Exposé zur Diplomarbeit

Orthologes Clustering auf multipartiten Proteingraphen

Enrico Maier

Betreuer: Prof. Dr. Ulf Leser

April 2007 - September 2007

Hintergrund
Die Identifikation von Proteinen unterschiedlicher Spezies, welche an den selben biologischen Prozessen beteiligt sind, ist eines der fundamentalen Probleme der Proteomforschung. Eine wichtige Vorraussetzung ist die Identifikation orthologer Proteine. Die Orthologie ist ebenso wie die Paralogie eine Form der Homologie, welche durch eine hohe funktionale Übereinstimmung zwischen Proteinen unterschiedlicher Spezies gekennzeichnet ist. Die Unterscheidung nach paraloge und orthologe Proteine erfolgt anhand der Entstehung der Homologie, so ist ersteres auf eine Genduplikation innerhalb eines Genoms und letzteres auf einer Speziation zweier Spezies aus einem gemeinsamen Vorfahren zurückzuführen [4]. Allerdings beruht diese Unterscheidung auf Hypothesen, da der genaue Verlauf der Evolution nicht bekannt ist.

Aktuelle Programme ([1], [6], [7], [9]) nutzen für die Identifikation orthologer Proteine Sequenzähnlichkeiten zwischen den Proteinen unterschiedlicher Proteome. Sie führen eine all-against-all BLAST Suche aus und markieren Proteinpaare als ortholog, wenn sie der jeweils beste BLAST Hit des Anderen sind. Diese paarweisen Cluster werden nach bestimmten applikationsbedingten Regeln mit paralogen Sequenzen ergänzt, welche durch eine ebenfalls hohe Sequenzähnlichkeit zu den orthologen Proteinpaar gekennzeichnet sind. Auf den paarweisen Orthologenbeziehungen werden bei [1], [6] und [9] zur Bestimmung orthologer Beziehungen zwischen mehr als zwei Proteomen heuristische Clusteralgorithmen ausgeführt. Im Ergebnis steht eine Menge von Clustern, welche sowohl orthologe als auch paraloge Proteine enthalten.

Zielstellung
Die Ergebnisse von [1], [6], [7] und [9] weisen für viele Applikationen und Forschungsvorhaben unerwünschte Eigenschaften auf. So enthalten viele Cluster mehr bzw. weniger Proteine als vorhandene Proteome oder ein und dasselbe Protein tritt in mehreren Clustern auf. Als Grundannahme dieser Arbeit soll ein optimales Matching zwischen den Proteinen unterschiedlicher Spezies hinsichtlich der Minimierung einer paarweisen Distanzfunktion gefunden werden. Das Ergebnis entspricht einer Menge eindeutiger Cluster, derart dass in jedem Cluster genau ein Protein jeder Spezies auftritt und die Cluster keine gemeinsamen Proteine besitzen.

Vorgehen

Die Proteine der unterschiedlichen Proteome können als multipartiter gewichteter Graph betrachtet werden. Die Partitionierung der Proteine entspricht der Spezieszugehörigkeit und die Wichtung der Kanten ist durch die Distanz zwischen den Proteinsequenzen bestimmt. Ein optimales Matching zwischen genau zwei Spezies entspricht dem paarweisen linearen Zuordnungsproblem, besser bekannt als Heiratsproblem, und eine Lösung wird mit Hilfe des Ungarischen Algorithmus gefunden [5]. Die Suche nach einen optimalen Matching zwischen mehr als zwei Spezies entspricht dem multidimensionalen Zuordnungsproblem und ist in der kombinatorischen Mathematik als Multi Index Assignment Problem bekannt. Dies gehört zu der Klasse der NP-vollständigen Problemen [8]. Eine Lösung für derartige Probleme kann nur mit Hilfe einer Heuristik gefunden werden, die einen Kompromiss zwischen der Berechnung in vertretbarer Zeit und der Qualität des Ergebnisses realisiert. In dieser Arbeit wird eine Heuristik von Bandelt et al. [3] implementiert und auf ihrer Eignung für das Finden eines optimalen Matchings zwischen den Proteinen vieler Spezies untersucht. Als Distanzfunktion wird zunächst die invertierte Sequenzähnlichkeit genutzt. Außerdem wird eine Verbesserung der Distanzfunktion in Bezug auf die Identifizierung wahrer Orthologer angestrebt. Dafür werden zusätzliche Parameter wie z.B. gemeinsame Domänen oder die Zugehörigkeit zu einer gemeinsamen Proteinfamilie der Distanzfunktion hinzugefügt.

