A (Partial) Tour of Bioinformatics

Three examples to illustrate
the interplay of statistics,
computer science, and biology

Contents

- What is bioinformatics?
- Why it exist?
- 3 examples to illustrate some aspects of last 3 years
- How bioinformatics changes biological research, now and future?

What is bioinformatics?

- DEFINITION: Research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, organize, archive, analyze, or visualize such data.
- Personal definition: Using statistical and computational tools to understand biological problems

Why bioinformatics exist?

- Able to do too much too fast
  - 10 years ago, 2000 SNP analysis (2000 PCRs) in 3 days
  - now, 6,000,000 SNP analyses (30 x 200,000) in 3 days
- Computer scientists, biologists, and statisticians cannot communicate to each other

Why bioinformatics exist?

- Able to do too much too fast
- Computer scientists, biologists, and statisticians cannot communicate to each other

A biophysicist talks
physics to the biologists
and biology to the physicists,
but then he meets another biophysicist,
they just discuss women.

-- Anonymous
but famous

Distribution of Posters in Asia-Pacific Bioinformatics Conference 2006

- Sequence analysis – 30%
- Microarray and experimental analysis – 30%
- Molecular modeling - 15%
- System biology – 10%
- Literature Analysis – 5%
- Others – 10%

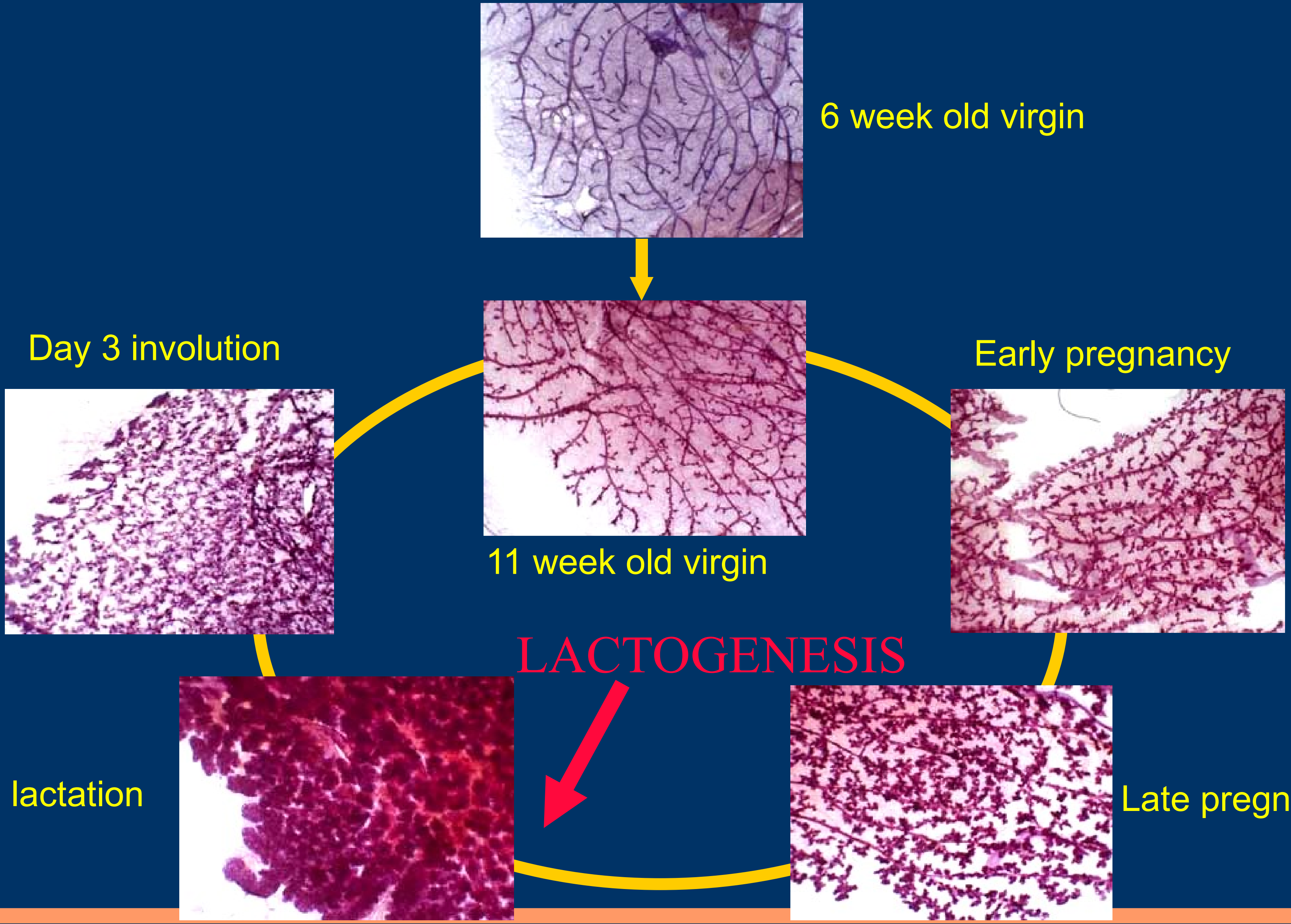
Taming of the Omics

- -ome: the entire collection of
- -omics: the study of -ome

- Genomics: the study of genome (genes)
- Proteomics: the study of proteome (proteins)

- Transcriptomics (expressed genes)
- Metabolomics (metabolism)
- Physiomics (organ physiology)
- Biblomics (human written literature)

Example 1: Mouse lactation biology



6 week old virgin

11 week old virgin

Early pregnancy

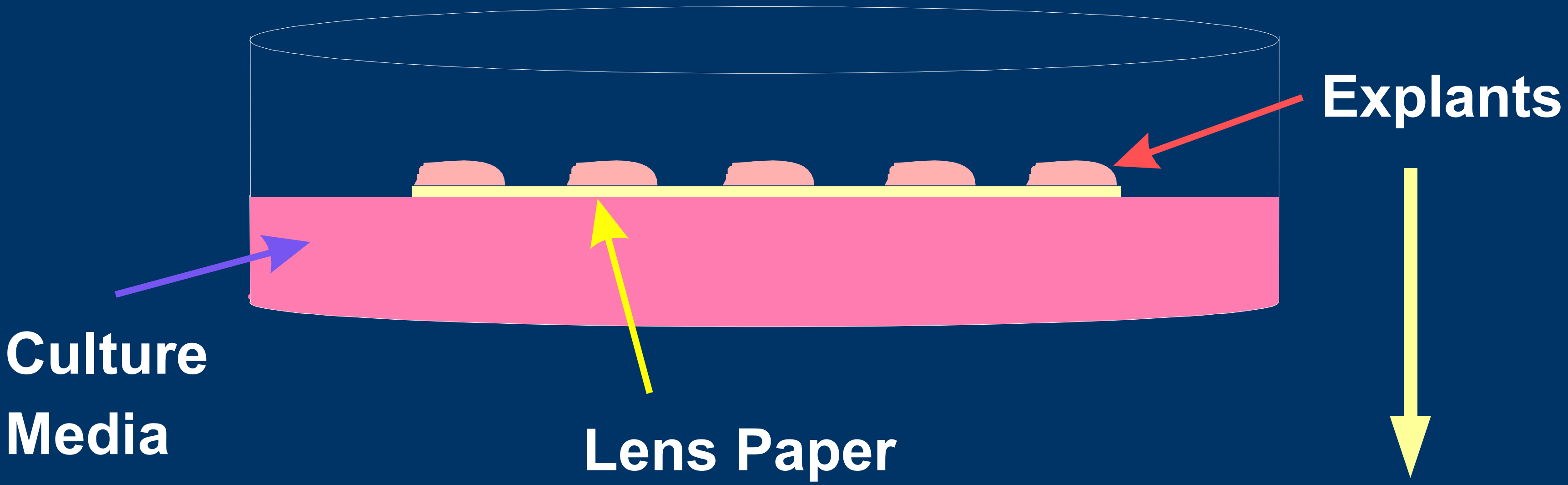
Late pregnancy

LACTOGENESIS

2 days lactation

Day 3 involution

Example 1: Mouse lactation biology



Containing:

insulin 胰島素

glucocorticoid 糖皮质激素

prolactin 催乳素

To look at crucial gene expressions, also known as marker genes (eg, a-casein)

Example 1: Mouse lactation biology

- Main problem in lactation by culture model
- Is tissue culture representative of the actual animal?
  - It is very very difficult to repeat tissue culture experiments on real mouse

Example 1: Mouse lactation biology

- Using 4 microarray datasets to get an idea of what happens in the mammary gland (animal)
  - About 100 Affymetrix microarrays
- Comparing tissue culture microarrays with animal microarrays between day 17 pregnancy and day 1 lactation
- Using protein-protein interactions mined from text to combine with microarrays for better understanding of my questions

Example 1: Mouse lactation biology

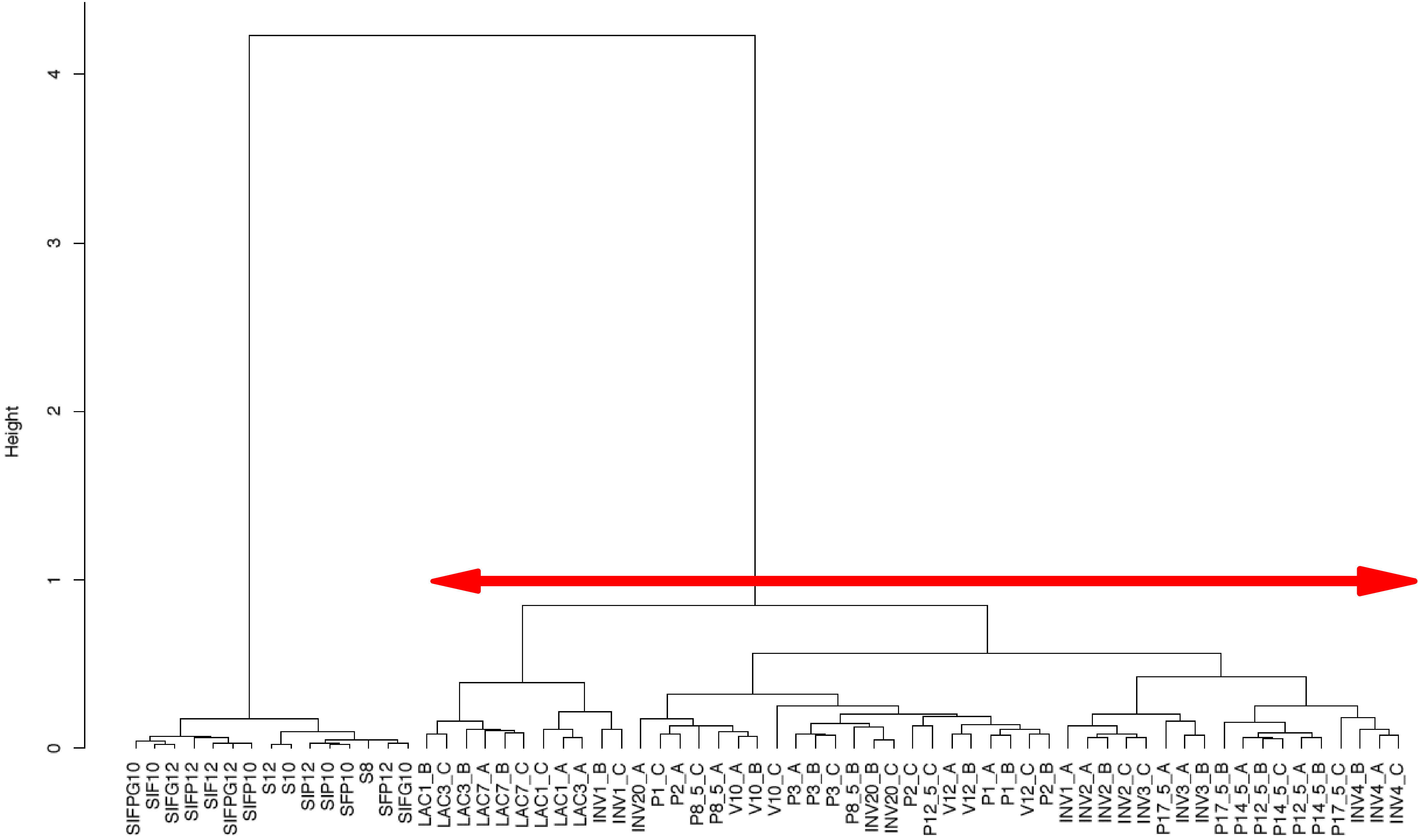
- How to do it?

