Paradigm

Some slide sources:
- Josh Stuart (UCSC) AACR 12 slides
- Carl Edward Rasmussen (Cambridge) Factor graphs slides
Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM

Charles J. Vaske, Stephen C. Benz, J. Zachary Sanborn, Dent Earl, Christopher Szeto, Jingchun Zhu, David Haussler, Joshua M. Stuart,
UC Santa Cruz, CA, USA
Bioinformatics 2010
The challenge of cancer

• 25% of breast cancer patients show amplification/over-expression of ERBB2 – can be treated by trastuzumab
• But <50% of them show any improvement
• Why?
  - We do not understand enough about the cancer process
  - What we call cancer types are actually composites of many subtypes
  - Each patient is different
• Can integrating multiple data sources help?
• Can taking a pathway perspective to cancer help?
• Can patient-specific perspective help?
Overview

- Input: patient expression and copy number profile; curated networks of cancer related pathways
- Goal: Infer patient-specific gene activity in each pathway
- Model via factor graphs
First - some background

Copy number profiles
Biological networks
Factor graphs
Copy number changes in cancer

Cancer cells undergo numerical and structural aberrations
Numerical changes: gain/loss of chromosomal segments or whole chromosomes
Copy number profiles

One patient

933 patients

TCGA breast invasive carcinoma (BRCA) segmented copy number (delete germline cnv) • N = 933

https://genome-cancer.ucsc.edu/
Biological Networks
Gene expression

proB → expression → γ-glutamyl kinase

γ-glutamyl kinase → catalyzes

2.7.2.11 → Produces

- substrates:
  - glutamate
  - ATP

- products:
  - ADP
  - γ-glutamyl phosphate

Gene

Positive interaction

Protein

EC (reaction) number

1.5.1.2

Compound
Metabolic Pathway: Proline Biosynthesis

- **proB**
  - expression
  - catalyzes gamma-glutamyl kinase
  - inhibits

- **proA**
  - expression
  - catalyzes gamma-glutamylphosphate reductase

- **proC**
  - expression
  - catalyzes 1-pyrroline-5-carboxylate reductase

1. **glutamate**
   - ATP
   - ADP
   - catalyzes 2.7.2.11
     - gamma-glutamyl phosphate
     - NADPH; H⁺
     - NADP; Pi
     - glutamate gamma-semialdehyde

2. **spontaneous**
   - H₂O
     - catalyzes 1.5.1.2
       - 1-pyrroline-carboxylate
       - NADPH
       - NADP

3. **proline**
Methionine Biosynthesis in *E. coli*

**Diagram Description:**
- **MetB operon**
  - Expression regulated by **MetL** and **MetB**.
- **Holorepressor**
  - Regulated by **MetC** and **MetE**.
  - Expresses **aspartate kinase II** and **homoserine dehydrogenase II**.
- **MetL operon**
  - **MetL** and **MetJ** express **metL and metJ** respectively.
  - **MetR** activates **metR**.
- **Expression**
  - **aspartate semialdehyde deshydrogenase**
  - **Cystathionine-gamma-synthase**
  - **Cystathionine-beta-lyase**
  - **Homocysteine methylation**
  - **Homocysteine transmethylation**
- **Metabolites**
  - **L-aspartate**
  - **L-Aspartate (4-P)**
  - **Homoserine**
  - **Homocysteine**
  - **Succinyl-CoA**
  - **Succinate**
  - **Cystathionine**
  - **Cysteine**
  - **5-Methyl THF**
  - **L-Methionine**
  - **L-Adenosyl-L-Methionine**

**Sources:**
- **http://www.ebi.ac.uk/research/pfmp**
- **ABDBM © Ron Shamir**
Shortcut Representation
G1/S signaling pathway

http://www.cs.tau.ac.il/~spike/
Genetic network controlling early development of sea urchin endomesoderm.
Eric Davidson
Biological network resources

• Many great databases summarizing
  - Specific pathways: everything we know about interactions of genes/proteins in pathway X of species Y
  - Protein interactions of various types (some with confidence values)

• Lots of data

• Very noisy
A typical hairball
Factor Graphs

- $x_1,..x_n$ variables over domains $A_1,..A_n$
- $g(x_1,..x_n) : A_1 \times .. \times A_n \rightarrow R$
  \[ g(x_1,..x_n) = f_A(x_1, x_2) \cdot f_B(x_2, x_3, x_4) \cdot f_C(x_3, x_4) \]
- $g$ factors into local functions
  \[ g(x_1,..x_n) = \prod_j f_j(X_j) \]
- Each $X_j$ is a subset of $\{x_1,..x_n\}$
- $f_j(X_j)$ is a function with arguments from $X_j$
- A factor graph has a variable node for each variable $x_i$, a factor node for each local function $f_j$ and an edge-connecting $x_i$ to $f_j$ iff $x_i$ is an argument of $f_j$.
- A factor graph expresses the structure of the factorization
Marginal functions

