

Technologierevolution in der Genomforschung

Von Thomas Hankeln, Hans Zischler und Erwin R. Schmidt

Neue revolutionäre Sequenziertechnologien versprechen für Medizin und Biologie einzigartige Einsichten in das Erbmateriale von Lebewesen. In wenigen Jahren werden wir alle die DNA-Sequenz unseres eigenen individuellen Genoms bei moderaten Kosten entschlüsseln können. Die riesigen Datenmengen schaffen aber auch Probleme für die bioinformatische Verarbeitung und Interpretation. Das Nukleinsäureanalytik-Kompetenzzentrum der Uni Mainz plant im Verbund mit dem Forschungsschwerpunkt „Rechner-gestützte Forschungsmethoden in den Naturwissenschaften“ die zentrale Etablierung der neuen Verfahren.

Manche Revolutionen dauern etwas länger. Immerhin 32 Jahre ist es her, dass Fred Sanger und Alan Coulson sowie unabhängig ein zweites Team von Walter Gilbert und Alan Maxam praktikable Verfahren zur Sequenzierung von DNA vorstellten. Nur 5386 DNA-Bausteine (Nukleotide bzw. Basenpaare des DNA-Doppelstrangs) umfasste das erste vollständig sequenzierte Genom des Bakteriophagen PhiX174. Bis heute hat die Sanger-Methodik, eigentlich eine nachgeahmte DNA-Replikation im Reagenzgefäß, überlebt: mit ihr wurden von 1998 bis 2003 mit einem Milliarden-Dollar-Einsatz die ersten Versionen des 3 Milliarden Nukleotide umfassenden Humangenoms erstellt. Die Nachteile des Sanger-Ansatzes, die der kostengünstigen Generierung von noch größeren Sequenzdatenmengen entgegenstehen, wurden bald offensichtlich. Es sind dies die für das Verfahren notwendige Auftrennung der Nukleinsäurefragmente im elektrischen Feld, die geringe Parallelisierbarkeit und die aufwändige Probenvorbereitung. Zwar wird die Sanger-Methode nicht aussterben und auch weiterhin zur Entschlüsselung kleiner DNA-Moleküle verwendet werden. Doch völlig neue, sehr viel schnellere und im Großmaßstab extrem kostengünstige Sequenziermethoden („Next-Generation-Sequencing“, NGS) gelangten vor etwa 3 Jahren zur Anwenderreife und revolutionieren derzeit die Genomforschung.

NGS: Entdecke die Möglichkeiten!

Mit dem beabsichtigten Publicity-Effekt konnte mit NGS kürzlich das Genom des DNA-Helix-Entdeckers und Nobelpreisträgers James D. Watson innerhalb weniger Wochen für etwa 350.000 Dollar entschlüsselt werden. Die ebay-Auktion einer Humangenom-Sequenzierung im Mai 2009 zum Einstiegspreis von 68000 USD war hingegen wohl (noch) ein reiner Werbegag der „Personal Genomics“-Firma KNOXE, zeigt aber klar die Richtung an. Ziel der Humangenetik ist es, die vielen kleinen und großräumigen Unterschiede unseres Erbmateriale zu identifizieren, also die durch Mutation ausgetauschten oder gar zusätzlich eingefügten oder deletierten Nukleotidpositionen. Es sind nämlich diese Unterschiede mit einer Frequenz von etwa 1 auf 1000 Bausteinen (gilt für einzelne Nukleotid-Austausche), die uns als Lebewesen individuell machen. Seit Januar 2008 läuft das „1000 genomes project“. Dessen Ziel ist es, durch Sequenzierung der Genome von Asiaten, Afrikanern und Europäern alle weltweit biomedizinisch relevanten Genunterschiede aufzuspüren. Dies wird die Grundlage für eine individualisierte Medizin darstellen, bei der Medikamente auf den Genotyp abgestimmt verabreicht werden, um ihre Wirksamkeit zu verbessern und Nebenwirkungen auszuschließen. Auch wird der Katalog der Genvarianten durch Assoziationsstudien die Identifizierung solcher Gene erheblich beschleunigen, die für komplexe genetische Erkrankungen (Diabetes, Bluthochdruck, Krebs etc.) verantwortlich sind. So konnte kürzlich z. B. ein komplettes Genom von Blutkrebszellen sequenziert werden:

der Vergleich mit dem Genom aus gesunden Zellen des Patienten zeigte 8 tumorspezifische Mutationen in Genen, die zuvor nie als krebsrelevant aufgefallen waren.

Doch nicht nur die Medizin, sondern alle Lebenswissenschaften sind elektrisiert ob der neuen Möglichkeiten: Molekulare Systematiker sequenzieren große Genmengen, um aufgrund von Sequenzen Stammbäume von Tieren, Pflanzen und Bakterien zu erstellen. Dabei entdecken sie, dass klassische Gruppierungen von Tieren (z. B. anhand des Merkmals „Körpersegmentation“) völlig neu sortiert und Lehrbücher umgeschrieben werden müssen, weil die genetische Verwandtschaft suggeriert, dass solche zunächst komplex erscheinenden morphologischen Eigenschaften während der Evolution mehrfach konvergent entstehen oder verloren gehen können. Sogar neue weitläufige Verwandte der Wirbeltiere, zuvor fälschlicherweise den Würmern zugeordnet, werden durch Genomanalysen entdeckt. Dies hat natürlich erheblichen Einfluss darauf, wie wir die Evolution unseres eigenen Genoms betrachten. Schon werfen Evolutionsbiologen gar einen direkten Blick in die Vergangenheit unserer Spezies: So haben Forscher aus uralten DNA-Stücken aus Knochen eine erste Version des Genoms des Neanderthalers per NGS rekonstruiert. Da die endogene Neandertaler-DNA stark fragmentiert vorliegt und nicht völlig problemlos von kontaminierender moderner DNA zu unterscheiden ist, läuft die Auswertung noch. Wird es sich bestätigen, dass sich der Neanderthaler vor etwa 600000 Jahren von unserer eigenen Evolutionslinie abgetrennt hat und bis zu seinem Aussterben vor 25000 Jahren quasi als Parallelgesellschaft und ohne Genaustausch mit Homo sapiens existiert hat?

Auch Ökologen haben Spannendes vor: NGS ermöglicht es, bislang eher exotische und auf Genomebene kaum bekannte Spezies aus wichtigen Ökosystemen zu analysieren. So wurde soeben ein Genkatalog von Riff-bildenden Korallen mit NGS erstellt, um daraus Gene zu extrahieren, die für eine Anpassung von Korallen an sich ändernde Klimabedingungen wichtig sind. Auch in der Mikrobiologie ist NGS ein Durchbruch: die relativ kleinen Genome von Bakterien (nur ca. 1/1000 des Humangenoms) werden zu Hunderten sequenziert und verglichen. Es zeigt sich eine ungeahnte Variabilität: so hat ein Stamm des Modell-Darmbakteriums Escherichia coli, der seit 20 Jahren im Labor der Gentechnologen ohne irgendeine Sicherheitsproblematik verwendet wurde, im Vergleich zu verwandten, aber pathogenen E. coli-Stämmen in der Tat mehr als 1000 Gene verloren, was ganz klar seine Gutmütigkeit erklärt. Andere Genomvergleiche zeigen, dass Bakterien durch den „Horizontalen Gentransfer“ aber auch natürlicherweise DNA aus der Umwelt aufnehmen, in ihrem Genom etablieren und daraus neue Eigenschaften entwickeln können.

Weitere Anwendungen findet die NGS z.B. in der qualitativen und quantitativen Analyse von Genprodukten wie RNA-Molekülen (Transkriptom). Durch „deep sequencing“ ist es sogar möglich, nur in sehr geringer Kopienzahl (statistisch < 1 Molekül pro Zelle) vorliegende Transkripte, sicher nachzuweisen. Legt man zugrunde, dass unser typisches Säugetier-Genom etwa 25 000 Gene beinhaltet, so entfällt nur ein geringer Prozentsatz unserer Erbsubstanz (ca. 1.5 %) auf die Kodierung dieser Information. Mit der NGS-Transkriptomanalyse können gezielt diejenigen Bereiche des Genoms erfasst werden, die funktionell wichtig sind und somit Aussagen über die Genaktivität und deren Regulation in verschiedenen Entwicklungsstadien und Geweben eines Organismus gewonnen werden. Da sich die Anzahl und Sequenz der Gene in nah verwandten Tiergruppen nur marginal unterscheidet, muss davon ausgegangen werden, dass die phänotypische Unterschiedlichkeit von Spezies ganz maßgeblich auf dem Muster der Genaktivität und deren Regulation in einem räumlich-zeitlichen Kontext beruht. Zur Regulation von Genen müssen die entsprechenden DNA-Abschnitte mit Proteinfaktoren interagieren. Diese Proteine können ‚quasi bei der Arbeit‘ im Komplex mit ihren Nukleinsäure-Interaktionspartnern chemisch verbunden und danach mit

spezifischen Antikörpern isoliert werden (Chromatin-Immunpräzipitation). Die gebundenen Sequenzen werden dann im NGS-Verfahren qualitativ und quantitativ erfasst und mit Genomsequenzen abgeglichen, sodass der genaue Ort einer Proteininteraktion im Genom festgelegt werden kann. Datenbanken mit Mustern der DNA-Proteininteraktion in ganzen Genomen befinden sich derzeit im Aufbau.

