

# Biodatabases

Jarno Tuimala / Eija Korpelainen  
CSC

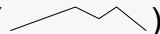


## Topics of the talk

- What data are stored in biological databases?
- What constitutes a good database?
- Nucleic acid sequence databases
- Amino acid sequence databases
- Genome databases
- Microarray databases
- Some current research trend (integration)

Modified from a Finnish slide by Eija Korpelainen

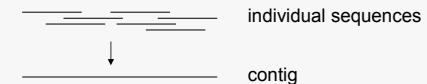
## Data types

- Sequences (... ATG GCT TTC ...)
- Motifs (A-X-[GT]-T)
- Mutations, SNPs (ACG T/A ACG)
- Gene expression profiles (  )
- Interactions (XRCC1 + PolB)
- Transcription factor binding sites (TATAA)
- etc.

Modified from a Finnish slide by Eija Korpelainen

## Some sequence terminology...

- Contig
  - several sequences are put together to form a single, longer sequence
  - typically results from sequencing projects

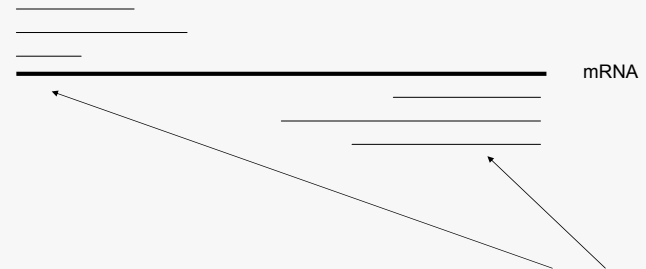


- Genomic sequence
  - a sequence that includes all elements of a genome, such as intron-exon structure

## Some sequence terminology...

- CDS
  - Coding sequence
- ORF
  - Open reading frame, a part of the genome that is transcribed into RNA

## Expressed sequence tags - ESTs



**EST:** A Short, 300-500 bp single run sequence either from the 5'- or 3'-end of the mRNA. Sequencing project like HUGO typically produce thousands of EST-sequences at a time, and are the largest submitters.

## SNPs

- ACGTACGT
- ACG**G**ACGT
- These might have effect on human disease predisposition, but then again, might not have any effect
- Used in gene mapping (finding disease genes), population genetics, etc.

## What makes a good database?

- Quality
  - Manual (slow)
  - No overlap between entries
  - Reliable
  - Some data might be missing
- Coverage
  - Automatic (fast)
  - Overlapping entries
  - Errors, biases
  - Up-to-date

Modified from a Finnish slide by Eija Korpelainen

## Database types

- Flat files (semi-structured text files)
  - Traditionally used for sequence databases
  - large indexes needed
- XML database
  - Typically extensions of flat files
- Relational databases
  - Used for gene expression and genome databases

## Genome databases: Ensembl, UCSC, MapViewer



## What are genome databases?

- Genome databases contain, well, genomic information collected from many sources.
  - Genome assembly
  - Gene predictions
  - Known genes, mRNA, ESTs, proteins
  - Genetic maps, markers and polymorphisms
  - Gene expression and phenotypes
  - Annotations
  - Interspecies homologues

## Why genome databases?

- Genome structure
- Gene identification
- Complete catalog or blueprint
- Rapid identification of proteins
- Genetic, transcriptome, proteome analysis
- Comparative genomics

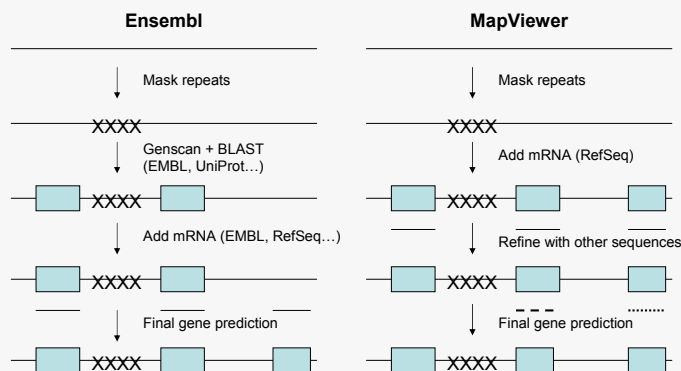
## Primary genome databases

- Ensembl
  - <http://www.ensembl.org>
  - 19 species (Chordates!)
- UCSC Genome Browser
  - <http://genome.ucsc.edu/>
  - 28 species (Insects!)
- NCBI MapViewer
  - <http://www.ncbi.nlm.nih.gov/mapview/>
  - 38 species (Plants, Fungi!)

## There's no single truth

- Number of human genes:
  - 24 194 (Ensembl)
  - 23 951 (UCSC)
  - 26 626 (MapViewer)
  - 24 625 (RefSeq mRNAs)
- And all use (almost) the same genomic assembly from 2004!
- So where is the difference?

## Gathering data



## Some considerations

- Selection of the database
  - Organism content
  - Speed (MapViewer can be slow)
- Organism specific databases can be more up-to-date than general databases
- Genome databases are not a one stop shop for all information, other databases like EMBL and UniProt are still needed

# Ensembl front page

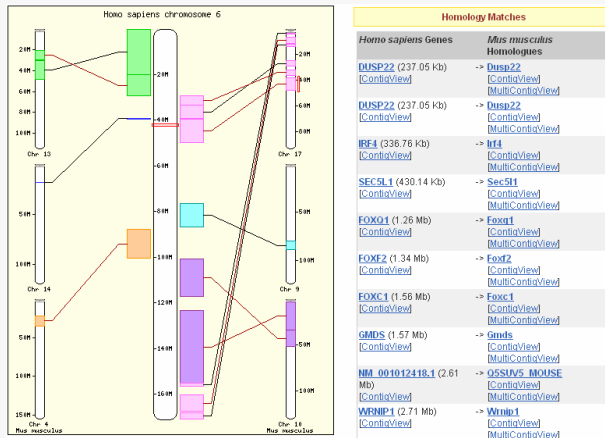
# Quick search results

# Gene View

# Gene View



## Synteny View



## Mining Ensembl

- A simple solution for mining data from Ensembl is WWW-based BioMart tool
  - For example, promoter sequences can be retrieved this way
- Direct queries from the database are also allowed using SQL

## MartView – select genome

**Select the dataset for this query**

Ensembl 33

Homo sapiens genes (NCBI35)

**Using MartView**

After choosing a DATASET above, select some FILTERS on the next page and then which data you want to EXPORT from the OUTPUT page. At any stage the COUNT button will calculate the number of entries you can expect in the final output.

MartView can generate a number of different types of output, including sequence and tabulated list data. Multiple output formats, including HTML, text and Microsoft Excel, are also supported.

**Summary**

- start
  - Not yet initialised
- filter
  - Not yet initialised
- output
  - Not yet initialised

## MartView - Filter

**REGION:**

- Chromosome: 1
- Base pair: Start 1, End 10000000
- Band: Start p36.33, End p36.33
- Marker: Start, End
- Encode type: Manual Picks
- Encode region: 11:115962315:116462315
- In encode region: Only

**GENE:**

- Disease genes: Only
- ID list limit: HGNC Symbol(s) (List: XPRCC1)

start  
Dataset: Homo sapiens genes  
34294 Entries Total  
filter  
Not yet initialised  
output  
Not yet initialised

Note, the number of genes passing the filter will appear here.

## MartView - output

new START FILTER OUTPUT export

bioMart

count help

Summary

Dataset: Homo sapiens genes  
34294 Entries Total

filter  
HGNC Symbol(s):  
Uploaded  
1 Entries pass Filters

output  
Sequences  
1 Results in Output

Select the Attribute Page

Sequences  
Features  
Structures  
SNPs  
Sequences

To Export (all in 5'-3' direction):

Unspliced (Transcript)  
Flank (Transcript)  
Flank-coding region (Transcript)  
5' UTR  
Exon sequences (Transcript)  
cDNA sequences  
Peptide

Unspliced (Gene)  
Flank (Gene)  
Flank-coding region (Gene)  
3' UTR  
Exon sequences (Gene)  
Coding sequence

Upstream flank: 100  
Downstream flank: 100

Header Information

Gene Attributes  
Chromosome  
Ensembl Gene ID (versioned)  
External Gene DB  
Description

Ensembl Gene ID  
External Gene ID  
Gene Family  
Sequence Type

What do you want to output?

