LSM2241
From sequence and structure to networks and pathways

Greg Tucker-Kellogg
dbsgtk@nus.edu.sg

28 October 2015

Outline

Molecular profiling

Organising genes into functional groups

Detecting interactions between genes and proteins

Representing genome wide data in networks

Pathway and interaction databases

Roundup and next week

Topic

Molecular profiling

Organising genes into functional groups

Detecting interactions between genes and proteins

Representing genome wide data in networks

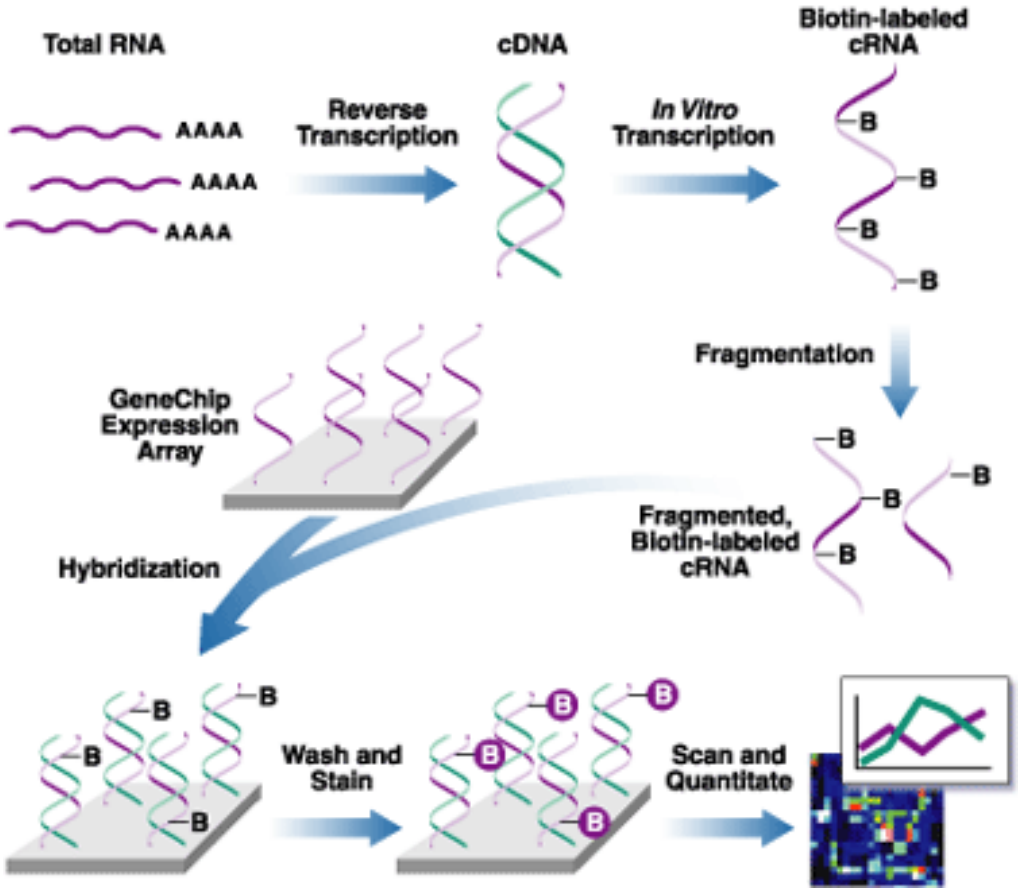
Pathway and interaction databases

Roundup and next week

We are swimming in data

- We can measure gene expression on a genome-wide scale
- We can sequence the genome of an organism in days (or hours, or minutes)
- We can use structural genomics to predict the structures of millions of proteins from thousands of known structures

A simple study



An old (now outdated) Affymetrix "GeneChip" to measure transcription of every gene

We need to *annotate* molecular profiling data, or it is impossible to interpret!

1. Start with some human cells in culture
2. Treat cells with a drug or protein
3. Measure gene expression of >20,000 genes transcripts using a "gene chip".
4. Compare gene expression in treated versus untreated cells using fancy statistics
5. Identify a large number (>1000) genes showing significant expression differences7
6. What does this list of genes tell us about the experiment we just did? Or "What's so special about this list of genes?"

Topic

Molecular profiling

Organising genes into functional groups

Detecting interactions between genes and proteins

Representing genome wide data in networks

Pathway and interaction databases

Roundup and next week

Taxonomies of function

Why taxonomies help

- Taxonomies of species help us interpret evolutionary events
- Taxonomies of structure help us compare and classify domains
- Taxonomies of literature help us search the library
- Taxonomies of *function* should help us understand functional genomics

The problem

- Scientists, not nature, assign categories of function
- Categories conflict
  - Is it a kinase, or a protein involved in cell motility?
  - It's a floor wax *and* a dessert topping!
- Hierarchies of function are what scientists decide make sense

What is an Ontology?

An *Ontology* can be seen as a type of taxonomy or controlled vocabulary.

