Computational Biology

Karsten Borgwardt

Machine Learning and Computational Biology Research Group
Max Planck Institute for Intelligent Systems &
Max Planck Institute for Developmental Biology, Tübingen
Eberhard Karls Universität Tübingen

Machine Learning Summer School
September 4, 2013

The Need for Machine Learning in Computational Biology

BGI Hong Kong, Tai Po Industrial Estate, Hong Kong

High-throughput technologies:

- Genome and RNA sequencing
- Compound screening
- Genotyping chips
- Bioimaging

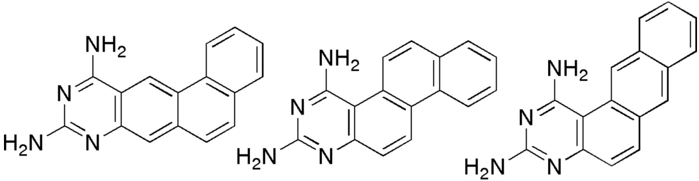
Molecular databases are growing much faster than our knowledge of biological processes.

Classic Bioinformatics: Focus on Molecules

- Large collections of molecular data
  - Gene and protein sequences
  - Genome sequence
  - Protein structures
  - Chemical compounds
- Focus: Inferring properties of molecules
  - Predict the function of a gene given its sequence
  - Predict the structure of a protein given its sequence
  - Predict the boundaries of a gene given a genome segment
  - Predict the function of a chemical compound given its molecular structure

Predicting Function from Structure

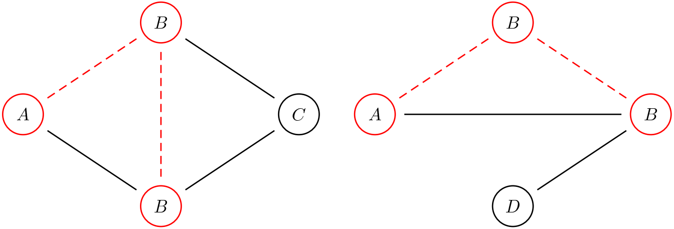
- Structure-Activity Relationship



25 **26** **27**

Source: Joska T M , and Anderson A C Antimicrob. Agents Chemother. 2006;50:3435-3443

Measuring the Similarity of Graphs

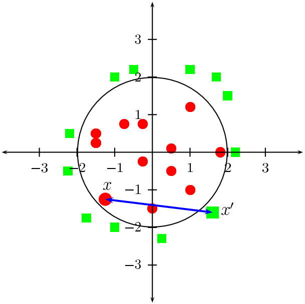
- How similar are two graphs?
  - How similar is their structure?
  - How similar are their node labels and edge labels?
- 

Graph Comparison

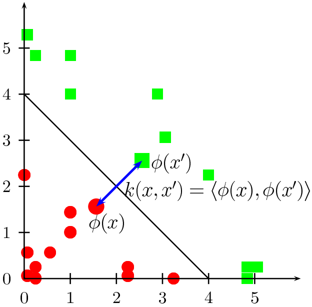
1. Graph isomorphism and subgraph isomorphism checking
   - Exact match
   - Exponential runtime
2. Graph edit distances
   - Involves definition of a cost function
   - Typically subgraph isomorphism as intermediate step
3. Topological descriptors
   - Lose some of the structural information represented by the graph **or**
   - Exponential runtime effort
4. Graph kernels (Gärtner et al, 2003; Kashima et al. 2003)
   - Goal 1: Polynomial runtime in the number of nodes
   - Goal 2: Applicable to large graphs
   - Goal 3: Applicable to graphs with attributes

Graph Kernels I

- Kernels
  - Key concept: Move problem to feature space $\mathcal{H}$.
  - Naive explicit approach:
    - Map objects $\mathbf{x}$ and $\mathbf{x}'$ via mapping ϕ to $\mathcal{H}$.
    - Measure their similarity in $\mathcal{H}$ as $\langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle$.
  - **Kernel Trick**: Compute inner product in $\mathcal{H}$ as kernel in input space $k(\mathbf{x}, \mathbf{x}') = \langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle$.

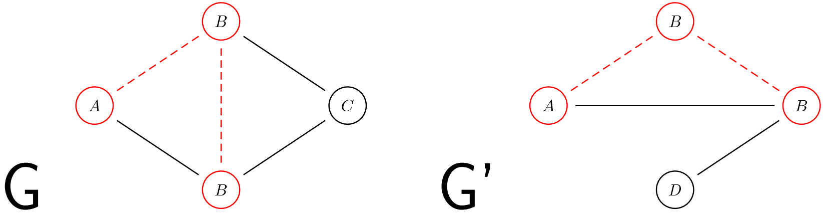


$\mathbb{R}^2 \quad \Rightarrow \quad \mathcal{H}$

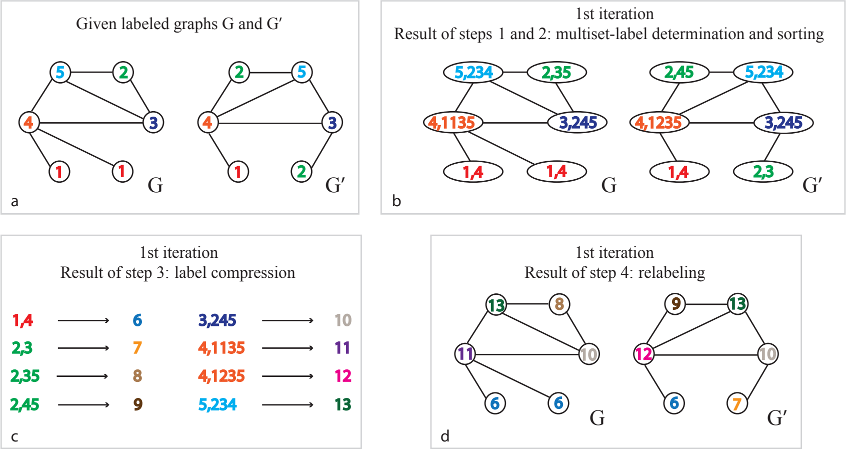


Graph Kernels II

- Graph kernels
  - Kernels on pairs of graphs
    (**not** pairs of nodes)
  - Instance of R-Convolution kernels (Haussler, 1999):
    - Decompose objects $\mathbf{x}$ and $\mathbf{x}'$ into substructures.
    - Pairwise comparison of substructures via kernels to compare $\mathbf{x}$ and $\mathbf{x}'$.
  - A graph kernel makes the whole family of kernel methods applicable to graphs.



Weisfeiler-Lehman Kernel (Shervashidze and Borgwardt, NIPS 2009)



Given labeled graphs G and G′

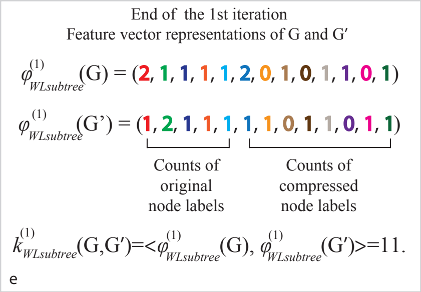
1st iteration
Result of steps 1 and 2: multiset-label determination and sorting

1st iteration
Result of step 3: label compression

1,4 ⟶ 6
2,3 ⟶ 7
2,35 ⟶ 8
2,45 ⟶ 9
3,245 ⟶ 10
4,1135 ⟶ 11
4,1235 ⟶ 12
5,234 ⟶ 13

1st iteration
Result of step 4: relabeling

Weisfeiler-Lehman Kernel (Shervashidze and Borgwardt, NIPS 2009)



End of the 1st iteration
Feature vector representations of G and G′

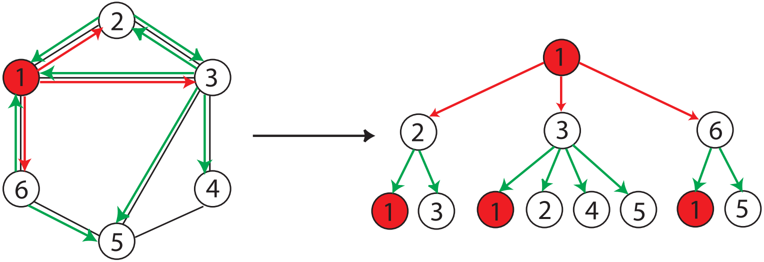
$$\varphi^{(1)}_{WLsubtree}(G) = (2, 1, 1, 1, 1, 2, 0, 1, 0, 1, 1, 0, 1)$$

$$\varphi^{(1)}_{WLsubtree}(G') = (\underbrace{1, 2, 1, 1, 1}_{\text{Counts of original node labels}}, \underbrace{1, 1, 0, 1, 1, 0, 1, 1}_{\text{Counts of compressed node labels}})$$

$$k^{(1)}_{WLsubtree}(G,G')=<\varphi^{(1)}_{WLsubtree}(G), \varphi^{(1)}_{WLsubtree}(G')>=11.$$

