Technische Universität Berlin. Fakultät I
Institut für Sprache und Kommunikation

Fachgebiet Audiokommunikation

Magisterarbeit

Zur Sonifikation genetischer Daten

Emanuel Schuchart
[redacted]

Berlin, den 2. 10. 2011

Erstgutachter: Prof. Dr. Stefan Weinzierl
Zweitgutachter: Dr. Hans-Joachim Maempel

Eidesstattliche Erklärung

Hiermit versichere ich, dass ich die vorliegende Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Alle Ausführungen, die anderen veröffentlichten oder nicht veröffentlichten Schriften wörtlich oder sinngemäß entnommen wurden, habe ich kenntlich gemacht.

Die Arbeit hat in gleicher oder ähnlicher Fassung noch keiner anderen Prüfungsbehörde vorgelegen.

_____ _____

Ort, Datum Emanuel Schuchart

… … … … … …

Inhaltsverzeichnis

1. Einleitung.....1

2. Hintergrund.....2

2.1 Molekulargenetik.....2

2.1.1 DNA.....2
2.1.2 Der genetische Code.....3
2.1.3 Struktur von DNA und Sequenztypen.....4
2.1.4 Klassische DNA-Sequenzierung.....6
2.1.5 Sequenziermethoden der nächsten Generation.....9

2.2 Auditory Display und Sonifikation.....12

2.2.1 Definitionen.....13
2.2.2 Auditory Display.....15
2.2.3 Sonifikation.....17
2.2.4 Systematisierung.....23
2.2.5 Sonifikation und Kunst – Sonifikation als Kunst 26

3. Sonifikation Genetischer Daten – Status Quo.....29

3.1 In der Wissenschaft.....29

3.2 In der Kunst.....35

4. Möglichkeiten zur Sonifikation genetischer Daten.....41

4.1 Einsatzszenarien.....41

4.2 Genetische Daten – Dimensionen und Darstellung.....43

4.2.1 DNA-Sequenzen.....43
4.2.2 Gene und Genversionen.....48
4.2.3 Chromosom / Chromosomabschnitt.....49

4.3 Vorschläge für Sonifikationen.....50

4.3.1 Sequenz-Sonifikation.....50
4.3.2 Repeat-Finder.....50
4.3.3 Transcript-Fingerprint.....52
4.3.4 Chromosom-Fingerprint.....55
4.3.5 Akustischer DNA-Browser.....57
4.3.6 Data Mining mittels MBS.....59

5. Programm.....61

5.1 Vorüberlegungen.....61

5.1.1 Warum ein neues Programm.....61
5.1.2 Weitere Vorüberlegungen.....63

5.2 Input.....64

5.2.1 Beschreibung der Dateien.....64
5.2.2 Vor- und Nachteile der Aufteilung.....67
5.2.3 Weitere Problematische Aspekte.....68

5.3 Funktionsumfang.....70

5.3.1 Grafiken....70
5.3.2 Sonifikationen....74
5.3.3 Weitere Funktionen....76
5.4 Implementierung....77
5.4.1 Klassenstruktur einer Grafik....77
5.4.2 Datenbehandlung....79
5.4.3 Sonifikationsklassen....80
5.4.4 Navigations- und Hilfsklassen....80
6. Diskussion....82
6.1 Diskussion des Programms....82
6.2 Zur Zusammenarbeit im Rahmen des Projekts....84
6.3 Zusammenfassung und Fazit....88
7. Literatur....90

1. Einleitung

New projects, such as the Sloan Digital Sky Survey [...] and the Human Genome Project [...], are yielding huge data sets that must be managed and explored.
([1], p. 5 in Version von 2010)

Die Darstellung von Daten mittels Audio wird, von einigen Pionieren abgesehen, seit etwa 20 Jahren systematisch und mit wachsendem Eifer erforscht, hat sich aber bisher kaum gegen den Primat der Visualisierung durchsetzen können. Molekulargenetik ist zwar schon ein etwas älterer Zweig, doch sind auch dort die letzten 20 Jahre von immensen Fortschritten geprägt.
Die Idee, dass die Molekulargenetik als eine extrem datenintensive Wissenschaft von Entwicklungen im Bereich Auditory Display profitieren könnte, ist, wie das einleitende Zitat aus einem Bericht der ICAD, in dem sie sich um Forschungsgelder bewirbt, belegt, schon eine Weile in der Welt, trotzdem sind sich die beiden Felder bisher selten begegnet. Die Begegnung beider Gebiete ist der Kern dieser Arbeit.

Im ersten Teil werden die Hintergründe beider Felder beleuchtet. Im Bereich Auditory Display wird dabei auch die Suche nach endgültigen Definitionen sowie die Kontroverse im Hinblick auf den Umgang mit künstlerischen Projekten reflektiert.
Der folgende Teil ist eine Bestandsaufnahme. Die bislang umgesetzten Sonifikationen molekulargenetischer Daten werden zusammengetragen, die zugrunde liegenden Prinzipien werden analysiert und Gemeinsamkeiten werden herausgearbeitet. Dabei werden sowohl wissenschaftlich als auch künstlerisch motivierte Ansätze betrachtet.
Der dritte Teil wirft einen systematischen Blick auf die Molekulargenetik und untersucht, welche Datendimensionen es gibt, wie diese bisher dargestellt werden. Die Betrachtungen erfolgen mit dem Ziel, eine Art Möglichkeitsraum zu erarbeiten, der eine Einschätzung erlaubt, ob und wo Auditory Display im Bereich der Molekulargenetik Vorteile verspricht. Ausgehend von den erarbeiteten Möglichkeiten werden zudem exemplarisch einige Sonifikationen skizziert, die sich auf verschiedene Aspekte der Daten und verschiedene Anwendungsfelder konzentrieren.
Am Anfang der Beschäftigung mit der Thematik stand ein Kompositionsprojekt mit genetischen Daten der französischen Komponistin Clara Maïda, die auf der Suche nach einem Programm war, das ihr die kreative Arbeit mit den großen Datenmengen erleichtern würde. Einige der vorgeschlagenen Sonifikationen sind im Zusammenhang mit diesem Projekt entstanden. Das Programm mit seinen Funktionen steht im Mittelpunkt eines weiteren Teils dieser Arbeit.
Im letzten Teil schließlich werden die Ergebnisse und der Prozess diskutiert.

2. Hintergrund

In diesem Teil werden die Grundlagen erläutert, die zum Verständnis dieser Arbeit notwendig sind. Zu Beginn steht eine knappe Einführung in die gentechnischen und molekularbiologischen Grundlagen. Anschließend wird das Gebiet 'Auditory Display' im Allgemeinen und Sonifikation im Besonderen vorgestellt. Neben der Einführung der wichtigsten Begriffe und Techniken liefert das Kapitel einen Beitrag zur Suche nach einer allgemein anerkannten, schlüssigen Definition und blickt auf Anwendungen von Daten-Klang-Transformationen im Bereich der Kunst.

2.1 MOLEKULARGENETIK

Wissensbasis dieses Kapitels waren die folgenden beiden Bücher: [2], [3]. Auf alle weiteren Quellen wird im Text verwiesen. Die Fragen, die beantwortet werden, sind:

- Was ist DNA und was ist der genetische Code?
- Wie wird er von der Zelle gelesen und übersetzt?
- Wie ist die DNA strukturiert und welche wichtigen Regionen gibt es?
- Wie erfolgt die Sequenzierung also das Transkribieren des Codes seitens des Menschen?

2.1.1 DNA

Seit 1944 ist eindeutig nachgewiesen, dass die Erbinformation nahezu aller Lebewesen auf langen Makromolekülen, der Desoxyribonukleinsäure oder kurz DNA.[1,] gespeichert ist
Die Bausteine dieser Makromoleküle sind die sogenannten Nukleotide. Nukleotide bestehen aus 3 Teilen - einer Phosphorsäure, einem Zucker mit 5 Kohlenstoffatomen (einer Pentose) und einer Nukleinbase. In DNA treten vier verschiedene Nukleinbasen auf: Adenin, Guanin, Cytosin oder Thymin. Der Zucker, in DNA ist es Desoxyribose, ist das Zentrum des Nukleotids und bildet eine Art Ring, mit dem Nukleinbase und Phosphatrest über die Kohlenstoffatome verbunden sind. Die Base sitzt am C-Atom 1' und das Phosphat am C-Atom 5'. Die Verkettung zweier Nukleotide erfolgt über eine Phosphordiesther-Bindung zwischen dem Phosphatrest des einen und dem 3'-Ende des C-Atoms des anderen Nukleotids.[2] Jede Nukleinsäure hat also ein freies 5'- und ein freies 3'-Ende. Die Verlängerung erfolgt immer am 3'-Ende.
Die Nukleinbasen sind bisher nur mit dem Zucker verbunden, hängen sozusagen 'in der Luft'.

1 DNA ist das Akronym des englischen 'Deoxyribonucleic Acid'.
2 Dabei wird ein Teil der Phosphorsäure abgespalten.

2.1.1 DNA

In der Zelle ist DNA normalerweise doppelsträngig, das heißt, zwei der langen Ketten sind parallel verknüpft. Die Verknüpfung erfolgt über Wasserstoffbrücken, die sich zwischen zwei Basen bilden. Dabei sind je zwei Basen komplementär. Adenin (A) bildet nur mit Thymin (T), Guanin (G) nur mit Cytosin(C) Wasserstoffbrücken aus. Dadurch sind auch beide Stränge komplementär. Kennt man die Sequenz des einen Strangs, so kennt man auch die des anderen. Die Doppelstrangstruktur macht die DNA sehr stabil. Die räumliche Struktur dieses Doppelstrangs ist die berühmte Doppelhelix.

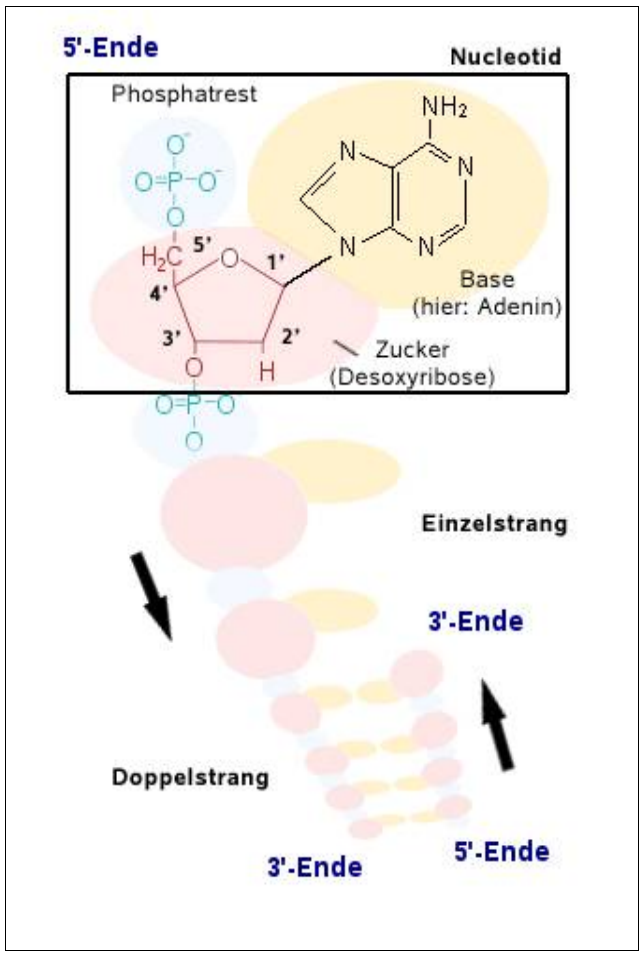

Abbildung 1: Schematischer Aufbau eines DNA-Moleküls

Die gesamte Erbinformation eines Lebewesens, das Genom, ist bei Eukaryonten[3] auf mehrere solcher Doppelstränge aufgeteilt und befindet sich locker gepackt als Komplex aus DNA und verschiedenen Gerüstproteinen im Zellkern[4]. Diesen Komplex, in dem die DNA in mehrere Stufen aufgewickelt ist, nennt man Chromatin. Bei der Mitose verdichten sich die Chromatinfäden sehr stark und werden als Chromosomen sichtbar.
Die Anzahl und auch die Länge der Chromosomen ist für verschiedene Arten sehr verschieden und hängt nicht unbedingt von der Komplexität des jeweiligen Organismus ab.

2.1.2 Der genetische Code

Wie im vorhergehenden Abschnitt erläutert, ist DNA eine Kette von Nukleotiden mit einer der vier Basen A, T, G oder C. Diese vier verschiedenen Nukleotide funktionieren wie ein Alphabet mit vier Buchstaben respektive Zeichen. Je drei Basen, ein Triplett, bilden ein Wort, ein sogenanntes Codon. Die meisten dieser Wörter stehen für eine Aminosäure, einige wenige, die Start- und Stop-Codons, haben syntaktische Funktion. Aus vier Buchstaben lassen sich $4^3 = 64$ verschiedene Wörter bilden. Die Natur verwendet zum Bau von Proteinen jedoch nur 20 Aminosäuren. Der genetische Code verfügt also über mehr Signifikanten, als er Signifikate hat.

3 Lebewesen mit Zellkern.
4 Wichtig zu erwähnen ist hierbei, dass jede Zelle des Körpers einen vollständigen Satz des Erbgutes beinhaltet

Es gibt zwei Möglichkeiten, was mit einer solchen Überzahl an Wörtern geschehen kann: Entweder tragen sie keine Bedeutung, oder sie haben die gleiche Bedeutung wie ein anderes Wort. Im Fall des genetischen Codes gilt das Zweite, sprich, mehrere Codons codieren dieselbe Aminosäure. Der genetische Code ist somit ein degenerierter Code. Die Aufteilung der Redundanz ist dabei keineswegs gleichmäßig. Einige Aminosäuren werden von nur einem Codon codiert, andere von sechs. Umgekehrt stehen nahezu alle Codons für genau eine einzige Aminosäure. Die größte Ausnahme ist das Triplet 'ATG', das sowohl Methionin als auch ein Startsignal für die Proteinsynthese codieren kann[5]. Drei weitere Codons codieren gar keine Aminosäure, sondern markieren das Ende einer Sequenz. Daher heißen sie auch Stop-Codons. Eine Code-Sequenz ist immer nur ein kurzer Abschnitt des DNA-Moleküls. Start- und Stop-Codons sind daher essentiell, um den Beginn und das Ende einer solchen funktionalen Sequenz anzuzeigen.
Die Aminosäuren bilden wiederum eine Art Alphabet. Sie sind aber keine abstrakten Codewörter, sondern die realen Bausteine aller Proteinen. Ein Protein ist immer eine Kette von Aminosäuren. Deren Abfolge ist entscheidend für die räumliche Struktur, die physikalischen und chemischen Eigenschaften und somit für die Funktion des Proteins.

Das Lesen des genetischen Codes in der Zelle erfolgt in drei Stufen. Zunächst synthetisiert eine RNA-Polymerase anhand des jeweils komplementären Strangs des Abschnitts, der ein Protein codiert, eine RNA-Kopie. RNA ist ebenfalls eine Nukleinsäure und der DNA sehr ähnlich. Statt Desoxyribose haben die Nukleotide aber den Zucker Ribose, die Base Thymin ist durch Uracil ersetzt, und RNA ist zumeist einzelsträngig.
Der fertige RNA-Strang, der eine inhaltlich exakte Kopie der DNA-Sequenz enthält, wird auch Prä-mRNA genannt. Im nächsten Schritt wird diese Prä-mRNA gespleißt, das heißt, es werden bestimmte Abschnitte (vgl. 2.1.3) herausgeschnitten und die Schnittstellen verbunden. Dabei kann eine bestimmte Prä-mRNA oft auf verschiedene Arten gespleißt werden, bildet in diesen Fällen also die Basis verschiedener Proteine. Das Ergebnis des Spleißens ist die sogenannte mRNA (kurz für Messenger-RNA), die nun den Zellkern verlässt und die codierten Informationen zu den Ribosomen trägt. Die Ribosomen decodieren die Codon-Information der mRNA und bauen so eine Kette aus Aminosäuren, ein Protein, zusammen. Schon eine fälschlich eingebaute Aminosäure kann die Struktur des Proteins derart modifizieren, dass es komplett dysfunktional wird. Das wiederum bedeutet, dass schon eine einziges falsches Nukleotid in der DNA für eine solche Fehlfunktion verantwortlich sein kann.

2.1.3 Struktur von DNA und Sequenztypen

Wie bisher erläutert, ist DNA eine lange Kette von vier verschiedenen, durch ihre Base charakterisierten Nukleotiden. Je drei Nukleotide bilden ein Codewort für eine Aminosäure und diese wiederum sind die Bausteine der Proteine.
Bei den höher entwickelten Lebewesen, präziser bei Eukaryonten, ist es nun so, dass nur ein sehr kleiner Teil des gesamten Genoms tatsächlich Baupläne für Proteine enthält. Von Spezies zu Spezies sind die Unterschiede sehr groß. Beim Menschen sind es tatsächlich nur etwa drei

5 In Prokaryonten gibt es zwei weitere Codons mit dieser Doppelbedeutung, wobei ATG auch dort am häufigsten als Startsignal verwendet wird.

Prozent der gesamten DNA, die in Proteine übersetzt werden. Diesen Teil bezeichnet man als protein-codierende DNA. Die codierenden Abschnitte sind fast über die gesamte Länge der Chromosomen verteilt. Lediglich an den Enden (Telomere) und dem Zentromer findet sich gar keine codierende DNA. Diese Aufsplitterung der Erbinformation ist sogar innerhalb des Bauplans für ein spezifisches Protein noch vorhanden. Das heißt, dass auch die Basensequenz, die ein bestimmtes Protein codiert, meist nicht 'am Stück' auf dem DNA-Strang vorliegt, sondern von im Vergleich zum codierenden Stück langen Abschnitten unterbrochen wird, die für den Bau des spezifischen Proteins nicht verwendet werden. Das heißt jedoch nicht, dass diese Abschnitte ausschließlich aus nicht-codierender DNA bestehen. Sie können vielmehr sehr wohl Teile des Codes eines anderen Proteins enthalten.

Die gesamte, ein Protein codierende Sequenz nennt man CDS (für 'Coding Sequence') (vgl. [4][6]), die Abschnitte, die eine CDS unterbrechen, nennt man Introns (ebd.). Die Bezeichnung Intron rührt daher, dass diese Sequenzinformationen den Zellkern nicht verlassen. Sie werden zwar zunächst in RNA übersetzt, vor dem Verlassen jedoch durch das bereits angesprochene Spleißen entfernt. Das Gegenstück zum Intron ist das Exon (ebd.). Als Exon bezeichnet man eine DNA-Sequenz, die, in RNA übersetzt, den Zellkern verlässt. Exons und CDSs sind sich definitorisch zwar ähnlich jedoch nicht exakt dasselbe, denn ein Exon kann neben Teilen der codierenden Sequenz (der CDS) auch noch 5'UTR und 3'UTR enthalten. UTRs, kurz für 'untranslated regions', sind regulatorische Sequenzen. Die Zahlen geben an, an welchem Ende der Sequenz der jeweilige Abschnitt liegt. Diese kurzen Abschnitte sind Teil der mRNA, und liefern zur Proteinsynthese nötige Steuerinformationen für die Ribosomen.
Die Verwendung des Begriffs 'Gen' wurde bisher bewusst vermieden. Der Grund dafür ist, dass im Gegensatz zu den oben genannten Begriffen keine allgemein gültige, eindeutige Definition gegeben werden kann. Wie Helen Pearson in einem Artikel in 'Nature' ([5]) erörtert, ist der klassische Genbegriff, nach dem ein Gen ein für eine phänotypische Ausprägung verantwortlicher, sprich, ein Protein codierender Abschnitt auf der DNA ist, nicht annähernd weit genug gefasst, um alle in der Natur auftretenden Phänomene und Konstellationen zu erfassen. Ein und dieselbe DNA-Sequenz kann beispielsweise Teil der CDS zweier verschiedener Proteine sein, oder der Code für ein komplettes Protein kann sich innerhalb eines Introns eines anderen 'Gens' befinden. Für die Expression eines Codes entscheidende regulatorische Sequenzen können sogar auf einem anderen Chromosom liegen als der Code selbst. Eine ausführliche Beschreibung dieser und anderer Konstellationen findet sich zusammen mit einer neuen Definition für 'Gen' in [6]. Im Kontext dieser Arbeit bezeichnet der Begriff 'Gen' einen DNA-Abschnitt, innerhalb dessen sich der komplette Code für ein oder mehrere komplette, verwandte Proteine befindet. Ein Gen kann also mehrere Versionen haben. Die CDS eines bestimmten Proteins wird als 'Gen-Version' oder Transcript bezeichnet. Liegen codierende Abschnitte zweier nicht verwandter Proteine im selben DNA-Abschnitt, wird von zwei Genen gesprochen.

Neben den bereits genannten Sequenztypen (CDS, Intron, Exon) gibt es zahlreiche weitere Klassen von DNA-Sequenzen. Zu den Klassen mit bekannter Funktion gehören Binding Sites und regulatorische Sequenzen wie Promoter und Enhancer sowie RNA-Gene, also Sequenzen, die den Bauplan für funktionale RNA-Moleküle enthalten.

6 Hier finden sich die Richtlinien für die Annotation molekulargenetischer Sequenzen. Zusätzlich finden sich Definitionen für viele Regionstypen.

Eine andere Klasse, deren Funktion bislang jedoch nur teilweise klar ist, bildet die sogenannte repetitive DNA. Sie ist dadurch gekennzeichnet, dass sich eine kurze Sequenz mehrmals, teils sogar hundertfach wiederholt. Man unterscheidet zwischen lokalen Wiederholungen (Tandem Repeats), die 'am Stück' vorliegen, und solchen, die über das Genom verteilt sind (Interpersed Repeats). Bei Tandem Repeats, man spricht je nach Länge der wiederholten Einheit auch von Mikro- und Minisatelliten, werden für den genetischen Fingerabdruck herangezogen. Repetitive DNA macht insgesamt etwa 40 % des menschlichen Genoms aus, und gerade die lokalen Wiederholungen sind ein großes Problem bei der Genomsequenzierung.

Insgesamt lässt sich bei weitem nicht das gesamte Erbgut funktional oder formal kategorisieren und man kann annehmen, dass ein Teil der DNA tatsächlich aus zufälligen, funktionslosen Sequenzen besteht. Wie groß dieser Teil jedoch ist, kann man nicht vorhersagen, und es ist nicht ausgeschlossen, dass einige funktionale Sequenztypen noch nicht entdeckt wurden.
Im weiteren Verlauf der Arbeit spielen hauptsächlich die Regionen CDS, Intron und Exon sowie die Tandem Repeats eine Rolle.

2.1.4 Klassische DNA-Sequenzierung

DNA-Sequenzen stehen als Basis für Sonifikation mit im Zentrum dieser Arbeit. Die Frage, wie diese Daten zustande kommen, ist zum Verständnis eigentlich nicht zwingend notwendig. Bei der Beschäftigung mit der Thematik schien mir diese Frage für ein vollständiges Bild aber trotzdem nicht ungeklärt bleiben zu können. Dieser und der folgende Abschnitt geben daher einen Überblick darüber.

DNA-Sequenzierung ist das Transkribieren der Basensequenz 'von außen'. Der technische Fortschritt auf diesem Gebiet ist rasant, und heutige Sequenzierautomaten ermöglichen Lesegeschwindigkeiten, die vor zehn Jahren noch unvorstellbar waren.
Zunächst soll das klassische Sanger-Verfahren vorgestellt werden, das bis zu Beginn dieses Jahrtausend quasi die einzige weithin verwendete Methode war, und das auch heute noch in kleinen Labors verwendet wird. Die Sequenzierung eines DNA-Abschnitts mittels der klassischen Methode umfasst folgende fünf Arbeitsschritte:

- Herstellung von kurzen Fragmenten
- Amplifikation dieser Fragmente
- Herstellung eines Sets von Nested Fragments
- Lesen der Sequenz
- Zusammensetzen der Teile

Der erste Schritt ist auch für die neuen Verfahren verbindlich, da alle bis dato realisierten Methoden nur sehr kurze DNA-Moleküle transkribieren können. Bei der klassischen Sanger-Methode liegt die maximale Länge beispielsweise bei ca. 300 Basenpaaren (Bp). Ein Abschnitt von 200 kBp (200.000 Bp) muss also zunächst in kleine Teile zerlegt werden. Diese Zerlegung erfolgt entweder physikalisch mittels Ultraschall oder molekularbiologisch mit Hilfe von Restriktionsnukleasen. Das sind Enzyme, die einen DNA-Strang bei einer bestimmten kurzen Sequenz zerschneiden. Es gibt Restriktionsnukleasen, deren Schneidesequenz lediglich drei

Basen lang ist, bei anderen sind es acht Basen. Je kürzer diese Trigger-Sequenz, desto häufiger taucht sie statistisch gesehen in einer langen DNA-Sequenz auf. Mittels verschiedener Restriktionsnukleasen lässt sich also ein DNA-Abschnitt in eine Reihe unterschiedlich langer Fragmente zerschneiden.

Nachdem man den zu sequenzierenden Abschnitt in mehr oder minder zufällige kurze Fragmente zerlegt hat, muss man diese Fragmente zunächst vervielfältigen. Ein Weg zur Vervielfältigung von DNA–Sequenzen ist die Klonierung. Das zu amplifizierende Stück wird dabei in das Erbgut eines sich schnell replizierenden Organismus integriert. In der Praxis wird oft mit Bakterien vom Typ E. Coli gearbeitet. Um das DNA-Fragment in die DNA des Bakteriums zu integrieren, bedient man sich eines Transporters, auch Vektor genannt. Natürliche Vektoren sind Bakteriophagen und Plasmide, es gibt aber auch eine Reihe von künstlichen Vektoren wie BAC, YAC oder PAC. Wenn sich das mit dem Fragment infizierte Bakterium teilt, repliziert es dabei auch die eingeschleuste DNA. Die Klonierung spielt vor allem beim Anlegen von Fragmentbibliotheken (Klonbibliotheken) eine große Rolle.
Die zweite Möglichkeit DNA zu vervielfältigen nutzt das Enzym, das auch bei der Zellteilung dafür verantwortlich ist – die DNA-Polymerase. Dieses Enzym komplettiert, genügend Baumaterial in Form freier Nukleotide vorrausgesetzt, einen Einzelstrang, indem es den Strang in Richtung seines 5'-Endes entlangläuft und jeweils das Nukleotid mit der komplementären Base anlagert.[7] Vorraussetzung dafür ist, dass ein kurzer Abschnitt des Strangs schon doppelsträngig vorliegt, da die Polymerase stets ein freies 3'-Ende braucht. Diesen kurzen, doppelsträngigen Abschnitt nennt man auch Primer. Man kann den Primer entweder auf die Sequenz abstimmen, wenn Teile der Sequenz des Fragments bekannt sind, oder man hängt an die Fragmente mittels DNA-Ligase ein künstlich hergestelltes Primertemplate an.
Für die Vervielfältigung mittels Polymerase Chain Reaction (PCR) wird das zu amplifizierende, mit einem Primer versehene Fragment zunächst denaturiert. In der Lösung befindet sich eine große Überzahl der Primerkomplementärsequenzen. Das führt dazu, dass sich die Einzelstränge nach dem Denaturieren in den meisten Fällen mit den Primertemplates verbinden anstatt mit ihrem ursprünglichen Komplementärstrang. Ausgehend vom Primer vervollständigt die Polymerase beide Einzelstränge wieder zu Doppelsträngen. Diese werden dann wieder denaturiert und die Einzelstränge wieder komplettiert. So lassen sich beliebig viele Kopien eines Fragments herstellen.

Der letzte Schritt vor dem eigentlichen Sequenzieren ist die Erstellung eines Sets von 'Nested Fragments'. Was das ist, lässt sich am besten erklären, indem man beschreibt, wie man es erzeugt. Das Set identischer DNA-Fragmente, die durch Amplifikation erzeugt wurden, wird auf vier Lösungen aufgeteilt, die jeweils DNA-Polymerasen und eine große Anzahl der vier Nukleotide enthält. Pro Lösung ist ein Teil jeweils einen Nukleotids manipuliert. Ihnen fehlt die Hydroxylgruppe am 3'-Ende des Moleküls, was zur Folge hat, dass kein weiteres Nukleotid 'angebaut' werden kann. Außerdem sind sie fluoreszierend markiert.
Die DNA-Stränge werden nun wieder denaturiert, und die Polymerasen beginnen, die Einzelstränge zu Doppelsträngen zu vervollständigen, wobei die Synthese immer dann abbricht, wenn sie ein manipuliertes Nukleotid einbauen. Bei hinreichend großer Zahl an identischen Fragmenten und einer gut ausbalancierten Anzahl manipulierter Nukleotide wird

7 Komplementäre Nukleotide lagern sich durchaus temporär selbst an einen Strang an. Die Wasserstoffbrücken allein sind jedoch sehr instabil. Nur die beidseitige Phosphordiestherbindung garantiert, dass das Nukleotid auf Dauer angelagert bleibt.

die Synthese bei jeder Position der Sequenz mindestens einmal abbrechen. Man erhält vier Lösungen mit doppelsträngigen Teilfragmenten unterschiedlicher Länge, wobei jede mögliche Länge mindestens einmal vertreten ist. Eine bestimmte Länge kann dabei nur in einer der Lösungen vorhanden sein, da die Synthese nur an einer manipulierten Base abbrechen kann. Genau das ist ein 'Set of Nested Fragments'. An einem Ende eines Fragments ist immer die Primersequenz, am anderen Ende eine manipulierte Base.

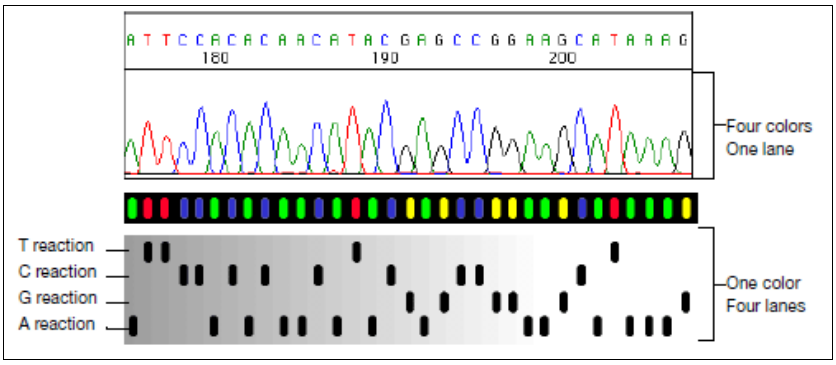

Abbildung 2: Vergleich von Gelelektrophorese bzw. Kapillarelektrophorese mittels einfarbig und mehrfarbig markierter Nukleotide; (Quelle: [7], p.9)

Um die DNA-Fragmente nun zu transkribieren verwendet man die Gel-Elektrophorese. Die von Natur aus negativ geladenen DNA-Moleküle unterschiedlicher Länge laufen in Richtung eines Plus-Pols durch ein Gel. Dieses Gel hat eine netzähnliche Struktur, und kurze DNA-Moleküle können sich schneller durch die Poren des Gels bewegen als längere. Auf diese Weise werden die Fragmente sozusagen der Länge nach sortiert. Gleich lange Moleküle laufen auch gleich schnell. Aufgrund der fluoreszierenden Markierung lässt sich Position der Fragmente feststellen. Aus einem Vergleich der Gelelektrophoresebilder der vier Lösungen lässt sich die Sequenz relativ genau ablesen.
Die Weiterentwicklung dieses Verfahrens zur Kapillar-Elektrophorese führte zu einer deutlichen Beschleunigung und weitgehenden Automatisierung des Leseprozesses. Die DNA-Fragmente durchlaufen das Gel nicht auf einer Platte, sondern in einer Kapillare, an deren Ausgang ein Laser die Fluoreszenz der jetzt verschiedenfarbig markierten Basen erregt. Ein Computer kann die Basen so eine nach der anderen anhand des Lichtblitzes identifizieren.

Im letzten Schritt, der Assembly, muss nun aus den bekannten Teilfragmenten, den sogenannten 'reads' die Sequenz des ursprünglichen, langen DNA-Abschnitt zusammengesetzt werden. Das erfolgt weitgehend automatisiert mittels verschiedener Algorithmen. Je größer die einzelnen Schnipsel dabei sind und je mehr sie überlappen, desto einfacher und weniger rechenintensiv ist dieser Prozess[8]. Die größten, bisher zum Teil unlösbaren Probleme bereitet die repetitive DNA. Enden zwei Fragmente mit Teilen desselben Minisatelliten, so ist es unmöglich festzustellen, wie lang der repetitive Abschnitt tatsächlich ist (vgl. Abb. 3). Die Assembly hat an dieser Stelle dann eine Lücke. Interpersed Repeats, also räumlich verteilte Wiederholungen, können dazu führen, dass zwei Fragmente als benachbart interpretiert werden, obwohl sie nicht benachbart sind, sondern nur zufällig im selben Repeat geschnitten

8 Moderne Kapillar-Elektrophorese-Methoden erlauben 'reads' mit bis zu 900 Bp Länge (vgl. [7], Seite 138 ff.).

wurden. Neben dem geringeren Rechenaufwand haben längere Teilstücke auch den Vorteil, dass die Wahrscheinlichkeit von Schnitten innerhalb eines Repeats geringer ist als bei kurzen Stücken.

Abbildung 3: Schematische Darstellung des Problems eines Mikrosatelliten am Ende eines DNA-Fragments

Obwohl sie durch die hohe Sequenziergeschwindigkeit und den niedrigen Preis ein Fortschritt sind, haben die im folgenden Abschnitt vorgestellten neuen Sequenziermethoden teils das Problem, dass sie nur sehr kurze 'reads' produzieren – eine Tatsache, die die Bioinformatik bei der Assembly vor große Herausforderungen stellt.

2.1.5 Sequenziermethoden der nächsten Generation

DNA-Sequenzierung mittels Gel- und Kapillarelektrophorese ist für Forschungsprojekte in denen es um die Entschlüsselung großer Mengen von DNA geht, wie beispielsweise die vergleichende Genomik, zu langsam und zu teuer. Die Entschlüsselung des menschlichen Genoms mittels des Sanger-Verfahrens nahm mehrere Jahre und unzählige Labors weltweit in Anspruch. In den letzten Jahren sind eine Reihe neuer Methoden entwickelt worden, die derartige Projekte für ein einzelnes Labor möglich machen. Eine sehr gute Übersicht über 'Next Generation Sequencing'-Sequenziermethoden findet sich in [8]. Drei dieser Methoden sollen hier kurz vorgestellt werden.

Das erste Verfahren ist das sogenannte 'true Single Molecule Sequencing' oder kurz tSMS™ der Firma Helicos BioScience (vgl. [9]). Hierbei werden viele kurze, einzelsträngige DNA-Fragmente mittels eines angehefteten Primers auf einer Platte, dem sogenannten Floater,

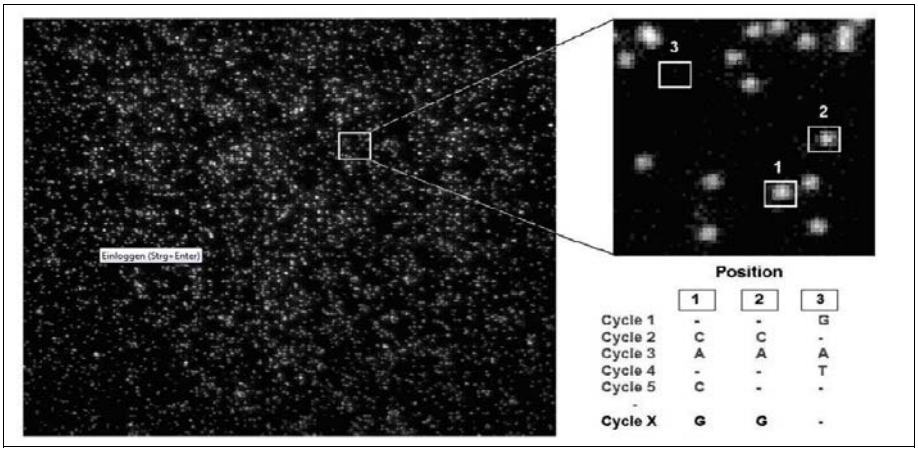

Abbildung 4: Roh-Output eines Helicos-Systems, (Quelle: [9], p.1)

fixiert. Nun werden eine große Anzahl eines fluoreszierend markierten Nukleotids und DNA-Polymerasen über den Floater gespült. Die Polymerasen werden den komplementären Strang, der bisher nur aus dem Primer besteht, um eine Base verlängern, wenn das gerade vorhandene Nukleotid zu dem des Einzelstrangs 'passt'. Passt es bei einem Strang nicht, so wird dieser auch nicht verlängert. Überall dort, wo eine Base eingebaut wurde, wird eine Fluoreszenz zu sehen sein und vom Computer registriert (vgl. Abb. 3).
Anschließend werden die überschüssigen Nukleotide sowie die fluoreszierenden Markierungen entfernt und der Prozess beginnt mit einer anderen Base erneut.
Ein Vorteile des Verfahrens ist, dass auf den Floatern bis zu einer Milliarde DNA-Fragmente gleichzeitig fixiert und analysiert werden können. Außerdem sind weder Amplifikation noch ein Set von 'Nested Fragments' notwendig, da die Fragmente Base für Base ausgelesen werden können. Der Nachteil ist jedoch die 'Read'-Länge von durchschnittlich 30 Basen, die sehr hohe Anforderungen an die Assembly-Routinen.

Das zweite Verfahren stammt von der Firma Pacific Bioscience und wird als SMRT™-Technology ('Single Molecule Real Time') vermarktet ([10]). Auf dem Boden eines Lochs in einer Aluminumplatte wird eine einzelne DNA-Polymerase zusammen mit einem einzelsträngigen, mit einem Primer versehenen DNA-Fragment fixiert. Darunter befindet sich ein Laser, der genau auf die Position des Polymerase-Moleküls fixiert ist, wo physikalisch der Einbau eines neuen Nukleotids erfolgt. Die Platte, auf der viele solcher Löcher integriert sind, liegt in einer Lösung mit den vier Nukleotiden. Diese sind mit vier Farben fluoreszierend markiert. Die Polymerasen komplettieren nun Base für Base ihren Einzelstrang. Immer wenn ein Nukleotid eingebaut wird, erregt der Laser die Fluoreszenz und die Optik registriert einen farbigen Lichtblitz. Die Folge der Lichtblitze liefert also direkt die Basenfolge. Das Verfahren ermöglicht das Lesen der Basen in der Geschwindigkeit, in der sie von der Polymerase am Einzelstrang fixiert werden. Ein Entfernen von Markern oder überschüssigen Nukleotiden ist nicht notwendig. Zudem ist wie bei tSMS™ auch hier keine Amplifikation der DNA-Fragmente nötig. Die 'Read'-Länge ist praktisch dadurch begrenzt, dass der energiereiche Laser die DNA-Polymerase mit der Zeit zerstört. Theoretisch sind 'Reads' mit mehr als 1000 Basen Länge möglich (vgl. [10]).

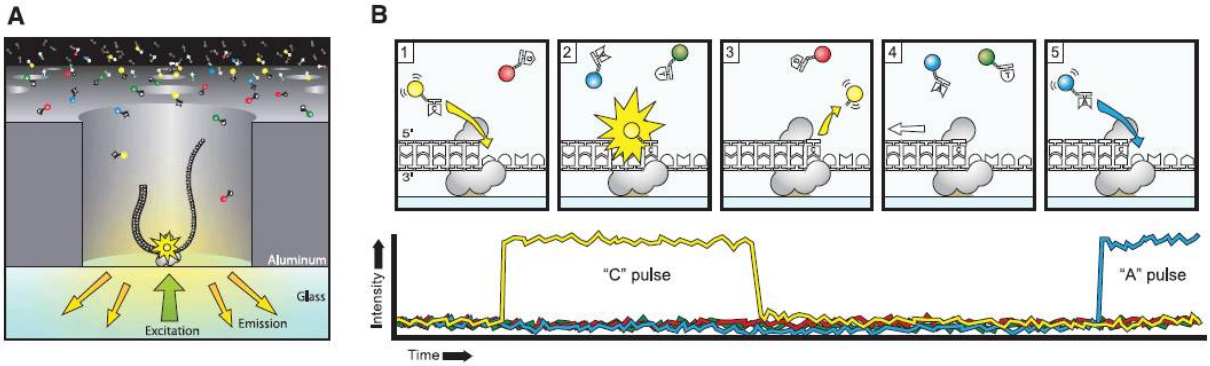

Abbildung 5: Funktionsweise des SMRT™-Systems; links: schematische Darstellung der Lochplatte; rechts: sequenzieller Einbau der Nukleotide und das zugehörige Zeitsignal der Lichtintensität (Quelle: [10], p. 134)

Das dritte Verfahren basiert auf Nanoporen und wird derzeit unter anderem von Forschern in Oxford entwickelt ([11]). Die Technologie benötigt weder Amplifikation noch fluoreszierende Labels und erlaubt, zumindest theoretisch, unbegrenzt lange 'Reads'. Ein DNA-Molekül wandert aufgrund seiner negativen Ladung Base für Base durch eine Mikropore in einem ansonsten hochisolierenden Material wie beispielsweise Graphen. Das gescheit entweder 'am Stück', oder die einzelnen Basen werden direkt über der Pore mittels einer Exonuklease vom DNA-Molekül abgeschnitten. Längs der Pore fließt ein Strom. Tritt nun ein Molekül, hier ein Nukleotid, durch die Pore hindurch, wird dieser Strom moduliert. Diese Modulation ist messbar und für die verschiedenen Basen charakteristisch. Die Modulationsfolge entspricht also genau der DNA-Sequenz. Auf einem Chip kann wie bei den anderen Verfahren auch eine große Anzahl solcher Poren integriert werden. Genaue Daten zu Sequenziergeschwindigkeit und -kosten liegen bisher nicht vor, doch spricht Clive Brown, Vizepräsident für Entwickelung und Informatik von Oxford Nanopore Technologies, im Interview mit Bio-IT World von einer Datenrate von *„[...]many tens of Megabytes/second.“ ([12], von Webseite)* für einen derartigen DNA-Chip, und Jay Flatley, CEO von Illumina, traut der Technologie zu, ein menschliches Genom für weniger als 1000 $ komplett zu lesen (zit. in [13]).

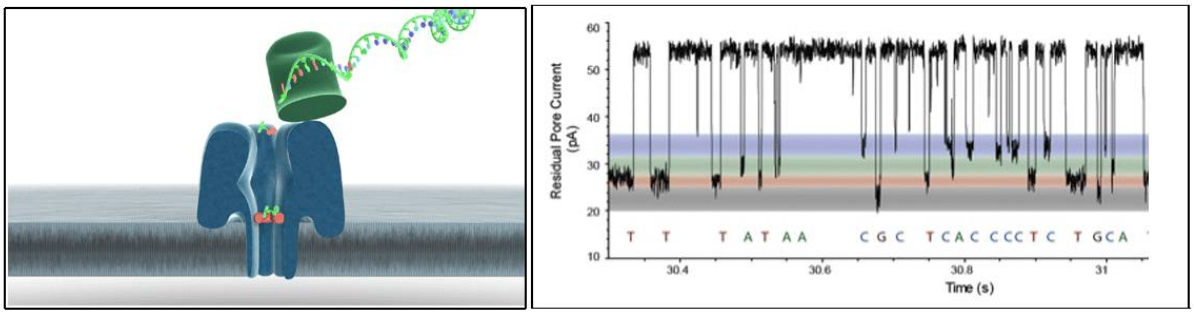

Abbildung 6: Darstellung des Nanopore-Systems; links : Schema einer Pore; rechts: Messkurve des Systems (Quelle links und rechts: [11], von Webseite)

2.2 AUDITORY DISPLAY UND SONIFIKATION

[...]it is straightforwardly unreasonable to assume that visualization channels alone will be able to accommodate such incredible amounts of information. The need to offload information to other perceptual channels is very apparent. [...] Since, under normal circumstances, humans perceive largely with both vision and audition, the auditory channel is a natural choice for this offloading. The simultaneous perception of visual and auditory information can have a synergistic effect on the brain's ability to comprehend an event. Furthermore, some types of information that are communicable through sound are not easily represented visually, and vice versa. ([14], p. 453)

Wie das Auge ist auch das menschliche Ohr ein hoch entwickeltes und potentes Sinnesorgan. Neben den drei Raumdimensionen empfängt und dekodiert unser auditives System Informationen über Dauer, Frequenz und Lautstärke teils einer Vielzahl verschiedener Schallereignisse zur gleichen Zeit. Der Mensch hat zudem die Möglichkeit, sich bewusst auf einzelne Ereignisse zu konzentrieren und andere weitgehend auszublenden. Man kann in einem Restaurant beispielsweise am Nebentisch lauschen, auch wenn am eigenen Tisch geredet wird.
Das menschliche Gehör ist also ein sehr hochentwickeltes Sinnesorgan und einer der Hauptkanäle der menschlichen Kommunikation. Das prominenteste Beispiel ist natürlich die gesprochene Sprache. Doch in unserem lassen sich zahllose andere Situationen finden, in denen Schallereignisse verschiedenster Art zur Übermittelung von Informationen verwendet werden.

