Mixed Graphical Models with Applications to Integrative Genomics

Genevera I. Allen

Dobelman Family Junior Chair,
Department of Statistics and Electrical and Computer Engineering, Rice University,
Department of Pediatrics-Neurology, Baylor College of Medicine,
Jan and Dan Duncan Neurological Research Institute, Texas Children’s Hospital.

April 1, 2016
Motivation: Integrative Cancer Genomics

The Cancer Genome Atlas

Understanding genomics to improve cancer care

Sequence Changes

- Mutations (SNP)
- Methylation
- Epigenetics
- Copy Number Variation
- MicroRNA Expression

Functional Genetics

Gene Expression
Motivation: Integrative Cancer Genomics

The Cancer Genome Atlas

Understanding genomics to improve cancer genomics care

Mutations (SNP)
- \( \sim 100K - 20\) Million
- Binary / Categorical

Copy Number Variation
- \( \sim 20K - 200K \)
- Continuous

Methylation
- \( \sim 30K - 450K \)
- Bounded Continuous

Gene Expression
- \( \sim 1K - 10K \)
- Continuous (array)
- Counts (Sequencing)

MicroRNA Expression
- \( \sim 1K - 10K \)
- Continuous (array)
- Counts (Seq)
Statistical Data Integration

Uses probabilistic models to jointly model multiple types of measurements taken on the same set of subjects.
Motivation: Integrative Cancer Genomics

Statistical Data Integration

Uses probabilistic models to jointly model multiple types of measurements taken on the same set of subjects.

Objective:

- Joint inference on samples.
  - Harness power across multiple data sources.

- Discover relationships between different types features.
1 Review: Markov Networks

2 Graphical Models via Exponential Families
   • Motivation
   • Graphical Models via Exponential Families
   • Results

3 Mixed Graphical Models
   • Mixed MRFs
   • Mixed Recursive Chain Graphical Models
   • Results
Markov Networks

- Represent dependencies in multivariate distributions.
- Undirected graphical models or pair-wise Markov Random Fields.
- \( X = (X_1, X_2, \ldots, X_p) \) is a random vector.
- Denote the graph \( G \) by a set of vertices \( V = \{1, \ldots, p\} \) and edges, \( E \subseteq V \times V \): \( G = \{V, E\} \).
- Lack of edges denote conditional independence given all other nodes (pairwise conditional independence).
- Captures direct dependencies between random variables.

\[ A \perp B \mid C \]
Markov Networks

- **Local Markov Property**: Conditional dependencies defined by node-neighborhoods, or the set of nodes connected to a given node via an edge.

- **Global Markov Property**: Pairwise conditional dependencies and neighborhoods jointly define the global dependence structure (formally defined by separators).

- **Hammersley-Clifford Theorem**: Density on graph factorizes according to sufficient statistics on cliques.
Review: Markov Networks

Graphical Models via Exponential Families
- Motivation
- Graphical Models via Exponential Families
- Results

Mixed Graphical Models
- Mixed MRFs
- Mixed Recursive Chain Graphical Models
- Results
Motivation: Networks from RNA-Sequencing Data

Gaussian Graphical Models have been widely used to infer genomic networks from microarray data:

Applications of Inferred Networks: Visualizing data, discovering biomarkers (hubs), regulatory pathways, potential drug targets.
Motivation: Networks from RNA-Sequencing Data

Next generation sequencing technology is rapidly replacing the microarray.

Gaussian Graphical Models not appropriate for next generation sequencing (RNA-seq) data!
Graphical Models from Count or Other Data Types?

- **Gaussian Graphical Model.**
  - Conditional & marginal distributions are Gaussian, jointly multivariate Gaussian.
  - Conditional independence given by zeros in inverse covariance matrix.
  - Sparse Graphical Model Estimation. (Meinshausen & Bühlmann, 2006; Yuan & Lin, 2007; Banerjee et al., 2008; Friedman et al., 2008)

- **Ising & Potts Model.**
  - Assumes node-conditional distributions are binomial / multinomial.
  - Sparse Graphical Model Estimation. (Ravikumar et al., 2010)
Mixed Conditional Gaussian Model & Chain Graphs.

