Zusammenfassung

Mining Protein Family Specific Residue Packing Patterns From Protein Structure Graphs

von

Jun Huan, Wei Wang, Deepak Bandyopadhyay, Jack Snoeyink, Jan Prins,
Alexander Tropsha

geschrieben von Martin Jess

Einleitung:

Es wird im folgenden dargelegt wie durch Vereinfachung von Proteinstrukturen in einen Graphen das Problem der Charakterisierung von Proteinfamilien schnell und automatisch gelöst werden kann.
Die Kernidee dieser Arbeit besteht im erstellen von Graphen möglichst effizienter Natur, auf diese dann ein *subgraph mining algorithm* angewendet wird. Mit den so gewonnenen Informationen über Untergraphen und ihrer Verbindung zu den 3-D Strukturmustern wird eine *Support Vector Machine* trainiert um dann automatisch Klassifikationen vorzunehmen.
Diese *Support Vector Machine* wird durch ein Experiment getestet um so weitere Verbesserungen und Anwendungsfelder zu finden.

Problembeschreibung:

Im Bereich der Proteindatenbanken wird sich die Existenten von Proteinfamilien zu nutze gemacht um Proteine zu klassifizieren.
(Proteinfamilie etc.)
Proteinfamilien besitzen im allgemeinen bestimmte Struktureigenschaften wie ähnliche aktive Zentren oder andere funktionsbestimmende Strukturen. So das man über die Beziehung von Proteinfamilien zu Strukturmotiven sehr einfach eine vielzahl Informationen auf noch nicht näher charakterisierte Proteine übertragen kann. Durch einen automatisch Erkennung von Strukturmotiven und Proteinfamilie kann man vielleicht die Proteinklassifikation, Proteinfunktionsvorhersage oder die Proteinfaltungsvorhersage verbessern.

Methode:

Zum vereinfachen und besseren Nutzbarkeit werden die Proteine nach ihrer Primärstruktur in Graphen abgelegt. Es werden drei verschiedene

Graphenmodelle eingführt, der *Contact Distance Graph (CD)*, der *Delaunay tessellation Graph (DT)* und der *Almost Delaunay Graph (AD).*
Als Knoten dienen die Aminosäuren die über die Koordinaten der C_α -Atome und dem Namen sowie der Position in der Primärstruktur klassifiziert werden (Bsp. Asp112 an Position x,y,z).
Es werden Kanten eingeführt die je nach Art des Graphen nach drei verschieden Methoden bestimmt werden. In allen Graphen gibt es die Kanten für *direct bonds, d.h.* alle „echten“ Bindungen in der Proteinstruktur.
Die Kanten werden im *Contact Distance Graph (CD)* um alle Kanten erweitert für C_α -Atomen im Abstand von δ um den Knoten.
Im *Delaunay tessellation Graph (DT)* werden Kanten, für alle Knoten die vier C_α -Atome die ein Tetraeder mit einer leeren Kugel darin ergeben , erweitert.
Der *Almost Delaunay Graph (AD)* ist der *Delaunay tessellation Graph (DT)* erweitert mit weiteren Delaunay Kanten bei denen die Positionen der C_α -Atome um den Parameter c verändert wird.

Das Grundprogramm besteht aus zwei Hauptkomponenten, 1. die Berechnung aller oben beschriebenen Graphentypen für die in der Proteindatenbanken gespeicherten Proteine, 2. die Identifikation der Untergraphen.

(canonical adjacency matrix/ Graphenberechnung etc.)

Als Algorithmus für die Untergraphen suche dient der in *Listing 1* gezeigte *subgraph mining algorithm*.

FFSM

1: $S \leftarrow$ { the CAMs of the frequent nodes }
2: $P \leftarrow$ { the CAMs of the frequent edges }
3: FFSM-Explore(P, S);

FFSM-Explore (P, S)

1: **for** $X \in P$ **do**
2: **if** $(X.isCAM)$ **then**
3: $S \leftarrow S \cup \{X\}$
4: $C \leftarrow$ { all matrices $M \mid X$ is submatrix of M}
5: remove CAM(s) from C that is either infrequent or not optimal
6: FFSM-Explore(C, S)
7: **end if**
8: **end for**

Listing 1 ***subgraph mining algorithm***

(Post Processing / coherent subgraph etc.)

Auswertung:

Klassifikation durch Support Vector Machine (SVM), da diese zwei entscheidende Vorteile hat, 1. hoch-dimensionale Datenmengen und 2. verschiedene Sets für kernel learing functions.
Im Experiment wird das `libsvm` Programm mit radius kernel benutzt. (http://www.csie.ntu.edu.tw/~cjlin/libsvm/)

Experiment:

Folgende Voraussetzungen wurden gemacht ...

(Binary Classification of SCOP families)
(Protein Family Classification using Coherent Subgraph Counts as Variables)
(Signature Identification in Eukaryotic Serine Protease)

Auswertung: