Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. Bioinformatics, 2010

C.J.Vaske et al.

May 22, 2013

Presented by: Rami Eitan

Complex Genomic Rearrangements

- Cancer tissue experience molecular changes
- Varied genomic data available
  - copy number variations
  - mutations, gene expression
- Stratification of cancers can improve:
  - diagnosis
  - prognosis
  - risk assessment
  - response to treatment

Complex Genomic Rearrangements

- Genetic alterations differ between patients
- Pathways often are common

Pathways

- What is a pathway?

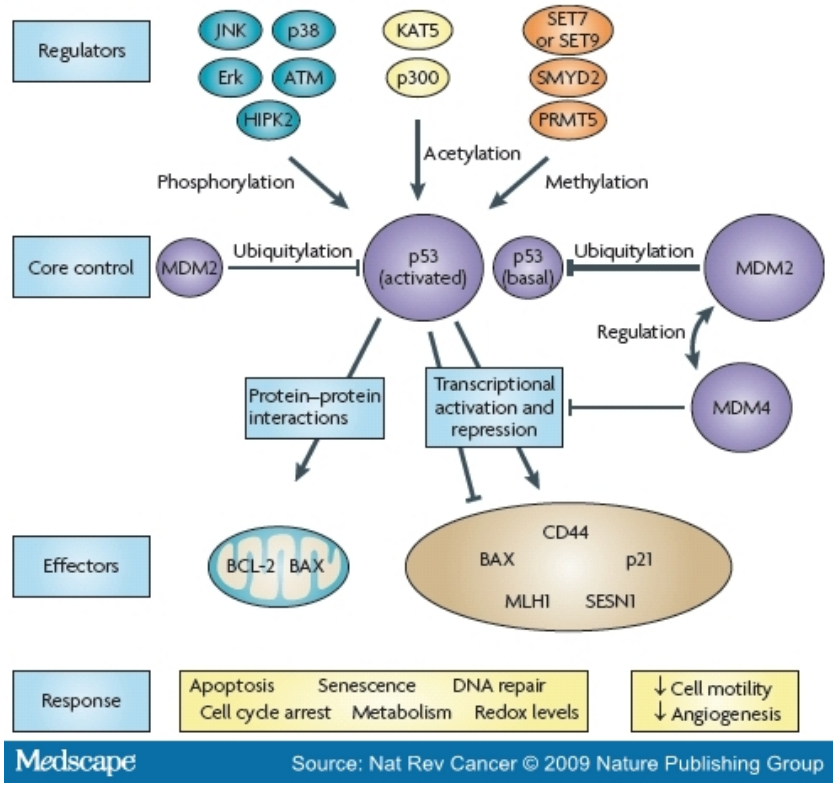


Figure : The P53 pathway

Pathways

- A set of interactions between entities, logically grouped together around a biological process.
- Protein-coding genes, small molecules, complexes, gene families, abstract processes
- Available databases: Reactome, KEGG, NCI

Motivation

- Integrative analysis of cancer genome data
  - Copy number variations, gene expressions
- Leverage pathway information to find frequently occurring pathway perturbations
  - NCI pathway interaction database, KEGG etc.

Observed Data

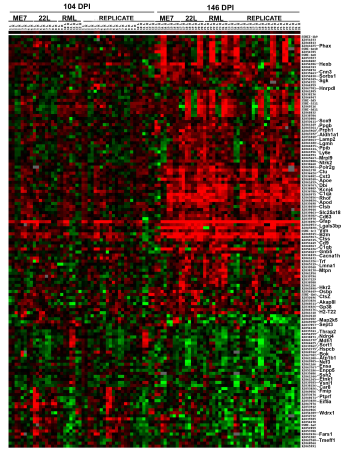


Figure : Gene expression

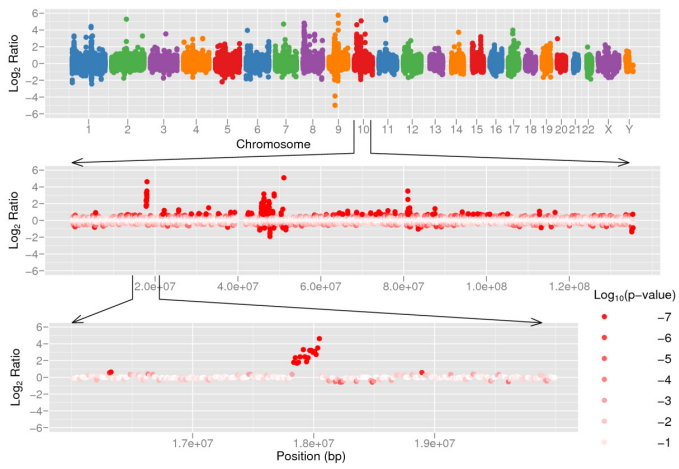
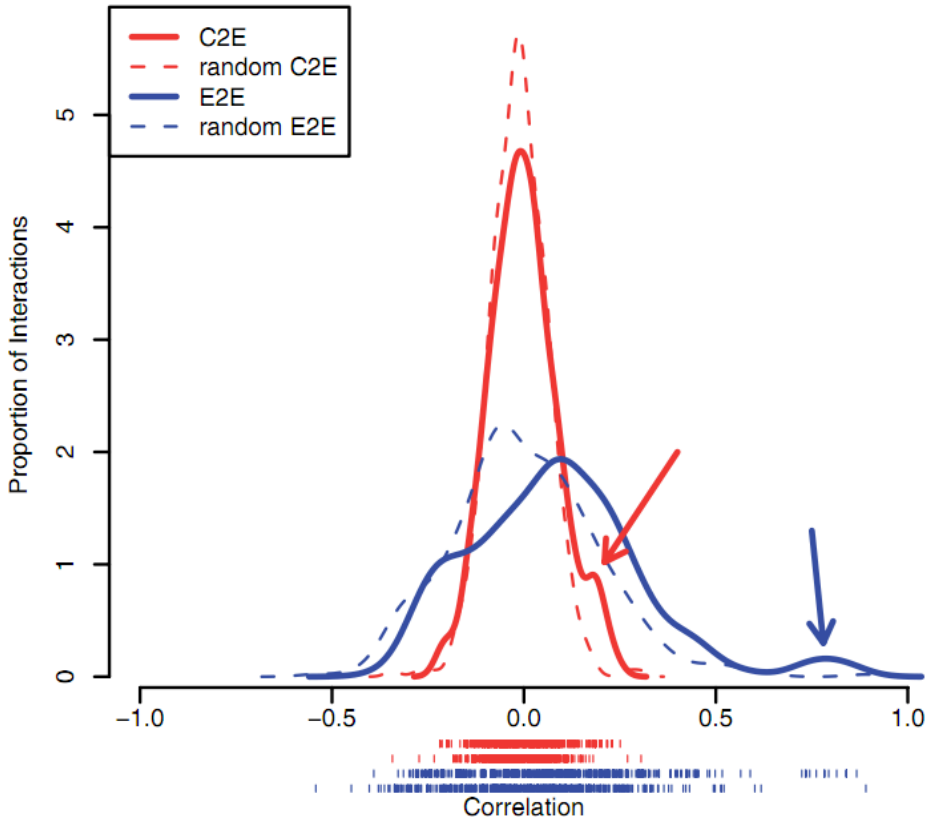


