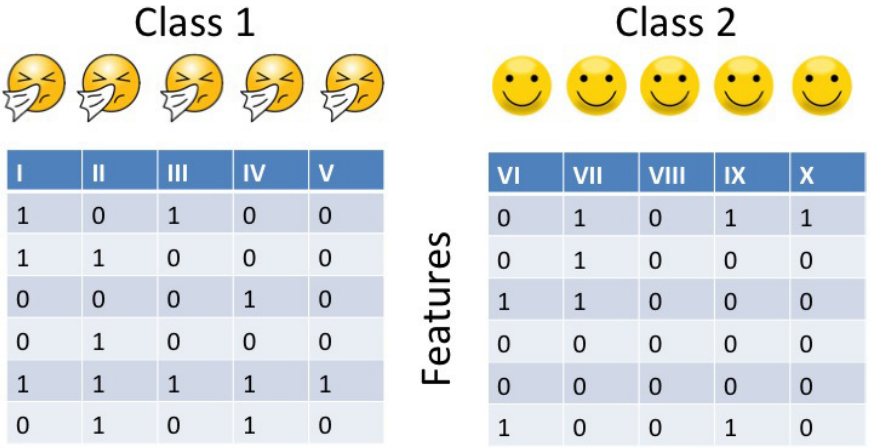
Significant Pattern Mining

Karsten Borgwardt

ETH Zürich

TU Dortmund, November 12, 2015

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

Features

Class 2

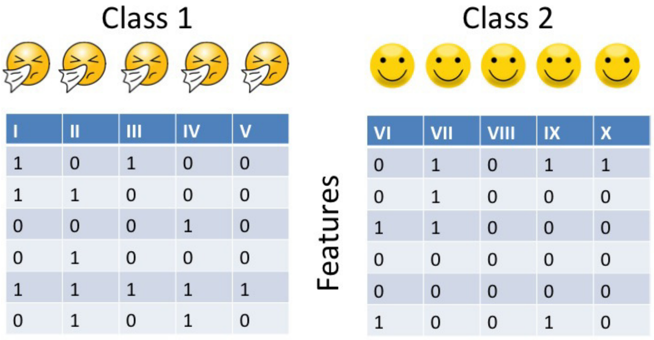
VI	VII	VIII	IX	X
0	1	0	1	1
0	1	0	0	0
1	1	0	0	0
0	0	0	0	0
0	0	0	0	0
1	0	0	1	0

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



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

Features

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

- 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

Fisher's exact test

- Contingency Table

	$S = 1$	$S = 0$	
$\mathbf{y} = 1$	a	$n_1 - a$	n_1
$\mathbf{y} = 2$	$x - a$	$n - n_1 - x + a$	$n - n_1$
	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).

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, α 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, α per cent of all candidate patterns will be false positives.
- If we do correct for multiple testing, e.g. via Bonferroni correction ($\frac{\alpha}{\#\mathit{tests}}$), 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* $\bar{\mathcal{U}}$.
- **Tarone's strategy**: Ignore untestable hypotheses $\bar{\mathcal{U}}$ 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* $\bar{\mathcal{U}}$.
- **Tarone's strategy**: Ignore untestable hypotheses $\bar{\mathcal{U}}$ 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* $\bar{\mathcal{U}}$.
- **Tarone's strategy**: Ignore untestable hypotheses $\bar{\mathcal{U}}$ 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 $\bar{\mathcal{U}}$.

