Moderne Methoden zur Rekonstruktion von Transkriptomen

Modern methods for transcriptome reconstruction

Rätsch, Gunnar; Bohnert, Regina

Friedrich-Miescher-Laboratorium für biologische Arbeitsgruppen in der Max-Planck-Gesellschaft, Tübingen

Korrespondierender Autor

E-Mail: gunnar.raetsch@tuebingen.mpg.de

Zusammenfassung

Die Entwicklung neuer Sequenziertechnologien mit hohem Durchsatz macht es möglich, die Gesamtheit an Transkripten von unter bestimmten Bedingungen exprimierten Genen zu messen. Um die bei diesen Verfahren entstehenden großen Datenmengen auszuwerten, werden akkurate und effiziente computergestützte Methoden benötigt. Unsere Arbeitsgruppe versucht, mithilfe modernster Algorithmen aus dem Bereich des „maschinellen Lernens" Transkriptomdaten zu analysieren, um beispielsweise den Zusammenhang zwischen genetischer Information und Erscheinungsbild eines Individuums zu verstehen.

Summary

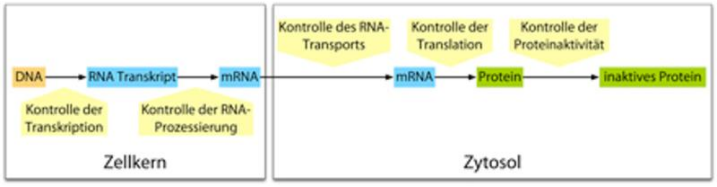
The development of novel high-throughput sequencing technologies allows the determination of the complete set of RNA-transcripts expressed under a given condition. Accurate and efficient computational methods are needed to uncover the full potential of the immense amount of data that is generated by these technologies. Our research group focuses on the analyses of transcriptome data using modern „Machine Learning" algorithms, providing a better insight into the relation of genetic information and phenotypic traits of individuals.

Expression und Verarbeitung von RNA-Transkripten

Durch die Entwicklung neuer Hochdurchsatzverfahren für die Nukleinsäure-Sequenzierung, mit deren Hilfe die Erbinformation (DNA) einer Zelle mittlerweile in kürzester Zeit entziffert werden kann, kam es zu einem enormen Anstieg verfügbarer Genomsequenzen in den Datenbanken. Heutzutage ist es möglich, die jeweiligen Genome von 1000 Menschen (siehe http://www.1000genomes.org) oder von 1001 jeweiligen Individuen einer Pflanzenart, beispielsweise der Ackerschmalwand (siehe http://www.1001genomes.org und [1]), zu rekonstruieren. Mithilfe der daraus hervorgehenden umfassenden DNA-Bibliotheken kann man zwar eine exakte Aussage darüber treffen, wie die DNA-Sequenz des einzelnen Individuums lautet, jedoch kann man nicht lesen und verstehen, welche genomischen Bereiche dafür verantwortlich sind, dass bestimmte, das Erscheinungsbild eines Organismus bedingende zelluläre Prozesse ablaufen. Deswegen ist es, neben der Bestimmung und Untersuchung der Genome, ebenso wichtig, die genomischen Bereiche zu identifizieren, die in einer bestimmten Zelle abgelesen werden, und zu erforschen, wie die genetische Information anschließend

prozessiert wird. Um dies zu erreichen, ermöglichen auch hier die neuen und schnellen Sequenziertechnologien eine Bestimmung des sogenannten Transkriptoms: Das Transkriptom ist die Gesamtheit derjenigen RNA-Moleküle, die als Vorlage zur Herstellung von Proteinen dienen oder direkt an der Regulation der Genexpression beteiligt sind.

Im Gegensatz zur statischen Genomsequenz hängt die Zusammensetzung des Transkriptoms vom Entwicklungsstadium, dem Zelltyp, äußeren Einflüssen und vielem mehr ab und ändert sich dynamisch während des gesamten Lebens einer Zelle (**Abb. 1**). Durch kontrollierte Variation in der Umgebung oder im Erbgut lassen sich Veränderungen im Transkriptom herbeiführen. Aus der Beobachtung dieser Veränderungen können wichtige Rückschlüsse auf die Arbeitsmechanismen der zu Grunde liegenden Prozesse, die bei der Genexpression ablaufen, gezogen werden. Beispielsweise kann die Übersetzung eines Gens gehemmt werden, damit anschließend beobachtet werden kann, wie sich die Expression anderer Gene verändert (siehe zum Beispiel [2, 3]). Mehrere solcher Experimente tragen dazu bei, Genregulationsnetzwerke zu beschreiben, was ein wichtiger Schritt für das Verständnis eines gesamten biologischen Systems ist.



