

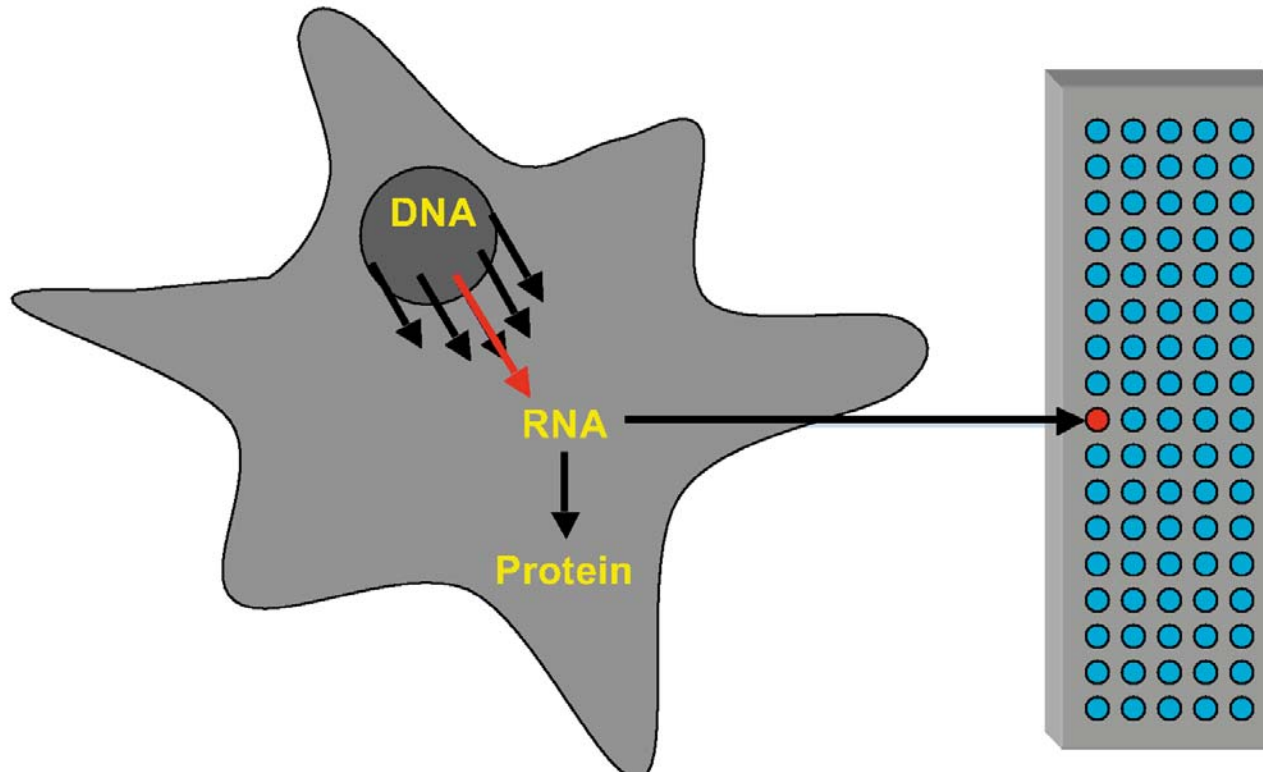
Microarray Annotation

Marc Zapatka

Computational Oncology Group
Dept. Theoretical Bioinformatics
German Cancer Research Center

2008-03-04

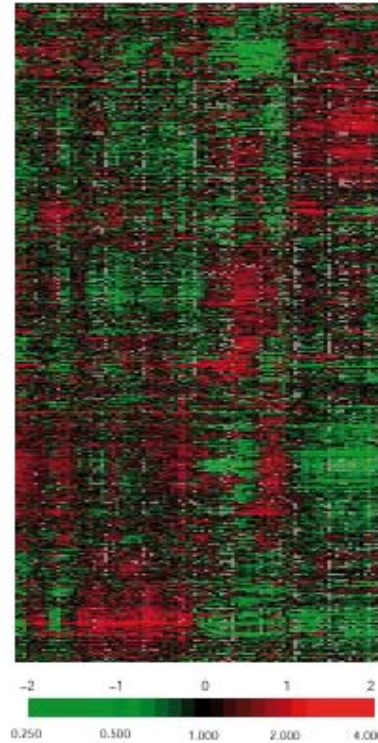
Biological Setting



There might be a correlation, but

- measured signal of DNA-microarrays \nRightarrow amount of protein
- amount of protein \nRightarrow activation status or effect of protein

Information in microarray data



Different levels and types of information

- Gene expression levels
- Gene annotations
- Sample annotations

Why do we need microarray clone annotation?

- Often, the result of microarray data analysis is a list of genes.
- The list has to be summarized with respect to its biological meaning. For this, information about the genes and the related proteins has to be gathered.
- If the list is small (let's say, 1–30), this is easily done by reading database information and/or the available literature.
- Sometimes, lists are longer (100s or even 1000s of genes). Automatic parsing and extracting of information is needed.
- To get complete information, you will need the help of an experienced computational biologist (aka bioinformatician). However, there is a lot that you can do on your own.

Primary databases

- Sequence databases

Information on genes and encoded proteins

e.g. database accession number, nucleotide and protein sequences, database cross references, and a sequence name that may or may not give a hint to the function. To find a sequence in another database, use sequence comparison tools like BLAST.

- Prominent examples of sequence data bases

The redundant databases **EMBL**, **GenBank**, and **DDBJ**.

They cover whole genome sequencing data, directly submitted sequences, sequences reported in support of patent applications and much more. Because they are so large, nobody cares about the quality of the data.

Everybody having internet access can deposit sequence information there. Errors introduced long time ago will stay there forever.

Curated databases

- In contrast, some databases are curated. That means that biologists will get the information first and compare them with literature before it goes into the database. Thus, the database is of high quality, but it takes some time until a newly discovered sequence is entered.
- Because information is only entered by curators, annotation can be unified. Rules can be put in place that say, e.g., that all enzymes cutting off phosphates are called phosphatases, not 'phosphate hydrolases'. A very famous curated database is Amos Bairoch's SWISSPROT (<http://www.expasy.org/sprot>).

Some further database examples

Meta databases collect further information and relate them to primary databases.

Examples are:

- **OMIM** (online mendelian inheritance in man) for disease-related genes
- **EntrezGene** for genomic location (integrates information from LocusLink and from genes annotated on Reference Sequences from completely sequenced genomes)
- **PFAM** for protein domain structure
- **GeneCards** for comprehensive information from other databases on human genes.

The relation of clone information to genes and proteins

- Microarrays are produced using information on *expressed sequences* as EST clones, cDNAs, partial cDNAs etc.

The relation of clone information to genes and proteins

- Microarrays are produced using information on *expressed sequences* as EST clones, cDNAs, partial cDNAs etc.
- At the other end, functional information is generated (and available) for *proteins*. Hence, there is a need to map a clone sequence ID to a protein ID. This is non-trivial.

