GCB 2003

Tutorial: Annotation
Introduction and Overview

• From Sequence to Gene
  – Gene Modelling

• From Protein to Function
  – Automated whole Genome Annotation (PEDANT)
  – Manual Annotation (Funcat/GO) -> Functional Classification

• Backbone Genome Projects
  – Fungi, Plant
  – Functional Modules (PPIs and Complexes)
  – Comparative Tasks

• “unsequenceable” Genomes (EST Sputnik)

• Research Architecture for Functional & Comparative Analysis

• Conclusions & Future Perspectives

GATTACAGATTACA

01 METABOLISM
01.01 amino acid metabolism
  01.01.01 amino acid biosynthesis
    01.01.01.01 assimilation of ammonia, biosynthesis of the glutamate group
    01.01.01.02 biosynthesis of the glutamate group (proline, hydroxyprolin, arginine, glutamine, glutamate)
    01.01.01.02.01 biosynthesis of proline
    01.01.01.02.02 biosynthesis of hydroxyproline
    01.01.01.02.03 biosynthesis of arginine
A Definition for Genome Annotation

Extraction, definition, and interpretation of features on the genome sequence, derived by integrating computational tools and biological knowledge.
Why do we annotate genomes?

- A genome is a large and complex object that carries a large amount of non-obvious information
  - Annotation identifies informative facets of genome
    - Looking for genes, classes of RNA, repeats, other genetic elements
- The collection of all genetic elements identified need to be ordered, classified and described
  - What use is just a collection of DNA or peptide sequences?
    - Describe each element in context with contemporary biology
    - Allows us to ask questions on “what is a genome and how does it work?”
- There is a genomics “Nirvana”
  - To understand the repertoire of genetic elements that are available within a genome
    - How can this genome do what it does? – biochemical pathways, disease resistance etc
  - To contrast the gene sets of different genomes and relate genes to biology
Genome annotation or gene annotation?

- A fundamental dichotomy underlying the genomics philosophy
  - Gene annotation
    - To describe a gene or protein in terms that can be understood by scientists
  - Genome annotation
    - To describe a gene or protein in terms that can be understood by scientists
    - To describe a gene or protein in context with all other genes and proteins from the same organism
    - To keep all of the information within a common ontology
      - Structuring of data
      - Machine readability
Genome Projects as Starting Point for Annotation

The flow of information

Traditional

Phenotype
(e.g. PEP+ADP -> Pyruvate + ATP)

↓

Isolation of the respective protein
(pyruvate kinase)

↓

Protein sequence analysis
(Edman degradation)

↓

Cloning and sequencing of
the corresponding gene

In the genomics area

DNA sequencing of a
genome

↓

Gene prediction

↓

Bioinformatic analysis of
genes/proteins and genetic elements

↓

Biochemical analysis of
proteins
Size of the TrEMBL Database
## Number of Genes versus Gene Density

<table>
<thead>
<tr>
<th>Organism</th>
<th>Genome Size</th>
<th>Number of Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Streptomyces coelicolor</strong></td>
<td>8.7 Mbp</td>
<td>7825</td>
</tr>
<tr>
<td><strong>Saccharomyces cerevisiae</strong></td>
<td>12.1 Mbp</td>
<td>6300</td>
</tr>
<tr>
<td><strong>Rice</strong></td>
<td>420 Mbp</td>
<td>50,000-60,000</td>
</tr>
<tr>
<td><strong>Homo sapiens</strong></td>
<td>3100 Mbp</td>
<td>~35,000</td>
</tr>
</tbody>
</table>
### Number of Genes versus Gene Density

<table>
<thead>
<tr>
<th></th>
<th>Genome size</th>
<th>Number of genes</th>
<th>Introns per gene</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yeast</td>
<td>12 Mbp</td>
<td>6300</td>
<td>0.04</td>
</tr>
<tr>
<td>Fruit fly</td>
<td>180 Mbp</td>
<td>13600</td>
<td>3</td>
</tr>
<tr>
<td><em>Homo sapiens</em></td>
<td>3100 Mbp</td>
<td>~35,000</td>
<td>9</td>
</tr>
</tbody>
</table>
Gene modelling in higher eukaryotes is a major challenge.