Figure : Copy number

Motivation

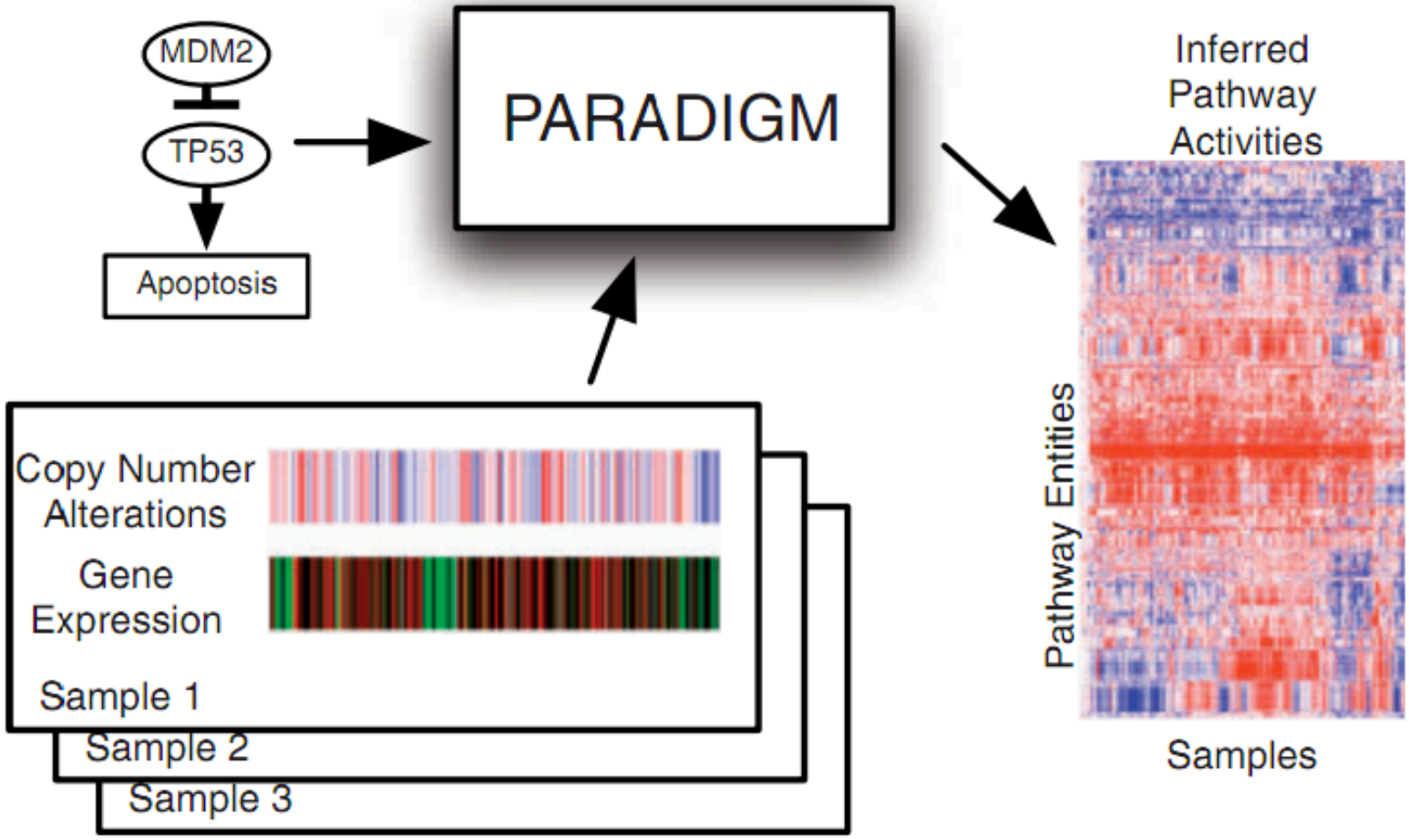
- Pathway information contains information on how genes are supposed to behave



Input

- Infer integrated pathway activity (IPA)
- Produce a matrix A. A_{ij} is the inferred activity of entity i in patient j

PARADIGM


MDM2
TP53
Apoptosis
PARADIGM
Inferred
Pathway
Activities
Copy Number
Alterations
Gene
Expression
Sample 1
Sample 2
Sample 3
Pathway Entities
Samples

Factor graph

- Factor graph is a probabilistic graphical model.
- Variables, factors.