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 $k$
```

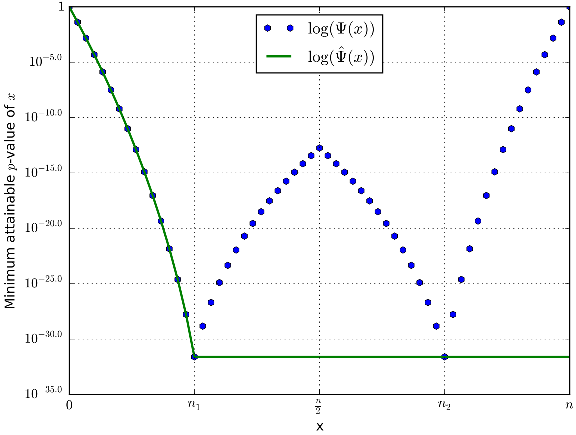
Mining Significant Patterns

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

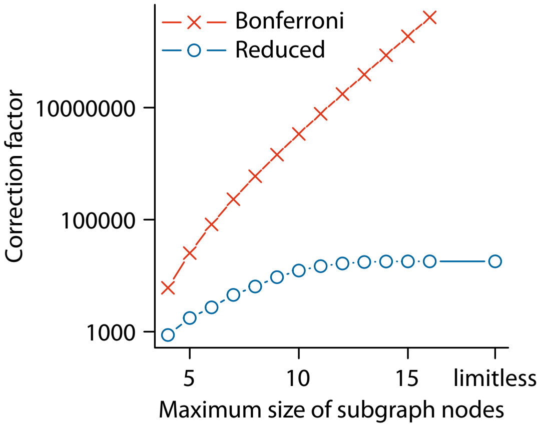
- For $0 \leq x \leq n_1$, the minimum p-value $p_{min}(S)$ decreases monotonically with x.
- One can use *frequent itemset mining* to find all S that are testable at level α, with frequency $\psi^{-1}(\alpha)$.
- They propose to use a decremental search strategy:

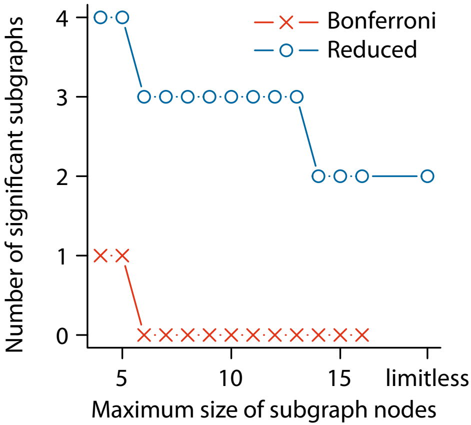
```
procedure TERADA'S DECREMENTAL SEARCH (LAMP)
    $k :=$ "very large";
    while $k > m(k)$ do
        $k := k - 1$;
        $m(k) :=$ frequent itemset mining$(D, \psi^{-1}(\frac{\alpha}{k}))$;
    return $k + 1$
```

Mining Significant Patterns

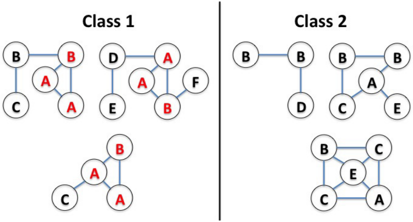


Example: PTC dataset (Helma et al., 2001)





Significant Subgraph Mining (Sugyiama 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.

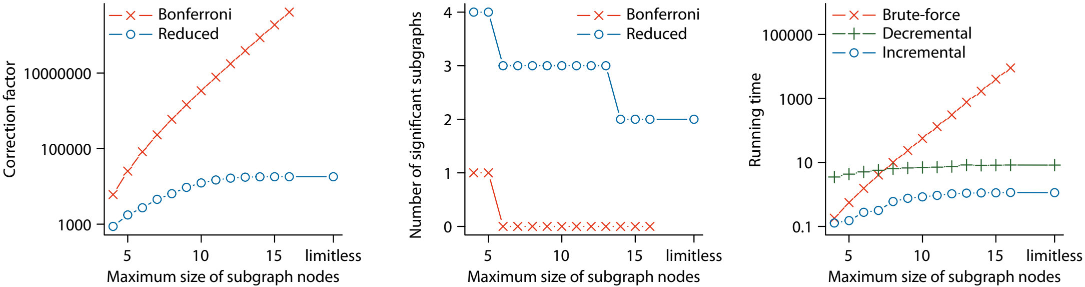
Significant Subgraph Mining (Sugyiama et al., SDM 2015)

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 $\psi(\theta)$
```

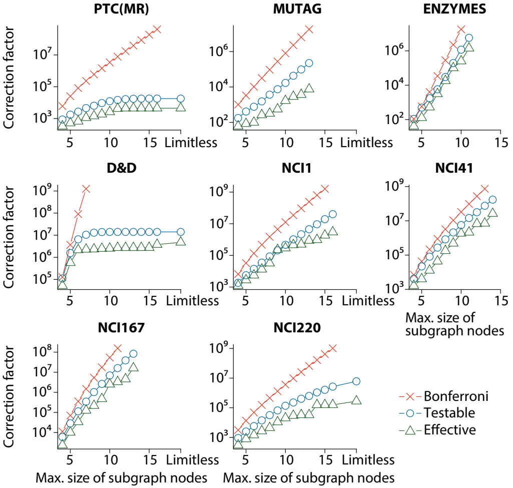
- $\frac{\alpha}{\psi(\theta)}$ is the maximum correction factor, such that subgraphs with frequency θ can be significant at level $\psi(\theta)$.

