

FERNFACHHOCHSCHULE SCHWEIZ

Semesterarbeit

**Strukturelle Untersuchungen der
Genetischen Netzwerke von Alzheimer
und Schizophrenie**

Network-Analysis

Authoren:

Patrick Bernhardsgrüter
Luca Hostettler

Eingereicht bei:

Ao. Prof. Dr. habil. Matthias Dehmer

Abstract

Der menschliche Organismus besteht aus einer Vielzahl an Netzwerken. Das Zusammenspiel von Genen in unseren Zellen wird heute noch aktiv erforscht. Es gibt Studien, welche sich dem Ziel annehmen, gewisse Ähnlichkeiten und Synergien genetischer Krankheiten zu erforschen. Genetische Krankheiten werden in der Regel vererbt und befinden sich bereits bei der Geburt im Menschen.

Diese Arbeit entstand im Rahmen des Moduls **Network Analysis** der **Fernfachhochschule Schweiz** unter der Leitung von **Ao. Prof. Dr. habil. Dehmer**. Es werden in dieser Arbeit keine medizinischen Aussagen getroffen. Der zentrale Teil der Arbeit behandelt die Analyse von grossen Netzwerken und das Vergleichen von lokalen und globalen Messwerten der Netzwerke.

In dieser Arbeit werden Beziehungen zwischen den relevanten Genen von Alzheimer und Schizophrenie analysiert und miteinander verglichen. Dabei wird zuerst eine Datenerhebung der relevanten und verwandten Gene der Krankheiten dokumentiert.

Als nächstes werden die Gene auf ihre Nachbarn, Gen-Fusionen oder gemeinschaftliche Experimente untersucht und daraus bilden sich die ersten Netzwerke. Die Netzwerke werden anhand von verschiedenen Massen untersucht und miteinander verglichen.

Besonders bei der Bildung von Communities und den Zentralitätsmassen können wir bei den betroffenen Genen eine Ähnlichkeit feststellen.

Inhaltsverzeichnis

1 Einleitung	1
1.1 Genetische Krankheiten	1
2 Ausgangslage	2
2.1 Fragestellung	2
3 Theoretische Grundlagen	3
3.1 Grundlagen der Graphentheorie	3
3.1.1 Knoten (Nodes, Vertices)	3
3.1.2 Kanten (Edges)	3
3.1.3 Adjazenz und Inzidenz	3
3.1.4 Isomorphie von Graphen	4
3.1.5 Grad eines Knotens	4
3.1.6 Adjazenzmatrix	4
3.2 Communities und Module	4
3.2.1 Anzahl Edges	5
3.2.2 Anzahl Nodes	5
3.2.3 Degree	5
3.2.4 Density	5
3.2.5 Betweenness Centrality	6
3.2.6 Closeness Centrality	6
3.2.7 Local Clustering Coefficient	6
3.2.8 Average Clustering Coefficient [global]	7
3.2.9 Wiener Index [global]	7
4 Krankheiten	8
4.1 Alzheimer	8
4.2 Schizophrenie	8
4.3 Bestehende Literatur	9
4.3.1 Barabasi	9
4.3.2 Kwang-Il Goh	9
4.3.3 Genes And Diseases	9
5 Netzwerk Analyse	10
5.1 Beschaffung der Daten	10
5.1.1 String-DB Kanten Score	10
5.1.2 String-DB Gen	11
5.1.3 Daten Aufbereitung	12
5.2 Analyse der Netzwerke	13
5.2.1 Jupyter Notebook mit pandas und networkx	13
5.2.2 Gesamt Netzwerk Analyse mit graph-tool	14
6 Ergebnisse	17
6.1 Netzwerk Kennzahlen	17
6.2 Darstellung Gesamtes Netzwerk	17
6.3 Darstellung Gen-Interaktionen bei Alzheimer	19
6.4 Darstellung Gen-Interaktionen bei Schizophrenie	20
6.5 Vergleich der Messwerte einzelner Gene (Alzheimer vs. Schizophrenie)	21
6.6 Messwerte relevanter Gene im gesamten Netzwerk gemessen	23
6.7 Messwerte relevanter Gene aus gefilterten Netzwerken gemessen	25
6.8 Messwerte relevanter Gene normiert	26
7 Diskussion	27
7.1 Auswertung der Netzwerk-Kennzahlen	27
7.2 Auswertung der einzelnen Gene	27
7.3 Ausblick	28
8 Anhang	29

1 Einleitung

Eine genetische Krankheit deutet auf eine Mutation oder einen Defekt der Gene hin. Dies kann auftreten, wenn ein Teil der DNA sich gegenüber der normalen DNA des Gens unterscheidet, was einen Fehler im Code darstellt. Normalerweise werden diese schon bei der Geburt dem Kind übergeben und es trägt diese Defekte ein Leben lang in sich. Es gibt wenige Fälle, in welchen Genmutation, sogenannte Reduktionen oder Multiplikationen auch später passieren können [1].

Ein Gen befindet sich an einer gewissen Position in unserer DNA. Die DNA wiederum ist Teil eines von 23 Paaren von Chromosomen, welche sich im Nukleus einer Zelle befinden. Die verschiedenen Chromosome beinhalten fast alle über 1000 Gene.

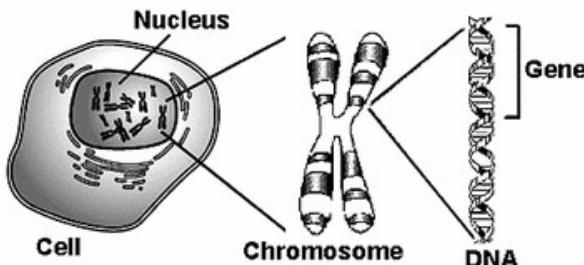


Abbildung 1: Zelle Chromoson DNA

Unsere DNA wird in 4 Basis-Bausteine aufgeteilt und mit diesen werden alle Gene erstellt. Es sind:

- Adenine
- Guanine
- Cytosine
- Thymine

In der Kodierung der Gene sehen wir nur die Kombinationen der ersten Buchstaben: "AGC, AGT, AAT...".

In unserem Körper befinden sich über 20'000 verschiedene Gene [2]. Die meisten Gene besitzen Baupläne um funktionale Moleküle (Proteine) herzustellen, welche der Körper benötigt. Es gibt ebenfalls Gene, welche keine dieser Baupläne besitzen (Non-Coding Gen), deren Aufgabe wurde bis heute noch nicht genau geklärt [1]. In der Literatur werden sie auch als *Junk DNA* bezeichnet.

Wir gehen davon aus, dass über 99% der Gene in allen Menschen gleich sind und dass weniger als 1% der verschiedenen Gene uns zu diesem einzigartigen Individuum machen, dass uns ausmacht.

Die Chromosome, DNA und Gene werden oft auch als das "Buch des Lebens" vorgestellt. Dabei erhält ein Kind von den beiden Elternteilen jeweils eine Hälfte und hat neu sein eigenes Buch des Lebens mit den nötigen Bauplänen.

1.1 Genetische Krankheiten

Bei den genetischen Krankheiten handelt es sich um Mutationen / Veränderungen in der DNA Struktur, was dazu führen kann, dass ein für den Körper notwendiges Protein nicht mehr genügend oder gar nicht mehr produziert werden kann.

Es gibt komplexe genetische Krankheiten, wie psychische Krankheiten (mental illnesses) welche oftmals nicht nur auf ein fehlerhaftes Gen zurückzuschliessen sind, sondern eine ganze Reihe von Genen betreffen.

Es gibt viele Studien, welche sich mit den Genmutationen von kranken Organismen beschäftigen. Dabei werden die Kodierungen und Zusammenhänge analysiert und festgehalten. Ein Sammlung dieser Studien befindet sich auf dem Netzwerk der NCBI (National Center For Biotechnology Information)¹ welche wiederum Links zu weiteren medizinischen Datenbanken wie PubMed² oder MedGen³ hat.

¹<https://www.ncbi.nlm.nih.gov/>

²<https://pubmed.ncbi.nlm.nih.gov/>

³<https://www.ncbi.nlm.nih.gov/medgen>

2 Ausgangslage

Grundlage für diese Arbeit sind die betroffenen Gene der beiden Krankheiten Alzheimer und Schizophrenie der NCBI Webseite. Dazu werden alle Gene des homo sapiens Genom von der String-DB heruntergeladen.

Im Theorie-Teil der Arbeit werden die Formeln und mathematischen Grundlagen für die Netzwerk Analyse nochmals repetiert und vorgestellt.

In Kapitel 4 werden die Krankheiten genauer vorgestellt und die relevanten und verwandten Gene aufgelistet. Eine detaillierte Auflistung und Beschreibung der Gene befindet sich im Anhang.

Danach werden die Netzwerke mit Python analysiert; dazu müssen die Daten zuerst aufbereitet werden um danach verschiedene Masse zu berechnen. Im Anhang der Arbeit befindet sich ebenfalls der gesamte Code, welcher im Rahmen dieser Arbeit entstanden ist.

Zum Schluss folgen die Erkenntnisse und Resultate sowie die Diskussion der relevanten Stellen der Arbeit.

2.1 Fragestellung

Insgesamt sollen folgende Fragen beantwortet werden:

Wie sehen Netzwerke genetischer Krankheiten aus?

Welche strukturellen Ähnlichkeiten besitzen die Gen-Netzwerke von Schizophrenie und Alzheimer?

3 Theoretische Grundlagen

In der Netzwerkanalyse werden Charakteristiken von Graphen bevorzugt auf numerische Werte abgebildet, um eine Kategorisierung und Vergleichbarkeit zu erhalten. Dazu sind unterschiedliche Masse notwendig. Es wird grundsätzlich zwischen lokalen und globalen Massen unterschieden. Die lokalen Masse beziehen sich immer auf einen spezifischen Knoten im Netzwerk, wohingegen die globalen Masse das gesamte Netzwerk auf einen Wert abbilden.

3.1 Grundlagen der Graphentheorie

Graph

Ein Graph oder ein Netzwerk G ist ein Paar von Mengen

$$G = (V, E).$$

Dabei ist V eine Menge mit beliebig vielen Elementen, den sogenannten Knoten (engl. Vertices) von G . Mit E wird die Menge aller Kanten (engl. Edges) bezeichnet. Eine Kante verbindet zwei (im Allgemeinen unterschiedliche) Knoten miteinander. Die hier gewählten Bezeichnungen sind in der Literatur so üblich; sie sind Abkürzungen der entsprechenden englischen Wörter: vertex für Knoten und edge für Kante. Man beachte, dass diese Definition eines Graphen auch den Fall eines oder mehrerer alleinstehender Knoten umfasst, wohingegen Kanten ohne Knoten nicht möglich sind: Es gibt kein alleinstehendes Kantenende und keine alleinstehende Kante [3].

Es können dabei auch mehrere Kanten zwischen zwei Knoten existieren. Oft wird für solche Graphen der Begriff *Multigraph* verwendet. Es besteht auch die Möglichkeit, dass sich ein Knoten selbst referenziert. Generell kann bei Graphen zwischen gerichtet und ungerichtet unterschieden werden.

3.1.1 Knoten (Nodes, Vertices)

Als Knoten oder Ecke bezeichnet man in der Graphentheorie ein Element der Knotenmenge eines Graphen. Knoten können verschiedene Attribute und Informationen besitzen und sind alleinstehend, sowie auch in Verbindung mit Kanten in einem Graph aufzufinden.

3.1.2 Kanten (Edges)

Als Kante wird die Verbindung zwischen zwei Knoten bezeichnet. Diese kann, ebenfalls wie die Knoten verschiedene Attribute bzw. eine Gewichtung haben. Weiter kann eine Kante entweder ungerichtet oder gerichtet sein. Ist sie gerichtet, so spielt die Notation der Verbindung eine grosse Rolle. Wird beispielsweise (a,b) geschrieben, bedeutet dies, dass die Verbindung vom Knoten a zum Knoten b läuft, aber nicht umgekehrt.

3.1.3 Adjazenz und Inzidenz

Zwei Knoten, die durch eine Kante verbunden sind, oder zwei Kanten, die einen gemeinsamen Knoten besitzen, nennt man benachbart oder adjazent. Gehört ein Knoten zu einer Kante, so nennt man die beiden inzident.

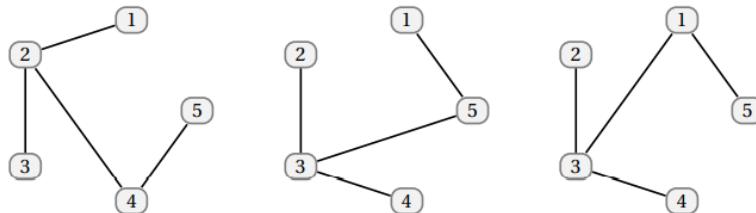


Abbildung 2: Unterschiedliche Graphdarstellungen [3]

Graphen besitzen oft eine Bezeichnung ihrer Knoten und Kanten. Ohne diese wären die in Abbildung 2 gleich. Der linke Graph in Abbildung 2 hat beispielsweise folgende Bezeichnung:

$$V = \{1, 2, 3, 4, 5\} \text{ und } E = \{\{1, 2\}, \{2, 3\}, \{2, 4\}, \{4, 5\}\}$$

3.1.4 Isomorphie von Graphen

Haben zwei Graphen G und G' die gleiche Anzahl von Knoten und gibt es darüber hinaus eine eindeutige Zuordnung der Knoten von G und G' , gemäß der die Kanten von G den Kanten von G' entsprechen, so nennt man die beiden Graphen isomorph und schreibt in diesem Fall $G \sim G'$ [3].

3.1.5 Grad eines Knotens

Der Grad eines Knotens sagt aus, mit wie vielen Kanten dieser verbunden ist. Bei einem ungerichteten Graphen werden alle Verbindungen eines Knotens, bei einem gerichteten Graphen die jeweils ein- bzw. ausgehenden Verbindungen zur Ermittlung des Knotengrades gezählt. Der Knoten Nummer 4 des linken Graphen auf Abbildung 2 hat daher den Grad 3.

