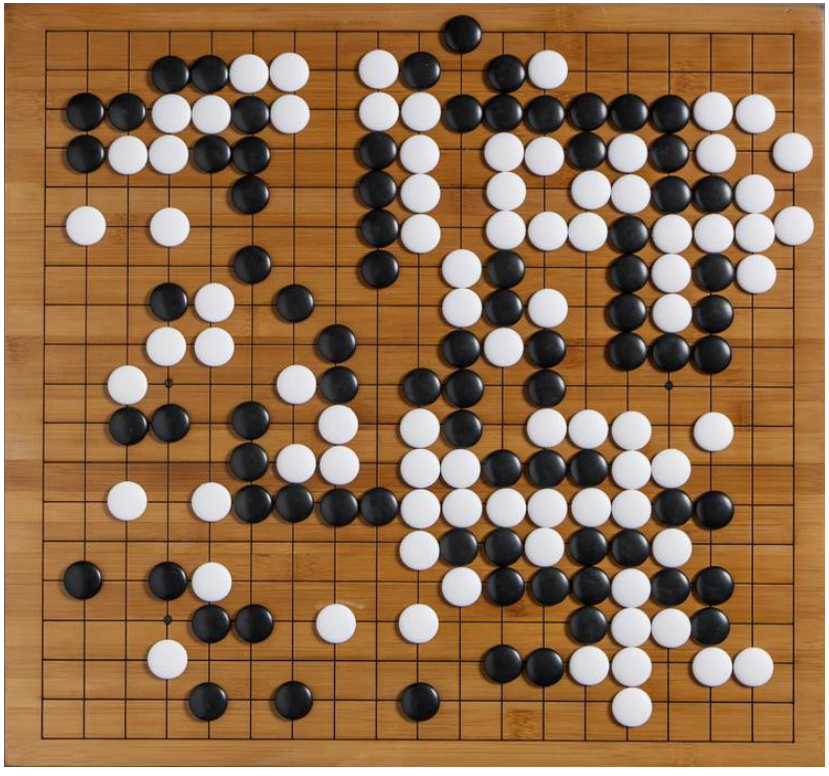
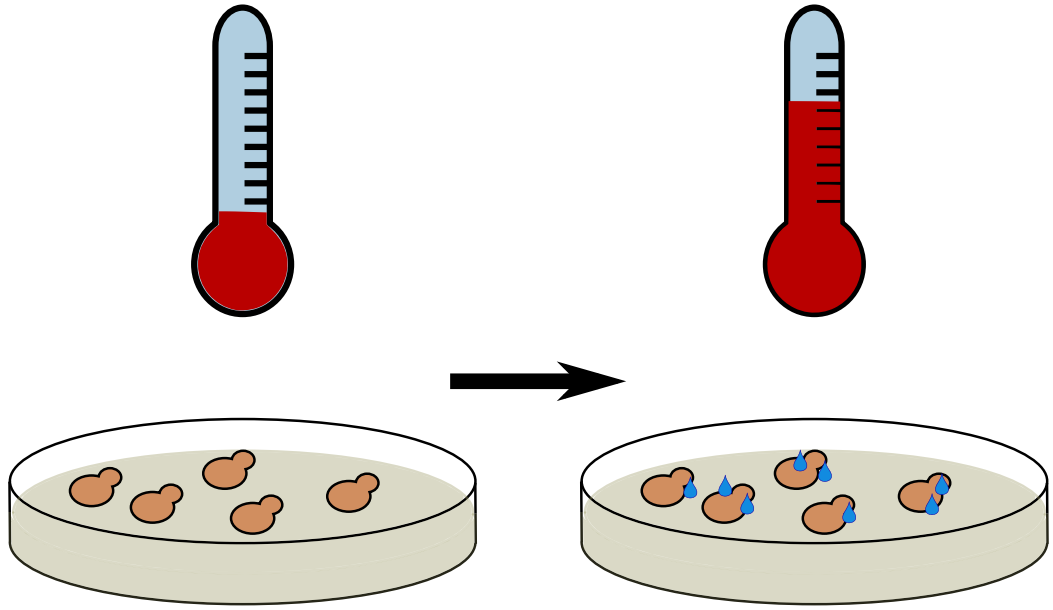


"Gene Ontology"

Wie passt sich Hefe an Hitzestress an?



Mit RNA-seq können wir herausfinden, welche Gene bei hohen Temperaturen stärker exprimiert werden. Eine Liste mit diesen Genen finden Sie auf der Wiki: `signifikanteGene.txt`. Dort gibt es auch eine Liste mit allen gemessenen Genen von Bäckerhefe: `alleGene.txt`

Wie können wir herausfinden was für Funktionen sich diese Gene teilen?

Funktionelle Annotationen für Gene

Zentrale Fragen für die Analyse von Gen-Funktionen:

- Welche Funktionen haben meine Gene?
- Was für Funktionen gibt es überhaupt und was für eine Beziehung haben sie zueinander?

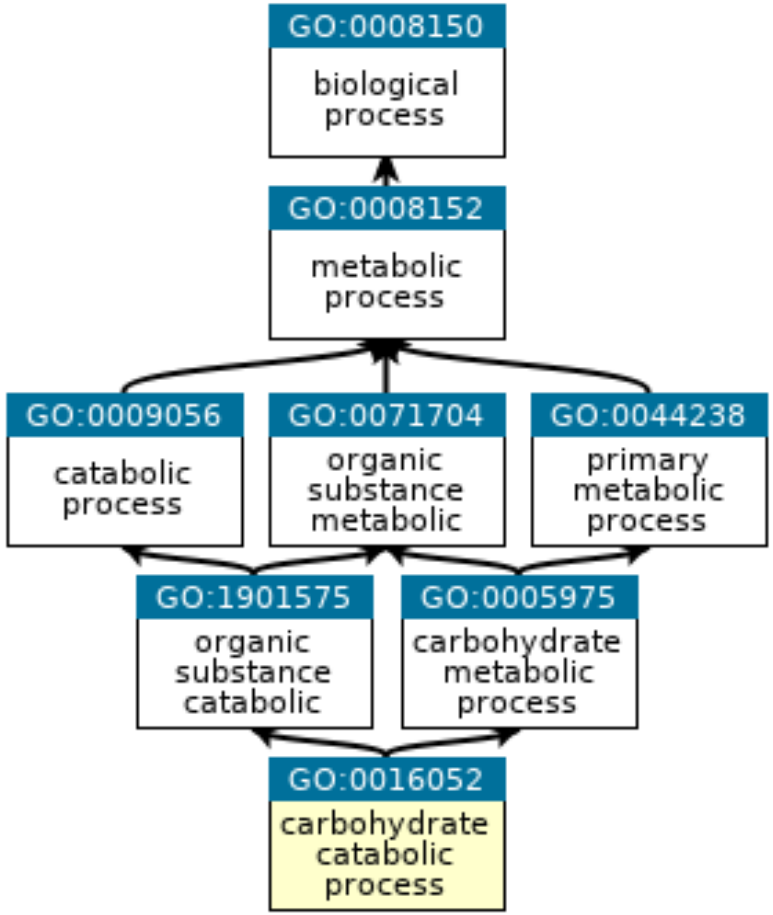
Funktionen von Genen werden mit GO-Termen annotiert. Als Beispiel schauen wir uns einmal die Annotationen des Genes `YJL052W` an. Die Annotationen können wir auf `http://amigo.geneontology.org` sehen.

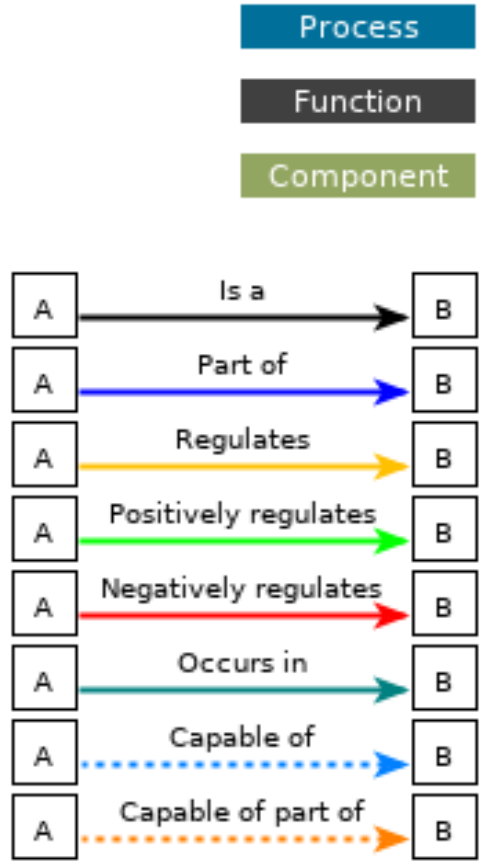
GO-Annotationen sind in drei Bereiche oder Ontologien geteilt

- **BP (Biological Process)**: In was für einem Stoffwechselweg spielt das Gen eine Rolle? Was für eine Funktion übernimmt es in der Zelle?
- **MF (Molecular Function)**: Wie interagiert das Protein mit anderen Molekülen? An was für Reaktionen ist es beteiligt? Hat es eine strukturelle Funktion?
- **CC (Cellular Component)**: Wo findet man das Protein in der Zelle? In was für Proteinkomplexen ist es enthalten?

Diese Ontologien sind **DAGs**

- **Directed**: Die Kanten im Netzwerk sind gerichtet.
- **Acyclic**: Es dürfen keine Kreise im Netzwerk entstehen.
- **Graph**: Es handelt sich um einen Graphen, die Knoten sind durch Kanten verbunden.

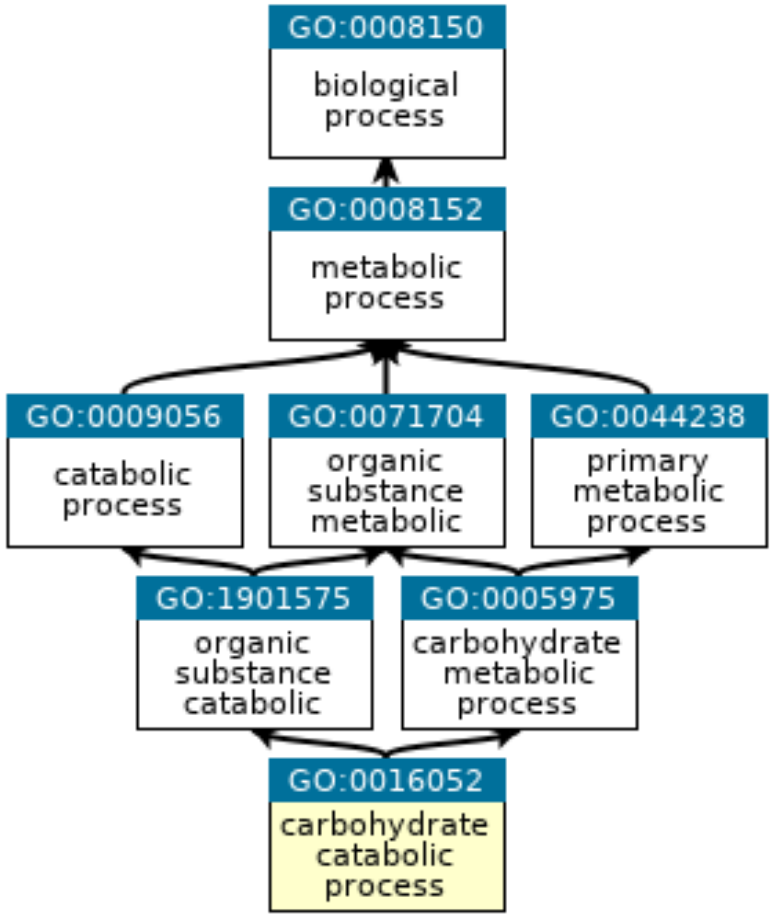




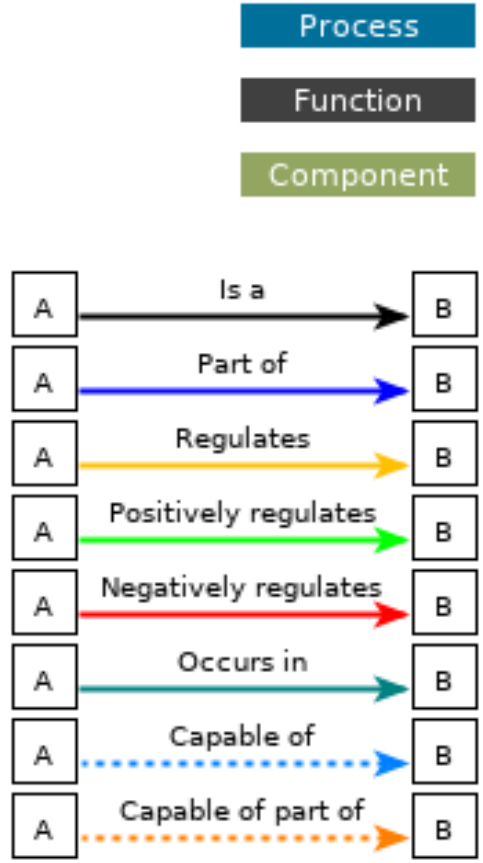
Diese Ontologien sind **DAGs**

Aber Ontologien sind keine Bäume, denn

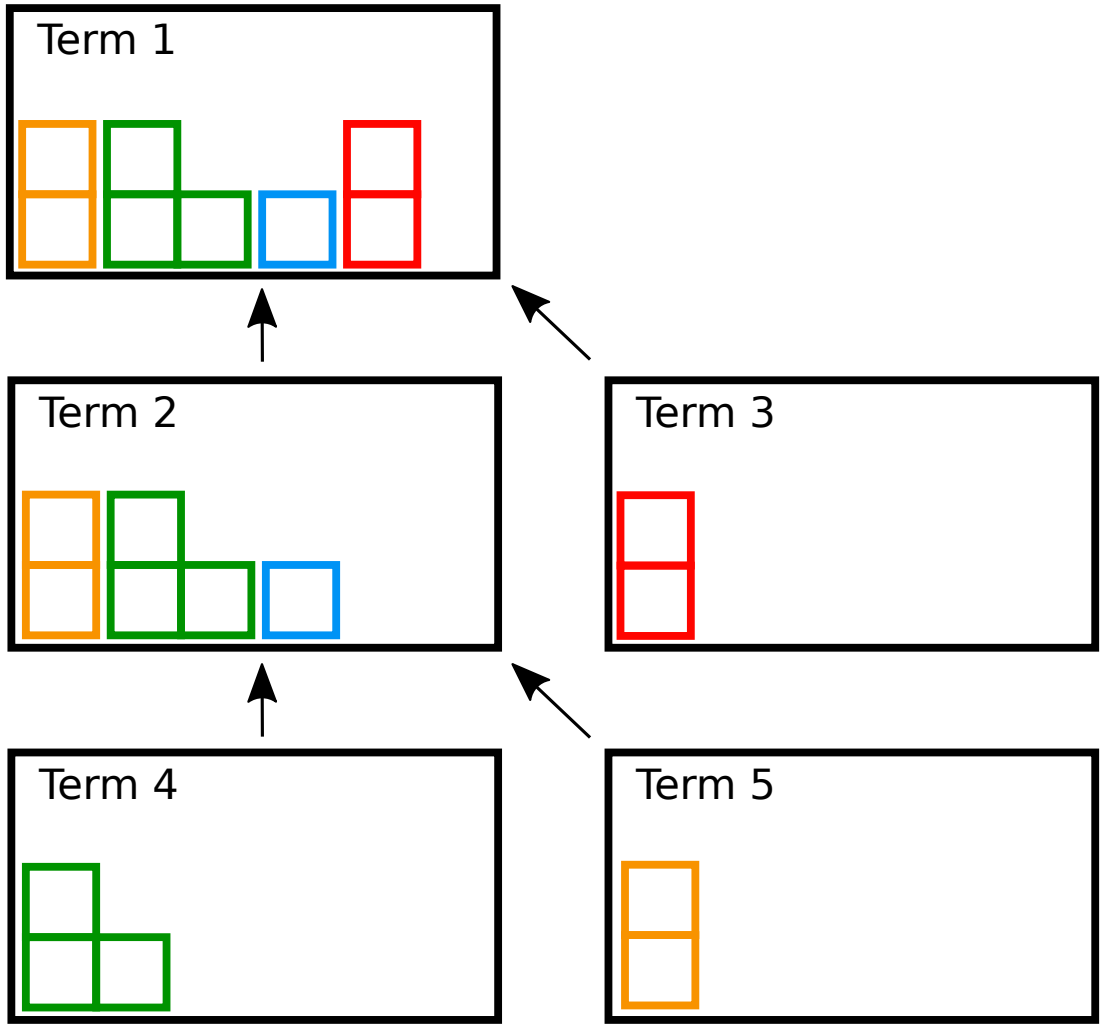
- sie können mehr als zwei Kinder und
- mehr als einen Elternknoten haben.



QuickGO - https://www.ebi.ac.uk/QuickGO



Die Annotationen werden für Eltern-Terme übernommen.



Wann ist ein Term signifikant angereichert?

Für den Term `GO:0016052` ("carbohydrate catabolic process") sind 10 von den signifikanten Genen annotiert. Ist das viel? Ist das wenig?

Die Verteilung der Gene die für diesen Term annotiert sind sehen Sie hier:

	signifikant	nicht signifikant
im Term	10	76
nur in anderen Termen	99	5489

Wann ist ein Term signifikant angereichert?

	signifikant	nicht signifikant
im Term	10	76
nur in anderen Termen	99	5489

Hierbei handelt es sich um eine 2x2 Kontingenztabelle. Wir wollen wissen, ob die Werte auf der Diagonalen größer sind, als wir es durch Zufall erwarten würden.
In anderen Worten: Überlappen die signifikanten Gene und die Gene, die in einem bestimmten Term sind, mehr als man es durch Zufall erwarten würde?

Wann ist ein Term signifikant angereichert?

	signifikant	nicht signifikant
im Term	10	76
nur in anderen Termen	99	5489

Mit dem exakten Fisher-Test können wir genau das testen. In R sieht das so aus:

```
fisher.test(rbind(c(10,76),c(99,5489)),alternative="greater")
```

GO:0016052 ("carbohydrate catabolic process") hätte demnach einen p-Wert von ungefähr $4 * 10^{-6}$.

DAVID: The Database for Annotation, Visualization and Integrated Discovery

Als erstes Tool schauen wir uns den Webserver DAVID an: `https://david.ncifcrf.gov/` (oder der dritte Eintrag bei Google)

Auf der Seite gehen wir oben auf `Start Analysis`.

Unter dem Reiter `Upload` können wir die `signifikanteGene.txt` hochladen. Hier wählen wir die Optionen `ENSEMBL_TRANSCRIPT_ID` und `Gene List`. Wir laden die Liste mit `Submit List` hoch.

DAVID: The Database for Annotation, Visualization and Integrated Discovery

Wir laden auch eine weitere Genliste als "Hintergrund" hoch. Damit werden dann unsere signifikanten Gene verglichen. Hier laden wir die Datei alleGene.txt hoch. Bei dieser Liste wählen wir die Option "Background" und laden auch diese hoch.

Wir wählen die signifikanten Gene im `List manager` und alle Gene bei `Background`.

Wir entfernen rechts den Haken bei `Check Defaults`. Stattdessen wählen wir erstmal nur `GOTERM_BP_ALL`, `GOTERM_MF_ALL` und `GOTERM_CC_ALL` unter `Gene Ontology` aus.

DAVID: The Database for Annotation, Visualization and Integrated Discovery

Zuerst schauen wir uns die Ergebnisse im `Functional Annotation Chart` an. Hier sind alle Ergebnisse, nach p-Wert sortiert, aufgelistet. Diese p-Werte wurden mit einem modifizierten Fisher-Test (EASE Score) errechnet.

Hier können wir schon interessante Terme wie "cellular response to heat" und "protein refolding" erkennen.

DAVID: The Database for Annotation, Visualization and Integrated Discovery

Als nächstes öffnen wir das `Functional Annotation Clustering`. Hier sind die gleichen Terme anhand der Gene, die in ihnen sind, geclustert. Für jeden Cluster können wir uns die Gene mit G anzeigen lassen.

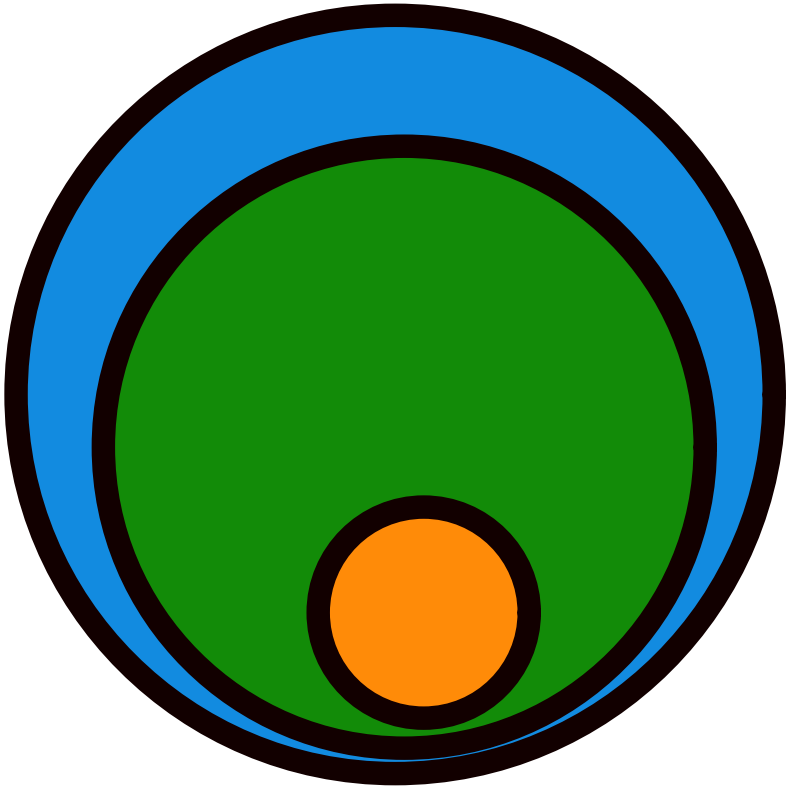
Hier fällt auf, dass sehr viele ähnliche Terme signifikant sind. DAVID bietet auch die Möglichkeit nur die spezifischste Annotation pro Gen zu testen. Hierfür muss man dann z.B. `GO_TERM_BP_DIRECT` auswählen. So werden die Cluster kleiner, aber es können auch interessante Treffer verloren gehen.

DAVID: The Database for Annotation, Visualization and Integrated Discovery

Mit DAVID kann man auch andere interessante Kategorien testen:

- Proteindomänen
- Stoffwechselwege
- Publikationen

Warum ist der Hintergrund wichtig?

