Bachelorarbeit: Explorationsphase

Ben Haladik

29. Oktober 2015

Inhaltsverzeichnis

1 Wichtige Fragen **2**

2 Datensatz **2**
2.1 Proteine von Tim 2
2.2 Proteine aus der Literatur 3
2.3 Datenbanken und große Datensätze für die Analyse 3
2.4 Proteinwahl anhand von EC-Termen 3

3 Methoden **4**
3.1 Datengewinnung aus der PTGL 4
3.2 Pipeline 4

1 Wichtige Fragen

Um die Bachelorarbeit gut schreiben zu können sind einige wichtige Fragen zu beantworten. Zur Erinnerung der Titel der Arbeit soll lauten:
Bioinformatische Anwendung von Graphlets auf Proteinstrukturtopologiegraphen zur Ähnlichkeitsanalyse von Proteinen.
Wie bei vielen naturwissenschaftlichen Arbeiten ist er sehr sperrig und passt nicht in eine Zeile. Folgende Fragen sind für die Implementierung zu beantworten:

Welche Metrik soll ich benutzen, um die berechneten Vektoren zu vergleichen? Hier bietet es sich an in die Liste vorhandener Strukturalignment-Methoden zu schauen, was direkt die nächsten zwei Fragen liefert:

Wie clustere ich die Vektoren?

Welche Methoden soll ich benutzen, um sie mit dem `graphletAnalyser` zu vergleichen? An sich wäre es klug, diese drei Fragen zusammen zu beantworten. Bei der Recherche nach Strukturalignment-Methoden werde ich bestimmt auf einige Datensätze stoßen, die hofffentlich öffentlich verfügbar sind. Ich sollte auch eine Idee davon bekommen, welche Metriken in der Bioinformatik üblicherweise benutzt werden. Die große Frage ist natürlich und leider die, die von dieser Arbeit nicht beantwort werden kann:

Was ist ein biologisch sinnvolles Strukturalignment?

Dieser Frage will ich mich im Laufe der Arbeit nähern. Ich bin so aufgeregt!

2 Datensatz

2.1 Proteine von Tim

4-Helix-Bundles Cytochrome b562 (PDB: 1QPU) und human growth hormone (PDB: 1HGU)

Globin-Fold Hier gibt es viele Proteine mit niedriger Sequenzidentität und hoher struktureller Ähnlichkeit. Gute Kandidaten sind Hemoglobin, Myoglobine und Phycocyanine

TIM-Barrels Die Pyruvat-Kinase hat eine TIM-Barrel-Domäne (PDB: 1A3W). Des weiteren wurde die Triosephosphat-Isomerase vorgeschlagen (7TIM).

2.2 Proteine aus der Literatur

Ein Artikel mit dem Titel: *Protein Structure Comaprison by Alignment of Distance Matrices* (URL: `http://www.sciencedirect.com/science/article/`

pii/S0022283683714890) scheint laut Abstract gute Kandidaten liefern zu können. Ein weiterer Artikel behandelt die Frage, ob es eine einziges korrektes Strukturalignment gibt/geben kann und ist unter der URL: http://onlinelibrary.wiley.com/doi/10.1002/pro.5560050711/abstract zu finden.
Der Artikel: *Structural alignment of proteins by a novel TOPOFIT method, as a superimposition of common volumes at a topomax point* unter der URL: http://onlinelibrary.wiley.com/doi/10.1110/ps.04672604/full. beschreibt eine Strukturalignment-Methode, die versucht die Substrukturen mit der minimalen Abweichung zu finden.

2.3 Datenbanken und große Datensätze für die Analyse

Bisher scheinen sich zwei große Datenbanken für dei Analyse zu eignen.

ASTRAL ist ein Datensatz, der prinzipiell zu SCOPe gehört. In ihm werden - basierend auf Sequenzen - Proteine mit großer Ähnlichkeit aufbewahrt, die nur geringe bzw. keine Homologie aufweisen und große strukturelle Ähnlichkeit besitzen.

SISYPHUS ist ein Datensatz der ebenfalls mit SCOPe assoziiert ist. Hierin befinden sich Proteine mit *nicht-trivialen Beziehungen* , die zusammen gruppiert werden. Hierzu gehören beispielsweise Proteine, die sich durch zyklische Vertauschungen unterschieden, oder sogenannte Chamäleon-Sekundärstrukturen aufweisen, die sich je nach Umgebung ändern. Der Datensatz ist unter der URL: http://www.spice-3d.org/sisyphus/index.jsp zu erreichen. Leider wird er seit etwa 2009 nicht mehr aktualisiert.

2.4 Proteinwahl anhand von EC-Termen

Es könnte eine gute Idee sein, anhand von EC-Termen Proteine auszuwählen. Die Methode kann getestet werden, indem man zwei oder mehrere verschiedene Gruppen von Proteinen wählt, die jeweils mit dem selben EC-Term assoziiert sind, dabei aber geringe Homolgien aufweisen. Wenn die Proteine, die mit den selben EC-Termen assoziiert sind, im selben Cluster landen, ist die Methode sinnvoll. Dies folgt dem Dogma, das die Funktion eines Proteins aus der Struktur folgt.

CATH ermöglicht die Suche nach EC-Termen. Nach den ersten Tests könnte es sinnvoll sein, automatisiert zu suchen. Bisher stellt CATH seine API jedoch noch nicht zur Verfügung. Vielleicht ist ein Umweg über die PDB möglich.

3 Methoden

Der Artikel *Advances and pitfalls of protein structure alignment* (URL: http://www.sciencedirect.com/science/article/pii/S0959440X09000621) scheint

ein guter Punkt für den Anfang zu sein. In ihm werden verschiedene Strukturalignment-Methoden verglichen und bewertet! GEILOMAT!!!!

Den Algorithmus DALI könnte man sich genauer anschauen. Vielleicht taugt er zum Vergleich mit dem graphletAnalyser

3.1 Datengewinnung aus der PTGL

Die Daten aus der PTGL werden mit bash Skripten unter Verwendung der REST API (URL: `http://ptgl.uni-frankfurt.de/api/`) gewonnen

3.2 Pipeline

Schön wäre es, eine Pipeline zu haben, die die Suche nach *Graphlets* automatisiert. Der Workflow könnte folgendermaßen aussehen.

1. Die Pipeline erhält EC-Terme und eine Zahl x, die angibt, wie viele Proteine fr jeden EC-Term gewählt werden sollen. Dann durchsucht sie eine Datenbank. Dies wird wahrscheinlich GO werden, da CATH noch keine API zur Verfügung stellt. Von dort werden die ersten x PDB-IDs extrahiert.

2. Mit Hilfe der REST API werden aus der PTGL die Proteingraphen extrahiert, deren PDB-IDs zuvor gesammelt wurden.

3. Die Proteingraphen werden dem `graphletAnalyser` übergeben. Dieser berechnet die *Graphlets* und clustert sie anhand einer Metrik, die noch bestimmt werden muss.