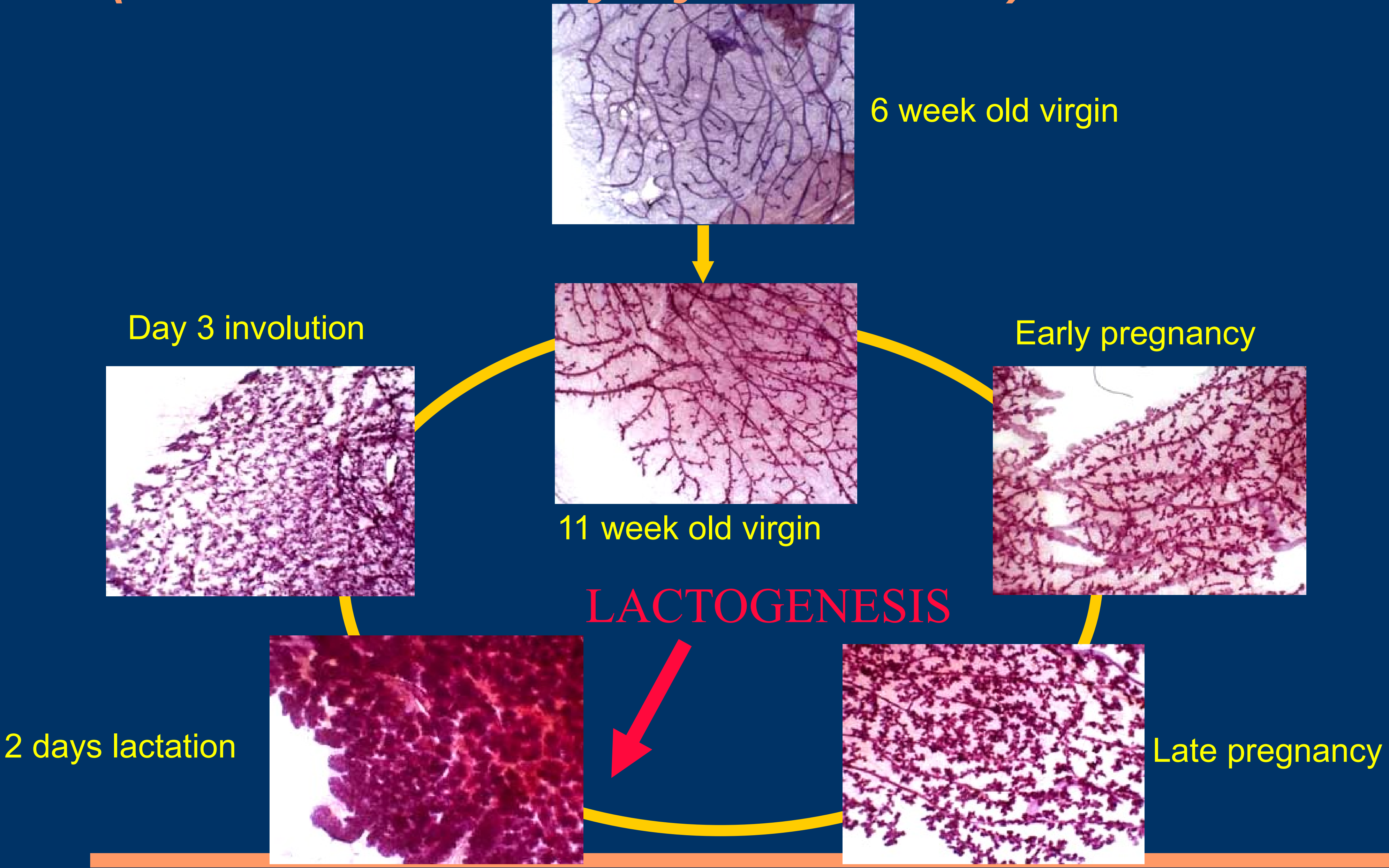
Supporting Data Analysis with Literature Analyses

Maurice Ling

Murine Lactation Cycle
(controlled mainly by hormones)



