Genomsequenzierung für Anfänger

Philipp Pagel

8. November 2005

1 DNA Sequenzierung

Heute wird DNA üblicherweise mit der sogenannten *Sanger* (oder *chain-termination*- oder Didesoxy-) Methode sequenziert dessen wesentliche Schritte im Folgenden schematisch dargestellt werden. (Biologisch versierte Leser mögen verzeihen, dass hier stark vereinfacht wird.)

Zunächst wird ein kleines (ca. 15 – 20 bp) Stück einzelstängiger DNA hergestellt, dessen Sequenz am Anfang (5' Ende) des kodierenden Strangs der zu sequenzierenden DNA vorkommt – der sogenannte *Primer*. Dieser ist somit komplementär zum korresponierenden Abschnitt des Komplementärstrangs (*template* Strang) und wird unter geeigneten Bedingungen an diesen binden (hybridisieren):

```
Primer:     5'        ACGGCTATGC 3'
Template:   3' ...AGGTGCCGATACGTACGGCTTGCTTAACTGACTGGTACCCATG... 5'
```

Nun gibt man eine sogenannte DNA-Polymerase, sowie ein Gemisch der vier Nukleotide (A, T, C und G) hinzu und lässt die Polymerase ihre Arbeit tun. Dabei hängt diese Schritt für Schritt weitere Nukleotide an den Primer an – und zwar nicht zufällig, sondern streng nach der Vorlage (dem Template). Dadurch entsteht mit der Zeit ein vollständiger DNA-Strang aus dem Primer:

```
Primer:     5'        ACGGCTATGCATGCCGAACGAATTGACTGACCATGGGTAC... 3'
Template:   3' ...AGGTGCCGATACGTACGGCTTGCTTAACTGACTGGTACCCATG... 5'
```

Wenn man nun zusätzlich zu den vier normalen Nukleotiden A, T, C und G eine geringe Menge A* hinzufügt, welches chemisch so modifiziert ist, dass es *nicht verlängert* werden kann (*Di*desoxy-ATP), so kommt es an den Stellen, an denen ein solches A* zufällig eingebaut wird, zu einem Kettenabbruch:

```
Primer:     5'        ACGGCTATGCATGCCGAACGAA* 3'
Template:   3' ...AGGTGCCGATACGTACGGCTTGCTTAACTGACTGGTACCCATG... 5'
```

oder

Primer: 5’ ACGGCTATGCATGCCGAACGAATTGACTGACCA* 3’
Template: 3’ ...AGGTGCCGATACGTACGGCTTGCTTAACTGACTGGTACCCATG... 5’

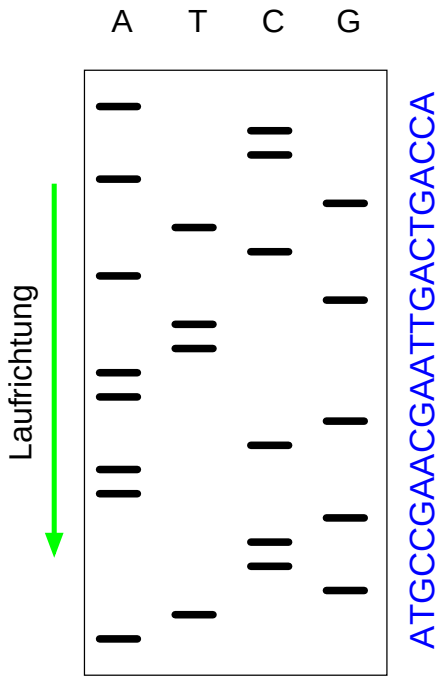
oder

Primer: 5’ ACGGCTATGCA* 3’
Template: 3’ ...AGGTGCCGATACGTACGGCTTGCTTAACTGACTGGTACCCATG... 5’

oder ...

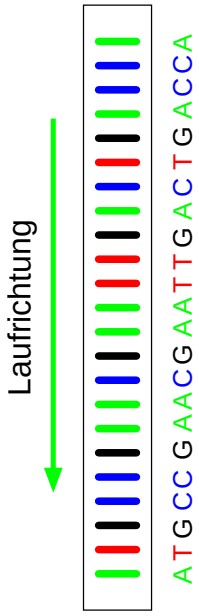
In unserem Reagenzglas befindet sich nun also ein Gemisch von neugebildeten DNA Molekülen, von denen wir wissen, dass sie mit einem A enden. Das Gleiche wiederholen wir in drei weiteren Reagenzgläsern analog mit den übrigen Nukleotiden. So erhalten wir vier separate Reaktionen, die jeweils Kettenabbrüche an je einer der vier Basen enthalten.

Diese vier DNA Lösungen werden mittels Gelelektrophorese aufgetrennt. Dabei wandern kurze Moleküle schneller durch das Agarosegel, als längere. Diese Information kann man verwenden, um die Sequenz zu bestimmen. Für unser Beispiel ergäbe sich etwa folgendes Bild:

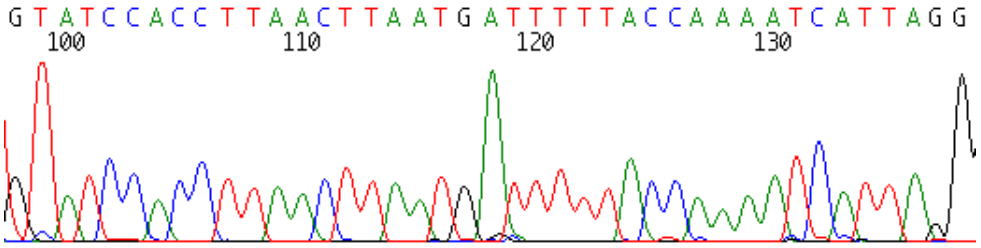


Das kürzeste Molekül auf dem Gel (ganz unten) findet sich in der ersten Spur (Spalte) – also ist die erste Base ein A. Das nächstlängere ist in der zweiten Spur: ein T. Es folgen ein G, ein C usw. Wenn wir die Banden nun schrittweise nach oben verfolgen, erhalten wir die gesuchte Basensequenz.

Man kann das Ganze noch übersichtlicher gestalten, indem man die DNA Moleküle in jeder der vier Reaktionen mit einem anderen Fluoreszenzfarbstoff markiert, zusammenmischt und dann gemeinsam in einer Spur laufen lässt:



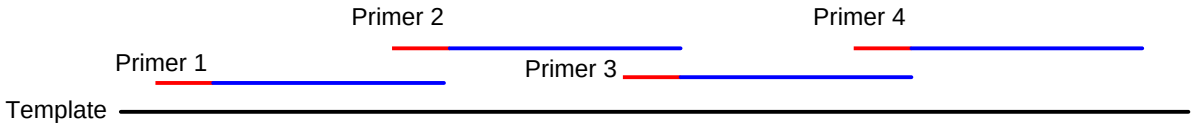
Heute verwendet man meist Sequenziergeräte, die die obigen Schritte automatisch ablaufen lassen. Die Auftrennung erfolgt dabei in einer Kapillare und ein optischer Sensor registriert die verschiedenfarbigen Banden. Am Ende erhält man Intensitätskurven für die vier Kanäle, sowie eine daraus vom Computer generierte Basensequenz. Hier ein Auschnitt:



Quelle: http://mit.edu/7.02/resources/Blast-tutorial/callseq-right.shtml

Mit zunehmender Länge, wird das Signal schwächer und somit unzuverlässiger. Daher kann man mit dieser Methode nur ca. 500 – 1000 Basen in einem *Read* (oder *Run*) sequenzieren.

Um längere DNA-Moleküle zu sequenzieren (die codierende Sequenz für ein durchschnittliches Protein kann leicht mehrere tausend Basen lang sein) benötigt man also mehrere Reads mit unterschiedlichen Primern:

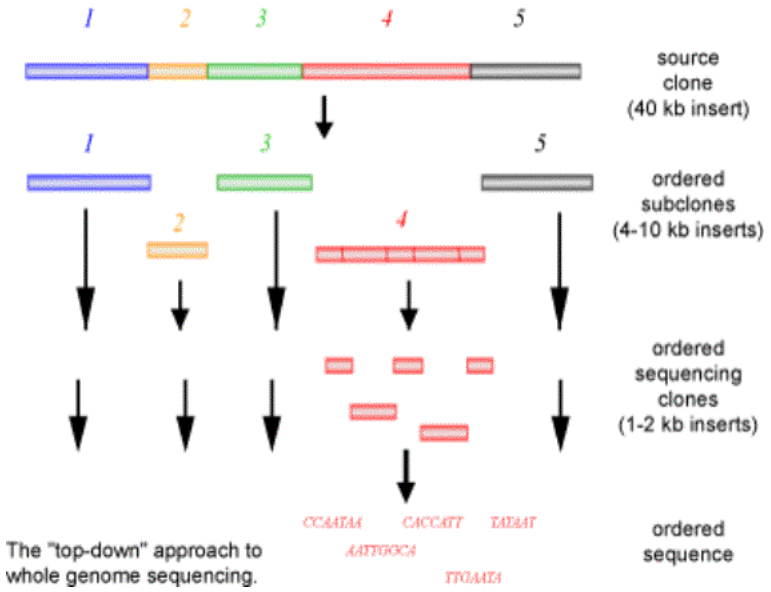


Dies ist kein Problem, wenn man die passenden Sequenzen für die einzelnen Primer bereits kennt. Sequenziert man jedoch eine unbekannte DNA, so muß man das Ergebnis

des ersten Runs abwarten, um eine Primersequenz für den nächsten Run zu bestimmen. Dieser Vorgang wird als *primer walking* bezeichnet und ist durchaus praktikabel, wenn insgesamt nur wenige Reads nötig sind um die gewünschte Sequenz zu bekommen. Im Falle ganzer Genome ist dies aber viel zu umständlich und langwierig.

2 Kleinere Häppchen – BAC sequencing

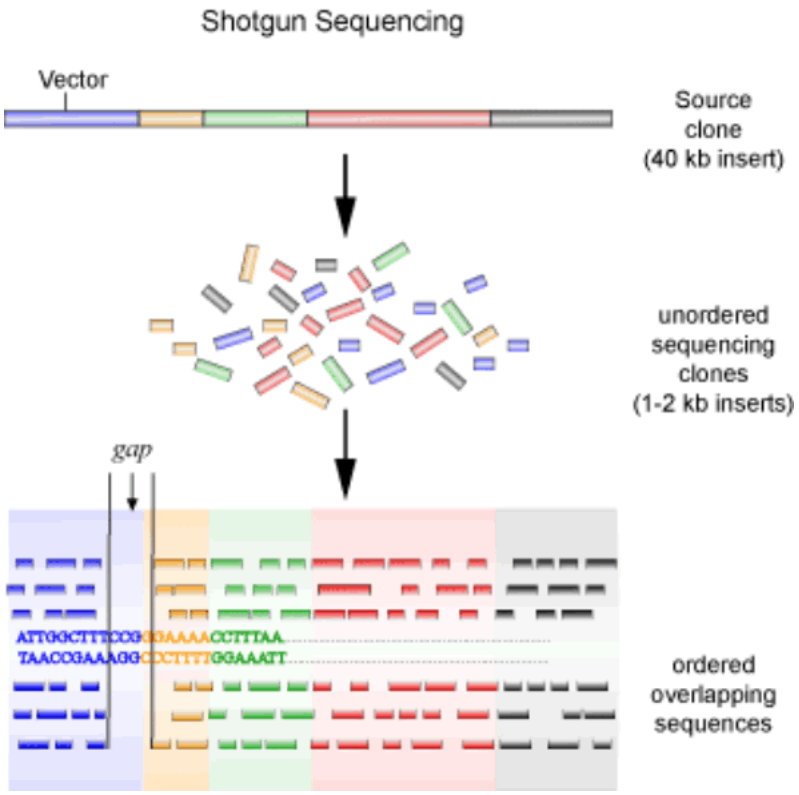
Um ein ganzes Genom zu sequenzieren bietet es sich an, die DNA in eine Vielzahl deutlich kürzerer Abschnitte („Klone“, ≈ 20 000 – 200 000 bp) zu zerlegen, die man dann parallel sequenziert und am Ende zusammensetzt (*assembliert*). Hierzu werden die einzelnen Abschnitte z.B. in sogenannte *bacterial artificial chromosomes* (BACs) kloniert. Diese zerlegt man weiter in Subklone, die dann mittels Primer-Walking sequenziert werden. Die Reihenfolge der Klone und Subklone wird dabei anhand bekannter Markersequenzen (sequence tags), deren relative Lage zueinander man aus vorher durchgeführten Mapping-Experimenten kennt (z.B. Radiation hybrid maps) bestimmt. Dabei bezeichnet man die aus den einzelnen Reads zusammengesetzten Sequenzen als *Contigs*, die ihrerseits zu *Supercontigs* (oder *Scaffolds*) zusammengefügt werden.