e

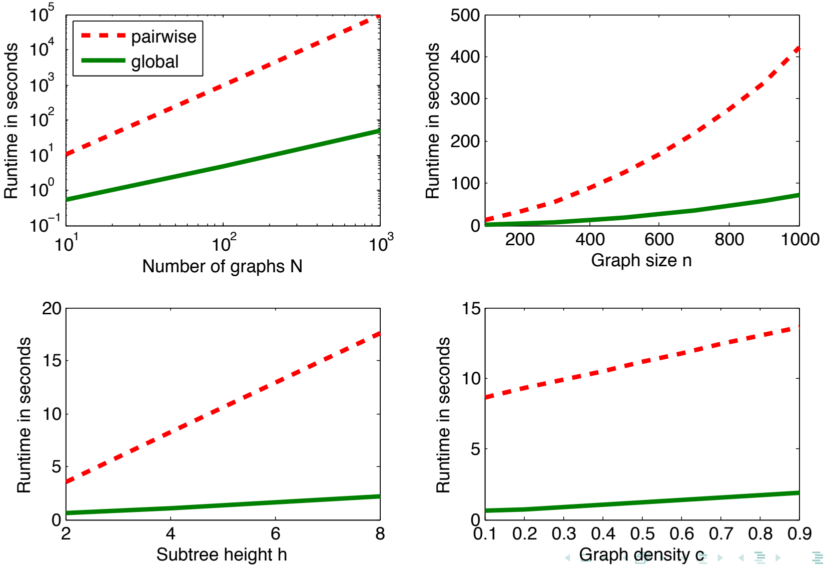
Subtree-like Patterns



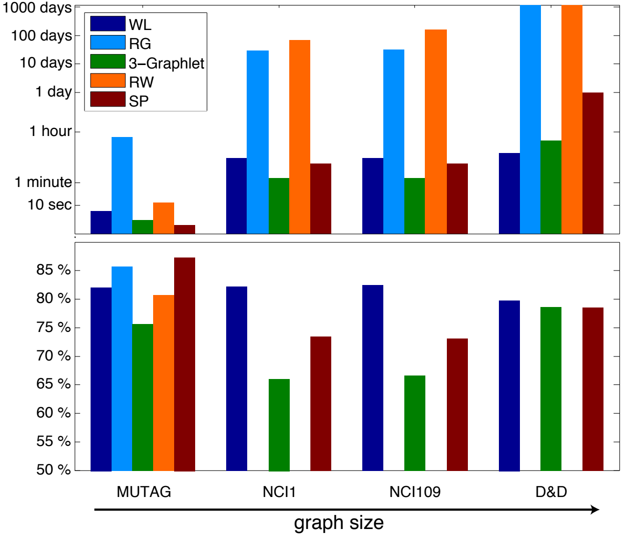
Weisfeiler-Lehman Kernel: Theoretical Runtime Properties

- Fast Weisfeiler-Lehman kernel (NIPS 2009 and JMLR 2011)
  - **Algorithm**: Repeat the following steps h times
    1. Sort: Represent each node v as sorted list L_v of its neighbors ($O(m)$)
    2. Compress: Compress this list into a hash value $h(L_v)$ ($O(m)$)
    3. Relabel: Relabel v by the hash value $h(L_v)$ ($O(n)$)
- Runtime analysis
  - per graph pair: Runtime $O(m\ h)$
  - for N graphs: Runtime $O(N\ m\ h + N^2\ n\ h)$ (naively $O(N^2\ m\ h)$)

Weisfeiler-Lehman Kernel: Empirical Runtime Properties

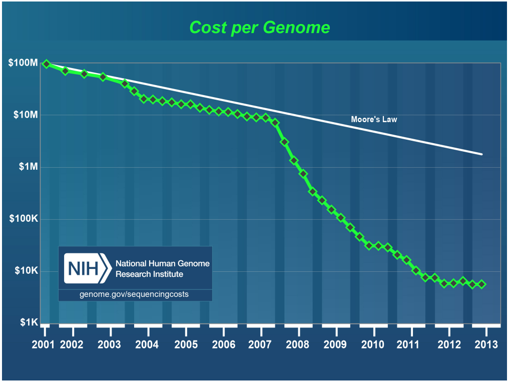


Weisfeiler-Lehman Kernel: Runtime and Accuracy



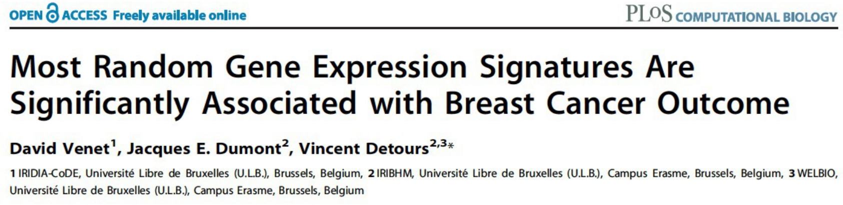
Modern Bioinformatics: Focus on Individuals

- High-throughput technologies now enable the collection of molecular information *on individuals*
  - Microarrays to measure gene expression levels
  - Chips to determine the genotype of an individual
  - Sequencing to determine the genome sequence of an individual



Phenotype Prediction

- Goal: Predict breast cancer outcome from gene expression levels
- Current results are not satisfying in terms of stability and prediction performance



OPEN ACCESS Freely available online

PLoS COMPUTATIONAL BIOLOGY

Most Random Gene Expression Signatures Are Significantly Associated with Breast Cancer Outcome

David Venet[1], Jacques E. Dumont[2], Vincent Detours[2,3]*

[1] IRIDIA-CoDE, Université Libre de Bruxelles (U.L.B.), Brussels, Belgium, [2] IRIBHM, Université Libre de Bruxelles (U.L.B.), Campus Erasme, Brussels, Belgium, [3] WELBIO, Université Libre de Bruxelles (U.L.B.), Campus Erasme, Brussels, Belgium

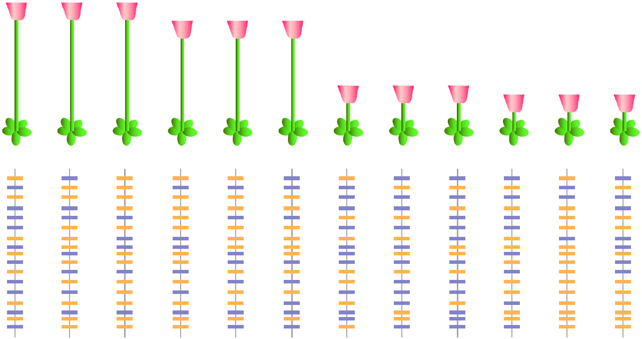
Phenotype Prediction

Nature News, March 2009

- 'Genetic test predicts eye color in Dutch men with 90% accuracy' (Liu et al., Current Biology 2009)
- Special setting: Candidate genes were already known beforehand
- Other phenotypes: Large genetics consortia try to detect candidate genes (e.g. diabetes, autism, depression, drug response, plant growth)

Genetics: Association Studies

- Genome-Wide Association Studies (GWAS)

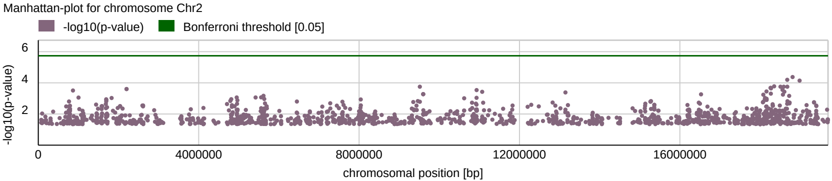


bco D. Weigel

- One considers genome positions that differ between individuals, that is *Single Nucleotide Polymorphisms (SNPs)* (more general: genetic locus or genomic variant).
- Problem size: 10^5-10^7 SNPs per genome, 10^2 to 10^5 individuals

Genetics: Manhattan Plots

- The standard statistical analysis in Genetics: Generating a Manhattan plot of association signals

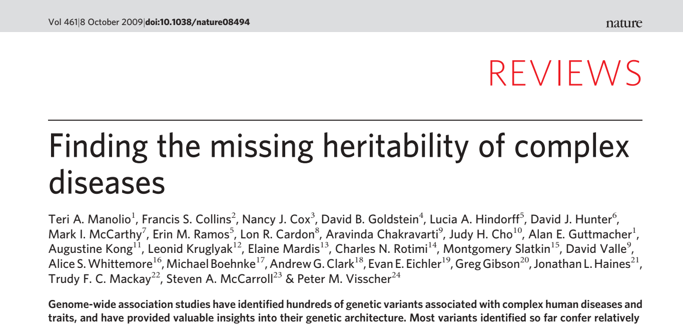


Phenotype: Flower color-related trait of *Arabidopsis thaliana*

- A plot of genome positions versus p-values of association/correlation.

Genetics: Missing Heritability

- More than 1200 new disease loci were detected over the last decade.
- The phenotypic variance explained by these loci is disappointingly low:



Vol 461|8 October 2009|**doi:10.1038/nature08494** nature

REVIEWS

Finding the missing heritability of complex diseases

Teri A. Manolio[1], Francis S. Collins[2], Nancy J. Cox[3], David B. Goldstein[4], Lucia A. Hindorff[5], David J. Hunter[6], Mark I. McCarthy[7], Erin M. Ramos[5], Lon R. Cardon[8], Aravinda Chakravarti[9], Judy H. Cho[10], Alan E. Guttmacher[1], Augustine Kong[11], Leonid Kruglyak[12], Elaine Mardis[13], Charles N. Rotimi[14], Montgomery Slatkin[15], David Valle[9], Alice S. Whittemore[16], Michael Boehnke[17], Andrew G. Clark[18], Evan E. Eichler[19], Greg Gibson[20], Jonathan L. Haines[21], Trudy F. C. Mackay[22], Steven A. McCarroll[23] & Peter M. Visscher[24]

Genome-wide association studies have identified hundreds of genetic variants associated with complex human diseases and traits, and have provided valuable insights into their genetic architecture. Most variants identified so far confer relatively

Manolio et al., Nature 2009

Genetics: Potential Reasons for Missing Heritability

Polygenic architectures

- Most current analyses neglect additive or multiplicative effects between loci → need for **systems biology perspective**

Small effect sizes

- Not detectable with small sample sizes

Phenotypic effect of other genetic, epigenetic or non-genetic factors

- Genetic properties ignored so far, e.g. rare SNPs
- Chemical modifications of the genome
- Environmental effect on phenotype

Moving to a Systems Biology Perspective

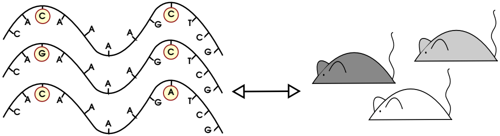
- Multi-locus models:
  - Algorithms to discover trait-related systems of genetic loci
- Increasing sample size:
  - Algorithms that support large-scale genotyping and phenotyping
- Deciding whether additional information is required:
  - Tests that quantify the impact of additional (epi)genetic factors

Machine Learning in Genetics II

Moving to a Systems Biology Perspective

- Multi-locus models:
  - Efficient algorithms for discovering trait-related SNP pairs (KDD 2011, Human Heredity 2012)
  - Efficient algorithms for discovering trait-related SNP networks (ISMB 2013)
- Increasing sample size:
  - Large-scale genotyping in *A. thaliana* (Nature Genetics 2011)
  - Automated image phenotyping of guppy fish (Bioinformatics 2012)
  - Automated image phenotyping of human lungs (IPMI 2013)
- Deciding whether additional information is required:
  - Assessing the stability of methylation across generations of *Arabidopsis* lab strains (Nature 2011)

Multi-Locus Models: Discovering Trait-Related Interactions



Problem statement

- Find the pair of SNPs most correlated with a binary phenotype

$$\underset{i,j}{\operatorname{argmax}} \, |r(\mathbf{x}_i \odot \mathbf{x}_j, \mathbf{y})|$$

- $\mathbf{x}_i$ and $\mathbf{x}_j$ represent one SNP each and $\mathbf{y}$ is the phenotype; $\mathbf{x}_i, \mathbf{x}_j, \mathbf{y}$ are all n-dimensional vectors, given n individuals.
- There can be up to $n = 10^7$ SNPs, and order 10^{14} SNP pairs.
- Existing approaches: Greedy selection, Branch-and-bound strategies or index structures → low recall or worst-case $O(n^2)$ time

Difference in Correlation for Epistasis Detection

- We phrase epistasis detection as a difference in correlation problem:

$$\operatorname*{argmax}_{i,j} |\rho_{cases}(\mathbf{x}_i, \mathbf{x}_j) - \rho_{controls}(\mathbf{x}_i, \mathbf{x}_j)|. \tag{1}$$

- Different degree of linkage disequilibrium of two loci in cases and controls

The Lightbulb Algorithm (Paturi et al., COLT 1989)

Maximum correlation

- The lightbulb algorithm tackles the maximum correlation problem on an $m \times n$ matrix A with binary entries:

$$\operatorname*{argmax}_{i,j} |\rho_{\mathbf{A}}(\mathbf{x}_i, \mathbf{x}_j)|. \tag{2}$$

Quadratic runtime algorithm

- As in epistasis detection, the problem can be solved by naive enumeration of all n^2 possible solutions.