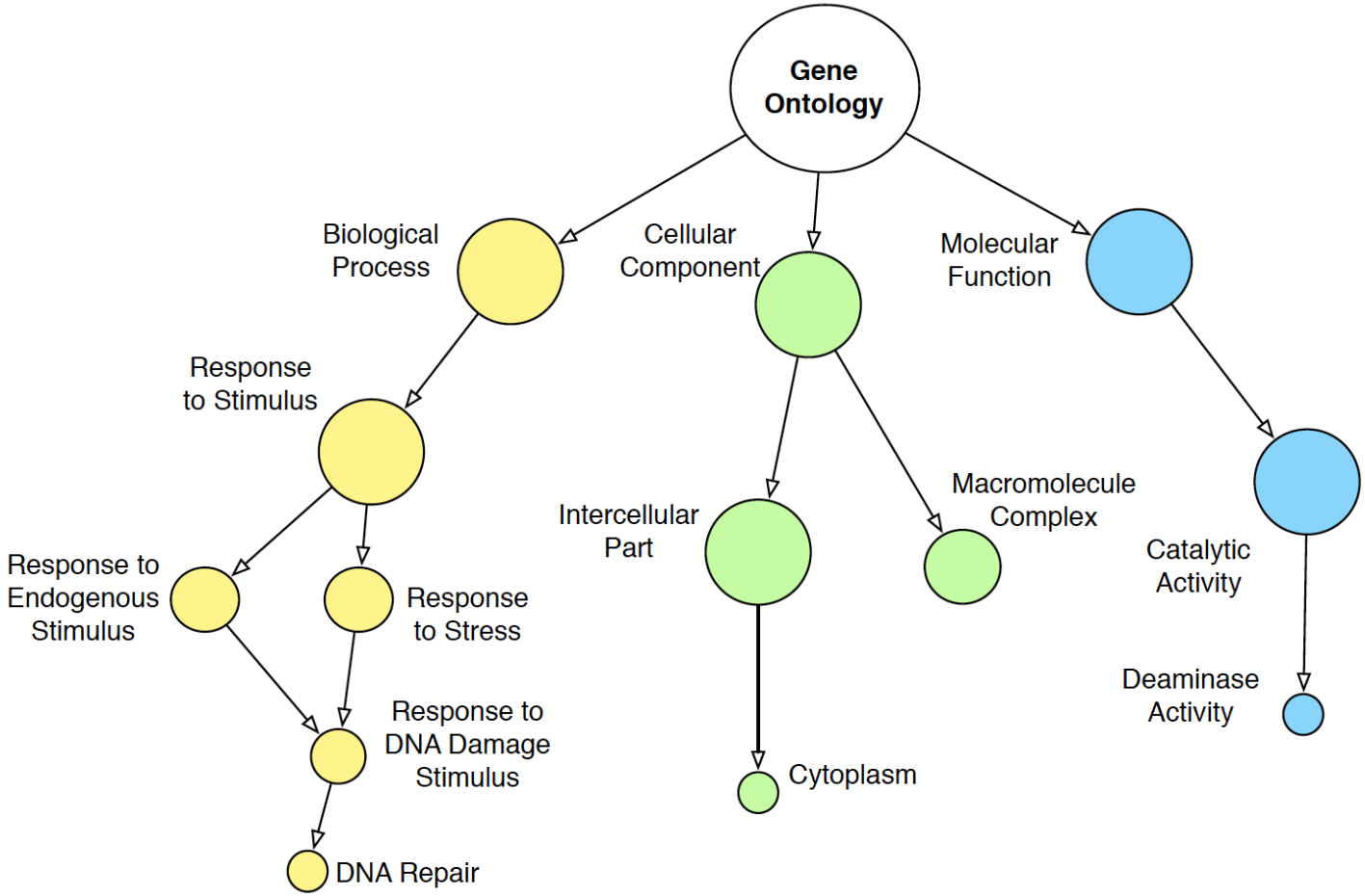
- They are a *formal* specification of concepts, usually within a domain, and the relationships between those concepts
- Meant to be readable and interpretable by humans
- Meant to be readable and interpretable by computers
- Computers can use the structure of ontologies to infer implicit knowledge by reasoning

The Gene Ontology

The Gene Ontology is the most widely used ontology in basic biomedical research

- Three top level domains
  - Molecular Function (MF)
  - Biological Process (BP)
  - Cellular Compartment (CC)
- Everything else is below these levels, with a *relationship* of children to parent terms
  - is a
  - part of
  - develops from
  - regulates
  - negatively regulates
  - positively regulates

graphical example of the Gene Ontology



From the PhD thesis of Sara Mostafavi (U. Toronto, 2011)

Gene ontology annotations are coded by categories of evidence

	Experimental Evidence Codes
EXP	Inferred from Experiment
IDA	Inferred from Direct Assay
IPI	Inferred from Physical Interaction
IMP	Inferred from Mutant Phenotype
IGI	Inferred from Genetic Interaction
IEP	Inferred from Expression Pattern
	Reviewed Computational Analysis Evidence Codes
ISS	Inferred from Sequence or Structural Similarity
ISO	Inferred from Sequence Orthology
ISA	Inferred from Sequence Alignment
ISM	Inferred from Sequence Model
IGC	Inferred from Genomic Context
RCA	Inferred from Reviewed Computational Analysis
	Author Statement Evidence Code
TAS	Traceable Author Statement
NAS	Non-traceable Author Statement
	Curator Statement Evidence Codes
IC	Inferred by Curator
ND	No biological Data available
	Automatically Assigned Evidence Codes
IEA	Inferred from Electronic Annotation

With genes in sets, you can identify functions using *guilt by association*

What do the genes I found in my experiment have in common?

- What *terms* do they have in common?
- Are there some terms that occur more often than expected by chance?
- There are many such methods

The wrong way to assess functional association

- Suppose we see that a lot of genes in our list belong to a particular category of function?
- Does this mean the function is meaningful?
  - NO
  - Maybe there are a lot of genes in that category overall

Topic

Molecular profiling

Organising genes into functional groups

Detecting interactions between genes and proteins

Representing genome wide data in networks

Pathway and interaction databases

Roundup and next week

The goals of protein interaction studies

Given	Any protein or gene
Find	What proteins (or genes) interact with it
Where	"interact" has a meaning that we understand and agree

Using *bait* to find interacting proteins

- Most methods rely on the use of one protein or gene, called a *bait*, to identify other proteins or genes that interact with it
- Each bait requires a genetic manipulation to do the experiment
- Creating a bait construct can be tedious, but methods are being improved to allow automation

Two example methods

The yeast two-hybrid system identify interactions by activating a reporter through protein-protein binding events

Tandem Affinity Purification identify protein interactions by purifying protein-protein binding events

Synthetic lethal studies identify interacting *genes* by screening deletion libraries using a knockout

