

Data Mining in the Life Sciences

The Path to Personalized Medicine

Karsten Borgwardt

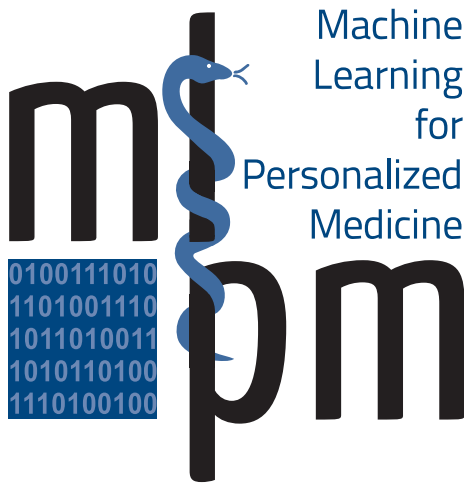
Machine Learning and Computational Biology Research Group
Max Planck Institute for Intelligent Systems &
Max Planck Institute for Developmental Biology, Tübingen
Eberhard Karls Universität Tübingen



Machine Learning for Personalized Medicine
ITN Summer School
September 23-27, 2013



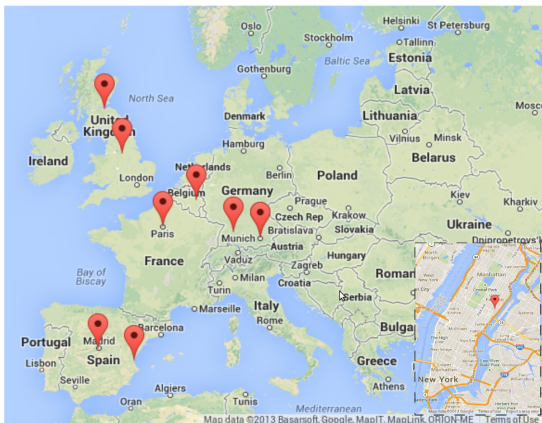
MAX-PLANCK-GESELLSCHAFT



Our Initial Training Network: In Numbers

- ▶ 6 countries: Belgium, France, Germany, Spain, United Kingdom, United States
- ▶ 13 ESRs + 1 ER
- ▶ 12 labs at 10 nodes, 8 academic partners and 2 industrial nodes (Siemens, Pharmatics)
- ▶ Duration: 4 years, January 2013 — December 2016
- ▶ Funding: Up to 3.75 million EUR
- ▶ 3 years of funding per student
- ▶ 2 three-month secondments per student
- ▶ 4 annual summer schools (Tübingen, UK, France, Spain)

Our Initial Training Network: As a Map



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

Our Initial Training Network: Who is Who? The PIs

- ▶ Belgium: University of Liège (Prof. Kristel Van Steen)
- ▶ France: ARMINES (Prof. Jean-Philippe Vert), INSERM (Prof. Florence Demenais)
- ▶ Spain: UC3 Madrid (Prof. Fernando Perez-Cruz), CIPF Valencia (Prof. Joaquin Dopazo)
- ▶ United Kingdom: University of Sheffield (Prof. Neil Lawrence, Prof. Magnus Rattray - *now Manchester*), Pharmatics (Dr. Felix Agakov)
- ▶ United States: MSKCC (Prof. Gunnar Rätsch)
- ▶ Germany: Siemens (Prof. Volker Tresp), Max-Planck-Society (Prof. Bertram Müller-Myhsok, Prof. Bernhard Schölkopf and Prof. Karsten Borgwardt)

Our Initial Training Network: Who is Who? The ESRs

MPG (Borgwardt)	Mr Felipe Llinares-López
MPG (Borgwardt)	Mr Carl-Johann Simon-Gabriel
MPG (Schölkopf)	Mr James McMurray
MPG (Müller-Myhsok)	Ms Meiwen Jia
Siemens	Mr Cristóbal Esteban
U Sheffield	Mr Max Zwiebele
U Liège	Ms Ramouna Fouladi
ARMINES Paris	Mr Yunlong Jiao
INSERM Paris	Mr Yuanlong Liu
UC3 Madrid	Ms Mélanie Fernández Pradier
CIPF Valencia	Mr Cancut Cubuk
MSKCC	Mr Yi Zhong

Our Initial Training Network: Our Topics

- ▶ Research Goal A.1: Biomarker discovery
- ▶ Research Goal A.2: Data Integration
- ▶ Research Goal B.1: Causal Mechanisms of Disease
- ▶ Research Goal B.2: Gene- Environment Interactions

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

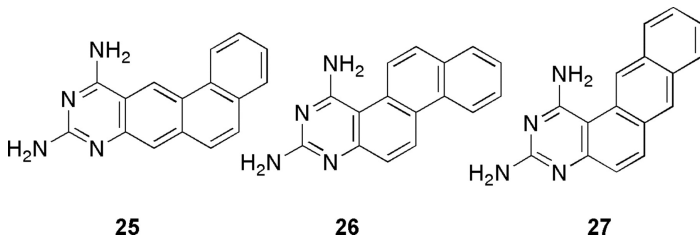
Molecular databases are growing much faster than our knowledge of biological processes.

- ▶ 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

Example: Predicting Function from Structure

► Structure-Activity Relationship

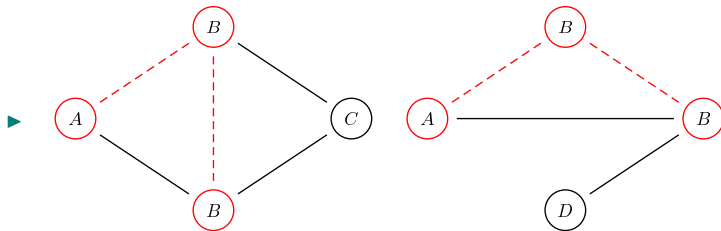


Source: Joska T M , and Anderson A C Antimicrob. Agents Chemother. 2006;50:3435-3443

► Fundamental idea: Similarity in structure implies similarity in function

Measuring the Similarity of Graphs

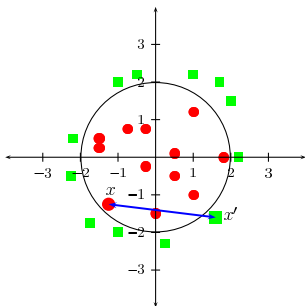
- ▶ How similar are two graphs?
 - ▶ How similar is their structure?
 - ▶ How similar are their node labels and edge labels?



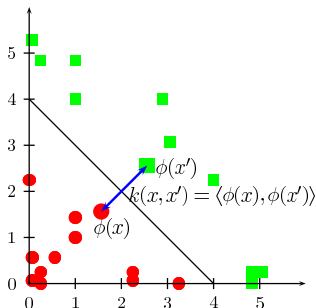
1. Graph isomorphism and subgraph isomorphism checking
 - ▶ Exact match
 - ▶ Exponential runtime
2. Graph edit distances
 - ▶ Involves definition of a cost function
 - ▶ Typically subgraph isomorphism as intermediate step
3. Topological descriptors
 - ▶ Lose some of the structural information represented 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

► 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 $k(\mathbf{x}, \mathbf{x}') = \langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle$.

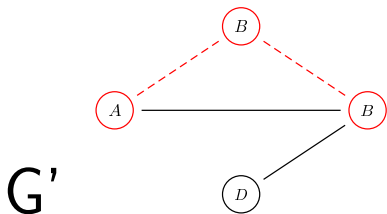
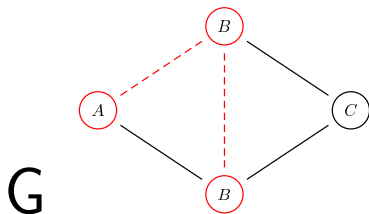


$\mathbb{R}^2 \Rightarrow \mathcal{H}$

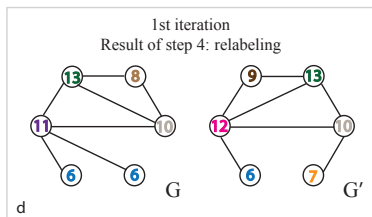
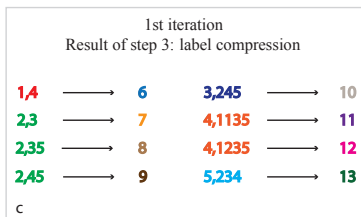
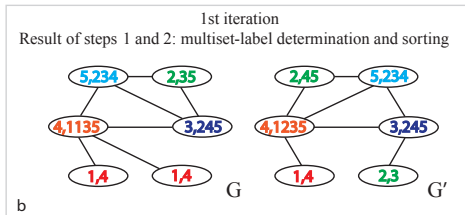
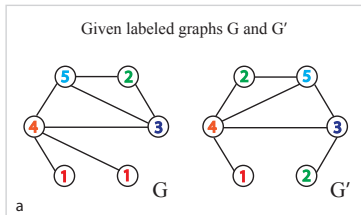


▶ Graph kernels

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



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



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

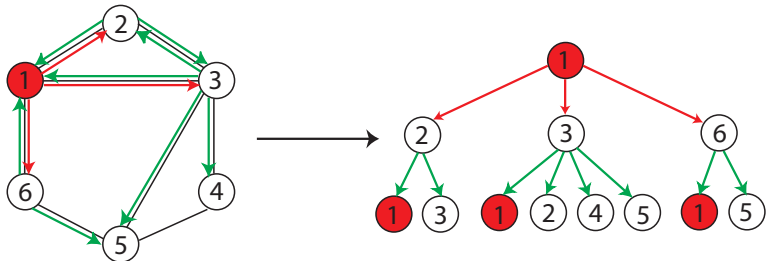
$$\phi_{WLsubtree}^{(1)}(G) = (2, 1, 1, 1, 1, 2, 0, 1, 0, 1, 1, 0, 1)$$

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

$$k_{WLsubtree}^{(1)}(G, G') = \langle \phi_{WLsubtree}^{(1)}(G), \phi_{WLsubtree}^{(1)}(G') \rangle = 11.$$

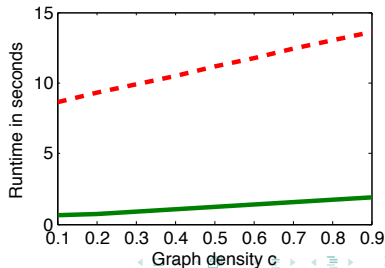
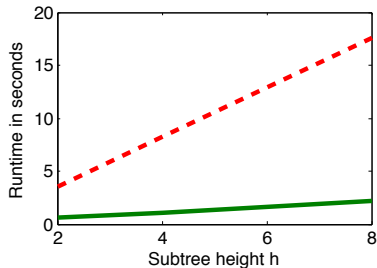
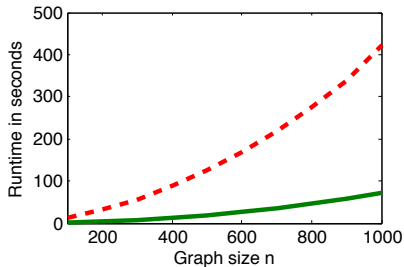
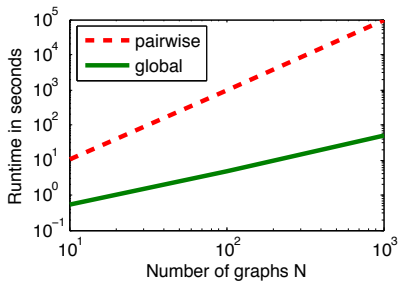
e

Subtree-like Patterns

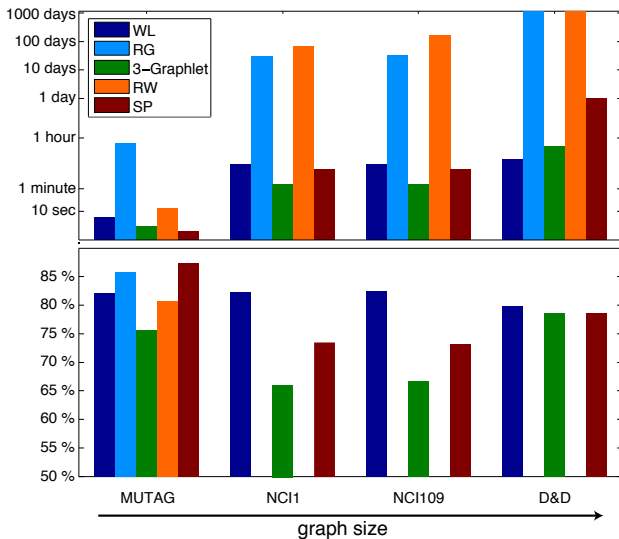


- ▶ **Fast Weisfeiler-Lehman kernel (NIPS 2009 and JMLR 2011)**
 - ▶ **Algorithm:** Repeat the following steps h times
 1. **Sort:** Represent each node v as sorted list L_v of its neighbors ($O(m)$)
 2. **Compress:** Compress this list into a **hash value** $h(L_v)$ ($O(m)$)
 3. **Relabel:** Relabel 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



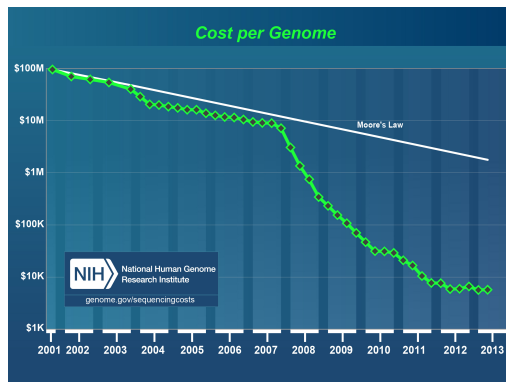
Weisfeiler-Lehman Kernel: Runtime and Accuracy



- ▶ Modern Bioinformatics: Focus on Individuals

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



- ▶ 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 Venet¹, Jacques E. Dumont², Vincent Detours^{2,3*}

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

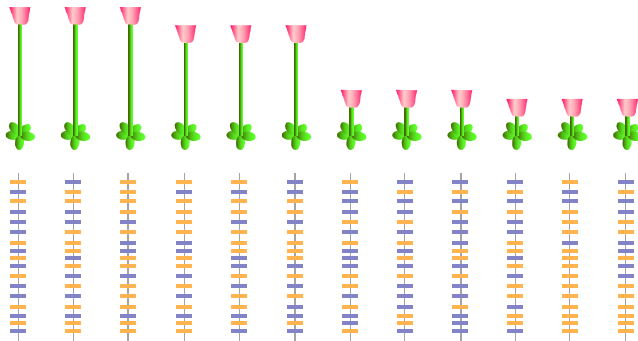
Source: Venet et al., PLoS Comp Bio 2011

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)



► Genome-Wide Association Studies (GWAS)

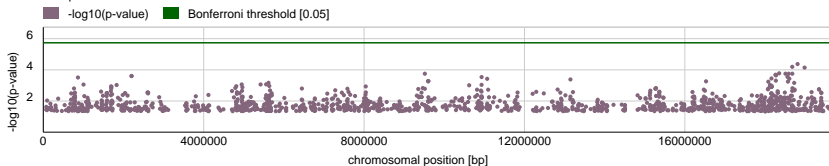


bco D. Weigel

- 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

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

Manhattan-plot for chromosome Chr2



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

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

- ▶ 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. Manolio¹, Francis S. Collins², Nancy J. Cox³, David B. Goldstein⁴, Lucia A. Hindorf⁵, David J. Hunter⁶, Mark I. McCarthy⁷, Erin M. Ramos⁵, Lon R. Cardon⁸, Aravinda Chakravarti⁹, Judy H. Cho¹⁰, Alan E. Guttmacher¹, Augustine Kong¹¹, Leonid Kruglyak¹², Elaine Mardis¹³, Charles N. Rotimi¹⁴, Montgomery Slatkin¹⁵, David Valle⁹, Alice S. Whittemore¹⁶, Michael Boehnke¹⁷, Andrew G. Clark¹⁸, Evan E. Eichler¹⁹, Greg Gibson²⁰, Jonathan L. Haines²¹, Trudy F. C. Mackay²², Steven A. McCarroll²³ & Peter M. Visscher²⁴

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

Missing genetic component

- ▶ Heritability in common traits
 - ▶ few > 50% (e.g. type I diabetes, fetal haemoglobin levels)
 - ▶ some 20-30% (e.g. Crohn's disease, lipid levels)
 - ▶ most < 20% (e.g. autism, height, schizophrenia)
- ▶ Explained heritability is phenotypic variance explained by known variants over variance explained by all (even unknown) variants.

Wrong models?

- ▶ Lander (2011) and Zuk et al. (2012) speculate that heritability estimates could be inflated: 'Phantom heritability'
- ▶ Current estimates ignore gene-gene interactions and gene-environment interactions.

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**

Moving to a Systems Biology Perspective

- ▶ Multi-locus models:
 - ▶ Efficient algorithms for discovering trait-related SNP pairs (KDD 2011, Human Heredity 2012)
- ▶ 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)

Examples of Epistasis

- ▶ Epistasis is conjectured to be one source of missing heritability (Manolio et al., 2009)
- ▶ Genetic interactions are one indicator that epistasis is a major factor in the genotype-phenotype relationship (e.g. Boone et al., 2007)
- ▶ Pairs of genes have been reported to affect complex diseases such as breast cancer (Ashworth et al., 2011):
 - ▶ Loss of either BRCA1 or BRCA2 tumor suppressor gene function in cells triggers a cell-cycle arrest at the G2/M checkpoint that can be suppressed by the inactivation of P53 (Connor et al., 1997 and Liu et al., 2007).
 - ▶ Loss of VHL (Von Hippel-Lindau tumor suppressor) function normally causes cellular senescence, but inactivation of a second tumor suppressor, RB (Retinoblastoma), can suppress this process (Young et al., 2008).

