Information Management for Genome Level Bioinformatics

Norman Paton and Carole Goble
Department of Computer Science
University of Manchester
Manchester, UK
<norm, carole>@cs.man.ac.uk

Structure of Tutorial

- Introduction - why it matters.
- Genome level data.
- Genomic databases.
- Modelling challenges.
- Integrating biological databases.
- Analysing genomic data.
- Summary and challenges.

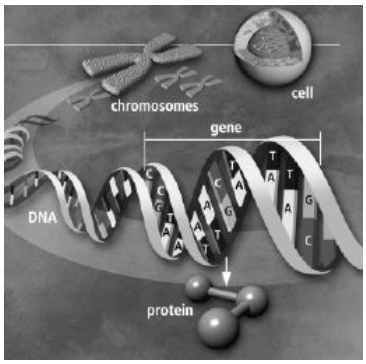
Information Management for Genome Level Bioinformatics

Norman Paton and Carole Goble
Department of Computer Science
University of Manchester
Manchester, UK
<norm, carole>@cs.man.ac.uk

Structure of Tutorial

- Introduction - why it matters.
- Genome level data.
- Modelling challenges.
- Genomic databases.
- Integrating biological databases.
- Analysing genomic data.
- Summary and challenges.

What is the Genome?



All the genetic material in the chromosomes of a particular organism.

What is Genomics?

- The systematic application of (high throughput) molecular biology techniques to examine the whole genetic content of cells.
- Understand the meaning of the genomic information and how and when this information is expressed.

What is Bioinformatics?

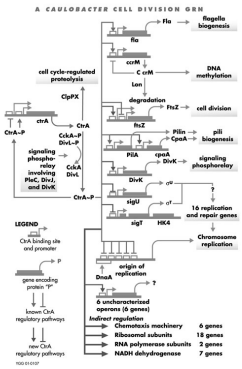
- "The application and development of computing and mathematics to the management, analysis and understanding of the rapidly expanding amount of biological information to solve biological questions"
- Straddles the interface between traditional biology and computer science

Human Genome Project

- The systematic cataloguing of individual gene sequences and mapping data to large species-specific collections
- "An inventory of life"
- June 25, 2000 draft of entire human genome announced
- Mouse, fruit fly, c. elegans, …
- Sequence is just the beginning

http://www.nature.com/genomics/human/papers/articles.html

Functional Genomics

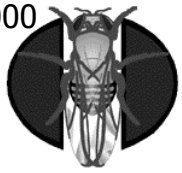


- An integrated view of how organisms work and interact in growth, development and pathogenesis
- From single gene to whole genome
- From single biochemical reactions to whole physiological and developmental systems

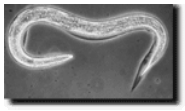
- What do genes do?
- How do they interact?

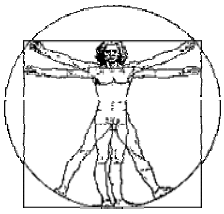
Comparative Genomics

~14,000

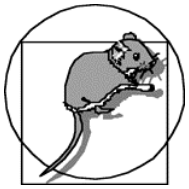


~9,000





~31,000

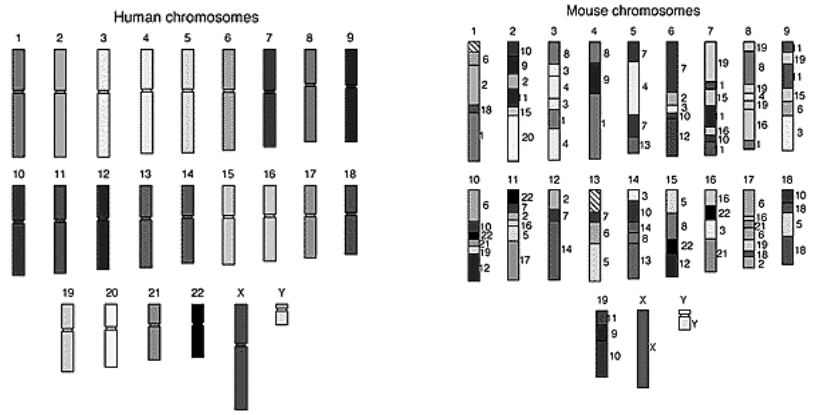


~30,000

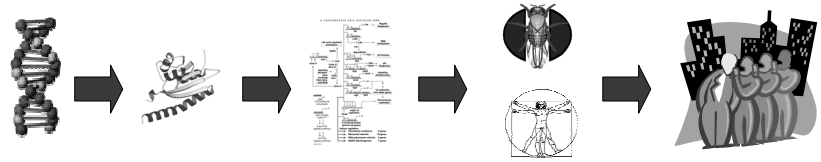
~6,000

http://wit.integratedgenomics.com/GOLD/

Of Mice and Men

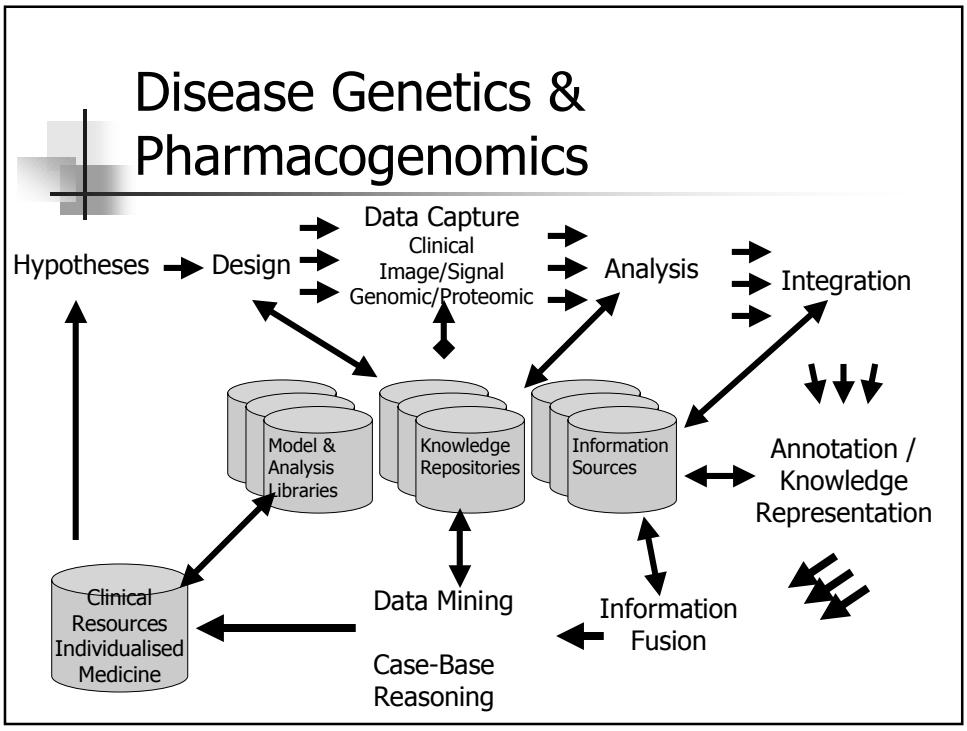


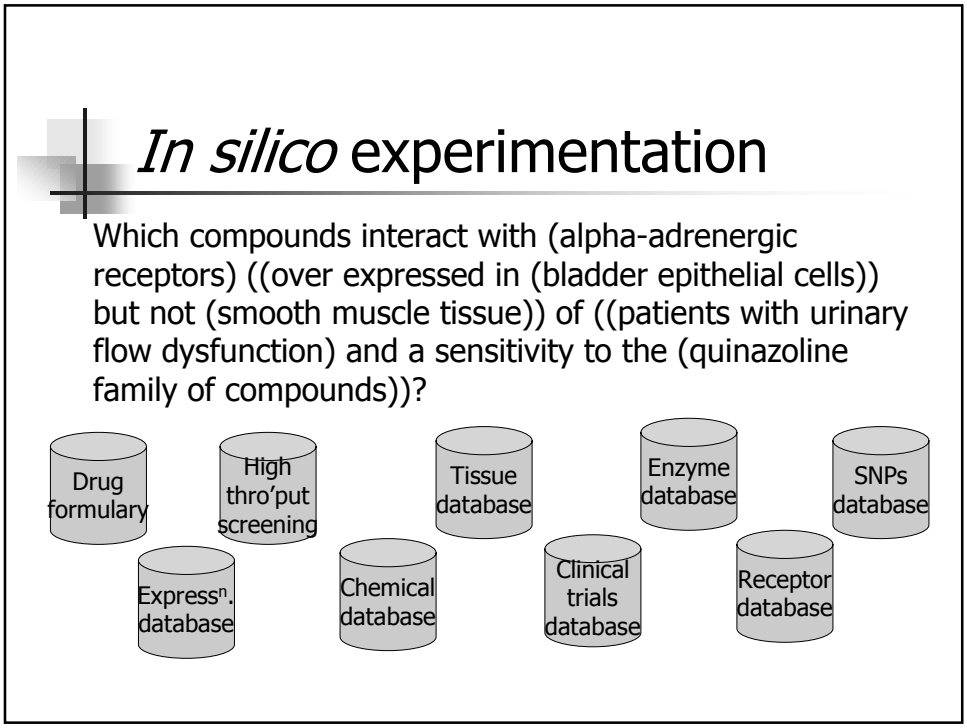
Genotype to Phenotype



DNA protein function organism population

- Link the observable behaviour of an organism with its genotype
- Drug Discovery, Agro-Food, Pharmacogenomics (individualised medicine)


Disease Genetics & Pharmacogenomics
Data Capture
Clinical
Image/Signal
Genomic/Proteomic
Hypotheses
Design
Analysis
Integration
Model & Analysis Libraries
Knowledge Repositories
Information Sources
Annotation / Knowledge Representation
Clinical Resources Individualised Medicine
Data Mining
Case-Base Reasoning
Information Fusion


In silico experimentation
Which compounds interact with (alpha-adrenergic receptors) ((over expressed in (bladder epithelial cells)) but not (smooth muscle tissue)) of ((patients with urinary flow dysfunction) and a sensitivity to the (quinazoline family of compounds))?
Drug formulary
High thro'put screening
Tissue database
Enzyme database
SNPs database
Express^{n}. database
Chemical database
Clinical trials database
Receptor database

A Paradigm Shift

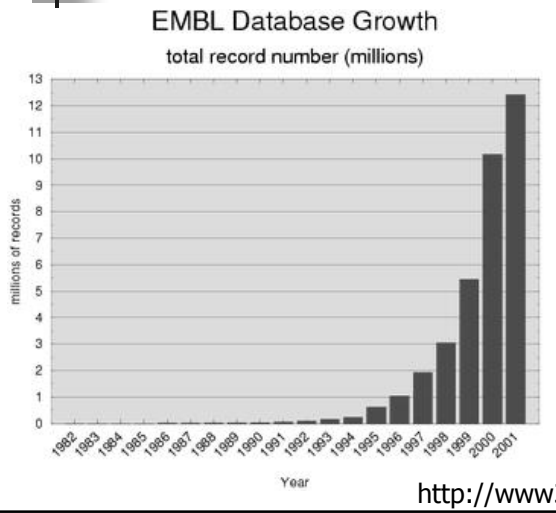
Hypothesis-driven

Hunter gatherers

Collection-driven

Harvesters

Size,
complexity, heterogeneity, instability



- EMBL July 2001
  - 150 Gbytes
- Microarray
  - 1 Petabyte per annum
- Sanger Centre
  - 20 terabytes of data
  - Genome sequences increase 4x per annum

http://www3.ebi.ac.uk/Services/DBStats/

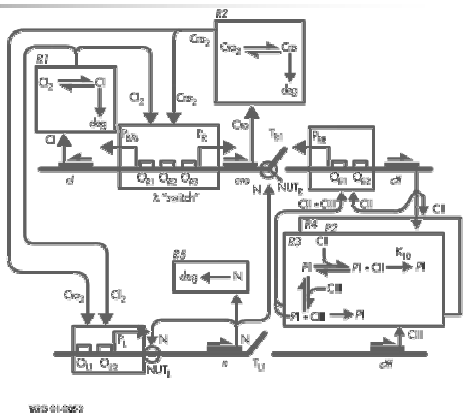
High throughput experimental methods

- Micro arrays for gene expression
- Robot-based capture
- 10K data points per chip
- 20 x per chip

- Cottage industry -> industrial scale

Complexity,
size, heterogeneity, instability

- Multiple views
- Interrelated
- Intra and inter cell interactions and bio-processes



"Courtesy U.S. Department of Energy Genomes to Life program (proposed) DOEGenomesToLife.org."

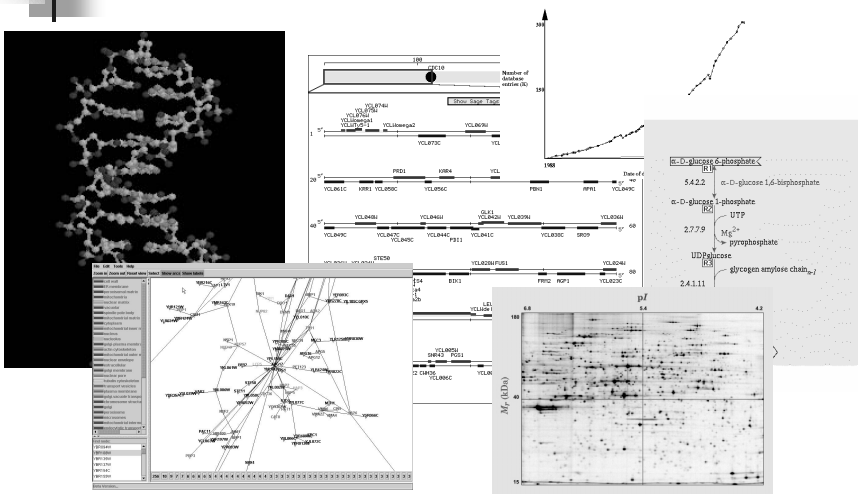
Heterogeneity
size, complexity, instability

- Multimedia
  - Images & Video (e.g. microarrays)
  - Text "annotations" & literature
- Over 500 different databases
  - Genomic, proteomic, transcriptomic, metabalomic, protein-protein interactions, regulatory bio-networks, alignments, disease, patterns & motifs, protein structure, protein classifications, specialist proteins (enzymes, receptors), …
  - Different formats, structure, schemas, coverage…
  - Web interfaces, flat file distribution,…

Instability
size, complexity, heterogeneity

- Exploring the unknown
  - At least 5 definitions of a gene
  - The sequence is a model
  - Other models are "work in progress"
- Names unstable
- Data unstable
- Models unstable

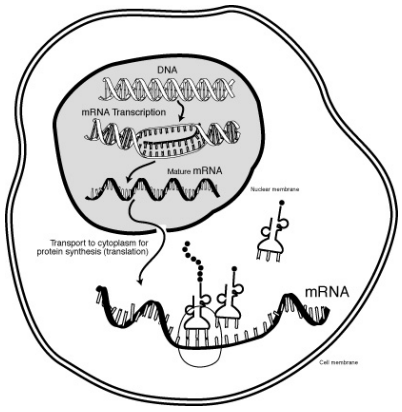
Genome Level Data



Biological Macromolecules

- *DNA*: the source of the program.
- mRNA: the compiled class definitions.
- Protein: the runtime object instances.