Yeast two hybrid system

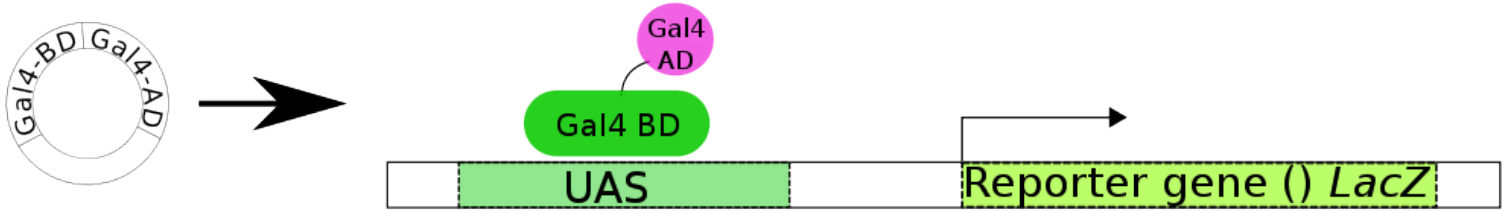
Premise

- Binding domains and activating domains of transcription factors are modular, but do not have to bind directly to activate transcription
- Proteins of interest can be positioned between binding and activating domains

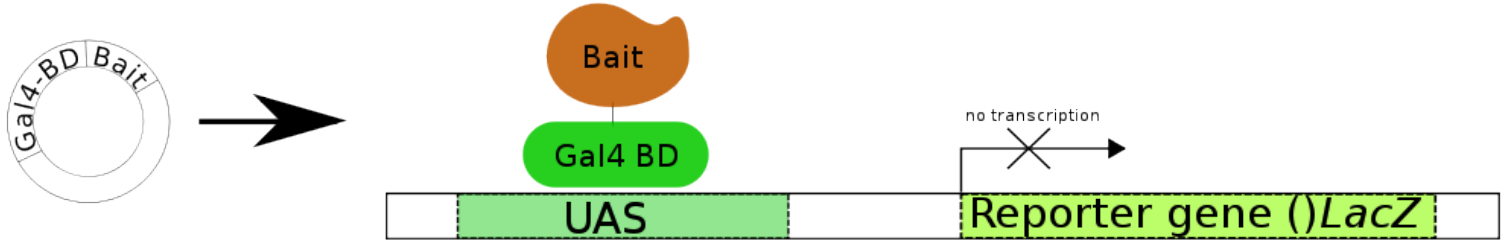
System

Genetically engineered yeast with fusion proteins and a metabolic requirement that can be used for screening

Yeast two-hybrid assay (1)



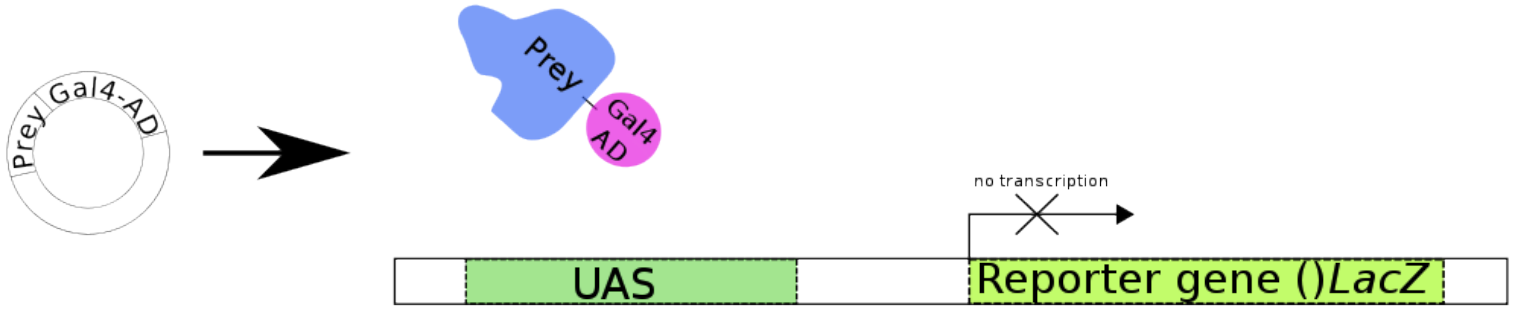
A. Regular transcription of the reporter gene



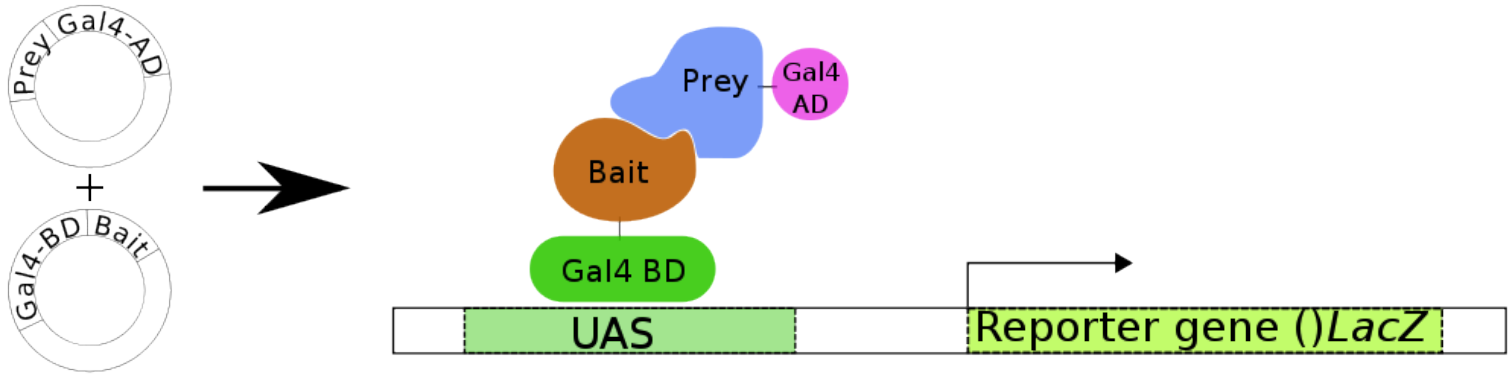
B. One fusion protein only (Gal4-BD + Bait) - no transcription

