V9 funktionelle Annotation

- Analyse von Gen-Expression
- Funktionelle Annotation: Gene Ontology (GO)
- Signifikanz der Annotation: Hypergeometrischer Test
- Annotationsanalysen z.B. mit NIH-Tool DAVID
- Ähnlichkeit von GO-Termen automatisch bestimmen
- OMIM-Datenbank

In Vorlesung 9 beschäftigen wir uns mit der downstream-Analyse von experimentellen Rohdaten. Ein typisches Transkriptomik- oder Proteomik-Experiment ergibt eine Menge von hoch- oder runterregulierten Genen. Funktionelle Annotation bedeutet, dass man aus diesem Ergebnis biologische Einblicke gewinnen möchte. Oft testet man dafür mit dem hypergeometrischen Test die Anreicherung von funktionellen Ausdrücken aus der Gene Ontology oder die Mitgliedschaft in biochemischen Pfaden in der KEGG Datenbank (https://www.genome.jp/kegg/) oder in Reactome (https://reactome.org/).

Ausgangslage

Daten aus Microarray-Analyse wurden ursprünglich als sehr „verrauscht“ angesehen.

Mittlerweile wurden jedoch sowohl die experimentellen Schritte wie auch die Datenauswertung gründlich verfeinert.

Microarray-Analyse ist daher heute eine (zwar teure, aber zuverlässige) Routine-Methode, die in allen großen Firmen verwendet wird.

Heute wird die MA-Analyse zunehmend durch RNA-seq ersetzt.

Die Datenaufbereitung kann in beiden Fällen folgende Schritte enthalten: Normalisierung, Logarithmierung, Clustering, evtl. Ko-Expressionsanalyse, **Annotation der Genfunktion (Inhalt von V9).**

Sehr wichtig ist es immer, die Signifikanz der Ergebnisse zu bewerten.

Gentleman et al. Genome Biology 5, R80 (2004)

Wir schließen in dieser Vorlesung unmittelbar an das Ergebnis von Vorlesung 8 an. Dort fehlte noch der letzte Punkt, die „Annotation der Genfunktion“:

Beispiel: differentielle Gen-Expression für ALL-Patienten

Input:

Genexpressionsdaten für Patienten mit akuter lymphatischer Leukämie (ALL).

Alle ALL-Patienten haben chromosomale Veränderungen.

Der Therapieerfolg ist jedoch sehr unterschiedlich.

Hintergrundinformation:

- Eine Gruppe von Patienten (ALL1/AF4) hat eine genetische Translokation zwischen den Chromosomen 4 und 11.
- Eine zweite Gruppe von Patienten (BCR/ABL) hat eine genetische Translokation zwischen den Chromosomen 9 und 22.
- Die Krankheitsursachen + optimale Therapie können für die beiden Gruppen verschieden sein.

Ziel:

Identifiziere Gene, die zwischen den beiden Gruppen differentiell exprimiert werden.

Beispiel für die Anwendung der `Bioconductor`-Software (siehe Ref unten, bisher mehr als 12000 mal zitiert).

Gentleman et al. Genome Biology 5, R80 (2004)

Zur Motivation verwenden wir ein Beispiel zu Patienten mit akuter lymphatischer Leukämie (ALL) aus dem Bioconductor-Paper von 2004 (mehr als 12000 mal zitiert, https://genomebiology.biomedcentral.com/articles/10.1186/gb-2004-5-10-r80).

Bioconductor (https://www.bioconductor.org/) ist ein ganz wichtige Ressource für viele Bereiche der Bioinformatik. Es stellt knapp 2000 Programme zur Verfügung und definiert sich als

„Bioconductor provides tools for the analysis and comprehension of high-throughput genomic data. Bioconductor uses the R statistical programming language, and is open source and open development. “

ALL wird durch genetische Veränderungen verursacht. Es scheint jedoch 2 verschiedene Ursache zu geben, eine genetische Translokation zwischen den Chromosomen 4 und 11 bzw. eine genetische Translokation zwischen den Chromosomen 9 und 22. Wie unterscheiden sich die Expressionsmuster beider Patientengruppen?

Auswahl der differentiell exprimierten Gene

```
> f <- factor(as.character(eset$mol))
> design <- model.matrix(~f)
> fit <- lmFit(eset, design)
> fit <- eBayes(fit)
> topTable(fit, coef = 2)
```

Bioconductor Kommandos

	ID	M	A	t	*p*-value	B
1016	1914_at	−3.1	4.6	−27	5.9e-27	56
7884	37809_at	−4.0	4.9	−20	1.3e-20	44
6939	36873_at	−3.4	4.3	−20	1.8e-20	44
10865	40763_at	−3.1	3.5	−17	7.2e-18	39
4250	34210_at	3.6	8.4	15	3.5e-16	35
11556	41448_at	−2.5	3.7	−15	1.8e-15	34
3389	33358_at	−2.3	5.2	−13	3.3e-13	29
8054	37978_at	−1.0	6.9	−10	6.5e-10	22
10579	40480_s_at	1.8	7.8	10	9.1e-10	21
330	1307_at	1.6	4.6	10	1.4e-09	21

Figure 1
Limma analysis of the ALL data. The leftmost numbers are row indices, ID is the Affymetrix HGU95av2 accession number, M is the log ratio of expression, A is the log average expression, and B is the log odds of differential expression.

Differential expression (D.E.) = $\log_2(R) / \log_2(G)$

Log ratio M : $M = \log_2(R) / \log_2(G)$; $M = 1$ -> zweifach D.E.

Wie signifikant ist dies? -> bewerte mit statistischem Test.

Vergleiche Gen-Expression in den beiden Gruppen.

Fokussiere auf Gene mit stark unterschiedlicher Expression.

Wähle z.B. alle Gene mit p-Wert < 0.05 aus.

Es bleiben 165 Gene übrig.

Gentleman et al. Genome Biology 5, R80 (2004)

165 Gene wurden als differentiell exprimiert ($p < 0.05$) zwischen beiden Patientengruppen identifiziert.

Differentielle Gen-Expression als Heatmap visualisieren

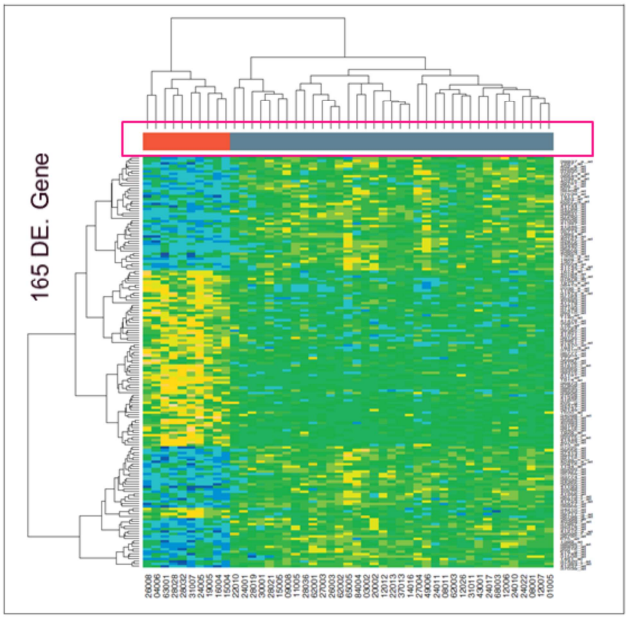


Figure 2
Heat map (produced by the Bioconductor function heatmap()) of the ALL leukemia data.

Mit einem Abstandsmaß und einem Cluster-Algorithmus werden die Ähnlichkeiten zwischen den Patienten (x-Achse) und den einzelnen Gene (y-Achse) erfasst.

Die beiden Patienten-Gruppen haben deutlich unterschiedliche Expressionsprofile (rot/grau).

Gelb: stark hochreguliert

Blau: stark runterreguliert

...ntleman et al. Genome Biology 5, R80 (2004)

Wie im lila Kasten markiert, werden die 47 gezeigten Patienten aufgrund der Expressionsprofile der 165 DE Gene in eine rote Gruppe mit 10 Patienten und in eine graue Gruppe mit 37 Patienten geclustert. Diese beiden Patienten-Gruppen entsprechen den Gruppen mit unterschiedlichen chromosomalen Translokationen. Man könnte also die Expressionsprofile als diagnostischen Biomarker für die genetische Veränderung verwenden.

Zuordnung von Gen-Funktion

```
> f <- factor(as.character(eset$mol))
> design <- model.matrix(~f)
> fit <- lmFit(eset, design)
> fit <- eBayes(fit)
> topTable(fit, coef = 2)
```

Bioconductor Kommandos

	ID	M	A	t	*p*-value	B
1016	1914_at	−3.1	4.6	−27	5.9e-27	56
7884	37809_at	−4.0	4.9	−20	1.3e-20	44
6939	36873_at	−3.4	4.3	−20	1.8e-20	44
10865	40763_at	−3.1	3.5	−17	7.2e-18	39
4250	34210_at	3.6	8.4	15	3.5e-16	35
11556	41448_at	−2.5	3.7	−15	1.8e-15	34
3389	33358_at	−2.3	5.2	−13	3.3e-13	29
8054	37978_at	−1.0	6.9	−10	6.5e-10	22
10579	40480_s_at	1.8	7.8	10	9.1e-10	21
330	1307_at	1.6	4.6	10	1.4e-09	21

Figure 1
Limma analysis of the ALL data. The leftmost numbers are row indices, ID is the Affymetrix HGU95av2 accession number, M is the log ratio of expression, A is the log average expression, and B is the log odds of differential expression.