Sonification is the use of nonspeech audio to convey information.
([1], p. 4 in Dokument von 2010).

Dies ist der erste Satz des NSF-Reports der ICAD[9] und eine oft zitierte Definition von Sonifikation. Der Begriff Sonifikation ist, zumindest meiner Alltagserfahrung nach, relativ wenigen Menschen vertraut. Das Konzept der Sonifikation ist jedoch, zumindest so man es einzig oben stehender Definition gemäß interpretiert, eigentlich eine Entwicklung der Evolution und schon seit langer Zeit Teil der Natur, bedienen sich doch zahlreiche Tierarten verschiedenartigster Schallereignisse, die von Artgenossen wahrgenommen und interpretiert werden. Sie kommunizieren via Audio, ohne sich jedoch einer Sprache nach menschlichem Vorbild zu bedienen. Auch der Mensch wird sich zunächst wohl mittels reiner Laute verständigt haben, ehe sich das darauf basierende, komplexe Sprachsystem herausbildete. Natürlich kann man den Begriff Sprache weiter fassen und jegliche der Kommunikation dienende Lautäußerung seitens eines Lebewesens als eine Art Sprache bezeichnen. Sie wäre dann per oben stehender Definition keine Sonifikationen mehr. Doch Kirchenglocken, die die volle Stunde signalisieren und zum Gottesdienst rufen, oder die mannigfaltigen Trompetensignale von Armeen sind Beispiele, die der oben genannten Definition genügen und schon lange Teil des menschlichen Lebens sind.
Das Konzept ist also schon sehr alt, der Terminus 'Sonifikation' ebenso wie das Bemühen um eine systematische Verwendung des Prinzips in der Wissenschaft auf Grundlage einer konsistenten Theorie hingegen sind noch vergleichsweise jung[10].

9 International Community for Auditory Display. Die erste Konferenz fand im Jahr 1992 statt.
10 Einen Überblick über geschichtlichen Entwicklung von ersten Anwendungen bis zur Mitte der 80er Jahre bietet Steven P. Frysinger in [15].

Dieses Kapitel nähert sich dem Thema Sonifikation von wissenschaftlicher und von künstlerischer Seite. Der erste Abschnitt beschäftigt sich mit den Definitionen, die auf dem Gebiet weder einheitlich noch eindeutig sind. Ausgehend vom Ergebnis dieser Diskussion wird zunächst ein Überblick über das Gebiet des Auditory Display insgesamt und anschließend über Sonifikation als Teilbereich desselben gegeben. Die drei in der Literatur am häufigsten erwähnten Sonifikationsprinzipien werden anhand von Beispielen vorgestellt. In der Folge werden die Bemühungen und Herausforderung hinsichtlich eines kohärenten, theoretischen Überbaus geschildert, bevor das Kapitel mit einer Betrachtung von Sonifikation und Kunst schließt.

2.2.1 Definitionen

Wie schon in der Einleitung zu diesem Kapitel erwähnt, ist die Definition von Sonifikation aus dem NSF-Report der ICAD von 1997 (vgl. [1]), die Sonifikation als Verwendung von nicht-sprachlichen Audiosignalen zur Informationsübermittelung beschreibt, die am weitesten Verbreitete. Dennoch ist sie meiner Ansicht nach nicht unproblematisch. Dafür gibt es mehrere Gründe. Erstens folgt auf die erste direkt eine weitere Definition:

> *[...], sonification is the transformation of data relations into perceived relations in an acoustic signal for the purposes of facilitating communication or interpretation. ([1], p .5)*

Diese ist inhaltlich viel spezifischer und enger gefasst, denn hier geht es nicht mehr um Information, sondern um Daten. Gemäß der ersten Definition wäre auch ein Handy-Klingeln eine Sonifikation, denn es ist ein Audiosignal, es ist in den meisten Fällen nicht-sprachlich und es transportiert die Information, dass jemand anruft. Nach der zweiten Definition ist es jedoch keine Sonifikation, denn es liegen keine Daten[11] zugrunde, deren inhärente Strukturen und Relationen auf akustische Weise durch das Klingeln nachgebildet werden sollen. Beide Definitionen sind also inhaltlich verschieden und umfassen mithin nicht dieselben Dinge. Sonifikation der ersten Art ist etwas anderes als Sonifikation der zweiten.
Neben dem Begriff Sonifikation findet man auch immer den Begriff 'Auditory Display'. Die Wissenschaftscommunity trägt den Term genauso in ihrem Namen wie der erste Konferenzbericht der ICAD von 1992, der als Buch mit dem Titel 'Auditory Display' ([16]) und nicht 'Sonification' erschienen ist. Im Buch 'Ecological Psychoacoustics' findet man im Kapitel 'Ecological Psychoacoustics and Auditory Display: Hearing, Grouping, Meaning Making' gleich zu Beginn dazu folgende Definition:

> *Auditory display is the use of non-speech sound to communicate information to a listener. ([17] p. 150)*

Das Kapitel stammt von Gregory Kramer und Bruce N. Walker, die auch beim NSF-Report als Autoren genannt werden. Diese Definition von 'Auditory Display' ist fast wortgleich mit der für Sonifikation aus dem NSF-Report. Daraus könnte man schließen, dass beide Begriffe synonym sind. In dem Kapitel heißt es jedoch weiter:

11 Auf eine genaue Definition von Daten und Informationen soll hier verzichtet werden. Das Beispiel sollte basierend auf dem umgangssprachlichen Verständnis der Begriffe nachvollziehbar sein.

2.2.1 Definitionen

The field of auditory display encompasses everything from alarms to complex sonifications. (ebd.)

Dies wiederum impliziert, dass Sonifikation eine Art eines Auditory Displays ist. Auch eine andere Veröffentlichung, an der Gregory Kramer als Autor beteiligt ist, bestätigt die Annahme, dass es sich um zwei verschiedene Dinge handelt.

A more specific type of auditory display technique called sonification involves "the transformation of data relations into perceived relations in an acoustic signal for the purposes of facilitating communication or interpretation" (Kramer et al., 1999, p. 3). ([18], p. 17)[12]

Die Definitionen sind also offensichtlich unscharf und mehrdeutig. Thomas Hermann weist in [19] ebenfalls auf diese Probleme hin und schlägt eine neue Definition vor, die Sonifikation anhand von vier Axiomen definiert, die eine Daten-Klang-Transformation erfüllen muss. Sonifikation ist bei Hermann ein umfassender Begriff für 'Verklanglichung', der mit Visualisierung auf einer Stufe steht. Auditory Display beschreibt er als Gesamtheit aus Abhörsituation, Wiedergabetechnik und Interaktionsmöglichkeiten. Jedoch ist auch nach Hermanns Definition ein Handy-Klingeln spätestens dann eine Sonifikation, wenn verschiedenen Anrufern unterschiedliche Klingeltöne zuordnet werden, denn auch die Wahl eines Klingeltons ist ein Mapping des Paramters 'Anrufer' auf ein bestimmtes 'Earcon' (s.u.). Eine solche Definition ist sehr breit, und der Kern von Sonifikation wird so meiner Ansicht nach nicht getroffen. Ein Windows-System-Sound und die Transformation eines multidimensionalen Datensatzes in einen Audiostrom wären begrifflich nicht voneinander getrennt.

Die meiner Ansicht nach gelungensten Definitionen stammen von Alberto de Campo. 'Auditory Display' beschreibt er, so wie es auch der Name der ICAD impliziert, als Überbegriff für das gesamte Feld:

Auditory Display is the rendering of data and/or information into sound designed for human listening. ([20], p. 1)

Diese Definition ist der allgemeinen ICAD-Definition sehr ähnlich, unterscheidet sich aber hinsichtlich des Zwecks, der hier nicht mehr explizit eine Übermittlung, sondern 'nur noch' die Umwandlung der Information oder Daten ist. Sie schließt also künstlerische Ansätze nicht von vornherein aus. Zudem ist Auditory Display nach de Campo nicht mehr auf nicht-sprachliche Audiosignale beschränkt, was auch Thomas Hermann in seiner Dissertation an der ICAD-Definition kritisiert hatte ([21], p. 23).
Den Überbegriff Auditory Display teilt de Campo in zwei Kategorien, nämlich 'Auditory Information Display' und 'Sonification'.

Auditory Information Display is the rendering of well-understood information into sound designed for communication to human beings. It includes speech messages such as in airports and train stations, auditory feedback sounds on computers, alarms and warning systems, process monitoring systems, etc. ([20], p. 1) [13]

12 Die innerhalb des Zitats gegebene Definition stammt aus ebenfalls aus [1] .

13 Hier muss angemerkt werden, dass der letzte Punkt, nämlich die Prozessüberwachung, nur bedingt in diese Kategorie gehört, da einige 'Auditory Displays' auf diesem Gebiet hinsichtlich des Inputs, der angewandten Technik und des Outputs klar den Sonifikationen zuzuordnen sind.

> *Sonification or Data Sonification is the rendering of (typically scientific) data into (typically non-speech) sound designed for human auditory perception. The informational value of the rendering is often unknown beforehand, particularly in data exploration.*
> *(ebd.)*

Auch diese Definition gibt als Zweck 'nur' die Wahrnehmung seitens des Menschen und nicht eine Analyse an, künstlerische Ansätze sind also weiterhin nicht ausgeschlossen.
Diese Definitionen auch nicht über alle Kritik erhaben[14], erscheinen mir jedoch am schlüssigsten und präzisesten von allen angeführten Beispielen. Im nächsten Abschnitt wird zunächst ein Überblick über das gesamte Feld des Auditory Display gegeben.

2.2.2 Auditory Display

Nach de Campo ist Auditory Display der Überbegriff für alle Formen und Facetten der Verklanglichung von Informationen. Auch wenn es im Vergleich mit Visualisierung und Grafikdesign wohl noch ein kleines Feld ist, so umfasst es doch eine Vielzahl teils sehr unterschiedlicher Aspekte.
Es gibt verschiedene Möglichkeiten, innerhalb des Gebietes Kategorien zu bilden, anhand derer sich ein akustisches Display klassifizieren lässt. Eine naheliegende Möglichkeit ist die Einteilung anhand des Zwecks. Sie bietet einen guten Überblick über die Vielseitigkeit des Gebiets. Die folgende Übersicht ist teilweise an die in [22] vorgeschlagene Einteilung angelehnt. Einige Punkte sind aus

- **Benachrichtigung / Alarm / Warnung**

  Beispiele für diese Gruppe sind hier Handy-Klingeltöne, Feuerwehrsirenen oder das Geräusch, mit dem in der U-Bahn das Schließen der Türen signalisiert wird.

- **Akustisches Feedback /Statusanzeige / Objektanzeige**

  Ein gutes Beispiel für eine Statusanzeige ist das Geräusch beim Schließen einer Autotür via Fernbedienung. Der Sound signalisiert, dass der Befehl angekommen ist und ausgeführt wurde. Weitere Beispiel findet man im Bereich der HCI (Human-Computer Interaction). Ein Fenster mit einer Fehlermeldung erscheint zum Beispiel häufig begleitet von einem Ton. Der Trend zur Verkleinerung der Displays wird den Anteil der Information, die ein Gerät akustisch transportiert, wohl mittelfristig erhöhen (vgl. [23]).

14 Auch diese Definition ist zu breit. Man kann beispielsweise einen Notentext als einen Datenbestand ansehen. Für einen Menschen, der des Notenlesens nicht mächtig ist, ist eine notierte Melodie wenig aussagekräftig. Bestenfalls kann er grobe Aussagen über die notierte Melodie machen, indem er sich an visuellen Clues orientiert. Ein Musiker gibt die notierten Daten mittels seines Instruments akustisch wieder. Gemäß der Definition handelt es sich also auch hier um eine Sonifikation, da eine 'Wissenschaftlichkeit' der Daten zwar als typisch nicht jedoch zwingend deklariert wird.

- **Monitoring von Prozessen**

  Ein Beispiel aus der Praxis ist die Überwachung des Herzschlags und anderer physiologischer Parameter während einer OP. Ein akustisches Display ist sinnvoll, wenn es gilt, mehrere Prozesse gleichzeitig zu überwachen oder wenn die Augen bereits 'beschäftigt' sind (vgl. [24], p. 24).

- **Unterstützung sehbehinderter Menschen**

  Hier reicht die Bandbreite der Anwendung von akustischen Computermenüs, über das akustische Äquivalent einer Grafik bis hin zur akustischen Brille, die es dem trainierten Nutzer erlaubt, visuelle Eigenschaften seiner Umwelt via Audio wahrzunehmen (ebd., p. 23).

- **Hilfsmittel zur Datenanalyse und Datenpräsentation**

  Das empfindliche Gehör des Menschen kann genutzt werden, um Daten auf inhärente Muster zu untersuchen. Ein Auditory Display kann dabei sowohl das einfache Gegenstück zu einer grafischen Darstellung als auch ein echtes Werkzeug zum interaktiven Data-Mining eines Datenbestands sein.

- **Kunst**

Man kann Auditory Displays auch anhand ihrer Form, ihres Inputs oder der zugrunde liegenden Transformation klassifizieren. Diese und weitere Möglichkeiten der Kategorienbildung diskutiert Stephen Barrass in seiner Dissertation [25] unter der Überschrift 'Approaches to Auditory Display'. Die meisten in der Literatur zu findenden Kategorien von Auditory Displays sind jedoch Mischformen mehrerer dieser Aspekte. E.D. Mynatt unterscheidet in [26] 'Auditory Icons', 'Earcons' und 'Sonification' als verschiedene 'Auditory Display Techniques'. Nees und Walker nehmen in [22] eine ähnliche aber erweiterte Einteilung in 'Common Auditory Display Classes' vor. Hier finden sich neben 'Earcons' und 'Audification' beispielsweise auch 'Auditory Graph', 'Alarm' und 'Audio in arts and entertainment'[15]. Thomas Hermann und Helge Ritter sprechen in [27] von 'Auditory Icons', 'Earcons', 'Audification' und 'Parameter Mapping' als *„Sonification approaches[...]mainly based on[...] four techniques“* ([27], p. 1).
Alle Einteilungen sind sich ähnlich, aber nicht identisch und besonders hinsichtlich des Begriffs 'Sonification' teils widersprüchlich[16]. Alle Autoren verwenden jedoch unter anderem die Begriffe 'Auditory Icon' und 'Earcon'. Da sie in der Literatur so häufig neben 'Sonification' und 'Parameter Mapping' genannt werden und spezifische, klar definierte Formen von Auditory Displays sind, sollen sie hier kurz erläutert werden.
'Auditory Icons' sind kurze Sounds mit ikonischem Verhältnis von Signifikat und Signifikant. Sie sind Ende der 80er Jahre aus der Forschung zur Mensch-Maschine-Interaktion entstanden

15 Hier zeigt sich deutlich, dass mehrere Aspekte zur Bildung von Kategorien herangezogen werden. Ein Alarm ist klar eine zweck-basierte Bezeichnung, während ein Auditory Icon durch seine Form definiert ist und durchaus Alarmfunktion haben kann.

16 Während Mynatt 'Sonification' als Kategorie neben 'Auditory Icons' und 'Earcons' stellt, bilden diese für Hermann und Ritter Untergruppen von 'Sonification'.

und fanden damals in Apples SonicFinder Anwendung ([28]). Ein Alltagsbeispiel für ein Auditory Icon ist der Windows-Systemsound für das Leeren des Papierkorbs – ein kurzes Geräusch, dass klingt, als werde Papier zerknüllt. Ein Auditory Icon wird im Idealfall nur anhand des Weltwissens des Hörers verstanden.
Earcons sind den Auditory Icons ähnlich. Auch sie sind in den 80er im Umkreis der HCI entstanden. Meera M. Blattner et al. bezeichnen Earcons in [29] als akustischen Gegenpart zu visuellen Icons, stellen aber klar, dass sie Icon in seiner weiter gefassten Bedeutung verwenden, also auch abstrakte Symbole als Icon bezeichnen. Ein Earcon ist ein kurzes akustisches Ereignis mit einem lexikalischen Bezug zu seiner Bedeutung, wodurch es sich vom Auditory Icon unterscheidet. Ein einfaches Beispiel ist der Windows-Vista-Sound zur Signalisierung eines niedrigen Akkustatus, ein kurzer, ein- und ausgeblendeter, glockenartiger Klang. Wenn man die Bedeutung nicht kennt, wird man die Information nicht verstehen. Earcons sind nicht auf einzelne Klänge beschränkt, sondern können auch kurze Melodien sein. Einzelne Earcons können außerdem zu kurzen 'Sätzen' kombiniert werden. Ein Beispiel für eine solche Kombination findet sich in [29] auf Seite 30. Vorteil von Earcons gegenüber Auditory Icons ist die viel größere Bandbreite an vermittelbaren Informationen., Nachteil ist der nötige Lernaufwand.

2.2.3 Sonifikation

Sonifikation nach der Definition von de Campo ist die Transformation von zumeist wissenschaftlichen Daten zu Sound zum Zweck der Wahrnehmung seitens des Menschen. Sie kann sowohl zur reinen Darstellung der Daten, also als akustischen Äquivalent zu einer Grafik, als auch zum Data Mining eingesetzt werden.
Sieht man von der Imitation von Alltagsgeräuschen ab, so werden in der Literatur im wesentlichen drei Techniken für die Daten-Klang-Transformation unterschieden: Audifikation, Parameter Mapping und Model Based Sonification (MBS). Auch de Campo erwähnt diese Einteilung als *„sonification approaches“ ([20], p. 1)*. Im Folgenden werden die drei verschiedenen Strategien vorgestellt.

Audifikation

Audifikation wird oft als die einfachste Form der Sonifikation bezeichnet. Jeder einzelne Datenpunkt wird als aktuelle Amplitude einer Schalldruckverlaufs interpretiert[17]. Theoretisch können alle Daten derart transformiert werden, für eine sinnvolle, praktisch umsetzbare Audifikation müssen sie jedoch einige Vorraussetzungen erfüllen. Zum ersten sollten die Daten um einen festen Wert oszillieren. Eine stetig steigende Entwicklung innerhalb der Daten würde nur zu Klickgeräuschen zu Beginn und zum Ende des Datensatzes führen. Liegt in den Daten sowohl eine Oszillation als auch ein stetiger Trend vor, so muss der oszillierenden Anteil für

17 Streng genommen ist Audifikation also ein Parameter Mapping von Datenwert auf Schwingungsamplitude in einer sehr feinen Auflösung. Sie wird jedoch in der Literatur immer als separater Punkt neben den beiden anderen geführt.

eine Audifikation isoliert werden. Ein Beispiel einer solchen Transformation beschreibt David Worrall in [30][18] anhand von Aktienkursen. Die Eingangsdaten sind die Tagesschlussstände des Australischen XAO-Indexes einer Periode von 22 Jahren. Um den generellen Anstieg des Indexes von 1000 auf zwischenzeitlich fast 7000 Punkten zu eliminieren, bedient er sich des linearen Net Returns, der die relative Veränderung zum Vortag beschreibt. Das Ergebnis ist ein Datensatz der um 0 herum verteilt ist.
Eine weitere Anforderung für eine sinnvolle Audifikation ist eine große Menge an Datenpunkten. Will man beispielsweise Frequenzkomponenten bis zu 4000 Hz hörbar machen, so braucht man aufgrund des Nyquist-Theorems 8000 Datenpunkte pro Sekunde. Zwei Sekunden Sound brauchen bei einer Maximalfrequenz von 8000 Hz schon 32.000 Datenpunkte, und diese Audifikation ist eher ein akustischer Fingerabdruck des Datenbestands[19] als eine Darstellung, die eine detaillierte Analyse hinsichtlich versteckter Pattern oder Periodizitäten erlaubt. Mittels Time-Stretching-Techniken kann ein Mangel an Eingangsdaten ein Stück weit kompensiert werden, doch selbst dann ist eine einigermaßen große Datenbasis notwendig.
Für eine nützliche Audifikation vorteilhaft, wenn auch nicht zwingend ist eine den Daten inhärente, ausgeprägte Periodizität. Rein stochastische Daten lassen sich zwar auch audifizieren, und Veränderungen im Rauschspektrum erlauben natürlich Rückschlüsse auf die Daten. Es dürfte jedoch schnell ermüdend für den Hörer werden, wenn er ausschließlich mit verschiedenem breitbandigen Rauschen konfrontiert ist.

Am sinnvollsten ist der Einsatz von Audifikation für Daten von Schwingungs- oder Wellenphänomenen, wie sie beispielsweise bei einem EEG anfallen. Dort werden Potentialschwankungen im Gehirn mittels auf dem Kopf verteilter Elektroden gemessen. Man betrachtet dabei verschiedene Frequenzbänder im Bereich zwischen 1 Hz und 30 Hz, die für verschiedene Gehirnzustände und Aktivierungslevel charakteristisch sind. Langzeit-EEGs können bis zu 36 Stunden dauern, während derer ununterbrochen gemessen wird. De Campo et al. schlagen in [31] Audifikation als Hilfsmittel vor, diese Langzeit-EEGs auszuwerten. Die Beschleunigung, die notwendig ist, um die EEG-Frequenzen in einen für den Menschen wahrnehmbaren Frequenzbereich zu transformieren, birgt zugleich den Vorteil, eine große Menge an Daten schnell sichten zu können. Der von den Autoren entwickelte 'EEG-Screener' bietet einen variablen Beschleunigungsfaktor zwischen 10 und 360. Die verschiedenen Frequenzbänder werden beibehalten und ihre Amplitude kann individuell eingestellt werden. In einem informellen Versuch mit Experten wurden vor allem mögliche Verbesserungen des Tools ermittelt. Laut der Experten ist jedoch ein sehr gründliches Training nötig, um die Audifikation effektiv verwenden zu können.
Ein anderes Feld, in dem Audifikation gut zur Datenanalyse eingesetzt werden kann, ist die Seismologie. Florian Dombois hat sich in mehreren Arbeiten intensiv mit dem Thema beschäftigt (vgl. [32],[33])[20]. Seismologische Daten sind Messungen von Körperwellen in der Erdkruste, die rund um die Uhr in vielen Stationen weltweit durchgeführt werden. Dombois konzentriert sich bei seiner Audifikation auf die z-Komponente der dreidimensionalen

18 Hier findet sich auch eine sehr ausführliche Übersicht über bisherige Anwendungen von Audifikation.
19 Das trifft auch auf Worralls Experiment zu. Die Wahl einer Abtastrate von 8000 Hz führte zu Sounds von etwa 0.7 Sekunden Länge. Obwohl laut Worrall eine leichte Amplitudenmodulation hörbar ist, sind die Sounds letztlich zu kurz, und er greift auf eine Frequenzmodulation zurück.
20 In seinen Beiträgen findet sich jedoch eine gute Übersicht über die bisherigen Arbeiten auf diesem Gebiet, die im Übrigen auch in vielen Publikationen zur Sonifikation angeführt werden.

Bewegung. Die auftretenden Frequenzen bewegen sich laut Dombois im Bereich von 0.3 mHz bis etwa 20 Hz[21]. Daher ist wie schon bei den EEG-Daten eine Transformation in den hörbaren Bereich notwendig. Das Eingangssignal umfasst 17 Oktaven, das menschliche Gehör überstreicht 'nur' etwa 10. Eine lineare Frequenztransformation wird also immer nur einen Teil des Eingangssignal hörbar abbilden. Neben der Frequenztransformation ist auch eine Dynamikanpassung nötig, da teilweise sehr hohe Amplituden auftreten können. Dombois hat mittels dieser Audifikation zahlreiche Daten verschiedener Beben und Messstationen analysiert und führt, auf Basis seiner eigenen Wahrnehmung, eine ganze Reihe von Features an, die sich anhand der Audifikation detektieren lassen. Beispiele dafür sind die Art der Plattengrenze, die Entfernung der Messstation zum Zentrum des Bebens und sogar die Messstation selbst. Quantitative Versuche wurden bisher noch nicht durchgeführt, um diese Ergebnisse zu verifizieren. Doch auch wenn sich diese Effekte letztlich nicht nachweisen ließen, bleibt Audifikation im Bereich der Seismologie ein vielversprechender Ansatz, denn sie gibt dem Menschen die Möglichkeit, sich schnell einen Überblick über die Messdaten zu verschaffen , und die Fähigkeit des menschlichen Gehörs, Signalanteile von Interesse selbst bei sehr geringem Rauschabstand noch 'durchzuhören', ist ein klarer Vorteil gegenüber einer visuellen Darstellung.

Parameter Mapping

Die Mehrzahl aller in der Literatur zu findenden Sonifikationen basieren in irgendeiner Form auf einem Parameter Mapping. Das Grundprinzip ist eine Zuordnung einer primären oder sekundären Datendimension[22] auf eine (psycho-)akustische Wahrnehmungsdimension[23]. Thomas Hermann et al. sprechen in dem Zusammenhang von direktem und indirektem Mapping ([35], p. 161).
Ein Mapping kann direkt auf eine auditiv wahrnehmbare Größe wie zum Beispiel die Tonhöhe erfolgen, es kann aber auch auf einen Steuerparameter eines Klangerzeugers abgebildet werden, der mehr als eine Wahrnehmungsdimension betrifft. Ein Beispiel dafür ist die Frequenz einer Amplitudenmodulation. Abhängig vom Wert wird der Mensch hier eine Änderung der Lautstärke, der Frequenzzusammensetzung oder der Klangfarbe wahrnehmen.
Mappings können für Daten aller Skalenniveaus erfolgen. Nominal wäre zum Beispiel ein Timbre-Mapping, denn Timbre lässt sich nicht als einfacher Zahlenwert ausdrücken. Vor allem Abbildungen auf Verhältnisgrößen wie Tonhöhe bergen große Herausforderungen und erfordern viel Sorgfalt und psychoakustisches Wissen, da weder die menschliche Wahrnehmung nicht linear noch frei von Wechselwirkungen der Dimensionen ist.
Parameter Mapping ist nicht an große Datenmengen gebunden und kann im Gegensatz zur Audifikation auch mehrdimensionale Datensätze in einer einzelnen Sonifikation darstellen, da auf verschiedene akustische Wahrnehmungsdimensionen gleichzeitig abgebildet werden kann.

21 Vereinzelt sind auch höhere Frequenzen möglich, die als tiefes Grollen direkt hörbar sind.

22 Sekundäre Datendimension meint hier jede aus den Rohdaten mittels statistischer Verfahren extrahierte Information wie zum Beispiel ein Mittelwert, eine Häufigkeitsverteilung oder eine Autokorrelation; primär sind die Daten selbst.

23 Eine häufig in der Literatur zu findende Abbildung dieser Audiodimensionen zusammen mit einer ersten Einstufung, für welche Art von Daten die Dimensionen jeweils geeignet sind, bietet John B. Krygier in [34].

Die akustischen Dimensionen beeinflussen sich jedoch hinsichtlich der Wahrnehmung gegenseitig teils sehr stark, weshalb der Zahl der gleichzeitig darbietbaren Dimensionen praktische Grenzen gesetzt sind[24].

Auf Parameter Mapping basierende Sonifikationen lassen sich grob in zwei Gruppen einteilen. Die erste Gruppe sind akustische Pendants zu bereits existierenden, visuellen Graphen wie Histogrammen oder Liniendiagrammen. Man nennt sie daher auch 'Auditory Graphs'. Eine einfache Form wäre die Abbildung der x-Achse auf die Zeit und der y-Achse auf die Tonhöhe. Im Jahr 2005 fand das 'First International Symposium on Auditory Graphs' statt, dass sich nur mit diesem Thema befasst. Die Beiträge zum Symposium finden sich gesammelt in den Proceedings der ICAD 2005 ([36]), eine gute Quellensammlung zu diesem Bereich bietet die School of Psychology des Georgia Institute of Technology ([37]).
Die zweite Gruppe von Mappings hat kein visuelles Vorbild, sondern sucht nach akustischen Darstellungen für Daten, die sich nicht adäquat oder zielführend visuell darstellen lassen. Ein sehr gutes Beispiel für eine derartige Sonifikation bieten Christian Dayé et al. in [38]. Aufgabenstellung war die Evaluation eines Toilettensystems für an Multipler Sklerose erkrankte Patienten. Das Toilettensystem bietet unter anderem einen in Höhe und Neigungswinkel mittels Fernbedienung verstellbaren Toilettensitz. Um zu untersuchen, wie die Nutzer mit der Toilette und den Bedienelementen zurechtkommen, ist man auf sensorisch erfasste Daten angewiesen, da sich eine Beobachtung mittels Kamera wegen des Schutzes der Privatsphäre verbietet. Die Sensoren erfassen neben den Knopfdrücken der Nutzer auch Höhe und Winkel des Sitzes sowie den Schließstatus der Tür. Eine grafische Darstellung aller Parameter ist grundsätzlich möglich, doch wenig übersichtlich (vgl. [38], p. 330). Die Autoren schlagen alternativ eine Sonifikation vor, die Parameter Mapping und Auditory Icons kombiniert.

> *In case the door is open, a colored noise similar to diffuse ambient outside noise plays which fades out when the door is closed. When the buttons of the remote control are pressed, height control and tilt control events (up or down) play looping glissandos, up or down, identified by different basic pitch, glissando rate, and timbre. The alarm is represented by a doorbell-like sound. For the flush button, we chose (for obvious reasons) a noise burst as the representing sound event. Current (static) tilt and height are both represented as soft background drones and turned louder when there are changes; then they recede again. ([38], p. 331)*

Der Vorgang des Toilettengangs lässt sich so akustisch nachvollziehen. Die Daten werden in perzeptiv gut getrennten Audioströmen dargeboten (Rauschen, Töne, Noise-Burst) und dürften sich ohne Probleme zeitgleich verfolgen lassen. Dieses Beispiel zeigt deutlich das Potential von Sonifikation, für manche Daten die vorteilhaftere Darstellung zu sein.
Das zweite Beispiel für ein Parameter Mapping, vorgestellt von Petr Janata und Edward Childs in [39], stammt aus dem Bereich des Börsenhandels. Ziel ist eine akustische Unterstützung des Traders bei der simultanen Überwachung verschiedener Kursverläufe, die bislang weitestgehend visuell erfolgt. Die Autoren betrachten vier Aspekte eines Kursverlaufs. Die relative Bewegung des Kurses wird mittels zweier Stakkato-Töne dargestellt. Je nach Kursbewegung ist der Zweite niedriger oder höher als der Erste. Der User hat die Möglichkeit,

24 Die maximal verträgliche Anzahl hängt von der Wahl der Dimensionen, von den Charakteristika der Input-Daten wie Wertebereich und Schwankung und in hohem Maße auch vom Design an sich ab.

mittels eines Thresholds die minimale Änderung einzustellen, die für ihn von Relevanz ist. Der zweite Aspekt ist die absolute Kursbewegung, also die Bewegung in Bezug auf einen vom User festzulegenden Referenzwert. Sie wird mittels dreier Töne sonifiziert. Der Erste, die Referenz, ist konstant. Der zweite Ton ist der Kurswert beim letzten Abspielen der Sonifikation und der dritte Ton schließlich ist der aktuelle Kurswert. Die Tonhöhen ergeben sich aus dem Kursunterschied zum Referenzkurs und dem vom User gewählten Verhältnis von Kursunterschied pro Halbton. Für den dritten Aspekt, die Annäherung an ein Kursziel, wird ein vierter Ton verwendet, der den drei eben beschriebenen nachfolgt. Je näher der Kurs seinem Ziel kommt, desto lauter und länger ist dieser vierte Ton. Seine Tonhöhe ergibt sich aus dem Kursunterschied zur Referenz. Der letzte Aspekt schließlich ist der 'Bid-ask stack'. Er wird als Tremolo aus zwei Noten dargestellt, deren Intervall sich aus dem höchsten Gebot und dem günstigsten Angebot ergibt. Die Länge des Tremolos hängt von der sogenannten 'show' ab, also von der Menge an Handelsanfragen, die für den Kurs momentan bestehen. Alle vier Aspekte lassen sich durch die Verwendung unterschiedlicher Instrumente und Register für mehrere Kurse zeitgleich darstellen.
Diese eigentlich sehr einfache Sonifikation ist zeitdiskret und weitestgehend direkt. Die Autoren erwähnen, dass ein Prototyp des Tools seit 2002 durchgehend von echten Händlern in Live-Situationen verwendet wird, und führen einiges an positivem Feedback dererseits an. Des Weiteren haben sie einen Versuch durchgeführt, bei dem der visuelle Kanal der Probanden durch eine Primäraufgabe gefordert war, während sie eine Sekundäraufgabe (die Überwachung von Kursverläufen) mit oder ohne akustische Unterstützung durchführen mussten. Die akustische Unterstützung erwies sich dabei als hilfreich.

Model Based Sonification

Die Model Based Sonification (MBS) wurde maßgeblich von Thomas Hermann an der Universität Bielefeld entwickelt ([21], [27]), und wird seither regelmäßig in der Literatur als eigenständige Technik zur Sonifikation genannt.
Die Eingangsdaten werden bei diesem Ansatz nicht als Steuerdaten für Parameter der Audiodomäne, sondern als physikalische oder chemische Kennwerte eines virtuellen Modells verwendet, mit dem der Benutzer mittels eines Interfaces interagieren kann, *„[...](in other words: the data is used to build an instrument or sound-capable object, while the playing is left to the user)“([19], p1)*. Die Modelle sind in einem Gleichgewichtszustand und produzieren von allein keinen Sound. Erst eine äußere Anregung führt zu einer Reaktion des Systems. Diese Reaktion ist bestimmt durch die Anregung sowie durch den Aufbau und die Kennwerte des System und somit letztlich abhängig von den zu untersuchenden Daten. Ziel der MBS ist es, einige der Nachteile des Parameter Mapping zu vermeiden und von der Fähigkeit des Menschen zu profitieren, aus der akustischen Antwort eines Gegenstands oder eines Systems sehr präzise Schlussfolgerungen auf dessen Struktur ziehen zu können. Außerdem verspricht ein virtuelles, physikalisches Modell intuitive Möglichkeiten der Interaktion. Die MBS ist in erster Linie für die explorative Datenanalyse multidimensionaler Datensätze gedacht.

Ein einfaches Beispiel eines für ein derartiges Modell ist ein Feder-Masse-System. Die Kennwerte sind die Masse, die Federsteife und ein Dämpfungsterm. Dieses Modell könnte also

genau einen Datenvektor mit drei Elementen repräsentieren. Ein solches Feder-Masse-System bildet die Grundlage für die von Hermann vorgeschlagenen 'Data Sonograms'. Jeder Datenpunkt ist durch einen Schwinger repräsentiert, der an einem Punkt P in einem n-dimensionalen Koordinatensystem entsprechend des Datenvektors 'befestigt' ist. Trifft eine virtuelle Schallwelle auf einen Datenpunkt, fängt dieser an zu schwingen und emittiert so seinerseits (virtuellen) Schall. Der User hat über ein 2D-Interface die Möglichkeit, sowohl den Ursprung der Anregungswelle als auch die Abhörposition zu bestimmen. So kann ein Datensatz beliebiger Größe im Hinblick auf seine örtliche Verteilung untersucht und etwa 'outlier' isoliert werden. Die Feder-Masse-Kennwerte können dabei für alle Punkte gleich sein oder von Elementen des jeweiligen Datenvektors bestimmt werden. Das Modell ist für Datensätze beliebiger Dimension verwendbar.
Das zweite Beispiel, das 'Particle Trajectories Sonification Modell', platziert zunächst für jeden Datenvektor eine Masse im n-dimensionalen euklidischen Raum, wobei n wieder der Anzahl der Dimensionen eines Datenvektors entspricht. Die Massen sind diesmal fest fixiert und über aber eine virtuelle, gravitationsartige Kraft aus, wodurch lokale Potentialtrichter entstehen. In diesen Raum werden nun seitens des Users virtuelle Testpartikel mit einer bestimmten Geschwindigkeit v geschossen. Sie bewegen sich durch den Raum und werden nach und nach von den Potentialtrichtern eingefangen, um dort schließlich in eine von der Tiefe des Trichters abhängende oszillierende Bewegung überzugehen[25]. Das Soundsignal des Modells ist die Summe der Signale der einzelnen Partikel und diese wiederum ergeben sich mittels Frequenz-Mapping aus der kinetischen Energie der Partikel zum Zeitpunkt t. Die oszillierenden Bewegungen innerhalb eines Potentialtrichters haben feste Frequenzanteile im Spektrum zur Folge. Die Potentiale der einzelnen Massen haben eine einstellbare *„interaction length“* ([21], p. 89), die bedingt, dass um Cluster von Punkten ein gemeinsamer lokaler Potentialtrichter entsteht, der natürlich größer ist als der eines einzelnen Datenpunktes. Zwei Cluster mit verschiedenen Gesamtmassen und somit unterschiedlich 'großen' Potentialtrichtern sind so als zwei unterschiedliche Frequenzenkomponenten hörbar. Die 'interaction length' erlaubt auch Rückschlüsse hinsichtlich der Dichte der Cluster. Für ein optimales Funktionieren dieses Modells sollten möglichst viele Partikel von verschiedenen Punkten in den Raum geschossen werden, denn sie müssen sich möglichst statistisch auf die Potentialtrichter verteilen, um den gesamten Datenbestand zu erfassen. Auch dieses Modell lässt sich für beliebige Datensätze anwenden. Es erlaubt eine Clustersuche ohne statistische Aufbereitung.

In seiner Dissertation stellt Hermann insgesamt sechs Modelle vor, die jeweils verschiedene Aspekte der Daten und des menschlichen Gehörs in den Mittelpunkt stellen ([21], p. 146). Er merkt auch an, dass lediglich die Sonifikation von 'Particle Trajectories' kein gleichmächtiges visuelles Pendant hat. Die MBS ist sicherlich ein interessanter Ansatz, der bei bestimmten Aspekten der explorativen Datenanalyse von Nutzen sein kann. Es gibt bisher jedoch kein Beispiel für einen praktischen Einsatz eines Modells. Das von Hermann angenommene intuitive Verständnis für die Modelle bzw. deren akustisches Feedback ist aus meiner Sicht nur teilweise Realität. Während die beiden hier vorgestellten Modelle hinsichtlich Konfiguration und Funktionsweise auch für den Laien leicht nachvollziehbar sind, ist für die Sonifikation von McMC (Markov chain Monte Carlo) oder die 'Growing Neural Gas Sonification' Hintergrundwissen oder eine ausführliche Einführung zum Verständnis notwendig.
Die Entwicklung einer modellbasierten Sonifikation ist verglichen mit einer Audifikation sehr aufwendig. Die Modelle selbst können sehr rechenintensiv sein und eine Echtzeit-Anwendung

25 Das wird durch die Wahl einer harmonischen Potentialfunktion erreicht.

unmöglich machen. Vorteilhaft ist die Tatsache, dass ein einmal entwickeltes Modell für viele verschiedene Datensätze verwendet werden kann.

2.2.4 Systematisierung

Until we understand more about what makes sonification successful, the field will remain mired in ad hoc trial-and-error design.. ([1] ,1997)

While Auditory Data Representation is still not a standard feature of commercial spreadsheet packages, the historical foundation suggests that such an outcome is possible. What is lacking is scientifically-informed guidance in the use of sounds for data representation. ([15], p. 3 , 2005)

Auditory graph design and implementation often has been subject to criticisms of arbitrary or a theoretical decision-making processes in both research and application. Despite increasing interest in auditory displays coupled with more than two decades of auditory graph research, no theoretical models of how a listener processes an auditory graph have been proposed. ([40], p. 1, 2007)

Diese drei Zitate sollen illustrieren, dass im Bereich Auditory Display noch einige Systematisierungs- und Grundlagenarbeit nötig ist, ehe ihm die erhoffte Aufmerksamkeit zuteil werden kann. Die Notwendigkeit einer konsistenten Theorie war offenbar schon früh erkannt, die Fortschritte dahingehend sind anscheinend aber eher klein, da sich die Zitate, obschon eine Dekade umspannend, im Kern nicht unterscheiden.
Es gibt im Wesentlichen zwei Gründe, die einen umfassenden theoretischen Überbau notwendig machen. Zum einen ist eine Theorie für eine effektive, nachvollziehbare und sachliche, also nicht von persönlicher Intention gesteuerte Entwickelung von Auditory Displays unabdingbar. Zwei Designer sollten, konfrontiert mit demselben Problem, anhand systematischer Analyse und Bearbeitung der Aufgabe mittels desselben systematischen Vorgehens vielleicht nicht notwendigerweise zu identischen, auf jeden Fall doch aber zu möglichst idealen, der Aufgabenstellung gerecht werdenden Lösungen gelangen können, ohne auf Versuch und Irrtum als Paradigma angewiesen zu sein.
Den zweiten Grund, der eine umfassende Theorie und eine Design-Systematik unverzichtbar macht, könnte man mit wissenschaftlicher Validität und Erwartungskongruenz überschreiben. Ein Auditory Display wird vom Entwickler mit bestimmten Abbildungseigenschaften geplant[26]. Im Bereich 'Auditory Information Display' (nach [20], vgl. 2.2.1) bzw. 'Auditory Graph' (vgl. [40]) geht es darum, dass ein bestimmter Dateninput X, durch das Display in wie auch immer gearteter Form dargeboten, beim Rezipienten zu einer vorhersagbaren Wahrnehmung Y führt. Der Designer eines Alarmsystems erstellt sein Display mit dem Ziel, dass der Hörer die Ausgabe des Systems im Moment einer Alarmsituation eben auch als Alarm interpretiert. Basiert das Design nun einzig darauf, was der Designer selbst als Alarm wahrnimmt, ist die Funktionalität für andere Benutzer nicht automatisch gegeben. Ein konkretes Beispiel für eine derartige Problematik stammt aus der Medizin. Hier wurden eine Reihe von Alarmsounds

26 Auch der Entwickler einer Sonifikation zum Data Mining wird zumeist eine vage Idee der Phänomene haben, die mittels des Displays gesucht werden sollen und wird dies beim Design berücksichtigen.

(kurze Melodien) in einen IEC-Standard für medizinisches Equipement integriert, ohne sie zuvor hinsichtlich ihrer Wahrnehmbarkeit und Unterscheidbarkeit zu untersuchen. Eine Evaluation dieser Sounds von Wissenschaftlern der Universität von Queensland ergab, dass überlappende Alarme nahezu nicht mehr zu unterscheiden, die Sounds also mithin für ihre Aufgabe nicht geeignet sind ([41]). Sollen akustische Monitorsysteme auch in hoch sicherheitskritischen Kontexten wie der Luftfahrt oder der Atomenergie zur Anwendung kommen, ist ein wissenschaftlich valides, evaluierbares Design unumgänglich.
Ähnlich hoch sind die Anforderungen im Bereich des 'Auditory Graph Designs'. Weder die Zuordnung von Daten- zu Audiodimension noch die Wahl der Polaritäten oder die Mappingfunktionen dürfen auf Intuition basieren oder mit Verweis auf 'Common Sense' erfolgen. Wie Walker und Kramer in [42] darlegen, können beispielsweise die 'intuitiven' Polaritäten von Sehenden und Blinden für dasselbe Mapping genau umgekehrt sein. Im Experiment war dies unter anderem für das Mapping-Paar 'Geldmenge – Frequenz' zu beobachten. Sehende Probanden assoziierten einen höheren Ton mit mehr Geld, Blinde hingegen mit weniger, was für einen sehenden Designer wohl kaum vorstellbar gewesen wäre. Der Grund für die Diskrepanz liegt, so Walker und Kramer, wohl in der andersartigen internen Interpretation des Gehörten. Sehende greifen vermutlich auf ein imaginiertes Liniendiagramm zurück, während die blinden Versuchsteilnehmer die Sonifikation als ikonisierte Alltagsgeräusche von Geld interpretierten. Das Klimpern auf den Tisch fallender Münzen steht für sie im Gegensatz zum dumpfen Geräusch, mit dem ein Bündel Scheine auf dem Tisch aufschlägt. Wie dieses Beispiel zeigt, müssen Mappingfunktionen, sollen sie eine korrekte Darstellung garantieren, immer mit der Zielgruppe eines Displays ermittelt und möglichst validiert werden.

