ulm university universität
uulm

Innovative Barcode-Konzepte für DNA-Sequenzierverfahren der zweiten Generation

DISSERTATION

zur Erlangung des akademischen Grades eines

DOKTOR-INGENIEURS

(DR.-ING.)

der Fakultät für Ingenieurwissenschaften,
Informatik und Psychologie der Universität Ulm

von

David Viktor Kracht

aus Illertissen

Gutachter:	Prof. Dr.-Ing. Martin Bossert Prof. Dr. Daniel Keim
Amtierende Dekanin:	Prof. Dr. Tina Seufert

Ulm, 29. April 2016

Vorbemerkung

Die dargelegte Arbeit entstand im Rahmen der Forschungstätigkeit als akademischer Mitarbeiter am Institut für Nachrichtentechnik der Universität Ulm (ehemals Institut für Telekommunikationstechnik und Angewandte Informationstheorie, TAIT). Der interdisziplinäre Kontext der untersuchten Thematik gründet in einem Forschungsprojekt des Schwerpunktprogramms InKoMBio (Informations- und Kommunikationstheorie in der Molekularbiologie, SPP 1395) der Deutschen Forschungsgemeinschaft (DFG). Das Leitbild des Forschungsschwerpunktes ist die gezielte Förderung eines fächerübergreifenden Ideentransfers zwischen Wissenschaftlern auf dem Gebiet der Lebens- und Ingenieurwissenschaften. Ein essentieller Aspekt des abgeschlossenen Projekts mit der Bezeichnung *„IRseq - Improving the Reliability of RNA-seq: Approaching Single-Cell Transcriptomics to Explore Individuality in Bacteria“* (BO867/30) ist die Steigerung der Zuverlässigkeit bei der Sequenzierung von DNA und RNA durch Konzepte der Kanalcodierung. Bei der Realisierung dieses Ziels hatten sogenannte *Barcodes*, als Molekülketten synthetisierte Codeworte aus DNA, eine wichtige Schlüsselfunktion. Ausgehend von etablierten Ansätzen zur Markierung natürlicher DNA mit Barcodes mussten hierzu innovative Konzepte zur Umsetzung von Barcodes entwickelt werden. Die vorliegende Dissertationsschrift umfasst eine interdisziplinäre Beschreibung der neuen Ansätze und deren Umsetzung. Einige Ergebnisse der präsentierten Arbeit wurden bereits zuvor veröffentlicht, siehe hierzu:

- David Kracht und Steffen Schober, Using the Davey-MacKay code construction for barcodes in DNA sequencing, *Proceedings of the International Symposium on Turbo Codes and Iterative Information Processing (ISTC)*, S. 142–146, 2014.

- David Kracht und Steffen Schober, Insertion and deletion correcting DNA barcodes based on watermarks, *BMC Bioinformatics*, Bd. 16(1), S. 1–14, 2015.

Das in dieser Arbeit enthaltene Farbschema wurde aus Rücksicht auf Leser mit Farbenfehlsichtigkeit und Farbsinnstörung genutzt.

Danksagung

Mein besonderer Dank gilt meinem Doktorvater Prof. Dr.-Ing. Martin Bossert, da er mich als Stipendiat in seinem Institut aufgenommen hatte und mir somit die Möglichkeit bot, ein Teil seiner wissenschaftlichen Gemeinschaft zu werden - einer Gemeinschaft in welcher nicht streng und ausschließlich akademischer Erfolg eingefordert wurde, sondern hauptsächlich Selbstständigkeit und Eigenverantwortlichkeit als Kernkompetenzen gefördert wurden.

Weiter möchte ich im Kontext meiner Dissertation Herrn Prof. Dr. Daniel Keim von der Universität Konstanz meine große Dankbarkeit äußern, dass er meine Arbeit begutachtete und damit zu meiner Promotion entscheidend beitrug. Natürlich sei an dieser Stelle den weiteren Mitgliedern des Prüfungskomitees zu danken, so dem Prüfungsvorsitzenden Herrn Prof. Dr. Hans Armin Kestler und Jun.-Prof. Dr. Jens Anders.

Ich möchte an dieser Stelle die Gelegenheit nutzen und allen Personen des Instituts für Nachrichtentechnik danken, für die schöne Zeit und lustigen Momente, die ich mit ihnen teilen durfte. Es war eine tolle Gemeinschaft und ich habe nicht nur das Fachliche, sondern insbesondere alles Zwischenmenschliche als sehr wertvoll empfunden. Speziellen Dank möchte ich an Dr.-Ing. Katharina Schilling und Dr.-Ing. Henning Zörlein richten, für das Lektorat meiner Arbeit und deren konstruktive Kritik.

Abschließend möchte ich mich bei meinen Eltern bedanken, dass Sie mir einen Grundstein für die Bildung gelegt haben, ganz besonders weil Sie keinen akademischen Hintergrund aufweisen konnten, dies aber trotzdem kein Hindernis für mich bedeuten musste.

David Kracht
Ulm, April 2016

Kurzfassung

Die Sequenzierung von DNA wurde zur Jahrtausendwende durch die Entwicklung der Sequenzierverfahren zweiter Generation revolutioniert. Die massive Parallelisierung als Kernkonzept ermöglicht seither eine stetig steigende Effizienz und damit sinkende Kosten für die Analyse von DNA und RNA. Grundlegend für die Parallelisierung ist auch der Einsatz von synthetisch erzeugten DNA-Sequenzen, *Oligonukleotide* genannt, die im Verbund mit nativen Molekül-Fragmenten eigenständige Einheiten, sogenannte *Templates*, bilden. Die Einbettung von zusätzlichen Sequenzen zur Markierung der Templates vor der Verarbeitung begründet das Feld der *Barcodes*. Das Auftreten von *Sequenzfehlern* und quantitativen Effekten im experimentellen *Protokoll* der Sequenzierung, wie in der *PCR*, motivieren die Anwendung von Fehlerkorrekturmechanismen und die Normalisierung von Zählgrößen durch den Einsatz von Barcodes.

Die dargelegte Arbeit umfasst zwei unterschiedliche Konzepte zum Einsatz von Barcodes, wobei zwei gänzlich verschiedene Anwendungsgebiete betrachtet werden: Zur gemeinsamen Sequenzierung von unterschiedlichen Proben, Multiplexing genannt, wird das Konzept der Watermark Codes für die Verwendung als Barcodes vorgeschlagen. Basierend auf dem ursprünglichen Prinzip von Davey und MacKay aus dem Jahre 2000 wird das Konzept für die DNA-Sequenzierung angepasst und die Eignung des Verfahrens anhand praxistauglicher Codes gezeigt.

Den zweiten Themenkomplex bildet der Einsatz von *Zufallsbarcodes*. Auf Basis der zufälligen Kombination von speziellen Oligonukleotiden können mit moderatem Aufwand sehr viele unterschiedliche Zufallsbarcodes erzeugt werden, die zur Zählung von Molekülen verwendet werden. Als Erweiterung von zwei bereits bekannten Konzepten wird ein verallgemeinertes Verfahren zur Implementierung von Zufallsbarcodes vorgestellt und experimentell anhand der Illumina Technologie evaluiert.

Abstract

Since the turn of the millennium, DNA-Sequencing has been revolutionised by the upcoming next-generation sequencing methods. The massive parallelisation, as a central concept, provides a steadily increasing efficiency and dropping costs on the analysis of DNA and RNA. For this parallel strategy an integration of synthetic DNA sequences, the oligonucleotides, is essential, which are used to build short separable compounds with native DNA fragments, the so called templates. The concept of barcodes is involved in the embedding of additional sequences into the oligonucleotides to label the compound-molecules prior to experimental processing. The existence of sequence errors and quantitative effects during the sequencing protocol, e.g. the PCR, gives the motivation for the application of error correction and the normalization of molecule-counts via labelling.

The presented work includes two diverse concepts for barcoding, within two entirely different tasks: For the joint sequencing of several probes, known as multiplexing, the concept of Watermark Codes is proposed: Based on the original principle given by Davey and MacKay in the year 2000, an adaptation for DNA sequencing is given as a proof of principle study.

The second topic is the application of random barcodes. Based on the stochastic combination of well-defined oligonucleotides, random barcodes can give a cost-efficient generation of diverse sequences to be used for counting molecules. As the generalisation of two known concepts, a novel method is proposed to produce random codes, which are evaluated via the Illumina sequencing technology.

Inhaltsverzeichnis

1

Einleitung und Kontext

Fehlererkennung und Fehlerkorrektur sind essentielle Grundpfeiler der modernen Speicherung von Daten und heutiger Übertragung von Informationen über unzuverlässige Kommunikationskanäle. Dass die Rückgewinnung von übertragener Information durch die Integration von zusätzlicher Redundanz in Nutzdaten verbessert wird, war schon vor dem Jahre 1948 bekannt, jedoch konstatiert die Veröffentlichungen von Shannon [151] aus besagtem Jahr den Beginn der Ära der Informations- und Codierungstheorie. Basierend auf dem Begriff der Entropie, als mittleren Informationsgehalt, beschrieb Shannon sowohl die Kapazität eines Kanals, als auch die Rate einer Übertragung und postulierte, dass eine fehlerfreie Übertragung über jeden Kanal prinzipiell möglich sei, wenn ihre Rate die Kapazität des Kanals nicht übersteigt. Die dafür nötige unbeschränkte Verlängerung der zu übertragenen Sequenz führt dabei zu einer theoretischen Perspektive. Die theoretische obere Schranke der Kanalkapazität wurde im Kanalcodiertheorem beschrieben und ein Erreichen dieser besonderen Grenze war ab 1948 ein definiertes Ziel für viele praktische Anwendungen der Kanalcodierung.

Nur einige Zeit später, im Jahre 1953, entdecken die späteren Nobelpreisträger Watson und Crick [171] die molekulare Struktur der DNA, die bis heute als universelles Speichermedium für jede Form von Leben gilt. Die Wissenschaftler begründeten damit ein neu aufkommendes Bewusstsein im Bereich der Molekularbiologie. Das doch sehr diffuse Abbild der DNA, das mittels Röntgenstrukturanalyse zuvor schon von Franklin und Gosling [55] erzeugt wurde, führten Watson und Crick jedoch zum Strukturmodell der Doppelhelix, das heute in jedem Biologielehrbuch zu finden ist. Sie legten mit ihrer Erkenntnis zur DNA als redundantes (sich komplementär ergänzendes) *Polymer* aus vier unterschiedlichen Molekülen, *Nukleotide* genannt, den Grundstein für die gezielte Analyse der darin enthaltenen Information. Die von Sanger und Coulson um das Jahr 1975 entwickelte DNA-Sequenzierung [141, 142] lieferte darauf die erste skalierbare Technik zur Auswertung von DNA-Molekülen. Sie ermöglichten die erste Abbildung der DNA als textuelle Sequenz von Symbolen und damit einen Ansatz zum direkten Lesen der Erbinformation. Trotz fortwährender Weiterentwicklung der Sanger-Sequenzierung existieren technologische und konzeptionelle Hindernisse, die eine stetige Steigerung der Effizienz (Durchsatz sequenzierter Nukleotiden) limitieren. In der Mitte der 90er Jahre leiteten neue Ansätze zur Sequenzierung der zweiten Generation (engl. *next-generation sequencing*, NGS) einen Paradigmenwechsel im Bereich der DNA-Analyse ein. Die reversible Unterbrechung der DNA-Synthese von Tsien et al. im Jahre 1991 [165] oder das Konzept zur zufälligen Bindung und Vervielfältigung von DNA an definierten Oberflächen, entwickelt von Kawashima et al. [86], sind nur beispielhafte Ingredienzien, welche sich in der parallelen Sequenzierung unter dem Technologiebegriff Solexa/Illumina vereinen. Die Jahrtausendwende ist ferner die Geburtsstunde unterschiedlichster Sequenziertechnologien, die sich der massiven Parallelisierung bedienen, um einen enormen Durchsatz an DNA-Molekülen zu ermöglichen und damit den Weg für die fortschreitende Entschlüsselung der

DNA bereiten. Der beschriebene Wandel in der Sequenzierung von DNA impliziert aber auch neue Herausforderungen, die stellvertretend für die gesamte Molekularbiologie des beginnenden 21. Jahrhunderts sind: Die Biologie ist zunehmend auf den interdisziplinären Transfer von Konzepten verschiedenster Forschungsfelder angewiesen, um eine einsetzende Flut von Daten mit Bedeutung zu versehen. Die Bereiche der Systembiologie, welche Konzepte der Mathematik und Systemtheorie zur Analyse und Modellierung von Lebensvorgängen nutzt, oder die Bioinformatik, welche computergestützte Lösungen für komplexe Probleme der Lebenswissenschaften bietet, sind neue Grundpfeiler der heutigen Molekularbiologie. Sowohl das Feld der Informationstheorie als auch die Kanalcodierung tragen hierfür Methoden für ein immer detaillierteres Modell des Lebens bei.

Blickt man für den zuvor skizzierten Zeitrahmen auf das Gebiet der Kanalcodierung, so zeigen sich ausgehend von Shannons Veröffentlichung eine Vielzahl von unterschiedlichsten Konzepten und Konstruktionen wie Information in Codeworte integriert, aber auch Methoden wie die in den Codes enthaltene Information durch Decodierung (möglichst effizient) zurückgewonnen werden kann. Hierzu folgender beispielhafter Kurzüberblick: Der im Jahre 1949 entdeckte Golay-Code [62] ist der bisher einzig-bekannte nicht-triviale perfekte Code. Perfekt als Struktureigenschaft bedeutet, dass sich zwei beliebige Codeworte stets um exakt dieselbe Anzahl von Symbolen unterscheiden. Im Jahre 1950 entwarf Richard Hamming den gleichnamigen Hamming-Code [69], um Fehler bei der Zeiterfassung mittels Lochkarten zu korrigieren. Er ist auch Namensgeber für die sogenannte Hamming-Metrik, welche den Unterschied zweier Sequenzen als Abstandsbegriff bewertet. Der minimale *Hamming-Abstand* von Codeworten kann seither als Charakterisierung der minimalen Korrekturfähigkeit eines Codes genutzt werden. Nach Irving Reed und David Muller wurden die 1954 vorgeschlagenen Reed–Muller-Codes [120, 135] benannt, welche unter anderem von 1969-1977 von der NASA zur Kommunikation zu Raumfahrtsonden des Mariner Vorhabens eingesetzt wurden [130]. Erste Arbeiten zu Product-Codes [45], der Multiplikation von zwei Codes, wurden von Peter Elias ab 1954 veröffentlicht. Die Klasse der sogenannten Faltungscodes wurde 1955 ebenfalls von Elias eingeführt [46] und beschreiben durch die auf einem Strom von Daten basierende Erzeugung von Codesymbolen eine gänzlich neue Form der Codierung. Methoden zur trellisbasierten Decodierung von Faltungscodes wurden als Viterbi-Algorithmus [167] oder BCJR-Algorithmus [12] bekannt. Im Jahre 1960 präsentierten Irving Reed und Gustave Solomon die gleichnamigen Reed-Solomon-Codes [136] (RS-Codes), welche in enger Verbindung zu den BCH-Codes [21, 75] (Bose-Chaudhuri-Hocquenghem-Codes) stehen, über die im gleichen Jahr publiziert wurde. Die Beschreibung einer effizienten Decodierung der BCH-Codes durch den Berlekamp-Massey-Algorithmus [17, 109] bereitete den Weg für den Einsatz von RS-Codes im Voyager-Programm der NASA aus dem Jahr 1977 oder die kommerzielle Implementierung für die Fehlerkorrektur der Compact Disk (CD), die im Jahr 1982 eingeführt wurde. Die LDPC-Codes (engl. *Low-Density-Parity-Check-Codes*), oder auch Gallager-Codes genannt, wurden 1962 von Robert Gallager [58, 59] entworfen und gerieten wegen der komplexen Implementierung bald in Vergessenheit, bis sie Ende der 90er Jahre eine Renaissance erfuhren. Zwischenzeitlich wurde 1965 das Konzept der verketteten Codierung von Dave Forney [51, 52] vorgeschlagen, mit dem Ziel der freien Kombination von bekannten Codes neue längere Codes zu erzeugen, die einerseits eine bessere Fehlerkorrektur aufweisen, andererseits eine moderat steigende Komplexität der Decodierung erfordern. In Forneys Dissertation [52] konnte er somit für längere Codes die Theorie von Shannon durch Simulationen bestätigen, auch wenn die Raten seiner Codes eine deutliche Lücke zur Kanalkapazität aufwiesen. Es dauerte schließlich mehr als 30

Jahre bis, durch die sogenannten Turbo-Codes [18] (basierend auf Product-Codes) oder die Neubetrachtung der LDPC-Codes [106], die Kanalkapazität für bestimmte Übertragungsszenarien heutzutage nahezu erreichbar ist: Turbo-Codes sind beispielsweise gegenwärtig Teil der Mobilfunkstandards wie 3G(UMTS,HSDPA) und 4G(LTE,WiMAX), LDPC-Codes finden sich in der Spezifikation von DVB-S2 (engl. *Digital Video Broadcasting–Satellite 2*) sowie dem als IEEE 802.11n bekannten WLAN-Standard und erlauben zuverlässige und extrem hochratige Übertragungen.

Obwohl Shannons implizite Zielvorgabe der Kanalkapazität von 1948 heute als quasi erreicht angesehen werden kann, ergeben sich für die Kanalcodierung immer neue Einsatzfelder fernab vom strikten Fokus auf die Kanalkapazität oder herkömmlichen Übertragungswegen: Die Forschung auf dem Feld des Quantencomputers stellt beispielsweise bisher als sicher geltende Kryptosysteme wie RSA in Frage und führt zu einer Neubewertung [124] von codebasierten Konzepten wie dem McEliece-Kryptosystem [111] von 1978. Im Bereich der Molekularbiologie wurde der Einsatz von Codes für die Integration von Daten in DNA in Erwägung gezogen, wobei der Übertragungskanal durch die natürliche (fehlerbehaftete) Replikation der DNA beschrieben wird. In [13] wurde beispielsweise die theoretische Kapazität einer Datenspeicherung in DNA untersucht. Während derartige Vorhaben momentan weitestgehend konzeptioneller Art sind, bietet der Bereich der DNA-Sequenzierung der zweiten Generation praktische Einsatzszenarien für fehlerkorrigierende Codes. Die Verwendung sogenannter *Barcodes*, synthetisch erzeugte DNA-Sequenzen zur Markierung von nativen DNA-Fragmenten, ist aktueller Stand der Technik in allen neuen Sequenziertechnologien. Durch die konkrete Konzeptionierung und Implementierung von Barcodes im molekularen Kontext besteht jedoch erheblicher Gestaltungsspielraum für neue Ideen oder zuweilen ungewöhnliche Konzepte.

Die vorliegende Arbeit befasst sich mit dem Einsatz von zwei unterschiedlichen Konzepten zur Markierung von DNA-Fragmenten durch Barcodes im Kontext der Sequenzierung. Dabei werden zwei vollständig unterschiedliche Einsatzfelder beschritten: Die Verwendung von fehlerkorrigierenden Codes zur parallelen Sequenzierung von unterschiedlichen Proben, auch Multiplexing genannt, folgt dabei gänzlich anderen Prinzipien und Kriterien als die Implementierung von *Zufallsbarcodes*, welche zur Zählung von Molekülen im Bereich der RNA-Sequenzierung eingesetzt werden. Während für erstere Anwendung dedizierte Sequenzen für die Markierung entworfen werden, welche einen möglichst hohen Schutz gegen *Sequenzfehler* bieten sollen, steht für Zufallsbarcodes eine möglichst hohe Anzahl unterschiedlicher Sequenzen im Fokus. Da der Aufwand für spezifische symbolweise Erzeugung von synthetischen Molekülen nicht unerheblich ist, bedient man sich für letzteres Einsatzgebiet meist der zufälligen Kombination von kürzeren, speziell entworfenen, Sequenzen, um möglichst viele unterschiedliche Realisierungen von Barcodes zu erzeugen. In diesem Zusammenhang wird auch von einer hohen *Diversität* der zufälligen Moleküle gesprochen. Aspekte der Fehlerkorrektur sind natürlich auch für den sinnvollen Einsatz der Zufallsbarcodes erforderlich, spielen jedoch neben der Diversität eine untergeordnete Rolle.

Für den Einsatz zum Multiplexing wird in dieser Arbeit das von Davey und MacKay in [40, 41] vorgestellte Konzept der *Watermark Codes* für Barcodes vorgeschlagen und in angepasster Form dargelegt. Das Prinzip der Watermark Codes wurde ursprünglich entworfen, um die Synchronisierung und Decodierung für die kontinuierliche Übertragung über einen binären Kanal zu ermöglichen, wenn in diesem sowohl Ersetzungen als auch Einfügungen oder Löschungen von Bits

auftreten. Mögliche Szenarien, die einer praktischen Realisierung dieser Fehler entsprechen sind jedoch rar, weil allgemeine Kommunikationssysteme meist Synchronisation von den Aspekten der Kanalcodierung trennen. Die beschriebenen Fehler sind im Bereich der DNA-Sequenzierung jedoch faktisch zu finden, daher stellt die Adaption des Konzepts eine praxisrelevante Anwendung der Watermark Codes dar. In diesem Manuskript wird ein Machbarkeitsnachweis und eine mögliche Konzeption von Barcodes auf Basis der Watermark Codes vorgestellt.

Eine praktische Umsetzung beinhaltet der zweite erarbeitete Themenkomplex der *Zufallsbarcodes*: Aus der Kombination zweier bekannter Konzepte zur zufälligen Markierung von Molekülen wird ein neuartiges Verfahren zur Erzeugung von Zufallsbarcodes für die RNA-Sequenzierung vorgestellt, experimentell auf molekularer Ebene implementiert und evaluiert. In fächerübergreifender Zusammenarbeit mit dem Kooperationspartner[1] aus dem Bereich der Molekularbiologie, erfolgte die Konzeption und Realisierung der neuen Ansätze. Die vorliegende Arbeit umfasst eine Beschreibung dieser neuen Ideen und deren interdisziplinäre Umsetzung aus der Perspektive eines Ingenieurs.

[1] ZIEL (Zentralinstitut für Ernährungs- und Lebensmittelforschung), Abteilung Mikrobiologie, TU München

Die Arbeit gliedert sich folgendermaßen:

Kapitel 2 befasst sich mit den nötigen Grundlagen zum Verständnis dieses Manuskripts: Neben der Einführung der verwendeten Notation und der Beschreibung von Definitionen und ausgewählten Konzepten der Ingenieurwissenschaften in Abschnitt 2.1, enthält 2.2 eine kurze Einführung in die Bereiche Molekularbiologie und Biotechnologie. Ausgehend vom sogenannten *Zentralen Dogma der Molekularbiologie* werden einzelne wichtige Teilaspekte der Biologie sowie biotechnologische Werkzeuge und Begriffe erklärt, die für den Themenkomplex der DNA-Sequenzierung hilfreich sind.

Den ersten Teil der Arbeit bildet **Kapitel 3**, das die Verwendung von Watermark Codes als Barcodes umfasst. In Abschnitt 3.1 werden Modelle und Annahmen dargelegt, die eine verallgemeinerte Beschreibung von Barcodes im Kontext der Watermark Codes beinhalten. Hauptaugenmerk liegt auf der Sequenzrepräsentation für Nukleotide und die Anpassung der eingesetzten Hidden Markov-Modelle, ausgehend von [40, 41]. Die optimale Decodierung und probabilistische Suche nach Sequenzen in Abschnitt 3.2 bildet die Motivation für die verwendete Codierung und Decodierung der Barcodes auf Grundlage der Watermark Codes. Die konkrete Umsetzung von Codierschemata für die Erstellung praxistauglicher Barcodes und deren simulative Evaluation bildet den Inhalt von Abschnitt 3.3. Eine Zusammenfassung und kritische Fragen schließen das Kapitel in 3.4 und geben mögliche Anknüpfungspunkte für weiterführende theoretische und praktische Aspekte.

Der zweite große Teilbereich umfasst die Beschreibung und Implementierung von neuartigen Zufallsbarcodes und ihre Anwendung für die RNA-Sequenzierung, dargelegt in **Kapitel 4**. Neben dem Entwurf und der experimentellen Erzeugung der Zufallsmoleküle in Abschnitt 4.1, werden in Abschnitt 4.2 notwendige Aspekte der Bioinformatik beschrieben, die durch die neuen Rahmenbedingungen des gezeigten Verfahrens bedingt sind. Die Validierung und Analyse der experimentellen Umsetzung des vorgestellten Konzepts bilden den Kern von Abschnitt 4.3. Zusätzlich zur Auswertung der beobachteten Zufallsmoleküle, der Analyse der Diversität und deren Auswirkung für die Zählung von Molekülen, werden zwei interessante Teilaspekte näher beschrieben, die sich durch den Einsatz der Zufallsbarcodes zeigen: Zum einen ist dies der Effekt der Selbst-*Hybridisierung* mit möglichen Implikationen für eine Weiterentwicklung der Methodik, zum anderen die vermehrte Sequenzierung von RNA-Fragmenten mit deren nativen 3'-Molekülenden. Der gezeigte Ansatz zur Detektion von signifikanten Häufungen von modifizierten 3'-Enden in den Sequenzierdaten bildet den Abschluss des methodischen Teils. Mit weiterführenden Themen im Kontext der vorgestellten Zufallsbarcodes schließt dieses Kapitel in 4.4.

Eine abschließende Zusammenfassung bildet **Kapitel 5**. Zusätzlich werden die dargelegten Konzepte in den momentanen Stand der Technik im Bereich der Sequenzierung von DNA und RNA eingestuft und ein möglicher Ausblick für die aktuell voranschreitenden Entwicklungen gegeben.

2

Grundlagen

Interdisziplinäres Arbeiten erfordert eine grundlegende Basis aus Zusammenhängen und Methoden unterschiedlicher Forschungsgebiete. Die Arbeit im Bereich der *Barcodes* für die Sequenzierung von DNA und RNA benötigt daher sowohl Konzepte der Ingenieurwissenschaften, als auch ein Basiswissen aus der Molekularbiologie und Biotechnologie, um wesentliche Problemstellungen zu erkennen und adäquate Lösungsansätze finden zu können.

Das folgende Kapitel umfasst elementare Zusammenhänge und Definitionen aus dem Bereich der Stochastik und der Statistik, dargelegt in Abschnitt 2.1.1, Grundlagen der Informationsübertragung und Kanalcodierung in 2.1.2 und wichtige Themen bezüglich der Sequenzähnlichkeit und dem Bereich der musterbasierten Suche, zusammengefasst in 2.1.3.

Der zweite Abschnitt umfasst einen grundlegenden Zugang zur Informationsspeicherung und Übertragung in der Biologie sowie ausgewählte Aspekte der Technologie der Molekularbiologie: Dafür wird das *Zentrale Dogma der Molekularbiologie* eingeführt und in dessen Kontext die elementaren Sequenzen der Biologie in einen groben Zusammenhang gestellt. In Abschnitt 2.2.2 werden in der Arbeit verwendete biotechnologischen Begriffe und Werkzeuge vereinfacht beschrieben. Die Einsatzmöglichkeiten von Barcodes werden abschließend in 2.2.3 durch die Erläuterung der Sequenzierung von DNA und RNA motiviert.

2.1 Konzepte der Ingenieurwissenschaften

Die Beschreibung von mathematischen Zusammenhängen und Konzepten macht eine einheitliche Notation unabdingbar. Deshalb werden zu Beginn einige Konventionen festgehalten.

Zur Unterscheidung der unterschiedlichen Klassen an Zahlen bezeichnet $\mathbb{R}$ die Menge der reellen Zahlen, $\mathbb{Z}$ alle ganzen Zahlen und $\mathbb{N}$ die natürlichen Zahlen. Weitere endliche algebraische Strukturen mit q Werten sind die Restklassenringe $\mathbb{Z}_q$ (ganzer Zahlen $\mathbb{Z}$ modulo q) und die Primkörper (Galois-Felder) $\mathbb{F}_q$, mit Primzahlpotenz q. Explizit sind Mengen durch geschweifte Klammern gekennzeichnet, z. B. $\mathbb{Z}_q = \{0, 1, 2, .., q-1\}$, während $\emptyset$ die leere Menge darstellt. Symbolisiert werden Mengen durch kalligrafische Buchstaben wie $\mathcal{A}$, wobei $|\mathcal{A}|$ deren Kardinalität anzeigt. Während Operatoren bzw. Funktionale durch nicht-kursive Bezeichner definiert werden, sind veränderliche Symbole durch kursive Schreibweise zu erkennen. Im Kontext von Zufallsvariablen und Wahrscheinlichkeiten bekommt die Orthografie von Symbolen eine erweiterte Bedeutung: Großbuchstaben wie X werden für Zufallsvariablen verwendet, Kleinbuchstaben wie x für deren konkrete Realisierung. Zur Darstellung von Folgen (Sequenzen) dient im Allgemeinen eine Abbildung $x : \mathbb{Z} \to \mathcal{A}, i \mapsto x_i$, wobei einem Index i ein Zahlenwert $x_i \in \mathcal{A}$ zugewiesen wird. Se-

quenzen werden in vektorisierter Form durch fett gedruckte Symbole dargestellt: Eine Sequenz der Länge n ist beispielsweise $\mathbf{x} = x_1 x_2 .. x_n$. Zur besseren Übersicht werden gegebenenfalls Kommata und Klammern ergänzt, so z. B. für $(x_1, x_2, .., x_n)$. Dabei ist $\mathbf{x} \in \mathcal{A}^n = \mathcal{A} \times \mathcal{A} \times \cdots \times \mathcal{A}$ ein Element aller möglichen n-dimensionalen Sequenzen über dem Alphabet einer Menge $\mathcal{A}$. Für die Notation in fett gedruckten Großbuchstaben existieren drei Bedeutungen, welche aus dem Kontext eindeutig ersichtlich sind: In Bezug auf Wahrscheinlichkeiten ist $\mathbf{X}$ eine mehrdimensionale Zufallsvariable oder es bezieht sich auf eine herkömmliche Sequenz von Symbolen und ist damit äquivalent zur Notation $\mathbf{x}$. Des Weiteren ist die Notation als Matrix $\mathbf{X} = \{X_{i,j}\}$ gebräuchlich, in welcher $X_{i,j}$ die Werte der i-ten Zeile und j-ten Spalte referenzieren. Eine verkürzte Schreibweise als X_{ij} ist ebenfalls gebräuchlich. An einigen Stellen der vorliegenden Arbeit ist es für die Notation von Vorteil auf zusätzliche Buchstaben zu verzichten, um die Lesbarkeit von Formalismen zu verbessern. Aus diesem Grund tritt gelegentlich das Symbol $*$ als Platzhalter auf, wenn der mögliche Wertebereich an der eingesetzten Stelle klar ersichtlich ist.

2.1.1 Wahrscheinlichkeitstheorie und Statistik

Die folgenden ausgewählten Aspekte basieren (falls nicht näher spezifiziert) auf Erklärungen in Standardwerken, wie [28, 37, 49, 132, 133].

Grundlagen, Allgemeines

Im Sinne der diskreten Wahrscheinlichkeitsrechnung beschreiben $X, Y \in \mathcal{A}$ zwei beispielhafte Zufallsvariablen, welche Werte aus einem Alphabet $\mathcal{A}$ annehmen können. Die konkreten Werte der Zufallsvariablen repräsentieren eine Abbildung (Bewertung) für den Ausgang eines zugrunde liegenden Zufallsexperiments: Betrachtet man X und Y, so beschreiben $\Pr(X = x)$ und $\Pr(Y = y)$ unabhängig voneinander die Wahrscheinlichkeiten, dass konkrete Werte x, y auftreten (realisiert) werden. Mit $\Pr(X = x, Y = y)$ ist die Verbundverteilung definiert. Für die Randverteilungen gilt

$$\Pr(X = x) = \sum_{y \in \mathcal{A}} \Pr(X = x, Y = y) \text{ bzw. } \Pr(Y = y) = \sum_{x \in \mathcal{A}} \Pr(X = x, Y = y).$$

Die Werte der Verteilungen sind stets größer null und summieren sich für alle Werte des Alphabets zu eins. Sind X und Y voneinander stochastisch unabhängig, so gilt

$$\Pr(X = x, Y = y) = \Pr(X = x) \cdot \Pr(Y = y).$$

Für existierende Abhängigkeit kann, falls $\Pr(X = x) > 0$, zusätzlich die bedingte Wahrscheinlichkeit

$$\Pr(Y = y | X = x) = \frac{\Pr(X = x, Y = y)}{\Pr(X = x)} = \frac{\Pr(X = x | Y = y) \Pr(Y = y)}{\Pr(X = x)}$$

definiert werden, einen Wert y zu erhalten unter der Bedingung, dass auch x präsent ist. Bei der Definition entspricht die letzte Gleichheit dem Satz von Bayes. Die Rolle von X und Y ist in diesem Zusammenhang austauschbar.

Definition 1 (Erwartungswert, Kovarianz, Varianz, Variationskoeffizient)**:** Der Erwartungswert einer diskreten Zufallsvariablen $Z \in \mathcal{A}$ wird berechnet als

$$\mathrm{E}(Z) = \sum_{z \in \mathcal{A}} z \cdot \Pr(Z = z).$$

Für die Summe und das Produkt von n stochastisch unabhängigen Z_i mit $i \in \{1, 2, .., n\}$ gilt

$$\mathrm{E}\left(\sum_i Z_i\right) = \sum_i \mathrm{E}(Z_i) \text{ und } \mathrm{E}\left(\prod_i Z_i\right) = \prod_i \mathrm{E}(Z_i).$$

Basierend auf dem Erwartungswert existiert für zwei Zufallsvariablen X, Y die Kovarianz

$$\mathrm{Kov}(X, Y) = \mathrm{E}\Big([X - \mathrm{E}(X)] \cdot [Y - \mathrm{E}(Y)]\Big) = \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y).$$

Für die Varianz einer Variable Z gilt $\mathrm{Var}(Z) = \mathrm{Kov}(Z, Z)$. Betrachtet man die Summe von n Variablen Z_i, so ergibt sich für die Varianz

$$\mathrm{Var}\left(\sum_i Z_i\right) = \sum_{i,j} \mathrm{Kov}(Z_i, Z_j),$$

welche sich bei paarweiser Unkorreliertheit für $\{Z_i\}$, wegen $\mathrm{Kov}(Z_i, Z_j) = 0$, für $i \neq j$, vereinfachen lässt zu $\mathrm{Var}(\sum_i Z_i) = \sum_i \mathrm{Var}(Z_i)$. Für eine Zufallsvariable Z mit $\mathrm{E}(Z) \neq 0$ existiert die relative Standardabweichung

$$\mathrm{VarK}(Z) = \frac{\sqrt[2]{\mathrm{Var}(Z)}}{\mathrm{E}(Z)},$$

auch Variationskoeffizient genannt. Dessen empirischer Wert findet in der deskriptiven Statistik als relatives Streuungsmaß Verwendung, um eine Aussage über die Variabilität von Messgrößen zu treffen.

Das kleinste nicht-triviale Alphabet, welches sich zur Beschreibung einer Zufallsvariable eignet, hat die Größe $|\mathcal{A}| = 2$ und umfasst nur zwei dichotome (komplementär ergänzende) Werte. Die Verteilung die sich mit zwei Werten beschreiben lässt, ist die Bernoulli-Verteilung, aus welcher sich weitere Verteilungen ableiten lassen. Zur einfacheren Beschreibung gilt im Folgenden letztlich $\mathcal{A} = \{0, 1\}$.

Definition 2 (Bernoulli-Verteilung)**:** Dichotome Zufallsvariablen $X \in \{0, 1\}$ mit Bernoulli-Verteilung werden charakterisiert durch einen Parameter $p = \Pr(X = 1)$, der sogenannten Wahrscheinlichkeit für einen *Erfolg*. Die Verteilung von X ist definiert durch

$$\Pr(X = x) = p^x (1 - p)^{1-x}, \text{ mit } x \in \{0, 1\}.$$

Des Weiteren gilt $\mathrm{E}(X) = p$ und $\mathrm{Var}(X) = p(1 - p)$.

Basierend auf der Bernoulli-Verteilung existiert die Binomialverteilung.

Definition 3 (Binomialverteilung)**:** Ist $K = \sum_{i=1}^{n} X_i$ die Summe aus n Zufallsvariablen einer Bernoulli-Verteilung mit einem Parameter p, so folgt K einer Binomialverteilung mit

$$\Pr(K = k) = \binom{n}{k} p^k (1-p)^{n-k}, \text{ für } 0 \leq k \leq n.$$

Des Weiteren gilt $\mathrm{E}(K) = np$ und $\mathrm{Var}(K) = np(1-p)$.

Für einige Anwendungen, insbesondere Beweisführungen der Wahrscheinlichkeitstheorie, sind die äußeren Bereiche der Binomialverteilung von direktem Interesse. Für die (inverse) kumulative Verteilung

$$\Pr(K \geq k) = \sum_{\kappa=k}^{n} \Pr(K = \kappa)$$

sind beispielsweise explizite Schranken bekannt (vgl. hierzu [7, 37]).

Definition 4 (Schranken der Binomialverteilung)**:** Für gegebenen Parameter p einer Binomialverteilung und Werte $0 < k < n$ sei $\hat{p} = k/n$, und es gilt

$$(n+1)^{-2} \cdot \mathrm{e}^{-n \cdot \mathrm{d_{KL}}(\hat{p}||p)} \leq \Pr(K \geq k) \leq \mathrm{e}^{-n \cdot \mathrm{d_{KL}}(\hat{p}||p)}$$

für $0 < p < \hat{p} < 1$. Dabei ist $\mathrm{d_{KL}}(\hat{p}||p) = p \ln(p/\hat{p}) + p' \ln(p'/\hat{p}')$ die Kullback-Leibler-Divergenz mit $p' = 1 - p$ und $\hat{p}' = 1 - \hat{p}$.

Die Bernoulli-Verteilung ist auch in weiteren Verteilungen wiederzufinden, so z. B. in der Geometrischen Verteilung.

Definition 5 (Geometrische Verteilung)**:** Ist eine Bernoulli-verteilte Zufallsvariable die entscheidende Größe für eine wiederholte Evaluation der Zufallsgröße bis zum Erfolgsereignis, so folgt die Zufallsvariable L, als Anzahl an Wiederholungen bis zum Erfolg, einer Geometrische Verteilung mit

$$\Pr(L = l) = (1-p)^l p, \text{ für } 0 \leq l.$$

Es gilt $\mathrm{E}(L) = (1-p)/p$ und $\mathrm{Var}(L) = (1-p)/p^2$.

Ausgehend von der eindimensionalen Binomialverteilung existiert eine Verallgemeinerung für mehrere Dimensionen, welche Multinomialverteilung genannt wird. Hierzu werden verschiedene Typen dichotomer Ereignisse betrachtet und die Anzahl der Erfolge für diese Typen genauer unterschieden.

Definition 6 (Multinomialverteilung)**:** Sei n die Anzahl aller Erfolge aus m unterschiedlichen Typen von dichotomen Ereignissen mit Erfolgswahrscheinlichkeiten $p_1, p_2, .., p_m$, und zusätzlich $\sum_{i=1}^{m} p_i = 1$, so beschreibt

$$\Pr(K_1 = k_1, K_2 = k_2, .., K_m = k_m) = \frac{n!}{k_1! k_2! \cdots k_m!} p_1^{k_1} p_2^{k_2} \cdots p_m^{k_m}$$

die Wahrscheinlichkeit jeweils k_i Erfolge vom Typ i zu erhalten. Die Zufallsvariablen K_i sind durch die Nebenbedingung $\sum_{i=1}^{m} K_i = n$ stochastisch voneinander abhängig. Für die

Kovarianz gilt

$$\mathrm{Kov}(K_i, K_j) = \begin{cases} -np_ip_j, & \text{für } i \neq j \\ -np_i(1-p_i), & \text{für } i = j. \end{cases}$$

Der letzte Fall entspricht der Varianz von K_i. Der Erwartungswert ist $\mathrm{E}(K_i) = np_i$.

Eine letzte hier beschriebene Verteilung ist die diskrete Uniformverteilung.

Definition 7 (Diskrete Uniformverteilung)**:** Eine diskrete Zufallsvariable $X \in \mathcal{A}$ mit Alphabet der Größe $q = |\mathcal{A}| \geq 2$ ist uniformverteilt (gleichverteilt), falls $\Pr(X = x) = 1/q$ für alle $x \in \mathcal{A}$ erfüllt ist. Die Auftrittswahrscheinlichkeit ist damit für alle Werte der Zufallsgröße identisch.

Diskrete Markov-Ketten und Hidden Markov Modell

Für die stochastische Modellierung von realen Vorgängen kann das Konzept der Zufallsvariable um eine Zeitkomponente erweitert werden. Anstatt nur einer Zufallsvariablen wird eine geordnete Menge (Sequenz) $\{X_t\}$ betrachtet, mit einem zusätzlichen diskreten Zeitparameter $t \in \{0, 1, 2, ..\}$. Im Kontext einer diskreten Definition von X_t beschreiben die möglichen Werte der Zufallsvariablen einen stochastischen Prozess, der sich auf einen endlichen Zustandsraum $\mathcal{X}$ beschränkt. Konkret realisierte Werte $x_t \in \mathcal{X}$ werden durch Übergänge in diesem Zustandsraum beschrieben. Gelten für X_t besondere Einschränkungen der Übergangswahrscheinlichkeiten, so ist eine äquivalente Beschreibung durch eine sogenannte Markov-Kette möglich.

Definition 8 (Markov-Kette)**:** Gilt für einen stochastischen Prozess für alle $t \geq 0$ gleichermaßen

$$\Pr(X_{t+1} = x_{t+1} | X_t = x_t, .., X_0 = x_0) = \Pr(X_{t+1} = x_{t+1} | X_t = x_t),$$

so ist er äquivalent als Markov-Kette beschreibbar. Dieses Kriterium wird auch als Markov-Eigenschaft oder Gedächtnislosigkeit bezeichnet, da der Zustand zum Zeitpunkt $t+1$ nur durch den Wert x_t bedingt ist. Die somit definierten stationären Übergangswahrscheinlichkeiten

$$P_{ji} = \Pr(X_{t+1} = j | X_t = i), \text{ mit } 0 \leq P_{ji} \leq 1 \text{ und } \sum\nolimits_j P_{ji} = 1$$

bestimmen zusammen mit einer Anfangsverteilung $p_{x_0} = \Pr(X_0 = x_0)$ mit $i, j, x_0 \in \mathcal{X}$ die Markov-Kette vollständig. Die Wahrscheinlichkeit einer Sequenz $x_0x_1..x_n \in \mathcal{X}^{n+1}$ ist somit

$$\begin{aligned} \Pr(X_n = x_n, .., X_1 = x_1, X_0 = x_0) &= \prod_{t=0}^{n-1} \Pr(X_{t+1} = x_{t+1} | X_t = x_t) \cdot \Pr(X_0 = x_0) \\ &= P_{x_n x_{n-1}} \cdots P_{x_2 x_1} P_{x_1 x_0} p_{x_0}, \end{aligned}$$

was einem möglichen Pfad einer Kette entspricht.

Eine Erweiterung des Konzepts der Markov-Kette um den Zusatz einer *versteckten* Dimension stellt das Hidden Markov Modell (HMM) dar. Nun folgende Auszüge beziehen sich auf die in [132, 133] gegebene Übersicht und darin enthaltene Definitionen.

Definition 9 (Hidden Markov Modell, HMM)**:** Das Hidden Markov Modell (HMM) besteht aus einem zweigeteilten Zufallsprozess $\{(X_t, Y_t)\}$ mit Zeitparameter t, $X_t \in \mathcal{X}$ und $Y_t \in \mathcal{Y}$. Ein HMM ist somit über einem erweiterten Zustandsraum $\mathcal{X} \times \mathcal{Y}$ definiert, wobei konzeptionell nur Beobachtungen y_t der zweiten Komponente existieren. Die Abfolge der Zustände x_t ist, obwohl sie die Beobachtungen y_t bedingen, nicht sichtbar. Dabei entspricht X_t einem Markov-Prozess und die Markov-Eigenschaft ist hierfür gegeben (vgl. Definition 8). Der kombinierte Zufallsprozess ist jedoch nur durch ein HMM beschreibbar, wenn zusätzlich

$$\Pr(Y_t = y_t | Y_{t-1} = y_{t-1}, .., Y_0 = y_0, X_t = x_t, .., X_0 = x_0) = \Pr(Y_t = y_t | X_t = x_t)$$

gilt, was auch als zweite Markov-Eigenschaft bezeichnet wird. Zusätzlich zu den Übergangswahrscheinlichkeiten $\{P_{ji}\}$ des Prozesses X_t existieren durch die zweite Eigenschaft sogenannte stationäre Emisionswahrscheinlichkeiten

$$Q_{ki} = \Pr(Y_t = k | X_t = i) \text{ mit } 0 \leq Q_{ki} \leq 1 \text{ und } \sum\nolimits_k Q_{ki} = 1 \text{ für } i \in \mathcal{X}, k \in \mathcal{Y}.$$

Nur vom konkreten Zustand i abhängig, ist hierfür $q_i(k) = Q_{ki}$ eine gängige Notation. Die Beschreibung des HMMs wird durch die Definition einer Anfangsverteilung $p_{x_0} = \Pr(X_0 = x_0)$ vervollständigt. Die Parameter des Modells sind $\mathcal{H} = \{P_{ji}, Q_{ki}, p_{x_0}, \mathcal{X}, \mathcal{Y}\}$. Die Wahrscheinlichkeit einer möglichen Realisierung von $\mathbf{x} = x_0 x_1 .. x_n$ und $\mathbf{y} = y_0 y_1 .. y_n$ ist

$$\Pr(\mathbf{Y} = \mathbf{y}; \mathbf{X} = \mathbf{x} | \mathcal{H}) = [q_{x_n}(y_n) P_{x_n x_{n-1}}][q_{x_{n-1}}(y_{n-1}) \cdots P_{x_2 x_1}][q_{x_1}(y_1) P_{x_1 x_0}][q_{x_0}(y_0) p_{x_0}].$$

Ein HMM beschreibt den Zusammenhang der folgenden drei Größen: Der beobachtbare Prozess $\mathbf{Y}$, der verborgene Prozess $\mathbf{X}$ und die Parametrisierung $\mathcal{H}$. Ist $\mathcal{H}$ bekannt, so kann die Wahrscheinlichkeit $\Pr(\mathbf{Y} = \mathbf{y} | \mathcal{H})$ oder $\Pr(X_t = x_t | \mathbf{y}, \mathcal{H})$ für unterschiedliche Fragestellungen von Interesse sein. Für Berechnungen zum HMM liefert die Forward-Backward Prozedur einen effizienten Ansatz.

Konzept 1 (Forward-Backward Prozedur)**:** Der naive Ansatz

$$\Pr(\mathbf{Y} = \mathbf{y} | \mathcal{H}) = \sum_{\mathbf{x} \in \mathcal{X}^n} \Pr(\mathbf{Y} = \mathbf{y}; \mathbf{X} = \mathbf{x} | \mathcal{H})$$

zur Berechnung der Randverteilung lässt sich durch die Markov-Eigenschaften auf die sukzessive Berechnung einer sogenannten Vorwärts-Metrik

$$F_t(j) = \Pr(Y_0 = y_0, Y_1 = y_1, .., Y_t = y_t \ ; \ X_t = j | \mathcal{H})$$

abbilden. Sie entspricht der Wahrscheinlichkeit aller Beobachtungen $y_0 y_1 y_2 .. y_t$ bis zum Zeitpunkt t und dem Aufenthalt in Zustand j. Im Folgenden gelte $i, j, x_0 \in \mathcal{X}$ und $1 \leq t \leq n$. Ausgehend von

$$F_0(x_0) = \Pr(Y_0 = y_0; X_0 = x_0) = [q_{x_0}(y_0) p_{x_0}]$$

erfolgt die induktive Berechnung über

$$F_t(j) = q_j(y_t) \Big[\sum\nolimits_i P_{ji} F_{t-1}(i) \Big].$$

Letztlich ergibt sich $\Pr(\mathbf{y}|\mathcal{H}) = \sum_j F_n(j)$ als Resultat. Zusätzlich zur dieser Wahrscheinlichkeit kann noch die Rückwärts-Metrik

$$B_t(i) = \Pr(Y_{t+1} = y_{t+1}, Y_{t+2} = y_{t+1}, .., Y_n = y_n | X_t = i; \mathcal{H})$$

definiert werden. Sie entspricht der Wahrscheinlichkeit von Beobachtungen $y_{t+1}y_{t+2}..y_n$ nach dem Zeitpunkt t, bedingt durch den Aufenthalt in Zustand i. Ausgehen von $B_n(i) = 1$ für alle $i \in \mathcal{X}$ gibt

$$B_{t-1}(i) = \sum\nolimits_j q_j(y_t) B_t(j) P_{ji}$$

eine induktive Berechnungsvorschrift.

Eine Bewertung der bedingten Aufenthaltswahrscheinlichkeit $\Pr(X_t = x_t|\mathbf{y}, \mathcal{H})$ ist durch die zuvor definierten Metriken wie folgt möglich:

$$\Pr(X_t = x_t|\mathbf{y}, \mathcal{H}) = \frac{F_t(x_t)B_t(x_t)}{\Pr(\mathbf{Y}|\mathcal{H})} = \frac{F_t(x_t)B_t(x_t)}{\sum_{x_t \in \mathcal{X}} F_t(x_t)B_t(x_t)}. \tag{2.1}$$

Stichproben und Statistik

Die mathematische Statistik ist die Anwendung der Wahrscheinlichkeitstheorie auf konkrete Realisierungen von Zufallsereignissen. Als Modell für die Beschreibung von Konzepten soll im Folgenden das Urnen-Modell dienen. Die Urne enthalte m unterschiedliche Typen von Kugeln, unterscheidbar durch ihren Wert $i \in \{1, 2, .., m\}$. Eine Stichprobe vom Umfang n besteht aus Werten $\{x_1, x_2, ..x_n\}$, wobei x_* den konkreten Wert einer Kugel beschreibt. Die Grundgesamtheit der Kugeln sei dabei so groß angenommen, dass die Entnahme einer Kugel keinen Einfluss auf nachfolgende Entnahmen hat.

Definition 10 (Relative Häufigkeit)**:** Sei $k_i = |\{x_* : x_* = i\}|$ die absolute Häufigkeit des Typs i in einer Stichprobe, so entspricht $\hat{p}_i = k_i/n$ der relativen Häufigkeit. Ist X die Zufallsvariable für den Wert einer entnommenen Kugel, so kann $\hat{p}_i$ als Schätzwert für die Wahrscheinlichkeit $p_i = \Pr(X = i)$ verwendet werden.

Die relative Häufigkeit stellt sich jedoch teilweise als schlechter Schätzwert dar: Die Wahrscheinlichkeit von nicht beobachtbaren (unterrepräsentierten) Typen wird, relativ im Verhältnis sichtbarer (überrepräsentierten) Typen der Stichprobe, auf deren Wahrscheinlichkeiten verteilt. Dadurch entsteht eine systematische Unterschätzung von nicht beobachteten und eine Überschätzung von sichtbaren Typen. In [63] wurde der Good-Turing Schätzer als alternative Methode vorgestellt, welche sich mit der Problematik seltener Beobachtungen befasst.

Konzept 2 (Good-Turing Schätzer)**:** Die alternative Repräsentation einer Stichprobe vom Umfang

$$n = \sum_{i=1}^{m} k_i = \sum_{r=0}^{m' \leq m} r n_r$$

durch die Anzahl n_r von Typen (einer Gruppe) mit gleicher Auftrittshäufigkeit r ist der Zugang zum Good-Turing Schätzer. Grundlage für das zentrale Theorem des Schätzers ist

die Annahme eines multinomialen Modells und die Betrachtung der Erwartungswerte

$$\mathrm{E_n}(n_r|\mathcal{H}) = \sum_{i=0}^{m'} \binom{n}{r} p_i^r (1-p_i)^{n-r}$$

unter einer hypothetischen Verteilung $\mathcal{H} = \{p_i\}$. Die Wahrscheinlichkeit $\binom{n}{r} p_i^r (1-p_i)^{n-r}$, dass die in n_r enthaltene Anzahl des Typs i gleich r ist, führt über den Satz von Bayes zur Wahrscheinlichkeit

$$\Pr(p_i' = p_i|\mathcal{H}) = \frac{p_i^r (1-p_i)^{n-r}}{\sum_i p_i^r (1-p_i)^{n-r}},$$

dass die unbekannte relative Häufigkeit p_i' der hypothetischen entspricht. Weiter ergibt sich in [63] der Erwartungswert

$$\mathrm{E}(p_i'|\mathcal{H}) = \frac{r+1}{n+1} \cdot \frac{\mathrm{E_{n+1}}(n_{r+1})}{\mathrm{E_n}(n_r)} \approx \frac{r+1}{n} \cdot \frac{\tilde{n}_{r+1}}{\tilde{n}_r} \to \hat{p}_i,$$

welcher zur Schätzung von Auftrittswahrscheinlichkeiten $\hat{p}_i$ genutzt wird [57]. Ausgehend von $\hat{p}_0 = {}^{n_1}/_n$ berechnet sich $\hat{p}_i$ auf Basis einer Glättung (engl. *smoothing*), welche für jedes r einen Wert $\tilde{n}_r$ aus $\{n_r\}$ interpoliert. Die in dieser Arbeit verwendete Glättung (vgl. [57]) nutzt eine lineare Regression von $\log(r)$ zu $\log(n_r)$. Für Verteilungen, die sich in log-log Darstellung (abschnittsweise) durch eine Gerade annähern lassen, ist diese Glättung ein adäquates Verfahren.

Bei der relativen Häufigkeit sowie der Good-Turing Methode handelt es sich im weitesten Sinn um Stichprobenfunktionen, welche auf Grundlage von Modellannahmen Informationen über unbekannte Parameter des Modells enthalten, d.h sie schätzen diese Parameter. Während erstere Stichprobenfunktion (relativen Häufigkeit) aus einer einfachen Summation besteht, ist die Berechnung der zweiten (Good-Turing) nicht trivial. Dennoch stellen beide auf Basis einer Hypothese $\mathcal{H}$ (der Multinomialverteilung) eine Abbildung $\{x_1, x_2, ..x_n\} \to \{\hat{p}_i\}$ dar. Diese Abbildung wird auch Punktschätzer genannt.

Eine zweite Art Stichprobenfunktionen ist die Teststatistik (auch Prüffunktion genannt), welche im Rahmen eines Hypothesentests verwendet wird, um eine Hypothese $\mathcal{H}$ (oder mehrere) anhand einer Stichproben zu überprüfen.

Konzept 3 (Hypothesentests)**:** Ein Hypothesentest gliedert sich wie folgt:

- Erstens, die Formulierung einer Hypothese, z. B. der sogenannten Null-Hypothese $\mathcal{H}_0$: „Die Grundgesamtheit an Kugeln (einer Urne) ist dichotom und Bernoulli-verteilt mit Parameter $p = p_0$“. Alternativ wird meist eine Hypothese H_1 formuliert, die beispielsweise $p > p_0$ postuliert. Der Test prüft letztlich die Widerlegung von $\mathcal{H}_0$.
- Formulierung einer adäquate Teststatistik, als konkrete Zufallsvariable, die eine Entscheidung über $\mathcal{H}_0$ erlaubt: So z. B. $K = \sum_{j=1}^{n} X_j$, als Anzahl von Erfolge für eine zufällige Stichprobe $\{X_j\}$. Dabei ist K eine Zufallsvariable und k ist die zugehörige Realisierung für konkrete Ereignisse $\{x_1, x_2, .., x_n\}$.
- Festlegung eines Signifikanzniveaus α (beispielsweise .05), was der vorab akzeptierten Wahrscheinlichkeit entspricht die Hypothese $\mathcal{H}_0$ fälschlicherweise abzulehnen.
- Erfassen der eigentlichen Stichprobe $\{x_1, x_2, ..x_n\}$.

- Integration der Teststatistik in eine bedingte Wahrscheinlichkeit: Für die gegebene Binomialverteilung gilt

$$\text{p-Wert}[\mathcal{H}_0] := \Pr(K \geq k|\mathcal{H}_0) = \sum_{\kappa=k}^{n} \binom{n}{\kappa} p^{\kappa}(1-p)^{n-\kappa}$$

  als Wahrscheinlichkeit, dass eine zufällige Evaluation K der Teststatistik unter $\mathcal{H}_0$ größere Werte annimmt als das k der konkreten Stichprobe $\{x_1, x_2, ..x_n\}$.
- Fällung einer Entscheidung: Eine Hypothese $\mathcal{H}_0$ kann signifikant (unter Signifikanzniveaus α) abgelehnt werden, falls p-Wert$[\mathcal{H}_0] \leq \alpha$ gilt.

Oft wird eine gemachte Stichprobe (durchgeführtes Experiment) für mehrere Hypothesentests verwendet. Dabei zeigt sich folgender Sachverhalt: Je mehr Tests auf einem Datensatz erfolgen, je größer wird die Wahrscheinlichkeit, dass eine der Null-Hypothesen widerlegbar wird, obwohl sie zutreffend ist (vgl. [1, 150]): Dabei ist das Ergebnis nicht zwangsläufig signifikant, sondern unter Umständen nur eine (absehbare) statistische Schwankung. Das zugehörige Phänomen wird Alphafehler-Kumulierung genannt, da die Wahrscheinlichkeit zum Fehlschluss durch jeden Test vergrößert wird. Eine einfache, aber auch sehr konservative Methode diesem Problem zu begegnen, beinhaltet die Bonferroni-Korrektur.

Definition 11 (Bonferroni-Korrektur)**:** Seien $\{\mathcal{H}_{0,a}\}$ für $a \in \{1, 2, .., A\}$ eine Reihe von Null-Hypothesen, die durch mehrfache Anwendung in einer Teststatistik zu p-Werten führt, so ist das angepasste Signifikanzniveau $\alpha_A^* = \alpha/A$ mit Vergleich p-Wert$[\mathcal{H}_{0,a}] \leq \alpha_A^*$ ein konservatives Kriterium zur Ablehnung der Null-Hypothesen unter der Bonferroni-Korrektur. Begründet ist das Kriterium durch folgende Überlegung: Sei $\mathcal{K}$ die Menge aller korrekten Hypothesen und E_a stellvertretend für ein Ereignis p-Wert$[\mathcal{H}_{0,a}] \leq \alpha_A^*$, so gilt

$$\Pr\Big(\bigcup_{a\in\mathcal{K}} E_a\Big) \leq \sum_{a\in\mathcal{K}} \Pr(E_a) \leq |\mathcal{K}| \cdot \alpha_A^* \leq A \cdot \alpha_A^* = \alpha.$$

Linksseitig steht die Wahrscheinlichkeit mindestens eine Hypothese aus $\mathcal{K}$ fälschlicherweise abzulehnen. Dabei ist die Wahrscheinlichkeit der Vereinigung der Ereignisse E_a kleiner als die Summe aller Einzelwahrscheinlichkeiten (Boolesche Ungleichung). Unter Einhaltung der Korrektur ist auf jeden Fall gewährleistet, dass der beschriebene Fehler kleiner als α ist. Dabei liefert α_A^* die konservativste (aller möglichen) Restriktionen.

2.1.2 Informationsübertragung und Kanalcodierung

Im folgenden Abschnitt werden grundlegende Konzepte der Informationsübertragung und Kanalcodierung dargestellt. Bei den Erklärungen handelt es sich um eine Auswahl an Aspekten, die detailliert in Standardwerken wie [22, 107, 140] zu finden sind. Für umfangreichere Informationen zum Themenkomplex Kanalcodierung wird auf die erwähnten Quellen verwiesen. Des Weiteren wird auf Ausführungen zu expliziten Codekonstruktionen und detaillierte Beschreibungen zur Decodierung verzichtet.

.3Für die Beschreibung der Konzepte bietet sich ein Modell der Übertragung an, welches in Abb. 2.1 dargestellt ist (vgl. dazu [140]).

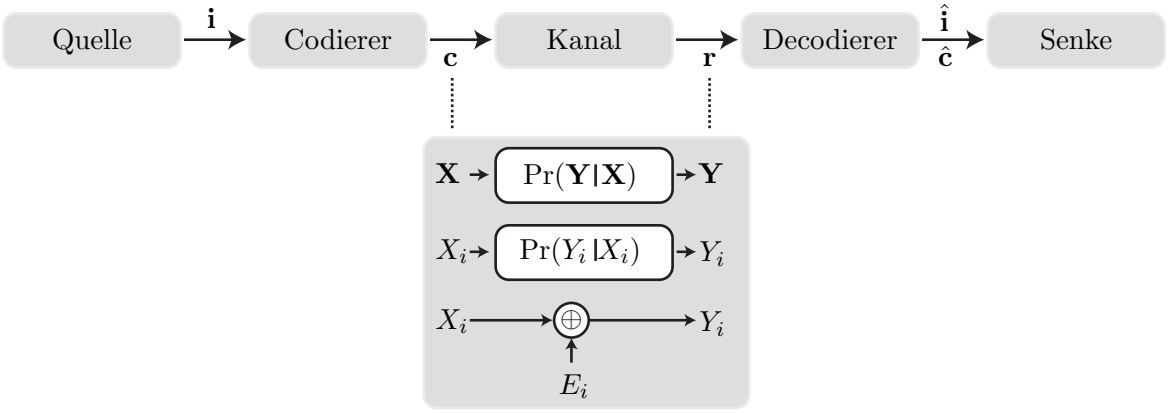

Abb. 2.1 Übertragungsmodell der Kanalcodierung (oben, nach [22]). Kanalmodelle (unten): Allgemeiner probabilistischer Kanal, Kanal ohne Gedächtnis, additiver q-wertiger Kanal.

Konzept 4 (Modell des diskreten probabilistischen Kanals)**:** Ursprung einer Information, welche über einen Kanal übertragen wird, ist die Quelle: Sie erzeugt zufällig eine Sequenz, das Informationswort $\mathbf{i} = i_1 i_2 ... i_k \in \mathcal{A}_\mathbf{i}^k$. Dabei beschreibt $\mathcal{A}_\mathbf{i}$ das Informationsalphabet und M sei die konkrete Anzahl gültiger Informationsworte $\{\mathbf{i}_1, \mathbf{i}_2, .., \mathbf{i}_M\}$ der Länge k. Der Codierer bildet $\mathbf{i}$ auf ein Codewort $\mathbf{c} = c_1 c_2 ... c_n \in \mathcal{A}_\mathbf{c}^n$ ab, welches über den probabilistischen Kanal übertragen wird. Die Codeworte $\{\mathbf{c}_1, \mathbf{c}_2, .., \mathbf{c}_M\} = \mathcal{C} \subseteq \mathcal{A}_\mathbf{c}^n$ sind dabei eine Teilmenge aller möglichen Sequenzen der Länge n. Die Rate des Codierers ist definiert als Kenngröße $R = {}^{(\log_{|\mathcal{A}_\mathbf{c}|} M)}/_n \leq 1$. Der Kanal erzeugt das Empfangswort $\mathbf{r} = r_1 r_2 ... r_n \in \mathcal{A}_\mathbf{r}^n$ auf Basis einer bedingten Wahrscheinlichkeit

$$\Pr(\mathbf{Y} = \mathbf{r} \mid \mathbf{X} = \mathbf{c}),$$

als weitere probabilistische Komponente des Modells. Die Größen $\mathbf{X}, \mathbf{Y}$ sind vektorwertige Zufallsvariablen und bilden eine wahrscheinlichkeitstheoretische Beschreibung von Symbol-Ersetzungen. Der Decodierer entscheidet letztlich auf Grundlage von $\mathbf{r}$, welches der Codeworte $\hat{\mathbf{c}} \in \mathcal{C}$ vor dem Kanal aufgetreten ist und schätzt somit ebenfalls die ursprüngliche Information $\hat{\mathbf{i}}$. Eine Senke vervollständigt das Bild.

Auf Grundlage der Übergangswahrscheinlichkeit $\Pr(\mathbf{Y}|\mathbf{X})$ lässt sich gegebenenfalls die Kanalkapazität explizit formulieren. Ohne tiefer auf die Theorie einzugehen, ist einer der elementarsten Grundpfeiler der Kanalcodierung, dass eine fehlerfreie Übertragung von Information (prinzipiell) möglich ist, solange die Rate des Übertragungssystems kleiner als die Kanalkapazität ist. Das *Kanalcodiertheorem* [151] liefert hierzu den elementaren Beweis.

Ausgehend von der sehr allgemein formulierten Übergangswahrscheinlichkeit lassen sich konkrete Realisierungen von Kanalmodellen ableiten.

Definition 12 (Diskreter Kanal ohne Gedächtnis)**:** Für einen gedächtnislosen Kanal mit Alphabeten $\mathcal{X}, \mathcal{Y}$ für Symbole am Ein- bzw. Ausgang gilt

$$\Pr(Y_i = y_i | \mathbf{X} = x_1 x_2 .. x_n) = \Pr(Y_i = y_i | X_i = x_i), \text{ für alle } i \in \{1, 2, .., n\}$$

mit $x_i, X_i \in \mathcal{X}$ und $y_i, Y_i \in \mathcal{Y}$. Des Weiteren gilt $\Pr(\mathbf{Y} = \mathbf{y} | \mathbf{X} = \mathbf{x}) = \prod_i \Pr(Y_i = y_i | X_i = x_i)$.

Besteht für Alphabete eine Beschränkung $\mathcal{X} = \mathcal{Y} = \{0, 1, 2, .., q-1\} = \mathcal{A}_q$ auf q ganzzahlige Werte, so ist eine Beschreibung durch Übergangsmatrizen $P_{rc} = \Pr(Y = r|X = c)$ mit $c, r \in \mathcal{A}_q$ möglich.

Definition 13 (Symmetrischer q-wertiger Kanal)**:** Für einen q-wertigen symmetrischen Kanal gilt

$$P_{rc} = \Pr(Y = r|X = c) = \begin{cases} 1 - p_e, & \text{für } c = r, \\ p_e/(q-1), & \text{für } c \neq r, \end{cases}$$

mit $c, r \in \mathcal{A}_q$ und p_e als gleiche Wahrscheinlichkeit für eine Symbolersetzung, unabhängig von der Position und dem Symbol c. Eine sinnvolle Beschreibung ist gegeben für $0 \leq p_e < 1/2$.

Oft vereinfacht die Ausnutzung der algebraischen Strukturen der Alphabete die Beschreibung des Kanals, beispielsweise für den additiven q-wertigen Kanal.

Definition 14 (Additiver q-wertiger Kanal)**:** Wählt man für das gemeinsame Alphabet des Kanals $\mathcal{A}_q = \mathbb{Z}_q$ als Restklassenring modulo[1] q (kurz mod q), so ist der Einfluss des Kanals als komponentenweise Addition einer Zufallsgröße beschreibbar (siehe Abb. 2.1). Für konkrete Sequenzen $\mathbf{c}, \mathbf{r}$ ergeben sich die Ersetzungen $r_i \to c_i$ durch $e_i \equiv r_i \ominus c_i$, wobei $\mathbf{e} \equiv \mathbf{r} \ominus \mathbf{c}$ als Fehlerwort bezeichnet wird. Der symmetrische Kanal ist beispielsweise charakterisiert durch $\Pr(E_i \not\equiv 0) = p_e$ und damit unabhängig vom Eingangssymbol. Die Stellen im Fehlerwort $\mathbf{e} \equiv e_1 e_2 .. e_n$ ungleich null zeigen dabei die Positionen (und Art) der Ersetzungen an.

Erweiterte algebraische Strukturen sind eine essentielle Grundlage zur Beschreibung von Codes und den damit verbundenen Berechnungen. Die sogenannten Primkörper sind dazu ein elementarer Bestandteil.

Definition 15 (Primkörper)**:** Mit dem Wert q als Primzahl (oder Primzahlpotenz) ist die Menge $\mathbb{F}_q = \{0, 1, .., q-1\}$ und zugehörigen Rechenoperationen ein Primkörper, auch *Galois-Feld* genannt. Die Menge der Elemente mit den Rechenoperationen Addition $\oplus$ und Multiplikation $\odot$ modulo q folgen den Axiomen eines Körpers [28]: Sowohl für $\oplus$ als auch $\odot$ existieren neutrale Elemente, d. h.

$$\forall x \in \mathbb{F}_q \Rightarrow (x \oplus 0), (x \odot 1) \in \mathbb{F}_q.$$

Es gilt

$$\forall x \in \mathbb{F}_q \Rightarrow x \oplus x^{-1} \equiv 1 \text{ und } \forall x \in \mathbb{F}_q \setminus \{0\} \Rightarrow x \odot x^{-1} \equiv 1$$

für inverse Elemente x^{-1}. Neben der Abgeschlossenheit, $\forall x, y \in \mathbb{F}_q \Rightarrow (x \oplus y), (x \odot y) \in \mathbb{F}_q$, gelten Kommutativ-, Assoziativ- und Distributivgesetz.

Eine mögliche Form zur Beschreibung von Codes sind die sogenannten Blockcodes.

Definition 16 (Blockcode)**:** Ein Blockcode $\mathcal{C}(\mathbb{F}_q, n, k, d_{\min})$ ist gekennzeichnet durch eine grundlegende algebraische Struktur, beispielsweise dem Galois-Feld $\mathbb{F}_q$, und zusätzlichen Parametern: Die Menge der Codeworte $\mathcal{C} \subseteq \mathbb{F}_q^n$ beinhaltete dabei q^k Elemente der Länge n. Die *Dimension* des Codes ist definiert durch k, n wird auch als *Codelänge* bezeichnet. Die Rate des Codes ist $R = k/n$ (vgl. Konzept 4). Der letzte Parameter ist die *Mindestdistanz* $d_{\min}$,

[1] Rechenoperationen mit Modul sind durch $\equiv$ (im Vergleich zu $=$) als Zuweisung gekennzeichnet

welche sich auf die kleinste Anzahl verschiedener Stellen bezieht, die zwei beliebige Codeworte voneinander unterscheidet.

Das der Mindestdistanz von Codes zugrundeliegende Maß wird auch *Hamming-Abstand* genannt.

Definition 17 (Hamming-Abstand und Gewicht)**:** Der Hamming-Abstand zwischen zwei Sequenzen $\mathbf{x}, \mathbf{y}$ der Länge n ist definiert als

$$\mathrm{d_h}(\mathbf{x}, \mathbf{y}) = \Big|\Big\{i \in \{1, 2, .., n\} : x_i \neq y_i\Big\}\Big|. \tag{2.2}$$

Basierend darauf ist das *Hamming-Gewicht* einer Sequenzen $\mathbf{x}$ definiert als $\mathrm{w_h}(\mathbf{x}) = \mathrm{d_h}(\mathbf{x}, \mathbf{0})$, wobei $\mathbf{0}$ die Null-Sequenz darstellt. Ferner gilt $\mathrm{w_h}(\mathbf{x} + \mathbf{y}) = \mathrm{d_h}(\mathbf{x}, \mathbf{y})$.

Es existieren weitere Kriterien, die einen Code klassifizieren, wie z. B. die Linearität.

Definition 18 (Linearer Code)**:** Ein Code heißt linear, falls für beliebige $\mathbf{c}, \mathbf{c}' \in \mathcal{C}$ und alle $i \in \mathbb{F}_q$ gilt, dass $(\mathbf{c} + i \cdot \mathbf{c}') \in \mathcal{C}$, d. h. eine Linearkombination von Codeworten stets ein Codewort ergibt.

Für lineare Codes gilt somit $\mathbf{0} \in \mathcal{C}$. Ferner gilt Gleichheit von minimalem Abstand und Gewicht:

$$d_{\min} = \min_{\mathbf{c}_i \neq \mathbf{c}_j} \{\mathrm{d_h}(\mathbf{c}_i, \mathbf{c}_j)\} = \min_{\mathbf{c}} \{\mathrm{w_h}(\mathbf{c})\}, \text{ mit } \mathbf{c}_i, \mathbf{c}_j, \mathbf{c} \in \mathcal{C}. \tag{2.3}$$

Eine Möglichkeit die Struktur eines Codes zu umschreiben bietet die folgende Eigenschaft.

Definition 19 (Distanz-Verteilung)**:** Die Distanz-Verteilung eines Codes $\mathcal{C}$ bezüglich einer Metrik d entspricht den relativen Häufigkeiten

$$D_d = \frac{1}{|\mathcal{C}|}\Big|\Big\{(i, j) : \mathrm{d}(\mathbf{c}_i, \mathbf{c}_j) = d,\ \mathbf{c}_i, \mathbf{c}_j \in \mathcal{C}\}\Big\}\Big|.$$

Dabei ist d der Distanzoperator und d ein konkreter Wert. Insbesondere gilt $D_0 = 1$ und zusätzlich $\sum_d D_d = |\mathcal{C}|$. Die vorliegende Beschreibung verallgemeinert die Definition in [107] für beliebige Metriken d.

Eine sinnvolle Interpretation der Häufigkeiten D_d ist die erwartete Anzahl von Codeworten mit Abstand d zu einem zufällig ausgewählten Codewort (uniforme Verteilung für $\mathcal{C}$ vorausgesetzt). Eine besondere Rolle spielt die Distanz-Verteilung für die sogenannten nicht-linearen Codes. Denn für lineare Codes ist die Verteilung der Hamming-Abstände und Gewichte äquivalent. Die Mindestdistanz

$$d_{\min} = \arg\min_{d>0} \{d : D_d \neq 0\}$$

ist die minimale Separierung von Codeworten. Der Abstand gibt somit auch eine Information darüber, wie umfangreich die Anzahl an Fehlern sein darf, dass eine eindeutige Decodierung garantiert ist.

Definition 20 (Korrekturfähigkeit)**:** Für einen Code $\mathcal{C}$ mit Mindestdistanz $d_{\min}$ (Hamming-Abstand) können $t \leq \lfloor (d_{\min}-1)/2 \rfloor$ Ersetzungen eindeutig korrigiert werden, wohingegen mindestens $d_{\min} - 1$ Ersetzungen detektierbar sind (vgl. [22]). Dabei entspricht $\lfloor . \rfloor$ der Abrundung[2].

Ohne konkret auf die Konstruktionen einzelner Varianten von Blockcodes einzugehen, existiert die Möglichkeit gegebene Codes (als Abbildung verstanden) miteinander zu verknüpfen, um beispielsweise Codeeigenschaften zu kombinieren. Die sogenannte *Codeverkettung* beschreibt das grundlegende Prinzip dazu.

Konzept 5 (Einfache Codeverkettung)**:** Betrachtet man zwei Codes ganz allgemein als Abbildungen

$$\mathcal{C}_1(n_1,k_1,d_1) : \mathcal{A}_{\mathbf{i}_1}^{k_1} \rightarrow \mathcal{A}_{\mathbf{c}_1}^{n_1},\ \mathbf{i}_1 \mapsto \mathbf{c}_1 \quad \text{und} \quad \mathcal{C}_2(n_2,k_2,d_2) : \mathcal{A}_{\mathbf{i}_2}^{k_2} \rightarrow \mathcal{A}_{\mathbf{c}_2}^{n_2},\ \mathbf{i}_2 \mapsto \mathbf{c}_2$$

von einem Informationswort $\mathbf{i}_* \in \mathcal{A}_{\mathbf{i}_*}^{k_*}$ auf ein Codewort $\mathbf{c}_* \in \mathcal{A}_{\mathbf{c}_*}^{n_*}$ mit beliebigen Alphabeten, so ist eine einfache Verkettung $\mathcal{C}_\mathrm{v} = \mathcal{C}_1 \circ \mathcal{C}_2$ der Codes möglich, falls $|\mathcal{A}_{\mathbf{c}_1}| = |\mathcal{A}_{\mathbf{i}_2}^{k_2}|$ gilt. Bei der Verkettung wird die erste Abbildung $\mathcal{C}_1$ als *äußerer Code* bezeichnet, $\mathcal{C}_2$ nennt man *innerer Code*. Betrachtet man ein Codewort $(\mathcal{C}_2(\tilde{c}_1)\mathcal{C}_2(\tilde{c}_2)..\mathcal{C}_2(\tilde{c}_{n_1})) \in \mathcal{C}_\mathrm{v}$, so besteht dieses aus der komponentenweisen Codierung der Codesymbole $c_j \in \mathcal{A}_{\mathbf{c}_1}$ des äußeren Codes für $j \in \{1,2,..,n_1\}$. Die Abbildung $\tilde{c} : \mathcal{A}_{\mathbf{c}_1} \rightarrow \mathcal{A}_{\mathbf{i}_2}^{k_2}$ ist dabei eine Eins-zu-eins-Abbildung von Symbolen des äußeren Codes auf Informationsworte des inneren Codes. Die Parameter des verketteten Codes $\mathcal{C}_\mathrm{v}(n_\mathrm{v},k_\mathrm{v},d_\mathrm{v})$ sind durch die Parameter der einzelnen Codes gegeben als $n_\mathrm{v} = n_1 n_2$, $k_\mathrm{v} = k_1 k_2$ und $d_\mathrm{v} \geq d_1 d_2$ (vgl. [140]). Dabei ist die Rate des neuen Codes $R_\mathrm{v} = {}^{(k_1 k_2)}/_{(n_1 n_2)}$. Die verallgemeinerte Codeverkettung und das Grundprinzip der Partitionierung des inneren Codes bietet ein mächtiges Konzept zur facettenreichen Rekombination von Codes zu längeren, neuen Strukturen. Das verallgemeinerte Prinzip ermöglicht für gleiche Codeparameter generell eine Erzeugung von Codes mit besseren Distanzeigenschaften, als es durch das einfache Schema der Fall wäre.

Unabhängig vom Einsatz eines bestimmten Codes zur Informationsübertragung existieren unterschiedliche Decodierprinzipien, die sich durch Formalismen der Wahrscheinlichkeitstheorie verallgemeinert darstellen lassen. Folgender Überblick bezieht sich auf die Standardwerke [22, 60].

Sei $\Pr(\mathbf{C} = \mathbf{c})$ die A-priori-Wahrscheinlichkeit für Codeworte $\mathbf{c} \in \mathcal{C}$ und $\Pr(\mathbf{Y} = \mathbf{r}|\mathbf{X} = \mathbf{c})$ die Übergangswahrscheinlichkeit des Kanals, welche der Beobachtung von Empfangsworten $\mathbf{r}$ zu Grunde liegt, so lässt sich die sogenannte A-posteriori-Wahrscheinlichkeit durch den Satz von Bayes formulieren zu

$$\Pr(\mathbf{X} = \mathbf{c}|\mathbf{r}) = \frac{\Pr(\mathbf{Y} = \mathbf{r}; \mathbf{X} = \mathbf{c})}{\Pr(\mathbf{Y} = \mathbf{r})} = \frac{\Pr(\mathbf{Y} = \mathbf{r}|\mathbf{X} = \mathbf{c})}{\Pr(\mathbf{Y} = \mathbf{r})} \Pr(\mathbf{C} = \mathbf{c}). \tag{2.4}$$

Sie liefert eine generelle Grundlage für Ansätze der allgemeinen Bayesschen Statistik und Decodierung.

[2] ganzzahlige Abrundung $\lfloor x \rfloor = \max\{\hat{x} \in \mathbb{Z} : \hat{x} \leq x\}$

Definition 21 (Maximum-a-posteriori-Decodierung, MAP)**:** Das Prinzip des MAP-Decodierers sind folgende Maximierungsprobleme:

$$\hat{\mathbf{c}} = \arg\max_{\mathbf{c}\in\mathcal{C}} \Big\{\Pr(\mathbf{X}=\mathbf{c}|\mathbf{r})\Big\} \text{ und } \hat{c}_i = \arg\max_{c_i\in\mathbb{F}_q} \Big\{\Pr(X_i=c_i|\mathbf{r})\Big\}.$$

Die Decodierentscheidung $\hat{\mathbf{c}}$ (bzw. $\hat{c}_i$) ist dabei die Sequenz (das Symbol) mit größter Wahrscheinlichkeit, bedingt durch ein bestimmtes Empfangswort $\mathbf{r}$. Generell ermöglicht die symbolweise Entscheidung ($\hat{c}_i$) Sequenzen $\hat{c}_1\hat{c}_2..\hat{c}_n \rightarrow \hat{\mathbf{c}} \notin \mathcal{C}$, die keine Codeworte sind. Dies ist möglich, weil die Struktur des Codes nicht explizit berücksichtigt wird. Ein Decodierergebnis das keinem gültigen Codewort entspricht wird auch als Decodierversagen bezeichnet.

