Machine Learning for Personalized Medicine

Karsten Borgwardt

ETH Zürich, D-BSSE

IMM, Universität Zürich, October 25, 2018

Machine Learning and Personalized Medicine

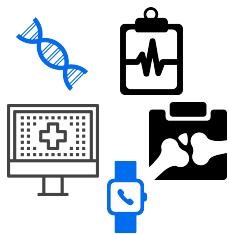
Goals

- Machine Learning tries to detect statistical dependencies in large datasets.

Machine Learning and Personalized Medicine

Goals

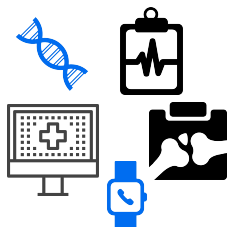
- Machine Learning tries to detect statistical dependencies in large datasets.



Machine Learning and Personalized Medicine

Goals

- Machine Learning tries to detect statistical dependencies in large datasets.



- Personalized Medicine tries to exploit wealth of health data for improved diagnosis, prognosis and therapy decisions, tailored to the properties of each patient.

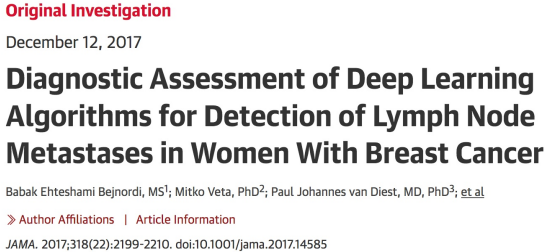
Machine Learning in Medicine

Key Topics

Machine Learning in Medicine

Key Topics

- Automation of diagnoses



Original Investigation

December 12, 2017

Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer

Babak Ehteshami Bejnordi, MS[1]; Mitko Veta, PhD[2]; Paul Johannes van Diest, MD, PhD[3]; et al

» Author Affiliations | Article Information

JAMA. 2017;318(22):2199-2210. doi:10.1001/jama.2017.14585

Machine Learning in Medicine

Key Topics

- Automation of diagnoses
- Biomarker discovery

Machine learning of neural representations of suicide and emotion concepts identifies suicidal youth

Marcel Adam Just, Lisa Pan, Vladimir L. Cherkassky, Dana L. McMakin, Christine Cha, Matthew K. Nock & David Brent

Nature Human Behaviour **1**, 911–919 (2017)
doi:10.1038/s41562-017-0234-y
Download Citation

Received: 06 February 2017
Accepted: 04 October 2017
Published online: 30 October 2017

Machine Learning in Medicine

Key Topics

- Automation of diagnoses
- Biomarker discovery

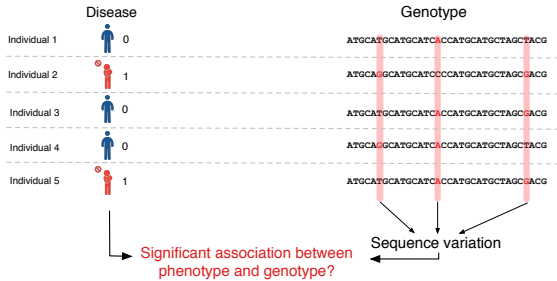
Roche to **buy Flatiron** Health for $1.9 billion to expand cancer care ...
Reuters - 15.02.2018
Roche to **buy Flatiron** Health for $1.9 billion to expand cancer care portfolio ... S) said on Thursday it would **buy** the rest of U.S. cancer data company **Flatiron** Health for $1.9 billion to speed development of cancer medicines and support its efforts to ... Privately held **Flatiron**, backed by Alphabet Inc (GOOGL.

- Biomedical data management

Machine Learning in Medicine

Key Topics

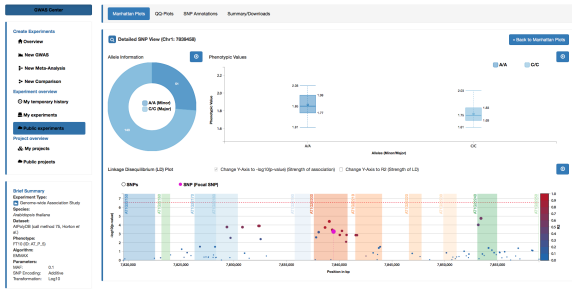
- Automation of diagnoses
- Biomarker discovery
  1. A new technique: Combinatorial Association Mapping
- Biomedical data management



Machine Learning in Medicine

Key Topics

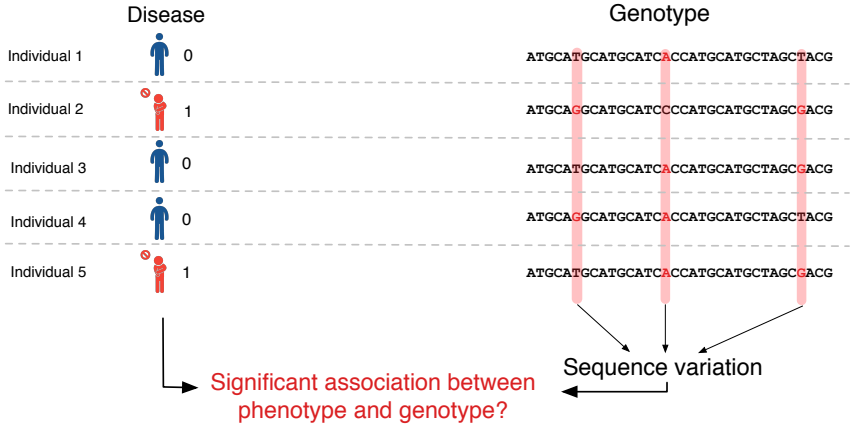
- Automation of diagnoses



- Biomarker discovery
  1. A new technique: Combinatorial Association Mapping
- Biomedical data management
  2. Software development for the Life Sciences: easyGWAS