- For $g(x_1,..x_n)$ define **marginal function** $g_i(x_i)$
- $g_i(a) = \text{sum of } g \text{ over all configurations with } x_i=a$
- E.g. if $h(x_1,x_2,x_3)$

$$\sum_{\sim\{x_2\}} h(x_1, x_2, x_3) = \sum_{x_1 \in A1} \sum_{x_3 \in A3} h(x_1, x_2, x_3)$$

- So

$$g_i(x_i) = \sum_{\sim\{x_i\}} g(x_1,..,x_n)$$

- For $g$ corresponding to joint probability distribution this will be the marginal prob (perhaps up to a normalization factor)
Rasmussen’s slides on factor graphs

http://mlg.eng.cam.ac.uk/teaching/4f13/1718/
The algorithm

- Sum-product theorem: If the factor graph for some function $f$ has no cycles, then

$$p(w) = \prod_i m_{f_i\rightarrow w}(w)$$

- Alg called sum-product alg or message passing or belief propagation.

- Compute from the leaves in - and then out to get all marginals
variable to local function:
\[
\mu_{x \rightarrow f}(x) = \prod_{h \in n(x) \setminus \{f\}} \mu_{h \rightarrow x}(x)
\]

local function to variable:
\[
\mu_{f \rightarrow x}(x) = \sum_{\sim \{x\}} \left( f(X) \prod_{y \in n(f) \setminus \{x\}} \mu_{y \rightarrow f}(y) \right)
\]
Alg for factor graphs with cycles

- Iterate until change is $< \varepsilon$ or the num of iterations exceeds an input limit
- No guarantee, but often practical
- Numerous applications from optimization, statistics, information theory (Kalman filtering, error correcting codes...), ML (HMM, BN,..)!

Sources:
- Kschischang et al. IEEE Trans Inf Theo 01
- Leoliger IEEE Sig Proce Magazine 04
Back to Paradigm
Integration key to interpret gene function

- Expression not always an indicator of activity
- Downstream effects often provide clues

**Inference:**
- **TF is ON** (expression reflects activity)
- **TF is OFF** (high expression but inactive)
- **TF is ON** (low-expression but active)
Integration key to interpret gene function

• Need multiple data modalities to get it right.

BUT, targets are amplified

Expression -> TF ON

Copy Number -> TF OFF

Lowers our belief in active TF because explained away by CN evidence.
Integration Approach: Detailed models of gene expression and interaction

- CDKN2A (P16/INK4A): Homozygous deletion, homozygous mutation in 52%
- CDKN2B: Homozygous deletion in 47%
- CDKN2C: Homozygous deletion in 2%
- Amplification in 18%
- Amplification in 2%
- CDK4, CCND2, CDK6
- RB1: Homozygous deletion, mutation in 11%
- CDKN2A (ARF), homzygous deletion, mutation in 49%
- RB signaling altered in 78%
- G1/S progression
- RTK/RAS/PI(3)K signaling altered in 88%
- Mutation, amplification in 45%
- Mutation, amplification in 8%
- Mutation, amplification in 13%
- Mutation, homzygous deletion in 4%
- NF1: Mutation in 18%
- Mutation, homzygous deletion in 18%
- RAS: Mutation in 2%
- PI(3)K: Mutation in 15%
- PTEN: Homozygous deletion in 36%
- AKT: Amplification in 2%
- FOXO: Mutation in 1%
- Proliferation survival translation
- Activated oncogenes
- p53 signaling altered in 87%
- Senescence
- TP53: Mutation, homozygous deletion in 35%
- Apoptosis
Integration Approach: Detailed models of expression and interaction

Two Parts:

1. Gene Level Model
   (central dogma)
2. Interaction Model
   (regulation)
Assumptions

• If gene A is an activator and is upstream from B in some pathway,
  - their expression should be correlated
  - The copy number of A and the expression of B should be correlated, though less

• Is this observable on real data?
A is activator of B according to NCI cancer pathways; GBM data
E2E: correlation of exp(A) and exp(B)
C2E: correlation of CN(A) and exp(B)
Histogram for 462 A,B pairs
Factor Graph representation of a pathway

**Variables** - states of entities in a cell, (e.g. a particular mRNA or complex) represent the *differential state* of each entity in comparison with a ‘control’ or normal level.

Each of $X = \{X_1, \ldots, X_n\}$ is a random variable taking values -1, 0, or 1.

**Factors** - interactions and information flow between these entities. Constrain the entities to take biologically meaningful values. j-th factor $\phi_j(X_j)$ is a probability distribution over a subset $X_j$ of $X$.