1 gene found!

## MartView – promoter sequences

SEQUENCES:

Type of Sequence to Export (all in 5'-3' direction):

Unspliced (Transcript)  
Flank (Transcript)  
Flank-coding region (Transcript)  
5' UTR  
Exon sequences (Transcript)  
cDNA sequences  
Peptide

Unspliced (Gene)  
Flank (Gene)  
Flank-coding region (Gene)  
3' UTR  
Exon sequences (Gene)  
Coding sequence

Upstream flank: 1000  
Downstream flank: 100

Select transcript flank.

Specify the flank and its length.

## DNA microarray databases

## DNA microarrays

- Microarrays are used in studies assessing gene expression of hundreds or thousands of genes at a time.
  - mRNA is detected semi-quantitatively
  - DNA -> mRNA -> protein (-> money)
- One microarray typically yields data about 25000 genes (each having >2 associated variables)
- One small study might contain 10-20 microarrays
  - A rather large dataset in the end (> 100 MBs)



## DNA microarray example



Red: high expression, green: low expression, yellow: equal expression

## MIAME

- There are international standards for the microarray data
  - MIAME = minimum information about microarray experiment
  - Store wet-lab procedure, sample identities, document basic bioinformatic analyses
- Major databases aim to comply with the standard
- Standard should facilitate easier use of the data by other researchers

## Principal databases

- ArrayExpress
  - European (EBI) effort
- GEO
  - American (NCBI) effort
- Stanford
  - Stanford University database

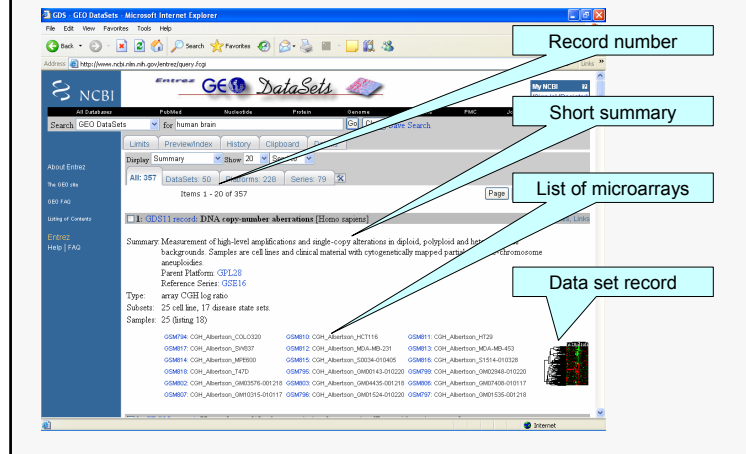
## NCBI GEO

Free text query

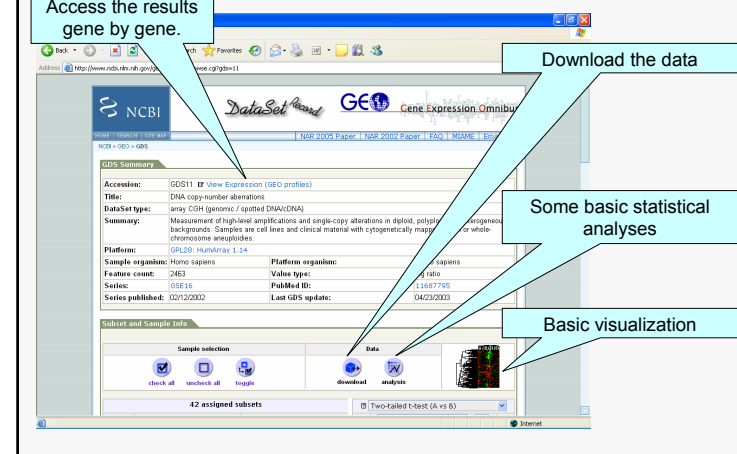
Browse

Submit data (needs an account)