Genexpression ist ein stark regulierter Prozess, der in verschiedenen Bereichen der Zelle stattfindet. Im Zellkern werden bestimmte Bereiche des Genoms als Vorlage benutzt, um RNA-Transkripte herzustellen. Nach deren weiterer Prozessierung werden sie als Boten-RNA (mRNA) aus dem Zellkern in das Zytosol transportiert, wo die Übersetzung in Proteine stattfindet, die wiederum zelluläre Prozesse beeinflussen und Aufgaben in der Zelle übernehmen können.

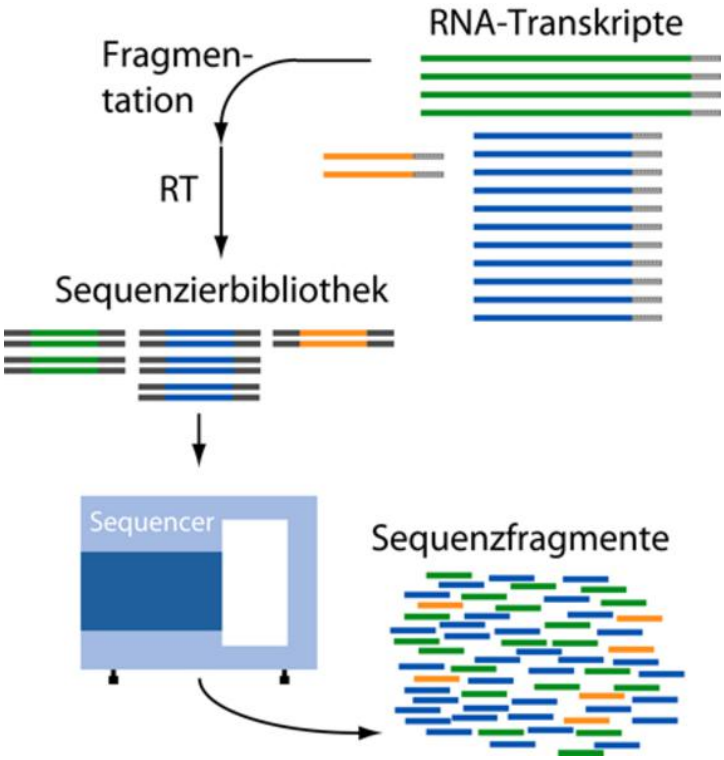
Durch die inzwischen nicht nur schnelle, sondern auch kostengünstig gewordene Sequenzierung großer Mengen von RNA-Transkripten lässt sich eine Überprüfung des Transkriptoms in beispielloser Auflösung realisieren. In der Arbeitsgruppe werden zu diesem Zweck computergestützte Methoden entwickelt. Ziel ist es, Transkriptome möglichst realitätsnah im Computer zu rekonstruieren und selbst kleine Veränderungen in der Konzentration oder Zusammensetzung von RNA-Transkripten messen zu können.

Hochdurchsatz-Sequenzierung von Transkriptomen

Die Fortschritte in der Entwicklung der Sequenziertechnologie haben den gesamten Bereich der Lebenswissenschaften nachhaltig beeinflusst. Dieser Trend hält mit der Entwicklung von neuen Hochdurchsatz-Sequenzierungsverfahren an und wird auch in Zukunft die Forschung in diesem Bereich revolutionieren und voranbringen.

Die Techniken zur Sequenzierung von Genomen (DNA) und Transkriptomen (RNA) sind sich sehr ähnlich. Zunächst werden die zu sequenzierenden DNA- oder RNA-Moleküle fragmentiert und entsprechend ihrer Länge gefiltert. Aus der Gesamtheit der so entstandenen Fragmente wird durch geeignete Modifikation und Vervielfältigung eine Sequenzierbibliothek erstellt (**Abb. 2**). Mithilfe von Sequenzierungsmaschinen, die auf dem *Sequencing-by-Synthesis*-Prinzip beruhen, können diese Bibliotheken innerhalb weniger Tage ausgelesen werden. Es entstehen beispielsweise beim Illumina Genome Analyzer II (Illumina Inc.) 100 bis 200 Millionen Sequenzfragmente mit einer Länge von 40 bis 200 Nukleotiden. Das entspricht fast dem zehnfachen Umfang

der menschlichen Genomsequenz - aber die Kosten dafür sind nur noch ein Zehntausendstel der Kosten, die das Sequenzieren des ersten menschlichen Genoms vor rund 10 Jahren benötigt hat. Aus dieser Fülle von Daten lassen sich sehr genaue Informationen über das untersuchte Genom oder Transkriptom ableiten.



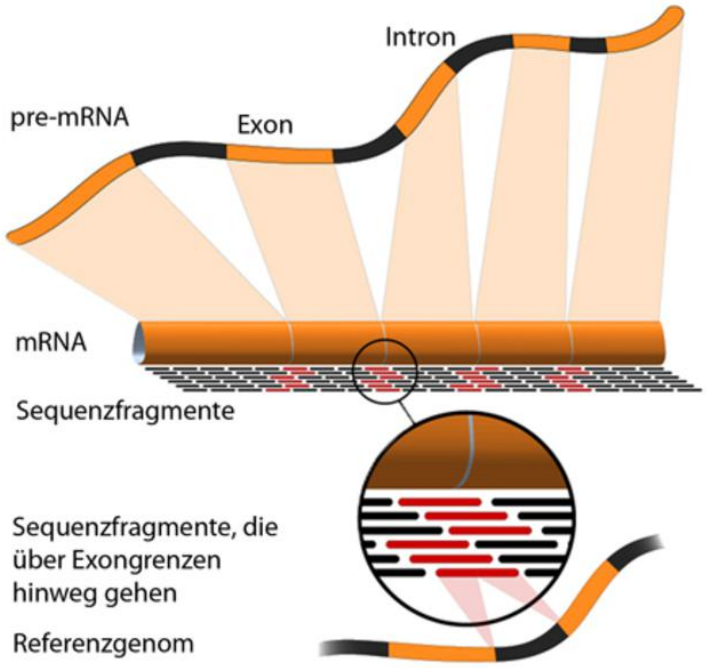
Die wesentlichen Schritte beim Sequenzieren eines Transkriptoms. Nach der Fragmentierung wird die RNA durch reverse Transkription (RT) in komplementäre DNA umgeschrieben und dann, vor der Sequenzierung, geeignet modifiziert. Das Ergebnis sind Millionen von Sequenzfragmenten, die den ursprünglichen Transkripten entsprechen.
© Friedrich-Miescher-Laboratorium/Bohnert

Die neuen Sequenziertechnologien haben aber im Moment trotz ihrer Vorzüge noch einige Nachteile. Es können zum Beispiel nicht vollständige RNA-Moleküle, sondern nur Fragmente sequenziert werden, und die Sequenzierfehlerrate liegt deutlich über der traditioneller Verfahren. Des Weiteren sind sowohl an der Vorbereitung als auch an der eigentlichen Sequenzierung biochemische Verfahren beteiligt, die die Konzentration von RNA-Fragmenten unerwünscht verändern können. Diese Eigenschaften erschweren die Verwertung der resultierenden Sequenzen für die Rekonstruktion und Quantifikation des Transkriptoms deutlich. Außerdem stellen die bei der Sequenzierung entstehenden Datenmengen eine sehr große Herausforderung für die nachfolgenden Analysen dar. Bisherige Methoden zur Verarbeitung der entstehenden Sequenzen kommen sehr leicht an ihre Grenzen, sowohl bezüglich der Genauigkeit als auch der Geschwindigkeit der Datenverarbeitung.

Maschinelles Lernen als Methode für die Verarbeitung von Transkriptomsequenzen

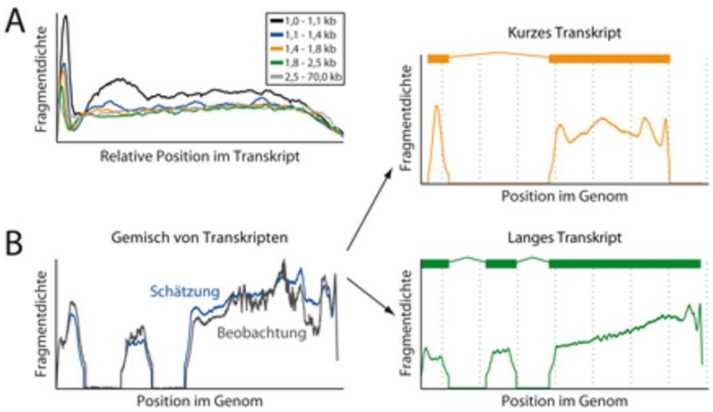
In der Arbeitsgruppe werden Genome und Transkriptome mithilfe des so genannten „maschinellen Lernens" analysiert. Dieses relativ junge Forschungsfeld vereint Methoden aus der künstlichen Intelligenz, der Statistik und der mathematischen Optimierung. Maschinelles Lernen beschäftigt sich mit der Analyse komplexer statistischer Phänomene, wie beispielsweise dem der Prozessierung von RNA-Transkripten in der Zelle. Dazu

werden empirische Beobachtungen, die sogenannten Lernstichproben, analysiert, um dann präzise Vorhersagen über das untersuchte Phänomen treffen zu können. Häufig werden hierzu die genauen und effizienten, kernbasierten Lernalgorithmen benutzt [4], die mittels einer sogenannten Kernfunktion leicht an das jeweilige Problem angepasst werden können. Diese Methoden wurden in der Forschungsgruppe dahingehend weiter entwickelt, dass sie nunmehr auch für die Analyse von Genom- und Transkriptomdaten geeignet sind. Dafür war es einerseits erforderlich, Kernfunktionen zu entwickeln, die aus Sequenzen diskriminative Informationen extrahieren können [5, 6], und andererseits, dass diese Methoden auch sehr große Datenmengen, effizient verarbeiten können [7]. Nur durch diese methodischen Entwicklungen war es möglich, das volle Potenzial der vorhandenen Daten für die Transkriptomanalyse auszuschöpfen und beispielsweise das Gen-Erkennungsprogramm *mGene* zu entwickeln [8], das die bisher genaueste Genomannotation des Fadenwurms *Caenorhabditis elegans* berechnen konnte.



Sequenzfragmente entstehen natürlicherweise auch aus RNA-Transkripten, aus denen in vorangegangenen Schritten die nicht kodierenden Bereiche (Introns) entfernt wurden. Dadurch wird das korrekte Wiederfinden der Sequenzfragmente im Referenzgenom erschwert. Dank neuer Alignmentstrategien kann dieses Problem gelöst werden.
© Wikipedia - http://en.wikipedia.org/wiki/RNA-Seq

Diese Methoden eignen sich ausgezeichnet für die Analyse von Sequenzdaten aus Transkriptomen, da sie die den Daten inhärenten Unsicherheiten, Fehler und Verzerrungen modellieren und ausgleichen können. Beispielsweise wurde eine neue Methode für das *Alignment*, das heißt für das Auffinden des Ursprungs von Sequenzfragmenten im Referenzgenom, entwickelt. Bei dieser Aufgabe ist es nämlich besonders schwierig, Sequenzfragmente, die sich über mehrere Exons erstrecken, im Genom wiederzufinden (**Abb. 3**). Durch die Kombination eines Alignmentalgorithmus mit computergestützten Vorhersagen der Position von Exon/Intron-Grenzen (Spleißstellen) und einem neuen Lernansatz [9] konnte die Alignmentfehlerrate von Sequenzfragmenten, die durch Spleißen hervorgegangen sind, von 14 % auf weniger als 1,8 % reduziert werden. In einem kürzlich vorgenommenen Vergleich mit anderen Alignmentstrategien (zum Beispiel *TopHat* [10]) wurde gezeigt, dass diese Methode mit Abstand am akkuratesten solche Alignments bestimmen kann. Diese hohe Genauigkeit wurde durch den Einsatz intelligenter Algorithmen erreicht, die sich ohne menschliches Zutun an die Eigenschaften der Daten anpassen können.



Die Fragmentdichte ist innerhalb eines Transkripts nicht gleichförmig verteilt und ist beispielsweise abhängig von der Länge des Transkripts (**A**). Besonders schwierig sind Fälle, bei denen in einem Bereich des Genoms mehrere alternative Transkripte abgelesen werden können (**B**; alternative Transkripte in diesem Beispiel in grün und orange dargestellt), aber in der Fragmentdichte nur als Summe zu beobachten sind. Der entwickelte Algorithmus verteilt die Gesamtfragmentdichte auf diese alternativen Transkripte in einer Weise, dass der Unterschied zwischen geschätzter und beobachteter Dichte möglichst gering ist.

Ein weiterer wichtiger Schritt bei der Untersuchung von Transkriptomdaten ist die Quantifizierung der untersuchten Transkripte, um mengenspezifische Unterschiede innerhalbs des Transkriptoms zu finden oder auch um verschiedene Transkriptome vergleichen zu können. Hier kann es, bedingt durch molekularbiologische Aufbereitungsschritte vor dem Sequenzieren, zu Abweichungen bei den eigentlichen Molekülkonzentrationen kommen. Um diese Verzerrungen bei der Quantifizierung zu berücksichtigen, wurde eine neue Methode, basierend auf einem Optimierungsansatz, entwickelt, die eine deutlich genauere Bestimmung der Konzentration von Mischungen gleichzeitig auftretender RNA-Transkripten erlaubt (**Abb. 4**).