Der Grad eines Knotens ist folgendermassen definiert:

Es sei $G = (V, E)$ ein Graph. Für jeden Knoten $v \in V$ definieren wir den Grad von v als die Anzahl der von v ausgehenden Kanten und schreiben dafür $d(v)$:

$$d(v) = |\{ \{v, w\} | \{v, w\} \in E \}|.$$

Ein Graph, bei dem alle Knoten den konstanten Grad k haben, heisst k -regulär. Einen Knoten vom Grad 0 nennt man isoliert[3].

3.1.6 Adjazenzmatrix

Eine vollständige Beschreibung eines Netzes setzt voraus, dass seine Verbindungen im Auge behalten werden. Der einfachste Weg, dies zu erreichen, ist die Erstellung einer vollständigen Liste der Verbindungen. Für mathematische Zwecke wird ein Netzwerk oft durch seine Adjazenzmatrix dargestellt. Die Adjazenzmatrix eines gerichteten Netzes mit N Knoten hat N Zeilen und N Spalten, deren Elemente sind:

$A_{ij} = 1$, falls eine Verbindung vom Knoten i zum Knoten j besteht.

$A_{ij} = 0$, falls keine Verbindung vom Knoten i zum Knoten j besteht.

Wird eine Adjazenzmatrix eines gerichteten Graphen generiert werden, so ist diese nicht symmetrisch. Für einen gewichteten Graphen werden nicht 0 oder 1 für die Verbindungen der Knoten in der Adjazenzmatrix notiert, sondern die Gewichte der Verbindungen.

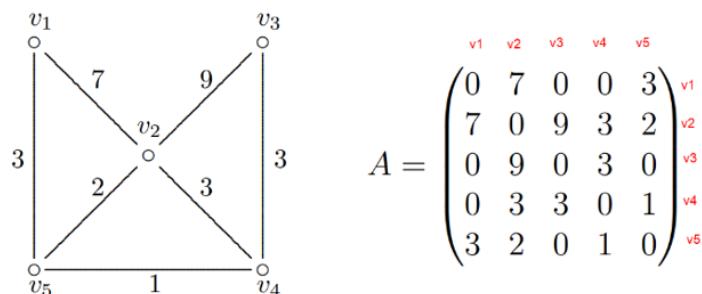


Abbildung 3: Adjazenzmatrix

3.2 Communities und Module

Communities können als eine Teilmenge von Knoten definiert werden, die dicht miteinander und lose mit den Knoten der anderen Communities desselben Graphen verbunden sind. Eine Segmentierung eines Graphen in Communities ist dahingehend wichtig, da eine Zuteilung zu einer bestimmten Klasse eine bessere Analyse der Graphstruktur erlaubt. Das Problem der Erkennung von Communities erfordert die Aufteilung eines Netzes in Gemeinschaften von dicht verbundenen Knoten, wobei die Knoten, die zu verschiedenen Gemeinschaften gehören, nur spärlich verbunden sind. Genaue Formulierungen dieses Optimierungsproblems sind bekanntermaßen rechnerisch schwer lösbar. Für die Suche nach Communities sind verschiedene Verfahren bekannt, jedoch läuft es bei grossen Netzen immer auf ein Approximationsverfahren hinaus, welches je nach gewähltem Algorithmus besser oder schlechtere Ergebnisse liefert. Auch die Rechenzeit unterscheidet sich stark von Algorithmus zu Algorithmus [4].

3.2.1 Anzahl Edges

Bei ungerichteten Graphen ist die Kantenzahl $m(G)$ eines gegebenen Graphen $G = (V, E)$ die Anzahl seiner Kanten, bzw. die Summe der Vielfachheiten der einzelnen Kanten, wenn es sich um einen Graphen mit Mehrfachkanten handelt.

Man kann sie auch als Mächtigkeit $|E|$ der Kantenmenge E sehen.

Um die Kantenzahl anhand einer Adjazenzmatrix zu berechnen, müssen nur alle Einträge addieren und noch durch 2 geteilt werden. Dieses Verfahren funktioniert auch für Graphen mit Mehrfachkanten.

Handelt es sich bei einem Graphen um einen **vollständigen Graphen**, so kann die Anzahl Edges m folgendermassen berechnet werden:

$$m = \binom{n}{2} = \frac{n(n-1)}{2} = \Delta_{n-1},$$

also der Dreieckszahl Δ_{n-1} .

3.2.2 Anzahl Nodes

Die Anzahl der Nodes alleine sagt nicht so gut wie nichts über einen Graphen aus, da zwei Graphen mit der selben Knotenzahl grundlegend verschieden sein können, durch ihre unterschiedlichen Knotenverbindungen.

3.2.3 Degree

Eine Schlüsseleigenschaft eines jeden Knotens ist sein Grad, der die Anzahl der Verbindungen zu anderen Knoten darstellt. Der Grad kann die Anzahl der Mobiltelefonkontakte einer Person im Anrufdiagramm darstellen (d. h. die Anzahl der verschiedenen Personen, mit denen die Person gesprochen hat) oder die Anzahl der Zitate, die eine Forschungsarbeit im Zitationsnetzwerk erhält.

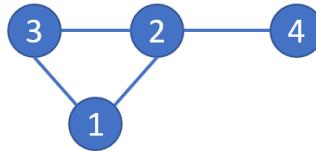


Abbildung 4: Einfacher Graph

Wir bezeichnen mit k_i den Grad des i -ten Knotens im Netz. Für die in Abbildung 4 dargestellten ungerichteten Netze haben wir beispielsweise $k_1 = 2, k_2 = 3, k_3 = 2, k_4 = 1$. In einem ungerichteten Netz kann die Gesamtzahl der Verbindungen, L , als Summe der Knotengrade ausgedrückt werden:

$$L = \frac{1}{2} \sum_{i=1}^N k_i$$

Hier korrigiert der Faktor $1/2$ die Tatsache, dass in der Summe jede Verbindung zweimal gezählt wird. Zum Beispiel wird die Verbindung zwischen den Knoten 2 und 4 in Bild 4 einmal im Grad von Knoten 1 und einmal im Grad von Knoten 4 gezählt[5].

3.2.4 Density

Die "Netzdichte" beschreibt den Anteil der potenziellen Verbindungen in einem Netz, die tatsächlich bestehen. Eine "potenzielle Verbindung" ist eine Verbindung, die potenziell zwischen zwei "Knoten" bestehen könnte - unabhängig davon, ob sie tatsächlich besteht oder nicht. Diese Person könnte jene Person kennen; dieser Computer könnte mit jenem verbunden sein. Ob sie tatsächlich miteinander verbunden sind, ist irrelevant, wenn es sich um eine potenzielle Verbindung handelt. Im Gegensatz dazu ist eine "tatsächliche Verbindung" eine, die tatsächlich besteht. Diese Person kennt diese Person; dieser Computer ist mit diesem Computer verbunden.

Die potenziellen Verbindungen E_p berechnen sich aus:

$$E_p = \frac{n(n-1)}{2}$$

und die Network Density D somit aus den möglichen Edges geteilt durch die effektiv vorhandenen Edges E

$$\text{UndirectedNetworkDensity} = \frac{\text{TotalEdges}}{\text{TotalPossibleEdges}} = \frac{\text{Cardinality}}{\text{Size}} = \frac{m}{n(n - 1)}$$

3.2.5 Betweenness Centrality

Die Betweenness Centrality misst, wie oft ein Knoten auf dem kürzesten Weg zwischen anderen Knoten liegt und ist definiert durch:

Sei $G = (V, E)$ ein Netz.

$$C_B(v_k) = \sum_{v_i, v_j \in V, v_i \neq v_j} \frac{\sigma_{v_i v_j}(v_k)}{\sigma_{v_i v_j}},$$

wobei σ für die Anzahl der kürzesten Wege von v_i nach v_j steht und sigma für die Anzahl der kürzesten Wege von v_i nach v_j , die v_k einschliessen. In der Tat kann die Menge

$$\frac{\sigma_{v_i v_j}(v_k)}{\sigma_{v_i v_j}}$$

als die Wahrscheinlichkeit angesehen werden, dass v_k auf einem kürzesten Weg liegt, der v_i mit v_j verbindet [6].

Was sie uns sagt: Dieses Mass zeigt, welche Knoten "Brücken" zwischen Knoten in einem Netz sind. Dazu werden alle kürzesten Pfade identifiziert und dann gezählt, wie oft jeder Knoten auf einem dieser Pfade liegt.

Wann ist es zu verwenden? Um die Knoten zu finden, die den Fluss in einem System beeinflussen.

3.2.6 Closeness Centrality

In einem verbundenen Graphen $G = (V, E)$ bestimmt die Closeness Centrality, wie nahe ein bestimmter Knoten an allen anderen Knoten ist. Dafür wird die Distanz d zwischen dem untersuchten Knoten v_k und allen anderen N Knoten v_i aufsummiert. Davon wird dann der Kehrwert verwendet.

$$C_C(V_k) = \frac{1}{\sum_{i=1}^N d(v_k, v_i)}$$

Damit ein Vergleich von unterschiedlichen Graphen anhand der Closeness Centrality stattfinden kann, muss der jeweilige errechnete Wert normiert werden. Dies wird durch Multiplikation des Kehrwerts mit der Anzahl Knoten N erreicht.

$$C_C(V_k) = \frac{N}{\sum_{i=1}^N d(v_k, v_i)}$$

3.2.7 Local Clustering Coefficient

Der Clustering-Koeffizient gibt an, in welchem Maße die Nachbarn eines bestimmten Knotens miteinander verbunden sind. Für einen Knoten i mit dem Grad k_i ist der lokale Clustering-Koeffizient definiert als

$$C_i = \frac{2L_i}{k_i(k_i - 1)}$$

wobei L_i die Anzahl der Verbindungen zwischen den k_i Nachbarn des Knotens i darstellt. Man beachte, dass C_i zwischen 0 und 1 liegt:

$C_i = 0$, wenn keiner der Nachbarn von Knoten i eine Verbindung zueinander hat.

$C_i = 1$, wenn die Nachbarn des Knotens i einen vollständigen Graphen bilden, d. h. sie alle miteinander verbunden sind.

C_i ist die Wahrscheinlichkeit, dass zwei Nachbarn eines Knotens eine Verbindung zueinander herstellen. $C = 0,5$ bedeutet also, dass die Wahrscheinlichkeit, dass zwei Nachbarn eines Knotens

miteinander verbunden sind, bei 50% liegt.

Zusammenfassend lässt sich sagen, dass C_i die lokale Verbindungsichte des Netzes misst: Je dichter die Nachbarschaft eines Knotens i miteinander verbunden ist, desto höher ist sein lokaler Clustering-Koeffizient [5].

3.2.8 Average Clustering Coefficient [global]

Der Grad der Clusterung eines gesamten Netzwerks wird durch den durchschnittlichen Clustering-Koeffizienten $\langle C \rangle$ erfasst, der den Durchschnitt von C_i über alle Knoten $i = 1, \dots, N$ darstellt [5],

$$\langle C \rangle = \frac{1}{N} \sum_{n=1}^N C_i$$

Entsprechend der probabilistischen Interpretation ist $\langle C \rangle$ die Wahrscheinlichkeit, dass zwei Nachbarn eines zufällig ausgewählten Knotens eine Verbindung zueinander herstellen.

Während die zuvor gesehene Gleichung für ungerichtete Netzwerke definiert ist, kann der Clustering-Koeffizient auch auf gerichtete und gewichtete Netzwerke verallgemeinert werden.

3.2.9 Wiener Index [global]

Der Wiener-Index dient zur Abbildung von Strukturinformationen eines Graphen auf eine Zahl und gehört somit zu den globalen Massen. Es werden die Summen aller Abstände (kürzeste Wege) von jedem Node zu jedem anderen Node in die Berechnung mit einbezogen. Der Wiener Index ist definiert durch

$$W = \frac{1}{2} \sum_{i,j}^N d_{ij}$$

4 Krankheiten

Nachfolgend werden die beiden zu analysierenden Krankheiten Alzheimer und Schizophrenia und deren involvierten Gene aufgezeigt. Dabei ist zu jedem Gen eine Beschreibung aus der StringDB vorhanden, diese wurde in der Originalsprache Englisch belassen. Sie finden die detaillierte Beschreibung der Gene im Anhang unter [Beschreibung der Gene](#).

4.1 Alzheimer

Bei Alzheimer handelt es sich um eine Genmutation, welche vererbt werden kann. Es handelt sich dabei um die häufigste Form der Demenz. Es handelt sich um eine fortschreitende Krankheit, die mit einem leichten Gedächtnisverlust beginnt und möglicherweise zum Verlust der Fähigkeit führt, ein Gespräch zu führen und auf die Umwelt zu reagieren. Bei der Alzheimer-Krankheit sind Teile des Gehirns betroffen, die das Denken, das Gedächtnis und die Sprache steuern. Sie kann die Fähigkeit einer Person, alltägliche Aktivitäten zu verrichten, ernsthaft beeinträchtigen [7]. Dabei führt eine Mutation der APP, PSEN1, PSEN2 Gene zu der Produktion von gefährlichen Amyloid Beta Peptide Proteinen im Gehirn (Ablagerungen) Folgende Gene stehen mit der Krankheit Alzheimer in direkter Verbindung, wobei die identifizierten Gene als Hauptverursacher der Krankheit gelten:

Relevante und verwandte Gene	
APP	Amyloid-beta A4 protein
HFE	Hereditary hemochromatosis protein
MPO	Myeloperoxidase
NOS3	Nitric Oxide Synthase 3
PLAU	Urokinase-type plasminogen activator
ABCA7	ATP binding cassette subfamily A member 7
PSEN1, PSEN2	Presenilin 1 / 2
APOE	Apolipoprotein E

Tabelle 1: Auflistung der betroffenen Gene von Alzheimer

4.2 Schizophrenie

Schizophrenie ist eine chronische und schwere psychische Störung, von der weltweit 20 Millionen Menschen betroffen sind [8]. Schizophrenie ist gekennzeichnet durch Störungen des Denkens, der Wahrnehmung, der Emotionen, der Sprache, des Selbstbewusstseins und des Verhaltens. Häufig treten Halluzinationen (Hören von Stimmen oder Sehen von Dingen, die nicht da sind) und Wahnsvorstellungen (feste, falsche Überzeugungen) auf. Schizophrenie ist weltweit mit erheblichen Behinderungen verbunden und kann die schulischen und beruflichen Leistungen beeinträchtigen. Bei Menschen mit Schizophrenie ist die Wahrscheinlichkeit, früh zu sterben, 2-3-mal höher als in der Allgemeinbevölkerung [9]. Dies ist häufig auf vermeidbare körperliche Erkrankungen wie Herz-Kreislauf-Erkrankungen, Stoffwechselkrankheiten und Infektionen zurückzuführen.