Als Datengrundlage werden zwanzig Bakterienproteome der OrthoMCL Datenbank[10] verwendet. Die paarweisen Sequenzähnlichkeiten werden mit Hilfe des BLAST Algorithmus[2] berechnet. Hierfür wird das Programm blastall des NCBI-Toolkits[11] auf jeweils zwei Proteomen im FASTA-Format ausgeführt. Als expectation-value wird der Wert 10.0 angegeben, so dass BLAST nur die Sequenzähnlichkeiten berechnet und nicht vermutete zufällige Treffer herausfiltert. Anschließend werden alle redundanten Informationen(z.B. ProteinA ProteinB 2e-10 55 entspricht ProteinB ProteinA 2e-10 55) und alle Sequenzähnlichkeiten, welche einen geringeren Wert als 10 % Übereinstimmung aufweisen, aus dem BLAST Ergebnis herausgefiltert. Dies dient vor allem der Einschränkung der zu betrachtenden Proteinmenge.

Entsprechend der Heuristik von Bandelt et al.[3] wird zunächst eine suboptimale Clustermenge als Startlösung berechnet. Hierfür wird eine Kette von Teillösungen y^1, y^2,..., y^k konstruiert, wobei die Teillösung y^k die vorherigen Teillösung y^{k-1} um die Proteine der Spezies k mit Hilfe des Bipartiten Matchings erweitert. Diese Kettenkonstruktion wird für jede Spezies genau einmal gestartet und die beste der k gefundenen Lösung wird als Startlösung deklariert. Für die Optimierung der gefundenen suboptimalen Lösung wird iterativ jeweils ein Speziesindex L aus der Indexmenge K ausgewählt. Analog teilt sich damit die Lösung x in die beiden Teillösungen x^L und $x^{K\setminus L}$. Ein optimales Bipartites Matching findet die optimale Rekombination der Cluster der beiden Teillösungen und verbessert die Kosten einer gefundenen Lösung. Die Iteration wird solange fortgesetzt bis sich die Kosten einer gefundenen Lösung nicht mehr verbessern.

Das optimale Bipartite Matching wird mit Hilfe des Ungarischen Algorithmus[5] gefunden. Derzeit stellt dieser Algorithmus das schnellste Verfahren zum Finden eines optimalen Bipartiten Matchings dar. Er hat eine Laufzeitkomplexität von $O(n^3)$ und einen Speicherplatzbedarf von n^2, wobei n der Mächtigkeit der größeren Itemmenge entspricht. Der Algorithmus ist bereits im Text Clustering Toolkit[12] der Universität von Dublin implementiert und wird als API in die Applikation eingebunden.

Für die Gütemessung des Verfahrens werden synthetische Cluster den Proteomen zugefügt. Es werden aus einen externen Proteom Proteinsequenzen entnommen und einer mehrfache Punktmutation ausgesetzt. Diese mutierten Sequenzen werden anschließend den zu untersuchenden Proteomen hinzugefügt. Diese Cluster weisen sehr hohe paarweise Sequenzähnlichkeiten auf und sollten bei einem optimalen Matching auch als Cluster identifiziert werden. Ein weiteres Gütekriterium ist die Verbesserung der Kosten des Matchings im Vergleich zu den Kosten der Startlösung, welche als Ergebnis eines greedy-basierten Verfahrens betrachtet werden kann. Die Gütemessung erfolgt in Abhängigkeit von der Anzahl der zu untersuchenden Spezies und der Anzahl der Iterationen zur Optimierung der Startlösung.

Literatur

[1] Alexeyenko, A., I. Tamas, et al. (2006). "Automatic clustering of orthologs and inparalogs shared by multiple proteomes." Bioinformatics **22**(14): e9-15.

[2] Altschul, S. F., W. Gish, et al. (1990). "Basic local alignment search tool." J Mol Biol **215**(3): 403-10.

[3] Bandelt, H.-J., M. A., et al. (2004). "Local search heuristics for multi-index assignment problems with decomposable costs." Journal of the Operational Research Society **55**: 694-704.

[4] Fitch, W. M. (1970). "Distinguishing homologous from analogous proteins." Syst Zool **19**(2): 99-113.

[5] Frank, A. (2004). "On Kuhn's Hungarian Method – A tribute from Hungary." Egrervary Research Group.

[6] Li, L., C. J. Stoeckert, Jr., et al. (2003). "OrthoMCL: identification of ortholog groups for eukaryotic genomes." Genome Res **13**(9): 2178-89.

[7] Remm, M., C. E. Storm, et al. (2001). "Automatic clustering of orthologs and inparalogs from pairwise species comparisons." J Mol Biol **314**(5): 1041-52.

[8] Spieksma, F. C. R. (2001). "Multi Index Assignment Problems: Complexity, Approximation, Applications." Maastricht University, Department of Mathematics.

[9] Tatusov, R. L., M. Y. Galperin, et al. (2000). "The COG database: a tool for genome-scale analysis of protein functions and evolution." Nucleic Acids Res **28**(1): 33-6.

[10] http://orthomcl.cbil.upenn.edu/ORTHOMCL_DB/. 02/2007

[11] ftp://ftp.ncbi.nlm.nih.gov/blast/executables/LATEST-BLAST/. 03/2007

[12] http://mlg.ucd.ie/content/view/18/. 03/2007