Scale of the problem

- ▶ Typical datasets include order $10^5 - 10^7$ SNPs.
- ▶ Hence we have to consider order $10^{10} - 10^{14}$ SNP pairs.
- ▶ Enormous multiple hypothesis testing problem.
- ▶ Enormous computational runtime problem.

Exhaustive enumeration

- ▶ Only with special hardware such as Cloud Computing or GPU implementations (e.g. Kam-Thong et al., EJHG 2010, ISMB 2011, Hum Her 2012)

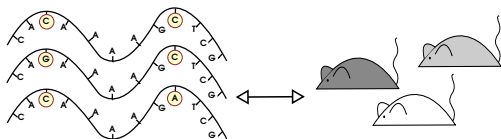
Filtering approaches

- ▶ Statistical criterion, e.g. SNPs with large main effect (Zhang et al., 2007)
- ▶ Biological criterion, e.g. underlying PPI (Emily et al., 2009)

Index structure approaches

- ▶ fastANOVA, branch-and-bound on SNPs (Zhang et al., 2008)
- ▶ TEAM, efficient updates of contingency tables (Zhang et al., 2010)

Multi-Locus Models: Discovering Trait-Related Interactions



Problem statement

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

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

- ▶ \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 m -dimensional vectors, given m 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 \rightarrow low recall or worst-case $O(n^2)$ time

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

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

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

Maximum correlation

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

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

Quadratic runtime algorithm

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

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. Rank all pairs according to their counter value

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.

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

$$\begin{pmatrix} \mathbf{A} & \mathbf{A} \\ \mathbf{B} & \mathbf{1} - \mathbf{B} \end{pmatrix} \quad (3)$$

- ▶ and on

$$\begin{pmatrix} \mathbf{A} & \mathbf{1} - \mathbf{A} \\ \mathbf{B} & \mathbf{B} \end{pmatrix} \quad (4)$$

- ▶ is identical to

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

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:

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

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

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

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

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

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

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 .

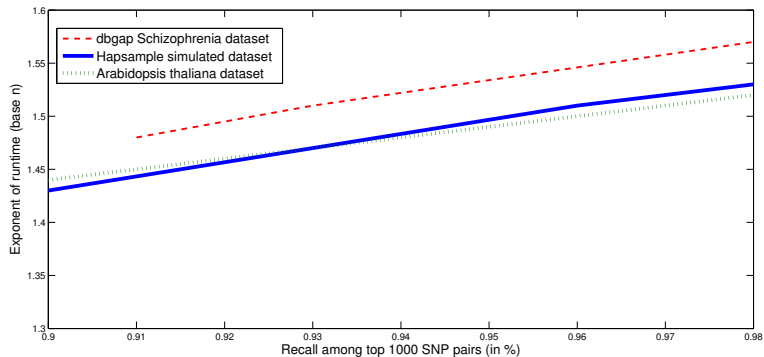
Results on *Arabidopsis* SNP dataset

# 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



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

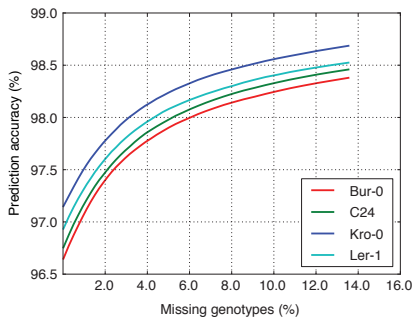
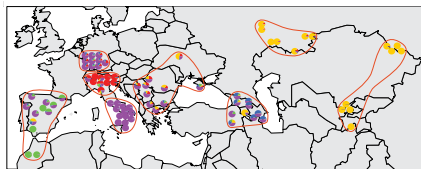
Road ahead

- ▶ We are performing the official SNP-SNP interaction discovery analysis for the international headache genetics consortium (Clinical Migraine)
- ▶ Our methods will be used in further consortia:
 - ▶ Psychiatric diseases such as autism, schizophrenia, depression

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)

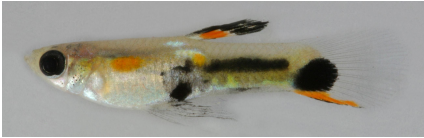


Setup

- ▶ 80 fully sequenced 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)

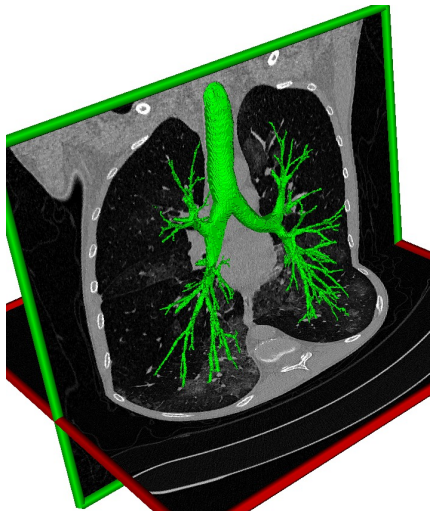


Setup

- ▶ 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



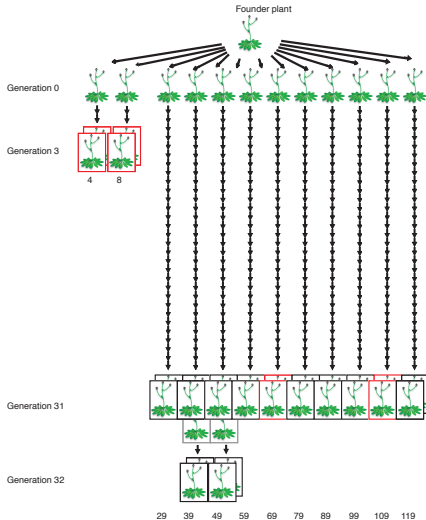
Setup

- ▶ 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):

The screenshot displays the easyGWAS web application interface. At the top, the browser address bar shows the URL: <https://easygwas.tuebingen.mpg.de/gwa/results/manhattan/view/daa5518d-4310-4ff1-96ff-6c699d3db7a5/>. The page header includes the easyGWAS logo and the text: "An integrated interspecies platform for performing and comparing genome-wide association studies" and "MACHINE LEARNING AND COMPUTATIONAL BIOLOGY RESEARCH GROUP".

