

Применение методов машинного обучения для анализа генетического полиморфизма популяций

Марьяновская Т.А.^{1*}, Щербаков Д.Ю.^{1, 2}

¹ Новосибирский государственный университет, Новосибирск, Россия

² Лимнологический институт СО РАН, Иркутск, Россия

* t.maryanovskaya@g.nsu.ru

Ключевые слова: микроэволюция; популяционная генетика; искусственные нейронные сети

Мотивация и цель: Организмы эволюционируют в составе сообществ разных видов, и для изучения микроэволюции было разработано много моделей в рамках дисциплины популяционной генетики. Применение методов глубокого обучения становится все более актуальным в биологии из-за большого объема накопленных генетических данных. Анализ микросателлитов представляет определенные трудности из-за их сложной структуры. Это обусловлено переменной длиной микросателлитных последовательностей и их повторяющейся природой. Изучение и анализ микросателлитов требуют специальных методов и подходов для интерпретации полученных данных в генетических исследованиях.

Применение глубокого обучения в генетике популяций открывает новые возможности для анализа генетических данных и для анализа микроэволюционных процессов. Целью этой работы является исследование популяционной динамики у организмов со смешанной стратегией размножения с помощью методов глубокого обучения.

Методы и алгоритмы: В данном исследовании было использовано стохастическое моделирование. Спецификация модели представляет собой сверточную нейронную сеть с 6 внутренними и 2 внешними слоями. Обучение этой модели производилось на данных, полученных от модифицированной модели Райта–Фишера, включающей дополнительное бесполое размножение.

Результаты: В данной работе была разработана модель, представляющая одну из первых попыток применения глубокого обучения для задач популяционной генетики. Модель осуществляет анализ экспериментальных данных и прогнозирует соотношение полового и бесполого размножения у организмов, способных изменять свою стратегию размножения, таких как гидры, некоторые ракообразные, моллюски и другие. Проведены вычислительные эксперименты, подтверждающие устойчивость модели к шуму в экспериментальных данных, обусловленному возможными ошибками в процессе их получения, а также к ограничениям выборки. Кроме того, модель продемонстрировала высокую точность и скорость вычислений по сравнению с существующими математическими моделями.

Выводы: Наш результат демонстрирует возможности, которые открываются для исследования микроэволюционных процессов и их влияния на разнообразие организмов в результате объединения компьютерных моделей популяционных процессов и модели нейронной сети. Этот подход будет полезным инструментом для дальнейших исследований в этой области.

Application of machine learning methods for analyzing genetic polymorphism in populations

Maryanovskaya T.A.^{1, 2*}, Sherbakov D.Yu.^{1, 2}

¹ Novosibirsk State University, Novosibirsk, Russia

² Limnological Institut, SB RAS, Irkutsk, Russia

* t.maryanovskaya@ng.su.ru

Key words: microevolution; population genetics; deep learning

Motivation and Aim: Organisms evolve within communities of different species, and many models have been developed within the discipline of population genetics to study microevolution. The application of deep learning methods is becoming increasingly relevant in biology due to the large volume of accumulated genetic data. Microsatellite analysis poses certain challenges due to their complex structure characterized by variable lengths and repetitive nature. Studying and analyzing microsatellites require special methods and approaches for interpreting the data obtained in genetic research. The application of deep learning in population genetics opens up new possibilities for genetic data analysis and for analyzing microevolutionary processes. The objective of this work is to investigate the population dynamics in organisms with mixed reproductive strategies using deep learning methods.

Methods and Algorithms: Stochastic modeling was utilized in this study. The model specification involves a convolutional neural network with 6 internal and 2 external layers. Training of this model was carried out on data obtained from a modified Wright-Fisher model, incorporating additional asexual reproduction.

Results: This work presents a model that represents one of the first attempts to apply deep learning for tasks in population genetics. The model analyzes experimental data and predicts the ratio of sexual to asexual reproduction in organisms capable of altering their reproductive strategy, such as hydras, some crustaceans, mollusks, and others. Computational experiments confirmed the model's robustness to noise in experimental data, stemming from potential errors in their acquisition, as well as to sampling constraints. Furthermore, the model demonstrated high accuracy and computational speed compared to existing mathematical models.

Conclusion: Our results demonstrate the opportunities offered for studying microevolutionary processes and their impact on organism diversity through the integration of computer models of population processes and neural network models. This approach serves as a valuable tool for further research in this field.