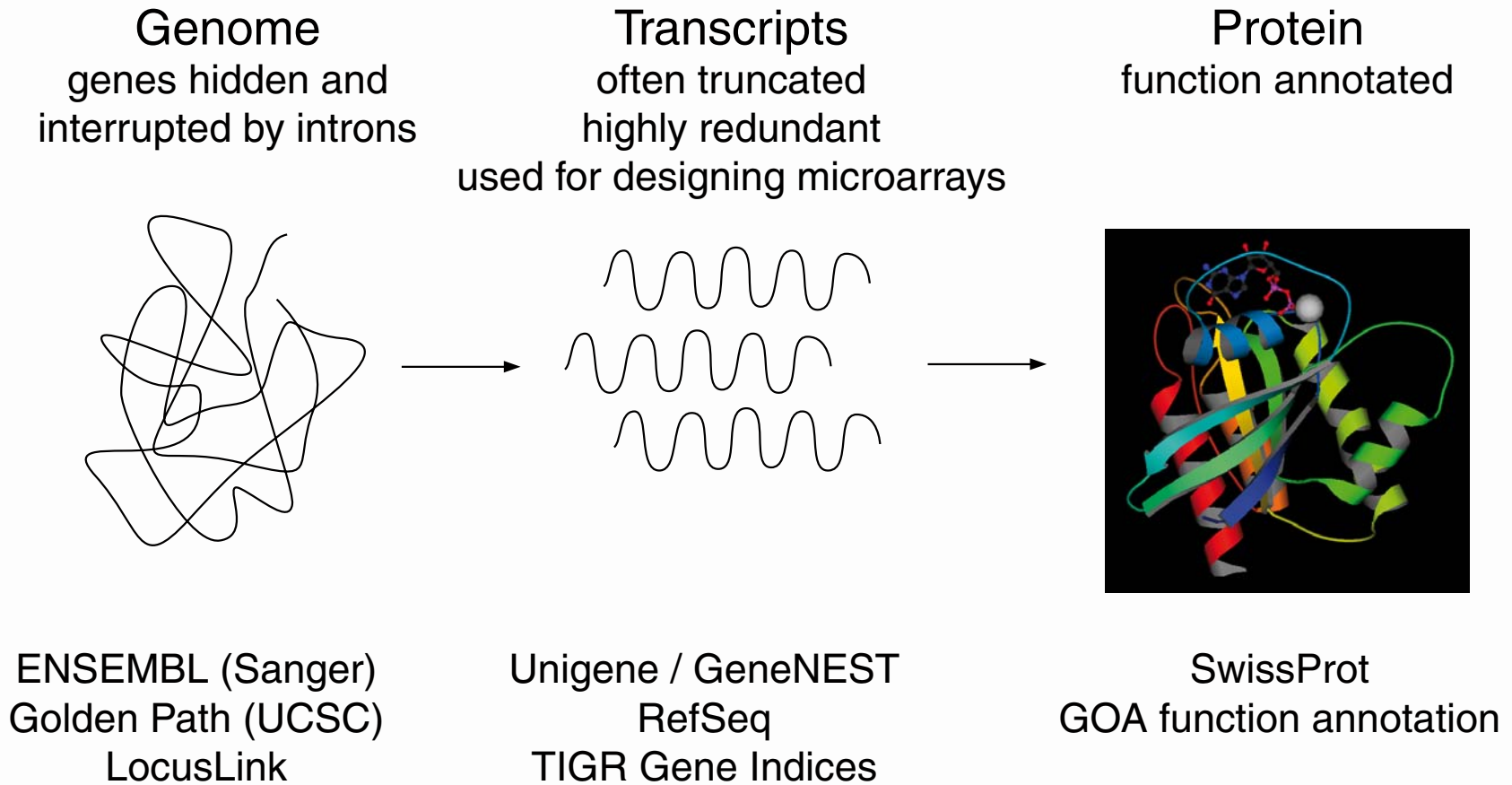
The relation of clone information to genes and proteins

- Microarrays are produced using information on *expressed sequences* as EST clones, cDNAs, partial cDNAs etc.
- At the other end, functional information is generated (and available) for *proteins*. Hence, there is a need to map a clone sequence ID to a protein ID. This is non-trivial.
- First, there are usually hundreds of ESTs (and several cDNA sequences) that map to the same gene. The Database *Unigene* tries to resolve this clustering by sequence clustering.

The relation of clone information to genes and proteins II

- *Locus Link*(superseeded by Entrez Gene)
 - Quite but not stable repository of genomic loci, supposed to be a single gene
 - Emphasis on well-characterised loci (not complete)
- **Entrez Gene**
 - Contains results of RefSeq, model organism databases and NCBI databases
 - Increased taxonomy scope over Locus Link
 - Improved access tools
 - **Unique stable tracked** integers as **identifiers** (even over organisms)
 - Outlinks to protein names, gene structure and sequence, functional annotation (domain content, CDD), GO, KEGG, HIV Interactions, OMIM, homology information

- There are other projects like RefSeq (NCBI) or TIGR Gene Indices. According to the cross-references available for a certain microarray, one or the other may be advantageous.



The Human Genome Sequence

- With the completion of the human genome sequence, you'd think that such ambiguities can be resolved. In fact, that is not the case.

The Human Genome Sequence

- With the completion of the human genome sequence, you'd think that such ambiguities can be resolved. In fact, that is not the case.
- Part of the problem is due to the fact that it is hard to predict gene structure (intron/exon) without knowing the entire mRNA sequence, which happens for about two-thirds of all genes.

The Human Genome Sequence

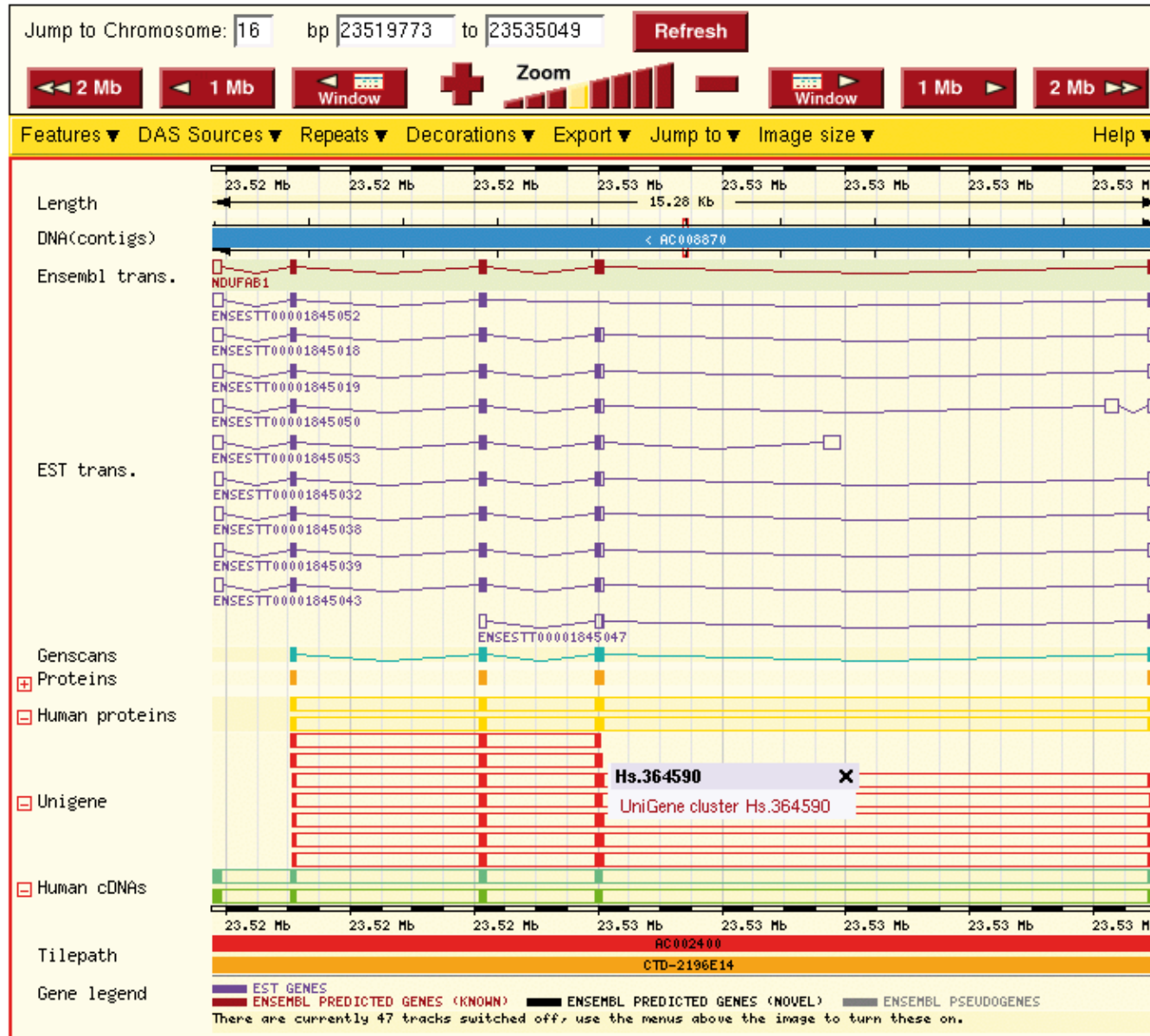
- With the completion of the human genome sequence, you'd think that such ambiguities can be resolved. In fact, that is not the case.
- Part of the problem is due to the fact that it is hard to predict gene structure (intron/exon) without knowing the entire mRNA sequence, which happens for about two-thirds of all genes.
- Then, there are errors in the assembly (putting together the sequence snippets). A typical symptom is that a gene appears to map to multiple loci on the same chromosome, with very high sequence similarity.

