ETH zürich

Die ‘Daten-Medizin’

Karsten Borgwardt

Department Biosysteme, ETH Zürich

Esslingen, 13.4.2018

Maschinelles Lernen und personalisierte Medizin

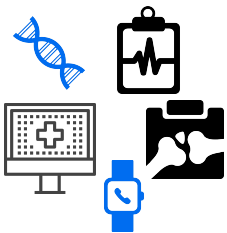
Ziele

- Mit maschinellem Lernen sollen statistische Abhängigkeiten in großen Datensätzen erkannt werden.

Maschinelles Lernen und personalisierte Medizin

Ziele

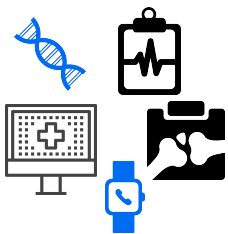
- Mit maschinellem Lernen sollen statistische Abhängigkeiten in großen Datensätzen erkannt werden.



Maschinelles Lernen und personalisierte Medizin

Ziele

- Mit maschinellem Lernen sollen statistische Abhängigkeiten in großen Datensätzen erkannt werden.



- Personalisierte Medizin versucht Gesundheitsdaten für verbesserte Diagnosen, Vorhersagen und Therapieentscheidungen zu nutzen, individuell angepasst auf die Eigenschaften eines jeden Patienten.

Maschinelles Lernen in der Medizin

Hauptthemen

Maschinelles Lernen in der Medizin

Hauptthemen

- Automatisierung der Diagnostik

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

Maschinelles Lernen in der Medizin

Hauptthemen

- Automatisierung der Diagnostik
- Entdeckung von Biomarkern

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

Maschinelles Lernen in der Medizin

Hauptthemen

- Automatisierung der Diagnostik
- Entdeckung von Biomarkern

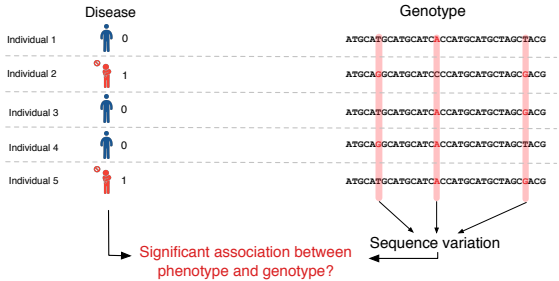
Roche to **buy Flatiron** Health for $1.9 billion to expand cancer care ...
Reuters - 15.02.2018
Roche to **buy Flatiron** Health for $1.9 billion to expand cancer care portfolio ... S) said on Thursday it would **buy** the rest of U.S. cancer data company **Flatiron** Health for $1.9 billion to speed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

- Biomedizinisches Datenmanagement

Maschinelles Lernen in der Medizin

Hauptthemen

- Automatisierung der Diagnostik
- Entdeckung von Biomarkern
  1. Methodenentwicklung: Kombinatorische Assoziationssuche
- Biomedizinisches Datenmanagement



Maschinelles Lernen in der Medizin

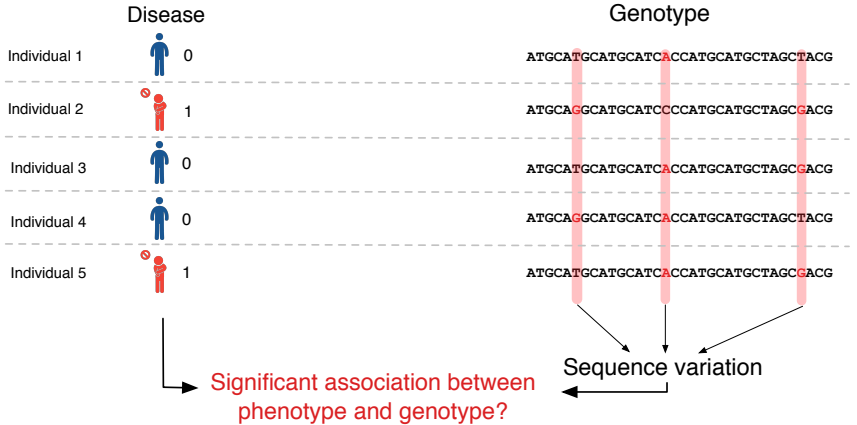
Hauptthemen

- Automatisierung der Diagnostik

- Entdeckung von Biomarkern
  1. Methodenentwicklung: Kombinatorische Assoziationssuche
- Biomedizinisches Datenmanagement
  2. Softwareentwicklung: easyGWAS

Kombinatorische Assoziationssuche

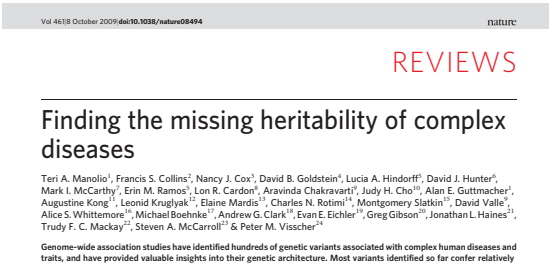
Assoziationssuche: Korrelationen zwischen Phänotyp und Genotyp



Eine genomweite Assoziationsstudie (GWAS) untersucht, ob Varianten im Genom (in Form von Einzelnukleotid-Polymorphismen, SNPs) mit Veränderungen des Phänotyps korreliert sind.

