UBBO-EMMIUS-GYMNASIUM

Schuljahr 2022/2023

Seminarfacharbeit im Kurs sf48

Dietmar Meyer

Wege zur Mathematisierung genetischer Verwandtschaft

Sidra Omar

Note: _____

Punktzahl: _____

Unterschrift des Kursleiters

11. Dezember 2023

Inhaltsverzeichnis

1 Einleitung

Bereits im 18 Jhr. widmeten sich zahlreiche Wissenschaftler den Ansätzen der Evolutionsbiologie und der Klassifikation von Lebewesen. In seinem Hauptwerk „On the origin of species“ hat Charles Darwin mit seiner Theorie, dass alle Lebewesen vom gemeinsamen Vorfahren abstammen, eine kausale und logische Sicht auf die Evolution und ihre Entwicklung geschaffen. Letzteres beschreibt die Studie der Phylogenie. Die Phylogenie bezieht sich auf die Stammesgeschichte bzw. die Verwandtschaftsverhältnisse zwischen verschiedenen Organismen und ihren gemeinsamen Vorfahren. Basierend auf ihre phänotypische[1] Merkmalen wurden Lebewesen erstmals von Carl von Linné klassifiziert. Doch die Entdeckung der genetischen Struktur war ein bedeutender/ausschlaggebender Durchbruch in diesem Wissenschaftsbereich. So wuchs mit der Möglichkeit zur Genomsequenzierung das Interesse (und die Notwendigkeit), Veränderungen der DNA-Sequenzen zu verfolgen und Einblicke in die Evolution von Arten zu gewinnen. In den letzten Jahrzehnten nahmen allerdings die faszinierenden Fortschritte der Biotechnologie und Bioinformatik dermaßen zu, dass die Notwendigkeit zur Entwicklung eines neuen Bereichs, der sich als Mischung aus Mathematik und Biologie definieren lässt, hervorgebracht wurde. Die Disziplin der Biomathematik beschäftigt sich mit der Anwendung der Mathematik auf Probleme biologischer Natur, sodass Lösungsansätze basierend auf Modellierungen für komplexe biologische Fragestellungen entwickelt werden können.
Denn es ist nicht nur die wissenschaftliche Leidenschaft zur Verfolgung des längst Verborgenem hinter der Genetik, sondern die variablen Möglichkeiten und Chancen dieses Forschungsgebiets. In dieser Seminararbeit wird der evolutionäre Abstand zwischen zwei DNA-Sequenzen anhand verschiedenster mathematischer Modellierungen gemessen. Es wird die Bedeutung dieser Messungen für die Biologie untersucht, ggf. werden Herausforderungen und Grenzen der angewendeten mathematischen Techniken diskutiert.

2 Biologische Hintergründe

Um die zentrale Fragestellung dieser Seminarfacharbeit beantworten, muss der molekulare Aufbau der Desoxyribosenukleinsäure, kurz DNA, sowie der Basen, wie diese repliziert und an die nächste Generation weitergegeben werden und auftretende Mutationsarten und ihre Auswirkungen verstanden und erläutert werden. Die DNA befindet sich in jedem Zellkern einer eukaryotischen Zel-

[1] Ein Phänotyp ist die Menge aller Merkmale auf morphologischer und physiologischer Ebene eines Organismus

le[2]und speichert die Erbinformationen.

2.1 Molekularer Aufbau der DNA

Nach der Konstruktion von Watson und Crick ist die Desoxyribonukleinsäure ein Makromolekül bestehend aus einem spiralig gewundenen Doppelstrang, der die Struktur einer um ihrer eigenen Achse gedrehten Strickleiter hat. Diese Struktur wird aufgrund ihres Aussehens als Doppelhelix bezeichnet.
Die Grundbausteine der DNA sind die Nukleotide, die aus jeweils einem Zucker-

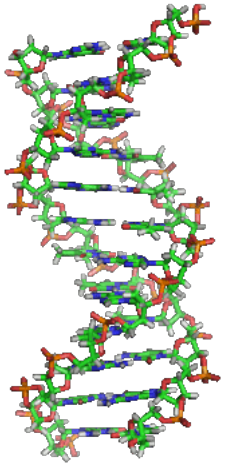


Abbildung 1: DNA als Doppelhelix

molekül (Desoxyribose), einer Phosphatgruppe und einer der vier organischen Basen Adenin, Thymin, Cytosin, Guanin bestehen. Das Rückgrat der Doppelhelix besteht aus einer alternierenden Verknüpfung von Phosphatmolekülen und Desoxyribosemolekülen, wobei die Phosphatgruppe abwechselnd an dem dritten und fünften Kohlenstoffatom (C-Atom) bindet. An einem Einzelstrang bleibt am Ende die Bindungsmöglichkeit am dritten C-Atom, am anderen Strang die Bindungsstelle am fünften C-Atom frei.
Dementsprechend hat jeder Strang eine Richtung, sodass die beiden Stränge antiparallel zueinander verlaufen. Man unterscheidet zwischen dem Strang mit dem 3´-Ende, wo die Bindungsstelle am dritten C-Atom des letzten Desoxyribose-Moleküls frei ist, und dem Strang mit dem 5‘-Ende, an dem das fünfte C-Atom des letzten Desoxyribose-Moleküls frei ist, was bedeutet, dass das 3´-Ende eines Stranges mit dem 5´-Ende des anderen Stranges verbunden ist. Die stickstoffhaltige Base ist immer an dem ersten C-Atom des Desoxyribosemoleküls gebunden. Die Verknüpfung der beiden Einzelstränge geschieht über die Basen der gegenüberliegenden Nukleotide. Aufgrund der chemischen Strukturen und

[2]Ein Eukaryot: ein Lebewesen, das einen Zellkern in seinen Zellen hat.

den polaren Gruppen der parallelen Basen können nur eine lange Purin-Base, Adenin oder Guanin, mit einer kürzeren Pyrimidin-Base, Cytosin oder Thymin, miteinander verknüpft werden. Dabei können nur zwei Konstellationen zustande kommen: die zueinander komplementären Basen Thymin und Adenin bilden zwei Wasserstoffbrückenbindungen[3] untereinander, die zwei anderen komplementären Basen, Guanin und Cytosin, bilden drei Wasserstoffbrückenbindungen.