- Mixed Chain Graphical Models. (Lauritzen and Wermuth (1989); Frydenberg (1990); Lauritzen (1996)).
  - Continuous variables conditioned on all combos discrete variables are multivariate Gaussian.
  - Number of parameters scale exponentially.
  - Can chain recursively.
  - Have the block concentration Markov property.

Graph selection for Mixed Chain Graphs:

- Learning the Structure of Mixed Graphical Models (Lee and Hastie (2012)).
- High-Dimensional Mixed Graphical Model (Cheng, Levina, Zhu (2013)).
Examples:

- Gaussian, Bernoulli, Poisson, Binomial, Negative Binomial, Exponential, ...

\[ P(Z) = \exp(\theta B(Z) + C(Z) - D(\theta)) \]

- \( \theta \) is the canonical parameter.
- \( B(Z) \) is the sufficient statistic.
- \( C(Z) \) is the base measure.
- \( D(\theta) \) is the log-partition function.
Graphical Models via Exponential Families

For a random vector $X = (X_1, X_2, \ldots X_p)$, suppose:

- Node-conditional distributions are univariate exponential family densities.
- Cliques are of order at most $k$.

**Theorem**

Joint Density necessarily has the form:

$$P(X) = \exp \left\{ \sum_s \theta_s B(X_s) + \sum_{s \in V} \sum_{t \in N(s)} \theta_{st} B(X_s)B(X_t) + \sum_{s \in V} \sum_{t_2, \ldots, t_k \in N(s)} \theta_{s\ldots t_k} B(X_s) \prod_{j=2}^{k} B(X_{t_j}) + \sum_s C(X_s) - A(\theta) \right\}$$

$N(s)$ denotes the neighborhood of node $s$ & $A(\theta)$ is the log-normalization term. (YRAL, 2012; YRAL, 2015)
Graphical Models via Exponential Families

Special Case:
- Pairwise interactions.
- Linear sufficient statistics $B(X_s) = X_s$.

**Joint Density**

$$P(X) = \exp \left\{ \sum_{s} \theta_s X_s + \sum_{(s,t) \in E} \theta_{st} X_s X_t + \sum_{s} C(X_s) - A(\theta) \right\}.$$ 

**Node-Conditional Density**

$$P(X_s | X_{V \setminus s}) \propto \exp \left\{ \left( \theta_s + \sum_{t \in N(s)} \theta_{st} X_t \right) X_s + C(X_s) \right\},$$

i.e. a Generalized Linear Model with response $X_s$ and predictors $X_{V \setminus s}$. 

G. I. Allen (Rice & BCM)
Integrative Networks
April 1, 2016 10 / 28
Graphical Models via Exponential Families

Example of Poisson Graphical Model (Count Data):

\[ P(X) = \exp \left\{ \sum_s \theta_s X_s + \sum_{(s,t) \in E} \theta_{st} X_s X_t + \sum_s \log(X_s!) - A(\theta) \right\}. \]

- Technical conditions needed to ensure proper densities.
  - \( \theta_{st} \leq 0 \) (only negative conditional dependencies!)

- Variations that permit negative & positive dependencies:
  - Truncation, Sub-linear, Quadratic, and approximations to these.

(YRAL, 2013; AL, 2012; AL, 2013)
Results: TCGA Breast Cancer microRNA

MicroRNAs:
- Short RNAs ($\approx 22$ base pairs).
- Post-transcriptional regulators.
- Predominately inhibitors of gene expression.

