HUMBOLDT-UNIVERSITÄT ZU BERLIN

Exposé zur Bachelorarbeit

Analyse von Krankheitsnetzwerken auf Basis einer zuvor vereinheitlichten Krankheitsontologie

Autor: Nico Borgsmüller
Betreuung: Yvonne Mayer
Prof. Dr. Ulf Leser

Ziel dieser Bachelorarbeit ist die Erstellung und Visualisierung einer Krankheitsontologie basierend auf Daten der "Disease Ontology" [1] und die Abbildung von Krankheits-Mutations-Relationen aus den Datenbanken OMIM [2], COSMIC [3],GAD [4] und GeneView [5] auf diese Ontologie. Anschließend soll eine Netzwerkanalyse auf Krankheits-Gen Ebene durchgeführt und mit Daten anderer Paper [6] [7] [8] verglichen werden.

1 Einleitung

Mutations-Krankheits-Relationen können auf verschiedene Arten erhoben werden. Zum einen dadurch, dass wissenschaftliche Mitarbeiter Publikationen lesen und in diesen erwähnte Relationen extrahieren und zu der jeweiligen Datenbank hinzufügen. Dadurch werden neue, validierte (experimentelle) Erkentnisse den Datenbanken hinzugefügt. Dies ist bei den Datenbanken OMIM [2], GAD [4] und teilweise auch bei der Datenbank COSMIC [3] der Fall. Eine andere Möglichkeit ist, dass experimentelle Datensätze automatisch nach Relationen durchsucht werden. So nutzt COSMIC neben der manuellen Kuration auch Daten aus genomweiten Assoziationsstudien von Krebspatienten des "Cancer Genome Project" (CGP). Hierbei werden statistische Zusammenhänge zwischen Mutationen und Krankheiten in die Datenbank aufgenommen, die allerdings nicht experimentell validiert worden sind. Dabei werden nur mit Krebs assoziierte Mutationen und Gene aufgenommen. GeneView benutzt Machine Learning Algorithmen um in Publikationen aus der Datenbank Pubmed [9] Mutations-Krankheits-Relationen zu finden (Text-Mining). Dadurch werden (experimentell) validierte Relationen gefunden. Allerdings können mittels GeneView auch Relationen gefunden werden, die in noch keiner anderen Datenbank enthalten sind, da sie noch nicht manuell kuriert worden sind. In dieser Arbeit verwenden wir GeneView so, dass wir von einer Relation ausgehen, wenn eine Mutation und eine Krankheit in demselben Satz erwähnt werden.

Da es in Publikationen keine einheitliche Krankheits- und Mutationsnomenklatur gibt müssen die Daten aus Publikationen erst standardisiert werden. In GeneView werden die Krankheiten auf UMLS-IDs, SNPs auf DBSNP-IDs und Gene mittels GNAT [10] auf Entrez Gene IDs abgebildet. In OMIM findet eine Abbildung auf den MIM Code statt. In GAD werden Gene auf Basis der HGNC Richtlinien [11] und Krankheiten an Hand der MeSH-Struktur [12] standardisiert. In COSMIC werden die Daten an ein eigenen Klassifikationssystems [2], welches speziell für die Datenbank entwickelt wurde, angepasst.
Möchte man aber mit Daten aus verschiedenen Datenbanken arbeiten, stellen die verschiedenen Nomenklaturen ein Problem dar. Nicht nur bei Krankheitsnamen wie z.B. "breast cancer" , welcher auch unter "breast carcinoma" oder "malignant breast melanoma" mit drei verschiedenen UMLS-IDs geführt wird, gibt es Mehrdeutigkeiten. Auch wird im UMLS die "Ebene" der Krankheitsklassen nicht beachtet. So werden z.B. "breast cancer", "cancer" und "male breast cancer" als verschiedene, gleichwertige Krankheiten angeführt, obwohl "male breast cancer" eine Unterklasse von "breast cancer" ist, welches wiederum eine Unterklasse von "cancer" ist.