Die allgemeinen Berechnungen zur MAP-Entscheidung setzten ein umfassendes Wissen über auftretende Verteilungen voraus, das meistens nicht gegeben ist oder durch Annahmen ersetzt wird. Das Prinzip der Maximum-Likelihood-Decodierung beruht auf derartig vereinfachende Annahmen.

Definition 22 (Maximum-Likelihood-Decodierung, ML)**:** Betrachtet man (2.4), so kann der Wert von $\mathbf{r}$ und damit $\Pr(\mathbf{Y}=\mathbf{r})$ als unveränderliche Konstante betrachtet werden, die keinen Einfluss auf eine Maximierung nimmt. Wird zusätzlich $\Pr(\mathbf{C}=\mathbf{c})$ als konstant angenommen zu $1/|\mathcal{C}|$ (uniforme A-priori-Wahrscheinlichkeit), so ist die MAP-Entscheidung äquivalent zu

$$\hat{\mathbf{c}} = \arg\max_{\mathbf{c}\in\mathcal{C}} \Big\{\Pr(\mathbf{Y}=\mathbf{r}|\mathbf{X}=\mathbf{c})\Big\},$$

was auch als Maximum-Likelihood-Entscheidung bezeichnet wird. Basierend auf einer Modellannahme (Hypothese) zum Kanal $\Pr(\mathbf{Y}|\mathbf{X})$ entscheidet der ML-Decodierer für das Codewort $\hat{\mathbf{c}}$, welches das Empfangswort $\mathbf{r}$ am besten erklärt.

Zu einer Vielzahl von Kanalmodellen existiert für das probabilistische Entscheidungsproblem eine äquivalente geometrische (distanzbasierte) Formulierung: So ist beispielsweise für einen diskreten Kanal ohne Gedächtnis (vgl. Definition 12) ein Fehlerereignis im Allgemeinen weniger wahrscheinlich als ein korrekt übertragenes Symbol. Im Kontext von Codeworten lässt sich der Vergleich von Symbolen auf den Hamming-Abstand beziehen und die Entscheidung auf eine Minimierung der Distanz (engl. *Minimum-Distance-Decoding, MD*) projizieren.

Definition 23 (Minimum-Distance-Decodierung, MD)**:** Die ML-Decodierung lässt sich (abhängig vom Kanalmodell), als Problem der Distanzminimierung, auf die Entscheidung

$$\hat{\mathbf{c}} = \arg\min_{\mathbf{c}\in\mathcal{C}} \Big\{\mathrm{d}(\mathbf{r},\mathbf{c})\Big\}$$

abbilden. Dabei ist d eine für das Kanalmodell adäquate Metrik, meist der Hamming-Abstand. Wird eine beschränkende Nebenbedingung auf maximale Distanzen $t \leq \lfloor (d_{\min}-1)/2 \rfloor$ in die Decodierung integriert, also $\hat{\mathbf{c}} = \arg\min_{\mathbf{c}}\{\mathrm{d}(\mathbf{r},\mathbf{c}) \leq t\}$, so handelt es sich dann um *Bounded-Minimum-Distance-Decoding (BMD)*. Existiert kein Codewort mit maximalem Abstand t zum Empfangswort $\mathbf{r}$, so entspricht $\hat{\mathbf{c}} = \emptyset$, d. h. es kann keine Entscheidung getroffen werden, was auch als Decodierversagen bezeichnet wird. Eine Implementierung des BMDs ist in Form eines sogenannten Standard-Arrays möglich. Für lange Codes ist der Ansatz jedoch sehr komplex und impraktikabel.

Betrachtet man den Hamming-Abstand als eine Art Kostenfunktion, so trägt jeder Symbolunterschied mit einheitlichen Kosten zur Distanz zweier Sequenzen bei. Verallgemeinert man das Konzept der Kosten auf uneinheitliche Werte und variable Sequenzlängen, so beschreitet man das Feld der *Sequenzähnlichkeit und musterbasierte Suche*.

2.1.3 Sequenzähnlichkeit und musterbasierte Suche

Das grundlegende Problem der approximativen musterbasierten Suche kann in folgender kompakter Form definiert werden:

Definition 24 (Musterbasierte Suche)**:** Sei eine kürzere Sequenz $\mathbf{P} = P_1P_2..P_m$ ein Muster und $\mathbf{T} = T_1T_2..T_n$ eine Referenz-Sequenz ($m < n$) mit Symbolen $P_i, T_j \in \mathcal{A}$ aus einem gemeinsamen Alphabet $\mathcal{A}$, so interessiert bei der musterbasierten Suche die Identifikation von $\mathbf{P}$ innerhalb $\mathbf{T}$. Dabei kann entweder lediglich die Lage von Anfang oder Ende des Musters oder die exakte Abbildung von $\mathbf{P}$ auf $\mathbf{T}$ von Interesse sein.

Geht man von dem Spezialfall aus, dass die Sequenz $\mathbf{T}$ vollständig aus einer Transformation von $\mathbf{P}$ entstanden ist (oder umgekehrt), so lässt sich bezüglich der zwei Folgen ein Distanzmaß formulieren, welches als *Editierdistanz* [99] oder Levenshtein-Distanz bekannt ist.

Definition 25 (Editierdistanz)**:** Nach [99] existiert für alle Sequenzen $\mathbf{P}$ und $\mathbf{T}$ ein zweistelliges Funktional $\mathrm{d_e}(\mathbf{P}, \mathbf{T})$ das definiert wird als die kleinste Anzahl beliebiger Kombinationen von Ersetzungen, Einfügungen und Löschungen, welche zu einer Transformation von $\mathbf{P}$ zu $\mathbf{T}$ führen oder umgekehrt. Das Funktional $\mathrm{d_e}$ ist eine Metrik und heißt Editierdistanz.

Bei der Editierdistanz sind Löschungen mit einer faktischen Verkürzung und Einfügungen durch die Verlängerung der Sequenzen verbunden. Für zwei Sequenzen mit Längen m und n gilt darüber hinaus $|m - n| \leq \mathrm{d_e} \leq \max\{m, n\}$. Für Sequenzen gleicher Länge gilt $\mathrm{d_e} \leq \mathrm{d_h}$, d. h. der Hamming-Abstand ist eine obere Schranke für die Editierdistanz.

Neben Levenshteins Definition der Distanz von 1965 gibt er keine explizite Berechnungsvorschrift für das nach ihm benannte Maß. Die Berechnung der Metrik dient jedoch oft als Lehrbuchbeispiel zur Veranschaulichung von Ansätzen der dynamischen Programmierung [14] und ist durch die Bezeichnung Levenshtein-Algorithmus geprägt:

Konzept 6 (Levenshtein-Algorithmus)**:** Die Berechnung der Editierdistanz zweier Sequenzen $\mathbf{P}$ und $\mathbf{T}$ der Längen m und n erfolgt auf Basis folgender rekursiver Operationen (vgl. [66]) auf einer vorab definierten Matrix $\mathbf{D}$ mit $D_{00} = 0$, $D_{i0} = i$ für $1 \leq i \leq m$ und $D_{0j} = j$ für $1 \leq j \leq n$:

$$D_{ij} = \min \begin{pmatrix} I(P_i \neq T_j) & + & D_{i-1,j-1} \\ 1 & + & D_{i,j-1} \\ 1 & + & D_{i-1,j} \end{pmatrix}, \text{ für } 1 \leq i \leq m, 1 \leq j \leq n,$$

mit sukzessiver spalten- bzw. zeilenweiser Propagierung. Dabei repräsentieren die Zeilen in der Minimierung (von oben nach unten) Entsprechung bzw. Ersetzung (mit Indikator $I = 1$ für $P_i \neq T_j$, sonst 0), Einfügung und Löschung eines Symbols. Der Wert D_{ij} beschreibt die Editierdistanz der Präfixe $P_1P_2..P_i$ und $T_1T_2..T_j$. Die eigentliche Editierdistanz ist D_{mn}.

Wie schon aus Definition 25 ersichtlich, handelt es sich bei der Distanzberechnung um ein Minimierungsproblem, für welches jeder Transformation Kosten zugeordnet werden. Diese Kosten sind für alle Transformationen einheitlich zu 1 gewählt, was bei einer Vertauschung der Sequenzen **P** und **T** zum gleichen Minimum führt und die Symmetrie der Distanz impliziert. In manchen Anwendungen ist es jedoch erforderlich, das klar definierte Distanzmaß durch Ähnlichkeitsmaße (ohne Distanzeigenschaft) zu substituieren.

Durch die Verallgemeinerung der Kosten gelangt man in den Themenkomplex der *Alignments*, einem Teilgebiet der musterbasierten Suche. Die Problembeschreibung des *Sequenzalignment* löst sich dabei vom strikten Distanzbegriff und der kompletten Transformation einer Muster-Sequenz in eine Referenz-Sequenz oder umgekehrt. Vielmehr rücken speziellere Aspekte in den Fokus, so dass beispielsweise ein Muster **P** als Teilsequenz einer Referenzsequenz **T** angenommen wird. Die approximative Suche nach Vorkommen eines Musters in einem bestimmten Kontext ist eine gängige Fragestellung, so z. B. in Internetsuchmaschinen oder eben im Bereich der Bioinformatik. Ausgangspunkt für ein breites Spektrum an Algorithmen zum Sequenzalignment ist der Needleman-Wunsch-Algorithmus [122] zur Berechnung des sogenannten globalen Alignments.

Konzept 7 (Needleman-Wunsch-Algorithmus)**:** In Anlehnung an [170] kann die Berechnung der minimalen Kosten eines globalen Sequenzalignments von **P** auf **T** mittels Rekurrenzen definiert werden, die auf einer vorab initialisierten Matrix **D** ausgeführt werden. Die Matrixeinträge sind wie folgt definiert:

$$\begin{bmatrix} D_{00} = & 0 \\ D_{i0} = & \min_{x \leq i}\{D_{i-x,j} + \alpha_i(x)\} \\ D_{0j} = & \min_{x \leq j}\{D_{i,j-x} + \alpha_j(x)\} \end{bmatrix} \text{ und } D_{ij} = \min \begin{pmatrix} D_{i-1,j-1} + \beta(P_i, T_j) \\ \min_{x \leq i}\{D_{i-x,j} + \alpha_i(x)\} \\ \min_{x \leq j}\{D_{i,j-x} + \alpha_j(x)\} \end{pmatrix},$$

für $1 \leq i \leq m$ und $1 \leq j \leq n$. Dabei beschreibt $\alpha_*(x) \geq 0$ eine allgemeine Funktion der Kosten für Einfügungen (Index i) und Löschungen (Index j) der Länge x, $\beta(P_i, T_j) \geq 0$ hingegen die Kosten für eine Ersetzung von T_j durch P_i. Die minimalen Kosten für die Abbildung von **P** auf **T** ergeben sich, nach sukzessiver Berechnung der Rekurrenzen, als D_{nm}.

Unter Verwendung zusätzlicher Matrizen ist eine explizite Nachverfolgung der Transformationen möglich, welche mit den minimalen Kosten verbunden sind. Es ist gängig die Kosten invertiert als (positive) *Scores* zu beschreiben, was eine Repräsentation des Optimierungsproblems als Maximierung zur Folge hat. Eine Äquivalenz der Beschreibungen wurde in [149] bewiesen. Für einen Bezug auf Konzept 6 wird fortwährend das Minimierungsproblem formuliert.

Der Needleman-Wunsch-Algorithmus lässt sich auf vielfältige Weise modifizieren, wodurch sich beispielsweise durch vereinheitlichte (affinen) Kostenfunktionen effizientere Inkarnationen des Grundkonzepts finden lassen [6, 64, 74, 149, 166] oder sich andere Lösungen für weiter Fragestellungen bieten, wie dem lokalen Sequenzalignment [154]. Eine weitere, hier hervorzuhebende Variante bildet das semi-globale Sequenzalignment.

Definition 26 (Semi-globale Alignment)**:** Der Needleman-Wunsch-Algorithmus kann in einfacher Weise zur Suche von semi-globalen Sequenzalignments genutzt werden. Für eine semi-globale musterbasierte Suche wird angenommen, dass ein Muster $\mathbf{P}$ als vollständige Teilsequenz in einer Sequenz $\mathbf{T}$ (auch mehrfach) enthalten ist. Da die Sequenz $\mathbf{T}$ neben dem Muster noch zusätzliche Symbole enthält (welche unabhängig von $\mathbf{P}$ sind), werden Einfügungen am Anfang bzw. Ende nicht berücksichtigt. Diese Nebenbedingung lässt sich in Konzept 7 integrieren, indem $D_{0j} = 0$ gesetzt wird, für $1 \leq j \leq n = |\mathbf{T}|$, und die gesamte letzte Zeile der Matrix $\mathbf{D}$ für die Suche des minimalen Wertes berücksichtigt wird. Somit ergibt sich $\min_i\{D_{mi}\}$ als Kosten der semi-globalen Alignments.

Für das semi-globale Alignment mit ganzzahligen Kosten existiert ein sehr effizientes Äquivalent der Kostenberechnung unter Ausnutzung von Bit-Arithmetik. Die Effizienz beruht dabei auf der Zusammenführung von Vergleichsoperationen in Bit-Mustern. Basierend auf Ideen [42, 43] zur exakten Suche nach Sequenzen mittels Bit-Operationen wurde Anfang der 90er Jahre der Bitap Algorithmus (Bit-Parallel Algorithm for Approximate Pattern Matching) als approximatives Pendant vorgestellt. Vorarbeiten in [11] zur Erkennung der Einbettung eines Musters in eine Referenz-Sequenz unter Hamming-Metrik wurden in [179] konsequent für die Editierdistanz verallgemeinert.

Konzept 8 (Bitap Algorithmus)**:** Sei $\mathbf{P} = P_1P_2..P_m$ ein Muster und $\mathbf{T} = T_1T_2..T_n$ eine Referenz-Sequenz mit Symbolen $P_i, T_j \in \mathcal{A}$. Die Grundlage für eine Arithmetik auf Bitebene ist eine Rechnerarchitektur, die ein binäres Wort $\mathcal{W}$ mit mindestens m Bit bereitstellt, sowie die Operatoren *Verschiebung*[3] ($\ll$) und logische Verknüpfungen *Und* (`and`) bzw. *Oder* (`or`). Sei ferner $\phi_{\mathbf{P}} : \mathcal{A} \to \mathcal{W}$ eine Projektion, die jedem Symbol des Alphabets $\mathcal{A}$ eine Bitfolge zuordnet, welche dessen Auftreten in $\mathbf{P}$ symbolisiert, so zeigt $\phi_{\mathbf{P}}(T_j)$ parallel an allen Positionen i eine logische eins, wo P_i und T_j übereinstimmen. Die Bit-Muster $\phi_{\mathbf{P}}$ und die Reduktion der zum Alignment nötigen Berechnungen auf Bit-Operationen ist das Kernkonzept von Bitap. So gilt für die exakte Suche nach [11] folgende rekursive Berechnungsvorschrift

$$R_j = (R_{j-1} \ll 1) \texttt{ and } \phi_{\mathbf{P}}(T_j)$$

mit $R_j \in \mathcal{W}$ und $R_0 = 0$. Dabei ist R_j der Zustandsspeicher nach der Verarbeitung des Symbols T_j und $R_j|_i$ (das Bit i in R_j) ein Indikator, ob eine Gleichheit $P_1P_2..P_i = T_{j-i+1}T_{j-i+2}..T_j$ besteht. Anders ausgedrückt enthält $R_j|_i$ die Information ob T_j eine Erweiterung des sonst fehlerfreien Präfixes bezüglich $\mathbf{P}$ darstellt. Für die approximative Suche [179] ist eine Erweiterung des Zustandsspeichers notwendig, um Abweichungen im Präfix zuzulassen. Dafür wird ein k-dimensionaler Vektor $\mathbf{R}_j \in \mathcal{W}^k$ benötigt, das wiederum den Zustand nach der Verarbeitung des Symbols T_j beschreibt. Der Wert k bestimmt dabei die maximal erlaubten Kosten für ein semi-globales Alignment. Dabei zeigt das Bit $\mathbf{R}_j[\kappa]|_i$ an, ob T_j der Abschluss eines Präfixes der Länge i ist, der Modifikationen der Kosten κ bezüglich $\mathbf{P}$ aufweist. Der Operator $[\kappa]$ ermöglicht den Zugriff auf die Dimension κ des Vektors. Die rekursive Berechnungsvorschrift

[3] Links-Verschiebung wird (abweichend gängiger Standards) mit Addition einer 1 an Bitposition 0 definiert.

Alg. 2.1 BITAP-ALGORITHMUS

Input : $\mathbf{P} = P_1P_2..P_m, \mathbf{T} = T_1T_2T_3..T_n, k, \{\kappa_\mathrm{s}, \kappa_\mathrm{i}, \kappa_\mathrm{d}\}$
Output : $\mathcal{I}$

$\mathcal{I} \leftarrow \emptyset$
$\mathbf{R}', \mathbf{R}'' \in \mathcal{W}^k$
$\mathbf{R}', \mathbf{R}'' \leftarrow \mathbf{0}$

for $\pi^\mathrm{end} \leftarrow 1$ **to** n **do**
  for $\kappa \leftarrow 0$ **to** k **do**
    $\mathbf{R}'[\kappa] = \phi_\mathbf{P}(T_{\pi^\mathrm{end}})$ `and` $\mathbf{R}''[\kappa] \ll 1$
    if $\kappa - \kappa_\mathrm{i} \geq 0$ **then** $\mathbf{R}'[\kappa] = \mathbf{R}'[\kappa]$ `or` $\mathbf{R}''[\kappa - \kappa_\mathrm{i}]$
    if $\kappa - \kappa_\mathrm{s} \geq 0$ **then** $\mathbf{R}'[\kappa] = \mathbf{R}'[\kappa]$ `or` $\mathbf{R}''[\kappa - \kappa_\mathrm{s}] \ll 1$
    if $\kappa - \kappa_\mathrm{d} \geq 0$ **then** $\mathbf{R}'[\kappa] = \mathbf{R}'[\kappa]$ `or` $\mathbf{R}'[\kappa - \kappa_\mathrm{d}] \ll 1$
    if $\mathbf{R}'[\kappa]|_m = 1$ **then** $\mathcal{I} \leftarrow \mathcal{I} \cup \{(\pi^\mathrm{end}, \kappa)\}$
  `flip`$(\mathbf{R}', \mathbf{R}'')$

Alg. 2.1 Bitap-Algorithmus: Zur Reduktion des Speicherbedarfs erfolgt die alternierende Nutzung von Zustandsspeichern $\mathbf{R}', \mathbf{R}'' \in \mathcal{W}^k$ (Wechsel durch die Methode `flip`). Die Rückgabe des Algorithmus besteht aus der Menge $\mathcal{I}$ an Einträgen $(\pi^\mathrm{end}, \kappa)$, zusammengesetzt aus End-Positionen bezüglich $\mathbf{T}$ und Kosten κ, die ein semi-globales Sequenzalignment von $\mathbf{P}$ ergeben.

der Elemente $\kappa \in \{0, 1, 2, ..., k\}$ der Vektoren lässt sich angeben als

$$\begin{aligned}
\mathbf{R}_j[\kappa] = \phi_\mathbf{P}(T_j) \ \texttt{and}\ & (\mathbf{R}_{j-1}[\kappa] \ll 1) \\
\texttt{or}\ & \mathbf{R}_{j-1}[\kappa - \kappa_\mathrm{i}] && (2.5)\\
\texttt{or}\ & (\mathbf{R}_{j-1}[\kappa - \kappa_\mathrm{s}] \ll 1) && (2.6)\\
\texttt{or}\ & (\mathbf{R}_j[\kappa - \kappa_\mathrm{d}] \ll 1), && (2.7)
\end{aligned}$$

wobei $\kappa_\mathrm{s}, \kappa_\mathrm{i}, \kappa_\mathrm{d} \geq 0$ die Kosten für Ersetzung bezüglich P_i und T_j, respektive Einfügungen und Löschungen enthalten. Die Zeilen (2.5) bis (2.7) werden nur ausgewertet, falls $\kappa - \kappa_* \geq 0$ gilt. Das Wort $\mathbf{R}_j[\kappa]$ aggregiert alle Indikatoren über Erweiterungen bestehender Präfixe mit den Kosten κ. Die Positionen aller semi-globalen Alignments von $\mathbf{P}$ auf $\mathbf{T}$ mit maximalen Kosten k können somit angegeben werden als $\{j : \mathbf{R}_j[\kappa]|_m = 1, \text{mit } 0 \leq \kappa \leq k\}$, wobei j die Enden bezüglich der Referenz $\mathbf{T}$ anzeigen. Der gesamte Algorithmus ist in Alg. 2.1 als Pseudoimplementierung beschrieben.

Der Bitap-Algorithmus birgt neben der Nutzung der wesentlich schnelleren Bit-Arithmetik Einschränkungen gegenüber der herkömmlichen Implementierung. Neben der Beschränkung auf ganzzahlige Kosten liefert der Algorithmus nur die Endposition der Übereinstimmung des Musters mit der Referenzsequenz. Eine genaue Beschreibung der zur Übereinstimmung nötigen Sequenzoperationen ist nicht enthalten.

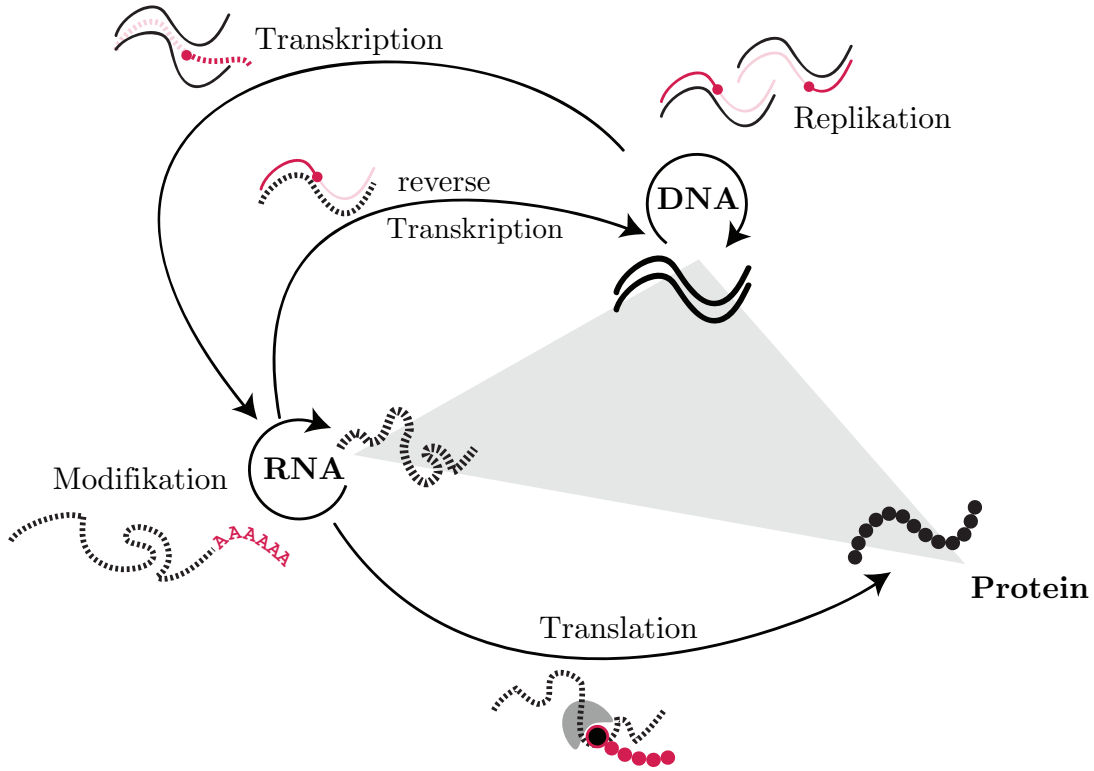

Abb. 2.2 Zentrales Dogma der Molekularbiologie: Abstrahierte Darstellung des Informationsflusses zwischen DNA, RNA, Proteinen und die damit verbundenen Mechanismen bzw. Vorgänge.

2.2 Biologie und Biotechnologie

Spricht man im Bereich der Biologie von Informationsspeicherung, so ist die damit verbundene Datenstruktur im physikalischen Sinne immer eine lineare. Für die Information an sich existieren die drei bekannten Repräsentationen DNA, RNA und Proteine, deren sequenzielle Struktur, als Ketten von Molekülen (*Polymere*) eine grundlegende Gemeinsamkeit darstellt. Möchte man den Informationsaustausch, oft auch Informationsfluss genannt, zwischen den materialisierten Formen der Information kompakt darstellen, bietet sich das sogenannte *Zentrale Dogma der Molekularbiologie* als Grundvorstellung an. Die von Francis Crick in den 50er Jahren entworfene und vorgestellte Hypothese [39] wurde 1970 veröffentlicht [38] und wird in ihrem Kern teils sehr kritisch bewertet: Zum einem bietet sie Studenten der Biologie einen übersichtlichen Zugang zu den Wirkungszusammenhängen des Lebens auf molekularer Ebene, zum anderen bot sie über die Jahre hinweg wegen ihrer dogmatisierten Abstraktion eine große Angriffsfläche für Wissenschaftler, die sich einer facettenreicheren Perspektive der Biochemie verschrieben haben (siehe dazu [119, 160]).

Ohne den Anspruch auf absolute Vollständigkeit des Informationstransfers soll im ersten Teil dieses Abschnitts die Darstellung Cricks genutzt werden, um den biologischen Hintergrund dieser Arbeit in einer überschaubaren Weise darzulegen. Dazu werden die biologischen Sequenzen charakterisiert und der wesentliche Transfer von Information beschrieben. Ausgehend von den natürlichen Vorgängen in Zellen werden im zweiten Teil biotechnologische Werkzeuge beschrie-

ben, welche natürliche Mechanismen nutzen, um auf molekularer Ebene eine Einflussnahme zu ermöglichen. Der letzte Teil motiviert die Einsatzmöglichkeiten von Barcodes im Rahmen der Sequenzierung von DNA und RNA. Eine ausführliche und vollständige Einführung der in reduzierter Form dargelegten Zusammenhänge ist beispielsweise im Standardwerk [5] zu finden, welches als Grundlage für die folgenden Erklärungen dient.

2.2.1 Informationstransfer in der Biologie

Das in Abb. 2.2 dargestellte Dreieck ist eine gängige Illustration zum Zentralen Dogma der Molekularbiologie: Die Ecken repräsentieren dabei die essentiellen Informationseinheiten DNA, RNA und Proteine. Die Pfeile entsprechen bekannten Mechanismen, welche einen Transfer unter den Einheiten ermöglichen.

Begriff 1 (DNA)**:** Die wohl beständigste Informationseinheit stellt die genetische Information in Form der *Desoxyribonukleinsäure* (engl. *DNA*) dar. Die Struktur eines DNA-Moleküls (vgl. Abb. 2.3) ist bestimmt durch den paarweisen Verbund von langen Ketten sogenannter *Nukleotide* (*nt* als Maßeinheit), Moleküle die in ihrer Abfolge die Erbinformation codieren. Einzelne Ketten werden auch als DNA-Stränge bezeichnet, welche auf Basis von Wasserstoffbrückenbindungen der Nukleotide einen Verbund als Leiterstruktur bilden. Die Holme der Struktur (auch Rückgrat genannt) werden dabei von einer alternierenden Abfolge von Zucker (Desoxyribose) und Phosphat(-gruppen) gebildet, welche gleichförmig in jedem der verschiedenen Nukleotide enthalten ist. Unterschieden werden die Nukleotide durch Anteile einer stickstoffhaltigen Base, welche unter anderem die Ausbildung der Wasserstoffbrückenbindung bestimmt. Es existieren die Basen Adenin (A), Guanin (G), Cytosin (C) und Thymin (T), die abgekürzt als Alphabet des genetischen Codes verstanden werden. Die chemische und strukturelle Eigenschaft der Zucker-Phosphat-Folge in Verbindung mit benachbarten Basen der Nukleotide verleiht jedem DNA-Einzelstrang eine klar definierte Orientierung. Als einheitliche Konvention zur Referenzierung der Richtung der Molekülketten hat sich eine Nummerierung anhand der fünf Kohlenstoffatome der Desoxyribose etabliert (siehe Abb. 2.3): So ist an Position 5 ein Phosphatrest gebunden, während an Position 3 eine OH-Gruppe (Hydroxyl) vorhanden ist. Dadurch definierte Enden eines Einzelstrangs werden auch als 5'(„fünf-Strich“)-Ende bzw. 3'(„drei-Strich“)-Ende bezeichnet. Die Enden haben für die Synthese von Molekülketten eine maßgebliche Bedeutung, weshalb die (5'→3')-Kette auch als Vorwärtsstrang und die (3'→5')-Folge als Rückwärtsstrang bezeichnet wird. Dabei ist eine Paarung der Ketten nur möglich, wenn die Richtungen zueinander antiparallel verlaufen. Beide Stränge bilden dann, als dreidimensionale Struktur, eine in sich verdrehte Doppelhelix, mit sich gegenüberliegenden Basen in der Mittelachse. Hinsichtlich der Struktur und Valenzen der Bindungskräfte können sich dabei nur Adenin (A) und Thymin (T) bzw. Guanin (G) und Cytosin (C) gegenüberliegen, was als komplementäre Basenpaarung benannt wird. Betrachtet man die Sequenzen der Nukleotide von Vorwärts- und Rückwärtsstrang zueinander, so geht die eine Folge aus dem Komplement der anderen hervor, wenn zusätzlich die Leserichtung geändert wird. Man sagt, die Sequenzen sind revers komplementär zueinander.

Die DNA ist bei allen Lebewesen prinzipiell gleich aufgebaut, dennoch lassen sich hinsichtlich der Lokalisation und der erweiterten Struktur der DNA zwei Gruppen von Organismen unterscheiden. Die *Eukaryoten* sind Organismen, welche die DNA im Zellkern bündeln. An dieser

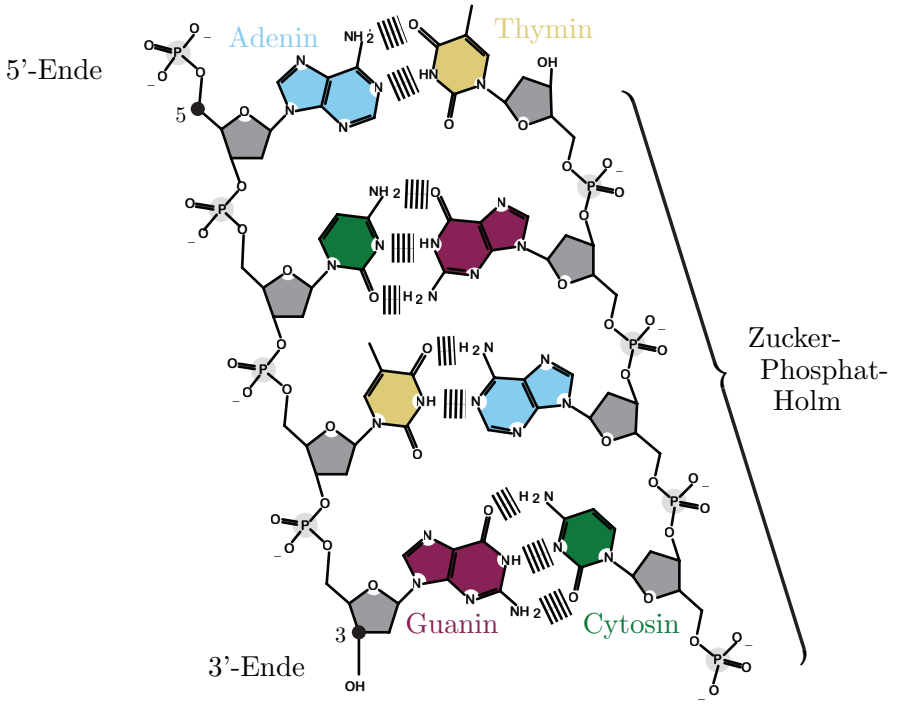

Abb. 2.3 Struktur der DNA (nach [178]): Standard-Nomenklatur der chemischen Elemente. Komplementäre Basenpaarungen werden durch Wasserstoffbrückenbindungen gehalten.

Bündelung sind weitere Struktureinheiten beteiligt, die Nucleosomen, welche von DNA umwickelt sind. Den *Prokaryoten*, wie beispielsweise den Bakterien, fehlt sowohl ein Zellkern als auch eine erweiterte Struktur. Auf Grund des anderen zellulären Aufbaus, ist die Zellteilung und Replikation bei Prokaryoten deutlich einfacher strukturiert. Folgende Erklärungen beziehen sich daher auf Prokaryoten.

Begriff 2 (DNA-Replikation)**:** Ein elementarer Informationstransfer in Abb. 2.2 ist die DNA-Replikation, welche sowohl die Erhaltung der Erbinformation für den eigenen Organismus als auch die Weitergabe an Nachkommen beinhaltet. Das grundlegende Konzept der Replikation von Sequenzinformation ist eine *matrizengesteuerte Polymerisation*, welche auf der unterschiedlichen Stärke der chemischen Bindung zur verhältnismäßig schwachen Wasserstoffbrückenbindung beruht. Eine reversible Trennung der DNA-Einzelstränge bietet dabei relativ stabile Molekülketten, die als Vorlage (auch *Matrize* oder *Template* genannt) für eine komplementäre Ergänzung von Nukleotiden dienen. Eine freie Synthese eines Einzelstrangs ist im natürlichen Sinne nicht möglich. In Abb. 2.4 ist eine vereinfachte Darstellung des Ablaufs der DNA-Replikation dargestellt, welche sich in die Phasen Initiation, *Elongation* und Terminierung unterteilen lässt: Initial findet (an nicht näher erklärten Regionen) eine enzymatische Entwindung und Auftrennung des Doppelstrangs statt, welcher in zwei Replikationsgabeln mündet. Diese werden durch spezielle Proteine am unmittelbaren schließen der Wasserstoffbrücken (*Hybridisierung* genannt) gehindert und dienen als Matrizen für die antiparallele und komplementäre Synthese. Die Synthese erfolgt dabei im Allgemeinen nicht

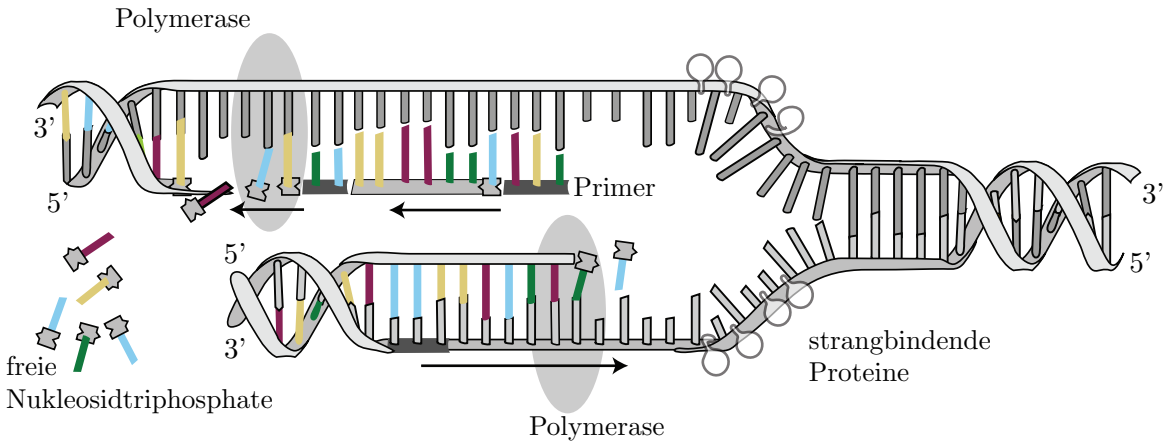

Abb. 2.4 Replikation von DNA (nach [175]): Antiparallele und komplementäre Synthese von doppelsträngiger DNA an Replikationsgabeln. An der Synthese sind unter anderem freie Nukleosidtriphosphate beteiligt, welche Vorstufen-Bausteine der Nukleotide darstellen.

kontinuierlich, sondern in Etappen. Diese Etappen werden definiert durch kurze Ketten von einsträngigen Nukleotiden, welche *Primer* genannt werden. Diese Primer hybridisieren (wenn komplementär passend) an den Replikationsgabeln und bilden somit doppelsträngige Bereiche, die als Startpunkt für das Enzym DNA-*Polymerase* dient, welches die komplementäre Erweiterung mit Nukleotiden katalysiert. Diese Vervollständigung wird auch Elongation genannt. Bedingt durch die chemische Struktur der DNA ist auch die Polymerase richtungsgebunden und ermöglicht eine Synthese nur in (5'→3')-Orientierung. Da die Orientierungen der komplementären Stränge antiparallel zueinander sind, kann die Synthese an einem Teilstrang nur diskontinuierlich erfolgen. In einem letzten Schritt werden Primer enzymatisch entfernt und dadurch entstandene Lücken komplementär ergänzt. Das Enzym *Ligase* schießt temporär entstandene Lücken der Zucker-Phosphat-Holme und stellt damit die stabile Struktur des Verbunds aus Matrize und synthetisiertem Strang sicher. Die Terminierung der Replikation wird durch bestimmte Sequenzen und daran bindende Proteine gesteuert, welche einen synchronisierten Stopp der Synthese für beide Replikationsgabeln sicherstellt.

Das Schlüssel-Schloss-Prinzip, das die chemische Struktur und die Wasserstoffbrückenbindungen komplementärer Nukleotide vorgibt, ist nicht ausreichend um eine Sequenz robust über eine Vielzahl von Replikationen zu konservieren. Zudem ist die mittlere Geschwindigkeit von circa 1000nt/s, in welcher die Synthese abläuft, beachtlich. Aus dem Grund existiert für diesen Informationstransfer eine Kaskade von zwei Fehlerkorrekturmechanismen: Eine unmittelbar an der Synthese beteiligte Korrekturleseaktivität ermöglicht die Reduktion der Fehlerrate von 10^{-5} für die (5'→3')-Polymerisation auf 10^{-7}. Ein zusätzliches Fehlpaarung-Korrektursystem bereinigt Sequenzfehler unabhängig von der DNA-Replikation (nachträglich) und reduziert die Auftrittswahrscheinlichkeit eines Fehler auf 10^{-9}.

Die in der DNA gespeicherte Information ist nicht nur zum Selbstzweck der Replikation darin enthalten, sondern sie codiert auch Vorschriften zur Synthese von komplexeren Molekülen, welche wiederum erweiterte Aufgaben in der Zelle erfüllen. Um diese Vorschriften *ausdrücken* (*exprimieren*) zu können, bedienen sich alle lebenden Organismen einer weiteren Klasse von Sequenzen, der sogenannten RNA.

Begriff 3 (RNA)**:** Eine weitere Informationseinheit (in Abb. 2.2) stellt die *Ribonukleinsäure* (engl. RNA) dar. Ähnlich der DNA besteht die RNA aus einer Kette von Nukleotiden, welche jedoch als einzelsträngige Form bekannt ist. Hinsichtlich der chemischen Struktur der Bestandteile der Ribonukleinsäure wird der Molekülverbund gestützt durch eine alternierende Abfolge des Zuckers Ribose (statt Desoxyribose) und den bereits bekannten Phosphatgruppen. Wie bei der DNA ist die Information der Sequenz in Basen encodiert, welche sich an die Ribose-Moleküle angliedern. Die Basen A, G, C sind strukturgleich in DNA und RNA zu finden, wohingegen die Base Uracil (U) an die Stelle des Thymins (T) tritt. Für Uracil (U) ist eine komplementäre Bindung an Adenin (A) möglich. Die RNA ist aufgrund der fehlenden Doppelstruktur um einiges variabler in der Bildung von dreidimensionalen Formationen, was durch mögliche Hybridisierung von Basen innerhalb eines Moleküls weiter begünstigt wird. Dies kann zu Rückfaltungen von Molekülen führen. Die erweiterte Komplexität in der *Sekundärstruktur* kann dabei bereits als funktionale Komponente verstanden werden.

RNA-Moleküle können vielseitige Aufgaben erfüllen, die hier nur beispielhaft erörtert werden: RNA kann als direkter Überträger genetischer Information dienen und eine unmittelbare Vorlage für Proteine darstellen, oder aber sie ist mittelbar als funktionale Einheit an der Synthese des Proteins beteiligt. Hinsichtlich ihrer Funktionalität kann die RNA durch einen Präfix näher spezifiziert werden: so ist die *mRNA* (engl. *messenger RNA*) beispielsweise eine Matrize für ein Protein; die *rRNA* encodiert die Struktur der sogenannten *Ribosomen*, das eine essentielle katalytische Funktion beim Aufbau von Proteinen aus Aminosäuren besitzt; oder die *tRNA* (Transfer-RNA), welche als Hilfs- und Trägermolekül für Aminosäuren agiert.

Begriff 4 (Transkription)**:** Der Informationstransfer, welcher in Abb. 2.2 den Übergang von DNA in RNA beschreibt nennt man *Transkription*. Dieser Vorgang ist dem Konzept der DNA-Replikation sehr ähnlich. Einzelsträngige DNA wird ebenfalls als Vorlage für eine *matrizengesteuerte Polymerisation* verwendet. Dabei zeigen sich jedoch gewisse Unterschiede: Das Endprodukt ist keine beständige doppelsträngige und komplette Kopie der Matrizen, sondern ein begrenzter Umschrieb (*Transkript*) einer lokalen Sequenzinformation in einzelsträngiger Form. Die Region auf einem Strang der DNA, welche als Vorlage für das Transkript dient, nennt man auch *Gen*. Transkripte sind nicht als beständige Informationseinheiten zu sehen, sondern werden als *Massen-Wegwerf-Produkt* in der Zelle verschlissen. So wird im Allgemeinen meist eine Vielzahl identischer Transkripte von einem Gen erzeugt. Die Synthese der RNA beinhaltet hierzu Steuermechanismen (als Bindungsstellen der DNA), welche die Aktivität der Transkription bestimmt. In Abb. 2.5 ist die Erzeugung von RNA vereinfacht dargestellt: Nach regionalen Entwinden und Auftrennen der DNA (10-20 Basen) wird der codierende Strang durch das Enzym RNA-Polymerase komplementär durch Ribonukleotide ergänzt. Bei der Elongation werden, wie schon bei der DNA, freie Nukleosidtriphosphate zur Synthese der RNA verbunden. Die Synthese verläuft wiederum in (5'→3')-Richtung. Sowohl der Start der Polymerisation als auch das Ende (Terminierung) wird durch Sequenzen vor und nach dem Transkript definiert. Das fertige RNA-Molekül verliert wie die RNA-Polymerase nach der Transkription die Bindung zum DNA-Strang.

Wie die Replikation von DNA ist die Transkription kein fehlerfreier Vorgang. Mit einer Bewegungsgeschwindigkeit von circa 50nt/s erzeugt die RNA-Polymerase im Mittel alle $10^4 - 10^5$ Symbole einen Fehler [16]. Wegen der fehlenden erweiterten Fehlerkorrektur unterscheidet sich die Fehlerwahrscheinlichkeit jedoch um den Faktor $10^4 - 10^5$ vom der DNA-Replikation. Der all-

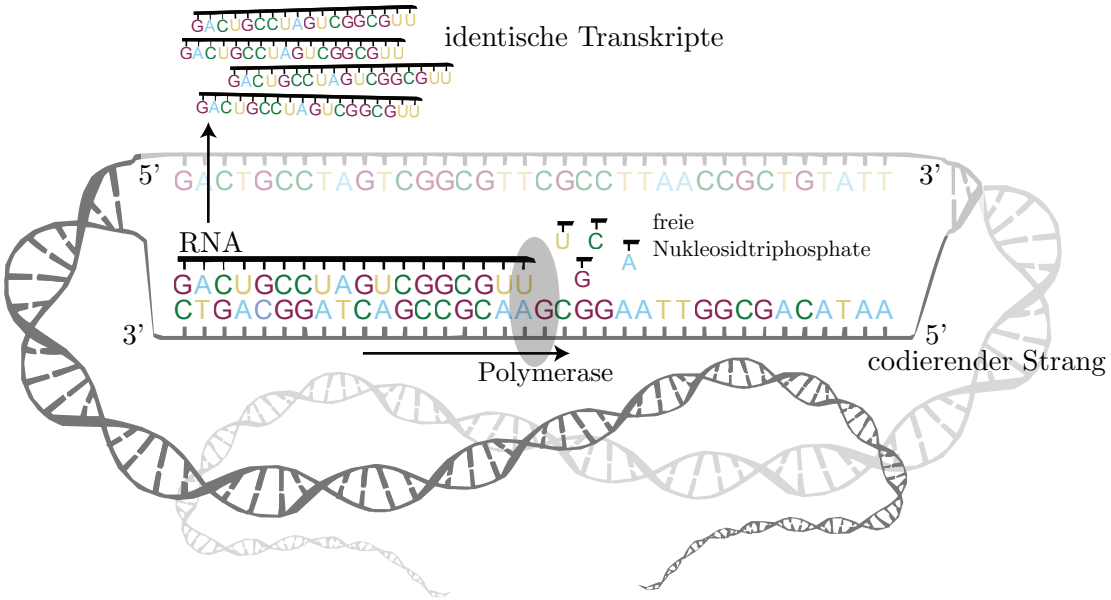

Abb. 2.5 Transkription von DNA zu RNA (nach [173]): Synthese von RNA am codierenden Strang der DNA.

gemein latente Charakter der RNA und die redundante Erzeugung vieler RNA-Moleküle macht die Präsenz von Fehlern für eine robuste Zellaktivität tolerabel.

Codiert eine RNA (als mRNA) die Sequenzinformation für ein Protein, so enthält dessen Beschreibung zusätzliche Redundanz, die eine Fehlertoleranz im Informationstransfer ermöglicht. Die Struktur der Redundanz ist auch unter dem Namen *genetischer Code* bekannt geworden.

Begriff 5 (Genetischer Code, Codons)**:** Der genetische Code ist definiert als eine Abbildung von Nukleotiden zu Aminosäuren, welche Bausteine der Proteine sind. Dazu ist die Information für eine Aminosäure in Gruppen von jeweils drei Nukleotiden (*Codons*) codiert. Auf Grundlage der vier Basen A,G,C und U der mRNA sind somit $4^3 = 64$ Codeworte möglich, jedoch sind lediglich 20 unterschiedliche Aminosäuren bekannt. Folglich existieren einige synonyme Codons, die für die gleiche Aminosäure codieren. In Abb. 2.6 ist der genetische Code als sogenannte Code-Sonne dargestellt. Die Leserichtung der Codons ist dabei von innen (5'-Ende) nach außen (3'-Ende), wobei die Abkürzungen der entsprechen Aminosäuren am Rand zu finden sind.

Die zuvor beschriebene theoretische Abbildung des genetischen Codes muss in der Zelle auch biochemisch umgesetzt werden. Die in der mRNA enthaltene Information muss *translatiert*, also übersetzt werden.

Begriff 6 (Translation)**:** Der Informationsfluss in Abb. 2.2, welcher die Synthese von Proteinen auf Basis von mRNA beschreibt heißt Translation. Die Translation als Reaktion findet an den *Ribosomen* statt, einem Molekülverbund aus rRNA, welcher die Bildung der Protein-Sequenzen katalysiert. Dieser Ablauf ist in Abb. 2.6 (rechts) schematisch illustriert und findet zeitgleich an einem oder mehreren mRNA-Molekülen statt: Für die Initiation der Synthese einer neuen Aminosäure-Sequenz ist (unter anderem) ein bestimmtes Signal der mRNA nötig,

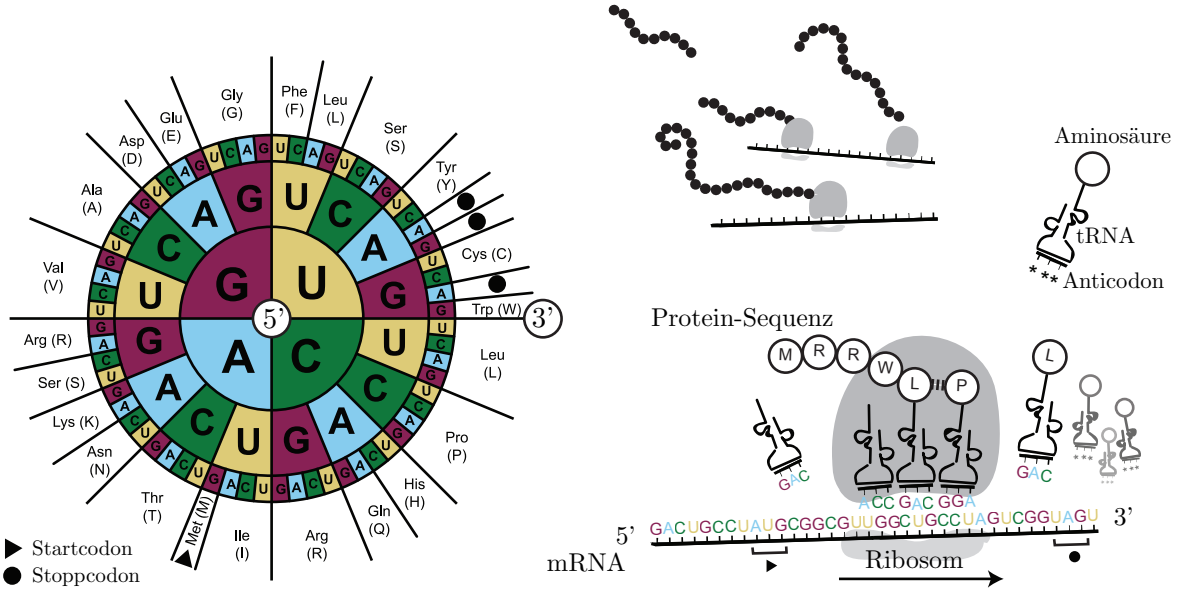

Abb. 2.6 Genetische Code als Code-Sonne (links, nach [172]) und Translation von mRNA zum Protein (rechts, nach [174]). Translation läuft mehrfach, zeitgleich und parallel, ab.

das Startcodon (AUG) genannt wird. Diese Sequenz ermöglicht die Fusion der Teilgruppen der Ribosomen zu einem funktionalen Synthesekomplex. An der Initiation sowie dem weiteren Aufbau der Protein-Kette sind freie tRNA-Moleküle beteiligt. Nach dem Schlüssel-Schloss-Prinzip bildet ein Codon und das an der tRNA befindliche *Anticodon* ein komplementäres Paar von Nukleotiden, das den Transfer der Sequenzinformation von mRNA auf ein Protein ermöglicht. Die Kettenbildung ist dabei in drei Schritte gegliedert: Die tRNA wird erkannt, sie wird an das Ribosomen gebunden und die Kette wird verlängert. Der Vorgang wird sukzessive (an der mRNA entlang) fortgeführt bis ein Teminierungssignal, das sogenannte Stoppcodon, erreicht wird. Durch diese Sequenz wird eine abschließende katalytische Aktivität initiiert, die das Protein vervollständigt und die Ablösung des Ribosoms von der mRNA bedingt.

Translation als Informationstransfer ist ebenfalls fehlerbehaftet: Ein Ribosomen gleitet sehr schnell am mRNA-Strang entlang und erzeugt dabei im Mittel circa 20 Aminosäuren pro Sekunde, wobei durchschnittlich ein Fehler pro 10^3 Symbolen auftritt [146].

Die hohe Fehlerrate bei Proteinen ist für Stoffwechselvorgänge eher von geringerer Bedeutung. Ähnlich wie für die mRNA, werden nicht einzelne konkrete Moleküle erzeugt, die nur fehlerfrei eine bestimmte Aufgabe in der Zelle übernehmen können, vielmehr existiert ein fließender Übergang zwischen voll-funktional und nicht funktional. Zusätzlich ist das Molekülgefüge in der Zelle einer hohen Dynamik unterworfen: Gene werden gleich mehrfach transkribiert; Transkripte werden beständig translatiert; anderseits sind Proteine ebenso wie mRNA einer ständigen Modifikation und dem Abbau (*Degradation*) unterworfen, der eine nachhaltige Fortpflanzung von Fehlern erschwert.

Begriff 7 (RNA-Modifikation)**:** Ein Informationstransfer in Abb. 2.2 von RNA zu RNA ist hier, im Gegensatz zur ursprünglichen Formulierung des zentralen Dogmas der Molekularbiologie, etwas anders interpretiert. Die Anschauung in [38] definiert diesen Transfer

über die RNA-Replikation, welche bei RNA-Viren eine Vervielfältigung von Sequenzinformation ermöglicht. In der hier gezeigten Perspektive soll die funktionale Modifikation an der RNA-Sequenz als Informationsverarbeitung gesehen werden. Der Begriff *Posttranskriptionale Modifikation* umfasst eine breites Feld an Wirkungsmechanismen, die meist im Kontext von Eukaryoten verstanden werden, da bei ihnen der Aufbau von protein-codierender mRNA eine mehrschichtige Informationsverarbeitung beinhaltet. Der hier beschriebene direkte Weg DNA→mRNA→Protein ist nur Prokaryoten zu Eigen. Dennoch existieren auch für Organismen ohne Zellkern *Posttranskriptionale Regulationen* [128], welche ein sehr junges Forschungsfeld bieten. Die *Polyadenylierung* ist eine dieser Regulationen. Sie umfasst die Erweiterung von RNA-Molekülen durch lange Ketten der Base Adenin (A) an deren 3'-Ende, die dann auch als Poly-(A)-3'-Enden bezeichnet werden. Diese Veränderungen werden mit der Stabilität und dem Abbau von mRNA bei Prokaryoten in Verbindung gebracht und sind für Eukaryoten schon länger bekannt.

Für Prokaryoten birgt der Mechanismus Polyadenylierung eine Vielzahl unbeantworteter, biologisch höchst interessanter Fragen (siehe hierzu [36, 143]). Für die Wirkungszusammenhänge und offene Themen bezüglich der Polyadenylierung wird auf die in [137, 155] gegebenen Übersichten referenziert.

Die in Abb. 2.2 hervorgehobenen Pfade der Repräsentation und Verarbeitung von Information in einer Zelle ist nur eine hervorgehobene Ebene im vielschichtigen Netzwerk der Wirkungsmechanismen, die mit voranschreitender Forschung immer umfangreicher dokumentiert und verstanden werden.

2.2.2 Biotechnologische Werkzeuge und Begriffe

Die im vorangehenden Abschnitt aufgeführten Vorgänge sind Teil der Lebensprozesse einer Zelle. In der Molekularbiologie spricht man auch von *in vivo* (lat. im Lebendigen) ablaufenden Prozessen. Viele der natürlichen Wirkungsmechanismen lassen sich aber auch *in vitro* (lat. im Glas) reproduzieren und damit als biotechnologische Werkzeuge verwenden, um einen gezielten Informationstransfer von außen zu ermöglichen. Hier werden die wichtigsten Begriffe kurz dargelegt.

Begriff 8 (Oligonukleotid)**:** Ein *Oligonukleotid* (auch Oligomer genannt) ist ein kurzes Polymer aus wenigen Nukleotiden (<200nt), das als kurze DNA- oder RNA-Sequenz bestimmte Aufgaben erfüllt. Die Synthese von Oligonukleotiden ermöglicht es, ausgehend von einer textuellen Sequenz (auf dem Papier), eine physikalische Repräsentation als Nukleotid-Kette zu erstellen. Diese Moleküle haben als *technische Sequenz* meist eine klar definierte Aufgabe: Die Ketten können beispielsweise als neue Primer für Polymerasen fungieren oder als *Adapter* bzw. DNA-Sonden eine lokale Bindungsstelle zur Bildung von Doppelsträngen (Hybridisierung) darstellen.

Manche Anwendungen machen es erforderlich, dass Oligonukleotide als Präfix oder Postfix an bestehenden (native) Sequenzen angeheftet werden oder zu komplexen Sequenzstrukturen kombiniert werden. Die *Ligation* beschreibt den dazu nötigen Vorgang der Konkatenation von biologischen Sequenzen.

Begriff 9 (Ligation)**:** Das in vivo vorkommende Enzym Ligase katalysiert in der gleichnamigen Ligation den Verbund von zwei Nukleotid-Ketten durch die Bildung von starken chemischen Bindungen. Unter Einsatz von Stoffwechselenergie (energiereicher Nukleosidtriphosphate) ermöglicht die Ligation den Aufbau der Zucker-Phosphat-Holmstruktur der DNA und RNA. Auf Grund des Energieaufkommens und der Struktur der Nukleotid-Ketten sind spezielle 3'-/5'-Enden für die Reaktion nötig.

Ein großer Teil der biotechnologischen Werkzeuge sind, wie die Ligase, nativ vorkommende Enzyme, die aus unveränderten Organismen isoliert werden können. Heute gängiger ist die industrielle Produktion durch die Übertragung der für die Enzyme codierenden Gene auf Bakterien (wie z. B. E. coli), welche über ihren normalen Stoffwechsel hinaus für die zusätzliche Transkription und Translation der gewünschten Stoffe verändert werden.

Ein weiteres bedeutungsvolles Enzym, das es ermöglicht den Informationstransfer DNA $\rightarrow$ RNA umzukehren (vgl. Abb. 2.2), ist die *reverse Transkriptase*. Der zugehörige Mechanismus wird *reverse Transkription* (RT abgekürzt) genannt und wurde zuerst für Viren wie dem bekannten AIDS-Virus (HIV) nachgewiesen.

Begriff 10 (Reverse Transkription zu cDNA, RT)**:** Die reverse Transkription (RT) beschreibt die Übersetzung von RNA-Sequenzen zu *cDNA* (engl. *complementary DNA*). Der Prozess läuft dabei ähnlich der zuvor beschriebenen DNA-Replikation ab, wobei die RNA nun als Vorlage für die matrizengesteuerte Polymerisation dient. Der oft als einstufiger Vorgang illustrierte Prozess beinhaltet jedoch zwei Ebenen, die von unterschiedlichen Komponenten der reverse Transkriptase ermöglicht werden: Auf eine grundlegende komplementäre RNA-DNA-Hybridisierung durch eine RNA-abhängige Polymerase folgt eine RNA-DNA-Hybridisierung durch eine DNA-abhängige Komponente.

Im Vergleich zu den in 2.2.1 aufgeführten Arten des Sequenztransfers werden bei der RT $10^1 - 10^2$nt/s [35] synthetisiert. Zudem existiert keine erweiterte Korrektur, sodass sich eine mittlere Wahrscheinlichkeit von $10^{-4}-10^{-3}$ für eine fehlerhafte Base ergibt. Für die biotechnologische Verwendung ist dieser Sachverhalt äußerst kritisch, wenn die exakte Sequenzinformation der RNA von besonderer Bedeutung ist. Oft jedoch stellt die RT die einzige Möglichkeit dar, um RNA über die mittelbare Transformation zur cDNA zu analysieren, wenn grundlegende Analysemethoden nur für DNA implementiert sind.

Ein Einsatzbereich der RT ist unter anderem die sogenannte *reverse Transkriptase-Polymerase-Kettenreaktion* (RT-PCR) als zweistufige Kombination aus RT und der allgemeinen *Polymerase-Kettenreaktion* (engl. *Polymerase Chain Reaction*, *PCR*).

Begriff 11 (Polymerase-Kettenreaktion, PCR)**:** Die Polymerase-Kettenreaktion (PCR) nutzt das Konzept der DNA-Replikation (vgl. Abb. 2.2), um DNA-Moleküle kontrolliert in vitro zu vervielfältigen. Notwendig für die matrizengesteuerte Polymerisation ist, wie bereits erwähnt, das Auftrennen der DNA-Doppelstränge. Dieser Vorgang impliziert das Lösen der Wasserstoffbrückenbindungen zwischen den komplementären Basenpaaren und dies kann auch unspezifisch durch Erhitzung erfolgen, was als *Aufschmelzen* bezeichnet wird. Durch diesen Sachverhalt wird es möglich steuernd in die enzymatische Reaktion einzugreifen und diese wiederholt in *Zyklen* ablaufen zu lassen. Problematisch ist hierbei, dass die Temperatur

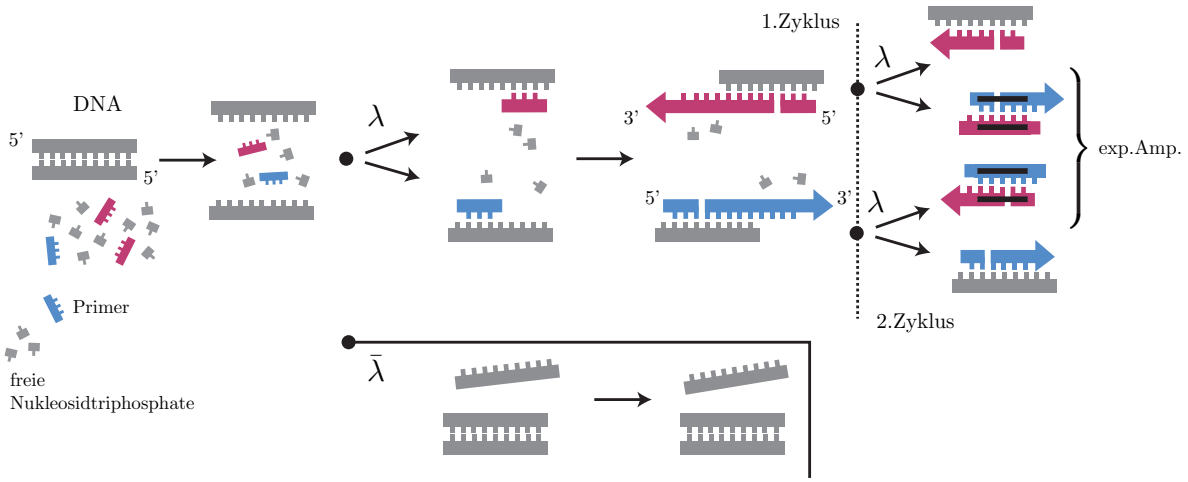

Abb. 2.7 Polymerase-Kettenreaktion, PCR (nach [177]): Primer flankieren einen zu vervielfältigenden Abschnitt der DNA. Mit Wahrscheinlichkeit λ resultiert ein Zyklus in einer Verdopplung des Abschnitts durch die DNA-Polymerase ($\bar{\lambda}$ Wahrscheinlichkeit des Gegenereignisses): Die PCR führt zur quasi-exponentiellen Verdopplung eines Abschnitts; andere Synthesevorgänge verlaufen linear.

neben dem gewünschten Aufschmelzen auch zerstörerisch auf andere Molekülverbindungen wirkt, so z. B. auf das Zucker-Phosphat-Gerüst der Peptid-Ketten oder auf das Replikation-Enzym Polymerase selbst. Kernpunkt für den biotechnologischen Erfolg der PCR war die Entdeckung von thermostabilen Enzymen (in thermophilen Organismen), wie die sogenannte *Taq*-Polymerase oder weitere Vertreter (z. B. *Pwo/Pfu*-Polymerase), welche zusätzlich über Korrekturmechanismen verfügen. Durch den Einsatz dieser Enzyme ist es möglich die Reaktion automatisiert in einem abgeschlossen Molekülgefüge durchzuführen.

Der idealisierte Ablauf der PCR ist in Abb. 2.7 dargestellt: Im Allgemeinen umfasst der Prozess 10-50 Zyklen (hier: zwei exemplarisch) zu jeweils drei Schritten. Zuerst wird die doppelsträngige DNA auf über 90°C erhitzt (das Aufschmelzen trennt Wasserstoffbrücken), um darauf folgend die Temperatur in kurzer Zeit um ca. 30°C zu verringern, was die Wiederausbildung der Brücken (Rehybridisierung) hemmt. DNA-Primer, kurze speziell synthetisierte Oligonukleotide, die als Reagenz beigefügt sind, werden in einem zweiten Schritt aktiv, der sogenannten Primer-Hybridisierung. Es existieren zwei Typen von Primern, welche für die unterschiedlichen Stränge spezifiziert sind, denn sie bilden Teilsequenzen der DNA nach, die als *Komplement* den Anfang und das Ende der zu vervielfältigenden Sequenz bestimmen. Auf Grundlage komplementärer Basenpaarungen hybridisieren die Primer und bilden den 5'-Startpunkt für die Polymerase. Die Hybridisierung ist dabei eher als Zufallsprozess zu verstehen, der sowohl von Temperatur (Brownsche Molekularbewegung) als auch von der Konzentration der Primer und der DNA abhängig ist. So erfolgt eine Verdopplung eines DNA-Moleküls im Mittel mit einer Wahrscheinlichkeit λ (auch Effizienz genannt), mit $\bar{\lambda} = 1 - \lambda$ verbleibt ein einzelsträngiges oder rehybridisiertes DNA-Molekül. Eine erfolgreiche Bindung des Primers initiiert die Elongation (vgl. Begriff 2). Die DNA-Polymerase übernimmt in (5'→3')-Richtung die matrizengesteuerte Polymerisation. DNA-Primer verbleiben im Replikat und bilden nach Zyklus 2 die Randbereiche der kettenmäßig amplifizierten DNA-Fragmente: Eine quasi-exponentielle Amplifikation erfolgt prinzipiell nur für die Sequenzen zwischen den Primer-Enden. Mit Syntheseraten in der Größenordnung von $10^2 - 10^3$ Nukleotide pro Sekunde reicht die mittlere Fehlerwahrscheinlich-

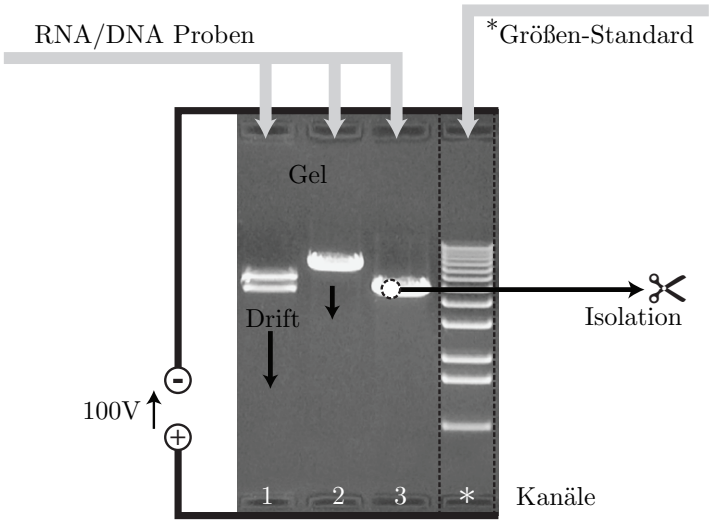

Abb. 2.8 Gelelektrophorese (nach [176]), beispielhaftes Schema.

keit von $10^{-5} - 10^{-4}$ [163] für das *Taq*-Enzym und bis zu niedrigen Fehlerraten von 10^{-6} [112] für speziellere thermostabile Enzyme.

Die PCR ist eine der elementarsten Technologien der modernen Molekularbiologie und findet in verschiedensten experimentellen Protokollen Verwendung, bei welchen eine Anreicherung von DNA notwendig ist. Als Anreicherung ist dabei zum einen die unspezifische Vervielfältigung aller DNA gemeint, wenn eine geringe Ausgangsmenge an Moleküle beispielsweise eine Sequenzierung verhindert. Zum anderen existiert eine spezifische Amplifizierung einer bestimmten Menge an DNA-Fragmenten, um durch die gezielte Verschiebung des Mengenverhältnisses eine Art *Reinigung* zu bewirken. Für eine spezifische Reaktion muss der Anfang und das Ende der Ziel-DNA bekannt und als Primer synthetisiert sein. Für eine unspezifische Anwendung wird die Ligation von definierten Adapter-Oligonukleotiden, als Bindestellen für PCR-Primer, genutzt.

Um Primer für Bereiche im *Genom* spezifizieren zu können muss die Sequenzinformation der DNA vor dem Experiment hinreichend bekannt sein. Es existieren vielfältige Ansätze zur Sequenzierung von DNA, die prinzipiell auf der komplementären Ergänzung von Nukleotiden beruhen. Im Folgenden soll der Paradigmenwechsel bei der Sequenzierung kurz skizziert werden. Dabei bildet der gesteuerte Abbruch der DNA-Replikation zusammen mit der Analyse der Länge der Replikate den Kern der Sequenzierung der ersten Generation. Die sogenannte *Gelelektrophorese* ist für die klassische Sequenzierung ein integraler biotechnologischer Bestandteil.

Begriff 12 (Gelelektrophorese)**:** Die Länge (und Reinheit) von Nukleotid-Ketten (DNA oder RNA) kann durch die Gelelektrophorese (vgl. Abb. 2.8) präzise bestimmt werden. Basis für diesen Prozess ist die Bewegung von Ladungsträgern in einem elektrischen Feld und deren längenabhängige Trägheit. Dabei bildet das *Gel* (meist ein Polymer) eine Art molekulares Sieb, durch welches die gelösten Nukleotid-Ketten diffundieren können. Die Diffusion wird angetrieben durch die Anziehungskraft des positiv geladenen Pols (Anode) und der negativen Ladung der Nukleinsäuren (Phosphat-Gruppen) der Nukleotide, wobei das Gel die Bewegung hemmt. Dabei driften kurze Molekülketten schneller als lange. Unter Verwendung von meh-

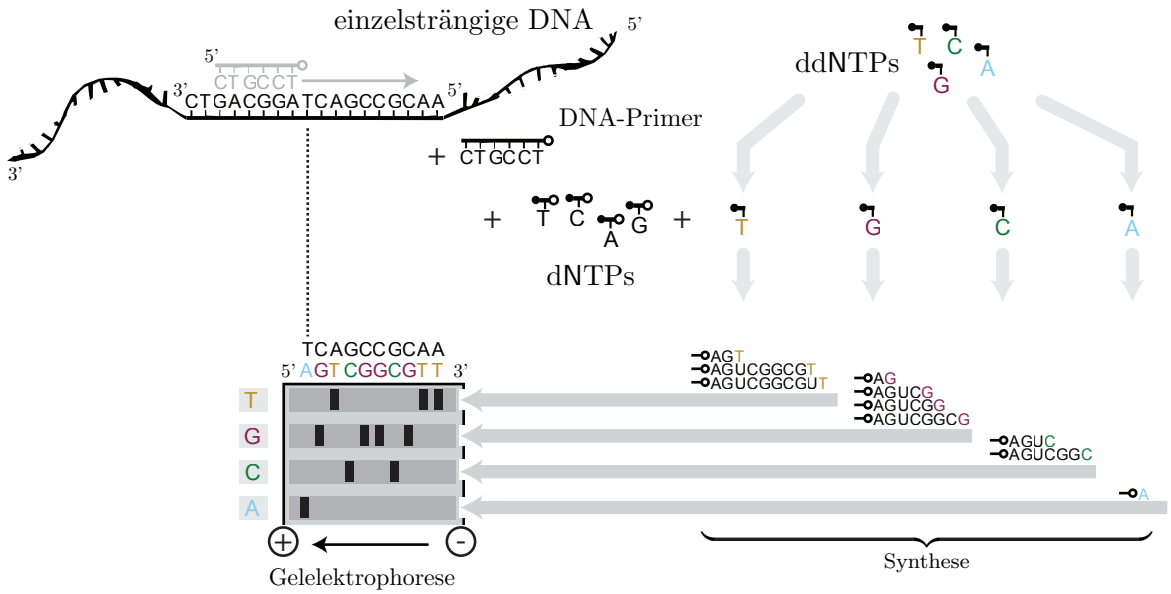

Abb. 2.9 Sanger-Sequenzierung (nach [5]): Sequenziert wird die dem Primer nachfolgende Sequenz. Durch die einzelne Zugabe von ddNTPs wird in unterschiedlichen Synthese-Reaktionen ein zufälliger nukleotid-bedingter Kettenabbruch erreicht. Somit wir eine Abhängigkeit zwischen Länge der Moleküle und Symbolen erzeugt. Die Position der Nukleotide wird über die Gelelektrophorese ermittelt (einzelne Sequenzen stehen stellvertretend für viele identische Moleküle).

reren getrennten Kanälen ist eine parallele und vergleichende Analyse von unterschiedlichen Proben möglich. Der Einsatz von speziellen Farbstoffen und Größen-Standards (DNA- oder RNA-Gefüge bekannter Längenverteilung) ermöglicht, über den Vergleich von Banden (Fortschrittslinien), Moleküle bestimmter Länge zu identifizieren und zu isolieren. Dabei ist eine Auflösung von einzelnen Nukleotiden generell möglich.

Die Gelelektrophorese ermöglicht mit weiteren Entwicklungen die ursprüngliche Form der DNA-Sequenzierung.

Begriff 13 (DNA-Sequenzierung durch Synthese: 1. Generation)**:** In Abb. 2.9 ist der klassische Ansatz der DNA-Sequenzierung nach Sanger [141, 142] illustriert. Basis hierzu ist eine standardmäßige (in vitro) DNA-Replikation (vgl. Begriff 2), welche isoliert in vier Reaktionen mit unterschiedlichen Zusammensetzungen an Nukleotid-Bausteinen durchgeführt wird. Dabei setzen sich die Moleküle zusammen aus allen vier nativen (nicht-terminierenden) Desoxyribonukleosidtriphosphate dNTP (Standard-Synthese) und jeweils einer veränderten terminierenden Nukleinsäure Didesoxyribonukleosid ddNTP, wobei N die Basen A,G,C,T symbolisiert. Die Einfügung eines ddNTP Moleküls führt dabei zum Abbruch der Doppelstrangsynthese. Ausgehend von einem einheitlichen Primer, der in allen Reaktionen den gleichen Start-Punkt der matrizengesteuerten Polymerisation bestimmt, befinden sich in jedem Reaktiongefüge letztlich unterschiedlich lange Ketten, die jedoch alle auf dem gleichen Symbol enden. Das Verhältnis zwischen terminierenden und nicht-terminierenden Basen bestimmt dabei die mittlere Anzahl dieses Symbols in der Kette, die im Allgemeinen geometrisch verteilt ist. Eine getrennte Gelelektrophorese der vier Synthese-Reaktionen ermöglicht über die Länge der Ketten auf die Position der Symbole eines Typs zu schließen.

Die heutige Revision der Sanger-Sequenzierung, wegen ihrer unübertroffenen Zuverlässigkeit noch immer genutzt, beinhaltet technische Neuerungen, arbeitet jedoch nach demselben Prinzip wie das ursprüngliche Verfahren aus den 70er Jahren. Trotz eines hohen Grades der heutigen Automatisierung verhindern das irreversible Konzept des Kettenabbruchs und die aufwendige Verarbeitung einen hohen Durchsatz an sequenzierten Symbolen.

Das Kernkonzept neuer Hochdurchsatz-Methoden zur DNA-Sequenzierung ist hingegen eine beständige Fortführung einer Kettenbildung in Verbindung mit der simultanen ortsgebundenen Detektion von molekularen Reaktionen bei der Doppelstrangbildung. Es existieren zahlreiche komplexe Technologien zur Sequenzierung, ein Übersichtsartikel [114] bietet hierzu umfassende Details.

Begriff 14 (DNA-Sequenzierung: 2. Generation)**:** Anstelle einer spezifischen Hybridisierung von Primern, die einen fixen Startpunkt bei der Sanger-Sequenzierung definiert, tritt bei den Verfahren zur *Sequenzierung der 2. Generation* (auch *Next-Generation Sequencing*, NGS, genannt) das Konzept einer unspezifischen Parallelisierung. Dieses Konzept benötigt eine spezielle biotechnologische Prozessierung des Ausgangsmaterials der Nukleotid-Ketten: Die Fragmentierung der Ketten in kürzere Moleküle, sowie die Erweiterung der Fragmente durch die Ligation spezieller Oligonukleotide sind nötige Maßnahmen, um unabhängige Einheiten der Zielsequenz zu erzeugen, welche dann parallel sequenziert werden. Eine unabhängige Einheit wird meist als *Template* bezeichnet, das darin enthaltene ursprüngliche Fragment oft auch als *Insert*. Die Gesamtheit aller sequenzierfähigen Moleküle wird *Sequenzier-Library* genannt und vom experimentellen *Protokoll* wird als Library-Herstellung gesprochen. Den erwähnten Oligonukleotiden kommen im NGS unterschiedliche Schlüsselfunktionen zu: Die Vereinzelung bzw. räumliche Separierung von Molekülen ist dabei eine Aufgabe. Anders als bei der klassischen Sequenzierung, welche getrennte Reaktionen nutzt, wird im NGS für jedes Molekül eine räumliche Trennung beabsichtigt, um den Vorgang der Synthese schrittweise für jedes Molekül nachzuverfolgen. Die Vereinzelung erfolgt dabei unter anderem durch eine zufällig verteilte Hybridisierung der Oligonukleotide an extra dafür synthetisierten Bindestellen. Diese Bindestellen sind letztlich Teil der Geometrie eines bildgebenden Verfahrens (z. B. an einer Glasoberfläche), welche genutzt wird um die Doppelstrangerzeugung zu initiieren und zu detektieren. Um die an sich unsichtbare Reaktion sichtbar zu machen, bedient man sich modifizierter Nukleotid-Bausteinen oder veränderter Hybridisierung-Reaktionen. Dabei ist der Einsatz von verschiedenen Fluoreszenzfarbstoffen für unterschiedliche Moleküle und die Anregung zur Lichtemission eine Option die Doppelstrangbildung zu analysieren. Der Prozess verläuft diskret in *Sequenzier-Zyklen*, wobei eine reversible Terminierung der Synthese eine Synchronisierung auf Symbolebene ermöglicht. Abhängig von der eingesetzten Technologie besteht eine Obergrenze an Sequenzier-Zyklen, die genutzt werden können. Das digitale Abbild der DNA wird als *Read* bezeichnet, dessen Länge durch die sogenannte *Sequenzierlänge* vorgegeben ist. Die Größenordnung der Sequenzierlänge liegt abhängig vom Verfahren im Bereich von $10^2 - 10^4$nt.

Die zur Sequenzierung nötigen Oligonukleotide erfüllen mit ihren Bindungsstellen meist noch erweiterte Aufgaben: So ist in der Library-Herstellung vor der eigentlichen Sequenzierung gegebenenfalls eine Vervielfältigung der Templates notwendig. Hierzu können Teile der Oligonukleotide als Hybridisierungs-Regionen für PCR-Primer dienen. Zusätzlich zur experimentellen Funktionalität im Protokoll bieten die erwähnten Oligonukleotide weiteren Gestaltungsraum, um beispielsweise *Barcode-Templates* mit Zusatzinformationen zu versehen. Eine mögliche Repräsentation stellen die Barcodes dar, deren Einsatz im Folgenden motiviert wird.

2.2.3 Sequenzierung und Einsatz von Barcodes

Die DNA-Sequenzierung der 2. Generation bietet ein breites Spektrum an Möglichkeiten zur Analyse von DNA und RNA. Meist beruhen die unterschiedlichen Ansätze auf Anpassungen der Library-Herstellung, so zum Beispiel der Einsatz der reversen Transkription zur Sequenzierung von RNA über cDNA. Die Abbildung von Molekülen auf eine lesbare Sequenz erfolgt dann letztendlich durch dieselbe Technologie. Am Ende der Abbildung steht für jedes prozessierte Template eine textuelle Sequenz mit der darin enthaltenen nativen Molekülsequenz des Inserts. Das Insert selbst ist der Startpunkt für eine anschließende Datenverarbeitung, in deren Zusammenhang der Begriff *Referenzgenom* eine zentrale Bedeutung spielt.