The main navigation bar contains: easyGWAS, Home, GWA-Experiments, Data Center, Download Center, a search bar, FAQ, logged in as: kborgwardt, and Settings. A left sidebar menu lists: CREATE EXPERIMENTS (Overview, Create new GWAS, Tutorial), EXPERIMENT OVERVIEW (My temporary history, My experiments, Shared experiments, Public experiments), and Brief summary (Species: Arabidopsis thaliana, AtPolyOB (cal method 75, Horton et al.), SNPs: All SNPs selected, Phenotypes: Width 22, Additional factors: PCA: 5, Algorithm: Linear Regression).

The main content area shows "Manhattan Plots" for the selected experiment. It includes a "Download Summary Statistics" button and "GWA result options". Two Manhattan plots are displayed:

- Manhattan-plot for chromosome 1:** The y-axis is $-\log_{10}(p\text{-value})$ and the x-axis is "chromosomal position [bp]". A horizontal green line indicates the Bonferroni threshold [0.05].
- Manhattan-plot for chromosome 2:** The y-axis is $-\log_{10}(p\text{-value})$ and the x-axis is "chromosomal position [bp]". A horizontal green line indicates the Bonferroni threshold [0.05].

Each plot shows blue dots representing $-\log_{10}(p\text{-value})$ and a green horizontal line for the Bonferroni threshold [0.05]. The plots show a distribution of points across the chromosome, with some points reaching the threshold.

How can Machine Learning contribute to Statistical Genetics?

- ▶ 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 (Pasaniuc B et al., Nature Genetics 2012)
- ▶ By measuring the 'added value' of further molecular factors

- ▶ Future of Bioinformatics: Personalized Medicine

- ▶ **Personalized Medicine**
 - ▶ Tailoring medical treatment to the molecular properties of a patient
- ▶ **Biomarker Discovery**
 - ▶ Detecting molecular components that are indicative of disease outbreak, progression or therapy outcome
- ▶ **Biomarker**
 - ▶ The term 'biomarker', short for 'biological marker', refers to a broad subcategory of medical signs — that is, objective indications of medical state observed from outside the patient — which can be measured accurately and reproducibly (Strimbu and Tavel, 2010).

- ▶ **Producing molecular data: Sequencing costs**
 - ▶ USD 300,000,000 cost of sequencing a human genome in 2001
 - ▶ USD 5,000 cost of sequencing a human genome in 2011
- ▶ **Storing molecular data: Electronic health records**
 - ▶ 4% U.S. hospitals with fully operational electronic health records in 2008
 - ▶ 22% U.S. hospitals with fully operational electronic health records in 2009
 - ▶ 50% U.S. population that had medical information recorded in electronic health records in some form in 2010
- ▶ **Using molecular data: Products**
 - ▶ 13 prominent examples of personalized medicine drugs, treatments and diagnostics products available in 2006
 - ▶ 72 prominent examples of personalized medicine drugs, treatments and diagnostics products available in 2011

Source: http://www.ageofpersonalizedmedicine.org/personalized_medicine/case/

▶ Examples of success

- ▶ In Germany, for 33 drugs, a corresponding diagnostic molecular test has been approved (as of August 21, 2013).
- ▶ For 25 of these drugs, the test is even required.
- ▶ Drugs for HIV/AIDS, cancer (e.g. lung, breast, leukemia, lymphoma), epilepsy, cystic fibrosis
- ▶ Tests on diverse biomarkers: genetic properties, deletions of genes, types of cell receptors, overexpression of specific genes, chromosomal deletions, presence of antibodies, presence of particular types of virus
- ▶ Common consequence: Drug is administered or not

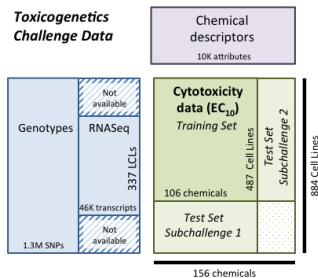
Source: Association of Research-based Pharmaceutical Companies, <http://vfa.de/personalisiert>

- ▶ U.S. FDA lists 121 drugs with pharmacogenomic information in their labels.
- ▶ Biomarkers may include gene variants, functional deficiencies, expression changes, chromosomal abnormalities.

Source: <http://www.fda.gov/drugs/scienceresearch/researchareas/pharmacogenetics/ucm083378.htm>

Personalized Medicine: Phenotype Prediction

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



Source: <https://www.synapse.org>

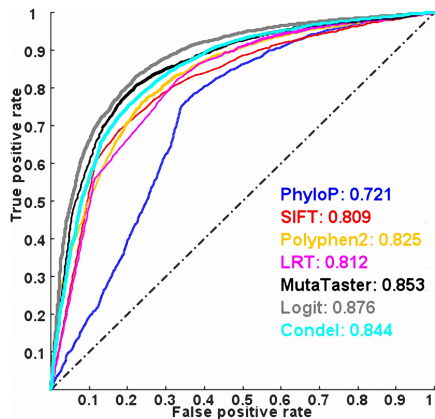
- ▶ Combining Genetics and Biochemistry

Loss-of-function (LoF) mutations (MacArthur et al., Science 2012)

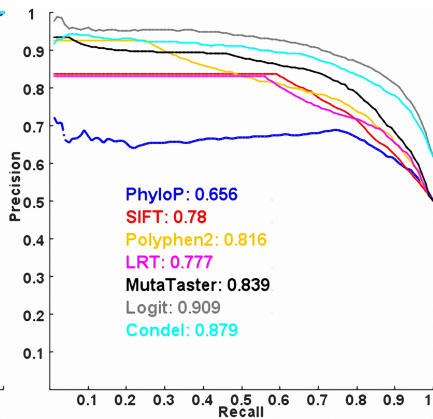
- ▶ MacArthur et al. assess 2951 putative LoF variants obtained from 185 human genomes to determine their true prevalence and properties.
- ▶ Human genomes typically contain **approx. 100 genuine LoF variants** with **approx. 20 genes** completely inactivated.