Combinatorial Association Mapping

Association Mapping: Mapping Phenotypes to the Genome



A genome-wide association study (GWAS) examines whether variation in the genome (in form of single nucleotide polymorphisms, SNPs) correlates with variation in the phenotype.

Association Mapping: Missing Heritability

- Since 2001: More than 70,963 trait-related loci from GWAS (GWAS catalog, September 19, 2018)
- Problem: Phenotypic variance explained still disappointingly low

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

REVIEWS

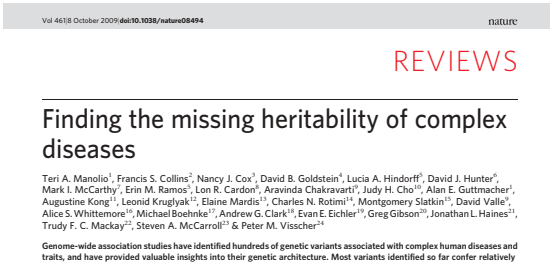
Finding the missing heritability of complex diseases

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

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

Association Mapping: Missing Heritability

- Since 2001: More than 70,963 trait-related loci from GWAS (GWAS catalog, September 19, 2018)
- Problem: Phenotypic variance explained still disappointingly low



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

REVIEWS

Finding the missing heritability of complex diseases

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

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

- Potential reasons:
  - Polygenic architectures of complex diseases
  - Small effect sizes
  - Incomplete integration of important genetic, epigenetic or non-genetic properties

Association Mapping: Missing Heritability

- Since 2001: More than 70,963 trait-related loci from GWAS (GWAS catalog, September 19, 2018)
- Problem: Phenotypic variance explained still disappointingly low

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

REVIEWS

Finding the missing heritability of complex diseases

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

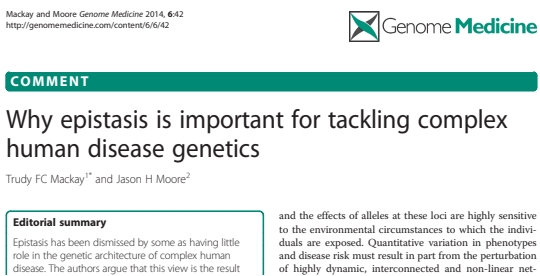
Genome-wide association studies have identified hundreds of genetic variants associated with complex human diseases and traits, and have provided valuable insights into their genetic architecture. Most variants identified so far confer relatively

- Potential reasons:
  - Polygenic architectures of complex diseases → Epistasis
  - Small effect sizes
  - Incomplete integration of important genetic, epigenetic or non-genetic properties

Association Mapping: Missing Heritability

Epistasis as a Potential Reason

- Most current analyses neglect interactive effects between loci
- Need for approaches for **combinatorial association mapping**



Mackay and Moore *Genome Medicine* 2014, **6**:42
http://genomemedicine.com/content/6/6/42

Genome Medicine

COMMENT

Why epistasis is important for tackling complex human disease genetics

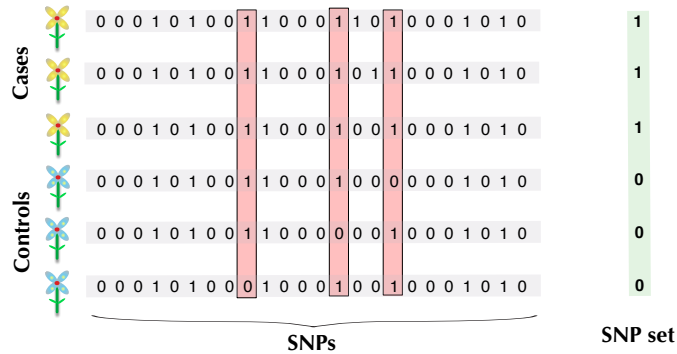
Trudy FC Mackay[1*] and Jason H Moore[2]

Editorial summary

Epistasis has been dismissed by some as having little role in the genetic architecture of complex human disease. The authors argue that this view is the result

and the effects of alleles at these loci are highly sensitive to the environmental circumstances to which the individuals are exposed. Quantitative variation in phenotypes and disease risk must result in part from the perturbation of highly dynamic, interconnected and non-linear net-

Combinatorial Association Mapping



- **Computational challenge**: Combinatorial explosion of the number of candidate sets
- **Statistical challenge**: Combinatorial explosion of the number of association tests
- Concrete example: Even for pairs of 10^6 features, order 10^{12} hypotheses!