From WikiPedia: Two hybrid screening

Yeast two-hybrid assay (2)



C. One fusion protein only (Gal4-AD + Prey) - no transcription



D. Two fusion proteins with interacting Bait and Prey

From WikiPedia: Two hybrid screening

Tandem Affinity Purification

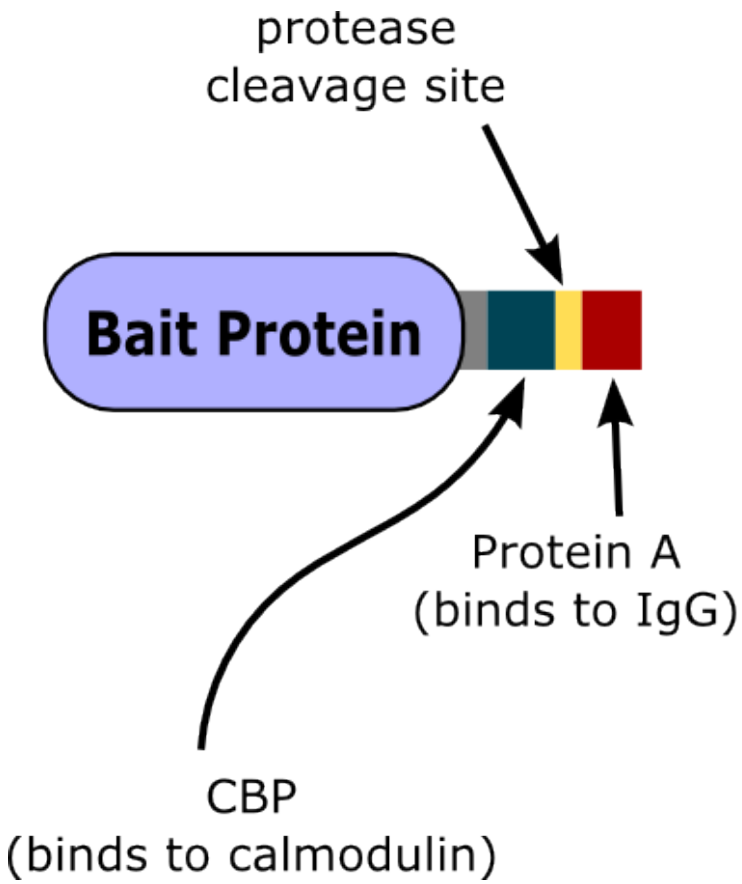
Premise

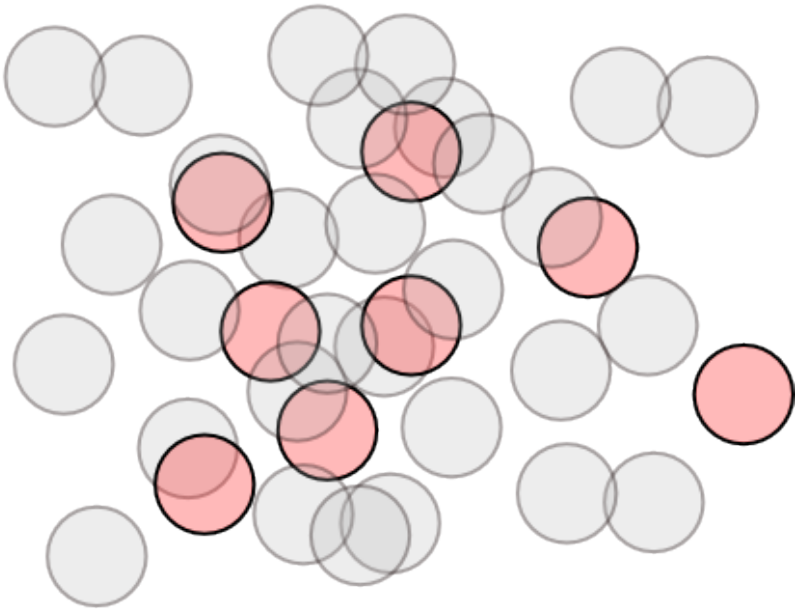
Proteins that interact with each other should purify together under native binding conditions

System

1. Engineer fusion proteins with tandem affinity tag
2. (over) express fusion proteins in cells, release proteins under native conditions
3. Purify the protein and its binding partners with successive use of immobilized beads
4. Use mass spectrometry to identify binding partners