6 week old virgin

11 week old virgin

Early pregnancy

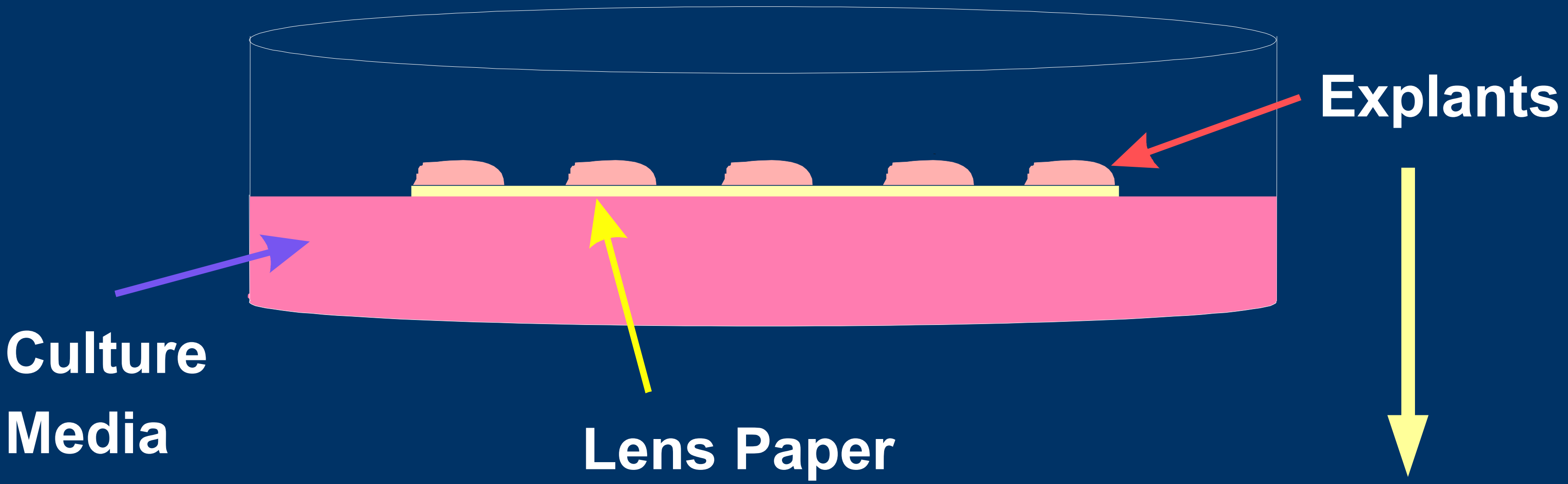
Late pregnancy

LACTOGENESIS

2 days lactation

Day 3 involution

Explant (Tissue) Culture Model



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)

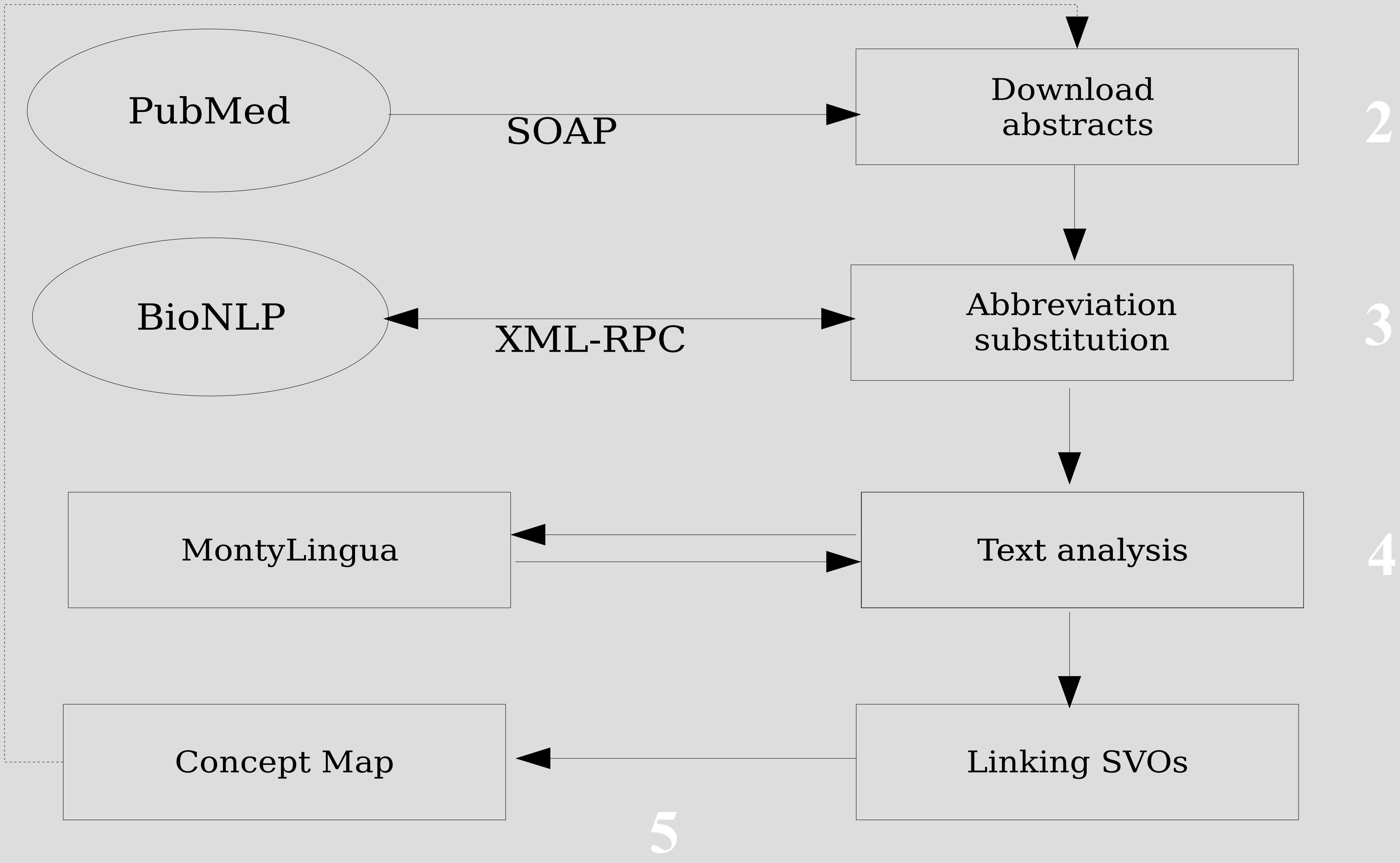
Main problems in lactogenesis

- Is tissue culture representative of the actual animal?
  - It is very very difficult to repeat tissue culture experiments on real mouse
- We infer the actions of insulin, glucocorticoid, and prolactin by expression of maker genes (milk protein genes)
  - No idea about the molecular pathways

My main approach

- 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

Muscorian – pipeline for text analysis



Step 2: Replace “long form” with abbreviations

Previous experiments using dominant negative constructs and gene ablation in mice suggested that two phosphoinositide phosphatases, SH2 domain-containing inositol 5'-phosphatase 2 (SHIP2) and phosphatase and tensin homolog deleted on chromosome 10 (PTEN) negatively regulate this insulin signaling pathway.