Assoziationssuche: Fehlende Erblichkeit

- Seit 2001: $>$ 59,000 Phänotyp-assoziierte Genomregionen durch GWAS (GWAS Catalog, 6.3.2018)
- Problem: Erklärte Varianz des Phänotyps enttäuschend gering



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

Assoziationssuche: Fehlende Erblichkeit

- Seit 2001: $>$ 59,000 Phänotyp-assoziierte Genomregionen durch GWAS (GWAS Catalog, 6.3.2018)
- Problem: Erklärte Varianz des Phänotyps enttäuschend gering

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

- Mögliche Gründe:
  - Polygenische Architektur komplexer Krankheiten
  - Kleine Effektstärken
  - Unvollständige Integration wichtiger genetischer, epigenetischer und nichtgenetischer Eigenschaften

Assoziationssuche: Fehlende Erblichkeit

- Seit 2001: $>$ 59,000 Phänotyp-assoziierte Genomregionen durch GWAS (GWAS Catalog, 6.3.2018)
- Problem: Erklärte Varianz des Phänotyps enttäuschend gering

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

- Mögliche Gründe:
  - Polygenische Architektur komplexer Krankheiten $\rightarrow$ Epistasis
  - Kleine Effektstärken
  - Unvollständige Integration wichtiger genetischer, epigenetischer und nichtgenetischer Eigenschaften

Assoziationssuche: Fehlende Erblichkeit

Epistase als möglische Ursache

- Die meisten aktuellen Untersuchungen vernachlässigen interaktive Effekte zwischen SNPs
- Bedarf an Methoden für **kombinatorische Assoziationssuche**

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

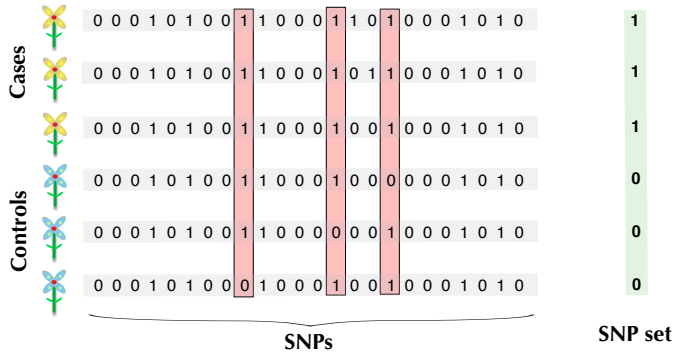
Trudy FC Mackay[1*] and Jason H Moore[2]

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-

Kombinatorische Assoziationssuche



- Informatische Herausforderung: Kombinatorische Explosion der Anzahl der Kandidatenmengen
- Statistische Herausforderung: Kombinatorische Explosion der Anzahl der Assoziationstests

Kombinatorische Assoziationssuche

Das Problem der Alphafehler-Kumulierung

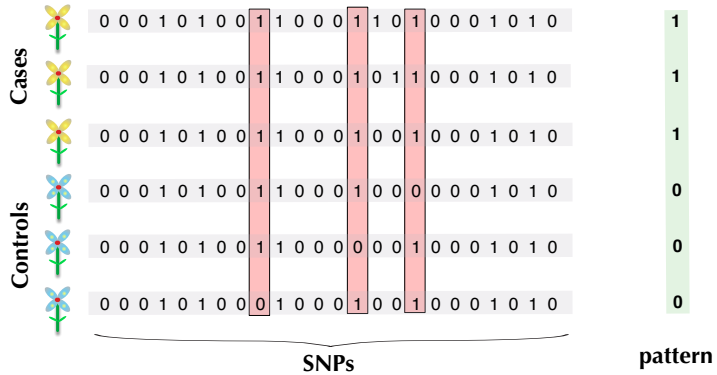
- Was, wenn wir Assoziationen von Gruppen von s SNPs mit dem Phänotyp betrachten?
- Dies führt multiplem Hypothesentesten: Jede der k SNP-Mengen entspricht einer zu testenden Hypothese ($k \in O(f^s)$), wobei f die Anzahl der SNPs ist.
- Falls nicht berücksichtigt, könnten selbst bei zufällig generierten Daten α Prozent aller SNP-Mengen als assoziiert betrachtet werden.
- Es ist zwingend notwendig, multiples Testen zu berücksichtigen, z. B. durch Kontrolle der family-wise error rate!
- Falls berücksichtig, z. B. durch Bonferroni-Korrektur ($\frac{\alpha}{k}$), könnte die gesamte Teststärke verloren gehen.

Kombinatorische Assoziationssuche

Das Problem der Alphafehler-Kumulierung

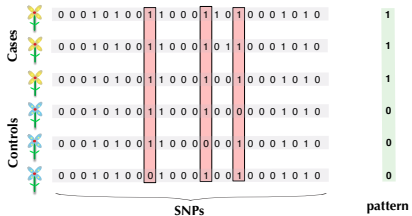
- Was, wenn wir Assoziationen von Gruppen von s SNPs mit dem Phänotyp betrachten?
- Dies führt multiplem Hypothesentesten: Jede der k SNP-Mengen entspricht einer zu testenden Hypothese ($k \in O(f^s)$), wobei f die Anzahl der SNPs ist.
- Falls nicht berücksichtigt, könnten selbst bei zufällig generierten Daten α Prozent aller SNP-Mengen als assoziiert betrachtet werden.
- Es ist zwingend notwendig, multiples Testen zu berücksichtigen, z. B. durch Kontrolle der family-wise error rate!
- Falls berücksichtig, z. B. durch Bonferroni-Korrektur ($\frac{\alpha}{k}$), könnte die gesamte Teststärke verloren gehen.
- **Lange als unlösbares Dilemma angesehen**

Kombinatorische Assoziationssuche als Data-Mining-Problem



- Variablenselektion: Finde die Variablen, die Klassen von Objekten unterscheiden
- Kombinationssuche: Finde **Kombinationen binärer Variablen** höherer Ordnung, um die Klassen zu unterscheiden