Links gezeigt ist dieselbe Tabelle wie zwei Folien zuvor.

Nun interessiert uns, welche Funktionen diese Gene in der Zelle ausüben.

Verwende dazu Informationen aus der `Gene Ontology` über diese Gene.

Gentleman et al. Genome Biology 5, R80 (2004)

Heute interessiert uns aber, ob wir mehr biologischen Einblick in die 165 DE-Gene bekommen können.

Die Gene Ontology (GO)

Ontologien sind strukturierte Vokabulare.

Die Gene Ontology hat 3 Bereiche:
- biologischer Prozess (BP)
- molekulare Funktion (MF)
- zelluläre Komponente (Lokalisation).

Hier ist ein Teil der BP-Ontologie gezeigt.

Oben ist der allgemeinste Ausdruck (root).

Rot: Blätter des Baums (sehr spezifische GO-Terme)

Grün: gemeinsame Vorgänger.

Blau: andere Knoten.

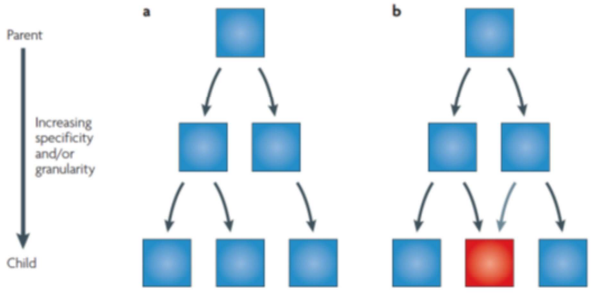
Linien: „Y ist in X enthalten“-Beziehungen

Dissertation Andreas Schlicker (UdS, 2010)

Die Gen-Ontologie (http://geneontology.org/) besitzt 3 Bereiche: biologische Prozesse (BP), molekulare Funktion (chemische Details, MF), sowie die zelluläre Komponente, in der man das kodierte Protein findet.

Jeder Bereich startet mit einem Wurzelknoten oben. Darunter sind Kind-Knoten angeordnet, deren Funktion je nach „Tiefe“ der Anordnung zunehmend stärker spezialisiert ist. Alle Kind-Knoten erben die Funktionen ihrer Vorfahren.

Baum vs. azyklische Graphen



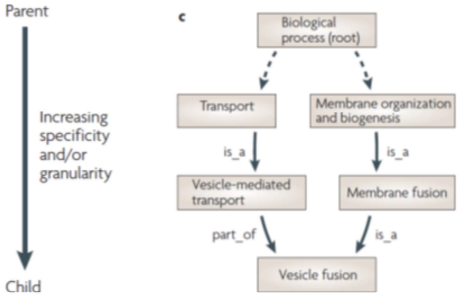
a | Einfacher **Baum**, in dem jedes Kind genau ein Elternteil hat.
Die Kanten sind vom Elternteil zum Kind gerichtet.

b | In einem **gerichteten azyklischen Graph** (DAG) kann jedes Kind ein oder mehrere Eltern haben. Hier besitzt z.B. der rote Knoten 2 Eltern.
Azklisch heisst, dass der Graph keine gerichteten Zyklen enthält.

Rhee et al. (2008) Nature Rev. Genet. 9: *509*

Die Gene Ontology besitzt die Topologie eines gerichteten azyklischen Graphen, wobei Kind-Knoten mehrere Eltern-Knoten besitzen können.

Die Gene Ontology ist ein directed acyclic graph



Der Knoten *vesicle fusion* besitzt in der BP Ontologie mehrere Eltern.

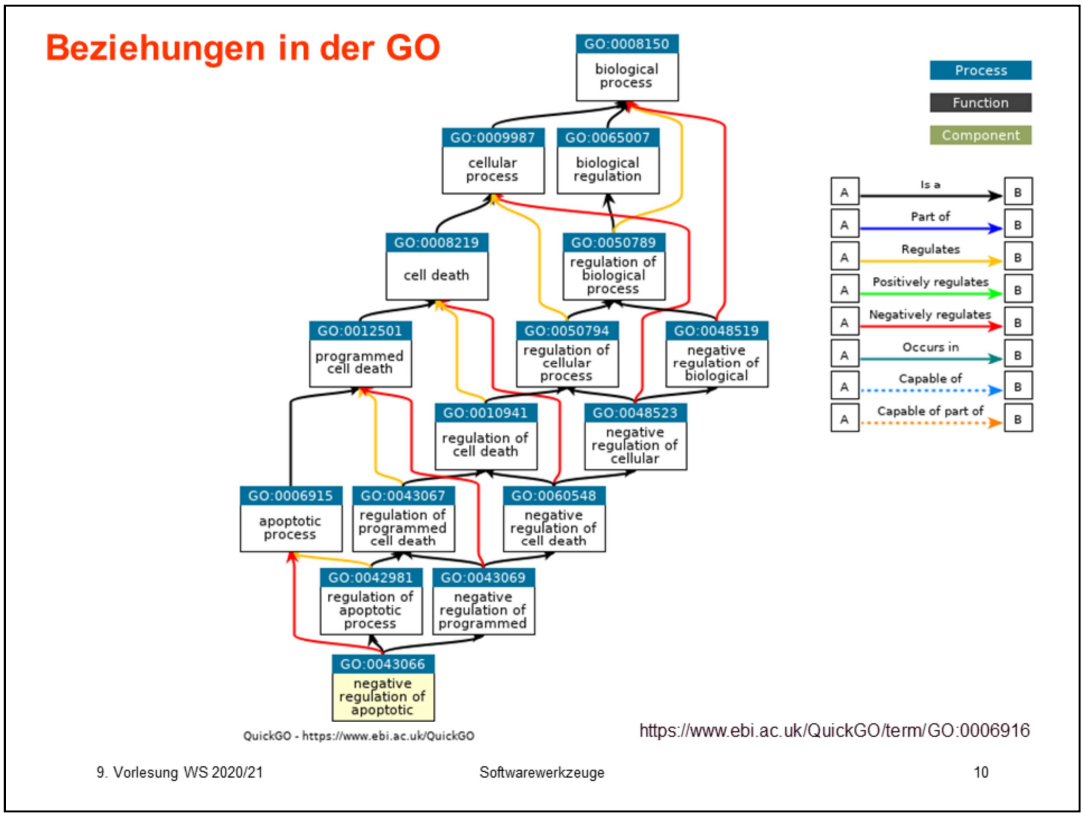
Gestrichelte Kanten: andere dazwischen liegende Knoten sind nicht gezeigt.
Root : keine Kanten zeigen zu diesem Knoten hin. Er hat mindestens ein Kind.
Leaf node : ein Endknoten ohne Kinder.

Tiefe eines Knotens: Länge des längsten Pfades von root zu diesem Knoten.
Höhe eines Knotens: Länge des längsten Pfades von diesem Knoten zu einem Endknoten.

Rhee et al. (2008) Nature Rev. Genet. 9: *509*

Dieses Beispiel zeigt, dass der Blattknoten „vesicle fusion“ (der z.B. in Endozytose und Exozytose und im vesikulären Transport zwischen verschiedenen Kompartments vorkommt) zwei Äste mit Vorläuferknoten besitzt.

Der linke Ast konzentriert sich auf die Vesikel, der rechte Ast auf die Membranprozesse. Obwohl die Pfeile in dieser Abbildung nach unten zeigen, sollten sie in entgegengesetzter Richtung gelesen werden. Z.B. ist „vesicle fusion“ ein „Teil_von“ „vesicle-mediated transport“, nicht umgekehrt.



Hier sind die Pfeile korrekt nach oben gerichtet. Es gibt 5 verschiedene Beziehungen zwischen Knoten, siehe rechts oben. Alle Ausdrücke (außer der Wurzelknoten) besitzen eine „is a“ Beziehung zu einem anderen Term, z.B. ist „GO:1904659:Glucose Transport” auch ein “GO:0015749:Monosaccharid Transport”. Die Gene Ontology verwendet mehrere andere Beziehungen, siehe rechts oben, z.B. reguliert „GO:0043069:negative regulation of programmed cell death” den “GO:0012501:programmed cell death”. „Is a“ Beziehungen verbinden Terme mit dem direkt darüberliegenden Term. Regulatorische Beziehungen können Ebenen überspringen.

Gene Ontology (GO) - Konsortium

Berkeley Bioinformatics Open-source Project (BBOP)
British Heart Foundation - University College London (BHF-UCL)
dictyBase
EcoliWiki
FlyBase
GeneDB
UniProtKB-Gene Ontology Annotation @ EBI (UniProtKB-GOA)
GO Editorial Office at the European Bioinformatics Institute
Gramene
Institute of Genome Sciences, Univ. of Maryland
J Craig Venter Institute
Mouse Genome Informatics (MGI)
Rat Genome Database (RGD)
Reactome
Saccharomyces Genome Database (SGD)
The Arabidopsis Information Resource (TAIR)
WormBase
The Zebrafish Information Network (ZFIN)

Der Aufbau der Gene Ontology wird seit 2002 vom NIH mit etwa 3 Million US$ pro Jahr unterstützt (https://projectreporter.nih.gov/project_info_history.cfm?aid=9209989&icde=0). Viele verschiedene Konsortien tragen zur Gene Ontology bei. Man merkte nämlich, dass Gen-Funktionen oft zwischen verwandten Organismen konserviert sind. Daher ist es unpraktisch, die Gen-Funktionen in Organismen-spezifischen Datenbanken zu sammeln.