Correct genestructure:

Extended exon:

Missing exon:

Additional exon:

Missing intron:

Extended genemodel:
Gene Modelling in Eukaryotes

This model has to be split into two genes
From Protein to Function

- From Sequence to Gene
  - Gene Modelling

- From Protein to Function
  - Automated whole Genome Annotation (PEDANT)
  - Manual Annotation (Funcat/GO) -> Functional Classification

- Backbone Genome Projects
  - Fungi, Plant
  - Functional Modules (PPIs and Complexes)
  - Comparative Tasks

- “unsequenceable” Genomes (EST Sputnik)

- Research Architecture for Functional & Comparative Analysis

- Conclusions & Future Perspectives
## Number of Genes versus Complexity of Organisms

<table>
<thead>
<tr>
<th>Genome size</th>
<th>Number of genes</th>
<th>Introns per gene</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Yeast</strong></td>
<td>12 Mbp</td>
<td>6300</td>
</tr>
<tr>
<td><strong>Fruit fly</strong></td>
<td>180 Mbp</td>
<td>13600</td>
</tr>
<tr>
<td><strong>Homo sapiens</strong></td>
<td>3100 Mbp</td>
<td>(\sim35,000)</td>
</tr>
</tbody>
</table>
Context determines the biological function: but this context is not in the sequence!

- Posttranslational modification
- Regulation
  - Condition of expression
  - Chemical modification
  - Degradation/cleavage
  - Export (condition dependent)
- Sub/extracellular location
- Protein/protein interaction
- Protein/DNA interaction
- Time/condition dependent occurrence (control of expression)

→ proteins do need in-depth annotation
Automatic Annotation of Proteins

Predictive Methods

Experimental Resources

Annotated Protein
Database Resources

- Data repositories
  - EMBL
  - GenBank
- Manually curated protein databases
  - PIR
  - SwissProt
  - MIPS
- Protein Structure databases
  - SCOP
  - CATH
- Metabolic pathway databases
  - KEGG
  - WIT
Tools for Interpretation of Individual Sequences

- Sequence homology searches
  - BLAST
  - FASTA
- Detection of protein functional sites
  - PROSITE
  - Pfam
  - PRINTS
  - ProDom
  - SMART
  - TIGRFAMs
  - InterPro
  - COGS
  - Panther
PEDANT: Protein Extraction, Description, and Analysis Tool

- Automatic analysis of complete genomes
- Multiple functional and structural categories
- Comprehensive genome browser
- DNA and protein viewers
- Manual annotation possible
- All complete genomes available
- User sequence sets can be directly submitted

(Frishman et al., Bioinformatics, 17, 44-57, 2000)
The PEDANT Architecture

Processing UNIT
- Bioinformatics interface
  - Orpheus wrapper
  - Blast wrapper
  - HMMER wrapper
  - BLIMPS wrapper
  - Predator wrapper

Parser
- Orpheus parser
- Blast parser
- HMMER parser
- BLIMPS parser
- Predator parser

Relational DBMS
- Primary data tables
  - Genomes, contigs, ESTs
  - Exons, genes, ORFs, tRNAs, proteins
  - Results (blast, pfam, prosite, etc.)

- Secondary data tables
  - Properties of input DNA sequences
  - Properties of derived genetic elements
  - Parsed output of bioinformatics methods

User interface
- CGI scripts, JavaScript
- Reports
- Functional categories
- Structural categories
- DNA views
- Protein views
- Statistics
- Search
- Help
- Manual annotation
PEDANT genome database

- Largest publicly available database on annotated genomes
- Completely sequenced genomes
  - 96 eubacteria, 16 archebacteria, 8 eukaryotes
- Genomes in progress
  - 58 eubacteria, 3 archebacteria, 16 eukaryotes
- 4-5 releases per year
- 350 Gb of data
- Integrated with BioRS™ retrieval system (Biomax Informatics AG) for fast searches and data mining
Pitfalls in Annotation – Information Transfer

- The information transfer from well-characterized proteins to predicted gene products has to be done carefully since:
  - Similar sequence does not always imply similar structure or function.
  - The annotation of the database protein might be incomplete or even wrong.
Manual Annotation of Proteins

Predictive Methods

Experimental Resources

Human Evaluation of Data

Annotated Protein

GCB '03 Tutorial Annotation
Requirements for a Functional Classification Scheme

- Human usability
- Computer readability
- Breadth and depth of coverage
- Stability and extendibility
- Independence of organism
Functions of the Gene Products of *E. coli*

- Functional classification scheme used for annotation of the first published genomes (Riley, 1993).
- Developed for *E. coli* gene products.
- One functional class per gene/gene product.
- Updated in GenProtEC and EcoCyc.
- Adapted e.g. in SubtiList.
The MIPS Functional Catalogue

- Organized hierarchically with up to six levels.
- ~1500 categories (developed in collaboration with BIOMAX).
- Genes can belong to multiple categories.
- Categories are independent of organism.
- Currently 9 organisms incorporated: yeast, human, *A. thaliana*, ...

