Exposé einer Diplomarbeit

# Graphbasierte Vorhersage von Proteinfunktionen

Christian Brandt

21. November 2007
Betreuer: Prof. Dr. Ulf Leser

## Einführung

Es gibt in der Bioinformatik eine Reihe von Methoden, um Funktionen von Proteinen vorherzusagen. Sie basieren unter anderem auf Aminosäuresequenzen, 3D-Strukturen, DNA-Sequenzen oder Interaktionsnetzwerken [8]. In einem solchen Netzwerk stellen die Knoten Proteine und die Kanten wechselseitige Interaktionen dar. Mit der zunehmenden Verfügbarkeit von Protein-Protein-Interaktionsdaten (PPI) werden diese für die Bestimmung von Proteinfunktionen interessanter. Schon Schwikowski et al. [9] zeigen, wie sie zur Vorhersage genutzt werden können.

Bekannte netzwerkbasierte Ansätze betrachten beispielsweise die Nachbarschaft eines Knotens, um seine Funktion vorauszusagen [6]. Andere stützen sich auf die Hypothese, daß stark verbundene Subgraphen (Cluster) funktionale Module darstellen, denen eine spezifische biologische Funktion zugeschrieben werden kann [10]. Innerhalb eines solchen Moduls können dann Funktionsannotationen übertragen werden. Wieder andere nutzten dafür die Orthologiebeziehungen zwischen Proteinen in Netzwerken verschiedener Organismen [4].

Der Begriff der Proteinfunktion ist stark kontextabhängig und nicht klar definiert. Um Proteinen automatisiert Funktionen zuschreiben zu können, ist daher ein standardisierter Wortschatz nötig. Das Gene-Ontology-Projekt (GO) [2] bietet ein solches kontrolliertes Vokabular zur Annotation von Genen und Genprodukten.

### Ausgangspunkt der Arbeit

Jaeger und Leser beschreiben in ihrem Artikel [3], wie konservierte Teilnetzwerke in PPI-Netzwerken verschiedener Organismen identifiziert werden können. Mit Hilfe dieser Teilnetzwerke lassen sich GO-Annotationen einzelner Proteine speziesübergreifend vorhersagen. Das Verfahren ist in folgende Schritte gegliedert:

1. Finden orthologer Proteine aufgrund von Alignmentscores ihrer Aminosäuresequenzen mit anschließendem multipartiten Matching
2. Erkennung konservierter Cluster, d.h. zusammenhängender Subgraphen, die in allen betrachteten Netzwerken vorkommen
3. Berechnung von Ähnlichkeitsscores für die GO-Annotationen orthologer Proteine und von Ähnlichkeitsscores für die Cluster
4. Bestimmung der Proteingruppen, deren Ähnlichkeitsscore signifikant unter dem jeweiligen Clusterscore liegt
5. Vorhersage von Funktionen für die wenig oder gar nicht annotierten Proteine dieser Gruppen

Dieser Prozess soll in der Diplomarbeit modifiziert werden. Die Orthologie zweier Proteine wird nicht am Anfang festgelegt, sondern verschiedene Informationen über funktionstragende Beziehungen zwischen Proteinen fließen als typisierte Kanten in ein Netzwerkmodell ein. Sequenzähnlichkeit ist dann eine Kantenart. Ein noch festzulegender Algorithmus sucht darin dann konservierte Subgraphen.

### Weitere verwandte Arbeiten

Motive in einem Netzwerk sind Subgraphen, die sehr viel häufiger dort als in zufälligen Netzwerken auftreten. Jin Chen et al. [1] stellen eine Methode vor, welche Motive in einem PPI-Netzwerk findet und annotiert, um Proteinfunktionen vorherzusagen. Sie basiert auf einer Heuristik zur Erkennung topologischer Motive (ohne Label) und auf Ähnlichkeitsscores zwischen GO-Termen, Proteinen und ganzen Subgraphen.

