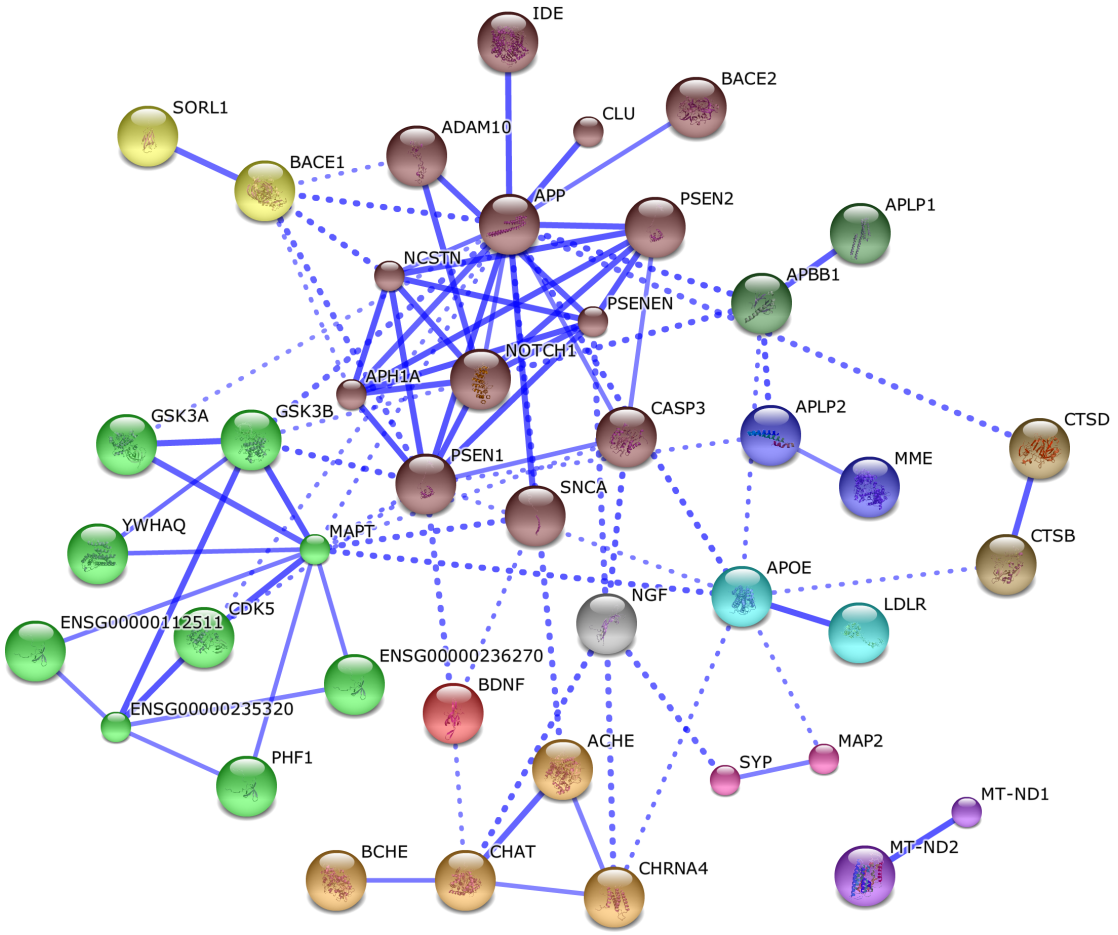
Large-scale integration of data and text



Lars Juhl Jensen

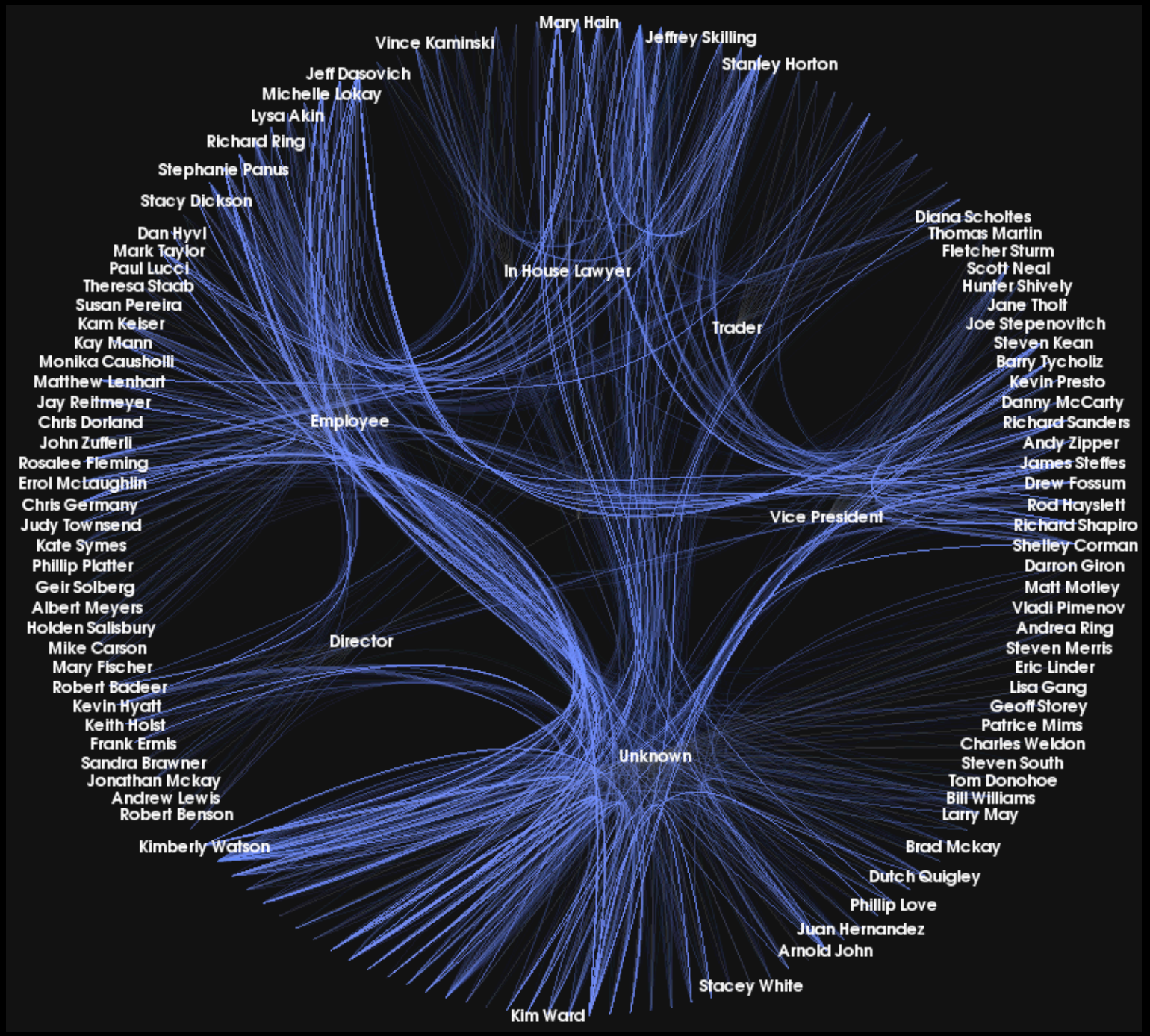
The Novo Nordisk Foundation
Center for Protein Research

FACULTY OF HEALTH SCIENCES
UNIVERSITY OF COPENHAGEN

interaction networks

association networks

guilt by association


Mary Hain
Vince Kaminski
Jeffrey Skilling
Stanley Horton
Jeff Dasovich
Michelle Lokay
Lysa Akin
Richard Ring
Stephanie Panus
Stacy Dickson
Dan Hyvl
Mark Taylor
Paul Lucci
Theresa Staab
Susan Pereira
Kam Keiser
Kay Mann
Monika Causholli
Matthew Lenhart
Jay Reitmeyer
Chris Dorland
John Zufferli
Rosalee Fleming
Errol McLaughlin
Chris Germany
Judy Townsend
Kate Symes
Phillip Platter
Geir Solberg
Albert Meyers
Holden Salisbury
Mike Carson
Mary Fischer
Robert Badeer
Kevin Hyatt
Keith Holst
Frank Ermis
Sandra Brawner
Jonathan Mckay
Andrew Lewis
Robert Benson
Kimberly Watson
In House Lawyer
Trader
Employee
Vice President
Director
Unknown
Diana Scholtes
Thomas Martin
Fletcher Sturm
Scott Neal
Hunter Shively
Jane Tholt
Joe Stepenovitch
Steven Kean
Barry Tycholiz
Kevin Presto
Danny McCarty
Richard Sanders
Andy Zipper
James Steffes
Drew Fossum
Rod Hayslett
Richard Shapiro
Shelley Corman
Darron Giron
Matt Motley
Vladi Pimenov
Andrea Ring
Steven Merris
Eric Linder
Lisa Gang
Geoff Storey
Patrice Mims
Charles Weldon
Steven South
Tom Donohoe
Bill Williams
Larry May
Brad Mckay
Dutch Quigley
Phillip Love
Juan Hernandez
Arnold John
Stacey White
Kim Ward

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)

