



# Machine Learning for Personalized Medicine

Karsten Borgwardt

ETH Zürich

Fraunhofer-Institut Kaiserslautern, September 30, 2016

# 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.

# The Evolution of Bioinformatics

- Classic Bioinformatics: Focus on Molecules

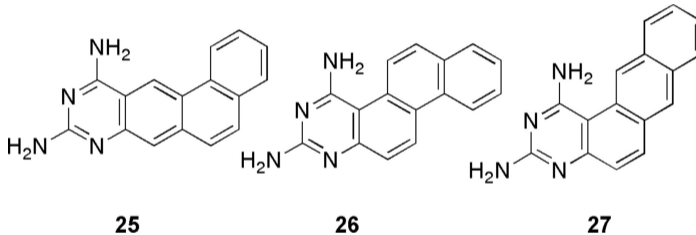
# Classic Bioinformatics: Focus on Molecules

- 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

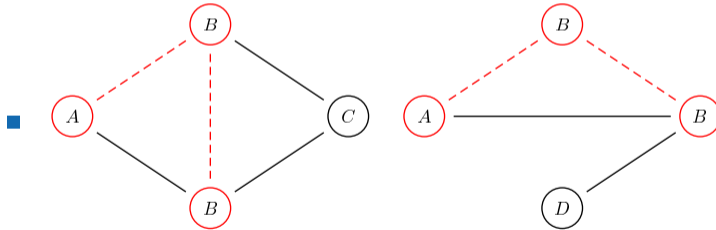


Source: Joska T M , and Anderson A C Antimicrob. Agents Chemother. 2006;50:3435-3443

### ■ 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?



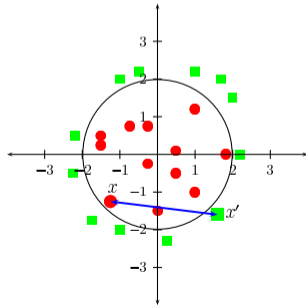
# Graph Comparison

- 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

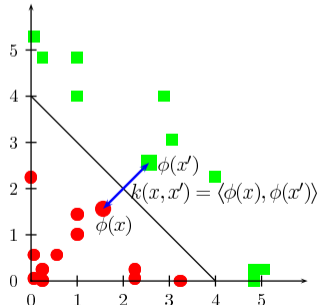
# Graph Kernels I

## ■ Kernels

- Key concept: Move problem to feature space  $\mathcal{H}$ .
- Naive explicit approach:
  - Map objects  $\mathbf{x}$  and  $\mathbf{x}'$  via mapping  $\phi$  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  $k(\mathbf{x}, \mathbf{x}') = \langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle$ .



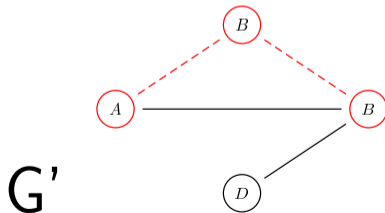
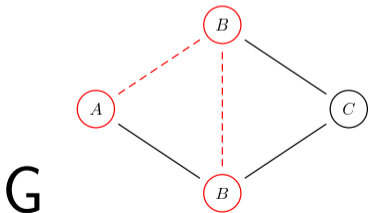
$\mathbb{R}^2 \Rightarrow \mathcal{H}$



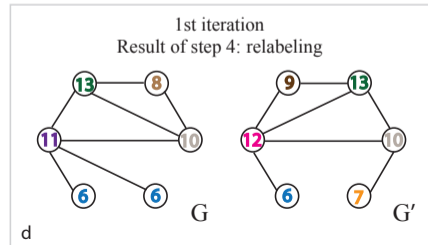
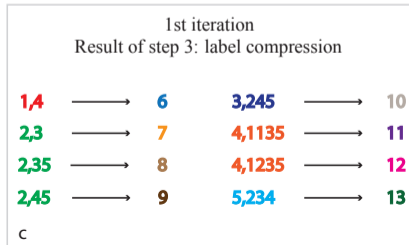
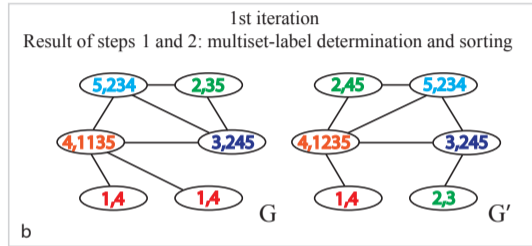
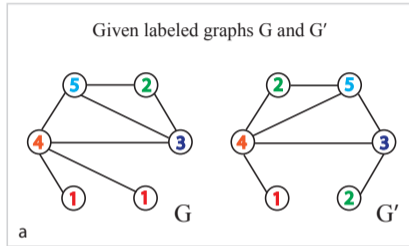
## Graph Kernels II

### ■ 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)



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

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

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

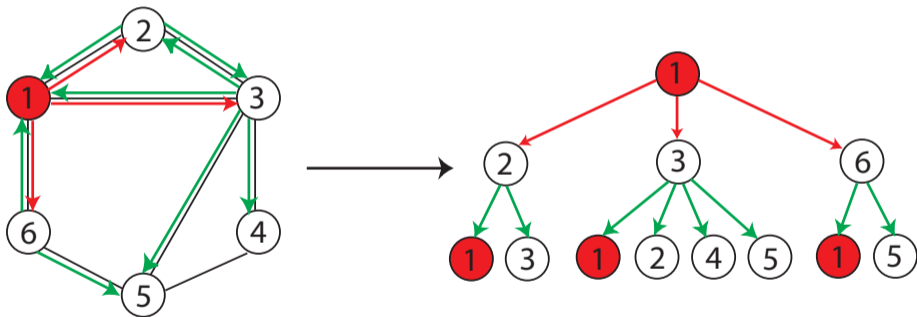
$$\phi_{WLsubtree}^{(1)}(G') = (1, 2, 1, 1, 1, 1, 1, 0, 1, 1, 0, 1, 1)$$