Relevante und verwandte Gene	
APOL2, APOL4	Apolipoprotein L
CHI3L1	Chitinase 3 Like 1
COMT	Catechol-O-Methyltransferase
DAOA	D-Amino Acid Oxidase Activator
DISC2	Disrupted In Schizophrenia
DRD3	Dopamine Receptor D3
HTR2A	5-Hydroxytryptamine Receptor 2A
MTHFR	Methylenetetrahydrofolate Reductase
RTN4R	Reticulon 4 Receptor
SYN2	Synapsin II
SHANK3	SH3 And Multiple Ankyrin Repeat Domains 3
DISC1	DISC1 Scaffold Protein
RBM12	RNA Binding Motif Protein 12
NRXN1	Neurexin 1
SLC1A1	Solute Carrier Family 1 Member 1
PRODH	Proline Dehydrogenase 1
NRG1	Neuregulin 1

Tabelle 2: Auflistung der betroffenen Gene von Schizophrenie

4.3 Bestehende Literatur

4.3.1 Barabasi

Im Kapitel 9 aus dem Network Science Book von Barabasi deutet er auf die Relevanz von Netzwerken und Communities (Modulen) im Betracht auf das Verstehen von Krankheiten. Weiter meint er, dass Proteine, welche in derselben Krankheit vorkommen, eine Tendenz haben, miteinander zu interagieren.

"Communities play a particularly important role in understanding human diseases. Indeed, proteins that are involved in the same disease tend to interact with each other." [5]

In unserer Arbeit wollen wir uns dieser Interaktion der Gene annähern und verstehen, was genau mit einer Interaktion gemeint ist. Zusätzlich wollen wir herausfinden, ob es ähnliche Communities oder Strukturen in den beiden Krankheiten gibt.

4.3.2 Kwang-II Goh

Auch Kwang-II Goh et al. deuten auf einen Zusammenhang zwischen Genen hin, welche in derselben Krankheit vorkommen. Sie sprechen in ihrer Arbeit einer Ähnlichkeit der Transkripte der Gene.

"Genes associated with similar disorders show both higher likelihood of physical interactions between their products and higher expression profiling similarity for their transcripts" [10]

Es fällt nun auch der Begriff physical interaction welchen wir an dieser Stelle noch gar nicht behandelt haben. Im Kapitel 5 bei der Vorstellung der String-DB werden wir näher auf den Begriff der physischen Interaktion eingehen.

4.3.3 Genes And Diseases

Eine eher allgemeine Feststellung aus dem Buch "Genes and Diseases" vom National Center for Biotechnology Information stärkt erneut die Behauptung der Relevanz von Gruppen / Netzwerken von Genen, welche die Beeinflussung der menschlichen Gesundheit zuständig sind.

"In all these cases, no one gene has the yes/no power to say whether a person has a disease or not. It is likely that more than one mutation is required before the disease is manifest, and a number of genes may each make a subtle contribution to a person's susceptibility to a disease" [1]

Es existiert kein bekannter Fall, in welchem nur ein einzelnes Gen für die Ausprägung einer Krankheit gesorgt hat.

5 Netzwerk Analyse

5.1 Beschaffung der Daten

Zuerst mussten die für die Krankheit relevanten Gene identifiziert werden. Diese Gene wurden mit der MedGen Datenbank der NCBI von [7] und [11] geholt.

Schizophrenia (SCZD)
MedGen UID: 48574 • Concept ID: C0036341 • Mental or Behavioral Dysfunction
Synonyms: SCHIZOPHRENIA WITH OR WITHOUT AN AFFECTIVE DISORDER; SCZD
SNOMED CT: Schizophrenic disorders (191526005); Schizophrenia (58214004)
Modes of inheritance: Heterogeneous (HPO) Autosomal dominant inheritance (HPO, OMIM)
Genes (locations): APOL2 (22q12.3); APOL4 (22q12.3); CHI3L1 (1q32.1); COMT (22q11.21); DAOA (13q33.2); DISC2 (1q42.2); DRD3 (3q13.31); HTR2A (13q14.2); MTHFR (1p36.22); RTN4R (22q11.21); SYN2 (3p25.2)
Related genes: SHANK3, DISC1, RBM12, NRXN1, SLC1A1, PRODH, NRG1
HPO: HP:0100753
Monarch Initiative: MONDO:0005090
OMIM®: 181500

Abbildung 5: MedGen Eintrag in der NCBI Datenbank

Danach konnten die Netzwerk Daten (Knoten und Kanten) von der String-DB [12] beschafft werden. Die Dateien wurden nach dem Organismus *Homo Sapiens* gefiltert.

Die gesamte String-DB beinhaltet über 14094 Genome und über 65 Millionen Proteine.

INTERACTION DATA			
File	Description	Access	
9606_protein.links.v11.5.txt.gz (72.7 Mb)	protein network data (scored links between proteins)		
9606_protein.links.detailed.v11.5.txt.gz (115.5 Mb)	protein network data (incl. subscores per channel)		
9606_protein.links.full.v11.5.txt.gz (133.6 Mb)	protein network data (incl. distinction: direct vs. interologs)		
9606_protein.physical.links.v11.5.txt.gz (12.0 Mb)	protein network data (scored links between proteins)		
9606_protein.physical.links.detailed.v11.5.txt.gz (15.1 Mb)	protein network data (incl. subscores per channel)		
9606_protein.physical.links.full.v11.5.txt.gz (16.2 Mb)	protein network data (incl. distinction: direct vs. interologs)		

Abbildung 6: String-DB Download Verzeichnis

Wie in der Abbildung 6 zu sehen ist, sind die komprimierten Dateien immer noch sehr gross. Diese Dateien beinhalten alle Gene inklusive alle Verbindungen des *Homo Sapiens* Genoms.

Zusätzlich wurden die lokalen Netzwerke der relevanten und verwandten Gene 4 der beiden Krankheiten heruntergeladen. Mit der Suche eines Genes aus unserer Liste erhalten wir das Gen mit 10 weiteren Verbindungen. Die Kanten zu den anderen Genen sind sortiert nach dem Combined-Score der Kanten.

5.1.1 String-DB Kanten Score

Die Kanten welche eine Verbindung von 2 Genen repräsentiert, können auf verschiedene Arten entstehen. Es wird unterschieden nach dem Zusammenhang welche nochmals eingeteilt in physisch oder funktionell ist. Diese Werte ergeben dann den Combined-Score.

Physische Zusammenhänge:

- Gen-Fusion
- Gen-Nachbarschaft

Funktionelle Zusammenhänge:

- Interaktion-Experimente
- Kuratiertes Wissen
- Datenbanken
- Literatur

5.1.2 String-DB Gen

Suchen wir in der String-DB Beispielsweise nach dem Gen ÄPP/ „Amyloid-Beta A4 Protein“, so finden wir folgende Daten.

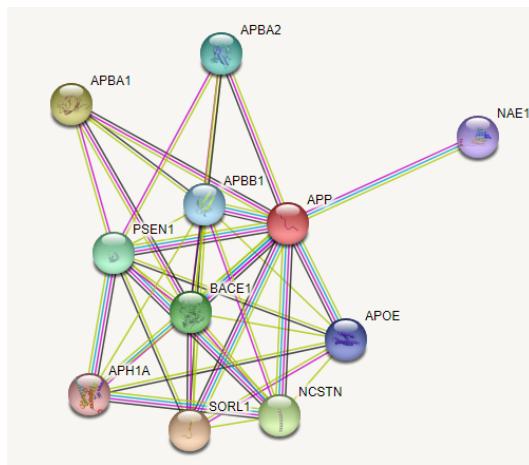


Abbildung 7: Das Lokale Netzwerk des APP Genes

Die Kanten-Scores des APP Genes mit den anderen Genen aus dem Netzwerk sind alle beinahe bei 1. In der Legende finden wir die Ursprünge der Kanten und können sogar mehr darüber herausfinden.

Zum Beispiel die hat die Kante $\{\{APP, APOE\}, \dots\}$ einen Score von 0.998 welche sich aus Coexpression (0.071), Experiments (0.871), Databases (0.720) und Textmining (0.972) berechnet.

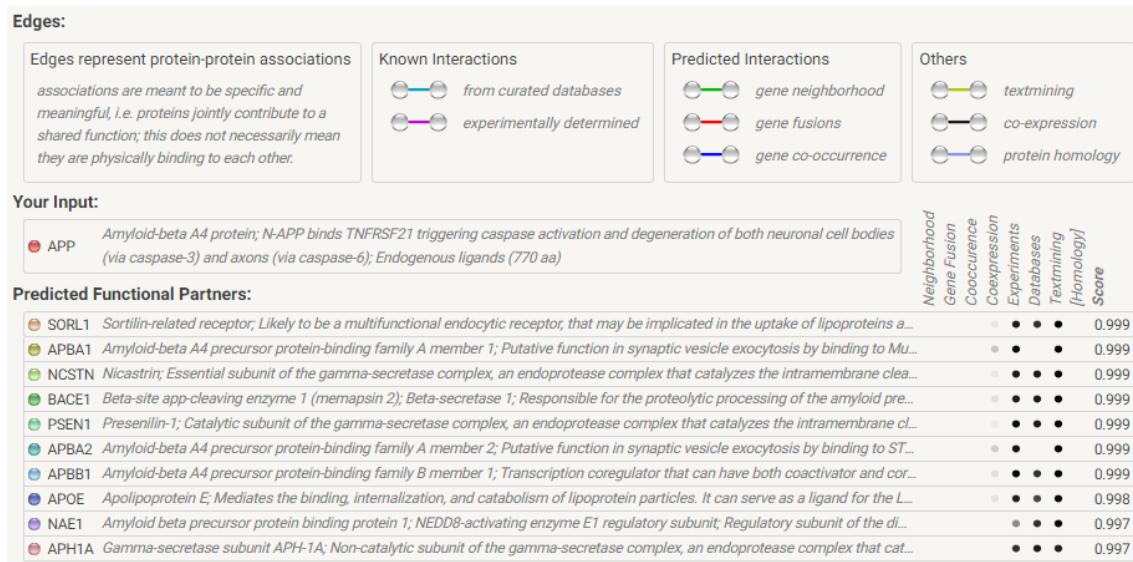


Abbildung 8: Die Legende der Kanten vom APP Gen

5.1.3 Daten Aufbereitung

Nach dem Herunterladen der TSV Dateien von der String-DB konvertierten wir sie mit Pandas zu CSV Dateien. Mit dem CSV Dateiformat kann Pandas / Networkx besser umgehen.

```

1      # Prepare Downloaded TSV Files (Convert to CSV)
2      path = "data/**/*/*.tsv"
3      for fname in glob.glob(path):
4          csv_table = pd.read_table(fname, sep='\t')
5          file_name = Path(fname).stem
6          file_path = Path(fname).resolve().parent
7          csv_table.to_csv(f'{file_path}/{file_name}.csv', index=False)

```

Listing 1: Konvertiere StringDB TSV-Files zu CSV

Für die bessere Bearbeitung der Netzwerke, transformieren wir in dem Datenaufbereitungs Skript ebenfalls die ID von der Protein ID (9606.ENSP00000284981) zum Kürzel **APP**.

Für das erstellten wir im Pandas Dataframe eine neue Spalte und wendeten eine Lambda-Funktion auf die Spalte an, welche den Wert **APP** für den Key **9606.ENSP00000284981** aus einem vorbereiteten Dictionary ausliest.

```

1      ad_dict = {
2          # Alzheimer Disease
3          "APP": "9606.ENSP00000284981",
4          ...
5      }
6
7      # Prepare Data (Create Label Columns)
8      def get_key(val):
9          val = all_nodes.loc[val]["preferred_name"]
10         if val:
11             return val
12
13         return ""
14
15     df_ad["Source"] = df_ad.apply(lambda row: get_key(row["SourceId"]), axis=1)
16     ...
17
18     df_ad.to_csv('data/ad/ad_network_full_with_labels.csv', index=False)
19     ...

```

Listing 2: Schreibe Lesbare Lables in die Source und Target Spalten der Pandas Dataframes

Zum Schluss wurden die Dataframes wieder als CSV exportiert um später als Basis für die Analyse mit Networkx und Graph-Tool zu dienen.

Als nächstes haben wir nun folgende Ordner und Dateistruktur vorliegen. Die Liste ist nicht abgeschlossen und verkürzt zur besseren Lesbarkeit:

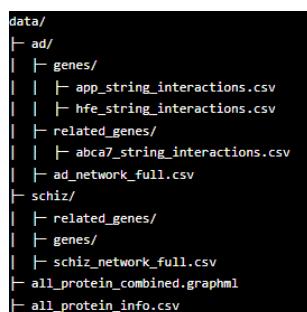


Abbildung 9: Dateistruktur nach dem Aufbereiten der Daten

5.2 Analyse der Netzwerke

5.2.1 Jupyter Notebook mit pandas und networkx

Die Funktion zur Berechnung der Graph Massen aus `networkx` wurden in einer Funktion gespeichert, welche die Vorbereitete Datei aus Kapitel 5.2 als Parameter erhält.

