Pathway Reconstruction

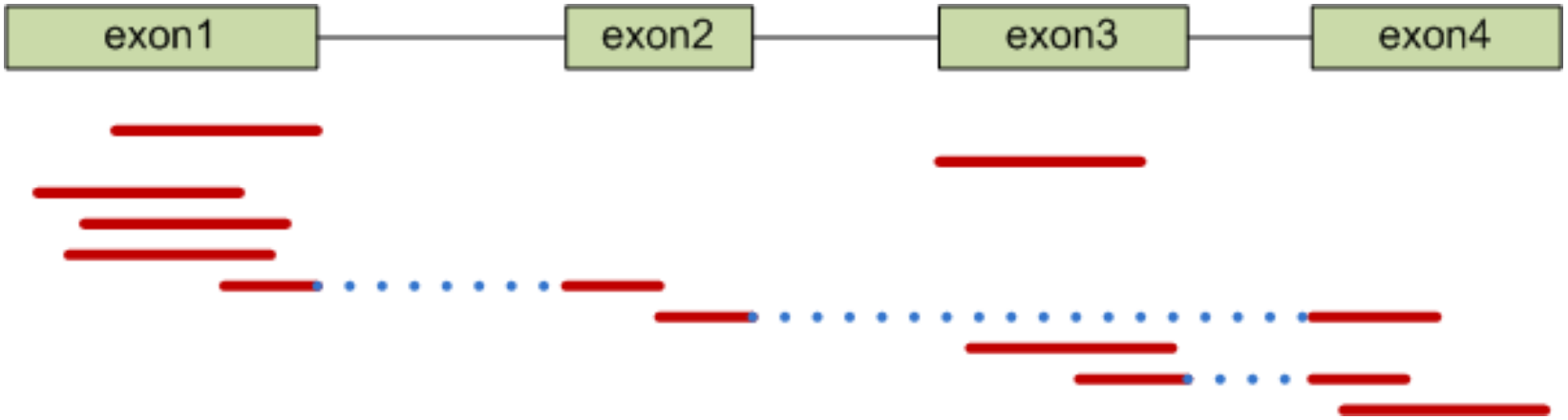
CENG 734

Advanced Topics in Bioinformatics

Fall 2011-2012

Quiz #1 (from IsoLasso paper)

- Consider the following gene and the mapped single end reads obtained from a next generation sequencing machine. The mapped sequences are shown in red lines.