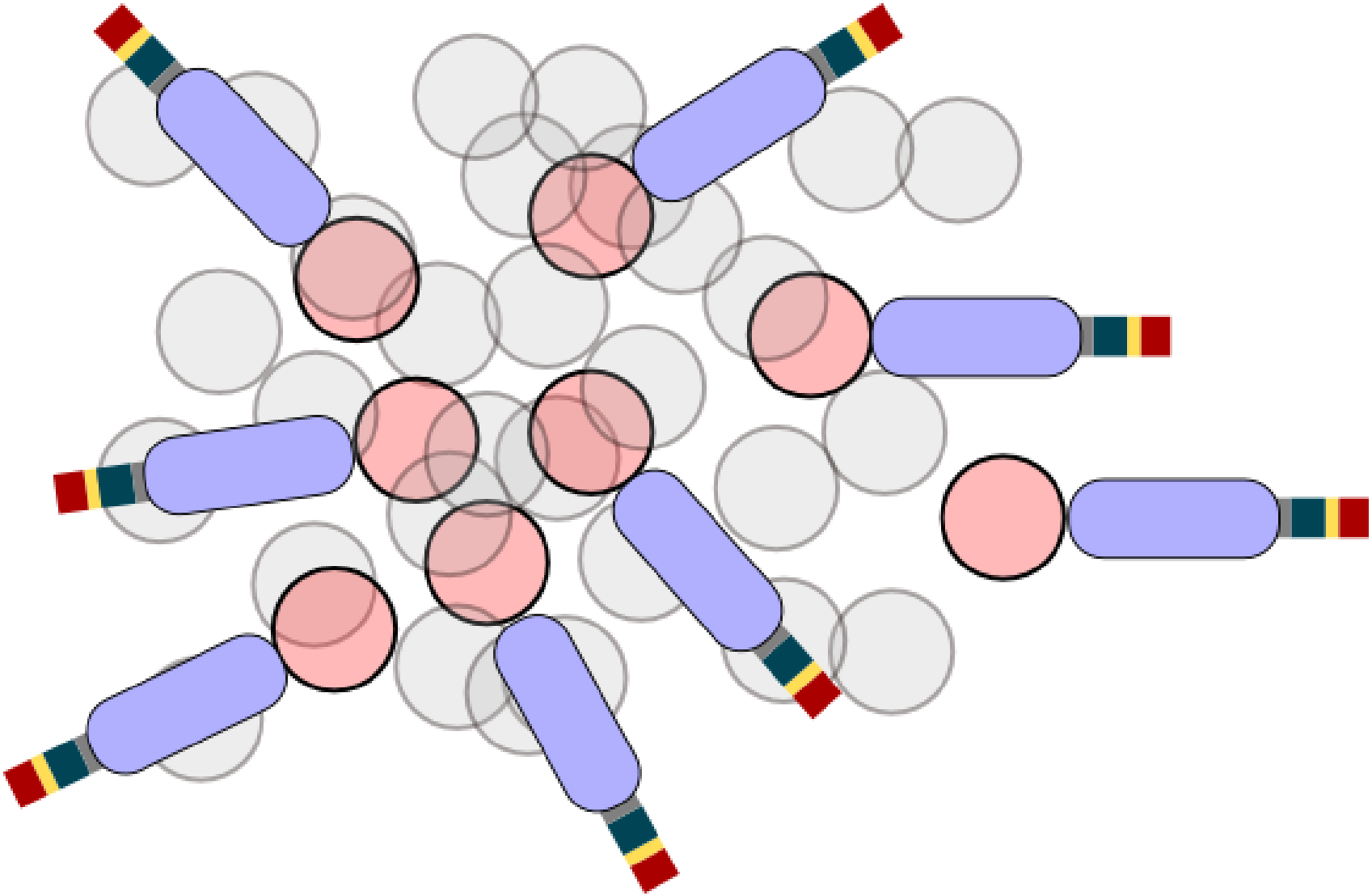
Tandem Affinity Purification



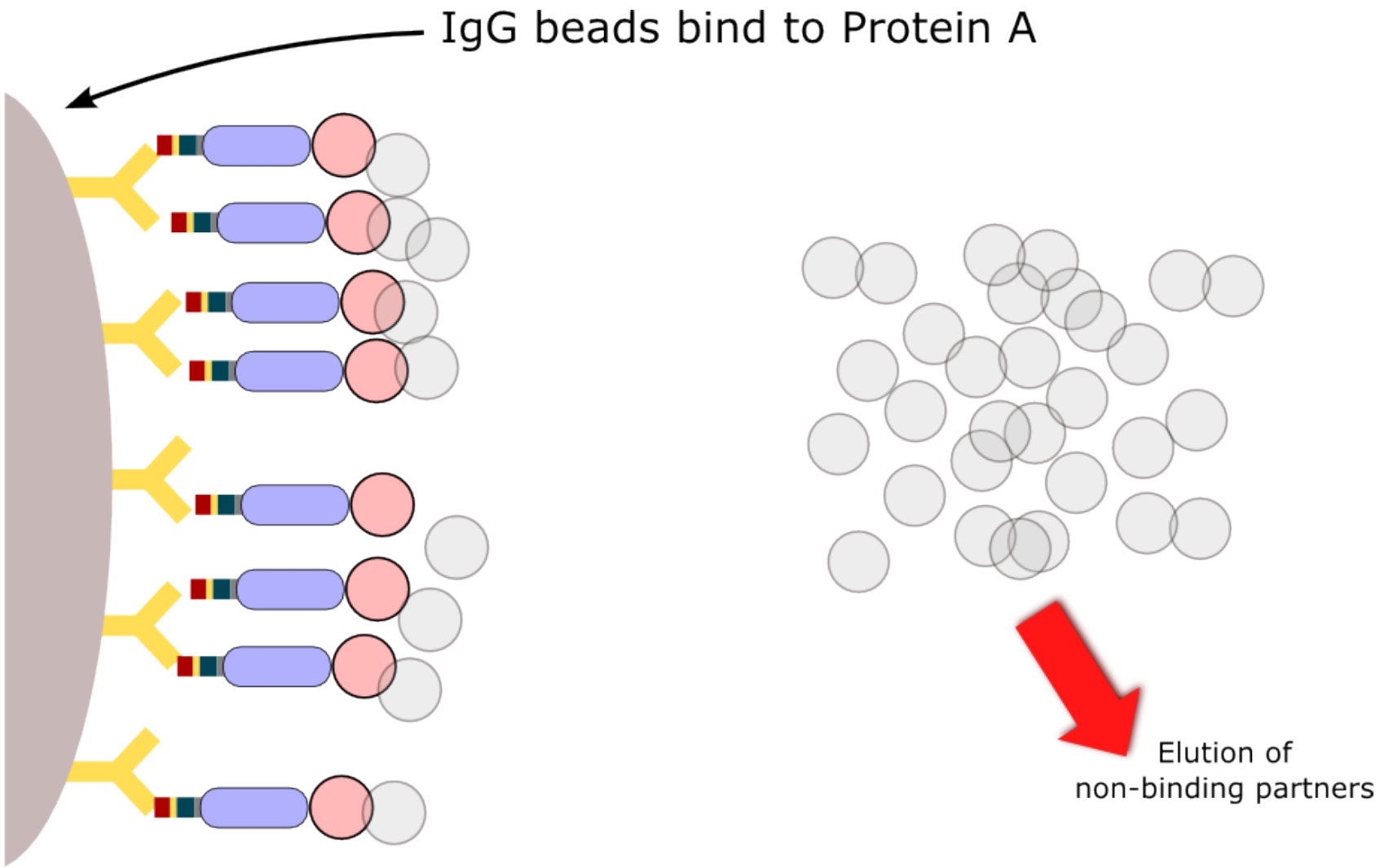


Express tagged protein
in cells with binding targets

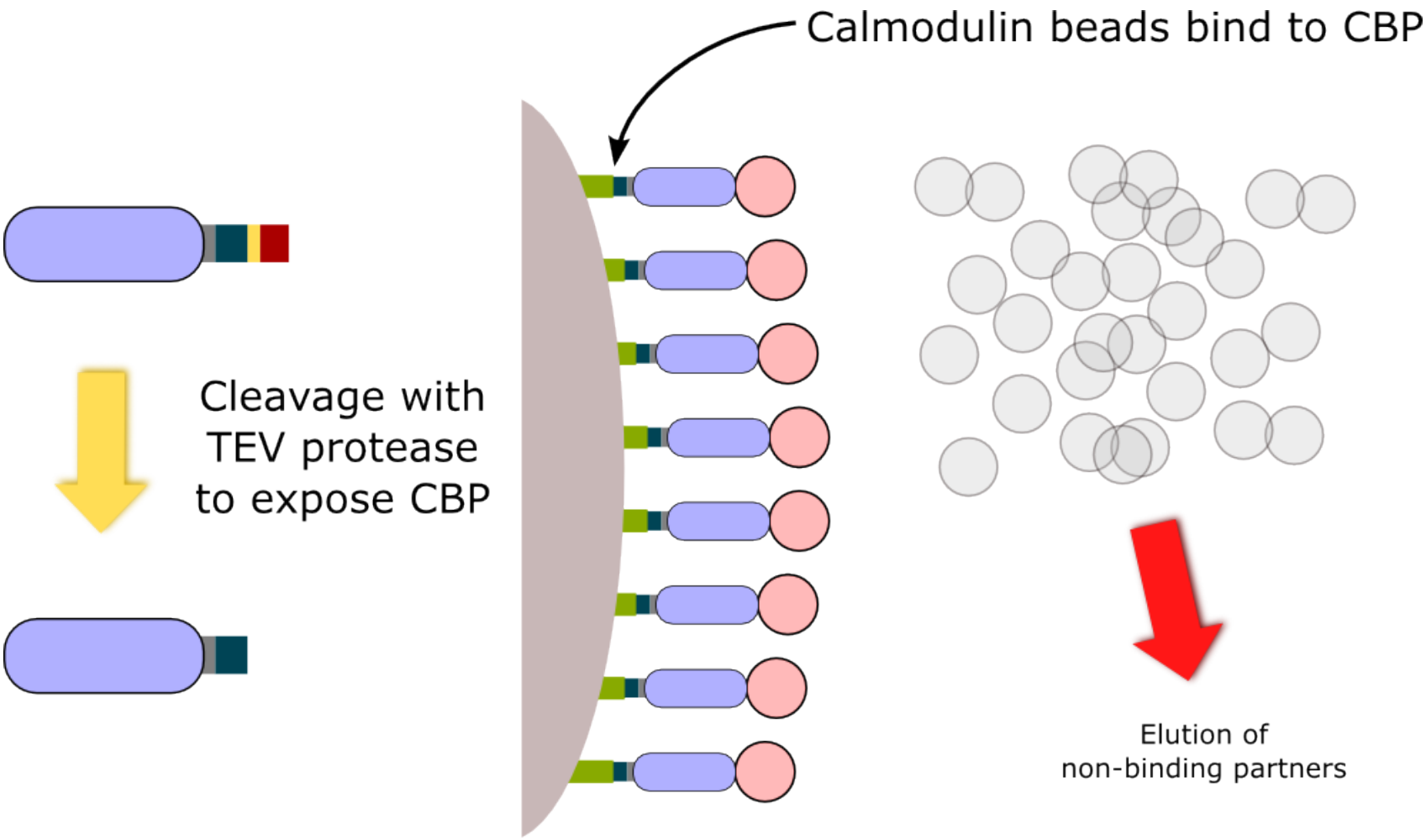
Tandem Affinity Purification (cont'd)



Tandem Affinity Purification (cont'd)



Tandem Affinity Purification (*fin.*)



The results of these methods

These methods are powerful

Given	A bait protein (or query gene) of interest
Determine	A list of binding partners (or genetic interactors) under experimental conditions

They are also limited

- No guarantee to get all the binding partners
- No guarantee that all the binding is correct
- For any individual interaction, a good test is to invert the system (use the finding as a new bait, and see if you can capture the original bait)