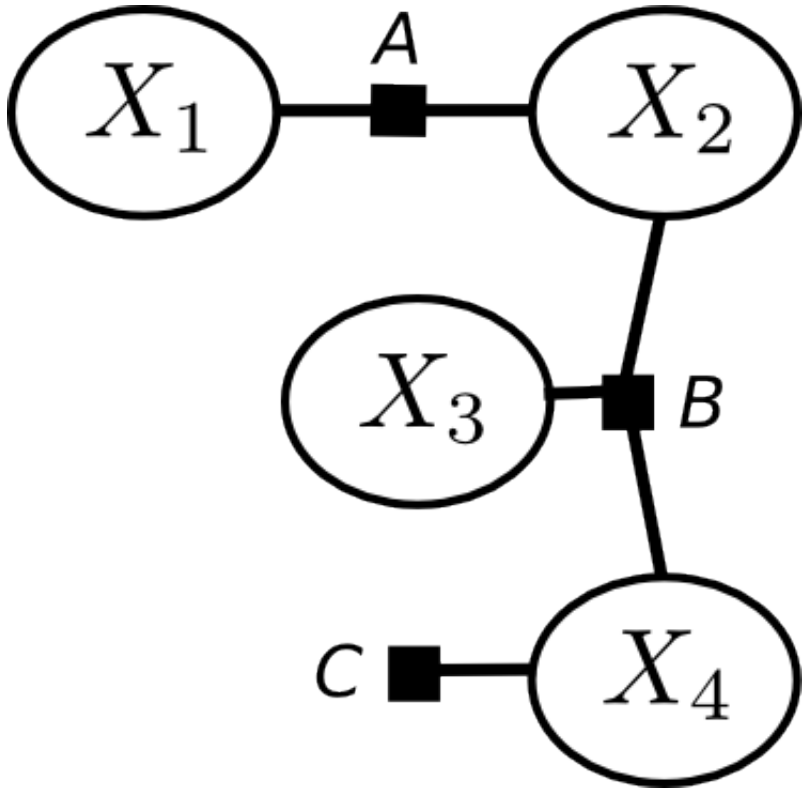
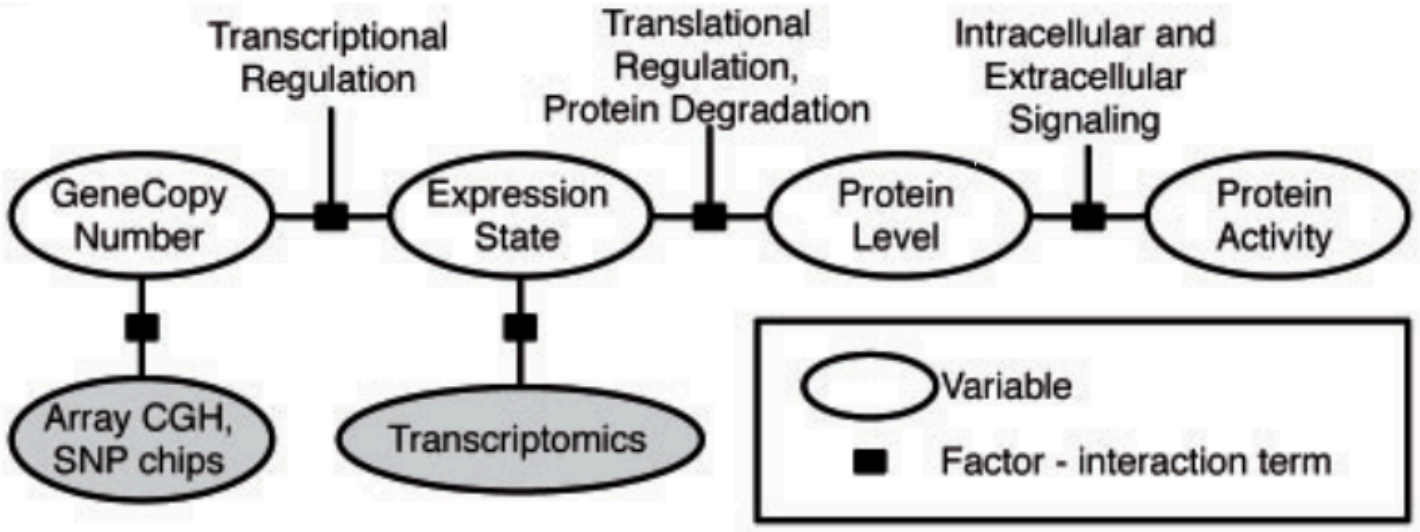


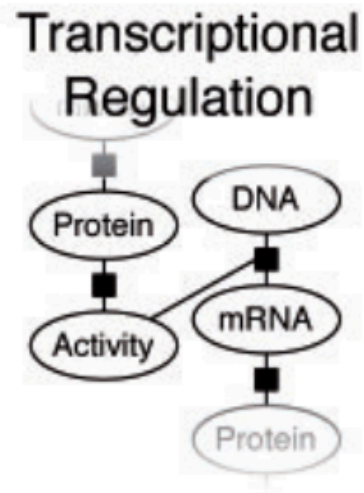
Figure : A simple factor graph

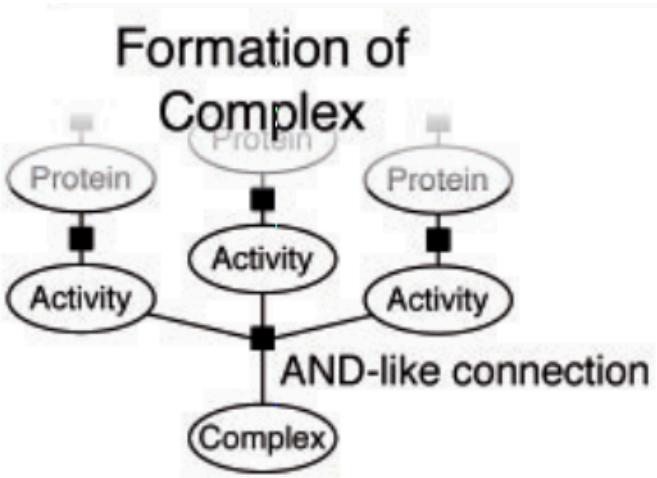
PARADIGM Model

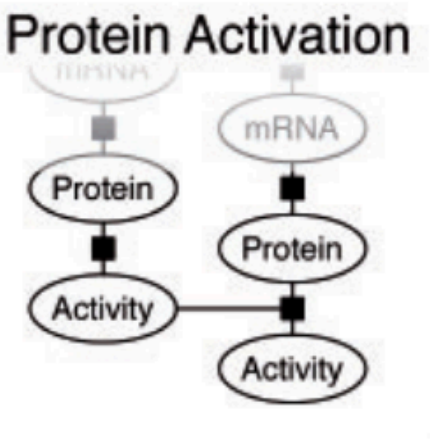
- Factor graph representation of various entities corresponding to a single gene

