Die funktionserhaltende, integrative Genselektion: Eine Methode zur Reduktion von krankheitsbezogenen Gensätzen auf ihre Schlüsselkomponenten

Dissertation

zur Erlangung des Doktorgrades der Naturwissenschaften

(Dr. rer. nat.)

dem Fachbereich Mathematik und Informatik der

Philipps-Universität Marburg vorgelegt

von

Catharina Lippmann

geboren in Marburg.

Marburg, Mai 2020

Erstgutachter: Prof. Dr. Alfred Ultsch, Philipps-Universität Marburg
Zweitgutachter: Prof. Dr. Dominik Heider, Philipps-Universität Marburg

Eingereicht am 05.05.2020
Tag der mündlichen Prüfung am 07.10.2020
Vom Fachbereich Mathematik und Informatik der Philipps-Universität Marburg, Hochschulkennziffer (1180), als Dissertation angenommen am 25.09.2020.

„We are drowning in information but starved for knowledge."
John Naisbitt [Naisbitt, 1982, p. 24]

Eigenständigkeitserklärung

Hiermit versichere ich, Catharina Lippmann, die vorliegende Dissertation mit dem Titel „Die funktionserhaltende, integrative Genselektion: Eine Methode zur Reduktion von krankheitsbezogenen Gensätzen auf ihre Schlüsselkomponenten“ selbst und ohne fremde Hilfe verfasst, nicht andere als die in ihr angegebenen Quellen oder Hilfsmittel benutzt, alle vollständig oder sinngemäß übernommenen Zitate als solche gekennzeichnet, sowie die Dissertation in der vorliegenden oder einer ähnlichen Form noch bei keiner anderen in- oder ausländischen Hochschule anlässlich eines Promotionsgesuchs oder zu anderen Prüfungszwecken eingereicht zu haben. Ferner erkläre ich zuvor keine Promotion versucht zu haben.

_____

Marburg, den _____

Lebenslauf

Der Lebenslauf enthält persönliche Daten. Er ist deshalb nicht Bestandteil der Online-Veröffentlichung.

Publikationsliste:

Originalarbeiten (Papers):

[Lippmann et al., 2019] **Lippmann, C., Lötsch, J., & Ultsch, A.**: Computational functional genomics-based reduction of disease-related gene sets to their key components, *Bioinformatics, Vol. 35*(14), pp. 2362-2370. doi: 10.1093/bioinformatics/bty986, **2019**.

[Kringel et al., 2018] **Kringel, D., Lippmann, C., Parnham, M. J., Kalso, E., Ultsch, A., & Lötsch, J.**: A machine-learned analysis of human gene polymorphisms modulating persisting pain points to major roles of neuroimmune processes, *European Journal of Pain, Vol. 22*(10), pp. 1735-1756. doi: 10.1002/ejp.1270, **2018**.

[Lötsch et al., 2017] **Lötsch, J., Lippmann, C., Kringel, D., & Ultsch, A.**: Integrated Computational Analysis of Genes Associated with Human Hereditary Insensitivity to Pain. A Drug Repurposing Perspective, *Frontiers in Molecular Neuroscience, Vol. 10*(252), pp. doi: 10.3389/fnmol.2017.00252, **2017**.

Übersichtsarbeiten (Reviews):

[Lippmann et al., 2018] **Lippmann, C., Kringel, D., Ultsch, A., & Lötsch, J.**: Computational functional genomics-based approaches in analgesic drug discovery and repurposing, *Pharmacogenomics, Vol. 19*(9), pp. 783-797. **2018**.

[Lötsch et al., 2015] **Lötsch, J., Knothe, C., Lippmann, C., Ultsch, A., Hummel, T., & Walter, C.**: Olfactory drug effects approached from human-derived data, *Drug Discovery Today, Vol. 20*(11), pp. 1398-1406. doi: https://doi.org/10.1016/j.drudis.2015.06.012, **2015**.

Konferenzbeitrag:

[Lippmann et al., 2015] **Lippmann, C., Lötsch, J., & Ultsch, A.**: Understanding the Biological Functions of Gene Sets, *Proc. European Conference on Data Analysis (ECDA)*, pp. 28-29, Colchester, **2015**.

Danksagung

Diese Arbeit entstand über einen längeren Zeitraum und obwohl ich sie selbstständig geschrieben habe, waren viele direkt oder indirekt daran beteiligt. Daher möchte ich mich herzlich bei allen bedanken, die mich in dieser Zeit unterstützt haben.

Zuerst gebührt mein Dank meinem Doktorvater, Herrn Prof. Ultsch. Danke für Dein in mich gesetztes Vertrauen, für lösungsorientierte und konstruktive Diskussionen, die Anleitung zum wissenschaftlichen Arbeiten und für die Möglichkeit diese Arbeit in Deiner Arbeitsgruppe zu schreiben. Ganz besonders danke ich Dir auch für die Beruhigung bei meinen Panikattacken.

Für die Übernahme der Aufgaben des Zweitgutachters möchte ich mich herzlich bei Herrn Prof. Heider bedanken. Vielen Dank auch für Ihre konstruktive Kritik, die nochmal zum Nachdenken angeregt und einen neuen Blickwinkel eröffnet hat.

Eine sehr große Hilfe und Unterstützung war auch Herr Prof. Lötsch von der Goethe Universität Frankfurt. Ihnen gebührt Dank für das Korrekturlesen der biologischen Aspekte meiner Arbeit, für die Einladung nach Frankfurt ans Fraunhofer Institut, wo ich mich in Ihrer Arbeitsgruppe sehr wohl gefühlt habe, und für die konstruktive Zusammenarbeit, die in mehreren Papers mündete. Besonders danke ich Ihnen für die unfassbar schnellen Antworten auf E-Mails und die Unterstützung auch nach Beendigung des Beschäftigungsverhältnisses.

Dank schulde ich außerdem auch ganz besonders meinen Eltern und Andre für stundenlanges Korrekturlesen und die hilfreichen Kommentare! Ebenso danke ich Euch für die drängelnden Nachfragen nach neuen Kapiteln aber auch die Zeit, die ihr mir gelassen habt, die moralische Unterstützung, das Bekochen und vor allem dafür, dass ihr meine nicht immer guten Launen ausgehalten habt.

Für weitere Anmerkungen bin ich meinem ehemaligen Arbeitskollegen Michael dankbar. Schade, dass Du nicht mehr mit mir im Büro gesessen hast!

Ebenso möchte ich mich bei Dir, Florian, bedanken für die unkomplizierte und schnelle Hilfe bei diversen technischen Problemen vor allem nach der Umstellung von Windows 7 auf Windows 10 in den letzten Zügen der Arbeit!

Für die netten und aufmunternden Worte zwischendurch auf dem Flur möchte ich mich bei Dir, Oliver, bedanken. Das war manchmal der Motivationsschub, der gerade gefehlt hatte.

Ein großes Dankeschön nicht zuletzt auch Dir, Katja, für ein offenes Ohr bei größeren und kleineren Sorgen und Deine Warmherzigkeit. Ich habe mich immer sehr auf unsere morgendlichen Gespräche gefreut und wusste so immer schon, dass mein Tag gut starten wird.

Dank für die Finanzierung der Doktorandenstelle gebührt dem Fraunhofer Institut IME-TMP in Frankfurt am Main.

Kurzzusammenfassung

Durch den technischen Fortschritt der letzten Jahre werden in immer kürzerer Zeit immer größere Mengen von Daten mit tausenden und abertausenden Merkmalen gesammelt [Stańczyk/Jain, 2017], [H. Liu/Motoda, 2012]. Um diese unüberschaubar große Datenflut nutzbringend einzusetzen, werden computergestützte Auswertungsmethoden benötigt, die die Wissenschaftler bei der Extraktion von nützlichen Informationen bzw. Wissen unterstützen [Fayyad et al., 1996]. Ein Ansatz dazu sind Methoden der „Feature Selection".

In der vorliegenden Arbeit wird ein solcher Algorithmus beispielhaft für Gensätze entwickelt, die anhand von aktuellem Wissen über die genetische Architektur von Merkmalen oder Krankheiten gefunden wurden. Es ist dabei nicht notwendig, numerische Messwerte aus Experimenten für die einzelnen Gene zu kennen, da ein integrativer Ansatz verfolgt wird, der die Gene Ontology Wissensbasis [Ashburner et al., 2000] als Grundlage für das Kriterium zur Auswahl der wichtigsten Gene verwendet. Die hier vorgestellte funktionserhaltende, integrative Genselektion reduziert eine Menge von Genen auf ihre wichtigsten Elemente, indem für jedes Gen ein Score berechnet wird, der die Wichtigkeit der Gene beschreibt. Dieser Score wird mithilfe der Annotationen der Gene zu den signifikanten, biologischen Prozessen in der polyhierarchisch organisierten Gene Ontology Wissensbasis ermittelt. Der sich ergebende gerichtete, azyklische Graph (DAG) von signifikanten, biologischen Prozessen beschreibt die Genfunktionen des Datensatzes von Genen. Mit dem Gen-Score können die Gene in eine Rangfolge entsprechend ihrer Wichtigkeit gebracht werden. Die ersten k^* Gene bilden eine optimale Teilmenge, wobei diejenige Teilmenge der Gene ausgewählt wird, die die beste funktionserhaltende Eigenschaft hat. Die Funktionserhaltung wird dabei über Precision und Recall bzw. deren Verrechnung zum F_1-Maß bezüglich der Reproduktion des gesamten DAGs mit der gewählten Teilmenge bewertet.

Mit der funktionserhaltenden, integrativen Genselektion konnte für die untersuchten Gensätze der ursprüngliche DAG jeweils mit Recall und Precision von etwa 70% reproduziert werden, wobei nur etwa 5% der ursprünglichen Gene verwendet wurden.

Abstract

Recently, due to the technical progress and development more and more data with thousands and thousands of features is collected in increasingly less time [Stańczyk/Jain, 2017], [H. Liu/Motoda, 2012]. In order to make use of this unmanageable flood of data, computer-aided evaluation methods are needed to support scientists in extracting useful information or knowledge [Fayyad et al., 1996]. One approach to this are methods of "feature selection".

In the present work, such an algorithm is developed exemplarily for sets of genes found on the basis of current knowledge about the genetic architecture of features or diseases. It is not required to have numerical measurements from experiments for the individual genes, since an integrative approach is pursued, which uses the Gene Ontology Knowledge Base [Ashburner et al., 2000] as a basis for the criterion for the selection of the most important genes. The function-preserving, integrative gene selection presented here reduces a set of genes to their most important elements by calculating a score for each gene that describes the importance of the genes. This score is determined using the annotations of the genes to the significant biological processes in the polyhierarchically organized Gene Ontology knowledge base. The resulting directed acyclic graph (DAG) of significant biological processes describes the gene functions of the set of genes. With the gene score, genes can be ranked according to their importance. The first k^* genes form an optimal subset, whereby the subset of genes is selected that has the best function-preserving property. The preservation of function is evaluated by precision and recall and their combination to the F_1 measure, respectively, regarding the reproduction of the entire DAG with the selected subset.

With the function-preserving, integrative gene selection, the original DAG could be reproduced with recall and precision of about 70% for each of the examined data sets, using only about 5% of the original genes.

Inhaltsverzeichnis

1 Einleitung

Infolge der Digitalisierung von fast allen Lebensbereichen, rasant gestiegener Computerleistungen und der fortschreitenden Entwicklung von Technologien zur Datenerhebung werden heute in immer kürzerer Zeit immer größere Mengen von digitalen Daten generiert und gespeichert [H. Liu/Motoda, 2012]. Sowohl Menge, Varietät als auch Komplexität der Daten wachsen exponentiell [Critchlow/van Dam, 2016]. Allein in den letzten drei Jahren wurden bereits mehr Daten generiert als bisher in der gesamten Geschichte der Menschheit [Deshpande/Kumar, 2018]. Um diese unüberschaubar große Datenflut nutzbringend einzusetzen, werden computergestützte Auswertungsmethoden benötigt, die die Wissenschaftler bei der Extraktion von nützlichen Informationen bzw. Wissen unterstützen [Fayyad et al., 1996].

Ein Ansatz sind Algorithmen zur Datenreduktion, die aus der Menge der Daten anhand eines Kriteriums eine minimale, repräsentative Teilmenge auswählen [H. Liu/Motoda, 2012]. Durch die Auswahl einer solchen Teilmenge werden bezüglich des Kriteriums irrelevante und redundante Daten entfernt und nur die wichtigsten Elemente beibehalten, die die Gesamtmenge beschreiben [H. Liu/Motoda, 2012]. Dadurch wird die Menge der Daten reduziert, bereinigt und vereinfacht, was zu präziseren Ergebnissen und besserer Verständlichkeit führen kann [H. Liu/Motoda, 2012].

In der vorliegenden Arbeit wird ein solcher Algorithmus zur Datenreduktion beispielhaft für Gensätze entwickelt, da als Grundlage für das Kriterium zur Auswahl der wichtigsten Gene eine exzellente Datengrundlage [Gaudet/Dessimoz, 2017] zur Verfügung stand: die Gene Ontology Wissensbasis [Ashburner et al., 2000]. Diese ist eine sehr gut gepflegte und formal strukturierte Quelle von Wissen über biologische Prozesse, molekulare Funktionen und zelluläre Komponenten von Genfunktionen [Gaudet/Dessimoz, 2017]. Es werden Konzepte der Wissensrepräsentation und Ideen der Datenbionik vereint, um eine Methode, die sogenannte funktionserhaltende, integrative Genselektion, zu entwickeln, die aus einer unüberschaubar großen Menge von Genen eine optimale Teilmenge von wichtigen Genen herausfiltert.

Wissensrepräsentation und –verarbeitung ist ein Kerngebiet der Künstlichen Intelligenz, das versucht intelligentes menschliches Verhalten nachzubilden [Lämmel/Cleve, 2012]. Es beschäftigt sich damit, wie Wissen symbolisch dargestellt und in automatisierter Weise manipuliert werden kann, so dass logische

Schlussfolgerungen möglich sind [Brachman/Levesque, 2004]. Dazu muss Wissen für den Computer zugänglich gemacht, also so formalisiert werden, so dass es maschinell verarbeitbar ist [Lämmel/Cleve, 2012].

Die Datenbionik beschäftigt sich mit der Informationsverarbeitung in der Natur [Ultsch, 2014]. Prinzipien und Methoden, die dafür in der Natur erkannt wurden, werden in Computerprogrammen umgesetzt [Ultsch, 2014]. Dabei spielt das Prinzip der Optimierung eine wichtige Rolle [Bein et al., 2005]. Optimierung, die dank der Evolution bereits in der Natur vorgenommen wurde, sollte möglichst auch im Computer nachempfunden werden [Bein et al., 2005]. Dabei stellt die „Kunst des Übersehens, die zu einem neuen Sehen des sonst Unsichtbaren führt"[Schwemmer, 2008/2009, p. 14] eine wichtige Methode des wissenschaftlichen Fortschritts dar [Ultsch, 2019]. Diese Abstraktion hilft dabei, die wichtigsten Informationen herauszufiltern, sich nicht im Detail zu verlieren und möglichst verständliche Ergebnisse zu liefern [Ultsch, 2019]. Insbesondere für die Suche nach neuartigem, bisher unerkanntem und nützlichem Wissen in gegebenen Datensammlungen bieten sich datenbionische Methoden an [Ultsch, 2014].

In dieser Arbeit wird basierend auf der Wissensrepräsentation ein Score entwickelt, der die Wichtigkeit der einzelnen Objekte einer Menge von Daten bewertet. Anhand des Scores kann eine Rangfolge der Objekte bestimmt werden. Diese Rangfolge dient anschließend als Auswahlkriterium für die Objekte der optimalen Teilmenge mit Hilfe eines datenbionischen Algorithmus. Durch die Reduktion auf eine wichtige Teilmenge wird das Ziel verfolgt, eine möglichst verständliche, aber trotzdem repräsentative Darstellung der Gesamtmenge zu finden [Ultsch, 2019].

Wie bereits erwähnt, wird als Wissensrepräsentation die Gene Ontology [Ashburner et al., 2000] verwendet. Die Repräsentation dieses Wissens in Form eines gerichteten, azyklischen Graphen ermöglicht es Regeln aufzustellen, die logisches Schließen erlauben. Ergänzt wird die Gene Ontology durch die Gene Ontology Annotationen, mit deren Hilfe Gene und Genfunktionen in Verbindung gebracht werden können [The Gene Ontology Consortium, 2018a]. Durch eine Überrepräsentationsanalyse werden für einen Satz von Genen die Prozesse bzw. biologischen Funktionen bestimmt, in die diese Gene signifikant involviert sind. Für jeden Satz von Genen ergibt sich dadurch ein azyklischer, gerichteter Teilgraph, der das Wissen über diesen Gensatz repräsentiert. Um eine Rangfolge

der Gene zu erstellen, wird ein Score pro Gen errechnet, der sich aus der Struktur dieses Teilgraphen und den grundlegenden Ideen der Functional Abstraction [Ultsch/Lötsch, 2014a] ergibt. Anhand dieser Rangfolge werden mögliche optimale Teilmengen erstellt und mit einer datenbionischen Methode, die sich an der ökonomischen Idee der berechneten ABC-Analyse [Ultsch/Lötsch, 2015] orientiert, bewertet. Die Teilgraphen, die zu den jeweiligen in Frage kommenden Teilmengen durch weitere Überrepräsentationsanalysen gefunden werden, werden mit dem ursprünglichen Teilgraphen verglichen, wodurch sich ein Maß für die Optimalität der Größe der jeweiligen Teilmenge ergibt. In dieser optimalen Teilmenge sollen so wenig Gene wie möglich enthalten sein, aber gleichzeitig möglichst viele Prozesse des ursprünglichen Gensatzes erhalten bleiben. Die Funktionen der Gesamtmenge der Gene werden somit bei der Reduktion auf eine Teilmenge beibehalten, was durch die Integration einer Wissensbasis in den Reduktionsprozess ermöglicht wird.

Diese Arbeit ist wie folgt gegliedert: Das zweite Kapitel dient enzyklopädisch als Nachschlagewerk für eine Reihe grundlegender Definitionen, die für die funktionserhaltende, integrative Genselektion relevant sind. In Kapitel 3 werden Vorarbeiten anderer Autoren präsentiert und eine Systematik zu ihrer Strukturierung gegeben, um die in Kapitel 4 vorgestellte Methode in den Stand der wissenschaftlichen Forschung einzuordnen. Kapitel 4 umfasst neben der Vorgehensweise auch die wichtigsten Ergebnisse der funktionserhaltenden, integrativen Genselektion, die in Kapitel 5 kritisch diskutiert werden. Kapitel 6 fasst diese Arbeit zusammen und gibt einen Ausblick auf mögliche Ansätze zur Verbesserung bzw. Erweiterung der Methode.

Die wichtigsten Ergebnisse dieser Dissertation konnten bereits erfolgreich peer-reviewed publiziert werden: [Lippmann et al., 2019].

2 Grundlagen

In diesem Kapitel werden wichtige grundlegende Konzepte, auf die die vorliegende Arbeit aufbaut, eingeführt.

Die statistischen Grundlagen werden hauptsächlich für die Überrepräsentationsanalyse in Kapitel 2.5 wichtig sein. Grundlagen aus der Graphentheorie werden für die Beschreibung von Wissensbasen in Form einer Ontologie genutzt. Die datenbionischen Grundlagen ermöglichen in Kapitel 4.2 die Auswahl einer optimalen Teilmenge. Die Grundlagen der Wissensrepräsentation beschreiben wie Wissen formal dargestellt werden kann und werden für die Definition der Gene Ontology benötigt. Die grundlegenden Definitionen aus der Informationstheorie werden ebenfalls in Kapitel 4.2 in die Berechnung des Gen-Scores zur Erstellung einer Rangfolge der untersuchten Gene abhängig von ihrer Wichtigkeit einfließen. In den biologischen Grundlagen werden die biologischen Hintergründe umrissen, die in dieser Arbeit vor allem in Bezug auf die Gene Ontology Wissensbasis verwendet werden. Zudem wird der datenbionische Algorithmus, die berechnete ABC-Analyse, dessen Idee zur Bestimmung der optimalen Teilmenge verwendet wird, vorgestellt.

2.1 Statistische Grundlagen

Sofern nicht anders gekennzeichnet sind die folgenden Definitionen im Wesentlichen aus [Fahrmeir et al., 2016] entnommen.

Definition 2.1.1: Grundgesamtheit, Population
Die Menge aller für eine Fragestellung relevanten Objekte, an denen interessierende Größen erfasst werden, wird Grundgesamtheit oder Population genannt.

Definition 2.1.2: Stichprobe
Eine Stichprobe ist eine endliche Teilmenge der Grundgesamtheit.

Definition 2.1.3: Merkmal, Variable
Ein Merkmal oder eine Variable ist eine interessierende Größe.

Definition 2.1.4: Merkmalsausprägung
Die Merkmalsausprägung ist ein konkreter Wert einer Variablen für ein bestimmtes Objekt aus der Grundgesamtheit oder Stichprobe.

Definition 2.1.5: Zufallsvorgang
Ein Zufallsvorgang führt zu einem von mehreren sich gegenseitig ausschließenden Ergebnissen. Es ist vor der Durchführung ungewiss, welches Ergebnis tatsächlich eintreten wird.

Definition 2.1.6: Ereignisraum, Ereignis
Der Ereignisraum $\Omega = \{\omega_1, \dots, \omega_n\}$ ist die Menge aller Ergebnisse ω_i, $i = 1, \dots, n$, eines Zufallsvorgangs. Teilmengen von Ω heißen (Zufalls-)Ereignisse.

Definition 2.1.7: Wahrscheinlichkeit
Die (Eintritts-)Wahrscheinlichkeit für das Eintreten eines Ereignisses A wird durch $P(A)$ bezeichnet.

Definition 2.1.8: bedingte Wahrscheinlichkeit
Seien $A, B \subset \Omega$ und $P(B) > 0$, dann ist die bedingte Wahrscheinlichkeit von A unter B definiert als $P(A|B) \coloneqq \frac{P(A \cap B)}{P(B)}$.

Definition 2.1.9: Zufallsvariable, Realisierung
Eine Variable X, deren Werte oder Ausprägungen die Ergebnisse eines Zufallsvorgangs sind, heißt Zufallsvariable X. Die Zahl $x \in \mathbb{R}$, die X bei einer Durchführung des Zufallsvorgangs annimmt, heißt Realisierung oder Wert von X.

Definition 2.1.10: diskrete Zufallsvariable, Wahrscheinlichkeitsverteilung einer diskreten Zufallsvariablen
Eine Zufallsvariable X heißt diskret, falls sie nur endlich oder abzählbar unendlich viele Werte $x_1, x_2, \dots, x_k, \dots$ annehmen kann. Die Wahrscheinlichkeitsverteilung von X ist durch die Wahrscheinlichkeiten $P(X = x_i) = p_i, i = 1, 2, \dots, k, \dots$ gegeben.

Definition 2.1.11: Wahrscheinlichkeitsfunktion einer diskreten Zufallsvariable
Die Wahrscheinlichkeitsfunktion $f(x)$ einer diskreten Zufallsvariable X ist für $x \in \mathbb{R}$ definiert durch $f(x) = \begin{cases} P(X = x_i) = p_i, x = x_i \in \{x_1, x_2, \dots, x_k, \dots\} \\ 0, \text{ sonst.} \end{cases}$

Definition 2.1.12: hypergeometrisch verteilte, diskrete Zufallsvariable, hypergeometrische Verteilung
Eine diskrete Zufallsvariable X heißt hypergeometrisch verteilt mit Parametern n, M und N, kurz $X \sim H(n, M, N)$, wenn sie die Wahrscheinlichkeitsfunktion

$$f(x) = \begin{cases} \dfrac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}, x \in T \\ 0, \text{ sonst} \end{cases}$$

besitzt. Dabei ist T durch $\{\max(0, n-(N-M)), \dots, \min(n, M)\}$ gegeben. Die Verteilung heißt hypergeometrische Verteilung. Sie ergibt sich, wenn aus einer endlichen Grundgesamtheit von N Einheiten, von denen M eine Eigenschaft A besitzen, n-mal rein zufällig, aber ohne Zurücklegen gezogen wird. Die Zufallsvariable X modelliert dann die Anzahl der gezogenen Objekte mit der Eigenschaft A.

Definition 2.1.13: binomialverteilte, diskrete Zufallsvariable, Binomialverteilung

Eine diskrete Zufallsvariable X heißt binomialverteilt mit den Parametern n und π, kurz $X \sim B(n, \pi)$, wenn sie die Wahrscheinlichkeitsfunktion

$$f(x) = \begin{cases} \binom{n}{x}\pi^x(1-\pi)^{n-x}, & x = 0,1, \dots, n \\ 0, \text{ sonst} \end{cases}$$

besitzt. Die Verteilung heißt Binomialverteilung. Sie ergibt sich, wenn aus n unabhängigen Wiederholungen eines Zufallsvorgangs mit nur zwei möglichen Ausgängen mit konstanter Wahrscheinlichkeit π die Summe der Treffer, also nur eines Ausgangs, gebildet wird.

Definition 2.1.14: statistisches Testproblem, Nullhypothese, Alternative, einseitiges Testproblem

Ein statistisches Testproblem besteht aus einer Nullhypothese H_o und einer Alternative H_1, die sich gegenseitig ausschließen und Aussagen über die gesamte Verteilung oder über bestimmte Parameter des interessierenden Merkmals in der Grundgesamtheit beinhalten. Falls H_0 : „$\leq$“ gegen H_1 : „$>$“ bzw. H_0 : „$\geq$“ gegen H_1 : „$<$“ getestet wird, spricht man von einem einseitigen Testproblem.

Definition 2.1.15: statistischer Test, Prüfgröße, Teststatistik

Das geeignete Instrumentarium zur Lösung eines statistischen Testproblems liefert ein statistischer Test. Dieser stellt eine formale Entscheidungsregel dar, mit der es möglich sein soll zu unterscheiden, ob das in der Stichprobe beobachtete Verhalten ein reines Zufallsprodukt ist oder den Schluss auf die Grundgesamtheit zulässt. Ein solcher statistischer Test basiert auf einer Zufallsvariablen, der

sogenannten Prüfgröße bzw. Teststatistik. Diese ist so konstruiert, dass sie für das interessierende Testproblem sensibel ist und dass ihre Verteilung unter der Nullhypothese bekannt ist.

Definition 2.1.16: Test zum Signifikanzniveau α, Signifikanztest, Signifikanzniveau
Ein statistischer Test heißt Test zum Signifikanzniveau α, $0 < \alpha < 1$, oder Signifikanztest, falls die bedingte Wahrscheinlichkeit, dass die Alternative angenommen wird, obwohl die Nullhypothese wahr ist, kleiner als das Signifikanzniveau α ist; also $P(H_1 \text{ angenommen} | H_0 \text{ wahr}) \leq \alpha$ gilt.

Definition 2.1.17: p-Wert
Der p-Wert ist definiert als die Wahrscheinlichkeit unter H_0 den beobachteten Wert der Prüfgröße oder einen in Richtung der Alternative extremeren Wert zu erhalten. Ist der p-Wert kleiner als das vorgegebene Signifikanzniveau α, so wird H_0 verworfen. Ansonsten wird H_0 beibehalten.

Definition 2.1.18: 2x2-Kontingenztablle
Die allgemeine Form einer zweidimensionalen 2x2-Kontingenztablle ist gegeben durch Tabelle 2.1 [Everitt, 1992]. Eine Grundgesamtheit von N Objekten wird durch zwei Merkmale A und B, die jeweils zwei Merkmalsausprägungen haben, beschrieben, indem die Auftrittshäufigkeiten a, b, c und d der Kombinationen der Merkmalsausprägungen in die Tabelle eingetragen werden [Everitt, 1992].

Tabelle 2.1: Allgemeine Form einer zweidimensionalen 2x2-Kontingenztabelle. (Darstellung angelehnt an [Everitt, 1992].)

		Merkmal B		Summe
		Ausprägung B_1	Ausprägung B_2	
Merkmal A	Ausprägung A_1	a	b	$N_1 = a + b$
	Ausprägung A_2	c	d	$c + d$
Summe		$a + c$	$b + d$	$N = a + b + c + d$

Definition 2.1.19: Exakter Test nach Fisher, hypergeometrischer Test
Der exakte Test nach Fisher für 2x2-Kontingenztabellen geht auf die Theorien von R.A. Fisher [Fisher, 1935] zurück. Der Test ist exakt in dem Sinne, dass die exakte Wahrscheinlichkeitsverteilung der beobachteten Häufigkeiten berechnet

wird [Everitt, 1992]. Für feste Zeilen- und Spaltensummen der Häufigkeiten entspricht die Wahrscheinlichkeitsverteilung dem Ziehen ohne Zurücklegen aus einer endlichen Grundgesamtheit, das mit der hypergeometrischen Verteilung modelliert wird [Everitt, 1992].

Seien A und B zwei Merkmale, die jeweils nur zwei Merkmalsausprägungen annehmen, wobei die Ausprägungen von A als zwei unterschiedliche Populationen angesehen und diejenigen von B mit „Erfolg“ und „Misserfolg“ kodiert seien.