Significant Subgraph Mining on PTC Dataset

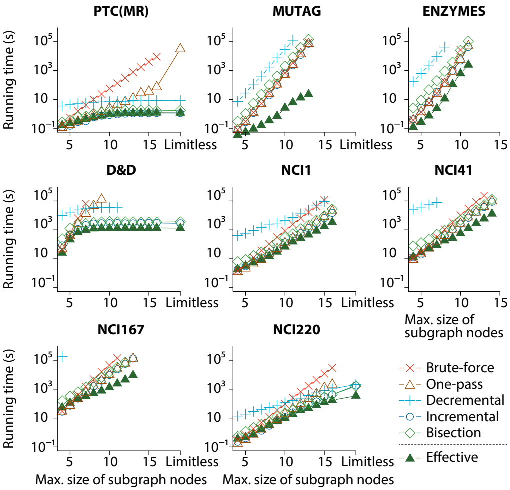


Dataset from Helma et al. (2001)

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 α, 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 δ^* by α-quantile of $p_{min}^{(j)}$, and increase θ accordingly.
   - Process all children of S with frequency $\geq \psi^{-1}(\delta^*)$.
5. **Output**: Corrected significance threshold δ^*.

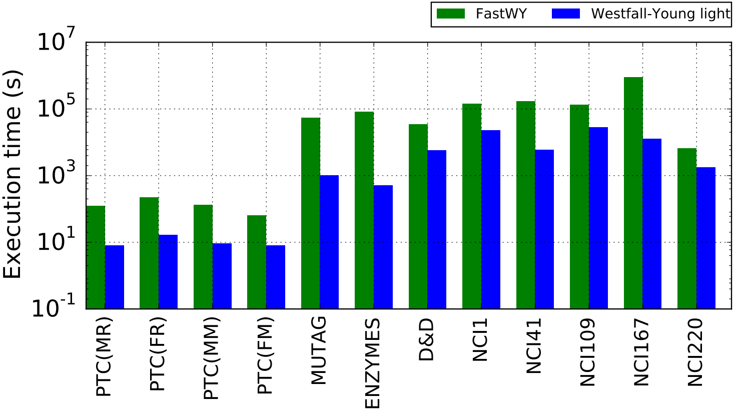
Westfall-Young light (Llinares-Lopez et al., KDD 2015)

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

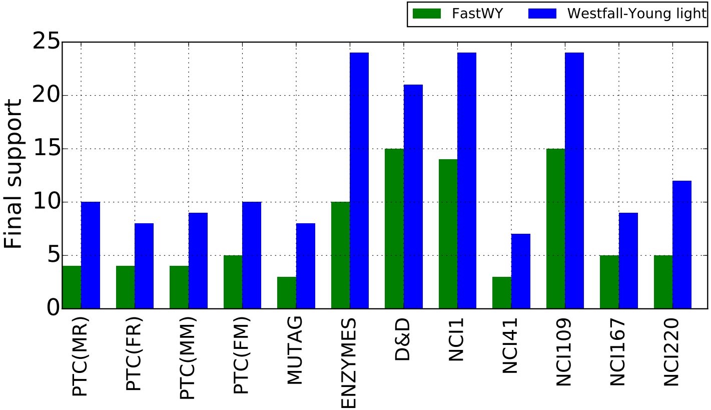
Westfall-Young light

- Runtime



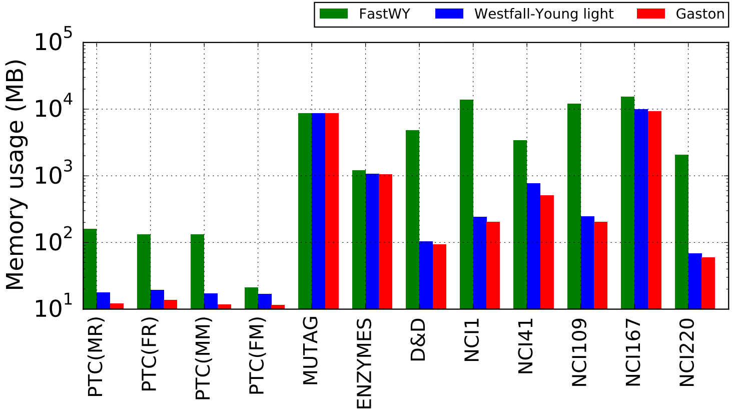
Westfall-Young light

- Final frequency threshold (support)



