UNIVERSITÄT ZU LÜBECK
INSTITUT FÜR
NEURO- UND BIOINFORMATIK

Der Effekt von nicht zufälligen Verteilungen in codierenden Daten auf die Mutationsstabilität des genetischen Codes

The effect of non-random pattern distributions in coding data on mutation stability of the genetic code

Bachelorarbeit

im Rahmen des Studiengangs
Medizinische Informatik
der Universität zu Lübeck

vorgelegt von
Berit Klaucke

ausgegeben und betreut von
PD Dr. rer. nat. Amir Madany Mamlouk

Lübeck, den 11. Februar 2018

Im Focus das Leben

Erklärung

Ich versichere an Eides statt, die vorliegende Arbeit selbstständig und nur unter Benutzung der angegebenen Hilfsmittel angefertigt zu haben.

Lübeck, den 11. Februar 2018

Kurzfassung

Die Eigenschaft des genetischen Codes, durch Mutation bedingte Polaritätsänderung gering zu halten, wurde in vorangegangenen Publikationen durch statistische und biochemische Untersuchungen gezeigt. Daraus konnte die Schlussfolgerung gezogen werden, dass der genetische Code optimiert ist, um den Einfluss von Translationsfehlern, Punkt- und Leserastermutationen zu minimieren. Nukleotide und Tripletts kommen in natürlichen Sequenzen nicht gleich verteilt vor. Miteinbezogen wurde dies bisher jedoch nur mit Daten aus dem ersten, humanen Chromosom. Dabei wurde nicht beachtet, dass nur ein sehr kleiner Teil der Sequenzen codierend ist. Diese Arbeit erweitert die Analysen auf das gesamte humane Genom und differenziert zwischen Daten aus codierenden und kompletten Sequenzen. Es wird gezeigt, dass sich codierende Verteilungen negativ auf die polaritätskonservierende Eigenschaft des genetischen Codes auswirken. Positive Auswirkungen zeigen hingegen die Verteilungen der kompletten Sequenzen. Insgesamt bleibt der genetische Code auch mit Beachtung der codierenden a priori Wahrscheinlichkeiten zu 99,99% effektiver als zufällige Codes. Allerdings weisen die Ergebnisse darauf hin, dass der genetische Code nicht dem Code entspricht, der die Auswirkung von Mutationen am stärksten minimiert.

Abstract

Previous Publications have used statistical and biochemical analyses to show that the genetic code is able to minimize the change of polar requirement caused by mutations. It was deduced that the genetic code is optimised to abate the consequences of mistranslation, point and frameshift mutations. The fact that in natural sequences nucleotides and triplets are not evenly distributed, was only shown with data from the first human chromosome analysis. It was not, however, considered that only a minute part of the DNA is encoding. This thesis thus expands the analyses to the complete genome and differentiates the data from coding material and that complete sequences. It is hereby shown, that the distribution in the coding sections has a negative effect on the genetic code's ability to conserve polarity, while in the distribution of complete sequences a positive effect on the same is found. In conclusion, the genetic code, even with the coding data, remains 99.99% more effective than random codes. The results also imply that the genetic code might not be the code that minimises the consequences of mutation the most.

Inhaltsverzeichnis

1 Einleitung

Der genetische Code ist kein Zufallsprodukt. Er ist optimiert, um Fehler durch Mutationen zu minimieren. Zu diesem Ergebnis kamen mehrere vorangegangene wissenschaftliche Arbeiten [5, 8–10]. Haig und Hurst fanden 1991 in einem Set aus 10.000 zufälligen Codes nur zwei Codes, die Punktmutationen besser minimieren können als der natürliche Code [9]. Die Analyse dahinter baut darauf auf, dass die Polarität der einzelnen Aminosäuren bei der Proteinfaltung eine entscheidende Rolle spielt. Da die Umgebung der Proteine, das Cytosol, zu 70% aus Wasser besteht, zeigt sich bei der Proteinfaltung die Tendenz der hydrophoben Moleküle, zu denen die nicht-polaren Aminosäuren gehören, sich im Inneren der Proteins zu clustern. Schon kleine Änderungen der Polarität weniger Aminosäuren können so deutliche Auswirkungen auf die Faltung des Proteins haben (Alexander et al., 2007) [4]. Mutationen, die zu starken Änderungen der Polarität der Aminosäuren führen, haben somit starke Auswirkung auf die Faltung des Proteins und damit auch auf dessen Funktion. Haig und Hurst (1991) [9] konnten feststellen, dass der genetische Code die Eigenschaft aufweist, Änderungen der Polarität bei Punktmutationen zu minimieren. 1998 erweiterten Freeland und Hurst die Arbeit von Haig und Hurst, indem sie Translationsfehler und einen Transitions/ Transversions Bias miteinbezogen. Damit konnten sie zeigen, dass es sogar in einem Set von 1.000.000 zufälligen Codes nur einen besseren Code gibt. 2017 zeigten Geyer und Madany [8], dass der genetische Code auch die Auswirkung von Leserasterverschiebungen minimiert. Im Verhältnis zu den zufälligen Codes minimiert der Genetische Code die Auswirkungen der kombinierten Mutationen sogar noch besser. Diese drei Arbeiten gehen alle von der impliziten Annahme aus, dass Nukleotide und Codons mit gleicher Wahrscheinlichkeit auftreten. Auf natürlichen Sequenzen ist aber keine Gleichverteilung der Nukleotide und Codons gegeben. Um die Untersuchungen in ein natürlicheres Szenario zu setzen, wurde die Arbeit von Keser (2016) [10] um die Gewichtung der a priori Wahrscheinlichkeiten erweitert. Keser [10] konnte damit das bisherige Ergebnis sogar noch einmal verstärken. Bisher zeigte sich also die Tendenz, dass, je näher das Analysemodell an natürliche Gegebenheiten angepasst wird, der genetische Code stärkere Minimierungstendenzen zeigt. Es ist verlockend, daraus zu schließen, dass der genetische Code nicht einfach nur 'ein' Code ist, welcher Mutationsfehler minimiert, sondern, dass er von allen 'der beste' Code ist. Diese Arbeit möchte die Analysen vorangegangener Arbeiten nun noch weiter an ein reales Szenario anpassen, indem bei der Gewichtung natürlicher a priori Wahrscheinlichkeiten zwischen kompletten und codierenden Sequenzen differenziert wird.

2 Methoden

2.1 Polaritätsänderungen durch Punktmutationen

Um die Auswirkungen von Punktmutationen zu quantifizieren, führten Haig und Hurst (1991) [9] ein Maß der mittleren quadratischen Abweichung ein. Im Weiteren soll dies in Konsistenz mit den vorangegangen Arbeiten, auf welchen diese hier aufbaut, mit MS (mean square) abgekürzt werden.
Sei $P(c_i)$ der Polaritätswert des Codons c_i und sei $M^j(c_i)$ die $j$te Mutatation des Codons c_i, folglich sei $P(M^j(c_i))$ der Polaritätswert der $j$te Mutatation des Codons c_i. Weiter sei m_i die Anzahl der möglichen Mutationen M des Codons c_i. Da die 3 Stopcodons nicht mitberechnet werden, gibt es 61 Codons, die mutiert werden können. Die Polaritätswerte der einzelnen Aminosäuren werden aus chemischen Untersuchungen von Woese et al. (1966) [13] übernommen. Die genauen Werte können in der Tabelle 2.1 abgelesen werden. Dann ist die quadratische Abweichung für eine Mutation M definiert durch:

$$D_M := \sum_{i=1}^{61} \sum_{j=1}^{m_i} (P(c_i) - P(M^j(c_i)))^2.$$

Der Skalierungsfaktor F ist definiert als die Anzahl möglicher Mutationen über alle Codons $F := \sum_{i=1}^{61} m_i$. Bei Punktmutationen sind die Skalierungsfaktoren jeweils $F1 = 174, F2 = F3 = 176$ an der ersten, zweiten und dritten Stelle, da es jeweils $61 * 3 = 183$ mögliche Mutationen gibt, von denen die entstehenden Stopcodons abgezogen werden müssen. Die mittlere quadratische Abweichung ist dann an erster, zweiter und dritter Stelle MS1, MS2, MS3:

$$MS1 = \frac{D_1}{F_1}, \quad MS2 = \frac{D_2}{F_2}, \quad MS3 = \frac{D_3}{F_3}.$$

Über alle drei Positionen ist die mittlere quadratische Abweichung MS0 definiert als:

$$MS0 = \frac{D_1 + D_2 + D_3}{F_1 + F_2 + F_3}.$$

2.2 Polaritätsänderungen mit Transition/Transversion Bias

Verschiedene Mutationstypen treten unterschiedlich häufig auf. Bekannt ist, dass Transitionen häufiger vorkommen, als Transversionen. Dabei wird von einer Transition gesprochen, wenn eine Pyrimidinbase (Uracil, Thymin und Cytosin) mit einer anderen

Aminosäure	Abk.	PW	Aminosäure	Abk.	PW	Aminosäure	Abk.	PW
Alanin	Ala	7.0	Glycin	Gly	7.9	Prolin	Pro	6.6
Arginin	Arg	9.1	Histidin	His	8.4	Serin	Ser	7.5
Asparaginsäure	Asp	13.0	Isoleucin	Ile	4.9	Threonin	Thr	6.6
Asparagin	Asn	10.0	Leucin	Leu	4.9	Tryptophan	Trp	5.2
Cystein	Cys	4.8	Lysin	Lys	10.1	Tyrosin	Tyr	5.4
Glutaminsäure	Glu	12.5	Methionin	Met	5.3	Valin	Val	5.6
Glutamin	Gln	8.6	Phenylalanin	Phe	5.0			

Tabelle 2.1: Polaritätswerte (PW) der proteinogenen Aminosäuren gemessen von Woese et al.(1966) [13]

Pyrimidinbase ersetzt wird, beziehungsweise wenn eine Purinbase (Guanin und Adenin) gegen eine andere Purinbase ausgetauscht wird. Bei Transversionen werden Purinbasen durch Pyrimidinbasen ersetzt, oder umgekehrt [6]. Um das Verhältnis von Transitionen zu Transversionen abzubilden, definierten Freeland und Hurst (1998) [5] eine weitere Formel zur Berechnung. Diese Formel teilt D auf Transitionen S und Transversionen V auf:

$$S_M := \sum_{i=1}^{61} \sum_{j=1}^{m_i^s} (P(c_i) - P(M^{j,s}(c_i)))^2, \quad V_M := \sum_{i=1}^{61} \sum_{j=1}^{m_i^v} (P(c_i) - P(M^{j,v}(c_i)))^2.$$

Dabei sei S_M die quadratische Abweichung aller Transitionen und V_M die quadratische Abweichung aller Transversionen. Transitionen können so anders gewichtet werden als Transversionen. Die gewichtete quadratische Abweichung aller Transitionen S_M und Transversionen V_M ist dann:

$$W^w := wS_M + V_M.$$

Der Skalierungsfaktor F muss ebenfalls separat berechnet werden.

$$F^w := w \sum_{i=1}^{61} m_i^s + \sum_{i=1}^{61} m_i^v$$

Die gewichtete mittlere quadratische Abweichung ist dann an erster, zweiter und dritter Stelle WMS1, WMS2, WMS3:

$$WMS1 = \frac{W_1^w}{F_1^w}, \quad WMS2 = \frac{W_2^w}{F_2^w}, \quad WMS3 = \frac{W_3^w}{F_3^w}.$$