DNA → mRNA → Protein



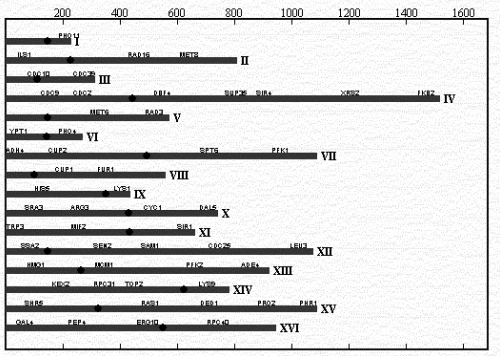
© Access Excellence@the National Health Museum

Biological Teaching Resources:
http://www.accessexcellence.com/

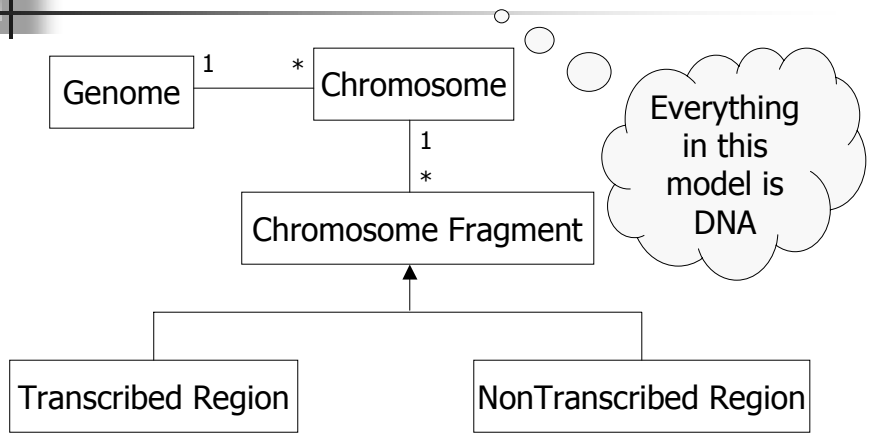
Genome

- The *genome* is the entire DNA sequence of an organism.

The yeast genome (*Saccharomyces cerevisiae*).
A friendly fungus: brewer's and baker's yeast.
`http://genome-www.stanford.edu/Saccharomyces/`



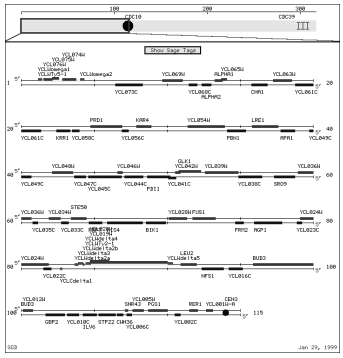
A Genome Data Model



Chromosome

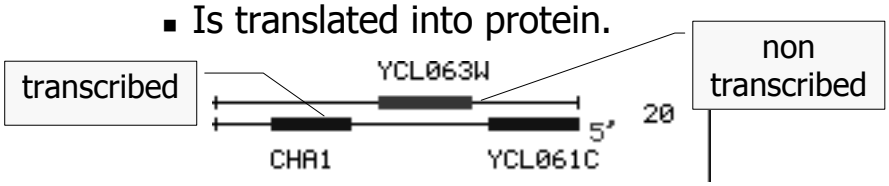
- A *chromosome* is a DNA molecule containing genes in linear order.

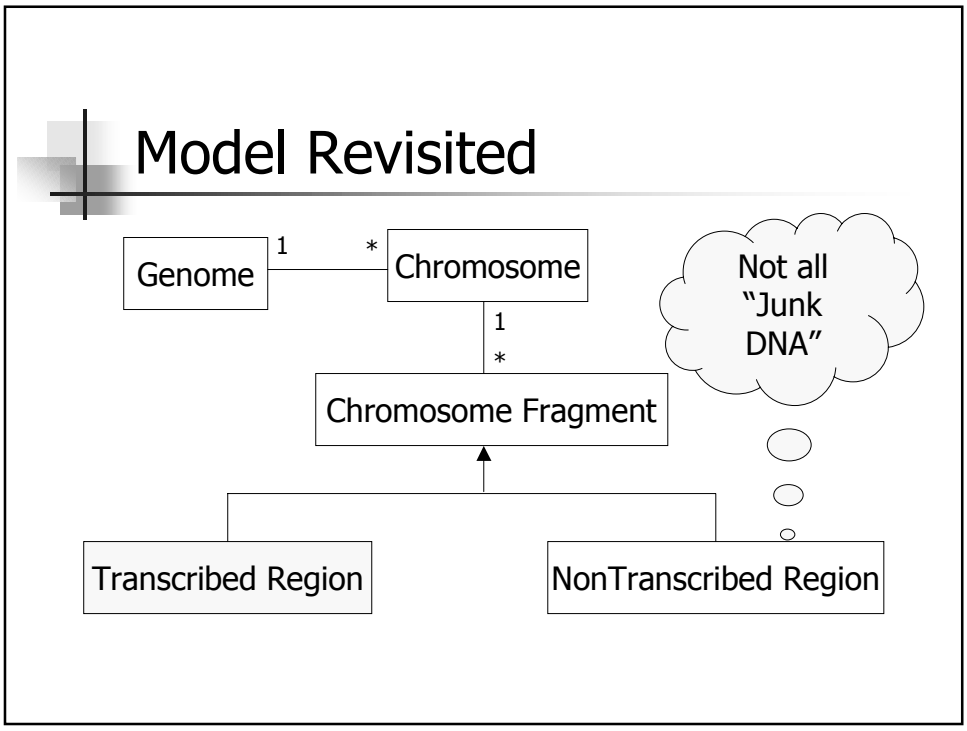
Chromosome III from yeast. Genes are shown shaded on the different strands of DNA.

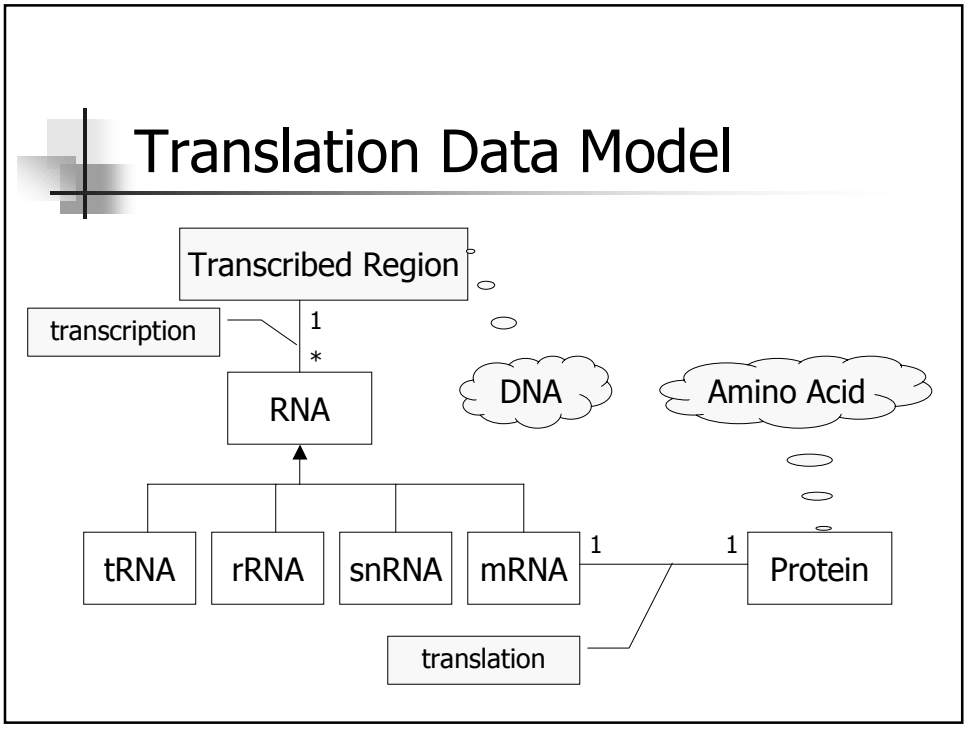


Gene

- A *gene* is a discrete unit of inherited information.
- A gene is *transcribed* into RNA which either:
  - Functions directly in the cell, or
  - Is translated into protein.




Model Revisited
Genome
1
*
Chromosome
1
*
Chromosome Fragment
Not all "Junk DNA"
Transcribed Region
NonTranscribed Region


Translation Data Model
Transcribed Region
transcription
1
*
RNA
DNA
Amino Acid
tRNA
rRNA
snRNA
mRNA
1
1
Protein
translation

Transcription

- In *transcription*, DNA is used as a template for the creation of RNA.

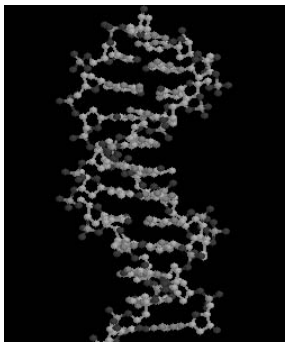
DNA		RNA	
A	Adenine	A	Adenine
C	Cytosine	C	Cytosine
G	Guanine	G	Guanine
T	Thymine	U	Uracil

Translation

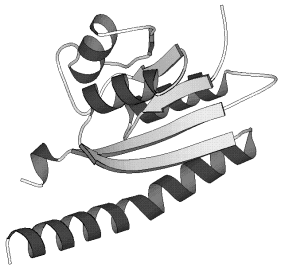
- In *translation* a protein sequence is synthesised according to the sequence of an mRNA molecule.
  - Four nucleic acids contribute to mRNA.
  - Twenty amino acids contribute to protein.

CODONS	**Amino Acid**
AAA, AAG	Lysine (Lys)
GCU, GCC, GCA, GCG	Alanine (Ala)
…	…

Molecular Structures



The double helix of DNA (http://www.bio.cmu.edu/Programs/Courses/)

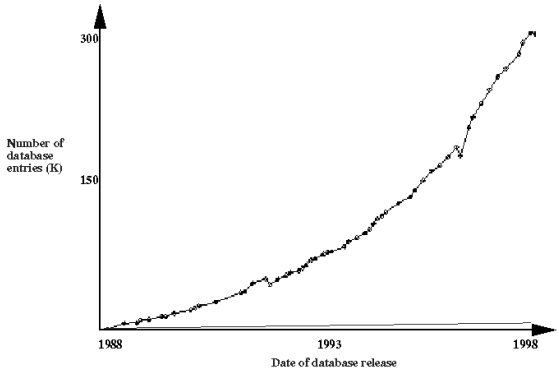


An abstract view of a globular Protein of unknown function (Zarembinski *et al*., *PNAS* **95** 1998)

Genome Facts

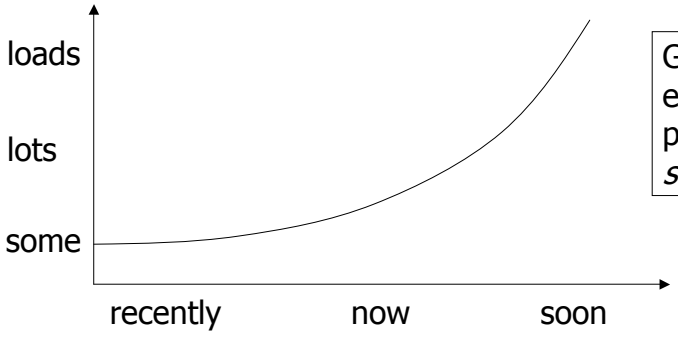
	Chromosomes	Genes	Base Pairs
Human	22 + X,Y	25000+	3.2 billion
Yeast	16	6000	12 million
E Coli	1	3500	4.6 million

Growth in Data Volumes



Non-redundant growth of sequences during 1988-1998 (black) and the corresponding growth in the number of structures (red).

General Growth Patterns



Growth in experimental production of *stuff*.

An emphasis on quantity could lead to oversights relating to complexity

Making Sense of Sequences

- The sequencing of a genome leaves two crucial questions:
  - What is the individual behaviour of each protein?
  - How does the overall behaviour of a cell follow from its genetic make-up?

In yeast, the function of slightly over 50% of the proteins has been detected experimentally or predicted through sequence similarity.

Reverse Engineering

- The genome is the source of a program by an inaccessible author, for which no documentation is available.
- Functional genomics seeks to develop and document the functionality of the program by observing its runtime behaviour.

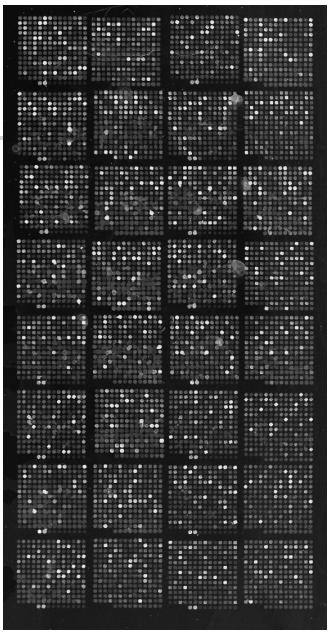
Funct

The "omes"

- Genome: the total DNA sequence of an organism (static).
- Transcriptome: a measure of the mRNA present in a cell at a point in time.
- Proteome: a measure of the protein present in a cell at a point in time.
- Metabolome: a record of the metabolites in a cell at a point in time.

Transcriptome

- Microarrays (DNA Chips) can measure many thousands of transcript levels at a time.
- Arrays allow transcript levels to be compared at different points in time.

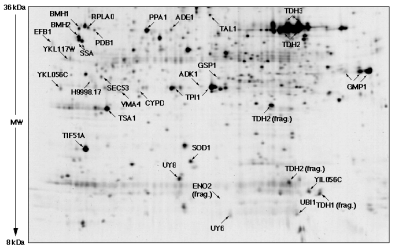


Transcriptome Features

- Loads of data:
  - Comprehensive in coverage.
  - High throughput.
- Challenging to interpret:
  - Normalisation.
  - Clustering.
  - Time series.

Proteome

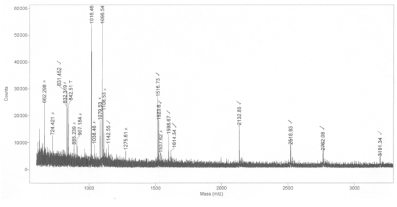
- Most proteome experiments involve separation then measurement.
- 2D Gels separate a sample according to mass and pH, so that (hopefully) each spot contains one protein.



Proteome Database: http://www.expasy.ch/

Mass Spectrometry

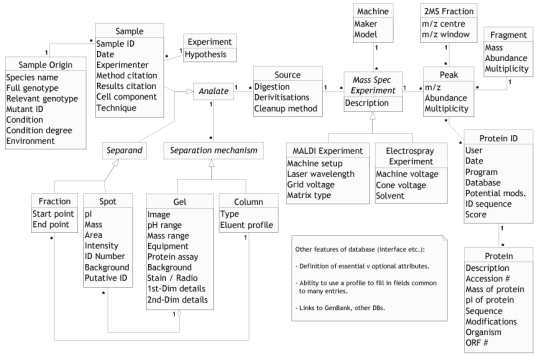
- Individual spots can be analysed using (one of many) mass spectrometry techniques.
- This can lead to the identification of specific proteins in a sample.



Mass spec. results for yeast. (http://www.cogeme.man.ac.uk)

Modelling Proteome Data

- Describing individual functional data sets is often challenging in itself.



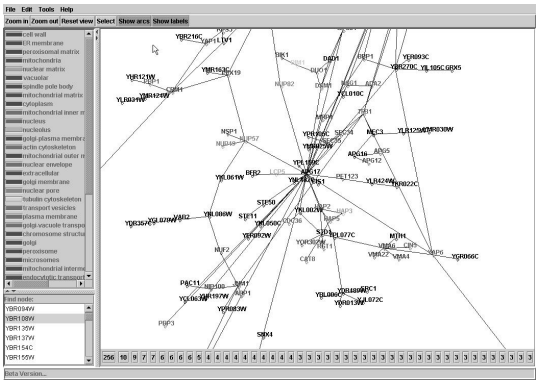
Proteome Features

- Moderate amounts of data:
  - Partial in coverage.
  - Medium throughput.
- Challenging to interpret:
  - Protein identification.