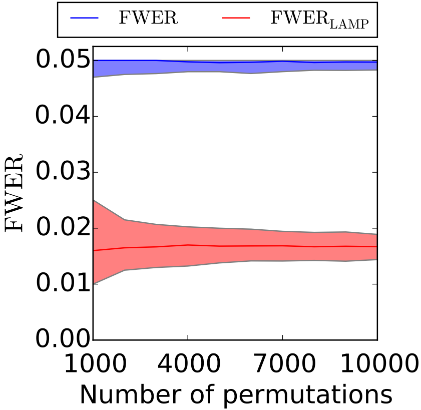
Westfall-Young light

- Peak memory usage

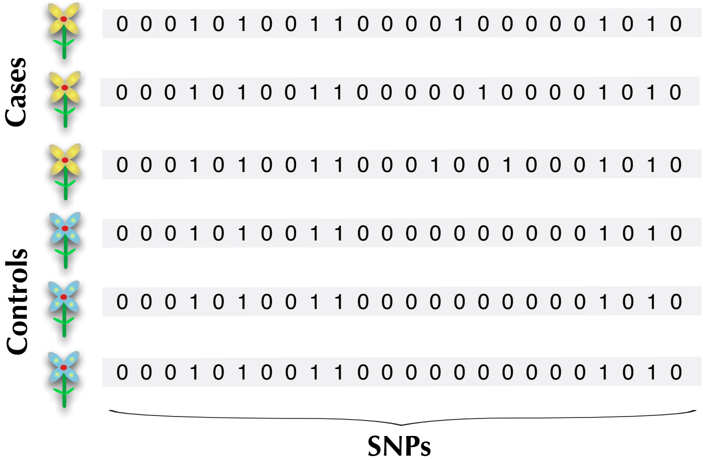


Westfall-Young light

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



FAIS: Finding Intervals That Exhibit Genetic Heterogeneity



Cases

0 0 0 1 0 1 0 0 1 1 0 0 0 0 1 0 0 0 0 0 1 0 1 0

0 0 0 1 0 1 0 0 1 1 0 0 0 0 0 1 0 0 0 0 1 0 1 0

0 0 0 1 0 1 0 0 1 1 0 0 0 1 0 0 1 0 0 0 1 0 1 0

Controls

0 0 0 1 0 1 0 0 1 1 0 0 0 0 0 0 0 0 0 0 1 0 1 0

0 0 0 1 0 1 0 0 1 1 0 0 0 0 0 0 0 0 0 0 1 0 1 0

0 0 0 1 0 1 0 0 1 1 0 0 0 0 0 0 0 0 0 0 1 0 1 0

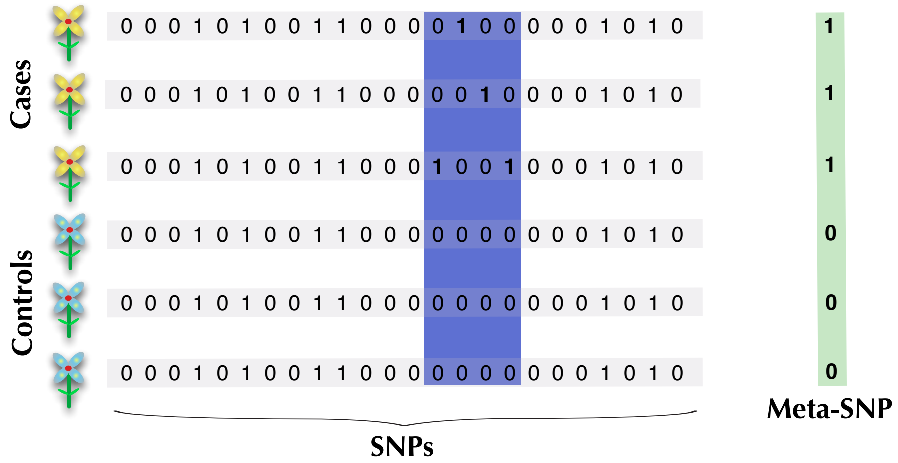
SNPs

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 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



