



# Machine Learning for Personalized Medicine

Karsten Borgwardt

ETH Zürich, Department Biosystems

Kartause Ittingen, November 2, 2015

# Why do we need Machine Learning in Systems Biology and Personalized Medicine?

# Recent News: Predicting Sexual Orientation



The screenshot shows the top portion of a news article on the Nature website. The header is dark red with the 'nature' logo in white. Below the logo is the tagline 'International weekly journal of science'. A navigation bar contains links for 'Home', 'News & Comment', 'Research', 'Careers & Jobs', 'Current Issue', and 'Archive'. A secondary navigation bar highlights 'News & Comment', 'News', '2015', 'October', and 'Article'. The article title is 'Epigenetic 'tags' linked to homosexuality in men'. The sub-headline reads 'Twin study reveals five DNA markers that are associated with sexual orientation.' The author is 'Sara Reardon'. The date is '08 October 2015' and it was updated on '12 October 2015'. A 'Rights & Permissions' button is visible at the bottom of the article preview.

**nature** International weekly journal of science

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive

News & Comment > News > 2015 > October > Article

NATURE | NEWS

Epigenetic 'tags' linked to homosexuality in men

Twin study reveals five DNA markers that are associated with sexual orientation.

**Sara Reardon**

08 October 2015 | Updated: 12 October 2015

 [Rights & Permissions](#)

## Recent News: Predicting Sexual Orientation

Claim that sexual orientation can be predicted (Source: Science— DOI: [10.1126/science.aad4686](https://doi.org/10.1126/science.aad4686))

- At ASHG 2015, the Vilain lab from UCLA claimed that certain methylation patterns in the human genome are predictive for sexual orientation.
- Tuck Ngun from this lab considered methylation patterns at 140,000 regions in the DNA of 37 pairs of male identical twins who were discordant and 10 pairs who were both homosexual.
- They reported to have identified five regions in the genome where the methylation pattern appears very closely linked to sexual orientation.
- The team reached 70% prediction accuracy when splitting the discordant twin pairs into 2 groups, one for training, one for testing.

# Recent News: Predicting Sexual Orientation

Criticisms (by Ed Yong, The Atlantic)

- Sample size is very small.

# Recent News: Predicting Sexual Orientation

## Criticisms (by Ed Yong, The Atlantic)

- Sample size is very small.
  - Ngun: “Yes, we were underpowered.”

# Recent News: Predicting Sexual Orientation

## Criticisms (by Ed Yong, The Atlantic)

- Sample size is very small.
  - Ngun: “Yes, we were underpowered.”
- Overfitting on the test set.

# Recent News: Predicting Sexual Orientation

## Criticisms (by Ed Yong, The Atlantic)

- Sample size is very small.
  - Ngun: “Yes, we were underpowered.”
- Overfitting on the test set.
  - Ngun: “All models (from the very first to the final one) were built using JUST the training data... If performance was unsatisfactory, we remade the model by selecting a different set of predictors/features/data based on information from the TRAINING set and then reevaluating on the test set.”



# Recent News: Predicting Sexual Orientation

## Criticisms (by Ed Yong, The Atlantic)

- No correction for multiple testing

# Recent News: Predicting Sexual Orientation

## Criticisms (by Ed Yong, The Atlantic)

- No correction for multiple testing
  - Ngun: “We are not testing whether each of the 6000 marks/loci are significantly associated with sexual orientation. If we had done that, multiple testing correction would have certainly been warranted. But we didn't. The single test we did was to ask whether the final model we had built was performing better than random guessing.”

# Recent News: Predicting Sexual Orientation

## Lessons we should learn

- 1 Predicting complex traits from high-dimensional molecular data is (becoming) a reality. Low sample size is still an important obstacle.
- 2 It is important to build predictors that generalize to unseen data and to avoid overfitting.
- 3 When searching high-dimensional spaces for higher-order associations, multiple testing correction is an enormous problem.

# Recent News: Predicting Sexual Orientation

## Lessons we should learn

- 1** Predicting complex traits from high-dimensional molecular data is (becoming) a reality. Low sample size is still an important obstacle.
  - Roqueiro, Witteveen et al. Bioinformatics/ISMB, 2015; Sugiyama and Borgwardt, NIPS 2015
- 2** It is important to build predictors that generalize to unseen data and to avoid overfitting.
  - Grimm et al., Human Mutation 2015
- 3** When searching high-dimensional spaces for higher-order associations, multiple testing correction is an enormous problem.
  - Sugiyama et al., SDM 2015; Llinares-Lopez et al., Bioinformatics/ISMB 2015, KDD 2015

# Overfitting and Generalization: Deleteriousness Prediction

# Deleteriousness Prediction

## Assessing the impact of missense variants

- Given the availability of more and more sequencing data on individual patients, one fundamental question to ask is: **Is a variant at a particular position in the genome deleterious?**

# Deleteriousness Prediction

## Assessing the impact of missense variants

- Given the availability of more and more sequencing data on individual patients, one fundamental question to ask is: **Is a variant at a particular position in the genome deleterious?**
- Even when restricting ourselves to missense variants that cause an amino acid change, one is usually left with tens of thousands of these variants.