Man betrachte zwei voneinander und untereinander unabhängige Stichproben $X_1, X_2, \dots, X_{n_1}$, $X_i \sim B(1; p_1)$, $\forall\, i \in \{1, \dots, n_1\}$ und $Y_1, Y_2, \dots, Y_{n_2}$, $Y_j \sim B(1; p_2)$, $\forall\, j \in \{1, \dots, n_2\}$ des binären Merkmals B aus den zwei Populationen des Merkmals A. Für die Summe dieser Zufallsvariablen gilt dann:

$$X \coloneqq \sum_{i=1}^{n_1} X_i \sim B(n_1; p_1), Y \coloneqq \sum_{i=1}^{n_2} Y_i \sim B(n_2; p_2).$$

Sei $Z \coloneqq X + Y$. Der exakte Test von Fisher nutzt die Tatsache, dass die Zeilensummen $n_1$und n_2 in der 2x2-Kontingenztabelle Tabelle 2.2 durch die Stichprobenumfänge n_1 und n_2 festgelegt sind.

Tabelle 2.2: Kontingenztabelle zweier Merkmale A und B.

		Merkmal B		Summe
		Erfolg	Misserfolg	
Merkmal A	Population 1	X	$n_1 - X$	n_1
	Population 2	$Y = Z - X$	$n_2 - Y$	n_2
Summe		Z	$n_1 + n_2 - Z$	$n = n_1 + n_2$

Bedingt auf die Gesamtzahl $Z = z$ der Erfolge (d.h. die Spaltensummen werden auch festgehalten) ist X die einzige Zufallsvariable, da die anderen Einträge der Tabelle durch die Realisation x von X und die Randsummen bestimmt sind. X ist dann hypergeometrisch verteilt, $X \sim H(z, n_1, n)$, mit $n = n_1 + n_2$. Es gilt nach Definition 2.1.12

$$P(X = x | Z = z) = \frac{\binom{n_1}{x}\binom{n - n_1}{z - x}}{\binom{n}{z}}.$$

Wird die Nullhypothese $H_0: p_1 \leq p_2$ gegen $H_1: p_1 > p_2$ getestet, ergibt sich der linksseitige p-Wert als

$$\text{p-Wert}_{\text{links}} = \sum_{i=\max(0,n_1-x)}^{x} P(X = i|Z = z), \text{ falls } x \leq z \cdot \frac{n_1}{n}$$

[Looney/Hagan, 2015]. Für den Test von $H_0: p_1 \geq p_2$ gegen $H_1: p_1 < p_2$, ergibt sich der rechtsseitige p-Wert entsprechend als

$$\text{p-Wert}_{\text{rechts}} = \sum_{i=x}^{\min(n,M)} P(X = i|Z = z), \text{ falls } x \geq z \cdot \frac{n_1}{n}$$

[Looney/Hagan, 2015]. Der einseitige exakte Test nach Fisher wird auch hypergeometrischer Test genannt [Piegorsch, 2015, p. 279].

2.2 Graphentheoretische Grundlagen

Die Definitionen dieses Abschnitts sind im Wesentlichen aus [Hartmann, 2015] entnommen, sofern nicht anders gekennzeichnet.

Definition 2.2.1: Euklidischer Abstand
Sind $x = (x_1, \dots, x_n)$ und $y = (y_1, \dots, y_n)$ zwei Elemente des $\mathbb{R}^n$, so heißt die Metrik $d(x, y) = \sqrt{(x_1 - y_1)^2 + \cdots + (x_n - y_n)^2}$ euklidische Metrik [Walz, 2016]. Sie überträgt den intuitiven Abstandsbegriff aus dem $\mathbb{R}^2$ und $\mathbb{R}^3$, den euklidischen Abstand, in die allgemeinere Situation des $\mathbb{R}^n$ [Walz, 2016].

Definition 2.2.2: endlicher, gerichteter Graph, gerichtete Kante, Anfangspunkt, Endpunkt
Ein endlicher, gerichteter Graph $G := (V, E)$ besteht aus einer endlichen Menge von Knoten $V = V(G)$ und aus einer endlichen Menge von geordneten Paaren, $E = E(G) = \{(x_i, x_j): x_i, x_j \in V\}$, den gerichteten Kanten von G. Die Richtung der Kante (x_i, x_j) zeigt dabei von x_i nach x_j. x_i heißt Anfangspunkt und x_j Endpunkt der gerichteten Kante (x_i, x_j).

Definition 2.2.3: gerichtete Kantenfolge
Existiert für zwei Knoten x_1 und x_k eines gerichteten Graphen G mit $V(G) = \{x_1, \dots, x_k, \dots, x_n\}$ eine Folge $x_1, (x_1, x_2), x_2, \dots, x_i, (x_i, x_{i+1}), x_{i+1}, \dots, x_{k-1}, (x_{k-1}, x_k), x_k$ mit $i \in \{1, \dots, n\}$ von

Knoten und gerichteten Kanten, die diese Knoten miteinander verbinden, dann heißt die Folge der gerichteten Kanten $(x_1, x_2), \dots, (x_{k-1}, x_k)$ gerichtete Kantenfolge von x_1 nach x_k.

Definition 2.2.4: gerichteter Pfad
Ein gerichteter Pfad von x_1 nach x_n ist eine gerichtete Kantenfolge, in der alle Knoten verschieden sind.

Definition 2.2.5: gerichteter Kreis
Ein gerichteter Kreis ist eine gerichtete Kantenfolge, in der alle Knoten bis auf Anfangs- und Endpunkt verschieden sind.

Definition 2.2.6: azyklischer, gerichteter Graph (DAG)
Ein gerichteter Graph, in dem es keinen gerichteten Kreis gibt, heißt azyklischer, gerichteter Graph (DAG).

Definition 2.2.7: Adjazenzmatrix
Ein gerichteter Graph kann in Form einer Adjazenzmatrix repräsentiert werden. Ein Element a_{ij} der Adjazenzmatrix ist genau dann 1, falls es eine gerichtete Kante von Knoten x_i nach Knoten x_j gibt, und hat sonst den Wert 0.

Definition 2.2.8: Wurzel
Existiert in einem DAG nur ein Knoten x_j, dessen zugehörige Spaltensumme $\sum_i a_{ij}$ der Adjazenzmatrix den Wert 0 hat, bezeichnet man den Knoten x_j als die Wurzel des DAGs. Im Folgenden wird immer von einem DAG mit Wurzel ausgegangen, falls nicht explizit anders gekennzeichnet.

Definition 2.2.9: Elter, Kind
Existiert eine gerichtete Kante von x_i nach x_j, so heißt x_i Elter von x_j und x_j Kind von x_i.

Definition 2.2.10: Vorfahre, Nachfahre
Existiert ein gerichteter Pfad von x_i nach x_j, dann heißt x_i Vorfahre von x_j und entsprechend x_j Nachfahre von x_i.

Definition 2.2.11: Blatt
Knoten ohne Nachkommen heißen Blätter.

2.3 Grundlagen der Wissensrepräsentation und der Datenbionik

Definition 2.3.1: Daten
Daten sind eine Menge von unabhängigen, isolierten Fakten, Messungen, Objekten, Worten, Zahlen oder Symbolen [Wang et al., 2001].

Definition 2.3.2: Information
Informationen sind organisierte, strukturierte, gegebenenfalls adjustierte und analysierte Daten, die in einen Kontext gesetzt werden [Wang et al., 2001].

Bemerkung 2.3.1:
Die Definition von Wissen ist komplex, umstritten und kann auf viele Arten und Weisen interpretiert werden [Wang et al., 2001]. Ein ausführlicher Versuch einer übersichtlichen Auflistung der verschiedenen sprachlichen Verwendungen des Begriffs „Wissen“ wird in [Riethmüller, 2012] gemacht. In dieser Arbeit wird jedoch eine eher technische Definition nach [Ultsch, 1987] verwendet:

Definition 2.3.3: Wissen
Repräsentiertes Wissen ist eine symbolische Repräsentation von Objekten, Fakten und Regeln in einer operationalen Form für einen Interpreter mit Symbolverarbeitungskompetenz [Ultsch, 1987], wobei Objekte, Fakten und Regeln und deren symbolische Repräsentation im Sinne der Prädikatenlogik [Dengel, 2011] verstanden werden und ein Interpreter mit Symbolverarbeitungskompetenz z.B. ein Mensch oder Beweissysteme für Prädikatenlogik erster Stufe wie Prolog [Clocksin/Mellish, 2012; Sterling et al., 1994] ist [Ultsch, 1987].

Definition 2.3.4: Wissensbasis
Eine Wissensbasis ist eine organisierte Sammlung von Wissen zusammen mit Operationen, um auf das Wissen zuzugreifen und es manipulieren zu können [Ultsch, 1987].

Definition 2.3.5: kontrolliertes Vokabular
Ein kontrolliertes Vokabular ist eine endliche Liste von Begriffen [Lassila/McGuinness, 2001], die eindeutig festgelegt sind [Merkl/Waack, 2013].

Definition 2.3.6: Taxonomie
Eine Taxonomie ist eine Hierarchie von Begriffen aus einem kontrollierten Vokabular, die Elemente in einer Über-/Unterordnung darstellt [Jerroudi, 2010].

Definition 2.3.7: Ontologie
Wissen wird häufig in Form einer Ontologie repräsentiert [Brewster/O'Hara, 2007]. Der Begriff der Ontologie, wörtlich „die Lehre vom Seienden" [Dengel, 2011], stammt aus der Philosophie und bezeichnet dort die Beschäftigung mit real existierenden Dingen [Merkl/Waack, 2013].
In der Informatik werden für gewöhnlich die Definition von [Gruber, 1993] als explizite Spezifikation einer Konzeptualisierung und diejenige von [Uschold/Gruninger, 1996] als gemeinsames Verständnis eines Interessensgebiets zusammengefasst zur Definition: „Eine Ontologie ist eine formale, explizite Spezifikation einer gemeinsamen Konzeptualisierung." [Dengel, 2011, p. 65]. Dabei wird eine gemeinsame Konzeptualisierung verstanden als ein Modell von Objekten, Begriffen und anderen Entitäten, von denen angenommen wird, dass sie in einem bestimmten Interessensgebiet existieren [Guarino et al., 2009], und den zwischen ihnen geltenden Relationen [Baader et al., 2009]. Eine formale, explizite Spezifikation bedeutet, dass das Modell in einer eindeutigen Sprache spezifiziert sein soll, die sowohl menschen- als auch maschinenverständlich ist [Baader et al., 2009].

Bemerkung 2.3.2:
In der Regel wird diese formale Definition einer Ontologie umgesetzt, indem eine Begriffshierarchie in Form einer Taxonomie mit zusätzlichen Relationen zwischen den Begriffen versehen wird und Regeln aufgestellt werden, die die Gegebenheiten in einer Domäne beschreiben, immer wahr sein müssen und mit Hilfe derer logische Schlüsse gezogen werden können [Dengel, 2011].

Definition 2.3.8: berechnete ABC-Analyse
Seien $\{x_i\}_{i=1}^{n}$ die Realisationen einer diskreten Zufallsvariable X. Die berechnete ABC-Analyse [Ultsch/Lötsch, 2015] ist ein datenbionisches Verfahren, das die Menge $\{x_i\}_{i=1}^{n}$ in drei Teilmengen A, B und C einteilt, um die wichtigsten Realisationen zu identifizieren [Ultsch/Lötsch, 2015]. In Teilmenge A werden möglichst die (oftmals) wenigen, wichtigen Realisationen zusammengefasst [Juran, 1975], die also aus ökonomischer Sicht mit minimalem Aufwand maximalen Ertrag ermöglichen [Ultsch/Lötsch, 2015]. Teilmenge B enthält diejenigen Realisationen, wo Aufwand und Ertrag weitestgehend proportional zueinander sind [Ultsch/Lötsch, 2015] und Teilmenge C die (meist) vielen, trivialen Realisationen [Juran, 1975], durch die nur mit großem Aufwand zusätzlicher Ertrag gewonnen werden kann [Ultsch/Lötsch, 2015].

Es wird eine graphische Darstellung der gegebenen Daten, die ABC-Kurve, im $\mathbb{R}^2$ erstellt [Ultsch/Lötsch, 2015]. Die Realisationen werden dazu der Größe nach absteigend–sortiert, es gilt also bis auf Umbenennung $x_i \geq x_{i+1} \forall\, i \in \{1, \dots, n\}$ [Ultsch/Lötsch, 2015]. Auf der x-Achse wird $\frac{i}{n}, i \in \{1, \dots, n\}$ aufgetragen, wodurch der Aufwand, den der zugehörige Ertrag kostet, dargestellt wird [Ultsch/Lötsch, 2015]. Die zugehörigen Werte der y-Achse, den Ertrag, bilden die kumulierten, anteiligen Werte der einzelnen Realisationen $\sum_{j=1}^{k} \frac{x_j}{\sum_{i=1}^{n} x_{i_i}}, \forall\, k \in \{1, \dots, n\}$ [Ultsch/Lötsch, 2015]. Um eine stetige [Forster, 2016] ABC-Kurve zu erhalten, werden die Werte zwischen den empirischen Datenpunkten mit Hilfe von quadratischen Splines [Friedrich/Pietschmann, 2010] interpoliert [Ultsch/Lötsch, 2015].

Die Grenzen zwischen den Teilmengen A, B und C werden rechnerisch bestimmt [Ultsch/Lötsch, 2015]. Die erste Grenze A_x ist der x-Wert des Punktes auf der ABC-Kurve, der den minimalen, euklidischen Abstand zum Punkt $(0, 1)$ hat, dem Punkt von minimalem Aufwand ($x = 0$) und maximalem Ertrag ($y = 1$) [Ultsch/Lötsch, 2015]. Die zweite Grenze B_x wird ganz ähnlich bestimmt. Sie ist der x-Wert des Punktes auf der ABC-Kurve, der den minimalen, euklidischen Abstand zum Punkt $(x_0, 1)$ hat, wobei x_0 bestimmt wird als derjenige Punkt auf der x-Achse, in dem die erste Ableitung [Forster, 2016] der ABC-Kurve gleich 1 ist, also Aufwand exakt dem Ertrag entspricht [Ultsch/Lötsch, 2015]. Die Grenze zwischen Teilmengen A und B, ist gegeben als $t_{AB} = \min(A_x, B_x)$ [Ultsch/Lötsch, 2015]. Die Teilmenge A enthält dann alle Realisationen für die gilt $A = \{x_i | x_i \leq icdf(t_{AB}), \forall\, i \in \{1, \dots, n\}\}$, wobei $icdf(p)$ die Quantilfunktion [Steland, 2013] zur stetigen ABC-Kurve ist [Ultsch/Lötsch, 2015]. Die Grenze zwischen Teilmengen B und C ist entsprechend gegeben als $t_{BC} = \max(A_x, B_x)$ [Ultsch/Lötsch, 2015]. Teilmenge C enthält alle Realisationen für die gilt $C = \{x_i | x_i > icdf(t_{BC}), \forall i \in \{1, \dots, n\}\}$ [Ultsch/Lötsch, 2015].

2.4 Informationstheoretische Grundlagen

Definition 2.4.1: Information Retrieval System, Information Retrieval
Ein Information Retrieval System erlaubt einem User katalogisierte und indizierte Sammlungen von Dokumenten, Bildern oder Tonaufnahmen zu durchsuchen [Chowdhury, 2010]. Ziel von Information Retrieval ist das Heraussuchen

von relevanten Informationen bzw. Dokumenten aus einem Information Retrieval System [Chowdhury, 2010].

Definition 2.4.2: Precision, Recall
Um die Güte eines Information Retrieval zu beurteilen, werden meist zwei Maße, Precision und Recall, verwendet [Baeza-Yates/Ribeiro, 2011]. Sie werden definiert als

$\text{Precision} = \frac{\text{Anzahl gefundener, relevanter Informationen}}{\text{Anzahl gefundener Informationen}}$ und

$\text{Recall} = \frac{\text{Anzahl gefundener, relevanter Informationen}}{\text{Anzahl relevanter Informationen in der Sammlung}}$ [Salton/McGill, 1984]. Die Ergebnisse des Information Retrieval können auch in Form einer 2x2-Kontingenztabelle ausgedrückt werden, siehe Tabelle 2.3.

Tabelle 2.3: 2x2-Kontingenztabelle für ein Information Retrieval System (Darstellung angelehnt an [Chowdhury, 2010])

		Information		Summe
		Gefunden	Nicht gefunden	
Information	Relevant	tp	fp	$tp + fp$
	Nicht relevant	fn	tn	$fn + tn$
Summe		$tp + fn$	$fp + tn$	$tp + fp + fn + tn$

Precision und Recall können dann formuliert werden als

$$\text{Precision} = \frac{tp}{tp + fp} \tag{2.1}$$

und

$$\text{Recall} = \frac{tp}{tp + fn}, \tag{2.2}$$

wobei tp für richtig positiv, also gefunden und relevant, fp für falsch positiv, also gefunden aber nicht relevant, fn für falsch negativ, also nicht gefunden aber relevant und tn für richtig negativ, also nicht gefunden und auch nicht relevant steht [Chowdhury, 2010].

Precision misst die Qualität der Resultate und Recall deren Vollständigkeit [Pentreath, 2015, p. 136].

Definition 2.4.3: F_β-Maß
Das F_β-Maß kombiniert Precision und Recall zu einer einzigen Maßzahl und erlaubt es Precision und Recall verschiedene Gewichte zuzuordnen [Baeza-Yates/Ribeiro, 2011]. Es ist definiert als

$$F_\beta = (1 + \beta^2) \cdot \frac{\text{Precision} \cdot \text{Recall}}{\beta^2 \cdot \text{Precision} + \text{Recall}},$$

wobei β die relative Wichtigkeit von Precision und Recall definiert [Baeza-Yates/Ribeiro, 2011]. Ist $\beta = 0$, so wird nur Precision, für $\beta = \infty$, nur Recall in die Berechnung einbezogen [Baeza-Yates/Ribeiro, 2011]. Für $\beta = 0.5$ wird Recall als halb so wichtig erachtet wie Precision [Baeza-Yates/Ribeiro, 2011].

Die am häufigsten verwendete Gewichtung ist $\beta = 1$, woraus sich gleiche Gewichte für Precision und Recall ergeben [Baeza-Yates/Ribeiro, 2011]. Das F_1-Maß bildet das harmonische Mittel von Precision und Recall [Baeza-Yates/Ribeiro, 2011]. Es ist entsprechend definiert als $F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$ [Baeza-Yates/Ribeiro, 2011]. Der Wert des F_1-Maßes ist nur dann hoch, wenn sowohl Precision als auch Recall einen hohen Wert haben [Baeza-Yates/Ribeiro, 2011]. Es kann als bester Kompromiss zwischen Precision und Recall angesehen werden [Baeza-Yates/Ribeiro, 2011].

Definition 2.4.4: Informationsgehalt
Um Information numerisch messen zu können, wird der Begriff des Informationsgehalts definiert. Sei X eine diskrete Zufallsvariable, die die Werte $x_1, x_2, \dots, x_n$ mit den Wahrscheinlichkeiten $p_1, p_2, \dots, p_n$ annimmt, wobei $\sum_{i=1}^{n} p_i = 1$ gilt [Spătaru, 2013], dann ist der Informationsgehalt einer Realisierung von X definiert als

$$IC(x_i) = c \cdot \log_a\left(\frac{1}{p_i}\right) = -c \cdot \log_a(p_i),\ \forall i \in \{1, \dots, n\}$$

wobei a und c positive Konstanten sind [Applebaum, 1996].

Definition 2.4.5: Shannon-Entropie
Die Entropie $\mathrm{H}(X)$ der Zufallsvariable X, nach Claude Shannons Definition in [Shannon, 1948] auch Shannon-Entropie genannt, ist definiert als das mit den Eintrittswahrscheinlichkeiten gewichtete arithmetische Mittel des Informationsgehalts

$$\mathrm{H}(X) = \sum_{i=1}^{n} p_i \, IC(x_i) = -c \cdot \sum_{i=1}^{n} p_i \log_a(p_i),$$

wobei die Konvention $0 \cdot \log(0) = 0$ angenommen wird, um Wohldefiniertheit zu garantieren [Karmeshu/Pal, 2003].

2.5 Biologische Grundlagen

Bemerkung 2.5.1:
Den Begriff des Gens zu definieren ist nicht trivial, da darunter je nach Betrachtungsweise Verschiedenes verstanden werden kann. Einen Überblick liefert [Graw, 2015b, pp. 6-7]. In dieser Arbeit wird der Begriff des Gens in folgendem Sinne verstanden:

Definition 2.5.1: Gen
Ein Gen ist ein Abschnitt auf der Desoxyribonukleinsäure (DNA) eines Chromosoms, der bestimmte Proteinbausteine codiert oder eine bestimmte Regulationsfunktion hat [Anhäuser et al., 2001].

Definition 2.5.2: Genprodukt
Die in einem Gen gespeicherte Erbinformation erlaubt es einer Zelle ein Makromolekül oder möglicherweise mehrere verschiedene Makromoleküle zu erzeugen [Thomas, 2017]. Ein solches Makromolekül wird Genprodukt genannt und kann entweder ein Protein (häufigster Fall) oder eine nicht-codierende Ribonukleinsäure (RNA) sein [Thomas, 2017].

Bemerkung 2.5.2
Im Rahmen dieser Arbeit wird, obwohl biologisch nicht ganz korrekt, zwischen Gen und Genprodukt der Einfachheit halber nicht unterschieden. Es wird mit Namen von Genen gearbeitet es sei aber darauf hingewiesen, dass die biologischen Prozesse in der Regel nicht über die Gene sondern über deren Produkte realisiert werden [Lötsch, 2019]. Die Namen von Gen und Genprodukt sind nicht unbedingt identisch [Lötsch, 2019]. Beispielsweise codiert das Gen *OPRM1* das Genprodukt μ-Opioid Rezeptor, welches als Protein mit „MOR“ abgekürzt wird [Lötsch, 2019]. Da wir uns in dieser Arbeit aber auf die Gene Ontology beziehen,

verwenden wir durchgehend die Namen der Gene, während die Namen der kodierten Genprodukte einfach nachgeschlagen werden können [Lötsch, 2019].

Definition 2.5.3: molekulare Maschine
Ein Genprodukt kann als eine molekulare Maschine agieren. Das heißt es kann eine chemische Aktion, also eine Aktivität, vollführen [Thomas, 2017].

Definition 2.5.4: makromolekularer Komplex
Genprodukte von verschiedenen Genen können sich zu größeren molekularen Maschinen zusammenschließen, was makromolekularer Komplex genannt wird [Thomas, 2017].

Definition 2.5.5: Genfunktion
Genprodukte oder makromolekulare Komplexe führen zelluläre Prozesse oder Aktivitäten aus, die im Folgenden auch als Genfunktion bezeichnet werden, obwohl eigentlich nicht das Gen selbst eine Funktion ausführt, sondern das Genprodukt bzw. ein makromolekularer Komplex [Thomas, 2017].

Definition 2.5.6: Gene Ontology Wissensbasis
Das Ziel des Gene Ontology Consortiums [Ashburner et al., 2000], [The Gene Ontology Consortium, 2017] ist es, ein strukturiertes, dynamisches, präzise definiertes, allgemein gültiges und kontrolliertes Vokabular zu gewährleisten, mit dem die Genfunktionen von Genprodukten in jedem beliebigen Organismus beschrieben werden können [Ashburner et al., 2000]. Das Resultat dieser Bemühungen ist die Gene Ontology Wissensbasis. Sie besteht aus zwei Komponenten: der Gene Ontology (siehe Definition 2.5.7) und den Gene Ontology Annotationen (siehe Definition 2.5.11) [The Gene Ontology Consortium, 2018c].

Definition 2.5.7: Gene Ontology (GO)
Die Gene Ontology (GO) ist die zurzeit erfolgreichste biologische Ontologie [Hastings, 2017]. Sie liegt in Form dreier DAGs vor [Gaudet et al., 2017], die verschiedene biologische Aspekte repräsentieren: (i) Molekulare Funktionen (MF), von Genen ausgeführte Aktivitäten auf molekularer Ebene, (ii) biologische Prozesse (BP), größere Prozesse ausgeführt durch multiple molekulare Aktivitäten, und (iii) zelluläre Komponenten (CC), die Orte in der zellulären Struktur, an denen ein Gen seine Funktion erfüllt [Ashburner et al., 2000], [The Gene Ontology Consortium, 2018d].

Definition 2.5.8: GO Term, Relation
Ein GO Term ist ein Begriff, der eine Genfunktion beschreibt [The Gene Ontology Consortium, 2018d]. Die zulässige Menge all dieser Begriffe bildet die Knoten in den drei DAGs der GO und die gerichteten Kanten repräsentieren die Relationen zwischen den GO Termen [The Gene Ontology Consortium, 2018d]. Die in dieser Arbeit verwendeten Relationen sind is a und part of.

Definition 2.5.9: is a Relation
Die is a Relation formt die grundlegende Struktur der GO [The Gene Ontology Consortium, 2018d]. Haben zwei GO Terme T_1 und T_2 die Relation T_2 is a T_1, bedeutet dies, dass Knoten T_2 ein Subtyp von Knoten T_1 ist [The Gene Ontology Consortium, 2018d]. Es existiert dann eine gerichtete Kante von T_1 nach T_2 und der Elter T_1 beschreibt eine Genfunktion, von der das Kind T_2 nur ein Detail beschreibt [The Gene Ontology Consortium, 2018d].

Definition 2.5.10: part of Relation
Die part of Relation beschreibt Teile-Ganzes-Beziehungen zwischen zwei GO Termen. Haben zwei GO Terme T_1 und T_2 die Relation T_2 part of T_1, bedeutet dies, dass Kind T_2 notwendigerweise Teil von Elter T_1 ist [The Gene Ontology Consortium, 2018d] und entsprechend eine gerichtete Kante von T_1 nach T_2 existiert.

Bemerkung 2.5.3:
Die Relationen is a und part of zwischen den Knoten in der GO ermöglichen logisches Schließen in dem Sinne, dass die biologische Korrektheit innerhalb eines Pfades durch das Gene Ontology Consortium sichergestellt wird. Gilt beispielsweise T_1 is a T_2 und T_2 is a T_3, dann gilt auch T_1 is a T_3. Analog können auch für part of Relationen und die Kombination von is a und part of Relationen Schlüsse gezogen werden [The Gene Ontology Consortium, 2018d].

Definition 2.5.11: Gene Ontology Annotation (GO Annotation), Evidenzcode
Eine Gene Ontology Annotation (GO Annotation) ist eine Zuordnung zwischen einem spezifischen Gen und einem zugehörigen GO Term [The Gene Ontology Consortium, 2018a]. Die GO Annotation wird immer mit einem Evidenzcode angegeben, der kodiert wie die Zuordnung belegt wird. Z.B. ob ein Biokurator die Zuordnung händisch überprüft hat, der Beweis nur auf wissenschaftlicher Literatur beruht oder ein Experiment durchgeführt wurde

[The Gene Ontology Consortium, 2018a]. Eine Übersicht der Evidenzcodes mit entsprechender Erklärung findet sich unter [The Gene Ontology Consortium, 2018b].

Definition 2.5.12: Überrepräsentationsanalyse auf Grundlage der Gene Ontology Wissensbasis (ORA)

Die Überrepräsentationsanalyse wurde für die statistische Auswertung von Mengen von Genen entwickelt [Backes et al., 2007]. In dieser Arbeit wird nur die Überrepräsentationsanalyse auf Grundlage der Gene Ontology Wissensbasis (ORA) behandelt, da diese die am häufigsten verwendete Datengrundlage ist [Bennett/Bushel, 2017]. Die ORA stellt mit Hilfe der GO Annotationen fest, ob ein bestimmter GO Term für die gegebene Menge der Gene in Bezug auf eine Referenzmenge [Backes et al., 2007] über- oder unterrepräsentiert ist [Bennett/Bushel, 2017], d.h. ob zu einem GO Term signifikant häufiger oder seltener Gene aus der gegebenen Menge annotiert sind als durch reinen Zufall zu erwarten gewesen wäre [Backes et al., 2007]. Dazu nutzt die ORA einen statistischen Test, meist den exakten Test nach Fisher, den χ^2-Test [Fahrmeir et al., 2016] oder den Binomialtest [Fahrmeir et al., 2016] [Khatri/Drăghici, 2005]. Diese sind geeignet, da sie die hypergeometrische Verteilung verwenden bzw. approximieren, die die Wahrscheinlichkeitsverteilung für dieses Problem modelliert [Khatri/Drăghici, 2005]. Das Resultat einer ORA sind die signifikanten GO Terme, die die Genfunktionen des gegebenen Gensatzes beschreiben [Lippmann et al., 2018].

3 Verwandte Arbeiten zur Reduktion einer Menge von Genen auf ihre wichtigsten Elemente

In diesem Kapitel wird ein Überblick über verschiedene bestehende Verfahrenstechniken zur Reduktion einer Menge von Genen auf ihre wichtigsten Elemente gegeben. Außerdem werden Beispiele verwandter Arbeiten vorgestellt.

Durch den technischen Fortschritt der letzten Jahre werden einerseits mehr und mehr Daten mit tausenden und abertausenden Merkmalen gesammelt, andererseits werden immer noch passende Methoden gesucht, die es erlauben herauszufinden, welche der Daten essentiell und welche unbrauchbar oder irrelevant sind [Stańczyk/Jain, 2017]. Diese Art von Methoden wird unter dem Begriff der „Feature Selection" zusammengefasst. Das Hauptziel von Feature Selection ist das Entfernen von Features, die nicht informativ also irrelevant oder redundant sind, um dadurch Dimensionsreduktion zu erreichen, Wissen zu entdecken und gesammelte Daten zu erforschen [Stańczyk/Jain, 2017]. Die Anwendungsbereiche von Feature Selection-Verfahren sind vielfältig. Sie werden beispielsweise für Textmining (z.B. [Forman, 2003], [T. Liu et al., 2003]), Bildverarbeitung (z.B. [Muštra et al., 2012], [Swets/Weng, 1995]), Data Mining (z.B. [H. Liu/Motoda, 2012]), Machine Learning (z.B. [Vafaie/De Jong, 1992], [Hall, 1999], [Khalid et al., 2014]), Mustererkennung (z.B. [Pudil et al., 1994], [Kittler, 1975]) [Mahajan/Singh, 2016], [Jović et al., 2015]) und auch in der Bioinformatik eingesetzt.

In der Bioinformatik werden diese Verfahren unter anderem genutzt, um Daten aus Microarray Experimenten [Tan/Lynch, 2012] zu analysieren [Hira/Gillies, 2015]. Diese Daten liegen für gewöhnlich in Form einer zweidimensionalen Matrix mit n Zeilen und m Spalten vor [Huawen Liu et al., 2010]. Die m Spalten repräsentieren m verschiedene Gene und jede Zeile (1 bis n) enthält eine Stichprobe, für die die Expression dieser Gene gemessen wurde [Huawen Liu et al., 2010]. Die Einträge der Matrix bilden die Expressionswerte. Sie beschreiben für jede Stichprobe die Quantität der mRNA-Moleküle [Graw, 2015a], die von der DNA auf den Microarray-Chips [Tan/Lynch, 2012] abgelesen werden [Tan/Lynch, 2012], also die erhöhte oder erniedrigte Expression der getesteten Gene [Tan/Lynch, 2012].

Expressionswerte aus Microarray Experimenten werden analysiert, um hauptsächlich drei verschiedene Probleme zu lösen [Tarca et al., 2006]:

Erstens die Unterscheidung von gesunden und kranken Stichproben [Tarca et al., 2006], z.B. kanzeröses Gewebe und nicht-kanzeröses Gewebe (z.B. in [Furey et al., 2000]).

Zweitens die Klassifikation [Runkler, 2015b] von Stichproben [Tarca et al., 2006], d.h. es ist bereits für andere Stichproben z.B. deren Typ von Krebs bekannt und eine neue Stichprobe kann anhand der Expressionswerte der einzelnen Gene einer der vorgegebenen Klassen zugeordnet werden (z.B. in [Singh et al., 2002]).

Drittens die Clusterung [Runkler, 2015a] von Stichproben [Tarca et al., 2006] bei denen noch keine Klassen vorgegeben sind, sondern Gruppen in den Stichproben gesucht werden, die ähnliche Expressionswerte der einzelnen Gene haben (z.B. in [de Souto et al., 2008]).

Eine häufig auftretende Schwierigkeit bei der Analyse von biologischen Datenmatrizen stellt ihre Größe dar [Dramiński et al., 2007]. Eine sehr kleine Anzahl von Stichproben (z.B. Versuchspersonen) wird einer sehr viel größeren Menge von Merkmalsausprägungen (z.B. Genen) gegenübergestellt [Dramiński et al., 2007]. Für gewöhnlich werden weniger als hundert Stichproben $n < 100$ gemessen [Huawen Liu et al., 2010]. Für jede Stichprobe werden dabei je nach Quelle tausende bis zehntausende [Huawen Liu et al., 2010], zehntausende [Tan/Lynch, 2012] oder sogar hunderttausende [Hira/Gillies, 2015] Gene $m >> 1.000$ analysiert. Bei der komplexen Prozedur zur Datenerhebung können die Daten einerseits durch Messfehler verunreinigt werden [Huawen Liu et al., 2010], andererseits neigen Klassifikatoren [Runkler, 2015b] für Probleme mit kleinem n und sehr großem m zu Overfitting [Huawen Liu et al., 2010]. Das bedeutet, dass der Algorithmus, der anhand der bekannten Klassen „lernt" wie neue Stichproben bewertet werden, „auswendig lernt" und neue Stichproben entsprechend falsch klassifiziert [Paris et al., 2004]. Das Phänomen ist als „Fluch der hohen Dimension" bekannt [Bellman, 1961]. Es besagt, dass eine verlässliche Datenanalyse nur dann stattfinden kann, wenn die Anzahl der Stichproben exponentiell zur Dimension der erhobenen Daten gesteigert wird [Bellman, 1961]. Im Falle der Microarray Experimente müsste also die Anzahl der Stichproben exponentiell gesteigert werden, was mit hohen Kosten verbunden wäre. Um dies zu umgehen, ist es notwendig, die Dimension des betrachteten Datensatzes zu reduzieren, also die Menge der Gene einzuschränken [Hira/Gillies, 2015].

Indem irrelevante und/oder redundante Gene aus der betrachteten Menge entfernt werden, können nicht nur die Berechnungskomplexität gesenkt, besser generalisierbare Modelle erstellt, die Größe des benötigten Speicherplatzes reduziert und bei Klassifikationsproblemen [Runkler, 2015b] die Lern-Performance gesteigert werden [J. Tang et al., 2014], sondern auch das Verständnis der Prozesse verbessert werden, die durch die Daten beschrieben werden [Guyon et al., 2003].

Die zur Reduktion von Mengen von Genen verwendeten Verfahren können anhand ihrer Vorgehensweise in zwei Gruppen eingeteilt werden: Genextraktion und Genselektion [Hira/Gillies, 2015].

Da die in dieser Arbeit vorgestellte Methode beispielhaft anhand der Gene Ontology Wissensbasis entwickelt wird, werden im Folgenden Arbeiten anderer Autoren vorgestellt, die ebenfalls Verfahren vorgeschlagen haben, wie man eine Menge von Genen auf ihre wichtigsten Elemente reduzieren kann. Eine systematische Übersicht zur Einordnung der funktionserhaltenden, integrativen Genselektion wird in Abbildung 3.1 gegeben.

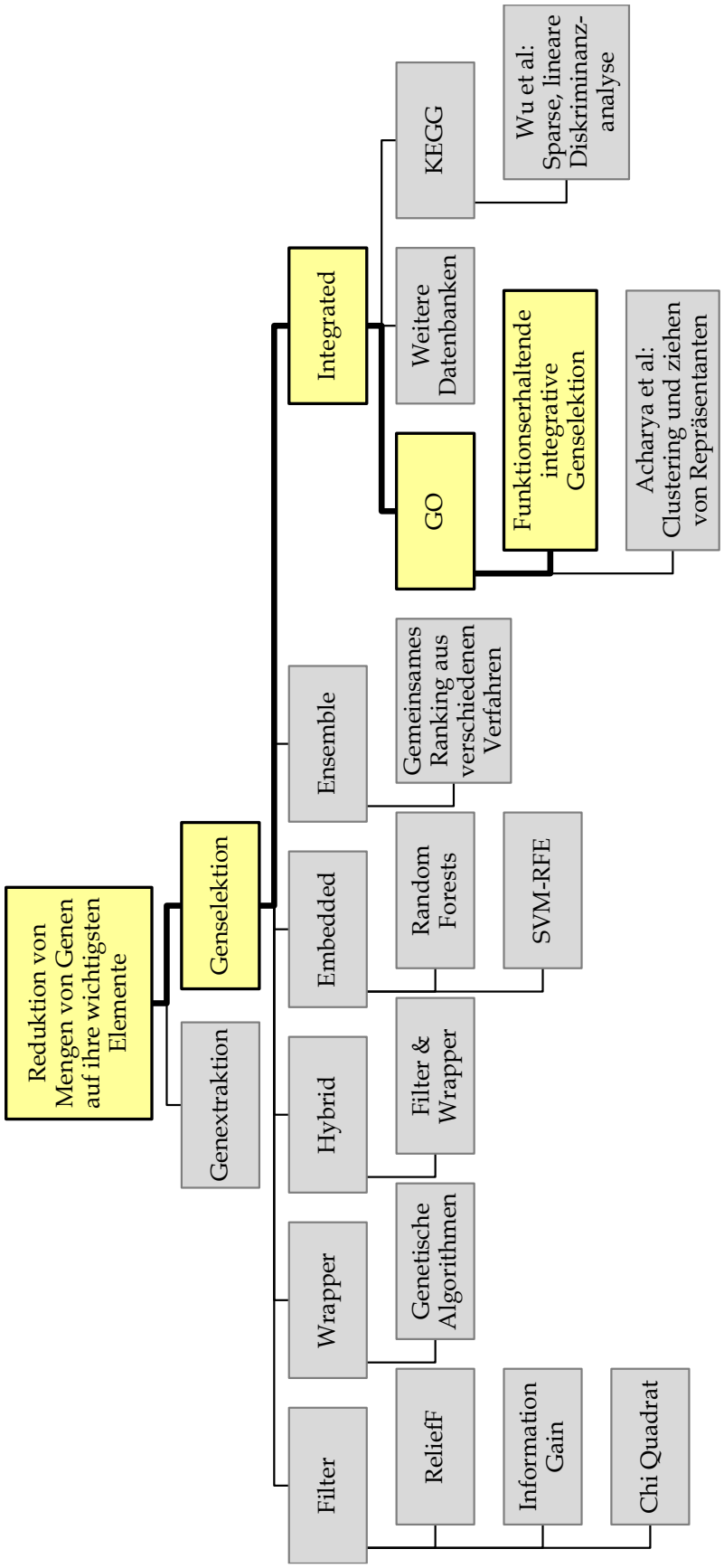


Abbildung 3.1: Systematische Übersicht von Verfahren zur Reduktion von Mengen von Genen auf ihre wichtigsten Elemente mit einer Auswahl von Beispielen.

3.1 Genextraktion

Bei der Genextraktion aus Expressionsdaten werden im Gegensatz zur Genselektion neue künstliche Gene als reduzierte Menge erzeugt, die durch Kombination oder Transformation der ursprünglichen Gene gebildet wurden [Jović et al., 2015; J. Tang et al., 2014]. Die Expressionsdaten der Gene werden dabei von einem höherdimensionalen Vektorraum in einen niederdimensionalen abgebildet [J. Tang et al., 2014]. Dadurch wird die weitere inhaltliche Analyse der Gene schwieriger, da den transformierten Genen keine direkte Bedeutung mehr zugemessen werden kann [J. Tang et al., 2014]. In diesem Sinne ist die Genselektion der Genextraktion im Bewahren besserer Verständlichkeit und Interpretierbarkeit überlegen [J. Tang et al., 2014]. Wird andererseits nur eine kleine Menge von Genen ausgewählt und waren die originalen Gene sehr divers, hat Genselektion den Nachteil, dass ein größerer Informationsverlust auftritt als bei Genextraktion, da in jedem Fall einige Gene beim Selektionsprozess ausgeschlossen [Khalid et al., 2014], bei der Genextraktion hingegen alle Gene zur Berechnung der neuen Gene einbezogen werden [J. Tang et al., 2014].

In dieser Arbeit sollen die wichtigsten Repräsentanten aus der Menge der Gene ausgewählt werden, damit anschließend z.B. weitere biologische Analysen mit diesen Genen vorgenommen werden können. Da aber künstliche Gene nicht für weitere Experimente verwendet werden können, wird die Genextraktion im Folgenden nicht näher betrachtet.

3.2 Genselektion

Bei der Genselektion werden Gene typischerweise mit einem beliebigen Maß bewertet, das es erlaubt, die wertvollsten Gene auszuwählen [Grasnick et al., 2019].

Die Verfahren der Genselektion können anhand der zugrundeliegenden Modelle in sechs Untergruppen unterteilt werden: Filter, Wrapper, Hybrid, Embedded, Ensemble und Integrated [Grasnick et al., 2019].

Filter
Filtermethoden selektieren Gene nur anhand der intrinsischen Eigenschaften der Daten [Saeys et al., 2007] mit Hilfe eines Suchverfahrens. Am häufigsten werden

Rangfolgen der Gene erstellt und die ersten k, also die k wichtigsten, ausgewählt [Saeys et al., 2007]. Dadurch sind Filter unabhängig vom Problem, für das die Genselektion durchgeführt wird [Jović et al., 2015]. Filtermethoden sind schneller und weniger rechenintensiv als Wrapper [Hira/Gillies, 2015]. Andererseits ist ihre Güte, die sie bei Klassifikationsproblemen erreichen, nicht so hoch wie die von Wrappern [Jović et al., 2015].

Wrapper

Wrapper wenden zuerst ein Suchverfahren an, um verschiedene Teilmengen der Gesamtmenge zu generieren [Saeys et al., 2007]. Auf alle gefundenen Teilmengen wird ein Lernalgorithmus angewendet, dessen Performance als Gütekriterium verwendet [Mahajan/Singh, 2016] und anhand dessen die optimale Teilmenge von Genen ausgewählt wird. Wrapper-Genselektions-Methoden sind dementsprechend von der Aufgabe abhängig, für die die Genselektion durchgeführt werden soll. Zum Beispiel bei Klassifikationsproblemen wertet der Wrapper Teilmengen der Gene basierend auf der Performance des Klassifikators aus, bei Clustering-Problemen anhand der Performance eines Cluster-Algorithmus [Jović et al., 2015]. Durch die Integration des Lernalgorithmus können Wrapper eine höhere Güte bei der Problemlösung erreichen als Filter, da auch Abhängigkeiten zwischen Genen berücksichtigt werden können [Mahajan/Singh, 2016]. Wrapper können aber gerade wegen der Integration des Lernalgorithmus nicht ohne weiteres auf andere Modelle übertragen werden [Mahajan/Singh, 2016] und sind zudem durch das Testen aller generierten Teilmengen auch sehr rechenintensiv [Saeys et al., 2007].

Hybrid

Hybride Genselektionsmethoden sind Kombinationen von Wrapper- und Filtermethoden, die versuchen die guten Eigenschaften beider Methoden auszunutzen [Jović et al., 2015]. Zuerst wird der Suchalgorithmus einer Filtermethode angewendet, um mögliche Teilmengen mit einer festgelegten Kardinalität zu finden [Huan/Lei, 2005]. Anschließend wird mit dem Lernalgorithmus einer Wrappermethode aus den vorgeschlagenen Teilmengen die optimale Teilmenge ausgewählt [Huan/Lei, 2005]. Hybride Genselektionsmethoden erreichen dadurch für gewöhnlich eine hohe Güte bei der Problemlösung und sind trotzdem effizient in der Berechnung [Jović et al., 2015]. Hybride Verfahren sind weniger anfällig für Overfitting als Wrapper, aber ebenso abhängig vom behandelten Problem [Mahajan/Singh, 2016]. Dementsprechend sind sie auch

nicht einfach auf andere Probleme anzupassen, beachten jedoch Abhängigkeiten zwischen Genen [Mahajan/Singh, 2016].

Embedded
Bei der Embedded Genselektion wird die optimale Teilmenge von Genen bereits während der Ausführung des Lernalgorithmus ausgewählt [Jović et al., 2015], [Hira/Gillies, 2015]. Dadurch sind embedded Methoden, ebenso wie Wrapper, stark abhängig vom Lernalgorithmus und können nicht direkt auf andere Genselektions-Probleme übertragen werden [Hira/Gillies, 2015]. Sie sind dabei allerdings weniger rechenintensiv als Wrapper [Saeys et al., 2007]. Genselektion ist bei diesen Methoden im Grunde nur ein Nebenprodukt der Lösung des Problems, für das die Genselektion genutzt wird [P. Yang et al., 2010].

Ensemble
Anstatt eine einzelne Genselektionsmethode zu verwenden, nutzen Ensemble-Methoden mehrere Genselektionsmethoden und resultieren in derjenigen Teilmenge, die in den meisten Methoden die besten Ergebnisse erzielt [Saeys et al., 2007]. Die Idee dahinter ist, dass es oft nicht nur eine einzige universell beste Genselektionsmethode gibt [Yang et al., 2005]. Ensemble-Methoden sind nach Konstruktion flexibel und robust [Saeys et al., 2007].

Integrated
Integrative Genselektion verwendet Domänenwissen aus externen Wissensbasen, wie der GO oder KEGG Pathways [M. Kanehisa/Goto, 2000], zur Auswahl der Gene, was eine bessere biologische Interpretierbarkeit ermöglicht [Grasnick et al., 2019].

In dieser Arbeit wird ein Ansatz einer integrativen Genselektion auf Grundlage der GO verfolgt, weshalb einige vorangegangene Arbeiten auf diesem Gebiet im Folgenden ausführlicher vorgestellt werden.

Eine Übersicht einer Auswahl von häufig verwendeten Filter-, Wrapper-, hybriden, embedded und ensemble Genselektions-Methoden, findet sich in Tabelle C. im Anhang.

3.3 Verwandte Verfahren zur Genselektion, die biologisches Wissen integrieren

Viele Verfahren nutzen die GO als Wissensbasis, so wird zum Beispiel in [Qi/Tang, 2007] ein Verfahren beschrieben, das die Gene Ontology Annotationen für die Genselektion zur Klassifikation von Microarray Daten ausnutzt. Jedem GO Term wird das Mittel der Expressionswerte derjenigen Gene zugewiesen, die zu diesem GO Term in der GO annotiert sind. Die GO Terme werden nach diesem Kriterium absteigend sortiert. Aus den zum ersten GO Term annotierten Genen wird dasjenige Gen ausgewählt, das den größten Expressionswert hat. Dieses Gen wird in die Menge der relevanten Gene aufgenommen und nicht mehr für die restliche Analyse verwendet. Mit den Expressionswerten der restlichen Gene werden die mittleren Expressionswerte der GO Terme neu berechnet und wiederum dasjenige Gen ausgewählt, das den größten Expressionswert hat und gleichzeitig an den GO Term mit dem aktuell höchsten mittleren Expressionswert annotiert ist. Daraus entsteht eine Rangfolge der Gene. Für jede Teilmenge der ersten k Gene, mit $k = 1, \dots, Anz\ aller\ Gene$, wird die Güte der Klassifikation gemessen und dann die kleinste Teilmenge mit der besten Güte als endgültige Teilmenge ausgewählt.

Ein anderer Ansatz, der sogar als Java Plattform „SoFoCles“ implementiert wurde, versucht ebenfalls ein Klassifikationsproblem mit Hilfe der GO zu lösen [Papachristoudis et al., 2010]. Damit soll eine gewöhnliche Genselektion um die biologisch relevantesten Gene erweitert werden, sofern diese sonst nicht ausgewählt worden wären [Papachristoudis et al., 2010]. Als gewöhnliche Genselektion sind die Filtermethoden basierend auf dem Chi-Quadrat-Algorithmus, Information Gain, und ReliefF implementiert. Die gefundene Menge von Genen wird anschließend um semantisch ähnliche Gene ergänzt. Semantische Ähnlichkeit wird dabei über verschiedene Gewichtungen des Informationsgehalts der einzelnen GO Terme und/oder die DAG Struktur der GO definiert. Die mit dieser Maßzahl als sehr ähnlich zu den selektierten Genen befundenen Gene, die zuvor nicht ausgewählt wurden, werden anschließend ergänzend in die relevante Teilmenge aufgenommen. Wie viele Gene ausgewählt werden sollen, muss vom User vorgegeben werden.

Mitra und Ghosh haben einen Ansatz zur Lösung eines Clusterproblems unter Einbeziehung der GO entwickelt [Mitra/Ghosh, 2012]. Die Gene aus einem Microarray-Experiment werden zunächst anhand ihrer Expressionswerte geclustert, anschließend wird für die Gene in jedem einzelnen Cluster die biologische Relevanz bestimmt und das relevanteste Gen als Repräsentant auswählt. Als Clusteralgorithmus wird CLARANS [Ng/Han, 2002] und in einem Folgepaper, [Ghosh et al., 2014], Fuzzy CLARANS [Ghosh/Mitra, 2013] verwendet. Die Autoren merken aber an, dass theoretisch ein beliebiger Clusteralgorithmus verwendet werden könne. Die biologische Relevanz der Gene wird mit Hilfe einer ORA auf Basis der GO für jedes Cluster einzeln festgestellt. Cluster, zu denen mindestens ein signifikanter GO-Term gefunden wurde, werden als biologisch relevant angesehen. Die reduzierte Menge der Gene wird durch Auswahl eines Gens aus jedem Cluster gebildet. Es wird jeweils dasjenige Gen ausgewählt, dessen Expressionswert am ähnlichsten zum Mittelwert aller Expressionswerte der Gene in diesem Cluster ist.

Wu et al. [Wu et al., 2009] schlagen einen Algorithmus zur Genselektion basierend auf der KEGG Pathway Datenbank [M. Kanehisa/Goto, 2000] vor. Für Expressionsdaten aus einem Microarray Experiment werden die beteiligten Gene zunächst in Gruppen eingeteilt. Zusätzlich werden unwichtige Gene mithilfe einer Gewichtung entfernt, wodurch sich eine Genselektion ergibt. Die Gruppen bestimmen sich dabei durch die Zugehörigkeit der Gene zu je einem KEGG Pathway. Für jede Gruppe wird ein gemeinsamer Expressionswert als Linearkombination der Expressionswerte aller in der Gruppe zusammengefassten Gene berechnet. Die optimalen Gewichte der Linearkombination werden mit Hilfe von linearer Diskriminanzanalyse [Backhaus et al., 2016] geschätzt, die um eine weitere Bedingung ergänzt wurde, sodass nur wenige der Gewichte ungleich Null geschätzt werden. Diese Bedingung beruht auf der Annahme, dass viele der Gene in einem KEGG Pathway keinen Effekt haben, also unwichtig sind. Durch die Gewichtung mit Null wird automatisch eine Genselektion durchgeführt, die die wichtigen Gene beibehält, so dass nur die Gene mit Gewichten ungleich Null als reduzierte Menge übrig bleiben.

Ein weiteres Verfahren, das zur Klassifikation von Microarray Daten entwickelt wurde, wird in [Fang et al., 2014] beschrieben. Es integriert sowohl GO als auch KEGG Pathways. Es wird zunächst wie bei [Papachristoudis et al., 2010] eine

Genselektion anhand der Expressionswerte mit Hilfe der Filtermethode berechnet, die Information Gain verwendet. Gene, deren Information Gain gleich Null ist, werden für die weitere Analyse nicht berücksichtigt. Anschließend werden die Annotationen der übrigen Gene mit Hilfe der Wissensbasen bestimmt. Durch eine Assoziationsanalyse [Bankhofer/Vogel, 2008], in die sowohl Annotationen als auch Expressionswerte einfließen, werden verschiedene (nicht unbedingt disjunkte) Gruppen in den Genen gefunden und nach ihrer Interessantheit geordnet, wobei die Interessantheit der Gruppen als durchschnittlicher Information Gain der beteiligten Gene gegeben ist. Aus jeder Gruppe wird jeweils das Gen, das den höchsten Information Gain hat, als Repräsentant gewählt. Dadurch entsteht ein Ranking der Gene. Für jede Teilmenge der ersten k Gene, $k \in \{1, \dots, Anz\ aller\ Gene\}$, wird wiederum die Güte der Klassifikation gemessen und dann die kleinste Teilmenge mit der besten Güte als endgültige Teilmenge ausgewählt.

Alle zuvor vorgestellten Arbeiten verwenden Expressionsdaten der Gene aus Laborexperimenten zur quantitativen Bewertung der Gene bei der Genselektion. Liegen aber für einen Gensatz keine Expressionsdaten vor, da dessen gemeinsame Thematik z.B. nur durch Literaturrecherche eines Experten belegt wird, können diese Verfahren nicht angewendet werden. Der in dieser Arbeit präsentierte Ansatz verzichtet daher auf Expressionswerte. Der folgende, für diese Arbeit relevante Ansatz von Acharya et al. beschreibt ebenfalls ein Vorgehen, für das zur Auswahl der Gene keine Expressionsdaten erforderlich sind. Eine ausführlichere Diskussion der Methode findet sich in Kapitel 5.2.

Zunächst wird in [Acharya et al., 2017] für eine Menge von Genen eine ORA durchgeführt, die signifikante GO Terme in der GO liefert. Zu jedem signifikanten GO Term wird ein „struktureller Informationsgehalt“ berechnet, der sich aus der Position des GO Terms und der Anzahl seiner Nachfahren im DAG errechnet, der aus der ORA resultiert. Es wird eine Matrix erstellt, die die Gene Ontology Annotationen beschreibt. Ihre Einträge sind Null, sofern ein Gen in der jeweiligen Zeile nicht zu dem GO Term der jeweiligen Spalte annotiert ist. Ist das Gen zum GO Term annotiert, enthält die Matrix den strukturellen Informationsgehalt des entsprechenden GO Terms. Die Gene aus dieser Matrix werden anschließend unter Verwendung des Partitioning around Medoids Algorithmus

[Kaufman/Rousseeuw, 2009] in k_i, $i \in \{1, \dots, \sqrt{Anzahl\ Gene}\}$ verschiedene Cluster eingeteilt, sodass funktionell ähnliche Gene, die also zu den gleichen GO Termen annotiert sind, in Clustern zusammengefasst werden. Mit Hilfe eines Silhouette-Plots [Rousseeuw, 1987] wird die optimale Anzahl an Clustern $k^* \in \{k_1, \dots, k_{\sqrt{Anzahl\ Gene}}\}$ bestimmt. Die reduzierte Menge der Gene wird schließlich durch Auswahl von Repräsentanten der einzelnen Cluster gebildet. Die Repräsentanten sind diejenigen Gene, die von allen Genen des gleichen Clusters am wenigsten weit entfernt, also in der Mitte der Cluster liegen.

4 Funktionserhaltende, integrative Genselektion

In diesem Kapitel wird ein eigener datenbionischer Ansatz, die funktionserhaltende, integrative Genselektion vorgestellt. Sie dient der Gewinnung einer kleinen, verständlichen Menge von relevanten Objekten aus einer für Menschen unüberschaubar großen Menge von Objekten. Dabei wird der datenbionischen Idee entsprochen, eine möglichst optimale Teilmenge zu finden, indem a priori Wissen in die Methode integriert wird. Es wird also im Prinzip ein Information Retrieval System erstellt, das das Wissen aus einer zugrundeliegenden Wissensbasis ausnutzt. Das vorgeschlagene Verfahren wird anhand der gut gepflegten Gene Ontology Wissensbasis entwickelt, es ist aber durchaus denkbar, es auf andere Wissensbasen, die in einer entsprechenden Form vorliegen, zu übertragen.

Im Zusammenhang mit dem Anwendungsbeispiel der Gene Ontology Wissensbasis eignet sich das Verfahren zur Genselektion. Die Bewertung der Gene basiert nicht auf numerischen Messwerten aus Experimenten, daher ist das Verfahren auch für Datensätze verwendbar, die nur durch Fachexpertise oder Literaturrecherche gewonnen wurden.

Es wird eine möglichst kleine Teilmenge der Gene gesucht, die die biologischen Prozesse der Menge aller Gene möglichst vollständig reproduziert. Dazu wird für jede gefundene Teilmenge eine ORA durchgeführt und der resultierende DAG mit dem DAG, der aus einer ORA mit allen Genen stammt, verglichen. Die Teilmengen werden dabei gezielt ausgesucht. Es wird mit Hilfe der GO Annotationen ein Gen-Score für jedes Gen berechnet, der die Wichtigkeit jedes Gens bestimmt. Die ersten k Gene bilden für jedes k je eine Teilmenge. Welches k optimal ist, wird mit Hilfe einer Idee aus der ABC-Analyse [Ultsch/Lötsch, 2015] und Precision- und Recall-Kurve bestimmt.

4.1 Verwendete Datensätze

Zur praktischen Anwendung und Überprüfung der Methode wurden verschiedene Datensätze herangezogen. Hauptsächlich wurde die Idee anhand einer Menge von Genen, die mit Schmerz assoziiert sind, entwickelt. Es konnte aber gezeigt werden, dass sich die Methode auch auf andere Gensätze übertragen lässt.

Der Satz der Gene, die mit Schmerz zusammenhängen, im Folgenden Schmerzgene genannt, stammt größtenteils aus der Pain Genes Database (http://www.jbldesign.com/jmogil/enter.html [LaCroix-Fralish et al., 2007]). Diese Datenbank beinhaltet hauptsächlich Gene, die in mindestens drei unabhängigen Studien zur Modulation von Schmerz an transgenen Mäusen gefunden wurden [Lippmann et al., 2018]. Die Studien wurden dabei durch eine Suche in der Literaturdatenbank MEDLINE/PubMed (https://www.ncbi.nlm.nih.gov/pubmed/) nach „'Pain Measurement' AND 'Mice, Knockout'" und nach „(mutant OR knockout OR knock-out OR deficient OR transgenic) AND (mice OR mouse) AND (pain OR *nocicepti* OR hyperalgesi* OR allodyni*) NOT review" identifiziert [LaCroix-Fralish et al., 2007]. Zusätzlich zu diesen Genen enthält der Datensatz Gene, die ursächlich mit hereditären Krankheiten, die mit extremen Schmerzen einhergehen, in Verbindung gebracht werden (zusammengefasst z.B. in [Lötsch et al., 2017]), und Gene, die für Arzneimittelzielstrukturen [Müller-Esterl, 2017] von zugelassenen analgetischen Medikamenten oder neuen in klinischen Phasen der Entwicklung befindlichen Analgetika kodieren [Lötsch/Geisslinger, 2011]. Dies ergab eine Liste von insgesamt $n_G = 540$ Schmerzgenen [Lippmann et al., 2018].

