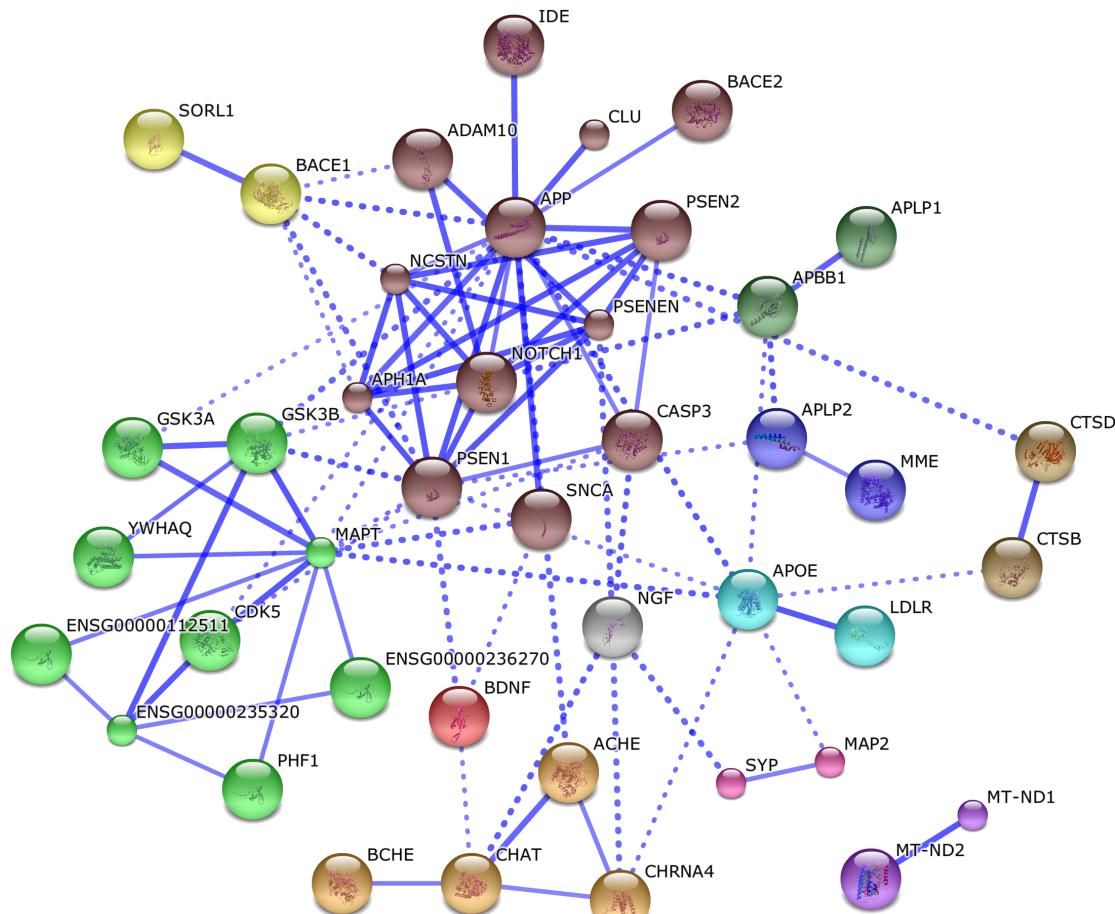


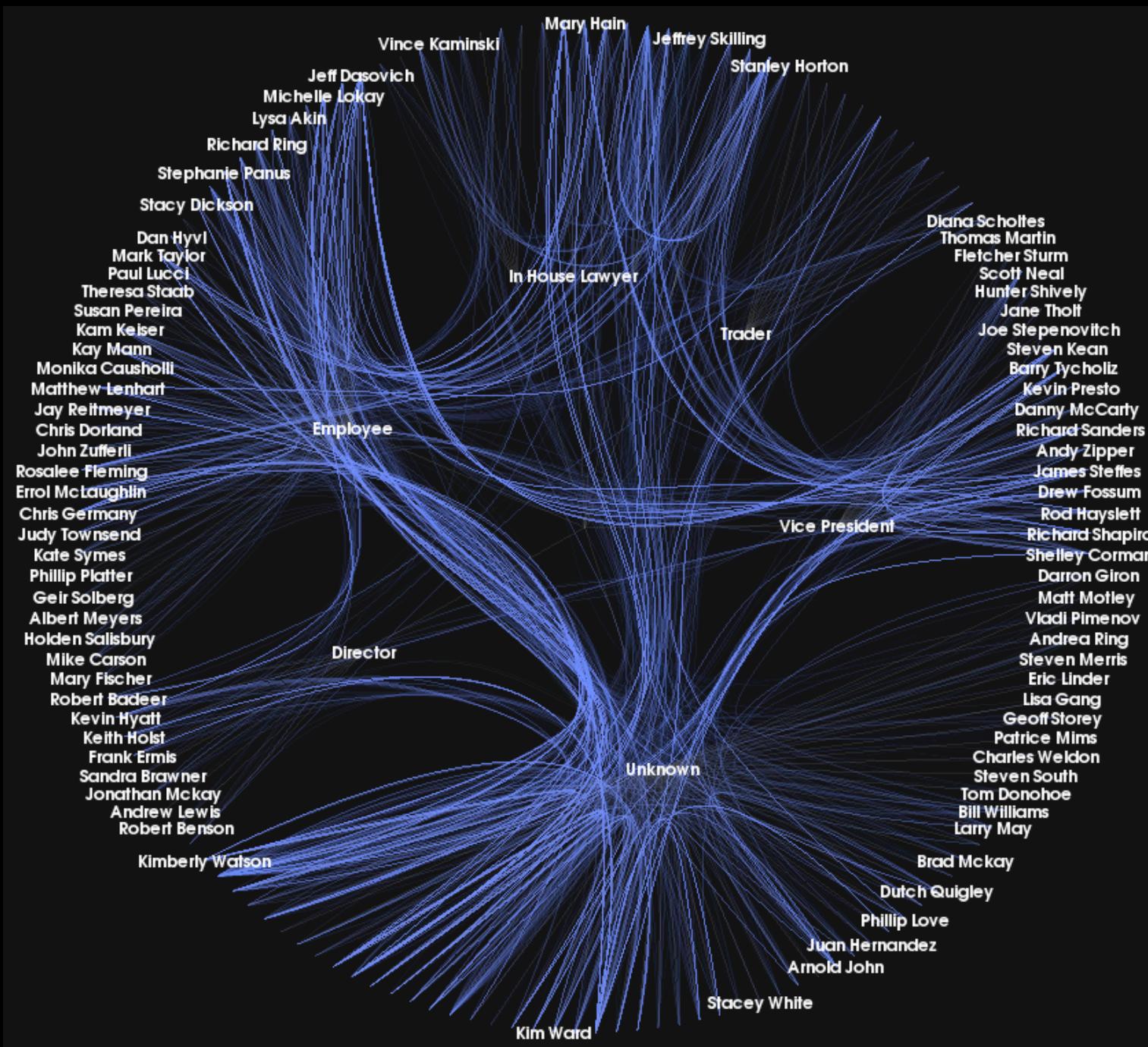
STRING: Protein association networks



interaction networks

association networks

guilt by association



protein networks

STRING

9.6 million proteins

common foundation

Exercise 1

Go to <http://string-db.org/>

Query for human insulin receptor (INSR)
using the *search by name* functionality

Make sure you are in *evidence* view
(check the buttons below the network)

