COMPGI10 – Intelligent Systems in Bioinformatics

Lecture 1: Introduction to high-throughput and systems biology

Dr Kevin Bryson

K.Bryson@cs.ucl.ac.uk

This part of the course …

- Up to now you have been looking at ‘individual biological parts’ (DNA, protein, etc.)

- The application of machine learning is valuable in these cases for making predictions about these parts, for instance predicting the secondary structure of a protein from its amino acid sequence.

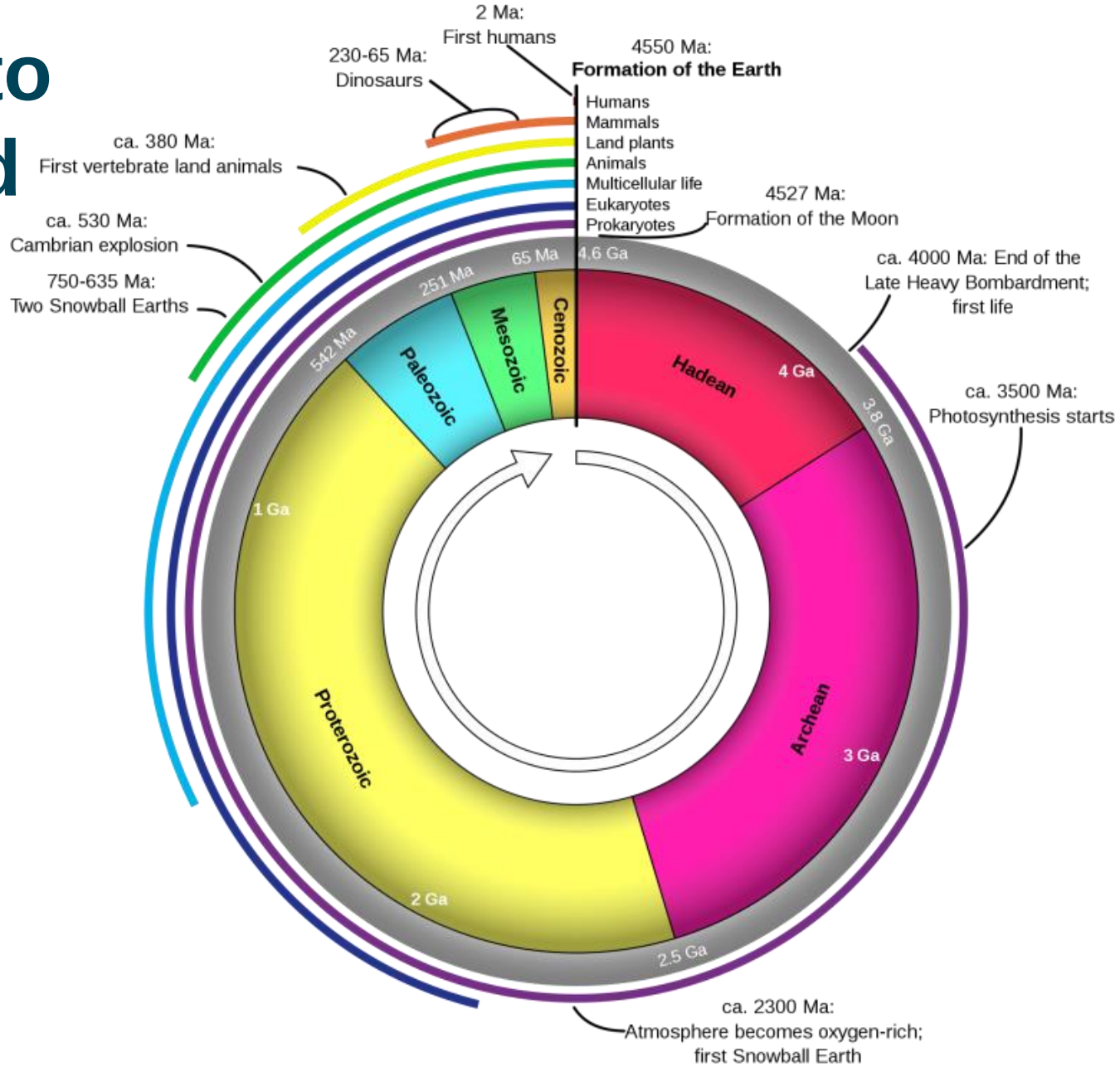
- But understanding biological systems requires more than just a reductionist understanding of the individual parts that compose them.
- In this part of the course we will look at complete biological systems, techniques that apply to these and how machine learning is useful.

Learning Outcomes

After this lecture, I would hope you:

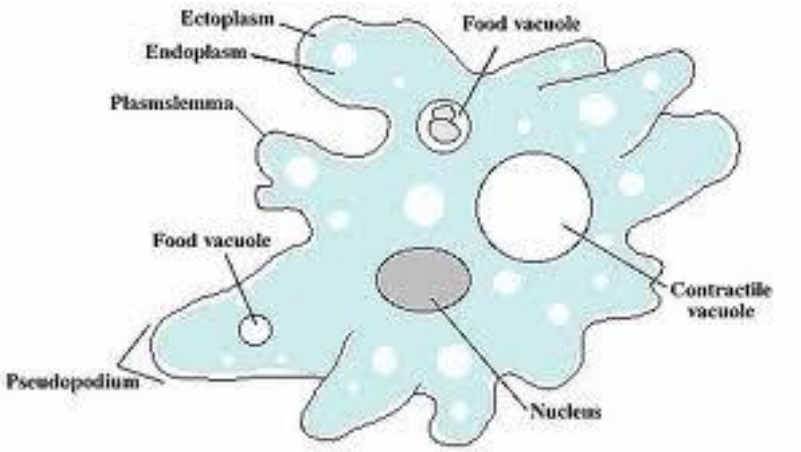
- Understand what we mean by a ‘biological system’ and comprehend the complexity of such a system.
- Appreciate that one way to understand such systems is in terms of their parts and the interactions between these parts.
- Understand different types of high-throughput (HT) experimental techniques to globally characterize the parts.
- Understand different types of biological networks to characterize the interactions between the parts.
- *All this provides a natural problem domain for machine learning.*

Our aim – to understand something that took 4 billion years to evolve!



A simple example from 1.2 billion years ago … the humble slime mould

Slime moulds are single cell organisms without any nervous system ...



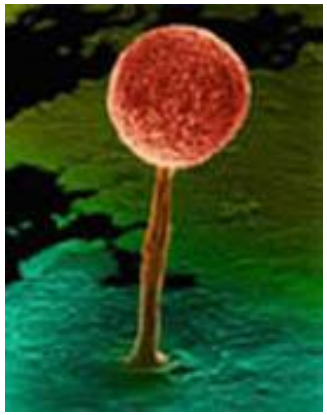
Slime moulds are very simple biological systems - just a bag of chemicals surrounded by a membrane:

... but they can solve mazes!

And they can find the optimal paths between multiple sources of food (like a distribution networks ...)

It's not rocket science ... but you can see why they have evolved to have this relatively complex behaviour.

Slime moulds are very simple biological systems - just a bag of chemicals surrounded by a membrane:

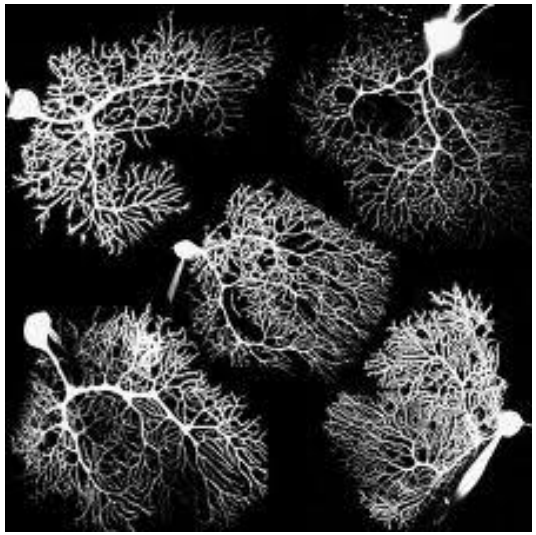


They have also ‘learnt’ how to collaborate with each other when required and show ‘self-sacrifice’ for the common good ...

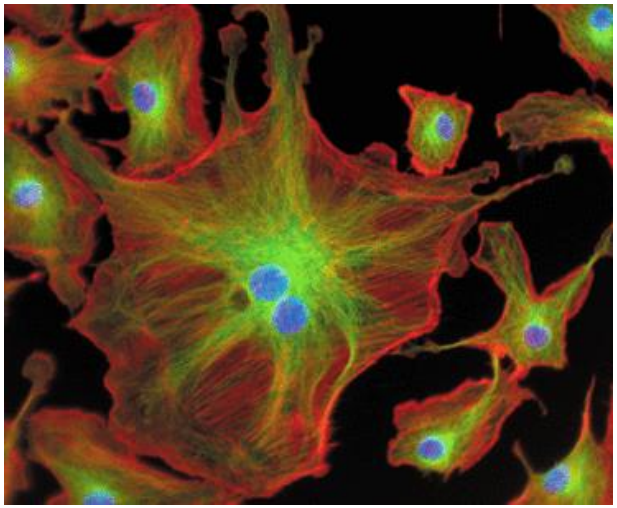
Aim of Systems Biology and HT technology

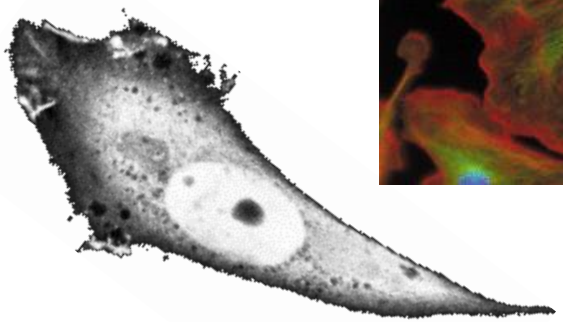
- These relatively complex behaviours are clearly at the ‘systems-level’.
  - They cannot be understood by studying a single gene or single protein structure.
  - They arise from complex interactions between different biological parts (molecules, cells, populations)

Before continuing ...
A question ...



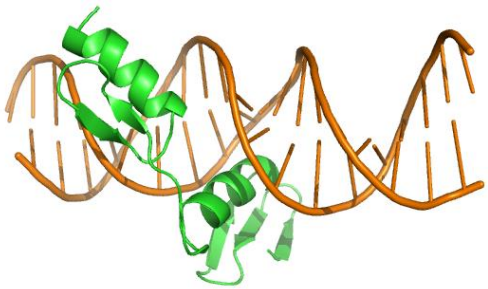
- All the cells making up our bodies all share an identical ‘blueprint’ of instructions (the genome sequence).
- But all our cells look and behave very differently ... how does this happen if they all have the same instructions for proteins?

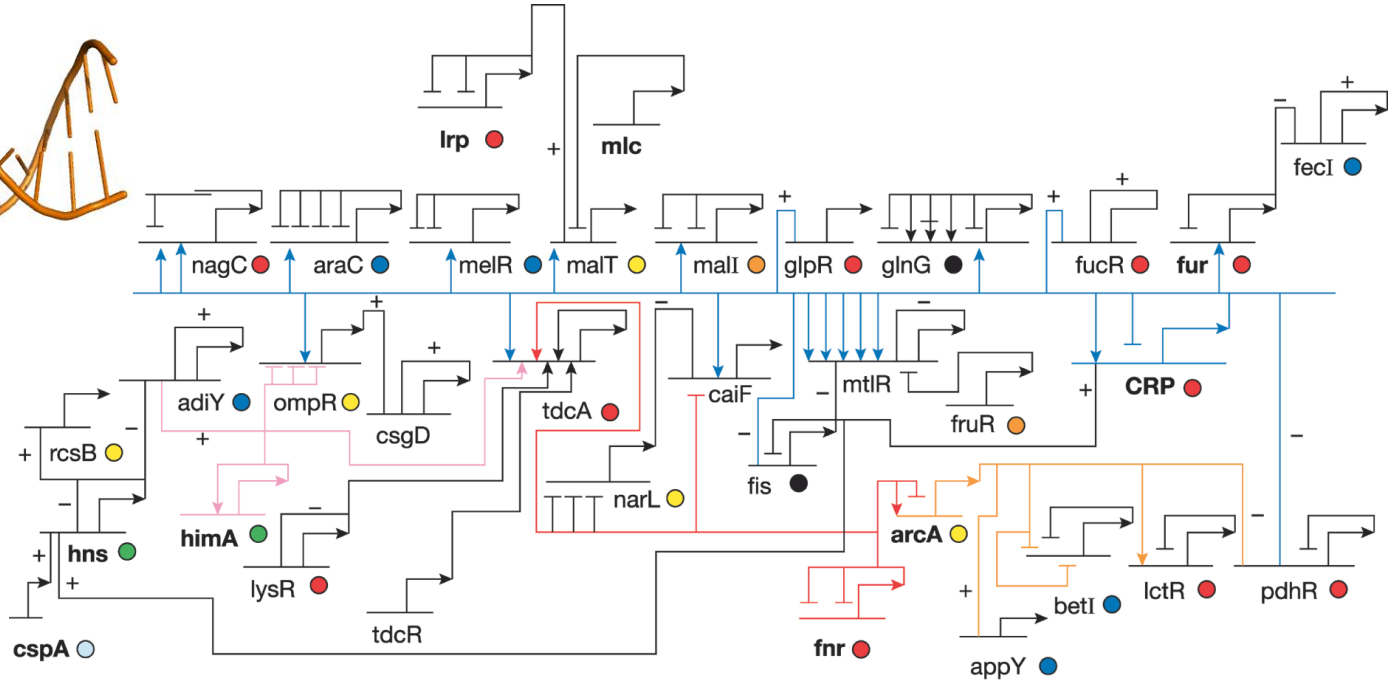




- Discuss in pairs for a couple of minutes …

During development each cell receives different ‘signals’ from chemicals released from other cells ... and an intricate ‘genetic program’ where proteins bind to DNA and turn on/off particular genes finally gives all the different cell types.





Biological Network Data – *biological processes* are accomplished by proteins working together in *biological networks* – each protein carrying out a specific *molecular function*

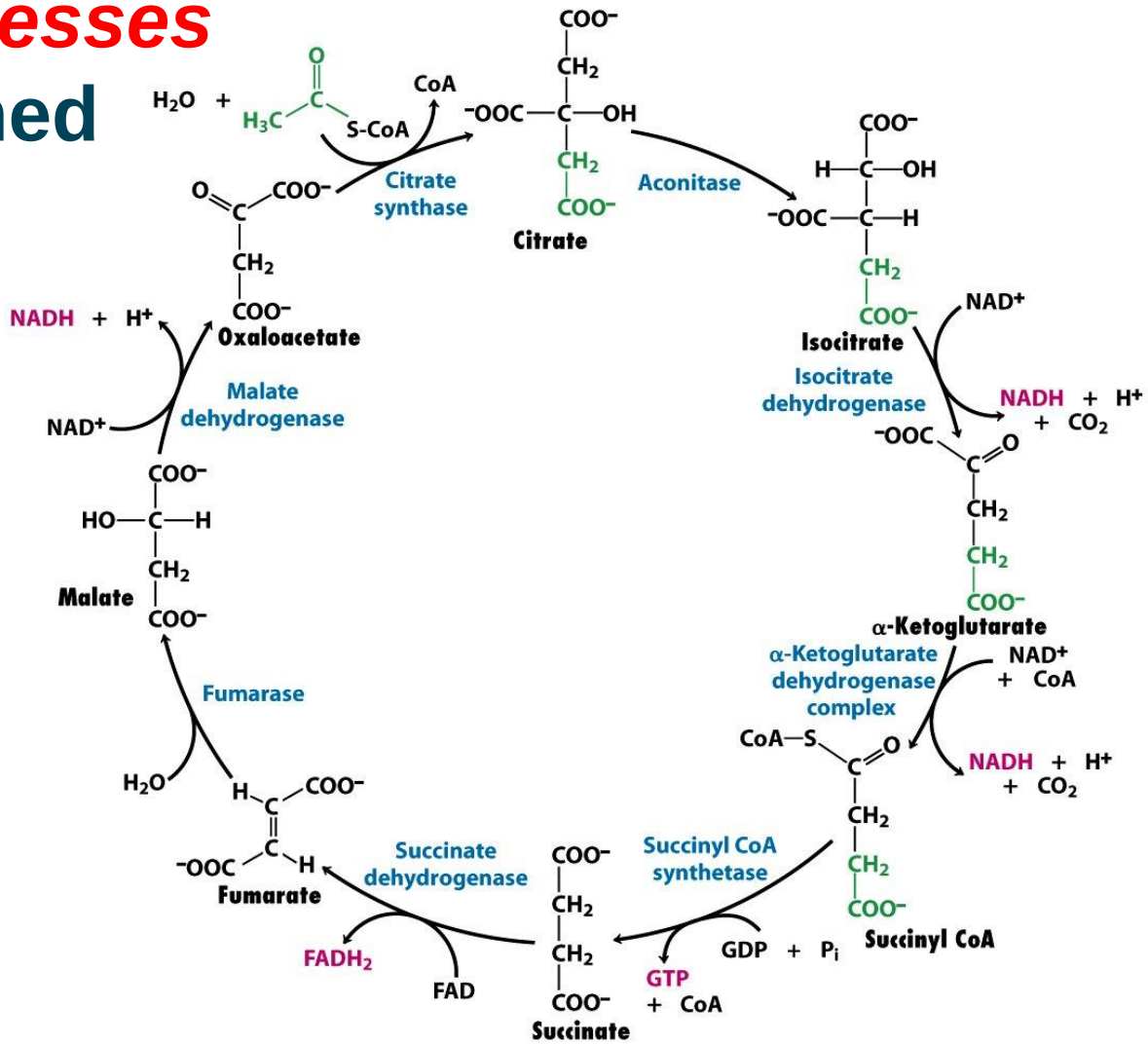
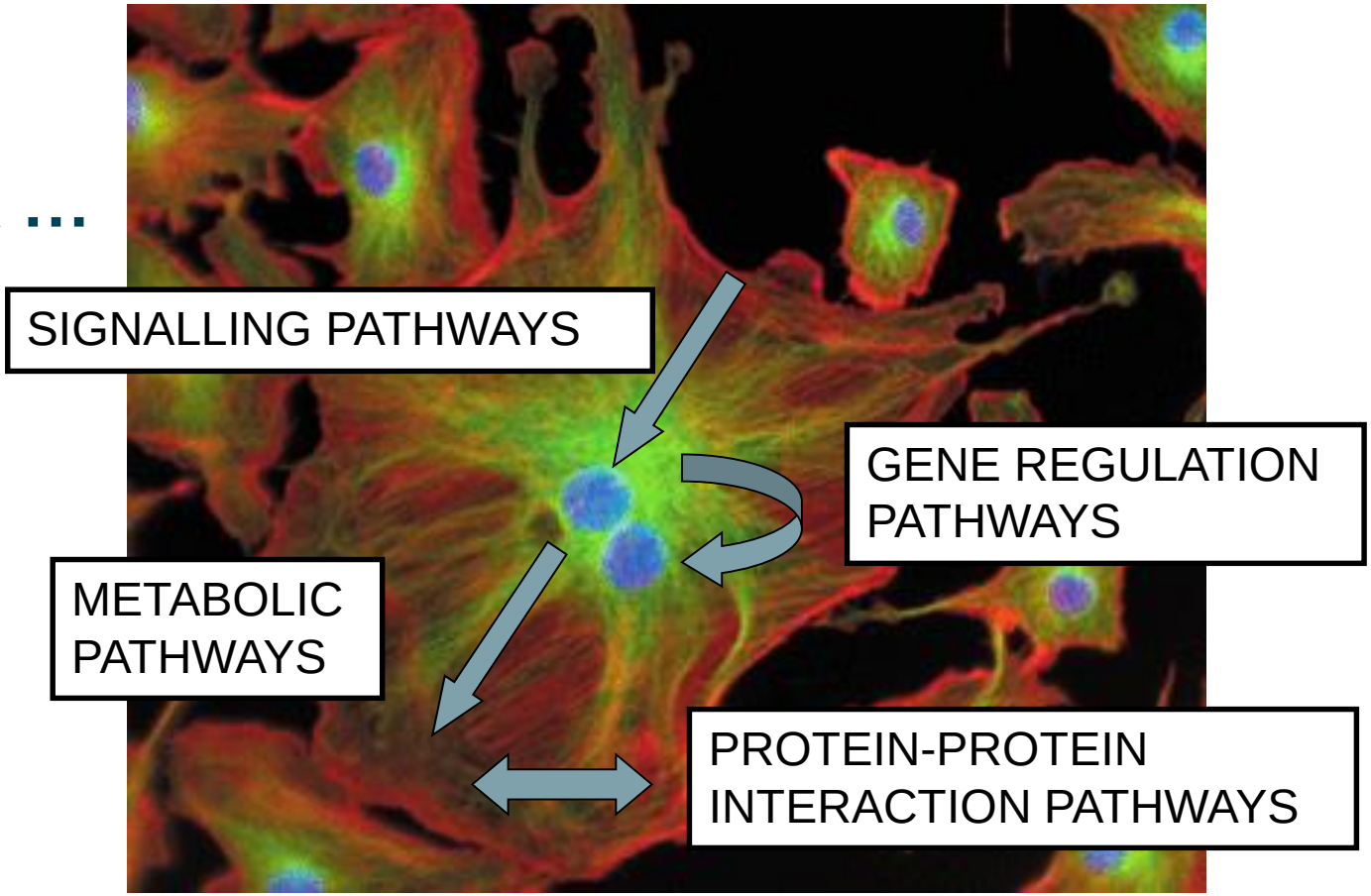


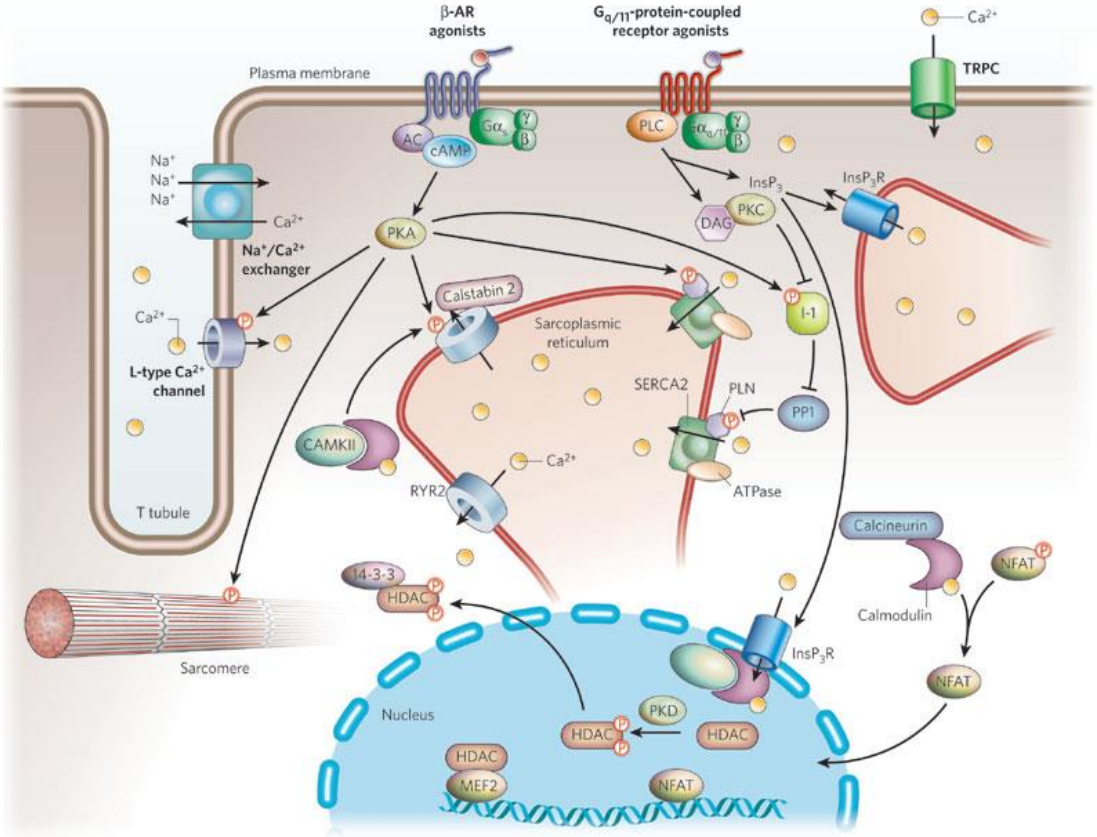
Figure 17-15
Biochemistry, Sixth Edition
© 2007 W. H. Freeman and Company

Biological Network Data helps integrate diverse types of data about biological components …