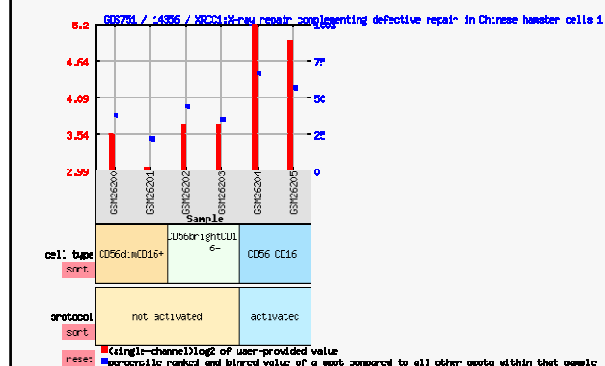
## GEO results



## Dataset record



## GEO – expression profile



## UCSC- access expression data

- UCSC genome browser has a possibility to visualize gene expression pattern in several tissues (Gene Sorter).
  - color coding as for microarray example (red and green)
- Gene Sorter can be used for other things, such as genomic proximity analyses, also.

[illegible]

# Structural databases

# Gene Sorter

Human chr19:48,739,303-48,771,555 - Gene Sorter v1.17 - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Search Favorites Home Mail News RSS Feeds

Address http://genome.ucsf.edu/cgi-bin/glsort?org=human&hg170base\_search=ucsc|line\_order=gc|chr19:48739303-48771555 Go Links

## UCSC Human Gene Sorter

genome: Human assembly: May 2004 search: NM\_006297 Goal

sort by: GO Similarity configure filter (now off) display: 50 output: sequence text

#	Name	total bases	exon bases	intron bases	UTR bases	BLASTP E-Value	Rankscore	Genome Position	Description
1	XBCO1		n/a	n/a	n/a	n/a	n/a	chr19:48,755,429	X-box repair excision complementing protein 1
2	LIG4		n/a	n/a	n/a	n/a	n/a	chr13:107,661,641	DNA ligase IV
3	REV1L		n/a	n/a	n/a	n/a	n/a	chr2:99,520,465	REV1-like
4	AY040777		n/a	n/a	n/a	n/a	n/a	chr9:32,977,140	Aspraton (Forkhead-associated domain homologue-like protein) (PHA-HIT)
5	BECAT1		n/a	n/a	n/a	0.992079	n/a	chr17:38,490,250	breast cancer 1, early onset isoform 1
6	DDI1		n/a	n/a	n/a	n/a	n/a	chr17:40,840,312	damage-specific DNA binding protein 1
7	DBP2		n/a	n/a	n/a	n/a	n/a	chr11:47,205,214	damage-specific DNA binding protein 2 (dBP2)
8	LIG3		n/a	n/a	n/a	0.0161879	n/a	chr17:30,343,795	ligase III, DNA, ATP-dependent
9	POLL		n/a	n/a	n/a	n/a	n/a	chr10:103,333,296	polymerase (DNA directed), lambda-
10	EROC2		n/a	n/a	n/a	n/a	n/a	chr2:127,749,539	excision repair cross-complementing rodent
11	MGEH		n/a	n/a	n/a	n/a	n/a	chr2:47,581,962	msh2 homolog 2
12	MEB3		n/a	n/a	n/a	0.16254	n/a	chr2:47,933,840	msh2 homolog 6
13	PAPF1		n/a	n/a	n/a	0.993399	n/a	chr1:222,878,083	poly (ADP-ribosyl) polymerase family, member 1
14	PARP4		n/a	n/a	n/a	n/a	n/a	chr1:23,939,009	poly (ADP-ribosyl) polymerase family, member 4
15	POLH		n/a	n/a	n/a	n/a	n/a	chr6:43,671,640	polymerase (DNA directed), eta
16	PAVS1T1		n/a	n/a	n/a	n/a	n/a	chr14:67,344,314	PAVS1-like 1 isoform 3

Similarity by GO ontology (checks whether the genes belong to the same pathway)

