Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: 7.3.2018
Manuscript Category

Aldabi

Projekt 3 Dokumentation

Lucas Rieckert, Maximilian Otto, Johanna Eitel

Projektbeginn: 24.1.2018; Projektabgabe: 7.3.2018, 23:59 Uhr

Abstract

Motivation: Unsere Motivation hinter dem Projekt, liegt darin aus homologen Gensequenzabschniten eines Virusgenoms einen pylogenetischen Baum zu erstellen um ihre Verwandschaftsverhaeltnisse und somit Verbreitungswege darzustellen.
Results: Als Ergebnis erhalten man zwei phylogenetische Bäume, welche auf der Basis von zwei unterschiedlichen Distanzmethoden erstellt wurden.
Contact: lackyluck@hotmail.de maxotto45@gmail.comeitel.johanna@yahoo.de

1 Introduction

Unsere Aufgabe für das dritte Praktikum bestand darin, aus homologen genetischen Sequenzen eines Virusgenoms einen phylogenetischen Baum zu erstellen, der die Verwandschaftsverhaeltnisse dieser Sequenzen wiederspiegelt. Zunächst werden die Distanzen der Gensequenzen berechnet, um dann den phylogenetischen Baum mittels der "unweighted pair group method using arithmetic averages", kurz: UPGMA, zu ermitteln. Mittels UPGMA erstellten wir einen phylogenetischen Baum im NEXUS-Format und hatten div. Programme zur freien VerfÃ¼gung, dieses Dateiformat grafisch ausgeben zu lassen. Die von uns verwendete Programmiersprache ist Java, da wir damit alle am meisten Erfahrung haben. Das Darstellungsprogramm des Baumes ist "FigTree".

2 Vorgehen

Zunächst benoetigten wir geeignete genetische Sequenzen, die wir uns aus einer Datenbank herunterladen sollten. Dazu hatten wir mehrere Datenbanken mit verschiedenen Viren zur Auswahl. Wir entschieden uns fuer den SARS Coronavirus und haben uns aus der NCBI-Datenbank 19 DNA-Sequenzen des SARS Coronavirus entnommen. Genauer entschieden wir uns fuer die Sequenzabschnitte, die das Protein "orf1ab" codierten, da dieses in jedem SARS Coronavirus vorhanden ist und es dazu sehr viele, gut dokumentierte Daten gibt. Das Produkt dieses Genabschnittes "orf1ab" ist ein Polyprotein, zusammengestzt aus zweien, welche eine Replikase bilden. Wir haben zunächst versucht direkt mit den Aminosäuresequenzen zu arbeiten, allerdings gab es bei diesen für unseren Geschmack zu wenige Unterschiede, sodass der Baum die Sequenzen, die eigentlich alle gleichweit voneinander entfernt sind, erst nacheinander geclustert hätte und das den Eindruck erweckt hätte, die Sequenzen würden weiter von einander entfernt liegen als sie es in Wirklichkeit tun. Allerdings waren auch die Neukleotidsequenzen sehr ähnlich. Mit weitaus grösseren Datensätzen sähe das Resultat vermutlich ergiebiger aus.

Die von uns gesammelten Gensequenzen für das Protein orf1ab führten wir dann in einer langen FASTA-Datei zusammen. Diese FASTA-Datei liegt in der Repository unter dem Namen "sequences.fasta" im Ordner "Sequenzdateien" im master-branch vor. Diese Datei übergaben wir dann dem Multiplen-Sequenz-Alignment-Online-Tool Muscle um uns aus den Sequenzen ein Multiples-Sequenz-Alignment berechnen zu lassen. Die von Muscle ausgegebene Alignmentdatei liegt im Repository unter dem Namen "aln-fasta.txt" vor. Dieses iterative Programm erstellt ein Alignment fÃ¼r UPGMA-Baueme, und errechnet daraus auch eine Distanzmatrix nach dem Kimura-Modell und versucht durch Profile, welche wiederholt zum MSA aligniert werden, bessere Distanzen zu erreichen, solange bessere Scores herauskommen. Das Alignment ist in Abbildung 1 zum Teil dargestellt. Diese Datei haben wir dann von unserem Programm mittels eines File-Readers einlesen lassen.

Um aus diesen alignierten Sequenzdatein mittels UPGMA einen phylogenetischen Baum erstellen zu koennen, mussten wir zunaechst Funktionen implementieren, welche die Distanz zwischen zwei alignierten Sequenzen berechnen. Dabei entschieden wir uns zunaechst fuer die in der Vorlesung behandelte P-Distanz, da diese recht simpel zu implementieren war und das direkte Verhaeltnis der Unterschiede der Sequenzen zu der Laenge der Sequenzen beschreibt, indem sie die Anzahl der sich unterscheidenden Stellen durch die Laenge der Sequenzen teilt. Als zweite Methode zur Distanzberechnung entschieden wir uns fuer die Methode von Jukes-Cantor, da diese zusaetzlich zur p-Distanz ebenfalls die grobe Mutationswahrscheinlichkeit einer Base als Faktor in die Berechnung mit einbezieht. So haben wir zwei unterschiedliche Verfahren zur Berechnung von paarweisen Distanzen von Nukleotidsequenzen, aus denen zwei unterschiedliche Bäume entstehen. Die aus den Berechnungen hervorgegangenen Distanzmatrixen sind in der Abbildung 2 und 3 dargestellt.

Mittels dieser Methoden zur Distanzberechnung haben wir dann den UPGMA-Algorithmus implementiert. Je nach Übergabeparameter (1 oder

2) wurde entweder die p-Distanz oder die Jukes-Cantor-Distanz benutzt. Wir führten das Programm mit jedem der Parameter aus um beide Bäume zu generieren. Die Berechnung der Distanzmatrizen ist in der Funktion distance bzw. distance2 implementiert. Aus diesen haben wir dann den kleinsten Wert herausgesucht und die dazugehörigen Knoten bzw. Cluster zu einem Cluster zusammengefügt, indem wir die kombinierten Metadaten der Ursprungsknoten/-cluster - gespeichert in der Array-List 'header' - als neuen Eintrag an dieses anfügen. Danach muss auch die Distanzmatrix angepasst werden, dies geschieht anhand der Funktion updateDistMatrix, die die zu den ursprünglichen Knoten gehörenden Zeilen und Spalten löscht und die passenden neuen anhängt. So verfahren wir iterativ bis sich in 'header' nur noch ein Eintrag befindet. Dieser enthält alle Informationen für den Tree-Block unserer NEXUS-Datei. Die von unserem Programm erstellte NEXUS-Datei übergaben wir dann Figtree und liessen uns die Bäume darstellen. Die von Figtree erstellten Bäume sind sowohl in Abbildung 4 und 5 dargestellt sowie im Repository im Branch "Bericht" hinterlegt.

3 Conclusion

Zusammenfassung: Die Ergebnisse der oben genannten Schritte sind in den Abbildungen 4 und 5 abgebildet. Abbildung 4 ist auf Basis der p-Distanz entstanden und Abbildung 5 auf Basis der Distanzmatrix nach Jukes-Cantor. Die Blaetter des Baumes, die Herkunftslaender der zugehoerigen Daten der Replikase-Sequenzen, sind farblich markiert, was die Verwandtschaft dieser Viren, bzw. ihre Ausbreitung, verdeutlichen soll. Die Baeume sind jedoch trotz unterschiedlicher Werte der zugrunde liegenden Matrizen fast identisch. Leider sind die Distanzen zwischen den Gensequenzen sehr gering, was sich vermutich auch auf die Baumdarstellung auswirkt. Jedoch ist zu erkennen, dass Daten, die aus einem Land kommen, wahrscheinlicher zusammen liegen, bzw. nicht sehr weit ueber Subtrees voneinander entfernt sind. Die meisten Daten dazu wurden aus China erhoben, weshalb diese auch zusammen liegen, bis auf die spaeter erfassten, diese liegen ueberraschend weit entfernt. Aus unserem Ergebnis ist zu schliessen, dass die zugrunde liegende Datensammlung ausschlaggebend fuer die Struktur des Baumes ist, zumal manche Sequenzen identisch waren. Je groesser die Datensammlung, desto komplexer wird der Baum, da wahrscheinlicher groessere Unterschiede auszumachen sind. Groessere Abweichungen wuerden auch mit deutlich zu unterscheidenden Sequenzen einhergehen. Bei den Versuchen, mit der Aminosaeuresequenz Baeume zu erstellen, zeigte sich auch in den Alignmentscores noch kleinere Unterschiede, vermutlich auch, weil diese Sequenzen noch weniger divergieren als die Gensequenzen dieser Abschnitte, moeglicherweise liegt dem die Wobble-Base-Hypothese zu Grunde und es spielt somit auch eine Rolle, was fuer Sequenzdaten gesammelt wurden um direkte Unterscheide festzustellen. Jedenfalls wird die Baumdarstellung durch die unterscheidlichen Verfahren der Distanzberechnung fast nicht beeinflusst, weshalb wir zu dem Schluss kamen, dass unsere gesammelten Daten zu aehnlich sind und quantitativ nicht genuegen. Es kam die Idee auf, andere verfahren zur Distanzberechnung zu implentieren, bspw. Kimaru's, um groessere Unterschiede der Distanzmatrizen zu gewaehrleisten, jedoch wuerde sich dies nicht sehr deutlich auf die Aenderungen der Baumstrukturen auswirken. Nachdem wir den Hinweis erhielten, eine Grundannahme des UPGMA bei der Baumdarstellung miteinzubeziehen, korrigierten wir zunaechst die erste Annahme des ultra-metrischen Matrixaufbaus, bei dem die Distanz zweier Blaetter zueinander groesser Null zu sein hat und entdeckten noch einen Fehler bei der synaktischen Erstellung des Baum-Strings, wobei die Distanzen u.a. falsch zugeordnet wurden. Die "Branch-Lenghts" sind fuer eine korrekte Darstellung auf gerade Zahlen transformiert. Die Auswahl der Sequenzen ist fuer Baeume, welche evolutionaere Daten aufzeigen koennen, von enormer Bedeutung und beeinflusst am staerksten das Resultat.

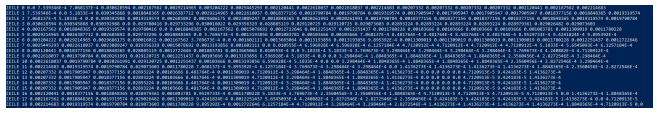

Abbildung 2: Die p-Distanz der alignierten Sequenzen

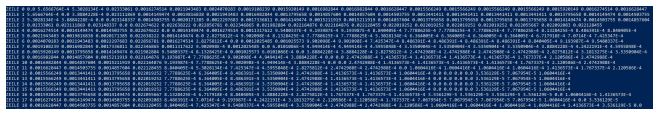

Abbildung 3: Die Distanzmatrixwerte der Sequenzen nach dem Jukes-Cantor-Modell

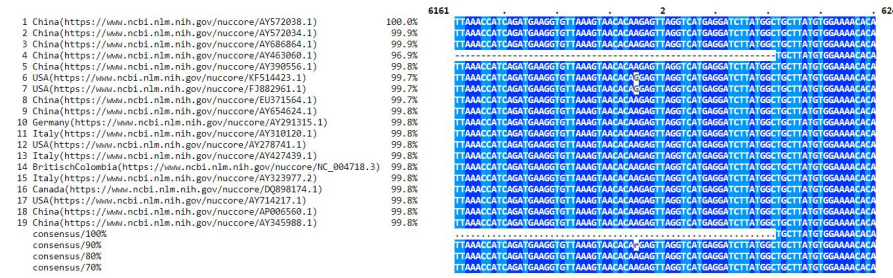

Abbildung 1: Alignment der 19 Sequenzen an zufaelliger Stelle, dargestellt durch "MView".

