PROJEKT 3: PAARWEISE SEQUENZALIGNMENTS

Name: Dang Quynh Tram Nguyen
Matrikelnummer: 5311561
Studiengang: Bioinformatik

Inhaltsverzeichnis

1. Einleitung

In der Zytologie beschäftigt man viel mit den Sequenzen aus mehreren Genomen. Man beobachtet nicht nur ihren Stamm, ihr Verhalten und ihre Funktion, sondern auch die Beziehungen zwischen ihnen. Dabei ist das Alignment dieser Sequenzen benötigt, um ihre Ähnlichkeit in den Strukturen sowie im Allgemein zu suchen.

In diesem Projekt geht es um die Sequenzaligments. Zuerst werden zwei empfohlenen Algorithmen Needleman-Wunsch und Smith-Waterman dafür beschreibt. Danach kommen die Beispiele, damit die Funktion der Algorithmen genau erklärt und verglichen werden. Zunächst wird die Ähnlichkeit im Alignment besprochen, ob sie eine biologische Bedeutung aufweist. Am Ende werden die Parameter, die an die Algorithmen teilnehmen, analysiert, ob ihre Veränderung auf das Ergebnis bewirken.

2. Algorithmen

Um die Ähnlichkeit sowie die biologische Relevanz der Sequenzen zu beobachten, werden die Sequenzen durch die Algorithmen ausgerichtet. Die Algorithmen helfen, das beste Alignment der Sequenzen zu finden, dabei die höchste Sequenzidentität erreicht ist.

Zuerst wird je Element der Sequenzen miteinander verglichen. Diesen Schritt kann man unter einer Matrix beschreiben. Die Anzahl der Spalten und Zeilen der Matrix entsprechen jeweils die Länge beider Sequenzen. Jeder Index, wobei zwei verglichen Elemente identisch sind, bekommt einen Punkt (Dot). Dieser Verlauf ist Dot-Plot-Algorithmus benannt. Aber der Überfluss der möglichen Alignments als Ergebnisse verwirren bei der Beobachtung. Deswegen sind die entwickelten Algorithmen, die genau das beste Alignment ergeben, benötigt. Zwei typischen darin sind Needleman-Wusch- und Smith-Waterman-Algorithmus. Sie berechnen die Scores der Punkte und dadurch kann man das beste Alignment suchen.

1. Needleman-Wunsch-Algorithmus

Das Needleman-Wunsch-Algorithmus ist der Algorithmus mit Score. Wie oben beschrieben, wird das Score in jedem Index der Matrix berechnet. Dafür sind die Werte von Gap-Penalty, Match und Missmatch notwendig, mit den das Score erhöht oder vermindert wird. Zuerst wird ein Gap (Abk. „-“) vor jeder Sequenz addiert, damit die Matrix eine Spalte und eine Reihe mehr enthält. Diese Spalte und Reihe werden dann initialisiert. Der Algorithmus läuft, Das Score eines Index ist das Maximum der temporalen Scores von oben, links und diagonal vom Index. Die temporalen Scores von oben und links sind jeweils die Subtraktion des Gap-Penalty-Wertes vom Score an dieser Stelle. Der temporale Score von diagonal ist die Addition des Match- oder Missmatch-Wertes und der Score an dieser Stelle. Beim Berechnen wird die Richtung des Maximums auch gespeichert, damit man später das Alignment in der Matrix finden kann. Außerdem wird der finale Score in der Score-Matrix auch gespeichert. Der Algorithmus wird klar bei der Implementierung im Quellecode beschreibt.

Manche zusätzlichen Funktionen werden auch implementieren, um das Berechnen der Scores und die Rückgabe der Vorgänger zu helfen.

2. Smith-Waterman-Algorithmus

Ähnlich wie den Verlauf des Needleman-Algorithmus aber mit schöner Matrix ist der Smith-Waterman-Algorithmus. Der Score jedes Indexes ist das Maximum von 0 und der temporalen Scores von links, oben und diagonal. Der Algorithmus wird genauer im Code des Anhangs beschreibt.

In diesem Algorithmus wird der beste Score gespeichert, statt der finale Score wie im Needleman-Wunsch. Dies ist als Vorteil beim lokalen Alignment. Aus dem Index des besten Scores kann man die identischen Teilbereiche der Sequenzen finden. Diese können die ähnlichen Gene beider Sequenzen

sein bzw. weisen die enge Verwandtschaft der Sequenzen auf. Je mehr identische Teilsequenzen gefunden sind, desto enger Verwandtschaft der Sequenzen beweist es.

3. Testen mit dem Onkogen

Proto-Onkogene sind die Gene, die das normale Zellwachstum, die Zellteilung und Entwicklung verschiedener Gewebe steuern. Wenn diese Gene mutiert werden, nennt man die Onkogene. Diese fördern die Zellteilung und da folgt den Überschuss der Produkte im Körper. Es verursacht oft den Krebs.

Bei diesem Test werden die Onkogene KRAS des Menschen (NM_004985.3) und der Maus (NM_021284.7) ausgenutzt. Die Sequenzen werden erst gelesen und unter String gespeichert. Dann werden sie mit beiden oberen Algorithmen ausgerichtet, dabei eine Substitutionsmatrix für Nukleotiden benötigt ist. Diese Matrix ist unter einem Dictionary gespeichert. Darin ist jeder Schlüssel ein String von zwei verglichenen Elementen der Sequenzen und sein Wert ist ein Score des Vergleichs. In dieser Matrix bekommt je Match 1 und je Missmatch -1.

Nach den Verläufen beider Algorithmen ergibt es zwei besten Alignements. Dann wird die Sequenzidentität jedes gefundenen Alignments bestimmt. Jede Stelle im Alignment wird nun verglichen und dadurch wird der Anteil die Summe der Übereinstimmung im ganzen Alignment berechnet.

Um zwei Algorithmen zu vergleichen, vergleicht man anhand der besten Alignment von zwei Algorithmen manche generelle Werte wie des finalen Scores, die Länge der eigenen Sequenz nach dem Alignment sowie die Sequenzidentität.

Die Anteile der Übereinstimmung in beider Alignments sind ganz ähnlich (68,74 % ~ 68,72%). Die Länge der Alignments sind genau wie einander (5467 Basenpaare). Und dieselben ausgerichteten Sequenzen im Vergleich zweier Alignments haben auch gar keinen großen Unterschied. Die Scores sind aber anders, weil jeder Algorithmus anderer Score ergibt. Wie oben beschreibt, speichert der Needleman-Wunsch Algorithmus der finale Score, während der Smith-Waterman das beste Score behält. Zusammenfassend sind das beste Alignment zweier Sequenzen trotz Anwenden unterschiedlicher Algorithmen gar nicht ändert.

4. Signifikanz von Ähnlichkeit

Bei der Studie möchte man nicht nur die besten Alignments der Sequenzen wissen, sondern auch ihre biologische Bedeutung entdecken, ob zwei Sequenzen sowie die Genome verwandt sind oder die Ähnlichkeit einfach einen Zufall in der Natur gekommen ist. Dabei gelten die Alignments der Aminosäuresequenz TRA2 von *C.elegans* (Fadenwurm) in dem Protein des menschlichen PTCH2 Gene und in dem Protein eyeless von *D.melanogaster* (Fruchtfliege) als Antwort der Signifikanz der Ähnlichkeit im Alignment.

TRA-2 ist ein geschlechtsbestimmendes Gen des Fadenwurmes. Seine Expression ist deutlicher bei Hermaphroditen als bei Männern. Diese Hermaphroditen neigen später dazu, Weibchen zu werden. Das Protein eyeless von *D.melanogaster* spielt eine Rolle bei der Entwicklung des Nervensystems. Es exprimiert viele Strukturen, die das zentrale Nervensystem bilden. Nervensystem kann die sexuelle Entwicklung kontrollieren.

Um das Alignment anzufangen, ist die Substitutionsmatrix BLOSUM62 benötigt, dabei es erwünscht ist, höchstens 62% der Sequenzidentität zu erreichen. Danach wird das beste Alignment mittels zwei Algorithmen mit Gap-Penalty von -5 gefunden.

Das Ziel ist, dass den biologischen Hintergrund der Alignments des Gens TRA2 mit dem Protein PTCH2 und Eyeless zu beobachten. Dabei versucht man ca. 500 Zufallalignments jedes Algorithmus, in denen das Gen TRA2 behält ist und die Aminosäuresequenzen in PTCH2- und Eyeless-Protein permutiert werden. Nach dem Versuch erhält man die Ergebnisse unter Histogrammen und Währscheinlichkeit der Alignments, die gleich oder besser als das Original sind.