"Disease Ontology" [1] (DO) ist eine Ontologie für menschliche Krankheiten welche neben einer vereinheitlichten Nomenklatur auch eine hierarchische Klassenunterteilung der Krankheiten und deren Beziehung zueinander beinhaltet. Über eine Abbildung von Datenbank-spezifischen Krankheitstermen auf das DO-Konzept ist es möglich, Daten aus verschiedenen Datenbanken zu vereinheitlichen und so in zukünftigen Forschungen datanbankübergreifend zu arbeiten.

Mit Hilfe von derart normalisierten Mutations-Relations-Paare ist es möglich ein datenbankübergreifendes Krankheitsnetzwerk zu erstellen. In einem entsprechenden Netzwerkgraphen sind die Krankheiten Knoten und zwei Krankheiten sind mittels einer Kante verbunden, falls sie mit der selben Mutation assoziiert werden. In einer ersten Analyse der Daten im Rahmen meines Studienprojekts "Auswertung von Krankheits-Mutations/Gen-Relationen in GeneView" hat sich aber herausgestellt, dass in den zu Grund liegenden Daten (GeneView, GAD,GWAS und COSMIC) ein Großteil der Mutationen nur mit einer Krankheiten assoziiert wird. Da diese Mutationen in einem Netzwerkgraphen nicht auftauchen würden, da sie ja auf Grund nur einer Assoziation keine Kanten bilden, bietet sich für eine Netzwerkanalyse stattdessen eine Betrachtung der Krankheits-Gen-Relationen an. Dafür werden Mutationen, die auf einem Gen liegen, zusammengefasst. Dadurch ist davon auszugehen, dass so wesentlich mehr Gene mit mehr als einer Krankheit assoziiert werden als dies bei Mutations-Relationen der Fall ist. Des weiteren spricht für eine Analyse auf Gen-Ebene, dass verschiedene Mutationen auf demselben Gen ohnehin Auswirkungen auf dasselbe Gen-Produkt haben. Außerdem setzten therapeutische Ansätze fast immer am Gen-Produkt an, nicht bei den zu Grunde liegenden Mutationen.
In einem entsprechenden Netzwerkgraphen sind die Krankheiten weiterhin die Knoten, allerdings werden sie mittels einer Kante verbunden, falls sie mit dem selben Gen assoziiert werden. Diese Art von Netzwerk wurde als erstes in [6] vorgestellt und "Human Disease Network" (HDN) genannt. Eine Visualisierung des HDN stellt Abbildung 1 auf Seite 3 dar.

Auf Basis von Krankheits- und Gennetzwerken lassen sich vielfältige Rückschlüsse über die zu Grunde liegende Mechanismen und Wechselwirkungen der einzelnen Netzwerk-