Wir haben die Daten dazu in verschiedene Dateien aufgeteilt und untersuchen die einzelnen Netzwerke der Gene sowie das gesamte Netzwerk aller Gene.

Zuerst lesen wir die Dateien der einzelnen, kleinen Netzwerke ein und ziehen über `networkx` die Messwerte heraus.

```

1     G = nx.from_pandas_edgelist(df, source="Source", target="Target",
2                                   create_using=Graphtype)
3     s_closeness_dict[file_name] = np.average(list(nx.closeness_centrality(G).values()))
4     s_betweenness_dict[file_name] = np.average(list(nx.betweenness_centrality(G).values()))
5     s_edges_dict[file_name] = nx.number_of_edges(G)
6     s_nodes_dict[file_name] = nx.number_of_nodes(G)
7     s_avg_clustering_dict[file_name] = nx.average_clustering(G)
8     s_wiener_dict[file_name] = nx.wiener_index(G)
9
10    print_graph_analytics(G, file_name)
11    plot_graph(G, file_name)

```

Listing 3: Berechne Globale Netzwerkmasse mit networkx

Dasselbe machen wir auch für die Schizophrenie Gen-Dateien. In den nachfolgenden Codeblocks werden gewisse Werte normalisiert. Dies machen wir mit folgender Funktion: $e^{-(x-y)^2}$

Die nächste Funktion berechnet die Graph-Masse für alle einzelnen Gen-Netzwerke.

```

1     def calculate_graph_measures(graph, relevant_nodes):
2         # betweenness_centrality
3         betweenness_centrality = nx.betweenness_centrality(graph)
4         betweenness_centrality_filtered = dict(
5             filter(lambda elem: elem[0] in relevant_nodes, betweenness_centrality.items()))
6
7         # # closeness_centrality
8         closeness_centrality = nx.closeness_centrality(graph)
9         closeness_centrality_filtered = dict(filter(lambda elem: elem[0] in relevant_nodes,
10                                         closeness_centrality.items()))
11
12         # # average_degree_connectivity
13         avg_local_degree = nx.algorithms.assortativity.average_degree_connectivity(graph)
14         avg_local_degree_filtered = dict(filter(lambda elem: elem[0] in relevant_nodes,
15                                         avg_local_degree.items()))
16         return {
17             'nodes': nx.number_of_nodes(graph),
18             'edges': nx.number_of_edges(graph),
19             'density': nx.density(graph),
20             'avg_clustering': nx.algorithms.average_clustering(graph),
21             'wiener_index': nx.algorithms.wiener_index(graph),
22             'betweenness_centrality': np.average(list(betweenness_centrality_filtered.values())),
23             'closeness_centrality': np.average(list(closeness_centrality_filtered.values())),
24             'average_degree_connectivity': np.average(list(avg_local_degree_filtered.values())),
25             'local_clustering_coefficient': nx.algorithms.approximation.average_clustering(graph)
26         }

```

Listing 4: Berechne Graph-Masse aus den einzelnen Gen-Netzwerken

Nun wollen wir noch die Werte aus dem Schizophrenie und die des Alzheimer Netzwerkes miteinander vergleichen.

```

1 Graphtype = nx.Graph()
2 G = nx.from_pandas_edgelist(df_ad, source="Source", target="Target",
3                             create_using=Graphtype, edge_attr=True)
4 relevant_genes = ["APP", "HFE", "MPO", "NOS3", "PLAU"]
5 measures_AD = calculate_graph_measures(G, relevant_genes)
6
7 G = nx.from_pandas_edgelist(df_schiz, source="Source", target="Target",
8                             create_using=Graphtype, edge_attr=True)
9 relevant_genes = ["APOL2", "APOL4", "CHI3L1", "COMT", "DAOA", "DISC1",
10                  "DRD3", "HTR2A", "MTHFR", "RTN4R", "SYN2"]
11 measures_schiz = calculate_graph_measures(G, relevant_genes)
12
13 for i in range(len(measures_AD)):
14     normalized = normalize_values(list(measures_AD.values())[i],
15                                    list(measures_schiz.values())[i])
16     print(f'Comparing measure {list(measures_AD.keys())[i]}: \
17           {list(measures_AD.values())[i]} to \
18           {list(measures_schiz.values())[i]} -> normalized: {normalized}')

```

Listing 5: Vergleiche alle Graph-Masse vom Schizophrenie Netzwerke mit dem Alzheimer Netzwerk

5.2.2 Gesamt Netzwerk Analyse mit graph-tool

Da die meisten Python Programme und Bibliotheken wie `networkx` oder `igraph` zum Teil direkt mit Python implementiert wurden sind sie nicht wirklich geeignet für grosse Netzwerke. Beim den ersten Testversuchen mit `networkx` lief der Code mehrere Stunde lang, um nur schon den durchschnittlichen Clustering Coefficient zu berechnen.

Bei der Recherche nach leistungsoptimiertem Code stiessen wir auf `graph-tool` [13]. Der gesamte Code ist in C++ programmiert und verspricht eine Memory- und Operationszeit-Leistung einer puren C, C++ Bibliothek.

Die Installation wurde mittels Docker durchgeführt⁴. Der Docker Container verfügt über eine Jupyter Notebook Schnittstelle, welche direkt vom lokalen Host aus geöffnet werden kann.

```

1 from graph_tool.all import *
2 g = load_graph("all_protein_combined.graphml")
3 print(g)

```

Listing 6: Graph-Tool Import Network and Print Graph Stats

Output:

```
<Graph object, undirected, with 19382 vertices and 5968680 edges,
 1 internal vertex property, 4 internal edge properties, at 0x7f63072dce80>
```

Die Abfrage dauerte über 5 Minuten.

Wie wir in der Ausgabe sehen können, wurde der gesamt Graph erfolgreich eingelesen. Er beinhaltet **19'382** Knoten und **5'968'680** Kanten!

⁴<https://git.skewed.de/count0/graph-tool/-/wikis/installation-instructions#installing-using-docker>

Als nächstes rechnen wir die Zentralitätswerte.

```

1   # Centrality
2   vp_betweenness, ep_betweenness = graph_tool.centrality.betweenness(g)
3   print(f"vertex p betweenness = {vp_betweenness}")
4   print(f"edge p betweenness = {ep_betweenness}")
5
6   vp_closeness = graph_tool.centrality.closeness(g)
7   print(f"vertex p closeness = {vp_closeness}")

```

Listing 7: Ausgabe der Zentralitätsmasse

Output:

```

vertex p betweenness = <VertexPropertyMap object with value type 'double',
for Graph 0x7f68a260f9d0, at 0x7f688807ca30>
edge p betweenness = <EdgePropertyMap object with value type 'double',
for Graph 0x7f68a260f9d0, at 0x7f688807caf0>
vertex p closeness = <VertexPropertyMap object with value type 'double',
for Graph 0x7f68a260f9d0, at 0x7f688807cee0>

```

Die Werte werden als Vertex und Edge PropertyMaps ausgegeben. Dies sind Klassen (Hashmaps) der Graph-Tool Library um die Geschwindigkeit zu erhöhen. In dem Kapitel 6 werden die Resultate noch dargestellt.

Wir wenden uns nun dem Clustering des gesamten Netzwerkes zu. Die Berechnung des lokalen Clustering Koeffizienten wird nach Watts-Strogatz[14] gemacht.

$$c_i = \frac{|\{e_j k\}|}{k_i(k_i - 1)} : v_j, v_k \in N_i, e_j k \in E$$

Wir geben davon einfach den Durchschnitt über alle Nodes aus.

Zusätzlich rechnen wir noch den globalen Clustering Coeffizienten nach Newmann-Structure[15].

$$c = 3 \times \frac{\text{number of triangles}}{\text{number of connected triplets}}$$

```

1   # Clustering
2   local_clustering = graph_tool.clustering.local_clustering(g)
3   print(graph_tool.stats.vertex_average(g, local_clustering))
4   global_clustering = graph_tool.clustering.global_clustering(g)
5   print(global_clustering)

```

Listing 8: Berechnen und ausgeben der Clustering Werte

Output:

```
(0.1952338728921458, 0.0006214113968411895)
(0.19299713533367516, 0.0010956810919999344)
```

Zum Schluss berechnen wir mit Graph-Tool noch Communities mit `minimize_blockmodel_dl`. Die Funktion ist eine High-Level Funktion welche auf einer tieferen Schicht den Markov-Chain Monte Carlo Algorithmus anwendet. Diese Operation hat eine Laufzeitkomplexität von $O(V * \ln^2 * V)$ bei der Berücksichtigung von 19'382 Knoten im Netzwerk sind das ≈ 2.872 Millionen Operationen.

```
1  # Draw with Modularity
2  state = graph_tool.inference.minimize.minimize_blockmodel_dl(g)
3  print(state)
4  # state.draw(pos=g.vp.pos, output="blockmodel.svg")
```

Listing 9: Communities mit Minimize Blockmodel

Output:

```
<BlockState object with 19382 blocks (394 nonempty), degree-corrected,
 for graph <Graph object, undirected, with 19382 vertices and 5968680 edges,
 1 internal vertex property,
 4 internal edge properties, at 0x7f63072dce80>,
 at 0x7f62edfd6520>
```

Die Ausgabe des SVG Bildes ist unter Kapitel 6.2 Abbildung [10](#).

6 Ergebnisse

In diesem Kapitel werden die Resultate aus der Netzwerk-Analyse vorgestellt.

Zuerst erfolgt ein Überblick über das gesamte homo sapiens Genom-Netzwerke. Zugleich werden die Kennzahlen der beiden Netzwerke (Alzheimer und Schizophrenie) miteinander verglichen.

Zum Schluss erfolgt die Analyse der einzelnen Gen-Netzwerke.

6.1 Netzwerk Kennzahlen

	Volles Netzwerk	Alzheimer	Schizophrenia
Nodes	19382	6089	7779
Edges	5968680	11310	16212
Density	0.015889254	0.000610196	0.000535889
Avg. Clustering	0.192997135	0.402805845	0.383712370
Wiener Index		47421635	88595362
Betweenness Centrality	0.000053691	0.241132202	0.086519757
Closeness Centrality	0.491657973	0.568955981	0.476481870
Local Clustering	0.195233872	0.396	0.357
Clustering ohne 0 Werte		0.979506707	0.912255052

Tabelle 3: Übersicht der Netzwerk Metriken

6.2 Darstellung Gesamtes Netzwerk

Das Netzwerk wurde mit der Stochastischen Blockmodell Inferenz unter Verwendung des Markov Chain Monte Carlo Algorithmus in Partitionen eingeteilt.

