Erkennung nichtkodierender RNAs mithilfe einer Support Vector Machine

Themenvorschlag für eine Bachelor-Arbeit

Felix Kühnl

14. Oktober 2016

Biologischer Hintergrund. Während man noch vor einigen Jahren annahm, ein Großteil der DNA wäre funktionslos, weil daraus keine Proteine synthetisiert würden, weiß man heute, dass dieser Teil des Genoms dennoch transkribiert, also in RNA übersetzt wird. Diese unter dem Begriff *nichtkodierende RNAs* zusammengefassten Moleküle sind u. a. für die Genregulation von großer Bedeutung. Im Gegensatz zu mRNA, die Proteine kodiert, ist für nichtkodierende RNAs häufig nicht die exakte Abfolge von Basen maßgeblich für die Funktion, sondern die sogenannte *Sekundärstruktur*. Damit bezeichnet man die Menge der intramolekularen Basenpaare, die die RNA nach der Transkription ausbildet. Dadurch ergibt sich eine zweidimensionale Skizze der räumlichen Struktur, mit der man häufig die Funktion der RNA erklären kann.

Mit der Entdeckung der Bedeutung der nichtkodierenden RNAs entstand auch der Bedarf nach effizienten Methoden, um genomweit nichtkodierende RNAs identifizieren zu können. Eine solche ist im Programm *RNAz* implementiert. Konkret berechnet das Tool für einzelne Fenster eines multiplen Sequenzalignments verschiedener Spezies einige statistische Werte aus den Sekundärstrukturen der Sequenzen. Anschließend wird auf Grundlage der so berechneten Werte mittels einer *Support Vector Machine* (*SVM*) entschieden, ob es sich bei dem momentanen Sequenzabschnitt um eine nichtkodierende RNA handelt.

Umfang der Arbeit. Die zur Klassifizierung verwendeten statistischen Werte bzw. Sekundärstrukturen werden mittels eines thermodynamischen Energiemodells berechnet. Das zugrunde liegende Modell wurde jedoch in der Zwischenzeit verbessert, sodass die momentan in RNAz verwendete Berechnungsgrundlage veraltet ist. Zwar steht eine aktualisierte Entwicklerversion von RNAz bereits

zur Verfügung, jedoch macht die Anpassung des Energiemodells ein erneutes Training der integrierten SVM erforderlich. Dies, ebenso wie die Zusammenstellung eines geeigneten Trainingsdatensatzes aus biologischen Datenbanken, soll im Rahmen der Bachelor-Arbeit erfolgen. Idealerweise sollte der Trainingsprozess mittels Skripten so implementiert werden, dass er in Zukunft weitestgehend automatisiert ablaufen kann, falls das zugrunde liegende Energiemodell erneut aktualisiert wird.

Erwartete Fähigkeiten. Für die erfolgreiche Bearbeitung sind u. a. folgende Kompetenzen erforderlich bzw. zu erwerben:

- sicherer Umgang mit der Programmiersprache *C* und der Linux-Shell *bash*
- Grundkenntnisse in einer Skriptsprache (*Perl*, *Python*, ...)
- Benutzung des Versionskontrollsystems *git*
- idealerweise Vorkenntnisse im Bereich maschinelles Lernen und SVMs
- idealerweise Hintergrundwissen in Genetik und Evolutionsbiologie
- das Verfassen der Arbeit in englischer Sprache und mit dem Textsatzsystem LaTeX ist wünschenswert

Natürlich ist die Aneignung fehlender Vorkenntnisse bei entsprechender Motivation auch im Laufe der Arbeit möglich.

Kontakt. Du hast Interesse an der Bearbeitung des o. g. Themas? Dann wende dich bitte an Felix Kühnl oder Prof. Peter F. Stadler.