# Deleteriousness Prediction

## Assessing the impact of missense variants

- Given the availability of more and more sequencing data on individual patients, one fundamental question to ask is: **Is a variant at a particular position in the genome deleterious?**
- Even when restricting ourselves to missense variants that cause an amino acid change, one is usually left with tens of thousands of these variants.
- This motivated the development of a large number of computational tools to predict the deleteriousness of missense variants.



# Deleteriousness Prediction

## Assessing the impact of missense variants

- A multitude of definitions of deleteriousness, datasets and algorithms for deleteriousness prediction (SIFT, POLYPHEN, MUTATIONTASTER, LRT, GERP, FatHMM) exists.

# Deleteriousness Prediction

## Assessing the impact of missense variants

- A multitude of definitions of deleteriousness, datasets and algorithms for deleteriousness prediction (SIFT, POLYPHEN, MUTATIONTASTER, LRT, GERP, FatHMM) exists.
- For the practitioner, it is extremely hard to choose among this plethora of approaches.

# Deleteriousness Prediction

## Assessing the impact of missense variants

- A multitude of definitions of deleteriousness, datasets and algorithms for deleteriousness prediction (SIFT, POLYPHEN, MUTATIONTASTER, LRT, GERP, FatHMM) exists.
- For the practitioner, it is extremely hard to choose among this plethora of approaches.
- **Our goal: To provide the cleanest and most comprehensive comparative evaluation of different deleteriousness predictors on a wide variety of datasets** (Grimm et al., Human Mutation 2015).

# Deleteriousness Prediction

## Assessing the impact of missense variants

- A multitude of definitions of deleteriousness, datasets and algorithms for deleteriousness prediction (SIFT, POLYPHEN, MUTATIONTASTER, LRT, GERP, FatHMM) exists.
- For the practitioner, it is extremely hard to choose among this plethora of approaches.
- **Our goal: To provide the cleanest and most comprehensive comparative evaluation of different deleteriousness predictors on a wide variety of datasets** (Grimm et al., Human Mutation 2015).
- We compared 10 methods on 5 widely used datasets.

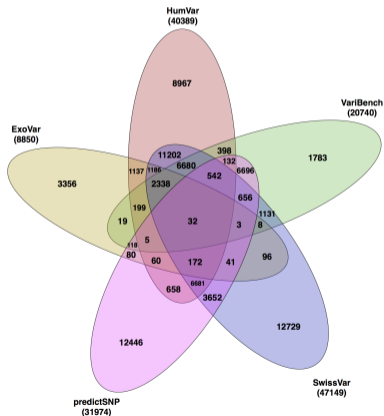
# Deleteriousness Prediction

## Two Major Types of Circularities

- **Type 1 Circularity:** The common benchmark datasets used for training and testing tools overlap to a large degree.
- **Type 2 Circularity:** Most proteins contain only deleterious or only neutral variants. A naive majority class vote within a protein gives (artificially) excellent results.