Previous experiments using dominant negative constructs and gene ablation in mice suggested that two phosphoinositide phosphatases, SHIP2 (SHIP2) and PTEN (PTEN) negatively regulate this insulin signaling pathway.

Previous experiments using dominant negative constructs and gene ablation in mice suggested that two phosphoinositide phosphatases, SHIP2 and PTEN negatively regulate this insulin signaling pathway.

17 Ways of writing PI3K in 25000 abstracts

- 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

Step 3: Text analysis with MontyLingua

- MontyLingua – text analysis engine by Hugo Liu (MIT Media Labs)
- Does all the normal text processing tasks
  - Tagging – label each word as noun, adverb etc
  - Chunking – breaking sentence into phrases
  - Stemming – reducing words to root word
- Outputs subject-verb-object-objects (SVOOS)

Step 3: Text analysis with MontyLingua

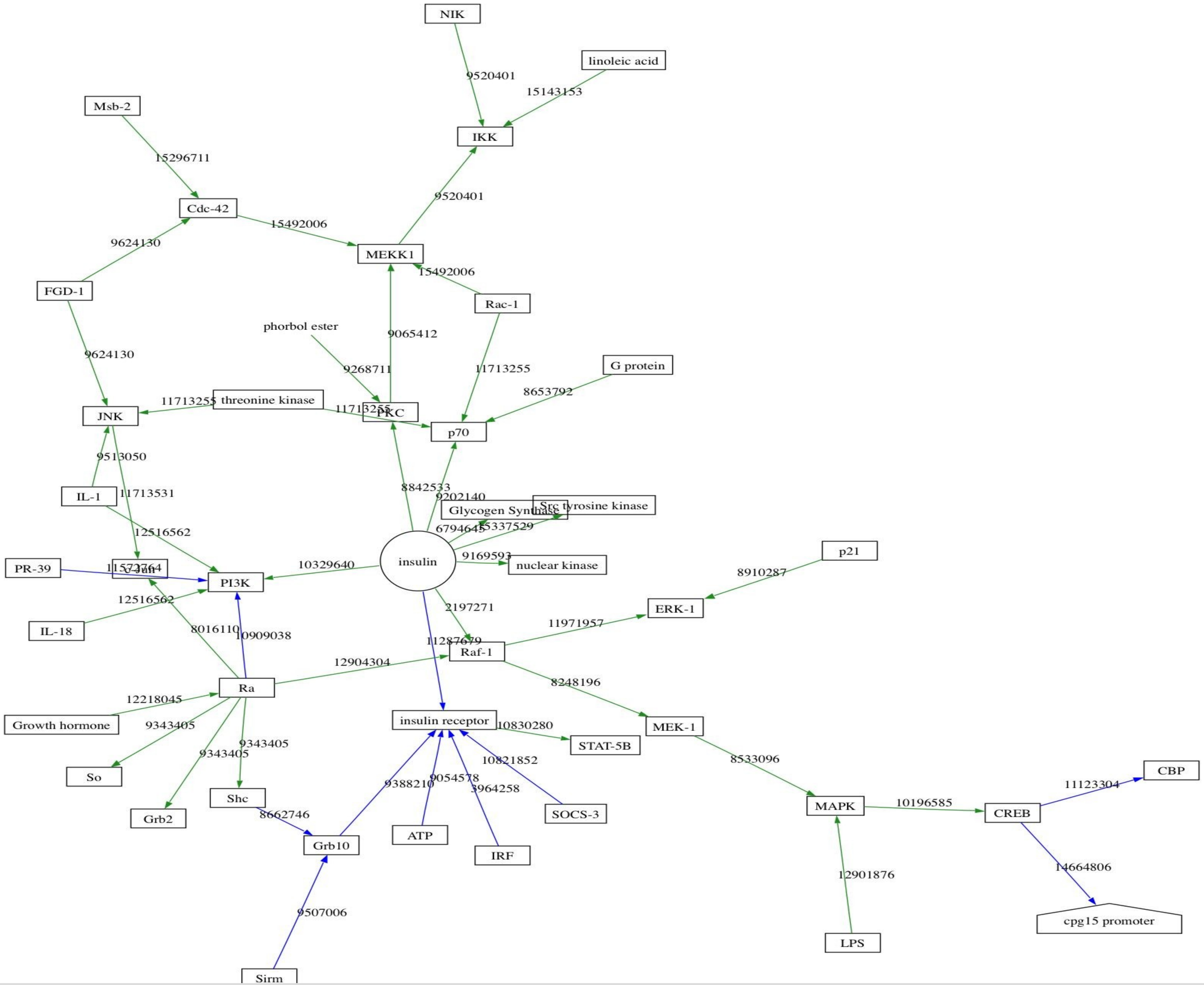
Previous experiments using dominant negative constructs and gene ablation in mice suggested that two phosphoinositide phosphatases, SHIP2 and PTEN negatively regulate this insulin signaling pathway.