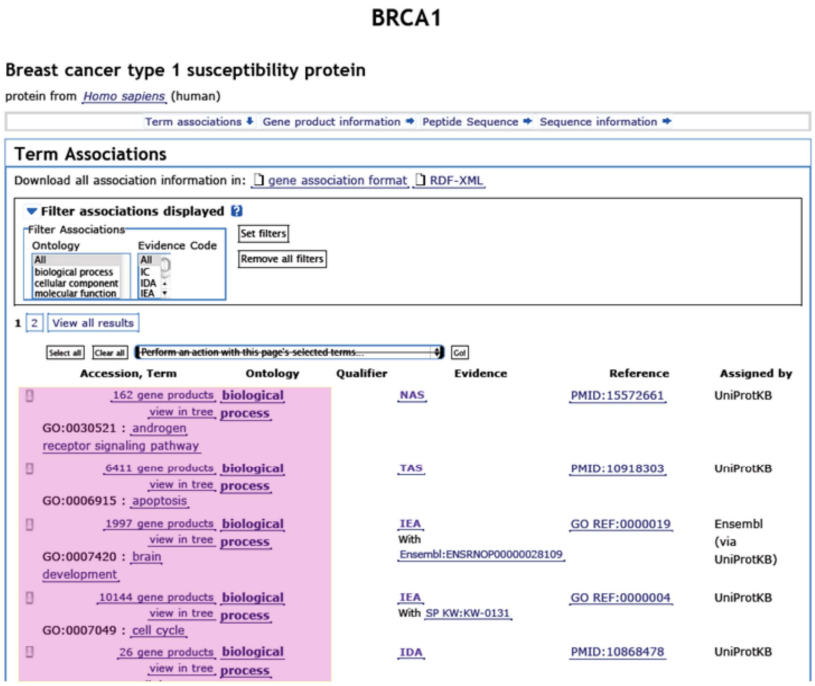
Woher stammen die Gene Ontology Annotationen?

Ontology	Property	Value
	Valid terms	44411 (Δ = -97)
	Obsoleted terms	2947 (Δ = 23)
	Merged terms	2056 (Δ = 91)
	Biological process terms	29112
	Molecular function terms	11118
	Cellular component terms	4181
Annotations	**Property**	**Value**
	Number of annotations	7,975,639
	Annotations for biological process	3,069,526
	Annotations for molecular function	2,455,089
	Annotations for cellular component	2,451,024
	Annotations for evidence PHYLO	4,163,423
	Annotations for evidence IEA	1,978,576
	Annotations for evidence EXP	759,654
	Annotations for evidence OTHER	791,743
	Annotations for evidence ND	241,978
	Annotations for evidence HTP	40,265
	Number of annotated scientific publications	159,963

http://geneontology.org/stats.html

Diese Statistik stammt von der Gene Ontology Webseite, Stand Juni 2020.

Beispiel: GO-Annotation für humanes BRCA1-Gen



Einzelne GO-Terme, mit denen das Brustkrebs-Gen BRCA1 annotiert ist.

Dies ist ein Beispiel für das „Brustkrebs-Gen“ BRCA1. BRCA1 gehört zum „androgen receptor signaling pathway“, zu Apoptose, brain development, Zellzyklus etc. Aufgelistet unter „Reference“ ist die Quelle für die Annotation, also aufgrund einer Publikation (PMID steht für Pubmed Identifier) oder durch IEA.

Signifkanz von GO-Annotationen

Sehr **allgemeine GO-Terme** wie z.B. “cellular metabolic process“ werden vielen Genen im Genom zugeordnet.

Sehr **spezielle Terme** gehören jeweils nur zu wenigen Genen.

Man muss also vergleichen, wie **signifikant** das Auftreten jedes GO-Terms in einer Testmenge an Genen im Vergleich zu einer zufällig ausgewählten Menge an Genen derselben Größe ist.

Dazu verwendet man meist den **hypergeometrischen Test**.

Dissertation Andreas Schlicker (UdS, 2010)

Wenn wir nun den Ergebnissen der Differentielle-Expression-Analyse biologische Bedeutungen zuordnen möchten, ist es vermutlich nicht hilfreich, wenn die Hälfte aller hochregulierten Gene einen „metabolischen Prozess“ unterstützen. Das ist nämlich eine sehr allgemeine Funktion.

Wenn jedoch einige der 165 DE Gene die Annotation „**purine nucleotide biosynthetic process** “ tragen würden, wäre das ein viel spezifischerer GO term (0006164).

Es gibt aktuell 202 menschliche Gene mit dieser Annotation, siehe http://amigo.geneontology.org/amigo/search/bioentity?q=*:*&fq=regulates_closure:%22GO:0006164%22&sfq=document_category:%22bioentity%22 und dann nach Homo Sapiens filtern. Wir müssen also die statistische Signifikanz bestimmen, dass z.B. von diesen 202 Genen 12 in ALL hochreguliert sind (hypothetisches Beispiel).

Vorbemerkung

Zieht man aus einer Urne mit n Kugeln insgesamt k Kugeln **ohne Beachtung der Reihenfolge**, so gibt es hierfür genau

$$\frac{n\cdot(n-1)\cdot(n-2)\cdot\ldots\cdot(n-k+1)}{k!}=\frac{n!}{k!\cdot(n-k)!}=\binom{n}{k}$$ Möglichkeiten

Zu Beginn liegen n Kugeln in der Urne. Es gibt also n Möglichkeiten.

Bei der zweiten Ziehung gibt es nur noch $n - 1$ Möglichkeiten.

Man zieht insgesamt k Kugeln. Bei der letzten hat man noch $n - k + 1$ Möglichkeiten, da man noch bereits $k - 1$ Kugeln fehlen.

Bei der Lottoziehung „6 aus 49“ wird die Reihenfolge nicht beachtet. Nach der Ziehung werden alle gezogenen Zahlen in aufsteigender Reihenfolge genannt. Es ist also egal ob man die 5 vor oder nach der 19 gezogen hat.
Insgesamt kann die kleinste Kugel an k Positionen auftauchen, die zweitkleinste Kugel an k-1 Positionen … deshalb dividiert man durch k!

http://www.schule-bw.de/

Bei der Annotation von Gene Ontology-Termen werden wir die Reihenfolge (ihre Numerierung) der Gene nicht beachten. Deshalb gilt die Analogie zur Lottoziehung von Zahlen. Um vom linken Term in der Gleichung auf den mittleren Term zu gelangen, multiplizieren wir oben und unten mit (n – k)! Diesen mittleren Ausdruck definieren wir als das mathematische Symbol „n über k“ (rechter Term in der Gleichung)

Hypergeometrischer Test

$$\text{p-Wert} = \sum_{i=k_\pi}^{min(n,K_\pi)} \frac{\binom{K_\pi}{i}\binom{N-K_\pi}{n-i}}{\binom{N}{n}}$$

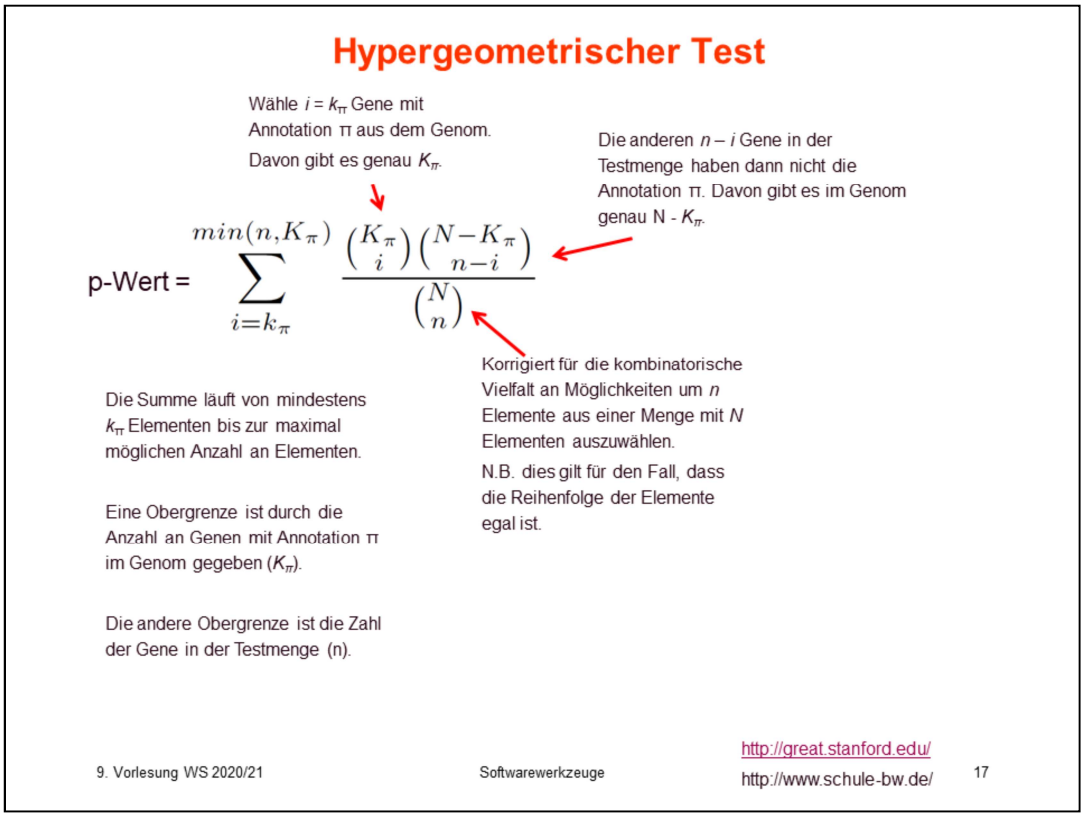
Der hypergeometrische Test ist ein statistischer Test, der z.B. überprüft, ob in einer vorgegebenen Testmenge an Genen eine biologische Annotation π gegenüber dem gesamten Genom statistisch signifikant angereichert ist.

