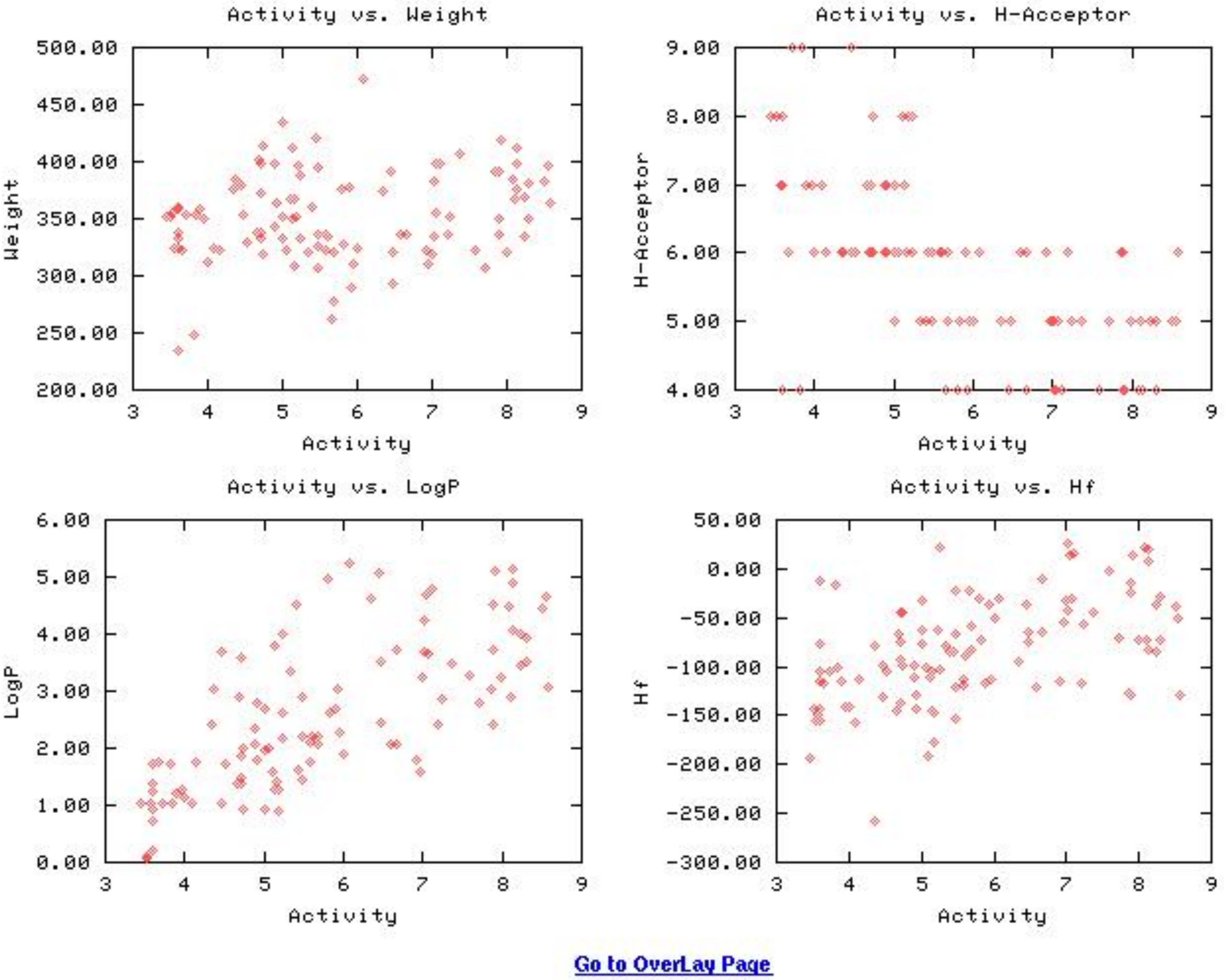
Support PMML Based Analytics




Decision Tree generated by WEKA

```
Options: -B 10 -W weka.classifiers.j48.J48 -- -C 0.25 -M 2
Regression by discretization
Class attribute discretized into 10 values
Subclassifier: weka.classifiers.j48.J48
J48 pruned tree
------------------
NCSP3-R22 <= 1
|   NS-R13 <= 27
|   |   H-Donors <= 2
|   |   |   H-Acceptor <= 5
|   |   |   |   Weight <= 248.299: '(-inf-3.962]' (2.0)
|   |   |   |   Weight > 248.299
|   |   |   |   |   LogP <= 2.687: '(5.498-6.01]' (6.0/1.0)
|   |   |   |   |   LogP > 2.687
|   |   |   |   |   |   Hf <= -51.57: '(4.986-5.498]' (2.0)
|   |   |   |   |   |   Hf > -51.57: '(5.498-6.01]' (2.0/1.0)
|   |   |   H-Acceptor > 5
|   |   |   |   H-Acceptor <= 6
|   |   |   |   |   Weight <= 384.51
|   |   |   |   |   |   LogP <= 1.734
|   |   |   |   |   |   |   NCSP3-R22 <= 0: '(4.474-4.986]' (3.0/1.0)
|   |   |   |   |   |   |   NCSP3-R22 > 0: '(4.986-5.498]' (3.0/1.0)
|   |   |   |   |   |   LogP > 1.734
|   |   |   |   |   |   |   Weight <= 338.378: '(5.498-6.01]' (6.0/2.0)
|   |   |   |   |   |   |   Weight > 338.378
|   |   |   |   |   |   |   |   Weight <= 374.497: '(4.474-4.986]' (3.0)
|   |   |   |   |   |   |   |   Weight > 374.497: '(3.962-4.474]' (3.0)
|   |   |   |   |   Weight > 384.51
|   |   |   |   |   |   LogP <= 2.461: '(4.986-5.498]' (4.0)
|   |   |   |   |   |   LogP > 2.461
|   |   |   |   |   |   |   LogP <= 3.781: '(4.474-4.986]' (2.0)
|   |   |   |   |   |   |   LogP > 3.781: '(4.986-5.498]' (2.0/1.0)
```

