Bericht zur Funktionellen Klassifikation

Praktikum Genomorientierte Bioinformatik

von Anton Smirnov & Uli Köhler - Gruppe 7

17. Juli 2014

Einleitung

Seit dem Aufkommen von High-Throughput-Sequencing steht die Medikamentenentwicklung vor einer neuen Herausforderung:

Da Genome, cDNAs und EST immer schneller und kostengünstiger sequenziert weren können, gibt es viele Millionen Genprodukte als potentielle Targets für Medikamente.

Laborexperimente, die unter diesen Genprodukten das richtige Target identifizieren, sind aufgrund der großen Menge an Kandidaten und der teuren Experimente vollkommen unrealistisch - daher muss die Bioinformatik *in-silico* die potentiellen Targets auf einige wenige Kandidaten reduzieren, die dann im Labor untersucht werden können.

Da bei der Medikamentensuche gerade Proteine mit einer bestimmten Funktion ermittelt werden müssen, kann dazu versucht werden, Proteinen möglichst genau ihre Funktion zuzuordnen. Hierfür können vor allem Kriterien wie die Orthologie zu anderen Proteinen herangezogen werden, deren Funktion bekannt ist. Um solche Familien von orthologen Proteinen zu klassifizieren, wurden im Laufe der Zeit verschiedene Bioinformatik-Datenbanken veröffentlicht, die es ermöglichen, neue – womöglich nur vorhergesagte – Proteine, auf potentielle Funktionen zu untersuchen, und somit auch weitere Informationen über die Genprodukte, wie beispielsweise die Lokation in der Zelle, zu gewinnen. Damit kann die Zahl der Laborexperimente, die für eine Drug-Target-Suche notwendig ist, stark eingeschränkt werden.

Im vorliegenden Bericht untersuchen und vergleichen wir exemplarisch am Protein *Nesprin-1 / SYNE1* verschiedene Datenbank- und literaturbasierte Methoden zur Funktionellen Klassifikation.

Aufgabe 1

Für den Blat-Search wird eine Proteinsequenz im FASTA-Format benötigt. Dafür muss zuerst die längste Isoform identifiziert werden. UniProt listet das gegebene Protein unter `http://www.uniprot.org/uniprot/Q8NF91`. Die längste Proteinvariante ist mit 8797 AS die Isoform 1, deren Sequenz unter `http://www.uniprot.org/uniprot/Q8NF91.fasta`. Beim USCS Genome Browser `http://genome.ucsc.edu` muss für die Blat Search ein Assembly des Humangenoms gewählt werden. Es ist davon auszugehen, dass das aktuellste Assembly *Feb. 2009 (GRCh37/hg19)* die beste Datenqualität hat, daher haben wir ebendieses Assembly für die Suche gewählt.

Tabelle 1: Ergebnisse der UniProt-Suche

Score	**Start**	**End**	**Identity**	**Chromosom**	**Strand**	**Start**	**End**	**Span**
23834	1	7993	100.0	6	+-	152477044	152949466	472423
287	260	7995	75.0	14	++	64428242	64675670	247429
24	1371	1378	100.0	14	+-	56314333	56314356	24
15	5813	5817	100.0	3	++	181604808	181604822	15
12	4028	4033	83.4	6	+-	152660628	152660645	18
12	1311	1314	100.0	14	+-	56314525	56314536	12
12	5782	5785	100.0	3	++	181604715	181604726	12
12	5866	5869	100.0	3	++	181604967	181604978	12

Der erste Hit hat einen sehr hohen Score und kann deswegen als signifikant betrachtet werden. Die weiteren Hits haben nicht nur eine sehr geringe Länge, sondern insbesondere auch einen extrem geringen Score. Die Sequenzidentität ist bei den weiteren Hits zwar teilweise 100%, dies liegt allerdings an der sehr geringen Länge der Hits.

Die Blat-Ergebnisse zeigen daher den Genlocus auf HSA Chromosom 6, Position 152477044-152949466.

Für die Analyse der Splicevarianten muss nun ebendieser Locus im UCSC Genome Browser geöffnet werden. Hierfür muss darauf geachtet werden, dass dasselbe Assembly verwendet wird, allerdings enthält die Auflistung der Blat-Resultate bereits einen *browse*-Link. In Abbildung 1 ist zu sehen, dass mit der Default-Filterung sehr viele Informationen angezeigt werden.

Darum wurden zuerst alle Information per *hide all* ausgeblendet, und danach lediglich die *Blat sequence = dense* sowie die *Human mRNAs = Full* eingeblendet, da die mRNAs (nicht aber die prä-mRNAs) eben die Splicevarianten darstellen. Das gefilterte Resultat ist in 2 zu sehen.

Allerdings sind jetzt alle mRNAs aus dem entsprechenden genomischen Bereich zu sehen. Die mRNAs können also auch andere Genprodukte haben als SYNE1. Daher muss zuerst ein Filter angewendet werden, der nur diejenigen mRNAs *included*, die das Proteinprodukt=*SYNE1* haben. Dabei trat das Problem auf, dass nur 3 sich nicht überlappende mRNAs (vermutlich Proteindomänen von SYNE1) angezeigt wurden (BC039121, BC150289 und BC090927). Nach Anwendung des Filters Gene=*SYNE1* statt dem Proteinproduktfilter zeigen sich die zu erwartenden sich überlappenden mRNAs, die die Splicevarianten darstellen.

Bei näherer Betrachtung ergibt sich, dass die zuvor nicht angezeigten mRNAs als Proteinprodukt *nesprin-1* eingetragen sind. Dies ist allerdings, wie aus UniProt/SwissProt ersichtlich, ein Alias für SYNE1. Wir vermuten daher, dass Protein-Aliases im UCSC Genome Browser, möglicherweise aber auch nur in GenBank bzw. der UCSC-GenBank-Schnittstelle nicht korrekt gemappt werden. Bei komplexen Proteinsuchen kann dies von Nachteil sein.

Übrig bleiben, wie in 3 sichtbar, die folgenden mRNAs – die mit (A)-(D) markierten Gruppen könnten auf 4 verschiedene Domänen hindeuten, da sie sich nicht (nicht einmal annähernd) überlappen:

- BC090927 (A)
- JQ740786 (B)

