Exposé zur Bachelorarbeit
VISUALISIERUNG VON PROTEIN-PROTEIN INTERAKTIONS-NETZWERKEN

Jens Pöthig

Humboldt Universität zu Berlin
Institut für Informatik
11. März 2009

1 Einführung

Neben den Nucleinsäuren, Polysacchariden und Lipiden gehören Proteine zu der wichtigsten Gruppe der organischen Biomoleküle und bilden einen integralen Bestandteil der in lebenden Organismen stattfindenden Prozesse[1]. Sie bestehen aus einer Sequenz von Aminosäuren, wobei die genaue Reihenfolge als Primärstruktur bezeichnet wird. Helix- und blattartige Verbindungen dieser Peptidketten bilden die Sekundärstruktur, räumliche Faltung die Tertiärstruktur[2]. In Organismen erfüllen Proteine eine Vielzahl von Funktionen. Hierzu gehören unter anderem die enzymatische Katalyse (nahezu alle Enzyme sind Proteine, die die chemische Umsetzung von Stoffen in Organismen steuern), die Erzeugung und Übertragung von Nervenimpulsen (spezifische Reize von Nervenzellen werden durch Rezeptorproteine vermittelt) sowie die Kontrolle von Wachstum und Differenzierung (Proteine regeln die koordinierte und zeitlich abgestimmte Expression von Geninformationen).

Die medizinische Forschung beschäftigt sich fortwährend mit der Entschlüsselung der Erbinformation einer Vielzahl von Lebewesen. Nach dem aktuellen Stand sind im menschlichen Organismus etwa 25000 Gene für die Kodierung von Proteinen verantwortlich[3]. Die Bioinformatik bietet eine Reihe von Möglichkeiten die Funktionen dieser Proteine vorherzusagen[4]. Das geschied zum Beispiel durch die Integration und Analyse von heterogenen Datenquellen (Genexpressionsdaten, Daten aus Text-Mining, Proteininteraktionsdaten), die RNA Strukturanalyse, sowie den Vergleich und die Klassifikation von 3D Proteinstrukturen unter Verwendung vielfältigster Algorithmen und Modelle.

Neben der Analyse von DNA- und Aminosäure-Sequenzen sowie der Molekülstrukturanalyse (Folding) ist die Betrachtung von Protein - Protein - Interaktionsnetzwerken ein vielversprechender Forschungszweig[5, 6]. Da es die Technik ermöglicht in kürzester Zeit die Erbinformation verschiedener Spezies zu sequenzieren, wächst die dem wissenschaftlichen Umfeld zur Verfügung stehende Datenbasis stetig. Sind innerhalb einer Spezies Annotationen vorhanden, tragen Proteininteraktionsdaten entscheidend dazu bei, über Orthologien

auf die Funktionen von Proteinen einer anderen Spezies zu schließen, die Annotationen zu übertragen und somit den Informationsgehalt zu steigern[7]. Die Erkenntnisse, die bei diesem Prozess gewonnen werden dienen vor allem dem tieferen Verständnis der äußerst komplexen biochemischen Vorgänge in Organismen und im Zuge dessen der anwendungsspezifischen Entwicklung von Medikamenten bzw. der Verfeinerung von neuartigen Therapiemöglichkeiten[8].

2 Projektrahmen und Ziele

Zur Untersuchung von Protein-Protein-Interaktionsnetzwerken stehen der Öffentlichkeit diverse Datenquellen zur Verfügung. Diese bieten einen unterschiedlichen Umfang an Funktionalität und Informationsgehalt.

Wie in den Artikeln von Jäger und Leser beschrieben[9, 10], soll in einem in der Arbeitsgruppe laufenden Projekt versucht werden, bisher unbekannte Funktionen von Proteinen mit einer sehr hohen Genauigkeit vorherzusagen und somit die Annotierung zu verbessern.

Im Rahmen dieser Forschungsarbeit ist eine MySQL Datenbank entstanden, die PPI-Netzwerkdaten von 7 verschiedenen Quellen (BIND[1], BioGRID[2], DIP[3], HPRD[4], IntAct[5], MINT[6] und MIPS[7]) integriert um eine möglichst breite Informationsbasis zu schaffen. Folgende Spezies sind erfasst: *Caenorhabditis elegans, Drosophila melanogaster, Homo sapiens, Mus musculus, Rattus norvegicus und Saccharomyces cerevisiae*. Um die Analyse und Auswertung der Daten zu erleichtern und so zum Beispiel schwach annotierte Bereiche zu finden bietet sich eine visuelle Repräsentation der Netzwerke an[11].