Combinatorial Association Mapping: Statistics vs. Data Mining

Different Views on pairwise Interactions

- Statistical Approach:

$$y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_{1,2} x_1 x_2 + \epsilon;$$

  where y is a phenotype, x_1 and x_2 are SNPs, represented as integers, ϵ is random noise.

- Data Mining Approach:

$$r(y, x_1 x_2)$$

  where x_1 and x_2 are assumed to be binary, r is measure of statistical dependence.

Relative Advantages

- Pro Statistics: Considers marginal effects, not limited to binary variables.
- Pro Data Mining: Scaling to arbitrary-order interactions is easier.

Combinatorial Association Mapping

Multiple Hypothesis Testing Problem

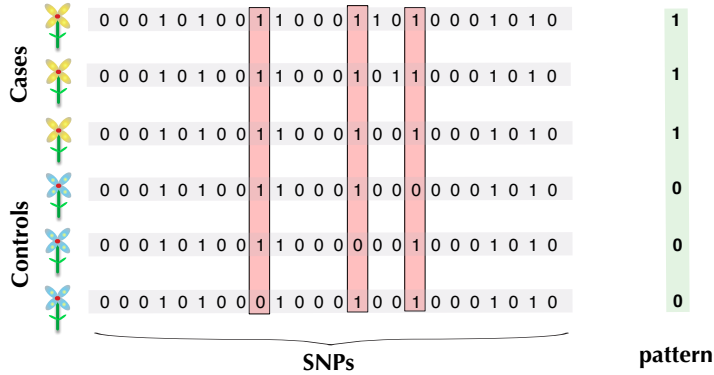
- What if we consider associations of groups of s SNPs with the phenotype?
- This leads to an enormous multiple testing problem: Any of the k SNP sets would correspond to a hypothesis that is tested ($k \in O(f^s)$), where f is the number of SNPs.
- If unaccounted for, α per cent of all SNP sets might be considered significantly associated by random chance.
- It is imperative to control for multiple testing, e.g. the family-wise error rate!
- If accounted for, e.g. by Bonferroni correction ($\frac{\alpha}{k}$), we might lose all statistical power.

Combinatorial Association Mapping

Multiple Hypothesis Testing Problem

- What if we consider associations of groups of s SNPs with the phenotype?
- This leads to an enormous multiple testing problem: Any of the k SNP sets would correspond to a hypothesis that is tested ($k \in O(f^s)$), where f is the number of SNPs.
- If unaccounted for, α per cent of all SNP sets might be considered significantly associated by random chance.
- It is imperative to control for multiple testing, e.g. the family-wise error rate!
- If accounted for, e.g. by Bonferroni correction ($\frac{\alpha}{k}$), we might lose all statistical power.
- **Long considered unsolvable dilemma**

Combinatorial Association Mapping as a Data Mining Problem

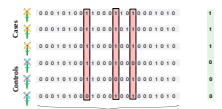


- Feature Selection: Find features that distinguish classes of objects
- Pattern Mining: Find higher-order **combinations of binary features**, so-called *patterns*, to distinguish one class from another

Combinatorial Association Mapping as a Data Mining Problem

Pattern

- D is a dataset of n patients. The i-th patient is represented by a binary vector $\mathbf{d}^{(i)} \in \{0,1\}^f$ and a class label $y_i \in \{0,1\}$.
- We choose a subset $\mathcal{S}$ of all features $\mathcal{F}$ in a dataset: $\mathcal{S} \subseteq \mathcal{F}$.
- Then an object $\mathbf{d}^{(i)}$ includes the pattern $\mathcal{S}$ if $\prod_{t \in \mathcal{S}} d^{(i)}(t) = 1$, otherwise not.



Problem Statement: Significant Pattern Mining

- We want to find all subsets $\mathcal{S}$ such that there is a statistically significant association between $\prod_{t \in \mathcal{S}} d^{(i)}(t)$ and y_i for $i \in \{1, \ldots, n\}$, while controlling the family-wise error rate at level α.

Significant Pattern Mining

Tarone's trick

- Contingency table for testing enrichment of a pattern in one of two classes

	Pattern present	Pattern absent	
y=0	a	$n_1 - a$	n_1
y=1	$x - a$	$n - n_1 - x + a$	$n - n_1$
	x	$n - x$	n

- A popular choice is Fisher's exact test to test whether the pattern 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).