The Lightbulb Approach

Lightbulb algorithm

1. Given a binary matrix $\mathbf{A}$ with m rows and n columns.
2. Repeat l times:
    - Sample k rows
    - Increase a counter for all pairs of columns that match on these k rows.
3. The counters divided by l give an estimate of the correlation $P(\mathbf{x}_i = \mathbf{x}_j)$.

Subquadratic runtime

- With probability near 1, the lightbulb algorithm retrieves the most correlated pair in $O(n^{1+\frac{\ln c_1}{\ln c_2}} \ln^2 n)$, where c_1 and c_2 are the highest and second highest correlation score.

Difference Between the Epistasis and Lightbulb Problem Setting

Discrepancies

- Difference in correlation
- SNPs are non-binary in general
- Pearson's correlation coefficient

Step 1: Difference in Correlation

Theorem

- Given a matrix of cases $\mathbf{A}$ and a matrix of controls $\mathbf{B}$ of identical size.
- Finding the maximally correlated pair on

$$\begin{pmatrix} \mathbf{A} & \mathbf{A} \\ \mathbf{B} & \mathbf{1} - \mathbf{B} \end{pmatrix} \tag{3}$$

- and on

$$\begin{pmatrix} \mathbf{A} & \mathbf{1} - \mathbf{A} \\ \mathbf{B} & \mathbf{B} \end{pmatrix} \tag{4}$$

- is identical to

$$\underset{i,j}{\operatorname{argmax}} |\rho_{\mathbf{A}}(\mathbf{x}_i, \mathbf{x}_j) - \rho_{\mathbf{B}}(\mathbf{x}_i, \mathbf{x}_j)|. \tag{5}$$

Step 2: Locality Sensitive Hashing (Charikar, 2002)

Given a collection of vectors in $\mathbb{R}^m$ we choose a random vector $\mathbf{r}$ from the m-dimensional Gaussian distribution. Corresponding to this vector $\mathbf{r}$, we define a hash function $h_{\mathbf{r}}$ as follows:

$$h_{\mathbf{r}}(\mathbf{x}_i) = \begin{cases} 1 & \text{if } \mathbf{r}^\top \mathbf{x}_i \geq 0 \\ 0 & \text{if } \mathbf{r}^\top \mathbf{x}_i < 0 \end{cases} \tag{6}$$

Theorem

For vectors $\mathbf{x}_i, \mathbf{x}_j$, $Pr[h_{\mathbf{r}}(\mathbf{x}_i) = h_{\mathbf{r}}(\mathbf{x}_j)] = 1 - \frac{\theta(\mathbf{x}_i, \mathbf{x}_j)}{\pi}$, where θ is the angle between the two vectors.

Step 3: Pearson's Correlation Coefficient

Link between correlation and cosine

Karl Pearson defined the correlation of 2 vectors $\mathbf{x}_i, \mathbf{x}_j$ in $\mathbb{R}^m$ as

$$\rho = \frac{cov(\mathbf{x}_i, \mathbf{x}_j)}{\sigma_{\mathbf{x}_i}\sigma_{\mathbf{x}_j}}, \tag{7}$$

that is the covariance of the two vectors divided by their standard deviations. An equivalent geometric way to define it is:

$$\rho = cos(\mathbf{x}_i - \bar{\mathbf{x}}_i, \mathbf{x}_j - \bar{\mathbf{x}}_j), \tag{8}$$

where $\bar{\mathbf{x}}_i$ and $\bar{\mathbf{x}}_j$ are the mean value of $\mathbf{x}_i$ and $\mathbf{x}_j$, respectively.

The Lightbulb Epistasis Algorithm (Achlioptas et al., KDD 2011)

Algorithm

1. Binarize original matrices $\mathbf{A}_0$ and $\mathbf{B}_0$ into $\mathbf{A}$ and $\mathbf{B}$ by locality sensitive hashing.
2. Compute maximally correlated pair $\mathbf{p}_1$ on $\begin{pmatrix} \mathbf{A} & \mathbf{A} \\ \mathbf{B} & \mathbf{1}-\mathbf{B} \end{pmatrix}$ via lightbulb.
3. Compute maximally correlated pair $\mathbf{p}_2$ on $\begin{pmatrix} \mathbf{A} & \mathbf{1}-\mathbf{A} \\ \mathbf{B} & \mathbf{B} \end{pmatrix}$ via lightbulb.
4. Report the maximum of $\mathbf{p}_1$ and $\mathbf{p}_2$.