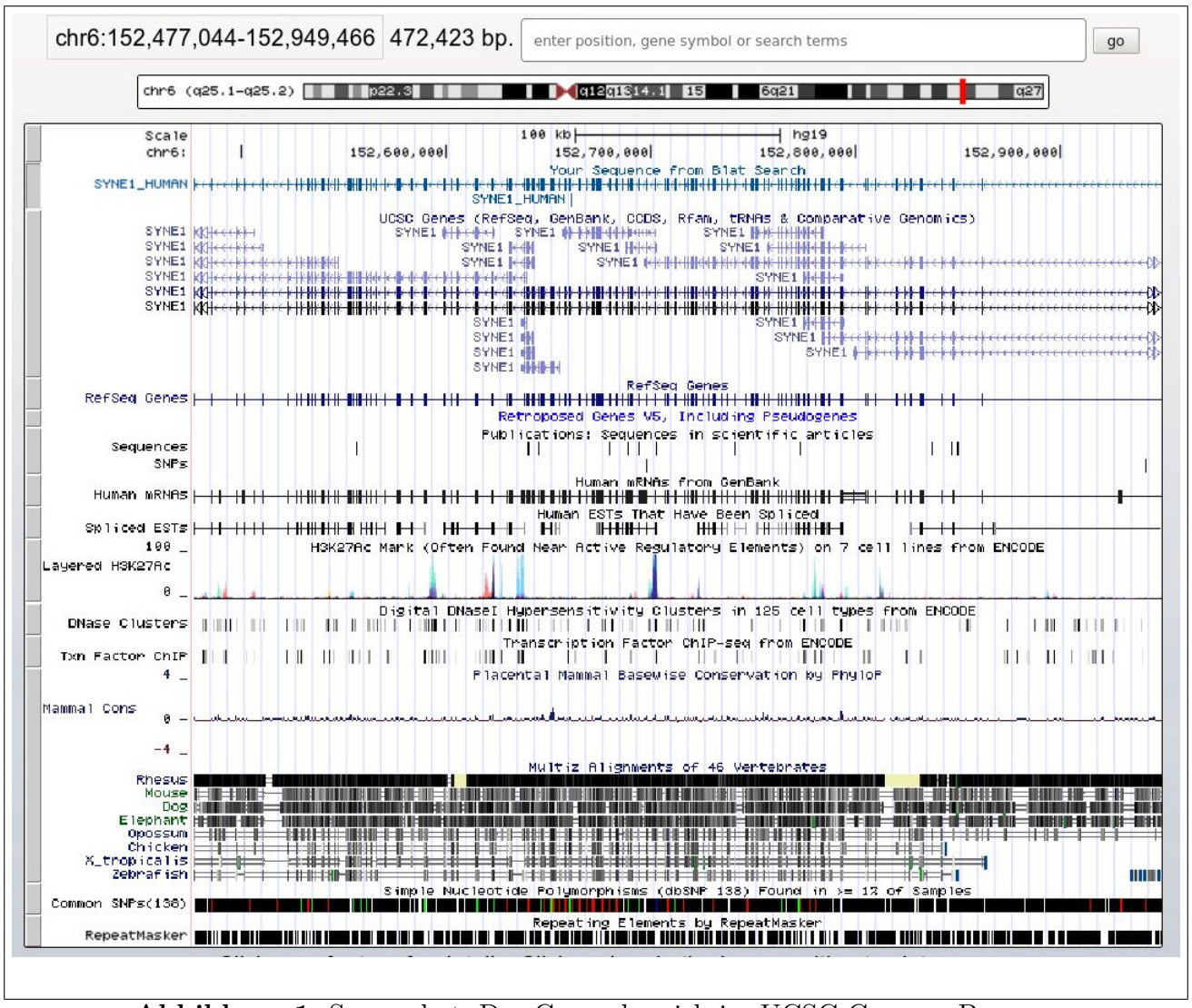

Abbildung 1: Screenshot: Der Genombereich im UCSC Genome Browser

- JQ740784 (B)
- BC150289 (B)
- JQ754365 (B)
- JQ754364 (B)
- JQ740785 (B)
- BC039121 (C)
- JQ740783 (D)

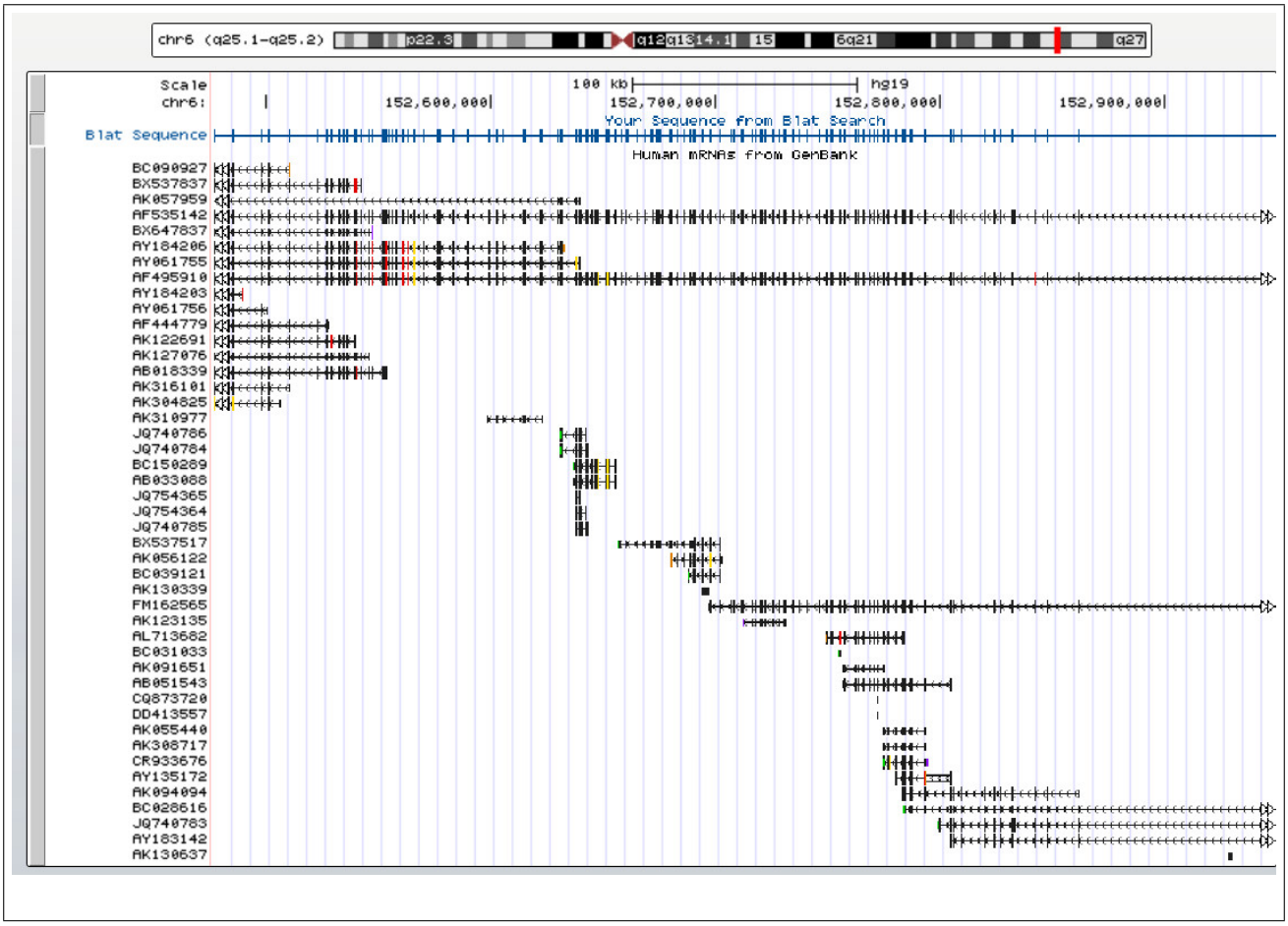

Abbildung 2: Screenshot: Der Genombereich im UCSC Genome Browser, gefiltert

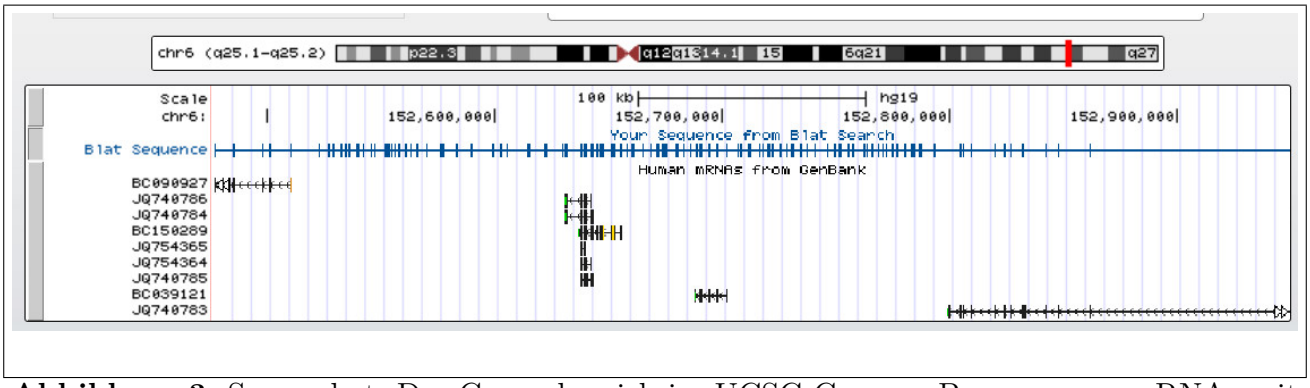

Abbildung 3: Screenshot: Der Genombereich im UCSC Genome Browser, nur mRNAs mit SYNE1-Genprodukt

Aufgabe 2

Suche in UniProt: *gene:syne1 AND organism:human*, restrict term „SYNE1 “to gene name, Entry Q8NF91, 9 Isoformen. Copypaste aus Align-Feld in Clustal Omega `http://www.ebi.ac.uk/Tools/msa/clustalo/` mit `http://www.uniprot.org/uniprot/Q63128.fasta`. Das Multiple