Significant Pattern Mining

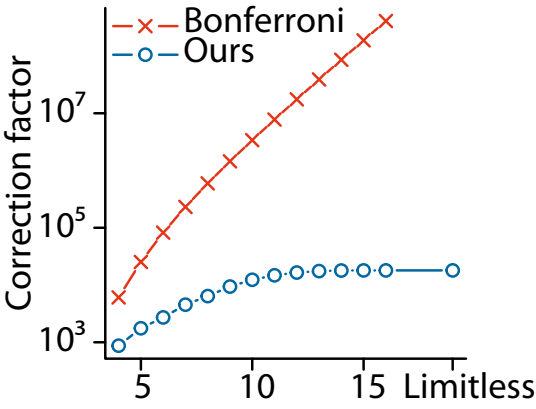
Tarone's trick

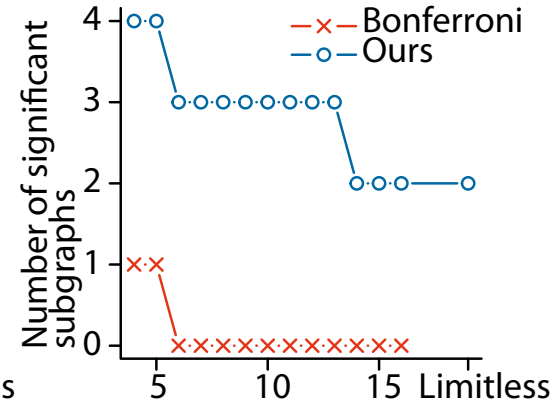
- Contingency table for testing enrichment of a pattern in a class

	Pattern present	Pattern absent	
y=0	a	$n_1 - a$	n_1
y=1	$x - a$	$n - n_1 - x + a$	$n - n_1$
	x	$n - x$	n

- Tarone (1990) noted that when working with discrete test statistics, e.g. Fisher's exact test, there is a minimum p-value that a pattern can achieve.
- There are many untestable hypotheses whose minimum p-value is not smaller than $\frac{\alpha}{k}$.
- Only the remaining $m(k)$ testable hypotheses can reach significance at all.
- One can correct for $m(k)$ instead of k. As often $m(k) << k$, this greatly improves statistical power.

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





Significant Pattern Mining

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 $\frac{\alpha}{k}$.
- $m(k)$ is a function of k and we require $k \geq m(k)$ to correct for all testable hypotheses.
- Then the optimization problem is

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

Significant Pattern Mining

Tarone's approach (1990)

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

procedure `Tarone`
  $k := 1$;
  while $k < m(k)$ **do**
    $k := k + 1$;
  return k

Significant Pattern Mining

Tarone's approach (1990)

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

```
procedure Tarone
    k := 1;
    while k < m(k) do
        k := k + 1;
    return k
```

- How to efficiently compute $m(k)$ without running through all $O(f^s)$ possible hypotheses?

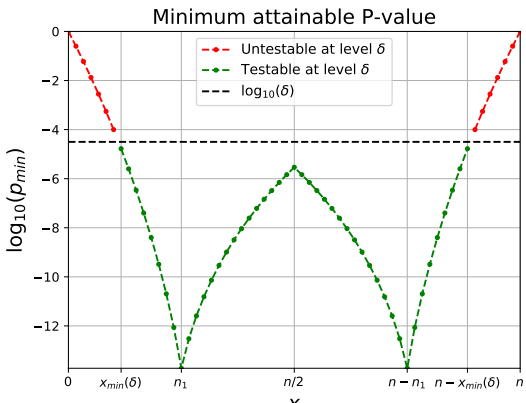
Significant Pattern Mining

Data mining challenge

- How to efficiently find $m(k)$ without running through all $O(f^s)$ possible hypotheses?
- Solution: Minimum p-value is determined by the frequency of a pattern.
- One can use frequent pattern mining algorithms from Data Mining to enumerate all patterns that pass a certain p-value threshold (Terada et al., PNAS 2013):
  - `frequent itemset mining`(D, θ) enumerates all patterns in a dataset D of frequency at least θ.

Significant Pattern Mining

- Frequency versus minimum p-value



Significant Pattern Mining

Tarone's approach with frequent itemset mining

- 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`(D, α)
  $k := 1$;
  while $k < m(k)$ **do**
    $k := k + 1$;
    $m(k) :=$ frequent itemset mining$(D, \phi(\frac{\alpha}{k}))$;
  return k

Significant Pattern Mining

Tarone's approach with frequent itemset mining

- 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`(D, α)
  $k := 1$;
  while $k < m(k)$ **do**
    $k := k + 1$;
    $m(k) :=$ frequent itemset mining$(D, \phi(\frac{\alpha}{k}))$;
  return k

- Note: $\phi(\frac{\alpha}{k})$ is the minimum frequency of a pattern that is testable at level $\frac{\alpha}{k}$.

Significant Pattern Mining

Tarone's approach with frequent itemset mining

- 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`(D, α)
  $k := 1$;
  while $k < m(k)$ **do**
    $k := k + 1$;
    $m(k) :=$ frequent itemset mining$(D, \phi(\frac{\alpha}{k}))$;
  return k

- Note: $\phi(\frac{\alpha}{k})$ is the minimum frequency of a pattern that is testable at level $\frac{\alpha}{k}$.
- For small k, $\phi(\frac{\alpha}{k})$ is small. Frequent itemset mining will be extremely expensive!

From Significant Pattern Mining to Combinatorial Association Mapping

Questions unanswered in 2014

1. How to efficiently find the optimal k? (SDM 2015)

From Significant Pattern Mining to Combinatorial Association Mapping

Questions unanswered in 2014

1. How to efficiently find the optimal k? (SDM 2015)
   - We proposed an efficient search strategy with early termination criterion (when $m(k) > k$).