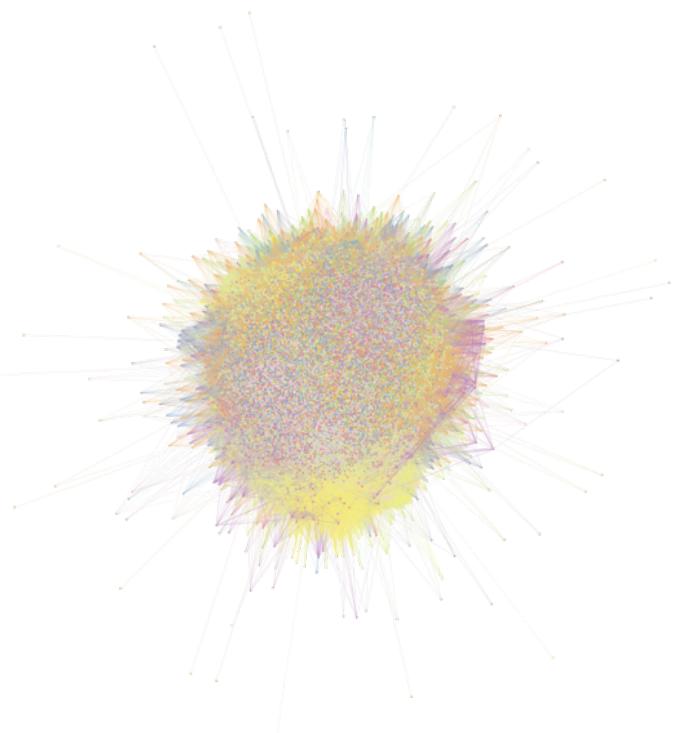


Abbildung 10: Homo Sapiens Genom Netzwerk

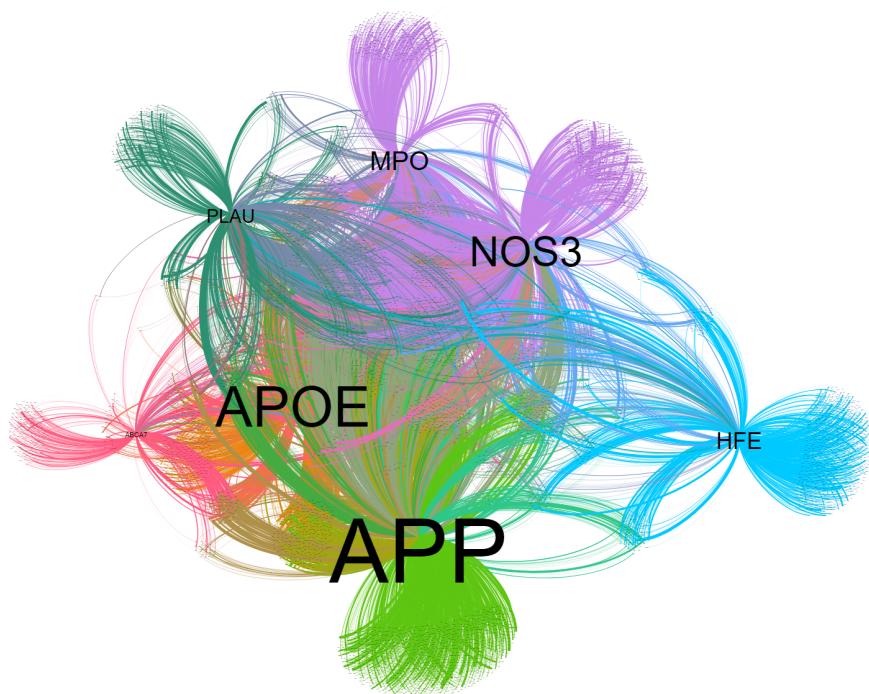


Abbildung 11: Netzwerk der Alzheimer Gene

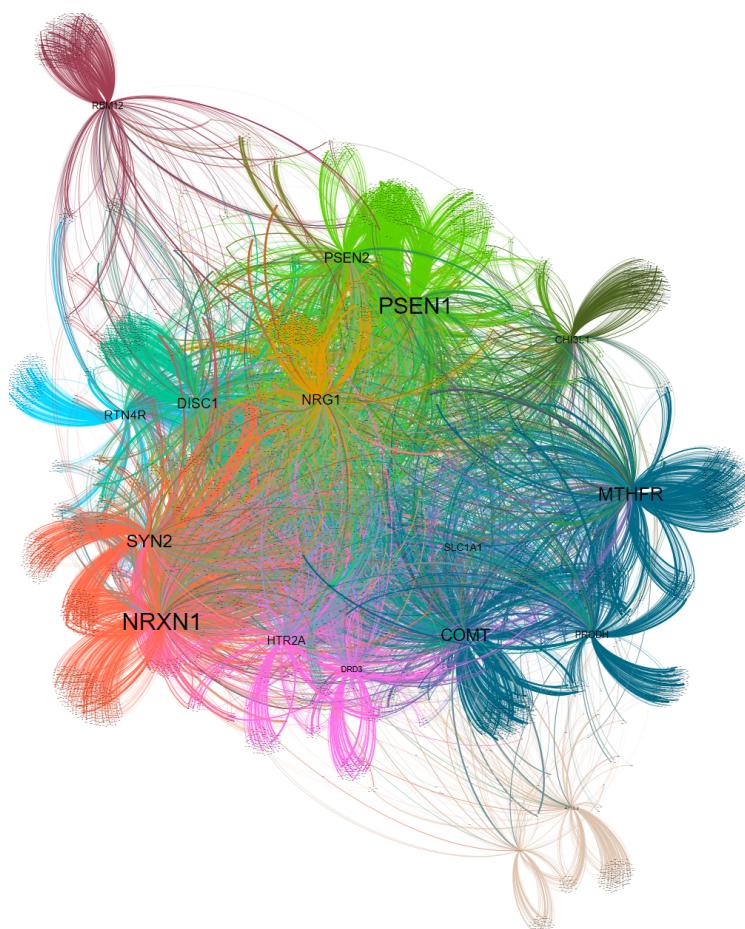


Abbildung 12: Netzwerk der Schizophrenie Gene

6.3 Darstellung Gen-Interaktionen bei Alzheimer

Nachfolgende Graphen zeigen für jedes relevante Gen der Krankheit Alzheimer die häufigsten Interaktionen mit anderen Genen. Es wurden jeweils immer die Zehn aktivsten Gene plus das zu untersuchende Gen in einem Graph dargestellt.

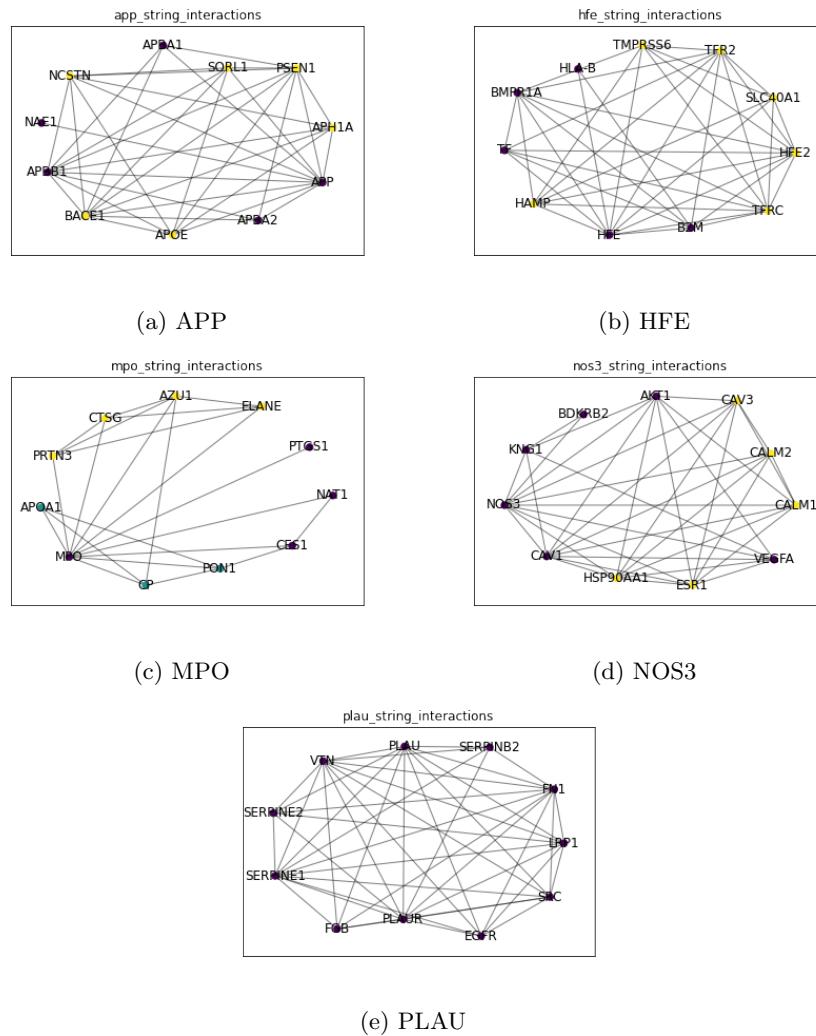


Abbildung 13: Involvierte Gene bei Alzheimer

6.4 Darstellung Gen-Interaktionen bei Schizophrenie

Nachfolgende Graphen zeigen für jedes relevante Gen der Krankheit Schizophrenie die häufigsten Interaktionen mit anderen Genen. Es wurden jeweils immer die Zehn aktivsten Gene plus das zu untersuchende Gen in einem Graph dargestellt.

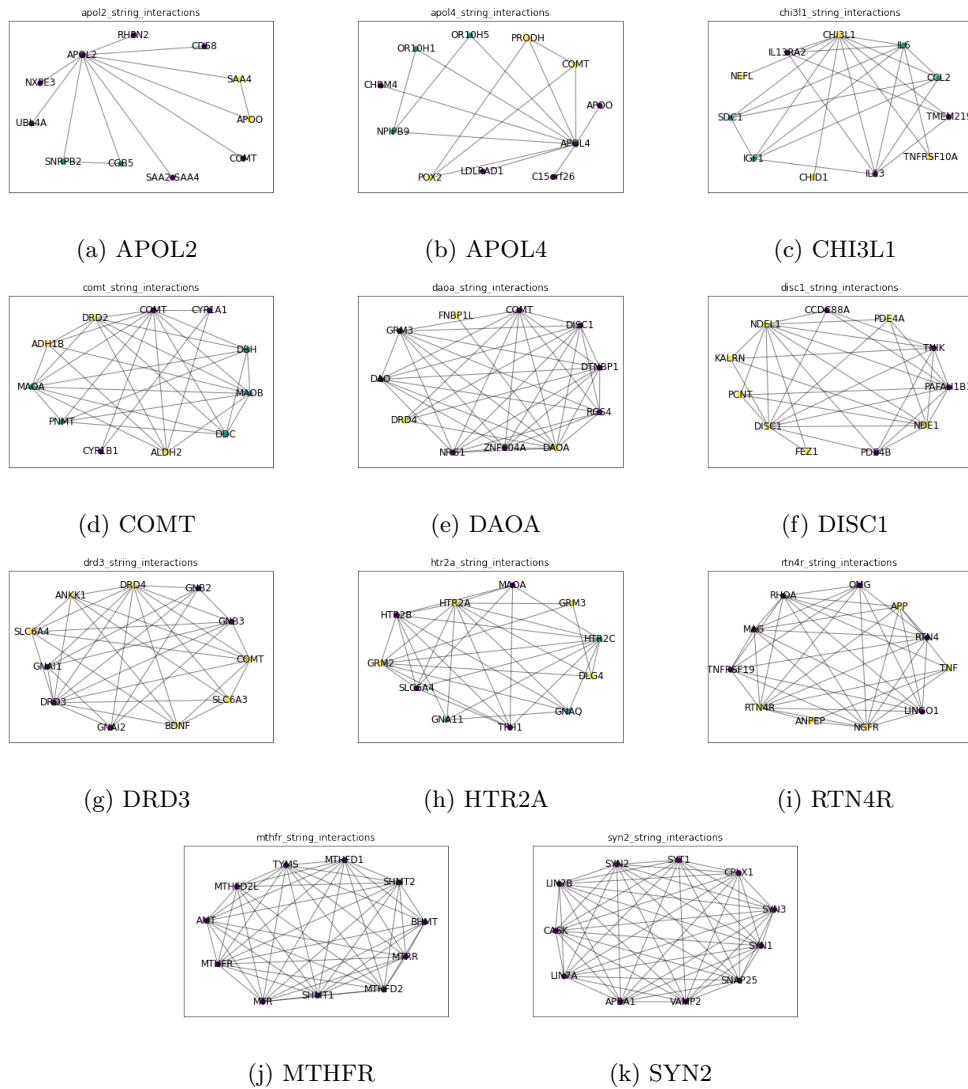


Abbildung 14: Involvierte Gene bei Schizophrenie (1)

6.5 Vergleich der Messwerte einzelner Gene (Alzheimer vs. Schizophrenie)

Die im vorherigen Kapitel gezeigten Graphen wurden in diesem Kapitel auf ihre Messwerte hin untersucht. Dabei sind die Messwerte *Anzahl Nodes*, *Anzahl Edges*, *Closeness Centrality*, *Betweenness Centrality*, *durchschnittliches Clustering*, sowie *Wiener Index* ersichtlich. Für Alzheimer wurde das Kürzel "AD" und für Schizophrenie das Kürzel "Schiz" verwendet.