Die Validierung der hier vorgeschlagenen Methode wurde an drei weiteren Datensätzen durchgeführt. Dies beinhaltete erstens eine Menge von $n_G = 2\,954$ Genen, die mit micro RNAs (miRNAs [Graw, 2015b]) interagieren [Ultsch/Lötsch, 2014b], zweitens eine Menge von $n_G = 719$ Genen, die ursächlich mit verschiedenen Krebsarten in Verbindung gebracht werden [Futreal et al., 2004], im Folgenden „Krebsgene" genannt, und drittens eine Menge von $n_G = 387$ Genen, die mit substanzgebundener Abhängigkeit assoziiert wurden [Lötsch/Ultsch, 2016], im Folgenden als „Suchtgene" bezeichnet.

4.2 Datenanalyse

Gen-Score

Um die Wichtigkeit der Gene in der untersuchten Menge numerisch bewerten und entsprechend eine Rangfolge erstellen zu können, wird für jedes Gen ein Gen-Score berechnet. Dieser setzt sich aus verschiedenen Maßzahlen zusammen. Der Idee von [Ultsch/Lötsch, 2014a] folgend werden deshalb während der ORA neben den p-Werten der signifikanten GO Terme weitere Maßzahlen berechnet

[Lippmann et al., 2018], die die Relevanz der GO Terme bewerten [Ultsch/Lötsch, 2014a]. Die für diese Arbeit interessierenden Maßzahlen sind Certainty, Information Value und Remarkableness [Ultsch/Lötsch, 2014a].

Die Maßzahl **Certainty** drückt aus, wie sicher es ist anzunehmen, dass ein GO Term T_i die gegebene Menge der Gene beschreibt [Ultsch/Lötsch, 2014a]. Sie ist definiert als

$$\begin{aligned}\text{Cert}(T_i) &= p(\text{es ex. ein Term mit kleinerem p-Wert}) \\ &= \frac{\left|\{T_k : \text{p-Wert}(T_k) < \text{p-Wert}(T_i)\}\right|}{n_T},\end{aligned}$$

wobei n_T die Anzahl signifikanter GO Terme ist und $\text{Cert}(T_i) \in [0;1]$, $i \in \{1, \dots, n_T\}$ [Ultsch/Lötsch, 2014a].

Der **Information Value** eines GO Terms T_i bestimmt, wie informativ dieser GO Term ist [Ultsch/Lötsch, 2014a]. Er wird mithilfe der einzelnen Summanden der Shannon-Entropie (siehe Definition 2.4.5) definiert [Ultsch/Lötsch, 2014a]. Wählt man die Konstanten c und a in (Definition 2.4.4) gleich der eulerschen Zahl e ergibt sich der Information Value als

$$\text{Info}(T_i) = -e \cdot p_i \cdot \ln(p_i), i \in \{1, \dots, n_T\}$$

mit $p_i = n_G(T_i)/n_G$, wobei $n_G(T_i)$ die Anzahl der Gene aus der gegebenen Menge, die zum GO Term T_i annotiert sind, und n_G die Anzahl der Elemente der gegebenen Menge der Gene ist [Ultsch/Lötsch, 2014a]. Mit Hilfe der eulerschen Zahl e und des natürlichen Logarithmus wird der Information Value eines GO Terms auf das Intervall $[0;1]$ normiert [Ultsch/Lötsch, 2014a].

Die Maßzahl **Remarkableness** fasst Certainty und Information Value in einer Maßzahl, die die Wichtigkeit jedes GO Terms beschreibt, zusammen. Sie berechnet sich als Produkt von Certainty und Information Value [Ultsch/Lötsch, 2014a] in Prozent. Für einen GO Term T_i gilt somit

$$\text{Rem}(T_i) = \text{Cert}(T_i) \cdot \text{Info}(T_i) \cdot 100$$

[Ultsch/Lötsch, 2014a].

Der **Gen-Score** eines Gens G_i berechnet sich als Summe der Remarkableness jedes GO Terms, zu dem das Gen annotiert ist. Sei $G = \{G_1, \dots, G_{n_G}\}$ eine Menge

von Genen und $T = \{T_1, \dots, T_{n_T}\}$ die Menge von signifikanten GO Termen, die aus einer ORA von G resultieren, mit Remarkableness-Werten $R = \{\mathrm{Rem}(T_1), \dots, \mathrm{Rem}(T_{n_T})\}$. Sei weiterhin $M \in \mathbb{R}^{n_G} \times \mathbb{R}^{n_T}$ die Matrix, die die Annotationen der Gene $G_1, \dots, G_n$ zu den GO Termen $T_1, \dots, T_{n_T}$ repräsentiert. Das heißt es gilt $M[i,j] = 1$, $i \in \{1, \dots, n_G\}$, $j \in \{1, \dots, n_T\}$, falls Gen G_i zu GO Term T_j annotiert ist, sonst ist $M[i,j] = 0$. Dann ist der Gen-Score definiert als $S := M \cdot R$, d.h. Gen G_i hat den Score

$$S_i = M[i,\cdot] \cdot R = \sum_{k=1}^{n_T} M[i,k] \cdot \mathrm{Rem}(T_k). \tag{4.1}$$

Der Gen-Score erlaubt also eine Gewichtung der Gene abhängig von der Wichtigkeit der Genfunktionen, an denen die Gene beteiligt sind.

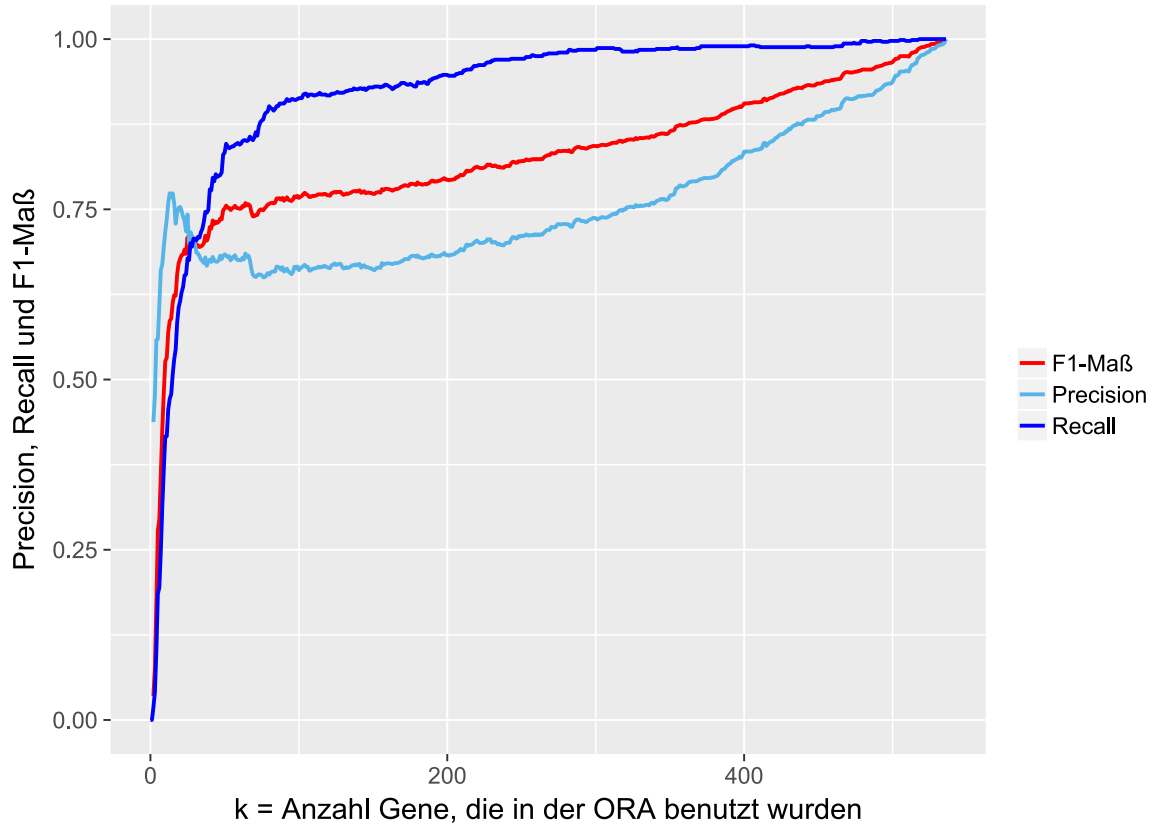
Auswahl der optimalen Teilmenge

Anhand des Gen-Scores werden die Gene einer Menge $G = \{G_1, \dots, G_{n_G}\}$ in absteigender Reihenfolge sortiert. Teilmengen, die auf ihre Optimalität hin überprüft werden sollen, werden aus den ersten k, $k \in \{1, \dots, n_G\}$ Genen gebildet, so dass sich eine Folge von ineinander liegenden Teilmengen ergibt. Für jede der n_G Teilmengen wird mit Hilfe der ORA die zugehörige Menge signifikanter GO Terme berechnet und Precision und Recall bezüglich der signifikanten biologischen Prozesse bestimmt, die aus der ORA mit der gesamten Menge der Gene resultieren. Zudem wird auch das F_1-Maß berechnet, das eine Kombination von Precision und Recall darstellt. Wird die Anzahl k der pro Teilmenge verwendeten Gene gegen F_1-Maß, Precision und Recall aufgetragen, wird die Qualität der Reproduktion der GO Terme des ursprünglichen DAGs in Abhängigkeit zur Anzahl der dazu verwendeten Gene dargestellt (siehe Abbildung 4.1). Es ergeben sich empirische Kurven für F_1-Maß, Precision und Recall.

Die optimale Teilmenge der Gene wird als Minimum von zwei verschiedenen Verfahren bestimmt. Einerseits wird der Schnittpunkt von Precision- und Recall-Kurve bestimmt, andererseits ein optimaler Punkt auf der F_1-Kurve.

Der Schnittpunkt der beiden Kurven (siehe Abbildung 4.1) liefert den optimalen Wert k'. Die Teilmenge der Größe k' erhält so viel wie möglich des funktionell

Abbildung 4.1: Empirische Precision-, Recall- und F_1-Maß-Kurven der signifikanten GO-Terme für den Datensatz der Schmerzgene. Die Kurven zeigen Precision, Recall und F_1-Maß für die signifikanten GO Terme, die aus ORAs der Teilmengen mit $k \in \{1, \dots, n_G\}$ der bestbewerteten Schmerzgene resultieren, bezüglich der GO-Terme, die aus der ORA mit der Gesamtmenge der Schmerzgene resultierten.

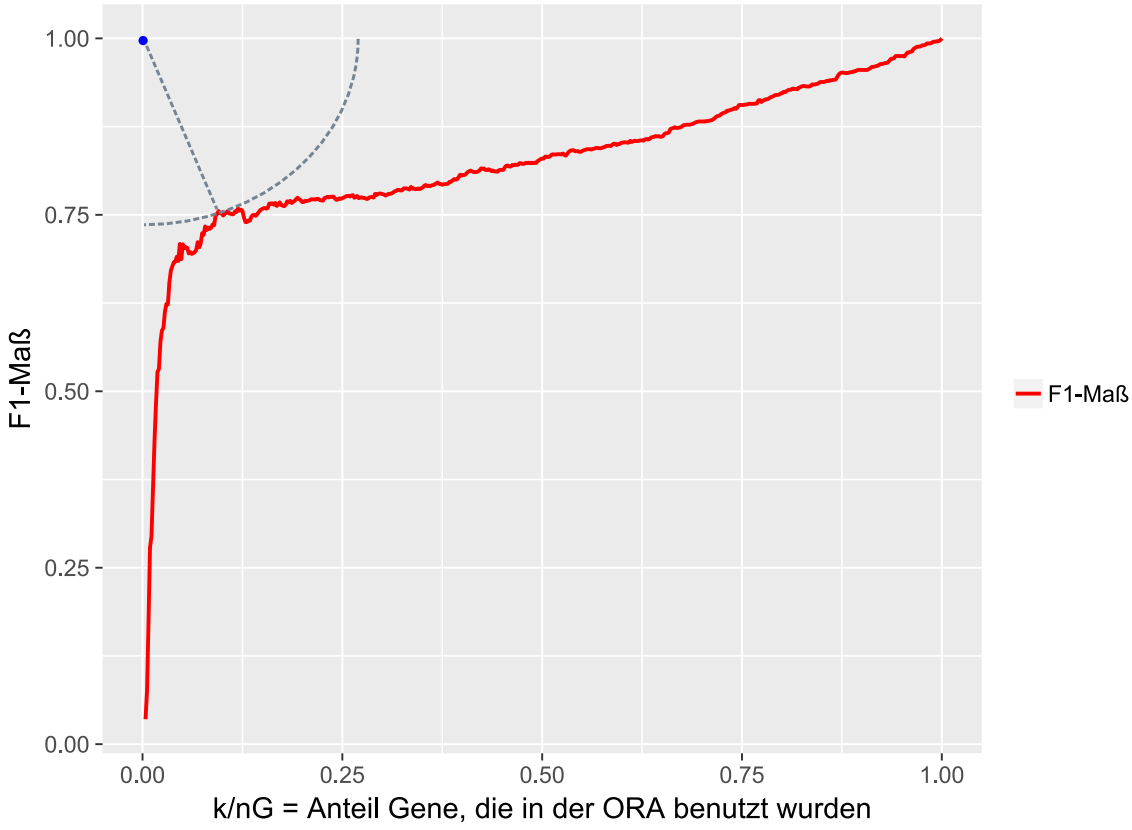


genomischen Bildes der gesamten Menge der Gene, während sie den kleinsten Trade-off zwischen für den DAG der Teilmenge relevanten GO Termen und der Reproduktion von GO Termen des ursprünglichen DAGs hat. Der Schnittpunkt ist auch gerade derjenige Punkt, in dem der DAG der Teilmenge die gleiche Anzahl GO Terme beinhaltet wie der DAG aller Gene. Dies folgt unmittelbar aus den Definitionen von Precision und Recall. Bezogen auf die Anzahl der GO Terme aus der ORA mit der gesamten Menge der Gene n_T beziehungsweise der Teilmenge $n_{T'}$, kann man die Formeln (2.1) und (2.2) aus Definition 2.4.2 umschreiben als $\text{Precision} = \frac{tp}{tp+fp} = \frac{n_{T' \cap T}}{n_{T' \cap T} + n_{T' \setminus T}} = \frac{n_{T' \cap T}}{n_{T'}}$ und $\text{Recall} = \frac{tp}{tp+fn} = \frac{n_{T' \cap T}}{n_{T' \cap T} + n_{T \setminus T'}} = \frac{n_{T' \cap T}}{n_T}$, wobei $n_{T' \cap T}$ die Anzahl der GO Terme ist, die in beiden DAGs enthalten sind, $n_{T' \setminus T}$ die Anzahl der GO Terme, die ausschließlich im DAG der Teilmenge, und $n_{T \setminus T'}$ die Anzahl der GO Terme, die ausschließlich im DAG der

gesamten Menge vorkommen. Aus den beiden Formeln ergibt sich, dass Precision gleich Recall ist, falls $n_T, = n_T$ gilt, d.h. die beiden DAGs die gleiche Größe haben.

In [Lippmann et al., 2019] wurde der Schnittpunkt der Precision- und Recall-Kurve als optimaler, leicht zu berechnender Punkt gewählt, durch den die Größe der Teilmenge k' festgelegt wurde. Dieses Verfahren wurde mit Hilfe der Idee der berechneten ABC-Analyse [Ultsch/Lötsch, 2015] erweitert, um die Methode

Abbildung 4.2: Empirische F_1-Maß-Kurve der signifikanten GO-Terme für den Datensatz der Schmerzgene (Ordinate) und Anteil der verwendeten $\frac{k}{n_G}, k \in \{1, \dots, n_G\}$ bestbewerteten Gene (Abszisse). In grau markiert: kürzester euklidischer Abstand der F_1-Maß-Kurve zum Punkt $(0, 1)$.



weiter zu verbessern. Gemäß der berechneten ABC-Analyse [Ultsch/Lötsch, 2015] wird der Punkt $(k = 0, F_1 = 1)$ als ökonomisches Optimum angesehen, da in diesem Punkt minimal wenige Gene verwendet werden, um die maximale Güte der Reproduktion der GO Terme des ursprünglichen DAGs gemessen an F_1 zu erhalten. Daher kann der optimale Punkt auf der F_1-Kurve $(k'', F_1(k''))$

analog zur Idee der berechneten ABC-Analyse [Ultsch/Lötsch, 2015] als derjenige Punkt, der den geringsten euklidischen Abstand zu $(k = 0, F_1 = 1)$ hat, bestimmt werden. In diesem Punkt wird mit minimalem Aufwand, also mit möglichst wenig Genen k'', ein maximaler Ertrag, also ein möglichst hoher Wert für F_1, erzielt. Dazu wird jeweils die Anzahl der Gene in der Teilmenge k mit der Anzahl der Gene in der Gesamtmenge n_G normiert und der Abstand des Punktes dieses Anteils und dem zugehörigen F_1-Maß zum Punkt (0,1) bestimmt. Die optimale Anzahl an Genen k'' analog zur berechneten ABC-Analyse wird entsprechend definiert als $k'' \coloneqq \min(k) : dist\left(\left(\frac{k}{n_G}, F_1(k)\right), (0,1)\right) = \min_k \sqrt{\left(0 - \frac{k}{n_G}\right)^2 + \left(1 - F_1(k)\right)^2}$, $k \in \{1, \dots, n_G\}$ (siehe Abbildung 4.2).

Um möglichst wenig Gene aus der gesamten Menge der Gene zu verwenden und der Idee der „Kunst des Übersehens, die zu einem neuen Sehen des sonst Unsichtbaren führt“[Schwemmer, 2008/2009, p. 14] zu folgen, wird die Größe der optimalen Teilmenge als Minimum beider Verfahren, $k^* = \min(k', k'')$, bestimmt.

Robustheit der Methode

Um zu testen, ob die funktionelle, integrative Genselektion robust gegen kleine Veränderungen der Daten ist, wurde ein Cross-Validierungs-Experiment durchgeführt. Dazu wurden von den $n_G = 540$ Schmerzgenen je 10 Stichproben von $90\%, 80\%, 70\%$ und 60% der Gene gezogen, was zu Teilmengen der Schmerzgene von 482, 429, 375 und 322 Genen führte. In diesen Teilmengen wurden wie zuvor die wichtigsten Gene gesucht und die Ergebnisse mit denen der gesamten Menge der Schmerzgene verglichen.

4.3 Resultate

Die ORA der Menge der $n_G = 540$ Schmerzgene resultierte in einem DAG von $n_T = 761$ signifikanten GO Termen im DAG der biologischen Prozesse der GO. Für die ORA wurden eine p-Wert-Schranke von 1% und nur manuell kuratierte Annotationen verwendet. Die Anzahl, die minimal an einen signifikanten GO

Term annotiert sein sollte, wurde auf 2 gesetzt und zur Korrektur für multiples Testen wurde die Bonferroni-Korrektur verwendet. Da für 4 der 540 Gene keine manuell kuratierten Annotationen bekannt waren, wurde die Analyse mit nur 536 Genen durchgeführt. Der komplexe, vollständige DAG der $n_T = 761$ GO Terme ist online in [Lippmann et al., 2019] veröffentlicht. Ein beispielhafter Ausschnitt des gesamten DAGs wird in Abbildung B.1 im Anhang gezeigt.

Für jedes der $n_G = 540$ Schmerzgene wurde der Gen-Score nach Formel (*4.1*) aus den Ergebnissen der ORA berechnet. Anschließend wurden Teilmengen aus den ersten $k \in \{1, \dots, 540\}$ Genen mit den höchsten Gen-Scores gebildet und für jede Teilmenge eine ORA mit den gleichen Parametern wie zuvor durchgeführt.

Die resultierenden Mengen der GO Terme wurden mit Precision und Recall mit den GO Termen aus dem DAG, der aus der ORA aller Schmerzgene resultierte, verglichen (siehe Abbildung 4.1) und der Schnittpunkt[1] der beiden sich ergebenden Kurven berechnet. Zudem wurde auch das F_1-Maß für jede der k Teilmengen berechnet und der Abstand der F_1-Kurve zum Punkt $(0, 1)$ bestimmt. Der Schnittpunkt von Precision- und Recall-Kurve lag bei $k' = 29$ Genen. Der geringste Abstand zur F_1-Kurve lag bei $k'' = 52$ Genen. Es wurde eine Teilmenge von $k^* = \min(k', k'') = 29$ Schmerzgenen als optimale Teilmenge gefunden, mit der der DAG aller Schmerzgene am besten reproduziert werden konnte (siehe Tabelle C. im Anhang). Die Werte von F_1-Maß, Precision und Recall in den durch die beiden Verfahren als optimal bestimmten Punkten k' und k'' können Tabellen 4.1 entnommen werden.

Eine ORA der $k^* = 29$ besten Gene lieferte (mit den gleichen Parametern wie zuvor) $n_T = 767$ signifikante GO Terme. Precision und Recall im Vergleich zum DAG aller Schmerzgene lagen bei 70.14% bzw. 70.70%. Es konnte also mit nur 5.41% der Schmerzgene mehr als zwei Drittel des funktionell genomischen Bildes des Schmerzes reproduziert werden.

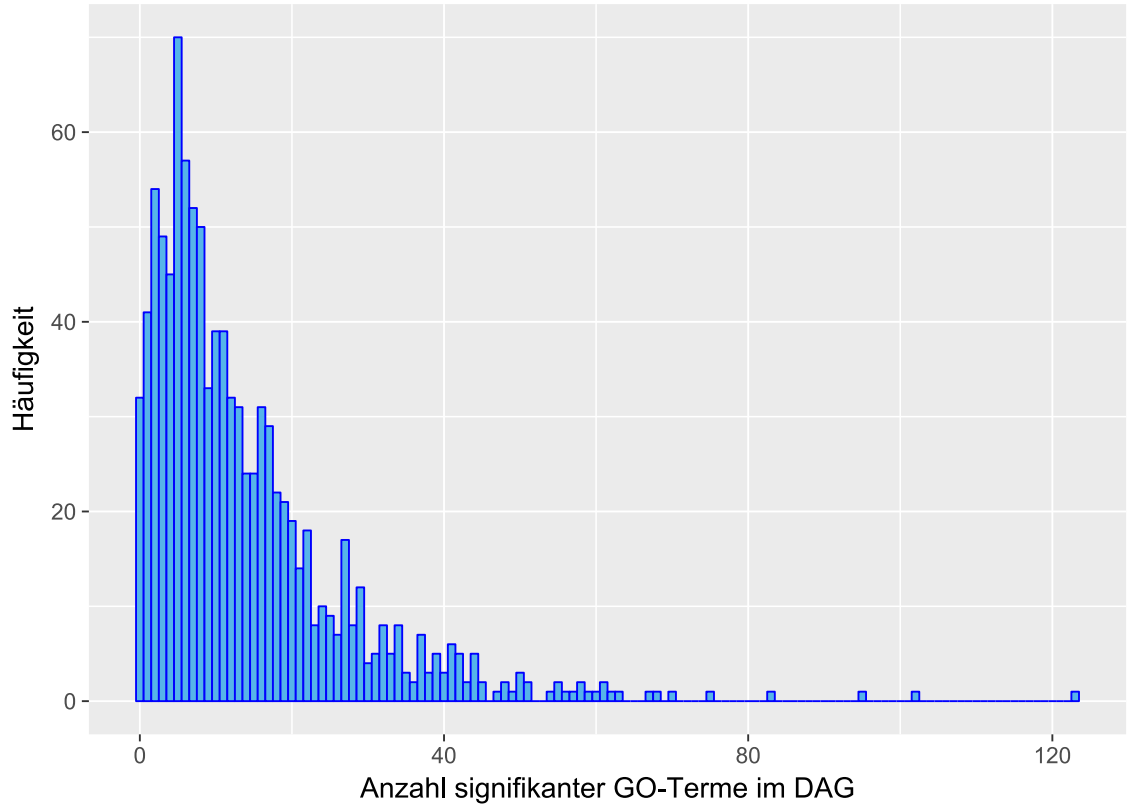
Validierung: Zufallsexperiment

Um festzustellen, ob die Ergebnisse tatsächlich durch die Auswahl der $k^* = 29$ Genen mit hohem Gen-Score erklärt werden können oder ob jede gleich große

[1] Es wurde der betragsmäßig kleinste Abstand zwischen den Werten von Precision und Recall für jedes k berechnet und nicht der tatsächliche theoretische Schnittpunkt der beiden Kurven, da nur ein ganzzahliger Wert für die Anzahl von Genen in der optimalen Teilmenge sinnvoll ist.

Teilmenge von beliebigen Schmerzgenen einen DAG liefert, der den ursprünglichen DAG aller Gene genauso gut reproduziert, wurden 1 000 Teilmengen aus $k^* = 29$ zufällig gewählten Schmerzgenen gebildet und wie zuvor analysiert. In den 1 000 Versuchen wurden eine durchschnittliche Precision von 93.03% und ein durchschnittlicher Recall von 1.77% (maximaler Recall in den 1 000 Versuchen: 13.40%) beobachtet. Der Wert von Precision nahe 100% ergibt sich daher, dass es nur wenige GO Terme gab, zu denen signifikant mehr oder signifikant weniger der $k^* = 29$ Gene annotiert waren, als rein zufällig zu erwarten wäre.

Abbildung 4.3: Verteilung der Anzahl von signifikanten GO-Termen, die aus 1 000 ORAs mit zufällig gezogenen Teilmengen von $k^* = 29$ der 540 Schmerzgene resultieren; Maximum bei nur etwa 5 GO-Termen in der Ontologie der biologischen Prozesse der GO.



Das bedeutet, dass die Anzahl der falsch positiven GO Terme nahezu Null und somit Nenner und Zähler in Formel (2.1) in Definition 2.4.2 nahezu identisch waren. Recall hingegen war wegen der wenigen signifikanten GO Terme sehr niedrig, da viele GO Terme aus dem ursprünglichen DAG nicht mit der zufälligen Teilmenge von Genen reproduziert werden konnten. Die Verteilung der Anzahl der GO Terme aus den 1 000 wiederholten ORAs für die zufälligen Teilmengen

von $k^* = 29$ der Schmerzgene zeigt ein Maximum bei nur 5 GO Termen (siehe Abbildung 4.3), wobei keine der ORAs auf $n_T > 760$ GO Terme kam, die mit den $k^* = 29$ bestbewerteten Schmerzgenen erzielt wurden.

Ergebnisse der Robustheits-Analyse

Von den $n_G = 540$ Schmerzgenen wurden je 10 Stichproben von 90%, 80%, 70% und 60% der Gene gezogen, was zu Teilmengen der Schmerzgene von 482, 429, 375 und 322 Genen führte. In diesen Teilmengen wurden wie zuvor die wichtigsten Gene gesucht und die Ergebnisse mit denen der gesamten Menge der Schmerzgene verglichen. Es wurden in den 10 Versuchen durchschnittlich je 27.0, 24.5, 20.1 und 19.5 Gene in den Teilmengen als wichtigste Gene identifiziert. Die zuvor gefundenen 29 wichtigsten Schmerzgene konnten mit durchschnittlich 95.53%, 97.00%, 91.78% und 88.27% Precision und durchschnittlich 93.80%, 88.67%, 90.03% und 77.94% Recall wiedergefunden werden.

Validierung: Weitere Datensätze

In der Menge der $n_G = 2\,954$ Gene, die mit miRNAs interagieren (siehe Kapitel 4.1), wurden für 2 787 Gene Annotationen zu GO Termen gefunden, mit denen die nachfolgende Analyse erfolgte. Eine ORA, die mit den gleichen Parametern durchgeführt wurde wie für die Schmerzgene, lieferte $n_T = 840$ signifikante GO Terme in der Ontologie der biologischen Prozesse. Die Menge der Gene, die mit miRNAs interagieren, konnte auf eine Teilmenge von nur $k^* = 28$ Gene reduziert werden. Die ORA der $k^* = 28$ bestbewerteten Gene lieferte einen DAG von $n_T = 838$ biologischen Prozessen, wobei Werte von 70.53% und 70.36% für Precision bzw. Recall erreicht wurden (siehe Abbildung B.2). Für eine Teilmenge von $k = 28$ zufällig ausgewählten, mit miRNAs interagierenden Genen hingegen wurde in 1 000 Versuchen eine durchschnittliche Precision von 62.89% und ein durchschnittlicher Recall von 0.34% (maximaler Recall in 1 000 Versuchen: 4.17%) beobachtet. Die Verteilung der Anzahl der GO Terme aus den 1 000 ORAs mit $k = 28$ Genen hatte ein Maximum bei nur einem GO Term (siehe Abbildung B.3). In diesem Experiment wurden aber überhaupt nur in 204 der 1 000 Versuche signifikante GO Terme durch die ORA gefunden, während 796 Versuche beendet wurden, ohne einen signifikanten biologischen Prozess unter den gegebenen Parametern zu finden.