Sequenzalignment konnte nach wenigen Minuten Wartezeit über die folgende URL (die möglicherweise nur temporär gültig ist) abgerufen werden. `http://www.ebi.ac.uk/Tools/services/web/toolresult.ebi?jobId=clustalo-I20131105-144703-0076-68492607-pg` In 4 ist das

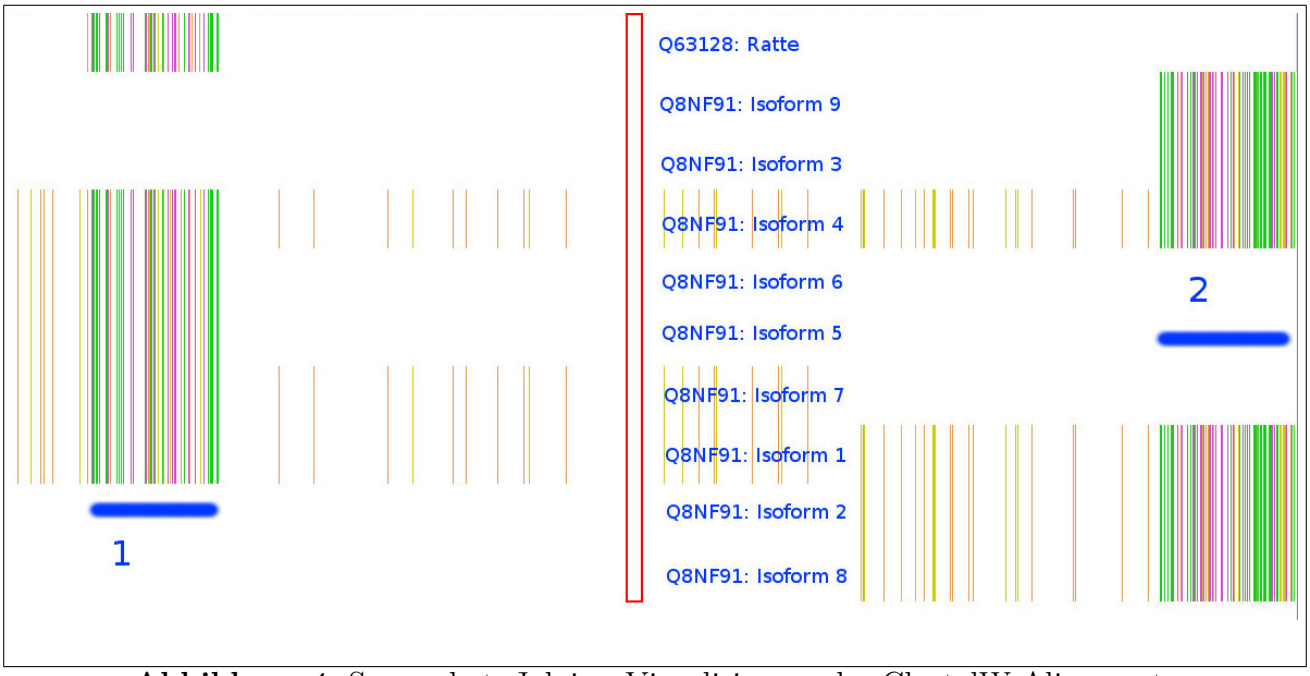

Abbildung 4: Screenshot: Jalview-Visualisierung des ClustalW-Alignments

mit Jalview visualisierte multiple Sequenzalignment zu sehen. Insgesamt existiert kein Bereich im Protein, der über alle untersuchten Proteine vollständig konserviert ist. Bei der Betrachtung muss beachtet werden, dass einige Isoformen signifikant kürzer sind als die längste Isoform Q8NF91-1, daher sind beispielsweise bei Isoform 2 die großen Bereiche ohne Matches vor allem dadurch zu erklären, dass die entsprechende kurze Isoform lediglich lang genug ist, um Matches im Bereich 2 zu haben.

Die Isoformen des Human-SYNE1 haben teilweise erhebliche Unterschiede.

Wie aus der Zusammenfassung `http://www.uniprot.org/uniprot/Q8NF91#section_seq` ersichtlich ist, sind bei vielen Isoformen im Vergleich zu der längsten Isoform fehlende Bereiche vorhanden, was auch in UniProt ersichtlich ist.

Ähnlich zueinander sind insbesondere die Isoformen 2 und 8 (die ersten ~5500 AS fehlen) sowie Isoform 3 und 9 (die ersten ~7800 AS fehlen).

UniProt annotiert zudem den Bereich 2 als „Alpha-Domäne“, die wiederum die KASH-Domäne enthält – diese enthält eine Transmembran-Domäne.

Aufgabe 3

Die neun Isoformen von Q8NF91 sind vergleichsweise wenig stark konserviert. Nur einige Bereiche sind stark konserviert, wobei es keinen Bereich gibt, der bei allen neun Isoformen konserviert ist. Die Bereiche 1 und 2 sind bei einer signifikanten Anzahl von Isoformen konserviert, insbesondere fällt jedoch auf, dass der Bereich 1 auch im Protein aus *Rattus Norvegicus* konserviert ist. Im

Humanprotein sprechen diese Phänomene für eine domänenbezogene Konservierung. Der Bereich 1 scheint jedoch für eine funktionelle Klassifizierung am relevantesten zu sein, da er auch über Organismengrenzen hinaus konserviert ist.

Ein Vergleich mit orthologen Sequenzen kann sinnvoll sein, da die Funktion bei Orthologen sehr stark konserviert ist – durch ein multiples Alignment aller bekannten Orthologen zeigen sich stark konservierte Regionen, die wieder auf die Bereiche hinweisen, die für die Funktion des Proteins essentiell sind.

Es ist auch moglich, dass wie im Falle von SYNE1 innerhalb des Gens mehrere funktionelle Domänen vorhanden sind. Zwischen den Orthologen kann somit die Funktion von einem Subset der Domänen konserviert sein. Evolutionär kann die Gesamtfunktion des Gens somit adaptiert werden.

