Freie Universität Berlin

MAX-PLANCK-GESELLSCHAFT

AlgoBio WS 16/17
Functional enrichment

Annalisa Marsico

11.01.2017

Exakter Test nach Fisher

Der exakter Test nach Fisher ist ein statistischer Signifikanztest, der bei der Analyse von Kontingenztabellen verwendet wird.
Er wird verwendet, wenn die Probengrößen klein sind.

Er ist nützlich für kategorische Daten, die aus der Klassifizierung von Objekten auf zwei Arten erhalten werden.
Er wird verwendet, um die Signifikanz der Assoziation zwischen zwei Arten von Klassifikationen zu schätzen.

Beispiel:
Erste Klassifizierung: diff. expr. Gene / nicht-diff. expr. Gene
Zweite Klassifizierung: Gene mit einer spezifischen Funktion (z.B. DNA repair / Gene mit anderen Funktionen)

Exakter Test nach Fisher

Wir wollen wissen, ob es eine Assoziation zwischen diesen beiden Klassifikationen gibt.

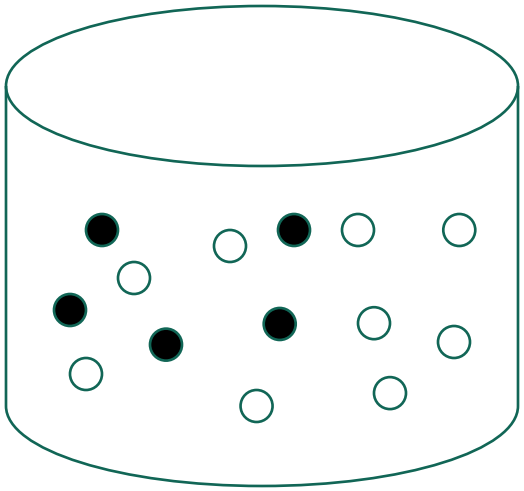
Dies wird in einer Kontingenztabelle dargestellt, und unter der Nullhypothese der Unabhängigkeit können die Zahlen in den Zellen der Tabellen mit einer *hypergeometrischen Verteilung* modelliert werden.

Wir müssen die hypergeometrische Verteilung einführen.

Die hypergeometrische Verteilung

Die hypergeometrische Verteilung ist eine diskrete Wahrscheinlichkeitsverteilung, die die Wahrscheinlichkeit von ***k*** Erfolgen in ***n*** Ziehungen, ohne Ersatz, aus einer endlichen Population der Größe ***N***, die genau ***m*** Erfolge enthält, beschreibt.

Urnenmodell:



N = Gesamtzahl der Kugeln
m = Anzahl der weißen Kugeln
n = Anzahl der Ziehungen ohne Ersatz
X = Anzahl der gezogenen weißen Kugeln

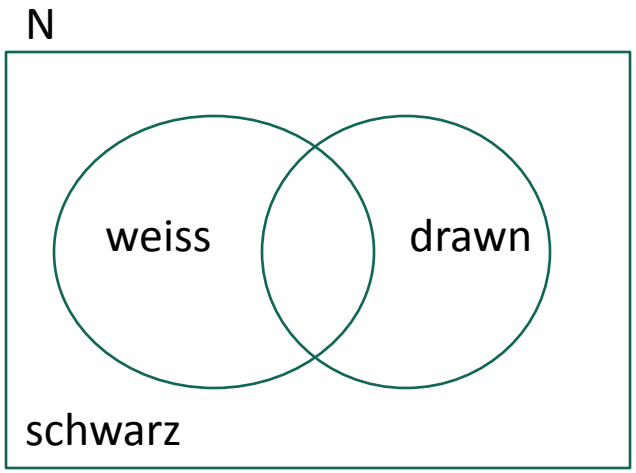
Wir wollen P(X=k) berechnen

Die hypergeometrische Verteilung

Die hypergeometrische Verteilung ist eine diskrete Wahrscheinlichkeitsverteilung, die die Wahrscheinlichkeit von ***k*** Erfolgen in ***n*** zeichnet, ohne Ersatz, aus einer endlichen Population der Größe ***N***, die genau ***m*** Erfolge enthält, beschreibt.

