Exposé zur Masterarbeit

Konservierte Cluster in Protein-Interaktionsnetzwerken

Samira Jaeger

Betreuer: Prof. Dr. Ulf Leser, Prof. Dr. Knut Reinert

April 2006 - September 2006

Motivation

Protein-Protein Interaktionen (PPIs) bilden die Grundlage für alle biologischen Prozesse. Sie lassen sich durch verschiedene experimentelle Methoden detektieren und charakterisieren [5]. In jüngster Zeit führten Fortschritte in der Entwicklung von Hochdurchsatz-Methoden zu einer enormen Zunahme von experimentellen Daten, die eine verbesserte Klassifizierung von Protein Interaktionen ermöglichen. Es wurden bereits Proteom-weite systematische Proteininteraktions-Analysen für einige Modellorganismen, wie *S. cerevisiae*, *H. pylori*, *D. melanogaster* und *E.Coli* durchgeführt. Diese Analysen führten zur Erstellung umfangreicher Interaktionsnetzwerke, die komplexe Beziehungen zwischen Proteinen eines Organismus darstellen. Für Mensch und Maus sind im Moment noch keine vollständigen Interaktionsnetzwerke erstellt worden, aber in Anbetracht des großen Potentials von PPIs zum Verständnis von Krankheitsmechanismen und Signalwegen sind spezifische small-scale Proteinnetzwerke (Huntington-Disease, TGF-Signalweg) vorhanden. Außerdem wurden zusätzlich durch bioinformatische Analysen Interaktionsinformationen aus Hypothesen gesteuerten Studien und durch Identifikation konservierter orthologer Interaktionen gesammelt und erfasst [9].

An vielen biologischen Prozessen, wie Signaltransduktion oder Transkription, sind häufig zwischen 5 und 25 Proteine, die Proteinkomplexe bilden, beteiligt. Diese sind in Interaktionsnetzwerken als Gruppen von stärker verbundenen Knoten (Cluster) zu finden [8]. Vergleiche von Netzwerken verschiedener Organismen ermöglichen das Identifizieren von Clustern, die in den Organismen konserviert sind [7]. Unter der Annahme, dass gleiche Proteincluster auch gleiche biologische Funktionen besitzen, ist es möglich bekannte Annotationen eines Clusters eines Organismus, auf noch nicht charakterisierte, aber konservierte Cluster zu übertragen. Ebenso können da-

durch uncharakterisierte Proteine annotiert und die Evolution von interagierenden Systemen untersucht werden.

Ziel

Die Zielstellung dieser Masterarbeit besteht einerseits in der Identifizierung möglichst großer konservierter Teilinteraktionsnetzwerke ("fast" Cluster), die in mehreren Netzwerken enthalten sind. Als Daten kommen dabei *S. cerevisiae*, *D. melanogaster*, *E.Coli*, *H. sapiens* und *M. musculus* in Frage. Des Weiteren werden vorhandene funktionelle Annotationen der Cluster zwischen den betrachteten Organismen quantitativ verglichen, um die in der Einleitung genannte Hypothese, dass strukturelle Konservierung auch funktionelle Konservierung bedingt, zu validieren. Die Möglichkeit zur Übertragung von funktionaler Annotationen auf noch uncharakterisierte Proteine wird an ausgewählten Beispielen untersucht.

Vorgehen

Die Masterarbeit wird in drei Bereiche unterteilt:

- Datenaufbereitung
- Algorithmus zur Detektion konservierter Teilnetzwerke
- Untersuchung der funktionalen Annotationen

Datenaufbereitung

Für die algorithmische Identifizierung konservierter Subnetzwerke müssen die PPI-Netzwerke der zu vergleichenden Organismen zunächst in einem einheitlichen Eingabeformat bereitgestellt werden. Die entsprechenden Interaktionsdaten können von Datenbanken wie z.B. DIP [6] und BIND [3] heruntergeladen werden. Die dort enthaltenen Netzwerke variieren in ihrem Umfang und sind in Größen von 292 Interaktionen zwischen 202 Proteinen (Maus) bis zu 21000 Interaktionen und 7050 Proteinen (*Drosophila*) enthalten. Die Daten sind in der Regel in verschiedenen Formaten verfügbar, wie XIN, Tabulator-Separierte und PSI-MI (MIF) Formate, deren Hauptstrukturen ähnlich sind, die sich aber im Informationsgehalt und in den Detailbeschreibungen unterscheiden. XIN- und PSI-MI (MIF) Formate basieren auf XML und enthalten neben den eigentlichen Interaktionen auch Informationen über experimentelle Detektionsmethoden und Cross-Referenzen zu anderen Datenbanken. Diese Formate können mit Hilfe XML-Parsern aufbereitet werden, um das

jeweilige Interaktionsnetz für jeden Organismus als Adjazenzmatrix zu generieren. Neben den Interaktionsdaten ist außerdem die Speicherung von EC-Nummern [4] und Aminosäuresequenzen der Proteine geplant. Für die letzte Phase der Arbeit, der Untersuchung der Cluster, werden funktionale Annotationen wie Gene Ontology [2] benötigt, diese werden ebenfalls im Zuge der Datenaufbereitung erfasst und gespeichert.

Algorithmus zur Detektion konservierter Teilnetzwerke

Die Repräsentierung der Proteininteraktionsnetzwerke erfolgt mit Hilfe von Adjazenzmatrizen, deren Einträge (i,j) $\neq 0$ für zwei Proteine I und J sind, wenn eine Interaktion zwischen diesen beiden Proteinen vorhanden ist.

Zur Identifizierung gemeinsamer Subgraphen werden wir eine Abwandlung des Apriori-Algorithmus [1] anwenden. Da wir an den Subgraphen interessiert sind, die in allen Netzwerken vorkommen, wird ein Support von 100% verlangt.

Zu Beginn müssen die Proteine ermittelt werden, die in allen Netzwerken der Organismen vorkommen. Um dies umzusetzen benötigen wir ein Identitätsmaß, da Proteine mit gleicher Funktion in verschiedenen Organismen häufig unterschiedlich bezeichnet sind. Grundlage für ein solches Maß könnten EC-Nummern und Sequenzdaten bilden, wobei ein bestimmter Schwellenwert bei dem Vergleich von zwei Proteinsequenzen erreicht werden muss, damit zwei Proteine als identisch betracht werden können.

Proteine/Knoten in Proteininteraktionsnetzwerken sind eindeutig gekennzeichnet, somit können auch die Interaktionen/Kanten in den Graphen eindeutig, beispielsweise über die inzidenten Knoten, spezifiziert werden. Dadurch können Subgraphen auch an Hand einer Menge von Kanten repräsentiert werden.
Der Algorithmus zur Detektion konservierter Teilnetzwerke wird folgendermaßen funktionieren:

1. Basierend auf den ermittelten Proteinen, werden alle paarweisen Interaktionen bzw. Kanten zwischen diesen erfasst und auf ihre Häufigkeit überprüft. Diese muss 100% betragen, sonst werden die Kanten nicht weiter berücksichtigt.

2. Die Erweiterung der Kantenmengen der Größe k (für $k \geq 2$), um ein Element, erfolgt in zwei Schritten.

    a) Erzeugung eines nächst größeren (k+1)-Sets aus zwei Mengen der Größe k, deren Elemente sich nur um eine Kante unterscheiden. Dabei sollte diese Kandidatenkante zu mindestens einer, der in der k-Menge enthaltene, Kanten adjazent sein.

b) Überprüfung des erweiterten Kantensets auf 100% Support, falls dieser nicht gegeben ist, wird das erweiterte Set verworfen und im weiteren Verlauf des Algorithmus nicht weiter berücksichtigt.

Auf diese Weise werden kleinere Subgraphen, repräsentiert durch Kantensets, zu größeren zusammengesetzt, in dem diese jeweils um eine Nachbarkante erweitert werden, wenn die entsprechende Interaktion in allen Organismen vorhanden ist. Der Algorithmus endet, wenn keine zwei Kantensets zu einem größeren zusammengesetzt werden kann.

Die detektierten Subgraphen bzw. Cluster, dargestellten durch Kantensets, können mit Hilfe graphischer Tools wie beispielsweise *Cytoscape* visualisiert werden.

Untersuchung der funktionalen Annotationen

Unter der Voraussetzung, dass in den Proteininteraktionsnetzen Cluster von Proteinen bzw. gemeinsame Subgraphen zu finden sind, werden wir anschließend die Funktionsannotationen dieser Cluster untersuchen. Dies erfolgt mit Hilfe der *Gene Ontology* Annotationen, die während der Datenaufbereitung bereits ermittelt wurden. Zunächst wird überprüft, ob die gefundenen Cluster in den jeweiligen Organismen ein einheitliches funktionales Bild liefern. Dabei wird ein Cluster für die weitere Analyse nur berücksichtigt, wenn der durchschnittliche Abstand aller Funktionen der Proteine einen bestimmten Schwellenwert, gemessen über den GO Annotationsbaum, nicht überschreitet. Die einheitliche molekulare Funktion für den Cluster stellt dann der kleinste gemeinsame GO Vorfahrterm dar.