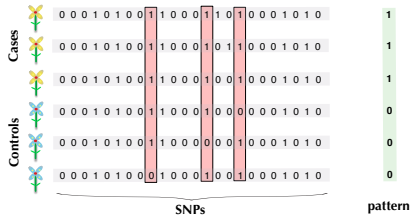
Kombinatorische Assoziationssuche als Data-Mining-Problem



Kombination

- In unserem Datensatz D sei der i-te von n Patienten dargestellt durch einen binären Vektor $\mathbf{d}^{(i)} \in \{0,1\}^f$ und ein Klassenlabel $y_i \in \{0,1\}$.
- Wir wählen eine Teilmenge $\mathcal{S}$ aller Variablen $\mathcal{F}$ in einem Datensatz: $\mathcal{S} \subseteq \mathcal{F}$.
- Dann enthält ein Objekt $\mathbf{d}^{(i)}$ die Kombination $\mathcal{S}$ genau dann, wenn $\prod_{t \in \mathcal{S}} d^{(i)}(t) = 1$.

Kombinatorische Assoziationssuche als Data-Mining-Problem



Problemstellung: Suche nach signifikanten Kombinationen

- Wir wollen alle Mengen $\mathcal{S}$ finden, so dass es eine statistisch signifikante Assoziation zwischen $\prod_{t \in \mathcal{S}} d^{(i)}(t)$ und y_i für $i \in \{1, \ldots, n\}$ besteht, unter Kontrolle der Family Wise Error Rate auf Level α.

Suche nach signifikanten Kombinationen

Tarones Trick

- Kontingenztabelle

	Kombination vorhanden	Kombination fehlt	
y=0	a	$n_1 - a$	n_1
y=1	$x - a$	$n - n_1 - x + a$	$n - n_1$
	x	$n - x$	n

- Häufig wird der exakte Test nach Fisher gewählt, um zu testen, ob eine Kombination in einer von zwei Klassen überrepräsentiert ist.
- Die übliche Methode zur Berechnung der p-Werte basiert auf der hypergeometrischen Verteilung und setzt voraus, dass die Randhäufigkeiten (x, n_1, n) fix sind.

Suche nach signifikanten Kombinationen

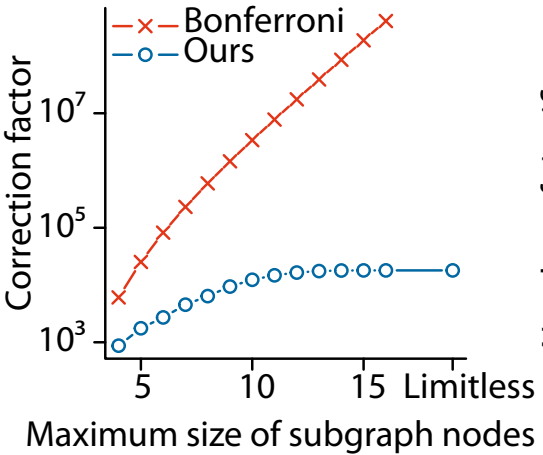
Tarones Trick

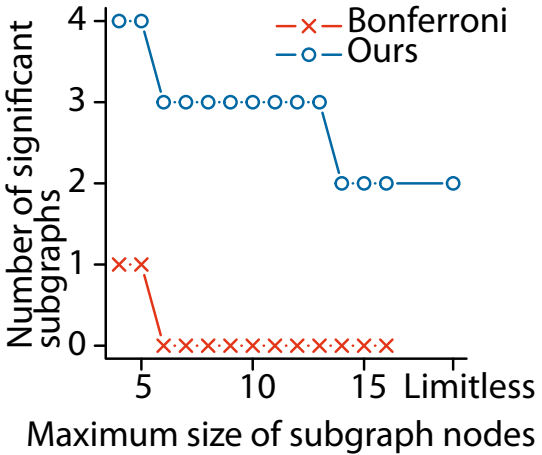
- Kontingenztabelle

	Kombination vorhanden	Kombination fehlt	
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) erkannte, dass es bei diskreten Teststatistiken, wie z. B. beim exakten Test nach Fisher, einen minimalen p-Wert gibt, den eine Kombination annehmen kann.
- Es gibt viele nicht testbare Hypothesen, deren minimaler p-Wert nicht kleiner ist als $\frac{\alpha}{k}$.
- Nur die übrigen $m(k)$ testbaren Hypothesen können überhaupt signifikant sein.
- Man kann für $m(k)$ statt k Tests korrigieren. Da oft $m(k) << k$, verbessert das die Teststärke sehr.

Beispiel: PTC Datensatz (Helma et al., 2001)





Suche nach signifikanten Kombinationen

Tarones Methode (1990)

- Sei k die Anzahl der Tests, für die korrigiert werden soll.
- $m(k)$ ist die Anzahl der testbaren Hypothesen zum Signifikanzniveau $\frac{\alpha}{k}$.
- $m(k)$ ist eine Funktion von k, und wir benötigen $k \geq m(k)$, um für alle testbaren Hypothesen zu korrigieren.
- Das Optimierungsproblem lautet dann

$$\begin{aligned} &\min k \\ &\text{so dass } k \geq m(k) \end{aligned}$$

Suche nach signifikanten Kombinationen

Tarones Methode (1990)

- Sei k die Anzahl der Tests, für die korrigiert werden soll.
- $m(k)$ ist die Anzahl der testbaren Hypothesen zum Niveau $\frac{\alpha}{k}$.

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

Suche nach signifikanten Kombinationen

Tarones Methode (1990)

- Sei k die Anzahl der Tests, für die korrigiert werden soll.
- $m(k)$ ist die Anzahl der testbaren Hypothesen zum Niveau $\frac{\alpha}{k}$.