- Look at the transcriptome globally – clustering by correlations

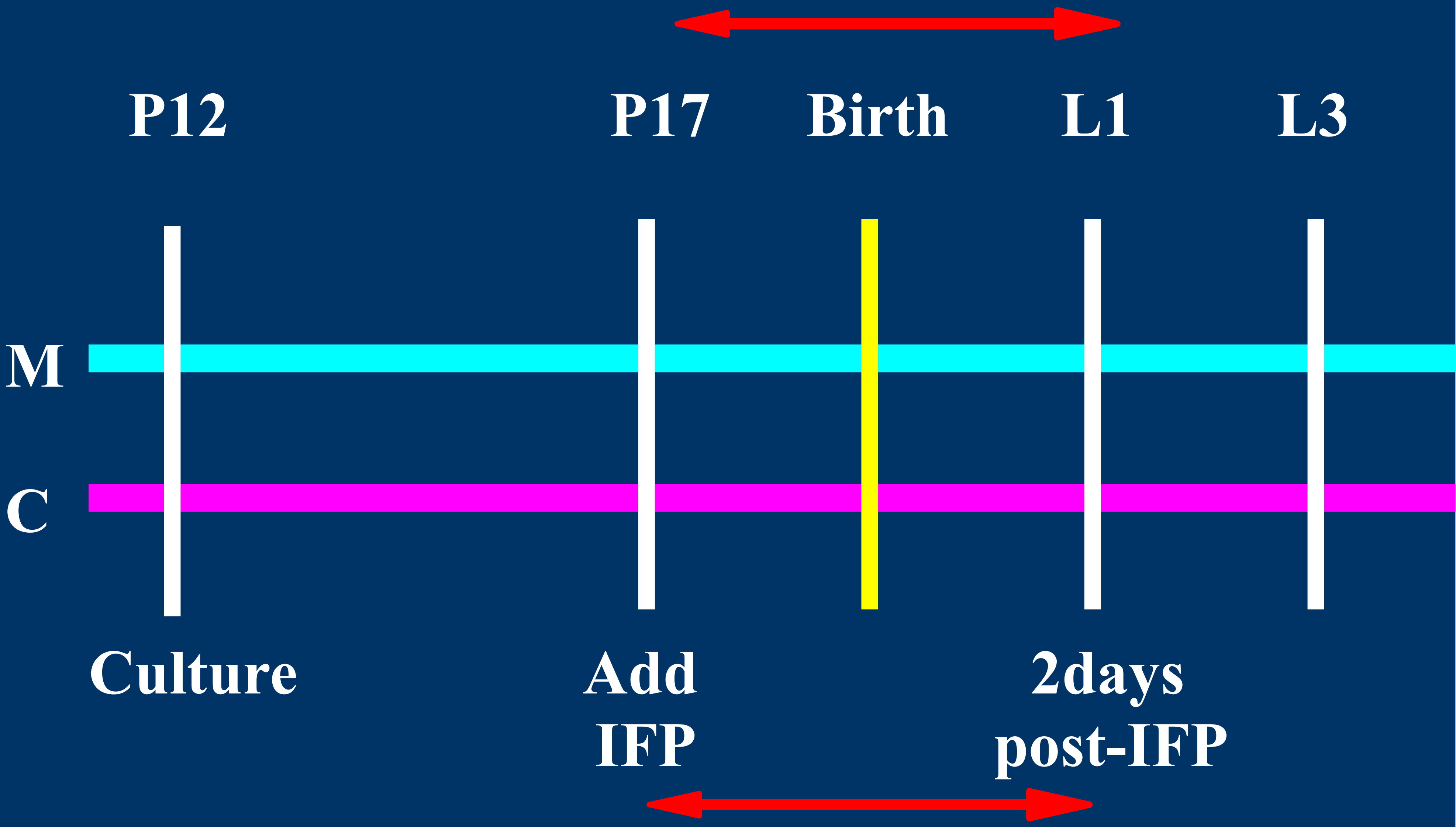
- Take the 6 differentials (ratios) and do permutation pair-wise comparison
- Simplest test: two-sided Sign test

- Used a stronger test – two-sided Wilconox exact rank sign test with continuity correction

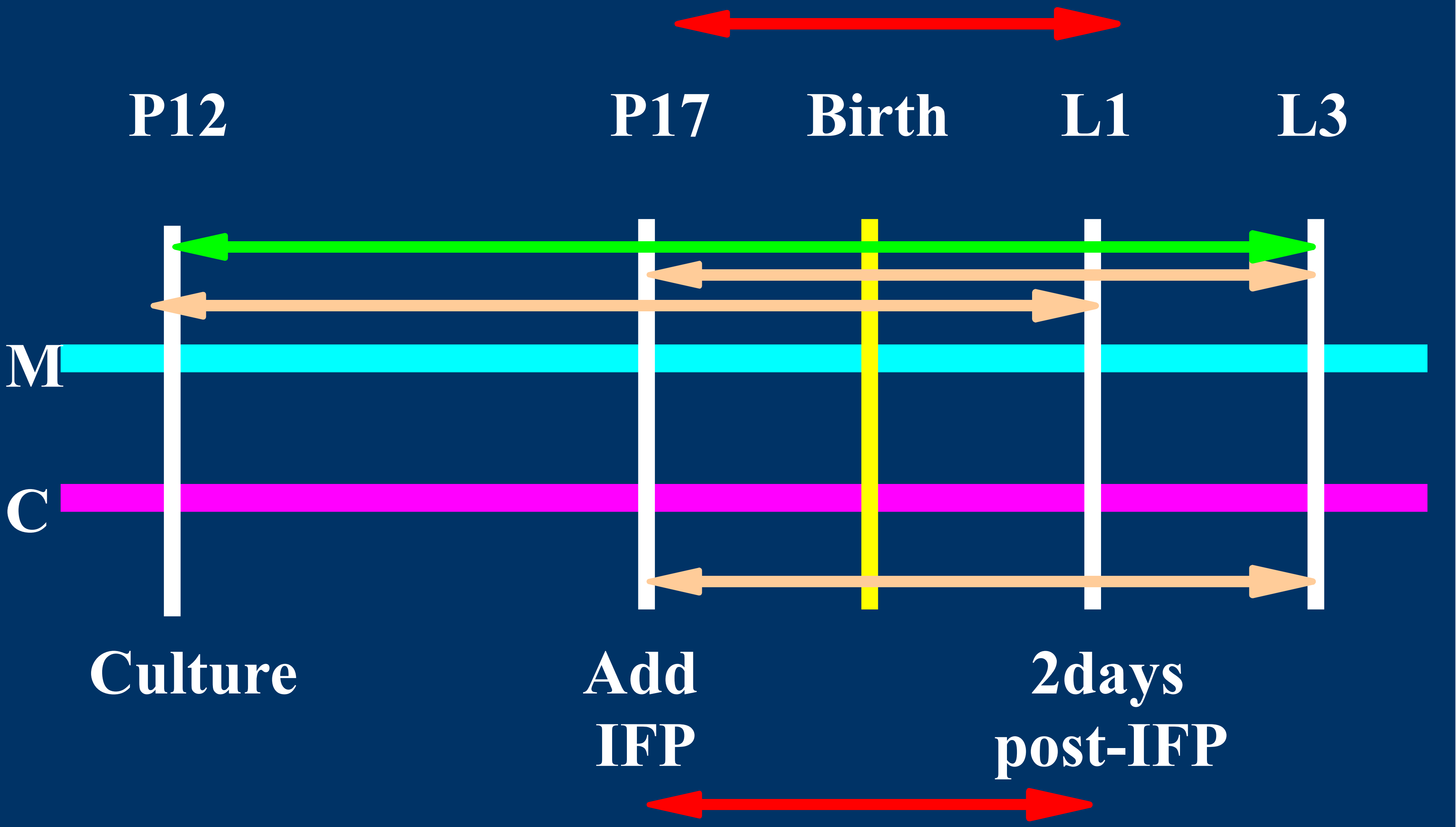
Example 1: Mouse lactation biology



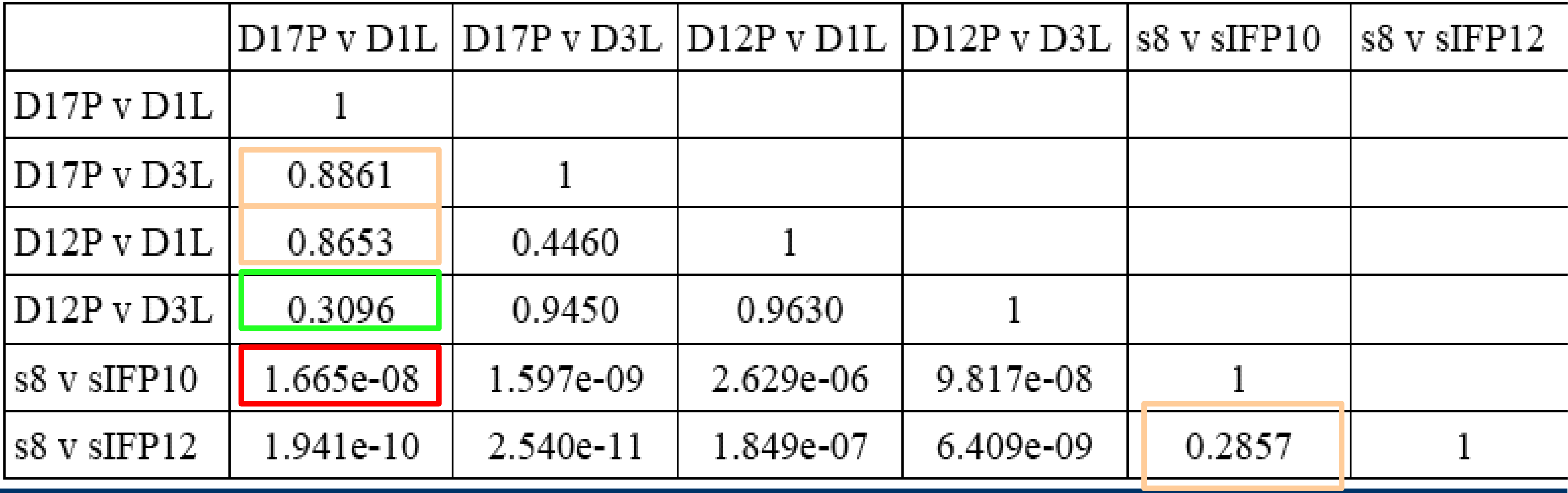
Example 1: Mouse lactation biology



Example 1: Mouse lactation biology



Example 1: Mouse lactation biology



	D17P v D1L	D17P v D3L	D12P v D1L	D12P v D3L	s8 v sIFP10	s8 v sIFP12
D17P v D1L	1					
D17P v D3L	0.8861	1				
D12P v D1L	0.8653	0.4460	1			
D12P v D3L	0.3096	0.9450	0.9630	1		
s8 v sIFP10	1.665e-08	1.597e-09	2.629e-06	9.817e-08	1	
s8 v sIFP12	1.941e-10	2.540e-11	1.849e-07	6.409e-09	0.2857	1

- These, plus a few other consistent results suggest that culture model does not mimic mammary gland function in the mouse well but it still has some merits.

Example 1: Mouse lactation biology

- What else can I do?

- Use microarray as a “fishing expedition” for interesting stuffs
  - new hypotheses?

Example 1: Mouse lactation biology

- Generate a co-expression network based on Pearson's correlation coefficient

- Related genes (ie, in the same pathway) exhibit similar expression patterns
  - positive feedback => positive correlation
  - negative feedback => negative correlation
- Rule: $r > 0.75$, $r < -0.75$

- Correlation network
- But, how do I know what are novel (new) stuffs?

Example 2: Statistically collating information about interactions

- The Bibliographical Problem
  - Examples 2 & 3

- Microarray analysis requires knowledge of thousands of genes and proteins

- 1.2 million new papers in 2006, 1.7 million in 2007, 1.6 million in Jan-July 2008
- 18.7 million papers in PubMed today

- ~20% of interaction/localization knowledge in databases

Example 2: Statistically collating information about interactions

- Guilt by association

- If 2 names (protein names) appear in X number of text more than random chance, there is more than random chance that these 2 proteins are related

- The larger X is, the higher probability of relatedness
- RESULT: weighted graph

Example 2: Statistically collating information about interactions

- How to do it?

- Get probability of the 2 names appearing in the same text by random

$$\frac{\#\text{ X mentions}}{\#\text{ text}} \times \frac{\#\text{ Y mentions}}{\#\text{ text}}$$

Example2: Statistically collating information about interactions

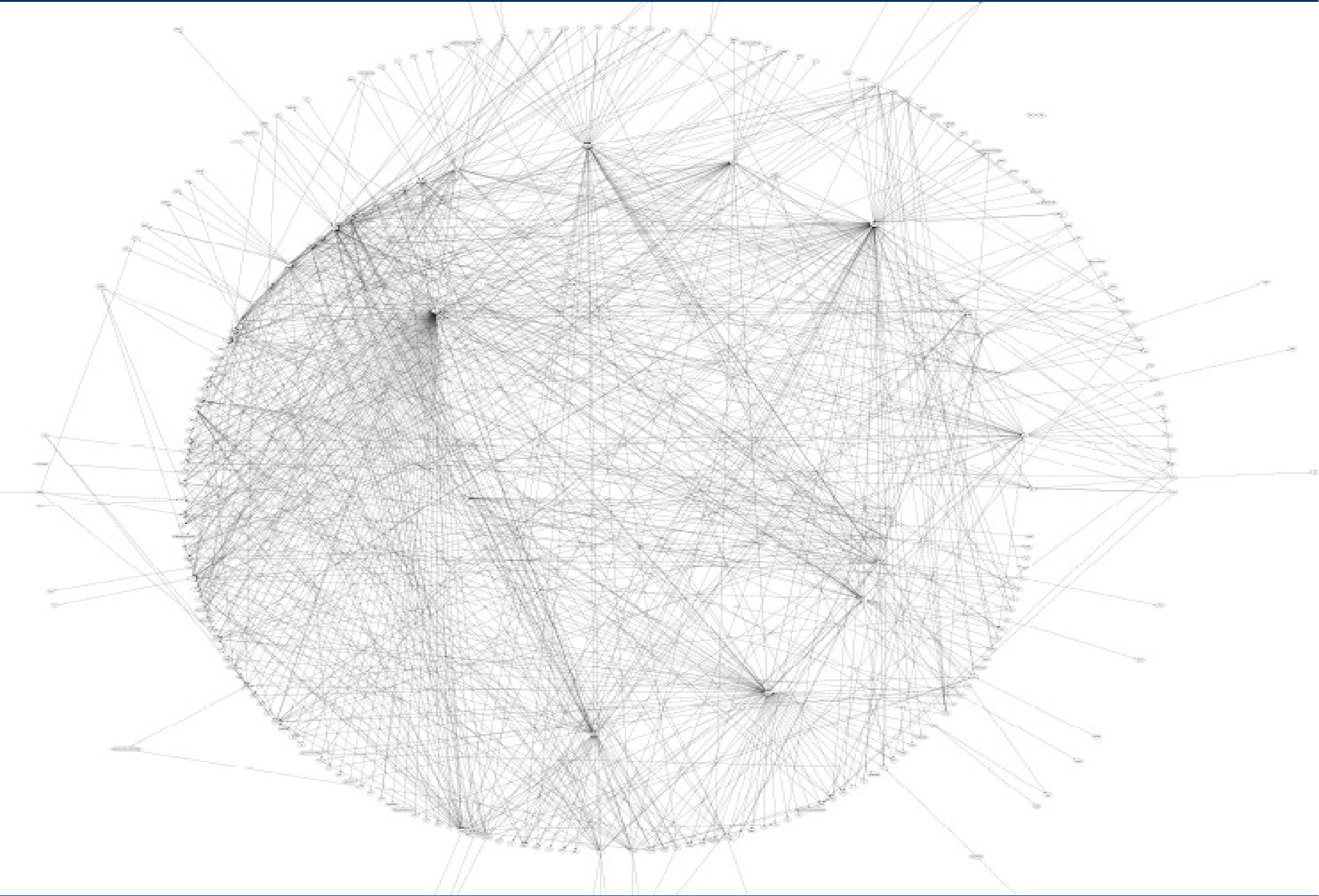
- Model on Poisson distribution
  - Based on binomial distribution for rare events
- Count the number of times these 2 names actually occurs in the same text
- Run tests at 95% or 99% confidence
- Do this for each pair of proteins
- (optional) correct for multiple tests