In der Menge der $n_G = 719$ Krebsgene (vgl Kapitel 4.1), wurden für 692 Gene Annotationen zu GO Termen gefunden, mit denen die nachfolgende Analyse erfolgte. Eine ORA, die mit den gleichen Parametern durchgeführt wurde wie für die Schmerzgene, lieferte $n_T = 687$ signifikante GO Terme in der Ontologie der biologischen Prozesse. Die Menge der Krebsgene konnte auf eine Teilmenge von $k^* = 39$ Gene reduziert werden. Die ORA der $k^* = 39$ bestbewerteten Gene lieferte einen DAG von $n_T = 594$ biologischen Prozessen, wobei Werte von 82.32% und 70.97% für Precision bzw. Recall erreicht wurden (siehe Abbildung B.4). Für eine Teilmenge von $k = 39$ zufällig ausgewählten Krebsgenen hingegen wurde in 1 000 Versuchen eine durchschnittliche Precision von 97.32% und ein durchschnittlicher Recall von 4.98% (maximaler Recall in 1 000 Versuchen: 20.32%) beobachtet. Die Verteilung der Anzahl der GO Terme aus den 1 000 ORAs mit $k = 39$ Genen hatte ein Maximum bei 3 GO Termen (siehe Abbildung B.5). In diesem Experiment wurden in 983 der 1 000 Versuche signifikante GO Terme durch die ORA gefunden, während 17 Versuche beendet wurden, ohne einen signifikanten biologischen Prozess unter den gegebenen Parametern zu finden.

In der Menge der $n_G = 387$ Suchtgene (vgl. Kapitel 4.1), wurden für 381 Gene Annotationen zu GO Termen gefunden, mit denen die nachfolgende Analyse erfolgte. Eine ORA, die mit den gleichen Parametern durchgeführt wurde wie für die Schmerzgene, lieferte $n_T = 454$ signifikante GO Terme in der Ontologie der biologischen Prozesse. Die Menge der Suchtgene konnte auf eine Teilmenge von nur $k^* = 14$ Gene reduziert werden. Die ORA der $k^* = 14$ bestbewerteten Gene lieferte einen DAG von $n_T = 383$ biologischen Prozessen, wobei Werte von 79.37% und 66.96% für Precision bzw. Recall erreicht wurden (siehe Abbildung B.6). Für eine Teilmenge von $k = 14$ zufällig ausgewählten Suchtgenen hingegen wurde in 1 000 Versuchen eine durchschnittliche Precision von 58.77% und ein durchschnittlicher Recall von 0.59% (maximaler Recall in 1 000 Versuchen: 7.27%) beobachtet. Die Verteilung der Anzahl der GO Terme aus den 1 000 ORAs mit $k = 14$ Genen hatte ein Maximum bei nur einem GO Term (siehe Abbildung B.7). In diesem Experiment wurden aber überhaupt nur in 479 der 1 000 Versuche signifikante GO Terme durch die ORA gefunden, während 521 Versuche beendet wurden, ohne einen signifikanten biologischen Prozess unter den gegebenen Parametern zu finden.

Tabellen 4.1: Übersicht von Precision, Recall und F_1-Maß für die optimale Größe der Teilmenge je Verfahren. Fett markiert: Ausgewähltes Optimum $k^* = min(k', k'')$ pro Datensatz bei der funktionserhaltenden, integrativen Genselektion.

Verfahren:	Schnittpunkt Precision- und Recall-Kurve			
Maß: / Datensatz:	Anz Gene k'	Precision	Recall	F_1-Maß
Schmerz	**29**	**70.1434**	**69.6035**	**69.8724**
Sucht	19	68.9956	75.6168	72.1546
Krebs	63	75.6168	70.3571	72.8922
miRNA	**28**	**70.5251**	**70.6965**	**70.6107**

Verfahren:	Abstand der F_1-Kurve zu (0,1)			
Maß: / Datensatz:	Anz Gene k''	Precision	Recall	F_1-Maß
Schmerz	52	68.2203	84.6255	75.5425
Sucht	**14**	**79.3734**	**66.9604**	**72.6404**
Krebs	**39**	**82.3232**	**70.9724**	**76.2276**
miRNA	37	67.8571	76.9048	72.0982

5 Diskussion

In diesem Kapitel wird die im vorhergehenden Kapitel vorgestellte Methode der funktionserhaltenden, integrativen Genselektion kritisch diskutiert. Sie wird mit einem ähnlichen Verfahren von [Acharya et al., 2017] verglichen, wobei Vor- und Nachteile aufgezeigt werden. Zudem werden die erzielten Ergebnisse erörtert und validiert.

5.1 Diskussion der Methode

In der vorliegenden Arbeit wurde eine Methode entwickelt, um eine Menge von Objekten auf eine optimale Teilmenge von relevanten Objekten zu reduzieren. Dabei soll die Menge aller Objekte den Begriffen einer Ontologie zugeordnet werden können, die in Form eines DAGs Wissen über die Objekte repräsentiert. Weiterhin wird angenommen, dass die Menge aller Objekte unüberschaubar groß und damit für einen Betrachter unverständlich ist. Die gesuchte Teilmenge sollte daher möglichst klein sein, so dass anhand der wenigen, herausgefilterten Objekte die Menge aller Objekte verstanden und beschrieben werden kann. Gleichzeitig sollte durch die Objekte in der Teilmenge auch so viel wie möglich des DAGs, also des Wissens über die gesamte Menge, reproduziert werden können.

Für das Anwendungsbeispiel von Sätzen von Genen, deren Funktionen in der Gene Ontology Wissensbasis beschrieben werden, filtert die funktionserhaltende, integrative Genselektion aus einer großen Menge von Genen die kleinstmögliche Menge von relevanten Genen heraus, so dass die biologischen Prozesse und Funktionen, die die Gene der gesamten Menge ansprechen, möglichst gut erhalten bleiben.

Dieser Idee könnte man den unweigerlichen Informationsverlust, den eine Selektion mit sich bringt, entgegenstellen. Allerdings ist die „Kunst des Übersehens, die zu einem neuen Sehen des sonst Unsichtbaren führt“ [Schwemmer, 2008/2009, p. 14] eine wichtige Methode des wissenschaftlichen Fortschritts [Ultsch, 2019]. So können durch die Auswahl der wichtigsten Elemente einer Menge redundante und irrelevante Informationen entfernt werden und dadurch

Rechenzeit und –kosten verringert, Vorhersagemodelle aufgrund der bereinigten Daten verbessert und gleichzeitig ein besseres Verständnis der Daten ermöglicht werden [Guyon et al., 2003].

Da die Methode unabhängig von der Quelle der Gensätze konzipiert wurde, wurde im Vergleich zu existierenden Methoden, die anhand von experimentellen Messwerten optimale Teilmengen von Genen selektieren, ein neues Bewertungskriterium benötigt. Daher wurde ein integrativer Ansatz verfolgt, der die Struktur einer Ontologie als Wissensbasis ausnutzt und anhand derer eine Bewertung der Gene vorgenommen werden kann.

Dass die Methode nicht auf experimentelle Messwerte angewiesen ist, sollte keinesfalls als Nachteil angesehen werden. Die Methode schließt die Analyse experimentell gefundener Gensätze nicht aus, sondern kann zusätzlich auch Mengen von Genen, die nur durch Literaturrecherche oder Fachexpertise zusammengestellt wurden, auf ihre wichtigsten Elemente hin untersuchen. Dadurch kann auf Experimente verzichtet werden, wodurch unter anderem Kosten und Zeit gespart werden können. Allerdings ist zu bemerken, dass die Methode kritisch von der Qualität der zugrundeliegenden Wissensbasis abhängt. Jede Auswertung kann nur so gut sein wie die Datengrundlage. Gleiches gilt aber ebenso für Methoden, die experimentelle Messwerte zur Selektion heranziehen. Bei der Erhebung experimenteller, insbesondere klinischer Daten ist stets mit Messungenauigkeiten zu rechnen [Stommel/Wills, 2004]. So wurde z.B. festgestellt, dass die Ergebnisse von Microarray-Experimenten kritisch von vielen verschiedenen Faktoren wie z.B. der Raumtemperatur abhängen [Jaksik et al., 2015]. Daher könnte ein (systematischer) Fehler in den experimentellen Daten sich leicht in den Ergebnissen fortsetzen, wogegen die hier vorgestellte Methode wegen der von den Messwerten unabhängigen Bewertung resistent ist.

Es spricht theoretisch nichts dagegen die Methode auch auf andere Wissensbasen zu übertragen, da die verwendeten Maße dem Konzept der Wissensrepräsentation und –verarbeitung in der künstlichen Intelligenz entsprechend nicht inhaltlich an das Wissen der Gene Ontology Wissensbasis gebunden sind. Sie wurden allgemeingültig anhand der Struktur Gene Ontology gewählt, so dass die Wissensrepräsentation austauschbar ist, sofern nur die Struktur übereinstimmt. Findet sich also eine bessere Wissensbasis, kann dies direkt durch die hier vorgestellte Methode ausgenutzt werden.

Für die Anwendung der funktionserhaltenden, integrativen Genselektion auf Mengen von Genen wurde als Wissensbasis die Gene Ontology Wissensbasis ausgewählt. Dadurch konnte die Methode an verschiedenen Datensätzen von Genen getestet werden. Ein Problem der Gene Ontology Wissensbasis ist, dass sie nie vollständig sein kann, da durch Forschung immer wieder neue Erkenntnisse gewonnen werden, andererseits wird sie dadurch ständig aktualisiert und repräsentiert den neusten Stand der Forschung [Gaudet/Dessimoz, 2017]. Ein weiterer Nachteil der Gene Ontology, der sich direkt aus der Unvollständigkeit ergibt, ist die unterschiedlich intensive Forschung. Genfunktionen, die leicht zu untersuchen sind oder die potentiell interessanter erscheinen als andere, werden entsprechend stärker erforscht, so dass auch mehr Annotationen gefunden werden [Gaudet/Dessimoz, 2017]. Dadurch verzerrt sich die Verteilung der Annotationen, die zum Beispiel in der ORA verwendet wird [Gaudet/Dessimoz, 2017]. Es kann also sein, dass zukünftige Analysen mit einer aktualisierten Wissensbasis andere Ergebnisse liefern [Gaudet/Dessimoz, 2017]. Allerdings ist davon auszugehen, dass eine Wissensbasis im Laufe der Zeit durch jeden Erkenntnisgewinn die Realität besser modelliert und daher auch die Ergebnisse von Analysen wie der funktionserhaltenden, integrativen Genselektion verbessert werden. Eine andere Schwierigkeit, die sich aus der Verwendung der Gene Ontology Wissensbasis ergibt ist, dass die GO Terme nicht in allen Bereichen der Struktur der GO gleichermaßen detailliert sind, auch wenn sie gleich viele Vorfahren haben [Gaudet/Dessimoz, 2017]. Dagegen hilft es informationsbasierte Maße einzusetzen, anstatt die Längen der Pfade zwischen den GO-Termen als Maß zu betrachten [Gaudet/Dessimoz, 2017], was in der vorliegenden Arbeit getan wurde. Ebenso zu beachten sind die unterschiedlichen Evidenzen, mit denen die Annotationen zu den GO Termen erstellt werden [Gaudet/Dessimoz, 2017]. Diese sind unterschiedlich zuverlässig [Gaudet/Dessimoz, 2017]. In der vorliegenden Arbeit wurde dies berücksichtigt, indem nur manuell kuratierte Annotationen verwendet wurden, die den höchsten Grad an Zuverlässigkeit haben [Gaudet/Dessimoz, 2017].Werden all diese Schwierigkeiten beachtet, ist die Gene Ontology Wissensbasis eine formidable Quelle von biologischem Wissen [Gaudet/Dessimoz, 2017] und nicht umsonst die zurzeit erfolgreichste biologische Ontologie [Hastings, 2017].

Die Gewichtung der GO Terme nach ihrer Wichtigkeit wird mit Hilfe des Remarkableness-Wertes bestimmt. Dies lag nahe, da Remarkableness zur Functional Abstraction [Ultsch/Lötsch, 2014a] verwendet wurde und sich dort bereits

als geeignetes Maß für die Wichtigkeit von GO-Termen herausgestellt hatte [Ultsch/Lötsch, 2014a]. Remarkableness ist definiert als Produkt aus Certainty und Information Value (siehe Kapitel 4.2). Certainty drückt die Sicherheit aus, mit der ein GO Term die Menge der analysierten Gene beschreibt. GO Terme, die der hypergeometrische Test mit sehr kleinem p-Wert bewertet, werden als besonders sicher angesehen, GO Terme mit größerem p-Wert als unsicherer. Dies ergibt sich daraus, dass die p-Werte der GO Terme, die aus einer ORA resultieren, direkt auch die Effektstärke der ORA beschreiben [Thrun, 2018]. Der Information Value bestimmt, wie informativ dieser GO Term im Sinne der Shannon-Entropie ist. Beide Werte nutzen die Struktur der Ontologie aus, sind dabei aber trotzdem unabhängig von den Pfadlängen zwischen den GO-Termen wie in [Gaudet/Dessimoz, 2017] empfohlen.

Nachdem die Gene mit Hilfe des Gen-Scores bewertet und in eine ihrer Wichtigkeit entsprechenden Reihenfolge gebracht wurden, wird eine optimale Teilmenge gesucht. Dabei soll ein Benutzer insofern unterstützt werden, als er keine eigene Vorgabe machen muss, wie viele Gene er als optimal erachtet und eine mathematische, eindeutige und reproduzierbare Auswahl getroffen wird. Zunächst wird der Schnittpunkt der Precision- und Recall-Kurve als optimaler, leicht zu berechnender Punkt gewählt. Anschließend wird die als datenbionisches Optimalitätskriterium bewährte ABC-Analyse herangezogen, die Aufwand gegen Ertrag abwägt. In diesem Fall wird also die Anzahl der verwendeten Gene gegen den Wert des F_1-Maßes aufgetragen und der Punkt des kürzesten Abstands zum Punkt (0 Gene, 100% F_1-Maß) auf der sich ergebenden, empirischen F_1-Kurve bestimmt. Da möglichst wenige Gene in der optimalen Teilmenge sein sollen, wird die kleinere Teilmenge von Genen aus den zwei Verfahren als optimale Teilmenge ausgewählt.

Da sich das F_1-Maß aus Precision und Recall zusammensetzt, kann das Ergebnis der ABC-Analyse detaillierter bewertet werden und von einem Experten in bekannten Größenordnungen eingeschätzt werden. Zukünftig wäre es auch vorstellbar, die Pareto-Front [Gandibleux et al., 2012] der Lösungen zu betrachten, wodurch ein Anwender der integrativen, funktionserhaltenden Genselektion individueller entscheiden könnte, ob bei der Auswahl der optimalen Menge von Genen mehr Wert auf Precision oder Recall bezüglich der Reproduktion des DAGs gelegt werden soll.

Bei der Bestimmung der optimalen Menge von Genen wurde ursprünglich in [Lippmann et al., 2019] in der Matrix M, die die Annotationen der Gene zu den GO Termen repräsentiert und in der Remarkableness als Gewichtung der Terme eingetragen ist (siehe Kapitel 4.2 in Formel (4.1)), als erste einfache Idee die Zeilensumme als Gen-Score für jedes Gen gewählt, wodurch eine Rangfolge der Gene festgelegt werden konnte. Auf diese Matrix M kann aber jeder Genselektionsalgorithmus angewendet werden, der keine Klassenlabels voraussetzt, also unüberwachte Algorithmen [Solorio-Fernández et al., 2019]. Beispielsweise können der Gewichtsvektoren der Random Forest Methode zum Erstellen einer Rangfolge der Gene verwendet werden (vgl. Abschnitt 5.3; insbesondere Tabellen 5.1). Andere unüberwachte Algorithmen werden beispielsweise genutzt, um Cluster zu erstellen. Dafür muss aber die Clusteranzahl vorgegeben werden, die in diesem Fall nicht bekannt ist. Außerdem beinhalten die Cluster zunächst nur ähnliche Mengen von Genen, die zusammengefasst wurden. Es wird dadurch aber noch keine relevante Teilmenge bestimmt. Es müssten also Repräsentanten der Cluster ausgewählt werden, wobei die gleiche Schwierigkeit wie bei [Acharya et al., 2017] auftritt, dass entweder die Form der Cluster vorgegeben wird und mittlere Elemente als Repräsentant ausgewählt werden - was nicht sehr robust und inhaltlich nicht unbedingt valide ist - oder andererseits die Form von Clustern nicht vorgegeben ist, aber dann nicht klar ist, was ein guter Repräsentant eines unförmigen Clusters ist (vgl. Abschnitt 5.2).

Um zu messen wie gut die mit der funktionserhaltenden, integrativen Genselektion gefundenen Teilmengen von Genen den DAG reproduzieren, der aus der ORA der Menge aller Gene resultiert, wurden die im Information Retrieval üblichen Maße Precision und Recall [Baeza-Yates/Ribeiro, 2011] verwendet. Sie bewerten die Teilmengen aus den k bestbewerteten Genen jeweils durch den Anteil der gefundenen und relevanten GO Terme an allen gefundenen GO Termen (Precision) bzw. durch den Anteil der gefundenen und relevanten GO Terme an allen relevanten GO Termen (Recall) [Gaudet/Dessimoz, 2017].

Um Precision und Recall zu einem Maß zusammenzufassen wird im Information Retrieval oft das F_1-Maß verwendet [Zhai/Massung, 2016]. Es kann als Versuch, einen bestmöglichen Kompromiss zwischen Precision und Recall zu finden, interpretiert werden [Baeza-Yates/Ribeiro, 2011]. In der vorliegenden Arbeit wurde mit Hilfe des F_1-Maßes der optimale Punkt unter Berücksichtigung der Anzahl der Gene in der Teilmenge gefunden. Bei der Berechnung des F_1-Maßes

ist kritisch, dass nur drei der vier Felder der Kontingenztabelle (vgl. Tabelle 2.3) verwendet werden. Die Anzahl der nicht gefundenen und auch nicht relevanten, also der richtig negativen Objekte, wird nicht berücksichtigt. Allerdings wurde zur Validierung die optimale Anzahl der Gene auch durch Matthews Correlation Coefficient [Matthews, 1975] bestimmt, der alle vier Felder der Kontingenztabelle berücksichtigt. Die Versuche lieferten die gleichen Anzahlen von Genen, so dass die Wahl auf das einfacher zu berechnende F_1-Maß fiel.

Bei der Berechnung des Gen-Scores werden Zeilensummen der Matrix gebildet, die die Annotationen der Gene zu den signifikanten Termen beschreibt. Das heißt Gene, die zu vielen Termen annotiert sind, haben tendenziell einen höheren Gen-Score als solche, die nur zu sehr wenigen Termen annotiert sind. Daher liegt die Annahme nahe, dass das Verfahren nur Gene herausfiltert, die zu besonders vielen Termen annotiert sind. Dieses Argument ist nicht ganz von der Hand zu weisen. Allerdings wurde bei der Validierung mit dem Random-Forests-Verfahren, das mit Hilfe der zuvor erwähnten Matrix eine andere Auswahl von Genen trifft, keine Verbesserung sondern eine Verschlechterung in Precision und Recall bezüglich der Reproduktion des Teil-DAGs, der die signifikanten Terme für die Schmerzgene enthält, festgestellt. Zudem konnte zumindest für die Schmerzgene gezeigt werden, dass die getroffene Auswahl auch bio-medizinisch relevante Gene beinhaltet.

5.2 Diskussion bestehender Verfahren

Die Idee, die Wichtigkeit der Gene mithilfe der Remarkableness der Terme, zu denen sie annotiert sind, zu bestimmen und anhand dieser Werte rechnerisch eine optimale Teilmenge auszuwählen, ist neu. Diese Idee beruht darauf, dass Gensätze, die anhand von aktuellem Wissen über die genetische Architektur von Merkmalen oder Krankheiten bestimmt wurden, ähnliche funktionelle Eigenschaften haben. Diese funktionellen Eigenschaften sind in der Gene Ontology abgebildet. Deren Struktur wurde zur Berechnung des Gen-Scores zur Bewertung der Wichtigkeit eines jeden Gens zur Reproduktion des Teil-DAGs ausgenutzt, der die funktionellen Eigenschaften für die gesamte Menge der Gene repräsentiert.

Bisher wurden zwar schon mehrere Verfahren vorgeschlagen, die die Gene Ontology oder eine andere Wissensbasis ausnutzen, aber meist wurden zusätzlich auch numerische Messergebnisse benötigt. Von den in Kapitel 3 vorgestellten Arbeiten anderer Autoren verwendet nur das Verfahren von [Acharya et al., 2017] - ebenso wie die in dieser Arbeit behandelte Methode - keine Expressionsdaten der Gene aus Laborexperimenten zur quantitativen Bewertung der Gene bei der Genselektion. Liegen für einen Satz von Genen numerische Messwerte vor, können sie aber dennoch mit beiden Methoden ausgewertet werden. Dies stellt einen Vorteil der funktionserhaltenden, integrativen Genselektion und der Methode von [Acharya et al., 2017] gegenüber den anderen Verfahren dar, da die Menge der Datensätze, die analysiert werden können, erweitert wird.

Die Idee von [Acharya et al., 2017] Cluster von funktionell ähnlichen Genen zu bilden und anschließend einen Repräsentanten zu wählen, erscheint interessant und sinnvoll, allerdings ist die Umsetzung verbesserungswürdig. Das Verfahren Gene, die „in der Mitte“ eines Clusters liegen, als Repräsentant zu wählen ist nicht besonders robust, da bei einer kleinen Veränderung der Auswahl der Stichproben, die für den Klassifikator als Training verwendet werden, nicht unbedingt wieder die gleichen Gene die Mitte der Cluster bilden [Huawen Liu et al., 2010]. Zudem wird durch die Verwendung des euklidischen Abstands als Abstandsmaß implizit eine sphärische Form der Cluster angenommen [Kuri-Morales/Aldana-Bobadilla, 2010], die nicht unbedingt gegeben sein muss, für nicht-sphärische Clusterformen ist die „Mitte“ eines Clusters unter Umständen gar nicht eindeutig definiert.

Für die funktionserhaltende, integrative Genselektion hingegen konnte in Abschnitt 4.2 gezeigt werden, dass es sehr robust gegenüber Veränderungen des Inputs ist.

Leider war es nicht möglich den hier vorgestellten Ansatz mit dem von Acharya et al. noch genauer zu vergleichen. In [Acharya et al., 2017] wurden zwei Beispieldatensätze, Experimente zur Genexpression für Hefe und für einen Datensatz, der verschiedene Krebsarten beinhaltet, präsentiert. Aus diesen Beispieldatensätzen wurden mit dem vorgeschlagenen Algorithmus Gene selektiert. Da der hier vorgeschlagene Algorithmus nur menschliche Gene verwendet, war nur der zweite Datensatz von Interesse für diese Arbeit. Folgt man den Schritten wie in [Acharya et al., 2017] angegeben, erhält man allerdings schon beim ersten Schritt – der ORA mit Hilfe des Programms [Mi et al., 2013], das auf

der GO Webseite (http://geneontology.org/) zur Verfügung steht – Ergebnisse, die nicht mit den im Paper dokumentierten übereinstimmen. Es wird angegeben, dass zu den 5565 Genen, die mit verschiedenen Krebsarten assoziiert wurden, nur 147 signifikante GO Terme (davon 71 BP, 42 MF, 34 CC) gefunden wurden - bei einer p-Wert-Schranke von 0.5[sic!]. Beim Nachvollziehen der Ergebnisse wurden zu den 5565 Genen mit den voreingestellten Standardparametern (p-Wert Schranke von 0.05; Bonferroni-Korrektur für multiples Testen, vollständige Version der GO, Homo Sapiens) allein im BP-DAG 8724 signifikante GO-Terme gefunden - bei einer schärferen p-Wert Schranke von 0.05, die nicht vom User eingestellt werden konnte. Um weniger signifikante GO-Terme zu erhalten, gibt es drei Möglichkeiten: 1. Eine schärfere p-Wert Schranke verwenden, 2. Eine GO-Slim-Version verwenden, also eine ausgedünnte GO in der viele der detaillierteren Knoten fehlen [The Gene Ontology Consortium, 2019], 3. Eine schärfere Korrektur für multiples Testen verwenden. Es fehlen Angaben, ob eine Korrektur für multiples Testen durchgeführt oder eine GO-Slim-Version verwendet wurde. Verwendet man die Bonferroni Korrektur, eine GO-Slim-Version und die p-Wert Schranke von 0.05, erhält man immer noch 234 signifikante GO-Terme allein in BP. Diese große Diskrepanz ist möglicherweise durch die unterschiedlichen Zeitpunkte der Verwendung der GO zu erklären (Einreichen des Papers von [Acharya et al., 2017] im August 2017 – wann die angegebenen Webseiten verwendet wurden, ist im Paper nicht angegeben; Nachvollziehen der Ergebnisse 19.12.2018), dies ist aber recht unwahrscheinlich. Da man nun raten müsste, wie die Parameter gewählt wurden um auf die angegebenen Zwischenergebnisse zu kommen, kann dieses Verfahren nicht zum Vergleich mit dem hier vorgestellten Algorithmus herangezogen werden. Verwendet man den Datensatz mit dem hier vorgestellten Algorithmus, erhält man eine Menge von 58 Genen, mit Precision und Recall jeweils über 71%, im Paper wird eine optimale Menge von 40 Genen angegeben, was größenordnungsmäßig übereinstimmt, allerdings werden die einzelnen 40 Gene nur als abgekürzte Symbole aufgelistet, die nicht eindeutig sind, so dass ein Vergleich der Ergebnisse auch hier nicht exakt möglich ist. Der Schnitt der beiden Mengen enthält nur ein einziges Gen. Die Idee hinter der Clusterung von [Acharya et al., 2017] ist allerdings auch nicht die funktionell wichtigsten Gene zu finden, sondern Gene, die anschließend für eine gute Clusterung verwendet werden können bzw. diejenigen, die Klassen für einen nachfolgenden Klassifikator bilden.

Das Verfahren kann also als teils verbesserungswürdiger, alternativer Ansatz verstanden, aber nicht zum direkten Vergleich mit der funktionserhaltenden, integrativen Genselektion herangezogen werden.

5.3 Diskussion der Resultate

Mit der vorgeschlagenen Methode, um genetische Information auf eine relevante Teilmenge von Genen zu reduzieren, gelang es mehr als zwei Drittel der signifikanten biologischen Prozesse, zu denen die 540 Schmerzgene annotiert sind, mit nur etwa 5% der gesamten Menge der Schmerzgene zu reproduzieren. Der Idee der vorliegenden Analyse folgend, bilden die $k^* = 29$ bestbewerteten Schmerzgene (siehe Tabelle C.) die relevantesten Gene, die mit der Krankheit, d.h. Schmerz, assoziiert werden. Tatsächlich beinhaltet die Teilmenge verschiedene aus der biomedizinischen Forschung bekannte Schlüsselgene. Beispielsweise geben das Opioid- und Toll-like System, die durch *OPRM1* bzw. *TLR4* repräsentiert werden, einen Hinweis auf die Interaktion des Immunsystems mit chronischen Schmerzen mit Fokus auf neuroimmunen Crosstalk am sogenannten „Glial-Opioid Interface" [Tian et al., 2012]. Weitere der $k^* = 29$ Gene weisen auch auf Immun-Prozesse hin welche als gemeinsamer Nenner chronischer Schmerzzustände gefunden werden [D. Kringel et al., 2018]. Diese stehen im besonderen Interesse öffentlich finanzierter, konzertierter Forschungen z.B. [Dario Kringel/Lötsch, 2015], was das Interesse der Schmerz- und Analgesieforschung zeigt. Eine deutliche Wichtigkeit der $k^* = 29$ Gene für Schmerz wird auch durch ihre Rolle als Arzneimittelzielstrukturen von Analgetika in der aktiven Entwicklung und Forschung belegt. Eine Abfrage nach „Lead Compounds AND Under Active Development AND Condition = ‚Pain'" der Thomson Reuters Integrity Datenbank (https://integrity.thomson-pharma.com, Zugriff 22 Mai 2018), einer kommerziellen Datenbank, die bereits zugelassene oder in klinischen Entwicklungsphasen befindliche Analgetika listet, identifizierte 15 der 29 bestbewerteten Schmerzgene (51,72%) als Arzneimittelzielstrukturen von neuen Analgetika, aber nur 142 der restlichen 511 Schmerzgene (27,79%). Die bestbewerteten Schmerzgene waren also gegenüber allen Schmerzgenen als Arzneimittelzielstrukturen für neue Medikamente statistisch signifikant überrepräsentiert. (Der exakte Test nach Fischer wies einen p-Wert von $p = 0.01027$ aus.)

Der vorliegende Ansatz zur Reduktion der genetischen Informationen, die relevant für ein Merkmal oder eine Krankheit sind, bietet eine Alternative zu vorherigen Vorschlägen, um die Datensätze auf ihre wichtigsten Repräsentanten zu reduzieren.

Zur weiteren Validierung wurde untersucht, ob das Verfahren der funktionserhaltenden, integrativen Genselektion vergleichbar gut ist wie andere Verfahren, die nur eine Matrix mit Gewichtungen der Wichtigkeit als Input benötigen. Viele Methoden (z.B. [Neumann et al., 2017], [Singh et al., 2002], [Qi/Tang, 2007]Qi) verwenden allerdings nicht nur eine Matrix, sondern benötigen auch a priori eine bekannte Klassenzugehörigkeit. Diese war für die gegebenen Datensätze nicht gegeben, so dass diese Verfahren sich nicht zum Vergleich eignen.