„Graphics *reveal* data. Indeed graphics can be more precise and revealing than conventional statistical computations.“ Edward R. Tufte, *The Visual Display of Quantitative Information*.

In diesem Sinne ist es Ziel dieser Arbeit sein, unter bestimmten Auswahlkriterien, die zur Verfügung stehenden PPI-Netzwerke aus der Datenbank zu extrahieren, Netzwerkmetriken zu berechnen und die Netzwerke grafisch darzustellen[12, 13].

Die zur Verfügung stehenden Daten stellen eine im biologischen Kontext attribuierte Knoten-/Kantenmenge dar. Eine reichhaltigere grafische Darstellung dieser Netzwerke erfordert jedoch, die strukturinhärenten Netzwerkmerkmale zu parametrisieren und bereitzustellen, um damit weitere Layout-Algorithmen ausführen zu können. Diese Algorithmen erzeugen nach bestimmten Vorgaben eine visuelle Repräsentation der Knoten-/Kantenmenge. Unter der Vielzahl der existierenden Layouts ist das Force-directed oder auch Spring-Embedded-Layout wohl das am häufigsten angewendete[14]. Hierbei werden bestimm-

[1] http://www.bind.ca/ Biomolecular Interaction Network Database
[2] http://www.thebiogrid.org/ General Repository for Interaction Datasets
[3] http://dip.doe-mbi.ucla.edu/ Database of Interacting Proteins
[4] http://www.hprd.org/ Human Protein Reference Database
[5] http://www.ebi.ac.uk/intact
[6] http://mint.bio.uniroma2.it/ Molecular INTeraction database
[7] http://mips.gsf.de/genre/proj/yeast/ munich information center for protein sequences

te Abstandsmaße zwischen den Knoten als Positionierungskriterium verwendet. Weiterhin kann jedes zur Verfügung stehende Knoten-/Kantenattribut in einen Visualisierungskontext gesetzt werden. Das in der Bioinformatik etablierte Werkzeug Cytoscape[15, 16] eignet sich hervorragend für diese Aufgaben. Somit soll im Rahmen dieser Arbeit ein Plugin zum MySQL-Datenimport für Cytoscape implementiert werden, welches die oben genannte Funktionalität bereitstellt.

3 Strategie

Um die aufgestellten Ziele zu erreichen sind die folgenden Arbeitsschritte nötig. Im Rahmen der Arbeit wird ausschließlich ein Plugin für die Plattform Cytoscape implementiert. Hierfür wird Java als Programmiersprache verwendet. Die statistischen Berechnungen übernimmt Rserve, ein R-Client für Java[17]. Für die visuelle Repräsentation der Daten ist Cytoscape verantwortlich, das heißt, Layout-Algorithmen sind nicht Teil dieser Arbeit.

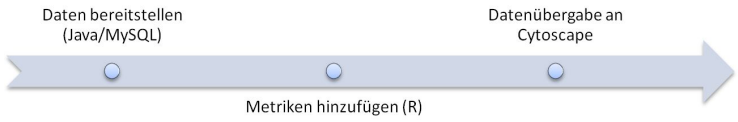

Abbildung 1: Bearbeitungsschritte

Zunächst wird über JDBC eine Verbindung zur MySQL Datenbank hergestellt, sowie der Datenbankstatus abgerufen. Dem Benutzer ist die Wahl des Servers überlassen, die Serveradresse des Lehrstuhls ist voreingestellt. Als nächstes werden auf Basis der zur Verfügung stehenden Tabellen die Statements zum extrahierenden der PPI Netzwerke vorbereitet. Dem Benutzer wird die Möglichkeit gegeben, die Netzwerke genauer zu spezifizieren. Das heißt, es können zum Beispiel nur Netzwerke einer bestimmten Spezies extrehiert werden, oder die Detektionsmethode kann ein Kriterium sein. Bestimmte Voreinstellungen werden implementiert. Ein nützliches Feature wäre an dieser Stelle, dem Benutzer einen SQL-Query-Editor bereitzustellen, damit nach benutzerspezifischen Wünschen ein Resultset generiert werden kann. Der nächste Schritt ist die Auswahl der zu berechnenden Netzwerkmetriken. Voreingestellt sind Degree-Centrality, Betweenness-Centrality, Closeness-Centrality, sowie PageRank. Weitere noch festzusetzende Metriken stehen ebenfalls zur Wahl.

