Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: 12th of Jabuary 2018
Manuscript Category

Algorithmische Bioinformatik

Projekt 3

Paul Vogler, Tobias Mechura & Franziska Rau

Abstract

Projekt 3 des Moduls Algorithmische Bioinformatik
Ziel ist das Erstellen und die Visualisierung eines phylogenetischen Baums anhand von homologen Sequenzen eines Virus.

1 Einführung

H3N2 ist ein Subtyp der Virengattung Influenzavirus A, dessen Name sich von den beiden Arten von Proteinen auf der Oberfläche des Virus, Hämagglutinin (H) und Neuraminidase (N), ableitet.
Dieser Viren Subtyp tritt saisonal auf und kann Gene für interne Proteine mit anderen Subtypen austauschen.
Bei Schweinen sind weltweit drei Influenza-A-Subtypen (H1N1, H3N2 und H1N2) im Umlauf. Schweine können menschliche Influenzaviren in sich tragen, die sich mit H5N1 kombinieren (genetische Reassoziation) und so mutieren, dass sie leicht auf den Menschen übertragbar sind.
H3N2 entstand durch Antigen-Shift aus H2N2.
Im Verlauf des Praktikums wurde das Oberflächenprotein Hämagglutinin im Wirt Schwein in Europa untersucht. Haemagglutinin führt zur Agglutination der Erythrozythen im Wirt und macht etwa 80 Prozent der Virushülle aus.
Hämagglutinin vermittelt die Anheftung an den Rezeptor Neuraminsäure auf der Wirtszelle, sodass das Innere des Virions durch die Endosomenmembran ins Zytosol geschleust wird.

2 Ansatz

Aus einer selbstgewählten RNA, DNA oder Protein Sequenz werden zunächst geeignete Bereiche extrahiert und mittels multiplen Alignments und der Berechnung von Distanzen ein phylogenetischer Baum erstellt.
Als Kontruktionsverfahren für den Baum dient der UPGMA Algorithmus.
Die resultierenden Bäume werden im NEXUS Format ausgegeben und können so visualisiert werden.

3 Methoden

Zur Umsetzung des Projektes wurden Protein Sequenzen aus der NCBI Influenzavirus Datenbank[1] extrahiert.
In der Datenbank ist eine spezialisierte Suche möglich. Es wurde der Sequenztyp Protein gewählt und um die Suche einzugrenzen wurden die Parameter Influenza A, Subtyp H3N2, Protein Haemagglutinin (HA) und Region Europa ausgewählt. Desweiteren wurde nach Vollständigkeit der Sequenzen sortiert.
Die NCBI Influenza Datenbank bietet verschiedene Tools zur Weiterverarbeitung der Sequenzen an, unter anderem das Herunterladen einer .FASTA Datei und die Erstellung eines multiplen Alignments.
Zum Anpassen der .FASTA Datei wurden die Metadaten "region" und "year" ausgewählt.
Da es sich bereits um homologe Sequenzen handelte, wurde anhand des NCBI tools ein Multiples Alignment erstellt. Insgesamt wurden 95 Sequenzen aligniert mit einer Alignmentgesamtlänge von 567 Aminosäuren.
Das MSA wird unter dem "sum-of-pairs" Kriterium[2] durchgeführt, dazu wird für jedes Sequenzpaar der score berechnet und über alle Paare summiert. Standardmäßig wurde ein gewichteter Summenalgorithmus implementiert.
Das Alignment gibt die entsprechende Position der Aminoäure in der Sequenz und die aus den Sequenzen generierte Consensus-Sequenz an. Angezeigt werden die unterschiedlichen Aminosäuren und gaps pro Sequenz.
Anhand der alignierten Sequenzen wurden anschließend die paarweisen Distanzen berechnet und die Distanzmatrizen erstellt. Zuerst wurde die paarweise Hamming Distanz berechnet, diese beschreibt die minimale

[1] `https://www.ncbi.nlm.nih.gov/genomes/FLU/Database/nph-select.cgi?go=database`
[2] `https://www.ncbi.nlm.nih.gov/CBBresearch/Schaffer/msa.html`

Stellendistanz der Sequenzen.
Danach wurde anhand der BLOSUM62 Matrix eine weitere Distanzmatrix erstellt. Bei BLOSUM62 handelt es sich um eine Matrix, die individuelle Substitutionen bewertet. Sie basiert auf homologen Proteinsequenzen und es werden sowohl die absoluten Häufigkeiten, als auch die Häufigkeiten aller Übergänge in allen Paaren betrachtet.
Zum erstellen der phylogenetischen Bäume wurde der UPGMA Algorithmus verwendet. Dieser sucht sich im ersten Schritt den minimalen Wert der Eingabematrix, fügt die beiden entsprechenden Elemente zu einem neuen Knoten zusammen und entfernt die Elemente aus der Matrix. Die Distanz der beiden Elemente wird durch 2 geteilt, um so die Kantenlängen zwischen den Knoten zu bestimmen. Die Distanz vom neuen Knoten zu allen anderen wird wie folgt berechnet:

$$d_{zw} = \frac{N_x d_{xw} + N_y d_{yw}}{N_x + N_y}$$

Wobei N die Anzahl der Sequenzen im Cluster ist und d die Distanz der Knoten.
Nachfolgend wird wieder der kleinste Wert aus der Matrix genommen und zu einem neuen Knoten zusammengefügt. Zur Berechnung der Distanz von 2 Clustern wird folgende Formel verwendet:

$$d_{xy} = \frac{1}{N_x + N_y} \sum_{i \in X, j \in Y} d_{ij}$$

Die resultierenden Bäume werden im NEXUS Format ausgegeben, um sie so zu visualisieren. Die NEXUS Datei enthält den Baum, die Kantenlängen und die Metadaten.
Zur Visualisierung diente die Software FigTree[3], welche die Nexus Datei einliest und eine Grafik daraus erstellt. Der dargestellte phylogenetische Baum lässt sich entsprechend anpassen: Einfärben der Kanten anhand der Metadaten und Erstellen einer Legende.

4 Diskussion

Da es sich bei der derzeitigen saisonalen Grippe um den Virus Influenza A H3N2 handelt, wurde dieser für das Projekt ausgewählt. Dementsprechend wurde dann gezielt nach einer Datenbank gesucht, die viele Influenza Daten und Sequenzen enthält.
Nach Fund der NCBI Datenbank Influenza konnte dann die Suche eingeschränkt werden. Zunächst wurden verschiedene Filter ausprobiert, um die Anzahl der Sequenzen einzuschränken. Beim Wirt Mensch wurden meist weit über 2000 Sequenzen ausgegeben, deshalb wurde sich auf den Wirt Schwein festgelegt, da dieser ebenfalls Träger des Influenza A H3N2 Virus sein kann. Die Auswahl des Proteins fiel auf Haemagglutinin, da dieses ein wichtigtes Oberflächen Protein des Virus ist, ebenso wie Neuraminidase.
Desweiteren bestand eine große Auswahl an geographischen Regionen. Nach einigen Suchanfragen wurde die Region Europa ausgewählt, da zum Beispiel bei der Region Afrika nur sehr wenige Seqeunzen vorlagen. Nach weiterem Filtern nach Vollständigkeit, betrug die Gesamtsequenzanzahl 95.
Die ausgewählten Sequenzen wurden anschließend über das NCBI Webseitentool multipel aligniert.
Anhand des multiplen Alignments, konnte festgestellt werden, dass die verwendeten Sequenzen sehr ähnlich sind (siehe Fig. 1).Anschließend wurden, auf dem Alignment basierend, die paarweisen Distanzen berechnet, einmal mit der Hamming Distanz (siehe Fig. 2) und dann mit Hilfe einer modifizerten BLOSUM62-Matrix (siehe Fig. 3).

[3] http://tree.bio.ed.ac.uk/software/figtree/

Die Hamming-Distanz ist definiert über die Anzahl Unterschiede zweier Aminosäuren an derselben Stelle in den zu vergleichenden Sequenzen.
Die BLOSUM62-Matrix hat für alle möglichen AminosÃ¤urekombinationen einen score gespeichert, der für jede Stelle in den Sequenzen addiert wird. Allerdings können sich auch negative Distanzen ergeben, die in einem phylogenetischen Baum nicht darstellbar sind. Dieses Problem wurde gelöst, indem die Werte der Substitutionsmatrix mit -1 multipliziert und dann der kleinste Wert, 11 addiert wurde. Außerdem wurde die Matrix um den Eintrag für eine gap erweitert, damit diese nicht einfach ignoriert wird.

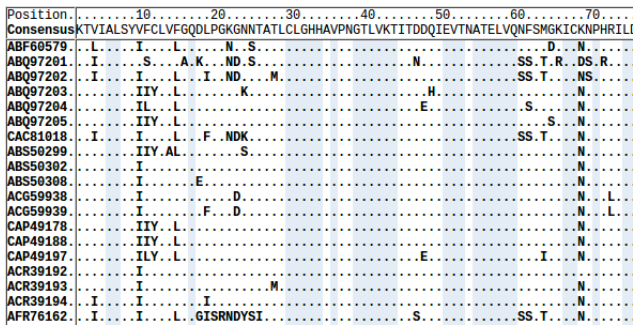

Fig. 1: Ausschnitt MSA
Die Sequenzen ähneln sich sehr und haben nur wenige Abweichungen voneinander. Insgesamt befindet sich meistens nur eine gap in einigen Seqeunzen, was darauf schließen lässt, dass die Sequenzen auch annähernd gleich lang sind.

0.0	5.0	19.0	16.0	18.0	12.0	16.0	15.0	13.0	83.0
5.0	0.0	24.0	21.0	23.0	17.0	21.0	20.0	18.0	86.0
19.0	24.0	0.0	7.0	29.0	27.0	28.0	28.0	26.0	94.0
16.0	21.0	7.0	0.0	26.0	24.0	25.0	24.0	24.0	92.0
18.0	23.0	29.0	26.0	0.0	25.0	26.0	25.0	26.0	94.0
12.0	17.0	27.0	24.0	25.0	0.0	17.0	13.0	14.0	86.0
16.0	21.0	28.0	25.0	26.0	17.0	0.0	12.0	20.0	92.0
15.0	20.0	28.0	24.0	25.0	13.0	12.0	0.0	19.0	90.0
13.0	18.0	26.0	24.0	26.0	14.0	20.0	19.0	0.0	84.0
83.0	86.0	94.0	92.0	94.0	86.0	92.0	90.0	84.0	0.0

Fig. 2: Hamming Distanzen
Die Hamming-Distanz ergibt sich aus der Anzahl an unterschiedlichen Aminosäuren innerhalb von zwei Sequenzen.

0.0	3197.0	3255.0	3247.0	3263.0	3236.0	3248.0	3246.0	3232.0	3558.0
3197.0	0.0	3273.0	3265.0	3281.0	3254.0	3266.0	3264.0	3250.0	3566.0
3255.0	3273.0	0.0	3202.0	3308.0	3296.0	3297.0	3299.0	3282.0	3598.0
3247.0	3265.0	3202.0	0.0	3300.0	3288.0	3289.0	3289.0	3278.0	3594.0
3263.0	3281.0	3308.0	3300.0	0.0	3301.0	3301.0	3299.0	3294.0	3612.0
3236.0	3254.0	3296.0	3288.0	3301.0	0.0	3273.0	3256.0	3248.0	3592.0
3248.0	3266.0	3297.0	3289.0	3301.0	3273.0	0.0	3260.0	3272.0	3603.0
3246.0	3264.0	3299.0	3289.0	3299.0	3256.0	3260.0	0.0	3264.0	3603.0
3232.0	3250.0	3282.0	3278.0	3294.0	3248.0	3272.0	3264.0	0.0	3567.0
3558.0	3566.0	3598.0	3594.0	3612.0	3592.0	3603.0	3603.0	3567.0	0.0

Fig. 3: BLOSUM62 Distanzen
Durch die modifizierte Substitutionsmatrix ergeben sich nur positive scores, aber auch deutlich größere als bei der Hamming Distanzberechnung.

Deutlich zu erkennen ist, dass Werte der Hamming Distanzmatrix wesentlich kleiner sind, als die der BLOSUM Distanzmatrix. Das liegt daran, dass die veränderte BLOSUM62-Matrix Anwendung fand. Die Hamming Distanz ist im Gegensatz die einfache Summe der Unterschiede der jeweiligen Sequenzen und deshalb auch geringer.
Nachdem der UPGMA Algorithmus angewandt wurde, konnten die Bäume mittels der Software Figtree visualisiert werden (siehe Fig. 4 bis 7)

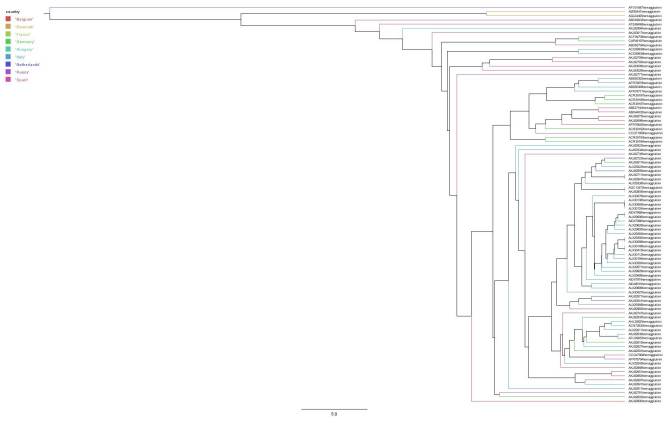

Fig. 4: Phylogenetischer Baum basierend auf Hammingdistanz mit den Metadaten "country"

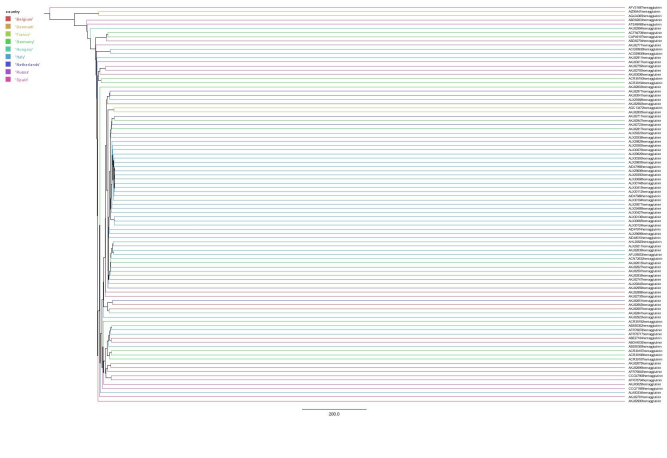

Fig. 5: Phylogenetischer Baum basierend auf BLOSUM Distanz mit den Metadaten "country"

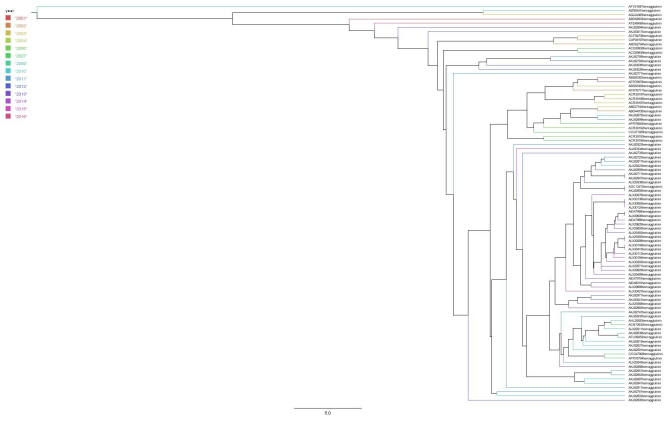

Fig. 6: Phylogenetischer Baum basierend auf Hammingdistanz mit den Metadaten "year"

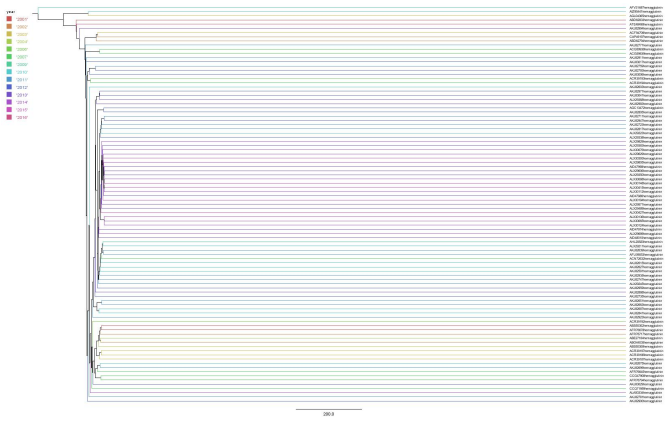

Fig. 7: Phylogenetischer Baum basierend auf BLOSUM Distanz mit den Metadaten "year"

Die distanzmatrixbasierten Bäume stellen die verschiedenen Haemagglutinin Taxa dar, zum einen mit den Metadaten "country" (Fig. 4 & 5) und zum anderen mit den Metadaten "year" (Fig. 6 & 7).
Die Distanzen stellen den größten Unterschied in den phylogenetischen Bäumen dar. In den Bäumen mit BLOSUM-Distanz ist extrem deutlich sichtbar, dass die evolutionäre Distanz zwischen den Sequenzen nur geringfügig variiert und die verschiedenen Aminosäuren innerhalb aller Sequenzen eine ähnliche Wirkung im Protein haben und somit das Protein in seiner Funktion nicht komplett verändert wurde. Desweiteren ist in den auf BLOSUM-Distanz basierenden Bäumen klar zu sehen, dass Proteinsequenzen aus geographisch verwandten Regionen mehr gemeinsam haben als jene aus geographisch weiter entfernten Regionen. Auch bei Betrachtung der Jahrgänge wird dies bestätigt.
Bei den phylogenetischen Bäumen die auf der Hamming-Distanz basieren können diese Aussagen auch geäußert werden. Jedoch befinden sich dort deutlich mehr Ausreißer, als dass man sich sicher sein könnte, korrekte Annahmen getroffen zu haben.

4.1 Probleme

Zuerst wurde mit der Programmiersprache Java gearbeitet, die Implementierung von UPGMA ergab allerdings einige Probleme. Deshalb wurde zur Programmiersprache Python gewechselt, da sich dort die Verarbeitung von Datensätzen als leichter erwies. Der Code konnte jedoch größtenteils einfach umgeschrieben werden.

4.2 Arbeitsaufteilung

Größtenteils wurde für das Projekt gemeinsam an einem PC gearbeitet. Lediglich kleinere Aufgaben wurden vereinzelt alleine bearbeitet.