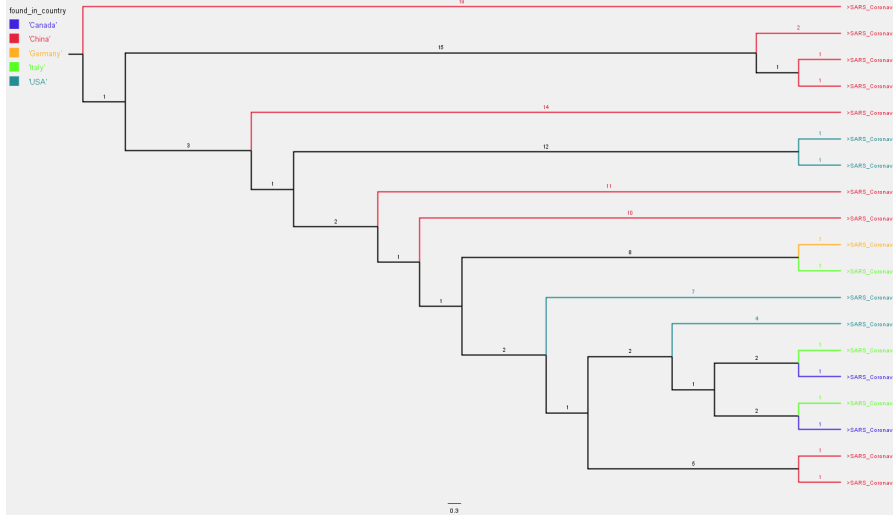

Abbildung 4: Der korrigierte, ausgegebene Baum auf Basis der P-Distanz-Matrix.

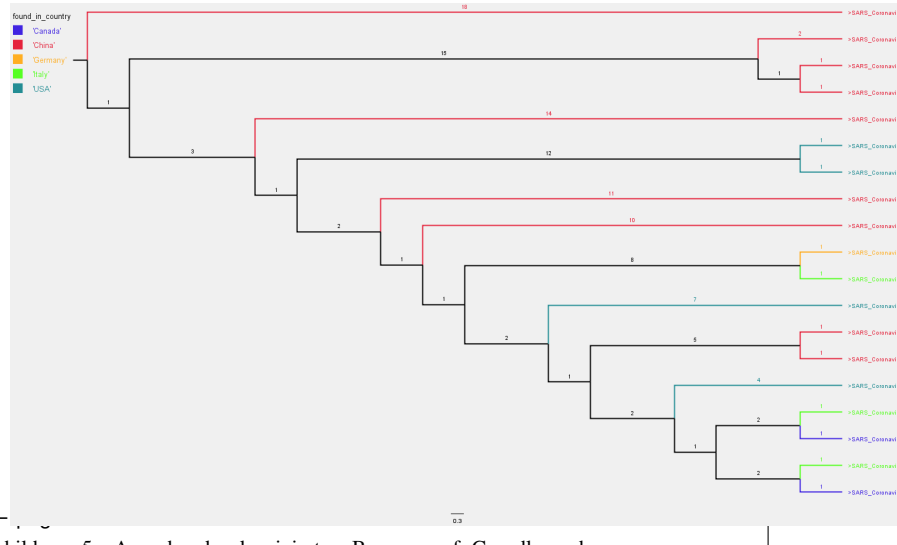

Abbildung 5: Ausgabe des korrigierten Baumes auf Grundlage der Distanzmatrix nach Jukes-Cantor.