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

- Wie kann $m(k)$ effizient berechnet werden, ohne alle $O(f^s)$ möglichen Hypothesen zu durchlaufen?

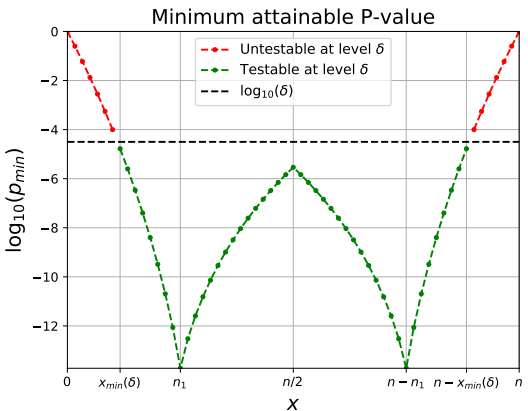
Suche nach signifikanten Kombinationen

Data-Mining Problem

- Wie soll man $m(k)$ effizient berechnen, ohne alle $O(f^s)$ möglichen Hypothesen zu durchlaufen?
- Lösung: Die Häufigkeit einer Kombination bestimmt deren minimalen p-Wert.
- Algorithmen zur Suche von häufigen Kombinationen aus dem Data-Mining können benutzt werden, um alle Kombinationen zu finden, deren minimaler p-Wert einen gewissen Schwellwert unterschreitet (Terada et al., PNAS 2013):
  - `frequent itemset mining`(D, θ) findet alle Kombinationen in einem Datensatz D mit einer Häufigkeit von mindestens θ.

Suche nach signifikanten Kombinationen

- Häufigkeit versus minimaler p-Wert



Suche nach signifikanten Kombinationen

Tarones Methode mit frequent itemset mining

- Sei k die Anzahl der Tests für die korrigiert werden soll.
- $m(k)$ ist die Anzahl der testbaren Hypothesen mit Signifikanzniveau $\frac{\alpha}{k}$.

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

Suche nach signifikanten Kombinationen

Tarones Methode mit frequent itemset mining

- Sei k die Anzahl der Tests für die korrigiert werden soll.
- $m(k)$ ist die Anzahl der testbaren Hypothesen mit Signifikanzniveau $\frac{\alpha}{k}$.

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

- Hinweis: $\phi(\frac{\alpha}{k})$ ist die minimale Häufigkeit einer Kombination, die mit Niveau $\frac{\alpha}{k}$ testbar ist.

Suche nach signifikanten Kombinationen

Tarones Methode mit frequent itemset mining

- Sei k die Anzahl der Tests für die korrigiert werden soll.
- $m(k)$ ist die Anzahl der testbaren Hypothesen mit Signifikanzniveau $\frac{\alpha}{k}$.

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

- Hinweis: $\phi(\frac{\alpha}{k})$ ist die minimale Häufigkeit einer Kombination, die mit Niveau $\frac{\alpha}{k}$ testbar ist.
- Für kleine k ist $\phi(\frac{\alpha}{k})$ klein. Frequent itemset mining wird extrem rechenaufwändig!

Von der Suche nach signifikanten Kombinationen zur kombinatorischen Assoziationssuche

Fragen, die 2014 noch unbeantwortet waren

1. Wie kann das optimale k effizient gefunden werden? (SDM 2015)

Von der Suche nach signifikanten Kombinationen zur kombinatorischen Assoziationssuche

Fragen, die 2014 noch unbeantwortet waren

1. Wie kann das optimale k effizient gefunden werden? (SDM 2015)
   - Wir präsentierten eine effiziente Suchstrategie mit Kriterium zum vorzeitigen Abbruch (falls $m(k) > k$).

Von der Suche nach signifikanten Kombinationen zur kombinatorischen Assoziationssuche

Fragen, die 2014 noch unbeantwortet waren

1. Wie kann das optimale k effizient gefunden werden? (SDM 2015)
   - Wir präsentierten eine effiziente Suchstrategie mit Kriterium zum vorzeitigen Abbruch (falls $m(k) > k$).
2. Kombinationen stehen in Untermengen-/Obermengenbeziehungen. Wie kann diese Abhängigkeit der Tests berücksichtigt werden? (KDD 2015)

Von der Suche nach signifikanten Kombinationen zur kombinatorischen Assoziationssuche

Fragen, die 2014 noch unbeantwortet waren

1. Wie kann das optimale k effizient gefunden werden? (SDM 2015)
   - Wir präsentierten eine effiziente Suchstrategie mit Kriterium zum vorzeitigen Abbruch (falls $m(k) > k$).
2. Kombinationen stehen in Untermengen-/Obermengenbeziehungen. Wie kann diese Abhängigkeit der Tests berücksichtigt werden? (KDD 2015)
   - Wir führen Westfall-Young Permutationen durch, um die Abhängingkeit zu berücksichtigen.
   - Durch dynamische Aktualisierung des Häufigkeitsschwellwerts benötigen wir auch bei 10,000 Permutationen lediglich 1 einzige Anwendung von frequent itemset mining.

Von der Suche nach signifikanten Kombinationen zur kombinatorischen Assoziationssuche

Fragen, die 2014 noch unbeantwortet waren

3. Können wir Effizienz und Teststärke beibehalten, wenn wir kategorische Kovariaten wie Alter und Geschlecht berücksichtigen? (NIPS 2016)

Von der Suche nach signifikanten Kombinationen zur kombinatorischen Assoziationssuche

Fragen, die 2014 noch unbeantwortet waren

3 Können wir Effizienz und Teststärke beibehalten, wenn wir kategorische Kovariaten wie Alter und Geschlecht berücksichtigen? (NIPS 2016)

- Wir erweiterten Tarones Trick auf den Cochran-Mantel-Haenszel-Test für stratifizierte Kontingenztabellen.