<table>
<thead>
<tr>
<th>01</th>
<th>METABOLISM</th>
</tr>
</thead>
<tbody>
<tr>
<td>01.01</td>
<td>amino acid metabolism</td>
</tr>
<tr>
<td>01.01.01</td>
<td>amino acid biosynthesis</td>
</tr>
<tr>
<td>01.01.01.01</td>
<td>assimilation of ammonia, biosynthesis of the glutamate group</td>
</tr>
<tr>
<td>01.01.01.01.01</td>
<td>assimilation of ammonia</td>
</tr>
<tr>
<td>01.01.01.01.02</td>
<td>biosynthesis of the glutamate group (proline, hydroxyprolin, arginine, glutamine, glutamate)</td>
</tr>
<tr>
<td>01.01.01.01.02.01</td>
<td>biosynthesis of proline</td>
</tr>
<tr>
<td>01.01.01.01.02.02</td>
<td>biosynthesis of hydroxyproline</td>
</tr>
<tr>
<td>01.01.01.01.02.03</td>
<td>biosynthesis of arginine</td>
</tr>
</tbody>
</table>
Proteome Comparison

Proteome comparison based on the high-level functional categories

- Metabolism
- Transcription
- Transport facilitation
- Cellular communication/signal transduction
- Cell growth, cell division, and DNA synthesis
- Energy
- Protein synthesis
- Protein destination
- Intracellular transport
- Cellular biogenesis
- Cell rescue, defence, death, and ageing
- Ionic homeostasis
Gene Ontology

- Classification scheme, particularly used for annotation of eukaryotes like yeast (SGD), fruit fly (FlyBase) and mouse (MGD).
- Goal: A dynamic, controlled vocabulary that can be applied to all eukaryotes even as knowledge of gene and protein roles in cells is accumulating and changing (The Gene Ontology Consortium, 2000).
- The GO Consortium is developing three ontologies: biological process, molecular function and cellular component.
- GO terms are organized in acyclic graphs, i.e. a ‘child’ can have many ‘parents’.
- GO consists of ~16,000 terms in total.
GO Annotation of the Fruit Fly Genome

Mif et al. (2003). Genome Research, 13, 2118-2128.

Figure 2: Coverage of Drosophila proteins classified by FlyBase and PANTHER. (A–C) Classification coverage for molecular function categories. (A) FlyBase associated 6301 proteins (red area) with at least one GO molecular function. (B) PANTHER associated 4682 proteins (blue area) with a GO molecular function. The light grey area indicates proteins that had a PANTHER HMM, but were not associated with a GO term (see text for details), whereas the dark grey area indicates proteins that did not hit any PANTHER HMMs. (C) Venn diagram illustrating the overlap between proteins classified by FlyBase and PANTHER. (D–F) Classification coverage for biological process categories. (D) number of proteins classified by FlyBase, (E) number classified by PANTHER, and (F) the overlap between sets of proteins classified by the two methods.
Comparison of Functional Annotation Schemes -FuncWheels-

Rison et al., 2000
Comparison of functional Annotation Schemes

Fig. 3A–F. Coverage of the ‘Combination Scheme’ (CS) illustrated using FuncWheels. A EcoCyc; B TIGR; C Subtilist; D MIPS; E KEGG; F WIT. In these FuncWheels, nodes in the CS not represented by the illustrated scheme are blanked out. CS functions present and absent can be identified by reference to Fig. 2.

node. Two of the nodes under the TIGR ‘Cell envelope’ node could have been mapped to CS 4.1.1: ‘biosynthesis of surface polysaccharides and lipopolysaccharides’ and ‘lipoproteins’. The former was more accurately mapped to ‘Surface polysaccharides/antigens’.
Current Development in Functional Annotation

