idw - Informationsdienst Wissenschaft
Nachrichten, Termine, Experten

Press release

Fraunhofer-Institut für Rechnerarchitektur und Softwaretechnik FIRST
Mirjam Kaplow M.A.

02/23/2007
http://idw-online.de/en/news197531

Research projects, Research results
Biology, Information technology, Medicine, Nutrition / healthcare / nursing
transregional, national

Vorhersagen für die Proteinfabrik

Die Gensuchmaschine mSplicer kann proteincodierende Bereiche auf den Genen des Fadenwurms C. elegans um 40 % exakter bestimmen als bisherige Verfahren

Noch ist es eine Vision: Aus den rund drei Milliarden Buchstaben des menschlichen Genoms auf Knopfdruck exakt diejenigen Abschnitte herauszufiltern, die für den Bau von Proteinen zuständig sind. Was für das menschliche Genom noch in der Zukunft liegt, ist Wissenschaftlern der Fraunhofer- und der Max-Planck-Gesellschaft für das Genom des Fadenwurms Caenorhabditis elegans nun gelungen: Sie können mit hoher Genauigkeit Exons und Introns, d. h. proteincodierende und nicht codierende Abschnitte erkennen. Die Ergebnisse des Kooperationsprojekts werden am 23. Februar 2007 in der Zeitschrift PLoS Computational Biology publiziert:
http://dx.doi.org/10.1371/journal.pcbi.0030020.eor

Der einen Millimeter lange Caenorhabditis elegans gehört zu den bestuntersuchten Organismen der Welt. Sein Genom ist seit 1998 vollständig sequenziert. Dennoch ist die Annotation des Genoms, d. h. die Lokalisierung seiner Gene und die Bestimmung der entsprechenden Proteine, bei weitem noch nicht vollständig. Sie wird fortlaufend überarbeitet und vervollständigt (www.wormbase.org). Ziel des Forschungsprojekts ist es, die bestehende, aber noch nicht komplett durch Experimente belegte Annotation des Fadenwurms zu verbessern. Dazu wählten die Forscher moderne Verfahren des maschinellen Lernens. Mit ihrer Hilfe sollten Exons und Introns in der genetischen Information des Fadenwurms identifiziert werden. Die Ergebnisse der Forschungsarbeiten zeigen, dass Verfahren des maschinellen Lernens um 40% exaktere Ergebnisse liefern als herkömmliche Methoden und insbesondere als die zur Zeit der Experimente gültige Annotation (Wormbase WS120). Verfahren des maschinellen Lernens können somit wesentlich zu einer Verbesserung bestehender Annotationen nicht nur bei C. elegans, sondern auch bei anderen Organismen beitragen und die korrekte Entschlüsselung genetischer Informationen erheblich beschleunigen.

Methode und Verfahren

Um ihre Ergebnisse zu belegen, gingen die Wissenschaftler in mehreren Schritten vor: Zunächst wurden die eingesetzten Algorithmen anhand bereits entschlüsselter mRNA-Sequenzen trainiert. mRNA-Moleküle (mRNA = Messenger-Ribonukleinsäure) transportieren die genetische Information der DNA und codieren die ihr entsprechenden Proteine. Während des Trainings lernen die Algorithmen die Muster für die Übersetzung von DNA in mRNA. Diese Muster helfen, die verschiedenen Teile der Gensequenz voneinander zu unterscheiden. Dabei spielt die Erkennung der Grenzen zwischen Exons und Introns, den sogenannten Spleißstellen, eine entscheidende Rolle.

Nach einer Trainingsphase wurden die Algorithmen zur Vorhersage von fertiger mRNA aus DNA eingesetzt und die Er-gebnisse mit bestehenden Datenbanken verglichen. Mit einer Genauigkeit von bis zu 95% konnte mSplicer alle Exons und Introns korrekt vorhersagen.

Auffällig war, dass die Ergebnisse nur in bis zu 50% mit der bestehenden Annotation des Genoms von C. elegans übereinstimmten. Eine Evaluation der Wormbase Annotation Version WS 120 mithilfe von später verfügbaren

Informationen (basierend auf Wormbase Version WS 150) bestätigte, dass WS 120 in 18% der untersuchten Fälle ungenau war, während von mSplicer nur 10-13% der Fälle nicht exakt übersetzt wurden. Darüber hinaus belegen biologische Laborex-perimente mit 20 Genen, bei denen WS 120 und mSplicer in hohem Maße voneinander abwichen, die Überlegenheit des algorithmischen Verfahrens. Es lieferte in 75% aller Fälle richtige Vorhersagen, während die bestehende Annotation in keinem der untersuchten Fälle korrekt war.

Auf Grundlage der Ergebnisse wurde eine neue Annotation von C. elegans entwickelt. Sie ist im WWW unter www.msplicer.org zum Download verfügbar.

In einem weiteren Schritt wurde mSplicer mit zwei weiteren State-of-the-art Verfahren zur Vorhersage von Exons und Introns verglichen: SNAP und ExonHunter. Diese Verfahren basieren auf sogenannten generativen Modellen, die versuchen, die Struktur der untersuchten Daten zu modellieren. mSplicer hingegen beruht auf diskriminativen Methoden: Der Algorithmus lernt "den Unterschied" zwischen richtigen und falschen Vorhersagen und unterscheidet sie anhand einer Trennfunktion. Je nach Auswahl der zugrundeliegenden Sequenzen erreichten SNAP und ExonHunter eine Genauigkeit bei der Vorhersage von Exons und Introns von nur 82,6 bzw. 90,2%. Die neu entwickelte Methode mSplicer kann eine Genauigkeit von 95.2% erzielen.

mSplicer wird seit 2003 im Rahmen eines Kooperationsprojekts zwischen der Fraunhofer- und der Max-Planck-Gesellschaft entwickelt. Der Schwerpunkt liegt auf einer engeren Verzahnung von Grundlagen- und angewandter Forschung.

Weitere Informationen erteilen Ihnen gern die zuständigen Projektleiter von Fraunhofer FIRST, Prof. Dr. Klaus-Robert Müller, vom Max-Planck-Institut für Biologische Kybernetik, Prof. Dr. Bernhard Schölkopf, und vom Friedrich-Miescher-Laboratorium, Dr. Gunnar Rätsch.

Pressekontakt:
Mirjam Kaplow, Leiterin Institutskommunikation Fraunhofer FIRST;
Tel.: 030/6392-1808; -1823
E-Mail: mirjam.kaplow@first.fraunhofer.de

Gunnar Rätsch, Leiter der Arbeitsgruppe "Machinelles Lernen in der Biologie"; Tel.: 07071/601 -820; -801
E-mail: Gunnar.Raetsch@tuebingen.mpg.de