Philipps Universität Marburg

Masterarbeit in Bioinformatik

Prof. Dr. Dominik Heider
dominik.heider@uni-marburg.de

Thema: Klassifikation und Sequenzvorhersage bei bakteriellen Replikationsursprünge (Machine Learning)

Der zentrale Schritt des Lebenszyklus bakterieller Zellen ist die Vervielfältigung ihrer genetischen Information. Essentiell dafür ist der Replikationsursprung, ein spezifischer Abschnitt auf der DNA, der als Signalprozessor fungiert: Hier werden Informationen über den Nahrungszustand der Zelle, ihrer genetischen Integrität und ihrer Umgebung gesammelt und daraus berechnet, ob mit der Replikation begonnen werden soll oder nicht. Das ist möglich, da der Replikationsursprung im Prinzip eine Ansammlung verschiedener, spezifischer Proteinbindestellen, sogenannte DNA-Motive, enthält, die es den betreffenden Proteinen ermöglichen, miteinander und mit der DNA zu interagieren [1].

In den letzten Jahrzehnten wurden einige für die korrekte Funktion des Replikationsursprungs notwendige Motive und Proteine identifiziert. Jedoch ist es bis heute nicht möglich, aus den bisher bekannten Bestandteilen einen funktionalen, bakteriellen Replikationsursprung herzustellen. Das deutet darauf hin, dass einige essentielle Motive weiterhin unbekannt sind. Neueste Fortschritte in der automatischen Motiverkennung unter Nutzung von Machine Learning-Ansätzen [2, 3], sollten diese Erkenntnislücke füllen können.

In diesem Projekt sollen Machine Learning-Ansätze zur Klassifikation von DNA-Sequenzen auf Motiv-Ebene und zur Motiverkennung genutzt werden und für die Gebrauch mit kleineren Datensätzen (~300 Sequenzen) angepasst werden. Dazu bieten sich Modelle wie Random Forests und Support Vector Machines an. Auf diese Weise sollen essentielle Bestandteile bakterieller Replikationsursprünge gefunden werden. Erfolgreiche Vorhersagemodelle sollen schließlich zu einem Web-basierten, öffentlichen Webtool weiterentwickelt werden.

Literatur

[1] Messer, W. (2002). The bacterial replication initiator DnaA. DnaA and oriC: The bacterial mode to initiate DNA replication. FEMS Microbiology Reviews, 26(4):355–374.

[2] Vidovic MM-C, Görnitz N, Müller K-R, Rätsch G, Kloft M (2015) SVM2Motif—Reconstructing Overlapping DNA Sequence Motifs by Mimicking an SVM Predictor. PLoS ONE 10(12): e0144782. https://doi.org/10.1371/journal.pone.0144782

[3] Ghandi M, Lee D, Mohammad-Noori M, Beer MA (2014) Enhanced Regulatory Sequence Prediction Using Gapped *k*-mer Features. PLoS Comput Biol 10(7): e1003711. https://doi.org/10.1371/journal.pcbi.1003711

Bitte melden Sie sich bei Interesse per Email bei: **20.04.17**
Prof. Dr. Dominik Heider - dominik.heider@uni-marburg.de