Auffinden von Proteinkomplexen in PPI-Datenbanken durch Cliquensuche

Exposé zur Bachelorarbeit

Sebastian Günther

21. Juli 2015

1 Hintergrund

Proteine spielen in allen zellulären Prozesse eine entscheidende Rolle. Sie sind die sowohl strukturell wie auch funktional komplexesten der vier bekannten Makromolekülarten (Kohlenhydrate, Lipide, Nukleinsäuren, Proteine). Ihre funktionale Diversität folgt direkt aus den chemischen und physikalischen, durch den strukturellen Aufbau bedingten, Eigenschaften. Proteine bestehen aus kovalent gebundenen, geordneten Aminosäureketten (Primärstruktur), die dazu tendieren sich lokal in immer wiederkehrende Formen (Sekundärstruktur: Helices, Sheets, Loops) zu legen. Durch Anziehungs- und Abstoßungskräfte zwischen verschieden geladenen Seitenketten der Molekülkette wird sie in ihre spezifische, aktive Form gefaltet (Tertiärstruktur) [KB10, S. 67f].
Durch unterschiedliche Anordnung der Aminosäuren ergeben sich eine Vielzahl an räumlichen Strukturen, so dass durch die räumliche Lage der funktionalen Gruppen nur sehr spezifische Moleküle gebunden werden können. Diese Spezifität brachte evolutionäre Vorteile, so ist es beispielsweise möglich mehrere chemische Prozesse mit Beteiligung von Proteinen gleichzeitig ablaufen zu lassen, da Proteine substrat- und reaktionsspezifisch sind und nicht an den „falschen“ Partner binden [KB10, S. 27]. Auch konnten erst mit Hilfe von Proteinen chemische (und somit auch biologische) Prozesse genau gesteuert werden, indem diese als Regulatoren bestimmte Reaktionen ermöglichen (Aktivatoren) oder bremsen (Inhibitoren). Weitere Vorteile der Nutzung von Proteinen liegen in der Koppelung von energieproduzierenden Prozessen mit energieverbrauchenden Prozessen und der räumlichen Eingrenzung chemischer Reaktionen in der Zelle [KB10, S. 20f]. Weiterhin besetzen Proteine Schlüsselfunktionen bei Transport und temporärer Speicherung von Energie innerhalb der Zelle indem sie unter anderem als Protonenpumpen und Ionenkanäle den Transport von Protonen (H^+) durch die lipide Trennwand zwischen mitochondrialer Matrix und mitochondrialem intermembranalem Raum ermöglichen [KB10, S. 21ff]. Auch die interzelluläre Kommunikation (endokrines System) läuft großteils mit Hilfe von Proteinen, den Proteohormonen und Peptidhormonen, ab. Außerdem sind Proteine maßgeblich beteiligt an der Zellabwehr und dem Aufbau intra- und extrazellulärer Strukturen mit speziellen Eigenschaft, wie zum Beispiel der Muskelkontraktion (durch Myosin und Aktin) [KB10, S. 28ff].
Die Untersuchung von Proteinen, insbesondere beim Menschen, wie sie unter anderem durch das „Human Proteome Project“ [Org15] betrieben wird, ist wichtig, da Proteine wichtige funktionale Einheiten der Zelle sind und daher nur mit Verständnis der Proteinfunktionen und Kenntnis ihres Vorkommen und der Interaktion mit anderen Proteinen die grundlegenden Lebensfunktionen verstanden werden können. Dieses Wissen kann dann bei der Entwicklung von Therapien zur Krankheitsbekämpfung eingesetzt werden, um zum Beispiel gezielt Proteine anzugreifen, die nur in kranken Zellen auftreten, nicht

aber in gesunden.