Quelle: http://www.bioteach.ubc.ca/Bioinformatics/GenomeProjects

3 Whole genome shotgun sequencing

Das whole genome shotgun Verfahren hat die Genomsequenzierung revolutioniert. Es verzichtet auf vorheriges Wissen über die Reihenfolge von Klonen. Stattdessen wird die gesamte Genomsequenz in einer großen Zahl zufällig plazierter Reads sequenziert („shotgun“). Im Anschluß werden die einzelnen Reads anhand ihrer Überlappungen von einem Computerprogramm zusammengesetzt.



Quelle: http://www.bioteach.ubc.ca/Bioinformatics/GenomeProjects

Da die Reads zufällig über das Genom verteilt sind, reicht es nicht aus, genau die Anzahl Basen zu sequenzieren, die das Genom enthält. Dabei würden einige Bereiche mehrfach sequenziert werden, andere hingegen garnicht. Aus diesem Grund ist es erforderlich, eine höhere mittlere Abdeckung (*coverage*) zu erreichen. Doch wie hoch muss diese sein?

3.1 Wieviele Reads?

Wie viele Reads benötigt werden hängt in erster Linie von den vorgegebenen Qualitätsansprüchen ab. D.h. Wie groß darf die Wahrscheinlichkeit sein, dass eine Base garnicht, nur einmal, nur zweimal etc. sequenziert wurde?

Beispiel: Ein Genom von 5 Mb ($5 \cdot 10^6$ Basen) soll mit der shotgun Methode sequenziert werden. Ein Read habe eine Länge von 500 Basen. Wieviele Reads werden benötigt, damit 99 % der Basen mindestens zweimal sequenziert wurden?

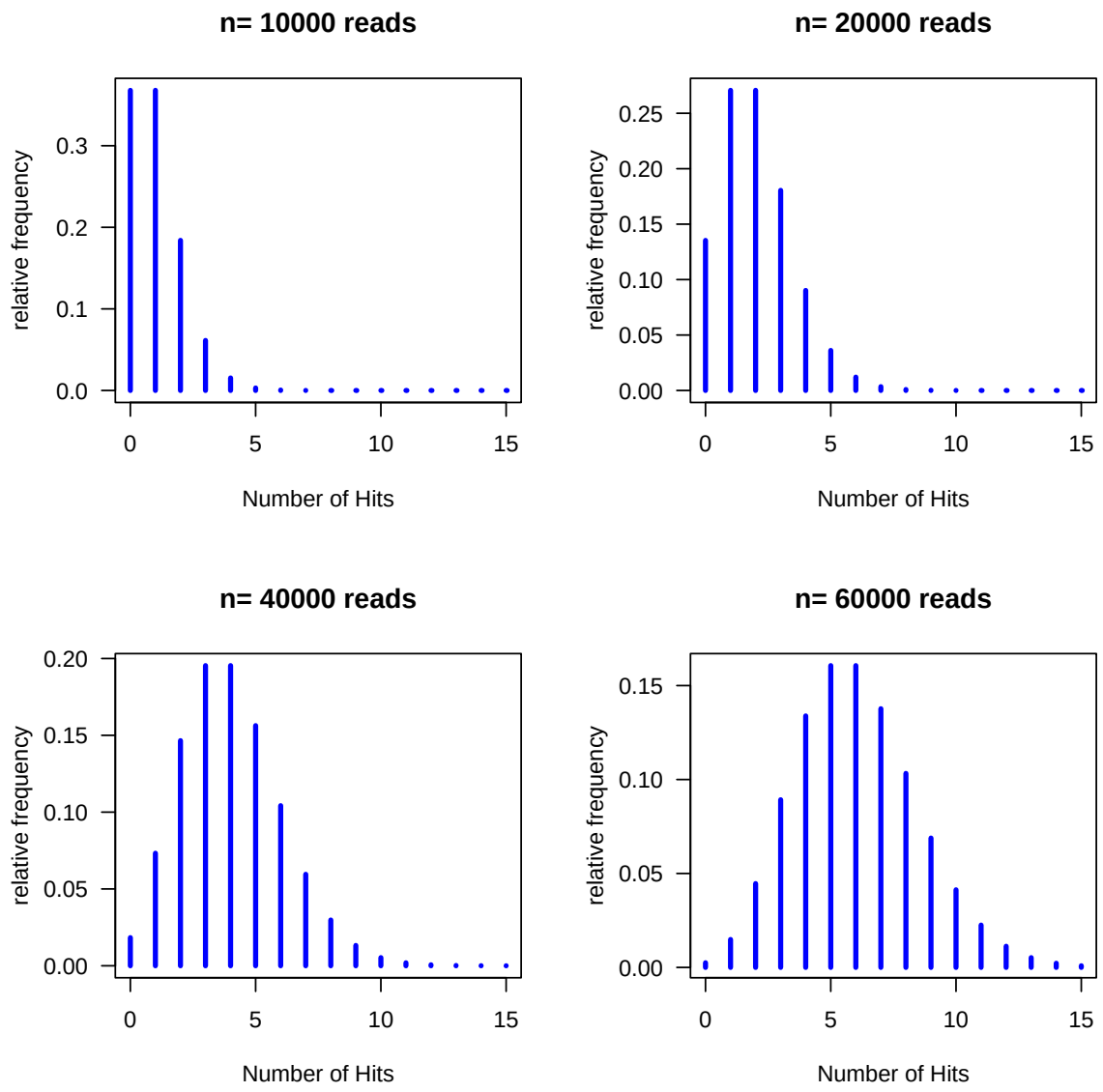
Die Wahrscheinlichkeit p, dass eine bestimmte Base von einem zufälligen Read abgedeckt wird ist angenähert

$$p \approx \frac{500}{5 \cdot 10^6} = 1 \cdot 10^{-4}$$

D.h. einfache Abdeckung entspricht $5 \cdot 10^6/500 = 10\,000$ Reads. Die Wahrscheinlichkeit, dass eine Base bei n Reads genau k mal sequenziert wurde (k Hits) folgt einer Binomialverteilung:

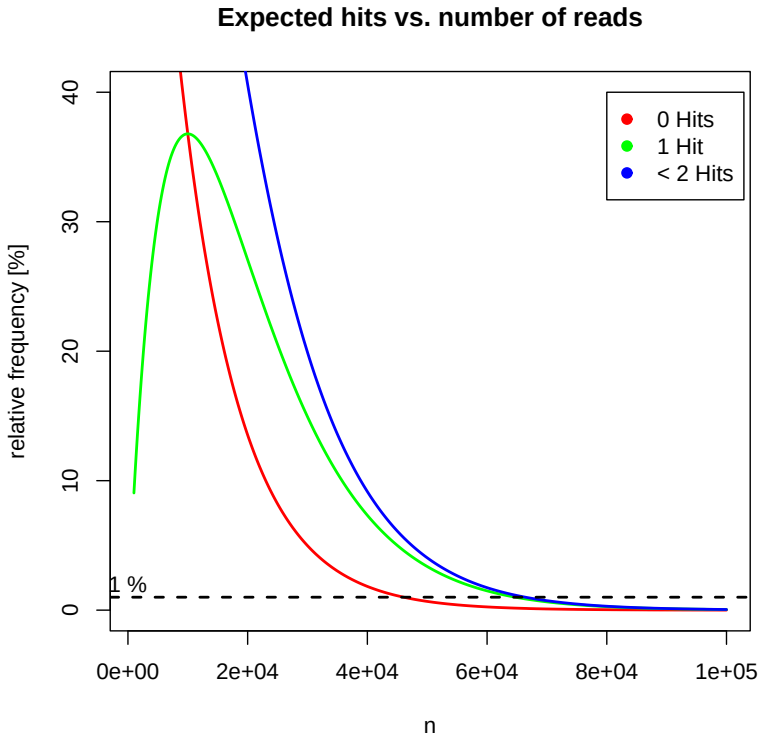
$$P(k|p,n) = \binom{n}{k} p^k (1-p)^{n-k}$$

Betrachten wir die Wahrscheinlichkeitsverteilung für eine zunehmende Zahl von Reads:



Wie man sieht, werden bei 10 000 Reads ca. 37 % der Basen von keinem Read abgedeckt. Bei 20 000 Reads sind es nur noch ca. 13.5 %, bei 40 000 ca. 1.8 %. Bei 60 000 Reads werden nur noch 0.25 % nicht sequenziert und ca. 1.5 % werden nur von einem Read abgedeckt.

Unser Ziel war es, dass 99 % der Basen mindestens zweimal sequenziert wurden – also maximal 1 % der Basen Null oder einen Hit hatten. Der folgende Graph zeigt die Anzahl der zu erwartenden Basen mit keinem, einem oder weniger als zwei Hits in Anhängigkeit von der Anzahl der Reads:



Die blaue Kurve erreicht die gewünschte 1 % Marke bei ca. 65 000 Reads (64744 um genau zu sein); nur noch 0.15 % werden nicht abgedeckt. D.h. wir benötigen mindestens diese Anzahl von Reads um unsere Qualitätsansprüche zu erfüllen. Dies entspricht einer mittleren Abdeckung von 6.5 fach.

Bei der angenommenen Genomgröße von 5 Mb sind 0.15 % aber immerhin 7500 unsequenzierte Basen – d.h. maximal auch ebensoviele Lücken (*Gaps*). Zwar kann man die Reihenfolge der entstandenen Contigs, wie schon bei der herkömmlichen Methode beschrieben, meist anhand von Markern festlegen, ein Schließen der Gaps bleibt dennoch wünschenswert. Anstatt nun die Anzahl der Shotgun Reads zu erhöhen ist es aber ökonomischer, die Gaps durch gezielte Sequenzierung zu eliminieren.

In der Praxis trifft man zusätzlich auf Regionen, die aus verschiedenen Gründen schwer zu sequenzieren sind und bei keiner noch so großen Anzahl von Shotgun Reads abgedeckt werden. Solche Problemfälle lassen sich (wenn überhaupt) nur durch gezielte Sequenzierung (mit bestimmten Tricks) lösen.