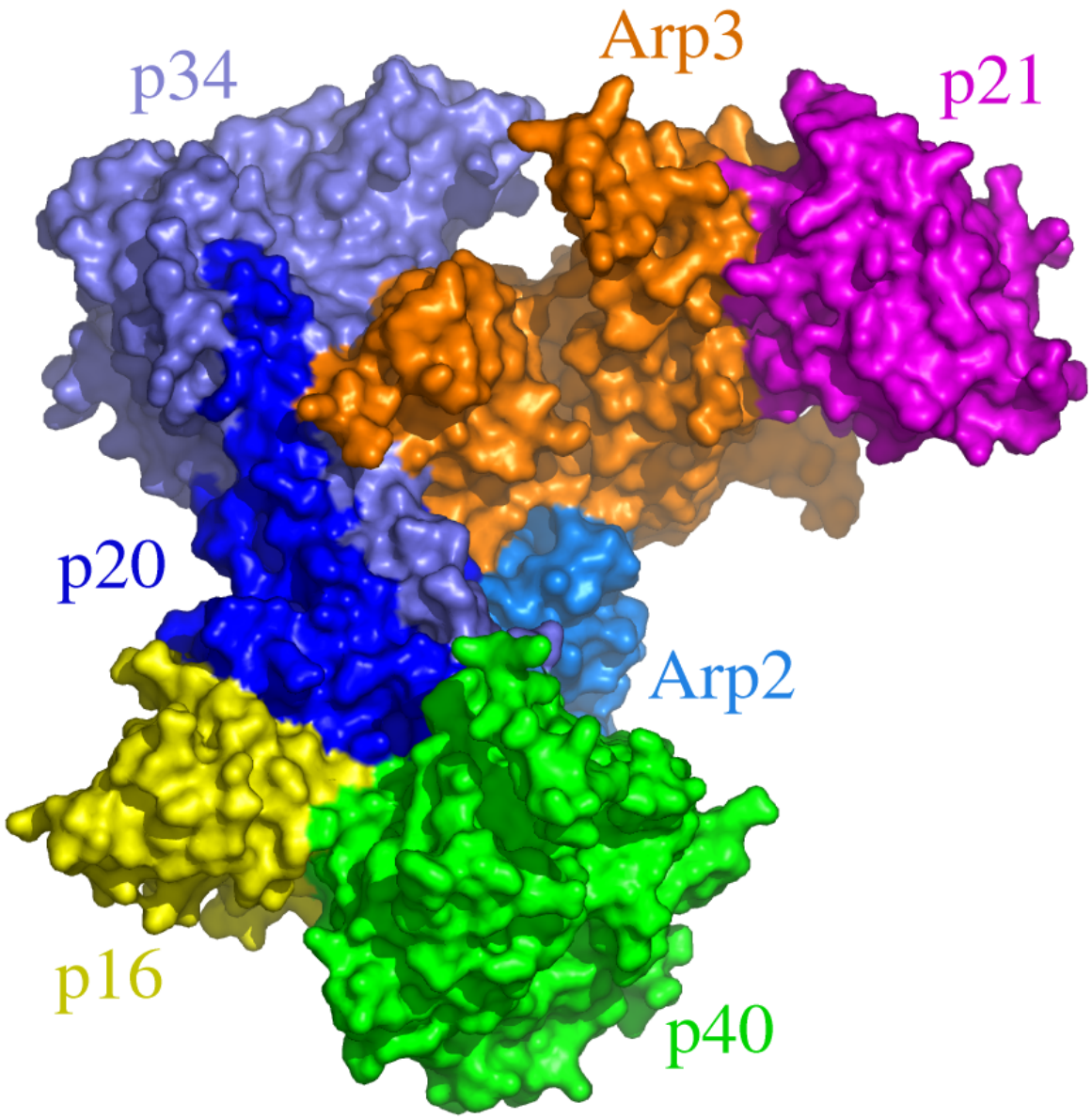
Why are there multiple lines connecting the same to two proteins?

curated knowledge

(what we know)