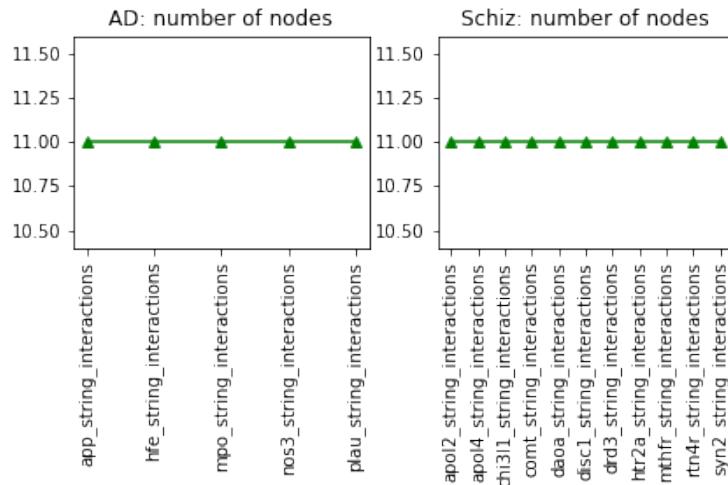


Abbildung 15: Anzahl Nodes

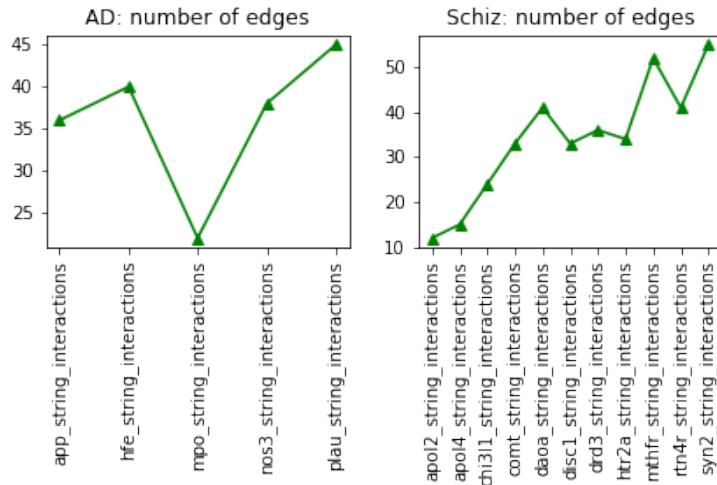


Abbildung 16: Anzahl Edges

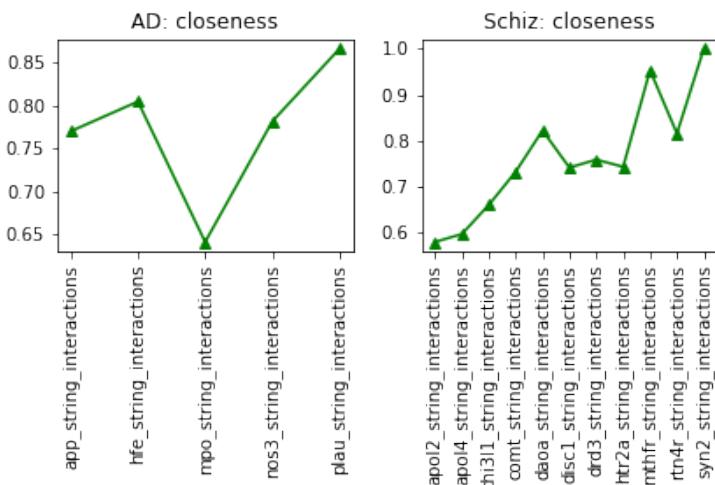


Abbildung 17: Closeness

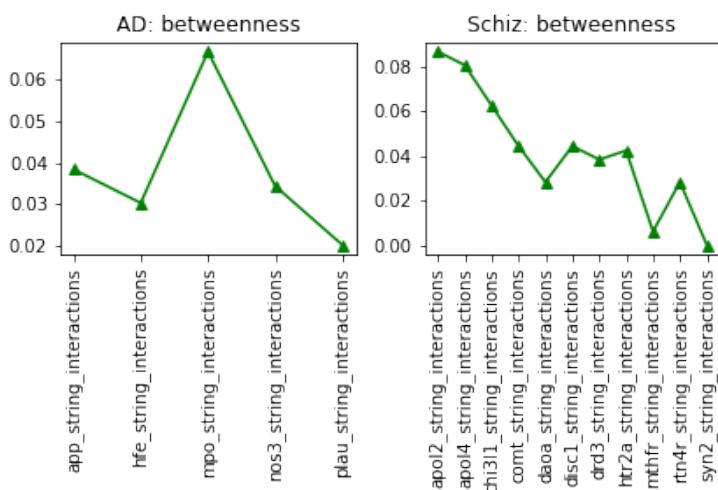


Abbildung 18: Betweenness

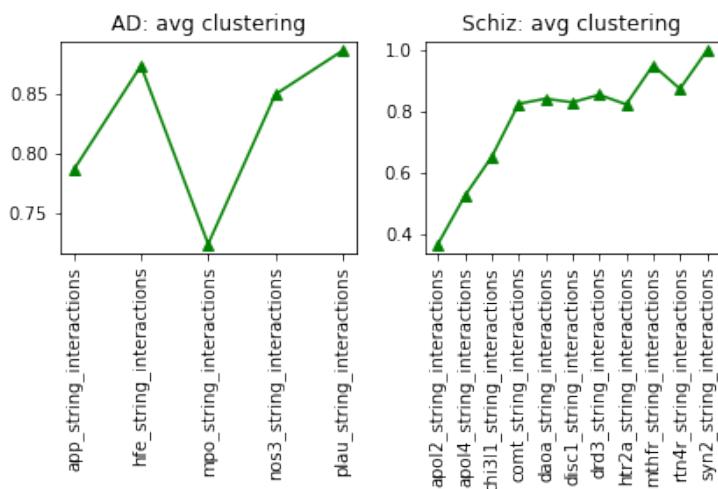


Abbildung 19: Avg. Clustering Coefficient

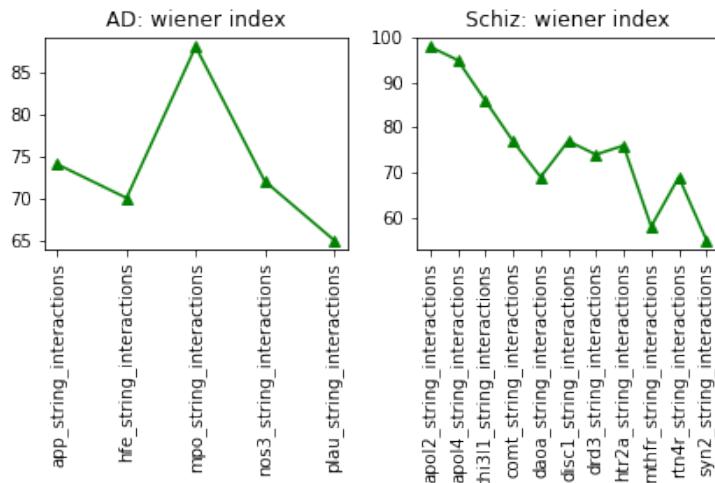


Abbildung 20: Wiener Index

6.6 Messwerte relevanter Gene im gesamten Netzwerk gemessen

Das folgende Kapitel zeigt die Auswertung der Kennzahlen *Degree*, *Closeness Centrality*, *Betweenness Centrality*, und *local Clustering* für die relevanten Gene im gesamten Netzwerk des menschlichen Genoms.

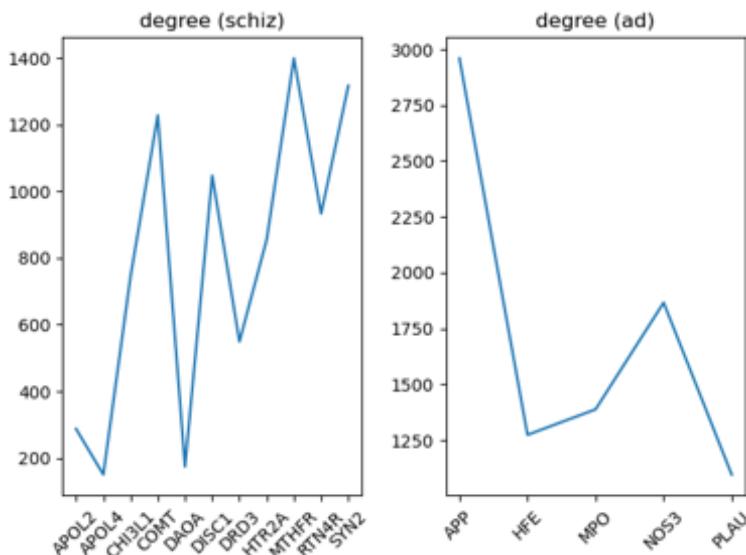


Abbildung 21: Ganzes Netz: Degree

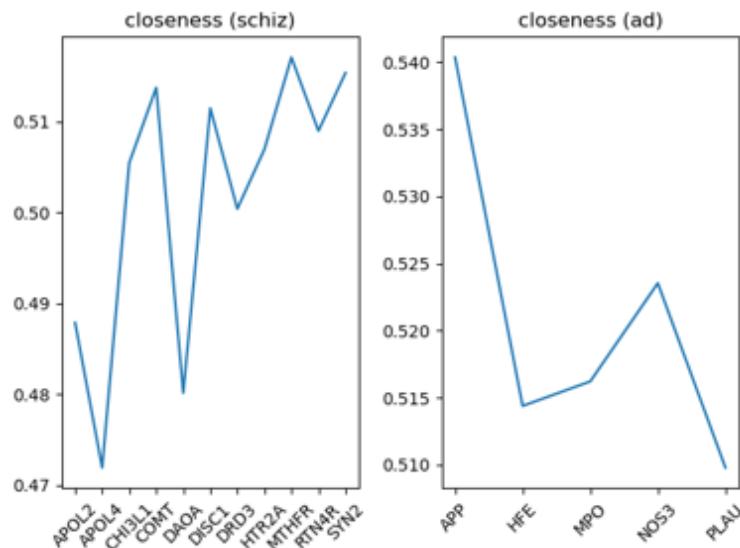


Abbildung 22: Ganzes Netz: Closeness

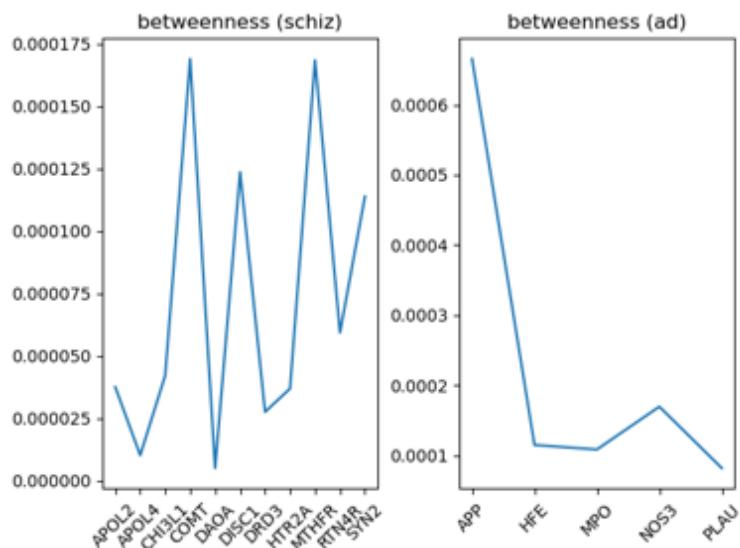


Abbildung 23: Ganzes Netz: Betweenness

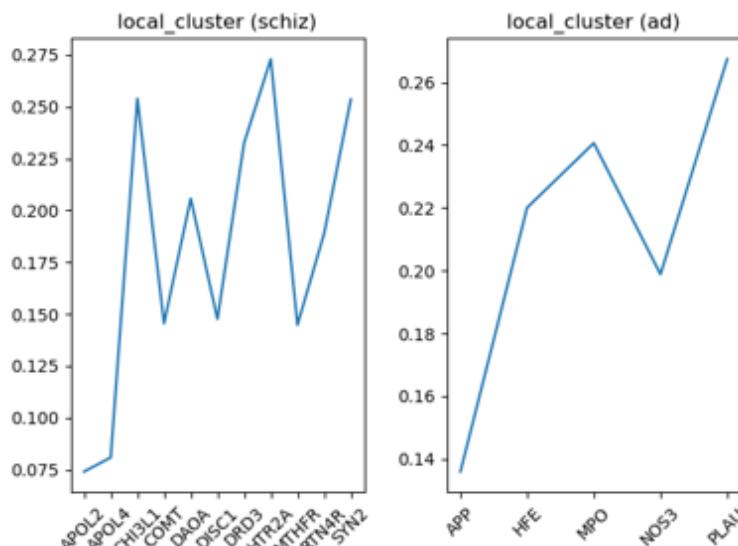


Abbildung 24: Ganzes Netz: Local Clustering

6.7 Messwerte relevanter Gene aus gefilterten Netzwerken gemessen

Das folgende Kapitel zeigt die Auswertung der Kennzahlen *Degree*, *Closeness Centrality*, *Betweenness Centrality*, und *local Clustering* für die relevanten Gene in den gefilterten Sub-Netzwerken der beiden Krankheiten.

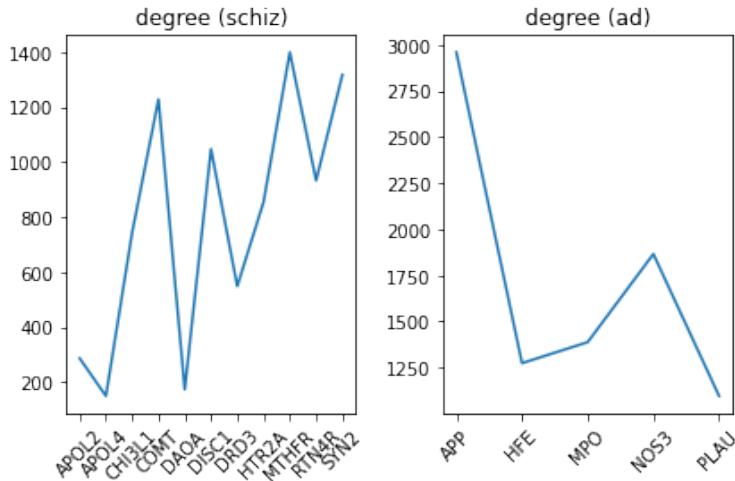


Abbildung 25: Gefiltertes Netz: Degree

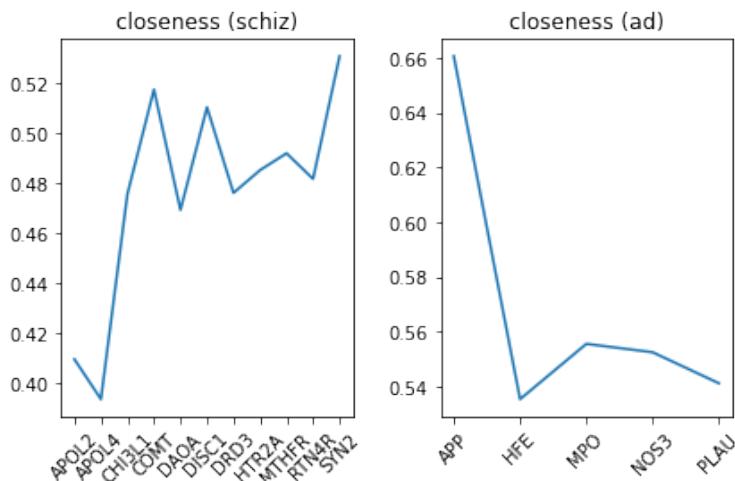


Abbildung 26: Gefiltertes Netz: Closeness

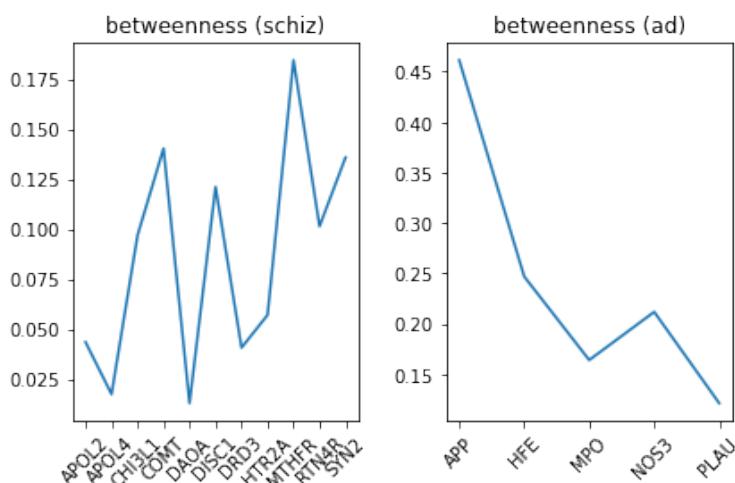


Abbildung 27: Gefiltertes Netz: Betweenness

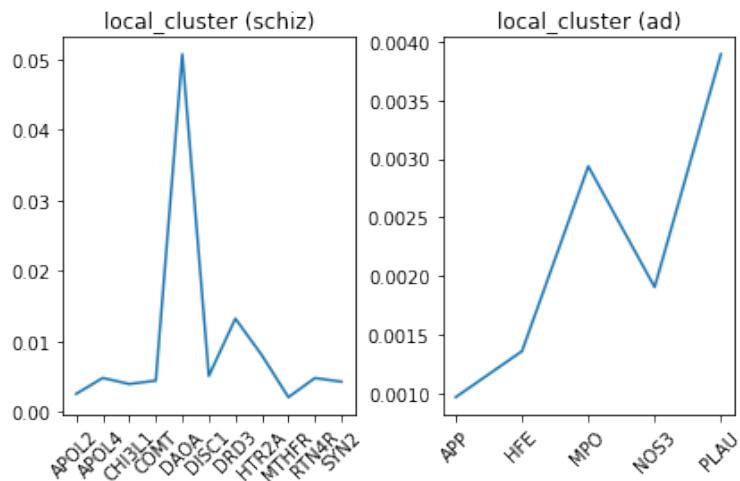


Abbildung 28: Gefiltertes Netz: Local Clustering

6.8 Messwerte relevanter Gene normiert

Die *Closeness Centrality*, *Betweenness Centrality* und das durchschnittliche *Clustering* sind normiert und als Vergleich zwischen dem gesamten Netzwerk und den gefilterten Netzen aufgeführt.