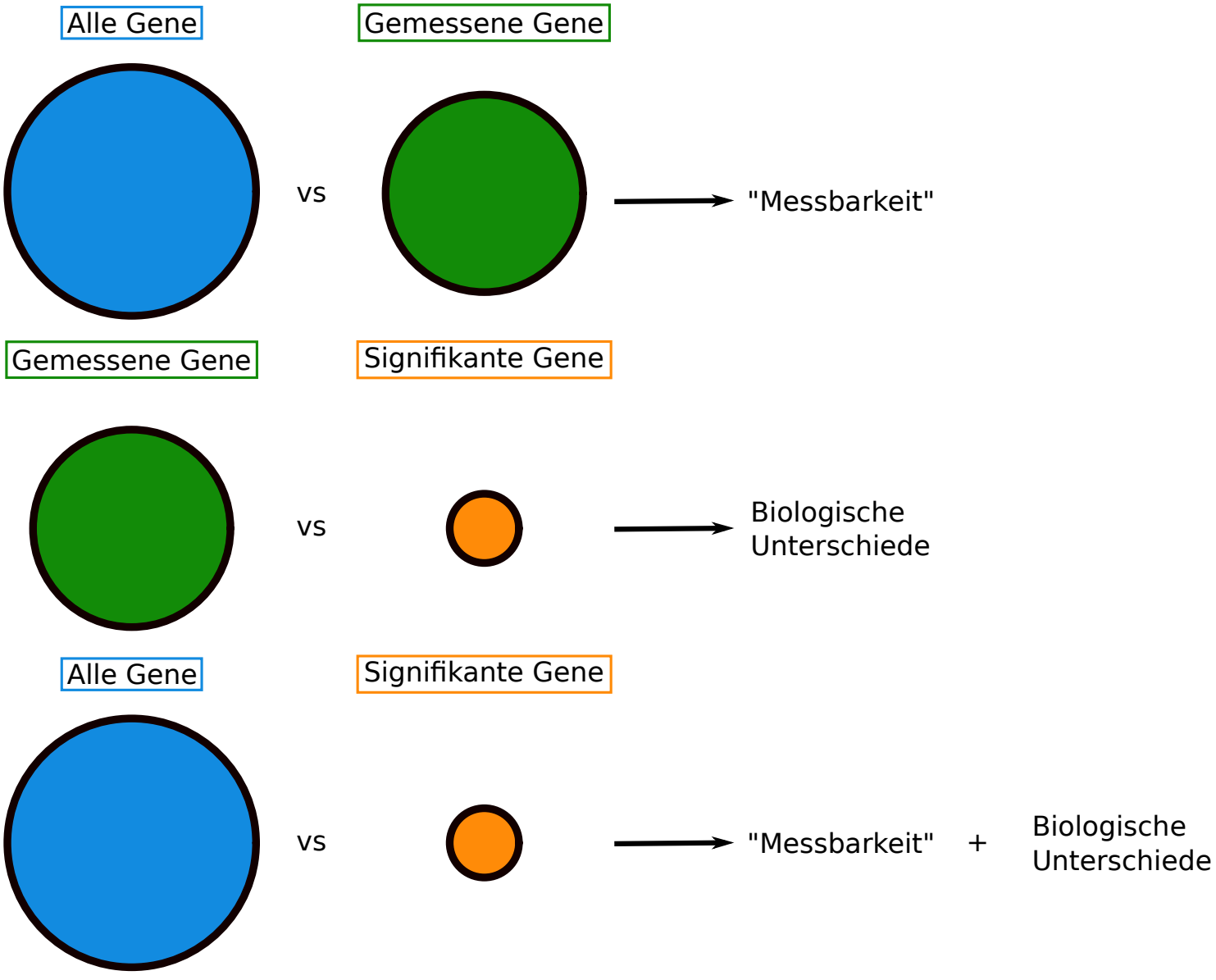


Alle Gene

Gemessene Gene

Signifikante Gene

Warum ist der Hintergrund wichtig?



GO-Enrichment Analysen mit topGO in R

Tests für funktionelle Anreicherungen in R zu machen, hat einige Vorteile:

- Man kann fast alle Parameter anpassen.
- Man kann alle Zwischenschritte analysieren und auf alle relevanten Daten zugreifen.
- Die Analysen können leichter automatisiert werden.
- Man kann leichter mit den Ergebnissen weiterarbeiten.
- Ohne großen Aufwand können hunderte von Genlisten getestet werden.

Zur Vorführung von topGO laden Sie sich bitte die Dateien topGOVorfuehrung.R und topGODaten.RData von der Wiki herunter. Die Vorführung ist **nicht** klausurrelevant! Die Folien sollten Sie aber kennen!

Große Terme sind nicht immer interessant!

