Abstract

This thesis presents innovative computational methods for analyzing count data from two disciplines: molecular biology and sociophysics. In the context of molecular biology, two novel probabilistic graphical models are introduced, enabling cell phenotyping in high-throughput sequencing data. Statistical algorithms based on Markov chain Monte Carlo sampling are proposed to estimate the models' parameters. In the field of sociophysics, a new agent-based model is derived, allowing for the approximation of citation indicators. The parameters of the agent-based model are sought using simple optimization procedures.

The first contribution is a hierarchical Bayesian model called Celloscope. This model utilizes knowledge about marker genes to decompose mixtures of cell types in spatial transcriptomics data. Comprehensive analysis of simulated data indicated that Celloscope outperforms other approaches in this task. Importantly, the model effectively identified known brain structures in mice, particularly demonstrating its ability to distinguish between inhibitory and excitatory neurons.

The second model described and derived in this study, ST-Assign, is used for simultaneous cell phenotyping based on single-cell RNA sequencing data and the decomposition of cell-type mixtures in spatial transcriptomics data. Its effectiveness was demonstrated on simulated data, as well as by integrating data from the mouse forebrain produced in two different experiments.

Finally, the problem of "negative citations" was analyzed with the use of data from the *Stack Exchange* platform. The developed agent-based model allows for approximating citation scores and drawing more universal conclusions about issues such as the evaluation of scientific achievements. The obtained results show the negligible impact of introducing "negative citations" on the most popular citation indicator, namely the Hirsch index. Furthermore, it is revealed that "negative citations" indicate areas of interest within the scientific community.

Overall, this thesis contributes novel tools for analyzing complex biological and social data, providing insights that can enhance our understanding of phenomena and patterns within these domains.

Keywords: probabilistic graphical model, transcriptomics, MCMC sampling, statistical modeling, science of science

Streszczenie

Niniejsza praca wprowadza nowatorskie statystyczne metody obliczeniowe służące do analizy danych zliczeniowych w dwóch obszarach badawczych: biologii molekularnej i socjofizyce. W kontekście biologii molekularnej przedstawiono dwa innowacyjne modele probabilistyczne, które umożliwiają fenotypowanie komórek na podstawie danych z sekwencjonowania o wysokiej przepustowości. W celu estymacji parametrów modelu, zaproponowano algorytmy statystyczne oparte na próbkowaniu Monte Carlo łańcuchami Markowa. Natomiast, w dziedzinie socjofizyki, wyprowadzono nowy model agentowy, umożliwiający przybliżanie wskaźników cytowalności. Parametry modelu agentowego szukane są przy użyciu prostych procedur optymalizacyjnych.

Pierwszą wprowadzoną metodą jest hierarchiczny model Bayesowski o nazwie Celloscope. Model ten wykorzystuje wiedzę na temat genów markerowych do dekompozycji mieszanek typów komórek w danych transkryptomiki przestrzennej. Dogłębna analiza w oparciu o dane symulowane wykazała, że Celloscope osiąga w tym zadaniu lepsze wyniki niż inne modele. Co ważne, Celloscope skutecznie zidentyfikował znane struktury mózgu myszy, w szczególności potrafił różnicować pomiędzy neuronami hamującymi a pobudzającymi.

Drugi opisany i wyprowadzony w pracy model, ST-Assign, służy do jednoczesnego fenotypowania pojedynczych komórek na podstawie danych sekwencjonowania RNA oraz dekompozycji mieszanek typów komórek w danych transkryptomiki przestrzennej. Wykazano jego skuteczność na danych symulowanych, a także integrując dane z przedniej części mózgu myszy powstałe w efekcie dwóch różnych eksperymentów.

Przeprowadzono także analizę problemu "negatywnych cytowań" na podstawie danych z serwisu *Stack Exchange*. Opracowany model agentowy pozwala na przybliżanie wskaźników cytowalności oraz na wysnucie uniwersalnych wniosków na temat zagadnień, takich jak ocena osiągnięć naukowych. Uzyskane wyniki pokazują minimalny wpływ wprowadzenia "negatywnych cytowań" na najbardziej popularny wskaźnik cytowalności, czyli indeks Hirscha. Ponadto okazuje się, że "negatywne cytowania" wskazują obszary zainteresowania społeczności naukowej.

Niniejsza rozprawa dostarcza nowe narzędzia do analizy danych biologicznych i społecznych, dając szansę na dogłębne zrozumienie zjawisk i wzorców w tych obszarach badawczych.

Słowa kluczowe: model probabilistyczny, transkryptomika, próbkowanie MCMC, modelowanie statystyczne, naukometria