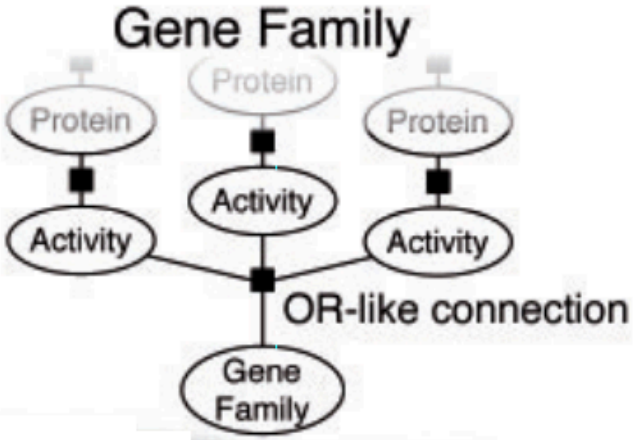


PARADIGM Model: Gene Interactions



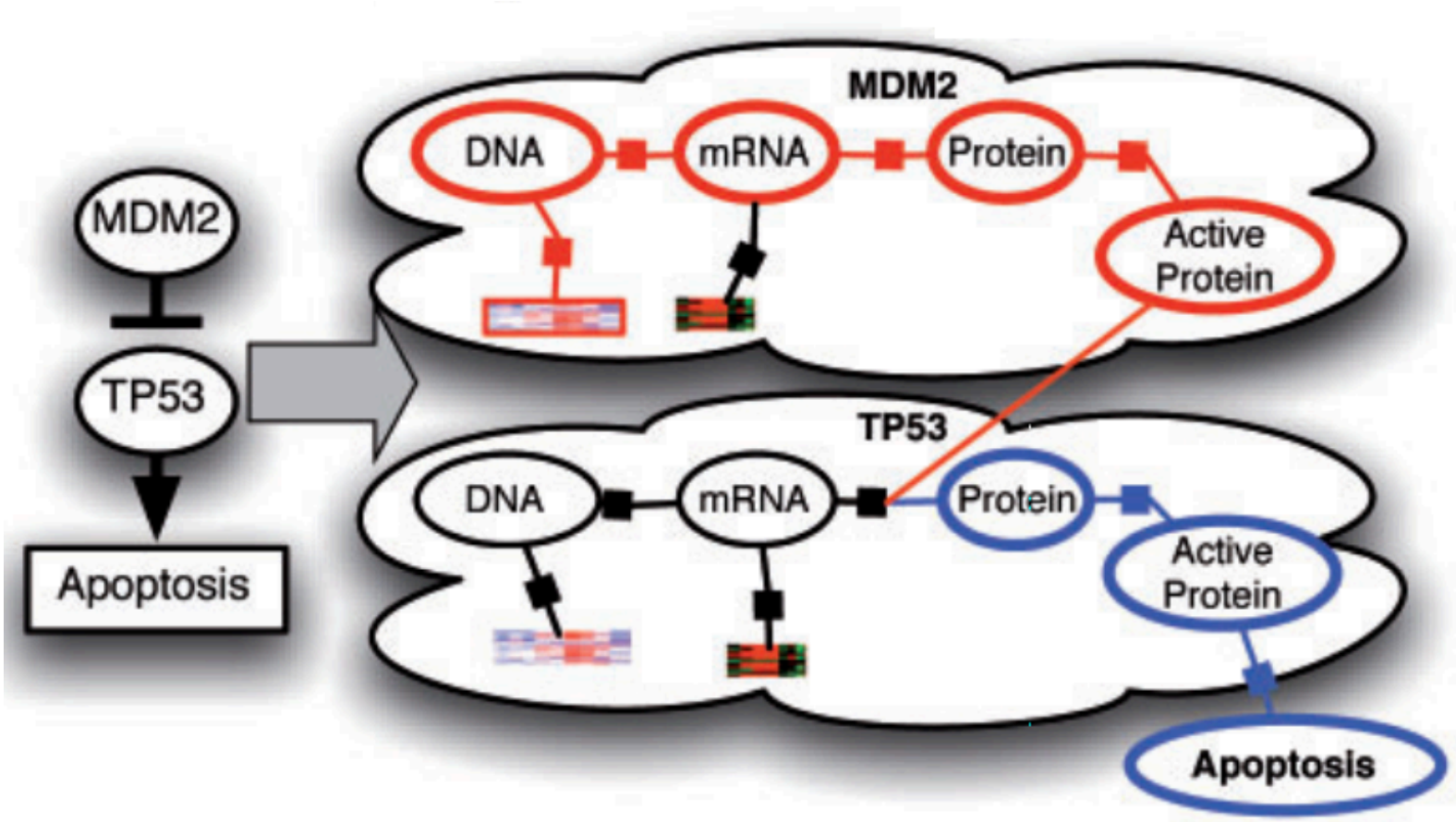






PARADIGM Model:

- A factor graph for a pathway



Model Specification

- Convert an NCI pathway into a factor graph
  - NCI pathway to Bayesian network
    - Directed network
    - Each variable takes values of -1 (de-activation), 0 (normal), 1 (activation)
      - mRNA: over expression for activation
      - Copy number variations: more than two copies for activations
    - Probability distribution of each node
      - Labeled edges for positive/negative interactions
      - Set the value of the child node as weighted votes from its parents

Model Specification

- Converting the Bayesian network to a factor graph
  - Assign a factor to each group of variables consisting of a node and its parents

$$P(X) = \frac{1}{Z} \prod_{j=1}^{m} \phi_j (X_j)$$

$$\phi_i(x_i, \text{Parents}(x_i)) = \begin{cases} 1-\epsilon & x_i \text{ is the expected state from Parents}(x_i) \\ \frac{\epsilon}{2} & \text{otherwise.} \end{cases}$$

  - Z: normalization constant
  - ε = 0.001

Inference

- Observed variables: copy number variations, gene expressions
- Unobserved variables: protein, protein activity, overall pathway activity state
- Learn models with EM algorithm
  - E step: Infer the probabilities of the unobserved variables
  - M step: Change parameters to to maximize the likelihood given the probabilities