Protein Interactions

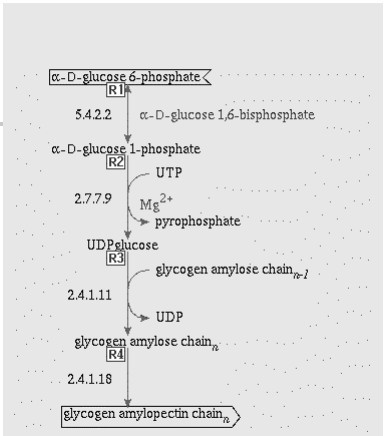
- Experimental techniques can also be used to identify protein interactions.

Protein-protein interaction viewer highlighting proteins based on cellular location (http://img.cs.man.ac.uk/gims)



Metabolome

- A metabolic pathway describes a series of reactions.
- Such pathways bring together collections of proteins and the small molecules with which they interact.



Glucose metabolism in yeast from WIT (http://wit.mcs.anl.gov/WIT2/)

Summary: Genome Level Data

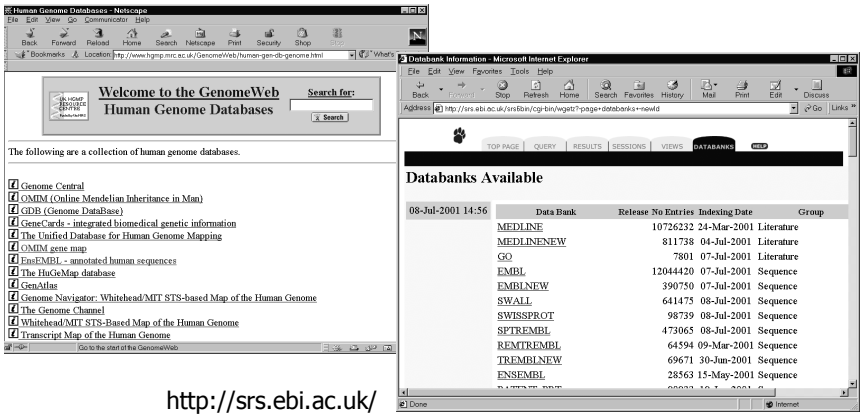
- Genome sequencing is moving fast:
  - Several genomes fully sequenced.
  - Many genomes partially sequenced.
- The sequence is not the whole story:
  - Many genes are of unknown function.
  - Developments in functional genomics are yielding new and challenging data sets.

Useful URLs on Genomic Data

- Nature Genome Gateway:
  - http://www.nature.com/genomics/
- UK Medical Research Council Demystifying Genomics document:
  - http://www.mrc.ac.uk/PDFs/dem_gen.pdf
- Genomic glossary:
  - http://www.genomicglossaries.com/
- Teaching resources:
  - http://www.iacr.bbsrc.ac.uk/notebook/index.html

Genomic Databases

http://www.hgmp.mrc.ac.uk/GenomeWeb/genome-db.html



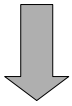
http://srs.ebi.ac.uk/

Key points

1. What do the databases contain?
   - Broad vs deep
   - Primary vs secondary
2. What are the database services?
   - Architecture & Web browsing paradigm
3. How are the databases published?
4. How are the data represented?
   - Annotation
5. How are the databases curated ?

A Paradigm Shift

Publishing journals



Publishing data

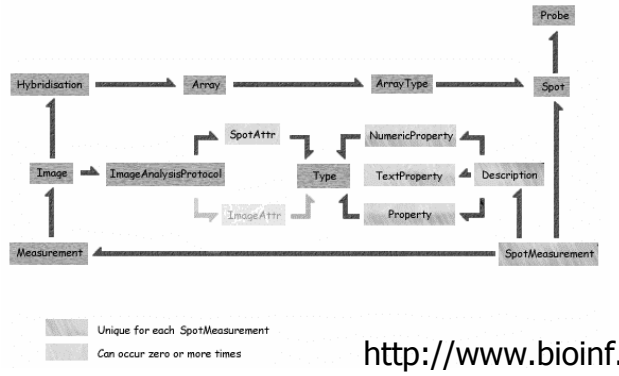
Re-analysable

Broad vs Deep Databases

- Broad: Clustered around data type or biological system across multiple species
  - Sequence: protein (Swiss-Prot), nucleotide (EMBL), patterns (Interpro) …
  - Genomic: transcriptome (MaxD), pathway (WIT)…
- Deep: Data integrated across a species
  - *Saccharomyces cerevisiae* MIPS, SGD, YPD
  - Flybase, MouseBase, XXXBase …

Broad Example – MaxD

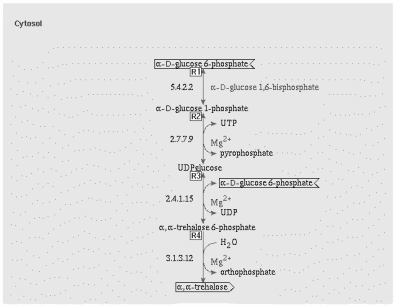
MaxD is a relational implementation of the ArrayExpress proposal for a transcriptome database.



http://www.bioinf.man.ac.uk

Broad Example - WIT

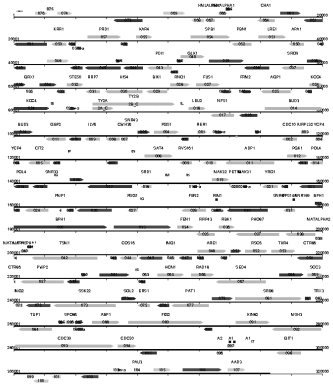
WIT is a WWW resource providing access to metabolic pathways from many species.



http://wit.mcs.anl.gov

Deep Example - MIPS

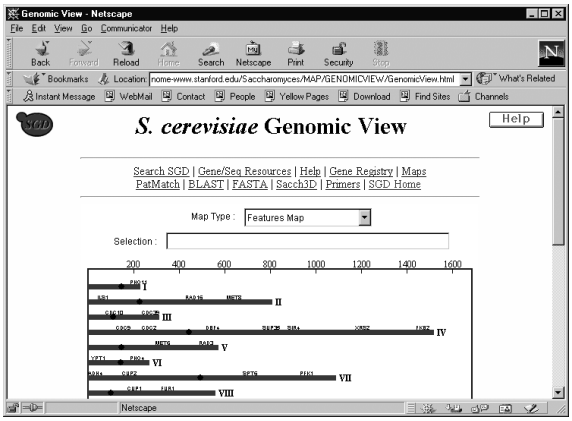
MIPS is one of several sites providing access, principally for browsing, to both sequence and functional data.



http://www.mips.biochem.mpg.de/

Deep Example - SGD

SGD contains sequence, function and literature information on *S. cerevisiae.* mostly for browsing and viewing.



http://genome-www.stanford.edu/Saccharomyces/

Primary Databases

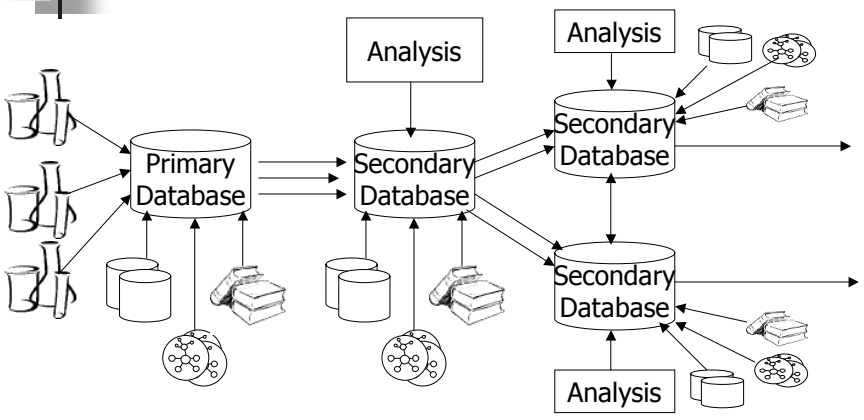
- Primary source generated by experimentalists.
- Role: standards, quality thresholds, dissemination
  - Sequence databases: EMBL, GenBank
  - Increasingly other data types: micro-array …

Secondary databases

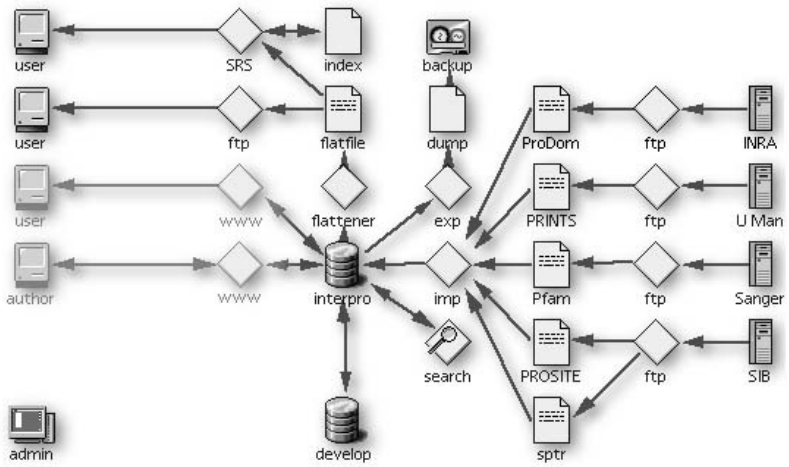
- Secondary source derived from repositories, other secondary databases, analysis and expertise.

1. Role: Distilled and accumulated specialist knowledge. Value added commentary called "annotation"
   - Swiss-Prot, PRINTS, CATH, PAX6, Enzyme, dbSNP…
2. Role: Warehouses to support analysis over replicated data
   - GIMS, aMAZE, InterPro…

Database collection flows

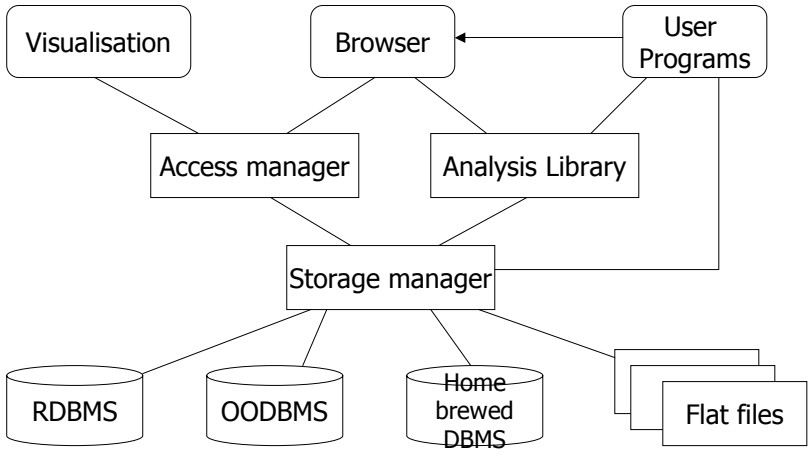


InterPro Data Flow



http://www.ebi.ac.uk/interpro/dataflow_scheme.html

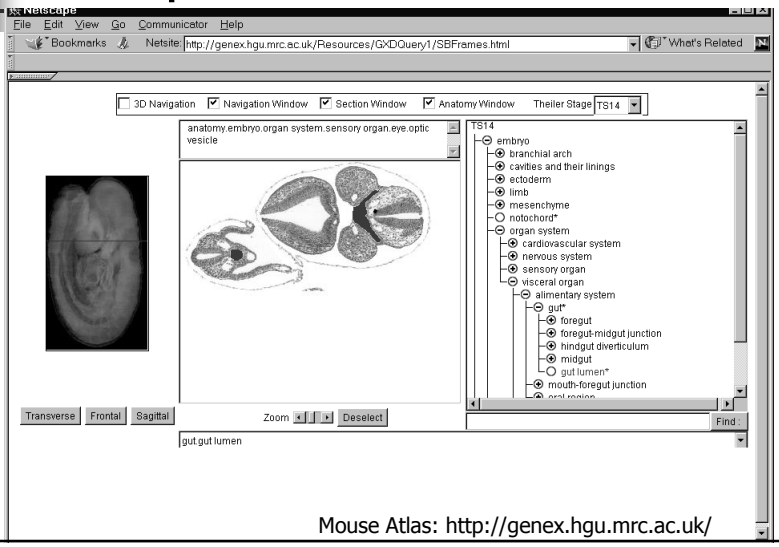
Services: Architecture



How do I use a database?

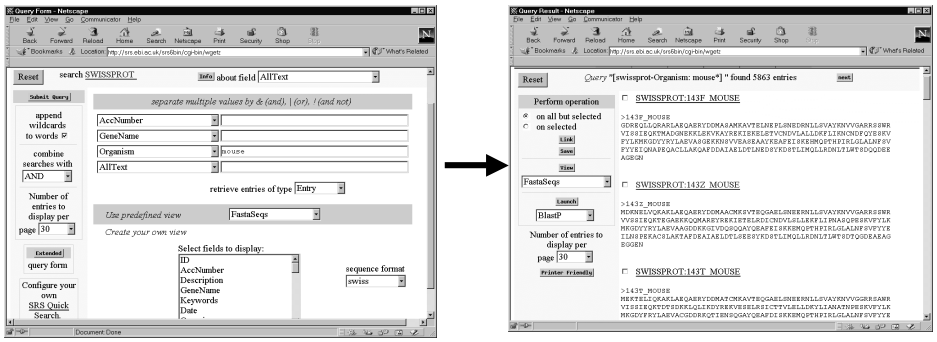
- Web browser
  - Cut and paste, point and click
  - Query by navigation
  - Results in flat file formats or graphical
  - Screen-scrapping
- Perl scripts over downloaded flat files
  - The most popular form
  - XML formats taking hold
- Beginnings of API's in Corba
  - But still limited to call-interface rather than queries