Topic

Molecular profiling

Organising genes into functional groups

Detecting interactions between genes and proteins

Representing genome wide data in networks

Pathway and interaction databases

Roundup and next week

Graphs and networks

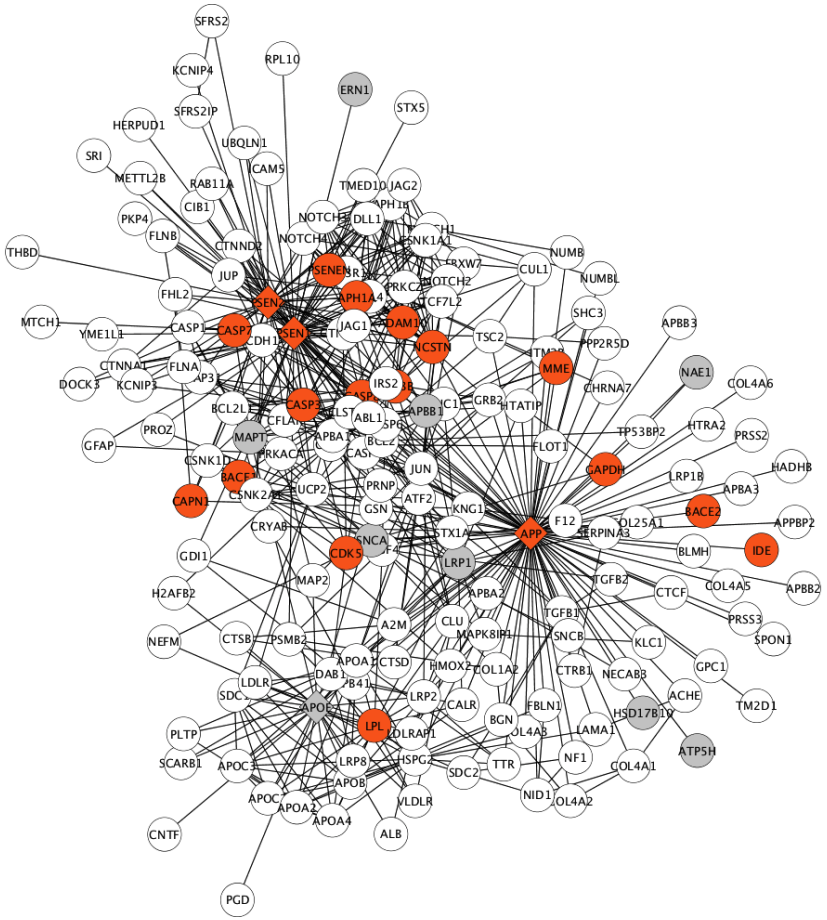
Properties of a graph

Must have

- **Nodes**
- **Edges**

May have

- Weights
- Labels
- Directions



Why represent genome wide data as networks?

- The same representation can be used for very different types of data!
  - Protein interactions
  - protein-compound interaction data
  - Facebook friends
  - Twitter trends
- *Guilt by association* recognizes that genes are similar. The edges of gene networks can recognize when gene **pairs** are similar
- Many tools are available to operate on networks

What do we look for in networks?

Hubs genes that are connected to a *lot* of others

Clusters or cliques groups of genes that are connected to each other

Functional overlap How to regions of a network, or subnetworks, relate to functional annotation?

Pathway overlap How do regions or subnetworks relate to pathways?

Cytoscape: Network visualization and analysis

- Load any interaction data
- Add annotations from disparate data sources
- Look at connectivity in the network

Topic

Molecular profiling

Organising genes into functional groups

Detecting interactions between genes and proteins

Representing genome wide data in networks

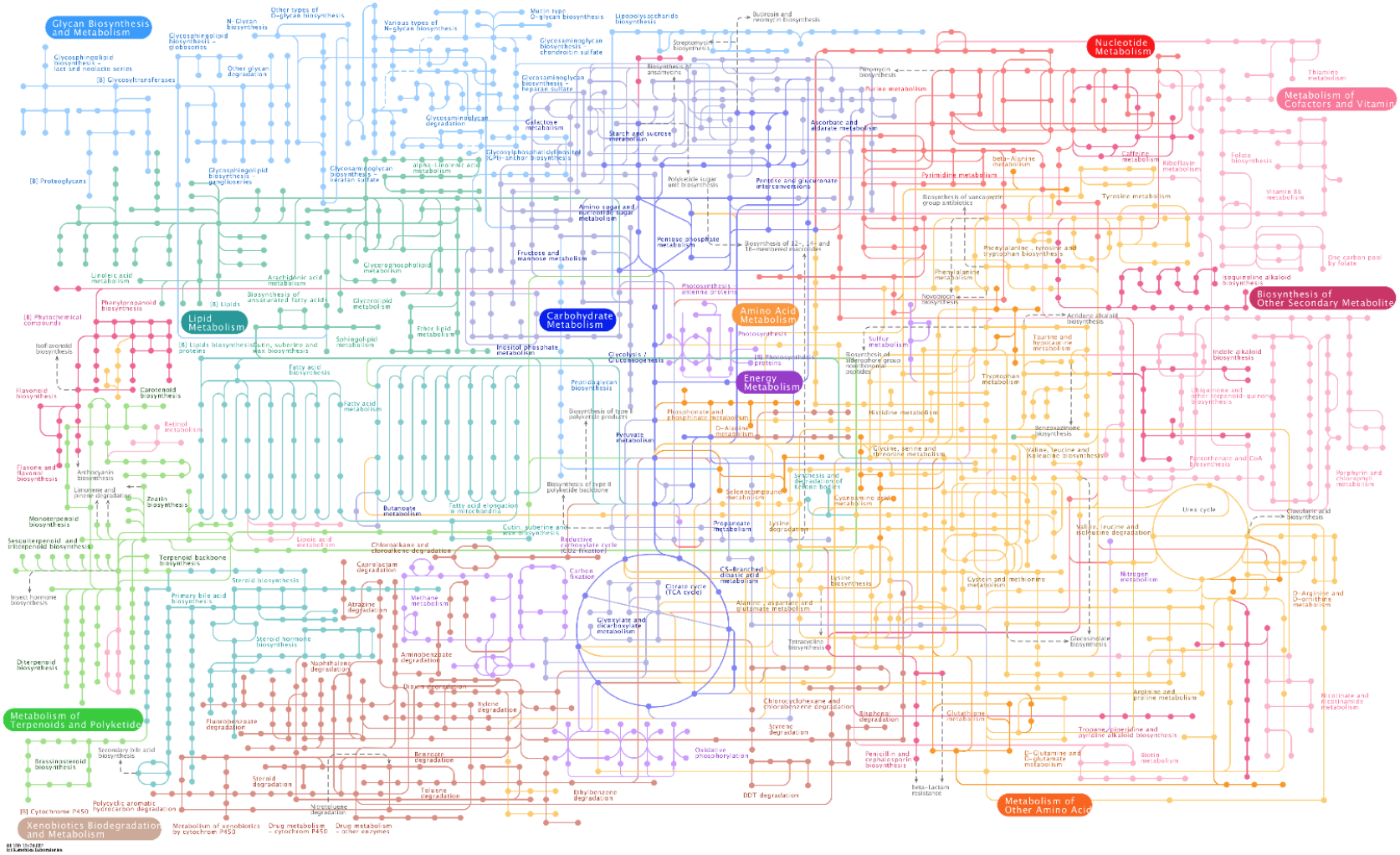
Pathway and interaction databases

Roundup and next week

STRING - both known and *predicted* protein-protein interactions

- Unlike IntAct, STRING, includes predictions of interactions
- A companion database, STITCH, includes interactions between proteins and chemicals
- Another companion datababase, eggNOG, uses functional associations to classify genes into groups of orthologs