Der Bedarf an Forschung ist also immens und erstreckt sich über viele Gebiete. Man kann mit Design-Methodologie, Psychoakustik und Werkzeugen drei Stoßrichtungen unterscheiden (vgl. [1], Kap. 3). Die Design-Methodologie gibt dem Designer einen festen Leitfaden, die Psychoakustik liefert die Wissensbasis für die Wahrnehmung und die Werkzeuge, Sonifikationsumgebungen oder sonifikationsaffine Programmiersprachen, vereinfachen den Prozess der Erstellung und machen so Sonifikation einer breiteren Zielgruppe praktisch zugänglich. Sie erlauben zudem auch Abstraktion von Problemklassen und die Erstellung von Bibliotheken mit Sonifikations-Templates und erhöhen idealerweise durch eine gemeinsame technische Basis den Austausch innerhalb der Sonifikations-Community.
In allen drei Feldern gibt es Forschungsbemühungen. Stephen Barrass, der sich bereits in seiner Dissertation mit einer Design-Systematik auseinandergesetzt hat (vgl. [25]), schlägt in [43] ein Vorgehen mittels Design-Pattern vor, ein Ansatz der sich in ähnlicher, umfassenderer Form auch in [44] findet. Alberto de Campo stellt im schon mehrfach referenzierten Konferenzbeitrag [20] seine Design-Space Map vor, die den Entwickler ausgehend von den Eigenschaften der Eingangsdaten bei der Auswahl der geeignetsten Sonifikationsmethode unterstützen soll. Keith V. Nesbitt versucht in [45] einen theoretischen Rahmen zu entwickeln, der für alle Modalitäten (sehen, hören, fühlen)[27] anwendbar ist und somit auch einen intermodalen Vergleich von Displays möglich macht. Eoin Brazil und Mikael Fernström schließlich argumentieren gegen die Abstraktion eines allgemeingültigen Designvorgehens.

27 Ein 'Odeur-tory Display' wird vorerst wohl nicht entwickelt werden.

2.2.4 Systematisierung

There is no single methodological framework that can deal adequately with the complex socio-cultural context of auditory display design in a coherent and non-reductionist manner. ([46], p. 1)

Ihrer Meinung nach sollten alle Entwicklungsstufen während des Designprozesses validiert werden, eine Forderung, die die Entwicklung natürlich um einiges aufwendiger macht.

Die Psychoakustik war bereits vor Aufkommen von Auditory Displays ein etabliertes Forschungsfeld, und Entwickler von Sonifikationen können auf einen beachtlichen Kanon an Wissen zurückgreifen. Jedoch ist die menschliche Hörwahrnehmung hochgradig non-linear und die Dimensionen beeinflussen sich gegenseitig teils sehr stark. Viele Fragestellungen, die im Zuge des Designs einer Sonifikation aufgekommen sind, waren bisher für Psychoakustiker nicht die Priorität. Wissenschaftler aus dem Bereich der Sonifikation unternehmen daher immer wieder Exkurse in die Wahrnehmungsforschung, um diese Fragen zu klären ([18], [47], [48]).

Hinsichtlich der Werkzeuge sollen hier zwei Projekte erwähnt werden. 'SonEnvir' ist ein gemeinsames Projekt von vier Universitäten der Stadt Graz (Österreich).

SonEnvir is intended to allow for very quick development of new sonification designs, based on a library of existing sound processes that have multiple perceptually independent control parameters. ([49], p. 1)

Am Anfang standen grundsätzliche Überlegungen hinsichtlich der Funktionalität des Programms. Anschließend wurden für Einsatzszenarien in verschiedenen Wissenschaftsfeldern funktionale Prototypen mit dem Ziel erstellt, langsam einen umfassenden Korpus an Sonifikationsansätzen zu erarbeiten[28].

Das zweite Projekt läuft unter dem Namen 'SoniPy' (vgl. [51], [52]). Die Autoren stellen zunächst fest, dass die Mehrzahl der bisherigen Ansätze in dieser Richtung auf einer Musik-Programmiersprache basierten, was ihrer Ansicht nach jedoch dazu geführt hat, dass die Möglichkeiten der statistischen Datenaufbereitung limitiert sind. Das SoniPy-Framework basiert daher auf der Programmiersprache Python. Das Ziel ist, eine flexible, modulare Plattform bereitzustellen, die es erlaubt, bereits bestehende Routinen aus anderen Bereichen wie zum Beispiel der Datenverarbeitung zu integrieren oder neue Module zu entwickeln und der Community zugänglich zu machen.

Neben der Forschung in den eben beschriebenen Teilfeldern besteht weiterhin die Notwendigkeit der Entwicklung einzelner Sonifikationen, denn 'erfolgreiche' Sonifikationen könnten beispielsweise hinsichtlich ihrer Funktionsweise analysiert werden und die Erstellung von Design-Patterns unterstützen.

A possible reason might have been that the community has not yet developed a sufficient number of proven-to-work prototypes for patterns to emerge from common practice. ([44], p. 2)

Desweiteren können erfolgreich eingesetzte Sonifikationen, auch wenn sie weder anhand eines standardisierten Designprozesses noch innerhalb eines festen Frameworks entwickelt worden sind, einen Beitrag zur Verbreitung von Auditory Displays leisten.

28 Alle Arbeiten sind auf der Webseite des Projekts ([50]) zu finden.

2.2.4 Systematisierung

Für die Entwicklung einer Sonifikation gibt es auch ohne eine feste Design-Guideline eine Reihe grundlegender Überlegungen und Fragestellungen, die in Betracht gezogen werden müssen. Nees und Walker bieten in [22] und [40] systematische Überblicke, die sich allerdings stark unterscheiden. Die vier wesentlichen Teilaspekte sind:

- Aufgaben/Eigenschaften des Displays[29]
- Eingangsdaten (Dimensionen, Wertebereich etc.)
- User (Vorerfahrung, Musikalität, Hörvermögen etc.)
- Nutzungssituation (Wo und wie hört man? Wie lange hört man? etc.)

Durch umfassende Betrachtung dieser Aspekte lassen sich eine Reihe von Anforderungen an ein Display formulieren, die dessen Form maßgeblich mitbestimmen werden. Ein solches analytisches Vorgehen ist bei der Entwicklung einer Sonifikation einem Vorgehen nach dem Prinzip 'Trial and Error' vorzuziehen.

2.2.5 Sonifikation und Kunst – Sonifikation als Kunst

Throughout the history of music, reproducing data by means of sound has played a role, without however being explicitly referred to as "sonification" ([53], p. 1)

Auditory Display im Allgemeinen und Sonifikation im Speziellen sind, wie in der Einleitung erwähnt, junge Forschungsgebiete. Das Konzept, Informationen mittels des Audiokanals zu übertragen, ist jedoch eine Entwicklung der Evolution und war somit von jeher Teil des menschlichen Lebens. Nicht verwunderlich ist es also, dass dieses Konzept auch in der Musik wiederzufinden ist. Ein Beispiel dafür ist die Affektenlehre, nach der verschiedenen Gefühle und Affekte durch verschiedene Tonarten oder andere kompositorische Mittel ausgedrückt werden. Der Inhalt von Liedtexten im Madrigal oder die Handlung einer Oper wurden so zum Teil in der Musik selbst 'lesbar' codiert. Mittels der oben eingeführten Terminologie ausgedrückt, handelt es sich um ein Parameter Mapping von nominalen Datendimensionen auf eine melodische Phrase oder einen Tonvorrat.
Ein weiteres Beispiel ist die maßgeblich durch Richard Wagner begründete Leitmotivtechnik. Ein Leitmotiv als kurze charakteristische, melodisch oder harmonisch geprägte Einheit ist fest und eindeutig mit einer Person, einem Gegenstand oder einem Gefühl verbunden. Leitmotive müssen zunächst gelernt werden, was natürlich zumeist automatisch vonstatten geht. Nach der Lernphase kann der Zuhörer bei Auftauchen eines Leitmotivs einen Teil der folgenden Handlung antizipieren, da er beispielsweise weiß, wer gleich auf die Bühne tritt. In der Fachsprache für Sonifikation ausgedrückt handelt es sich bei Leitmotiven um nichts anderes als um Earcons.
Zieht man die allgemeine ICAD-Definition[30] heran, so sind beide Beispiele zweifelsfrei Sonifikationen. Enger gefasste Definitionen von Sonifikation sprechen meist von Daten als Grundlage der Transformation. In der Musik speziell des letzten Jahrhunderts finden sich zahlreiche Werke mit diesem Paradigma.

29 Dieser Punkt kann beispielsweise die vage Forderung an ein Display beinhalten, möglichst wenig 'nervig' zu sein.

30 *„Sonification is the use of nonspeech audio to convey information.“ ([1], p. 4 in pdf von 2010)*

2.2.5 Sonifikation und Kunst – Sonifikation als Kunst

Starting in the 1950s, composers began to refer to extra-musical entities, particularly scientific data, on a different level. ([54], p. 3)

Volker Straebel trägt in [54] Beispiele aus dem Bereich der instrumentellen Musik zusammen, und ein Projekt der Universität der Künste in Bern ist dabei, alle derartigen Ansätze zu sammeln ([53], [55]). Ein Vergleich mit der von Frysinger beschriebenen Geschichte der wissenschaftlichen Sonifikation ([15]) zeigt, dass die Entwicklung in beiden Bereichen etwa im gleichen Zeitraum erfolgte, wobei der 'Durchbruch' auf wissenschaftlicher Seite, markiert durch die Gründung der ICAD, später als der in der Musik erfolgte.
Trotz dieser parallelen Entwicklung und der zahlreichen Kompositionen, die eindeutig Techniken der Sonifikation verwenden, ist die Frage, ob diesen Arbeiten das Label 'Sonifikation' verliehen werden sollte, bis dato nicht eindeutig beantwortet. An dieser Stelle sei zunächst auf die Überschrift dieses Abschnitts hingewiesen, in der bewusst von Kunst und nicht von Musik gesprochen wird. Es gibt aus meiner Sicht zwei Argumente dafür, dem allgemeineren 'Kunst' gegenüber 'Musik' den Vorzug zu geben. Zum Ersten gibt es im Deutschen den Begriff 'Klangkunst', der von Musik definitorisch einigermaßen klar abgegrenzt ist. Ein Teil der in [55] gesammelten Arbeiten ist diesem Genre zuzurechnen. Zum Zweiten ist der Begriff Musik meiner Ansicht nach im allgemeinen Verständnis weit weniger 'gedehnt' als der Begriff Kunst[31], birgt also eher die Gefahr, einer Arbeit aus ästhetischen Gründen verweigert zu werden als das Attribut 'Kunst'.
Die Frage ist also, ob künstlerische Arbeiten auch als Sonifikationen bezeichnet werden sollten, sofern sie auf derartigen Techniken beruhen.
Thomas Hermann spricht sich im Zusammenhang seines Versuchs einer Neudefinition (vgl. [19]) gegen eine Inklusion künstlerischer Arbeiten aus.

Same as some data visualizations may be 'viewed' as art, sonifications may be heard as 'music'[5], yet this use differs from the original intent. ([19], p. 3)

Hermann definiert vier Axiome, die eine Daten-Sound-Transformation erfüllen muss, um als Sonifikation bezeichnet werden zu können. Durch künstlerische Arbeiten sieht Hermann vor allem die Klarheit und Transparenz der Transformation verletzt.

If sonification-like techniques are employed to obtain a specific musical or acoustic effect without transparency between the used data and details of the sonification techniques, it might, for the sake of clarity, better be denoted as 'data-inspired music', or 'data-controlled music' than as sonification. ([19], p. 5)

Die Argumentation ist nachvollziehbar, führt jedoch konsequenterweise dazu, dass eine Klanginstallation, in der seismische Daten ähnlich der in 2.2.3 beschriebenen Methode audifiziert werden, genau dann eine Sonifikation ist, wenn mittels eines Begleithefts oder anderer Kanäle über die Methode der Audifikation informiert wird. Die gleiche Installation ohne derartige Information wäre hingegen 'data-controlled music'. Auch wenn Hermann also künstlerische Ansätze begrifflich von Sonifikation zu trennen sucht, ist dies in seiner Definition letztlich nicht konsequent verankert.
In der Tat ist der einzige Weg, künstlerischen Arbeiten das Label 'Sonifikation' sicher zu

31 Mir fällt es beispielsweise sehr schwer, einige Werke der seriellen Musik tatsächlich als 'Musik' zu bezeichnen ohne den nicht nur die zeitliche Herkunft anzeigenden Begriff 'zeitgenössisch' voranzustellen.

versagen, die Festlegung eines wissenschaftlichen Zwecks wie in der enger gefassten ICAD-Definition[32]. Für alle anderen möglichen Forderungen, die sich formulieren lassen, sind künstlerische Ansätze vorstellbar, die diese Forderungen erfüllen. Ist es nun wirklich sinnvoll, eine von zwei identischen Daten-Sound-Transformationen nur deshalb nicht als Sonifikation zu bezeichnen, weil es dem Ersteller nicht um wissenschaftliche, sondern vielleicht um emotionale oder metaphysische Erkenntnisse oder, etwas profaner, um die Befriedigung seiner Neugier ging? Und welches Label vergibt man, wenn ein streng lineares Mapping vorliegt, die Intention jedoch unbekannt ist?
Das Gebiet des Auditory Display hat insgesamt eine, teils vielleicht unbewusste Tendenz, die Kunst als Teil seiner selbst zu akzeptieren. Nees und Walker sprechen in ihrer Kategorisierung von *„Sonification as art“ ([22], p. 3)*. Volker Straebel arbeitet in [54] Parallelen zwischen der wissenschaftlichen Literatur zu Sonifikation und der Weltsicht der Romantik heraus und verweist unter anderem auf den Titel des Kapitels 'Listening to the earth sing' des Buches 'Auditory Display' ([56]) als Indiz für eine gedankliche Nähe auch wissenschaftlicher Arbeiten zur Kunst. Auch der ICAD-Wettbewerb 'Listening to the mind listening', wo es um Sonifikation von EEG-Daten ging, hat trotz der ausführlichen, durchaus wissenschaftlichen Dokumentationen eine klar künstlerische Komponente ([57]). Paul Vickers und Bennett Hogg erarbeiten in [58] einen *„The Ars Informatica–Ars Musica Æsthetic Perspective Space“ (ebd., p. 5)* und stellen fest „*The difference, then, between sonification and musical composition is largely one of perspective.“ (ebd. p. 6).* Auch Hotchkiss und Wampler sehen in ihrer Arbeit von 1991 keine klare Trennlinie zur Kunst.

> *The sound representations described above are music, even though they may initially sound much different to the listener than any music one has ever heard before. ([14], p. 7).*

Ein definitorischer Ausschluss künstlerischer Werke ist, wie oben erwähnt nur schwerlich möglich. Er entspräche weder der entwicklungsgeschichtlichen Realität, insofern als dass Sonifikation nicht ein wissenschaftliches Verfahren ist, das sich Künstler zueigen gemacht haben, noch würde er der in der Community vorhandenen Tendenz gerecht.
Wenn ein Werk sich von einer wissenschaftlichen Sonifikation nur dadurch abhebt, dass es eben nicht für die wissenschaftliche Arbeit gedacht ist, sollte es als Sonifikation bezeichnet werden können. Auch bei Werken, in denen neben einem Datensatz der Zufall das Ergebnis determiniert, ist die Bezeichnung Sonifikation durchaus noch angemessen, wird doch sozusagen der Zufall sonifiziert. Im Falle von Arbeiten, die neben der Sonifikation selbst kompositorische Anteile des Künstlers enthalten, könnte man von sonifikationsbasiert sprechen.

32 *[...], sonification is the transformation of data relations into perceived relations in an acoustic signal for the purposes of facilitating communication or interpretation. ([1], p .5)*

3. Sonifikation Genetischer Daten – Status Quo

In diesem Teil werden die bisherigen Ansätze von Sonifikation genetischer Daten untersucht. Zunächst wird ein Blick auf die wissenschaftliche Seite geworfen. Die bisher publizierten Ansätze werden vorgestellt und hinsichtlich eines möglichen praktischen Einsatzes bewertet.
Der zweite Abschnitt gibt einen Überblick darüber, wo und wie genetische Daten Grundlage für Kompositionen waren und arbeitet anschließend einige gemeinsame Eigenschaften von 'Gen-Musik' heraus.
Bezüglich dieser Aufteilung muss angemerkt werden, dass die vorgenommene Trennung zwischen Wissenschaft und Kunst in der 'Gen-Musik-Szene' sehr schwierig ist und offenbar mehrheitlich so gar nicht vorgenommen wird. In Quellensammlungen zum Thema (vgl. 3.2, p. 35) finden sich sowohl Ansätze, die in dieser Arbeit als wissenschaftlich klassifiziert werden, als auch rein künstlerische. Der Übersichtlichkeit des Kapitels halber wird hier trotzdem versucht, beide Gebiete voneinander zu trennen. Dazu werden zwei Kriterien herangezogen. Ein Ansatz wird genau dann als wissenschaftlich eingeordnet wenn:

- der Urheber selbst in erster Linie die Anwendung in der Wissenschaft als Einsatzgebiet seines Systems erwähnt und
- der Output des System nicht nach ästhetisch-musikalischen Gesichtspunkten manipuliert wird.

Musik, die mit Hilfe genetischer Algorithmen entstanden ist, bleibt bei den Betrachtungen ausgeklammert, da genetische Algorithmen zwar Prinzipien der Evolution modellieren, jedoch keine genetischen Sequenzdaten verwenden. Eine gute Einführung zu diesem Thema bietet [59].

3.1 In der Wissenschaft

In der Wissenschaft hat die Sonifikation genetischer Daten bislang keine große Verbreitung gefunden[33].
Die erste Publikation zu dem Thema stammt aus dem Jahr 1984 von Nobuo Munakata und Kenshi Hayashi. Munakata beschäftigt sich bis heute mit dem Thema und bietet auf seiner Webseite [61] eine Einführung in das Gebiet und zahlreiche Soundbeispiele. Der verwendete Ansatz zur Sonifikation, ein Pitch-Mapping im tonalen Bereich via MIDI, das jeder Base und/oder Aminosäure einen Ton zuordnet, ist so etwas wie eine Blaupause für den Großteil aller realisierten DNA-Sonifikationen, seien sie nun wissenschaftlich oder künstlerisch. Bei

33 Garcia-Ruiz und Gutierrez-Pulido geben in [60] eine Übersicht über Applikationen von Auditory Display für chemischer Daten, in der auch einige der hier genannten Beispiele aufgeführt sind.

Hayashi und Munakata bildet eine Base die Tonika, und die anderen werden auf Intervalle wie Quinte und Sexte abgebildet. Ohno und Ohno verwenden 1986 einen ähnlichen Ansatz ([62]), sie ordnen jedoch jeder Base jeweils zwei benachbarte Tonhöhen zu und erzeugen so eine komplette Tonleiter. Die Motivation der Autoren war weniger eine nützliche Form der Darstellung sondern ein Beleg für die These, dass Gene und Musik auf ähnlichen organisatorischen Prinzipien beruhen.

Ross D. King und Colin G. Angus gehen für ihre 'Protein Music' ([63]) einen Schritt weiter. Zur Oberstimme, die sich aus der Basenfolge ergibt, kommt eine auf den Aminosäuren basierende Bassstimme. Auch dafür wird ein Pitch-Mapping verwendet[34], jedoch wird nicht jeder einzelnen Aminosäure eine Tonhöhe, oder besser ein Intervall zur Tonika zugeordnet. Das wäre bei 22 verschiedenen Aminosäuren zwar möglich, ergäbe aber einen für eine Basslinie eher atypisch Tonvorrat. King und Angus wählen daher sieben physikalisch-chemische Eigenschaften der Aminosäuren, denen sie je ein Intervall zugeordnen. Die Aminosäuren haben dabei oft nicht nur eine dieser Eigenschaften sondern meist mehrere. Daraus leiten King und Angus ab, wie viele Bassnoten pro Triplet gespielt werden und welche. Die Autoren sagen über ihr gewähltes Mapping :

> *[...] timing and ordering of notes for each residue was found to allow properties and residues to be identified as well as sounding harmonious[...] ([63] p. 2)*

Belege für die Behauptung fehlen genauso wie eine Beschreibung eines konkreten Anwendungsfeldes in der wissenschaftlichen Arbeit. Die Autoren weisen darauf hin, dass sie ihren Ansatz bei der Analyse einiger Gene verwendet haben, ohne jedoch zu spezifizieren, wie genau und mit welchem Erfolg. Außerdem merken sie an, dass sich ein Song, der basierend auf einem Serotoninrezeptor mittels ihres Algorithmus erstellt wurde, mehr als 100.000 mal weltweit verkauft hat. Obwohl also vor einem wissenschaftlichen Hintergrund entstanden, hat 'Protein Music' wohl doch eher in der Kunst ihre Bestimmung gefunden.
Ein für eine konkrete Aufgabe im Kontext der DNA-Sequenz-Analyse erstelltes Tool stammt von Sook Young Won ([64]). Ziel seiner in Lisp programmierten Applikation ist eine akustische Unterstützung bei der Suche nach CpG-Inseln. Das sind Abschnitte im Genom mit einer hohen Dichte von Di-Nukleotiden aus Cytosin und Guanin. Diese Bereiche sind unter Anderem deshalb von besonderem Interesse, da sie ein Indikator für codierendes Material sind. Won verwendet ein sehr einfaches Pitch-Mapping. Die Basen A und T werden durch einen tiefen, G und C durch einen hohen Ton im Intervall einer Dodezime repräsentiert. Die Tondauer für eine einzelne Base beträgt 50 Ms., für Cluster aus C oder G wird die Tondauer verdoppelt. CpG-Inseln sind so als *„[...]groups of high-pitched notes with long durations,[...]“ ([64], p. 2)* hörbar. Neben Tonhöhe und Tondauer trägt auch das Timbre Informationen, das in Abhängigkeit vom Grad der Gesichertheit des sonifizierten Sequenzabschnitts mittels der Wahl verschiedener MIDI-Instrumente verändert wird. Dieser Ansatz hat im Vergleich zu den drei vorhergehenden ein klares Anwednungsszenario. Es ist jedoch eher unwahrscheinlich, dass Wissenschaftler in der Praxis ein Chromosom manuell nach CpG-Inseln durchsuchen, da diese Arbeit vom Computer deutlich schneller ausgeführt werden kann als vom Menschen[35].

34 Interessant ist anzumerken, dass die Autoren von 'Audification' sprechen, was weder dem heutigen noch dem damaligen Definitionsstand entspricht.

35 Wollte man etwa das menschliche Chromosom 21 mit seinen rund 34 Millionen Basen bei einer Rate von 20 Basen pro Sekunde einmal komplett durchhören, bräuchte man fast 20 Tage.

Trotzdem ist diese Sonifikation als Unterstützung oder Ersatz einer visuellen Darstellung der Ergebnisse CpG-Analyse durchaus in der Praxis vorstellbar.
Charlie Cullen und Eugene Coyle haben ebenfalls ein Tool erstellt, das Basen- und Aminosäuresequenzen mittels Pitch-Mapping sonifiziert ([65]). Die Basenfolge ist bei ihnen die Grundlage für den Bass, und den Aminosäuren werden ganze Akkorde anstatt einzelner Töne zugeordnet. Diese Zuordnung ist flexibel und kann nach Belieben vom User vorgenommen werden. Die Autoren schreiben über ihren Ansatz:

> *[...], it is proposed that this research uses the sonified output of DNA/RNA sequences for pattern matching and analysis purposes rather than as frameworks for composition.*
> *([65],* p. 2).

Eine konkretere Beschreibung, wie genau diese Arbeiten mittels der Sonifikation durchgeführt werden könnten, bleiben sie jedoch schuldig. In ihrem Fokus steht laut eigener Aussage vor allem der Rhythmus, den sie in Sonifikationen allgemein als unterrepräsentiert ansehen. Doch bis auf die Wahl zwischen einem 3/4-Takt, der ob der Triplettstruktur des genetischen Codes ohnehin inhärent ist, und einem 4/4-Takt mit einem stummen letzten Viertel findet sich wenig dazu im Text. Die Autoren bescheinigen dem 4/4-Takt *„[...] to be an even better way of defining codon base groupings[...]“* (ebd.); belegt wird das jedoch nicht, und es ist eher zweifelhaft, ob die Darstellung einer DNA-Sequenz als Melodie für Pattern-Matching-Aufgaben wirklich Vorteile bringt. Es wäre zu untersuchen, inwieweit der Mensch in der Lage ist, zwei quasi-zufällige Melodien aus etwa 20 bis 100 Noten zu unterscheiden und aufeinander zu beziehen.
Aus dem Jahr 2007 stammt ein Ansatz von Rie Takahashi und Jeffrey H. Miller. Der Titel ihres Beitrags, 'Conversion of amino-acid sequence in proteins to classical music: search for auditory patterns' (vgl. [66]), ist ein gutes Beispiel für die im Bereich Sonifikation genetischer Daten schwierige Trennung der Sphären Wissenschaft und Kunst. Auch wenn sich im Text Passagen finden, die eine Einordnung der Abreit als künstlerisch motiviert rechtfertigen würden, so gab die Phrase 'search for auditory patterns' des Titels schließlich den Ausschlag für das Attribut 'wissenschaftlich'. Zum Zweck der Arbeit äußern sich die Autoren folgendermaßen:

> *By converting genomic sequences into music, we hope to achieve several goals, which include investigating sequences by the vision impaired. Another aim is to attract young people into moleculargenetics by using the multidisciplinary approach of fusing music and science. ([67], p. 3)*

Weitere Beispiele geben sie jedoch nicht. Sie streben eine Sonifikation an, die einerseits *„[...] reasonable to a musician's ear [...]“ (ebd., p. 2)* klingt und gleichzeitig *„[...] faithful to the science of the protein sequences.[...]“ (ebd.)* bleibt. Zweiteres ist immer erfüllt, wenn weder Zufall noch Manipulation seitens des Menschen das Ergebnis des Mappings beeinflussen. Ersteres, also eine 'musikalische Qualität', kann bei einem einfachen Mapping aller Aminosäuren auf verschiedene Tonhöhen aufgrund der sich ergebenden Tonsprünge von bis zu 20 Halbtönen zumindest nach abendländischen Maßstäben nicht erreicht werden. Um genau dieses Problem zu umgehen, weisen Takahashi und Miller jeder Aminosäuren einen Akkord zu. Ähnliche Aminosäuren[36] teilen sich dabei denselben Grundton und unterscheiden sich nur durch in der Umkehrung des Akkords. Das Ergebnis ist ein reduzierter Tonumfang von 13 Tönen und,

36 Worauf die Ähnlichkeit im Detail beruht, wird nicht erklärt.

aufgrund der mittels der Akkorde ins Ergebnis gebrachten Harmonien, eine erhöhte Musikalität im Vergleich zu einer einstimmigen Melodie. Die Notendauer der Akkorde (von Ganzton bis Achtelnote) ergibt sich aus der relativen Häufigkeit der jeweiligen Aminosäure im Humangenom.
Die von den Autoren erwähnte Eignung ihrer Sonifikation als Hilfsmittel für sehbehinderte Menschen ist, wie im Übrigen bei allen bisher vorgestellten Ansätzen, natürlich vorstellbar, doch fehlt auch hier eine wissenschaftliche Evaluation. Es wurde auch keine Analyse der Aufgaben unternommen, bei denen das Program einem sehbehinderten Menschen assistieren soll. Die Sonifikation basiert also nicht auf ihrem späteren Einsatzszenario und geht nicht auf dementsprechende ergonomische Erfordernisse oder wahrnehmungstheoretische Grundlagen ein. Der spielerischer Zugang zum Thema Genetik für Kinder oder die ebenfalls im Paper erwähnte Grundlage zur musikalischen Improvisation sind als Einsatzgebiet der Methode wohl realistischer.
Ein klar definiertes Einsatzszenario, nämlich die Steuerung von Proteinsyntheseprozessen in Pflanzen und auch Tieren mittels auf Quantenoszillation beruhender Musik, hat der Ansatz von Joel Sternheimer. Die Quellenlage hierzu ist dürftig und wissenschaftlich kaum belastbar. Auf einer japanischen Webseite [69] findet sich eine Sammlung von Artikeln und Forschungsberichten, die mit dem Konzept zusammenhängen[37]. Die Beiträge in Englisch sind, so ist es auf der Webseite angegenen, alles Artikel japanischer Zeitungen, wobei jedoch keine Autoren kenntlich gemacht sind. Sucht man eine wissenschaftliche Darstellung der in diesen Artikeln erwähnten Experimente, muss man auf die französische Sektion von [69] zurückgreifen. Dort finden sich dann Autoren, Messwerte und grafische Darstellungen von Ergebnissen. Die Lektüre insbesondere von 'Musical Tomatoes', 'Musical Bred' und 'Music and „Musical Medicine“', lohnt sich, da in knapper Form unglaublich scheinende Ergebnisse berichtet werden. Zwei Beispiele:

> *Les tomates "musicales" étaient (tableau ci-dessus) environ deux fois plus nombreuses que les tomates témoin, et même trois fois plus nombreuses pour les 10 pieds les plus proches de la source sonore. Elles étaient de plus, en poids, en moyenne 2,5 fois plus grosses.*
> *([70], von Webseite)*[38]

> *Cheese made from the milk was, according to a testing panel, very delicious. The cheese made of the milk was also sold in a shop in Paris and the sales became 6 times more than usual during 2 weeks when the experiment has been carried out.*[39]
> *([71], von Webseite)*

Joel Sternheimers Methode fußt auf der These, dass bei der Synthese von Proteinen, genauer beim Andocken einer Aminosäure an die wachsende Peptidkette, durch Quantenoszillation für die jeweilige Aminosäure spezifische Frequenzen erzeugt werden. Da die Aminosäuren eine nach der anderen an das wachsende Makromolekül angehängt werden, bilden auch die emittierten Frequenzen eine Sequenz, eine sogenannte 'Protéodie'. Die Frequenzen lassen sich messen und in den hörbaren Bereich transformieren. So erhält man für jedes Protein eine

37 Diese Texte findet man zusammen mit einer US-Amerikanischen Patentschrift und einiger abgedruckter themenbezogener E-Mail-Korrespondenz von Sternheimer auch auf [68].

38 *Die 'musikalischen' Tomaten waren etwa zweimal so zahlreich wie die Tomaten der Vergleichsgruppe. Im Bereich von 10 Fuß um die Schallquelle herum waren es sogar dreimal mehr. Außerdem waren sie im Durchschnitt 2,5 mal schwerer. (Übersetzung des Zitats durch den Autor dieser Arbeit)*

39 Die referenzierte Milch stammt von Kühen, die mit Sternheimers Gen-Musik beschallt wurden.

individuelle Tonfolge. 'Hört' man nun diese Tonfolge (Hören im Sinne von, befindet man sich ihren Schallfeld), so kommt es laut Sternheimer zu Resonanzphänomenen, die die Produktion und Funktion des jeweils zu hörenden Proteins beeinflussen können[40]. Die Resultate, so wie sie in den oben angeführten Artikeln und Berichten beschrieben werden, wären in der Tat bahnbrechend. Bei kritischer Lektüre des Experiments zum Wachstum von Tomaten könnte vielleicht irritieren, dass die Pflanzen der Kontrollgruppe und der Versuchsgruppe unterschiedlich viel Wasser erhalten haben, die Ergebnisse also auch darauf zurückzuführen seien könnten. Insgesamt wirkt der Bericht dem Augenschein nach aber wissenschaftlich. Gleiches gilt für den Bericht über Brot ([73]). In den dort vorgestellten Experimenten wurde ein am Gärprozess beteiligtes Protein mittels Sternheimers Sonifikation gesteuert. Das behandelte Brot war laut Bericht größer und schmackhafter als das Kontrollbrot. Es stellt sich also die Frage, warum es so wenige Informationen zu Sternheimer und seiner Theorie gibt, und warum sie bis heute anscheinend weitgehend unbekannt ist[41]. Einen Hinweis liefert ein Essay von Jacques Lavau ([74]). Er analysiert darin eingehend einen Bericht Sternheimers zum Thema Quantenoszillation und bezeichnet ihn als inhaltsleer und wissenschaftlich nicht haltbar. Er bezichtigt Sternheimer des Betrugs und wirft ihm vor, mittels clever erfundener Forschungsergebnisse öffentliche Gelder zu akquirieren. Eine qualifizierte Bewertung der Theorie Sternheimers ist mir angesichts unzulänglichen Wissens im Bereich der Quantenphysik nicht möglich. Daher will ich die angegebenen Quellen pro und contra Sternheimer weitgehend unkommentiert nebeneinander stehen lassen. Insgesamt sind die wirklich erstaunlichen Ergebnisse im Zusammenspiel mit dem nach 16 Jahren immer noch extrem niedrigen Verbreitungsgrad der Theorie Gründe für berechtigte Zweifel.

Alle bisherigen Beispiele hatten die Sequenz von Basen oder Aminosäuren im Fokus. Zwei weitere Ansätze befassen sich mit der Faltungsstruktur der von Makromolekülen, die sich natürlich aus letztendlich aus der Sequenz ergibt. 'PROMUSE', vorgestellt in [75], ist ein Tool für das sogenannte 'Structure Alignment' von Proteinen, bei dem verschiedene Proteine anhand ihrer räumlichen Struktur verglichen werden. Dazu werden meist 3D-Renderings visuell übereinandergelegt. 'PROMUSE' erweitert das visuelle Display des Alignment-Tools 'RasMol' um die auditive Dimension und gibt insgesamt vier Eigenschaften eines Alignments in je drei Ausprägungen akustisch wieder. Die hier verwendete Technik könnte man als parametrisierte Earcons bezeichnen. Jeder Eigenschaft ist ein Bestandteil eines Jazz-Quartetts zugeordnet[42]. Die drei möglichen Ausprägungen werden entweder durch verschiedene Spielweisen oder verschiedene Instrumente repräsentiert. Die Polarität eines Proteins bestimmt beispielsweise die Akkordbegleitung, und die drei Ausprägungen, hier im einzelnen 'hoch', 'mittel' und 'niedrig', geben den Ausschlag für die Wahl von E-Gitarre, Marimba oder E-Piano. Die so ermittelten Instrumente spielen kurze vorher festgelegte Phrasen, die Inputdaten steuern also nicht, was gespielt wird, sondern lediglich wie. Die Phrasen sind so komponiert, dass sie alle harmonisch und rhythmisch zueinander passen. Das Ergebnis jedweder Kombination klint wie ein kurzer Ausschnitt eines Jazz-Songs. Auf dieser Grundlage haben die Autoren ein Experiment durchgeführt. Alignments zweier Proteine mussten anhand der vier gewählten

40 Mit dieser Theorie erklärt er auch das umstrittene Phänomen, das Milchkühe mehr Milch produzieren, wenn sie Mozart hören. Ein Teil der Melodie eines an der Milchproduktion beteiligten Proteins hat nämlich Ähnlichkeit mit der Melodie eines Mozartwerkes.

41 Die Theorie ist keineswegs komplett verschwunden. Ein Artikel in 'Sud Ouest' aus dem Jahre 2010 berichtet von einer erfolgreichen Bekämpfung der Weinrebenkrankheit 'Esca' mittels Sternheimers Methode [72].

42 Die Bestandteile sind Bass, Begleitung, Drums und Melodie

Eigenschaften beurteilt werden, und das Abschneiden der Versuchspersonen wurde für rein visuelle, rein akustische und audio-visuelle Darbietung verglichen. Die Ergebnisse zeigen einerseits einen erhöhten Anteil richtiger Klassifikationen bei Darstellungen mit akustischer Komponente im Vergleich zu rein visuellen und andererseits wenig Unterschiede zwischen bimodaler und rein akustischer Präsentation. Die Sonifikation hilft dem Menschen also offensichtlich bei der gestellten Aufgabe. Doch ist dies wirklich ein Beleg für eine sinnvolle Applikation von Sonifikation?

Eine mögliche Erklärung für das Ergebnis des Experiments ist die Tatsache, dass die akustische Komponente lediglich 3x4 also 12 verschiedenen Earcons respektive Vokabeln verwendet, von denen nie mehr als vier zur gleichen Zeit dargeboten werden, und deren Darbietung in einer segregationsfreundlichen Art und Weise geschieht. Die Autoren verwenden unter Anderem Panning und Lautstärke um die Stimmen zu trennen (vgl. [75], p. 5). Die akustische Darstellung liefert also schon fertige Kategorien, während die visuelle Komponente eine Analyse und Klassifizierung des Gesehenen erfordert. Es bestanden also ungleiche Vorraussetzungen für beide Sinne. Gleiche Vorraussetzungen wären zum Beispiel dann gegeben, wenn visuell auch eine Kombination aus vier Icons mit je einer von drei Fraben gezeigt würde. Das Ergebnis ist meiner Meinung nach also wenig aussagekräftig. Desweiteren stellt sich die Frage, wozu ein menschlicher Reviewer überhaupt nötig ist, wenn die Input-Daten bereits kategorisiert und klassifiziert sind. Genau das ist jedoch die Vorraussetzung für ein nominales Mapping der obigen Form, denn nur so kann das Programm den passenden Clip abspielen. Obwohl hier also im Gegensatz zu allen anderen bisher vorgestellten Ansätzen eine Evaluation der Methode versucht wurde, denke ich, dass der Ansatz zumindest so wie er beschrieben wurde lediglich im Bereich der Darstellung von Proteinstrukturen und Alignments für sehbehinderte Menschen einen Nutzwert hat.

Eine weitere Methode aus dem Jahr 2010 verwendet Sonifikation im Bereich von Sekundärstrukturen von RNA-Molekülen. Die relativ instabile, einzelsträngige RNA kann sich auf viele verschiedene Arten auf sich selbst zurückfalten und so stabilisieren. Florian Grond et al. stellen in [76] eine Methode vor, diese Sekundärstrukturen zu sonifizieren. Obwohl es natürlich Selektionsalgorithmen gibt, muss ein großer Teil von RNA-Sekundärstruktur-Datenbanken vom Wissenschaftler selbst durchsucht werden. Die Sonifikation soll die visuelle Darstellung nicht ersetzen, sondern erweitern und bei die manuellen Suche so unterstützen. Die Eingangsdaten, die Länge der RNA und die Klammernotation der jeweiligen Faltung, werden auf die Frequenz und die Zeit abgebildet. Der Output ist ein Obertonspektrum, dessen Peaks die Klammerfolge der Abstraktionsebenen und somit auch die Sekundärstruktur der RNA wiedergeben[43]. Jede Faltungsvariante hat ihren eigenen Sound, und ähnliche Strukturen erzeugen ähnliche Sounds. Die Ähnlichkeiten sind dabei visuell nicht immer sofort offensichtlich. Die akustische Darstellung ist hier klarer.

Von allen Ansätzen ist dieser wohl der wissenschaftlich fundierteste, da die Autoren einen sehr konkreten und auch plausiblen Anwendungsfall benennen. Desweiteren haben sie zunächst einen Anforderungskatalog an ihre Sonifikation erarbeitet und diese dann entlang der so ermittelten Randbedingungen entwickelt (vgl. [76] p. 2-3). Eine Evaluation, inwieweit die Sonifikation ihr Ziel erfüllt, wurde von den Autoren in Aussicht gestellt.

43 Eine präzisere Darstellung des Mappings ist nur möglich, wenn man mit der Klammernotation vertraut ist. Das würde den Rahmen des Abschnitts sprengen.

Insgesamt ist die Sonifikation genetischer Daten weder in der Praxis verbreitet noch ist sie wissenschaftlich fundiert bearbeitet worden. Viele der Ansätze gleichen sich in ihrem Grundprinzip des Pitch-Mappings und darin, dass sie eher anhand von Möglichkeiten als von Erfordernissen entwickelt wurden. Die Funktion folgt der Form. Auf die Wahrnehmung der Sonifikationen seitens des Menschen wurde so gut wie gar nicht eingegangen. Die Frage, ob eine Darbietung einer Sequenz als Melodie wirklich eine gut verständliche Form für den Menschen ist, wurde in keiner der Veröffentlichungen diskutiert, sondern aufgrund der Nähe zur Musik als gegeben angenommen. Einige der vorgestellten Ansätze lassen sich im Bildungsbereich sicher gut verwenden, da sie einen neuen, multimodalen Zugang zu Sequenzdaten eröffnen. In der praktischen Arbeit im Gebiet jedoch ist ihr Einsatz eher unwahrscheinlich.

3.2 IN DER KUNST

Wie andere Wissenschaftszweige auch war und ist die Molekulargenetik Quelle der Inspiration für Kunst aller Art, also auch für Musik. Insbesondere DNA-Sequenzen selbst haben einige Charakteristika, die die Assoziation zur Musik naheliegend machen. Zum Ersten ist eine DNA-Sequenz ähnlich einer einstimmigen Melodie eine Sequenz diskreter Bausteine, wobei zufälligerweise drei der vier Basen (a,c und g) sogar mit Buchstaben abgekürzt werden, die auch Notennamen sind. Das Lesen des genetischen Codes erfolgt unter natürlichen Bedingungen sequenziell, und auch der Zusammenbau des Proteins geschieht Aminosäure für Aminosäure. Diese Parallelen wurden grundsätzlich bereits 1979 von Gerald Hofstädter in seinem Buch 'Gödel, Escher, Bach' [77] gezogen, in dem er einen DNA-Strang mit einem Magnetband und ein Ribosom mit dem Lesekopf verglich, was ihn zu einem viel zitierten Autor im Bereich Sonifikation genetischer Daten gemacht hat.
Ohno, dessen Ansatz zur Sonifikation im Abschnitt zur Wissenschaft bereits vorgestellt wurde, vergleicht DNA und Musik und sieht für beide in der variierten Wiederholung kurzer Abschnitte die Grundlage des komplexeren Ganzen ([78]). Während er diese These aufgrund seiner wissenschaftlichen Tätigkeit im Bereich der DNA argumentativ unterfüttern kann, bleiben seine Ausführungen im Bereich der Musik sehr spekulativ. Der Artikel ist jedoch auf jeden Fall ein gutes Zeugnis dafür, dass der Wissenschaftler Ohno von einer künstlerischen Perspektive der DNA fasziniert war, und dass der Weg von der DNA-Sequenz zur Melodie keine abwegige Assoziation ist.
Wentian Lee führt in [79] eine statistische Gemeinsamkeit von Musik und DNA an, die letzendlich durch die von Ohno diskutierte modifizierte Wiederholung entsteht. Diese Gemeinsamkeit ist eine Langzeitkorrelation mit einem Power-Spektrum in Form eines 1/f -Rauschens. Lee wirft direkt im Abstract die Frage auf, ob es irgendwelche inhärenten Verbindungen zwischen Musik und DNA geben könnte, zieht jedoch letztendlich keine konkreten Schlüsse. Dass ein solches Spektrum in Musik überhaupt vorliegt wird zudem in [80] in Frage gestellt. Auch dies ist jedoch ein Beispiel für einen Wissenschaftler, der objektiv nachweisbare Parallelen zwischen Molekulargenetik und Musik sieht.
Neben diesen 'weltlichen' Brücken ist sicherlich auch die Nähe von der Genetik zum Leben an sich eine Eigenschaft, die das Thema für Menschen künstlerisch interessant macht.

Bill Clinton nannte den genetischen Code *„[...]the language in which god created life“ ([81], von Webseite)*. Die Musik als Mittler von Informationen und Gefühlen, die wir nicht in Worte zu fassen vermögen, könnte vielleicht auch die endlosen Ketten von Basen anders erschließen als die Wissenschaft und uns ein tieferes Verständnis der Schöpfung respektive des Wunders des Lebens ermöglichen. Derartigen Hoffnungen begegnet man öfter im Bereich von DNA-Musik, und ich werde später darauf zurückkommen. Zunächst jedoch folgt ein Überblick über die kompositorischen Annäherungen an die Molekulargenetik. Im Internet finden sich einige teils sehr gut kommentierte Übersichten ([82], [83], [84][44]). Alle drei Quellen sind US-Amerikanisch, listen jedoch auch internationale Beiträge auf. Es finden sich dort auch einige der Ansätze, die in dieser Arbeit als wissenschaftlich eingestuft wurden. Insgesamt geben alle drei einen guten Überblick, komplett sind sie allerdings nicht.

Die Bandbreite von Ansätzen für Musik aus DNA und Proteinen ist etwas breiter als im wissenschaftlichen Kontext. Auch hier jedoch bilden Mappings verschiedener Art die Grundlage der Mehrheit der Kompositionen.

In der ersten Gruppe solcher Mappings basiert auf den Basen und Aminosäuren selbst. Die erste Umsetzung nach diesem Prinzip, das sich quasi gar nicht von den ersten wissenschaftlichen Ansätzen unterscheidet, stammt von David Deamer, seines Zeichens Chemieprofessor an der UCSC, der 1982 den Basen die 4 Noten A, C, G und E zuordnet.[45] Ein anderes Beispiel für diese Herangehensweise ist die Arbeit Todd Bartons aus dem Jahr 2001 ([86]). Auch Barton bildet die Basen auf Tonhöhen ab, verwendet jedoch die derart erhaltenen Melodien als Rohmaterial für einen kompositorischen Prozess. Er verlängert die Noten teils extrem, lässt sie überlappen, vergrößert oder verkleinert gelegentlich Intervalle und setzt Crescendi und Decrescendi ein. Das Ergebnis ist ein ambientartiger Klangteppich.

Für die zweite Gruppe von Mappings bilden nicht die Basen und Aminosäuren selbst, sondern ihre physikalischen und chemischen Eigenschaften den Input. Ein Beispiel dafür sind die Arbeiten von Peter Gena ([87], [88]), der ein MAX/MSP Patch namens 'DNA Mixer' programmiert hat. Die betrachteten Datendimensionen sind die Gesamtzahl der Wasserstoff-bindungen im Codon, die Molekulargewichte und die Dissoziationskonstanten der Aminosäuren. Mittels dieses Inputs werden Pitch, Intensität und Dauer der Tonereignisse ermittelt. Das Programm verwendet bis zu acht Instrumente gleichzeitig, die anhand der jeweiligen Oberklasse einer Aminosäure gewählt werden können. Gena nimmt keine nachträglichen Manipulationen am Score vor, sondern belässt ihn so, wie ihn das Programm erzeugt. Der 'DNA Mixer' erlaubt allerdings die Wahl unterschiedlichster virtueller Instrumente. Daher haben die Kompositionen eine sehr große Bandbreite hinsichtlich ihres Charakters und lassen sich nicht allgemein beschreiben. Der 'DNA Mixer' kam bereits in mehreren teils multimedialen Werken zum Einsatz (siehe [88]), darunter in dem multimedialen Projekt 'Genesis' (s.u., vgl. [89]).

Bei David Lindsay, unter anderem Kolumnist für die 'New York Press', steht der Moment der Dissoziation eines Strangs unter natürlichen Bedingungen im Fokus des Interesses. Die Wasserstoffbrücken zwischen den komplementären Basen werden nacheinander aufgebrochen. Die Trennung des Paars G-C dauert dabei etwas länger, da hier drei solcher Bindungen aufzulösen sind. Zudem, so behauptet Lindsay unter Berufung auf nicht genannte Quellen, laufe die Auftrennung in Gruppen von 4 Basenpaaren ab. Dann folge eine Pause. Die Auftrennung eines Strangs erfolgt in beide Richtungen zugleich.

44 Hier sind viele der Links leider verwaist.

45 Auf Youtube findet sich dazu das Video eines Vortrags, den Deamer beim Bay Area Festival of Science 'Wonderfest 2007' gegeben hat ([85]).

Dissoziierende DNA bildet so für Lindsay eine Art Rythmusmaschine, und das Ergebnis seiner Sonifikation ist ein Drum-Solo ([90]).

Neben den rein sequenzbasierten Mappings gibt es auch andere Ansätze, die Komponisten für die Arbeit mit genetischen Daten verwenden. Alexander Mihalic arbeitet für seine Komposition, ohne es selbst direkt so zu benennen, im Prinzip mit Earcons ([91]). Anhand chemischer und physikalischer Eigenschaften der einzelnen Atome einer Aminosäure erstellt er jeweils eine melodische Phrasen mit einer Länge zwischen acht Sekunden und einer Minute. Dabei verwendet er die melodischen Bausteine, die er bereits für ein anderes Projekt entwickelt hatte, bei dem er die chemischen Elemente in Musik übersetzt hat. Die Phrase einer Aminosäure ergibt sich also teilweise aus der Phrase ihrer molekularen Bausteine. Die fertigen Phrasen sind bereits die Komposition. Die konkrete Ausgestaltung in der Zeit obliegt dem Interpreten selbst[46].
Linda Long nutzt für ihre Arbeiten Proteinstrukturen, die sie mittels Röntgen-Kristallografie ermittelt, als Grundlage eines Mapping, das sie allerdings nicht näher beschreibt. Die offensichtlich MIDI-basierten, kurzen Ausschnitte ihrer CDs 'Music of the Body' und 'Music of the Plants', die es auf ihrer Webseite zu hören gibt ([92]), wirken wie konventionelle Entspannungsmusik. Gary Burd beschreibt es in seiner Rubrik 'Best of the Web' in 'The Biochemist' recht treffend.

> *They all have a sort of dreamy underbelly, with a piano plinking over the top, and are certainly of a relaxing and soothing nature.([93], p. 40)*

Die Komponistin Susan Alexjander hat für ihre Arbeit mit dem Genetiker David Deamer (s.o.) kooperiert. Mit Hilfe ultravioletten Lichts wurden die Absorptionsspektren der einzelnen Basen gemessen und dann in den hörbaren Frequenzbereich transformiert. Die größten Peaks bildeten die Grundlage für eine mikrotonale Skala für jede Base. Diese Skalen sind das Tonmaterial für Alexjanders Kompositionen, die unter dem Titel 'Sequencia' auch auf CD erhältlich sind. Auf ihrer Webseite ([94]) findet man die Online-Version des IEEE-Beitrags ([95]), in dem sie das Vorgehen eingehend beschreibt. Interessant sind auch ihre Ausführungen über Musik als allem Inhärenten und Gleichnis des Lebens selbst, sowie ihre Spekulationen zur Wirkung ihrer Musik. Sie habe ihr Werk 'Sequencia' zunächst einmal komponiert *„[...] because that is what a composer does in the world; write music.*“ (Onlineversion von [95]). Doch, sagt sie weiter, habe das konsistente Feedback vieler Hörer ein Nachdenken angeregt, ob es eine Art ganzheitliche Form des Erkennens seitens der Hörer gibt.

> *People report feelings of connectedness, familiarity. 'I feel right at home,' they say. It is tempting to speculate that the body is recognizing itself, and is communicating this to the psyche. (ebd.)*

Auf der Webseite finden sich zwei Ausschnitte aus der CD. In einem dominieren Flächen hoher synthetischer Streicher, im anderen sind eher glockenartige Timbres zu hören, wobei beide Beispiele sehr ambienthaft und im Sinne des abendländischen Musikverständnisses harmonisch sind.

46 Natürlich kann auch eine Aminosäurefolge das Werk in seiner Form determinieren.

Die bis hierhin vorgestellten Werke sind Beispiele und sollen wie eingangs erwähnt exemplarisch die Bandbreite von 'Gen-Musik' hinsichtlich Technik, Ergebnis und Motivation illustrieren. Ein systematischer Blick auf die 'Gen-Musik-Szene' insgesamt offenbart eine Reihe gemeinsamer Charakteristika, die im Folgenden zusammengetragen werden sollen.
Zum ersten sind es nicht selten Wissenschaftler aus dem Umfeld der Molekularbiologie oder Chemie, die sich mit DNA musikalisch auseinandersetzen, oder an derartigen Projekten partizipieren[47]. Zum Zweiten ist Zahl der Arbeiten, denen man wirklichen künstlerischen Schöpfungsdrang attestieren kann, eher kurz. Einfache Online-Tools zum MIDI-Mapping wie der 'PROM Composer' von L.Y. Han et al. ([96]) sind, auch wenn sie das Wort Musik verwenden, sicher nicht zu solchen zu rechnen.
Auch wenn es vielleicht unangemessen ist, über künstlerische Qualität zu urteilen, kann man doch außerdem feststellen, dass selbst Arbeiten, die als CD erschienen sind, wie die von Linda Long (s.o.) oder Aurora Sanchez Sousa ([97]), etwas flach, beliebig und im schlimmsten Fall 'billig' wirken. Gerade Letztgenannte verweist auf ihrer Webseite auf eine beachtliche Zahl von Erwähnungen ihrer Arbeit in der Presse. Das verwundert etwas angesichts der Tatsache, dass die Basis hier ein einfaches Pitch-Mapping der Basen war. Dieser Ansatz war zu der Zeit (Jahr 2000) keineswegs mehr neu, und auch das Ergebnis ist wenig beeindruckend und aufgrund seiner stark westlich-tonal geprägten Charakteristik überaus konventionell.[48]
Die meisten der hier vorgestellten Werke sind zumindest einmal von der Presse rezipiert und besprochen worden. Die Bandbreite reicht dabei von wissenschaftlichen Fachblättern wie 'The Biochemist' bis hin zur 'New York Times'. Interessanterweise, und das ist ein weiteres gemeinsames Merkmal, ist 'Gen-Musik' jedoch nicht Teil allgemeiner Abhandlungen über genetisch inspirierte Kunst. In keiner der vier hier herangezogenen Quellen ([98], [99], [100], [101]) findet sich auch nur eine Erwähnung von Musik oder Klangkunst. Alle Beiträge reflektieren und besprechen Kunstprojekte, die Genetik als Thema oder Inspiration haben, ohne den Begriff 'Kunst' in irgendeiner Weise zuvor einzugrenzen. Die gegebenen Beispiele gehen dementsprechend von Malerei und Fotografie über Installationen multimedialen Charakters bis hin zu Manipulationen von genetischem Material selbst. Musik jedoch fehlt gänzlich. Das Multimedia-Projekt 'Genesis' von Eduardo Kac (vgl. [89]), in dessen Zentrum eine in die DNA eines Virus eingeschleuste Bibelsequenz steht, wird beispielsweise in allen vier Aufsätzen besprochen. Die Musik zu eben jener Installation stammt vom oben vorgestellten Komponisten Peter Gena und seinem 'DNA Mixer' und erfüllt zweifelsohne die Axiome 'ist Kunst' und 'basiert auf molekulargenetischen Daten'. Trotzdem findet dieser Aspekt der Installation in keinem der Artikel Erwähnung. Vielleicht ist eine Begriffsunschärfe der Grund für dieses Fehlen, und die Autoren subsumieren Musik nicht unter Kunst, sondern sehen beides auf einer Stufe nebeneinander. Es wäre jedoch auch möglich und denkbar, dass sie die wenigen musikalischen Ansätze mit ästhetischer und kreativer Substanz als zu uninteressant, zu wenig kontrovers oder zu wenig künstlerisch angesehen haben.
Vielleicht liegt der Grund für das Fehlen aber auch in der Intention, die fast allen Werken der Gen-Musik gemein ist, und die sich von den Intentionen, die Werken wie 'Genesis' zugrunde liegt, deutlich unterscheidet. Alle Arbeiten, die in den vier Quellen zu 'Kunst und Genetik'

47 Ob dieses Phänomen in irgendeiner Weise für die Genetik spezifisch ist, würde einer breiteren Recherche bedürfen und kann daher hier nicht beantwortet werden.

48 Die hohe Anzahl an Presseechos könnte vielleicht auch damit zu erklären sein, dass die Arbeit von einem Redakteur der Associated Press (AP) in einem Artikel beschrieben wurde, und somit in einer gut vernetzten Infrastruktur lag. Eine einfache Google-Suche einer längeren Phrase stützt diese These, da sie in verschiedensten Quellen auftaucht (u. a. bei CBS New unter anderem Namen).

vorgestellt werden, haben eine Art Message oder einen ideologischen Überbau und thematisieren soziale und ethische Fragen. Dabei scheint der Tenor eher skeptisch bis kritisch zu sein. Im Artikel von Madoff ([99]) werden Craig Venter und Francis Collins, zwei an der Entschlüsselung des Humangenoms in Schlüsselfunktionen beteiligte Wissenschaftler, mit unabhängigen Aussagen zitiert, laut derer sie, sinngemäß wiedergegeben, über eine negative Einstellung seitens der Kunst gegenüber der Genetik enttäuscht sind.

> *'The way in which this instruction book is put together, the dance within the cell that alters in microseconds in response to thousands of different circumstances and makes the right decision to keep that cell healthy. It's stunning. There are plenty of ethical issues to address, but I wish more of that sense of awe was honored.'*
> *(Dr. Francis Collins, zit. in [99])*

Genau diese Wertschätzung des Wunders des Lebens findet sich in vielen Statements von Komponisten von Gen-Musik. Todd Barton begann die Präsentation zu seiner Musik im Smithonian mit den Worten:

> *I stand in awe of the genetic code, its deep countrapuntal complexity born of utter simplicity: just four bases: A,T,C and G spin out patterns of chemical recipes for the creation of life! ([86], von Webseite),*

und endete nach einer knappen Beschreibung seiner Vorgehensweise mit:

> *So. . . what does this all mean? For me, as a composer, it is a touchstone of curiosity. A springboard for metaphor. A map of wonder. (ebd.).*

Susan Alexjander schreibt auf ihrer Webseite über ihre CD „*SEQUENCIA is a unique and lyrical celebration of the miracle of life.*“ ([102], von Webseite), und auf der Startseite von Linda Longs Homepage steht unter anderem:

> *Molecular Music is computer generated music that creates a gateway into understanding life, allowing us to perceive nature in her many forms through sound. Music is not limited by words or logic and so may communicate with us on deeper levels, freeing us to directly experience the essence of all living matter and giving us an insight into the voice of nature. ([92], von Webseite).*

Diesen und auch anderen, hier nicht exemplarisch angeführten Komponisten geht es augenscheinlich um eine Abbildung der Natur in musikalischer Form, auf dass der Hörer die Grundlage allen Lebens sinnlich erfahren könne. Gen-Musik hat mehrheitlich keine gesellschafts- oder sozialkritische Intention, sondern eher eine deskriptive und vermittelnde. Die Zielsetzungen wissenschaftlicher Ansätze und künstlerischer Auseinandersetzungen mit dem Thema sind sich also sehr ähnlich. Den Wissenschaftler interessiert vielleicht der konkrete Anteil an C und G, den Musiker das 'Wesen' eines Proteins.
Das letzte gemeinsame Charakteristikum von Gen-Musik oder besser der 'Gen-Musik-Szene' ist eine gewisse Affinität zur Esoterik, die sich im Umfeld vieler Arbeiten unterschiedlich stark ausgeprägt beobachten lässt. Blickt man beispielsweise bei 'Amazon' auf die Artikel, die Käufer von Susan Alexjanders CD sonst noch gekauft haben, findet man Titel wie 'Chrystal

Bowl Chackra Chants' und 'Seven Secrets of Sound Healing'[49], und die Beschreibung von Linda Longs CD 'Music of the Body' enthält unter anderem folgendes Zitat:

> *Its [Anm.: die CD] soothing and uplifting properties make it ideal for use by health professionals and alternative medicine practitioners to accompany their therapeutic work (e.g. massage therapy with essential oils, reiki and music therapy), and by individuals to enhance meditation, mind-body visualization techniques, and bodywork e.g. yoga. ([103], von Webseite).*

Wie im schon erwähnt, streben Komponisten von Gen-Musik, so sie ihre Projekte nicht nur aus purer Neugier und Freude an der Idee betreiben, danach, das faszinierende und wunderbare des 'Systems DNA' zu erfassen und hörbar zu machen. Dieser Wunsch wird nicht selten von der Idee begleitet, dass die resultierende Musik auf den Körper, aus dem sie letztlich stammt, in irgendeiner Form zurückwirken könnte[50], ein Wunsch, den zumindest Joel Sternheimer (vgl. 3.1) als Realität bezeichnen würde.

> *Sternheimer warns scoffers to be careful tinkering with the tunes because they can affect people as well. 'Don't ask a musician to play them', he says. 'You must be very careful.' Sternheimer says that one of his musicians had difficulty breathing after playing the tune for cytochrome C too often. ([105], p. 2)*

Die Ergebnisse lassen sich wie folgt zusammenfassen. Die Gen-Musik ist ein überschaubares Feld, entsteht oft als Nebenbeschäftigung von Wissenschaftlern und wird als Kunst wenig rezipiert. Ihre Intention ist oftmals ein sinnlicher Zugang zum Urgrund des Lebens. Sie orientiert sich mehrheitlich sehr stark am tonalen System des Abendlandes und ist, und diese Aussage ist natürlich rein subjektiv, qualitativ oft eher durchschnittlich.

49 Stand: September 2011

50 Die Hoffnung auf eine solche positive Wirkung ist, neben purer Eitelkeit, wohl auch die Lebensgrundlage von Internetseiten wie Yourdnasong.com [104]. Ab 259 Britischen Pfund aufwärts bekommt man dort zwei Musik-CDs und einen Handy-Klingelton basierend auf der eigenen DNA – und das alles im Musikstil seiner Wahl.

4. Möglichkeiten zur Sonifikation genetischer Daten

Dieser Teil versucht herauszuarbeiten, welche Anwendungsmöglichkeiten es für Sonifikation im Bereich Molekulargenetik gibt. Zu Beginn werden die verschiedenen Arbeitsbereiche der Bioinformatik hinsichtlich Aufgaben untersucht, bei denen der Einsatz von Sonifikation prinzipiell denkbar und sinnvoll ist. Im Anschluss geht es um die Fragen, welche Datendimensionen und Informationen im Bereich der Molekulargenetik vorliegen, wann diese von Interesse sind und wie sie bisher im wissenschaftlichen Alltag dargestellt werden.
Im dritten Kapitel schließlich werden basierend auf den gewonnenen Erkenntnissen konkrete Sonifikationen vorgeschlagen. Einige dieser Vorschläge sind im Rahmen der Zusammenarbeit mit Fr. Maïda realisiert worden. So dies der Fall ist, werden die Ergebnisse diskutiert.

4.1 EINSATZSZENARIEN

Um mögliche Anwendungsgebiete für die Sonifikation molekulargenetischer Daten zu finden, soll zunächst ein Blick auf die Arbeitsfelder der Bioinformatik geworfen werden. Obwohl sie noch eine recht junge Disziplin ist, hat sie schon eine enorme Bedeutung, denn ihre Aufgabe ist es, die riesige Menge an Sequenzdaten so aufzubereiten, dass sie einen Nutzen haben. Das Arbeitsgebiet Bioinformatik lässt sich in verschiedene Aufgabenbereiche aufteilen.

Assembly

Die Assembly ist der Prozess, die einzelnen, kurzen 'reads', die jede bislang realisierte Sequenziertechnik liefert, zu möglichst langen, kontinuierlichen Sequenzen zusammenzusetzen. Dazu verwendet man Algorithmen, die die 'reads' nach überlappenden Sequenzabschnitten durchsuchen. Eine vollständige, lückenlose Rekonstruktion eines Chromosoms ist wegen der repetitiven DNA kaum zu erreichen.
Der Output der Mehrzahl von Next-Generation-Sequenzierautomaten besteht aus im Vergleich zur klassischen Sanger-Methode aus sehr kurzen 'reads' mit einer Länge von unter 100 Basenpaaren. Diese Tatsache stellt die Bioinformatik vor große Herausforderungen und führt unausweichlich zu einer fragmentierteren Assembly ([106]).
Die Assembly findet hochautomatisiert statt.

Alignment

Alignment ist der Vergleich zweier oder mehrerer Sequenzen und kommt bei der Assembly und bei der Untersuchung von genetischer Verwandschaftsbeziehungen zum Einsatz. Alignments werden zumeist computergestützt durchgeführt. Im Kern wird immer ein Kennwert dafür ermittelt, wie stark zwei Sequenzen übereinstimmen, wobei es gibt eine ganze Reihe

verschiedener Kennwerte gibt. Eine einfaches Maß wäre zum Beispiel das Verhältnis aus übereinstimmenden Basen zur Gesamtlänge der zu vergleichenden Abschnitte. Andere Algorithmen geben komplett übereinstimmenden kurzen Abschnitten mehr Gewicht als verteilten Einzeltreffern. Trotz der hohen Automation ist der Mensch als Reviewer unverzichtbar, sei es zur Qualitätsprüfung eines Algorithmus, zum manuellen Alignment, wenn sich Alignment-Kriterien nicht adäquat als Algorithmus formulieren lassen und natürlich zur Interpretation der Ergebnisse.

Sequenzanalyse

Ziel der Sequenzanalyse ist es, aus einer Basenfolge Rückschlüsse hinsichtlich einer möglichen Funktion des DNA-Abschnitts zu ziehen. Anhand bestimmter statistischer und struktureller Eigenschaften kann eine Sequenz beispielsweise auf die Frage hin analysiert werden, ob sie codierendes Material enthält. Ein simples Beispiel dafür ist die Betrachtung der drei Stop-Codons. In einer reinen Zufallsequenz müsste in etwa jedes 21. Codon ein Stop-Codon sein. Ein Fehlen derselben über einen längere Abschnitt ist daher ein Indiz dafür, dass die entsprechende Sequenz ein Gen enthält. Das gilt so jedoch nur für Prokaryoten, deren CDS 'am Stück' vorliegen, also nicht durch Introns unterbrochen sind. Bei Eukaryoten ist die Suche komplexer, jedoch gibt es auch hier eine Vielzahl von Indikatoren wie zum Beispiel Promoter-Strukturen, bestimmte Pattern an den Grenzen zwischen Introns und Exons und eine Reihe rein statischer Eigenschaften der Sequenz (vgl. 4.2.1).

Dateninfrastruktur

Ein Teilbereich der Bioinformatik beschäftigt sich mit der Speicherung, Annotierung und Darstellung der riesigen Datenmengen.
In den drei großen Sequenzdatenbanken, GenBank, DDBJ und EBML[51], die sich täglich aktualisieren und austauschen, waren im August 2011 142.284.608 Sequenzen mit insgesamt 130.671.233.801 Basen gespeichert ([107]). Wissenschaftler aus aller Welt können mittels FTP-Upload oder E-Mail Sequenzen in diese Datenbanken einstellen, Annotationen vornehmen oder in der Datenbank nach Sequenzen suchen. All das erfordert eine gute IT-Infrastruktur und allgemein anerkannte, einheitliche Standards wie zum Beispiel das FASTA-Datei-Format. Die Entwicklung dieser Standards und der Infrastruktur ist ebenfalls ein Teil der Bioinformatik.

Zumindest ein Teil des eben beschriebenen Aufgabenkatalogs findet sich auch unter den möglichen Einsatzgebieten von Auditory Display (vgl. 2.2.2), und so kann man, abgesehen von der Kunst, im Wesentlichen zwei Einsatzfelder für Auditory Display im Bereich Genetik festmachen. Visuellen Displays können um eine akustische Komponente erweitert werden, die den visuellen Kanal entlastet und ergänzt, beziehungsweise für sehbehinderte Menschen vollständig ersetzt, und die Molekulargenetik könnte Sonifikation für die explorative

51 GenBank vom National Center for Biotechnology Information (NCBI), DNA Data Bank of Japan (DDBJ), European Molecular Biology Laboratory (EBML)

Datenanalyse einsetzen, die einerseits die Sequenzanalyse, aber auch Alignments und auf Gen- oder Chromosomstatistiken beruhende Homologie-Suchen umfasst.
Einer der in 3.1 vorgestellten Ansätze dient explizit dem Data Mining. Die anderen Ansätze jedoch haben den Schwerpunkt auf der Darstellung. Den Input bildet dabei mehrheitlich die DNA-Sequenz in ihrer Buchstabenrepräsentation.

4.2 GENETISCHE DATEN – DIMENSIONEN UND DARSTELLUNG

In diesem Kapitel wird Zweierlei zusammengetragen. Der erste Teil der Fragestellung lautet: welches sind die Datendimensionen genetischer Daten, und bei welchen Fragestellungen sind diese von Interesse? Dazu werden drei verschiedene 'Einheiten' oder Datenklassen unterschieden - eine Sequenz, ein Gen und ein Chromosom. Jede Klasse hat einige exklusive Datendimensionen. Allen gemeinsame Dimensionen sind abhängig von der Klasse von unterschiedlicher Relevanz. Die zweite Frage ist die nach der Art der Darstellung genetischer Daten, da ein Blick auf die Wege der visuellen Repräsentation durchaus Impulse für eine Umsetzung in eine Sonifikation liefern kann.
Die grundlegende Klasse ist eine Sequenz. Gene und Chromosomen respektive lange DNA-Abschnitte sind natürlich immer auch Sequenzen, haben also auch alle entsprechenden Datendimensionen. Daher nimmt der Bereich der Sequenz hier mit Abstand den größten Raum ein.

Alle Beispiele und Werte beziehen sich auf das Humangenom, da Frau Maïda, für die im Zuge dieser Arbeit das Programm entsteht, zunächst mit diesem arbeiten wird. Die Datendimensionen selbst sind für genetisches Material weitestgehend universell.
Informationen über Proteine sowie die Sekundärstrukturen der RNA werden nicht berücksichtigt, da sie den Umfang dieser Arbeit sprengen würden.

4.2.1 DNA-Sequenzen

Eine DNA-Sequenz ist eine Folge aus Nukleotiden. Einmal sequenziert liegt sie in Dateien als eine Folge der Buchstaben a, t, g und c vor, man kann sie aber natürlich auch mittels geometrischer Formen oder Farben darstellen. Allgemein ausgedrückt handelt es sich um lange Folgen von vier verschiedenen Zuständen. Vier verschiedene Zustände lassen sich binär mittels 2 bit codieren, eine DNA-Sequenz kann also in einen 2-bit Datenstrom übersetzt werden. Um alle 64 möglichen Triplets zu codieren benötigt man 6 bit, für die 20 proteinogenen Aminosäuren noch 5.
Das menschliche Genom hat eine Länge von rund 3,1 Milliarden Basenpaaren. Eine Textdatei dieser Sequenz hätte einer Größe von etwa 739 MB. Das Erbgut ist auf 24 Chromosomen (inklusive X und Y) mit Längen zwischen 50 Millionen und 250 Millionen Nukleotiden aufgeteilt.
Man unterscheidet eine ganze Reihe von Sequenzklassen (vgl. 2.1), deren Längen sich teils

erheblich unterscheiden.

'Read'	10 – 1000	vgl. 2.1
Chromosom	50 Mio.(Chr. 21) -250 Mio (Chr. 1)	([108], Build GRCh37.p3)
Gen (inklusive Introns)	14 - ca. 2.500.000 , Median: 14.000	([109], Tab. 1, [110], p. 38)
Gen (Coding Sequence)	(keine Quelle) - 80.780 , Median: 1100	([110], p. 38)
Intron	< 10 – 500.000 , Median: 1023	([109] Tab. 1, [111], p 3 , [110], p. 38)
Exon	<10 – 17.106 , Median: 122[52]	([109], Tab. 1, [110], p. 38)

Tabelle 1: Übersicht über die Längen verschiedener Klassen von Regionen

Neben diesen Sequenztypen gibt es eine Reihe regulatorischer Sequenzen, deren Länge zumeist im Bereich unter 100 Bp, teils unter 10 Bp liegt (vgl. [112]). Der Wertebereich von Sequenzlängen umfasst also insgesamt elf Zehnerpotenzen, und auch in mehrheitlich kurzen Sequenzklassen gibt es Einzelfälle, die um ein Vielfaches länger sind.

Für jede Sequenz, gleich wie lang, lassen sich eine Reihe statistischen Angaben machen. Die grundlegendsten sind die Häufigkeiten der vier Basen, der Triplets und der codierten Aminosäuren. Ein weiteres, oft ermitteltes Feature einer Sequenz ist ihr CG-Gehalt, der Hinweise darauf liefert, ob die Sequenz codiert oder nicht, und der bereits als Input einer Sonifikation verwendet wurde (vgl. 3.1).
Neben diesen grundlegenden gibt es eine Reihe weiterer statistischer Dimensionen, die für Sequenzen ermittelt und als Indizien bezüglich der Funktion der Sequenz genutzt werden. Ein umfassender Überblick über solche statistischen Methoden und Kennwerte mit dem Fokus auf der Identifizierung codierender Sequenzen findet sich zum Beispiel in [113]. Robertas Damaševičius schlägt in [114] eine zweistufige Analyse von DNA-Sequenzen vor, und die zwei Stufen sind gleichzeitig gute Kategorien der Datendimensionen.

> *The complexity metrics, which can be used to analyze DNA sequences, can be broadly categorized into two groups: information-theoretic metrics (Shannon entropy, Rényi entropy,Kolmogorov complexity) and frequency analysis metrics (time delayed mutual informationand autocorrelation function). ([114], p. 6)*

Die erste Gruppe ermittelt also den Informationsgehalt für eine beliebige Sequenz. Wäre diese komplett zufällig, hätte jede Base an jeder Position dieselbe Wahrscheinlichkeit und wie oben schon erwähnt einen Informationsgehalt von 2 bit. DNA-Sequenzen sind jedoch nicht komplett zufällig. Für Sequenzen mit einem hohen Anteil an repetitiver DNA kann der mittlere

52 Hier anzumerken: *„About 80-85% exons on each chromosome were found to be less than 200 bp in length“* ([109], von Webpage)

Informationsgehalt deutlich niedriger liegen (vgl. ebd.). Informationstheoretische Maße wie die Shannon-Entropie u. a. können nützliche Hinweise auf den Typ der Sequenz liefern.
Die Gruppe der frequenzbezogenen Maße ist im Hinblick auf Sonifikation besonders interessant. Man kann grob zwei Arten von Frequenz-Phänomenen unterscheiden, die im Zusammenhang mit DNA-Sequenzen eine Rolle spielen. Zum Ersten finden sich in DNA-Sequenzen teilweise Langzeitkorrelationen, die sich über sogar über Chromosomen erstrecken können und deren Spektren stets den Charakter eines 1/f -Rauschens haben (vgl. [79]). Die zweite Art von Frequenzphänomenen sind eine Reihe verschiedener, in DNA-Sequenzen feststellbarer Periodizitäten ([113], [114], [116]), die sehr verschiedene Ursachen haben. Einige sind beispielsweise die Folge der 'Packung' der DNA. Die am meisten zitierte und verwendete ist die 3'-Periodizität codierender Sequenzen. Ihre Ursache ist zum einen die Tatsache, dass Codons immer aus drei Basen bestehen und zum anderen der sogenannte Codon-Bias. Der genetische Code ist degeneriert, das heißt, mehrere Codons codieren dieselbe Aminosäure. Innerhalb einer Spezies gibt es jedoch Präferenzen für bestimmte Codons, während andere, obschon sie dieselbe Aminosäure codieren, weit weniger häufig vorkommen (vgl. [117]). Daraus resultiert ein Übergewicht bestimmter Codons, und somit auch bestimmter DNA-Patterns, die immer im Abstand von 3*n auftreten. Die 3'-Periodizität findet sich in den Exons aller Spezies, und sie ist ein sehr guter Indikator dafür, ob eine Sequenz codierend ist oder nicht[53]. Es würde zu weit führen, hier alle Periodizitäten und sämtliche Möglichkeiten der Frequenzauswertung im Detail vorzustellen, und es soll genügen, sie und die verschiedenen Informationsgehalte als Datendimensionen einer Sequenz festzuhalten.

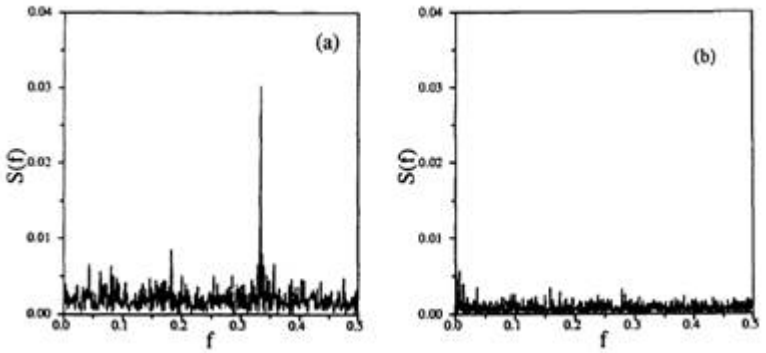

Abbildung 7: Fourierspektrum eines codierenden (links) und eines nicht codierenden (rechts) DNA-Abschnitts von Backhefe.
(Quelle: [115]*, p. 2)*

Abschließend soll noch ein Blick auf verschiedene Formen der Sequenzdarstellung geworfen werden, denn eine Symbolfolge ist nur eine Möglichkeit von vielen, eine DNA-Sequenz zu visualisieren. Bei längeren Sequenz wird diese Darstellung schnell unübersichtlich und Ähnlichkeiten sind für den Menschen schwer bis gar nicht erkennbar. Daher gibt es diverse

53 Das P-3-Feature und natürlich auch andere Periodizitäten lassen sich auf verschiedene Weisen feststellen bzw. quantifizieren. Eine naheliegende Möglichkeit ist eine Fouriertransformation, wie in [115] beschrieben. Die Autoren zerlegen die Sequenz zunächst in vier digitale Sequenzen für die einzelnen Basen. Diese werden dann fouriertransformiert und zum Gesamtspektrum aufsummiert.

Ansätze zur Sequenzdarstellung, die sich eventuell auch für Sonifikation als nutzbar erweisen könnten. Abbildung 8 zeigt eine mögliche Visualisierung, die an ein grob quantisiertes Zeitsignal erinnert. Ein visueller Vergleich der beiden gezeigten 450 Basen langen Sequenzen fällt jedoch schwer.

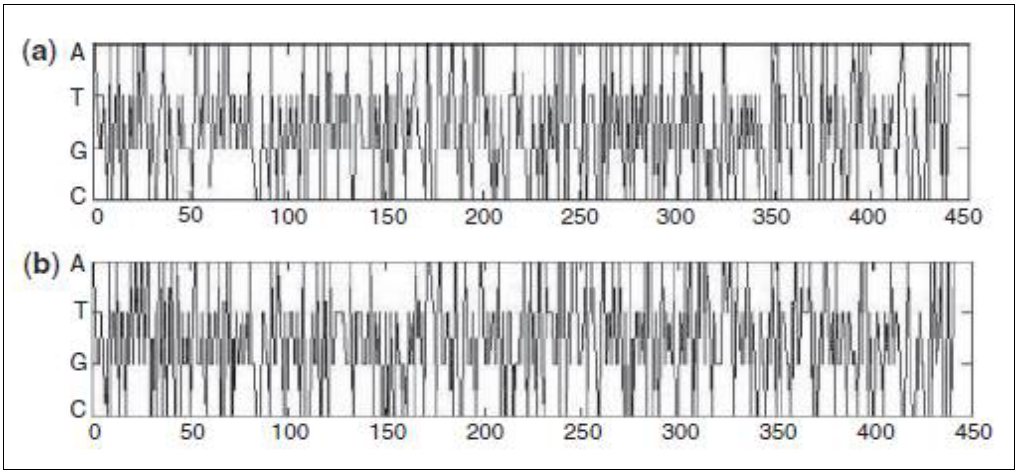

Abbildung 8: Komplette Coding-Sequence des β-Globin -Proteins von Mensch (oben) und Opossum (unten); (Quelle: [118], p.1112)

Eine Reihe verschiedener Ansätze stellt DNA-Sequenzen in Form einer Kurve dar. Allen diesen Ansätzen ist gemein, dass sie jeder Base der Sequenz abhängig von ihrer Position einen Punkt im zwei- oder dreidimensionalen Raum zuordnen, und diese Punkte dann zu einer Kurve verbinden. Die Abbildung sollte dabei möglichst umkehrbar sein und die Länge der Sequenz reflektieren. Yonghui Wu et al. schlagen in [119] die 'DB-Curve' vor. Der Name DB-Curve, kurz für 'Dual-Base-Curve', reflektiert die Tatsache, dass sich die Kurve auf jeweils ein Basenpaar konzentriert. Adenin wird beispielsweise der Vektor (-1,1) zugewiesen und Cytosin der Vektor (1,1), während Guanin und Thymin beide durch (0,1) repräsentiert werden, im resultierenden Bild also nicht mehr zu differenzieren sind. Ein Kurvenpunkt ergibt sich aus der Summe aller vorhergehenden Vektoren. Eine Sequenz, die lediglich aus G und T besteht, wäre in dieser Darstellung eine vertikale Linie, und die Länge der Sequenz bliebe eindeutig sichtbar. Um eine Sequenz komplett darzustellen, braucht man also zwei DB-Kurven.
Die Z-Curve, vorgestellt in [120], berechnet den Punkt einer Base im dreidimensionalen Koordinatensystem anhand der absoluten Häufigkeiten der Basen innerhalb des Teilabschnitts von Sequenzbeginn bis zur aktuellen Base. Die Autoren beweisen in der Arbeit die Eindeutigkeit und Umkehrbarkeit dieser Transformation. Sie geben weiterhin ein Beispiel, in dem sie anhand der Z-Curve eines Chromosoms verschiedene, zutreffende Schlussfolgerungen ziehen.
Die DV-Curve von Zhu-Jin Zhang ([118]) ordnet jeder der vier Basen ein spezifisches Vektorenpaar zu. Die x-Komponente jedes Vektors ist 1, während die y-Komponente entweder -1 oder 1 sein kann. Auch für diese Abbildung lässt sich die Eindeutigkeit und Stetigkeit belegen. Abbildung 9 zeigt die DV-Kurven desselben Gens verschiedener Spezies. Ähnlichkeiten und Unterschiede fallen unmittelbar ins Auge.

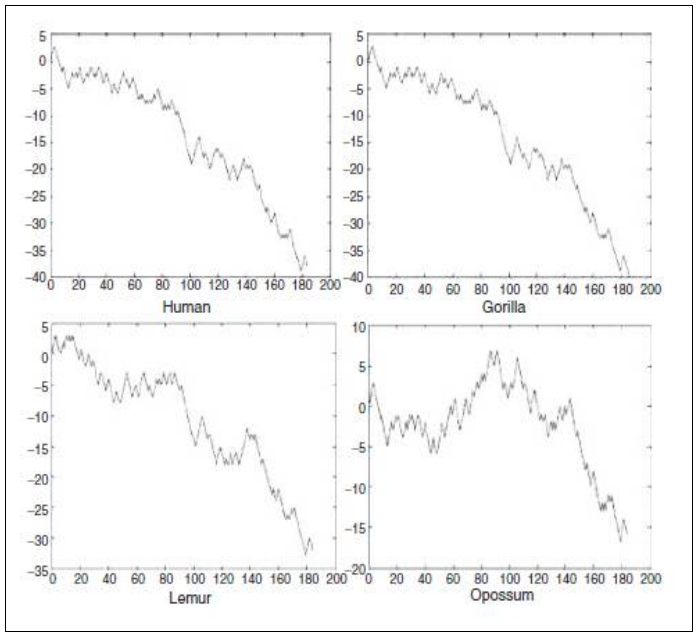

Abbildung 9: DV-Kurve für Mensch, Gorilla, Lemur und Opossum; (Quelle: [118], p. 1116)

Die Graphen haben außerdem, zumindest in diesem Fall, Ähnlichkeit mit der Hüllkurve eines Spektrums. Es wäre durchaus denkbar, dass eine auf der DV-Transformation einer Sequenz aufbauende Sonifikation einen Vergleich auch längerer Sequenzen anhand spektraler Charakteristika wie beispielsweise 'Genformanten' ermöglicht.

Neben Kurven im Euklidischen Koordinatensystem gibt es auch andere Ansätze zur Sequenzdarstellung, wie den von H. Joel Jeffrey ([122])[54]. Er verwendet die DNA-Sequenz als pseudo-zufälligen Input eines 'Chaos Game'. Jeder Base ist wird ein Eckpunkt eines Quadrats zugeordnet. Willkürlich wird ein innerhalb des Quadrats ein erster Bildpunkt gewählt. Der nächste Bildpunkt wird genau auf die Hälfte zwischen dem letzten Punkt und der Ecke der jeweiligen Base gesetzt. Das Ergebnis ist ein zweidimensionales musterartiges Bild. Die Länge einer Sequenz lässt sich nur noch bedingt anhand der Punktmenge ablesen und die Reihenfolge ist nicht ersichtlich, was aber nicht zwingend ein Nachteil sein muss. Repetitive Sequenzabschnitte werden regelmäßige Musterkomponenten erzeugen und lassen sich so gut identifizieren.
Yinhe Cao et al. schlagen in [123] eine Visualisierung vor, die, anders als alle bisherigen Beispiele, nicht mehr die Basensequenz mit ihren vier möglichen Zuständen selbst, sondern eine daraus mittels Transformation ermittelte Sequenz als Input verwendet. Die Basis der Methode ist die 'Recurrence Time'. Zunächst wird eine Wortbreite w definiert. Anschließend wird die zu transformierende Sequenz Base für Base analysiert. Jede Base ist immer der 'Anfangsbuchstabe' eines Wortes der Länge w. Die 'Recurrence Time' ist der Abstand (in Basen) vom Original eines Wortes zur ersten buchstabengetreuen Wiederholung. In einer Sequenz aaaaaaaaaaaa wären bei w = 3 also beispielsweise alle 'Recurrence Times' gleich 1, da jedes Wort unmittelbar wiederholt wird. Tritt ein Wort in einer Sequenz nur einmal auf, so ist

54 Eine Weiterentwickelung der Idee findet sich [121].

die 'Recurrence Time' gleich -1. Das Ergebnis ist eine Folge von Zahlen, mit der Länge N[55] - w. In einem zweiten Schritt werden alle -1 entfernt. Die resultierende Sequenz erlaubt laut der Autoren über die Auswertung mittels Häufigkeitsverteilungen Erkenntnisse zu allen Periodizitäten und zu repetitiven DNA-Abschnitten. Die Frage, inwieweit diese Sequenz mittels einer Sonifikation erschlossen werden kann, könnte durchaus ein lohnender Untersuchungsgegenstand sein.

Es würde hier zu weit führen, alle verschiedenen Visualisierungen von DNA-Sequenzen vorzustellen. Die angeführten Ansätze sollen vielmehr die Bandbreite illustrieren und zeigen, dass eine Betrachtung solcher Ansätze auch Impulse für eine Darstellung mittels Sonifikation abseits der bekannten MIDI-Melodien gibt.

4.2.2 Gene und Genversionen

Eine Genversion ist neben ihrer Sequenz mit den entsprechenden Datendimensionen vor allem durch ihre innere lokale Struktur, also durch die Lage und Länge von Exons und Introns charakterisiert. Im Humangenom liegt die Anzahl der Exons pro Genversion zwischen 1 und 148, die durchschnittliche Zahl pro Gen ist 8.8 und der Median bei 7 (vgl. [110]). Die Darstellung in visueller Form erfolgt meist ähnlich der in Abbildung 10.

Abbildung 10: Screenshot der Gen-Darstellung von ENSEMBL (Quelle: [108], Havanna Gene OTTHUMG00000055853 auch MOXD1)

Introns sind im Durchschnitt mehr als zehn mal so lang wie Exons (vgl. Tab. 1). Eine Darstellung wird sich also ob ihrer beschränkten Auflösung auf einen Aspekt konzentrieren müssen.
Die Visualisierung der Strukturen verschiedener Genversionen wie in Abbildung 10 erlaubt einen schnellen Vergleich. Gemeinsame und exklusive Exons lassen sich mit einem Blick erkennen, was bei einer Darstellung der Coding-Sequenz nicht so einfach möglich wäre.
Neben den einzelnen Längen von Exons und Introns sind auch die Länge der Coding-Sequence als ganzes, die durchschnittlichen Längen und die Länge der gesamten Genversion (inklusive Introns) als relevante Strukturmerkmale festzuhalten. Auch Lage und Länge der regulatorischen Sequenzen eines Gens gehören zu den Strukturinformationen. Eine Betrachtung derselben ist aber im Umfang dieser Arbeit nicht möglich und wird daher ausgeklammert.

55 Länge der Originalsequenz.

4.2.3 Chromosom / Chromosomabschnitt

Ein Chromosom oder ein langer DNA-Abschnitt[56] ist zum einen, wie auch eine Genversion eine natürlich eine Basensequenz, für die sich alle sequenzspezifischen Datendimensionen angeben lassen. Es lassen sich aber auch unabhängig von der Sequenz selbst einige Aussagen machen. Ensembl (vgl. [108]) gibt unter dem Punkt 'Chromosome Statistics' unter anderem die Gesamtzahl verschiedener Gen-Typen (RNA-Gen, Pseudo-Gen etc.) sowie die Anzahl an SNPs[57] im Chromosom an, insgesamt also neun Variablen. Robertas Damaševičius nennt in [114] in einer Zusammenfassung des Chromosoms 22 neben der Anzahl der Gene noch den GC-Gehalt, den Anteil an Repeats und den Anteil an ALU-Repeats. Offensichtlich fallen Zusammenfassungen je nach Fokus sehr unterschiedlich aus. Theoretisch kommen sowohl für ein Chromosom, als auch für einen längeren Abschnitt die absoluten und relativen Häufigkeiten aller Regionstypen des Feature-Tables (vgl. [4]) sowie sämtliche weitere annotierte Charakteristika als Datendimension in Betracht. Eine absolute Dimensionsanzahl anzugeben ist schwer. Ensembl ([108]) bietet für die 'Region View' etwa 300 optionale Anzeigen, wobei die meisten dieser Angaben sind absolute oder relative Häufigkeiten von Regionen oder Polynukleotiden sind.
Neben diesen Angaben kann für ein Chromosom auch die lokale Verteilung von Regionen als Input für eine Sonifikation verwendet werden. Eine Darstellung erfolgt meist ähnlich der in Abbildung 11.