Example Visualisation

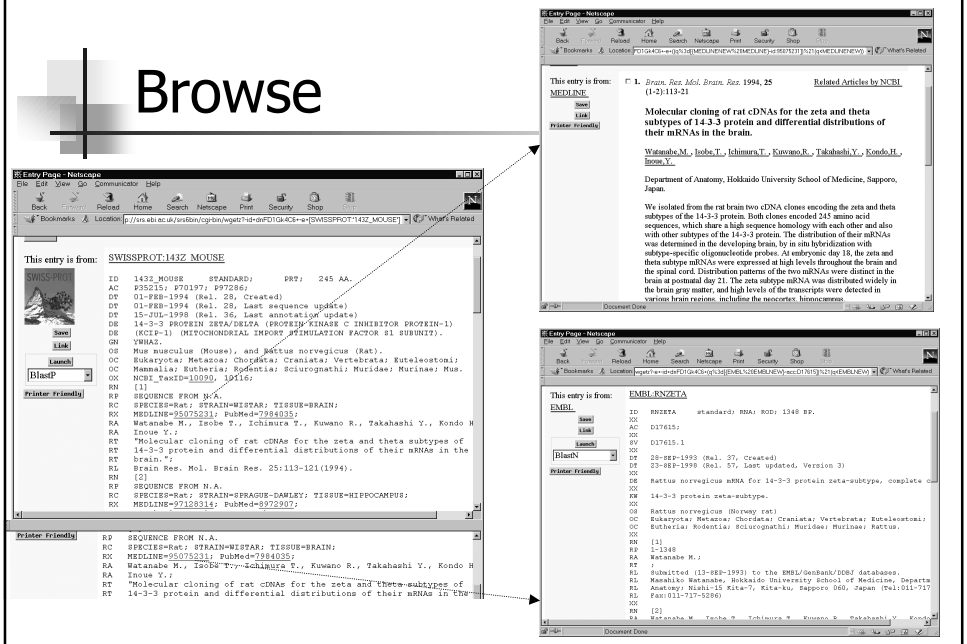


Mouse Atlas: http://genex.hgu.mrc.ac.uk/

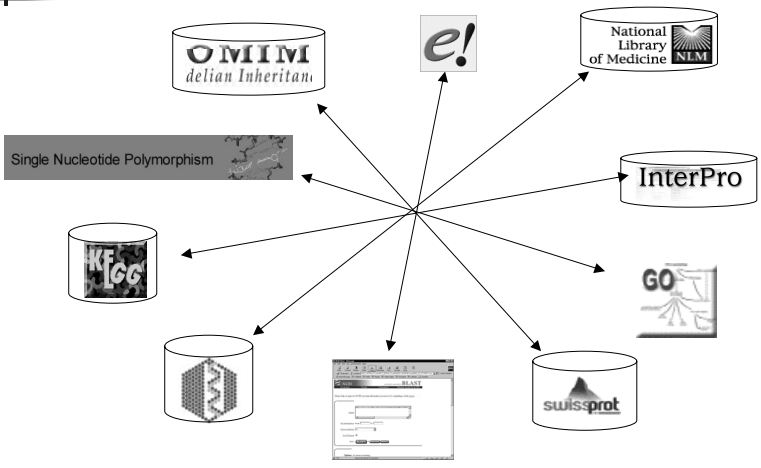
Query and Browse



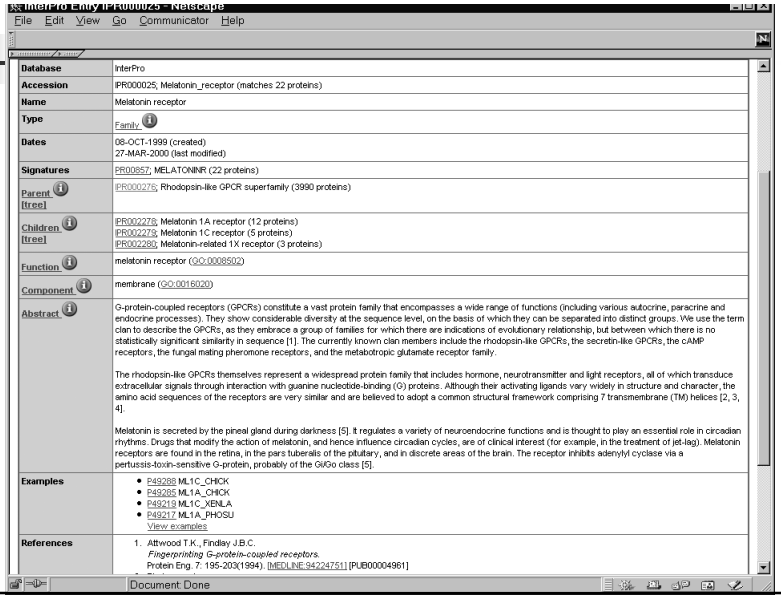
Browse



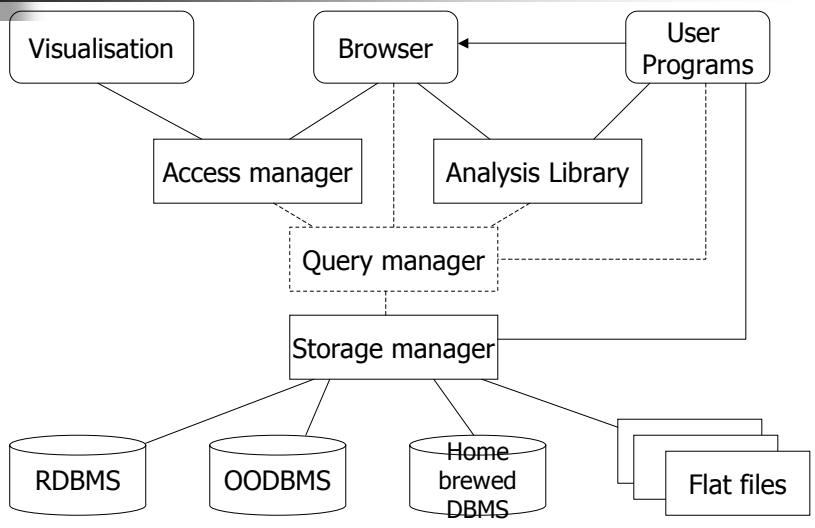
Inter-database referential integrity



Inter-database references



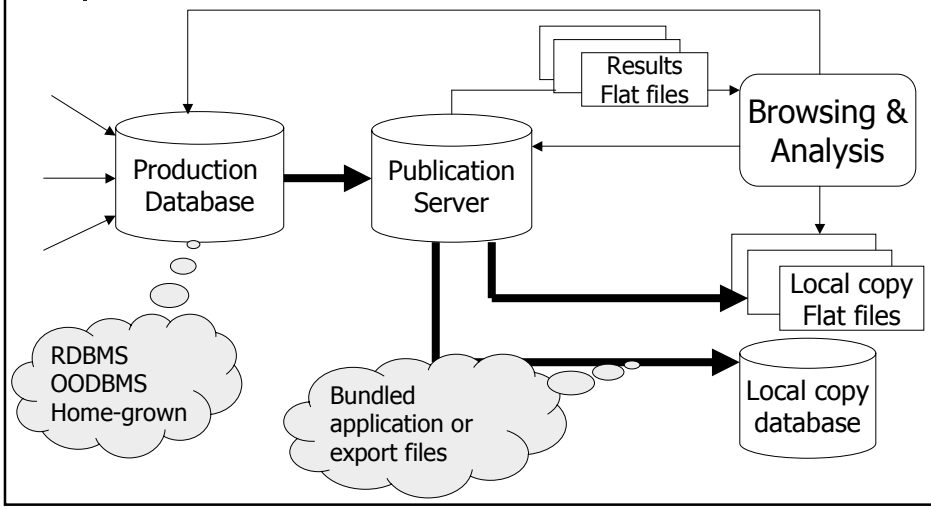
Query based retrieval?



Query Expressions

- A query interface through a web browser or command line
  - AceDB language (SGD)
  - Icarus (SRS)
  - SQL?
- API's generally don't allow query submission

Two (& three) tier delivery



EMBL Flat File Format part 1

```
ID   TRBG361    standard; RNA; PLN; 1859 BP.
AC   X56734; S46826;
SV   X56734.1
DT   12-SEP-1991 (Rel. 29, Created)
DT   15-MAR-1999 (Rel. 59, Last updated, Version 9)
DE   Trifolium repens mRNA for non-cyanogenic beta-glucosidase
KW   beta-glucosidase.
OS   Trifolium repens (white clover)
OC   Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
OC   Spermatophyta; Magnoliophyta; eudicotyledons; core eudicots; Rosidae;
OC   eurosids I; Fabales; Fabaceae; Papilionoideae; Trifolium.
RN   [5]
RP   1-1859
RX   MEDLINE; 91322517.
RA   Oxtoby E., Dunn M.A., Pancoro A., Hughes M.A.;
RT   "Nucleotide and derived amino acid sequence of the cyanogenic
RT   beta-glucosidase (linamarase) from white clover (Trifolium repens L.).";
RL   Plant Mol. Biol. 17:209-219(1991).
```

EMBL Flat File Format part 2

```
DR   AGDR; X56734; X56734.
DR   MENDEL; 11000; Trirp;1162;11000.
DR   SWISS-PROT; P26204; BGLS_TRIRP.
FH   Key             Location/Qualifiers
FH
FT   source          1..1859
FT                   /db_xref="taxon:3899"
FT                   /organism="Trifolium repens"
FT                   /tissue_type="leaves"
FT                   /clone_lib="lambda gt10"
FT                   /clone="TRE361"
FT   CDS             14..1495
FT                   /db_xref="SWISS-PROT:P26204"
FT                   /note="non-cyanogenic"
FT                   /EC_number="3.2.1.21"
FT                   /product="beta-glucosidase"
FT                   /protein_id="CAA40058.1"
```

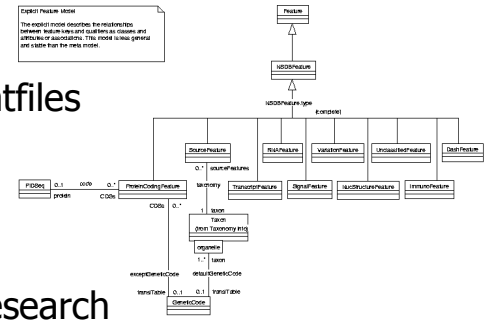
EMBL Flat File Format part 3

```
FT                   /translation="MDFIVAIFALFVISSFTITSTNAVEASTLLDIGNLSRSSFPRGFI
FT                   FGAGSSAYQFEGAVNEGGRGPSIWDTFTHKYPEKIRDGSNADITVDQYHRYKEDVGIMK
FT                   DQNMDSYRFSISWPRILPKGKLSGGINHEGIKYYNNLINELLANGIQPFVTLFHWDLPQ
FT                   VLEDEYGGFLNSGVINDFRDYTDLCFKEFGDRVRYWSTLNEPWVFSNSGYALGTNAPGR
FT                   CSASNVAKPGDSGTGPYIVTHNQILAHAEAVHVYKTKYQAYQKGKIGITLVSNWLMPLD
FT                   DNSIPDIKAAERSLDFQFGLFMEQLTTGDYSKSMRRIVKNRLPKFSKFESSLVNGSFDF
FT                   IGINYYSSSYISNAPSHGNAKPSYSTNPMTNISFEKHGIPLGPRAASIWIYVYPYMFIQ
FT                   EDFEIFCYILKINITILQFSITENGMNEFNDATLPVEEALLNTYRIDYYYRHLYYIRSA
FT                   IRAGSNVKGFYAWSFLDCNEWFAGFTVRFGLNFVD"
FT   mRNA            1..1859
FT                   /evidence=EXPERIMENTAL
XX
SQ   Sequence 1859 BP; 609 A; 314 C; 355 G; 581 T; 0 other;
     aaacaaacca aatatggatt ttattgtagc catatttgct ctgtttgtta ttagctcatt        60
     cacaattact tccacaaatg cagttgaagc ttctactctt cttgacatag gtaacctgag       120
     tcggagcagt tttcctcgtg gcttcatctt tggtgctgga tcttcagcat accaatttga       180
     aggtgcagta aacgaaggcg gtagaggacc aagtatttgg gataccttca cccataaata       240
      etc....
```

XML embraced

- Side effect of publication through flat files and textual annotation
- XML for distribution, storage and interoperation,
  - e.g. BLASTXML, Distributed Annotation System
- Many XML genome annotation DTDs:
  - Sequence: BIOML, BSML, AGAVE, GAME
  - Function: MAML, MaXML
  - http://www.bioxml.org/
- I3C vendors attempt to coordinate activities and promote XML for integration
  - http://i3c.open-bio.org

Move to OO interfaces



- OO API's to RDMS or flatfiles
- Corba activity
  - EMBL Corba Server
  - OMG: Life Sciences Research
- OMG not yet taken hold

http://corba.ebi.ac.uk/EMBL_embl.html
http://lsr.ebi.ac.uk/

Annotation and Curation

"the elucidation and description of biologically relevant features [in a sequence]"

1. Computationally formed – e.g. cross references to other database entries, date collected;
2. Intellectually formed – the accumulated knowledge of an expert distilling the aggregated information drawn from multiple data sources and analyses, and the annotators knowledge.

Annotation Distillation

Expressed Sequence Tags	millions
nrdb	503,479
TrEMBL	234,059
Swiss-Prot	85,661
InterPro	2990
PRINTS	1310

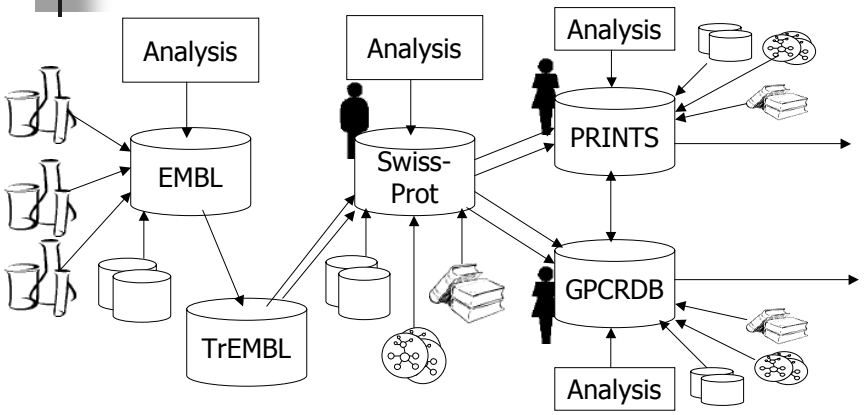
Swiss-Prot Annotation

```
ID   PRIO_HUMAN     STANDARD;      PRT;   253 AA.
AC   P04156;
DE   MAJOR PRION PROTEIN PRECURSOR (PRP) (PRP27-30) (PRP33-35C) (ASCR).
OS   Homo sapiens (Human).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
OX   NCBI_TaxID=9606;
RN   [1]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=86300093 [NCBI, ExPASy, Israel, Japan]; PubMed=3755672;
RA   Kretzschmar H.A., Stowring L.E., Westaway D., Stubblebine W.H., Prusiner S.B., Dearmond S.J.
RT   "Molecular cloning of a human prion protein cDNA.";
RL   DNA 5:315-324(1986).
RN   [6]
RP   STRUCTURE BY NMR OF 23-231.
RX   MEDLINE=97424376 [NCBI, ExPASy, Israel, Japan]; PubMed=9280298;
RA   Riek R., Hornemann S., Wider G., Glockshuber R., Wuethrich K.;
RT   "NMR characterization of the full-length recombinant murine prion protein, mPrP(23-231).";
RL   FEBS Lett. 413:282-288(1997).
CC   -!- FUNCTION: THE FUNCTION OF PRP IS NOT KNOWN. PRP IS ENCODED IN THE HOST GENOME AND IS
CC       EXPRESSED BOTH IN NORMAL AND INFECTED CELLS.
CC   -!- SUBUNIT: PRP HAS A TENDENCY TO AGGREGATE YIELDING POLYMERS CALLED "RODS".
CC   -!- SUBCELLULAR LOCATION: ATTACHED TO THE MEMBRANE BY A GPI-ANCHOR.
CC   -!- DISEASE: PRP IS FOUND IN HIGH QUANTITY IN THE BRAIN OF HUMANS AND ANIMALS INFECTED WITH
CC       NEURODEGENERATIVE DISEASES KNOWN AS TRANSMISSIBLE SPONGIFORM ENCEPHALOPATHIES OR PRION
CC       DISEASES, LIKE: CREUTZFELDT-JAKOB DISEASE (CJD), GERSTMANN-STRAUSSLER SYNDROME (GSS),
CC       FATAL FAMILIAL INSOMNIA (FFI) AND KURU IN HUMANS; SCRAPIE IN SHEEP AND GOAT; BOVINE
CC       SPONGIFORM ENCEPHALOPATHY (BSE) IN CATTLE; TRANSMISSIBLE MINK ENCEPHALOPATHY (TME);
CC       CHRONIC WASTING DISEASE (CWD) OF MULE DEER AND ELK; FELINE SPONGIFORM ENCEPHALOPATHY
CC       (FSE) IN CATS AND EXOTIC UNGULATE ENCEPHALOPATHY(EUE) IN NYALA AND GREATER KUDU. THE
CC       PRION DISEASES ILLUSTRATE THREE MANIFESTATIONS OF CNS DEGENERATION: (1) INFECTIOUS (2)
CC       SPORADIC AND (3) DOMINANTLY INHERITED FORMS. TME, CWD, BSE, FSE, EUE ARE ALL THOUGHT TO
CC       OCCUR AFTER CONSUMPTION OF PRION-INFECTED FOODSTUFFS.
CC   -!- SIMILARITY: BELONGS TO THE PRION FAMILY.
DR   HSSP; P04925; 1AG2. [HSSP ENTRY / SWISS-3DIMAGE / PDB]
DR   MIM; 176640; -. [NCBI / EBI]
DR   InterPro; IPR000817; -.
DR   Pfam; PF00377; prion; 1.
DR   PRINTS; PR00341; PRION.
KW   Prion; Brain; Glycoprotein; GPI-anchor; Repeat; Signal; Polymorphism; Disease mutation.
```