Von der Suche nach signifikanten Kombinationen zur kombinatorischen Assoziationssuche

Fragen, die 2014 noch unbeantwortet waren

3 Können wir Effizienz und Teststärke beibehalten, wenn wir kategorische Kovariaten wie Alter und Geschlecht berücksichtigen? (NIPS 2016)
   - Wir erweiterten Tarones Trick auf den Cochran-Mantel-Haenszel-Test für stratifizierte Kontingenztabellen.

4 Können wir neue kombinatorische Assoziationsverfahren entwickeln, die auf Tarones Trick basieren? (ISMB 2015, OUP Bioinformatics 2017, ISMB 2018)

Von der Suche nach signifikanten Kombinationen zur kombinatorischen Assoziationssuche

Fragen, die 2014 noch unbeantwortet waren

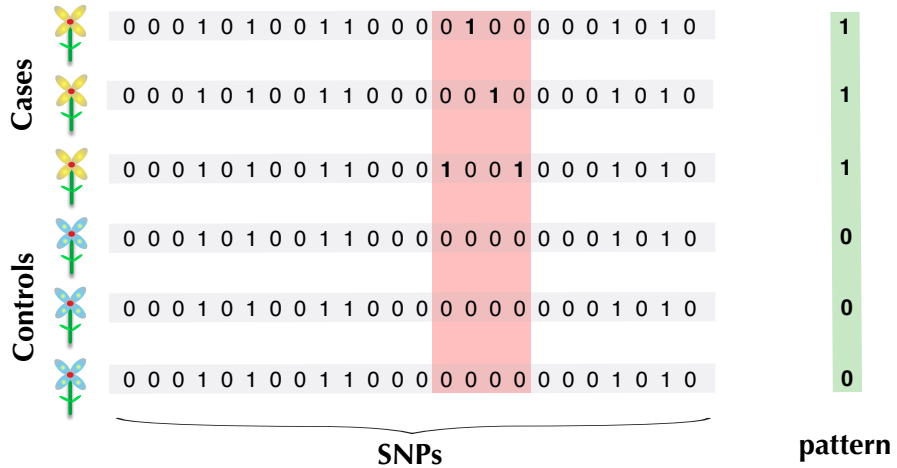
3 Können wir Effizienz und Teststärke beibehalten, wenn wir kategorische Kovariaten wie Alter und Geschlecht berücksichtigen? (NIPS 2016)

- Wir erweiterten Tarones Trick auf den Cochran-Mantel-Haenszel-Test für stratifizierte Kontingenztabellen.

4 Können wir neue kombinatorische Assoziationsverfahren entwickeln, die auf Tarones Trick basieren? (ISMB 2015, OUP Bioinformatics 2017, ISMB 2018)

Kombinatorische Assoziationskartierung für die Erkennung von genetischer Heterogenität

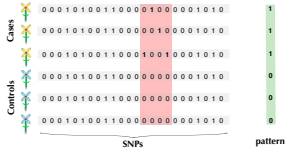
Erkennung von genetischer Heterogenität



Erkennung von genetischer Heterogenität

Genetische Heterogenität

- Genetische Heterogenität bezeichnet das Phänomen, dass mehrere verschiedene Gene oder Sequenzvarianten zu demselben Phänotyp führen können.



- Spezielles Problem, das hier betrachtet wird: Finde ein Intervall im Genom, so dass die Eigenschaft, dass
  - eine seltene Variante,
  - ein rezessiver Genotyp, oder
  - ein Minderheitsallel

  in diesem Intervall liegt, mit dem Phänotyp assoziiert ist.

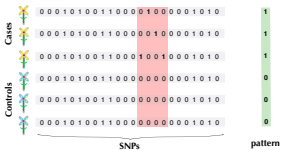
Erkennung von genetischer Heterogenität

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

- **Aktueller Stand der Forschung**: Beschränkung der Suche auf Intervalle, die Genen oder Exons entsprechen (Lee et al., AJHG 2014).
- Unser Ziel ist es, nach Intervallen zu suchen, in denen genetische Heterognität vorkommen könnte, und dabei
  - beliebige Start- und Endpunkte der Intervalle zu erlauben,
  - die Alphafehler-Kumulierung des inhärenten multiplen Testproblems zu korrigieren, und
  - Teststärke und Recheneffizient beizubehalten.

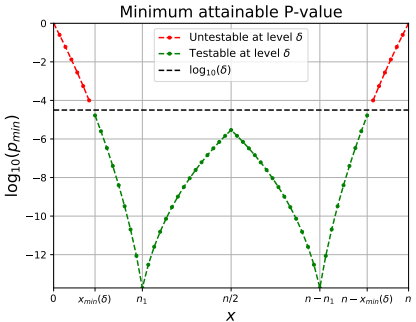
FAIS: Suche nach Intervallen mit genetischer Heterogenität

Erkennung von genetischer Heterogenität als Fragestellung der Kombinationssuche



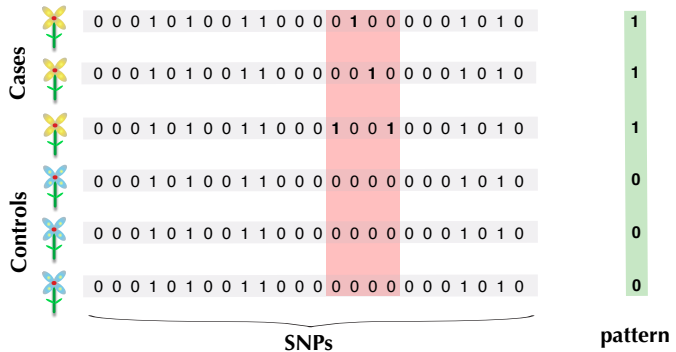
- Wir modellieren die Suche als Problem der Kombinationssuche: Gegeben ein Intervall, dann besitzt ein Individuum eine Kombination, wenn es mindestens eine seltene Variante in diesem Intervall hat.
- Ein Intervall wird durch seinen maximalen Wert dargestellt. Je länger das Intervall, desto höher ist die Wahrscheinlichkeit, dass dieses Maximum 1 ist.
- Assoziation wird mit dem exakten Test nach Fisher gemessen, und die Family Wise Error Rate mit Tarones Methode beschränkt.

