Was das Leben wissen muss

G(e)nomforschung (Teil 1)

Röbbe Wünschiers, Universität Köln

In den vergangenen sieben Jahren wurden mehrere Dutzend Genome sequenziert. Computerprogramme und Menschenhirne analysieren seither die angefallenen Sequenzdaten und kommen zumindest zu einem Schluss: je höher ein Organismus evolutiv entwickelt ist, desto mehr scheinbar unnütze DNA trägt er mit sich herum. Aber wie viel DNA ist überhaupt notwendig? Das kleinste bekannte Genom des autonomen pathogenen Bakteriums *Mycoplasma genitalium* kodiert für 480 Proteine und 37 RNAs. Lässt sich mit weniger Genen das Leben bestreiten? Welchen Informationsgehalt trägt die DNA überhaupt, und wird es bald möglich sein, künstliches Leben zu schaffen? Die Totalsynthese von Viren jedenfalls ist bereits gelungen. Forschungsgruppen in aller Welt suchen nach Antworten auf das informationstheoretische Rätsel des Lebens.

Der Begriff Genom ist nicht so jung wie man aufgrund der neuen Wortschöpfung „Genomics“ meinen könnte. Er wurde in den dreißiger Jahren des vergangenen Jahrhunderts von dem deutschen Botaniker Hans Karl Winkler als Zusammenführung der Wörter „Gen“ und „Chromosom“ geschaffen. Erst zwei Jahrzehnte zuvor, im Jahre 1909, wurde von dem dänischen Populationsforscher und Züchter Wilhelm Johannsen der Begriff „Gen“ geprägt (siehe auch [1]). Der Genbegriff war damals aber nicht unbedingt mit den Chromosomen verknüpft, sondern stellte eine rein formale Einheit der Vererbung eines Merkmals von einer Generation auf die nächstfolgende Generation dar (Vererbungseinheit). Damit ist der Begriff Genom älter als die Entdeckung von Oswald Avery im Jahre 1944, dass die DNA die Erbinformation trägt.

Während früher die molekulare Analyse der Erbinformation auf einzelne Gene beschränkt war, werden seit 1995 komplette Genome sequenziert (Abb. 1). Dies resultiert in einer immensen Datenflut, die es zu analysieren gilt. Aber wie sieht die Datenmenge von organismischer Seite aus? Wie viel Erbinformation ist für ein Lebewesen überhaupt notwendig?

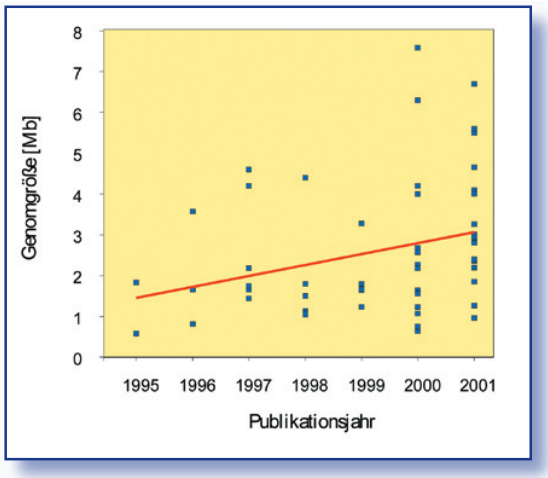
Der Autor

Dr. Röbbe Wünschiers studierte Biologie in Marburg und promovierter dort über „Eigenschaften, Regulation und Funktion einer Hydrogenase im Wasserstoffmetabolismus der einzelligen Grünalge *Scenedesmus obliquus*“. Seit Jan. 2002 ist R. Wünschiers an der Universität Köln, forscht dort im Bereich der Genom- und Genexpressionsdatenanalyse und unterrichtet am Cologne University Bioinformatic Center (CUBIC) Genetik. Für die CLB schreibt R. Wünschiers seit 1997 biologisch orientierte Beiträge.

Was ist Leben?