Expectation Maximization

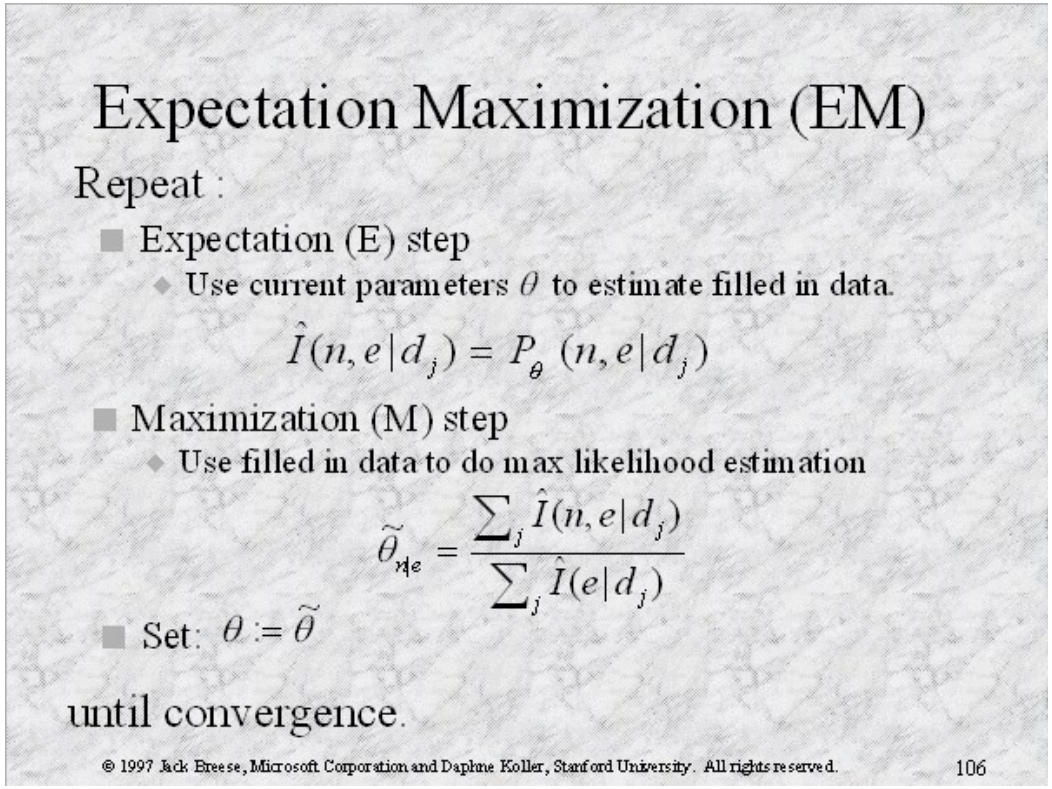


Figure : EM algorithm

Log-likelihood Ratio Test

- Test statistic for assessing entity i's activity given data D

$$L(i,a) = \log\left(\frac{P(D,x_i=a|\Phi)}{P(D,x_i\neq a|\Phi)}\right) - \log\left(\frac{P(x_i=a|\Phi)}{P(x_i\neq a|\Phi)}\right)$$

$$= \log\left(\frac{P(D|x_i=a,\Phi)}{P(D|x_i\neq a,\Phi)}\right).$$

  - The probabilities can be obtained by performing inference on the factor graph

$$P(x_i=a,D|\Phi)=\frac{1}{Z}\prod_{j=1}^{m}\sum_{\mathbf{S}\sqsubset A_i(a)\cup D X_j}\phi_j(\mathbf{S})$$

$$P(x_i=a|\Phi)=\frac{1}{Z}\prod_{j=1}^{m}\sum_{\mathbf{S}\sqsubset A_i(a) X_j}\phi_j(\mathbf{S})$$

Significance assessment

- Permutate the labels of the observed data
- 'Within' permutation: choosing random genes from the same pathway
- 'Any' permutation: choosing any random genes
- 1000 permutations of each type are used to determine null distribution

Decoy paths

- Create decoy paths by replacing genes with random genes
- Maintain the same structure
- All complexes and abstract processes remain the same

Log-likelihood Ratio Test

- Aggregating over multiple values entity i takes

$$IPA(i) = \begin{cases} L(i,1) & \text{L(i,1)>L(i,-1) and L(i,1)>L(i,0)} \\ -L(i,-1) & \text{L(i,-1)>L(i,1) and L(i,-1)>L(i,0)} \\ 0 & \text{otherwise.} \end{cases}$$

Dataset

- Breast cancer copy number and gene expression data
- TCGA Glioblastoma copy number and gene expression data
- Pathways from NCI pathway interaction database (PID)

Results - breast cancer

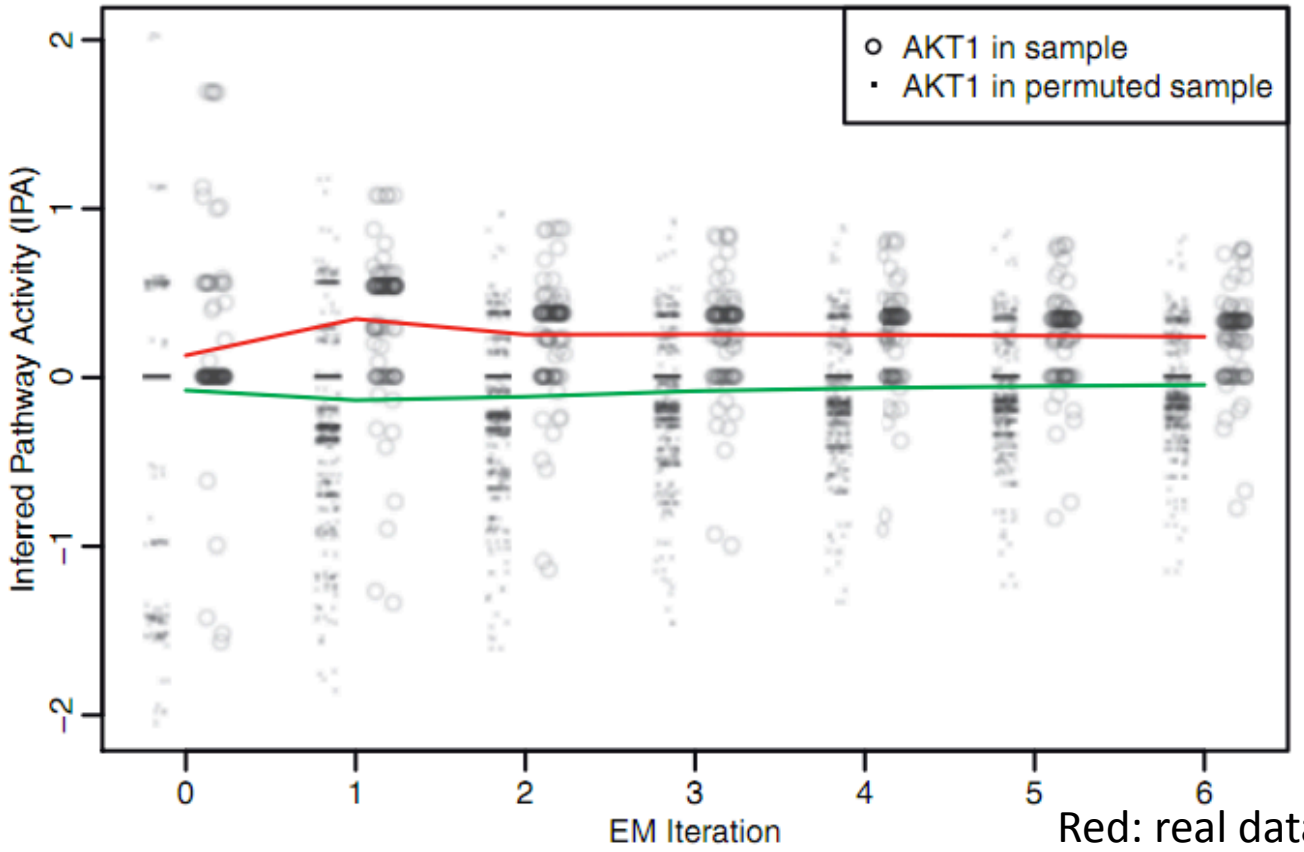
- Breast Cancer dataset:
  - 56172 IPA's (7%) found to be significantly higher
  - 497 significant entities per patient on average
  - 103 out of 127 pathways had at least one entity altered in 20% or more of the patients

