Intelligente Algorithmen zur Analyse zellulärer Spleißmechanismen

Intelligent Algorithms for the Analysis of Cellular Splicing Mechanisms

Rätsch, Gunnar

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

Korrespondierender Autor/in

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

Zusammenfassung

Neueste Technologien erlauben vielfältige Messungen an biologischen Systemen, die zu einer immer größer werdenden Datenmenge und -vielfalt führen. Das volle Potenzial lässt sich dabei nur durch gründliche Auswertung ausschöpfen. Neben der elektronischen Organisation stellt die effiziente, automatisierte Analyse dieser Beobachtungen eine große, konzeptionelle Herausforderung dar. Mithilfe moderner Techniken des *maschinellen Lernens* wird am Friedrich-Miescher-Laboratorium beispielsweise das komplexe Phänomen des zellulären Spleißens von Boten-RNA (mRNA) im Zellkern analysiert. Dabei ist man besonders an der Vorhersage alternativen Spleißens und dem damit verbundenen, tieferen Verständnis genetischer Regulationsmechanismen interessiert.

Summary

Novel technologies allow for many measurements on biological systems, leading to fast-growing amounts and variety of data. In order to tap the full potential of the available data a thorough analysis is demanded. Apart from the electronic data organisation, an efficient and automatic analysis is a great conceptual challenge. Using modern *Machine Learning Methods*, researchers at the Friedrich Miescher Laboratory are analysing for example the complex phenomenon of cellular messenger RNA splicing. Their particular interest is the prediction of alternative splicing and a deeper understanding of its regulation mechanisms.

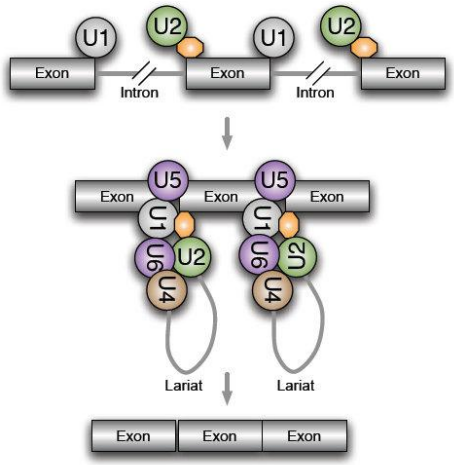
Die Verarbeitung der Boten-RNA

Spleißen ist ein wichtiger Schritt während der Verarbeitung von Ribonukleinsäure (RNA), der im Zellkern von Eukaryoten stattfindet. Beim Spleißen entsteht aus der durch Transkription entstandenen prä-mRNA, die so genannte Introns und Exons enthält, die reife mRNA. Dabei werden die Introns entfernt und die angrenzenden Exons miteinander verknüpft. Die so entstandene reife mRNA kann dann z.B. in Proteine übersetzt werden. Anfänglich ist man davon ausgegangen, dass das Spleißen stets zum gleichen Ergebnis führt, d.h. dass unabhängig von äußeren Einflüssen die gleiche mRNA entsteht und somit jedes Gen, das sich aus Exons und Introns zusammensetzt, am Ende nur ein und dasselbe Protein kodiert. Neuere Ergebnisse zeigen jedoch, dass Spleißen häufig von äußeren Faktoren abhängt und damit unterschiedlich ablaufen kann. Deshalb können

aus einem Gen viele verschiedene, modulare mRNA-Sequenzen und folglich auch verschiedene Proteine entstehen. Es wird geschätzt, dass mindestens 60% der menschlichen Gene alternativ gespleißt werden und dass es im Durchschnitt mehr als fünf Spleißformen pro Gen gibt. Diese Spleißformen haben zum Teil einen wesentlichen Einfluss auf die Entwicklung und das korrekte Ablaufen biochemischer Prozesse. So sind beispielsweise mehrere Krankheiten bekannt, die durch abnormales Spleißen verursacht werden. Ein genaueres Verständnis der zu Grunde liegenden Mechanismen ist daher auch von medizinischer Bedeutung.

Mechanismus des Spleißens

Das Spleißen findet in den meisten Fällen in einem großen Komplex aus RNA und Proteinen statt (**Abb. 1**). Dieser Komplex (Spleißosom) besteht aus mehreren Teilkomplexen und den snRNPs (small nuclear ribonucleoproteins), die wiederum aus verschiedenen Proteinen und snRNAs (small nuclear RNAs) zusammengesetzt sind. Die snRNPs erkennen die Grenzen zwischen Exons und Introns (Spleißstellen) und entfernen die Introns auf der prä-mRNA. Während des Spleißens interagieren so genannte spleißosomale Proteine mit den snRNPs oder regulativen Elementen auf der prä-mRNA und beeinflussen so die Erkennung der Spleißstellen abhängig vom Zustand der Zelle. Auf diese Weise kann unter anderem das Auslassen ganzer Exons in bestimmten Geweben reguliert werden, was zu drastischen Veränderungen der Funktionsweise des resultierenden Proteins führen kann.