Über alle drei Positionen ist die mittlere quadratische Abweichung WMS0 definiert als:

$$WMS0 = \frac{W_1^w + W_2^w + W_3^w}{F_1^w + F_2^w + F_3^w}.$$

2.3 Polaritätsänderungen durch Leserastermutation

Leserastermutationen sind Leserasterverschiebungen, die durch Insertion oder Deletion eines oder mehrerer Nukleotide entstehen. Dadurch wird das Leseraster nach rechts (+1) beziehungsweise nach links (-1) verschoben. Da durch Leserasterverschiebungen nicht nur ein Triplett falsch translatiert wird, sondern alle auf auf den Mutationspunkt folgenden, haben sie besonders schwerwiegende Auswirkungen auf das resultierende Protein. Geyer und Madany [8] definierten den MS Score auch für Leserastermutationen:

$$rMS = \frac{D_r}{F_r}, \quad lMS = \frac{D_l}{F_l}, \quad fMS = \frac{D_r + D_l}{F_r + F_l}.$$

Wobei D_r, D_l analog zu D in Abschnitt 2.1 definiert sind. Der Skalierungsfaktor entspricht wieder der Anzahl möglicher Mutationen über alle Codons $F_r = F_l = 232$ (= 61 Tripletts*4 Nukleotide -12 Stopcodons)

2.4 Polaritätsänderungen mit a priori Wahrscheinlichkeiten Gewichtung

Die vier verschiedenen Nukleotide und die 64 Tripletts kommen nicht mit gleicher Wahrscheinlichkeit in DNA Sequenzen vor. Lediglich analog der Häufigkeit, in der ein Triplett vorkommt, kann auch maximal eine Mutation dieses Tripletts vorkommen. Das Gleiche gilt auch für Nukleotide. Somit haben natürliche Verteilungen eine direkte Relevanz bei der Untersuchung der Effizienz des genetischen Codes. Keser (2016) [10] erweiterte die Berechnungen um die Gewichtung mit a priori Wahrscheinlichkeiten. Dabei entspricht die Gewichtung G jeweils den skalierten a priori Wahrscheinlichkeiten. Da G skaliert wird, muss der Skalierungsfaktor F nicht angepasst werden. Der Gewichtungsfaktor G kann an alle hier vorgestellten quadratischen Abweichungen appliziert werden. Die Berechnung der quadratischen Abweichung bei Punktmutationen erfolgt dann folgendermaßen:

$$D_M := \sum_{i=1}^{61} \sum_{j=1}^{m_i} (P((c_i) - P(M^j(c_i)))^2 * \mathbf{G}).$$

Alle anderen quadratischen Abweichung werden äquivalent zu der oben gezeigten, durch Multiplikation der einzelnen quadratischen Abweichungen mit dem Gewichtungsfaktor gewichtet.

Durch die Triplett a priori Gewichtung (TA) wird erreicht, dass Polaritätsdifferenzen durch Mutationen von Tripletts, die in natürlichen Sequenzen selten vorkommen, weniger stark in den MS Wert eingehen, als solche von Codons, die besonders häufig vorkommen. Ebenso werden bei Punktmutationen durch Nukleotid a priori Gewichtung (NA) die einzelnen Mutationsnukleotide entsprechend ihres natürlichen Vorkommens gewichtet. Beide a priori Gewichtungen können durch Multiplikation kombiniert werden. Auch

bei Leserastermutationen wird die Häufigkeit, mit der die einzelnen Polaritätsdifferenzen vorkommen, realistisch abgebildet, indem TA und NA beachtet werden. Gemeint sind damit die TA der ursprünglichen Codons, gemeinsam mit der NA des Nukleotids, welches neu in das leserasterverschobene Triplett kommt. Doch bei Leserasterverschiebungen ist zusätzlich zu beachten, dass die NA und TA nicht unabhängig voneinander sind. Zum Beispiel kann die NA für „A“ nach dem Triplett „ATG“ anders sein, als nach dem Triplett „GCT“. Das gleiche gilt auch für Nukleotide, die dem Triplett voranstehen. Um Leserasterverschiebungen möglichst realitätsnah zu simulieren, wird bei der a priori gewichteten rMS, lMS, fMS Berechnung die aus den natürlichen Sequenzen entnommene „gemeinsame“ a priori Wahrscheinlichkeit, also die a priori Wahrscheinlichkeit eines Triplett mit direkter Nukleotid Nachbarschaft appliziert. Diese a priori Wahrscheinlichkeit wurde von Keser als „triplet transition“ a priori Wahrscheinlichkeit bezeichnet und daher mit TT abgekürzt. Um Konsistenz zu erhalten, wird die Abkürzung hier beibehalten, mit dem Hinweis darauf, dass mit „transition“ hier die englische Bezeichnung für Übergang gemeint ist und nicht der Mutationstyp einer Transition. Die TT Gewichtung wird verwendet, um den Verschiebung des Rasters vom ursprünglichen Codon zum leserasterverschobenen Codon zu gewichten und enthält bereits TA, NA und das Abhängigkeitsverhältnis der beiden Gewichtungen zueinander, wobei zur genaueren Analyse des Effekts die Gewichtungen zunächst auch einzeln betrachtet werden.
Alle drei Gewichtungen können kombiniert werden. Eine solche Kombination simuliert ein natürliches Szenario am Besten. Zur genaueren Analyse des Effekts werden die Gewichtungen aber zunächst auch einzeln betrachtet.

Keser untersuchte die Verteilungen der kompletten DNA Sequenzen, jedoch werden von diesen nur ca. 2% translatiert und somit vom genetischen Code codiert . Bevor die Translation stattfindet, werden große Teile der Sequenz durch Splicing entfernt. Die komplette DNA Sequenz enthält sogenannte Exons und Introns. Nach dem Splicing sind nur noch Exons in den Sequenzen enthalten. Exons bestehen wiederum aus untranslatierten Regionen und codierenden Sequenzen (CDS). Die CDS machen nur einen sehr geringen Teil der kompletten Sequenzen aus. Für die a priori gewichtete Berechnung der mutationsbedingten Polaritätsabweichung, welche die Optimierung des genetischen Codes beurteilen soll, erscheint es sinnvoll, die a priori Wahrscheinlichkeiten nur aus dem codierenden Teil der Sequenzen zu erheben. Daher untersucht diese Arbeit, wie sich a priori Wahrscheinlichkeiten speziell der CDS auf die Polaritätsberechnungen auswirken.

Die hier verwendeten Daten der kompletten DNA Sequenzen wurden von der NCBI (National Center for Biotechnology Information) Nukleotid Datenbank [11] zu Verfügung gestellt. Die Daten zu den codierenden Sequenzen stammen aus der vom Concensus CDS [12] bereitgestellten Datenbank. Das Concensus CDS Projekt ist die Zusammenarbeit mehrerer Gendatenbanken, welche gemeinsam die CDS zusammentragen, überprüfen und allgemein zu Verfügung stellen, um möglichst komplette und valide Daten zu liefern. Die Berechnung der TA und TT erfolgte im Leseraster. Daher wurden nur Sequenzen eingearbeitet, die mit einem Startcodon beginnen und mit einem Stopcodon enden. Alternatives Splicing wurde nicht beachtet.

Keser untersuchte die Anzahl effizienterer Codes nur bei Gewichtung des menschlichen Chromosom eins. Dabei drängt sich die Frage auf, ob die Ergebnisse für das gesamte menschliche Genom übertragbar sind. Daher wurden in dieser Arbeit die a priori Wahrscheinlichkeiten aus Chromosom 1 bis 22, sowie Chromosom X und Y erhoben. Zudem wurde untersucht, ob die a priori Wahrscheinlichkeiten sich auf allen Chromosomen gleich verhalten, oder ob jedes Chromosom individuell auf die Optimierung der Polaritätserhaltung des genetischen Codes evaluiert werden muss. Ebenso wurde die Untersuchung der Anzahl effizienterer Codes in dieser Arbeit mit NA, TA, TT des gesamten humanen Genoms durchgeführt.

2.5 A priori Wahrscheinlichkeiten und Stopcodon

Mit den a priori Wahrscheinlichkeiten ist eine Untersuchung der Nonsense Mutationen möglich. Als Nonsense Mutationen werden Mutationen bezeichnet, bei denen Stopcodons entstehen. Diese haben besonders starke Auswirkungen auf das Protein, da nicht nur an der mutierten Stelle keine Aminosäure translatiert wird, sondern die Translation abgebrochen wird. Alle Codons nach dem Mutationspunkt werden folglich nicht translatiert.

Die einzelnen Codons c_i, die durch Mutation zu Stopcodons führen, werden mit der Häufigkeit ihres Auftretens in natürlichen Sequenzen gewichtet und ins Verhältnis zu allen möglichen Mutationen m gesetzt. Die Anzahl möglicher Leserastermutationen ist $m = 61 * 4 = 244$ und die Anzahl möglicher Punktmutationen ist $m = 61 * 3 = 183$. N_M entspricht dem Verhältnis von Nonsense Mutationen zu allen allen möglichen Mutation.

$$N_M := \frac{\sum_{i=1}^{m_i} c_i * G}{m}.$$

Die Anzahl i möglicher Nonsense Mutationen durch Punktmutationen beträgt $i = 9$ an der ersten Position und $i = 7$ an zweiter und dritter Stelle. Durch Leserasterverschiebungen sind in beide Richtungen jeweils $i = 12$ verschiedene Nonsense Mutationen möglich.

2.6 Kombinierter Score

Die kombinierte mittlere quadratische Abweichung wird in dieser Arbeit definiert wie bei Keser (2016) [10]:

$$GMS = \frac{W_1 + W_2 + W_3 + D_r + D_l}{F_1^w + F_2^w + F_3^w + F_r + F_l}.$$

Wobei die WMS Scores mit einem Bias $w = 1$ den MS Scores entsprechen. Der GMS Score fasst zusammen, wie effektiv ein Code alle Mutationstypen minimieren kann, wobei

ein möglichst kleiner Wert einer geringen Veränderung der Polarität entspricht. Daraus folgt, dass der entsprechende Code gegenüber den berücksichtigten Mutationstypen optimiert ist.

2.7 Generierung des zufälligen Codesets

Der natürliche Code wird mit 1.000.000 Millionen zufälligen Codes hinsichtlich des MS Scores verglichen. Zur besseren Vergleichbarkeit, wird in dieser Arbeit das gleiche Codeset verwendet mit dem Geyer [7] und Keser [10] ihre Untersuchungen durchführten. Alle zufälligen Codes wurden so generiert, dass das Redundanzlevel des natürlichen Codes erhalten bleibt. Dafür wurden die Codonsets entsprechend dem natürlichen Code beibehalten und die Zuordnung der Aminosäuren zu den Codonsets permutiert. Mit dieser Methode können 20! ($\geqslant 2,4 \times 10^{18}$) verschiedene Codes generiert werden. Im Verhältnis zu den möglichen Codes ist das Set von eine Millionen Codes eine verhältnismäßig kleine Stichprobe. Es ist daher möglich, dass „bessere“ Codes übersehen wurden.