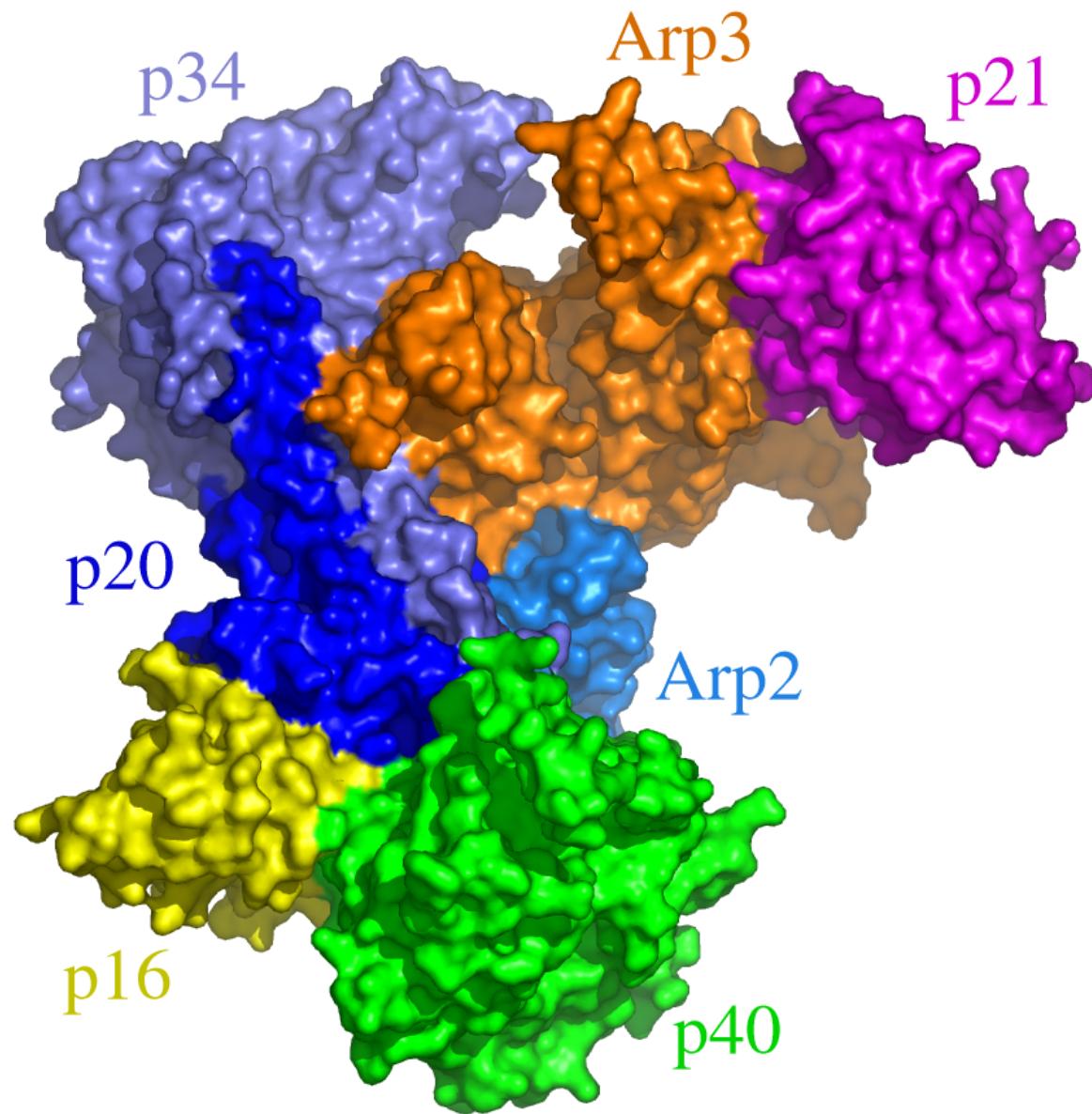
Why are there multiple lines connecting the same to two proteins?

curated knowledge

(what we know)