Neben der molekularen Funktion können auch die biologischen Prozesse betrachtet werden. Die detektierten Cluster sollten bestimmten biologischen Prozessen entsprechen. Um nähere Informationen über diese zu erhalten, werden die biologischen Prozesse erfasst, die mit allen, in den Clustern enthaltenen, Proteinen assoziiert sind. Sind keine identischen Prozesse vorhanden, könnte stattdessen ebenfalls der kleinste gemeinsame GO Vorfahrterm verwendet werden.

Die ermittelten Funktionen und assoziierten biologischen Prozesse der Cluster werden entsprechend repräsentiert und können anschließend zwischen den Clustern der verschiedenen Organismen verglichen werden.

Um zu überprüfen, ob die strukturelle Konservierung mit der funktionellen Konservierung korreliert, wird die Hypothese falsifiziert. Dabei wird die Struktur eines Netzwerk beibehalten und bei einem zweiten verändert (z.B. 1, 5, 10, 25, 50 % der Interaktionen). Dann werden gemeinsame Cluster identifiziert, deren Funktionen ermittelt und verglichen. Anschließend wird getestet, ob der Korrelationsfaktor

signifikant geringer ist.

An ausgewählten Beispielen wird die Möglichkeit zur Übertragung funktionaler Annotationen bzw. der Funktionsvorhersage noch uncharakterisierter Proteine untersucht. Besitzen, in einem konservierten Cluster, viele Proteine dieselbe Funktion, kann dies darauf hinweisen, dass die restlichen bisher unklassifizierten Proteine ebenfalls über diese bestimmte Funktion verfügen. Basierend auf dieser Annahme, wird eine Funktion auf uncharakterisierte Proteine eines Clusters übertragen, wenn

- der durchschnittliche Abstand aller Funktionen des Clusters einen Schwellenwert nicht überschreitet und dieser somit berücksichtigt wird
- mindestens die Hälfte der annotierten Proteine, diese bestimmte Funktion besitzen

Die Genauigkeit der Annotationsübertragung kann mit der Kreuzvalidierung getestet werden. Dabei werden nur die Proteine mit bekannten Annotationen berücksichtigt und in k Submengen, möglichst gleicher Größe, unterteilt. Die Annotationen einer der Gruppen werden jeweils mit Hilfe der restlichen k-1 Gruppen übertragen. Die Annotationen der jeweils getesteten Gruppe sind bekannt und somit kann überprüft werden, wie zuverlässig die oben genannte Methode der Annotationsübertragung bzw. Funktionsvorhersage ist.

Literatur

[1] Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules in large databases. In *VLDB*, pages 487–499, 1994.

[2] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nature Genetics*, 25(1):25–29, May 2000.

[3] Gary D. Bader, Doron Betel, and Christopher W. V. Hogue. BIND: the biomolecular interaction network database. *Nucleic Acids Research*, 31(1):248–250, 2003.

[4] Nomenclature Committee of the International Union of Biochemistry and Molecular Biology. *Enzyme Nomenclature*. Academic Press, 1992. ISBN: 0–122–27164–5.

[5] Eric M. Phizicky and Stanley Fields. Protein-protein interactions: Methods for detection and analysis. *Microbiological Reviews*, 59(1):94–123, March 1995.

[6] Lukasz Salwínski, Christopher S. Miller, Adam J. Smith, Frank K. Pettit, James U. Bowie, and David Eisenberg. The database of interacting proteins: 2004 update. *Nucleic Acids Research*, 32(Database-Issue):449–451, 2004.

[7] Roded Sharan, Silpa Suthram, Ryan M. Kelley, Tanja Kuhn, Scott McCuine, Peter Uetz, Taylor Sittler, Richard M. Karp, and Trey Ideker. Conserved patterns of protein interaction in multiple species. *PNAS*, 102(6):1974–1979, February 2005.

[8] Victor Spirin and Leonid A. Mirny. Protein complexes and functional modules in molecular networks. *PNAS*, 100(21):12123–12128, 2003.

[9] U. Stelzl, M.Lalowski U. Worm, C. Haenig, F.H. Brembeck, H. Goehler, M. Stroedicke, M. Zenker, A. Schoenherr, S. Koeppen, J. Timm, S. Mintzlaff, C. Abraham, N. Bock, S. Kietzmann, A. Goedde, E. Toksöz, A. Droege, S. Krobitsch, B. Korn, W. Birchmeier, Hans Lehrach, and E. E. Wanke. A human protein-protein interaction network: A resource for annotating the proteom. *Cell*, 122:957–968, September 2005.