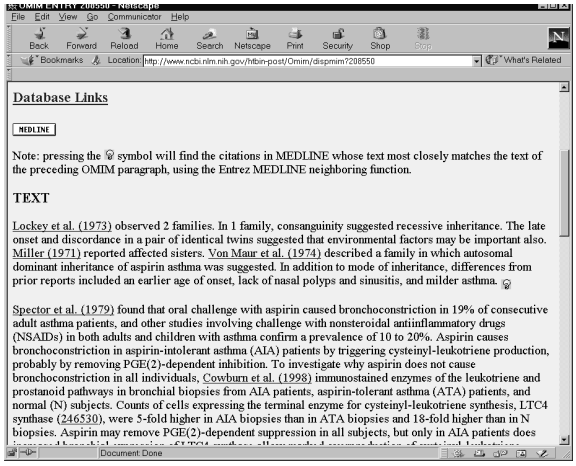
PRINTS Annotation

```
gc; PRION
gx; PR00341
gt; Prion protein signature
gp; INTERPRO; IPR000817
gp; PROSITE; PS00291 PRION_1; PS00706 PRION_2
gp; BLOCKS; BL00291
gp; PFAM; PF00377 prion
bb;
gr; 1. STAHL, N. AND PRUSINER, S.B.
gr; Prions and prion proteins.
gr; FASEB J. 5 2799-2807 (1991).
gr;
gr; 2. BRUNORI, M., CHIARA SILVESTRINI, M. AND POCCHIARI, M.
gr; The scrapie agent and the prion hypothesis.
gr; TRENDS BIOCHEM.SCI. 13 309-313 (1988).
gr;
gr; 3. PRUSINER, S.B.
gr; Scrapie prions.
gr; ANNU.REV.MICROBIOL. 43 345-374 (1989).
bb;
gd; Prion protein (PrP) is a small glycoprotein found in high quantity in the brain of animals infected with
gd; certain degenerative neurological diseases, such as sheep scrapie and bovine spongiform encephalopathy (BSE),
gd; and the human dementias Creutzfeldt-Jacob disease (CJD) and Gerstmann-Straussler syndrome (GSS). PrP is
gd; encoded in the host genome and is expressed both in normal and infected cells. During infection, however, the
gd; PrP molecules become altered and polymerise, yielding fibrils of modified PrP protein.
gd;
gd; PrP molecules have been found on the outer surface of plasma membranes of nerve cells, to which they are
gd; anchored through a covalent-linked glycolipid, suggesting a role as a membrane receptor. PrP is also
gd; expressed in other tissues, indicating that it may have different functions depending on its location.
gd;
gd; The primary sequences of PrP's from different sources are highly similar: all bear an N-terminal domain
gd; containing multiple tandem repeats of a Pro/Gly rich octapeptide; sites of Asn-linked glycosylation; an
gd; essential disulphide bond; and 3 hydrophobic segments. These sequences show some similarity to a chicken
gd; glycoprotein, thought to be an acetylcholine receptor-inducing activity (ARIA) molecule. It has been
gd; suggested that changes in the octapeptide repeat region may indicate a predisposition to disease, but it is
gd; not known for certain whether the repeat can meaningfully be used as a fingerprint to indicate susceptibility.
gd;
gd; PRION is an 8-element fingerprint that provides a signature for the prion proteins. The fingerprint was
gd; derived from an initial alignment of 5 sequences: the motifs were drawn from conserved regions spanning
gd; virtually the full alignment length, including the 3 hydrophobic domains and the octapeptide repeats
gd; (WGQPHGGG). Two iterations on OWL18.0 were required to reach convergence, at which point a true set comprising
gd; 9 sequences was identified. Several partial matches were also found: these include a fragment (PRIO_RAT)
gd; lacking part of the sequence bearing the first motif,and the PrP homologue found in chicken - this matches
gd; well with only 2 of the 3 hydrophobic motifs (1 and 5) and one of the other conserved regions (6), but has an
gd; N-terminal signature based on a sextapeptide repeat (YPHNPG) rather than the characteristic PrP octapeptide.
```

The "Annotation Pipeline"



"Un"Structured Literature



Database Links

MEDLINE

Note: pressing the symbol will find the citations in MEDLINE whose text most closely matches the text of the preceding OMIM paragraph, using the Entrez MEDLINE neighboring function.

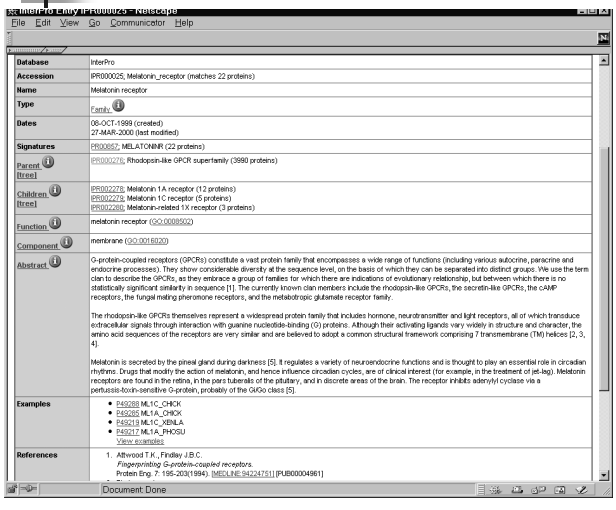
TEXT

Lockey et al. (1973) observed 2 families. In 1 family, consanguinity suggested recessive inheritance. The late onset and discordance in a pair of identical twins suggested that environmental factors may be important also. Miller (1971) reported affected sisters. Von Maur et al. (1974) described a family in which autosomal dominant inheritance of aspirin asthma was suggested. In addition to mode of inheritance, differences from prior reports included an earlier age of onset, lack of nasal polyps and sinusitis, and milder asthma.

Spector et al. (1979) found that oral challenge with aspirin caused bronchoconstriction in 19% of consecutive adult asthma patients, and other studies involving challenge with nonsteroidal antiinflammatory drugs (NSAIDs) in both adults and children with asthma confirm a prevalence of 10 to 20%. Aspirin causes bronchoconstriction in aspirin-intolerant asthma (AIA) patients by triggering cysteinyl-leukotriene production, probably by removing PGE(2)-dependent inhibition. To investigate why aspirin does not cause bronchoconstriction in all individuals, Cowburn et al. (1998) immunostained enzymes of the leukotriene and prostanoid pathways in bronchial biopsies from AIA patients, aspirin-tolerant asthma (ATA) patients, and normal (N) subjects. Counts of cells expressing the terminal enzyme for cysteinyl-leukotriene synthesis, LTC4 synthase (246530), were 5-fold higher in AIA biopsies than in ATA biopsies and 18-fold higher than in N biopsies. Aspirin may remove PGE(2)-dependent suppression in all subjects, but only in AIA patients does

- Biology is knowledge based
- The insights are in the literature

Semi-Structured



Database	InterPro
Accession	IPR000025, Melatonin_receptor (matches 22 proteins)
Name	Melatonin receptor
Type	Family
Dates	08-OCT-1999 (created) 27-MAR-2000 (last modified)
Signatures	PR00687; MELATONINR (22 proteins)
Parent [tree]	PR00237; Rhodopsin-like GPCR superfamily (3990 proteins)
Children [tree]	PR00278; Melatonin 1A receptor (12 proteins) PR00279; Melatonin 1C receptor (5 proteins) PR00280; Melatonin-related 1X receptor (3 proteins)
Function	melatonin receptor (GO:0008502)
Component	membrane (GO:0016020)
Abstract	G-protein-coupled receptors (GPCRs) constitute a vast protein family that encompasses a wide range of functions (including various autocrine, paracrine and endocrine processes). They show considerable diversity at the sequence level, on the basis of which they can be separated into distinct groups. We use the term clan to describe the GPCRs, as they embrace a group of families for which there are indications of evolutionary relationship, but between which there is no statistically significant similarity in sequence [1]. The currently known clan members include the rhodopsin-like GPCRs, the secretin-like GPCRs, the cAMP receptors, the fungal mating pheromone receptors, and the metabotropic glutamate receptor family. The rhodopsin-like GPCRs themselves represent a widespread protein family that includes hormone, neurotransmitter and light receptors, all of which transduce extracellular signals through interaction with guanine nucleotide-binding (G) proteins. Although their activating ligands vary widely in structure and character, the amino acid sequences of the receptors are very similar and are believed to adopt a common structural framework comprising 7 transmembrane (TM) helices [2, 3, 4]. Melatonin is secreted by the pineal gland during darkness [5]. It regulates a variety of neuroendocrine functions and is thought to play an essential role in circadian rhythms. Drugs that modify the action of melatonin, and hence influence circadian cycles, are of clinical interest (for example, in the treatment of jet-lag). Melatonin receptors are found in the retina, in the pars tuberalis of the pituitary, and in discrete areas of the brain. The receptor inhibits adenylyl cyclase via a pertussis-toxin-sensitive G-protein, probably of the Gi/Go class [5].
Examples	• P49288 ML1C_CHICK • P49285 ML1A_CHICK • P49219 ML1C_XENLA • P49217 ML1A_PHOSU View examples
References	1. Attwood T.K., Findlay J.B.C. *Fingerprinting G-protein-coupled receptors.* Protein Eng. 7: 195-203(1994). MEDLINE 94224751 [PUB00004961]

- Schemaless Descriptions
- Evolving
- Non-predictive
- The structured part of the schema is open to change
- Hence flat file mark up's prevalence

Typical Database Services

1. Browsing ✔✔✔
2. Visualisation ✔✔✔
3. Querying ✔
4. Analysis -
5. API ✔

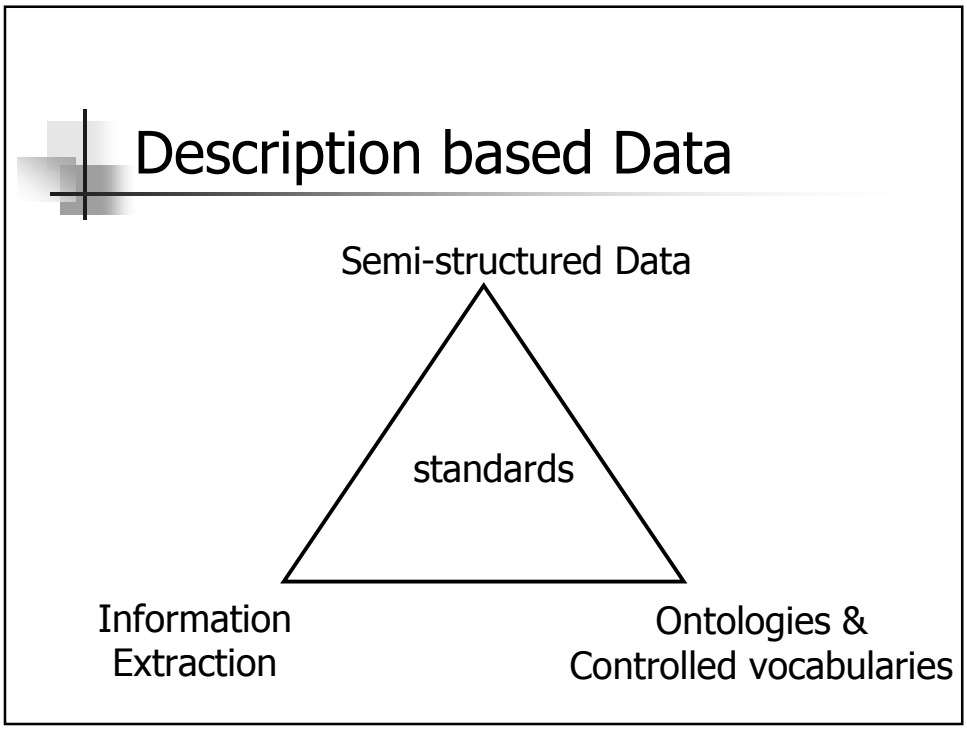
Focus on a person sitting in front of a Web browser pointing and clicking

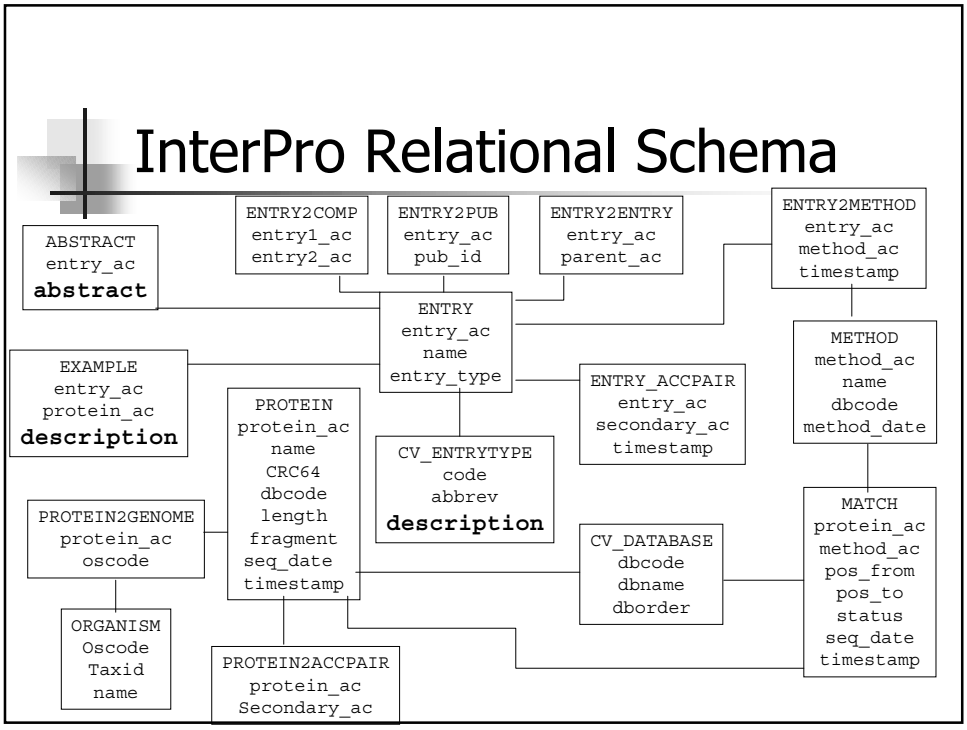
Typical Genomic Databases

	Single Genome	Multiple Genomes	Sequence	Function
Browse	✔	✔	✔	✔✔
Visualise	✔	✔	✔	✔✔
Query				
Analyse				