Experiments: *Arabidopsis* SNP dataset

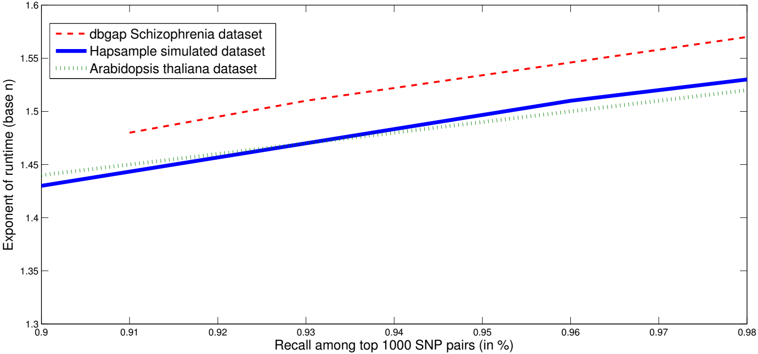
Results on *Arabidopsis* SNP dataset

# SNPs	Measurements	Pairs	Exponent	Speedup	Top 10	Top 100	Top 500	Top 1K
100,000	8,255,645	8,186,657	1.38	611	1.00	0.86	0.82	0.80
100,000	52,762,001	51,732,700	1.54	97	1.00	1.00	0.99	0.98

Runtime

- Runtime is empirically $O(n^{1.5})$.
- Epistasis detection on the human genome would require 1 day of computation on a typical desktop PC.

Experiments: Runtime versus Recall



Multi-Locus Models: Discovering Trait-Related Interactions

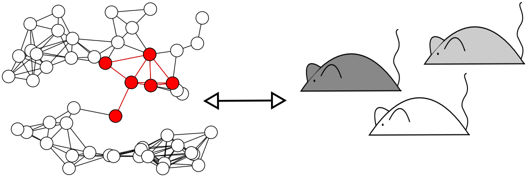
Alternative: Engineering approach

- Use parallel computing power of Graphical Processing Units for interaction discovery (Kam-Thong et al., ISMB 2011 & Human Heredity 2012)
- Similar speed-up as with Lightbulb algorithm

Road ahead

- We got the approval to perform the official SNP-SNP interaction discovery analysis for:
  - The international lung disease genetics consortium COPDGene
  - The international headache genetics consortium (Clinical Migraine)
- Our methods will be used in further consortia:
  - Psychiatric diseases such as autism, schizophrenia, depression

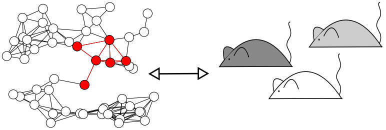
Multi-Locus Models: Discovering Trait-Related Networks



Network information

- What about models with more than 2 SNPs?
- Additive models are hard to interpret, multiplicative models are hard to compute.
- Can the growing knowledge about gene and protein networks be exploited to improve multi-locus mapping?

Multi-Locus Models: Discovering Trait-Related Networks

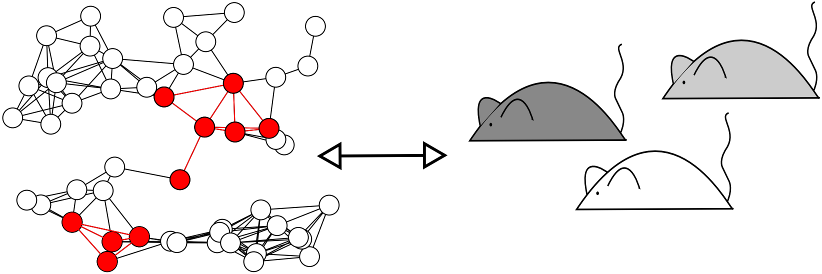


- Edges between SNPs near the same gene or SNPs in interacting genes
- c_i is the association score of SNP i, $f_i = 1$ if SNP i is selected, $f_i = 0$ if not.
- Find a set of SNPs with maximum total score:

$$\underset{\boldsymbol{f} \in \{0,1\}^n}{\operatorname{argmax}} \; \boldsymbol{c}^\top \boldsymbol{f}$$

  such that
  - the selected SNPs form a connected subgraph and
  - $\boldsymbol{f}$ is sparse.
- NP-complete problem: Maximum Weight Connected Subgraph Problem (Lee and Dooly, 1993)

Multi-Locus Models: Discovering Trait-Related Networks



Multi-Locus Models: Discovering Trait-Related Networks

Our formulation (Azencott et al., ISMB 2013)

- Networks are incomplete → Connectedness needs not be strictly enforced, but merely rewarded by a Graph Laplacian regularizer $\boldsymbol{f}^\top \boldsymbol{L} \boldsymbol{f} = \sum_{i \sim j} (f_i - f_j)^2$, where $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{W}$.
- The SNP subnetwork selection problem is then:

$$\underset{\boldsymbol{f} \in \{0,1\}^n}{\operatorname{argmax}} \quad \underbrace{\boldsymbol{c}^\top \boldsymbol{f}}_{\text{association}} \; - \underbrace{\lambda \; \boldsymbol{f}^\top \boldsymbol{L} \boldsymbol{f}}_{\text{connectivity}} - \underbrace{\eta \;\; ||\boldsymbol{f}||_0}_{\text{sparsity}}$$

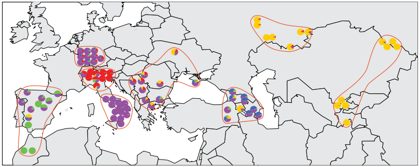
- This is a min-cut problem, for which efficient algorithms exist (we use Boykov and Kolmogorov, IEEE TPAMI 2004).
- Much faster and recovers four times more phenotype-related genes in *A. thaliana* than network-constrained Lasso models

