Die großen Geheimnisse kleiner Dinge

G(e)nomforschung (Teil 2)

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? Forschungsgruppen in aller Welt suchen nach Antworten auf das informationstheoretische Rätsel des Lebens.

Informationstheorie

Für die Informationstheorie ist die Biologie ein relativ altes Territorium. Es wurde insbesondere durch die Gedanken von Erwin Schrödinger über die Bedeutung der Entropie in der Biologie eröffnet [2]. Damit wurde ein Zusammenhang zwischen Leben und Information geschaffen. Wir vermuten heute, dass in dem Erbgut, dem Genom, alle Informationen enthalten sind, die der Organismus benötigt. Das Genom wird von Generation zu Generation weitergereicht und erfährt infolge von Mutationen Veränderungen, die letztlich die Grundlage der biologischen Evolution bilden. Wir können also versuchen, das Problem der Komplexität der Lebewesen auf das Problem der Komplexität der Genome, bzw. deren Informationsgehalt, zu reduzieren.

Wie kann der Informationsgehalt einer DNA-Sequenz beschrieben werden? Einem gängigen Ansatz liegt die Informationstheorie nach Claude Shannon zugrunde [19]. Danach ist Informationsmenge die Zahl binärer ja/nein-Entscheidungen, die man im Mittel benötigt, um eine bestimmte Symbolabfolge zweifelsfrei zu identifizieren. Bevor wir ein Symbol kennen, besteht eine Unsicherheit (*uncertainty*) ob des Symbols. Kennen wir das Symbol, so sinkt unsere Unsicherheit und wir haben Information erhalten. Information ist also eine Verringerung der Unsicherheit. Bei einer Auswahl aus vier Symbolen haben wir eine Unsicherheit von vier Symbolen pro Symbol. Es ist einfacher von Information in bit zu reden. Eine ja/nein-Entscheidung entspricht somit einem bit (1 oder 0). Nehmen wir folgende Symbole, also die vier Nukleotide an:

A C G T (2.1)

Um die Anzahl der notwendigen binären Fragen pro Nukleotid zu ermitteln berechnen wir:

$$\log_2 (M) \qquad (2.2)$$

Dabei bezeichnet M die Anzahl der Symbole (Nukleotide). Das heißt, im Falle der Nukleotide aus (2.1) belegt jedes Nukleotid 2 bit, was sich wie folgt veranschaulichen lässt:

A → 00 C → 01 G → 10 T → 11 (2.3)

Daher lässt sich die folgende Sequenz schreiben als:

ACATGAAC → 00010011100000001 (2.4)

Die Sequenz belegt somit 16 bit. Es macht aber durchaus Sinn zu berücksichtigen, mit welcher Wahrscheinlichkeit ein Nukleotid vorkommt. Nach Umformung erhält man:

$$\begin{aligned} \log_2(M) &= -\log_2(M^{-1}) \\ &= -\log_2\left(\frac{1}{M}\right) = -\log_2(P) \end{aligned} \qquad (2.5)$$

sodass $P=1/M$ die Wahrscheinlichkeit des Auftretens eines Nukleotids ist. Für die vier Nukleotide muss somit gelten:

$$\sum_{i=1}^{4} P_i = 1 \qquad (2.6)$$

Die Überraschung (*surprisal,* u_i) das i-te Nukleotid zu sehen beträgt entsprechend zu (2.2),

$$u_i = -\log_2(P_i) \qquad (2.7)$$

Nähert sich P_i gegen 0, so werden wir sehr überrascht sein, weil die Häufigkeit von Nukleotid i sehr gering ist. Ungewissheit ist die durchschnittliche Überraschung über das Auftreten eines Nukleotids in einer unendlich (wegen der Statistik) langen DNA-Sequenz. Nehmen wir an, wir haben eine DNA-Sequenz mit N Nukleotiden. Nehmen wir weiter an, dass das i-te Nukleotid N_i-mal in der Sequenz vorkommt, sodass gilt:

$$N = \sum_{i=1}^{4} N_i \qquad (2.8)$$

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.

Dann haben N_i-Fälle die Überraschung u_i. Die durchschnittliche Überraschung für alle N Nukleotide beträgt demnach:

$$\frac{\sum_{i=1}^{4} N_i u_i}{\sum_{i=1}^{4} N_i} \quad \rightarrow \quad \text{mit (2.8)} \quad \rightarrow \quad \sum_{i=1}^{4} \frac{N_i}{N} u_i \tag{2.9}$$