Urnenmodell:

N = Gesamtzahl der Kugeln
m = Anzahl der weißen Kugeln
n = Anzahl der Ziehungen ohne Ersatz
X = Anzahl der gezogenen weißen Kugeln



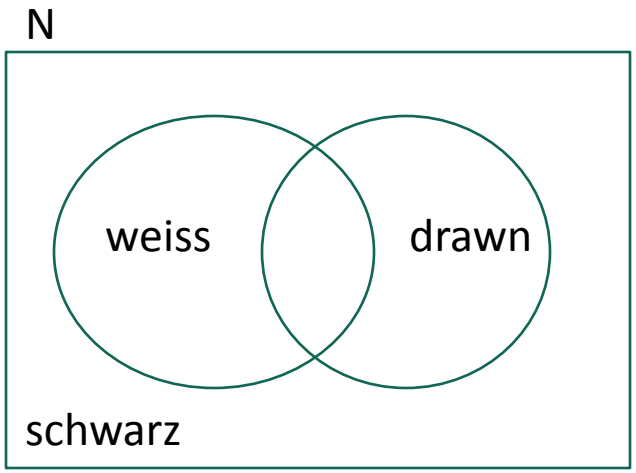
	drawn	not drawn	
weiss (Erfolg)	**k**	**m-k**	**m**
schwarz (Misserfolg)	**n-k**	**N-n-(m-k)**	**N-m**
	n	**N-n**	**N**

Die hypergeometrische Verteilung

Die hypergeometrische Verteilung ist eine diskrete Wahrscheinlichkeitsverteilung, die die Wahrscheinlichkeit von k Erfolgen in n zeichnet, ohne Ersatz, aus einer endlichen Population der Größe N, die genau m Erfolge enthält, beschreibt.

Urnenmodell:

N = Gesamtzahl der Kugeln
m = Anzahl der weißen Kugeln
n = Anzahl der Ziehungen ohne Ersatz
X = Anzahl der gezogenen weißen Kugeln



$$P(X = k) = \frac{\binom{m}{k}\binom{N-n}{m-k}}{\binom{N}{n}}$$

Wahrscheinlichkeit um genau k weiße Kugeln zu ausziehen

Exakter Test nach Fisher

Null-Hypothese: Es gibt keine Assoziation zwischen den beiden Klassen

Die Wahrscheinlichkeit, die Daten unter der Nullhypothese zu beobachten, ist

$$P(X \leq k) = \sum_{k=0}^{n} \frac{\binom{m}{X}\binom{N-n}{m-X}}{\binom{N}{n}}$$

χ^2 – Test für statistische Unabhängigkeit

Er testet, wie wahrscheinlich ist, dass eine beobachtete Assoziation zwischen Klassen durch Zufall beobachtet wird

$\boldsymbol{H_0}$: Es gibt keine Assoziation zwischen Klassen
$\boldsymbol{H_1}$: Es gibt eine Abhängigkeit zwischen Klassen

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

$E_i \equiv E_{r,c} = (n_r * n_c)/N$ sind die erwarteten Zählungen in jeder Zelle der Tabelle; n_r Gesamtzahl der Zeilen; n_r Gesamtzahl der Spalten.

χ^2 – Test für statistische Unabhängigkeit

Er testet, wie wahrscheinlich ist, dass eine beobachtete Assoziation zwischen Klassen durch Zufall beobachtet wird

$\boldsymbol{H_0}$: Es gibt keine Assoziation zwischen Klassen
$\boldsymbol{H_1}$: Es gibt eine Abhängigkeit zwischen Klassen

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

Die Freiheitsgrade einer χ^2 Statistik mit *c* Spalten and *r* Zeilen sind

$$df = (r-1)(c-1)$$

Gene Ontology für Datenanalyse

Ontologien bieten kontrollierte, konsistente Vokabulare,
um Konzepte und Beziehungen zu beschreiben,
wodurch Wissensaustausch ermöglicht wird

Gene Ontologien (GO) sind Ontologien für die Molekularbiologie –
ein kontrolliertes Vokabular zur Beschreibung von Funktionen
und Prozessen von Genen und Genprodukten

Ziele des GO Projekts

1. Design strukturierter Vokabeln, die Aspekte der Molekularbiologie beschreiben

2. Unterstützung Annotation von Gen-Produkten mit spezifischen Begriffen

3. Bereitstellung von Datenbankzugriff über diese allgemeinen Begriffe zu Genproduktanmerkungen und zugehörigen Sequenzen.