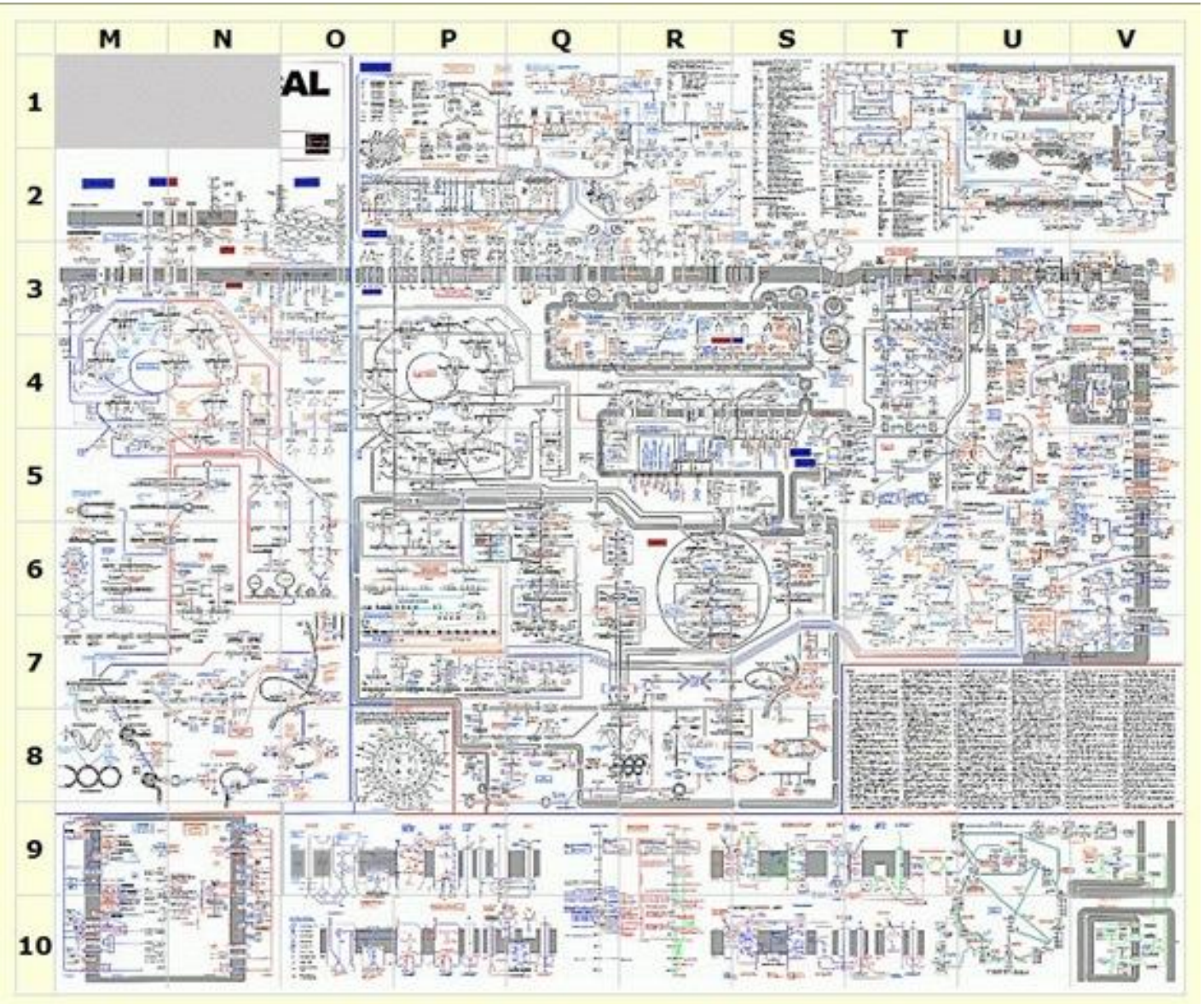


- Create the connectivity graph between the exons based on the two conditions described in the IsoLasso paper.
- Which isoforms can be inferred from this connectivity graph based on the algorithm described in the IsoLasso paper?
- Make a guess on which isoform is expressed at the highest level. Justify your guess.

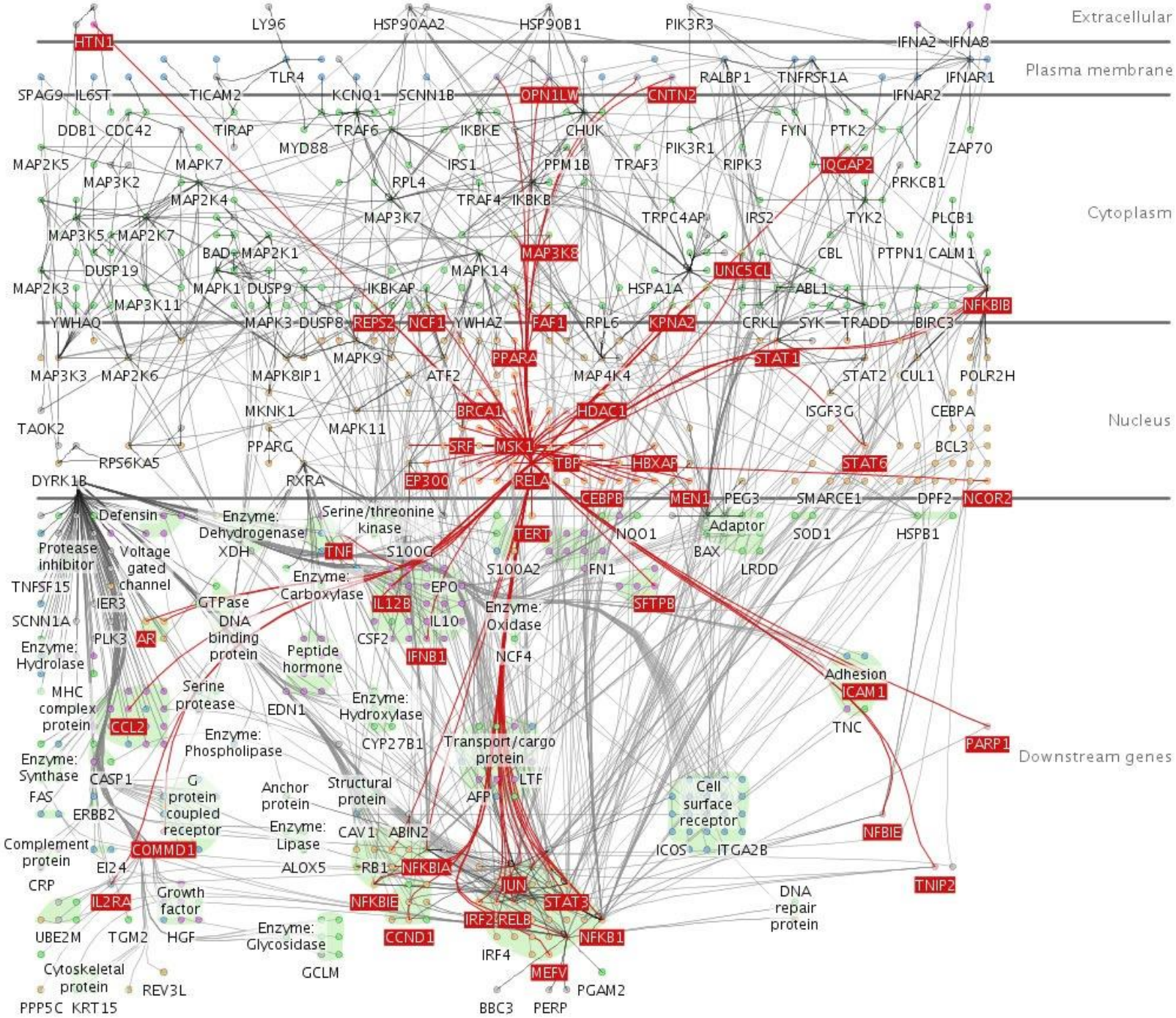
Systems Biology

- The wiring diagram of the cell
  - Highly dynamic
  - Very complex

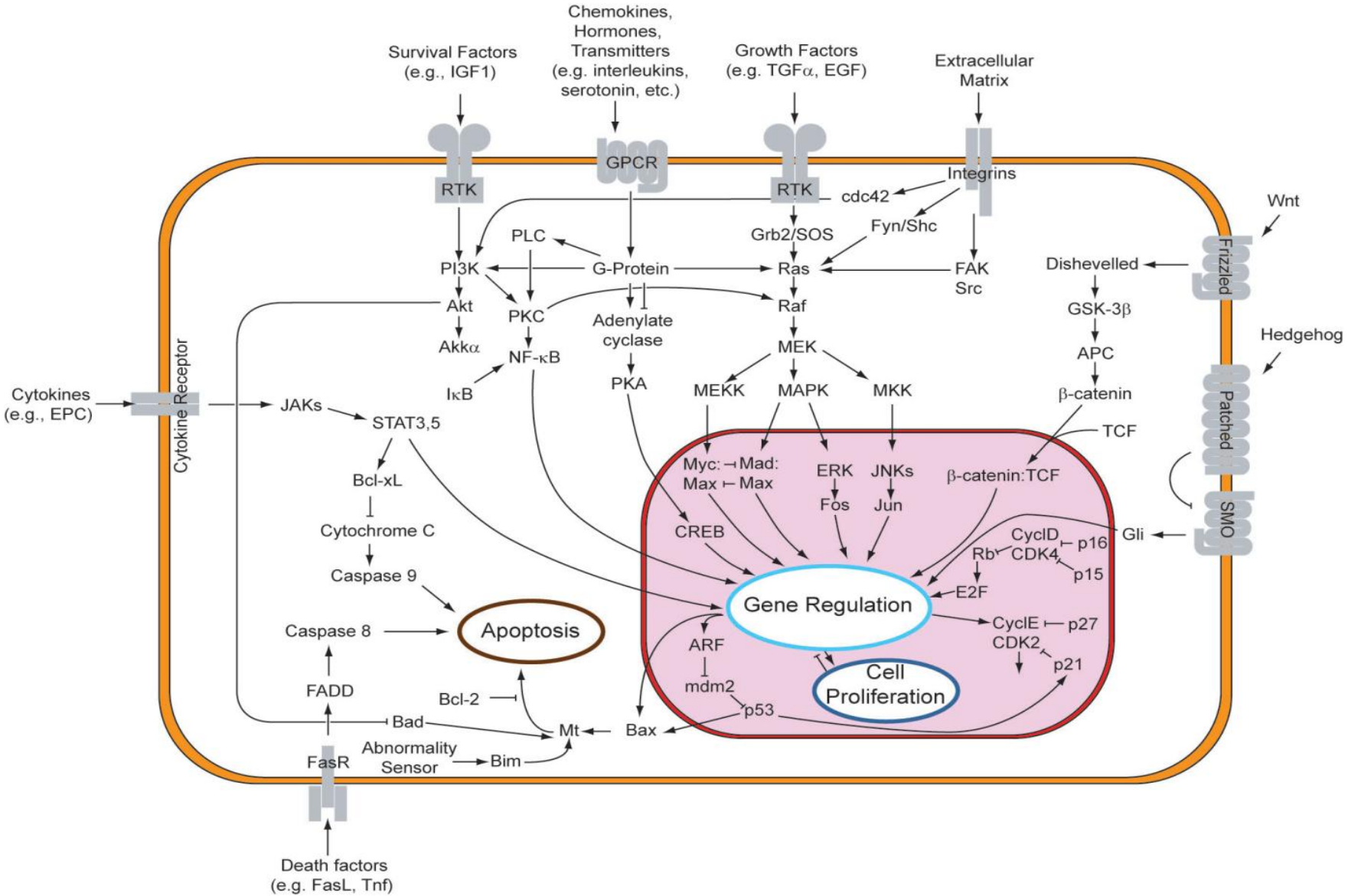
Systems biology



Roche Applied Science's "Biochemical Pathways" series of wall charts.


Extracellular
Plasma membrane
Cytoplasm
Nucleus
Downstream genes
HTN1
LY96
HSP90AA2
HSP90B1
PIK3R3
IFNA2
IFNA8
IFNAR1
IFNAR2
TLR4
RALBP1
TNFRSF1A
SPAG9
IL6ST
TICAM2
KCNQ1
SCNN1B
OPN1LW
CNTN2
DDB1
CDC42
TIRAP
TRAF6
IKBKE
CHUK
FYN
PTK2
ZAP70
MYD88
PIK3R1
MAP2K5
MAPK7
IRS1
PPM1B
TRAF3
RIPK3
IQGAP2
MAP3K2
RPL4
PRKCB1
MAP2K4
TRAF4
IKBKB
MAP3K7
TRPC4AP
IRS2
TYK2
PLCB1
MAP3K5
MAP2K7
BAD
MAP2K1
MAP3K8
CBL
PTPN1
CALM1
DUSP19
UNC5CL
MAP2K3
MAPK1
DUSP9
IKBKAP
MAPK14
HSPA1A
ABL1
NFKBIB
MAP3K11
YWHAQ
MAPK3
DUSP8
REPS2
NCF1
YWHAZ
FAF1
RPL6
KPNA2
CRKL
SYK
TRADD
BIRC3
MAPK9
PPARA
STAT1
MAP3K3
MAP2K6
MAPK8IP1
ATF2
MAP4K4
STAT2
CUL1
POLR2H
MKNK1
BRCA1
HDAC1
ISGF3G
CEBPA
TAOK2
MAPK11
SRF
MSK1
BCL3
PPARG
RPS6KA5
TBP
HBXAP
STAT6
DYRK1B
RXRA
EP300
RELA
CEBPB
MEN1
PEG3
SMARCE1
DPF2
NCOR2
Defensin
Enzyme: Dehydrogenase
Serine/threonine kinase
TERT
NQO1
Adaptor
SOD1
HSPB1
Protease inhibitor
Voltage gated channel
XDH
TNF
S100G
S100A2
FN1
BAX
LRDD
TNFSF15
Enzyme: Carboxylase
EPO
IER3
GTPase
IL12B
Enzyme: Oxidase
SFTPB
SCNN1A
IL10
DNA binding protein
PLK3
AR
CSF2
IFNB1
NCF4
Enzyme: Hydrolase
Peptide hormone
Adhesion
ICAM1
MHC complex protein
Serine protease
EDN1
Enzyme: Hydroxylase
CCL2
TNC
PARP1
Enzyme: Phospholipase
CYP27B1
Transport/cargo protein
Enzyme: Synthase
CASP1
LTF
FAS
G protein coupled receptor
Anchor protein
Structural protein
AFP
Cell surface receptor
ERBB2
CAV1
ABIN2
NFBIE
Complement protein
Enzyme: Lipase
COMMD1
ICOS
ITGA2B
EI24
ALOX5
RB1
NFKBIA
TNIP2
CRP
JUN
DNA repair protein
IL2RA
Growth factor
NFKBIE
STAT3
IRF2
RELB
UBE2M
TGM2
HGF
Enzyme: Glycosidase
CCND1
NFKB1
IRF4
GCLM
MEFV
Cytoskeletal protein
REV3L
PGAM2
PPP5C
KRT15
BBC3
PERP

Signaling pathways



This week

- SPINE: a framework for signaling-regulatory pathway inference from cause-effect experiments by Ourfali et al. ISMB/ECCB 2007

The problem

- Given a directed network of protein-protein and protein-DNA interactions and knockout-effect data in terms a gene expression experiments assign a sign (activation/repression) to each node or edge in the network that will best explain the gene expression experiments.

Definitions

- Knockout-pair: If a gene i is effected in terms of expression level (p-value ≤ 0.02) when a gene j is knocked out, i and j are named as a knockout pair.
- The directed network of interactions contain PPI and protein-DNA interactions and each interaction is assigned a reliability, $r(e)$, and a sign, $sgn(e)$, (the sign is the output of the method proposed in the paper)
  - 0: activation
  - 1: repression

Definitions

- $e(s,t)$ indicates the effect of gene s on gene t on wild type, i.e., when no gene is knocked out.
- A path π starting with s and ending with t and consisting of k edges explains a knockout pair (s,t) if
  - π is a path on the input graph with no cycles
  - the last edge on the path is a protein-DNA interaction
  - an intermediate gene on the path and the last gene, i.e., (p_i,t) is also a knockout pair.

If a path π satisfies these conditions it is called the k-explanatory of a knockout pair (s,t)

Definitions

- For a path π, $s(\pi)$ and $t(\pi)$ indicate its source and target.
- N_π is the indicator variable denoting whether the path explains its source and target as a knockout pair.
- An explanatory path is also said to be *consistent* if it satisfies:
  - the aggregate sign (addition mod 2) along the path is equal to $e(s,t)$
  - Every suffix of that path is also consistent wrt its source and target.

Definitions

- M_π is the indicator variable denoting whether an explanatory path is consistent
- The sign is defined for an edge but in the paper they assign signs to nodes, meaning that all the edges outgoing from that node have the same sign as the node. This approach reduces the number of variables to be inferred.

Optimization problem

- Find an assignment of signs to nodes/edges in the input network that will maximize the **expected** number of pairs that have at least one consistent path.
  - The expectation computation takes edge weights into account

More definitions

- Probability of a path π, $p(\pi)$, is given by the product of reliabilities of its edges.
- The variable $K_{s,t}$ indicates that there is at least one explanatory path which is consistent.