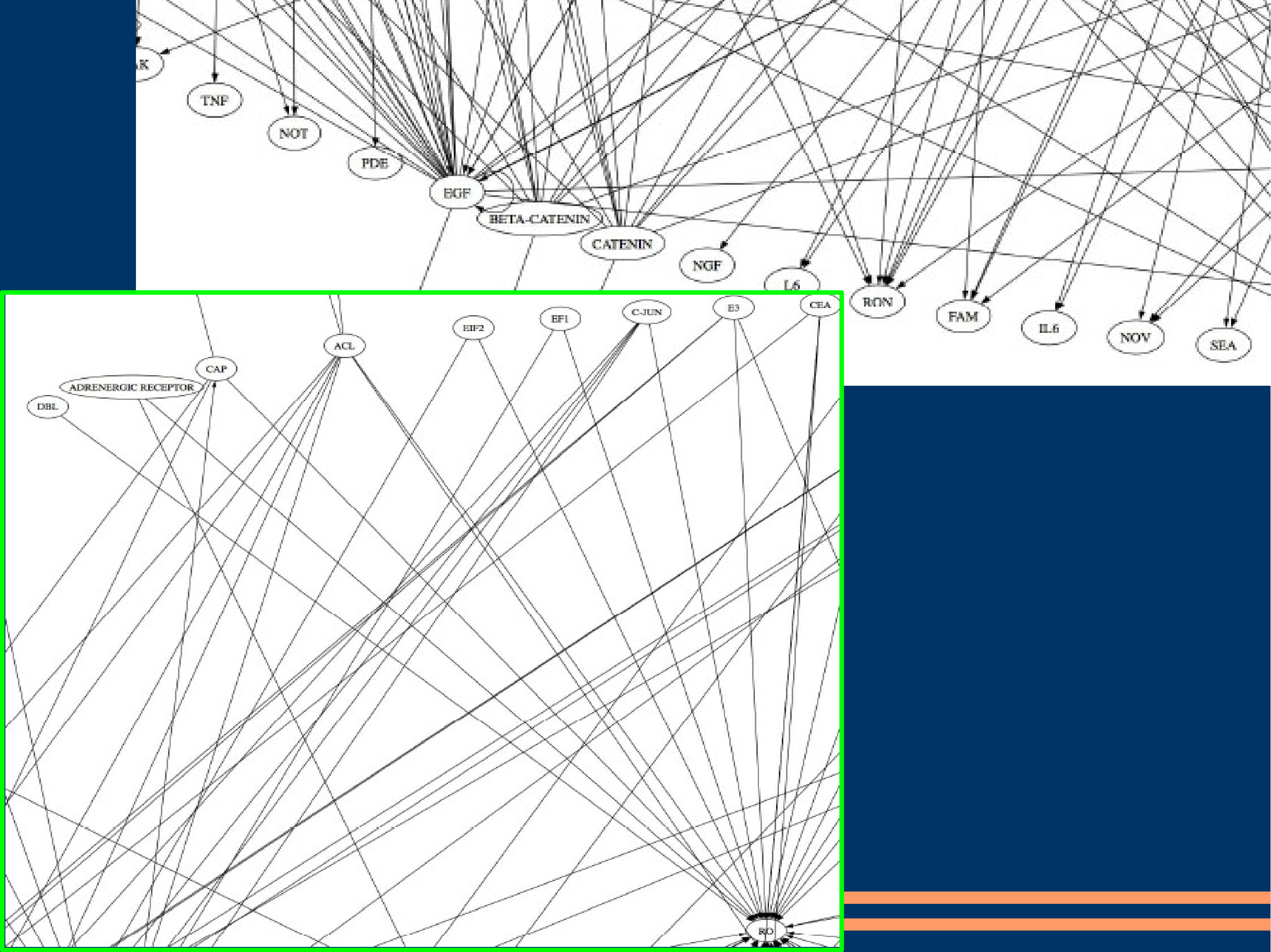
Example 2: Statistically collating information about interactions

- An example
- 1 million papers
  - “insulin” appears in 1000 papers
  - “MAP kinase” appears in 100 papers
- Random probability = 0.0000001 (1 in a million)

- If there are 1 in 1 million, p = 0.5
- If there are 5 in 1 million, p = 0.000001

Network of 1000 Proteins




AK
TNF
NOT
PDE
EGF
BETA-CATENIN
CATENIN
NGF
L6
RON
FAM
IL6
NOV
SEA
EIF2
EF1
C-JUN
E3
CEA
ACL
CAP
ADRENERGIC RECEPTOR
DBL
RO