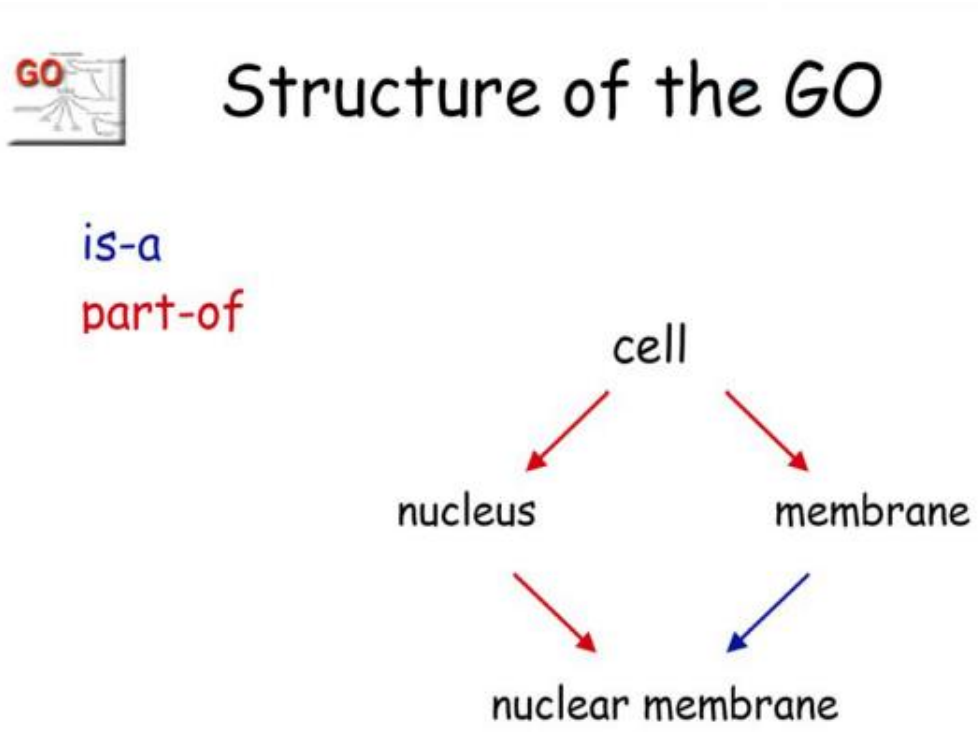
Schlüsselaspekte

1. Jedes Konzept muss sorgfältig definiert werden

2. Die minimale Datenstruktur einer Ontologie ist ein Directed Acyclic Graph (DAG)

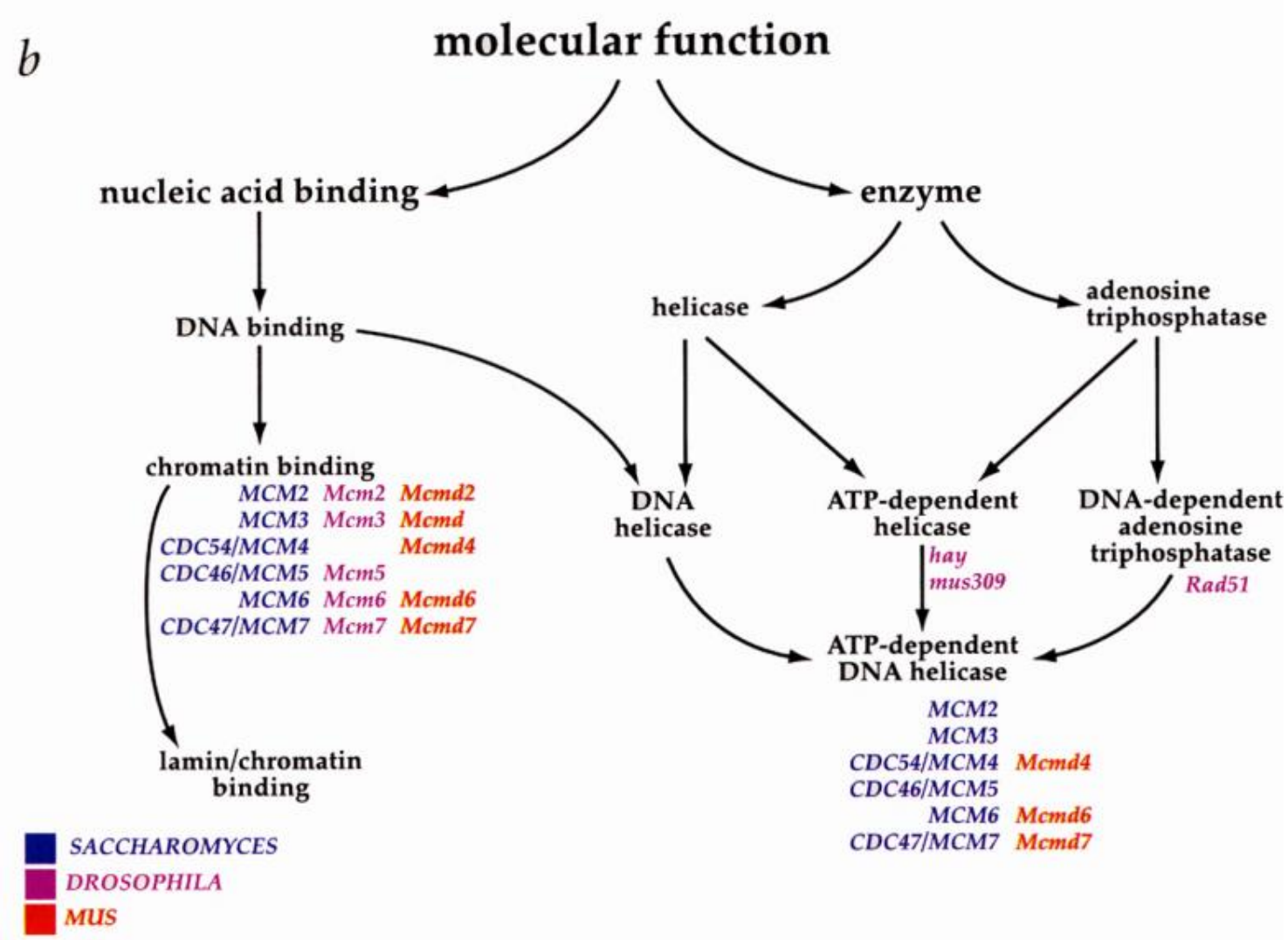
3. GO ist keine Datenbank von Genprodukten, Proteinen, Domänen und Motiven. Es definiert auch keine evolutionären Beziehungen.

Ontologie Gebäude



1. Directed Acyclic Graph (DAG): Jedes Kind kann einen oder mehrere Eltern haben
2. Beziehungen zwischen Begriffen werden definiert
3. Alle Begriffe sind mit einer ID verknüpft

Beispiel für eine molekulare Funktion



http://amigo.geneontology.org/

Referenzen

Quantil-Normalisierung

- “A comparison of normalization methods for high density oligonucleotide array data based on variance and bias”, Bolstad et al., *Bioinformatics* 2003

RNA-Sequenzierung

- “Mapping and quantifying mammalian transcriptomes by RNA-Seq”, Mortazavi et al., *Nature Methods* 2008
- https://www.youtube.com/watch?v=C8RNvWu7pAw (Statistics for Genomics: Introduction to RNA-Seq)

PCA

- https://www.youtube.com/watch?v=WGXcIlJhc58 (PCA problem formulation)
- https://www.youtube.com/watch?v=N5ynBdHqnGU (PCA algorithm)