Nariai et al. [7] beschreiben, wie sie eine Reihe heterogener Informationen, wie zum Beispiel PPI-Daten, Genexpressionsdaten, Proteinmotive und Knock-out-Phänotypdaten, in einen Graphen integrieren. Eine gewichtete Kante steht dort für die Evidenz einer funktionalen Ähnlichkeit der verbundenen Proteine. Mit einem auf Bayes-Netzwerken basierenden Ansatz sagen sie dann Funktionen voraus.

## Ziele

Das Hauptanliegen der Arbeit ist eine gute Vorhersage von Proteinfunktionen. Dafür wird ein Graphmodell erarbeitet, das verschiedene Indizien funktionaler Ähnlichkeit von Proteinen in ein Netzwerk integriert. In diesem Netzwerk sollen Teilnetzwerke gesucht werden, die in unterschiedlichen Spezies die gleiche Rolle spielen. Idealerweise stellt ein solches Teilnetzwerk ein bestimmtes funktionales Modul dar, das in den betrachteten Arten konserviert ist. Mit Hilfe einer Heuristik sollen diese Subgraphen erkannt werden. Für die spärlich annotierten Knoten darin werden anschließend Annotationen vorgeschlagen, die aus den entsprechenden Teilnetzwerken anderer Spezies abgeleitet wurden. Die Qualität der Vorhersagen soll mit einem Kreuzvalidierungsverfahren überprüft werden.

## Vorgehen

Die Diplomarbeit unterteilt sich in folgende Etappen:

- Graphmodell festlegen
- Algorithmus zur Erkennung der Subgraphen
- Umsetzen des Algorithmus
  - Datenaufbereitung
  - Subgraphen finden und Funktionsvorhersage
  - Evaluation

### Graphmodell

Ein Teil der Aufgabe ist das Aufstellen eines geeigneten Modells, in das sich die verschiedenen Informationen integrieren lassen. Es könnte etwa so aussehen: Für jede Spezies gibt es ein Interaktionsnetzwerk. Die Knoten stellen Proteine, die Kanten Interaktionen dar. Weiterhin gibt es gewichtete Kanten zwischen Proteinen unterschiedlicher Spezies, wenn ihre Sequenzen einen hohen Alignmentscore haben oder wenn ihre GO-Annotationen eng verwandt sind. Zwischen den Proteinen innerhalb einer Spezies gibt es noch Kanten, wenn ihre Gene auf dem Chromosom nahe beieinander liegen. Andere Informationen, die auf ähnliche Funktionen hindeuten, lassen sich leicht durch weitere Kanten hinzufügen. Abbildung 1 zeigt ein Beispiel des beschriebenen Modells. Die Proteine $b_1$ und $b_2$ werden dort als orthologes Paar eingestuft. Da die Gene von $b_1$ und $c_1$ sowie die von $b_2$ und $c_2$ benachbart sind, werden auch $c_1$ und $c_2$ als funktional ähnlich angesehen.

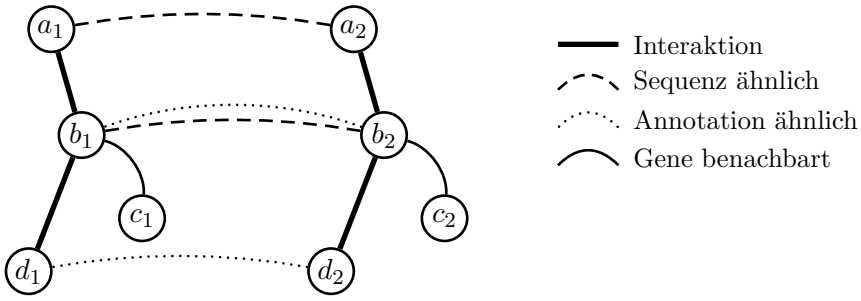


Abbildung 1: Hypothetischer, konservierter Subgraph. Gleiche Buchstaben stehen für orthologe Proteine. Der Index nummeriert die Spezies.

Einige Parameter, die für die Erstellung des Graphen nötig sind, werden während der Arbeit festgelegt oder angepaßt, wie zum Beispiel der Schwellwert für Sequenzähnlichkeit, ab dem eine Kante zwischen den Proteinen gezogen wird.

