Technische Berichte

Autor(en): Andre Dau; Johannes Krugel

Titel des Berichts: tt-analyze and tt-generate: Tools to Analyze and Generate Sequences with Trained Statistical Properties

Abstract: Algorithms working on sequences are influenced by the statistical properties of the sequences. Algorithms for fragment assembly for example usually produce a worse result if there are many repetitions. Also the space usage and running time of many data structures and algorithms depend on the statistical properties of the underlying text. We implemented tt-analyze, a tool to analyze sequences for certain statistical properties, among others the entropy, the number and distribution of different substrings, and the repeat structure. Besides, we also designed and implemented tt-generate, a tool to generate synthetic sequences with certain predefined properties, using models such as a Markov process, a discrete autoregressive process, and a repeat model. In bioinformatics these models have primarily been used to analyze given sequences, whereas here, we use them to also generate synthetic ones. The respective parameters of the models can be defined manually or be learned from given training data. The combination of both tools allows to generate sequences that are similar to real world sequences with respect to certain properties. This will allow to investigate the performance of algorithms under to some extent realistic, yet controlled conditions, and to determine the degree of dependence from parameters of the underlying sequence. Both tools have an extensible design which allows the integration of new modules for other statistical properties or generating models with the same programming interface.

Stichworte: models; genome analysis; efficient sequence analysis; efficient algorithms; machine learning; Markov process

Jahr: 2011
Sprache: en

Occurences:
- Einrichtungen > Fakultäten > Fakultät für Informatik > Technische Berichte > 2011

entries: