

# LABORWELT

Nr. 6 / 2008 – 9. Jahrgang

## Laborautomation

Metabolom-Analyse von  
Industrie-Mikroorganismen  
im Hochdurchsatz

Automatisierte Optimierung  
von Zellkultur-Parametern  
und Produktionsstämmen

Großer Sonderteil:  
Medizintechnik und  
Laboratoriumsmedizin



Marktübersicht:  
DNA-Synthese

BIOCOM

# Next Generation-Bioinformatik

Dr. Kerstin A. Stangier, Dr. Yadhu Kumar, GATC Biotech AG, Konstanz

Heute ist es keine Herausforderung mehr, mit Sequenziergeräten der neuesten Generation – sogenannten Next Generation-Sequenzern – Sequenzrohdaten im Gigabasen-Bereich in wenigen Tagen zu generieren. In den letzten Jahren ist die Leistungsfähigkeit dieser Geräte regelrecht explodiert. Während ein klassisches, auf der Sanger-Technologie basierendes System im Durchschnitt täglich eine Megabase produziert, schafft ein Next Gen-Sequencer mindestens 400 Megabasen pro Tag. Die Gerätehersteller entwickeln kontinuierlich neue Reagenzien, mit deren Hilfe die Leseweite erhöht und somit die Leistungsfähigkeit der Geräte weiter gesteigert wird.

Die Leselänge macht den wesentlichen Unterschied zwischen den Technologien hinsichtlich der bioinformatischen Analyse aus. Sie variiert zum Beispiel bei den Short-read-Technologien von 50 bis 100 bp (Illumina Genome Analyzer II, Applied Biosystems SOLiD System), über rund 250 bp oder 400 bp (Roche Diagnostics GS FLX) bis zu 1.100 bp der klassischen Sanger-Technologien (z.B. Applied Biosystems 3730xL).

Aufgabe der Bioinformatik ist es, effiziente Methoden zur Auswertung dieser gigantischen Datenmengen und heterogenen Datentypen anzubieten. Dennoch hinkt die Entwicklung der Bioinformatik dem rasanten Fortschritt der Sequenziertechnologien noch hinterher.

Es stellen sich viele Fragen, wie etwa: Welche verfügbaren Software-Tools können die Datenmengen verarbeiten und analysieren? Wie ist die Qualität der Analysen einzustufen? Ist es sinnvoll, Rohdaten für zukünftige Forschung zu speichern? Die derzeit von den Geräteherstellern mitgelieferten Software-Tools bieten noch keine optimale Lösung für alle Anwendungen. Dies trifft ganz besonders für Analysesoftware von

Short Reads sowie bei „Hybrid-Projekten“ zu, bei denen mehrere Technologien kombiniert werden. Zudem macht die Weiterentwicklung der Applikationsprotokolle (z.B. Paired-End-Analysen) eine kontinuierliche Anpassung und Verbesserung der verwendeten Algorithmen notwendig. Daher hat der Sequenzierdienstleister GATC Biotech aus Konstanz – Anwender der Technologien von Roche, Illumina und Applied Biosystems mit einer jährlichen Kapazität von mehr als einer Tera-Base – einige auf dem Markt verfügbare Analysetools getestet.

Dabei ging GATC unter anderem insbesondere der Frage nach, ob die Ergebnisse einfach in andere Programme übertragen und benutzerfreundlich für Visualisierungen aufbereitet werden können.

## Assemblierung und Alignment

Bei der Assemblierung werden die Daten komplett neu zu einer möglichst langen Sequenz zusammengesetzt. Alignments/Mappings

basieren auf einer Referenzsequenz. Zunächst erscheint dies einfacher als eine *de novo*-Assemblierung. Mapping ist wie das Zusammensetzen eines Puzzles, wobei sich einige Puzzle-Teile (reads) jedoch extrem ähnlich sind (Repeats). Ein Teil stimmt mit der Referenzsequenz überein, andere Abschnitte sind unterschiedlich aufgebaut (strukturelle Variationen, Segment-Duplikationen), oder Stücke des re-sequenzierten Genoms sind völlig neu (z.B. Insertionen).