1.1 Proteinkomplexe

Proteine agieren in einer Vielzahl der Fälle nicht alleine, sondern bilden mit anderen Proteinen Komplexe (Quartärstruktur), sogenannte Oligomere und Multimere, mit oftmals wenigen manchmal aber auch einer großen Anzahl an „Subunits“, also einzelnen Proteinmolekülen. So besitzt beispielsweise die HIV-1 Aspartyl Protease nur 2 Subunits, die F_0-Komponente der ATP-Synthase 13 Subunits und in „Human large 60S ribosomal subunit“ wurden 57 Subunits gefunden [KB10, S. 165f].
Proteine können nicht nur mit anderen Proteinen Komplexe bilden, sondern auch mit RNA (Ribosome), DNA (Nukleosome) und Lipiden (Kernpore). Dies ist für die Arbeit aber nicht von Relevanz.

1.2 Methoden zur Bestimmung von Proteinkomplexen

Es gibt direkte und indirekte Methoden zur Bestimmung von Proteinkomplexen. Für die direkte Bestimmung wird die Struktur des Komplexes mittels experimenteller Methoden aufgeklärt, wie Massenspektrometrie [u.a. Ho+02], Kristallstrukturanalyse [siehe bspw. Dei+84] und Kernspinresonanzspektrospkopie [siehe bspw. Sre+09]. Zur indirekten Bestimmung werden zunächst Interaktionen zwischen jeweils zwei Proteinen untersucht. Lassen sich so Proteine finden, die alle (oder ein großer Teil) paarweise interagieren, kann daraus auf einen möglichen Proteinkomplex geschlossen werden. Daher werden im Folgenden einige experimentelle Methoden zur Bestimmung von Protein-Protein-Interaktionen betrachtet, die sich hinsichtlich ihrer Sensitivität und Spezifität und somit ihrer Zuverlässigkeit unterscheiden.

- *Affinitätschromatographie.* Hierfür wird ein Ligand an eine stationäre Phase gekoppelt. Mit diesem wird dann das bindende Protein aus einem Gemisch gefangen. Danach wird das Ausfallprodukt getrennt und bspw. mittels SDS-PAGE oder Massenspektrometrie bestimmt. [siehe Ben74]

- *Tandem Affinity Purification.* Proteinreinigung bei der zwei Affinitätschromatographien hintereinander ausgeführt werden. [siehe Pui+01]

- *Yeast2Hybrid.* Ein Transkriptionsfaktor (Protein) wird in sein aktives und bindendes Zentrum geteilt. Diese beiden Teile werden jeweils an eines der auf PPI zu testenden Proteine gebunden. Wenn die beiden Proteine nun interagieren, ist

der TF aktiviert und kann das „Reporter-Gen“ exprimieren (bspw. fluoreszierende Zellen). [siehe You98]

- *Co-Immunpräzipitation.* Für ein bekanntes Protein in einem vermuteten Proteinkomplex wird ein spezifischer Antikörper entweder auf einer stationären Phase aufgebracht (indirekt) oder zu dem Proteingemisch (direkt) gegeben. Bei der direkten Methode wird die stationäre Phase in das Gemisch gegeben und die spezifisch bindenden Proteine werden gefangen. Bei der indirekten Methode wird nach einer Weile ein Träger mit A/G-Protein umhüllt in das Gemisch gegeben, dieses bindet das an den Antikörper und der Proteinkomplex ist gefangen. [siehe PF95]

2 Zielsetzung

Da es bereits eine große Zahl von PPI-Datenbanken und eine kaum überschaubare Anzahl an Veröffentlichungen experimentell bestimmter Proteinkomplexe gibt, aber keine umfassende Datenbank humaner Proteinkomplexe existiert, ist das Ziel dieser Arbeit die Bereitstellung einer umfassenden Proteinkomplexdatenbank. Diese Datenbank dient dem Vergleich experimentell ermittelter Proteinkomplexe mit Komplexdaten, die entweder aus bereits bestehenden Proteinkomplexdatenbanken stammen oder rechnerisch mittels (Quasi-)Cliquensuche in dem durch die PPI induzierten Graphen ermittelt wurden. Dies geschieht unter den Annahme, dass (fast) vollständig verbundene Teilgraphen des aus Proteinen als Knoten und Interaktionen als Kanten gebildeten Interaktionsgraphen Proteinkomplexe anzeigen, da in diesen (fast) alle Mitgliederproteine miteinander interagieren. Außerdem soll die Datenbank einen Qualitätsindex bieten, anhand dessen sich die Güte und somit die Zuverlässigkeit der gefundenen Proteinkomplexe bestimmen lässt.