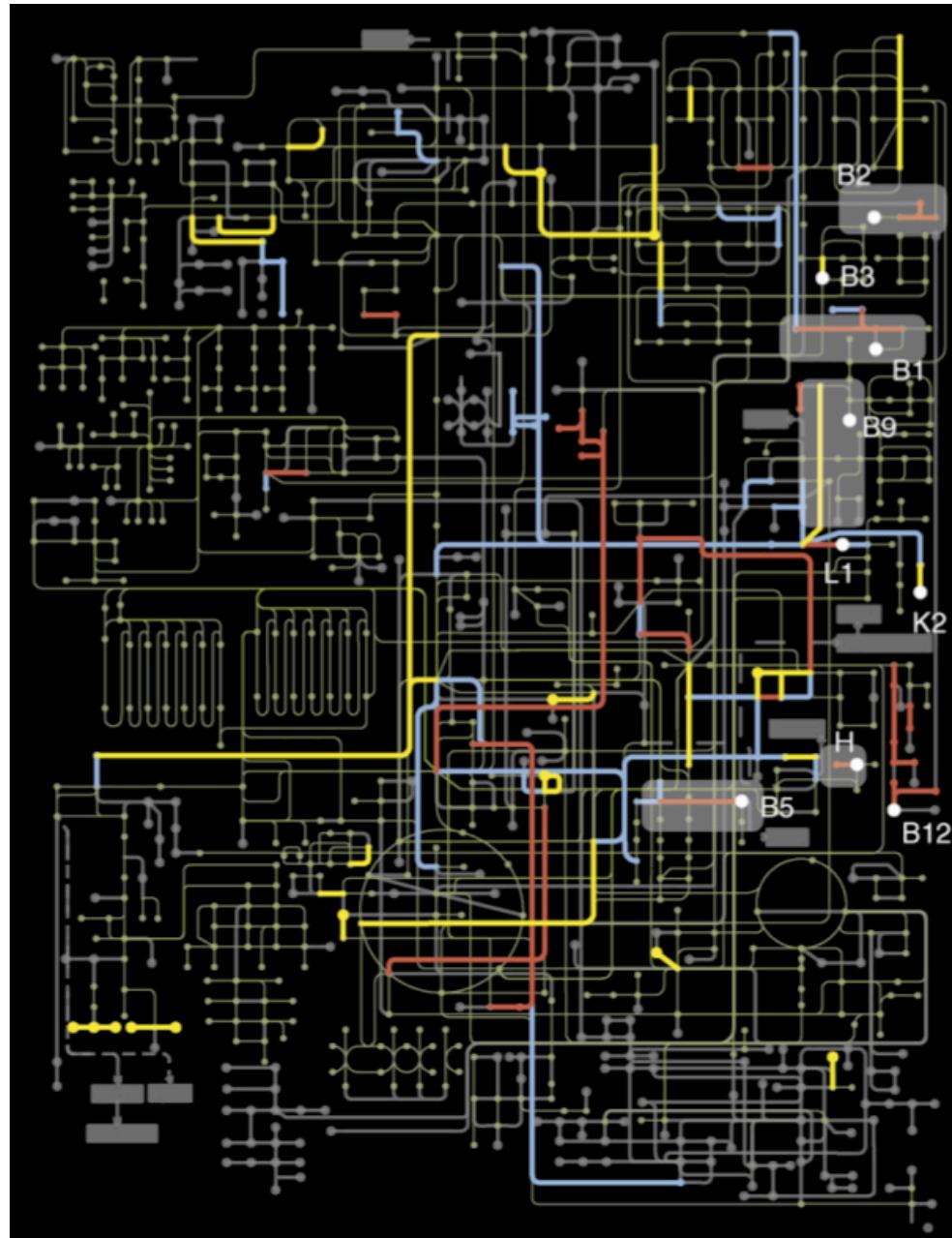
protein complexes

3D structures



pathways

metabolic pathways



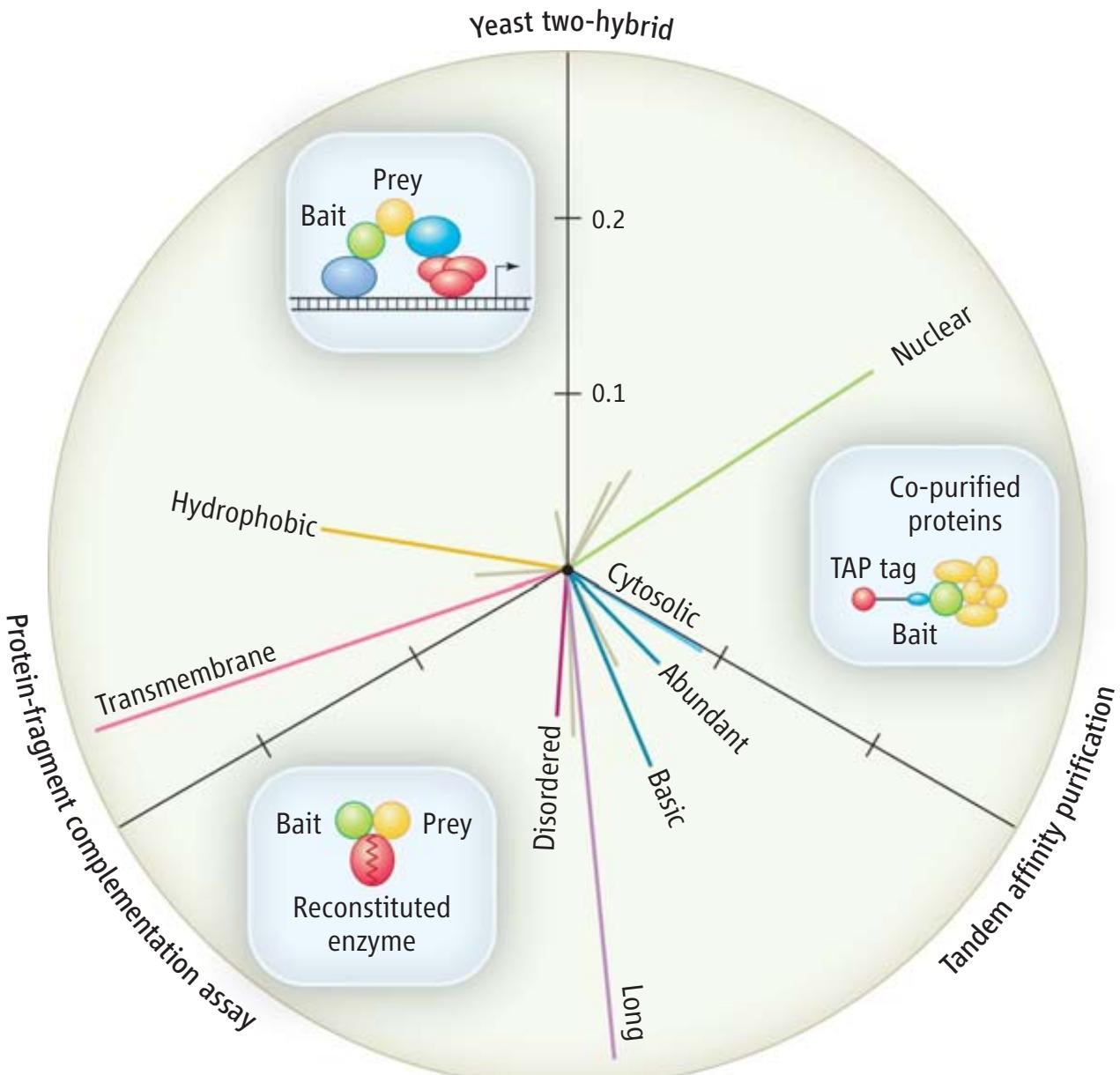
signaling pathways

very incomplete

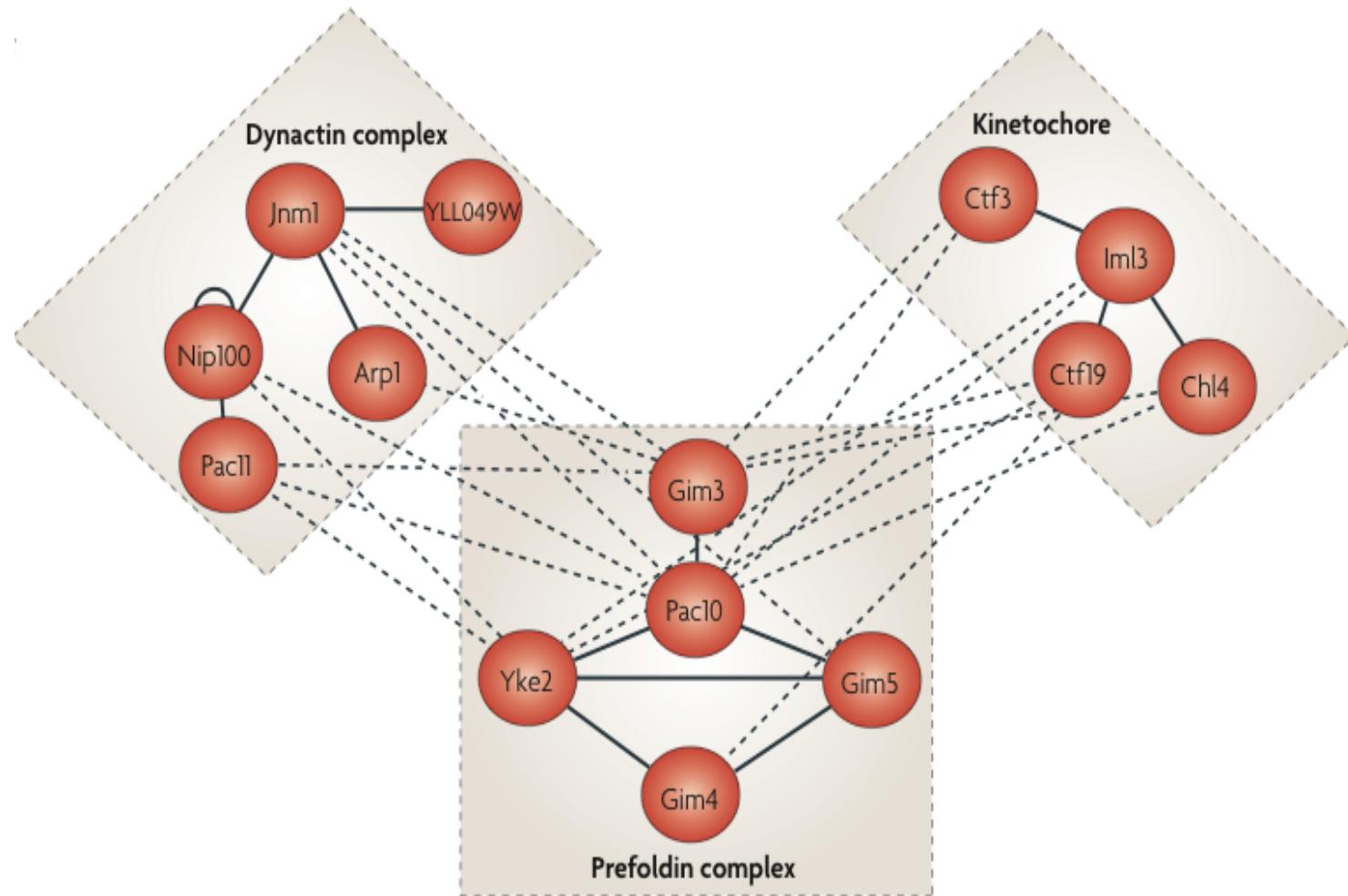
experimental data

(what we measured)

physical interactions



genetic interactions



gene coexpression

microarrays

RNAseq

Exercise 2

(Continue from where exercise 1 ended)

Which types of evidence support the interaction between INSR and IRS1?

Click on the interaction to view the popup, which has buttons linking to full details

Which types of experimental assays support the INSR–IRS1 interaction?

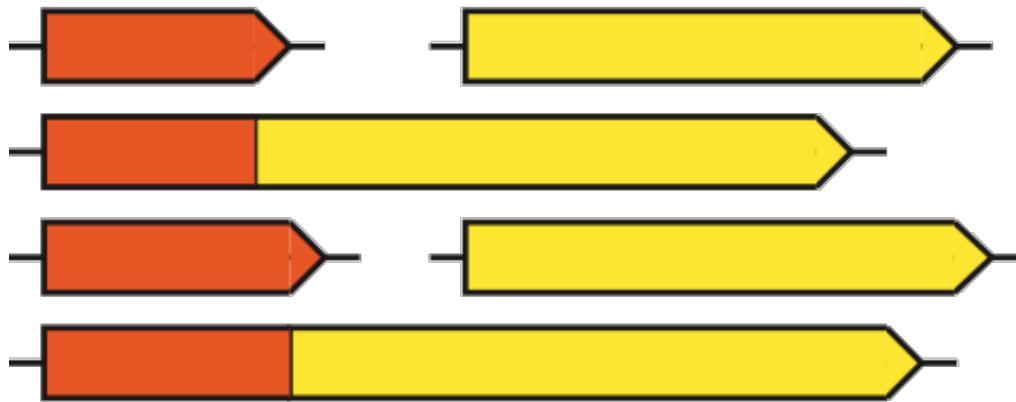
predictions

(what we infer)

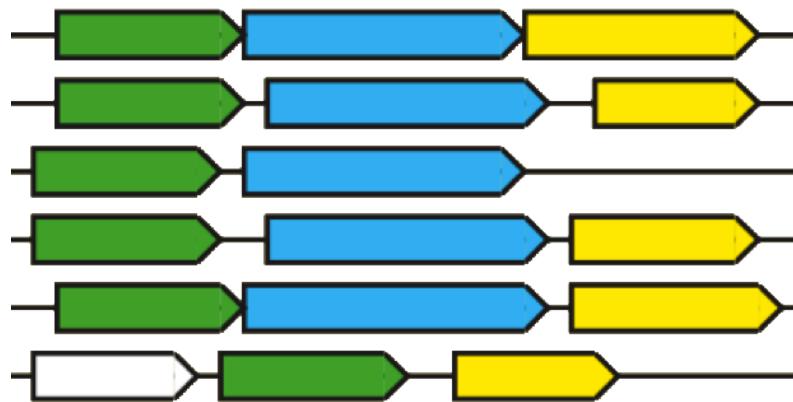
genomic context

evolution

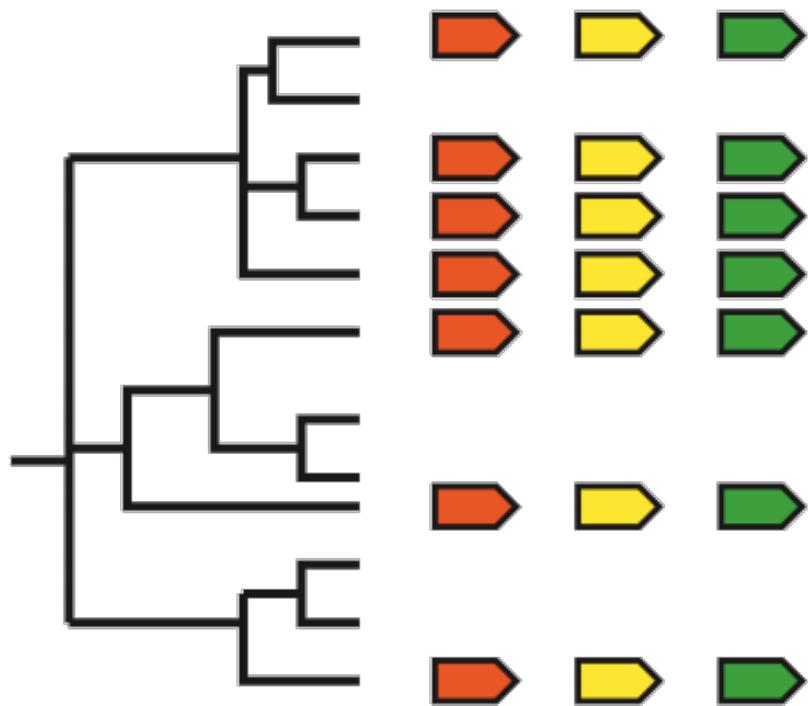
gene fusion



gene neighborhood



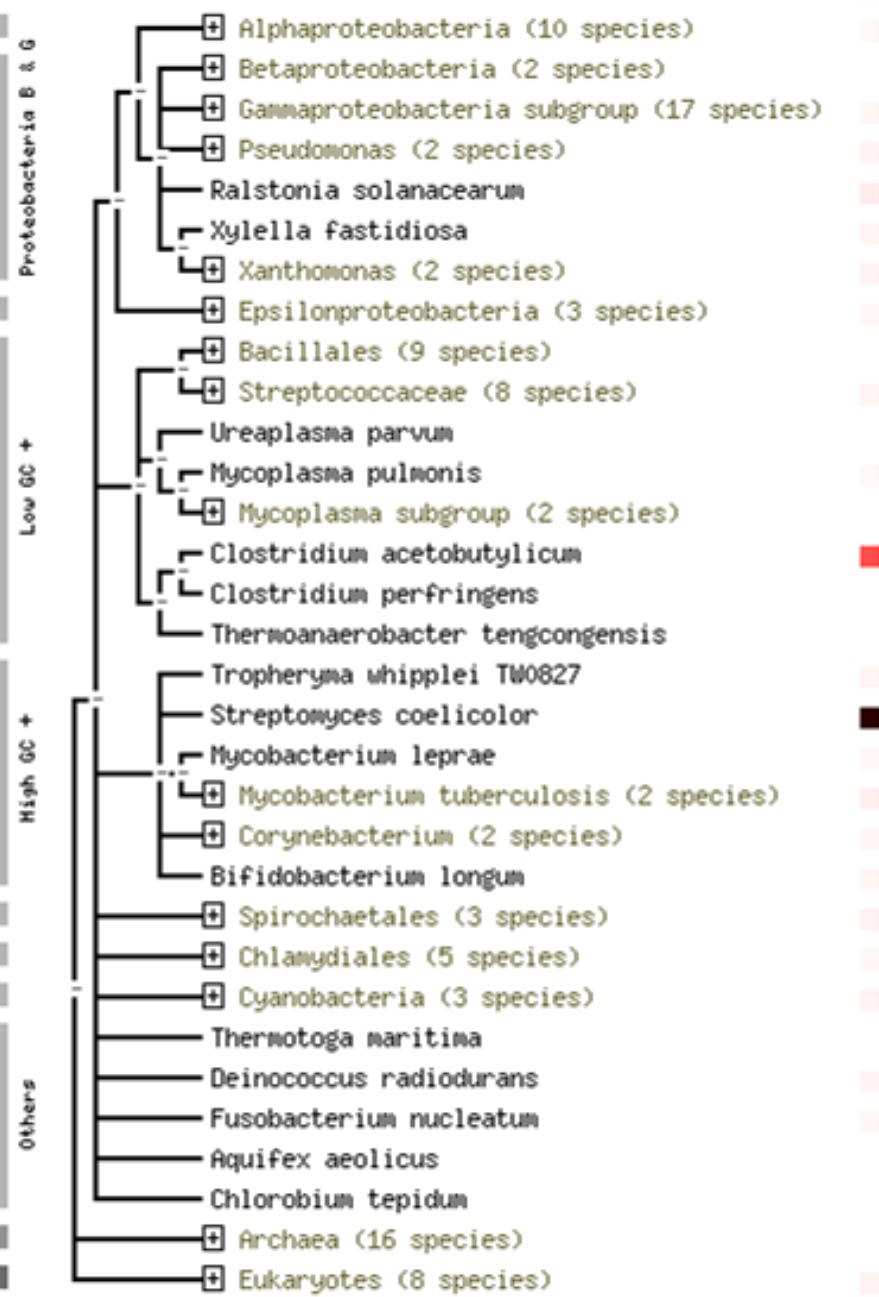
phylogenetic profiles

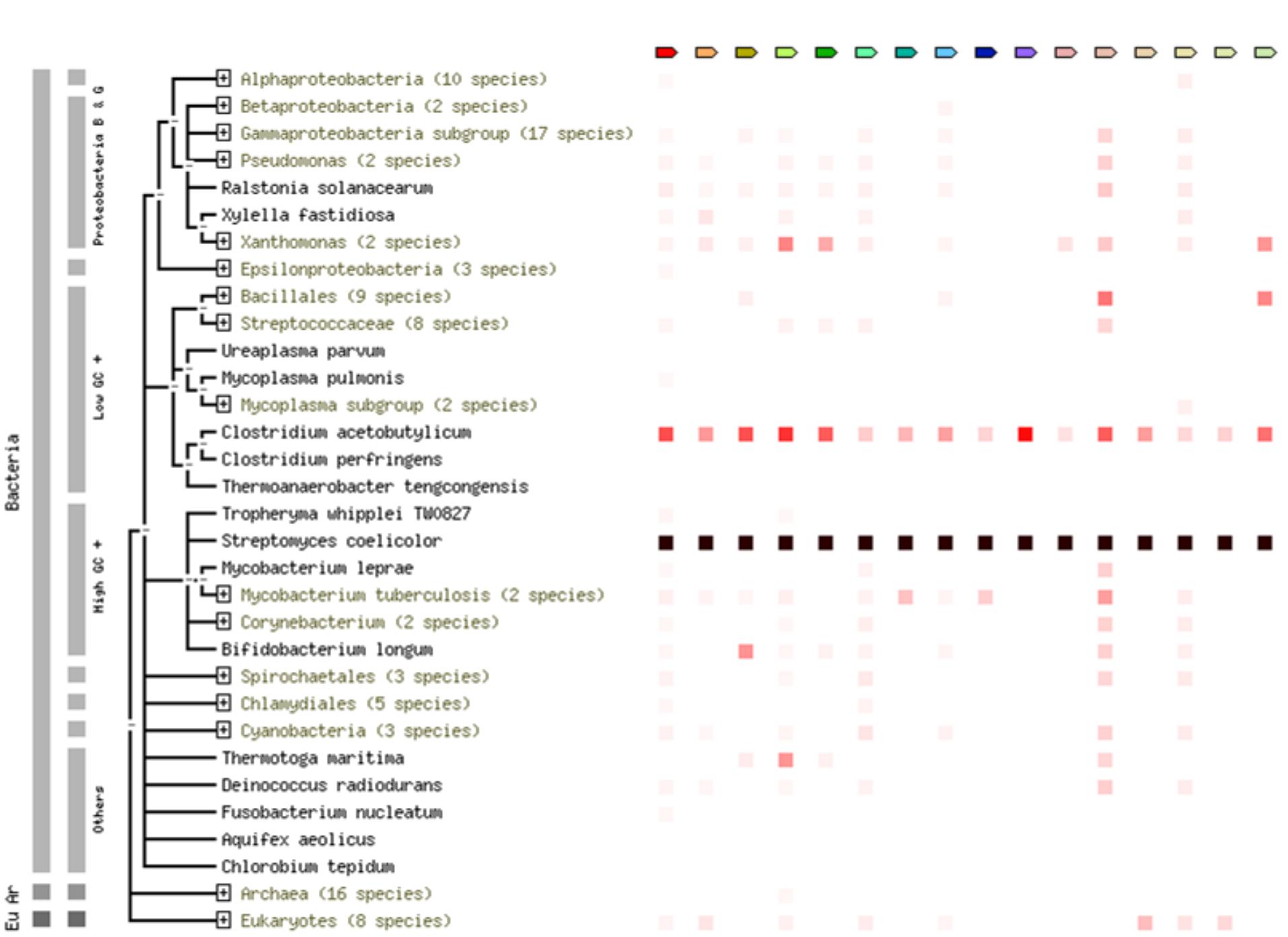


a real example

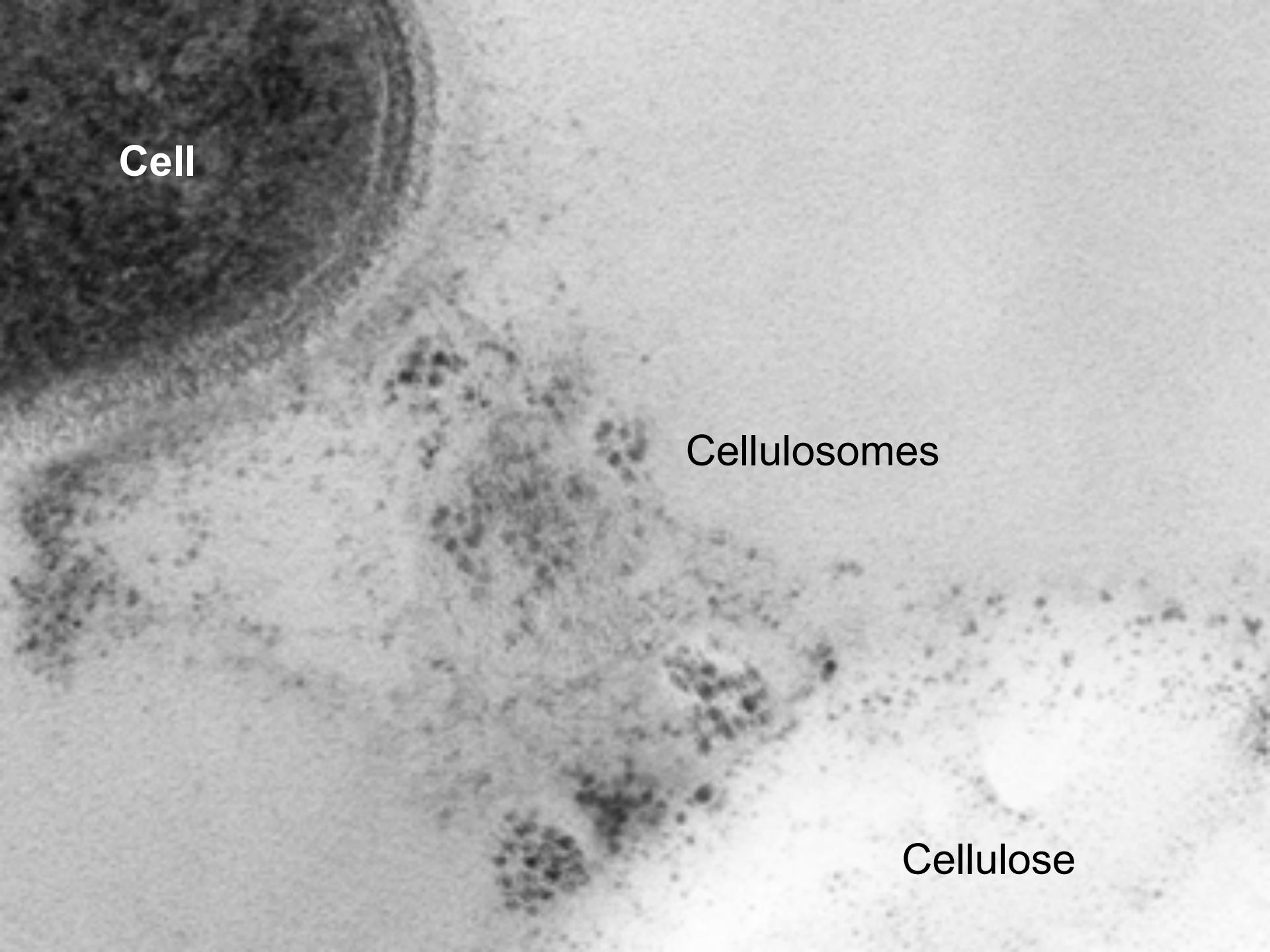
Bacteria

Eu Ar





- ▶ Putative secreted cellulase (973 aa)
- ▶ Putative secreted beta-galactosidase (933 aa)
- ▶ Putative secreted arabinosidase (824 aa)
- ▶ Putative secreted cellulase (890 aa)
- ▶ Secreted endoglucanase (747 aa)
- ▶ Putative secreted esterase (706 aa)
- ▶ Hypothetical protein SC04853 (136 aa)
- ▶ Putative secreted protease (781 aa)
- ▶ Hypothetical protein SC06611 (186 aa)
- ▶ Hypothetical protein SC00396 (424 aa)
- ▶ Putative conserved DNA-binding protein (290 aa)
- ▶ Putative transcriptional regulator (206 aa)
- ▶ Hypothetical protein SC00964 (143 aa)
- ▶ Putative secreted esterase (505 aa)
- ▶ Putative secreted protein (213 aa)
- ▶ Putative ribonuclease inhibitor (89 aa)

A black and white transmission electron micrograph showing a cross-section of a plant cell wall. Dark, electron-dense regions represent cellulose microfibrils, which are organized into larger, irregularly shaped structures labeled 'Cellulosomes'. A large, roughly circular, dark area in the upper left is labeled 'Cell'.