### Algorithmus

Die Bestimmung der Subgraphen soll als Optimierungsproblem formuliert werden. Der Kern der Arbeit ist die Entwicklung eines Approximationsalgorithmus für das zugehörige Suchproblem. Verschiedene bekannte Ansätze und lokale Optimierungsstrategien werden dafür untersucht.

### Umsetzung

Die für die Implementation nötigen Interaktionsdaten und Annotationen werden vom Lehrstuhl Wissensmanagement in der Bioinformatik[1] in Form einer Datenbank zur Verfügung gestellt. Die Proteine werden dabei über ihre UniProt-ID [11] identifiziert. Folgende Spezies sind erfasst: Caenorhabditis elegans, Drosophila melanogaster, Homo sapiens, Mus musculus, Rattus norvegicus und Saccharomyces cerevisiae. In der Arbeit wird eine konkrete Auswahl getroffen. Die Daten für die chromosomale Lokation von Genen können von NCBI Gene[5] bezogen werden.

Nachdem der Graph konstruiert und die Teilnetzwerke ähnlicher Funktion erkannt wurden, werden für die wenig annotierten Proteine Funktionen, d.h. GO-Terme, vorgeschlagen.

Die Ergebnisse sollen durch Kreuzvalidierung überprüft werden. Dafür werden bei einer Teilmenge der bekannten Proteine die Annotionen weggelassen. Die Verbleibenden dienen als Eingabe. Die Genauigkeit des Verfahrens ergibt sich aus dem Verhältnis richtiger Vorhersagen zur Gesamtzahl der Vorhersagen. Das Ganze wird mit weiteren Teilmengen wiederholt.

Die Implementierung aller Datenstrukturen und Methoden erfolgt in Java.

## Literatur


[1] Jin Chen, Wynne Hsu, Mong Li Lee, and See-Kiong Ng. Labeling network motifs in protein interactomes for protein function prediction. In *ICDE*, pages 546–555. IEEE, 2007.

[2] Gene Ontology Consortium. The Gene Ontology (GO) project in 2006. *Nucleic Acids Research*, 34(Database issue):D322–D326, January 2006.

[3] Samira Jaeger and Ulf Leser. High-precision function prediction using conserved interactions. In *GCB*, 2007. (to appear).

[4] Brian P. Kelley, Roded Sharan, Richard M. Karp, et al. Conserved pathways within bacteria and yeast as revealed by global protein network alignment. *Proceedings of the National Academy of Sciences*, 100(20):11394–11399, September 2003.

[5] Donna Maglott, Jim Ostell, Kim D. Pruitt, and Tatiana Tatusova. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Research*, 33(Database Issue):D54–D58, January 2005.


[1] `http://www.informatik.hu-berlin.de/forschung/gebiete/wbi`

[6] Jason McDermott, Roger Bumgarner, and Ram Samudrala. Functional annotation from predicted protein interaction networks. *Bioinformatics*, 21(15):3217–3226, August 2005.

[7] Naoki Nariai, Eric D. Kolaczyk, and Simon Kasif. Probabilistic protein function prediction from heterogeneous genome-wide data. *PLoS ONE*, 2(3):e337, March 2007.

[8] Gaurav Pandey, Vipin Kumar, and Michael Steinbach. Computational approaches for protein function prediction: A survey. Technical Report TR 06-028, Department of Computer Science and Engineering, University of Minnesota, Twin Cities, 2006.

[9] Benno Schwikowski, Peter Uetz, and Stanley Fields. A network of protein-protein interactions in yeast. *Nature Biotechnology*, 18(12):1257–1261, December 2000.

[10] Victor Spirin and Leonid A. Mirny. Protein complexes and functional modules in molecular networks. *Proceedings of the National Academy of Sciences*, 100(21):12123–12128, October 2003.

[11] The UniProt Consortium. The Universal Protein Resource (UniProt). *Nucleic Acids Research*, 35(Database issue):D193–D197, January 2007.