Data:
- 544 breast cancer tumor samples, 524 microRNAs.
Results: TCGA Breast Cancer microRNA

- hsa-mir-512-2
- hsa-mir-489
- hsa-mir-525
- hsa-mir-3941
- hsa-mir-618
- hsa-mir-449a
- hsa-mir-526b
- hsa-mir-512-1
- hsa-mir-449b
- hsa-mir-520b
- hsa-mir-517a
- hsa-mir-138-2
- hsa-mir-624
- hsa-mir-4326
- hsa-mir-632
- hsa-mir-3944
- hsa-mir-520d
- hsa-mir-514-1
- hsa-mir-514-3
- hsa-mir-520h
- hsa-mir-520g
- hsa-mir-499
- hsa-mir-676
- hsa-mir-581
- hsa-mir-580
- hsa-mir-3940
- hsa-mir-3927
- hsa-mir-320e
- hsa-mir-517b
- hsa-mir-518c
- hsa-mir-519c
- hsa-mir-449c
- hsa-mir-514-2
- hsa-mir-211
- hsa-mir-206
- hsa-mir-133a-2
- hsa-mir-133b
- hsa-mir-3150
- hsa-mir-3690
- hsa-mir-133a-1
- hsa-mir-891a
- hsa-mir-451
- hsa-mir-658
- hsa-mir-3652
- hsa-mir-3199-1
- hsa-mir-3194
- hsa-mir-607
- hsa-mir-378b
- hsa-mir-144
- hsa-mir-551b
- hsa-mir-184
- hsa-mir-224
- hsa-mir-378c
- hsa-mir-200a
- hsa-mir-342
- hsa-mir-126
- hsa-mir-145
- hsa-mir-643
- hsa-mir-22
- hsa-mir-99b
- hsa-let-7f-2
- hsa-mir-139
- hsa-mir-1284
- hsa-mir-142
- hsa-mir-187
- hsa-mir-1-2
- hsa-mir-183
- hsa-mir-934
- hsa-mir-21
- hsa-mir-452
- hsa-mir-203
- hsa-mir-200b
- hsa-mir-200c
- hsa-mir-148a
- hsa-mir-3622a
- hsa-mir-1254
- hsa-mir-143
- hsa-mir-1277
- hsa-mir-605
- hsa-mir-545
- hsa-mir-301b
- hsa-mir-33b
- hsa-mir-3664
- hsa-mir-150
- hsa-mir-10b
- hsa-mir-100
- hsa-mir-3176
- hsa-let-7c
- hsa-mir-636
- hsa-mir-29a
- hsa-mir-1251
- hsa-mir-1269
- hsa-mir-3619
- hsa-mir-3199-2
- hsa-mir-585
- hsa-mir-3651
- hsa-mir-579
- hsa-mir-556
- hsa-mir-210
- hsa-mir-639
- hsa-mir-592
- hsa-mir-3174
- hsa-mir-147b
- hsa-mir-3610
- hsa-mir-196a-2
- hsa-mir-486
- hsa-mir-196a-1
- hsa-mir-141
- hsa-mir-2277
- hsa-mir-1247
- hsa-mir-3136
- hsa-mir-1537
- hsa-let-7b
- hsa-let-7a-2
- hsa-mir-577
- hsa-mir-375
- hsa-mir-135b
- hsa-mir-103-1
- hsa-mir-767
- hsa-let-7a-3
- hsa-mir-548o
- hsa-mir-30d
- hsa-mir-92a-2
- hsa-mir-548s
- hsa-mir-7-2
- hsa-mir-25
- hsa-mir-338
- hsa-mir-149
- hsa-mir-138-1
- hsa-mir-29c
- hsa-mir-3678
- hsa-let-7a-1
- hsa-mir-548t
- hsa-mir-1224
- hsa-mir-516a-1
- hsa-mir-522
- hsa-mir-527
- hsa-mir-516a-2
- hsa-mir-518a-2
- hsa-mir-518a-1
- hsa-mir-518e
- hsa-mir-520a
- hsa-mir-518b
- hsa-mir-519d
- hsa-mir-520f
- hsa-mir-519a-2
- hsa-mir-548o
- hsa-mir-30d
- hsa-mir-92a-2
- hsa-mir-548s
- hsa-mir-7-2
- hsa-mir-25
- hsa-mir-338
- hsa-mir-149
- hsa-mir-138-1
- hsa-mir-29c
- hsa-mir-3678
- hsa-let-7a-1
- hsa-mir-548t
- hsa-mir-1224
- hsa-mir-516a-1
- hsa-mir-522
- hsa-mir-527
- hsa-mir-516a-2
- hsa-mir-518a-2
- hsa-mir-518a-1
- hsa-mir-518e
- hsa-mir-520a
- hsa-mir-518b
- hsa-mir-519d
- hsa-mir-520f
- hsa-mir-519a-2
- hsa-mir-548o
- hsa-mir-30d
- hsa-mir-92a-2
- hsa-mir-548s
- hsa-mir-7-2
- hsa-mir-25
- hsa-mir-338
- hsa-mir-149
- hsa-mir-138-1
- hsa-mir-29c
- hsa-mir-3678
- hsa-let-7a-1
- hsa-mir-548t
- hsa-mir-1224
- hsa-mir-516a-1
- hsa-mir-522
- hsa-mir-527
- hsa-mir-516a-2
- hsa-mir-518a-2
- hsa-mir-518a-1
- hsa-mir-518e
- hsa-mir-520a
- hsa-mir-518b
- hsa-mir-519d
- hsa-mir-520f
- hsa-mir-519a-2
- hsa-mir-548o
- hsa-mir-30d
- hsa-mir-92a-2
- hsa-mir-548s
Results: TCGA Breast Cancer microRNA