Cell

Cellulosomes

Cellulose

complications

many databases

different formats

different identifiers

variable quality

not comparable

not same species

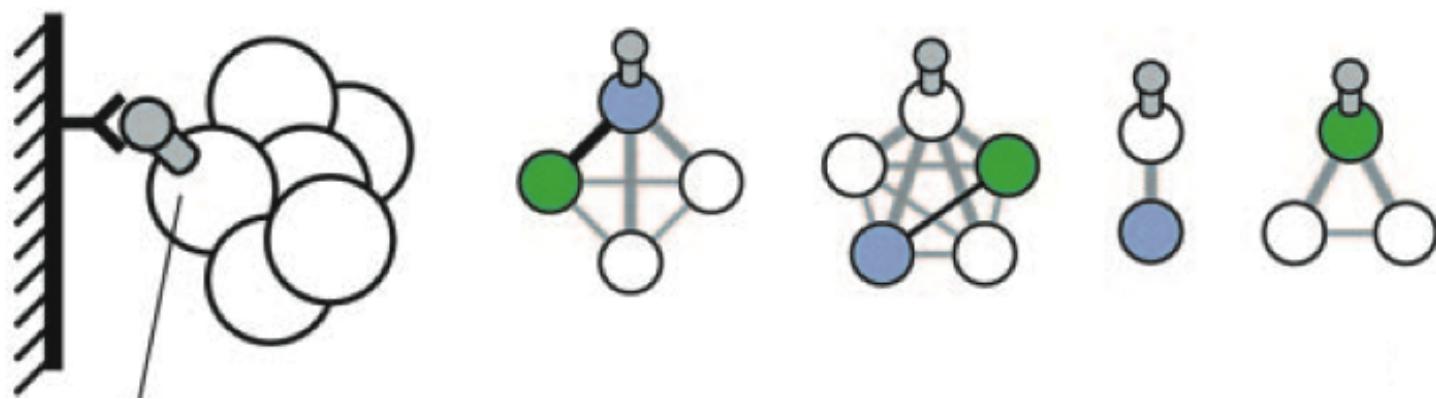
hard work

parsers

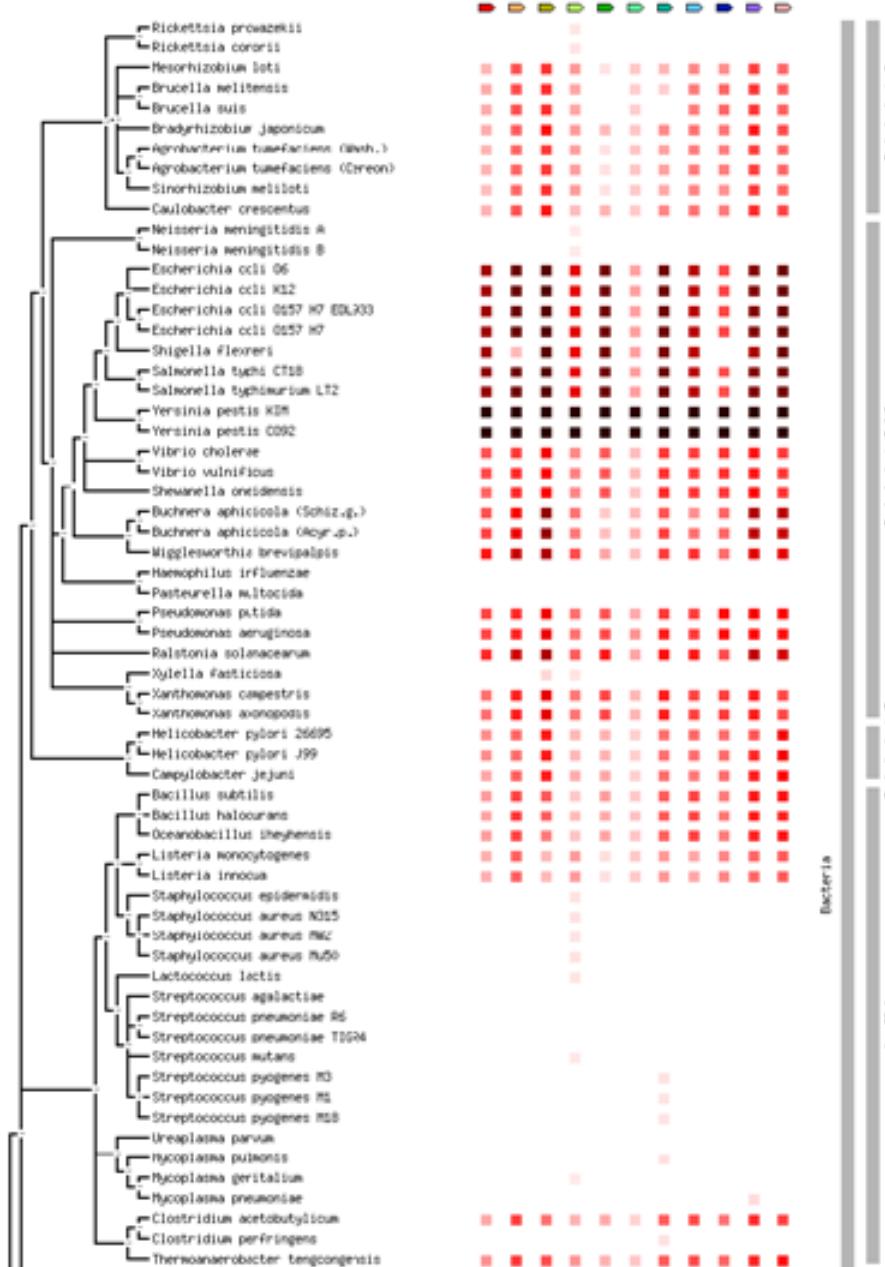
mapping files

quality scores

affinity purification

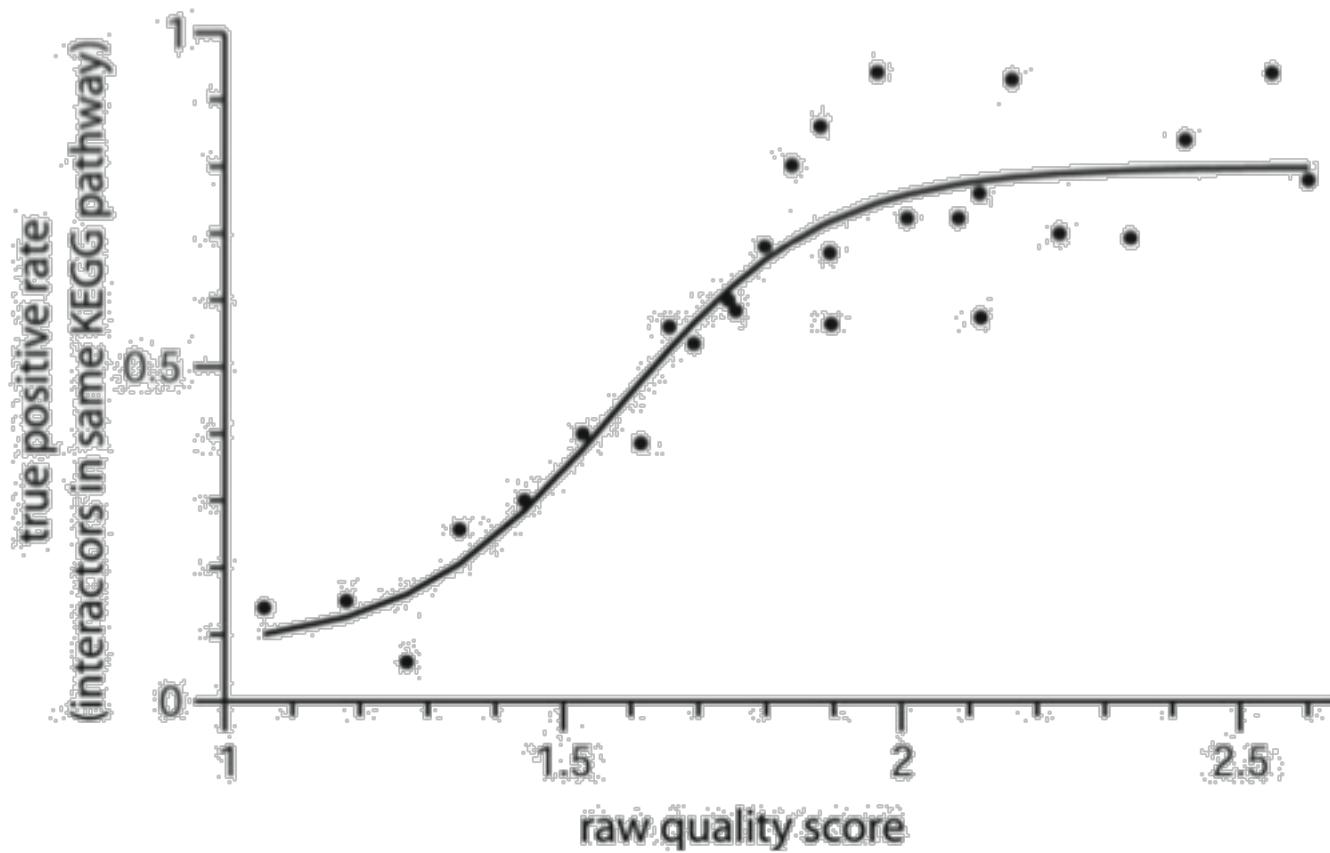


phylogenetic profiles



score calibration

gold standard



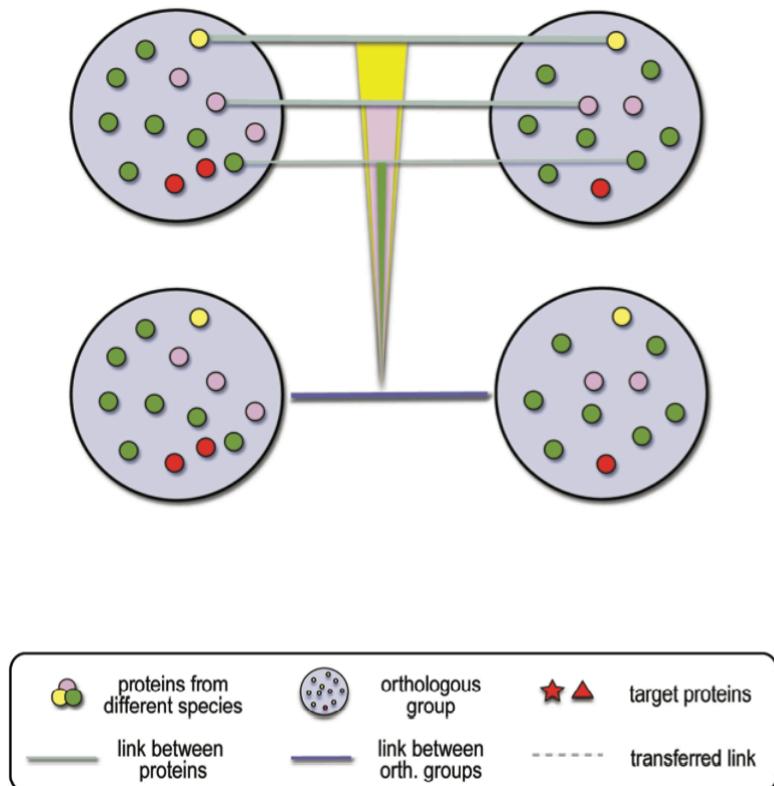
implicit weighting by quality

common scale

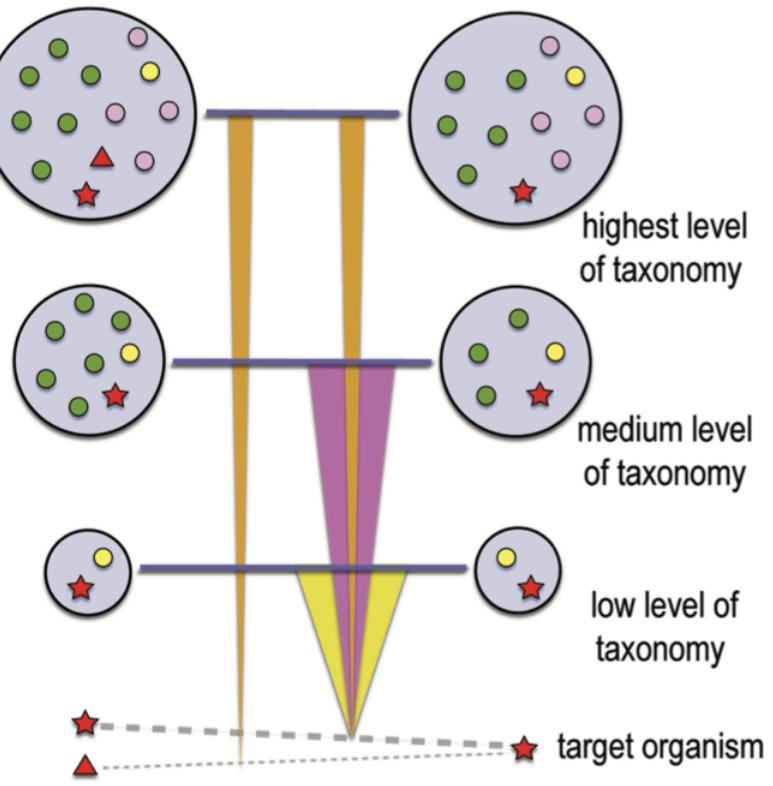
homology-based transfer

orthologous groups

Step 1: Combining individual protein-protein links into links between orthologous groups.



Step 2: Transferring orthologous group links back to the protein level



missing most of the data

Exercise 3

(Continue from where exercise 2 ended)

Change the network to the *confidence* view

Change the confidence cutoff to 0.9; any changes in proteins or interactions shown?

Turn off all but *experiments*; what changes?