Für eine unendlich lange DNA-Sequenz wird die Häufigkeit N_i/N gleich der Wahrscheinlichkeit P_i des i-ten Nukleotids. Mit dieser Substitution und der Substitution von u_i durch (2.7) erhalten wir:

$$H = -\sum_{i=1}^{4} P_i \log_2 P_i \quad \text{[bit pro Nukleotid]} \tag{2.10}$$

wobei H die durchschnittliche Überraschung ist und auch als *Entropie* bezeichnet wird. Dies ist die berühmte Formel von Claude Shannon, dem Begründer der Informationstheorie [19]. Nehmen wir an, die vier Nukleotide treten in einer Sequenz mit den folgenden Wahrscheinlichkeiten auf:

$$P_A = \frac{1}{2} \quad P_C = \frac{1}{4}$$
$$P_G = \frac{1}{8} \quad P_T = \frac{1}{8} \tag{2.11}$$

Nach (2.7) beträgt die Überraschung dann

$u_A = 1$ Bit $\quad u_C = 2$ Bit $\quad u_G = 3$ Bit $\quad u_T = 3$ Bit (2.12)

Die durchschnittliche Überraschung pro Nukleotid:

$$H = \frac{1}{2}\times 1 + \frac{1}{4}\times 2 + \frac{1}{8}\times 3 + \frac{1}{8}\times 3 = 1{,}75 \tag{2.13}$$

Rekodiert man nun die Nukleotide mit binären Symbolen, deren Anzahl der Anzahl der Überraschungs-Bits entspricht:

A → 1 $\quad$ C → 01 $\quad$ G → 000 $\quad$ T → 001 (2.14)

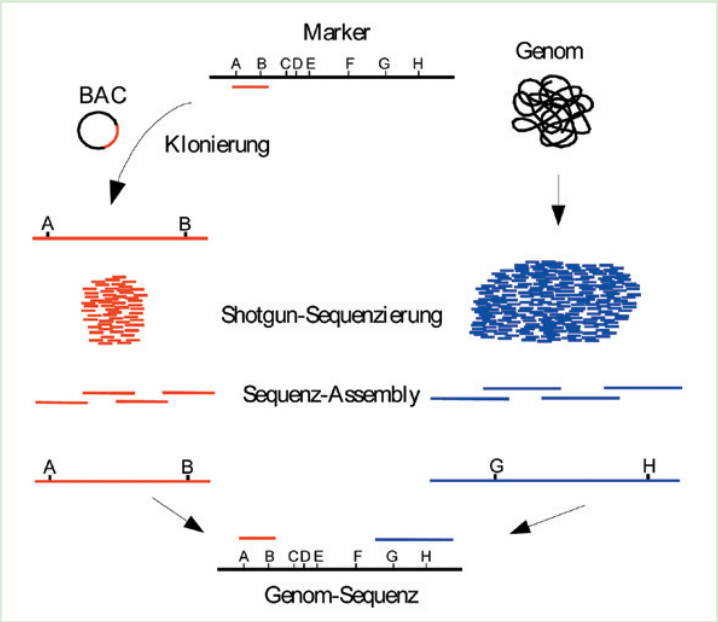
Dann kann die Sequenz aus (2.4) unter Anwendung der Nukleotidwahrscheinlichkeiten aus (2.11) wie folgt geschrieben werden:

ACATGAAC → 10110010001101 (2.15)

Sequenzierung von Genomen

Derzeit dominieren zwei Methoden zur Sequenzierung von Genomen: die clone-by-clone (Klon-nach-Klon) Sequenzierung und die whole-genome-shotgun-Sequenzierung (Gesamtgenom-Schrotflinten-Methode). Sie unterscheiden sich vor allem in dem zeitlichen Aufwand, der im ersten Fall im Labor und im anderen Fall vor dem Computer aufgewendet werden muss. Bei der Sequenzierung des menschlichen Genoms kamen beide Methoden zu Einsatz: Das Human Genome Project Consortium verfolgte den clone-by-clone Weg, während Celera Genomics die whole-genome-shotgun Methode wählte.

Für die clone-by-clone Methode wird das Genom (in Realität mehrere 100 Genome) zunächst enzymatisch in rund 150 Basenpaare große Fragmente partiell „verdaut". Der Verdau, vermittelt durch DNA-spaltende Restriktionsenzmye, erfolgt derart, dass sich überlappende Fragmente ergeben. Diese Fragmente werden in künstliche bakterielle Chromosomen (BACs, bacterial artificial chomosomes) kloniert, in Bakterien vervielfacht und anschließend nach Markern durchsucht. Solche Marker können zum Beispiel Schnittstellen für Restriktionsenzme sein. Anhand dieser „Kartierung", die in einer „physikalischen Karte" (physical map) mündet, kann bestimmt werden, welche Fragmente sich überlappen. Die Fragmente werden nun nach der Shotgun-Methode in noch kleinere Fragmente zerlegt, die sequenziert werden. Schließlich werden zunächst die BAC-Fragment Sequenzen zusammengefügt, die, aufgrund der Markerinformationen, zur Genomsequenz zusammengefügt werden können.



