Suche und Clustering mit Textreferenzen zu PDB-Strukturen (Related Proteins in Columba)

Exposé einer Bachelorarbeit

Marcus Pankalla

Betreuer: Prof. Dr. Ulf Leser,
Institut für Informatik, Humboldt-Universität zu Berlin
Prof. Dr.-Ing. Knut Reinert,
Institut für Informatik, Freie Universität Berlin

Bearbeitungszeit: 8 Wochen

Motivation
Primäre Datenbanken wie SWISSPROT [1], PDB [2] und ENZYME [3] halten selten alle Daten über die verzeichneten Proteine bereit, so dass man oft gezwungen ist, die interessanten Daten aus mehreren Quellen zusammenzutragen. Die integrierte Proteinannotationsdatenbank Columba [4] verfolgt das Ziel, alle relevanten Daten auf einen Blick darzustellen und für detailliertere Informationen auf die jeweiligen Quellen zu verweisen. Die Basis von Columba sind PDB-Proteinstrukturen, zu denen zusätzlich weitere Daten, wie z.B. die Klassifikation nach SCOP [5] und CATH [6] oder Informationen über die beteiligten Stoffwechselwege (KEGG [7]), annotiert werden.
Aktuelle Daten und Erkenntnisse über Proteine werden zunächst in wissenschaftlichen Fachzeitschriften veröffentlicht. Diese Artikel werden in Datenbanken wie PubMed [8] gesammelt und annotiert und stellen eine relevante Quelle für neue, aktuelle Informationen zu Proteinen oder Proteinstrukturen dar.

Zielsetzung
Das Ziel dieser Bachelorarbeit ist es, Einträgen in Columba die Abstracts zu einem oder mehreren Veröffentlichungen zuzuordnen, sofern eine solche Information verfügbar ist. Die Abstracts sollen komplett in Columba gespeichert werden. Zusätzlich indexieren wir diese Texte mit Hilfe von *tsearch2* [9], so dass sich die Datenbank über das Web-Interface danach durchsuchen lässt. Weiterhin sollen die Abstracts, und damit indirekt die Proteinstrukturen, in verwandte Gruppen eingeteilt werden. Verwandtschaft wird hierbei über die Ähnlichkeit von Texten definiert. Das Ziel dieses *Clusterings* ist es, zu einer PDB-Proteinstruktur mehrere ähnliche Proteine zu finden.

Vorgehen
Zunächst müssen zu den Proteinstrukturen in Columba relevante PubMed-Einträge gefunden werden. Dazu bieten sich zwei Möglichkeiten an:
- viele Artikel sind bereits in den Quelleinträgen der PDB-Datenbank über PubMed-IDs verlinkt, jedoch nicht in Columba.
- Durchsuchen der SWISSPROT-Datenbank nach PubMed/Medline-IDs zu einem Protein, die Verknüpfung PDB ↔ Swissprot ist bereits in Columba vorhanden.

Beide Vorgehensweisen werden angewendet. Dazu werden Dumps der beiden Datenbanken verwendet, die offline mit einem Parser-Programm nach PubMed-IDs durchsucht werden. Die gefundenen Ergebnisse werden dann in Columba zusammen mit einem Verweis auf die jeweilige Quelldatenbank eingetragen. Dabei ist zu beachten, dass eine PDB-Struktur oft mehrere Proteinketten (chains) oder sogar unterschiedliche Proteine (z.B. ein Ligand – Rezeptor Komplex)

enthalten kann. Nach Möglichkeit soll die Literatur den einzelnen Chains zugeordnet werden. Die Artikel, die über die oben beschriebene Methode gefundenen werden, werden ebenfalls in Columba abgelegt und über das Web-Interface durchsuchbar gemacht.
Das Clustering der Artikel findet ebenfalls extern statt, und zwar mit Hilfe der Java-Bibliothek Weka [10]. Weka bietet mehrere Clustering-Algorithmen an, in dieser Arbeit wird ein einfacher *k-means* Algorithmus verwendet. Dazu müssen die Texte vorher aufbereitet werden:

- Zerlegung der Texte in *Token*, d.h. jedes Wort im Text wird beim Clustering als ein Attribut angesehen
- Herausfiltern häufig vorkommender Wörter (stopwords)
- Rückführung auf die Stammform (Bsp.: fold, folds, folding → fold)
- Speichern dieser Daten in ein Weka-spezifisches Dateiformat (ARFF, enthält Attribute und Daten)

Die aus dem Clustering erhaltenen Ergebnisse werden ebenfalls in Tabellen der Datenbank gespeichert. Dadurch lassen sich indirekt über die PubMed-Einträge auch die PDB-Strukturen, die Chains und die Proteine den unterschiedlichen Clustern zuordnen. Dabei kann es durchaus vorkommen, dass ein PDB-Eintrag in mehreren Clustern vorkommen kann, da die dazu gehörigen Abstracts zu verschiedenen Clustern gehören.

Im Verlauf dieser Arbeit sollen mehrere Experimente durchgeführt werden, um das Ergebnis des Clusterings zu verfeinern und zu verbessern. Zum Beispiel können zusätzliche Annotationen aus der PubMed-Datenbank übernommen werden. Dazu gehören Medical Subject Headings (MeSH), eine Art Schlagwortliste, und eine Liste der mit dem Artikel zusammen hängenden Chemikalien (chemicals).

Auf Grund der großen Menge der Abstracts können wir keine exakte Validierung der Ergebnisse vornehmen. Vielmehr soll ein Vergleich mit bekannten Klassifizierungen (SCOP, CATH) durchgeführt werden. Eine weitere Einordnung ist PDB-Cluster [11], wobei hier das Clustering auf Sequenzähnlichkeit beruht. Dabei werden wir überprüfen, wie nah unsere Ergebnisse an den vorhandenen liegen. Da das in dieser Arbeit vorgestellte Verfahren keine Nachbildung dieser Konzepte sein soll, dient der Vergleich eher als *proof-of-concept*, so dass sich die Ergebnisse mehr oder weniger stark unterscheiden können.

Referenzen

[1] ExPASy – Swiss-Prot
http://www.expasy.org/sprot/
[2] RCSB PDB – Protein Data Bank
http://www.rcsb.org/pdb/
[3] ENZYME – Enzyme nomenclature database
http://www.expasy.org/enzyme/
[4] Columba – A Database of Protein Structure Annotation
http://www.columba-db.de
[5] SCOP - Structural classification of protein folds
http://scop.berkeley.edu
[6] CATH – Protein Structure Classification
http://www.biochem.ucl.ac.uk/bsm/cath/cath.html
[7] KEGG: Kyoto Encyclopedia of Genes and Genomes
http://www.genome.ad.jp/kegg/
[8] Entrez PubMed
http://www.ncbi.nlm.nih.gov/entrez
[9] Tsearch2 - full text extension for PostgreSQL
http://www.sai.msu.su/~megera/postgres/gist/tsearch/V2/
[10] Ian H. Witten and Eibe Frank (2005) "Data Mining: Practical machine learning tools and techniques", 2nd Edition, Morgan Kaufmann, San Francisco, 2005. http://www.cs.waikato.ac.nz/ml/weka/
[11] PDB-CLUSTERS
http://www.rcsb.org/pdb/redundancy.html
http://www.rcsb.org/pdb/newsletter/2003q4/focus_clusters.html