Bye bye Sanger: die neue Generation von Sequenzierverfahren

Das Wesen von NGS besteht in dem Verzicht auf langsame und kostspielige Abläufe der Sanger-Methode. Abgeschafft wurden so die klassische Klonierung von DNA, durch die man genug homogenes Erbmateriale für die Sequenzierung erhielt, und der unpraktische und langsame Sequenz-Leseschritt, nämlich die Auftrennung der Sequenzierprodukte durch Gelelektrophorese. Die Durchführung der Sequenzierreaktion im Mikro- oder zukünftig gar Nanomaßstab senkt zudem den Reagenzienverbrauch drastisch. Drei NGS-Verfahren konkurrieren derzeit auf dem Markt, basieren aber auf unterschiedlichen Prinzipien und werden z. T. unterschiedliche Anwendungsgebiete in der Genomforschung haben:

- Das Verfahren von 454 Life Sciences/ Roche basiert auf der bereits Anfang der 1990er Jahre von Ronaghi und Kollegen erdachten Pyrosequenzierung (Abb. 1A). Zunächst wird die zu sequenzierende DNA (z. B. ein komplettes Genom oder Kopien der Gentranskripte = cDNA) physikalisch zerlegt. Die Bruchstücke werden einzeln an 20µm-Beads gekoppelt und daran heftend durch eine Polymerasekettenreaktion (PCR) klonal vermehrt (d. h. eine Kugel trägt viele identische Kopien eines bestimmten DNA-Moleküls). Die Beads werden nun einzeln in Löcher einer sog. Picotiter-Platte gefüllt. Bei 1 Mio. Wells pro Platte können eben so viele verschiedene Moleküle gleichzeitig sequenziert werden. Bei der eigentlichen Sequenzreaktion wird wie bei Sanger ein Einzelstrang des zu sequenzierenden DNA-Moleküls als Vorlage genommen und mit Hilfe eines Primers und einer Polymerase zum Doppelstrang ergänzt („sequencing-by-synthesis“). Immer wenn ein Nukleotid richtig (d. h. komplementär zum Vorlagenstrang) eingebaut wird, kann das dabei freigesetzte Pyrophosphat (PPi) mit Hilfe eines in den Wells befindlichen Enzymsystems zunächst in den Energielieferanten ATP und dann in einen zu messenden Lichtblitz umgewandelt werden (Abb.1A, B). Nacheinander werden die vier DNA-Bausteine A, G, C und T hinzu gegeben (Sequenzierzyklus 1). Zwischen den einzelnen Nukleotidzugaben erfolgen Waschschrte, die das System „zurücksetzen“. Wenn der Matrizenstrang z. B. ein A enthält, leuchtet das Well ausschließlich bei der Zugabe von T auf. Danach erfolgt der nächste Sequenzierzyklus, wieder mit den vier Zugabeschritten. Wenn hier z. B. der Lichtblitz bei G-Zugabe entsteht und dreimal so stark ist wie der Blitz in Zyklus 1, so lautet die Sequenz bis hierhin „TGGG“. Die Darstellungsform solcher Sequenzdaten („Flowgram“) ist in Abb. 1C gezeigt. Durchschnittlich etwa 400 Nukleotide (geplant: bis 1000) können so derzeit pro Well gelesen werden. Diese Leselänge kommt dem Sangerverfahren nahe und ermöglicht eine weniger problematische Assemblierung von Teilsequenzen eines Genoms. Die 454-Technologie gilt daher insbesondere als Methode der Wahl für die sogenannte *de novo*-Sequenzierung bisher völlig unbekannter DNA. Aber auch die Re-Sequenzierung bekannter Genome (Mensch, Bakterienstämme etc.) und die Sequenzierung von ganzen Transkriptomen lassen sich mit der sehr vielseitigen 454-Technik durchführen. Die generierten Datenmengen sind mit bis zu 500 Mega-Basenpaaren (MBp) Sequenzinformation pro 10 h-Gerätelauf bereits erheblich (zum Vergleich: ein klassischer Sanger-Kapillarsequencer schafft etwa 1 MBp pro Tag), liegen jedoch unterhalb der beiden anderen Systeme, die daher für reine quantitative Applikationen (Auszählen von DNA-Schnipseln z.B. zur Messung der Genexpression) besser geeignet sind. Schwächen hat die 454-Technologie prinzipbedingt beim Lesen langer Homopolymer-Abschnitte, die jedoch in den wichtigen kodierenden Genbereichen nicht so häufig sind: es ist

schwierig, die Lichtintensität nach Einbau z. B. von 20 C-Nukleotiden gegenüber nur 19 Bausteinen zu diskriminieren. Dennoch liegt insbesondere bei ausreichend hoher Redundanz (d.h. mehrfachem Sequenzieren derselben DNA-Region) die Lesegenauigkeit bei etwa 99 %.

- Das 2006 eingeführte Solexa/Illumina -Verfahren auf dem „Genome Analyzer II“ produziert mit max. 120 Mio. Sequenzierungen pro Lauf deutlich höhere Datenmengen (20 GB), erzeugt aber dabei geringere Leseweiten von maximal 75-100 Nukleotiden (geplant: 150). Solche „short reads“ sind bei der Assemblierung zu längeren Sequenzabschnitten bekannt problematisch. Daher wird das Verfahren zumeist angewandt, wenn bereits eine Referenzsequenz (z. B. das Humangenom) als Vorlage zum Zuordnen der Sequenzschnipsel zur Verfügung steht. Die Solexa-Methode vermehrt die zu sequenzierenden DNA-Moleküle nicht an Kügelchen, sondern führt eine PCR-Vervielfältigung auf einem soliden Träger durch („bridge PCR“). Die Sequenzierung verfeinert den Nobelpreis-belohnten Sanger-Trick: an einem einzelsträngigen DNA-Molekül unbekannter Sequenz als Vorlage wird der komplementäre Strang repliziert und dabei ein Terminator-Nukleotid eingebaut (Abb. 2). Durch den Terminator wird die DNA-Synthese gestoppt, und die Fluoreszenzmarkierung (vier unterschiedliche Farben für die vier Nukleotide) wird aufgenommen. Danach werden der Terminator und die Fluoreszenz am Ende des entstehenden Strangs entfernt, die Synthese und damit die nächste Sequenzierrunde können beginnen.

- Im Gegensatz zu den beiden „sequencing-by-synthesis“-Verfahren setzt das Ende 2007 eingeführte SOLiD-System (ABI) auf „sequencing-by-ligation“ (Abb. 3). Wie bei 454 werden die zu sequenzierenden Moleküle zunächst einzeln an 1µm-Beads gekoppelt und durch eine Amplifikation in einer Wasser-in-Öl-Emulsion (emPCR) klonal vermehrt. Die Beads mit den Sequenziervorlagen werden auf einen soliden Träger aufgebracht (bis zu 300 Mio.). Zu dem zu sequenzierenden Vorlagen-Strang wird dann ein Gemisch farbmarkierte Oktamer-Nukleotide mit unterschiedlichen Sequenzen zugegeben. Diese Oligonukleotide „tasten“ den unbekanntes DNA-Strang quasi ab: passt ein solches Oligonukleotid exakt zur Vorlage, so wird es durch Ligation fest gebunden. Seine Fluoreszenzfarbe zeigt sodann die Sequenz des gebundenen Moleküls an. Im nächsten Sequenzierzyklus wird die Fluoreszenz entfernt und ein neues Oligonukleotid gebunden, das wiederum Basen entziffert. Die Leseweite dieser Technik ist derzeit auf etwa 35-50 Bausteine beschränkt (Ziel für 2010: 75 Bp), jedoch erlaubt ein kompliziertes Farbkodierungsverfahren bei der Sequenzierung eine extrem gute Lesegenauigkeit. Durch die hohe Zahl parallel sequenzierter Schnipsel liefert SOLiD insgesamt die höchste Datenmenge pro Lauf (30 GB). Aufgrund der Kürze der „reads“ ist das System letztlich am besten für Re-Sequenzierungen mit einer Referenzsequenz als Vorlage geeignet. Die extrem hohe Anzahl parallel sequenzierter Moleküle ermöglicht es zudem, über einen breiten Bereich quantitative Aussagen zum Vorhandensein bestimmter Sequenzen zu treffen. SOLiD gilt daher als Verfahren, mit dem man in Zukunft die differenzielle Expression von Genen z. B. unter pathologischen Bedingungen extrem kostengünstig und hoch auflösend messen kann und so die gegenwärtig verwendeten, oftmals schwer standardisierbaren Mikroarray-Technologien ersetzt.

Natürlich macht diese Revolution technologisch nicht halt. Auch NGS wird konstant weiterentwickelt: schon liest man über die Next-Next (3rd) Generation-Technologie, die wie die derzeit praktikablen Verfahren jedoch vermutlich noch 3-5 Jahre bis zur Marktreife braucht. Firmen wie Pacific Biosciences, Visigen, Complete Genomics oder Oxford Nanopore Technologies sind bereits zum Teil von den etablierten NGS-Unternehmen als Zukunftsinvestition aquiriert worden. Die geplanten Methoden nehmen dabei Abschied von der klonalen DNA-Amplifikation per PCR, die gegenwärtig noch unverzichtbar für die Herstellung ausreichender Mengen an Sequenziermatrizen ist. Vorteil ist dabei nicht nur eine

weitere Kosten- und Arbeitsreduktion, sondern vor allem die Einzelmolekül-Sequenzierung, bei der man Hunderttausende von Nukleotiden quasi „in einem Rutsch“ liest. Diese Abkehr von der Sequenzierung „gestückelter“ DNA ermöglicht die für den Genetiker wichtige zweifelsfreie Identifizierung der auf ein und demselben DNA-Strang gekoppelt liegenden Sequenzunterschiede („Haplotypen“). Die Probleme seitens der bioinformatischen Aufarbeitung von Sequenzdaten nehmen mit der Sequenzlänge ab, zudem entfallen die „Kinderkrankheiten“ der ersten NGS-Verfahren wie z.B. eine zum Ende einer Sequenz abnehmende Qualität der Sequenzierergebnisse, die Fehler beim Lesen einer längeren Abfolge identischer Basen und die durch PCR-Amplifikation generierten Fehlerartefakte.

Wie funktioniert „NNGS“? Beim Verfahren von Pacific Biosciences wird in optischen Kammern aus der Halbleitertechnik („zero-mode wave guide“) von etwa 100 nm Durchmesser und einem Reaktionsvolumen von nur 20 Zeptolitern (10^{-21} l) ein einzelner DNA-Matrizenstrang durch ein einziges DNA-Polymeraseenzym repliziert. Die 4 Nukleotide für die DNA-Synthese sind jeweils an ihrem endständigen (gamma)-Phosphat unterschiedlich fluoreszenzmarkiert. Bei der Replikationsreaktion ist nun die Verweildauer eines „richtigen“, zum Vorlagenstrang komplementären Bausteins in der ZMW-Kammer um Größenordnungen länger (einige 10 msec), als die Anwesenheit der in den ZMW ebenso hinein diffundierenden, aber gerade „nicht passenden“ Nukleotide. Die Fluoreszenz des eingebauten Nukleotids wird nach Laseranregung in Echtzeit erfasst, das Fluorophor danach in dem normalen Abspaltungsprozeß von PPi freigesetzt und entfernt. Das nächste passende Nukleotid kann dann ohne störenden Hintergrund durch den ersten Sequenzierschritt detektiert werden. Die Leselänge soll dabei bereits etliche Tausend Basenpaare betragen (Geschwindigkeit: 10 Nukleotide/sec), wobei wohl der Laser durch fortschreitende Zerstörung der Polymerase noch ein begrenzender Faktor ist. Durch „Stroboskop-Sequenzieren“ (zeitweiliges Ausschalten des Lasers für einige sec) können offenbar entlang eines DNA-Moleküls immer wieder intervallmäßig Abschnitte gelesen werden, was die Rekonstruktion der Gesamtsequenz von Chromosomen z. B. in den problematischen repetitiven Bereichen des Genoms (s. u.) erheblich erleichtert.

Das zweite aussichtsreiche NNGS-Verfahren, als „nanopore sequencing“ bezeichnet, kommt ganz ohne DNA-Vervielfältigung oder Fluorophore aus. Es macht sich die Eigenschaft von Hämolyysin zu Nutze, das sich als bakterielles Exotoxin mit seiner Beta-Faltblattstruktur aus 7 Einheiten in biologische Membranen einlagert und dort stabile Poren bildet. Bereits 1996 fanden Deamer und Branton, dass sich durch die 2.6 nm weiten Poren nach Anlegen eines elektrischen Feldes einzelne Nukleinsäuremoleküle schleusen lassen und sich dies durch Änderungen des Stromflusses verfolgen lässt. Mittlerweile konnten die Wissenschaftler das Hämolyysin so gentechnisch modifizieren, dass es innerhalb der Pore die 4 Nukleotide der DNA durch einen charakteristischen Stromfluss unterscheiden kann. Auch das bei Säugern verbreitete und für die Diagnose von Krebsgenen medizinisch sehr wichtige 5-Methylcytosin („fünfter Baustein der DNA“) kann einzig mit dieser Technik direkt detektiert werden. Allerdings sind gegenwärtig die Nukleinsäuren noch immobilisiert, wobei man beim Sequenzieren natürlich den DNA-Strang beim Wandern durch die Nanopore entziffern möchte.

Bioinformatische Herausforderungen

So elegant und einfallsreich die neuen Hochdurchsatz-Sequenzierverfahren auch sind, so groß ist die Herausforderung für die Bioinformatik, diese Datenmengen zu speichern, sie auszuwerten und sinnvolles Wissen über das Genom und seinen Geninhalt daraus zu generieren. In der Zeitschrift *Nature Biotechnology* wurde dieses Unterfangen kürzlich mit

dem Versuch verglichen, seinen Wasserdurst aus einem Feuerwehrschauch heraus zu stillen. Aufgrund der seriellen Fluoreszenzmessung in jedem Sequenzierschritt produzieren alle NGS-Techniken Bild-Rohdaten in bisher so nicht gekanntem Umfang. Das SOLiD-System erfordert z. B. 15 Terabyte an Speicher für die reine Arbeitsumgebung sowie 30-40 TB Festplattenspeicher plus Bandspeicherung für mittel- und langfristige Datenarchivierung. Dabei ist eine Langzeitlagerung dieser wertvollen Rohdaten ratsam, da die gegenwärtigen NGS-Verfahren (wie auch zuvor Sanger) durchaus reaktionsbedingte systematische Fehler machen. Charakteristisch für die 454-Pyrosequenzierung sind z. B. falsche Abschätzungen der Nukleotidzahl in Homopolymer-Nukleotidabschnitten, artifizielle Baseninsertionen und Fehler durch unterschiedliche Matrizen-Moleküle an einem Bead. Die sog. Base-Calling-Algorithmen werden jedoch stetig verbessert und eine Re-Analyse alter Läufe damit sinnvoll. Parallel zu dem Problem der Archivierung stoßen die althergebrachten Laborprotokollbücher an ihre Grenzen und müssen bei starker Auslastung der Sequenziermaschinen z. B. im Multi-User-Betrieb einer Serviceeinheit durch professionelle (und teure) Labor-Managementsysteme ersetzt werden.

Entscheidend ist jedoch, dass eine Gemeinsamkeit von Sanger-Technik und NGS leider weiter besteht: alle derzeitigen Verfahren haben methodisch bedingte Beschränkungen in ihrer Leselänge (zwischen 35 Nukleotiden bei SOLiD bis etwa 1000 Nukleotiden bei Sanger). Daher müssen die DNA-Moleküle komplexer Genome zunächst physikalisch zerstückelt und dann in Millionen relativ kleinen Stücke sequenziert werden („Shotgun“-Verfahren, Abb. 4A). Erst der Computer kann aus diesen Schnipseln durch Erkennen von Überlappungen der Nukleotidabfolge zwischen den Schnipseln wieder die gewünschte Sequenz in originaler Moleküllänge rekonstruieren. Gemeinhin wird dies mit einem Puzzle aus Millionen von Teilen verglichen, ohne dass man allerdings das fertige Bild vor Augen hat (Abb. 4B). Diese *de novo*-Assemblierung zusammenhängender DNA aus Sequenzschnipseln zu sog. Contigs (= contiguous sequence) ist als NP-schweres informatisches Problem ohne exakte Lösung bekannt. Mathematisch haben wir es hier mit einem „shortest common superstring“-Problem bzw. Hamilton-Graph zu tun (Abb. 5A). Die Teilsequenzen sind dabei als Knoten dargestellt, eine Sequenzüberlappung zwischen ihnen als Kante. Ziel der Assemblierung ist es, den kürzesten Weg zu rekonstruieren, bei dem alle Knoten nur einmal angesteuert werden. Abb. 5A zeigt, dass schon bei Auftreten geringfügiger Sequenz-Doppelungen zwischen Reads mehrere Rekonstruktionspfade möglich sind. Die nunmehr von NGS erzeugten mittel-langen bis sehr kurzen Reads verschärfen dieses Problem: je kürzer und mehr die Schnipsel sind, desto problematischer gestaltet sich der Assemblierungsvorgang. So wurde gezeigt, dass 750 Bp lange Sanger-Reads das *Neisseria*-Bakteriengenom immerhin in 59 Contigs von meist > 1000 Nukleotiden Länge assemblieren können, während 70 Bp-short reads mehr als 1800 Contigs produzieren. Hier tut also Algorithmenentwicklung Not, die auf die Verarbeitung von Millionen kurzer Reads hin speziell abgestimmt ist.