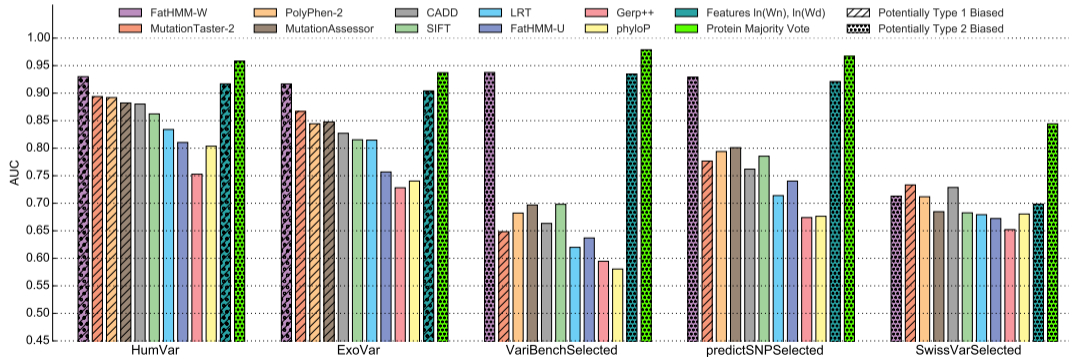
# Deleteriousness Prediction

## Type 1 Circularity



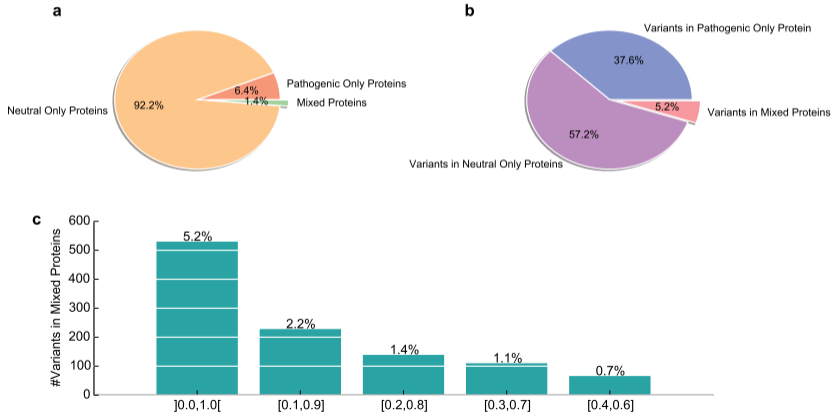
# Deleteriousness Prediction

## Comparative Evaluation



# Deleteriousness Prediction

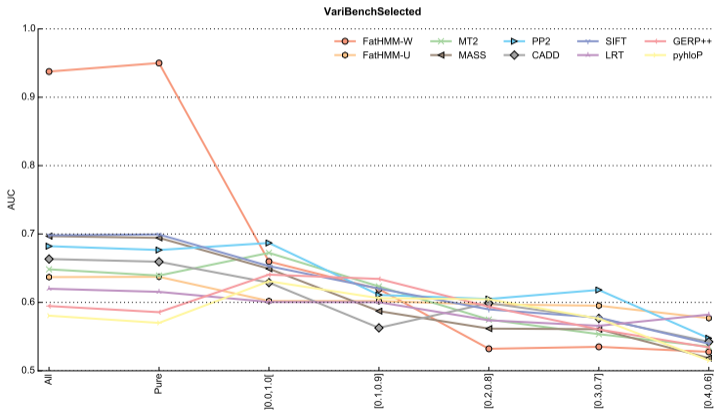
## Fraction of 'mixed' proteins in VariBenchSelected





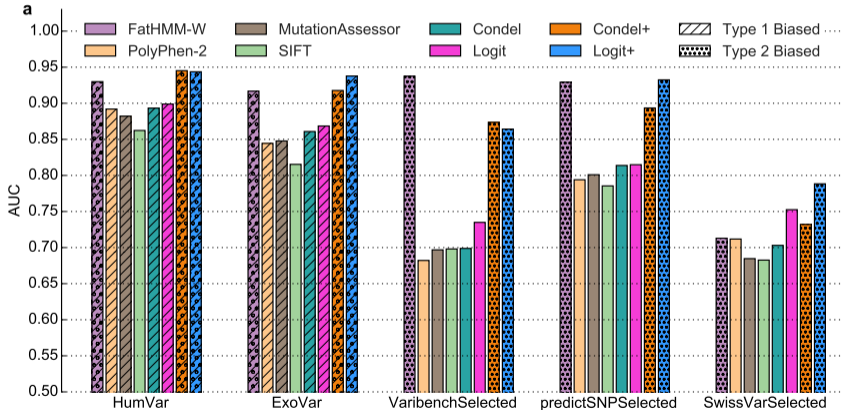
# Deleteriousness Prediction

Type 2 Circularity: Predictive performance versus neutral/deleterious ratio



# Deleteriousness Prediction

## Impact of Circularity on Combining Predictors



# Deleteriousness Prediction

## Conclusions

- A comparative evaluation of deleteriousness prediction tools is complicated by two types of circularity:

# Deleteriousness Prediction

## Conclusions

- A comparative evaluation of deleteriousness prediction tools is complicated by two types of circularity:
- Type 1 circularity can only be avoided by cleanly separating training and test dataset.

# Deleteriousness Prediction

## Conclusions

- A comparative evaluation of deleteriousness prediction tools is complicated by two types of circularity:
- Type 1 circularity can only be avoided by cleanly separating training and test dataset.
- Type 2 circularity can only be avoided by stratifying training and test dataset with respect to protein membership.