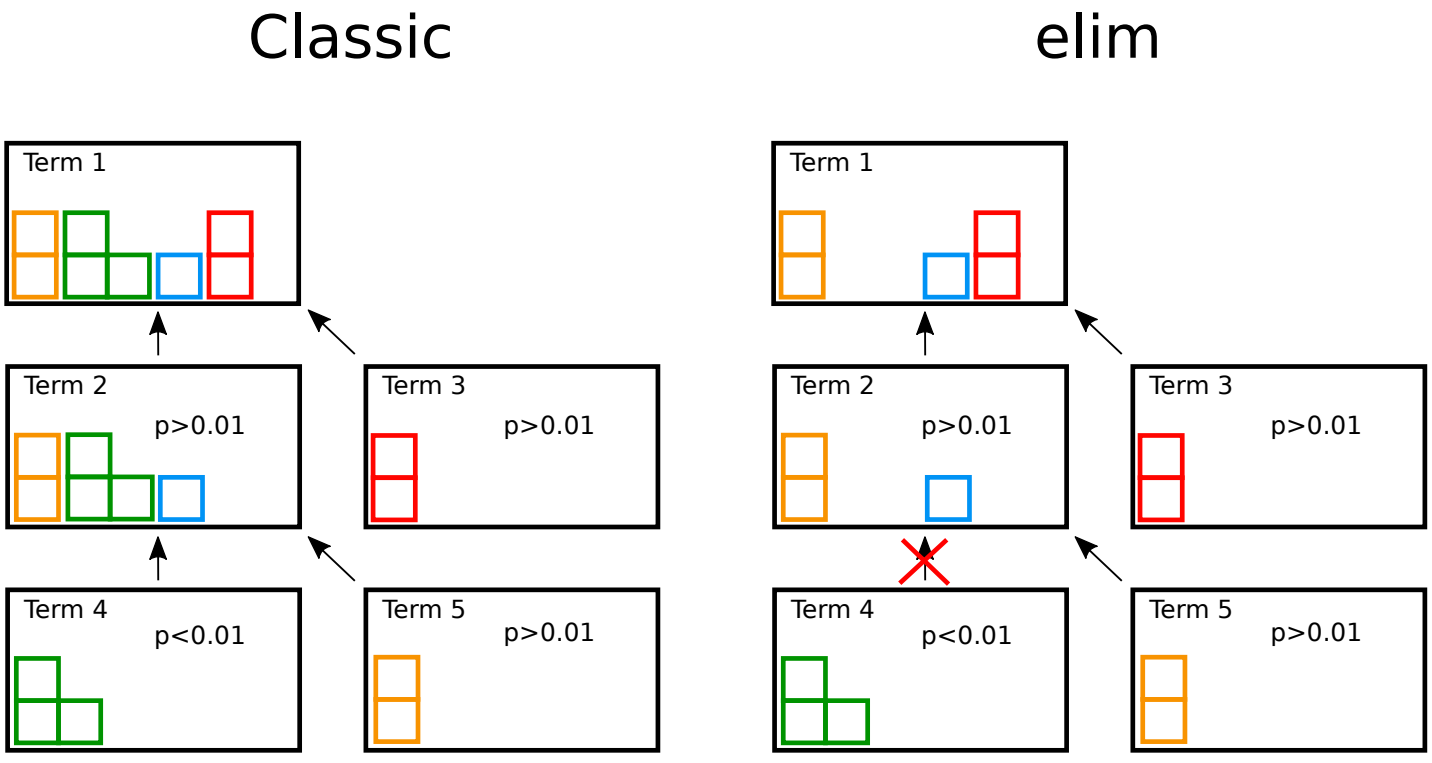
GO-Terme mit vielen annotierten Genen können mit relativ schwachen Anreicherungen sehr kleine p-Werte erreichen. Biologisch sind sie aber oft nicht interessant, da sie schwer interpretierbar sind.

In David wurde dieses Problem mit den "direkten" Annotationen (`GO_TERM_BP_DIRECT`, `GO_TERM_MF_DIRECT`, `GO_TERM_CC_DIRECT`) gelöst.

topGO hat eine noch bessere Lösung: den `elim`-Algorithmus.

Der elim-Algorithmus

Der elim-Algorithmus gibt Annotationen nur an Elternterme weiter wenn die kleineren Terme nicht signifikant waren:



Dadurch haben größere Terme oft größere p-Werte als mit dem klassischen Algorithmus, aber man verliert auch keine biologisch interessanten Ergebnisse.

Zusammenfassung

- GO-Terme werden genutzt um die Funktion von Genen systematisch zu beschreiben.
- Diese Terme sind in drei Ontologien geteilt. Eine Ontologie ist kein Baum, sonder ein DAG.
- Mit dem exakten Test nach Fisher kann man testen ob eine Gruppe von Genen sich häufiger eine bestimmte Funktion teilt, als man durch Zufall erwarten würde.
- DAVID und topGO sind Tools um die Teste durchzuführen.
- Mit dem elim-Algorithmus kann man kleinere stärker angereicherte Terme bevorzugen.
- Wenn man gegen den falschen Hintergrund testet, kann das zu Fehlern erster und zweiter Art führen.