Aufgabe 4

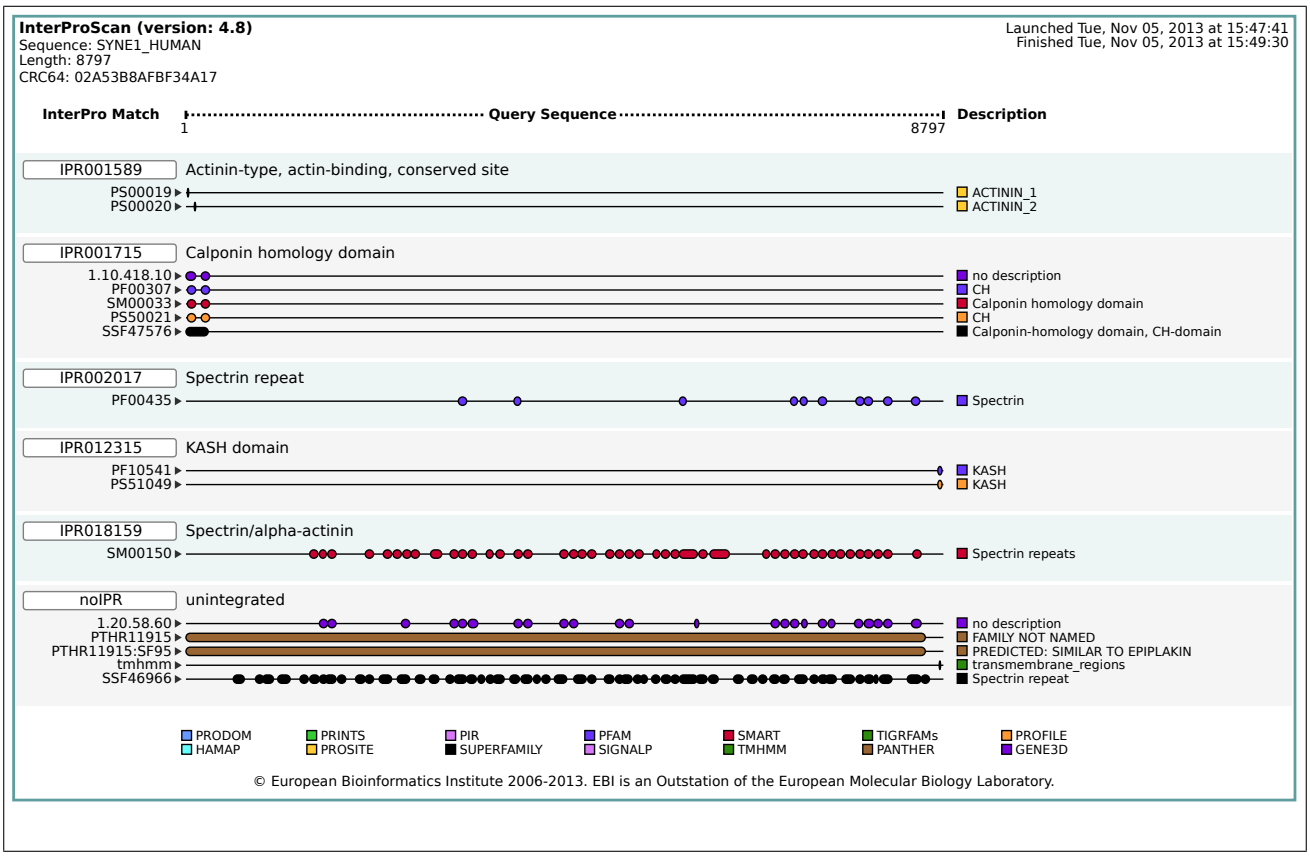

Abbildung 5: Screenshot: InterProScan über die Q8NF91-1 Isoform

Die längste Variante ist, wie in Aufgabe 1 begründet, die Isoform 1 des UniProt-Proteins Q8NF91.

Bei der Betrachtung der Resultate fällt zuerst auf, dass die KASH-Domain, die bereits in Aufgabe 2 in UniProt aufgefallen war, auch im InterProScan angezeigt wird.

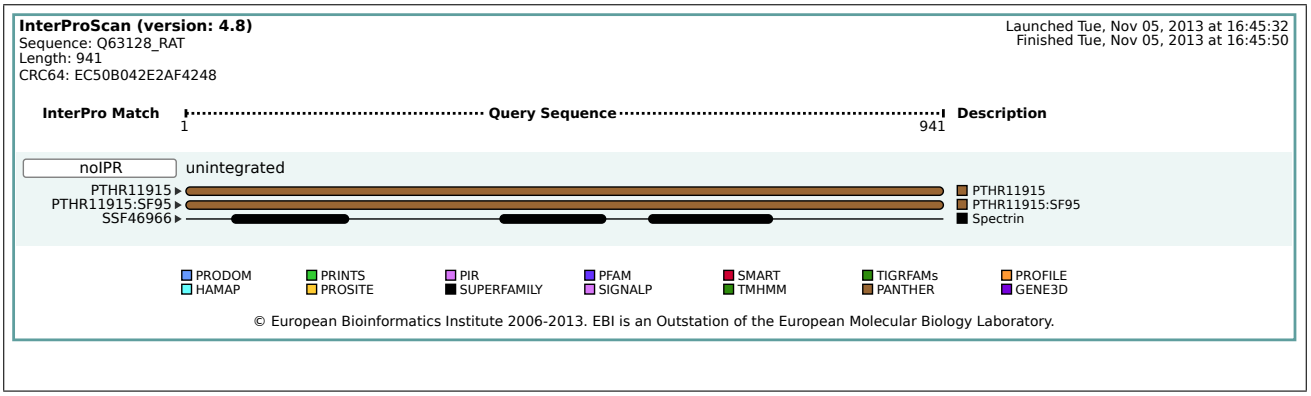

Abbildung 6: Screenshot: InterProScan über Q63128_RAT

Allerdings scheint der stark konservierte Bereich 2 (siehe Aufgabe 2 4) bei der durchgeführten InterProScan-Suche keinerlei entsprechende Domänen zu haben. Es gibt zwar eine signifikante Zahl weiterer Domänen, unter anderem auch die „Calponin Homology Domain“, die sich allerdings in der Proteinsequenz vor dem Bereich 2 befindet.

Aus diesem Grund wurde die InterProScan-Suche mit dem *Rattus Norvegicus*-Protein wiederholt (siehe 6.

Hier fällt auf, dass die Anzahl der Hits sehr gering ist – insbesondere sind die beiden Hits PTHR11915 und PTHR11915:SF95 nicht konklusiv, da sie auch im Human-IPR-Scanresultat über die gesamte Länge der Sequenz auftreten. Auch die drei Spectrin-Domänen, aus dem *Rattus Norvegicus*-Scan zeigen sich in großer Stückzahl beim Q8NF91-Scan und können daher nicht als konklusiv angesehen werden.

Zudem ist auffällig, dass die einzigen Hits bei *Rattus Norvegicus* keinen Interpro-Identifier besitzen – vermutlich, da sich bisher nicht manuell reviewt wurden.

Diese Evidenzen könnten dafür sprechen, dass eine bisher nicht annotierte Domäne in SYNE1 enthalten ist (ebendie in Bereich 1). Da Q63128 in Uniprot bisher nicht manuell reviewt wurde, ist nicht zu erwarten, dass ein systematischer Fehler bei unserer Analyse zu der konservierten Domäne geführt haben.

Aufgabe 5

Für die Annotationen wurden die Texte der angegebenen Publikationen herangezogen. Biologisch relevante Termini wurden GO Terms zugeordnet.[1]

Annotation der Isoform 1 von SYNE1

Abstract von PMID 19874786, die Experimente beziehen sich auf Mäuse, aufgrund der Ähnlichkeit des Proteins wird aber auf *Homo Sapiens* geschlossen:

- *GO:0003779* actin binding

[1] Mithilfe der in 1 vorgestellten Software wurde unsere Annotation gegenüber der am vergangenen Dienstag abgegebenen Version noch geringfügig verbessert. Wir bitten, dies zu entschuldigen.

- *GO:0034993* SUN-KASH complex
- *GO:0022008* neurogenesis
- *GO:0001764* neuron migration

GO-Terme wie *brain development* wurden nicht getrennt annotiert, da diese Terme in anderen, spezifischeren Termen wie *neurogenesis* bereits (als Ancestor) enthalten sind. Introduction von PMID 19874786:

- *GO:0034993* SUN-KASH complex
- *GO:0022008* neurogenesis
- *GO:0001764* neuron migration
- *GO:0090286* cytoskeletal anchoring at nuclear membrane
- *GO:0005640* nuclear outer membrane

Results von PMID 19874786:

- *GO:0021670* lateral ventricle development
- *GO:0021819* layer formation in cerebral cortex
- *GO:0021766* hippocampus development
- *GO:0021694* cerebellar Purkinje cell layer formation
- *GO:0030901* midbrain development
- *GO:0003360* brainstem development
- *GO:0021766* dynactin binding
- *GO:0005635* nuclear envelope
- *GO:0005737* cytoplasm (Aufgrund der direction PPI mit Dynein/Dynaktin im Cytoplasma)

Annotation von CPG2

PMID 15541315:

- *GO:2000369* regulation of clathrin-mediated endocytosis
- *GO:0031623* receptor internalization
- *GO:0048167* regulation of synaptic plasticity
- *GO:0043197* dendritic spine

- *GO:0045211* postsynaptic membrane
- *GO:0014069* postsynaptic density
- *GO:2000311* regulation of alpha-amino-3-hydroxy-5-methyl-4-isoxazole propionate selective glutamate receptor activity
- *GO:2000310* regulation of N-methyl-D-aspartate selective glutamate receptor activity

Aufgabe 6

Über die Suche von Proteindomänen können Proteinfamilien identifiziert werden, zu denen das gesuchte Protein gehört. Die Proteinfunktion kann durch Proteindomänen somit nicht direkt abgeleitet werden, sondern nur aufgrund der Orthologie innerhalb einer Familie. Dafür ist es allerdings notwendig, dass mindestens eine der Domänen des Proteins bekannt (und funktionell annotiert) ist – ansonsten können keine Annotationen abgeleitet werden. Zudem können Proteindomänen nur für einen kleinen Teil der GO-Terms Evidenzen liefern – das liegt daran, dass sich nicht alle (GO-)Annotation aus der Proteinstruktur ableiten lassen. Beispielsweise ist nur aufgrund der Zugehörigkeit zu einer Proteinfamilie nicht bekannt, in welchem Zellkompartiment das Protein aktiv ist, oder ob es wie im Fall von CPG2 mit *clathrin-coated vesicles* interagiert. Je komplexer die tatsächliche biologische Funktion eines zu untersuchenden Proteins, desto wichtiger ist es allerdings, sowohl die strukturbasierten Annotationen (v.a. Proteindomänen) als auch andere (meist literaturbasierte) Annotationen einzubringen. Insbesondere lässt sich vermuten, dass für eine korrekte *in-silico*-Vorhersage von Proteinfunktionen in der Regel weder die literaturbasierte- noch die Sequenzbasierte (→insbesondere Domänen) Vorhersage alleine ausreicht, es sei denn die Literatur enthält bereits alle aus der Domänenstruktur ableitbaren Informationen.