Um der Frage nach der minimal notwendigen Erbinformation, also des minimalen Genoms, eines Lebewesens nachgehen zu können, muss zunächst geklärt werden, was ein Lebewesen ist. So haben Viren zwar erheblich kleinere Genome als zum Beispiel die einfachsten Bakterien, die wir kennen. Allerdings zählen sie nicht, wie wir noch sehen werden, zu den Lebewesen. Was ist Leben? Obwohl die Frage so einfach klingt, haben schon viele Wissenschafter und Philosophen ihrer Beantwortung ihr Leben gewidmet. Es mag dabei selbstverständlich klingen, dass sich zwar die Biowissenschaftler ausgiebig der *Natur* des Lebens widmen, dass es jedoch überwiegend anderen Disziplinen überlassen ist, das Wesen des Lebens zu formulieren. Hier spielt vor allem die Physik eine große Rolle, die mit den Gedanken von Erwin Schrödinger den Entropie- und den damit zusammenhängenden Informationsbegriff in die Biologie einführte [2]. Die grundlegende Frage, was eigentlich Leben ist, sollte keineswegs als rein akademisches Problem abgetan werden. Sie ist für die Formulierung des Gentechnikgesetztes ebenso relevant wie für die Suche nach Lebensspuren im Universum. Wie könnten Wissenschaftler auf dem Mars nach Leben suchen, wenn sie nicht klare Definitionen haben, aus denen sie Suchmuster ableiten können. Und auch für die Suche nach dem kleinsten möglichen Genom ist der Lebensbegriff maßgebend. Kompliziert wird die Definitionssuche dadurch, dass es etwa so viele Kurzbeschreibungen des Lebens gibt wie Autoren, die sich daran versucht haben. Jede Betrachtung wird durch das Fachgebiet und den persönlichen Hintergrund des Autors gefärbt und getrübt. Ein Blick in die Brockhaus

Abbildung 1: Größe der publizierten Genomsequenzen. Mit jedem Jahr werden mehr Genome, aber auch größere Genome sequenziert. Daten aus [32].



Aufbau von Genomen

Prokaryontische Genome (von Bakterien und Archaebakterien) bestehen aus einem zirkularen Chromosom das nicht in einem Zellkern kompartimentiert ist. Zusätzlich zu dem Chromosom liegen oft noch Plasmide vor. Dies sind kleine zirkulare DNA-Moleküle, die leicht zwischen Bakterien ausgetauscht werden können und zusätzliche Gene tragen. Bakterielle Chromosomen und Plasmide sind in der Regel sehr dicht mit Genen belegt, die sich teilweise auch überlappen. Repetitive Sequenzen und Gen-unterbrechende Introns sind äußerst selten.

Eukaryontische Genome sind in der Regel auf mehrere lineare Chromosomen aufgeteilt und äußerst komplex aufgebaut. Die eukaryontischen Chromosomen befinden sich im Zellkern und werden von einer großen Anzahl unterschiedlicher Proteine in einer kompakten Struktur gehalten. Die meisten Gene sind von Introns unterbrochen (Abbildung 6). In „Geninseln" liegen viele Gene dicht zusammen, während große Sequenzbereich eher wie eine „Wüste" wirken. Das menschliche Genom besteht zu rund 70 Prozent aus DNA zwischen diesen Geninseln (intergenic sequences), wovon etwa die Hälfte repetitive Sequenzen sind. Etwa 30 Prozent des Genoms sind Introns, 0,5 Prozent tRNA und rRNA kodierenden Gene sowie lediglich ein bis zwei Prozent Protein-kodierenden Gene.

Enzyklopädie lehrt uns: „Naturwissenschaftlich ist Leben ein für Lebewesen (Organismen) eigentümliches Geschehen, das sich von der unbelebten Natur nicht ausreichend durch einzelne Merkmale, sondern nur als ein komplexes System (Ganzheit) von Eigenschaften unterscheiden lässt." [3]. Der französische Molekularbiologe und Nobelpreisträger Jacques Monod hat dieses „eigentümliche Geschehen" in den drei Eigenschaften Teleonomie (Lebewesen sind Objekte, die mit einem Plan ausgestattet sind), autonome Morphogenese (Freiheit gegenüber äußeren Kräften und Bedingungen) und Invarianz (unveränderte Reproduktion und Übertragung ihrer eigenen Information) zusammengefasst [4]. Eine praktikablere, anschaulichere und heute gebräuchliche Definition des Lebendigen gibt uns der Nobelpreisträger Christian de Duve [5]. Seiner Ansicht nach sind es sieben Eigenschaften, die das lebende System ausmachen:

1. *Synthese* seiner Bestandteile aus Material der Umgebung
2. Aufnahme und Verwandlung von *Energie*
3. *Katalyse* chemischer Reaktionen
4. Steuerung der Lebensprozesse, um eine getreue *Reproduktion* zu sichern
5. *Isolation*, um mit der Umwelt einen kontrollierten Stoffaustausch zu ermöglichen
6. *Regulation* der Lebensprozesse, um Änderungen der Umweltbedingungen begegnen zu können
7. *Vermehrung*

Diese sieben Eigenschaften sind notwendig und hinreichend für die Existenz des Lebens an sich und für sein Fortbestehen. Sie sind allen Lebewesen gemeinsam. Die Minimalbedingungen des Lebens sind demnach eine halb durchlässige (semipermeable) Membran, ein stoffwechselaktives Plasma und ein zur identischen Selbstvermehrung befähigter Genapparat. Diese Fähigkeiten sind erst mit einer Zelle zu bewerkstelligen. Das Phänomen des Lebens ist somit auf das Engste an die Existenz einer wie auch immer gearteten Zelle gebunden. Aufgrund dieser Eigenschaften sind Viren eindeutig von der Welt des Lebendigen ausgeschlossen. Sie sind als Einheit nicht lebensfähig – ja, sie stehen nicht einmal in einem Energie- oder Stoffaustausch mit der Umwelt. Sie sind eine Art „Trittbrettfahrer des Lebendigen", die vollständig auf ihren Wirt angewiesen sind und ohne ihn nicht existieren könnten und würden.