Selected Findings:

- Let-7c: Regulates breast cancer metastatic.
- MiR-10b: Over-expressed in triple negative (ER-, PR-, Her2-) breast cancer patients.
- MiR-375: Blocking slows down tumor progression in ER+ patients.
- MiR-10-b & MiR-375 estimated to be conditionally independent.

(AL, 2013)
Results: TCGA Breast Cancer microRNA

Comparisons: Poisson vs. Gaussian Graphical Models

Sub-Linear Poisson Graphical Model

Gaussian Graphical Model
Review: Markov Networks

Graphical Models via Exponential Families
- Motivation
- Graphical Models via Exponential Families
- Results

Mixed Graphical Models
- Mixed MRFs
- Mixed Recursive Chain Graphical Models
- Results
Motivation: Integrative Genomics

Graphical Models via Exponential Families
Motivation: Integrative Genomics

No general multivariate density that directly parameterizes dependencies for mixed variables exists!
Mixed Markov Random Fields

Building Mixed MRFs:

- $p$-variate random vector: $X = (X_1, X_2, \ldots, X_p)$ where $X_s \in \mathcal{X}_s$ potentially all distinct types.

- Node-Conditional Distribution: $P(X_s|X_{\setminus s})$ a univariate exponential family.

$$P(X_s|X_{\setminus s}) = \exp \left\{ g_s(X_{\setminus s}; \theta) B_s(X_s) + C_s(X_s) - D_s(X_{\setminus s}; \theta) \right\}$$

- Canonical parameter, $g_s(X_{\setminus s}; \theta)$, a function of other nodes (like GLMs).

- $B_s(X_s)$ is the sufficient statistic.

- $C_s(X_s)$ is the base measure.

- $D_s(X_{\setminus s}; \theta)$ is the log-partition function.
Theorem

Joint Density has the form:

\[
P(X) = \exp \left\{ \sum_{s} \theta_s B_s(X_s) + \sum_{s \in V} \sum_{t \in N(s)} \theta_{st} B_s(X_s)B_t(X_t) \\
+ \sum_{s \in V} \sum_{t_2, \ldots, t_k \in N(s)} \theta_{s \ldots t_k} B_s(X_s) \prod_{j=2}^{k} B_t(X_{t_j}) + \sum_{s} C_s(X_s) - A(\theta) \right\}
\]