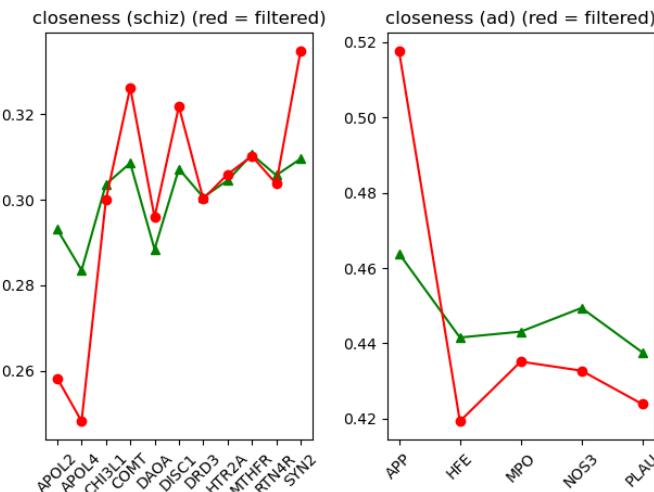


Abbildung 29: Normiert: Closeness

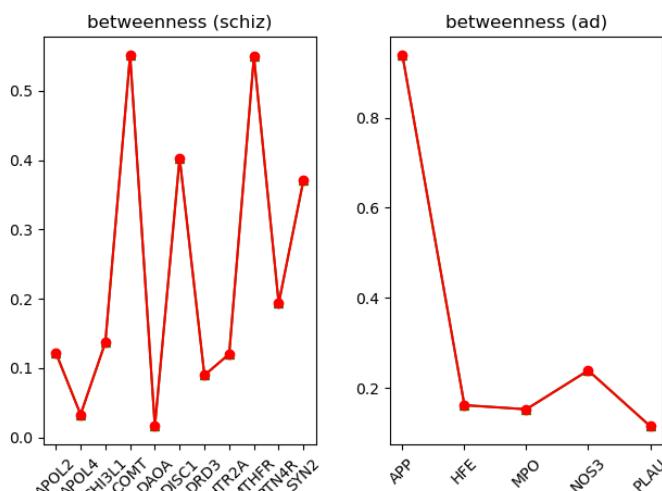


Abbildung 30: Normiert: Betweenness

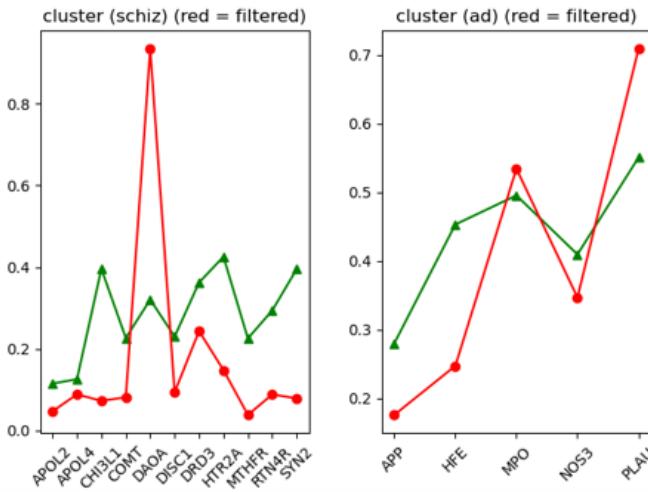


Abbildung 31: Normiert: Clustering

7 Diskussion

7.1 Auswertung der Netzwerk-Kennzahlen

Vergleichen wir die Kennzahlen aus Tabelle 3, zeigt sich, dass die separierten Netzwerke von Alzheimer (≈ 31 Prozent) und Schizophrenie (≈ 40 Prozent) der Nodes des gesamten Netzwerkes ausmachen, was einen beachtlichen Anteil darstellt. Betrachtet man die Anzahl Kanten im Vergleich ergibt sich ein ganz anderes Bild. Die Edges von Alzheimer machen gerade mal ≈ 0.19 Prozent und die Edges von Schizophrenie ≈ 0.27 Prozent des gesamten Netzes aus. Die identifizierten Gen-Interaktionen der Subnetze machen also nur einen sehr geringen Bruchteil aller festgestellten Interaktionen aus. Die beiden Subnetze weisen eine ähnliche Dichte und eine noch ähnlicheres durchschnittliches Clustering mit 0.403 und 0.384 auf. Dies deutet auf eine strukturelle Ähnlichkeiten der untersuchten Krankheiten hin.

7.2 Auswertung der einzelnen Gene

Die Einzelgraphen der relevanten Gene in Kapitel 6.3 und 6.4 weisen einen sehr unterschiedlichen Vernetzungsgrad auf. Dies reicht von Graphen, welche sehr sparse sind (Abbildung 13 MPO, Abbildung 14 APOL2) bis zu nahezu vollständigen Graphen (Abbildung 14 SYN2).

In Kapitel 6.5 wurden für die Einzelgraphen die Messwerte *Anzahl Nodes*, *Anzahl Edges*, *Closeness Centrality*, *Betweenness Centrality*, *Durchschnittliches Clustering* und der *Wiener Index* als Liniendiagramme dargestellt. Die Graphen mit einer hohen Vernetzung (SYN2) weisen charakteristisch einen tieferen Wert für den Wiener Index auf, da die Summe der kürzesten Wege sehr gering ausfällt. Auch die tiefe Betweenness Centrality ist auf die hohe Vernetzung zurückzuführen. Die Messwerte der beiden Graphen der Gene *MPO* und *APOL2*, welche nur sehr sparse sind, weisen im Vergleich sehr ähnliche Werte auf.

Führen wir diesen Vergleich mittels der in Kapitel 6.6 ersichtlichen Messwerte fort, so ist ein Unterschied erkennbar, da das Gen *APOL2*, welches in den Einzelgraphen die höchste Betweenness Centrality aufwies und aus der Sicht des gesamten Graphen nun eine eher tiefe Betweenness Centrality aufweist. Dies röhrt wohl daher, dass durch die Hinzunahme der restlichen Gene des menschlichen Genoms viele parallele Pfade zu dem Knoten *APOL2* existieren.

Für die beiden Krankheiten Alzheimer und Schizophrenie wurden alle Knoten jeweils als Subgraph aus dem gesamten Netzwerk herausgelöst und separat untersucht. Dies ist in Kapitel 6.7 ersichtlich. Im Vergleich zu den Messwerten des gesamten Netzes weisen die Messwerte der gefilterten Netze ähnliche Ausprägungen in den Werten auf, jedoch mit unterschiedlichen Amplituden. Einzig der lokale Clustering Coefficient des gefilterten Netzwerks von Schizophrenie weicht stark von dem des gesamten Netzwerks ab. Dabei ist lediglich ein einziges Gen (DAOA) mit einem hohen Clustering Coefficient vorhanden. Dieser Unterschied ist auch in den Plots des Kapitel 6.8 (Abbildung 31) gut ersichtlich. Die restlichen normierten Werte zeigen eine sehr hohe Übereinstimmung zwischen den unterschiedlichen Graphen.

7.3 Ausblick

Für weitere Forschungsarbeiten auf dem Gebiet der Netzwerk Analyse mit genetischen Netzwerken empfehlen wir, weiter auf die Scores der Kanten aus der String-DB Bezug zu nehmen. Es wäre sinnvoll, die Kantengewichte stärker hervorzuheben.

Die Berechnung des gesamten Netzwerkes dauert sehr lange, daher empfehlen wir hier, die Verwendung von **Graph-Tool** weiter anzustreben, oder sogar noch eine weitere schnellere Alternative zu testen. Eine weitere Möglichkeit wäre es, wie in unserer Arbeit, die Aufteilung des Graphen in die einzelnen Communities. Dies erfordert jedoch meistens mehr Domänen-Wissen oder bereits erforschte Zusammenhänge.

Mit den Datenbeschaffungs- und Analyse-Skripts ist nun eine solide Grundlage geschaffen, um ebenfalls weitere genetische Netze von anderen Krankheiten oder Organismen zu untersuchen. Durch Erweiterung der Analyse-Scripts um weitere Kantenauswertungen könnte eine Basis für zukünftige Forschungsarbeiten geschaffen werden.

Die Jupyter-Notebooks sowie die Basis-Daten befinden sich im Anhang dieser Arbeit.

8 Anhang

Literatur

- [1] N. C. f. B. Information (US), *Genes and Disease*. National Center for Biotechnology Information (US), 1998.
- [2] “The Human Genome Project.” <https://www.genome.gov/human-genome-project>.
- [3] A. Krischke and H. Röpcke, *Graphen und Netzwerktheorie: Grundlagen - Methoden - Anwendungen ; mit 137 Bildern und zahlreichen Beispielen*. Quantitative Methoden, Fachbuchverlag Leipzig im Carl Hanser Verlag, 2015.
- [4] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, “Fast unfolding of communities in large networks,”
- [5] A.-L. Barabasi, *Network Science*. Cambridge, United Kingdom: Cambridge University Press, 1st edition ed., Aug. 2016.
- [6] F. Emmert-Streib, S. Moutari, and M. Dehmer, *Mathematical Foundations of Data Science Using R*. De Gruyter Oldenbourg, May 2020.
- [7] “Alzheimer disease - MedGen - NCBI.”
- [8] “Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017 - The Lancet.”
- [9] T. M. Laursen, M. Nordentoft, and P. B. Mortensen, “Excess early mortality in schizophrenia,” *Annual Review of Clinical Psychology*, vol. 10, pp. 425–448, 2014.
- [10] K.-I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal, and A.-L. Barabási, “The human disease network,” *Proceedings of the National Academy of Sciences*, vol. 104, pp. 8685–8690, May 2007. Publisher: National Academy of Sciences Section: Physical Sciences.
- [11] “Schizophrenia disease - MedGen - NCBI.”
- [12] “STRING: functional protein association networks.” <https://string-db.org/>.
- [13] “graph-tool: Efficient network analysis with python.”
- [14] D. J. Watts and S. H. Strogatz, “Collective dynamics of ‘small-world’ networks,” *Nature*, vol. 393, pp. 440–442, June 1998. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 6684 Primary_atype: Research Publisher: Nature Publishing Group.
- [15] “The Structure and Function of Complex Networks | SIAM Review | Vol. 45, No. 2 | Society for Industrial and Applied Mathematics.”
- [16] Y. Mao, D. W. Fisher, S. Yang, R. M. Keszycki, and H. Dong, “Protein-protein interactions underlying the behavioral and psychological symptoms of dementia (BPSD) and Alzheimer’s disease,” *PloS One*, vol. 15, no. 1, p. e0226021, 2020.
- [17] EMBL-EBI, “Protein-protein interaction networks | Network analysis of protein interaction data.” <https://www.ebi.ac.uk/training/online/courses/network-analysis-of-protein-interaction-data-an-introduction/protein-protein-interaction-networks/>.
- [18] J. Xia, M. J. Benner, and R. E. W. Hancock, “NetworkAnalyst - integrative approaches for protein–protein interaction network analysis and visual exploration,” *Nucleic Acids Research*, vol. 42, pp. W167–W174, July 2014.
- [19] J. Piñero, J. Saich, F. Sanz, and L. I. Furlong, “The DisGeNET cytoscape app: Exploring and visualizing disease genomics data,” *Computational and Structural Biotechnology Journal*, vol. 19, pp. 2960–2967, 2021.
- [20] J. S. Goldman, S. E. Hahn, J. W. Catania, S. Larusse-Eckert, M. B. Butson, M. Rumbaugh, M. N. Strecker, J. S. Roberts, W. Burke, R. Mayeux, and T. Bird, “Genetic counseling and testing for Alzheimer disease: Joint practice guidelines of the American College of Medical Genetics and the National Society of Genetic Counselors,” *Genetics in Medicine*, vol. 13, pp. 597–605, June 2011.
- [21] R. Karbalaei, M. Allahyari, M. Rezaei-Tavirani, H. Asadzadeh-Aghdaei, and M. R. Zali, “Protein-protein interaction analysis of Alzheimer’s disease and NAFLD based on systems

biology methods unhide common ancestor pathways,” *Gastroenterology and Hepatology from Bed to Bench*, vol. 11, no. 1, pp. 27–33, 2018.