Cases

0 0 0 1 0 1 0 0 1 1 0 0 0 0 **1** 0 0 0 0 0 1 0 1 0 **1**

0 0 0 1 0 1 0 0 1 1 0 0 0 0 0 **1** 0 0 0 0 1 0 1 0 **1**

0 0 0 1 0 1 0 0 1 1 0 0 0 **1** 0 0 **1** 0 0 0 1 0 1 0 **1**

Controls

0 0 0 1 0 1 0 0 1 1 0 0 0 0 0 0 0 0 0 0 1 0 1 0 **0**

0 0 0 1 0 1 0 0 1 1 0 0 0 0 0 0 0 0 0 0 1 0 1 0 **0**

0 0 0 1 0 1 0 0 1 1 0 0 0 0 0 0 0 0 0 0 1 0 1 0 **0**

SNPs

Meta-SNP

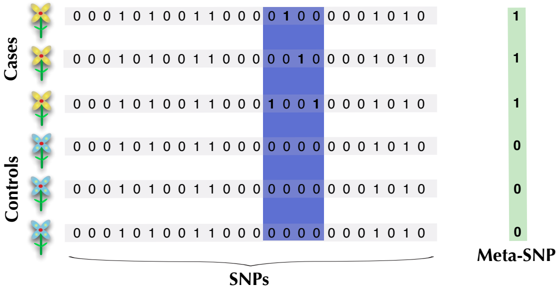
FAIS: Finding Intervals That Exhibit Genetic Heterogeneity

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

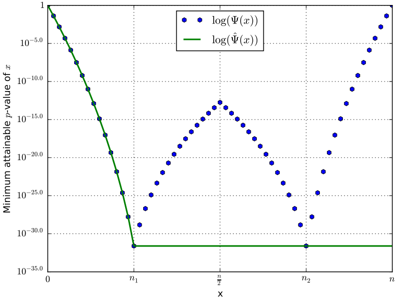
Finding trait-associated genome **segments** with at least one minor allele



	SNPs	Meta-SNP
Cases	0 0 0 1 0 1 0 0 1 1 0 0 0 0 1 0 0 0 0 0 1 0 1 0	1
	0 0 0 1 0 1 0 0 1 1 0 0 0 0 0 1 0 0 0 0 1 0 1 0	1
	0 0 0 1 0 1 0 0 1 1 0 0 0 1 0 0 1 0 0 0 1 0 1 0	1
Controls	0 0 0 1 0 1 0 0 1 1 0 0 0 0 0 0 0 0 0 0 1 0 1 0	0
	0 0 0 1 0 1 0 0 1 1 0 0 0 0 0 0 0 0 0 0 1 0 1 0	0
	0 0 0 1 0 1 0 0 1 1 0 0 0 0 0 0 0 0 0 0 1 0 1 0	0

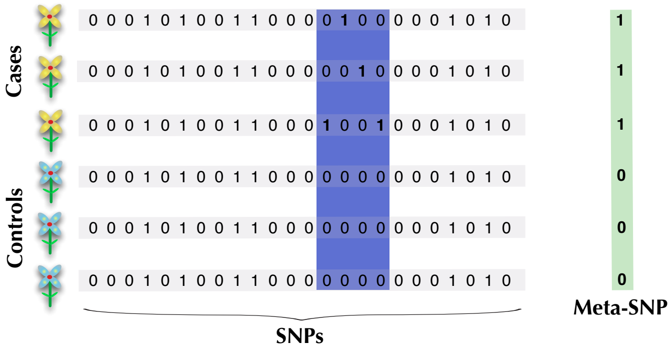
- An interval is represented by its maximum value. The longer an interval, the more likely it is that this maximum is 1.

FAIS: Finding Intervals That Exhibit Genetic Heterogeneity



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

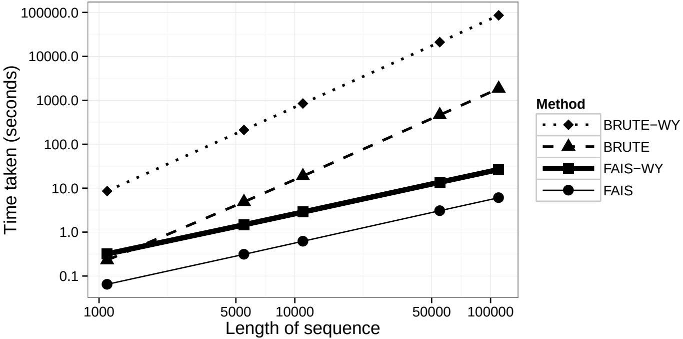
FAIS: Finding Intervals That Exhibit Genetic Heterogeneity