Broad Databases

Deep Databases


Description based Data
Semi-structured Data
standards
Information Extraction
Ontologies & Controlled vocabularies

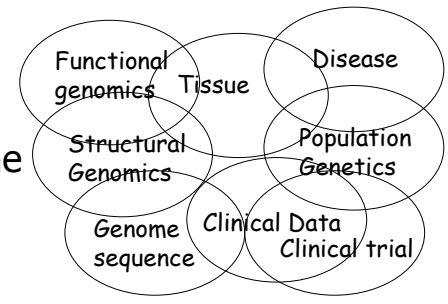

InterPro Relational Schema
ABSTRACT
entry_ac
abstract
ENTRY2COMP
entry1_ac
entry2_ac
ENTRY2PUB
entry_ac
pub_id
ENTRY2ENTRY
entry_ac
parent_ac
ENTRY2METHOD
entry_ac
method_ac
timestamp
ENTRY
entry_ac
name
entry_type
METHOD
method_ac
name
dbcode
method_date
EXAMPLE
entry_ac
protein_ac
description
PROTEIN
protein_ac
name
CRC64
dbcode
length
fragment
seq_date
timestamp
ENTRY_ACCPAIR
entry_ac
secondary_ac
timestamp
CV_ENTRYTYPE
code
abbrev
description
PROTEIN2GENOME
protein_ac
oscode
CV_DATABASE
dbcode
dbname
dborder
MATCH
protein_ac
method_ac
pos_from
pos_to
status
seq_date
timestamp
ORGANISM
Oscode
Taxid
name
PROTEIN2ACCPAIR
protein_ac
Secondary_ac

Controlled Vocabularies

```
ID  PRIO_HUMAN     STANDARD;      PRT;   253 AA.
DE  MAJOR PRION PROTEIN PRECURSOR (PRP) (PRP27-30) (PRP33-35C) (ASCR).
OS  Homo sapiens (Human).
OC  Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC  Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
CC  -!- FUNCTION: THE FUNCTION OF PRP IS NOT KNOWN. PRP IS ENCODED IN THE HOST GENOME AND IS
CC       EXPRESSED BOTH IN NORMAL AND INFECTED CELLS.
CC  -!- SUBUNIT: PRP HAS A TENDENCY TO AGGREGATE YIELDING POLYMERS CALLED "RODS".
CC  -!- SUBCELLULAR LOCATION: ATTACHED TO THE MEMBRANE BY A GPI-ANCHOR.
CC  -!- DISEASE: PRP IS FOUND IN HIGH QUANTITY IN THE BRAIN OF HUMANS AND ANIMALS INFECTED
WITH
CC  NEURODEGENERATIVE DISEASES KNOWN AS TRANSMISSIBLE SPONGIFORM ENCEPHALOPATHIES OR PRION
CC  DISEASES, LIKE: CREUTZFELDT-JAKOB DISEASE (CJD), GERSTMANN-STRAUSSLER SYNDROME (GSS),
CC  FATAL FAMILIAL INSOMNIA (FFI) AND KURU IN HUMANS; SCRAPIE IN SHEEP AND GOAT; BOVINE
CC  SPONGIFORM ENCEPHALOPATHY (BSE) IN CATTLE; TRANSMISSIBLE MINK ENCEPHALOPATHY (TME);
CC  CHRONIC WASTING DISEASE (CWD) OF MULE DEER AND ELK; FELINE SPONGIFORM ENCEPHALOPATHY
CC  (FSE) IN CATS AND EXOTIC UNGULATE ENCEPHALOPATHY(EUE) IN NYALA AND GREATER KUDU. THE
CC  PRION DISEASES ILLUSTRATE THREE MANIFESTATIONS OF CNS DEGENERATION: (1) INFECTIOUS (2)
CC  SPORADIC AND (3) DOMINANTLY INHERITED FORMS. TME, CWD, BSE, FSE, EUE ARE ALL THOUGHT TO
CC  OCCUR AFTER CONSUMPTION OF PRION-INFECTED FOODSTUFFS.
CC  -!- SIMILARITY: BELONGS TO THE PRION FAMILY.
KW  Prion; Brain; Glycoprotein; GPI-anchor; Repeat; Signal; Polymorphism; Disease mutation.
```

Controlled Vocabularies

- Data resources have been built introspectively for human researchers
- Information is machine readable not machine understandable
- Sharing vocabulary is a step towards unification



Ontologies in Bioinformatics

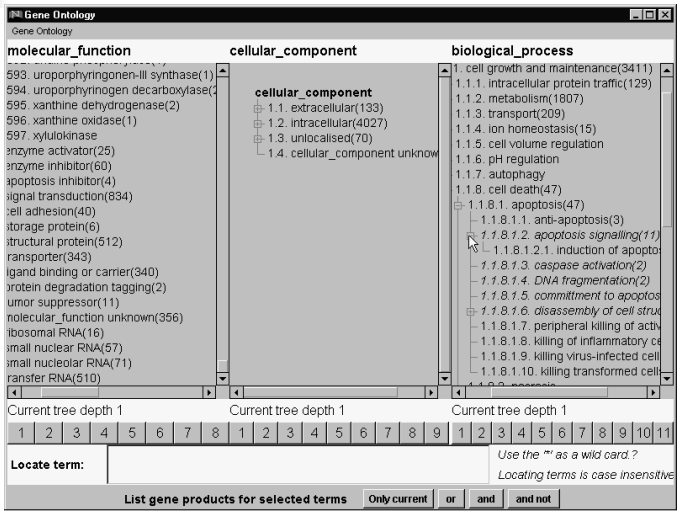
- Controlled vocabularies for genome annotation
  - Gene Ontology, MGED, Mouse Anatomy …
- Searching & retrieval
  - above + MeSH
- Communication framework for resource mediation
  - TAMBIS
- Knowledge acquisition & hypothesis generation
  - Ecocyc, Riboweb
- Information extraction & annotation generation
  - EmPathIE and PASTA
- BioOntology Consortium (BOC)

Gene Ontology

- Controlled vocabularies for the description of the molecular function, biological process and cellular component of gene products.
- Terms are used as attributes of gene products by collaborating databases, facilitating uniform queries across them.
- ~6,000 concepts

http://www.geneontology.org/

Gene Ontology



How GO is used by databases

1. Making database cross-links between GO terms and objects in their database (typically, gene products, or their surrogates, genes), and then providing tables of these links to GO;

2. Supporting queries that use these terms in their database;

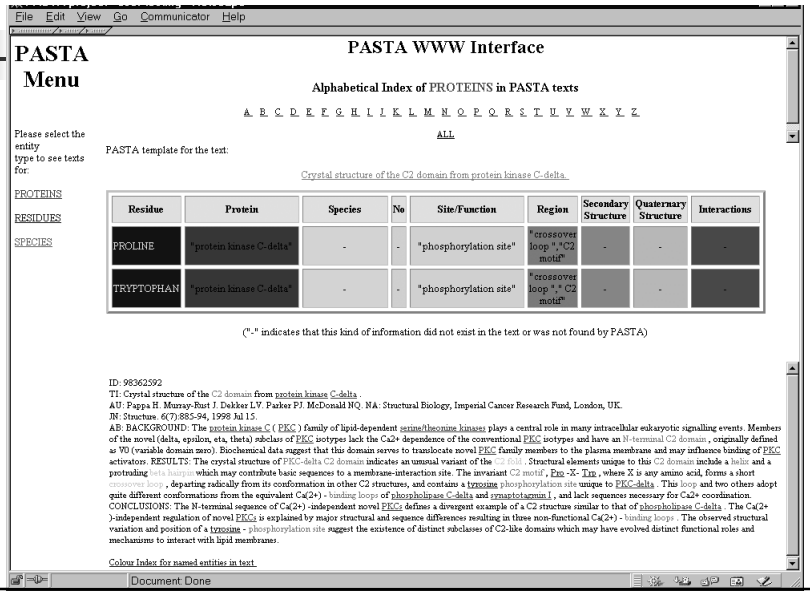
Information Extraction

- Annotation to annotation
  - Irbane: SWISS-PROT to PRINTS annotations
  - Protein Annotators Workbench
- From online searchable journal articles
  - EMPathIE: Enzyme and Metabolic Path Information Extraction
  - PASTA Protein structure extraction from texts to support the annotation of PDB
    - http://www.dcs.shef.ac.uk/research/groups/nlp/
  - PIES Protein interaction extraction system
  - BioPATH http://www.lionbioscience.com/

Research on Term Extraction in Biology

- Rule based (linguistics)
  - terminology lexicons derived from biology databases and annotated corpora
- Hybrid (statistics & linguistics)
  - pattern extraction, information categorisation using clustering, automated term recognition
- Machine Learning (Decision Trees, HMM)
- Text in Biology (BRIE & OAP) 2001
  - http://bioinformatics.org/bof/brie-oap-01/
- Natural language processing of biology text
  - http://www.ccs.neu.edu/home/futrelle/bionlp/

PASTA Protein Structure



Summary (1)

- Sequence data has a good data abstraction: the sequence
- No obvious or good abstractions for functional genomic data yet
  - Descriptive models
  - Unstable schemas
  - Retain all results in primary database just in case (e.g. microarray images)

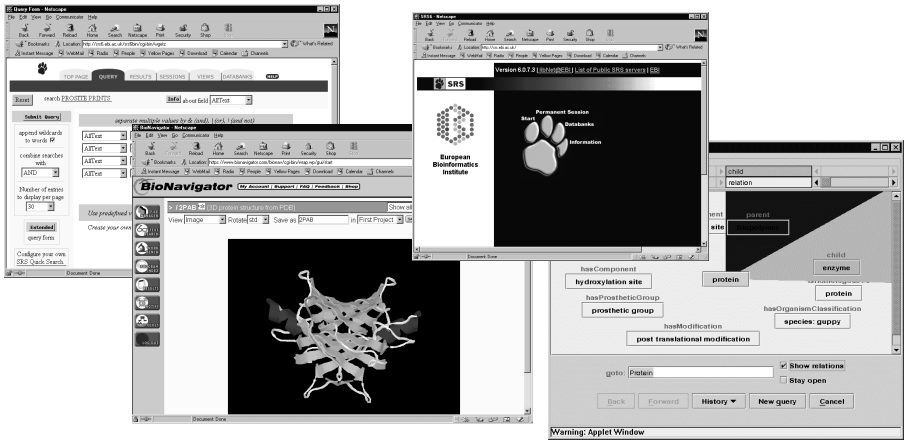
Summary (2)

- Reliance on description
  - Semi-structured data
  - Controlled vocabularies
  - Text extraction
- High value on expert curation
  - "Knowledge" warehouses
  - Labour intensive

Summary (3)

- Current dominant delivery paradigms
  - Document publication & flat files
  - Web browsing & interactive visualisation
  - Human readable vs machine understandable
- High connectivity between different databases for making links between pieces of evidence
  - Poor mechanisms for maintaining the connectivity
  - Integration considered essential

Biological Database Integration



Motivation

- Quantity of biological resources:
  - Databases.
  - Analysis tools.
- Databases represented in Nucleic Acids Research, January 2001 = 96.
- Many meaningful requests require access to data from multiple sources.

Difficulties

- All the usual ones:
  - Heterogeneity.
  - Autonomy.
  - Distribution.
  - Inconsistency.
- And a few more as well:
  - Focus on interactive interfaces.
  - Widespread use of free text.

Example Queries

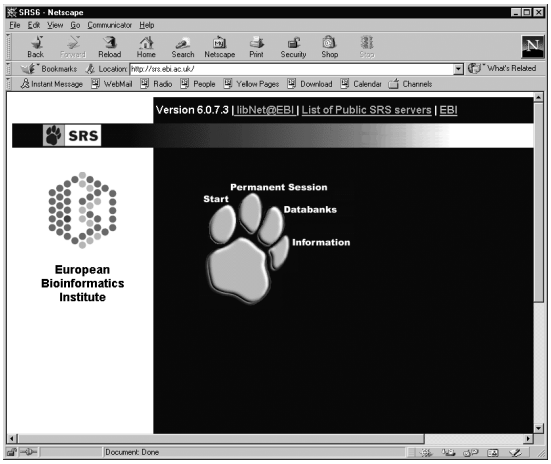
- Retrieve the motifs of proteins from *S. cerevisiae*.
- Retrieve proteins from *A. fumigatus* that are homologous to those in *S. cerevisiae.*
- Retrieve the motifs of proteins from *A. fumigatus* that are homologous to those in *S. cerevisiae.*

Possible Solutions

- Many different approaches have been tried:
  - *SRS*: file based indexing and linking.
  - *BioNavigator*: type based linking of resources.
  - *Kleisli*: semi-structured database querying.
  - *DiscoveryLink*: database oriented middleware.
  - *TAMBIS*: ontology based integration.
- Some standards are emerging:
  - OMG Life Sciences.
  - I3C.