Ausblick

Die in der Gruppe entwickelten Methoden bilden die Grundlage für zahlreiche Projekte, in denen Transkriptome und deren Unterschiede untersucht werden. In Zusammenarbeit mit Gruppen am Max-Planck-Institut für Entwicklungsbiologie, der Universität Tübingen, der Universität Gießen, der University of Utah und dem European Neuroscience Institute in Göttingen werden diese Methoden bereits angewandt, um Transkriptome von Menschen, Mäusen, Fischen, Würmern und Pflanzen zu untersuchen und zu vergleichen. Um auch anderen Forschern die Anwendung dieser Methoden zu ermöglichen, bietet das Friedrich-Miescher-Laboratorium Webdienste unter http://galaxy.tuebingen.mpg.de an. Mithilfe dieser Dienste lassen sich Genome annotieren, Sequenzfragmente alignieren und Transkriptkonzentrationen mithilfe von Transkriptomsequenzen bestimmen.

Die Anwendung der neuen Sequenziertechnologien und der neu entwickelten Methoden ermöglichen es, sehr genaue Abbilder von Transkriptomen im Computer zu erzeugen und deren Veränderung unter verschiedenen experimentellen Bedingungen festzustellen. Die genaue Kenntnis der Transkriptome erlaubt außerdem eine Anwendung von Lernverfahren, die an Hand der gemessenen Daten erlernen können, wie das Transkriptom aus dem Erbgut und anderen Faktoren gebildet wird. Es hat sich außerdem gezeigt, dass auch epigenetische Informationen und äußere Einflüsse das Transkriptom sehr stark beeinflussen. In der Arbeitsgruppe wird deswegen angestrebt, detailliertere Modelle zur genauen Vorhersage von Transkriptomen unter Zuhilfenahme solcher Informationen zu entwickeln. Diese Modelle können dann dazu benutzt werden, ein tieferes

Verständnis der Plastizität von Transkriptomen zu gewinnen und damit in Zukunft auch einige biologische Experimente *in silico* durchzuführen zu können.

Originalveröffentlichungen

[1] **D. Weigel:**

Die 1001 Genome der Ackerschmalwand.
Jahresbericht der Max-Planck-Gesellschaft, München (2009).

[2] **S. Laubinger, T. Sachsenberg, G. Zeller, W. Busch, J. Lohmann, G. Rätsch, D. Weigel:**

Dual roles of the nuclear cap binding complex and serrate in pre-mRNA splicing and microRNA processing in *Arabidopsis thaliana*.
Proceedings of the National Academy of Sciences USA 105, 8795–8800 (2008).

[3] **G. Zeller, S. R. Henz C. K. Widmer, T. Sachsenberg, G. Rätsch, D. Weigel, S. Laubinger:**

Stress-induced changes in the *Arabidopsis thaliana* transcriptome analyzed using whole genome tiling arrays.
The Plant Journal 58, 1068–1082 (2009).

[4] **B. Schölkopf:**

Statistische Lerntheorie und Empirische Inferenz.
Jahrbuch der Max-Planck-Gesellschaft, München (2004).

[5] **A. Ben-Hur, C. S. Ong, S. Sonnenburg, B. Schölkopf, G. Rätsch:**

Support vector machines and kernels for computational biology.
PLoS Computational Biology 4, e1000173 (2008).

[6] **S. Sonnenburg, G. Rätsch, B. Schölkopf:**

Large scale genomic sequence SVM classifiers.

In: S. Dzeroski (ed.) Proceedings of the 22nd International Conference on Machine Learning, ICML. ACM Press, New York (2005).

[7] **S. Sonnenburg, G. Rätsch, K. Rieck:**

Large scale learning with string kernels.
In: L. Bottou, O. Chapelle, D. DeCoste, J. Weston (eds.) Large Scale Kernel Machines, pp 73–103. MIT Press, Cambridge, MA (2007).

[8] **G. Schweikert, A. Zien, G. Zeller, J. Behr, C. Dieterich, C. S. Ong, P. Philips, F. De Bona, L. Hartmann, A. Bohlen, N. Krüger, S. Sonnenburg, G. Rätsch:**

mGene: accurate SVM-based gene finding with an application to nematode genomes.
Genome Research 19, 2133–2143 (2009).

[9] **F. De Bona, S. Ossowski, K. Schneeberger, G. Rätsch:**

Optimal spliced alignments of short sequence reads.
Bioinformatics 24, i174–180 (2008).

[10] **C. Trapnell, L. Pachter, S. L. Salzberg:**

TopHat: Discovering splice junctions with RNA-Seq.
Bioinformatics 25, 1105–1111 (2009).