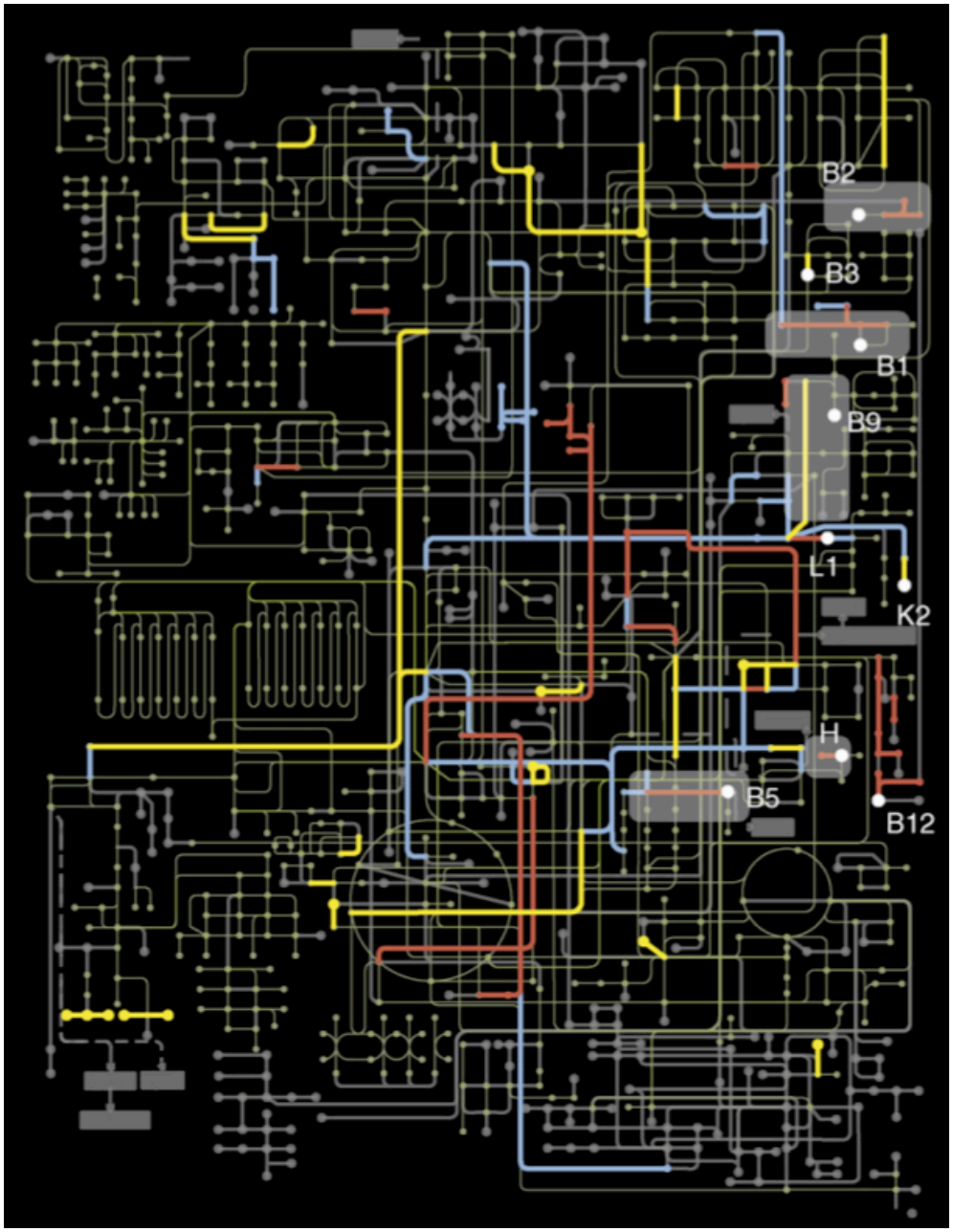
protein complexes

3D structures


p34
Arp3
p21
p20
Arp2
p16
p40

pathways

metabolic pathways



Letunic & Bork, *Trends in Biochemical Sciences*, 2008

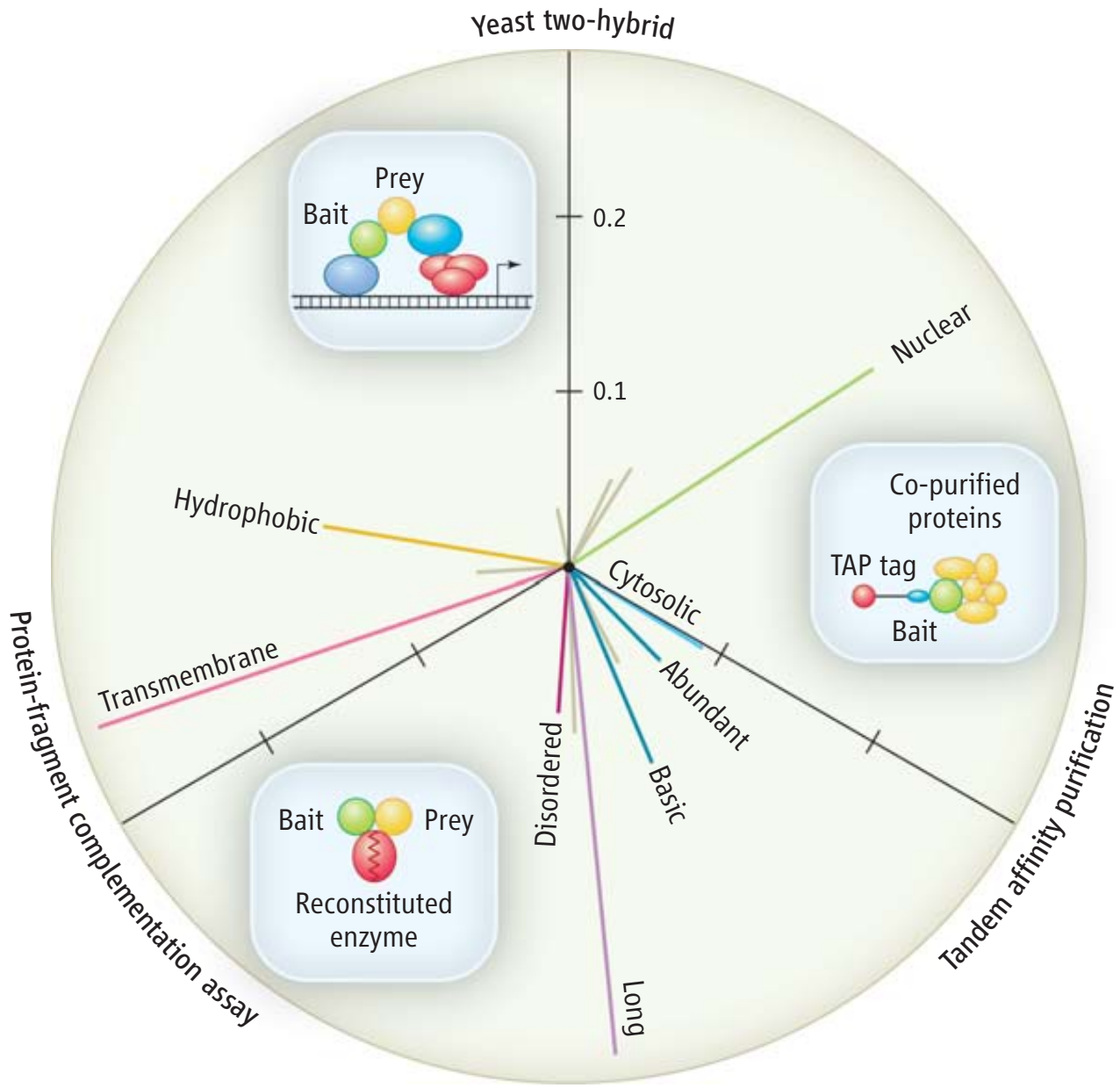
signaling pathways

very incomplete

experimental data

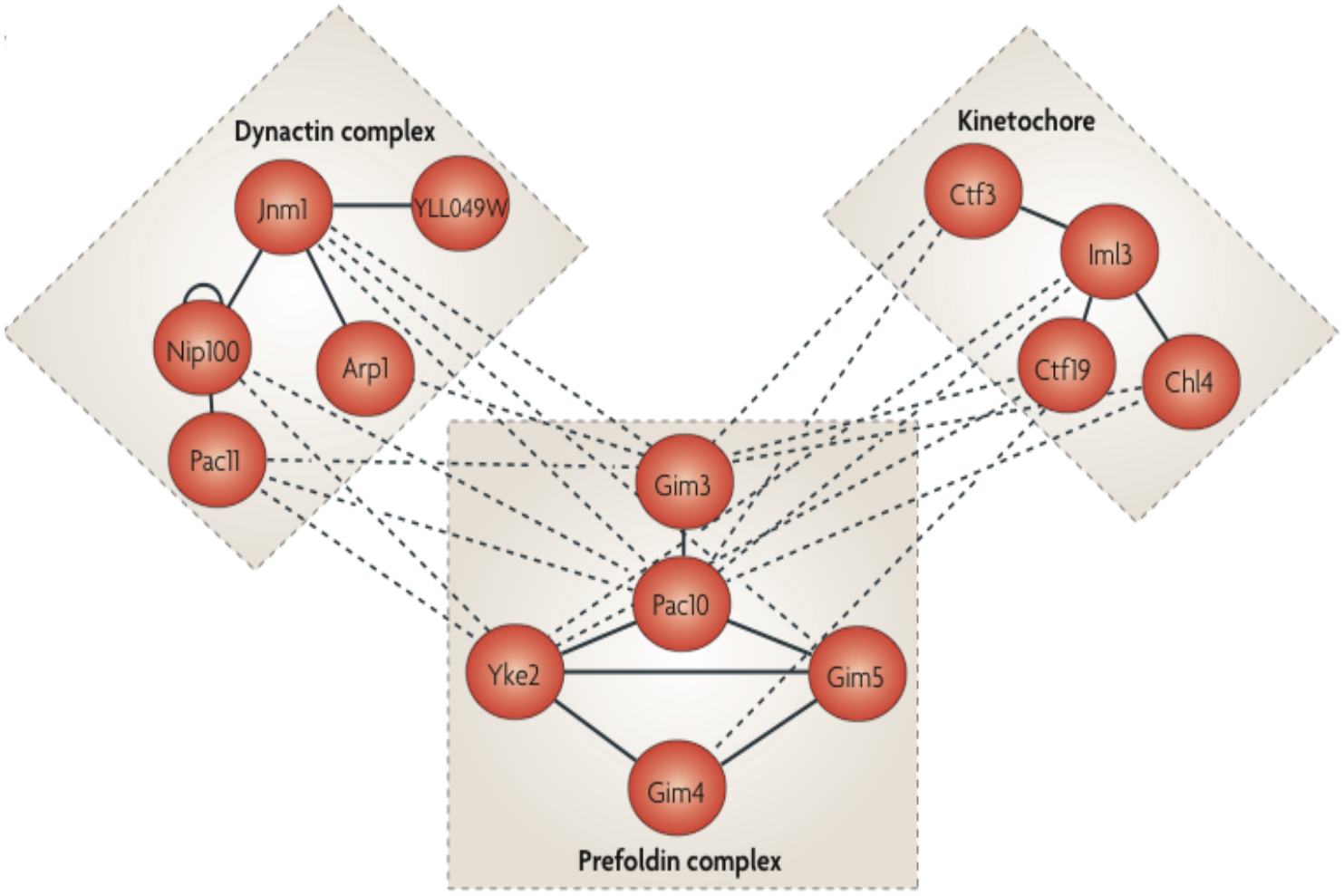
(what we measured)

physical interactions



Jensen & Bork, *Science*, 2008

genetic interactions



Beyer et al., *Nature Reviews Genetics*, 2007

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?

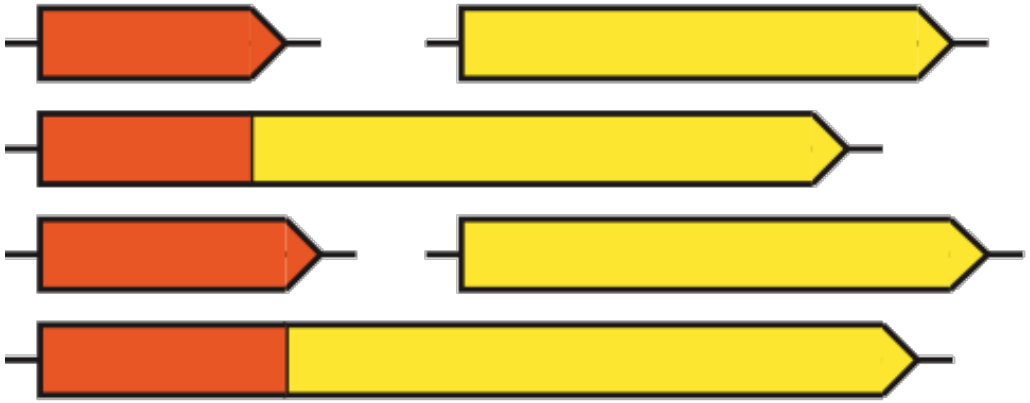
predictions

(what we infer)

genomic context

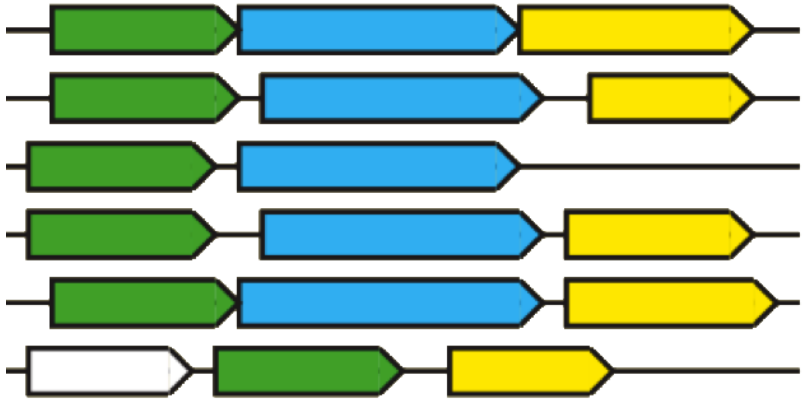
evolution

gene fusion



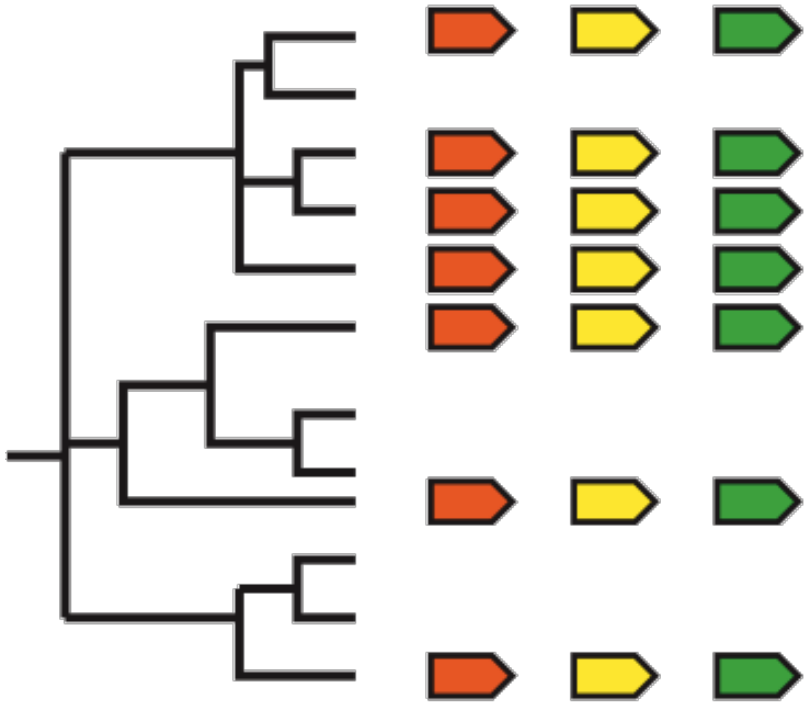
Korbel et al., *Nature Biotechnology*, 2004