Pathway databases: KEGG



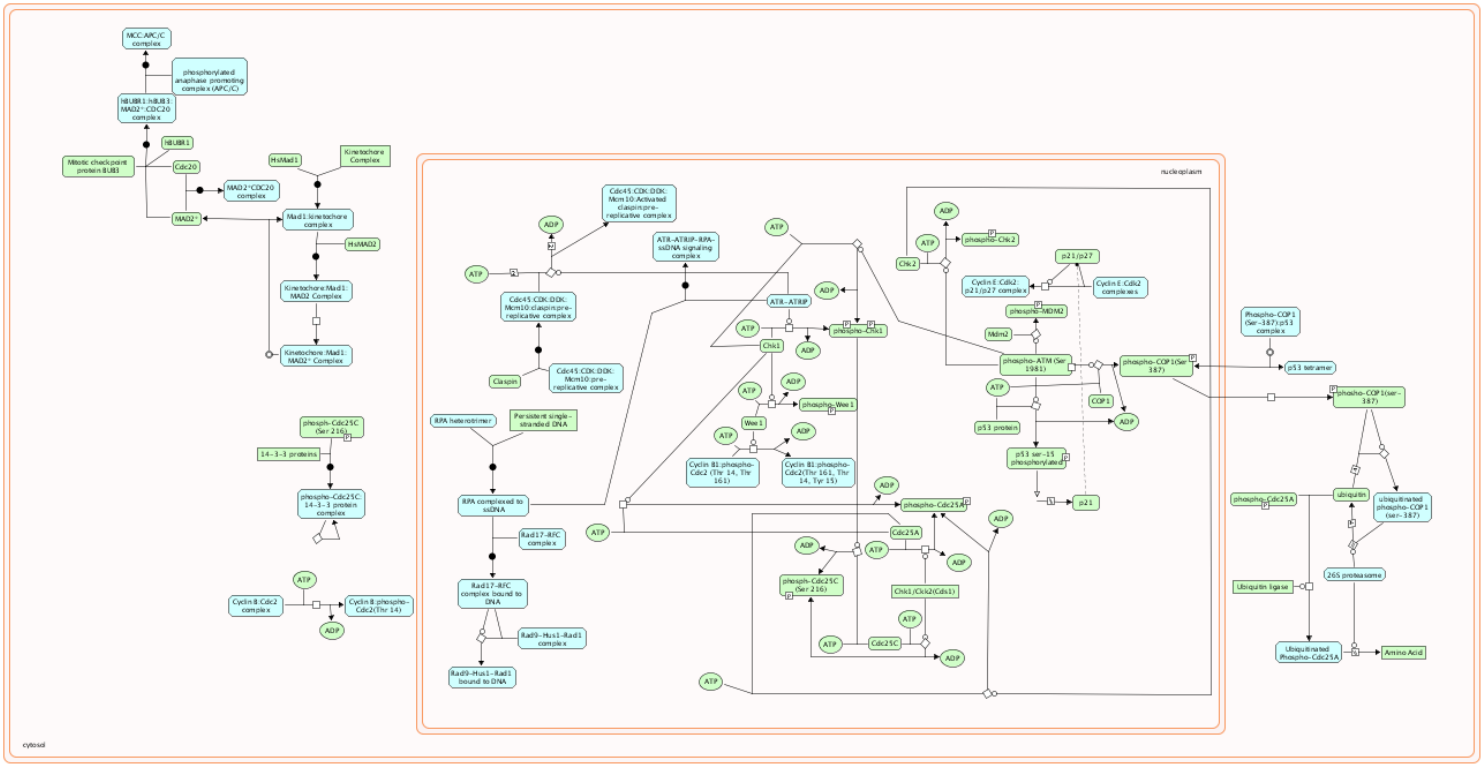
KEGG's Pathway Database

- The most well-known part of KEGG is the pathway database
- A graphical diagrams representing molecular interaction data *with* cellular processes
- Contrast this with the Gene Ontology
- Explore at `http://www.genome.jp/kegg/pathway.html`

Pathway database: Reactome

- Reactome is a US funded pathway database resource
- All open source, accessible data
- Peer reviewed, curated
- Nice visual representations
- Excellent tutorial published in proteomics (Haw et al. 2011)

Reactome example



Cell cycle checkpoint in reactome

Topic

Molecular profiling

Organising genes into functional groups

Detecting interactions between genes and proteins

Representing genome wide data in networks

Pathway and interaction databases

Roundup and next week

What we have learned I

- Whole genome studies create a surfeit of data
- The tools of bioinformatics provide bridges between different forms of large scale experimental data
- Using categories of gene function can help to interpret large scale studies
- "Guilt by association" can be used to infer functional involvement in large scale genomic studies
- High throughput interaction data are generated from a variety of strategies
- Interaction data is often represented as a graph or network
- Network analysis tools allow us to explore the structure of networks, and overlay additional information on top of them
- Pathway databases provide more specialized, but biologically comprehensible, representations of interaction data

Next week

- Next week we will look at where bioinformatics is going
- We will discuss personal genomics
- We will look at my genomic data from 23 And Me

Bibliography I

Haw, Robin et al. (2011). “Reactome pathway analysis to enrich biological discovery in proteomics data sets.” In: *Proteomics* 11.18, pp. 3598–613. DOI: 10.1002/pmic.201100066 (cit. on p. 36).