Abbildungsverzeichnis

1	Zelle Chromoson DNA	1
2	Unterschiedliche Graphdarstellungen [3]	3
3	Adjazenzmatrix	4
4	Einfacher Graph	5
5	MedGen Eintrag in der NCBI Datenbank	10
6	String-DB Download Verzeichnis	10
7	Das Lokale Netzwerk des APP Genes	11
8	Die Legende der Kanten vom APP Gen	11
9	Dateistruktur nach dem Aufbereiten der Daten	12
10	Homo Sapiens Genom Netzwerk	17
11	Netzwerk der Alzheimer Gene	18
12	Netzwerk der Schizophrenie Gene	18
13	Involvierte Gene bei Alzheimer	19
14	Involvierte Gene bei Schizophrenie (1)	20
15	Anzahl Nodes	21
16	Anzahl Edges	21
17	Closeness	22
18	Betweenness	22
19	Avg. Clustering Coefficient	22
20	Wiener Index	23
21	Ganzes Netz: Degree	23
22	Ganzes Netz: Closeness	24
23	Ganzes Netz: Betweenness	24
24	Ganzes Netz: Local Clustering	24
25	Gefiltertes Netz: Degree	25
26	Gefiltertes Netz: Closeness	25
27	Gefiltertes Netz: Betweenness	25
28	Gefiltertes Netz: Local Clustering	26
29	Normiert: Closeness	26
30	Normiert: Betweenness	26
31	Normiert: Clustering	27

Tabellenverzeichnis

1	Auflistung der betroffenen Gene von Alzheimer	8
2	Auflistung der betroffenen Gene von Schizophrenie	8
3	Übersicht der Netzwerk Metriken	17

Code Listings

1	Konvertiere StringDB TSV-Files zu CSV	12
2	Schreibe Lesbare Labels in die Source und Target Spalten der Pandas Dataframes	12
3	Berechne Globale Netzwerkmasse mit networkx	13
4	Berechne Graph-Masse aus den einzelnen Gen-Netzwerken	13
5	Vergleiche alle Graph-Masse vom Schizophrenie Netzwerke mit dem Alzheimer Netzwerk	14
6	Graph-Tool Import Network and Print Graph Stats	14
7	Ausgabe der Zentralitätsmasse	15
8	Berechnen und ausgeben der Clustering Werte	15
9	Communities mit Minimize Blockmodel	16

Beschreibung der Gene aus MedGen

Alzheimer Disease

Genes:

APP: Amyloid-beta A4 protein; N-APP binds TNFRSF21 triggering caspase activation and degeneration of both neuronal cell bodies (via caspase-3) and axons (via caspase-6); Endogenous ligands

HFE: Hereditary hemochromatosis protein; Binds to transferrin receptor (TFR) and reduces its affinity for iron-loaded transferrin; Belongs to the MHC class I family

MPO: Myeloperoxidase; Part of the host defense system of polymorphonuclear leukocytes. It is responsible for microbicidal activity against a wide range of organisms. In the stimulated PMN, MPO catalyzes the production of hypohalous acids, primarily hypochlorous acid in physiologic situations, and other toxic intermediates that greatly enhance PMN microbicidal activity; Belongs to the peroxidase family. XPO subfamily

NOS3: Myeloperoxidase; Nitric oxide is a reactive free radical which acts as a biologic mediator in several processes, including neurotransmission and antimicrobial and antitumoral activities. Nitric oxide is synthesized from L-arginine by nitric oxide synthases. Variations in this gene are associated with susceptibility to coronary spasm. Alternative splicing and the use of alternative promoters results in multiple transcript variants. [provided by RefSeq, Oct 2016]

PLAU: Urokinase-type plasminogen activator; Specifically cleaves the zymogen plasminogen to form the active enzyme plasmin

Related Genes:

ABCA7: The protein encoded by this gene is a member of the superfamily of ATP-binding cassette (ABC) transporters. ABC proteins transport various molecules across extra- and intra-cellular membranes. ABC genes are divided into seven distinct subfamilies (ABC1, MDR/TAP, MRP, ALD, OABP, GCN20, White). This protein is a member of the ABC1 subfamily. Members of the ABC1 subfamily comprise the only major ABC subfamily found exclusively in multicellular eukaryotes. This full transporter has been detected predominantly in myelo-lymphatic tissues with the highest expression in peripheral leukocytes, thymus, spleen, and bone marrow. The function of this protein is not yet known; however, the expression pattern suggests a role in lipid homeostasis in cells of the immune system. [provided by RefSeq, Jul 2008]

PSEN1, PSEN2: Alzheimer's disease (AD) patients with an inherited form of the disease carry mutations in the presenilin proteins (PSEN1 or PSEN2) or the amyloid precursor protein (APP). These disease-linked mutations result in increased production of the longer form of amyloid-beta (main component of amyloid deposits found in AD brains). Presenilins are postulated to regulate APP processing through their effects on gamma-secretase, an enzyme that cleaves APP. Also, it is thought that the presenilins are involved in the cleavage of the Notch receptor such that, they either directly regulate gamma-secretase activity, or themselves act as protease enzymes. Two alternatively spliced transcript variants encoding different isoforms of PSEN2 have been identified. [provided by RefSeq, Jul 2008]

APOE: The protein encoded by this gene is a major apoprotein of the chylomicron. It binds to a specific liver and peripheral cell receptor, and is essential for the normal catabolism of triglyceride-rich lipoprotein constituents. This gene maps to chromosome 19 in a cluster with the related apolipoprotein C1 and C2 genes. Mutations in this gene result in familial dysbetalipoproteinemia, or type III hyperlipoproteinemia (HLP III), in which increased plasma cholesterol and triglycerides are the consequence of impaired clearance of chylomicron and VLDL remnants. [provided by RefSeq, Jun 2016]

Schizophrenia

Genes:

APOL2: This gene is a member of the apolipoprotein L gene family. The encoded protein is found in the cytoplasm, where it may affect the movement of lipids or allow the binding of lipids to organelles. Two transcript variants encoding the same protein have been found for this gene. [provided by RefSeq, Jul 2008]

APOL4: This gene encodes a member of the apolipoprotein L family. The encoded protein may play a role in lipid exchange and transport throughout the body, as well as in reverse choleste-

rol transport from peripheral cells to the liver. Alternative splicing results in multiple transcript variants. [provided by RefSeq, Sep 2020]

CHI3L1: Chitinases catalyze the hydrolysis of chitin, which is an abundant glycopolymer found in insect exoskeletons and fungal cell walls. The glycoside hydrolase 18 family of chitinases includes eight human family members. This gene encodes a glycoprotein member of the glycosyl hydrolase 18 family. The protein lacks chitinase activity and is secreted by activated macrophages, chondrocytes, neutrophils and synovial cells. The protein is thought to play a role in the process of inflammation and tissue remodeling. [provided by RefSeq, Sep 2009]

COMT: Chitinases catalyze the hydrolysis of chitin, which is an abundant glycopolymer found in insect exoskeletons and fungal cell walls. The glycoside hydrolase 18 family of chitinases includes eight human family members. This gene encodes a glycoprotein member of the glycosyl hydrolase 18 family. The protein lacks chitinase activity and is secreted by activated macrophages, chondrocytes, neutrophils and synovial cells. The protein is thought to play a role in the process of inflammation and tissue remodeling. [provided by RefSeq, Sep 2009]

DAOA: The protein encoded by this gene is a major apoprotein of the chylomicron. It binds to a specific liver and peripheral cell receptor, and is essential for the normal catabolism of triglyceride-rich lipoprotein constituents. This gene maps to chromosome 19 in a cluster with the related apolipoprotein C1 and C2 genes. Mutations in this gene result in familial dysbetalipoproteinemia, or type III hyperlipoproteinemia (HLP III), in which increased plasma cholesterol and triglycerides are the consequence of impaired clearance of chylomicron and VLDL remnants. [provided by RefSeq, Jun 2016]

DISC2: DISC2 is thought to specify a noncoding RNA molecule antisense to DISC1 (MIM 605210). Both genes were found to be disrupted by a translocation in a large schizophrenia (MIM 181500) kindred. [supplied by OMIM, Jul 2002]

DRD3: This gene encodes the D3 subtype of the five (D1-D5) dopamine receptors. The activity of the D3 subtype receptor is mediated by G proteins which inhibit adenylyl cyclase. This receptor is localized to the limbic areas of the brain, which are associated with cognitive, emotional, and endocrine functions. Genetic variation in this gene may be associated with susceptibility to hereditary essential tremor 1. Alternative splicing of this gene results in transcript variants encoding different isoforms, although some variants may be subject to nonsense-mediated decay (NMD). [provided by RefSeq, Jul 2008]

HRT2A: This gene encodes one of the receptors for serotonin, a neurotransmitter with many roles. Mutations in this gene are associated with susceptibility to schizophrenia and obsessive-compulsive disorder, and are also associated with response to the antidepressant citalopram in patients with major depressive disorder (MDD). MDD patients who also have a mutation in intron 2 of this gene show a significantly reduced response to citalopram as this antidepressant downregulates expression of this gene. Multiple transcript variants encoding different isoforms have been found for this gene. [provided by RefSeq, Sep 2009]

MTHFR: The protein encoded by this gene catalyzes the conversion of 5,10-methylenetetrahydrofolate to 5-methyltetrahydrofolate, a co-substrate for homocysteine remethylation to methionine. Genetic variation in this gene influences susceptibility to occlusive vascular disease, neural tube defects, colon cancer and acute leukemia, and mutations in this gene are associated with methylenetetrahydrofolate reductase deficiency. [provided by RefSeq, Oct 2009]

RTN4R: This gene encodes the receptor for reticulon 4, oligodendrocyte myelin glycoprotein and myelin-associated glycoprotein. This receptor mediates axonal growth inhibition and may play a role in regulating axonal regeneration and plasticity in the adult central nervous system. [provided by RefSeq, Jul 2008]

SYN2: This gene is a member of the synapsin gene family. Synapsins encode neuronal phosphoproteins which associate with the cytoplasmic surface of synaptic vesicles. Family members are characterized by common protein domains, and they are implicated in synaptogenesis and the modulation of neurotransmitter release, suggesting a potential role in several neuropsychiatric diseases. This member of the synapsin family encodes a neuron-specific phosphoprotein that selectively binds to small synaptic vesicles in the presynaptic nerve terminal. Polymorphisms in this gene are associated with abnormal presynaptic function and related neuronal disorders, including autism, epilepsy, bipolar disorder and schizophrenia. Alternative splicing of this gene results in multiple transcript variants. The tissue inhibitor of metalloproteinase 4 gene is located within an intron of this gene and is transcribed in the opposite direction. [provided by RefSeq, Feb 2014]

Related Genes:

SHANK3: This gene is a member of the Shank gene family. Shank proteins are multidomain scaffold proteins of the postsynaptic density that connect neurotransmitter receptors, ion channels, and other membrane proteins to the actin cytoskeleton and G-protein-coupled signaling pathways. Shank proteins also play a role in synapse formation and dendritic spine maturation. Mutations in this gene are a cause of autism spectrum disorder (ASD), which is characterized by impairments in social interaction and communication, and restricted behavioral patterns and interests. Mutations in this gene also cause schizophrenia type 15, and are a major causative factor in the neurological symptoms of 22q13.3 deletion syndrome, which is also known as Phelan-McDermid syndrome. Additional isoforms have been described for this gene but they have not yet been experimentally verified. [provided by RefSeq, Mar 2012]

DISC1: This gene encodes a protein with multiple coiled coil motifs which is located in the nucleus, cytoplasm and mitochondria. The protein is involved in neurite outgrowth and cortical development through its interaction with other proteins. This gene is disrupted in a t(1;11)(q42.1;q14.3) translocation which segregates with schizophrenia and related psychiatric disorders in a large Scottish family. Alternate transcriptional splice variants, encoding different isoforms, have been characterized. [provided by RefSeq, Jul 2008] **RBM12:** This gene encodes a protein that contains several RNA-binding motifs, potential transmembrane domains, and proline-rich regions. This gene and the gene for copine I overlap at map location 20q11.21. Alternative splicing in the 5' UTR results in four transcript variants. All variants encode the same protein. [provided by RefSeq, Nov 2010]

NRXN1: This gene encodes a single-pass type I membrane protein that belongs to the neurexin family. Neurexins are cell-surface receptors that bind neuroligins to form Ca(2+)-dependent neurexin/neuroligin complexes at synapses in the central nervous system. This complex is required for efficient neurotransmission and is involved in the formation of synaptic contacts. Three members of this gene family have been studied in detail and are estimated to generate over 3,000 variants through the use of two alternative promoters (alpha and beta) and extensive alternative splicing in each family member. Recently, a third promoter (gamma) was identified for this gene in the 3' region. Mutations in this gene are associated with Pitt-Hopkins-like syndrome-2 and may contribute to susceptibility to schizophrenia. [provided by RefSeq, Aug 2016]

SLC1A1: This gene encodes a member of the high-affinity glutamate transporters that play an essential role in transporting glutamate across plasma membranes. In brain, these transporters are crucial in terminating the postsynaptic action of the neurotransmitter glutamate, and in maintaining extracellular glutamate concentrations below neurotoxic levels. This transporter also transports aspartate, and mutations in this gene are thought to cause dicarboxylic amino aciduria, also known as glutamate-aspartate transport defect. [provided by RefSeq, Mar 2010]

PRODH: This gene encodes a mitochondrial protein that catalyzes the first step in proline degradation. Mutations in this gene are associated with hyperprolinemia type 1 and susceptibility to schizophrenia 4 (SCZD4). This gene is located on chromosome 22q11.21, a region which has also been associated with the contiguous gene deletion syndromes, DiGeorge and CATCH22. Alternatively spliced transcript variants encoding different isoforms have been found for this gene. [provided by RefSeq, Aug 2010]

NRG1: The protein encoded by this gene is a membrane glycoprotein that mediates cell-cell signaling and plays a critical role in the growth and development of multiple organ systems. An extraordinary variety of different isoforms are produced from this gene through alternative promoter usage and splicing. These isoforms are expressed in a tissue-specific manner and differ significantly in their structure, and are classified as types I, II, III, IV, V and VI. Dysregulation of this gene has been linked to diseases such as cancer, schizophrenia, and bipolar disorder (BPD). [provided by RefSeq, Apr 2016]