Expectation calculation

$$E\left(\sum_{(s,t)\in X} K_{s,t}\right) = \sum_{(s,t)\in X} E(K_{s,t}) = \sum_{(s,t)\in X} p(K_{s,t}=1)$$

- Given a collection of explanatory paths for a pair (s,t), Π, the indicator M_{Π} variable indicates whether all these paths are consistent with this pair.
- The probability of all paths in a collection is the product of all the edges in these paths.
  - for overlapping paths and edge is counted once

Expectation calculation contd...

- The probability that at least one path is consistent in a set of paths that explain a knockout pair is:

$$p(K_{s,t} = 1) = \sum_{1 \le i \le n} (-1)^{i-1} \sum_{\Gamma \in P_{\Pi}^{i}} p(\Gamma) M_{\Gamma}$$

Maximization of Expectation

- Integer programming formulation
  - Maximize a linear objective function with respect to linear constraints
- Maximize $\sum_{(s,t)\in X} p(K_{s,t} = 1)$
- Subject to:
  - Path constraints: determining explanatory paths that are consistent
  - Knockout pair constraints: determining the expectation of knockout pairs

Path constraints

$$M_\pi \leq \sum_{e \in \pi} sgn(e) - 2d_\pi + 1 - e(s(\pi), t(\pi)) \leq 1$$

$$M_\pi - M_{q_\pi} \leq 0$$

$$0 \leq \sum_{\gamma \in \Gamma} 2M_\gamma - 2|\Gamma| M_\Gamma \leq 2|\Gamma| - 1$$

Confidence assignment

- Assign a confidence to each protein showing how confident we are on its sign assignment.
- Find the optimum value using the original sign assignment
- Flip the sign assignment and recompute the optimum value.
- If the difference between these two values are greater then a defined threshold we state that assignment as a confident assignment.
- A knockout pair is confidently explained if all the nodes along the consistent explanatory path are confident nodes.

Speedup heuristic

- Infer model variables iteratively in a number of runs
- Each iteration starts with a set of confident variables and try to assign signs to the remaining variables that maximize the objective function
- The iterations starts at paths of length 1 and at each iteration the maximum explanatory path length is increased by 1 until a predefined threshold.

The algorithm

$SPIN(G, X, k)$

For i in $1 \ldots k$ **loop**

$(solution, O_{max}) = Optimize(G, X, i)$

C = MeasureConfidence$(solution, O_{max})$

Foreach variable $v \in solution$ **loop**

If $C_v > \epsilon$ **then**

Update the sign of v in G

End if

End loop

End loop

Function: Optimize

$Optimize(G, X, k)$

Find explanatory paths of length $\leq k$

Foreach unexplained knockout pair $x \in X$ **loop**

  Foreach explanatory path π for x **loop**

    Construct linear equations for π

  End loop

  Construct linear equations for the constraints of x

  Construct linear equations that calculate $p(K_x = 1)$

End loop

Maximize the expected number of explained knockout pairs subject to the constraints

Return the optimal solution and its value

Function: MeasureConfidence

MeasureConfidence($Solution, O_{max}$)

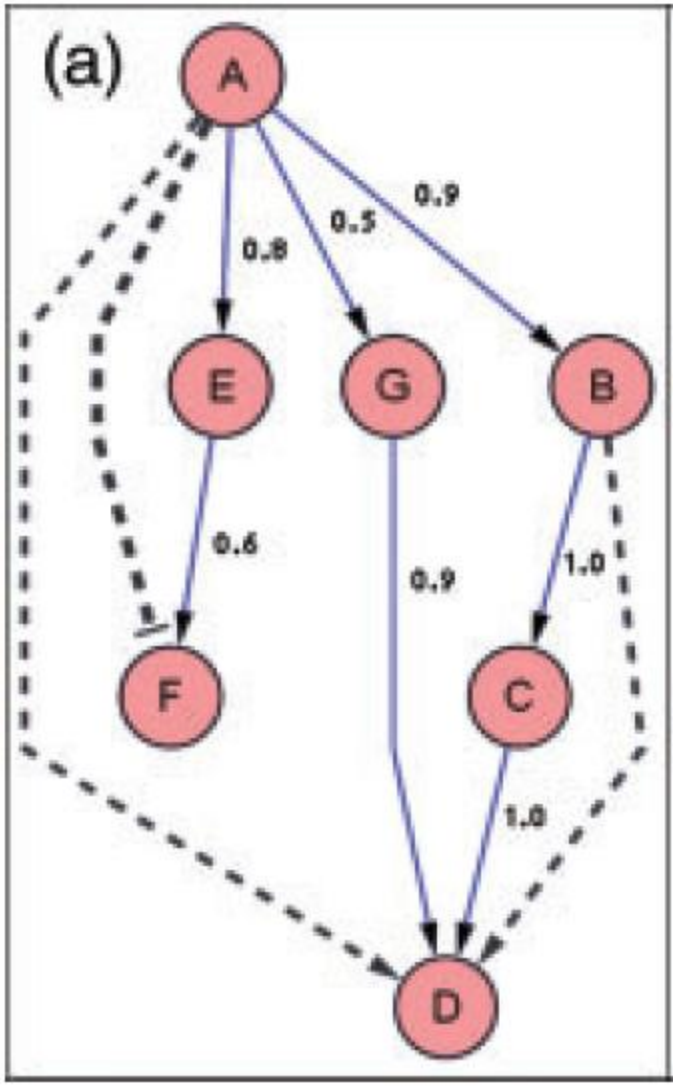
Foreach variable $v \in Solution$ **loop**

