Statistical Data Integration
Using Networks to Integrate a Variety of Big Biomedical Data

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.

May 21, 2015
Biomedical Data on a Massive Scale

- ADNI (Alzheimer's Disease Neuroimaging Initiative)
- CURE (Cancer United for Research in Epilepsy)
- ABIDE (Autism Brain Imaging Data Exchange)
- Human Connectome Project
- The Cancer Genome Atlas

- 33 cancer types
- Over 11,000 patients
- ~10 types of molecular profiling per sample
Biomedical Data on a Massive Scale

The Cancer Genome Atlas

Mutations (SNP)

Methylation

Epigenetics

Sequence Changes

Copy Number Variation

MicroRNA Expression

Gene Expression

Functional Genetics
Biomedical Data on a Massive Scale

The Cancer Genome Atlas

Understanding genomics to improve cancer care

Mutations (SNP)
- ~100K – 20 Million
- Binary / Categorical

Methylation
- ~30K – 450K
- Bounded Continuous

Copy Number Variation
- ~20K – 200K
- Continuous

MicroRNA Expression
- ~1K – 10K
- Continuous (array)
- Counts (Seq)

Gene Expression
- ~1K – 10K
- Continuous (array)
- Counts (Sequencing)
Uses probabilistic models to jointly model multiple types of measurements taken on the same set of subjects.
Statistical Data Integration

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

Advantages:

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

- Discover relationships between different types of biomedical info.
  - Ex: EHRs, genomics, medical imaging.
Statistical Data Integration: Markov Networks

Why Networks?

- Visualize Big Data.
- Discover relationships between biomarkers.
- Model complex systems.

Markov Networks:

- Undirected graphical models based on conditional dependencies.
- Probabilistic model - proper multivariate distribution.
Networks for Different Data Types

Existing (Markov) Network Types:

1. Gaussian Graphical Models (Continuous-Valued).

2. Ising Models (Binary-Valued).

Networks for Different Data Types

Existing (Markov) Network Types:

1. Gaussian Graphical Models (Continuous-Valued).
2. Ising Models (Binary-Valued).

Modules from lung cancer somatic mutation network.
Networks for Different Data Types

Existing (Markov) Network Types:

1. Gaussian Graphical Models (Continuous-Valued).
2. Ising Models (Binary-Valued).
Networks for Different Data Types

Existing (Markov) Network Types:

1. Gaussian Graphical Models (Continuous-Valued).
2. Ising Models (Binary-Valued).

What about count-valued data? Others?

RNA-sequencing data? Text data? Geographic coordinates?
Networks for Different Data Types

**Our Framework**: Graphical Models via Exponential Families.

- **Assumption**: Conditional distributions are Exponential Families.
  - Ex: Gaussian, Bernoulli, Poisson, Exponential, Negative Binomial, etc.

  A lot of math . . .

- **Theorem**: Joint network distribution exists and has a closed form!
  - Dependencies parameterized by products of sufficient statistics.
  - Strong statistical guarantees for network inference.
  - Fast, parallelizable algorithm to learn network structure.
Networks for Different Data Types

Our Framework: Graphical Models via Exponential Families.
Networks for Different Data Types

Our Framework: Graphical Models via Exponential Families.

Lung Cancer Gene Expression Network (via RNA-Seq).
Integrated Network Models

Mutations (SNP) → Methylation

Copy Number Variation

Gene Expression

MicroRNA Expression
Integrated Network Models

Block-Directed Graphical Models via Exponential Families.

Assumptions:
▶ Conditional distributions are Exponential Families.
▶ Variables belong to known groups and the directionality of dependencies between groups is known.

A lot of math . . .

Theorem: Joint integrated network distribution exists and has a closed form!
▶ Dependencies parameterized by products of sufficient statistics from different distributions.
▶ Strong statistical guarantees for network inference.
▶ Fast, parallelizable algorithm to learn network structure.
Integrated Network Models

Block-Directed Graphical Models via Exponential Families.
Integrated Network Models

Block-Directed Graphical Models via Exponential Families.

Mutations (SNP) → Methylation

Copy Number Variation

Gene Expression

MicroRNA Expression
Applications & Implications

Implication

First multivariate distribution for mixed data types.
Applications & Implications

Glioblastoma Integrated Network.
Applications & Implications

Breast Cancer Integrated Mutation-Gene Expression Network.

Blue nodes: RNA-sequencing
Yellow nodes: genomic mutations
Applications & Implications

UCLA CART
CENTER FOR AUTISM RESEARCH AND TREATMENT
Applications & Implications

ADNI
ALZHEIMER'S DISEASE NEUROIMAGING INITIATIVE

CURE
EPILEPSY

G. I. Allen (Rice & BCM)
Acknowledgments

Allen Group:

- Manjari Narayan
- Frederick Campbell
- Yue Hu
- John Nagorski
- Yulia Baker
- Qijia Jiang
- Connor Barnhill

Collaborators:

- Pradeep Ravikumar, UT Austin
- Eunho Yang, IBM Watson
- Zhandong Liu, BCM
- Matthew Anderson, BCM
- Ying-Wooi Wan, BCM

www.stat.rice.edu/~gallen/