3 Vorgehensweise

Zunächst wird untersucht inwiefern verschiedene PPI-Datenbanken integriert werden können. Danach wird aus den Interaktionsdaten ein Graph mit den Proteinen als Knoten und Interaktionen als Kanten erstellt und die Kanten nach der Zuverlässigkeit der experimentellen Bestimmungsmethode gewichtet. Da durch die experimentellen Methoden die Interaktionsdaten verrauscht sind und es daher zu fälschlicherweise fehlenden oder bestehenden Kanten kommen kann, ist es angebracht nach Quasi-Cliquen, also „fast vollständig verbundenen“ Teilgraphen zu suchen.
Diese stellen die gesuchten Komplexe dar, die in einem abschließenden Schritt noch mit

einem, aus den Kantengewichten der durch die Quasiclique induzierten Teilgraphen berechneten, Qualitätsscore versehen werden.
Zum Schluss werden die rechnerisch ermittelten Komplexe mit den bestehenden Daten aus den Komplexdatenbanken Corum [Rue+09], PDB [Ber+00] und PCDq [Kik+12] integriert, um somit eine umfassende Proteinkomplexdatenbank zu erhalten.

3.1 Zuverlässigkeit von experimentell bestimmten PPI-Interaktion

Da im Verlauf der Arbeit jedem Proteinkomplex ein Qualitätsscore zugeordnet werden soll, der eine qualitative Aussage darüber trifft, wie hoch die Wahrscheinlichkeit des Auftretens des ermittelten Proteinkomplexes in einem lebenden Organismus ist, muss zunächst eine Bewertung der angewandten PPI-Bestimmungsmethoden vorgenommen werden, so dass jeder Interaktionen ein Score zugeordnet werden kann.
Auf Grund intrinsischer Fehler der experimentellen Methoden ist es allerdings schwierig Aussagen über die Zuverlässigkeit der gefundenen Interaktionen zu treffen. So liegt die geschätzte Fehlerrate bei Y2H in der Größenordnung von 50% [SSM03]. Hinzu kommt, dass nur ein geringer Anteil an Interaktionen von mehr als einer Methode bestimmt werden konnten (ca. 2400 von 80000 bei Hefe).
In [SSM03] sieht man, dass die True-Positive-Rate, also der Anteil wirklich stattfindender Interaktionen, bei allen biochemischen und immunologischen Bestimmungsmethoden bei über 80% liegt und somit weit zuverlässiger ist als Y2H (60-70% in „small-scale-Versuchen“ und ca 50% bei „high-throughput-Experimenten“).
Konkrete Scores für eine Vielzahl experimenteller Methoden werden in [Sch+12] vergeben und dienen als Grundlage der Scores in der Arbeit.

3.2 Integration von PPI-Datenbanken

Es gibt bereits eine Vielzahl an Proteininteraktions- wie auch einige Proteinkomplexdatenbanken, siehe Tabelle 1 . Als Datengrundlage werden händisch kuratierte Datenbanken verwendet. Also solche in denen ausschließlich PPI oder Komplexe enthalten sind, die durch experimentelle Methoden entdeckt wurden oder deren computergestütztes Auffinden durch Literaturvergleich verifiziert wurde.

Es muss untersucht werden ob es überhaupt sinnvoll ist alle oder eine Auswahl der oben genannten PPI-Datenbanken zu integrieren oder ob es nicht ausreichend ist eine aktuelle gut kuratierte Datenbank als Grundlage für die nächsten Schritte zu verwenden. Dazu muss zunächst bestimmt werden zu welchem Grad sich die Datenbestände überdecken. Bei einer hohen Überdeckung kann von einer Integration der Daten abgesehen werden.

