Berlin, den 19.06.2002

Clusterung großer Proteinsequenzdatenmengen

(Originaltitel: Large Scale Clustering of Protein Sequences)

Antje Krause

Schlagworte: Sequenzdatenbanken; Proteinfamilien; Hierarchische Clusterung; Wirbeltierevolution

Zusammenfassung:
Mit dem enormen Wachstum biologischer Sequenzdatenbanken wird die Verarbeitung dieser Daten mehr und mehr zum Problem. Proteinsequenzdatenbanken enthalten heute über eine halbe Million verschiedener Sequenzen. Es ist daher naheliegend, eine Gruppierung evolutionär verwandter Sequenzen durchzuführen. Dies bietet viele Vorteile. Eine häufig gestellte Frage ist z.B. die Identifikation ähnlicher Sequenzen zu einer bisher unbekannten Sequenz. Diese Aufgabe kann schneller durchgeführt werden, wenn nur ein Sequenzvergleich pro Gruppe (Cluster) im Gegensatz zu einem Sequenzvergleich pro Datenbanksequenz durchgeführt werden muss. Der Informationsgehalt des gefundenen Clusters ist zudem höher, da sämtliche Informationen der im Cluster enthaltenen Sequenzen herangezogen werden. Darüber hinaus kann eine geclusterte Sequenzdatenbank die Auswahl von Kandidaten für die zeit- und kostenintensive Strukturbestimmung erleichtern. Ferner können die Sequenzen eines Clusters direkt zur Analyse ihrer evolutionären Beziehung herangezogen werden. Die biologisch sinnvolle Gruppierung von Proteinsequenzen kann auf der Basis von Sequenzähnlichkeit erfolgen. Eine der einfachsten und häufig verwendeten Methoden hierzu ist Single-Linkage Clustering. Ausgehend von einem Paar von Datenpunkten mit größter Ähnlichkeit werden sukzessive Datenpunkte bzw. bereits erzeugte Cluster zusammengefaßt. Die resultierende Hierarchie kann auch als Baum betrachtet werden. Die Blätter entsprechen hierbei den Datenpunkten, während die Wurzel ein großes Cluster mit allen Datenpunkten bildet. Alle Ebenen dazwischen entsprechen Clustermengen an verschiedenen Ähnlichkeitsstufen. Es ist jedoch nicht klar, welche dieser Ebenen einer sinnvollen Gruppierung der Daten entspricht bzw. ob es überhaupt eine Ebene gibt, die alle Daten sinnvoll gruppiert.

In dieser Arbeit werden verschiedene Methoden zur automatischen Gruppierung großer Proteinsequenzdatenmengen präsentiert und evaluiert. Zunächst wird basierend auf der iterative Datenbanksuchmethode SYSTERS (SYSTEmatic Re-Searching) eine mengentheoretischen Clusterung (SYSTERS1) abgeleitet. Die nachfolgende Methode SYSTERS2 ändert die Sicht der Daten auf einen graph-basierten Ansatz. Hierbei liegt der Schwerpunkt zunächst auf der Verbesserung der Qualität der Eingabedaten, insbesondere der paarweisen Distanzen der Sequenzen. Darauf aufbauend wird dann ein Single-Linkage Clustering an verschiedenen statischen Schwellwerten durchgeführt. Es stellt sich heraus, daß es keinen eindeutigen Schwellwert für alle Proteinfamilien gibt, da der Grad an Sequenzähnlichkeit innerhalb verschiedener Proteinfamilien stark variiert. Aufgrund dieses Ergebnisses wurde die ebenfalls graph-basierte Methode SYSTERS3 entwickelt, die eine Gruppierung der Sequenzdaten in Superfamilien- und Familien-Cluster erzeugt. Hierbei werden aufgrund der inneren Struktur des Single-Linkage Baumes zunächst Superfamilien abgeleitet. Für jede Superfamilie wird anschließend der entsprechende Distanzgraph an geeigneten Stellen weiter in Familien-Cluster getrennt. SYSTERS3 ist damit völlig unabhängig von statischen benutzerdefinierten Schwellwerten.

Der zweite Teil der Arbeit widmet sich der Rekonstruktion der Phylogenese der Wirbeltiere. Um verschiedene Hypothesen über weitreichende Gen- bzw. Genomduplikationen auf dem Weg zu den Wirbeltieren testen zu können, bedarf es zunächst wohl-separierter Proteinfamilien, die jeweils nur einen Repräsentanten in den Wirbellosen haben. Als Grundlage dienen hier die vorhergesagten Proteinsequenzen der komplett sequenzierten Genome von Fruchtfliege, Fadenwurm und Bäckerhefe. Im Gegensatz zu anderen Ansätzen ist die hier entwickelte Methode desweiteren in der Lage, auch Datensätze nicht komplett zur Verfügung stehender Genome einzubinden (z.B. Mensch, Maus, Ratte, Lanzettfischchen). Die resultierende Clustermenge COPSE (Clusters of Orthologous and Paralogous SEquences) bildet eine hilfreiche Basis zur Analyse der Wirbeltierevolution sowie zur funktionellen Annotation.

Beide Clustermengen wurden zusammen mit weiteren Informationen im Internet zur Verfügung gestellt.