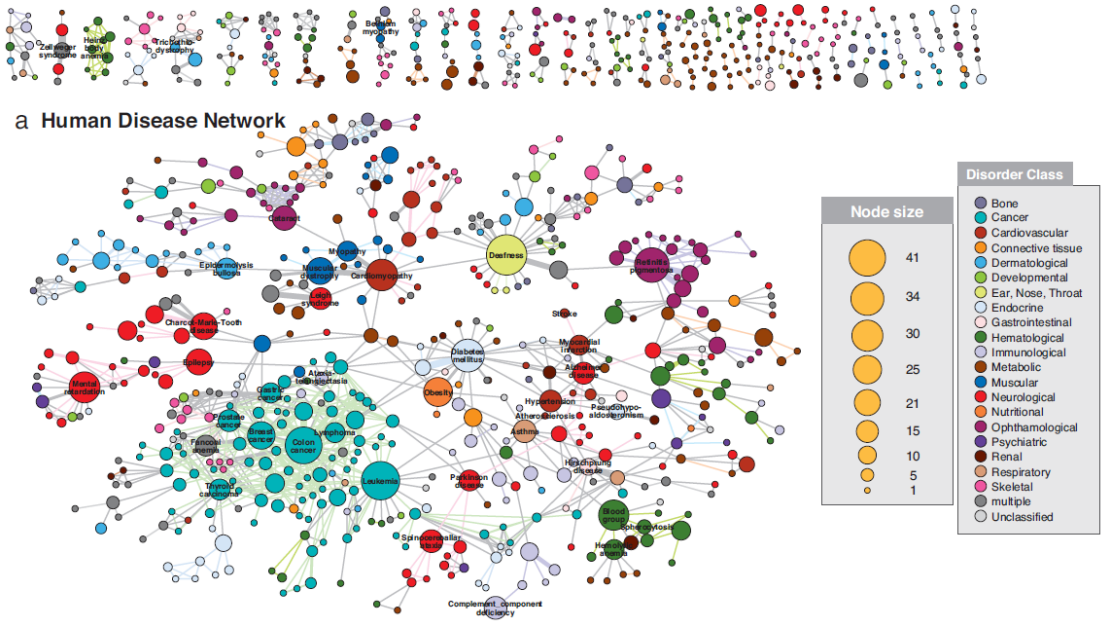

Abbildung 1: Human Disease Network. Jeder Knoten stellt eine Krankheit dar, eingefärbt nach der zugehörigen Krankheitsklasse. Kanten zwischen Knoten der gleichen Krankheitsklasse sind entsprechend der Krankheitsklasse eingefärbt, Kanten zwischen Krankheiten verschiedener Klassen sind grau dargestellt. Die Knotengröße ist proportional zu der Anzahl der mit der Krankheit assoziierten Gene. Die Kantengröße ist proportional zu der Anzahl an gemeinsam assoziierten Genen. Bild entnommen aus [6]

.

Komponenten ziehen wie z.B. die Existenz von krankheitsspezifischen, funktionellen Modulen auf Gen-Ebene ([13]). So zeigen Gene, die mit der selben Krankheit assoziiert sind, ähnliche Genexpressionsprofile ([14]) und ihre Produkte interagieren mit einer größeren Wahrscheinlichkeit miteinander ([15],[16]).

2 Ziele

Ziel dieser Bachelorarbeit ist zum einen die Erstellung und Visualisierung einer Krankheitsontologie basierend auf Daten der “Disease Ontology“ [1]. Die Ontologie soll es ermöglichen die Krankheits-Gen-Relationen aus verschiedenen Quellen (öffentliche Datenbanken, Text-Mining) auf eine einheitliche Nomenklatur abzubilden und so datenbankübergreifend zu arbeiten.
Auf Basis der so standardisierten Krankheits-Gen-Relationen sollen drei Netzwerke erstellt werden. Eines basiert auf Krankheits-Gen-Relationen aus den öffentlichen Datenbanken OMIM [2], COSMIC [3] und GAD [4], ein weiteres aus Text-Mining Daten aus Gene-View [5] und ein drittes kombiniert beide Datensätze. Die Netzwerke werden dann

analysiert und miteinander verglichen. Hierbei soll die Topologie der Netzwerke sowie die Bildung von Subclustern untersucht werden. Anschließend sollen die Netzwerke mit Ergebnissen aus [6] [7] , deren Datenbasis nur OMIM war, und [8] verglichen werden, um eine Aussage darüber treffen zu können, wie sich die Hinzunahme weiterer (und aktuellerer) Daten auswirkt.

3 Herangehensweise

Ausgangspunkt der Arbeit wird eine Datei mit allen Relationen der Datenbanken OMIM [2], COSMIC [3] und GAD [4] sowie eine Datei mit durch GeneView [5] gewonnenen Relationen sein. Die Datei beinhaltet für jede Relation folgende Daten: Mutation, Gen, Krankheit und Quelldatenbank. Dabei sind bereits Mutationen zu dbSNP-IDs, Gene zu Entrez Gene IDs und Krankheiten zu UMLS-IDs standardisiert. In den UMLS-IDs stehen aber teilweise verschiedene IDs für dieselbe Krankheit, so wird z.B. zwischen "breast cancer" (UMLS-IDs: C0006142; C0153555; C1458155), "breast carcinoma" (UMLS-ID: C0678222) und "malignant breast melanoma"(UMLS:C0346787) unterschieden. Außerdem geben die UMLS-IDs, wie bereits in der Einleitung erwähnte, keinen Aufschluss über die hierachische Ordnung der Krankheiten untereinander.

