Learning Regulatory Links in Cancer Through Integrated Pathway Analysis with Paradigm

Andrew Sedgewick
Summer work with Five3 Genomics LLC
Advisor: Takis Benos

Abstract
High-dimensional omics profiling provides a detailed molecular view of individual cancers. We extended the Paradigm algorithm, a pathway analysis method for combining multiple omics data types with 10307 interactions curated from the literature, to learn the strength and direction of curated interactions. Using genomic and mRNA expression data from 1936 samples in The Cancer Genome Atlas (TCGA) cohort, we learned interactions that gave support for and relative strength of curated links. Gene set enrichment found that targets of the strongest interactions were significantly enriched for apoptosis and cell morphogenesis, and that strong regulators were significantly associated with phosphorylation. Within the TCGA breast cancer cohort we assessed different interaction strengths between breast cancer subtypes, and found interactions associated with the MYC pathway and the ER alpha network to be among the most differential between basal and luminal A subtypes. Learning links separately under an Naive Bayesian assumption produced gene activity predictions that, when clustered, found groups of patients with better separation in survival than both the original version of Paradigm and a version without the assumption.

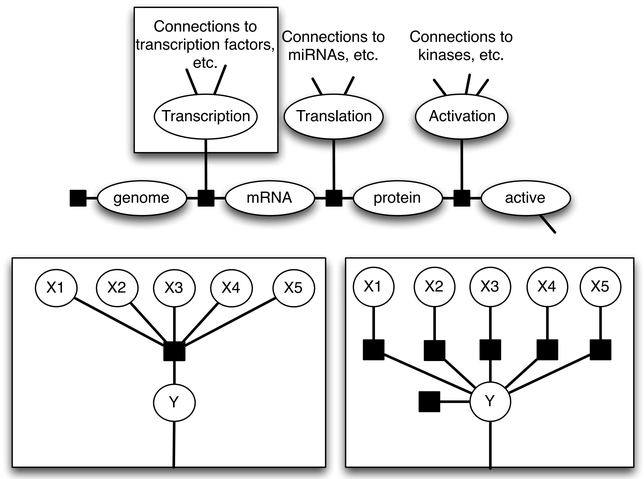

Figure 1 Factor graphs learned by Paradigm. Previously regulatory node states were determined by a vote of regulators, we now can learn a full conditional probability table or we can learn conditional probabilities of individual links and use a Naive Bayes assumption to calculate the likelihood of the child node given the parents.

$$G_{parent-child} = 2N \sum_{i,j} P(X_i, Y_j) \ln \frac{P(X_i, Y_j)}{P(X_i)P(Y_j)}$$

Equation 1 We use a G-test to find the significance of a regulatory link. This tests if a parent node, X, and child node, Y, are statistically independent where i and j are settings of each node. This statistic follows the χ^2 distribution.

$$WPMI_{i,j} = P(X_i, Y_j) \ln \frac{P(X_i, Y_j)}{P(X_i)P(Y_j)}$$

Equation 2 The weighted point-wise mutual information (WPMI) tells us how much a given setting (i,j) of parent and child nodes contributes to the G statistic (eq. 1).

