Philipps Universität Marburg

Bachelor- oder Masterarbeit in Bioinformatik

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

Thema: Graphische DNA-Darstellungen als Feature Encoding für tiefe neuronale Netze (Machine Learning, Bioinformatik)

Eine der wichtigsten Arbeitsschritte im Einsatz von Methoden des maschinellen Lernens ist die Auswahl eines feature encodings: Daten müssen so dargestellt werden, dass die in diesen enthaltenen Informationen für ein Modell gut auslesbar sind. Das gilt insbesondere in der Bioinformatik, die sich mit DNA- oder Proteinsequenzen beschäftigt. Die übliche Darstellungsform dieser Moleküle sind Strings von Buchstaben, die wiederum für einzelne Nukleotide bzw. Aminosäuren stehen. Jedoch gehen so z.B. Informationen zu den physikalisch-chemischen Eigenschaften und Interaktionen zwischen Nukeotiden verloren.

Seit der Entdeckung des genetischen Codes gab es einige Versuche, alternative graphische Darstellungsformen für DNA-Sequenzen zu etablieren, wie z.B. die Chaos Game Representation [1], Hilbert Curves [2], Wavelet-Analysen [3] und Random Walk Plots [4] . Diese konnten sich in der Vergangenheit zwar nicht gegen die sehr einfach handhabbare Stringdarstellung durchsetzen, jedoch liegen seit einigen Jahren mit z.B. *convolutional neural networks* (CNNs) kraftvolle Methoden zur Analyse und Klassifikation von Bilddaten vor. Durch diese technische Entwicklung werden graphische Darstellungsformen nun jedoch wieder relevant und verhelfen Klassifikationsmodellen zum Teil zu sehr guter Performanz.

In diesem Projekt sollen, nach einer Literaturrecherche, verschiedene visuelle DNA-Darstellungsformen implementiert werden. Auf der Basis von DNA-Benchmark-Datensätze sollen diese Darstellung dann daraufhin untersucht werden, wie gut sie sich zur Klassifikation der Sequenzen mit Hilfe von CNNs und anderen Klassifikationsmethoden (wie z.B. random forests) eignen.

Literatur

[1] Jeffrey, H. J. (1990). Chaos game representation of gene structure. Nucleic Acids Research, 18(8), 2163–2170. https://doi.org/10.1093/nar/18.8.2163

[2] Anders, S. (2009). Visualization of genomic data with the Hilbert curve. Bioinformatics, 25(10), 1231–1235. https://doi.org/10.1093/bioinformatics/btp152

[3] Haimovich, A. D., Byrne, B., Ramaswamy, R., & Welsh, W. J. (2006). Wavelet Analysis of DNA Walks. Journal of Computational Biology, 13(7), 1289–1298. https://doi.org/10.1089/cmb.2006.13.1289

[4] Leong, P. M., & Morgenthaler, S. (1995). Random walk and gap plots of DNA sequences. Bioinformatics, 11(5), 503–507. https://doi.org/10.1093/bioinformatics/11.5.503

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