gene neighborhood



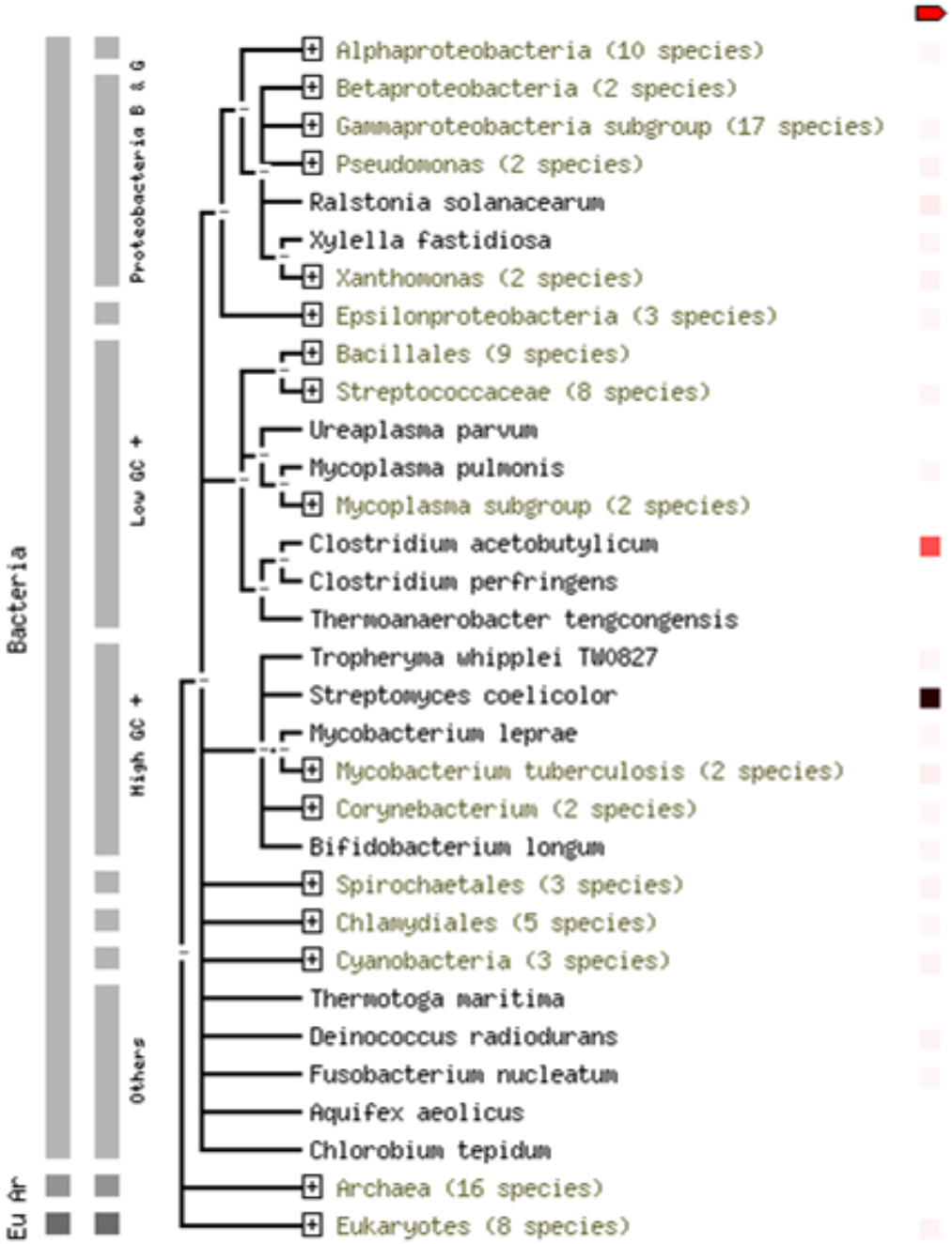
Korbel et al., *Nature Biotechnology*, 2004