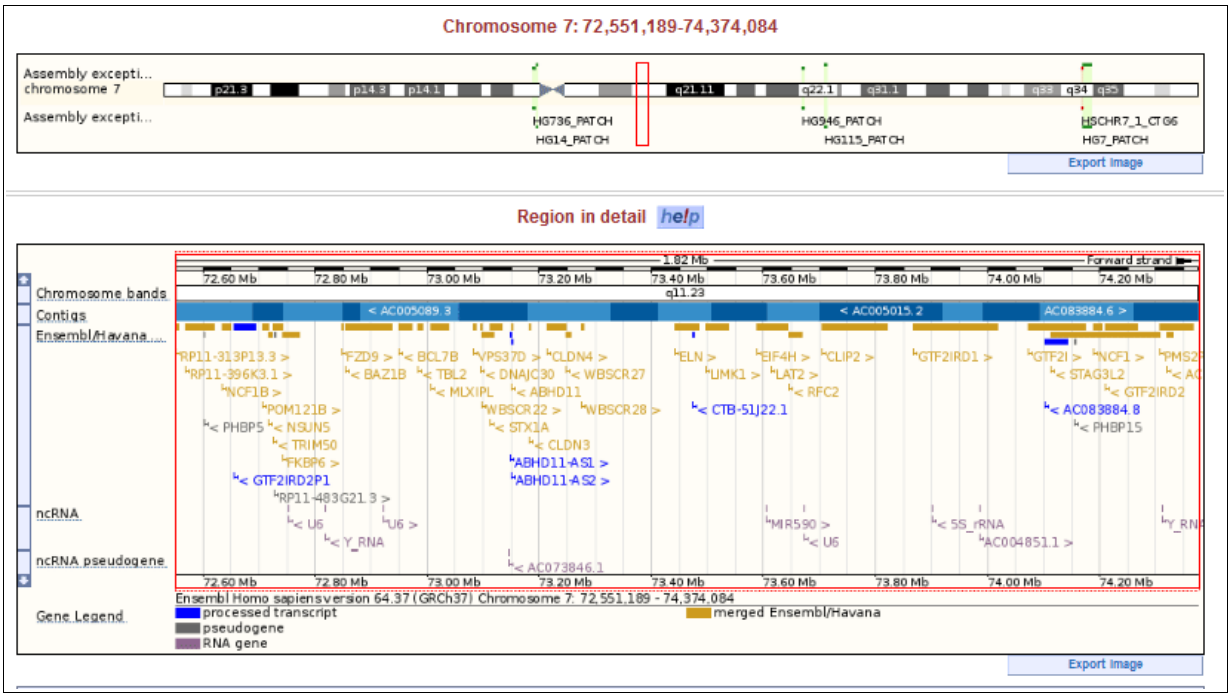

Abbildung 11: Screenshot des Ensembl-Browsers. Der gezeigte Abschnitt des Chromosoms 7 ist 1.822.895 Basen lang.

56 Lang meint hier 100.000 Basen und mehr. Gemessen am gesamten Chromosom ist man aber immer noch im Promille-Bereich.
57 Single-Nucleotide-Polymorphism

4.3 VORSCHLÄGE FÜR SONIFIKATIONEN

Im folgenden Kapitel werden eine Reihe von Möglichkeiten zur Sonifikation molekulargenetischer vorgestellt. Zweck, Input, Transformation und zu erwartender Klangcharakter werden beschrieben und mögliche Erweiterungen diskutiert. Ein Teil davon ist im Zuge des Projekts für Fr. Maïda praktisch implementiert worden, während andere Vorschläge rein theoretisch sind.

4.3.1 Sequenz-Sonifikation

Auf der Sonifikation einer DNA-Sequenz als eine Folge einzelner Noten oder Akkorde basieren bereits eine ganze Reihe von Arbeiten mit sowohl wissenschaftlichem als auch künstlerischem Hintergrund (vgl. 3.1). Allen Sonifikationen dieser Art ist gemein, dass sie in der Praxis nur für relativ kurze DNA-Abschnitte einsetzbar sind, da jede Base durch einen separaten Ton repräsentiert wird. Bei einer Geschwindigkeit von 240 Bpm bräuchte man für ein durchschnittliches Exon etwa 30 Sekunden. Als Argument für diese Art der Sonifikation wird unter anderem angeführt, dass eine Melodie eine dem Menschen wohlbekannte Form ist , was den Zugang und die Analyse erleichtert. Die resultierende Melodie (im Sinne von Tonfolge) ist jedoch an sich sehr monoton, da sie, abgesehen möglicherweise von einem integrierten Codon-Akzent, von sich aus keinerlei Variabilität in Lautstärke oder Timbre besitzt. Der Tonvorrat ist, abhängig vom Mapping, teils sehr eingeschränkt, und auch eine 'musikalische' Progression mit tonalem Zentrum und Spannungsbogen wird in den wenigsten Fällen das Ergebnis sein. Inwieweit jedoch eine derartige quasi-zufällige Tonfolge wirklich eine geeignete Darstellung ist, wurde bisher nicht untersucht. Eine essentielle Frage ist die nach der Fehlerrate in Abhängigkeit von Mapping, Trainingsstand, musikalischer Vorbildung und anderen Parametern. Um derartige Sonifikationen zum Vergleich beziehungsweise zum Alignment zu verwenden muss weiterhin untersucht werden, inwieweit nacheinander gehörte Melodien überhaupt zu unterscheiden sind, und ob eine gleichzeitige Darbietung hinreichend frei von psychoakustischen Artefakten.
Im Programm wurde eine Basis-Variante einer solchen Sequenz-Sonifikation implementiert. Jeder Base und auch jeder Aminosäure kann eine Tonhöhe innerhalb von fünf Oktaven zugewiesen werden. Das Tempo ist variabel einstellbar. Die Klangausgabe erfolgt mittels des General-MIDI-Synthesizers des Betriebssystems. Die zunächst angedachte Erweiterung um eine Art Melodie-Bank, in der sich verschiedene Melodien gleichzeitig oder gegeneinander verschoben abspielen lassen, wurde schlussendlich nicht implementiert.

4.3.2 Repeat-Finder

Spectral techniques have not, so far, been extensively used to detect repetitive sequences even though it would seem that they are ideally suited for detecting periodic patterns.
([124], p. 2)

Der Repeat-Finder basiert auf einer Audifikation von DNA-Sequenzen. Er dient dazu, sich

einen schnellen Überblick darüber zu verschaffen, ob in einem längeren Sequenzabschnitt Tandem Repeats (TR) enthalten sind. Dazu wird die hohe Empfindlichkeit des menschlichen Ohrs für tonale Anteile in einem verrauschten Signal genutzt. Tonalität setzt immer eine Periodizität vorraus. Genau das ist das Hauptcharakteristikum der TRs im Gegensatz zu anderen Sequenzabschnitten. Die Überlegungen gehen hier zunächst von idealen Repeats, also solchen ohne jegliche Abweichungen innerhalb des Patterns aus.
Die Audifikation ist denkbar einfach. Den vier Basen, die je einem Sample einer digitalen Audiodatei entsprechen, werden zunächst beliebig vier Quantisierungsstufen zugeordnet. Eine Sequenz von 100.000 Basen hätte bei einer Abtastrate von 44.1 kHz beispielsweise eine Länge von etwas über zwei Sekunden. Basensequenzen sind zwar nicht komplett zufällig, haben jedoch, zumal in nicht-codierenden Bereichen, chaotischen Charakter. Grundsätzlich wird eine Audifikation also Rauschcharakter haben, und periodische DNA-Abschnitte werden, so sie nicht von dem Rauschen verdeckt werden, als tonale Anteile wahrnehmbar sein. Je nach Repeat-Pattern und dem gewählten Mapping von Basen auf Sample-Wert ergibt sich für die Repeats ein anderes spezifisches Obertonspektrum. Die Grundfrequenz hängt immer von der Abtastfrequenz und der Repeat-Länge ab.

$$f_{Ton} = \frac{f_{Abtast}}{l_{Repeat}}$$

Ein Repeat der Periode 8 ergäbe bei einer Abtastrate von 10.000 Hz einen tonalen Anteil mit der Grundfrequenz 1250 Hz. Die für seine Wahrnehmbarkeit wichtige Dauer des tonalen Anteils ergibt sich aus der Abtastfrequenz und der Zahl der Wiederholungen oder aus Abtastfrequenz und absoluter Länge.

$$t_{repeat} = \frac{n}{f_{Ton}} = \frac{l_{absolut}}{f_{Abtast}}$$

Würde der Repeat aus obigem Beispiel 6 mal wiederholt, hätte er also eine Länge von insgesamt 48 Basen, ergäbe sich für das resultierende tonale Ereignis eine Dauer von 4,8 Ms. Ein Repeat mit drei Wiederholungen eines 20 Basen langen Patterns ergäbe eine Länge von 6 Ms und eine Frequenz von 500 Hz, 12 Wiederholungen dieses Patterns hätten eine Länge von 20 Ms. Auch wenn jedoch rein rechnerisch im Spektrum ein tonaler Anteil vorhanden ist, ist seine Wahrnehmbarkeit nicht automatisch gegeben, sondern hängt von seiner Dauer und seiner 'Reinheit' ab. Ein idealer Repeat hat keine Rauschkomponente. Seine Lautheit und seine Wahrnehmbarkeit hängt nur von der Signalenergie an sich ab, und davon, wie stark er eventuelle durch Nachverdeckung maskiert wird. Oft sind Repeats jedoch nicht ideal, sondern variieren von Wiederholung zu Wiederholung in Länge und Basenfolge (vgl. [124]). Diese Varianz äußert sich spektral durch einen Rauschanteil, der umso höher ausfällt, je inkonsistenter die einzelnen Repeats sind, und der die tonalen Komponenten schließlich verdecken wird. Mit dieser Methode werden sich also nur hinreichend lange und 'reine' repetitive Abschnitte erkennen lassen.
Im Zuge des Projekts für Frau Maïda wurde eine Basisversion des Repeat-Finders implementiert und anhand einer Testdatei erprobt. Die Datei enthält, eingebettet in zufällig gewählte Sequenzen des X-Chromosoms, sowohl künstliche erstellte ideale und nicht ideale

Repeats verschiedener Topologie, als auch einen Telomerabschnitt des Y-Chromosoms, der eindeutig repetitiven Charakter hat. Alle in der Datei enthaltenen Repeats waren für mich hörbar.[58] Allerdings waren mir die Anzahl und Lage bekannt, was sich positiv auf die Perzeption ausgewirkt haben dürfte.
Die vorgeschlagene Audifikation erlaubt das schnelle Sichten langer Sequenzen hinsichtlich auf Tandem Repeats. Die Bandbreite der theoretisch erkennbaren Repeats ergibt sich aus der Abtastfrequenz. Die Grundfrequenz der resultierenden tonalen Anteile gibt einen Hinweis auf die Länge des Patterns, ihre Dauer auf die Länge des gesamten Repeats. Mittels der den vier Basen beliebig zuweisbaren Quantisierungsstufen lässt sich sogar, wie am Beispiel des Telomers des Y-Chromosoms ein Stück weit die Art des Repeats erkennen.[59] Ein Muster **ctctctctctctct** ist beispielsweise nicht hörbar, wenn sowohl c als auch t eine Amplitude von 0 zugeordnet ist. Der Abschnitt ist dann Stille. Natürlich führen auch regelmäßige Nullwerte zu einer tonale Komponente. Man könnte also auch versuchen, Basen, die man ausblenden möchte, anstatt 0 einen zufälligen Samplewert zuzuweisen. Ein Repeat **gccctgccctgccctgcccg** wäre mit a = 1 und t=-1 ein Rauschen für g/c = [Zufall] und ein klangliches Ereignis für g = 1 und c = -1. Die Zusammensetzung des Repeats ließe sich so rein auditiv untersuchen. Andererseits erhöht eine derartige Zugabe von Rauschen natürlich die Gefahr der Verdeckung der tonalen Anteile.
Ein Kritikpunkt an dieser Audifikation ist sicher der Klangcharakter. Das sehr ungleichmäßige, breitbandige Rauschen ist extrem ermüdend für Ohren[60]. Ohne experimentelle Untersuchung des Repeat-Finders, am besten mit Fachleuten, kann jedoch kein abschließendes Urteil getroffen werden.

4.3.3 Transcript-Fingerprint

Der Transcript-Fingerprint ist eine Art parametrisiertes Earcon eines Transcripts, also einer Gen-Version, das, ähnlich dem RNA-Browser aus [76] (vgl. 3.1), als auditive Komponente von Datenbank-Interfaces verwendet werden kann. Es lassen sich einige Anforderungen für eine solche Sonifikation formulieren. Der Klang sollte sich natürlich aus den Datendimensionen des Transcripts ergeben, und verwandte Transcripts sollten einen ähnlichen Klang erzeugen. Die Länge des Klangs sollte nicht oder nur äußerst grob auf die Zeit abgebildet werden, da die möglichen Längen eine viel zu hohe Bandbreite haben. Die 'Abhörbarkeit' eines exakten Datenwertes oder exakter Relationen aus dem dargebotenen Klang ist hingegen nicht notwendigerweise erforderlich.
Als Hauptinput für die Sonifikation dient die ein Transcript eindeutig charakterisierende lokale Struktur, also die absolute oder relative Längenverteilung von Introns und Exons bezogen auf die erste Base.

---

58 Auf der beiliegenden CD-ROM befinden sich die Dateien 'Repeat-Testinput.txt' und 'Repeat-Testinput-dry.txt'. Letztere enthält die Repeats ohne die dazwischen liegenden DNA-Sequenzen. Die korrespondierenden Audiodateien sind 'Repeat-Testinput-Audio.wav' und 'Repeat-Testinput-dry-Audio.wav'.

59 Auf der beiliegenden CD-ROM befindet sich die Datei 'Telomer-Y.txt'. Diese wurde einmal mit einer Einstellung a = 1, t = -1, g = 0, c = 0 und einmal mit der Einstellung a,t,g = 0, c =1 audifiziert. Die entsprechenden Audiodateien sind 'Telomer-Y.wav' und 'Telomer-Y-Nur_C.wav'.

60 Ein Versuch mit einem Flanger-Effekt konnte die 'Nervigkeit' ein wenig mildern ohne klanglichen Phänomene zu beeinträchtigen.

Aufgrund der hohen Standardabweichungen der Längen von Introns und Exons (vgl. [109]) ist es eher unwahrscheinlich, dass zwei nicht verwandte Transcripts mit vier oder mehr Exons identische lokale Strukturen aufweisen. Ganz ausgeschlossen ist der Fall natürlich nicht, aber mehrheitlich sind ähnlichen lokalen Strukturen ein Indikator für Genverwandtschaft. Auf den Strukturdaten basierende Earcons werden also die oben formulierte Forderung nach Korrelation klanglicher und struktureller Ähnlichkeit erfüllen. Neben der lokalen Struktur als Hauptinput sollten jedoch optional auch weitere, vor allem sequenzbezogene Datendimensionen hinzufügbar sein.
Die Wertebereiche der Längen sind nicht unproblematisch, denn sie umfassen bis zu sechs Zehnerpotenzen (bei Introns) während der Median im Bereich von 10^2 liegt. Jegliche Abbildung, sei sie visuell oder akustisch, steht also immer vor der Entscheidung, wie sie skaliert. Das Display kann entweder die extremeren Werte clippen oder den Wertebereich auf die gegebene Auflösung des Displays[61] herunterquantifizieren. Für ein Display ohne Informationsverlust muss eine Zoom-Funktion integriert werden. Neben dem großen Wertebereich an sich ist auch das Längenverhältnis von Exons und den durchschnittlich zehnmal längeren Introns eine Schwierigkeit. Zwei Exons mit 100 und 120 Basen Länge, wären aus der Perspektive eines 2000 Basen langen Introns etwa gleichlang. Diese Beziehung gilt jedoch auch nicht immer, denn Transcripts mit längeren Exons als Introns sind nicht ausgeschlossen. Für eine eindeutige, präzise Darstellung sollten also Introns und Exons getrennt behandelt werden. Die im Folgenden skizzierte Sonifikation beschränkt sich zunächst auf die Exons.
Die Transformation ist ein einfaches Parameter-Mapping von Länge in Basen auf Frequenz. Die Frequenz ist gut geeignet, denn erstens sind die Wertebereiche von Exonlängen und wahrnehmbarer Frequenz, einige extreme Längen ausgeklammert, nahezu identisch, und zweitens erlaubt das Ohr im Frequenzbereich eine Auflösung von bis zu 620 Punkten. Eine der Inputdimension 'Länge' vielleicht gedanklich nähere Abbildung auf die Position im Stereobild hätte deutlich weniger Stufen (< 120 bei einer Abbildungsbreite von 120°), und auch die Intensität scheint ob der ebenfalls geringeren Auflösung und den Verdeckungsphänomenen weniger geeignet.
Etwa die Hälfte aller Exon ist kürzer als 130 Basen (vgl. 4.2.1), und ein direktes Länge-Frequenz-Mapping würde zu sehr tiefen, teils nicht wahrnehmbaren Frequenzen führen. Es ist also nötig, eine lineare Verschiebung nach oben vorzunehmen. Da die absoluten Werte nicht erkennbar sein müssen, ist eine solche Verschiebung auch problemlos möglich. Als tiefster 'sinnvoller' Ton soll hier die Sub-Oktave des Kammertons, also 220 Hz festgelegt werden[62]. Als Referenz könnte man den gewählten Basiston optional einblenden.
Exons mit einer Länge von über 10.000 werden mit 10.000 Hz sonifiziert. Um die Information über die tatsächliche Länge nicht komplett zu verlieren, könnte man statt eines durchgehenden Tons ein Folge kurzer Sinuspulse darbieten, deren Wiederholungsfrequenz von der tatsächlichen Länge abhängt. Alle Frequenzkomponenten sollten mit der gleichen Lautheit wahrgenommen werden, was eine A-Kompensation notwendig macht.
Ein Transcript aus fünf Exons führt nach der skizzierten Transformation zu einem Klang aus fünf Komponenten mit den Grundfrequenzen f = Länge + 220. Inwieweit obertonhaltige Klänge reinen Sinuskomponenten hinsichtlich der Deutlichkeit des Displays überlegen sind, müsste experimentell ermittelt werden. Weitere Datendimensionen, zum Beispiel der CG-Gehalt der Exons, könnten etwa mittels Amplitudenmodulation oder Timbrevariation der

61 Das können Graustufen, Pixel oder Frequenzen sein.
62 Die verhältnismäßig hohe Wahl gewährleistet ein Funktionieren des Displays auch unter schlechten Abhörbedingungen wie Laptoplautsprecher.

einzelnen Komponenten, anhand von Rauschanteilen oder mittels impulshafter Ereignisse in den akustischen Fingerabdruck integriert werden. Dabei muss jedoch darauf geachtet werden, den Klang nicht zu überfrachten.
Die Sonifikation, soweit hier umrissen, bietet Informationen über die Anzahl der Exons in einem Transcript und den Wertebereich der Längen. Ist ein extrem langes Exon beteiligt, ist das unmittelbar offensichtlich. Zwei Genversionen lassen sich anhand dieser Sonifikation bedingt vergleichen, denn ähnliche Strukturen führen zu einem ähnlichen Klang. Unterscheiden sich zwei Transcripts nur um ein Exon, ist dies unmittelbar zu erkennen. Die Sonifikation kann leicht um eine Zoomfunktion erweitert werden, indem man jede Länge mit einem Skalierungsfaktor multipliziert. Das ist insbesondere dann sinnvoll, wenn in einer Genversion viele Exons ähnlicher Länge vorkommen. Weiterhin ließen sich so auch Längenunterschiede erkennen, die ohne Skalierung unterhalb der Frequenzunterschiedsschwelle lägen.
Informationen über die Reihenfolge der Exons sind in den Klängen nicht mehr enthalten. Dieser Informationsverlust lässt sich vermeiden, indem man die einzelnen Klangkomponenten nicht gleichzeitig, sondern zeitlich versetzt in der Reihenfolge ihres Auftretens darbietet. Die Regelung der zeitlichen Verschiebung sollte dann vom User steuerbar sein. Durch den Einsatz unterschiedlicher Timbres in Kombination mit einer Verteilung im Stereopanorama wäre weiterhin die Darbietung mehrerer Transcripts zur gleichen Zeit möglich, was für Vergleichszwecke klar vorteilhaft wäre. Auch in diesem Fall wäre eine zeitversetzte Darbietung vorzuziehen.

Ausgehend von dem Prinzip der Abbildung von Länge auf Frequenz wäre auch eine Sonifikation denkbar, die sowohl Introns mit einbezieht als auch die Reihenfolge der Exons erhält. Die Gesamtlänge des Transcripts könnte auf einen Frequenzbereich abgebildet werden. Dabei sollte entsprechend der Mel-Skala transformiert werden. Das Grundsignal ist ein rosa Rauschen. Mittels Filterung werden Exon-Bereiche amplifiziert und Intron-Bereiche gedämpft. Aufgrund der beschriebenen Längenverhältnisse zwischen Introns und Exons wäre die Auflösung der Exons natürlich sehr grob, und ein Exon mit 50 Basen und eins mit 100 würden bei dieser Transformation perzeptiv sehr ähnliche spektrale Anteile erzeugen. Eine Abgrenzung von Bereichen mit vielen kurzen Exons gegenüber einem einzigen langen Exon könnte beispielsweise mittels einer Rauhigkeit im entsprechenden Frequenzbereich erreicht werden.

Beide vorgeschlagenen Sonifikationen basieren auf einem Mapping von Ort auf Frequenz. Die Synthese erfolgt einmal additiv mittels Klanggeneratoren und einmal substraktiv durch Filterung. Das Ergebnis sind Vielklänge oder Arpeggien, in denen die Ortsstruktur des jeweiligen Transcripts enthalten ist, ohne dass sie rein auditiv exakt zu deduzieren wäre. Diese parametrisierten Earcons können eine visuelle Datenbankschnittstelle unterstütze. Für Geräte mit sehr kleinen Displays sowie für sehbehinderte User könnten sie auch der Hauptzugang zu den Datenbanken und ein nützliches Tool zur Ähnlichkeitsabschätzung von Transcripts sein.

Im Zuge der Arbeiten am Programm für Fr. Maïda wurde eine einfache Version der additiven Methode implementiert. Optionale Einstellungen oder Zoom sind nicht möglich. Auch eine ist A-Kompensation und ein akzeptables GUI sind integriert.

4.3.4 Chromosom-Fingerprint

It is well known that human chromosomes are very different among themselves [Venter et al., 2001]. Putting aside the obvious differences in size, there are also divergences in the density and spatial location of genes, the types of genes, organization of Alu repeats [Grover et al., 2003] and the distribution of CpG islands [Chen et al., 2002].
([109], von Webseite)

Dem Zitat zufolge ist ein Chromosom unter anderem durch die Ausprägungen der im Zitat genannten Datendimensionen charakterisiert und also auch identifizierbar. Eine Sonifikation mit ebenjenen Daten als Input sollte ein für das jeweilige Chromosom charakteristisches Ergebnis produzieren. Die Daten könnten also zur Erstellung eines klanglichen Fingerabdrucks des Chromosoms, ähnlich dem der Transcripts (vgl. 4.3.3), verwendet werden.
Die hier vorgeschlagenen Sonifikation wurde von Fr. Maïda im Zuge ihrer kompositorischen Arbeit entwickelt. Als Input dienen die Häufigkeiten der Längen von Introns, Exons und 'Nicht-Gen-Abschnitten'. Die entsprechenden Histogramme werden als Spektrum uminterpretiert, sprich, die Länge in Basenpaaren wird auf eine Frequenz abgebildet und die Häufigkeit, mit der eine Länge im Chromosom auftritt, auf die Intensität. Je mehr Regionen mit einer bestimmten Länge in einem Chromosom vorhanden sind, desto größer ist die Amplitude dieser Frequenzkomponente.
Für jeden Regionstyp wird zunächst ein Einzelspektrum erstellt. Die Summe aller Einzelspektren ergibt das Gesamtspektrum für das Chromosom. Dieses Vorgehen erlaubt eine Gewichtung der Einzelspektren, um etwa große Unterschiede in der Maximalamplitude auszugleichen oder bestimmte Regionstypen durch Amplifikation zu betonen. Viele Exons haben eine Länge von unter 100 Basen (vgl. 4.2.1), das entsprechende Spektrum wird also eine hohe Energie im sehr tiefen Frequenzbereich haben. Eine generelle Verschiebung nach oben um 100 oder 200 Hz wäre also durchaus sinnvoll. Auch eine A-Kompensation sollte integriert werden.
Bei verhältnismäßig Gen-armen Chromosomen wird das Ergebnis der Sonifikation ein Klang aus vielen Sinuskomponenten, ohne einen festen Grundton. Ein Gen-reiches Chromosom wird zu einer rauschhaften Sonifikation führen, die im Bereich um 200 Hz einen hohen Energieanteil hat.
Eine wichtige Frage ist, inwieweit sich derartige Spektren abgesehen von den durch den Genanteil bedingten Differenzen überhaupt unterscheiden. Michael Deutschs und Manyuan Longs Beitrag zur Intron-Exon-Struktur von Eukaryoten ([111]) ermöglicht eine etwas belastbarere Abschätzung als oben angeführtes Zitat. Die Autoren vergleichen verschiedene Spezies, darunter Mensch, Maus, Huhn und Kresse anhand der Längenstatistiken für Introns und Exons. An ihren Histogrammen und Angaben lässt sich zweierlei feststellen. Zum Ersten liegt die deutliche Mehrheit aller auftretenden Längen für alle untersuchten Spezies betragsmäßig in einem für den Menschen hörbaren Frequenzbereich, und zum Zweitens weisen die Histogramme insbesondere der Introns teils deutliche Unterschiede auf. Unterschiedliche Spezies werden also unterschiedliche Spektren haben. Die Spektren der Histogramme aus Abbildung 12 werden auditiv klar zu unterscheiden sein.

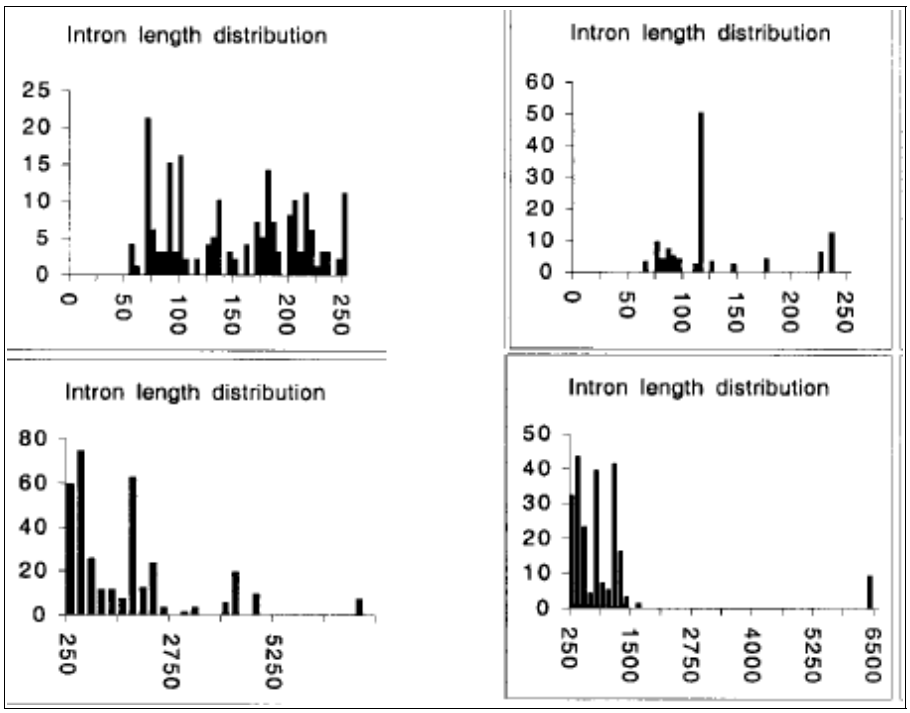

Abbildung 12: Längenverteilung der Introns von Ratte (links) und Huhn (rechts). Die Autoren haben die Darstellungen für eine Spezies sind in zwei Diagramme unterteilt, um entlang aller Längen eine gute visuelle Auflösung der Häufigkeit zu gewährleisten.(Quelle: [111]), p. 4;die hier abgebildeten Teile wurden aus Platzgründen aus verschiedenen Abbildungen der Originale extrahiert und zusammengesetzt. An den einzelnen Graphen selbst wurden keine Manipulationen vorgenommen)

Inwieweit dies auch für verschiedene Chromosomen derselben Spezies oder für gleiche Chromosomen unterschiedlicher Individuen der Fall ist, wäre eine interessante Fragestellung. Selbst wenn sich die Gesamtspektren mehrheitlich klanglich sehr ähnlich sein sollten, so sind in Teilbändern Unterschiede zu erwarten. Vielleicht ist es, eine Lernphase vorausgesetzt, sogar möglich, Chromosomen anhand eines solchen Spektrums zu erkennen oder Verwandtschaftsbeziehungen zu deduzieren.

Im Rahmen der Arbeit für Fr. Maïda wurde eine einfache Version des vorgeschlagenen Fingerabdrucks implementiert (vgl. 5.3.2). Das Tool verwendet die Dateien von Fr. Maïda, in denen alle Regionen[63] mitsamt ihrer Länge aufgeführt sind. Diese Dateien sind jedoch nicht allgemein zugänglich (vgl. 5.2.1). Mir selbst lagen bei der Niederschrift der Arbeit lediglich die Dateien der Chromosom X und Y vor. Beide Spektren[64] sind sehr unterschiedlich, was jedoch vor allem in der relativen Genarmut des Y-Chromosoms begründet liegt. Daher sind mir weitere Aussagen hinsichtlich der Differenzierbarkeit verschiedener Chromosomen nicht möglich.

63 Introns, CDSs und 'Intergenic Material'.

64 Die Audiodateien befinden sich auf der beiliegenden CD-ROM (Fingerabdruck-Chromosom-X.wav und Fingerabdruck-Chromosom-Y.wav). Die zur Sonifikation herangezogenen Regionen sind Introns und Exons ohne Gewichtung.

4.3.5 Akustischer DNA-Browser

Die Zielstellung des akustischen DNA-Browsers ist eine akustische Unterstützung der schematischen Anzeige ganzer Chromosomen und längerer Abschnitte von DNA (vgl. 4.2.3). Die Grundidee ähnelt somit dem bereits in 3.1 vorgestellten RNA-Browser ([76]).

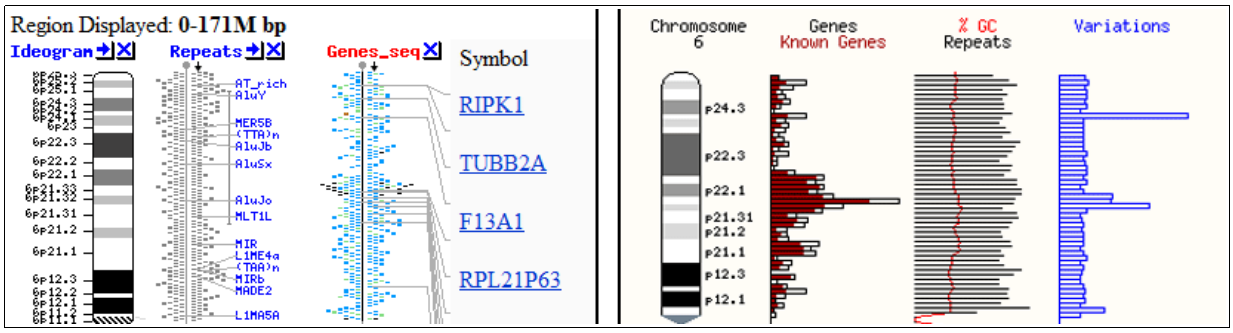

Abbildung 13: Visualisierung eines Teils des menschlichen Chromosoms 6; links: Screenshot des NCBI-Map Viewers ([125]), rechts Screenshot von Ensembl ([108])

Abbildung 13 zeigt Ausschnitte der schematischen Darstellung eines ganzen Chromosoms und Abbildung 11 ist die Darstellung eines längeren Abschnitts. Wie in 4.2.3 beschrieben, gibt es angefangen von statistischen Daten der Sequenz bis hin zu Auswertungen von Gen- und Repeat-Dichten eine Vielzahl von verfügbaren Informationen über einen längeren Abschnitt von DNA. Auch wenn jede dieser Informationen für sich genommen gut visuell darstellt werden kann, ist doch der verfügbare Raum und damit die Anzahl der darstellbaren Dimensionen durch die Größe des Bildschirms begrenzt. Gerade bei kleinen Geräte wie Netbooks oder Smart-Phones kann diese Einschränkung ein produktives Arbeiten erschweren. Weiterhin ist das Auge in seiner Aufmerksamkeit limitiert. Zwanzig Kurvendiagramme auf einem Display erfordern einiges an Konzentration. Es scheint also durchaus sinnvoll, einen Teil der Daten akustisch darzustellen und so den visuellen Kanal zu entlasten. Eine absolute Darstellung von Zahlenwerten ist dabei glücklicherweise mehrheitlich nicht notwendig[65], denn auch die visuelle Darstellung zeigt in erster Linie relative Verhältnisse (vgl. Abbildung 13). Es ist zwar erkennbar, welche Abschnitte einen hohen Genanteil haben und welche nicht, die genaue Größe jedoch nicht.

Der erste Schritt für akustische Erweiterung der visuellen Darstellungen ist die Definition einer Fensterbreite in Basen. Dieses Fenster legt den Abschnitt der DNA-Sequenz fest, dessen Parameter ausgewertet und sonifiziert werden sollen. Diese Fensterung ist äquivalent zu einer Darstellung mittels Balken wie im rechten Teil von Abbildung 13. Die ermittelten Parameter werden dann mittels Parameter-Mapping sonifiziert.

Die Steuerung erfolgt über ein interaktives, visuelles Display. Mittels Maus oder Keyboard kann der User das Sonifikationsfenster entlang der visuellen Darstellung verschieben und so in verschiedene Sequenzabschnitte 'reinhören'. Zudem gibt es die Möglichkeit, mittels einer Abspielfunktion die gesamte Sequenz von links nach rechts zu durchlaufen. Dabei können entweder diskrete Klänge für jedes Fenster erstellt werden, oder die Sonifikation erfolgt,

65 Wie in 2.2.4 erläutert, stellt eine Sonifikation, die absolute Zahlenwerte darstellen soll, eine große Herrausforderung dar.

insofern die Rechenkapazität es zulässt, kontinuierlich.
Die Konfiguration des gesamten Display, also der visuellen, wie auch der akustischen Komponente erfolgt über ein Menü, in dem der User festlegt, welche Informationen in welcher Modalität dargestellt werden sollen. Für die akustische Darstellung steht eine Bibliothek von Audioströmen zur Verfügung, die der User den Datendimensionen zuordnen kann. Jeder Audiostrom verfügt dabei über mindestens einen durch Daten steuerbaren Parameter.
Für eine sinnvolle Sonifikation ist es entscheidend, dass die kombinierten Audioströme möglichst perzeptiv getrennt sind, sich also gegenseitig hinsichtlich der Wahrnehmung nicht oder nur wenig beeinflussen. Die folgende Tabelle bietet ein Beispiel, das sich an den Elementen der Sonifikation in [38] orientiert, da die dort gewählten Audioströme, soweit keine extremen Lautstärken auftreten, gut getrennt sind.

Audiostrom	**steuerbarer Parameter**	**Inputvariable**
Breitbandrauschen	Lautheit	Entropie
tieffrequenter Klang	Tonhöhe und Energie in Obertönen	Anzahl der protein-codierenden Gene
hochfrequenter Ton / Klang	Lautheit und Rauhigkeit	Zahl der RNA-Gene
tieffrequentes Klopfen	Wiederholungsfrequenz	Zahl der ALU-Sequenzen
hochfrequentes Klicken	Wiederholungsfrequenz,	Zahl der Tandem Repeats

Tabelle 2: Mögliche Daten-Audiostrom-Konfiguration

Bei einer praktischen Umsetzung dieses hier nur theoretisch entwickelten akustischen Displays wird man zwei große Probleme lösen müssen. Zum Ersten ist nicht jeder User mit psychoakustischen Wechselwirkungen vertraut. Es besteht daher die Gefahr, dass Konfigurationen gewählt werden, in denen die Audioströme perzeptiv stark interagieren, wodurch die Qualität des Displays leidet. Um diesem Problem vorzubeugen, könnte man vorgefertigte Konfigurationen von Audioströmen wie in Tabelle 2 anbieten. Auch eine einfache Konfigurationskontrolle durch den Computer ist denkbar. Jeder Audiostrom der Bibliothek könnte bezüglich verschiedener Variablen wie Tonhöhe, Frequenzbandbreite, Rauschanteil, Impulshaftigkeit und anderer Maße quantifiziert werden. Bei einer Kombination zweier Ströme wird anhand dieser Einordnung dann verglichen, inwieweit die Ströme klanglich übereinstimmen. Gibt es zu viele Ähnlichkeiten, gibt der Computer eine Warnung aus.
Das Hauptproblem aber, das für eine praktische Umsetzung zu lösen ist, ist die sehr variable Zeit, die zur Berechnung der Sounds nötig ist. Sie hängt, abgesehen natürlich vom Rechner selbst, zunächst von den gewählten Informationen und der Fensterbreite ab. Sollen für eine Sequenz der Länge 100.000 die Häufigkeiten aller Basen, die Shannon-Entropie sowie die Autokorrelation der Sequenz akustisch dargestellt werden, so ist vermutlich etwas Zeit für die

Berechnung dieser Inputvariablen notwendig. Die akustische Darstellung von Gendichte und GC-Gehalt derselben Sequenz dürfte hingegen für moderne Rechner, eine adäquate Datenbankstruktur vorrausgesetzt, keine Herausforderung sein. Ebenfalls entscheidend für die Performance des vorgeschlagenen Tools ist, wie schnell der Zugriff auf die Sequenzdaten erfolgen kann. Erfolgt dieser über das Internet, so ist eine Echtzeitsonifikation von beliebigen Fenstern kaum mehr möglich. Hier müssten für eine feste Fensterbreite und feste Fensterpositionen alle darstellbaren Variablen vorab berechnet und in einer Datenbank abgelegt werden, sodass über das Internet lediglich die Zahlenwerte, nicht aber die Sequenz selbst transferiert werden muss. Der User, der nach wie vor mittels der Maus über der schematischen Darstellung navigiert, würde dann jeweils das Fenster hören, über dem sich der Cursor gerade befindet. Bei hinreichend feiner Auflösung und überlappenden Fenstern könnte auch web-basiert eine kontinuierlich wirkende Sonifikation erreicht werden, bedürfte allerdings einigen Aufwands seitens der Datenbankbetreiber.
Trotz dieser praktischen Einschränkungen ist eine Erweiterung visueller Sequenz-Browser um eine akustische Komponente vielversprechend, denn es könnten mehr Informationen über den DNA-Abschnitt gleichzeitig dargestellt werden, die Nutzung auf Geräten mit kleinem Display würde komfortabler und die Accessibility der Browser würde deutlich erhöht.

4.3.6 Data Mining mittels MBS

Die Model Based Sonification wurde in 2.2.3 als Ansatz zum Data Mining mittels Sonifikation vorgestellt. Die beiden angeführten Beispiele hatten gemeinsam, dass den Datenpunkten abhängig von ihrem Datenvektor eine bestimmte Position in einem n-dimensionalen Raum zugeordnet wurde. Diese Modelle sind prinzipiell immer dann einsetzbar, wenn die 'räumliche' Struktur eines Datensatzes von Interesse ist. Ein Beispiel für eine solche Analyse im Bereich der Molekulargenetik stammt von Patrick Hoffman et al. ([126]). Die Autoren verteilen ihre Datenpunkte, hier Introns und Exons, in einem zweidimensionalen Raum. Für jede Sequenz ermitteln sie zunächst die Häufigkeiten aller vier Basen auf jeder der drei Positionen eines ORF[66] und erstellen so für jede Region einen Datenvektor mit zwölf Elementen. Das Prinzip ihrer Visualisierung beschreiben die Autoren wie folgt:

> *We developed a radial visualization(Radviz)[sic!], similar in spirit to parallel coordinates (lossless visualization), in which n-dimensional data points are laid out as points equally spaced around the perimeter of a circle. One end of n springs are attached to these n perimeter points. The other ends of the springs are attached to a data point. The spring constant Ki equals the values of the i-th coordinate of the fixed point. Each data point is then displayed where the sum of the spring forces equals 0. ([126], von Webseite)*

Introns liegen in dieser Visualisierung näher um die Kreismitte herum, die Exons sind stärker verteilt. s gibt in codierendem Material also für einzelne Basen and bestimmten Positionen höhere Wahrscheinlichkeiten als in nicht-codierender DNA, eine Tatsache, die mit dem bereits im Zusammenhang mit der 3'-Periodizität angesprochenen Codon-Bias zu erklären ist (4.2.1).

66 Open Reading Frame – Legt fest welche Base die erste eines Codons ist.

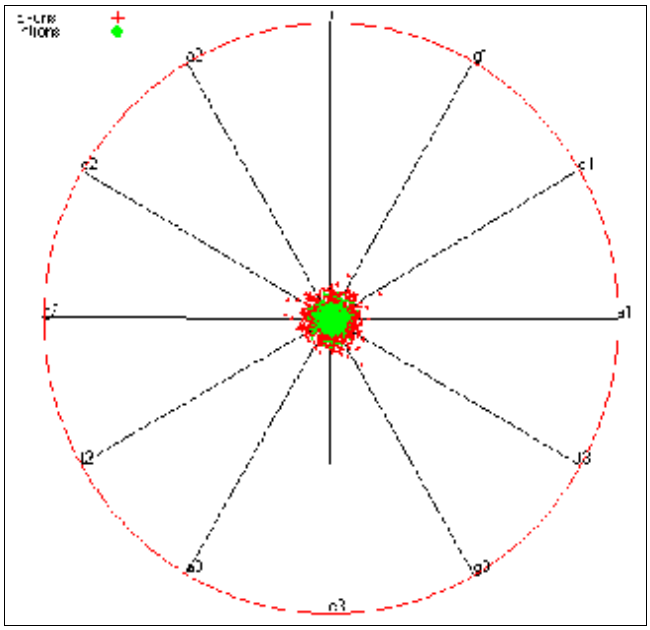

Abbildung 14: Radial Visualization von Exons (rot) und Introns (grün); (Quelle: [126], von Webseite)

Die Radial Visualization ließe sich direkt in die Audiodomäne übertragen. Die Anregung erfolgt im Kreismittelpunkt, und die Datenpunkte antworten abhängig von ihrem Typ mit einem hohen oder tiefen kurzen Klang, wenn die von der virtuellen Welle erreicht werden. Die Antworten würden aufgrund des durchschnittlich höheren Abstands auch später beim Rezeptor, der ebenfalls im Kreismittelpunkt sitzt, eintreffen. Die Autoren wenden das Visualisierungsprinzip auf Datenvektoren höherer Dimension an, indem sie entsprechend mehr Punkte auf dem Umkreis verteilen. Die Visualisierung ist allerdings immer auf zwei Dimensionen beschränkt. Mittels Sonifikation nach dem Prinzip des 'Data Sonogramms' als auch der 'Particle Trajectory' hat diesen Nachteil nicht, könnte also weitere Verteilungsphänomene zutage fördern.
Das angeführte Beispiel soll illustrieren, dass es auch in der Genetik Anwendungsfelder für Model Based Sonification gibt. Sie liegen vor allem bei vergleichenden Analysen, wenn beispielsweise Gemeinsamkeiten für ein bestimmtes Gen verschiedener Spezies gesucht werden.

5. Programm

Der ursprüngliche Anstoß zu dieser Arbeit war ein Projekt zur Sonifikation von DNA-Sequenzen der französischen Komponistin Clara Maïda ([127]). Sie war auf der Suche nach einer Schnittstelle zwischen der reinen Molekulargenetik und der Musik und benötigte ein Programm, um die große Menge an Daten zu sichten, hinsichtlich interessanter Phänomene zu untersuchen und schließlich zu sonifizieren. Dieses Programm ist Thema dieses Teils der Arbeit.
Nach einer Reflexion der Vorüberlegungen, die sich auf die entgültige Form des Programms ausgewirkt haben, werden die Input-Daten vorgestellt. Da diese nicht unproblematisch waren und im Verlauf der Arbeit an dem Programm einige Schwierigkeiten bereiteten sowie zu wissenschaftlich falschen Darstellungen führen, nimmt der entsprechende Abschnitt einigen Raum ein. Es folgt eine Übersicht über die Funktionalität des Programm zur Visualisierung und zur Sonifikation, gefolgt von einer knappen Beschreibung der Implementation selbst.