("use" "Previous experiment" "dominant negative")
("construct" "dominant negative" "and gene ablation" "in mouse")
("suggest" "mouse" "that two phosphoinositide phosphatase SHIP2 and PTEN")
("regulate" "two phosphoinositide phosphatase SHIP2 and PTEN" "insulin signal pathway")

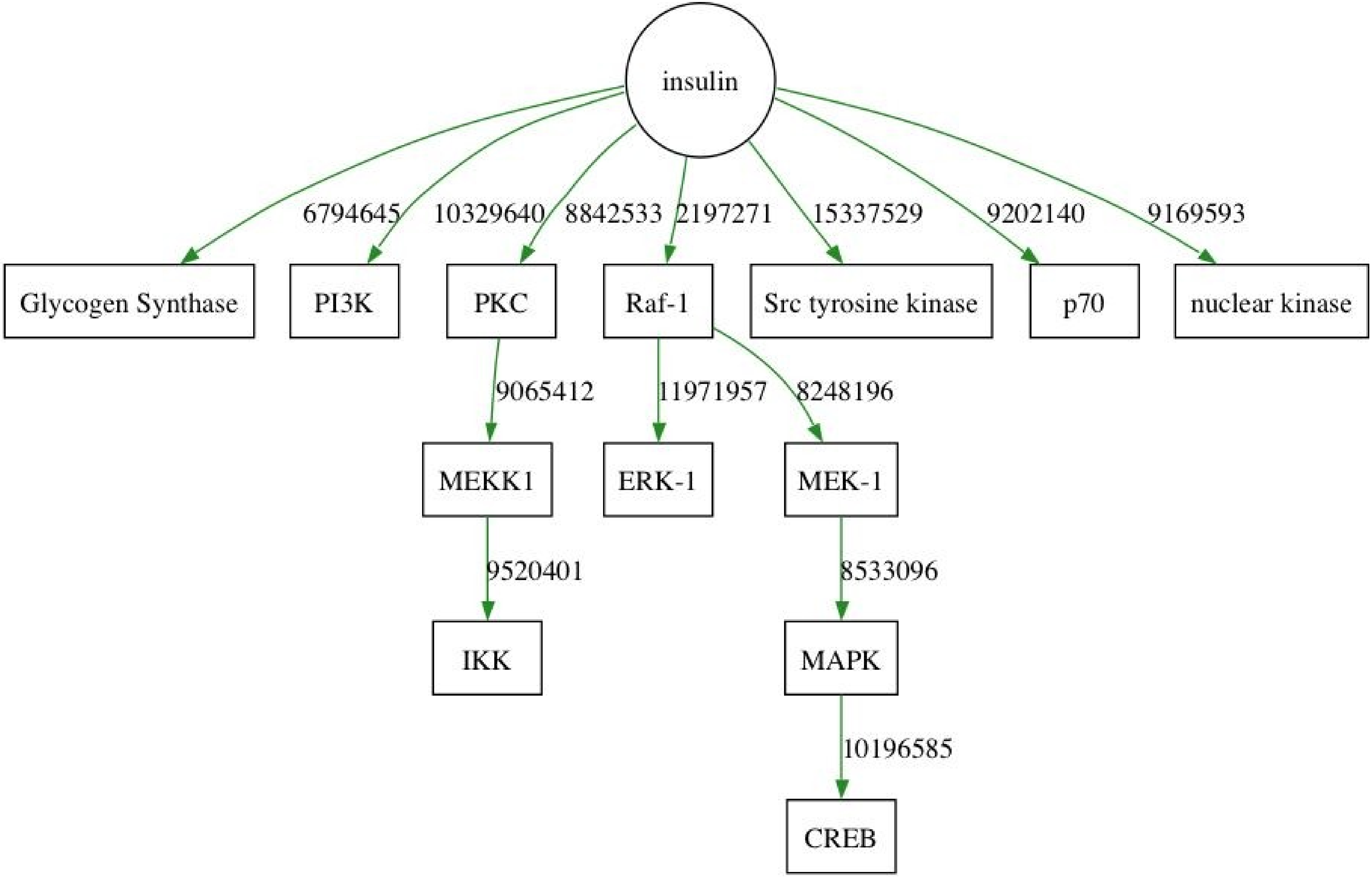
Step 4: Mining Interactions from SVOs

```
subject_protein = object_protein = list of proteins

for subject_protein in protein_entities$_{1 \text{ to } n}$
  for object_protein in protein_entities$_{1 \text{ to } n}$
    insert (pmid, subject_protein, object_protein) into
      entity_bind_SVO
      from select pmid
      from (select * from SVO where verb = 'bind')
      where subject is containing subject_protein
      and object is containing object_protein
return  entity_bind_SVO
```


NIK
linoleic acid
9520401
15143153
Msb-2
IKK
15296711
9520401
Cdc-42
15492006
9624130
MEKK1
15492006
FGD-1
Rac-1
phorbol ester
9065412
9624130
9268711
11713255
G protein
8653792
11713255
threonine kinase
11713255
JNK
PKC
p70
9513050
8842533
IL-1
11713531
9202140
Glycogen Synthase
Src tyrosine kinase
6794645
15337529
12516562
9169593
p21
insulin
PR-39
11573764
c-Jun
10329640
nuclear kinase
8910287
PI3K
12516562
ERK-1
2197271
IL-18
11971957
8016110
10909038
11287679
Raf-1
12904304
Ra
8248196
12218045
Growth hormone
9343405
insulin receptor
10830280
MEK-1
9343405
STAT-5B
9343405
So
10821852
8533096
9054578
9388210
3964258
Shc
11123304
CBP
Grb2
8662746
MAPK
10196585
CREB
SOCS-3
ATP
Grb10
IRF
14664806
12901876
9507006
cpg15 promoter
LPS
Sirm

Trial run of Insulin activation



Evaluation Method

- Using a known dataset
  - Compares the output of a system to model answers
- Sample the output and compare with original
- Have a set of inputs and examine the outputs

- Precision = # correct output / # output
- Recall = # correct output / # correct in dataset
- 1 – precision ==> wrong information
- 1 – recall ==> missing information

How good is my system?