Counts of original node labels
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

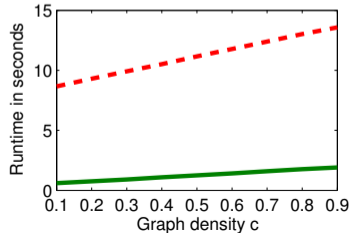
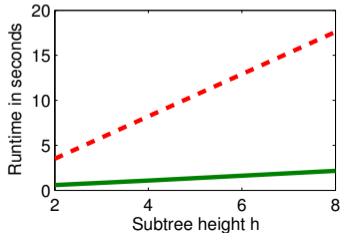
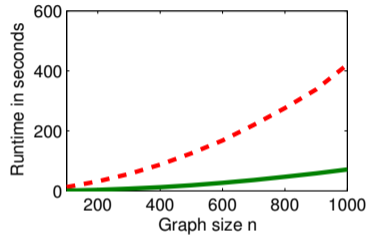
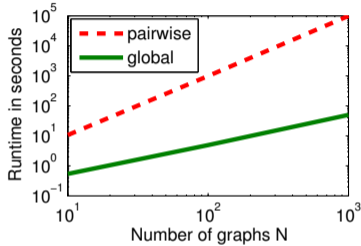




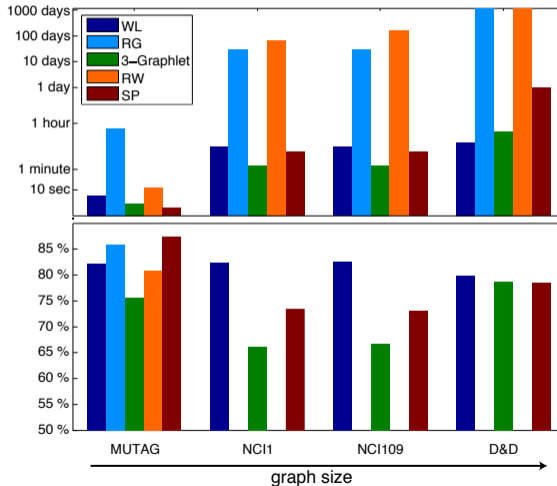
# Weisfeiler-Lehman Kernel: Theoretical Runtime Properties

- 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

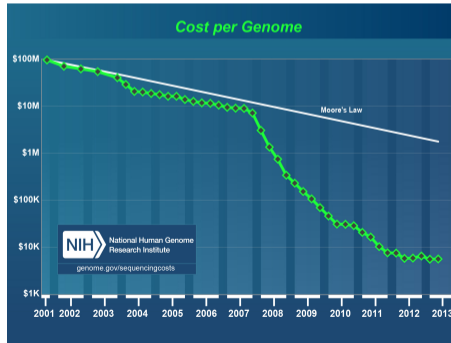


# The Evolution of Bioinformatics

- Modern Bioinformatics: Focus on Individuals

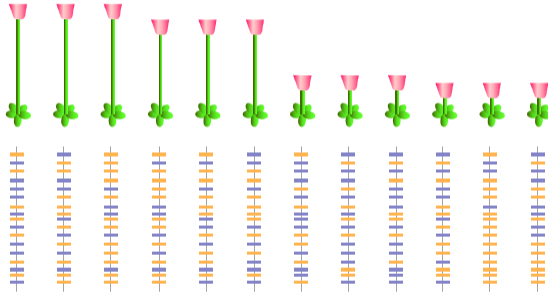
## Modern Bioinformatics: Focus on Individuals

- 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



## Genetics: Association Studies

### ■ Genome-Wide Association Studies (GWAS)

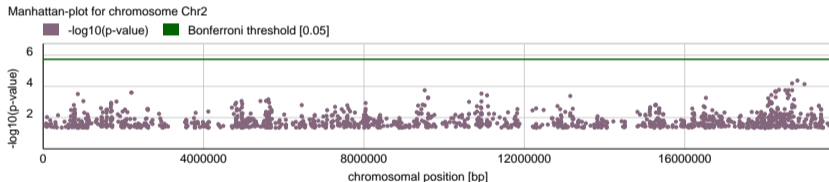


bco D. Weigel

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

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



## 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

# Machine Learning in Genetics I

## 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: Epistasis discovery (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)

# Machine Learning in Genetics II

## Moving to a Systems Biology Perspective

- Multi-locus models:
  - Efficient algorithms for discovering trait-related SNP pairs: Epistasis discovery (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)

## Epistasis: Computational Bottlenecks

### 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.

## Epistasis: Common approaches in the literature

### 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)

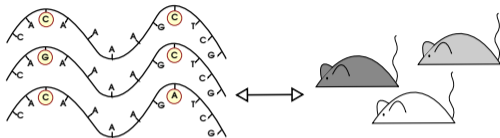
### Filtering approaches

- 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: 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:

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

- 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:

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

### 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 **A** and a matrix of controls **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  $\theta(\mathbf{x}_i, \mathbf{x}_j)$  is the angle between the two vectors  $\mathbf{x}_i$  and  $\mathbf{x}_j$ .