O_v = maximum objective given $sgn(v) = (1 - sgn(v))$

$C_v = O_{max}$ - O_v

End loop

Toy example: The input



Explanatory paths and knockout pair indicators

$$\pi_1 = A \rightarrow B \rightarrow C \rightarrow D\ (0.9).$$

$$\pi_2 = B \rightarrow C \rightarrow D\ (1).$$

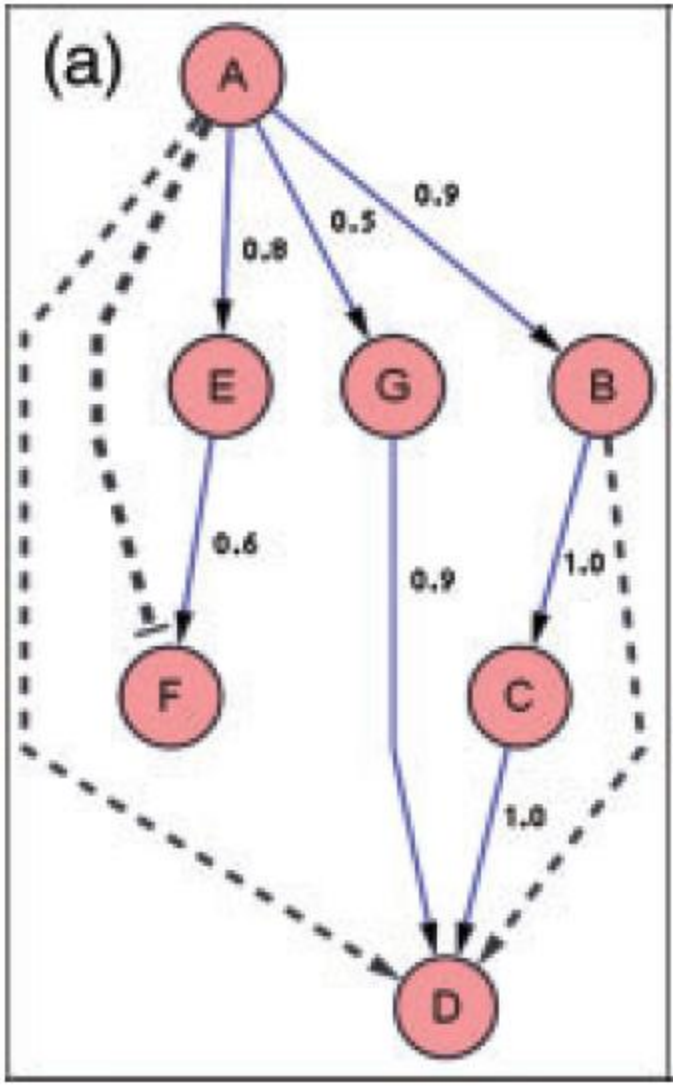
$$\pi_3 = A \rightarrow G \rightarrow D\ (0.45).$$

$$\pi_4 = A \rightarrow E \rightarrow F\ (0.8).$$

$$K_{A,D} \equiv (M_{\pi_1} \vee M_{\pi_3})$$

$$K_{B,D} \equiv M_{\pi_2}$$

$$K_{A,F} \equiv M_{\pi_4}$$



The total expectation

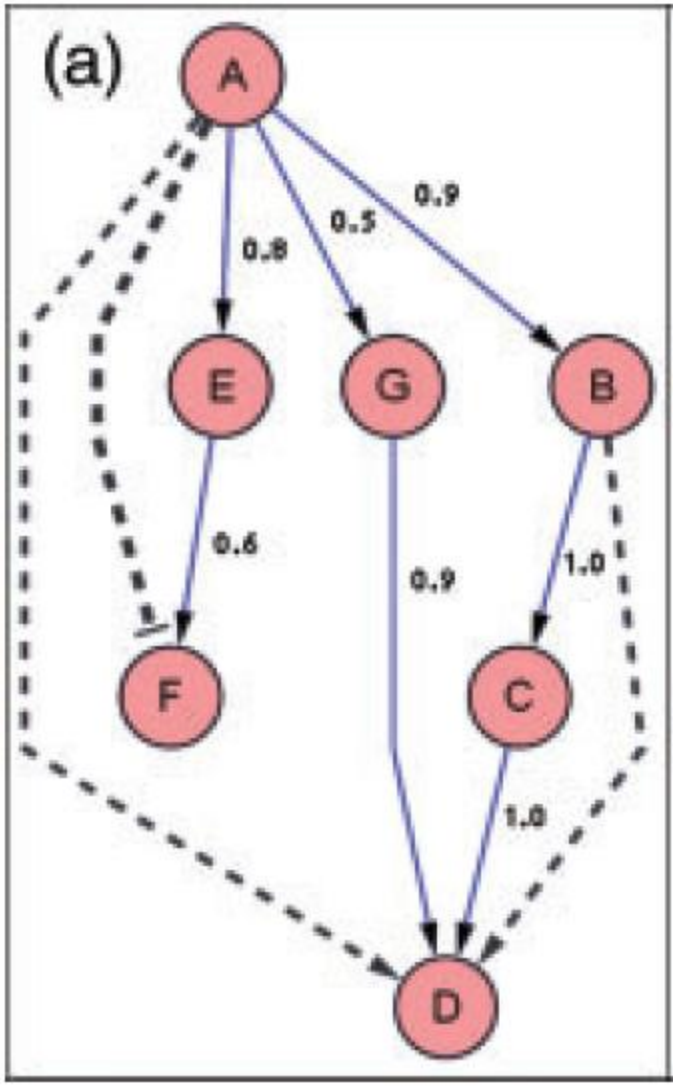
$$E\Big(\sum_{(s,t)\in X} K_{s,t}\Big)$$