Für die Entstehung des Lebens gab es zwei grundlegende Voraussetzungen: die *chemische Evolution*, welche die Baustoffe des Lebens zur Verfügung stellte und die *Selbstorganisation*, infolge welcher reproduktive Einheiten entstanden. Der Selbstorganisation muss – in welcher Form auch immer – eine Information zugrunde liegen.

Informationsgehalt von DNA

Dass sich die Frage nach dem Leben nicht in einer einfachen Definition beantworten lässt, liegt nicht zuletzt in der Fülle seiner Erscheinungsformen. Ebenso komplex wie das Leben sind auch die Lebensformen. Die Komplexität eines Lebewesens wollen wir hier gleichsetzten mit dem Informationsgehalt, den es trägt. Zumindest auf der Ebene des Erbgutes gibt es heute verschiedene Ansätze, die Themen der Komplexität und des Informationsgehaltes von Lebewesen anzusprechen. Im Folgenden werden die wichtigsten Theorien und Verfahren vorgestellt, welche alle die Genomsequenz, also die lineare Abfolge der Nukleotide Adenin (A), Thymin (T), Guanin (G) und Cytosin(C) in dem Polymer DNA zur Grundlage haben.

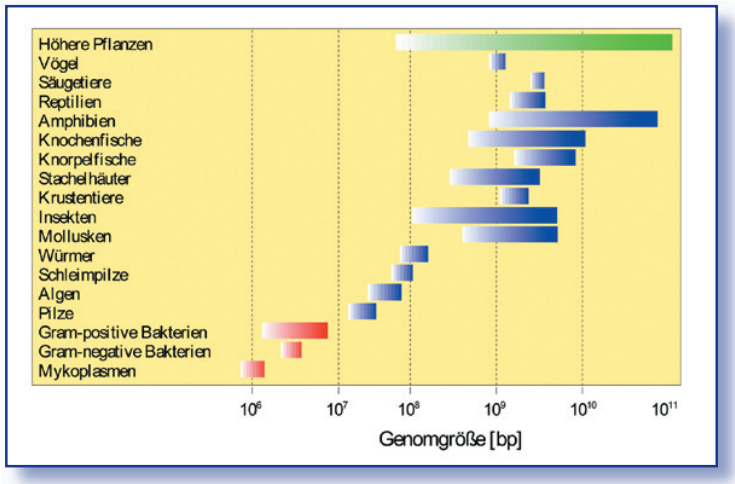


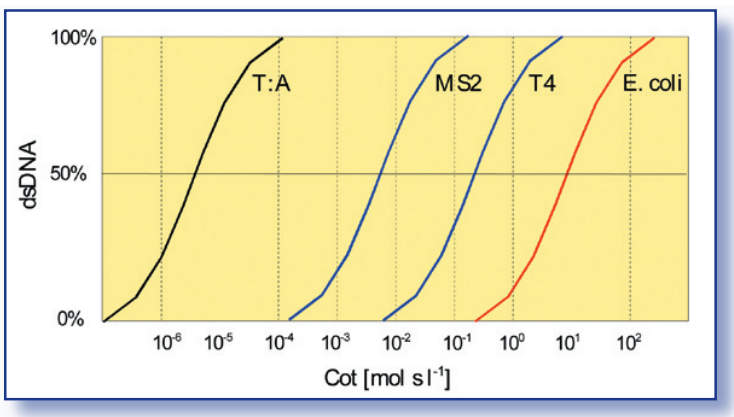
Abbildung 2: Genomgröße unterschiedlicher Organismengruppen. Daten aus [33].

C-Wert Paradox

Einen neuen Zugang zur Frage nach der Komplexität von Lebewesen lieferten die Ergebnisse der Molekularbiologie [1]. Mit relativ einfachen Methoden konnte schon in den 50er Jahren die Größe eines Genoms bestimmt werden. Die Größe des haploiden Genoms (Genom mit einem einfachen Chromosomensatz) wurde als C-Wert (*C-value*) bezeichnet [6]. Der Begriff hat heute an Bedeutung verloren und wird durch den Begriff Genomgröße (*genome size*) ersetzt. Der C-Wert wird entweder in Mega-Basenpaaren (Mb) oder in Picogramm DNA (pg) angegeben. Die erste umfassende Untersuchung zum C-Wert einer Reihe von Organismen wurde 1951 veröffentlicht [7]. Abbildung 2 gibt eine moderne Version der gefundenen Daten wieder. Es wird deutlich, dass einfache Prokaryonten wie die Mykoplasmen viel kleinere Genome haben als strukturell komplexere Prokaryonten, einfache Eukaryonten, und so weiter. Die größten Genome sind bei den Amphibien und Höheren Pflanzen zu finden. Der eukaryontische Protist *Amoeba dubia* macht eine Ausnahme und hat mit 670 Megabasenpaaren das größte bisher beobachtete Genom [8]. Insgesamt aber besteht eine positive Korrelation zwischen der strukturelle Komplexität eines Organismus und seiner Genomgröße. Allerdings ist diese Korrelation nur auf die einfacheren Organisationsebenen beschränkt. Bei den hochorganisierten eukaryontischen Vielzellern ist diese Korrelation zwischen der Organisationsstufe und der Genomgröße nicht mehr klar zu beobachten. Man spricht dabei vom C-Wert Paradox (*C-value paradox*), welches erstmals Anfang der 70er Jahre beschrieben wurde [9].

Was ist die Ursache für dieses Paradox? Die große Frage in der damaligen Zeit war, ob große Genome eine größere Zahl unterschiedlicher Gene beherbergen als kleine Genome (also komplexer, informativer sind), oder ob einfach die gleichen Gene in einer höheren Kopienzahl vorliegen (also redundanter sind). Während dieser Frage heute vergleichsweise einfach (aber kostenaufwendig) durch die Totalsequenzierung von Genomen nachgegangen werden kann, bestand diese Möglichkeit in den 70er Jahren nicht.

Abbildung 3: Reassoziationskurven (Cot-Kurven) von genomischer DNA aus unterschiedlichen Quellen. T:A bezeichnet ein künstliches Genom, das nur aus dem einem Basenpaar Thymin:Adenin besteht (schwarzer Graph). MS2 und T4 bezeichnen die Cot-Kurven der Genome von Bakteriophagen (blaue Graphen). E. coli bezeichnet die Cot-Kurve des Bakteriums Escherichia coli (roter Graph). Daten aus [10].



Cot-Analyse

Eine etablierte Methode zur Charakterisierung der Komplexität von Genomsequenzen ist die Reassoziationskinetik von denaturierter DNA [10]. Zur Analyse der Reassoziationskinetik wird die genomische DNA zunächst durch Scherkräfte in etwa 400 Basenpaare lange Abschnitte zerlegt. Die entstandenen DNA-Doppelhelix-Fragmente, bestehend aus den beiden komplementären Einzelsträngen die über Wasserstoffbrückenbindungen miteinander verbunden sind, werden anschließend in wässriger Lösung durch Hitzeeinwirkung denaturiert (siehe auch: [11][12]). Zur Denaturierung, bei der die beiden Einzelstränge der Doppelhelix voneinander getrennt werden, muss eben so viel Energie zugeführt werden, dass alle Wasserstoffbrückenbindungen aufgehoben werden. Während eines nachfolgenden Abkühlprozesses renaturiert die DNA wieder. Der Renaturierungsprozess ist von der zufälligen Kollision der komplementären Einzelstränge abhängig und folgt daher einer Kinetik der 2. Ordnung:

$$\frac{\Delta[A]}{\Delta t} = -k\,[A]\,[B] \qquad (1.1)$$

Dabei sind *[A]* und *[B]* die Konzentrationen komplementärer einzelsträngiger Sequenzen und *k* die Ratenkonstante der Reaktion. Da für doppelsträngige DNA *[A] = [B]* gilt, ergibt das Integral von (1.1):

$$\int_{[A]_0}^{[A]_t} \frac{\Delta[A]}{\Delta t} \rightarrow \frac{1}{[A]} = \frac{1}{[A]_0} + k\,t \qquad (1.2)$$

Dabei ist $[A]_0$ die Anfangskonzentration von *A*. Im Experiment wird dem Renaturierungsgemisch an definierten Zeitpunkten eine Probe entnommen und chromatographisch mittels einer Hydroxyapatitmatrix in Einzelstrang- (ssDNA, *single strand DNA*) und Doppelstrang-DNA (dsDNA, *double strand DNA*) getrennt. Die DNA-Quantifizierung erfolgt photospektrometrisch bei 260 Nanometern. Das Verhältnis *f* von dsDNA zu ssDNA beträgt:

$$f = \frac{[A]}{[A]_0} \qquad (1.3)$$

Nach Einsetzten von Gleichung (1.3) in (1.2) erhält man:

$$f = \frac{1}{1 \quad [A]_0 k\,t} \qquad (1.4)$$

Die DNA-Konzentrationen beziehen sich immer auf einzigartige, nicht-repetitive Sequenzen, da die Kollision von nicht-komplementären Sequenzen nicht zur Renaturierung führt. Daher gilt, wenn C_0 die Anfangskonzentration an Basenpaaren in der Lösung ist:

$$[A]_0 = \frac{C_0}{x} \qquad (1.5)$$

Dabei beschreibt x die Anzahl einzigartiger Sequenzen (in Basenpaaren) und gilt als Maß für die Komplexität von DNA. Zum Beispiel hat die repetitive Sequenz $(ACGT)_n$ eine Komplexität von 4, wohingegen das Genom von *Escherichia coli* mit 4,7 Millionen nicht-repetitiven Sequenzen eine Komplexität von 4,7 Millionen aufweist. Die Kombination von Gleichungen (1.4) und (1.5) ergibt:

$$f = \frac{1}{1 + C_0 t k x^{-1}} \qquad (1.6)$$

Wenn die Hälfte aller Moleküle in der Lösung renaturiert ist (f=0,5), dann gilt:

$$C_0 t_{1/2} = \frac{x}{k} \qquad (1.7)$$

wobei $t_{½}$ die benötigte Zeit beschreibt. Die Ratenkonstante k ist charakteristisch für die Frequenz der ssDNA-Kollisionen unter den gegebenen Reaktionbedingungen. k ist unabhängig von der Komplexität der DNA und, für kurze DNA-Fragmente, von der Länge der DNA (daher wird die genomische DNA vor dem Experiment in etwa 400 Basenpaare lange Fragmente geteilt). Aufgrund dessen ist $C_0t_{½}$, unter gegebenen Versuchsbedingungen nur von der Komplexität x der DNA abhängig. Es reicht demnach aus, die Anfangskonzentration der ssDNA C_0 und die Zeit $t_{½}$ zu kennen, die die DNA benötigt, um zur Hälfte renaturiert zu sein, um deren Komplexität x zu bestimmen.

Aufgrund experimentell ermittelter Daten wird das Verhältnis von ssDNA zu dsDNA gegen die Renaturierungszeit aufgetragen (Abbildung 3). Die Renaturierungsreaktion einer individuellen DNA-Spezies wird eindeutig durch die half-completion time, das heißt durch die Zeit, in der die Hälfte aller ssDNA-Moleküle renaturiert ist, beschrieben. Wie aus der Gleichung 1.7 hervorgeht ist dieser Wert, der das Produkt aus der DNA-Ausgangskonzentration C_0 und $t_{½}$ und daher $C_0t_{½}$ genannt wird, proportional zur Komplexität der eingesetzten DNA. Normalerweise wird bei einer Komplexitätsbestimmung eine genomische DNA-Probe von *Escherichia coli* parallel untersucht, um die versuchsabhängige Variabel k zu eliminieren (siehe Gleichung (1.7)). Es wird dabei davon ausgegangen, dass die Komplexität des *Escherichia coli* Genoms 4,2 Millionen Basenpaare beträgt:

Komplexität von Genom X =

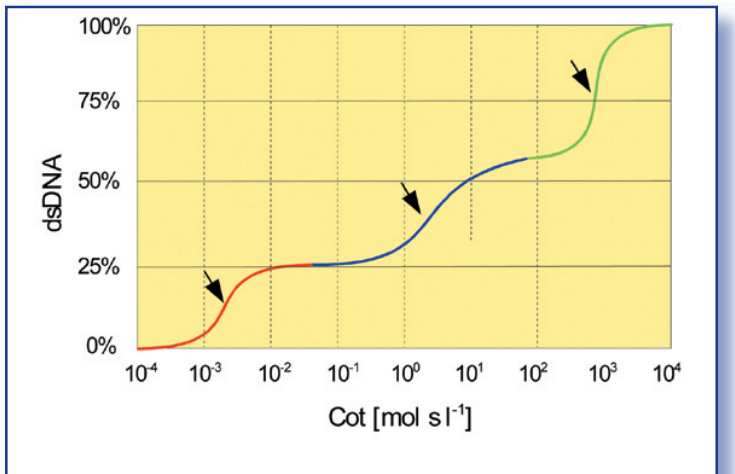
$$\frac{(C_0 t_{1/2} \text{ von Genom X}) \times (4.2 \times 10^6\, bp)}{C_0 t_{1/2} \text{ von } E.coli} \qquad (1.8)$$

Es kann durchaus vorkommen, dass genomische DNA aus Sequenzen unterschiedlicher Komplexität besteht. In diesem Fall kann eine C_0t-Kurve entstehen, wie sie in Abbildung 4 zu sehen ist. Im dargestellten Fall besteht das Genom aus drei unterschiedlich komplexen Komponenten: einzigartige, durchschnittlich-repetitive und hoch-repetitive Sequenzen. Hoch-repetitive Sequenzen kommen im Genom definitionsgemäß in einer hohen Kopienzahl vor und finden deshalb im Renaturierungsexperiment schnell einen Partner. Sie reassoziieren schnell, haben daher einen geringen $C_0t_{½}$ Wert und zeichnen sich folglich durch eine sehr geringe Komplexität aus. Umgekehrt liegt die Situation bei einzigartigen Sequenzen, die genau einmal im Genom auftreten und die entsprechend länger benötigen, um im Reaktionsgemisch ihren Partner zu finden. Dies schlägt sich in einem hohen $C_0t_{½}$ Wert und einer hohen Komplexität nieder (Abbildung 4). Unterschiedliche Komplexitätskomponenten sind typisch für eukaryontische Genome. Der Anteil nicht-repetitiver DNA schwankt in eukaryontischen Genomen erheblich (Tabelle 1).

Tabelle 1: Anteil hoch-repetitiver, durchschnittlich-repetitiver und nicht-repetitiver DNA-Anteile am Genom unterschiedlicher Organismen. Daten aus [34].

Organismus	Sequenzverteilung [%]		
	Hoch Repetitiv	Mittel Repetitiv	Einzelkopien
Bakterien			99,7
Mensch	8	13	70
Maus	10	25	60
Arabidopsis	10	27	55
Baumwolle	8	27	61
Mais	20	40	30
Weizen	4	83	10
Tomate	13	14	73

Abbildung 4: Exemplarische Reassoziationskurve (Cot-Kurve) eukaryontischer DNA. Die Pfeile markieren die $Cot_{½}$-Werte für die jeweiligen Teilabschnitte: rot: hoch-repetitiv, blau: durchschnittlich-repetitiv, grün: nicht-repetitiv. Die Komplexität der Teilabschnitte berechnet sich nach der Gleichung 1.8 im Text.



	Schnell	Medium	Langsam
Anteil des Genoms [%]	25	30	45
$Cot_{½}$	0,0013	1,9	630
Komplexität [bp]	500.000	350	1

Es ist an dieser Stelle erwähnenswert, dass die Cot-Analyse in Form des Cot-based cloning and sequencing (CBCS) ein Revival feiert [13]. Wissenschaftler nutzen die Methode, um im Vorfeld einer Genomsequenzierung informationsarme repetitive Elemente durch Hybridisierung zu entfernen und die Sequenzierung zunächst auf die verbleibende Sequenzinformation zu beschränken. Bedenkt man, dass das Genom vieler Organismen zum überwiegenden Teil aus repetitiven Elementen besteht (zum Beispiel 98 Prozent bei der Zwiebel), dann wird ersichtlich, dass viel Zeit und Geld gespart werden kann.

Mit Hilfe der Cot-Analyse konnte das oben angesprochene C-Wert Paradox weitgehend aufgeklärt werden: Vor allem bei Eukaryonten mit einem großen Genom korreliert die Genomgröße nicht mit dem Organisationsgrad beziehungsweise der Kompexität des Organismus, da ein Großteil des Genoms aus repetitiven Sequenzen besteht. Die genomische Information ist also redundant. Das C-Wert Paradox hat sich allerdings in ein bis heute nur teilweise geklärtes C-Wert Rätsel (*C-value enigma*) gewandelt [14][15]. Nach wie vor ist weitgehend unklar, warum einige Organismen sich ein großes Genom mit überwiegend repetitiven Sequenzen leisten. Immerhin kostet der Erhalt eines großen Genoms dem Organismus viel Energie (Strukturproteine, Regulation, Nukleotide).

G-Wert Paradox

Nach der Lösung des C-Wert Paradox trat ein neues Paradox zutage: das G-Wert Paradox (*G-value paradox*) [16]. Der G-Wert (*g-value*) bezeichnet die Anzahl aller Gene in einem haploiden Genom. In der Regel besteht dieser Wert aus der Summe der Anzahl der identifizierten bekannten Gene und der Anzahl der vorhergesagten offenen Leserahmen (ORFs, *open reading frames* = Startcodon bis Stopcodon). Wiederum korreliert der G-Wert nicht mit dem Organisationsgrad der Organismen (Abbildung 5) [16].