Daher sollen die UMLS-Terme, soweit möglich, auf die Daten der "disease ontology" [1] abgebildet werden. Hierzu werden die Daten der "disease ontology" im .obo-Format heruntergeladen und in Python mit Hilfe des Pakets "networkx" [17] visualisiert. In der .obo Datei ist für jede enthaltene Krankheit ihre zugehörigen DOID-ID aufgeführt, welche UMLS-ID/s der DOID-ID entsprechen und welche Krankheit die hierarchisch nächsthöhere Krankheit ist. Dabei sollen zum einen die Mehrfachnennungen der UMLS-IDs aufgelöst werden. So sollen z.B. allen Relationen mit den UMLS-IDs für "Lipoma face NOS", "Cutaneous Lipoma" und "Cutaneous Lipomatous Neoplasm" auf die Krankheit "skin lipoma" mit der DOID-ID 10188 abgebildet werden. Krankheiten, für die keine Abbildung auf entsprechende DOID-IDs stattfinden kann, da sie in der DO nicht beinhaltet sind, sollen weiterhin mit ihrer UMLS-IDs geführt werden. Bei Krankheiten, die auf eine DOID-ID abgebildet werden können, soll dies auf die tiefste hierarchische Ebene geschehen. Das heißt, wenn z.B. ein Gen mit der Krankheit "amyloidosis" assoziiert wird, dann soll dieses Gen auf alle Krankheiten, die hierarchisch auf der tiefsten Ebene unter "amyloidosis" liegen, abgebildet werden. Dies wären "paramyloidosis", "cerebral amyloid angiopathy", "transthyretin amyloidosis", "familial visceral amyloidosis", "primary cutaneous amyloidosis" und "finnish type amyloidosis". Für Relationen mit Krankheiten oberhalb der tiefsten Ebene wurde oder konnte die genaue Art der Krankheit experimentell nicht bestimmt werden. Daher wird die zugehörige Mutation auf alle Unterklassen der Krankheit, die bestimmt wurde, abgebildet. Dadurch soll verhindert werden, dass spezielle Krankheiten mit allgemeinenen Krankheitsklassen verglichen werden.

Anschließend soll auf Basis der öffentlichen Daten aus OMIM, COSMIC und GAD ein Krankheitsnetzwerk erstellt werden. Zum Vergleich wird auch ein Netzwerk auf Basis der Daten aus GeneView sowie ein Krankheitsnetz auf Basis beider Datensätze (öffentlich Daten und Text-Mining Daten) erstellt. Die Umsetzung der Netzwerkgraphen soll mit