SRS

Sequence Retrieval System
http://srs.ebi.ac.uk/

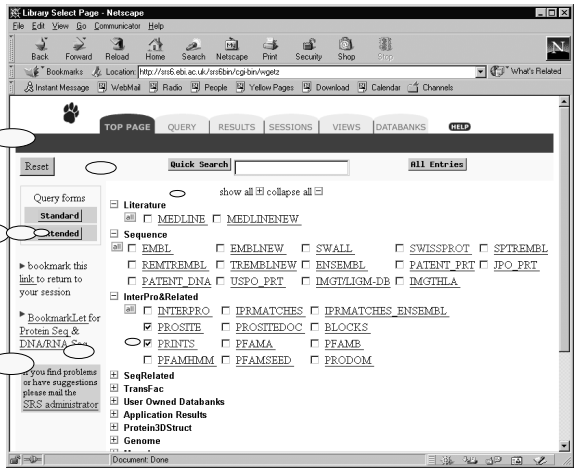


SRS In Use

List of Databases

Search Interfaces

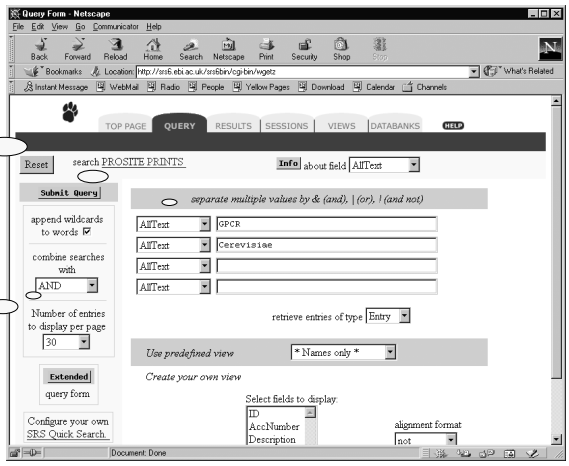
Selected Databases



Searching in SRS

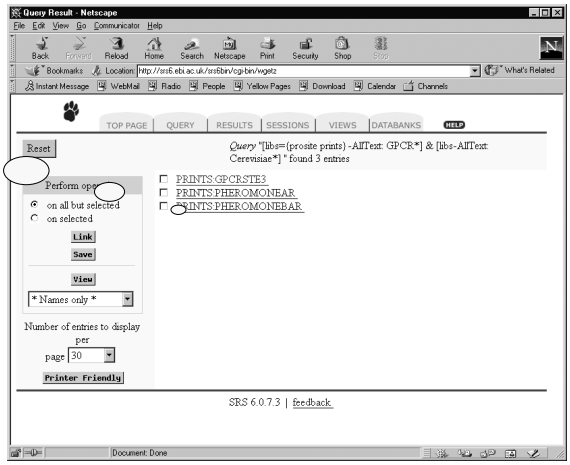
Search Fields

Boolean Condition



SRS Results

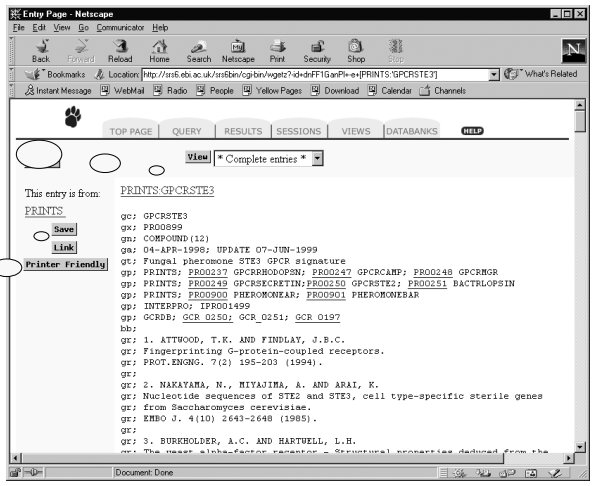
Links to Result Records



PRINTS Database Record

File Format from Source

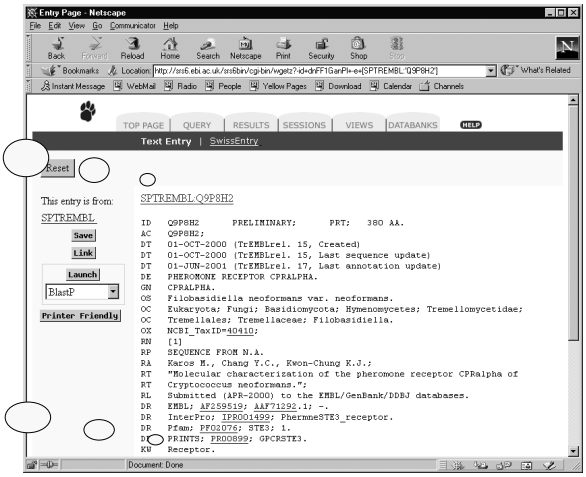
Link to Other Databases



Link Following

Related record from SPTREMBL

Reference back to PRINTS



Features of SRS

- Single access point to many sources.
- Consistent, if limited, searching.
- Fast.
- No global model, so suffers from N^2 problem linking sources.
- No reorganisation of source data.
- Minimal transparency.

BioNavigator

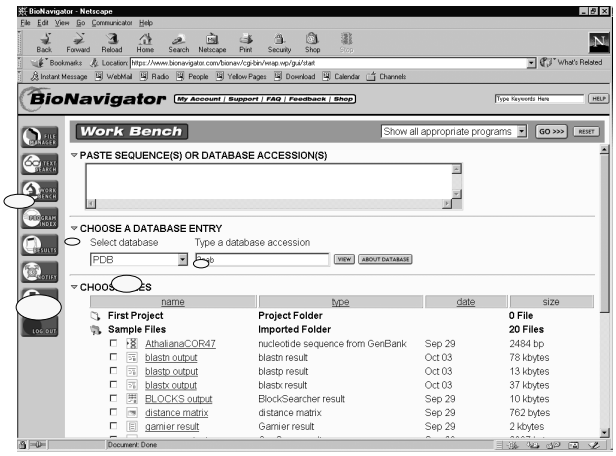
- BioNavigator combines data sources and the tools that act over them.
- As tools act on specific kinds of data, the interface makes available only tools that are applicable to the data in hand.

Online trial from:
https://www.bionavigator.com/

Initiating Navigation

Select database

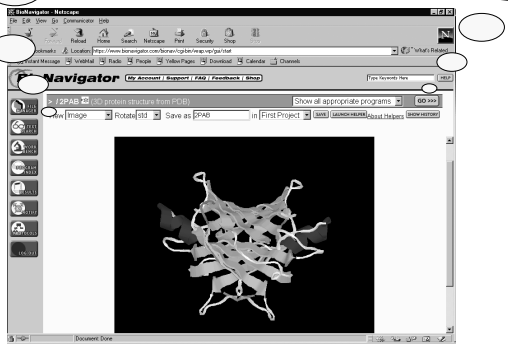
Enter accession number



Viewing Selected Data

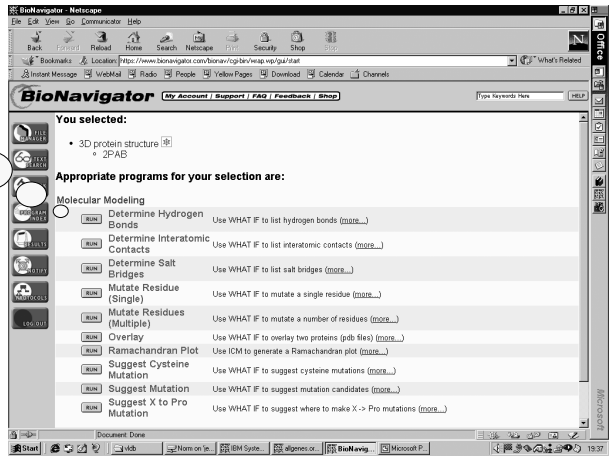
Relevant display options

Navigate to related programs



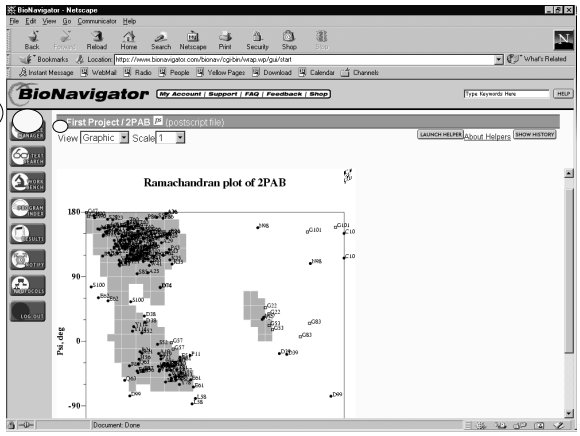
Listing Possible Applications

Programs acting on protein structures



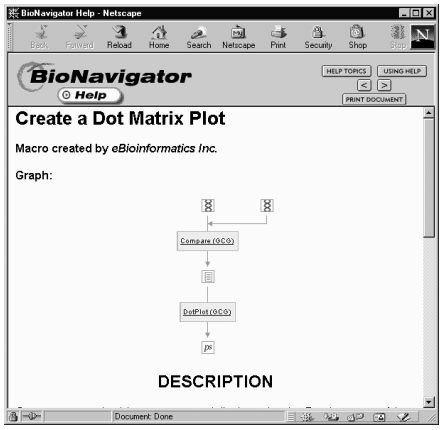
Viewing Results

Several views of result available



Chaining Analyses in Macros

Chained collections of navigations can be saved as macros and restored for later use.



Features of BioNavigator

- Single access point for many tools over a collection of databases.
- Easy-to-use interface.
- Not really query oriented.
- User selects order of access.
- Possible to minimise exposure to file formats.

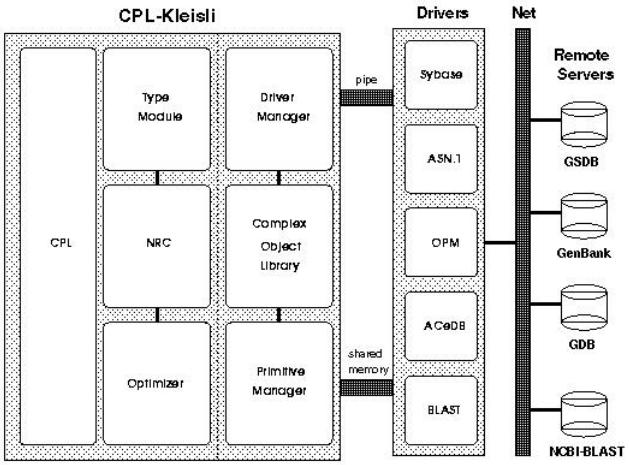
Kleisli

- Many biological sources make data available as structured flat files.
- Such structures can be naturally represented and manipulated using complex value models.
- Kleisli uses a comprehension-based query language (CPL) over such models.

Architecture

Kleisli supports client-side wrapping of sources, which surface to CPL as functions.

Online demos: http://sdmc.krdl.org.sg:8080/kleisli/demos/



Queries

- Queries can refer to multiple sources by calling driver functions.
- Example: *Which motifs are components of guppy proteins?*

```
{m |
   \p <- get-sp-entry-by-os("guppy"),
   \m <- go-prosite-scan-by-entry-rec(p)}
```

Query calls drivers from two sources

Features of Kleisli

- Query-oriented access to many sources.
- Comprehensive querying.
- No global model as such.
- Not really a user level language.
- Some barriers to optimisation.

L. Wong, Kleisli: its Exchange Format, Supporting Tools and an application in Protein Interaction Extraction, Proc. BIBE, 21-28, IEEE Press, 2000.
S.B. Davidson, et al., K2/Kleisli and GUS: Experiments in integrated access to genomic data sources, IBM Systems Journal, 40(2), 512-531, 2001.

DiscoveryLink

- DiscoveryLink ≅ Garlic + DataJoiner applied to bioinformatics.
- In contrast with Kleisli:
  - Relational not complex value model.
  - SQL not CPL for querying.
  - More emphasis on optimisation.
  - Wrappers map sources to relational model.

DiscoveryLink Example

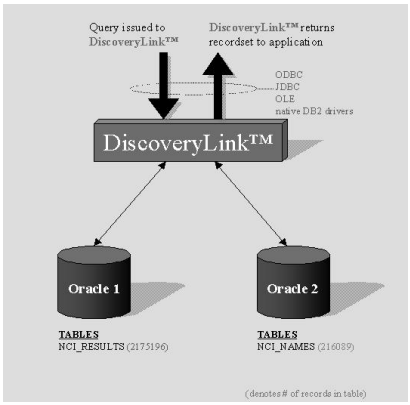
- Not much to see: SQL query ranges over tables from different databases.

```
SELECT a.nsc, b.compound_name, …
FROM nci_results a, nci_names b
WHERE panel_number = [user selected]
AND cell_number = [user selected]
AND a.nsc = b.nsc
```

Description online: www.ibm.com/discoverylink

On Relational Integration

- Relational model has reasonable presence in bioinformatics.
- More commercial than public domain sources are relational.
- Wrapping certain sources as relations will be challenging.

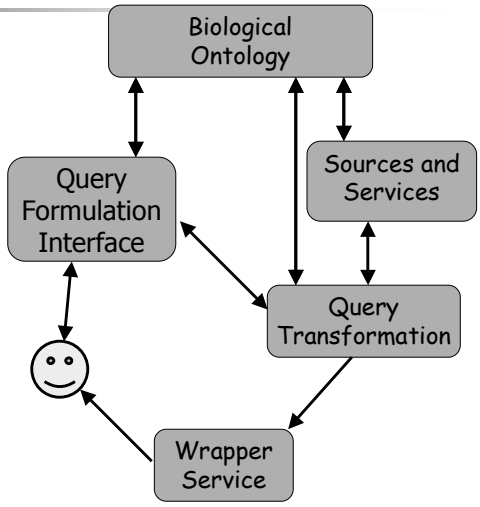


TAMBIS

- TAMBIS = Transparent Access to Multiple Bioinformatics Information Sources.
- In contrast with Kleisli/DiscoveryLink:
  - Important role for global schema.
  - Global schema = domain ontology.
  - Sources not visible to users.

TAMBIS Architecture

- Ontology described using Description Logic.
- Query formulation = ontology browsing + concept construction.
- Wrapper service = Kleisli.

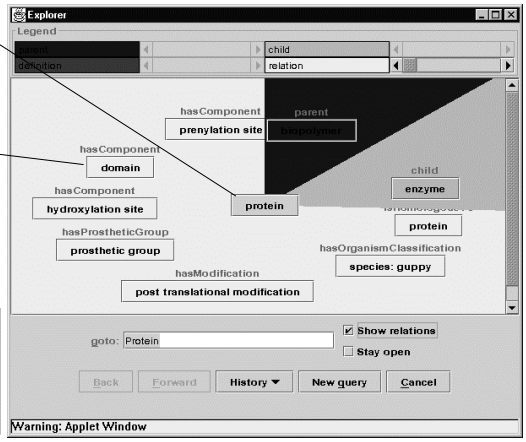


Ontology Browsing

Current Concept

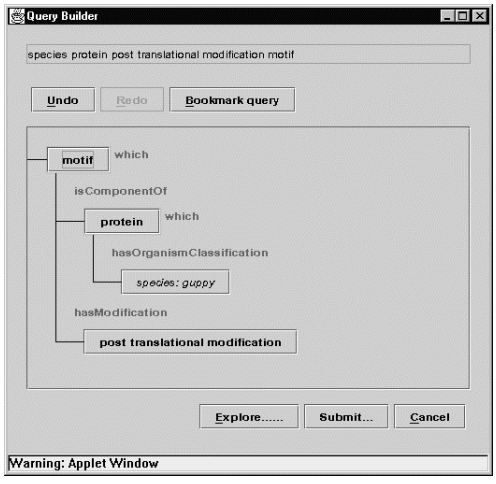
Buttons for changing current concept

Online demo: http://img.cs.man.ac.uk/tambis



Query Construction

Query = "Retrieve the motifs that are both components of guppy proteins and associated with post translational modification.



Genome Level Integration

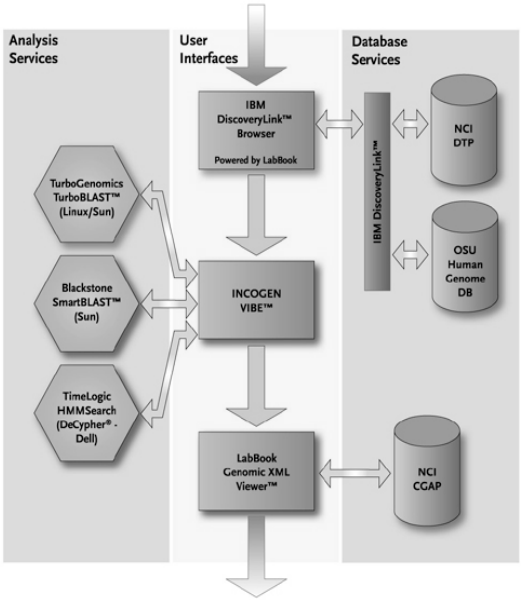
- Few integration proposals have focused on genome level information sources.
- Possible reasons:
  - Most mature sources are gene-level.
  - Lack of standards for genome-level sources.
  - Species-specific genome databases are highly heterogeneous.
  - There are few functional genomics databases.