FAIS: Suche nach Intervallen mit genetischer Heterogenität



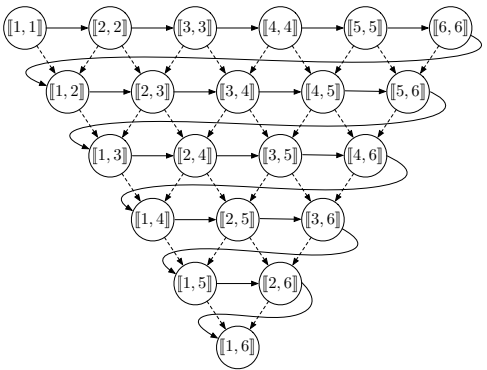
- Falls zu viele Individuen eine bestimmte Kombination besitzen, dann ist das entsprechende Intervall nicht testbar.

FAIS: Suche nach Intervallen mit genetischer Heterogenität



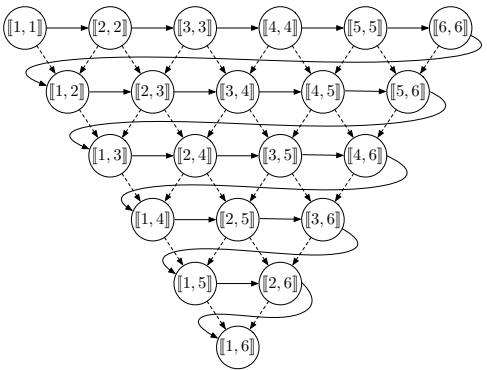
- **Pruning-Kriterium:** Falls eine Kombination zu häufig ist, um testbar zu sein, dann ist kein Superintervall testbar.

FAIS: Suche nach Intervallen mit genetischer Heterogenität



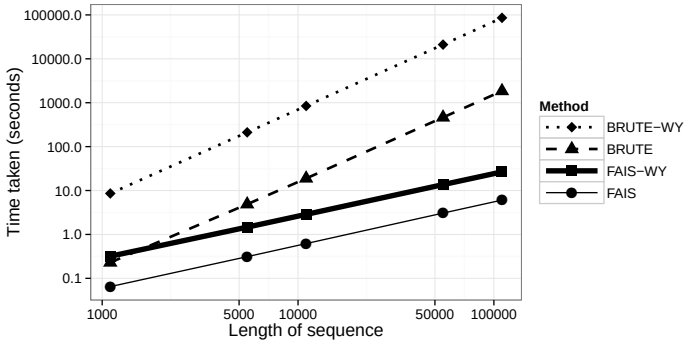
- **Suchstrategie:** Wir suchen Intervalle steigender Länge l und verwerfen nicht-testbare Superintervalle.

FAIS: Suche nach Intervallen mit genetischer Heterogenität



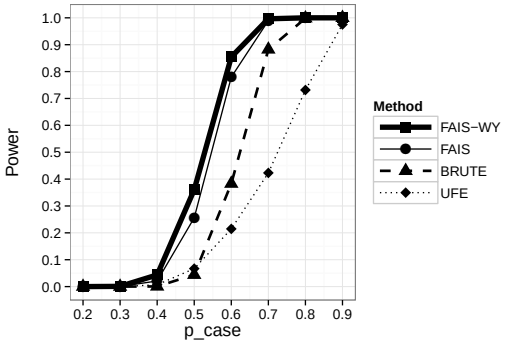
- **Suchstrategie:** Konkret überspringen wir ein Intervall der Länge l, falls mindestens eines seiner beiden Subintervalle der Länge $l - 1$ zu häufig ist, um testbar zu sein.

FAIS: Suche nach Intervallen mit genetischer Heterogenität



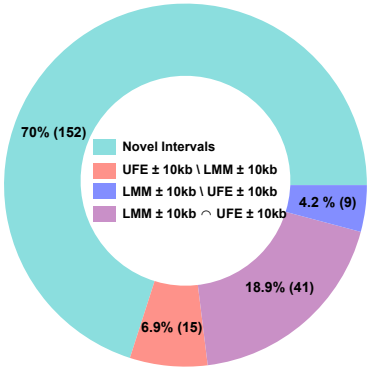
- Unsere Methode FAIS (Fast Automatic Interval Search) erzielt in Simulationen eine bessere Laufzeit als die 'brute-force' Intervallsuche.

FAIS: Suche nach Intervallen mit genetischer Heterogenität



- Unsere Methode FAIS (Fast Automatic Interval Search) erzielt in Simulationen eine bessere Teststärke als die 'brute-force' Intervallsuche und univariate Methoden.

FAIS: Suche nach Intervallen mit genetischer Heterogenität



- FAIS entdeckt 217 signifikante Intervalle bei 21 binären Phänotypen von *Arabidopsis thaliana*, mit 214,051 SNPs und bis zu 194 Linien (Atwell et al., Nature 2010).
- 70% wären mit univariaten Methoden (UFE and LMM) übersehen worden.