- Sei *N* die Anzahl an Genen im Genom.
- Sei *n* die Anzahl an Genen in der Testmenge.
- Sie K_π die Anzahl an Genen im Genom mit der Annotation π.
- Sei k_π die Anzahl an Genen in der Testmenge mit der Annotation π.

Der hypergeometrische p-Wert drückt die Wahrscheinlichkeit aus, dass k_π oder mehr **zufällig** aus dem Genom ausgewählte Gene auch die Annotation π haben.

http://great.stanford.edu/

Man benutzt oft den hypergeometrischen Test um einen p-Wert für die statistische Signifikanz von GO-Termen zu erhalten. Der Bruch drückt aus, welcher Anteil an kombinatorischen Möglichkeiten auf zufällige Weise zu einer Verteilung von Genen mit/ohne Annotation π wie in der Testmenge führt.

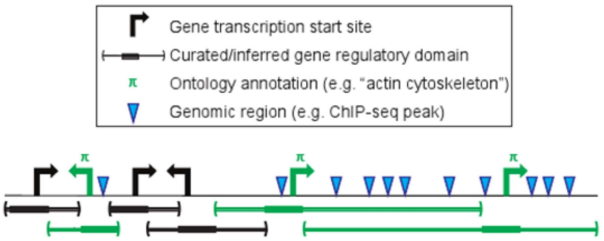


Man muss die Formel folgendermassen interpretieren:

Im Nenner steht die kombinatorische Vielfalt um n Gene aus einer grossen Menge mit N Genen auszuwählen. Im Zähler stehen zwei Terme. Wir betrachten einen bestimmten GO-Term π. Insgesamt sind K_π Gene in der Gesamtmenge von N Genen mit diesem Term annotiert. Der linke Terme drückt die Anzahl an Möglichkeiten aus, i Gene mit dieser Annotation aus K_π auszuwählen. In unserem Fall enthält die Testmenge i = k_π solcher Gene. Wir summieren über die Anzahl an Möglichkeiten, dass mindestens k_pi Gene diese Annotation besitzen. Die obere Schranke der Summe ist daher zum einen die Grösse der Testmenge (n) und zum anderen die Anzahl an Genen mit Annotation π in der Gesamtmenge. Der zweite Term berücksichtigt die anderen Gene in der Testmenge, die die Annotation π nicht besitzen. Davon gibt es n – i. Der Bruch drückt also aus, welcher Anteil an kombinatorischen Möglichkeiten auf zufällige Weise zu einer Verteilung von Genen mit/ohne Annotation π wie in der Testmenge führt. Wenn es viele solche Fälle gibt, wäre der p-Wert recht hoch und die statistische Signifikanz daher gering.

Beispiel

p-Wert = $\sum_{i=k_\pi}^{min(n,K_\pi)} \frac{\binom{K_\pi}{i}\binom{N-K_\pi}{n-i}}{\binom{N}{n}}$



$$p = \sum_{i=3}^{min(3,3)} \frac{\binom{3}{3}\binom{3}{0}}{\binom{6}{3}} = \frac{1 \cdot 1}{\frac{6 \cdot 5 \cdot 4}{1 \cdot 2 \cdot 3}} = \frac{1}{20}$$

Hypergeometric test over genes
N = 6 total genes
K_π = 3 genes annotated with π
n = 3 genes with an associated genomic region
k_π = 3 genes annotated and with a genomic regio
P-value = 0.05

Frage: ist die Annotation π in der Testmenge von 3 Genen signifikant angereichert?

Ja! p = 0.05 ist (knapp) signifikant.

http://great.stanford.edu/

Dies ist ein sehr einfaches Beispiel, das man mit der Hand ausrechnen kann. „3 über 3“ und „3 über 0“ sind jeweils als 1 definiert.

Anwendung auf ALL-Beispiel

```
> ll <- mget(geneNames(esetSel),
  hgu95av2LOCUSID)
> ll <- unique(unlist(ll))
> mf <- as.data.frame(GOHyperG(ll))[, 1:3]
> mf <- mf[mf$pvalue < 0.01, ]
> mf
```

Bioconductor Kommandos

	p-values	goCounts	intCounts
GO:0045012	1.1e-08	12	6
GO:0030106	6.4e-03	8	2
GO:0004888	7.0e-03	523	16
GO:0004872	9.7e-03	789	21

```
> as.character(unlist(mget(row.names(mf),
  GOTERM)))
[1] "MHC class II receptor activity"
[2] "MHC class I receptor activity"
[3] "transmembrane receptor activity"
[4] "receptor activity"
```

Figure 3
Hypergeometric analysis of molecular function enrichment of genes selected in the analysis described in Figure 1.

Die signifikanteste Anreicherung ergibt sich für MHC Klasse 2 Rezeptoraktivität.

6 von 12 Genen im Genom mit dieser Annotation sind in den 2 ALL-Klassen differentiell exprimiert.

Gentleman et al. Genome Biology 5, R80 (2004)

Wenn man nun die Anreicherung von GO-Termine in den 165 differentiell exprimierten GO-Termen für die 2 Gruppen von ALL-Patienten berechnet, sieht man dass die Major Histocompatibility Complex (MHC)-Proteine des Immunsystems in den beiden Gruppen unterschiedlich stark exprimiert werden. Insgesamt gibt es 12 Gene mit Annotation MHC2 im Genom. Davon sind 6 differentiell exprimiert. Bei den MHC1 sind es 2 von 8. Somit hat man nun einen Ansatzpunkt um die evtl. unterschiedliche Erfolgschance von Krebstherapien in beiden Patientengruppen nachvollziehen zu können.

NIH Tool David: Tool für Annotation der Genfunktion

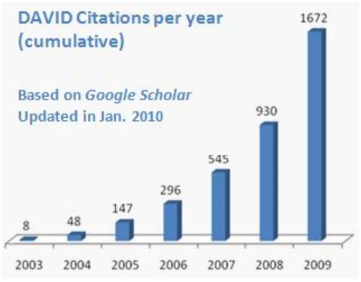
PROTOCOL

Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources

Da Wei Huang[1,2], Brad T Sherman[1,2] & Richard A Lempicki[1]

[1]Laboratory of Immunopathogenesis and Bioinformatics, Clinical Services Program, SAIC-Frederick Inc., National Cancer Institute at Frederick, Frederick, Maryland 21702, USA. [2]These authors contributed equally to this work. Correspondence should be addressed to R.A.L. (rlempicki@mail.nih.gov) or D.W.H. (huangdawei@mail.nih.gov)

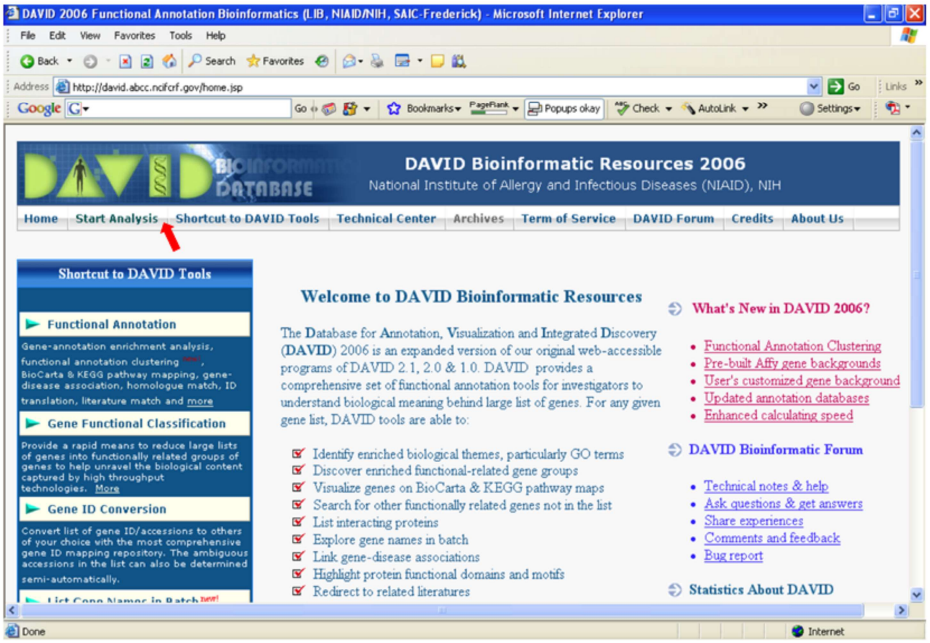
Published online 18 December 2008; doi:10.1038/nprot.2008.211



Huang et al. Nat. Protocol 4, 44 (2009)

Es gibt viele Online-Tools, mit denen man funktionelle Annotation von Gen-Mengen durchführen kann. Eines dieser Tool heisst DAVID vom NIH (https://pubmed.ncbi.nlm.nih.gov/19131956/ wurde bereits über 25000 mal zitiert).