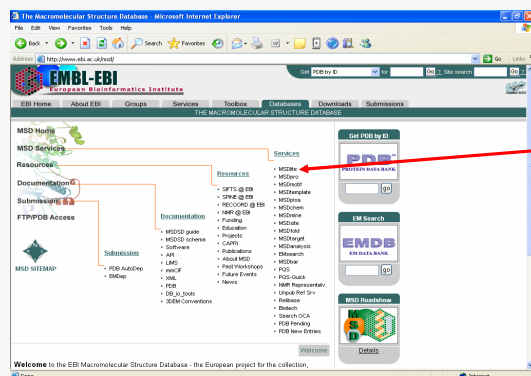
Internet

# PDB and MSD

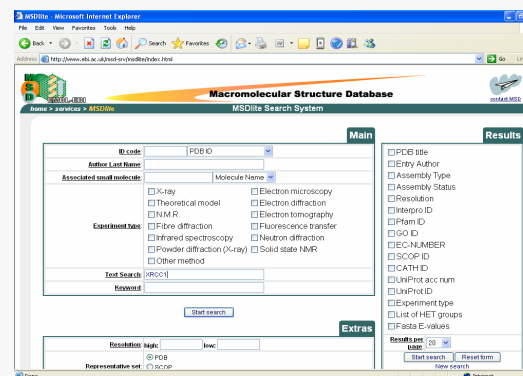
- PDB contains structures of biological macromolecules.
  - Mainly proteins, but also DNA and RNA structures
- MSD is also a collection of biological structures, but it extends the PDB data format, and circumvents some problems.

- PDB contains structures of biological macromolecules.
  - Mainly proteins, but also DNA and RNA structures
- MSD is also a collection of biological structures, but it extends the PDB data format, and circumvents some problems.