From Significant Pattern Mining to Combinatorial Association Mapping

Questions unanswered in 2014

1. How to efficiently find the optimal k? (SDM 2015)
   - We proposed an efficient search strategy with early termination criterion (when $m(k) > k$).
2. Patterns are in subset/superset relationships. How to account for this dependence between tests? (KDD 2015)

From Significant Pattern Mining to Combinatorial Association Mapping

Questions unanswered in 2014

1. How to efficiently find the optimal k? (SDM 2015)
   - We proposed an efficient search strategy with early termination criterion (when $m(k) > k$).
2. Patterns are in subset/superset relationships. How to account for this dependence between tests? (KDD 2015)
   - We perform Westfall-Young Permutations to take the dependence into account.
   - By dynamically updating the frequency threshold, we only require 1 single application of frequent itemset mining even for 10,000 permutations.

From Significant Pattern Mining to Combinatorial Association Mapping

Questions unanswered in 2014

3. Can we retain efficiency and statistical power when accounting for categorical covariates such as age and gender? (NIPS 2016)

From Significant Pattern Mining to Combinatorial Association Mapping

Questions unanswered in 2014

3. Can we retain efficiency and statistical power when accounting for categorical covariates such as age and gender? (NIPS 2016)
   - We extended Tarone's trick to the Cochran-Mantel-Haenszel-test for stratified contingency tables.

From Significant Pattern Mining to Combinatorial Association Mapping

Questions unanswered in 2014

3. Can we retain efficiency and statistical power when accounting for categorical covariates such as age and gender? (NIPS 2016)
   - We extended Tarone's trick to the Cochran-Mantel-Haenszel-test for stratified contingency tables.
4. Can we develop new combinatorial association mapping approaches based on Tarone's trick? (ISMB 2015, OUP Bioinformatics 2017, ISMB 2018)

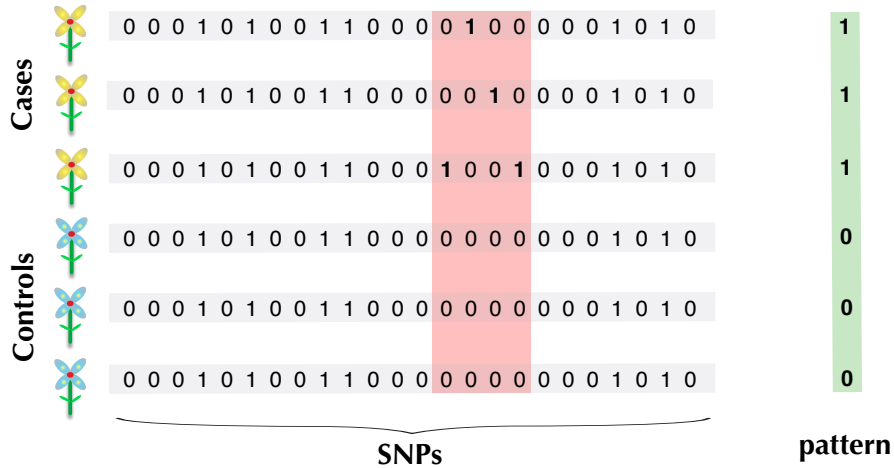
From Significant Pattern Mining to Combinatorial Association Mapping

Questions unanswered in 2014

3. Can we retain efficiency and statistical power when accounting for categorical covariates such as age and gender? (NIPS 2016)
   - We extended Tarone's trick to the Cochran-Mantel-Haenszel-test for stratified contingency tables.
4. Can we develop new combinatorial association mapping approaches based on Tarone's trick? (ISMB 2015, OUP Bioinformatics 2017, ISMB 2018)

Combinatorial Association Mapping for Genetic Heterogeneity Discovery

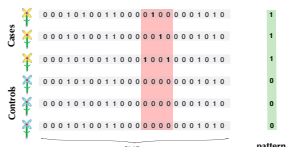
Genetic Heterogeneity Discovery



Genetic Heterogeneity Discovery

Genetic heterogeneity

- Genetic heterogeneity refers to the phenomenon that several different genes or sequence variants may give rise to the same phenotype.



- Specific problem studied here: Find a genomic interval such that having
  - a rare variant,
  - a recessive genotype, or
  - a minor allele

  in this interval is associated with the disease phenotype.

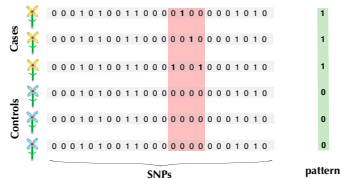
Genetic Heterogeneity Discovery

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

- **Current state of the art**: Restrict search to intervals that correspond to genes or exons (Lee et al., AJHG 2014).
- 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.