NIH Tool David



Der Benutzer von DAVID sollte eine Liste an Genen hochladen, die z.B. in einem Transkriptom-Assay differentiell exprimiert waren. Man kann DAVID aber auf beliebige Gen-Listen anwenden. Die Interpretation der Ergebnisse hängt dann damit zusammen, wie diese Gene ausgewählt wurden.

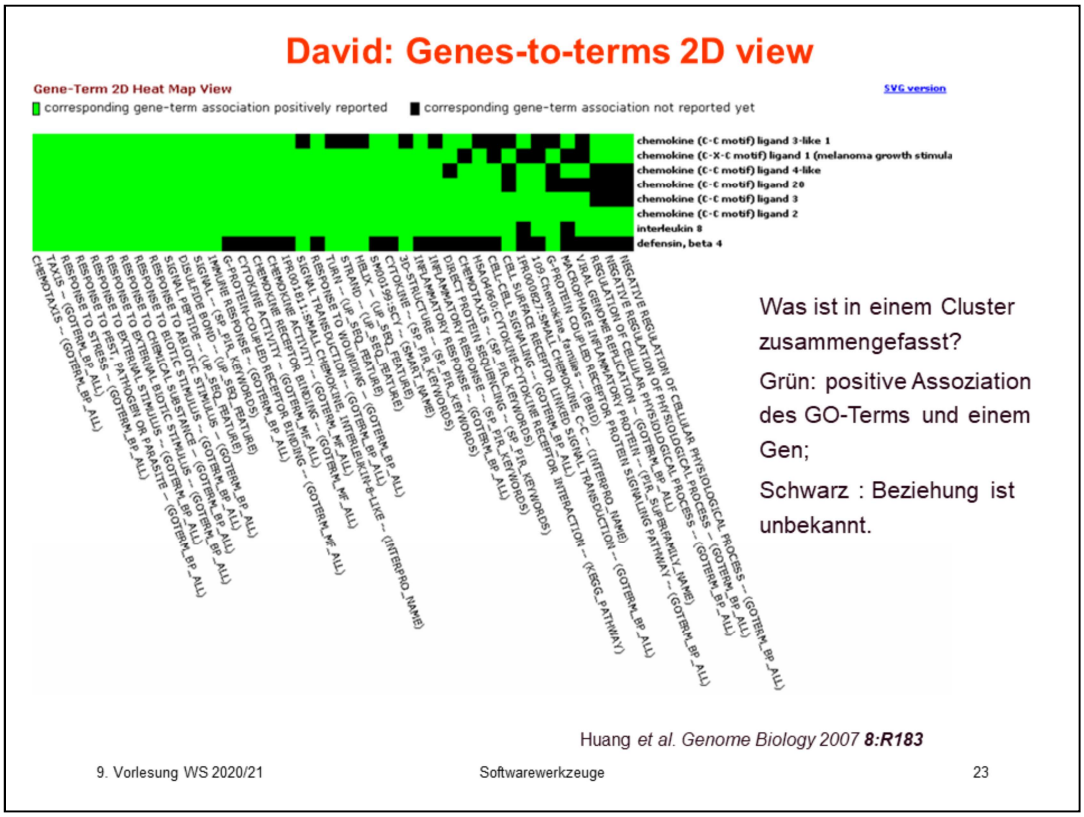
Automatische funktionelle Annotation

++ Highly Applied
+ Relevant

	Functional Annotation Chart	Functional Annotation Clustering	Functional Annotation Table	Gene Functional Classification	Gene Name Batch Viewer	Gene ID Conversion Tool	DAVID Knowledgebase	DAVID API
Initial glance of major biological functions associated with my gene list	++	++	++	+	+			
Which biological terms/functions are specifically enriched in my gene list?	++	++						
View the genes in my list on related biological pathways	++	++						
Which diseases are associated with my gene list?	++	++						
Which protein functional domains are associated with my gene list?	++	++						
Which other genes frequently interact with the genes in my list?	++	++						
How to group the highly redundant annotations into group?		++						
What are the major gene functional groups in my gene list?				++				
View related annotation and related genes on a single graphic view		++		++				
What are other functionally similar genes in genome, but not in my list?	+	+		++	++			
What are other annotations functionally similar to my interesting one?	++	++						
What are the gene names in my list?			+		++			
How to convert my gene IDs to other type of IDs?	+					++		
How to directly link to DAVID functions ?								++
How can I download DAVID data for in-house study?	+	+	+	+			++	

https://david.ncifcrf.gov/content.jsp?file=documentation.html

Dies sind typische Punkte, die Benutzer interessieren könnten.



Eine Menge von Chemokin-Liganden hat eine große Anzahl an GO-Termen gemeinsam (grün eingefärbt). Die Abwesenheit eines GO-Terms bezeichnet man am besten mit „Beziehung ist unbekannt“, da es schwierig ist, experimentell zu zeigen, dass ein Gen/Protein eine bestimmte Funktion definitiv nicht hat. Man müsste dazu ja alle möglichen Bedingungen überprüfen.

Korrektur für multiples Testen

Meist testet man sehr viele GO-Terme nacheinander.

Für jeden Test gilt im Prinzip das übliche $p < 0.05$ Kriterium.

Wenn man aber z.B. 20 Tests auf demselben Datensatz von differentiell exprimierten Genen macht, ist dann die Wahrscheinlichkeit für einen Zufallstreffer $20 \times 0.05 = 1$.

Daher muss man eine **Korrektur für multiples Testen** anwenden.

http://great.stanford.edu/

Die Korrektur für multiples Testen muss man sehr oft anwenden, nicht bloss bei der funktionellen Annotation von Genen.

Bonferroni Korrektur

Carlo Bonferroni (1892-1960) hat die „Bonferroni-Korrektur“ gar nicht erfunden. Sie benutzt aber die von ihm entwickelten Ungleichungen.

Die einfachste **Korrektur für multiples Testen** ist die **Bonferroni**-Korrektur. Dabei multipliziert man einfach den *p*-Wert mit der Anzahl an Tests. Bei 1.0 sättigt die Methode.

Die Bonferroni-Korrektur kontrolliert die **family-wise error rate**, d.h. die Wahrscheinlichkeit eine oder mehr falsche Entdeckungen zu machen.

Die Methode ist sehr konservativ, da sie alle p-Werte als unabhängig betrachtet.

Dies ist jedoch bei der Gen-Kategorie-Analyse nicht der typische Fall.

Die Bonferroni-Korrektur führt zu einer geringen statistischen Power.

Sebastian Bauer, Gene Category Analysis Methods in Molecular Biology 1446, 175-188 (2017); wikipedia.org

http://great.stanford.edu/

Die sogenannte Bonferroni-Korrektur ist eine sehr strenge Methode. Man „verliert“ oft recht viele Ergebnisse, die anschließend an die Korrektur nicht mehr signifikant sind.

Wenn man andererseits sehr kleine individuelle p-Werte erhalten hat, kann man durchaus die strenge Bonferroni-Korrektur anwenden, um den Leser zu beeindrucken, dass die Ergebnisse auch nach der Bonferroni-Korrektur immer noch sehr signifikant sind.

Benjamini Hochberg: erwartete false discovery rate

- Die Benjamini–Hochberg-Methode kontrolliert die Falscherkennungsrate (**false discovery rate** FDR). Das ist der Anteil an fälschlicherweise zurückgewiesenen Nullhypothesen zu den zurückgewiesenen Nullhypothesen insgesamt.
- Dies hat einen positiven Effekt auf die statistische Power. Allerdings hat man weniger Kontrolle über die false discoveries.
- Die FDR-Methode wird von der Amerikanischen Physiologischen Gesellschaft als Society “best practical solution to the problem of multiple comparisons” angesehen.
- Solche weniger konservativen Korrektur ergeben üblicherweise mehr signifikante Terme. Es hängt von der Situation ab, ob dies erwünscht ist.

Sebastian Bauer, Gene Category Analysis Methods in Molecular Biology 1446, 175-188 (2017)

http://great.stanford.edu/

In diesem schönen Video wird die BH-Methode motiviert:
https://www.youtube.com/watch?v=K8LQSvtjcEo

Benjamini Hochberg Korrektur: how to Kochrezept

0. Wähle eine FDR-Prozentschranke Q aus. Je nach Anwendung kann man FDR auf Werte zwischen 1% und 25% setzen.

1. Sortiere die einzelnen p-Werte in aufsteigender Reihenfolge.
2. Ordne den p-Werten jeweils einen Rang zu. Der signifikanteste p-Wert hat Rang 1, der nächste Rang 2 etc.
3. Berechne den kritischen BH-Wert für jeden einzelnen p-Wert mit der Formel *(i/m)Q*, wobei:
   - *i* = der Rang des einzelnen p-Werts,
   - *m* = Gesamtanzahl an Tests,
   - *Q* = die gewählte false discovery rate
4. Vergleiche die Original p-Werte mit den kritischen BH-Werten aus Schritt 3. Finde den größten p-Wert, der kleiner als der kritische BH-Wert ist.

https://www.statisticshowto.com/benjamini-hochberg-procedure/

Schritte 1 – 4 sind die hauptsächlichen Schritte der Benjamini Hochberg-Methode.

Schritt 0 wurde ergänzt, da die Auswahl der FDR-Schranke am Anfang geschehen sollte, nicht erst, nachdem man die Ergebnisse erhalten hat.

Benjamini Hochberg Korrektur: how to Kochrezept

Als ein Beispiel zeigt diese Liste an Daten die teilweisen Ergebnisse aus 25 Tests mit ihren jeweiligen p-Werten in Spalte 2.

In Schritt 1 wurde die Ergebnisse nach aufsteigenden p-Werten geordnet und mit Rängen versehen. Die rechte Spalte enthält den kritischen BH-Wert mit FDR=25%.

Z.Blautet dieser für „depression“ (1/25) · 0.25 = 0.01.

Der fett markierte p-Wert (für Children) ist der höchste p-Wert, der kleiner als der kritische Wert ist: .042 < .050. **Alle** Variablen oberhalb werden als significant angesehen, auch wenn ihr p-Wert nicht kleiner als der kritische Wert ist.

Variable	P Value	Rank	(I/m)Q
Depression	0.001	1	0.01
Family History	0.008	2	0.02
Obesity	0.039	3	0.03
Other health	0.041	4	0.04
Children	**0.042**	**5**	**0.05**
Divorce	0.060	6	0.06
Death of Spouse	0.074	7	0.07
Limited income	0.205	8	0.08

z.B. sind Obesity und Other Health für sich nicht significant (e.g. .039 > .03). Mit der BH-Koorektur würde man sie jedoch als significant ansehen, d.h. die Nullhypothese für diese Variablen zurückweisen.

https://www.statisticshowto.com/benjamini-hochberg-procedure/

Dies ist ein einfaches Beispiel für die Berechnung von FDR-korrigierten p-Werten. Die Prozedur ist wirklich sehr einfach. Davor braucht niemand von Ihnen Angst haben. Die zweite Spalte enthält die p-Werte, die man z.B. aus dem hypergeometrischen Test für diese Variable (oder GO-Term) erhalten hat. Mit der gewählten FDR-Schranke füllt man dann einfach die rechte Spalte mit den kritischen Werten (I/m)xQ auf. Die individuellen p-Werte gehen hier nicht ein. Wenn die individuellen p-Werte sehr klein sind, haben sie natürlich eine bessere Chance, kleiner als der kritische BH-Wert zu sein. Wenn man mehr und mehr Datenpunkte hat, fallen die aus dem statistischen Test erhaltenen p-Werte normalerweise immer kleiner aus. Andererseits wird auch der kritische Wert proportional zur Anzahl *m* der Tests immer kleiner. Dadurch bestraft man die Durchführung von vielen Tests auf demselben Datensatz.

Vergleich von GO-Termen

Die hierarchische Struktur der GO-Ontologie ermöglicht es, Proteine miteinander zu vergleichen, die mit verschiedenen GO-Termen annotiert sind.

Dies geht so lange, wie die Terme Beziehungen zueinander haben.

Nahe beieinander liegende Terme im GO-Graphen (d.h. mit wenigen dazwischen liegenden Termen) sind tendentiell **semantisch ähnlicher** zueinander als solche, die weiter voneinander entfernt sind.

Man könnte einfach die **Anzahl an Kanten** zwischen 2 Knoten als Maß für ihre Ähnlichkeit nehmen.

Dies ist jedoch problematisch, da verschiedene Regionen der GO-Ontologie unterschiedlich dicht mit Termen abgedeckt sind.

Gaudet, Škunca, Hu, Dessimoz
Primer on the Gene Ontology,
https://arxiv.org/abs/1602.01876

Wir haben bereits die Topologie der Gene Ontology vorgestellt und wie man signifikant angereicherte GO-Terme identifizieren kann. Bis jetzt haben wir uns stets mit einzelnen GO-Termen beschäftigt. Nun werden wir uns anschauen, wie man die Ähnlichkeit verschiedener GO-Terme numerisch bewerten kann.

Messe funktionelle Ähnlichkeit von GO-Termen

Die **Wahrscheinlichkeit eines Knoten** *t* drückt man so aus:

Wieviele Gene besitzen die Annotation *t* relativ zur Häufigkeit der Wurzel?

$$p_{anno}(t) = \frac{occur(t)}{occur(root)}$$

Die Wahrscheinlichkeit hat Werte zwischen 0 und 1 und nimmt zwischen den Blättern bis zur Wurzel monoton zu.

Aus der Wahrscheinlichkeit *p* berechnet man den **Informationsgehalt** jedes Knotens:

$$IC(t) = -\log p(t)$$

Je seltener ein Knoten ist, desto höher sein Informationsgehalt.

Schlicker PhD dissertation (2010)

Es gibt 2 Gruppen von Ansätzen um den Informationsgehalt (IC) von GO-Termen auszudrücken: Annotations-basiert und Topologiert-basiert.

Die hier vorgestellte Definition von p_anno gehört zu den Annotations-basierten Methoden.

Messe funktionelle Ähnlichkeit von GO-Termen

Die Menge an gemeinsamen Vorgängern (**common ancestors** (CA)) zweier Knoten t_1 und t_2 enthält alle Knoten, die auf einem Pfad von t_1 zum Wurzel-Knoten **UND** auf einem Pfad von t_2 zum Wurzelknoten liegen.

Der **most informative common ancestor** (MICA) der Terme t_1 und t_2 ist der Term mit dem höchsten Informationsgehalt in CA.

Normalerweise ist das der gemäß dem Abstand nächste gemeinsame Vorgänger.

https://repositorio.ul.pt/bitstream/10451/14140/1/07-6.pdf

Eine Art um die semantische Ähnlichkeit zwischen zwei (oder mehr) GO-Termen zuzuordnen ist es, ihre gemeinsamen Vorläufer-Knoten zu berücksichtigen. Intuitiv wäre der „nächste“ gemeinsame Vorläufer vermutlich am aussagekräftigsten. Aufgrund der DAG-Struktur der Gene Ontology kann es jedoch mehrere gemeinsame „nächste“ Nachbarn geben, die entweder auf derselben Ebene der GO-Hierarchie liegen oder dieselbe Pfadlänge zu den beiden GO-Termen aufweisen. Stattdessen wählt man meist unter allen gemeinsamen Vorläufer-Knoten denjenigen mit höchstem Informationsgehalt (IC). Dieser wird most informative common ancestor genannt.

Messe funktionelle Ähnlichkeit von GO-Termen

Aus dem Abstand zum **most informative common ancestor** (MICA) kann man die funktionelle Ähnlichkeit der GO Terme t_1 und t_2 definieren:

$$sim_{Rel}(t_1,t_2) = \frac{2 \cdot IC(MICA)}{IC(t_1)+IC(t_2)}$$

MICAs, die nahe bei t_1 und t_2 liegen, erhalten daher eine höhere Bewertung als diejenigen Knoten, die weiter entfernt (d.h. weiter oben) im GO-Graphen liegen.

Man normiert den IC des MICA-Knoten mit der Summe der ICs der beiden betrachteten GO-Terme.

Da man oben das Attribut eines Knotens hat und unten die Summe über die Attribute von 2 Knoten, wird der Zähler mit 2 multipliziert.

Im Extremfall, wenn IC(MICA) = IC (t1) = IC (t2), kann die Ähnlichkeit maximal gleich 1 sein.

Messe funktionelle Ähnlichkeit von GO-Termen

Zwei Gene oder zwei Mengen an Genen *A* und *B* haben jedoch meist jeweils mehr als eine GO-Annotation. Betrachte daher die Ähnlichkeit aller Terme *i* und *j*:

$$s_{ij} = sim(GO_i^A, GO_j^B), \forall i \in 1,...,N, \forall j \in 1,...,M.$$

und wähle daraus in den Reihen und Spalten jeweils die Maxima

$$rowScore(A,B) = \frac{1}{N}\sum_{i=1}^{N} \max_{1 \le j \le M} s_{ij}, \qquad GOscore_{avg}^{BMA}(A,B) = \frac{1}{2} \cdot (rowScore(A,B) + columnScore(A,B))$$

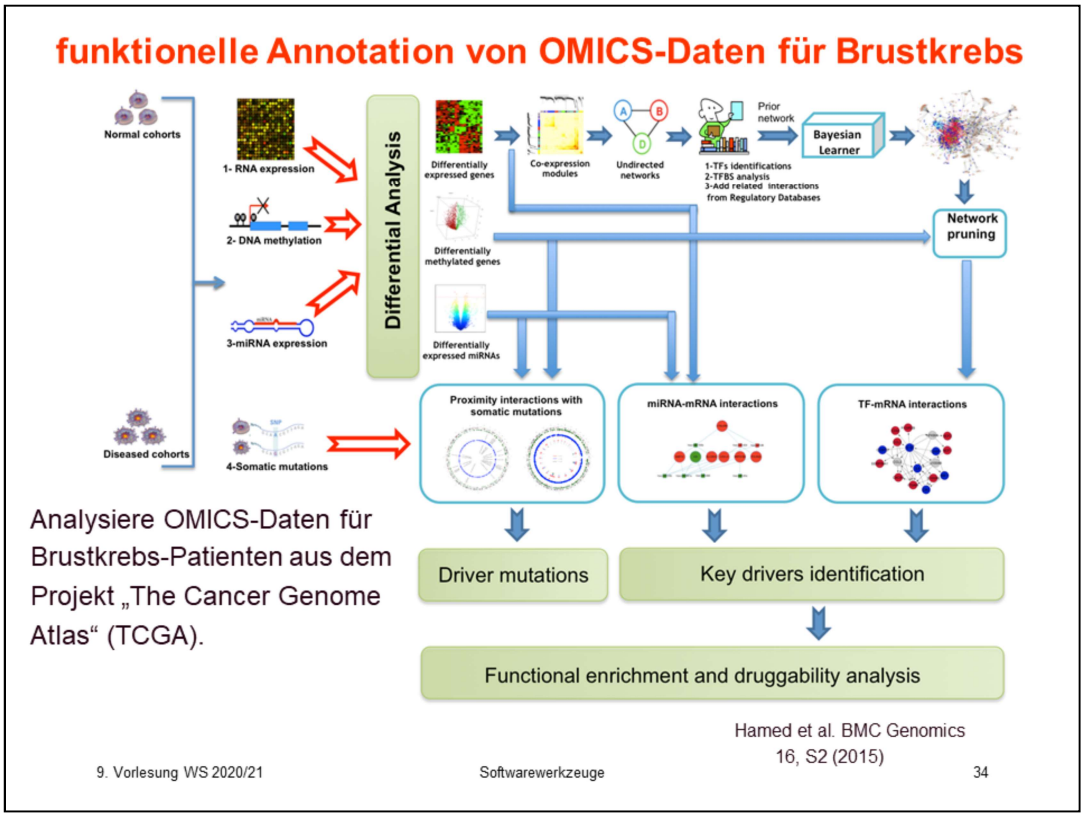
$$columnScore(A,B) = \frac{1}{M}\sum_{j=1}^{M} \max_{1 \le i \le N} s_{ij}. \qquad GOscore_{max}^{BMA}(A,B) = max(rowScore(A,B), columnScore(A,B))$$

Aus den Scores für den BP-Baum und den MF-Baum wird der *funsim*-Score berechnet.