$$= \sum_{(s,t)\in X} p(K_{s,t} = 1) =$$

$$= p(K_{A,D} = 1) + p(K_{B,D} = 1) + p(K_{A,F} = 1) =$$

$$= p(\pi_1)M_{\pi_1} + p(\pi_3)M_{\pi_3} - p(\{\pi_1, \pi_3\})M_{\{\pi_1,\pi_3\}}$$

$$+ p(\pi_2)M_{\pi_2} + p(\pi_4)M_{\pi_4}$$



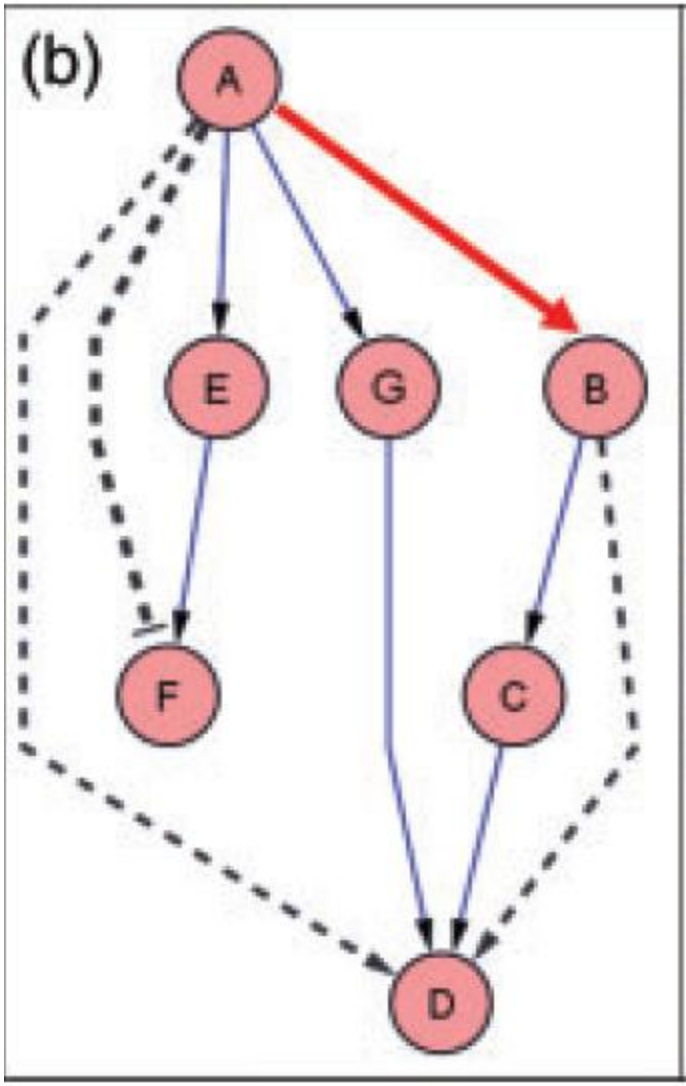
Solution (edge assignment)

$$E\Big(\sum_{(s,t)\in X} K_{s,t}\Big)$$

$$= \sum_{(s,t)\in X} p(K_{s,t}=1) =$$

$$= p(K_{A,D}=1) + p(K_{B,D}=1) + p(K_{A,F}=1) =$$

$$= p(\pi_1)M_{\pi_1} + p(\pi_3)M_{\pi_3} - p(\{\pi_1,\pi_3\})M_{\{\pi_1,\pi_3\}}$$
$$+ p(\pi_2)M_{\pi_2} + p(\pi_4)M_{\pi_4}$$



Solution (node assignment)