5.1 VORÜBERLEGUNGEN

5.1.1 Warum ein neues Programm

Bevor man die Arbeit an einem neuen Programm beginnt, sollte man zunächst überprüfen, inwieweit bestehende Programme den eigenen Anforderungskatalog erfüllen könnten, und was genau ein neues Programm notwendig macht. Diese Überlegungen sollen im Folgenden kurz dargelegt werden.
Gen-Browser und Programme zur DNA-Visualisierung gibt es viele. Eine Gruppe bilden die webbasierten Tools wie der 'UCSC Genome Browser' der Universität Santa Cruz ([128]), Ensembl ([108]) eine Zusammenarbeit von EBMI (European Bioinformatics Institut) und dem Welcome Trust Sanger Institute, und der 'Map Viewer' des NCBI (National Center for Biotechnology Information) [125]. Alle drei greifen auf die großen Online-Datenbanken (GenBank, DDBJ, EBML) genetischer Daten zu und bieten visuelle Darstellungen von Sequenzen, Proteinen und einer Vielzahl an Annotationen. Die große Informationsfülle, die über diese Browser erschlossen werden kann, ist im wissenschaftlichen Betrieb unabdingbar, einen Laien wird sie jedoch eher überfordern. Ensembl beispielsweise bietet im Konfigurationsdialog allein für das 'Main Panel' weit über 100 Anzeigeoptionen. Fr. Maïda hatte ihren Fokus von vornherein auf einige wenige Features beschränkt. Die Optionsfülle der wissenschaftlichen Browser ist da eher hinderlich. Ein weiterer Punkt, der die webbasierten Browser für Fr. Maïdas Arbeit weniger geeignet scheinen ließ, ist die Navigation. Zwar bieten alle Browser die Möglichkeit zu zoomen und entlang eines Strangs zu navigieren; das allerdings erfolgt, zweifelsohne bedingt durch die Tatsache, dass die Systeme webbasiert sind, weder in Echtzeit noch stufenlos. Jeder Navigations-befehl wird bei Ensembl zunächst in

einem kleinen Popup-Fenster ausgewählt und bestätigt. Anschließend wird eine neue Darstellung geladen, wobei die Anzeige während des Ladens auch bei DSL-Verbindungen kurz völlig verschwindet. Ein Chromosom auf der Suche nach 'Auffälligkeiten', die künstlerisch inspirierend sein könnten, zu überfliegen, ist praktisch unmöglich, ist aber nebenbei gesagt auch nicht Ziel dieser Systeme[67].

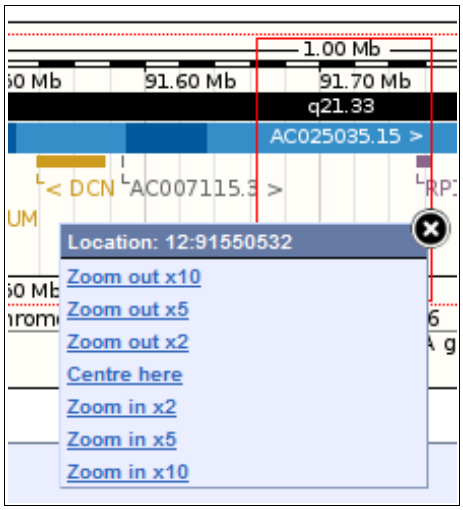

Abbildung 15: Navigationsdialog von Ensembl (Quelle: Screenshot , [108])

Neben den webbasierten Applikationen gibt es auch Tools, die auf dem User-Computer installiert werden. Beispiele sind 'Sockeye' ([129]) oder das 2010 vorgestellte 'MochiView' ([130]). Da die dargestellten Sequenzen und die Annotationen bei diesen Systemen im Arbeitsspeicher oder auf der Festplatte liegen, ist hier Echtzeitnavigation möglich. 'MochiView' hat zudem den Anspruch, ein visuell ansprechendes und benutzerfreundliches Programm zu sein.

> *We present MochiView, a platform-independent Java software that integrates browsing of genomic sequences, [...] in a visually-appealing and user-friendly application. ([130], p 1)*

In der Tat bietet das Tool hohe Funktionalität und hat eine ansprechende Optik. Die Navigation funktioniert im Demonstrationsvideo stufenlos und ohne Ruckeln. Das Programm arbeitet mit einer Datenbank auf der Festplatte, die es weitestgehend selbst organisiert und auf der Webseite zum Programm finden sich Links zu Seiten, von denen ein Basisdatenbestand heruntergeladen werden kann. Trotzdem schien ein neues Programm schlussendlich der bessere Weg.
Frau Maïda hatte in manchen Belangen schon frühzeitig sehr konkrete Vorstellungen von der Funktionalität des Programms. Die Auftretenswahrscheinlichkeiten von Basen und Triplets und die Längen von Regionen spielten für sie von Beginn an eine sehr zentrale Rolle. Besonders letzteres steht bei anderen Tools weitab vom Fokus. Weiterhin hatte sie sich hinsichtlich des Inputs auf Dateien festgelegt, die sie von einem befreundeten Bioinformatiker aus Straßburg erhält[68]. Diese sind auf ihre Vorstellungen und Interessen hin zugeschnitten und beinhalten nur die Informationen, mit denen Fr. Maïda arbeiten wollte. Annotationen, Gen-Bezeichungen oder Versionsinformationen fehlen hingegen. Diese Dateien genügen allerdings keinem anerkannten Standard, gängige Browser-Tools können sie folglich nicht lesen.
Auch hinsichtlich der Visualisierung standen die Grundlagen zu Beginn schon fest. Eine Grafik sollte beispielsweise die Regionen als gleichbreite Balken unterschiedlicher Höhe nebeneinander stellen. Diese Form der Darstellung findet sich so in keiner anderen Applikation. Die Länge eines DNA-Abschnitts wird immer in der gleichen Dimension, zumeist die X-Achse, wie deren Abfolge abgebildet. Die Darstellung, die Frau Maïda für die Sequenzen ersonnen hat, weicht zwar nicht wesentlich von der gängigen Darstellung mittels farbiger Blöcke respektive Buchstaben ab, jedoch hatte sie insbesondere hinsichtlich der Nummerierung der Basen einige spezielle Wünsche, die andere Tools in der Form nicht bieten.

67 Dennoch sei an dieser Stelle erlaubt, auf die wenig ansprechende Optik insbesondere des 'MapViewers' vom NCBI hinzuweisen.
68 Nähere Information zu diesem Kontakt habe ich nicht erhalten.

Der zweite große Aspekt des Programms schließlich, die Sonifikation, ist in keinem der verfügbaren Tools auch nur ansatzweise integriert.
Vor der Wahl, ein bestehendes Programm wie 'MochiView' zu erweitern oder eine komplett neue Applikation zu erstellen, erschien das Zweite als die bessere Lösung. Ein Aufbauen auf bestehender Software würde sehr viele, nicht unwesentliche Modifikation am Basisprogramm erfordern. Das birgt aber die Gefahr, im Verlaufe der Arbeit auf architektonische Besonderheiten des Basisprogramms zu stoßen, die sich nur schwer mit dem gewünschten vereinbaren lassen. Auch zeitlich ist eine Neuimplementierung nicht unbedingt von Nachteil, da der zu erweiternde Code auch zunächst analysiert und verstanden werden muss.
Schließlich wären bei einer Erweiterung eines bestehenden Programms deutlich mehr Konzessionen und Kompromisse seitens Frau Maïda nötig gewesen, denn ein wirklich komplett nach ihren Vorgaben funktionierendes Programm ließe sich durch Modifikation vorhandener Implementierungen wohl nicht erstellen.

5.1.2 Weitere Vorüberlegungen

Eine wichtige Frage ist die nach der Programmiersprache. Nach einer kurzen Überlegung in Richtung LISP fiel die Wahl schnell auf JAVA. Es gibt mehrere Gründe für diese Wahl. Der erste ist die Tatsache, dass Fr. Maïdas Kontakt die Input-Dateien mittels JAVA erstellt, und ihr schien es eine gute Wahl, dasselbe Programm zu verwenden. Ein zweites Argument für JAVA ist die Plattformunabhängigkeit. Ich arbeite mit einem Windows-Rechner, während Fr. Maïda einen Mac verwendet. Das Programm sollte mit so wenig Aufwand wie möglich auf beiden Systemen lauffähig sein. Ein weiterer Vorteil von JAVA ist die große Anzahl an bestehenden Klassenbibliotheken, die für viele Anwendungen grundlegende Funktionalität zur Verfügung stellen. Eine spezialisiertere Sprache wie OpenMusic böte sicher vielseitigere Möglichkeiten im Bereich der Klangsynthese, jedoch wäre die Erstellung der sehr individuell geschneiderten Grafiken sicher komplizierter, wenn nicht unmöglich. Inwieweit sich die Entscheidung für JAVA als praktikabel erwiesen hat, wird Teil der Diskussion sein.
Ebenfalls Gegenstand von Reflexion war die Zielgruppe, also die Frage danach, wer das Programm schlussendlich verwenden wird. Diese Frage hat naturgemäß Einfluss auf die Funktionalität und vor allem auf das User-Interface. Grundsätzlich könnte ein Tool zur Visualisierung und vor allem zur Sonifikation genetischer Daten für eine breitere Zielgruppe interessant sein, sei es für Wissenschaftler, die in diesem Bereich arbeiten oder für Komponisten, die genetische Daten als Teil ihres Schöpfungsprozesses verwenden möchten. Es gab allerdings ein Faktum, dass die Zielgruppe des erstellten Programms von vornherein auf Fr. Maïda beschränkt - den Input. Die Dateien, mit denen das Programm arbeitet, sind wie schon angesprochen extra für Fr. Maïda kompiliert worden und nur ihr zugänglich. Sie organisieren die aus den öffentlichen Datenbanken extrahierten Informationen in einer Art und Weise, die abseits Fr. Maïdas Arbeit von wenig wissenschaftlichem Wert ist und weichen hinsichtlich Format von den gängigen Dateistrukturen in diesem Gebiet so stark ab, dass ein auf sie zugeschnittenes Programm mit anderen Formaten nichts anfangen kann. Ein Verzicht auf diese Dateien kam aber nicht in Betracht, da die Extraktion der enthaltenen Informationen aus den Datenbanken einigen Aufwand erfordert. Die Alternative einer doppelten Implementation schließlich, einmal zugeschnitten auf Fr. Maïda und ihre Dateien und einmal auf die gängigen Formate, hätte den Umfang der Arbeit bei weitem gesprengt. Das Programm

ist daher zu nahezu 100 Prozent Fr. Maïdas Vorstellungen gemäß entwickelt worden.
Die Tatsache, dass Fr. Maïda alleiniger Nutzer des Programms sein würde, hatte direkte Auswirkungen auf das Design. Da sie ein programmiertechnischer Laie ist, musste das User-Interface intuitiv und möglichst frei von Kommandozeilen sein. Andererseits konnte auf eine Hilfe-Funktion und auf Erklärungen zu den einzelnen Buttons in Form von etwa Popup-Fenstern verzichtet werden, da Fr. Maïda die Funktionalitäten bekannt sein würden.
Eine weitere Prämisse war Echtzeitfähigkeit. Das Tool sollte in jedem Fall eine ruckelfreie und unmittelbare Visualisierung auch bei Navigation durch große Datenmengen bieten. Zu Beginn der Implementierung wurden hierfür noch Vorkehrungen wie Grafik-Puffer verwendet, die aber im weiteren Verlauf zu Gunsten selektiver Zeichenroutinen entfallen konnten.

5.2 INPUT

Dieses Kapitel wirft einen detaillierten Blick auf die Input-Dateien. Deren Struktur ist alles andere als optimal. Sie bereitete während der Implementierung immer wieder Probleme und erforderte einen nicht unbeträchtlichen Teil an Quellcode, der mit der eigentlichen Aufgabe des Programms nichts zu tun hat.

5.2.1 Beschreibung der Dateien

Diese Dateien sind, abhängig davon, welches Format Fr. Maïdas MacBook ihr beim Download vorgeschlagen hat, entweder im Format '.txt' oder '.rtf'. Es gibt vier verschiedene Typen. Für jedes der menschlichen Chromosomen liegt von jedem Typ eine Datei vor. Die Typen lassen sich entweder inhaltlich oder der Form nach zu zwei Gruppen zusammenfassen. In den folgenden Erläuterungen verwende ich die Namenssystematik, mit der ich die Daten erhalten habe. In allen Beispielen handelt es sich um das X-Chromosom.

CHRX-Frequ.rtf (W-General)

In dieser Datei sind alle Regionen des direkten Strangs des jeweiligen Chromosoms in zweizeiligen Einheiten aufgelistet. In der ersten Zeile einer Einheit findet sich jeweils der Regionstyp, die absolute Position der ersten und der letzten Base auf dem Strang sowie die Länge der Region. In der zweiten Zeile stehen die Auftretenswahrscheinlichkeiten der vier Basen mit einer Genauigkeit von drei Nachkommastellen.
In manchen Regionen, insbesondere in Introns befinden sich Basen, die bisher nicht eindeutig identifiziert wurden. In den entsprechenden Sequenzdateien steht an diesen Stellen ein *n*. Es handelt sich teilweise um längere Abschnitte. Die Summe der Wahrscheinlichkeiten der vier Basen in dieser Region ist dennoch nicht kleiner als 1. Die Wahrscheinlichkeiten beziehen sich

also auf die Gesamtzahl der bekannten Basen und nicht auf die Länge der Region an sich[69].
Befinden sich in einer Region nur unbekannte Basen finden sich in Zeile der Wahrscheinlichkeiten vier Fragezeichen.
Es werden nur drei Typen von Regionen unterschieden: 'Inter', 'Intron' und 'CDS'. 'Inter' ist dabei keine offiziell definierte Region im Sinne des Feature Tables (vgl. [4]), sondern bezeichnet alle Abschnitte, die weder Intron noch CDS sind.[70]

```
intron 4 776267..776723 457
0,175 0,407 0,265 0,153
CDS J 776724..776798 75
0,187 0,333 0,307 0,173
intron 5 776799..777238 440
 0,070 0,393 0,380 0,157
```

Abbildung 16: Ausschnitt aus einer W-General-Datei

CDSs, die nur einen Teil eines Proteins codieren, werden als CDS J bezeichnet. Die gesamte Codesequenz eines Proteins kann sich also aus mehreren CDS J zusammensetzen. Die Introns einer Genversion sind in der Datei durchnummeriert.
Für das X-Chromosom hat die W-General-Datei einen Umfang von 27736 Zeilen und eine Größe von 802 KB.

CHRX-Seq.txt (S-General)

Diese Datei korrespondiert mit dem Typ W-General. Sie listet exakt die gleichen Regionen mit Typ, Beginn, Ende und Länge auf. Allerdings enthält sie für jede Region die komplette Basenfolge in Form der Buchstaben *a, t, g* und *c* sowie *n* für unbekannte Basen. In einer Zeile stehen dabei maximal 60 Basen.
Die Datei umfasst für das X-Chromosom über 2.7 Millionen Zeilen und hat eine Größe von 159 MB.

```
CDS J 3494730..3494790 61
atgagggaaggaagtcggcctgggtgcggccccttggcccttagcgcgcagtcccttagg
g
intron 1 3494791..3494888 98
gtctcctggaaacccggcgcatgcggccctgagggcccgctgacccaccgggtgcccatc
tgcgacagggttcctatggcgtgggtggagcagcacag
```

Abbildung 17: Ausschnitt aus einer S-General-Datei

69 Die Gesamtzahl der bekannten Basen ist nicht vermerkt und kann nur durch Zählen ermittelt werden.
70 Zu Beginn des Projekts war geplant, noch weitere Regionen zu differenzieren, u.a. 'repeat_region', 'misc_feature' oder '5'UTR'. Jedoch wurde dieser Plan seitens Fr. Maïda später wieder fallengelassen.

CHRX-CDS-Freq-3.rtf (W-CDS)

Diese Datei listet alle CDSs des Stranges in ihrer Reihenfolge auf. Dabei sind auf mehrere Exons verteilte Sequenzen bereits zusammengefasst. Jeder Eintrag besteht bei deaktiviertem automatischem Zeilenumbruch aus zwei Zeilen. Die erste enthält zunächst den CDS-Typ, gefolgt von einer Liste aller entsprechender Exon-Grenzen[71]. Ist die Grenze nicht genau bekannt, so wird das mittels < oder > markiert. Es werden zwei Typen unterschieden. *CDS* bezeichnet eine Genversion, deren codierende Sequenz nicht von Introns unterbrochen wird, *CDS J* steht für Genversionen, die auf mehrere Exons verteilt sind. Abschluss der ersten Zeile bildet die Länge der gesamten Sequenz.
Die zweite Zeile beinhaltet die Auftretenswahrscheinlichkeiten aller 64 möglichen Codons in alphabetischer Reihenfolge. Die Genauigkeit beträgt auch hier drei Nachkommastellen.
Für das X-Chromosom hat diese Datei einen Umfang von 2190 Zeilen und eine Größe von 609 KB.

```
CDS J join(<2798730..2799021,2799811..2799845) 327
0,046 0,000 0,046 0,009 0,019 0,000 [...]* 0,019 108,000
```

*Abbildung 18: Ausschnitt aus einer W-CDS-Datei; * Hier wurden formal gleiche Einträge aus Platzgründen weggelassen.*

CHRX-CDS-Seq.rtf (S-CDS)

Diese Datei korrespondiert mit dem Typ W-CDS, sprich, sie listet dieselben Einheiten auf. Die erste Zeile eines Eintrags ist mit der entsprechenden Zeile aus W-CDS identisch. Daran schließt sich die Basensequenz des CDS in der gleichen Form wie im Typ S-General an. Diese Basenfolge ist die direkte Grundlage für ein Protein.
Für das X-Chromosom hat diese Datei einen Umfang von 26068 Zeilen und eine Größe von 1.7 MB.

```
CDS J join(<2798730..2799021,2799811..2799845) 327
ggaaaggtgtcgcctctttttctgtggaaagaggaagctcatgagcgcgaaacagcaggg
gacggagggcg [...]* gagaagaatccatt
```

*Abbildung 19: Ausschnitt aus einer S-CDS-Datei; * Hier wurde ein Sequenzabschnitt aus Platzgründen weggelassen.*

71 Erste und letzte Base des jeweiligen Abschnitts.

Alle für Fr. Maïda relevanten Informationen sind also auf vier Dateien verteilt. Die Verteilung ist nicht exklusiv, sondern es gibt direkte und indirekte Redundanz. Direkte Redundanz meint, dass dieselbe Information in mehr als einer Datei als Zahl abgelegt ist. Indirekt redundant sind all die Informationen, die in den Daten einer Datei enthalten sind, jedoch durch Berechnung erst extrahiert werden müssten.
Tabelle 3 gibt einen Überblick über die Verteilung der Informationen in den vier Dateien.

	W-General	S-General	W-CDS	S-CDS
% der Basen pro CDS	■	■	■	■
% der Basen pro Gen	■	■	■	■
% der Basen pro Intron	■	■	■	■
% der Basen pro Inter	■	■	■	■
% der Codons pro Gen[72]	■	■	■	■
Grenzen von CDS	■	■	■	■
Grenzen Intron	■	■	■	■
Grenzen Inter	■	■	■	■
Grenzen Gen	■	■	■	■

■ *: enthalten* ■ *: berechenbar* ■ *: nicht enthalten*

Tabelle 3: Übersicht über die Verteilung der Informationen in den vier Dateien

5.2.2 Vor- und Nachteile der Aufteilung

Die Aufteilung hat Vor- und Nachteile. Aus der Datei S-General lassen sich prinzipiell alle Informationen berechnen (vgl. Tab. 3). Diese Berechnung kann jedoch wie im Fall der Wahrscheinlichkeiten der Basen oder Codons in einer Genversion sehr rechenintensiv sein. Für eine Darstellung dieser Werte als in Echtzeit scroll- und zoombares Liniendiagramm wäre eine JIT-Berechnung zu langsam. Die Alternative, alle Informationen direkt beim Laden der Datei zu berechnen, würde wiederum zu einer unverhältnismäßig langen Ladezeit führen. Die einmalige Berechnung und Speicherung häufig verwendeter Informationen ist, zumal da sie keiner Änderung unterliegen, also durchaus sinnvoll.
Auch die Trennung von Sequenzdaten einerseits und Auftretenswahrscheinlichkeiten andererseits auf separate Dateien ist zweckmäßig, da die Sequenzdaten um ein Vielfaches größer sind als die Wahrscheinlichkeiten, aber längst nicht immer benötigt werden. Eine Datei vom Typ W-General lässt sich problemlos via Mail versenden, während eine Sequenzdatei einigen Traffic verursacht. Für eine Überprüfung, ob das Programm auch das anzeigt, was in den Dateien steht, kann es nötig sein, Regionen in der Datei in einem Texteditor zu suchen. Das geht bei den kleinen General-Dateien deutlich schneller als in einer Sequenz-Datei. Auch das

72 Da die Sequenzen aus den Regionen Inter und Intron nicht in Proteine übersetzt werden, ist die Triplet-Information für diese Abschnitte zumindest in diesem Zusammenhang nicht relevant.

bloße Öffnen einer Datei vom Typ S-General nimmt mit WordPad, dem Windows Editor und Star Office bis zu einer halben Minute in Anspruch.
Die Aufteilung von Sequenzen und Wahrscheinlichkeiten auf zwei Dateien ist also sinnvoll.
Die Aufteilung auf vier Dateien ist hingegen weniger vorteilhaft. Tabelle 3 zeigt, dass die Dateien vom Typ W-CDS mit der Wahrscheinlichkeit der Codons eigentlich nur eine einzige berechnungsintensive Information exklusiv enthalten. Diese hätte auch in den Typ W-General integriert werden können. Stattdessen wurden gleich zwei neue Dateitypen verwendet, die mittels eigener Routinen geparst werden müssen. Die Unterscheidung zwischen General- und CDS-Dateien führt außerdem dazu, dass viele Routinen wie das Erkennen gemeinsamer Regionsgrenzen oder die Identifizierung von Regionskopien und Genversionen doppelt und nach unterschiedlichen Methoden implementiert werden mussten (siehe dazu auch 6.2).

5.2.3 Weitere Problematische Aspekte

Neben der unnötigen Aufteilung auf vier Dateien gibt es zwei weitere große Probleme innerhalb der Dateien. Zum Ersten hat ein Gutteil der Gene verschiedene Genversionen beziehungsweise Transcripts (vgl. 2.1.3). Manche Versionen unterscheiden sich nur um wenige Basen, andere haben nur wenige Basen gemeinsam. In den großen Datenbanken haben alle diese Versionen denselben Gennamen und sind so sofort als verwandt erkennbar.
In den hier verwendeten Dateien sind Genversionen jedoch nicht markiert. Fr. Maïda war die Information um Genversionen verständlicherweise wichtig, also musste für alle Dateien eine Versionserkennung implementiert werden. Im Typ W-General wird eine Version immer dann attestiert, wenn zwischen zwei Regionen des Typs 'CDS' weder 'Intron' noch 'Inter' liegen[73]. Im Typ W-CDS ist dies nicht möglich, da die Liste ausschließlich aus Genversionen besteht. Hier erfolgt die Identifizierung anhand der Versionsgrenzen. Liegt der Beginn einer Genversion vor dem Ende der vorhergehenden, so werden beide als Versionen desselben Gens interpretiert. Beide Vorgehensweisen klassifizieren den Großteil der Versionen korrekt. Gerade die zweite Methode ist jedoch weder wissenschaftlich begründet noch unfehlbar. Sie scheitert beispielsweise systematisch immer dann, wenn ein 'Nested Gen' vorliegt. Das ist ein Gen, dass sich komplett innerhalb des Introns eines anderen Gens befindet[74]. Die RNAs dieser beiden Transcripts haben keine einzige Base gemeinsam, und obschon der Code im gleichen Abschnitt von des Chromosoms zu finden ist, handelt es sich um zwei verschiedene Gene. Die hier verwendete Methode interpretiert ein 'Nested Gen' hingegen als Version des umschließenden Gens. Das ist wissenschaftlich falsch. Zur Vermeidung solcher Fehler wäre eine präzise Indikation der Genversionen in den Dateien notwendig. Man könnte beispielsweise neben dem Regionstyp jeweils den Namen des Gens angeben, unter dem die Sequenz in den Online-Datenbanken zu finden ist.
In den Dateien finden sich zum Teil exakte Kopien von Genversionen, sprich, identische Gruppen von Regionen finden sich teils mehrfach hintereinander[75]. Auch diese Kopien sind

73 Ein 'Inter' ist genau der Bereich zwischen zwei Genen. Das Fehlen ist also ein klarer Hinweis, dass man sich noch im selben Abschnitt des Chromosoms befindet.
74 Eine systematische Analyse der Nested-Genes beim Menschen findet sich in [131].
75 Als Begründung habe ich von Fr. Maïda nur den vagen Verweis auf „Kalkulationen von Wissenschaftlern“ erhalten.

nicht gekennzeichnet, müssen aber als solche erkannt, von einer Genversion unterschieden und ignoriert werden, da sie komplett redundant sind. Diese Erkennung ist keineswegs trivial. Zur Verdeutlichung der Problematik seien zwei Genversionen angenommen, also zwei Einträge in W-General bzw. W-CDS. Beide Versionen bestehen aus zehn Exons, und der einzige Unterschied ist, dass das letzte Exon der zweiten Genversion um eine Base kürzer ist als das der ersten. Gäbe es diesen Unterschied nicht, handelte es sich um eine Kopie, die zu ignorieren wäre. Die Erkennung solcher Kopien musste auch wieder doppelt implementiert werden. Im Typ W-CDS sind Kopien zwei identische Einträge direkt hintereinander. Der Vergleich kann somit einfach durch einen Abgleich von zwei direkt einlesbaren Zeichenketten erfolgen. Im Typ W-General hingegen beginnt eine Kopie einer Exon-reichen Genversionen teils zwanzig Einträge später, und die Entscheidung, ob es sich um eine Kopie handelt, kann erst getroffen werden, wenn die zweite Version komplett gelesen ist. Die genauen Algorithmen sollen hier nicht ausgeführt werden. Es ist jedoch einige temporäre Speicherung erforderlich, und der Ladevorgang wird durch diese Analyse, die nebenbei bemerkt bei jedem Öffnen erneut erfolgt, deutlich verlangsamt.

Meiner Ansicht nach wäre eine Einteilung in zwei Dateien die bessere Wahl gewesen. Die Datei W-General könnte um eine Kopfzeile nach dem Vorbild von W-CDS, ergänzt um den Namen des Gens, erweitert werden. Die folgende Zeile könnte die Codon-Wahrscheinlichkeiten der Genversion enthalten. Daran anschließen würden sich die Auflistung der einzelnen CDS J und Introns der Genversion entsprechend der jetzigen Form von W-General. Die Größe dieser kombinierten Datei läge für das X-Chromosom immer noch unter 2 MB, sie wäre also immer noch hinreichend 'handlich'. Die Dateien vom Typ S-CDS könnten komplett entfallen, da sie nur ein Ausschnitt der S-General-Datei sind.

Die Struktur der Dateien lag zu Beginn der Implementierung nur teilweise vor. W-General und S-General standen von Anfang an zur Verfügung, während W-CDS und S-CDS zu einem deutlich späteren Zeitpunkt ergänzt wurden. Die Behandlung der ersten beiden war da bereits weitgehend fertiggestell. Typ W-General und Typ W-CDS haben ihren Fokus auf verschiedenen Aspekten des Chromosoms, werden in unterschiedlichen Fenstern visualisiert und können über das Hauptmenü separat geöffnet werden. Es ist also möglich, mit den CDS-Dateien unabhängig von den General-Dateien zu arbeiten und umgekehrt.
Die Visualisierung der W-Dateien dient als Tor zu den entsprechenden S-Dateien, die nicht manuell geöffnet werden können. Sie werden beim Laden von W-General/W-CDS automatisch gesucht und für den Zugriff aufbereitet.

5.3 FUNKTIONSUMFANG

Das Kapitel beschreibt die Funktionen des Programms. Zunächst folgt ein Überblick über die verschiedenen Grafiken und zur Verfügung stehenden Navigationsmöglichkeiten. Anschließend werden die verschiedenen Sonifikationen vorgestellt.

5.3.1 Grafiken

Das Program bietet drei verschiedene Typen von Grafiken. Als Vorlage dafür dienten jeweils Skizzen von Fr. Maïda. Anhand dieser Skizzen besprachen wir, welche Möglichkeiten der Navigation und der Manipulation es geben sollte und wie die Steuerung dafür aussehen könnte. Aus Platzgründen finden sich bei den Beschreibungen nur Ausschnitte der Screenshots. Die vollständigen Ansichten finden sich im Anhang.

Regionsansicht

Die Regionsansichten sind Balkendiagramme. Die dargestellten Regionen sind entlang der X-Achse in der Reihenfolge ihres Auftretens auf dem direkten Strang des Chromosoms aufgereiht. Die Höhe der Balken ergibt sich aus der Länge der Regionen in Basen. Es gibt drei Versionen der Regionsansicht, die sich anhand der verwendeten Daten und der Auswahl der Regionen unterscheiden.
Die 'General-View' stellt die Daten der General-Dateien dar. Die Regionen sind Inter, Introns und CDSs. Jede der Regionen kann ausgeblendet oder mittels einer Art Solo-Button exklusiv dargestellt werden. Ausgeblendete Regionen sind in hellgrau, fallen also nicht aus der Anzeige. Versionen desselben Gens werden hintereinander dargestellt, und die Farben aller bis auf der ersten Version werden zur eindeutigen Darstellung automatisch aufgehellt. Neben den Regionen lassen sich die Auftretenswahrscheinlichkeiten der Basen Region als Liniendiagramme einblenden.
Für die 'CDS-View' werden die CDS-Dateien verwendet. Die Balken entsprechen hier Abschnitten des Chromosoms, die codierendes Material enthalten. Sie werden durch das Programm anhand der Daten ermittelt. Enthält ein Abschnitt mehr als eine Version eines Gens so lässt sich ein weiteres Fenster, die Genansicht, öffnen, in dem die einzelnen Versionen nach dem gleichen Prinzip als Balkendiagramm dargestellt sind. In die 'CDS-View' können die Wahrscheinlichkeiten aller Codons als Liniendiagramm einblendet werden. Es wäre technisch durchaus möglich, alle 64 Codons zeitgleich darzustellen. Die Anzeige wäre aber kaum aussagekräftig und wurde daher auf acht Codons zur gleichen Zeit beschränkt. Die Auswahl erfolgt mittels manueller Eingabe des Triplets.
Beide Regionsansichten haben einen vertikalen Zoom, da die Längen der Regionen einen Wertebereich von bis zu sechs Zehnerpotenzen überdecken[76]. Die Wahrscheinlichkeiten sind ebenfalls zoombar, um die Bildschirmhöhe optimal ausnutzen zu können. Auch die Breite der Balken lässt sich frei und in Echtzeit einstellen.

76 Eine Darstellung in logarithmierter Form wurde aufgegeben, da der Großteil der Regionen dann wiederum gleichlang erschien. Ein Logarithmus bildet jedoch die Grundlage des Verhaltens des Zoom-Sliders, dessen Wertinkrement dynamisch mit seinem aktuellen Wert wächst.

5.3.1 Grafiken

Entlang der X-Achse ermöglicht ein Rollbalken die Navigation durch das Chromosom. Entlang der Y-Achse ist keine Navigation möglich.

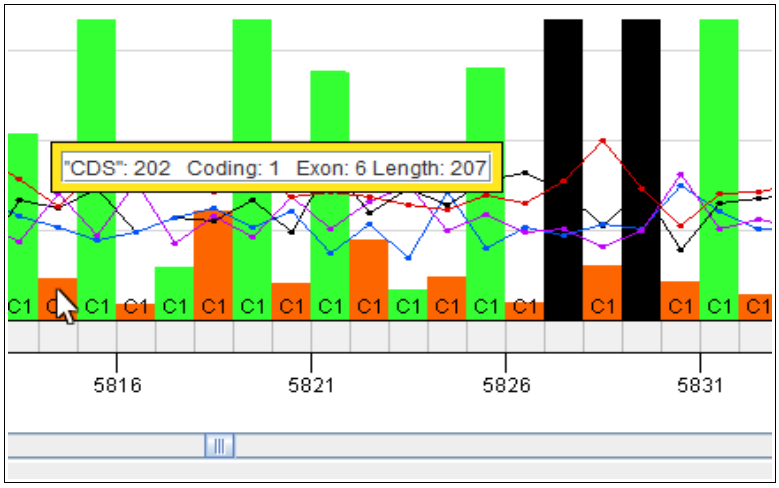

Abbildung 20: Screenshot der 'General-View' mit eingeblendetem Info-Popup; die Angabe 'C_' gibt an, wieviele identische Kopien dieser Region in der Datei enthalten sind. 'C1' meint dabei das Original selbst.

Alle Regionsansichten sind interaktiv und erlauben über einen Mausklick auf einen bestimmten Balken den Zugang zu den entsprechenden Detailinformationen. In der 'General-View' kann mittels Mausklick auf die grauen Balken unterhalb der Balken die Sequenzansicht der jeweiligen Region geöffnet werden. In der 'CDS-View' erfolgt dies über einer Auswahl in einem Popup-Fenster.

Sequenzansicht

Sequenzansichten gibt es einerseits für die Basenfolgen, andererseits für die Codon- beziehungsweise Aminosäurefolgen einer Region. Die Einheiten sind entlang der X-Achse als farbige Boxen aufgereiht. Stehen zwei oder mehrere identische Einheiten hintereinander, so werden sie zu einer Box zusammengefasst. Unterhalb der Boxen findet sich eine Angabe des Element an sich, und darüber, aus wievielen gleiche Elementen der Cluster besteht.
Weiterhin enthält die Grafik Angaben sowohl über die absolute Position der Einheit innerhalb der Region als auch innerhalb des Chromosoms. Umfasst die Region mehrere Exons, so werden die Exon-Grenzen mittel vertikaler, grauer Balkens markiert.
Die Ansicht kann entlang der X-Achse gezoomt werden. Bei der Darstellung von Basensequenzen lassen sich einzelne Basen ausblenden oder auf 'solo' schalten. Für die Aminosäuren ist dieselbe Funktionalität bisher nicht implementiert, da dafür eine kleinteilige GUI erforderlich wäre.
Von einem allgemeinen Standpunkt aus wäre es wohl erstrebenswert, alle Sequenzen (Basen/Codons/Aminosäuren) in einer Grafik darzustellen, Frau Maïda legte jedoch Wert auf eine Trennung der Aspekte.

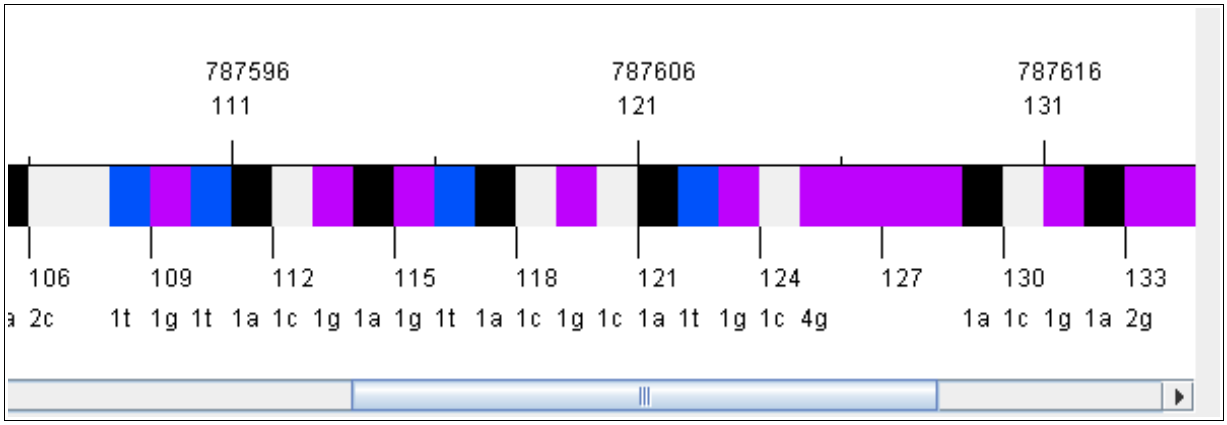

Abbildung 21: Screenshot der Sequenzansicht (hier für die Basen); die Base Cytosin ist ausgeblendet und daher hellgrau. Die oberste Zahlenreihe enthält die Basennummern bezogen auf das gesamte Chromosom, die zweite Zahlenreihe zählt die Basen ab Anfang der Region Die erste Zahlenreihe unter der Grafik zählt ebenfalls von Beginn der Region, jedoch in Dreierschritten. Die unterste Reihe ist die Angabe der Base zusammen mit der Angabe wie oft sie hintereinander folgt.

Gen-Ansicht

Die Gen-Ansicht stellt die verschiedenen Versionen[77] eines einzelnen Gens dar. Sie lässt sich nur mittels Mausklick auf einen hellblauen Balken in der 'CDS-View' öffnen. Für Gene, die nur eine Version haben, kann keine Gen-Ansicht geöffnet werden.
Auf der linken Seite werden für jede Version die einzelnen Exons übereinander 'gestapelt' dargestellt. Die waagerechten, weißen Linien markieren die Grenzen der einzelnen Exons. Diese Darstellung, in der Form von Fr. Maïda erbeten, erlaubt zwar einen schnellen Vergleich der Gesamtlängen von Transcripts, ist aber sonst wenig aussagekräftig, wenn nicht gar stark irreführend. Unterscheiden sich zwei Genversionen direkt an ihrem Anfang etwa um ein Exon, so werden beide Balken dem Augenschein nach nichts mehr gemeinsam haben. Das wird offensichtlich, betrachtet man in Abbildung 22 die erste und die letzte Version.
Für einen besseren Überblick über die lokalen Struktur der Genversion ist daher auf der rechten Seite die Exon-Intron-Struktur abgebildet. Exons sind blau und unabhängig von ihrer relativen Länge mindestens 1 Pixel hoch. Diese Darstellung entspricht, mit dem Unterschied, dass sie hier vertikal erfolgt, der gängigen Praxis. Sie erlaubt einen schnellen Vergleich der Ortsstrukturen der verschiedenen Transcripts.
Zusätzlich lassen sich die Wahrscheinlichkeiten aller 64 Codons als Liniendiagramm einblenden. Auch hier gilt jedoch eine Beschränkung auf zehn Codons zur selben Zeit. Die Auswahl erfolgt mittels manueller Eingabe des Triplets in die Textfelder auf der rechten Seite, wo auch die Farben festgelegt werden.
Die Darstellung auf der linken Seite und auch die Wahrscheinlichkeiten lassen sich unabhängig vertikal skalieren[78].

77 Hier soll nochmals darauf hingewiesen werden, dass die Attributierung 'Version' algorithmisch vorgenommen wird und nicht korrekt ist.
78 Aufgrund von Rundungsfehlern wird Zahlenangabe der Achsen fehlerhalft, sobald ein kleiner Wertebereich (ca. < 20) abgebildet wird

5.3.1 Grafiken

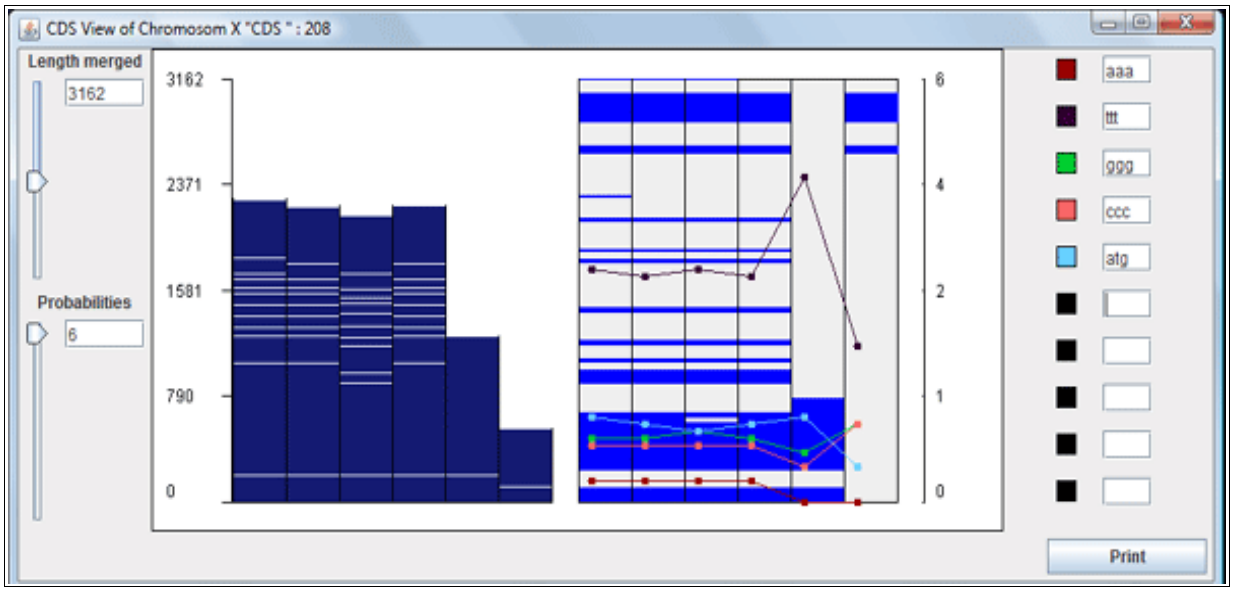

Abbildung 22: Screenshot der Genansicht eines Gens mit sechs Versionen; es sind 5 Wahrscheinlichkeiten eingeblendet.

Spektrum-Ansicht

Die Spektrum-Ansicht zeigt ein Histogramm der Regionslängen zusammen mit einem an ein Spektogramm angelehntes Grauwertbild[79]. Da das Histogramm als Frequenzspektrum interpretiert wird, ist die Darstellung der Längen logarithmisch zur Basis 10 und auf den Bereich 20-20.000 limitiert.

Die Darstellung weicht vom Standard eines normalen Spektrums oder Histogramms insofern ab, als dass sie um 90° gegen den Uhrzeigersinn gedreht und anschließend an der jetzt senkrechten X-Achse gespiegelt ist. Der Grund für diese Abweichung liegt in der Bitte Frau Maïdas, die Anzeige solle möglichst der ihr geläufigen Spektogramm-Ansicht von Pro-Tools entsprechen, auch wenn eigentlich keine Zeitdimension vorliegt. Dieser Wunsch ist auch der Grund für das Grauwertbild, das dem Histogramm eigentlich keinerlei neue Informationen hinzufügt.

Keine der beiden Achsen ist direkt scroll- oder zoombar. Es gibt aber vier Schieberegler zur

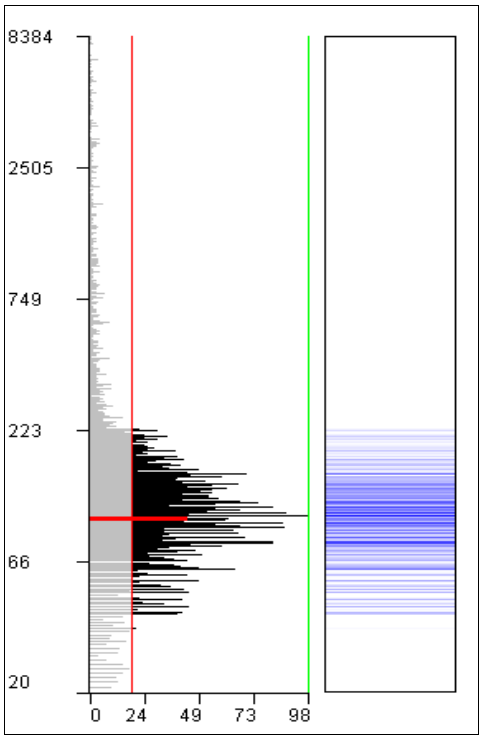

Abbildung 23: Screenshot der Spektrum-Ansicht; der Threshhold liegt bei 19

79 Hier in blau.

Kontrolle von Parametern der Sonifikation, die sich auch direkt auf die Darstellung auswirken. Zum einen gibt es einen High- und einen Low-Cut, mit Hilfe derer sich eine mini- und eine maximal zu berücksichtigende Länge festlegen lässt. Eine Einstellung des Low-Cut auf 50 bewirkt beispielsweise, dass alle Regionen, die kürzer als 50 Basen sind, weder angezeigt noch sonifiziert werden. Der Regler 'Minimum' legt fest, wie oft eine bestimmte Länge mindestens auftreten muss, um in der Sonifikation berücksichtigt zu werden. 'Maximum' legt analog eine Maximalhäufigkeit fest, die differenziert werden soll, wirkt also in etwa wie ein Limiter.
Das Histogramm setzt sich aus den Einzelhistogrammen der Regionstypen zusammen, die sich einzeln an- und abschalten lassen. Zudem gibt es eine Art 'Gain', mittels dessen sich der Anteil eines Typs zum Gesamthistogramm verstärken lässt.

5.3.2 Sonifikationen

Innerhalb des Programms sind vier verschiedene Sonifikationen implementiert.

Sequenz – Melodie

Die Sonifikation der Frequenz folgt dem bereits beschriebenen Grundprinzip des Pitch-Mappings im MIDI-Framework. Sowohl Basen- als auch Aminosäuresequenzen können als Input verwendet werden. Das Mapping ist frei wählbar und bietet sieben Oktaven sowie eine Stummschaltung. Man kann jeder Sequenz ihr eigenes Mapping zuordnen. Auch verschiedene Mappings für dieselbe Sequenz zur gleichen Zeit sind möglich, da die Sequenzen in unabhängigen Fenstern geöffnet werden. Neben den Tonhöhen lässt sich noch die Geschwindigkeit in Bpm festlegen. Stehen mehrere identische Basen oder Aminosäuren hintereinander, wird die gespielte Note, analog zur visuellen Darstellung, um die Anzahl der Wiederholungen verlängert, nicht aber neu angeschlagen.
Die Steuerung der Sonifikation ist im Sequenzfenster integriert. Es gibt einen 'Play/Stop'- , einen 'Save'- und einen 'Pitch'-Button. Über den 'Save'-Button kann die Sequenz als .mid-File exportiert werden. Der 'Pitch'-Button öffnet ein separates Fenster, in dem sich die Tonhöhen festlegen lassen.