Kombinatorische Assoziationssuche: Zusammenfassung und Ausblick

Zusammenfassung

- Die Suche nach signifikanten Kombinationen galt lange als unlösbar.

Kombinatorische Assoziationssuche: Zusammenfassung und Ausblick

Zusammenfassung

- Die Suche nach signifikanten Kombinationen galt lange als unlösbar.
- Wir haben die kombinatorische Assoziationssuche auf mehreren Ebenen verbessert, wodurch genomweite kombinatorische Assoziationssuche (z. B. für die Erkennung von genetischer Heterogenität) ermöglicht wird.

Kombinatorische Assoziationssuche: Zusammenfassung und Ausblick

Zusammenfassung

- Die Suche nach signifikanten Kombinationen galt lange als unlösbar.
- Wir haben die kombinatorische Assoziationssuche auf mehreren Ebenen verbessert, wodurch genomweite kombinatorische Assoziationssuche (z. B. für die Erkennung von genetischer Heterogenität) ermöglicht wird.

`www.significant-patterns.org`

Kombinatorische Assoziationssuche: Zusammenfassung und Ausblick

Zusammenfassung

- Die Suche nach signifikanten Kombinationen galt lange als unlösbar.
- Wir haben die kombinatorische Assoziationssuche auf mehreren Ebenen verbessert, wodurch genomweite kombinatorische Assoziationssuche (z. B. für die Erkennung von genetischer Heterogenität) ermöglicht wird.

`www.significant-patterns.org`

Kommende Herausforderungen

- Wie kann genetische Heterogenität in biologischen Netzwerken entdeckt werden?
- Wie kann die False Discovery Rate kontrolliert werden?
- Wie können nichtbinäre Features und nichtbinäre Phänotypen behandelt werden?

Biomedizinische Softwareentwicklung

easyGWAS

- Wir haben easygwas.org (Grimm et al., 2017), eine Cloud-Plattform für genomweite Assoziationsstudien (1362 Benutzer am 10. April 2018), entwickelt:

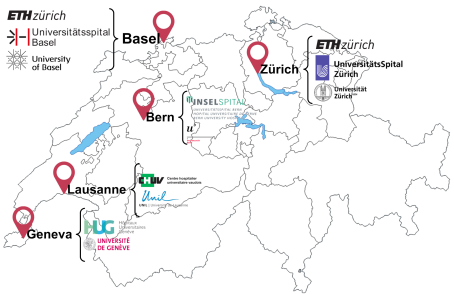
Biomarkerentdeckung für Sepsis

Personalized Swiss Sepsis Study

- Konsortium aus 22 Forschungsgruppen und 5 Universitätskrankenhäusern in der Schweiz
- Ziel: Vorhersage von Sepsis und der damit zusammenhängenden Sterblichkeit
- Vorgehen: Integration von klinischen und molekularen Daten zur gemeinsamen Biomarkerentdeckung

Adrian Egli
PI SPHN
Clinical Microbiology, University Hospital Basel

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



- Dauer:
  3 Jahre
  (2018-2021)
- Gesamtbudget:
  5.3 Millionen CHF

Vorhersage von Sepsis

Hintergrund: Was ist Sepsis und warum ist sie relevant?

- Sepsis ist ein lebensbedrohliches Organversagen, verursacht durch eine fehlregulierte Reaktion des Körpers auf eine Infektion (Singer et al., 2016).
- Nach der Blutentnahme kann es noch 24 bis 48h dauern, eine Bakterienspezies im Blut zu identifizieren (Osthoff et al., 2017).
- Nach Ausbruch der Erkrankung steigt das Sterberisiko mit jeder Stunde, um die die Antibiotikabehandlung verzögert wird (Ferrer et al., 2014).

Vorhersage von Sepsis

Hintergrund: Was ist Sepsis und warum ist sie relevant?

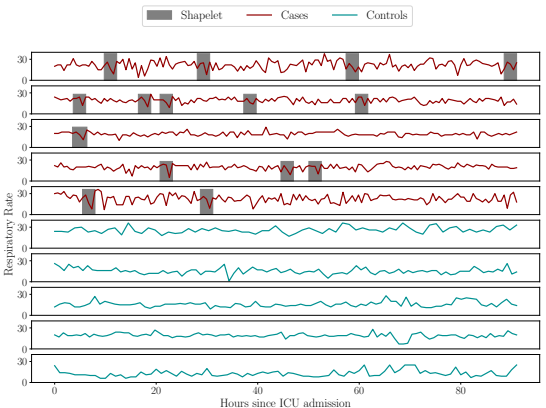
- Sepsis ist ein lebensbedrohliches Organversagen, verursacht durch eine fehlregulierte Reaktion des Körpers auf eine Infektion (Singer et al., 2016).
- Nach der Blutentnahme kann es noch 24 bis 48h dauern, eine Bakterienspezies im Blut zu identifizieren (Osthoff et al., 2017).
- Nach Ausbruch der Erkrankung steigt das Sterberisiko mit jeder Stunde, um die die Antibiotikabehandlung verzögert wird (Ferrer et al., 2014).

→ Die ersten Stunden sind von entscheidender Wichtigkeit.
→ Derzeit ist der Organschaden bereits fortgeschritten, wenn Sepsis diagnostiziert wird.
→ **Schnellere Erkennung und Behandlung von Sepsis** und bessere Identifikation von Hochrisikogruppen könnte von höchster klinischer Bedeutung sein.

Vorhersage von Sepsis

- Datensatz: MIMIC (https://mimic.physionet.org)
- Labels:
  - Fall: Sepsis-3-Kriterien (Singer et al., 2016) während AUfenthalts auf der Intensivstation (mindestens 4 Stunden nach Aufnahmen) erfüllt, wobei der Begriff 'suspicion of infection' wie in (Seymour et al., 2016) definiert wird.
  - Kontrolle: kein Infektionsverdacht (zumindest nicht während Aufenthalt auf Intensivstation und den 2 vorangegangenen Wochen)
  - Anstieg des SOFA-Scores, gemessen als der maximale SOFA-Score im 3d-Fenster um den Infektionsverdacht (–2 bis +1 Tage) im Vergleich zum Grundwert (baseline SOFA im 3d-Fenster davor).
- Features: Zeitreihen mit Herzfrequenz, systolischem Blutdruck und Atemfrequenz während intensivmedizinischer Behandlung
- Stichprobengröße: 355 intensivmedizinische Behandlungen im Fall, 21,079 Kontrollen (Zufallsstichprobe 355)

Vorhersage von Sepsis



- Wir entdecken Kombinationen in Zeitreihen mit der Atemfrequenz, die statistisch signifikant mit Sepsis assoziiert sind (Bock et al., ISMB 2018).

Data-Mining in den Lebenswissenschaften

Data-Mining in Genetik, Medizin und den Lebenswissenschaften

Ausblick

- Automatisierung, Biomarkerentdeckung und biomedizinsches Datenmanagement werden Hauptforschungsthemen bleiben.

Data-Mining in Genetik, Medizin und den Lebenswissenschaften

Ausblick

- Automatisierung, Biomarkerentdeckung und biomedizinsches Datenmanagement werden Hauptforschungsthemen bleiben.
- Datenwachstum in drei Dimensionen wird extreme neue Herausforderungen im Data-Mining in Genetik und Medizin generieren:
  - Datensätze von Individuen einer ganzen Population
  - Lebenslange Aufzeichnung des Gesundheitszustandes
  - Höchstaufgelöste Information über den Gesundheitszustand

Data-Mining in Genetik, Medizin und den Lebenswissenschaften

Ausblick

- Automatisierung, Biomarkerentdeckung und biomedizinsches Datenmanagement werden Hauptforschungsthemen bleiben.
- Datenwachstum in drei Dimensionen wird extreme neue Herausforderungen im Data-Mining in Genetik und Medizin generieren:
  - Datensätze von Individuen einer ganzen Population
  - Lebenslange Aufzeichnung des Gesundheitszustandes
  - Höchstaufgelöste Information über den Gesundheitszustand
- Wie können diese Daten ausgewertet und genutzt werden?

Data-Mining in Genetik, Medizin und den Lebenswissenschaften

Ausblick

- Automatisierung, Biomarkerentdeckung und biomedizinsches Datenmanagement werden Hauptforschungsthemen bleiben.
- Datenwachstum in drei Dimensionen wird extreme neue Herausforderungen im Data-Mining in Genetik und Medizin generieren:
  - Datensätze von Individuen einer ganzen Population
  - Lebenslange Aufzeichnung des Gesundheitszustandes
  - Höchstaufgelöste Information über den Gesundheitszustand
- Wie können diese Daten ausgewertet und genutzt werden?
- Viele Zweige der Lebenswissenschaften stehen vor sehr ähnlichen oder analogen Problemen.

Data-Mining in Genetik, Medizin und den Lebenswissenschaften

Ausblick

- Automatisierung, Biomarkerentdeckung und biomedizinsches Datenmanagement werden Hauptforschungsthemen bleiben.
- Datenwachstum in drei Dimensionen wird extreme neue Herausforderungen im Data-Mining in Genetik und Medizin generieren:
  - Datensätze von Individuen einer ganzen Population
  - Lebenslange Aufzeichnung des Gesundheitszustandes
  - Höchstaufgelöste Information über den Gesundheitszustand
- Wie können diese Daten ausgewertet und genutzt werden?
- Viele Zweige der Lebenswissenschaften stehen vor sehr ähnlichen oder analogen Problemen.

Viele Möglichkeiten für Data-Mining in den Lebenswissenschaften

Vielen Dank

- Marie-Curie-Initial Training Network for 'Machine Learning for Personalized Medicine' (mlpm.eu, 2013-2016)
- Starting Grant (Temporäre ERC-Ersatzmaßnahmen des SNSF)
- Alfried-Krupp-Förderpreis für junge Hochschullehrer
- SPHN-PHRT Driver Project 'Personalized Swiss Sepsis Study'

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

Referenzen I


C. E. Bonferroni, *Teoria statistica delle classi e calcolo delle probabilita* (1936). Published: (Pubbl. d. R. Ist. Super. di Sci. Econom. e Commerciali di Firenze. 8) Firenze: Libr. Internaz. Seeber. 62 S. (1936).


R. Ferrer, *et al.*, *Critical Care Medicine* **42**, 1749 (2014).


D. G. Grimm, *et al.*, *The Plant Cell* **29**, 5 (2017).


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-López, *et al.*, *KDD* (2015), pp. 725–734.


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


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


M. Osthoff, *et al.*, *Clinical Microbiology and Infection: The Official Publication of the European Society of Clinical Microbiology and Infectious Diseases* **23**, 78 (2017).

Referenzen II


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


C. W. Seymour, *et al.*, *JAMA* **315**, 762 (2016).


M. Singer, *et al.*, *JAMA* **315**, 801 (2016).


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


R. E. Tarone, *Biometrics* **46**, 515 (1990).


A. Terada, *et al.*, *Proceedings of the National Academy of Sciences* **110**, 12996 (2013).


P. H. Westfall, *et al.*, *Biometrics* **49**, 941 (1993).