Begriff 15 (Referenzgenom)**:** Das Referenzgenom ist eine mögliche Repräsentation der Nukleotid-Abfolge des faktischen Genoms eines Organismus. Grundlage für die Erstellung eines Referenzgenoms sind sogenannte *de novo* Sequenzierungen: Basierend auf überlappende Reads ist es möglich größere Bereiche der ursprünglichen DNA aus den Fragmenten zu rekonstruierten, was auch als *Assemblierung* bezeichnet wird. Ein Referenzgenom kann jedoch zum einen Lücken enthalten, die durch die Integration späterer Sequenzierungen geschlossen werden können, andererseits kann es durch Fehler in der Sequenzierung oder natürliche Variation in der DNA des Organismus zu Mehrdeutigkeiten kommen. Die Referenz ist somit nicht absolut definierbar, sondern als wandelbare Arbeitshypothese zu verstehen.

Sobald ein hinreichend zuverlässiges Referenzgenom für den sequenzierten Organismus bekannt ist, wird es möglich Inserts aus unterschiedlichsten Sequenzierungen des gleichen Organismus auf die Referenzsequenz zu alignieren, um über den Abgleich diverse Fragestellungen zu beantworten. Das Sequenzalignment ist ein Optimierungsproblem das eine approximative Zuordnung (engl. *Mapping*) von zwei Sequenzen beschreibt (vgl. Abschnitt 2.1.3). Geht man von einer korrekten Referenz aus, so ist ein approximatives Verfahren auf Grund von Modifikationen in den sequenzierten Nukleotid-Ketten nötig.

Begriff 16 (Sequenzfehler)**:** Der Begriff *Sequenzfehler* umfasst in dieser Arbeit die Menge an symbolweisen Modifikationen, die dazu führen, dass ein Read nur approximativ auf die zu erwartenden, beinhaltenden Nukleotid-Ketten abgebildet werden kann. Auftretende Modifikationen sind dabei neben *Ersetzungen* von Nukleotiden auch die *Einfügung* oder *Löschung* von Symbolen. Während erstere als Substitutionen im Sinne eines standardmäßigen vierwertigen Kanalmodells (vgl. Definition 13) verstanden werden können, führen letztere zu einer faktischen Verlängerung oder Verkürzung von Sequenzen. Der Ursprung der Sequenzfehler ist vielschichtig: Einerseits existieren natürliche *Mutationen*, sogenannte *Punktmutationen*, die durch Fehler in Informationsspeicherung und -transfer der lebenden Zelle bedingt sind (vgl. Fehlerraten in 2.2.1) und somit die Sequenz des Inserts beeinflussen. Ein weitaus größerer Anteil an Veränderungen ist jedoch durch biotechnologische Verfahren induziert (RT und PCR in 2.2.2) und betrifft nicht nur das Insert im Template sondern auch die Oligonukleotide. Dabei kommt fehlerbehafteten Prozessen bei der Sequenzierung eine wesentliche Bedeutung zu. Einige Technologien sind beispielsweise für eine hohe Rate an Einfügung oder Löschung bekannt, die eine adäquate approximative Zuordnung vor besondere Herausforderungen stellt.

Begriff 17 (Mapping)**:** Für das Mapping von Inserts auf ein Referenzgenom besteht mittlerweile ein breites Repertoire an vorgeschlagenen Ansätzen. Der in [50] gegebene Überblick bietet eine umfassende Auflistung aktueller Algorithmen. Die aus der Bioinformatik im Mittel mehr als 100 mal pro Jahr zitierten Algorithmen wie BLAT [87], Bowtie(2) [95, 96], BWA [100], MAQ [101], SOAP [102] oder TopHat [164] nutzen meist einen im Voraus berechneten Index (des Referenzgenoms) basierend auf der Burrows-Wheeler-Transformation, Tries oder Suffixtries für eine schnelle Suche von kurzen, teilweisen Übereinstimmungen mit dem Insert. Die Parametrisierung des Index, die berücksichtigten Typen von Fehlern und die genauen Strategien zur Fortsetzung eines Sequenzalignments sind dabei sehr unterschiedlich definiert, was eine direkte Gegenüberstellung der Algorithmen schwierig gestaltet. Auch folgt das Mapping von Inserts aus RNA oder DNA unter Umständen anderen Prinzipien. Im Allgemeinen gilt eine erfolgreiche Zuordnung von $80 - 95\%$ [50] der sequenzierten Reads als realistisch.

Einer Sequenzierung zugrundeliegende Fragen haben generell eine *qualitative Komponente*, welche um eine *quantitative Dimension* erweitert werden kann: Die Sequenzierung von DNA ist meistens darauf ausgerichtet einen genetischen Vergleich oder einen Nachweis über genetische Veränderungen zu führen. Eine zuverlässige Erkennung dieser Modifikationen und deren Abgrenzung zu verfahrenstechnischen Sequenzfehlern ist dabei ein Kernaspekt. Ob ein bestimmter Abschnitt der DNA hingegen mehrfach in den Sequenzierdaten auftritt, spielt dabei eine untergeordnete Rolle. Anders hingegen ist es, wenn die *Transkriptomik* (Analyse der Transkription) über die Sequenzierung von RNA analysiert werden soll. Hierzu ist die ursprünglich in der Zelle vorliegende Anzahl von mRNA Molekülen (oder deren Verhältnis) von Interesse, um mittelbar Aussagen über die Transkription von Genen zu machen. Ein Schritt in der Library-Herstellung der dabei einen signifikanten quantitativen Einfluss auf das Verhältnis der initial vorliegenden Templates hat ist die PCR. Der nachteilige quantitative Effekt wird auch als *PCR-Bias* bezeichnet.

Begriff 18 (PCR-Bias)**:** Abstrahiert man die in Abb. 2.7 illustrierte PCR, so lässt sich der PCR-Bias besser erklären. Vereinfacht gesehen ist die zyklische Replikation der PCR für jedes DNA-Molekül ein individueller stochastischer Prozess, der sich auf die Wahrscheinlichkeit $0 \leq \lambda \leq 1$ einer Verdopplung reduzieren lässt. Sei N_0 eine initiale Ganzzahl und N_i für alle $i > 0$ eine Zufallsvariable, welche die Anzahl der Moleküle im Zyklus i definiert, so beschreibt

$$N_{i+1} = \sum_{j=0}^{N_i} (1 + I_j), \quad \text{mit } \Pr(I_j = x) = \begin{cases} \lambda, & \text{für } x = 1, \\ \bar{\lambda}, & \text{für } x = 0, \end{cases}$$

die Anzahl der Moleküle im Folgezyklus. Dabei ist $\bar{\lambda} = 1 - \lambda$ und I_j eine Bernoulli-verteilte Zufallsvariable (Definition 2), welche eine Verdopplung anzeigt. Ist jedes I_j eine unabhängige gleichverteilte Zufallsvariable, so beschreibt N_i einen sogenannten Galton–Watson Prozess [156]. Bekannt ist weiterhin für den Erwartungswert und die Varianz von N_n, dass

$$\mathrm{E}(N_n) = N_0 \lambda_*^n \quad \text{und} \quad \mathrm{Var}(N_n) = \lambda \lambda_*^{n+1} (\lambda_*^n - 1)$$

gilt, mit $\lambda_* = \mathrm{E}(I_j) = 1 + \lambda$ (vgl. [70]). Während die erwartete Anzahl an Molekülen exponentiell pro Zyklus wächst, gilt gleiches auch für die Varianz (für $0 < \lambda < 1$). In Abb. 2.10 sind

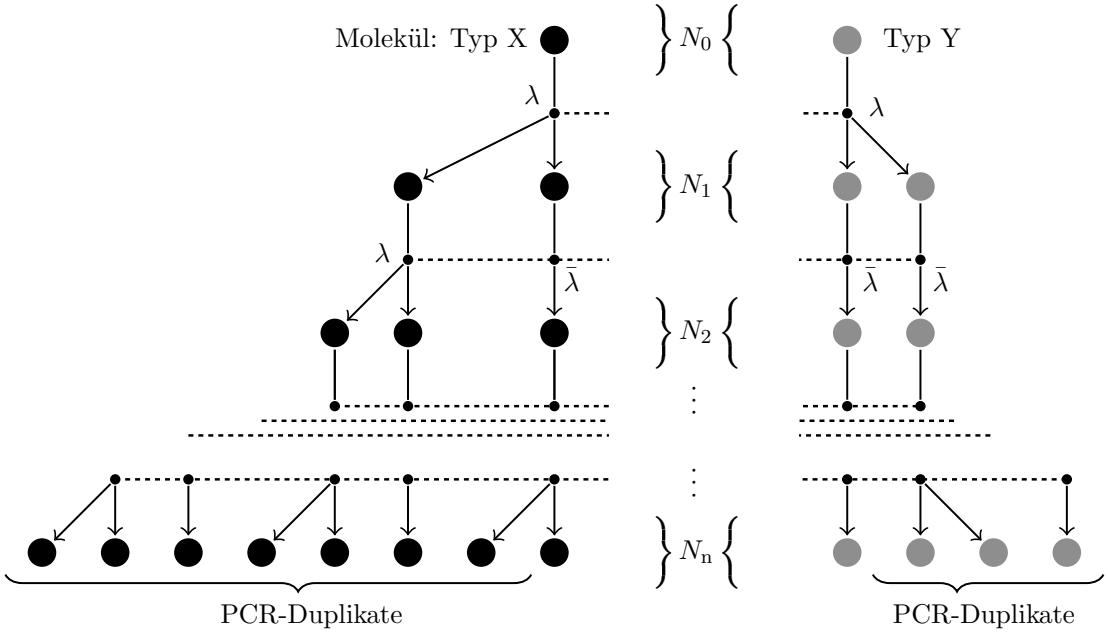

Abb. 2.10 PCR-Bias: Ungleichförmige Vervielfältigung von Fragmenten als stochastischer Prozess (exemplarische Verläufe). Ausgehend von gleichen Ausgangsmengen N_0 an unterschiedlichen Typen von Molekülen, ist der Ausgang der Kettenreaktion nur im Mittel identisch zu $\mathrm{E}(N_n) = N_0(1+\lambda)^n$. Dabei steht λ (als Effizienz) für die Wahrscheinlichkeit einer Molekül-Replikation.

zwei mögliche Verläufe des Zufallsprozesses dargestellt, welche das Kernproblem der zufälligen Replikation verdeutlichen.

Für die reale enzymatische Reaktion der PCR ist weiterhin bekannt, dass deren Effizienz (λ) über die Zyklen variiert. Aufgrund von physischen Limitationen gilt sicher $\lambda \to 0$ für große n. Des Weiteren gibt es nachweisbare Zusammenhänge, welche die Effizienz mit der Sequenzeigenschaft von Molekülen in Verbindung bringen (vgl. beispielsweise [2, 4, 182]). Der gesamte Kontext der ungleichförmigen Vervielfältigung in der PCR wird unter dem Begriff *PCR-Bias* zusammengefasst.

Möchte man auf der Grundlage einer Sequenzierung der *PCR-Duplikate* auf die initiale Anzahl von Molekülen rückschließen, so ist eine zuverlässige Aussage ohne Zusatzinformation nur bedingt möglich. Ein solcher Zusatz kann durch Oligonukleotide in Form von Barcodes bei der Library-Herstellung erreicht werden.

Einsatz von Barcodes und Anforderungen

Zwei Einsatzmöglichkeiten von Barcodes sind zum einen die dedizierte (explizite) Anwendung zum sogenannten *Multiplexing* und zum anderen die erweiterte Verwendung als *Zufallsbarcodes* zur *PCR-Korrektur*. Beispielhafte Publikationen die den Einsatz von Barcodes zum Multiplexing thematisieren sind [19, 29, 32, 54, 68, 85, 92, 115, 118], die Anwendung von Zufallsbarcodes in der Sequenzierung wird unter anderem in [82, 88, 153] ausgeführt.

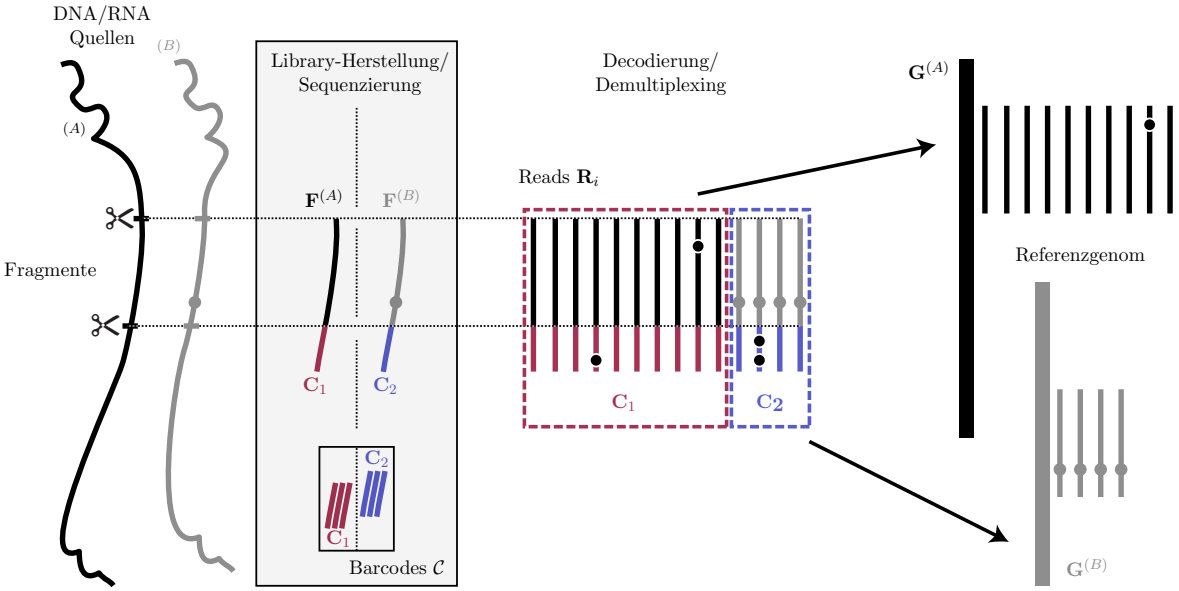

Abb. 2.11 Barcode-Multiplexing: Die Verwendung von Barcodes zur gemeinsamen Sequenzierung unterschiedlicher Proben und der anschließenden Separierung der Quellen (dargestellt als Rechtecke). Barcodes sollten durch die Korrekturfähigkeit robust gegenüber Sequenzfehlern (als Punkte symbolisiert) sein, um eine fehlerhafte Zuordnung der Proben zu vermeiden.

Begriff 19 (Multiplexing)**:** Für eine Vielzahl von Analysen stellt der enorme Durchsatz der neuen Sequenziertechnologien eine Überkapazität dar. Um die hohe Anzahl an möglichen Reads pro Sequenzierlauf parallel für mehrere experimentelle Auswertungen nutzen zu können, bietet sich das Multiplexing an, das beispielhaft in Abb. 2.11 dargestellt ist: Fragmente $\mathbf{F}^{(A)}$ und $\mathbf{F}^{(B)}$ von unterschiedlichen experimentellen Proben (A/B) werden in der Library-Herstellung mit unterschiedlichen Barcodes aus einem Satz $\mathcal{C}$ an Codeworten versehen und als Templates gemeinsam sequenziert. Die resultierenden Reads $\mathbf{R}_i$ werden durch die Decodierung der Barcodes im Anschluss klassifiziert (um die Barcodes verkürzt) und den ursprünglichen genomischen Quellen $\mathbf{G}^{(A)}$ und $\mathbf{G}^{(B)}$ zugeordnet (*Demultiplexing* genannt). Die Robustheit der Barcodes gegenüber Sequenzfehlern sollte dabei möglichst groß sein, um ein Übersprechen der unterschiedlichen Proben zu verhindern. Die Anzahl benötigter Codeworte liegt dabei in der Größenordnung der unterschiedlichen Proben.

Während das Multiplexing als Methode zur Unterscheidung von Reads aus mehreren Experimenten genutzt wird, kann das Konzept der Zufallsbarcodes als Erweiterung des Multiplexing zur Zählung von Molekülen in einer Probe verwendet werden.

Begriff 20 (Zufallsbarcodes zur PCR-Korrektur)**:** Für einige Anwendungen stellt die ungleichmäßige Vervielfältigung von Molekülen durch die PCR ein elementares Problem dar, welches durch den Einsatz von Zufallsbarcodes gemindert werden kann. In Abb. 2.12 ist die PCR-Korrektur vereinfacht dargestellt: Ausgangspunkt ist eine quantitative Analyse für eine Quelle von RNA-Molekülen. Dazu existiert ein bestimmtes RNA-Fragment $\mathbf{F}$ vor der Library-Herstellung unter Umständen mehrfach. Anders als beim Multiplexing werden Barcodes als Fusion von Oligonukleotiden eingesetzt um Fragmente damit zufällig zu markieren. Durch

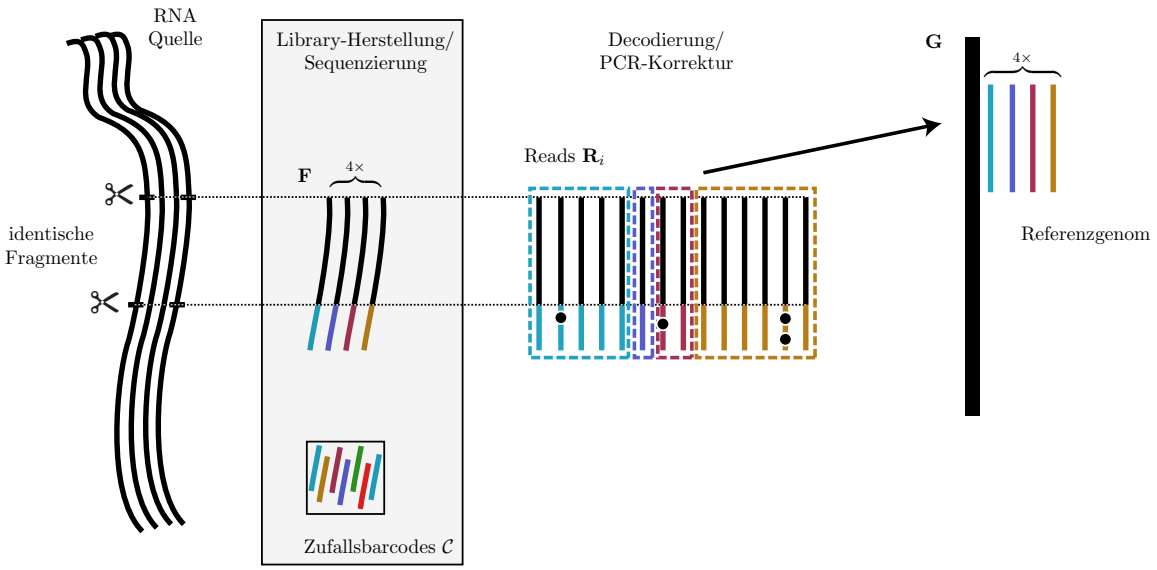

Abb. 2.12 Zufallsbarcodes zur PCR-Korrektur: Die Markierung von identischen Fragmenten mit Zufallsbarcodes ermöglicht die Korrektur von PCR-Duplikaten unabhängig von einer ungleichmäßigen Replikation. Die Robustheit der Zählung gegenüber Sequenzfehlern (als Punkte symbolisiert) kann durch die Verwendung von Fehlerkorrektur in den Codeworten erhöht werden.

den Einsatz von Zufallsbarcodes $\mathcal{C}$ mit einer hohen *Diversität* soll die Wahrscheinlichkeit, dass zwei identische Fragmente mit dem gleichen Codewort versehen werden möglichst klein sein. Dabei ist die Diversität die Kombination aus einer hohen Anzahl unterschiedlicher Codeworte, die in ihrem Vorkommen möglichst uniform an der Markierung beteiligt sind. Der Idealfall ist in Abb. 2.12 dargestellt: Durch die eindeutige zufällige Markierung der vier identischen Fragmente $\mathbf{F}$ ist eine fehlerfrei Korrektur der PCR-Duplikate in den Reads $\mathbf{R}_i$ möglich. Die Korrektur ist dabei unabhängig von der ungleichmäßigen Vervielfältigung. Um eine Robustheit der Zählung von Molekülen zu gewährleisten, ist auch im Fall der Zufallsbarcodes eine Fehlerkorrektur erforderlich. Die Anzahl benötigter Codeworte liegt dabei deutlich über der Größenordnung der erwarteten identischen Fragmente, welche markiert werden sollen.

Prinzipiell ist für das Multiplexing und die PCR-Korrektur der Einsatz der gleichen Barcodes möglich. Da jedoch der Aufwand der dedizierten synthetischen Erzeugung und Verwendung von Barcodes nicht unerheblich ist, bedient man sich in der Praxis öfters der molekularen Kombination von kürzeren Oligonukleotiden, um die Diversität kostengünstig zu vergrößern. Mögliche Ansätze dazu wurden z. B. in [77] diskutiert. Durch die Kombination von Oligonukleotiden ergeben sich jedoch spezielle Einschränkungen hinsichtlich der Struktur an möglichen Codeworten und zusätzliche Restriktionen für die synthetischen Sequenzen.

3

Barcodes als Watermark Codes

Parallelisierung ist ein wesentliches Grundprinzip der DNA-Sequenzierung in ihrer zweiten Generation und ermöglicht, seit dem Aufkommen dieser neuen Technologie zum Jahrhundertwechsel, eine stetige Steigerung des Durchsatzes. Dieser wird dabei durch Länge und Anzahl von DNA-Templates bestimmt, welche in einem Verarbeitungsvorgang des Sequenziergerätes in eine digitale Sequenzrepräsentation abgebildet werden können. So lässt sich der momentan mögliche Durchsatz von Symbolen mittlerweile in den Größenordnungen von Giga und Tera angeben. Für eine Vielzahl von Sequenzierungen ist eine derartige Kapazität jedoch nicht nötig, um spezifische Nachweise zu erbringen oder Hypothesen zu verifizieren. Diese Überkapazität für bestimmte Anwendungen motiviert die Idee der effizienten parallelen Nutzung von Sequenziergeräten für unterschiedliche Sequenzierexperimente. Das *Multiplexing* stellt das dafür mögliche Schlüsselkonzept dar: DNA-Fragmente unterschiedlichster Experimente/Organismen (allgemein Proben) werden durch individuelle Barcode-Sequenzen (synthetische DNA) markiert und gemeinsam (im Verbund) sequenziert. Die Barcodes dienen im Anschluss an die Sequenzierung dazu die Menge der Ausgabesequenzen den einzelnen Quellen zuzuordnen. Diese Trennung als Schritt der Nachverarbeitung der Sequenzierdaten wird auch *Demultiplexing* genannt. Die Robustheit der Multiplex-Verfahren hinsichtlich experimenteller Fehlschlüsse, die durch eine falsche Zuordnung von Templates induziert werden, beruht im Wesentlichen auf der Robustheit der verwendeten Barcodes und wie gut diese an die Sequenzierplattform und die darin auftretenden Fehlercharakteristika angepasst sind.

In ersten Publikationen [19, 85, 115] die den spezifischen Einsatz von Barcodes zur Sequenzierung beschreiben, wurde das Konzept der Kanalcodierung und Fehlerkorrektur zur Auswahl von Codeworten nicht integriert. In anderen Bereichen der Mikrobiologie, beispielsweise den sogenannten Microarrays, wurden für die Konzeption von Oligonukleotiden [15, 104] bereits *Hamming-Codes* [69] in Erwägung gezogen (vgl. [29]). Spätere Ansätze nutzten diese Konstruktionen der klassischen Kanalcodierung dann explizit zur Erzeugung von Barcodes, wie beispielsweise in [32, 68]. Des Weiteren wurden auch andere klassische Codes, wie die *BCH-Codes* [21, 75] in [92], als Vorbild für Barcodes verwendet. Neben dem Einsatz der beispielhaft genannten linearen Codes wurde für kurze Barcode-Sequenzen auch die Klasse der nicht-linearen Codes in Betracht gezogen, so zum Beispiel in den Ansätzen von [54] oder [118], welche randomisierte Suchalgorithmen nutzen, um DNA-Sequenzen mit maximalem Abstand zueinander zu finden.

Vom konzeptionellen Standpunkt aus sind die bisher genannten Ansätze zur Erzeugung von Barcodes darauf ausgelegt Codeworte zu finden, die bestimmte Anforderung bezüglich des Hamming-Abstands erfüllen und somit eine klar definierte Korrektur von Symbolersetzungen zu ermöglichen. Jedoch existieren einige Sequenziertechnologien, die bekanntermaßen neben Ersetzungen eine außerordentlich hohe Anzahl an Einfügungen oder Löschungen als Sequenz-

fehler produzieren können. Diese Art von Modifikationen der Sequenzen werden auch *Indels* genannt und sind mit einer faktischen Vergrößerung oder Verkleinerung der Sequenzlängen verbunden: Die Plattformen namens *Roche 454 Pyrosequencing* [61], *Pacific Biosciences (PacBio)* [33] oder *Applied Biosystems (Ion Torrent PGM)* [23] sind für derartige Sequenzfehler bekannt (siehe auch Übersichtsartikel [105, 152, 181]). Speziell für diese Geräte ist der Einsatz von Codes auf Basis des Hamming-Abstands suboptimal und es ist unerlässlich, die Korrektur aller Sequenzfehler zu berücksichtigen.

Als wegweisende Ansätze für die Erstellung von Codes im Kontext der DNA, welche als robust gegenüber Indels gelten, sind die Publikationen von Ashlock et al. [8, 9] zu nennen. Sie befassen sich mit dem Entwurf und dem Einsatz von Greedy-Algorithmen (später evolutionäre Algorithmen) zur Suche nach kurzen Codes mittleren Umfangs mit hoher Editierdistanz. Diese Art von Suchalgorithmen ist sehr aufwändig und nur für kleine Mengen kurzer Codeworte sinnvoll einsetzbar. Zur Berechnung des Abstands zweier Codeworte berücksichtigt die gewählte Metrik der Editierdistanz neben Ersetzungen auch Einfügungen bzw. Löschungen und ist damit besser geeignet für Übertragungswege mit dieser Art von Sequenzfehlern. Leider ist bis heute keine Codekonstruktion bekannt, die eine komplexere Partitionierung von Codeworten in dieser Metrik liefert. So existieren, ausgehend von den genannten Publikationen, vielversprechende Ansätze wie [3, 48, 129], die sich mit der randomisierten Suche nach Codes mit hoher Editierdistanz befassen (siehe Übersicht in [48]).

Unabhängig von der theoretischen Korrekturfähigkeit der fokussierten Codes in Editier-Metrik wurde die praktische Decodierung von Codeworten im Kontext von weiteren Symbolen an den Grenzen der Barcodes in vielen Veröffentlichungen nicht thematisiert. Praktisch existiert bei der Sequenzierung ja nicht nur das Codewort, sondern zusätzlich dazu noch Nukleotide der Ziel-DNA oder weitere technisch bedingte Oligonukleotide. In [29] wurde letztlich gezeigt, dass die Maximierung der Editierdistanz in einem erweiterten Sequenzkontext suboptimal (oder sogar falsch) ist und als korrektes Kriterium die sogenannte *Sequenz-Levenshtein Distanz* eingeführt. Das in [29, 30] erklärte und berücksichtigte Phänomen beschreibt den Einfluss von zusätzlichen Symbolen an einem Ende eines Barcodes und das Auftreten von Indels als zusätzliche Reduktion der Editierdistanz bei der Decodierung. Die genannte Sequenz-Levenshtein Distanz berücksichtigt diesen Effekt im Abstandsbegriff und führt damit, als Distanzmaß, zu Codeworten mit erweiterter Editierdistanz. Der generalisierte Fall der zweiseitigen Einbettung von Barcodes wird jedoch auch von der Sequenz-Levenshtein Distanz nicht berücksichtigt. Denn ganz allgemein betrachtet sind beide Blockgrenzen der Barcodes nach der Sequenzierung nicht mehr exakt definierbar.

Das folgende Kapitel bietet eine gänzlich andere Perspektive auf das Thema der Barcodes für die Fehlerkorrektur von allgemeinen Sequenzfehlern und der Einbettung von Barcodes in Templates. Anstatt einer strikt auf Distanz basierten Auswahl von Barcodes und deren Decodierung wird das Konzept der sogenannten *Watermark Codes* genutzt, um Barcodes zu erzeugen und eine probabilistische Decodierung zu ermöglichen. Das Prinzip der Watermark Codes wurde ursprünglich von Davey und MacKay [40, 41] vorgestellt, um die Synchronisierung und Decodierung für die kontinuierliche Übertragung über einen binären Kanal zu ermöglichen, wenn in diesem sowohl Ersetzungen als auch Einfügungen oder Löschungen von Bits auftreten können. Der Einsatz von Watermark Codes als einzelne kurze Barcodes erfordert jedoch Anpassungen der Modelle und birgt zusätzliche Einschränkungen für die Konstruktion von praktikablen Co-

deworten. Watermark Codes wurden in [72, 73] bereits im Kontext der DNA verwendet, jedoch mit der Intention synthetische Daten im Genom von lebenden Zellen zu integrieren und über den sogenannten evolutionären Kanal (den Generationswechsel) zu erhalten. Der Kern der benannten Ansätze weist bei detaillierter Betrachtung nur eine augenscheinliche Ähnlichkeit mit der hier gezeigten Anwendung der Watermark Codes auf.

Dieses Kapitel gliedert sich wie folgt: In Abschnitt 3.1 wird die Verallgemeinerung von Modellen für die Beschreibung von Sequenzen im Wertebereich der Nukleotide dargelegt. Ausgehend von der Darstellung der Sequenzrepräsentation der Templates und der damit verbundenen Einbettung von Codeworten in 3.1.1, wird das Kanalmodell der Sequenzierung in 3.1.2 eingeführt. Jenes Modell dient letztlich als Vorbild für ein Hidden Markov-Modell, das in 3.1.3 näher spezifiziert wird und das Kernelement der Decodierung der Watermark Codes darstellt. Die optimale Decodierung und probabilistische Mustersuche in 3.2.1 bildet die Motivation für die Erklärungen der Codierung und Decodierung der Barcodes basierend auf den Watermark Codes in Abschnitt 3.2.2 respektive 3.2.3. Konkrete Realisierungen von Codierschemata (in Abschnitt 3.3.1) und Zusatzbedingungen (3.3.2), die eine Suche nach praxisrelevanten Barcodes bedingen, bilden den ersten Teil von Abschnitt 3.3. Die Charakterisierung der Eigenschaften der gefundenen Barcodes ist der Inhalt von Abschnitt 3.3.3, an welchen sich in 3.3.4 die simulative Evaluation von Barcodes anschließt. Letztlich endet dieses Kapitel in 3.4 mit einer Zusammenfassung und kritischen Fragen für weiterführende Themen. Bestandteile dieses Kapitels zur Nutzung von Watermark Codes im Kontext der DNA-Sequenzierung wurden in [90, 91] veröffentlicht.

3.1 Modelle und Annahmen

Betrachtet man das als Konzept 4 auf Seite 16 beschriebene und in Abb. 2.1 illustrierte Übertragungsmodell der Kanalcodierung, so lässt sich dieses nicht ohne Weiteres auf die Sequenzierung von Nukleotiden und den Einsatz von Barcodes übertragen. In diesem Abschnitt wird ein erweitertes kommunikationstheoretisches Modell eingeführt, um das Multiplexing mittels Barcodes und deren fehlerbehaftete Sequenzierung adäquat zu beschreiben. Unter der Annahme eines statistisch unabhängigen Auftretens von allgemeinen Sequenzfehlern bei der Sequenzierung, wird dieses Verhalten danach in einem vereinfachten Kanalmodell beschrieben.

3.1.1 Sequenzrepräsentation und Einbettung von Codeworten

Die Übertragung von Information beim Multiplexing in der Sequenzierung ist zweigeteilt und besteht nicht nur aus der Übertragung und Decodierung von reinen Codeworten: Vielmehr ist die Sequenz des Codewortes als Barcode eine Art Metainformation, welche in den Kontext eines Templates (siehe Begriff 14) eingebettet ist. In Abb. 3.1 ist die Sequenzierung als Übertragungsmodell dargestellt: Sei $\mathbf{c} \in \mathcal{C} \subseteq \mathcal{A}^n$ ein Barcode eines bestimmten Codes $\mathcal{C}$ der Länge n aus dem Alphabet $\mathcal{A} = \{\mathsf{A}, \mathsf{G}, \mathsf{C}, \mathsf{T}\}$ der Nukleotide, so erhält man die (theoretisch) übertragene Sequenz

$$\mathbf{t} = (\underbrace{t_1 t_2 t_3 ...}_{\mathbf{p}_1} \underbrace{t_{\delta+1} .. t_{\delta+n}}_{\mathbf{c}=c_1 c_2 ... c_n} \underbrace{...t_{L-1} t_L}_{\mathbf{p}_2}) \in \mathcal{A}^L \tag{3.1}$$

des Templates einer Länge L als Verbund eines Präfixes $\mathbf{p}_1$ und Postfixes $\mathbf{p}_2$, mit Barcode $\mathbf{c}$ dazwischen. Die übertragene Sequenz enthält neben einem Codewort noch ein Insert $\mathbf{f}$ (hier

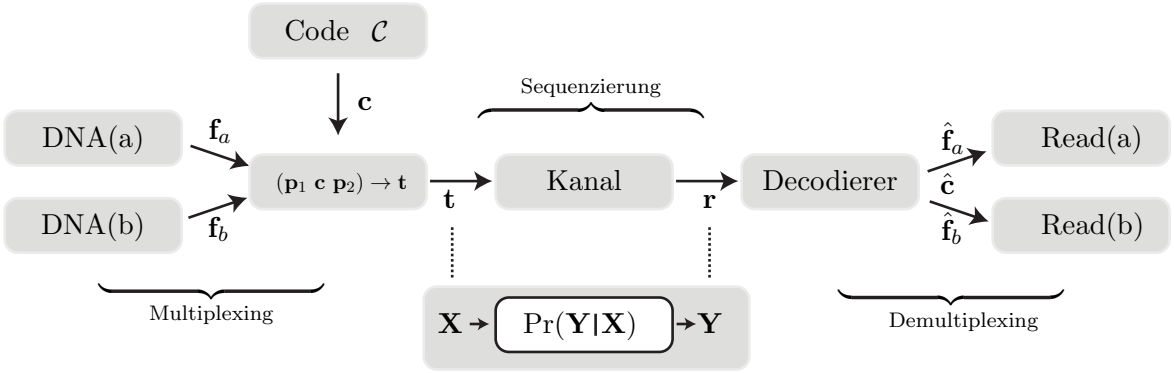

Abb. 3.1 Einbettung von Codeworten (vgl. mit Standard Kanalmodell Abb. 2.1): Modell des Multiplexings, der Sequenzierung und des Demultiplexings.

beispielsweise ein DNA-Fragment $\mathbf{f}_a$ oder $\mathbf{f}_b$), das ganz allgemein entweder in $\mathbf{p}_1$ oder $\mathbf{p}_2$ enthalten sein kann. Hierbei stehen die Indexe a und b stellvertretend für unterschiedliche Klassen von Fragmenten (Proben), die mit unterschiedlichen Codeworten $\mathbf{c} \in \mathcal{C}$ markiert werden. Neben einem DNA-Fragment (einer Probe) und den Barcodes (Kennzeichnungen) können noch weitere Oligonukleotide Teil des Templates $\mathbf{t}$ sein. Im hier vorgestellten Modell wird allen Teilsequenzen, die kein Codewort darstellen, keine weitere Bedeutung beigemessen, als dass deren Vorhandensein und zufällige Länge eine Verschiebung $\delta \in \{0, 1, .., L - n\}$ zur Folge hat, die als Realisierung einer Zufallsvariablen verstanden werden kann. Zur weiteren Vereinfachung sei angenommen, dass sowohl δ als auch die Sequenzen $\mathbf{p}_1$ und $\mathbf{p}_2$ zufällig gewählt seien, gleichverteilt auf ihrer jeweiligen Wertemenge. Die Sequenz $\mathbf{t}$ ist dabei ein theoretisches Konstrukt, das als idealisierte Repräsentation einer Molekülkette gesehen werden kann, die (mit heutigem Kenntnisstand) nicht direkt beobachtet werden kann. Die Sequenzierung des Templates resultiert letztlich in der (beobachtbaren) Sequenz $\mathbf{r} \in \mathcal{A}^{L'}$, dem Read einer Länge L'. Der Übergang von Template $\mathbf{t}$ zum Read $\mathbf{r}$ wird dabei von der Übergangswahrscheinlichkeit $\Pr(\mathbf{Y} = \mathbf{r}|\mathbf{X} = \mathbf{t})$ beschrieben. Die Aufgabe des Decodierers ist es auf Basis von $\mathbf{r}$ eine Entscheidung zu treffen, welches Codewort $\hat{\mathbf{c}}$ darin enthalten ist, um damit das Template der ursprünglichen Quelle zuzuordnen. Die Klassifikation der Reads anhand von Barcodes wird auch als Demultiplexing bezeichnet.

3.1.2 Kanalmodell der Sequenzierung

Geht man von einem einfachen Substitutionsmodell aus, wie es für Barcodes beispielsweise in [118] verwendet wurde, so besteht der Kanal aus einer probabilistischen Abbildung

$$\mathcal{A}^L \to \mathcal{A}^{L'},\ \mathbf{t} \mapsto \mathbf{r}, \tag{3.2}$$

mit $L' = L$. Ferner gilt für jede Position $i \in \{1, 2, .., L\}$ eine direkte Abhängigkeit

$$\mathcal{A} \to \mathcal{A},\ t_i \mapsto r_i,$$

die durch eine Übergangswahrscheinlichkeit $\Pr(Y = r_i|X = t_i)$ beschrieben wird. Ein weitaus allgemeineres Modell für Sequenzfehler erhält man, wenn man für alle $i \in \{1, 2, .., L\}$ und

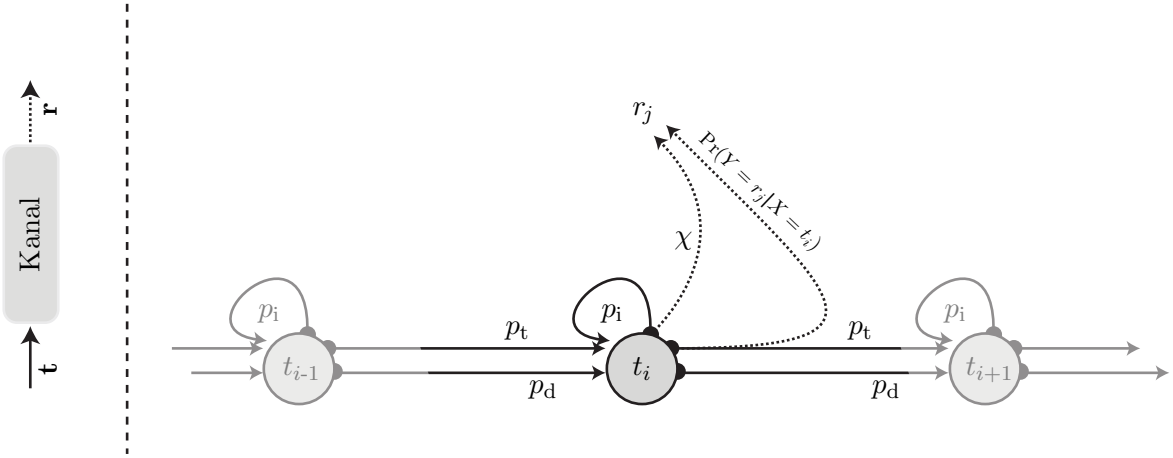

Abb. 3.2 Kanalmodell der Sequenzierung, Modellierung als Zustandsautomat (nach [90]): zufällige Zustandsänderungen (durgehende Linien), Ausgaben von Symbolen (unterbrochene Linien).

für alle $j \in \{1, 2, .., L'\}$ allgemeine symbolweise Abbildungen $t_i \mapsto r_j$ berücksichtigt und die Gleichheit $i = j$ respektive $L' = L$ vernachlässigt. Im Wesentlichen erhält man dadurch das Kanalmodell, welches für den binären Fall von Davey und MacKay [41] vorgestellt wurde. Das Kanalmodell lässt sich als probabilistischer Zustandsautomat (siehe Abb. 3.2) implementieren, der sich durch die Menge

$$\mathcal{H} : \{p_\mathrm{i}, p_\mathrm{d}, \Pr(Y|X)\} \tag{3.3}$$

an Parametern charakterisieren lässt. Dabei beschreibt p_i und p_d die Wahrscheinlichkeit einer Einfügung respektive Löschung und $\Pr(Y = r_j|X = t_i)$ ist die Wahrscheinlichkeit das Symbol r_j an Position j im Read zu beobachten, wenn t_i an Position i im Template dieses Symbol bedingt. Dabei werden die Symbole t_i ebenso wie die Positionen i, j als Zustandsvariablen betrachtet, welche durch zufällige Ereignisse verändert werden. Neben dem Wechsel von Zuständen können Zustandsübergänge die Ausgabe eines Symbols r_j bedingen. Es existieren, ausgehend vom initialen Zustand $i = j = 1$, drei mögliche Zufallsereignisse:

- Löschung (Ereignis $p_\mathrm{d}[i{+}{+}]$): Mit Wahrscheinlichkeit p_d erfolgt lediglich das Inkrement (dargestellt als $+{+}$) der Position i und keine Ausgabe eines Symbols r_j.
- Einfügung (Ereignis $p_\mathrm{i}[r_j = \chi, j{+}{+}]$): Mit Wahrscheinlichkeit p_i erfolgt die Ausgabe eines zufälligen Symbols $\chi \in \mathcal{A}$ aus dem Alphabet der Nukleotide gefolgt vom Inkrement $j{+}{+}$.
- Ersetzung (Ereignis $p_\mathrm{t}[r_j : \Pr(Y = r_j|X = t_i), i{+}{+}, j{+}{+}]$): Mit einer Wahrscheinlichkeit $p_\mathrm{t} = 1 - p_\mathrm{d} - p_\mathrm{i}$ erfolgt die Ausgabe eines zufälligen Symbols r_j auf Grundlage der einheitlichen bedingten Wahrscheinlichkeit $\Pr(Y|X)$, gefolgt von $i{+}{+}$ und $j{+}{+}$.

Abhängig von den Parametern $\mathcal{H}$ erlaubt das Modell für eine endliche Sequenz $\mathbf{t}$ gegebenenfalls eine unbeschränkte Länge bei der Realisierung von $\mathbf{r}$. Beschränkt man jedoch die möglichen Zustände (und Übergänge) des dargelegten Zustandsautomates, so ermöglicht das eine approximative Beschreibung eines endlichen Ensembles von Sequenzen $\mathbf{r}$ mittels Hidden Markov-Modell (HMM), wie als binärer Beschreibung in [40] gezeigt wurde. Das im Folgenden dargelegte HMM ist der Kern für die später gezeigte Decodierung von Barcodes.

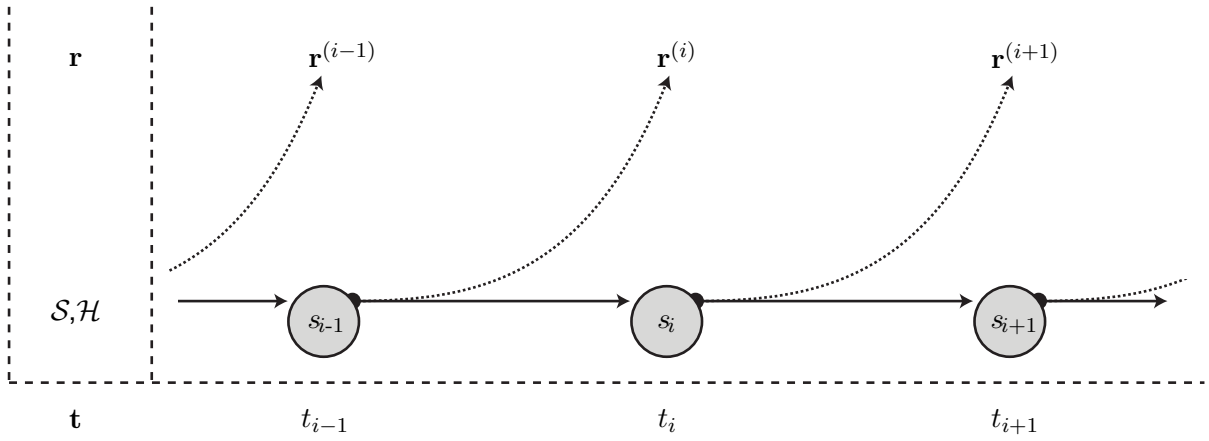

Abb. 3.3 HMM des Sequenzierkanals (nach [90, 91])

3.1.3 Hidden Markov-Modell der Sequenzierung

Betrachtet man den Zustandsautomat des letzten Abschnitts, so lassen sich die dargelegten Zustände auf einen Zustandsraum $\mathcal{S}$ und die Sequenz $\mathbf{t}$ (als zusätzlichen Modellparameter) reduzieren. Folgendes Beispiel soll die geänderte Perspektive motivieren: Für die symbolweise Abhängigkeit $t_i \leftrightsquigarrow r_j$, mit $i = j$, führt eine einzelne Einfügung vor Position i dazu, dass t_i versetzt, anhand von r_{i+1} beobachtet wird. Gleiches gilt für die Löschung eines Symbols vor Position i, welche einer Verschiebung der Abhängigkeit von t_i auf r_{i-1} entspricht. Die differenzielle Betrachtung der Positionen i und j führt zu folgender Definition: Sei $i \in \{1, 2, .., L\}$, als Position i in der Sequenz $\mathbf{t}$, der Zeitparameter eines HMMs (vgl. Definition 9), so beschreibt der sogenannte Drift $s_i \in \mathcal{S} \subseteq \mathbb{Z}$ die Zahl der Einfügungen abzüglich der Anzahl Löschungen vor Berücksichtigung von Symbol t_i. Die ganzzahligen Werte der Menge $\mathcal{S}$ definieren die möglichen nicht-sichtbaren Zustände (engl. *hidden states*) des Modells. Ein Drift s_i impliziert somit einen Zusammenhang der Symbole $t_i \leftrightarrow r_{i+s_i}$. Zusätzlich dazu beschreibt

$$\mathbf{r} = r_1 r_2 ... r_L = \mathbf{r}^{(1)} \mathbf{r}^{(2)} ... \mathbf{r}^{(L)}$$

eine Partitionierung von $\mathbf{r}$ in Subsequenzen $\mathbf{r}^{(i)}$, welche als mehrstellige Beobachtungen des HMMs verstanden werden. Anders als in der herkömmlichen Definition eines HMMs ist die Beobachtung der Sequenz $\mathbf{r}^{(i)}$ nicht nur durch einen einzigen unsichtbaren Zustand bedingt, sondern abhängig vom Verbundereignis $(s_{i-1}, s_i) \in \mathcal{S} \times \mathcal{S}$ und dem Symbol t_i. Für eine weitere Parametrisierung des HMMs sind die möglichen beobachtbaren Sequenzen $\mathcal{R}$ und deren Limitierung von weiterem Interesse. Allgemein gilt

$$\mathbf{r}^{(i)} \in \left\{ \epsilon, t_i, \bar{t}_i, \chi t_i, \chi \bar{t}_i, \chi\chi t_i, \chi\chi \bar{t}_i, .., \underbrace{\chi\chi \cdots \chi}_{I} t_i, ... \right\} = \mathcal{R},$$

wobei entweder keine Beobachtung vorliegt, was durch das leere Symbol ϵ illustriert wird, oder die Sequenz besteht aus einer Anzahl an $I \geq 0$ eingefügten zufälligen Symbolen $\chi \in \mathcal{A}$, gefolgt von einem Symbol, das entweder identisch zu t_i ist oder durch $\bar{t}_i \in \mathcal{A} \setminus \{t_i\}$ ersetzt wurde. Betrachtet man die Menge $\mathcal{R}$, so entspricht diese in der unbeschränkten Mächtigkeit der durch den Zustandsautomat in Abb. 3.2 beschriebenen Folgen.

Um eine approximative Beschreibung des Kanals zu erhalten, beschränkt man die maximale Anzahl an aufeinanderfolgenden Einfügungen auf eine feste Anzahl $I \in \mathbb{N}$. Dadurch wird es möglich eine geschlossene (endliche) Formulierung für die Wahrscheinlichkeit anzugeben, dass eine Sequenz $\mathbf{r}^{(i)}$ in einer Länge l realisiert wird. Eine auf I Einfügungen limitierte Längenverteilung kann angegeben werden als

$$p_l(L = l) = \alpha_l(l) \cdot p_\mathrm{d} + \alpha_l(l-1) \cdot p_\mathrm{t}, \tag{3.4}$$

wobei wie zuvor $p_\mathrm{t} = 1 - p_\mathrm{d} - p_\mathrm{i}$ gilt. Darin enthalten ist

$$\alpha_\ell(l) = \begin{cases} 1, & \text{für } l = 0, \\ {}^{p_\mathrm{i}^l}\!/_{(1-p_\mathrm{i}^I)}, & \text{für } 0 < l < I, \\ 0, & \text{sonst }, \end{cases} \tag{3.5}$$

die bedingte Wahrscheinlichkeit l Einfügungen zu beobachten, wenn die Sequenz $\mathbf{r}^{(i)}$ in einer endlichen Länge ℓ vorliegt. Dabei kann I als approximativer Parameter des gesamten Modells gesehen werden. Die Verteilung (3.5) wurde implizit in [41] verwendet. Bezugnehmend auf (3.4) kann eine Sequenz der Länge l letztlich durch l Einfügungen und der Löschung von t_i modelliert werden oder durch $l-1$ führende zufällige Symbole und einer Ergänzung von t_i oder $\bar{t}_i$ als letztes Symbol erklärt werden.

Die Übergangswahrscheinlichkeiten des HMMs können auf Basis der Längenverteilung direkt definiert werden zu

$$\Pr(S_i = s_i | S_{i-1} = s_{i-1}) = p_l(L = l),$$

wobei die Länge $l = s_i - s_{i-1}$ durch die Realisierung zweier aufeinanderfolgender Drift-Zustände determiniert ist. Bedingt durch die direkte Abhängigkeit der Beobachtungen von den Zustandsänderungen lässt sich eine strikt separierte Formulierung der Emmisionswahrscheinlichkeiten nicht kompakt darstellen. Geschlossen lässt sich jedoch die Verbundverteilung

$$\Pr\Big(\mathbf{R}^{(i)} = \mathbf{r}^{(i)}, S_i = s_i | S_{i-1} = s_{i-1}, t_i\Big)$$

formulieren. Sie ist die Wahrscheinlichkeit der Beobachtung von $\mathbf{r}^{(i)}$ und dem Modell in einem Drift-Zustand s_i, wenn zuvor der Drift s_{i-1} vorgelegen hat. Die Abhängigkeit vom Symbol t_i kann dabei als zusätzlicher Parameter $\mathbf{t}$ des HMMs verstanden werden (vgl. Abb. 3.3). Im Folgenden wird (wenn nicht benötigt) auf explizite Formalismen (Großbuchstaben) für Zufallsvariablen verzichtet. Verkürzt lässt sich der Wert der Wahrscheinlichkeit darstellen als Größe

$$Q(l, r_l^{(i)} | t_i) = \Big[\alpha_l(l) \cdot {}^1\!/_{q^l}\Big] \cdot p_\mathrm{d} + \Big[\alpha_l(l-1) \cdot {}^1\!/_{q^{l-1}}\Big] \cdot p_\mathrm{t} \cdot \Pr(r_l^{(i)} | t_i), \tag{3.6}$$

mit $0 \leq l \leq I + 1$ und $t_i, r_l^{(i)} \in \mathcal{A}$. Dabei entsprechen die geklammerten Ausdrücke der Wahrscheinlichkeit eine zufällige Einfügung der Länge l bzw. $l-1$ zu beobachten, die Gleichverteilung der Einfügungen $\chi \in \mathcal{A}$ vorausgesetzt. Für das Symbolalphabet der Nukleotide gilt $q = |\mathcal{A}| = 4$. Die bedingte Wahrscheinlichkeit $\Pr(r_l^{(i)} | t_i)$ beschreibt die Ersetzung von Symbolen t_i als q-wertigen Substitutionskanal (siehe Definition 12). Die durch Q beschriebene Wahrscheinlichkeit dient im Folgenden als Ausgangspunkt für Ansätze zur Decodierung und den Entwurf von Barcodes für den in Abschnitt 3.1.2 dargelegten Sequenzierkanal.

3.2 Decodierung und Codierung

Die Maximierung des Likelihoods eines Reads, bedingt durch das zuvor gezeigte Modell und der vollständigen Sequenz eines Templates, dient als Einstieg in den anschließenden methodischen Teil zur Anwendung des gezeigten HMMs. Darauf folgend wird die probabilistische Erkennung von eingebetteten Sequenzen erläutert, welche die Grundlage für die auf Watermark Codes basierenden Barcodes ist. Das Codierprinzip und die damit verbundene suboptimale symbolweise Decodierung der Barcodes bilden den Abschluss des Abschnitts 3.2.

3.2.1 Optimale Decodierung und probabilistische Mustersuche

Sei $\mathbf{r} \in \mathcal{A}^{L'}$ im Folgenden ein beispielhafter Read der Länge L' dessen Realisierung sich durch das gegebene HMM mit Parametern $\mathcal{H}$ beschreiben lässt, so resultiert der Ansatz

$$\hat{\mathbf{t}} = \arg\max_{\mathbf{t}\in\mathcal{A}^L} \Big\{ \Pr(\mathbf{r}|\mathbf{t}, \mathcal{H}) \Big\} \tag{3.7}$$

in der als ML-Entscheidung (vgl. Definition 22) bekannten Strategie, um das als wahrscheinlichst geltende Template $\hat{\mathbf{t}} \in \mathcal{A}^L$ zu ermitteln. Hierbei wird implizit angenommen, die Länge L der ursprünglichen Sequenz $\mathbf{t}$ sei bekannt. Auf Basis von Konzept 1, der Forward-Backward Prozedur, lässt sich das Argument der Maximierung durch die Vorwärts-Metrik

$$F_i(s) = \Pr(r_1 r_2 .. r_{i-1+s}, S_{i-1} = s|\mathbf{t}, \mathcal{H})$$

ermitteln. Sie entspricht der Wahrscheinlichkeit, dass der Drift-Zustand s erreicht wurde und die Sequenz bis zum Symbol r_{i-1+s} beobachtbar war, bevor das Symbol t_i berücksichtigt wurde. Für die vollständige Abbildung von $\mathbf{t}$ auf $\mathbf{r} = (\mathbf{r}^{(0)}\mathbf{r}^{(1)}\mathbf{r}^{(2)}...\mathbf{r}^{(L)})$ gilt

$$F_1(s) = \Pr(\mathbf{r}^{(0)} = \epsilon, S_0 = s) = \begin{cases} 1, & \text{für } s = 0, \\ 0, & \text{für } s \neq 0 \text{ und } s \in \mathcal{S}. \end{cases} \tag{3.8}$$

Dabei spielt das Hilfskonstrukt $\mathbf{r}^{(0)}$, als leeres Präfix symbolisiert durch ϵ, bei der Partitionierung von $\mathbf{r}$ keine Rolle, denn das Modell befindet sich vor dem Index $i = 1$ sicher im Zustand $s_0 = 0$. Für die explizite Berechnungsvorschrift, vgl. (3.6), gilt

$$F_{i+1}(s) = \sum_{l=0}^{I+1} F_i(s - \delta_l) \cdot Q(l, r_{i+s}|t_i) \tag{3.9}$$

mit $\delta_l = l - 1$ und den Nebenbedingungen $1 \leq i \leq L$ bzw. $1 \leq i + s \leq L'$. Die Bedingungen stellen sicher, dass das Symbol r_{i+s} Teil der Sequenz $\mathbf{r}$ ist und implizieren damit den maximal möglichen Zustandsraum $\mathcal{S}$. Letztlich ergibt sich $\Pr(\mathbf{r}|\mathbf{t}, \mathcal{H}) = F_{L+1}(L' - L)$ für das Argument der Maximierung (3.7). Die Differenz $L' - L$ beschreibt dabei den verbleibenden Drift nach der Zuordnung aller L Symbole t_i, d. h. wenn $\mathbf{t}$ komplett in $\mathbf{r}$ abgebildet wurde.

Betrachtet man, ungeachtet der Tatsache dass $\mathbf{t}$ in (3.1) nur teilweise bekannt ist, die Komplexität der Berechnungen des Likelihoods $F_{L+1}(L - L')$, so entspricht dieser der Anzahl an gültigen Partitionierungen von $\mathbf{t}$ bedingt durch die maximale Länge I einer Partition. Geht man von einer einfachen Implementierung aus, so ergeben sich für jedes t_i mitunter maximal

$i(I+1)+1$ nach (3.9) zu evaluierende Zustände, berechnet als Summe von $I+2$ Elementen. Konservativ betrachtet ergeben sich dadurch

$$\left[L/2(L+1)(I+1)+L\right](I+2) \tag{3.10}$$

Berechnungen für jedes $\mathbf{t}$. Der Ansatz alle möglichen Sequenzen der Länge L zu evaluieren ist jedoch mit exponentieller Komplexität verbunden und daher nur für sehr kurze Sequenzen realisierbar. Des Weiteren sind Bestandteile des Templates die keinen Barcode darstellen für das in Abschnitt 3.1.1 als Demultiplexing dargelegte Problem der Klassifikation von Reads unerheblich.

Ein weitaus einfacherer Ansatz, der durch die Beschreibung mittels HMM ermöglicht wird, ist die probabilistische Erkennung von eingebetteten Codeworten, welche als Motivation für das Konzept der Watermark Codes dienen soll.

Erkennung von eingebetteten Codeworten

Betrachtet man den in (3.1) dargestellten Verbund des Templates $\mathbf{t}$, so kann die Decodierung mit Bezug auf ein eingebettetes Codewort $\mathbf{c} \in \mathcal{C}$ der Länge $n \leq L$ wie folgt angepasst werden: Dazu adaptiert man die Vorwärts-Metrik zu

$$F_i(s) = \Pr(r_1 r_2 .. r_{i-1+s}, S_{i-1} = s | \mathbf{c}, \mathcal{H}) \tag{3.11}$$

und bedingt diese auf ein Codewort $\mathbf{c}$, mit Index i (Zeitparameter des HMMs) relativ zum Codewort, d. h. die Partitionierung von $\mathbf{r}$ wird in Abhängigkeit von $c_1, c_2, .., c_i, .., c_n$ beschrieben anstatt $\mathbf{t}$. Die neu definierte Größe entspricht somit der Wahrscheinlichkeit, dass der Drift-Zustand s erreicht wurde und die Sequenz bis zum Symbol r_{i-1+s} beobachtbar war, bevor das Codesymbol c_i berücksichtigt wurde.

Geht man von einer Abbildung von $\mathbf{c} \rightarrow \mathbf{r} = (\mathbf{r}^{(0)}\mathbf{r}^{(1)}\mathbf{r}^{(2)}...\mathbf{r}^{(L)})$ aus, so besteht wegen der Einbettung im Gegensatz zu (3.8) eine Unsicherheit über das Präfix $\mathbf{r}^{(0)}$ vor der ersten zu c_1 gehörigen Partition $\mathbf{r}^{(1)}$ des Reads. Ist über der Länge δ des Präfixes $\mathbf{p}_1 = t_1 t_2 t_3 ... t_\delta$ in (3.1) keine Information (A-priori-Verteilung) bekannt, so ist eine vereinfachende Annahme der Gleichverteilung

$$F_1(s) = \Pr(\mathbf{r}^{(0)}, S_0 = s) = \begin{cases} 1/L', & \text{für } 0 \leq s \leq L', \\ 0, & \text{für } s < 0 \text{ oder } L' < s \end{cases} \tag{3.12}$$

möglich. Ist eine Verteilung für die mögliche Position von Codeworten im Template bekannt, kann diese in (3.12) integriert werden. Vergleichbar mit dem semi-globalen Sequenzalignment (vgl. Definition 26) wird jedes Präfix beliebiger Länge durch (3.12) die gleiche Wahrscheinlichkeit zugeordnet. Analog zu (3.9) erfolgt die sukzessive Berechnung von $F_i(s)$ für alle i, s im Bereich $1 \leq i \leq L$ für $1 \leq i+s \leq L'$.

Zusätzlich zur Vorwärts-Metrik wird an dieser Stelle noch die Rückwärts-Metrik (vgl. Konzept 1) als Größe

$$B_i(s) = \Pr(r_{i+1+s} ... | S_i = s, \mathbf{c}, \mathcal{H}) \tag{3.13}$$

eingeführt. Sie entspricht der Wahrscheinlichkeit das Postfix des Reads beginnend mit r_{i+1+s} zu beobachten, bedingt durch den Aufenthalt im Zustand s, bei der Berücksichtigung des Codesymbols c_i. Ein explizite Berechnungsvorschrift für (3.13) ist

$$B_{i-1}(s) = \sum_{l=0}^{I+1} B_i(s+\delta_l) \cdot Q(l, r_{i+s+\delta_l}|c_i) \tag{3.14}$$

mit $\delta_l = l-1$ und Nebenbedingungen $1 \leq i \leq L$ bzw. $1 \leq i+s+\delta_l \leq L'$. Das Symbol $r_{i+s+\delta_l}$ ist somit immer Element von $\mathbf{r}$, was den gültigen Zustandsraum bestimmt. Ähnlich wie für das Präfix kann für den Fall der Einbettung im Allgemeinen keine Aussage über die Länge oder den Sequenzinhalt des Postfixes nach Berücksichtigung des Codesymbols c_n getroffen werden. Analog zu (3.12) ist die Annahme von $B_n(s) = 1$ als bedingte Wahrscheinlichkeit für $-n \leq s \leq L-n$ gerechtfertigt. Eine Berücksichtigung von A-priori-Informationen zur Sequenz des Postfixes ist hier ebenfalls möglich. Anhand der Gleichung (2.1) auf Seite 13 ermöglicht

$$\Pr(S_i = s|\mathbf{r}, \mathbf{c}, \mathcal{H}) = \frac{F_{i+1}(s)B_i(s)}{\Pr(\mathbf{r}|\mathbf{c}, \mathcal{H})} \tag{3.15}$$

die Schätzung des Drift-Zustandes zu

$$\hat{s}_i = \arg\max_{s \in \mathcal{S}} \{F_{i+1}(s)B_i(s)\}, \tag{3.16}$$

wobei der konstante Likelihood $\Pr(\mathbf{r}|\mathbf{c}, \mathcal{H})$ für die Maximierung von (3.15) unerheblich ist. Dabei stehen die Werte $\hat{s}_0$ und $\hat{s}_n$ stellvertretend für die geschätzte Position des Anfangs und des Endes der Repräsentation des Codewortes im beobachteten Read.

3.2.2 Barcodes als Watermark Codes

Zu Beginn des vorliegenden Abschnitts werden die grundlegende Idee hinter den Watermark Codes und die Rolle des namensgebenden *Wasserzeichens* motiviert. Im zweiten Teil wird das konkrete Übertragungsschema der Barcodes als strukturierte Codierung durch einfache Codeverkettung (vgl. Konzept 5) beschrieben.

Grundlegende Idee des Wasserzeichens

Ein einführendes Beispiel soll im Folgenden dazu dienen, die Idee des Wasserzeichens und die Implikationen für ein erweitertes Kanalmodell zu veranschaulichen: Der einfachste (nicht triviale) Fall der Nutzung von Barcodes ist die Beschränkung auf eine Menge $\mathcal{C} = \{\mathbf{c}^{(a)}, \mathbf{c}^{(b)}\}$ von zwei Codeworten der Länge n (vgl. Abb. 3.1). Ein standardmäßiges Vorgehen zur Wahl der Codeworte wäre im Bereich der Kanalcodierung die Maximierung der Korrekturfähigkeit (vgl. Definition 20) bezüglich des Hamming-Abstands und damit eine Maximierung von $\mathrm{d_h}(\mathbf{c}^{(a)}, \mathbf{c}^{(b)})$. Das Konzept der Watermark Codes zeigt jedoch ein anderes Vorgehen: Neben den eigentlichen Codeworten existiert das sogenannte Wasserzeichen $\mathbf{w}$, eine zufällige (nicht näher definierte) Sequenz der Länge n aus dem Alphabet des Codes. Das Wasserzeichen dient als Prämisse für die Auswahl der Codeworte: Denn die eigentlichen Codeworte werden mit minimalem Hamming-Abstand (ungleich Null) zur Sequenz $\mathbf{w}$ gewählt und erzeugen in diesem Beispiel einen Code mit minimaler Rate größer Null und minimal $n-2$ identischen Symbolen zu $\mathbf{w}$. Auf Basis dieser

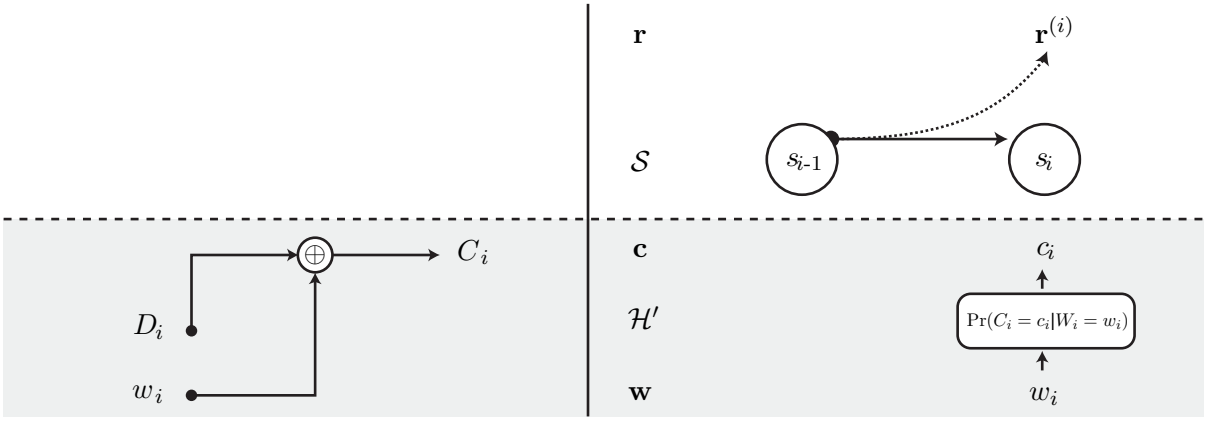

Abb. 3.4 HMM für Codeworte und Wasserzeichen: Komponentenweise Addition von einem festen Wasserzeichen $\mathbf{w}$ und eine dünnbesetzte Repräsentation des Information im Codewort als Zufallsvariable D_i zum probabilistischen Codesymbol C_i (links). Erweiterung des HMMs um eine Zusätzliche bedingte Wahrscheinlichkeit um die mittleren Unterschiede von Wasserzeichen und Codeworten zu berücksichtigen. Die dadurch angepasste Parametrisierung wird als $\mathcal{H}'$ bezeichnet.

fixen Symbole ist analog zu (3.15) die Berechnung von $\Pr(s_i|\mathbf{r}, \mathbf{w}, \mathcal{H})$ und damit die Schätzung der Position, sowohl für das Codewort $\mathbf{c}^{(a)}$, als auch $\mathbf{c}^{(b)}$ im Read möglich.

Die Parametrisierung $\mathcal{H}$ des ursprünglichen HMMs (3.3) ist jedoch aus probabilistischer Sicht nicht adäquat für die Diskrepanz von Wasserzeichen und Codeworten angepasst. Dazu folgende Perspektive: Nutzt man als Alphabet für die erwähnten Sequenzen $\mathcal{A} = \mathbb{Z}_q$, so lässt sich die Erzeugung der Codeworte als additiver q-wertiger Kanal (vgl. Definition 14) darstellten. Sei $\mathbf{c}^{(x)} \equiv \mathbf{d}^{(x)} \oplus \mathbf{w}$ (komponentenweise Addition modulo q) die Vorschrift zur Erzeugung der Codeworte von zuvor, so ist $\mathbf{d}^{(x)}$ für $x \in \{a, b\}$ jeweils eine *dünnbesetzte* Sequenz mit Hamming-Gewicht $\mathrm{w_h}(\mathbf{d}^{(x)}) = 1$. Die Auswahl welches x für die Codierung verwendet wird, ist zufällig von der Informationsquelle bestimmt. Ohne Einschränkungen gelte beispielsweise $d_a^{(a)} \neq 0$ und $d_b^{(b)} \neq 0$ mit $a \neq b$. Betrachtet man die zufällige Erzeugung eines Codewortes $\mathbf{C}$ (als Zufallsvariable), so entspricht das i-te Codesymbol C_i der Addition von Symbol w_i des Wasserzeichens und der Zufallsvariablen D_i (abhängig von der Auftrittswahrscheinlichkeit der dünnbesetzten Sequenzen $\mathbf{D}^{(x)}$). Schließlich besteht für die Positionen a und b prinzipiell die Möglichkeit, dass ohne zusätzlichen Kanal $w_a \neq c_a$ oder $w_b \neq c_b$ gilt (siehe Abb. 3.4, links). Der durch die Codierung beschriebene Einfluss lässt sich probabilistisch als zusätzliche Übergangswahrscheinlichkeit im HMM berücksichtigen (siehe Abb. 3.4, rechts). Die neue Parametrisierung des HMMs für die Verwendung bezüglich des Wasserzeichens wird als $\mathcal{H}'$ bezeichnet.

Ausgehend vom Beispiel mit zwei Codeworten existiert eine Vielzahl von Konfigurationen des gezeigten Prinzips, das im Folgenden kurz zusammengefasst wird: Für die Schätzung der Position δ von Codeworten als Einbettung (3.1) in eine längere Sequenz wird die Ähnlichkeit (in Hamming-Metrik) zu einem vorab definierten Wasserzeichen genutzt. Die übertragene Information wird als erweiterter Substitutionsfehler im Modell probabilistisch berücksichtigt (Parametrisierung $\mathcal{H}'$). Die eigentliche Decodierung der Codeworte geschieht im Anschluss daran basierend auf der geschätzten Position der Einbettung, der möglichen Codeworte selbst und einem HMM mit ursprünglicher Parametrisierung $\mathcal{H}$.

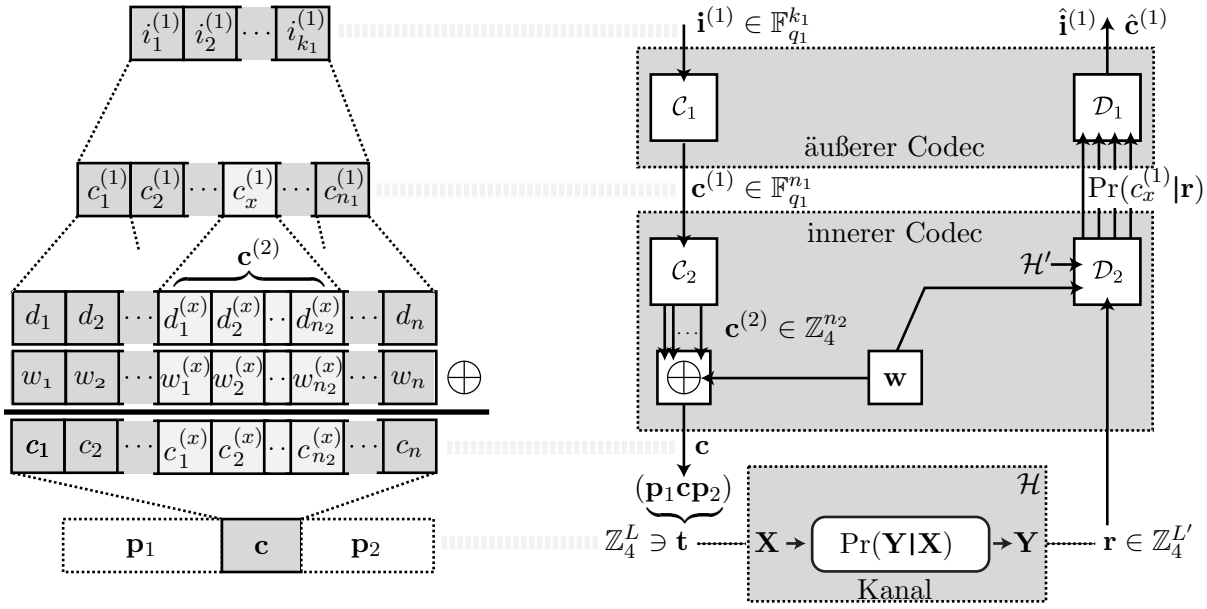

Abb. 3.5 Übertragungsschema für Barcodes als Watermark Codes (nach [91]): Blockbild der Codierung/Decodierung basierend auf einer einfachen Codeverkettung (Codecs) und dem Kanal für allgemeine Sequenzfehler (rechts). Repräsentation der Sequenzen der einzelnen Stufen der Übertragung auf Symbolebene (links).

Strukturierte Codierung durch Codeverkettung

Die strukturierte Konstruktion der Barcodes nutzt das Konzept der einfachen Codeverkettung (vgl. Konzept 5 auf Seite 19), das in ähnlicher Weise schon in [41] genutzt wurde. Das Blockbild in Abb. 3.5 zeigt das komplette Übertragungsschema für Barcodes. Prinzipiell kann man die Codierung und Decodierung in zwei Blöcke (hier Codecs genannt) untergliedern, welche im Sinne der Codeverkettung als äußerer und innerer Codec bezeichnet werden. Ausgangspunkt für die Codierung ist ein Informationswort $\mathbf{i}^{(1)}$ der Länge k_1 und ein äußerer Code mit Parametern $\mathcal{C}_1(\mathbb{F}_{q_1}, n_1, k_1, d_1)$. Die k_1 Informationssymbole des Alphabets der Größe q_1 (Galois-Feld, vgl. Definition 15) ergebene $q_1^{k_1}$ unterschiedliche Codeworte $\mathbf{c}^{(1)} \in \mathbb{F}_{q_1}^{n_1}$ der Länge n_1. Dieser Code sollte dabei eine maximale Korrekturfähigkeit bezüglich Symbolersetzungen bieten, d. h. die redundanten Symbole des äußeren Codes mit der Rate $R_1 = {}^{k_1}/_{n_1}$ sollten so strukturiert sein, dass der minimale Hamming-Abstand d_1 zwischen äußeren Codeworten maximiert ist.

Für den inneren Code gilt sozusagen das inverse Prinzip dazu: Jedes Codesymbol $c_x^{(1)} \in \mathbb{F}_{q_1}$ der äußeren Codeworte wird durch den Code $\mathcal{C}_2$ auf innere Codeworte $\mathbf{c}^{(2)} \in \mathbb{Z}_4^{n_2}$ abgebildet. Die inneren Codeworte entsprechen den dünnbesetzten Sequenzen $\mathbf{d}^{(x)}$ des letzten Abschnitts, realisiert in einer ganzzahligen Repräsentation $\mathbb{Z}_4$ des Alphabets der Nukleotide. Anders als im Absatz zuvor bilden hier n_1 Teilsequenzen der Länge n_2 eine strukturierte, dünnbesetzte Sequenz $\mathbf{d}$ der Länge $n = n_1 \cdot n_2$. Dabei ist n_2 mit $q_1 < 4^{n_2}$ der freie Parameter des inneren Codes. Für die Abbildung $\mathcal{C}_2 : \mathbb{F}_{q_1} \to \mathbb{Z}_4^{n_2},\ c_x^{(1)} \mapsto \mathbf{d}^{(x)}$ existieren für ein definiertes n_2 mehrere umkehrbare Zuordnungen $\mathcal{C}_2$. Die tatsächlich genutzten inneren Codes der Länge n_2 zeichnen sich dadurch aus, dass die hinzugefügte Redundanz durch eine maximal mögliche Anzahl von Null-Symbolen repräsentiert wird. Dabei soll die Null-Sequenz $\mathbf{0}$ (anders als im Ansatz von [41]) nicht Element des Codes sein. Somit gilt für alle diese atypischen inneren Codes eine minimale

Mindestdistanz von $d_2 = 1$. Die Rate des inneren Codes entspricht $R_2 = {}^{\log_2(q_1)}/_{(n_2 \log_2(4))}$, mit Bit als zugrundeliegende Informationseinheit.

Das letztendliche Barcodewort $\mathbf{c} = c_1 c_2 .. c_n$ resultiert schließlich aus der komponentenweisen Addition $\mathbf{c} \equiv (\mathbf{d} \oplus \mathbf{w}) \in \mathbb{Z}_4^n$ der abschnittsweise dünnbesetzten Codesequenz $\mathbf{d}$ mit der Sequenz $\mathbf{w}$ des Wasserzeichens. Das Wasserzeichen ist in dem dargelegten Schema der Codierung ein weiterer frei wählbarer Parameter, der hauptsächlich einen Einfluss auf die Verteilung der Symbole der Barcodes hat und später noch mit Bedeutung versehen wird. Für die Repräsentation der Codeworte aus $\mathbb{Z}_4$ als Sequenzen von Nukleotiden dient eine beliebige Eins-zu-eins-Abbildung $\mathbb{Z}_4 \mapsto \{\mathsf{A}, \mathsf{G}, \mathsf{C}, \mathsf{T}\}$. Betrachtet man die fertige Menge $\mathcal{C}$ an Barcodes als Verknüpfung $\mathcal{C} = \mathcal{C}_1 \circ \mathcal{C}_2$ (vgl. Konzept 5), so existieren insgesamt $q_1^{k_1}$ unterschiedliche Codeworte der Länge $n = n_1 \cdot n_2$ mit minimalem Hamming-Abstand $d \geq d_1 \cdot d_2$. Die Rate der Barcodes ergibt sich zu $R = ({}^{k_1}/_{n_2}) \cdot ({}^{\log_2(q_1)}/_{\log_2(4)})$.

3.2.3 Decodierung der Barcodes

Der Decodierer gliedert sich analog zur Codierung in zwei Blöcke (siehe Abb. 3.5): Ein innerer Decoder $\mathcal{D}_2$ nutzt das in 3.1.3 gegebene HMM in zwei Stufen. Unter Berücksichtigung der erweiterten Parametrisierung $\mathcal{H}'$ und der Verwendung des Wasserzeichens $\mathbf{w}$ wird in einem ersten Schritt die Verteilung der Drifts an den Blockgrenzen der inneren Codeworte eines Barcodes geschätzt. In einem zweiten Schritt erfolgt anhand dieser Verteilungen die Berechnung von symbolweisen Likelihoods $\Pr(\mathbf{r}|c_x^{(1)})$, welche normalisiert in Form einer A-posteriori-Verteilung $\Pr(c_x^{(1)}|\mathbf{r})$ im äußeren Decoder $\mathcal{D}_1$ zur Entscheidung für das wahrscheinlichst enthaltene Codewort $\hat{\mathbf{c}}^{(1)}$, respektive dem geschätzten Informationswort $\hat{\mathbf{i}}^{(1)}$, beitragen.