Gen-Fingerabdruck

Mit dieser Methode wird der Aufbau bzw. die Struktur einer Genversion sonifiziert (vgl. 4.3.3). Längen und Entfernungsangaben werden auf die Frequenz abgebildet. Die Soundproduktion erfolgt mittels einfacher additiver Synthese von Sinustönen. Es wird immer ein Audioclip im PCM-Format erstellt, der dann mittels der Standard-Soundkarte des Rechners ausgegeben wird. Alle Sinuskomponenten tragen mit der gleichen Energie zum Gesamtsignal bei. Eine A-Kompensation ist nicht integriert.
Das Programm bietet zwei Möglichkeiten, die Längenstruktur zu 'lesen', die einen jeweils anderen Klangcharakter hervorbringen. Die erste Möglichkeit betrachtet die absoluten Längen. Ein Exon der Länge 300 erzeugt also die Frequenzkomponente 300 Hz. Das Ergebnis liefert einen Eindruck davon, ob eine Genversion viele oder wenige Exons umfasst und wie variabel

diese in ihrer Länge sind. Ausreißer und Längencluster sind hörbar. Die zweite implementierte Möglichkeit mappt die kumulativen Länge auf die Frequenz. Die erste Frequenz ergibt sich aus der Länge des ersten Exons, die zweite aus der ersten plus der Länge des zweiten Exons etc..
Die verwendete Abstastfrequenz ist 22050 Hz. Sollten sich aus den Berechnungen Frequenzen über 10.000 Hz ergeben, so werden diese nicht sonifiziert.
Die rudimentäre GUI ist in das Informationsfenster der Genversionen integriert.

Chromosom - Spektrum

Das Chromosom-Spektrum ist eine Art akustischer Fingerabdruck eines Chromosoms (vgl. 4.3.4). Es basiert auf den Mappings 'Länge - Frequenz' und 'absolute Häufigkeit – Amplitude'. Längen über 20.000 werden generell nicht berücksichtigt und bei der Erstellung der Daten für das Spektrum als 20.001 Hz zusammengefasst. Längen unterhalb von 20 Hz werden statistisch ausgewertet, jedoch nicht sonifiziert. Mittels der schon im Zusammenhang mit der entsprechenden Grafik beschriebenen Regler lassen sich Häufigkeiten unterhalb einer Grenze ausklammern und solche überhalb einer Maximalhäufigkeit clippen. Ebenfalls einstellbar ist der zur Sonifikation herangezogene Frequenzbereich. Amplituden- und Frequenzbereich sind jedoch global wirksam, also nicht für die vier Regionstypen einzeln steuerbar. Das Spektrum setzt sich mittels einfacher Summation aus den Teilspektren der vier Regionstypen zusammen. Jeder Typ kann einzeln deaktiviert werden.
Da die Berechnung eines Clips teils einige Sekunden in Anspruch nimmt, wurde auf eine automatische Aktualisierung bei Änderung der Parameter verzichtet. Die Erstellung des jeweils aktuellen Soundsamples erfolgt nur durch Drücken des Update-Buttons.
Ein Export des Audioclips war zum Zeitpunkt der Niederschrift nicht implementiert, da Fr. Maïda an den Zahlenwerten des Spektrums interessiert war und die Audioausgabe nur zur groben Überprüfung benötigte.

Audifikation – Repeat Finder

Bei dieser Sonifikation bildet jede einzelne Base ein Sample eines digitalen Audioclips (vgl. 4.3.2). Der Wert für eine Base lässt sich mittels eines Sliders in fünf Stufen inklusive der Null einstellen. Auch die Abstastfrequenz ist frei wählbar. Das Programm erzeugt einen Audioclip, dessen Länge sich aus der Länge der Sequenz und der Abtastfrequenz ergibt. Der Input ist ein Textfile, dass ausschließlich die Sequenz enthält. Zeilenumbrüche werden ignoriert, Zeichen, die keiner der vier Basen entsprechen bekommen den Samplewert 0.
Der Repeat-Finder kann unabhängig von den Daten Fr. Maïdas verwendet werden und ist über die Menüleiste des Programm zu öffnen.

5.3.3 Weitere Funktionen

Neben den bereits beschriebenen Grafikfenstern gibt es noch zwei verschiedene Informationsfenster, eines für die Regionen aus W-General (Intron, CDS, Inter) und eines für eine Region aus W-CDS (Genversion). Prinzipiell bieten beide Fenster ähnliche Informationen, jedoch ist das Fenster der Genversionen wesentlich ausführlicher als das der anderen Regionen. Es enthält die Auftretenswahrscheinlichkeiten aller Codons (auch als Absolutwert), die Auflistung der Grenzen der Exons sowie ein Strukturschema der Genversion, vergleichbar mit der Anzeige in der Gen-Ansicht.

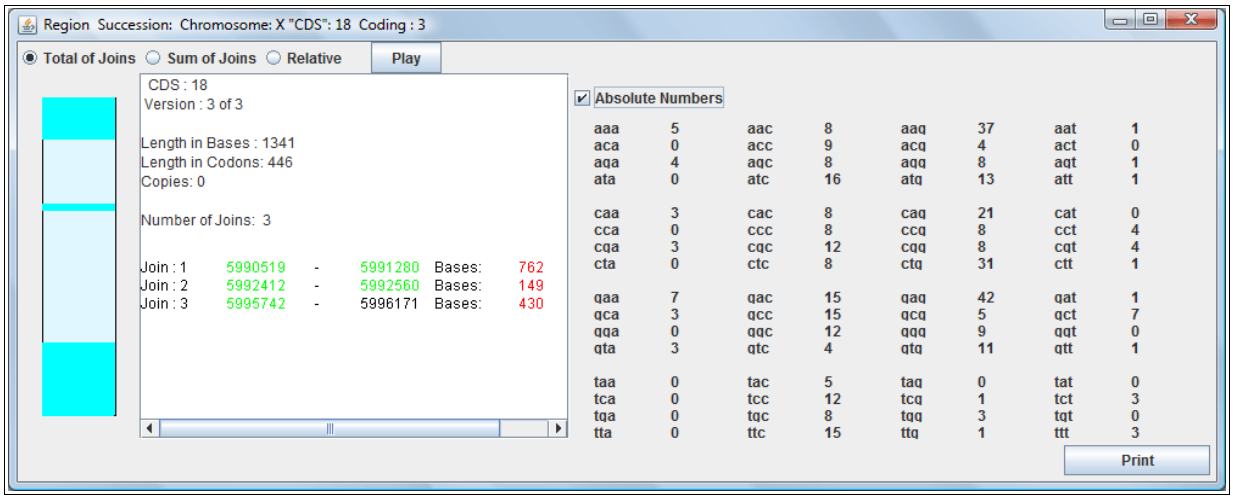

Abbildung 24: Informationsfenster einer Genversion

Das Programm bietet die Möglichkeit, nahezu jedes Grafik- und Informationsfenster auszudrucken. Dabei verwendet es teils die gleichen Zeichenroutinen wie die Ausgabe auf dem Bildschirm, teils wurden aber modifizierte Methoden geschrieben. Desweiteren bietet das Programm die Möglichkeit, eine Reihe von Textdateien zu einem Chromosom zu speichern. In diesen Textdateien stehen Regionslängen, Typen und die Häufigkeit des jeweiligen Auftretens. Im Grunde sind sie dem Input sehr ähnlich, und lediglich die Häufigkeit des Auftretens einer bestimmten Länge ist in den Originaldateien nicht enthalten. Diese Dateien bilden den Input für ein OpenMusic-Programm, dass Fr. Maïda mit einem Kooperationspartner entwickelt, und in dem mit den Spektren gearbeitet und komponiert werden soll.

5.4 IMPLEMENTIERUNG

Dieses Kapitel beschäftigt sich mit der konkreten Implementierung, wobei nicht alle Einzelheiten erläutert, sondern lediglich die wichtigsten Konzepte, Klassen und der Programmablauf vorgestellt werden. Alle JAVA-spezifischen Bezeichnungen wie Klassen- oder Funktionsnamen sind zur besseren Erkennbarkeit in **`Courier New`** geschrieben.

Insgesamt umfasst das Programm rund 55 aktive Klassen und ca. 10.000 Zeilen Code. Beim Öffnen erscheint eine Menüleiste, über die sich die verschiedenen Dateien laden sowie das Color-Menü öffnen lassen. Schließen dieser Leiste bewirkt gleichzeitig das Schließen des Programms. Alle Grafiken werden auf Wunsch Fr. Maïdas in einem separaten Fenster geöffnet, die auch die weiterführende Funktionalität enthalten. Eine Menüleiste für die einzelnen Grafiken wurde nicht integriert, da Fr. Maïda als einziger Nutzer feststand, und so kein besonderes Augenmerk auf Accessibility gelegt wurde.

Ein den Programmcode prägendes Paradigma betrifft den Ort von Methoden. Der Code für eine Funktionalität befindet sich immer im Element, aus dessen GUI die Funktionalität aufgerufen wird. Wird beispielsweise mittels eines Mausklicks in eine Grafik ein Fenster geöffnet, so befinden sich die dazu benötigten Methoden in der Grafik-Klasse, sind also nicht in einer Controller-Klasse enthalten. Das ist entgegen des gängigen Prinzips einer MVC-Architektur und liegt hauptsächlich in meiner mangelnden Programmierpraxis begründet (vgl. 6.1). Die Programmierung erfolgte mittels der Eclipse IDE, Build id: 20100218-1602 .

5.4.1 Klassenstruktur einer Grafik

Jede der grafischen Darstellungen ist ein Komplex mit folgenden Klassen.

Chef-Klasse (`___Chef`)[80]:

Diese Klasse, abgeleitet von `Chef`, ist ein Container, der sowohl die Grafik selbst als auch die Steuerelemente enthält und anordnet. Das GUI-Layout der Grafik wird in dieser Klasse definiert. Auch die Funktionalität der Steuerelemente findet sich hier, insoweit als dass sie nicht in den Elementen selbst gekapselt ist.
Die Klasse `Chef` erweitert die Klasse `JPanel`, muss also immer in einen `JFrame` eingebunden werden. Dieser wird beim Öffnen der jeweiligen Grafik erzeugt. Der Code dazu befindet sich in der Methode, die das Öffnen initiiert. Die Chef-Klasse selbst erhält eine Referenz auf dieses Fenster, sodass es aus ihr heraus geschlossen werden kann.

80 In den Klammern ist die Schreibweise angegeben, die man innerhalb des Programms findet. ___ ist der Teil, der den Grafiktyp angibt, also zum Beispiel `SuccessionData`

Grafik-Klasse (`___Graph`):

Diese Klasse, abgeleitet von `GraphWindow`, enthält in der überladenen Methode `paintComponent()` die eigentlichen Zeichenroutinen. Werden, wie beispielsweise in der 'General-View' verschiedene Informationen dargestellt (Balkendiagramm, Liniendiagramm), so sind die Routinen in separate Methoden zerlegt.
Die Grafik-Klasse instanziiert die Achsen und gibt diesen bei Bedarf den Befehl zum Zeichnen. Zur Kommunikation mit den Achsen-Klassen enthält sie außerdem einige Methoden, die beispielsweise für einen gegebenen Bildpixel den entsprechenden Datenwert des Modells zurückgeben.
Für eine ruckelfreie Darstellung wird stets nur der sichtbaren Teil einer Grafik sowie ein kleiner Puffer an den Rändern gerendert. Vor dem Zeichnen muss also ermittelt werden, welcher Datenpunkt gerade am linken Rand zu sehen ist. Dies erfolgt anhand der Position des Rollbalkens und des Abbildungsfaktors in x-Richtung[81]. Aus dieser Systematik ergibt sich die Forderung nach Lückenlosigkeit der Darstellung. Es ist nicht möglich, Elemente des dargestellten Datenvektors wegzulassen, da sonst eine Bestimmung des am linken Bildrand befindlichen Datenpunktes nicht mehr ohne Weiteres möglich wäre.
Neben den Zeichenroutinen enthält die Grafik-Klasse auch die Print-Routinen. Teilweise verwenden Druck- und Bildschirmausgabe dieselben Methoden, teilweise werden auch unterschiedliche Methoden verwendet, die sich allerdings nur marginal unterscheiden. Die Möglichkeit zu drucken war nicht Teil der ursprünglich besprochenen Funktionsumfangs, und Fr. Maïda formulierte diesen Wunsch erst sehr spät. Daher wurde die Druckfunktionalität 'schnellstmöglich' implementiert, was derartige Code-Dopplungen wenn auch nicht entschuldigt so doch erklärt.
Die Grafik-Klasse, ein `JPanel`, wird immer von der Chef-Klasse instanziiert und im Bildschirm-Layout platziert.

Info-Pool (`___InfoPool`):

Für jede Grafik gibt es eine Instanz der Klasse `GraphInfoPool`. Die Instanziierung erfolgt in der Chef-Klasse, die eine Referenz an die Grafik-Klasse übergibt.
Der Info-Pool entstand in erster Linie als ein Hilfsmittel zum Programmieren. Alle Klassen einer Grafik verwenden den gleichen Info-Pool und 'registrieren' sich dort[82]. Alle wichtigen Variablen wie zum Beispiel Abbildungsfaktoren sind in den Info-Pool ausgelagert oder dort als Referenz vorhanden. So können sie von überall identisch adressiert werden. Sowohl in der Grafik-Klasse als auch in einem Navigationsmodul wie dem `NavigationAssistant` funktioniert die folgende Codezeile und liefert stets dasselbe Objekt:

```
infoPool.chefRef
```

Der Vorteil ist einerseits, dass den Konstruktoren anstatt einer Reihe verschiedener Parameter nur eine Referenz übergeben werden muss und trotzdem alle Variablen zur Verfügung stehen.

81 Der Abbildungsfaktor ist beispielsweise die Balkenbreite der 'General-View'.
82 Der Info-Pool enthält also Referenzen aller Klassen.

Eine spätere Erweiterung einer Grafik-Klasse etwa um einen Zoom in y-Richtung erfordert keine Änderung des Konstruktors. Alle dazu nötigen Variablen sind im allgemeinen Info-Pool bereits deklariert und können verwendet werden.
Der Info-Pool ist außerdem immer eine Instanz derselben Klasse, während die Chef- und Grafik-Klassen als abgeleitete Klassen trotz der gemeinsamen Basis sehr unterschiedlich sein können. Lägen die Variablen in einer dieser Klassen, so müsste ein Navigationsmodul, dass von beiden verwendet werden soll, auch zwei Konstruktoren enthalten. Lagerte man die Variablen wiederum in die Parent-Klassen aus, so könnte zwar auf verschiedene Konstruktoren verzichtet werden, allerdings müssten dann Casting-Operation erfolgen.
Bestes Beispiel für die Funktionalität des Info-Pools ist der Transformationsfaktor `XstepWidth`, der einen Datenwert auf die x-Koordinate eines Pixels abbildet. Er wird beim Rendern verwendet, kann mittels eines Sliders oder mittels manueller Eingabe geändert werden und wird auch für die Berechnung des einem Bildpixel entsprechenden Datenpunktes benötigt. Diese häufig verwendete Variable befindet sich im Info-Pool und kann von jeder Klasse verwendet werden, die den Info-Pool kennt. Das ermöglichte ein schnelles, komfortables Programmieren.

5.4.2 Datenbehandlung

Die Behandlung der Daten ist im Programm leider nicht einheitlich. Die Daten liegen in eigenen Klassen (`GeneralData`, `GeneralSuccessionData`, `CDSData`), die jeweils auch die Methoden zum Parsen sowie zum Zugriff und zur statistischen Auswertung bereitstellen. Sie sind von der Klasse `GraphData` abgeleitet, die allerdings nur ein leerer Stub ohne abstrahierte Funktionalität ist. Die Instanziierung einer Datenklasse erfolgt in der korrespondierenden Chef-Klasse, die das zu parsende File als Argument übergibt.
Ein einzelner Datenpunkt ist immer eine Region, zu der verschiedene Informationen zur Verfügung stehen. Daher wurden für diese Einheiten entsprechend dem Paradigma der OOP die Klassen `CDS` und `ChromoRegion` erstellt, welche von der Klasse `Region` abgeleitet sind. Sie enthalten Informationen wie Start und Ende der Region, die Anzahl der Introns und Exons sowie deren Grenzen, den Regionstyp, die Anzahl der Kopien dieser Region in den Eingangsdaten und auch die Basensequenz selbst beziehungsweise die Adresse derselben auf der Festplatte. Weiterhin stellen sie Zugriffsmethoden für diese Werte zur Verfügung und enthalten einige Hilfsalgorithmen, wie den zur Erstellung einer Codonsequenz aus einer Basensequenz.
Sequenzen, soweit in den Arbeitsspeicher geladen, liegen als int-Array vor. Die komplette Sequenz eines Chromosoms wird allerdings wegen der Größe von bis zu 250 MB nicht geladen. Abschnitte werden vielmehr mittels der JAVA-Klasse `MemoryMappedFile` direkt von der Festplatte gelesen. Die Regionsklassen enthalten dann die nur die Byteadresse desentsprechenden Abschnitts.

5.4.3 Sonifikationsklassen

Die für eine Sonifikation nötigen Methoden liegen jeweils in einer eigenen Klasse. Im Einzelnen sind dies `CDSSound`, `SpectrumSound` und `SuccessionSound` sowie `RepeatFinder`. Diese Klassen enthalten mit Ausnahme von `SpectrumSound` auch die jeweiligen Steuerelemente und deren Layout. Zur Sounderstellung und Ausgabe dienen die Methoden der JAVA Sound API, die sowohl Funktionalität für 'digitales Audio' als auch für MIDI bereitstellt. In beiden Bereichen hat man die zur Ausgabe die Wahl zwischen Streaming und der Erstellung eines Clips, respektive einer Sequenz. Zweiteres schien programmiertechnisch weniger problematisch und weniger systemabhängig. Außerdem sind so weniger Erweiterungen für eine Speicherung der Clips nötig als bei einer Ausgabe mittels Streaming. Für die Sequenz-Sonifikation wird also die MIDI-Sequenz zunächst vollständig erzeugt und dann abgespielt. In allen anderen Fällen wird ein PCM-Audioclip berechnet. Die Länge hängt beim Repeat-Finder von der Länge der Sequenz und der gewählten Abtastfrequenz ab, für Chromosom-Spektrum und den Gen-Fingerabdruck ist die Dauer softwareseitig auf eine Sekunde festgelegt.
Die additive Synthese erfolgt in beiden Fällen ohne Look-Up-Table, sprich, für jede beteiligte Frequenz und jedes Sample muss eine Sinusoperation berechnet werden. Im Falle des Gen-Fingerabdrucks ist dies kein großes Problem, da hier nur wenige Frequenzen am Klang beteiligt sind. Beim Chromosom-Spektrum tragen jedoch teils 10.000 Frequenzen zum Ergebnis bei. Hier kann die Berechnung also einige Sekunden in Anspruch nehmen.
Um ein unkorreliertes Signal zu erhalten, wird beim Chromosom-Spektrum für jede Frequenz eine zufällige Phase zwischen 0 und 2*Pi erzeugt. Dies erfolgt einmalig bei der Instanziierung der Klasse `SpectrumSound`. Ohne die Zufallsphase enthielt der resultierende Sound, soweit eine hinreichend große Zahl an tieferen Frequenzen beteiligt war, am Anfang und Ende ein extrem energiereiches Knacken. Die Zufallsphase wurde zunächst auch für den Gen-Fingerabdruck angewendet. Dort wirkte sie sich aufgrund der wenigen beteiligten Frequenzen negativ aus und hatte ein durch den Schnitt in der Sinuswelle bedingtes Initialknacken zur Folge. Daher sind beim Gen-Fingerabdruck alle Sinuskomponenten phasengleich.

5.4.4 Navigations- und Hilfsklassen

Neben den bisher vorgestellten gibt es noch eine Reihe weiterer, oft verwendeter Klassen. Eine Gruppe, die man als allgemeine Hilfsklassen bezeichen könnte, stellt Funktionen zum Öffnen und Speichern von Dateien bereit oder enthält die Basisfunktionen zum Drucken und zum programmweiten Management der Farben[83].
Die zweite Gruppe sind die Navigationsklassen. JAVA stellt eine große Zahl vorgefertigter GUI-Module (Buttons, Slider, Scrollbar) zur Verfügung. Die Verwendung derselben ist jedoch nicht immer komfortabel und erfordert teils etwas länglichen Code. Daher wurden relativ zu Beginn der Arbeit einige Navigationselemente erstellt, die bestimmte Funktionalitäten kapseln,

83 Die Farben sind stets in allen Fenstern identisch und können im Menü 'Colors' verändert werden.

und die als komplettes Element in der GUI platziert werden können[84]. Die folgende Tabelle gibt eine kurze Übersicht dieser Klassen.

`Sliderbox`	– ein Slider; manuelle, überprüfte Eingabe des Wertes möglich
`NavigationAssistant`	– eine Scroll-Bar mit optionaler `Sliderbox` für eine Dimension;
`ProbabilityUnit`	– Farbwahlfeld, Textfeld mit Überprüfung zur Eingabe des Codons, optionale Anzeige der entsprechenden Aminosäure
`DataDisplayManager`	– Panel aus Solo/Mute-Switches für beliebig viele Elemente einer beliebigen Grafik

Tabelle 4: Übersicht wichtiger Navigationsklassen

84 Ziel war, diese Elemente soweit wie möglich 'Plug-and-Play'-esk zu erstellen. Besonders bei der Sliderbox ist dies gut gelungen, da sie neben der Instanziierung nur zwei Zeilen Code nötig macht, um voll zu funktionieren.

6. Diskussion

6.1 Diskussion des Programms

Für die Diskussion des Programms soll zwischen zwei Aspekten differenziert werden. Der Erste ist die Qualität des Programms in Bezug auf seine Benutzeroberfläche und seine Funktionalität.
Allein durch Betrachten der einzelnen Grafikfenster lässt sich grob eine Reihenfolge der Implementierung ablesen. Während die GUI der 'General-View' und der Sequenzfenster noch sehr aufgeräumt wirkt, ist sie in der 'CDS-View' und in der Genansicht weniger ausgefeilt. Slider sind unverhältnismäßig lang, und Buttons haben nicht die gleiche Größe und sind teilweise nicht bündig angeordnet. Das macht für die Funktion natürlich zunächst keinen Unterschied, schmälert aber doch den Gesamteindruck.
Weiterhin gibt es einige kleine Bugs. Das Popup-Fenster der Regionsansichten verschwindet nicht, wenn der Cursor die Grafik zur Seite oder nach unten verlässt, die Beschriftungen der y-Achse werden rundungsbedingt immer ungenauer, je kleiner der insgesamt überstrichene Wertebereich ist, eine Änderung der Farben wird von offenen Grafiken zum Teil erst bei einer Navigationsbewegung übernommen und ein Abbruch des 'Datei Öffnen'-Dialogs führt im Falle der 'General-View' zu einem Teilabsturz.
Neben diesen Bugs gibt es auch einige Funktionalität, deren Fehlen schon beim ersten Benutzen auffällt, wie zum Beispiel eine Loop-Funktion für den Sound der Spektrum-Sonifikation oder die Möglichkeit, verschiedene Einstellungen sowie die zugehörigen Clips der Spektrum-Ansicht zu Vergleichszwecken zwischenzuspeichern. Auch ein konsistentes System der Fehlermeldung und ein Fortschrittsbalken beim Laden von Dateien wären wünschenswert.
Vor allem die Sonifikationen im Programm sind extrem rudimentär und böten viel Raum für Erweiterung. Das ist für ein Projekt, dessen eigentliche Intention die Sonifikation ist, kein gutes Zeugnis. Allerdings musste in die Visualisierung und insbesondere in die diversen, datenbezogenen Funktionen bereits derart viel Arbeit investiert werden, dass für eine Ausgestaltung der Sonifikationen keine Zeit mehr blieb.
Positiv anzumerken ist zum Ersten, dass alle von Fr. Maïda geforderte Funktionalität umgesetzt wurde. Auch die selbst gemachten Vorgaben wurden weitgehend erfüllt. In den Grafiken kann man flüssig und schnell navigieren, und die Bedienelemente sind hinreichend groß, um dem User nicht allzu präzise Cursor-Steuerung abzuverlangen, aber reduziert genug, um der eigentlichen Grafik möglichst viel des verfügbaren Raums zu geben. Alle Fenster sind präzise und nach einem einheitlichen System betitelt und orientieren sich an einem einheitlichen Design.

Der zweite zu diskutierende Aspekt, die Implementierung selbst, ist stark verbesserungsfähig. Der Code an sich ist gut formatiert, und die Methoden und Variablen sind weitgehend

selbsterklärend benannt. Die Kommentierung des Codes beschränkt sich jedoch auf ein absolutes Mindestmaß und wird bei einer Arbeit mit dem Code nur wenig Hilfestellung leisten. In den Klassen befinden sich eine Reihe veralteter Funktionen, die jedoch nicht als solche markiert sind, und im Projekt sind einige nicht verwendete Klassen enthalten.
Die Struktur des Programms läuft streckenweise sowohl dem objektorientierten Ansatz als auch einer Modell-View-Controller-Architektur zuwider. Die Datenkapselung ist zugunsten einer schnelleren Programmierung ignoriert worden, und die Deklaration von Variablen als `private` oder `public` erfolgte ohne größere Reflexion. Weiterhin ist die Abstraktion von Variablen und Methoden innerhalb des Programms unzureichend, was zu einigen Code-Doppelungen geführt hat. Insbesondere die rein datenbezogenen Methoden wie die Auswertung in Bezug auf Kopien, die Ermittlung gemeinsamer Grenzen mitsamt deren farbiger Darstellung und die Kompilierung der neuen Dateien sind nach dem Prinzip 'so-schnell-wie-möglich' geschrieben worden, da hier von Anfang an davon auszugehen war, dass diese Methoden nie wieder und für niemanden eine Relevanz haben oder gar weiterentwickelt werden. Aber auch zentralere Klassen wurden zum Teil 'on-the-fly' geschrieben, was zu umständlichen, schwer nachzuvollziehenden Interdependenzen zwischen Klassen geführt hat. Ein besonders negatives Beispiel ist die Achsenklasse, insbesondere die Beschriftung. Hier wurde sehr früh ohne jegliche Planung eine erste Version geschrieben und dann im weiteren Verlauf nur noch modifiziert. Die Folgen sind die rundungsfehlerbehafteten Labels, die verwirrende Methodenschachtelung innerhalb der Achsenklasse und die umständliche Kommunikation mit der Grafik-Klasse.
Das oben beschriebene Gerüst der Grafiken ist zwar weitgehend einheitlich umgesetzt, jedoch gibt es auch davon Abweichungen. Die erste Version der 'General-View' war in Teilen noch anders strukturiert und beinhaltete einige Funktionen zur Pufferung der Grafikausgabe. Im späteren Verlauf wurden diese durch effektivere Methoden unnötig gemacht, jedoch sind Teile der alten Struktur immer noch im Code enthalten. Das sieht man unter Anderem daran, dass bei der 'General View' ein anderer Info-Pool verwendet wird als bei den restlichen Grafiken.
Dennoch ist die Systematik soweit gediehen, dass sie die Erstellung der später hinzugefügten Grafiken (Codon-Sequenz, Genansicht) deutlich erleichterte. Ebenfalls als sehr praktisch erwiesen sich die relativ zu Beginn der Arbeit erstellten Navigationsmodule, die nur sehr wenig geändert werden mussten und die Programmierung sehr beschleunigten.

Obwohl also nicht völlig durcheinander, gibt es in Bezug auf die Implementierung einigen Raum für Verbesserungen. Der Hauptgrund für diese vielen Unzulänglichkeiten ist die Tatsache, dass ich zu Beginn des Projektes noch nie eine einzige Zeile JAVA-Code produziert hatte und sich meine Programmiererfahrung auf einen Grundkurs in C sowie einige Matlab-Projekte im Rahmen des Studiums beschränkte. Größere Projekte wie dieses waren komplettes Neuland für mich. Angesichts des sehr umfangreichen Projekts ging es mir darum, schnell funktionierende Teile zu erstellen. Vor der Entscheidung, eine grafische Darstellung zu programmieren oder zunächst eine eingehende Analyse der beteiligten Parameter vorzunehmen, entschied ich mich für das Erste, wobei angesichts meiner geringen Erfahrungen keineswegs gegeben ist, dass eine Analyse zu besseren Ergebnissen geführt hätte.
Mit Fortschreiten der Arbeit wuchs mein Verständnis, und die Implementierung wurde strukturierter. Die Beseitigung von Altlasten habe ich jedoch immer zugunsten neuer Funktionalität hintan gestellt.
Mit Vorliegen der Visualisierungen wurden einige Probleme der Eingangsdaten offenbar, die vorher weder mir noch Fr. Maïda bewusst gewesen waren. Beispiele dafür sind die nicht

markierten Genversionen und -kopien oder die Tatsache, dass die Wahrscheinlichkeiten aller Basen immer 1 ist, auch wenn unbekannte Basen in der Region vorhanden sind. Diese neuen Kenntnisse und Fr. Maïdas spezielle Wünsche hinsichtlich der Darstellung wie die farbige Darstellung von Regionsgrenzen im Infofenster führten dazu, dass immer mehr Programmierzeit in längliche, aber anspruchslose Routinen investiert werden musste, die mit Sonifikation nichts zu tun hatten, was zu einer zunehmenden Demotivation meinerseits führte. In der Folge sind diese Code-Teile nach dem Motto 'möglichst schnell funktional' programmiert worden. Ob dieses Vorgehen komplett negativ zu bewerten ist, ist angesichts der gegen null tendierenden Relevanz dieses Codes für die Welt und des bestehenden Zeitdrucks - Fr. Maïda hatte einen Abgabetermin für ihre Komposition - fraglich.

Realistisch gesehen wird dieses Programm wohl nie erweitert werden. Teile des Codes und des gewählten Layouts sind zwar durchaus brauchbar, eine strukturierte Neuimplementierung der Sonifikationen, die auf die standardisierten, für alle zugänglichen Daten zugeschnitten ist, ist aber zweifelsohne vielversprechender.

6.2 Zur Zusammenarbeit im Rahmen des Projekts

Die Zusammenarbeit zwischen Fr. Maïda und mir begann mit einem ersten Treffen. Fr. Maïda stellte mir die zu dem Zeitpunkt vorliegenden Dateien vor (W-General und S-General) und die Funktionalität des Programms wurde abgesteckt. Dabei lag der Fokus zunächst auf der Visualisierung. Fr. Maïda hatte hier sehr konkrete Vorstellungen, die sie als Bleistiftskizzen fixiert hatte (siehe Abbildung 25). Hinsichtlich der Sonifikationen war zu dem Zeitpunkt lediglich das Konzept des Spektrums aus den Regionslängen vorhanden.

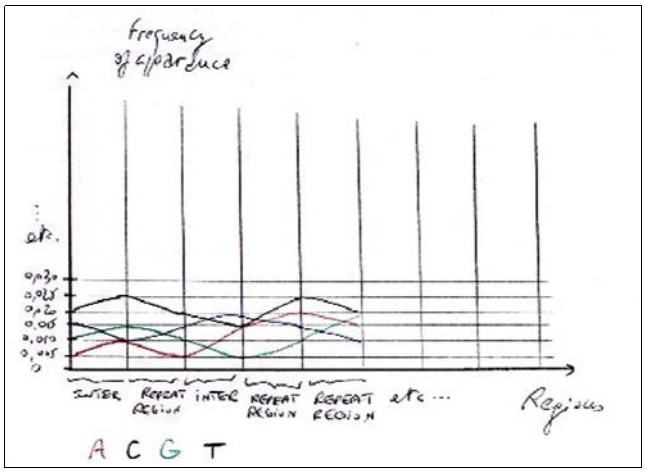

Abbildung 25: Skizze von Fr. Maïda zur Darstellung der Wahrscheinlichkeiten der Basen in den Regionen

Es fanden im weiteren Verlauf noch zwei Treffen sowie ein ausgedehntes Telefonat statt, im wesentlichen erfolgte die Kommunikation aber über E-Mails. Sprache der Wahl war Englisch, was die Kommunikation um einiges erschwerte und zu zahlreichen sprachbedingten Missverständnissen führte.
Die Zusammenarbeit war, neben dieser Sprachbarriere, von zwei großen Problemen geprägt. Das erste war das Programm selbst. Während des gesamten Prozesses hat mindestens ein Drittel aller implementierten Funktionen auf Fr. Maïdas Computer nicht richtig funktioniert. Dieses Problem zog sich von der Anzeige über das Einlesen oder Schreiben einer Datei bis hin zur Klangsynthese durch alle Bereiche des Programms. Bei nahezu jeder neuen Version war also Fehlersuche notwendig, die dadurch deutlich erschwert wurde, dass viele der Probleme spezifisch auf Fr. Maïdas MacBook auftraten, während das Programm auf Windows-Rechnern und auch deutschen MacBooks einwandfrei funktionierte. Die Suche musste also direkt auf Fr. Maïdas Computer erfolgen[85]. Das war ebenfalls nicht problemfrei. Das zur Analyse notwendige JAVA-Log ist auf einem Mac über die 'Console' einsehbar. Das Öffnen dieses Fensters auf Fr. Maïdas Computer nahm teils zehn Minuten in Anspruch, manchmal reagierte das Fenster auch gar nicht. Bei einem Treffen machte uns Andre Bartezki dankenswerterweise auf die Möglichkeit aufmerksam, das JAVA-Log über die Programmzeile aufzurufen, was auch gut funktionierte. Allerdings notierte sich Fr. Maïda die dazu nötige Kommandozeile fehlerhaft und teilte bei nächster Gelegenheit mit, das Log würde bei ihr nicht funktionieren; eine Missverständnis, das ca. einer Woche und mehrerer E-Mails zur Klärung bedurfte.
Die Gründe, aus denen Programmteile nicht funktionierten, waren ebenfalls vielseitig und gehen nur zum Teil auf die verschiedenen Plattformen zurück. In zwei Fällen fanden sich in der W-General-Datei 'ungewöhnliche' Zeilen[86], die in der gesamten Datei je nur einmal auftraten, in der korrespondierenden S-General-Datei aber nicht enthalten waren. Die Parsing-Routinen waren auf diese Art von Fehler nicht vorbereitet, da die Dateien per Augenschein eine feste Struktur hatten, und auch seitens Fr. Maïdas Kontakt keine genauere Spezifikation erfolgte. Das Ergebnis waren fehlerhafte Anzeigen, die allerdings nur bei genauem Vergleich der Grafiken mit den Dateien zu entdecken waren. Eine klarere Kommunikation bzw. Spezifikation der Dateien seitens Fr. Maïdas Kontakt hätte hier viel Zeit sparen können.
Fr. Maïda hat, um sich zu orientieren und ihre Ideen zu konkretisieren, manuell mit den Dateien gearbeitet. Dabei hat sie Notizen eingefügt, die sie anschließen nicht wieder komplett entfernt hat. In einem Fall hat sie zudem versehentlich einen Zeilenumbruch gelöscht. Diese Änderungen führten ebenfalls wieder auf ihrem Computer zu fehlerhafter Darstellung (Sequenzansicht und Regionsansicht stimmen nicht überein) oder Programmabstürzen aufgrund unabgefangener 'Exceptions'. Gerade in diesem Fall dauerte es sehr lange, bis die Fehlerquelle identifiziert war, da meine Dateien nicht verändert waren und ich gleiches für die Ihrigen annahm.
Die Fehlersuche insgesamt war sehr mühsam. Insofern das JAVA-Log keinen Aufschluss gab, mussten manuell Ausgaben ins Programm integriert werden. Eine neue Programmversion musste also zu Fr. Maïda geschickt werden, die ihrerseits einen Log-Ausdruck zurückschicken musste. Diese Art der Ferndiagnose war sehr enervierend und eine Problemsuche vor Ort hätte viel Zeit sparen können. Gegen Anfang des Projekts hatten wir ein derartiges Debugging-Treffen, das Fr. Maïda am Ende aber mit den Worten kommentierte, sie hätte keine Zeit für so

85 Eine Fehlerquelle waren beispielsweise die französischen Pfadnamen auf Fr. Maïdas Computer.
86 In einer älteren Version von 'W-General' fand sich einmalig '_____', in einer aktuellen Version findet sich einmalig ' ? ? ?' anstelle der Zeile mit den Auftretenswahrscheinlichkeiten. Die Erklärung ist vermutlich, dass in der betreffenden Region nur unidentifizierte Basen vorliegen.

etwas und ich solle doch solche Probleme auf meinem Computer lösen – eine Forderung, die angesichts beschriebener Probleme realitätsfern war.

Die zweite große Problematik der Zusammenarbeit war die zunehmende Kollision von Interessen. Auf der einen Seite stand der Wunsch Fr. Maïdas nach einem perfekt auf ihre Bedürfnisse und Prioritäten angepassten Programm, auf der anderen Seite mein Wunsch nach wissenschaftlicher Konsistenz und einem eventuellen Nutzen des Tools über Frau Maïdas Arbeit hinaus. Diese beiden Vorstellungen waren mit Fortschreiten der Arbeit immer weniger miteinander zu vereinbaren.
Wissenschaftliche Konsistenz war, ihrer eigenen Aussage gemäß, durchaus auch Fr. Maïdas Bestreben, doch gab es mehrere Situationen, in denen ihre Entscheidungen dieser Prämisse klar widersprachen. Das beste Beispiel ist die bereits erläuterte Problematik der Genversionen und der 'Nested Genes', die anhand der gelieferten Daten nicht voneinander zu trennen sind (vgl. 5.2.3). Ich habe Fr. Maïda mehrfach auf diese Problematik aufmerksam gemacht. Obwohl durchaus wissenschaftlich belegt ist, dass auch im menschlichen Genom derlei Gene auftreten ([131]), bestritt sie mit dem Hinweis, ich solle doch noch einmal die Grundlagen durchlesen, zunächst deren Existenz und verlegte sich angesichts vorgelegter Belege darauf, das Phänomen als vage Hypothese zu betrachten, die man zugunsten einer Fokussierung auf allgemein anerkanntes Wissen doch besser ausblenden sollte.
Mit Fortschreiten der Arbeit wurde auch immer klarer, dass das Programm abseits der Arbeit Fr. Maïdas keinerlei Nutzen haben würde. Es ist zu spezialisiert, teilweise wissenschaftlich falsch und enthält eine Reihe von irreführenden Darstellungen. Ein Beispiel dafür ist die Farbigkeit der CDS-Grenzen im Informationsfenster (vgl. Abbildung 26). Ob eine Grenze farbig ist oder nicht, hängt davon ab, ob dieselbe Base in einer oder mehreren anderen Version desselben Gens nochmals als Grenze auftritt. Diese Art der Darstellung kann zu Fehlinterpretationen führen. Sind nämlich sowohl die Anfangs- als auch die Endbase eines Exons grün[87] dargestellt, könnte man annehmen, dass das Exon so in der Form in einer anderen Genversion noch einmal auftritt. Dies ist aber nicht zwingend der Fall, da die Anfangsbase und Endbase in einer Version zu einem, in einer anderen Version zu zwei verschiedenen Exons gehören können.

Join : 1	3494930	-	3495283	Bases:	354
Join : 2	3495891	-	3495968	Bases:	78
Join : 3	3496097	-	3496208	Bases:	112
Join : 4	3496310	-	3496454	Bases:	145
Join : 5	3496561	-	3496612	Bases:	52

Abbildung 26: Beispiel für die farbige Darstellung von Regionsgrenzen

Fr. Maïda forderte des Weiteren auch eine farbige Darstellung der Längen. Die Länge eines Exons sollte immer genau dann rot dargestellt werden, wenn mindestens eine der Grenzen des Exons auch in einer weiteren Version desselben Gens mindestens einmal als Grenze eines beliebigen Exons fungiert. Es fällt einerseits schwer, sich vorzustellen, welchen praktischen Nutzen diese Information wohl haben könnte und was sie wertvoller macht als beispielsweise die Information darüber, ob ein Exon in mehreren Versionen unverändert auftritt. Das eigentliche Problem mit der Wahl der Farbe der Länge nach eben beschriebener Vorgabe ist aber ihre einhundertprozentige Redundanz. Die Länge wird nämlich immer genau dann rot

87 Tritt eine Base nur einmal als Grenze auf, so ist sie schwarz, bei 1 – 3 mal ist sie grün und bei mehr als drei mal blau.

sein, wenn mindestens eine der Grenzen grün ist und es liegt somit kein informationeller Mehrwert vor. Der Fakt jedoch, dass die Länge in Rot dargestellt wird, vermittelt den Eindruck, dass hier wichtige Information vorliegt. Eine farbige Darstellung der Grenzen an sich ist durchaus sinnvoll, doch wäre die Information ob ein Exon genau oder beinahe wiederholt wird objektiv erkenntnisreicher.
Für mich war besonders die eben beschriebene farbige Darstellung der Grenzen eine Quelle der Frustration. Die Funktionalität hat in meinen Augen wenig informationellen Wert, ist teils irreführend und hat zudem absolut nichts mit Sonifikation zu tun. Die zur Darstellung nötigen Informationen mussten zunächst aus den Daten extrahiert werden, was einer Reihe monotoner Vergleichsroutinen bedurfte. Da Fr. Maïda diese Form der Darstellung auch für das Informationsfenster der W-General-Datei erbat, mussten alle statistischen Auswertungen doppelt implementiert werden. Weiterhin hatten wir insbesondere hinsichtlich der Farbe der Exon-Längen ein großes Missverständnis, was dazu führte, dass einige Algorithmen komplett umsonst geschrieben wurden. Die Funktionalität im Ganzen hat circa zwei Wochen in Anspruch genommen und umfasst etwa 500 Zeilen Code.
Ein anderes ähnliches Beispiel einer an sich wenig anspruchsvollen, gleichwohl viel Zeit in Anspruch nehmenden Funktionalität, deren Nutzlosigkeit abseits des einen speziellen Anwendungsfalles von vornherein feststand und die zudem auch wieder wissenschaftlich fehlerhaft ist[88], sind die Textdateien, die das Programm speichert. Sie dienen als Input für ein OpenMusic-Patch, das Fr. Maïda eine weitere Arbeit mit dem Spektrum der Regionslängen in einer musikaffineren Umgebung erlauben soll. Hier soll nicht mehr auf alle Einzelheiten eingegangen werden, jedoch wurde auch da viel Zeit in relative profanen Code investiert, ohne dass dadurch im Hinblick auf Sonifikation genetischer Daten irgendein bleibender Nutzen geschaffen wurde.

Die Arbeit für Fr. Maïda war vor allem gegen Ende eine reine unbezahlte Dienstleistung und von einer Zusammenarbeit kann eigentlich nicht gesprochen werden, zumal Fr. Maïda wenig Zweifel daran ließ, dass das Künstlerische allein ihre Aufgabe sei. Zwar gab ihr Kunstprojekt den Anstoß für die systematische Beschäftigung mit dem Thema Sonifikation genetischer Daten, allerdings ist die Vorstellung, dass mit der auf reine Dienstleistungsaufgaben verwendeten Zeit und Energie auch ein weiterverwendbares, erweiterbares und auf die allgemein zugänglichen Daten zugeschnittenes Tool hätte entwickelt werden können, frustrierend.

88 Mit Betrachtung der zu dem Zeitpunkt bereits fertiggestellten Visualisierungen des Programms wurde Fr. Maïda bewusst, dass genetische Daten „ganz schön viele“ sind. Daher entschloss sie sich, nur noch jeweils die erste Version eines Gens zu berücksichtigen, wobei die erste Version die ist, die in der Datei als erstes steht.

6.3 ZUSAMMENFASSUNG UND FAZIT

Diese Arbeit beschäftigte sich mit der Sonifikation genetischer Sequenzdaten. Nach knappen Einführung in die Molekulargenetik wurde zunächst ein Überblick über das Gebiet 'Auditory Display' gegeben. Neben einer bloßen Darstellung wurde dabei auch versucht, einen Beitrag zur Suche nach geeigneten Definitionen zu leisten. Im Zuge dessen wurde auch erörtert, inwieweit künstlerische Arbeiten als Sonifikation bezeichnet werden sollten.
Anschließend wurden die bestehenden Ansätze der Sonifikation genetischer Daten in Wissenschaft und Kunst gesammelt und hinsichtlich ihrer Charakteristik und ihrer Gemeinsamkeiten analysiert.
Es folgte eine systematischen Betrachtung des Felds Molekulargenetik hinsichtlich der typischen Arbeitsbereiche und der verwendeten Daten und derer Dimensionen. Basierend auf den Ergebnissen dieser Betrachtungen wurden dann gedanklich sechs verschiedene, mögliche Sonifikationen entwickelt.
Anstoß zur Arbeit war ein Kompositionsprojekt von Clara Maïda, die für ihre Arbeit ein adäquates Programm benötigte. Vier der vorgeschlagenen Sonifikationen, die zum Teil auf Ideen von Fr. Maïda basieren, wurden im Zuge der Arbeit an dem Programm in einfacher Form implementiert.