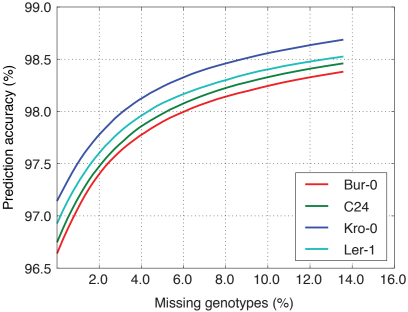
Multi-Locus Models: Current Work

Other important aspects

- Including prior knowledge on relevance of SNPs (Limin Li et al., ISMB 2011)
- Accounting for relatedness of individuals (Rakitsch et al., Bioinformatics 2012)
- Measuring statistical significance
- Predicting multiple correlated phenotypes jointly

Increasing Sample Size: Genotyping (Cao et al., Nat. Gen. 2011)





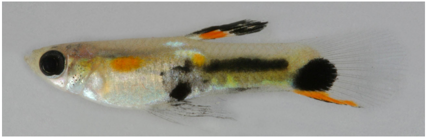
Setup

- 80 fully sequences genomes from *A. thaliana* (3 million SNPs)
- 4 strains with 250.000 SNPs
- Can we predict the remaining SNPs?

Result

- Employed BEAGLE to predict missing SNPs in 4 strains
- Missing sites can be accurately predicted (>96% accuracy)

Increasing Sample Size: Phenotyping (Karaletsos et al., Bioinf. 2012)



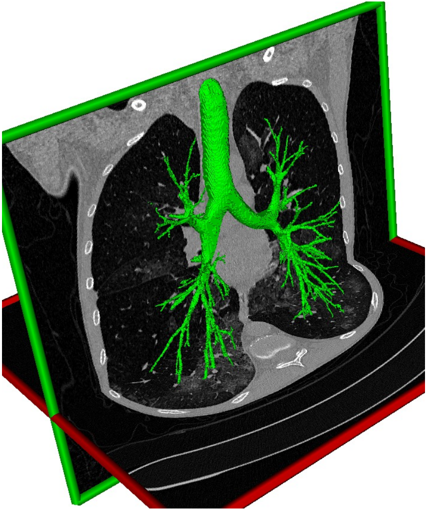
Setup

- Guppy image collections
- Re-occurring color patterns are phenotypes
- How to phenotype the guppies automatically?

Result

- Proposed Markov Random Field for pattern discovery
- Recovers color patterns found by manual annotation

Increasing Sample Size: Phenotyping (Feragen et al., IPMI 2013)



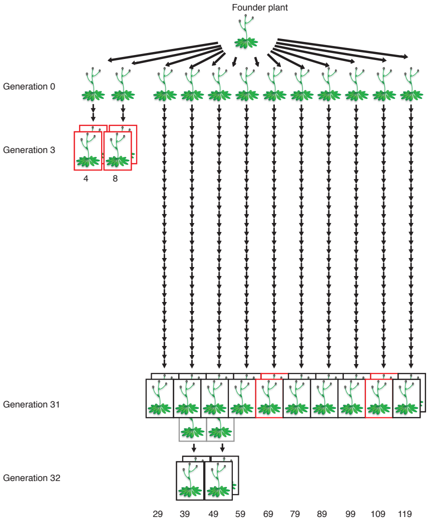
Setup

- Collections of CT-scans of human lungs
- Structural differences may be linked to disease (COPD)
- How to measure differences in lung structure?

Result

- Proposed novel, efficient similarity measure on geometric trees (tree kernel)

Additional Factors: Epigenetic Influences (Becker et al., Nature 2011)



Setup

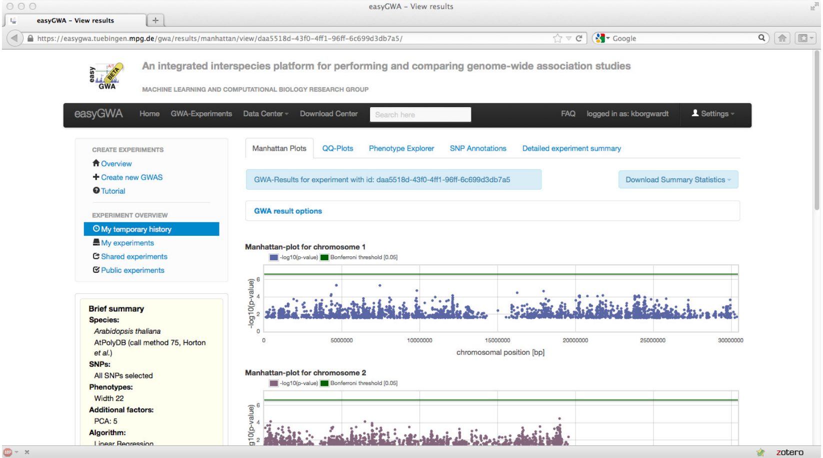
- 33 generations of lab strains of *A. thaliana*
- How stable is the methylation state of genome positions across generations?