Increase the number of *interactors* shown to 50; how many proteins do you get? Why?

text mining

>10 km



too much to read

exponential growth

~40 seconds per paper

computer

as smart as a dog

teach it specific tricks

What we say to dogs

Okay, Ginger! I've had it!
You stay out of the garbage!
Understand, Ginger? Stay out
of the garbage, or else!



what they hear



named entity recognition

comprehensive lexicon

cyclin dependent kinase 1

CDC2

orthographic variation

expansion rules

prefixes and suffixes

CDC2

hCdc2

flexible matching

spaces and hyphens

cyclin dependent kinase 1

cyclin-dependent kinase 1

“black list”

SDS

information extraction

co-mentioning

counting

within documents

within paragraphs

within sentences

scoring scheme

$$C_{ij} = \sum_{k=1}^n \delta_{dijk} w_d + \delta_{pijk} w_p + \delta_{sijk} w_s$$

$$S_{ij} = C_{ij}^\alpha \left(\frac{C_{ij} C_{..}}{C_{i..} C_{..j}} \right)^{1-\alpha}$$

score calibration

NLP

Natural Language Processing

part-of-speech tagging

what you learned in school

pronoun

pronoun

verb

preposition

noun

semantic tagging

grammatical analysis

Gene and protein names

Cue words for entity recognition

Verbs for relation extraction

[nxexpr The expression of
[nxgene the cytochrome genes
[nxpg CYC1 and CYC7]]]
is controlled by
[nxpg HAP1]