Bei der whole-genome-shotgun Methode wird das gesamte Genom direkt in kleine Fragmente zerlegt und sequenziert. Auf eine Kartierung des Genoms wird also verzichtet. Stattdessen musste Celera Genomics die Weltklasse Mathematiker für sich gewinnen, um aus der ungeheuren Menge aus kurzen Sequenzen die Genomsequenz zusammen zu puzzeln.

Beide Methoden haben Vor- und Nachteile. Insbesondere bei Chromosomabschnitten mit einem hohen Anteil repetitiver Sequenzen (zum Beispiel Chromosom Y) schneidet die aufwendigere clone-by-clone Methode zur Zeit noch besser ab.

Im Gegensatz zu (2.4) belegt die Sequenz jetzt nur noch 14 bit. Der Informationsgewinn stammt aus der Kenntnis der relativen Häufigkeiten der Nukleotide. Der verwendete Code wird nach seinem Erfinder als Fano Code bezeichnet und zeichnet sich dadurch aus, dass zwischen den Symbolen keine Trennzeichen angegeben werden müssen.

Dieses Beispiel sollte verdeutlichen, wie die Informationstheorie zur Bestimmung des Informationsgehaltes von einer DNA-Sequenz angewendet werden kann. Informationstheoretische Untersuchungen von Genen und Genomen haben bereits interessante Ergebnisse hervorgebracht. Aufgrund der beobachteten Frequenzverteilung der Nukleotide benötigt die Kodierung eines Nukleotids 1,95 bit, gegenüber 2 bit bei einer Zufallssequenz. In Protein-kodierenden Sequenzen scheint die Entropie generell geringer zu sein.

Bestimmte Sequenzen des menschlichen Genoms (*human splice acceptor sites*), die für das Splicen von Introns und Exons verantwortlich sind (Abbildung 6), tragen eine Information von etwa 9,4 bit verteilt auf 40 Basenpaare (0,235 bit pro Nukleotid) [20]. Warum gerade 9,4 bit? Es ergab sich, dass die *human splice acceptor sites* im Mittel 812 bp voneinander getrennt waren. Die Information, die benötigt wird, diese Sequenzen zu finden beträgt also $\log_2 812 = 9{,}7$ bit. Diese Ergebnisse zeigen, dass es einen einfachen Zusammenhang zwischen der Art und der Anzahl von Bindungsstellen auf der einen Seite und der Größe des Genoms auf der anderen Seite gibt. Später konnte gezeigt werden, dass dies eine generelle Eigenschaft von DNA-Bindungsstellen ist [21].

Weiterhin ermöglicht die Informationstheorie die Validierung statistischer Modelle über die Verteilung von Nukleotiden in Genomen [22]. Solche Modelle finden auch bei der Suche nach Genen und regulativen Einheiten in sequenzierten Genomen eine große Rolle (Gen-Annotierung) und bilden auf der großen Spielwiese der Bioinformatik eine Schnittstelle zwischen statistischer Physik und Genetik.

Die alleinige Betrachtung des Informationsgehaltes von Genomen wird uns aber kaum eine Antwort auf die minimal notwendige Größe von Genomen liefern. Dazu wird es notwendig sein, mehr Informationen über den Informationsgehalt der Lebensprozesse zusammenzutragen und beschreiben zu können. Daher ist die Suche nach dem kleinsten notwendigen Genom noch rein empirischer Natur und, wie im folgenden dargestellt wird, vor allem von *try-and-error* geprägt.

Minimal-Genom Projekt

Die ersten Experimente, um eine Idee von der kleinsten notwendigen Genomgröße eines Lebewesens zu erhalten, führte der japanische Biologe Mitsuhiro Itaza in den Laboren von Mitsubishi mit dem Bakterium *Bacillus subtilis* durch [23]. Diesen, im Jahr 1995 veröffentlichten Ergebnissen, lag noch kein sequenziertes Genom zugrunde. Nach dem Zufallsprinzip schalteten die Mitarbeiter um Itaza Gene von *Bacillus subtilis* aus und berechneten auf Basis dieser Daten eine minimal notwendige Ausstattung aus 256 Genen. Aufgrund der Ende 1997 veröffentlichten Genomsequenz von *Bacillus subtilis* kennen wir heute die Gesamtzahl aller Gene: rund 4100. Es liegt auf der Hand, dass die detaillierte Kenntnis der Genomsequenz und der Annotation (Suche und Beschreibung der Gene) der Suche nach dem minimalen Genom zugute kommt.