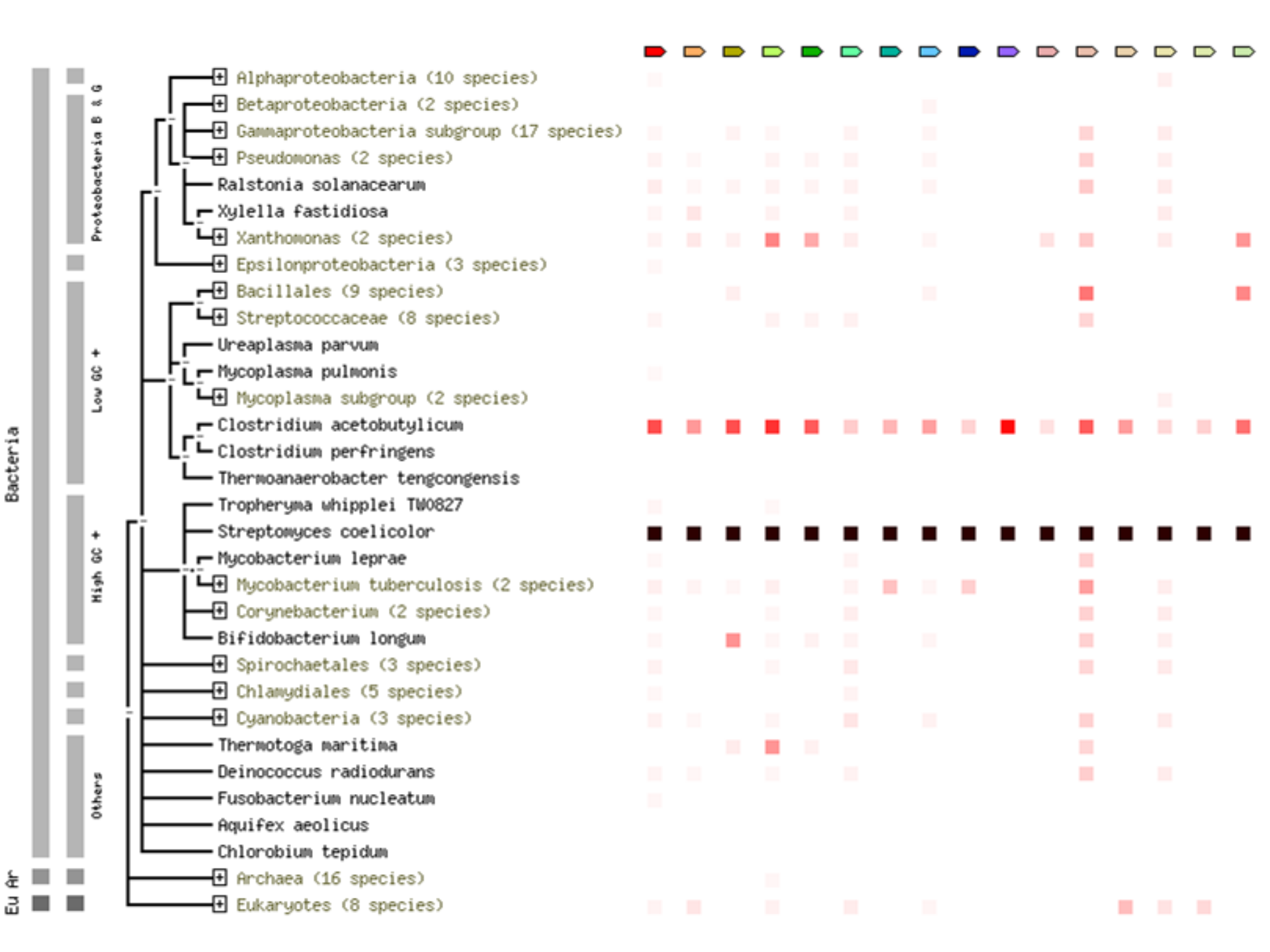
phylogenetic profiles

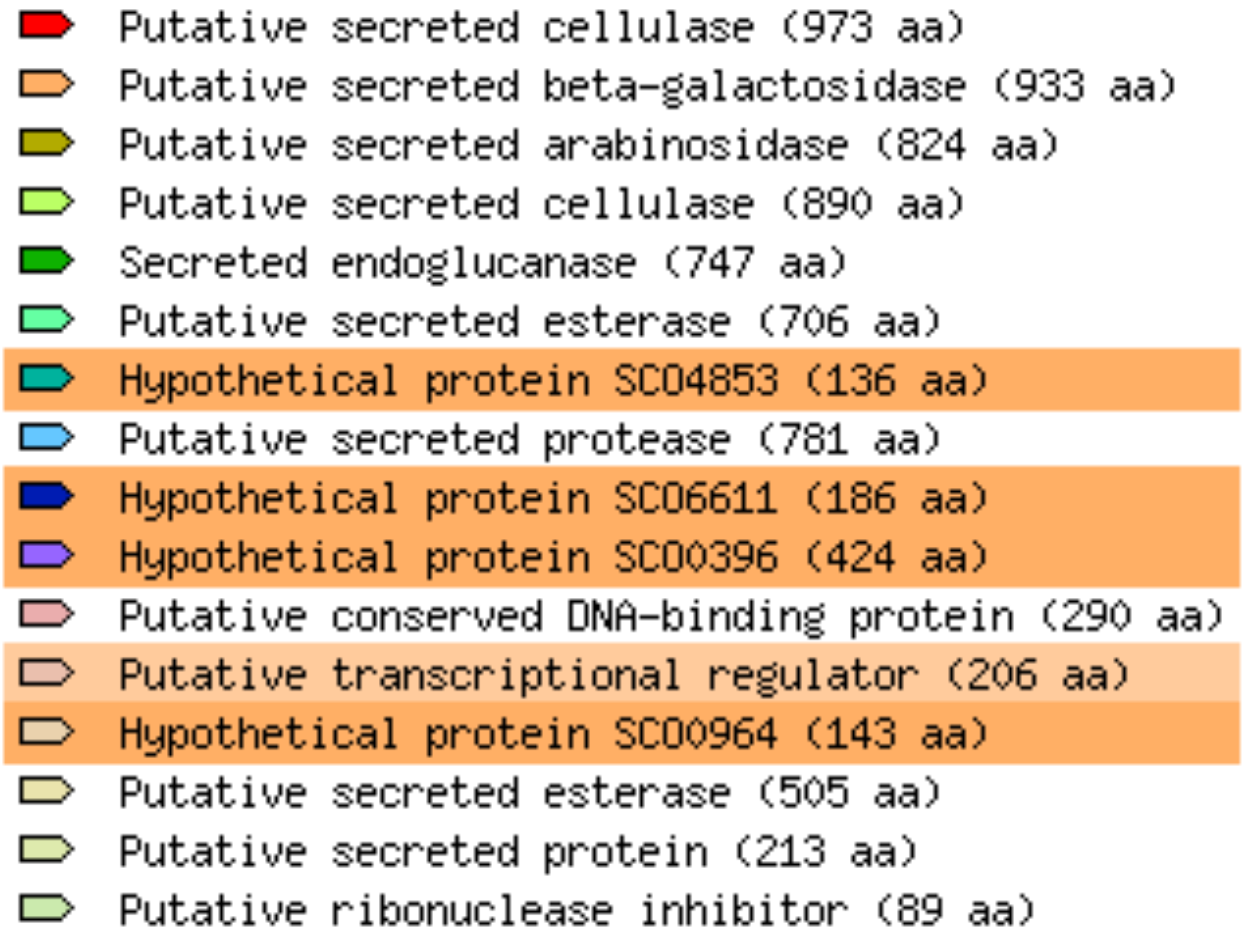


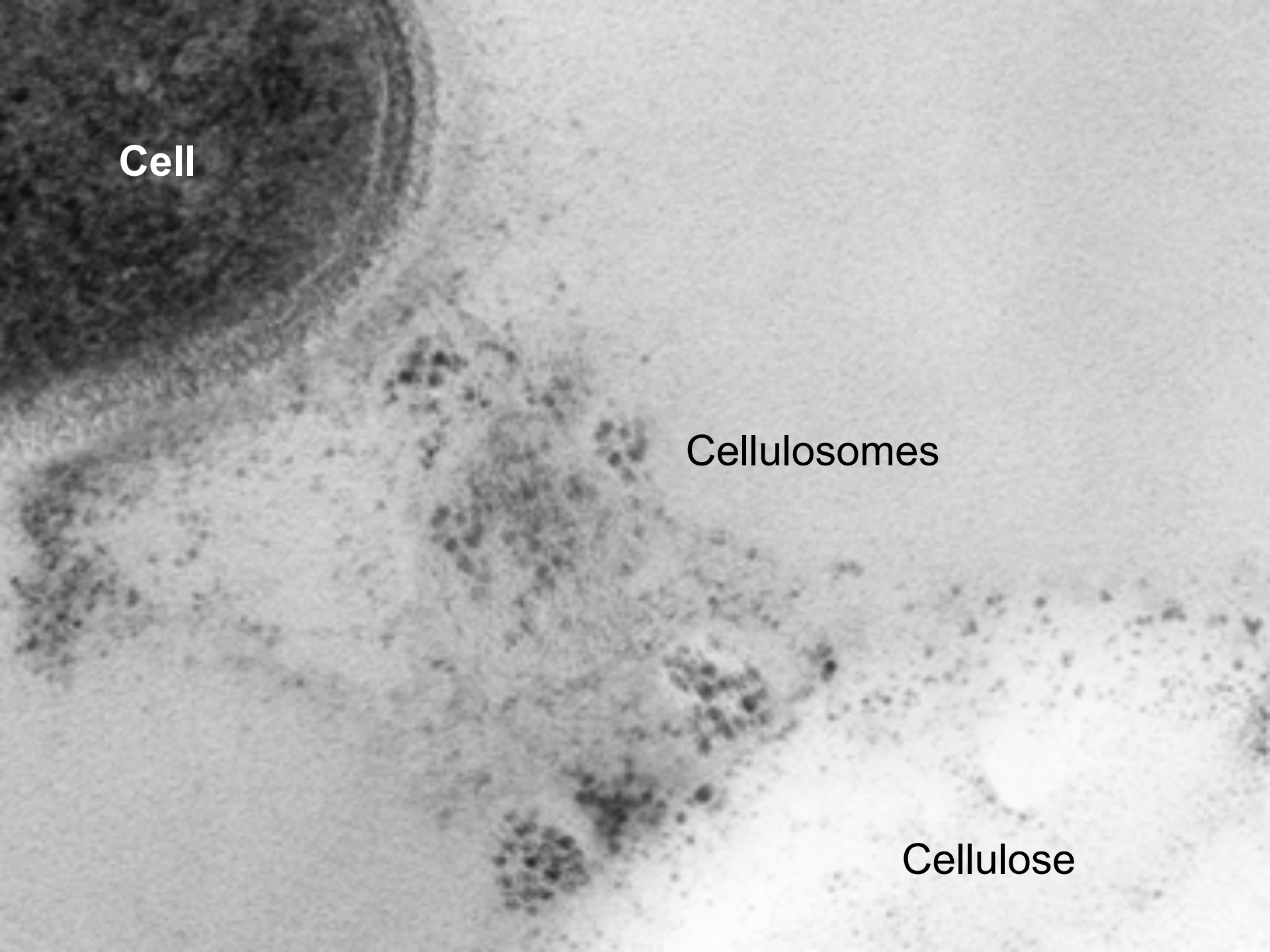
Korbel et al., *Nature Biotechnology*, 2004

a real example


Bacteria
Eu
Ar
Proteobacteria B & G
Low GC +
High GC +
Others
Alphaproteobacteria (10 species)
Betaproteobacteria (2 species)
Gammaproteobacteria subgroup (17 species)
Pseudomonas (2 species)
Ralstonia solanacearum
Xylella fastidiosa
Xanthomonas (2 species)
Epsilonproteobacteria (3 species)
Bacillales (9 species)
Streptococcaceae (8 species)
Ureaplasma parvum
Mycoplasma pulmonis
Mycoplasma subgroup (2 species)
Clostridium acetobutylicum
Clostridium perfringens
Thermoanaerobacter tengcongensis
Tropheryma whipplei TW0827
Streptomyces coelicolor
Mycobacterium leprae
Mycobacterium tuberculosis (2 species)
Corynebacterium (2 species)
Bifidobacterium longum
Spirochaetales (3 species)
Chlamydiales (5 species)
Cyanobacteria (3 species)
Thermotoga maritima
Deinococcus radiodurans
Fusobacterium nucleatum
Aquifex aeolicus
Chlorobium tepidum
Archaea (16 species)
Eukaryotes (8 species)


Bacteria
Eu
Ar
Proteobacteria B & G
Low GC +
High GC +
Others
Alphaproteobacteria (10 species)
Betaproteobacteria (2 species)
Gammaproteobacteria subgroup (17 species)
Pseudomonas (2 species)
Ralstonia solanacearum
Xylella fastidiosa
Xanthomonas (2 species)
Epsilonproteobacteria (3 species)
Bacillales (9 species)
Streptococcaceae (8 species)
Ureaplasma parvum
Mycoplasma pulmonis
Mycoplasma subgroup (2 species)
Clostridium acetobutylicum
Clostridium perfringens
Thermoanaerobacter tengcongensis
Tropheryma whipplei TW0827
Streptomyces coelicolor
Mycobacterium leprae
Mycobacterium tuberculosis (2 species)
Corynebacterium (2 species)
Bifidobacterium longum
Spirochaetales (3 species)
Chlamydiales (5 species)
Cyanobacteria (3 species)
Thermotoga maritima
Deinococcus radiodurans
Fusobacterium nucleatum
Aquifex aeolicus
Chlorobium tepidum
Archaea (16 species)
Eukaryotes (8 species)


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 SCO4853 (136 aa)
Putative secreted protease (781 aa)
Hypothetical protein SCO6611 (186 aa)
Hypothetical protein SCO0396 (424 aa)
Putative conserved DNA-binding protein (290 aa)
Putative transcriptional regulator (206 aa)
Hypothetical protein SCO0964 (143 aa)
Putative secreted esterase (505 aa)
Putative secreted protein (213 aa)
Putative ribonuclease inhibitor (89 aa)