Alle in dieser Arbeit durchgeführten Berechnungen wurden mit Hilfe von MatLab und dem von Keser(2016) [10] zu Verfügung gestellten Java Framework durchgeführt. Das Java Framework wurde zu diesem Zweck überarbeitet und weiterentwickelt.

3 Ergebnisse

3.1 A priori Wahrscheinlichkeiten in humanen DNA Sequenzen

Zunächst wurden die a priori Wahrscheinlichkeiten in den kompletten Sequenzen der Chromosomen 1 bis 22, sowie für Chromosom X und Y, einzeln ermittelt.

Von Freeland und Hurst (1998) [5] werden implizit indifferente a priori Wahrscheinlichkeiten vorausgesetzt. Zur Betrachtung der Veränderung, welche die Berücksichtigung der natürlichen a priori Wahrscheinlichkeiten bewirkt, wird die mittlere Abweichung der indifferenten zu den natürlichen Wahrscheinlichkeiten berechnet. Sowohl für komplette Sequenzen, als auch für codierende Sequenzen wird dies mit NA und TA Wahrscheinlichkeiten gezeigt. Tabelle 3.1 zeigt zusammenfassend die jeweils mittlere Abweichung zur Gleichverteilung. Sowohl die a priori Wahrscheinlichkeiten der kompletten Sequenzen, als auch die der codierenden Sequenzen, weichen von den indifferenten a priori Wahrscheinlichkeiten ab.

	NA		TA	
	komplette Seq.	CDS	komplette Seq.	CDS
indifferente a priori	25		1,64	
mittlere Abweichung	4,93	2,57	0,83	0,88
max. Abweichung	6,79	4,82	0,92	1,12
min. Abweichung	1,20	1,79	0,72	0,78

Tabelle 3.1: Statistik über die Abweichung der natürlichen a priori Wahrscheinlichkeiten zu indifferenten a priori Wahrscheinlichkeiten in Prozent

Um die Fragestellung zu beantworten, ob eine Verallgemeinerung über die einzelnen Chromosomen möglich ist, muss die Variationsbreite der Chromosomen bezüglich ihrer a priori Wahrscheinlichkeiten, betrachtet werden. Die gleiche Vorgehensweise wird für codierende Sequenzen gewählt.

Um die Variationsbreite der Chromosomen zu beurteilen, wurde die Standardabweichung der Mittelwerte der 24 einzelnen Gewichtungseinheiten der a priori Wahrscheinlichkeiten berechnet. Tabelle 3.2 zeigt diese exemplarisch für die NA Wahrscheinlichkeiten. Da die entsprechenden Tabellen für TA und TT sehr groß sind und deren genaue Betrachtung hier keine weitere Erkenntnis bringt, sind sie hier nicht angegeben. Abschließend werden die mittlere Standardabweichung sowie die maximalen und minimalen Standardabwei-

chungen herangezogen, um einen zusammenfassenden Überblick über die Variationsbreite der 24 Chromosomen (hinsichtlich ihrer a priori Wahrscheinlichkeiten) zu bekommen.

Tabelle 3.3 zeigt, dass die Variation der a priori Verteilung unter den einzelnen Chromosomen relativ gering ist. Die a priori Wahrscheinlichkeiten der einzelnen Chromosomen sind nicht gleich, können aber als ähnlich bezeichnet werden. Eine Verallgemeinerung über alle Chromosomen bezüglich der Auswirkung der a priori Gewichtung kann als vertretbar angenommen werden. Daher wird im Weiteren über die Effekte gesprochen, die sich im Mittel der Chromosomen zeigen. Es wird explizit darauf hingewiesen, wenn die Veränderung der Optimierungseffekte bei den einzelnen Chromosomen nicht in dieselbe Richtung geht. Wenn nicht anders angegeben, kann davon ausgegangen werden, dass die jeweilige Tendenz auf allen Chromosomen gleich ist.

	T	C	A	G
komplette Sequenzen				
Mittelwert	29, 32	20, 68	29, 21	20, 79
SD	±1, 25	±1, 24	±1, 27	±1, 28
codierende Sequenzen				
Mittelwert	21,95	25,58	26,32	26,15
SD	±1, 36	±1, 83	±1, 76	±1, 26

Tabelle 3.2: Statistik über die mittlere NA Wahrscheinlichkeiten und die jeweiligen Standardabweichungen (SD) in Prozent.

	NA		TA	
	komplette Seq.	CCDS	komplette Seq.	CCDS
Mittelwert	25		1,64	
mittlere SD	1,26	1,55	0,13	0,22
max. SD	1,28	1,83	0,35	0,70
min. SD	1,24	1,26	0,00	0,00

Tabelle 3.3: Statistik über die mittlere Standardabweichung (SD) der natürlichen a priori Wahrscheinlichkeiten zwischen den einzelnen Chromosomen in Prozent

Zusammenfassend konnten drei Beobachtungen gemacht werden:

- Die a priori Wahrscheinlichkeiten der natürlichen Sequenzen weichen von den indifferenten a priori Wahrscheinlichkeiten ab.
- Die a priori Wahrscheinlichkeiten der codierenden Sequenzen sind anders als die, der kompletten Sequenzen.
- Die a priori Wahrscheinlichkeiten auf den einzelnen Chromosomen weichen kaum voneinander ab.

3.2 Punktmutationen

In dem Set von einer Millionen zufälligen Codes befinden sich 125 Codes, die bei Punktmutationen effizienter sind als der natürliche genetische Code. Als „effizienter“ werden hier Codes bezeichnet, die einen niedrigeren MS0 Wert haben. Der MS0 Wert ist ein Indikator für die Abweichung der Polarität von der durch Mutation codierten Aminosäure zu der Aminosäure, die von dem fehlerlosen Codon codiert wird. Je kleiner der MS0 Wert, desto ähnlicher sind die Polaritätswerte. Die Wahrscheinlichkeit P allein durch Zufall einen stärker konservierenden Code als den natürlichen zu finden, wird durch die Anzahl der effizienteren Codes im zufälligen Codeset eingeschätzt. Bei Gleichverteilung der Basen und Tripletts beträgt $P_0^u = 0,000125$. Dies reproduziert das Ergebnis von Freeland und Hurst, die in einem anderen zufälligen Codeset 124 effizientere Codes fanden. Daraus folgerten Freeland und Hurst, dass der genetische Code durch natürliche Selektion zur Minimierung der Effekte von Punktmutationen geformt sein könnte.

Werden nun die Nukleotid a priori Wahrscheinlichkeiten aus kompletten Nukleotid Sequenzen in die Berechnung des MS0 Wertes miteinbezogen, so ergibt sich eine Wahrscheinlichkeit von $P_0^k = 0,000305$. Diese Wahrscheinlichkeit ist mehr als doppelt so hoch wie die ohne Gewichtung angenommene.

Stärkere Effizienz hat der natürliche Code, wenn die a priori Wahrscheinlichkeiten der codierenden Sequenzen der menschlichen DNA miteinbezogen werden. Bei der Gewichtung aus den gesamten CDS ist $P_0^c = 0,000103$ kleiner als P_0^u, was für eine Verstärkung des Optimierungseffektes spricht. Allerdings zeigt sich diese Verstärkung nicht auf jedem einzelnen Chromosom. Um die Verstärkung des Minimierungseffektes besser einordnen zu können, ist die Betrachtung der MS0 Werte hilfreich. Dabei zeigt sich, dass der MS0 Wert sich um 1% verringert, während der mittlere MS0 Wert der zufälligen Codes um $0,2\%$ ansteigt. Die Abweichung dieser Werte ist also sehr gering. Die Verbesserung hängt sowohl mit einem etwas geringeren MS0 Wert zusammen als auch damit, dass unter Gewichtung der Nukleotid a priori Wahrscheinlichkeiten der CDS zufällige Codes sich tendenziell gegenläufig zum natürlichen Code verhalten. Das heißt, dass während der MS0 Wert des natürlichen Codes etwas sinkt, der mittlere MS0 Wert der zufälligen Codes etwas ansteigt. Die genauen Werte sind Tabelle 3.4 zu entnehmen.

Abbildung 3.1 zeigt die Verteilung der MS0 Werte der Codes aus dem Set der 1.000.000 zufälligen Codes. In den drei verschiedenen Graphen werden jeweils die $MS0^c$ Histogramme der drei Gewichtungen abgebildet (gemeint sind die a priori Gewichtungen basierend auf CDS: Nukleotid a priori (NA), Triplett a priori (TA) und die Kombination beider (NA+TA)). Dabei ist in jedem Graph auch das Histogramm der MS0 Werte unter Indifferenzannahme der Nukleotide und Tripletts abgebildet, um die Veränderung zu veranschaulichen, welche die a priori Gewichtung auf die Verteilung der MS0 Werte der zufälligen Codes und den MS0 Wert des natürlichen Codes hat. Der MS0 Wert des natürlichen Codes ist jeweils markiert. Die kumulative Frequenz links der Markierung gibt die Anzahl der zufälligen Codes an, die effizienter sind als der natürliche. Durch

Gewichtung	MS0 Wert	Mittelwert
keine Gewichtung	$5,194$	$9,411 \pm 1,511$
CCDS		
NA	$5,145$	$9,427 \pm 1,511$
TA	$5,927$	$9,618 \pm 1,404$
NA+TA	$5,846$	$9,634 \pm 1,410$

Tabelle 3.4: Statistik über die Verteilung der MS0 Werte. Vergleich zwischen der Verteilung mit indifferenten a priori Wahrscheinlichkeiten zu den a priori Wahrscheinlichkeiten der Codierenden Sequenzen

diese Anzahl wird die Wahrscheinlichkeit P effizientere Codes als den natürlichen, zu finden veranschlagt.

In Abbildung 3.1 ist zu sehen, dass die Verteilung der MS0 Werte sich bei TA und NA+TA der codierenden Sequenzen ähnlich verhält. Unter Berücksichtigung der TA aus codierenden Sequenzen steigt zum einen der mittlere MS0 Wert der zufälligen Codes um 2% ($MS0^c = 9,618$), zum anderen erhöht sich der MS0 Wert des natürlichen Codes um 14% auf $MS0^c = 5,927$. Ähnlich verändert sich die MS0 Verteilung bei NA+TA. Auch die Wahrscheinlichkeit, zufällig einen effizienteren Code zu finden, steigt bei beiden Gewichtungen an. Bei TA Gewichtung steigt die Wahrscheinlichkeit um fast das Achtfache der ungewichteten Wahrscheinlichkeit auf $P_0^c = 0,002426$ (die Erhöhung P_0^c ist auf allen Chromosomen zu beobachten). Bei NA+TA Gewichtung steigt sie auf $P_0^c = 0,002036$ (ebenfalls auf allen Chromosomen).

Die Berücksichtigung der TA und NA+TA Gewichtungen aus kompletten Sequenzen bewirkt hingegen eine Reduktion der Effizienzwahrscheinlichkeit. So sinkt P_0 bei TA auf $P_0^k = 0,000024$ um fast $\frac{1}{13}$. Auf jedem einzelnen Chromosom ist eine Verstärkung des Optimierungseffekts zu beobachten. Bei NA+TA Gewichtungen sinkt die Wahrscheinlichkeit $P_0^k = 0,000067$. Dieser Effekt ist jedoch nicht auf jedem einzelnen Chromosom zu beobachten.

Zusammenfassend senkt die Berücksichtigung der TA und NA+TA Gewichtung kompletter Sequenzen die Wahrscheinlichkeit, effizientere Codes zu finden. Die NA Gewichtung kompletter Sequenzen steigert diese Wahrscheinlichkeit. Bei der Berücksichtigung der a priori Gewichtungen der codierenden Sequenzen verhält sich die Auswirkung auf die Wahrscheinlichkeiten gegenteilig. Die NA Gewichtung der CDS senkt die Wahrscheinlichkeit, effektivere Codes zu finden, während die NA+TA Gewichtung die Wahrscheinlichkeit steigert.

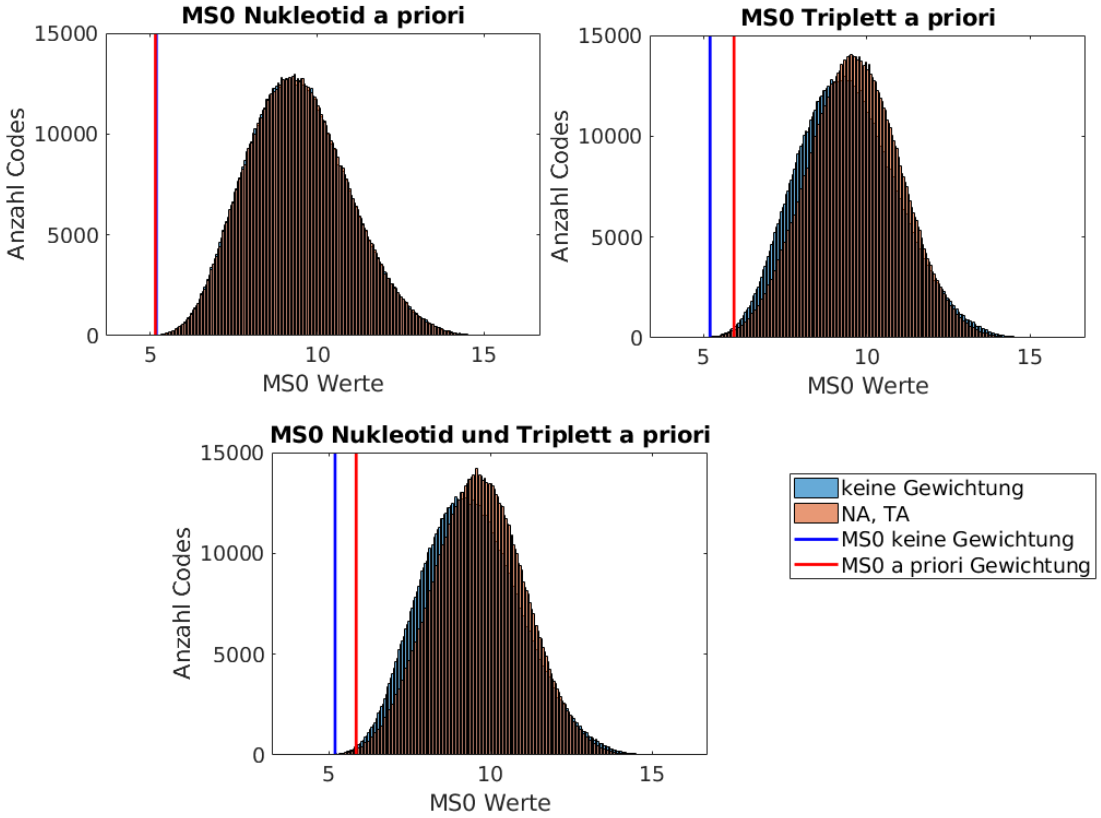

Abbildung 3.1: Histogramme für die MS0 Werte der Codes im Set der 1.000.000 zufälligen Codes. Die einzelnen Plots zeigen die MS0 Werte mit den jeweiligen a priori Gewichtungen der codierenden Sequenzen. Blau sind die Histogramme der Berechnungen ohne a priori Gewichtungen dargestellt, rot die gewichteten Histogramme. Die x-Achse ist die Skala der MS0 Werte. Die y-Achse gibt die Anzahl der Codes an, die den jeweiligen MS0 Wert haben. Der blaue Strich markiert den MS0 Wert des natürlichen Codes ohne Gewichtung. Der rote Strich markiert den MS0 Wert des natürlichen Codes mit a priori Gewichtung.Die kumulative Frequenz des Bereiches links des natürlichen Codes gibt die Anzahl der Codes an, die effizienter sind.

Der MS0 Wert setzt sich aus den MS Werten der einzelnen Positionen im Codon zusammen. Tabelle 3.5 zeigt an den einzelnen Positionen im Codon, sowie für das gesamte Codon, die Anzahl der Codes, die einen geringeren MS Wert haben als der natürliche Code. Für die einzelnen positionsabhänigen MS Werte wirkt sich, wie bei den Berechnungen für den MS0 Wert, die Miteinbeziehung der TA und NA+TA Gewichtungen kompletter Sequenzen abmildernd auf die Wahrscheinlichkeit, effizientere Codes zu finden, aus. Nur für MS2 NA+TA zeigt sich eine Steigerung der Wahrscheinlichkeit. Die a priori Gewich-

Gewichtung	MS0		MS1		MS2		MS3	
keine	125		3086		221300		74	
	ges	CCDS	ges	CCDS	ges	CCDS	ges	CCDS
NA	305	103	3094	2693	288709	212920	83	73
TA	24	2426	2058	4136	178845	446592	62	56
NA+TA	67	2036	1993	3570	247532	425755	63	56

Tabelle 3.5: Statistik über die Anzahl der effektiveren Codes im Set der 1 000 000 zufälligen Codes in Abhängigkeit der a priori Gewichtungen (Nukleotid a priori -NA, Triplett a priori -TA, beide in Kombination NA+TA). Aus der Anzahl der effektiveren Codes ergibt sich die Wahrscheinlichkeit P allein durch Zufall einen effektiveren Code, als den natürlichen zu finden.

tungen der codierenden Sequenzen wirkt sich für die erste und die zweite Base ebenso aus, wie bei der Gesamtbeurteilung des Codons. Nur bei Nukleotid Gewichtung hat sie eine senkende Wirkung. Für die zweite Position gilt dies jedoch ebenfalls lediglich für das Mittel, wobei es einzelne Chromosomen gibt, auf denen die NA Gewichtung einen leicht steigernden Effekt hat. Für die dritte Base wird die Wahrscheinlichkeit bei allen Gewichtungen weiter verstärkt. Bei der NA Gewichtung gilt dies nur für das Mittel, nicht für jedes einzelne Chromosom. Die Verteilung der MS Werte der zufälligen Codes und das Verhältnis des natürlichen Codes zu den zufälligen Codes kann in Tabelle 3.6 und Abbildung 3.2, Abbildung 3.3, Abbildung 3.4 im Einzelnen betrachtet werden.

	a priori	CCDS a priori		
	Indifferenz	NA	TA	NA+TA
MS1 nat. Code	$4,88$	$4,80$	$5,17$	$5,08$
MS1 Mittelwert	$12,05 \pm 2,80$	$12,10 \pm 2,82$	$11,85 \pm 2,49$	$11,92 \pm 2,52$
MS2 nat. Code	$10,56$	$10,49$	$12,47$	$12,32$
MS2 Mittelwert	$12,63 \pm 2,60$	$12,61 \pm 2,60$	$12,82 \pm 2,56$	$12,81 \pm 2,56$
MS3 nat. Wert	$0,14$	$0,14$	$0,13$	$0,13$
MS3 Mittelwert	$3,59 \pm 1,50$	$3,59 \pm 1,50$	$4,21 \pm 1,93$	$4,19 \pm 1,92$

Tabelle 3.6: Statistik über die Verteilung der MS Werte an den einzelnen Positionen im Codon. Vergleich zwischen der Verteilung mit indifferenten a priori Wahrscheinlichkeiten zu den Verteilungen mit a priori Wahrscheinlichkeiten der codierenden Sequenzen

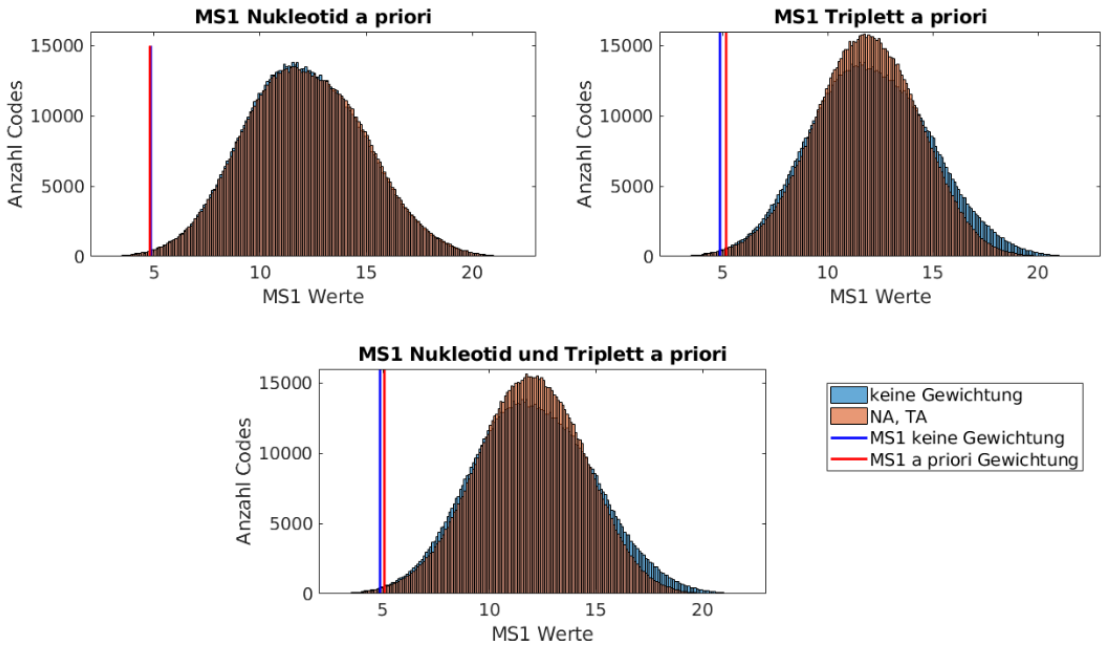

Abbildung 3.2: MS1 Histogramme

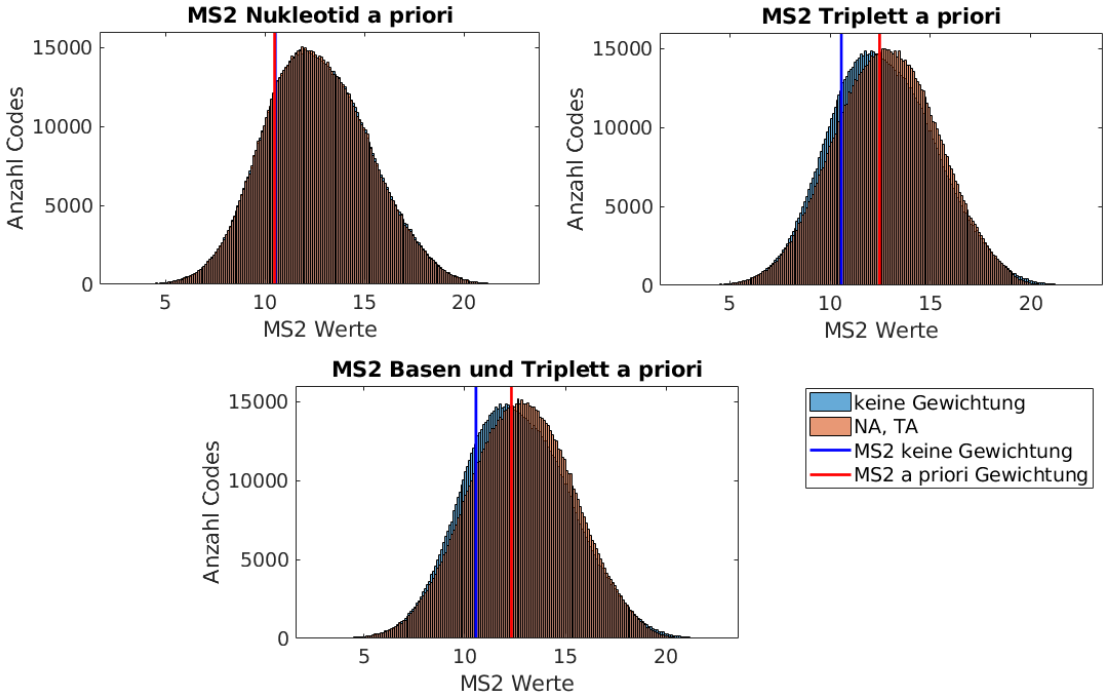

Abbildung 3.3: MS2 Histogramme

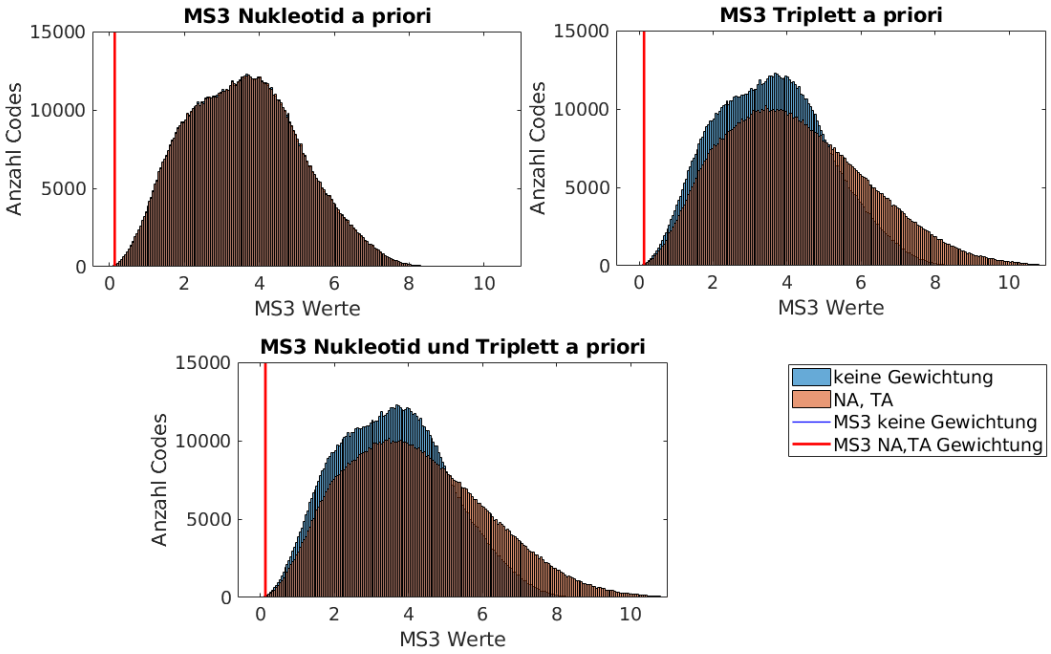

Abbildung 3.4: MS3 Histogramme

3.2.1 entstehende Stopcodons

Mit der Berücksichtigung der a priori Gewichtungen der kompletten Sequenzen steigt die Wahrscheinlichkeit, dass eine Mutation eine Nonsense-Mutation ist. Dies gilt für alle Positionen der Punktmutation, bei allen drei Gewichtungen. Die Anzahl neuer Stopcodons ist in Tabelle 3.7 aufgeführt. Dabei entspricht die Wahrscheinlichkeit einer Nonsense Mutation dem Quotienten der Anzahl neuer Stopcodons dividiert durch die Anzahl aller Mutationscodons(=183).

Anders verhält es sich bei den Gewichtungen der codierenden Sequenzen. Wie Tabelle 3.7 zu entnehmen ist, entstehen mit der NA Gewichtung bei Mutation der ersten Base, sowie unter der TA und der NA+TA Gewichtung bei Mutation der zweiten und dritten Base weniger neue Stopcodons als unter Annahme der a priori Gleichverteilung.

3.3 Leserastermutationen

Keser konnte zeigen, dass mit der Berücksichtigung der Triplett Übergang a priori (TT) Gewichtung aus Chromosom 1, die Wahrscheinlichkeit, effizientere Codes zu finden, deutlich geringer ist, als ohne Gewichtung. Diese Arbeit bestätigt das Ergebnis von Keser für alle humanen Chromosomen. Im Mittel wird die Wahrscheinlichkeit auf $P_f^k = 0,000028$ reduziert.

a priori Gewichtung der kompletten Sequenzen			
	Base 1	Base 2	Base 3
keine	9	7	7
NA	10,80	7,59	7,59
TA	11,27	7,86	7,65
NA+TA	13,52	8,36	7,25

a priori Gewichtung der codierenden Sequenzen			
	Base 1	Base 2	Base 3
keine	9	7	7
NA	7,88	7,34	7,34
TA	13,04	4,34	5,34
NA+TA	11,42	4,55	5,59

Tabelle 3.7: Statistik über die Anzahl durch Mutationen entstehender Stopcodons an den einzelnen Codon Positionen Base 1 bis 3. Vergleich zwischen der Verteilung mit indifferenten a priori Wahrscheinlichkeiten zu den Verteilungen mit a priori Wahrscheinlichkeiten der kompletten und der codierenden Sequenzen. Die Wahrscheinlichkeit, dass durch Punktmutation ein Stopcodon entsteht, entspricht der Anzahl neuer Stopcodons dividiert durch die Anzahl aller entstanden Codons (=183).

Auf codierenden Sequenzen bewirkt die a priori Gewichtung hingegen, dass die Wahrscheinlichkeit, effizientere Codes zu finden, deutlich steigt. Aus der Tabelle 3.9 können die genauen Werte entnommen werden. Es lässt sich nun beobachten, dass rechts und links Leserasterverschiebungen nicht mehr symmetrisch optimiert sind. Die Wahrscheinlichkeiten zufällige, effizientere Codes zu finden, sind für die beiden Mutationsrichtungen verschieden. Bei rechts-Leserasterverschiebungen zeigt der natürliche Code sich stabiler als bei links-Leserasterverschiebungen. Lediglich auf Chromosom Y, bei TT Gewichtung der codierenden Sequenzen und rechts-Leserasterverschiebungen, sinkt die Wahrscheinlichkeit, effizientere Codes zu finden. Bei allen anderen Chromosomen steigen die Wahrscheinlichkeiten sowohl für rechts als auch für links-Leserasterverschiebungen. Abbildung 3.5 veranschaulicht den Anstieg der MS Werte des natürlichen Codes unter Berücksichtigung der TT Gewichtung codierender Sequenzen. Auch der Mittelwert der zufälligen Codes steigt durch diese Gewichtung, ebenso wie die Anzahl zufälliger Codes, die effizienter sind.

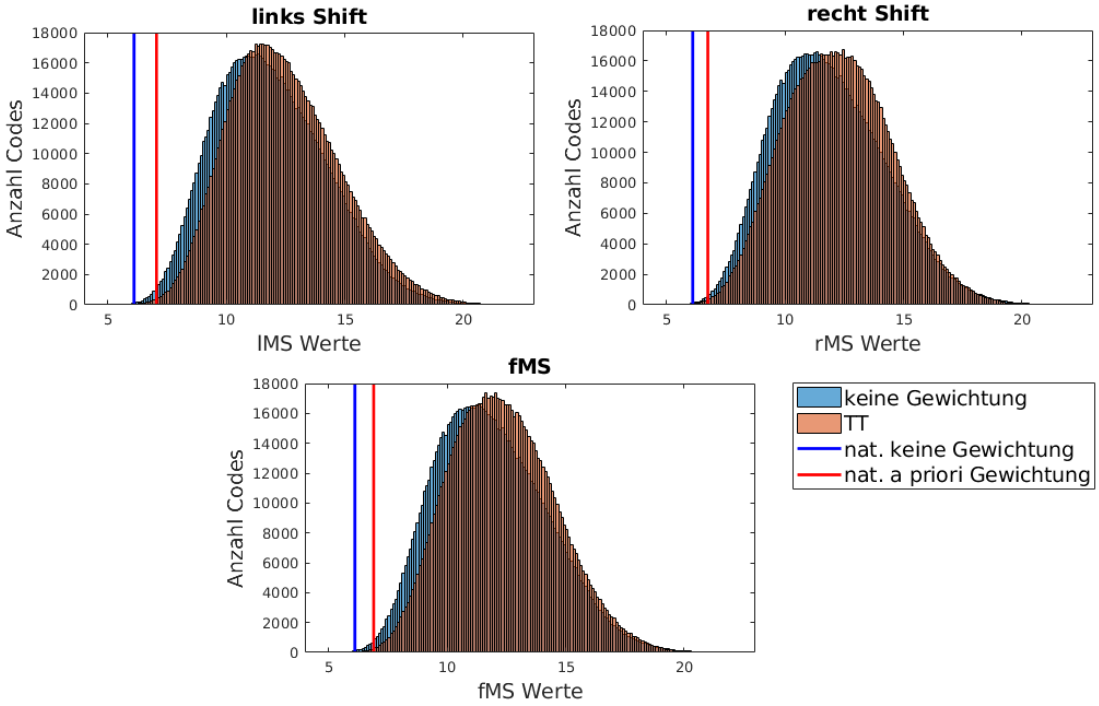

Abbildung 3.5: Leserastermutation Histogramme

	ohne Gewichtung	TT Gewichtung lMS	rMs	fMS
Mittel±SD	$11,87 \pm 2,33$	$12,42 \pm 2,30$	$12,26 \pm 2,23$	$12,34 \pm 2,21$
nat. Code	6,11	7,06	6,75	6,90

Tabelle 3.8: Verteilung bei Leserastermutationen. Mittelwert der zufälligen Codes und Standartabweichung, sowie der MS Wert des natürlichen Codes werden in Abhängigkeit von Gewichtung und Mutationsrichtung gezeigt. Ohne a priori Gewichtung sind die MS Werte an für rMS=lMS=fMS gleich.

Gewichtung	lMS	rMs	fMS
keine	267	267	267
TT komplette S	28	28	28
TT CCDS	1502	1191	965

Tabelle 3.9: Statistik über die Anzahl der effizienteren Codes aus einem Set von 1.000.000 Codes bei Leserastermutationen, ohne a priori Gewichtungen und mit Triplett Übergangs a priori(TT) Gewichtungen kompletter und codierender Sequenzen. Aus der Anzahl kann auf die Wahrscheinlichkeit P, allein durch Zufall einen effizienteren Code, als den natürlichen zu finden, geschlossen werden.

3.3.1 entstehende Stopcodons

Durch Leseraster Mutationen besteht, unter Indifferenzannahme der a priori Gewichtungen, eine Wahrscheinlichkeit von $\frac{12}{244}$ dafür, dass die Mutation eine Nonsense Mutation ist. Tabelle 3.10 kann entnommen werden, dass bei TT Gewichtung der kompletten Sequenzen die Wahrscheinlichkeit steigt, dass durch Leserastermutationen Stopcodons entstehen. Links-Leserasterverschiebungen auf codierenden Sequenzen haben eine geringere Wahrscheinlichkeit neue Stopcodons zu erzeugen. Hingegen haben rechts-Leserasterverschiebungen auf CCDS ein verstärkte Wahrscheinlichkeit dafür.

Gewichtung	lMS	rMs
keine	12	12
TT komplette S	13,92	13,92
TT CCDS	10,88	14,36

Tabelle 3.10: Statistik über die Anzahl der neu entstehenden Stop Codes bei Leserastermutationen. Aus der Anzahl kann auf die Wahrscheinlichkeit P für ein neues Stop Codon geschlossen werden, indem die Anzahl der neu entstanden Stopcodons durch die Anzahl aller möglichen mutierten Codons (=244)f dividiert wird.

3.4 Kombination der Mutationen

Bei der kombinierten Bewertung von Leseraster- und Punktmutationen kam Kesers zu dem Ergebnis, dass Minimierungseffekte sich verstärken, wenn die a priori Wahrscheinlichkeiten des kompletten Chromosom 1 gewichtet werden. Die Untersuchungen dieser Arbeit haben ergeben, dass die a priori Gewichtungen aller kompletten Chromosomen Sequenzen die Minimierungseffekte verstärken. Nur die NA Gewichtung bildet eine Ausnahme davon. Die NA Gewichtung der kompletten Sequenzen bewirkt sowohl bei Chromosom 1 als auch bei allen anderen humanen Chromosomen eine Erhöhung der Wahrscheinlichkeit, zufällige effizientere Codes zu finden. Die stärksten Optimierungshinweise zeigt die Gewichtung mit TA+TT. In Tabelle 3.11 sind zu allen Gewichtungskombinationen die Anzahl der effizienteren Codes aufgelistet.

Mit den Gewichtungen der codierenden Sequenzen verringert sich der Hinweis auf Optimierung hingegen deutlich. Bei allen Gewichtungen und auf jedem einzelnen Chromosom steigt die Wahrscheinlichkeit, zufällig effizientere Codes zu finden. Die stärksten Optimierungstendenzen bewirken bei den codierenden Sequenzen die NA Gewichtung. Diese Optimierungstendenz bei der NA Gewichtung ist jedoch kein gesicherter Hinweis auf Optimierung, da es auch einzelne Chromosomen gibt, bei denen die NA Gewichtung keine Verbesserung im Vergleich zu den Ergebnissen ohne Gewichtungen darstellt.

Gewichtung	komplette S	CCDS
keine	37	
NA	58	35
TA	10	130
TT	4	72
NA+TA	17	113
NA+TT	11	63
TA+TT	3	281
NA+TA+TT	7	257

Tabelle 3.11: Statistik über die Anzahl der GMS effektiveren Codes im Set der 1.000.000 zufälligen Codes in Abhängigkeit der a priori Gewichtungen. (Nukleotid a priori -NA, Triplett a priori -TA,Triplett Nachbarschaft -TT) Aus der Anzahl der effektiveren Codes ergibt sich die Wahrscheinlichkeit P, allein durch Zufall einen effektiveren Code als den natürlichen zu finden.

3.5 Transition/ Transversion Bias

Freeland und Hurst erweiterten ihre Berechnungen um die Gewichtung des Transition/Transversion Bias. Sie fanden heraus, dass die Miteinberechnung des Bias die Wahrscheinlichkeit senkt, allein durch Zufall effizientere Codes als den natürlichen zu finden. Der Hinweis auf Optimierungstendenzen des genetischen Codes verstärkt sich mit Transition Bias. Bei einem Bias von zwei erreicht der natürliche Code nach Geyer sein Minimum der Wahrscheinlichkeit, effektivere zufällige Codes zu finden.

Da die Berechnungen hier um die Gewichtung mit den a priori Wahrscheinlichkeiten erweitert wurde, muss erneut untersucht werden, mit welchem Bias die Wahrscheinlichkeit, effektivere zufällige Codes zu finden, am stärksten minimiert wird. Da andere Arbeiten auf ein natürliches Transition Bias zwischen 1,7 und 5 hinweisen [7], wird der Suchraum des Bias Minimum hier auf einen Bereich von 1 bis 7 beschränkt. Da die NA+TA+TT Gewichtung Mutationen in codierenden Sequenzen am realistischsten simuliert, wird zur Vereinfachung hier nur die Betrachtung der Berechnungen mit dieser a priori Gewichtung im Vergleich zu Berechnungen ohne a priori Gewichtung durchgeführt. In Abbildung 3.6 ist zu sehen, dass ohne a priori Gewichtung das Minimum bei einem Bias von 2 erreicht wird. Dies entspricht dem Ergebnis zu dem Geyer kam. In der a priori gewichteten Berechnung wird das Minimum erst bei mit einem Bias von 4 erreicht. Weiterhin bleibt die Wahrscheinlichkeit, durch Zufall einen effektiveren Code zu finden, bei der a priori Gewichtung deutlich höher, als wenn eine a priori Indifferenz angenommen wird. Die einzelnen Wahrscheinlichkeiten können in Tabelle 3.12 betrachtet werden.

Abbildung 3.6: WMS0 Minimum - Die Anzahl effektiverer Codes in einem Set aus 1.000.000 zufälligen Codes, mit den Bias 1 bis 7 für a priori ungewichtete WMS0 Werte und NA+TA+TT gewichtete WMS0 Werte. Die x-Achse zeigt den Transition Bias. Die y-Achse die Anzahl effektiver Codes.

	WMS0	
Bias	keine S	CCDS
Gewichtung	a priori	a priori
1	0,000 125	0,002 036
2	0,000 032	0,000 809
3	0,000 034	0,000 587
4	0,000 047	0,000 568
5	0,000 070	0,000 594
6	0,000 098	0,000 651
7	0,000 126	0,000 717

Tabelle 3.12: Statistik über die WMS0 Wahrscheinlichkeit P allein durch Zufall einen effektiveren Code als den natürlichen zu finden in Abhängigkeit der a priori Gewichtungen NA+TA.

Geyers Arbeit brachte den deutlichen Hinweis darauf, dass der natürliche Code nicht nur für Punktmutationen, sondern auch für Leserastermutationen optimiert ist. Aber bisher wurde noch nicht untersucht, ob zufällige Codes, die sich bei Punktmutationen mit Transitions Bias konservierender als der natürliche Code verhalten, dieses Verhalten auch bei Leserastermutationen zeigen. Daher erweitert diese Arbeit die Untersuchung des Transition Bias Minimums auf den GMS Wert, der beide Mutationstypen berücksichtigt.

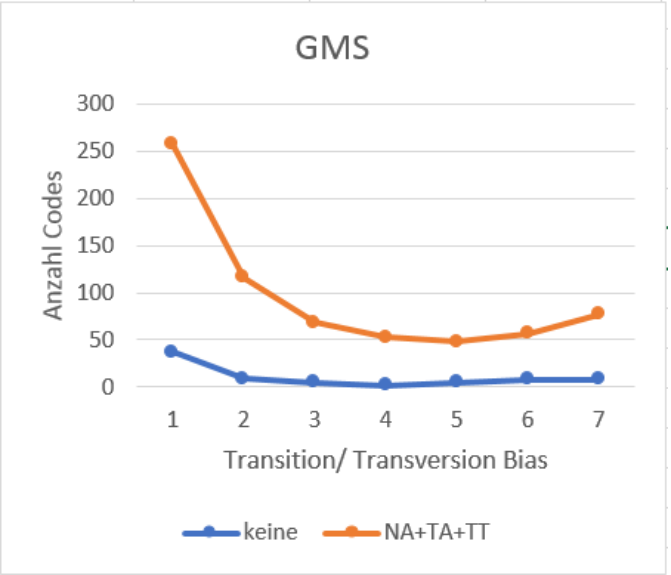

Abbildung 3.7: GMS Minimum - Die Anzahl effektiverer Codes in einem Set aus 1.000.000 zufälligen Codes, mit den Bias 1 bis 7, für a priori ungewichtete GMS Werte und NA+TA+TT gewichtete GMS Werte. Die x-Achse zeigt den Transition Bias, die y-Achse die Anzahl effektiverer Codes.

Abbildung 3.7 und Tabelle 3.13 zeigen, dass das Minimum der Berechnung, die nicht a priori gewichtet, erst bei einem Bias von 4 erreicht wird. Das Minimum bei Miteinbeziehung der a priori Gewichtung der CCDS wird bei einem Bias von 5 gefunden. Die Wahrscheinlichkeit durch Zufall effizientere Codes zu finden, ist bei einem Transition Bias im Bereich von 2 bis 5 für beide Messreihen geringer als ohne Bias. Zudem liegen die Bias Werte, bei denen das Minimum der Wahrscheinlichkeit gefunden wird, ebenfalls für beide im Bereich der Bias Werte, welche in anderer Literatur als das natürliche Bias angenommen werden. In Abbildung 3.8 sind die GMS Verteilungen der beiden am stärksten minimierten Berechnungen abgebildet. Die GMS Werte des natürlichen Codes sind markiert und geben Aufschluss darüber, dass der GMS Wert ohne a priori Gewichtung mit Bias 4 kleiner ist, als der GMS Wert mit NA+TA+TT Gewichtung und Bias 5.

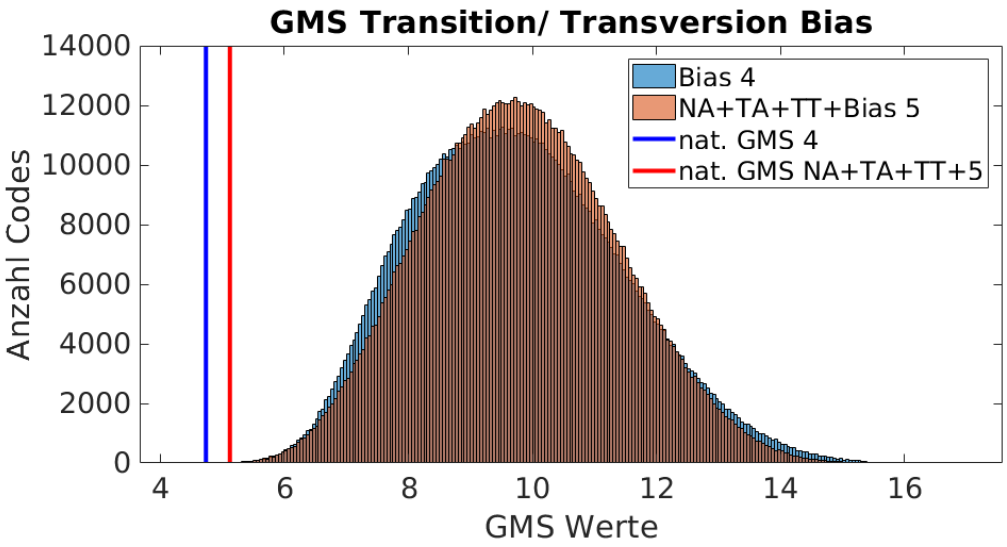

Abbildung 3.8: Histogramm der über das Transition Bias minimierten GMS Werte der Codes im Set der 1.000.000 zufälligen Codes. In blau sind die indifferent a priori gewichteten GMS Werte mit Bias 4 dargestellt. In rot die NA+TA+TT gewichteten GMS Werte mit Bias 5

	GMS	
Bias	keine S	CCDS
Gewichtung	a priori	a priori
1	0,000 037	0,000 257
2	0,000 009	0,000 116
3	0,000 005	0,000 068
4	0,000 002	0,000 053
5	0,000 005	0,000 048
6	0,000 008	0,000 057
7	0,000 008	0,000 077

Tabelle 3.13: Statistik über die GMS Wahrscheinlichkeit P, allein durch Zufall einen effektiveren Code als den natürlichen zu finden, in Abhängigkeit der a priori Gewichtungen NA+TA+TT.

4 Diskussion der Ergebnisse

4.1 Muster der natürlichen Sequenzen

In Abschnitt 3.1 wurde zunächst gezeigt, dass die a priori Wahrscheinlichkeiten der natürlichen Sequenzen, wie erwartet, nicht indifferent sind. Dies gilt sowohl für die a priori Wahrscheinlichkeiten der kompletten Sequenzen, als auch für die der CDS. Da komplette Sequenzen und CDS verschiedene a priori Wahrscheinlichkeiten haben, können die Ergebnisse, die Keser (2016) [10] für Berechnungen mit den Gewichtungen aus der kompletten Sequenz des ersten Chromosoms machte, folglich nicht einfach übertragen werden.
Außerdem wurde festgestellt, dass die a priori Wahrscheinlichkeiten der einzelnen Chromosomen ähnlich sind. Dadurch sind auch die Auswirkungen der a priori Wahrscheinlichkeiten auf dem humanen Genom ähnlich.

Durch die Erkenntnis, dass a priori Häufigkeiten auf allen Chromosomen ähnlich sind, ergibt sich, dass ein a priori Muster zu erkennen ist. Interessanterweise ist dieses Muster auf codierenden Regionen anders, als auf den kompletten Sequenzen. Diese Beobachtung wirft die weiterführende Frage auf, ob es möglich ist, durch a priori Mustererkennung codierende Sequenzen in der DNA ausfindig zu machen. Würden auch nicht codierende Sequenzabschnitte der kompletten DNA das gleiche a priori Muster zeigen, so wäre die Identifikation von CDS mit dieser Methode nicht erfolgreich. In so einem Fall wäre jedoch eine genauere Betrachtung der nicht CDS mit codierenden a priori Muster durchaus interessant. Aus diesen Überlegungen lässt sich die Frage folgern, ob so eventuell Erkenntnisse über die Funktion dieser nicht codierenden Regionen gewonnen werden können. Da in der nicht codierenden DNA sogenannte Pseudogene [3] enthalten sind, liegt die Vermutung nahe, dass „codierende“ Regionen der Pseudogene durch eine Mustererkennung zu finden wären. Zudem ist noch eine weitere Einschränkung einer a priori Mustererkennung codierender Sequenzen zu erwarten: Da die a priori Wahrscheinlichkeiten der CDS unterschiedlicher Chromosomen nur ähnlich und nicht gleich sind, ist es wahrscheinlich, dass die Suche nach codierenden Regionen durch einen Musterkennungsalgorithmus keine präzisen Ergebnisse liefern würde. Dies bleibt allerdings zu zeigen.

4.2 Optimierungseffekte bei Punktmutationen

In Abschnitt 3.2 konnten zunächst die Ergebnisse von Freeland und Hurst (1998) [5] reproduziert werden. Weiterhin konnte mit den TA und NA+TA Gewichtungen der kompletten DNA die Wahrscheinlichkeit, allein durch Zufall effizientere Codes zu finden, weiter gesenkt werden. Gesteigert wird die Wahrscheinlichkeit jedoch durch die NA Gewichtung der kompletten DNA. Diese Tendenzen entsprechen denen, die von Keser (2016) [10] für Chromosom 1 gezeigt wurden. Keser folgerte daraus, dass der genetische Code Optimierungstendenzen für das Triplett Vorkommen in natürlichen Sequenzen des ersten Chromosoms zeigt. Diese Folgerung kann nun für alle humanen Chromosomen bestätigt werden.

Die Tendenzen, welche Gewichtungen der CDS bewirken, verhalten sich genau entgegengesetzt. So wird durch die NA Gewichtung die Wahrscheinlichkeit leicht gesenkt (allerdings gibt es Chromosomen, für die das nicht gilt). Die TA und die NA+TA Gewichtung bewirkt, dass die Wahrscheinlichkeit, durch Zufall effizientere Codes zu finden, deutlich gesteigert wird. So ist diese Wahrscheinlichkeit bei der NA+TA Gewichtung der CDS 30 mal so hoch wie bei gleicher Gewichtung der kompletten Sequenzen. Die vollständigen Wahrscheinlichkeiten sind aus Tabelle 3.5 ersichtlich.

Überraschenderweise tritt also durch die Einschränkung der Gewichtung auf CDS, keine Verstärkung der Optimierungstendenzen ein, sondern im Gegenteil sogar eine Abschwächung. Trotz der abschwächenden Wirkung, die die CDS a priori Gewichtungen auf den Optimierungshinweis haben, bleibt es ein starker Optimierungshinweis. Besonders stark ist dieser an erster und dritter Stelle im Codon. Die Feststellung von Freeland und Hurst (1998) [5], dass für die 2. Base im Triplett keine signifikanten Hinweise auf Optimierung gezeigt werden können, wird auch mit den CDS a priori Gewichtungen nur noch weiter verstärkt.

4.2.1 Regulierung von Nonsense-Mutationen

In Abschnitt 3.2.1 wurde der Einfluss den die CDS und komplette Sequenzen a priori Gewichtungen auf das Auftreten von Nonsense-Mutationen haben, untersucht. Es wurde herausgefunden, dass die a priori Gewichtungen der kompletten Sequenzen das Auftreten von Nonsense-Mutationen begünstigen. Hingegen wirkt sich der Einfluss der CDS a priori Gewichtungen reduzierend auf das Auftreten von Nonsense-Mutationen aus. Neu entstehende Stopcodons haben nicht nur den Effekt, dass das mutierte Codon nicht translatiert wird, sondern sie bewirken den Abbruch der Translation, sodass alle Codons die auf das neue Stopcodon folgen, nicht mehr translatiert werden. Nonsense-Mutationen haben folglich relative schwerwiegende Folgen, sodass die Reduktion von Nonsense-Mutationen einen Optimierungsgewinn darstellen. Die reduzierende Auswirkung der a priori Gegebenheiten auf natürlichen CDS fügt sich gut in die Theorie ein, dass der genetische Code die Auswirkungen von Punktmutationen minimiert.

Weiterführend stellt sich hier die Frage, ob a priori Gegebenheiten auf natürlichen CDS auch den äquivalenten Effekt tendenziell mehr neue Startcodons zu erzeugen, hat. Da das Entstehen neuer Startcodons den Abbrucheffekt neuer Stopcodons auffangen könnte, könnten die Auswirkungen von Punktmutationen so noch weiter reduziert werden.

In die Theorie, dass der genetische Code die Auswirkungen von Punktmutationen minimiert, fügt sich ebenso gut der steigernde Einfluss, der kompletten DNA a priori Gewichtung auf die Entstehung von Stopcodons ein. Da nicht-codierende DNA im fehlerfreien Fall nicht translatiert werden sollte, erscheint die erhöhte Wahrscheinlichkeit, dass Stopcodons entstehen, unterstützend zur Fehlerminimierung. Allerdings passt es nicht zu den deutlichen fehlerminimierenden Tendenzen, welche aus den eben beschriebenen Minimierungstendenzen der Polaritätsabweichungen bei a priori Gewichtung der kompletten DNA klar hervorgehen.

4.3 Optimierungseffekte bei Leserastermutationen

Geyer und Madany (2017) [8] zeigten, dass der genetische Code auch Leserastermutationen hinsichtlich der Polaritätsänderungen optimiert. Unter der a priori Gewichtung der kompletten DNA konnte konnte dieses Ergebnis für alle Chromosomen sogar noch deutlich verstärkt werden (siehe Abschnitt 3.3).

Wie bei den Punktmutationen zeigt sich auch hier das überraschende Ergebnis, dass unter Berücksichtigung der a priori Gewichtungen der CDS, die Optimierungstendenzen deutlich abgeschwächt werden. Außerdem erweisen sich links-Leserasterverschiebungen nun weniger optimiert, als rechts-Leserasterverschiebungen.

Der von Geyer (2014) [7] und Keser (2016) [10] festgestellte Hinweis auf Minimierung der Fehler, wird mit Berücksichtigung der TT Gewichtung codierender Sequenzen zwar klar abgeschwächt, insgesamt bleibt eine Optimierungstendenz aber erhalten.

Aus den Ergebnissen von Abschnitt 3.3.1 kann darauf geschlossen werden, dass die Wahrscheinlichkeit eines neuen Stopcodons und somit eines Abbruchs der Translation bei kompletten Sequenzen deutlich höher ist, als bei CDS links-Leserastermutationen. Auf codierenden Sequenzen ist die Wahrscheinlichkeit des frühzeitigen Translationsabbruches durch entstehende Stopcodons bei Leserasterverschiebungen nach rechts größer, als bei denen nach links. Die Untersuchung auf die Effizienz des natürlichen Codes hatte gezeigt, dass rechts Rasterverschiebungen effizienter sind, als solche nach links. Dennoch sind es interessanterweise die rechts-Leserastermutationen bei denen die Wahrscheinlichkeit, Stopcodons zu erzeugen, erhöht ist. Und es sind die verhältnismäßig weniger polaritätserhaltenden links-Rastermutationen, welche die Erzeugung neuer Stopcodons reduzieren. Auf den ersten Blick scheint das widersprüchliche Optimierungstendenzen zu unterstützen. Doch diese Beobachtung stimmt sehr gut mit den Ergebnissen überein, zu denen Seligmann und Pollock (2004) [2] durch Analyse der mitochondrialen Primaten DNA kamen. Ihre Analyse ergab, dass Spezies mit instabiler rRNA häufiger rechts Stop-

codons (+1) außerhalb des Leserasters (die sie als versteckte Stopcodons bezeichnen) haben. Für -1 Stopcodons konnten sie dies nicht feststellen. In ihrer Arbeit untersuchten sie versteckte Stopcodons in ihrer Auswirkung auf Genexpression. Aus ihren Untersuchungen folgerten sie, dass das Vorkommen versteckter Stopcodons optimal ist, um keine Energie in die Translation leserasterverschobener und somit falscher Sequenzen zu verschwenden. Weiterhin schlossen sie aus taxonomischen Strukturen, dass rechts-Rastermutationen häufiger vorkommen als links-Rasterverschiebungen.
Dass hier die häufiger auftretende rechts-Leserasterverschiebung die polaritätserhaltendere ist, kann als weiterer Hinweis auf die Optimierungstendenzen des natürlichen Codes gesehen werden. Da sich eine Leserastermutation auf die gesamte Sequenz nach dem Mutationspunkt auswirkt, können auch relativ kleine Polaritätsänderungen pro Aminosäure insgesamt starke Auswirkungen haben. In diesem Kontext erscheint es den Hinweis auf Optimierungstendenzen sogar noch zu unterstützen, dass +1 versteckte Stopcodons häufiger vorkommen.

Durch Mutation entstehende Stopcodons sind schwerwiegender, je früher sie in einer Sequenz vorkommen und je weiter sie von einem Startcodon entfernt sind. Daher erscheint es eine interessante weiterführende Aufgabe, zu überprüfen an welcher Position in den Sequenzen die Stopcodons entstehen. Kann vielleicht festgestellt werden, dass die Vermeidungstendenz von Stopcodons zu Beginn der Sequenzen höher ist als zum Ende?

4.4 Kombination der Mutationen

Durch den kombinierten Score wird in den drei Gewichtungsreihen jeweils eine Verbesserung der Effizienz des natürlichen Codes erreicht.
Wieder zeigt sich, dass mit den a priori Gewichtungen der kompletten Sequenzen die von Geyer und Madany (2017) [8] beschriebenen Optimierungseffekte sogar noch verstärkt werden. Die a priori Gewichtungen der CDS hingegen zeigen deutlich schlechter optimierte GMS Werte als die ungewichteten. Dennoch bleibt trotz Verringerung der Optimierungstendenzen insgesamt der Hinweis auf eine Optimierung des natürlichen genetischen Codes zur Erhaltung der Polarität bei Punkt- und Leserastermutationen erhalten.

4.5 Transition/Transversion Bias

In Abschnitt 3.5 wurde festgestellt, dass unter Berücksichtigung der a priori Gewichtung der CDS eine minimale Wahrscheinlichkeit bei einem Transition Bias von 4 erreicht wird. Das Minimum Bias, der MS0 Berechnungen ohne a priori Gewichtung, ist nur halb so groß. Da zufällige Codes, welche die Eigenschaft haben Punktmutationen mit Transition Bias gut zu optimieren, nicht unbedingt auch die notwendigen Eigenschaften haben, um Leserastermutationen zu optimieren, wurde auch für den GMS nach dem Minimum Bias gesucht. Ohne Gewichtungen ist dieses bei 4 erreicht, mit a priori Gewichtung der CDS

bei 5. Diese beiden Bias Werte scheinen noch im Rahmen dessen zu liegen, was dem natürlichen Transition Bias entsprechen könnte. In verlässlicher Literatur konnte hier jedoch nur der Hinweis für ein natürliches Transition Bias, das bei ca. 2 liegt, gefunden werden. Sankoff, Cedrgren und Lapalme (1996) [1] geben einen Transition/Transversion Bias von 2,3 an. Ob das Transition/Transversion Bias von 5 natürliche Gegebenheiten widerspiegelt, bleibt zu klären.

Hier soll nicht beansprucht werden, dass von der Wahrscheinlichkeitsminimumssuche, auf das natürliche Transition/Transversion Bias geschlossen werden kann. Die Beweislage auf das natürliche Bias scheint aber auch nicht so eindeutig, dass nicht ein Blick auf die Minimum-Wahrscheinlichkeiten mit gewählten Bias gewagt werden könnte. Die Berücksichtigung des Transition Bias Minimum 4 bei der GMS Berechnung reduziert die Wahrscheinlichkeit auf $P^u = 0,000002$. Dies kommt dem Ergebnis der „eins aus einer Millionen“, von Freeland und Hurst schon sehr nahe. Unter Berücksichtigung der a priori Gewichtungen und dem Transition/Transversion Bias 5, wird die minimale Wahrscheinlichkeit $P^c = 0,000048$ erreicht. Dies ist deutlich von „eins aus einer Millionen“ entfernt, die Optimierungstendenz des natürlichen Codes bleibt aber trotzdem klar erkennbar.

Mit vermutlich realistischerem Transition/Transversion Bias 2 erreicht die a priori gewichtete Wahrscheinlichkeit $P^c = 0,000116$. Damit würde der natürliche genetische Code trotz der Abschwächung der Optimierungstendenz durch die a priori Gewichtung immer noch mehr als 99,99% der eine Millionen zufälligen Codes in ihrer Effizienz Polarität zu konservieren, übertreffen.

4.6 Zusammenfassung und Ausblick

In dieser Arbeit wurde untersucht, welchen Einfluss natürliche Verteilungen in der DNA auf die Optimierung des genetischen Codes haben. Mit Optimierung ist hier gemeint, wie stark der genetische Code die Veränderung der Polarität durch Mutationen der DNA minimiert. Von Minimierung wird hier auf Optimierung geschlossen, da minimierte Polariätsveränderung zu minimierten Änderungen der Proteinfaltung und somit letztlich zu minimierten Funktionsänderungen führen. Zur Analyse der Optimierung wurden Punktmutationen, Transition/Transversion Bias und Leserastermutationen miteinbezogen. Die natürlichen a priori Wahrscheinlichkeiten wurden für CDS und komplette Sequenzen differenziert analysiert. Es wurde untersucht, wie stark der natürliche genetische Code bei genannten Mutationen die Polarität der ursprünglichen Aminosäure beibehalten kann. Dies wurde ins Verhältnis gesetzt zu der Anzahl Codes, aus einem Set von einer Millionen zufälligen Codes, welche Polaritätsänderung stärker minimieren können als der natürliche Code. Es wurde gezeigt, dass a priori Wahrscheinlichkeiten der kompletten Sequenzen sich positiv auf die Optimierung des genetischen Codes auswirken. A priori Wahrscheinlichkeiten der CDS wirken sich hingegen negativ auf die Optimierung des genetischen Codes aus. Ohne Berücksichtigung des CDS a priori Wahrscheinlichkeiten zeigte sich der genetische Code besser optimiert als 99,9991% von einer Millionen zufälligen Codes. Mit

CDS a priori Gewichtung sinkt diese Wahrscheinlichkeit auf 99,9884%. Der genetische Code ist offensichtlich immer noch immer noch dahingehend optimiert, Fehler durch Mutation zu minimieren. Allerdings deuten die Ergebnisse darauf hin, dass der genetische Code nicht der Code ist, welcher Mutationsauswirkungen am besten minimiert. Dadurch wird die Frage aufgeworfen, ob das Minimum an Konservierung tatsächlich ein Optimum für den Menschen beutet? Entspricht das Optimum vielleicht viel mehr der perfekten Balance zwischen Fehlerresistenz und Flexibilität? Durch Mutationen könnten auch Verbesserungen entstehen. Und so könnte die Balance zwischen dem Beibehalten von funktionalen Strukturen in Kombination mit gelegentlichem Zulassen von Veränderungen das Optimum sein.

An dieser Stelle möchte ich noch einmal betonen, worauf auch Geyer und Madany (2017) [8] deutlich hinwiesen: Hier soll nicht behauptet werden, dass Polaritätskonservierung allein die evolutionstreibende Kraft war, die den genetischen Code geformt hat. Untersuchungen weiterer Faktoren sind hier sicherlich interessant.

Die Untersuchung des Translationsfehlers, die schon Freeman und Hurst, aber auch Geyer und Madany als wesentlich herausstellten, bleibt in dieser Arbeit noch offen. Im nächsten Schritt könnte der Translationsfehler noch miteinbezogen werden, um zu noch aussagekräftigeren Ergebnissen zu kommen.

Das wirklich überraschende Ergebnis dieser Arbeit ist, dass a priori Wahrscheinlichkeiten kompletter Sequenzen sich hingegen der Wirkung von CDS a priori Wahrscheinlichkeiten, positiv auf die Minimierung des genetischen Codes auswirken. Ausgerechnet Sequenzen, die nicht translatiert werden, die also in der Realität nicht für Aminosäuren codieren, sind in dem theoretischen Szenario, dass sie translatiert werden, besonders polaritätskonserviert. Der genetische Code zeigt also die Eigenschaft, die Polarität von nicht existenten Aminosäuren besonders gut zu konservieren. Diese Aussage erscheint mindestens auf den ersten Blick absolut unsinnig, spiegelt aber die hier gefundenen Ergebnisse wider. Daher stellt sich unweigerlich die Frage, ob es Zusammenhänge gibt, die dieses merkwürdige Phänomen erklären. Weist diese Optimierung vielleicht auf eine bisher unerforschte Funktion der nicht-codierenden Sequenzen hin?

Weiterführend wäre es auch interessant, zu untersuchen, wie natürliche Verteilungen sich auf andere Organismen auswirken. Der genetische Code ist, bis auf wenige Ausnahmen, universell. Allerdings stellt sich die Frage, ob auch seine mutationsminimierende Eigenschaft universell ist. Oder ist derselbe Code vielleicht für unterschiedliche Organismen unterschiedlich optimiert?

Literaturverzeichnis

[1] Sankoff, D. und Cedrgren, R. J. und Lapalme, G. . Frequency of Insertion-Deletion, Transversion, and Transition in the Evolution of 5S Ribosomal RNA . *Journal of Molecular Evolution 7* (1966), 133–149.

[2] Seligmann, H. und Pollock, D. D. . The Ambush Hypothesis: Hidden Stop Codons Prevent Off-Frame Gene Reading . *DNA and Cell Biology 23(10)* (2004), 701–5.

[3] Zhang, Z. und Gerstein, M. . Largescale analysis of pseudogenes in the human genome. *Current opinion in Genetics and Development Bd 14, Heft 4* (2004), 328.

[4] Alexander, P. A., He, Y., Chen, Y., Orban, J., and Bryan, P. N. The design and characterization of two proteins with 88identity but different structure and function. . *Proc Natl Acad Sci USA, 104(29)* (2007), 11963–8.

[5] Freeland, S. J. und Hurst, L. D. The Genetic Code Is One in a Million. *Journal of Molecular Evolution 47* (1998), 238–248.

[6] Freeman, W. H., Griffiths A. J. F., Gelbart WM, Miller J. H., et al. *The Molecular Basis of Mutation, Modern Genetic Analysis.* https://www.ncbi.nlm.nih.gov/books/NBK21322/, 1999.

[7] Geyer, R. Frameshift Mutations of the Genetic Code and Their Impact on the Polarity Conservation of Amino Acids. Universität zu Lübeck, 2014. Bachelorarbeit.

[8] Geyer, R. und Madany Mamlouk, A. On the Efficiency of the Genetic Code after Frameshift Mutations. *PeerJ Preprints 5:e3270v1* (2017), https://doi.org/10.7287/peerj.preprints.3270v1.

[9] Haig, D. und Hurst, L. D. A Quantitative Measure of Error Minimization in the Genetic Code. *Journal of Molecular Evolution 33* (1991), 412–417.

[10] Keser, S. The DNA is More than One in a Million. Universität zu Lübeck, 2016. Bachelorarbeit.

[11] National Center for Biotechnology Information, Jan. 2018. https://www.ncbi.nlm.nih.gov/.

[12] NCBI The Consensus CDS (CCDS) project, Jan. 2018. https://www.ncbi.nlm.nih.gov/projects/CCDS/CcdsBrowse.cgi.

[13] Woese, C. R., Dugre, D. H., Saxinger, W. C., and Dugre, S. A. . The Molecular Basis for the Genetic Code . *Proc Natl Acad Sci USA 55(4)* (1996), 966–974.