The Human Genome Sequence

- With the completion of the human genome sequence, you'd think that such ambiguities can be resolved. In fact, that is not the case.
- Part of the problem is due to the fact that it is hard to predict gene structure (intron/exon) without knowing the entire mRNA sequence, which happens for about two-thirds of all genes.
- Then, there are errors in the assembly (putting together the sequence snippets). A typical symptom is that a gene appears to map to multiple loci on the same chromosome, with very high sequence similarity.
- But there are also sequences that are nearly identical, but duplicated. This has happened not long ago in evolution by means of transposable elements.

Genomic mapping: ENSEMBL Browser

Detailed View



Some figures

- Currently, it's estimated that the human genome contains about 25,000 – 30,000 genes that code for 50,000 – 100,000 different transcripts (and thus, proteins).
- Unigene (human section) contains 83,896 clusters, but 46492 of them are of size 2 or less.
- RefSeq DNA contains 28,118 human sequences (3,295 EST's, 11,972 predicted seq., 17,708 mRNA's).
- ENSEMBL contains 22,205 predicted genes, 49,134 predicted transcripts. Fully computational methods like Genscan produce more than 65,000 predictions.
- Entrez Gene contains 38,603 genes.

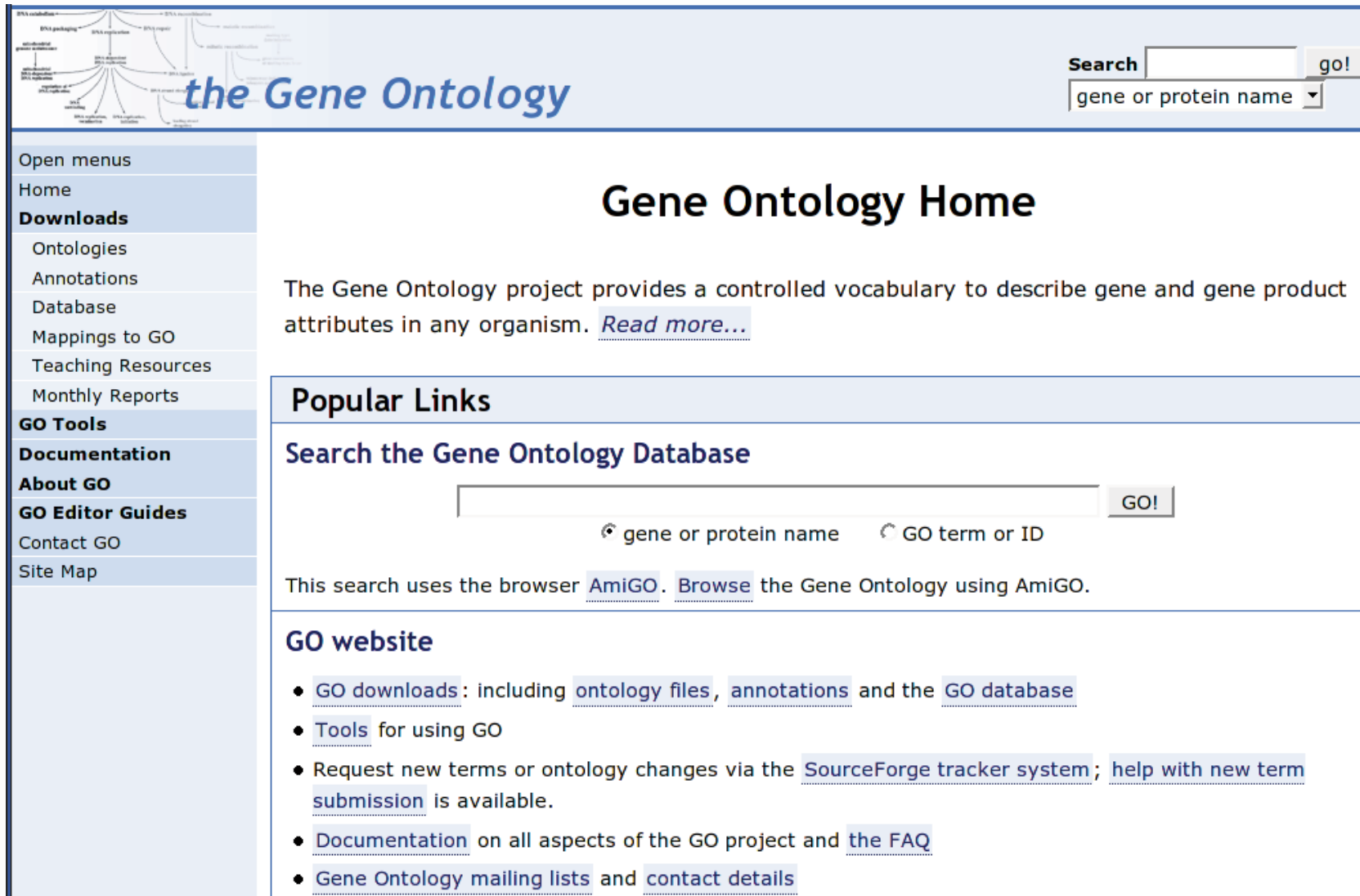
Function annotation

- Probably, the most important thing you want to know is what the genes or their products are concerned with, i.e. their **function**.
- Function annotation is difficult: Different people use different words for the same function, or may mean different things by the same word. The context in which a gene was found (e.g. “TGF β -induced gene”) may not be particularly associated with its function.
- Inference of function from sequence alone is error-prone and sometimes unreliable. The best function annotation systems (GO, SwissProt) use human beings who read the literature before assigning a function to a gene.