Ein bekannter und etablierter Algorithmus, der auch in R im Package „randomForest“ (https://cran.r-project.org/web/packages/randomForest/randomForest.pdf) bereits implementiert war, ist Random Forests. Random Forests wird genutzt, um Klassen für Gene anhand einer Kombination von Entscheidungsbäumen vorherzusagen [Breiman, 2001]. Jeder Entscheidungsbaum teilt die Gene in Klassen. Diejenige Klasse, in die ein Gen von den meisten Entscheidungsbäumen eingeteilt wurde, wird vom Random Forest verwendet [Breiman, 2001]. Dabei werden Gewichtsvektoren erstellt, die als Indikator für die Wichtigkeit der Gene herangezogen werden können. Für die Datensätze aus Kapitel 4.1 wurden die Ergebnisse in Tabellen 5.1 erzielt:

Für jeden der vier Datensätze ist die Auswahl der Teilmenge durch Random Forests größer als durch die funktionserhaltende, integrative Genselektion. Zudem sind auch alle Recall- und F_1-Maß-Werte kleiner als bei der funktionserhaltenden, integrativen Genelektion. Bei den Schmerzgenen, den Suchtgenen und den Genen, die mit miRNA assoziiert sind, sind auch alle Precision-Werte kleiner als bei der funktionserhaltenden, integrativen Genselektion. Bei den Krebsgenen, ist der Wert der Precision um 0.0795 größer, was 0.90% entspricht.

Es kann also festgehalten werden, dass für 3 der 4 Datensätze die Auswahl der funktionserhaltenden, integrativen Genselektion besser ist als die Auswahl mit Hilfe des Random Forests Verfahren und bei einem Datensatz ein ähnliches Ergebnis mit 27,78% weniger Genen erzielt werden kann.

Tabellen 5.1: Übersicht von Precision, Recall und F_1-Maß für die optimale Größe der Teilmenge je Verfahren, wobei die Gewichtsvektoren aus dem Random Forests-Verfahren als Bewertungskriterium herangezogen wurden. Fett markiert: Optimum $k^* = min(k', k'')$ pro Datensatz nach Random Forests.

Verfahren:	Schnittpunkt Precision- und Recall-Kurve			
Maß: / Datensatz:	Anz Gene k'	Precision	Recall	F_1-Maß
Schmerz	**40**	**68.9153**	**68.4626**	**68.6882**
Sucht	**26**	**62.3913**	**63.2159**	**62.8009**
Krebs	83	75.6996	74.6009	75.1462
miRNA	**51**	**70.4192**	**70.0000**	**70.2090**

Verfahren:	Abstand der F_1-Kurve zu (0,1)			
Maß: / Datensatz:	Anz Gene k''	Precision	Recall	F_1-Maß
Schmerz	58	68.4724	83.0486	75.0594
Sucht	38	62.0209	78.4141	69.2607
Krebs	**54**	**82.4027**	**70.6822**	**76.0938**
miRNA	67	68.0982	79.2857	73.2673

Random Forests ist selbstverständlich nicht der einzige Algorithmus, der keine Klassifikation der Daten voraussetzt. Es wäre daher möglich die funktionserhaltende, integrative Genselektion auch gegen andere Verfahren zu testen wie z.B. diverse Clusterverfahren mit anschließender Auswahl eines Repräsentanten wie z.B. auch in Kapitel 3 von [Acharya et al., 2017] vorgeschlagen. Dies würde allerdings den Rahmen dieser Arbeit sprengen und die Nachteile der Methode von [Acharya et al., 2017] wurden bereits im vorherigen Abschnitt 5.2 erläutert.

Zuletzt muss bemerkt werden, dass zur Interpretation der Ergebnisse der ORA und der Genselektion eine enge Zusammenarbeit zwischen Informatik- und Fachexperten stattfinden muss. In der vorliegenden Analyse wurde die biologische Plausibilität und Validität der Ergebnisse durch den Vergleich der bestbewerteten Schmerzgene mit den Arzneimittelzielstrukturen von zur Zeit in klinischen Phasen der Entwicklung befindlichen Analgetika bewertet. Dabei konnte ein deutlich höherer Anteil der bestbewerteten Schmerzgene an den Arzneimittelzielstrukturen festgestellt werden als für diejenigen Schmerzgene, die nicht durch die funktionserhaltende, integrative Genselektion ausgewählt wurden. Ähnliche fachspezifische Interpretationen waren weniger geeignet für die

Validierungs-Datensätze, daher wurde der Schmerzgendatensatz als hauptsächlicher Analysedatensatz ausgewählt.

6 Zusammenfassung und Ausblick

In der vorliegenden Arbeit wurde eine Methode, die funktionserhaltende, integrative Genselektion, zur Reduktion von Mengen von Objekten auf eine optimale Teilmenge von relevanten Objekten vorgestellt. Die Methode integriert dabei das Wissen und die Struktur aus einer zugrundeliegenden Ontologie in die Analyse, um die Objekte nach ihrer Wichtigkeit zu ordnen und anschließend die wichtigsten Objekte herauszufiltern. Als Anwendungsbeispiel wurde die Gene Ontology Wissensbasis [Ashburner et al., 2000] als Ontologie gewählt. Diese ist eine sehr gut gepflegte und formal strukturierte Quelle von Wissen über biologische Prozesse, molekulare Funktionen und zelluläre Komponenten von Genfunktionen [Gaudet/Dessimoz, 2017]. Für einen gegebenen Gensatz, der anhand von aktuellem Wissen über die genetische Architektur von Merkmalen oder Krankheiten gefunden wurde, werden mithilfe der Annotationen der Gene die signifikanten, biologischen Prozesse in der polyhierarchisch organisierten Gene Ontology Wissensbasis ermittelt. Der sich ergebende gerichtete, azyklische Graph (DAG) von signifikanten, biologischen Prozessen beschreibt die Genfunktionen des Gensatzes. Für jedes Gen in dieser Menge wird ein Gen-Score bestimmt, der sich aus der Position und dem Informationsgehalt der zuvor bestimmten Prozesse der Gene Ontology Wissensbasis ergibt. Mit Hilfe des Gen-Scores, der die Wichtigkeit der Gene beschreibt, können die Gene in eine Rangfolge gebracht werden. Aus den ersten k^* Genen wird die optimale Teilmenge von Genen bestimmt, indem die Teilmenge der Gene ausgewählt wird, die die beste funktionserhaltende Eigenschaft hat. Wie gut die Genfunktionen des DAGs der signifikanten, biologischen Prozesse erhalten werden, bestimmt sich über Precision und Recall bzw. deren Verrechnung zum F_1-Maß bezüglich der Reproduktion des DAGs mit der gewählten Teilmenge.

Für das Anwendungsbeispiel von Gensätzen, deren Funktionen in der Gene Ontology Wissensbasis beschrieben werden, filtert die funktionserhaltende, integrative Genselektion also aus einer großen Menge von Genen die kleinstmögliche Teilmenge von relevanten Genen heraus, so dass die biologischen Prozesse und Funktionen, die die Gene der gesamten Menge ansprechen, möglichst gut erhalten bleiben.

Für die untersuchten Datensätze konnte der ursprüngliche DAG jeweils mit Recall und Precision von etwa 70% reproduziert werden, wobei nur etwa 5% der ursprünglichen Gene verwendet wurden. In einem Zufallsexperiment mit 1 000

Wiederholungen lieferte jede andere, gleich große, zufällig ausgewählte Teilmenge aus der Menge aller Gene nur einen durchschnittlichen Recall von etwa 2% - in allen Versuchen wurden 15% nicht überschritten. Precision lag in allen Versuchen zwischen 58% und 98%. Der hohe Wert von Precision ergab sich daraus, dass es nur wenige GO Terme gab, zu denen signifikant mehr oder signifikant weniger der k^* Gene annotiert waren, als rein zufällig zu erwarten wäre. Das bedeutet, dass die Anzahl der falsch positiven GO Terme nahezu Null und somit Nenner und Zähler in Formel (2.1) in Definition 2.4.2 nahezu identisch waren. Recall hingegen war wegen der wenigen signifikanten GO Terme sehr niedrig, da viele GO Terme aus dem ursprünglichen DAG nicht mit der zufälligen Teilmenge von Genen reproduziert werden konnten.

Dass die gefundenen Gene in der Teilmenge als die wichtigsten Elemente der gesamten Menge angesehen werden können, wurde für schmerzrelevante Gene auch inhaltlich belegt. Es konnte durch einen Schmerzforscher die biologische Plausibilität validiert werden. Für die Hypothese die wichtigsten Gene gefunden zu haben spricht ebenso die Überrepräsentation der ausgewählten Gene im Vergleich zu der gesamten Menge der Gene in der Menge der Gene, die als Arzneimittelzielstrukturen für in klinischen Phasen der Entwicklung befindlichen Analgetika dienen. Mit der vorliegenden Methode war es möglich, eine Teilmenge schmerzrelevanter Gene zu finden, von denen mehr als die Hälfte bereits als mögliche Arzneimittelzielstrukturen für Analgetika erforscht werden.

Die funktionserhaltende, integrative Genselektion liefert eine Teilmenge von Genen, die zu untersuchen besonders interessant oder lohnenswert ist. Eine mögliche Anwendung der vorgestellten Methode könnte daher als Vorschlagssystem im Bereich der Drug Discovery liegen. Dabei könnte die Entwicklung von Medikamenten mit bekannten Arzneimittelzielstrukturen verbessert werden. Außerdem könnte versucht werden, insbesondere die durch die Genselektion identifizierten Gene, die noch nicht als Arzneimittelzielstrukturen genutzt wurden, durch neue Medikamente zu beeinflussen. Während es schwierig ist, diese Suche auf alle Elemente einer Menge von einigen hundert Genen auszudehnen, kann die Menge der zu untersuchenden Gene mit Hilfe der vorgestellten Methode auf weit unter hundert interessante Kandidaten eingegrenzt werden. Die deutlich kleinere Teilmenge der Gene kann als Ausgangspunkt für nachfolgende Laboruntersuchungen dienen.

Weiterhin wäre es interessant, die Methode zur Dimensionsreduktion in weiteren themenspezifischen Datensätzen von Genen anzuwenden und dadurch eine möglichst verständliche, aber trotzdem repräsentative Darstellung der Gesamtmenge der Gene zu finden.

Eine Idee für zukünftige Forschung ist es, weitere biologische Wissensbasen in die Analyse einzubinden. So konnte in [Richards et al., 2012] bei einem integrativen Ansatz gezeigt werden, dass die besten Ergebnisse mit einer Kombination von Wissen aus GO und Pub-Med-Literaturrecherche erzielt werden konnten, wobei die gleichzeitige Erwähnung von Genen in einem Fachartikel als Faktor gewertet wurde, dass diese Gene funktionell ähnlich seien [Richards et al., 2012]. Es wurden auch die KEGG- [Minoru Kanehisa et al., 2006] und MINT-Datenbank [Ceol et al., 2009] als Wissensbasen getestet, die aber keinen zusätzlichen Informationsgewinn lieferten [Richards et al., 2012]. Die Idee bei [Richards et al., 2012] war eine etwas andere als bei der funktionserhaltenden, integrativen Genselektion. Eine gegebene Liste von Genen wurde mit Hilfe von spektralem Clustering [Bolla, 2013] in Teilmengen aufgeteilt, so dass jede Teilmenge eine funktionell zusammenhängende Gruppe von Genen bildet. Es wurden also in der Liste der Gene Cluster gesucht und nicht wie in der vorliegenden Arbeit die wichtigsten Gene, die die Funktionen der Gesamtmenge möglichst gut abbilden. Dennoch hat sich „Mixture of Experts“ als Problemlösungsansatz bewährt, bei dem angenommen wird, dass ein Komitee aus verschiedenen Experten zu einer besseren Lösung kommt, als jeder einzelne allein [Alam, 2014, p. 81], da sich ihr Wissen komplementiert.

Anhang A. Beispiele verschiedener Genselektionsverfahren

Tabelle A.1: Detailliertere Übersicht einer Auswahl von Beispielen verschiedener Genselektionsverfahren.

Methode und Anwendungsbeispiele	**Typ**	**Beschreibung**
ReliefF [Kononenko, 1994] Beispiele: [Yuhang Wang et al., 2005], [P. Yang et al., 2010], [Papachristoudis et al., 2010], [Cai et al., 2014]	Filter	Die dem ReliefF-Algorithmus zugrunde liegende Idee ist es für einen gegebenen Datensatz, dessen Klassen bekannt sind, wiederholt zufällig Proben zu ziehen, anschließend deren nächste Nachbarn in der gleichen und allen anderen Klassen zu bestimmen und anhand dessen einen Gewichtsvektor für die Gene zu berechnen, der diejenigen Gene stärker gewichtet, die Proben unterschiedlicher Klassen voneinander unterscheiden und Proben gleicher Klassen zusammen gruppieren können [Yuhang Wang et al., 2005].
Information Gain Beispiele: [Yuhang Wang et al., 2005], [P. Yang et al., 2010], [Nguyen et al., 2015], [Papachristoudis et al., 2010]	Filter	Information Gain (IG) misst den Mehrwert eines Gens in Bezug auf eine Klassenvorhersage für eine neue Probe anhand der dazugewonnenen Information durch das Wissen um die Werte, die dieses Gen annimmt [Yu Wang et al., 2005]. Ist eine Menge von Klassen $\{c_i\}_{i=1}^{m}$ gegeben und ist V die Menge der möglichen Expressionswerte, die ein Gen annehmen kann, dann ist der IG eines Gens definiert als: $IG(g) \coloneqq$ $-\sum_{i=1}^{n} P(c_i)\log_2\big(P(c_i)\big)$ $+\sum_{v\in V}\sum_{i=1}^{m} P(g = v)P(c_i\|g = v)\log_2(P(c_i\|g = v))$
Chi-Quadrat Beispiele: [Yuhang Wang et al., 2005], [P. Yang et al., 2010], [Papachristoudis et al., 2010]	Filter	Mit dem Chi-Quadrat Test wird die Abweichung von der erwarteten Verteilung unter der Annahme, dass Gene und Klassen voneinander unabhängig sind, gemessen [Khalid et al., 2014]. Ist also die Chi-Quadrat Statistik für ein Gen besonders hoch, kann angenommen werden, dass die Annahme falsch war, dass also die Klassen mit Hilfe des Gens erklärt

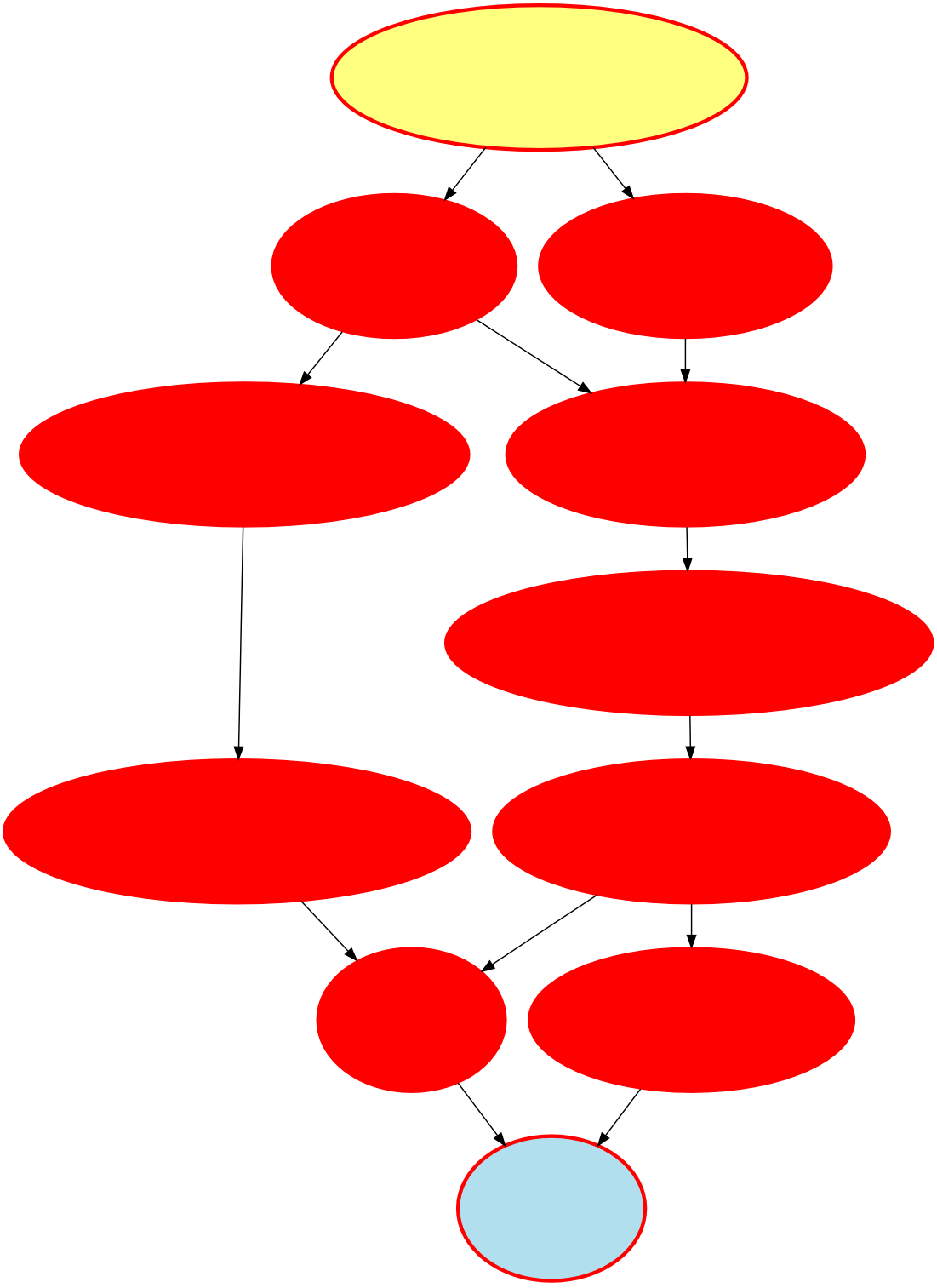
		werden können. Für Klassen $\{c_i\}_{i=1}^{m}$ und mögliche Expressionswerte des Gens V berechnet sich die Statistik als: $\chi^2(g) = \sum_{v \in V} \sum_{i=1}^{m} \frac{\left(N(g=v,c_i) - E(g=v,c_i)\right)^2}{E(g=v,c_i)}$, wobei $N(g=v,c_i)$ und $E(g=v,c_i)$ das beobachtete bzw. erwartete gemeinsame Auftreten der Klasse c_i und des Gens g, das den Wert v hat, sind [P. Yang et al., 2010].
Correlation-based Feature Selection[Hall, 1999] Beispiel: [Yu Wang et al., 2005]	Filter	Korrelationsbasierte Feature Selektion bewertet eine Teilmenge von Genen danach, wie gut die einzelnen Gene der Teilmenge Klassen vorhersagen können und beachtet dabei deren Korrelation [Hira/Gillies, 2015] mit dem Gedanken, dass gute Teilmengen Gene enthalten, die stark mit der Klassifikation, aber untereinander wenig korrelieren [Hall, 1999].
Minimum redundancy-maximum relevance (MRMR) [Ding/Peng, 2005] Beispiel: [Ding/Peng, 2005]	Filter	Die MRMR-Methode zielt darauf ab, maximal relevante und minimal redundante Teilmengen von Genen auszuwählen, um zwischen Klassen zu diskriminieren. [Huang et al., 2018]. Dabei wird die Korrelation der Gene untereinander berücksichtigt, um Redundanz zu vermeiden und Information Gain um maximale Relevanz zu erhalten [Ding/Peng, 2005].
Genetische Algorithmen [Sivanandam/Deepa, 2007] Beispiele: [Ooi/Tan, 2003], [Jirapech-Umpai/Aitken, 2005]	Wrapper	Genetische Algorithmen (GA) sind randomisierte Such- und Optimierungstechniken, die in Analogie zu Prinzipien der Evolution und natürlichen Selektion designt werden [Ooi/Tan, 2003]. In traditionellen GAs wird ein String fester Länge, der Nullen und Einsen enthält, als Repräsentation verwendet [Sivanandam/Deepa, 2007]. Von jeder Position des Strings wird angenommen, dass sie eine bestimmte Eigenschaft eines Individuums repräsentiert und der Wert dieser Position die Ausprägung dieser Eigenschaft beschreibt [Sivanandam/Deepa, 2007]. Im Falle der

		Genselektion steht jede Position des Strings für ein Gen und bestimmt, ob das Gen für die Genselektion ausgewählt wird oder nicht. Strings werden dann nach verschiedenen Verfahren (analog zu Vererbung und Mutation in der Genetik) miteinander kombiniert, wodurch eine neue „Generation“ von Strings entsteht [Sivanandam/Deepa, 2007]. Zur Genselektion wird deren Güte zur Lösung eines Klassifikationsproblems mit verschiedenen Methoden bestimmt [Sivanandam/Deepa, 2007]. GAs sind sehr robust und können komplexe und große Datenmengen nach Lösungen des Genselektionsproblems durchsuchen [Ooi/Tan, 2003].
Sequential forward selection Beispiel: [Inza et al., 2002], [Xiong et al., 2001]	Wrapper	Sequential Forward Selection ist ein deterministischer Suchalgorithmus, der mit einer leeren Teilmenge von Genen startet, alle Gene durchsucht und das Beste zur Teilmenge hinzufügt [Inza et al., 2002]. So werden sukzessive Gene auswählt, bis keine Verbesserung der Bewertungsfunktion, also z.B. die Güte der Klassifikation, mehr erreicht wird [Inza et al., 2002].
Filter + Genetischer Algorithmus Beispiele: [C.-H. Yang et al., 2010], [Chuang et al., 2011]	Hybrid	In [C.-H. Yang et al., 2010] wird ein hybrides Verfahren aus einer Kombination des Information Gain Filters und eines genetischen Algorithmus, in [Chuang et al., 2011] aus einer Kombination des korrelationsbasierten Filters und eines Genetischen Algorithmus vorgeschlagen. Zunächst wird für alle Gene eine Rangfolge mit der Filtermethode festgelegt, anschließend werden die Gene, die als relevant befunden wurden, durch den Genetischen Algorithmus noch weiter selektiert [C.-H. Yang et al., 2010], [Chuang et al., 2011]. Die Güte der gefundenen Teilmengen der Gene wurde mit Hilfe der Klassifikationsgüte eines Klassifikators bestimmt [C.-H. Yang et al., 2010],

		[Chuang et al., 2011] und die beste Teilmenge ausgewählt.
Support Vector Machine Recursive Feature Elimination (SVM-RFE) [Guyon et al., 2002] Beispiele: [Guyon et al., 2002], [Y. Tang et al., 2007]	Embedded	SVM-RFE bildet eine Rangfolge der Gene mit Hilfe der Gewichte aus der SVM [Y. Tang et al., 2007]. Dann wird wiederholt eine kleine Menge von Genen mit den kleinsten Gewichten, die also am wenigsten zur Klassifikation beitragen, aus der Menge der Gene entfernt und die Gewichte neu berechnet [Y. Tang et al., 2007]. Dies erfolgt solange, bis die vorgegebene Anzahl an Genen übrig bleibt [Y. Tang et al., 2007].
Random Forest [Breiman, 2001] Beispiele: [Díaz-Uriarte/Alvarez de Andrés, 2006], [Gunther et al., 2003]	Embedded	Random Forests sagt Klassen für Gene anhand einer Kombination von Entscheidungsbäumen vorher [Breiman, 2001]. Jeder Entscheidungsbaum teilt die Gene in Klassen und diejenige Klasse, in die ein Gen von den meisten Entscheidungsbäumen eingeteilt wurde, wird vom Random Forest verwendet [Breiman, 2001]. Um die Random Forests zur Genselektion für Klassifikationsprobleme zu verwenden, werden iterativ Random Forests auf die Menge von Genen angepasst, nachdem diejenigen Gene, die am wenigsten wichtig für die Klassifikation waren, aus der Menge entfernt wurden [Díaz-Uriarte/Alvarez de Andrés, 2006]. Die reduzierte Menge von Genen, die schließlich ausgewählt wird, minimiert den Fehler der Klassifikation [Díaz-Uriarte/Alvarez de Andrés, 2006].
Kombination verschiedener Methoden zur Genselektion Beispiele: [Nguyen et al., 2015], [Saeys et al., 2008]	Ensemble	Es werden mit verschiedenen Genselektionsverfahren Gene ausgewählt und bewertet, so dass eine Rangfolge der Gene für jedes Verfahren entsteht [Nguyen et al., 2015]. Anschließend werden die am häufigsten als wichtig bewerteten Gene als relevante Teilmenge ausgewählt [Saeys et al., 2008].

Anhang B. Ergänzende Abbildungen

Abbildung B.1: Beispielhafter Ausschnitt aus dem DAG, der aus der ORA der Schmerzgene resultiert. Rote Knoten und rot umrandete Knoten sind signifikant überrepräsentiert, Blätter sind zusätzlich blau markiert und sogenannte Headlines [Ultsch/Lötsch, 2014a] gelb. Diese Graphik wurde mit dem R Package dbtORA [Lippmann et al., 2018] erzeugt.



Alle folgenden Abbildungen ebenso wie die entsprechenden Abbildungen in Kapitel 4.2 und 4.3 für die Schmerzgene wurden mithilfe des R Packages „ggplot2“ (https://cran.r-project.org/package=ggplot2; [Wickham, 2016]) erstellt.

Abbildung B.2: Empirische Precision-, Recall- und F_1-Maß-Kurven der signifikanten GO-Terme für den Datensatz der mit miRNAs interagierenden Gene. Die Kurven zeigen Precision, Recall und F_1-Maß für die signifikanten GO Terme, die aus ORAs der Teilmengen mit $k \in \{1, \dots, n_G\}$ der bestbewerteten mit miRNAs interagierenden Gene resultieren, bezüglich der GO-Terme, die aus der ORA mit der Gesamtmenge der mit miRNAs interagierenden Gene resultierten.

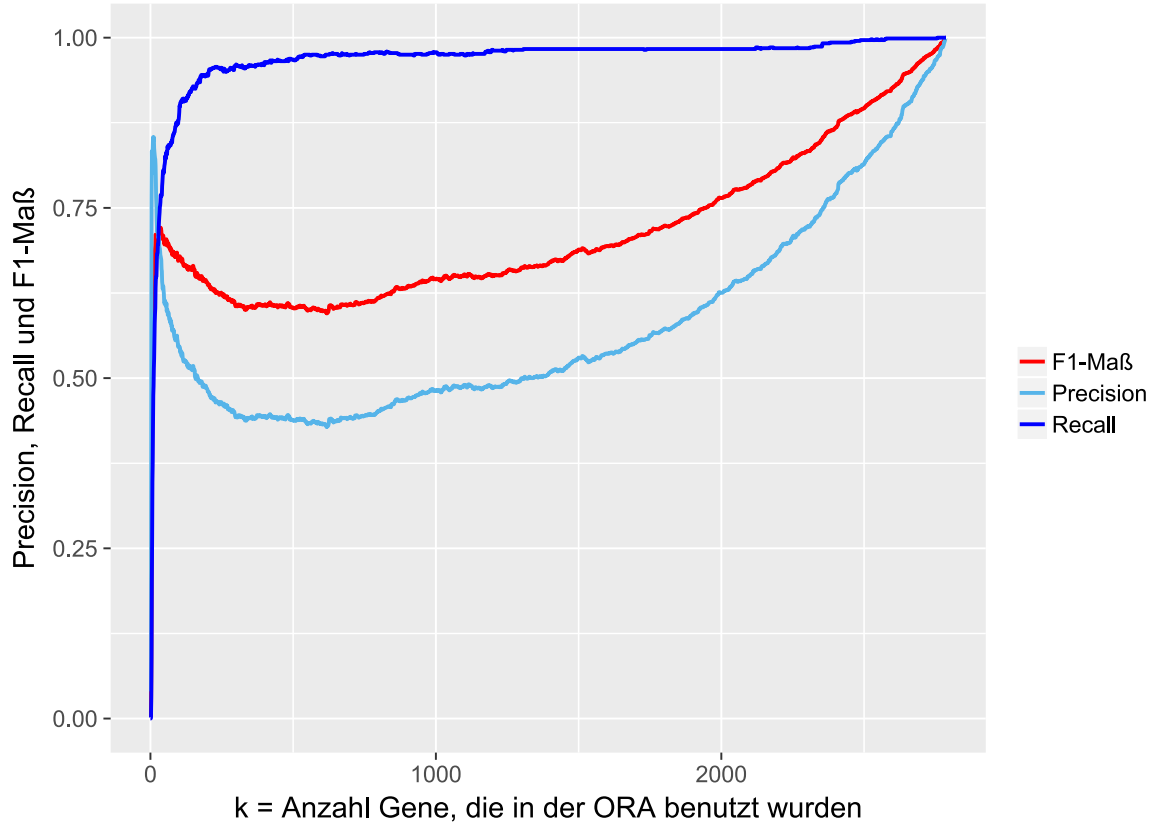


Abbildung B.3: Verteilung der Anzahl von signifikanten GO-Termen, die aus 1 000 ORAs mit zufällig gezogenen Teilmengen von $k^* = 28$ der 2954 mit miRNAs interagierenden Gene resultieren; Maximum bei nur etwa 1 GO-Term in der Ontologie der biologischen Prozesse der GO.