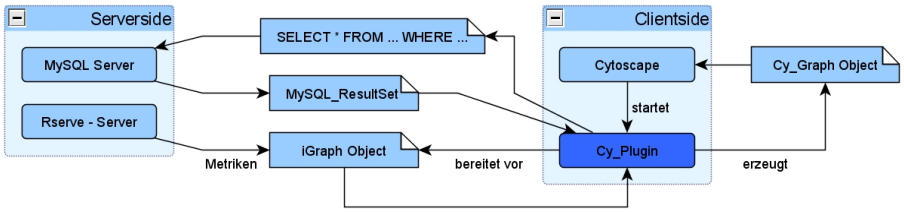

Abbildung 2: Ablauf

Zu diesem Zeitpunkt stehen alle Eingabeparameter fest und das Plugin kann anfangen zu arbeiten. Es wird ein SQL-Statement an den Server geschickt, worauf dieser nach einiger Zeit das Anfrageergebnis bereitstellt. An dieser Stelle stehen die Netzwerkdaten zur weiteren Bearbeitung bereit. Sie werden mit Rserve, einem R-Client für Java in das iGraph Format überführt, so können nun die verschiedenen Metriken berechnet und den bestehenden Daten hinzugefügt werden. Im Anschluss daran wird das Netzwerk an Cytoscape übergeben und steht dort zur weiteren Bearbeitung, sowie insbesondere der Visualisierung zur Verfügung. Cytoscape[18] bietet darüber hinaus auch die Möglichkeit das Netzwerk in verschiedenen Formaten zu exportieren.

Literatur

[1] CHRISTEN, Philipp ; JAUSSI, Rolf: *Biochemie*. Springer, 2005 (Springer-Lehrbuch XVI). – 635 S. – ISBN: 978-3-540-21164-8

[2] UETZ, Peter ; POHL, Ehmke: Protein-Protein- und Protein-DNA-Interaktionen. In: *Wink et al., Molekulare Biotechnologie, Wiley-VCH* (2004)

[3] SULTAN, Marc ; SCHULZ, Marcel H. ; RICHARD, Hugues ; MAGEN, Alon ; KLINGENHOFF, Andreas ; SCHERF, Matthias ; SEIFERT, Martin ; BORODINA, Tatjana ; SOLDATOV, Aleksey ; PARKHOMCHUK, Dmitri ; SCHMIDT, Dominic ; O'KEEFFE, Sean ; HAAS, Stefan ; VINGRON, Martin ; LEHRACH, Hans ; YASPO, Marie-Laure: A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. In: *Science* 321 (2008), Aug, Nr. 5891, 956–960. http://dx.doi.org/10.1126/science.1160342. – DOI 10.1126/science.1160342

[4] TRAJANOSKI, Z.: GEN-AU Projekt: Bioinformatik Integrationsnetzwerk. In: *Interuniversitäre Forschungsprojekte* (WS 05/06), S. 23–24

[5] SCHMITT, Stefan ; KUHN, Daniel ; KLEBE, Gerhard: A New Method to Detect Related Function Among Proteins Independent of Sequence and Fold Homology. In: *Journal of Molecular Biology* 323 (2002), Nr. 2, 387 - 406. http://dx.doi.org/DOI:10.1016/S0022-2836(02)00811-2. – DOI DOI: 10.1016/S0022–2836(02)00811–2. – ISSN 0022–2836

[6] ORF: *ORF ON Science – Neues Verständnis für Proteinfunktionen*. http://science.orf.at/science/news/24146

[7] MATHIVANAN, Suresh ; PERIASWAMY, Balamurugan ; GANDHI, TKB ; KANDASAMY, Kumaran ; SURESH, Shubha ; MOHMOOD, Riaz ; RAMACHANDRA, YL ; PANDEY, Akhilesh: An evaluation of human protein-protein interaction data in the public domain. In: *BMC Bioinformatics* 7 (2006), Nr. Suppl 5, S. S19. http://dx.doi.org/10.1186/1471-2105-7-S5-S19. – DOI 10.1186/1471–2105–7–S5–S19. – ISSN 1471–2105