The Gene Ontology system

- To overcome some of the problems, an annotation system has been created: Gene Ontology (<http://www.geneontology.org>). Ontology means here the art (or science) of giving everything its correct name.
- It represents a unified, consistent system, i.e. terms occur only once, and there is a dictionary of allowed words.
- Furthermore, terms are related to each other: the hierarchy goes from very general terms to very detailed ones.
- Subsections are
 - biological process
 - cellular component
 - molecular function

The Gene Ontology site



the Gene Ontology

Search go!
gene or protein name ▾

Gene Ontology Home

The Gene Ontology project provides a controlled vocabulary to describe gene and gene product attributes in any organism. [Read more...](#)

Popular Links

Search the Gene Ontology Database

GO!

🔍 gene or protein name 🔍 GO term or ID

This search uses the browser [AmiGO](#). [Browse](#) the Gene Ontology using AmiGO.

GO website

- [GO downloads](#): including [ontology files](#), [annotations](#) and the [GO database](#)
- [Tools](#) for using GO
- Request new terms or ontology changes via the [SourceForge tracker system](#); [help with new term submission](#) is available.
- [Documentation](#) on all aspects of the GO project and [the FAQ](#)
- [Gene Ontology mailing lists](#) and [contact details](#)

The Gene Ontology hierarchy

AmiGO

Last updated: 2005-10-09

serine-type endopeptidase inhibitor activity

Accession: GO:0004867

Ontology: molecular_function

Synonyms:

related: serpin

exact: serine protease inhibitor activity

exact: serine proteinase inhibitor activity

exact: serpin activity

Definition:

Stops, prevents or reduces the activity of serine-type endopeptidases, enzymes that catalyze the hydrolysis of nonterminal peptide linkages in oligopeptides or polypeptides; a serine residue (and a histidine residue) are at the active center of the enzyme.

Comment: None

Term Lineage

all : all (<215714)

① GO:0003674 : molecular_function (<160720)

① GO:0030234 : enzyme regulator activity (<2409)

① GO:0004857 : enzyme inhibitor activity (<771)

① GO:0030414 : protease inhibitor activity (<438)

① GO:0004866 : endopeptidase inhibitor activity (<414)

① **GO:0004867 : serine-type endopeptidase inhibitor activity (<274)**

[Graphical View](#)

Actual annotation

- Gene Ontology by itself is only a system for annotating genes and proteins. It does not relate database entries to a special annotation value.
- Luckily, research communities for several model organisms have agreed on entering Gene Ontology information into the databases. As this is done 'by hand', GO annotation for most organisms is far from complete.

The NetAffx System

- For Affymetrix arrays, annotation is provided by the supplier via the NetAffx system (<http://www.affymetrix.com/analysis/netaffx/>)

The screenshot displays the NetAffx website interface. At the top, the Affymetrix logo is visible on the left, and navigation links for 'PRODUCTS & APPLICATIONS', 'SUPPORT', 'NETAFFX', 'SCIENTIFIC COMMUNITY', and 'CORPORATE' are in the center. On the right, there is a link for 'アフィメトリス・ジャパンへはこちら'. The main content area is titled 'QUERY' and 'Getting Started'. It is divided into three sections: 'Exon Array', 'Expression', and 'Genotyping'. Each section contains a list of search options with brief descriptions. The 'Exon Array' section includes: 'Search all available information in the database for a particular term or identifier.', 'Search for Probesets using specific fields in the database for a term or identifier [Standard Query]', 'Search for Exon Clusters using specific fields in the database for a term or identifier [Standard Query]', 'Search for Transcript Clusters using specific fields in the database for a term or identifier [Standard Query]', 'Retrieve annotations for a probe list [Batch Query]', and 'Find probes that identically match your sequence(s) [Probe Match]'. The 'Expression' section includes: 'Search all available information in the database for a particular term or identifier. This is recommended as a starting point for your searches. [Quick Query]', 'Search specific fields in the database for a term or identifier [Standard Query]', 'Retrieve annotations for a probe list [Batch Query]', 'Find probe sets that align to your sequence(s) through BLAST [BLAST]', 'Find probes that identically match your sequence(s) [Probe Match]', and 'Query the UCSC Browser for genomic alignment [UCSC Query]'. The 'Genotyping' section includes: 'Search all available information in the database for a particular term or identifier. This is recommended as a starting point for your searches. [Quick Query]', 'Search specific fields in the database for a term or identifier [Standard Query]', 'Retrieve annotations for a probe list [Batch Query]', 'Query the UCSC public genome by position [UCSC Query]', and 'Search for SNPs between microsatellites [SNP Finder]'. At the bottom of the main content area, there is a 'Begin' button with a right-pointing arrow.

Alternative pre-compiled annotation

- The Computational Biology and Functional Genomics Laboratory at the Dana-Farber Institute has its own pre-compiled annotation for most commercial arrays (Affymetrix, Agilent, Incyte etc.): <http://compbio.dfci.harvard.edu/tgi/cgi-bin/magic/r1.pl>



The screenshot shows the website for the Computational Biology and Functional Genomics Laboratory. The header features a logo on the left, the text "Computational Biology and Functional Genomics Laboratory" in the center, and "The Gene Index Project" on the right. Below the header is a navigation bar with links: Home, Resourcerer, Gene Indices, Genomic Maps, and EGO. A secondary navigation bar includes Plant Resourcerer, Batch Search, QTL, Marker Search, What's New, and READ ME.

Resourcerer

Please note that we are currently rebuilding Resourcerer. Version 13.0 will be available soon.

RESOURCERER ([Genome Biology 2001 PDF](#)) provides annotation based on The Gene Indices (TGI) for commonly available microarray resources, including widely used clone sets and Affymetrix GeneChip Arrays.

RESOURCERER also allows comparisons between resources from the same species using TGI, UniGene, LocusLink, or RefSeq and between species using the EGO database.

RESOURCERER is updated every four months (March, July, November) following TGI and EGO updates. Requests to include new resources should be made at least one month prior to the update.

RESOURCERER data (single resource annotation) is available at [DFCI public ftp site](#).

The Gene Indices identifiers and processes are described at [TGI FAQ Page](#).

Resourcerer 12.0 July 2005 Release
Select a single resource in Data Set A:
Human: affy_HG-U95Av2

- Annotation for Data Set A
- Compare to another resource
- EMBL-EBI GO Slim Analysis
- Genome Mapping Analysis
- Pull 2KB Upstream Sequences

Submit

Data packages in Bioconductor

Bioconductor Task View: ChipManufacturer

Subview of

- [AnnotationData](#)

Subviews

- [AffymetrixChip](#)
- [AgilentChip](#)
- [ClontechChip](#)
- [GECip](#)
- [INDACChip](#)
- [QiagenChip](#)
- [RNG_MRCChip](#)
- [RocheChip](#)

Packages in view

Package	Maintainer	Title
adme16cod	Diego Diez	ADME Rat 16-Assay Bioarray Annotation Data
ag	Biocore Data Team	Affymetrix Arabidopsis Genome Array Annotation Data (ag)
agcdf	Biocore Data Team	agcdf
agprobe	Biocore Data Team	Probe sequence data for microarrays of type ag
arrayannotation	Manhong	Array annotation data of custom CDF