Genomvergleiche

Unter der Leitung von Craig Venter gelang es einer Arbeitsgruppe am amerikanischen Intitute for Genomic Research (TIGR) 1995 zum ersten mal, das Erbgut eines Organismus vollständig zu entschlüsseln: alle 1 830 138 Basenpaare des Bakteriums *Haemophilus influezae* KW20 (ein Bakterium, das die Atemwege des Menschen befällt) wurden sequenziert. Rund 1740 Gene konnten identifiziert werden. Kurze Zeit später wurden von Mitarbeitern desselben Instituts alle 580 074 Basenpaare des Bakteriums *Mycoplasma genitalium* sequenziert. Hier konnten 517 Gene identifiziert werden. 480 Gene kodieren für Proteine, 37 Gene für verschiedene RNAs (zum Beispiel transfer-RNA und ribosomale-RNA). Bis heute ist das Genom von *Mycoplasma genitalium* das kleinste bekannte eines sich selbstständig, unabhängig vermehrenden Organismus. Ist bei diesem Bakterium das untere Limit erreicht, oder besteht die Möglichkeit, dass eine Zelle mit noch weniger Genen auskommen könnte? Dieser Frage gehen mehrere Arbeitsgruppen im „Minimal Genom Projekt“ nach. Ein erster Eindruck wurde durch den Computer-gestützten Vergleich der Gene der ersten beiden sequenzierten Genome erhalten [24][25]. Der Vergleich basiert auf der Annahme, dass die Schnittmenge der ähnlichen, also sehr wahrscheinlich funktional verwandten, Gene für beide Organismen

Repetitive Sequenzen

Der hohe Anteil repetitiver Sequenzen in eukaryontischen Genomen ist ein besonderes Rätsel. Für eine große Familie von rund 300 Basenpaare langen repetitiven Sequenzen, den Alu-Sequenzen, wurde kürzlich eine Funktion vorgeschlagen. Sie sind mit über 500 000 Kopien über das gesamte menschliche Genom verteilt. Ihren Namen erhielt die Alu-Familie, da alle Sequenzen einen kurzen DNA-Abschnitt gemein haben, der von dem Restriktionsenzym AluI erkannt wird. Wie fast alle repetitiven Sequenzen wurden die Alu-Sequenzen lange als DNA-Müll (junk DNA) bezeichnet. Neueren Untersuchungen zur Folge könnte diesen Sequenzen jedoch eine wichtige Rolle bei der Zellteilung zukommen [35]. Bevor sich zwei Tochterzellen bilden, müssen die Chromosomen zunächst verdoppelt (Replikation) und dann voneinander getrennt und in die Tochterzellen „gezogen“ werden. An diesem Prozess sind eine Reihe von Proteinkomplexen beteiligt. Wie es scheint, heftet sich das Protein Kohäsin an die Alu-Sequenzen und beteiligt sich später an der Trennung der neuen von den alten Chromosomen.

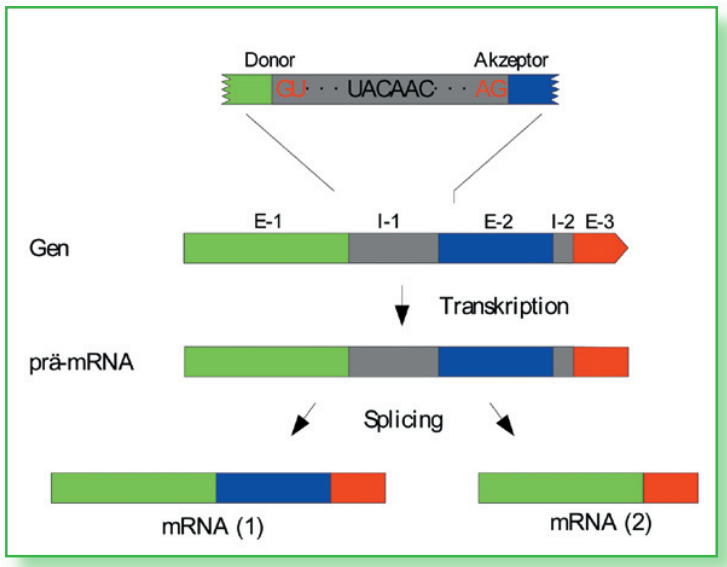
essentiell ist. Auf diese Weise wurden 256 Gene beschrieben, die für die grundlegenden Lebensprozesse notwendig sein sollten. Erstaunlicher Weise entspricht diese Zahl exakt der von Mitsuhiro Itaza postulierten Zahl. Eine genauere Analyse der Genprodukte (Proteine, Enzyme) ergab, dass die Nährstoffansprüche des minimalen Organismus groß wären: alle Aminosäuren (Bausteine der Proteine), Nukleotide (Bausteine der DNA und RNA), Fettsäuren (Bausteine der Fette) und viele Koenzyme (wie Vitamine) müssten im Nährmedium vorhanden sein [25]. In einer späteren Untersuchung wurden 21 vollständig sequenzierte Genome auf Proteinebene miteinander verglichen [26]. Dies führte zu einer Liste von 51 proteinkodierenden Genen, die allen 21 Genomen gemein, also konserviert waren. Rund 70 Prozent dieser konservierten Proteine sind am Aufbau der Ribosomen beteiligt, den zellulären Proteinfabriken. Es wurde sofort deutlich, dass diese 51 Proteine nicht für eine funktionsfähige Zelle ausreichen würden, da wichtige Enzyme des Energiestoffwechsel nicht in der Liste enthalten waren. Eine Quelle des Fehlers liegt darin begründet, dass Proteine, die sich in ihrer Aminosäuresequenz und Struktur stark voneinander unterscheiden, durchaus dieselbe Funktion erfüllen können. Diese Ergebnisse führten zu einem experimentellen Ansatz, um der Frage nach dem minimalen Genom nachzugehen.