## 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:

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

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

# 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

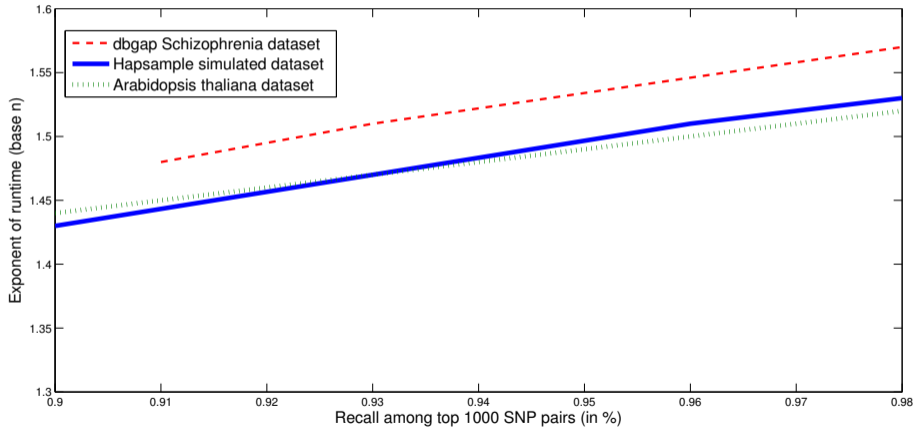
# SNPs	Measurements	Pairs	Exponent	Speedup	Top 10	Top 100	Top 500	Top 1K
100,000	8,255,645	8,186,657	1.38	611	1.00	0.86	0.82	0.80
100,000	52,762,001	51,732,700	1.54	97	1.00	1.00	0.99	0.98

### 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

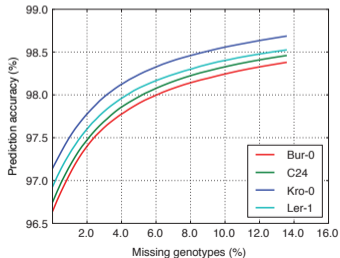
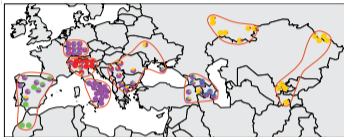


## Multi-Locus Models: Current Work

### 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 (Sugiyama et al., arxiv 2014)
- Modelling correlations between multiple phenotypes (Rakitsch et al., NIPS 2013)

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



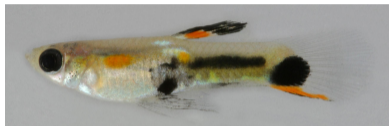
## Setup

- 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)

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



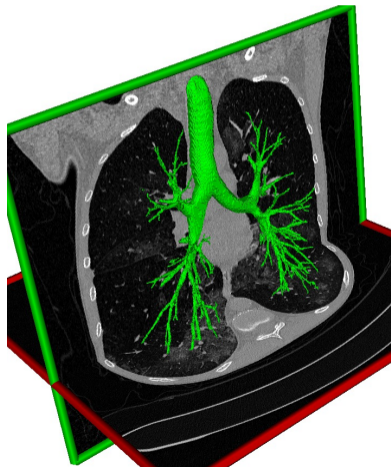
## Setup

- 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., NIPS 2013c)



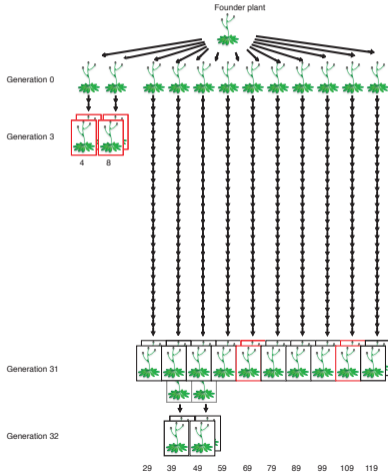
### Setup

- 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, Haggmann et al., PLoS Genetics, 2015)



## Setup

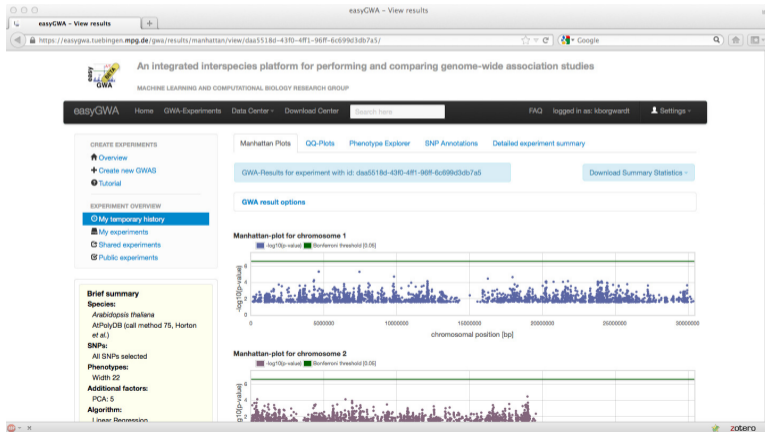
- 33 generations of lab strains of *A. thaliana*
- How stable is the methylation state of genome positions across generations?