- Testing using known dataset
  - 86% precision
  - 31% recall
- Sample output of extracted protein-protein binding
  - n = 135
  - 88% precision

Large Scale Binding Study (I)

- Given ~3600 proteins, there is a possibility of ~13 million (3600^2) binding interactions
- If more than 1 abstract suggest the same interaction, higher chance of correct interaction

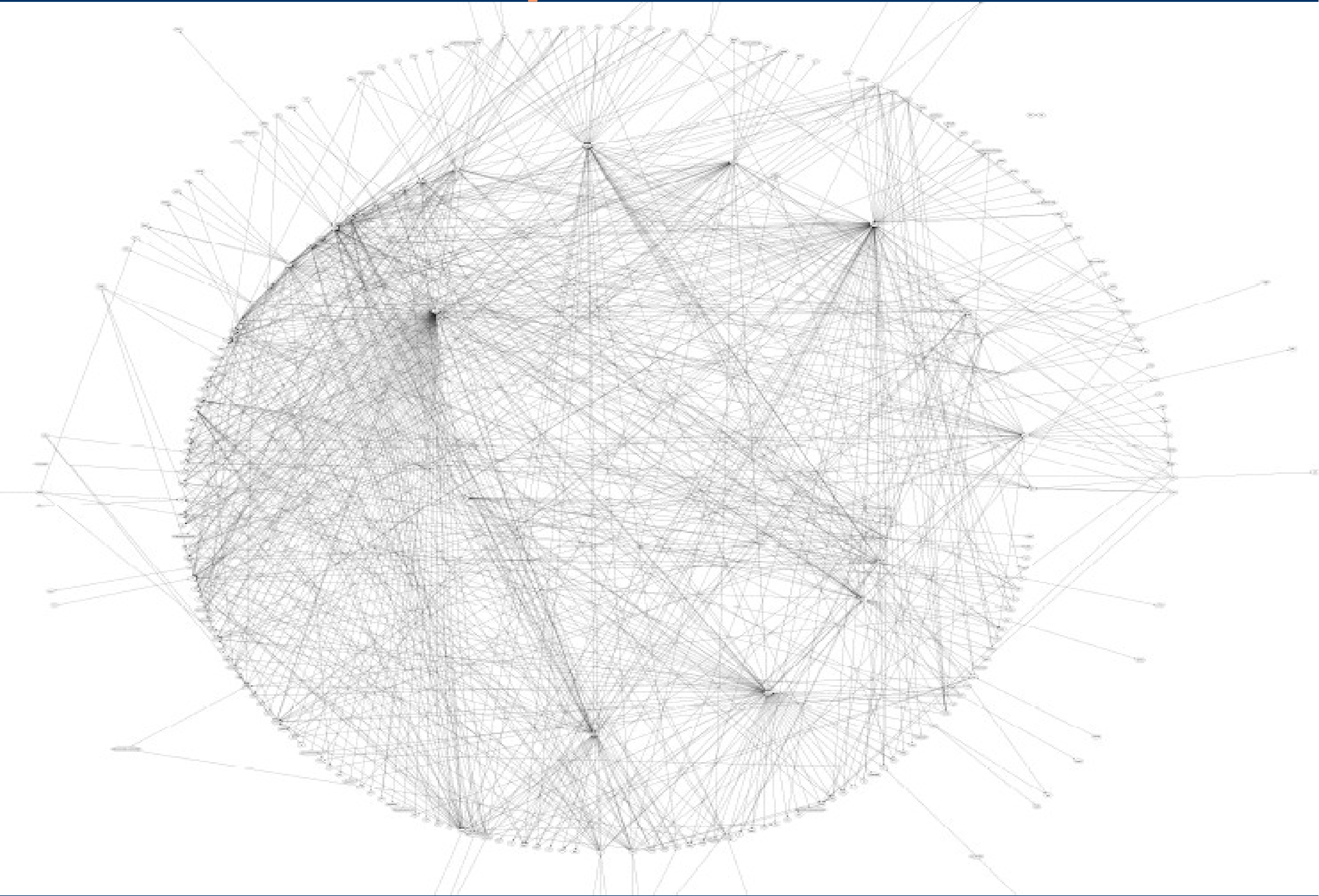
- P(1 male in 1 birth) = 1 – P(1 female in 1 birth) = $1 – 0.5^1 = 0.5$
- P(at least 1 male in 2 births) = 1 – P(both females) $= 1 – 0.5^2 = 0.75$
- P(at least 1 male in 3 births) = $1 – 0.5^3 = 0.875$

