TH BINGEN
University of Applied Sciences

prot-fin
Das SHAZAM für Proteinerkennung?

Praxisphase von Franz-Eric Sill
Durchgeführt bei: Prof. Asis Hallab, TH Bingen

Einleitung

Als SHAZAM ist eine Anwendung bekannt, die Musiktitel in Sekundenschnelle anhand von ebenso kurzen Tonaufnahmen erkennt. Der Kern des Algorithmus gleicht hierbei die Struktur der Aufnahme mit einer Datenbank ab, die mit Millionen von Songs gefüttert wurde und liefert den besten Treffer als Ergebnis [Wan03].
Doch wäre das auch für Proteine möglich?
Aktuell sind Alignments von den Aminosäureketten eine sehr populäre und effiziente Methode, um Proteine und deren funktionsgleiche Verwandte über Sequenzähnlichkeit zu erkennen. Bekannte Tools hierfür sind bspw. BLAST und DIAMOND. Das Problem hierbei ist, dass mit zunehmender evolutionärer Distanz zwischen Sequenzen gefundene Homologien nicht mehr signifikant von zufälliger Ähnlichkeit zu unterscheiden sind. Es gilt das Basiskonzept von Struktur und Funktion, dennoch wird hier keine Ähnlichkeit auf Basis der Struktur festgestellt, da die letztendliche Tertiärstruktur schwer vorhersehbar ist und lediglich die aufgetretenen Aminosäuren verglichen werden. Wenn diese Vorgehensweise also nicht immer ideal ist und SHAZAM Musik auf Basis struktureller Information erkennt, wäre es doch vielleicht möglich, dass der zugrundeliegende Algorithmus in den physikalischen Eigenschaften der Proteinsequenzen strukturelle Information findet, die spezifisch für das Protein und vielleicht auch seine Verwandten ist.

Methode

Vorbereitung: Voraussetzung für den Algorithmus ist ein numerischer Vektor, so wie es die digitale 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 [Kid+85]. In diesem Projekt wurden für Aminosäuren 10 Faktoren ermittelt, mit denen die physikalischen Eigenschaften eines Proteins am meisten korrelieren, sodass nun eine Aminosäuresequenz pro Faktor in einen numerischen Vektor übersetzt werden, wobei ein höherer absoluter Wert für mehr Relevanz des jeweiligen Faktors steht.

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. Die Frequenzen der auffälligsten Signale werden ausgewählt (Fig. 1), sodass über alle Intervalle eine sogenannte Constellation-Map entsteht, wobei das Reziproke einer Frequenz die Länge einer Periode in Aminosäuren angibt.

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 (Fig. 1).

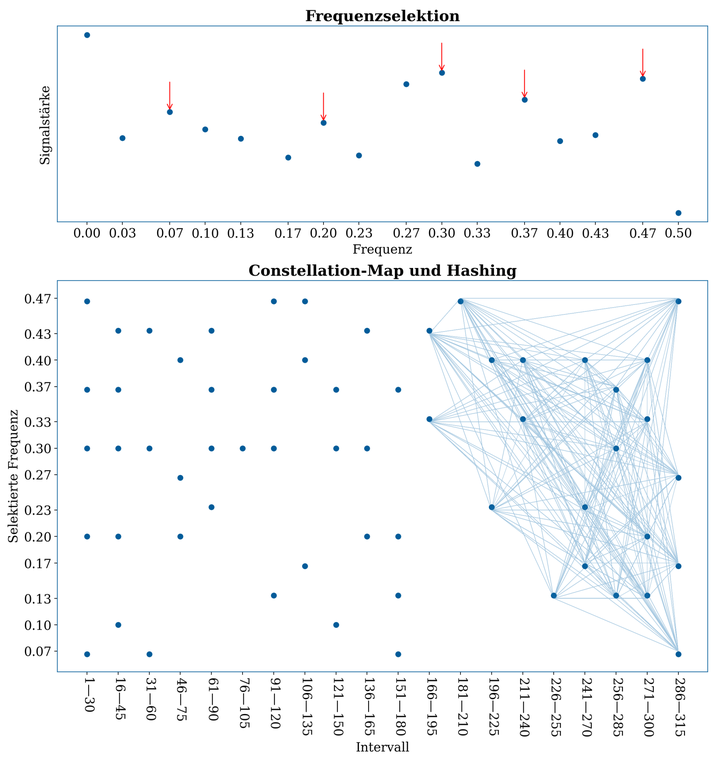

Fig. 1: Frequenzselektion und Hashing

Matching

Nachdem eine Datenbank mit den Hashes verschiedener Trainings-Proteine (TP) trainiert wurde, wird sie verwendet, um Proteine zu erkennen. Dazu werden die Constellation-Map einer Eingabesequenz mit denen der TP verglichen. Jedes TP erhält einen Score (S1), der repräsentiert, wie viele seiner Hashes in Position mit der Eingabe übereinstimmen.
Da sehr große Proteine potenziell kleine Proteine mit ihren Constellation-Maps als Subsequenz enthalten können, wird der Jaccard-Similarity-Index (JSI) aufmultipliziert, einem Maß, das die Übereinstimmung zweier Hash-Mengen A und B positionsunabhängig mit $\frac{|A \cap B|}{|A \cup B|}$ errechnet. Der Index nimmt somit einen Wert von 0 an, wenn beide Mengen disjunkt sind, und nähert sich der 1 je größer die Schnittmenge ist, was somit die Schwäche des S1 ausbügelt.

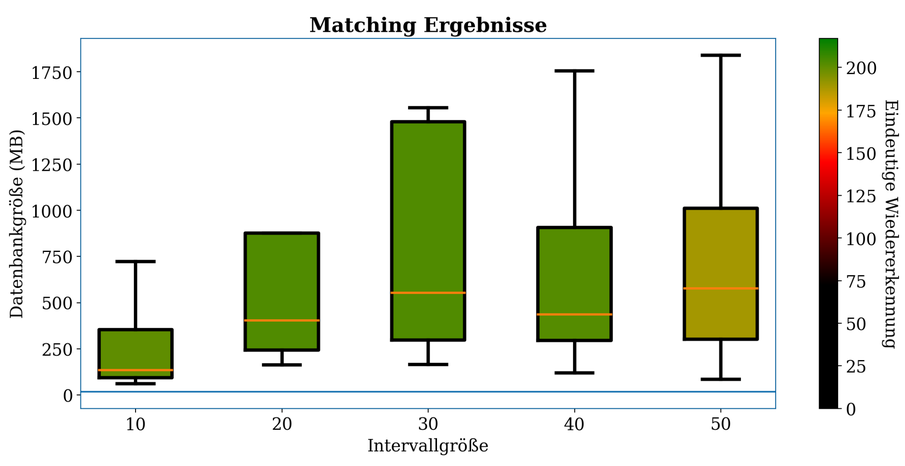

Fig. 2: Erzielte Ergebnisse

Performanz

Um den Algorithmus von Musik auf Proteine abzustimmen, wurden bei der Erstellung der Constellation-Map verschiedene Parameter für die STFT durchprobiert, nämlich die Intervallgröße, der Abstand, wie weit das Intervall weitergeschoben wird, und die Anzahl an maximal selektierter Frequenzen. Als Trainingsdaten wurden circa 40.000 Pflanzenproteine verwendet und nachfolgend 217 vollständige Sequenzen möglichst funktionsverschiedener davon als Eingabe für das Matching, wobei lediglich die Hydrophobizität der Kidera Faktoren betrachtet wurde.
In Fig. 2 sind die Matching-Ergebnisse abgebildet, wobei die Boxen, gruppiert nach Intervallgröße, die Datenbankgrößen für die verschiedenen Parameterkonfigurationen wiedergeben, inklusive Median. Die Färbung einer Box stellt dar, wie viele der Eingabeproteine im Mittel eindeutig erkannt wurden. Die Wiedererkennungsrate ist ziemlich hoch, aber die Datenbankgrößen ebenfalls. Verglichen mit der Eingabegröße (blaue horizontale Linie in Fig. 2) ist diese für nur einen betrachteten Kidera Faktor um ein Vielfaches größer. Das Matching ist dementsprechend langsam, weshalb es für Intervall 50 vorzeitig abgebrochen wurde.
In zukünftigen Experimenten muss eine drastische Reduktion der Speicherkomplexität erzielt werden, um die Anwendung auf alle Kidera Faktoren zu ermöglichen. Ist das erreicht, kann in prot-fin anstelle der Wiedererkennung einzelner Proteine die Identifikation von Gruppen funktionsähnlicher Proteine angegangen werden. Nach der bisherigen Entwicklung des Projekts und den vorläufigen Ergebnissen besteht dafür auf jeden Fall ein gewisses Potential, sodass vielleicht eine neue Alternative zu Sequenzalignments entsteht.

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.

[Wan03] Avery Wang. "An Industrial Strength Audio Search Algorithm." In: Jan. 2003.

Studiengang Angewandte Bioinformatik