Software packages in Bioconductor

Package	Maintainer	Title
affycoretools	James W. MacDonald	Functions useful for those doing repetitive analyses with Affymetrix GeneChips.
altcdfenvs	Laurent Gautier	alternative cdfenvs
annaffy	Colin A. Smith	Annotation tools for Affymetrix biological metadata
AnnBuilder	J. Zhang	Bioconductor annotation data package builder
annotate	Biocore Team	Annotation for microarrays
biomaRt	Steffen Durinck	Interface to BioMart databases (e.g. Ensembl)
Category	S. Falcon	Category Analysis
ChromoViz	Jihoon Kim	Multimodal visualization of gene expression data
DynDoc	Biocore Team	Dynamic document tools
ecolink	Laurent	Meta-data and tools for E. coli
gaggle	Paul Shannon	Broadcast data between R and Java bioinformatics programs
GeneR	Y. d'Aubenton-Carafa	R for genes and sequences analysis
GlobalAncova	R. Meister	Calculates a global test for differential gene expression between groups
globaltest	Jelle Goeman	Testing Association of Groups of Genes with a Clinical Variable
GOstats	S. Falcon	Tools for manipulating GO and microarrays.
goTools	Agnes Paquet	Functions for Gene Ontology database
KEGGSOAP	J. Zhang	Client-side SOAP access KEGG
matchprobes	Biocore Team	Tools for sequence matching of probes on arrays
nem	Florian Markowetz	Nested Effects Models to reconstruct phenotypic hierarchies
Resourcerer	Jianhua Zhang	Reads annotation data from TIGR Resourcerer or convert the annotation data into Bioconductor data package.
RMAPPER	VJ Carey	interface to mapper.chip.org
RSNPper	VJ Carey	interface to chip.org::SNPper for SNP-related data
simpleaffy	Crispin Miller	Very simple high level analysis of Affymetrix data

Bioconductor metadata packages

- These packages contain one-to-one and one-to-many mappings for frequently used chips, especially Affymetrix arrays.
- Information available includes gene names, gene symbol, database accession numbers, Gene Ontology function description, enzyme classification number (EC), relations to PubMed abstracts, and others.
- The data use the framework of the `annotate` package, so I will briefly explain how it works.

Environments in R

- To quickly find information on one subject in a long list, a data structure called *hash table* is frequently used in computer science.
- A hash table is a list of key/value pairs, where the key is used to find the corresponding value. To go the other way round, you have to use pattern matching, which is much slower.
- In R, hash tables are implemented as *environments*. For the moment, we do not care about the philosophy behind it and simply treat it as another word for hash table.

Setting up environments

To set up a new environment:

```
symbol.hash = new.env(hash=TRUE)
```

To create a key/value pair:

```
assign("1234_at", "EphA3", env=symbol.hash)
```

To list all keys of an environment:

```
ls(env=symbol.hash)
```

To get the value for a certain key:

```
get("1234_at", env=symbol.hash)
```

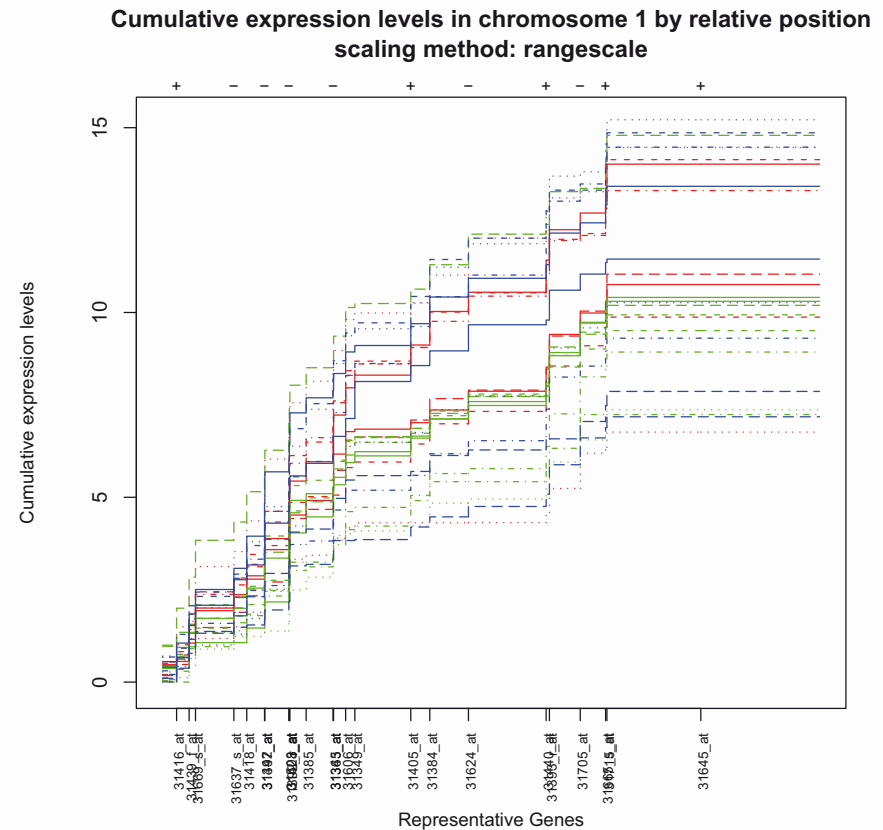
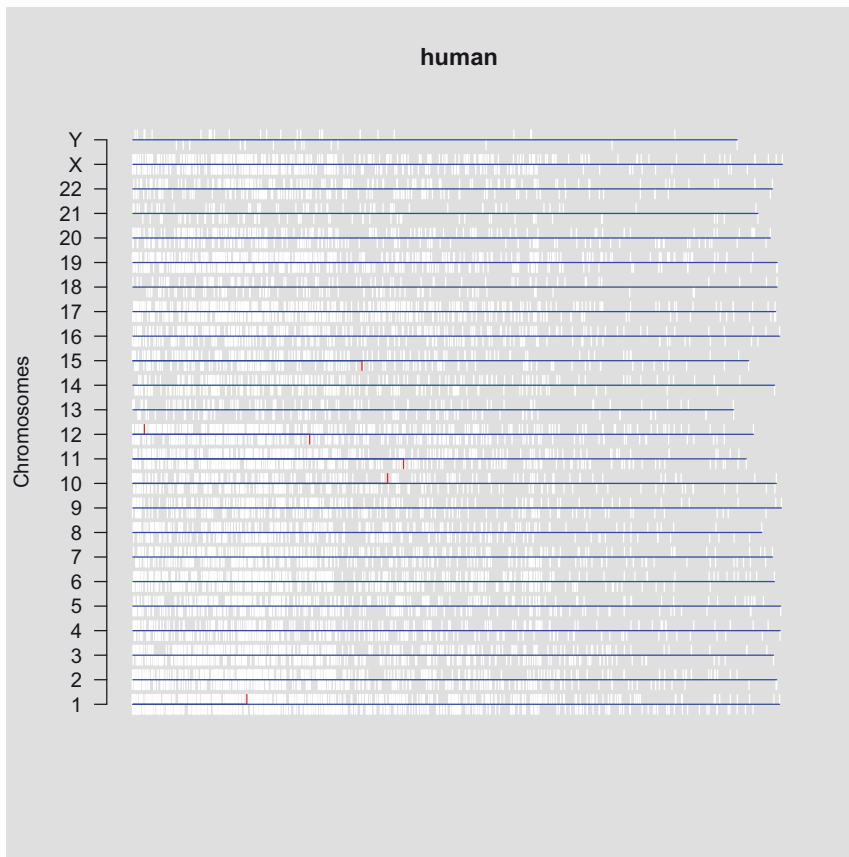

The annotate package

- That's all standard R. The annotate package gives one further function, `mget`, which retrieves more than one entry at a time, and definitions for special data, e.g. PubMed abstracts, or chromosomal location objects.
- ChromLoc objects are quite useful if you want to associate gene expression with certain positions on a chromosome, e.g. if aberration occurs in your samples.
- You can construct a ChromLoc object on your own (→ Vignette), or use the function `buildChromLocation`. For chip HGU95a_v2:

```
library(hgu95av2)
cl.95a = buildChromLocation("hgu95av2")
```

Plots for ChromLocation objects

- Plotting methods are available via library `geneplotter` or *ChromoVis*.