Cell
Cellulosomes
Cellulose

complications

many databases

different formats

different identifiers

variable quality

not comparable

not same species

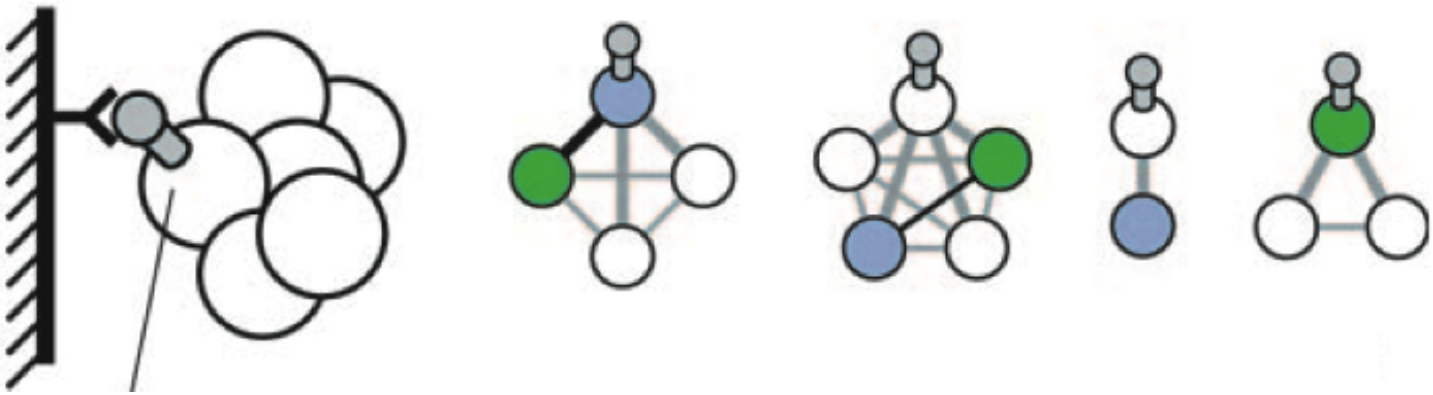
hard work

parsers

mapping files

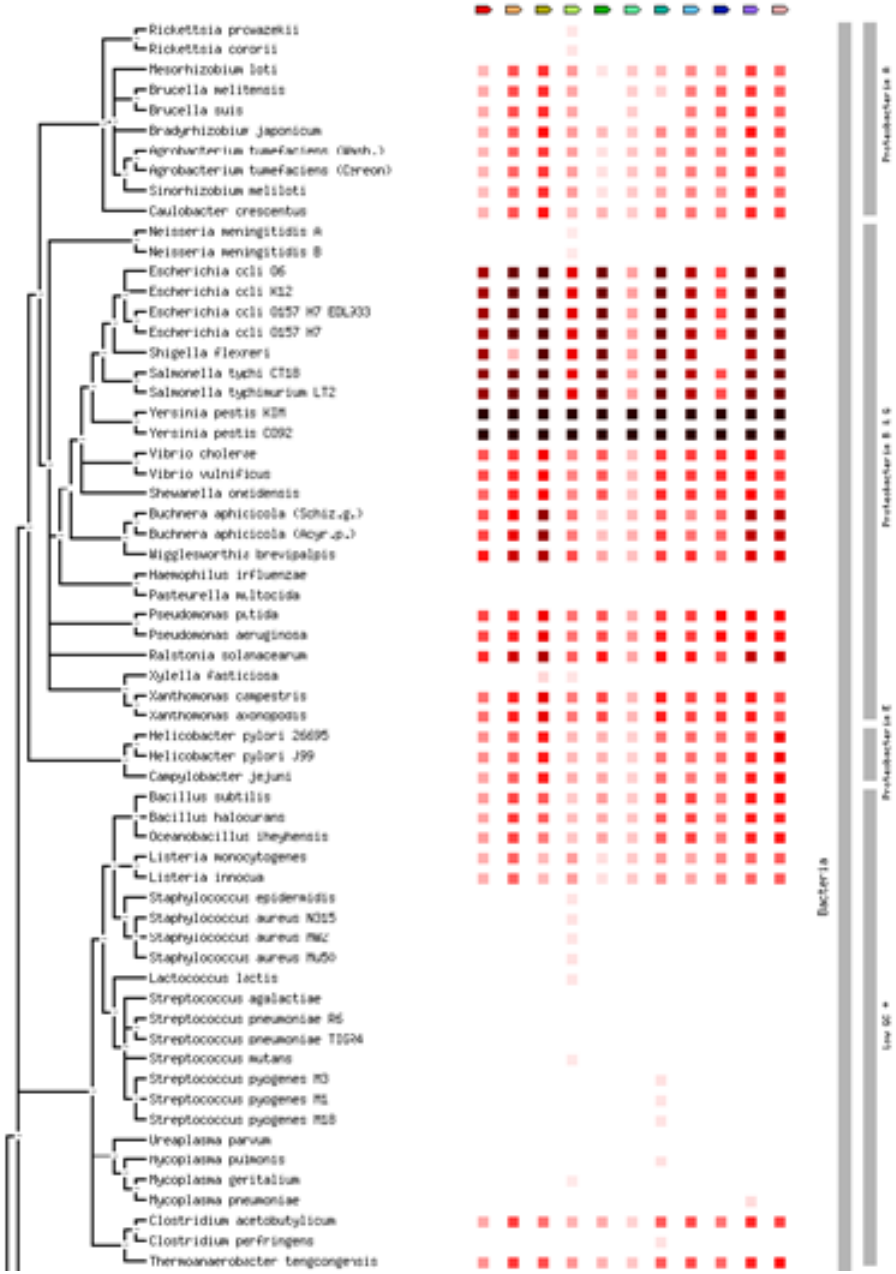
quality scores

affinity purification



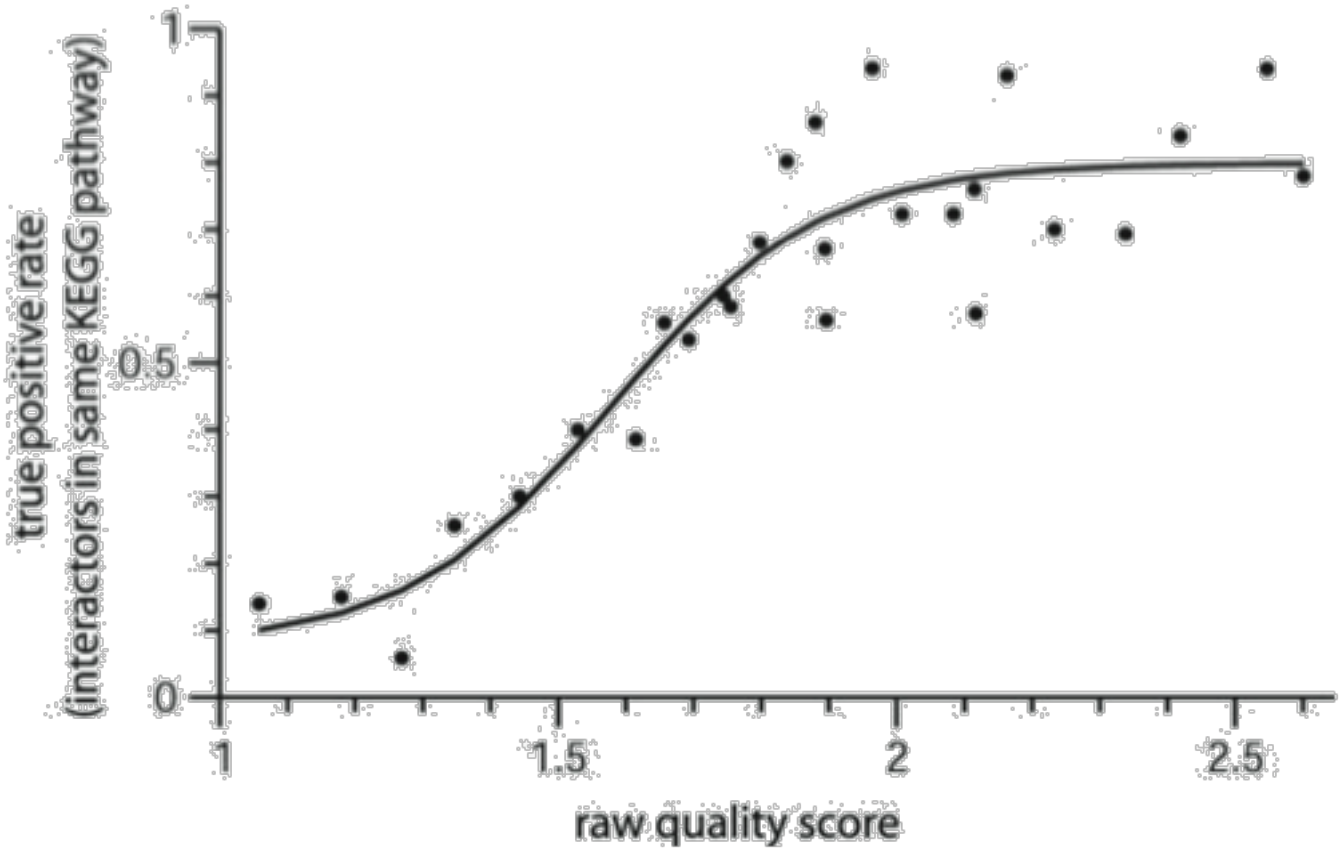
von Mering et al., *Nucleic Acids Research*, 2005

phylogenetic profiles


Rickettsia prowazekii
Rickettsia conorii
Mesorhizobium loti
Brucella melitensis
Brucella suis
Bradyrhizobium japonicum
Agrobacterium tumefaciens (UWash.)
Agrobacterium tumefaciens (Dcreon)
Sinorhizobium meliloti
Caulobacter crescentus
Neisseria meningitidis A
Neisseria meningitidis B
Escherichia coli O6
Escherichia coli K12
Escherichia coli O157 H7 EDL933
Escherichia coli O157 H7
Shigella flexneri
Salmonella typhi CT18
Salmonella typhimurium LT2
Yersinia pestis KIM
Yersinia pestis CO92
Vibrio cholerae
Vibrio vulnificus
Shewanella oneidensis
Buchnera aphicicola (Schiz.g.)
Buchnera aphicicola (Acyr.p.)
Wigglesworthia brevipalpis
Haemophilus influenzae
Pasteurella multocida
Pseudomonas putida
Pseudomonas aeruginosa
Ralstonia solanacearum
Xylella fastidiosa
Xanthomonas campestris
Xanthomonas axonopodis
Helicobacter pylori 26695
Helicobacter pylori J99
Campylobacter jejuni
Bacillus subtilis
Bacillus halodurans
Oceanobacillus iheyhensis
Listeria monocytogenes
Listeria innocua
Staphylococcus epidermidis
Staphylococcus aureus N315
Staphylococcus aureus MW2
Staphylococcus aureus Mu50
Lactococcus lactis
Streptococcus agalactiae
Streptococcus pneumoniae R6
Streptococcus pneumoniae TIGR4
Streptococcus mutans
Streptococcus pyogenes M3
Streptococcus pyogenes M1
Streptococcus pyogenes M18
Ureaplasma parvum
Mycoplasma pulmonis
Mycoplasma genitalium
Mycoplasma pneumoniae
Clostridium acetobutylicum
Clostridium perfringens
Thermoanaerobacter tengcongensis
Bacteria
Proteobacteria α
Proteobacteria β γ δ
Proteobacteria ε
Low GC +

score calibration

gold standard



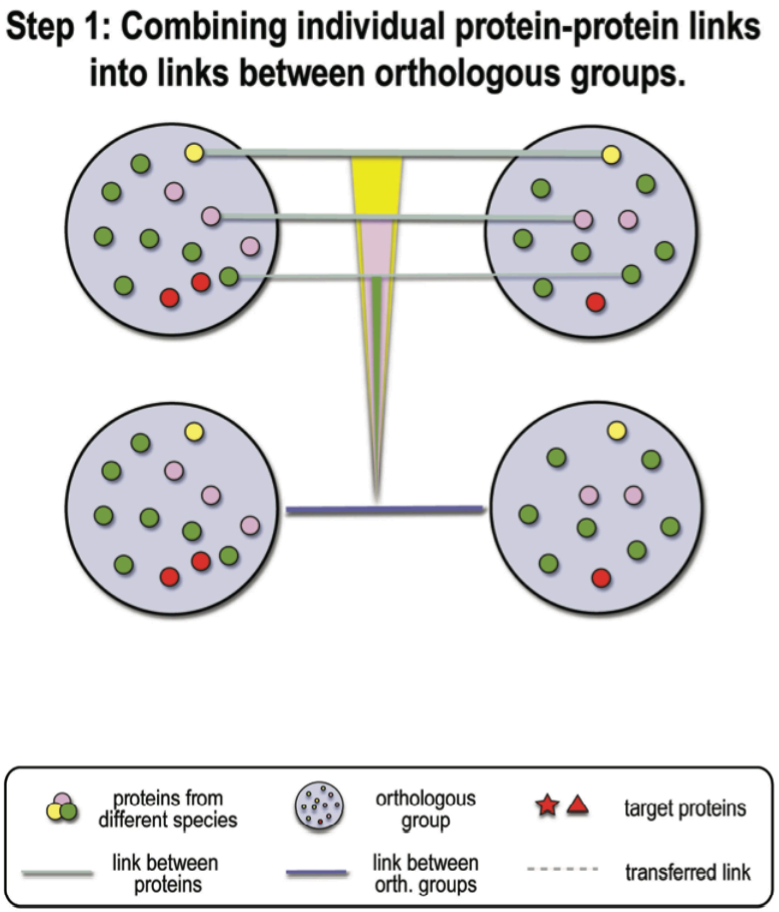
von Mering et al., *Nucleic Acids Research*, 2005