Example 2: Statistically collating information about interactions

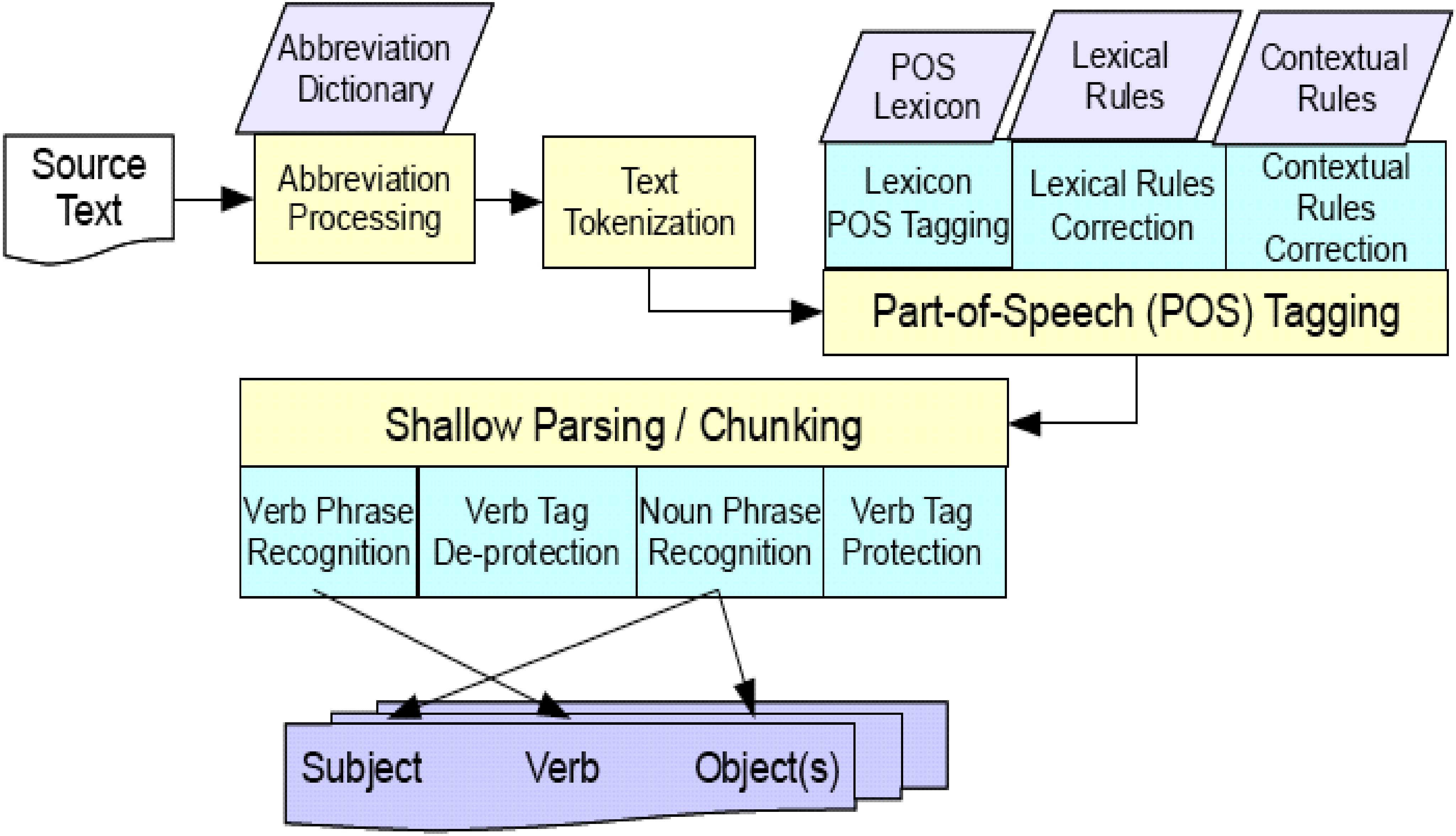
- How good is this?

- 1 mention in 10 million abstracts ~ 65% correct
- 5 in 10 million abstracts ~ 75% correct

- Very easy to increase “correctness”

- Advantage: fast and simple
- Disadvantage: have to know what to look for

Example 3: Computational linguistics information extraction



Example 3: Computational linguistics information extraction

- An Example ...
- Source text
  - Insulin activates phosphoinositide 3'-kinase
- Abbreviated text
  - Insulin activates PI3K
- Part-of-Speech Tagged Text
  - Insulin/NNP activates/VBZ PI3K/NNP
- Chunked Text
  - (NX Insulin/NNP NX) (VX activates/VBZ VX) (NX PI3K/NNP NX)
- Subject-Verb-Object Format
  - ['Insulin', 'activate', 'PI3K']

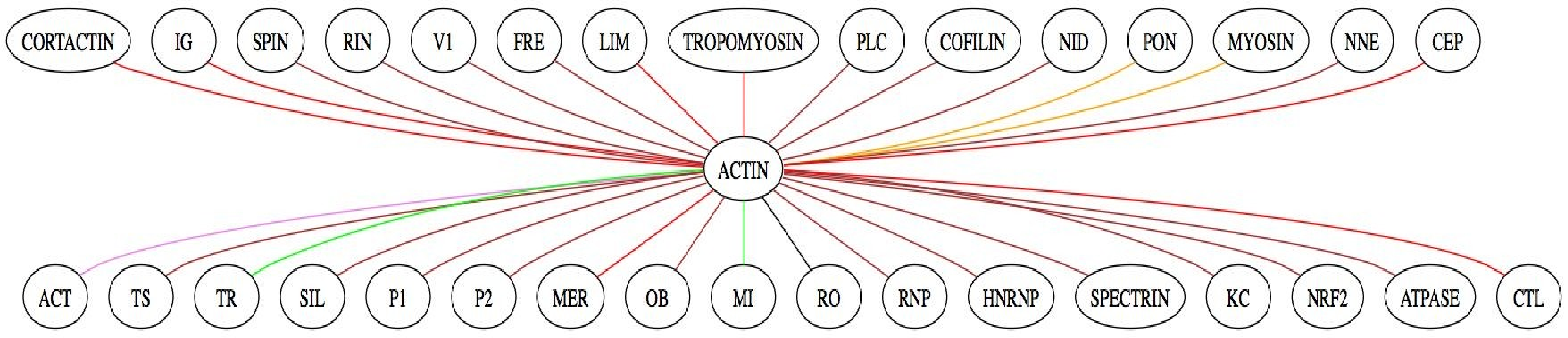
Example 3: Computational linguistics information extraction

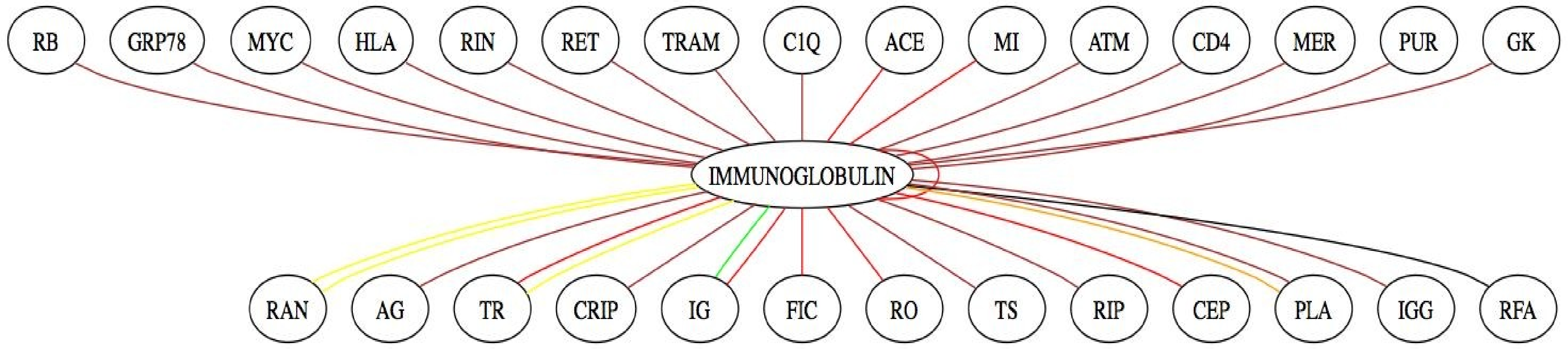
- Why need to abbreviate protein names?
  - Too many variations

- phosphatidylinositol-3-kinase
- phosphatidylinositol-3'-kinase
- phosphatidyl-inositol 3'-kinase
- phosphatidyl-inositol-3-kinase
- phosphatidylinositol 3-kinase
- phosphatidyl inositol 3-kinase
- phosphatidyl-inositol-3 kinase
- phosphatidylinositol 3'-kinase
- phosphatidyl inositol 3' kinase
- phosphatidylinositide 3-kinase
- phosphoinositide 3-kinase
- phosphotidylinositol-3-kinase
- phosphatidylinositol (PI) 3-kinase
- PI 3-kinase
- PI3-kinase
- PI-3K
- PI3K