Key types of biological network data …



Signalling Pathways – heart disease

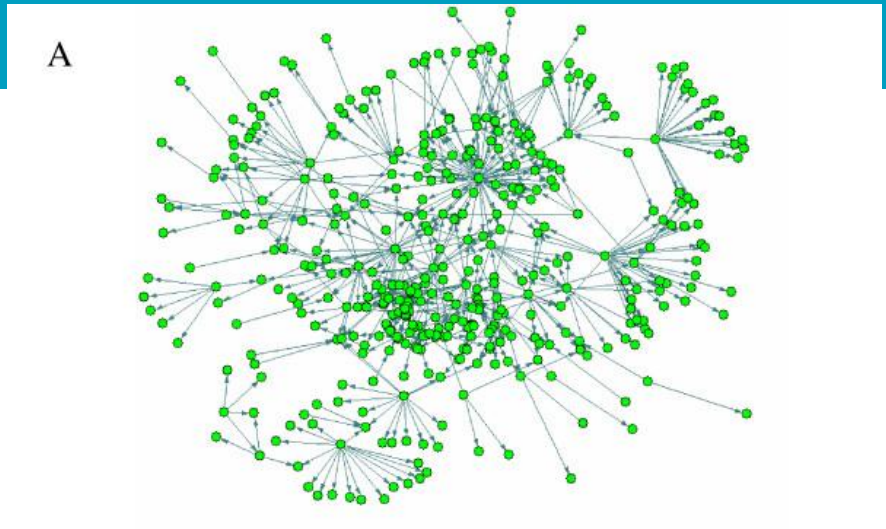


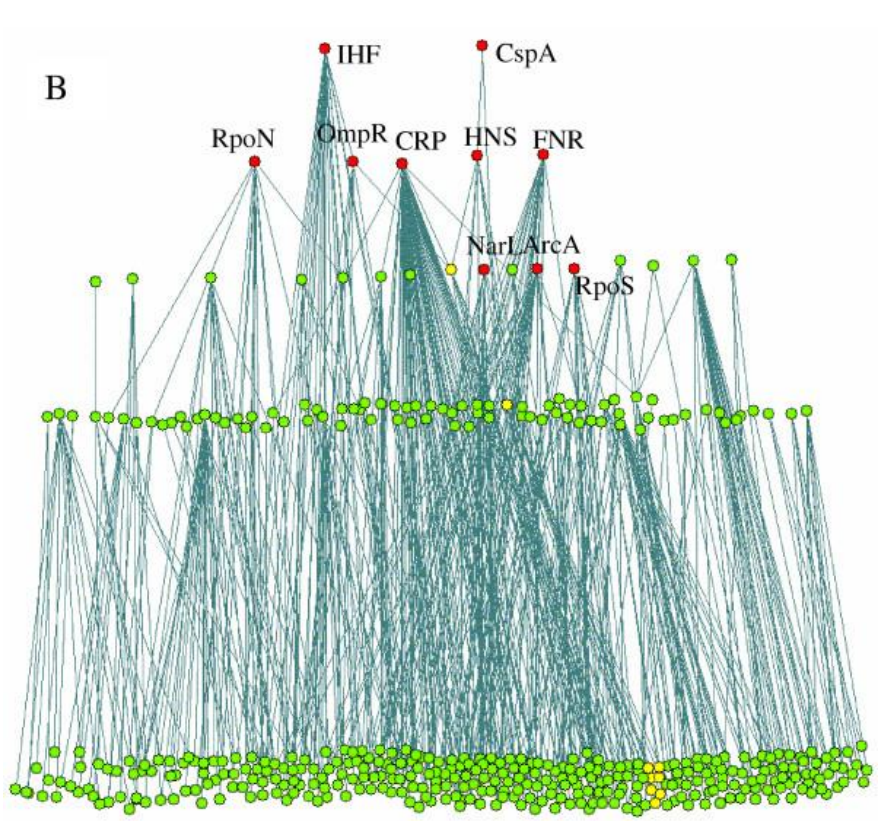
Nodes = Proteins

Edges = Activation/Inhibition (e.g., phosphorylation)

- Signalling pathways complex and ‘interwoven’.
- Involve gene regulation and protein/protein interaction, etc.

Tackling heart failure in the 21st Century (2008) *Nature* 451:919-928.





Gene Regulation Networks of *E. coli* *… a very simple organism !*

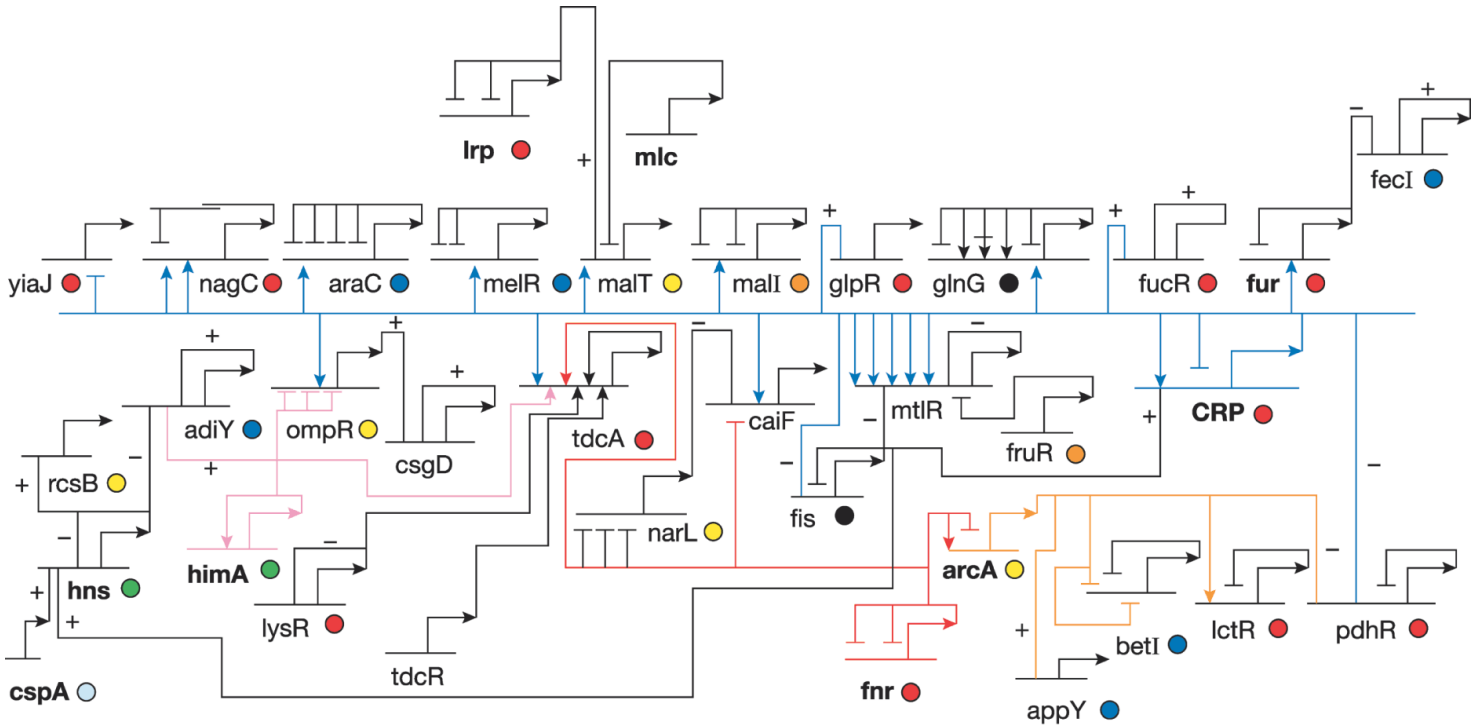
Nodes = Transcription Factors

Edges = Regulatory interaction

Gene regulation is hierarchical.

Hierarchical structure and modules in the Escherichia coli transcriptional regulatory network revealed by a new top-down approach (2004) *BMC Bioinformatics* **5:** 199

Annotated Gene Regulation Network of *E. coli*



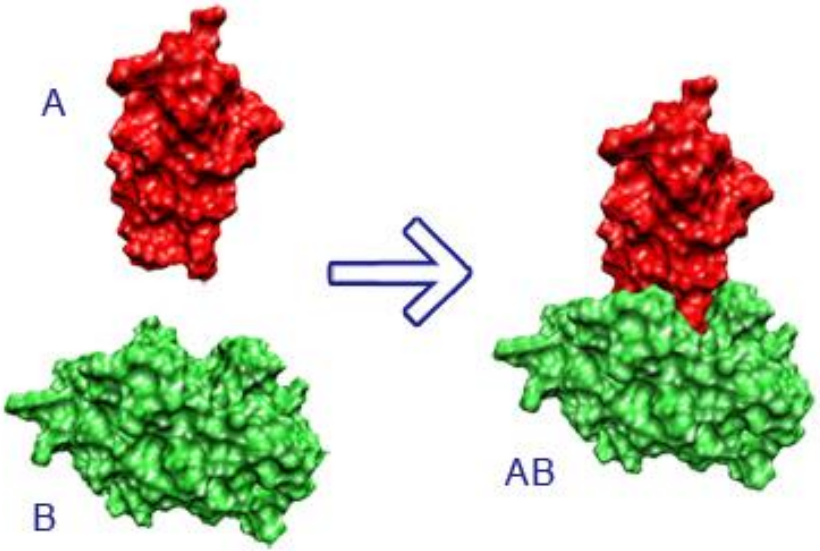
Expressed Gene

Activation

Repression

Charting gene regulatory networks: strategies, challenges and perspectives (2004) *Biochem. Journal* **381**:1-12.

Protein-Protein (P2P) Interaction Network

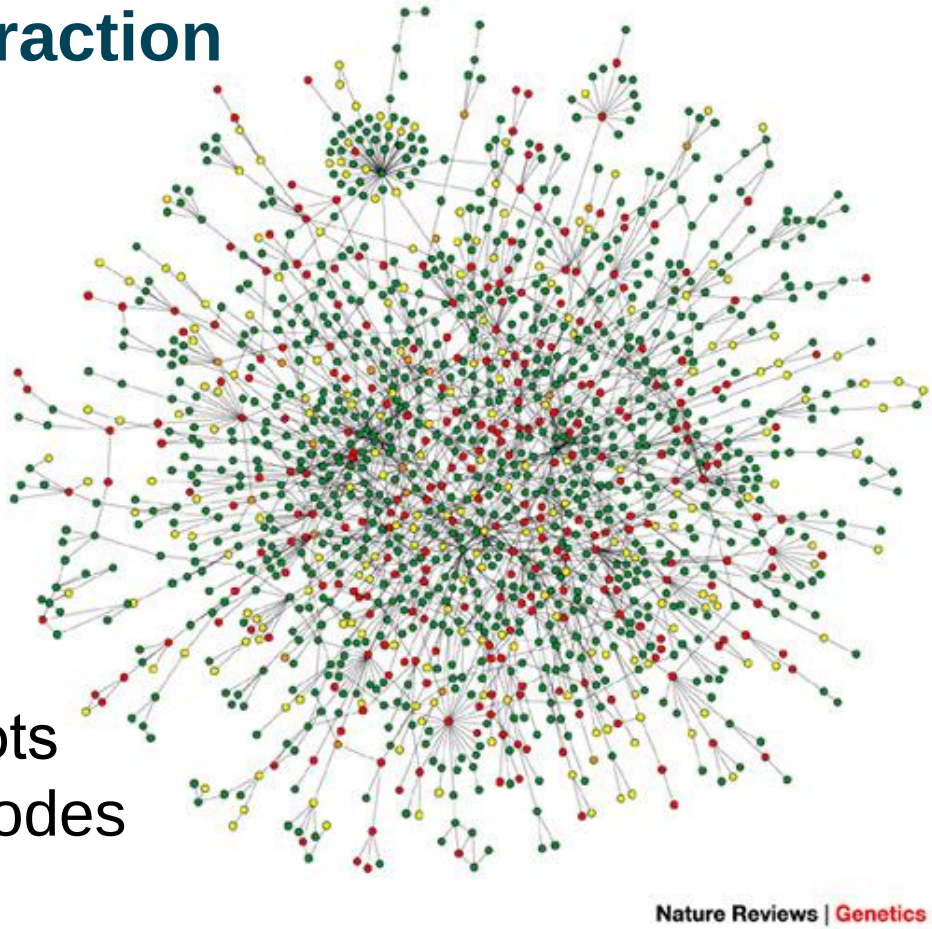


- Proteins interact with each other for a number of reasons:
  - Allosteric activation
  - Inhibitory binding (covering the active site)
  - Phosphorylation or modification of residues
  - Physical closeness of enzymes along metabolic pathway.
  - Etc.

Protein-Protein (P2P) Interaction Networks for Yeast

Nodes = Protein

Edges = Physical interaction between proteins



- Thought to be a ‘robust scale-free network’ (a few ‘fragile’ nodes with lots of connections and many nodes with few connections).
- Provides an indication of ‘functional modules’

Network biology: understanding the cell's functional organization (2004) Nature Reviews Genetics **5**:101-113.

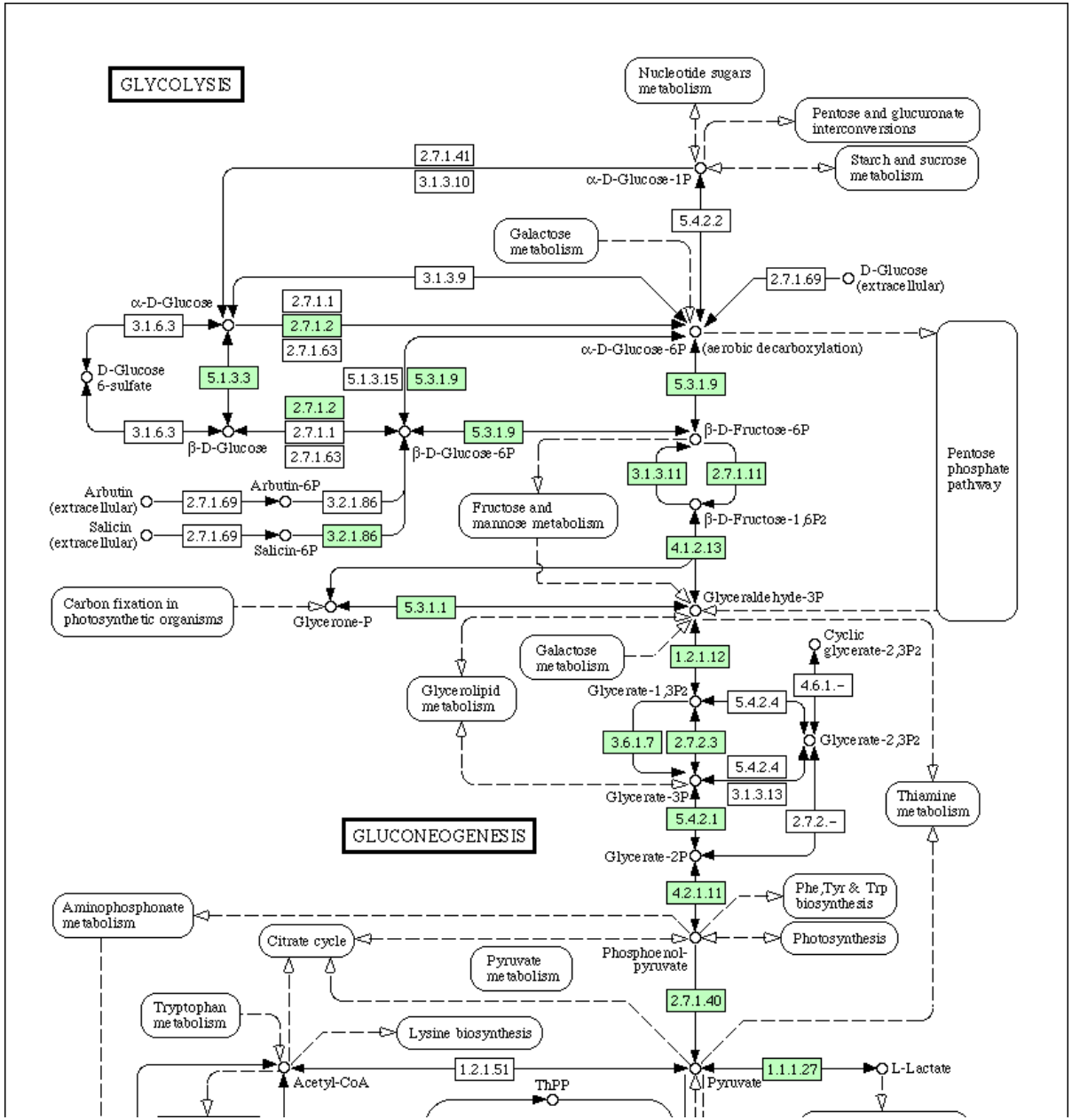
Metabolic Pathways

Node = metabolite (small molecule such as glucose)