How to get annotation for a set of genes

- Suppose you have found some interesting genes. The index in the matrix is in `index.int`. To get the gene names:

```
gnam.int = geneNames(exprset)[index.int]
```

- To find the description:

```
mget(gnam.int, env=hgu95av2GENENAME)
```

- To get EC Numbers (relating to KEGG pathways):

```
mget(gnam.int, env=hgu95av2ENZYME)
```

Some caveats

- Because of the non-unique matching of sequences to the genome, array features are sometimes annotated with more than one position:

```
a = ls(env=hgu95av2CHRL0C)
```

```
table(sapply(mget(a, env=hgu95av2CHRL0C),  
length))
```

1	2	3	4	5	6	7	9
11520	856	156	57	20	9	4	3

- For the 1000 or so sequences with more than one location, only the first one is used, although there is no warning. It should be desirable to resolve the ambiguities by hand, but nobody has done yet.

Some caveats

- Looking at the number of chromosomal annotations

```
table(sapply(mget(a, env=hgu95av2CHRL0C),  
            function(x){length(unique(names(x)))}))  
      0      1      2      3  
907 11662      55      1
```

There are even 56 probe sets on HGU95A_v2 that map to 2 or more chromosomes; however, most of these are located on some special extrachromosomal segment and annotated with “X” and “Y”.

- Special annotation package for Affymetrix arrays
- Provides simplified mappings between Affymetrix IDs and annotation data
- Relies on chip-level annotation packages created by **AnnBuilder**
- Supplies functions to produce mappings for almost all environments in a given annotation

- Enables to query the BioMart databases Ensembl, VEGA (Vertebrate Genome Annotation), dbSNP, sequence mart (Ensembl genome sequences)
- Two sets of functions
 - Information retrieval from BioMart databases
<http://www.biomart.org>
 - Functions to access Ensembl <http://www.ensembl.org>
- Supplies annotation of features on arrays concerning affy ids, locuslink, RefSeq, entrezgene, gene names, GO, OMIN, ...

Pattern matching

- To find something in character vectors or character lists, some pattern matching is required.
- If you have real full names, use `match`, e.g.

```
match("1234_at", rownames(exprs(exprset)) )
```
- This will give you the index of ‘ ‘1234_at’ ’. It works also with more than one gene:

```
match(gnam.int, rownames(exprs(exprset)) )
```

will give all indeces for genes in `gnam.int`.
- If you want to use regular expression matching, use `grep`.

Export of annotation to HTML

- `annotate` is able to export tables of gene annotations to HTML, which is much nicer to browse than text tables
- Suppose, from a t-test you have for some genes `igenes`: mean of genes in class 1, `igenes.gp1`, mean in class 2, `igenes.gp2`, and P-value `igenes.pval`. To construct pretty HTML output:

```
igenes.ll = mget(igenes, env=hgu95av2LOCUSID)
igenes.sym = mget(igenes, env=hgu95av2SYMBOL)
ll.htmlpage(igenes.ll, "HOWTO.igenes", "Some genes",
list(igenes,sym, igenes, round(igenes.gp1,3),
round(igenes.gp2,3),round(igenes.pval,3)))
```

The result

Ensembl Human Genome ... NICEPLOT view of Swiss-PT... BioConductor Linkage List

BioConductor Linkage List

Some genes

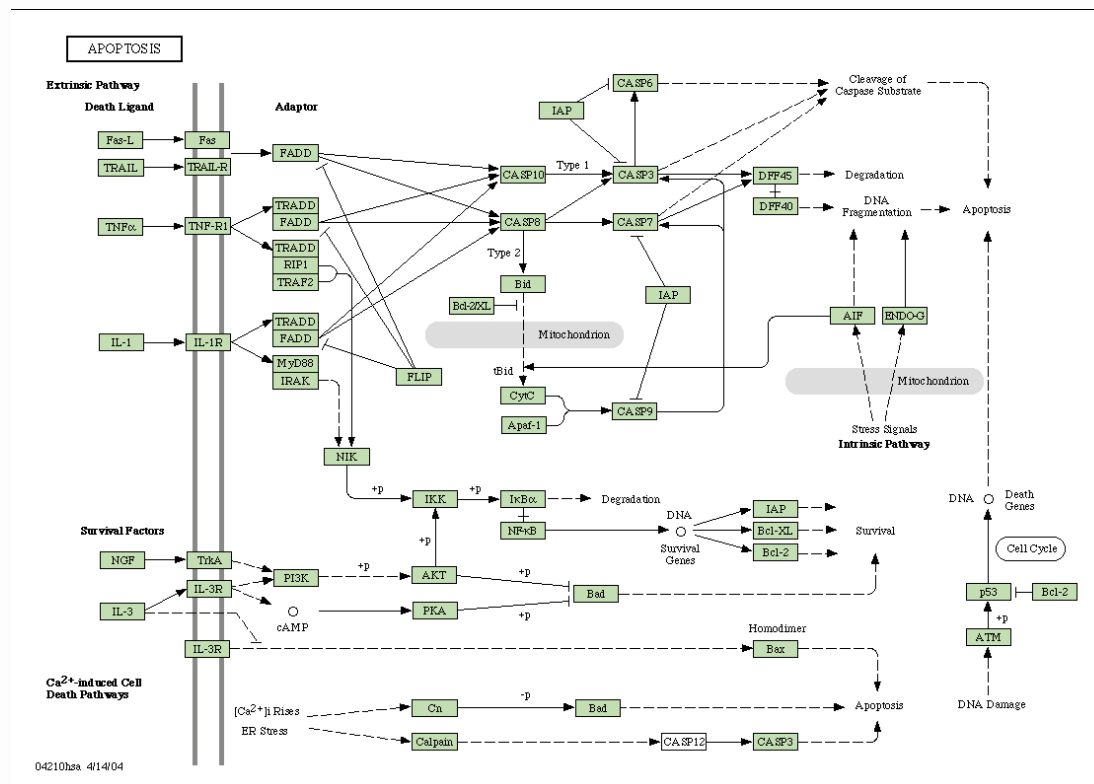
23378	KIAA0409	31484_at	145.869	153.948	0.635
221823	LOC221823	31485_at	150.41	153.703	0.892
4330	MN1	31486_s_at	13.057	16.238	0.447
9637	FEZ2	31487_at	82.982	27.448	0.311
27335	elF3k	31488_s_at	268.605	259.847	0.864
NA	NA	31489_at	0.886	0.479	0.873
6331	SCN5A	31490_at	200.904	194.797	0.767
841	CASP8	31491_s_at	22.029	23.582	0.606
27335	elF3k	31492_at	293.814	318.384	0.736
1442	CSH1	31493_s_at	29.719	32.583	0.82
NA	NA	31494_at	6.14	5.071	0.773
6846	XCL2	31495_at	118.936	113.031	0.714
6846	XCL2	31496_g_at	49.544	42.06	0.455
2543	GAGE1	31497_at	309.21	363.383	0.354
2578	GAGE6	31498_f_at	104.038	161.529	0.44
2215	FCGR3B	31499_s_at	163.479	132.496	0.448

Pathways

- For biological interpretation of function, most people want to use *pathways*
- A pathway is something like a bunch of interacting proteins and/or nucleic acids that allow for mass flux (metabolism) or information flux (signal transduction)
- The problem is that interaction information for proteins is quite rare (except for yeast)
- Some textbook pathways exist, but only few in computer-readable format

Pathway databases

- For metabolic pathways, some databases exist: KEGG (<http://www.genome.ad.jp/kegg/>), and EcoCyc (<http://ecocyc.org>), HumanCyc (<http://humancyc.org>) from BioCYC (<http://biocyc.org>)



Signal transduction information

- KEGG has some very limited information on signal transduction
- The database TRANSPATH wants to cover signal transduction. But information is incomplete, and you have to pay for part of the information (available via HNB)
- Other sources are www.biocarta.com and www.stke.org (requires registration)

Some software packages for function analysis

- There are some packages that allow to map gene expression profiles to biological information, like pathways.
- One example is GeneMAPP (www.genmapp.org) which also has a collection of user-contributed pathways.
- GoMiner (<http://discover.nci.nih.gov/gominer>) tries to find statistically significantly enriched terms in a gene list. This is, however, very crude and tends to favor annotations with very few total number of associated genes.
- Ingenuity (<http://www.ingenuity.com>) has its own database with interaction information, and software to infer pathways from microarray experiments. It seems to be quite capable, but is also expensive.

Dealing with GO annotations

- Since the annotation system is hierarchical, i.e. for each term there is a hierarchical list of more general terms, we can compare functions of genes on every level we wish.

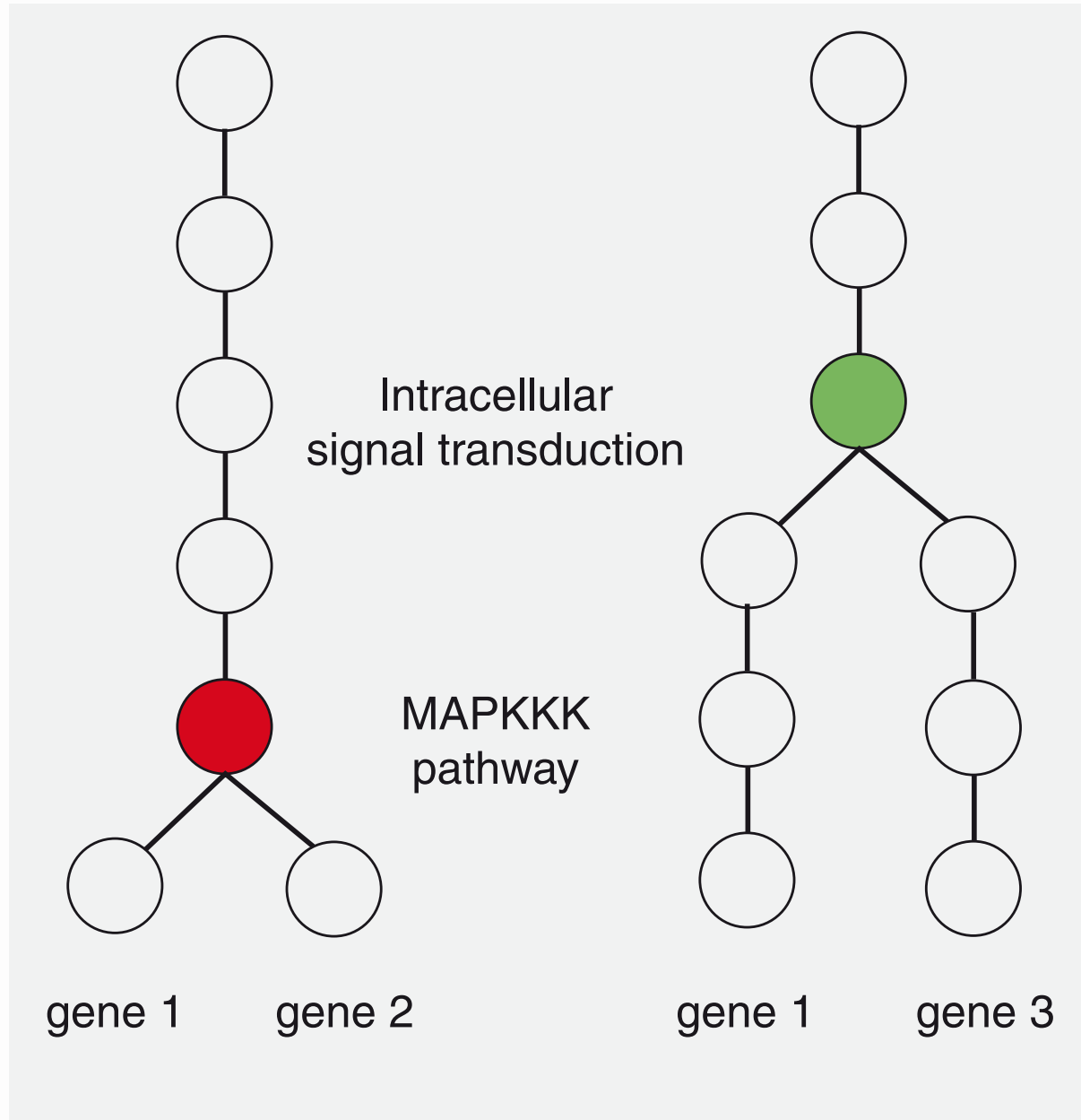
Dealing with GO annotations

- Since the annotation system is hierarchical, i.e. for each term there is a hierarchical list of more general terms, we can compare functions of genes on every level we wish.
- Technically, this amounts to the problem of finding the least common parent node between two genes of interest.