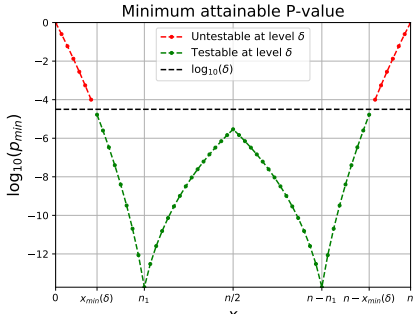
FAIS: Finding Intervals That Exhibit Genetic Heterogeneity

Genetic Heterogeneity Discovery as a Pattern Mining Problem



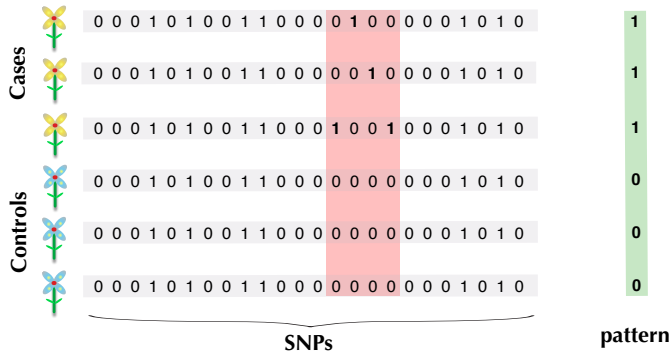
- We model the search as a pattern mining problem: Given an interval, an individual contains a pattern, if it has at least one minor allele in this interval.
- An interval is represented by its maximum value. The longer an interval, the more likely it is that this maximum is 1.
- Association is measured by Fisher's exact test, and we control the family-wise error rate using Tarone's method.

FAIS: Finding Intervals That Exhibit Genetic Heterogeneity



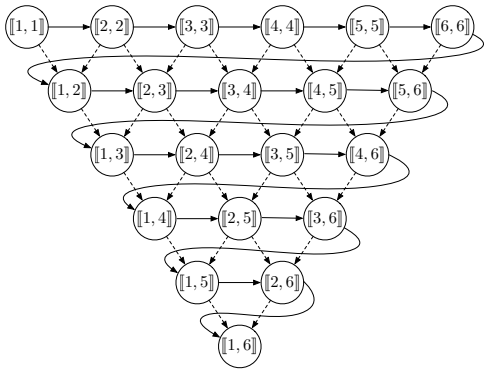
- If too many individuals have a particular pattern, the corresponding interval is not testable.

FAIS: Finding Intervals That Exhibit Genetic Heterogeneity



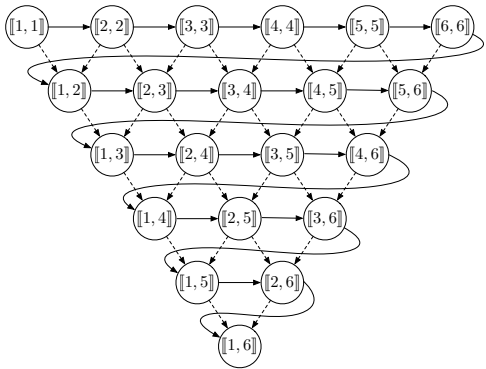
- **Pruning criterion:** If a pattern is too frequent to be testable, then none of the superintervals of the corresponding interval is testable.

FAIS: Finding Intervals That Exhibit Genetic Heterogeneity



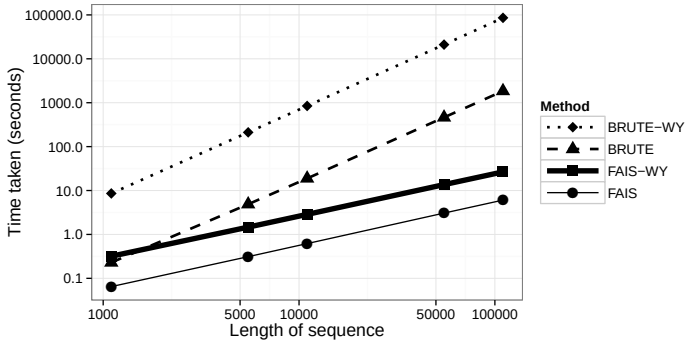
- **Search strategy:** We search intervals of increasing length l and prune untestable superintervals.

FAIS: Finding Intervals That Exhibit Genetic Heterogeneity



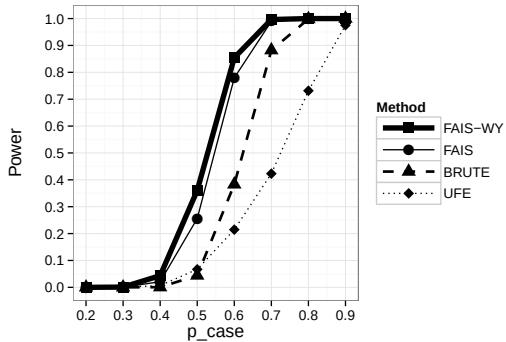
- **Search strategy:** Specifically, for each interval of length l, we prune it if at least one of its two length $l-1$ subintervals is too frequent to be testable.

FAIS: Finding Intervals That Exhibit Genetic Heterogeneity



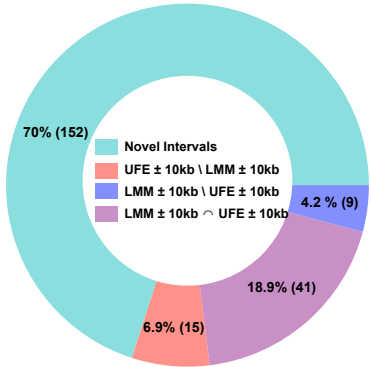
- Our method FAIS (Fast Automatic Interval Search) improves over the brute-force interval search in terms of runtime in simulations.