Reads aus repetitiven Strukturen sind in beiden Fällen kompliziert zu verarbeiten; einige bioinformatische Tools ordnen diese zufällig an den passenden Stellen an, andere sortieren sie ganz aus. Daher muss die eingesetzte Software – abhängig vom jeweiligen Projekt – sorgfältig ausgewählt oder sogar mehrere Tools nacheinander für eine sequentielle Analyse kritischer Bereiche in einer Pipeline zur Verfügung gestellt werden. Eine weitere Herausforderung stellt die Analyse von Paired-End-Reads dar. Da die Protokolle für die Erstellung von Banken mit großen Inserts immer besser werden, kann mit der Kombination von Libraries unterschiedlicher Insertgrößen die Problematik repetitiver Strukturen überwunden werden.

## Assemblierung

SeqMan NGen von DNASTAR kann auf einem Desktop-Computer Next Gen-Sequenzdaten von Illumina und Roche sowie Daten klassischer Sanger-Sequencer miteinander assemblieren. Eine Schnittstelle zu SeqMan Pro in der DNASTAR Lasergene Suite erlaubt eine Visualisierung und weitere Analyse. Datenfiles aus SeqMan NGen können in SeqMan Pro importiert werden, um etwa Dual-End-Sequenzdaten darzustellen. Der Benutzer kann Basen editieren, Qualitätswerte anschauen sowie SNPs detektieren und filtern.

Der mit dem Next Gen-System mitgelieferte GS De Novo Assembler (Newbler Assembler) von Roche/454 Life Science verarbeitet sowohl Single- als auch Paired-End-Reads. Auch Hybrid-Assemblies aus GS FLX-Daten und Sanger-Daten können generiert werden. Für den Export stehen verschiedene File-Formate zur Verfügung.

MIRA (Mimicking Intelligent Read Assembly)<sup>1</sup> eignet sich besonders für die Verarbeitung von Genomen mit vielen repetitiven Sequenzen. Die derzeit verfügbare Version kann GS FLX-Reads und Sanger-Reads aus *de novo*-Genomprojekten assemblieren.

Velvet<sup>2</sup> ist eine vom European Bioinformatics Institute (EMBL-EBI, Cambridge, UK) entwickelte Software zur *de novo*-Assemblierung von sehr kurzen Next Gen-Rohdaten, auch in Kombination mit GS FLX-Daten. Um Repeats aufzulösen und Contigs zu orientieren, können Paired-End-Reads und Sequenzen aus der klassischen Sanger-Sequenzierung eingelesen werden.

Edena (Exact De Novo Assembler)<sup>3</sup> ist ähnlich zu Velvet, kann derzeit aber keine Paired-End-Assemblies verarbeiten.



Abb. 1: Grafische Darstellung der Contig-Verteilung einer *de novo*-Assemblierung. Je größer der Kreis, desto größer das Contig. Es handelt sich um die Genomsequenz eines Bakteriums, sequenziert mit dem Roche GS FLX.

Bei Tests mit Datensätzen verschiedener Technologien oder deren Kombination und bei unterschiedlichen Organismengrößen zeigte sich, dass keines der Tools für alle Anwendungen das beste Ergebnis erzielt.

### Alignment / Mapping

Einige der genannten Tools, z.B. SeqMan NGen, können für *de novo*-Assemblierungen sowie Alignments eingesetzt werden. Neben SeqMan NGen und der Roche Diagnostics GS Reference Mapper-Software stehen auch für Alignments eine Reihe von Freeware-Tools zur Verfügung:

ELAND (Efficient Large-Scale Alignment of Nucleotide Databases) sucht in einer Referenzsequenz identische Bereiche zu kurzen DNA-Reads und erlaubt dabei bis zu zwei Abweichungen (InDels) werden nicht abgedeckt, da die Software in den gemappten Reads an solchen Stellen keine Lücken einfügt.

Mosaik (von Michael Stromberg, Boston University) verarbeitet eine große Bandbreite unterschiedlicher Readlängen und kann über InDel-Bereiche mappen, aber keine Paired-end-Informationen nutzen.

Maq (<http://maq.sourceforge.net/index.shtml>), ZOOM (Zillions of Oligos Mapped)<sup>4</sup> sowie SOAP (Short Oligonucleotide Alignment Program)<sup>5</sup> sind speziell für das Mapping von Short Reads konzipiert und können mit Insertionen und Deletionen umgehen.

Auch hier ergab die genaue Betrachtung, dass sich lediglich Tendenzen erkennen lassen, nicht aber eine Software der Problemlöser schlechthin ist. Je nach Organismus, Homologie zur Referenzsequenz und anderen beeinflussenden Parametern muss das eingesetzte Tool zur Analyse mit Bedacht gewählt werden.