## 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.org/) (<https://easygwas.org/>), a machine learning platform for analysing complex traits (Grimm et al., arXiv 2012):



# The Evolution of Bioinformatics

- Future of Bioinformatics: Personalized Medicine



# Personalized Medicine: Biomarker Discovery

## ■ 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).

## Personalized Medicine: Where We Stand

- Producing molecular data: Sequencing costs
  - USD 300,000,000 cost of sequencing a human genome in 2001
  - USD 1,000 cost of sequencing a human genome in 2014
- Storing molecular data: Electronic health records
  - 29% of U.S. physicians used an electronic health system in 2006
  - 93% reported actively using medical records in 2013
- Using molecular data: Products
  - 13 prominent examples of personalized medicine drugs, treatments and diagnostics products available in 2006
  - 113 prominent examples of personalized medicine drugs, treatments and diagnostics products available in 2014

Source: <http://www.ageofpersonalizedmedicine.org/>



# Significant Pattern Mining

Karsten Borgwardt

ETH Zürich

Fraunhofer-Institut Kaiserslautern, September 30, 2016

## Biomarker Discovery

Class 1



I	II	III	IV	V
1	0	1	0	0
1	1	0	0	0
0	0	0	1	0
0	1	0	0	0
1	1	1	1	1
0	1	0	1	0

Class 2



VI	VII	VIII	IX	X
0	1	0	1	1
0	1	0	0	0
1	1	0	0	0
0	0	0	0	0
0	0	0	0	0
1	0	0	1	0

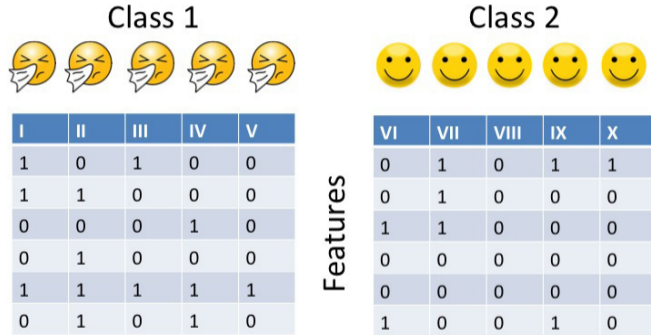
Features

# Biomarker Discovery as a Pattern Mining Problem

## Finding groups of disease-related molecular factors

- Single genetic variants, gene expression levels, protein abundancies are often not sufficiently indicative of disease outbreak, progression or therapy outcome.
- Searching for combinations of these molecular factors creates an enormous search space, and two inherent problems:
  - 1 Computational level: How to efficiently search this large space?
  - 2 Statistical level: How to properly account for testing an enormous number of hypotheses?
- The vast majority of current work in this direction (e.g. Achlioptas et al., KDD 2011) focuses on Problem 1, the computational efficiency.
- **But Problem 2, multiple testing, is also of fundamental importance!**

# Biomarker Discovery as a Pattern Mining Problem



- Feature Selection: Find features that distinguish classes of objects
- Pattern Mining: Find higher-order **combinations of binary features**, so-called *patterns*, to distinguish one class from another

# Pattern Mining

## Definition of a pattern

- A pattern is a set of dimensions  $S$ .
- A binary vector  $\mathbf{x}$  contains a pattern  $S$  if  $\prod_{i \in S} x_i = 1$ , where  $x_i$  is dimension  $i$  of  $\mathbf{x}$ .

## Definition of a frequent pattern

- Assume we are given  $n$  vectors  $\{\mathbf{x}_j\}_{j=1}^n$ .
- If at least  $\theta$  vectors contain pattern  $S$ , then  $S$  is a frequent pattern.
- $\theta$  is a pre-defined frequency threshold.

# Pattern Mining

## Frequent pattern mining

- Analogous definitions for frequent **subgraph** mining and frequent **substring** mining.
- Numerous branch-and-bound algorithms have been proposed for finding frequent patterns (Aggarwal and Han, 2014).

## Problem statement: Significant Pattern Mining

- Assume we are given  $n$  vectors  $\{\mathbf{x}_j\}_{j=1}^n$ , each of which has a binary class label  $y_j \in \{0, 1\}$ .
- Our goal is to find **all patterns  $S$  that are statistically significant enriched in one of the two classes  $y = 0$  or  $y = 1$ .**



# Statistical Significance and Testability

## Fisher's exact test

### ■ Contingency Table

	$ S \in \mathbf{x} $	$ S \notin \mathbf{x} $	
$y = 0$	$a$	$n_0 - a$	$n_0$
$y = 1$	$x - a$	$n - n_0 - x + a$	$n - n_0$
	$x$	$n - x$	$n$

- A popular choice is Fisher's exact test to test whether  $S$  is overrepresented in one of the two classes.
- The common way to compute  $p$ -values for Fisher's exact test is based on the hypergeometric distribution and assumes fixed total marginals  $(x, n_0, n)$ .

# Statistical Significance and Testability

## Multiple testing correction in pattern mining

- The number of candidate patterns grows exponentially with the cardinality of the pattern.

# Statistical Significance and Testability

## Multiple testing correction in pattern mining

- The number of candidate patterns grows exponentially with the cardinality of the pattern.
- If we do not correct for multiple testing,  $\alpha$  per cent of all candidate patterns will be false positives.

# Statistical Significance and Testability

## Multiple testing correction in pattern mining

- The number of candidate patterns grows exponentially with the cardinality of the pattern.
- If we do not correct for multiple testing,  $\alpha$  per cent of all candidate patterns will be false positives.
- If we do correct for multiple testing, e.g. via Bonferroni correction ( $\frac{\alpha}{\#tests}$ ), then we lose any statistical power.

# Statistical Significance and Testability

## Tarone's trick

- **Tarone's insight:** When working with discrete test statistics (e.g. Fisher's exact test), there is a minimum  $p$ -value that a given pattern can obtain, based on its total frequency.

# Statistical Significance and Testability

## Tarone's trick

- **Tarone's insight:** When working with discrete test statistics (e.g. Fisher's exact test), there is a minimum  $p$ -value that a given pattern can obtain, based on its total frequency.
- **Tarone's trick (1990):** Ignore those patterns in multiple testing correction, for which the minimum  $p$ -value is larger than the Bonferroni-corrected significance threshold.

## Statistical Significance and Testability

### Tarone's trick

- **Tarone's insight:** When working with discrete test statistics (e.g. Fisher's exact test), there is a minimum  $p$ -value that a given pattern can obtain, based on its total frequency.
- **Tarone's trick (1990):** Ignore those patterns in multiple testing correction, for which the minimum  $p$ -value is larger than the Bonferroni-corrected significance threshold.
- If the  $p$ -values are conditioned on the total marginals (e.g. in Fisher's exact test), Tarone's trick does not increase the Family Wise Error rate.

## Mining Significant Patterns

### Tarone's approach (1990)

- For a discrete test statistics  $T(S)$  for a pattern  $S$ , such as in Fisher's exact test, there is a minimum obtainable p-value,  $p_{min}(S)$ .
- For some  $S$ ,  $p_{min}(S) > \frac{\alpha}{m}$ . Tarone refers to them as *untestable hypotheses*  $\bar{U}$ .
- **Tarone's strategy:** Ignore untestable hypotheses  $\bar{U}$  when counting the number of tests  $m$  for Bonferroni correction.



## Mining Significant Patterns

### Tarone's approach (1990)

- For a discrete test statistics  $T(S)$  for a pattern  $S$ , such as in Fisher's exact test, there is a minimum obtainable p-value,  $p_{min}(S)$ .
- For some  $S$ ,  $p_{min}(S) > \frac{\alpha}{m}$ . Tarone refers to them as *untestable hypotheses*  $\bar{U}$ .
- **Tarone's strategy:** Ignore untestable hypotheses  $\bar{U}$  when counting the number of tests  $m$  for Bonferroni correction.
- If the  $p$ -values of the test are conditioned on the total marginals (as in Fisher's exact test), this does not affect the Family-Wise Error Rate.

## Mining Significant Patterns

### Tarone's approach (1990)

- For a discrete test statistics  $T(S)$  for a pattern  $S$ , such as in Fisher's exact test, there is a minimum obtainable p-value,  $p_{min}(S)$ .
- For some  $S$ ,  $p_{min}(S) > \frac{\alpha}{m}$ . Tarone refers to them as *untestable hypotheses*  $\bar{U}$ .
- **Tarone's strategy:** Ignore untestable hypotheses  $\bar{U}$  when counting the number of tests  $m$  for Bonferroni correction.
- If the  $p$ -values of the test are conditioned on the total marginals (as in Fisher's exact test), this does not affect the Family-Wise Error Rate.
- **Difficulty:** There is an interdependence between  $m$  and  $\bar{U}$ .

# Mining Significant Patterns

## Tarone's approach (1990)

- Assume  $k$  is the number of tests that we correct for.
- $m(k)$  is the number of testable hypotheses at significance level  $\frac{\alpha}{k}$ .
- Then the optimization problem is

$$\begin{aligned} & \min k \\ & \text{s. t. } k \geq m(k) \end{aligned}$$

# Mining Significant Patterns

## Tarone's approach (1990)

- Assume  $k$  is the number of tests that we correct for.
- $m(k)$  is the number of testable hypotheses at significance level  $\frac{\alpha}{k}$ .

**procedure** TARONE

$k := 1;$

**while**  $k < m(k)$  **do**

$k := k + 1;$

**return**  $k$

## Mining Significant Patterns

Terada's link to frequent itemset mining (Terada et al., PNAS 2013)

- For  $0 \leq x \leq n_1$ , the minimum p-value  $p_{min}(S)$  decreases monotonically with  $x$ .
- One can use *frequent itemset mining* to find all  $S$  that are testable at level  $\alpha$ , with frequency  $\psi^{-1}(\alpha)$ .
- They propose to use a decremental search strategy:

**procedure** TERADA'S DECREMENTAL SEARCH (LAMP)

$k :=$  "very large";

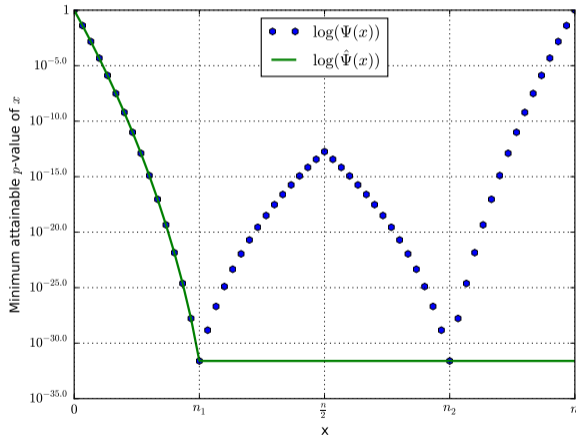
**while**  $k > m(k)$  **do**

$k := k - 1$ ;

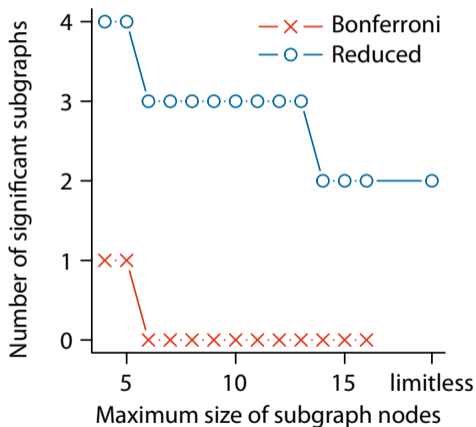
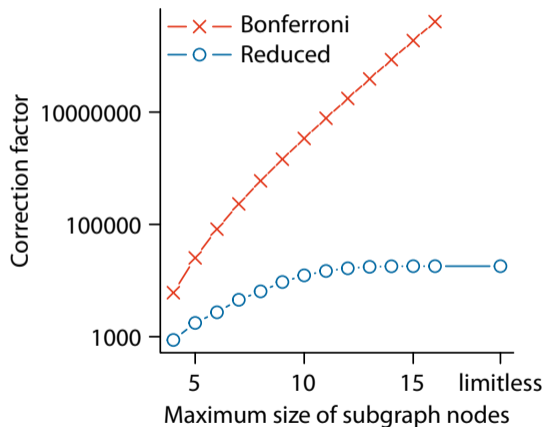
$m(k) :=$  frequent itemset mining( $D, \psi^{-1}(\frac{\alpha}{k})$ );

**return**  $k + 1$

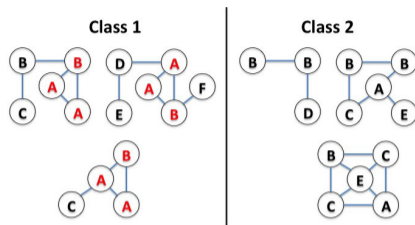
## Mining Significant Patterns



## Example: PTC dataset (Helma et al., 2001)



## Significant Subgraph Mining (Sugiyama et al., SDM 2015)



### Significant Subgraph Mining

- Each object is a graph.
- A pattern is a subgraph in these graphs.
- Typical application in Drug Development: Find subgraphs that discriminate between molecules with and without drug effect.
- Counting all tests (= all patterns) requires exponential runtime in the number of nodes.



## Significant Subgraph Mining (Sugiyama et al., SDM 2015)

### Incremental search with early stopping

- **procedure** INCREMENTAL SEARCH WITH EARLY STOPPING

$\theta := 0$

**repeat**

$\theta := \theta + 1; FS_{\theta} := 0;$

**repeat**

find next frequent subgraph at frequency  $\theta$

$FS_{\theta} := FS_{\theta} + 1$

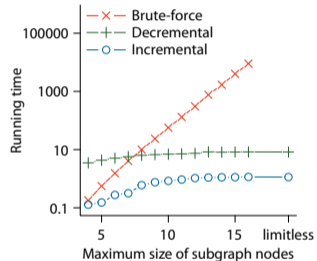
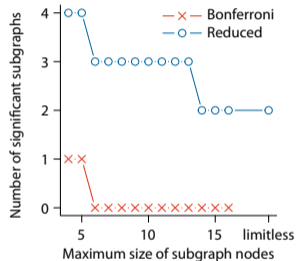
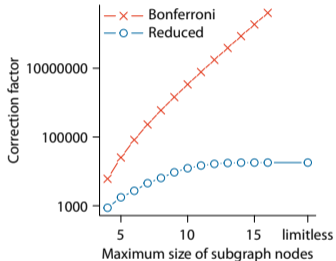
**until** (no more frequent subgraph found) or  $(FS_{\theta} > \frac{\alpha}{\psi(\theta)})$

**until**  $FS_{\theta} \leq \frac{\alpha}{\psi(\theta)}$

**return**  $\psi(\theta)$

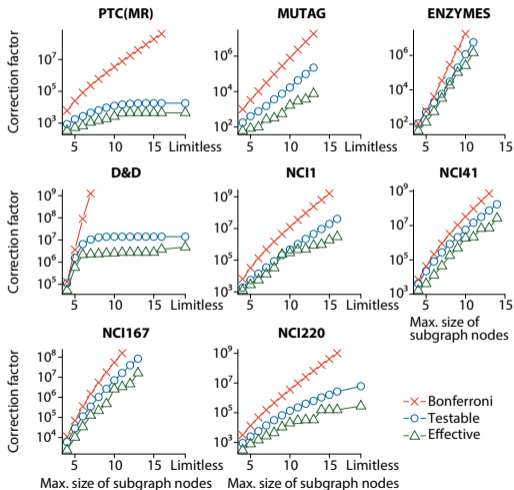
- $\frac{\alpha}{\psi(\theta)}$  is the maximum correction factor, such that subgraphs with frequency  $\theta$  can be significant at level  $\psi(\theta)$ .

# Significant Subgraph Mining on PTC Dataset

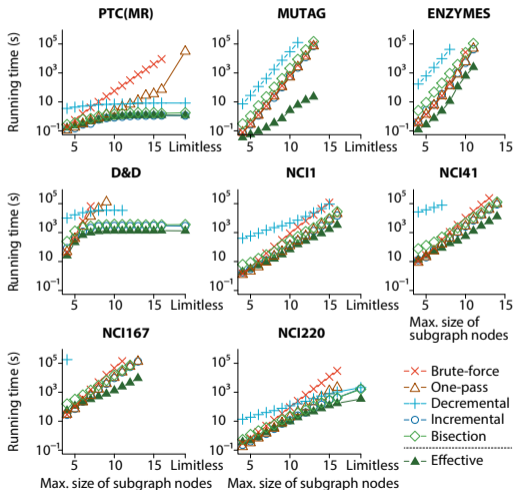


Dataset from Helma et al. (2001)

# Significant Subgraph Mining: Correction Factor



# Significant Subgraph Mining: Runtime



## Westfall-Young light (Llinares-Lopez et al., KDD 2015)

### Dependence between hypotheses

- As patterns are often in sub-/superpattern-relationships, they do not constitute independent hypotheses.
- Informally: The underlying number of hypotheses may be much lower than the raw count.
- Westfall-Young-Permutation tests (Westfall and Young, 1993), in which the class labels are repeatedly permuted to approximate the null distribution, are one strategy to take this dependence into account.
- **Computational problem: How to efficiently perform these thousands of permutations?**
- There is one existing approach, FastWY (Terada et al., ICBB 2013), which suffers from either memory or runtime problems.