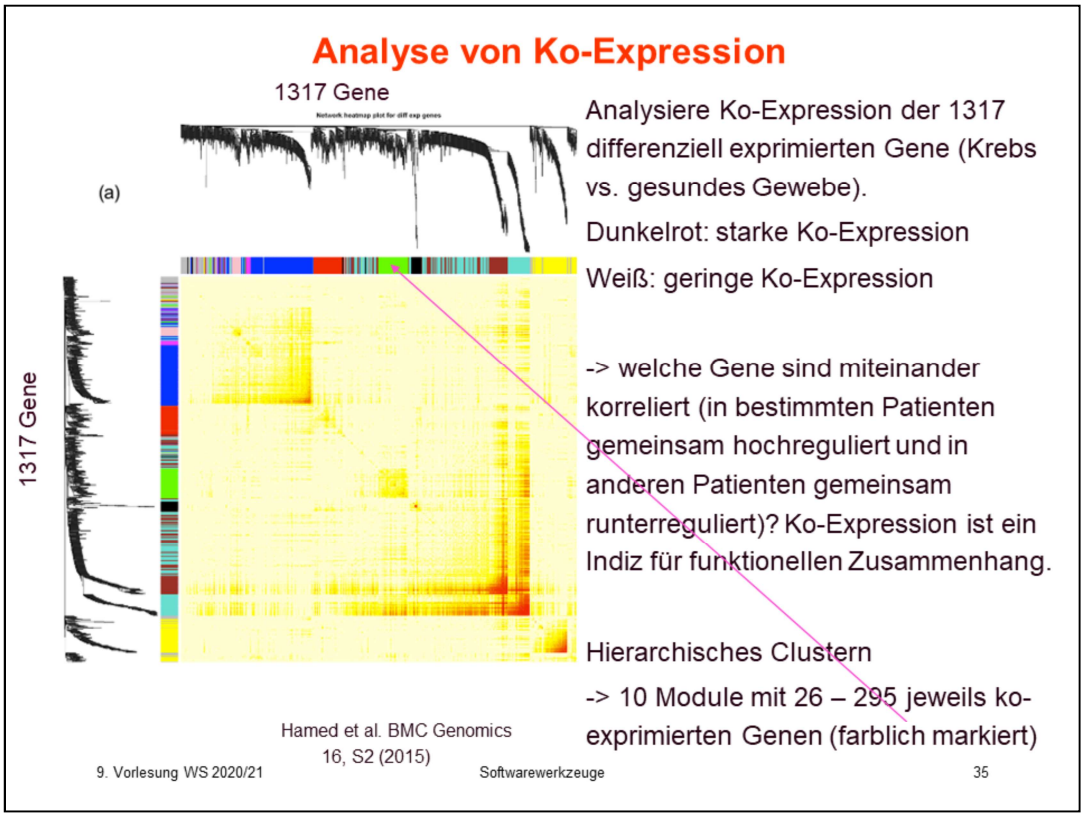
$$funsim(A,B) = \frac{1}{2} \cdot \left[\left(\frac{BPscore}{\max(BPscore)} \right)^2 + \left(\frac{MFscore}{\max(MFscore)} \right)^2 \right]$$

Schlicker PhD dissertation (2010)

Wir können dieses Prinzip fortsetzen. Dadurch gelangen wir von der semantischen Ähnlichkeit zwischen 2 GO-Termen zu der semantischen/funktionellen Ähnlichkeiten zwischen 2 Genen. Meist hat ein Gene Annotationen mit mehreren GO-Termen. Damit ist nicht gemeint 1 BP-Annotation, 1 MF-Annotation und 1 CC-Annotation, sondern z.B. 5 BP-Annotationen. Wenn wir nun die funktionellen Ähnlichkeiten zwischen 2 Genen bewerten möchten, können wir dazu betrachten, wie hoch die Ähnlichkeiten zwischen all ihren GO-Termen sind. Wir wissen jedoch nicht, ob der erste GO-Term von Gen1 am besten zu dem ersten, zweiten, … Term von Gen2 passt. Deshalb betrachten wir einfach alle Kombinationsmöglichkeiten und ordnen die GO-Terme von Gen1 auf der x-Achse einer Matrix an und die GO-Terme von Gen2 auf der y-Achse. Dann füllen wir die Einträge der Matrix mit den paarweisen Ähnlichkeiten zwischen ihren GO-Termen (siehe vorige Folie).



Dies ist ein Beispiel aus unserer Forschung zu Brustkrebs (siehe https://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-16-S5-S2). Der Doktorand Mohamed Hamed analysierte TCGA-Daten für Transkriptom, Methylom, miRNA-Expression und somatische Mutationen. Wir konzentrieren uns auf das, was mit den mRNA-Expressionsdaten gemacht wurde.



Zunächst wurde die 1317 differentiell exprimierten Gene in Tumor vs. gesundem Gewebe bestimmt. Für alle Paare von Genen wurde der Grad an Ko-Expression bestimmt. Damit wurden die Gene dann in 10 ko-exprimierte, disjunkte Module geclustert.

Gibt es angereicherte Genfunktionen in diesen Modulen?

	Module	Gene count	Top GO category	Top KEGG categories	Key driver count	Key drivers
TF- mRNA interactions	black	41	Regulation of transcription	Pathways in cancer, Renal cell carcinoma	5	SORBS3, ZNF43, ZNF681, RBMX, POU2F1
	blue	247	Nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	Cell cycle, Prostate cancer, Melanoma	9	AR, BRCA1, ESR1, JUN, MYB, RPN1, E2F1, E2F2, PPARD
	brown	195	Anatomical structure morphogenesis	Leukocyte transendothelial migration	5	TMOD3, CREB1, POU5F1, SP3, TERT
	green	110	Cellular macromolecule metabolic process	Endometrial cancer, Insulin signaling pathway	15	B4GALT7, OS9, CDC34, MAN2C1, MYO1C, SH3GLB2, INPP5E, PLXNB1, USF2, PPP1R12C, CDK9, DAP, E4F1, E2F4, USF1
	grey	148	Anatomical structure development	Sulfur metabolism	18	AHCTF1, NQO2, FGFR2, CCDC130, ABCG4, BIRC6, CA6, SP4, RNF2, SPRR1B, C16orf65, DNAJC5G, SNCAIP, GRIK5, SLC6A4, SMAD1, DAD1, POU4F2
	magenta	26	Regulation of metabolic process	p53 signaling pathway, Alzheimer's disease	3	ATF6, NGEF, POGK
	pink	30	Transcription initiation from RNA polymerase II promoter	Basal transcription factors	4	CCDC92, TMEM70, RNF139, E2F5
	red	93	Regulation of cellular process	Endometrial cancer, Neurotrophin signaling pathway	14	ATP1B1, STAT3, ABCB8, MYC, TGFB1, SP1, TP53, PCGF1, SUMF2, GTF3A, IPO13, GMPPA, HTR6, TGIF1
	turquoise	295	Regulation of cellular metabolic process	p53 signaling pathway, Pancreatic cancer, Apoptosis	2	UBL5, RNF111
	yellow	132	Immune system process	Chemokine signaling pathway, Natural killer cell mediated cytotoxicity	19	APOC1, CD2, CD79B, LRRC28, DAPK1, FAM124B, EML2, LAP3, TSPAN2, FCRL3, ELMO1, SLC7A7, RASSF5, SLC31A2, TRAF3IP3, GALNT12, ITGA4, SPI1, TFAP2A
	Total	1317				

Am stärksten angereicherter GO term (kleinster p-value) unter den Genen dieses Moduls

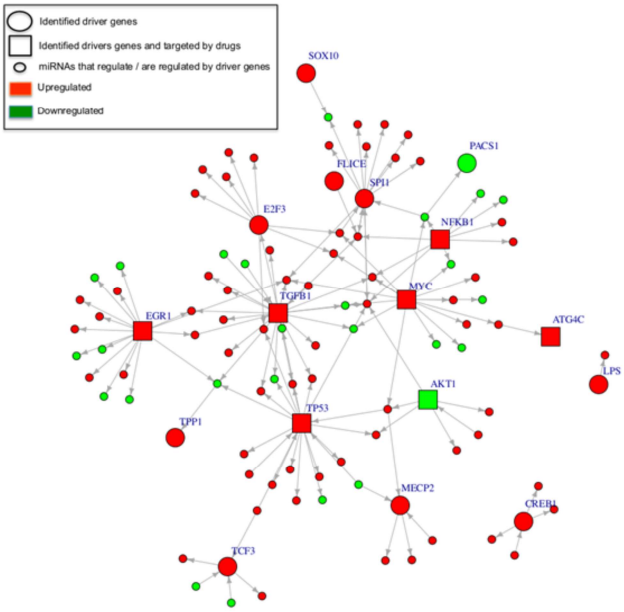
Wieviele Gene sind key drivers?