Datenbank	URL	Anzahl an Interaktionen/Komplexen
PDB	http://www.rcsb.org/pdb/home/home.do	4989 Komplexe
Corum	http://mips.helmholtz-muenchen.de/genre/proj/corum	2084 Komplexe
PCDq	http://h-invitational.jp/hinv/pcdq/	1266 Komplexe
DIP	http://dip.doe-mbi.ucla.edu/	74962
HPRD	http://www.hprd.org/	39240
HIPPIE	http://cbdm.mdc-berlin.de/tools/hippie/	193486
STRING	http://string-db.org	
BioGRID	http://thebiogrid.org/	359110
IntAct	http://www.ebi.ac.uk/intact/	349690
CPDB	http://consensuspathdb.org/	238766
APID	http://bioinfow.dep.usal.es/apid/index.htm	322579

Tabelle 1: Aktuelle PPI- und Proteinkomplexdatenbanken

Einen Überblick über aktuelle Protein-Protein-Interaktionsdatenbanken bietet [KP11]. Bei der Integration werden die PPI-Scores mit „noisy OR“ [Die93] aggregiert. Das „noisy OR“ stellt eine Verallgemeinerung des logischen „ODER“ dar und berechnet sich unter den Annahmen i) es sind alle Ursachen U_i für ein Ereignis X bekannt, ii) wenn eine Ursache nicht eintritt ($\neg U_i$), hat sie keine Auswirkungen auf X und iii) unabhängige Wahrscheinlichkeiten q_i für jedes U_i, zu

$$\mu\left(X|U_1 \dots U_j, \not{U}_{j+1} \dots \not{U}_k\right) = 1 - \prod_{i=1}^{j} q_i \tag{1}$$

3.3 Vorbereitung der Cliquensuche

Aus den integrierten Interaktionsdaten wird ein Graph erstellt, dessen Knoten die Proteine und dessen Kanten Interaktionen sind. Die Kanten werden gewichtet mit den PPI-Interaktionsscores. Der Graph wird durch eine Adjazenzmatrix repräsentiert, mit den Kantengewichten als Einträge a_{ij}.

3.4 (Quasi-)Cliquensuche

Eine *Clique* ist eine Untermenge U der Knoten V eines ungerichteten Graphen $G = (V, E)$, so dass ihr induzierter Teilgraph $H = (U, F)$ vollständig ist. Eine Clique ist *maximal* wenn man ihr keine weiteren Knoten $v \in V$ hinzugefügt werden können, so dass $U \cup v$ eine Clique ist. U wird größte Clique (engl. *maximum clique*) genannt, wenn es keine Clique gibt, die mehr Elemente als U hat.
Eine γ-Quasi-Clique ist ein Teilgraph $H = (U, F)$ des Graphen $G = (V, E)$ mit $0 \leq \gamma \leq 1$, so dass gilt $|F| \geq \gamma \cdot \binom{|U|}{2}$. Eine erweiterte Definition gibt auch noch eine untere Grenze für den Grad jedes einzelnen Knotens an mit $deg_U(u) \geq \lambda \cdot (|U| - 1), \forall u \in U$ [BHB08].
Cliquen fester Größe k lassen sich mit einem Brute-Force-Algorithmus einfach finden, indem jeder Teilgraph mit k Knoten untersucht wird ob er eine Clique bildet. Dieser Algorithmus läuft in $\mathcal{O}(n^k k^2)$, ist also polynomial wenn k konstant ist.
Die Suche maximaler Cliquen ist NP-hart. Wenn also $P \neq NP$ gilt, gibt es keinen in Polynomialzeit laufenden Algorithmus, der für beliebige Graphen alle maximalen Cliquen auflistet. Der *Bron-Kerbosch-Algorithmus* [BK73] (und Variationen) mit einer Worst-Case-Laufzeit von $\mathcal{O}(3^{n/3})$ ist der bekannteste und meist genutzte Algorithmus um in Real-World-Graphen (oftmals nicht-dicht [engl. *sparse*]) alle maximalen Cliquen zu finden [CK08].
Da für $\gamma = \lambda = 1$ Quasi-Cliquen zu echten Cliquen werden, ist die Suche nach Quasi-Cliquen also mindestens genau so hart wie die Suche nach Cliquen.
Um Quasi-Cliquen zu finden, werden Abwandlungen heuristischer Verfahren zur Suche von Cliquen angewandt, *Dynamic Local Search for Maximum Cliques* und *Reactive Local Search* [BHB08].
Die genannten Algorithmen werden in Java implementiert. Sollte die Laufzeit der Suche einen noch festzulegenden Rahmen überschreiten, ist zu überlegen ob die Suche begrenzt wird oder eine Parallelisierung des Algorithmus erreicht werden kann.