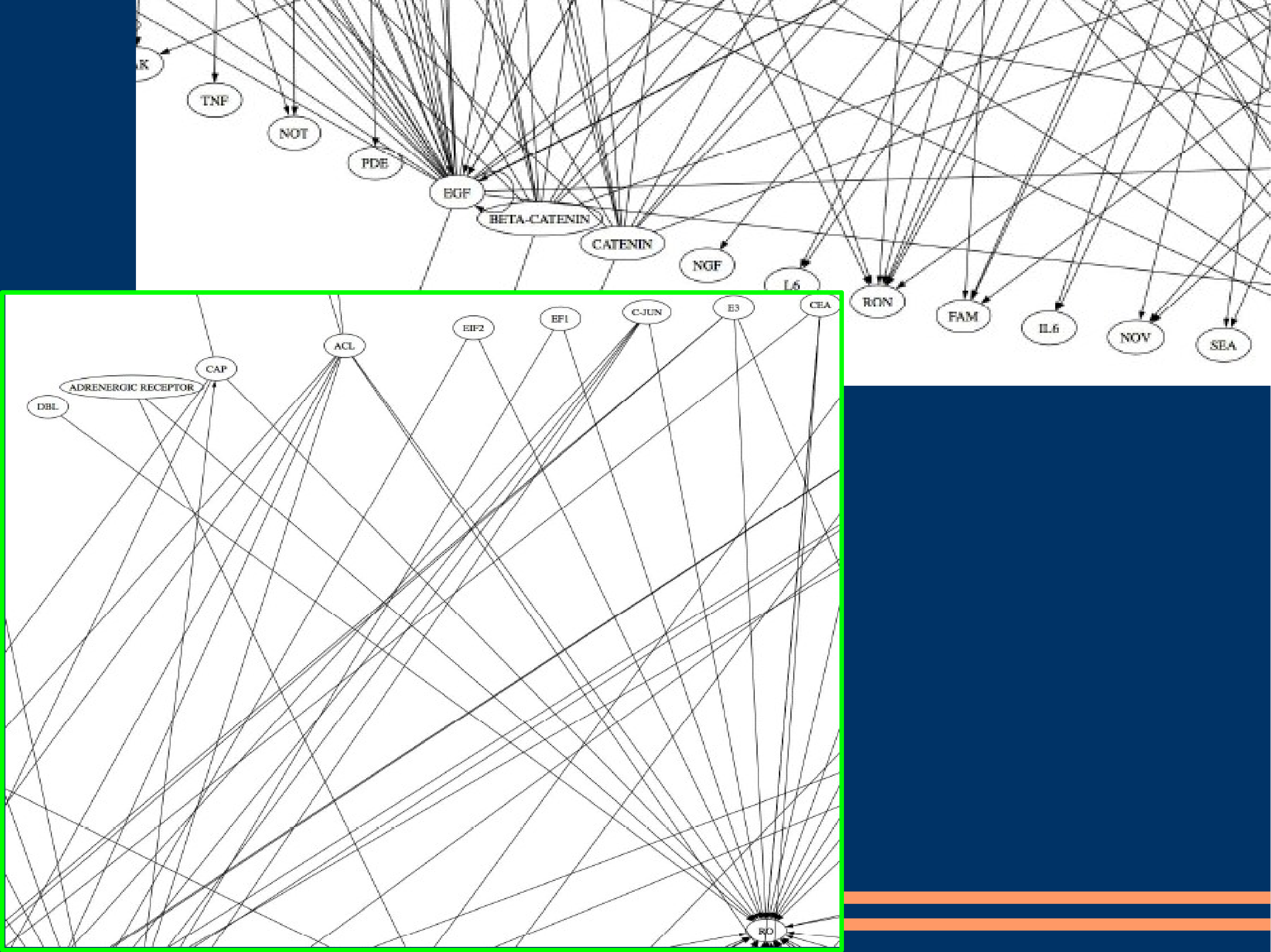
Large Scale Binding Study (II)

- P(binding interaction that is correct) = 0.82
- P(binding interaction that is wrong) = 0.18
- P(at least 1 correct in 2) = 96.8%
- P(at least 1 correct in 3) = 99.4%

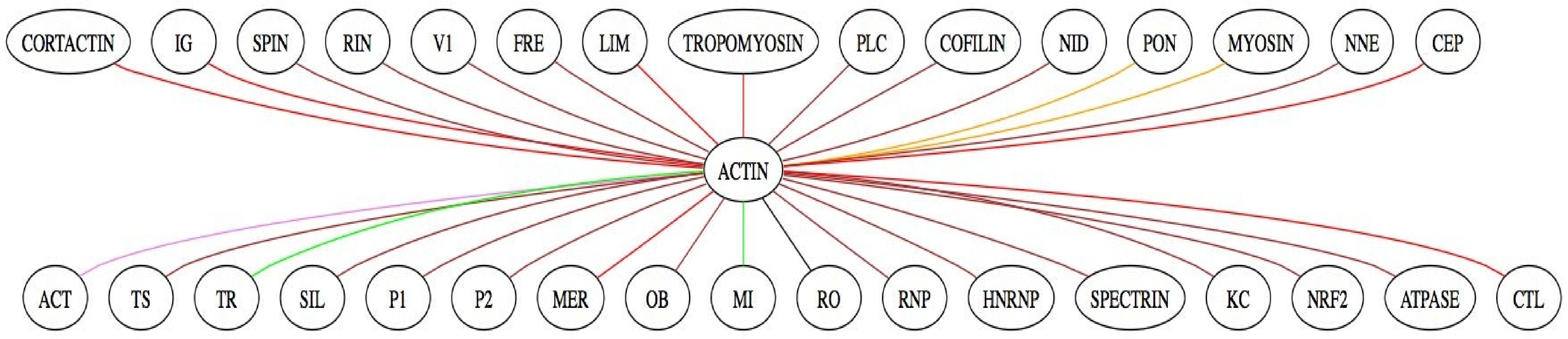
# of multiples	1	2	3	4	5	>5
P(at least 1 correct)	82.00%	96.40%	99.40%	99.90%	99.98%	>99.99%
# interactions	7049	1297	516	235	164	572