Mutanten Analyse

Der experimentelle Versuch, Hinweise auf die minimal notwendige Genomgröße zu erhalten, basiert auf der *Knockout-* oder *Deletionsmutanten-Analyse*. Hierbei werden Mutanten (genetisch veränderte Organismen) erzeugt, in denen jeweils ein Gen zerstört (Knockout durch Transposon-Mutagenese) oder entfernt (deletiert) beziehungsweise durch ein Markergen ersetzt ist. In beiden Fällen kann das Gen seine Funktion nicht mehr erfüllen, das heißt, es kommt nicht zur Synthese des entsprechenden Proteins oder der ribosomalen RNA. Im Falle von *Mycoplasma genitalium* mussten also 517 Mutanten erzeugt werden [27]. Tatsächlich können aber nicht alle Mutanten überleben, da in einigen Fällen essentielle Gene betroffen sind. Eine Analyse von 1291 zufällig erzeugten, lebensfähigen Mutanten ergab, dass 93 Gene zerstört wurden. Dies ergibt 480-93=387 essentielle proteinkodierende Gene. Eine weiterführende Untersuchung mit einem Vergleich des Genoms des nahe verwandten *Mycoplasma pneumoniae* führte schließlich zu der Vermutung, dass 265-350 der 480 proteinkodierenden Gene von *Mycoplasma genitalium* unter Laborbedingungen essentiell sind. Erstaunlicherweise befinden sich unter diesen rund 100 Gene, deren Funktion bis heute völlig unbekannt ist. Dies zeigt einmal mehr, wie wenig wir von den lebenswichtigen Vorgängen in einer Zelle wissen.

Eine vergleichbare Untersuchung wird derzeit vom „*Genome Deletion Project*“ Konsortium mit der höher entwickelten, eukaryontischen Bäckerhefe (*Saccharomyces cerevisiae*) durchgeführt. Gegenwärtig sind etwa 95 Prozent der circa 6200 Gene ausgeschaltet worden [28]. Um nicht jede Mutante einzeln untersuchen zu müssen, wurden sie mit einer Art molekularen Barcode ersetzt. Jedes Gen, dass aus dem Genom entfernt wird, wird durch ein Markergen (Antibiotikaresistenz) plus einer Identifizierungssequenz (eben dem Barcode) ersetzt. Der Barcode ist zusätzlich von zwei konservierten Sequenzen flankiert (Abbildung 7). Jetzt werden alle Mutanten gemeinsam in einem Fermenter unter bestimmten Bedingungen angezogen. Nach etwa 30 Generationen wird die DNA aller Zellen isoliert und die Identifizierungssequenz per Polymerase-Kettenreaktion (PCR, siehe auch [11]) amplifiziert. Mittels eines DNA-Chips (siehe auch [12]) können nun alle Identifizierungssequenzen der Hefezellen nachgewiesen werden, die überlebt haben. Aufgrund der bisherigen Forschungsergebnisse sind etwa 19 Prozent aller Gene für das Wachstum essentiell. Dies entspricht fast 1200 Genen. Die Zerstörung von rund 66 Prozent aller untersuchten Gene zeigte keinen messbaren Effekt auf das Zellwachstum, während etwa 15 Prozent der Deletionsmutanten ein verändertes Wachstumsverhalten zeigten. Betroffen sind unerwartet viele Gene, die den Proteinbiosynthese-Apparat und die zellulare Organisation betreffen. Gene des Zellmetabolismus (Stoffwechsel) sowie unbekannte Gene sind in einem weit geringeren Maße betroffen, als deren prozentualer Anteil am Genom erwarten lassen würde (Abbildung 8).

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.