Edge = conversion of one metabolite to another via enzyme

This shows a small part of glucose breakdown in a bacteria

The concentration of enzymes are regulated via the gene regulatory network (i.e. all the pathways are linked)



Oxford Journals | Life Sciences | Nucleic Acids Research | Database Summary Paper - Mozilla Firefox

www.oxfordjournals.org/our_journals/nar/database/subcat/6/27

OXFORD JOURNALS CONTACT US MY BASKET MY ACCOUNT

Nucleic Acids Research

ABOUT THIS JOURNAL CONTACT THIS JOURNAL SUBSCRIPTIONS CURRENT ISSUE ARCHIVE SEARCH

Oxford Journals › Life Sciences › Nucleic Acids Research › Database Summary Paper

2014 NAR Database Summary Paper

- Nucleotide Sequence Databases
- RNA sequence databases
- Protein sequence databases
- Structure Databases
- Genomics Databases (non-vertebrate)
- Metabolic and Signaling Pathways
  - ChemProt
  - Prokaryotic genome databases
  - Enzymes and enzyme nomenclature
  - Metabolic pathways
  - Protein-protein interactions
  - Signalling pathways
    - CR Cistrome
    - IUPHAR-DB
    - KBDOCK
    - NetworKIN
    - P2CS
    - PepCyber:P~Pep
    - PhosPhAt
    - PID
    - PRRDB
    - Quorumpeps
    - RegPhos
    - REPAIRtoire
    - SPIKE
    - UCSD-Nature Signaling Gateway Molecule Pages
- Human and other Vertebrate Genomes
- Human Genes and Diseases

Where to find out about this biological network data?

The NAR database provides a number of biological network resources.

Another source of information about biological networks is the PathGuide resource

- Biological pathway resources given at: http://www.pathguide.org/

- Currently 547 resources for:
  - Signalling Pathways
  - Gene Regulation Pathways
  - Protein-Protein Interactions
  - Metabolic Pathways
  - Protein-Compound Interactions

Pathguide: the pathway resource list - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://www.pathguide.org/

Pathguide: the pathway resource list

Home BioPAX cBio MSKCC

Pathguide » the pathway resource list

Navigation

- Protein-Protein Interactions
- Metabolic Pathways
- Signaling Pathways
- Pathway Diagrams
- Transcription Factors / Gene Regulatory Networks
- Protein-Compound Interactions
- Genetic Interaction Networks
- Protein Sequence Focused
- Other

Complete Listing of All Pathguide Resources

News

Find: KEGG

Pathguide: the pathway resource list - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://www.pathguide.org/

Pathguide: the pathway resource list

GeneNet - Genetic Networks	Details	Free	
GenMAPP - Gene MicroArray Pathway Profiler	Details	Free	
GO - Gene Ontology	Details	Free	
GOLD.db - Genomics of Lipid-associated Disorders	Details	Free	
HMDB - Human Metabolome Database	Details	Free	
IMG - Integrated Microbial Genomes	Details	Free	
Indigo - Gene Neighborhoods and Codon Usage	Details	X	
IntEnz - Integrated relational Enzyme database	Details	Free	
iPath - Invitrogen iPath	Details	$	
JWS Online - Online Cellular Systems Modelling	Details	Free	SBML
KEGG - Kyoto Encyclopedia of Genes and Genomes	Details		BioPAX
LIGAND - Database of Chemical Compounds and Reactions in Biological Pathways	Details	Free	
Malaria - Malaria Parasite Metabolic Pathways	Details	Free	
MetaCore - MetaCore pathway database	Details	$	
MetaCyc - Metabolic Pathway Database	Details		BioPAX SBML
MetNetDB - Metabolic Network Exchange	Details	Free	
Millipore Pathways - Your online source for visualizing metabolic and signaling	Details	Free	

Find: KEGG Next Previous Highlight all Match case

Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://www.genome.jp/kegg-bin/xml/xmlview?-v+0.6.1+ko+0001

KEGG BioPAX

http://www.geno...+0.6.1+ko+00010

```
<?xml version="1.0"?>
<!DOCTYPE pathway SYSTEM "http://www.genome.jp/kegg/xml/KG
<!-- Creation date: Apr 22 2009 12:02:47 +0900 (JST) -->
<pathway name="path:ko00010" org="ko" number="00010"
         title="Glycolysis / Gluconeogenesis"
         image="http://www.genome.jp/kegg/pathway/ko/ko000
         link="http://www.genome.jp/dbget-bin/show_pathway
    <entry id="1" name="ko:K00128" type="ortholog" reactio
        link="http://www.genome.jp/dbget-bin/www_bget?ko+K
        <graphics name="K00128" fgcolor="#000000" bgcolor=
             type="rectangle" x="262" y="940" width="46" h
    </entry>
    <entry id="2" name="ko:K01895" type="ortholog" reactio
        link="http://www.genome.jp/dbget-bin/www_bget?ko+K
        <graphics name="K01895" fgcolor="#000000" bgcolor=
             type="rectangle" x="119" y="908" width="46" h
    </entry>
    <entry id="3" name="ko:K00129" type="ortholog" reactio
        link="http://www.genome.jp/dbget-bin/www_bget?ko+K
        <graphics name="K00129" fgcolor="#000000" bgcolor=
             type="rectangle" x="262" y="961" width="46" h
    </entry>
    <entry id="4" name="cpd:C00033" type="compound"
        link="http://www.genome.jp/dbget-bin/www_bget?compound+C00033">
        <graphics name="C00033" fgcolor="#000000" bgcolor="#FFFFFF"
             type="circle" x="119" y="951" width="8" height="8"/>
    </entry>
    <entry id="5" name="path:ko00030" type="map"
        link="http://www.genome.jp/kegg/pathway/ko/ko00030.html">
        <graphics name="Pentose phosphate pathway" fgcolor="#000000" bgcolor="#FFFFFF"
             type="roundrectangle" x="629" y="336" width="62" height="237"/>
    </entry>
```

Find: Com Next Previous Highlight all Match case Reached end of page, continued from top

Network data is much more complex than sequence data – many standards employing different XML-languages such as SBML, PSI-MI, BioPAX, etc.

Biological networks/pathways provide ‘global’ views about how the biological parts/components within a cell interact …

- But what information is available on the biological components themselves?

- Well all the classic data (sequences, structures, function annotation, etc.) is generally available via established databases (UniProt, PDB, etc.) as you would have covered with David Jones.

- In addition, there is **high-throughput ‘-omics’ data** that provides ‘global’ information about particular types of components within a cell (e.g. the concentrations of all mRNAs within a cell in a particular situation).

Introduction to HT –omics data

- Gen*omics*
- Transcript*omics*
- Prote*omics*
- Metabol*omics*

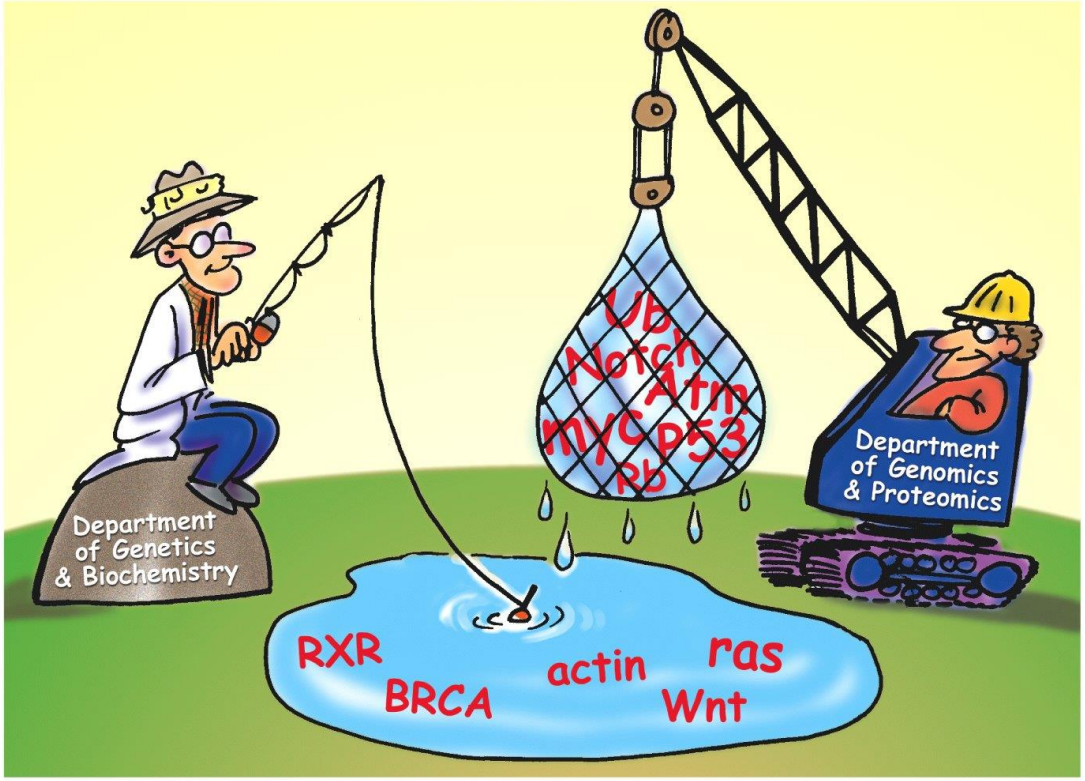


Figure 2-46 Cell and Molecular Biology, 4/e (© 2005 John Wiley & Sons)

Overview ... recall the central dogma of molecular biology ...

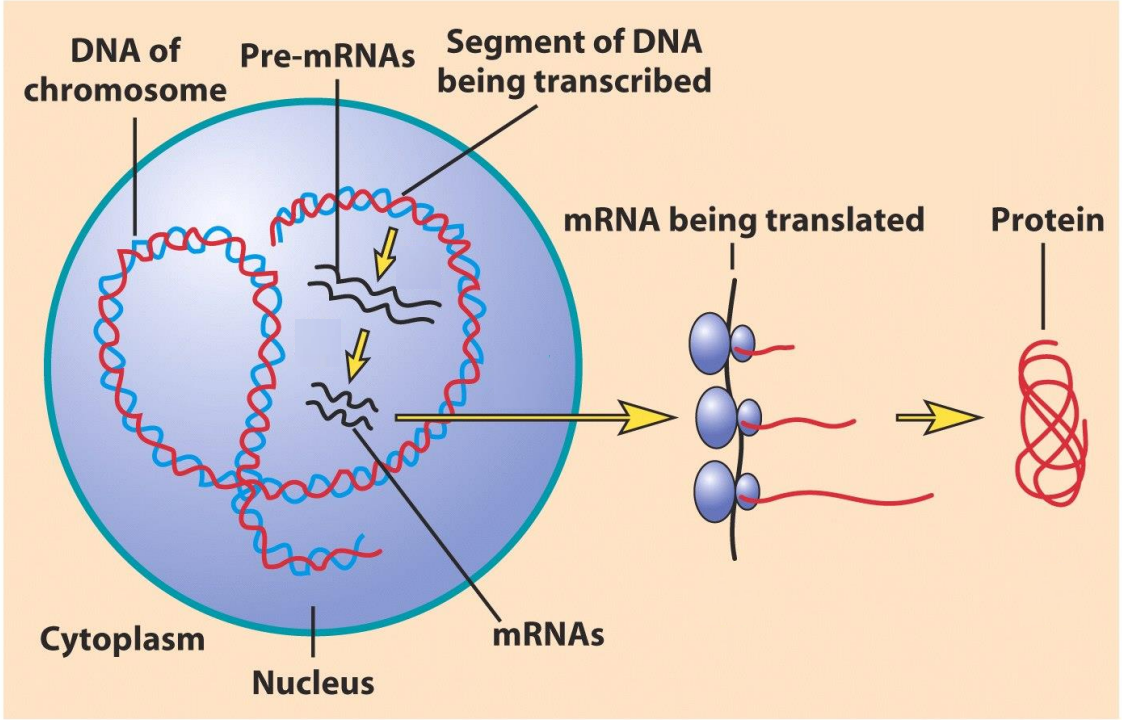


Figure 11-2 Cell and Molecular Biology, 4/e (© 2005 John Wiley & Sons)

