Technische Hochschule Bingen
Fachbereich 2 – Technik, Informatik und Wirtschaft
Angewandte Bioinformatik (B. Sc.)

prot-fin, ein toller Titel

Bachelorarbeit
abgegeben am: 26.08.2024
von: Franz-Eric Sill

Dozent: Prof. Dr. Asis Hallab

Zusammenfassung

...

Abstract

...

Literatur

[Kid+85] Akinori Kidera u. a. “Statistical analysis of the physical properties of the 20 naturally occurring amino acids”. In: *Journal of Protein Chemistry* 4.1 (Feb. 1985), S. 23–55. ISSN: 1573-4943. DOI: 10.1007/BF01025492. URL: https://doi.org/10.1007/BF01025492.

Abbildungsverzeichnis

Tabellenverzeichnis

Inhaltsverzeichnis

1 Einleitung

...

2 Material und Methoden

2.1 Grundalgorithmus

Vorbereitung: Voraussetzung für den Algorithmus ist ein numerischer Vektor, so wie es das Spektrum einer Tonspur bei SHAZAM darstellt. Um dies im proteinischen Kontext zu erreichen, wird in prot-fin auf sogenannte Kidera-Faktoren zurückgegriffen. Diese Faktoren stammen aus einem Forschungsprojekt von Akinori Kidera, welches 1985 publiziert wurde. Inhalt des Projekts war die statistische Faktorenanalyse von 188 physikalischen Eigenschaften der 20 natürlichen Aminosäuren zur Ermittlung von 10 dieser Eigenschaften, durch die die anderen aufgrund hoher Korrelation erklärt werden können [vgl. Kid+85]. In Tabelle 1 sind diese dargestellt. Folglich kann eine Aminosäuresequenz pro Faktor in einen numerischen Vektor übersetzt werden, wobei ein höherer absoluter Wert für mehr Relevanz des Faktors steht.

Tabelle 1: Kidera-Faktoren

Beschreibung	**A**	**C**	**D**	**E**	**F**	**G**	**...**
Helix/bend preference	-1.56	0.12	0.58	-1.45	-0.21	1.46	...
Side-chain size	-1.67	-0.89	-0.22	0.19	0.98	-1.96	...
Extended structure preference	-0.97	0.45	-1.58	-1.61	-0.36	-0.23	...
Hydrophobicity	-0.27	-1.05	0.81	1.17	-1.43	-0.16	...
Double-bend preference	-0.93	-0.71	-0.92	-1.31	0.22	0.1	...
Partial specific volume	-0.78	2.41	0.15	0.4	-0.81	-0.11	...
Flat extended preference	-0.2	1.52	-1.52	0.04	0.67	1.32	...
Occurrence in alpha region	-0.08	-0.69	0.47	0.38	1.1	2.36	...
pK-C	0.21	1.13	0.76	-0.35	1.71	-1.66	...
Surrounding hydrophobicity	-0.48	1.1	0.7	-0.12	-0.44	0.46	...

Sammeln von Strukturdaten: Das Extrahieren von struktureller Information aus den erhaltenen Vektoren basiert auf der Short-Time-Fourier-Transformation (STFT), welche den Vektor intervallweise auf periodische Signale untersucht, wie z.B. dem wiederholten Auftreten von hydrophoben Aminosäuren im gleichen Abstand oder in der Musik ein Refrain oder dem Rhythmus. Da für eine STFT negative Werte kritisch sind, sind die Vektoren so normalisiert, dass das um 1 inkrementierte absolute Minimum der Tabelle 1 auf jeden Wert addiert wird. Die Frequenzen der auffälligsten Signale werden ausgewählt, also den lokalen Maxima, sodass über alle Intervalle eine sogenannte Constellation-Map entsteht.

Hashing: Die erhaltene Map wird nun elementweise gehashed, um einen effizienten Vergleich mit anderen Maps zu ermöglichen. Um das zu erzielen wird jede ausgewählte Frequenz mit jeder weiteren Frequenz der Folgeintervalle gepaart. Es werden also Kanten gebildet, wodurch die Map zu einem Graphen wird. Jede dieser Kanten bildet nun

einen Hash, also einer Kombination aus den beiden Frequenzen/Kantenenden und der Kantenlänge. In einer Hashmap, also der Datenbank, wird sich folgend für den Hash die Position der Kante in der Constellation-Map gemerkt. Sollte ein Hash mehrfach vorkommen, so gilt dies nur für die letzte Position.