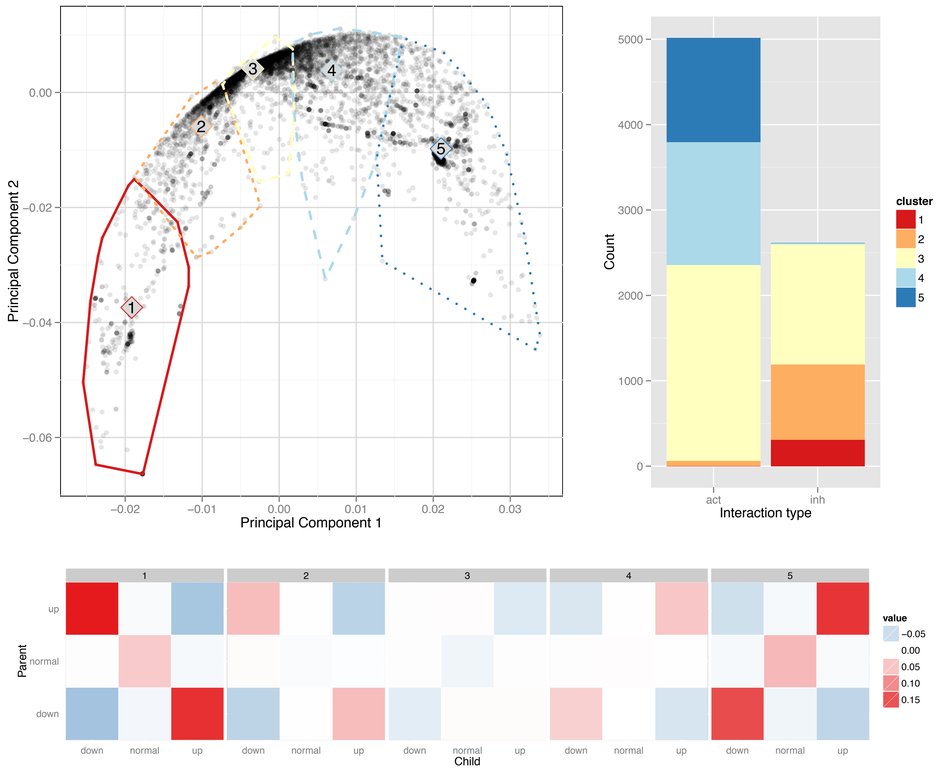

Figure 2 a. Principal component analysis of regulatory links in the TCGA cohort. Each point is the projection of the 9 WPMI scores for a link onto the top two principal components. The convex hulls show the membership of k-means clustering performed on the (unprojected) wpmi scores, and the cluster numbers are placed at the centroid of each cluster. **b.** Cluster membership of links labeled as activation and inhibition in the pathway. **c.** Heatmaps of the WPMI values of the centroids of the clusters show a range from strong inhibition (1) to strong activation (5).

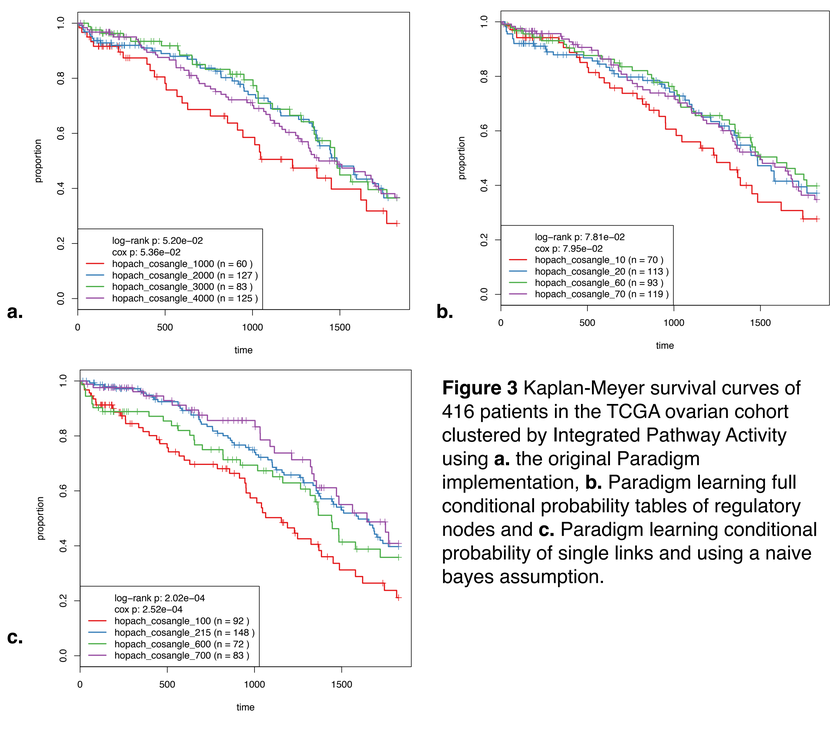

Figure 3 Kaplan-Meyer survival curves of 416 patients in the TCGA ovarian cohort clustered by Integrated Pathway Activity using **a.** the original Paradigm implementation, **b.** Paradigm learning full conditional probability tables of regulatory nodes and **c.** Paradigm learning conditional probability of single links and using a naive bayes assumption.

Parent	Child	p-val Basal	p-val Luminal	direction
ERK1 (family)	PTGS2:txreg	2.04e-5	.146	↓
MYC/MAX/MIZ-1 (complex)	BCL2:txreg[a]	6.89e-4	.364	↑
JUN dimer (complex)	NTS:txreg	2.66e-3	.395	↑
IL2/IL2R_alpha/JAK1/LCK/JAK3 (complex)	SOCS3:actreg	9.95e-3	.174	↓
HIF1A/ARNT_(complex)	BNIP3:txreg	6.04e-3	.462	↑
MYC/MAX (complex)	ENO1:txreg	.528	1.41e-31	↑
E2/ERA dimer/ PCNA (complex)	TFF1:txreg[a]	.137	8.12e-32	↑
Myb/GATA1 (complex)	GATA1:txreg	.996	2.22e-29	↑
ER2/ERA dimer/AIB1 (complex)	SOD1:txreg	.754	2.63e-29	↑
ERBB4	ERBB:actreg	.216	1.54e-19	↑

[a] intermediate node

Table 1. Regulatory links with p < .05 in either Basal or Luminal breast cancer tumors, but not both.

Parent	Child	g score
FOXA1	SFTPA (family):txreg	3247.197 ↑
HNF1A	HNF4A (family):txreg	3208.440 ↑
GATA1	alpha-globin (family):txreg	3065.885 ↑
ONECUT1	HNF1B (family):txreg	3008.945 ↑
p53 tetramer (complex)	MDM2:txreg[a]	2931.148 ↑
KLF4	Preproghrelin and prepro-des-Gln14-ghrelin (family):txreg	2914.620 ↑
PDX1	NR5A2 (family):txreg	2872.275 ↑
p53 tetramer (complex)	SFN:txreg[a]	2811.958 ↑
ER alpha homodimer (complex)	alpha tubulin (family):txreg	2781.369 ↑
FOXM1	CENPA:txreg	2739.028 ↑

[a] intermediate node

Table 2. Regulatory links that with the highest g test score across the entire TCGA cohort. p-values for all links listed are less than 1e-323

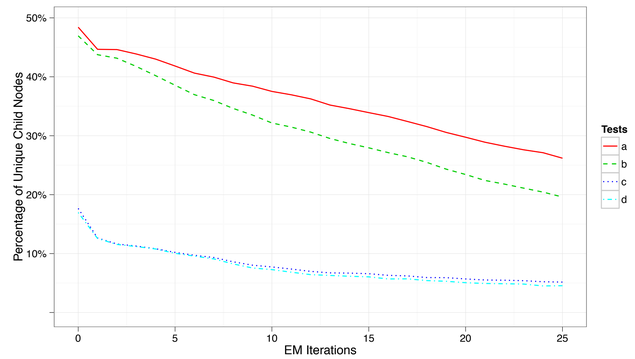

Table 4. Percentage of unique child nodes with at least two parents that fail the following tests at each EM step of a Paradigm run learning a full conditional probability table: **a.** a test of conditional independence given the child (this is the Naive Bayes assumption) **b.** conditional independence and at least one parent is significantly linked to the child **c.** conditional independence and the direction of interactions is ambiguous **d.** all of the above

Conclusions

- We can learn the strength and direction of regulatory links with the Paradigm algorithm
- The strongest links and the most differential links between tumor types can identify biologically relevant genes and interactions.
- Our Naive Bayes assumption holds for many of the regulation nodes, and could only be causing only a small proportion of false positives in activatior or inhibitor identification.
- Using prior knowledge of direction interaction is still necessary.