## Westfall-Young light (Llinares-Lopez et al., KDD 2015)

## The Algorithm

- 1 Input:** Transactions  $D$ , class labels  $\mathbf{y}$ , target FWER  $\alpha$ , number of permutations  $j_p$ .
- 2** Perform  $j_p$  permutations of the class label  $\mathbf{y}$  and store each permutation as  $\mathbf{c}_j$ .
- 3** Initialize  $\theta := 1$  and  $\delta^* := \psi(\theta)$  and  $p_{min}^{(j)} := 1$ .
- 4** Perform a depth first search on the patterns:
  - Compute the  $p$ -value of pattern  $S$  across all permutations, update  $p_{min}^{(j)}$  if necessary.
  - Update  $\delta^*$  by  $\alpha$ -quantile of  $p_{min}^{(j)}$ , and increase  $\theta$  accordingly.
  - Process all children of  $S$  with frequency  $\geq \psi^{-1}(\delta^*)$ .
- 5 Output:** Corrected significance threshold  $\delta^*$ .

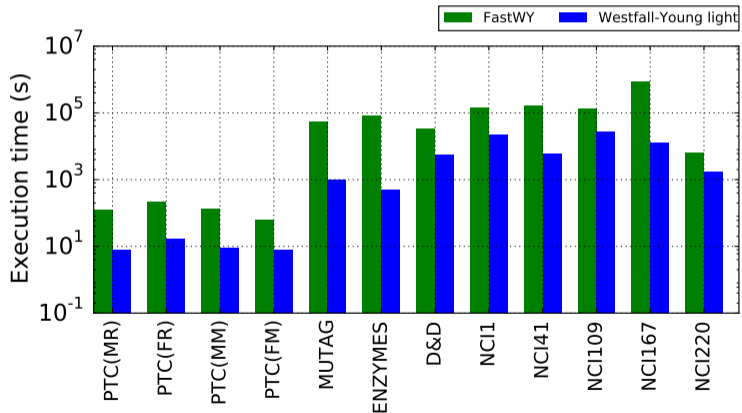
## Westfall-Young light (Llinares-Lopez et al., KDD 2015)

### Speed-up tricks of Westfall-Young light

- Follows incremental search strategy rather than decremental search strategy of FastWY
- Performs only one iteration of frequent pattern mining
- Does not store the occurrence list of patterns
- Does not compute the upper  $1 - \alpha$  quantile of minimum p-values exactly.
- Reduces the number of cell counts that have to be evaluated
- Shares the computation of p-values across permutations

## Westfall-Young light

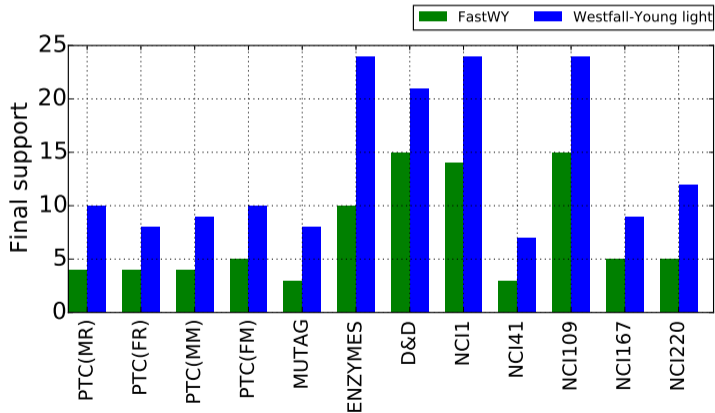
## ■ Runtime





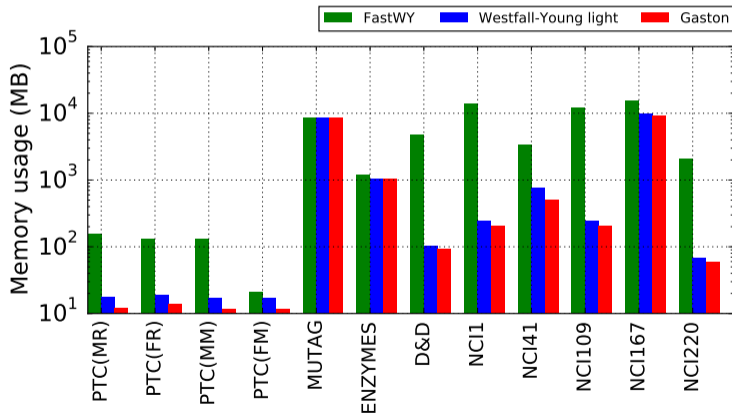
## Westfall-Young light

- Final frequency threshold (support)