Die ersten Versuche Sequenzdaten mittels Sound darzustellen stammen aus den 80er Jahren des vergangenen Jahrhunderts. Die darin verwendete Methode des Mappings von Basen oder Aminosäuren auf Tonhöhen im temperierten System wurde in den folgenden Dekaden immer wieder aufgegriffen und liegt der Mehrzahl aller Arbeiten zugrunde. Es gibt nur wenige wissenschaftliche Ansätze, von denen wiederum nur ein Teil mit Hinblick auf ein konkretes Anwendungsszenario entstanden ist. Ein Einsatz dieser Sonifikationen in der Praxis, wie er etwa bei EEG-Daten oder Börsenkursen zumindest vereinzelt belegt ist, findet bisher aber nicht statt. Die Zahl künstlerisch motivierter Ansätze ist insgesamt etwas größer und weist eine höhere Bandbreite hinsichtlich der verwendeten Transformationen auf. Den Autoren geht es oft darum, das Wunder des Lebens über die Musik gefühlsmäßig erfahrbar zu machen.
Zwischen beiden Feldern existieren nicht unerhebliche Schnittmengen hinsichtlich der Verfahrensweisen und der Ergebnisse. Oft sind Wissenschaftler auch an künstlerischen Umsetzungen beteiligt oder verwenden in wissenschaftlichen Publikationen eindeutig der Kunst zuzuordnende Termini wie 'Musik'.

Die Molekulargenetik ist eine extrem datenintensive Wissenschaft. Es gibt einer Vielzahl verschiedener direkter und indirekter Datendimensionen, die je nach Forschungsbereich von Interesse sind, eine Tatsache, die sich in den bisherigen Ansätzen zur Sonifikation kaum wiederfindet.
Die Darstellung der genetischer Daten erfolgt bisher weitestgehend visuell, doch wie die immer noch andauernde Entwicklung und Erprobung neuer Darstellungsmethoden und -konzepte zeigt, sind noch längst nicht für alle Anwendungsfälle optimale Formen gefunden.
Die in dieser Arbeit vorgeschlagenen Beispiele, der akustische DNA-Browser und der Transcript-Fingerprint, zeigen, ohne die Gesamtheit aller Möglichkeiten auch nur annähernd abzudecken, dass eine Einbeziehung des Audiokanals eine für die Darstellung fruchtbare, neue Perspektiven eröffnende Erweiterung sein könnte. Gerade die Tatsache, dass das Gebiet der Molekulargenetik noch relativ jung ist und nur wenige wirklich tradierte Darstellungsnormen

kennt, könnte für die Einbindung von Auditory Displays vorteilhaft sein, da einer Akzeptanz dieses 'Exots' weniger historisch gewachsene Konventionen im Wege stehen.
Data-Mining mittels statistischer Verfahren ist schon jetzt Teil der Molekulargenetik. Verschiedene Maße zum Informationsgehalt und zu Periodizitäten erlauben Rückschlüsse auf die Art und Funktion der untersuchten Sequenz. Durch die immer schneller und günstiger werdende Sequenzierung wird die vergleichende Genomik und die damit verbundene explorative Datenanalyse in Zukunft stark an Bedeutung gewinnen. Sonifikationen, die sich in anderen Feldern für eine bestimmte Aufgabe wie beispielsweise eine Clusteranalyse als zweckmäßig erwiesen haben, lassen sich auch im Bereich der Genetik applizieren, wie die Gegenüberstellung der Radial-Visualisierung mit den Modellen der MBS deutlich gezeigt hat. Ein einfaches, auf Audifikation beruhendes Beispiel zum Data-Mining ist der in dieser Arbeit vorgestellte Repeat-Finder.

Zusammenfassend kann man festhalten, dass die Einsatzmöglichkeiten von Auditory Display im Bereich der Molekulargenetik bisher bei weitem nicht ausgeschöpft sind. Insbesondere im Bereich der Darstellung sollte ein Einsatz des Audiokanals etwa in Form des hier vorgeschlagenen akustischen DNA-Browsers weiter verfolgt werden.

Im Hinblick auf die Zusammenarbeit von Wissenschaft und Kunst herrscht im Bereich der Sonifikation genetischer Daten eine große Nähe, eine Tatsache, die Robert Scott Root-Berstein als ein großer Kritiker der heutzutage als selbstverständlichen Trennung der Sphären Wissenschaft und Kunst sicherlich positiv bewerten würde.

> *The sciences and arts were once, not so very long ago, considered to be very similar, certainly complementary, and sometimes even overlapping ways of understanding the world. No longer. Today we accept such generalizations as that the sciences are objective, analytical, and rational whereas the arts are subjective, emotional, and based on intuition.*
> *([132] , p. 49)*

Auch diese Arbeit entspringt letztendlich einer Kooperation von Wissenschaft und Kunst. Diese Kooperation ist aber nur bedingt als erfolgreich zu bezeichnen.
Es ist einerseits sicherlich so, dass der von wissenschaftlichen Zwängen und Regeln freie Umgang mit Material sehr kreative Ergebnisse zur Folge haben kann, die sich, einmal entdeckt, keineswegs zwingend einer wissenschaftlich korrekten Applikation verweigern müssen.
Andererseits dient der strenge wissenschaftliche Rahmen, der Reproduzierbarkeit und Objektivität fordert, zumindest in bestimmten Gebieten elementaren Sicherheitsbedürfnissen. Eine Sonifikation zum Monitoring eines Atomkraftwerks muss die Daten nachweislich korrekt und unverändert Darstellen und vorhersagbare Wahrnehmungen erzeugen.
Eine direkte Kooperation von Wissenschaft und Kunst kann aufgrund der unterschiedlichen inhaltlichen und ästhetischen Zielstellungen, wie im Falle dieses Projekts sehr problematisch sein. Ein für beide Seiten akzeptabler Kompromis ist kaum zu erzielen, denn künstlerische Freiheit passt nicht in das enge, aber notwendige Korsett wissenschaftichen Arbeitens. Eine Offenheit der Wissenschaft – die Offenheit der Kunst für Inspiration aller Art ist fraglos vorhanden- , praktikable Techniken und Ideen aus der Kunst zu importieren, kann aber durchaus nützlich sein.

7. Literatur

[1] "Sonification Report: Status of the Field and Research Agenda", http://www.icad.org/websiteV2.0/References/nsf.html, pdf-Datei von 2010: http://digitalcommons.unl.edu/cgi/viewcontent.cgi?article=1443&context=psychfacpub&sei-redir=1#search=%22Soni%20ication%20Report%3A%20Status%20Field%20Research%20Agenda%22, zuletzt besucht: 17.9.2011

[2] Desmond S. T. Nicholl (2008): "An Introduction to Genetic Engineering", bei: CambridgeUniversity Press; ISBN: 978-0-521-85006-3

[3] Rolf Knippers (2006): "Molekulare Genetik", bei: Thieme; ISBN: 978-3-13-477009-4

[4] "The DDBJ/EMBL/GenBank Feature Table", http://www.ncbi.nlm.nih.gov/collab/FT/#7.2, zuletzt besucht: 26.9.2011

[5] Helen Pearson, "What is a gene?", 2006, in: Nature, Vol. 441, pp. 398-401

[6] Mark B. Gerstein et al., "What is a gene, post-ENCODE? History and updated definition", 2007, in: Genome Res. 17, pp. 669-681

[7] Applied Biosystems (Hg.), "DNA Sequencing by Capillary Electrophoresis - Applied Biosystems Chemistry Guide | Second Edition", 2009, in: Applied Biosystems, online: http://www3.appliedbiosystems.com/cms/groups/mcb_support/documents/generaldocuments/cms_041003.pdf

[8] Michael L. Metzker, "Sequencing Technologies - the Next Generation", 2010, in: Nature Reviews Genetics Vol. 11-1, pp. 31–46

[9] "True Direct DNA Measurement", http://www.helicosbio.com/Portals/0/Documents/Helicos%20tSMS%20Technology%20Primer.pdf, zuletzt besucht: 29.9.2011

[10] John Eid et al., "Real-Time DNA Sequencing fromSingle Polymerase Molecules", 2009, in: Science Vol. 323 Heft 5910, pp. 133–138

[11] Oxford Nanopore Technology, http://www.nanoporetech.com/, zuletzt besucht: 29.9.2011

[12] Kevin Davies; Clive Brown (Interview), "What Can Brown Do For Oxford Nanopore?", 2009, in: bei: Bio IT World; URL: http://www.bio-itworld.com/NGS-Brown.html, zuletzt besucht: 29.9.2011

[13] "Illumina and Oxford Nanopore Enter into Broad Commercialization Agreement", http://www.reuters.com/article/2009/01/12/idUS49869+12-Jan-2009+BW20090112, zuletzt besucht: 28.9.2011

[14] Robert S. Hotchkiss; Cheryl L. Wampler, "Auditorialization of Scientific Information", 1991, in: Proceedings of the 1991 ACM/IEEE, pp. 453–461

[15] Steven P. Frysinger, "A Brief History of Auditory Data Representation to the 1980s", 2005, in: Proceedings of the 11th International Conference on Auditory Display, pp. 410–413

[16] Gregory Kramer (Hg.) (1994): "Auditory Display", bei: Addison Wesley, ISBN: 978-0201626032

[17] Bruce N. Walker; Gregory Kramer, "Ecological Psychoacoustics and Auditory Displays: Hearing Grouping, Meaning Making", 2004, in: John G. Neuhoff "Ecological Psychoacoustics", Elsevier Academic Press, ISBN: 978-0125158510, pp. 150–175

[18] John G. Neuhoff et al., "Pitch and Loudness Interact in Auditory Displays: Can the Data Get Lost in the Map?", 2002, in: Journal of Experimental Psychology: Applied Copyright 2002 Vol. 8, No. 1, pp. 17–25

[19] Thomas Hermann, "Taxonomy and Definitions for Sonification and Auditory Display", 2008, in: Proceedings of the 14th International Conference on Auditory Display

[20] Alberto de Campo, "A Data Sonification Design Space Map", 2007, in: Proceedings of the 2nd International Workshop on Interactive Sonification

[21] Thomas Hermann, "Sonification for Exploratory Data Analysis", 2002, Dissertation, Universität Bielefeld

[22] Michael A. Nees; Bruce N. Walker, "Auditory Interfaces and Sonification", 2009, in: in: "The Universal Access Handbook" (2009) Constantine Stephanidis (Hrg.), CRC Press, ISBN: 978-0805862805, Preprint

[23] Ashley Walker; Stephen Brewster, "Spatial Audio in small screen device displays", 2000, in: Personal Technologies 4 (2), pp. 144-154

[24] Stephen Barrass; Gregory Kramer, "Using Sonification", 1999, in: Multimedia Systems (Multimedia Systems) 7-1, pp. 23–31

[25] Stephen Barrass, "Auditory Information Design", 1997, Dissertation, The Australian National University

[26] Elizabeth D. Mynatt, "Designing with Auditory Icons", 1994, in: Proceedings of the 2nd International Conference on Auditory Display, pp. 109–120

[27] Thomas Herrmann, Helge Ritter, "Listen to your Data: Model Based Sonification for Data Analysis", 1999, in: Advances in intelligent computation and multimedia systems (1999), pp. 189–194

[28] William W. Gaver, "The SonicFinder: a prototype interface that uses auditory icons", 1989, in: Human-Computer Interaction 4, pp. 67 - 94

[29] Meera M. Blattner et al. , "Earcons and Icons: Their Structure and Common Design Principles", 1989, in: Human-Computer Interaction 4-1, pp. 11–44

[30] David Worrall, "Using Sound to Identify Correlations in Market Data", 2009, in: in: Lecture Notes in Computer Science, 202-218, bei: Springer ISBN: 978-3642124389, pp. 202-218

[31] Alberto de Campo et al., "New Sonification Tools for EEG Data Screening and Monitoring", 2007, in: Proceedings of the 13th International Conference on Auditory Display, pp. 536-542

[32] Florian Dombois, "Using Audification in Planetary Seismology", 2001, in: Proceedings of the 7th International Conference on Auditory Display, pp. 227-230

[33] Florian Dombois, "Auditory Seismology: On Free Oscillations, Focal Mechanisms, Explosions and Synthetic Seismograms", 2002, in: Proceedings of the 8th International Conference on Auditory Display

[34] "Sound and Geographic Visualization", http://krygier.owu.edu/krygier_html/krysound.html, zuletzt besucht: 25.9.2011

[35] Thomas Herrman et al., "Vocal sonification of pathologic EEG Features", 2006, in: Proceedings of the 12th International Conference on Auditory Display, pp. 158-163

[36] Eoin Brazil (Hg.), Proceedings of the 11th International Conference on Auditory Display (ICAD2005), 2005

[37] "Georgia Tech Sonification Lab -- auditory graphs", http://sonify.psych.gatech.edu/research/auditorygraphs/, Sourcelist, zuletzt besucht: 25.9.2011

[38] Christian Dayé et al., "Sonification as a tool to reconstruct user's actions in unobservable areas", 2005, in: Proceedings of ICAD 05 - 11th Meeting of the International Conference on Auditory Display, pp. 328–331

[39] Petr Janata; Edward Childs, "Marketbuzz: Sonification of real-time financial data", 2004, in: Proceedings of ICAD 04-10th Meeting of the International Conference on Auditory Display

[40] Michael A. Nees; Bruce N. Walker, "Listener, Task and Auditory Graph: Toward a Conceptual Model of Auditory Graph Comprehension", 2007, in: Proceedings of the 13th International Conference on Auditory Display, pp. 266–273

[41] Philippe Lacherez et al. , "Overlapping Melodic Alarms Are Almost Indiscriminable", 2007, in: Human Factors 49-4, pp. 637–645

[42] Bruce N. Walker; David M. Lane, "Psychophysical Scaling of Sonification Mappings : A Comparision of Visually Impaired and Sighted Listeners", 2001, in: Proceedings of the 7th International Conference on Auditory Display, pp. 90–94

[43] Stephen Barrass, "Sonification Design Patterns", 2003, in: Proceedings of the 9th International Conference on Auditory Display, pp. 170–175

[44] Christopher Frauenberger et al., "Pattern Design in the Context Space - A Methodological Framework for Auditory Display Design", 2007, in: Proceedings of the 13th International Conference on Auditory Display, pp. 513–518

[45] Keith V. Nesbitt, "A Framework to Support the Designers of Haptic, Visual and Auditory Displays", 2005, in: GOTHI'05 - Guidance on Tactile and Haptic Interactions, pp. 54-64

[46] Eoin Brazil; Mikael Fernström, "Empirically Based Auditory Display Design", 2009, in: Proceedings of the SMC 2009 - 6th Sound and Music Computing Conference, pp. 7-12

[47] Stephen Barrass, "A Perceptual Framework for the Auditory Display of Scientific Data", 2005, in: ACM Transactions on Applied Perception, Vol. 2, No. 4, pp. 389–402

[48] Bruce N. Walker, "Magnitude Estimation of Conceptual Data Dimensions for Use in Sonification", 2002, in: Journal of Experimental Psychology: Applied Vol. 8, No. 4, pp. 211–221

[49] Alberto de Campo et al., "Designing a Generalized Sonification Environment", 2004, in: Proceedings of ICAD 04-10th Meeting of the International Conference on Auditory Display

[50] "SonEnvir - The SonEnvir Project", http://sonenvir.at/, zuletzt besucht: 25.9.2011

[51] David Worrall, "Overcoming the Software Inertia in Data Sonification Reseach using the SoniPy Framework", 2007, in: Proceedings of the Inaugural International Conference on Music Communication Science (ICoMCS), online: http://www.avatar.com.au/papers/SonipyICoMCS07.pdf

[52] David Worrall et al., "SoniPy: The Design of an Extendable Software Framework for Sonification Research and Auditory Display", 2007, in: Proceedings of the 13th International Conference on Auditory Display, pp. 445–452

[53] Andi Schoon; Florian Dombois, "Sonification in Music", 2009, in: Proceedings of the 15th International Conference on Auditory Display

[54] Volker Straebel, "The Sonification Metaphor in Instrumental Music and Sonification's Romantic Implecations", 2010, in: Proceedingsof the 16th International Conference on Auditory Display, pp. 287-294

[55] "Sonifyer wissen", http://www.sonifyer.org, zuletzt besucht: 21.9.2011

[56] Chris Hayward, "Listening to the earth sing", 1992, in: in: Gregory Kramer, "Auditory Display" (1994), bei: Addison-Wesley, ISBN: 978-0201626032

[57] "CMJ Reviews - International Conference on Auditory Display 2004: Listening to the Mind Listening", in: Computer Music Journal 29-1, URL: http://www.computermusicjournal.org/reviews/29-1/childs-icad04.html, zuletzt besucht: 30.9.2011

[58] Paul Vickers; Bennett Hogg, "SONIFICATION ABSTRAITE/SONIFICATION CONCRE` TE:AN ÆSTHETIC PERSPECTIVE SPACE FOR CLASSIFYING AUDITORY DISPLAYS IN THE ARS MUSICA DOMAIN", 2006, in: Proceedings of the 12th International Conference on Auditory Display, pp. 210–216

[59] Eduardo R. Mirinda; John A. Biles (2007): "Evolutionary Computer Music", bei: Springer, ISBN: 978-1-84628-599-8

[60] M.A. Garcia-Ruiz; J.R. Gutierrez-Pulido , "An overview of auditory display to assist comprehension of molecular information", 2006, in: Interacting with Computers 18-4, pp. 853–868

[61] "Gene Music and Sangen Studio", http://www.toshima.ne.jp/~edogiku/index.html, zuletzt besucht: 16.9.2011

[62] Susumo Ohno; Midori Ohno, "The all pervasive principle of repetitious recurrence governs not only coding sequence construction but also human endeavor in musical composition", in: Immogenetics Vol. 24 No. 2

[63] Ross D. King; Colin G.Angus, "PM - Protein music", 1996, in: Cabios Applications Note Vol. 12-3, pp. 251-252

[64] Sook Young Won, "Auditory Display of Genome Data: Human Chromosome 21", 2005, in: Proceedings of ICAD 05-Eleventh Meeting of the International Conference on Auditory Display

[65] Charlie Cullen; Eugene Coyle, "Rhythmic Parsing of Sonified DNA and RNA Sequences", 2003, in: ISSC: Irish Signals Systems Conference

[66] Rie Takahashi; Jeffrey H. Miller, "Conversion of amino-acid sequence in proteins to classical music: search for auditory patterns", 2007, in: Genome Biology 8:405

[67] Rie Takahashi; Jeffrey H. Miller, "Conversion of amino-acid sequence in proteins to classical music:search for auditory patterns", 2007, in: Genome Biology 8:405

[68] "Joel Sternheimer -- Protein Music -- French Patent # 2,136,737", http://www.rexresearch.com/sternheimer/sternheimer.htm, zuletzt besucht: 16.9.2011

[69] "Protein Music", Informationen zu Joel Sternheimers Theorie, http://www.bekkoame.ne.jp/~dr.fuk/, zuletzt besucht: 16.9.2011

[70] "Régulation Épigénétique de la Biosynthèse des Protéines Appliquée à la Culture de Fruits et Légumes:Compte-rendu d'Expérience en Jardin Potager", http://www.bekkoame.ne.jp/~dr.fuk/TomateFranceF.html, zuletzt besucht: 16.9.2011

[71] "What will happen to milking cows that listen to the radio?- the reason why they can do "good work" with favorite Mozart -",Informationen zu Joel Sternheimers Theorie, http:// www.bekkoame.ne.jp/~dr.fuk/Series5E.html, zuletzt besucht: 16.9.2011

[72] Christian Boisson, "Quand la musique soigne", 2010, in: Sud Ouest, Online: http://www.sudouest.fr/2010/09/20/quand-la-musique-soigne-189789-788.php, zuletzt besucht: 20.9.2011

[73] Pédro Ferrandiz, "Procédé de "Régulation Épigénétique de la Synthèse Protéique": Essai en Panification", 1993, in: Industries des céréales, No 85, online: http://www.bekkoame.ne.jp/~dr.fuk/Pain.Pedro.html

[74] "Une théorie basée à 100% sur l'hypnose... du lecteur confiant et naïf !", http://jacques.lavau.perso.sfr.fr/Theorie_fondee_sur_l_hypnose.html, zuletzt besucht: 16.9.2011

[75] Marc D. Hansen et al., "PROMUSE: A System For Multi-Media Data Presentation of Protein Structural Alignments", 1999, in: Pacific Symposium on Biocomputing 4, pp. 368–379

[76] Florian Grond et al., "Browing RNA-Structures by Interactive Sonification", 2010, in: Proceedings of ISon 2010, 3rd Interactive Sonification Workshop

[77] Douglas R. Hofstadter (1995 (1. Auflage 1979)): "Gödel, Escher, Bach: Ein Endloses Geflochtenes Band", bei: Klett-Cotta, ISBN: 978-3423300179

[78] Susumu Ohno, "Repetition as the Essence of Life on this Earth: Music and Genes", 1987, in: Haematol Blood Transfus (Haematology and blood transfusion) 31, pp. 511–518

[79] Wentian Lee, "Large-Scale Patterns in DNA Texts", 1999, in: Manuskript für: Scientific America (nicht veröffentlicht)

[80] James S. Walker; Amanda J.Potts, "Time-Frequency Spectra of Music", 2003, in: Advances in Analysis - Proceedings of the 4th International ISAAC Congress, pp. 487–493

[81] "White House Press Release", http://www.ornl.gov/sci/techresources/Human_Genome/project/clinton2.shtml, zuletzt besucht: 15.92011

[82] "DNA & Protein Music", http://www.nslij-genetics.org/dnamusic/, Sourcelist, zuletzt besucht: 15. 9. 2011

[83] "Gene2Music: Previous Work", http://www.mimg.ucla.edu/faculty/miller_jh/gene2music/previouswork.html, Sourcelist, zuletzt besucht: 15.9.2011

[84] "Genetic Music Sourcepage", http://www.whozoo.org/mac/Music/Sources.htm, Sourcelist, zuletzt besucht: 15.9.2011

[85] "Deamer on DNA-Music", http://www.youtube.com/watch?v=lEkcHhXC5i4, Youtube Video, zuletzt besucht: 15.9.2011

[86] "The Human Genome Project Presentation", http://www.charmian.com/toddbarton/index.html, Präsentation für das Smithsonian, National Museum of Natural History,Washington, DC, USA, zuletzt besucht: 15.9.2011

[87] Peter Gena; Charles Strom, "Musical Synthesis of DNA Sequences", 1995, in: Sixth International Symposium on Electronic Art, Montreal und in XI Colloquio di Informatica Musicale, Univeristà di Bologna

[88] Peter Gena; Charles Strom, "A Physiological Approach to DNA Music", 2001, in: The Proceedings of CADE 2001: the 4th Computers in Art and Design Education Conference

[89] "Genesis", http://www.ekac.org/geninfo.html, besucht: 04.11

[90] "A Striking Resemblance: DNA Dissiation as Rythmic Event", http://www.lazslo.com/dnaarticle, zuletzt besucht: 15.9.2011

[91] Alexander Mihalic, "DNA and composition", http://galileo.cincom.unical.it/Music/workshop/articoli/mihalic.pdf, zuletzt besucht: 12.05.2011

[92] "Molecular Music", http://www.molecularmusic.com/, zuletzt besucht: 15.9.2011

[93] Gary Burd, "Best of the web - www.molecularmusic.com", 2002, in: The Biochemist Vol. 24 No. 6

[94] "Our Sound Universe: the Music of Susan Alexjander, aka Susan Alexander", http://www.oursounduniverse.com/index.html, zuletzt besucht: 15.9.2011

[95] Susan Alexjander, David Deamer, "The infrared frequencies of DNA bases: science and art", 1999, in: Engineering in Medicine and Biology Magazine, IEEE 18:2, pp. 74–79

[96] "ProM --- Protein Music", http://amas.cz3.nus.edu.sg/music/index.html, zuletzt besucht: 15.9.2011

[97] "Genoma Music", http://www.genomamusic.com/inicio.html, zuletzt besucht: 15.9.2011

[98] Joe Davis et al., "Arts and Genetics", 2002 (Druckversion), in: Encyclopedia of Life Science

[99] Stephen H. Madoff, "The Wonders Of Genetics Breed A New Art", 26.5.2002, in: The New York Times

[100] André Brodyk, "Genetic Art & Culture", 2001, in: AIHLE 6th Annual Conference 2001

[101] Ingeborg Reichle, "Kunst im Zeitalter der technischen Reproduzierbarkeit des Menschen - Zur Rezeption der Gentechnik in der zeitgenössischen Kunst", 2001, von: kunsttexte.de ISSN: 1618 - 8101, http://edoc.hu-berlin.de/kunsttexte/download/bwt/reichle.pdf, zuletzt besucht: 15.9.2011

[102] "Our Sound Universe: the Music of Susan Alexjander, aka Susan Alexander" - Music Samples, http://www.oursounduniverse.com/music_samples.html, zuletzt besucht: 15.9.2011

[103] "Music of the Body", http://www.molecularmusic.com/musicofthebody.htm, Zitat, zuletzt besucht: 15.9.2011

[104] "Your DNA Song Ltd", http://www.yourdnasong.com/, zuletzt besucht: 15.9.2011

[105] Andy Coghlan, "Good Vibrations Give Plants Excitations", 1994, in: New Scientist Magazine Vol 142 Issue 1927

[106] Mihai Pop; Steven L. Salzberg, "Bioinformatics challenges of new sequencing technology", 2008, in: Trends Genet. 2008 March ; 24(3): 142–149

[107] "Genetic Sequence Data Bank August 15 2011", ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt, zuletzt besucht: 25.9.2011

[108] "Ensembl Genome Browser", http://www.ensembl.org/index.html, zuletzt besucht: 13.09.2011

[109] Meena Kishore Sakharkar et al., "Distributions of exons and introns in the human genome", 2004, in: In Silico Biology 4, 0032, pp. 387–393

[110] International Human Genome Sequencing Consortium, "Initial sequencing and analysis of the human genome", 2001, in: Nature Vol. 409, 860–921

[111] Michael Deutsch; Manyuan Long, "Intron–exon structures of eukaryotic model organisms", 1999, in: Nucleic Acids Research , Vol. 27, No. 15, 3219–3228

[112] Jacek Majewski; Jurg Ott, "Distribution and Characterization of Regulatory Elements in the Human Genome", 2002, in: Genome Research 2002-12, pp. 1827–1836

[113] James W. Fickett; Chang-Shung Tung, "Assessment of protein coding measures", 1992, in: Nucleic Acids Research, Vol. 20, No. 24, pp. 6441–6450

[114] Robertas Damaševičius, "Complexity Estimation of Genetic Sequences Using Information-Theoretic and Frequency Analysis Methods", 2009, in: Informatica, Vol. 20, No. 4, pp. 13-30

[115] Shrish Tiwari et al., "Prediction of probable genes by Fourieranalysis of genomic sequences", 1997, in: CABIOS Vol. 13 no. 3, 263–270

[116] Kaifu Chen et al. , "A novel DNA sequence periodicity decodes nucleosome positioning", 2008, in: Nucleic Acids Research, Vol. 36, No. 19, pp. 6228–6236

[117] Stephen T. Eskesen et al., "Periodicity of DNA in exons", 2004, in: BMC Molecular Biology 2004, 5:12

[118] Zhu-Jin Zhang, "DV-Curve: a novel intuitive tool for visualizing and analyzingDNA sequences", 2009, in: Bioinformatics Vol. 25 no. 9, pp. 1112–1117

[119] Yonghui Wu et al., "DB-Curve: a novel 2D method of DNA sequence visualization and representation", 2003, in: Chemical Physics Letters 367, pp. 170–176

[120] Chun-Ting Zhang et al., "The Z curve database: a graphic representationof genome sequences", 2003, in: Bioinformatics Vol. 19 no. 5, pp. 593–599

[121] Jonas S. Almeida and Susana Vinga, "Universal sequence map (USM) of arbitrary discrete sequences", 2002, in: BMC Bioinformatics 2002, 3

[122] H. Joel Jeffrey, "Chaos game representation of gene structure", 1990, in: Nucleic Acids Research, Vol. 18, No. 8, pp. 2163–2170

[123] Yinhe Cao et al., "Recurrence time statistics: Versatile tools forgenomic DNA sequence analysis", 2005, in: Journal of Computational Biology June 3, 677–696

[124] Deepak Sharma et al., "Spectral Repeat Finder (SRF): identification of repetitive sequences using Fourier transformation", 2004, in: Bioinformatics Vol. 20 no. 9, pp. 1405–1412

[125] Map Viewer, http://www.ncbi.nlm.nih.gov/projects/mapview/, zuletzt besucht: 13.09.2011

[126] Patrick Hoffman et al., "DNA Visual and Analytic Data Mining", 1997, in: VIS '97 Proceedings of the 8th conference on Visualization

[127] Clara Maïda, http://claramaida.com/, zuletzt besucht: 13.09.2011

[128] "UCSC Genome Browser", http://genome.ucsc.edu/index.html, zuletzt besucht: 13.09.2011

[129] Stephen B. Montgomery, "Sockeye: A 3D Environment for Comparitive Genomics", 2004, in: Genome Research 14, pp. 956–962

[130] Oliver R. Homann; Alexander D. Johnson, "MochiView: versatile software for genome browsing and DNA motif analysis", 2010, in: BMC Biology 2010, 8:49

[131] Peng Yu et al., "Nested genes in the Human Genome", 2005, in: Genomics 86

[132] Robert Scott Root-Bernstein, "The Sciences and Arts share a common creative Aesthetic", 1997, in: Tauber, Alfred I. "The elusive synthesis: Aesthetics and science" (1997), bei: Kluwer Academic Publ, ISBN: 978-0792347637

Anhang 1 - Visualisierungen

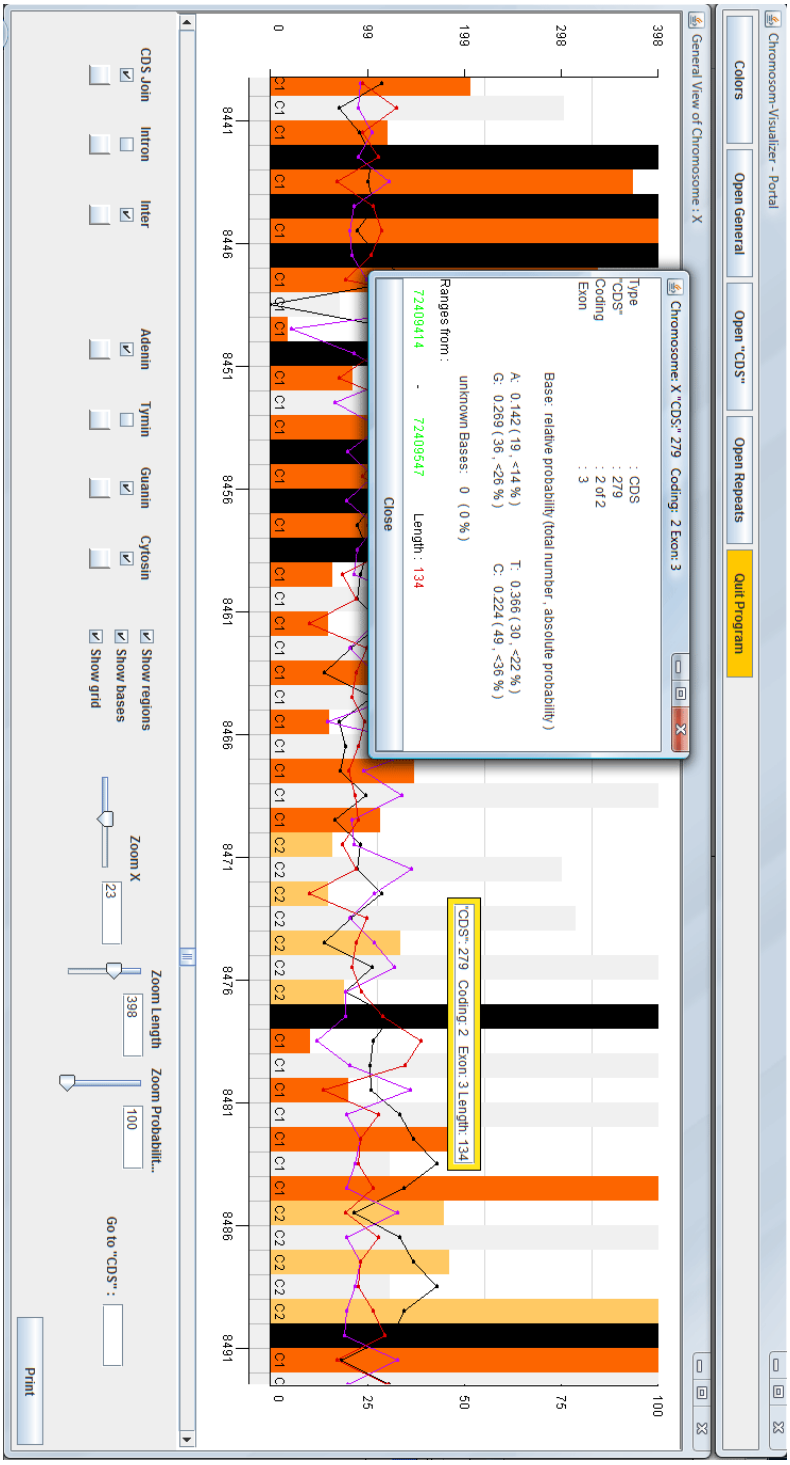

Abbildung 27: Screenshot der 'General-View' inklusive des Informationsfensters und der Menüleiste

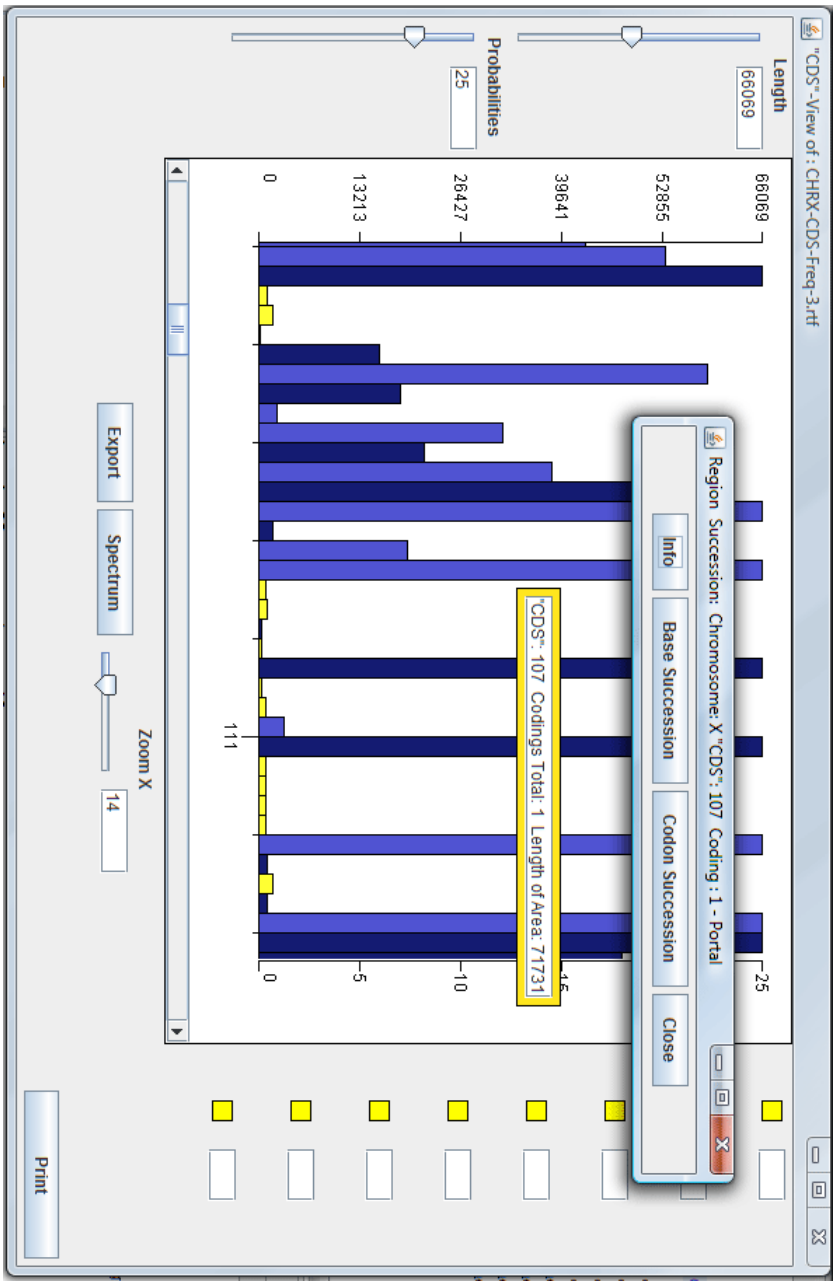

Abbildung 28: Screenshot der 'CDS-View' inklusive des Portals und des Info-Popups

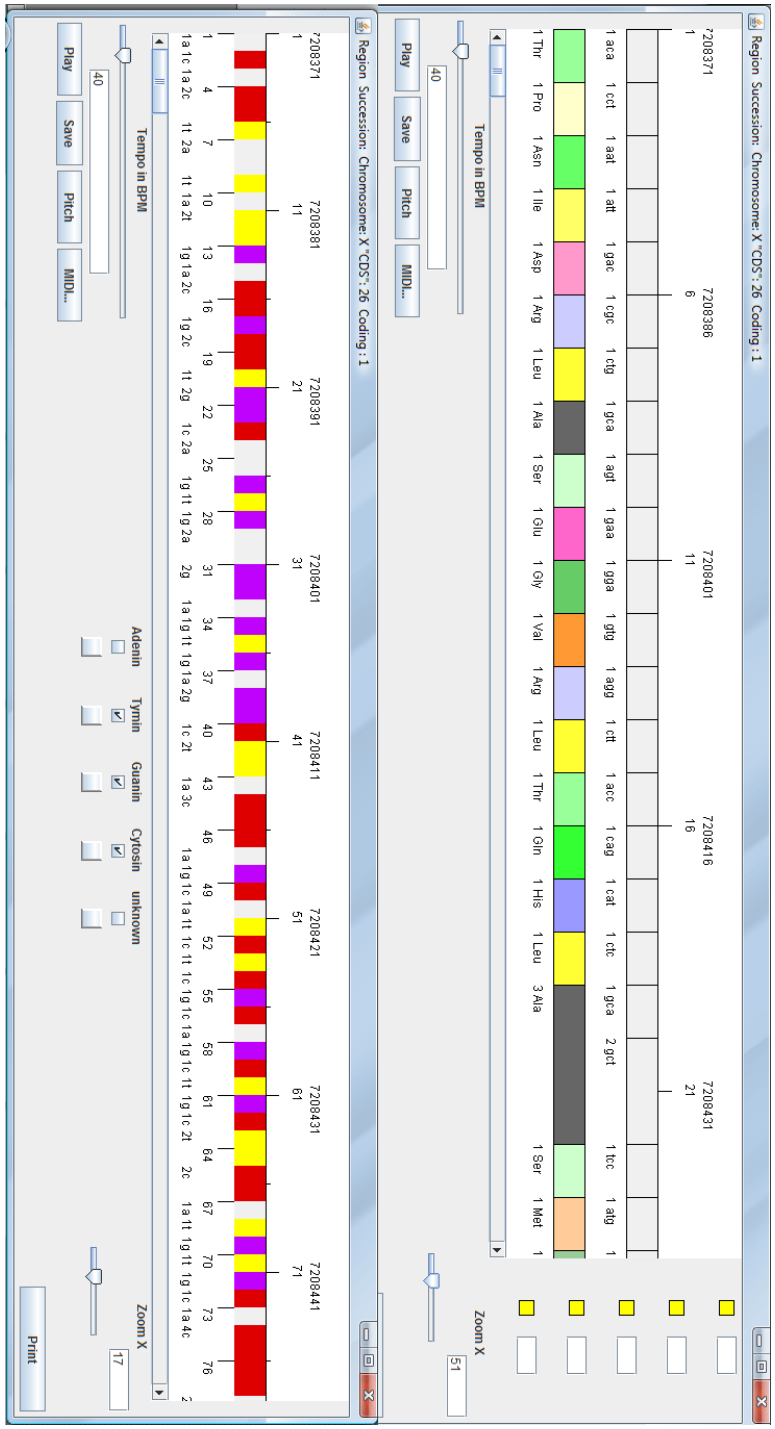

Abbildung 29: Screenshot der Sequenzansichten derselben Region für Basen und Codons.

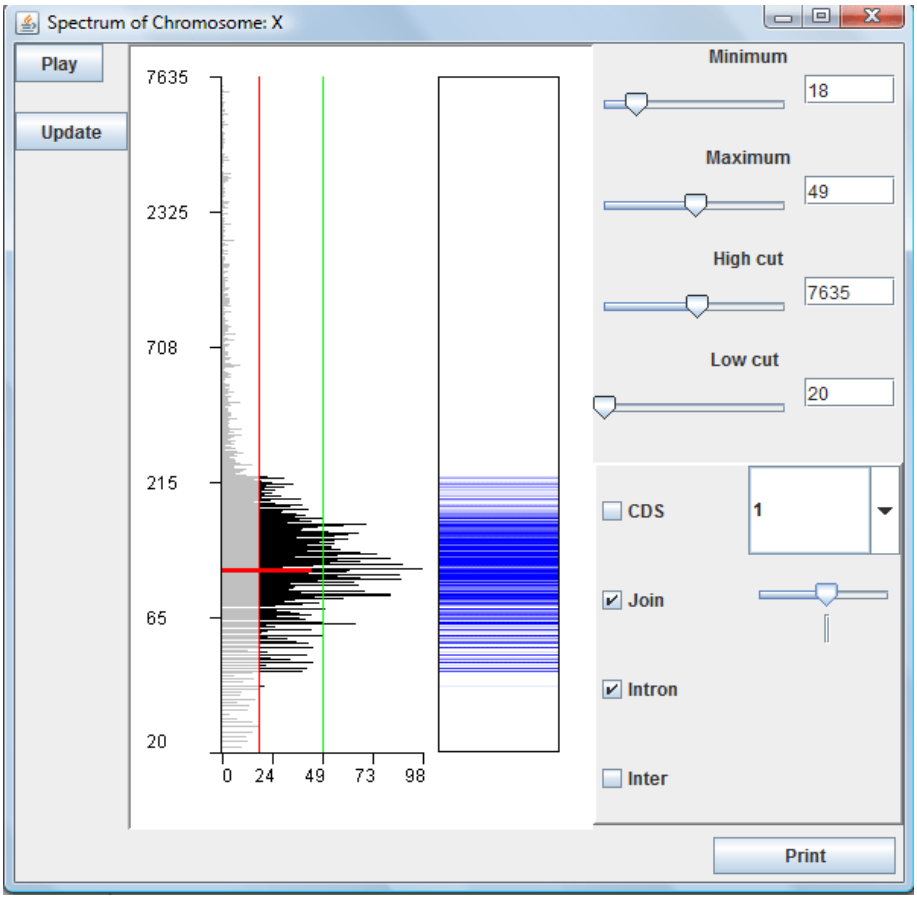

Abbildung 30: Screenshot der Spektrum-View.

Anhang 2 – Inhalt der CD-ROM

- Die Arbeit als Pdf-Datei (Magisterarbeit-Emanuel-Schuchart-214176.pdf)

- Citavi-Projekt-Magisterarbeit; in der Arbeit referenzierte Literatur haben eine rote Marke, auf der CD-ROM als Pdf vorliegende Dokumente haben eine Markierung.

- Audiodateien:

  Repeat-Finder:

  1) Repeat-Testinput-dry-Audio.wav
  2) Repeat-Testinput-Audio.wav
  3) Telomer-Y.wav
  4) Telomer-Y-Nur_C.wav

  Spektrum-Sonifikation:

  1) Fingerabdruck-Chromosom-X.wav
  2) Fingerabdruck-Chromosom-Y.wav

- Programm als lauffähige .jar-Datei (Programm-Fr_Maida-Final.jar)

- Source-Code des Programms

- Inputdateien für das Programm der Chromosomen X und Y des Menschen:

  1) CHRX-Frequ.rtf
  2) CHRX-Seq.txt
  3) CHRX-CDS-Freq-3.rtf
  4) CHRX-CDS-Seq.rtf

  5) CHRY-Frequ.rtf
  6) CHRY-Seq.txt
  7) CHRY-CDS-Freq-3.rtf
  8) CHRY-CDS-Seq.rtf

- Inputdateien für den Repeat-Finder

  1) Repeat-Testinput.txt
  2) Repeat-Testinput-dry.txt
  3) Telomer-Y.txt