- ▶ Deleterious Variants ('Loci under purifying selection', loss-of-fitness variants)
 - ▶ Assessing the functional impact of sequence variants
 - ▶ Binary classification whether a variant has a deleterious effect or not
 - ▶ Commonly used features:
 - ▶ Conservation scores
 - ▶ Sequence features
 - ▶ Biochemical and physicochemical features
 - ▶ Structural features, annotation-based features
 - ▶ Recent empirical comparison by [Li et al., Plos Genetics 2013](#) of various predictors and meta-predictors:
 - ▶ PolyPhen-2 (Polymorphism Phenotyping v2) ([Adzhubei et al., N Meth 2010 and 2013](#))
 - ▶ MutationTaster ([Schwarz et al., N Meth 2010](#))
 - ▶ SIFT ([Sim et al., Nucleic Acid Research 2012](#))
 - ▶ LRT ([Chun and Fay, Genome Research 2009](#))
 - ▶ Two combined models (CONDEL and logit)

► Empirical Comparison on ExoVar Dataset (10-fold cross-validation)

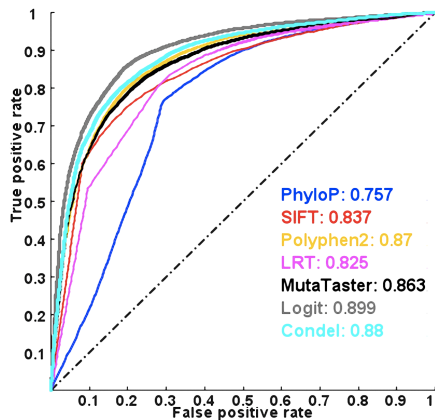


(a)

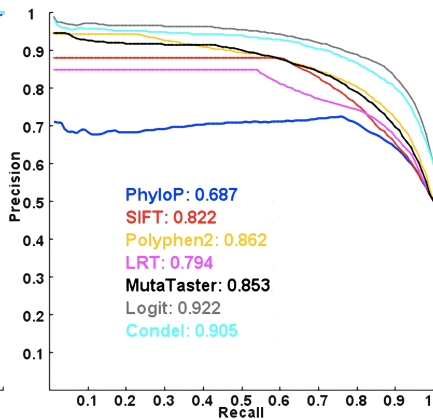


(b)

► Empirical Comparison on HumVar Dataset (10-fold cross-validation)



(a)

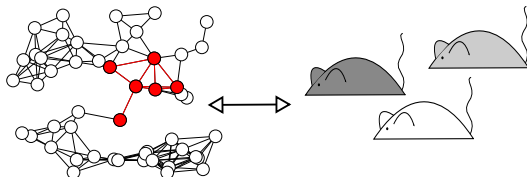


(b)

- ▶ Meta-Predictors outperform single predictors
- ▶ Room to define new, better meta-predictors
- ▶ Other areas that will receive attention (Wu et al., 2013):
 - ▶ Deleterious effect of non-coding mutations
 - ▶ Deleterious rare variant prediction
 - ▶ Disease-specific prioritization

- ▶ Bromberg et al. examine the structural impact of sequence variants in healthy and diseased individuals with SNAP (Bromberg & Rost, NAR 2007) and make two observations:
 - ▶ The first is expected: coding variants reported in disease-related databases significantly alter the function of affected proteins.
 - ▶ The second is surprising: the genomes of healthy individuals appear to carry many variants that are predicted to have some effect on function.
- ▶ They draw two conclusions:
 - ▶ Diseases may be extreme phenotypic variations and often attributable to one or a few severely functionally disruptive variants.
 - ▶ Nondisease phenotypes potentially arise through combinations of many variants whose effects are weakly nonneutral (damaging or enhancing) to the molecular protein function but fall within the wild-type range of overall physiological function.

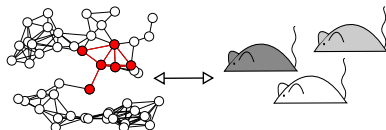
- ▶ Combining Genetics and Biological Network Analysis



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



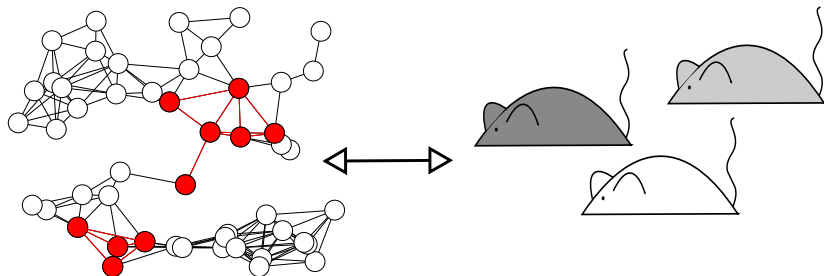
- ▶ Edges between SNPs near the same gene or SNPs in interacting genes
- ▶ c_i is the association score of SNP i , $f_i = 1$ if SNP i is selected, $f_i = 0$ if not.
- ▶ Find a set of SNPs with maximum total score:

$$\operatorname{argmax}_{\mathbf{f} \in \{0,1\}^n} \mathbf{c}^\top \mathbf{f}$$

such that

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

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

$$\mathbf{f}^\top \mathbf{L} \mathbf{f} = \sum_{i \sim j} (f_i - f_j)^2, \text{ where } \mathbf{L} = \mathbf{D} - \mathbf{W}.$$