Parametrisierung $\mathcal{H}'$

Kernelement der Parametrisierung $\mathcal{H}'$ ist die probabilistische Beschreibung der inneren Codierung durch $\Pr(C_i = c_i|w_i)$, als Wahrscheinlichkeit ein Barcodesymbol c_i zu erhalten, wenn ein Symbol w_i an Position i des Wasserzeichens vorliegt (vgl. Abb. 3.4). Diese Wahrscheinlichkeit stellt kaskadiert mit der Übergangswahrscheinlichkeit $\Pr(Y = r_j|X = c_i)$ der Parametrisierung $\mathcal{H}$, vgl. (3.3), einen Meta-Kanal für die Beschreibung der effektiven Substitution von w_i durch das Empfangssymbol r_j dar. Geht man von einer Gleichverteilung der äußeren Codesymbole in $\mathbf{c}^{(1)}$ (vgl. Abb. 3.5) aus, so ist für jedes $\alpha \in \{1, 2, .., n_2\}$ der Wert von $\Pr(C_{\alpha+\beta \cdot n_2}|W_{\alpha+\beta \cdot n_2})$ für alle $\beta \in \{0, 1, .., n_1 - 1\}$ identisch und definiert durch die Auftrittshäufigkeit der Symbole $c_\alpha^{(2)}$ der inneren Codeworte $\mathbf{c}^{(2)} \in \mathcal{C}_2$. Ferner wird (analog zum Ansatz von Davey und MacKay [41]) vereinfachend davon ausgegangen, dass die Symbole d_i der dünnbesetzten Sequenz für alle $i \in \{1, 2, .., n\}$ identisch verteilt sind. Somit ergibt sich der reduzierte Ausdruck

$$\underline{\Pr(r_j|w_i)} = \sum_{c_i \in \mathbb{Z}_4} \Pr(r_j|c_i) \Pr(c_i|w_i) \tag{3.17}$$

für die (mittlere) effektive Substitutionwahrscheinlichkeit eines Symbols w_i. Daraus resultiert die veränderte Parametrisierung $\mathcal{H}'$ des HMMs und mit der Äquivalenz $r_l^{(i)} = r_j$ ergibt sich durch

$$Q'(l, r_l^{(i)}|w_i) = \left[\alpha_l(l) \cdot {}^1/_{q^l}\right] \cdot p_\mathrm{d} + \left[\alpha_l(l-1) \cdot {}^1/_{q^{l-1}}\right] \cdot p_\mathrm{t} \cdot \underline{\Pr(r_l^{(i)}|w_i)} \tag{3.18}$$

eine modifizierte Variante Q' von (3.6) bezogen auf das Wasserzeichen $\mathbf{w}$.

Innere Decodierung

Als erster Schritt der inneren Decodierung $\mathcal{D}_2$ (vgl. Abb. 3.5) erfolgt zunächst, analog zu (3.11) und (3.13), die Berechnung der Vorwärts-Metrik $F_i(s) = \Pr(r_1 r_2..r_{i-1+s}, S_{i-1} = s|\mathbf{w}, \mathcal{H}')$ und der Rückwärts-Metrik $B_i(s) = \Pr(r_{i+1+s}...|S_i = s, \mathbf{w}, \mathcal{H}')$, jedoch auf Grundlage der modifizierten Parameter aus dem letzten Abschnitt. Basierend auf den Größen F_i und B_i wird für jedes mögliche Symbol $c_x^{(1)} \in \mathbb{F}_{q_1}$ des äußeren Codes folgender bedingter Likelihood

$$\Pr(\mathbf{r}|c_x^{(1)}) = \sum_{(s_\alpha, s_\beta)\in\mathcal{S}^2} F_{\langle i_x}(s_\alpha) \cdot \pi_{\langle i_x}\Big\{s_\alpha, s_\beta|\mathbf{d}^{(x)} \oplus \mathbf{w}^{(x)}, \mathcal{H}\Big\} \cdot B_{i_x\rangle}(s_\beta) \tag{3.19}$$

berechnet. Dabei entspricht die Abbildung $\langle i_x = 1 + (x-1)\cdot n_2$ der Anfangsposition des Blocks x mit Bezug auf die absolute Position i im globalen Kontext des gesamten Codewortes und $i_x\rangle = x \cdot n_2$ liefert dessen letzte Position. Die Zuordnung der Sequenzen wird anhand Abb. 3.5 deutlich. Die Werte von s_α und s_β stehen für mögliche Drifts (Zustände des HMMs) und definieren über $\langle i_x + s_\alpha$ und $i_x\rangle + s_\beta$ hypothetische Bezugsgrenzen des äußeren Codesymbols $c_x^{(1)}$ im Kontext des Reads $\mathbf{r}$. Der Wert $\pi_{\langle i_x}$ entspricht dabei einer Fortsetzung der Vorwärts-Metrik ausgehend von der Position $\langle i_x$ (im Codewort) und dem Zustand s_α bis zur Position $i_x\rangle$ und dem Zustand s_β. Die Evaluation von π beruht dabei auf der Berechnungsvorschrift (3.11) mit Parametrisierung $\mathcal{H}$, respektive der Größe Q. Der Wert von π entspricht dabei der Wahrscheinlichkeit die Sequenz $\mathbf{r}$ innerhalb der genannten Bezugsgrenzen zu beobachten, bedingt durch den Aufenthalt in einem bestimmten Anfangszustand und der Sequenz $\mathbf{c}^{(x)} \equiv \mathbf{d}^{(x)} \oplus \mathbf{w}^{(x)}$, dem x-ten Abschnitt des Barcodewortes $\mathbf{c}$, abhängig von $c_x^{(1)}$.

Zur inneren Decodierung wurde in [41] eine Alternative mit reduziertem Aufwand vorgestellt, die in angepasster Weise im Folgenden dargelegt wird: Auf Grundlage der Schätzung der Drift-Zustände mittels (3.16) ist es möglich den Zustandsraum zu beschränken. Sei $\hat{s}_0$ der geschätzte Drift vor der Berücksichtigung des eingebetteten Barcodes, so definiert $\mathcal{S}^* = [\hat{s}_0 - \Delta, \hat{s}_0 + \Delta]$ einen Drift-Korridor der Größe $2\Delta + 1$ um $\hat{s}_0$. Von der Varianz des tatsächlichen Drifts und der Größe Δ ist es abhängig, wie groß letztlich die Wahrscheinlichkeit ist, dass eine Realisierung einer beobachteten Sequenz innerhalb von $\mathcal{S}^*$ liegt. Ist sichergestellt, dass dies der Fall ist, so ist die Abweichung der approximativen Berechnung von $\Pr(\mathbf{r}|c_x^{(1)})$ vernachlässigbar. In [41] wird die Varianz des Drifts pro Symbol definiert als $p_\mathrm{d}/(1-p_\mathrm{d})$ (für $p_\mathrm{i} = p_\mathrm{d}$) und eine Empfehlung für die Wahl der Größe Δ als vielfaches Produkt der Codewortlänge und der Standardabweichung des Drifts gegeben.

Äußere Decodierung

Prinzipiell wäre für eine optimale gesamte Decodierung im Sinne der Maximierung des Likelihoods (vgl. ML Definition 22) folgender Ansatz gegeben:

$$\hat{\mathbf{c}}^{(1)} = \arg\max_{\mathbf{c}^{(1)}\in\mathcal{C}_1} \Big\{\Pr(\mathbf{r}|\mathbf{c}^{(1)})\Big\}.$$

Eine äußere Decodierung würde durch dieses Konzept nicht erforderlich werden. Aus Gründen der Komplexität dieses optimalen Ansatzes (vgl. 3.2.1) wird die Struktur des äußeren Codes nicht für die innere Verarbeitung berücksichtigt. Unter Vernachlässigung der Abhängigkeiten der Symbole $c_x^{(1)}$ untereinander wird angenommen, dass für die A-posteriori-Wahrscheinlichkeiten

folgende Proportionalität gilt

$$\Pr(c_x^{(1)}|\mathbf{r}) \propto \Pr(\mathbf{r}|c_x^{(1)}) \cdot \frac{\Pr(c_x^{(1)})}{\Pr(\mathbf{r})},$$

wobei der rechtsseitige Bruch für alle $x \in \{1, 2, .., n_1\}$ eine Konstante sei. Die letztendliche äußere Decodierung $\mathcal{D}_1$ besteht aus der Maximierung

$$\hat{\mathbf{c}}^{(1)} = \arg \max_{\mathbf{c}^{(1)} \in \mathcal{C}_1} \Big\{ \prod_{x=1}^{n_1} \Pr(c_x^{(1)}|\mathbf{r}) \Big\}$$

der symbolweisen A-posteriori-Wahrscheinlichkeiten. Ein Decodierversagen, $\hat{\mathbf{c}} \notin \mathcal{C}_1$, ist durch strikte Evaluation aller möglichen Codeworte nicht möglich.

Komplexität der Decodierung

Die Komplexität der zweistufigen symbolweisen (suboptimalen) Decodierung lässt sich folgendermaßen vereinfachend abschätzen: Geht man für den ersten Schritt der inneren Codierung davon aus, dass für die Längen des Templates $\mathbf{t}$ und des Reads $\mathbf{r}$ näherungsweise $L \approx L'$ gilt, so lässt sich für die möglichen Tupel (i, s) der nicht trivialen Auswertung von $F_i(s)$ und $B_i(s)$ eine Ordnung von $\mathcal{O}(nLI)$ Operationen angeben. Dabei ist n die Länge des enthaltenen Barcodewortes und I entspricht der maximalen Anzahl an (modellierten) Einfügungen. Berücksichtigt man die weniger komplexe Alternative zur approximativen Berechnung der A-posteriori-Wahrscheinlichkeiten $\Pr(c_x^{(1)}|\mathbf{r})$, basierend auf einem reduzierten Zustandstraum der Größe Δ, so ergeben sich für die symbolweise Evaluation der Vorwärts-Metrik π die Ordnung von $\mathcal{O}(n_2 \Delta I)$ Berechnungen. Für alle Symbole der Sequenzen der Länge n_1 aus $\mathbb{F}_{q_1}$ ergeben sich somit Berechnungen in $\mathcal{O}(q_1 n \Delta I)$ für den zweiten Teil der inneren Decodierung. Abschließend sind für die äußere Decodierung $q_1^{k_1}$ Vergleiche von Produkten der Länge n_1 nötig. Der damit verbundene Aufwand ist in der Ordnung von $\mathcal{O}(n_1 q_1^{k_1})$.

3.3 Relevante Codes

Die forcierte Anwendung der ursprünglichen Watermark Codes ist die Übertragung eines quasikontinuierlichen Stroms von binären Symbolen, in Blöcken der Größe von $N = 2500 - 6000$ Bits [41] und dem Einsatz von LDPC-Codes [58] als äußere Codes in Längen von $100 - 888$ Symbolen. Das Wasserzeichen ist in diesem Zusammenhang ein zyklisches Bit-Muster der Länge N als Hintergrundsequenz im Strom der Bits. Die probabilistische Nutzung des Wasserzeichens beschränkt sich dabei auf die vereinzelte Resynchronisation eines Decodierfensters, wenn ein Synchronisationsfehler des Decoders erkannt wird. Die zuvor beschriebene Decodierung unterscheidet sich wesentlich von der ursprünglichen Konzeption. Gleiches gilt auch für den Aufbau der Codeverkettung.

Für den Einsatz des Codierschemas der in 3.1.1 gezeigten diskontinuierliche Übertragung von einzelnen endlichen Blöcken (als Sequenzierung von Nukleotiden) ergeben sich zusätzliche Eigenschaften, die für eine Anwendung von Watermark Codes als Barcodes berücksichtigt werden müssen. Zusätzlich zu der gezeigten Anpassung auf das vierwertige Kanal-Alphabet der DNA und die Einbettung von Codeworten ergeben sich noch andere Aspekte, die hier kurz aufgelistet und im Folgenden berücksichtigt werden:

- Die Sequenzierlänge der eingesetzten Technologie (vgl. Begriff 14) bildet eine faktische Obergrenze für die Länge L der Templates, die in einen Read abgebildet werden können. Da man neben der Markierung des Reads durch den eingebetteten Barcode hauptsächlich an der Sequenz der Inserts als Informationseinheit interessiert ist, sollte die Länge n des Barcodes möglichst kurz sein. Eine repräsentative Auswahl von Längen für Barcodes unterschiedlichster Konzepte sind beispielsweise 8nt [118], 10nt [48], 12nt [29, 30] oder 25nt [159], um nur einige zu nennen. Der speziellen Auswahl kurzer Sequenzen wird in Abschnitt 3.3.1 Rechnung getragen.
- Für den Einsatz der Barcodes zum Multiplexing (vgl. Begriff 19) ist neben der Länge und Anzahl der Codeworte (der Rate im eigentlichen Sinne) auch die Korrekturfähigkeit von Bedeutung. Für distanzbasierte Decodierverfahren (vgl. Definition 23) lässt sich über die minimale Distanz der Codeworte (in Hamming- oder Editier-Metrik) eine klare untere Schranke für die Korrekturfähigkeit angeben. Bedingt durch die Codekonstruktion ist es jedoch nur sehr eingeschränkt möglich, den minimalen Abstand zweier Codeworte eines Watermark Codes zu optimieren. Betrachtet man hingegen die Distanz-Verteilung (vgl. Definition 19) und mittlere Distanz als erweitertes Maß für Codes, so ergibt sich durch die Wahl des inneren Codes und des Wasserzeichens Optimierungspotential, welches in Abschnitt 3.3.2 umgesetzt wird. Das mittlere Hamming-Gewicht des inneren Codes und biotechnologisch begründete Bedingungen an die Sequenz der Barcodes bilden hierzu wichtige Nebenbedingungen. Ein Suchverfahren nach adäquaten Codes, die sich praktisch als Barcodes verwenden lassen, ist der Inhalt des letzten Teils von 3.3.2.

Den Abschluss dieses Abschnitts 3.3 bildet die Bewertung der speziellen Eigenschaften der gefundenen Barcodes in 3.3.3, sowie eine in silico Anwendung von beispielhaften Codes in einer Simulation der Sequenzierung von DNA in 3.3.4.

3.3.1 Codierschemata für Barcodes

Die Wahl eines Codierschemas für Watermark Codes beinhaltet zahlreiche Freiheitsgrade, die hier schrittweise spezifiziert und eingeschränkt werden, um einem sinnvollen Einsatz als Barcodes zu entsprechen. Eine Zielgröße ist die Codelänge der Barcodes, welche auf die Werte $n = n_1 n_2 \in \{12, 14, .., 21, 24, 25\}$ beschränkt wird, wobei für die Faktorisierung in Codelängen des äußeren und inneren Codes $n_1 \in \{3, 4, 5, 6\}$ und $n_2 \in \{4, 5, 6, 7, 8\}$ gelte. Die Wahl des äußeren Codes $\mathcal{C}_1$ ist für das in Abb. 3.5 gezeigte Übertragungsschema prinzipiell beliebig möglich, jedoch erscheint die Maximierung der Korrekturfähigkeit im Kontext des Hamming-Abstands (vgl. Definition 20) als sinnvoll. Für kurze Codes (wie gefordert) bietet die Sammlung der *best known linear codes* [65] eine Referenz zur Erzeugung von Codes mit den besten, momentan bekannten Distanzeigenschaften. Basierend auf dem Galois-Feld $\mathbb{F}_{q_1}$ der Kardinalitäten $q_1 \in \{2, 3, 4, 5, 6, 7, 8, 9\}$, der Länge n_1 und der Dimension k_1 sei $\mathcal{C}_1(\mathbb{F}_{q_1}, n_1, k_1, d_1)$ ein möglicher Code der maximal möglichen Mindestdistanz d_1. Für $\mathcal{C}_1(\mathbb{F}_5, 5, 2, 4)$ wäre dies beispielsweise ein *BCH-Code* [21], für $\mathcal{C}_1(\mathbb{F}_7, 5, 3, 3)$ eine Variante des *Hamming-Codes* [69]. Die im ursprünglichen Watermark Konzept genutzte Klasse der *LDPC-Codes* [106] ist, falls von der Mindestdistanz gegeben, auch eine Teilmenge der hier evaluierten äußeren Codes kurzer Länge.

3.3.2 Zusatzbedingungen für Barcodes und Suche nach Codes

Ein wesentlicher Punkt für die Anwendung von Barcodes ist die Anzahl verfügbarer Codes mit bestimmten Eigenschaften. Vergleicht man kommerziell verfügbare Ensembles von Barcodes, so erreichen deren Anzahl meist eine Größenordnung von 10^2 (z. B. 96 Barcodes des *NEBNext®* Produkt von New England Biolabs). Theoretische Ansätze schlagen meist erheblich umfangreichere Sätze von Codeworten in Ordnungen von 10^3 vor (z. B. 1544 Barcodes in [68], vgl. dazu [48]). Bei der Verwendung von größeren Ensembles von Barcodes ist eine Kosten-Nutzen-Abwägung relevant, weil eine individuelle Synthetisierung von Oligonukleotide einen erheblichen Aufwand darstellt, der durch eine spezielle Anforderung des Multiplexings gerechtfertigt sein sollte. Um den aktuellen Ansprüchen des Multiplexing gerecht zu werden, wird das Codierschema aus 3.3.1 um Codes reduziert für welche $7^2 = 49 \leq |\mathcal{C}_1| < 1000$ nicht erfüllt ist und die Anzahl $q_1^{k_1}$ nicht der geforderten Menge an Barcodes entspricht.

Betrachtet man zusätzlich den inneren Code als Abbildung $\mathcal{C}_2 : \mathbb{F}_{q_1} \mapsto \mathbb{Z}_{q_2}^{n_2}$ (mit $q_2 = 4$ für Barcodes), so gilt $q_1 < q_2^{n_2}$ für alle Schemata aus 3.3.1 und es existieren sogar mehrere umkehrbare Abbildung $\mathcal{C}_2$. Um die effektive Substitutionsrate in (3.17) durch den inneren Code zu minimieren, wird das mittlere Hamming-Gewicht

$$\overline{\mathrm{w_h}}(\mathcal{C}_2) = \frac{1}{q_1} \sum_{\mathbf{c} \in \mathcal{C}_2} \frac{\mathrm{d_h}(\mathbf{c}, \mathbf{0})}{n_2}$$

als Anteil der Symbole ungleich Null, berücksichtigt. Dabei ist n_2 die Länge und $q_1 = |\mathcal{C}_2|$ die Kardinalität des inneren Codes. Mit $\mathrm{d_h}$ ist die Hammingdistanz zur Nullsequenz $\mathbf{0}$ bezeichnet, vgl. (2.2). Im Folgenden ist

$$\mathcal{C}_2^* = \Big\{ \mathcal{C}_2 : \min_{\mathcal{C}_2} \{ \overline{\mathrm{w_h}}(\mathcal{C}_2) \} \Big\}, \text{ mit } \mathcal{C}_2 : \mathbb{F}_{q_1} \mapsto Z_{q_2}^{n_2} \setminus \mathbf{0}$$

eine Menge von Codes (Abbildungen), für welche das mittlere Hamming-Gewicht minimal ist. Dabei sei die Nullsequenz $\mathbf{0}$ kein Element aller möglichen Codes $\mathcal{C}_2^*$ (basierend auf den Analysen zum Fehlerschutz durch den inneren Code in [27]). Das maximal auftretende Hamming-Gewicht für Codes aus $\mathcal{C}_2^*$ ist hierbei

$$w_{\max} = \arg \min_{w \leq n_2} \Big\{ w : q_1 \leq N(w) = \sum_{i=1}^{w} \Pi(i) \Big\}, \tag{3.20}$$

mit $\Pi(i) = \binom{n_2}{i}(q_2 - 1)^i$. Hierbei ist $N(w)$ die maximale Anzahl der geforderten Codeworte mit Gewicht kleiner w. Ist die Nebenbedingung $q_1 \leq N(w)$ der Minimierung nicht mit Gleichheit erfüllt, so gilt

$$|\mathcal{C}_2^*| = \binom{\Pi(w_{\max})}{q_1 - N(w_{\max} - 1)}$$

für die Anzahl (bezüglich $\overline{\mathrm{w_h}}$) äquivalenter innerer Codes. Andernfalls existiert nur ein einziger Code $\mathcal{C}_2$, der diese Bedingung erfüllt. Für die Barcodes gelte zusätzlich, dass das mittlere Hamming-Gewicht auf jeden Fall kleiner als 0.3 ist und

$$\mathcal{C}_2^{(.3)} = \Big\{ \mathcal{C}_2 \in \mathcal{C}_2^* : \mathrm{w_h}(\mathcal{C}_2) < 0.3 \Big\}$$

bezeichnet die Menge möglicher Realisierungen. Die empirisch ermittelte Beschränkung des mittleren Hamming-Gewichts auf 0.3 bedeutet im übertragenen Sinne, dass ein Symbol des

Wasserzeichens w_i im Mittel mit Wahrscheinlichkeit größer 0.7 identisch als Codesymbol c_i im Barcode repräsentiert ist. Der Parameter $\overline{\mathrm{w_h}}(\mathcal{C}_2)$ bestimmt im Wesentlichen die Qualität des ersten Schritts der inneren Decodierung in 3.2.3 und beschränkt zusätzlich die maximale Rate der Barcodes.

Unter Berücksichtigung der Barcodelänge, der Anzahl an möglichen Codeworten und der Bedingung an, im Mittel, dünnbesetzten inneren Codes, verbleiben für die Codierschemata mit den Parametern q_1, k_1, n_1 und n_2 insgesamt 73 Konfigurationen. Zur Realisierung von konkreten Barcodes ist jedoch zusätzlich noch die explizite Auswahl eines möglichen Codes aus $\mathcal{C}_2^{(.3)}$ und die Wahl eines Wasserzeichens $\mathbf{w} \in \mathbb{Z}_4^n$ nötig, für welche im Folgenden weitere Kriterien gegeben werden, die einen Suchalgorithmus für Barcodes motivieren.

Das Wasserzeichen und die Editierdistanz

Für die auf dem HMM basierende innere Decodierung spielt der Abstandsbegriff für die Barcodes $\mathcal{C}$ eine doppelte Rolle: Für die Lokalisation (erster Decodierschritt) der Codeworte, respektive der effektiven Substitution von Symbolen, ist deren Abstand zum Wasserzeichen relevant. Die letztendliche probabilistische Bewertung und Unterscheidbarkeit der einzelnen beobachtbaren Sequenzen (zweiter Decodierschritt) macht einen großen Abstand der Codeworte zueinander unverzichtbar. Betrachtet man beide Perspektiven im Kontext des Hamming-Abstandes, so ist die Distanzverteilung (vgl. Definition 19) der Codeworte

$$D_{d,\mathbf{w}}^{(\mathrm{h})} = \frac{1}{|\mathcal{C}|}\Big|\Big\{(i,j) : \mathrm{d_h}(\mathbf{c}_i \oplus \mathbf{w}, \mathbf{c}_j \oplus \mathbf{w}) = d,\ \mathbf{c}_i, \mathbf{c}_j \in \mathcal{C}\Big\}\Big|$$

unabhängig von der Wahl des Wasserzeichens $\mathbf{w} \in \mathbb{Z}_4^n$. Gleiches gilt für die mittlere Distanz

$$\frac{1}{|\mathcal{C}|}\sum_{\mathbf{c}\in\mathcal{C}} \mathrm{d_h}(\mathbf{c} \oplus \mathbf{w}, \mathbf{w})$$

der Codeworte zum Wasserzeichen selbst. Sie ist lediglich von der Gewichtsverteilung des inneren Codes und der Auftrittshäufigkeit der äußeren Codesymbole abhängig. Zudem limitiert die Anforderung an das minimale Gewicht beim inneren Code eine Maximierung des Hamming-Abstands der Barcodeworte. Betrachtet man die Addition (komponentenweise, modulo 4) des Wasserzeichens als eine Art Verschiebung (Permutation) des Codes im Raum der Sequenzen, so ist der Hamming-Abstand invariant gegenüber dieser Art von Transformation. In Abb. 3.6 ist der Einfluss des Wasserzeichens sehr einfach abstrahiert.

Blickt man auf die Abstände von Barcodeworten zueinander, in Editier-Metrik, so ist die Distanzverteilung

$$D_{d,\mathbf{w}}^{(\mathrm{e})} = \frac{1}{|\mathcal{C}|}\Big|\Big\{(i,j) : \mathrm{d_e}(\mathbf{c}_i \oplus \mathbf{w}, \mathbf{c}_j \oplus \mathbf{w}) = d,\ \mathbf{c}_i, \mathbf{c}_j \in \mathcal{C}\Big\} \tag{3.21}$$

durch die Addition eine Wasserzeichens $\mathbf{w}$ veränderlich. Dieser Effekt soll im Folgenden kurz beschrieben werden, um das gegebene Vorgehen zu motivieren. Betrachtet man zwei unterschiedliche Sequenzen $\mathbf{d}_i$ und $\mathbf{d}_j$ des inneren Codierers (vgl. Abb. 3.5), so gilt der monotone Zusammenhang $\mathrm{d_e}(\mathbf{d}_i, \mathbf{d}_j) \leq \mathrm{d_e}(\mathbf{d}_i \oplus \mathbf{w}, \mathbf{d}_j \oplus \mathbf{w})$ nicht allgemein für $\mathbf{d}_i, \mathbf{d}_j, \mathbf{w} \in \mathbb{Z}_{q_2}^n$ ($q_2 = 4$ für Barcodes). Jedoch zeigen empirische Untersuchungen (hier nicht aufgeführt), dass für innere Codes mit minimalem Hamming-Gewicht im Mittel eine Vergrößerung der Editierdistanz möglich ist. Eine vergrößerte Kardinalität q_2 des Alphabets im inneren Code zeigt dabei einen

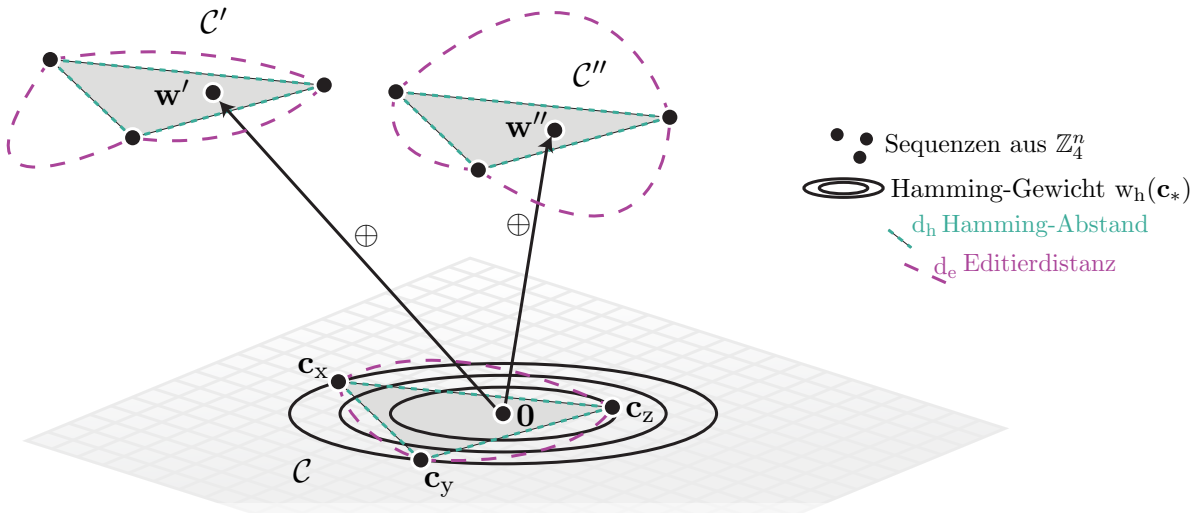

Abb. 3.6 Der Einfluss des Wasserzeichens auf Distanzen (schematisch): Die Distanzverteilung bezüglich des Hamming-Abstands bleibt für Barcodes $\mathcal{C}^* = \{\mathbf{c} \oplus \mathbf{w}^*, \mathbf{c} \in \mathcal{C}\}$ von der Addition eines Wasserzeichens $\mathbf{w}^* \in \mathbb{Z}_4^n$ unbeeinflusst. Für die Editierdistanz ist hingegen eine Variation der paarweisen Abstände der Codeworte durchaus möglich. Für die gezeigte schematische Illustration gilt jedoch stets $\mathrm{d_e}(\mathbf{c}_i, \mathbf{c}_j) \leq \mathrm{d_h}(\mathbf{c}_i, \mathbf{c}_j)$ für $\mathbf{c}_i, \mathbf{c}_j \in \mathcal{C}^*$ mit dem Hamming-Abstand als obere Schranke. Größere Distanzen sind in der Illustration durch längere Strecken der Kanten schematisiert.

positiven Einfluss auf den Zugewinn an mittlerer Editierdistanz eines Codes durch die Addition eines zufälligen Wasserzeichens. Unabhängig von der Wahl des Alphabets gilt jedoch für alle Barcodes und Wasserzeichen unabhängig voneinander $\mathrm{d_e}(\mathbf{c}_i \oplus \mathbf{w}, \mathbf{c}_j \oplus \mathbf{w}) \leq \mathrm{d_h}(\mathbf{c}_i, \mathbf{c}_j)$ (vgl. Eigenschaften zu Definition 25), was die Steigerung der paarweisen Editierdistanz begrenzt.

Neben der im nächsten Abschnitt erklärten Filterung von Barcodeworten soll die Maximierung der mittleren Editierdistanz

$$\overline{\mathrm{d_e}}(\mathcal{C}) = \frac{1}{|\mathcal{C}| - 1} \sum_{d>0} d \cdot D^{(\mathrm{e})}_{d,\mathbf{w}} \tag{3.22}$$

als Kriterium für die Auswahl eines Codes $\mathcal{C}$, als Kombination aus innerem Code und Wasserzeichen dienen. Dabei bildet sich das Mittel über $\sum_{d\neq 0} D^{(\mathrm{e})}_{d,\mathbf{w}} = |\mathcal{C}| - 1$ Codewortpaare, als Summe aller Häufigkeiten $D^{(\mathrm{e})}_{d,\mathbf{w}}$ mit $d \neq 0$, womit der Bezug von Codeworten auf sich selbst bei der Maximierung nicht berücksichtigt wird (siehe Eigenschaften für D_d in Definition 19). Anzumerken ist an dieser Stelle, dass die Qualität des Kriteriums für einen verbesserten Fehlerschutz der Codeworte von den konkreten Parametern $\mathcal{H}$ des Kanals und des HMMs abhängig ist. Die hier verwendete Editierdistanz stellt nur dann eine adäquate Metrik für Codeworte dar, wenn Substitutionen, Einfügungen und Löschungen mit gleicher Wahrscheinlichkeit auftreten. Speziell diese Parametrisierung wird Bestandteil der Evaluation in 3.3.4.

Biotechnologische Einschränkungen für die Sequenz der Barcodes

Für ein gegebenes Codierschema definiert die Auswahl eines Tupels $(\mathcal{C}_2, \mathbf{w})$ einen Code $\mathcal{C}$. Ob der komplette Code jedoch sinnvoll als Sequenzen für Barcodes verwendet werden kann, ist nicht sichergestellt. Denn es existieren de facto Standards bezüglich Sequenzstrukturen, welche ungeeignet für den biotechnologischen Einsatz in der Herstellung einer Sequenzier-Library sind

Alg. 3.1 Such-Algorithmus für Barcodes basierend auf Watermark Codes

Input : $\mathcal{C}_1, \mathcal{C}_2^{(.3)}, l_\mathrm{h}, \underline{r_\mathrm{GC}}, \overline{r_\mathrm{GC}}, N_\mathrm{iter}$
Output : $\mathcal{C}$

$(\mathcal{C}, \overline{\mathrm{d_e}}(\mathcal{C})) \leftarrow (\emptyset, 0)$
for 1 **to** N_iter **do**
  $(\mathcal{C}_2, \mathbf{w}) \leftarrow$ `rand.get`$(\mathcal{C}_2^{(.3)}, \mathbb{Z}_4^n)$

  $\mathcal{C}' \leftarrow \emptyset$
  foreach $\mathbf{c} \in \mathcal{C}_1 \circ \mathcal{C}_2$ **do**
    if `runLength`$(\mathbf{c} \oplus \mathbf{w}) \leq l_\mathrm{h}$ **then** $\mathcal{C}' \leftarrow \mathcal{C}' \cup \{\mathbf{c} \oplus \mathbf{w}\}$

  foreach $\pi \in \{\mathbb{Z}_4 \rightarrow \{A, G, C, T\}\}$ **do**

    $\mathcal{C}_\pi \leftarrow \emptyset$
    * **foreach** $(c_1 c_2 .. c_n) \in \mathcal{C}'$ **do**
      $\mathbf{c} \leftarrow [\pi(c_1)\pi(c_2)..\pi(c_n)]$

      if `gc`$(\mathbf{c}) < \underline{r_\mathrm{GC}}$ *or* `gc`$(\mathbf{c}) < \overline{r_\mathrm{GC}}$ **then** continue *↰
      if `selfComplement`$(\mathbf{c})$ **then** continue *↰

      $\mathcal{C}_\pi \leftarrow \mathcal{C}_\pi \cup \{\mathbf{c}\}$

  foreach $\pi \in \arg\max_\pi \{|\mathcal{C}_\pi| : |\mathcal{C}_\pi| \geq |\mathcal{C}|\}$ **do**
    if $\overline{\mathrm{d_e}}(\mathcal{C}_\pi) > \overline{\mathrm{d_e}}(\mathcal{C})$ **then** $(\mathcal{C}, \overline{\mathrm{d_e}}(\mathcal{C})) \leftarrow (\mathcal{C}_\pi, \overline{\mathrm{d_e}}(\mathcal{C}_\pi))$

oder zu einer erhöhten Fehlerwahrscheinlichkeit beim Vorgang der Sequenzierung führen. Vergleicht man Publikationen (wie beispielsweise [29, 30, 32, 48, 54, 68, 118, 153]), die sich mit der Konstruktion praxisrelevanter Barcodes befassen, so lassen sich grundsätzliche Leitsätze für Bedingungen zusammenfassen, die hier kurz beschrieben werden: Eine Sequenz sollte als Barcode nur dann eingesetzt werden, wenn der Anteil der darin enthaltenen Nukleotide Guanin (G) und Cytosin (C), auch *GC-Gehalt* genannt, zwischen 40%-60% liegt. Betrachtet man die frei wählbare Projektion $\pi : \mathbb{Z}_4 \mapsto \{\mathsf{A}, \mathsf{G}, \mathsf{C}, \mathsf{T}\}$ von numerischen Sequenzen auf Nukleotide, so kann der Anteil der Codeworte, die als Barcode verwendet werden können, durch eine Permutation möglicherweise vergrößert werden. Weiter müssen Codeworte ausgeschlossen werden, welche identisch zu ihrem Komplement sind oder Folgen von identischen Symbolen (*Homopolymere*) enthalten die länger als 2nt sind. Durch den Ausschluss von Sequenzen mit den beschriebenen Eigenschaften können experimentelle Probleme mit den gängigen Sequenziertechnologien verhindert werden. Die Filterung bezüglich inadäquater Codeworte führt natürlich zu einem Verlust in der Coderate.

Suche nach Barcodes

In Alg. 3.1 ist die randomisierte Suche nach Watermark Codes basierend auf den Zusatzbedingungen für Barcodes in Pseudoimplementierung dargestellt. Eingangsgrößen des Suchalgorithmus sind neben einem vorgegebenem äußeren Code $\mathcal{C}_1$ und der Menge $\mathcal{C}_2^{(.3)}$, an möglichen inneren Codes mit mittlerem Hamming-Gewicht kleiner 0.3, die maximale Länge $l_\mathrm{h} = 2$ für Homopolymere und der minimale (maximale) GC-Gehalt von $\underline{r_\mathrm{GC}} = 0.4$ (respektive $\overline{r_\mathrm{GC}} = 0.6$). In N_iter

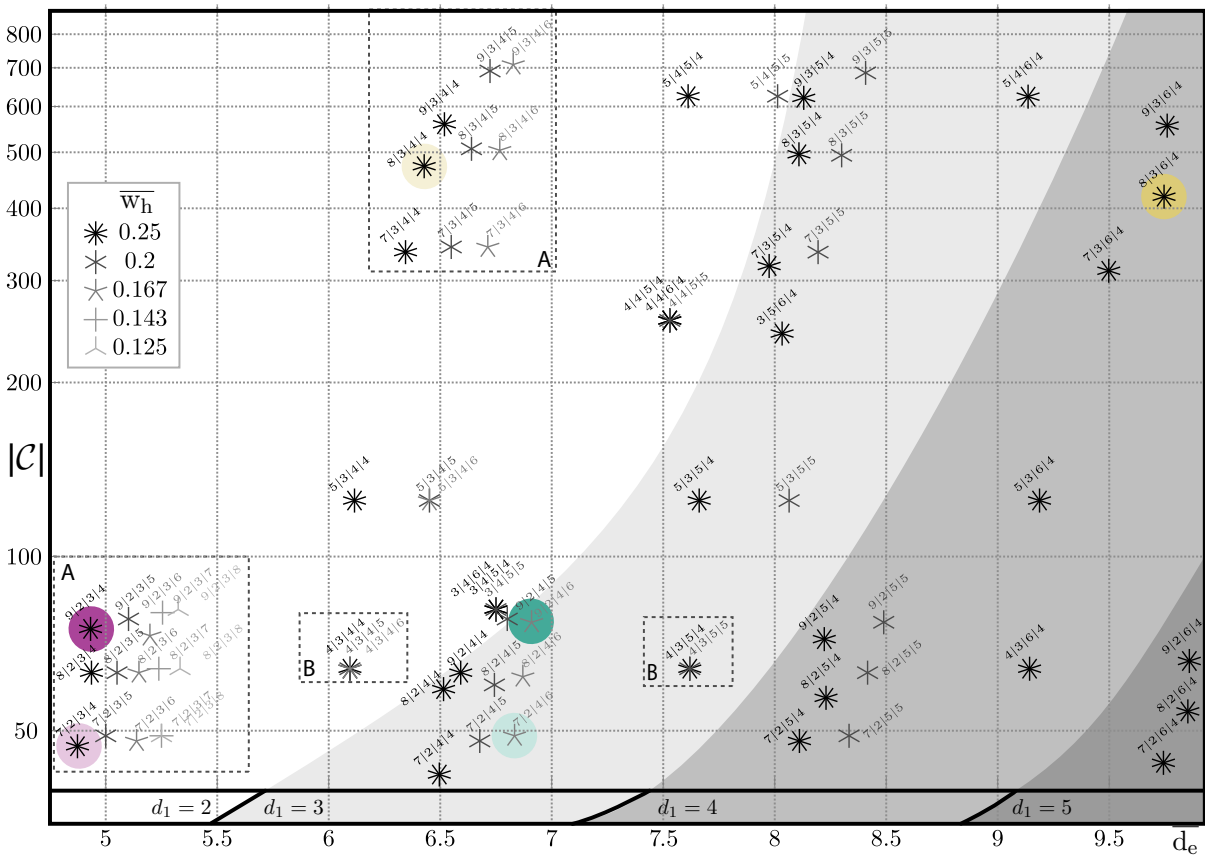

Abb. 3.7 Eigenschaften der untersuchten Barcodes basierend auf Watermark Codes (nach [91]): Markersymbole in Form von Sternen verknüpfen die mittleren Editierdistanz $\overline{d_e}$ mit der Anzahl $|\mathcal{C}|$ an nutzbaren Barcodeworten. Barcodes mit größerem mittlerem Hamming-Gewicht des inneren Codes sind durch Sternsymbole mit mehr Ecken gekennzeichnet. Die Annotation $q_1|k_1|n_1|n_2$ spezifiziert das Codierschema aus äußerem Code $\mathcal{C}_1(\mathbb{F}_{q_1}, n_1, k_1, d_1)$ und innerem Code $\mathcal{C}_2 : \mathbb{F}_{q_1} \mapsto \mathbb{Z}_4^{n_2}$.

Iterationen ermittelt die Suche eine Menge $\mathcal{C}$ an Barcodes mit maximaler Anzahl an Elementen (als erste Bedingung) und einer maximalen mittleren Editierdistanz (als zweite Nebenbedingung). Ausgehend von einer zufälligen Wahl $(\mathcal{C}_2, \mathbf{w})$ des inneren Codes und des Wasserzeichens erfolgt über `runLength` eine Filterung bezüglich der längsten in den Codeworten enthaltene Lauflänge. Für alle möglichen Projektionen π von numerischen Sequenzen auf Nukleotide erfolgt anschließend eine Maximierung hinsichtlich der Menge gültiger Barcodes, mit richtigem GC-Gehalt (durch das Funktional `gc`) und ohne mögliche Übereinstimmung mit dem eigenen Komplement (durch die Methode `selfComplement` als Indikator). Die letztendliche Ersetzung der favorisierten Menge $\mathcal{C}$ an Barcodes mit neuen Codes erfolgt auf Grundlage der mittleren Editierdistanz $\overline{d_e}(\mathcal{C})$.

Für jede der 73 Konfigurationen an äußeren Codes wurden $N_{\text{iter}} = 10^7$ zufällige, verkettete Codes evaluiert und nach den beschriebenen Bedingungen ausgewählt. Die Eigenschaften der gefundenen Codes (verbleibender Codeworte) werden im folgenden Abschnitt näher betrachtet.

3.3.3 Eigenschaften der gefundenen Barcodes

In Abb. 3.7 sind die resultierenden Barcodes und deren Eigenschaften in kompakter Form miteinander Verknüpft: Neben der mittleren Editierdistanz $\overline{\mathrm{d_e}}$ (als x-Achse) ist die Kardinalität $|\mathcal{C}|$ der Mengen an Codeworten (als y-Achse) aufgetragen. Das mittlere Hamming-Gewicht $\overline{\mathrm{w_h}}$ des enthaltenen inneren Codes ist als dritte Charakteristik der Barcodes durch die Auswahl unterschiedlicher Markersymbole illustriert. Für die 73 ausgewählten Codierschemata existieren fünf unterschiedliche Niveaus für $\overline{\mathrm{w_h}} \in \{0.125, 0.143, 0.167, 0.2, 0.25\}$, welche durch Sternsymbole mit 3-10 Strahlen dargestellt werden (ansteigend mit dem Gewicht der Codeworte). Zusätzlich zu der Symbolik ist für jeden Code eine knappe textuelle Notation $q_1|k_1|n_1|n_2$ eingefügt, welche die Parameter des äußeren Codes $\mathcal{C}_1(\mathbb{F}_{q_1}, n_1, k_1, d_1)$ enthält und die Abbildung des inneren Codes $\mathcal{C}_2 : \mathbb{F}_{q_1} \mapsto \mathbb{Z}_4^{n_2}$ spezifiziert. Wie zuvor definiert ist $q_1^{k_1}$ die maximale Anzahl durch die Codierung möglicher Codeworte und $|\mathcal{C}| \leq q_1^{k_1}$ die Anzahl (nach der Filterung) übrigen Barcodes. Die Länge der Barcodes ist $n = n_1 \cdot n_2$ und das mittlere Gewicht $\overline{\mathrm{w_h}}$ lässt sich über die Größen q_1, n_2 und den Anmerkungen zu (3.20) ableiten. Möchte man im Rahmen der Barcodes, als spezielle Form einer Informationsübertragung, den Begriff der Rate verwenden, so lässt sich

$$R = \frac{\log_2(|\mathcal{C}|)}{\log_2(4^n)} = \frac{1}{2n} \log_2(|\mathcal{C}|) \tag{3.23}$$

als Übertragungrate von Information (bezogen auf die Informationseinheit Bit) angeben. In Abb. 3.7 sind zusätzlich Ebenen eingefügt (grau schattiert), welche den minimalen Hamming-Abstand d_1 des äußeren Codes kennzeichnen: Für die untersuchten Klassen von Codes ergeben sich insgesamt vier verschiedene Stufen $d_1 \in \{2, 3, 4, 5\}$ und damit eine maximale Fehlerkorrektur von $t = 2$ Symbolen für die äußeren Codes in $q_1|2|6|4$ mit $q_1 \in \{7, 8, 9\}$. Des Weiteren enthält die Illustration farbige Markierungen, die beispielhaft ausgewählte Codes kennzeichnen, welche hier detailliert diskutiert werden. Alle 73 Barcodes sind als explizite Listen von Codeworten inklusive der konkreten Codierschemata (mit Wasserzeichen) im Anhang der Veröffentlichung [91] enthalten.

In der aufgeführten Darstellung lassen sich Gruppen gleicher äußerer Codes ausmachen (A in Abb. 3.7), welche für den Einfluss des inneren Codes klare Tendenzen zeigen, so z. B. für die Schemata $q_1|2|3|n_2$ und $q_1|3|4|n_2$ mit $n_2 \in \{7, 8, 9\}$. Eine Steigerung der Codewortlänge n_2 des inneren Codes stellt neben einer Verringerung der Rate (proportional zu $1/n_2$) eine Reduktion des mittleren Hamming-Gewichts dar. Zusätzlich dazu ermöglicht die vergrößerte Anzahl an Nullen in den inneren Codeworten in Verbindung mit dem Wasserzeichen eine Steigerung der mittleren Editierdistanz der Barcodes zueinander. Für die markierten Mengen von Codes ist im Allgemeinen ein gewisser Austausch von Coderate und Distanz zu erkennen, die durch die Maximierung von (3.22) bei der Suche induziert ist.

Unabhängig davon existieren jedoch einige Konstellationen von Codes für welche dieser Austausch für keine der evaluierten Realisierungen aus 3.3.2 sichtbar ist. Für das beispielhafte Schema $4|3|4|n_2$ mit Werten $n_2 \in \{4, 5, 6\}$ oder $4|3|5|n_2$ mit $n_2 \in \{4, 5\}$ (siehe B in Abb. 3.7) konnte (in der Suche) keine der 10^7 zufälligen Kombination aus innerem Code und Wasserzeichen die Distanz-Verteilung weiter verbessern. Dies ist insofern besonders, da sich die Länge der letztendlichen Barcodes um bis zu 8 Symbole unterscheiden. Betrachtet man alle möglichen Konstellationen bei denen die Länge n_2 keinen Einfluss auf die Lage der Codes innerhalb der $\overline{\mathrm{d_e}}$-$|\mathcal{C}|$-Ebene in Abb. 3.7 haben, so gilt grundsätzlich $q_1 \leq n_2$. Diese Bedingung ermöglicht,

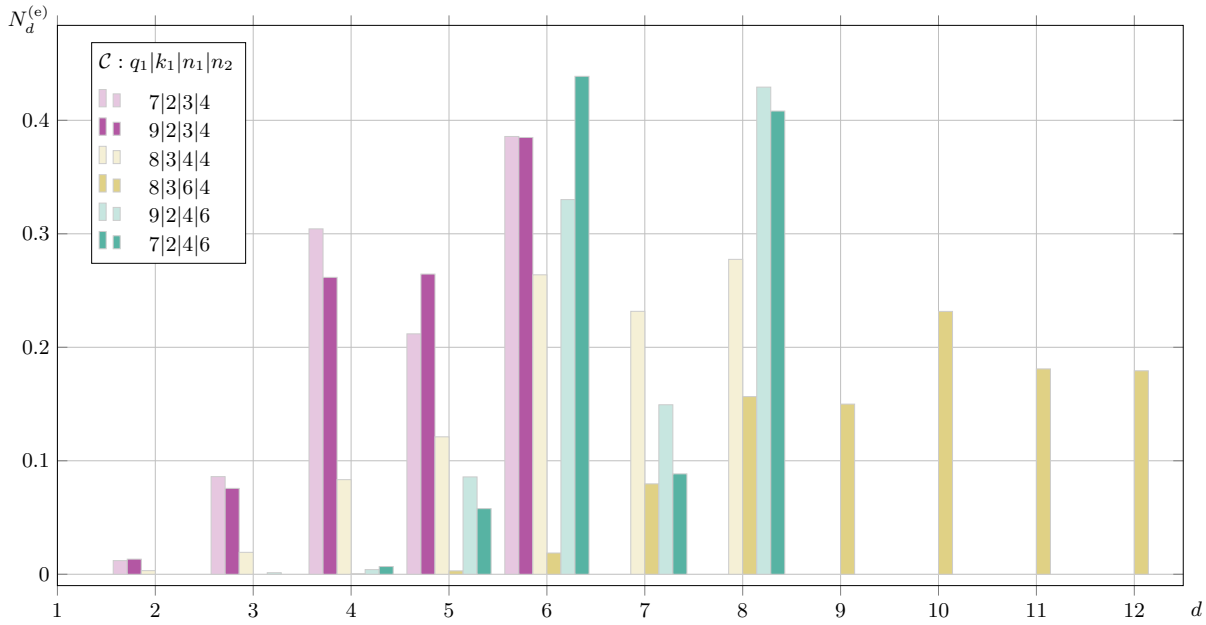

Abb. 3.8 Exemplarische Distanz-Verteilung für Barcodes auf Basis der Watermark Codes: Das Farbschema aus Abb. 3.7 identifiziert ausgewählte Codes. Für alle dargestellten Codes gilt $d \geq 2$.

dass jedes äußere Codesymbol auf eine exklusive Position im inneren Codewort abgebildet wird. Diese Konstellation bestimmt offenbar die Optimierung der Distanzeigenschaften der Barcodes.

Trotz der formulierten Suche nach Codeworten mit möglichst großer mittlerer Editierdistanz ist ein konkreter Blick auf die Verteilung der paarweisen Distanzen interessant. Um Codes unterschiedlicher Dimension (hier faktisch Kardinalität der Mengen) zu vergleichen, ist eine Normalisierung der Distanzverteilung sinnvoll. Im Folgenden ist

$$N_d^{(e)} = \frac{D_d^{(e)}}{|\mathcal{C}| - 1}, \text{ für } d \neq 0,$$

der relative Anteil von Paaren eines Codes, mit einer bestimmten Editierdistanz d. Die Berechnung des Anteils bezieht sich dabei auf die Ausdrücke (3.21) und (3.22) ohne Parametrisierung durch das Wasserzeichen. Im übertragenen Sinn der Definition 19 (der Distanzverteilung eines Codes) bestimmt $N_d^{(e)}$ für ein zufällig ausgewähltes Codewort bei der zufälligen (gleichverteilten) Auswahl eines weiteren verschiedenen Codewortes die Wahrscheinlichkeit eines mit Distanz d zu erhalten.

In Abb. 3.8 ist eine exemplarische Auswahl von Distanzverteilungen dargestellt: Zwei der kürzesten illustrierten Codes der Länge $n = 12$ (Parameter $q_1|2|3|4$ mit $q_1 \in \{7, 9\}$, violett in Abb. 3.8) sind neben Codes mit relativ hohen Raten ($8|3|n_1|4$ für $q_1 \in \{4, 6\}$, gelb in Abb. 3.8) und Längen von 16 und 24 Symbolen dargestellt. Zusätzlich ist die Distanzverteilung der Codeworte mit Parametern 7|2|4|6 und 9|2|4|6 (Abb. 3.8, petrol-blau) und moderater Rate gezeigt, welche sich im Vorgriff auf Abschnitt 3.3.4 als Codes mit der besten Fehlerkorrektur präsentierten. Bedingt durch die Konstruktion der Barcodes mit minimalem mittlerem Hamming-Abstand zu einem bestimmten Wasserzeichen und der Beschränkung $\mathrm{d_e} \leq \mathrm{d_h}$ (vgl. Definition 25), ist es nicht verwunderlich Paare von Barcodes mit relativ geringer Editierdistanz zueinander zu beobachten.

Nichtsdestotrotz sind für jeden gezeigten Code mindestens $d = 2$ Sequenzmodifikationen nötig, um zwei beliebige Codeworte ineinander zu überführen. Für Codes der Längen $n > 12$ ist der erwartete Abstand für 99% der zufälligen Codewortpaare $d \geq 3$. Für Codes mit $n \geq 20$ gilt mit 99% Wahrscheinlichkeit sogar $d \geq 5$. Geht man davon aus, dass die Blockgrenzen der Barcodes durch das darin enthaltene Wasserzeichen identifizierbar sind, so gilt die Korrekturfähigkeit $t = \lfloor {}^{(d-1)}/_{2} \rfloor$ (vgl. Definition 20 und [99]) auch in der Metrik der Editierdistanz und für die in dieser Arbeit gezeigte Einbettung von Barcodes.

Vergleicht man die hier evaluierten Codes mit den in [29] erzeugten Codes auf Basis der Sequenz-Levenshtein Distanz, so zeigen sich hinsichtlich der (quasi) garantierten Korrekturfähigkeit und der gegebenen Rate sicherlich drastische Unterschiede: Für die sogenannten Sequenz-Levenshtein Codes existieren mindestens 612 unterschiedliche Codeworte der Länge $n = 9$ und Mindestdistanz $d = 3$, respektive mehr als 554 unterschiedliche Sequenzen der Länge $n = 13$ und Distanz $d = 5$. Um 473 Barcodes auf der Basis von Watermark Codes zu erzeugen, die mit hoher Wahrscheinlichkeit eine Editierdistanz $d \geq 3$ aufweisen, sind $n = 16$ Symbole nötigt, 420 Codeworte mit (effektiver) Minimaldistanz $d \geq 5$ basieren auf einer Länge von $n = 24$. Betrachtet man die Coderate, so ist die Integration des Wasserzeichens in den evaluierten Codes (vorbehaltlich der Annahme vergleichbarer Mindestdistanzen) mit einer Verdopplung der Codelänge und damit einem Faktor $1/2$ hinsichtlich der Rate verbunden. Der spezifische Vergleich der Distanzeigenschaften der unterschiedlichen Klassen von Codes könnte auf Basis der Distanzverteilung geführt werden, welche für die Sequenz-Levenshtein Codes aus [29] jedoch nicht publiziert sind.

Letztlich kann an dieser Stelle nur die Struktur der verschiedenen Codes hinsichtlich unterschiedlicher Metriken verglichen werden. Würde die Decodierung von Watermark Codes auch auf Basis der Distanzminimierung (vgl. Definition 23) erfolgen, so wäre die Distanzverteilung der Codes sicherlich ein aussagekräftiges Maß. Ob sich darauf ein fairer Vergleich der konzeptionell andersartigen (mehrstufigen) Decodierung der Watermark Codes aufbauen lässt, ist schwer zu beurteilen. Um einiges einfacher scheint hierzu die konkrete Anwendung der Barcodes und die Schätzung von Decodierfehler, beispielsweise in einer Simulation.

3.3.4 In silico Anwendung

Im Bereich der Entwicklung neuer Barcodes ist es üblich vor der aufwändigen und kostenintensiven realen Sequenzierung eine *in silico* (lateinisch für *in Silizium*, also auf dem Chip platzierte) Anwendung in Form von Simulationen durchzuführen. Ähnlich wie in der Nachrichtentechnik wird hierzu auf Basis eines vorab definierten Kanalmodells ein Kanalsimulator implementiert, der den experimentellen Einfluss auf die Sequenzen von Nukleotiden imitiert und das Auftreten von Fehlern nachbildet. In Anlehnung an die Beschreibung des Kanalsimulators in [29] wird im Folgenden die Implementierung des hier genutzten Kanals näher erläutert, bevor die Parametrisierung und Auswertung der Decodierung beschrieben wird.

Kanalsimulator

Grundlage für den Kanalsimulator ist das in 3.1.1 beschriebene Konzept zur Einbettung von Codeworten $\mathbf{c} \in \mathcal{C}$ in

$$\mathbf{t} = (\mathbf{p}_1, \underbrace{c_1 c_2 ... c_n}_{\mathbf{c}}, \mathbf{p}_2) \in \mathcal{A}^L, \text{ mit } \mathcal{A} = \{\mathsf{A}, \mathsf{G}, \mathsf{C}, \mathsf{T}\},$$

als das zu sequenzierende Template und dessen Abbildung (vgl. 3.1.2)

$$\mathcal{A}^L \to \mathcal{A}^{L'}, \ \mathbf{t} \mapsto \mathbf{r}$$

auf einen Read $\mathbf{r}$ (empfange Sequenz), wobei $L \neq L'$ möglich ist. Zur Realisierung der Einbettung wird eine fixe Ganzzahl $l_\mathbf{p}$ gewählt, welche die Länge der Präfixe und Postfixe bestimmt. Die Sequenzen $\mathbf{p}_1, \mathbf{p}_2 \in \mathcal{A}^{l_\mathbf{p}}$ werden dann gleichverteilt aus der Menge der Nukleotid-Sequenzen der Länge $l_\mathbf{p}$ erzeugt. Die Abbildung $\mathbf{t} \mapsto \mathbf{r}$ wird darauf folgend durch einen probabilistischen Zustandsautomat erzeugt, der im Unterschied zum Kanalmodell in 3.1.2 auf vier unterschiedlichen Ereignissen basiert, die Positionen i und Symbole t_i als Zustände hält und die Sequenz $\mathbf{r}$ als Ausgabe erzeugt. Die vier möglichen Ereignisse sind hierbei die Folgenden: Symbolisiert durch `C` ist die korrekte Ausgabe (das Durchreichen) von t_i, das Symbol `S` kennzeichnet eine Ersetzung (Ausgabe von $\bar{t}_i \in \mathcal{A} \setminus t_i$) und mit `I` und `D` ist die Einfügung respektive Löschung eines Symbols annotiert. Anders als in der Beschreibung des Modells in Abschnitt 3.1.2 erfolgt die Implementierung ohne explizite bedingte Wahrscheinlichkeit für eine Ersetzung und unter der Annahme, dass jedes Symbol in einer Ersetzung (oder Einfügung) mit gleicher Wahrscheinlichkeit auftritt. Die Wahrscheinlichkeit für `C` oder `S` entspricht somit exakt $p_t = 1 - p_\mathrm{i} - p_\mathrm{d}$ in der Repräsentation des ursprünglichen Modells. Zudem wird ein einziger Parameter p_mod eingeführt, der den Fehlereinfluss des Kanalsimulators quantifiziert. Die Wahrscheinlichkeit für `C` entspricht $1 - p_\mathrm{mod}$ und die Wahrscheinlichkeit für eines der Ereignisses `S`, `D` oder `I` ist p_mod, wobei jedes Ereignis für sich gleich wahrscheinlich ist.

Anzumerken ist, dass eine äquivalente Beschreibung des probabilistischen Modells sowohl durch eine Wahrscheinlichkeit p_mod oder durch die zuvor spezifizierten Kanalparameter $\mathcal{H}$ mit $p_\mathrm{i}, p_\mathrm{d}$ und $\Pr(Y|X)$ möglich ist. Die Parametrisierung des HMMs zur Decodierung erfolgt auf Basis der bereits bekannten Repräsentation $\mathcal{H}$. Der uniforme Fehlerparameter p_mod beschreibt jedoch eine gleiche Verteilung aller möglichen Fehler im Sinne der Editierdistanz, indem jede symbolweise Operation mit gleicher Wahrscheinlichkeit auftritt. Der gezeigte Kanalsimulator ist damit ein sehr spezieller, aber geschlossen beschreibbarer Fall, der großen Menge möglicher Kanalrepräsentationen.

Schätzung des Decodierfehlers

Zur Schätzung des Decodierfehlers der erzeugten Barcodes wurde wie folgt verfahren: Für alle 73 Gruppen von Codes wurden Fehlerkurven erzeugt, die aus der mittleren Anzahl von fehlerhaft decodierten Barcodes pro Evaluationsschritt abgeleitet wurden. Der als P_e bezeichnete Mittelwert stellt eine Schätzung für die Fehlerwahrscheinlichkeit dar. Für jeden Evaluationsschritt wurde ein zufälliges Codewort aus der Menge eines Codes (gleichverteilt) ausgewählt, durch den Kanalsimulator in jeweils $l_\mathbf{p} = 50$ zufällige Symbole eingebettet, modifiziert und durch das in 3.2.3 beschriebene Verfahren decodiert. Ein Decodierfehler stellt hierbei eine Ungleichheit zwischen Eingangscodewort und Decodierergebnis dar. Zur Bewertung der Fehlerkorrektur wurde der Kanalparameter im Bereich $p_\mathrm{mod} \in \{0, 0.002, 0.005, .., 0.15, 0.16\}$ variiert.

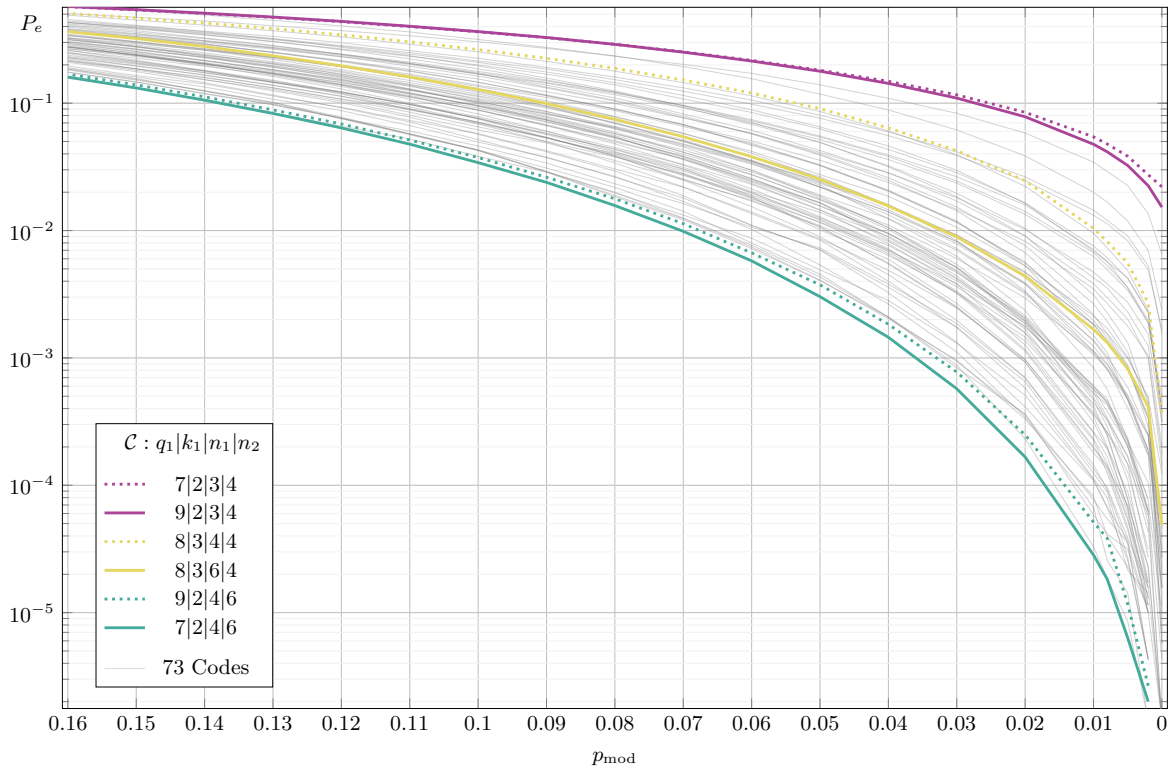

Abb. 3.9 Decodierfehler für unterschiedliche Kanalparameter der 73 Codes. Das Farbschema aus Abb. 3.7 kennzeichnet wiederum die beispielhaft ausgewählten und hervorgehobenen Codes.

Die vorliegende Simulation entspricht im Wesentlichen der in [29] gegebenen Parametrisierung für die Kanalfehlerwahrscheinlichkeit. Die kleinste gewählte Fehlerwahrscheinlichkeit der Simulation in [29] wurde jedoch mit 0.1 beziffert. Der hier angegebene Bereich für p_{mod} umfasst praxisrelevante Größenordnungen (vgl. beispielsweise Angaben in [105]), zusätzlich dazu wurde der fehlerfreie Fall mit in die Evaluation aufgenommen. Die Schätzung der Fehlerwahrscheinlichkeit P_e einer Menge von Barcodes bei einem bestimmten Wert p_{mod} basiert, abhängig von der Kanalqualität, auf bis mehr als 10^6 evaluierten Codeworten pro Datenpunkt.

Die in Abb. 3.9 hervorgehobenen Fehlerraten P_e beschränken sich auf die bereits in Abb. 3.7 farbig markierten Codes, um einen Überblick über die mögliche Fehlerkorrektur der vorgestellten Barcodes zu geben. Der Bereich aller evaluierten Fehlerkurven der 73 Codes (in der Darstellung ausgegraut) wird dabei begrenzt durch die hier in violett und petrol-blau dargestellten Kurven. Die in gelb gezeichneten Fehlerraten sind beispielhaft für Codes höherer Rate hervorgehoben, welche eine höhere Zahl an Codeworten ermöglichen. Im Folgenden sollen gegebene Beobachtungen anhand der farbig gekennzeichneten Kurven kurz ausgeführt werden. Dazu sind zwei spezielle Ereignisse interessant, die unterschiedliche Realisierungen der Sequenzen am Ausgang des Kanalsimulators bzw. Decodierergebnisse zusammenfassen: Zum einen das Auftreten eines Decodierfehlers, symbolisiert als Ereignis `E`, und die Beobachtung eines fehlerhaften Barcodes am Ausgang des Kanalsimulators, als Ereignis `F`. Betrachtet man einen Barcode der Länge n, so gilt $\Pr(\bar{\mathtt{F}}) = (1 - p_{\text{mod}})^n$ für den Fall das Codewort in unveränderter Form zu decodieren. Wäre das Codewort keine Subsequenz, in der Bedeutung der Einbettung, so ist für jeden sinnvollen Decodierer zu erwarten, dass die Wahrscheinlichkeit einer fehlerhaften Ent-

scheidung $\Pr(\mathtt{E}|\bar{\mathtt{F}}) = 0$ ist. Solange die Abbildung des verwendeten Codes eindeutig umkehrbar ist und keine weiteren Symbole an der Decodierentscheidung beteiligt sind, ist die Decodierung trivial. Blickt man auf die gezeigten Fehlerraten für $p_{\text{mod}} = 0$, so ist die eindeutige Decodierung in der beschriebenen Übertragung (Sequenzierung) von Barcodes offenbar nicht sicher. Für die Einbettung von Codeworten in eine Menge von Symbolen ist es nämlich möglich im Kontext eine Sequenz zu finden, die rein zufällig einem gültigen Codewort entspricht (oder ähnlich ist) und damit gilt auch für die optimale Decodierung von Codeworten $\Pr(\mathtt{E}|\bar{\mathtt{F}}) > 0$. Für die probabilistische Decodierung der Watermark Codes ist eine analytische Anschauung für $\Pr(\mathtt{E}|\bar{\mathtt{F}})$ nicht einfach zu definieren, und für distanzbasierte Verfahren ist eine theoretische Bewertung der Restfehlerwahrscheinlichkeit ebenfalls eher komplex. Für den Spezialfall eines Bounded-Minimum-Distance-Decoding (BMD, vgl. Definition 23) mit maximal tolerierter Distanz $t = 0$ (perfekte Übereinstimmung) ließe sich die Größenordnung der Wahrscheinlichkeit $\Pr(\mathtt{E}|\bar{\mathtt{F}})$ vereinfacht angeben als proportional zu $4^{-n}(L-n)|\mathcal{C}|$, wobei für gleiche Codewortlängen n ein linearer Zusammenhang mit der Anzahl an möglichen Codeworten zu erwarten ist.

Eine Beschreibung der Effekte für die hier verwendeten zweistufigen probabilistischen Decodierung ist jedoch vielschichtiger als eine Suche nach perfekten Übereinstimmungen mit Codeworten: Betrachtet man die Codes mit den Parametern $q_1|2|3|4$ der Länge $n = 12$, so ist die (marginal) bessere Korrekturfähigkeit des (höherratigen) Codes mit $q_1 = 9$ nur durch die Struktur des inneren Codes zu erklären. Wäre nur die rein zufällige Ähnlichkeit für den bedingten Fehler $\Pr(\mathtt{E}|\bar{\mathtt{F}})$ verantwortlich, so würde auch die geschätzte Wahrscheinlichkeit unter 75 Codeworten des Codes $9|2|3|4$ zufällig eine Übereinstimmung innerhalb der 100 umliegenden Symbole zu finden deutlich größer sein als für 47 Codeworte des Codes $7|2|3|4$. Der explizite äußere Code für beide Codeschemata ist der *duale Code* des *Wiederholungscode* (auch *Parity-Check-Code* genannt) der Länge 3 und für beide Alphabete der Größe q_1 ergibt sich der gleiche minimale Hamming-Abstand des äußeren Codes, und damit nur die Möglichkeit der Fehlererkennung ohne Korrektur. Es bleibt letztlich nur zu vermuten, dass für $q_1 = 9$ beim inneren Code eine ausbalancierte Verteilung der inneren Codesymbole ungleich Null vorliegt. Die Verteilung könnte besser durch die Parametrisierung der effektiven Substitutionswahrscheinlichkeit (3.17) des HMMs im ersten Decodierschritt modelliert sein. Anzumerken ist, dass in der Übergangswahrscheinlichkeit (3.17) lediglich von einer identischen Verteilung der durch den inneren Code substituierten Stellen des Wasserzeichens ausgegangen wird. Der Einfluss der äußeren Decodierung und damit der Effekt der unterschiedlichen Größen der möglichen Codeworte sind im Decodierergebnis insgesamt eher marginal. Seiteninformationen zu den Decodierergebnissen (nicht dargestellt) zeigen, dass die fehlerhafte innere Decodierung (erste Decodierschritt) für das schlechte Decodierergebnis dominant ist. Unabhängig von der spezifischen Wahl eines inneren Codes bieten diese sehr kurzen Codeworte und deren kurzes Wasserzeichen nur sehr beschränkte Möglichkeiten zur zuverlässigen Erkennung der eingebetteten Sequenzen. Selbst im fehlerfreien Fall ist für die gegebenen Bedingungen zu erwarten, dass in 100 decodierten Reads circa ein bis zwei Decodierfehler enthalten sind. Ein praktikabler Einsatz der gezeigten Codes in einer Länge von 12nt ist daher nur eingeschränkt.

Betrachtet man auf der anderen Seite die Codes $7|2|4|6$ bzw. $9|2|4|6$ der Länge $n = 24$ (besten gezeigte Fehlerkorrektur), so ist für die simulierte fehlerfreie Sequenzierung der Barcodes in $3 \cdot 10^6$ evaluierten Simulationsschritten keine einzige fehlerhafte Zuordnung beobachtbar. Das Wasserzeichen der Länge $n = 24$ und der verwendete innere Code, welcher im Mittel nur $1/6$ der Symbole des Wasserzeichens überlagert, ermöglicht offensichtlich eine sehr zuverlässige Lokalisierung der Sequenz des Barcodes im Read. Aus theoretischer Sicht ist für die dargestellte

Decodierung bei $p_{\text{mod}} = 0$ natürlich auch ein Restfehlerplateau zu erwarten, welches auf Basis der begrenzten Anzahl von Simulationsereignissen nicht zuverlässig quantifizierbar ist. Betrachtet man für die genannten Barcodes größere Fehlerwahrscheinlichkeiten p_{mod}, so ist damit über einen großen Bereich eine robuste Decodierung erlaubt. Anders als für die diskutierten Codes der Länge $n = 12$ zeigt sich bei genauerer Analyse der zwei Schritte der Decodierung (nicht illustriert), dass hier tatsächlich die Anzahl der unterschiedlichen möglichen Codeworte limitierend für die letztendliche Entscheidung ist. Betrachtet man das Ergebnis der Decodierung im Detail, so ist für $p_{\text{mod}} = 0.09$ die erwartete Anzahl an Sequenzfehler im Bereich des Barcodes zirka 2, wobei im Mittel von 100 verarbeiteten Sequenzen nur 2-3 nicht richtig lokalisiert oder decodiert wurden. Für eine mittlere Anzahl von nur einem Sequenzfehler im Barcode ($p_{\text{mod}} = 0.04$) ist die gleiche Nummer von fehlerhaften Zuordnungen in einer Größenordnung von annähernd 1000 Sequenzen zu erwarten. Möchte man die dargestellten Ergebnisse mit anderen Veröffentlichungen Barcodes vergleichen (beispielsweise [29]), so ist hierbei stets zu berücksichtigen, dass die exakte Lokalisation der Codewortgrenzen in der hier gezeigten Evaluierung von Fehlerereignissen mit berücksichtigt wurde. In Simulationen zu distanzbasierten Barcodes wird die explizite Einbettung von Codeworten meist nicht berücksichtigt, obwohl sie dem realen Anwendungsfall entspricht. Möchte man die vorgeschlagenen Barcodes mit bester Fehlerkorrektur hinsichtlich der Coderate bewerten, so zeigt sich für $q_1 = 7$ ein sehr niedriger Wert von $R = 0.117$ bzw. eine Rate $R = 0.131$ für $q_1 = 9$.

Für klassische Szenarien der Kanalcodierung ist die Anzahl der Codeworte eines Codes nur implizit durch die gegebene Rate von Bedeutung und letztlich liegt der Fokus faktisch auf den Distanzeigenschaften und der Fehlerkorrektur. In der Anwendung von Codes als Barcodes für das Multiplexing ist die Menge der zur Verfügung stehenden Codeworte eine weitere zu berücksichtigende Qualität, die bestimmt wie viele DNA-Proben parallel sequenziert werden können. Bisher analysierte Codeschemata bieten weniger als 100 unterschiedliche Codeworte und befinden sich damit in der Größenordnung aktuell verwendeter Multiplexing-Anwendungen (vgl. [48]). Für mögliche Anforderungen zukünftiger Verfahren bietet das Prinzip der Watermark Codes eine skalierbare Konstruktion von Barcodes mit größerem Umfang: Die Codes 8|3|6|4 mit 419 Sequenzen (Rate $R = 0.181$) oder 8|3|4|4 mit 473 Codeworten ($R = 0.278$) sind hierfür beispielhaft in Abb. 3.9 hervorgehoben. Für diese Codes lassen sich der Einfluss der äußeren Codierung und die Auswirkung des längeren Wasserzeichens erkennen. Sowohl für den ersten Schritt der inneren Decodierung als auch für die Separierung der äußeren Codeworte stehen für den Code der Länge $n = 24$ zwei Blöcke mit 4 zusätzlichen Symbolen zur Verfügung, was eine erheblich zuverlässigere Decodierung erlaubt.

Welches Codeschema für welche Größe der Multiplexinganwendung und Kanalqualität am besten geeignet ist, lässt sich abschließend nicht allgemeingültig beantworten. Die Menge an geeigneten kurzen Codes für die äußere Codierung ist letztlich beschränkt und die Umsetzung der verketteten Codekonstruktion damit sehr stark eingeschränkt. Die innere Codierung lässt jedoch eine Vielzahl von Freiheitsgraden die individuell angepasst werden können, um damit die mögliche Fehlerkorrektur positiv zu beeinflussen. Für die gezeigte Auswahl an Codes, die in dieser Arbeit evaluiert wurden, lässt sich jedoch ein empfohlenes Minimum für die Länge der Barcodes formulieren. Für den sinnvollen Einsatz des Wasserzeichens zur Lokalisierung der Barcodes und eine vereinfachte zweistufige Decodierung ist eine minimale Einflusslänge von ca. 18-22nt nötig. Für eine niedrigere Anzahl an Referenzsymbolen ist die gezeigte probabilistische Dekodierung wenig zielführend.

3.4 Zusammenfassung, kritische Fragen und weiterführende Themen

Dieses Kapitel befasste sich mit der Adaption des Konzepts Watermark Codes für den Einsatz als Barcodes im Bereich der DNA-Sequenzierung. Neben der Modellierung des Anwendungsszenario des Multiplexings als eine Art Übertragungskanal und die damit verbundene Einbettung von Codeworten in Abschnitt 3.1 wurden in 3.2 essentielle Anpassungen der ursprünglichen Methodik erklärt, um eine Deocodierung von einzelnen kurzen Barcodes zu ermöglichen. Denn im ursprünglich formulierten Konzept von Davey und MacKay steht eigentlich eine kontinuierliche Übertragung über einen binären Kanal im Fokus. Zusätzlich zu der formalen Beschreibung der notwendigen Modifikationen zur Codierung und Decodierung wurden in 3.3 praxisrelevante Barcodes untersucht. Dadurch wurden gängige Anforderung für biotechnologisch kompatible Oligonukleotide in die Suche nach Barcodes integriert. Den Abschluss des Kapitels bildet eine in silico Anwendung ausgewählter Codeschemata und die Bewertung der Decodierfehler für unterschiedliche Kanalparameter. In der Simulation wurde gezeigt, dass es für den spezifizierten Sequenzierkanal möglich ist Watermark Codes als Barcodes zu nutzen, um damit die Aufgabe des Multiplexings zu erfüllen. Das Konzept der Watermark Codes ist prinzipiell geeignet für den Einsatz im Bereich der DNA-Sequenzierung der zweiten Generation und bietet einen konstruktiven Ansatz für Erzeugung von Barcodes, die eine Fehlerkorrektur für allgemeine Sequenzfehler, wie Einfügungen und Lösungen, bieten. Außerdem enthält das Prinzip eine Decodiervorschrift, die eine Lokalisation von Barcodes in einem gänzlich unbekannten Kontext von Nukleotiden ermöglicht.

Kritische Fragen

Die für die gegebenen formalen Definitionen und Simulationen implizit als bekannt geltende Parametrisierung des Kanalmodells ist die Krux der praktischen Umsetzung der Watermark Codes. Wie für alle modellbasierten Ansätze, ist die Aussagekraft von Vorhersagen (hier Likelihoods) nur so gut, wie die darunterliegenden Annahmen zu den tatsächlichen empirischen Beobachtungen passen. Leider existieren für das Kanalmodell der DNA-Sequenzierung nur wenig verbindliche Aussagen. Neben den werbewirksam angepriesenen Fehlerraten der Technologieführer sind aussagekräftige Publikationen über robuste Kanalschätzungen nicht zu finden. Das Grundproblem in diesem Kontext ist wohl ein mangelnder Konsens darüber, wie solche Daten korrekt erhoben werden. Die Evaluation von Sequenzalignments ist eine gängige Methode, um Informationen über Sequenzfehler zu erhalten, so beispielsweise in [44, 117]. Dazu werden die Sequenzen der nativen DNA-Fragmente auf ein Referenzgenom abgebildet und die dazu notwendigen Modifikationen als Sequenzfehler bewertet. Hierbei ist das Problem der Überanpassung (engl. *overfitting*) von Fehlerstatistiken durch den Einsatz vordefinierter Kostenfunktionen jedoch kritisch zu bewerten. Die Kommutativität einer Ersetzung durch eine Einfüge- und Löschoperation verdeutlichen dazu einen Freiheitsgrad, der eine Kanalschätzung erschwert.

Eine weitere starke Annahme liegt in der dargelegten Formulierung des Sequenzierkanals mittels unabhängigen und identisch verteilten Operationen auf den Symbolen der Templates. Bezieht man sich auf empirische Daten bezüglich Symbolersetzungen, so existieren Hinweise [113, 121], dass diese von einem größeren Kontext als nur einem Symbol abhängig sind. Ferner ist zu vermuten, dass Einfügungen und Löschungen nicht unabhängig, sondern auch gehäuft, auftreten. Vorbehaltlich einer existierenden probabilistischen Beschreibung der hypothetischen Phänomene dieser Modifikationen, ist eine Integration dieser Abhängigkeiten in das hier beschriebene

Hidden Markov-Modell durchaus möglich, jedoch ist dadurch eine erhebliche Vergrößerung der Komplexität der Modelle und der Decodierung zu erwarten.

Weiterführende Themen

Der essentielle Schritt für eine praktische Realisierung der Watermark Codes als Barcodes in der DNA-Sequenzierung ist zunächst die Validierung des Kanalmodells und gegebenenfalls die Erweiterung des beschriebenen HMMs. Hierzu wäre die exklusive Sequenzierung einer Menge von Oligonukleotide zur Kalibrierung und zum Training des HMMs nötig. Für die Auswahl der dafür verwendeten Oligonukleotiden könnte eine sehr große Anzahl zufälliger Sequenzen dienen, die eine möglichst große Editierdistanz zu einander aufweisen und keine Sequenzkonstellationen enthalten, die nachweislich zu negativen experimentellen Seiteneffekten führen. Der durch die Erzeugung dieser Oligonukleotide erforderliche Aufwand ist dabei nicht unerheblich. Wäre letztlich eine adäquate Beschreibung der Sequenzierung durch ein HMM gefunden, die sich auch durch empirische Daten aus Folgeexperimenten bestätigen lässt, so könnte auf Basis dieses Modells eine zusätzliche Anpassung der Barcodes erfolgen. Ausgehend von der hier eingesetzten Editierdistanz könnten allgemeinere Kostenfunktionen Potential für eine verbesserte Trennung der Barcodes bei der Decodierung beinhalten. Der Einsatz der gewichteten Editierdistanz als „probabilistische Metrik“ in [108, 123] oder die in [31] genutzte *Change Probability* könnten für eine Anpassung von Barcodes auf den Kanal dienen. Die gegenseitig bedingte Auswahl von äußerem und innerem Code ist ein weiterer Ansatzpunkt für eine Adaption der Codierung an bekannte Kanalparameter.