type and direction

complex sentences

anaphoric references

it

summary

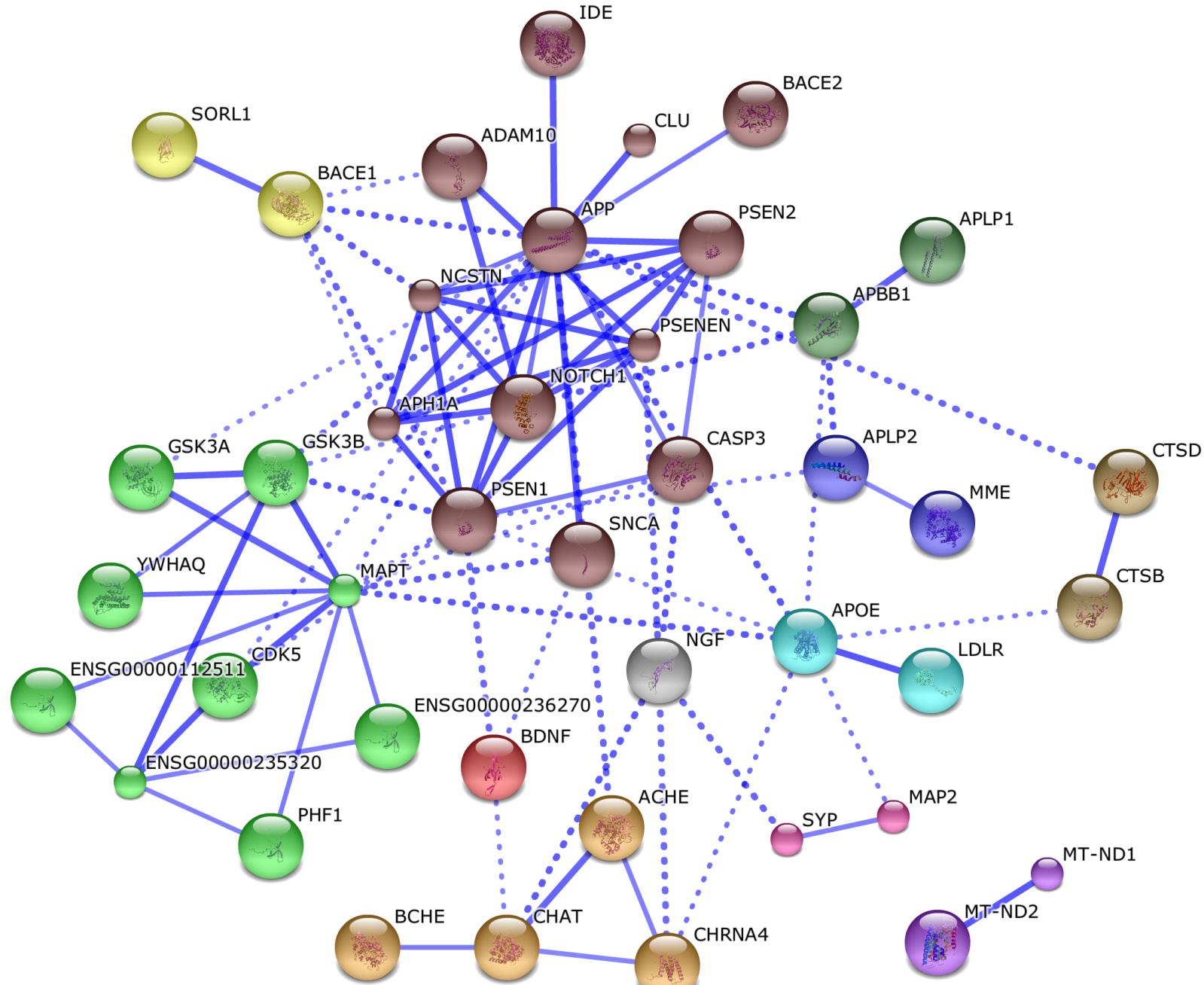
association networks

heterogeneous data

common identifiers

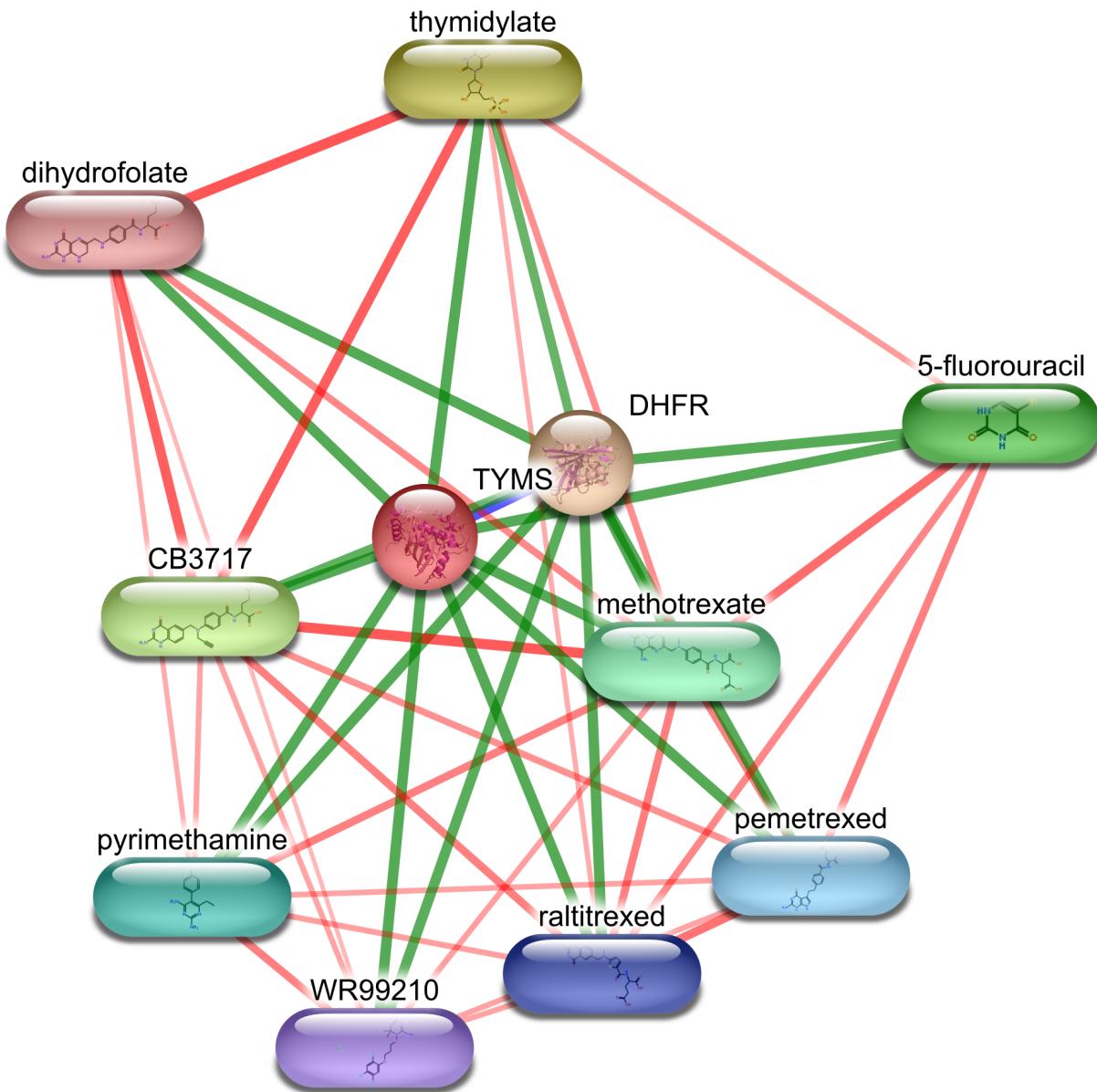
quality scores

protein networks



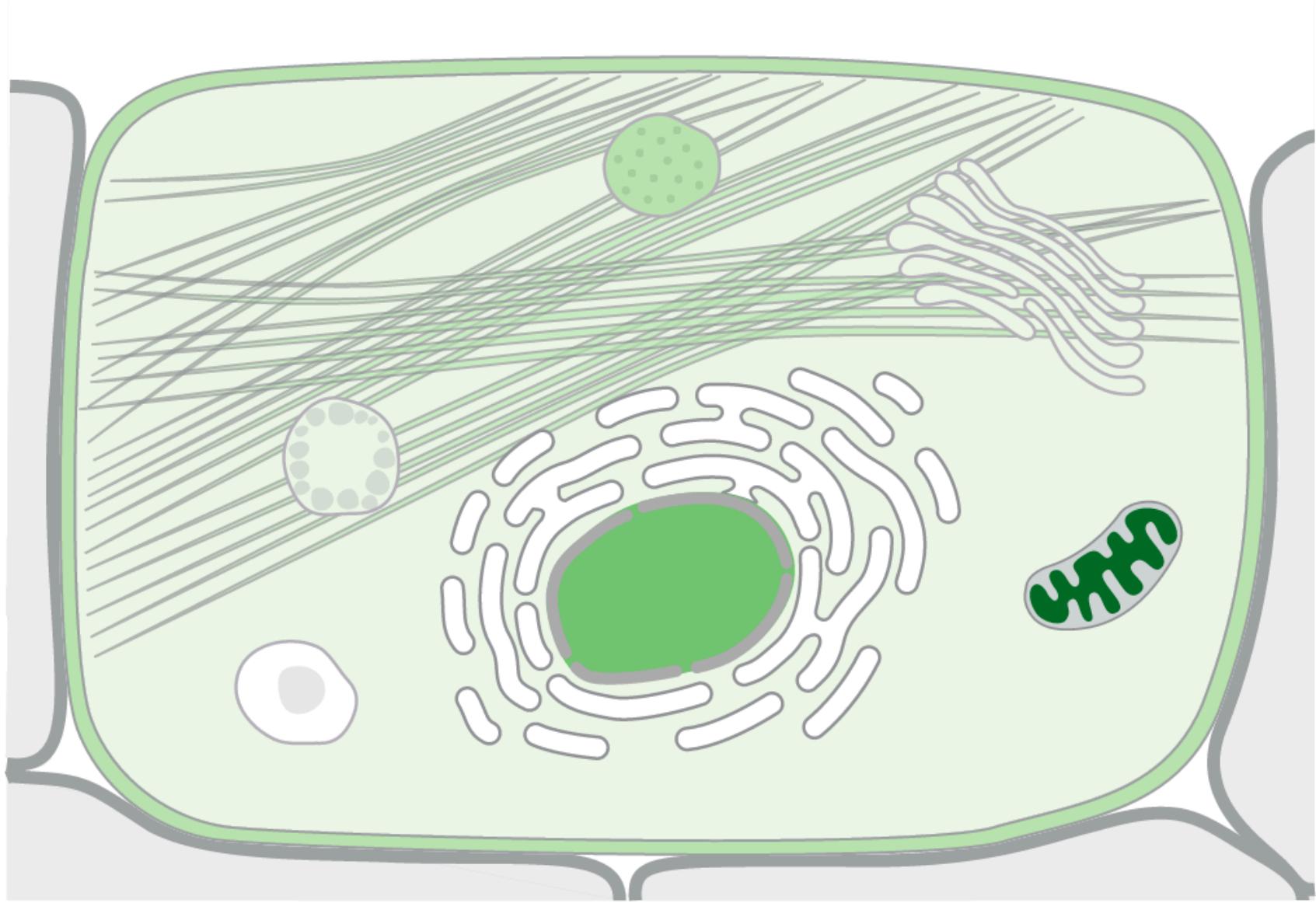
STITCH

chemical networks



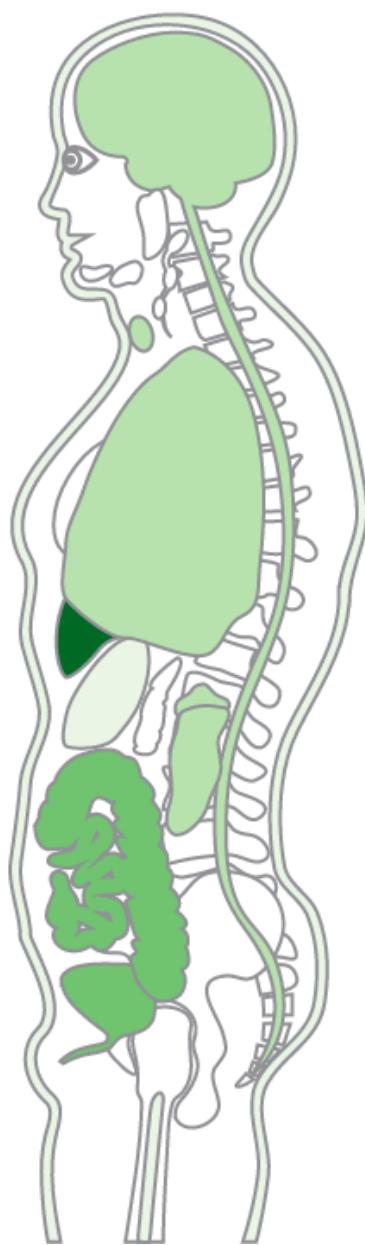
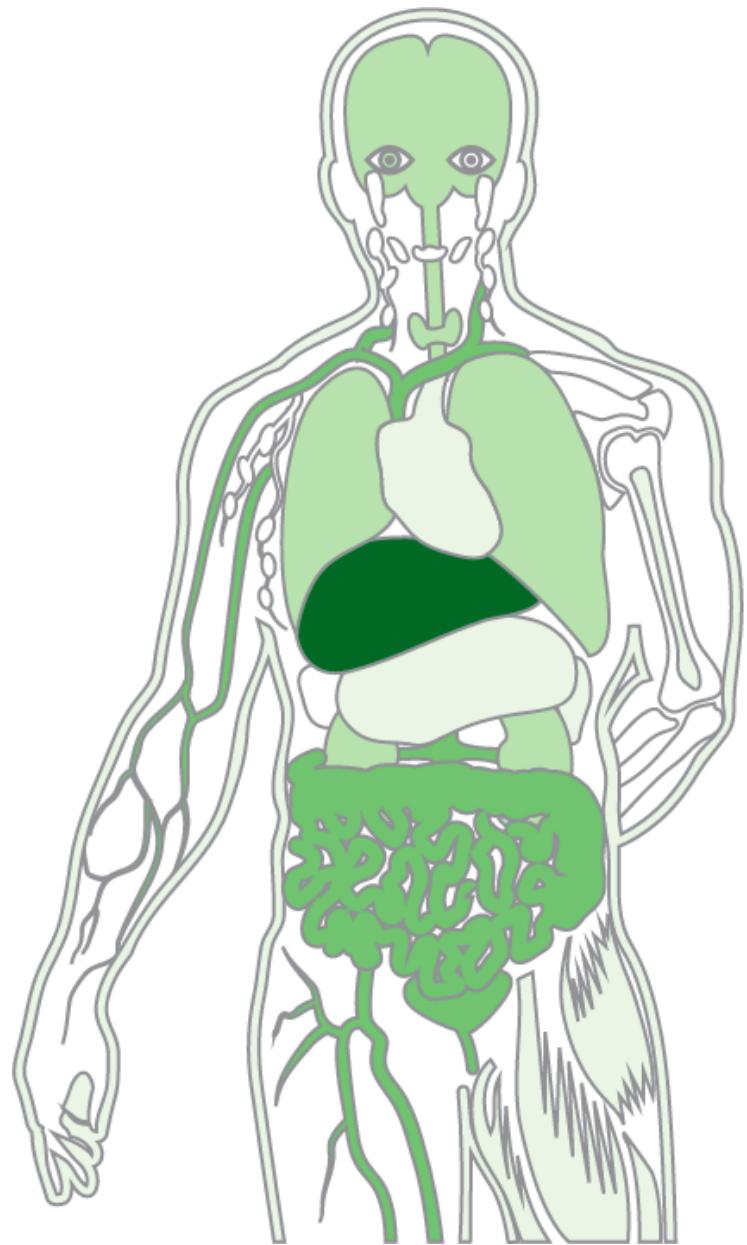
COMPARTMENTS

subcellular localization



TISSUES

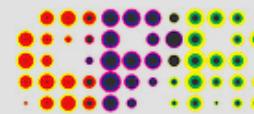
tissue expression



DISEASES

disease associations

DISEASES



Disease-gene associations mined from literature

[Search](#)[Downloads](#)[About](#)

Human genes for idiopathic pulmonary fibrosis

Idiopathic pulmonary fibrosis [DOID:0050156]

A idiopathic interstitial pneumonia which is a distinctive type of chronic fibrosing interstitial pneumonia with thick scarring in the lung creating a honeycomb appearance. The main symptoms start insidiously as shortness of breath on exertion, cough, and diminished stamina. Other common complaints include weight loss and fatigue. The level of oxygen in the blood decreases, and the skin may take on a bluish tinge (called cyanosis) and the ends of the fingers may become thick or club-shape. In most people, symptoms worsen over a period ranging from about 6 months to several years.

Synonyms: idiopathic pulmonary fibrosis, DOID:0050156, FIBROCYSTIC PULMONARY DYSPLASIA, IDIOPATHIC PULMONARY FIBROSIS, FAMILIAL, cryptogenic fibrosing alveolitis ...

Text mining

[Next >](#)

Name	Z-score	Confidence
TGFB1	4.1	★★★★★
SFTPC	3.9	★★★★★
MUC1	3.7	★★★★★
SFTPД	3.4	★★★★★
ELMOD2	3.3	★★★★★
FN1	3.2	★★★★★
TERT	3.1	★★★★★
SFTPA2	3.0	★★★★★
MMP7	2.9	★★★★★
CTGF	2.9	★★★★★

Exercise 4

Open <http://tissues.jensenlab.org>

Look up tissue associations for insulin (INS)

Open <http://diseases.jensenlab.org>

Search for insulin receptor (INSR)

What is the strongest associated disease?

Inspect the underlying text-mining evidence