Annotation of Functional Modules

Functional Annotation of Genomes

Annotation of Genes/Proteins
Functional Modules

Modules are units that are expected to share functions related to their common cellular activity. Members of modules interact directly or indirectly.
Backbone Genome Projects and Data Integration

- **From Sequence to Gene**
  - Gene Modelling

- **From Protein to Function**
  - Automated whole Genome Annotation (PEDANT)
  - Manual Annotation (Funcat/GO) $\rightarrow$ Functional Classification

- **Backbone Genome Projects**
  - Fungi, Plant
  - Functional Modules (PPIs and Complexes)
  - Comparative Tasks

- “unsequenceable” Genomes (EST Sputnik)
- Research Architecture for Functional & Comparative Analysis
- Conclusions & Future Perspectives
Yeast, as well as *Neurospora*, *Arabidopsis* and Rice, are species specific genome databases. Entries are connected to updated DNA-contig sequences.

- Entries are CDS and any other genetic elements.
- Additionally to these MIPS manually annotated genome databases, genomes of related species are subjected to PEDANT analysis to enable comparative genomics.
## Synopsis of the Attribute Catalogues

<table>
<thead>
<tr>
<th>CATALOGUE</th>
<th>Distinct CATEGORIES</th>
<th>ENTRIES</th>
</tr>
</thead>
<tbody>
<tr>
<td>FUNCAT</td>
<td>261</td>
<td>16164</td>
</tr>
<tr>
<td>LOCALIZATION</td>
<td>49</td>
<td>7200</td>
</tr>
<tr>
<td>COMPLEX</td>
<td>1049 (783 mass spec.)</td>
<td>8494</td>
</tr>
<tr>
<td>PHENOTYPE</td>
<td>142</td>
<td>3041</td>
</tr>
<tr>
<td>ENZYME</td>
<td>531</td>
<td>1169</td>
</tr>
<tr>
<td>PROTEIN CLASS</td>
<td>149</td>
<td>1058</td>
</tr>
<tr>
<td>TRANSPORT/MEMBRANE</td>
<td>235</td>
<td>844 (De Hertogh et al. 2002)</td>
</tr>
</tbody>
</table>
Features at their Genomic Loci

Saccharomyces cerevisiae - Chromosome II [316613 bp] (from 1 to 35000)

Clickable map:

Upper colour: ORF class
Lower colour: type

Types colours
- Protein
- RNA (tRNA, snRNA, ...)
- Centromere
- LTR
- Ty element
- Repeat unit
- Miscellaneous features

ORF classification colours
- Known proteins
- Similarity to known proteins
- Similarity to unknown proteins
- No similarity to other proteins
- Questionable ORF

From 1 to 35000 base pairs per pixel 50 redraw Show comparison
Entry Annotation

CYGD Annotation 'FUNCTIONAL_CATEGORIES' for entry YER146c

Selected
FUNCTIONAL_CATEGORY:

67.04.01.01

FUNCTIONAL_CATEGORY
(Single Selection!):

63.21.99 other complex cofactor binding
63.99 other binding function
65. STORAGE PROTEIN
65.01 storage facilitating proteins
65.02 stored proteins
67. TRANSPORT FACILITATION
67.01 channel / pore class transporters
67.01.01 ion channels
67.01.03 gap-junction forming protein family (e.g. connexin, innexin)
67.01.05 nuclear pore forming protein
67.01.07 general bacterial poren family (e.g. omp1, sugar poren, ompf)
67.01.09 pore-forming toxins (e.g. diphtheria toxin, hemolysin)
67.01.09 other channels and pores forming proteins
67.04 ion transporters
67.04.01 cation transporters
67.04.01 heavy metal ion transporters (Cu, Fe, etc.)

PMID:

Selected Evi:

901 information accessibility
901.01 reviewable
901.01.01 journal literature (TAS)
901.01.01.01 PubMed uid
901.01.01.01 non PubMed uid
901.01.02 database (TAS) or (IEA)
901.01.02.01 open access (Web)
901.01.02.02 restricted access

EVI (Multiple Selection!):

update  clear your modifications
Functional Modules: Protein-Protein Interactions and Complexes

- Yeast
- Mammalian
Why care for PPI?

• PPIs happen for a reason:
  – Assembling structural entities
  – Signalling pathways
  – Protein trafficking
  – Building functional enzymes from subunits
  – Facilitate protein folding
  – Protein modifications

• PPIs link proteins of common function

• Potential drug targets

• PPI networks help understanding of cellular organisation
Yeast Protein-Protein Interactions

- About 10,600 binary interactions are annotated:
  - Physical (~9100; Co-immunoprecipitation, 2-Hybrid – physical interaction of 2 hybrid proteins in nucleus)
  - Genetic (~1500; synthetically lethal, suppression phenotypes, e.g. overexpression of protein X suppresses phenotype of mutant Y)
- More than 1000 annotated complexes can be split in multiple interactions describing higher order complexity.
Mammalian PPI
- Special Considerations

- Mammalian Genomes are **huge compared to yeast** ⇒ *we can't just go through all ORFs*
- Interactions among mammalian orthologs usually work across the species boundary. This is reflected by a large number of cross-species experiments in the literature (human → mouse, rat → dog, ...) ⇒ *We are annotating a group of organisms!*
- We don't have complete genomes for many mammals ⇒ *we can't easily assign ORF codes to many proteins*
PPI – Functional Projection
Fungal Kingdom

Ascomycota

- Aspergillus nidulans
- Aspergillus niger
- Fusarium graminearum
- Neurospora crassa
- Saccharomyces cerevisiae
- Candida albicans
- Schizosaccharomyces pombe
- Ustilago maydis

Basidiomycota

- Agaricus bisporus (White button pizza mushroom)

Hedges SB, 2002
Upcoming Genomes

- About 50 fungal species with medical, commercial, evolutionary or agricultural interest, are proposed to be sequenced in the next 3 years.
Comparison of species increases prediction probability

- Sequencing projects of related species allows comparative genomics
  - Validation of gene models
    - Spurious or rapid evolving genes
  - Identification of regulatory elements

![Genetic Comparison Diagram](image)
Genomes annotated at MIPS/Biomax

- Bacteria
  - *Bacillus subtilis* 168 (Biomax)
  - *Helicobacter pylori* KE 26695
  - *Listeria innocua* Clip 11262
  - *Listeria monocytogenes* EGD

- Archaea
  - *Thermoplasma acidophilum* (Ruepp et al., 2000)

- Lower Eukaryotes
  - *Saccharomyces cerevisiae* (Bussey et al., 1997)
  - *Neurospora crassa* (Galagan et al., 2003)

- Plants
  - *Arabidopsis thaliana* (AGI, 2000)

- Mammals
  - *Homo sapiens* (Biomax)
“unsequenceable” Genomes (EST Sputnik)

- From Sequence to Gene
  - Gene Modelling
- From Protein to Function
  - Automated whole Genome Annotation (PEDANT)
  - Manual Annotation (Funcat/GO) -> Functional Classification
- Backbone Genome Projects
  - Fungi, Plant
  - Functional Modules (PPIs and Complexes)
  - Comparative Tasks

- “unsequenceable” Genomes (EST Sputnik)
- Research Architecture for Functional & Comparative Analysis
- Conclusions & Future Perspectives
Genome size and apparent organism complexity poorly correlated.
Psilotum nudum

- Genome size of ~250,000 Mbp
- Largest characterised genome (I think) anywhere
- **Family:** Psilotaceae *Kanitz.*
- Curious club-moss; botanically interesting as a very primitive vascular plant; herbaceous perennial devoid of roots but with only scale leaves on ribbed stems. Up to 50cm tall. Tropical
Other Model Plants (& Products)
Some Genomes are “Unsequenceable”

- Genomes have highly variable sizes
  - Large genomes are expensive to sequence
  - Large genomes are highly repetitive – low information potential
  - Bioinformatics based genome assembly is questionable

- Not all genomes are currently worthy of complete genome sequencing
  - Should be of immediate agricultural, medical, economic or scientific interest – this is a thin slice of all genomes

- We are still looking for an approximation for these genomes?
  - Genome sampling methods are available
    - Expressed sequence tag (EST) sequencing
    - BAC end sequencing
ESTs form a rather large sequence resource

- Some reasonable underlying biodiversity
- Major eukaryotic clades are at least superficially represented
- There has to be some bioinformatics potential with the resource
  - Proteomics
  - Comparative genomics
  - Marker development
Problem one: meaningful sequence clustering