Beschränkt man sich hingegen auf theoretische Erweiterungen der Codierung und Decodierung von Watermark Codes als Barcodes, so ist beispielsweise die in 3.2.3 beschriebene effektive Substitution von Symbolen des Wasserzeichens für allgemeine innere Codes nicht wie angenommen unabhängig und gleichverteilt. Eine von der Position abhängige bedingte Wahrscheinlichkeit würde den Einfluss des inneren Codes bei der Lokalisation des Wasserzeichens (erster Decodierschritt) korrekt abbilden. Verbunden mit einer komplexeren Formulierung der nötigen HMMs, ist für die genannte Anpassung ein besseres Decodierergebnis zu erwarten. Eine weitere mögliche theoretische Ergänzung wäre die Integration von Zuverlässigkeitswerten in die Decodierung. Denn für Sequenzierdaten existieren im Allgemeinen Qualitätswerte der einzelnen Symbole im Read, welche als Seiteninformationen für die innere Decodierung genutzt werden können. Bezogen auf die äußere Decodierung kann die Rückführung des Decodierergebnisses zum inneren Decoder eine Verbesserung der Lokalisierung der Codeworte bewirken. Verallgemeinert kann eine iterative Decodierung einen zusätzlichen Gewinn bei der Fehlerkorrektur bewirken. Hinsichtlich der Leistungsfähigkeit der Decodierung der Watermark Codes existieren Ansätze die eine weitere Aufwandsreduktion der inneren Decodierung zeigen [84] oder die Portierung der nötigen Berechnungen auf sehr leistungsfähigen Grafikprozessoren (engl. *graphics processing units, GPUs*) vorschlagen [26]. Derartige Optimierungen schaffen zusätzliche Kapazität, um die (offenbar notwendigen) komplexeren Modelle des Sequenzierkanals berechnen zu können.

Schlussbemerkung

Obwohl es für Ansätze wie [29, 48] im Allgemeinen möglich ist sehr gute Codes zu finden, die den Watermark Codes hinsichtlich der Rate oder der minimalen Editierdistanz überlegen sind, so existierten dennoch zwei wesentliche Punkte, die berücksichtigt werden sollten: Erstens, der

hier vorgestellte Ansatz für Barcodes beinhaltet eine implizite Methodik zur Detektion von Codeworten in gänzlich unbekannten Sequenzen. Die probabilistische Integration eines Wasserzeichens ermöglicht den freien Einsatz von Oligonukleotiden zur Markierung von DNA, ohne bei der Decodierung auf feste Präfixe oder Postfixe angewiesen zu sein. Diese Einbettung von Barcodes ist in rein distanzbasierten Ansätzen nicht verallgemeinert. Zweitens existiert eine klar definierte Decodiervorschrift, die auf Basis einer probabilistischen Beschreibung des Kanals ein diesbezüglich angepasstes Vorgehen darstellt. Sicherlich ist die Decodierung mittels HMM auch für Codes ohne Wasserzeichen möglich, dennoch beinhaltet das Verfahren in zwei Schritten (Lokalisation und Decodierung) einen Gewinn in Decodierkomplexität, speziell wenn Multiplexing-Szenarien und Sequenzierlängen nach oben skalieren. Sowohl für die Parallelisierung als auch für mögliche Sequenzierlängen ist hierfür über die letzten Jahre betrachtet eine dramatische Steigerung zu vermerken. Eine Trendwende ist diesbezüglich auch in nächster Zeit nicht zu erwarten. Hinsichtlich der Skalierbarkeit für eine massive Parallelisierung durch Multiplexing ist das Konzept der Watermark Codes eine adäquate Alternative zu klassischen Ansätzen zur Markierung von DNA durch Barcodes.

4

Neuartige Zufallsbarcodes und ihre Anwendung

RNA-Sequenzierung ist innerhalb der letzten Dekade zu einem der revolutionären Konzepte zur Analyse der Transkription geworden. Auf Basis der reversen Transkription und der Überführung von RNA-Molekülen in cDNA lassen sich für DNA etablierte Technologien auch für die Sequenzierung von RNA nutzen. Neben zahlreichen Vorteilen dieser Technik existieren auch einige Schwierigkeiten, wie beispielsweise für die exakte Quantifizierung von RNA-Molekülen [131, 169]. Zwei essentielle Herausforderungen sind dabei maßgeblich [82]: Zum einen die Effizienz der cDNA Synthese (vgl. Begriff 10), welche als initialer Vorgang erheblichen Einfluss auf die Anzahl der verwertbaren (doppelsträngigen) Transkripte hat und damit einen limitierenden Faktor bezüglich der Beobachtbarkeit von Molekülen bildet. Andererseits hat der sogenannte Amplifikations-Bias der PCR (vgl. Begriff 18) negative Auswirkungen auf die Genauigkeit von quantitativen Analysen. Um diesem Effekt adäquat entgegenzuwirken ist der Einsatz von Zufallsbarcodes (vgl. Abschnitt 2.2.3) Stand der aktuellen Technik. Die synthetisch erzeugten Sequenzen werden zur eindeutigen Markierung von RNA-Molekülen vor der PCR eingesetzt und ermöglichen eine Korrektur von Zählgrößen nach der Sequenzierung. Dabei ist die eindeutige Markierung eher als optimales Ziel zu verstehen. Die Anzahl an unterschiedlichen Markierungen und deren Verteilung zum Zeitpunkt der Markierung, zusammengefasst als Diversität bezeichnet, sind die entscheidenden Qualitätsmerkmale von Zufallsbarcodes und deren Eignung zur PCR-Korrektur. Hinsichtlich der Konstruktion und Integration von Codes existieren unterschiedliche Konzepte, welche in Abb. 4.1 schematisch dargestellt sind: In [77] wurden zwei Verfahren theoretisch ausgearbeitet, welche *einseitige* Zufallsbarcodes konkretisieren. Beim *multimeric linker labeling* (links im Bild) wird die zufällige Länge von repetitiven synthetischen Molekülsequenzen (Oligonukleotiden) zur Erzeugung der Diversität der Zufallsbarcodes genutzt. Die Phrase „*mer*“ bezieht sich dabei auf den Begriff Polymer, der den Molekülverbund beschreibt. Für das *tag labeling* (Bildmitte) wurden unterschiedliche Sequenzen fixer Länge zur Markierung von Molekülen vorgeschlagen. Letzteres wurde unabhängig von der Sequenzierung beispielsweise in [24, 25] zur Anwendung gebracht. Obwohl die *multimeric linker* auf der Basis von Wiederholungscodes prinzipiell eine Fehlerkorrektur ermöglichen, wurde in [77] nicht auf Aspekte der Kanalcodierung eingegangen. Auch neuere Formen [34, 56, 82, 88] des *tag labeling* nutzen nur selten Konzepte der Kanalcodierung, welche jedoch zur Steigerung der Robustheit von experimentellen Analysen durchaus sinnvoll wären. Eine *zweiseitige* Markierung von cDNA durch Zufallsbarcodes mit Fehlerkorrektur wurde in [153] vorgestellt (mitte-rechts in Abb. 4.1): Die zufällige paarweise Verknüpfung von einer kleinen Menge von Oligonukleotiden ermöglicht hierbei die aufwandsreduzierte Erzeugung einer quadratischen Anzahl an Molekülkombinationen und damit eine hohe Diversität der Zufallsbarcodes.

In diesem Kapitel werden neuartige Zufallsbarcodes vorgestellt, die als logische Erweiterung des *multimeric linker labeling* und *tag labeling* verstanden werden können. Bei der neuen Technik,

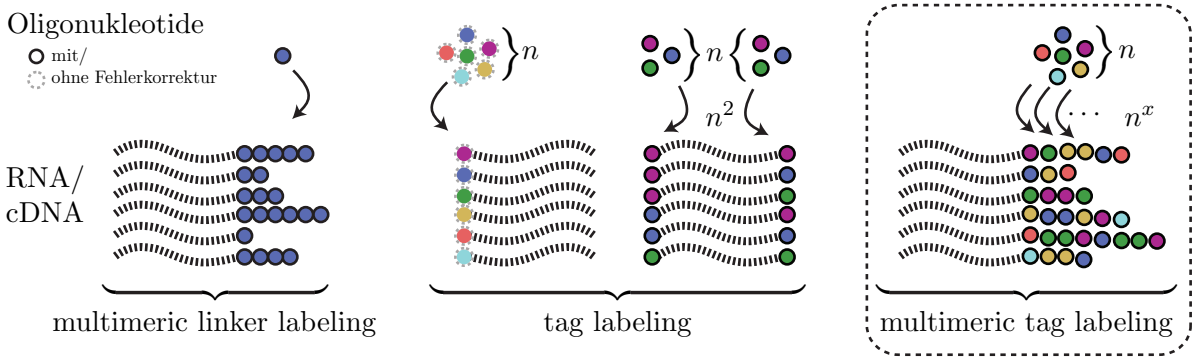

Abb. 4.1 Konstruktion von Zufallsbarcodes an RNA/cDNA: Auf Grundlage einer Basismenge an synthetisch erzeugten Sequenzen (Oligonukleotiden) ist eine Kombination von unterschiedlichen Markierungen möglich. Beim *multimeric linker labeling* wird durch die zufällige Kettenbildung eines Oligonukleotids (*linker*) Diversität über die Verteilung der Kettenlänge erreicht. Für das *tag labeling* muss gegebenenfalls eine große Menge unterschiedlicher Sequenzen synthetisiert werden, um eine unterschiedliche Markierung von Molekülen zu ermöglichen. Das Basisrepertoire an Oligonukleotiden kann durch eine zweiseitige Kombination von Markierungen erheblich reduziert werden. Der Einsatz von fehlerkorrigierenden Sequenzen ist dabei eine unabhängige Option. In dieser Arbeit wird das *multimeric tag labeling* als systematische Erweiterung vorgestellt.

multimeric tag labeling genannt, werden auf Basis von speziell entworfenen Barcode-Templates Zufallsbarcode variabler Länge erzeugt, die einer Fehlerkorrektur für Substitutionsfehler bieten. Als eine systematische Erweiterung bekannter Konzepte wird es somit möglich aus einer Basismenge an unterschiedlichen Oligonukleotiden die effektive Anzahl an Kombinationen beliebig zu exponentieren. Für die konkrete Anwendung des generischen Konzepts in der RNA-Sequenzierung finden die Barcode-Templates in erweiterten RT-Primern Verwendung, wodurch eine unmittelbare Markierung der RNA-Moleküle bei der Synthese von cDNA sichergestellt wird. Diese frühe Anheftung der Zufallsbarcodes an RNA-Moleküle reduziert zusätzliche Fehlerquellen bei der reversen Transkription auf ein Minimum und ermöglicht eine problemlose Integration in bestehende Protokolle, wie am Beispiel der Illumina TruSeq Small RNA Technologie [78, 79] demonstriert wird. Die experimentelle Umsetzung der gezeigten Konzepte wurde in Zusammenarbeit mit Experten im Gebiet der Molekularbiologie[1] erreicht.

Das vorliegende Kapitel gliedert sich wie folgt: Ausgehend vom Abschnitt 4.1 werden zunächst der Entwurf und die Erstellung der neuartigen Zufallsbarcodes dargelegt. Des Weiteren werden in 4.2 besondere Aspekte der Bioinformatik beschrieben, welche für die Anwendung der gezeigten Zufallsbarcodes notwendig sind. Die Auswertung beispielhafter RNA-Sequenzierungen des Modellorganismus *Escherichia coli* (kurz *E. coli*) bildet den Kern von 4.3. Der Abschnitt 4.4 beendet das vorliegende Kapitel mit einer Zusammenfassung und dem Verweis auf weiterführende Themen.

[1] ZIEL (Zentralinstitut für Ernährungs- und Lebensmittelforschung), Abteilung Mikrobiologie, TU München

4.1 Entwurf und Erstellung der Zufallscodes

Das in dieser Arbeit entwickelte Konzept des *multimeric tag labeling* wird exemplarisch für die sogenannte Illumina TruSeq Technologie[2] erklärt. Der wesentliche Bestandteil des Konzepts sind sogenannte Barcode-Templates (spezielle Oligonukleotide), die in zufälliger Kombination in das bestehende experimentelle Protokoll der Technologie integriert werden. Zusammenfassend dient das Protokoll zur Erzeugung von sequenzierfähiger cDNA aus RNA Fragmenten (vgl. Abschnitt 2.2.2). An diesen vorbereitenden Schritten sind fest vorgegebene technische Sequenzen beteiligt, die teilweise adaptiert oder ersetzt werden müssen, um Barcode-Templates zu integrieren. Zum besseren Verständnis wird in 4.1.1 zunächst das Standard Illumina TruSeq Protokoll erklärt (vgl. dazu Grundlagenteil Abschnitt 2.2.2). Darin auftretende technische Sequenzen sind in Anhang A.1 explizit aufgeführt. Die Integration und notwendige Anpassungen werden im Anschluss in 4.1.2 spezifiziert. Den Abschluss bildet der theoretische Entwurf und konkrete Parametrisierung von Zufallsmolekülen in Abschnitt 4.1.3. Das in dieser Arbeit entwickelte Konzept der Zufallsbarcodes wurde für ein konkretes Protokoll implementiert, kann jedoch prinzipiell unabhängig von einer bestimmten Sequenziertechnologie angewendet werden.

4.1.1 Standard TruSeq Sequenzier-Library

Das Illumina TruSeq Small RNA Protokoll wird genutzt um ein breites Spektrum von RNA-Molekülen für deren Sequenzierung aufzubereiten. Als Ausgangspunkt für die Erklärungen soll ein beispielhaftes RNA Fragment, im Folgenden als Insert bezeichnet, dienen (vgl. Abb. 4.2, links). Ein Fragment wird durch das Kürzel RNA symbolisiert, andere Sequenzen sind ebenfalls durch Bezeichner in Blockschrift zu erkennen. Vor der eigentlichen Sequenzierung durchläuft das Insert unter anderem diese Schritte:

1. Ligation des 3'-Adapters RA3 am 3'-Ende des RNA-Inserts mittels RNA Ligase: Eine spezielle Ligase in Verbindung mit einem adenylierten 5'-Ende des Adapters garantiert eine Abfolge RNA-RA3 und verhindert eine Schleifenbildung der RNA.
2. Ligation des 5'-Adapters RA5 am 5'-Ende des Inserts RNA-RA3 mittels RNA Ligase.
3. reverse Transkription wird initiiert durch Hybridisierung des RNA RT-Primer RTP am 3'-Ende des mit Adaptern flankierten RNA Moleküls (komplementär ergänzte Sequenzen werden durch Kleinbuchstaben symbolisiert).
4. Das einzelsträngige Molekül wird durch die reverse Transkriptase zur cDNA ergänzt.
5. Start der PCR-Reaktion zur Amplifikation der cDNA mit Standard PCR-Primer RP1 und PCR Index Primer RPIX (siehe PCR, Begriff 11). In Zyklus 1 entstehen verkürzte Zwischenprodukte, welche in der gesamten PCR-Reaktion zu vernachlässigen sind.
6. Fortführung der PCR (ab Zyklus 2): Quasi-exponentielle Vervielfältigung der cDNA-Moleküle mit größter Länge, welche beide Primer beinhalten. Der Prozess führt zu einer erheblichen Anreicherung der Zielsequenzen, im Vergleich zu anderen kürzeren cDNAs.

Details bezüglich der vorgegebenen Sequenzen sind in Anhang A.1 zu finden. Weiter Informationen zum Protokoll können den Datenblättern des Herstellers Illumina entnommen werden.

Die grundlegende Idee der hier vorgestellten Zufallsbarcodes ist es, das RNA-Insert am 3'-Ende durch eine Aneinanderreihung zufälliger Kombinationen von Barcode-Templates zu erweitern.

[2] Die TruSeq Small RNA Technologie ist ein eingetragenes Markenzeichen der Firma Illumina [78, 79].

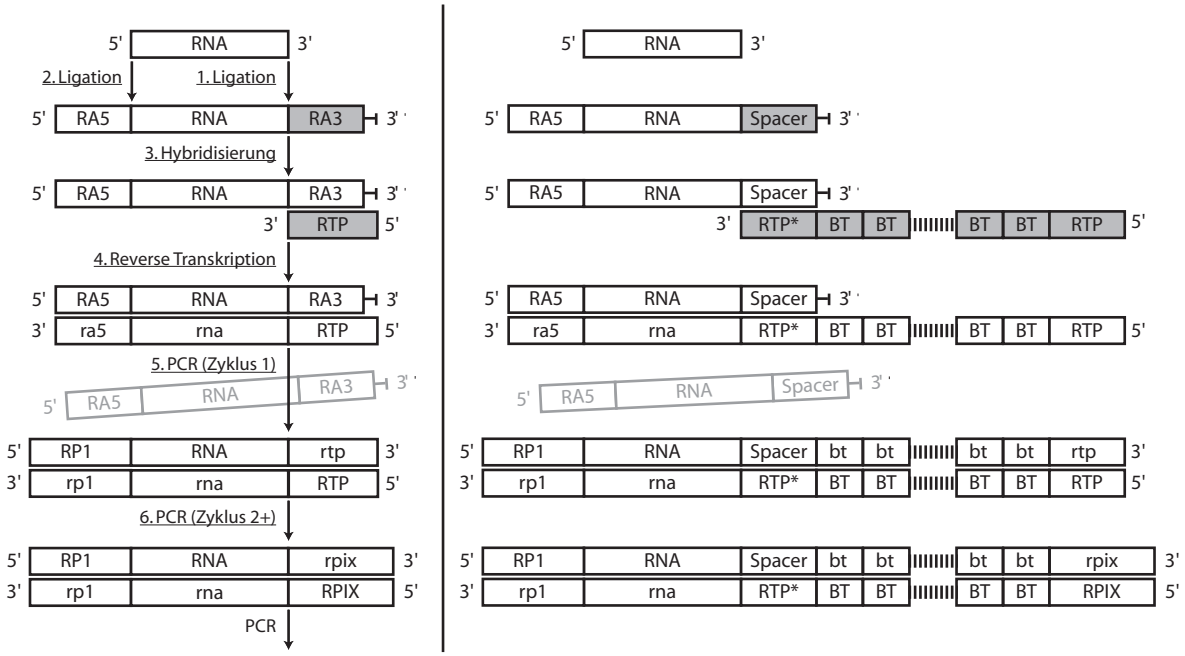

Abb. 4.2 Ausschnitt des TruSeq Protokolls, Standard (links), modifiziert (rechts): Die Substitution des RA3-Adapters durch den *Spacer* ermöglicht die Erweiterung des RNA-Inserts durch Barcode-Templates. Wegen der Verwendung eines modifizierten RT-Primers unter Beibehaltung der Sequenz RTP am 5'-Ende bleibt das Protokoll im Ablauf nach der reversen Transkription unverändert.

Das Kernelement zu der hier vorgeschlagenen Realisierung ist die Ersetzung des RT-Primers durch ein modifiziertes Molekül, welches als Träger eines randomisierten Postfixes dient. Im folgenden Abschnitt 4.1.2 wird zunächst implizit von der Existenz dieses modifizierten RT-Primers ausgegangen und dessen Einsatz im Protokoll dargelegt. Ausgehend von einer groben Vorstellung des RT-Primer wird in kurzen Absätzen der konkrete Entwurf dieses Moleküls präzisiert: Dazu gehört zum einen der theoretische Aufbau des RT-Primer als Kombination von ursprünglichen technischen Sequenzen mit Barcode-Templates. Die Formulierung der konkreten Anforderungen an die Barcode-Templates und die Suche nach gültigen Sequenzen bildet den Abschluss des folgenden Abschnitts. Die Bedingungen sind insbesondere wichtig, weil molekulare Wechselwirkungen im Allgemeinen sequenzabhängig sind, was zu unerwünschten Seiteneffekten führen kann.

4.1.2 RT-Primer als Zufallsmolekül aus Barcode-Templates

Die Integration einer zufälligen Kombination von Barcode-Templates ist durch eine geschickte Substitution des RT-Primers in Verbindung mit dem Ersatz des Adapters RA3 möglich, welcher im Folgenden als Spacer (Spacer) bezeichnen wird. Die dafür notwendigen Änderungen des Protokolls sind in Abb. 4.2 (rechts) dargestellt: Analog zum Standardverfahren werden RA5-Adapter und der später näher spezifizierte Spacer an das RNA-Insert ligiert. Der in einem separaten Prozess (siehe Abschnitt 4.1.3) erzeugte modifizierte RT-Primer trägt neben der Sequenz zur Hybridisierung mit dem Spacer (als RTP* bezeichnet) eine variable Folge von

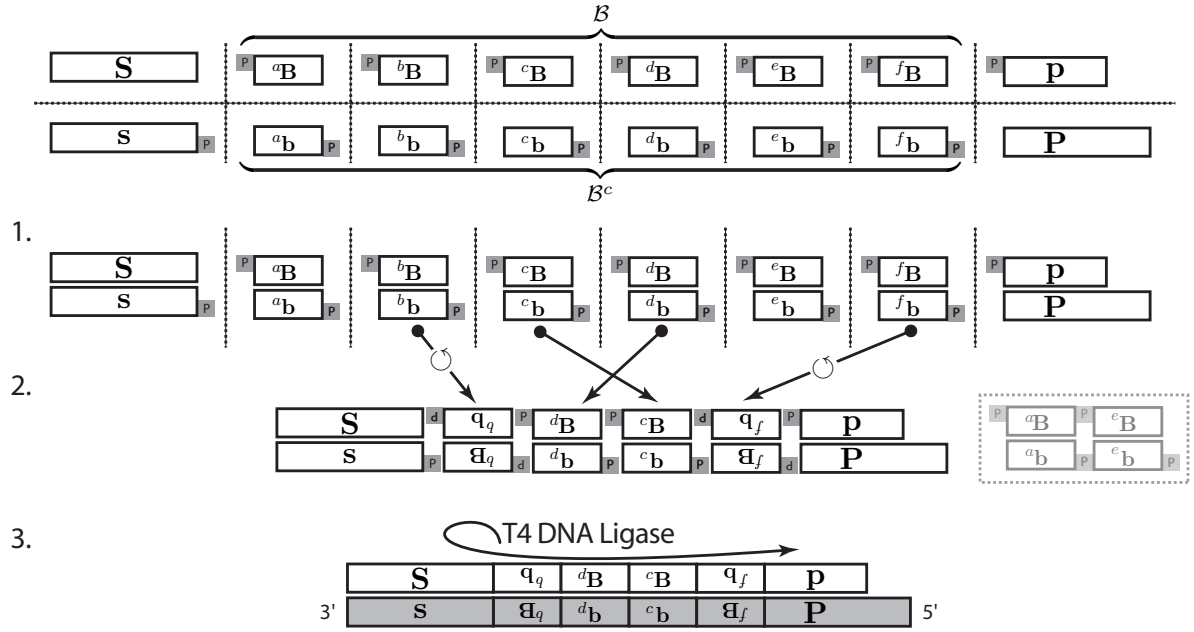

Abb. 4.3 Ligationsmodell für RT-Primer: 1. Hybridisierung synthetisierter DNA-Einzelstränge, 2. Zufällige Kombination im Gefüge aus DNA-Molekülen (ausgegraut: beispielhaftes Zwischenprodukt), 3. Ligation durch T4 DNA Ligase; RT-Primer als grauer Einzelstrang.

Barcode-Templates (BT genannt). An deren 5'-Ende befindet sich zusätzlich eine zum Standard RT-Primer RTP homologe Sequenz. Die Barcode-Templates BT symbolisieren unterschiedliche Repräsentanten einer definierten Basismenge an Oligonukleotiden. Der in der reverse Transkription elongierte untere Einzelstrang ist anhand der Sequenzen an den 3'-/5'-Enden nicht von dem Pendant des Standardvorgehens zu unterscheiden. Die von diesem Molekül ausgehende Hybridisierung und Elongation verlaufen in der PCR analog zum Standard-Protokoll. Auf die beschriebene Weise wird es möglich RNA-Inserts um randomisierte synthetische Sequenzen zu ergänzen, wobei der Spacer-Sequenz die Rolle einer fixen Präambel zukommt. Die Markierung der RNA-Moleküle mit Zufallsbarcodes findet somit unmittelbar und möglichst früh im experimentellen Ablauf statt.

Im folgenden Absatz werden der Aufbau und die Ligation des modifizierten RT-Primers weiter spezifiziert. Bei dem letztendlichen Primer handelt es sich um ein einzelsträngiges DNA-Molekül, das jedoch nicht direkt durch die Kombination von DNA-Einzelsträngen erzeugt wird. Vielmehr erfolgt die Ligation auf Basis von doppelsträngigen Molekülen und dem anschließenden Aufschmelzen und Filtern der gewünschten Primer. Die Erzeugung der klar definierten essentiellen Sequenzstruktur RTP*+BT+$\cdots$+BT+RTP (vgl. Abb. 4.2) stellt dabei spezielle Anforderungen an deren Bestandteile.

Ligationsmodell des RT-Primer

Die Erstellung des RT-Primers erfolgt in drei Schritten unter Verwendung synthetisierter DNA-Einzelstrangsequenzen[3] (siehe Abb. 4.3), wie: der Spacer $\mathbf{S} = S_1S_2S_3..S_m\rangle$ und dessen Kom-

[3] Die Symbole $\rangle$ und $\langle$ kennzeichnen 3'-Enden, Groß- und Kleinbuchstaben sind komplementäre Ergänzung.

plement (RTP*) als $\mathbf{s} = \langle s_m..s_3 s_2 s_1$, der Primer (RTP) als $\mathbf{P} = \langle P_n..P_3 P_2 P_1$ und dessen verkürztes Komplement $\mathbf{p} = p_1 p_2..p_o\rangle$ mit $o \leq n$ und zwei Mengen sich komplementär ergänzender Barcode-Templates $\mathcal{B}$ und $\mathcal{B}^c$ der Länge l, mit $l, m, n, o \in \mathbb{Z}$. Für alle Templates gilt außerdem

$$ {}^iB_1\, {}^iB_2..\, {}^iB_l\rangle = {}^i\mathbf{B} \in \mathcal{B} \quad \Leftrightarrow \quad \mathcal{B}^c \ni {}^i\mathbf{b} = \langle {}^ib_l..\, {}^ib_2\, {}^ib_1. $$

Die Spezifizierung der DNA-Einzelstränge ist sinnvoll, um nach der Hybridisierung der komplementären Gegenstücke überstehende (einzelsträngige) Enden zu ermöglichen. Mit dem eigentlichen Ziel der Hybridisierung zu DNA-Molekülen und einer anschließenden Ligation ist es nötig die komplementären Pendants so zu modifizieren, dass eine gerichtete Ligation begünstigt wird. Bekannt ist, dass die Verkürzung eines Strangs eine DNA-Ligation hemmt und eine *Phosphorylierung* der 5'-Enden die Reaktion begünstigt. Die Verkürzung des Komplements $\mathbf{p}$ und eine gezielte Phosphorylierung der mit $\mathbf{s}$ und $\mathbf{p}$ deklarierten Moleküle dienen der gerichteten Ligation und der Ausbildung klar definierter Enden des RT-Primers. Für die mit $\mathcal{B}$ und $\mathcal{B}^c$ bezeichneten Moleküle sind 3 Strategien zur Phosphorylierung der 5'-Enden denkbar. Eine gleichförmig gerichtete Ligation kann jedoch durch kein Vorgehen sichergestellt werden. Der dadurch entstehende Freiheitsgrad stellt eine besondere Herausforderung an die Auswahl der Barcode-Templates.

Im ersten Schritt des Verfahrens werden die komplementär ergänzenden Sequenzen zusammengefügt, welche zu doppelsträngigen DNA-Molekülen hybridisieren (vgl. Abb. 4.3). Das Gefüge aus resultierenden DNA-Molekülen unterläuft in einem zweiten Schritt eine zufällige Kombination. Dabei ist eine freie Rotation von Barcode-Templates möglich. Neben Molekülketten, welche durch Spacer- oder Primer-Sequenzen terminiert werden, existieren in diesem Schritt Zwischenprodukte, welche an beiden Enden durch weitere Moleküle ergänzt werden können. Eine beständige langkettige Form wird durch die abschließende Reaktion einer speziellen DNA-Ligase erreicht.

Neben der hier beschriebenen symmetrischen Form der Phosphorylierung, existiert zusätzlich eine asymmetrische Alternative, die sich jedoch im Experiment als nicht praktikabel erwiesen hat. Probleme, die aus einer asymmetrischen Phosphorylierung resultieren, sind in Anhänge A.2 und A.6 genauer beschrieben.

Der letztlich im symmetrischen Verfahren erzeugte RT-Primer liegt in hybridisierter (doppelsträngiger) Form vor. Um den funktionalen (unteren) Einzelstrang (vgl. Abb. 4.3) zu trennen, bedient man sich des Größenunterschiedes der Einzelstränge im Molekülverbund und der Gelelektrophorese. Die Möglichkeit die Größenunterschiede für eine konkrete Implementierung optimal zu nutzen wird im Abschnitt zur Größenseparierung (4.1.3) näher ausgeführt. Zunächst werden die allgemeinen Anforderungen an Barcode-Templates und die Suche nach validen Sequenzen diskutiert.

Anforderungen an rotations-immune Barcode-Templates

Neben der freien Rotation in der zuvor beschriebenen Ligation ergeben sich Gründe, welche zusätzliche Kriterien an Barcode-Templates motivieren:

1. Um eine Fehlerkorrektur von t Substitutionsfehlern zu ermöglichen, muss für den minimalen Hamming-Abstand $d_{\min}$ zwischen allen Templates $d_{\min} \geq 2t + 1$ erfüllt sein.

2. Fehlerhafte Teilhybridisierung von Sequenzen $\mathbf{S}, \mathbf{s}, \mathbf{P}, \mathbf{p}$ und den Sequenzen aus $\mathcal{B}$ und $\mathcal{B}^c$ können zu unbeabsichtigten Molekülstrukturen und unvorhersehbaren Abweichungen bezüglich der Primer-Ligation führen. Homologe Abschnitte und deren Komplemente sind in Kombinationen der Sequenzen zu vermeiden.
3. Die Sequenzen $\mathbf{S}, \mathbf{s}, \mathcal{B}$ und $\mathcal{B}^c$ sollten das experimentelle Protokoll und die Qualität der Sequenzierung nicht negativ beeinflussen. Homologe Abschnitte in Kombination mit Sequenzen der TruSeq Technologie sind zu vermeiden, ebenso wie Sequenzabfolgen mit ungünstigem GC-Gehalt oder Homopolymeren. Eine breite Übereinkunft bezüglich der Anforderung an Barcodes kann in [29, 30, 32, 48, 54, 153] gefunden werden.
4. Letztendlich sollte die Diversität der zufälligen RT-Primer möglichst groß sein, d. h. eine Maximierung der Menge an Barcode-Templates wird angestrebt.

Eine Illustration der problematischen Konstellationen bei der Kombination von Oligonukleotiden ist in Anhang A.3 gegeben.

In Anlehnung an das von Mir et al. [118] publizierte Verfahren zur *Randomized Construction of Codes* ist im Anhang A.4 ein verwandter Algorithmus dargestellt, der zur Suche nach Sequenzen mit den genannten Eigenschaften verwendet wird. Ein ähnliches Konzept zur Suche nach weniger restriktiven und spezifischen Barcodes wurde als Algorithmus *barcrawl* in [54] vorgestellt. Im nun folgenden Paragraph wird die Grundidee des erweiterten Suchalgorithmus skizziert. Für die detaillierte Pseudoimplementierung wird auf den Anhang verwiesen.

Konzept des Suchalgorithmus

Kurzgefasst handelt es sich bei dem Algorithmus um einen randomisierten Filterprozess unter Berücksichtigung der folgenden Prämissen:

- Die Suche wird auf eine Menge $\mathcal{B}$ von Codeworten reduziert, für welche gilt: ist ein Codewort $\mathbf{B} \in \mathcal{B}$, so ist das reverse Komplement[4] $\mathbf{B}'$ ebenfalls enthalten, d. h. $\mathbf{B} \in \mathcal{B} \Leftrightarrow \mathbf{B}' \in \mathcal{B}$.
- Ohne weitere Einschränkungen soll der Spacer aus einer Kombination von Sequenzen aus $\mathcal{B}$ bestehen, was die Suche auf eine geschlossene Menge $\mathcal{B}$ von Codeworten reduziert.

Für eine Menge zu prüfender Codeworte wird in zufälliger Reihenfolge eine zweistufige Filterung durchgeführt: In einer Vorfilterung werden die Anforderungen (siehe oben) für einzelne Codeworte sichergestellt. In einer zweiten Phase wird eine Liste von bereits gültigen Codeworten ergänzt, falls die genannten Kriterien für alle möglichen Codewort-Paarungen erfüllt sind. Dabei kommt der Rotationssymmetrie bei der Bildung von Kombinationen eine besondere Bedeutung zu. Das veranschaulichte Konzept und die daraus resultierenden *rotations-immunen Barcode-Templates* stellen eine Neuerung dar, welche hier erstmalig vorgestellt und experimentell zu Anwendung gebracht werden.

4.1.3 Herstellung konkreter RT-Primer

Für die experimentelle Umsetzung der vorangehenden allgemeinen Beschreibung müssen weitere Größen konkretisiert werden:

[4] Das reverse Komplement ${}^{i}\mathbf{B}'$ von ${}^{i}\mathbf{B}$ ist in Abb. 4.3 als rotierter Kleinbuchstabe ${}^{i}\mathbf{b}$ zu verstehen.

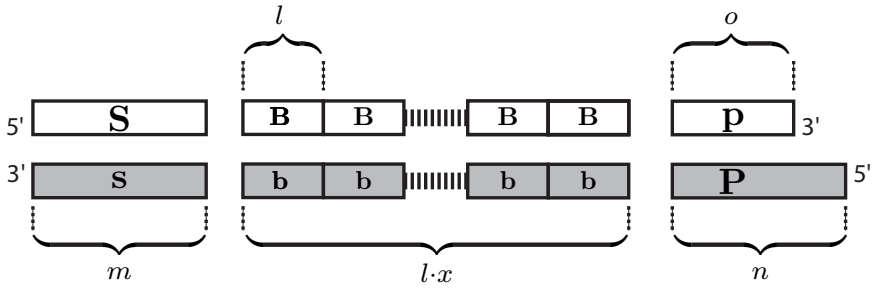

Abb. 4.4 Bestandteile der RT-Primer und deren Längen

- Neben den Anforderungen an die Barcode-Templates muss die Länge $l_{\mathrm{RT}}(x) = m+l\cdot x+n$ der zur Anwendung kommenden RT-Primer festgelegt werden (vgl. Abb. 4.4). In Bezug auf die Beschränkung einer maximalen Sequenzierlänge (vorgegeben durch die Sequenzierplattform) ist ein Kompromiss nötig, zwischen der geplanten Länge der RNA-Inserts und der Anzahl x an gewünschten Barcode-Templates. In Verbindung mit der Anzahl an unterschiedlichen Barcode-Templates definiert x die Anzahl möglicher Zufallsmoleküle.
- Des Weiteren können die Längen l, m, n, o der einzelsträngigen Bestandteile des RT-Primer variiert werden, um die Größenseparierung (Unterscheidbarkeit) der Zielmoleküle zu vergrößern. Eine geschickte Wahl der Länge m des Spacers ist hier sinnvoll.
- Für die Ligationsreaktion muss letztlich ein adäquates Mischungsverhältnis der physischen Moleküle $\{\mathbf{s}, \mathbf{S}\}$, $\{\mathbf{p}, \mathbf{P}\}$ und der Barcode-Templates $\{\mathbf{b}, \mathbf{B}\}$ getroffen werden.
- Die Größenselektion mittels Gelelektrophorese zur Isolierung der Moleküle mit gesuchter Länge bildet den Abschluss der Herstellung der modifizierten RT-Primer.

Konkrete Parametrisierung

Der Such-Algorithmus (siehe A.4) wurde für Codeworte der Länge $l = 9$, mit einem minimalen Hamming-Abstand von $d_{\mathrm{h}} = 5$ (Korrekturfähigkeit: 2 Substitutionen pro Template) spezifiziert. Weitere Parameter waren: ein GC-Gehalt zwischen $\underline{r_{\mathrm{GC}}} = 40\%$ und $\overline{r_{\mathrm{GC}}} = 60\%$, eine maximale Länge von Homopolymeren $l_{\mathrm{h}} = 3$ und Homologe Abschnitte beschränkt auf Längen $l_{\mathrm{s}} = 3$ (Codeworte untereinander) und $l_{\mathrm{t}} = 6$ (zu technische Sequenzen). Die Auflistung relevanter Sequenzen ist in A.1 zu finden. Die Suche führte zu einer Menge $\mathcal{B}$ von 44 Barcode-Templates (siehe A.5). Als Zufallsbarcodes sollen Kombinationen von $x = 4$ zufälligen Barcode-Templates pro Primer genutzt werden, wie sie in Abb. 4.3 (in grau) dargestellt sind. Eine spezielle Selektion (siehe Paragraph Größenselektion) dieser Moleküle aus den vielfältigen in der Ligation erzeugten Strukturen erfordert weitere Überlegungen bezüglich der Längen der eingesetzten Oligonukleotide.

Größenseparierung

Prinzipiell entstehen bei der zufälligen Ligation (siehe Abb. 4.4) unterschiedlichste einzelsträngige Moleküle, jedoch nur ein funktionaler RT-Primer aus $x = 4$ Barcode-Templates. Um diese Molekülkonstellation bestmöglich extrahieren zu können, ist eine allgemeine Betrachtung der Moleküllängen angebracht: Insgesamt entstehen acht Klassen an einzelsträngigen Molekülen der Längen

$$
\begin{aligned}
l_1(x) &= m+l \cdot x, & l_5(x) &= o+ l \cdot x + n,\\
l_2(x) &= m+ l \cdot x + m, & l_6(x) &= \quad l \cdot x + o,\\
l_3(x) &= m+ l \cdot x + o, & l_7(x) &= \quad l \cdot x + n\\
\underline{l_{\mathrm{RT}}(x) = l_4(x)} &= m+ l \cdot x + n, \quad \text{und} & l_8(x) &= \quad l \cdot x,
\end{aligned}
$$

mit $x \in \{0, 1, ..\}$ beteiligten Templates. Zur Reduzierung der Klassen wird vereinfachend (ohne Einschränkungen) $o = l$ gesetzt werden, wodurch folgende zusammengefasste Moleküllängen verbleiben:

$$
\begin{aligned}
l_{1/3}(x) &= m+l \cdot x, & l_{5/7}(x) &= l \cdot x + n\\
l_2(x) &= m+ l \cdot x + m, \quad \text{und} & l_{6/8}(x) &= l \cdot x.\\
\underline{l_{\mathrm{RT}}(x) = l_4(x)} &= m+ l \cdot x + n,
\end{aligned}
$$

Für eine effiziente Größenselektion muss sich die Länge l_{RT} der RT-Primer maximal von anderen Längen unterscheiden. Die minimalen Abstände δ_* von l_{RT} zu den einzelnen Längen sind

$$\delta_{1/3} = \min\{\quad \pm n \mod l\}, \tag{4.1}$$

$$\delta_2 = \min\{\pm(m-n) \mod l\}, \tag{4.2}$$

$$\delta_{5/7} = \min\{\quad \pm m \mod l\}, \tag{4.3}$$

$$\delta_{6/8} = \min\{\pm(m+n) \mod l\}. \tag{4.4}$$

So ergeben sich beispielsweise $[\pm l_{1/3}(x_1) \mp l_{\mathrm{RT}}(x_2) \mod l]$ als relevante Längenunterschiede für den minimalen Abstand $\delta_{1/3}$, wenn Vielfache x_1, x_2 der Länge l berücksichtigt werden. Im konkreten Fall mit $l = 9$ und dem Primer (RTP) der Länge $n = 21$ (vgl. A.1) ergibt sich ein limitiertes Optimierungspotential hinsichtlich der Länge m des Spacers. Dieser wird, wie bereits erwähnt, aus Barcode-Templates zusammengesetzt, wozu die Menge $\mathcal{B}$ um vier Elemente (zwei komplementäre Codepaare) reduziert wird (vgl. A.5). Die Codeworte ergeben den Spacer der Länge $m \leq 2l$. Eine Verkürzung auf $m = 15$ erscheint trotz $\delta_{6/8} = 0$ als sinnvoll, weil es sich bei Molekülen der Klasse 6/8 um latente Zwischenstufen handelt, von welchen angenommen werden kann, dass sie in der experimentellen Anwendung unbeteiligt sind, sprich nicht hybridisieren (weil Sequenzen RTP* bzw. RTP fehlen). Alle weiteren Abstände sind identisch mit $\delta_{1/3} = \delta_2 = \delta_{5/7} = 3$, was einer Selektion von 72nt ± 2nt für die Länge des RT-Primer entspricht[5].

Das hier gezeigte Konzept der Größenseparierung wurde in verallgemeinerter Form dargelegt und kann für unterschiedliche Parametrisierungen der RT-Primer universell angepasst und optimiert werden.

Für einen einfacheren Bezug auf ligierte Strukturen wird folgende vereinfachte Notation verwendet: Mit den Symbolen $\overline{\underline{\mathrm{S}}}$, $\overline{\underline{\mathrm{B}}}$ und $\overline{\underline{\mathrm{P}}}$ werden im Folgenden die DNA-Moleküle bezeichnet, welche aus Spacer, Barcode-Templates und Primer (RTP) gebildet werden und an der zufälligen Kombination des RT-Primers beteiligt sind. Die tatsächliche Orientierung der Oligonukleotide wird dabei, unter der Annahme einer wohldefinierten Ligation (zwischen den 5'- und 3'-Enden), als gegeben angesehen. Für einzelsträngige Strukturen wird die Lage der RNA-Sequenz in Abb. 4.2 als Bezugspunkt für die Leserichtung definiert und für die Unterscheidung von komplementären

[5] Länge des RT-Primer: 72nt = 15nt Spacer + 36nt (4 Barcode-Templates) + 21nt RTP.

Sequenzen Groß- und Kleinbuchstaben verwendet. Für den Bezug auf die in Abb. 4.3 dargestellte Sequenz wird somit `SbBBbp` verwendet. Diese Darstellung ist mit der Orientierung aus später folgenden Sequenzierungen konform. Für die kompakte Darstellung von Klassen von Sequenzen und Mustern wird im Folgenden eine Notation in Anlehnung an Reguläre Ausdrücke verwendet. Dabei stellt $\{\ \}^*$ eine optionale Wiederholung der Elemente der Menge dar, $[\ |\ ]$ die alternative Auswahl. Der Quantor $*$ kann dabei durch eine ganze Zahl spezifiziert werden.

Ligationsreaktion

Die Ligation entspricht im Grunde einer *Kettenpolymerisation* von DNA, an welcher zwei Klassen von Molekülen beteiligt sind: Zum einen die Klasse der Barcode-Templates $\underline{\overline{\mathtt{B}}}$, welche ohne Beteiligung anderer Oligonukleotide eine freilaufende nicht-terminierte Kettenbildung zeigt (siehe Anhang A.6); zusätzlich die Klasse $\{\underline{\overline{\mathtt{S}}}, \underline{\overline{\mathtt{P}}}\}$, welche das Kettenwachstum an zwei Enden einer $\underline{\overline{\mathtt{B}}}$-Kette terminieren kann. Prinzipiell ist es möglich durch (experimentell aufwändig zu bestimmende) Reaktionsparameter die Kinetik der Ligation zu beschreiben, was im Rahmen der vorliegenden Arbeit nicht erfolgte. Bezüglich einer möglichen Formulierung der speziellen Ligationsreaktion wird auf Literatur zum Thema Kettenpolymerisation, z. B. [162], verwiesen. Anstelle einer dynamischen Beschreibung des Reaktionsverlaufes soll hier folgendes vereinfachtes stochastisches (statisches) Modell treten, das die Auswahl der Mengenverhältnisse der Oligonukleotide für die Erzeugung der Primer begründet. Unter Vernachlässigung von Reaktionszeit und physikalischer Limitierungen, lässt sich die Längenverteilung der x-mere (Polymere mit x Barcode-Templates) als geometrische Verteilung beschreiben: Sei $p_{\mathtt{B}}$ eine als konstant angenommene Auftrittswahrscheinlichkeit für ein $\underline{\overline{\mathtt{B}}}$ und $1 - p_{\mathtt{B}}$ die Wahrscheinlichkeit das Kettenwachstum durch ein $\underline{\overline{\mathtt{S}}}$ oder $\underline{\overline{\mathtt{P}}}$ zu unterbrechen, so ist

$$\Pr(L = x) = p_{\mathtt{B}}^{x}(1 - p_{\mathtt{B}})$$

die Wahrscheinlichkeit ein Molekül der Form $[\underline{\overline{\mathtt{S}}}|\underline{\overline{\mathtt{P}}}]\{\underline{\overline{\mathtt{B}}}\}^{x}[\underline{\overline{\mathtt{S}}}|\underline{\overline{\mathtt{P}}}]$, mit x Barcode-Templates $\underline{\overline{\mathtt{B}}}$, zu beobachten. Der Erwartungswert kann angegeben werden als $\mathbb{E}(L) = {}^{p_{\mathtt{B}}}/_{(1-p_{\mathtt{B}})}$. Sei $x : 1$ das Verhältnis der $\underline{\overline{\mathtt{B}}}$ zu terminierenden Sequenzen, so ist $p_{\mathtt{B}} = {}^{x}/_{(x+1)}$ und der Erwartungswert der Kettenlänge $\mathbb{E}(L) = x$. Des Weiteren ist die Beobachtung von nicht-terminierten Molekülen in diesem Modell nicht vorgesehen: deshalb muss ein zweites $\{\underline{\overline{\mathtt{S}}}, \underline{\overline{\mathtt{P}}}\}$-Molekül implizit für jede Kettenbildung vorhanden sein. Für die Parametrisierung der an der Ligation beteiligten Mengen erscheint daher ein Verhältnis von $\underline{\overline{\mathtt{S}}} : \underline{\overline{\mathtt{B}}} : \underline{\overline{\mathtt{P}}}$ im Bereich zwischen $1/2 : x : 1/2$ und $1 : x : 1$ für die molare Zusammensetzung plausibel, um die Anzahl an x-meren zu maximieren. Für die konkrete Ligation wurde ein Mengenverhältnis von $1 : 4 : 1$ für die Oligonukleotide gewählt.

Größenselektion

Den Abschluss der Erzeugung der RT-Primer bildet letztlich die Größenselektion auf Basis der Gelelektrophorese (vgl. Begriff 12). Nach chemischer Unterbrechung der im Verborgenen ablaufenden Ligationsreaktion kann die Längenverteilung der erzeugten einzelsträngigen Moleküle sichtbar gemacht werden. In Abb. 4.5 ist die Größenselektion im Gel zu sehen, wobei die bereits berechnete Zielgröße von 72nt ($\pm$ 2nt) für die Isolation der RT-Primer verwendet wurde. Durch die zuvor diskutierte Größenseparierung kann davon ausgegangen werden, dass die ausgewählten Moleküle mit hoher Wahrscheinlichkeit den entworfenen und funktionalen RT-Primern entsprechen. Diese Moleküle wurden nach Aufreinigung im modifizierten Sequenzier-Protokoll eingesetzt.

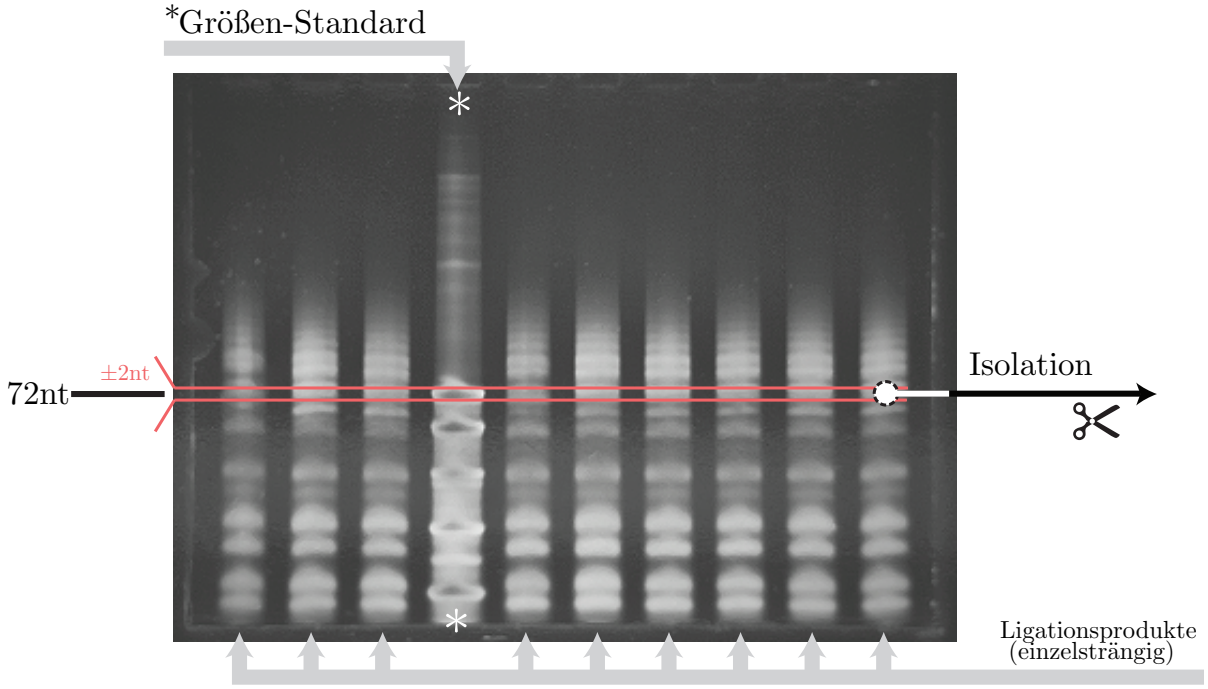

Abb. 4.5 Größenselektion der RT-Primer, vgl. [67, 94]: Ligationsprodukte (als einzelsträngige DNA) der Länge 72nt (± 2nt) werden durch die Gelelektrophorese (vgl. Begriff 12) isoliert. Die im Abschnitt zur Größenseparierung diskutierten Längen der Oligonukleotide führen zu den beobachtbaren strukturierten Bandenstrukturen.

4.2 Bioinformatik

Die Schnittstelle zwischen physikalischen Nukleotid-Ketten und einer digitalen Repräsentation der Molekülabfolgen ist die DNA-Sequenzierung. Mit der Wandlung einer physikalischen Polymer-Sequenz in lesbaren Text wird jedoch lediglich die Grundlage für den Informationsgewinn gelegt. Eine spezifische Aufbereitung der gewonnen Daten durch Konzepte der Bioinformatik ist für die Erzeugung von aussagekräftigen Ergebnissen unerlässlich.

Der vorliegende Teil umfasst Aspekte der Bioinformatik, welche speziell für die in Abschnitt 4.1 vorgestellten RT-Primer aus Barcode-Templates und deren konkrete Anwendung in der RNA-Sequenzierung vorgesehen sind. Die Anpassung des experimentellen Protokolls zur Verwendung der Zufallsbarcodes birgt sowohl Einschränkungen als auch zusätzliche Möglichkeiten, die durch eine adäquate Datenverarbeitung berücksichtigt werden: Zunächst wird die Illumina Sequenzierung vereinfacht als Schnittstelle zu formalen Sequenzen beschrieben und Besonderheiten des konkreten Anwendungsfalls dargelegt. Darauf basierend wird die Decodierung und die Filterung der Sequenzen auf valide Kombinationen des RT-Primers motiviert. Es wird eine schlichte und transparente Methode zur Korrektur von PCR-Duplikaten aufgezeigt und deren theoretische und praktische Grenzen diskutiert. Letztlich schließt der Abschnitt mit der Beschreibung eines alternativen Ansatzes zum Sequenzalignment und dem speziellen Fokus auf die Analyse der natürlichen 3'-Enden von RNA Fragmenten.

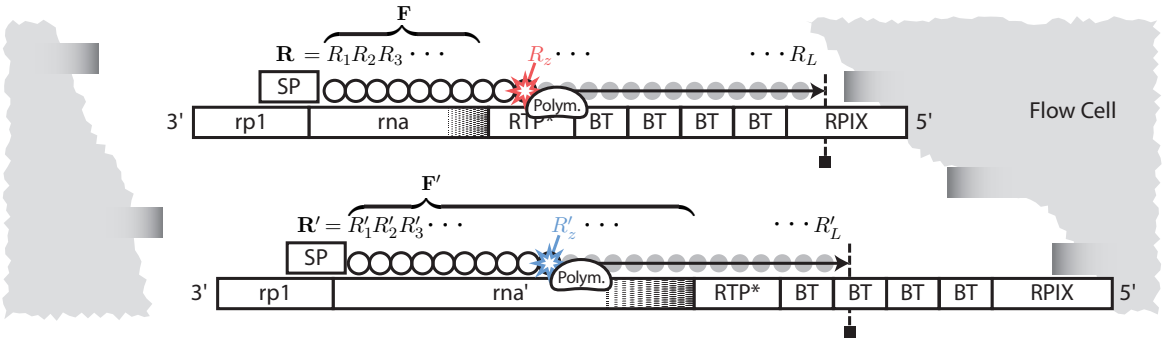

Abb. 4.6 Vereinfachte Darstellung der Illumina Sequenzierung bei Verwendung der modifizierten RT-Primer (aus Abb. 4.2): Sequenzierung durch Synthese zu doppelsträngigen DNA-Molekülen, initiiert durch Sequenzierprimer (SP) unter Einsatz speziell terminierter Nukleotide; Beispielhafte Erzeugung von Reads $\mathbf{R}$ und $\mathbf{R}'$ auf Basis von RNA-Inserts rna und rna' mit unterschiedlichen Längen (z bezeichnet den Sequenzierzyklus); Die Sequenzen $\mathbf{F}$ und $\mathbf{F}'$ entsprechen den Reads der RNA; Unvollständige Sequenzierung der Struktur des RT-Primers in $\mathbf{R}'$ auf Grund einer limitierenden Sequenzierlänge L. Die Sequenzierung relativ kurzer Inserts erhöht die Wahrscheinlichkeit der Beobachtung natürlicher 3'-Enden der RNA (schraffiert) und ermöglicht eine Analyse.

4.2.1 Besonderheiten der Sequenzierung

In Abb. 4.6 ist eine beispielhafte Sequenzierung illustriert, bei welcher die vorgestellten Zufallsbarcodes verwendet wurden (vgl. Abb. 4.2). Der eigentliche Vorgang der Sequenzierung (vgl. allgemeine Beschreibung in Begriff 14) findet an einer sogenannten *Flowcell* statt, einem Glasträger, der sowohl von Flüssigkeiten durchflossen, als auch durch ein bildgebendes Verfahren abgetastet werden kann. Eine weitere Besonderheit der Flowcell ist eine spezielle Oberfläche, die eine Vielzahl von Hybridisierungsregionen für Primer-Sequenzen bietet, wodurch eine räumlich gestreute Fixierung von einzelsträngigen DNA-Molekülen ermöglicht wird. Zwei stellvertretende Moleküle unterschiedlicher Länge sind in Abb. 4.6 zu sehen. Die eigentliche Sequenzierung besteht aus einer kontrollierten Elongation der vorliegenden Einzelstränge zu doppelsträngigen DNA-Molekülen: Initiiert durch einen Sequenzierprimer (SP) wird von einer DNA-Polymerase unter Einsatz speziell terminierter Nukleotide ein zyklischer stufenweiser Aufbau des komplementären Gegenstrangs erreicht. Der Einsatz unterschiedlich fluoreszierender Terminierungen bei den Nukleotiden ermöglicht durch entsprechende Bildgebung eine Klassifikation. Im Zyklus z wird somit parallel für alle sequenzierten Moleküle (relativ zum Sequenzierprimer) das z-te Nukleotid identifiziert und ein formales textuelles Symbol R_z erzeugt und gespeichert. Die Länge der digitalen Sequenzen (Reads) $\mathbf{R}$ und $\mathbf{R}'$ sind durch eine maximale Sequenzierlänge L limitiert. Die Reads können dabei messtechnisch bedingt von der tatsächlichen Abfolge der Moleküle abweichen. Zusätzlich zur Abfolge der Nukleotide bietet die Sequenzierung grundsätzlich noch Qualitätsinformationen über die Güte der Klassifikationen, welche als Wahrscheinlichkeitswert für eine Substitution interpretiert werden können. Diese Zusatzinformation ist hier nicht aufgeführt und wird zur Vereinfachung der Zusammenhänge in der gesamten Arbeit keine Verwendung finden. Des Weiteren soll zur Vereinfachung davon ausgegangen werden, dass ein sequenzierter Read nur in der dargestellten Leserichtung existiert, d. h. eine zweiseitige Sequenzierung soll nicht berücksichtigt werden. Für den Einsatz der modifizierten RT-Primer ergeben sich Einschränkungen als auch zusätzliche Möglichkeiten für die Analyse der Nukleotide:

- Von experimenteller Seite ist für eine effiziente Sequenzierung eine sinnvolle Größenselek-

tion der RNA-Inserts notwendig. Inserts sollten im Mittel gerade so lang sein, dass der RT-Primer vollständig im Read abgebildet ist (nicht zutreffend für $\mathbf{R}'$ in Abb. 4.6). Eine Standard-Sequenzierung sieht im Allgemeinen eine Insert-Länge vor, die größer ist als L, aber Sequenzierung über die Grenzen des Insert weitestgehend vermeidet.
- Von Seite der Bioinformatik ist eine robuste und effiziente Erkennung der RT-Primer gefragt, die auch ohne absolute Symbolpositionen eine sichere Klassifizierung von Barcode-Templates ermöglicht.
- Die beschriebene Notwendigkeit zur Selektion von kürzeren Inserts birgt zusätzlich Möglichkeiten zur Analyse von RNA im Hinblick auf deren natürliche 3'-Enden. Die Wahrscheinlichkeit systematische Modifikationen wie die Polyadenylierung (vgl. Begriff 7) zu beobachten wird damit drastisch erhöht.
- Neben einer zuverlässigen Identifikation des Inserts im Read ist zudem ein alternativer Ansatz zum Sequenzalignment notwendig, um RNA trotz möglicher Modifikation fehlertolerant einem Referenzgenom zuordnen zu können.

4.2.2 Decodierung, Filterung und Korrektur von PCR-Duplikaten

Eine klassische Decodierung im Sinne der Kanalkodierung ist für die in Abschnitt 4.1.2 erzeugten Zufallsbarcodes nicht gegeben. Zwei der Gesichtspunkte sollen im Folgenden als Motivation eines musterbasierten Ansatzes (vgl. Abschnitt 2.1.3) zur Erkennung und Klassifikation der RT-Primer dienen. Problematisch ist unter anderem:

- Die Codeworte $\mathcal{B}$ weisen im Allgemeinen, bis auf den minimalen Hamming-Abstand $d_{\min}$, keinerlei Struktur auf. Daher bleibt letztlich nur die Decodierung durch Distanzminimierung (vgl. Definition 23), um mittels $x|\mathcal{B}|$ Distanzberechnungen Blöcke der Länge x zu decodieren. Dieses Vorgehen setzt die Kenntnis der Blockgrenzen voraus, die jedoch durch die variable Länge der Inserts nicht gegeben sind.
- Die entworfenen zufälligen RT-Primer sind lediglich hypothetische Strukturen: Obwohl sie auf sorgfältigen molekularbiologischen Überlegungen beruhen, bleibt eine abschließende detaillierte Validierung unverzichtbar. Dazu sollten alle tatsächlich möglichen Oligonukleotide bei der Decodierung berücksichtigt werden.

Decodierung

Ein effizienter Ansatz, den genannten Punkten Rechnung zu tragen, ist in Alg. 4.1 dargestellt: Für die Eingangsgrößen Read $\mathbf{R}$, Barcode-Templates $\mathcal{B}$ mit genutzter Korrekturfähigkeit $t_{\mathcal{B}}$ und einer Menge von Sequenz-Distanz-Paaren $(\mathbf{X}, d_{\mathbf{X}})$ für Sequenzen $\mathbf{X} \in \{\mathbf{S}, \mathbf{p}, \mathbf{s}, \mathbf{P}\}$ des Primers wird als Ausgabegröße das Insert $\mathbf{F}$ und die decodierte Primer-Sequenz $\mathbf{C}$ ermittelt. Kernelement der Decodierung ist der Algorithmus Bitap (vgl. Konzept 8), der genutzt wird um im Read einzelnen Bereiche zu identifizieren, welche einen beschränkten Hamming-Abstand zu den bekannten Teilsequenzen aufweisen. Dabei ist das Vorgehen folgendes: In einer Menge $\mathcal{I}$ werden Tupel mit Positionen π_* abgelegt, die den Anfang, das Ende und die erkannten Teilsequenzen im Read definieren. Dazu werden in einem ersten Schritt alle möglichen Übereinstimmungen von Sequenzen $\mathbf{B}$ bzw. $\mathbf{X}$ innerhalb des Reads gesucht, wobei der Algorithmus `bitap.hamming` (vgl. Alg. A.4) k Substitutionen für die Rückgabe einer Übereinstimmung toleriert. In einem zweiten Schritt wird mittels `maxLength` anhand aller in $\mathcal{I}$ enthaltenen Grenzen der erste zusammenhän-

Alg. 4.1 Decodierung der RT-Primer

Input : $\mathbf{R}, \mathcal{B}, t_{\mathcal{B}}, \{(\mathbf{X}, d_{\mathbf{X}}) : \mathbf{X} \in \{\mathbf{S}, \mathbf{p}, \mathbf{s}, \mathbf{P}\}\}$
Output : $(\mathbf{F}, \mathbf{C})$

$\mathcal{I} \leftarrow \emptyset$
foreach $\mathbf{B} \in \mathcal{B}$ **do**
  $\mathcal{I} \leftarrow \mathcal{I} \cup \{(\pi_*^{\text{start}}, \pi_*^{\text{end}}, \mathbf{B})\} \leftarrow$ `bitap.hamming`$(\mathbf{B}, \mathbf{R}, k = t_{\mathcal{B}})$
foreach $(\mathbf{X}, d_{\mathbf{X}})$ **do**
  $\mathcal{I} \leftarrow \mathcal{I} \cup \{(\pi_*^{\text{start}}, \pi_*^{\text{end}}, \mathbf{X})\} \leftarrow$ `bitap.hamming`$(\mathbf{X}, \mathbf{R}, k = d_{\mathbf{X}})$

$[(\pi, \pi_1^{\text{end}}, \mathbf{Y}_1), (\pi_1^{\text{end}}, \pi_2^{\text{end}}, \mathbf{Y}_2)...(\pi_{l-1}^{\text{end}}, \pi_l, \mathbf{Y}_l)] \leftarrow$ `maxLength`$(\mathcal{I})$

$\mathbf{F} \leftarrow R_1 R_2..R_\pi$
$\mathbf{C} \leftarrow \mathbf{Y}_1 \mathbf{Y}_2..\mathbf{Y}_l$

gende Block maximaler Länge gesucht, welcher als Grundlage für das Decodierergebnis $\mathbf{C}$ und dessen Trennung vom Insert $\mathbf{F}$ dient.

Anzumerken ist, dass kein Vorwissen über die Struktur der RT-Primer für die Decodierung verwendet wird. Lediglich die Maximierung der längsten übereinstimmenden Sequenzen erscheint aus Sicht des Ligationsmodells von Abschnitt 4.1.2 als sinnvoll. Des Weiteren ist es über die Reduktion der genutzten Korrekturfähigkeit $t_{\mathcal{B}}$ (vgl. Definition 23) und der zulässigen Hamming-Abstände $d_{\mathbf{X}}$ möglich, die Sensitivität für die Erkennung der Oligonukleotide zu steuern. In der Anwendung wird $d_{\mathbf{X}} = 1$ gewählt und zur Ausnutzung der maximalen Fehlerkorrektur $t_{\mathcal{B}} = \lfloor (d_{\min}-1)/2 \rfloor = 2$ gesetzt, mit $d_{\min} = 5$ als minimaler Hamming-Abstand zwischen Barcode-Templates.

Filterung

Alle theoretisch möglichen Decodierergebnisse können mit der in Abschnitt 4.1.3 eingeführten Notation als $\{\mathtt{s}, \mathtt{S}, \mathtt{b}, \mathtt{B}, \mathtt{p}, \mathtt{P}\}^*$ beschrieben werden. Eine wohldefinierte Trennung der Inserts von synthetisch erzeugten Sequenzen der Oligonukleotide im Read beruht jedoch auf der Annahme der korrekten Hybridisierung der RT-Primer, nämlich ausschließlich an den dafür vorgesehenen Spacern. Für eine zuverlässige Erkennung des 3'-Endes des Inserts ist daher eine Beschränkung auf valide (dem Modell entsprechende) Sequenzen $\mathbf{C} \in \mathtt{S}\{\mathtt{B}, \mathtt{b}\}^x\mathtt{p}$ sinnvoll. Später folgende Abschnitte beschränken sich auf Muster mit $x \in \{1, 2, 3, 4, 5\}$ Barcode-Templates.

Zusätzlich zur Filterung der Primer $\mathbf{C}$ ist eine Einschränkung der Inserts $\mathbf{F}$ angebracht. Für die Korrektur der PCR-Duplikate und ein spezifisches Sequenzalignment ist ein Minimum an Symbolen nötig. Daher ist es sinnvoll, dass Inserts unter einer bestimmten Länge verworfen werden. In der späteren Anwendung werden Sequenzen mit weniger als 15nt nicht weiter berücksichtigt. Eine beispielhafte Menge von N sequenzierten Reads $\{\mathbf{R}_1, \mathbf{R}_2, .., \mathbf{R}_i, .., \mathbf{R}_N\}$ ergibt somit eine (mögliche) reduzierte Menge

$$\Phi = \Big\{(\mathbf{F}_1, \mathbf{C}_1), (\mathbf{F}_4, \mathbf{C}_4), (\mathbf{F}_6, \mathbf{C}_6), .., (\mathbf{F}_i, \mathbf{C}_i), .., (\mathbf{F}_N, \mathbf{C}_N)\Big\} \tag{4.5}$$

an gefilterten Tupeln von Inserts $\mathbf{F}_i$ und decodierten Sequenzen $\mathbf{C}_i$.

Korrektur von PCR-Duplikaten

Gestützt auf die Gleichheit von jeweils zwei Elementen der Tupel aus dem vorangehenden Abschnitt lassen sich die Abbildungen

$$\phi_{\mathbf{F}} : \mathcal{S} \mapsto \mathcal{N} \quad \text{und} \quad \phi_{\mathbf{C}} : \mathcal{S} \mapsto \mathcal{N} \tag{4.6}$$

formulieren, die aus der Menge der Sequenzen $\mathcal{S}$ in die Menge $\mathcal{N}$ abbilden. Die Menge $\mathcal{N}$ setzt sich dabei wiederum aus Teilmengen von ganzen Zahlen $\{1, 2, .., N\}$ zusammen, die Gruppen von Reads aus Φ referenzieren. Dabei liefert $\phi_{\mathbf{F}}(\mathbf{F}_i)$ die Indexe $*$ aller identisch decodierten Sequenzen $\mathbf{C}_*$, die für eine Sequenz $\mathbf{F}_i$ in Φ enthalten sind. Umgekehrt ergibt $\phi_{\mathbf{C}}(\mathbf{C}_i)$ die Indexe $*$ der Inserts $\mathbf{F}_*$, welche die Sequenzen $\mathbf{C}_i$ im Tupel tragen. Zum Beispiel bedeutete das für $\mathbf{F}_1 = \mathbf{F}_4 = \mathbf{F}_6$ und $\mathbf{C}_1 = \mathbf{C}_4 \neq \mathbf{C}_6$ unter anderen Tupeln in (4.5)

$$\Big\{\{1, 4\}, \{6\}\Big\} \subseteq \phi_{\mathbf{F}}(\mathbf{F}_1) \quad \text{und} \quad \Big\{\{1, 4\}\Big\} \subseteq \phi_{\mathbf{C}}(\mathbf{C}_1).$$

Die enthaltenen Teilmengen trennen hierbei Gruppen von PCR-Duplikaten. Für ein Insert $\mathbf{F}_i$ lassen sich ungeachtet der decodierten RT-Primer die zwei Zählgrößen

$$\check{n}(\mathbf{F}_i) = |\{(\mathbf{F}_j, \mathbf{C}_j) \in \Phi : \mathbf{F}_j = \mathbf{F}_i\}| \quad \text{und} \quad \hat{n}(\mathbf{F}_i) = 1 \tag{4.7}$$

angeben. Dabei ist $\check{n}$ die Anzahl aller identischer Reads mit $\mathbf{F}_i$ und damit eine obere Schranke für die Auftrittshäufigkeit eines Inserts vor der Erstellung der Sequenzier-Library. Wohingegen $\hat{n}$ der konservativen Betrachtung zugrunde liegt, die jedes Duplikat eines Inserts als PCR-Kopie klassifiziert und damit eine untere Schranke darstellt. Mit Integration der Zusatzinformation der decodierten Sequenzen (Zufallsbarcodes) lässt sich

$$\tilde{n}(\mathbf{F}_i) = |\phi_{\mathbf{F}}(\mathbf{F}_i)| \tag{4.8}$$

als korrigierte Anzahl von initialen Inserts angeben. Der Schätzwert $\tilde{n}$ entspricht der in $\phi_{\mathbf{F}}$ enthaltenden Gruppen von identischen Inserts mit gleichem Zufallsbarcode. Ähnlich dazu kann die korrigierte Auftrittshäufigkeit für decodierte Sequenzen durch $\tilde{n}(\mathbf{C}_i) = |\phi_{\mathbf{C}}(\mathbf{C}_i)|$ definiert werden. Im zuvor gegebenen Beispiel würde einer der Reads $\mathbf{R}_1$ oder $\mathbf{R}_4$ als Duplikat erkannt werden. Ein Rückschluss darüber, welcher der zwei Reads letztendlich ein PCR-Duplikat darstellt, kann dabei nicht getroffen werden (denn sie sind ja identisch). Auch werden mögliche Seiteninformationen (wie Qualitätsinformationen), welche die Reads weiter charakterisieren könnten, bei dieser Betrachtung verworfen.

4.2.3 Theoretische Grenzen der PCR-Korrektur

Die Qualität des beschriebenen Verfahrens zur Korrektur von PCR-Duplikaten wird im Wesentlichen bestimmt von zwei Größen: Erstens der Diversität der RT-Primer und zweitens den Sequenzfehlern in den Reads. Ausgehend von der Annahme eines idealisierten fehlerfreien Sequenzierexperiments wird in diesem Abschnitt zunächst der theoretische Einfluss der Diversität auf die Korrektur näher untersucht. Darauf folgend wird die Präsenz von Sequenzfehlern und deren Implikationen für die Erkennung von PCR-Duplikaten diskutiert.

Diversität der RT-Primer

Eine optimale Korrektur von Duplikaten wäre möglich wenn, erstens die Anzahl unterschiedlicher RT-Primer bei der Herstellung der Sequenzier-Library größer ist als die Zahl identischer Inserts, und zweitens jeder RT-Primer nur einmal zur Hybridisierung verfügbar wäre. Unabhängig vom Verlauf einer PCR würde das den Fall ausschließen, dass zwei identische (ursprünglich vorliegende) Moleküle als Duplikat klassifiziert und das Ausmaß der PCR überschätzt wird. Diese Veranschaulichung soll als Motivation der Diversität von Molekülen und den damit gegebenen Grenzen dienen.

Für einen theoretischen Zugang zur Überschätzung der PCR durch die Korrektur soll ein stochastisches Modell dienen, welches den Ansatz aus [56] für den nicht-uniformen Fall verallgemeinert. Dazu wird die Problematik auf folgendes (von der PCR unabhängige) Modell reduziert: Betrachtet wird lediglich ein Typ von Insert und eine Menge von n identischen Repräsentationen dieses Typs (dieser Sequenz). Dabei ist n die unbekannte Zielgröße der durch die Korrektur zu schätzenden Auftrittshäufigkeit. Des Weiteren existieren m unterscheidbare Typen (Sequenzen) von RT-Primern, welche zur einfacheren Repräsentation als Menge $\{1, 2, 3, .., i, .., m\}$ von ganzzahligen Indexen i dargestellt sind. Die Annahme eines unerschöpflichen Vorrats an Indexen eines Typs rechtfertigt eine Beschreibung mittels konstanten Auftrittswahrscheinlichkeiten $p_1, p_2, .., p_i, .., p_m$. Im Modell der Herstellung der Sequenzier-Library beschreibt p_i die Wahrscheinlichkeit, dass ein konkretes Insert mit einem RT-Primer des Indexes i hybridisiert. Die Annahme, dass dies ausnahmslos für alle n Inserts geschieht, erlaubt die Erklärung der Auftrittshäufigkeiten der RT-Primer mittels Multinomialverteilung (vgl. Definition 6). Beschreibt die Zufallsvariable A_i die Anzahl der Inserts mit Index i so gilt letztlich $n = \sum_i A_i$. Auf Basis einer Indikatorfunktion I (als $I(0) = 0$ bzw. $I(A) = 1$ für $A \in \mathbb{N}$) lässt sich mittels Zufallsvariable

$$\tilde{N} = \sum_{i=1}^{m} I(A_i) = \sum_{i=1}^{m} I(\lfloor \alpha_i A_i \rfloor) \quad \text{, mit } 1 \leq \alpha_i \text{ für } \alpha_i \in \mathbb{R} \tag{4.9}$$

eine theoretische stochastische Analogie zur Bestimmung der Anzahl an Fragmenten in (4.8) beschreiben. Dabei ist der Wert von $\tilde{N}$ invariant gegenüber diskreten Skalierungen[6] der Zufallsvariablen A_i. Die ungleichförmige Amplifikation der PCR (vgl. Begriff 18) lässt sich für einen Typ von Insert durch die Faktoren α_i beschreiben. Dennoch unterschätzt die korrigierte Zufallsvariable $\tilde{N}$ im Allgemeinen die ursprüngliche Auftrittshäufigkeit n in Abhängigkeit der Wahrscheinlichkeiten p_i. Mit $I_i = I(A_i)$ als Zufallsgröße der Indikatorfunktion lässt sich der Erwartungswert und die Varianz der stochastischen Zählgröße $\tilde{N}$ als

$$\mathrm{E}(\tilde{N}) = \sum_{i=1}^{m} 1 - (1 - p_i)^n \tag{4.10}$$

und

$$\mathrm{Var}(\tilde{N}) = \sum_{i=1}^{m} \mathrm{Var}(I_i) + \sum_{i \neq j} \mathrm{Cov}(I_i, I_j) \tag{4.11}$$

angeben, mit $\mathrm{Var}(I_i) = (1 - p_i)^n[1 - (1 - p_i)^n]$ und den Kovarianzen

$$\mathrm{Cov}(I_i, I_j) = (1 - p_i - p_j)^n - (1 - p_i)^n(1 - p_j)^n$$

[6] Skalierung ist definiert als $\lfloor \alpha_i A_i \rfloor = \max\{k \in \mathbb{Z} : k \leq \alpha_i A_i\}$.

der Größen I_i und I_j (vgl. Anhang A.8). Des Weiteren kann gezeigt werden, dass für eine Gleichverteilung der RT-Primer, $p_i = 1/m$, der Erwartungswert der Zählgröße $\tilde{N}$ maximiert wird. Die Gleichverteilung von Zufallsbarcodes ist die grundlegende Annahme in den Berechnungen und den Abschätzungen von Zählgrößen in [56]. Die hier dargelegte Theorie entspricht der realistischeren Verallgemeinerung für den nicht-uniformen Fall.

Die hergeleiteten wahrscheinlichkeitstheoretischen Größen beziehen sich auf den idealisierten Vorgang der Hybridisierung von RNA Molekülen und randomisierten RT-Primern. Wesentliche Größen des verwendeten Modells reduzieren sich auf die Zahl n an identischen Inserts und die Anzahl m, respektive die Verteilung $\{p_i\}$, der Zufallsbarcodes. Anzumerken ist dazu Folgendes:

- In der Realität ist davon auszugehen, dass neben der Auftrittswahrscheinlichkeit von Molekülen die chemische Struktur wesentlich an der Ausbildung von Bindungen beteiligt ist. Daher sind Annahmen der Gleichverteilung in der Verwendung von Zufallsbarcodes, wie sie unter anderem in [56, 77, 148, 153] getroffen werden, kritisch zu hinterfragen. Das hier verallgemeinerte nicht-uniforme Modell wird später (in 4.3.2) angewendet, um die Diversität der experimentell eingesetzten RT-Primer zu evaluieren.
- Eine weitere Diskrepanz des Modells zu den Beobachtungen aus Sequenzierdaten liegt in der Geschlossenheit. Sowohl eine vollständige Hybridisierung aller RNA-Moleküle mit Primern, als auch die umfassende Abbildung aller Moleküle in Reads ist experimentell nicht möglich. Die Annahme eines uniformen Zufallsprozesses für die Auswahl der zu beobachtenden Reads aus einer sehr großen Menge von Molekülen ist daher essentiell.
- Die theoretische Betrachtung nutzt ein fixes (beständiges) Repertoire an Typen von Molekülen: Unveränderliche Zufallsereignisse sind stellvertretende Abstraktionen für Moleküle und deren Verknüpfungen, welche letztlich als unabhängig von einer Sequenzeigenschaft angenommen werden. Tatsächlich sind Sequenzen durch das Experiment veränderlich, wodurch eine Korrektur der PCR-Duplikate von Sequenzfehlern und deren Erkennung negativ beeinflusst wird.

Eine ausführlichere Betrachtung wird dem letzten Punkt der Aufzählung gegeben, wie im Folgenden dargelegt wird.