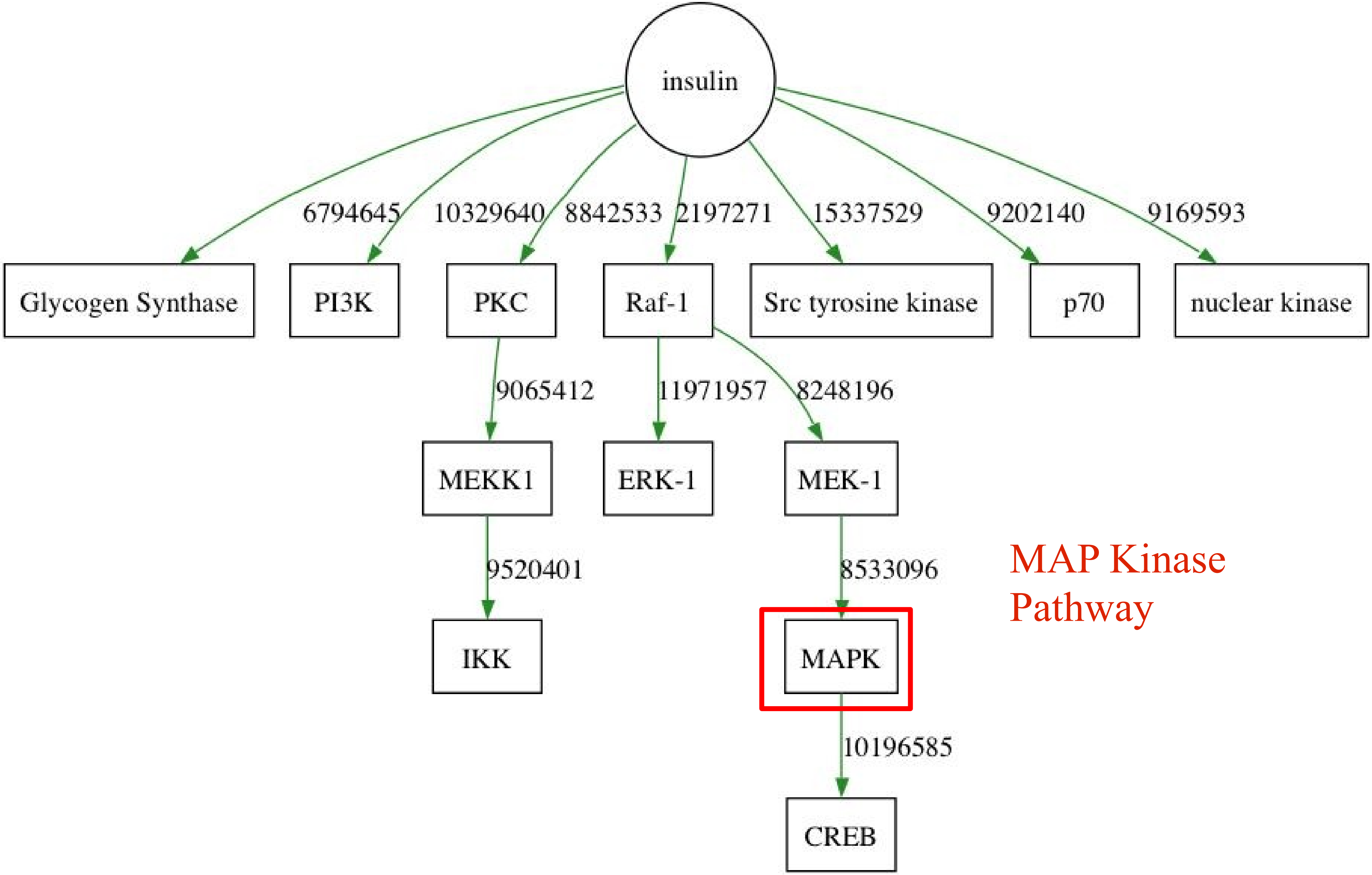
- GENE NORMALIZATION

What binds to Actin or Immunoglobulin?