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



- FAIS detects 217 significant intervals on 21 binary phenotypes from *Arabidopsis thaliana*, with 214,051 SNPs and up to 194 lines (Atwell et al., Nature 2010).
- 70% would have been missed by univariate approaches (UFE and LMM).

Combinatorial Association Mapping: Summary and Outlook

Summary

- Significant Pattern Mining was long considered an unsolvable problem.

Combinatorial Association Mapping: Summary and Outlook

Summary

- Significant Pattern Mining was long considered an unsolvable problem.
- We have improved Significant Pattern Mining on several levels that allow us to use it for instances of Combinatorial Association Mapping at a genome-wide scale, e.g. for Genetic Heterogeneity Discovery.

Combinatorial Association Mapping: Summary and Outlook

Summary

- Significant Pattern Mining was long considered an unsolvable problem.
- We have improved Significant Pattern Mining on several levels that allow us to use it for instances of Combinatorial Association Mapping at a genome-wide scale, e.g. for Genetic Heterogeneity Discovery.

`www.significant-patterns.org`

Combinatorial Association Mapping: Summary and Outlook

Summary

- Significant Pattern Mining was long considered an unsolvable problem.
- We have improved Significant Pattern Mining on several levels that allow us to use it for instances of Combinatorial Association Mapping at a genome-wide scale, e.g. for Genetic Heterogeneity Discovery.

`www.significant-patterns.org`

Next challenges

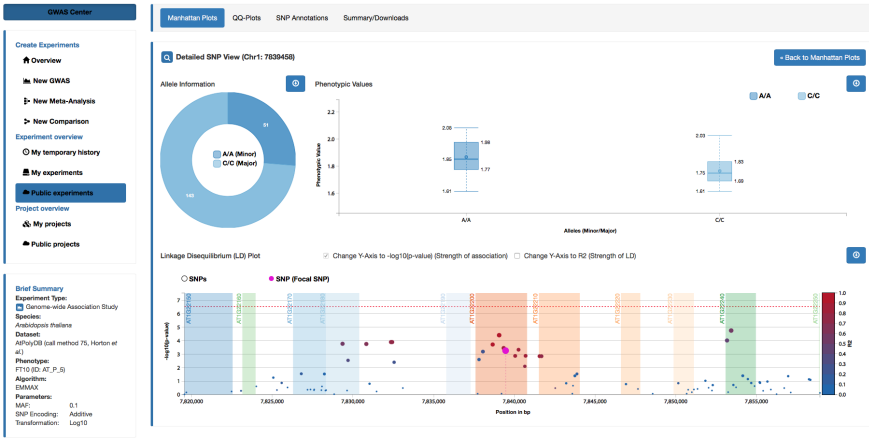
- How to detect genetic heterogeneity in biological pathways?
- How to control the False Discovery Rate?
- How to deal with non-binary features and non-binary phenotypes?

What's next?

Biomedical Software Development

easyGWAS

- We have been developing easygwas.org (Grimm et al., 2017), a cloud platform for genome-wide association studies (1580 users as of September 19, 2018):



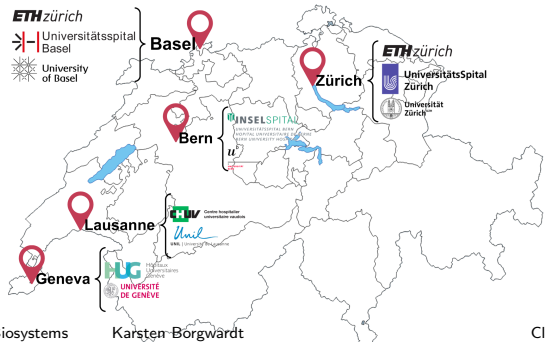
Biomarker Discovery for Sepsis

Personalized Swiss Sepsis Study

- Consortium of 22 research labs and 5 university hospitals in Switzerland
- Goal: Predict sepsis and sepsis-related mortality
- Approach: Integrate clinical data and molecular data for joint biomarker discovery

Adrian Egli
PI SPHN
Clinical Microbiology, University Hospital Basel

Karsten Borgwardt
PI PHRT
MLCB, D-BSSE, ETH Zürich



- Duration:
  3 years
  (2018-2021)
- Total funding:
  5.3 Million CHF

Predicting Sepsis

Background: What is sepsis and why is it relevant?

- Sepsis is a life-threatening organ dysfunction, caused by a dysregulated host response to infection (Singer et al., 2016).
- Identification of a bacterial species in blood still takes between 24h and 48h after blood sampling (Osthoff et al., 2017).
- From onset each hour of delayed effective antibiotic treatment increases mortality (Ferrer et al., 2014).

Predicting Sepsis

Background: What is sepsis and why is it relevant?