Wie sind diese Diskrepanzen zu erklären? Je mehr Organismen auf der Ebene ihrer Genome und ihres Metabolismus (Stoffwechsel) untersucht werden, desto mehr Regulationsmechanismen werden aufgedeckt. Im Folgenden sollen nur einige Beispiele gegeben werden, die für das G-Wert Paradox mitverantwortlich gemacht werden:

- **Kombinatorik** Je größer die Anzahl der Gene in einem Organismus ist, desto mehr Kombinationen von exprimierten Genen können gemeinsam komplexe Funktionen erfüllen. Geht man von 18 000 Protein kodierenden Genen aus (dies entspricht etwa der Situation beim Wurm Caenorhabditis elegans), so gibt es rund 162 Millionen paarweise Kombinationen. Einer kleiner Anstieg des G-Wertes führt bereits zu einer erheblich größeren Komplexität der resultierenden Proteinnetzwerke.
- **„Schweizer Messer Proteine“** Einige Proteine können mehr als nur eine Funktion erfüllen. Dies führt dazu, das trotz größerer biologischer Komplexität weniger Gene als erwartet ausreichen. Dies würde erklären, weshalb beispielsweise der G-Wert des menschlichen Genoms erheblich kleiner als erwartet ausfällt.
- **Alternatives Splicing** In Eukaryonten ist ein Protein-kodierendes Gen oft in Exons und Introns unterteilt. Nach der Transkription werden auf der mRNA-Ebene die Introns herausgetrennt und die Exons zusammengefügt (Abbildung 6). Dieser Prozess wird als Splicing oder mRNA-Reifung bezeichnet. Werden Exons unterschiedlich wieder zusammengefügt, so spricht man vom alternativen Splicen (Abbildung 6). Auf diese Weise kann ein

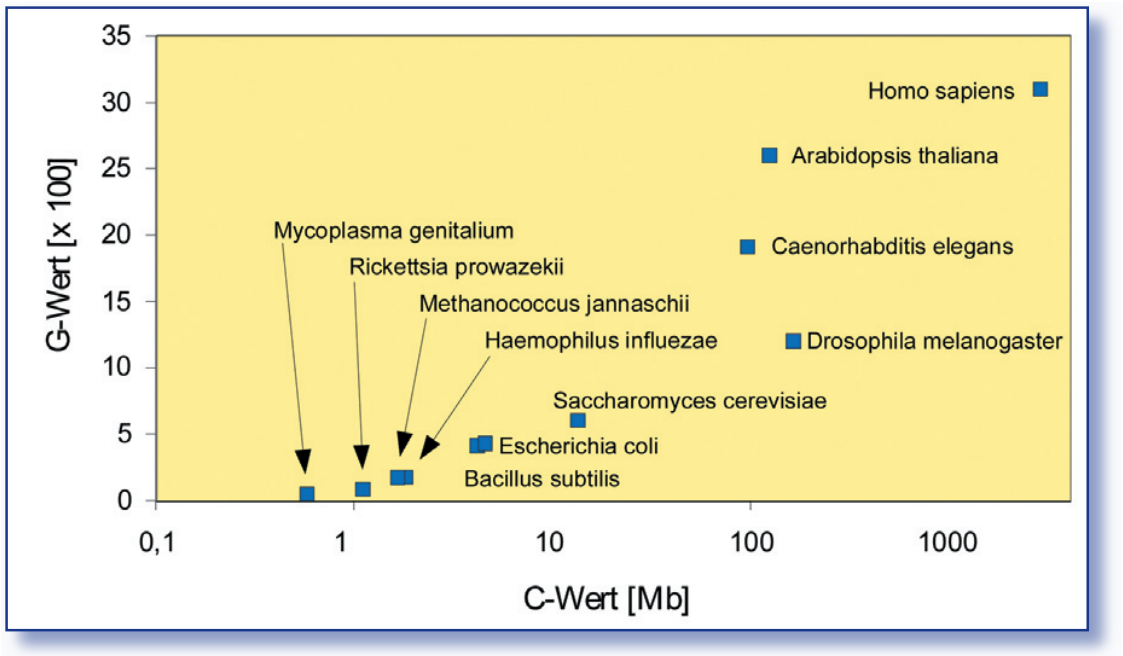


Abbildung 5: G-Wert Paradox. Der C-Wert, die Größe des Genoms, korreliert nicht mit dem G-Wert, der Anzahl aller Gene im Genom. Daten aus [33].

Gen für mehrere Proteine kodieren. Abschätzungen zufolge unterliegen 59 Prozent der Gene des menschlichen Genoms dem alternativen Splicing.

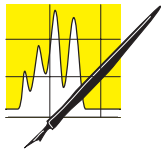
- **Posttranslationale Modifikation** Nach der Proteinbiosynthese werden die entstanden Proteine oftmals modifiziert. Beispielsweise kann von dem Protein ein Peptid abgetrennt oder ein Kohlenhydrat angehängt werden. Diese posttranslationalen Modifikationen führen wiederum dazu, dass ein Gen letztlich für mehr als eine Proteinsorte kodieren kann.
- **Promotoren** Eine Vielzahl von Genen scheint mehrere Regionen zu besitzen (Promotoren), die deren Expression steuern. Das heißt, ein Gen kann mehreren unterschiedlichen Regulationsmustern unterliegen.
- **Unterschätzung** Es ist nicht ausgeschlossen, eher sogar sehr wahrscheinlich, dass die Komplexität des Genoms über die Komplexität seiner Sequenz hinausgeht. Niemand würde heute abstreiten, dass die eigentliche Komplexität der Proteine in ihrer dreidimensionalen Struktur begründet liegt. In der Genomforschung sind diese oder ähnliche Formen der Komplexität, die über die reine Sequenzinformation hinausgehen, bisher kaum untersucht worden.

Eine ganz neue Form der Komplexität von Genomen, nämlich auf der Basis der RNA, schlägt der australische Biologe John Mattick von der „University of Queensland“ vor [17].

RNA: Motor der Komplexität

Das zentrale Dogma der Biologie aus den 50iger Jahren besagt, dass der Informationsfluss von der DNA über die RNA zum Protein stattfindet (siehe auch: [1]). Seitdem haben zahlreiche Versuche gezeigt, dass die RNA mehr Aufgaben erfüllt als den simplen Transfer von Information. Dazu zählen zahlreiche strukturelle und regulative Funktionen. Die RNA für diese Aufgaben wird von eigenen Genen auf der DNA kodiert. In der Zelle befindet sich jedoch auch eine große Menge RNA, der bislang jegliche Funktion abgesprochen wurde: die transkribierten Introns (Abbildung 6). Dies trifft vor allem für Eukaryonten zu. Die meisten Gene der höheren Organismen sind mosaikartig aufgebaut. Die als Exons bezeichneten Abschnitte kodieren ein Protein, die Introns tragen hingegen nicht zum Aufbau der betreffenden Proteine bei. Oft sind die Introns viel umfangreicher als die Exons, so dass letztere wie kleine Informationsstücke in ein Meer von Introns eingebettet erscheinen. Aus der mRNA werden die Introns mit speziellen Enzymen herausgetrennt. Nur intronfreie mRNA verlässt den Zellkern.

In der folgenden Ausgabe wird die Rolle der Informationstheorie in der Biologie näher beleuchtet. Wie kann der Informationsgehalt einer DNA-Sequenz beschrieben werden? Wie erhält man eine Idee von der kleinsten notwendigen Genomgröße? Und wenn sie bekannt ist, welche wissenschaftliche Erkenntnis können Forscher daraus gewinnen?



Literatur

[1] Wünschiers R, Wünschiers C & Borzner S (2000) CLB 51: 138-143
[2] Schrödinger, E. (1944) What is live? Cambridge University Press
[3] Brockhaus Enzyklopädie (1970) F. A. Brockhaus
[4] Jacque Monod (1971) Zufall und Notwendigkeit. R. Piper & Co. Verlag
[5] Christian de Duve (1991) Blueprint for a Cell: The Nature and Origin of Life. Neil Patterson Publishers
[6] Swift H (1950) Proc. Natl. Acad. Sci. USA 36: 643-654
[7] Mirsky AE & Ris H (1951) J. Gen. Physiol. 34: 451-462
[8] Wen-Hsiung Li (1997) Molecular Evolution. Sinauer Associates, Inc.
[9] Thomas CA (1971) Annu. Rev. Genet. 5: 237-256
[10] Britten RJ & Kohne DE (1968) Science 161: 529-40
[11] Zinn T, Wünschiers R & Borzner S (2000) CLB 51: 328-333
[12] Wünschiers R, Zinn T & Borzner S (2001) CLB 52: 260-266
[13] Peterson DG et al. (2002) Genome Research 12: 795-807
[14] Gregory TR (2000) Genome 43: 895-901
[15] Gregory TR (2001) Biol. Rev. 76: 65-101
[16] Hahn MW & Wray GA (2002) Evol. Develop. 4: 73-75
[17] Mattick JS (2001) EMBO Rep. 2: 986-991
[33] Benjamin Lewin (2000) Genes VII. Oxford University Press

Abbildung 6: In einem typischen eukaryontischen Gen ist die Protein-kodierende Sequenz (Exon, E-1-3) von mehreren Introns (I-1-2) unterbrochen. Nach der Transkription entsteht eine vorläufige mRNA (prä-mRNA), aus der die Introns während des Splicing-Prozesses heraus getrennt werden. Normalerweise werden nur Introns heraus getrennt (mRNA-1), jedoch können beim alternativen Splicing auch Exons entfernt werden (mRNA-2), was zur Bildung unterschiedlicher Proteine führen kann. Entscheidend für das Splicing sind Erkennungssequenzen im Intron. Die meisten Introns sind durch die Nukleotide GU und AG an den Intron-Grenzen gekennzeichnet. Innerhalb der Introns existieren ebenfalls Erkennungssequenzen, die am Splicing beteiligt sind. Die Intron-Exon-Grenzen werden als Donor und Akzeptor bezeichnet.