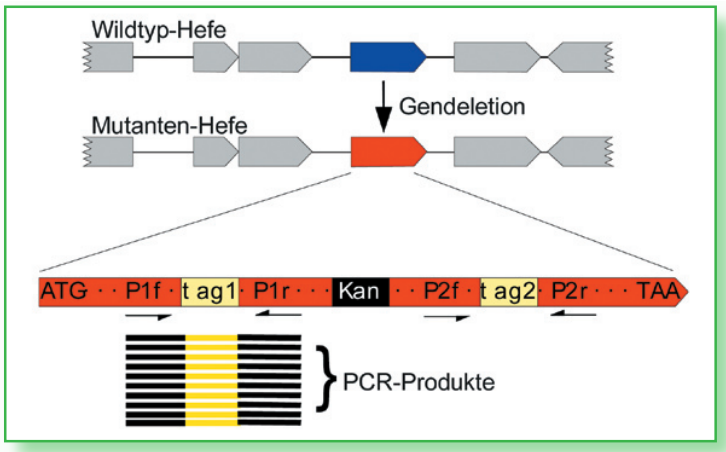
Warum zeigt ein Großteil der Gendeletionen keinen Effekt? Eine Möglichkeit ist, dass die betroffenen Gene in doppelter oder sogar mehrfacher Kopienzahl im Genom vorliegen. Dies kann beispielsweise durch eine im evolutiven Maßstab kürzlich erfolgte Genverdopplung der Fall sein. Es besteht somit die Möglichkeit, dass beide Gene noch dieselbe Funktion erfüllen. Auch gibt es von einigen Enzymen Isoformen, die sich lediglich in ihrer Regulation oder Aktivität unterscheiden. In solchen Fällen ist die Information im Genom redundant. Ein wichtiger Aspekt ist auch die Art und Weise, wie die Experimente durchgeführt werden. Unter Laborbedingungen können die natürlichen Umweltbedingungen nicht annähernd simuliert werden. Die Wahrscheinlichkeit ist groß, dass man nicht die Anzuchtbedingungen trifft, unter denen das Gen eine essentielle Rolle gespielt hätte. Auch lässt sich in Laborexperimenten kaum nachweisen, ob ein Gen einen Einfluss auf die Fitness eines Organismus hat. Die Fitness beschreibt die Anpassung eines Organismus an seinen Lebensraum und schlägt sich direkt in der Anzahl seiner fertilen Nachkommen wieder. Ist der Effekt eines Gens auf die Fitness seines Trägers nur gering, so wird sich dies nicht, oder erst nach vielen Generationen, und nur unter bestimmten (normalerweise unbekannten) Bedingungen nachweisen lassen.

Daher schlagen einige Wissenschaftler eine andere Route ein und versuchen die Entwicklung von Genomen in der Natur zu verfolgen, statt sich auf Laborexperimente zu verlassen.

Symbiose

Einiges über das minimale Genom kann von einem Blick in die Natur gelernt werden. Ein prominentes Forschungsobjekt in dieser Hinsicht sind Blattläuse. Wie sicherlich allen leidlich bekannt ist, saugen Blattläuse Pflanzensäfte. Die dort enthaltenen Kohlenhydrate sind teilweise aber schwer zu verdauen. Daher sind Blattläuse vor rund 200 Millionen Jahren eine Symbiose mit unter anderem gamma-Proteobakterien der Gattung *Buchnera* eingegangen. Die *Buchnera*-Bakterien leben in speziellen Zellen, den Bakteriozyten, in den Blattläusen und sind somit in einer sicheren Umwelt. Zusätzlich werden sie optimal mit Nährstoffen, wie mit eben jenen pflanzlichen Kohlenhydraten, versorgt. Die Blattläuse erhalten im Gegenzug niedermolekulare Kohlenhydrate, die sie problemlos verstoffwechseln können. Interessanter Weise werden Blattlausembryos von ihrer Mutter, und nicht aus der Umwelt, mit *Buchnera*-Bakterien infiziert. Das heißt, die Evolution der symbiontischen *Buchnera*-Bakterien beschränkt sich seit 200 Millionen Jahren auf optimale Anpassungen an die Bakteriozyten und es hat offenbar kein Genaustausch mit freilebenden *Buchnera*-Populationen stattgefunden. Neueren Untersuchungen zufolge hat während dieser Zeit eine drastische Reduktion der Genomgröße stattgefunden und dieser Prozess scheint noch nicht abgeschlossen zu sein [29]. Bestimmungen der Genomgröße von *Buchnera*-Bakterien aus sechs Blattlausarten ergaben Werte unterhalb der von *Mycoplasma genitalium*. Es sind somit die kleinsten bislang bekannten Genome überhaupt. Das kleinste *Buchnera*-Bakterien Genom ist lediglich 450 000 Basenpaare groß und kodiert Abschätzungen zur Folge für knapp 400 Proteine. Aber, im Gegensatz zu den einfachsten bekannten Lebewesen, den Mycoplasmen, deren Genom für rund 500 Proteine kodiert, können *Buchnera*-Bakterien nicht frei leben, sondern sind obligat an einen Wirt gebunden.