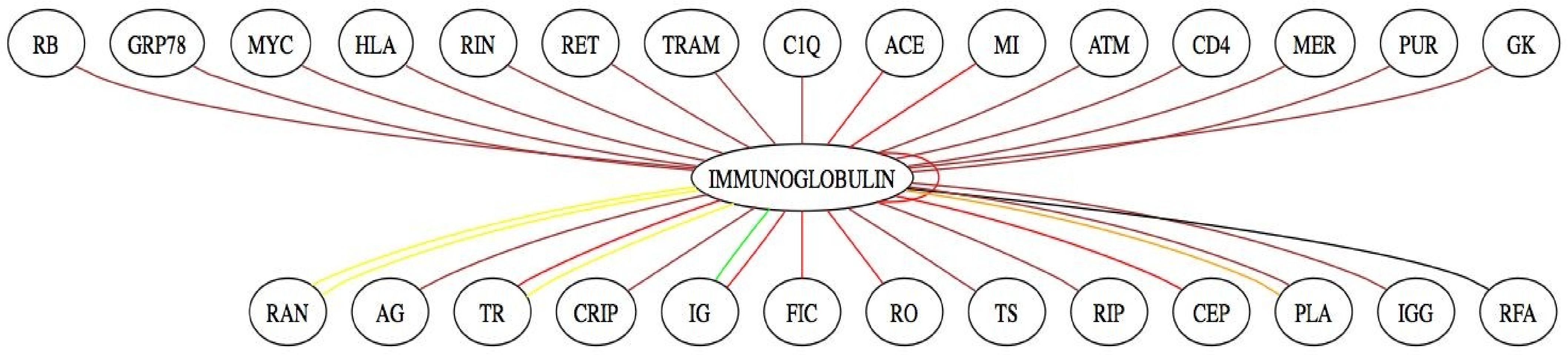
Activation Map of 1000 Proteins



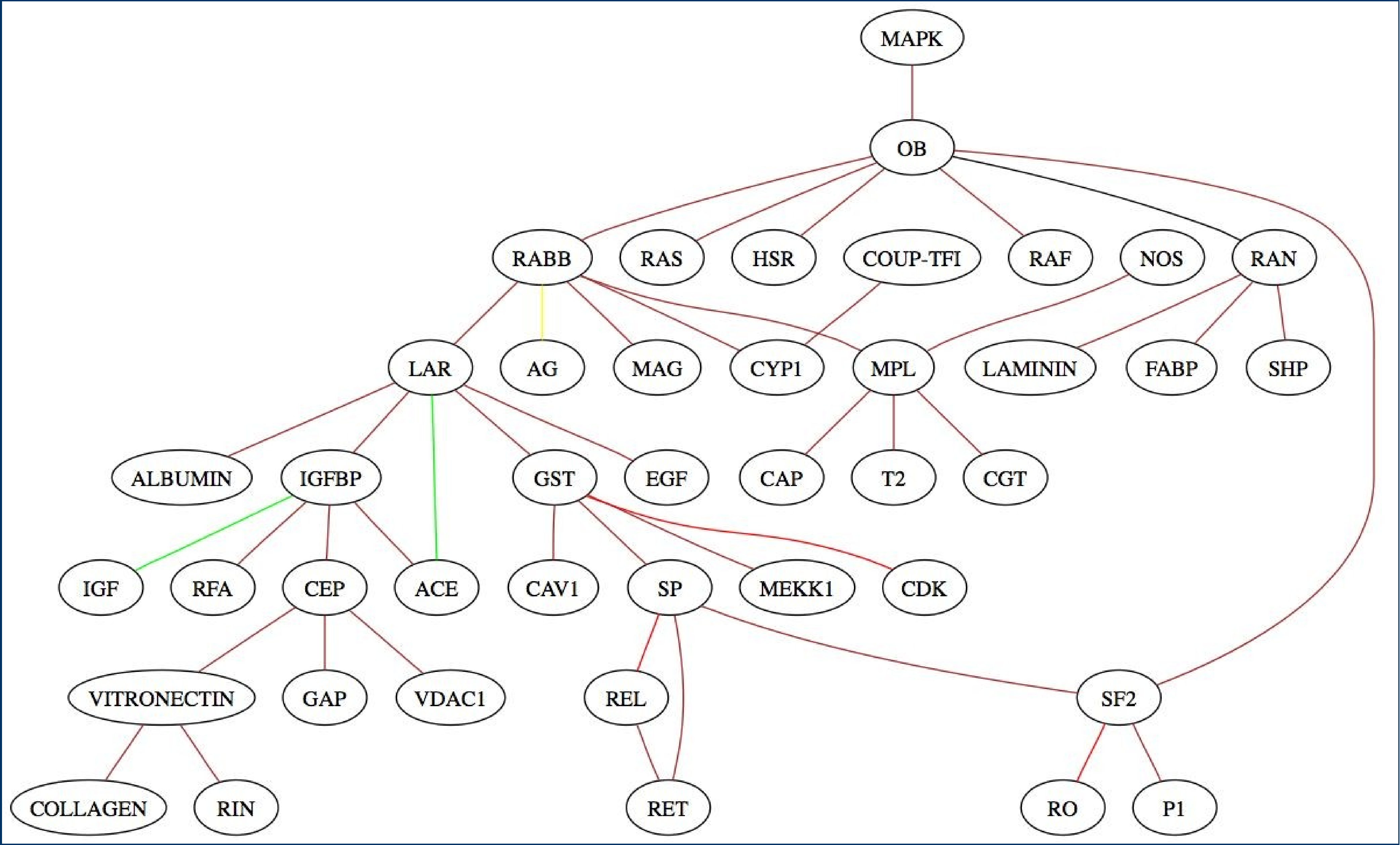

AK
TNF
NOT
PDE
EGF
BETA-CATENIN
CATENIN
NGF
L6
RON
FAM
IL6
NOV
SEA
DBL
ADRENERGIC RECEPTOR
CAP
ACL
EIF2
EF1
C-JUN
E3
CEA
RO

What binds to Actin or Immunoglobulin?