Deoxyribonucleic acid (DNA)

↓ transcription

Ribonucleic acid (RNA)

↓ splicing of introns

Messenger RNA (mRNA)

↓ translation

Amino acid string

↓ folding

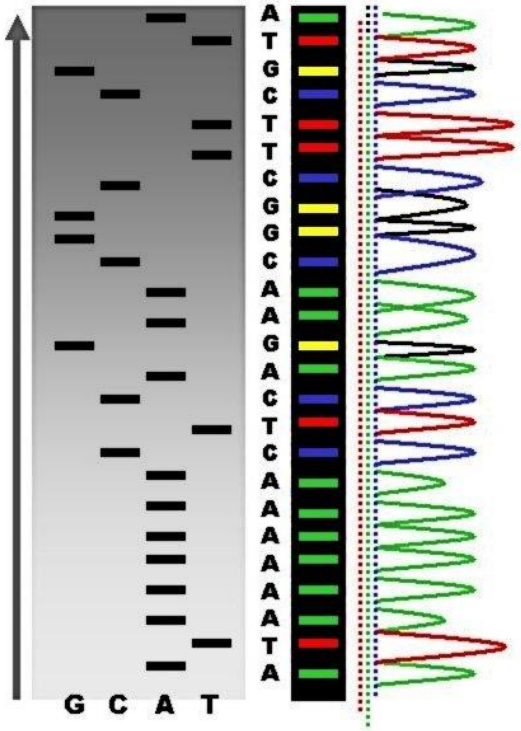
Protein

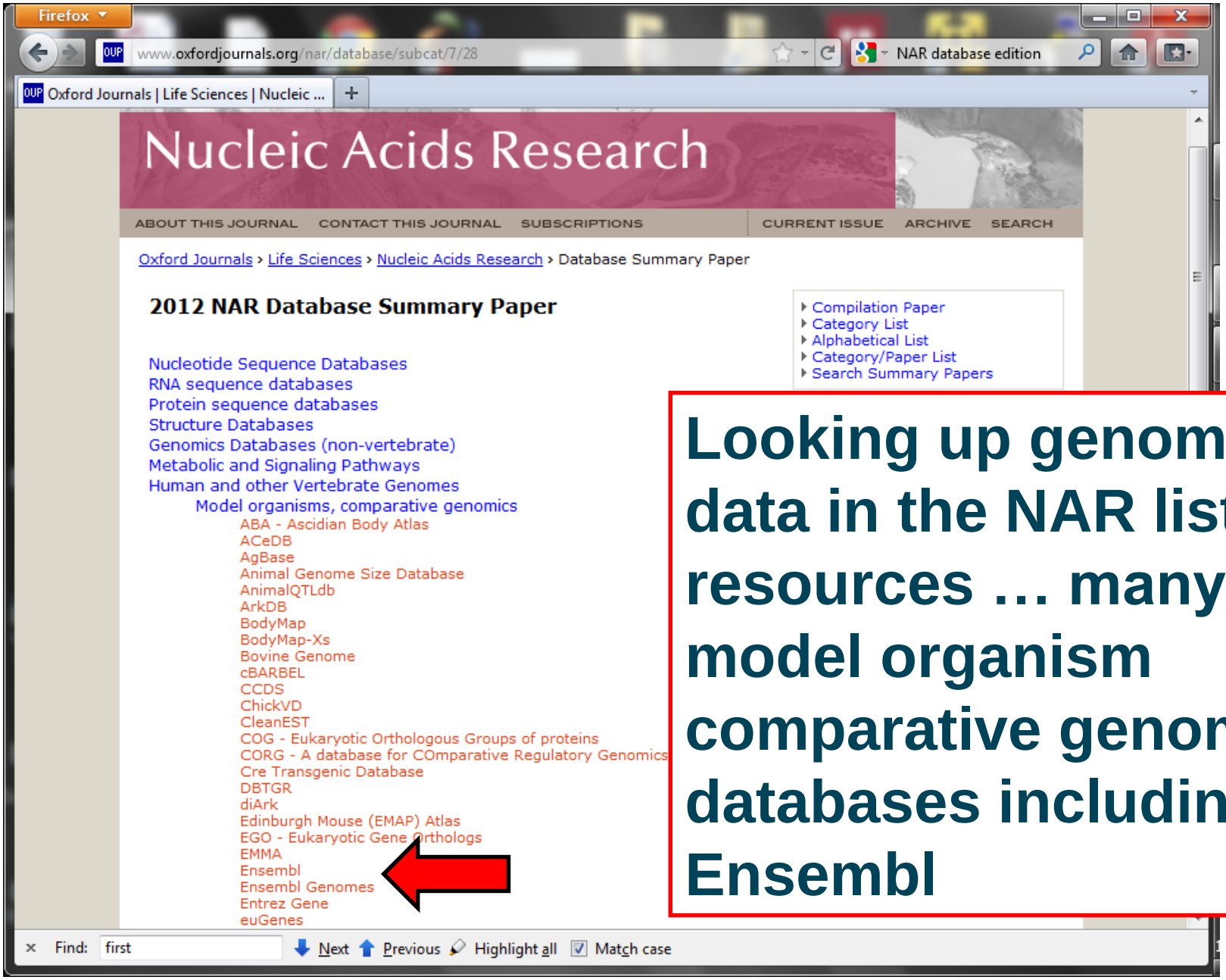
Gen-*omics*

- Genomics is about determining the complete sequence of the DNA genome of an organism.

- The first complete genome ever sequenced was the phage virus Φ-X174 with approx' 5000 bp in 1977 by Fred Sanger.
- The first 'free living' organism was the bacteria H. Influenza with approx' 1.8 Mbp in 1995.
- Human genome with approx' 3 Gbp was finally sequenced in April 2003.

The Sanger method has dominated nucleotide sequencing since 1977 ...
Sequence DNA fragments and then assembly these parts into a complete genome.

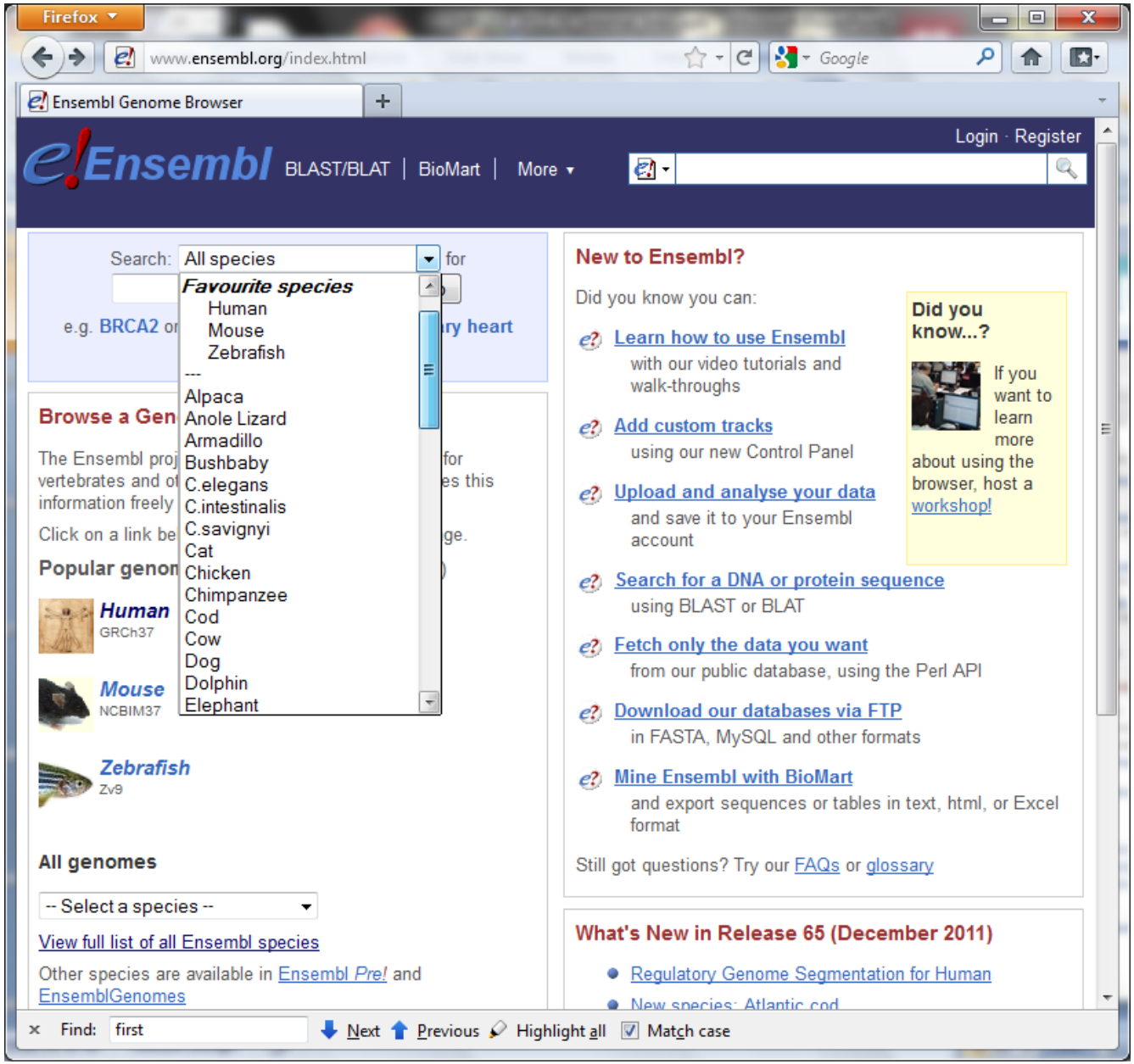


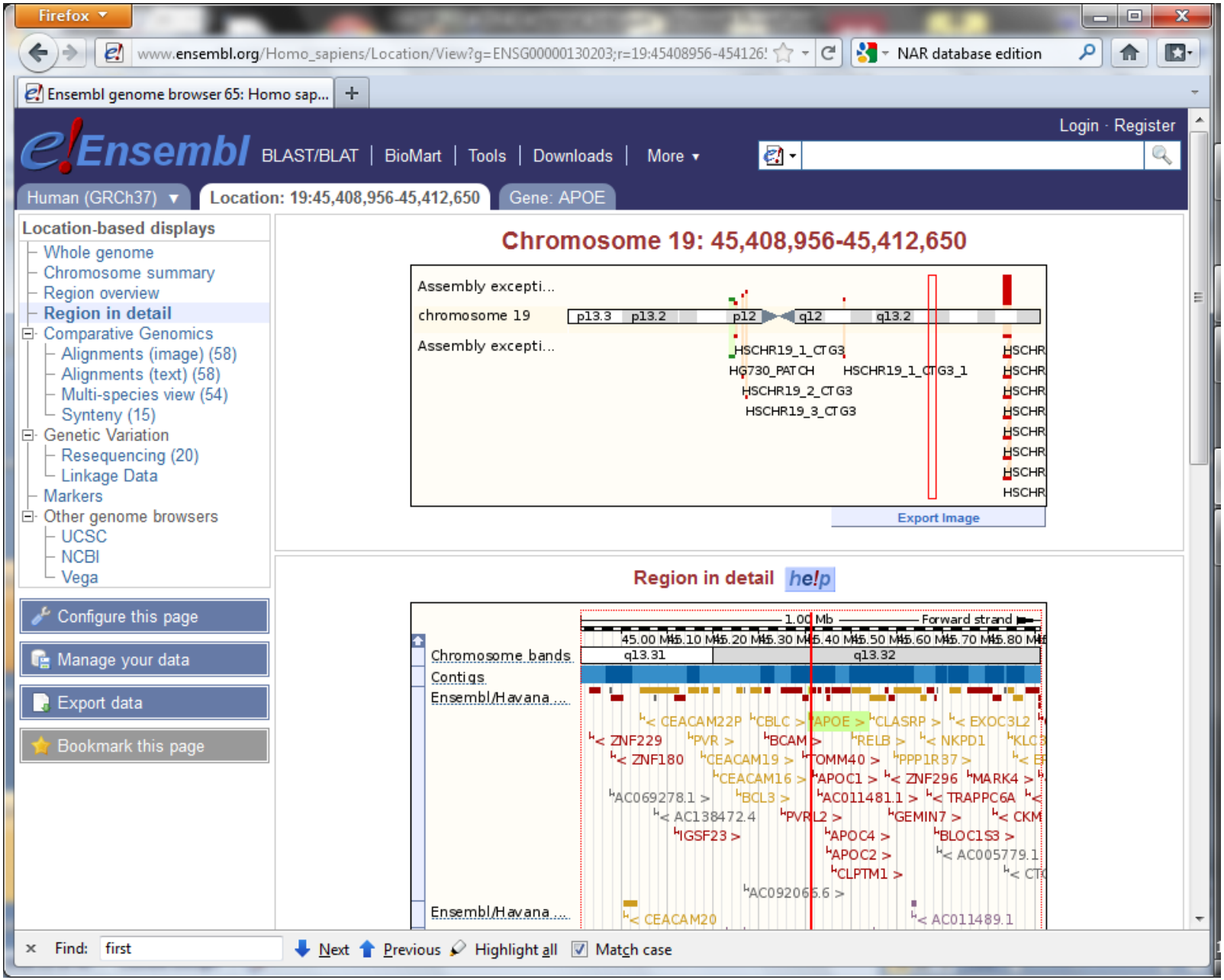


Looking up genomics data in the NAR list of resources … many model organism comparative genomics databases including Ensembl