Anmerkung zu Taxonomien Während der GO-Annotation wurde davon ausgegangen, dass das Mausmodell der untersuchten Proteine, das in den Publikationen dargestellt wird, äquivalent zum zu untersuchenden Humanmodell ist. Streng genommen kann davon nicht ausgegangen werden, allerdings stehen häufig für die Humanmodelle nur sehr wenige Daten zur Verfügung, da lediglich bei ohnehin erkrankten Individuen sequenzbasiert Schlüsse gezogen werden können (z.B. per cDNA-Bibliothek „Protein X wird bei Phänotyp Y mit reduzierter synaptischer Plastizität nicht oder nur reduziert exprimiert“→X *könnte* dem GO-Term *regulation of synaptic plasticity* zugeordnet werden). Bei Mausmodellen können, wie auch bei anderen Tiermodellen (z.B. *D. Rerio*), dagegen bestimmte Gene gezielt ausgeschaltet werden, um den Effekt zu beobachten. Dies ist beim Menschen nicht nur aus ethischer Sicht, sondern u.A. auch wegen des langen Generationszyklus schwierig bis unmöglich. Allerdings ist in den meisten Fällen ein mausbezogenes Modell genetisch ausreichend nahe an einem Humanmodell, sodass es sinnvoll für Humangenom-bezogene Analysen eingesetzt werden kann.

Aufgabe 7

Wir haben den Vergleich pro Gruppe nach Protein und nach allgemeinen Unterschieden sortiert. Aufgrund des begrenzten Platzes konnten wir nicht auf jeden einzelnen Unterschied eingehen, sondern haben exemplarisch für spezifische Ursachen repräsentative Unterschiede analysiert.

Gruppe 1, Allgemein Gruppe 1 hat beispielsweise *protein binding* annotiert. Dies betrachten wir als redundant, da es z.B. bereits in *actin binding* enthalten ist (per GO-Relation ist *actin binding* ein spezifischerer Unterbegriff (child) von *protein binding*). Desweiteren finden wir in PMID 15541315 keine Evidenz, nach der CPG2 (im Gegensatz zu SYNE1) ein Aktin-bindendes Protein ist. Leider hat die Gruppe 1 bei der uns vorliegenden Annotation weder GO-IDs noch Quellen annotiert, was die Zurückführung auf Ursachen teilweise schwierig macht.

Gruppe 1, SYNE1 Für *lamin-binding* bzgl. SYNE1 finden wir keine Evidenzen in PMID 19874786, da sich der Satz „The N-termini of SUN proteins have been shown to be in the nucleoplasm and to interact with nuclear lamins“unserer Auffassung nach lediglich auf SUN1 und SUN2 bezieht. SYNE1 ist allerdings ein KASH-Protein. Dasselbe gilt für Termini wie *nuclear migration*. Wir vermuten, dass Gruppe 1 nicht streng nach SUN1/SUN2/SYNE2-bezogenen Evidenzen und SYNE1-bezogenen Evidenzen differenziert hat. Zudem fällt auf, dass Gruppe 1 *dynein complex binding* annotiert hat. Wir fanden allerdings nur Evidenzen für *dynein binding*.

Gruppe 1, CPG2 Bei CPG2 fällt hier insbesondere auf, dass *SUN-KASH Complex* annotiert wurde. In PMID 15541315 finden wir hierfür keine Evidenzen, allerdings zeigt SwissProt eine automatisierte Annotation mit Ensembl als Evidenz an. Derartige automatische Annotationen, die sich auf Sequenzbasierte Ähnlichkeiten zurückführen lassen, haben wir nicht unsere Annotation aufgenommen.

Gruppe 2, SYNE1 *Interkinetic nuclear migration* wurde bei uns nicht annotiert, da wir in PMID 19874786 keine eindeutige Zuordnung zu den 4 Proteinen SUN1/SUN2/SYNE1/SYNE2 gefunden haben. INM ist zwar ein wichtiger Prozess für brain development und hängt mit SUN-KASH-Komplexen zusammen, muss aber deshalb noch nicht relevant für SYNE1 sein. Für *sarcomere* fanden wir zwar in Datenbanken wie UniProt Evidenzen, nicht aber in der Literatur.

Gruppe 2, CPG2 *regulation of endocytosis* ist unserer Auffassung nach ein zu allgemeiner Term, da, zumindest aufgrund der von uns untersuchten Literatur-Evidenzen *clathrin-mediated endocytosis* reguliert wird. Für eine positive Regulierung finden wir ebenso keine Evidenzen, unserer Auffassung nach kann aufgrund der vorliegenden Literatur nicht eindeutig entschieden werden, ob die Regulation positiv oder negativ ist.

Gruppe 3, SYNE1 Ähnlich zu Gruppe 2. Für *cerebral cortex cell migration* finden wir keine Evidenzen – dies könnte an einer Verwechslung mit dem *cerebellum* liegen, das orthographisch ähnlich ist, sich aber auf das Kleinhirn und nicht auf den Cerebralcortex bezieht. Zudem finden wir keine eindeutige Zuordnung zu SYNE1.

Gruppe 3, CPG2 Hier fällt insbesondere auf, dass im Gegensatz zu unserer *regulation of clathrin-mediated endocytosis* als Lokation *clathrin-coated vesicle* angenommen wurde. Dafür finden wir keine Evidenzen.

Gruppe 4, SYNE1 Ähnlich zu Gruppe 2, insbesondere finden wir keine Eindeutige Evidenz für *CH-Domain binding*. SYNE2 enthält eine CH-Domäne, aber deswegen bindet SYNE1 noch nicht an CH-Domänen.

Gruppe 4, CPG2 Ähnlich zu Gruppe 2, für regulation of transmembrane transport *regulation of transmembrane transport* finden wir keinerlei Evidenzen, in der Literatur, allerdings führt der *Gene expression atlas* diesen GO-Term auf.

Gruppe 5, Allgemein Im Gegensatz zu den meisten anderen Gruppen klassifiziert Gruppe 5 ihre Hits nach Quelle der Evidenz und (qualitativ) nach Stärke der Evidenz. Dies kann bei der Verifikation sehr hilfreich sein und wurde auch von uns nur eingeschränkt umgesetzt (wir haben lediglich die Abschnitte und PMIDs der Quellen angegeben).

Gruppe 5, SYNE1 Ähnlich zu Gruppe 2, allerdings fehlt *brain development*, das sich unserer Auffassung nach eher als das spezifischere *Cerebral cortex development* (siehe auch Gruppe 3) anwenden ließe.

Gruppe 5, CPG2 Ähnlich zu Gruppe 2, insbesondere fällt auf dass Gruppe 5 im Gegensatz zu den meisten anderen Gruppen eine Relation zu AMPA erkannt hat (der GO-Name des Terms ist allerdings nur abgekürzt angegeben, sowie ohne GO-ID. Bei Komplexeren Aufgabenstellungen könnte dies zu Verwechslungen führen). Wir finden allerdings keine Evidenz für Clustering von AMPA, sondern lediglich für die Regulierung.