Abbildung 7: Gendeletionsmutanten von der Hefe *Saccharomyces cerevisae*. Das zu deletierende („löschende") Gen (blau) wird durch ein anderes ersetzt (rot), welches ein für eine Antibiotikaresistenz (Kan) kodiert. Zusätzlich trägt das neue Gen zwei molekulare Barcodes (tag1 und tag2), die jeweils von Primerbindungsstellen (P1l P1r, P2l, und P2r) umgeben sind. Jede Mutante besitzt ihren eigenen molekularen Marker, der durch PCR-Primer (kleine Pfeile) amplifiziert und zur Identifikation verwendet werden kann.



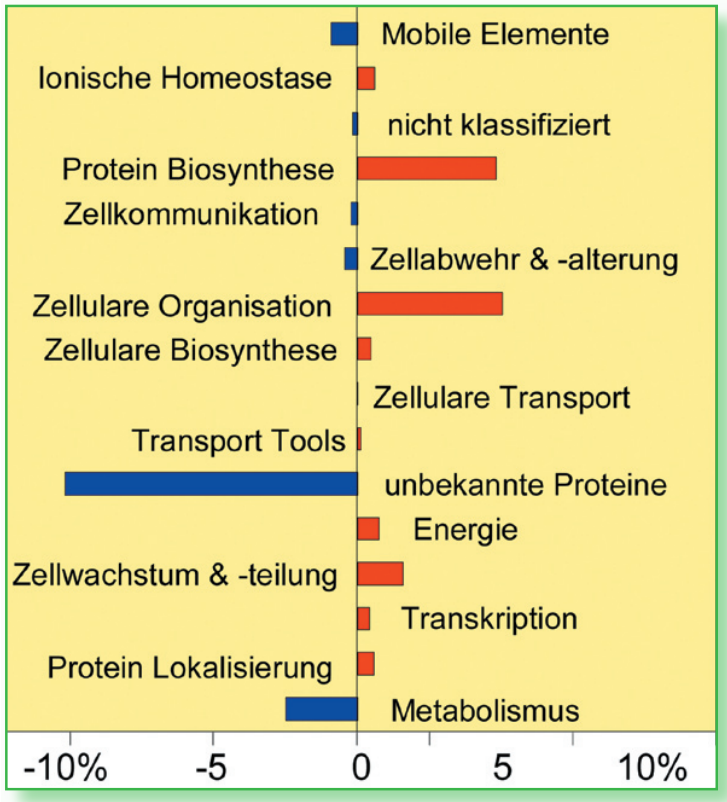
Künstliche Organismen

Fassen wir die bislang vorgestellten Ergebnisse zusammen, so schwankt die minimal notwendige Genomgröße zwischen 256 und rund 1200 Genen. Der bereits oben angesprochenen Schwäche, dass die Knockout- oder Deletionsmutanten-Analyse keine endgültige Aussage über die Funktion des betreffenden Gens zulässt, will dass Forschungsteam um Craig Venters nun auf andere Weise begegnen: durch die Synthese eines künstlichen Organismus. Nachdem durch die beschriebenen Experimente und zusätzliche Computermodelle der vermutlich minimale Gensatz gefunden wurde, muss er synthetisiert und zu einem Genom zusammengefügt werden. Man darf diese Arbeitsschritte getrost als die einfacheren bezeichnen. Die Synthese von Genen (also langkettigen Nukleotidpolymeren) wird heute bereits gewerblich angeboten und von vielen Wissenschaftlern als Dienstleistung genutzt. Kürzlich ist es einer Arbeitsgruppe an der „State University of New York" sogar gelungen, das rund 7500 Basenpaare große Genom des Polioviruses komplett synthetisch herzustellen und daraus infektiöse Viren zu erzeugen [30]. Abgesehen davon, dass

dies eine alarmierende Nachricht für die Initiative zur Bekämpfung des Bioterrorismus ist, zeigt dieser Versuch, dass es durchaus möglich ist synthetisch Genome zu assemblieren. Was beim Virus gelungen ist – eine funktionale Einheit zu erzeugen – dürfte im Falle eines künstlichen Bakteriums nicht so einfach sein. Mit einem nackten Genom, mit der gesamten Information die eine Zelle ausmacht, ist noch keine Zelle geboren. Vielmehr wird es notwendig sein, das Genom in eine bereits lebensfähige Zelle einzuführen. Nur dann steht das enzymatische Besteck zur Verfügung, das zur Expression der Gene benötigt wird. Zusätzlich wird es notwendig sein, aus der „Zellamme" ihr eigenes Genom zu entfernen. Auf der Ebene von eukaryontischen Zellen ist dies bereits möglich. Ohne großen experimentellen Aufwand kann beispielsweise einer Ei- oder Körperzelle der Zellkern (welcher in Form der Chromosomen das Genom enthält) entnommen und durch einen fremden ersetzt werden. Auf diese Art und Weise ist das Schaf Dolly „erzeugt" worden. Bei Bakterien jedoch ist das zirkulare Chromosom, welches häufig in vielfacher Kopienzahl vorliegt, nicht in einem Zellkern kompartimentiert. Methoden, um ein Bakterium zu „entgenomisieren", müssen also erst noch entwickelt werden.