	SNPs	Meta-SNP
Cases	0 0 0 1 0 1 0 0 1 1 0 0 0 0 1 0 0 0 0 0 1 0 1 0	1
Cases	0 0 0 1 0 1 0 0 1 1 0 0 0 0 0 1 0 0 0 0 1 0 1 0	1
Cases	0 0 0 1 0 1 0 0 1 1 0 0 0 1 0 0 1 0 0 0 1 0 1 0	1
Controls	0 0 0 1 0 1 0 0 1 1 0 0 0 0 0 0 0 0 0 0 1 0 1 0	0
Controls	0 0 0 1 0 1 0 0 1 1 0 0 0 0 0 0 0 0 0 0 1 0 1 0	0
Controls	0 0 0 1 0 1 0 0 1 1 0 0 0 0 0 0 0 0 0 0 1 0 1 0	0

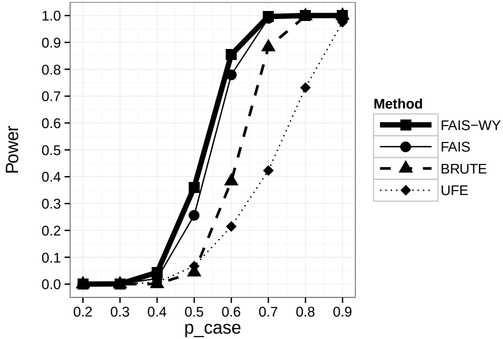
- **Pruning criterion 2:** If a pattern is too frequent to be testable, then none of the superintervals of the corresponding interval is testable.

FAIS: Finding Intervals That Exhibit Genetic Heterogeneity



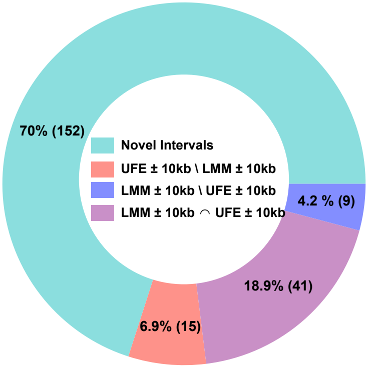
- Our method FAIS (Fast Automatic Interval Search) improves over the brute-force interval search in terms of runtime in simulations.

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).

FAIS: 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.

FAIS: 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?

FAIS: 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.

Significant Pattern Mining: Summary & Outlook

Summary

- We have shown how to enable significant pattern mining
  - in subgraph mining,
  - in association rule mining while taking dependence into account,
  - in interval-based genome-wide association mapping.

Outlook

- Pattern summarization
- Conditioning on covariates (Llinares-Lopez et al., arxiv 2015)
- Network-based genome-wide association mapping

Also of Interest...

...may be our latest work on graph kernels (Sugiyama & Borgwardt, NIPS 2015).

- We show that it is better to use a fixed-length random walk kernel

$$k_{fixed}(G, G') = \sum_{i,j=1}^{|V_\times|} [\sum_{k=0}^{l} A_\times^k]_{ij}$$

than a geometric random walk kernel

$$k_\times(G, G') = \sum_{i,j=1}^{|V_\times|} [\sum_{k=0}^{\infty} \lambda^k A_\times^k]_{ij} = \mathbf{e}^\top(\mathbf{1} - \lambda A_\times)^{-1}\mathbf{e}$$

as a baseline in comparative evaluations of graph kernels.

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

Machine
Learning
for
Personalized
Medicine

Sponsors:

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

References: `http://www.bsse.ethz.ch/mlcb`

References

- R. E. Tarone, *Biometrics* **46**, 515 (1990).
- P. H. Westfall, S. S. Young, *Statistics in Medicine* **13**, 1084 (1993).
- D. R. Nyholt, *American Journal of Human Genetics* **74**, 765 (2004).
- 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.
- M. Sugiyama, F. Llinares-López, N. Kasenburg, K. M. Borgwardt, *SIAM Data Mining* (2015).

References II

- 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.

- F. Llinares-López, *et al.*, *Bioinformatics* **31**, 240 (2015).