\(N(s)\) denotes the neighborhood of node \(s\) & \(A(\theta)\) is the log-normalization term. (*YRAL*, 2014)
Mixed Markov Random Fields

Special Case: Two partitions of variables \((X, Y)\) & pairwise graph.

\[
P(X, Y; \theta) = \exp \left\{ \sum_{s \in V_X} \theta_s B_X(X_s) + \sum_{s' \in V_Y} \theta_{s'} B_Y(Y_{s'}) + \sum_{(s, t) \in E_X} \theta_{st} B_X(X_s) B_X(X_t) + \sum_{(s', t') \in E_Y} \theta_{s't'} B_Y(Y_{s'}) B_Y(Y_{t'}) + \sum_{(s, s') \in E_{XY}} \theta_{ss'} B_X(X_s) B_Y(Y_{s'}) \right. \\
+ \left. \sum_{s \in V_X} C_X(X_s) + \sum_{s' \in V_Y} C_Y(Y_{s'}) - A(\theta) \right\}
\]

\[A(\theta) := \log \int_{X^p} \exp \left\{ \sum_{s \in V_X} \theta_s B_X(X_s) + \sum_{s' \in V_Y} \theta_{s'} B_Y(Y_{s'}) + \ldots + \sum_{s' \in V_Y} C_Y(Y_{s'}) \right\}
\]

\(B_X(\cdot), C_X(\cdot)\) sufficient stat. and base measure for node-conditional of \(X\)
\(B_Y(\cdot), C_Y(\cdot)\) sufficient stat. and base measure for node-conditional of \(Y\)

\(A(\theta)\) log-partition function
Mixed MRFs

Advantage:
- General mixed multivariate distribution exists!

Caveat:
- Stringent Normalizability Assumptions.
  - $A(\theta) < \infty$.
  - No distribution exists linking Poisson and Gaussian variables.
Mixed MRFs

Advantage:
- General mixed multivariate distribution exists!

Caveat:
- Stringent Normalizability Assumptions.
  - $A(\theta) < \infty$.
  - No distribution exists linking Poisson and Gaussian variables.

Solution:
- Chain rule of conditional probability: $P(X, Y) = P(Y|X)P(X)$. 
Mixed Chain MRFs: Elementary Construction

Partition \( p \) variables into two groups: \( X = \{Y, Z\} \):

\[
P(X) = P_1(Y|Z)P_2(Z)
\]

- \( P_1 \) is a **Conditional Markov Random Field** constructed via node-conditional exponential families.
  - Heterogeneous (Mixed).
  - Homogeneous.

- \( P_2 \) is a **Markov Random Field** constructed via node-conditional exponential families.
  - Heterogeneous (Mixed).
  - Homogeneous.

(YRABWL, 2015)
Mixed Chain MRFs: Elementary Construction

Homogeneous Elementary Mixed Chain MRF:

Heterogeneous Elementary Mixed Chain MRF:
Multiple Blocks: Recursive Mixed Chain Graphs

Idea: Recursively apply chain rule to blocks of variables.

\[ P(X, Y, Z) = P(X|Y, Z)P(Y|Z)P(Z) \]
Multiple Blocks: Recursive Mixed Chain Graphs

Will this yield a network with Markov properties?

- Yes! Block concentration Markov property. (Lauritzen, 1996)
- Special cases - global Markov property for network skeleton.

Blocks known & Partial ordering of blocks known:

- Blocked Directed Acyclic Graph (DAG):
  - Within Block: Undirected edges.
  - Between Blocks: Directed edges (no cycles!).
Multiple Blocks: Recursive Mixed Chain Graphs

Mutations (SNP)

Methylation

Copy Number Variation

Gene Expression

MicroRNA Expression

G. I. Allen (Rice & BCM)
Mixed Chain MRFs: Properties

How does this compare to the Mixed MRF?