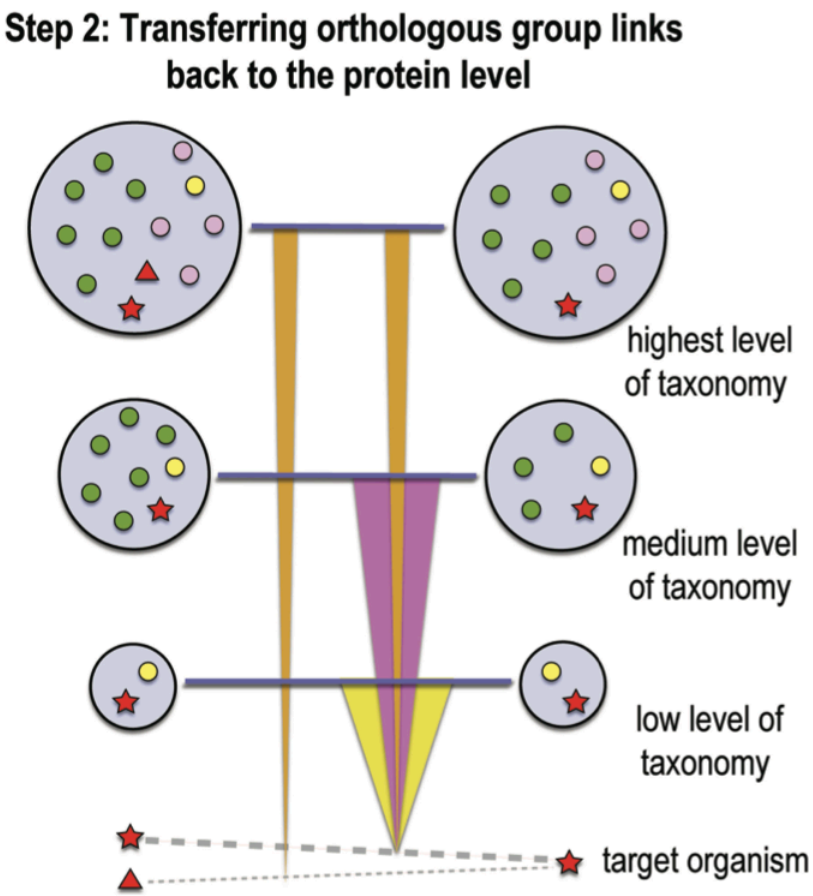
implicit weighting by quality

common scale

homology-based transfer

orthologous groups





Franceschini et al., *Nucleic Acids Research*, 2013

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
blah blah GINGER blah
blah blah blah blah blah
blah blah GINGER blah
blah blah blah blah...

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

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

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

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]