Das größte Problem bei der Assemblierung ist, dass komplexe Genome von eukaryotischen, vielzelligen Organismen oft viele Millionen sich nahezu perfekt wiederholender „repetitiver“ Sequenzen (> 95 % Ähnlichkeit) besitzen, deren mögliche Funktionen im Genom noch unklar sind (weswegen sie oft vielleicht etwas voreilig als Genom-Müll bezeichnet werden). Das Humangenom besteht zu fast 50 % aus repetitiver DNA, darunter > 1 Mio. Alu-Repeats (300 Nukleotide lang) und etwa 200 000 LINE1-Repeats (meist mehr als 1000 Nukleotide lang). Ist die Leselänge der Sequenziermethode N kleiner als die Repeatlänge n , so kann kein Standard-Assemblierungsalgorithmus hier mehr Ordnung schaffen, da er nicht erkennen kann, an welche Stelle des Genoms genau eine bestimmte repetitive Kopie gehört (Abb. 5B). im Hamilton-Graph verweist jeder Repeat-Knoten auf viele andere ebensolche Knoten. Jedes Repeat ist dann ein Bruchpunkt für die Assemblierung. Das Resultat sind extrem stark

fragmentierte Assemblies: so zeigen Simulationen, dass mit 50 Bp-Reads das Genom des Fadenwurms *Caenorhabditis elegans* (110 Mio. Nukleotide lang) überhaupt nur zur Hälfte und nur zu Contigs von max. 10 000 Nukleotiden zusammengesetzt werden kann (ein natürlich völlig unbefriedigendes Ergebnis). Sanger-Reads mit fast 1000 Bp Leseweite haben hingegen im Humangenom die meisten Repeat-Hürden erfolgreich genommen und Contigs in nahezu vollständiger Chromosomenlänge ermöglicht. Als Lösungsansatz für NGS-Verfahren bietet es sich seit kurzem an, sogenannte „paired end“-Information zu benutzen. Dabei werden die DNA-Moleküle eines Genoms vor der Sequenzierung so getrimmt, dass deren zu sequenzierende Enden sich in einem vom Experimentator bestimmten festen Nukleotidabstand befinden. Diese Abstandsinformation wird bei der Assemblierung genutzt, um repetitive Bereiche zu überbrücken. Doch auch neue algorithmische Möglichkeiten tun sich auf. Pevzner, Waterman und Kollegen haben z. B. anstatt des NP-schweren Hamilton-Pfads für die Assemblierung die Variante eines Euler-Pfad-Ansatzes („de Bruijn-Graph“) vorgeschlagen, der effiziente Lösungen in linearer Zeit verspricht. Hierbei werden anstatt der Knoten (Sequenzen) alle Kanten (Nukleotidüberlappungen) nur einmal begangen (Abb. 5C). Dabei werden Repeats quasi „verklebt“ und ihre Teilsequenzen bei der Erstellung der Gesamtsequenz mehrfach verwendet. Paradoxerweise erfordert diese Methode eine künstliche Zerlegung von Sequenzreads in überlappende Bruchstücke definierter Länge (k -mere, = Kanten des de Bruijn-Graphen; $k-1$ -mere = Knoten). Wenngleich sich auch dieser Ansatz mittlerweile aufgrund der „ungeklärten Herkunft“ der Sequenzreads von je einem der beiden komplementären DNA-Strängen als NP-schwer herausgestellt hat, konnten Brudno et al. kürzlich durch Anwendung bidirektionaler Graphen das Optimierungsproblem in polynomialer Zeit lösen.

Ungeachtet der sehr aktiven Forschung in diesem Bereich der Bioinformatik zeigen unsere Vergleiche implementierter Assemblierungswerkzeuge bisweilen extreme Unterschiede in ihrer Performance: so variierte kürzlich die Anzahl von erstellten Contigs eines 45 000 Reads umfassenden cDNA-Sequenzierungsprojekts zwischen 400 und 7000. Natürlich ist so ein Ergebnis durch Assemblierungsparameter (z. B. Länge und Match-Qualität der erforderlichen Überlappung zweier Sequenzen) bedingt. Oftmals verhalten sich gerade die von NGS-Unternehmen mitgelieferten „Komplettlösungen“ wie eine „Schwarze Box“ mit intern definierten Einstellungen. Auch geeignete Benchmarking-Datensätze sind rar, so dass es schwer fällt, das derzeit bestfunktionierende Tool zu identifizieren. Darüber hinaus stoßen kleine Arbeitsgruppen, selbst solche mit kleinen Cluster-Architekturen, bei umfangreicheren Assemblierungsprojekten rechnerisch schnell an Grenzen. Selbst die Alignierung kurzer Reads eines menschlichen Genoms an die in Datenbanken hinterlegte Human-Referenzsequenz dauert mehrere Tage. Hier wird es Aufgabe sein, die entsprechenden universitären Rechenzentren und die NGS-Lieferanten zur Anpassung der Firmensoftware auf die hauseigenen Systeme zu bewegen. Zudem erscheinen innovative Lösungen, wie die Portierung der Prozesse auf sehr schnelle, kostengünstige Grafikkarten eine Perspektive.

Vergessen werden darf jedoch nicht, dass die Erstellung von Contigs für Genomsequenzen oder mRNA-Transkripte erst den Anfang der bioinformatischen Analyse darstellt. Erst danach erfolgt die Suche nach dem Informationsgehalt der Sequenzen: welches Protein wird kodiert? Enthält die DNA eine funktionelle RNA-Sekundärstruktur? Gibt es unterschiedliche Spleißvarianten der mRNA? (u.v.m.). Diese *downstream*-Analyse erfordert erneut die intensive Zusammenarbeit von Molekularbiologen und Bioinformatikern. Hier wird dann auch ein entscheidendes Erfolgskriterium der neuen Sequenzieretechniken liegen. Die Zeitschrift *Nature Biotechnology* vergleicht daher eine NGS-Maschine ohne komplette Bioinformatik-Infrastruktur mit einer Stradivari-Geige ohne den passenden Bogen.

Genomforschung und NGS in Mainz

Die Genomforschung hat erhebliche Tradition in Mainz. Schon Ende der 70er Jahre wurden in der AG Schmidt damals noch mit der sogenannten „chemischen Sequenzierung“ nach Maxam und Gilbert kleine Genomabschnitte sequenziert. Dies geschah noch vollkommen manuell und entsprechend langsam tröpfelten die Sequenzdaten. Nach Umstellung auf die fluoreszenzbasierte Online-Sanger-Sequenzierung im Institut für Molekulargenetik 1990 war es soweit: mit der damals leistungsfähigsten Sequenziermaschine konnten wir uns an internationalen Genomprojekten beteiligen, so z. B. 1996 an der Entschlüsselung des 12.5 Mb großen Genoms der Bäckerhefe, einem Projekt, an dem mehr als 80 Gruppen beteiligt waren. Eine deutlich anspruchsvollere Größenordnung war das Humangenomprojekt, zu dessen nationaler Förderung sich die Bundesregierung leider etwas spät entschließen konnte. Wir in Mainz haben in diesem Projekt den damals innovativen Ansatz der „komparativen Genomik“ verfolgt, indem wir parallel einen humangenetisch interessanten Chromosomenabschnitt des Menschen (1 Mio Nukleotide auf Chromosom 11p15.3) und die entsprechende Region des Mausgenoms (Chromosom 7) sequenziert haben. Mit Hilfe des Vergleichs der beiden homologen Sequenzabschnitte konnten wir in den zunächst anonymen Sequenzen deutlich besser Genstrukturen erkennen, weil Genabschnitte meist der negativen Selektion unterliegen und damit evolutionär deutlich konservierter sind als Nicht-Gene. So erweisen sich Gene bei Maus und Mensch häufig in über 90% der Sequenzpositionen als identisch, während intergenische Bereiche mit weniger als 60 % Übereinstimmung manchmal kaum zu alignieren sind. Unser damaliger „Pilot-Ansatz“, der schon wenige Jahre später durch die Gesamt-Sequenzierung des Mausgenoms Bestätigung fand, war sehr erfolgreich. Wir konnten in relativ kurzer Zeit 15 Gene hochakkurat definieren, darunter auch viele völlig neue Gene, deren Funktionen auch heute noch nicht aufgeklärt sind.

Mit der Einführung der NGS-Technologie wird die Diversität der nun beantwortbaren biologischen Fragestellungen und die Zahl der Genomik-Projekte erheblich anwachsen:

- In der AG Zischler (Anthropologie) wird z. B. eine völlig neue Genklasse, piRNA bzw. PIWI-interagierende RNA genannt, untersucht. piRNA-Gene kodieren kleine RNA-Moleküle im Größenbereich von nur 21 –30 Nukleotiden. Ihnen kommt eine wichtige Funktion in der Abwehr „springender“ Nukleinsäuren (Transposons) zu, die über evolutionäre Zeiträume in der Lage sind, Genome regelrecht zu parasitieren. Indiz für das fortwährende Bombardement des Genoms durch solche Transposons ist, dass wir heute beim Menschen ca. 65 rezente Insertionen springender DNA kennen, deren Integrationen zur Zerstörung von Genen und somit zu genetischen Krankheiten geführt haben. Durch qualitative und quantitative Katalogisierung der vermutlich mehr als 50 000 piRNA-Genloci im Menschen und in nicht-humanen Primaten soll letztlich erforscht werden, auf welche Weise piRNA-Pools die Besiedlung von Genomen mit springender DNA kontrollieren können. Nur NGS-Methoden können mit angemessenem Aufwand die erforderlichen Daten produzieren.

- Die prähistorische Populationsgenetik steht im Mittelpunkt des Interesses der AG Burger (Anthropologie, Palaeogenetik). In hochreinen Spurenlabors untersuchen die Anthropologen alte DNA aus archäologischen Skeletten des Menschen und seiner Haustiere. Ziel der langwierigen Untersuchungen ist die Rekonstruktion der Besiedlungsgeschichte Europas und Zentralasiens im frühen Holozän. Durch NGS kann die bisherige Datenmenge etwa 20-fach erhöht werden, womit eine kleinräumig detaillierte Rekonstruktion prähistorischer demographischer Dynamik endlich möglich sein wird.

- Die AGs Lieb (Zoologie) und Hankeln (Molekulargenetik) arbeiten im Rahmen des DFG-Schwerpunktprogramms 1174 „Deep Metazoan Phylogeny“ mit Phylogenomik an der

Aufklärung der stammesgeschichtlichen Stellung z. T. sehr exotischer Tiergruppen, für die es bislang so gut wie keine Gen(om)information gibt. Hierzu zählen extrem seltene Molluskenarten aus der Tiefsee, Rädertierchen mit ungewöhnlichen Fähigkeiten zur Kryptobiose, asexuellen Vermehrung und Strahlungsresistenz und Kratzwürmer, die z. B. Fische parasitieren. Die durch NGS erhobenen Sequenzdaten sollen eine stabile Rekonstruktion phylogenetischer Bäume ermöglichen und einen Einblick in das interessante Genrepertoire dieser Exoten geben.

- In einem Kooperationsprojekt mit der Uni Frankfurt bearbeitet die AG Hankeln (Molekulargenetik) ökologische und evolutionsbiologische Fragestellungen bei Insekten der Gattung Chironomus (Zuckmücken), deren Larven in Gewässer-Ökosystemen mit manchmal mehr als 1000 Individuen pro m² einen Großteil der Biomasse ausmachen. Obgleich Chironomiden im Rahmen von OECD-Tests etablierte Modellorganismen der Ökotoxikologie darstellen, ist ihr Genom quasi unerforscht. Durch NGS-Transkriptomanalyse sollen Gene identifiziert werden, die eine klimatische Anpassung von Chironomiden steuern und möglicherweise ökologische Voraussagen des Verhaltens von Zuckmückenpopulationen angesichts des Klimawandels treffen lassen. Weiterhin sollen Gene extrahiert werden, die eine mögliche genetische Grundlage für Artbildungsprozesse und die auffällig divergente Evolution der Genomstruktur bei nahe verwandten Schwesterarten von Chironomus bilden.

- Ein aus der Sicht des Genomikers schwieriges Gewebe ist Knochen bzw. Knorpel, weil sich daraus nur sehr schwer qualitativ hochwertige RNA gewinnen lässt und damit die Transkriptomanalyse sehr erschwert ist. Sowohl im Rahmen des Deutschen Humangenomprojektes als auch als Mitglied im EU-Konsortium „EUROGrow“ geht es in der AG Schmidt um die Gene, die beim Aufbau von Knochen und Knorpel eine Rolle spielen. Viele dieser Gene sind für Skeletterkrankungen verantwortlich, so dass diese Transkriptomanalyse auch eine erhebliche medizinische Relevanz besitzt. Ein weiteres Ziel ist die Aufklärung der Genexpressionsmuster bei der Differenzierung mesenchymaler Stammzellen mittels NGS. Die *in vitro*-Differenzierung dieser Zellen zu Chondroblasten und Chondrozyten wird in der regenerativen Medizin von großem Interesse sein.

Das neue Zentrum für Nukleotidsequenzanalysen im Gigabasen-Maßstab soll an vorhandene Sequenzier- und Nukleinsäureanalyseeinrichtungen angegliedert werden. Diese über die Jahre aufgebauten apparativen und „know how“-Ressourcen bilden eine Grundlage, auf der die NGS-Technologie erfolgreich und ohne Anlaufschwierigkeiten aufgebaut werden kann. Die Methodik der Genom- und Transkriptomforschung inklusive der klassischen Sequenzierverfahren ist im Rahmen des Kompetenz-Zentrums für Nukleinsäureanalyse in den letzten 20 Jahren ständig fortentwickelt worden. Für die jetzt vorgesehene Ausbaustufe planen wir zunächst die Etablierung der 454-Technologie, die insbesondere eine *de novo*-Sequenzierung erlaubt. Dies ist für viele Forschergruppen der Johannes Gutenberg-Universität eine wichtige Voraussetzung. In einer zweiten Ausbaustufe, für die aber inzwischen schon sehr konkrete Planungen existieren, soll eine der für andere Anwendungen überlegenen Short-Read-Plattformen ebenfalls etabliert werden. Bei voller Auslastung werden hierbei bislang unvorstellbare Datenmengen erzeugt, deren Verarbeitung und Auswertung wiederum nur möglich ist, wenn entsprechend leistungsfähige Rechneranlagen mit entsprechender wissenschaftlicher Kompetenz dies übernehmen. Ein Kernstück des neuen NGS-Zentrums ist daher der Forschungsschwerpunkt „Rechnergestützte Forschungsmethoden in den Naturwissenschaften“, der einen Standortvorteil für die Universität in dieser Hinsicht darstellen wird.

Neue Technologien haben seit jeher den wissenschaftlichen Fortschritt in den Naturwissenschaften entscheidend mitbestimmt. Im Bereich der Genomforschung, die in den „life sciences“ einen immer wichtigeren und größeren Raum einnimmt, sind die neuen Sequenziertechnologien ein solcher zukunftsbestimmender Faktor, auf den wir auch in Mainz nicht verzichten können.

Übersichtsliteratur:

Pop, M. and Salzberg, S.L. (2008) Bioinformatics challenges of new sequencing technology. Trends in Genetics 24:142-149

Rothberg, J.M. and Leamon, J. H. (2008) The development and impact of 454 sequencing. Nature Biotechnology 26:1117-1124

Shendure, J. and Li, H. (2008) Next-generation DNA sequencing. Nature Biotechnology 26: 1135-1145

Abbildungsunterschriften und -legenden:

Abb. 1 Die 454-Pyrosequenzierung. A. Prinzip der Sequenzreaktion durch DNA-Synthese. B. Detektion der Lichtsignale auf Picotiterplatten-Ausschnitt. C. Darstellung des Sequenzierungsergebnisses als „Flowgram“.

Abb. 2 Die Solexa/Illumina-Sequenzierung. Das Verfahren beruht ebenfalls auf „sequencing-by-synthesis“. Charakteristisch sind die Befestigung der Matrizenmoleküle auf einem Träger und die Verwendung „reversibler Farbstoff-Terminatoren“ als DNA-Bausteine bei der Sequenzierung.

Abb. 3 Die ABI/SOLiD-Sequenzierung. A. Das Verfahren beruht auf „sequencing-by-ligation“. Im ersten Ligationszyklus wird ein komplexes Gemisch farbstoff-kodierter Oktanukleotide zugegeben. Bei der Hybridisierung und Ligation des passenden Moleküls werden die ersten zwei Positionen der unbekannt Sequenz (rot) detektiert. Danach werden die letzten 3 Nukleotide der in Zyklus 1 gebundenen Sonde inklusive des Fluorophors abgespalten und eine neue Ligationsrunde mit dem SONDENGEMISCH initiiert. Dabei werden erneut zwei Positionen in definiertem Abstand zu den ersten entschlüsselt (siehe Schema unten). Die Verwendung unterschiedlich langer Primer (Runde 2, 3 usw.) erlaubt das sukzessive Lesen aller Positionen. Dabei werden alle Positionen des Matrizenmoleküls zweimal gelesen, was zur Sequenzierungsgenauigkeit beiträgt. B. Dekodierung der gemessenen Farbsignale („color space“) in eine DNA-Sequenz. Ist das erste Nukleotid (pink) bekannt (weil es zu der bekannten Adaptersequenz gehört) und die erste detektierte Farbe ein GRÜN, so muss Nukleotid 2 ein C sein. In der n+1-Runde (Zeile 2) wird zuerst ROT gemessen und Position 2 ist bekanntermaßen C: Nukleotid 3 muss daher G sein.

Abb. 4 Genomsequenzierung im Shotgun-Verfahren. A. Herstellung der Teil-Sequenzierungen (Reads) B. Assemblierung der Gesamtsequenz in drei Schritten („overlap-layout-consensus“-Ansatz).

Abb. 5 Rekonstruktion der Gesamtsequenz durch Graphen-theoretische Verfahren. A. Hamilton-Graphen (H). Die Reads (S) bestehen in diesem einfachen Beispiel aus Trinukleotiden. Das Beispiel unten ist nicht eindeutig lösbar, da zwei Pfad-Möglichkeiten für die Assemblierung bestehen. B. Hamilton-Pfade sind problematisch, wenn eine Sequenz mit Repeats durchsetzt ist. C. Euler-Pfad als alternativer Assemblierungsansatz.

Summary

Life science research is currently being revolutionized by novel ultrahigh-throughput methods for deciphering genomic information of organisms. Within a few years, these new DNA sequencing technologies (termed “Next-Generation Sequencing“, NGS) will most probably allow us to know our own personal genome at reasonable costs, with enormous biomedical impact. The huge amounts of sequence data produced by NGS, however, create big challenges for bioinformatics methods to keep pace. The University of Mainz ‘Competence Center for Nucleic Acid Analysis’ and the research focus ‘Computational Sciences Mainz’ are joining forces to centrally establish NGS technology.