Example: $P(Y, X) = P(Y|X)P(X)$ for 2-block homogeneous, pairwise graphs.

$$P(X, Y; \theta) = \exp \left\{ \sum_{s \in V_X} \theta_s B_X(X_s) + \sum_{s' \in V_Y} \theta_{s'} B_Y(Y_{s'}) ight.$$ 

$$+ \sum_{(s, t) \in E_X} \theta_{st} B_X(X_s) B_X(X_t) + \sum_{(s', t') \in E_Y} \theta_{s't'} B_Y(Y_{s'}) B_Y(Y_{t'})$$

$$+ \sum_{(s, s') \in E_{XY}} \theta_{ss'} B_X(X_s) B_Y(Y_{s'}) + \sum_{s \in V_X} C_X(X_s) + \sum_{s' \in V_Y} C_Y(Y_{s'})$$

$$- A_{Y|X}(\theta) - A_X(\theta) \right\}$$

$A_{Y|X}(\theta)$ normalizability term for CRF.

$A_X(\theta)$ normalizability term for MRF.
Mixed Chain MRFs: Properties

What does this mean in terms of normalizability?

Theorems

1. Normalizable if CRFs and MRFs are normalizable.
2. Normalizability conditions strictly weaker than Mixed MRFs.

Implication: Dependencies permitted in Mixed Chain MRFs that are not in Mixed MRFs.
Mixed Chain MRFs: Properties

Gaussian-Poisson Chain MRF

Let $X$ be counts (Poisson) and $Y$ continuous (Gaussian).

- Mixed MRF does not permit dependencies between $X$ and $Y$.
- Mixed Chain MRF $P(X, Y) = P(Y|X)P(X)$ exists!
- Mixed Chain MRF $P(X, Y) = P(X|Y)P(Y)$ exists!

$$
\log P(X, Y) = \sum_{s \in V_Y} \left( \theta_s + \sum_{t \text{ s.t } (s,t) \in E_{XY}} \theta_{st} X_t \right) \frac{Y_s}{\sigma_s} + \sum_{(s,s') \in E_Y} \theta_{ss'} \frac{Y_s Y_{s'}}{\sigma_s \sigma_{s'}} - \sum_{s \in V_Y} \frac{Y_s^2}{2\sigma_s^2}
+ \sum_{X \in V_X} \theta_t X_t + \sum_{(t,t') \in E_X} \theta_{tt'} X_t X_{t'} - \sum_{t \in V_X} \log(X_t!) - A_{Y|X}(\theta(X)) - A_X(\theta).
$$
Mixed Chain MRFs: Properties

Implications:

- General, flexible class of MRF models.
  - Mixed variables.
  - Mixed edges (both directed and undirected).

- First multivariate density that directly parameterizes dependencies over mixed variables!
Graph Selection and Estimation

**Objective**: Given iid observations, seek to learn graph structure (selection) and parameters (estimation).

- Density factors into products of CRFs $\Rightarrow$ Can learn each CRF separately.
- Each CRF factors by node-neighborhood $\Rightarrow$ Can learn each node neighborhood separately.
- **Neighborhood Selection** - For each node:
  - Maximize penalized conditional likelihood $= \text{Mixed, penalized GLMs!}$

\[
\ell_1 \text{ regularized MLE (Gaussian node-conditional):}
\]

\[
\min_{\theta_{xy}, \theta_{xx}} \left\| X_s - X/s \theta_{xx} - Y \theta_{xy} \right\|^2_2 + \lambda_{xx} \left\| \theta_{xx} \right\|_1 + \lambda_{xy} \left\| \theta_{xy} \right\|_1,
\]
Graph Selection and Estimation

Theoretical Guarantees

Assumptions:
- Restricted Eigenvalue condition (on Fisher Information matrix).
- $\lambda \propto \sqrt{\log(p_x + p_y)/n} \& \ n \propto d^2 \log(p)$ (degree squared).

Result:
- Unique solution.
- With high probability, exactly recover the true edge structure.
- Consistent parameter estimation.
Simulation Study

- Samples generated via Gibbs sampling.
- Lattice structure
- $p = 72$: $p_Y = 36$, $p_Z = 36$
- Sample sizes: $n=50$, 72, 100 and 200.
Simulation Study

Figure: ROC curves for different types of models when $p_Y = 36, p_Z = 36$. 