Standardisation

- Most standards in bioinformatics have been de facto.
- The OMG has an ongoing Life Sciences Research Activity with Standardisation activities in: Sequence Analysis; Gene Expression; Macromolecular structure.
  - http://www.omg.org/homepages/lsr/
- XML approach: I3C
  - http://i3c.open-bio.org
- Open bio consortium
  - http://www.open-bio.org

I3C

- The Interoperable Informatics Infrastructure Consortium (I3C)
- Open XML-in, XML-out paradigm
- Services-based for accessing remote analysis services
- http://i3c.open-bio.org/



Business vs Biology Data Warehouses

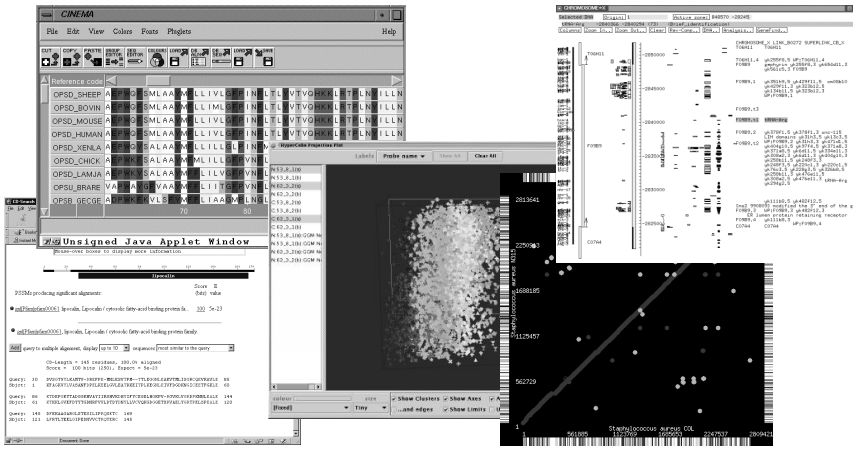
Classical Business	Biological Science
High number of queries over a priori known data aggregates	Query targets frequently change due to new scientific insights/questions
Pre-aggregation easy since business processes/models are straightforward, stable and know a priori	Pre-aggregation not easy since body of formal background knowledge is complex and growing fast
Data necessary often owned by enterprise	Most relevant data resides on globally distributed information systems owned by many organisations
Breakdown of data into N-cubes of few simple dimensions	Complex underlying data structures that are inherently difficult to reduce to many dimensions
Temporal view of data (week, month, year); snapshots	Temporal modelling important but more complex

Dubitzky et al, NETTAB 2001

Integrated Genomic Resources

- For yeast, by way of illustration:
  - MIPS (http://www.mips.biochem.mpg.de/).
  - SGD (http://genome-www.stanford.edu/Saccharomyces/).
  - YPD (http://www.proteome.com/).
- General features:
  - Integrate data from single species.
  - Limited support for analyses.
  - Limited use of generic integration technologies.

Analysing Genomic Data



Gene Level Analysis

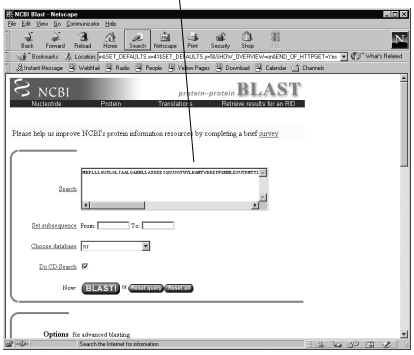
- Conventional bioinformatics provides the principal gene level analyses, such as:
  - Sequence homology.
  - Sequence alignment.
  - Pattern matching.
  - Structure prediction.

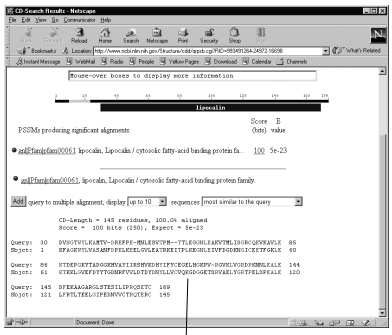
Sequence Homology

- Basic idea:
  - Organisms evolve.
  - Individual genes evolve.
  - Sequences are homologous if they have diverged from a common ancestor.
  - Comparing sequences allows inferences to be drawn on the presence of homology.
- Well known similarity search tools:
  - BLAST (http://www.ncbi.nlm.nih.gov/BLAST/).
  - FASTA (http://fasta.genome.ad.jp/).

Running BLAST

Search Sequence

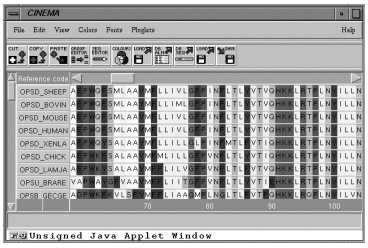




Aligned Result

Multiple Alignments

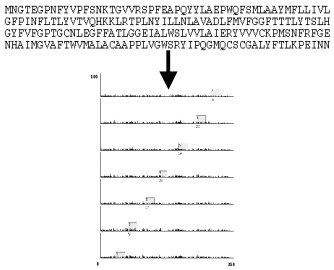
- Multiple sequences can be aligned, possibly with gaps or substitutions.
- Sequence alignment is important to the classification of sequences and to function.



CINEMA alignment applet: http://www.bioinf.man.ac.uk/dbbrowser/CINEMA2.1/

Pattern Databases

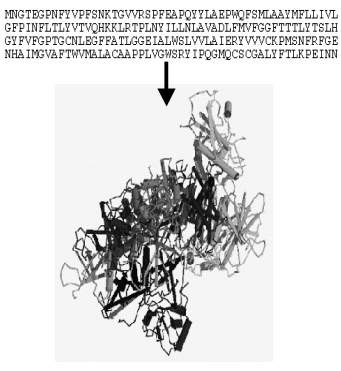
- Pattern databases are secondary databases of patterns associated with alignments.
- Conserved regions in alignments are known as motifs.



InterPro pattern database: (http://www.ebi.ac.uk/interpro/)

Protein Structure

- Structural data is important for understanding and explaining protein function.
- Predicting structure from sequence is an ongoing challenge (http://predictioncenter.llnl.gov/).



Relevance to Genome Level

- Making sense of sequence data needs:
  - Identification of gene function.
  - Understanding of evolutionary relationships.
- Genome level functional data is often understood in terms of the results of gene level analyses.
- Genome sequencing has given new impetus to gene level bioinformatics (e.g. in structural genomics http://www.structuralgenomics.org)

Genome Level Analysis

- Genome level analyses can be classified according to the data they use.
- Within a genome:
  - Individual genomic data sets.
  - Multiple genomic data sets.
- Between genomes.
  - Individual genomic data sets.
  - Multiple genomic data sets.

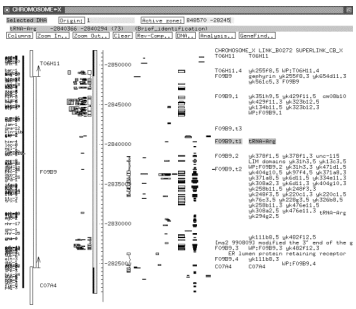
Some examples follow…

Sequencing

- Data management and analysis are essential parts of a sequencing project. Typical tasks:
  - Sequence assembly.
  - Gene prediction.
- Examples of projects supporting the sequencing activity:
  - AceDB (http://www.acedb.org/).
  - Ensembl (http://www.ensembl.org/).
- Providing systematic and effective support for sequencing will continue to be important.

ACeDB

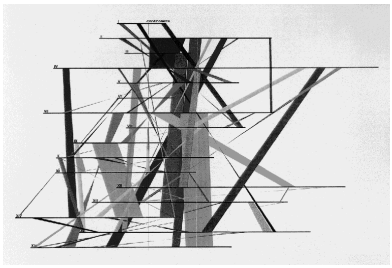
- ACeDB was developed for use in the C.Elegans genome project.
- Roles:
  - Storage.
  - Annotation.
  - Browsing.
- Semi-structured data model.
- Visual, interactive interface.



C.elegans Genome: (http://www.sanger.ac.uk/Projects/C_elegans/)

Sequence Similarity

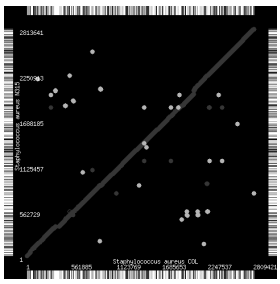
- Sequence similarity searches can be conducted:
  - Within genomes.
  - Between genomes.
- Challenges:
  - Performance.
  - Presentation.
  - Interpretation.



Visualisation of regions of sequence similarity between chromosomes in yeast.

Whole Genome Alignment

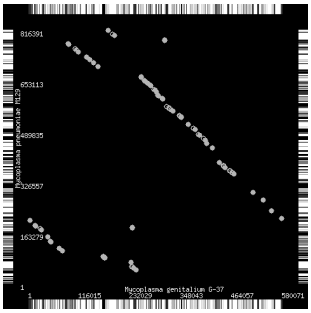
- Aligning genomes allows identification of:
  - Homologous genes.
  - Translocations.
  - Single nucleotide changes.
- Broader studies, for example, might focus on understanding pathogenicity.



Comparison of two *Staphyloccus* strains using MUMmer: (http://www.tigr.org/)

Another Genome Alignment

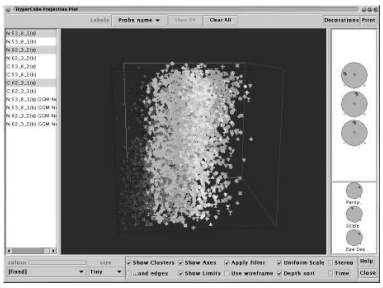
- Fast searching and alignment will grow in importance.
  - More sequenced genomes.
  - Sequencing of strains/individuals.
- Interpreting alignments requires other information.



Mycoplasma genitalium v *Mycoplasma pneumoniae*, A.L. Delcher, N. Acids Res. 27(11), 2369-2376, 1999.

Transcriptome

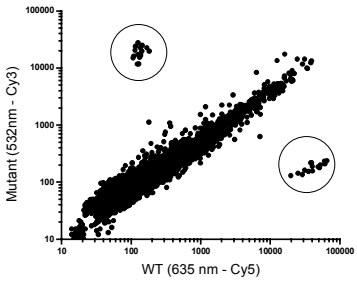
- Data sets are:
  - Large.
  - Complex.
  - Noisy.
  - Time-varying.
- Challenges:
  - Normalisation.
  - Clustering.
  - Visualisation.



maxd: http://www.bioinf.man.ac.uk/microarray/
GeneX: http://genex.ncgr.org/

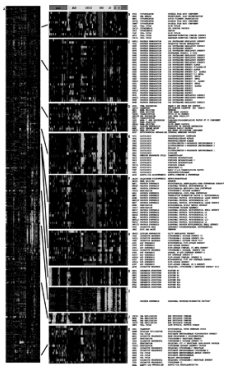
Transcriptome Results

- Dot plots allow changes in specific mRNAs to be identified.
- The example shows a comparison of two different yeast strains.



Transcriptome Clustering

- The key issue: what genes are co-regulated?
- Some techniques give absolute and some relative expression measures.
- Experiments compare expression levels for different:
  - Strains.
  - Environmental conditions.



Yeast clusters: M.B. Eisen et al., PNAS 95(25), 14863-14868, 1998.

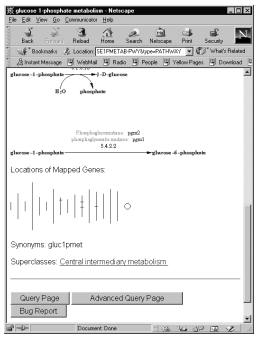
Proteome Analysis

- Driven directly from proteome-centred experiments:
  - Identification of proteins in samples.
  - Identification of post translational modifications.
- Grouping existing protein entries by:
  - Sequence similarity.
  - Sequence family.
  - Structural family.
  - Functional class.

CluS+TR:http://www.ebi.ac.uk/proteome/

Metabolome Analyses

- Analysis tasks include:
  - Searching for routes through pathways.
  - Simulating the dynamic behaviour of pathways.
  - Building pathways from known reactions.
- Other data can be overlaid on pathways (e.g. transcriptome).



EcoCyc (Frame Based): http://ecocyc.pangeasystems.com/

Integrative Analysis

- Analysing individual data sets is fine.
  - Specialist techniques often required.
  - Many research challenges remain.
- Analysing multiple data sets is necessary:
  - Understanding the whole story requires all the evidence.
  - Most important results yet to come?

Further Information

- IBM Systems Journal 40(2), 2001:
  - http://www.research.ibm.com/journal/sj40-2.html

Challenges

- The opportunities for partnership between information management providers & researchers, and biologists, is enormous.
- The challenges of genomic data are even greater than for sequence data.
- There are genuine research issues for information management.

Information representation

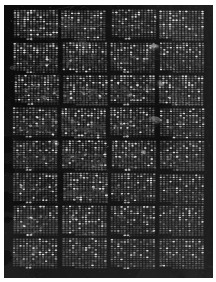
- Semi-structured description
  - Controlled vocabularies, metadata Complexity of living cells
- Context: genome is context independent and static; transcriptome, proteome etc are context-dependent and dynamic
  - Granularity: molecules to cells to whole organisms to populations

Information representation

- Spatial / temporal
  - Time-series data; cell events on different timescales
  - Gene expression spatially related to tissue

Representational forms

- A huge digital library
- Free text
  - literature & annotations
- Images
  - micro array
- Moving images
  - calcium ions waves, behaviour of transgenic mice



Quality & Stability

- Data quality
- Inconsistency, incompleteness
- Provenance
- Contamination, noise, experimental rigour
- Data irregularity
- Evolution

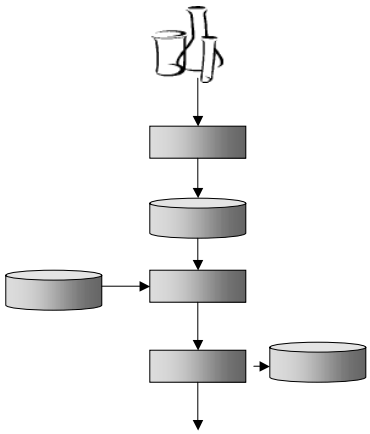
" … the problem in the field is not a lack of good integrating software, Smith says. The packages usually end up leading back to public databases. "The problem is: the databases are God-awful," he told BioMedNet.

If the data is still fundamentally flawed, then better algorithms add little"

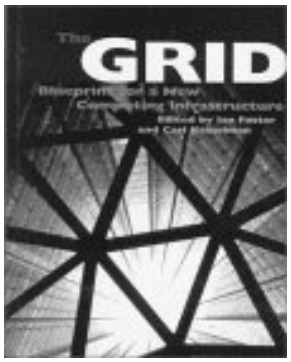
Temple Smith, director of the Molecular Engineering Research Center at Boston University, BioMedNet 2000

Process Flow

- Supporting the annotation pipeline
- Supporting *in silico* experiments
- Provenance
- Change propagation
- Derived data management
- Tracability



Interoperation



- Seamless repository and process integration & interoperation
  - The Semantic Web for e-Science
- Genome data warehouses for complex analysis
  - Distributed processing too time consuming
  - Perhaps GRIDs will solve this…?

Supporting Science

- Personalisation
  - My view of a metabolic pathway
  - My experimental process flows
- Science is not linear
  - What did we know then
  - What do we know now
- Longevity of data
  - It has to be available in 50 years time.

Prediction and Mining

- Data mining
- Machine learning
- Visualisation
- Information Extraction
- Simulation …

Final point

"Molecular biologists appear to have eyes for data that are bigger than their stomachs. As genomes near completion, as DNA arrays on chips begin to reveal patterns of gene sequences and expressions, as researchers embark on characterising all known proteins, the anticipated flood of data vastly exceeds in scale anything biologists have been used to."

(Editorial Nature, June 10, 1999)

Acknowledgements

- Help with slides:
  - Terri Attwood
  - Steve Oliver
  - Robert Stevens
- Funding:
  - UK Research Councils: BBSRC, EPSRC.
  - AstraZeneca.

Further information on bioinformatics: http://www.iscb.org/