Sequenzfehler im Read

Im Gegensatz zur Diversität von Zufallsbarcodes die, wie in idealisierter Form gezeigt wurde, generell für die Überschätzung der PCR-Duplikate verantwortlich ist, führen Sequenzfehler in Reads tendenziell zur einer Unterschätzung. Bezug nehmend auf die einführende Darstellung von Abb. 2.12 auf Seite 42 wird die komplexe Problematik anhand einer erweiterten Grafik in Abb. 4.7 näher erläutert. Anstatt einer Position in der genomischen Quelle $\mathbf{G}$ sollen nun zwei unterschiedliche Abschnitte A und B berücksichtigt werden, die eine hohe Ähnlichkeit aufweisen. So ist es biologisch möglich, dass sich zwei Fragmente $\mathbf{F}^{(A)}$ und $\mathbf{F}^{(B)}$ von unterschiedlichen Positionen im Genom in nur einem Symbol unterscheiden, d. h. die Editierdistanz (vgl. Definition 25) ist $\mathrm{d_e}(\mathbf{F}^{(A)}, \mathbf{F}^{(B)}) = 1$. Eine beispielhafte Sequenzierung könnte sich wie folgt darstellen: Zwei Inserts von Referenz A sind in der Sequenzier-Library mit RT-Primer $\mathbf{C}_1$ und $\mathbf{C}_2$ versehen worden (was eine PCR-Korrektur ermöglicht), ein weiteres Insert von Referenz B ist jedoch ebenfalls mit der Sequenz $\mathbf{C}_2$ versehen worden. Die drei Fragmente durchlaufen eine beispielhafte (fehlerbehaftete) Sequenzier-Library-Herstellung und werden letztendlich sequenziert und durch die Decodierung klassifiziert. Prinzipiell müssen bezüglich ihrer Position im Read zwei

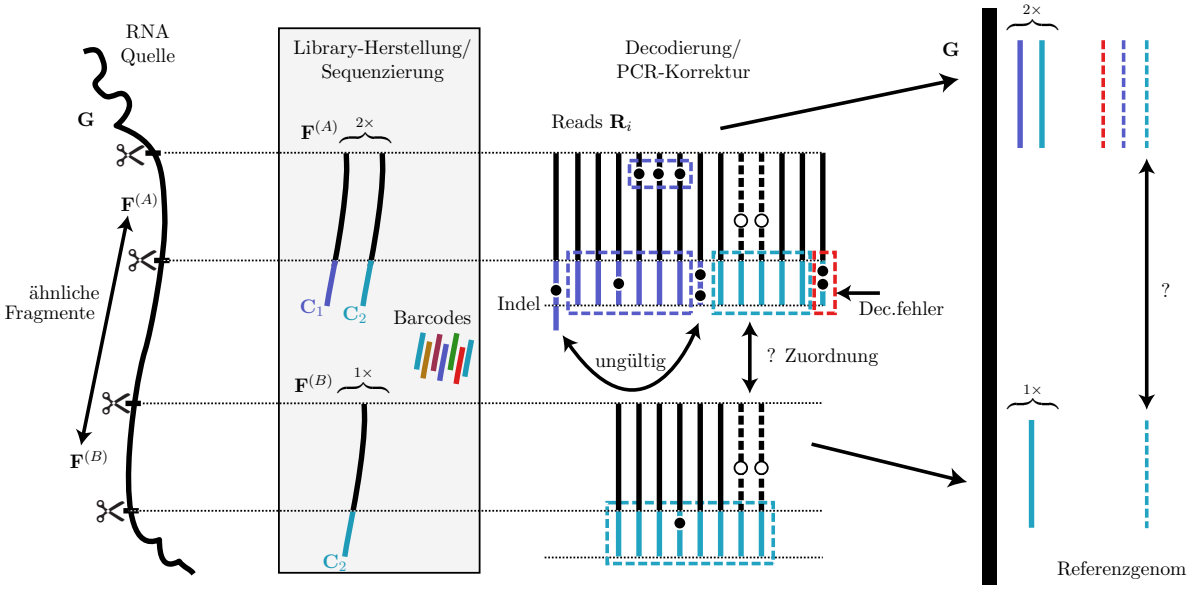

Abb. 4.7 Fehler in der Sequenzierung und deren Auswirkungen auf die PCR-Korrektur (Symbolik wie in Abb. 2.12): Ausgehend von zwei Sequenzen $\mathbf{F}^{(A)}$ und $\mathbf{F}^{(B)}$ von unterschiedlichen Positionen im Genom $\mathbf{G}$ werden drei Inserts durch Zufallsbarcodes $\mathbf{C}_1$ und $\mathbf{C}_2$ markiert, durch die PCR vervielfältigt und sequenziert. Hinsichtlich der Decodierung und Korrektur von PCR-Duplikaten ergeben sich unterschiedliche Probleme in der Klassifikation des Ursprungs und der initialen Auftrittshäufigkeit der Inserts. Der niedrige Abstand von $\mathbf{F}^{(A)}$ und $\mathbf{F}^{(B)}$ und die Anheftung desselben Barcodes $\mathbf{C}_2$ verhindert eine robuste Zuordnung oder Zählung im Fehlerfall. Fehler sind durch Kreise annotiert. Fehler in Inserts führen generell zu einer Unterschätzung des PCR. Fehler in Zufallsbarcodes führen durch die Korrekturfähigkeit nur bedingt zur Unterschätzung. Durch die Blockstruktur der Barcode-Templates können selbst Einfügungen oder Löschungen von Symbolen (Indels) als ungültige Sequenzen identifiziert werden.

Kategorien von Fehlern unterschieden werden. Fehler in der Region des Zufallsbarcodes und solche, die sich im Bereich des RNA-Fragments befinden.

Für Fehler in Zufallsbarcodes gilt: Auf Basis des minimalen Hamming-Abstands für Barcode-Templates können Symbolersetzungen in den Primer-Sequenzen in gewissem Umfang korrigiert werden (beispielsweise ein Fehler in Abb. 4.7). Auf Grund der Restriktionen bei der Filterung bezüglich der Blockstruktur ist sogar davon auszugehen, dass eine fehlerhafte Klassifikation durch Einfügungen oder Löschungen von Symbolen in gewissem Maße zu vernachlässigen ist. Übersteigt die Anzahl der Fehler jedoch die Korrekturfähigkeit kann das zu einem Decodierfehler und damit einer vergrößerten Anzahl an geschätzten Fragmenten führen. Der Umfang der PCR wird also unterschätzt.

Weitaus komplexer hingegen ist die Unterschätzung, die durch Fehler im Insert verursacht wird. Für Inserts kann im Allgemeinen keine minimale Distanz formuliert werden auf deren Basis eine Trennung von Sequenzen möglich ist. Wird eine strikte (vom Sequenzalignment unabhängige) Korrektur der PCR-Duplikate auf Basis von (4.8) vorgenommen, bedeutet das für jeden zusätzlichen Fehler im Insert eine Vergrößerung der geschätzten Anzahl an RNA-Molekülen oder sogar eine fehlerhafte Zuordnung zu einem anderen genomischen Ursprung. Es existieren einzelne Ansätze, die eine Korrektur von PCR-Duplikaten nach dem Sequenzalignment vorschlagen [148].

Alg. 4.2 Alternativer Alignment-Algorithmus

Input : $\mathbf{F}, \mathbf{G}, k_{\max}$
Output : $\mathcal{L}$

$\mathcal{I} \leftarrow \{(\pi_*^{\text{end}}, k_{\min})\} \leftarrow$ `bitap.lasso`$(\mathbf{F}, \mathbf{G}, k = k_{\max})$

$\mathcal{L} \leftarrow \emptyset$
foreach $(\pi_*^{\text{end}}, k_{\min}) \in \mathcal{I}$ **do**
  $\mathcal{L} \leftarrow \mathcal{L} \cup \{(i_1, i_2, j, F_i, G_j, idx)\} \leftarrow$ `align`$(\mathbf{F}, \mathbf{G}, \pi_*^{\text{end}}, k = k_{\min})$

Ob diese Verfahren jedoch einen zielführenden Ansatz für das grundlegende Problem der mangelnden Separierung der Ausgangssequenzen liefert bleibt dabei fragwürdig. Aus Gründen der Transparenz und der Nachvollziehbarkeit hinsichtlich der Verarbeitung von Inserts mit Fehlern wird in dieser Arbeit eine unabhängige Korrektur von Duplikaten verwendet.

4.2.4 Alternatives Sequenzalignment

Der folgende Abschnitt beschreibt eine alternative Methode des Sequenzalignments mit Fokus auf der Analyse von systematischen Modifikationen an den 3'-Enden der RNA. Standardmethoden zum Mapping (vgl. Begriff 17) nutzen zumeist Heuristiken zur Kürzung der Inserts oder ein partielles (lokales) Alignment, um die Qualität der Übereinstimmung mit der Referenzsequenz zu verbessern. Somit wird die Menge an Diskrepanzen zum Genom verringert und die Anzahl an erfolgreich zugeordneten Inserts vergrößert. Eine Veränderung der Inserts würde jedoch zum einen die Korrektur von PCR-Duplikaten verfälschen und andererseits die Beobachtung der systematischen Effekte herabsetzen.

Die grundlegende Idee des alternativen Ansatzes (dargestellt in Alg. 4.2) besteht in einer Aufwandsreduzierung eines klassischen semi-globalen Alignments (vgl. Definition 26) zur Berechnung aller optimalen Alignments bezüglich der Editierdistanz und einer reduzierten Repräsentation in einem sogenannten *Alignment-Lattice* $\mathcal{L}$. Dies geschieht in zwei Schritten: Zuerst wird eine modifizierte Version des Algorithmus Bitap (vgl. Konzept 8) genutzt, um die End-Positionen $\{\pi_*^{\text{end}}\}$ aller optimaler Alignments sehr effizient zu ermitteln. Danach wird ein klassisches dynamisches Programm genutzt, welches die exakte Alignierung von Symbolen auf Nukleotid-Ebene bestimmt. Die Erweiterung des Algorithmus `bitap.lasso` (vgl. Alg. A.5) ist folgende Aufwandsreduktion: Ausgehend von einer maximalen Anzahl $k = k_{\max}$ an zulässigen Fehlern wird k zur Laufzeit der Suche für jedes gefundene Alignment mit $k_{\min}$ ($< k$) Fehlern angepasst zu $k = k_{\min}$. Gleichzeitig wird die Rückgabe auf Positionen $\{\pi_*^{\text{end}}\}_{k_{\min}}$ mit exakt $k_{\min}$ Fehlern beschränkt. In Analogie zu einem Lasso-Seil wird die Ergebnismenge sukzessive zusammengezogen, sodass mit zusätzlicher Reduktion des Aufwands nur die Positionen bestehen bleiben, an welchen Übereinstimmungen von $\mathbf{F}$ in $\mathbf{G}$ mit maximal $k_{\min}$ Fehlern vorliegen. Ausgehend von den Endpositionen $\{\pi_*^{\text{end}}\}_{k_{\min}}$ besteht der zweite Schritt in einem herkömmlichen Sequenzalignment. Diese hier nicht formal beschriebene Methode namens `align` entspricht im Wesentlichen dem Needleman-Wunsch-Algorithmus (vgl. Konzept 7) mit vertauschter Leserichtung der Sequenzen $\mathbf{F}$ und $\mathbf{G}$, weil der Algorithmus `bitap.lasso` lediglich die letzten Symbole $G_{\pi_*^{\text{end}}}$ aller Übereinstimmungen ermittelt. Die Kosten des Needleman-Wunsch-Algorithmus sind so angepasst, dass die Kostenberechnungen des `bitap` Algorithmus exakt anhand von Sequenz-

operationen nachvollziehbar werden. Die Ausgabe der `align`-Methode besteht jedoch nicht aus einer Liste der Sequenzoperationen, sondern aus einer Menge von Tupeln, die Alignment-Lattice genannt wird. In diesen Tupeln beschreibt j die Position bezügliche des Genoms, $i_1 (= |\mathbf{F}| - i_2)$ und i_2 die Position relativ zum Anfang und zum Ende des Inserts $\mathbf{F}$, mit $|\mathbf{F}|$ als dessen Länge. Da es möglich ist mehrere gleichwertige Alignments für eine Sequenz $\mathbf{F}$ zu finden wird jedes Element im Alignment-Lattice mit einem Index idx versehen, der eine Auflösung von Mehrdeutigkeiten ermöglicht.

Die Beschränkung des Sequenzalignment auf schlichte Zuordnungen $F_i \leftrightarrow G_j$ von Symbolen ist mit einem Informationsverlust verbunden, der weitere Klärung bedarf: Generell ist es auf Basis der erwähnten Tupel nicht möglich Rückschlüsse über enthaltene Einfügungen oder Löschungen zu gewinnen. Es werden lediglich Positionen gespeichert, an welchen ein Symbol des Inserts mit einem Symbol des Genoms übereinstimmt oder nicht. Es findet eine Reduktion auf Gleichheit von Symbolen statt. Zu der vollständigen Repräsentation des exakten Verlaufs eines Sequenzalignment ist generell eine Erweiterung des Symbolvorrats nötig. Verwendet man beispielsweise $*$ als symbolischen Platzhalter (vgl. Notation Abschnitt 2.1), so könnte man eine Einfügung durch $G_j = *$ darstellen, respektive $F_i = *$ für eine Löschung verwenden. Eine derartige Differenzierung soll hier nicht getroffen werden. Obwohl die ignorierten Operationen Bestandteil des Optimierungsproblems Alignment sind (Definition 26), wird deren Auftreten für die Auswertung des hier vorgestellten Ansatzes nur mittelbar berücksichtigt. Anders als Ersetzungen haben Einfügung und Löschungen nicht nur an einem einzigen Tupel mit $F_i \neq G_j$ Anteil, sondern beeinflussen unter Umständen noch ein weiteres Tupel im Alignment-Lattice. Auf Basis der im Folgenden erklärten Projektionen findet bezüglich dieser Mehrdeutigkeit eine Randomisierung statt, sodass Einfügung und Löschungen im Mittel zu weniger Ereignissen $F_i \neq G_j$ führen als native Ersetzungen.

Prinzipiell enthält $\mathcal{L}$ eine Menge von Zuordnungen von Symbolen F_i (der Position i im Insert) zu Symbolen G_j (der Position j im Genom). Es existieren nun zwei unterschiedliche Projektionen für die Tupel aus $\mathcal{L}$, welche zur statistischen Auswertung genutzt werden:

- Sortiert man die Elemente in $\mathcal{L}$ nach dem Index j der Sequenz $\mathbf{G}$, so erhält man für jede Position im Genom eine Menge von Symbolen, die stellvertretend für die Überdeckung mit Inserts verstanden werden kann. Da $\mathcal{L}$ alle optimalen Sequenzalignments enthält, ist es möglich, dass für eine bestimmte Position j mehrere Tupel des gleichen Inserts vorliegen, d. h. eine Mehrdeutigkeit im Alignment existiert. Gleiches gilt für jede Einfügung bezüglich der Referenz. In diesen Fällen werden alle Mehrdeutigkeiten bis auf ein zufälliges Tupel verworfen. Dieses Vorgehen ermöglicht eine Zählung von Inserts auf der Ebene von einzelnen Nukleotiden und stellt sicher, dass Inserts zu jeder passenden Position im Genom zugerechnet werden, die Anzahl an Überdeckungen jedoch nicht durch Mehrdeutigkeit vergrößert wird. Für Einfügungen bedeutet dieses, dass sie im Mittel mit einer Rate kleiner eins eine Zuordnung $F_i \neq G_j$ zur Folge haben. Die beschriebene Art der Projektion ist schematisch in Abb. 4.8 illustriert.
- Sortiert man Tupel aus $\mathcal{L}$ nach dem Index i (also i_1 oder i_2) relativ zum Anfang oder dem Ende der Sequenzen $\mathbf{F}$, so erhält man für jede Position i im Allgemeinen Insert eine Menge von Tupel, die für eine Zuordnung zum Genom stehen. Für $i_1, i_2 \leq 15$ (vgl. Filterung auf 15nt in 4.2.2) ist die Anzahl an Beobachtungen (Tupel) für jede Position i identisch. Dennoch können wie zuvor an bestimmten Positionen Mehrdeutigkeiten auftreten: Es ist möglich, dass bestimmte Abschnitte im Genom mehrfach vorliegen, was zur Folge hat,

dass sich Zuordnungen natürlicherweise wiederholen (siehe 2 in Abb. 4.8). Gleiches gilt für Löschungen bei welchen für eine Position i im Insert mindestens zwei unterschiedliche Tupel vorliegen können. Damit die erwähnten Mehrdeutigkeiten lediglich einen mittleren Einfluss auf statistische Aussagen haben, wird die Auswahl von Tupeln ebenfalls randomisiert. Analog zu den Einfügungen ist es für Löschungen bezüglich der Referenz möglich, dass nicht jede Löschung durch eine Zuordnung $F_i \neq G_j$ in der Statistik repräsentiert ist.

Während die erstgenannte Projektion für quantitative Aussagen der Überdeckung einer Position j im Genom mit alignierten Inserts bestimmt ist (später genutzt in Abschnitt 4.3.3), bei welcher Symbolunterschiede keinen Einfluss auf Zählgrößen haben, ist die zweite Projektion für die Analyse von systematischen Modifikationen relativ zur Position i der Inserts gedacht (vgl. 4.3.4). Dabei macht die Art der beobachteten Sequenzvariationen prinzipiell einen Unterschied in der Bewertung der Rate von Modifikationen. Beschäftigt man sich mit anderen Publikationen, die basierend auf einem Sequenzalignment statistische Auswertungen zu Fehlerwahrscheinlichkeiten machen, so existieren verschiedenste Ansätze: Einfügungen und Löschungen werden beispielsweise nicht berücksichtigt oder nicht explizit ausgewertet [44, 113, 121, 148] oder lediglich durch mittlere Auftrittshäufigkeiten [47, 89, 117] charakterisiert. Einige Arbeiten [93, 139, 147] beziehen die Analyse von beobachteten Modifikationen zwar auf Positionen im Read oder Genom, jedoch lässt sich kein gemeinsamer Standard erkennen, wie diese Abbildung korrekt vorzunehmen ist. Allen genannten Ansätzen gemeinsam ist, dass für Berechnungen nicht alle alternativen Alignments explizit berücksichtigt werden. Der hier vorgestellte Ansatz auf Basis der Editierdistanz soll dazu eine pragmatische Alternative darstellen, die un-

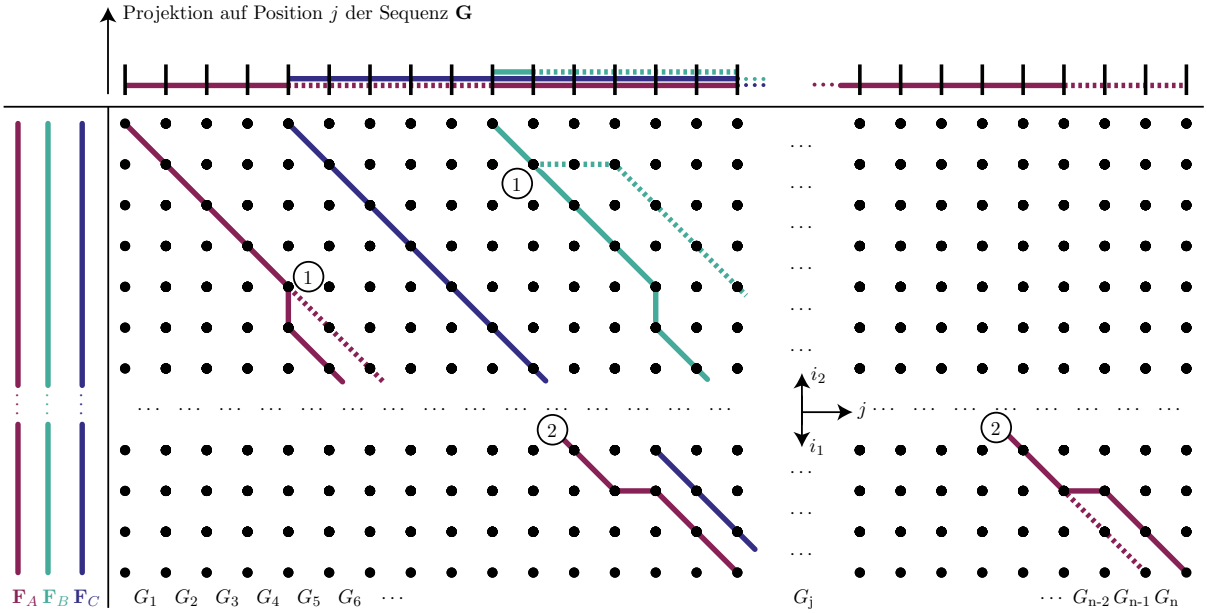

Abb. 4.8 Alignment-Lattice (Projektion auf j): Beispielhafte Zuordnung von drei Inserts $\mathbf{F}_A$,$\mathbf{F}_B$ und $\mathbf{F}_C$ auf eine Referenzsequenz $\mathbf{G}$. Eine diagonale Verbindung zwischen Gitterpunkten entspricht einem Tupel des Lattice $\mathcal{L}$. Es existieren u. U. alternative Sequenzalignments an einer bestimmten Position j (1), oder an unabhängigen Positionen (2). Für die gezeigte Projektion werden Einfügungen bezüglich der Referenz (vertikale Abschnitte) ignoriert und für mehrdeutige horizontale Abschnitte nur ein zufälliges Tupel für die Projektion übernommen. Dadurch wird die Anzahl der Diskrepanzen $F_i \neq G_j$ im gezeigten Ansatz im Mittel reduziert.

ter Berücksichtigung aller möglichen Sequenzalignments eine klare Zuordnung auf Symbolebene liefern.

4.3 Validierung und Analysen

Kompatibilität und Konformität mit den experimentellen Standards des TruSeq Small RNA Protokolls sind Hauptkriterien für den in Abschnitt 4.1 beschriebenen Entwurf der Barcode-Templates und den daraus erzeugten RT-Primern. Für eine Validierung der Zufallsbarcodes, deren Motivation die Korrektur von PCR-Duplikaten darstellt, erscheint es jedoch durchaus sinnvoll die Sequenzier-Standards hinsichtlich einer Größe zu variieren: der Anzahl an PCR Zyklen. Dieser Abschnitt umfasst die Evaluierung dreier technischer Replikate einer RNA-Sequenzierung unter Anwendung der neuen RT-Primer, bei welcher lediglich die Anzahl der PCR Zyklen variiert wurde. Ausgehend vom Standard [79] von 11 Zyklen wurde die Anzahl auf 15 und 25 erhöht, alle übrigen Parameter wurden bei der Herstellung der Sequenzier-Library jedoch konstant gehalten.

Beginnend mit einer Analyse aller decodierten Strukturen der RT-Primer in Abschnitt 4.3.1 und einer Diskussion von Diskrepanzen, die sich gegenüber der erwarteten Kombinationen zeigen, beschränken sich die darauffolgenden Teile auf wohldefinierte Primer. Basierend auf einer Schätzung der Auftrittswahrscheinlichkeiten der experimentell beobachtbaren Zufallsbarcodes und der in 4.2.3 gezeigten Theorie zur Diversität, werden zunächst in 4.3.2 die Grenzen der PCR-Korrektur für die vorliegenden Sequenzierungen abgeschätzt. Die konkrete Gegenüberstellung von Zählgrößen der Inserts in 4.3.3 offenbart eine Anomalie in der Anzahl der PCR-Duplikate, welche besondere Klärung bedarf. Ein offenbar nichtmonotones Verhalten zwischen beobachtbaren PCR-Duplikaten und der Anzahl der PCR Zyklen, legt die Hypothese der Selbst-Hybridisierung nahe, die daraufhin diskutiert wird. Der abschließende Abschnitt befasst sich mit einer beispielhaften Analyse, die durch die Verwendung der RT-Primer ermöglicht wurde: In Abschnitt 4.3.4 fällt der Fokus auf RNA Fragmente und deren 3'-Enden. Basierend auf der in 4.2.1 beschriebenen verstärkten Sequenzierung der nativen Enden der Inserts wird eine statistische Auswertung dargelegt, um Aufschluss über Art und Lokalisation der Modifikationen der RNA zu erhalten.

Details der Sequenzierungen

Die für die Sequenzierung genutzte RNA stammt von dem bakteriellen Modelorganismus *Escherichia coli*, genauer dem Bakterienstamm O157:H7 EDL 933, dessen Genom erstmalig im Jahre 2001 sequenziert [127] und später vervollständigt publiziert [97] wurde. Es umfasst in der finalen Version (Accession Number: CP008957.1) 5547323nt und enthält mehr als 5700 annotierte Gene. Zusätzlich befindet sich im Organismus das Plasmid pO157 mit 92-104 $\cdot 10^5$nt Länge und ca. 100 annotierten Genen, welches für die Analysen jedoch nicht berücksichtigt wurde. Eine detaillierte Charakterisierung des Organismus findet sich in [103].

Zur Gewinnung der RNA wurde das Bakterium unter Standardbedingungen in LB-Medium (für *Lysogeny Broth*-Medium) kultiviert und auf Basis einer Wachstumskurve (bei Ende des exponentiellen Wachstums) geerntet. Darauf folgte die Zerstörung der Zellstruktur und die chemische Isolation und Extraktion der RNA. Anschließend wurde der Anteil an ribosomaler RNA

(allgemein 90-95% aller RNA Moleküle) durch entsprechende Verfahren reduziert und restliche DNA-Moleküle entfernt. Die verbleibende RNA wurde daraufhin mittels Ultraschall fragmentiert, die Fragmente anschließend enzymatisch dephosphoryliert, repariert und letztlich wiederum einheitlich phosphoryliert. Die so vorbereiteten RNA-Fragmente sind der Ausgangspunkt für das weiterführende modifizierte Protokoll nach Abb. 4.2 unter Verwendung der speziellen RT-Primer. Um Sequenzierungen mit unterschiedlicher Anzahl von PCR Zyklen zu erhalten, werden drei parallel durchgeführte Prozesse zur Erzeugung der Sequenzier-Library auf Basis der gleichen RNA durchgeführt, welche im Folgenden mit 11Z, 15Z und 25Z abgekürzt werden. Zur genaueren Analyse der Effizienz der PCR bei unterschiedlichen Zyklenzahlen wurden parallel dazu sogenannte *SYBR-Green*-Kurven erzeugt (siehe Abb. 4.9 für Details), die eine Quantifizierung der PCR-Produkte pro Zyklus ermöglichen. Die Auswertung der Messungen lässt schlussfolgern, dass eine effiziente PCR auf ca. 19-20 Zyklen[7] beschränkt ist. So scheint die PCR für 11Z und 15Z in der exponentiellen Phase vorzuliegen, wohingegen die Kettenreaktion bei 20 Zyklen bereits als gesättigt erscheint und für 25Z definitiv zum Erliegen kommt. Dies wird für Abschnitt 4.3.3 weitere Bedeutung haben. Die letztlich amplifizierte cDNA wird für 11Z, 15Z und 25Z unabhängig voneinander mittels Gelelektrophorese auf eine Größe von 240-300nt angereichert, um Inserts der Länge 50-100nt (mit 36nt Barcode-Templates, 15nt Spacer und Illumina Primer) zu erhalten. Nach entsprechender Aufreinigung und Verdünnung der gelösten cDNA wurden die unterschiedlichen Sequenzier-Libraries zu Sequenzierung auf eine nominelle Molekülkonzentration vom 8-20 pM (picoMolar) verdünnt. Für die letztendliche Sequenzierung mit einer Sequenzierlänge 150nt diente ein Illumina MiSeq-Gerät.

[7] PCR nach Illumina-Standards

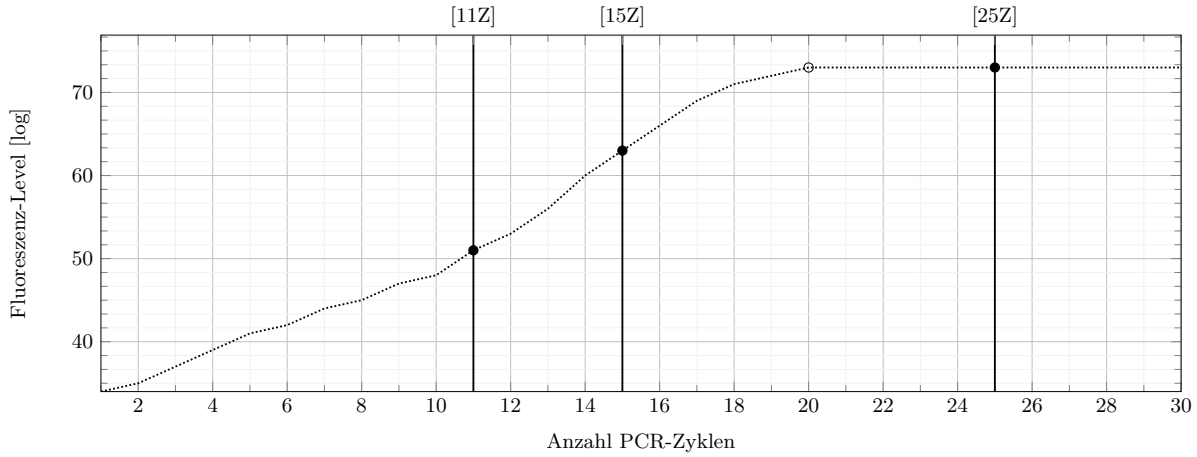

Abb. 4.9 Amplifikation-Kurve (SYBR Green, SmartCycler II): Grundlage der dargestellten Kurve ist die sogenannte quantitative Echtzeit-PCR: Ein Fluoreszenzfarbstoff der sich in doppelsträngiger DNA einlagert wird genutzt, um die Menge an DNA (pro Zyklus) durch ein bildgebendes Verfahren zu detektieren. Die Messungen werden genutzt, um für eine Amplifikation die unterschiedlichen Phasen der PCR zu bestimmen. Im Fall der dargestellten Messung ist zu sehen, dass sich die PCR für 11Z und 15Z noch in der exponentiellen Phase befindet (log-Skala der Fluoreszenz), wohingegen die Reaktion bei ca. 20 Zyklen zum Erliegen kommt. Es ist zu erwarten, dass sich die Zusammensetzung der 25Z Probe in den 5 vorangehenden Zyklen nicht maßgeblich verändert.

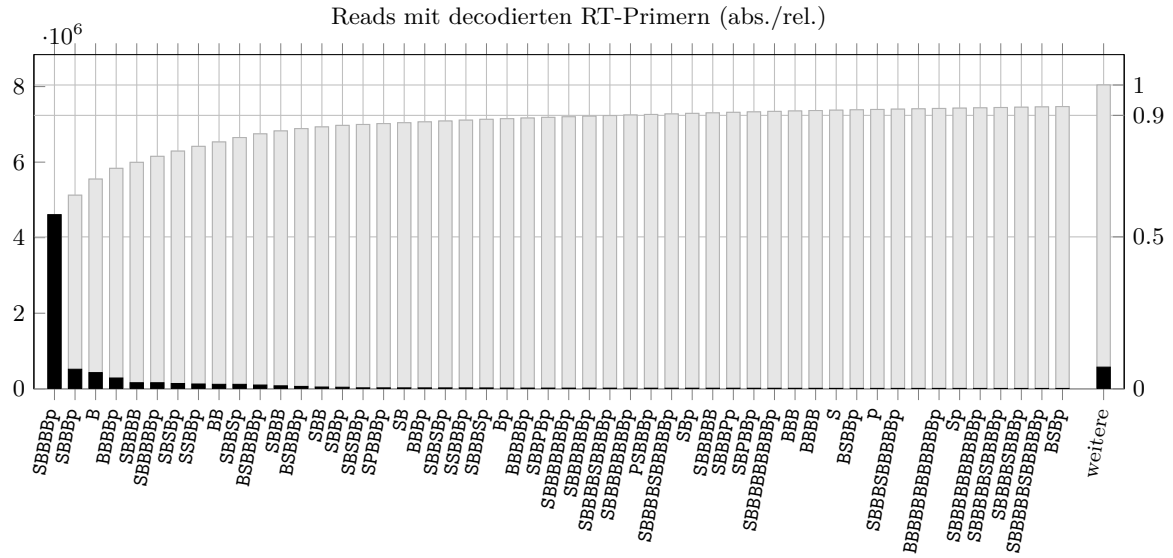

Abb. 4.10 Histogramm aller Reads bezüglich der decodierten RT-Primer (15Z): absolute Anzahl, relativer Anteil, ergänzt durch die kumulative Darstellung. Betrachtet werden die 50 häufigst decodierten Sequenzen (in absteigender Reihenfolge).

4.3.1 Kombinationen der RT-Primer

Ein erster Schritt in der Untersuchung des dargelegten Verfahrens ist die Analyse aller auftretender decodierten Sequenzen des RT-Primers. In Abb. 4.10 ist ein sortiertes Histogramm der für 15Z sequenzierten Reads zu sehen. Die Gruppierung der Balken basiert auf den bei der Decodierung (vgl. 4.2.2) beobachtbaren Mustern der Oligonukleotide. Zur Reduktion der Gruppen wurde auf die Unterscheidung der Orientierung der Barcode-Templates verzichtet: die Symbole `B` stehen stellvertretend für $\{\texttt{b}, \texttt{B}\}$. Die Orientierung aller anderen Sequenzen ist angezeigt. Die Grafik beschränkt sich auf die 50 häufigsten Kombinationen (in absteigender Reihenfolge), dargestellt in absoluter Anzahl und als relativer Anteil aller beobachteten Reads. Die Ergänzung der kumulierten Häufigkeiten deutet an, dass mindestens 90% aller Beobachtungen dargestellt sind.

Die Analyse zeigt eine unerwartete Vielfalt an decodierten Sequenzen, die es mit der Theorie zu vereinen gilt. Das Ergebnis legt nahe, dass die Selektivität der in Abb. 4.5 abgebildeten Gelelektrophorese eine weitaus größere Varianz aufweist, als in (4.1) bis (4.4) theoretisch gefordert wurde. Abgesehen von der niedrigen Effizienz bei der Auswahl der Primer auf die Zielgröße von 72nt (welche experimenteller Optimierung bedarf), können die Daten der Sequenzierungen mit zusätzlichem Aufwand für eine angepasste Validierung des Verfahrens dienen. Denn eine Vielzahl der möglichen Moleküle ungleich 72nt sind bezüglich der Prozesse im Protokoll nichtfunktional und damit schlichtweg nicht beobachtbar. Im Folgenden werden die in Abb. 4.10 dargelegten Molekülformationen und deren Diskrepanz zur Theorie diskutiert. Dazu folgende tabellarische Aufstellung:

- $\texttt{S}\{\texttt{B}\}^x\texttt{p}$: Neben der erwarteten Primer mit Muster `SBBBBP` (57, 3% Anteil) befindet sich zusätzlich eine nicht unerhebliche Menge an beobachteten Reads mit mehr oder weniger Barcode-Templates, wie z. B. `SBBBP` oder `SBBBBBP`. Unter Berücksichtigung der Diversi-

tät dieser außerplanmäßigen RT-Primer werden diese Reads in die folgende Evaluation (Abschnitt 4.3.2) integriert.

- {S, B, BB, p, etc.}: Dabei handelt es sich mit hoher Wahrscheinlichkeit um falsch-positive Ergebnisse der Decodierung für Reads mit zu langen Inserts. Ein niedriger Hamming-Abstand der synthetischen Oligonukleotide zum Genom von E. coli kann zu dieser fehlerhaften Zuordnung führen.
- {B, p}S{B}^{x}p: Hier handelt es sich um Erweiterungen von decodierten Sequenzen in Richtung des Inserts, die ebenfalls falsch-positive Ergebnisse sind. Diese Art Fehler könnte durch die Berücksichtigung der genomischen Referenz bei der Erstellung der Barcode-Templates unter Umständen vermieden werden. Die Barcode-Templates würden dadurch jedoch auf ein spezielles Genom angepasst.
- S{B}*: Resultat unvollständiger Sequenzierung des RT-Primers aufgrund von zu langen Inserts (vgl. Abb. 4.6). Prinzipiell könnten diese Reads zur Korrektur verwendet werden.
- {B}*p: RNA-Fragmente ohne Spacer können in der reversen Transkription nicht zu cDNA ergänzt werden (vgl. Abb. 4.2) und daher in der Sequenzierung nicht auftreten. Bei den beobachteten Sequenzen handelt es sich um Artefakte durch Sequenzfehler in der Sequenz des Spacers. Eine gesonderte Untersuchung der Sequenzen zeigt, dass es sich bei den hier vorliegenden decodierten Sequenzen um falsch-negative Ergebnisse bezüglich des Spacers handelt. Eine Vergrößerung des Decodierradius könnte die Erkennungsrate verbessern. Des Weiteren ist anzumerken, dass der beschriebene Fehler für eine größere Anzahl an PCR Zyklen wohl häufiger aufzutreten (siehe Abb. A.4) scheint.
- S{B, S}*p: Die in Abb. 4.3 dargestellte gerichtete Ligation der RT-Primer ist lediglich eine idealisierte Darstellung. Die Ausbildung der beobachtbaren Bindungen von Oligonukleotiden mit mehreren Symbolen S innerhalb der Struktur kann offenbar nicht ausgeschlossen werden. Weitere Optimierungen hinsichtlich der Reduzierung dieser Formationen sind beispielsweise durch eine Verkürzung der an den DNA Molekülen $\overline{\text{S}}$ und $\overline{\text{P}}$ beteiligten Einzelsträngige möglich.

Die Größenordnung der in Abb. 4.10 als „weitere“ gekennzeichneten Kombinationen kann mit $10^3 - 10^4$ beziffert werden und obwohl eine Vielzahl der erklärbaren Strukturen den Zweck der Zufallsbarcodes erfüllen, beschränken sich weitere Analysen auf klar definierte Sequenzen.

4.3.2 Qualität der PCR-Korrektur

Anhand der decodierten RT-Primer werden die Sequenzierungen in 5 Kategorien von Reads analysiert. Die Daten werden dabei unterteilt in Reads, welche Primer der Form SBp, SBBp, SBBBp, SBBBBp oder SBBBBBp beinhalten, alle weiteren Kombinationen werden dabei verworfen. Zunächst wird die Zusammensetzung der Reads quantifiziert. Dabei wird der Zahl aller Sequenzen die Anzahl unterscheidbarer Kombinationen von Barcode-Templates gegenübergestellt, d. h. der Vergleich von Sequenzierdaten mit und ohne PCR-Duplikate. Daran anschließend wird die Verteilung der RT-Primer genauer betrachtet und geschätzt. Basierend auf der Modellannahme aus 4.2.3 und der geschätzten Verteilung wird der theoretische Korrekturfehler analysiert. Für weitere Analysen der PCR-Korrektur werden nur Primer der Form SBBBBp berücksichtigt, was die ursprüngliche Intention widerspiegelt.

RT-Primer	11Z		15Z		25Z		max. Komb.
SBp	1400	**40**	14158	**40**	59306	**40**	**40**
SBBp	11385	1481	39793	1574	138979	**1600**	**1600**
SBBBp	124971	43493	516049	54733	719520	61055	64000
SBBBBp	499152	371167	4608564	1187520	4405569	1544436	2560000
SBBBBBp	189603	157867	158744	93857	168337	139877	102400000
weitere	281711		2705263		3955667		
Summe	1108222		8042571		9447378		

Tab. 4.1 Zusammensetzung der validen RT-Primer Kombinationen für die Sequenzierungen (11Z,15Z und 25Z): Zeilen beziehen sich auf die Formen valider Primer, weitere (ignorierte) Muster und die Summe aller sequenzierten Reads. Aufgeführt ist die Anzahl beobachtbarer RT-Primer (erster Wert) und die Anzahl unterscheidbarer RT-Primer (zweiter Wert). Die letzte Spalte zeigt die maximale Anzahl theoretisch möglicher Kombinationen für die Primer.

Valide Primer-Moleküle

Die Zusammensetzung der RT-Primer nach Filterung (vgl. 4.2.2) ist in Tab. 4.1 zusammengefasst. Gegenübergestellt werden für die unterschiedlichen Längen x der Barcode-Templates jeweils die Anzahl aller beobachtbarer Sequenzen und die Anzahl unterscheidbarer Kombinationen. Letztere gibt eine Einschätzung bezüglich der Diversität der auftretenden Zufallsbarcodes und damit der Abdeckung an beobachtbaren Sequenzen bezüglich aller 40^x möglichen Kombinationen. Dabei ist bei weitem nicht jede Kombination durch mindestens einen Read belegbar: Abgesehen von der Beobachtung aller 40 möglichen Barcode-Templates für die Kategorie SBp (ebenso 40^2 Paare von Barcode-Templates für 25Z) bleiben viele der möglichen Molekülformationen bei der Sequenzierung verborgen. Diese Tatsache ist der mangelnden Abdeckung an Beobachtungen geschuldet. Ein faktischer Nachweis, ob eine bestimmte Kombination von Oligonukleotiden tatsächlich als RT-Primer im Experiment verfügbar ist, könnte letztlich nur durch eine erhebliche Vergrößerung der Beobachtungsmenge erbracht werden. Für die Größenordnung von 10^8 möglichen Molekülen der Form SBBBBBp wäre dieser Nachweis mit heutiger Technik sehr aufwändig. Des Weiteren ist bezüglich Tab. 4.1 zu bemerken, dass die Gesamtzahl an Reads stark von dem Umfang der PCR abhängig ist (die niedrige Anzahl an Reads für 11Z ist dabei ein reproduzierbares Phänomen[8]).

Verteilung der RT-Primer und Korrekturfehler

Wie im Paragraph zuvor erwähnt kann die Existenz aller theoretisch möglichen RT-Primer durch die durchgeführten Sequenzierungen nicht belegt werden, weil die Anzahl an Ereignissen in der Größenordnung der Beobachtungen oder sogar darüber liegt. Möchte man für die vorliegenden Sequenzierungen die Qualität der PCR-Korrektur über die Zusammenhänge aus 4.2.3 bewerten (über die Verteilung von Zufallsbarcodes), so ergeben sich diverse Probleme, welche im Folgenden diskutiert werden:

Unabhängig davon, welche Verteilung bei der Bewertung entscheidend ist, stellt sich bei der Anzahl von möglichen Ereignissen und Beobachtungen die Frage nach der korrekten Methodik. Relative Häufigkeiten von Molekülen sind eine Möglichkeit etwas über deren Verteilung zu erfahren, wobei diese Schätzung zwei Schwächen zeigt (vgl. Definition 10): Nicht beobachtete

[8] Der Zusammenhang von PCR-Zyklen und Anzahl Reads ist systematisch/experimentell belegbar (nicht durch veränderte Parameter bedingt). Mögliche Ursachen bedürfen gesonderter molekular-technischer Analyse.

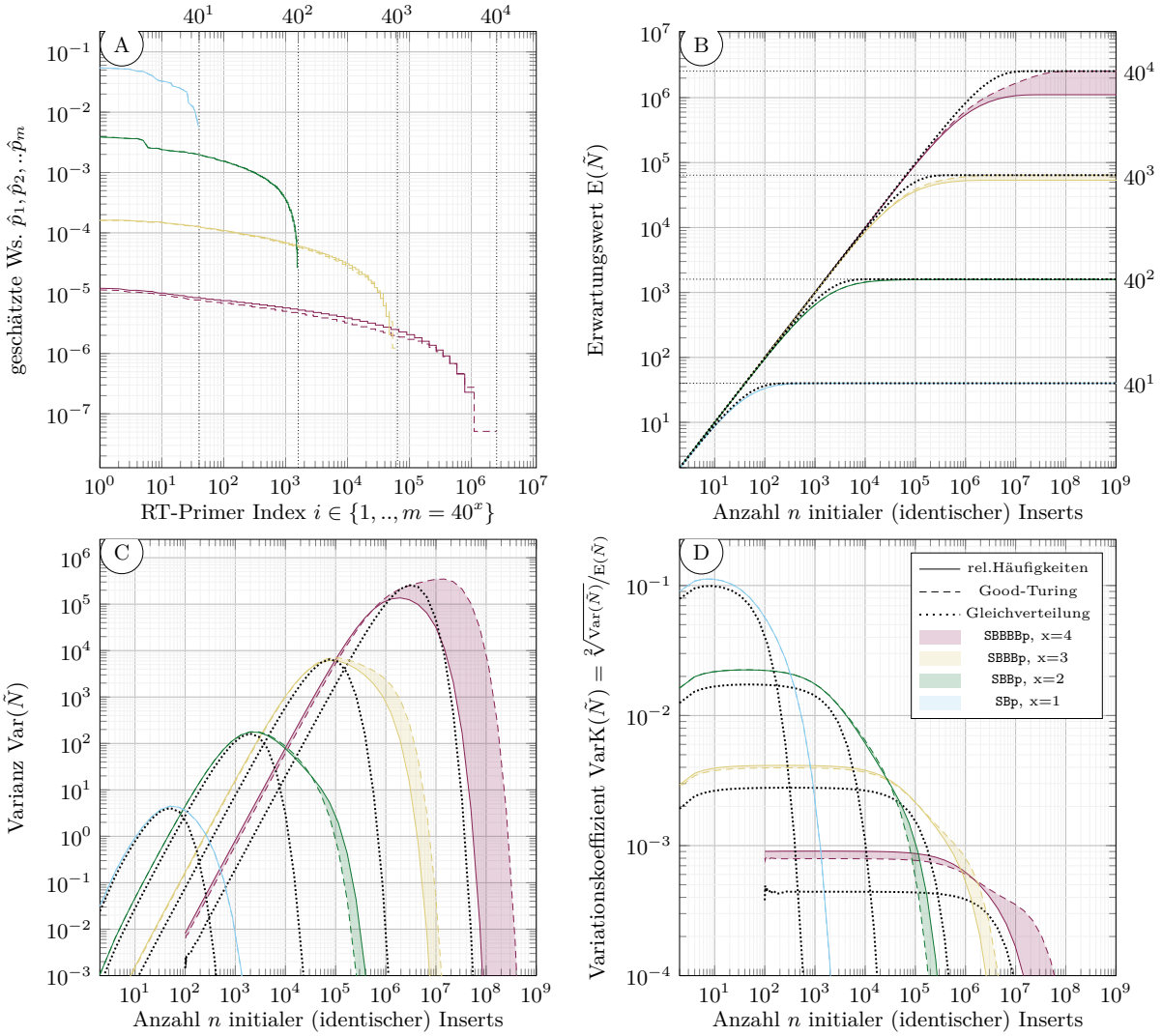

Abb. 4.11 Grenzen der PCR-Korrektur (15Z): (A) Schätzung bedingter Auftrittswahrscheinlichkeiten $\hat{p}_i$ von RT-Primern abhängig von der Anzahl x enthaltener Barcode-Templates `B`. Primer sind Indexen i zugeordnet (Wahrscheinlichkeit absteigend). Für die Schätzung der relativen Häufigkeiten ist m reduziert, die Good-Turing-Schätzung geht hingegen von einer umfassenden Verteilung aus. (A) und n als hypothetische Anzahl an identischer initialer Inserts (x-Achse) ist mit dem theoretischen Ansatz aus 4.2.3 Grundlage für: Erwartungswert (B) der korrigierten Zählgröße $\tilde{N}$ (Zufallsvariable), deren Varianz (C) und Variationskoeffizient (D). Um die Darstellung von numerischen Ungenauigkeiten zu vermeiden wurden Kurven beschnitten.

Ereignisse werden einerseits aggregiert und mit einer Wahrscheinlichkeit von Null bewertet, zusätzlich wird der Anteil nicht nachweisbarer Ereignisse verhältnismäßig auf die Wahrscheinlichkeit der beobachtbaren verteilt. Der Good-Turing-Schätzer (vgl. Konzept 2) bietet hierfür eine Alternative zur Korrektur von Verteilungen bei relativ seltenen Beobachtungen.

Ein weiteres Problem besteht in der Tatsache, dass die Verteilung der RT-Primer zum Zeitpunkt der Sequenzierung nicht gleichbedeutend ist mit den Wahrscheinlichkeiten, die bei der

reversen Transkription der Inserts vorgelegen haben. Betrachtet man eine idealisierte uniforme Verteilung von Molekülen, so lässt die PCR erwarten, dass sich die empirische Verteilung der Reads im Allgemeinen nicht gleichmäßiger repräsentiert, als die ursprüngliche Form. Da die Gleichverteilung das Optimum für den Einsatz von Zufallsbarcodes darstellt, führt folgende Näherung zu einer eher konservativen Bewertung der empirischen Verteilungen. Als Näherung kann man annehmen, dass die PCR keinen Einfluss auf die Verteilung der RT-Primer hat und die beobachteten Reads der Wahrscheinlichkeitsverteilung zum Zeitpunkt der reversen Transkription entsprechen. Dies wäre der Fall, wenn die Anzahl an Molekülen hinreichend groß wäre und wenn jedes Molekül (Insert und Zufallsbarcode) durch die PCR im Mittel gleich oft dupliziert würde. Diese Annahmen sollen im Folgenden dazu dienen eine konservative Abschätzung der Qualität der PCR-Korrektur auf Basis von Abschnitt 4.2.3 zu erstellen.

In Abb. 4.11 (A) sind geschätzte Auftrittswahrscheinlichkeiten $\{\hat{p}_i\}$ der RT-Primer auf Grundlage der Sequenzierdaten dargestellt. Dabei sind sowohl die relativen Häufigkeiten (konservative Schätzung) und die angepasste (moderate) Schätzung nach Good-Turing gezeigt. Die Good-Turing-Methode ordnet den nicht beobachteten RT-Primern eine korrigierte Wahrscheinlichkeit ungleich Null zu. Diese Korrektur hat natürlich Einfluss auf die folgenden theoretischen Ergebnisse: Um konservative und moderate Schätzung der Diversität der Primer gegenüberzustellen, sind daher jeweils Kurven für beide Verteilungen gezeigt. Flächen zwischen den Kurven können als ein Bereich der realistischen Schätzung von theoretischen Größen gesehen werden. Dabei sind die Grafiken wie folgt zu lesen: Einer hypothetische Anzahl identischer initialer Inserts n (x-Achse) werden theoretisch errechnete Werte (y-Achsen) zugeordnet, die zusätzlich auf der geschätzten Wahrscheinlichkeitsverteilung $\hat{p}_1, \hat{p}_2, .., \hat{p}_m$ der Primer und der Zufallsvariable $\tilde{N}$ (korrigierte Zählgröße) basieren. Die dafür relevanten Größen wurden in (4.9) bis (4.11) bereits definiert. Neben der erwarteten korrigierten Anzahl an Inserts sind sowohl die Varianz der Zählgröße $\tilde{N}$ als auch der Variationskoeffizient (vgl. Definition 1) dargestellt. Letzterer kann als relativer Fehler interpretiert werden, der abhängig von der mittleren Molekülanzahl die Unsicherheit in der Zählung (mit Zufallsbarcodes) beschreibt: Für Primer der Form `SBBBBp` bedeutet das konkret, dass mit der geschätzten Diversität für die korrigierte Zählgröße $\tilde{N}$ an Molekülen gilt, dass das Verhältnis der Standardabweichung zum Erwartungswert im Mittel kleiner als 10^{-3} ist. Für Molekülzahlen n in welchen der Einsatz von Zufallsbarcodes sinnvoll erscheint, d. h. für Bereiche in denen $\mathrm{E}(\tilde{N})$ und n einen annähernd linearen Zusammenhang beschreiben, ist die Zufallsgröße $\tilde{N}$ mit hoher Wahrscheinlichkeit nahe am Erwartungswert zu finden. Gleiches ist für realen Einsatz und statistische Auswertungen (wie z. B. in 4.3.3) zu erwarten. In den relevanten Bereichen von n ($n \ll m$) sind für die verschieden Klassen an Primern beim Variationskoeffizienten jeweils Plateaus zu erkennen, in welchen lediglich marginale Unterschiede bezüglich der unterschiedlichen Verteilungen zu sehen sind. Diese Plateaus kennzeichnen Parameter für den sinnvollen Einsatz der vorgestellten Zufallsbarcodes zu Korrektur von PCR-Duplikaten. Die gezeigten Zusammenhänge lassen weiter Analysen und Abschätzungen von Fehlerwahrscheinlichkeiten zu, auf welche an dieser Stelle nicht eingegangen wird. Weitere Darstellungen zum Vergleich der Evaluierung der Sequenzierungen 11Z und 25Z sind in Abb. A.5 und A.6 aufgeführt.

Grundlage für die Ergebnisse der folgenden Abschnitte ist die alternative Alignierung (siehe 4.2.4) der Sequenzierdaten basierend auf der in (4.5) definierten Liste Φ von decodierten Reads (für 11Z, 15Z und 25Z) und der genomischen Referenz $\mathbf{G}$ von E. coli [97]. Die Menge der decodierten Reads Φ (vgl. Tab. 4.1) wird hierfür weiter gefiltert, indem nur noch Primer

Zusammensetzung d. Inserts	**11Z**		**15Z**		**25Z**	
länger als 15nt	134885	(100%)	458296	(100%)	645499	(100%)
Alignierung erfolgreich	77856	(63,4%)	259136	(58,2%)	342788	(54,3%)
nicht aligniert	49344	(36,6%)	191464	(41,8%)	295012	(45,7%)
kürzer 15nt (nicht berücksichtigt)	504	-	5965	-	13828	-
Summe (unterschiedliche Inserts)	135389		464261		659327	

Tab. 4.2 Ergebnis der Alignierung für 11Z, 15Z und 25Z: Grundlage für das Sequenzalignment sind die in Tab. 4.1 zusammengefassten Reads, beschränkt auf Primer der Form `SBBBBp`. Die Größen beziehen sich auf unterscheidbare Sequenzen, d. h. eine Vielzahl der gezählten Inserts liegt mehrfach in den Sequenzierdaten vor. Verworfen wurden Insert welche kürzer als 15nt sind und nicht für die Alignierung berücksichtigt. Für Sequenzen die länger sind als 15 Symbole ist eine erfolgreiche Zuordnung zum Referenzgenom ein weiteres Unterscheidungsmerkmal.

der Form `SBBBBp` berücksichtigt werden, was die ursprüngliche Intention der hier angewendeten Zufallsbarcodes widerspiegelt. Für sämtliche in Φ enthaltene Inserts $\mathbf{F}$ werden alle optimalen Sequenzalignments zu $\mathbf{G}$ berechnet und im erwähnten Alignment-Lattice $\mathcal{L}$ abgelegt. Die Zusammensetzung der verbleibenden Inserts ist in Tab. 4.2 aufgeschlüsselt. Die relativ geringen Raten an erfolgreich alignierten Inserts von 54-63% (vgl. Raten in Begriff 17) sind hauptsächlich den modifizierten Enden der RNA (siehe Begriff 7) geschuldet, welche in der hier beschriebenen Analyse fehlertolerant verarbeitet werden. Alle erfolgreich alignierten Inserts bilden letztendlich eine Menge $\mathcal{L}$ an Zuordnungen von Symbolen F_i (der Position i im Insert) auf Symbolen G_j (der Position j im Genom). Die Projektion von $\mathcal{L}$ in j-Richtung wird im folgenden Abschnitt genutzt, um den Einfluss der PCR auf das Genom abzubilden.

4.3.3 Umfang der PCR-Duplikate und die Hypothese der Selbst-Hybridisierung

Im Abschnitt zuvor wurde die Qualität der PCR-Korrektur bewertet, indem theoretische Ergebnisse über deren Grenzen (vgl. Abschnitt 4.2.3) auf geschätzte Verteilungen angewendet wurden. In diesem Abschnitt soll der konkrete Einsatz der vorgestellten Zufallsbarcodes für quantitative Analysen veranschaulicht werden. In einem ersten Schritt wird das Ausmaß der PCR-Duplikate an beispielhaften Positionen im Referenzgenom von E. coli verdeutlicht. Dabei werden die Ergebnisse der verschiedenen Sequenzierungen 11Z, 15Z und 25Z direkt gegenübergestellt und verglichen. Die exemplarische Analyse der Reads für unterschiedliche Anzahl an PCR-Zyklen gibt einen Hinweis auf einen Effekt, der sich genomweit bestätigen lässt. Ausführungen zur Hypothese der Selbst-Hybridisierung als eine mögliche Erklärung für das Beobachtete schließt diesen Teil der Validierung.

Intensität der PCR-Duplikate und deren Korrektur

Ist durch die entsprechende Projektion (vgl. 4.2.4) des Lattices $\mathcal{L}$ sichergestellt, dass jedes Insert $\mathbf{F}$ genau eine Position j im Genom überdeckt, können die Tupel aus $\mathcal{L}$ anhand von Φ und der in (4.7) und (4.8) gegebenen Zählgrößen $\check{n}(\mathbf{F}), \hat{n}(\mathbf{F})$ und $\tilde{n}(\mathbf{F})$ gewichtet und somit (bezogen auf j) aufsummiert werden. Die Summen ergeben damit eine positionsabhängige Schätzung für die Anzahl von RNA-Fragmenten, die zu Beginn der Herstellung der Sequenzier-Library existierten. Dabei ist $\check{n}_j$ die Anzahl aller j überdeckender Fragmente ohne PCR-Korrektur, $\hat{n}_j$

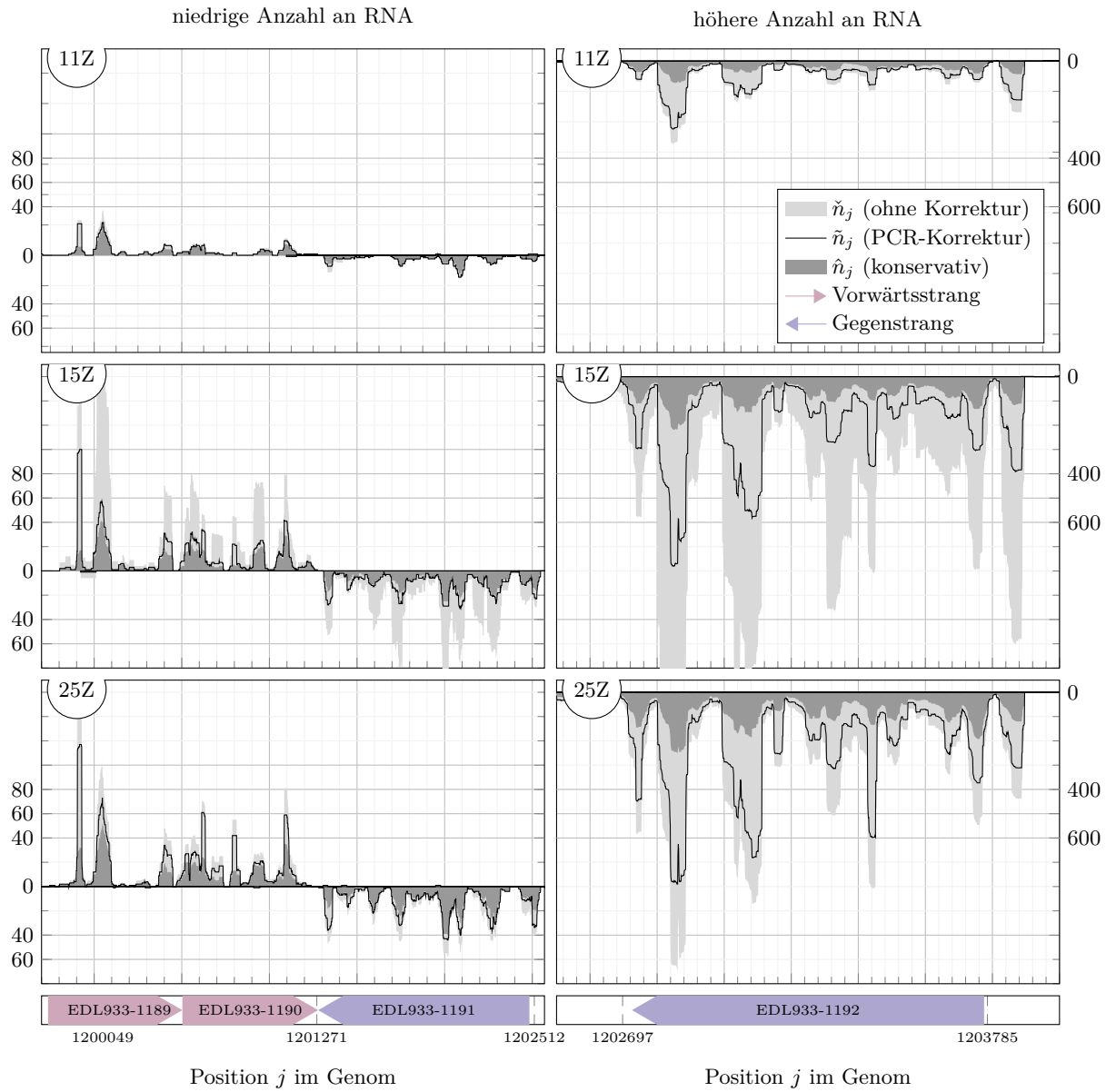

Abb. 4.12 Ausmaß der PCR-Duplikate (exemplarisch) anhand der geschätzten Zählgrößen $\check{n}_j, \hat{n}_j$ und $\tilde{n}_j$ (vgl. Abschnitt 4.2.2) strangspezifisch für 11Z,15Z und 25Z: Die Beobachtung von PCR-Duplikaten ist für 11Z auf Grund der niedrigen Gesamtanzahl von Reads herabgesetzt, daraus folgt $\tilde{n}_j \approx \hat{n}_j$. Für steigende Zyklenzahlen ist ein steter Zuwachs an PCR-Duplikate zu erwarten. Anomalie: Trotz vergleichbarer korrigierter Größen $\tilde{n}_j$ zeigen sich für 15Z im Vergleich zu 25Z deutlich mehr identische Reads.

entspricht der konservativen Zählung, die jedes Duplikat als PCR-Kopie bewertet und $\tilde{n}_j$ bezeichnet schließlich die korrigierte Anzahl auf Basis der Decodierung der RT-Primer.

In Abb. 4.12 sind die Zählgrößen $\check{n}_j, \hat{n}_j$ und $\tilde{n}_j$ (y-Achse) beispielhaft für verschiedene Positionen j im Genom (x-Achse) dargestellt. Insgesamt umfasst das Beispiel vier RNA-codierende Bereiche auf der DNA von E. coli, die bezeichnet werden mit EDL933-1189 bis EDL933-1192. Die Bereiche befinden sich sowohl auf dem Vorwärts- als auch Gegenstrang des Genoms und

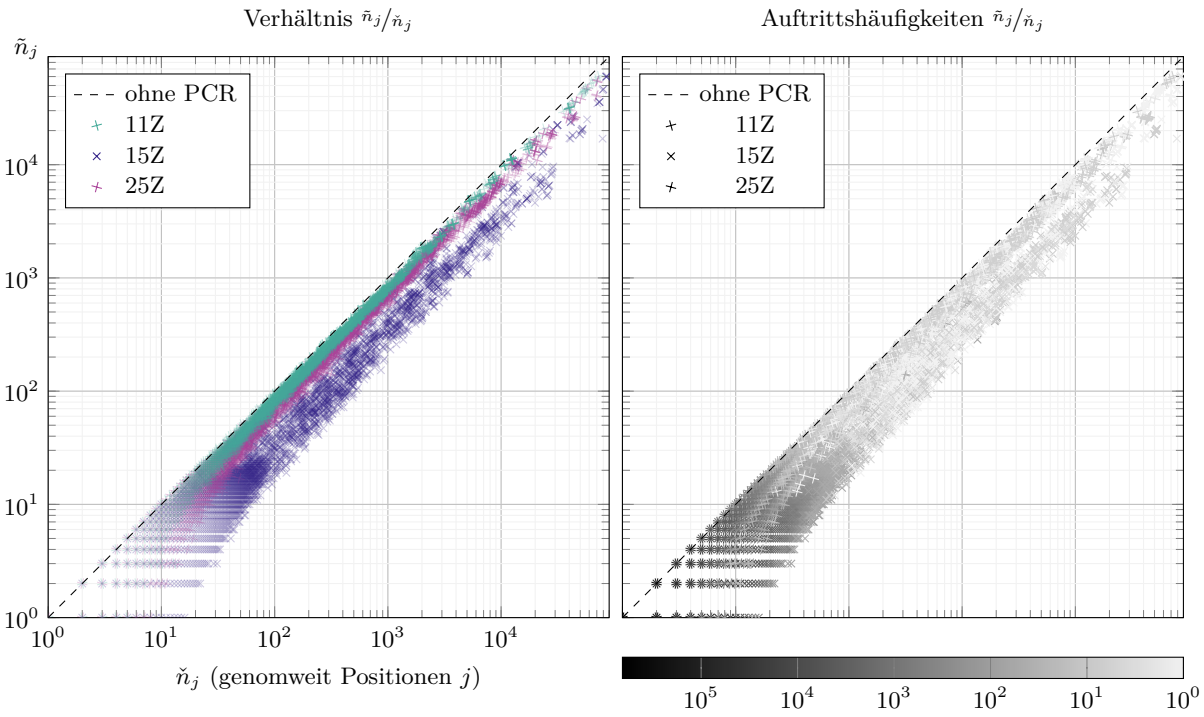

Abb. 4.13 Ausmaß der PCR (genomweit) für 11Z, 15Z und 25Z: Verhältnis $\tilde{n}_j/\check{n}_j$, der Anzahl korrigierter Inserts $\tilde{n}_j$ zur Zählgröße $\check{n}_j$ ohne PCR-Korrektur, bezogen auf alle Positionen j im Genom vom E. coli. Werte der Verhältnisse (links), Auftrittshäufigkeit in Graustufen (rechts).

weisen unterschiedliche Dynamik hinsichtlich der Transkription auf. Zur Darstellung der strangspezifischen Anzahl der alignierten Inserts wird sowohl die positive (Vorwärtsstrang) als auch negative (Gegenstrang) y-Halbachse genutzt. Eine Korrektur von PCR-Duplikaten zeigt für 11Z keine großen Veränderungen hinsichtlich der gezählten Inserts. Grund hierfür ist die experimentell bedingte niedrige Gesamtzahl an Reads für diese Sequenzierung (vgl. Tab. 4.1), wodurch die Wahrscheinlichkeit einen identischen Read zu beobachten (durch die vergleichbar geringe Menge an Beobachtungen) gemindert ist. Eine vergleichende Bewertung mit den Sequenzierungen 15Z und 25Z, welche deutlich höhere Ausgangsmengen an Sequenzen bereitstellen, ist daher nur bedingt möglich. Stellt man jedoch die Ergebnisse von 15Z und 25Z gegenüber, zeichnet sich ein unerwartetes Bild ab: Betrachtet man die Differenz der Anzahl aller Inserts $\check{n}_j$ und der durch die Korrektur verbleibenden $\tilde{n}_j$, kann das Ausmaß der PCR-Duplikate abgeschätzt werden. Zu erwarten wäre ein durch die Anzahl an PCR-Zyklen bedingter monotoner Zuwachs der PCR-Duplikate. Trotz ähnlicher Größen $\tilde{n}_j$ scheint für 15Z eine deutlich größere Anzahl an identischen Reads vorzuliegen, als das für 25Z der Fall ist. Dass es sich bei dem dargestellten Sachverhalt nicht um eine gesonderte Beobachtung an einem speziell gewählten Bespiel handelt, ist durch eine genomweite Analyse der PCR belegbar.

Zu diesem Zweck ist in Abb. 4.13 das Verhältnis von $\tilde{n}_j$ zu $\check{n}_j$ für jede Position j im Genom von E. coli in einer doppeltlogarithmischen Abbildung veranschaulicht. Jeder Punkt in der zweidimensionalen Ebene (links) steht für die Beobachtung von mindestens einem Verhältnis $\tilde{n}_j/\check{n}_j$ an einem bestimmten Nukleotid. Die Auftrittshäufigkeit der Verhältnisse (linkes Bild) ist zusätzlich als Projektion in Grauwerte integriert (rechts). Die Abbildung erfüllt die Funktion eines 2-dimensionalen Histogramms. Zum besseren Verständnis der Grafik folgender Hinweis: Wäre

eine Sequenzierung ohne PCR experimentell umgesetzt oder würden Zufallsbarcodes existieren, die ein eindeutiges Zählen von Inserts ermöglichen, so würden keine PCR-Duplikate in den Daten verbleiben, was durch die winkelhalbierende Gerade $\tilde{n}_j = \check{n}_j$ veranschaulicht wird. Welcher Korrekturschritt wie oft in den Sequenzierdaten auftritt ist die Kernaussage der Darstellung. Auch diese umfassende genomweite Analyse der PCR wirft die Frage auf, warum die Anzahl der PCR-Duplikate für eine erweiterte Anzahl von PCR-Zyklen rückläufig ist.

Hypothese der Selbst-Hybridisierung

Eine mögliche Erklärung für das beobachtbare Verhalten könnte das kontra-intuitive Phänomen sein, das in [110] als Rehybridisierung bei der PCR klassifiziert wurde. Bezugnehmend auf den sogenannten C_ot-Effekt [80] bestätigt [110] diesen Effekt, welcher in [157] für die konkurrierende PCR von einzelnen DNA-Primer-Konstellationen gezeigt wurde. Schon in [110] wurde auf die Bedeutung und Wichtigkeit des Effekts hingewiesen, speziell wenn mehr als ein spezifisches PCR-Produkt amplifiziert wird, wie es letztlich bei einer herkömmlichen PCR der Fall ist. Zunächst werden grundlegende Zusammenhänge der sogenannten Selbst-Hybridisierung erklärt, bevor denkbare Implikationen des genannten Effekts für Zufallsbarcodes erläutert werden. Mögliche positive Nebeneffekte der PCR im Kontext von Zufallsbarcodes blieben in anderen Publikationen zu diesem Thema bisher unberücksichtigt.

Die PCR ist eine zyklische Abfolge der drei Schritte (vgl. Begriff 11): Aufschmelzen, Hybridisierung und Elongation. Für die Effizienz der Reaktion ist maßgeblich, wie oft die Schritte in einer tatsächlichen Verdopplung eines Ausgangsmoleküls resultieren, d. h. eine gezielte Hybridisierung von Primern an den aufgetrennten Doppelsträngen auftritt und somit eine Komplettierung von Duplikaten durch die Polymerase initiiert wird. Es existieren eine Vielzahl von Einflussfaktoren auf die zuvor definierte Effizienz, wie z. B. Bindungsenergien oder Molekülkonzentrationen. Diese Einflussfaktoren sind jedoch keine fixen Größen. So unterliegt beispielsweise die Molekülkonzentration der Dynamik eines Ausgleichsprozesses, weil die PCR-Reaktion ein stoffmäßig geschlossenes System darstellt. Berücksichtigt man wie in [110, 157] das Verhältnis von identischen DNA-Molekülen gegenüber verfügbaren Primern, so lässt sich ein selbst-hemmender Einfluss von vielfach identisch vorliegenden Molekülen erklären. Dominierend wird dieser Effekt für spätere PCR-Zyklen, in welchen PCR-Produkte hinreichende Konzentrationen erreichen können, in denen eine Selbst-Hybridisierung (komplementärer Einzelstränge) mit der Hybridisierung von Primern konkurriert. Die Reduzierung der Effizienz für zahlreich vorliegende Moleküle resultiert somit in einer Normalisierung der initialen DNA-Fragmenten am Ende einer jeden PCR. Das heißt, Moleküle mit niedriger Konzentration werden im Vergleich zu Teilchen mit hoher Dichte verhältnismäßig stärker amplifiziert.

Das in Abb. 4.12 und 4.13 angedeutete Phänomen der reduzierten Anzahl von PCR-Duplikaten für 25Z kann der dargelegten Hypothese der Selbst-Hybridisierung entsprechen: Fragmente aus Inserts und zufälligen Barcode-Templates werden nur in einem beschränkten Umfang effizient amplifiziert. Begünstigt werden (in späteren PCR-Zyklen) Moleküle, welche im momentanen Gefüge unterrepräsentiert sind. Geht man davon aus, dass das Phänomen für Insert und Barcode-Templates gleichermaßen existiert, würde das ebenfalls eine Erklärung für die vergrößerte Diversität der beobachtbaren Kombination von RT-Primern bieten (vgl. Tab. 4.1, 1187520:1544436 Kombinationen `SBBBBp` für 15Z:25Z). Zur letztlichen Bestätigung der dargelegten Hypothese bleiben weitere Experimente offen, die nicht im Rahmen der vorliegenden

Arbeit zu sehen sind. Sollte sich der Effekt der Selbst-Hybridisierung im Kontext von Zufallsbarcodes bestätigen, hätte das weitreichende Implikationen zur möglichen Optimierung des Verfahrens zur Quantifizierung von RNA durch Sequenzierung. Das beschriebene nichtlineare Verhalten der PCR, welches als generelles Problem für standardmäßige quantitative Analysen gesehen wird, könnte unter Einsatz von Zufallsbarcodes genutzt werden, um den Dynamikumfang bei der Sequenzierung zu erhöhen: Auf Basis einer validen PCR-Korrektur wäre es somit gezielt möglich die Wahrscheinlichkeit zu vergrößern seltene Moleküle durch die Sequenzierung zu beobachten. Dafür sollte der Effekt der Selbst-Hybridisierung gezielt verstärkt werden.

Für die vorausgehenden Analysen der Sequenzierdaten hatte die konkrete Sequenz der Inserts eine untergeordnete Rolle. So enthält beispielsweise Abb. 4.12 und 4.13 keine Information darüber, welches Nukleotid beim Sequenzalignment eine Diskrepanz zum Referenzgenom aufweist oder wie häufig eine bestimmte Stelle im Genom von E. coli von einem nicht übereinstimmenden Symbol überdeckt ist. Für die bisher gezeigten quantitativen Betrachtungen ist lediglich entscheidend, ob ein Insert unter einer maximalen Abweichung (hier Editierdistanz) auf das Genom abgebildet werden kann, jedoch nicht welche Ersetzungen dazu nötig sind. Besonderes Augenmerk wurde dabei auf die Korrektur von mehrfach identisch auftretenden Insert gelegt, um eine verbesserte Schätzung der Anzahl ursprünglicher Moleküle zu erhalten. Diese Korrektur hat selbstverständlich für alle Arten von Statistiken, für welche Sequenzierdaten herangezogen werden, eine essentielle Bedeutung.

Der folgende Teil hingegen umfasst qualitative Analysen hinsichtlich der Sequenzfehler und deren Statistik. Die Idee via Sequenzalignment Aussagen über die Qualität der Sequenzierung und der dabei auftretenden Sequenzfehler zu treffen ist nun schon einige Jahre bekannt. In [44], einer der frühen Publikationen zu Fehlerquellen bei der DNA-Sequenzierung, wird beispielsweise ein sehr restriktives Alignment von relativ kurzen Reads der Länge 32nt genutzt, um Substitutionsraten in den Sequenzierdaten zu schätzen. Später wurden größere Sequenzierlängen und fehlertoleranteres Sequenzalignment herangezogen, um Fehlerstatistiken durch die Dimension der Positionsabhängigkeit zu erweitern und das Alignment durch gezieltes Beschneiden der Sequenzen zu verbessern [117]. So wurde nicht nur die Position innerhalb des Reads zum freien Parameter, sondern die Lage der Sequenz auf dem Genom zur relevanten Größe bei Auswertungen. Eine logische Erweiterung wird daraufhin in [113, 121] präsentiert, als der Kontext um Sequenzfehler und damit sequenzabhängige Fehler in Betracht gezogen werden. Der Schritt, von einem durch die Position im Genom bestimmten Fehler zur Annahme der Abhängigkeit des Sequenzfehlers vom umliegenden Kontext, ist der zentrale Punkt in den entsprechenden Ansätzen. Während in [121] zur Identifikation von fehlerträchtigen Sequenzen eine Art Schwellwert verwendet wird, nutzt [113] einen robusten Test auf Basis eines Hintergrundmodells, welcher in angepasster Form in dieser Arbeit genutzt wird. Mit der Fehlerschätzung bei der Sequenzierung von DNA ist gleichzeitig die Erkennung und Filterung von Fehlern zur Aufbereitung von Sequenzierdaten ein essentieller Themenkomplex. Ausgehend von einer Filterung auf Basis von Qualitätswerten (Seiteninformationen der Sequenziergeräte) [117] oder weiteren Klassifikationsmerkmalen [20] ist im Kontext der Fehlerkorrektur eine Tendenz zur Anpassung der Sequenzierexperimente zu erkennen. So stößt bei inhärenten und systematischen Fehlerquellen das reine Aufbereiten von Daten an Grenzen, die sich durch Anpassungen der Experimente verschieben lassen. In [47] wurde beispielsweise die beidseitig überlappende Sequenzierung genutzt, um bei der Zusammenführung von Reads (Konsensus genannt) Sequenzfehler zu detektieren. Die Veröffentlichung [47] nennt hingegen die PCR als maßgebliches Problem des vorgestellten

Ansatzes. Das Hindernis scheint in [148] durch den Einsatz von Zufallsbarcodes und Konsensus-Sequenzen überwunden zu werden, wobei einige Fragen zur Anwendbarkeit für RNA und deren modifizierten 3'-Enden offen bleiben.

4.3.4 RNA und ihre 3'-Enden

Die Analyse von Sequenzfehlern in RNA birgt besondere Herausforderungen, die nun kurz diskutiert werden: Anders als bei der DNA unterliegt das RNA-Molekül unter Umständen nach der Synthese einer Modifikation (siehe Begriff 7), die eine direkte Abbildung auf ein Referenzgenom erheblich erschweren kann. Eine Molekülveränderung ist die sogenannte Polyadenylierung, welche für Eukaryoten schon länger bekannt ist und für Prokaryoten [36, 143] eine Menge unbeantworteter, molekularbiologisch höchst aktueller Fragen [137, 155] aufweist. Unter den teils widersprüchlichen Facetten zur Bedeutung und Wirkung der Polyadenylierung bei Prokaryoten herrscht breite Übereinkunft, dass die Beobachtung von RNA mit erweiterten Polyadeny-3'-Enden im Wirkungszusammenhang des Abbaus von Molekülketten (Degradation) steht. Unabhängig von der exakten Bedeutung ist die Detektion des Phänomens für die Erforschung von essentieller Bedeutung. Dabei stellt die Sequenzierung von RNA ein neues Medium zur Beobachtung dieser molekularen Veränderungen dar.

Bei der Polyadenylierung handelt es sich um eine systematische, nicht dem Sequenzierexperiment geschuldeten Veränderung am Insert. Die bisher benannten Ansätze gehen von systematischen Fehlerquellen in der Sequenzierung aus, nicht von weiteren elementaren Modifikationen am Molekül. Daher ist anzunehmen, dass die Beobachtung von RNA Fragmente und deren modifizierte 3'-Enden durch eine Filterung stark begrenzt wird. Methoden zur Identifikation der Modifikationen [98] gewinnen zunehmend an Interesse und auch der Einsatz von DNA-Sequenzierung als Analysemethode gewinnt in jüngster Zeit zunehmend an Zuspruch [53, 76, 180]. Jedoch sind bei der Erkennung von Modifikationen meist einfache Heuristiken anzutreffen und der Einfluss von PCR-Duplikaten bleibt bei statistischen Auswertungen unberücksichtigt. Es existieren zudem eine Reihe bekannter Probleme, die eine Detektion des Phänomens erschweren [98]: Eine verminderte Qualität der Sequenzierverfahren an den Enden der Reads ist beispielsweise ein technisches Hindernis, eine ungleichmäßige und schwache Transkription ein biologisches Hemmnis. Zudem ist bekannt, dass die Polyadenylierung nicht notwendigerweise aus uniformen und konstanten Anhängseln besteht und aufgrund des Abbauprozesses eine äußerst latente und flüchtige Erscheinung ist. Für Prokaryoten kommt erschwerend hinzu, dass die Polyadenylierung als signifikant kürzer vermutet wird (14-60nt statt 80–200nt [143]) und die Dynamik des Abbaus als deutlich größer angenommen wird [155]. Von einem Zusammenhang zwischen Länge und Dynamik wird dabei ausgegangen.

In diesem Abschnitt stehen die RNA Fragmente und deren modifizierte 3'-Enden im Fokus. Die sequenzierte RNA des Modellorganismus E. coli bietet einen beispielhaften Blick auf die Polyadenylierung bei Prokaryoten. Basierend auf der in 4.2.1 beschriebenen gehäuften Repräsentation von nativen 3'-Enden von RNA in den Sequenzierdaten und gestützt durch die PCR-Korrektur (vgl. 4.1) wird hier ein weiteres neuartiges Verfahren dargelegt. Das vorgestellte Konzept ermöglicht zum einen die zuverlässige Schätzung von Fehlerwahrscheinlichkeiten abseits der Enden von Reads und zum anderen ein statistisches Verfahren zur Erkennung von signifikanten Häufungen von modifizierten 3'-Enden.