Neben den beiden häufigsten Anwendungen, der Assemblierung und dem Alignment, sind teilweise nachgelagerte Analysen erforderlich, wie zum Beispiel BLAST und Annotation. Die Daten müssen nicht nur kompatibel, sondern auch transferierbar sein.



**Abb. 2:** Grafische Darstellung mittels unterschiedlicher Schriftgrößen und -farben von SAGE-Tags (Serial Analysis of Gene Expression) eines Transkriptom-Experiments. Je größer die Schrift, umso häufiger wird das SAGE-Tag im Sample exprimiert. Unterschiedliche Farben repräsentieren unterschiedliche Tags. Die Sequenz stammt vom Mausgenom, sequenziert mit dem Illumina Genome Analyzer II.

### Datentransfer

Die ursprünglichen Rohdaten, die die Next Generation-Sequenzierer als Ergebnis generieren, sind Bilddateien mit riesigem Datenvolumen. Diese Dateien werden direkt in den Systemen zu Qualitäts- und Sequenzdaten verarbeitet. Aufgrund der extrem hohen Readzahl erreichen auch diese Files immer noch eine gewisse Größe. Zur weiterführenden Analyse müssen die Daten auf andere Computer transferiert werden, was dank der Entwicklung von Glasfaser-Kabel, High Speed-Internet und immer preiswerteren Datenspeichern mit sehr hoher Kapazität im Giga- und Tera-Bereich relativ einfach ist.

### Datenspeicherung

Obwohl Datenspeicher heutzutage erschwinglich sind, stellt sich die Frage, wie sinnvoll eine Speicherung von Sequenzrohdaten ist. Es gibt noch keinen allgemeingültigen Standard für Datenformate von Next Gen-Sequenzdaten, obwohl Initiativen wie das SRF-Projekt (<http://srf.sourceforge.net>) versuchen, einheitliche Formate

zu definieren und durchzusetzen. Auch gibt es keine Normen für analysierte Daten wie z.B. SNPs, InDels und Annotationen.

Die Sequenzierertechnologien und deren Leistungsfähigkeit, Kapazität und Anwendungsprotokolle entwickeln sich rasant weiter; damit werden die Sequenzierkosten weiter sinken. Daher kann es sogar preiswerter sein, Sequenzierprojekte zu wiederholen als Rohdaten vergangener Projekte zu speichern.

### Fazit

Das Ziel eines Projekts und bereits vorhandene Sequenzdaten bestimmen nicht nur die Auswahl der Sequenzierertechnologie oder Kombination zweier oder mehrerer Technologien, sondern sie diktiert auch den effizienten Einsatz der bioinformatischen Software.

Daher ist zusätzlich zur detaillierten Fachkenntnis von Stärken und Einschränkungen der unterschiedlichen Next Generation-Sequenzierertechnologien auch das Wissen um die zur Verfügung stehenden Analysetools der entscheidende kosten- und zeitsparende Faktor für die erfolgreiche Durchführung eines Projekts.



**Abb. 3:** Vergleich einer *de novo*-Assemblierung eines Bakteriengenoms, generiert mittels einer Technologie (Roche GS FLX, blau) bzw. mit kombinierten Technologien (Hybrid-Strategie mit Roche GS FLX und Illumina GA II, rot).

### Literatur

- [1] Chevreur, B., Wetter, T. and Suhai, S. (1999). Proceedings of the German Conference on Bioinformatics (GCB) 99, pp. 45-56.
- [2] D.R. Zerbino and E. Birney. Genome Research 18:821-829.
- [3] D. Hernandez, P. François, L. Farinelli, M. Osteras, and J. Schrenzel. Genome Research. 18:802-809, 2008.
- [4] Hao Lin, Zefeng Zhang, Michael Q Zhang, Bin Ma, Ming Li, Bioinformatics (6 August 2008), btn416.
- [5] Ruiqiang Li, Yingrui Li, Karsten Kristiansen and Jun Wang BIOINFORMATICS APPLICATIONS NOTE Vol. 24 no. 5 2008, pages 713-714

### Kontakt

GATC Biotech AG  
Dr. Kerstin A. Stangier  
k.stangier@gatc-biotech.com  
www.gatc-biotech.com