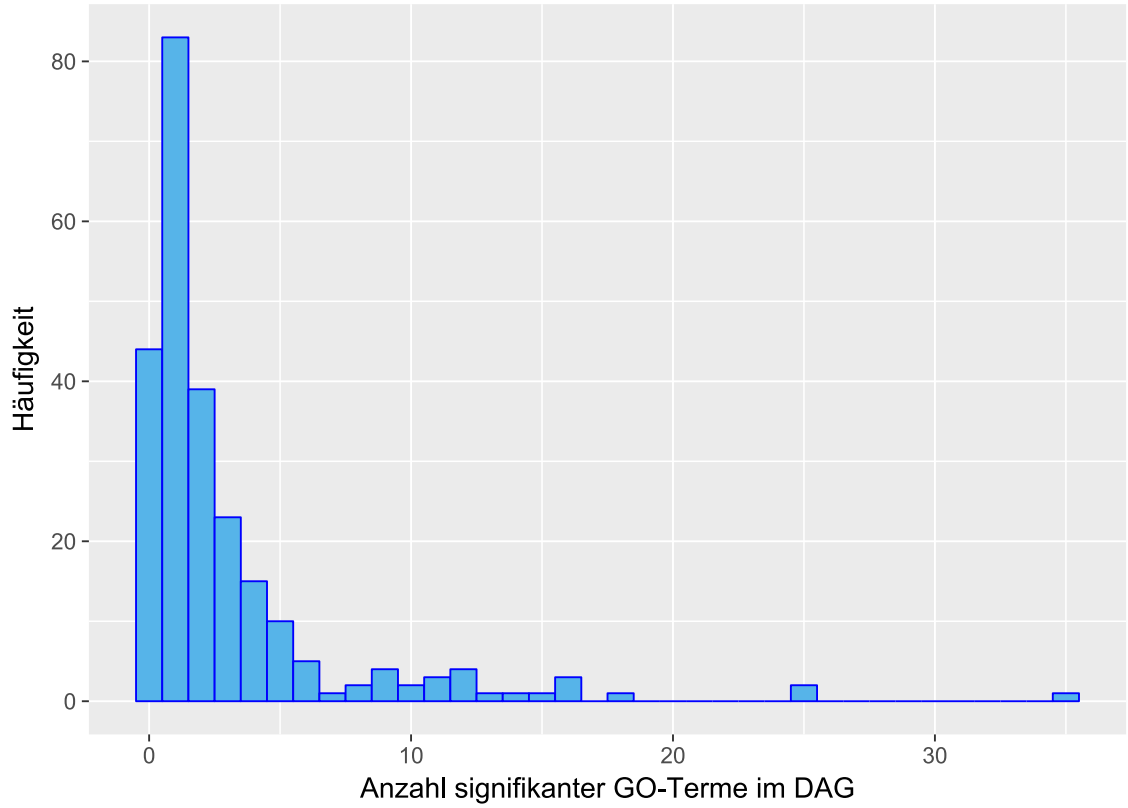


Abbildung B.4: Empirische Precision-, Recall- und F_1-Maß-Kurven der signifikanten GO-Terme für den Datensatz der Krebsgene. Die Kurven zeigen Precision, Recall und F_1-Maß für die signifikanten GO Terme, die aus ORAs der Teilmengen mit $k \in \{1, \ldots, n_G\}$ der bestbewerteten Krebsgene resultieren, bezüglich der GO-Terme, die aus der ORA mit der Gesamtmenge der Krebsgene resultierten.

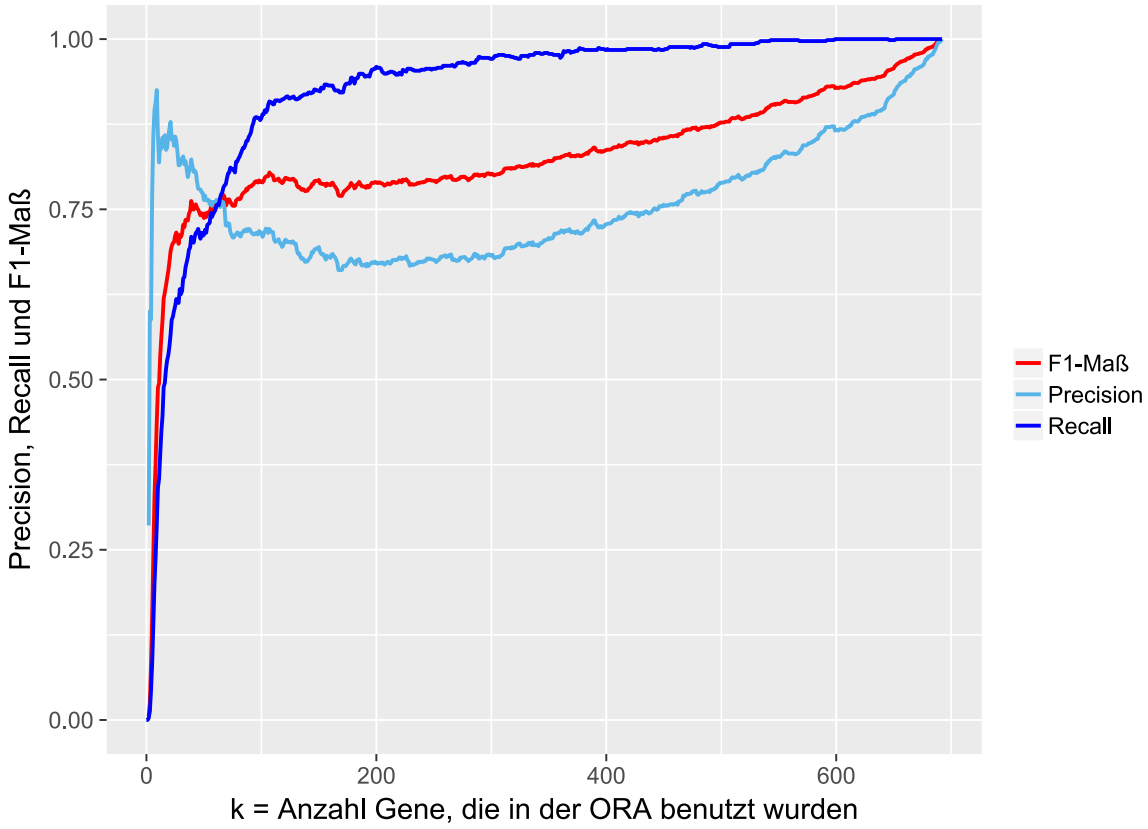


Abbildung B.5: **Verteilung der Anzahl von signifikanten GO-Termen, die aus 1 000 ORAs mit zufällig gezogenen Teilmengen von $k^* = 39$ der 719 Krebsgene resultieren; Maximum bei nur etwa 3 GO-Termen in der Ontologie der biologischen Prozesse der GO.**

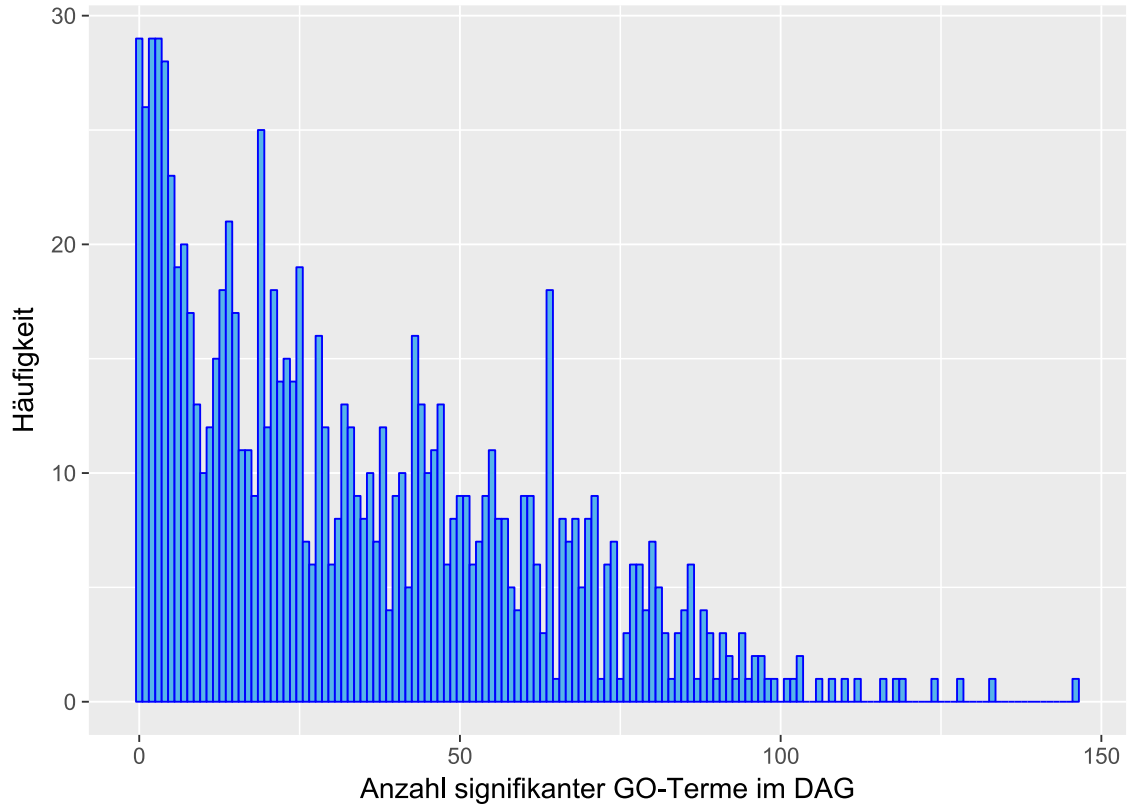


Abbildung B.6: Empirische Precision-, Recall- und F_1-Maß-Kurven der signifikanten GO-Terme für den Datensatz der Suchtgene. Die Kurven zeigen Precision, Recall und F_1-Maß für die signifikanten GO Terme, die aus ORAs der Teilmengen mit $k \in \{1, \ldots, n_G\}$ der bestbewerteten Suchtgene resultieren, bezüglich der GO-Terme, die aus der ORA mit der Gesamtmenge der Suchtgene resultierten.

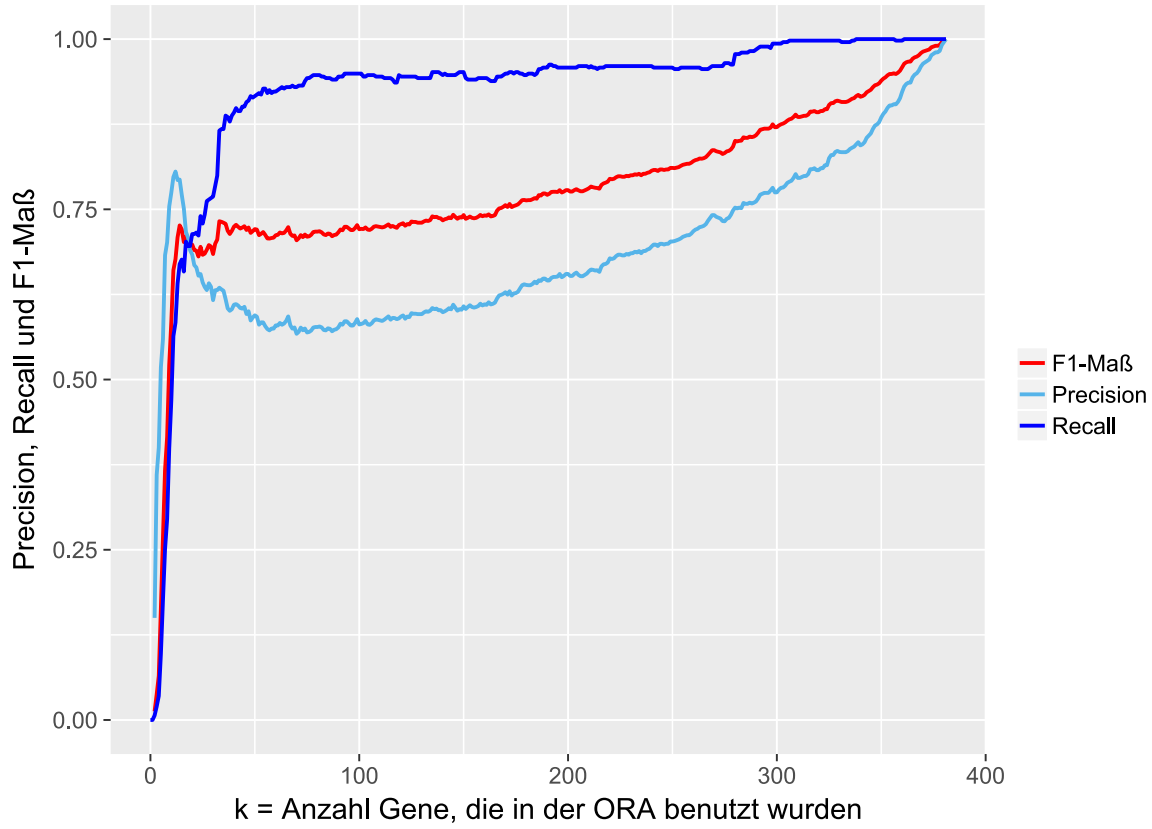
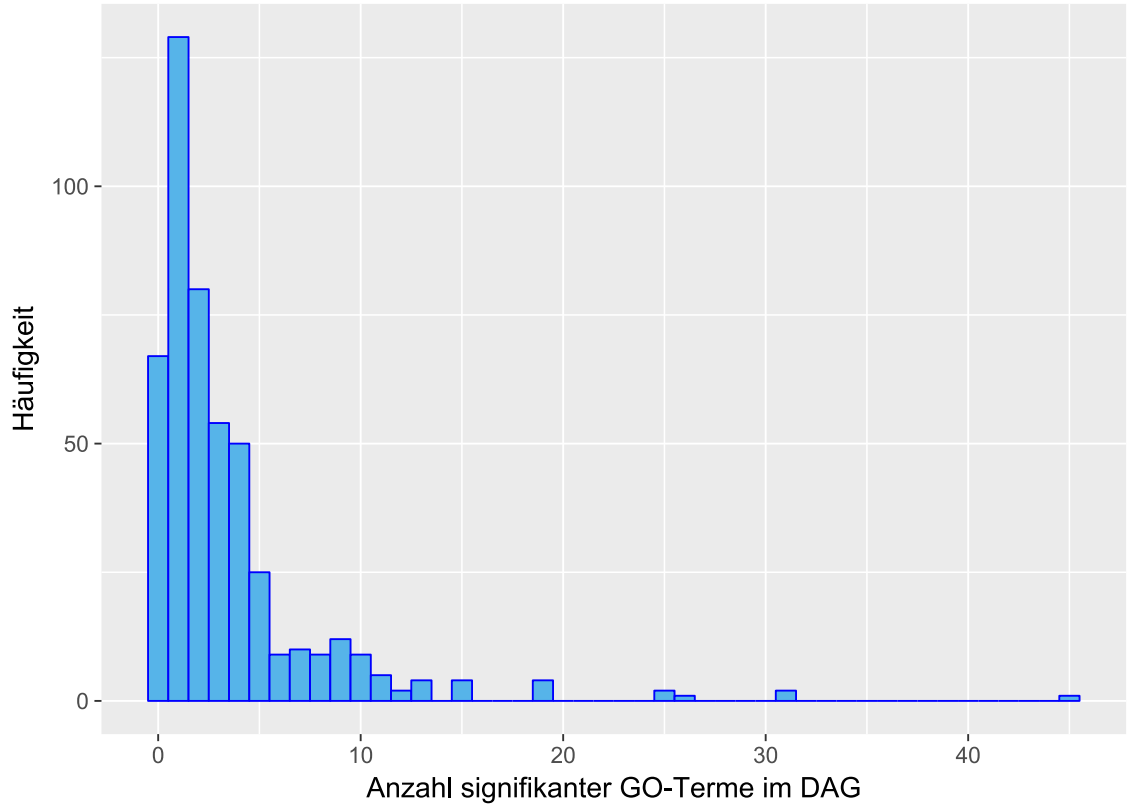


Abbildung B.7: Verteilung der Anzahl von signifikanten GO-Termen, die aus 1 000 ORAs mit zufällig gezogenen Teilmengen von $k^* = 14$ der 387 Suchtgene resultieren; Maximum bei nur etwa 1 GO-Term in der Ontologie der biologischen Prozesse der GO.



Anhang C. Liste der bestbewerteten Schmerzgene.

Tabelle C.1: Liste der $k^* = 29$ mit der funktionserhaltenden, integrativen Genselektion bestbewerteten Schmerzgene.

NCBI#	Gensymbol	Genbeschreibung
7124	*TNF*	Tumor necrosis factor
3569	*IL6*	Interleukin 6
2150	*F2RL1*	F2R like trypsin receptor 1
2149	*F2R*	Coagulation factor II thrombin receptor
3553	*IL1B*	Interleukin 1 beta
1813	*DRD2*	Dopamine receptor D2
348	*APOE*	Apolipoprotein E
7099	*TLR4*	Toll like receptor 4
1236	*CCR7*	C-C motif chemokine receptor 7
4067	*LYN*	LYN proto-oncogene, Src family tyrosine kinase
4988	*OPRM1*	Opioid receptor mu 1
5578	*PRKCA*	Protein kinase C alpha
1139	*CHRNA7*	Cholinergic receptor nicotinic alpha 7 subunit
972	*CD74*	CD74 molecule
150	*ADRA2A*	Adrenoceptor alpha 2A
177	*AGER*	Advanced glycosylation end-product specific receptor
54106	*TLR9*	Toll like receptor 9
729230	*CCR2*	C-C motif chemokine receptor 2
6352	*CCL5*	C-C motif chemokine ligand 5
5594	*MAPK1*	Mitogen-activated protein kinase 1
5027	*P2RX7*	Purinergic receptor P2X 7
958	*CD40*	CD40 molecule
1906	*EDN1*	Endothelin 1
5580	*PRKCD*	Protein kinase C delta
1812	*DRD1*	Dopamine receptor D1
3458	*IFNG*	Interferon gamma
154	*ADRB2*	Adrenoceptor beta 2
920	*CD4*	CD4 molecule
6366	*CCL21*	C-C motif chemokine ligand 21

Literaturverzeichnis

[Acharya et al., 2017] **Acharya, S., Saha, S., & Nikhil, N.**: Unsupervised gene selection using biological knowledge : application in sample clustering, *BMC bioinformatics, Vol. 18*, pp. 513. doi: 10.1186/s12859-017-1933-0, **2017.**

[Alam, 2014] **Alam, S.**: *Biologically-Inspired Techniques for Knowledge Discovery and Data Mining*, IGI Global, **2014**.

[Anhäuser et al., 2001] **Anhäuser, M., Arnheim, D. K., Becker-Follmann, D. J., & Bensel, J.**: Band 6: Flocculus bis Gzhelian-Stufe, In: Sauermost, R. (Ed.), *Lexikon der Biologie: in fünfzehn Bänden* (Vol. 6), Heidelberg, Spektrum Akademischer Verlag, **2001**.

[Applebaum, 1996] **Applebaum, D.**: Probability and Information: An Integrated Approach, (pp. 93-111), Cambridge University Press, **1996**.

[Ashburner et al., 2000] **Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., . . . Eppig, J. T.**: Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium, *Nature genetics, Vol. 25*(1), pp. 25-29. doi: 10.1038/75556, **2000.**

[Baader et al., 2009] **Baader, F., Horrocks, I., & Sattler, U.**: Description Logics, In: Staab, S. & Studer, R. (Eds.), *Handbook on Ontologies*, (pp. 21-43, doi: 10.1007/978-3-540-92673-3_1), Berlin, Heidelberg, Springer Berlin Heidelberg, **2009**.

[Backes et al., 2007] **Backes, C., Keller, A., Kuentzer, J., Kneissl, B., Comtesse, N., Elnakady, Y. A., . . . Lenhof, H.-P.**: GeneTrail—advanced gene set enrichment analysis, *Nucleic acids research, Vol. 35*, pp. W186-W192. doi: 10.1093/nar/gkm323, **2007.**

[Backhaus et al., 2016] **Backhaus, K., Erichson, B., Weiber, R., & Plinke, W.**: Diskriminanzanalyse, In: Backhaus, K., Erichson, B., Plinke, W. & Weiber, R. (Eds.), *Multivariate Analysemethoden: Eine anwendungsorientierte Einführung*, (pp. 215-282, doi: 10.1007/978-3-662-46076-4_5), Berlin, Heidelberg, Springer Berlin Heidelberg, **2016**.

[Baeza-Yates/Ribeiro, 2011] **Baeza-Yates, R., & Ribeiro, B. d. A. N.**: *Modern information retrieval: the concepts and technology behind search*, (second Edition ed.), Harlow u.a., Pearson Education Limited, **2011**.

[Bankhofer/Vogel, 2008] **Bankhofer, U., & Vogel, J.**: Datenanalyse und Statistik: Eine Einführung für Ökonomen im Bachelor, (pp. 261-272), Gabler Verlag, **2008**.

[Bein et al., 2005] **Bein, T., Hanselka, H., & Nuffer, J.**: Adaptronik — ein technischer Ansatz zur Lösung bionischer Aufgaben, In: Rossmann, T. & Tropea, C. (Eds.), *Bionik: Aktuelle Forschungsergebnisse in Natur-, Ingenieur- und Geisteswissenschaft*, (pp. 17-30, doi: 10.1007/3-540-26948-7_2), Berlin, Heidelberg, Springer, **2005**.

[Bellman, 1961] **Bellman, R. E.**: Computational Aspects of Dyamic Programming, *Adaptive Control Processes: A Guided Tour*, (2016 ed., pp. 85-99, doi: 10.1515/9781400874668-007), Princeton, Princeton University Press, **1961**.

[Bennett/Bushel, 2017] **Bennett, B. D., & Bushel, P. R.**: goSTAG: gene ontology subtrees to tag and annotate genes within a set, *Source Code for Biology and Medicine, Vol. 12*(1), pp. 6. doi: 10.1186/s13029-017-0066-1, **2017.**

[Bolla, 2013] **Bolla, M.**: *Spectral Clustering and Biclustering: Learning Large Graphs and Contingency Tables*, Wiley, **2013**.

[Brachman/Levesque, 2004] **Brachman, R., & Levesque, H.**: *Knowledge Representation and Reasoning*, Elsevier Science, **2004**.

[Breiman, 2001] **Breiman, L.**: Random forests, *Machine Learning, Vol. 45*(1), pp. 5-32. **2001.**

[Brewster/O'Hara, 2007] **Brewster, C., & O'Hara, K.**: Knowledge representation with ontologies: Present challenges—Future possibilities, *International Journal of Human-Computer Studies, Vol. 65*(7), pp. 563-568. doi: https://doi.org/10.1016/j.ijhcs.2007.04.003, **2007.**

[Cai et al., 2014] **Cai, H., Ruan, P., Ng, M., & Akutsu, T.**: Feature weight estimation for gene selection: a local hyperlinear learning approach, *BMC bioinformatics, Vol. 15*, pp. 70. doi: 10.1186/1471-2105-15-70, **2014.**

[Ceol et al., 2009] **Ceol, A., Chatr Aryamontri, A., Licata, L., Peluso, D., Briganti, L., Perfetto, L., . . . Cesareni, G.**: MINT, the molecular interaction database: 2009 update, *Nucleic acids research, Vol. 38*(suppl_1), pp. D532-D539. doi: 10.1093/nar/gkp983, **2009.**

[Chowdhury, 2010] **Chowdhury, G. G.**: *Introduction to Modern Information Retrieval*, Facet, **2010**.

[Chuang et al., 2011] **Chuang, L.-Y., Yang, C.-H., Wu, K.-C., & Yang, C.-H.**: A hybrid feature selection method for DNA microarray data, *Computers in Biology and Medicine, Vol. 41*(4), pp. 228-237. doi: https://doi.org/10.1016/j.compbiomed.2011.02.004, **2011.**

[Clocksin/Mellish, 2012] **Clocksin, W. F., & Mellish, C. S.**: *Programming in Prolog: Using the ISO Standard*, Springer Berlin Heidelberg, **2012**.

[Critchlow/van Dam, 2016] **Critchlow, T., & van Dam, K. K.**: Data-Intensive Science, (pp. 1-15), CRC Press, **2016**.

[de Souto et al., 2008] **de Souto, M. C. P., Costa, I. G., de Araujo, D. S. A., Ludermir, T. B., & Schliep, A.**: Clustering cancer gene expression data: a comparative study, *BMC bioinformatics, Vol. 9*, pp. 497-497. doi: 10.1186/1471-2105-9-497, **2008.**

[Dengel, 2011] **Dengel, A.**: Semantische Technologien: Grundlagen. Konzepte. Anwendungen, (pp. 31-33), Spektrum Akademischer Verlag, **2011**.

[Deshpande/Kumar, 2018] **Deshpande, A., & Kumar, M.**: Artificial Intelligence for Big Data: Complete guide to automating Big Data solutions using Artificial Intelligence techniques, (pp. 1-22), Packt Publishing, **2018**.

[Díaz-Uriarte/Alvarez de Andrés, 2006] **Díaz-Uriarte, R., & Alvarez de Andrés, S.**: Gene selection and classification of microarray data using random forest, *BMC bioinformatics, Vol. 7*(1), pp. 3. doi: 10.1186/1471-2105-7-3, **2006.**

[Ding/Peng, 2005] **Ding, C., & Peng, H.**: Minimum redundancy feature selection from microarray gene expression data, *Journal of Bioinformatics and Computational Biology, Vol. 3*(02), pp. 185-205. **2005.**

[Dramiński et al., 2007] **Dramiński, M., Rada-Iglesias, A., Enroth, S., Wadelius, C., Koronacki, J., & Komorowski, J.**: Monte Carlo feature selection for supervised classification, *Bioinformatics, Vol. 24*(1), pp. 110-117. doi: 10.1093/bioinformatics/btm486, **2007.**

[Everitt, 1992] **Everitt, B. S.**: *The Analysis of Contingency Tables, Second Edition*, Taylor & Francis, **1992**.

[Fahrmeir et al., 2016] **Fahrmeir, L., Heumann, C., Künstler, R., Pigeot, I., & Tutz, G.**: *Statistik: Der Weg zur Datenanalyse*, Springer Berlin Heidelberg, **2016**.

[Fang et al., 2014] **Fang, O. H., Mustapha, N., & Sulaiman, M.**: An integrative gene selection with association analysis for microarray data classification, *Intelligent data analysis, Vol. 18*(4), pp. 739-758. **2014.**

[Fayyad et al., 1996] **Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P.**: From data mining to knowledge discovery in databases, *AI magazine, Vol. 17*(3), pp. 37-37. **1996.**

[Fisher, 1935] **Fisher, R. A.**: *The design of experiments*, Oxford, England, Oliver & Boyd, **1935**.

[Forman, 2003] **Forman, G.**: An extensive empirical study of feature selection metrics for text classification, *Journal of Machine Learning Research, Vol. 3*(Mar), pp. 1289-1305. **2003.**

[Forster, 2016] **Forster, O.**: *Analysis 1: Differential- und Integralrechnung einer Veränderlichen*, Springer Fachmedien Wiesbaden, **2016**.

[Friedrich/Pietschmann, 2010] **Friedrich, H., & Pietschmann, F.**: *Numerische Methoden: ein Lehr- und Übungsbuch*, De Gruyter, **2010**.

[Furey et al., 2000] **Furey, T. S., Cristianini, N., Duffy, N., Bednarski, D. W., Schummer, M., & Haussler, D.**: Support vector machine classification and validation of cancer tissue samples using microarray expression data, *Bioinformatics, Vol. 16*(10), pp. 906-914. **2000.**

[Futreal et al., 2004] **Futreal, P. A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., . . . Stratton, M. R.**: A census of human cancer genes, *Nature reviews cancer, Vol. 4*(3), pp. 177. doi: 10.1038/nrc1299, **2004.**

[Gandibleux et al., 2012] **Gandibleux, X., Sevaux, M., Sörensen, K., & T'Kindt, V.**: *Metaheuristics for Multiobjective Optimisation*, Springer Berlin Heidelberg, **2012**.

[Gaudet/Dessimoz, 2017] **Gaudet, P., & Dessimoz, C.**: Gene Ontology: Pitfalls, Biases, and Remedies, In: Dessimoz, C. & Škunca, N. (Eds.), *The Gene Ontology Handbook*, (pp. 189-205, doi: 10.1007/978-1-4939-3743-1_14), New York, NY, Springer New York, **2017**.

[Gaudet et al., 2017] **Gaudet, P., Škunca, N., Hu, J. C., & Dessimoz, C.**: Primer on the Gene Ontology, In: Dessimoz, C. & Škunca, N. (Eds.), *The Gene*

Ontology Handbook, (pp. 25-37, doi: 10.1007/978-1-4939-3743-1_3), New York, NY, Springer New York, **2017**.

[Ghosh/Mitra, 2013] **Ghosh, S., & Mitra, S.**: Clustering large data with uncertainty, *Applied Soft Computing, Vol. 13*(4), pp. 1639-1645. doi: https://doi.org/10.1016/j.asoc.2012.12.036, **2013.**

[Ghosh et al., 2014] **Ghosh, S., Mitra, S., & Dattagupta, R.**: Fuzzy clustering with biological knowledge for gene selection, *Applied Soft Computing, Vol. 16*, pp. 102-111. doi: https://doi.org/10.1016/j.asoc.2013.11.007, **2014.**

[Grasnick et al., 2019] **Grasnick, B., Perscheid, C., & Uflacker, M.**: A Framework for the Automatic Combination and Evaluation of Gene Selection Methods, In: Fdez-Riverola, F., Mohamad, M. S., Rocha, M., De Paz, J. F. & González, P. (eds.), *Proc. Practical Applications of Computational Biology and Bioinformatics, 12th International Conference*, pp. 166-174, Springer International Publishing, Cham, **2019**.