-> Module hängen mit Prozessen zusammen, die bereits mit Brustkrebs in Verbindung gebracht werden (endometrial cancer, p53, Prostatakrebs ...)

Hamed et al. BMC Genomics 16, S2 (2015)

In der zweiten Spalte sind die 10 Module aufgelistet, in der dritten Spalte die Anzahl an Genen in jedem Modul. In der vierten und fünften Spalte stehen die signifikantesten angereicherten GO-Terme und KEGG-Pfade. Durch eine topologische Analyse kann man Transkriptionsfaktoren („key drivers“) bestimmen, die möglichst viele Gene in einem Modul regulieren.

Ergänze regulatorische Information -> key regulators



Differenziell exprimierte Gene aller Module

-> extrahiere regulatorische Interaktionen (TF -> Gen) aus öffentlichen Datenbanken wie JASPAR, TRED, MSigDB

Identifiziere 85 **key regulators:** 17 Transkriptionsfaktoren und 68 miRNAs, die zusammen alle differentiell exprimierten Gene regulieren.

31% der key regulators kodieren für Proteine, die Targets für bekannte Krebs-Medikamente sind!

Hamed et al. BMC Genomics 16, S2 (2015)

Der Graph zeigt die regulatorischen Verknüpfungen der 17 key Transkriptionsfaktoren. Man sieht (a), dass die key TFs ein eng miteinander verknüpftes Netzwerk aufspannen. (b) Ausserdem sind ein knappes Drittel der key TFs bereits Zielmoleküle für bekannte Krebs-Medikamente. Diese Prozedur identifiziert also erfolgreich viele Gene, die man bereits als wichtige Faktoren für Brustkrebs identifiziert hatte. Oft sind diese key TFs bei Brustkrebs mutiert.

GO ist unvollständig

Die Gen-Ontologie repräsentiert eine Auswahl des aktuell verfügbaren **Wissens**. Daher ist sie sehr **dynamisch**.

Die Ontologie wird ständig verbessert um die Biologie aller Organismen möglichst genau darzustellen.

Sobald neue Entdeckungen gemacht werden, werden diese in GO aufgenommen.

Allerdings stellt die Geschwindigkeit der aktuellen Forschung das GO-Konsortium vor die hohe Herausforderung, damit Schritt zu halten.

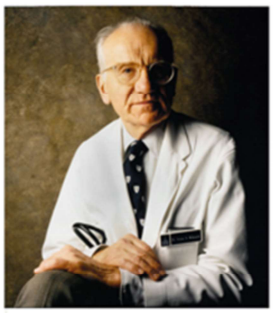
Auf jeden Fall ist die Information in GO notwendigerweise **unvollständig.**

Daher bedeutet fehlende Evidenz über (eine bestimmte) Funktion NICHT, dass diese Funktion nicht vorliegt.

Gaudet, Dessimoz,
Gene Ontology: Pitfalls, Biases, Remedies
https://arxiv.org/abs/1602.01876

Wenn man funktionelle Annotationen mit der Gene Ontology vornimmt, muss man sich immer bewusst sein, dass die Gene Ontology notgedrungen unvollständig ist. Sie bildet aus Kapazitätsgründen nur einen Teil des bekannten Wissens ab, außerdem ist eben noch nicht alles bekannt.

OMIM-Datenbank

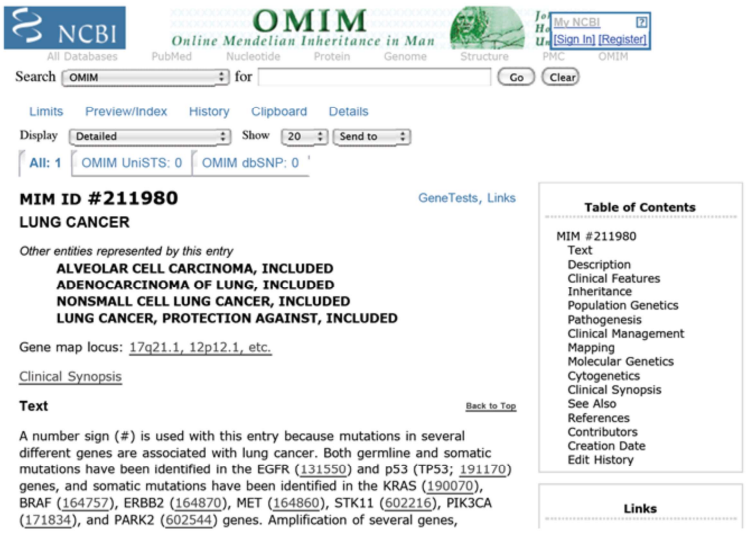


Victor McKusick (1921-2008),
Johns Hopkins Universität,
- begründete das Gebiet
Medical genetics
- gründete die Datenbank
Mendelian Inheritance in Man

9. Vorlesung WS 2020/21

OMIM ®, Online Mendelian Inheritance in Man ®.

OMIM is a comprehensive, authoritative, and timely compendium of human genes and genetic phenotypes.



Die OMIM-Datenbank (https://omim.org) ist ein Quasi-Standard für genetisch verursachte Krankheiten.

BIOINFORMATICS

Vol. 26 ECCB 2010, pages i561–i567
doi:10.1093/bioinformatics/btq384

Improving disease gene prioritization using the semantic similarity of Gene Ontology terms

Andreas Schlicker†, Thomas Lengauer and Mario Albrecht*

Max Planck Institute for Informatics, Department of Computational Biology and Applied Algorithmics, Campus E1.4, 66123 Saarbrücken, Germany

ONIM-Datenbank & UniProt Datenbank: GO-Annotationen für bekannte Krankheits-gene.

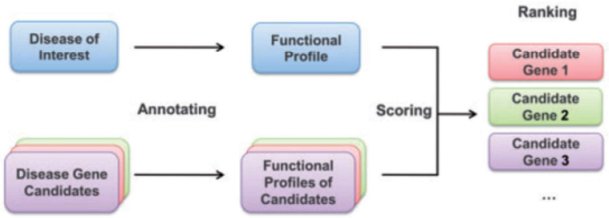


Fig. 1. Flow chart of the MedSim approach. First, the functional profiles of the disease of interest and the disease gene candidates are created using one of the annotation strategies. Afterwards, the functional profile of the disease is scored against each functional profile of a candidate, and the candidates are ranked according to this functional similarity score.

Schlicker et al. Bioinformatics 26, i561 (2010)

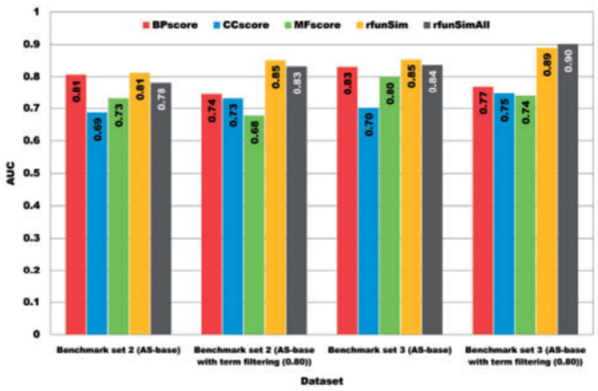
In der oberen Reihe werden einer Reihe von Krankheiten die damit verknüpften Gene zugeordnet und deren GO-Terme gesammelt.

In der unteren untersucht man ein Gen, das mit einer oder mehreren unbekannten Krankheiten in Verbindung stehen könnte.

Man sammelt die GO-Terme dieses Gens und vergleicht diese Liste mit den Listen der oberen Krankheiten. So kann man eine Priorisierung vornehmen, mit welchen Krankheiten dieses Gen in Verbindung stehen könnte.

Die Methode liefert recht genaue Vorhersagen, mit welchen Krankheiten Gene in Verbindung stehen könnten.

Die Sensitivität, d.h. die Anzahl der korrekten Vorhersagen relativ zur Anzahl aller Vorhersagen, beträgt 73%.



Schlicker et al. Bioinformatics 26, i561 (2010)

Hier wird anhand bekannter Gen-Krankheiten-Zuordnungen evaluiert, wie genau diese Methode funktioniert. Die BP-Terme der Gene Ontology (rot) sind aussagekräftiger als die MF-Terme (grün) und CC-Terme (blau). Am besten funktioniert eine Kombination aus den 3 Termen (rfunSim).

Zusammenfassung

- funktionelle Annotation für differentiell exprimierte Gene ist Standardverfahren

- man verwendet entweder Gene Ontology-Terme oder KEGG-Pfade

- signifikante An-/Abreicherung wird mit hypergeometrischem Test analysiert

- semantische Ähnlichkeit von GO-Termen erlaubt es, die funktionelle Ähnlichkeiten von Genen numerisch zu bewerten

Kein Kommentar.