3.5 Erstellung eines Qualitätsindex

Da dem Nutzer der Datenbank eine Möglichkeit gegeben werden soll, beurteilen zu können wie zuverlässig die gefundenen Proteinkomplexe sind, ist es notwendig die Daten mit einem Qualitätsscore zu versehen.
Dieser Index sollte sowohl eine Aussage treffen können über die Zuverlässigkeit des verwendeten experimentellen Verfahrens mit dem die Protein-Protein-Interaktion gefunden wurde wie auch über die Qualität des Komplexes, insbesondere ob der Komplex über (Quasi-)Cliquensuche gefunden wurde oder aus einer bestehenden kuratierten Datenbank stammt.

Dafür könnte ein Index aus folgenden Teilen aufgebaut sein:

1. Arithmetische Mittel und Standardabweichung aller Kantengewichte der Clique.
2. Anzahl der Proteine eines Komplexes, sortiert in noch festzulegende Kategorien.

Ähnliches wurde in [Kik+12] gemacht.

3.6 Vergleichende Statistik

Abschließend werden die Spezifizität und Sensitivität für verschiedene Parameter γ und λ der Quasi-Cliquen-Suche untersucht.

4 Related Work

Literatur

[Ben74] Heinz Bende. „Affinitäts-Chromatographie“. In: *Chemie in unserer Zeit* 8.1 (1974), S. 17–25. DOI: 10.1002/ciuz.19740080104. URL: http://dx.doi.org/10.1002/ciuz.19740080104 (siehe S. 2).

[Ber+00] Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov und Philip E. Bourne. „The Protein Data Bank“. In: *Nucleic Acids Research* 28.1 (2000), S. 235–242. DOI: 10.1093/nar/28.1.235. eprint: http://nar.oxfordjournals.org/content/28/1/235.full.pdf+html. URL: http://nar.oxfordjournals.org/content/28/1/235.abstract (siehe S. 4).

[BHB08] Mauro Brunato, Holger H Hoos und Roberto Battiti. „On effectively finding maximal quasi-cliques in graphs“. In: *Learning and Intelligent Optimization*. Springer, 2008, S. 41–55 (siehe S. 6).

[BK73] Coen Bron und Joep Kerbosch. „Algorithm 457: finding all cliques of an undirected graph“. In: *Communications of the ACM* 16.9 (1973), S. 575–577 (siehe S. 6).

[CK08] F. Cazals und C. Karande. „A note on the problem of reporting maximal cliques“. In: *Theoretical Computer Science* 407.1-3 (2008), S. 564–568. DOI: http://dx.doi.org/10.1016/j.tcs.2008.05.010. URL: http://www.sciencedirect.com/science/article/pii/S0304397508003903 (siehe S. 6).

[Dei+84] J. Deisenhofer, O. Epp, K. Miki, R. Huber und H. Michel. „X-ray structure analysis of a membrane protein complex: Electron density map at 3 Å resolution and a model of the chromophores of the photosynthetic reaction center from Rhodopseudomonas viridis“. In: *Journal of Molecular Biology* 180.2 (1984), S. 385–398. DOI: http://dx.doi.org/10.1016/S0022-2836(84)80011-X. URL: http://www.sciencedirect.com/science/article/pii/S002228368480011X (siehe S. 2).

[Die93] Francisco Javier Diez. „Parameter adjustment in Bayes networks. The generalized noisy OR-gate“. In: *Proceedings of the Ninth international conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc. 1993, S. 99–105 (siehe S. 5).