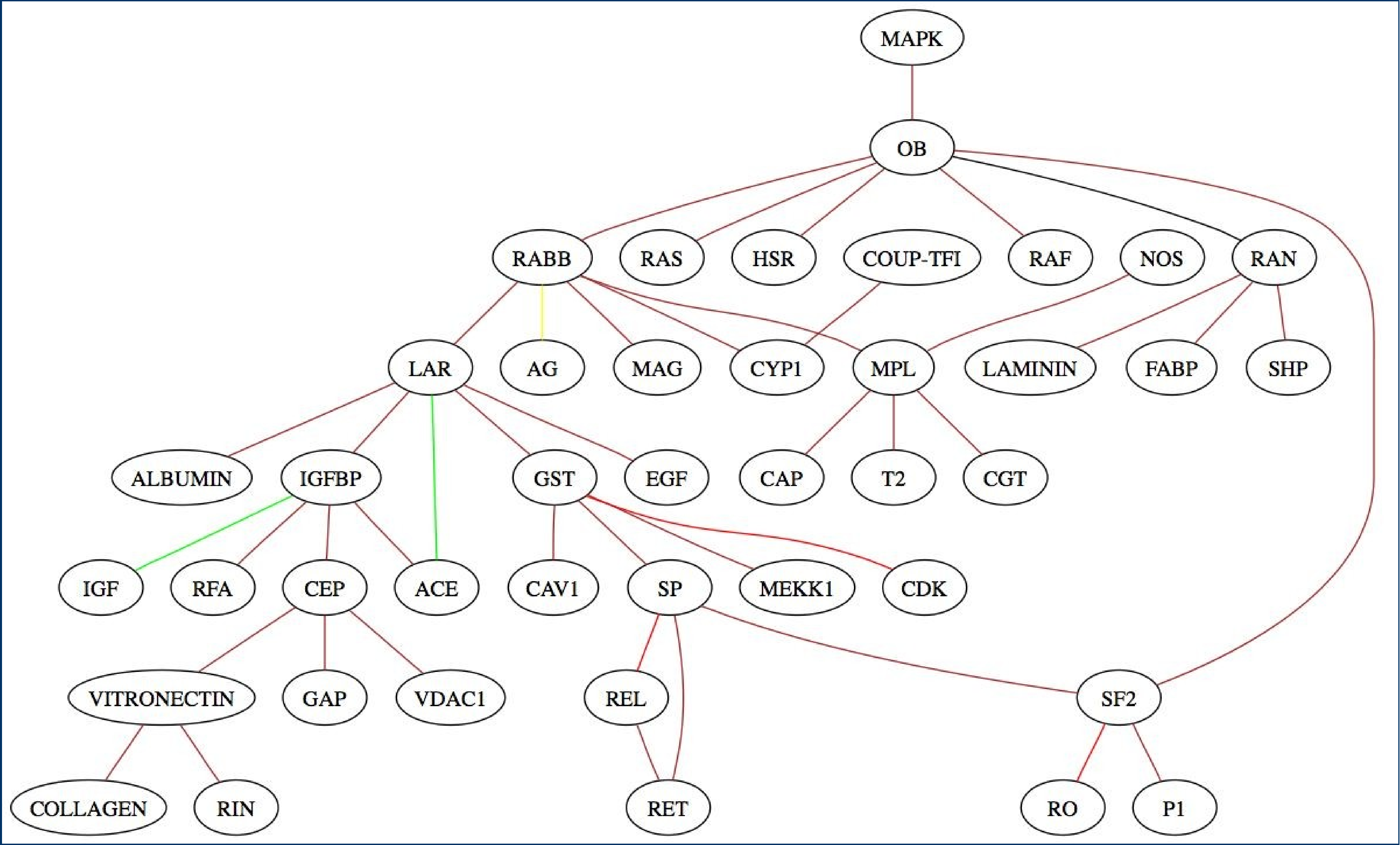


Binding Network of Insulin

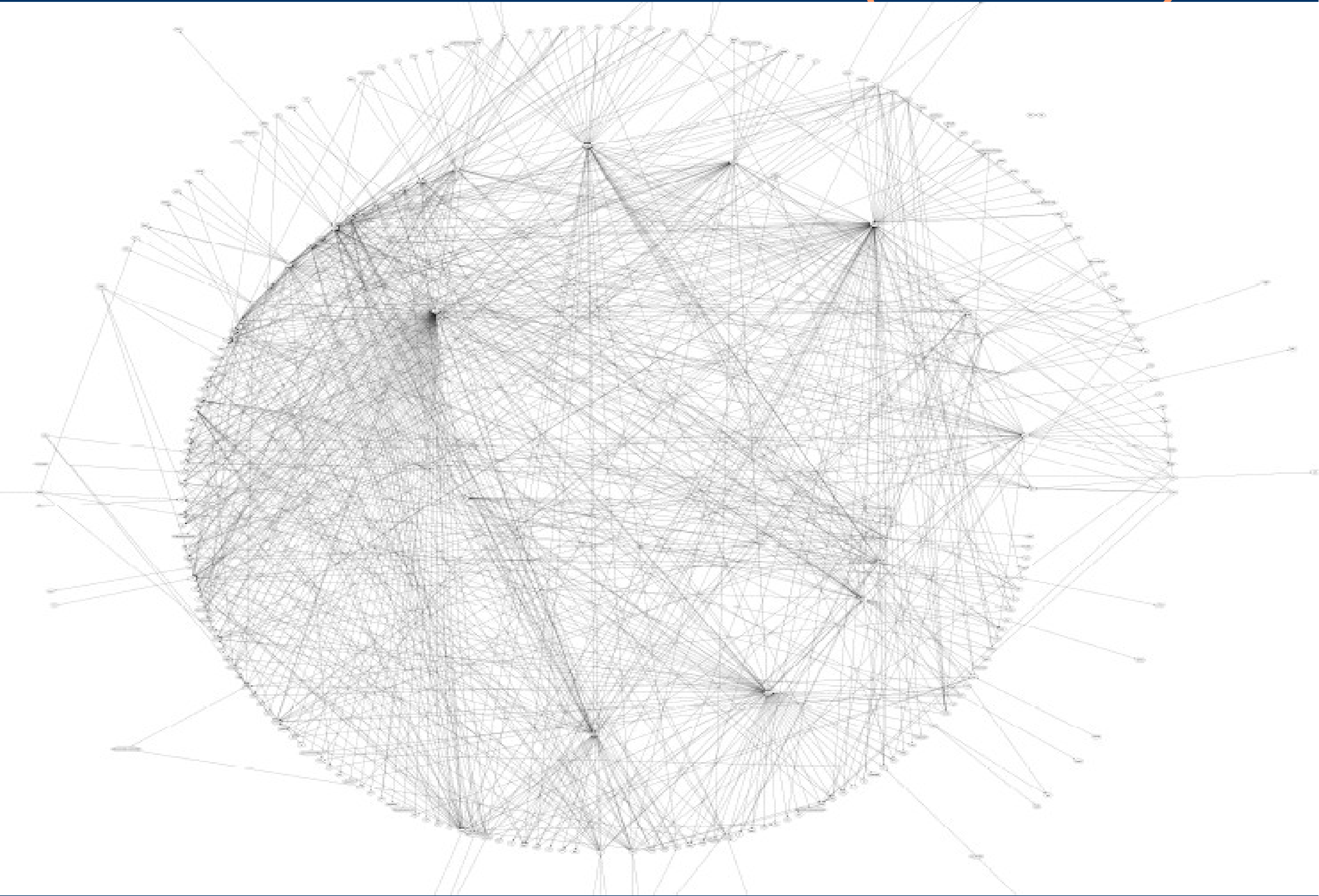


MAP Kinase
Pathway

Binding Network of MAPK



Network of 1000 Proteins (Annotate)



Example 3: Computational linguistics information extraction

- How good is this?

- Computational linguistics is about 90% accurate
  - Compare 60-75% accurate with statistics
  - Can know WHAT is the interaction

Lets merge all the examples

- Correlation map from transcriptomics (microarray)
- Co-occurrence map from bibliomics (literature)

- Use co-occurrence map as a sieve of what are known

- The filtrate are what are not known in the literature
  - perhaps new hypothesis
  - increase the strength by increase correlation coefficient

Epilogue: Information content

- more than 18 million papers published in biology science 1950 to today
  - 1 million in 2005 (about 15 million total)
  - 1.2 million in 2006 (about 16.3 million total)
  - 1.7 million from 2007 (more than 18 million)
- Gene Expression Omnibus (GEO) stores more than 7000 genomic experiments in this century
  - about 181000 biological samples
  - worth about SGD 270 million in expendables
  - about 400 man-years to do the experiments

Epilogue: Information content

- On 22nd August 2005, International Nucleotide Sequence Database Collaboration has 100 gigabases of DNA/RNA sequences
  - 1,000,000,000,000 bases
  - more than 165,000 organisms
  - human has 3,000,000,000 bases
  - 200000 different organisms
  - About 3 million sequences (900,000,000 bases) added every month in 2005

Epilogue: Changing face of Science (Biology)

- Biological research as I was taught a decade ago:
  - get a handle of known knowledge (literature review)
  - devise a research question
  - draw up hypothesis/hypotheses
  - conduct experiment and collect results
  - analyse results
  - how the results fit into the known knowledge

Epilogue: Changing face of Science (Biology)

- Biological research as I know now:
  - get a handle of known knowledge (literature review)
  - devise a research question
  - draw up hypothesis/hypotheses
  - see if what you want is already out there
  - conduct remaining experiment, collect results, and keep an eye on new data released by others
  - analyse results
  - how the results fit into the known knowledge

Epilogue

The improver of natural science absolutely
refuses to acknowledge authority, as such.
For him, skepticism is the highest of duties:
blind faith the one unpardonable sin.
-- Thomas Henry Huxley

If you cannot - in the long run -
tell everyone what you have been doing,
your doing has been worthless.
-- Erwin Schrodinger