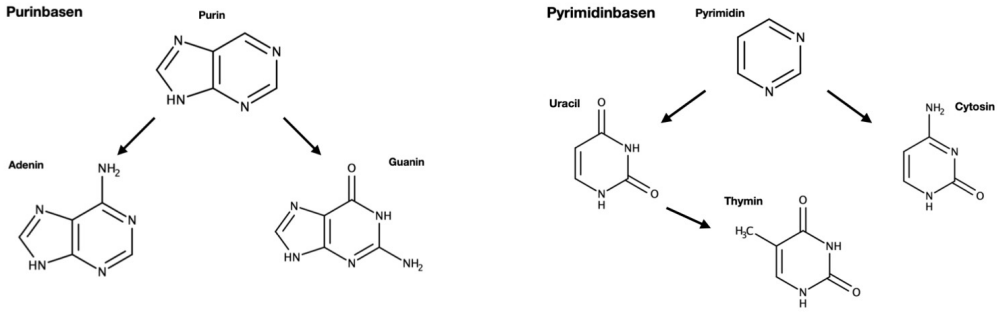


Abbildung 2: Chemische Struktur der Basen

2.2 Speicherung und Weitergabe der DNA

Die genetische Information wird durch die DNA-Sequenz, also die Abfolge der Nukleotide, getragen und gespeichert. Informationstragende Abschnitte bezeichnet man als Gen oder als codierenden Abschnitt.
Ausgehend von einer codierenden DNA-Sequenz wird bei der Proteinbiosynthese[4] aus einer speziellen und einzigartigen Aminosäurensequnez ein Protein entschlüsselt, das als Baustein des Körpers dient und eine Vielzahl an Funktionen erfüllt. Drei aufeinanderfolgende Basen, was als Basentriplett bezeichnet wird, können für eine spezifische Aminosäure codieren.
Die Weitergabe der genetischen Information erfolgt über die Eltern an ihre Nachkommen. Die Gameten[5] tragen einen haploiden Chromosomensatz. Durch die Fusion beider Gameten wird ein neuer diploider Chromosomensatz gebildet, das eine Kombination aus den beiden haploiden Chromosomensätze beider Eltern darstellt.

2.3 Replikation der DNA und Mutationsarten

Um eine exakte Kopie der DNA an die nächste Generation weiterzugeben, müssen die ursprünglichen Erbinformationen verdoppelt werden. Diesen Pro-

[3] Anziehungskraft zwischen einem Wasserstoffatom eines Moleküls und einem freien Elektronenpaar eines anderen Moleküls; stärkste intermolekulare Wechselwirkung.
[4] Prozess zur Produktion von Proteinen
[5] geschlechtliche Keimzellen

zess bezeichnet man als Replikation der DNA. Bei diesem semikonservativen Prozess dient der alte Einzelstrang als Vorlage des neu synthetisierten Komplementärstranges. Zunächt wird der Doppelstrang in zwei Einzelstränge gespalten. Der neue Einzelstrang wird kontinuierlich synthetisiert, indem komplementäre Nukleotide an den Nukleotiden des alten Einzelstranges angefügt werden. Es findet eine andauernde Verlängerung statt, sodass am Ende jede doppelsträngige DNA-Sequnez aus einem alten und einem neuen Einzelstrang besteht.

Die Replikation ist ein präziser Prozess, bei dem ständig die Fehlerquoten durch Reparaturmechanismen während des Prozesses gesenkt werden. Nachträgliche Reparatursysteme steigern weiterhin die Präzision der Kopie. Falls ein Fehler trotz der Reparaturmechanismen bestehen bleibt, dann könnte es zu Veränderungen in der DNA-Sequenz führen, was gravierende Auswirkungen auf die Merkmale, Überlebens- und Fortpflanzungsmöglichkeiten in der nächsten Generation haben kann. Eine solch vererbbare Veränderung nennt man Mutation.
Man unterscheidet zwischen mehreren Arten von Mutationen, die jeweils unterschiedliche Effekte/ Konsequenzen haben können. Wichtig zu erwähnen ist, dass manche auftretende Mutationen keine sichtbaren oder sogar vorteilhafte Auswirkungen auf die Merkmale haben können. Es ist stark abhängig von der Position und der Art der Mutation Für diese Facharbeit sind folgende Mutationsarten relevant:
Punktmutation: Substitution einer einzelnen Base innerhalb eines Einzelstranges durch eine andere Base, wobei man zwischen der Transition und der Transversion unterscheiden muss. Bei der Transition wird eine Purin-Base durch eine andere Purin-Base oder eine Pyrimidin-Base durch eine andere Pyrimidin-Base ersetzt wird. Eine Transversion ist eine Mutation, bei der eine Purin-Base durch eine Pyrimidin-Base oder umgekehrt ersetzt wird.

Insertion: Bei einer Insertion werden eine oder mehrere zusätzliche Basen in die DNA-Sequenz eingefügt. Dies kann zu einer Verschiebung des Leserasters führen und somit zu einer Veränderung der Aminosäuresequenz.

Deletion: Bei einer Deletion werden eine oder mehrere Basen aus der DNA-Sequenz entfernt. Dies kann ebenfalls zu einer Verschiebung des Leserasters führen und somit zu einer Veränderung der Aminosäurensequenz.