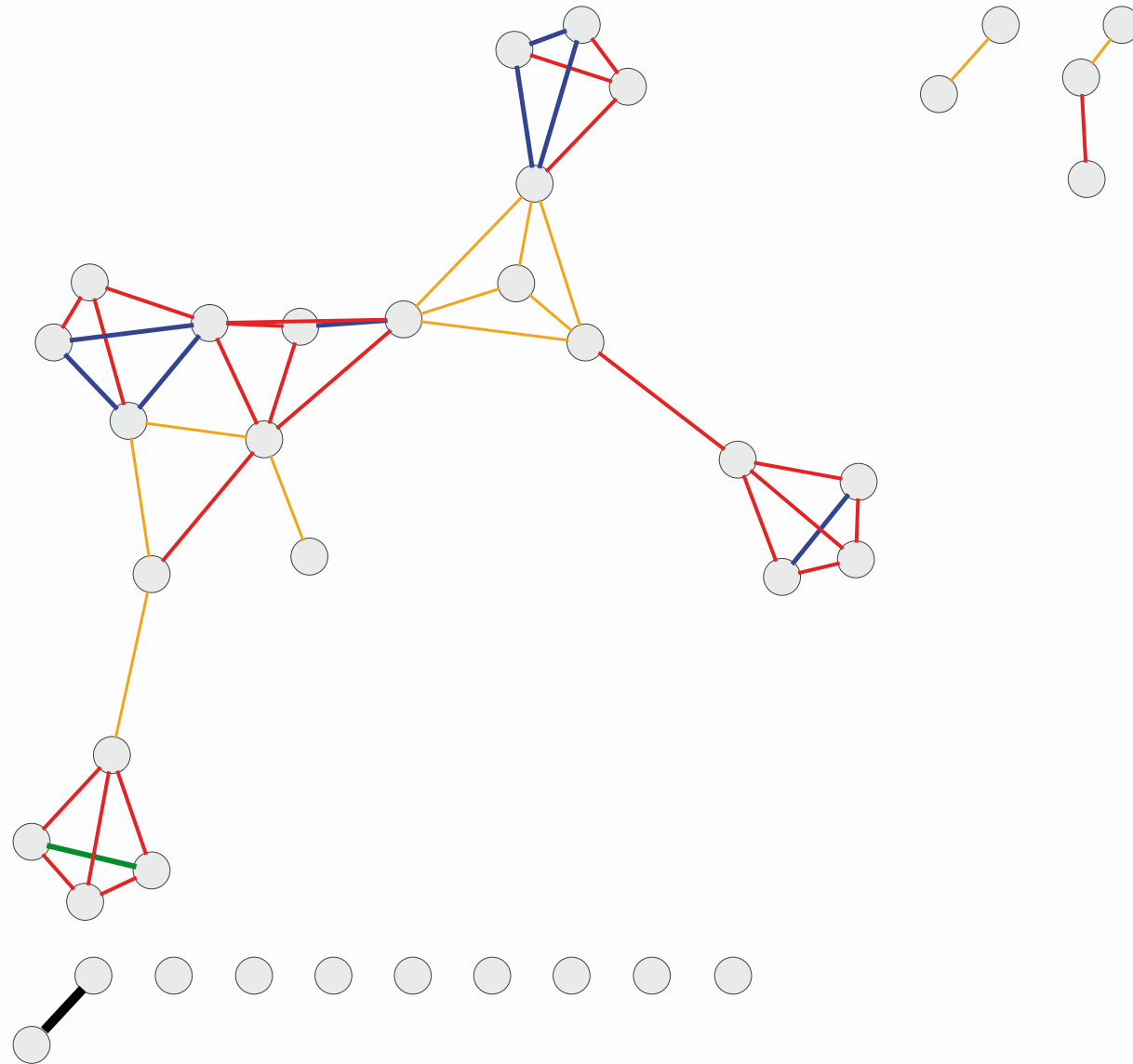
Dealing with GO annotations

- Since the annotation system is hierarchical, i.e. for each term there is a hierarchical list of more general terms, we can compare functions of genes on every level we wish.
- Technically, this amounts to the problem of finding the least common parent node between two genes of interest.
- This can be used to find clusters of functionally related genes in a list that comes out of some other analysis.

Comparing GO-annotated genes



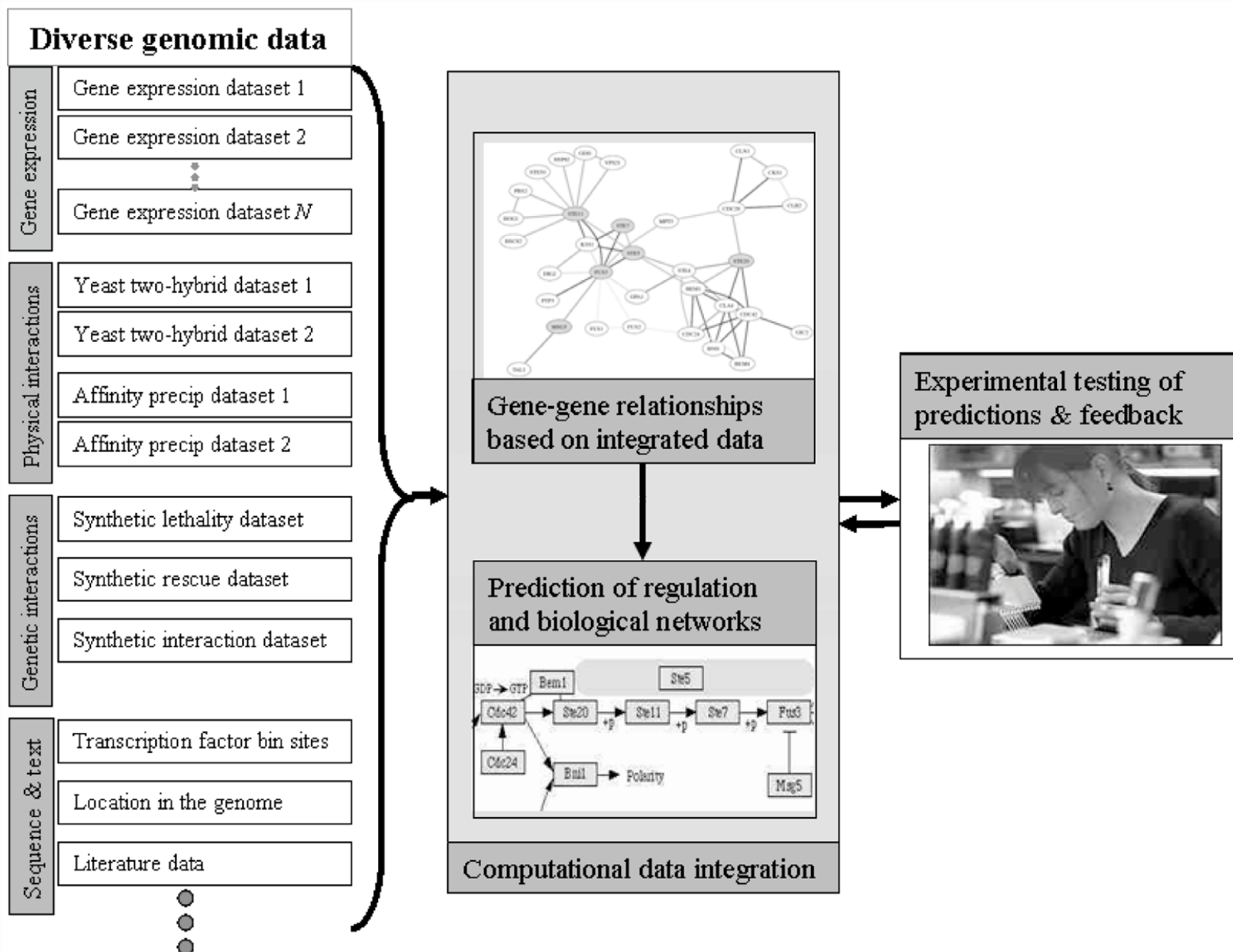
GO functional clusters as a graph



Graphs as analysis tools

- Graphs are quite useful for bioinformatic analysis, and have a long-standing history in sequence analysis.
- Recently, some functionality has been built into R to deal with graphs (`graph`, `Rgraphviz`, `RBGL`). Certainly, the most useful capability is to visualize graphs via `Rgraphviz`. The R package is an interface to the external program `graphviz` (from AT&T). Big graphs should be visualized by means of `ggobi`, however.
- Some other immediate use is to construct PubMed co-citation graphs for genes of interest. Functions for this exist. However, for many other applications the meaning of graphs or graph-theoretic algorithms is not clear, so a lot of work remains to be done.

Outlook: Integrated Analysis



Acknowledgements - Slides borrowed from

- Benedikt Brors
- Robert Gentleman

Thank you for your attention!