Prof. Dr. Thomas Hankeln

Thomas Hankeln, Jahrgang 1959, hat Biologie und Geographie an der Ruhr Universität Bochum studiert und dort 1990 promoviert. Nach seinem Wechsel nach Mainz hat er verschiedene Themen der Genomforschung bearbeitet und sich 1998 mit Arbeiten zur Molekularen Evolution von Genfamilien im Fach Genetik habilitiert. Seit 2001 ist Thomas Hankeln C3-Professor am Institut für Molekulargenetik der Johannes Gutenberg Universität. Gegenwärtige Forschungsthemen umfassen die Identifizierung von Genen durch vergleichende Genomanalyse, die Funktionsaufklärung solcher Gene in Tiermodellen sowie die Verwendung von Genomdaten zur phylogenetischen Systematik. Thomas Hankeln ist Mit-Gründer des seit 1998 bestehenden Biotechnologieunternehmens GENterprise GmbH, das im Bereich der Genomforschung Service-Dienstleistungen anbietet und F&E-Projekte durchführt.

Prof. Dr. Hans Zischler

Hans Zischler, Jahrgang 1957, studierte Biologie an den Universitäten Hohenheim und Tübingen. Er hat seine Doktorarbeit am MPI für Psychiatrie, München angefertigt und wurde 1991 promoviert. Nach Postdoc-Zeiten am MPI für Psychiatrie und an der LMU München übernahm er 1997 die Leitung der Arbeitsgruppe Primatengenetik am Deutschen Primatenzentrum in Göttingen. 2002 wurde er auf den Lehrstuhl für Anthropologie der Universität Mainz berufen. Seine Forschungsschwerpunkte liegen auf der Analyse von evolutionären Mustern und Prozessen innerhalb der Divergenz nicht-humaner Primaten und des Menschen.

Prof. Dr. Erwin R. Schmidt

Erwin R. Schmidt, Jahrgang 1949, studierte von 1967 bis 1973 Biologie und Chemie an der Justus Liebig-Universität in Gießen. 1975 hat er am dortigen Institut für Genetik promoviert und ist anschließend als DFG-Postdoc an die Ruhr-Universität Bochum in die Arbeitsgruppe Genphysiologie im Fachbereich Biologie gegangen. 1978 erfolgte der Wechsel in die Medizin ans Institut für Genetik. Dort habilitierte er 1985 in Genetik mit Arbeiten über die molekulare Struktur von repetitiven Genomelementen und ihre Rolle bei der Speziation. Nach drei Jahren als Professor C2 für Molekulargenetik in Bochum folgte er 1989 einem Ruf auf eine Professur für Molekulargenetik nach Mainz. 1992 erhielt er den Ruf auf eine Professur für Genetik in Stuttgart-Hohenheim. Nach einem Jahr in Stuttgart kehrte er zurück nach Mainz, um den Ruf auf die Professur für Molekulargenetik anzunehmen. 1994 wurde er Leiter des neu gegründeten Instituts für Molekulargenetik, gentechnologische Sicherheitsforschung & Beratung, an dessen Entstehung er wesentlich beteiligt war. 1998 gründete er zusammen mit weiteren Kollegen die Biotechnologiefirma GENTERprise, die in Kooperation mit Universitäten und anderen Forschungseinrichtungen mehrere nationale und internationale Forschungsprojekte durchgeführt hat. Seine Forschungsschwerpunkte liegen in der Gen- und Genomforschung, sowie der Nukleinsäureanalytik. Seit April letzten Jahres dient er darüber hinaus dem Fachbereich Biologie als Dekan.

Abb. 2 Solexa/Illumina-Sequenzierung

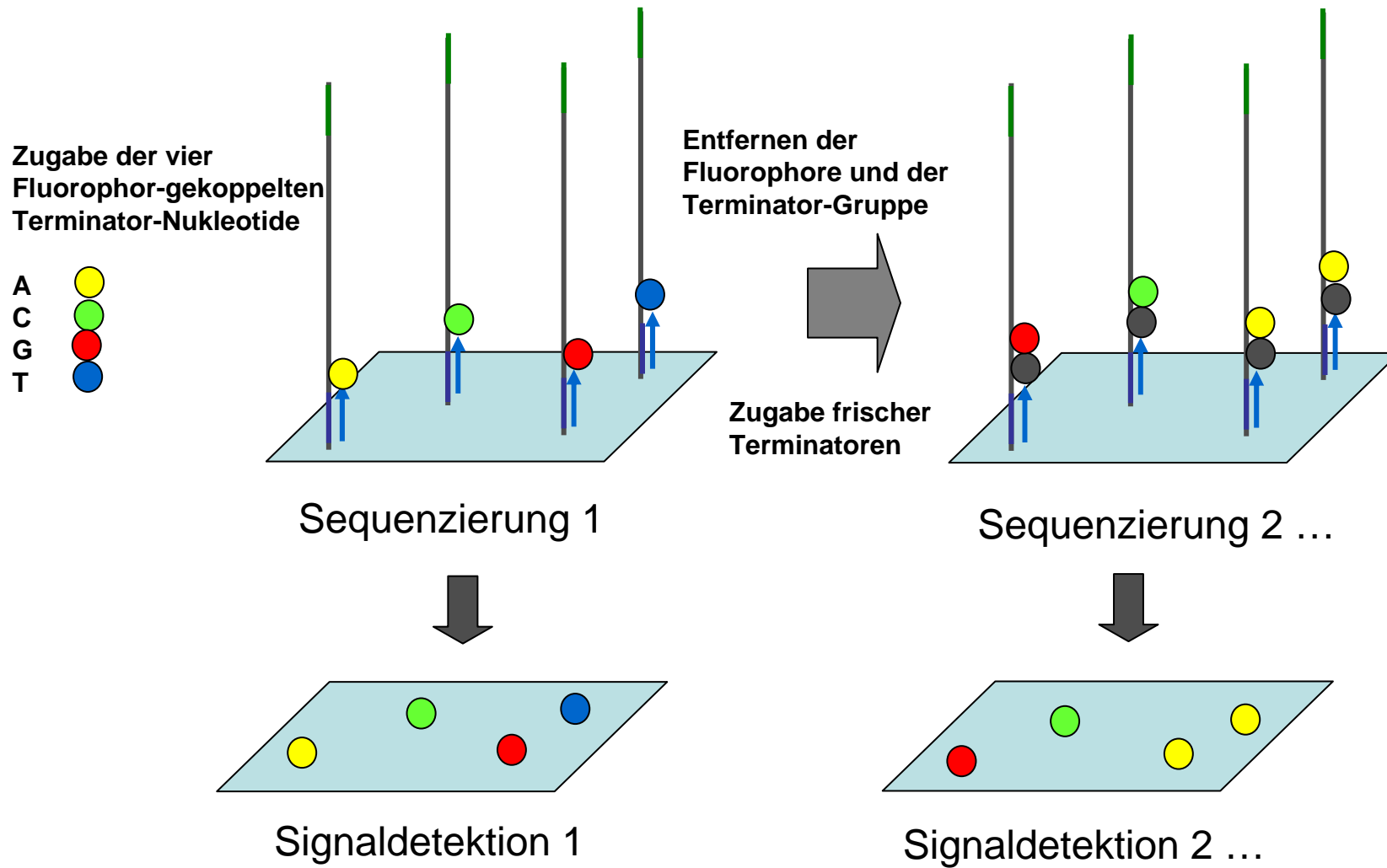
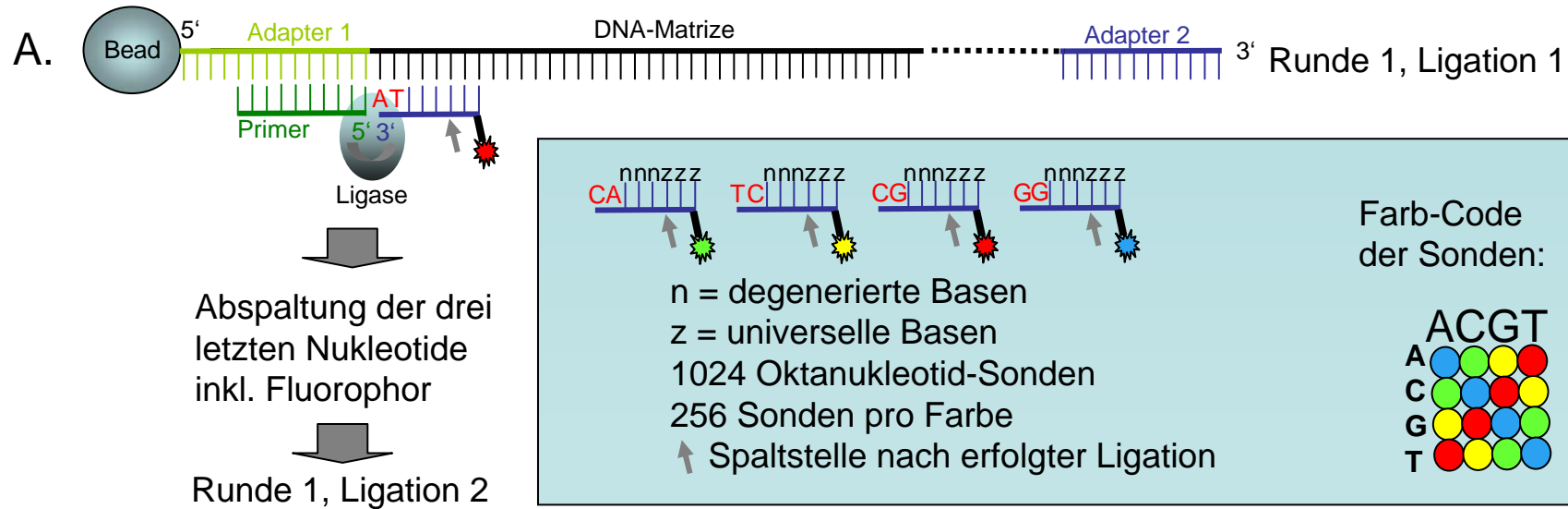


Abb. 3 SOLiD/ABI-Sequenzierung



Runde	Ligationszyklus					entzifferte Positionen (immer doppelt gelesen!)
	1	2	3	4	5	
1	Primer 1,2	6,7	11,12	16,17	21,22	}
2	Primer n-1 0,1	5,6	10,11	15,16	20,21	
3	Primer n+3 4,5	9,10	14,15	19,20	24,25	

und weitere ...