- Not all “unigenes” or “indices” are created equal!
  - What is the goal of the clustering?
    - Assembly of ESTs into biologically meaningful contigs
    - Optimal assemblies at cost of redundancy
- HPT (hashed position tree) clustering
  - Suffix tree based, fast
- CAP3 assemblies
  - Optimised for separation of paralogues
EST clustering

- The clustering rationale?
- Reduction in complexity of the EST collection – towards non-redundancy, in larger multi-member clusters, the consensus sequence is clean of sequencing errors, and is a more suitable frame to anchor analyses
Extraction of coding sequence from ESTs

- ESTs contain frameshifts through sequence errors.
- Most ESTs contain 3’ UTR sequence – this is non-coding but may contain clear ORFs.
- Coding sequence has clear codon usage bias, and in context with weighting of frameshift start and extensions can parsimoniously determine a single CDS in a low quality background.

No clear ORF fs fs fs No clear ORF
SNP mining large EST collections

\[ s(i, X) := f_1 \left( \sum_{k=1}^{N} P \left( e_i^{(k)} = X \right) \right) + \sum_{j=1}^{L} f_2 \left( \sum_{k=1}^{N} P \left( e_i^{(k)} = X \land e_j^{(k)} \neq c_j \right) \right), \quad X \in A \]

Figure 1: An EST alignment of \( N \) ESTs \( (e^{(1)}, \ldots, e^{(N)}) \) with consensus sequence \( c \). The consensus sequence has length \( L \). At the \( i \)th position we have the letters \( c_i \) in the consensus sequence and \( e_i^{(k)} \) at the \( k \)th EST, respectively.
Sputnik comparative genomics platform
Database Architecture for Functional & Comparative Analysis

- From Sequence to Gene
  - Gene Modelling
- From Protein to Function
  - Automated whole Genome Annotation (PEDANT)
  - Manual Annotation (Funcat/GO) -> Functional Classification
- Backbone Genome Projects
  - Fungi, Plant
  - Functional Modules (PPIs and Complexes)
  - Comparative Tasks
- “unsequenceable” Genomes (EST Sputnik)

- Research Architecture for Functional & Comparative Analysis
- Conclusions & Future Perspectives
Biological Information

Functional Modules

Biochemical Entities

Genes

PPI
Regulatory Cascades
Other Views ...

Proteins
mRNA
Regulatory Elements
...
Complex Example: PPI Networks
Data Management of Complex Biological Knowledge

- System with data management capabilities similar to databases required
- (Normalized) datamodels
- Entities / objects and relationships
- Functional indexing / ordering
- Retrieval mechanisms
Orthogonalization of Information
Example: PPI Data Model
• Axis ⇔ Entities
How to manage this?

- **Data Management:**
  - Relational Databases?
  - OO Data Models?
  - XML Databases?

- **Knowledge Management:**
  - Context specific information of biological entities
  - Semantic information management required
Heterogeneous Datasources

Manual Annotation

Genome DBs

Web Services

Analysis Tools

Function

Retrieval Systems

...
Mapping to Legacy Information

- access relational DBs
- use GenRE's persistence extension

Vocabulary:
- name
- function
- sequence
- notes
BioRS Integration of Heterogenous Systems
{ Middleware }

- TP-Monitors (Tuxedo, ...)
  - Procedure oriented
  - „OS“ controls transactions, resources, ...
- ORBs (CORBA, Java-RMI, ...)
  - Distributed objects
  - Communication backbones
- CTMs (Component Transaction Monitors)
  - Hybrids
  - Concurrency, transactions, distributions of objects and load, security, resources, ...
  - CTM (business objects) ⇔ DBMS (data)
J2EE

- CTM: J2EE application server
- Enterprise Java Beans (EJBs) (server side components)
- Open standard
- Defines THAT services are implemented but not HOW
- Open Source AppServer (JBoss)
- Open Source solutions around it
Application Logic / Information "Services"

- Business delegate / session facade design pattern

- ContigsList
- GeneticElementsList
- GeneticElement
- Protein
- ....
Integration Layer
Multi-Tier Architecture
Semantic Data / "Mindmaps"

Existing data can often be interpreted in different ways, e.g.:

- A transcription factor binds to a sequence
- On the other hand, a sequence may somehow regulate the transcription factor
Problem Relationships

- Relational Data Models

- OO UML Models
Semantic Representation with gMBL

- **Rules:**
  - always consist of two nouns (can be the same)
  - associations are represented by verbs
  - rules can be read in two directions