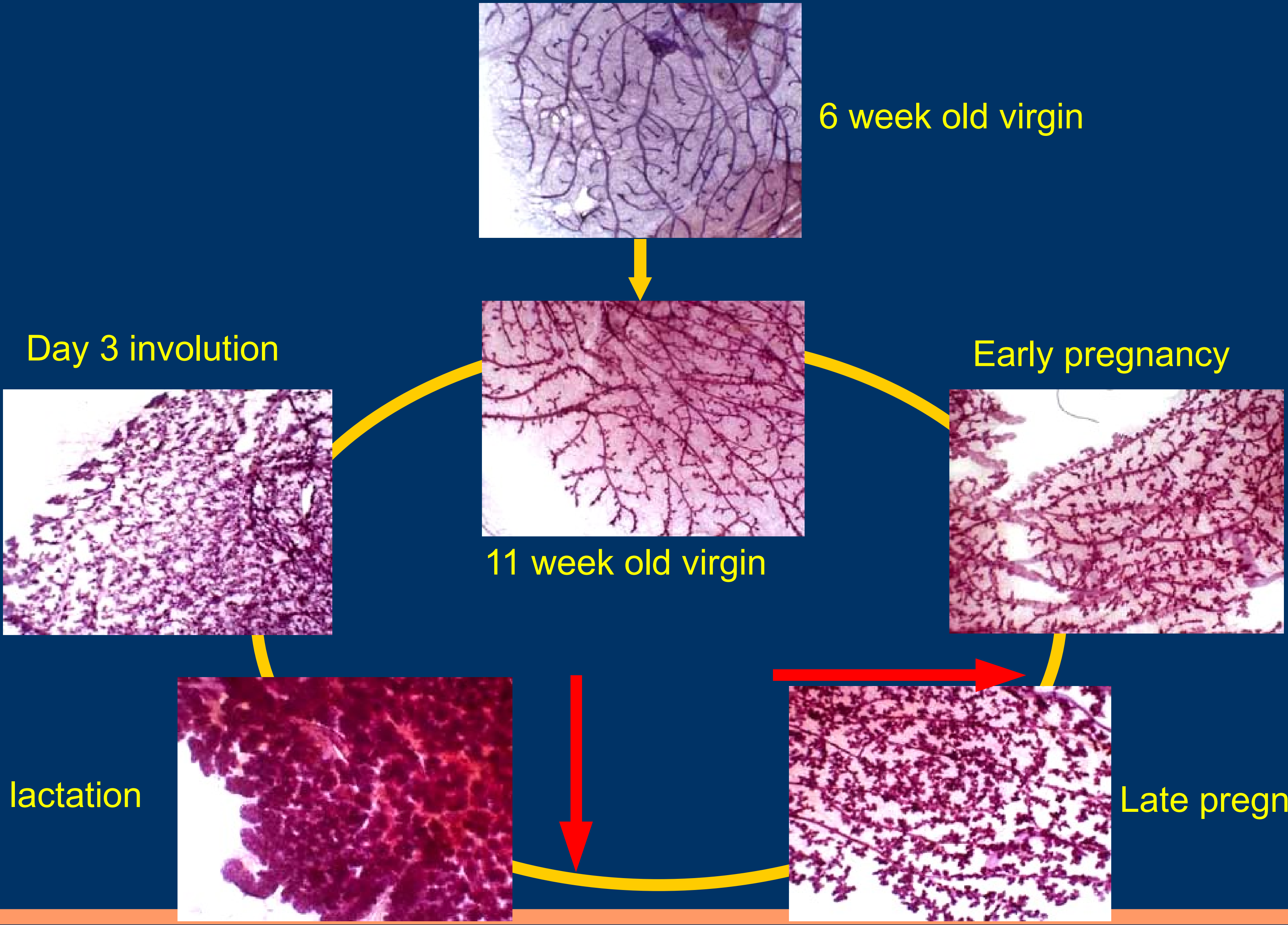


Binding Network of MAPK



Briefly describe microarray work (work in progress)

4 Sets of Microarrays (Public)



6 week old virgin

Day 3 involution

Early pregnancy

11 week old virgin

2 days lactation

Late pregnancy

From in vivo (animal) microarrays...

- Get an idea of the major biological processes in the mouse mammary gland
  - Pearson's co-expression network
  - Cluster co-expression network (MCODE in Cytoscape)
  - Use Gene Ontology over-representation
- Get co-expression network that is differentially expressed between Day 17 Pregnancy and Day 1 Lactation
- What are the major (statistically) biological processes that happens between Day 17 Pregnancy and Day 1 Lactation?

From in vitro (tissue culture) microarrays...

- Get a list of differentially expressed genes between control and hormone treatment
- Look at what these mean by Gene Ontology over-representation
- Compare it with Day 17 Pregnancy and Day 1 Lactation...
- Use mined protein-protein interaction maps to have an idea of what is already known
  - Hypothesize what is not known
  - Strengths and limitations of tissue culture?

Thank you for listening

Acknowledgements

- My supervisors
  - Kevin Nicholas, Christophe Leferve, Andrew Lonie, Lin Feng
- CRC for Innovative Dairy Products
- Dept. of Zoology, The University of Melbourne
- Institute of Bioinformatics, National Yang Ming University
  - 郭政儒， 鐘翊方