Bei der Zufallalignments mittels Needleman-Wunsch-Algorithmus ist die Wahrscheinlichkeit der guten bis besseren Scores (≥ dem Score des Alignment mit den originalen Sequenzen) sehr klein im Vergleich zu den schlechteren Scores. D.h. das beste Alignment von TRA2 und PTCH2 sowie Eyeless einzigartig bzw. nicht zufällig gekommen ist. Dies beweist auf eine Verwandtschaft zwischen die Genome unterschiedlicher Lebewesen.

Bei der Zufallalignments mittels Smith-Waterman-Algorithmus ergibt die berechnete Wahrscheinlichkeit von 1. Wegen des Berechnens mit dem besten Score, statt finalen Score wie bei Needleman-Wunsch, sind die gefundenen Scores der zufälligen Alignments deutlich besser als das mit den originalen Sequenzen. Aber ist die biologische Bedeutung des besten Alignments nicht verloren. Die besten Scores reflektieren in diesem Fall nur die besten lokalen Alignments, nicht die Alignment der ganzen Sequenzen.

Zusammenfassend weist die Ähnlichkeit der jeweiligen Sequenzen eine Beziehung der Genome der Organismen auf und dient als einen Vorteil bei Beobachtung, Entdeckung und Entwicklung in der Zytologie. In diesem Beispiel wurde das Gen TRA-2 auf dem menschlichen PTCH2 Gen und auf dem Eyeless-Gen mit hoher Identität gefunden. D.h. die Proteine aus dieser Gene vermutlich ähnliche Funktion exprimieren, die auf Geschlecht der somalischen Zellen betrifft.

5. Einfluss von Alignmentparametern

Um die Scorematrix bei den Algorithmen zu erfüllen, spielen die Werte des Matches, Missmatch und Gap-Penalty wichtige Rolle. Dies werden Alignmentparameter benannt. Eine Frage wird gestellt, ob die unterschiedliche Alignmentparameter zu den verschiedenen Ergebnissen führt. Um das Problem zu erklären, wurden 500 Alignments mit zwei Sequenzen aus dem File ‚rRNASeq.fasta' und zufälligen Parametern mittels des Smith-Waterman-Algorithmus durchgeführt. Inzwischen verändern die unterschiedlichen Matches und Missmatch die Substitutionsmatrix in jedem Alignment.

Am Ende werden alle Länge der Alignment sowie die Sequenzidentität im Alignment in zwei Liste gespeichert und unten Histogramme dargestellt. Im Hinblick fallen zwei Säulen, die viel höher als die anderen sind, aus den Graphiken auf. Das bedeutet, die Werte (Alignmentslänge von 72 bp und Sequenzidentität von 32,5%) meisten in Alignments gekommen sind bzw. die Ergebnisse sind, die am bestens das beste Alignment mit diesem Algorithmus aufweisen.

Missmatch und Gap-Penalty kontrollieren die Sequenzidentität des Alignment. Wenn diese (negativen) Werte sehr klein ist, bekommt das Score den Wert von 0, dabei das kontinuierliche Alignment mit Smith-Waterman-Algorithmus abgebrochen wird. Da folgt am Ende mit dem besten Score nur ein lokales Alignment. Umgekehrt halten der große Match-Wert die Scores positiv, damit das Alignment am Ende auf den ganzen Sequenzen ist und manche Nicht-Übereinstimmungen vernachlässig werden können.

6. Literatur

- C. Meckbach, Projekt 3: Paarweise Sequenzalignments, 2022.
- Berufsverband Deutscher Internistinnen und Internisten e.V., *Begriff: Proto-Onkogene und Onkogene*, Begriff » Glossar » Internisten im Netz » (internisten-im-netz.de)

- P. G. Okkema, J. Kimble, *Molekulare Analyse von Tra-2, einem geschlechtsbestimmenden Gen in C.elegans.*, Molekulare Analyse von Tra-2, einem geschlechtsbestimmenden Gen in C.elegans. - PMC (nih.gov)
- National Library of Medicine, *ey eyeless [Drosophila melanogaster (Fruchtfliege)]*, ey eyeless [Drosophila melanogaster (Fruchtfliege)] - Gen - NCBI (nih.gov)