B. Dekodierung der SOLiD-Farbsignale

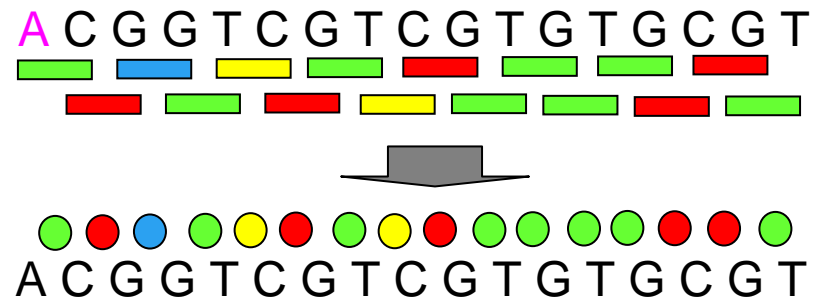
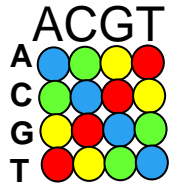


Abb. 4A Genom-Sequenzierung im Shotgun-Verfahren

Genom-DNA-Moleküle

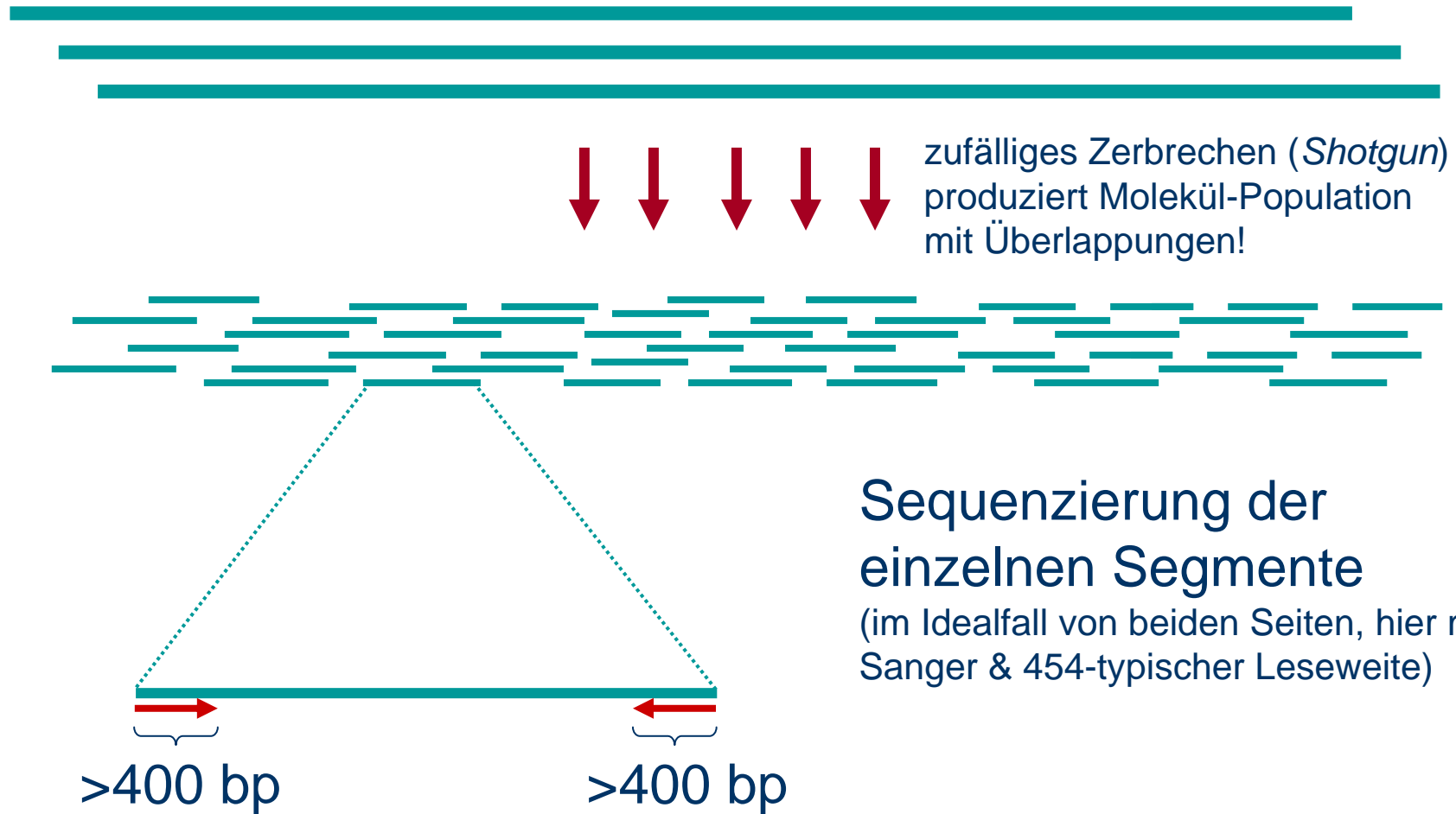


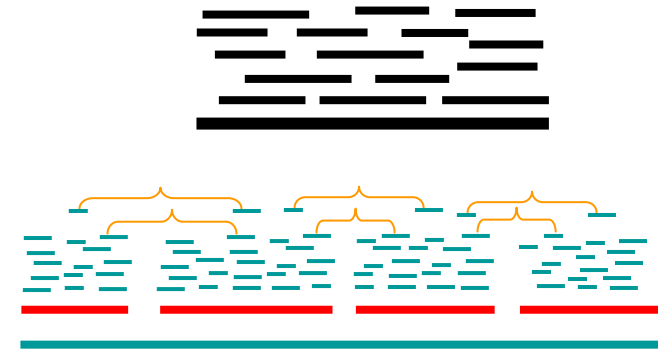
Abb. 4B Shotgun-Verfahren: Assemblierung der Gesamtsequenz in 3 Schritten

Overlap: finde überlappende Sequenz-Reads



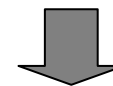
Layout: Vereinige Reads zu Contigs
und Contigs zu Supercontigs

(Klammern: "paired end"-Information hilft
bei der Orientierung der Contigs)



Consensus: Leite die DNA-
Sequenz ab und korrigiere
dabei Lesefehler

```
TAGATTACACAGATTACTGA TTGATGGCGTAA CTA  
TAGATTACACAGATTACTGACTTGGATGGCGTAAACTA  
TAG TTACACAGATTATTGACTTCATGGCGTAA CTA  
TAGATTACACAGATTACTGACTTGGATGGCGTAA CTA  
TAGATTACACAGATTACTGACTTGGATGGGGTAA CTA
```



...TAGATTACACAGATTACTGACTTGGATGGCGTAA CTA...