Wie steht es um die wissenschaftliche Erkenntnis, die von dem Minimal Genom Projekt erwartet werden kann? Nach dem gegenwärtigen Stand wird sich die Frage, wie viele Gene für das Leben mindestens notwendig sind, nur bedingt beantworten lassen. Da das künstliche Erbgut nur in einer funktionsfähigen Zelle „zum Leben erweckt" werden kann, hat das Projekt zunächst eine vollständige Zelle als Grundlage. Man wird zwar beobachten können, ob die Hybride aus der „entgenomisierten" Bakterienhülle und dem „minimalen Genom" überlebensfähig ist und sich vermehren kann. Die Frage aber, wie ein Bakterium, oder allgemeiner: eine Zelle, entsteht, wird auf diese Weise umgangen. Diese Sichtweise hat auch Einfluss auf die ethische Beurteilung des Projektes, wie sie von dem Mediziner und Philosophen Arthur Caplan, Direktor des Zentrums für Bioethik der „University of Pennsylvania" und wissenschaftlicher Beirat von Celera Genomics, vorangetrieben wird [31]. Sicherlich nicht ganz unbefangen, sieht Caplans Ethik-Kommission in dem Vorhaben einen wichtigen Schritt für die Gentechnologie, da man Organismen aus der bloßen Kenntnis der Gensequenz erzeugen könnte. Dass dies sicherlich nicht ohne weiteres gilt, zeigt die Tatsache, dass wir in unseren Zoos nicht auf vorzeitliche Tiere treffen. Auch steht es zur Erwägung, ob man die Technologien entwickeln möchte, welche die „Totalsynthese" von *Bacillus anthracis* (dem Milzbranderreger) *de novo* zulassen.

Abbildung 8: Einfluss von Gendelitionen auf das Wachstum der Hefe *Saccharomyces cerevisiae*. Gezeigt ist der Anteil betroffener Gene bei Hefen, die langsamer als der Wildtyp wachsen. Die Gene sind nach Gruppen sortiert. Die blauen Balken stellen Genkategorien dar, die im Vergleich zu ihrem prozentualen Anteil im Genom bei langsam wachsenden Hefemutanten unterrepräsentiert sind. Im Gegensatz dazu bezeichnen die roten Balken Genkategorien, die bei diesen Mutanten häufiger vorkommen, als man es von ihrer Präsenz im Genom erwarten würde. Null Prozent bedeutet, dass genauso viele Gene der entsprechenden Kategorie das Wachstum der Hefe betreffen, wie man es aufgrund ihrer Repräsentanz im Genom erwarten würde. Nach Daten aus [28].



Literatur

[18] Kapranov P et al. (2002) Science 296: 916-919
[19] Shannon CE (1948) Bell System Tech. J. 27: 379-423, 623-656
[20] Stephens RM & Schneider TD (1992) J. Mol. Biol. 228: 1124-1136
[21] Schneider TD (2000) Nucl. Acid Res. 28: 2794-2799
[22] Loewenstern D & Yianilos PN (1999) J. Comput. Biol. 6: 125-142
[23] Itaya M (1995) FEBS Lett. 362: 257-260
[24] Fraser CM et al. (1995) Science 270: 445-446
[25] Mushegian AR & Koonin EV (1996) Proc. Natl. Acad. Sci. USA 93: 10268-10273
[26] Huynen M & Bork P (1998) Proc. Natl. Acad. Sci. USA 95: 5849-5856
[27] Hutchison III CA et al. (1999) Science 286: 2165-2169
[28] Giaever G et al. (2002) Nature 418: 387-391
[29] Gil R et al. (2002) Proc. Natl. Acad. Sci. USA 99: 4454-4458
[30] Cello J, Paul AV & Wimmer E (2002) Science 297: 1016-1018
[31] Cho MK et al. (1999) Science 286: 2087, 2089-2090
[32] http://ncbi.nlm.nih.org
[33] Benjamin Lewin (2000) Genes VII. Oxford University Press
[34] http://www.ndsu.nodak.edu
[35] Hakimi M-A et al. (2002) Nature 418: 994-998