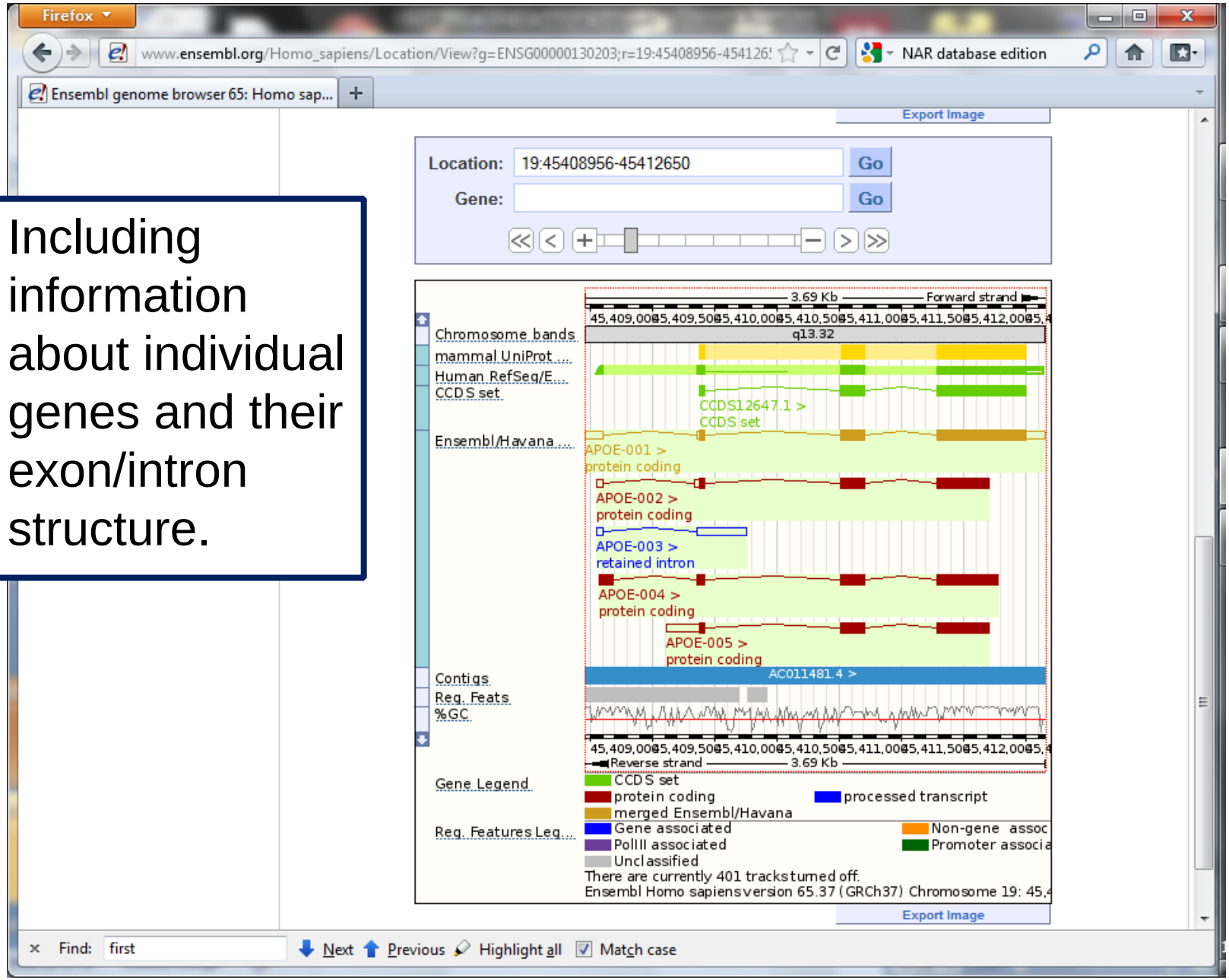
Ensembl genomics resource

- Provides sequence, function and other relevant data organized around the framework of an organism's genome.

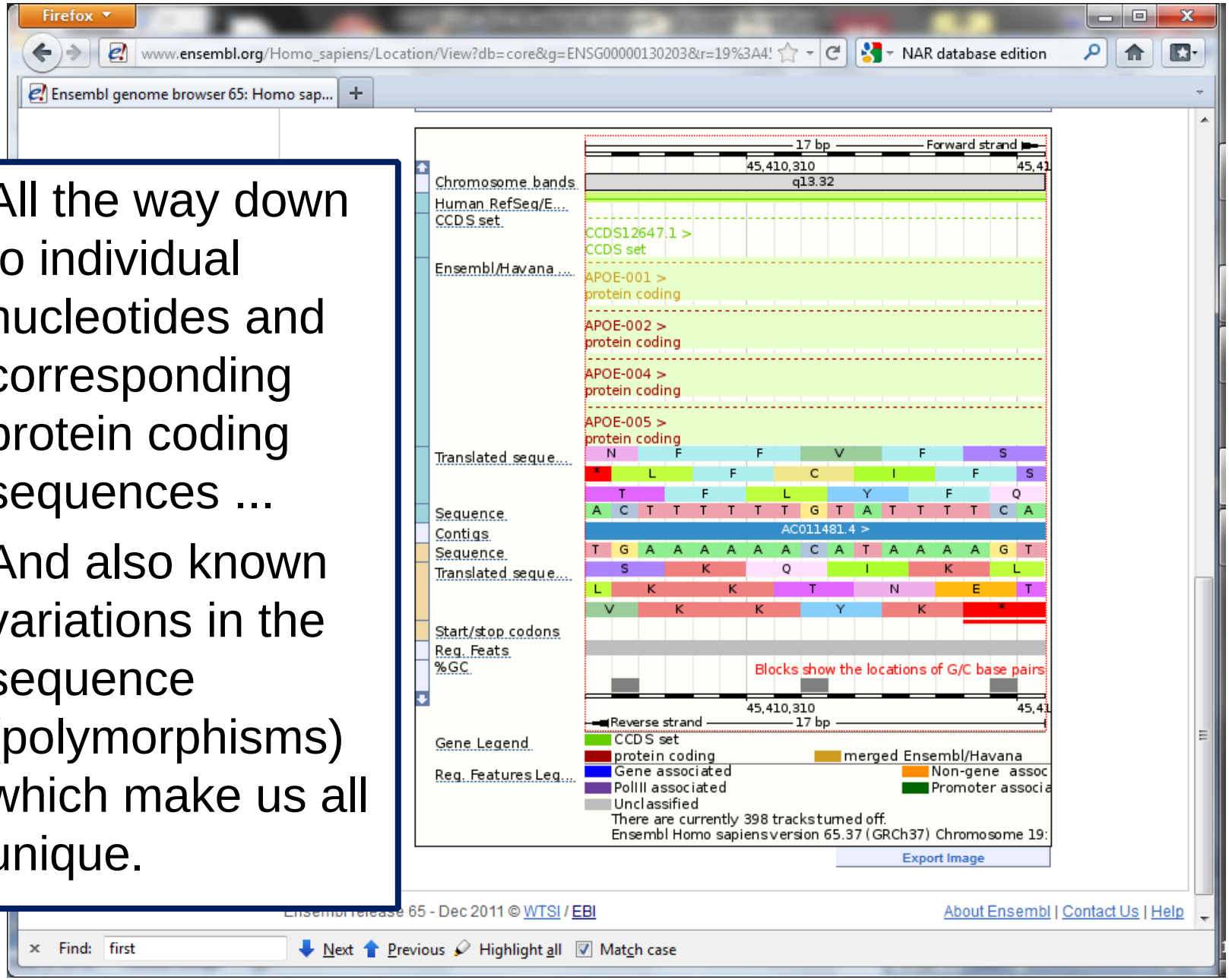



Firefox
www.ensembl.org/Homo_sapiens/Location/View?g=ENSG00000130203;r=19:45408956-454126
NAR database edition
Ensembl genome browser 65: Homo sap...
Login · Register
Ensembl
BLAST/BLAT | BioMart | Tools | Downloads | More
Human (GRCh37)
Location: 19:45,408,956-45,412,650
Gene: APOE
Location-based displays
Whole genome
Chromosome summary
Region overview
Region in detail
Comparative Genomics
Alignments (image) (58)
Alignments (text) (58)
Multi-species view (54)
Synteny (15)
Genetic Variation
Resequencing (20)
Linkage Data
Markers
Other genome browsers
UCSC
NCBI
Vega
Configure this page
Manage your data
Export data
Bookmark this page
Chromosome 19: 45,408,956-45,412,650
Assembly excepti...
chromosome 19
p13.3
p13.2
p12
q12
q13.2
Assembly excepti...
HSCHR19_1_CTG3
HG730_PATCH
HSCHR19_1_CTG3_1
HSCHR19_2_CTG3
HSCHR19_3_CTG3
Export Image
Region in detail
help
1.00 Mb
Forward strand
Chromosome bands
q13.31
q13.32
Contigs
Ensembl/Havana ...
CEACAM22P
CBLC
APOE
CLASRP
EXOC3L2
ZNF229
PVR
BCAM
RELB
NKPD1
ZNF180
CEACAM19
TOMM40
PPP1R37
CEACAM16
APOC1
ZNF296
MARK4
AC069278.1
BCL3
AC011481.1
TRAPPC6A
AC138472.4
PVRL2
GEMIN7
IGSF23
APOC4
BLOC1S3
APOC2
AC005779.1
CLPTM1
AC092065.6
Ensembl/Havana ...
CEACAM20
AC011489.1
Find: first
Next
Previous
Highlight all
Match case

- Including information about individual genes and their exon/intron structure.



- All the way down to individual nucleotides and corresponding protein coding sequences ...
- And also known variations in the sequence (polymorphisms) which make us all unique.



Genome variation within a species

- Even simple phenotypic traits result from complex interactions between many genes/proteins.

- The dog genome was initiated to try to unravel these relationships between:
genotype ←→ phenotype

Transcript-*omics*

- Transcriptomics is about globally measuring the expression level of all the different mRNAs within a tissue.
- Essentially giving you information about the genes being expressed in a tissue.

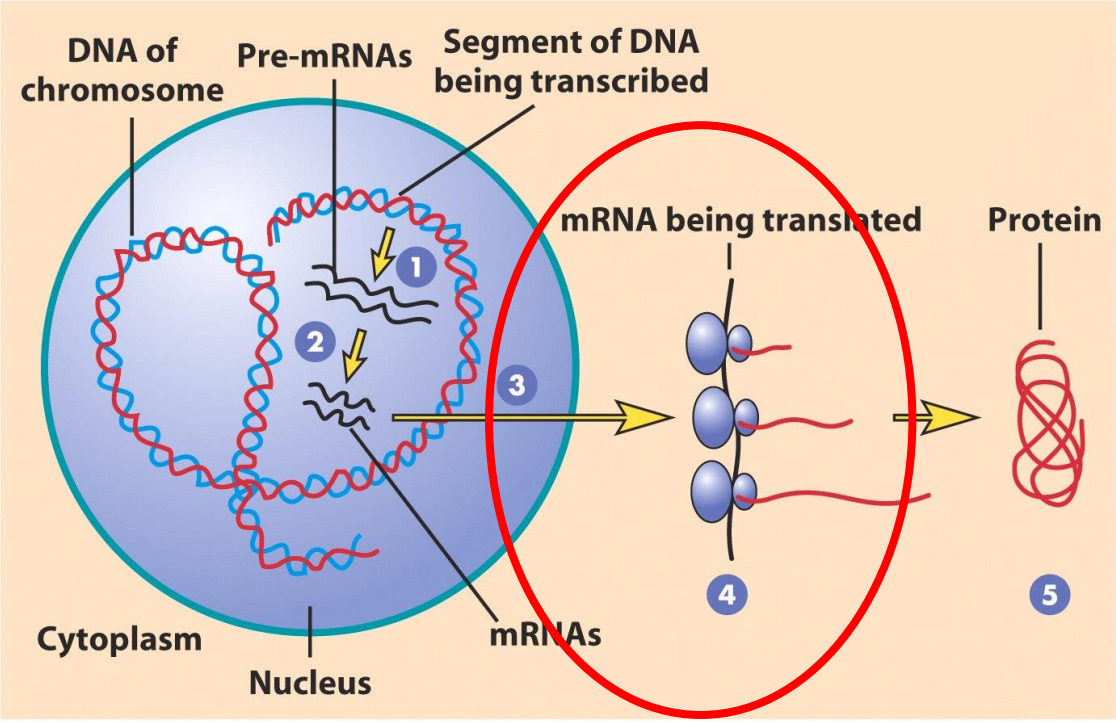
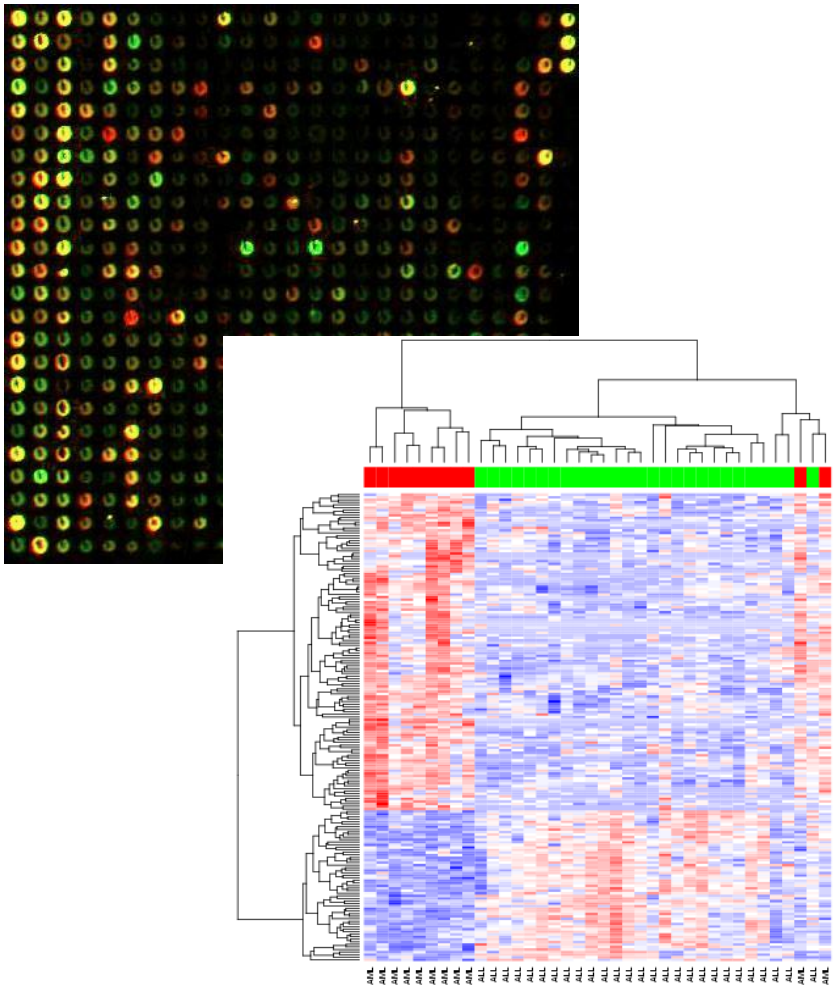