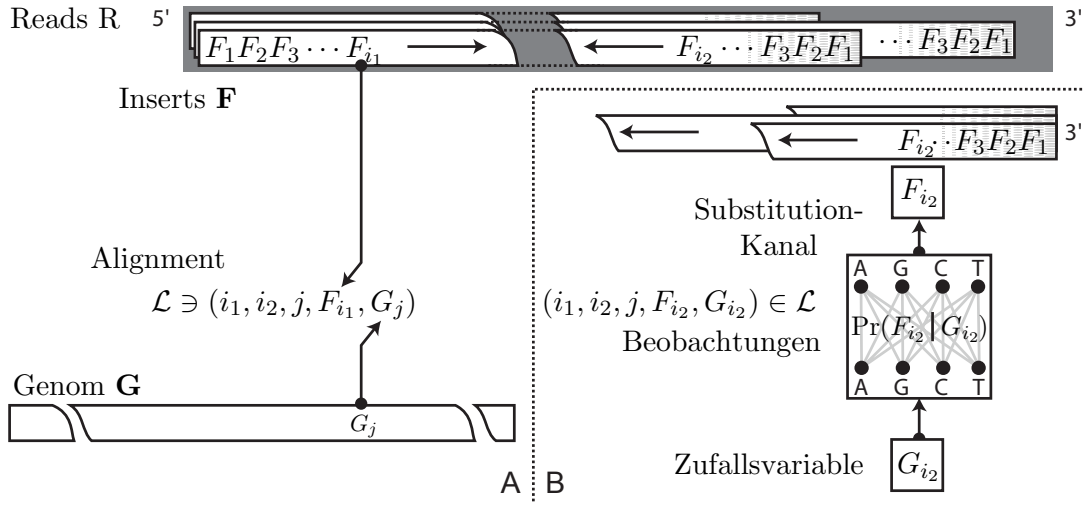

Abb. 4.14 Übergang vom Sequenzalignment zur Modellierung als Zufallsprozess: (A) Jedes Tupel des Alignment-Lattice $\mathcal{L}$ (vgl. Abb. 4.8) beinhaltet für jeweils ein Insert **F** eine explizite Zuordnung von Symbolen F_{i_1} zum Genom **G**. (B) Vernachlässigt man den Kontext (in **F** und **G**) und betrachtet die Positionen i_2 (bezüglich dem 3'-Ende der Inserts) unabhängig voneinander, so impliziert dieser Ansatz Verbundverteilungen von F_{i_2} und G_{i_2} in den Beobachtungen $\mathcal{L}$: auftretende Diskrepanzen im Sequenzalignment werden damit auf Basis von abhängigen Zufallsvariablen beschrieben. Dabei charakterisiert die bedingte Wahrscheinlichkeit $\Pr(F_{i_2}|G_{i_2})$ einen Substitutionskanal relativ zu einem verallgemeinerten Insert **F** und einer Position i_2. Auf Grund der Mindestlänge der Inserts von 15 ist für $i_1, i_2 \in \{1, 2, .., 15\}$ die Beobachtungsmenge einheitlich groß und allumfassend.

Substitutionsmodell mit relativem Bezug zur Position im Insert

Wie im letzten Abschnitt erwähnt ist die Position innerhalb des Reads zu einer definierten Größe bei der Analyse und Bewertung von Sequenzfehlern in der Sequenzierung geworden: Sogenannte Fehlerprofile für Reads gehören zum Standardrepertoire in Publikationen über Fehleranalysen (vgl. [44, 47, 89, 117, 121, 147, 168]). Obwohl sich allgemein eine Korrelation zwischen der Lage eines sequenzierten Nukleotids im Read und der Wahrscheinlichkeit einer Diskrepanz beim Alignment nachweisen lässt, existieren Hinweise, dass die lokale Varianz in den Fehlerprofilen spezifisch durch den Sequenzierzyklus (vgl. Abb. 4.6) beeinflusst wird [93, 117, 139]. Die Verwendung der kurzen Inserts unterschiedlicher Länge erlaubt das im Folgenden aufgeführte Substitutionsmodell (vgl. Konzept 4) mit relativen Bezug auf die 3'-Enden. Die variable Länge ermöglicht bei der Schätzung von positionabhängigen Fehlerwahrscheinlichkeiten zusätzlich eine Dekorrelation von systematischen vom Sequenzierzyklus bedingten Fehlern.

In Abb. 4.14 ist der Übergang vom Sequenzalignment zur Schätzung eines positionsabhängigen Substitutionsmodells dargestellt. Ausgehend von allen Reads, die auf das Referenzgenom abgebildet werden konnten, gilt für die im Alignment-Lattice $\mathcal{L}$ enthaltenen Zuordnungen von Symbolen (F_i, G_j) folgende Vereinfachung: Ausgehend von einer Position i in einem verallgemeinerten Insert **F** wird angenommen, dass eine Verbundverteilung $\Pr(F_i, G_j)$ der Zuordnungen existiert, die unabhängig von einer Position j im Genom ist. Geht man davon aus, dass die Verbundverteilung hinreichend gut geeignet ist, die beobachtbaren fehlerhaften Zuordnungen zu beschreiben, so ist die Modellierung der beteiligten Symbole gleichbedeutend mit der Verwendung unabhängiger Zufallsvariablen und der Beschreibung der Fehler durch einen einfachen Substitutionskanal (vgl. Definition 12). Abhängig von der Bezugsposition in **F**, können die

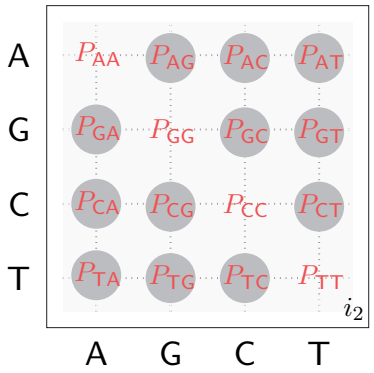

Abb. 4.15 Übergangsmatrix für Position i_2 relativ zum Ende der Inserts: Die Einträge der Matrix entsprechen der Wahrscheinlichkeit $P_{\mathsf{YX}} = \Pr(F_{i_2} = \mathsf{Y} | G_{i_2} = \mathsf{X})$, dass anstatt G_{i_2} im Sequenzalignment ein Symbol F_{i_2} beobachtet wird. Es gilt $P_{\mathsf{XX}} = 1 - \sum_{\mathsf{X} \neq \mathsf{Y}} P_{\mathsf{YX}}$, so kann mit Fokus auf Fehlerereignisse auf die Illustration der Diagonale verzichtet werden.

im Alignment enthaltenen Tupel als Beobachtungen relativ zum Anfang ($i = i_1$) oder zum Ende ($i = i_2$) verstanden werden. Betrachtet man die Modellierung bezogen auf die Position i_1 so ist die Anschauung äquivalent zu den Fehlerprofilen zuvor genannter Publikationen, weil i_1 sowohl im Read als auch im Insert die gleiche Position beschreibt. Für den Bezug auf die Position i_2 relativ zum Ende des Inserts unterscheidet sich die Bedeutung des dargestellten Substitutionsmodells: Ausgehend von Modifikationen am 3'-Ende, modelliert der angenommene Substitutionskanal nicht nur Fehler, die durch die Sequenzierung verursachte wurden, sondern enthält anteilig die Wahrscheinlichkeit für systematische Ersetzungen.

Eine formale Matrix der Übergangswahrscheinlichkeiten (für Position i_2) ist in Abb. 4.15 dargestellt. Zur Schätzung der empirischen Übergangswahrscheinlichkeiten werden die Tupel (i_1, i_2, j, f, g) für $1 \leq i_2 \leq 15$ aus dem Alignment-Lattice $\mathcal{L}$ ausgewählt und nach dem Index i_2 sortiert. Auf dieser Basis werden für jede Position i_2 die Auftrittshäufigkeit $\#_{f,g}^{(i_2)}$ der Tupel (f, g) bestimmt. Die letztendliche Schätzung der bedingten Übergangswahrscheinlichkeit ergibt sich als relative Häufigkeit zu

$$\Pr(F_{i_2} = f | G_{i_2} = g) = \frac{\#_{f,g}^{(i_2)}}{\sum_{f'} \#_{f',g}^{(i_2)}} \quad \text{mit } f, f', g \in \{\mathsf{A}, \mathsf{G}, \mathsf{C}, \mathsf{T}\}.$$

Durch die Beschränkung von i_2 auf die minimale Länge der Inserts ist sichergestellt, dass jeder lokalen Schätzung die gleiche Ausgangsmenge an $\sum_{f,g} \#_{f,g}^{(i_2)}$ Beobachtungen zu Grunde liegt. Alle resultierenden Matrizen sind in Abb. 4.16 beispielhaft für die Sequenzierung 15Z dargestellt. Es sind Übergangswahrscheinlichkeiten in Form von 15 positionsabhängigen Transitionsmatrizen, wobei $i_2 = 1$ die letzte Position der verallgemeinerten Inserts beschreibt.

Wie zu erwarten war, lässt sich für die hinteren Stellen der Inserts ($1 \leq i_2 \leq 5$) eine Präferenz für die Transition hin zum Symbol A erkennen. Beispielsweise erfolgt für die letzte Position im Insert eine Ersetzung von T durch A mit Wahrscheinlichkeit von über 8%. In Analogie dazu lässt sich für alle Sequenzierungen (siehe Abb. A.7 für 11Z/25Z) die Polyadenylierung in den Matrizen erkennen. Des Weiteren ist eine gehäufte Ersetzung zum Symbol T zumindest für die letzte Position zu erkennen. Ob es sich bei dieser Beobachtung um ein technisches Artefakt handelt oder inwieweit das Nukleotid Urcil, als RNA-Pendant zum Nukleotid Thymin, im

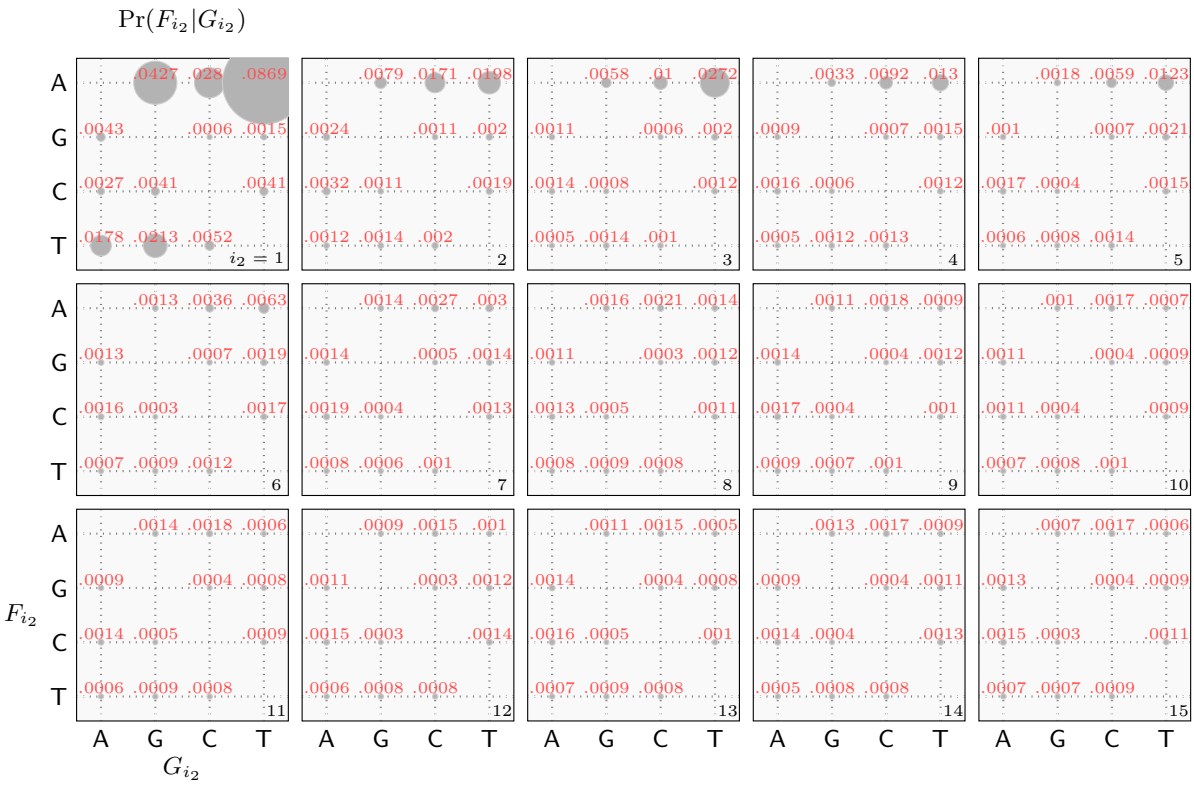

Abb. 4.16 Positionsabhängige Transitionsmatrizen (am 3'-Ende) der Inserts (15Z): Pr($F_{i_2}|G_{i_2}$) beschreibt geschätzte Substitutionswahrscheinlichkeiten. Für die letzten 5 Symbole der Inserts, mit Werten $i_2 \leq 5$, besteht eine starke Präferenz zur Ersetzung durch ein A bzw. T. Zur Mitte der Inserts, $i_2 > 10$, normalisiert sich diese Beobachtung. Das arithmetische Mittel der Transitionsmatrizen mit $i_2 > 10$ dient im Folgenden als Hintergrundmodell.

Wirkungszusammenhang der Polyadenylierung bei Prokaryoten steht, muss von mikrobiologischer Seite her weiter untersucht werden. An dieser Stelle kann nur auf die statistische Häufung innerhalb der Daten hingewiesen werden.

Signifikante Akkumulationen von modifizierten 3'-Enden

Die tatsächliche Einflusslänge der modifizierten 3'-Enden lässt sich durch das hier dargelegte statistische Substitutionsmodell nicht ermitteln und nur schwer abschätzen. Unabhängig von der Bestimmung einer exakten Längenverteilung der Modifikationen bieten die gezeigten Fehlerprofile der Inserts eine hinreichende Statistik zur Erkennung von signifikanten Häufungen von modifizierten 3'-Enden. Eine zuverlässige Identifikation und Annotation der Polyadenylierung im Genom kann dabei als grundlegender Schritt für weiterführende Analysen dienen. Dafür soll im Folgenden ein robustes Testverfahren dargelegt werden, das in ähnlicher Form in [113] zur Erkennung von systematischen Sequenzierfehlern verwendet wurde. Der Kern des hier gezeigten Verfahrens ist ein Hintergrundmodell für die Fehler im Inneren der Reads und abseits der Enden der Inserts. Betrachtet man die Transitionsmatrizen in Abb. 4.16, so scheint ein Einfluss der 3'-Modifikationen im unteren Drittel der Darstellung nicht mehr beobachtbar und

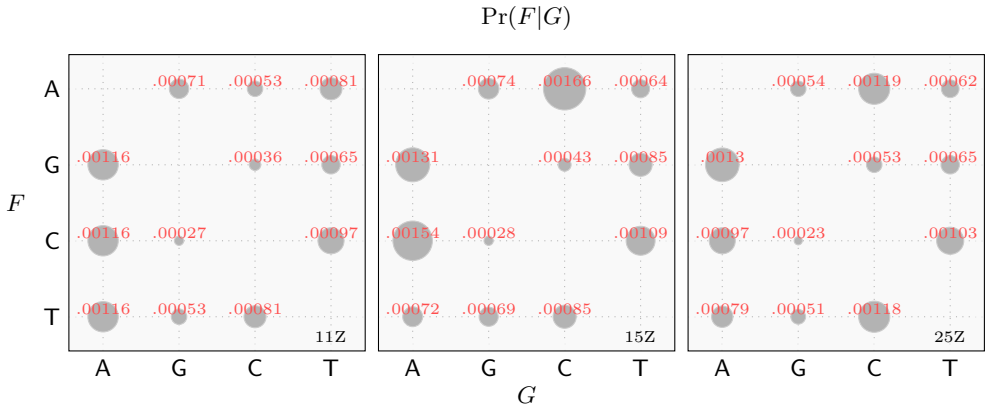

Abb. 4.17 Hintergrundmodell für nicht-systematischen Sequenzfehler. Transitionsmatrizen abseits der Enden der Inserts (für 11Z,15Z und 25Z): $\Pr(F|G)$ beschreibt die geschätzte Wahrscheinlichkeit, dass anstatt G im Sequenzalignment ein Symbol F beobachtet wurde, wenn die letzten 10 Symbole eines jeden Inserts nicht berücksichtigt werden. Bezogen auf den Index i_2 (relative zum 3'-Ende) wurden nur Positionen $10 \leq i_2 \leq 15$ statistisch ausgewertet.

es wird daher angenommen, dass das arithmetische Mittel der Matrizen (für $10 \leq i_2 \leq 15$) die Wahrscheinlichkeiten der nicht-systematischen Sequenzfehler in guter Näherung beschreibt. In Abb. 4.17 sind Transitionsmatrizen des Hintergrundmodells für die Sequenzierungen 11Z,15Z und 25Z dargestellt. Eine Reihe von Publikationen bestätigen vergleichbare Fehlerraten, wie beispielsweise $0.3 - 0.5\%$ in [139], $0.1 - 0.28\%$ [117], $0.28 - 0.3\%$ in [93] oder 0.26% in [113]. Das leicht reduzierte Auftreten von Fehlern der hier gezeigten Schätzungen ist möglicherweise der eingeschränkten Analyse abseits der (als fehlerträchtig bekannten) Enden der Sequenzierung geschuldet (vgl. dazu [89, 117]). Weiter weisen die in Abb. 4.17 gezeigten bedingten Wahrscheinlichkeiten darauf hin, dass erhebliche Unterschiede in den Ersetzungen bestimmter Ausgangssymbole besteht: Die in gewissem Maße über verschiedene Sequenzierungen hinweg reproduzierbare Größen zeigen beispielsweise $\Pr(\mathsf{C}|\mathsf{G}) \leq 0.029\%$ und $\Pr(\mathsf{G}|\mathsf{A}) \geq 0.11\%$.

Zur weiteren Vereinfachung des Hintergrundmodells wird anstelle von 16 unterschiedlichen Substitutionswahrscheinlichkeiten eine mittlere Fehlerwahrscheinlichkeit

$$p_e = \sum_{f \neq g} \Pr(F = f|G = g)\Pr(G = g) \quad \text{mit } F, G, f, g \in \{\mathsf{A}, \mathsf{G}, \mathsf{C}, \mathsf{T}\}$$

treten. Für die Berechnung dieser mittleren Wahrscheinlichkeit der fehlerhaften Zuordnung zweier Nukleotide F und G beim Sequenzalignment wird bezüglich der Symbole G (genomischen Referenz) von einer Gleichverteilung ausgegangen, d. h. $\Pr(G) = 1/4$. Für die Sequenzierungen 11Z, 15Z und 25Z ergeben sich somit abseits der 3'-Enden Fehlerwahrscheinlichkeiten p_e von 0.22%, 0.26% und 0.24%. Basierend auf folgender Annahme wird eine Hypothese formuliert: Ist ein Nukleotid F nicht Resultat eines systematischen Fehlers, wie z. B. eines modifizierten Endes, so ist die Wahrscheinlichkeit für eine Diskrepanz $F \neq G$ bei der Alignierung unabhängig vom Symbol G oder anderen Ereignissen. Basierend auf der Wahrscheinlichkeit p_e lautet die grundlegende Hypothese $\mathcal{H}_0$: *„Das Fehlerereignis ist Bernoulli-verteilt (vgl. Definition 2) mit Parameter p_e."* Diese Hypothese $\mathcal{H}_0$ wird anhand von Zählgrößen alignierter Nukleotide für das Referenzgenom getestet. In 4.3.3 wurde die korrigierte Größe $\tilde{n}_j$ für Nukleotide bereits definiert. Ähnlich der Berechnung der Anzahl $\tilde{n}_j$ von Inserts, die eine exakte Position j im Genom überdecken, werden zur Prüfung der Hypothese $\mathcal{H}_0$ lokaler Verhältnisse von Fehlern

bei den Zuordnungen auf Abschnitten gezählt. Dazu wird das Genom der Länge 5547323 in äquidistante Abschnitte $[j_a, j_a + \epsilon] \subset \{1, 2, .., 5547323\}$ der Größe ϵ geteilt für welche zwei neue Zählgrößen k_a und n_a (für Bereiche anstatt Positionen) definiert werden. Dabei entspricht

$$n_a = \sum_{j \in [j_a, j_a + \epsilon]} \tilde{n}_j$$

der Summe aller Nukleotide welche einen Abschnitt a überdecken. Der Wert für k_a ist die Summe aller Fehlerereignisse im Abschnitt a, wobei für die Zählung ebenfalls die PCR-Korrektur berücksichtigt wird. Zur Reduktion des Tests auf statistisch relevante Bereiche wird die Größe der Abschnitte zu $\epsilon = 10$ gewählt und es werden nur Ausschnitte des Genoms berücksichtigt in welchen $n_a \geq \epsilon$ und $\hat{p}_a = {}^{k_a}/_{n_a} > p_e$ gilt. Somit ergeben sich A relevante Abschnitte gekennzeichnet durch einen neuen Index $a \in \{1, 2, .., A\}$, wobei die Reihenfolge der Indexe keine Bedeutung hat. Für jeden Abschnitt a lässt sich damit ein p-Wert$[H_{0,a}]$ (siehe Konzept 3) unter eine unabhängigen Hypothese $H_{0,a}$ formulieren als

$$\text{p-Wert}[H_{0,a}] = \Pr(K \geq k_a | p_e, n_a) = \sum_{\kappa = k_a}^{n_a} \binom{n_a}{\kappa} p_e^{\kappa} (1 - p_e)^{n_a - \kappa}.$$

Dabei entspricht K der Zufallsvariablen zur Beschreibung der Anzahl der Fehler unter Binomial-Verteilung mit Parametern p_e und n_a. Für sehr kleine Wahrscheinlichkeiten birgt die numerische Auswertung der Summe Probleme gegenüber einer theoretischen analytischen Lösung. Zur Begrenzung der nummerischen Abweichung und zur zuverlässigeren Ordnung von Signifikanzwerten sind folgende Schranken (vgl. Definition 4) hilfreich:

$$\frac{\mathrm{e}^{-n_a \mathrm{d}_{\mathrm{KL}}(\hat{p}_a || p_e)}}{(n_a + 1)^2} \leq \Pr(K \geq k_a | p_e, n_a) \leq \mathrm{e}^{-n_a \mathrm{d}_{\mathrm{KL}}(\hat{p}_a || p_e)} \text{ mit } p_e < \hat{p}_a = {}^{k_a}/_{n_a} < 1.$$

Zur Bestimmung von Abschnitten a, welche eine signifikante Häufung von Fehlern beinhalten, wird auf Basis eines durch Bonferroni-Korrektur (vgl. Definition 11) angepassten Signifikanzniveaus $\alpha_A^* = {}^{\alpha}/_{A}$ ein gemeinsamer Ablehnungsbereich für jeden p-Wert$[H_{0,a}]$ definiert. Das von $\alpha = 0.05$ aus angepasstes Signifikanzlevel begrenzt bei der Verwendung einer gemeinsamen Datenbasis für A unabhängige Hypothesentests die α-Fehler-Kumulierung, d. h. es wird sichergestellt, dass die Wahrscheinlichkeit nur eine einzige Hypothese $H_{0,a}$ fälschlicherweise abzulehnen kleiner ist als $\alpha = 0.05$.

Generell ist es mit dem beschriebenen Verfahren möglich jegliche signifikante Häufung von systematischen Fehlern bei der Abbildung von Sequenzierdaten auf das Genom adaptiv und robust zu klassifizieren (Ähnliches wurde in [113] gezeigt). Mit Fokus auf 3'-Enden ist es angebracht die Analyse auf den relevanten Bereich der Inserts zu beschränken. Für die in Abb. 4.18 dargestellten Ergebnisse des Tests ist eine solche Einschränkung vorgenommen worden: Für die Berechnung der lokalen Verhältnisse ${}^{k_a}/_{n_a}$ wurden lediglich die letzten 5 Nukleotide aller Inserts berücksichtigt, d. h. es wurden nur Symbole F_{i_2} mit $i_2 \leq 5$ verwendet, an welchen die Modifikationen am prägnantesten sind (vgl. Abb. 4.16). Für das Sequenzalignment der Daten von 15Z ergeben sich unter den genannten Kriterien $A = 9533$ Abschnitte im Genom von E. coli, die für den Test berücksichtigt wurden. Die berechneten p-Werte sind in Abb. 4.18 in sortierter Reihenfolge aufgezeigt. Für eine bessere Darstellung wurden die p-Werte in einer log-Skala aufgetragen, sprich ein p-Wert von 10^{-10} ist an der Position -10^1 auf der y-Achse zu finden. Gleiches gilt für das korrigierte Signifikanzniveau, das bei $\log(\alpha_A^*) = -11.317$ zu

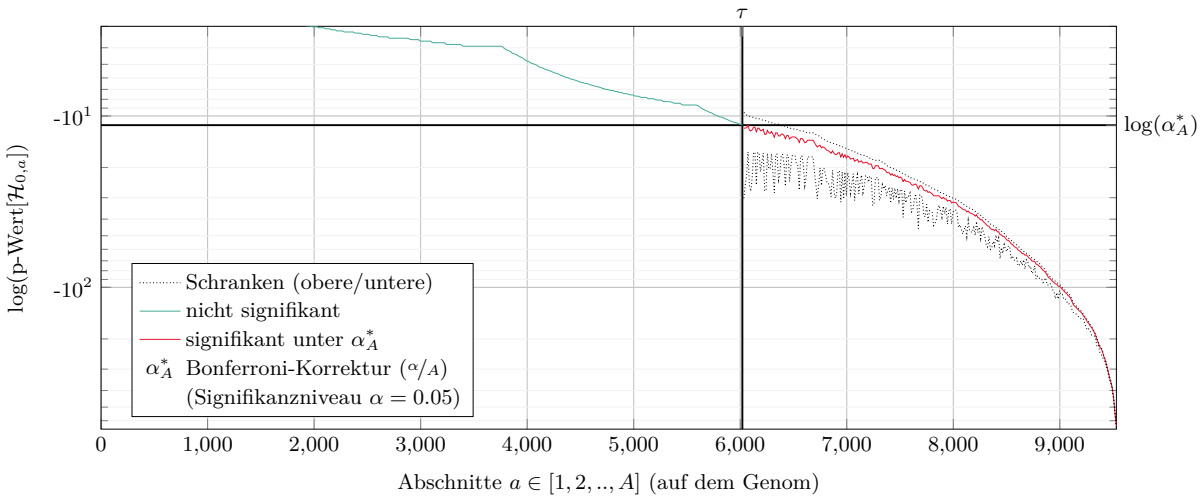

Abb. 4.18 Hypothesentest für signifikante Akkumulation von Fehlern an 3'-Enden (15Z): Logarithmische Signifikanzwerte $-\log(\text{p-Wert}[H_{0,a}])$ (sortiert), für $A = 9533$ Abschnitte a (der Länge $\epsilon = 10$) im Genom von E. coli. Für 3513 Abschnitte $a > \tau$ besteht eine signifikante Häufung von Fehlern, basierend auf dem Signifikanzniveau $\alpha^*_A = \alpha/A$ (Bonferroni-Korrektur) mit $\alpha = 0.05$. Zur Korrektur von numerischen Ungenauigkeiten dienen die angegebenen Schranken. Die Sortierung der Werte erfolgt für $a > \tau$ gemäß der oberen Schranke (relevant für p-Werte $< 10^{-100}$).

finden ist. Zur Korrektur von numerischen Ungenauigkeiten an den kleinsten vorliegenden p-Werten wurden die genannten Schranken bei der Bewertung für alle signifikanten Abschnitte (Abschnitte $a \geq \tau = 6022$) berücksichtigt (Sortierung anhand der oberen Schranke). Relevant wird die Beschränkung der Werte für p-Werte $< 10^{-100}$. Es zeigen sich insgesamt 3513 Bereiche mit signifikanter Modifikation der 3'-Enden.

Anhand des kompletten Sequenzalignment ist es mit einem sogenannten Genome-Browsers (z. B. IGV [138, 161]) möglich eine sorgfältige Einzelprüfung der gegebenen Bereiche durchzuführen. In Abb. 4.19 ist eine repräsentative Auswahl von zwei Abschnitten dargestellt, die durch den Test als signifikante Fehlerhäufung klassifiziert wurden. Welche der tausenden ermittelten Stellen tatsächlich im Wirkungszusammenhang der Polyadenylierung bei E. coli stehen, muss nun durch Expertise im Bereich der Molekularbiologie analysiert und verifiziert werden. Zu diesem Zweck wurden die gewonnenen Daten aller Sequenzierungen (11Z, 15Z und 25Z, vgl. dazu Abb. A.8 und A.9) in Form eines Datensatzes aufbereitet, der eine sukzessive Auswertung der Bereiche im Genom ermöglicht.

4.4 Zusammenfassung und weiterführende Themen

In diesem Kapitel wurden neuartige Zufallsbarcodes vorgestellt und deren Anwendung exemplarisch auf das Illumina TruSeq RNA Protokoll bezogen. Obwohl das innovative Konzept unabhängig von einer bestimmten Plattform einsetzbar ist, wurde zur Evaluation eine konkrete Technologie genutzt. Als Grundlage für die neue Barcode-Technologie wurden sogenannte Barcode-Templates entworfen, die ein Maximum an Flexibilität bei der Gestaltung von zufälligen Oligonukleotiden bereitstellen. Als Schlüsselelement für die Erzeugung von Zufallsbarcodes wurde

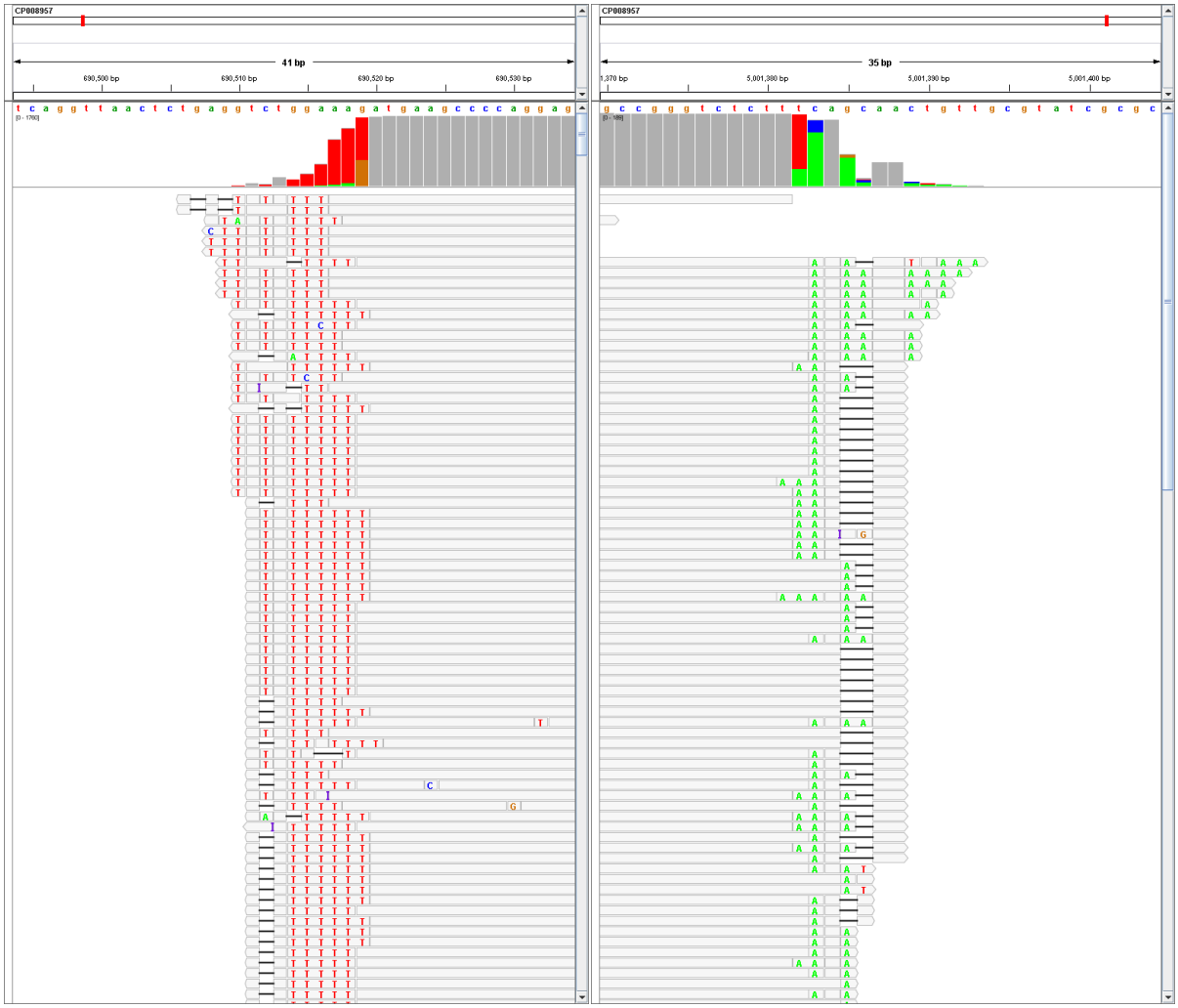

Abb. 4.19 Auswahl von zwei stellvertretenden Abschnitten mit signifikanter Polyadenylierung (15Z): Bildschirminhalt des Genom-Browsers IGV [138, 161] für das Sequenzalignment (Genom von E. coli). Farbige Buchstaben und horizontale Striche kennzeichnen Sequenzfehlern, graue Pfeile sind Übereinstimmungen. Links dargestellt ist die Überdeckung des Rückwärtsstrangs im Bereich der Nukleotide 690505←690515. Das *reverse Komplement* der Polyadenylierung erscheint als poly-T am Anfang in Leserichtung (links). Rechts ist die Polyadenylierung für Position 5001380→5001390 illustriert. Die relativen Häufigkeiten im Balkendiagramm (oben) lassen erkennen, dass es sich bei der Polyadenylierung nicht um eine strikt uniforme Erweiterung der Sequenzen handelt.

eine freie Ligation von Molekülen aufgezeigt, die besondere Anforderungen an die Barcode-Templates stellt. Die Rotationssymmetrie und die Fehlerkorrektur wurden als Kernelemente der rotations-immunen Barcode-Templates vorgestellt. Neben der konkreten theoretischen Parametrisierung zum einen und der experimentellen Erstellung der Zufallsbarcodes zum anderen, wurden besondere Aspekte der Bioinformatik näher erörtert: So wurde die Decodierung, Filterung und Korrektur von PCR-Duplikaten beschrieben sowie theoretische Grenzen der Korrektur hergeleitet. Es wurde ein alternatives Sequenzalignment eingeführt, um eine Analyse von systematischen Modifikationen der RNA-Moleküle an deren 3'-Enden zu ermöglichen. Den Abschluss bildet die Demonstration der entwickelten Konzepte anhand realer Sequenzierexperimente. Zur

Evaluation der Verfahren wurden drei technische Replikate einer RNA-Sequenzierung des Organismus E. coli durchgeführt, bei denen die Anzahl der PCR-Zyklen variiert wurde. Ausgehend von der Analyse der beobachtbaren Molekülkombinationen der Zufallsbarcodes wurde über geschätzte Verteilungen von validen RT-Primern die Korrektur von PCR-Duplikaten bewertet. Des Weiteren wurde das Ausmaß der PCR und deren Korrektur vergleichend für die Sequenzierungen gegenübergestellt. Eine Anomalie hinsichtlich der beobachtbaren Zählgrößen führte daraufhin zur Hypothese der Selbst-Hybridisierung. Abschließend fällt der Fokus auf die modifizierten 3'-Enden. Auf Grundlage eines Substitutionsmodells mit relativem Bezug auf die Inserts wurde ein statistischer Test vorgestellt, der eine robuste Klassifikation von Modifikationen ermöglicht. Die in diesem Kapitel dargelegten Anwendungsszenarien stellen nur eine begrenzte Auswahl der Möglichkeiten dar, die von dem innovativen Konzept ausgehen. Das gezeigte Verfahren auf Basis der Barcode-Templates bietet vielschichtige Ansatzpunkte für weiterführende Themenkomplexe. Dabei sind drei große Bereiche abzugrenzen.

Als Erstes ist die Optimierung der Herstellung der RT-Primer zu nennen. In 4.3.1 wurde die niedrige Effizienz hinsichtlich der Längenselektion der Oligonukleotide genannt. Hierzu sollte eine Revision und Verbesserung der experimentellen Größenselektion erfolgen, um die gezeigte theoretische Größenseparation auch praktisch zu nutzen. Die Gerichtetheit und Spezifität der Polymerisation könnten ferner durch Modifikationen der an der Ligation beteiligten Oligonukleotide vergrößert werden. Auch das Reaktionsgefüge der zugrundeliegenden Kettenpolymerisation der RT-Primer birgt zahlreiche Faktoren zur Optimierung der Ausbeute an funktionalen Zufallsbarcodes. Zur zielgerichteten Bearbeitung dieser Themenbereiche sind weitere Expertisen für diese Art von Polymerisation notwendig.

Ein zweiter Ansatzpunkt neben der Optimierung der dargelegten Verfahren ist die Prüfung und Revalidierung der gezeigten Analysen und den dabei auftretenden Effekten. Dazu wäre zu überprüfen, ob die Verteilung der Zufallsbarcodes in unterschiedlichen Losen der Ligation reproduzierbare Repräsentationen zeigen. Eine vergrößerte Datenerhebung (z. B. mittel Illumina HiSeq-Plattform) ist hierzu sinnvoll. Weitere Sequenzierungen wären auch für die Bestätigung der Hypothese zur Selbst-Hybridisierung nötig. Um das beobachtete Phänomen im Kontext der Zufallsbarcodes zu entschlüsseln und den Effekt zielgerichtet für die Zählung von Molekülen einzusetzen, sollten die experimentellen Pfade der Standardprotokolle der PCR verlassen und gegebenenfalls neue Protokolle entworfen werden.

Der Entwurf neuartiger Experimente ist schließlich die oberste Ebene. Die in 4.2.1 dargelegte neue Sicht auf die 3'-Enden der RNA Moleküle ist hier nur beispielhaft zu nennen. Experimente können mit Fokus auf die vermehrte Sequenzierung der RNA-Enden hin verändert werden. In einem dazu verwandten Kontext könnte der Zweck der Sequenzierung auf die reine Analyse von Zufallsbarcodes angepasst werden. Dies würde zum Zweck der Fehlerschätzung und Kalibrierung erfolgen: Anders als bei gängigen Verfahren zur Quantifizierung von Sequenzfehlern, welche auf einer genomischen Referenz als Symbolquelle und dem Sequenzalignment basieren, könnten synthetische Barcode-Templates mit Fehlerkorrektur einen deutlich transparenteren Ansatz liefern. Wäre sichergestellt, dass die dazu verwendeten Zufallsbarcodes ein hinreichendes Maß an lokaler Unsicherheit bieten, wäre eine Fehlerschätzung möglich, die durchaus mit Konzepten der Kanalschätzung in der Nachrichtentechnik beschreibbar wäre.

Um abschließend den eigentlichen Zweck der Zufallsbarcodes wieder aufzugreifen - die zuverlässige Korrektur von PCR-Duplikaten - so wäre die Einzelzell-Sequenzierung von RNA ein weiteres Einsatzfeld des neuartigen Konzepts. Ein Kernproblem bei der Einzelzell-Sequenzierung

stellt die geringe Masse der Ausgangsmoleküle (< 10 pg RNA bei Eukaryoten, deutlich geringer für Prokaryoten) dar, welche zur Detektion über eine bestimmte Schwelle hinweg amplifiziert werden muss. Für Eukaryoten existieren bereits Protokolle die sich mit dem schwierigen Themenkomplex befassen [71, 81, 134, 144, 158], wobei die reverse Transkription und eine spezielle (erweiterte) PCR meist die Hauptherausforderung stellen. Die herausragende Flexibilität und Diversität der gezeigten Zufallsbarcodes könnte dabei maßgeblich zu neuen Ansätzen der Einzelzell-Sequenzierung für Prokaryoten beitragen.

5

Zusammenfassung und Ausblick

Mit der dargelegten Arbeit wurden zwei neuartige Konzepte zur Erzeugung von Barcodes für Sequenzierverfahren der zweiten Generation vorgestellt. Ob das Prinzip der Watermark Codes, welches ursprünglich für die kontinuierliche Übertragung von sehr langen Codewortblöcken über einen binären Kanal mit Synchronisierungsfehlern entworfen wurde, auch für den Einsatz als Barcodes in der DNA-Sequenzierung taugt, wurde in der vorliegenden Arbeit erstmalig untersucht. Dabei kann die DNA-Sequenzierung durchaus als mustergültige Realisierung eines Kanals mit Synchronisierungsfehlern wie Einfügungen oder Löschungen von Symbolen verstanden werden, für welchen die sehr spezielle Form der Synchronisierung auf Basis des Wasserzeichens ursprünglich konzipiert wurde. Die Sequenzierung an sich schafft jedoch einige Rahmenbedingungen, die hierfür berücksichtigt werden mussten. In Kapitel 3 wurde das Modell der Einbettung von Codeworten in DNA-Templates dargelegt und die essentielle Anpassung der Modelle zur Decodierung gezeigt. Bedingt durch die Anforderung an kurze Codes und weiteren technischen Voraussetzungen an die Sequenz der Barcodes, wurde eine Auswahl an praktisch relevanten Codierschemata vorgestellt und evaluiert. Es konnte gezeigt werden, dass Watermark Codes für den Einsatz in der Sequenzierung prinzipiell geeignet sind. Auf Basis der probabilistischen Erkennung von Codeworten bietet das gezeigte Verfahren eine neuartige Option zu rein distanzbasierten Ansätzen zur robusten Markierung von DNA-Fragmenten im Multiplexing. Um die dargelegte Alternative optimal in der Praxis umzusetzen und Modellparameter gezielt anpassen zu können, bleibt letztlich noch eine zuverlässige Schätzung des Sequenzierkanals als weiterführendes Thema.

Betrachtet man die aktuellen Entwicklungen im Bereich der Sequenziertechnologie, so existieren bereits Ansätze der sogenannten dritten Generation: Basierend auf der individuellen Sequenzierung von DNA-Molekülen produzieren diese direkte Reads, die im Wesentlichen weniger von experimentellen Modifikationen der Moleküle abhängen und damit unter Umständen eine geringere experimentelle Varianz aufweisen. Eine vielversprechende beispielhafte Methode ist unter dem Begriff Nanoporen-Sequenzierung (Produktname *MinION* der Oxford Nanopore Technologies) bekannt, bei welcher eine künstliche Membran (aus Nanoporen) genutzt wird, um auf die Sequenz eines die Pore passierenden Polymers rückzuschließen. Sicherlich steckt diese Technik noch in ihren Kinderschuhen, was bisherige Analysen der Zuverlässigkeit der Ergebnisse zeigen [10, 83, 116]. Mit einer rapiden Weiterentwicklung dieser Technologie ist jedoch sicherlich zu rechnen. Ob die nächste Generation an Sequenziergeräten wegen extrem kleiner Fehlerdichten den Einsatz von fehlerkorrigierenden Barcodes gänzlich unnötig macht oder ob das Multiplexing durch weitere Technologiesprünge abgelöst wird, bleibt aus heutiger Sicht fraglich. Bis sich eine solche Wende in der Praxis vollzieht, ist der Einsatz von innovativen Konzepten für die bestehende und bereits etablierte Technologie angebracht.

In dieser Arbeit wurde zusätzlich ein vollkommen neuartiges Konzept für Barcode-Templates entworfen, welches ein Maximum an Flexibilität bei der kostengünstigen Gestaltung von zufälligen Oligonukleotiden bereitstellt. Kapitel 4 bietet hierzu die detaillierte Konzeption der Barcode-Templates zum Einsatz für das Illumina Protokoll, eine der momentan etablierten Technologien zur RNA-Sequenzierung. Die Rotationssymmetrie und die Integration von Fehlerkorrektur sind hervorzuhebende Eigenschaften der vorgestellten rotations-immunen Barcode-Templates, die eine freie Ligation von Zufallsbarcodes mit Fehlerschutz ermöglichen. Neben der experimentellen Implementierung von Barcodes für die Illumina-Technologie wurde unter anderem die Anwendung zur Korrektur von quantitativen Effekten der PCR evaluiert. Die Notwendigkeit der PCR für Sequenzierverfahren zweiter Generation motiviert dabei den Einsatz von Zufallsbarcodes für quantitative Ansätze wie der Transkriptom-Analyse.

Möchte man die Implementierung von Zufallsbarcodes in den Kontext der aktuellen Entwicklungen im Bereich der RNA-Sequenzierung einordnen, so zeigt dieses Manuskript ein neuartiges Konzept für eine konventionelle Technologie und gegenwärtige Anforderungen. Wagt man hingegen einen Blick in die Richtung von zukünftigen Technologien, so ist es sehr gut möglich, dass sich Rahmenbedingungen und Erfordernisse im Bereich der Sequenzierung von DNA und RNA bald neu positionieren. Die Sequenzierung der nächsten Generation spezifiziert unter anderem bereits die Verarbeitung von Molekülen ohne den Einsatz der PCR und damit verbundenen Modifikationen (siehe beispielsweise [125, 126]). Die direkte RNA-Sequenzierung [145] bietet hierzu einen beispielhaften Ansatz wie quantitative Analysen des Transkriptoms gänzlich ohne den fehlerträchtigen Schritt über cDNA oder PCR erfolgen können. Kommende Entwicklungen dieser Art könnten den Zufallsbarcodes die Motivation für den aktuellen Zweck entziehen. Nichtsdestotrotz bietet das vorgestellte Konzept der zufälligen Oligonukleotiden mit Fehlerkorrektur nützliche Eigenschaften für zukünftige Anwendungen abseits der Verwendung als Barcodes. Hinsichtlich einer Kanalschätzung für die Sequenzierung bietet das dargestellte Konzept beispielsweise eine kostengünstige Methode, um zufällige (und lokal strukturierte) Testsequenzen zu erzeugen, die neben einer großen Anzahl variabler Symbolkombinationen und damit einem hohem Informationsgehalt der Symbole zusätzlich eine Fehlerkorrektur ermöglichen.

Anhangsverzeichnis

A

Zufallsbarcodes und deren Anwendung

Der folgende Abschnitt enthält einzelne Aspekte und Ergänzungen zum Kapitel 4, der Zufallsbarcodes und deren Anwendung in der RNA-Sequenzierung.

A.1 Technische Sequenzen

Die Menge der berücksichtigten technischen Sequenzen $\mathcal{T}$ des TruSeq Protokolls (siehe [78]) sind im Folgenden aufgelistet:

RA5: `5'-GTTCAGAGTTCTACAGTCCGACGATC`
RA3: `5'-TGGAATTCTCGGGTGCCAAGG`
RTP: `5'-CCTTGGCACCCGAGAATTCCA`
RP1: `5'-AATGATACGGCGACCACCGAGATCTACACGTTCAGAGTTCTACAGTCCGA`
RPIX: `5'-CAAGCAGAAGACGGCATACGAGATxxxxxxGTGACTGGAGTTCCTTGGCACCCGAGAATTCCA`

mit `xxxxxx` als Index 1-12 der TruSeq Adapter
`ATCACG`
`CGATGT`
`TTAGGC`
`TGACCA`
`ACAGTG`
`GCCAAT`
`ATCACG`
`CAGATC`
`ACTTGA`
`GATCAG`
`TAGCTT`
`GGCTAC`
`CTTGTA`

Die Anwendung der Sequenzen ist in Abb. 4.2 auf Seite 78 illustriert.

A.2 Asymmetrische Phosphorylierung

Prinzipiell bestehen drei Möglichkeiten der Phosphorylierung (eine symmetrische und zwei asymmetrische) der in Abschnitt 4.1.2 mit $\mathcal{B}$ und $\mathcal{B}^c$ bezeichneten Moleküle. Die symmetrische

Phosphorylierung und eine homogene Ligation sind in Abb. 4.3 auf Seite 79 dargestellt. Eine asymmetrische Phosphorylierung resultiert in einer uneinheitlichen Ligationsreaktion, welche in Abb. A.1 modellhaft dargestellt ist.

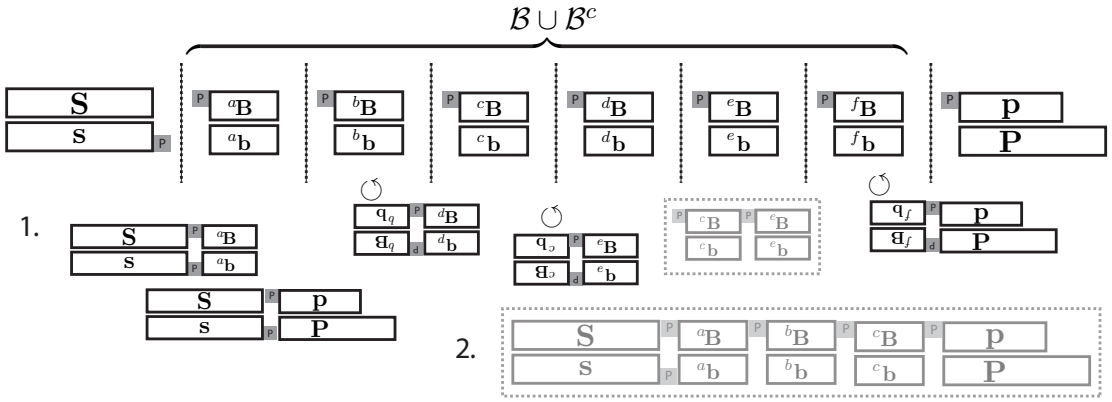

Abb. A.1 Problem der asymmetrischen Phosphorylierung: 1. Präferenz der Bildung von Dimeren ohne offener Phosphorylierung; 2. Ligation der beabsichtigten Strukturen (ausgegraut) ist weniger effizient.

Bei einsetzender Ligation und der beginnenden Bildung von Dimeren besteht eine Präferenz zur Bindung zweier phosphorylierten 5'-Enden der Oligonukleotide. Neben der Bildung von RT-Primer-Strukturen ohne Barcode-Templates entstehen somit Dimere als Zwischenprodukte, welche wegen fehlender offener Phosphorylierung nicht effektiv an einer weiterführenden Ligationsreaktion beteiligt sind. Somit ist eine effektive Ligation der beabsichtigten RT-Primer im asymmetrischen Kontext nicht möglich. Ein experimenteller Nachweis, der Rückschlüsse auf die unterschiedliche Effizienz der beschriebenen Verfahren zulässt, ist Anhang A.6 zu entnehmen.

A.3 Problematische Oligonukleotid-Konstellationen im RT-Primer

In Abb. A.2 sind wesentliche problematische Konstellationen von Sequenzen illustriert, welche zu negativen Nebeneffekten bei der Anwendung der Barcode-Templates führen können:

1. Zeigen zwei Barcode-Templates ${}^{i}\mathbf{B}$ und ${}^{j}\mathbf{B}$ einen Hamming-Abstand von d_h können lediglich $t = \lfloor (d_\mathrm{h}-1)/2 \rfloor$ Substitutionen bei der Decodierung eines fehlerhaften Templates ${}^{k}\mathbf{B}$ korrigiert werden.
2. Homologe Abschnitte in Kombinationen von Sequenzen können
   a) zu Problemen bei der Hybridisierung von Doppelsträngen und der Ligation des RT-Primers führen, oder
   b) Schleifenbildung im einsträngigen Zustand eines Moleküls verursachen.

   Die daraus resultierende Sekundärstruktur (räumliche Molekülstruktur) der Sequenzen kann den gleichförmigen Ablauf von Reaktionen im Protokoll verhindern.

Auf eine Darstellung der Probleme, die durch den GC-Gehalt (Ungleichgewicht an Nukleotide Guanin G und Cytosin C) oder lange Homopolymere (Folgen identischer Nukleotide) verursacht

werden, ist an dieser Stelle verzichtet worden. Die Publikationen [29, 30, 32, 48, 54, 153] dienen als Referenz für die Motivation der notwendigen Bedingungen für Oligonukleotide.

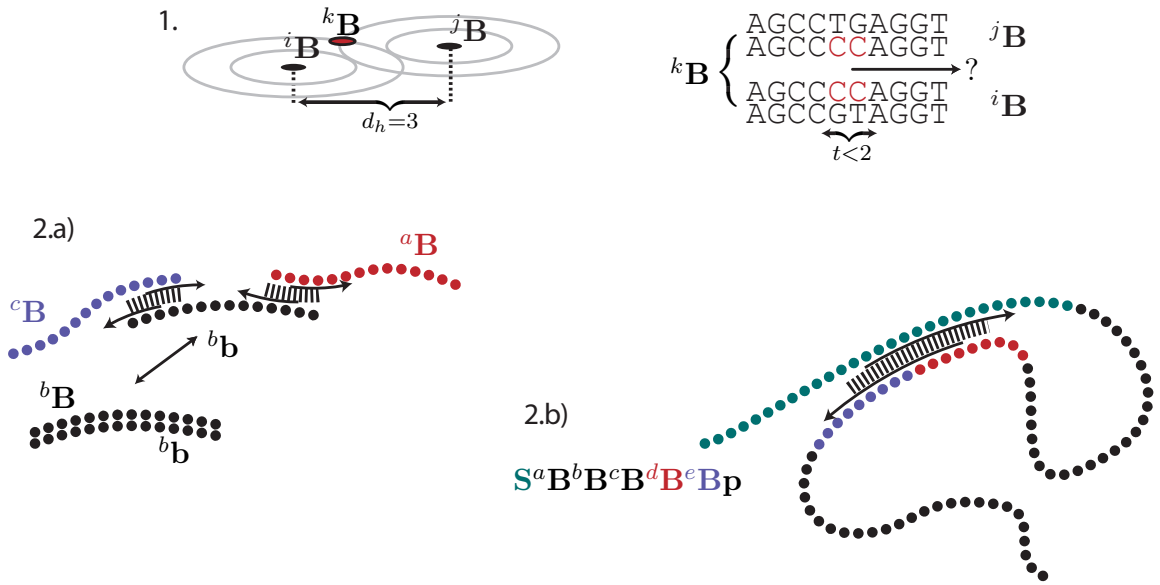

Abb. A.2 Problematische Konstellationen von Sequenzen: 1. Niedriger Hamming-Abstand impliziert niedrige Robustheit bei der Decodierung von Codeworten; 2.a) Treten ähnliche Sequenzen in den Barcode-Templates auf (normale/revers komplementäre Orientierung) kann die Entstehung der gewünschten DNA Moleküle behindert werden; 2.b) Schleifenbildung bei Auftreten von sogenannten invertierten Repeats im Kontext des gesamten Fragments in der Sequenzier-Library kann zu stabilen Sekundärstrukturen führen. Molekulare Wechselwirkungen sind als Leiterstruktur dargestellt, Orientierung der Sequenzen als Pfeil: 3'-Ende → 5'-Ende.

A.4 Such-Algorithmus für Barcode-Templates

Die in Alg. A.1 dargestellte Prozedur zur Auswahl geeigneter Barcode-Templates beinhaltet die folgenden Variablen, Parameter und Funktionen:

Parameter und Variablen:

l	Länge der gesuchten Codeworte
$\mathcal{P}$	Anforderungen in Form der folgenden Parameter:
d_h	Minimaler Hamming-Abstand zwischen Codeworten
$\mathcal{T}$	Menge an technischen Sequenzen
l_h	Maximale Länge (Lauflänge) von identischen Symbolen
l_s	Maximale Länge an Übereinstimmung einer Sequenz zu sich selbst
l_t	Maximale Länge an Übereinstimmung mit Sequenzen aus $\mathcal{T}$
$\underline{r_\mathrm{GC}}$	Untergrenze und
$\overline{r_\mathrm{GC}}$	Obergrenze für den GC-Gehalt
N_a	Anzahl von Iterationen (außen)
N_i	Anzahl von Iterationen (innen)
$\mathcal{M} = \{\mathsf{A},\mathsf{G},\mathsf{C},\mathsf{T}\}^l$	Menge aller Codeworte der Länge l
$\mathbf{B}, \mathbf{B}'$	Ein ausgewähltes Codewort $\mathbf{B} \in \mathcal{B}$ und dessen reverses Komplement $\mathbf{B}' \in \mathcal{B}^c$
$\mathbf{X}, \mathbf{Y}, \mathbf{Z}$	Variable Sequenzen zur Kombination von Teilsequenzen
$\mathcal{B}, \mathcal{B}_i$	Untermenge (Auswahl) von Codeworten

Operatoren:

$[\mathbf{XY}]$	Konkatenation von zwei Sequenzen $\mathbf{X}$ und $\mathbf{Y}$
`rand.get`$(\mathcal{M})$	Zufällige Auswahl eines Elements aus $\mathcal{M}$. Keine Mehrfachauswahl von gleichen Elementen in einer Serie.
`rand.reset`	Initialisieren der Funktion zur zufälligen Auswahl
`revCompl`$(\mathbf{X})$	reverse Komplement des Sequenz $\mathbf{X}$
`hDist`$(\mathbf{X}, \mathbf{Y})$	Hamming-Abstand zweier Sequenzen $\mathbf{X}$ und $\mathbf{Y}$
`runLength`$(\mathbf{X})$	Ermittelt die höchste Anzahl an aufeinanderfolgenden identische Symbolen in $\mathbf{X}$.
`isRepeat`$(\mathbf{Z}, l)$	Prüft, ob eine Teilsequenz der Länge l von $\mathbf{Z}$ wiederholt in $\mathbf{Z}$ auftritt (ein *Repeat* vorliegt): Ausgeschlossen vom Test werden identisch Positionen oder Subsequenzen die Bestandteil von Wiederholungen in $\mathbf{Z}$ sind, z. B. bei $\mathbf{Z} = [\mathbf{XX}]$.
`homology`$(\mathbf{X}, \mathbf{Y})$	Ermittelt die Länge der längsten Subsequenz von $\mathbf{X}$, die in $\mathbf{Y}$ enthalten ist.

Subfunktionen:

`validSelf`	Prüft, ob ein Codewort (in Kombination mit sich selbst) die in den Parametern $\mathcal{P}$ spezifizierten Anforderungen erfüllt.
`validMutual`	Prüft, ob ein Codewort in Kombination mit Codeworten aus einer bereits existierenden Menge die Anforderungen erfüllt.

Anmerkungen:

Die Anzahl der Iterationen N_i (der inneren Schleife) bestimmt die Tiefe der Suche in 4^l Elementen. Lässt sich die komplette Teilmenge $\mathcal{M}'$ an Sequenzen, welche durch `validSelf` nicht verworfen wird zwischenspeichern, so lässt sich der Rechenaufwand reduzieren. Sei $N_\mathrm{i} = |\mathcal{M}'|$ die Größe der genannten Teilmenge, so repräsentiert $\{\mathbf{B}\} \leftarrow$ `rand.get`$(\mathcal{M}')$ eine zufällige Abfolge aller vorab gefilterten Sequenzen. Die Anzahl der Iterationen N_a gibt die Breite der Suche an. Ist es nicht möglich durch alle Abfolgen $\{\mathbf{B}\}$ aus $\mathcal{M}'$ zu iterieren, lässt sich durch N_a der Aufwand beschränken. Die randomisierte Überprüfung gleicher Abfolgen sollte dabei vermieden

werden. Die Anzahl der Aufrufe kann insgesamt auf $N_\mathrm{a} \cdot N_\mathrm{i}$ beschränkt werden. Die Berechnung von `validSelf` skaliert linear mit 4^l, `validMutual` hingegen mit $N_\mathrm{a}|\mathcal{B}|$, wobei $|\mathcal{B}|$ der Menge an letztendlich gefundenen Barcode-Templates entspricht. Das Resultat der Suche sind möglichst große Mengen an Codeworten, welche die geforderten Restriktionen erfüllen.

Alg. A.1 Such-Algorithmus

Input : $l, n, N, \mathcal{P}$
Output : $\mathcal{B}$

for $i \leftarrow 1$ **to** N_a **do**
  $\mathcal{B}_i \leftarrow \emptyset$
  `rand.reset()`
  * **for** 1 **to** N_i **do**
    $\mathbf{B} \leftarrow$ `rand.get`$(\{\mathcal{M} = \mathsf{A,G,C,T}\}^l)$
    $\mathbf{B}' \leftarrow$ `revCompl`$(\mathbf{B})$
    if *not* `validSelf`$(\mathbf{B}, \mathbf{B}', \mathcal{P})$ **then** continue *↰
    if *not* `validMutual`$(\mathbf{B}, \mathbf{B}', \mathcal{P}, \mathcal{B}_i)$ **then** continue *↰
    $\mathcal{B}_i \leftarrow \mathcal{B}_i \cup \{\mathbf{B}, \mathbf{B}'\}$
$\mathcal{B} \leftarrow \arg\max_{\mathcal{B}_i}\{|\mathcal{B}_i|\}$

Prüfprozedur `validSelf`

Input : $\mathbf{B}, \mathbf{B}', \mathcal{P} : \{d_\mathrm{h}, \mathcal{T}, l_\mathrm{h}, l_\mathrm{s}, l_\mathrm{t}, \underline{r_\mathrm{GC}}, \overline{r_\mathrm{GC}}\}$
Output : **valid** (init as *false*)

if $\mathtt{gc}(\mathbf{B}) < \underline{r_\mathrm{GC}}$ *or* $\mathtt{gc}(\mathbf{B}) < \overline{r_\mathrm{GC}}$ **then** return
if `hDist`$(\mathbf{B}, \overline{\mathbf{B}'}) < d_\mathrm{min}$ **then** return
foreach $\mathbf{X}, \mathbf{Y} \in \{\mathbf{B}, \mathbf{B}'\}$ **do**
  if `runLength`$([\mathbf{XY}]) > l_\mathrm{h}$ **then** return
  if `isRepeat`$([\mathbf{XY}], l_\mathrm{s})$ **then** return
  foreach $\mathbf{Z} \in \{\mathcal{T}, \mathtt{revCompl}(\mathcal{T})\}$ **do**
    if `homology`$([\mathbf{XY}], \mathbf{Z}) > l_\mathrm{t}$ **then** return
valid $\leftarrow$ *true*

Prüfprozedur `validMutual`

Input : $\mathbf{B}, \mathbf{B}', \mathcal{P} : \{d_\mathrm{h}, \mathcal{T}, l_\mathrm{h}, l_\mathrm{s}, l_\mathrm{t}\}, \mathcal{B}_i$
Output : **valid** (init as *false*)

foreach $\mathbf{X} \in \mathcal{B}_i$ *and* $\mathbf{Y} \in \{\mathbf{B}, \mathbf{B}'\}$ **do**
  if `hDist`$(\mathbf{X}, \mathbf{Y}) < d_\mathrm{min}$ **then** return
  if `runLength`$([\mathbf{XY}]) > l_\mathrm{h}$ **then** return
  if `isRepeat`$([\mathbf{XY}], l_\mathrm{s})$ **then** return
  foreach $\mathbf{Z} \in \{\mathcal{T}, \mathtt{revCompl}(\mathcal{T})\}$ **do**
    if `homology`$([\mathbf{XY}], \mathbf{Z}) > l_\mathrm{t}$ **then** return
valid $\leftarrow$ *true*

A.5 Konkrete Barcode-Templates

Für die Suche aus Abschnitt A.4 wurden folgende Parameter verwendet:
$l = 9, d_{\mathrm{h}} = 5, \underline{r_{\mathrm{GC}}} = 40\%, \overline{r_{\mathrm{GC}}} = 60\%, l_{\mathrm{h}} = 3, l_{\mathrm{s}} = 3, l_{\mathrm{t}} = 6$ und $\mathcal{T}$ wie in A.1 beschrieben.

Resultat der Algorithmen ist folgende Menge aus 44 Barcode-Templates $\mathcal{B}$:

GAGCGTTTA, TAAACGCTC, CCTATGTGC, GCACATAGG, ATGGTGTCT, AGACACCAT, GAGCAACAG, CTGTTGCTC, CGAATACCT, AGGTATTCG, TTACCGAAG, CTTCGGTAA, CGCTTCTAT, ATAGAAGCG, GTTGTGGAC, GTCCACAAC, AGGGCAAAG, CTTTGCCCT, TAAGGAGAC, GTCTCCTTA, TAACCCGCT, AGCGGGTTA, TAAGAGTGG, CCACTCTTA, AGGCTTGAC, GTCAAGCCT, AGCATCCTC, GAGGATGCT, TAGAGCCAT, ATGGCTCTA, CCTTTAGCG, CGCTAAAGG, GTTAGACTC, GAGTCTAAC, ATTTCGTGG, CCACGAAAT, CCTGATAAC, GTTATCAGG, TACTGGTCT, AGACCAGTA, GCGGAGTAA, TTACTCCGC, TACCCTTGC, GCAAGGGTA

Eigenschaften der Barcode-Templatess:

Profil des GC-Gehalt	$0_0, 0_1, 0_2, 0_3, 24_4, 20_5, 0_6, 0_7, 0_8, 0_9$ [$\#\mathrm{CW}_{\#\mathrm{CG}}$]
min. GC-Gehalt	44.44 %
max. GC-Gehalt	55.55 %
Max. $d_{\mathrm{h}} = 9$ z. B. für	TTACCGAAG $\leftrightarrow$ GAGGATGCT
Min. $d_{\mathrm{h}} = 5$ z. B. für	AGACACCAT $\leftrightarrow$ AGACCAGTA
Profil für Homopolymere	$0_0, 0_1, 32_2, 12_3, 0_4, 0_5, 0_6, 0_7, 0_8, 0_9$ [$\#\mathrm{CW}_{\mathrm{len}}$]
Profil für Hamming-Abstand	$1_0, 8.5_5, 9.5_6, 10.8_7, 9.2_8, 5.1_9$ [mttl.$\#\mathrm{CW}_{d_{\mathrm{h}}}$]
Profil für Editierdistanz	$1_0, 0.4_3, 2_4, 12.2_5, 12.8_6, 11_7, 4.3_8, 0.4_9$ [mttl. $\#\mathrm{CW}_{d_{\mathrm{e}}}$]

Die Profile bestehen aus der (mittleren) Anzahl an Barcode-Templates (Codeworten) eines bestimmten Wertes (im Subskript). Für das Profil der Homopolymere ist das Maß die Lauflänge identischer Symbole, für die Profile über die Distanzen entspricht das Subskript dem jeweiligen Wert der Distanzfunktion zweier zufälliger Codeworte (vgl. dazu Definition 19).

Die durch die Unterstreichung xxxxxxxxx gekennzeichneten 4 Templates dienen zur Erzeugung

- des Spacers (Spacer):
  5'-TGCCCT GAGCGTTTA-3' als $\mathbf{S} = S_1 S_2 S_3 .. S_m\rangle$ mit $m = 15$
- des Komplements (RTP*):
  3'-ACGGGA CTCGCAAAT-5' als $\mathbf{s} = \langle s_m .. s_3 s_2 s_1$

Vergleiche dazu Abb. 4.2. Zusätzlich existieren noch die Sequenzen:

- Primer RTP :
  3'-ACCTTAAGAGCCCACGGTTCC-5' als $\mathbf{P} = \langle P_n .. P_3 P_2 P_1$ mit $n = 21$
- und das verkürzte Komplement :
  5'-TGGAATTCT-3' als $\mathbf{p} = p_1 p_2 .. p_o\rangle$ mit $o = l = 9$

A.6 Freilaufende Kettenbildung von Barcode-Templates

Zur Einschätzung und Veranschaulichung der Reaktionsgeschwindigkeit (Effizienz) der Ligation wurde eine Reihe identischer Experimente zur Ligation von Barcode-Templates (ohne Beteiligung anderer terminierender Oligonukleotide) durchgeführt. Dabei wurde die Reaktion lediglich nach unterschiedlicher Laufzeit chemisch angehalten und die Längenverteilung im Reaktionsprodukt mittels Gelelektrophorese sichtbar gemacht. In Abb. A.3 sind zwei Gele zu sehen, welche aus Barcode-Templates mit unterschiedlicher Phosphorylierung resultieren. Neben dem Größenstandard (gekennzeichnet durch $*$) zeigen Banden die Verteilung der entstandenen x-mere (Molekülketten mit x Barcode-Templates `B`) an. Die niedrige Effizienz des asymmetrischen Konzepts ist deutlich erkennbar: Nach 20 Stunden ist keine Kettenstruktur länger als drei Oligonukleotide im Gel nachweisbar (links). Für das symmetrische Verfahren (recht) ist nach 10 Minuten eine deutliche Dispersion der Verteilung zu immer länger wachsenden Ketten zu erkennen. Von einer freilaufenden und effizienten Kettenpolymerisation ist im symmetrischen Kontext auszugehen.

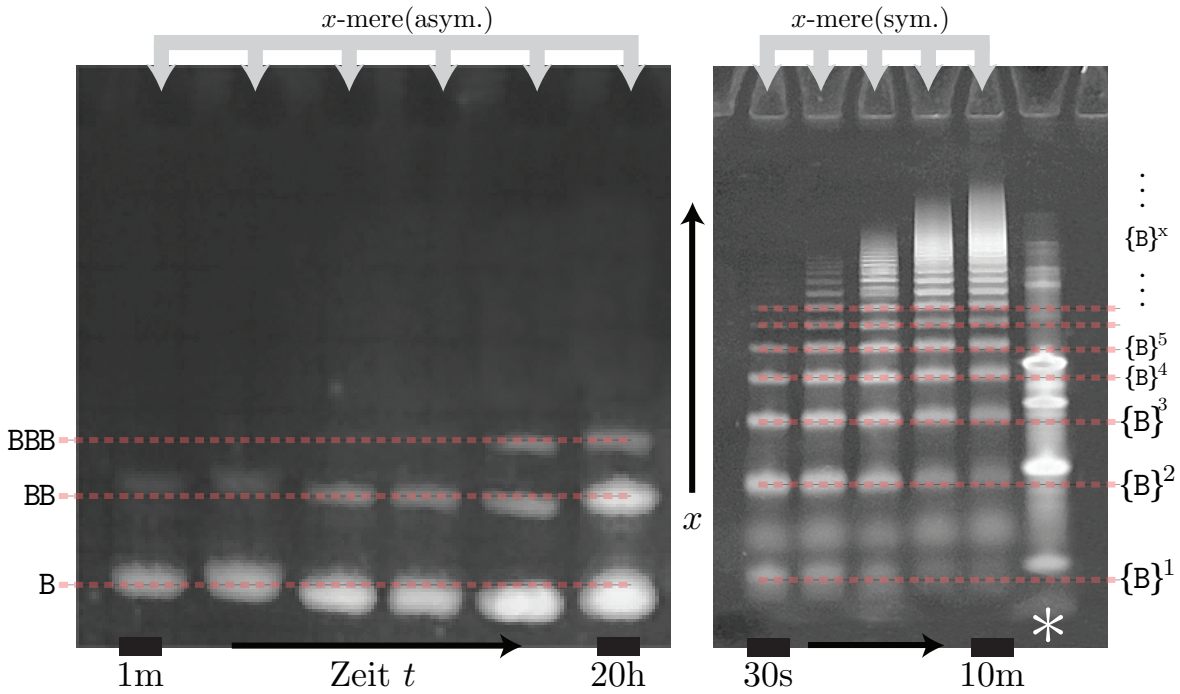

Abb. A.3 Freilaufende Ligation von Barcode-Templates mit asymmetrischer/symmetrischer Phosphorylierung (links/rechts), nach [94]: Längenverteilung der x-mere (Molekülketten mit x Barcode-Templates `B`) im Reaktionsprodukt durch Gelelektrophorese (vgl. Begriff 12) nach unterschiedlicher Reaktionslaufzeit (Größenordnungen der Reaktionsdauer). Neben den Ligationsprodukten wurde parallel ein Größenstandard (durch $*$ gekennzeichnet) verwendet um x-mere zu identifizieren.

A.7 Decodierung und alternatives Alignment

Alg. A.4 BITAP.HAMMING

Input : $\mathbf{P} = P_1P_2..P_m, \mathbf{T} = T_1T_2T_3..T_n, k$ // Kosten als Hamming-Abstand
Output : $\mathcal{I}$
$\mathcal{I} \leftarrow \emptyset$
$\mathbf{R}', \mathbf{R}'' \leftarrow \mathbf{0} \in \mathcal{W}^k$

for $\pi^{\text{end}} \leftarrow 1$ **to** n **do**
  for $\kappa \leftarrow 0$ **to** k **do**
    $\mathbf{R}'[\kappa] = \phi_{\mathbf{P}}(T_{\pi^{\text{end}}})$ `and` $\mathbf{R}''[\kappa] \ll 1$
    if $\kappa - \kappa_{\text{s}} \geq 0$ **then** $\mathbf{R}'[\kappa] = \mathbf{R}'[\kappa]$ `or` $\mathbf{R}''[\kappa - 1] \ll 1$
    if $\mathbf{R}'[\kappa]|_m = 1$ **then**
      $\pi^{\text{start}} = \pi^{\text{end}} - m$ // Anfang des Alignments
      $\mathcal{I} \leftarrow \mathcal{I} \cup \{(\pi^{\text{start}}, \pi^{\text{end}}, \mathbf{P})\}$ // Anfang-/Endposition d. Alignments & Zuordnung
  `flip`$(\mathbf{R}', \mathbf{R}'')$

Variante des Bitap-Algorithmus (vgl. Alg. 2.1): Zur Suche aller Vorkommen von Mustern $\mathbf{P}$ in $\mathbf{T}$ mit maximalem Hamming-Abstand k werden Kosten $\kappa_{\text{s}} = 1, \kappa_{\text{i}} = \kappa_{\text{d}} = 0$ definiert wodurch sich die bit-weisen Operationen reduzieren. Da keine Einfügungen und Löschungen berücksichtigt werden ist eine eindeutig Definition der Anfangspositionen π^{start} des Sequenzalignment möglich. Die Rückgabe enthält alle Begrenzungen der Einbeschreibungen von $\mathbf{P}$ in $\mathbf{T}$ und die Sequenz $\mathbf{P}$.

Alg. A.5 BITAP.LASSO

Input : $\mathbf{P} = P_1P_2..P_m, \mathbf{T} = T_1T_2T_3..T_n, k_{\max}$ // Kosten als Editierdistanz
Output : $\mathcal{I}$
$\kappa \leftarrow k_{\max}$ // variable maximale Kosten
$\mathcal{I} \leftarrow \emptyset$
$\mathbf{R}', \mathbf{R}'' \leftarrow \mathbf{0} \in \mathcal{W}^k$

for $\pi^{\text{end}} \leftarrow 1$ **to** n **do**
  for $\kappa \leftarrow 0$ **to** k **do**
    $\mathbf{R}'[\kappa] = \phi_{\mathbf{P}}(T_{\pi^{\text{end}}})$ `and` $\mathbf{R}''[\kappa] \ll 1$
    if $\kappa - \kappa_{\text{i}} \geq 0$ **then** $\mathbf{R}'[\kappa] = \mathbf{R}'[\kappa]$ `or` $\mathbf{R}''[\kappa - 1]$
    if $\kappa - \kappa_{\text{s}} \geq 0$ **then** $\mathbf{R}'[\kappa] = \mathbf{R}'[\kappa]$ `or` $\mathbf{R}''[\kappa - 1] \ll 1$
    if $\kappa - \kappa_{\text{d}} \geq 0$ **then** $\mathbf{R}'[\kappa] = \mathbf{R}'[\kappa]$ `or` $\mathbf{R}'[\kappa - 1] \ll 1$
  * **for** $k_{\min} \leftarrow 0$ **to** k **do**
    if $\mathbf{R}'[\kappa]|_m = 1$ **then**
      $k \leftarrow k_{\min}$ // Reduktion d. maximalen Kosten
      $\mathcal{I} \leftarrow \{(\pi^{\text{end}}, k_{\min})\} \cup \{(\pi_*^{\text{end}}, k) \in \mathcal{I} : k = k_{\min}\}$ // Reduktion d. Alignments
      break *↰
  `flip`$(\mathbf{R}', \mathbf{R}'')$
$\mathcal{I} = \{(\pi_*^{\text{end}}, k_{\min})\}$ // Alignments mit gleichen minimalen Kosten

Lasso-Variante des Bitap-Algorithmus (vgl. Alg. 2.1) mit Kosten $\kappa_{\text{s}} = \kappa_{\text{i}} = \kappa_{\text{d}} = 1$ (Editierdistanz). Die maximalen Kosten $k(\leq k_{\max})$ sind hier eine variable Größe, welche durch ein Sequenzalignment mit geringeren Kosten $k_{\min}$ reduziert wird. Hierdurch verringert sich der nötige Aufwand zur Suche aller optimalen Sequenzalignments. Eine solche Reduktion wird auch auf die Ergebnismenge $\mathcal{I}$ übertragen, sodass letztlich gilt $\mathcal{I} = \{(\pi_*^{\text{end}}, k_{\min})\}$.

A.8 PCR-Korrektur: Multinomiales Modell

Für gegebene $m, n \in \mathbb{N}$ und $p_1, p_2, .., p_i, .., p_m \in [0, 1]$ mit $\sum_i p_i = 1$ seien $A_1, A_2, .., A_m$ multinomial verteilte Zufallsvariablen mit

$$\Pr(A_1 = a_1, A_2 = a_2, .., A_m = a_m) = \frac{n!}{a_1! a_2! \cdots a_m!} p_1^{a_1} p_2^{a_2} \cdots p_m^{a_m}$$

und $\sum_i A_i = n$. Sei weiterhin $I_i = I(A_i)$ eine Zufallsvariable basierend auf der Indikatorfunktion

$$I(A) = \begin{cases} 0, & \text{für } A = 0, \\ 1, & \text{für } A > 0, \end{cases}$$

so lässt sich der Erwartungswert $\mathrm{E}(\sum_i I_i)$ als

$$\begin{aligned} \mathrm{E}(\textstyle\sum_i I_i) &= \textstyle\sum_i 1 \cdot \Pr(A_i \neq 0) \\ &= \textstyle\sum_i (1 - \Pr(A_i = 0)) \\ &= \textstyle\sum_i (1 - (1 - p_i)^n) = \sum_i \mathrm{E}(I_i) \end{aligned}$$

und die Varianz $\mathrm{Var}(\sum_i I_i) = \sum_{i,j} \mathrm{Kov}(I_i, I_j) = \sum_i \mathrm{Var}(I_i) + \sum_{i \neq j} \mathrm{Kov}(I_i, I_j)$ berechnen mit

$$\begin{aligned} \mathrm{Var}(I_i) &= \mathrm{E}(I_i^2) - \mathrm{E}^2(I_i) \\ &= \mathrm{E}(I_i) - \mathrm{E}^2(I_i) \\ &= \mathrm{E}(I_i)(1 - \mathrm{E}(I_i)) \\ &= (1 - (1 - p_i)^n)(1 - p_i)^n \end{aligned}$$

für die Varianzen und

$$\begin{aligned} \mathrm{Kov}(I_i, I_j) &= \mathrm{E}(I_i - \mathrm{E}(I_i)) \cdot \mathrm{E}(I_j - \mathrm{E}(I_j)) \\ &= \mathrm{E}(I_i \cdot I_j) - \mathrm{E}(I_i) \cdot \mathrm{E}(I_j) \\ &= \{\Pr(A_i \neq 0 \cap A_j \neq 0)\} - \Pr(A_i \neq 0) \Pr(A_j \neq 0) \\ &= \{\Pr(A_i \neq 0) + \Pr(A_j \neq 0) - \Pr(A_i \neq 0 \cup A_j \neq 0)]\} - \Pr(A_i \neq 0) \Pr(A_j \neq 0) \\ &= 1 - (1 - p_i)^n + 1 - (1 - p_j)^n - (1 - (1 - (p_i + p_j))^n) - \ldots \\ &\quad \ldots - (1 - (1 - p_i)^n)(1 - (1 - p_j)^n) \\ &= (1 - (p_i + p_j))^n - (1 - p_i)^n (1 - p_j)^n \end{aligned}$$

für die Kovarianzen, wenn $i \neq j$ gilt.