Joint probability distribution of all entities:

$$P(X) = \frac{1}{Z} \prod_{j=1}^{m} \phi_j(X_j)$$
Toy example
Integrated Pathway Analysis for Cancer

- Integrated dataset for downstream analysis
- Inferred activities reflect neighborhood of influence around a gene.
- Can boost signal for survival analysis (later: also mutation impact)
Model construction

• Create a directed graph. For each gene, nodes: DNA, exp, protein, active protein
  - Edges have positive/negative label
  - Pos edges DNA $\rightarrow$ exp $\rightarrow$ protein $\rightarrow$ active-protein
  - Pos/neg edges active-prot1 $\rightarrow$ prot2 using the pathway info
• For variable $x_j$ add a factor $\phi_j (X_j)$ where $X_j = \{x_j\} \cup$ Parents ($x_j$)
  • Expected value is set by majority vote of parent edges: positive: $+1*$state, negative: $-1*$state
• Other rules for AND, OR relations

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

• Observed variables:
  - DNA: copy number
  - mRNA: transcription level

• All values of the same data type from all samples are ranked from smallest to largest and mapped to [0,1]

• All variable values discretized to ternary values

• Different observed data D for each patient, same factor graph $\Phi$ (per pathway)

• Inference: Compute $P(x_i=a|\Phi)$ and $P(x_i=a,D|\Phi)$ using belief propagation
Inference

- D={x1=s1,...,xk=sk} the observed data for a patient
- Let $S \subseteq _D X$ represent the set of all possible assignments to a set of variables X that are consistent with the assignments in D
- Want to estimate the state of a hidden vbl $x_i$

\[ P(x_i=a|\Phi) = \frac{1}{Z} \prod_{j=1}^{m} \sum_{S \sqsubseteq A_i(a)X_j} \phi_j(S), \]

- Pr $x_i=a$ along with all the patient data:

\[ P(x_i=a,D|\Phi) = \frac{1}{Z} \prod_{j=1}^{m} \sum_{S \sqsubseteq A_i(a) \cup D X_j} \phi_j(S). \]
Learning

• Inference: Compute $P(x_i=a,D|\Phi)$ using belief propagation
• Learning values of parameters: EM (details skipped)

• More nodes for
  - complexes
  - gene families
  - …
IPA scores

- Compute **IPA**: integrated pathway activity score per gene similar to log-likelihood ratios.
- Use log likelihood ratio: how much the data $D$ increases our belief that the entity is up/down.

\[
L(i, \alpha) = \log\left(\frac{P(D, x_i=\alpha)}{P(D, x_i \neq \alpha)}\right) - \log\left(\frac{P(x_i=\alpha)}{P(x_i \neq \alpha)}\right)
\]

\[
= \log\left(\frac{P(D|x_i=\alpha)}{P(D|x_i \neq \alpha)}\right).
\]

- The IPA is $L$ value of the best option + its sign.

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

- Create 1000 permuted datasets where GE+CN of a gene are taken from a random sample and a random gene in the pathway. (“within”)
- Create 1000 permuted datasets where GE+CN of a gene are taken from a random sample and a random gene in the genome. (“any”)
- Empirical p-values are measured against the distribution of the scores obtained on the permuted datasets
Problem: Most NCI pathways are cancer related. Need surely negative pathways. Created “decoy pathways” – same topology on random genes. Ran Paradigm and SPIA (Tarca et al 09) on the combination of real and decoy pathways. Used each method to rank all the pathways. Computed ROC curve.
From: Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM
Bioinformatics | © The Author(s) 2010. Published by Oxford University Press. This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.5), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

X axis: patient sample IPAs for the PI3K pathway.
Red: real data mean+std
Grey: permuted data ("within").
IPAs on the right include AKT1, CHUK and MDM2.
ErbB2 pathway in breast cancer

For each node, ER status, IPAs, expression data and copy-number data are displayed as concentric circles, from innermost to outermost, respectively. The apoptosis node and the ErbB2/ErbB3/neuregulin 2 complex node have circles only for ER status and for IPAs, as there are no direct observations of these entities. Each patient’s data is displayed along one angle from the circle center to edge.
Clustering of IPAs for GBM patients.

Rows: 1755 entities with IPA > 0.25 in ≥75 of 229 samples.

Cols: samples

Cluster 4 signif. different from rest (Cox PH p < 2.11 x 10^{-5})
TCGA Ovarian Cancer Inferred Pathway Activities

Patient Samples (247)

Pathway Concepts (867)

Ovarian: FOXM1 pathway altered in majority of serous ovarian tumors

Patient Samples (247)

FOXM1 Transcription Network

Pathway Concepts (867)

FOXM1 central to cross-talk between DNA repair and cell proliferation in Ovarian Cancer

PARADIGM-SHIFT predicts the function of mutations in multiple cancers using pathway impact analysis

Sam Ng, Eric A. Collisson, Artem Sokolov, Theodore Goldstein, Abel Gonzalez-Perez, Nuria Lopez-Bigas, Christopher Benz, David Haussler, Joshua M. Stuart

UC Santa Cruz, CA, USA

Bioinformatics 2012
Pathway signatures of mutations

- Mutated genes are the focus of many targeted approaches.
- Some patients with “right” mutation don’t respond. Why?
- Many cancers have several “novel” mutations. Can these be targeted with current approaches?

- Pathway-motivated approach: A mutated gene can alter its neighborhood
  - Gain-of-function (GOF): new activity of gene/pathway
  - Loss-of-function (LOF): deactivated gene/pathway/function
- A mutation in a gene can be manifest through the effect on its neighbors in the pathway
PARADIGM-Shift: Pathway context of GOF and LOF events

Use pathways to predict the impact of observed mutations in patient tumors

Predicted Loss-Of-Function

Predicted Gain-Of-Function

High Inferred Activity

Low Inferred Activity

FG: focus gene
PARADIGM-Shift Predicting the Impact of Mutations On Genetic Pathways

Inference using all neighbors

Inference using downstream neighbors

Inference using upstream neighbors

High Inferred Activity

Low Inferred Activity

mutated gene

SHIFT
PARADIGM–Shift Calculation Overview
PARADIGM–Shift Calculation Overview

1. Identify Local Neighborhood
1. Identify Local Neighborhood

2a. Regulators Run

2b. Targets Run
PARADIGM–Shift Calculation Overview

1. Identify Local Neighborhood

2a. Run Regulators

2b. Run Targets

3. Calculate Difference

P-Shift Score

(LOF)
PS score

- $f$ - the focus gene.
- Want $PS(f) = \log(\text{observed}(f)/\text{expected}(f))$
- Expected - activity based on upstream regulators
- Observed - activity based on downstream targets
- R-run: Paradigm run with regulators only, T-run: targets only
- $D(R), D(T), D(f)$: observed data for regulators, targets, f only
- $LR(Y|x^a, Z) = P(Y|x^a, Z)/P(Y|x^{-a}, Z)$ likelihood ratio for data given $x=a$ vs. the alternative $x \neq a$
- $\Phi(T)$ model limited to $f$ & its targets, $\Phi(R)$ f & its regulators

$$PS(f) = \log\left( \frac{LR(D(T)|x^a_f, \Phi(T))}{LR(D(R)|x^a_f, \Phi(R))} \right)$$

$$= \log\left( \frac{LR(D(T), x^a_f | \Phi(T))}{LR(D(T), x^{-a}_f | \Phi(T))} \right) - \log\left( \frac{LR(D(R), x^a_f | \Phi(R))}{LR(D(R), x^{-a}_f | \Phi(R))} \right) - \text{prior}$$
PS score computation

- Use IPLs from PARADIGM: $PS(f) = IPL_T(f) - IPL_R(f)$
- Transforming to Z-scores showed improvement in practice:
  - Create 100 random samples for gene $f$ by shuffling data
  - Compute $PS(f)$ scores, average $\mu$ and std $\sigma$.
  - Z-normalize the score $s$ on the real data to $(s - \mu)/\sigma$
- Running P-Shift: 2k Paradigm runs for $k$ mutated genes per patient (typically $k = 10-30$).
- Paradigm runs are faster on the reduced neighborhoods.
P-Shift Predicts RB1 Loss-of-Function in GBM

9 of 185 GBM samples have RB1 mutation
RB1 Discrepancy Scores distinguish mutated vs non-mutated samples

Distribution of M-sep scores on 1000 background models: same topology with permuted gene data tuples

→ Distinction is significant

Mutated
Non-mutated

Background
TP53 Network in GBM

48 of 185 GBM samples have TP53 mutations
LOF – tumor suppressor
Gain-of-Function (LUSC - lung squamous carcinoma)

- P-Shift Score
- PARADIGM downstream
- PARADIGM upstream
- Expression
- Mutation

NFE2L2
17 of 184 LUSC samples have NFE2L2 mutations
A known proto-oncogene
PARADIGM-Shift gives orthogonal view of the importance of cancer mutations

Genes ordered by their m-sep scores. Some low frequency mutations appear impactful. They are not detectable by tools that seek frequent mutations as potential “driver mutations” (e.g. Mutsig)
Recap

• **Paradigm**: integrated analysis of GE and CN data, employing knowledge on cancer pathways
• Modeling by factor graphs - versatile, useful tool!
• Produces IPA (integrated pathway activity) scores
• Clustering the IPA scores matrix identifies groups with significantly separable survival
• Was used in several large cancer projects since
• **Paradigm-shift**: use the above + mutations
• Uses multiple Paradigm runs per focus genes to compute m-sep scores, identify GOF and LOF events
• **CircleMap**: smart visualization of multiple data levels