[8] CLARKE, Paul A. ; POELE, Robert te ; WOOSTER, Richard ; WORKMAN, Paul: Gene expression microarray analysis in cancer biology, pharmacology, and drug development: progress and potential. In: *Biochemical Pharmacology* 62 (2001), Nr. 10, 1311 - 1336. http://dx.doi.org/DOI:10.1016/S0006-2952(01)00785-7. – DOI DOI: 10.1016/S0006–2952(01)00785–7. – ISSN 0006–2952

[9] JAEGER, S. ; LESER, U.: High-Precision Function Prediction using Conserved Interactions. In: *German Conference on Bioinformatics (GCB)* (2007), S. 145–162

[10] JAEGER, Samira ; GAUDAN, Sylvain ; LESER, Ulf ; REBHOLZ-SCHUHMANN, Dietrich: Integrating protein-protein interactions and text mining for protein function prediction. In: *BMC Bioinformatics* 9 Suppl 8 (2008), Jul, S2. http://dx.doi.org/10.1186/1471-2105-9-S8-S2. – DOI 10.1186/1471–2105–9–S8–S2

[11] SUDERMAN, Matthew ; HALLETT, Michael: Tools for visually exploring biological networks. In: *Bioinformatics* 23 (2007), Oct, Nr. 20, 2651–2659. http://dx.doi.org/10.1093/bioinformatics/btm401. – DOI 10.1093/bioinformatics/btm401

[12] SCHREIBER, Falk: *Visualization*. http://dx.doi.org/10.1007/978-1-60327-429-6_23. Version: 2008

[13] HOLDEN, Brian J. ; PINNEY, John W. ; LOVELL, Simon C. ; AMOUTZIAS, Grigoris D. ; ROBERTSON, David L.: An exploration of alternative visualisations of the basic helix-loop-helix protein interaction network. In: *BMC Bioinformatics* 8 (2007), 289. http://dx.doi.org/10.1186/1471-2105-8-289. – DOI 10.1186/1471–2105–8–289

[14] HAN, Kyungsook ; BYUN, Yanga: Three-dimensional visualization of protein interaction networks. In: *Computers in Biology and Medicine* 34 (2004), Nr. 2, 127 - 139. http://dx.doi.org/DOI:10.1016/S0010-4825(03)00045-3. – DOI DOI: 10.1016/S0010–4825(03)00045–3. – ISSN 0010–4825

[15] SHANNON, Paul ; MARKIEL, Andrew ; OZIER, Owen ; BALIGA, Nitin S. ; WANG, Jonathan T. ; RAMAGE, Daniel ; AMIN, Nada ; SCHWIKOWSKI, Benno ; IDEKER, Trey: Cytoscape: a software environment for integrated models of biomolecular interaction networks. In: *Genome Res* 13 (2003), Nov, Nr. 11, 2498–2504. http://dx.doi.org/10.1101/gr.1239303. – DOI 10.1101/gr.1239303

[16] CLINE, Melissa S. ; SMOOT, Michael ; CERAMI, Ethan ; KUCHINSKY, Allan ; LANDYS, Nerius ; WORKMAN, Chris ; CHRISTMAS, Rowan ; AVILA-CAMPILO, Iliana ; CREECH, Michael ; GROSS, Benjamin ; HANSPERS, Kristina ; ISSERLIN, Ruth ; KELLEY, Ryan ; KILLCOYNE, Sarah ; LOTIA, Samad ; MAERE, Steven ; MORRIS, John ; ONO, Keiichiro ; PAVLOVIC, Vuk ; PICO, Alexander R. ; VAILAYA, Aditya ; WANG, Peng-Liang ; ADLER, Annette ; CONKLIN, Bruce R. ; HOOD, Leroy ; KUIPER, Martin ; SANDER, Chris ; SCHMULEVICH, Ilya ; SCHWIKOWSKI, Benno ; WARNER, Guy J. ; IDEKER, Trey ; BADER, Gary D.: Integration of biological networks and gene expression data using Cytoscape. In: *Nat Protoc* 2 (2007), Nr. 10, 2366–2382. http://dx.doi.org/10.1038/nprot.2007.324. – DOI 10.1038/nprot.2007.324

[17] GENTLEMAN, Robert ; IHAKA, Ross: *The R Project for Statistical Computing*. http://www.r-project.org/

[18] *Cytoscape: Analyzing and Visualizing Network Data*. http://www.cytoscape.org/