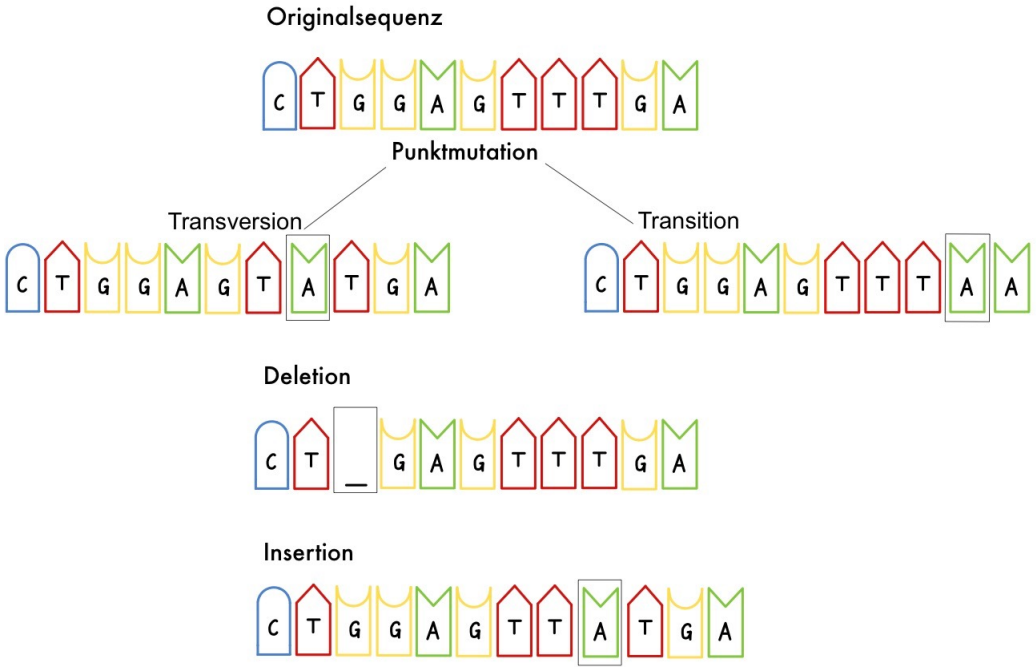


Abbildung 3: Arten von Mutationen

3 Hamming-Abstand

Der sogenannte Hamming-Abstand wird ursprünglich für Unterschiede zwischen binären Codes verwendet. Bezogen auf die Facharbeit wird der Hamming-Abstand durch Annahmen und Vereinfachungen so weit modifiziert und auf die Modellierung angepasst, dass es auf die biologische Problemstellung angewendet werden kann.
Der evolutionäre Abstand wird bei dieser Methode als die Anzahl an unterschiedlichen Basen zwischen den zu analysierenden Sequenzen definiert. Bei den hier eingesetzten Datensätzen sind die beiden zu vergleichenden Sequenzen auf als Vereinfachung gleiche Länge gebracht worden. Dementsprechend wird gleichzeitig die Annahme gemacht, dass im Verlauf der Evolution nur Punktmutationen mit gleicher Mutationswahrscheinlichkeit auftreten können, da sich bei einer Deletion die Längen der Sequenzen verändern würden.

3.1 Mathematische Definition

Der Hamming-Abstand kann wie folgt mit diesem Ausdruck beschrieben werden:

$$d_H : S \times S \to R, (S_1, S_2) \mapsto \frac{|\{S_{1,i} \neq S_{2,i,}, i \in (1, ..., n)\}|}{n}$$

Der Hamming-Abstand wird als d_H bezeichnet und hat als Eingabe die Menge aller Nukleotide in S_1 und S_2 und als Ausgabe die Anzahl der Positionen i an denen Unterschiede zwischen den Nukleotiden der Sequenzen auftreten.

Unter $|\{S_{1,i} = S_{2,i}, i \in \{1, ..., n\}\}|$ versteht man die Menge der Indizes[6] i, wo die i-ten Stellen von S_1 und die i-Stelle von S_2 ungleich sind. Die Indizes i deckt alle Positionen von 1 bis n ab, da die Variable n die Länge von S_1 und S_2 angibt.

3.2 Beispiel

Gegeben sind die beiden Sequenzen S_1: ATGGCACTGATT und S_2: ACGACACTGAAT.

Position	1	2	3	4	5	6	7	8	9	10	11	12
Sequenz 1	A	T	G	G	C	A	C	T	G	A	T	T
Sequenz 2	A	C	G	A	C	A	C	T	G	A	A	T
Wert	0	1	0	1	0	0	0	0	0	0	1	0

Abbildung 4: Zwei Sequenzen im Vergleich

Zunächst wird die Länge n der beiden Sequenzen bestimmt, was $n = 12$ ergibt, d.h. beide Sequenzen haben 12 Nukleotide. Daher ist $i \in 1, .., 12$. An Position i 2,4 und 11 weisen S_1 und S_2 Unterschiede auf. Das Distanzsmaß d_H beträgt also 3. Das Distanzmaß zwischen den Sequenzen S_3: ATGCGGATTTC und S_4: ATTCAGATTTC beträgt 2. Sie sind sich ähnlicher und damit evolutionär und verschwandtschaftlich näher aneinander. als S_1 und S_2.

3.3 Validation des Ergebnis

Eine höhere Anzahl an Mutationen bzw. an Unterschieden zwischen den Sequenzen bedeutet ein größeres Distanzmaß. Der ermittelte Hamming-Abstand stimmt also mit der aufgestellten Monotonieannahme überein. Mit der Simplizität und Schnelligkeit zur Berechnung des Hamming-Abstandes nimmt auch die Ungenauigkeit der Technik zu. Besonders wenn die Datensätze stark variieren, kann dieses distanzbasierte Verfahren unzureichend sein, da andere Mutationsarten wie Insertion oder Deletion ausgeschlossen werden. Zudem verlieren die zu vergleichenden Sequenzen durch die Annahme, dass die Abschnitte aller Sequenzen gleich lang sind, wertvolle Informationen und es kommt zu unpräzisen Ergebnissen. Besonders wichtig ist außerdem, dass der Abstand basierend auf beobachtbare und an den Sequenzen sichtbare Mutationen getroffen wird. Vergangene Mutationen werden nicht mit berücksichtigt. Daher ist es sinnvoll, eine andere Art der Modellierung heranzuziehen.

[6] in diesem Kontext bedeutet es die Postion des Nukleotids in der Sequenz

4 Markov-Modelle

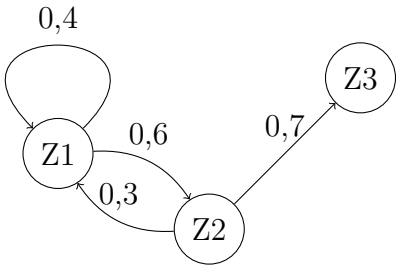
Die Markov-Modelle finden vielfältige Einsatzmöglichkeiten zum Lösen diverser Probleme, die auf Vorhersagen über zukünftige Ereignisse angewiesen sind, wie z.B. in der Finanzwelt (Berechnen von möglichen Transaktionen) oder eben in der Biologie. Daher eignet sich diese Methode zur Betrachtung der Abstandsbestimmung zweier Sequenzen gut, was im Folgenden dokumentiert wird.

Mathematische Definition: Ein Markov-Modell ist ein stochastischer Prozess, der die zeitliche Entwicklung von zufallsabhängigen Systemen beschreibt. Es stellt die Wahrscheinlichkeit einer Abfolge von Ereignissen oder Zuständen dar.
Gleichung:

$$P(X_{n+1} = e_{n+1}|X_0 = e_0, X_1 = e_1, \ldots, X_n = e_n) = P(X_{n+1} = e_{n+1}|X_n = e_n)$$

Dieser Prozess wird über den Zustandsraum, der Startverteilung und der konstanten Übergangswahrscheinlichkeit definiert. Der Zustandsraum M
$\{e_0, e_1, \ldots, e_n\}$ enthält alle abzählbaren Zustände der Markov-Kette. Die Startverteilung beschreibt, in welchem Zustand sich das System zu Beginn befindet. In der Formel wird es durch $X_0 = e_0$ gekennzeichnet. Die Übergangswahrscheinlichkeit gibt an, wie hoch die Wahrscheinlichkeit zum Wechsel von dem einen in den anderen Zustand ist. In der Formel wird die Übergangswahrscheinlichkeit durch die bedingte Wahrscheinlichkeit $P(X_{n+1} = e_{n+1}|X_n = e_n)$ dargestellt, was bedeutet, dass der nächste Zustand (X_{n+1}) gleich e_{n+1} ist, unter der Bedingung, dass der aktuelle Zustand (X_n) gleich e_n ist. Charakteristisch für die Markov-Modelle ist, dass diese Übergangswarscheinlichkeit nur vom aktuellen Zustand und nicht von vorherigen Zuständen abhängig ist. Man bezeichent diese Eigenschaft als "Gedächnislosigkeit".
Darstellungsformen:
Die Festlegung der endlichen Zustände, die das System annehmen kann, lassen sich mithilfe von Graphen darstellen. Typische Darstellungsformen sind z.B. Prozessdiagramme. Auch können sie in Übergangsmatrizen überführt werden, die den Zufallsprozess vereinfacht anzeigen.
Allgemeines Beispiel:
Angewandt an einem Beispiel lassen sich die Markov-Ketten genauer verdeutlichen. Als Beispiel wird die Autofahrt des Nachhausewegs während eines Staus herangezogen. Zustand 1 (Z1) ist hier der Stau. Zustand 2 ist die Auflösung des Staus und Zustand 3 die pünktliche Ankunft Nachhause. Die Übergangswahrscheinichkeiten werden durch die Kanten zwischen den Zuständen dargestellt.

Die Wahrscheinlichkeit, dass sich der Stau auflöst, also eine Zustandsänderung



$$\begin{pmatrix} 0,4 & 0,3 & 0 \\ 0,6 & 0 & 0 \\ 0 & 0,7 & 0 \end{pmatrix}$$

Abbildung 5: Markov-Kette zur Darstellung der Autofahrt auf dem Nachhauseweg während eines Staus und zugehörige Übergangsmatrix

von Z1 zu Z2, beträgt 60 Prozent. Dass sich der Stau nicht auflöst und bestehen bleibt hat eine Wahrscheinlichkeit von 40 Prozent und wird durch eine Schlaufe im Graph dargestellt. Eine Rückbildung des Staus ist zu 30 Prozent wahrscheinlich, eine Zustandsänderung von Z2 zu Z1. Eine pünktliche Ankunft Zuhause ist dementsprechend zu 70 Prozent wahrscheinlich. Die Summe der Wahrscheinlichkeit zum Übergang von einem Zustand in den nächstmöglichen, also die Summe aller ausgehenden Kanten von einem Knoten, entspricht immer 1. IN der Übergangsmatrix ist das die Summe jeder Spalte.

Übertragung auf die Biologie:

Durch seine zuletzt erläuterten Eigenschaften eignet sich das Markov-Modell mit einigen Modifikationen und Anpassungen an den biologischen Kontext gut zur Anwendung auf die Problemfrage dieser Facharbeit. Man nimmt an, dass es ein Substitutionsmodell ist, d.h. es werden nur Punktmutationen zur Veränderung der Sequenz berücksichtigt.

Die Sequenzen der verwendeten Datensätze entsprechen der gleichen Länge. Für unterschiedliche Mutationstypen (siehe Abbildung von Mutationstypen, 2.3) können unterschiedliche Mutationsraten definiert werden. Im Gegensatz zur Hamming-Methode berücksichtigt das Markov-Modell auch Mutationen, die den Datensätzen nicht entnommen werden können, wie z.B. Rückmutationen. Bei diesem stochastischen Analyseverfahren kann man also im Bezug auf DNA-Sequenzen in die Vergangenheit auf die phylogenetische Entwicklung blicken, aber auch Vorhersagen über nachfolgende Mutationen treffen. Basierend auf den oben gestellten Annahmen, wird zur mathematischen Modellierung mit dieser Technik ein modifizierter Abstandsbegriff definiert:

$$P_{XY}(t) = P(B_{1,t+t_0} = Y | B_{1,t_0} = X), X, Y \in A, C, G, T$$

$B_{1,t}$ ist in dem Fall die Abfolge bzw. die Zufallsvaribale. Sie gibt den Zustand der Base B an der i-ten Stelle Von S_1 zum Zeitpunkt t an. Die Gleichung zeigt die Übergangswahrscheinlichkeit von der Base vom aktuellen Zustand X in

den nächsten Zustand Y in einem festgelegten Zeitraum von t_0 bis $t + t_0$ an. Auf eine Übergangsmatrix übertragen sieht es wie folgt aus:

$$\begin{pmatrix} P(B_{1,t} = A|B_0 = A) & P(B_{1,t} = A|B_0 = C) & P(B_{1,t} = A|B_0 = G) & P(B_{1,t} = A|B_0 = T) \\ P(B_{1,t} = C|B_0 = A) & P(B_{1,t} = C|B_0 = C) & P(B_{1,t} = C|B_0) = G) & P(B_{1,t} = C|B_0 = T) \\ P(B_{1,t} = G|B_0 = A) & P(B_{1,t} = G|B_0 = C) & P(B_{1,t} = G|B_0 = G) & P(B_{1,t} = G|B_0 = T) \\ P(B_{1,t} = T|B_0 = A) & P(B_{1,t} = T|B_0 = C) & P(B_{1,t} = T|B_0 = G) & P(B_{1,t} = T|B_0 = T) \end{pmatrix}$$

Abbildung 6: Übergangsmatrix

$P(B_{1,t} = C|B_0 = A)$ gibt also die Wahrscheinlichkeit wieder, dass die Base A in Base C mutiert.
Es gibt einige Methoden, die eine Spezifizierung der Markov-Modelle darstellen. Sie wurden für die Zwecke der Genomforschung basierend auf den Markov-Modellen entwickelt. Diese werden in den folgenden Abschnitten erläutert werden.[7]

4.1 Jukes-Cantor-Modell

Das Jukes-Cantor-Modell, auch 1-Parameter-Modell genannt, ist ein mathematisches Modell in der Bioinformatik, das zur Schätzung der Evolutionsrate von DNA-Sequenzen von W. James Jukes und Christopher R. Cantor 1969 entwickelt worden ist. Es ist eine Spezifizierung von den Markov-Modellen, mit dem erst die Bestimmung des evolutionären Abstandes ermöglicht wird. Dabei haben die zu analysierenden Sequenzen die gleiche Substitutionwahrscheinlichkeit. Alle Mutationstypen der Punktmutation haben die gleiche Mutationsrate (in der Abbildung α), wenn A zu C oder zu G oder zu T mutiert ist die Wahrscheinlichkeit aller Substitutionen gleich und sie treten gleich häufig auf. Außerdem wird angenommen, dass die Entwicklung beider Sequenzen aus der Vorläufersequenz S_0 gleichgesetzt werden können.
Abstandsbegriff des Jukes-Cantor-Modells:

Abstand zwischen S_1 und $S_2 = 2 \cdot$ erwartete Anzahl der während der Entwicklung von S_1 aus $S_0 = 2 \cdot (r \cdot t)$

Gleichung: $d_{JC} = \frac{-3}{4} ln(1 - \frac{4}{3} \cdot p)$
Diese Gleichung gibt die Anzahl an eingetreten Mutationen bzw. nach diesem Abstandsbegriff den Abstand zwischen den Sequenzen an. Die Multiplikation am Anfang der Gleichung mit dem Faktor $\frac{3}{4}$ stammt von der Annahme, dass die Wahrscheinlichkeit einer auftretenden Mutation pro Zeit und pro Stelle gleich ist. p ist die prozentuale Abweichung. p ist der Quotient aus der Anzahl an unterschiedlichen Positionen zwischen den Sequenzen und der Anzahl der gesamten Nukleotiden einer Sequenz. Je größer p ist , also je größer der

[7] vgl. Sube, Frank und Walcher[3]

Unterschied zwischen den Sequenzen ist, desto größer wird die Anzahl der Substitutionen sein, die zwischen ihnen aufgetreten sind, und desto größer wird auch der evolutionäre Abstand d_{JC}.

Beispiel 1:

Gegeben seien die Sequenzen S_1: ATTGCTAG und S_2: CTTGATAG

Zunächst erfolgt die Ermittlung der prozentualen Abweichung p. Die Anzahl an Unterschieden in S_1 und S_2 beträgt 2, da sich die Sequenzen an Position 1 und 5 unterscheiden. Die Gesamtzahl an Nukleotiden beträgt 8, also ist $p = \frac{2}{8} \cdot 100\% = 25\%$. Dies kann man nun in die Formel setzen: $d_{JC} = \frac{-3}{4} \cdot ln(1\frac{-4}{3} \cdot 0.25) = 0.304$ Das heißt, dass es 0,304 Substitutionen pro Stelle zwischen den beiden Sequenzen gibt.

Beispiel 2:

Man kann auch ein Beispiel nehmen, wo die Sequenzen deutlich länger sind. Folgendes ist über den bekannt Die beiden Sequenzen S_1 und S_2 haben je 300 Nukleotide. Die Anzahl an unterschiedlichen Nukleotiden beträgt 37. Die prozentuale Abweichung $p = \frac{37}{300} \cdot 100\% = 12,3\%$ $d_{JC} = \frac{-3}{4} \cdot ln(1 - \frac{4}{3} \cdot 0.123) \approx 0,134$ Der Abstand zwischen den beiden Sequenzen beträgt 0,134. damit kann man sagen, dass die Sequnezen aus Beispiel 2 sich ähnicher sind als die Sequnezen aus Beispiel 1. Sie sind sich evolutionär ähnlicher und genetisch näher aneinander verwandt.

Validation des Ergebnis:

Beim Jukes-Cantor-Modell wird die Wahrscheinlichkeit des Mutationstyps der Transversion und Transition gleichgesetzt. Diese Annahme ist realitätsfern, denn Studien längst gezeigt, dass die Transition häufiger eintritt als die Transversion und somit eine höhere Mutationsrate hat. Diese Tatsache wird im nächsten Modell vorgestellt. [8]

4.2 Kimura-Modell

Ein weiteres mathematisches Modell ist das Kimura-Modell. Es wurde 1980 von Motoo Kimura entwickelt. Dabei gibt es verschiedene Komplexitätsstufen, z.B. entweder mit 2 oder mit 3 Parametern. Die Parameter stehen hier für die Mutationsraten der bereits genannten Arten an Mutationen(siehe 2.3). Denn im Gegenteil zum Jukes-Cantor-Modell wird beim Kimura-2-Parameter-Modell bei der Abstandsbestimmung Trasitionen und Transversionen differenziert. Grund dafür ist der Beweis, dass aufgrund der chemischen Struktur der Basen Transitionen in der Evolution schneller und häufiger auftreten als Transversionen.

Abstandsbegegriff des Kimura-Modells

[8] vgl. Sube, Frank und Walcher[3](S.87-95) und Haubold[1](S.26-28)

Abstand zwischen S_1 und $S_2 = 2\cdot$ erwartete Anzahl der Mutationen während der Entwicklung von S_1 aus $S_0 = (\alpha + 2\beta) \cdot 2 \cdot t \approx \frac{-1}{4} ln((1-2P-2Q)^2(1-4Q) = -\frac{1}{2}((1-2P-2Q)\sqrt{1-4Q})$

P ist die Häufigkeit, dass eine Transition zu sehen ist. Q sei die Häufigkeit, dass eine Transversion. Die erste Mutationsrate wird α genannt für die Transition, die zweite β für die Transversion.

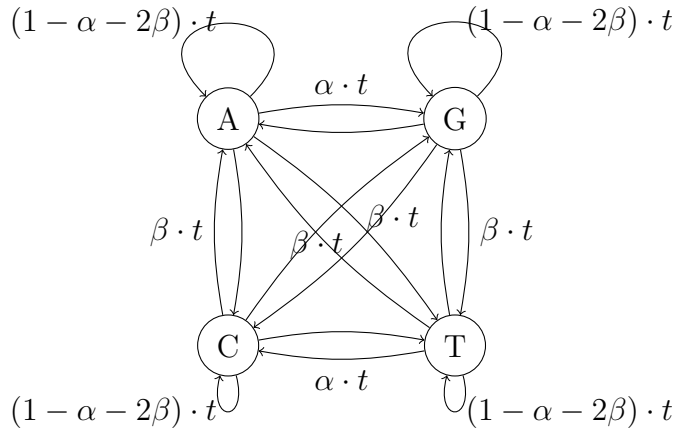


Abbildung 7: Graph zum Kimura-2-Parameter-Modell

Beispiel 1:
Für einen Vergleich der Modelle kann man auch das Beispiel 2 auf seite 11 von den Jukes-Cantor-Modell mit dem Abstandsbegriff des Kimura-2-Parameter-Modell berechnen. Die Länge der zu vergleichenden Sequenzen beträgt 300. Die Anzahl an unterschiedlichen Stellen zwischen den Sequnezen war 37. Die restlichen 263 Nukleotide sind identisch. Es seien davon 20 Transition und 17 Transversionen. Eingesetzt in der Formel beträgt der Abstand 0,205. Es ermöglicht eben eine realitätsnäheren Abstandsbegriff durch die Annahme, dass es meherer unterschiedliche Mutationsarten gibt. [9]

5 Angleichung mithilfe von Algorithmen

Der Angleichungsalgorithmus wird verwendet, um Überbestimmungsverfahren zu lösen. Hierbei gibt es nicht immer eine eindeutige Lösung, da ein solches System aus mehreren Gleichungen als Unbekannte besteht.
In den vorherigen mathematischen Modellierungen wurde der Abstand zwischen zwei DNA-Sequenzen lediglich in Betracht auf die Punktmutation ermittelt. Durch Algorithmen lassen sich jedoch die anderen Mutationsarten, Deletion und Insertion, in die Mathematik übertragen. Doch dadurch verändern

[9] vgl. Sube, Frank und Walcher[3](S.103-105)

sich ebenso die zuvor getroffenen Annahmen. Die Sequenzen sind nun nicht mehr zwingend gleich lang.
Zunächst werden zwei Sequenzen betrachtet: GACTTA und ACT. Erstmals gibt es keine Überscheidungen der Stellen, doch durch Verschiebung der zweiten Sequenz kann man erkennen, dass diese in der ersten wiederzufinden ist.

GACTTA
-ACT- -

Hier erkennt man die Mutationsarten Insertion und Deletion. In der oberen Sequenz hat eine Insertion an den Stellen 1,4 und 5 stattgefunden, während in der unteren eine Deletion an denselben Stellen stattgefunden hat.
In diesem Verfahren wird nun weiter vorgegangen um die Ähnlichkeit der Sequenzen zu ermitteln. Bei gleichlangen Abschnitten werden Lücken eingefügt, um die Ähnlichkeit aufzuweisen. Jedoch darf auch keine künstliche Vergrößerung des Abstands geschehen und die zuvorige Monotonieannahme muss berücksichtigt werden.

$$\begin{aligned} \text{GACCTAG} &\Rightarrow \text{GACCT-AG} \\ \text{GCCTGAG} &\Rightarrow \text{G-CCTGAG} \end{aligned}$$

Anschließend werden die DNA-Sequenzen mathematisch erfasst. Wie zuvor wird die Länge zweier zu vergleichenden Sequenzen S_1 und S_2 der Menge S, die aus den Datensätzen gewonnen wird, mit n_1, n_2 bezeichnet. Da die Abschnitte zuvor mit Lücken angeglichen worden sind, werden sie nun mit S_{1a} und S_{2a} bezeichnet und sind Teilmengen der Menge S_a mit den zugehörigen Längen n_{1a} und n_{2a}. Dabei ist eine Lücke an der gleichen Stelle beider Sequenzen ausgeschlossen.

5.1 Needlemann-Wunsch-Algorithmus

Bei diesem Verfahren erstellt man zunächst eine Matrix aller möglichen Paare der Menge S_a. Diese Matrix wird *Dotplotmatrix* genannt. Hier wird in der ersten Zeile S_1 notiert und in der ersten Spalte S_2. Zusätzlich wird eine *scoring-Tabelle* angelegt, in der die Bewertung für die Übereinstimmungen und Unterschiede festgelegt werden. Falls durch die Mathematisierung fälschlicherweise an der gleichen Stelle Lücken auftreten, wird eine sogenannte *Lückenstrafe* eingeführt und mit g bezeichnet, dabei gilt $g > 0$. Diese wird ebenfalls in der scoring-Tabelle notiert. Positive Werte zeigen hier die Übereinstimmungen und Lücken an, während negative Zahlen die Unterschiede darstellen.[10]
Im Folgenden werden die Sequenzen S_1 GTACCGA und S_2 ATACCTG verglichen. Die entstehende Matrix wird F genannt und $i = 1, ..., 7$ der Sequenz

[10] vgl. Sube, Frank und Walcher[3](S.115-122)

1 und $j = 1, ..., 7$ der Sequenz 2.

Für ein Paar, also eine Übereinstimmung (*engl. match*), wird der Wert +1 zu-

scoring-Tabelle
Übereinstimmung: +1
Unterschied: -1
Lücke: 1

	-	G	T	A	C	C	G	A
-								
A								
T								
A								
C								
C								
T								
G								

Abbildung 8: Dotplotmatrix F

geordnet. Bei Unterschieden (*engl. mismatch*) wird -1 notiert. Auch die Lücke g erhält den Bewertung 1. Nun wird der Algorithmus folgendermaßen angewendet: Im ersten Schritt werden die Abstände zum Ausgangspunkt 0 eingetragen.

Im zweiten Schritt wird folgendermaßen vorgegangen: $F_{i,j} = max \begin{Bmatrix} F_{i-1,j-1} + M_{i,j}, \\ F_{i-1,j-1} - g, \\ F_{i-1,j-1} - g \end{Bmatrix}$

scoring-Tabelle
Übereinstimmung: +1
Unterschied: -1
Lücke: 1

	-	G	T	A	C	C	G	A
-	0	← -1	-2	-3	-4	-5	-6	-7
A	↑-1							
T	-2							
A	-3							
C	-4							
C	-5							
T	-6							
G	-7							

Abbildung 9: Dotplotmatrix F, erster Schritt

Es wird in drei Richtungen ermittelt: $F(\nwarrow), F(\leftarrow) und F(\uparrow)$. Der erste einzutragende Wert ist -1, da $F_{1,1} = max\{0 + (-1), -1 - (1), -1 - (1)\} = -1$ Da an der Stelle 1 beider Sequenzen keine Übereinstimmung ist, wird -1 für die erste Gleichung genommen. Bei den Gleichung $F(\leftarrow) und F(\uparrow)$ wird 1 für g eingesetzt.

In diesem Schema wird nun weiterhin verfahren, bis die Matrix gefüllt ist

Die Bildung der Pfeile erfolgt analog zum Needleman-Wunsch-Algorithmus. Beginnend am höchsten Score (hier: 3) werden durch den diagonalen Weg die vier Basen gepaart:

T	A	C	C
⇳	⇳	⇳	⇳
T	A	C	C

	-	G	T	A	C	C	G	A
-	0	← -1	-2	-3	-4	-5	-6	-7
A	↑-1	←-1	0	-1	-2	-3	-4	-5
T	-2	-2	↖0	-1	-2	-3	-4	-5
A	-3	-3	-1	↖←1	0	-1	-2	-3
C	-4	-4	-2	0	↖2	1	0	-1
C	-5	-5	-3	-1	1	↖3	2	1
T	-6	-6	-2	-2	0	2	↖2	1
G	-7	-5	-3	-3	-1	1	3	←2

Abbildung 10: Dotplotmatrix F

Auswertung: Es gibt ein mögliches Alignment mit einem Score von 3 bezüglich der gewählten Bewertung.[11]

5.2 Edit-Abstand

Der Edit-Abstand, auch Levenshtein-Distanz genannt, wird grundsätzlich in der Bioinformatik für den Vergleich zweier Strings verwendet. Übertragen auf den biologischen Kontext dieser Facharbeit stellen die zu vergleichenden Sequenzen die Strings dar. Der Edit-Abstand zum Vergleich zweier DNA-Sequenzen stellt die minimale Anzahl an den drei möglichen Operationen, um die eine Sequenzen in die andere zu überführen, dar. Diese Operationen sind in diesem Fall die Mutationstypen, wobei das Ersetzen die Substitution oder Punktmutation, das Löschen Deletion und das Einfugen Insertion bedeuten. Die Sequenzen müssen allerding gleich lang sein.

Abstandbegriff: Nachdem man die beiden Sequenzen aligniert hat, kann man mithilfe der Levensthein-Distanz den Abstand beider messen. Die Festlegung der Kosten kann man mit der Biologie begründen. Eine Substitution hat einen kleineren Kosten als eine Deletion oder Insertion. Damit gilt $0 < w_S \leq w_L$, wobei bei identischen Basen keine Kosten entstehen.[12]

Edit-Abstand/Levenshtein-Distanz

Gegeben seien zwei DNA-Sequenzen $S_1, S_2 \in S$ mit Längen n_1, n_2. Der Abstand zwischen diesen beiden Sequenzen ist definiert durch

$$d(S_1, S_2 = min\{d(S_{1,n_1-1}, S_{2,n_2-1}) + w_S(S_{1,n_1-1}, S_{2,n_2-1}),$$
$$d(S_{1,n_1-1}, S_{2,n_2}) + w_L, d(S_{1,n_1}, S_{2,n_2-1}) + w_L\}$$

mit w_S als Gewichtung der durchgeführten Substitution
und w_L als Gewichtung der durchgeführten Deletion/Insertion.[13]

Der Abstand ist also die Summe aller Kosten pro Stelle.

[11] vgl. Sube, Frank und Walcher[3](S.119-122) und Likic[2](S.12-18)
[12] vgl. Sube, Frank und Walcher[3](S.141-146)

Beispiel:
gegeben sind die zwei Sequnezen: S1: TGGACACTA und S2: Für die Vereinfachung wurden die Kosten aller möglichen Mutationen , d.h.w_S und w_S gleich 1 gesetzt. Es kam dabei folgende Matrix raus, an der man an der letzten Spalte und letzten Zeile die minimale Anzahl an Operationen erkennt, die nötig wären, um die Sequenz S1 in S2 zu überführen. $d(S_1, S_2$ beträgt in diesem Beispel 3.

		G	C	T	A	T	A	C
	0	1	2	3	4	5	6	7
G	1	0	1	2	3	4	5	6
C	2	1	0	1	2	3	4	5
G	3	2	1	1	2	3	4	5
T	4	3	2	1	2	3	4	5
A	5	3	3	2	1	2	3	4
T	6	4	4	3	2	1	2	3
G	7	5	5	4	3	2	2	3
C	8	6	5	5	4	3	3	2

Abbildung 11: Matrix zur Berechnung des Edit-Abstandes

6 Fazit

Anhand dieser Arbeit wird deutlich, dass die Mathematik die Forschung in der Biologie erheblich vereinfacht und durch sie neue sich neue Erkenntnisse erfassen lassen. Jede angewandte Methode hat einen anderen Schwerpunkt, und somit kann man je nach Spezifizierung der Fragestellung die passende Methode in Betracht ziehen. Letztendlich gelangt man zur Feststellung, dass jeder gemessene Abstand anhand von Annahmen und Vereinfachungen kalkuliert wurde, sodass das Ergebnis jeder mathematischen Modellierung rein hypothetisch betrachtet werden darf. Auch variieren die gemessenen Abstände, da für jede Technik ein anderer Abstandsbegriff definiert wird. Die Einsatzmöglichkeiten der in dieser Seminarfacharbeit erwiesen Methoden sind breitgefächert und zeigen den Abstand zwischen zwei DNA-Sequenzen auf.
Die biomathematische Messung des evolutionären Abstandes findet praktische Anwendung in der Medizin, besonders in der Krebsforschung, wo die Bestimmung des evolutionären Abstands zwischen Tumorzellen zum Verständnis der Entstehung und Entwicklung der Krankheit essenziell ist, um personalisierte Therapieansätze zu entwickeln. Die Techniken werden auch benutzt, um die genetische Verwandtschaft zwischen verschiedenen Stämmen von Krankheitserregern zu ermitteln, sodass man mehr über die Ausbreitung, Bekämpfung und Prävention dieser Infektionskrankheiten erfahren kann.
Zusammengefasst lässt sich die genetische Verwandschaft daher sinnvoll durch

verschiedenste Modelle mathematisieren.

Literatur

[1] Bernhard Haubold. “MBI: Sequenzvergleich ohne Alignment”. In: (2013).

[2] Vladimir Likic. “The Needleman-Wunsch algorithm for sequence alignment”. In: *Lecture given at the 7th Melbourne Bioinformatics Course, Bi021 Molecular Science and Biotechnology Institute, University of Melbourne* (2008), pp. 1–46.

[3] Maike Sube, Martin Frank, and Sebastian Walcher. *Entwicklung und Evaluation von Unterrichtsmaterial zu Data Science und mathematischer Modellierung mit Schülerinnen und Schülern*. Tech. rep. Lehr-und Forschungsgebiet Mathematik, 2019.

https://docplayer.org/22936311-Phylogenetische-baeume.html (S.20)

Abbildungsverzeichnis

Versicherung zur selbstständigen Arbeit

Hiermit versichere ich, dass ich die Arbeit selbstständig angefertigt, keine anderen als die angegebenen Hilfs- mittel benutzt und die Stellen der Facharbeit, die im Wortlaut oder im Wesentlichen Inhalt aus anderen Werken entnommen wurden, mit genauer Quellenangabe kenntlich gemacht habe. [Verwendete Informationen aus dem Internet sind dem(r) Lehrer/in vollständig ausgedruckt in einem sepa- raten Ordner zur Verfügung gestellt worden.] je nach Absprache

Leer, den 27.03.2023
(Ort, Datum)

Sidra Omar

Veröffentlichungserklärung

Hiermit erkläre ich, dass ich damit einverstanden bin, wenn die von mir verfasste Facharbeit der schulinternen Öffentlichkeit (Bsp.: Schülerbibliothek, IServ) zugänglich gemacht wird.

Leer, den 27.03.2023
(Ort, Datum)

Sidra Omar