(d) Poisson-Ising Mixed MRF  (e) Poisson MRF-Ising CRF  (f) Poisson CRF-Ising MRF

(g) Gaus CRF-TPGM MRF  (h) Exp MRF-Ising CRF  (i) Gaus CRF-Poisson MRF
Simulation Study

Figure: ROC curves for 3 blocks of variables: binary (Ising, X), continuous (Gaussian, Y), and counts (Poisson, Z).
Case Study: Breast Cancer Genomics

Objective: Identify both between and within connections between mutation and expression biomarkers.

- Gene expression: TCGA Level III RNA-sequencing (counts).
- Mutations & Aberrations: Combination of TCGA Level II non-silent somatic mutation and TCGA Level III copy number variation (binary).
- 697 patients and 498 genes (329 expression biomarkers & 169 mutation biomarkers).
- Modeled via Poisson CRF- Ising MRF (mutations influence expression).
- Stability selection for model selection.

(YRABWL, 2015)
Case Study: Breast Cancer Genomics

Blue nodes: RNA-sequencing; Yellow nodes: genomic mutations
Discovery of Previously Indicated Links:

- **GATA3 mutation** linked to SLC39A6 expression.
  - Ratio of gene expression levels used to define breast cancer sub-types.

- **FGFR1 mutation** linked to PEG3 expression.
  - FGFR1 growth factors amplified in breast cancer work with PEG3 which modulates cancer progression.

- **STAT3 mutation** linked to ERBB2 expression.
  - Amplified in HERB2 sub-types and promotes cancer stem-cell proliferation.
Case Study: Breast Cancer Genomics

Novel Discoveries:

- **TP53** mutation linked to **ADAM6** expression.
  - TP53 a tumor suppressor gene & ADAM6 a long non-coding RNA over-expressed in breast cancer.

- **FGF3** mutation linked to **CCND1** expression.
  - FGF3 regulates estrogen expanding breast cancer stem cells & CCDN1 over-expression of hormone receptors in breast cancer.

- **PIK3CA** mutation linked to **CLEC3A** expression and **NAT1** expression.
  - PIK3CA an oncogene, CLEC3A affects tumor metastasis, and NAT1 a potential marker for estrogen receptor positive sub-type.
### Summary

<table>
<thead>
<tr>
<th>Graphical Models via Exponential Families</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stat: Extends Markov Networks for (almost) any data type!</td>
</tr>
<tr>
<td>Application: Poisson Graphical Models for RNA-Seq data.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Mixed Graphical Models</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stat: <strong>First direct multivariate density for mixed data types!</strong></td>
</tr>
<tr>
<td>Stat: Permits both directed AND undirected edges - flexible class of models.</td>
</tr>
<tr>
<td>Application: Can model connections within and between multiple types of biomarkers.</td>
</tr>
</tbody>
</table>
Future Work

- Mutations (SNP)
- Methylation
- Copy Number Variation
- MicroRNA Expression
- Gene Expression
Acknowledgments

Collaborators:

- **Pradeep Ravikumar**, University of Texas, Austin.
- **Eunho Yang**, IBM Watson Research Center.
- **Yulia Baker**, PhD Candidate, Statistics, Rice University.
- **Zhandong Liu**, Baylor College of Medicine.
- **Ying-Wooi Wan**, Baylor College of Medicine.
- **Matthew Anderson**, Baylor College of Medicine.
Software

**XMRF**: An R Package to Fit Markov Networks to High-Throughput Genomics Data.

**TCGA2STAT**: Simple TCGA Data Access for Integrated Statistical Analysis in R.
Major References


G. I. Allen and Z. Liu, ”A Log-Linear Graphical Model for Inferring Genetic Networks from High-Throughput Sequencing Data”, In IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2012.

Thank You!