Figure 11-2 Cell and Molecular Biology, 4/e (© 2005 John Wiley & Sons)

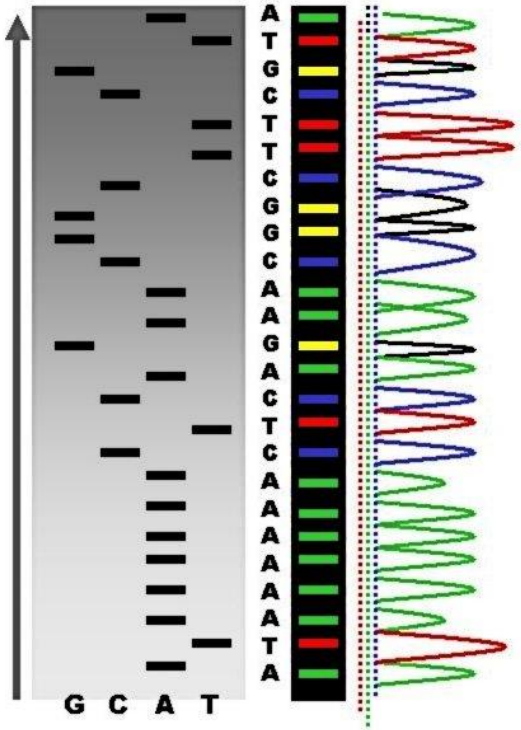
Transcriptomics is a very mature field with data standards and analysis techniques

- Provides gene expression levels (mRNA) under different conditions.
- Established databases such as GEO and ArrayExpress.
- Established data formats such as MAGE-ML, etc. http://www.fged.org/
- Extensive open-source software such as BioConductor.



Transcriptomics based on *Next Generation Sequencing (NGS)* is currently revolutionizing the field: massively parallel sequencing ... actually counting the mRNA molecules one at a time!

Next Generation Sequencing is like having 160,000,000 Sanger machines all in a row!



Prote-*omics*

- Proteomics is about globally measuring the concentration/activity of all the different proteins within a tissue.
- Giving you direct info' about what protein activities are present in the tissue.

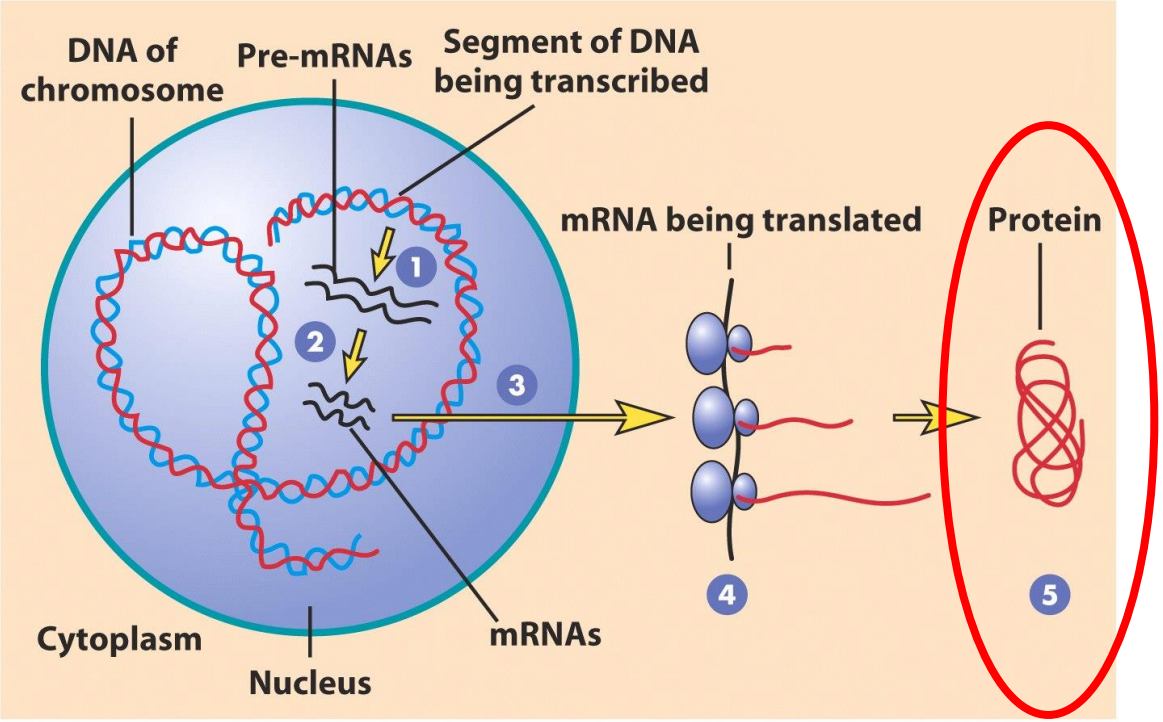
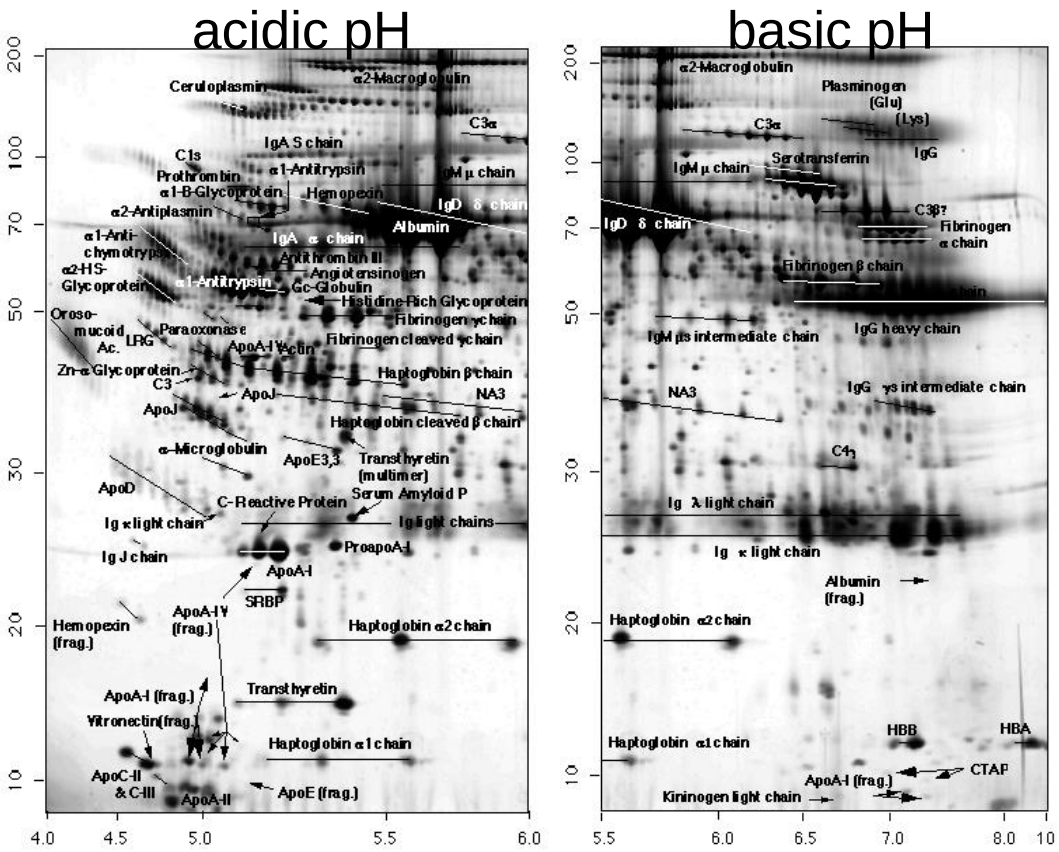


Figure 11-2 Cell and Molecular Biology, 4/e (© 2005 John Wiley & Sons)

Proteomics is starting to mature …

- Provides information about protein concentration under different conditions.
- Established techniques such as: 2D PAGE, Mass Spec', Protein Chips.
- Proteomics standards being developed: http://www.psidev.info/



Metabol-*omics*

- Metabolomics is about globally measuring the concentrations of all the metabolites (or small molecules) within a tissue or sample.

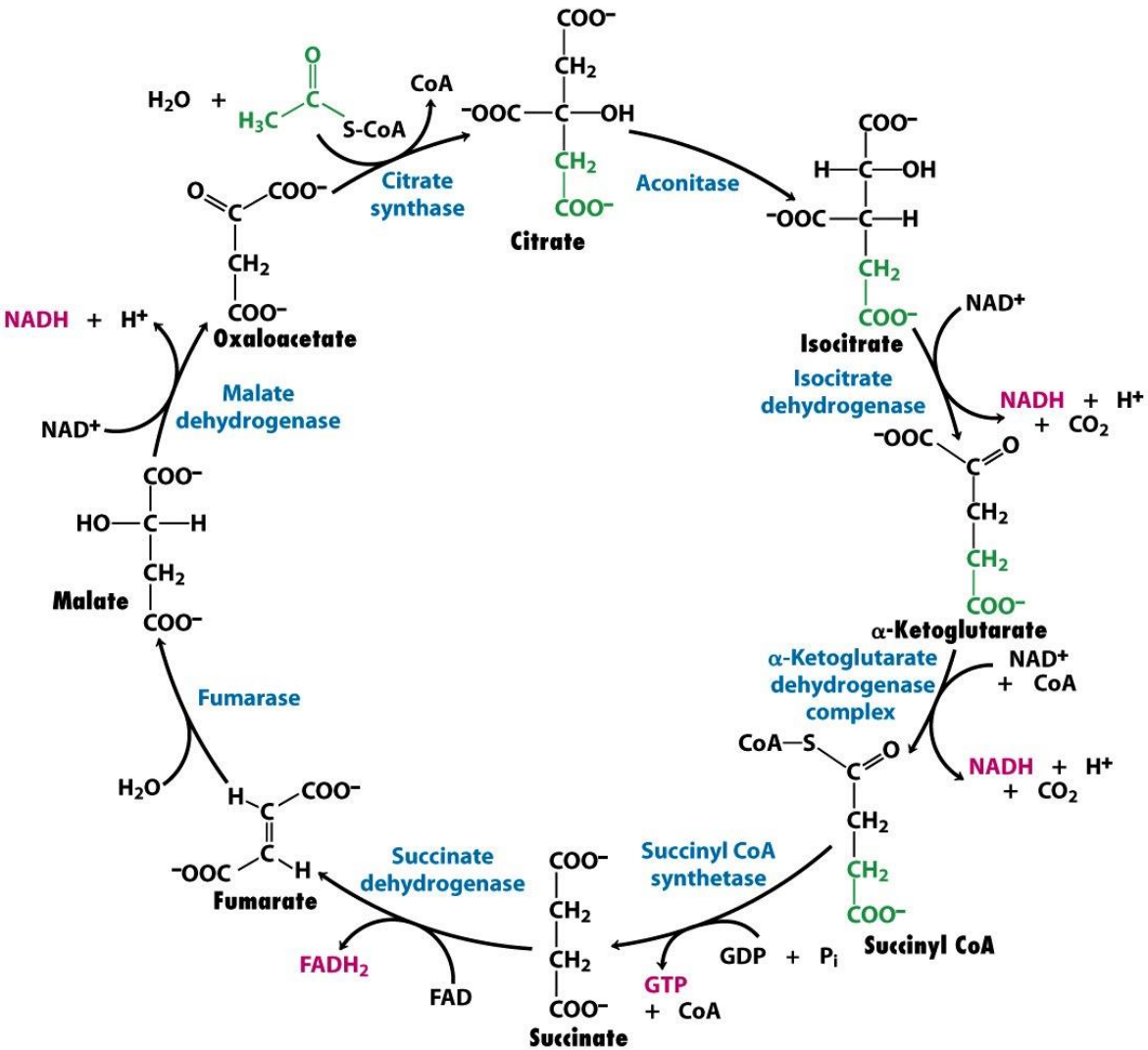
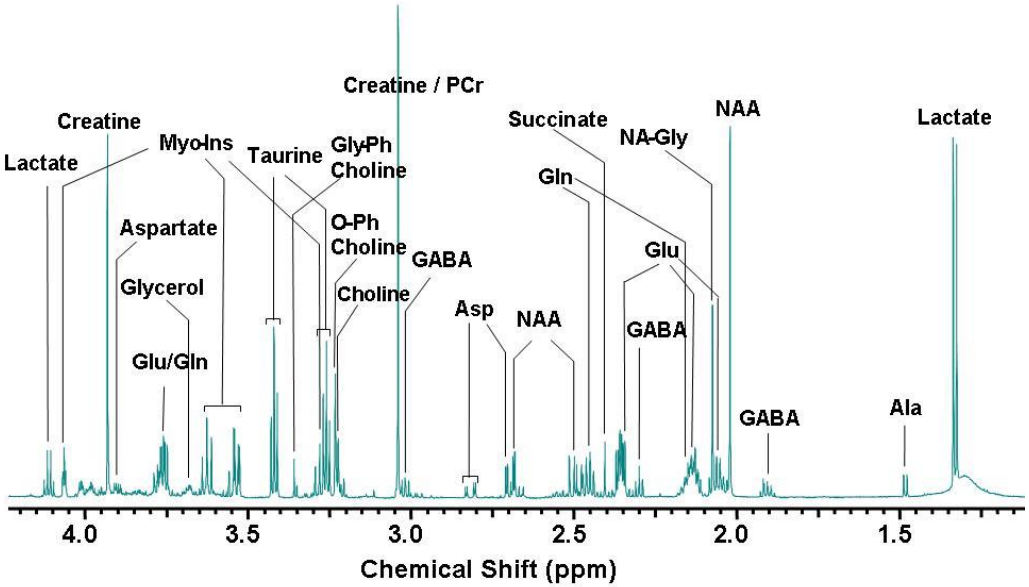