Result

- Position-specific methylation varies greatly
- Region-wide methylation is more stable

An Online Resource for Machine Learning on Complex Traits

- We published easyGWAS (https://easygwas.tuebingen.mpg.de/), a machine learning platform for analysing complex traits (Grimm et al., arXiv 2012):



easyGWA – View results

https://easygwa.tuebingen.mpg.de/gwa/results/manhattan/view/daa5518d-43f0-4ff1-96ff-6c699d3db7a5/

An integrated interspecies platform for performing and comparing genome-wide association studies

MACHINE LEARNING AND COMPUTATIONAL BIOLOGY RESEARCH GROUP

logged in as: kborgwardt

GWA-Results for experiment with id: daa5518d-43f0-4ff1-96ff-6c699d3db7a5

Manhattan-plot for chromosome 1

Manhattan-plot for chromosome 2

Brief summary
Species:
Arabidopsis thaliana
AtPolyDB (call method 75, Horton *et al.*)
SNPs:
All SNPs selected
Phenotypes:
Width 22
Additional factors:
PCA: 5
Algorithm:
Linear Regression

Summary and Outlook

How can Machine Learning contribute to Personalized Medicine?

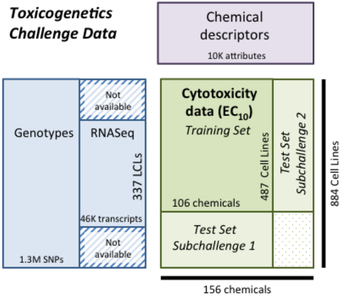
- By discovering relationships between groups of molecular components and functions of a system
- By allowing to efficiently collect and annotate large sample sizes of observations
- By measuring the 'added value' of further molecular factors

Outlook: Phenotype Prediction

- Scaling tests, models, algorithms to large, high-dimensional datasets, e.g. from Imaging, Epigenomics, Transcriptomics
- Learning across different data sources
- Analysing structured phenotypes (images, time series)
- More challenges for and applications of machine learning

The Road Ahead: Personalized Medicine

- Example: DREAM 8 NIEHS-NCATS-UNC DREAM Toxicogenetics Challenge
- Goal: Predict a reaction of a genotyped cell line to a chemical compound
- Joins molecule- and individual-centered bioinformatics



Toxicogenetics Challenge Data

Chemical descriptors

10K attributes

Genotypes

1.3M SNPs

Not available

RNASeq

46K transcripts

Not available

337 LCLs

Cytotoxicity data (EC_{10})

Training Set

106 chemicals

487 Cell Lines

Test Set Subchallenge 2

Test Set Subchallenge 1

884 Cell Lines

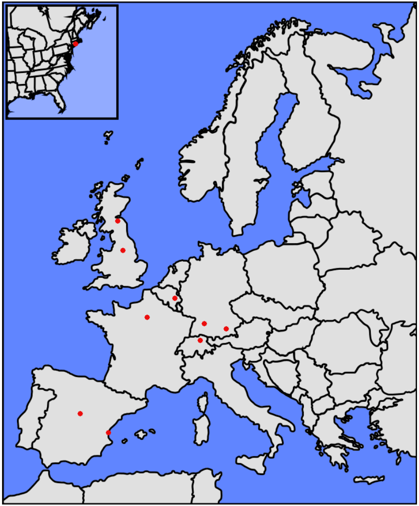
156 chemicals

Source: https://www.synapse.org

The Road Ahead: Marie Curie Initial Training Network

- Goal: Enable medical treatment tailored to patients' molecular properties
- Plan: Help to build a research community at the interface of Machine Learning and data-driven Medicine
- First step: Marie Curie Initial Training Network (ITN)
  - Topic: Machine Learning for Personalized Medicine (MLPM)
  - Duration: 4 years, started January 2013
  - 14 early-stage researchers in 12 labs at 10 nodes in 6 countries
  - 3.75 million EUR funding for PhD students and training events
  - Research programmes:
    - Biomarker Discovery
    - Data Integration
    - Causal Mechanisms of Disease
    - Gene-Environment Interactions
- Follow us on mlpm.eu

ITN on "Machine Learning for Personalized Medicine" (MLPM)



- Pharmatics, Edinburgh
- University of Sheffield
- University of Liège
- INSERM and ARMINES, Paris
- MPIs Tübingen
- MPI for Psychiatry, Munich
- Siemens Munich
- Universidad Carlos III de Madrid
- Prince Felipe Research Centre (CIPF) in Valencia
- MSKCC New York

Thank You

Postdocs and PhD students:

- Aasa Feragen
- Barbara Rakitsch
- Carl-Johann Simon-Gabriel
- Chloé-Agathe Azencott
- Damian Roqueiro
- Dominik Grimm
- Felipe Llinares Lopez
- Mahito Sugiyama
- Niklas Kasenburg

Sponsors:

- Krupp-Stiftung
- A.-v.-Humboldt-Stiftung
- DFG
- Det Frie Forskningsrad Denmark
- Marie-Curie-FP 7

Thank You

https://www.facebook.com/MLCBResearch

Main References

- C.-A. Azencott, D. Grimm, M. Sugiyama, Y. Kawahara, K. M. Borgwardt, *ISMB* (2013).
- P. Achlioptas, B. Schölkopf, K. Borgwardt, *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)* (2011), pp. 726–734.
- J. Cao, *et al.*, *Nature Genetics* **43**, 956 (2011). PMID: 21874002.
- T. Karaletsos, O. Stegle, C. Dreyer, J. Winn, K. M. Borgwardt, *Bioinformatics* **28**, 1001 (2012).
- C. Becker, *et al.*, *Nature* **480**, 245 (2011).
- D. Grimm, *et al.*, *arXiv:1212.4788* (2012).
- N. Shervashidze, K. M. Borgwardt, *Neural Information Processing Systems (NIPS)* pp. 1660–1668 (2009). **NIPS Outstanding Student Paper Award Winner.**