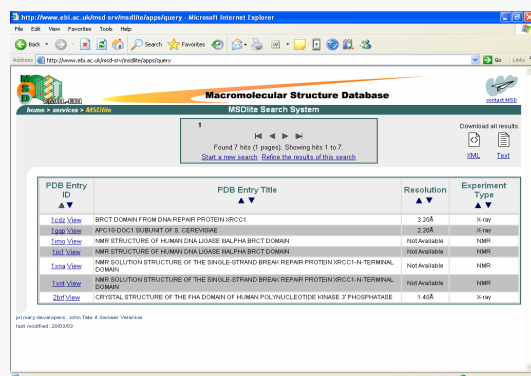
## MSD 1/5



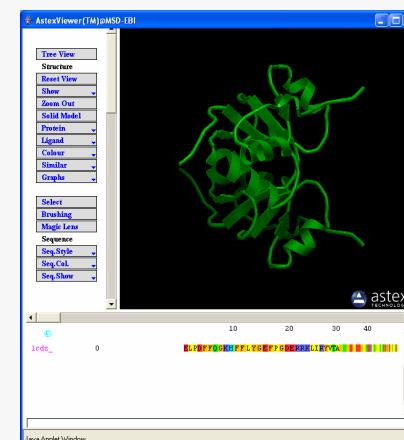
## MSD 2/5



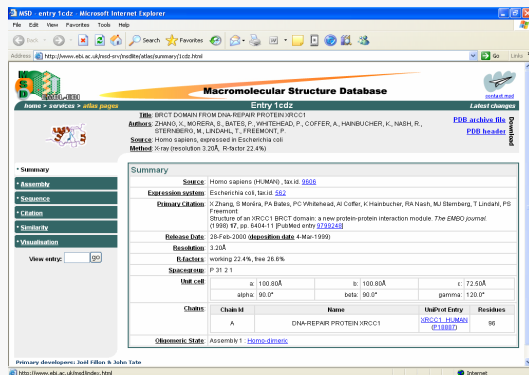
## MSD 3/5



## MSD 4/5



## MSD 5/5

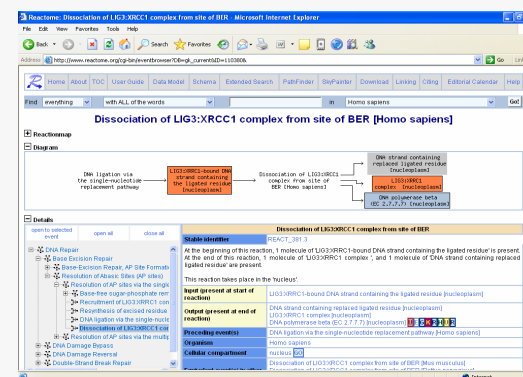


## Biological pathways

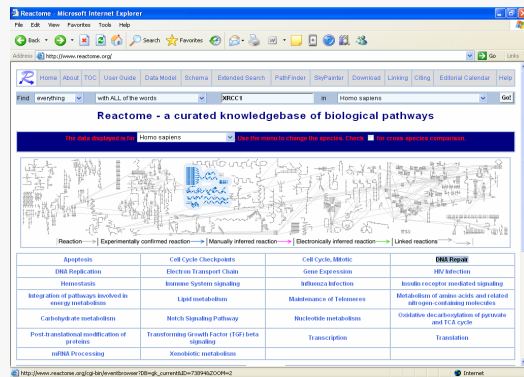
## Pathway databases

- Reactome
  - Curated
  - Pathways and reactions
- KEGG
  - Curated
  - Manually drawn pathway maps for molecular interactions and reactions
  - Used extensively
- Both contain data for several species

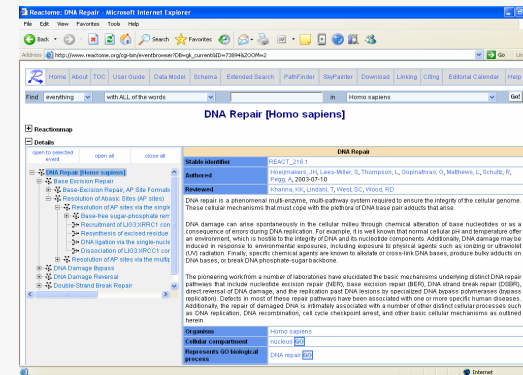
## Reactome - Find



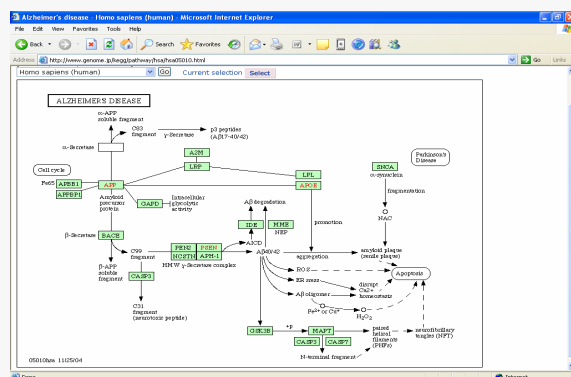
## Reactome – Highlight pathways



## Reactome – View entire pathway



## KEGG

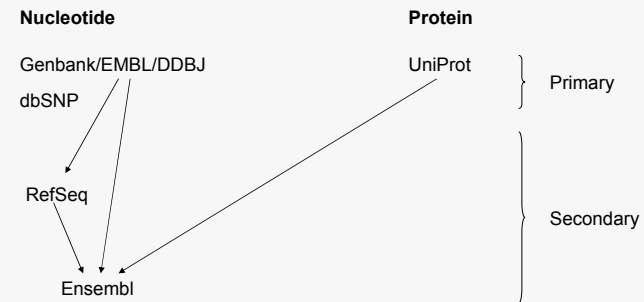


## Integrating databases

## Why integration?

- Data is distributed to several sources
  - That can prevent efficient access to data
- Genomics
  - Study of whole genomes, knowledge of gene content, expression etc. needed
- To get a better view to cells
  - Systems biology
  - Reductionism doesn't work by itself anymore, we need integration of knowledge
    - One PhD student, one gene ;(
  - Add protein studies, metabolomics, etc.

## Hierarchy of databases - an illustrative example



## About accession numbers

- Every sequence entry is individually labeled with an accession number. E.g., from Genbank you can always retrieve the same sequence, if you know the accession number.
- Accession number: alpha-numeric code
- ID: human readable sequence name
- Some examples:

XRCC1	HUGO ID
M36089	EMBL accession number
P18887	UniProt accession number
NM_006297	RefSeq, nucleotide sequence
NP_006388	RefSeq, protein sequence
Hs.98493	UniGene ID
ENSG00000073050	Ensembl, gene sequence
ENSO00000262887	Ensembl, protein sequence
7515	Locuslink ID, Entrez Gene GeneID

## Problems in integration

- Integration can't be based on accession numbers
  - Every databases use a different system
- Integration can't be based on sequences
  - Sequence is not unique
    - ACGT is a substring of ACGTACGTA and ACGTGGTATTGCTAG, so which gene does it actually represent?
- What about common terms (you wish!)