Das Spleißen erfolgt in mehreren, genau abgestimmten Schritten: Als Erstes binden die snRNPs U1 und U2 unter Einfluss verschiedener zusätzlicher Proteine u.a. an die beiden Spleißstellen. Dieser prä-spleißosomale Komplex verbindet sich danach mit dem U4-U5-U6 snRNP-Komplex – dadurch bildet sich das so genannte Lariat. Innerhalb dieses Komplexes gibt es RNA-RNA- und RNA-Protein-Umordnungen, die zur Freigabe der snRNPs U1 und U4, zur Entfernung des Lariats und zur Verknüpfung der beiden Exons führen. Trotz der relativ genauen Kenntnis des biochemischen Ablaufs ist die genaue in silico Spleißstellenerkennung nur mit Methoden des maschinellen Lernens möglich. Bei jedem der genannten Schritte kann es zu Wechselwirkungen mit spleißosomalen Proteinen kommen, die an eventuell vorhandene regulative Elemente auf der prä-mRNA binden können. Dadurch kann es zu alternativem Spleißen kommen, das wir nun mit unseren Analysetechniken vorhersagen und genauer verstehen können.

Maschinelles Erlernen des alternativen Spleißens

In unserer Arbeit untersuchen wir alternatives Spleißen mithilfe moderner Techniken des „Maschinellen Lernens". Dieses relativ junge Forschungsfeld vereint Themen der Künstlichen Intelligenz, der Statistik und der mathematischen Optimierung. Es beschäftigt sich mit der Analyse komplexer statistischer Phänomene, wie eben z.B. dem des alternativen Spleißens. Dazu werden empirische Beobachtungen, die so genannte Lernstichprobe, analysiert, um dann präzise Vorhersagen über das Phänomen treffen zu können. Häufig werden hierzu die genauen und effizienten kernbasierten Lernalgorithmen benutzt [3], die mittels eines so genannten Kerns leicht an das jeweilige Problem angepasst werden können. Durch die Entwicklung so genannter String-Kerne [6] lassen sich diese Methoden auch zur Klassifikation und Annotation von DNA-Sequenzen und insbesondere zur Analyse von Spleißstellen benutzen.

Mit solchen Methoden analysieren wir in einem ersten Schritt die unmittelbaren Umgebungen der Spleißstellen, die regulative Elemente enthalten können und an die die spleißosomalen Proteine binden. Die in dieser Region vorhandene Information reicht unseren Analysetechniken bereits aus, um präzise vorherzusagen, ob alternatives Spleißen an der betrachteten Spleißstelle auftritt oder nicht. Mit Experimenten am Fadenwurm *Caenorhabditis elegans* konnten wir bestätigen, dass auf diese Art mehr als 50% aller im Genom auftretenden

Exonauslassungen effizient gefunden werden können [2]. Neuere Ergebnisse unserer Arbeit lassen erkennen, dass sich diese Techniken sowohl auf andere Formen alternativen Spleißens (Intronauslassung, alternative Exonenden etc.) als auch auf andere Organismen übertragen lassen, sobald eine genügend große Anzahl von Beobachtungen als Lernstichprobe vorliegt. Wir stellten fest, dass sich bei evolutionär nah verwandten Organismen die Vorhersagealgorithmen direkt und ohne substanziellen Genauigkeitsverlust auf andere Organismen übertragen lassen (z.B. von *C. elegans* auf *C. remanei*).

Neben der korrekten Vorhersage der verschiedenen Spleißformen streben wir das genauere Verständnis der Regulationsmechanismen des Spleißens an. Ein wichtiger Ansatz besteht darin, die komplexen Hintergründe der Vorhersage genauer zu untersuchen und auf einfache, für Fachexperten allgemein verständliche Informationen zu reduzieren. Dazu haben wir am Friedrich-Miescher-Laboratorium neue Techniken entwickelt, mit denen man aus den kernbasierten Lernmethoden verständliche Informationen für den Biologen extrahieren kann [1, 5]. Unter anderem kann damit die Position oder sogar die Länge von relevanten, degenerierten Sequenzmotiven relativ zu den Spleißstellen bestimmt werden. Die so bestimmten Regionen können dann mit anderen informatischen oder biochemischen Methoden genauer untersucht werden, um auf diese Weise einen tieferen Einblick in die vielschichtigen Spleißmechanismen zu erlangen [2].

Zusammengefasst, helfen unsere Lernalgorithmen die komplexen Mechanismen des Spleißens, aber auch die der Transkription [7] und Translation, nachzubilden und besser zu verstehen. Wir planen, schon bald die Entwicklung neuer Algorithmen zur *ab initio* Erkennung von Genen inklusive aller ihrer Spleißformen abzuschließen und sie auf neu sequenzierte Genome, wie beispielsweise auf das Genom von Pristionchus pacificus, einem Fadenwurm [5], anwenden zu können. Des Weiteren haben wir bereits begonnen, diese Techniken auch auf das menschliche Genom anzuwenden, um auch dort zum Verständnis der komplexen Regulationsmechanismen beizutragen.

Originalveröffentlichungen

[1] **Rätsch, G., S. Sonnenburg, and C. Schäfer:**

Learning interpretable SVMs for biological sequence classification.
BMC Bioinformatics **7**, S9 (2005).

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

RASE: Recognition of alternatively spliced exons in *C. elegans*.
Bioinfomatics **21**, i369–i377 (2005).

[3] **Schölkopf, B.:**

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

[4] **Sommer, R. J.:**

Der Fadenwurm *Pristionchus pacificus* als Modellsystem für die Erforschung evolutionärer Prinzipien im Tierreich.
Jahrbuch der Max-Planck-Gesellschaft (2005).

[5] **Sonnenburg, S., G. Rätsch, C. Schäfer, and B. Schölkopf:**

Large scale multiple kernel learning.
Journal of Machine Learning Research **7**, in press (2006).

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

Large scale genomic sequence SVM classifiers.

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

[7] **Sonnenburg, S., A. Zien, and G. Rätsch:**

ARTS: Accurate recognition of transcription starts.
Bioinformatics **22**, in press (2006).