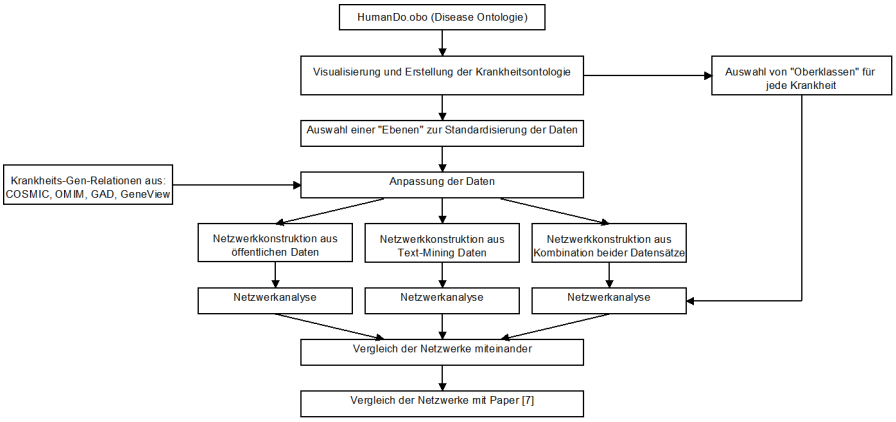

Abbildung 2: Herangehensweise als Flussdiagramms visualisiert

Python und dem Paket “networkx“ [17] durchgeführt werden. Bei den erstellten Netzwerkgraphen soll die “degree distribution“ und Bildung von Subclustern untersucht werden. Dafür werden auf Basis der erstellten Ontologie “Oberklassen“ auf einer hierarchischen Ebene ausgewählt. Diese befindet sich relativ weit “oben“ in der Ontologie und soll beschreiben zu welcher Art von Krankheitsklasse jede Krankheit gehört. Es könnte z.B. als Oberklasse “cancer“ gewählt werden, zu der dann alle Krankheiten zählen, welche in der Ontologie aus “cancer“ hervorgehen, wie z.B. “breast cancer“. Krankheiten, die nicht in der DO enthalten sind und denen daher auch keine Oberklasse zugeordnet werden kann, sollen der Krankheitsklasse “unclassified“ zugeordnet werden. Es ist zu erwarten, dass Krankheiten der Oberklassen (z.B. “cancer“) ein Subcluster bilden und von Krankheiten anderer Oberklassen (z.B. “neurological“) zu unterscheiden sind. Darüber hinaus soll untersucht werden, wie sich die “degree distribution“ und die Bildung von Subclustern in den Netzwerken basierend auf verschiedenen Datensätzen auswirkt. Die so gewonnenen Erkenntnisse sollen anschließend mit den Ergebnissen aus [6] [7] und [8] verglichen werden.
Die Arbeitsschritte, die im Rahmen der Bachelorarbeit vorgesehen sind, sind in Abbildung 2 dargestellt.

Literatur

[1] Schriml LM, Arze C, Nadendla S, Chang YW, Mazaitis M, Felix V, Feng G, Kibbe WA. Disease Ontology: a backbone for disease semantic integration. Nucleic Acids Res. 40 (2012).

[2] Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res. 33 (2004).

[3] Forbes SA, Tang G, Bindal N, et al. COSMIC (the Catalogue of Somatic Mutations

in Cancer): a resource to investigate acquired mutations in human cancer. Nucleic Acids Res. 38 (2010).

[4] Becker KG ,Barnes KC, Bright TJ, Wang SA. The Genetic Association Database. Nature Genetics 36 (2004).

[5] Thomas P. et al. GeneView: a comprehensive semantic search engine for PubMed. Nucleic Acids Res. Vol 40 (2012).

[6] Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabási AL. The human disease network. PNAS Vol 104 No. 21 (2007).

[7] Goh KI, Choi IG. Exploring the human diseasome: the human disease network. Briefings in Functional Genomics Vol 11 (2012).

[8] Barabási AL, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human diseases. Nature Vol 12 (2011).

[9] http://www.ncbi.nlm.nih.gov/pubmed

[10] Hakenberg J, Gerner M, Haeussler M, Solt I, Plake C, Schroeder M, Gonzalez G, Nenadic G and Bergman CM. The GNAT library for local and remote gene mention normalization. Bioinformatics, 27 (2011).

[11] Seal RL, Gordon SM, Lush MJ, Wright MW, Bruford EA. Genenames.org: the HGNC resources in 2011. Nucleic Acids Res. 39 (2011).

[12] Rogers FB. Medical subject headings. Bull Med Libr Assoc. 51 (Jan 1963).

[13] Suthram S, Dudley JT, Chiang AP, Chen R, Hastie JH, Butte AJ. Network-Based Elucidation of Human Disease Similarities Reveals Common Functional Modules Enriched for Pluripotent Drug Targets. PLoS Computational Biology Vol 6 (2010).

[14] Ge H, Walhout AJ, Vidal M. Integrating 'omic' information: A Bridge between genomics and system biology. Trends Genet 19 (2003).

[15] Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabási AL. Hierarchical organization of modularity in metabolic networks. Science 297 (2002).

[16] Hartwell LH, Hopfield JJ, Leibler S, Murray AW. From molecular to modular cell biology. Nature Vol 402 (1999).

[17] http://networkx.github.io/