- **Attributes:**
  - represent the data

```
Protein                    expressed from
                        expressed to
Gene
- name
- weight
- sequence
```

GCB '03 Tutorial Annotation 77
Data Management
gMBL Mapping

- Conversion from and to gMBL

Diagram:
- gMBL
  - XML Database
  - Object Binding
  - Data Binding
  - Database Schemes
  - Markup Languages
  - XSL / XSLT
Indexing

Catalogs

01 METABOLISM
  01.01 amino acid metabolism
  01.02 nitrogen and sulfur metabolism
  01.02.01 nitrogen and sulfur utilization
  ...
  02 ENERGY
  02.01 glycolysis and gluconeogenesis
  02.01.01 glycolysis methylglyoxal bypass
  ...
  03 CELL CYCLE AND DNA PROCESSING
  03.01 DNA processing
  03.01.01 cellular DNA uptake
  03.01.01.01 bacterial competence
  03.01.01.03 conjugational DNA transfer
  ...

Any other structuring mechanism:

e.g.: clustering
What can we do now?!

- Normalized datamodels
- Technology independent mapping of datasources against entities
- Different indexing mechanism to order information
Comparative Analysis

Localisation

Disease

Function
Interactions / Localisations
Interactions / Functions
Interactions / Function / Localisations

780 lipid particles
775 microsomes
770 vacuole
765 endosome
760 peroxisome
755 mitochondria
750 nucleus
745 transport vesicles
740 golgi
735 ER
730 cytoskeleton
725 cytoplasm
720 plasma membrane
715 cell periphery
710 cell wall
705 bud
701 extracellular

ENERGY
TRANSCRIPTION
METABOLISM
PROTEIN SYNTHESIS
CELL CYCLE AND DNA PROCESSING
CELL TRANSPORT
CELLULAR COMMUNICATION
CELL RESCUE, DEFENSE AND VIRULENCE
REGULATION WITH CELLULAR ENVIRONMENT
CELL FATE
GenRE: A Genome Research Environment
What have we seen so far?

- **From Sequence to Gene**
  - Gene Modelling

- **From Protein to Function**
  - Automated whole Genome Annotation (PEDANT)
  - Manual Annotation (Funcat/GO) -> Functional Classification

- **“unsequenceable” Genomes (EST Sputnik)**

- **Backbone Genome Projects**
  - Fungi, Plant
  - Functional Modules (PPIs and Complexes)
  - Comparative Tasks

- **Research Architecture for Functional & Comparative Analysis**
Conclusions

- We are now in a position to annotate genomes
  - Model genes and their structures
  - Describe genes in terms of structure, function and relation to other genomes
  - Assemble all the annotative attributes in a broad, extensible and generic computational infrastructure
  - Display all aspects of the genome back to a user through versatile display methods

- BUT
  - Analyses are centered towards single genomes
    - Cross genome interactivity – but still primitive
  - We are entering a post-genomics era
  - The reliability of some annotative foundations remains “questionable”
Future perspective

- Many more genomes
  - Genome sequencing has become routine
  - Taxonomic sampling of the related, the exotic and the important
    - Yet more bits of other genomes

- Functional and structural genomics
  - The reference annotation for known genes and proteins is changing
    - Experimentally validated gene structures
    - Precise spatial, temporal, functional and biochemical dissections of proteins
    - Identification of interaction partners, metabolites and regulators
    - Microarray based measurements of gene expression and localisation profiles
  - We need to adapt to and stay concurrent with systems biology

- Comparative analysis of core genomes and nodal genomes is of primary relevance
  - Which genes in species ‘A’ may be responsible for a phenotype not observed in closely related species ‘B’
  - Understanding of the evolution of genes and genomes

- Underlying questions that we can answer (biologists will ask) change
A final outlook?

- Contemporary genome annotation is a fair approximation
  - We apply latest methods
  - We exploit the newest experimental evidence
  - We place annotation into a common informatics frame
  - We are in a sound position to understand and annotate genomes

- We are not yet there!
  - For most genomes (eukaryotics) certainly >50% of genes are of unknown, undescribed or unmeasurable function and structure
  - This is a challenge
    - Stay abreast of the changes to the biological datasets
    - To keep our prediction methods tuned and refined
    - To keep our databases current and informative
    - To keep on top of newly sequenced genomes
  - Our jobs are safe