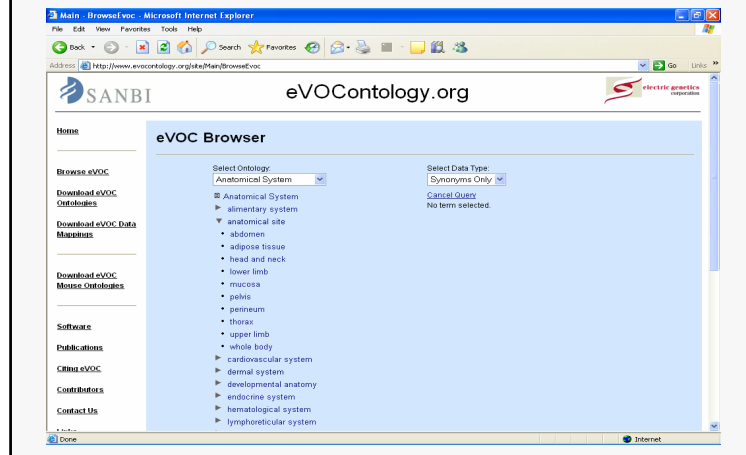
## Problems in semantic integration

- Differences in terminology
  - Vector
    - A line with a direction (math.)
    - Carrier of an infectious agent (biol., med.)
    - Virus or DNA molecule used for transferring genetic material to or from cells (biol.)
  - Breakfast cereal manufactured by Kellogg (food)
  - A rock band (music)
  - Ghost town (Final Fantasy VI)

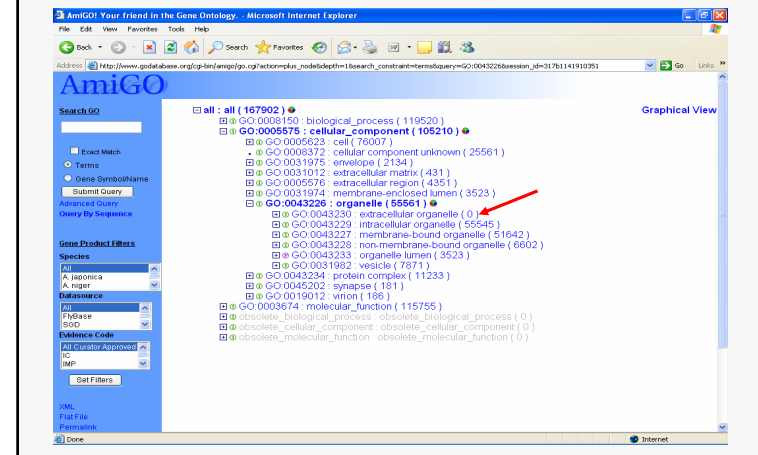
## Solutions to terminology

- Controlled vocabularies
  - A set list of terms that are used to describe certain elements
  - GO ontology: hierarchical ontology of gene functions, cellular localizations, etc.
  - eVOC ontology: describe elements of humans
- Ontologies
  - Knowledge representation systems
  - Use richer semantic terms to describe relationships between elements

## eVOC



## Gene Ontology (GO)





## Technical solutions to integration

- Data warehouse
  - All data put into the same database
- Federated database
  - Distributed processing of data
- Data grid
  - Shared databases

## Data warehouse

- Data is collected from several sources into a single database management system
- Data may be filtered or transformed to match the desired queries
- Data mart = subset warehouse for a special purpose
- Examples: EBI microarray data warehouse, Ensembl

## Warehouse - Ensembl

- Remember browsing and BioMart?
  - These are two different databases, and can return two different answers to the "same query"
  - Data behind browsing approach is normalized
  - Data in BioMart is denormalized
  - Sometimes the same gene can be returned several times for the same query even if it shouldn't; that's due to the normalization

## Warehouse – pros and cons

- Pros
  - Permits filtering and transformation
  - Might result to excellent query performance
  - Changes in remote sources do not directly affect the warehouse
- Cons
  - Heavy maintenance burden
  - Sanger center has ~1000 processors