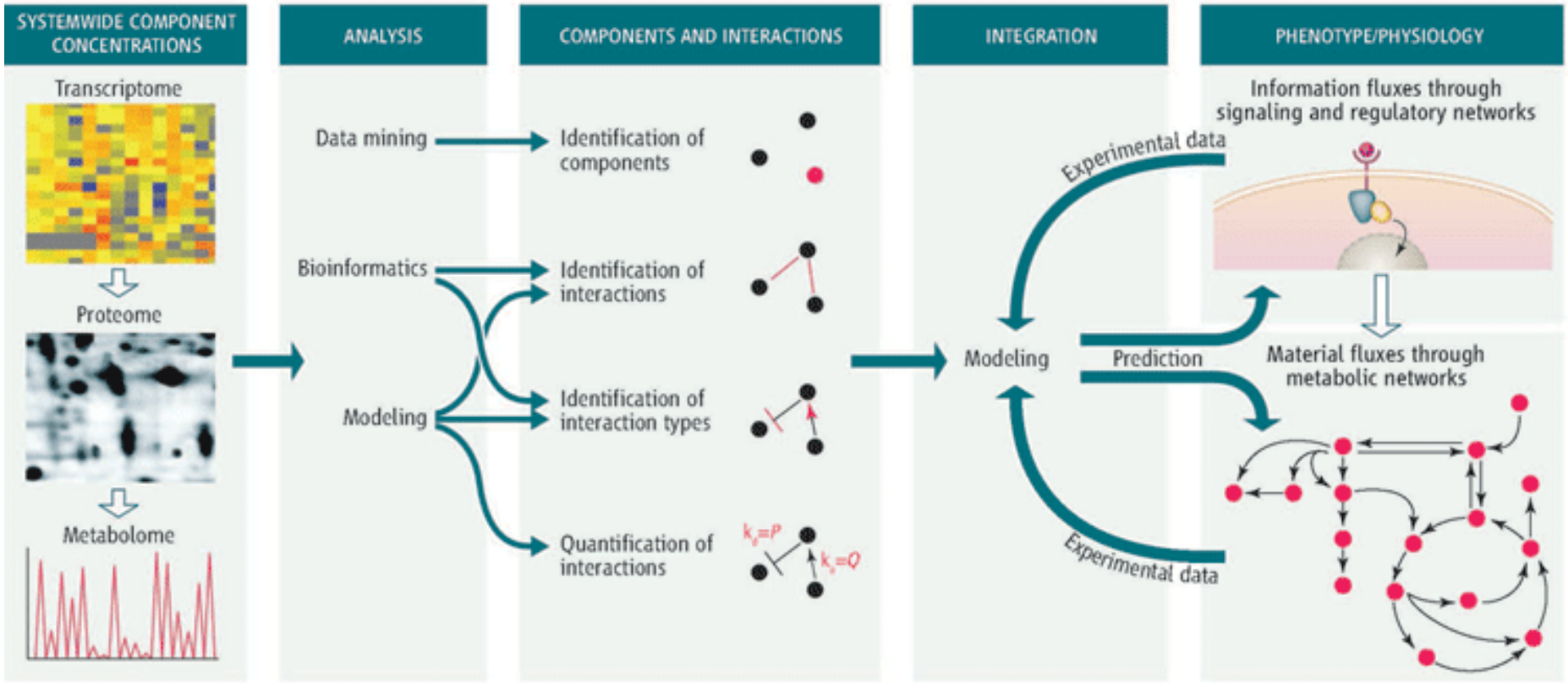
Figure 17-15
Biochemistry, Sixth Edition
© 2007 W. H. Freeman and Company

Metabolomics is also maturing rapidly …

- Provides information about small molecule metabolites under different conditions.
- Key technique is high-resolution NMR.
- Metabolomics data standards currently under development: http://msi-workgroups.sourceforge.net/

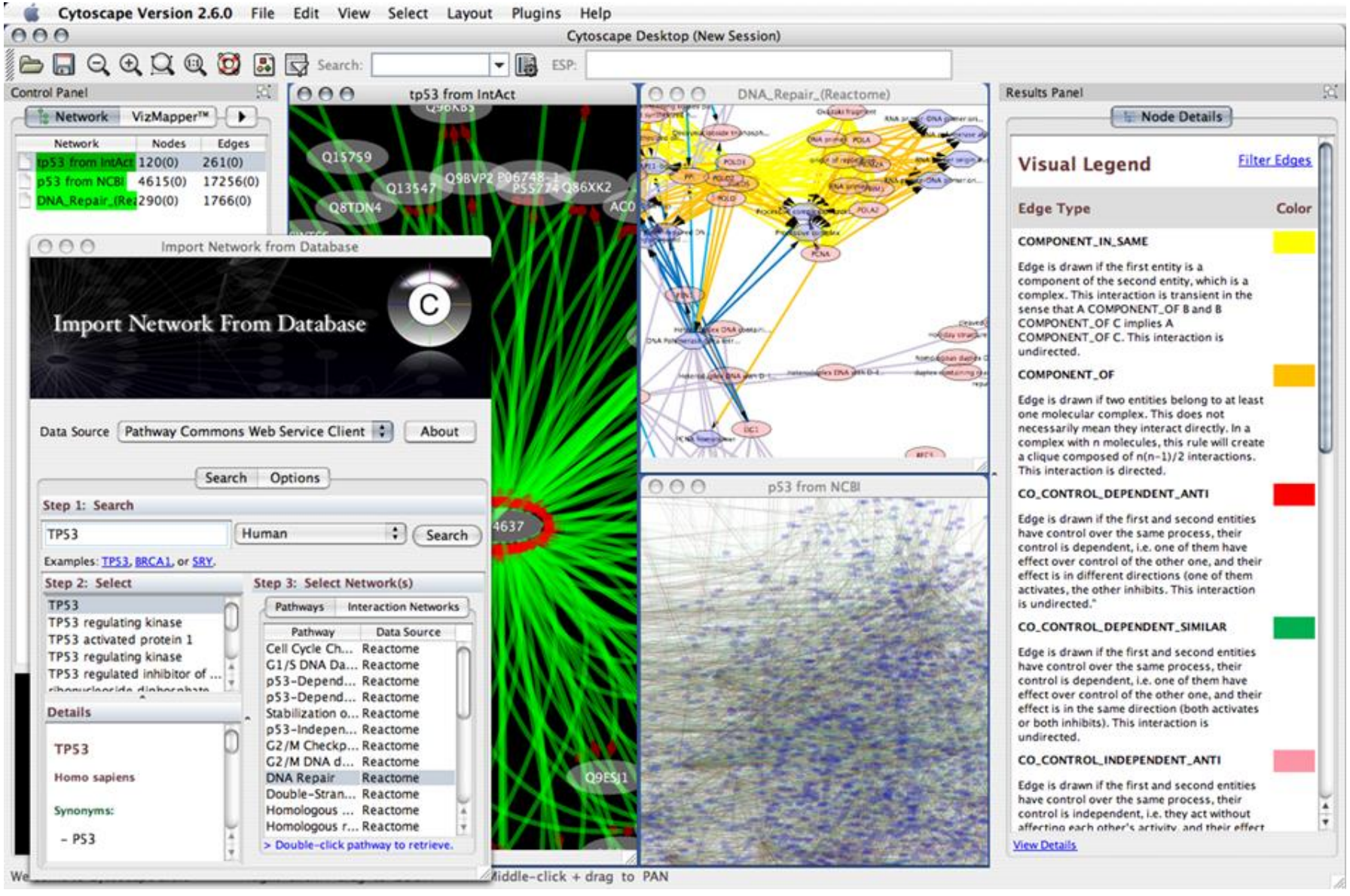


Biology is now rich with complex data – machine learning can be used to integrate and unravel complex patterns within this data …

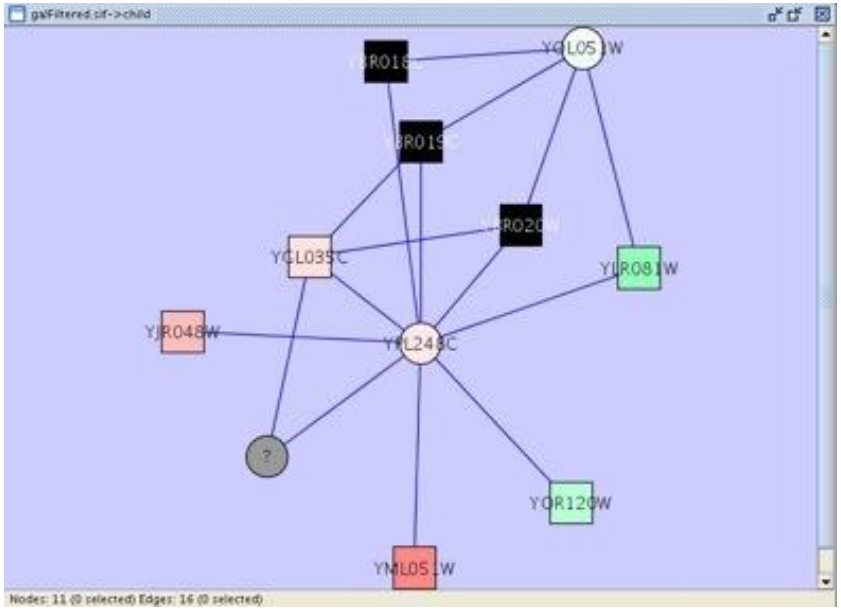


Getting closer to the whole picture (2007) *Science* **316**(5824), 550 – 551.

Integrating heterogeneous data with Cytoscape



Patterns within data (derived by hand by overlaying expression data onto gene/protein networks …



- YML051W (Gal80) is significantly down-regulated (red square).
- The black and green squares are significant up-regulation – these seem to be around YPL248C (Gal4).
- But Gal4 itself is not significantly changed (white circle).
- However, knowing Gal80 repressed Gal4 … it can be inferred that Gal80 is down-regulated, thus up-regulating Gal4 which then up-regulates all the genes it activates.

- The up-regulation of Gal4 *cannot* be detected via microarrays since most transcription factors have very low expression levels (below noise level).
- However taking the network information into account does strongly suggests that Gal4 is up-regulated.

Summary

- In this lecture we have introduced systems biology as trying to understand the behaviour of a complete biological system under study (or some aspect of this complete system).
- This requires a good understanding of the components making up a particular biological system (genes, proteins, metabolites, etc.) and how they interact.
- We introduced a number of high-throughput –omics technologies (genomics for getting a ‘parts list’, transcriptomics/proteomics/metabolomics to get a picture of the concentrations of various components).
- We also introduced biological network data (signalling pathways, gene regulation networks, protein-protein interaction networks and metabolite networks).