# Deleteriousness Prediction

## Conclusions

- A comparative evaluation of deleteriousness prediction tools is complicated by two types of circularity:
- Type 1 circularity can only be avoided by cleanly separating training and test dataset.
- Type 2 circularity can only be avoided by stratifying training and test dataset with respect to protein membership.
- A severe complication in practice is that many authors only publish their prediction tool, but not the features used to train the predictors. Retraining the models to ensure a clean, circularity-free prediction is practically impossible.

## Multiple Testing Problem: Biomarker Discovery

# Biomarker Discovery as a Pattern Mining Problem

## Finding groups of trait-related molecular factors

- Single genetic variants, gene expression levels, protein abundancies are often not sufficiently indicative of disease outbreak, progression or therapy outcome.



# Biomarker Discovery as a Pattern Mining Problem

## Finding groups of trait-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 (**patterns**) creates an enormous search space, and two inherent problems:

# Biomarker Discovery as a Pattern Mining Problem

## Finding groups of trait-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 (**patterns**) creates an enormous search space, and two inherent problems:
  - 1 Computational level: How to efficiently search this large space?

# Biomarker Discovery as a Pattern Mining Problem

## Finding groups of trait-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 (**patterns**) 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?

# Biomarker Discovery as a Pattern Mining Problem

## Finding groups of trait-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 (**patterns**) 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.

# Biomarker Discovery as a Pattern Mining Problem

## Finding groups of trait-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 (**patterns**) 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

| Class 1   |    |     |    |   | Class 2  |     |      |    |   |
|---|----|-----|----|---|--|-----|------|----|---|
|  |    |     |    |   |  |     |      |    |   |
| I   | II | III | IV | V | VI   | VII | VIII | IX | X |
| 1   | 0  | 1   | 0  | 0 | 0  | 1   | 0    | 1  | 1 |
| 1   | 1  | 0   | 0  | 0 | 0  | 1   | 0    | 0  | 0 |
| 0   | 0  | 0   | 1  | 0 | 1  | 1   | 0    | 0  | 0 |
| 0   | 1  | 0   | 0  | 0 | 0  | 0   | 0    | 0  | 0 |
| 1   | 1  | 1   | 1  | 1 | 0  | 0   | 0    | 0  | 0 |
| 0   | 1  | 0   | 1  | 0 | 1  | 0   | 0    | 1  | 0 |

Features

- 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

# 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

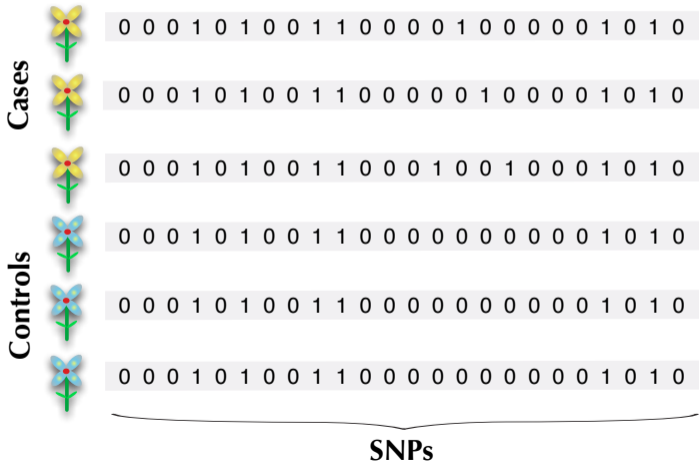
- **Tarone's insight:** When working with discrete test statistics (e.g. Fisher's exact test), there is a minimum  $p$ -value that a given pattern can obtain, based on its total frequency.
- **Tarone's trick (1990):** Ignore those patterns in multiple testing correction, for which the minimum  $p$ -value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

# 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.
- **Our work:** We showed how to efficiently find testable hypotheses in graph mining and association rule mining (Suygama et al., SDM 2015, Llinares-Lopez et al., KDD 2015).