Abb. 5A HAMILTON-GRAPHEN zur Sequenzrekonstruktion

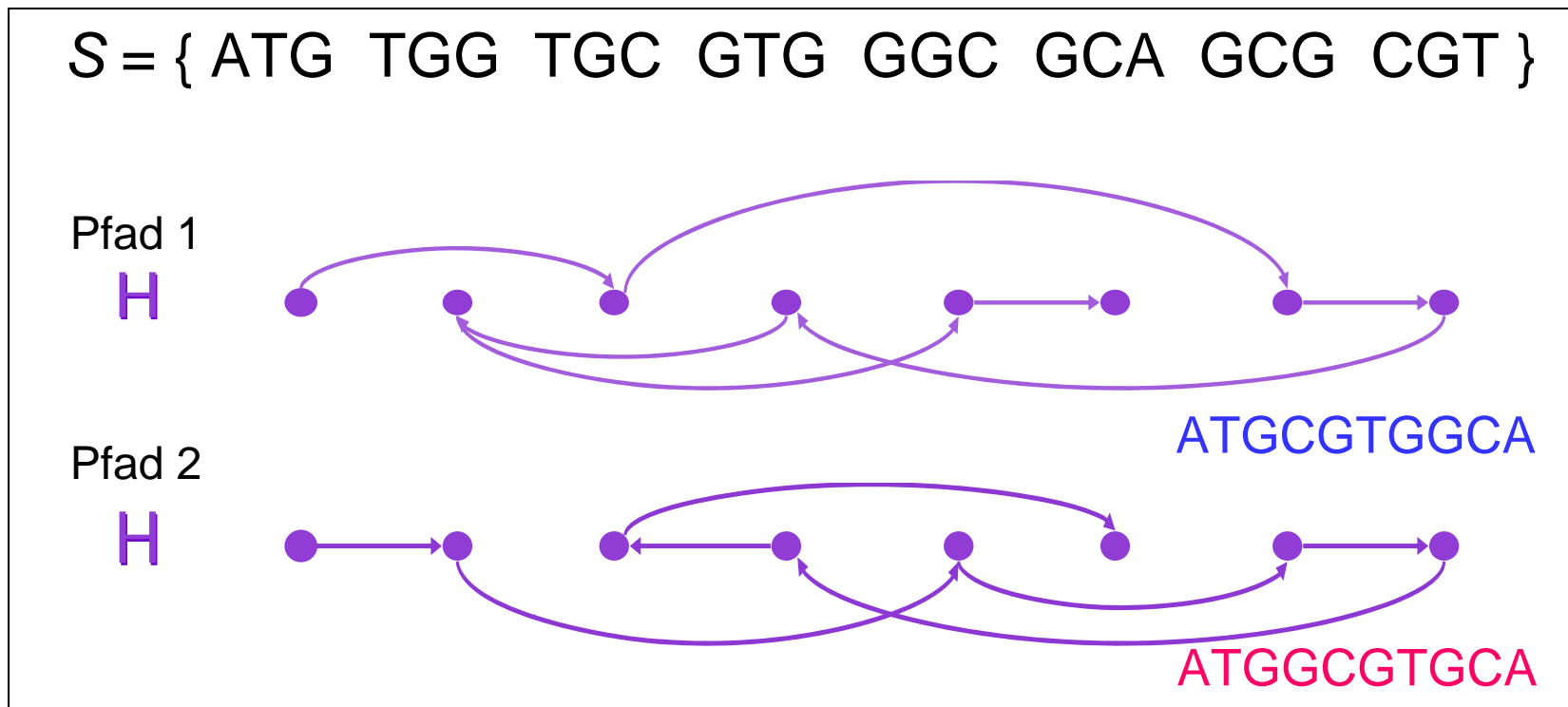
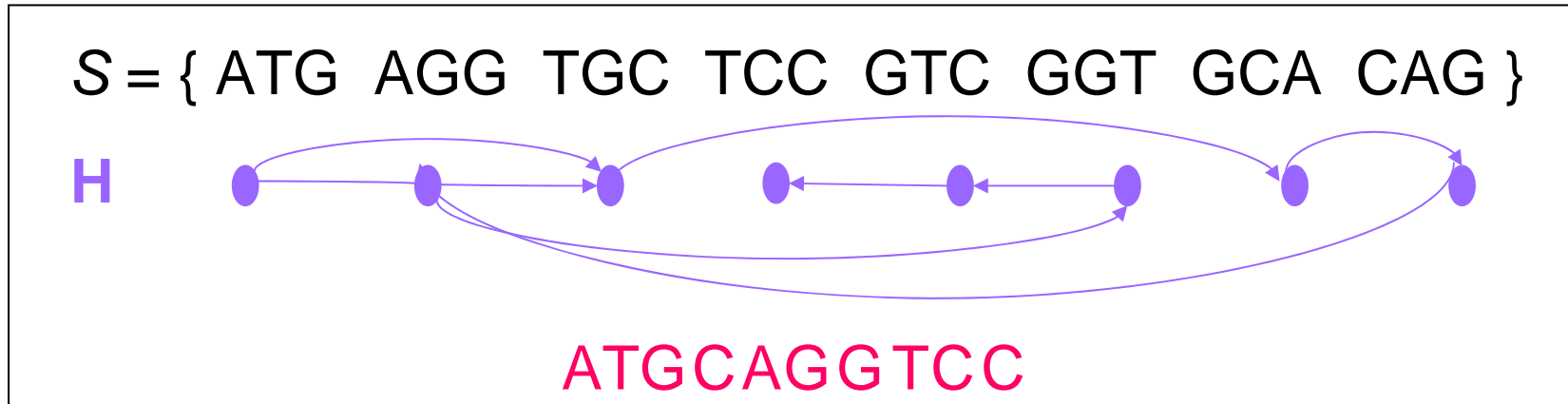
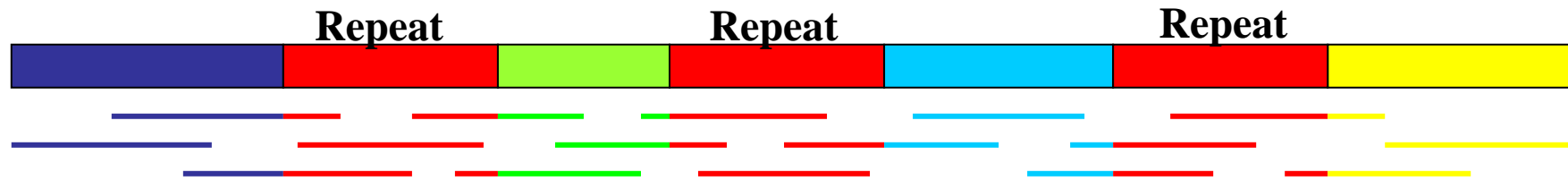
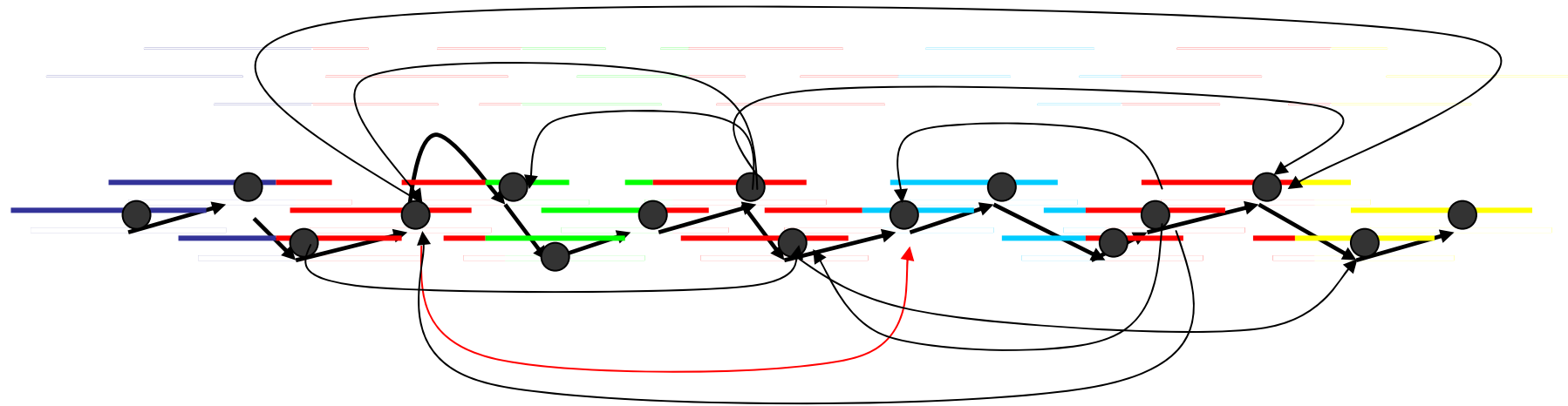


Abb. 5B Das HAMILTON PFAD-Problem der Genom-Assemblierung

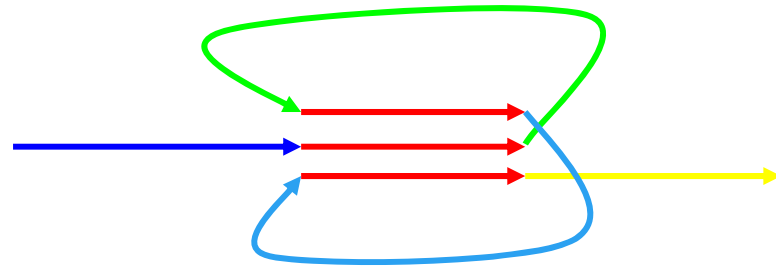
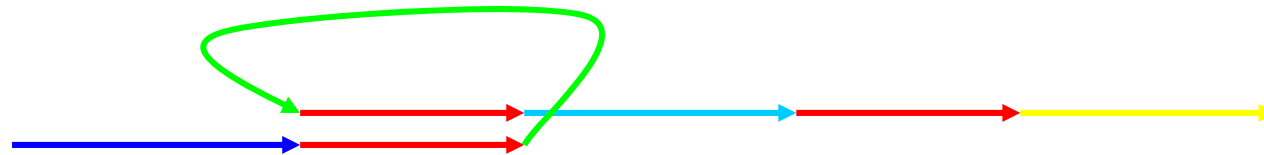


Jeder Sequenz-Read wird durch einen Knoten repräsentiert. Knoten aus Repeats sind mit vielen anderen Knoten assoziiert.

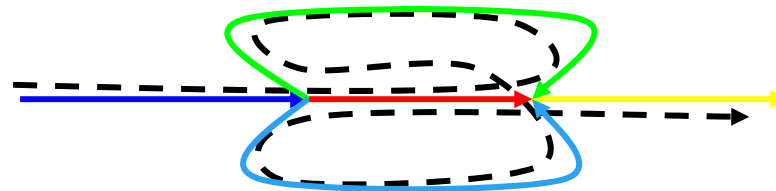


Finde den Pfad, der jeden *Knoten* einmal passiert:
Hamilton-Pfad-Problem (NP-schwer)

Abb. 5C Der EULER-PFAD-Lösungsansatz



“Verkleben” der Repeat-Kanten ergibt Fortschreiten des Pfads durch die gesamte Sequenz.



Finde einen Pfad, der jede *Kante* einmal passiert:

Euler-Pfad (linearer Aufwand)