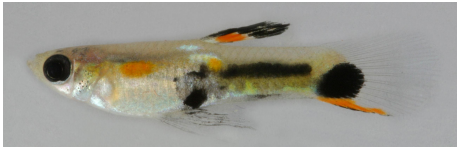
- ▶ The SNP subnetwork selection problem is then:

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

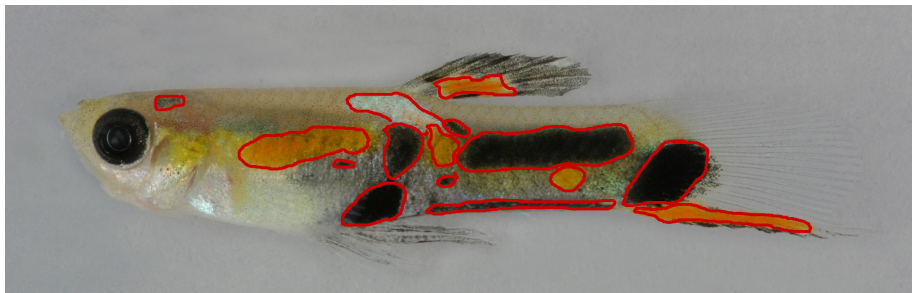
- ▶ This is a min-cut problem, for which efficient algorithms exist (we use Boykov and Kolmogorov, IEEE TPAMI 2004).
- ▶ Much faster and recovers four times more phenotype-related genes in *A. thaliana* than network-constrained Lasso models

- ▶ Combining Genetics and Bioimaging

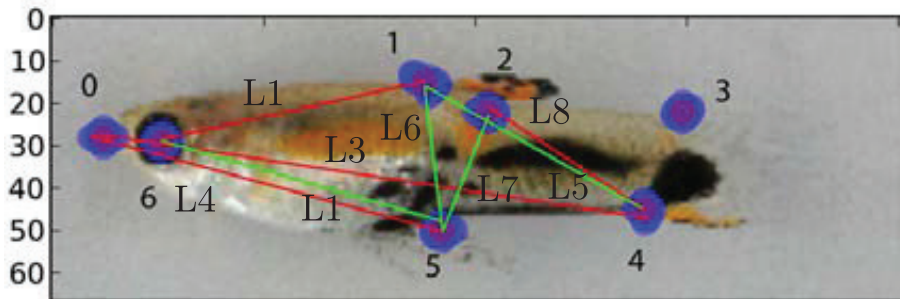
Bioimaging: Natural variation in male guppy fish



Bioimaging: Natural variation in male guppy fish

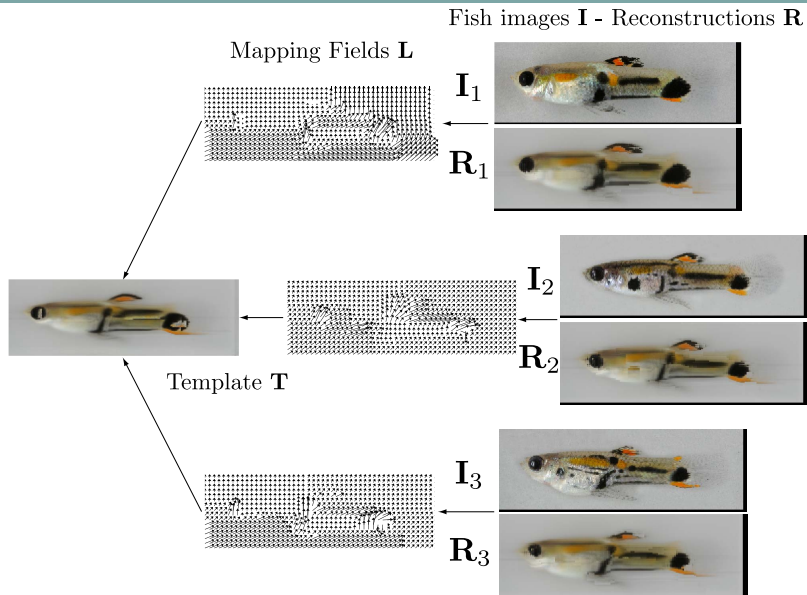


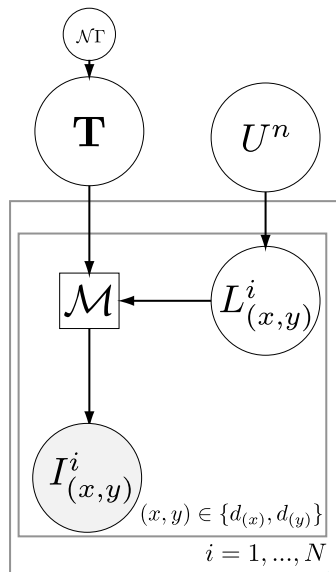
Bioimaging: From geometric measurements to shape deformations



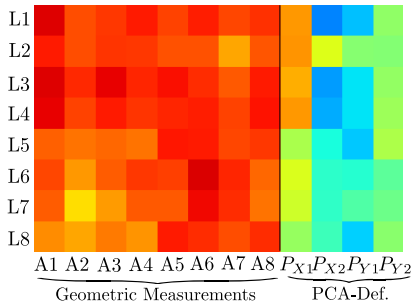
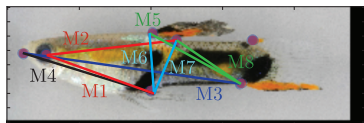
from Tripathi et al., 2009

Bioimaging: Reconstructing fish from a template

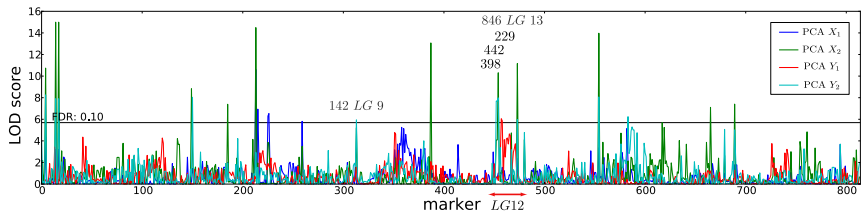




ShapePheno: From geometric measurements to shape deformations



ShapePheno: Association mapping of shape phenotypes



- ▶ Genetic information for thousands of patients suffering from *related* phenotypes is available
- ▶ **Biological question:** Is there a shared genetic basis of related diseases?
- ▶ **Machine learning task:** Are there features that are predictive of related phenotypes?
- ▶ A recent study by Lee et al. (Nature Genetics, August 2013)
 - ▶ Diseases: schizophrenia, bipolar disorder, major depressive disorder, autism spectrum disorders (ASD) and attention-deficit/hyperactivity disorder (ADHD)
 - ▶ The genetic correlation calculated using common SNPs was
 - ▶ high between schizophrenia and bipolar disorder (0.68 ± 0.04 s.e.),
 - ▶ moderate between schizophrenia and major depressive disorder (0.43 ± 0.06 s.e.), bipolar disorder and major depressive disorder (0.47 ± 0.06 s.e.), and ADHD and major depressive disorder (0.32 ± 0.07 s.e.),
 - ▶ low between schizophrenia and ASD (0.16 ± 0.06 s.e.) and
 - ▶ non-significant for other pairs of disorders as well as between psychiatric disorders and the negative control of Crohn's disease.

- ▶ Limitations of Phenotype Prediction

- ▶ Burga and Lehner argue that, although the *typical* phenotypic outcome of an individual's genome can be predicted, it is much more difficult to predict the actual outcome for a particular individual.
- ▶ Three reasons:
 - ▶ First, the outcome of mutations can be influenced by random (stochastic) processes.
 - ▶ Second, genetic variation present in one generation can influence phenotypic traits in the next generation, even if individuals do not inherit this variation.
 - ▶ Third, the environment experienced by one generation can influence phenotypic variation in the next generation.
- ▶ Long been appreciated by quantitative geneticists, although only recently studied at the molecular level
- ▶ Genotypes of individuals and the environment that they experience may not be sufficient to determine their phenotypes.

- ▶ Roberts et al. estimated the **capacity of whole-genome sequencing to identify individuals at clinically significant risk** (at least 10% positive predictive value) for 24 different complex diseases.
- ▶ Their estimates were derived from the analysis of large numbers of monozygotic twin pairs; twins of a pair share the same genomotype and therefore identical genetic risk factors.
- ▶ Their analyses indicate that:
 - ▶ (i) for 23 of the 24 diseases, the majority of individuals will receive negative test results,
 - ▶ (ii) these negative test results will, in general, not be very informative, as the risk of developing 19 of the 24 diseases in those who test negative will still be, at minimum, 50 - 80% of that in the general population, and
 - ▶ (iii) on the positive side, in the best-case scenario more than 90% of tested individuals might be alerted to a clinically significant predisposition to at least one disease.

- ▶ Queitsch et al. argue that the actual phenotype of an individual depends on its phenotypic robustness.
- ▶ **Phenotypic Robustness** is the ability of a given genotype to produce a constant phenotype, even when the organism is faced with genetic or environmental perturbations.
- ▶ Decreased phenotypic robustness significantly increases heritability of complex traits due to revealed, formerly **cryptic genetic variation** and increased penetrance of genetic variants
- ▶ The best-characterized master regulator of robustness is the molecular chaperone **HSP90**, which assists the proper folding and function of many key enzymes and transcription factors that govern growth and development.
- ▶ In humans, an increase in microsatellite mutations, transposon mobility, recombination rates, base-substitution mutation rate, and large duplications and deletions may indicate decrease in phenotypic robustness.

Outlier Detection

- ▶ Detect anomalies in large patient databases
- ▶ Must scale to large datasets of high-dimensional data

Sampling-Based Method (Mahito and Borgwardt, NIPS 2013a)

- ▶ Current Methods focus on efficient Nearest Neighbor Search via Indexing Structures
- ▶ New approach: Computer Nearest Neighbor among a small sample of points
- ▶ For outliers, it is much more unlikely to detect a similar point than for 'inliers'
- ▶ In an extensive empirical comparison, this sampling based approach is superior to the state-of-the-art methods in terms of runtime and efficacy
- ▶ The sample size can be optimized to maximize power.

Our Marie Curie Initial Training Network

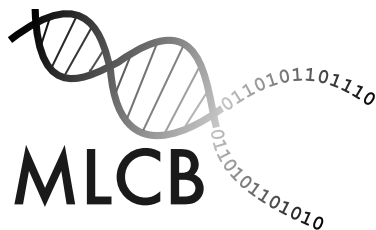
- ▶ Goal: Enable medical treatment tailored to patients' molecular properties
- ▶ Plan: Build a research community at the interface of Machine Learning and data-driven Medicine
- ▶ First step: [Marie Curie Initial Training Network \(ITN\)](#)
 - ▶ Topic: [Machine Learning for Personalized Medicine \(MLPM\)](#)
 - ▶ Duration: 4 years, started January 2013
 - ▶ 13 early-stage researchers + 1 postdoc 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](#)

Our ITN Research Projects

- SNP, gene, phenotype interaction models (4 projects)
- Heterogeneous data integration and decision support (3 projects)
- Disease subtype discovery (2 projects)
- Genome and transcriptome annotation (2 projects)
- Environmental and epigenetic effects (1 project)
- Feature selection (1 project)
- Adverse drug reaction (1 project)

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.**

"We are in a new era of the life sciences. . . but in no area of research is the promise greater than in the field of personalized medicine."

US Senator Edward M. Kennedy Remarks on the Senate's Consideration of the Genetic Information Nondiscrimination Act, April 24, 2008

Source: http://www.ageofpersonalizedmedicine.org/personalized_medicine/case/