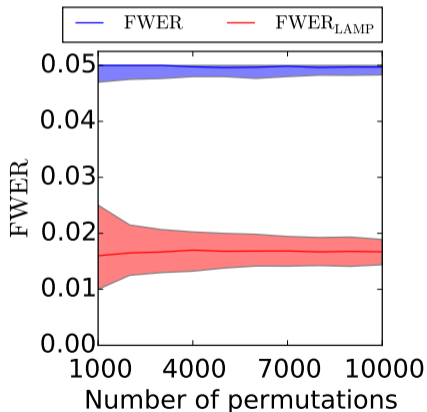
## Westfall-Young light

## ■ Peak memory usage

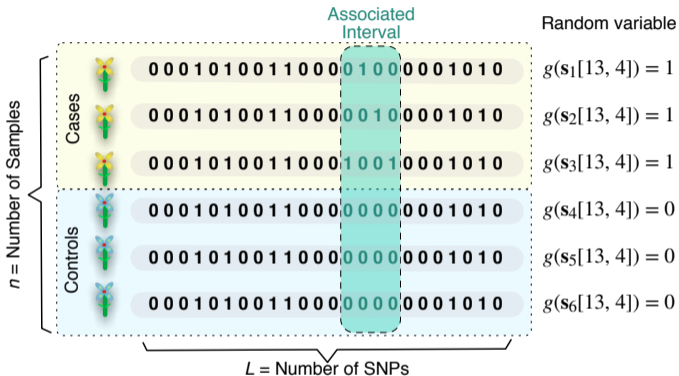


## Westfall-Young light

- Better control of the Family-wise error rate (Enzymes)



# Detecting “Genetic Heterogeneity” (Llinares et al., ISMB 2015b)



## Detecting intervals significantly associated with phenotypic variation

- Find subsequences which tend to contain at least one minor allele in one of the two phenotypic groups

## Education: 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, 2013-2016
  - 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](#)

# Summary

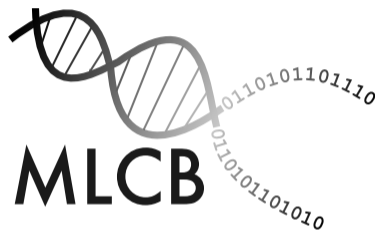
## Machine Learning in Bioinformatics

- Classic Bioinformatics
  - Comparing graphs in Chemoinformatics
- Current Bioinformatics
  - High-dimensional feature selection in Statistical Genetics
- Future Bioinformatics
  - Significant pattern mining for biomarker discovery in **Personalized Medicine**

## Thank You

### Postdocs and PhD students:

- Dean Bodenham
- Lukas Folkman
- Udo Gieraths
- Thomas Gumbsch
- Anja Gumpinger
- Xiao He
- Felipe Llinares Lopez
- Laetitia Papaxanthos
- Damian Roqueiro
- Caroline Weis



### Sponsors:

- Krupp-Stiftung
- Marie-Curie-FP 7
- Starting Grant (SNSF's ERC backup scheme)

# Thank You



[bsse.ethz.ch/mlcb](http://bsse.ethz.ch/mlcb)












# Machine Learning for Personalized Medicine

Karsten Borgwardt







ETH Zürich

Fraunhofer-Institut Kaiserslautern, September 30, 2016






## Main References I

-  C.-A. Azencott, D. Grimm, M. Sugiyama, Y. Kawahara, K. M. Borgwardt, *Bioinformatics (Oxford, England)* **29**, i171 (2013).
-  P. Achlioptas, B. Schölkopf, K. Borgwardt, *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)* (2011), pp. 726–734.
-  C. Becker, *et al.*, *Nature* **480**, 245 (2011).
-  J. Cao, *et al.*, *Nature Genetics* **43**, 956 (2011).
-  A. Feragen, N. Kasenburg, J. Petersen, M. de Bruijne, K. M. Borgwardt, *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013*. (2013), pp. 216–224.
-  D. Grimm, *et al.*, *arXiv:1212.4788* (2012).
-  D. G. Grimm, *et al.*, *Human Mutation* **36**, 513 (2015).

## Main References II

-  T. Kam-Thong, *et al.*, *Eur J Hum Genet* (2010).
-  T. Kam-Thong, B. Pütz, N. Karbalai, B. Müller-Myhsok, K. Borgwardt, *Bioinformatics (ISMB)* **27**, i214 (2011).
-  T. Kam-Thong, *et al.*, *Human Heredity* **73**, 220 (2012).
-  T. Karaletsos, O. Stegle, C. Dreyer, J. Winn, K. M. Borgwardt, *Bioinformatics* **28**, 1001 (2012).
-  L. Li, B. Rakitsch, K. Borgwardt, *Bioinformatics (Oxford, England)* **27**, i342 (2011). PMID: 21685091.
-  F. Llinares-López, M. Sugiyama, L. Papaxanthos, K. M. Borgwardt, *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, August 10-13, 2015*, L. Cao, *et al.*, eds. (ACM, 2015), pp. 725–734.

## Main References III

-  F. Llinares-López, *et al.*, *Bioinformatics* **31**, 240 (2015).
-  N. Shervashidze, K. M. Borgwardt, *NIPS*, Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, A. Culotta, eds. (MIT Press, Cambridge, MA, 2009), pp. 1660–1668.
-  M. Sugiyama, K. M. Borgwardt, *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013*. (2013), pp. 467–475.
-  M. Sugiyama, C. Azencott, D. Grimm, Y. Kawahara, K. M. Borgwardt, *Proceedings of the 2014 SIAM International Conference on Data Mining, Philadelphia, Pennsylvania, USA, April 24-26, 2014* (2014), pp. 199–207.
-  M. Sugiyama, F. Llinares-López, N. Kasenburg, K. M. Borgwardt, *SIAM Data Mining* (2015).

## Main References IV

-  R. E. Tarone, *Biometrics* **46**, 515 (1990).
-  A. Terada, M. Okada-Hatakeyama, K. Tsuda, J. Sese, *Proceedings of the National Academy of Sciences* **110**, 12996 (2013).
-  A. Terada, K. Tsuda, J. Sese, *IEEE International Conference on Bioinformatics and Biomedicine* (2013), pp. 153–158.
-  P. H. Westfall, S. S. Young, *Statistics in Medicine* **13**, 1084 (1993).