# FAIS: Finding Intervals That Exhibit Genetic Heterogeneity

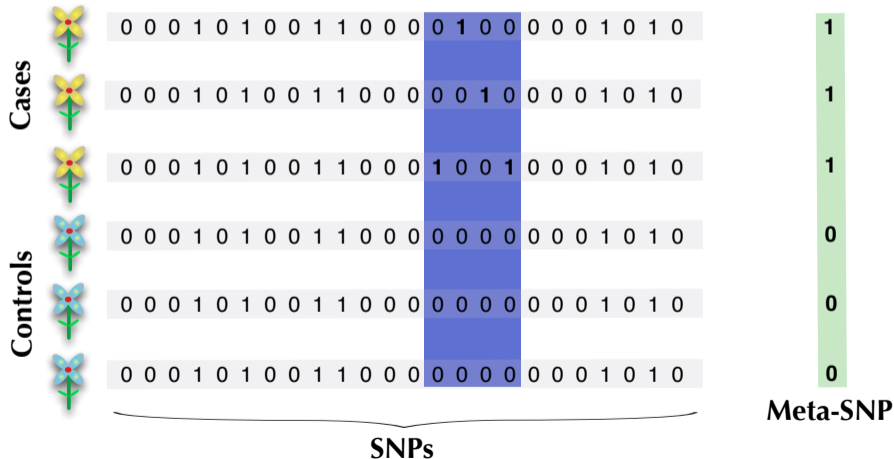


# FAIS: Finding Intervals That Exhibit Genetic Heterogeneity

## Genetic heterogeneity

- Genetic heterogeneity refers to the phenomenon that several different genes or sequence variants may give rise to the same phenotype.
- The correlation between each individual gene or variant and the phenotype may be too weak to be detected, but the group may have a strong correlation.
- The only current way to consider genetic heterogeneity is to consider fixed groups of variants. Genome-wide scans cause tremendous computational and statistical problems.

# FAIS: Finding Intervals That Exhibit Genetic Heterogeneity





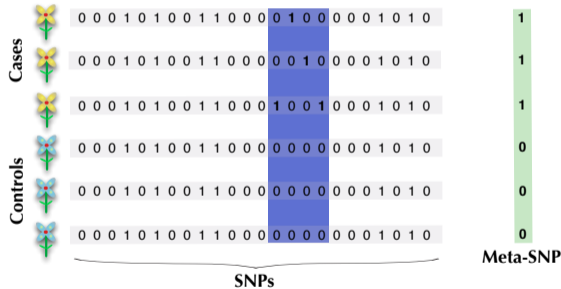
# FAIS: Finding Intervals That Exhibit Genetic Heterogeneity

## Fast Automatic Interval Search (Llinares-Lopez et al., ISMB 2015)

- Our goal is to **search for intervals that may exhibit genetic heterogeneity**, while
  - allowing for arbitrary start and end points of the intervals,
  - properly correcting for the inherent multiple testing problem, and
  - retaining statistical power and computational efficiency.
- We model the search as a **pattern mining problem**: Given an interval, an individual contains a pattern, if it has at least one minor allele in this interval.

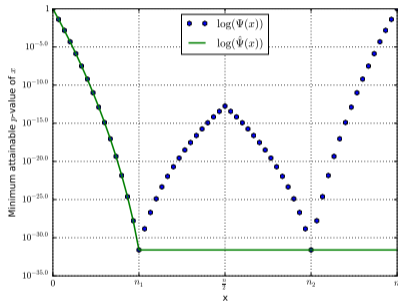
# FAIS: Finding Intervals That Exhibit Genetic Heterogeneity

Finding trait-associated genome **segments** with at least one minor allele



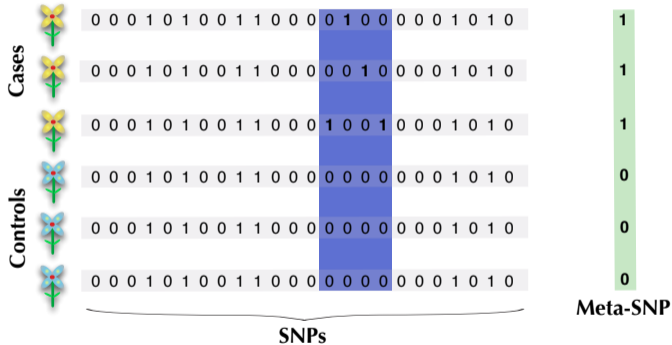
- An interval is represented by its maximum value. The longer an interval, the more likely it is that this maximum is 1.

# FAIS: Finding Intervals That Exhibit Genetic Heterogeneity



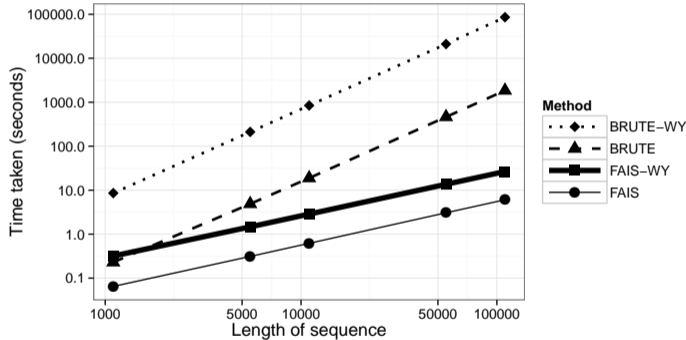
- **Pruning criterion 1:** If too many individuals have a particular pattern, the corresponding interval is not testable.