Dieses Verfahren wird in Abbildung 1 repräsentativ dargestellt, wobei rote Kanten die ignorierten Kanten abbilden.

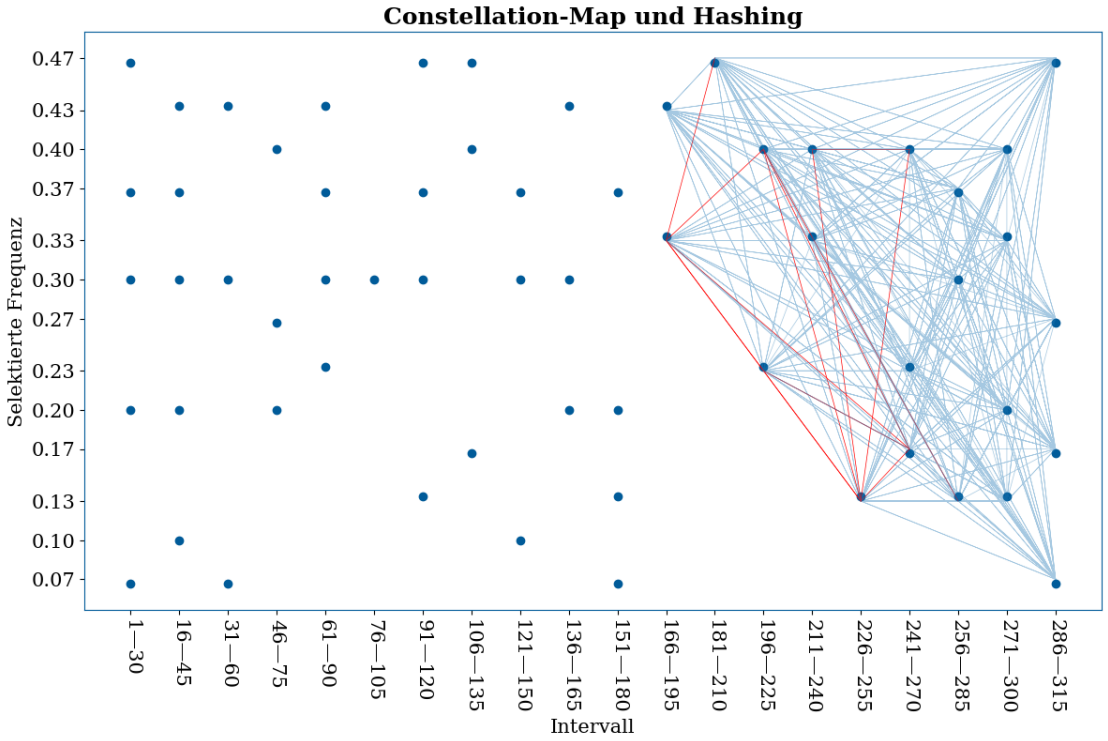

Abbildung 1: Constellation-Map und Hashing

Single-Protein-Matching: Nachdem eine Datenbank mit den Hashes verschiedener Trainings-Proteine (TP) trainiert wurde, wird sie verwendet, um Proteine zu erkennen. Dazu werden für die jeweilige Eingabesequenz von Aminosäuren die Hashes gebildet und deren Positionen gespeichert. Um nun die Ähnlichkeit der Constellation-Map der Eingabe mit denen der TP zu bestimmen, werden pro Eingabe-Hash die Differenzen zwischen dessen Position mit den Positionen der trainierten Hashes gebildet und global pro Protein gezählt. Diese Differenzen repräsentieren den Abstand der Kante in der Eingabe-Map zur Kante der jeweiligen TP-Map, also wie weit die Eingabe-Map verschoben wäre, sollte es sich bei dem TP um das Original handeln. Auf diese Weise sammeln sich pro TP mehrere solcher potentiellen Abstände, wobei nun der Abstand, der am häufigsten aufgetreten ist, offensichtlich die meiste Übereinstimmung in den Kanten zeigt. Diese Tatsache qualifiziert diese Maximalanzahl als geeigneten Score (S1) für ein Match.

Da es große Proteine mit sehr langen Aminosäuresequenzen kürzere Sequenzen kleinerer funktionsungleicher Proteine enthalten können, reicht der ermittelte Score alleine nicht aus, da in diesem Fall sehr viele Kanten der Eingabe-Map übereinstimmen würden, sodass trotz Mis-Match der nahezu maximale Score erreicht werden würde.

Um das zu umgehen, wird der Jaccard-Similarity-Index (JSI) verwendet, einem Maß, das die Übereinstimmung zweier Mengen A und B wie folgt bewertet:

$$JSI(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

Dieser Index nimmt einen Wert von 0 an, wenn beide Mengen disjunkt sind, und nähert sich der 1 je größer die Schnittmenge ist. Im Fall des Vergleichs zweier Constellation-Maps, also zwei Hash-Mengen, wird hier bewertet, wie viele Kanten sich die beiden Maps positionsunabhängig teilen. Durch diese Unabhängigkeit reicht der JSI alleine nicht als Score aus, sodass nur in Kombination/Multiplikation mit dem S1 ein robuster Score entsteht, da beide zusammen ihre Schwächen aufheben.

Family-Matching: Wenn prot-fin wie erwartet funktioniert, sollten beim Single-Protein-Matching die Matches mit den besten Scores funktionsähnliche Proteine sein. Ein weiterer Ansatz, solche Verwandten zu ermitteln, ist das Matching mit Proteinfamilien. Hierbei wird als Eingabe eine Tabelle akzeptiert, die lediglich eine Zuordnung von Protein-ID und Familie enthält. Um nun Matches zu finden, werden von allen Hashes einer Familie nur die behalten, die in allen Mitgliedern vorkommen. Anschließend wird die Datenbank nach Proteinen durchsucht, welche ebenfalls diese Hashes enthalten, wobei als Score diesmal nur die Anzahl infrage kommt, wie viele der Hashes enthalten sind. Die Idee hinter dieser Methode ist, dass Proteine derselben Proteinfamilie, also mit ähnlichen Funktionen, möglicherweise familienspezifische Kanten in der Constellation-Map haben.

2.2 Experiment 1: UniRef90 Sampling

Ein wichtiger Bestandteil des Algorithmus ist die Selektion signifikanter Frequenzen zur Erstellung der Constellation-Map. Es wäre möglich, einfach alle Frequenzen auszuwählen und die Signalstärke in den Hash einfließen zu lassen. Problem hierbei ist aber, dass diese Vorgehensweise zu wesentlich mehr Hashes und einer folglich sehr großen Datenbank führt, was wiederum das Scoring/Matching verlangsamt. Ein Anspruch an prot-fin ist, dass die Datenbankgröße die Eingabegröße nicht wesentlich übersteigt, wobei es sich bei der Eingabe um eine einfache FASTA-Datei handelt.

Diesem Problem soll durch ein Sampling-Experiment abgeholfen werden. Darin werden aus etwa 180 Millionen Sequenzen je ein zufälliges Intervall für die STFT ausgewählt, transformiert und die Signalstärken je Frequenz gemerkt. Um nun daraus eine Selektionsmethode abzuleiten, werden die Grenzquantile einer jeden Frequenz ermittelt, um signifikant seltene Signalstärken zu ermitteln. Folglich ist es möglich, für die Constellation-Map nur diejenigen Frequenzen zu behalten, welche in den Randzonen der Signalstärken liegen, sodass nicht nur Signale infrage kommen, die für eine besonders starke Ausprägung

eines Kidera-Faktors sprechen, sondern auch für den Fall der umgekehrten Ausprägung, wie z.B. Hydrophilie statt Hydrophobie.

Der Algorithmus wird daher insofern angepasst, dass bei der Frequenz-Selektion von den Maxima der Signalstärken nur die behalten werden, die die Grenzwerte über-/unterschreiten. Zudem wird beim Hashing je Frequenz noch die Information hinzugefügt, ob sie besonders stark oder schwach ist.

3 Ergebnisse

...

4 Diskussion

...

Eigenständigkeitserklärung

Ich bestätige, dass die eingereichte Arbeit eine Originalarbeit ist und von mir ohne weitere Hilfe verfasst wurde. Die Arbeit wurde nicht geprüft, noch wurde sie widerrechtlich veröffentlicht. Die eingereichte elektronische Version ist die einzige eingereichte Version.

Unterschrift	Ort und Datum

Erklärung zu Eigentum und Urheberrecht

Ich erkläre hiermit mein Einverständnis, dass die Technische Hochschule Bingen diese Arbeit Studierenden und interessierten Dritten zur Einsichtnahme zur Verfügung stellen und unter Nennung meines Namens (Franz-Eric Sill) veröffentlichen darf.

Unterschrift	Ort und Datum