Results - GBM

- GBM dataset:
  - 141682 IPA's (9%) found to be significantly higher
  - 616 significant entities per patient on average
  - 110 out of 127 pathways had at least one entity altered in 20% or more of the patients

EM Convergence

- Original data vs. permuted data



Red: real data

Green: permuted data

Results - decoy paths

Distinguishing decoy from real pathways

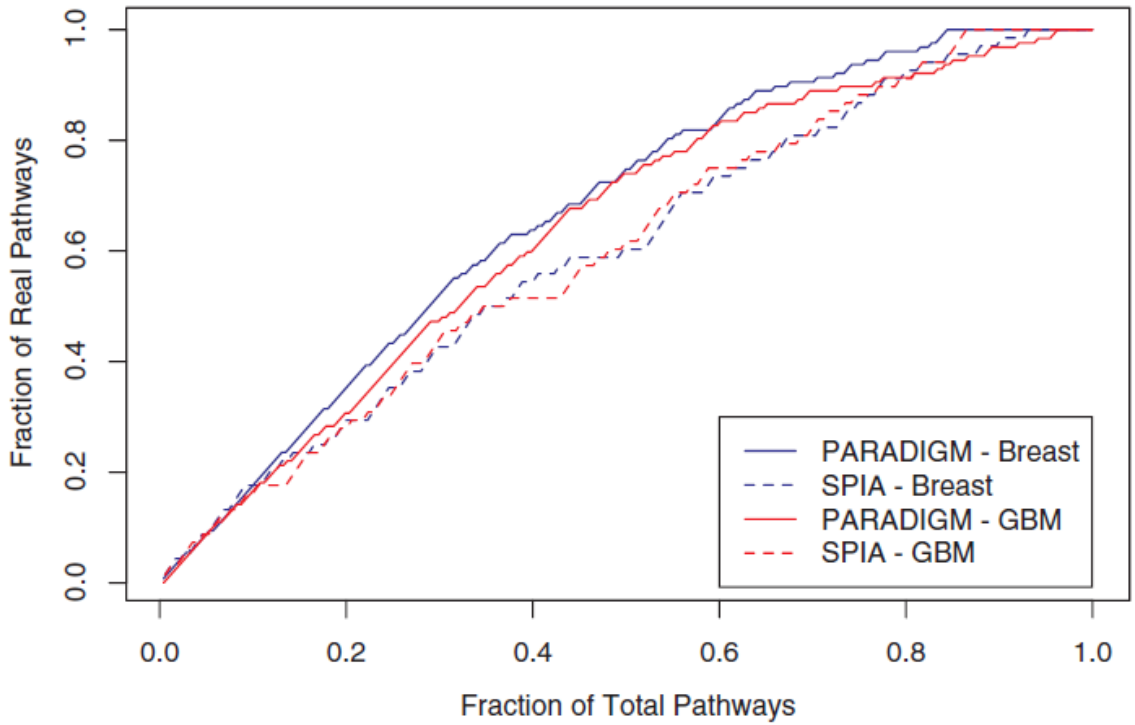


Figure : PARADIGM vs SPIA: FP rate

Results - decoy paths

- Distinguishing decoy from real pathways
- Breast cancer AUC:
  - PARADIGM: 0.669
  - SPIA: 0.602
- GBM AUC:
  - PARADIGM: 0.642
  - SPIA: 0.604

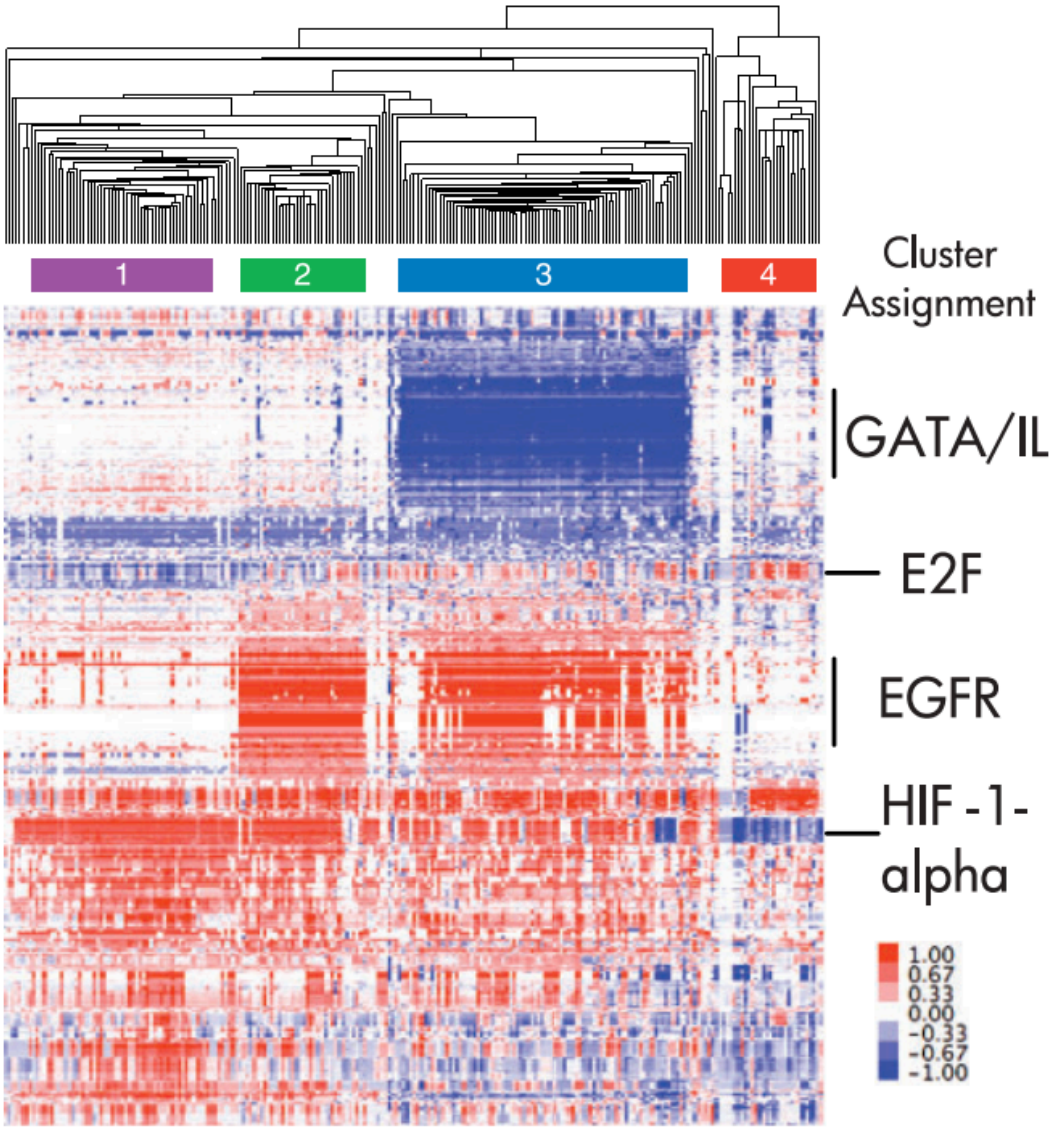
Top PARADIGM Pathways of Breast Cancer

Rank	Name	Avg.[a]	SPIA?[b]
1	Class I PI3K signaling events mediated by Akt	20.7	No
2	Nectin adhesion pathway	14.1	No
3	Insulin-mediated glucose transport	13.8	No
4	ErbB2/ErbB3 signaling events	12.1	Yes
5	p75(NTR)-mediated signaling	11.5	No
6	HIF-1-alpha transcription factor network	10.7	No
7	Signaling events mediated by PTP1B	10.7	No
8	Plasma membrane estrogen receptor signaling	10.6	Yes
9	TCR signaling in naive CD8+ T cells	10.6	No
10	Angiopoietin receptor Tie2-mediated signaling	10.1	No
11	Class IB PI3K non-lipid kinase events	10.0	No
13	Osteopontin-mediated events	9.9	Yes
12	IL4-mediated signaling events	9.8	No
14	Endothelins	9.8	No
15	Neurotrophic factor-mediated Trk signaling	9.7	No

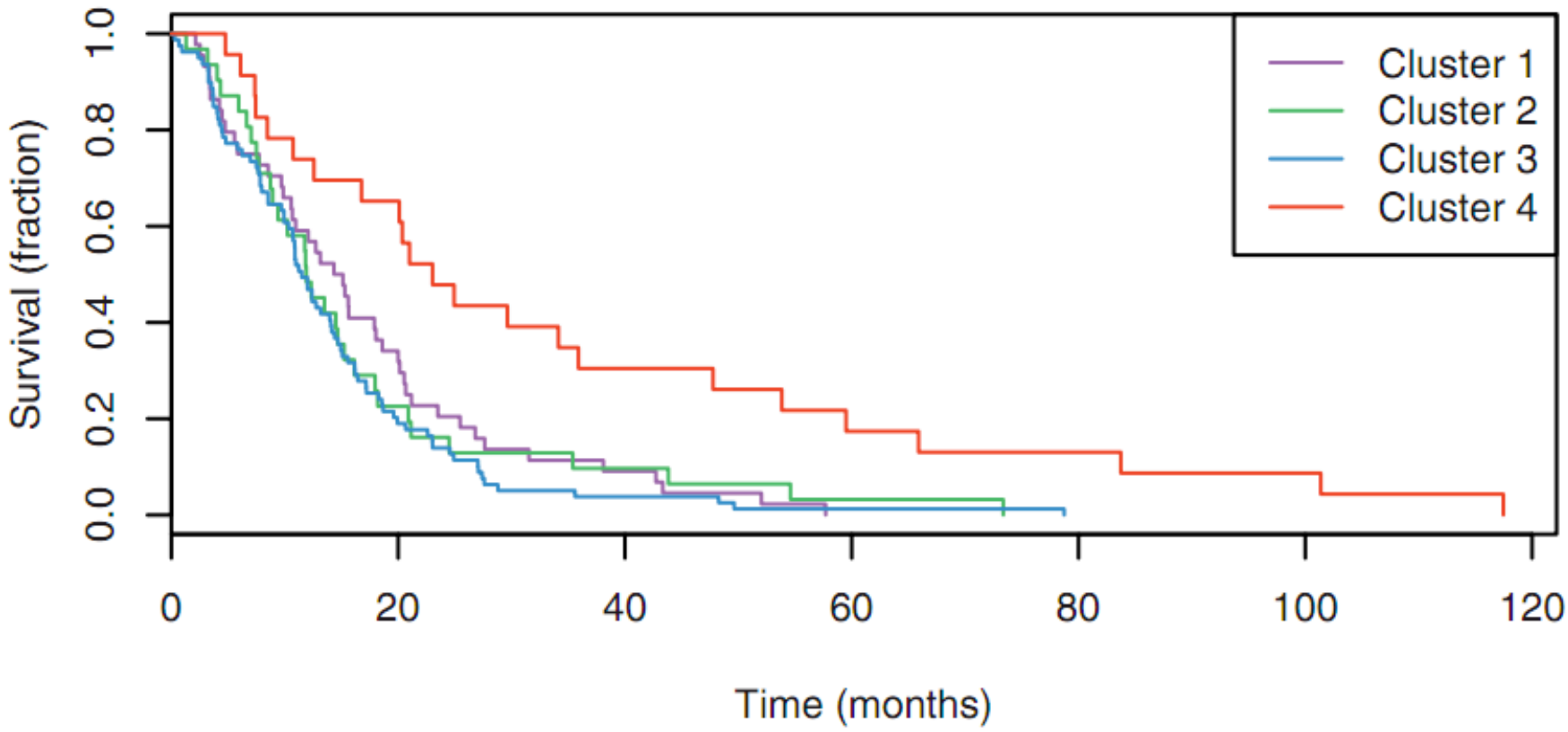
Top PARADIGM Pathways of Glioblastoma

Rank	Name	Avg.[a]	SPIA?[b]
1	Signaling by Ret tyrosine kinase	46.0	No
2	Signaling events activated by Hepatocyte GFR	43.7	No
3	Endothelins	42.5	Yes
4	Arf6 downstream pathway	42.3	No
5	Signaling events mediated by HDAC Class III	36.3	No
6	FOXM1 transcription factor network	35.9	Yes
7	IL6-mediated signaling events	33.2	No
8	FoxO family signaling	31.3	No
9	LPA receptor mediated events	30.7	Yes
10	ErbB2/ErbB3 signaling events	30.1	No
11	Signaling mediated by p38-alpha and p38-beta	28.1	No
12	HIF-1-alpha transcription factor network	27.6	Yes
13	Non-genotropic Androgen signaling	27.3	No
14	p38 MAPK signaling pathway	27.2	No
15	IL2 signaling events mediated by PI3K	26.9	No

Glioblastoma Subtypes



Survival Rates for Each Subtypes



Results - Patient vs permutation

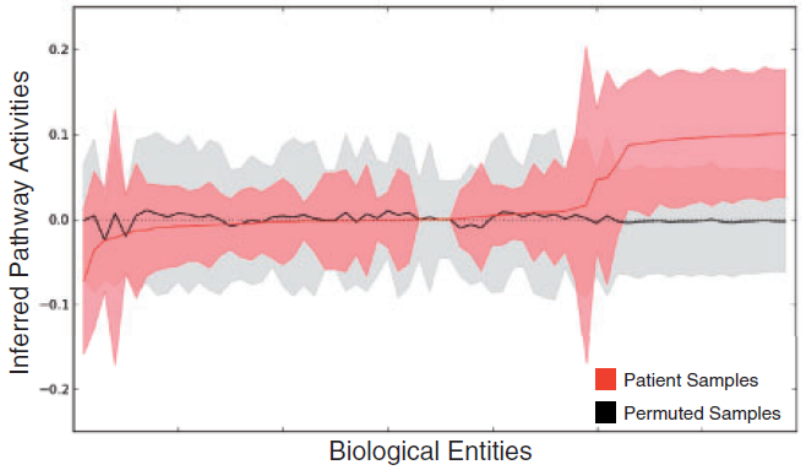


Fig. 6. Patient sample IPAs compared with 'within' permutations for Class I PI3K signaling events mediated by Akt in breast cancer. Biological entities were sorted by mean IPA in the patient samples (red) and compared with the mean IPA for the permuted samples. The colored areas around each mean denote the of SD each set. IPA's on the right include AKT1, CHUK and MDM2.

Figure : Patient vs permuted IPA's

Results - Patient vs permutation

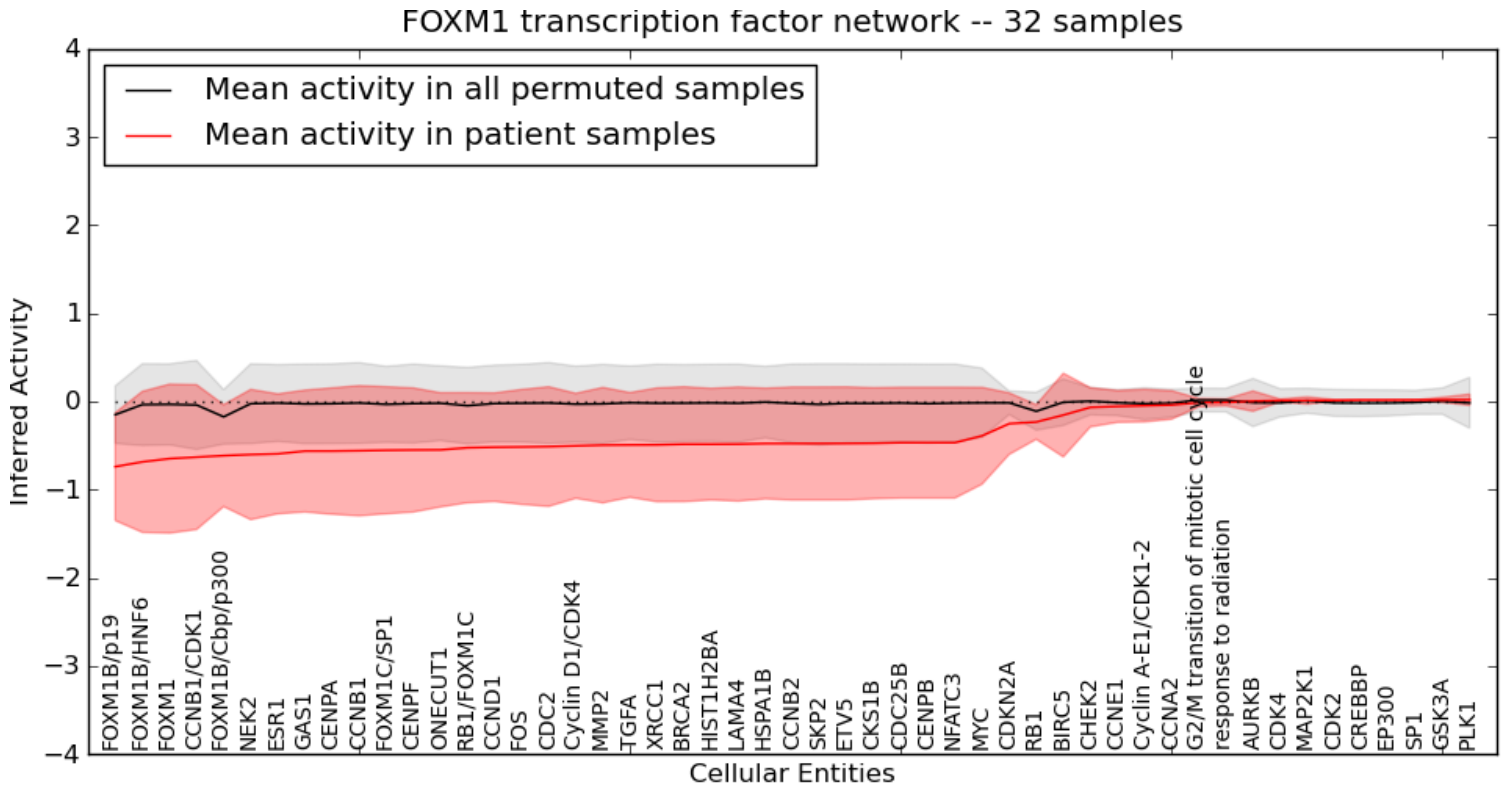


Figure : Patient vs permuted IPA's. Source: Broad Institute/Dana-Farber Cancer Institute/Harvard Medical School

Summary

- PARADIGM integrates different types of data, including gene-expression, copy number variation, and pathway database, in order to infer pathway activities for individual cancer patients.
  - Factor graph model for representing pathway and modeling datasets
  - Pathway activities inferred by PARADIGM can be used to identify cancer subtypes

Questions

Discussion

- Can the method be successfully expanded to more observed data?
- Instead of using the pathways as is, can this method be used to find new pathways and interactions?