# FAIS: Finding Intervals That Exhibit Genetic Heterogeneity



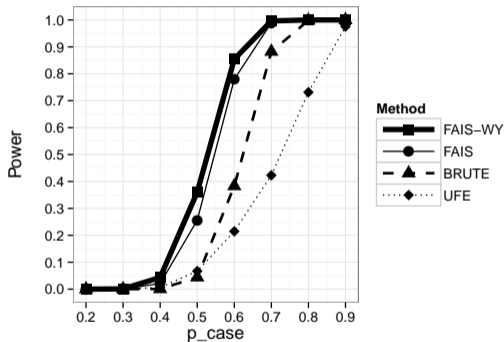
- **Pruning criterion 2:** If a pattern is too frequent to be testable, then none of the superintervals of the corresponding interval is testable.

# FAIS: Finding Intervals That Exhibit Genetic Heterogeneity



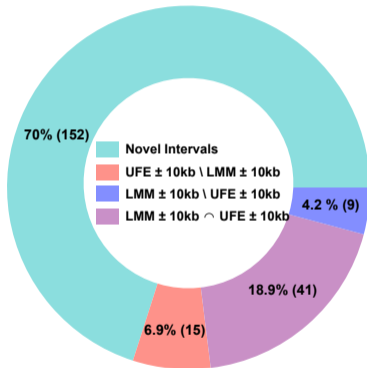
- Our method FAIS (Fast Automatic Interval Search) improves over the brute-force interval search in terms of runtime in simulations.

# FAIS: Finding Intervals That Exhibit Genetic Heterogeneity



- Our method FAIS (Fast Automatic Interval Search) improves over brute-force interval search and univariate approaches in terms of power in simulations.

# FAIS: Finding Intervals That Exhibit Genetic Heterogeneity



- Most significant intervals would have been missed by univariate approaches (UFE and LMM) on 21 binary phenotypes from *Arabidopsis thaliana* (Atwell et al., Nature 2010).

# Conclusions and Outlook

## Conclusions

- We can search for intervals that may exhibit genetic heterogeneity
  - efficiently,
  - without pre-defining the boundaries of intervals,
  - while properly correcting for multiple testing.



# Conclusions and Outlook

## Conclusions

- We can search for intervals that may exhibit genetic heterogeneity
  - efficiently,
  - without pre-defining the boundaries of intervals,
  - while properly correcting for multiple testing.

## Outlook

- Genetic heterogeneity discovery: How to extend our approach to human genetics?

# Conclusions and Outlook

## Conclusions

- We can search for intervals that may exhibit genetic heterogeneity
  - efficiently,
  - without pre-defining the boundaries of intervals,
  - while properly correcting for multiple testing.

## Outlook

- Genetic heterogeneity discovery: How to extend our approach to human genetics?
- **In General: Machine Learning and Data Mining will gain further importance in Systems Biology and Personalized Medicine.**

# Thank You

- Felipe Llinares Lopez
- Menno Witteveen
- Dean Bodenham
- Udo Gieraths
- Dominik Grimm
- Elisabetta Ghisu
- Anja Gumpinger
- Xiao He
- Laetitia Papaxanthos
- Damian Roqueiro
- Birgit Knapp





## Sponsors:

- Krupp-Stiftung
- Marie-Curie-FP 7
- SNSF Starting Grant (ERC backup)
- Horizon 2020

References: <http://www.bsse.ethz.ch/mlcb>

## References I

-  R. E. Tarone, *Biometrics* **46**, 515 (1990).
-  P. H. Westfall, S. S. Young, *Statistics in Medicine* **13**, 1084 (1993).
-  A. Terada, M. Okada-Hatakeyama, K. Tsuda, J. Sese, *Proceedings of the National Academy of Sciences* **110**, 12996 (2013).
-  D. G. Grimm, *et al.*, *Human Mutation* **36**, 513 (2015).
-  M. Sugiyama, F. Llinares-López, N. Kasenburg, K. M. Borgwardt, *SIAM Data Mining* (2015).
-  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.

## References II

-  F. Llinares-López, *et al.*, *Bioinformatics* **31**, 240 (2015).
-  S. Reardon, *Nature News* (2015). October 8, 2015.
-  E. Yong, *The Atlantic* (2015). October 10, 2015.
-  T. Ngun, *vizbang.tumblr.com* (2015). October 10, 2015.



# Mining Significant Patterns

## Fisher's exact test

### ■ Contingency Table

|         | $S = 1$ | $S = 0$           |           |
|---------|---------|-------------------|-----------|
| $y = 1$ | $a$     | $n_1 - a$         | $n_1$     |
| $y = 2$ | $x - a$ | $n - n_1 - x + a$ | $n - n_1$ |
|         | $x$     | $n - x$           | $n$       |

- 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  $(x, n_1, n)$ .

# Mining Significant Patterns

## Multiple Testing Problem

- Each  $S$  and contingency table corresponds to one hypothesis that is tested.
- To control the Family-Wise Error Rate (probability of detecting at least one false positive), we have to perform multiple testing correction.
- Without multiple testing correction, we will discover millions and billions of false positives in biomarker discovery.
- The classic approach is Bonferroni correction (1936), dividing the significance level  $\alpha$  by the number of tests  $m$ , that is,  $\frac{\alpha}{m}$ .



# 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*  $\bar{S}$ .
- **Tarone's strategy:** Ignore untestable hypotheses  $\bar{S}$  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  $\bar{S}$ .

# 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

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

# 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";

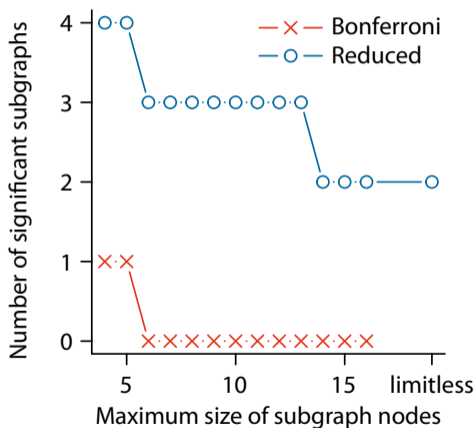
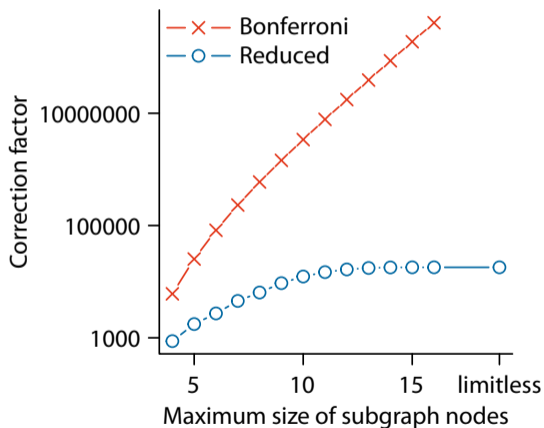
**while**  $k > m(k)$  **do**

$k := k - 1$ ;

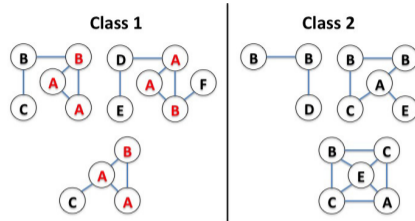
$m(k) :=$  frequent itemset mining( $D, \psi^{-1}(\frac{\alpha}{k})$ );

**return**  $k + 1$

## 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  $\theta$

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

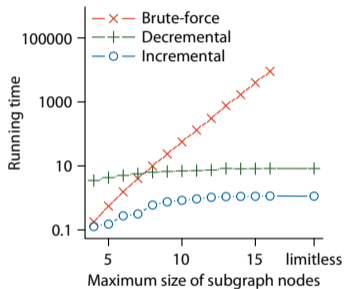
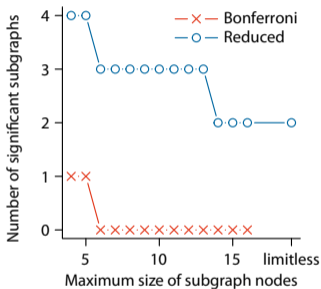
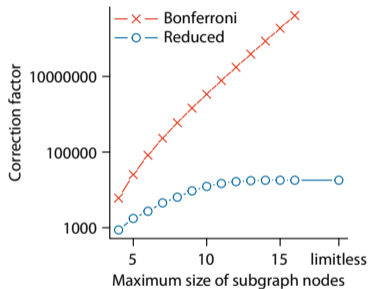
**until** (no more frequent subgraph found) or  $(FS_{\theta} > \frac{\alpha}{\psi(\theta)})$

**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  $\psi(\theta)$ .

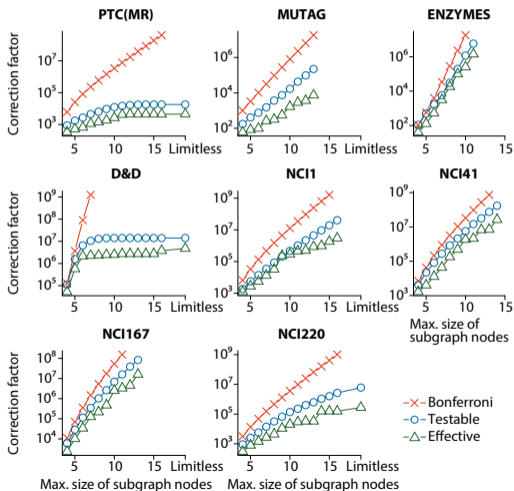
# Significant Subgraph Mining on PTC Dataset