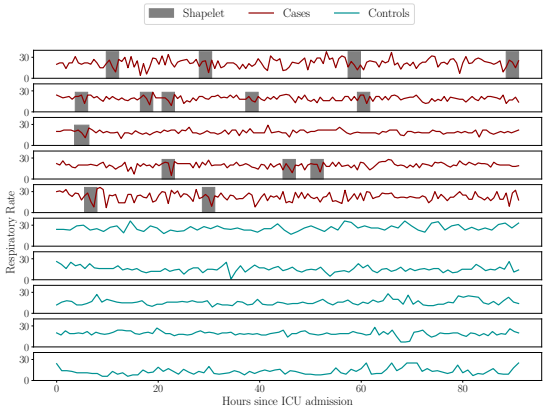
- Sepsis is a life-threatening organ dysfunction, caused by a dysregulated host response to infection (Singer et al., 2016).
- Identification of a bacterial species in blood still takes between 24h and 48h after blood sampling (Osthoff et al., 2017).
- From onset each hour of delayed effective antibiotic treatment increases mortality (Ferrer et al., 2014).

→ The first hours of sepsis are of critical importance.
→ Currently, when sepsis is detected, organ damage has already progressed.
→ **Detecting and treating sepsis earlier** and better identifying high-risk subgroups could be of highest clinical impact.

Predicting Sepsis

- Dataset: MIMIC (https://mimic.physionet.org)
- Labels:
  - Case if sepsis-3 criteria (Singer et al., 2016) fulfilled during ICU stay (at least 4 hours after admission) using the notion 'suspicion of infection' as defined in (Seymour et al., 2016)
  - Control if no suspicion of infection (at least not during ICU stay and the 2 preceding weeks)
  - SOFA increase evaluated as the maximum SOFA of the 3d window around suspicion of infection (–2 to +1 days) compared to baseline SOFA (3d window before that).
- Features: In-ICU time series of heart rate, systolic blood pressure, and respiratory rate.
- Sample size: 355 case ICU stays, 21,079 controls (sampling 355)
- Exclusion criteria: Age < 15, CareVue logging (insufficient detail), chartvalues or in/out-time unavailable.

Predicting Sepsis



- We detect patterns in respiratory rate time series that are statistically significantly associated with sepsis (Bock et al., Bioinformatics 2018).

Data Mining in the Life Sciences

Data Mining in Genetics, Medicine and the Life Sciences

Outlook

- Automation, biomarker discovery, and biomedical data management will remain key research topics.

Data Mining in Genetics, Medicine and the Life Sciences

Outlook

- Automation, biomarker discovery, and biomedical data management will remain key research topics.
- Data growth in three dimensions will pose extreme new challenges in Data Mining in Genetics and Medicine:
  - Population-scale datasets of individuals
  - Life-long recordings of health state
  - Highest-resolution information of the health state

Data Mining in Genetics, Medicine and the Life Sciences

Outlook

- Automation, biomarker discovery, and biomedical data management will remain key research topics.
- Data growth in three dimensions will pose extreme new challenges in Data Mining in Genetics and Medicine:
  - Population-scale datasets of individuals
  - Life-long recordings of health state
  - Highest-resolution information of the health state
- How to mine (handle and use) this data?

Data Mining in Genetics, Medicine and the Life Sciences

Outlook

- Automation, biomarker discovery, and biomedical data management will remain key research topics.
- Data growth in three dimensions will pose extreme new challenges in Data Mining in Genetics and Medicine:
  - Population-scale datasets of individuals
  - Life-long recordings of health state
  - Highest-resolution information of the health state
- How to mine (handle and use) this data?
- Many branches of the Life Sciences face very similar or analogous problems.

Data Mining in Genetics, Medicine and the Life Sciences

Outlook

- Automation, biomarker discovery, and biomedical data management will remain key research topics.
- Data growth in three dimensions will pose extreme new challenges in Data Mining in Genetics and Medicine:
  - Population-scale datasets of individuals
  - Life-long recordings of health state
  - Highest-resolution information of the health state
- How to mine (handle and use) this data?
- Many branches of the Life Sciences face very similar or analogous problems.

Plenty of opportunities for Data Mining in the Life Sciences

Thank you

- Starting Grant (ERC-Backup Scheme of the SNSF)
- Marie-Curie-Initial Training Network for ‘Machine Learning for Personalized Medicine’ (mlpm.eu, 2013-2016)
- Marie-Curie-Initial Training Network for ‘Machine Learning Frontiers in Precision Medicine’ (mlpm.eu, 2019-2022)
- Alfried-Krupp-Award for Young Professors
- SPHN-PHRT Driver Project ‘Personalized Swiss Sepsis Study’

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

References I


T. F. Mackay, J. H. Moore, *Genome Medicine* **6**, 42 (2014).


T. A. Manolio, *et al.*, *Nature* **461**, 747 (2009).


S. Lee, *et al.*, *The American Journal of Human Genetics* **95**, 5 (2014).


F. Llinares-Lopez, *et al.*, *KDD* (2015), pp. 725–734.


F. Llinares-Lopez, *et al.*, *Bioinformatics* **31**, i240 (2015).


F. Llinares-Lopez, *et al.*, *Bioinformatics (Oxford, England)* (2017).


L. Papaxanthos, *et al.*, *NIPS*, D. D. Lee, *et al.*, eds. (2016), pp. 2271–2279.


M. Sugiyama, *et al.*, *SIAM Data Mining*, S. Venkatasubramanian, J. Ye, eds. (SIAM, 2015), pp. 37–45.