What we do

• Frame biomedical problems as computational problems, especially on “omic” scale
  – What approaches ‘scale up’?
• Use and develop appropriate computational methods
• Apply to real-world problems, especially cancer
What we do

• Seek to minimize input in the form of “biological knowledge” – as much as possible tie everything to primary data

• Emphasize role for computational biology in questioning assumptions and frameworks for representation

• Seek to drive biomedical research through computation
What we don’t

• Medical Informatics
  – Patient Records
    • Entry, Storage, Retrieval, Privacy
• Biological Computation
Faculty

• Core of Lane Center faculty with strong research programs related to computational biology
• Many affiliated faculty
Particular Research Strengths

• Interaction network inference
  – Bar-Joseph, Xing, Kim, Kingsford, Lopez, Hinman, McManus, Roeder, Schwartz, Wu

• Bioimage informatics
  – Murphy, Rohde, Kovacevic, Yang

• Multiscale modeling
  – Schwartz, Langmead, Murphy

• Active learning
  – Carbonell, Murphy, Schneider
Computational Genomics
- Understanding the genetic basis of common diseases (diabetes, asthma, etc.)
- Gene network learning, evolutionary modeling
- Long-term goal: personalized medicine

Machine Learning
- Learning in high-dimensional space
- Probabilistic graphical models
- Feature selection, structured sparsity
- Fast optimization methods

Disease causing genetic variations
Genetic Basis of **Complex** Diseases

Association to intermediate phenotypes

Causal SNPs

Gene expression

Clinical records

Healthy

Cancer

Seyoung Kim
The eukaryotic genomes are huge collections of molecules packed into a small space.

“3C” experiments produce a graph where nodes are genome locations and edges give weights related to the # of times the locations were spatially close.

Q: Can we model the spatial structure from pairwise proximities measured from a population of cells?

e.g. Duggal et al., WABI 2013, and Duggal et al., Alg. Mol. Biol. 2012

Q: Can we develop efficient statistical tests to determine if a set of genomic positions are significantly close or not?

e.g. Wang et al., BCB 2013
Fast Genomic Sequence Analysis

Huge amount of genomic sequence data available now (a single public database has > 1,587 terra bases of sequence)

Our tool, called Sailfish, can estimate gene expression from next-generation sequencing data 20 – 70 times faster than existing algorithms.

Patro et al., arXiv:1308.3700

Speed will be crucial for personal genomics.

Q: How do you efficiently search for complex structures (e.g. spliced sequences, patterns of TF binding sites) when traditional sequence alignment is too slow?

Q: How can you transmit large collections of sequences (between collaborators, from the sequencer, to the cloud)?

Q: How can you compress sequence data so that you can still analyze it in its compressed form (“functional compression”)?
Modeling dynamic networks with Input – Output Hidden Markov Models

Hidden States (transitions between states form a tree structure)

Emissions (Distribution of expression values)

Fly development

Science 2010

Ziv Bar-Joseph
Russell Schwartz

Tumor evolution and heterogeneity

Simulation and model inference for macromolecular assembly

Models and algorithms for phylogenetics and population genetics
ACTIVE LEARNING FOR DRUG DISCOVERY

Murphy, Schneider, Langmead
Assumption/Framework

• Exhaustive experimentation will permit understanding of biological systems
  – We can always do whatever experiments/measurements needed
  – Drug development can be done by focusing initially on specific target and then checking toxicity of chosen drugs
Problem

- Drugs fail late in development because of unanticipated side effects
- Only real solution is to choose drugs with desired effect on target and no undesired effect
- This requires determining the effects of millions of potential drugs on tens of thousands of potential targets
Solution

- Build model to predict full matrix from whatever data we have
- Use active learning to choose new experiments

- This and other relevant topics covered in Course 02-750 Automation of Biological Research
Paradigm shift

• Exhaustive experimentation will permit understanding of biological systems

• Paradigm shift: Computer control over experiment choice – active learning

• New company, Quantitative Medicine, commercializing this technology
IMAGE-DERIVED GENERATIVE MODELS OF SUBCELLULAR ORGANIZATION

Murphy, Rohde
Assumption/Framework

• Words are a good way of representing information about the spatial organization of cells and the subcellular localization of proteins
How do we learn and represent

- the number, sizes, shapes, positions of subcellular structures
- the distribution of proteins across those structures
- how structures and distributions change between cell types, in presence of perturbagens, or over time?
Subcellular Location Analysis

“Where” (in which type of organelle) is the protein of interest located?

Lysosome?
Mitochondria?
Golgi?

......
Descriptive vs. Generative Models

• If the task is to test which of two (or more) possibilities is true, can use *descriptive features*
  – Is this an apple or orange? can be answered by measuring color or texture

• But if the task is to understand as much as possible, a *generative model* is better
  – What does an apple look like? requires a generative model
Alternative: Generative Modeling

- Human cognition
  - 555 examples
  - Learn
  - Mental model
  - Write
  - Generated examples

- Generative model
  - Training images
  - Statistical generative model
  - Generated image
CellOrganizer

Statistical Model

- Nuclear shape
- Cell shape
- Object pos. probability
- Microtubule distribution
- Object appearance
- Object positions
- Object number
- Object distribution

Training

Synthesis

Cell Images

Synthetic Images

Zhao & Murphy, Cytometry 2007
Synthetic movie of cell/nuclear shape dynamics

3D HeLa
Synthetic movie of cell and nuclear shape changes during neuronal differentiation

hr: 0
Big Future Issues

- Learning multiscale dynamics
- Learning deeper conditional structure (pattern causality)
  - Organelles on cell framework
  - Organelles on organelles
  - Proteins on organelles
  - Proteins on proteins
  - All of above on perturbagens
Pattern Causality

• Need variations on existing methods (such as Granger Causality, Convergent cross mapping) appropriate for images/spatial distributions
Paradigm shift

• Words are a good way of representing information about the spatial organization of cells and the subcellular localization of proteins

• Paradigm change: Generative statistical models
Summary

• Computational biology research requires investigators with deep knowledge of biology, computer science, math, statistics to develop rigorous approaches and reformulate paradigms for biomedical research.
Summary

• Opportunities for computer scientists to
  – help solve framed computational biology problems
  – help formalize solutions, e.g.
    • prove convergence
    • establish bounds