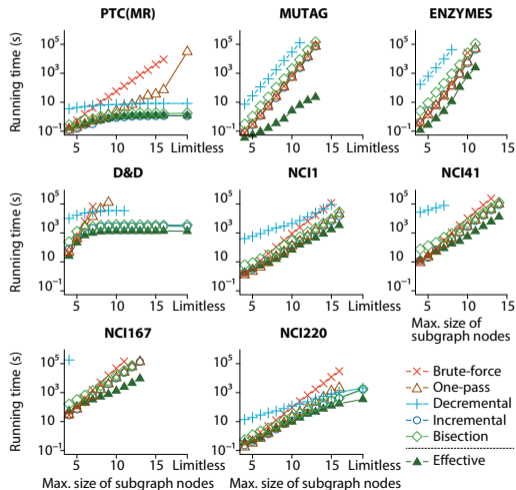
Dataset from Helma et al. (2001)



# 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  $\delta^*$ .

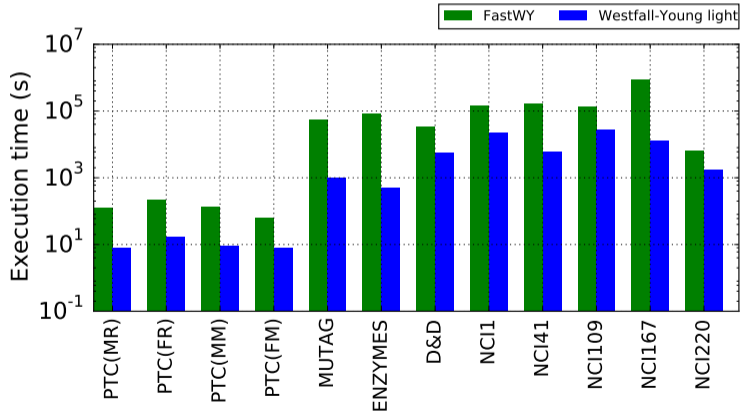
## Westfall-Young light (Llinares-Lopez et al., KDD 2015)

### 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

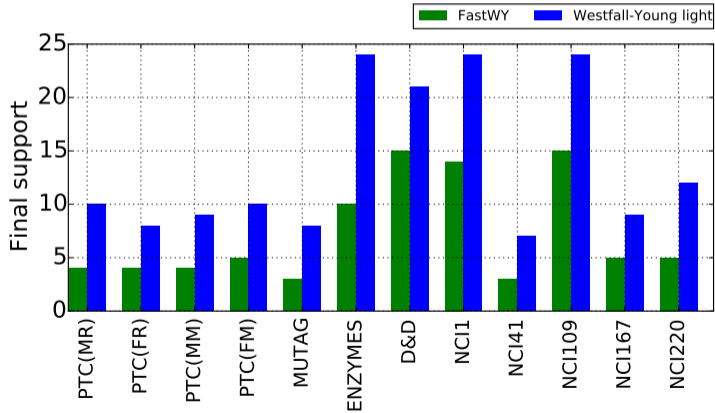
## Westfall-Young light

## ■ Runtime



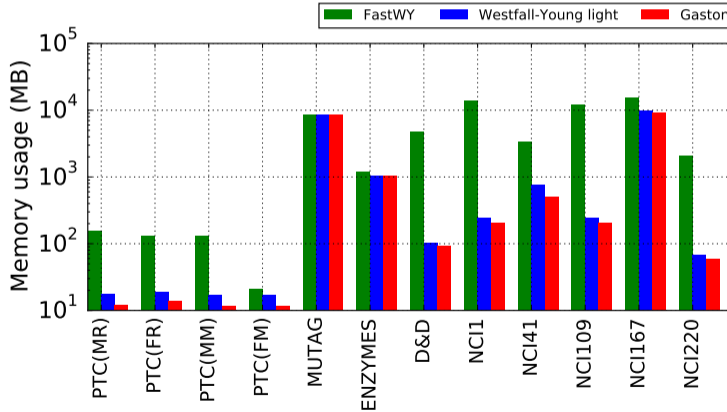
# Westfall-Young light

- Final frequency threshold (support)



## Westfall-Young light

## ■ Peak memory usage





## Westfall-Young light

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