[Ho+02] Yuen Ho, Albrecht Gruhler, Adrian Heilbut, Gary D Bader, Lynda Moore, Sally-Lin Adams, Anna Millar, Paul Taylor, Keiryn Bennett, Kelly Boutilier u. a. „Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry“. In: *Nature* 415.6868 (2002), S. 180–183 (siehe S. 2).

[KB10] Amit Kessel und Nir Ben-Tal. *Introduction to proteins: structure, function, and motion*. CRC Press, 2010 (siehe S. 1 f.).

[Kik+12] Shingo Kikugawa, Kensaku Nishikata, Katsuhiko Murakami, Yoshiharu Sato, Mami Suzuki, Md Altaf-Ul-Amin, Shigehiko Kanaya und Tadashi Imanishi. „PCDq: human protein complex database with quality index which summarizes different levels of evidences of protein complexes predicted from H-Invitational protein-protein interactions integrative dataset“. In: *BMC Systems Biology* 6.Suppl 2 (2012), S7. DOI: 10.1186/1752-0509-6-S2-S7. URL: http://www.biomedcentral.com/1752-0509/6/S2/S7 (siehe S. 4, 7).

[KP11] Tomas Klingström und Dariusz Plewczynski. „Protein–protein interaction and pathway databases, a graphical review“. In: *Briefings in bioinformatics* 12.6 (2011), S. 702–713 (siehe S. 5).

[Org15] Human Proteome Organization. *Human Proteome Project*. 2015. URL: http://www.thehpp.org/ (siehe S. 1).

[PF95] Eric M Phizicky und Stanley Fields. „Protein-protein interactions: methods for detection and analysis.“ In: *Microbiological reviews* 59.1 (1995), S. 94–123 (siehe S. 3).

[Pui+01] Oscar Puig, Friederike Caspary, Guillaume Rigaut, Berthold Rutz, Emmanuelle Bouveret, Elisabeth Bragado-Nilsson, Matthias Wilm und Bertrand Séraphin. „The tandem affinity purification (TAP) method: a general procedure of protein complex purification“. In: *Methods* 24.3 (2001), S. 218–229 (siehe S. 2).

[Rue+09] Andreas Ruepp, Brigitte Waegele, Martin Lechner, Barbara Brauner, Irmtraud Dunger-Kaltenbach, Gisela Fobo, Goar Frishman, Corinna Montrone und H-Werner Mewes. „CORUM: the comprehensive resource of mammalian protein complexes - 2009“. In: *Nucleic acids research* (2009), gkp914. URL: `http://mips.helmholtz-muenchen.de/genre/proj/corum` (siehe S. 4).

[Sch+12] Martin H. Schaefer, Jean-Fred Fontaine, Arunachalam Vinayagam, Pablo Porras, Erich E. Wanker und Miguel A. Andrade-Navarro. „HIPPIE: Integrating Protein Interaction Networks with Experiment Based Quality Scores“. In: *PLoS ONE* 7.2 (02/2012), e31826. DOI: `10.1371/journal.pone.0031826`. URL: `http://dx.doi.org/10.1371/journal.pone.0031826` (siehe S. 4).

[Sre+09] Sridhar Sreeramulu, Hendrik RA Jonker, Thomas Langer, Christian Richter, C Roy D Lancaster und Harald Schwalbe. „The human Cdc37.Hsp90 complex studied by heteronuclear NMR spectroscopy“. In: *Journal of Biological Chemistry* 284.6 (2009), S. 3885–3896 (siehe S. 2).

[SSM03] Einat Sprinzak, Shmuel Sattath und Hanah Margalit. „How reliable are experimental protein–protein interaction data?“ In: *Journal of molecular biology* 327.5 (2003), S. 919–923 (siehe S. 4).

[You98] KH Young. „Yeast two-hybrid: so many interactions,(in) so little time...“ In: *Biology of reproduction* 58.2 (1998), S. 302–311 (siehe S. 3).