Saric et al., *Proceedings of ACL*, 2004

type and direction

complex sentences

anaphoric references

it

summary

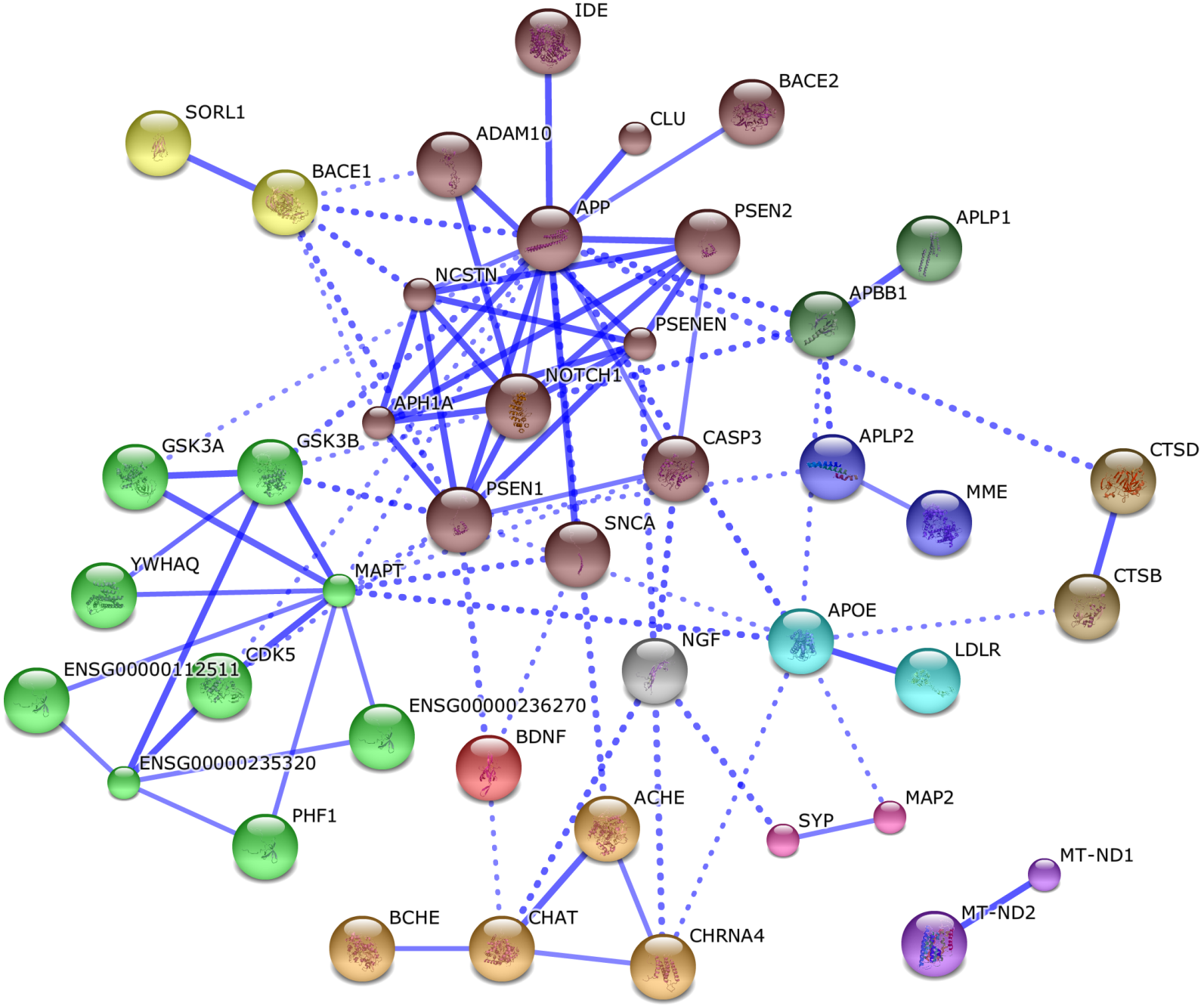
association networks

heterogeneous data

common identifiers

quality scores

protein networks

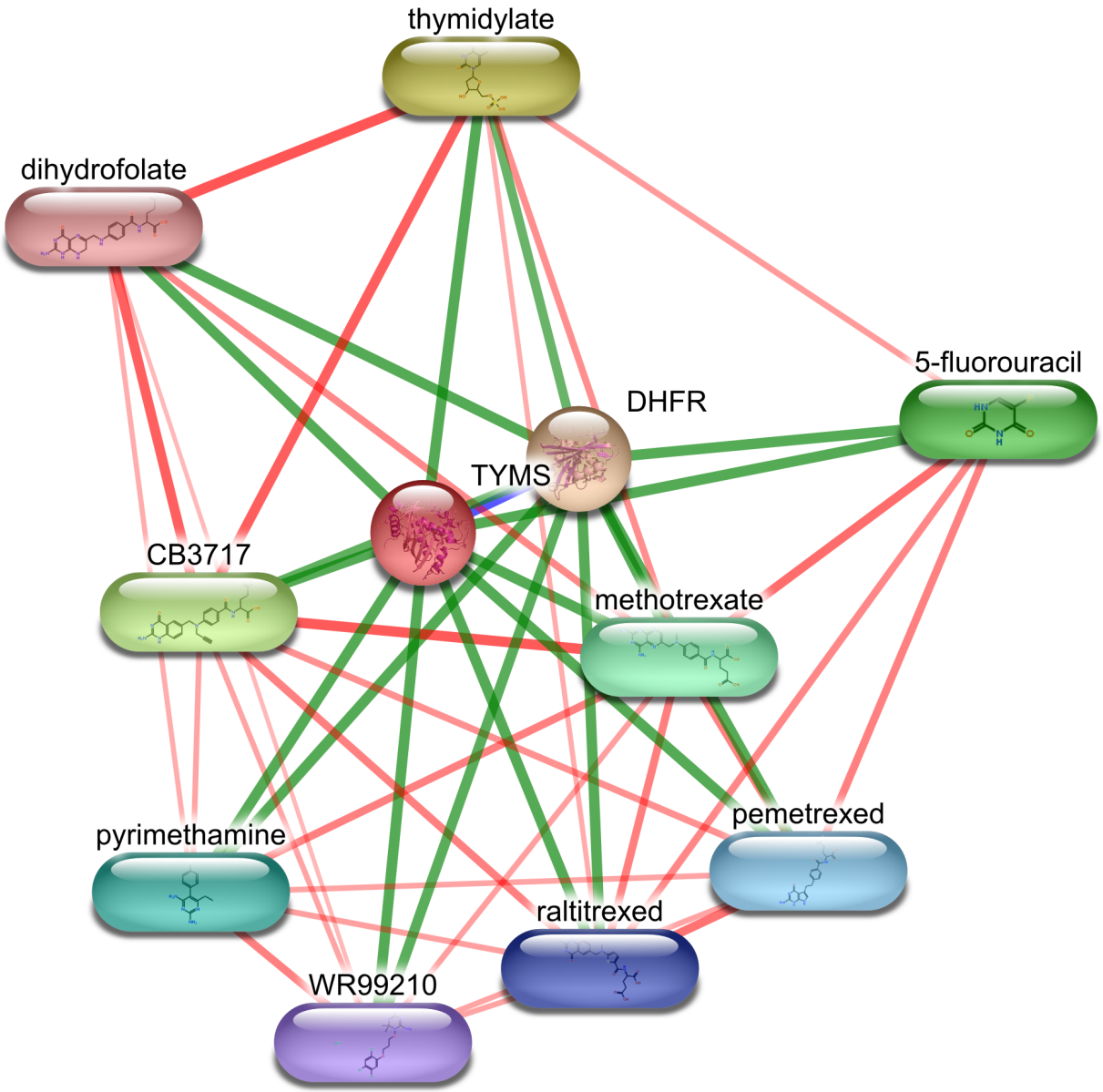


string-db.org

Szklarczyk et al., *Nucleic Acids Research*, 2015

STITCH

chemical networks

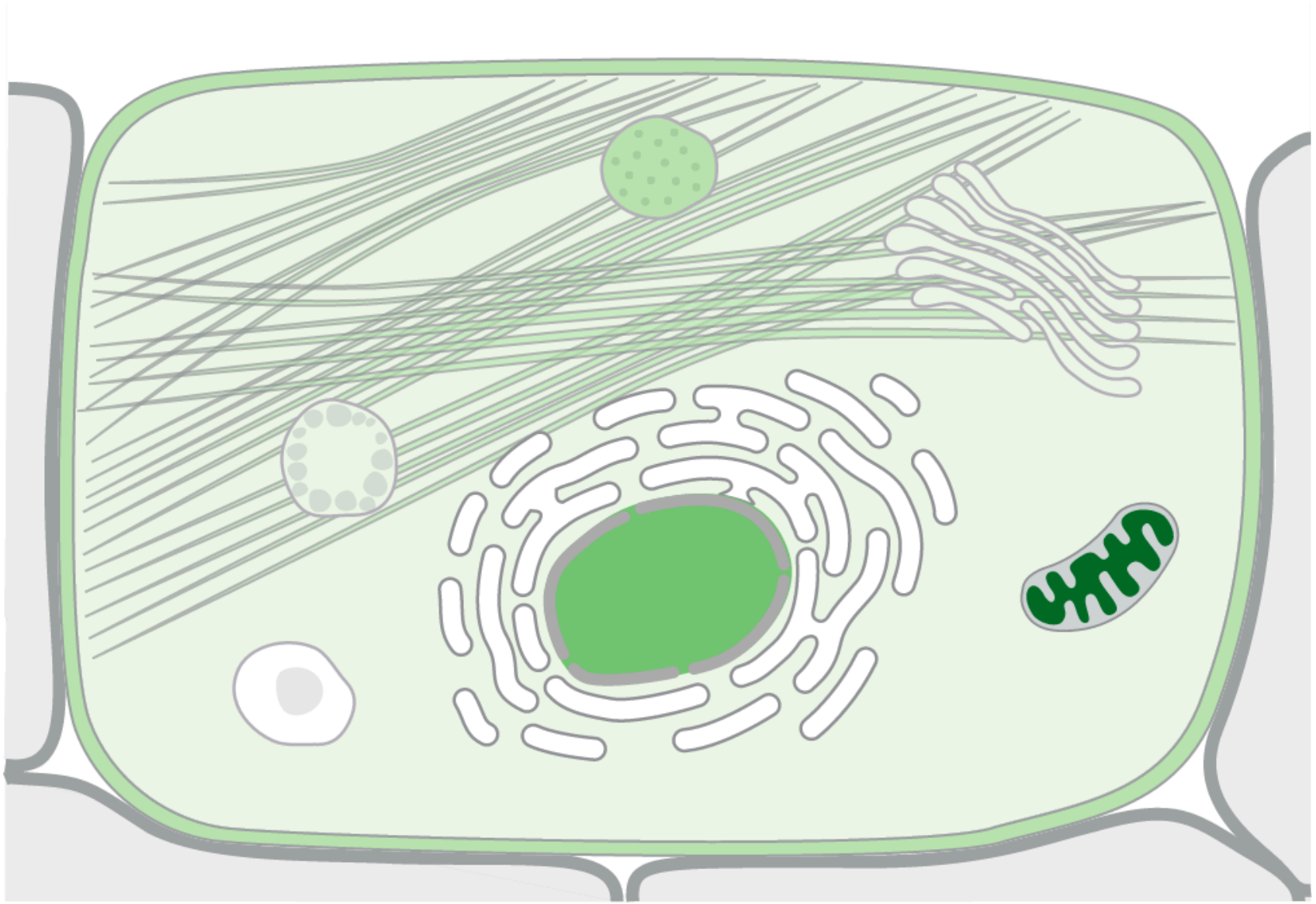


stitch-db.org

Kuhn et al., *Nucleic Acids Research*, 2014

COMPARTMENTS

subcellular localization

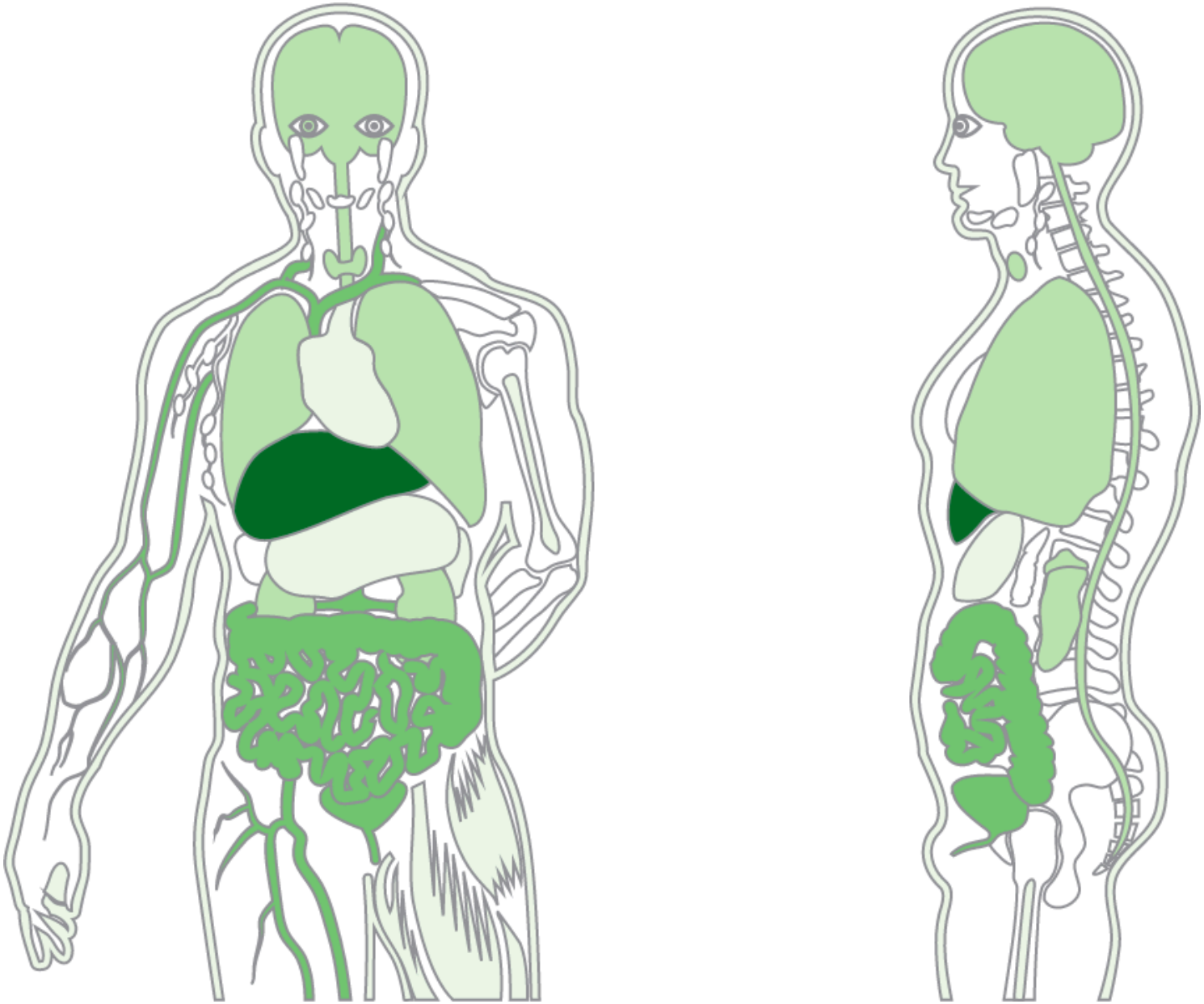


compartments.jensenlab.org

Binder et al., *Database*, 2014

TISSUES

tissue expression



tissues.jensenlab.org

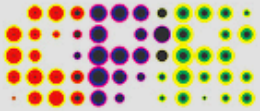
Santos et al., *submitted*, 2015

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	★★☆☆☆
SFTPD	3.4	★★☆☆☆
ELMOD2	3.3	★★☆☆☆
FN1	3.2	★★☆☆☆
TERT	3.1	★★☆☆☆
SFTPA2	3.0	★★☆☆☆
MMP7	2.9	★★☆☆☆
CTGF	2.9	★★☆☆☆

diseases.jensenlab.org

Frankild et al., *Methods*, 2015

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