N.T.M. 2017 · 25:459–483
DOI 10.1007/s00048-017-0179-2
Online publiziert 20. Oktober 2017

Big Data-Revolution oder Datenhybris?

Überlegungen zum Datenpositivismus der Molekularbiologie

Gabriele Gramelsberger

Big Data Revolution or Data Hubris? On the Data Positivism of Molecular Biology

Genome data, the core of the 2008 proclaimed big data revolution in biology, are automatically generated and analyzed. The transition from the manual laboratory practice of electrophoresis sequencing to automated DNA-sequencing machines and software-based analysis programs was completed between 1982 and 1992. This transition facilitated the first data deluge, which was considerably increased by the second and third generation of DNA-sequencers during the 2000s. However, the strategies for evaluating sequence data were also transformed along with this transition. The paper explores both the computational strategies of automation, as well as the data evaluation culture connected with it, in order to provide a complete picture of the complexity of today's data generation and its intrinsic data positivism. This paper is thereby guided by the question, whether this data positivism is the basis of the big data revolution of molecular biology announced today, or it marks the beginning of its data hubris.

Keywords: genome sequencing, automation, validation, human genome project, base-calling algorithms, big data

Genomdaten, Kernstück der 2008 ausgerufenen Big Data-Revolution der Biologie, werden voll automatisiert sequenziert und analysiert. Der Wechsel von der manuellen Laborpraktik der Elektrophorese-Sequenzierung zu DNA-Sequenziermaschinen und softwarebasierten Analyseprogrammen vollzog sich zwischen 1982 und 1992. Erst dieser Wechsel ermöglichte die Flut an Daten, die mit der zweiten und dritten Generation der DNA-Sequenzierer erheblich zunimmt. Doch mit diesem Wechsel verändern sich auch die Validierungsstrategien der Genomdaten. Der Beitrag untersucht beides – die Automatisierung und die damit verbundene Validierungskultur – um ein Bild der Komplexität der Datengenerierung und deren Datenpositivismus zu geben. Leitend ist dabei die Frage, ob dieser Datenpositivismus die Grundlage der aktuell angekündigten Big Data-Revolution der Molekularbiologie ist oder deren Datenhybris.

Schlüsselwörter: Gensequenzierung, Automatisierung, Validierung, Humangenomprojekt, *Base-calling* Algorithmen, Big Data

„Biologists are joining the big-data club"

… heißt es 2013 programmatisch in der Zeitschrift *Nature* (Marx 2013: 255) angesichts der wachsenden Datenmengen in der Molekularbiologie. Eine Graphik des Artikels zeigt eindrucksvoll den Anstieg der Datenflut

von wenigen Terabytes im Jahr 2008 auf 20 Petabytes bis 2012, darunter zwei Petabytes an Genomdaten (Marx 2013; Lander et al. 2001; Gandomi & Haider 2015). Allein in GenBank, der weltweit größten Datenbank für Gensequenzdaten, wurden bis August 2016 über 218 Milliarden Basen und 196 Millionen Gensequenzen gespeichert.[1] Bezieht sich der Begriff Big Data in erster Linie auf die schiere Quantität der Daten, so gehen die wachsenden Datenmengen doch zunehmend mit einem Wandel der Forschungspraxis einher. Sie ermöglichen es, Wissenschaft rein *in silico* zu betreiben. Für die Molekularbiologie bedeutet dies, Genomdaten automatisiert zu generieren und rein rechnerisch zu analysieren. Dies bringt einen neuen Typ von Biologen hervor, „who get neither their feet nor their hands wet" (Marx 2013: 260) und der dennoch immer mehr Daten und Informationen aus den bestehenden Genomdaten mit Hilfe avancierter Analysealgorithmen akkumuliert.

Diese Entwicklung hat das Interesse der Wissenschaftsforschung geweckt und umfangreiche Studien zu datengenerierenden Messmethoden wie den Genomsequenzierverfahren angeregt (Rabinow 1996; Kay 1988; García-Sancho 2012; Heather & Chain 2016), zur Datenpraxis im Umgang mit Datenbanken (Smith 1990; Leonelli 2012, 2014, 2016; Strasser 2012) sowie zur Nutzung von Analysealgorithmen und Simulationsmodellen zur Generierung von *in silico* Daten (O'Malley & Soyer 2012; Gramelsberger 2013). Der Wandel der Molekularbiologie vom Labor zum Computer ist jedoch nicht nur ein forschungspraktischer. Wie es in dem viel zitierten Artikel „The End of Theory" von Chris Anderson heißt:

> Petabytes [of data] allow us to say: "Correlation is enough". We can stop looking for models. We can analyze the data without hypotheses about what it might show. We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot. (Anderson 2008)

Damit verspricht die Big Data-Revolution eine neue Form induktiver Wissenschaft, deren Paradebeispiel der Webservice *Google Flu Trends* (GFT) ist. GFT ermöglicht anhand der Häufigkeit bestimmter Suchbegriffe Rückschlüsse auf Grippeepidemien. Dazu wurden in mehreren hundert Billionen Suchanfragen zwischen 2003 und 2008 nach Mustern der Häufigkeit bestimmter Suchbegriffe verteilt nach Regionen gesucht, sodass Wahrscheinlichkeitsaussagen über eine mögliche Grippeinfizierung in einer Region zu einem bestimmten Zeitpunkt getroffen werden konnten. Nach eigenen Angaben der Google-Forscher wurden diese Aussagen anhand staatlich erhobener Grippedaten durch die US-Seuchenschutzbehörde *Centers for Disease Control and Prevention* (CDC) evaluiert und es zeigte sich: „Whereas traditional systems require l–2 weeks to gather and

process surveillance data, our estimates are current each day“ (Ginsberg et al. 2009: 1014).

Doch was die Google-Forscher in *Nature* 2009 als Revolution der *Big Data Analytics* verkündet hatten (Ginsberg et al. 2009), erklärten David Lazer, Ryan Kennedy, Gary King und Alessandro Vespignani in ihrem 2014 erschienen *Science* Artikel „The Parable of Google Flu: Traps in Big Data Analysis“ zur Datenhybris. Eine Reihe von Fehlprognosen von GFT stellte den Big Data-Ansatz grundlegend in Frage. Die Datenhybris bestehe darin, so die Autoren, dass soziale Daten beliebige bis bizarre Korrelationen enthalten, beispielsweise hoch korrelierte Suchpaare wie „Grippe“ und „High School Basketball“. Das eigentliche Problem war jedoch das Overfitting des Ad-hoc-Vergleichs von 50 Millionen Suchbegriffen mit den Realdaten des CDC, um geeignete Vorhersagestrategien zu formulieren. „GFT developers, in fact, report weeding out seasonal search terms unrelated to the flu but strongly correlated to the CDC data, such as those regarding high school basketball. This should have been a warning that the big data were overfitting the small number of cases—a standard concern in data analysis“ (Lazer et al. 2014: 1203). So erstaunt es nicht, dass „even 3-week-old CDC data do a better job of projecting current flu prevalence than GFT“ (Lazer et al. 2014: 1204).

Die Datenhybris von GFT macht zweierlei deutlich: Zum einen, dass Vorhersagestrategien auf Basis korrelierter Daten durchaus Theorien enthalten, die als Datenmodelle in die Algorithmen integriert sind. Zum anderen stellt sich die Frage, ob hier überhaupt von „Daten“ gesprochen werden kann. Denn im Unterschied zu den CDC-Daten, die aus der Statistik konkreter Arztbesuche und -diagnosen resultieren, sind die Daten der sozialen Medien kaum objektivierbar. Wie Declan Butler bereits 2013 mutmaßte: „The press reports [about GFT] may have triggered many flu-related searches by people who were not ill […] a glitch attributed to changes in people's search behaviour“ (Butler 2013: 156). Andere Dienste zur Vorhersage von Grippewellen, die soziale Daten von Betreibern wie Twitter nutzen, haben das Problem, dass „most active Twitter users are young adults and so are not representative of the general public“ (Butler 2013: 156). Usergenerierte, informationsgeladene Daten sind tendenziös infiziert und wenig neutral. Forschung, die ihre Entdeckungen und Vorhersagen auf *glitches*, also Pannen, aufbaut, wäre jedoch der Horror jeder Naturwissenschaft.

Insbesondere letzterer Aspekt zeigt, dass der Begriff „Daten“ ontologisch indifferent verwendet wird. Vor diesem Hintergrund stellt sich die Frage nach möglichen Fallstricken der Molekularbiologie im Zeitalter von Big Data, die eine Datenhybris zur Folge haben könnten. Dies betrifft zum einen die Generierung von Genomdaten als Datengrundlage der Molekularbiologie, insofern DNA-Sequenzierer mit zunehmend komplexeren As-

semblern ausgestattet sind, also Algorithmen, die die Sequenzierung vollständig automatisieren. Diese ermöglichen eine immer schnellere Sequenzierung und sind damit für den Anstieg der Datenflut seit 2008 maßgeblich verantwortlich. Dies betrifft zum anderen die Analysealgorithmen, die aus eben diesen Genomdaten durch avancierte Analysestrategien immer mehr funktionelle Gendaten aus derselben Datengrundlage generieren. Genomdatenbanken erweisen sich aufgrund ihrer endlosen Re-Interpretierbarkeit als unerschöpfliche, *in silico* Datenreservoire (Leonelli & Ankeny 2010; Garcia-Sancho 2012; Leonelli 2016). Doch bioinformatische Forschung, die maßgeblich auf die Re-Interpretierbarkeit dieser Datenreservoire ausgerichtet ist, propagiert neben ihren eigenen Interpretationsproblemen die der Datengrundlage mit. Da der Zweck von Assembler-Algorithmen wie den Analysealgorithmen in der automatisierten Generierung und Interpretation umfangreicher Datenmengen besteht, ist das grundlegende epistemische Problem der bioinformatisch geleiteten Forschung – wie Roger Staden, einer der führenden Programmierer der ersten Sequenzierprojekte – bereits 1979 anmerkte: „One of the problems of using computers is that once the data has been filed away in the machine its quality is often forgotten: any data in the machine is taken to be correct although originally doubts may have existed" (Staden 1979: 2602).

Dieser Datenpositivismus ist Thema des vorliegenden Artikels. Um die epistemische Problematik nachvollziehen zu können, werden zunächst die Ursprünge der Automatisierung der DNA-Analyse in den 1980ern näher dargelegt. Die Geschichte der Gen-Sequenzierung, ihre wissenschaftlichen Medien sowie der durch den Computer induzierte epistemische Umbruch in der molekularbiologischen Forschung sind gut untersucht (Fox Keller 1995, 2000; Kay 1988, 2000; Rheinberger 2001b; Chadarevian 2002; Garćia-Sancho 2012).[2] Was jedoch fehlt, ist eine Analyse der informatischen Strategien, die den menschlichen Beobachter zunehmend automatisieren und gleichzeitig die Qualität der automatisierten Datengenerierung evaluieren. Denn aufgrund der vollständigen Automatisierung der DNA-Analyse stellt sich die berechtigte Frage, wie garantiert werden kann, dass die endlosen Reihen von As, Cs, Gs und Ts in den Datenbanken den tatsächlichen DNA-Sequenzen entsprechen. In der informatischen Antwort auf diese Frage liegt der Datenpositivismus und mit diesem eine mögliche Datenhybris begründet.

Automatisierung der DNA-Sequenzierung

Sanger-Methode der DNA-Sequenzierung

Zentral für die Geschichte der DNA-Sequenzierung sind die Forschungen des zweifachen Nobelpreisträgers Frederick Sanger (1918–2013), der in den 1970er Jahren mit der Kettenabbruchmethode das vorherrschende Verfahren der Sequenzierung entwickelte (Sanger et al. 1977; historisch rekonstruiert in Sanger 1988; Garcia-Sancho 2010). Die Idee, langkettige organische Moleküle (Polymere) nicht als Einheiten (Kolloide), sondern als Sequenzen einzelner und wiederkehrender Aminosäuren zu denken, verfolgte Sanger bereits in den 1940er Jahren, als er am Department of Biochemistry in Cambridge mit der analytisch-chemischen Erforschung von Insulin begann (Fruton 1972). Dabei verwendete er chemische Methoden, um einzelne Aminosäuren abzuspalten sowie chromatographische Methoden, um die Aminosäuren farblich sichtbar zu machen, und so die Sequenz von Insulin aufzuklären (Garcia-Sancho 2010: 11; Sanger 1949). Diese manuelle Aufklärung der Insulinstruktur gestaltete sich mühsam und dauerte bis Mitte der 1950er Jahre (Sanger 1959). Doch die aufwendigen Fragmentierungs- und Sortierungsstrategien – Auftrennen und chromatographisches Ordnen von Teilmolekülen nach Größen, biochemisches Markieren der Molekülenden sowie deren Sichtbarmachen durch Einfärbung – erlaubten es Sanger, dreißig Jahre später am Laboratory of Molecular Biology in Cambridge in modifizierter Weise DNA-Sequenzen zu bestimmen. Seine ersten Versuche waren dabei anfangs noch klassisch chemisch, doch dies änderte sich bald (Garcia-Sancho 2010; de Chadarevian 2002). Die Motivation lag in der materialen Problemstellung der Desoxyribonukleinsäure selbst begründet:

> Small sequences could be obtained, but at that time there was no way of locating them in the chains. The main problem with DNA was the very large size: the smallest pure DNAs that were available were the genomes of the single-stranded bacteriophages (such as øX174, which will be referred to as øX) of about 5,000 nucleotides, and these were rather large for testing out methods. Another difficulty was the absence of suitable degradative enzymes. […] However, the methods were both slow and laborious, and it seemed that if we were really going to be able to attack the vast sequences of genetic materials, then an entirely new approach was needed (Sanger 1988: 20).

Diese neue Methode bestand in der differenzierten Forcierung des Sequenzabbruchs durch die Replizierung mit Hilfe von Enzymen (Polymerase nach Berg et al. 1963; Wu & Kaiser 1968) eines Einzelstranges der DNA anhand geeigneter, nukleotidspezifischer Abbruchkriterien. „It was pos-

sible to prepare a C-specific digest, and similar digests to split at other residues. […] The ribo-substitution method worked reasonably well and we were able to determine a sequence of about 80 nucleotides" (Sanger 1988: 21). Die Ergebnisse wurden – und auch das war neu – durch Gel-Elektrophorese nach Größen separiert und durch radioaktive Markierung sichtbar gemacht. Durch die Gel-Elektrophorese wurde es möglich, Sequenzen von dreihundert Basenpaaren zu analysieren. Die Aufteilung in horizontale Linien sowie in vertikale Spalten für je ein Nukleotid (G, C, A, T) auf der Gelplatte machte den genetischen Code direkt ablesbar (vgl. Originalabbildung in Sanger et al. 1977) – ein Gelbild repräsentierte dabei ein klassisches Tabellenformat, das sich als Paradigma der Tabellierung in der relationalen Architektur der Datenbanken jener Jahre wiederfand (Codd 1970; Krajewski 2007).

Automatisierung der DNA-Sequenzierung

Das Ablesen der sich überlappenden Balken auf den Gelplatten mit dem bloßen Auge ermöglichte den Transfer der materiellen Informationen der Abfolge der vier Nukleotide in den semiotischen Raum des genetischen Codes von Buchstabenfolgen. An eben dieser Stelle setzte die Automatisierung der Datengenerierung ein. Doch es war nicht Sanger, sondern eine Gruppe von Caltech-Forschern um Leroy Hood, die Ende der 1970er Jahre die tabellarische Darstellung auf der zweidimensionalen Fläche der Gelplatten durch Aneinanderreihung der Balken linearisierten, jedes Nukleotid mit Fluoreszenzmethoden eigens farbkodierten und damit die datendifferenziellen Voraussetzungen schafften, die Sequenzen mit einem Laser einzulesen (Garcia-Sancho 2010). Was jedoch eingelesen wurde, waren nicht die überlappenden Balken als optisches Muster, wie es die Forscher mit bloßem Auge taten, sondern die durch den Laser angeregte Rückstrahlung der Farbkodierungen (*signal peaks*).[3] Messtechnisch bedeutete die Automatisierung der Sequenzierung folgendes: Zuerst wurden die DNA-Fragmente in einer Gelspur mit vier fluoreszierenden Färbemitteln gekennzeichnet; dann wurden mehrere DNA-Vorlagen des selben Genoms zu Vergleichszwecken in separaten Gelspuren angeordnet; schließlich wurden die fluoreszierendes Färbemittel durch einen Laser angeregt und dann die Emissionsintensität der Lasersignale auf vier Wellenlängen gemessen: „The laser and detectors scan the bottom of the gel continuously during electrophoresis in order to build a gel image in which each lane has a ladder-like pattern of bands of four different colors, each band corresponding to the fragments of a particular length" (Ewing et al. 1998: 175). Das Resultat war ein Messbild der DNA-Sequenzen in Form eines Gelbildes (vgl. Originalabbildung in Smith et al. 1986).

Doch dies war nicht das Ende, denn nun erst erfolgte die Computeranalyse des Messbildes, das heißt das messtechnische Gelbild wurde in mehreren Schritten in eine Chromatogramm-Datei umgewandelt: Zuerst wurden die Gelspuren voneinander separiert; dann wurde jedes Signals über sein Spurbreite summiert und daraus die „Spurdaten" (trace data) generiert, die durch gängige Signalverarbeitungsmethoden gefiltert und geglättet wurden; und schließlich wurde das so genannte *base-calling* mit Hilfe eines Algorithmus durchgeführt, der die Spurdaten in Basensequenzen übersetzte. Das Resultat war eine symbolische Darstellung der DNA-Sequenzen in Form eines Gel-Reads des DNA-Fragments.

Auch wenn die Auflösungsrate der Digitalisierung zu dieser Zeit erstaunlich gering war – „Typically, 8,000 points were taken for each of the four filter positions (total of 32,000 points) in a 13-h run; this rate of data acquisition gives 40 or more data points per DNA peak" (Smith et al. 1986: 676) –, so war der Automatisierungsschritt von der unmittelbaren, visuellen zur mittelbaren, signalverarbeitenden Datengenerierung nicht nur ein medientechnischer, sondern hatte epistemische Folgen. Dies zeigte sich auch an anderen Experimentaltechnologien. So beschreibt etwa Hans-Jörg Rheinberger (2001b) diesen Übergang für die Flüssigkeitsszintillation in den 1940er Jahren. Dabei handelte es sich um ein Messverfahren der Biologen, bei dem die durch die radioaktive α-Strahlung des Poloniums verursachten Lichtblitze gezählt wurden. Das Delegieren der Beobachtung an einen photoelektrischen Schaltkreis (Photomultiplier) sorgte dabei nicht nur für eine erste Flut an Daten, sondern hatte auch Folgen für die Experimentalkultur der Biologie, denn es ermöglichte, Experimentreihen automatisiert durchzuführen und so durch die Vergleichbarkeit der Daten deren Aussagegehalt zu verbessern. Die ersten DNA-Sequenzierer waren Ende der 1980er Jahre zwar noch weit entfernt von den „serial counts involving hundreds of samples" der Flüssigkeitsszintillation (Rheinberger 2001a: 158), doch dies änderte sich Anfang der 1990er Jahre mit der Miniaturisierung der Detektion durch die Kapillarelektrophorese. Damit wurde es möglich, bis zu 96 Proben parallel zu sequenzieren. Diese erste Generation der automatisierten Sequenzierer konnte DNA-Fragmente (Reads) bis zu 1000 Basen (1 Kilobase) handhaben und erzeugte pro Gerät täglich etwa 384 Kilobasen an Sequenzdaten. Gensequenzierung mit dieser ersten Generation DNA-Sequenzierer war aufwendig und teuer und wurde ausschließlich in den wenigen, sich konstituierenden Sequenzier-Zentren durchgeführt.

Informatische Strategien im Kontext der DNA-Sequenzierung

Shotgun-Methode

Eine Auflösung von tausend Basenpaaren war angesichts der immensen Länge der DNA-Sequenz eines Genoms nicht viel und dokumentierte das bereits von Sanger artikulierte Diskretisierungsproblem als „eternal one of fragmentation" (Sanger 1988: 20) – sowohl materiell als auch informationell: „Although DNA fragments larger than about 1.000 bp cannot always be stably inserted into the single-stranded phage vectors such as M13mp2, insertions greater than 350 bp in length are desirable in order to take full advantage of the resolving power of DNA sequencing gels" (Anderson 1981: 3015–3016). Wie sollten unter diesen Umständen die 4,6 Millionen Basenpaare des *Escherichia coli* Genoms, die 13 Millionen Basenpaare des Hefe-Genoms oder die 3 Milliarden Basenpaare des menschlichen Genoms entschlüsselt werden? Die Antwort lieferte der Computer. Bereits die Dekodierung des *Bacteriophagen øX174* hatte die Nutzung von Micro-Computern zur Datenauswertung erfordert (Staden 1979; Anderson 1981).[4]

> The whole of the DNA to be sequenced is shotgunned[5] into a suitable vector and cloned. Ideally the cloned fragments would be of at least 200 bases in length. The clones are then sequenced [by hand] and the computer used to collate the data. Collation involves searching for overlaps in the data. If the 5' end of the sequence from one gel reading is the same as the 3' end of the sequence from another the data is said to overlap. If the overlap is of sufficient length to distinguish it from being a repeat in the sequence the two sequences must be contiguous. The data from the two gel readings can then be joined to form one longer continuous sequence. To facilitate the search for overlaps all sequences derived from previous gel readings are stored in one master file. All new gel readings and their complements are compared with the sequences in the master file. Any sequences involved in overlaps are joined and new data is added into the Master file (Staden 1979: 2601–2602).

Bis zu 18.000 Nuklotide waren auf diese Weise Ende der 1970er Jahre mit Micro-Computern analysierbar. Die Shotgun-Methode schredderte im wahrsten Sinne des Wortes die zahlreichen Kopien der DNA-Vorlage in beliebige, kleine, sich überlappende Fragmente (Randomisieren). Aus den Überlappungen der Fragmente musste die ursprüngliche DNA-Abfolge der Sequenz rückgeschlossen werden (Rekombinieren). Dies erforderte umfangreiche Sequenzierungen für eine ausreichende Abdeckungsrate (Replizieren) des zu untersuchenden Genoms sowie die Analyse großer Datenmengen. Erschwerend kam jedoch hinzu, dass hohe Sequenzwieder-

holungsraten – im menschlichen Genom gut 50 Prozent, in Bakteriengenomen nur 1,5 Prozent – die Puzzlearbeit komplex gestalteten. Mit der Shotgun-Methode wurde zwar die Sequenzierung des menschlichen Genoms möglich, doch es liegt auf der Hand, dass dies nur durch den Wechsel vom menschlichen Experimentator zum Computer möglich wurde. Denn die neue Methode ersetzte die an der menschlichen Beobachtung ausgerichteten Strategien des Fragmentierens, Linearisierens und Sortierens durch bioinformatische Strategien des Replizierens, Randomisierens und Rekombinierens.

Automatisierung des Beobachters

Führend in der frühen Entwicklung von Computerprogrammen der Molekularbiologie war Roger Staden, der als Programmierer seit den späten 1970er Jahren eng mit Frederick Sanger am Laboratory of Molecular Biology in Cambridge zusammenarbeitete. Die frühen Computerprogramme von Staden – OVERLAP, XMATCH, FILINS – ähnelten dabei noch sehr der klassischen Papierarbeit mit Labortagebüchern, obwohl sich die Biologen zunehmend in „Operatoren" transformierten. Gel-Reads mussten manuell als Buchstabenfolge eingegeben, Übereinstimmungen aktiv vom Operator erfragt, Unsicherheiten in der Dekodierung per Hand definiert und die Zusammenfügung der Sequenzen initialisiert werden. „Note: it is often necessary to refer back to the experimental data to make these editing decisions" (Staden 1979: 2609). Da die Automatisierung der Gelbilder in Gel-Reads fehlerbehaftet war, musste die Datenqualität per Auge überprüft werden. Dazu entwickelte Staden Ende der 1970er Jahre einen *uncertainty code*, der Aussagen über mögliche Fehlinterpretationen als Einträge in eine Computerdatei erlaubte (Tab. 1).

Auf Basis dieses manuell einzupflegenden *uncertainty code* generierte das Programm einen *confidence value* (Konfidenzwert). Dieser Wert gab Auskunft über die Datenqualität, in welcher die Anzahl der Sequenzierungen pro Sequenz (Abdeckung) sowie der jeweilige *uncertainty code* einging. Aus diesen Konfidenzwerten ließen sich dann Bewertungen berechnen. „Well determined" entsprach dabei allen Werten gleich oder größer 75 Prozent Genauigkeit und daraus ergab sich eine Klassifikation der Datengenauigkeit:

1. Well determined on both strands and they agree. code = 0;
2. Well determined on the plus strand only. code = 1
3. Well determined on the minus strand only. code = 2
4. Not well determined on either strand. code = 3
5. Well determined on both strands but they disagree. code = 4 (Staden 1982: 4740).

Tab. 1 Uncertainty Code (nach Staden 1982: 4735)

Symbol	Meaning	
	probably	
1	″	C
2	″	T
3	″	A
4	″	G
D	″	C possibly
V	″	T ″
B	″	A ″
H	″	G ″
K	″	C ″
L	″	T ″
M	″	A ″
N	″	G ″
R	A for G	y
Y	C for T	y
5	A for C	
6	G for T	
7	A for T	
8	G for C	
-	A or G or C or T	
Else ″-		

Kurze Zeit später führte Staden für die Analyse der DNA-Daten die grundlegenden informatischen Konzepte der „Contigs“ (kurze Gel-Reads) und der „Konsensussequenz“ ein (1982).[6] Eine Konsensussequenz ist die durch Überlappung der Contigs automatisch generierte Sequenz, in deren Berechnung die Bewertungen der Konfidenzwerte eingeht. Mit dieser statistischen Darstellung der DNA-Sequenzen wurde es möglich, neu entschlüsselte Sequenzen quantitativ mit bereits bekannten Konsensussequenz zu vergleichen (Alignment). Ein Verfahren, das bis heute die Genomforschung charakterisiert.

Eliminierung des Beobachters

Noch wurden Datenlücken in der Sequenzierung materiell geschlossen (Anderson 1981: 3020), doch dies änderte sich in den darauffolgenden Jahren durch die Automatisierung jeglicher *editing decision* durch Algorithmen. Denn der Engpass in der Automatisierung der Datenanalyse war die visuelle Überprüfung der Gel-Reads. Der in der Einleitung skizzierte Datenpositivismus hat hier seinen Ursprung. Die Frage, welche Aussa-

gekraft nach diesem Wechsel von den Gelplatten und Gelfilmen zu den „trace representations of the band intensities“ (Dear & Staden 1972: 107) deren automatisierte Interpretation der Basensequenzen hatte, beschäftigte die Programmierer in den 1980ern und 1990er Jahren. In einem ersten Schritt definierte Staden mit seinem Kollegen Simon Dear ein „machine independent format for storing data derived from automatic sequencing instruments“ (Dear & Staden 1992: 107) für die prozessierten Tracedaten (Standard Chromatogram Format, SCF). Das SCF-Format wurde im Kontext von John Sulstons et al. *C. elegangs* Sequenzierungsprojekt entwickelt (Chadarevian 2002), das mit den ersten kommerziellen Sequenzieren arbeitete (ABI 373A und Pharmacia A.L.F.). SCF definierte Angaben, wie die Daten in einer Chromatogram-Datei aussehen sollten und welchen Konfidenzwert diese besitzen. In einem weiteren Schritt entwickelten beide für die *C. elegangs*-Sequenzierung ein Computerprogramm, das es den Forschern in den uneditierten Maschinendaten erlaubte, „to visually select left and right cutoff positions to denote the start and end of good data“ (Dear & Staden 1991: 3908).

Dieses Programm markierte den Übergang von der manuellen Evaluation der Gel-Reads anhand des *uncertainty code* zu einem computergestützten Verfahren. Noch weiter ging der kurz darauf entwickelte Consensus-Algorithmus von Staden und James Bonfield, insofern es sich um eine Software handelte, „to decide if conflicts between readings require human expertise to help adjudicate. […] so the time taken to check and edit a contig will be greatly reduced“ (Staden & Bonfield 1995: 1406). Damit verschob sich die angesprochene menschliche Expertise endgültig von der visuellen Kontrolle der Gelplatten anhand des *uncertainty code* auf die Kontrolle der codierten Sequenzen in den Dateien – allein durch den Befehl „next problem“ geleitet. Die epistemische Akteurialität bezüglich der Genauigkeit der Daten wurde damit komplett an die Software delegiert: Nicht der Forscher, sondern die Software „marks the changed character as having been edited by hand“ (Staden & Bonfield 1995: 1407). Schließlich erlaubten die *Estimating Sequence Quality* Methode von Bary Churchill und Michael Waterman (1992) sowie der *Phred Quality Score* von Brent Ewings und Phillip Green, „that human involvement in sequence data processing be […] eliminated“ (Ewing et al. 1998). Die visuelle Überprüfung wurde durch den statistischen Begriff der Genauigkeitsrelevanz ersetzt.

> We take the term accuracy to mean the probability that a given DNA base or sequence of bases in a finished sequence is identical to the corresponding base in the actual DNA molecule. We assume implicitly that there is a single true sequence and not a population of sequences being studied. The problems of errors that occur at the DNA

> preparation or cloning stages and of errors in data transcription are not addressed here. Systematic errors that occur in sequence determination are also beyond the current scope of these methods. Thus, the statistics described represent a bound on sequence accuracy (Churchill & Waterman 1992: 90).

Datenpositivismus – Big Data-Revolution oder Datenhybris?

Der Begriff des Datenpositivismus ist, in Anlehnung an Stadens eingangs zitierte Bemerkung von 1979, als informatisch registrierte, symbolische Markierung definiert (Datum), deren Existenz den Umstand ihrer Wahrscheinlichkeit verschleiert. Oder in anderen Worten: Was in Stadens *uncertainty code* noch eindeutig als „probably" und „possibly" definiert war, „to code for, and hence keep track of all types of uncertainty in the data" (Staden 1979: 2602; Tab. 1), wird zu einer Gesamtaussage wie „well determined", auf deren statistischer Relevanz (>75 %) als Datengrundlage datenanalytisch operiert wird, als ob diese eindeutig gegeben wäre. Diese statistische Datenbewertung ist für numerische Messwerte in Form von Fehlerunsicherheiten gang und gäbe, sie erhält aber im Kontext der Basensequenzen, die eine und nur eine automatisierte Übersetzung der messtechnischen Spurdaten in einen der vier Buchstaben der Nukleobasen nach Wahrscheinlichkeitsabschätzungen vornehmen, ein andere Bedeutung: am Ende steht ein konkreter Buchstabe.[7] Diese vermeintliche Exaktheit ist aus verschiedenen Gründen problematisch: Zum einen, weil bioinformatische Entdeckungen maßgeblich vom Vergleich mit bestehenden Konsensussequenzen und Referenzdatensätzen (Alignment) abhängen und deren Validierung alles andere als einfach und eindeutig ist. Zum zweiten, weil aktuelle Assembler-Algorithmen aus informatischen Gründen hohe Anforderungen an die Durchsatzraten der Sequenzierer stellen. Zum dritten, weil neue und schnellere Sequenzierergenerationen noch fehlerbehafteter sind als die Sanger-Methode. Dies zusammengenommen legt die Frage nahe, ob der Datenpositivismus der Molekularbiologie zu einer Big Data-Revolution führt oder ob sich hier nicht eine Datenhybris ankündigt.

Validierung von Genomdaten

„The future of human population genetics is […] rosy," konstatierten 2015 Ewan Birney und Nicole Soranzo und bezogen sich dabei auf den Abschluss des *1.000-Genom-Projekts* (Birney & Soranzo 2015: 53). Dieses internationale Projekt (2008 bis 2015) – Nachfolger des Humangenompro-

jekts (HPG, 1990 bis 2004) und des Sequenzierprojekts ENCODE (2003 bis 2012) – sammelte und sequenzierte Genome von verschiedenen Populationen weltweit, um eine Karte der menschlichen Genomvariationen zu erstellen, mit dem ambitioniertem Ziel: „Measurement of human DNA variation is an essential prerequisite for carrying out human genetics research. The 1000 Genomes Project represents a step towards a complete description of polymorphic human DNA sequence variation" (1000 Genomes Project Consortium 2010: 1070; International HapMap Consortium 2013). In der Pilotphase wurden dafür insgesamt 4,9 Terabasen als Datengrundlage generiert. Entdeckungen wurden im 1.000-Genom-Projekt rein rechnerisch durch den Vergleich mit dem Datensatz des Referenzgenoms NCBI36 erzielt. Der Realitätsgehalt dieser Entdeckungen steht und fällt daher mit der Güte des NCBI36-Referenzgenomdatensatzes.

NCBI36 ist die Endversion des menschlichen Genoms (International Human Genome Sequencing Consortium 2004), die 2004 in *Nature* veröffentlicht wurde und nach Aussage der maßgeblichen Sequenzierer eine aufwendige Prozedur des Qualitätschecks durchlaufen hat, um zukünftig als Referenzdatensatz zu fungieren (Schmutz et al. 2004). Aus dieser Evaluation resultierte das Label einer 99,99-prozentigen Genauigkeit (*Phred quality score* 40) gemäß dem 1997 vom *International Human Genome Consortium* etablierten „Bermuda-Standard". Doch dieser Standard verlangt lediglich die Veröffentlichung der Sequenzen, eine 99,99-prozentige Genauigkeit sowie die Vollständigkeit der Sequenzen. Wie dies in der Praxis erzielt werden soll, legen die Richtlinien des *International Human Genome Consortium* fest (2001). In ihrer Evaluationsstudie beschreiben Jeremy Schmutz, Jeremy Wheeler und Jane Grimwood wie sich die Genauigkeit für den Referenzgenomdatensatz NCBI36 definierte: „Compliance with the base-pair (bp) accuracy standard was measured by error probability assessments generated by DNA base-calling software and by examining discrepancies between overlapping clone sequences" (Schmutz et al. 2004: 365). Ohne Referenzdatensatz – „more than 2.8 billion base pairs of unique finished sequence has been generated" (Schmutz et al. 2004: 365) – gestaltete sich die Validierung schwierig. Daher wurde zum einen eine rein computer-basierte Qualitätsanalyse durchgeführt, indem die ursprünglichen Spurdaten (Gelbilder) neu analysiert wurden. Zum anderen wurde ein Goldstandard definiert, wofür 34 Megabasen des NCBI36-Datensatzes ausgewählt und aufwendig neu sequenziert wurden. Das Fazit lautete: „Our analysis indicates that all of the sequencing centres surveyed met the standards for 99.99 % accuracy over the time period studied." (Schmutz et al. 2004: 366) Doch der Vergleich der 34 Megabasen des NCBI36-Datensatzes mit den biologischen Vorlagen konnte nur über die automatisiert generierten und prozessierten Daten erfolgen. Dennoch wird dadurch eine

99,99-prozentige Übereinstimmung der Spurdaten mit den empirischen Daten suggeriert.

Fragt man, worauf sich dieses Vertrauen begründet, dann ist dies der *Phred* Standard für die *base-calling* Algorithmen der Sequenziermaschinen. Zur Erinnerung: *base-calling* ist die vierte und letzte Phase der Computeranalyse der Sequenzierung, in der die Spurdaten in eine Abfolge von (wahrscheinlichen) Basensymbolen übersetzt werden. Der *Phred* Quality Score wurde 1998 von Ewing und Green definiert, indem sie ihren *base-calling* Algorithmus mit dem des kommerziellen Anbieters *Applied Biosystems* für deren ABI PRISM Sequenzer verglichen (ABI 1996). Applied Biosystems war der erste kommerzielle Anbieter eines Sequenzierers und entstand als Ausgründung der erwähnten Caltech-Forschergruppe um Leroy Hood Mitte der 1970er Jahre (Chow-White & Garcia-Sancho 2012: 134–136). Der ABI *base-calling* Algorithmus ging dabei maßgeblich auf die Software der ersten Sequenzier-Maschine zurück (Connell et al. 1987). In den Worten von Ewing, LaDeana Hillier, Michael C. Wendl und Phil Green: „That software achieves impressive accuracy and remains the standard against which other methods must be judged" (1998: 176). Die Methode des *Phred base-calling* Algorithmus bestand aus folgenden Schritten: Zuerst wurde mathematisch, mit der Fourier-Methode ein idealisiertes Lasersignal für alle vier Basensignale generiert (*synthetic trace* respektive *predicted peak*), das jedoch dieselbe Lokalisierung wie die tatsächlichen Messsignale hatte; dann wurde das tatsächliche Signal gemessen (*observed peak*) und mit dem idealisierten Signal verglichen. „Peak matching consists of assigning an observed peak to each predicted peak. This is the most complex part of the base-calling procedure" (Ewing et al. 1998: 178). Abschließend wurde versucht, fehlende Signale aufzufinden. Die Genauigkeit des Algorithmus konnte dann,

> easily measured by aligning read sequences produced by that program to the correct sequence and tabulating discrepancies. […] We assessed the accuracy of the *phred* and ABI base-callers in several large sets of reads from cosmid clones sequenced in three laboratories (Ewing et al. 1998: 179). For each cosmid we created two FASTA formatted files, one containing the ABI base-called reads, and the other containing the *phred* base-called reads (as generated from the ABI-processed trace files) (Ewing et al. 1998: 180).

Das Resultat des Datenvergleichs erstaunt wenig: *„[P]hred* could attain an error rate of <1 in 10,000 by designating every base with quality value <40" (Ewing et al. 1998: 183). Was mit der „korrekten Sequenz" in obigem Zitat gemeint war, wird nicht ganz klar, da bereits die Inputdaten prozessierte Daten waren (Ewing et al. 1998: 177). Doch was bedeutet das für

den Realitätsgehalt? Was gezeigt wurde, war, dass im Vergleich zu anderen Algorithmen *phred* in punkto Genauigkeit der Interpretation der Chromatogramm-Dateien dominierte und dass man sich bezüglich letzterer auf das SCF-Format verließ. Dieses Format von Dear und Staden beschrieb, wie die Daten in einer Chromatogram-Datei auszusehen haben. Dazu gehörte die entscheidende Angabe über den Konfidenzwert für jede einzelne Übersetzung der Spurdaten in einen der vier DNA-Buchstaben (Dear & Staden 1992: 108). Doch die Lektüre der Originalbeschreibung des SCF weist eine erstaunliche Lücke auf, denn so Dear und Staden, „no instrument yet provides them [confidence values]" (Dear & Staden 1992: 109). Immerhin lief zu diesem Zeitpunkt das Humangenomprojekt bereits seit zwei Jahren, so dass dieser Befund nicht nur irritierend ist, sondern auch in direktem Widerspruch zur Aussage steht, dass die ABI-Software eine beeindruckende Akkuratesse aufweise (Ewing et al. 1998: 176).

Irgendwann zwischen 1982 und 1992, so ist zu vermuten, brach die direkte Validierungskette zwischen den empirischen Daten (Gelplatten, Gelfilmen) und den Maschinendaten (Gelbilder und Gel-Reads) ohne manuelle Codierung der Unsicherheiten und ohne Konfidenzwerte der Maschinendaten ab. Es bleibt nur das Vertrauen in die Algorithmen der Geräteherstellers, in adäquate experimentelle Designs der Sequenzierungsprojekte (hohe Abdeckungsrate) sowie in frei zugängliche Algorithmen und Daten. Denn die menschlichen Kapazitäten reichen angesichts der steigenden Datenmengen schon lange nicht mehr aus – „A single person can sequence 7 or 14 clones per day, each of which will give a sequence of about 250 bases and so produce a daily total of between 1,750 and 3,500 characters" (Staden 1982: 4731) –, um den Realitätsgehalt der Mega- und Terabytes an computergenerierten Daten ohne die Hilfe von Algorithmen zu überprüfen.

Rein Algorithmen-basierter Ansatz

Nicht nur reichen die menschlichen Kapazitäten schon lange nicht mehr aus, um die Datengenerierung manuell respektive visuell zu überprüfen; die informatischen Strategien der Shotgun-Methode sind nur dann aussagekräftig, wenn die Abdeckungsrate des zu untersuchenden Genoms entsprechend hoch ist – insbesondere wenn sie als *Whole Genom Shotgun* (WGS) auf die Entschlüsselung des menschlichen Genoms mit dessen 50-prozentiger Wiederholungsrate angewandt wird. Das heißt, „there is a greater risk of long-range misassembly" (Waterston et al. 2002: 3712). Wie groß das Risiko ist, kann einzig durch Algorithmen überprüft werden. Hinzukommt, dass die Delegation der DNA-Datengenerierung an die Algorithmen als Folge der experimentellen Strategie der Shotgun-Methode immer komplexere mathematische Verfahren integriert. Das ist der Preis

dafür, dass dadurch die Sequenzierung der drei Milliarden Basenpaare des menschlichen Genoms möglich wurde.

Deutlich wird diese Problematik am Wettrennen des staatlich finanzierten, internationalen Konsortiums des Humangenomprojekts mit dem kommerziellen Unternehmen *Celera Genomics* in den 1990er und frühen 2000er Jahren. *Celera* entwickelte mit dem *Celera Assembler* eine rein algorithmenbasierte WGS-Methode. Diese basierte auf einem graphentheoretischen Ansatz, der zwischen den Fragmenten Überlappungen als Graphen darstellte (Overlap Graph) und durch Koordinatenzuweisungen (*unitigger*) der Kanten der Graphen Ordnungen rekonstruierte (Myers et al. 2000). „Unitigs", also Subgraphen des Graphen aller Überlappungen, wurden dann zu Einheiten verbunden (*scaffolds*), indem aus den Verknüpfungsinformationen der Gel-Reads näherungsweise auf den Abstand der überlappenden DNA-Segmente (*contigs*) geschlossen wurde, die zusammen eine Konsensussequenz der DNA repräsentierten. Eine Konsensussequenz ist, wie bereits von Staden 1982 definiert, die berechnete Sequenz von DNA-Basen, die in der Summe am wenigsten von einer gegebenen Menge von entsprechenden Mustersequenzen abweicht. Dieser avancierte, algorithmische Ansatz (von DNA-Fragmenten über den Screener und Overlapper zum Unitigger sowie Scaffolder und schließlich zur Konsensussequenz) verband statistische Methoden mit heuristischen, stochastischen und approximativen Methoden.

Voraussetzung für diesen Ansatz wäre jedoch gewesen, die Anzahl der sequenzierten Fragmente von der üblichen 7- bis 9-fachen Abdeckung des Genoms auf eine 14- bis 18-fache Abdekung zu verdoppeln, um die Fehlerquoten zu reduzieren. Ohne diese hohe Abdeckung ist die Güte der Ergebnisse nicht zu gewährleisten und eben dies kennzeichnete die Datenhybris der *Celera* Sequenzierung (Venter et al. 2001). Denn im Unterschied zu den staatlichen Institutionen, die weitere Methoden (*hierarchical shotgun*) zur Absicherung der Resultate verwendeten, setzte *Celera Genomics* allein auf die WGS-Methode bei einer Abdeckung von 5,1 plus zusätzlicher 2,9-facher Abdeckung aus den öffentlichen Datenbanken des Humangenomprojekts (Schön 2002: 8–10). Zudem wurden gezielt Rohsequenzen anstelle zufälliger Verteilungen aus den Daten des Humangenomprojekts ausgewählt. Eine Analyse der *Celera* Daten ergab ein entsprechend fragwürdiges Bild: „Our analysis indicates that it is not possible to draw meaningful conclusions about the WGS approach because the authors did not perform an analysis of their own data by itself. Instead, they used an unorthodox approach to incorporate simulated data from the HGP" (Waterson et al. 2002: 3716). In anderen Worten, es wurde einiges getrickst in dem,

was als „The Sequence of the Human Genome" in *Science* angepriesen wurde (Venter et al. 2001).

Dies diskreditiert nicht die WGS-Methode, aber es macht zweierlei deutlich: Zum einen, dass die Zukunft der Molekularbiologie in noch mehr Statistik/Stochastik, Heuristik und Approximation bestehen wird, insbesondere wenn kommerzielle Forschung wie die von Craig Venter, dem Gründer von *Celera Genomics,* sich nichts Geringeres zum Ziel auserkoren hat, als „sequencing the ocean […] and the air" (Venter et al. 2004). Zum anderen, dass dieses Mehr an Statistik/Stochastik, Heuristik und Approximation neue und wesentlich schnellere Sequenziermethoden braucht, denn die Anforderungen, die ein rein Algorithmen-basierter Ansatz an die Sequenzierung stellt, sind mit der Sanger-Sequenzierung nicht mehr zu bewerkstelligen. „An individual human genome sequenced today using the Sanger method and capillary electrophoresis would take approximately 10,000 instrument days (e. g. 30 instruments for 1 year) to complete and would cost approximately $10 million" (Bentley 2006: 545). Engpass der Sanger-Sequenzierung ist die Elektrophorese, auch wenn sie als Kapillarelektrophorese von anfangs 96 auf bis zu 384 parallele Proben Mitte der 2000er Jahre erweitert wurde.

Next Generation Sequencer

Das Bestreben, die Elektrophorese ad acta zu legen, führte in den letzten Jahren zu einer Explosion vielfältigster Methoden für die DNA-Sequenzierung. Als alternative Ansätze zur Elektrophorese wurden die Rasterelektronenmikroskopie, die Massenspektrometrie, verschiedene Hybridsierungsmethoden, die pH-vermittelnde Sequenzierung (Post-Light-Sequenzierung), das *Nanopore Sequencing* oder das *Pyrosequencing* erforscht. Aus dieser Vielfalt der so genannten *Next Generation Sequencer*, deren Miniaturisierung auf Pico- und Nano-Ebene sowie deren Reduktion der Kosten für einzelne, wesentlich kleinere Geräte, generiert sich die seit 2008 ansetzende Flut an Genomdaten. Sequenzierung ist nicht mehr nur den großen Sequenzierzentren vorbehalten, sondern erobert sowohl den diagnostischen Alltag der Kliniken im Kontext der individualisierten Systemmedizin als auch die Laborforschung außerhalb der Molekularbiologie wie etwa in den Umweltwissenschaften. Diese zweite Generation an Sequenziergeräten produziert etwa 1 Gigabase Daten pro Tag und Maschine bei einer Read-Länge von wenigen Basen (*short read*) bis zu 10 Kilobasen, je nach Methode. Insbesondere das *Pyrosequencing* kombiniert mit Emulsions-PCR (Miniaturisierung) lässt sich hochgradig parallelisieren (Ronaghi et al. 1996, 1998; Margulies et al. 2005; Shendure et al. 2005). „The strategy entailed arraying several hundred thousand sequencing templates in

either picotiter plates or agarose thin layers, so that these sequences could be analyzed in parallel" (Schuster 2008: 16).

Damit kündigte sich das Zeitalter des *Real Time DNA Sequencing* an, also die dritte Sequenzierergeneration. Der Begriff „Real Time" bezieht sich auf die „DNA polymerase as a real-time sequencing engine" (Eid et al. 2009: 133). Damit wird die direkte Echtzeitentschlüsselung der DNA-Sequenzen Realität, indem einzelne Basenmaleküle in pico- und nanomolaren Konzentrationen mit Hilfe fluoreszierenden Substanzen detektiert werden. Oder in anderen Worten: „That is, direct observation of processive DNA polymerization with basepair resolution" (Eid et al. 2009: 133; Levene et al. 2003). Forscher können nun der Entschlüsselung zusehen, falls sie etwas Zeit mitbringen. Denn es braucht etliche Minuten, um eine einzelne Base zu entschlüsseln, da sich das Attribut Echtzeit auf die Enzymkinetik bezieht. Allerdings ist die Fehlerquote dieser neuen Sequenziertechnologien relativ hoch. In einer Machbarkeitsstudie wurden von 158 Basen eines künstlich hergestellten und nur aus zwei Basen bestehenden DNA-Strangs lediglich 131 Basen korrekt identifiziert. „The 27 errors consisted of 12 deletions, eight insertions, and seven mismatches" (Eid et al. 2009: 136). Die hohe Fehlerquote dieser neuen Technologie soll rein rechnerisch mit Korrekturalgorithmen und adäquaten Assembly-Strategien ausgeglichen werden (Koren et al. 2012):

> The instrument generates reads that average only 82.1 %–84.6 % nucleotide accuracy, with uniformly distributed errors dominated by point insertions and deletions […] far beyond the 5–10 % error rate that most genome assemblers can tolerate. […] Thus, there is a great potential advantage to the long, single-pass reads if the error rate can be algorithmically managed. […] Our PBcR (PacBio corrected Reads) algorithm, implemented as part of the Celera Assembler, trims and corrects individual long-read sequences by first mapping short-read sequences to them and computing a highly accurate hybrid consensus sequence: improving read accuracy from as low as 80 % to over 99.9 % (Koren et al. 2012: 693).

Konklusion

Die magische Zahl von 99,99-prozentiger Genauigkeit der Genomdaten, wie auch immer sie bei 5- bis 10-prozentiger Fehlerrate der Sequenzierer begründet wird, suggeriert Vertrauen in den Datenpositivismus der Molekularbiologie. Darüber wird gerne vergessen, dass selbst eine 99,99-

prozentige Genauigkeit „would still result in hundreds of errors in a microbial genome and hundreds of thousands of errors in a mammalian genome“ (Shendure et al. 2005: 1728). Dessen ungeachtet macht die skizzierte Entwicklung des informatischen Ansatzes die inhärente Eigendynamik als Forderung nach immer mehr Daten deutlich, insofern das intrinsische Diskretisierungsproblem der Fragmentierung der DNA-Sequenzierung (Sanger 1988:20) durch die Shotgun-Methode extrinsisch und randomisiert wird. Dies erfordert eine entsprechende Replizierung der DNA-Vorlagen (hohe Abdeckungsquote), was wiederum nur durch die komplette Automatisierung der Sanger-Sequenzierung auf Basis von Messbildern, Chromatographie-Dateien und Algorithmen möglich wurde. Dies hatte die Transformation des Datenbegriffs der Genomanalyse bezüglich der Genauigkeit der Daten in einen rein statistischen Datenbegriff zur Folge, der seinerseits eine Steigerung der Abdeckungsquote erforderte. Mit dem rein Algorithmen-basierten Ansatz der WSG-Methode erhöhte sich die Anforderung an die Abdeckungsquote ein weiteres Mal, um die Güte der Resultate zu verbessern. All dies ist mittlerweile mit der Sanger-Methode finanziell wie zeitlich kaum noch zu bewerkstelligen und führt unweigerlich zu neuen Generationen von Sequenzieren, die kleiner, kostengünstiger und hochgradig parallelisiert sind und damit schneller arbeiten. Doch dies geht wiederum mit einer erhöhten Fehlerquote der Sequenzierung einher, die rein rechnerisch durch noch mehr Statistik/Stochastik, Heuristik und Approximation kompensiert werden muss. In anderen Worten: Der indikative Gehalt der Genomdaten, also der Realitätsgehalt, der sich aus der messtechnischen Differenzbildung im Laborexperiment generiert und in der molekularbiologischen Fachliteratur „experimental data“ oder „empirical data“ genannt wird, lässt sich zunehmend schwerer feststellen und allenfalls als Wahrscheinlichkeitszuweisung. Diese wird jedoch in Aussagen wie „well determined“ oder *phred* Quality Score 40 transformiert und erhält dadurch indikativ anmutenden Aussagewert. Die Datenproblematik verstärkt sich zudem durch die Heterogenität der Daten aufgrund der sich permanent verändernden Verfassung der Datengrundlage, wie dies typisch für ein Großprojekt wie das 1,000-Genom-Projekt ist: „The heterogeneity of the sequence data (read lengths from 25 to several hundred base pairs (bp); single and paired end) reflects the diversity and rapid evolution of the underlying technologies during the project“ (1000 Genomes Project Consortium 2010: 1062).

Doch wenn die Datenbasis bereits in ihrem epistemischen Status problematisch ist, was bedeutet dies dann für die daraus abgeleiteten Funktionsdaten? Denn nicht nur die Assembler-Algorithmen werden zunehmend komplexer, sondern vor allem die zahlreichen Analysealgorithmen, die immer neue Formen der Analyse der sequenzierten Daten in den Kartierungs-

und Mappingprojekten immer schneller ermöglichen. Ein solch avancierter Analysealgorithmus ist beispielsweise der SIFT-Algorithmus (*Sorting Intolerant from Tolerant*), der die Lücke zwischen Mutationen und Phänotyp-Variationen algorithmisch schließt, indem er vorhersagt, ob die Substitution einer Aminosäure die Proteinfunktion beeinflusst oder nicht (Kumar et al. 2009). Es ist diese Art von rein rechnerisch rückgeschlossener Vorhersage auf Basis der Sequenzierdaten, die das aktuelle bioinformatische Wissen charakterisiert. Dabei wird vergessen, „kritisch zu hinterfragen, wo die konkrete biologische Relevanz der unermesslichen Fülle an Sequenzdaten liegt. Denn für die meisten Gene gibt es auch […] nach Ende des Human Genome Project nur eine „putative" – sprich bioinformatisch vermutete – Funktion" (Gabrielczyk 2009: 3).

Vor diesem Hintergrund ist der Datenpositivismus der Molekularbiologie mit Vorsicht zu genießen. Die Frage nach der epistemischen Umgestaltung des Genomverständnisses durch die informatischen Strategien und den damit verbundenen Aussagegehalten ist nicht so einfach zu beantworten. Zur Trias von Statistik/Stochastik, Heuristik und Approximation gesellt sich nun das „correlation ist enough" der Big Data-Analytik hinzu (Anderson 2008). Erste Studien prognostizieren bereits über die Korrelation signifikanter Genexpressionen die Rückfall- und Überlebensrisiken von Krankheiten wie beispielsweise die Studie „Gene expression profiling predicts clinical outcome of breast cancer" (van't Veer et al. 2002). Doch ähnlich wie die eingangs erwähnte Datenhybris der Prognosen von GFT sind solche Prognosen zweifelhaft, wie Liat Ein-Dor, Gad Getz, David Givol und Eytan Domany feststellen:

> Several microarray studies yielded gene sets whose expression profiles successfully predicted survival in breast cancer. However, the overlap between these gene sets was almost zero. […] first, many genes are correlated with survival; second, the differences between these correlations are small; and third, the correlation-based rankings of the genes depend strongly on the training set. These properties may indicate that the top 70 genes are not superior to others in predicting disease outcome. […] Our results imply that although the top 70 genes may provide good prediction, other groups of 70 genes may do the same (Ein-Dor et al. 2005: 173).

Mehr Beliebigkeit ist kaum denkbar und diese Beliebigkeit resultiert aus der gewählten Methode für die Datenanalyse anhand der Analysealgorithmen.[8] Doch es könnte noch einen weiteren Grund geben: Die Beliebigkeit könnte der biologische Relevanz geschuldet sein, dass individuelle Patientengene zu heterogen sind (Ein-Dor et al. 2005: 177). Dies ist eine offene Forschungsfrage, die jedoch für ihre Beantwortung weit mehr als tausend

Genome zum Vergleich benötigt. Ließe sich ersteres eventuell mit Unachtsamkeit in der Methodenwahl erklären oder eben als Fallstrick eines ontologisch indifferenten Gebrauchs des Datenbegriffs, so würde letzteres das Aus der Big Data-Revolution für die Molekularbiologie bedeuten. Denn dann stünde die Bioinformatik vor dem selben Problem wie *Google Flu Trends*: Ihre Daten wären kaum objektivierbar, um die versprochene Transformation des Korrelativen ins Indikative für ihre Vorhersagen zu gewährleisten.

Anmerkungen

1 GenBank wird vom US-National Institute of Health (NIH) betrieben und ist Teil der International Nucleotide Sequence Database Collaboration des NIH, der DNA DataBank of Japan (DDBJ) und des European Molecular Biology Laboratory (EMBL) (GenBank 2016).

2 Grundlegend für den „Big Data Club" der Molekularbiologie ist das zentrale Dogma des genetischen „Code-Scripts" (Schrödinger 1944; Crick 1970), gleichwohl es mittlerweile umstritten ist (Thieffry & Sarkar 1998; Falk 2010). Das Dogma tradiert sich stark verkürzt wie folgt: Die Desoxyribonukleinsäure (DNA) eines aktivierten Gens wird in Ribonukleinsäure (RNA) (Transkription) und diese in ein Protein übersetzt (Translation). Ausbuchstabiert bedeutet dies: Die vier Nukleotide der DNA (A Adenin, T Thymin, G Guanin, C Cytosin) werden in die vier Nukleotiden der RNA (Adenin, Guanin, Cytosin, Uracil) transkribiert. Jeweils ein Triplet von RNA-Nukleotiden (Codon) kodiert eine der zwanzig kanonischen Aminosäuren, aus welchen die Proteine gebildet werden (Fox Keller 1995, 2000; Kay 2000; de Chadarevian 2002). Der Code des Lebens ergibt sich also aus der Kombinatorik der 4^3 (64) möglichen Codons und daher scheint es nicht erstaunlich, dass die Sequenzierung der DNA zur Obsession der Molekularbiologie geworden ist – angefangen mit dem 5 Kilobasen (Kb) Genom des Bakteriophagen øX174 über das 4,6 Megabasen (Mb) Bakteriengenom von *Escherichia coli* bis hin zum 3 Gigabasen (Gb) Genom des Menschen.

3 Die Substitution radioaktiver durch fluoreszierender Markierungen war daher der entscheidende Schritt in Richtung Automatisierung. Ein weiterer war die von Joachim Messing und Kollegen vom Max-Planck-Institut für Biochemie und der Universität zu Köln bereits Mitte der 1970er Jahre entwickelte Methode, doppelsträngige DNA in so genannten „Vektor-Systemen" (E. coli Bakterien oder Bakteriophagen M13 Plasmide) für die DNA-Sequenzierung zu vervielfältigen (Messing und Vierira 1982; Heather & Chain 2016: 3; Karger & Guttman 2009). „Not only is this a means of obtaining single-stranded DNA, but it also solves the fractionation problem as fractionating is by cloning, which is really the ultimate method of purification and can be applied to a mixture of any complexity" (Sanger 1988: 23).

4 Mitte bis Ende der 1970er Jahre wurde die Computertechnologie für Konsumenten wie auch Labore mit den ersten Mikroprozessoren (Intel 4004), Micro-Computern und PCs (Apple II, Atari, Xerox Parc Alto, etc.), Laserdruckern (Xerox Parc) und Disketten erschwinglich.

5 „Shotgunned" bezieht sich auf die Aufspaltung eines Genoms in kurze Fragmente von hundert bis tausend Basenpaaren, die dann sequenziert werden. Das Problem besteht jedoch in der Zusammensetzung der einzelnen Basenpaare zum vollständigen Genom (Anderson 1981: 3015).

6 „A contig is a set of gel readings that are related to one another by overlap of their sequences. All gel readings belong to one and only one contig and each contig contains at least one gel reading. Using some simple rules all the gel readings in a contig can be summed to produce a continuous consensus sequence and the length of this sequence is the length of the contig. At any stage of a sequencing project the data will comprise a number of contigs each of which contains a set of gel readings. When a project is completed there will be only one contig and its consensus will be the finished sequence" (Staden 1982: 4732–4733).

7 Durch die Kommerzialisierung der DNA-Sequenzierer wird der Zugang zu den Rohdaten wie auch zu den *base calling* Algorithmen schwierig. Dadurch lässt sich die Güte der Daten nicht beurteilen und dies bringt die Forscher dazu, offene Community-Standards und -Formate einzuführen (Ferry & Sulston 2010: 93–95).

8 „The main lesson is that whenever any arbitrary decision (e. g. choice of training and test set) is taken throughout analysis of the data, one has to generate a large ensemble of the different ways in which this arbitrary decision could be taken, and perform a statistical analysis of the results obtained over this ensemble. A high sensitivity of the results to the arbitrary decisions may indicate that the conclusions, e. g. the list of survival-related genes, are not unequivocal" (Ein-Dor et al. 2005: 177).

Literatur

1000 Genomes Project Consortium 2010. A Map of Human Genome Variation from Population-scale Sequencing. *Nature* (467/7319): 1061–1073.

ABI 1996. *ABI PRISM, DNA sequencing analysis software. User's manual*. Foster City, CA: PE Applied Biosystems.

Anderson, Stephen 1981. Shotgun DNA Sequencing Using Cloned DNase I-generated Fragments. *Nucleic Acids Research* (9): 3015–3027.

Anderson, Chris 2008. The End of Theory: The Data Deluge Makes the Scientific Methods Obsolete. *Wired Magazine* 16.07.2008.

Bentley, David R. 2006. Whole-genome Re-sequencing. *Current Opinion in Genetics & Development* (16): 545–552.

Berg, P., H. Fancher und M. Chamberlin 1963. The Synthesis of Mixed Polynucleotides Containing Ribo- and Dexyribonucleotides by Purified Preparations of DNA Polymerase from Escherichia coli. In: Henry J. Vogel et al. (Hg.). *Symposium on Informational Macromolecules*. New York, London: Academic Press: 467–483.

Birney, Ewan und Nicole Soranzo 2015. The End of the Start for Population Sequencing. *Nature* (526): 52–53.

Bonfield, James K. und Roger Staden 1995. The Application of Numerical Estimates of Base Calling Accuracy to DNA Sequencing Projects. *Nucleic Acids Research* (23/8): 1406–1410.

Butler, Declain 2013. When Google Got Flu Wrong. *Nature* (494): 155–156.

Chadarevian, Soraya de 2002. *Designs for Life: Molecular Biology after World War II*. Cambridge: Cambridge University Press.

Chow-White, Peter A. und Miguel Garcia-Sancho 2012. Bidirectional Shaping and Spaces of Convergence: Interactions between Biology and Computing from the First DNA Sequencers to Global Genome Databases. *Science, Technology & Human Values* (37/1): 124–164.

Churchill, Gary A. und Michael S. Waterman 1992. The Accuracy of DNA Sequences: Estimating Sequence Quality. *Genomics* (14): 89–98.

Codd, Edgar F. 1970. A Relational Model of Data for Language Shared Data Banks. *Communications of the ACM* (13/6): 377–387.

Connell, Charles et al. 1987. Automated DNA Sequence Analysis. *BioTechniques* (5): 342–348.
Crick, Francis 1970. Central Dogma of Molecular Biology. *Nature* (227): 561–563.
Dear, Simon und Roger Staden 1991. A Sequence Assembly and Editing Program for Efficient Management of Large Projects. *Nucleic Acids Research* (19/14): 3907–3911.
Dear, Simon und Roger Staden 1992. A Standard File Format for Data from DNA Sequencing Instruments. *DNA Sequence* (3/2): 107–110.
Eid, J., A. Fehr, J. Gray et al. 2009. Real-time DNA Sequencing from Single Polymerase Molecules. *Science* (323): 133–138.
Ein-Dor, Liat, Gad Getz, David Givol und Eytan Domany 2005. Outcome Signature Genes in Breast Cancer: Is There a Unique Set? *Bioinformatics* (21/2): 171–178.
Ewing, Brent, LaDeana Hillier, Michael C. Wendl und Phil Green 1998. Base-calling of Automated Sequencer Traces Using *Phred.* I. Accuracy Assessment. *Genome Research* (8/3): 175–185.
Falk, Raphael 2010. What is a Gene? Revisited, Studies in History and Philosophy of Science. *Studies in History and Philosophy of Biological and Biomedical Sciences* (41/4): 396–406.
Ferry, Georgina und John Sulston 2010. *The Common Thread*. New York: Random House.
Fox Keller, Evelyn 1995. *Refiguring Life: Changing Metaphors in 20th Century Biology*. New York: Columbia University Press.
Fox Keller, Evelyn 2000. *The Century of the Gene*. Harvard: Harvard University Press.
Fruton, Joseph 1972. *Molecules and Life*. New York: Wiley Intersccience 1972.
Gabrielczyk, Thomas 2009. Editorial: Next-Generation-Publishing. *Laborwelt* (10/3): 3.
Gandomi, Amirund Murtaza Haider 2015. Beyond the Hype: Big Data Concepts, Methods, and Analytics. *International Journal of Information Management* (35/2): 137–144.
García-Sancho, Miguel 2010. A New Insight into Sanger's Development of Sequencing: From Proteins to DNA, 1943–1977. *Journal of the History of Biology* (43/2): 265–323.
García-Sancho, Miguel 2012. *Biology, Computing, and the History of Molecular Sequencing: From Proteins to DNA*, 1945–2000. London: Palgrave Macmillan.
GenBank 2016: *Homepage*. URL: https://www.ncbi.nlm.nih.gov/genbank/ (09.12.2016).
Ginsberg, Jeremy et al. 2009. Detecting Influenza Epidemics Using Search Engine Query Data. *Nature* (457): 1012–1014.
Gramelsberger, Gabriele 2013. Simulation and Systems Understanding. In: Hanne Andersen, Dennis Dieks, Wenceslao J. Gonzalez, Thomas Uebel und Gregory Wheeler (Hg.). *New Challenges to Philosophy of Science*. Dordrecht: Springer: 151–161.
Heather, James M. und Benjamin Chain 2016. The Sequence of Sequencers: The History of Sequencing DNA. *Genomics* (107/1): 1–8.
International HapMap Consortium 2013. The International HapMap Project. *Nature* (426): 789–796.
International Human Genome Consortium 2001. *Standard Finishing Practices and Annotation of Problem Regions for the Human Genome Project*. URL: https://www.genome.gov/10001812/ (10.09.2017).
International Human Genome Sequencing Consortium 2004. Finishing the Euchromatic Sequence of the Human Genome. *Nature* (431): 931–945.
Karger, Barry L. und Andras Guttman 2009. DNA Sequencing by Capillary Electrophoresis. *Electrophoresis* (30): 196–202.
Kay, Lilly 1988. Laboratory Technology and Biological Knowledge: The Tiselius Electrophoresis Apparatus, 1930–1945. *History and Philosophy of the Life Sciences* (10): 51–72.
Kay, Lilly 2000. *Who Wrote the Book of Life: A History of the Genetic Code*. Stanford: Stanford University Press.
Koren, Sergey, Michael C. Schatz, Brian P Walenz, Jeffrey Martin, Jason T Howard, Ganeshkumar Ganapathy, Zhong Wang, David A Rasko, W. Richard McCombie, Erich D. Jarvis und Adam M. Phillippy 2012. Hybrid Error Correction and de novo Assembly of Single-molecule Sequencing Reads. *Nature Biotechnology* (30): 693–700.
Krajewski, Markus 2007. In Formation. Aufstieg und Fall der Tabelle als Paradigma der Datenverarbeitung. *Nach Feierabend. Züricher Jahrbuch für Wissenschaftsgeschichte* (3): 37–55.

Kumar, Prateek, Steven Henikoff und Pauline C. Ng 2009. Predicting the Effects of Coding Non-synonymous Variants on Protein Function Using the SIFT Algorithm. *Nature Protocols* (4): 1073–1081.

Lander, Eric S., Lauren M. Linton, Bruce Birren et al. 2001. Initial Sequencing and Analysis of the Human Genome. *Nature* (409): 860–921.

Lazer, David, Ryan Kennedy, Gary King und Alessandro Vespignani 2014. The Parable of Google Flu: Traps in Big Data Analysis. Science (343): 1203–1205.

Leonelli, Sabina 2012. Introduction: Making Sense of Data-driven Research in the Biological and Biomedical Sciences. *Studies in History and Philosophy of Biological and Biomedical Sciences* (43): 1–3.

Leonelli, Sabina 2014. What Difference Does Quantity Make? On the Epistemology of Big Data in Biology. *Big Data & Society* (1): 1–11.

Leonelli, Sabina 2016. *Data-Centric Biology: A Philosophical Study*. Chicago, IL: Chicago University Press.

Leonelli, Sabina und Rachel A. Ankeny 2010. Re-thinking Organisms: The Impact of Databases on Model Organism Biology. *Studies in History and Philosophy of Biological and Biomedical Sciences* (43): 29–36.

Levene, M. J., J. Korlach, S. W. Turner, M. Foquet, H. G. Craighead und W. W. Webb 2003. Zero-mode Waveguides for Single-molecule Analysis at High Concentrations. *Science* (299): 682–686.

Margulies, Marcel, Michael Egholm, William E. Altman et al. 2005. Genome Sequencing in Microfabricated High-density Picolitre Reactors. *Nature* (437): 376–380.

Marx, Vivien 2013. Biology: The Big Challenges of Big Data. *Nature* (498): 255–260.

Messing, Joachim und J. Vierira 1982. A New Pair of M13 Vectors for Selecting either DNA Strand of Double-digest Restriction Fragments. *Gene* (19/3): 269–276.

Myers, Eugene W., Granger G. Sutton, Art L. Delcher et al. 2000. A Whole-Genome Assembly of Drosophila. *Science* (287): 2196–2204.

O'Malley, Maureen und Orkun S. Soyer 2012. The Roles of Integration in Molecular Systems Biology. *Studies in History and Philosophy of Biological and Biomedical Sciences* (43/1): 58–68.

Rabinow, Paul 1996. *Making PCR. A Story of Biotechnology*. Chicago, IL: Chicago University Press.

Rheinberger, Hans-Jörg 2001a. Putting Isotopes to Work: Liquid Scintillation Counters, 1950–1970. In: Bernward Joerges, Terry Shinn (Hg.). Instrumentation Between Science, State and Industry. Dordrecht: Springer: 143–174.

Rheinberger, Hans-Jörg 2001b. *Experimentalsysteme und epistemische Dinge. Eine Geschichte der Proteinsynthese im Reagenzglas.* Göttingen: Wallstein Verlag.

Ronaghi, M., S. Karamohamed, B. Pettersson, M. Uhlén und P. Nyrén 1996. Real-Time DNA Sequencing Using Detection of Pyrophosphate Release. *Analytical Biochemistry* (242): 84–89.

Ronaghi, Mostafa, Mathias Uhlén und Pål Nyrén 1998. A Sequencing Method Based on Real-Time Pyrophosphate. *Science* (281): 363–365.

Sanger, Frederick 1949. Some Chemical Investigations on the Structure of Insulin. *Cold Spring Harbor Symposia on Quantitative Biology: Amino Acids and Proteins* (14): 153–160.

Sanger, Frederick 1959. Chemistry of Insulin. *Science* (129): 1340–1344.

Sanger, Frederick 1988. Sequences, Sequences, and Sequences. *Annual Review of Biochemistry* (57): 1–28.

Sanger, Frederick, Steve Nicklen und Alan R. Coulson 1977. DNA Sequencing with Chain-terminating Inhibitors. *PNAS Proceedings of the National Academy of Sciences* (74): 5463–5467.

Schmutz, Jeremy, Jeremy Wheeler, Jane Grimwood et al. 2004. Quality Assessment of the Human Genome Sequence. *Nature* (429): 365–368.

Schön, Oliver 2002. *Systematische Verfahrensoptimierung im Bereich der Mega-Sequenzierung und ihr exemplarischer Einsatz zur Analyse des Humangenoms*. Dissertation, Technische Universität Carolo-Wilhelmina Braunschweig.

Schrödinger, Erwin 1944. *What is Life? The Physical Aspect of the Living Cell*. Cambridge: Cambridge University Press.

Schuster, Stephan C. 2008. Next-generation Sequencing Transforms Today's Biology. *Nature* (5/1): 16–18.
Shendure, Jay et al. 2005. Accurate Multiplex Polony Sequencing of an Evolved Bacterial Genome. *Science* (309): 1728–1732.
Smith, Temple F. 1990. The History of the Genetic Sequence Databases. *Genomics* (6): 702–707.
Smith, L. M., G. J. Porreca, N. B. Reppas, X. Lin, J. P. McCutcheon, A. M. Rosenbaum, M. D. Wang, K. Zhang, R. D. Mitra, G. M. Church 1986. Fluorescence Detection in Automated DNA Sequence Analysis. *Nature* (321): 674–678.
Staden, Roger 1979. A Strategy of DNA Sequencing Employing Computer Programs. *Nucleic Acids Research* (6/7): 2601–2610.
Staden, Roger 1982. Automation of the Computer Handling of Gel Reading Data Produced by the Shotgun Method of DNA Sequencing. *Nucleic Acids Research* (10/15): 4731–4751.
Strasser, Bruno J. 2012. Data-driven Sciences: From Wonder Cabinets to Electronic Databases. *Studies in History and Philosophy of Biological and Biomedical Sciences* (43/1): 85–87.
Thieffry, Denis und Sahotra Sarkar 1998. Forty Years under The Central Dogma. *Trends in Biochemistry* (23): 312–316.
van't Veer, Laura J., Hongyue Dai, Marc J. van de Vijver 2002. Gene Expression Profiling Predicts Clinical Outcome of Breast Cancer. *Nature* (415): 530–536.
Venter, Craig J., Mark D. Adams, Eugene W. Myers 2001. The Sequence of the Human Genome. *Science* (291): 1304–1351.
Venter, Craig J., Karin Remington, John F. Heidelberg et al. 2004. Environmental Genome Shotgun Sequencing of the Sargasso Sea. *Science* (304): 66–74.
Waterston, Robert H., Eric S. Lander und John E. Sulston 2002. On the Sequencing of the Human Genome. *PNAS Proceedings of the National Academy of Sciences* (99/6): 3712–3716.
Wu, Ray und Dale A. Kaiser 1968. Structure and Base Sequence in the Cohesive Ends of Bacteriophage Lambda DNA. *Journal of Molecular Biology* (35/3): 523–537.

Gabriele Gramelsberger
Zentrum für interdisziplinäre Wissenschafts- und Technikforschung
RWTH Aachen
Theaterplatz 14
52062 Aachen
Deutschland
gramelsberger@humtec.rwth-aachen.de