A.9 Validierung und weitergehende Analysen

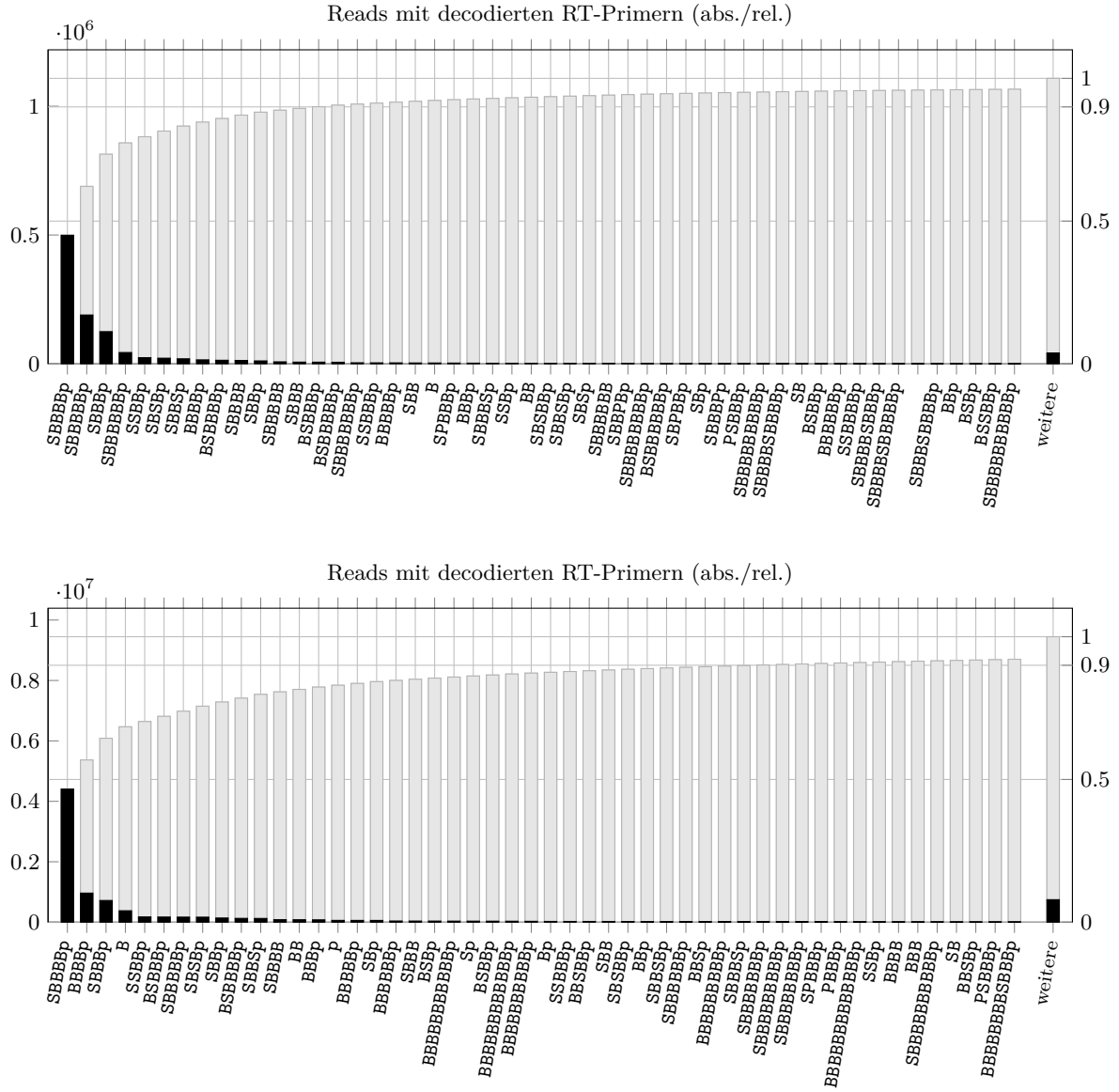

Abb. A.4 Histogramm aller Reads bezüglich der decodierten RT-primer (11Z oben, 25Z unten) in absoluter Anzahl und als relativer Anteil, ergänzt durch die kumulative Darstellung. Betrachtet werden die 50 häufigsten decodierten Sequenzen (in absteigender Reihenfolge).

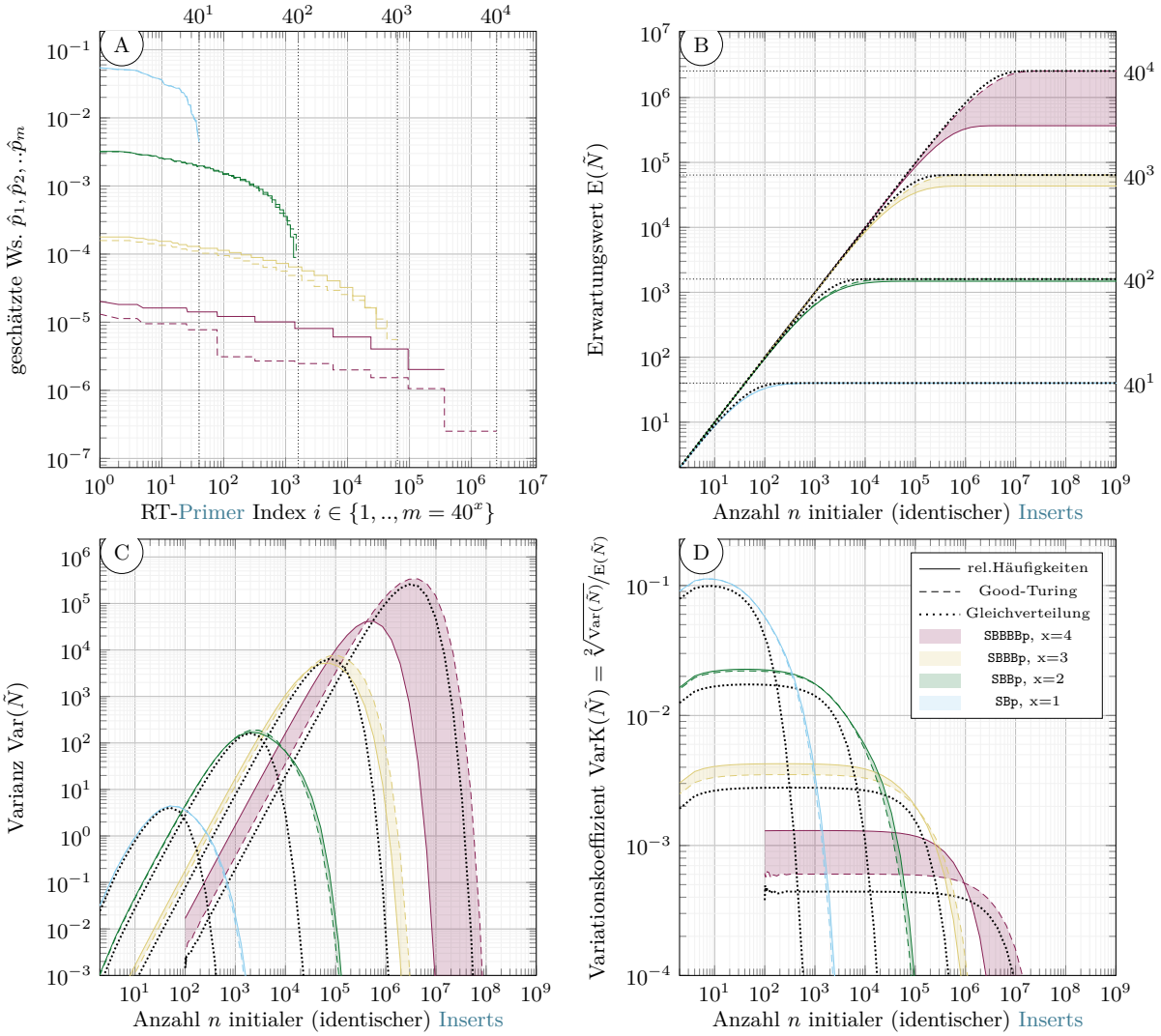

Abb. A.5 Grenzen der PCR-Korrektur (11Z): (A) Schätzung bedingter Auftrittswahrscheinlichkeiten $\hat{p}_i$ von RT-Primern abhängig von der Anzahl x enthaltener Barcode-Templates `B`. Primer sind Indexen i zugeordnet (Wahrscheinlichkeit absteigend). Für die Schätzung der relativen Häufigkeiten ist m reduziert, die Good-Turing-Schätzung geht hingegen von einer umfassenden Verteilung aus. (A) und n als hypothetische Anzahl an identischer initialer Inserts (x-Achse) ist mit dem theoretischen Ansatz aus 4.2.3 Grundlage für: Erwartungswert (B) der korrigierten Zählgröße $\tilde{N}$ (Zufallsvariable), deren Varianz (C) und Variationskoeffizient (D). Um die Darstellung von numerischen Ungenauigkeiten zu vermeiden wurden Kurven beschnitten.

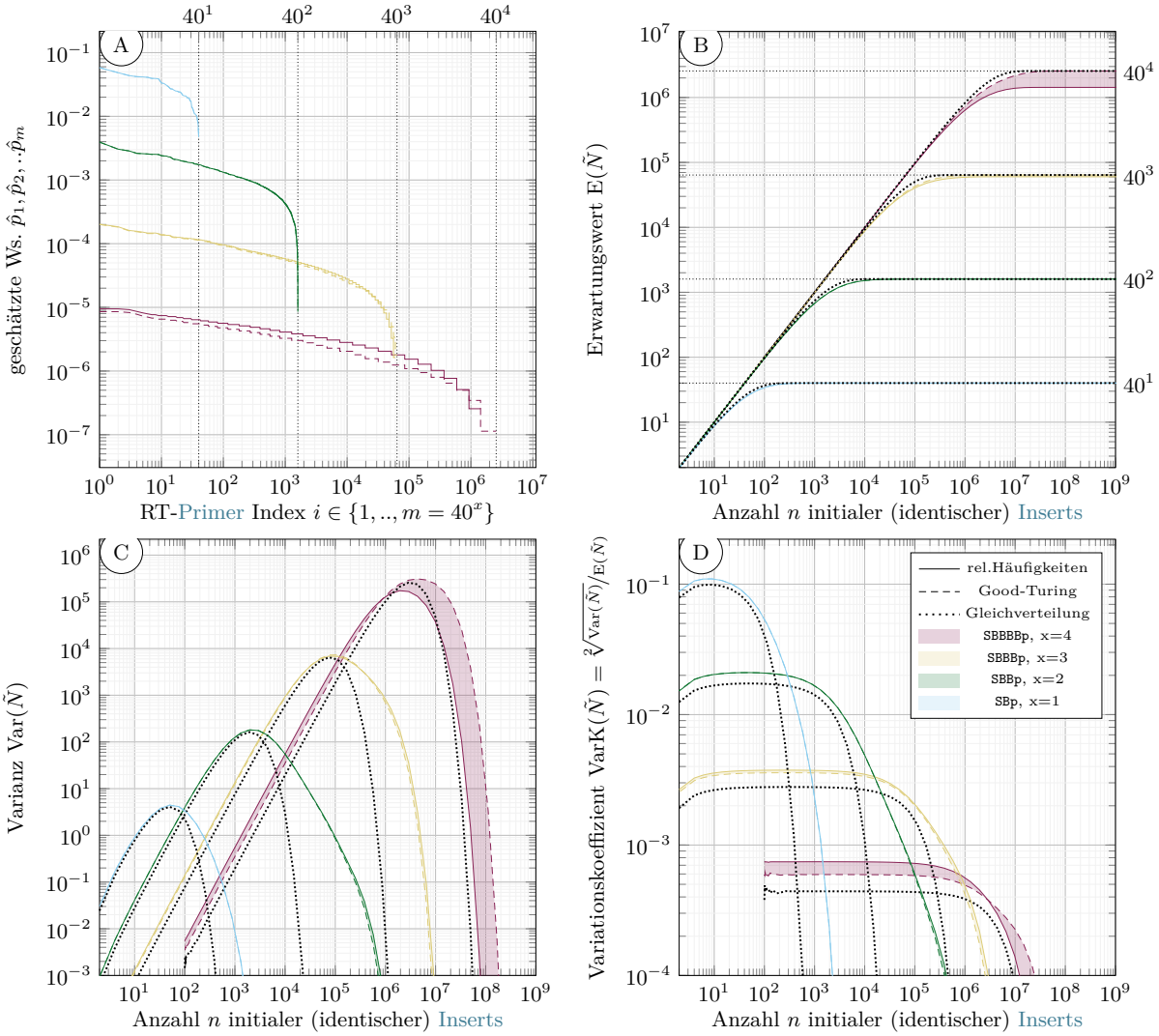

Abb. A.6 Grenzen der PCR-Korrektur (25Z): (A) Schätzung bedingter Auftrittswahrscheinlichkeiten $\hat{p}_i$ von RT-Primern abhängig von der Anzahl x enthaltener Barcode-Templates `B`. Primer sind Indexen i zugeordnet (Wahrscheinlichkeit absteigend). Für die Schätzung der relativen Häufigkeiten ist m reduziert, die Good-Turing-Schätzung geht hingegen von einer umfassenden Verteilung aus. (A) und n als hypothetische Anzahl an identischer initialer Inserts (x-Achse) ist mit dem theoretischen Ansatz aus 4.2.3 Grundlage für: Erwartungswert (B) der korrigierten Zählgröße $\tilde{N}$ (Zufallsvariable), deren Varianz (C) und Variationskoeffizient (D). Um die Darstellung von numerischen Ungenauigkeiten zu vermeiden wurden Kurven beschnitten.

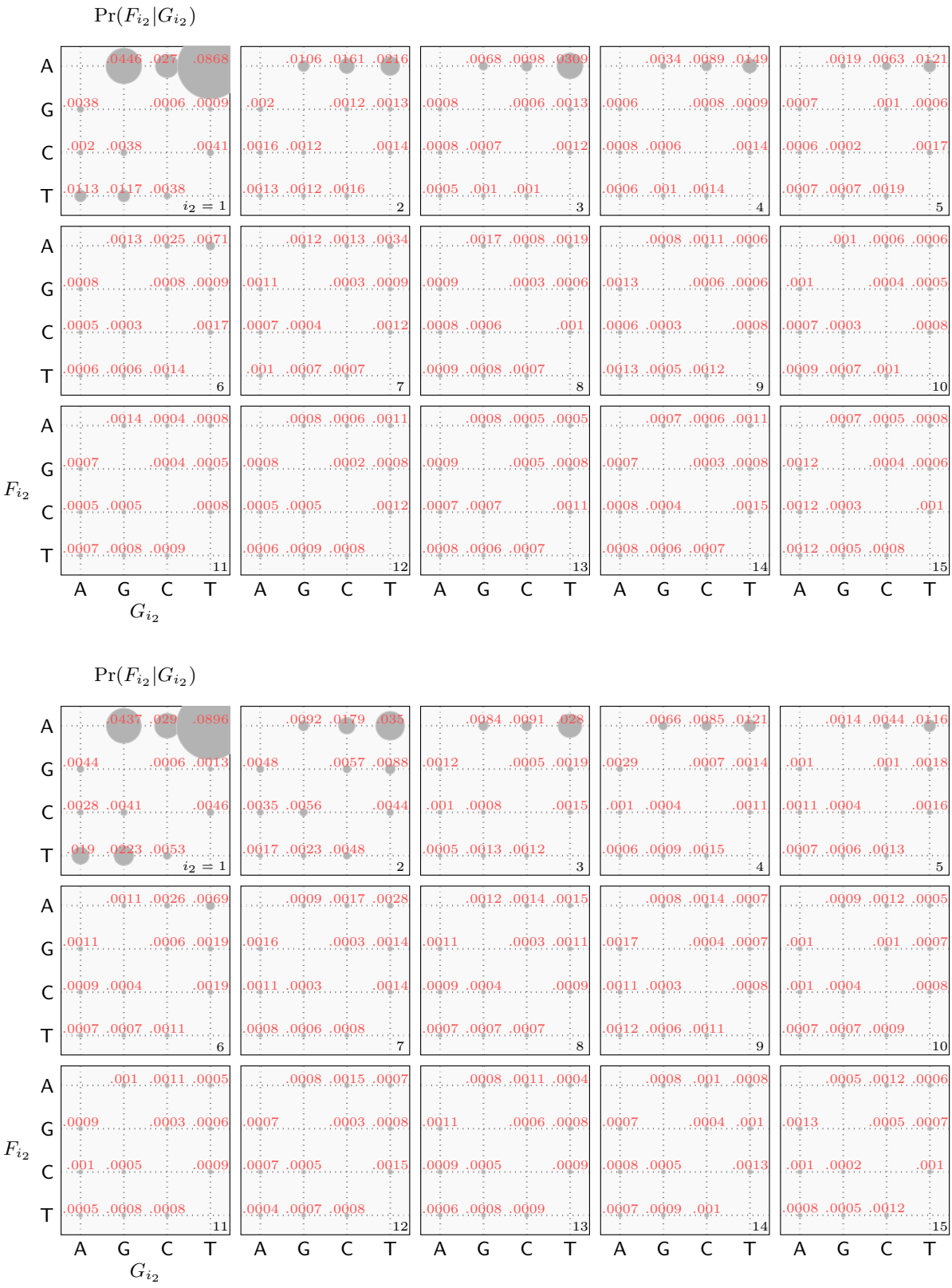

Abb. A.7 Positionsabhängige Transitionsmatrizen (am 3'-Ende) der Inserts (11Z oben, 25Z unten): Pr($F_{i_2}|G_{i_2}$) beschreibt geschätzte Substitutionswahrscheinlichkeiten. Für die letzten 5 Symbole der Inserts, mit Werten $i_2 \leq 5$, besteht eine starke Präferenz zur Ersetzung durch ein A bzw. T. Zur Mitte der Inserts, $i_2 > 10$, normalisiert sich diese Beobachtung. Das arithmetische Mittel der Transitionsmatrizen mit $i_2 > 10$ dient im Folgenden als Hintergrundmodell.

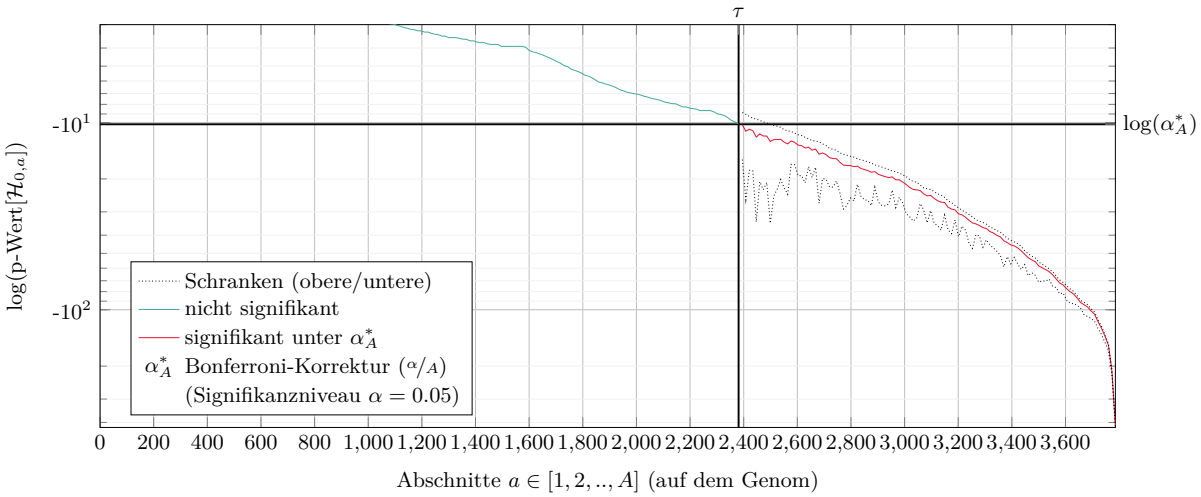

Abb. A.8 Hypothesentest für signifikante Akkumulation von Fehlern an 3'-Enden (11Z): Logarithmische Signifikanzwerte $-\log(\text{p-Wert}[H_{0,a}])$ (sortiert), für $A = 3793$ Abschnitte a (der Länge $\epsilon = 10$) im Genom von E. coli. Für 1412 Abschnitte $a > \tau$ besteht eine signifikante Häufung von Fehlern, basierend auf dem Signifikanzniveau $\alpha^*_A = \alpha/A$ (Bonferroni-Korrektur) mit $\alpha = 0.05$. Zur Korrektur von numerischen Ungenauigkeiten dienen die angegebenen Schranken. Die Sortierung der Werte erfolgt für $a > \tau$ gemäß der oberen Schranke (relevant für p-Werte $< 10^{-100}$).

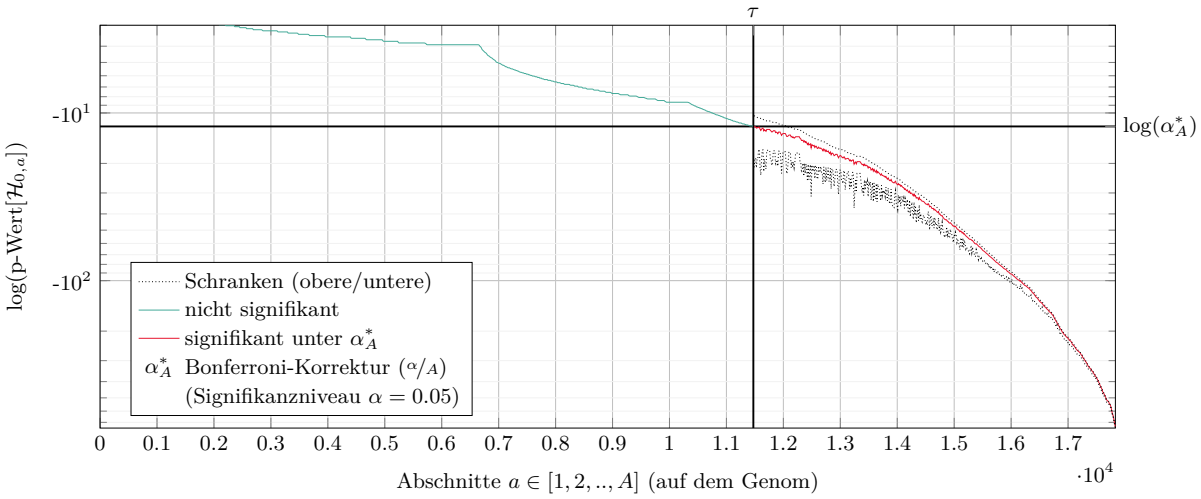

Abb. A.9 Hypothesentest für signifikante Akkumulation von Fehlern an 3'-Enden (25Z): Logarithmische Signifikanzwerte $-\log(\text{p-Wert}[H_{0,a}])$ (sortiert), für $A = 17831$ Abschnitte a (der Länge $\epsilon = 10$) im Genom von E. coli. Für 6358 Abschnitte $a > \tau$ besteht eine signifikante Häufung von Fehlern, basierend auf dem Signifikanzniveau $\alpha^*_A = \alpha/A$ (Bonferroni-Korrektur) mit $\alpha = 0.05$. Zur Korrektur von numerischen Ungenauigkeiten dienen die angegebenen Schranken. Die Sortierung der Werte erfolgt für $a > \tau$ gemäß der oberen Schranke (relevant für p-Werte $< 10^{-100}$).

Glossar

Adapter

Ein Adapter ist eine kurze synthetisch erzeugte DNA-Sequenz (Oligonukleotid, siehe Begriff 8), welche aufgrund von speziell beschaffenen Enden (engl. *sticky ends*) mit anderen DNA-Molekülen (Adapter und andere Oligonukleotide mit eingeschlossen) einen Molekülverbund eingehen kann. Adapter werden oft genutzt, um synthetisch erzeugte Moleküle an native DNA-Fragmente zu binden und damit eine funktionale Sequenzregion für weitere Prozesse zu bilden.

Seiten 32, 35, 77, 78, 143

Alignment-Lattice

Ein Alignment-Lattice im Kontext dieser Arbeit bezeichnet die Menge aller optimalen Sequenzalignments (siehe Semi-globales Alignment, Definition 26), die Sequenzierdaten auf ein Referenzgenom (siehe Begriff 15) abbilden. Der Name Lattice leitet sich dabei von der Zuordnung von Symbolpositionen der sequenzierten Reads auf Positionen im Referenzgenom ab, welche als Gitterstruktur dargestellt werden kann.

Seiten 93–95, 103, 109, 110

Aufschmelzen

Das Trennen der Wasserstoffbrückenbindungen des DNA-Doppelstrangs unter Einfluss von thermischer Energie wird auch Aufschmelzen genannt. Die Höhe der nötigen Temperatur zur Erzeugung von Einzelsträngen ist vom GC-Gehalt der DNA abhängig. Das Aufschmelzen der DNA ermöglicht unter anderem die biotechnologische Initiation der DNA-Replikation.

Seiten 33, 34, 79, 106, 138

Barcode

Ein Barcode im biotechnologischen Kontext der DNA-Sequnezierung ist eine synthetisch erzeugte DNA-Sequenz (Oligonukleotid, siehe Begriff 8), welche zur Markierung von nativen DNA- oder RNA-Fragmenten genutzt wird. Die Markierung kann einerseits spezifisch durch dedizierte Moleküle erfolgen, was als Multiplexing bezeichnet wird (siehe Begriff 19). Eine probabilistische Anwendung beschreiben die Zufallsbarcodes (siehe Begriff 20), welche verwendet werden, um quantitative Effekte für die Zählung von Molekülen auszugleichen.

Seiten iii, vii, 3–5, 7, 26, 37, 40–47, 49–52, 54–73, 81, 92, 114, 119, 120, 137, 143

Barcode-Templates

Ein Barcode-Template ist eine kurze synthetische DNA-Sequenz (Oligonukleotid, siehe Begriff 8), die Bestandteil eines Barcodes ist bzw. eine Teilsequenz eines größeren Verbunds aus Oligonukleotiden darstellt, die für die Erzeugung von Zufallsbarcodes genutzt wird.

Seiten 37, 76–85, 87, 88, 92, 96–101, 106, 114–116, 120, 124–129, 133, 134, 139

Diversität

Der Begriff der Diversität im Kontext der Zufallsbarcodes bedeutet eine hohe Anzahl an unterscheidbaren Sequenzen, deren Auftrittsverteilung im Experiment (als physisches Molekül) möglichst uniform ist. Die Gleichverteilung ist für die zufällige Markierung von Molekülen durch Zufallsbarcodes optimal.

Seiten 3, 5, 42, 75, 76, 81, 89–91, 96, 98, 100, 102, 106, 117, 143

Editierdistanz

Siehe Definition 25.

Seiten 21, 23, 44, 58, 60, 61, 63–67, 72, 91, 93, 95, 107, 128, 130

Elongation

Die Elongation ist ein Begriff der die komplementäre Ergänzungen auf Basis eines DNA-Einzelstrangs oder einer mRNA (bei der Proteinsynthese) umschreibt. Siehe dazu auch Begriffe wie 2, 4, 10 und 11.

Seiten 27–29, 34, 79, 86, 106

Eukaryot

Die Eukaryoten sind Organismen, welche ihre DNA in einem Zellkern bündeln. Es existieren ein- und mehrzellige Eukaryoten. Im Vergleich zu den Prokaryoten (ohne Zellkern), weisen Eukaryoten hinsichtlich der Differenzierung von Struktureinheiten der Zelle (der Differenzierung unterschiedlicher Zellen) und den Lebensvorgängen teilweise komplexere Zusammenhänge auf.

Seiten 26, 32, 108, 117, 138, 141

GC-Gehalt

Der GC-Gehalt ist ein Charakteristikum von Nukleotid-Sequenzen. Er bewertet den Prozentsatz der Nukleotide Guanin (G) und Cytosin (C) bezüglich aller enthaltenen Basen. Als lokaler oder globaler Wert ist der GC-Gehalt für weitere Eigenschaften der DNA bestimmend, so zeigen Abschnitte mit vielen G-C-Paarungen z. B. hinsichtlich der Wasserstoffbrückenbindungen eine höhere Stabilität, was Implikationen für das Aufschmelzen von DNA beinhaltet.

Seiten 62, 63, 81, 82, 124, 126, 128, 137

Gelelektrophorese

Die Gelelektrophorese ist ein biotechnologisches Verfahren zur Größenseparierung von DNA- oder RNA-Molekülen. Siehe dazu Begriff 12.

Seiten 35, 36, 80, 82, 84, 85, 97, 98, 129

Gen

Ein Gen ist ein Abschnitt der DNA, der als Informationseinheit die Struktur und die biologische Synthese von RNA codiert. Dabei können Gene als Vorlage für mRNA dienen und damit die Beschreibung für Proteine liefern oder die Informationen für andere funktionale RNA-Typen enthalten. Gene von Prokaryoten und Eukaryoten weisen grundsätzlich eine andere Struktur auf. Gene sind Informationseinheiten und Bestandteile des sogenannten Genoms.

Seiten 29, 31, 33, 39, 96

Genom

Die gesamte Sequenzinformation der DNA einer Zelle (RNA bei einem Virus) wird Genom genannt. Das Referenzgenom (vgl. Begriff 15) ist das aus DNA-Sequenzierungen und Computerunterstützung erstelltes Abbild des faktischen Genoms eines Organismus und dient als Arbeitshypothese und Referenz für die Mikrobiologie.

Seiten 35, 38, 39, 45, 71, 87, 91–96, 99, 102–105, 107–109, 111–116, 136–138, 142

Hamming-Abstand

Siehe Definition 17.

Seiten 2, 18–21, 23, 43, 44, 52, 54, 55, 58, 60, 61, 64, 65, 69, 80, 82, 87, 88, 92, 99, 124–126, 128, 130

Homologe Sequenz

Im Kontext dieser Arbeit werden zwei Sequenzen als homolog zueinander bezeichnet, wenn sie einen Abschnitt zuvor definierter Länge enthalten, welcher entweder identisch in der anderen Sequenz zu finden ist oder selbiges für dessen reverses Komplement gilt. Homologe Sequenzen und deren komplementäre Entsprechungen können zu einer (teilweisen) Hybridisierung in DNA-Doppelstränge führen. Die Funktion der Primer basiert auf dem Prinzip der homologen Sequenzen.

Seiten 79, 81, 82, 124

Homopolymer

Ein Homopolymer ist ein Polymer das im Kontext der DNA und RNA eine Molekülkette darstellt, welche aus der Aneinanderreihung ein und derselben Base besteht. Die Lauflänge dieser Folge wird auch als Homopolymerlänge bezeichnet. Längere Homopolymere zeigen bezüglich des Informationstransfers durch Polymerase und reverse Transkriptase negative Effekte hinsichtlich der lokalen Fehlerrate.

Seiten 62, 81, 82, 124, 128

Hybridisierung

Das schießen von Wasserstoffbrückenbindungen komplementärer Basenpaarungen wird auch als Hybridisierung bezeichnet (siehe beispielsweise Begriff 2). Für teilweise komplementäre Bereiche ist auch eine Teilhybridisierung zu doppelsträngigen Strukturen möglich.

Seiten 5, 27–29, 32–34, 37, 77–81, 86, 88, 90, 91, 96, 103, 106, 107, 116, 124, 138, 141, 143

Indel

Im Bereich der DNA-Sequenzierung bezieht sich der Begriff Indel auf Sequenzfehler, die neben Einfügungen und Löschungen auch Ersetzungen umfassen. Siehe hierzu Begriff 16.

Seiten 44, 92

Insert

Im Kontext der Sequenzierung bezeichnet das Insert den Bereich (Sequenzabschnitt) der nativen DNA oder RNA im Template. Neben anderen biotechnisch relevanten Abschnitten im Template ist die Bestimmung der Sequenzinformation des Inserts die eigentliche Motivation einer Sequenzierung. Siehe auch Begriff 14.

Seiten 37–39, 45, 58, 77–79, 82, 86–97, 99, 101–103, 105–113, 116, 133–135, 142

Kettenpolymerisation

Die Kettenpolymerisation ist ein allgemeiner Begriff einer chemischen Reaktion an welcher gleiche (oder unterschiedliche) elementare Polymere (Monomere) beteiligt sind, aus welcher eine wachsende Kette von längeren Polymeren resultiert. Im Kontext der Zufallsbarcodes bezieht sich die Kettenpolymerisation auf die Bildung von Ketten der Barcode-Templates.

Seiten 76, 84, 116, 129

Komplement

Auf Basis der Struktur und Valenzen der Bindungskräfte bilden die Nukleotide Adenin (A) und Thymin (T) bzw. Guanin (G) und Cytosin (C) komplementäre Basenpaarungen. Sequenzabschnitte die dieser komplementären Paarung entsprechen werden auch als Komplement bezeichnet. Siehe auch Aufbau der DNA in Begriff 1. Oft wird der Term Komplement stellvertretend für das reverse Komplement verwendet, welches die gegensinnige Orientierung und Leserichtung von Sequenzen mit einschließt.

Seiten 1, 26–29, 31, 33–35, 62, 63, 79–81, 83, 86, 106, 128, 141

Ligase

Die Ligase ist ein Enzym welches die Verbindung von DNA-Molekülen katalysiert und die Herstellung einer geschlossenen DNA-Holmstruktur ermöglicht. Siehe Begriff 9.

Seiten 28, 33, 77, 79, 80, 139

Ligation

Die durch das Enzym Ligase katalysierte Reaktion wird Ligation genannt. Siehe Begriff 9.

Seiten 32, 33, 35, 37, 77–85, 88, 99, 115, 116, 120, 124, 129, 140

nt

Die Einheit nt bezeichnet die Anzahl an Nukleotiden die eine DNA- oder RNA-Sequenz umfasst.
Seiten 26, 28, 29, 32, 33, 37, 58, 62, 69, 70, 83–85, 88, 94, 96–98, 103, 107, 108

Nukleotid

Nukleotide sind die Bausteine für die Nukleinsäuren DNA und RNA. Siehe hierzu Aufbau der DNA in Begriff 1 bzw. der RNA in Begriff 3.
Seiten 1, 5, 26–38, 44, 45, 47, 49, 54, 55, 57, 62, 63, 66, 67, 71, 85, 86, 93, 94, 105, 107, 109, 110, 112, 113, 115, 124, 138–143

Oligonukleotid

Ein Oligonukleotid, auch Oligomer genannt, ist ein kurzes synthetisch erzeugtes Polymer aus wenigen Nukleotiden, welches eine spezifische biotechnologischen Funktion erfüllt. Siehe Begriff 8.
Seiten vii, 32, 34, 35, 37, 38, 40–44, 46, 59, 71–73, 75–77, 79, 81–85, 87, 88, 98–100, 114, 116, 120, 124, 125, 129, 137, 141–143

PCR

Die PCR ist eine biotechnologische (kontrollierte) Umsetzung der, auch natürlich vorkommenden DNA-Replikation und dient unter anderem zum Zweck der Vervielfältigung von DNA-Molekülen. Siehe Begriff 11.
Seiten vii, 33–35, 37–42, 75, 77, 79, 89–92, 96, 97, 99–108, 113, 116, 117, 120, 121, 131, 133, 134, 140, 143

PCR-Duplikat

Eine in der PCR erzeugte Kopie eines DNA-Moleküls wird auch als PCR-Duplikat bezeichnet. Für quantitative Analysen ist die ungleichmäßige Erzeugung von PCR-Duplikaten (PCR-Bias) eine unvermeidbare Fehlerquelle. Siehe auch PCR-Bias in Begriff 18.
Seiten 40, 42, 85, 88–93, 96, 99, 102–106, 108, 115, 116

Phosphorylierung

Die Phosphorylierung beschreibt die umkehrbare Erweiterung eines organischen Moleküls mit einer Phosphatgruppe (siehe hierzu Aufbau der DNA in Begriff 1). Im Kontext der biotechnologischen Werkzeuge beinhaltet die Phosphorylierung von Sequenzen aus Nukleotiden die Bereitstellung von Stoffwechselenergie, um katalytische und enzymatische Prozesse zu ermöglichen. Die Ligation ist ein beispielhafter Prozess, der eine Phosphorylierung voraussetzt.
Seiten 80, 97, 123, 124, 129

Polyadenylierung

Die Polyadenylierung bezeichnet die Erweiterung von mRNA-Molekülen mit Ketten des Nukleotids Adenin (A). Die Ergänzung von Poly-(A)-3'-Enden stellt eine Modifikation von RNA dar (siehe Begriff 7), deren funktionalen Aspekte im Stoffwechsel von Zellen noch nicht vollständig geklärt sind. Die Polyadenylierung wird mit der Stabilität und dem Abbau von mRNA bei Prokaryoten in Verbindung gebracht.
Seiten 32, 87, 108, 110, 111, 114, 115, 141

Polymer

Ein Polymer bezeichnet eine chemische Struktur die aus Makromolekülen einer bestimmten Klasse von Molekülen gebildet werden. Der Begriff Polymer bezieht sich im Kontext dieser Arbeit ausschließlich auf die Molekülketten der DNA und RNA.
Seiten 1, 25, 32, 35, 75, 84, 85, 119, 124, 129, 139, 140

Polymerase

Die Polymerase ist ein Enzym, das die Polymerisation von Nukleotiden katalysiert. Es existieren viel unterschiedliche Klassen von Polymerase-Molekülen: Die DNA-Polymerase ist maßgeblich an

der matrizengesteuerten Polymerisation der DNA-Replikation beteiligt. Siehe Begriff 2. RNA-Polymerasen sind zum Beispiel in der Polyadenylierung von mRNA-Molekülen involviert.
Seiten 28, 29, 32–34, 86, 106, 139

Primer

Der Begriff Primer steht stellvertretend für ein Oligonukleotid welches, durch die Hybridisierung an einer einzelsträngigen Nukleotid-Sequenz, als Initiator für die matrizengesteuerte Polymerisation der DNA dient. Primer existieren als DNA- oder RNA-Moleküle und bedingen unter anderem die enzymatische Reaktion der DNA-Replikation (siehe Begriff 2) oder der reversen Transkription (siehe Begriff 10).
Seiten 28, 32, 34–37, 76–92, 96–104, 106, 116, 124, 128, 133, 134, 138, 142, 143

Prokaryot

Den Prokaryoten, wie beispielsweise den Bakterien, fehlt sowohl ein Zellkern als auch eine erweiterte strukturierte Bündelung der DNA. Auf Grund des anderen Aufbaus der prokaryotischen Zelle zeigen die zellulären Vorgänge in ihr einen teilweise einfacheren Aufbau. Die Zellteilung und Proteinsynthese ist bei Prokaryoten beispielsweise deutlich einfacher strukturiert als bei den Eukaryoten.
Seiten 27, 32, 108, 111, 117, 138, 140

Protokoll

Ein Protokoll im Bereich der Mikrobiologie ist eine exakte Beschreibung zur Durchführung eines bestimmten Experiments oder biotechnischen Vorgangs.
Seiten vii, 35, 37, 76–79, 81, 84, 85, 96–98, 114, 116, 117, 120, 123, 124, 142

Read

Ein Read im Kontext der DNA-Sequenzierung der zweiten Generation (siehe Begriff 14) entspricht dem textuellen Abbild eines Bereichs der DNA-Sequenz des Template. Experimentell bedingt entspricht der Read immer einer Teilsequenz des gesamten Templates und ist durch eine fest vorgegebene Sequenzierlänge beschränkt.
Seiten 37–39, 41, 42, 46, 47, 50–53, 56–58, 67, 69, 72, 86–89, 91, 95, 98–100, 102–105, 107–111, 119, 132, 137, 142

Repeat

Ein Repeat liegt in einer Folge von Nukleotiden vor, wenn ein bestimmter Abschnitt der Sequenz identisch an anderer Stelle wiederholt auftritt. Ein sogenannter invertierter Repeat liegt vor, wenn das reverse Komplement eines bestimmten Abschnitts identisch an anderer Stelle auftritt. Ein invertierter Repeat kann zur teilweisen Hybridisierung von einzelsträngigen Molekülen mit sich selbst führen und eine erweiterte Sekundärstruktur der Moleküle verursachen.
Seiten 125, 126

reverse Komplement

Als Erweiterung des Komplements beschreibt das reverse Komplement eine zusätzliche Umkehrung der Leserichtung der Sequenz. Bezüglich des Aufbaus der DNA (vgl. Begriff 1) ist die (5'-3')-Kette des Vorwärtsstrangs und die 3'-5'-Folge als Rückwärtsstrang revers komplementär zueinander. Das reverse Komplement basiert auf der Struktureigenschaft der DNA.
Seiten 26, 77, 80, 81, 115, 125, 126, 138, 139, 141

reverse Transkriptase

Die reverse Transkriptase ist ein Enzym, das die Synthese von sogenannter cDNA (engl. *complementary DNA*) auf Basis von RNA katalysiert. Siehe Begriff 10.
Seiten 33, 77, 139

reverse Transkription

Reverse Transkription (RT), siehe Begriff 10.
Seiten ix, 33, 38, 75–91, 96–102, 104, 106, 116, 117, 124, 132–134, 141, 142

Sekundärstruktur

Einzelsträngige Nukleotid-Sequenzen (DNA als auch RNA) können auf Basis von Bindungskräften erweiterte Molekülformationen, sogenannte Sekundärstrukturen, ausbilden. Sekundärstrukturen erfüllen am Beispiel der RNA-Moleküle funktionale Aspekte für den Zellstoffwechsel. Zusätzlich können unerwünschte Sekundärstrukturen zu nachteiligen Effekten bei der Effizienz biotechnologischer Prozesse führen.
Seiten 29, 124, 125, 141

Sequenzalignment

Der Bereich des Sequenzalignments ist ein Teilgebiet der musterbasierten Suche. In dieser Arbeit bezieht sich das Sequenzalignment auf das paarweise Alignment, im Speziellen das semi-globale Alignment (siehe Definition 26), mit dem Ziel der Abbildung von Sequenzierdaten (Reads) auf ein Referenzgenom. Siehe hierzu auch Begriff 17.
Seiten 22–24, 38, 39, 51, 71, 85, 87, 88, 92–96, 102, 103, 105, 107, 109, 110, 112–116, 130, 137

Sequenzfehler

Siehe Begriff 16.
Seiten vii, 3, 38, 39, 41–46, 54, 66, 70, 71, 89, 91, 93, 99, 107–109, 112, 115, 116, 139

Sequenzier-Library

Der Begriff Sequenzier-Library umfasst eine Menge von DNA-Fragmenten, die in Form von Inserts (im Verbund des Templates) direkt zur Sequenzierung eingesetzt werden können. Die Herstellung der Templates umfasst ein komplexes biotechnologisches Protokoll, welches auch als Library-Herstellung bezeichnet wird.
Seiten 37–41, 61, 89–91, 96, 97, 103, 125, 142

Sequenzierlänge

Die Sequenzierung von DNA ist auf Grund von technischen Limitierungen auf eine maximale Länge von Nukleotiden begrenzt. Für unterschiedliche Technologien gibt die Sequenzierlänge die jeweils maximale Anzahl an möglichen Symbolen an die gelesen werden können.
Seiten 37, 58, 73, 82, 86, 97, 107, 141

Spacer

Das sogenannte Spacer-Molekül beschreibt in dieser Arbeit ein Füllstück an DNA, das genutzt wird, um die neuartigen Zufallsbarcodes an eine native RNA-Sequenz zu koppeln. Die Markierung der RNA mit einem zufälligen Barcode erfolgt über die reverse Transkription, wobei der Sequenz des Spacers an der RNA zum einen die Funktion einer Bindungsstelle zukommt, wohingegen der Zufallsbarcode die Funktion eines RT-Primers erfüllt. Siehe hierzu Abschnitt 4.1.2.
Seiten 78–83, 88, 97, 99, 128

technische Sequenz

Als technische Sequenzen werden in dieser Arbeit Oligonukleotide bezeichnet, die einen fest vorgegebenen Zweck im Protokoll der Library-Herstellung erfüllen sollen. Zur problemlosen Integration von neuen synthetischen Nukleotid-Sequenzen in bestehende Protokolle, ist eine molekulare Interaktion mit bestehenden technischen Sequenzen zu vermeiden.
Seiten 32, 77, 78, 82, 123, 126, 139

Template

Ein Template ist ein DNA-Molekül, das als Vorlage sowohl für die PCR als auch für die Sequenzierung dient. Zur Erzeugung der funktionalen Einheit des Templates ist der Einsatz von Oligonukleotiden nötig, um beispielsweise native Nukleotid-Sequenzen für biotechnologische Methoden zugänglich zu machen. Meist enthält ein Template neben der nativen DNA (RNA) noch Adapter, die als Hybridisierungsregion für Primer dienen.

Seiten vii, 37–39, 41, 43–47, 50, 51, 57, 58, 67, 71, 119, 139, 141–143

Transkription

Siehe Begriff 4.

Seiten 29, 39, 75, 105, 108

Zufallsbarcode

Das herkömmliche Konzept der Barcodes setzt eine bekannte Menge an Codeworten voraus, die unverändert zur Markierung von Templates genutzt wird. Ein Zufallsbarcode nutzt die zufällige physikalische Kombination einer bekannten (kleinen) Menge von synthetischen Sequenzen, um eine große Anzahl variabler Sequenzen zur Markierung von Molekülen zu erhalten. Die Diversität der Zufallsbarcodes spielt bei deren Verwendung zur Zählung von Molekülen eine wichtige Rolle. Siehe auch Begriff 20.

Seiten vii, 3–5, 40–42, 75–77, 79, 82, 85–87, 89, 91, 92, 96, 99, 100, 102, 103, 106–108, 114–117, 120, 123, 137, 139, 142

Literaturverzeichnis

[1] H. Abdi, The Bonferonni and Šidák Corrections for Multiple Comparisons, N. J. Salkind (Hg.), Encyclopedia of Measurement and Statistics, S. 103–107, Thousand Oaks (CA): SAGE Publications, Inc., 3. Aufl., 2007.

[2] S. G. Acinas, R. Sarma-Rupavtarm, V. Klepac-Ceraj et al., PCR-Induced Sequence Artifacts and Bias: Insights from Comparison of Two 16S rRNA Clone Libraries Constructed from the Same Sample, *Applied and Environmental Microbiology*, Bd. 71(12), S. 8966–8969, 2005.

[3] A. Adey, H. G. Morrison, X. Xun et al., Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition, *Genome Biology*, Bd. 11(12), 2010.

[4] D. Aird, M. G. Ross, W.-S. Chen et al., Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries, *Genome Biology*, Bd. 12(2), 2011.

[5] B. Alberts, D. Bray, J. Lewis et al., Molekularbiologie der Zelle, John Wiley & Sons Inc., Hoboken/New Jersey, 5. Aufl., 2011.

[6] S. F. Altschul und B. W. Erickson, Optimal sequence alignment using affine gap costs, *Bulletin of Mathematical Biology*, Bd. 48(5), S. 603–616, 1986.

[7] R. Arratia und L. Gordon, Tutorial on large deviations for the binomial distribution, *Bulletin of Mathematical Biology*, Bd. 51(1), S. 125–131, 1989.

[8] D. Ashlock, L. Guo und F. Qiu, Greedy Closure Evolutionary Algorithms, Proceedings of the World on Congress on Computational Intelligence, Bd. 2, S. 1296–1301, 2002.

[9] D. Ashlock und S. K. Houghten, A Novel Variation Operator for More Rapid Evolution of DNA Error Correcting Codes, Proceedings of the IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), S. 1–8, 2005.

[10] P. M. Ashton, S. Nair, T. Dallman et al., MinION nanopore sequencing identifies the position and structure of a bacterial antibiotic resistance island, *Nature Biotechnology*, Bd. 33(3), S. 296–300, 2015.

[11] R. Baeza-Yates und G. H. Gonnet, A New Approach to Text Searching, *Communications of the ACM*, Bd. 35(10), S. 74–82, 1992.

[12] L. Bahl, J. Cocke, F. Jelinek et al., Optimal Decoding of Linear Codes for Minimizing Symbol Error Rate, *IEEE Transactions on Information Theory*, Bd. 20(2), S. 284–287, 1974.

[13] F. Balado, On the Shannon capacity of DNA data embedding, Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), S. 1766–1769, 2010.

[14] R. E. Bellman, The Theory of Dynamic Programming, *Bulletin of the American Mathematical Society*, Bd. 60(6), S. 503–515, 1954.

[15] A. Ben-Dor, R. Karp, B. Schwikowski et al., Universal DNA Tag Systems: A Combinatorial Design Scheme, *Journal of Computational Biology*, Bd. 7(3-4), S. 503–519, 2004.

[16] J. M. Berg, J. L. Tymoczko und L. Stryer, Biochemistry, W. H. Freeman and Company, New York, 5. Aufl., 2002.

[17] E. Berlekamp, Nonbinary BCH decoding, *IEEE Transactions on Information Theory*, Bd. 14(2), S. 242–242, 1968.

[18] C. Berrou und A. Glavieux, Near Optimum Error Correcting Coding And Decoding: Turbo-Codes, *IEEE Transactions on Communications*, Bd. 44(10), S. 1261–1271, 1996.

[19] J. Binladen, M. T. Gilbert, J. P. Bollback et al., The Use of Coded PCR Primers Enables High-Throughput Sequencing of Multiple Homolog Amplification Products by 454 Parallel Sequencing, *PLoS ONE*, Bd. 2(2), 2007.

[20] N. A. Bokulich, S. Subramanian, J. J. Faith et al., Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing, *Nature Methods*, Bd. 10(1), S. 57–59, 2013.

[21] R. C. Bose und D. K. Ray-Chaudhuri, On A Class of Error Correcting Binary Group Codes*, *Information and Control*, Bd. 3(1), S. 68–79, 1960.

[22] M. Bossert, Kanalcodierung (Informationstechnik), Teubner Verlag, Stuttgart, 2. Aufl., 1998.

[23] L. M. Bragg, G. Stone, M. K. Butler et al., Shining a Light on Dark Sequencing: Characterising Errors in Ion Torrent PGM Data, *PLoS Computational Biology*, Bd. 9(4), 2013.

[24] S. Brenner, M. Johnson, J. Bridgham et al., Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays, *Nature Biotechnology*, Bd. 18(6), S. 630–634, 2000.

[25] S. Brenner, S. R. Williams, E. H. Vermaas et al., In vitro cloning of complex mixtures of DNA on microbeads: Physical separation of differentially expressed cDNAs, *Proceedings of the National Academy of Sciences, PNAS*, Bd. 97(4), S. 1665–1670, 2000.

[26] J. A. Briffa, A GPU Implementation of a MAP Decoder for Synchronization Error Correcting Codes, *IEEE Communications Letters*, Bd. 17, S. 996–999, 2013.

[27] J. A. Briffa und H. G. Schaathun, Improvement of the Davey-MacKay construction, Proceedings of the International Symposium on Information Theory and Its Applications (ISITA), S. 1–4, 2008.

[28] I. N. Bronstejn, K. A. Semendjaev, G. Musiol et al., Taschenbuch der Mathematik, Verlag Harri Deutsch, Frankfurt am Main, 6. Aufl., 2005.

[29] T. Buschmann und L. V. Bystrykh, Levenshtein error-correcting barcodes for multiplexed DNA sequencing, *BMC Bioinformatics*, Bd. 14(1), S. 272, 2013.

[30] T. Buschmann, R. Zhang, D. E. Brash et al., Enhancing the detection of barcoded reads in high throughput DNA sequencing data by controlling the false discovery rate, *BMC Bioinformatics*, Bd. 15(1), S. 264, 2014.

[31] V. Buttigieg und J. A. Briffa, Improved Code Construction for Synchronization Error Correction, Proceedings of the International ITG Conference on Systems, Communications and Coding (SCC), S. 1–6, 2015.

[32] L. V. Bystrykh, Generalized DNA Barcode Design Based on Hamming Codes, *PLoS ONE*, Bd. 7(5), 2012.

[33] M. Carneiro, C. Russ, M. Ross et al., Pacific biosciences sequencing technology for genotyping and variation discovery in human data, *BMC Genomics*, Bd. 13(1), S. 375, 2012.

[34] J. A. Casbon, R. J. Osborne, S. Brenner et al., A method for counting PCR template molecules with application to next-generation sequencing, *Nucleic Acids Research*, 2011.

[35] J. M. Coffin, S. H. Hughes und H. E. Varmus, Retroviruses, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, 1997.

[36] S. N. Cohen, Surprises at the 3' End of Prokaryotic RNA, *Cell*, Bd. 80(6), S. 829–832, 1995.

[37] T. M. Cover und J. A. Thomas, Elements of Information Theory, John Wiley & Sons Inc., Hoboken/New Jersey, 2. Aufl., 2006.

[38] F. Crick, Central Dogma of Molecular Biology, *Nature*, Bd. 227(5258), S. 561–563, 1970.

[39] F. H. Crick, On Protein Synthesis, Symposia of the Society for Experimental Biology, Bd. 12, S. 138–163, 1958.

[40] M. C. Davey, Error-correction using Low-Density Parity-Check Codes, Dissertation, Universität Cambridge, 1999.

[41] M. C. Davey und D. J. MacKay, Reliable Communiation over Channels with Insertions, Deletions and Substitutions, *IEEE Transactions on Information Theory*, Bd. 47(2), S. 687–698, 2001.

[42] B. Dömölki, An algorithm for syntactical analysis, *Computational Linguistics*, Bd. 3(29-46), S. 151, 1964.

[43] B. Dömölki, A universal compiler system based on production rules, *BIT Numerical Mathematics*, Bd. 8(4), S. 262–275, 1968.

[44] J. C. Dohm, C. Lottaz, T. Borodina et al., Substantial biases in ultra-short read data sets from high-throughput DNA sequencing, *Nucleic Acids Research*, Bd. 36(16), 2008.

[45] P. Elias, Error-free Coding, Techn. Ber., Massachusetts Institute of Technology, Research Laboratory of Electronics, 1954.

[46] P. Elias, Coding for noisy channels, *IRE International Convention Record*, Bd. 4, S. 37–46, 1955.

[47] A. M. Eren, J. H. Vineis, H. G. Morrison et al., A Filtering Method to Generate High Quality Short Reads Using Illumina Paired-End Technology, *PLoS ONE*, Bd. 8(6), 2013.

[48] B. C. Faircloth und T. C. Glenn, Not All Sequence Tags Are Created Equal: Designing and Validating Sequence Identification Tags Robust to Indels, *PLoS ONE*, Bd. 7(8), 2012.

[49] W. Feller, An Introduction to Probability Theory and Its Applications, Bd. 1, John Wiley & Sons Inc., Hoboken/New Jersey, 3. Aufl., 1968.

[50] N. A. Fonseca, J. Rung, A. Brazma et al., Tools for mapping high-throughput sequencing data, *Bioinformatics*, 2012.

[51] G. D. Forney, Generalized Minimum Distance Decoding, *IEEE Transactions on Information Theory*, Bd. 12(2), S. 125–131, 1966.

[52] G. D. Forney Jr, Concatenated Codes, Dissertation, Massachusetts Institute of Technology, 1965.

[53] K. Fox-Walsh, J. Davis-Turak, Y. Zhou et al., A multiplex RNA-seq strategy to profile poly(A^+) RNA: Application to analysis of transcription response and 3' end formation, *Genomics*, Bd. 98(4), S. 266–271, 2011.

[54] D. N. Frank, BARCRAWL and BARTAB: software tools for the design and implementation of barcoded primers for highly multiplexed DNA sequencing, *BMC Bioinformatics*, Bd. 10(1), S. 362, 2009.

[55] R. E. Franklin und R. G. Gosling, Evidence for 2-chain Helix in Crystalline Structure of Sodium Deoxyribonucleate, *Nature*, Bd. 172, S. 156–157, 1953.

[56] G. K. Fu, J. Hu, P.-H. Wang et al., Counting individual DNA molecules by the stochastic attachment of diverse labels, *Proceedings of the National Academy of Sciences, PNAS*, Bd. 108(22), S. 9026–9031, 2011.

[57] W. A. Gale und G. Sampson, Good-turing frequency estimation without tears*, *Journal of Quantitative Linguistics*, Bd. 2(3), S. 217–237, 1995.

[58] R. G. Gallager, Low-Density Parity-Check Codes, *IRE Transactions on Information Theory*, Bd. 8(1), S. 21–28, 1962.

[59] R. G. Gallager, Low-Density Parity-Check Codes, Monograph, M.I.T. Press, 1963.

[60] R. G. Gallager, Information Theory and Reliable Communication, John Wiley & Sons Inc., Hoboken/New Jersey, 1968.

[61] A. Gilles, E. Meglécz, N. Pech et al., Accuracy and quality assessment of 454 GS-FLX Titanium pyrosequencing, *BMC Genomics*, Bd. 12(1), S. 245, 2011.

[62] M. J. E. Golay, Notes on Digital Coding, *Proceedings of the Institute of Radio Engineers*, Bd. 37(6), S. 657–657, 1949.

[63] I. J. Good, The Population Frequencies of Species and the Estimation of Population Parameters, *Biometrika*, Bd. 40(3-4), S. 237–264, 1953.

[64] O. Gotoh, An Improved Algorithm for Matching Biological Sequences, *Journal of Molecular Biology*, Bd. 162(3), S. 705–708, 1982.

[65] M. Grassl, Searching for linear codes with large minimum distance, Algorithms and Computation in Mathematics, Bd. 19, S. 287–313, Springer, Berlin/Heidelberg, 2006.

[66] D. Gussfield, Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology, Cambridge University Press, 1. Aufl., 1997.

[67] M. Hackhofer, Transcriptome sequencing of small cell populations in Escherichia coli O157:H7 str. EDL933 (EHEC), Bachelorarbeit, ZIEL, Zentralinstitut für Ernährungs- und Lebensmittelforschung, TU München, 2015.

[68] M. Hamady, J. J. Walker, J. K. Harris et al., Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex, *Nature Methods*, Bd. 5(3), S. 235–237, 2008.

[69] R. W. Hamming, Error Detecting and Error Correcting Codes, *The Bell System Technical Journal*, Bd. 29(2), S. 147–160, 1950.

[70] T. E. Harris, The Theory of Branching Processes, Dover Publications, Mineola, New York, 2002.

[71] T. Hashimshony, F. Wagner, N. Sher et al., CEL-Seq: Single-Cell RNA-Seq by Multiplexed Linear Amplification, *Cell Reports*, Bd. 2(3), S. 666–673, 2012.

[72] D. Haughton und F. Balado, BioCode: Two biologically compatible Algorithms for embedding data in non-coding and coding regions of DNA, *BMC Bioinformatics*, Bd. 14(1), S. 121, 2013.

[73] D. Haughton und F. Balado, A Modified Watermark Synchronisation Code for Robust Embedding of Data in DNA, Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), S. 1148–1152, 2013.

[74] D. S. Hirschberg, A Linear Space Algorithm for Computing Maximal Common Subsequence, *Communications of the ACM*, Bd. 18(6), S. 341–343, 1975.

[75] A. Hocquenghem, Codes correcteurs d'erreurs, *Chiffres*, Bd. 2, S. 147–156, 1959.

[76] C.-C. Hon, C. Weber, O. Sismeiro et al., Quantification of stochastic noise of splicing and polyadenylation in Entamoeba histolytica, *Nucleic Acids Research*, 2012.

[77] H. Hug und R. Schuler, Measurement of the Number of Molecules of a Single mRNA Species in a Complex mRNA Preparation, *Journal of Theoretical Biology*, Bd. 221(4), S. 615–624, 2003.

[78] Illumina, Inc., TruSeq Technology: Illumina Adapter Sequences Document, Document #1000000002694v00, online 11.08.2015 - `https://support.illumina.com/downloads/illumina-customer-sequence-letter.html`.

[79] Illumina, Inc., TruSeq Technology: TruSeq® Small RNA Library Prep Kit Reference Guide, Document #15004197v01, online 11.08.2015 - `https://support.illumina.com/downloads/truseq-small-rna-library-prep-guide-15004197.html`.

[80] M. A. Innis, D. H. Gelfand, J. J. Sninsky et al., PCR Protocols: A Guide to Methods and Applications, Academic Press, Inc. , New York, 1. Aufl., 2012.

[81] S. Islam, U. Kjällquist, A. Moliner et al., Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq, *Genome Research*, Bd. 21(7), S. 1160–1167, 2011.

[82] S. Islam, A. Zeisel, S. Joost et al., Quantitative single-cell RNA-seq with unique molecular identifiers, *Nature Methods*, Bd. 11(2), S. 163–166, 2014.

[83] M. Jain, I. T. Fiddes, K. H. Miga et al., Improved data analysis for the MinION nanopore sequencer, *Nature Methods*, Bd. 12(4), S. 351–356, 2015.

[84] X. Jiao und M. Armand, On a Reduced-Complexity Inner Decoder for the Davey-MacKay Construction, *ETRI Journal*, Bd. 34(4), S. 637–640, 2012.

[85] K. D. Kasschau, N. Fahlgren, E. J. Chapman et al., Genome-Wide Profiling and Analysis of Arabidopsis siRNAs, *PLoS Biology*, Bd. 5(3), 2007.

[86] E. Kawashima, L. Farinelli und P. Mayer, Method Of Nucleic Acid Amplification, Patent WO 1998/044151 A1, 1998.

[87] W. J. Kent, BLAT—The BLAST-Like Alignment Tool, *Genome Research*, Bd. 12(4), S. 656–664, 2002.

[88] T. Kivioja, A. Vähärautio, K. Karlsson et al., Counting absolute numbers of molecules using unique molecular identifiers, *Nature Methods*, Bd. 9(1), S. 72–74, 2012.

[89] J. J. Kozich, S. L. Westcott, N. T. Baxter et al., Development of a Dual-Index Sequencing Strategy and Curation Pipeline for Analyzing Amplicon Sequence Data on the MiSeq Illumina Sequencing Platform, *Applied and Environmental Microbiology, AEM*, Bd. 79(17), S. 5112–5120, 2013.

[90] D. Kracht und S. Schober, Using the Davey-MacKay code construction for barcodes in DNA sequencing, Proceedings of the International Symposium on Turbo Codes and Iterative Information Processing (ISTC), S. 142–146, 2014.

[91] D. Kracht und S. Schober, Insertion and deletion correcting DNA barcodes based on watermarks, *BMC Bioinformatics*, Bd. 16(1), S. 1–14, 2015.

[92] A. Krishnan, M. Sweeney, J. Vasic et al., Barcodes for DNA sequencing with guaranteed error correction capability, *Electronics Letters*, Bd. 47(4), S. 236–237, 2011.

[93] D. Laehnemann, A. Borkhardt und A. C. McHardy, Denoising DNA deep sequencing data—high-throughput sequencing errors and their correction, *Briefings in Bioinformatics*, 2015.

[94] R. Landstorfer, Laborbuch, ZIEL, Zentralinstitut für Ernährungs- und Lebensmittelforschung, TU München, 2015.

[95] B. Langmead und S. L. Salzberg, Fast gapped-read alignment with Bowtie 2, *Nature Methods*, Bd. 9(4), S. 357–359, 2012.

[96] B. Langmead, C. Trapnell, M. Pop et al., Ultrafast and memory-efficient alignment of short DNA sequences to the human genome, *Genome Biology*, Bd. 10(3), 2009.

[97] H. Latif, H. J. Li, P. Charusanti et al., A Gapless, Unambiguous Genome Sequence of the Enterohemorrhagic Escherichia coli O157:H7 Strain EDL933, *Genome Announcements*, Bd. 2(4), 2014.

[98] J. Y. Lee, J. Y. Park und B. Tian, Identification of mRNA Polyadenylation Sites in Genomes Using cDNA Sequences, Expressed Sequence Tags, and Trace, Methods in Molecular Biology, Bd. 419, S. 23–37, Springer, Berlin/Heidelberg, 2008.

[99] V. I. Levenshtein, Binary Codes Capable of Correcting Deletions, Insertions and Reversals, *Soviet Physics Doklady*, Bd. 10(8), S. 707–710, 1966.

[100] H. Li und R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform, *Bioinformatics*, Bd. 25(14), S. 1754–1760, 2009.

[101] H. Li, J. Ruan und R. Durbin, Mapping short DNA sequencing reads and calling variants using mapping quality scores, *Genome Research*, Bd. 18(11), S. 1851–1858, 2008.

[102] R. Li, Y. Li, K. Kristiansen et al., SOAP: short oligonucleotide alignment program, *Bioinformatics*, Bd. 24(5), S. 713–714, 2008.

[103] J. Y. Lim, J. W. Yoon und C. J. Hovde, A Brief Overview of Escherichia coli O157:H7 and Its Plasmid O157, *Journal of Microbiology and Biotechnology*, Bd. 20(1), S. 5, 2010.

[104] W. Liu, S. Wang, L. Gao et al., DNA Sequence Design Based on Template Strategy, *Journal of Chemical Information and Computer Sciences*, Bd. 43(6), S. 2014–2018, 2003.

[105] N. J. Loman, R. V. Misra, T. J. Dallman et al., Performance comparison of benchtop high-throughput sequencing platforms, *Nature Biotechnology*, Bd. 30(5), S. 434–439, 2012.

[106] D. J. MacKay und R. M. Neal, Near Shannon Limit Performance of Low Density Parity Check Codes, *Electronics Letters*, Bd. 32(18), S. 1645–1646, 1996.

[107] F. J. MacWilliams und N. J. A. Sloane, The Theory of Error Correcting Codes, North Holland Publishing Co., Amsterdam, 9. Aufl., 1977.

[108] M. F. Mansour und A. H. Tewfik, Convolutional Decoding in the Presence of Synchronization Errors, *IEEE Journal on Selected Areas in Communications*, Bd. 28(2), S. 218–227, 2010.

[109] J. L. Massey, Shift-Register Synthesis and BCH Decoding, *IEEE Transactions on Information Theory*, Bd. 15(1), S. 122–127, 1969.

[110] F. Mathieu-Daudé, J. Welsh, T. Vogt et al., DNA rehybridization during PCR: the 'C_0t effect' and its consequences, *Nucleic Acids Research*, Bd. 24(11), S. 2080–2086, 1996.

[111] R. J. McEliece, A Public-Key Cryptosystem Based On Algebraic Coding Theory, Techn. Ber., Deep Space Network Progress Report 42-44, S.114–116, 1978.

[112] P. McInerney, P. Adams und M. Z. Hadi, Error Rate Comparison during Polymerase Chain Reaction by DNA Polymerase, *Molecular Biology International*, Bd. 2014, 2014.

[113] F. Meacham, D. Boffelli, J. Dhahbi et al., Identification and correction of systematic error in high-throughput sequence data, *BMC Bioinformatics*, Bd. 12(1), S. 451, 2011.

[114] M. L. Metzker, Sequencing technologies—the next generation, *Nature Reviews Genetics*, Bd. 11(1), S. 31–46, 2010.

[115] M. Meyer, U. Stenzel, S. Myles et al., Targeted high-throughput sequencing of tagged nucleic acid samples, *Nucleic Acids Research*, Bd. 35(15), 2007.

[116] A. S. Mikheyev und M. M. Tin, A first look at the Oxford Nanopore MinION sequencer, *Molecular Ecology Resources*, Bd. 14(6), S. 1097–1102, 2014.

[117] A. E. Minoche, J. C. Dohm, H. Himmelbauer et al., Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and Genome Analyzer systems, *Genome Biology*, Bd. 12(11), 2011.

[118] K. Mir, K. Neuhaus, M. Bossert et al., Short Barcodes for Next Generation Sequencing, *PLoS ONE*, Bd. 8(12), 2013.

[119] M. Morange, What history tells us XIII. Fifty years of the Central Dogma, *Journal of Biosciences*, Bd. 33(2), S. 171–175, 2008.

[120] D. E. Muller, Application of Boolean Algebra to Switching Circuit Design and to Error Detection, *Transactions of the IRE Professional Group on Electronic Computers*, Bd. 3, S. 6–12, 1954.

[121] K. Nakamura, T. Oshima, T. Morimoto et al., Sequence-specific error profile of Illumina sequencers, *Nucleic Acids Research*, 2011.

[122] S. B. Needleman und C. D. Wunsch, A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins, *Journal of Molecular Biology*, Bd. 48(3), S. 443–453, 1970.

[123] P.-M. Nguyen, M. A. Armand und W. TONG, On the Watermark String in the Davey-MacKay Construction, *IEEE Communications Letters*, Bd. 17(9), S. 1830–1833, 2013.

[124] R. Overbeck, Structural Attacks for Public Key Cryptosystems based on Gabidulin Codes, Journal of Cryptology, Bd. 21(2), S. 280–301, Springer, Berlin/Heidelberg, 2008.

[125] F. Ozsolak, Third Generation Sequencing Techniques and Applications to Drug Discovery, *Expert Opinion on Drug Discovery*, Bd. 7(3), S. 231–243, 2012.

[126] C. S. Pareek, R. Smoczynski und A. Tretyn, Sequencing technologies and genome sequencing, Journal of Applied Genetics, Bd. 52(4), S. 413–435, Springer, Berlin/Heidelberg, 2011.

[127] N. T. Perna, G. Plunkett, V. Burland et al., Genome sequence of enterohaemorrhagic Escherichia coli O157:H7, *Nature*, Bd. 409(6819), S. 529–533, 2001.

[128] F. Picard, C. Dressaire, L. Girbal et al., Examination of post-transcriptional regulations in prokaryotes by integrative biology, *Comptes Rendus Biologies*, Bd. 332(11), S. 958–973, 2009.

[129] F. Qiu, L. Guo, T.-J. Wen et al., DNA Sequence-Based „Bar Codes" for Tracking the Origins of Expressed Sequence Tags from a Maize cDNA Library Constructed Using Multiple mRNA Sources, *Plant Physiology*, Bd. 133(2), S. 475–481, 2003.

[130] W. H. P. Quattrocchi, Informations-und Codierungstheorie, Springer, Berlin/Heidelberg, 1983.

[131] C. A. Raabe, T.-H. Tang, J. Brosius et al., Biases in small RNA deep sequencing data, *Nucleic Acids Research*, Bd. 42(3), S. 1414–1426, 2014.

[132] L. R. Rabiner, A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, *Proceedings of the IEEE*, Bd. 77(2), S. 257–286, 1989.

[133] L. R. Rabiner und B.-H. Juang, An Introduction to Hidden Markov Models, *IEEE ASSP Magazine*, Bd. 3(1), S. 4–16, 1986.

[134] D. Ramsköld, S. Luo, Y.-C. Wang et al., Full-Length mRNA-Seq from single cell levels of RNA and individual circulating tumor cells, *Nature Biotechnology*, Bd. 30(8), S. 777–782, 2012.

[135] I. Reed, A Class of Multiple-Error-Correcting Codes and the Decoding Scheme, *Transactions of the IRE Professional Group on Information Theory*, Bd. 4, S. 38–49, 1954.

[136] I. S. Reed und G. Solomon, Polynomial Codes Over Certain Finite Fields, *ournal of the Society for Industrial and Applied Mathematics*, Bd. 8(2), S. 300–304, 1960.

[137] P. Régnier und C. M. Arraiano, Degradation of mRNA in bacteria: emergence of ubiquitous features, *BioEssays*, Bd. 22(3), S. 235–244, 2000.

[138] J. T. Robinson, H. Thorvaldsdóttir, W. Winckler et al., Integrative Genomics Viewer, *Nature Biotechnology*, Bd. 29(1), S. 24–26, 2011.

[139] M. G. Ross, C. Russ, M. Costello et al., Characterizing and measuring bias in sequence data, *Genome Biology*, Bd. 14(5), 2013.

[140] R. Roth, Introduction to Coding Theory, Cambridge University Press, 2006.

[141] F. Sanger und A. R. Coulson, A Rapid Method for Determining Sequences in DNA by Primed Synthesis with DNA Polymerase, *Journal of Molecular Biology*, Bd. 94(3), S. 441–448, 1975.

[142] F. Sanger, S. Nicklen und A. R. Coulson, DNA sequencing with chain-terminating inhibitors, *Proceedings of the National Academy of Sciences, PNAS*, Bd. 74(12), S. 5463–5467, 1977.

[143] N. Sarkar, Polyadenylation of mRNA in prokaryotes, *Annual Review of Biochemistry*, Bd. 66(1), S. 173–197, 1997.

[144] Y. Sasagawa, I. Nikaido, T. Hayashi et al., Quartz-Seq: a highly reproducible and sensitive single-cell RNA sequencing method, reveals non-genetic gene-expression heterogeneity, *Genome Biology*, Bd. 14(4), 2013.

[145] E. E. Schadt, S. Turner und A. Kasarskis, A window into third-generation sequencing, *Human Molecular Genetics*, Bd. 19(2), S. 227–240, 2010.

[146] K. Schilling, Theoretical Aspects of Overlapping Genes, Dissertation, Fakultät für Ingenieurwissenschaften und Informatik der Universität Ulm, 2015.

[147] M. Schirmer, U. Z. Ijaz, R. D'Amore et al., Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform, *Nucleic Acids Research*, 2015.

[148] M. W. Schmitt, S. R. Kennedy, J. J. Salk et al., Detection of ultra-rare mutations by next-generation sequencing, *Proceedings of the National Academy of Sciences, PNAS*, Bd. 109(36), S. 14508–14513, 2012.

[149] P. H. Sellers, On the Theory and Computation of Evolutionary Distances, *SIAM Journal on Applied Mathematics*, Bd. 26(4), S. 787–793, 1974.

[150] J. P. Shaffer, Multiple Hypothesis Testing, *Annual Review of Psychology*, Bd. 46(1), S. 561–584, 1995.

[151] C. Shannon, A Mathematical Theory of Communication, *Bell System Technical Journal*, Bd. 27, S. 379–423, 623–656, 1948.

[152] J. Shendure und H. Ji, Next-generation DNA sequencing, *Nature Biotechnology*, Bd. 26(10), S. 1135–1145, 2008.

[153] K. Shiroguchi, T. Z. Jia, P. A. Sims et al., Digital RNA sequencing minimizes sequence-dependent bias and amplification noise with optimized single-molecule barcodes, *Proceedings of the National Academy of Sciences, PNAS*, Bd. 109(4), S. 1347–1352, 2012.

[154] T. F. Smith und M. S. Waterman, Identification of Common Molecular Subsequences, *Journal of Molecular Biology*, Bd. 147(1), S. 195–197, 1981.

[155] D. A. Steege, Emerging features of mRNA decay in bacteria, *RNA*, Bd. 6(8), S. 1079–1090, 2000.

[156] F. Sun, The Polymerase Chain Reaction and Branching Processes, *Journal of Computational Biology*, Bd. 2(1), S. 63–86, 1995.

[157] M. T. Suzuki und S. J. Giovannoni, Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR, *Applied and Environmental Microbiology*, Bd. 62(2), S. 625–630, 1996.

[158] F. Tang, C. Barbacioru, Y. Wang et al., mRNA-Seq whole-transcriptome analysis of a single cell, *Nature Methods*, Bd. 6(5), S. 377–382, 2009.

[159] E. Tapia, F. Spetale, F. Krsticevic et al., DNA Barcoding through Quaternary LDPC Codes, *PLoS ONE*, Bd. 10(10), 2015.

[160] D. Thieffry und S. Sarkar, Forty years under the central dogma, *Trends in Biochemical Sciences*, Bd. 23(8), S. 312–316, 1998.

[161] H. Thorvaldsdóttir, J. T. Robinson und J. P. Mesirov, Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration, *Briefings in Bioinformatics*, Bd. 14(2), 2013.

[162] B. Tieke, Makromolekulare Chemie: Eine Einführung, John Wiley & Sons Inc., Hoboken/New Jersey, 2014.

[163] K. R. Tindall und T. A. Kunkel, Fidelity of DNA Synthesis by the Thermus aquaticus DNA Polymerase, *Biochemistry*, Bd. 27(16), S. 6008–6013, 1988.

[164] C. Trapnell, L. Pachter und S. L. Salzberg, TopHat: discovering splice junctions with RNA-Seq, *Bioinformatics*, Bd. 25(9), S. 1105–1111, 2009.

[165] R. Tsien, P. Ross, M. Fahnestock et al., DNA Sequencing, Patent WO 91/06678, 1991.

[166] E. Ukkonen, On approximate string matching, Foundations of Computation Theory, Bd. 158, S. 487–495, Springer, Berlin/Heidelberg, 1983.

[167] A. J. Viterbi, Error Bounds for Convolutional Codes and an Asymtotically Optimum Decoding Algorithm, *IEEE Transactions on Information Theor*, Bd. 13(2), S. 260–269, 1967.

[168] X. V. Wang, N. Blades, J. Ding et al., Estimation of sequencing error rates in short reads, *BMC Bioinformatics*, Bd. 13(1), S. 185, 2012.

[169] Z. Wang, M. Gerstein und M. Snyder, RNA-Seq: a revolutionary tool for transcriptomics, *Nature Reviews Genetics*, Bd. 10(1), S. 57–63, 2009.

[170] M. S. Waterman, T. F. Smith und W. A. Beyer, Some Biological Sequence Metrics, *Advances in Mathematics*, Bd. 20(3), S. 367–387, 1976.

[171] J. D. Watson und F. H. Crick, Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid, *Nature*, Bd. 171(4356), S. 737–738, 1953.

[172] Wikipedia, Wikimedia Commons (public domain), Genetische Code-Sonne: Die Codierung der Aminosäuren (außen) durch die Basentripletts auf der mRNA ist von innen (5') nach außen (3') zu lesen, 2009, online 29.10.2015 - `https://upload.wikimedia.org/wikipedia/commons/7/70/Aminoacids_table.svg`.

[173] Wikipedia, Wikimedia Commons (public domain), Schematische Darstellung der beiden DNA-Stränge während der Transkription (sense und antisense) und des entstehenden RNA-Transkripts, 2009, online 28.10.2015 - `https://upload.wikimedia.org/wikipedia/commons/3/36/DNA_transcription.svg`.

[174] Wikipedia, Wikimedia Commons (public domain), Translation an einem Ribosom (Schematisch), 2011, online 02.11.2015 - `https://upload.wikimedia.org/wikipedia/commons/7/70/Ribosom_funktion.png`.

[175] Wikipedia, Wikimedia Commons, User: LadyofHats Mariana, Ruiz / tikurion Michael, Biech (public domain), Schematische Darstellung der DNA-Replikation, 2008, online 28.10.2015 - `https://upload.wikimedia.org/wikipedia/commons/6/69/DNA_replication_de.svg`.

[176] Wikipedia, Wikimedia Commons, User:Dr d12 (CC BY-SA 3.0), Assembling a Nacatamale, 2007, online 25.11.2015 - `https://upload.wikimedia.org/wikipedia/commons/e/e6/DNAgel4wiki.png`.

[177] Wikipedia, Wikimedia Commons, User:Enzoklop (CC BY-SA 3.0), Grafik der Polymerase-Kettenreaktion, 2013, online 30.10.2015 - `https://upload.wikimedia.org/wikipedia/commons/9/96/Polymerase_chain_reaction.svg`.

[178] Wikipedia, Wikimedia Commons, User:Madprime (CC BY-SA 3.0), Chemische Struktur der DNA mit farbigen Beschriftungen um die 4 Basen, die Phosphate und die Desoxyribose zu kennzeichnen, 2009, online 28.10.2015 - `https://upload.wikimedia.org/wikipedia/commons/f/f0/Chemische_Struktur_der_DNA.svg`.

[179] S. Wu und U. Manber, Fast text searching: allowing errors, *Communications of the ACM*, Bd. 35(10), S. 83–91, 1992.

[180] X. Wu, G. Ji und Q. Q. Li, Poly (A)-Tag Deep Sequencing Data Processing to Extract Poly (A) Sites, Methods in Molecular Biology, Bd. 1255, S. 39–48, Springer, Berlin/Heidelberg, 2015.

[181] X. Yang, S. P. Chockalingam und S. Aluru, A survey of error-correction methods for next-generation sequencing, *Briefings in Bioinformatics*, Bd. 14(1), S. 56–66, 2013.

[182] W. Zheng, L. M. Chung und H. Zhao, Bias detection and correction in RNA-Sequencing data, *BMC Bioinformatics*, Bd. 12(1), S. 290, 2011.

PUBLIKATIONEN:

„Insertion and deletion correcting DNA barcodes based on watermarks“,
D. Kracht und S. Schober,
BMC Bioinformatics, Bd. 16(1), S. 1-14, 2015.

„Using the Davey-MacKay code construction for barcodes in DNA sequencing“,
D. Kracht und S. Schober,
Proceedings of the International Symposium on Turbo Codes and Iterative Information Processing (ISTC), S. 142-146, 2014.

„Canalizing Boolean Functions Maximize the Mutual Information“,
J. G. Klotz, D. Kracht, M. Bossert und S. Schober,
IEEE Transactions on Information Theory, Bd.60(4), S. 2139-2147, 2014.

„Canalizing Boolean Functions Maximize the Mutual Information“,
J. G. Klotz, D. Kracht, M. Bossert und S. Schober,
Proceedings of International ITG Conference on Systems, Communication and Coding (SCC), S. 1-6, 2013

„Inferring Boolean functions via higher-order correlations“,
D. Kracht, M. Maucher, S. Schober, M. Bossert und H. A. Kestler
Computational Statistics, Bd. 29(1-2), S. 97-115, 2012.

„Detecting controlling nodes of boolean regulatory networks“,
S. Schober, D. Kracht, R. Heckel und M. Bossert,
EURASIP Journal on Bioinformatics and Systems Biology, Bd. 2011(1), S.1-10, 6, 2011