[Graw, 2015a] **Graw, J.**: Verwertung genetischer Informationen, *Genetik*, (pp. 55-108, doi: 10.1007/978-3-662-44817-5_3), Berlin, Heidelberg, Springer Berlin Heidelberg, **2015a**.

[Graw, 2015b] **Graw, J.**: Was ist Genetik?, *Genetik*, (pp. 1-20, doi: 10.1007/978-3-662-44817-5_1), Berlin, Heidelberg, Springer Berlin Heidelberg, **2015b**.

[Gruber, 1993] **Gruber, T. R.**: A translation approach to portable ontology specifications, *Knowledge acquisition, Vol. 5*(2), pp. 199-220. **1993.**

[Guarino et al., 2009] **Guarino, N., Oberle, D., & Staab, S.**: What Is an Ontology?, In: Staab, S. & Studer, R. (Eds.), *Handbook on Ontologies*, (pp. 1-17, doi: 10.1007/978-3-540-92673-3_0), Berlin, Heidelberg, Springer Berlin Heidelberg, **2009**.

[Gunther et al., 2003] **Gunther, E. C., Stone, D. J., Gerwien, R. W., Bento, P., & Heyes, M. P.**: Prediction of clinical drug efficacy by classification of drug-induced genomic expression profiles in vitro, *Proceedings of the National Academy of Sciences of the United States of America, Vol. 100*(16), pp. 9608-9613. doi: 10.1073/pnas.1632587100, **2003.**

[Guyon et al., 2003] **Guyon, I., Andr, #233, & Elisseeff**: An introduction to variable and feature selection, *J. Mach. Learn. Res., Vol. 3*, pp. 1157-1182. **2003.**

[Guyon et al., 2002] **Guyon, I., Weston, J., Barnhill, S., & Vapnik, V.**: Gene Selection for Cancer Classification using Support Vector Machines, *Machine Learning, Vol. 46*(1), pp. 389-422. doi: 10.1023/A:1012487302797, **2002.**

[Hall, 1999] **Hall, M. A.**,*Correlation-based feature selection for machine learning*, (PhD Thesis), Waikato University, New Zealand**, 1999**.

[Hartmann, 2015] **Hartmann, P.**: Graphentheorie, *Mathematik für Informatiker: Ein praxisbezogenes Lehrbuch*, (pp. 263-294, doi: 10.1007/978-3-658-03416-0_11), Wiesbaden, Springer Fachmedien Wiesbaden, **2015**.

[Hastings, 2017] **Hastings, J.**: Primer on Ontologies, In: Dessimoz, C. & Škunca, N. (Eds.), *The Gene Ontology Handbook*, (pp. 3-13, doi: 10.1007/978-1-4939-3743-1_1), New York, NY, Springer New York, **2017**.

[Hira/Gillies, 2015] **Hira, Z. M., & Gillies, D. F.**: A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data, *Advances in Bioinformatics, Vol. 2015*, pp. 198363. doi: 10.1155/2015/198363, **2015.**

[Huan/Lei, 2005] **Huan, L., & Lei, Y.**: Toward integrating feature selection algorithms for classification and clustering, *IEEE Transactions on Knowledge and Data Engineering, Vol. 17*(4), pp. 491-502. doi: 10.1109/TKDE.2005.66, **2005.**

[Huang et al., 2018] **Huang, X., Zhang, L., Wang, B., Li, F., & Zhang, Z.**: Feature clustering based support vector machine recursive feature elimination for gene selection, *Applied Intelligence, Vol. 48*(3), pp. 594-607. doi: 10.1007/s10489-017-0992-2, **2018.**

[Inza et al., 2002] **Inza, I., Sierra, B., Blanco, R., & Larrañaga, P.**: Gene selection by sequential search wrapper approaches in microarray cancer class prediction, *Journal of Intelligent & Fuzzy Systems, Vol. 12*(1), pp. 25-33. **2002.**

[Jaksik et al., 2015] **Jaksik, R., Iwanaszko, M., Rzeszowska-Wolny, J., & Kimmel, M.**: Microarray experiments and factors which affect their reliability, *Biology Direct, Vol. 10*, pp. 46-46. doi: 10.1186/s13062-015-0077-2, **2015.**

[Jerroudi, 2010] **Jerroudi, Z. E.**: *Eine interaktive Vorgehensweise für den Vergleich und die Integration von Ontologien*, Eul, **2010**.

[Jirapech-Umpai/Aitken, 2005] **Jirapech-Umpai, T., & Aitken, S.**: Feature selection and classification for microarray data analysis: Evolutionary methods for identifying predictive genes, *BMC bioinformatics, Vol. 6*(1), pp. 148. doi: 10.1186/1471-2105-6-148, **2005.**

[Jović et al., 2015] **Jović, A., Brkić, K., & Bogunović, N.**: A review of feature selection methods with applications, *Proc. 2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pp. 1200-1205, 25-29 May 2015, **2015**.

[Juran, 1975] **Juran, J. M.**: The non-Pareto principle; mea culpa, *Quality Progress, Vol. 8*(5), pp. 8-9. **1975.**

[Kanehisa/Goto, 2000] **Kanehisa, M., & Goto, S.**: KEGG: kyoto encyclopedia of genes and genomes, *Nucleic acids research, Vol. 28*(1), pp. 27-30. **2000.**

[Kanehisa et al., 2006] **Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K. F., Itoh, M., Kawashima, S., . . . Hirakawa, M.**: From genomics to chemical genomics: new developments in KEGG, *Nucleic acids research, Vol. 34*(Database issue), pp. D354-D357. doi: 10.1093/nar/gkj102, **2006.**

[Karmeshu/Pal, 2003] **Karmeshu, & Pal, N. R.**: Uncertainty, Entropy and Maximum Entropy Principle — An Overview, In: Karmeshu (Ed.), *Entropy Measures, Maximum Entropy Principle and Emerging Applications*, (pp. 1-53, doi: 10.1007/978-3-540-36212-8_1), Berlin, Heidelberg, Springer Berlin Heidelberg, **2003**.

[Kaufman/Rousseeuw, 2009] **Kaufman, L., & Rousseeuw, P. J.**: *Finding Groups in Data: An Introduction to Cluster Analysis*, Wiley, **2009**.

[Khalid et al., 2014] **Khalid, S., Khalil, T., & Nasreen, S.**: A survey of feature selection and feature extraction techniques in machine learning, *Proc. 2014 Science and Information Conference*, pp. 372-378, 27-29 Aug. 2014, **2014**.

[Khatri/Drăghici, 2005] **Khatri, P., & Drăghici, S.**: Ontological analysis of gene expression data: current tools, limitations, and open problems, *Bioinformatics, Vol. 21*(18), pp. 3587-3595. **2005.**

[Kittler, 1975] **Kittler, J.**: Mathematical methods of feature selection in pattern recognition, *International Journal of Man-Machine Studies, Vol. 7*(5), pp. 609-638. **1975.**

[Kononenko, 1994] **Kononenko, I.**: Estimating attributes: Analysis and extensions of RELIEF, In: Bergadano, F. & De Raedt, L. (eds.), *Proc. Machine Learning: ECML-94*, pp. 171-182, Springer Berlin Heidelberg, Berlin, Heidelberg, **1994**.

[Kringel et al., 2018] **Kringel, D., Lippmann, C., Parnham, M. J., Kalso, E., Ultsch, A., & Lötsch, J.**: A machine-learned analysis of human gene polymorphisms modulating persisting pain points to major roles of neuroimmune processes, *European Journal of Pain, Vol. 22*(10), pp. 1735-1756. doi: 10.1002/ejp.1270, **2018.**

[Kringel/Lötsch, 2015] **Kringel, D., & Lötsch, J.**: Pain research funding by the European Union Seventh Framework Programme, *European Journal of Pain, Vol. 19*(5), pp. 595-600. doi: 10.1002/ejp.690, **2015.**

[Kuri-Morales/Aldana-Bobadilla, 2010] **Kuri-Morales, A., & Aldana-Bobadilla, E.**: Finding Irregularly Shaped Clusters Based on Entropy, In: Perner, P. (Ed.), *Proc. Advances in Data Mining. Applications and Theoretical Aspects*, pp. 57-70, Springer Berlin, Heidelberg, **2010**.

[LaCroix-Fralish et al., 2007] **LaCroix-Fralish, M. L., Ledoux, J. B., & Mogil, J. S.**: The Pain Genes Database: An interactive web browser of pain-related transgenic knockout studies, *Pain, Vol. 131*(1-2), pp. 3.e1-3.e4. **2007.**

[Lämmel/Cleve, 2012] **Lämmel, U., & Cleve, J.**: *Künstliche Intelligenz*, Carl Hanser Verlag GmbH & Company KG, **2012**.

[Lassila/McGuinness, 2001] **Lassila, O., & McGuinness, D.**: The role of frame-based representation on the semantic web, *Linköping Electronic Articles in Computer and Information Science, Vol. 6*(5), **2001.**

[Lippmann et al., 2018] **Lippmann, C., Kringel, D., Ultsch, A., & Lötsch, J.**: Computational functional genomics-based approaches in analgesic drug discovery and repurposing, *Pharmacogenomics, Vol. 19*(9), pp. 783-797. **2018.**

[Lippmann et al., 2019] **Lippmann, C., Lötsch, J., & Ultsch, A.**: Computational functional genomics-based reduction of disease-related gene sets to their key components, *Bioinformatics, Vol. 35(14)*, pp. 2362-2370. doi: 10.1093/bioinformatics/bty986, **2019.**

[Liu et al., 2010] **Liu, H., Liu, L., & Zhang, H.**: Ensemble gene selection by grouping for microarray data classification, *Journal of biomedical informatics, Vol. 43*(1), pp. 81-87. doi: https://doi.org/10.1016/j.jbi.2009.08.010, **2010.**

[Liu/Motoda, 2012] **Liu, H., & Motoda, H.**: *Feature Selection for Knowledge Discovery and Data Mining*, Springer US, **2012**.

[Liu et al., 2003] **Liu, T., Liu, S., Chen, Z., & Ma, W.-Y.**: An evaluation on feature selection for text clustering, *Proc. Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, pp. 488-495, **2003**.

[Looney/Hagan, 2015] **Looney, S. W., & Hagan, J. L.**: *Analysis of Biomarker Data: A Practical Guide*, Wiley, **2015**.

[Lötsch, 2019] **Lötsch, J.**: Unterschied zwischen Gen, Genprodukt und Protein, Persönliche Kommunikation (05.03.2019). **2019.**

[Lötsch/Geisslinger, 2011] **Lötsch, J., & Geisslinger, G.**: Pharmacogenetics of new analgesics, *British journal of pharmacology, Vol. 163*(3), pp. 447-460. doi: 10.1111/j.1476-5381.2010.01074.x, **2011.**

[Lötsch et al., 2017] **Lötsch, J., Lippmann, C., Kringel, D., & Ultsch, A.**: Integrated Computational Analysis of Genes Associated with Human Hereditary Insensitivity to Pain. A Drug Repurposing Perspective, *Frontiers in Molecular Neuroscience, Vol. 10*(252), pp. doi: 10.3389/fnmol.2017.00252, **2017.**

[Lötsch/Ultsch, 2016] **Lötsch, J., & Ultsch, A.**: A computational functional genomics based self-limiting self-concentration mechanism of cell specialization as a biological role of jumping genes, *Integrative Biology, Vol. 8*(1), pp. 91-103. doi: 10.1039/C5IB00203F, **2016.**

[Mahajan/Singh, 2016] **Mahajan, S., & Singh, S.**: Review on feature selection approaches using gene expression data, *Imperial Journal of Interdisciplinary Research, Vol. 2*(3), **2016.**

[Matthews, 1975] **Matthews, B. W.**: Comparison of the predicted and observed secondary structure of T4 phage lysozyme, *Biochimica et Biophysica Acta (BBA) - Protein Structure, Vol. 405*(2), pp. 442-451. doi: https://doi.org/10.1016/0005-2795(75)90109-9, **1975.**

[Merkl/Waack, 2013] **Merkl, R., & Waack, S.**: *Bioinformatik Interaktiv: Grundlagen, Algorithmen, Anwendungen*, Wiley, **2013**.

[Mi et al., 2013] **Mi, H., Muruganujan, A., Casagrande, J. T., & Thomas, P. D.**: Large-scale gene function analysis with the PANTHER classification system, *Nature Protocols, Vol. 8*, pp. 1551-1566. doi: 10.1038/nprot.2013.092, **2013.**

[Mitra/Ghosh, 2012] **Mitra, S., & Ghosh, S.**: Feature Selection and Clustering of Gene Expression Profiles Using Biological Knowledge, *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), Vol. 42*(6), pp. 1590-1599. doi: 10.1109/TSMCC.2012.2209416, **2012.**

[Müller-Esterl, 2017] **Müller-Esterl, W.**: *Biochemie: Eine Einführung für Mediziner und Naturwissenschaftler - Unter Mitarbeit von Ulrich Brandt, Oliver Anderka,*

Stefan Kerscher, Stefan Kieß und Katrin Ridinger, Springer Berlin Heidelberg, **2017**.

[Muštra et al., 2012] **Muštra, M., Grgić, M., & Delač, K.**: Breast density classification using multiple feature selection, *automatika, Vol. 53*(4), pp. 362-372. **2012.**

[Naisbitt, 1982] **Naisbitt, J.**: *Megatrends: Ten New Directions Transforming Our Lives*, Warner Books, **1982**.

[Neumann et al., 2017] **Neumann, U., Genze, N., & Heider, D.**: EFS: an ensemble feature selection tool implemented as R-package and web-application, *BioData mining, Vol. 10*, pp. 21. doi: 10.1186/s13040-017-0142-8, **2017.**

[Ng/Han, 2002] **Ng, R. T., & Han, J.**: CLARANS: A Method for Clustering Objects for Spatial Data Mining, *IEEE Trans. on Knowl. and Data Eng., Vol. 14*(5), pp. 1003-1016. doi: 10.1109/tkde.2002.1033770, **2002.**

[Nguyen et al., 2015] **Nguyen, T., Khosravi, A., Creighton, D., & Nahavandi, S.**: A novel aggregate gene selection method for microarray data classification, *Pattern Recognition Letters, Vol. 60-61*, pp. 16-23. doi: https://doi.org/10.1016/j.patrec.2015.03.018, **2015.**

[Ooi/Tan, 2003] **Ooi, C. H., & Tan, P.**: Genetic algorithms applied to multi-class prediction for the analysis of gene expression data, *Bioinformatics, Vol. 19*(1), pp. 37-44. doi: 10.1093/bioinformatics/19.1.37, **2003.**

[Papachristoudis et al., 2010] **Papachristoudis, G., Diplaris, S., & Mitkas, P. A.**: SoFoCles: feature filtering for microarray classification based on gene ontology, *Journal of biomedical informatics, Vol. 43*(1), pp. 1-14. **2010.**

[Paris et al., 2004] **Paris, G., Robilliard, D., & Fonlupt, C.**: Exploring Overfitting in Genetic Programming, In: Liardet, P., Collet, P., Fonlupt, C., Lutton, E. & Schoenauer, M. (eds.), *Proc. Artificial Evolution*, Vol. 2936, pp. 267-277, Springer Berlin, Heidelberg, **2004**.

[Pentreath, 2015] **Pentreath, N.**: *Machine Learning with Spark*, Packt Publishing, **2015**.

[Piegorsch, 2015] **Piegorsch, W. W.**: *Statistical Data Analytics: Foundations for Data Mining, Informatics, and Knowledge Discovery*, Wiley, **2015**.

[Pudil et al., 1994] **Pudil, P., Novovičová, J., & Kittler, J.**: Floating search methods in feature selection, *Pattern Recognition Letters, Vol. 15*(11), pp. 1119-1125. doi: https://doi.org/10.1016/0167-8655(94)90127-9, **1994.**

[Qi/Tang, 2007] **Qi, J., & Tang, J.**: Integrating gene ontology into discriminative powers of genes for feature selection in microarray data, *Proceedings of the 2007 ACM symposium on Applied computing*, pp. 430-434, ACM, Seoul, Korea, **2007.**

[Richards et al., 2012] **Richards, A. J., Schwacke, J. H., Rohrer, B., Cowart, L. A., & Lu, X.**: Revealing functionally coherent subsets using a spectral clustering and an information integration approach, *BMC Systems Biology, Vol. 6*(Suppl 3), pp. S7-S7. doi: 10.1186/1752-0509-6-S3-S7, **2012.**

[Riethmüller, 2012] **Riethmüller, J.**: *Der graue Schwan: Prolegomena zum Wissen der Wissensgesellschaft*, Verlag Wilhelm Fink, **2012**.

[Rousseeuw, 1987] **Rousseeuw, P. J.**: Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, *Journal of Computational and Applied Mathematics, Vol. 20*, pp. 53-65. doi: https://doi.org/10.1016/0377-0427(87)90125-7, **1987.**

[Runkler, 2015a] **Runkler, T. A.**: Clustering, *Data Mining: Modelle und Algorithmen intelligenter Datenanalyse*, (pp. 109-130, doi: 10.1007/978-3-8348-2171-3_9), Wiesbaden, Springer Fachmedien Wiesbaden, **2015a**.

[Runkler, 2015b] **Runkler, T. A.**: Klassifikation, *Data Mining: Modelle und Algorithmen intelligenter Datenanalyse*, (pp. 89-107, doi: 10.1007/978-3-8348-2171-3_8), Wiesbaden, Springer Fachmedien Wiesbaden, **2015b**.

[Saeys et al., 2008] **Saeys, Y., Abeel, T., & Van de Peer, Y.**: Robust feature selection using ensemble feature selection techniques, *Proc. Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 313-325, Springer, **2008**.

[Saeys et al., 2007] **Saeys, Y., Inza, I., & Larrañaga, P.**: A review of feature selection techniques in bioinformatics, *Bioinformatics, Vol. 23*(19), pp. 2507-2517. doi: 10.1093/bioinformatics/btm344, **2007.**

[Salton/McGill, 1984] **Salton, G., & McGill, M. J.**: *Introduction to modern information retrieval*, (2. pr ed.), Auckland u.a., McGraw-Hill, Inc, **1984**.

[Schwemmer, 2008/2009] **Schwemmer, O.**: Vorlesung: Wahrheit und Wissenschaft. *Vorlesungsreihe: Zwei Kulturen oder Einheitswissenschaft? Zum Verhältnis von Natur- und Geisteswissenschaften.*(2008/2009, 01.12.2011). Zugriffsdatum 20.01.2019, 2019, von https://www.philosophie.hu-berlin.de/de/forschung/drittmittelprojekte/ernst-cassirer-nachlassedition/lehre

[Shannon, 1948] **Shannon, C. E.**: A Mathematical Theory of Communication, *Bell System Technical Journal, Vol. 27*(3), pp. 379-423. doi: doi:10.1002/j.1538-7305.1948.tb01338.x, **1948.**

[Singh et al., 2002] **Singh, D., Febbo, P. G., Ross, K., Jackson, D. G., Manola, J., Ladd, C., . . . Sellers, W. R.**: Gene expression correlates of clinical prostate cancer behavior, *Cancer Cell, Vol. 1*(2), pp. 203-209. doi: https://doi.org/10.1016/S1535-6108(02)00030-2, **2002.**

[Sivanandam/Deepa, 2007] **Sivanandam, S. N., & Deepa, S. N.**: *Introduction to Genetic Algorithms*, Springer Berlin Heidelberg, **2007**.

[Solorio-Fernández et al., 2019] **Solorio-Fernández, S., Carrasco-Ochoa, J. A., & Martínez-Trinidad, J. F.**: A review of unsupervised feature selection methods, *Artificial Intelligence Review, Vol. 1*, doi: 10.1007/s10462-019-09682-y, **2019.**

[Spătaru, 2013] **Spătaru, A.**: *Theorie Der Informationsübertragung: Signale und Störungen*, Vieweg+Teubner Verlag, **2013**.

[Stańczyk/Jain, 2017] **Stańczyk, U., & Jain, L. C.**: *Advances in Feature Selection for Data and Pattern Recognition*, Springer International Publishing, **2017**.

[Steland, 2013] **Steland, A.**: *Basiswissen Statistik: Kompaktkurs für Anwender aus Wirtschaft, Informatik und Technik*, Springer Berlin Heidelberg, **2013**.

[Sterling et al., 1994] **Sterling, L., Shapiro, L. S. E. Y., Shapiro, E. Y., coaut, S. E., & Warren, D. H. D.**: *The Art of Prolog: Advanced Programming Techniques*, MIT Press, **1994**.

[Stommel/Wills, 2004] **Stommel, M., & Wills, C.**: *Clinical Research: Concepts and Principles for Advanced Practice Nurses*, Lippincott Williams & Wilkins, **2004**.

[Swets/Weng, 1995] **Swets, D. L., & Weng, J. J.**: Efficient content-based image retrieval using automatic feature selection, *Proc. Computer Vision, 1995. Proceedings., International Symposium on*, pp. 85-90, IEEE, **1995**.

[Tan/Lynch, 2012] **Tan, D., & Lynch, H. T.**: *Principles of Molecular Diagnostics and Personalized Cancer Medicine*, Wolters Kluwer Health, **2012**.

[Tang et al., 2014] **Tang, J., Alelyani, S., & Liu, H.**: Feature selection for classification: A review, *Data Classification: Algorithms and Applications, Vol.*, pp. 37. **2014.**

[Tang et al., 2007] **Tang, Y., Zhang, Y., & Huang, Z.**: Development of Two-Stage SVM-RFE Gene Selection Strategy for Microarray Expression Data Analysis, *IEEE/ACM Transactions on Computational Biology and Bioinformatics, Vol. 4*(3), pp. 365-381. doi: 10.1109/TCBB.2007.70224, **2007.**

[Tarca et al., 2006] **Tarca, A. L., Romero, R., & Draghici, S.**: Analysis of microarray experiments of gene expression profiling, *American journal of obstetrics and gynecology, Vol. 195*(2), pp. 373-388. doi: 10.1016/j.ajog.2006.07.001, **2006.**

[The Gene Ontology Consortium, 2017] **The Gene Ontology Consortium**: Expansion of the Gene Ontology knowledgebase and resources, *Nucleic acids research, Vol. 45*(D1), pp. D331-D338. doi: 10.1093/nar/gkw1108, **2017.**

[The Gene Ontology Consortium, 2019] **The Gene Ontology Consortium**: Guide to GO subsets. (2019). Zugriffsdatum 28.12.2019, 2019, von http://geneontology.org/docs/go-subset-guide/

[The Gene Ontology Consortium, 2018a] **The Gene Ontology Consortium**: GO Annotations. (2018a). Zugriffsdatum 29.11.2018, 2018, von http://www.geneontology.org/page/go-annotations

[The Gene Ontology Consortium, 2018b] **The Gene Ontology Consortium**: Guide to GO Evidence Codes. (2018b). Zugriffsdatum 29.11.2018, 2018, von http://www.geneontology.org/page/guide-go-evidence-codes

[The Gene Ontology Consortium, 2018c] **The Gene Ontology Consortium**: Introduction to the GO resource. (2018c). Zugriffsdatum 29.11.2018, 2018, von http://geneontology.org/page/introduction-go-resource

[The Gene Ontology Consortium, 2018d] **The Gene Ontology Consortium**: Ontology Relations. (2018d). Zugriffsdatum 29.11.2018, 2018, von http://www.geneontology.org/page/ontology-relations

[Thomas, 2017] **Thomas, P. D.**: The Gene Ontology and the Meaning of Biological Function, In: Dessimoz, C. & Škunca, N. (Eds.), *The Gene Ontology Handbook*, (pp. 15-24, doi: 10.1007/978-1-4939-3743-1_2), New York, NY, Springer New York, **2017**.

[Thrun, 2018] **Thrun, M. C.**: *Projection Based Clustering through Self-Organization and Swarm Intelligence*, (Ultsch, A. & Hüllermeier, E. Eds.), Heidelberg, Springer, **2018**.

[Tian et al., 2012] **Tian, L., Ma, L., Kaarela, T., & Li, Z.**: Neuroimmune crosstalk in the central nervous system and its significance for neurological diseases, *Journal of Neuroinflammation, Vol. 9*(1), pp. 155. doi: 10.1186/1742-2094-9-155, **2012.**

[Ultsch, 1987] **Ultsch, A.**,*Control for knowledge-based information retrieval,* (Dissertation), Techn. Wiss. ETH Zürich, Zürich, (8353)**, 1987**.

[Ultsch, 2014] **Ultsch, A.**: Selbstorganisierende Systeme zur Entdeckung ungewöhnlicher Strukturen in Unternehmensdaten, In: Deggendorfer Forum zur digitalen Datenanalyse e.V. (Ed.), *Transparenz durch digitale Datenanalyse: Prüfungsmethoden für Big Data*, (pp. 37-52), Berlin, Erich Schmidt Verlag GmbH & Co, **2014**.

[Ultsch, 2019] **Ultsch, A.**: Abstraktion als Methode wissenschaftlichen Fortschritts, Persönliche Kommunikation (18.01.2019). **2019.**

[Ultsch/Lötsch, 2014a] **Ultsch, A., & Lötsch, J.**: Functional abstraction as a method to discover knowledge in gene ontologies, *PloS one, Vol. 9*(2), pp. e90191. **2014a.**

[Ultsch/Lötsch, 2014b] **Ultsch, A., & Lötsch, J.**: What do all the (human) micro-RNAs do?, *BMC Genomics, Vol. 15*(1), pp. 976. doi: 10.1186/1471-2164-15-976, **2014b.**

[Ultsch/Lötsch, 2015] **Ultsch, A., & Lötsch, J.**: Computed ABC Analysis for Rational Selection of Most Informative Variables in Multivariate Data, *PloS one, Vol. 10*(6), pp. e0129767. doi: 10.1371/journal.pone.0129767, **2015.**

[Uschold/Gruninger, 1996] **Uschold, M., & Gruninger, M.**: Ontologies: Principles, methods and applications, *The knowledge engineering review, Vol. 11*(2), pp. 93-136. **1996.**

[Vafaie/De Jong, 1992] **Vafaie, H., & De Jong, K.**: Genetic algorithms as a tool for feature selection in machine learning, *Proc. Tools with Artificial Intelligence, 1992. TAI'92, Proceedings., Fourth International Conference on*, pp. 200-203, IEEE, **1992**.

[Walz, 2016] **Walz, G.**: *Lexikon der Mathematik: Band 2: Eig bis Inn*, Springer Berlin Heidelberg, **2016**.

[Wang et al., 2001] **Wang, K., Hjelmervik, O. R., & Bremdal, B.**: *Introduction to Knowledge Management: Principles and Practice*, Tapir Academic Press, **2001**.

[Wang et al., 2005] **Wang, Y., Makedon, F. S., Ford, J. C., & Pearlman, J.**: HykGene: a hybrid approach for selecting marker genes for phenotype

classification using microarray gene expression data, *Bioinformatics, Vol. 21*(8), pp. 1530-1537. doi: 10.1093/bioinformatics/bti192, **2005.**

[Wang et al., 2005] **Wang, Y., Tetko, I. V., Hall, M. A., Frank, E., Facius, A., Mayer, K. F. X., & Mewes, H. W.**: Gene selection from microarray data for cancer classification—a machine learning approach, *Computational Biology and Chemistry, Vol. 29*(1), pp. 37-46. doi: https://doi.org/10.1016/j.compbiolchem.2004.11.001, **2005.**

[Wickham, 2016] **Wickham, H.**: *ggplot2: elegant graphics for data analysis*, Springer, **2016**.

[Wu et al., 2009] **Wu, M. C., Zhang, L., Wang, Z., Christiani, D. C., & Lin, X.**: Sparse linear discriminant analysis for simultaneous testing for the significance of a gene set/pathway and gene selection, *Bioinformatics, Vol. 25*(9), pp. 1145-1151. doi: 10.1093/bioinformatics/btp019, **2009.**

[Xiong et al., 2001] **Xiong, M., Fang, X., & Zhao, J.**: Biomarker identification by feature wrappers, *Genome research, Vol. 11*(11), pp. 1878-1887. doi: 10.1101/gr.190001, **2001.**

[Yang et al., 2010] **Yang, C.-H., Chuang, L.-Y., & Yang, C. H.**: IG-GA: a hybrid filter/wrapper method for feature selection of microarray data, *Journal of Medical and Biological Engineering, Vol. 30*(1), pp. 23-28. **2010.**

[Yang et al., 2010] **Yang, P., Zhou, B. B., Zhang, Z., & Zomaya, A. Y.**: A multi-filter enhanced genetic ensemble system for gene selection and sample classification of microarray data, *BMC bioinformatics, Vol. 11*(Suppl 1), pp. S5-S5. doi: 10.1186/1471-2105-11-S1-S5, **2010.**

[Yang et al., 2005] **Yang, Y. H., Xiao, Y., & Segal, M. R.**: Identifying differentially expressed genes from microarray experiments via statistic synthesis, *Bioinformatics, Vol. 21*(7), pp. 1084-1093. doi: 10.1093/bioinformatics/bti108, **2005.**

[Zhai/Massung, 2016] **Zhai, C. X., & Massung, S.**: *Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining*, Association for Computing Machinery and Morgan & Claypool Publishers, **2016**.