Gruppe 6, SYNE1 Ähnlich zu Gruppe 2.

Gruppe 6, CPG2 Ähnlich zu Gruppe 2, wobei im Gegensatz zu den meisten anderen Gruppen auch die Lokalisation im *dendritic shaft* erkannt wurde, für die sich in der gegebenen Literatur mehrere Evidenzen finden.

Zusammenfassung der Unterschiede und Gemeinsamkeiten

Beim Vergleich aller Gruppen zeigt sich, dass es eine Reihe von Termini gibt, die bei fast allen Gruppen gemeinsam auftreten, wie beispielsweise *SUN-KASH complex* oder *brain development*. Diese Termini wurden in der untersuchten Literatur meist wiederholt und eindeutig in Zusammenhang mit den untersuchten Proteinen aufgeführt.

Viele Termini treten dagegen nur bei einer geringen Gruppenanzahl auf – nimmt man an, dass unsere Annotation korrekt ist, lassen sich viele davon darauf zurückführen, dass falsche Relationen zwischen den Proteinen und in den Publikationen anderweitig vorkommenden Begriffen gezogen wurden, oder ein Mechanismus falsch oder nicht tiefgreifend genug verstanden

wurde. Genauso ließe sich allerdings annehmen, dass selbiges auf unsere Annotation zutrifft – beispielsweise, weil wir einfach Evidenzen übersehen haben. Eine eindeutige Annotation können wir aufgrund mangelnder Erfahrung nicht liefern – aus den Resultaten der Gruppen ließe sich zwar ein Kandidat ableiten, allerdings müsste dafür ein Kompromiss zwischen Spezifität und Sensitivität gefunden werden. Dies ist allerdings nicht eindeutig möglich.

Zusatzaufgabe 8

Erratum Während der Bearbeitung dieser Aufgabe fiel uns auf, dass der Link zum Bulk-Download von FunCat (links unten auf `http://mips.helmholtz-muenchen.de/proj/funcatDB/`) nicht mehr funktioniert, da das Verzeichnis auf dem FTP-Server nicht mehr existiert. Wir bitten darum, diese Information an die dafür zuständigen Personen weiterzuleiten[2].

Vorgehensweise Aus den Vorträgen und aus eigenen Erfahrungen ist bekannt, dass FunCat älter und weniger umfangreich als GO ist – vermutlich auch, da dem GO-Konsortium deutlich mehr Personal zur Verfügung steht. Daher wurden die GO-Termini von SYNE1 sowie weitere Termini aus den Publikationen, die in GO nicht gefunden wurden, in FunCat gesucht und zudem noch der (im Vergleich zu GO kleine) Term-Tree nach passenden Termini durchsucht wurde.

FunCat-Annotation Termini:

- 73.03.13 neuron
- 77.03.01.01.01 brain
- 20.09.18.09.01.05 synaptic vesicle endocytosis
- 16.01 protein binding
- 41.05.13 neurogenesis
- 70.03 cytoplasm

Diskussion der Ergebnisse Die FunCat-Annotation hat einen deutlich geringeren Umfang als die GO-Annotation. Dies liegt vor allem daran, dass viele Termini wie beispielsweise *dendritic spine* in FunCat nur sehr unspezifisch oder gar nicht vertreten sind (Beispielsweise als localization im *neuron*). Zudem viel auf, dass FunCat viele Begriffe wie z.B. *brain* mehrfach in verschiedenen Kontexten enthält. Daher müssen Termini und deren Zusammenhänge noch genauer als bei GO überprüft werden, da false positive hits häufig sind. Leider enthalten viele Termini bei FunCat (im Gegensatz zu GO) keine genaue Beschreibung, beispielsweise *41.05.13 neurogenesis*.

[2] Es ist zwar möglich, aus dem ECMAScript-Code des Treeviews die FunCat-Hierarchie abzuleiten, allerdings schien uns der Aufwand hierzu unverhältnismäßig hoch

Automatisierung des Workflows

Insbesondere während der Bearbeitung der Aufgabe 5 fiel uns auf, dass die manuelle Annotation, insbesondere mit GO-Termini einen signifikanten manuellen Arbeitsaufwand erfordert – zum Zeitpunkt der Bearbeitung war zudem die GO-Suche reproduzierbar langsam und dauerte für fast jeden gesuchten Begriff zwischen 10 und 120 Sekunden. Dies lies sich zu anderen Zeitpunkten zwar nur beschränkt reproduzieren, liegt aber vermutlich daran, dass AmiGO CGI-Webservices verwendet, die selbst bei leistungsfähigen Servern einen signifikanten Overhead erzeugen, der – insbesondere bei hoher Auslastung – zu hohen Response-Latenzen führt.

Da in GO nicht alle in den gegebenen Publikationen vorkommenden denkbaren Termini enthält, gab es zudem viele Suchen ohne passende Resultate. Der größte Zeitaufwand ergab sich somit nicht etwa aus dem Lesen der Publikationen, sondern dem Zuordnen von semantisch relevanten Begriffen.

Aufgrund vorhergehender Entwicklungen wie EXCERBT (`http://mips.helmholtz-muenchen.de/excerbt/`, bei dessen Entwicklung einer der Autoren dieses Berichtes beteiligt war) ist allerdings bekannt, dass sich aus linguistischer Sicht viele der Termini aus dem Fließtext der Publikationen automatisch verschiedenen Ontologien wie *GO* (wie im vorliegenden Fall) oder *SNOMED* (insbesondere bei humanmedizinischen Textkorpora) zuordnen lassen.

Obwohl von der automatischen Analyse viele Termini nicht bzw. inkorrekt erkannt werden, wäre eine Software, die nicht nur eine interaktivere Suche (bei der die Resultate noch während der Eingabe angezeigt werden), sondern insbesondere eine direkte Suche nach Termini aus dem Text heraus und eine Markierung der entsprechenden Textstellen bietet, eine Hilfe bei dearartigen Aufaben.

Die Liste der bestehenden Tools für GO (`http://www.geneontology.org/GO.tools.annotation.shtml`) scheint eine derartige Software nicht zu enthalten – die Webseite der Software *PubSearch* war zum Zeitpunkt der Erstellung dieses Berichtes offline (wie auch der Quellcode-Download).

GOCat (`http://eagl.unige.ch/GOCat/`) dagegen bietet keine interaktive Annotation und eine vergleichsweise langsame Suche. Testweise haben wir den Abstract der Publikation zu Negatome 2.0 (`http://www.ncbi.nlm.nih.gov/pubmed/24214996`) in *GOCat* (im Machine-Learning-Modus) eingefügt – in den 50 Ergebnissen finden sich viele *false positives* wie *GO:0044304: main axon*, zu denen es keine Entsprechung im Text zu geben scheint.

Um unser Konzept einer solchen Software verdeutlichen zu können, haben wir im Rahmen dieses Berichtes das Tool *GOText* (`http://gotext.localgrid.de`) entwickelt, das vor allem eine interaktive Suche nach Termen und Termpräfixen in den gesamten GO-Termini erlaubt (Gesucht wird aktuell mit dem Modus *case-insensitive prefix* in den GO-Feldern *name* und *def* getrennt, ohne Beschränkung auf *exhaustive matches*).