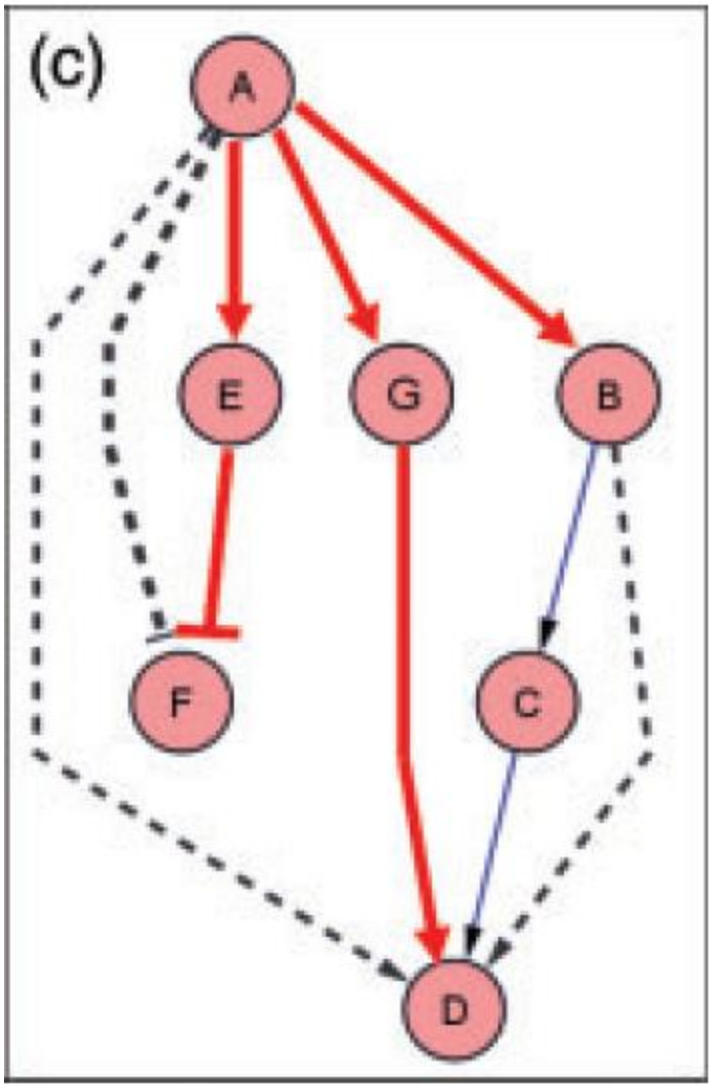
$$E\Big(\sum_{(s,t)\in X} K_{s,t}\Big)$$

$$= \sum_{(s,t)\in X} p(K_{s,t}=1) =$$

$$= p(K_{A,D}=1) + p(K_{B,D}=1) + p(K_{A,F}=1) =$$

$$= p(\pi_1)M_{\pi_1} + p(\pi_3)M_{\pi_3} - p(\{\pi_1,\pi_3\})M_{\{\pi_1,\pi_3\}}$$

$$+ p(\pi_2)M_{\pi_2} + p(\pi_4)M_{\pi_4}$$



Results

- On yeast network
- Small and large scale tests

Small scale experiments

Table 1. Cross-validation results for the yeast mating subnetwork

Method	Left out	Number of Trials	Correct	Incorrect	Undecided	Accuracy
Yeang *et al.*	1	103	97.1%	2.9%	0%	97%
SPINE—edge variant	1	103	98%	1%	1%	99%
Yeang *et al.*	5	200	96.5%	3.5%	0%	96.5%
SPINE—edge variant	5	200	96.9%	0.4%	2.7%	99%

Benefit of using “expectation”

Table 2. Performance in cross-validation on the yeast mating subnetwork

Method variant	Maximization criterion	Left out	Correct	Incorrect	Undecided	Accuracy
Edge variant	Expectation	1	98%	1%	1%	99%
Edge variant	Number	1	47.6%	0%	52.4%	100%
Node variant	Expectation	1	89.3%	10.7%	0%	89.3%
Node variant	Number	1	60.2%	9.7%	30.1%	86.1%
Edge variant	Expectation	1 (noisy)	93.5%	3.9%	2.6%	96%
Node variant	Expectation	1 (noisy)	88.3%	11.7%	0%	88.3%

A hidden knockout pair is considered to be successfully predicted if the expected number of explanatory paths with consistent signs is higher than the expected number of explanatory pathways with non-consistent signs. The accuracy represents the percentage of correct predictions among all predictions made.

Large scale experiments

Table 3. Cumulative contributions of different path lengths (levels) to the inference process

Level	Number of Explained knockouts	Number of Inferred signs
1	107	14
2	183	30
3	861	183

Large scale experiments

Table 4. Results of the comparison to known regulators, for both Yeang *et al.* and SPINE

Type	Number of Known	Number of Predicted	Number of Correct	Significance
Yeang *et al.* - Activators	119	31	28	0.002
Yeang *et al.* - Repressors	57	22	8	0.4
SPINE - Activators	120	37	32	0.003
SPINE - Repressors	60	22	12	0.02

Shown are the number of known activators and repressors that appear in the network, the number of sign predictions on this known set, the number of correct predictions and a hypergeometric *P*-value. The latter was computed separately for the activator and repressor predictions.

Alternative Solutions?

- What about using EM – Expectation Maximization?
- Other optimization algorithms?

Next week

- An Efficient Network Querying Method Based on Conditional Random Fields by Huang et al., Bioinformatics, Sep 2011.