Integrate with Open Source Analytics



Part 3. Strategy for the Americas

It’s the data.

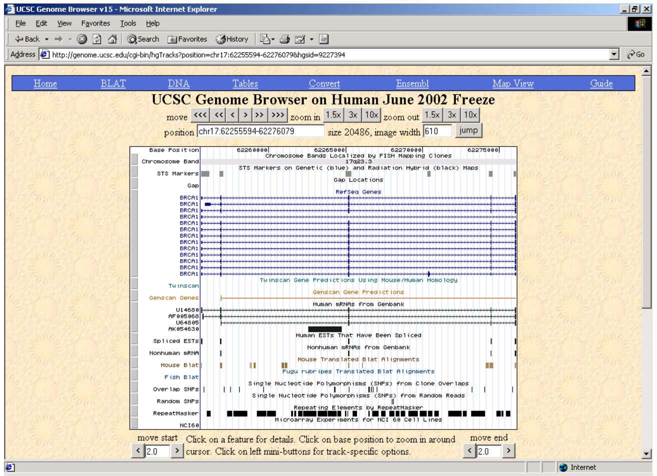
1. Think Small, Medium and Large

- ❑ Large Science: high energy physics, astronomy, …
- ❑ Medium science: sequencing an organism, biodiversity surveys, …
- ❑ Small science: creating interesting bioinformatics databases and resources, overlaying external data over your data to do new science.

2. Open Data:
Free & Controlled Biodata

- ❑ Lawrence Lessig of Stanford Law School has highlighted the battle between “free” and “controlled” web resources
- ❑ Genbank created a culture of free sequence data
- ❑ There is also a culture of proprietary data
- ❑ Consider part of your mission to create ***open data repositories***

Open Source Projects



UCSC Genome Browser v15 - Microsoft Internet Explorer

Address http://genome.ucsc.edu/cgi-bin/hgTracks?position=chr17:62255594-62276079&hgsid=9227394

UCSC Genome Browser on Human June 2002 Freeze

position chr17:62255594-62276079 size 20486, image width 610

move start 2.0 Click on a feature for details. Click on base position to zoom in around cursor. Click on left mini-buttons for track-specific options. move end 2.0

File Edit View Favorites Tools Help
Back Search Favorites History Go
Home BLAT DNA Tables Convert Ensembl Map View Guide
move <<< << < > >> >>> zoom in 1.5x 3x 10x zoom out 1.5x 3x 10x
jump
Base Position 62260000 62265000 62270000 62275000
Chromosome Bands Localized by FISH Mapping Clones
Chromosome Band 17q23.3
STS Markers on Genetic (blue) and Radiation Hybrid (black) Maps
STS Markers
Gap Locations
Gap
RefSeq Genes
BRCA1
Twinscan Gene Predictions Using Mouse/Human Homology
Twinscan
Genscan Gene Predictions
Genscan Genes
Human mRNAs from Genbank
Human ESTs That Have Been Spliced
Spliced ESTs
Nonhuman mRNAs from Genbank
Nonhuman mRNA
Mouse Translated Blat Alignments
Mouse Blat
Fugu rubripes Translated Blat Alignments
Fish Blat
Single Nucleotide Polymorphisms (SNPs) from Clone Overlaps
Overlap SNPs
Single Nucleotide Polymorphisms (SNPs) from Random Reads
Random SNPs
Repeating Elements by RepeatMasker
RepeatMasker
Microarray Experiments for NCI 60 Cell Lines
NCI60
Internet

- ❑ Open Source libraries
  - – Bioperl, Biojava, Biopython
- ❑ Open Source protocols
  - – DWTP, DAS, MOBY, OmniGene, G2G,…
- ❑ Open Source end-user applications
  - – Genquire, Generic Genome Browser, PyMol, Molecular DataSpace…

For More Information

- ❑ Data webs – www.dataspaceweb.net
- ❑ Data web servers – www.sourceforge.net/projects/dataspace
- ❑ Robert Grossman – grossman @ uic.edu or rlg @ opendata.biz