Hätte dieses Tools bereits vor der Bearbeitung der Aufgaben zur Verfügung gestanden, so hätten wir die GO-Annotation deutlich effizienter durchführen können, da bereits während der Eingabe des potentiellen Terms meist sichtbar gewesen wäre, dass in GO ein solcher Term nicht existiert. Wie im Screenshot 1 sichtbar, werden die Ergebnisse einer Suche in Matches auf den GO-Name und die GO-Termdefinition unterteilt. Bereits während der Eingabe werden in Echtzeit Ergebnisse aus einer speziell für diesen Zweck angepassten parallelisierten Levelled-Inverted-Index-Implementierung auf YakDB-Basis (YakDB wurde von den Autoren dieses Berichtes für die Verarbeitung von großen Bio- und Geoinformatischen Datensätzen konzipiert

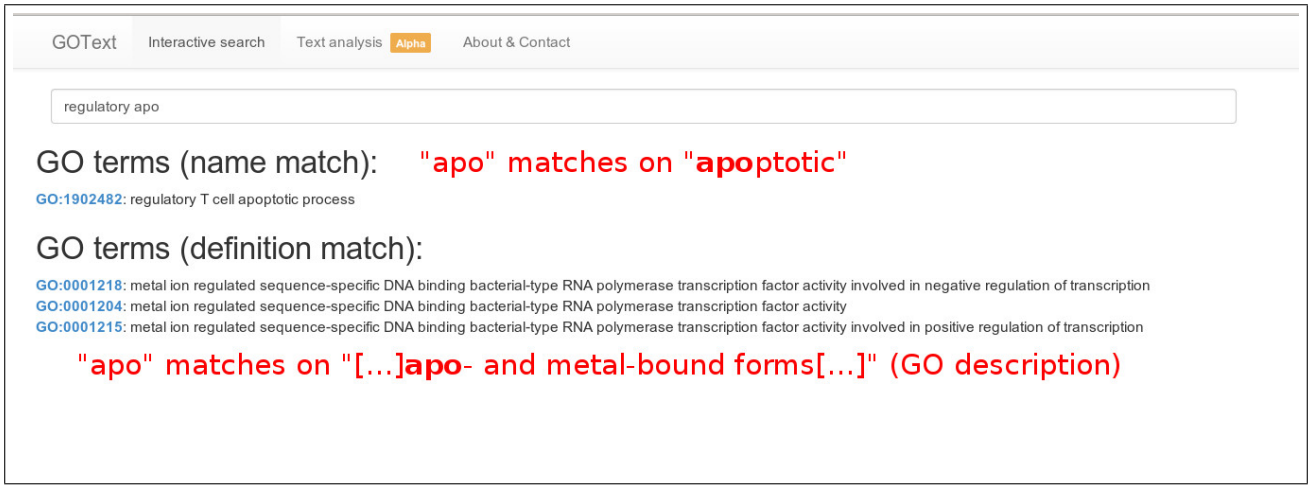

Abbildung 7: Screenshot: Die Termsuche in GOText

und entwickelt, ist zu diesem Zeitpunkt allerdings noch nicht öffentlich verfügbar). Hierdurch wurde eine gemittelte Antwortzeit für die interaktive Suche von unter 10 Millisekunden erreicht (nach Cache-Warmup).

Zudem bietet GOText einen Prototyp eines interaktiven Textannotators, der eine Serverseitige *Named Entity Recognition* mit GO-Terms erlaubt. Dessen Konzept basiert auf der Beobachtung, dass bestehende Tools wie GOCat zwar eine Liste an gefundenen Termini anzeigen, allerdings keine Evidenzen bieten, die anzeigen, wo im eingegebenen Text bestimmte Termini gefunden wurden. GOText zeigt daher nicht nur die gefundenen Termini, sondern auch Textstellen an, an denen diese gefunden wurden.

Methoden (Überblick)

Da für die Entwicklung für GOText nur etwa eine halber Arbeitstag zur Verfügung stand, ist GOText bei weitem kein fertiges Produkt, sondern lediglich ein Proof-of-Concept-Prototyp. Gezeigt wurde, dass eine schnelle interaktive Live-Suche in den GO-Termini möglich ist, und wie eine visuelle Annotationsunterstützung aussehen könnte.

In der gezeigten Implementierung wird nur ein kleiner Teil der im Text enthaltenden Terme gefunden – dies liegt vor allem daran, dass per *Natural Language Processing* lediglich *Noun phrases* aus dem Text extrahiert werden und diese ähnlich wie bei der interaktiven Suche in der Datenbank gesucht werden.

Allerdings erzeugt die *Noun phrase*-Erkennung für diesen Zweck sehr viele *false negatives*, unter anderem da sie im Falle von GOText teilweise mit Korpusbasierten Datenbanken durchgeführt wird, die nicht auf biomedizinische Texte trainiert sind. Eine Alternative wäre die Suche nach n-Grammen (die auch in EXCERBT implementiert ist), die zu einer deutlich größeren Anzahl an Hits führen würde – und somit zu einer höheren Sensitivität auf Kosten der Spezifität.

Im vorliegenden Fall war vor allem die geringe privat zur Verfügung stehende Serverkapazität für die Wahl der NER nur auf erkannte *noun phrases* ausschlaggebend. Eine Suche nach n-Grammen ist selbst bei EXCERBT, das über alle zu erkennenden Entitäten einen Inverted Index

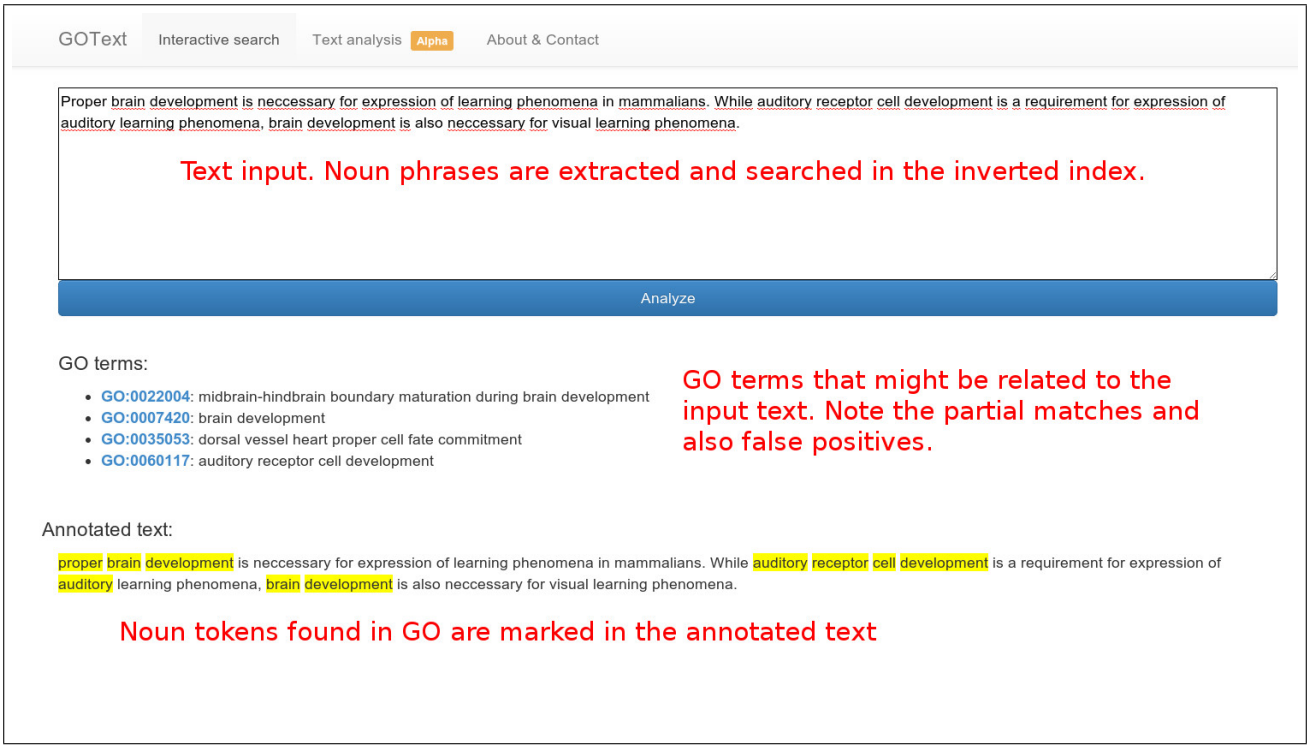

Abbildung 8: Screenshot: Die Textanalyse in GOText

auch über *multi-token-entities* erstellt[3], vergleichsweise Zeitaufwändig. Für GOText haben wir uns dagegen für einen Inverted Index, der ausschließlich aus Einzeltokens besteht, entschieden. Bei einer Suche nach Phrasen mit mehreren Tokens wird die Schnittmenge der Resultate der einzelnen Tokens berechnet. Diese Vorgehensweise benötigt deutlich weniger Speicherplatz und tendiert zur Verschiebung der Resultate in Richtung *false positives*, allerdings erzeugt sie insbesondere bei Entitäten mit einer großen Tokenanzahl (wie z.B. IUPAC-Identifiern) eine sehr hohe CPU-Last und müsste daher für eine produktive Verwendung deutlich besser untersucht werden.

Ein Vorteil des single-Token-Indizes ist, dass durch den geringeren Speicherplatzverbrauch deutlich weniger Disk-IO auftritt, wodurch die Gesamtperformance trotz der höheren CPU-Last steigen könnte. Eine quantitative Untersuchung der Abhängigkeit dieser Beobachtungen von verschiedenen Faktoren wie z.B. der Größe der Texte oder der Anzahl der bekannten Entitäten fand bisher nicht statt, da die für diesen Bericht zur Verfügung stehende Zeit für derartige Analysen nicht ausreicht.

Ausblick

Für Texte in der Größenordnung von durchschnittlichen PubMed-Abstracts funktioniert die Textanalyse von GOText sehr schnell und kann Resultate in nur wenigen Sekunden anzeigen. Dem gegenüber steht die geringe Erkennungsrate, die in der Regel nur offensichtliche Hits

[3]EXCERBT implementiert mehrere Algorithmen für die NER. Wir beziehen uns hier auf die NER mit präindizierten Entities, nicht auf diejenige mit präindizierten Dokumenten.

erkennt.

Eine mögliche produktiv einsetzbare Variante von GOText könnte zudem eine Benutzeroberfläche implementieren, die es dem Nutzer erlaubt, ein oder mehrere Tokens zu markieren und direkt in verschiedenen Ontologien bzw. Datenbanken zu suchen. Zudem bestünde die Möglichkeit, aufgrund der manuell annotierten Ontologie-Termini für ein Dokument[4] ähnliche Dokumente in einer Datenbank zu identifizieren. Dies wäre nicht nur für den Nutzer vorteilhaft, sondern könnte auch dazu genutzt werden, Systeme zur Relationserkennung (wie EXCERBT) besser zu trainieren.

Der Zweck der nutzerfreundlichen Implementierung eines solchen Systems wäre somit nicht nur primär die Verbesserung der Effizienz bei der Literaturannotation, sondern auch sekundär die Datenakquise zur Verbesserung der Genauigkeit von Tools zur automatisierten Relationsextraktion – langfristig könnten die Tools zur SLR und NER so auch speziell auf biomedizinische Texte trainiert werden, ohne dafür große Textkorpora erstellen zu müssen[5].

Daher sind wir der Auffassung, dass das Grundkonzept von GOText, eine interaktive Lösung zur Kombination von manueller und automatischer Annotation zu schaffen, weiter verfolgt werden sollte.

Konklusion

Wie auch bei der Genvorhersage zeigte sich bei der funktionellen Klassifikation eindeutig, dass nur eine Kombination aus manueller Annotation und automatischer Klassifikation zu verwertbaren, normalisierten Ergebnissen führen kann:

Bei rein automatischer Annotation und Klassifikation sind entweder Sensitivität oder Spezifität hoch genug, um die Resultate weiter verwerten zu können, allerdings ist eine rein manuelle Annotation bei der großen Zahl der zu analysierenden Proteine nicht realistisch durchführbar. Daher werden von Datenbanken wie UniProt alle Proteine soweit möglich automatisch annotiert und ausgewählte Proteine manuell reviewed (SwissProt).

Wie auf `http://www.ebi.ac.uk/uniprot/TrEMBLstats` sichtbar, sind fast $\frac{3}{4}$ der Proteine lediglich vorhergesagt und haben keine experimentellen Evidenzen – bei vielen dieser Proteine existiert vermutlich keine verwertbare Literatur, eine manuelle Annotation könnte daher nur die Resultate der automatischen Annotation überprüfen, würde aber in der Regel kaum neue Information hinzufügen.

Selbst bei Einsatz enormer personeller Ressourcen scheint es auch in Zukunft unmöglich, alle Einträge der Proteindatenbanken zu reviewen – UniProt hat alleine im Oktober 2013 über 3.5 Millionen neue Sequenzen aufgenommen, es ist sogar zu erwarten dass diese Zahl in Zukunft noch weiter steigt. Zudem müssen Annotationen bei der Veröffentlichung neuer Literatur gewartet werden, um neue Erkenntnisse mit in die Datenbanken aufzunehmen.

Selbst bei den manueller Annotation sind die Resultate, wie Aufgabe 7 gezeigt hat, keineswegs eindeutig. Dies liegt teilweise an der mangelnden Annotationserfahrung der Annotatoren und unterschiedlichen Prämissen bezüglich Sensitivität vs. Spezifität – es kann aber auch Ursachen

[4] Im Sinne der EXCERBT-Definition von *document*, insbesondere PubMed-Abstracts bzw. PMC-Fulltext-Papers

[5] Die meisten frei verfügbaren biomedizinischen Kopora, siehe z.B. `http://www.nactem.ac.uk/resources.php` haben einen geringen Umfang in der Größenordnung von 1000 Abstracts

wie nicht eindeutige sprachliche Konstrukten in der Literaturquellen oder nicht eindeutige Versuche haben. Ließe man also nur vollkommen eindeutige Resultate zurück (Spezifität auf Kosten der Sensitivität), würden nur sehr wenige oder gar keine Klassifikationen auf Proteine zutreffen.

Ebendiese Klassifikationen sind aber zur Vorhersage der Funktion unerlässlich. Da Versuche, die die Funktion eines Proteins eindeutig experimentell bestimmen, teuer und nicht ohne weiteres möglich sind, muss für den Großteil der bekannten Proteine auf die Analyse von homologen Zusammenhängen mit anderen Proteinen zurückgegriffen werden.

Wie auch bei der automatischen Annotation ist die manuelle Annotation zudem von Fehlern behaftet – bei Aufgabe 7 zeigte sich, dass wir für viele Annotationen der anderen Gruppen keine Evidenzen finden konnten. Dies könnte aber auch bedeuten, dass wir ebendiese schlicht übersehen oder falsch verstanden haben. Es wäre zwar möglich gewesen, weitere Literatur zu den beiden Proteinen zu recherchieren, die weitere Evidenzen liefern – allerdings war die für diese Aufgabe zur Verfügung stehende Zeit begrenzt und die weitere Recherche hätte vermutlich nur zu wenigen zusätzlichen Annotationen geführt.

Die sequenzbasierte Analyse ist eine weitere Komponente für die funktionelle Klassifikation, die Evidenzen liefert, die in der Literatur nur selten zu finden sind – die automatische Annotation funktioniert hier in der Regel besser als bei der Literatursuche. Allerdings hängt die Qualität der Resultate stark von der Qualität der Datenbank ab. Beim UCSC Genome Browser trat das Problem des Aliasing (in unserem Fall *Nesprin-1* vs. *SYNE1*). Dieses Problem ist aufgrund des Bekanntheitsgrades von UniProt und dem dort eingetragenen Alias unverständlich und lässt auf grundsätzliche Probleme mit dem Alias mapping des UCSC Genome Browser schließen.

Mit GOText wurde ein Prototyp eines Tools vorgestellt, dessen Konzept eine effizientere, teilweise automatisierte Annotation von Literatur ermöglicht und dazu neue Konzepte des Echtzeit-Texminings und der Suchmaschinentechnik umsetzt. Leider konnte kein fertiges Produkt präsentiert werden, sondern nur ein eingeschränktes Proof-of-Concept, vor allem zur Demonstration der Suchmaschinentechnik. Eine vollständige Implementierung könnte den Annotatoren unter Umständen einen höheren Durchsatz ermöglichen und zudem Textminingsysteme langfristig stark verbessern.

Wir hoffen, dass in Zukunft Konzepte wie EXCERBT und GOText nicht nur den Annotationsprozess vereinfachen, sondern möglicherweise auch eine standardisierte Grammatik zur Expression von Relationen entwickelt wird, die von Wissenschaftlern genutzt werden kann. Damit könnten die Probleme, die bei Textmining-Systemen beim *natural language processing* auftreten, umgangen werden – die Qualität von *in-silico*-Analysen würde damit signifikant steigen.