

Автоматизация анализа геномных данных и предиктивное моделирование в селекции растений с использованием интегрированной биоинформатической системы компании ООО “ПЛАСТИЛИН”

Е.И. Григорьева*, Д.К. Константинов, Н.А. Сапожников, А.Ж. Ревель-Муроз, Н.О. Ерофеева,
С.А. Брускин, К.В. Стрыгина, Д.О. Медведев

ООО “ПЛАСТИЛИН”, Москва, Россия

* grigoreva@plastilin.team

Современная геномика растений переживает этап экспоненциального роста объемов данных в связи с распространением технологий высокопроизводительного секвенирования. Эффективная работа с этими данными, их интеграция с фенотипической информацией и трансляция в практически значимые результаты представляет собой перспективу для развития селекционных программ. Для решения этих задач в компании ООО “ПЛАСТИЛИН” разработана биоинформатическая система для автоматизации рутинных вычислительных этапов и предиктивного моделирования.

Результаты: Ключевым элементом системы являются автоматизированные вычислительные пайплайны, обеспечивающие обработку первичных данных секвенирования, полученных с использованием различных технологий, включая полногеномное секвенирование (WGS), генотипирование по сайтам рестрикции (GBS/ddRAD) и таргетное SNP-генотипирование. Пайплайн выполняет полный цикл обработки сырых данных: оценку качества, фильтрацию, картирование на референсный геном, маркировку ПЦР-дубликатов и вызов генетических вариантов. Автоматизация этих процессов значительно сокращает время обработки данных и обеспечивает высокую воспроизводимость результатов, минимизируя ошибки, связанные с ручной обработкой. Важнейшим аналитическим блоком платформы является автоматизированный GWAS-пайплайн, предназначенный для поиска ассоциаций между генетическими маркерами и хозяйственно ценными признаками. В ходе анализа выполняются все необходимые шаги предобработки и непосредственно сам GWAS-анализ с выбором наиболее подходящей в каждом конкретном случае статистической модели, учитывающей структуру популяции и родственные связи между образцами. Результатом являются идентифицированные маркеры, статистически значимо связанные с изучаемыми признаками, такими, например, как урожайность, содержание белка или масла, устойчивость к болезням и абиотическим стрессам, параметры вегетационного периода и другие. Эти маркеры служат основой для разработки инструментов маркер-ориентированной и геномной селекции.

Центральным звеном системы является интегрированная база данных, которая аккумулирует и структурирует генотипические данные полученные в ходе собственных анализов, фенотипические описания образцов, собранные в различных условиях и в разные годы, информацию об условиях выращивания, а также постоянно пополняемую данными из мировых коллекций базу знаний об известных генах и маркерах, ассоциированных с признаками у различных культур. Такая интеграция данных важна для проведения комплексных анализов и построения точных предиктивных моделей.

Особая роль в представленной системе отводится модулям машинного обучения. ML-алгоритмы активно используются не только для вспомогательных процессов, но и для решения ключевых селекционных задач. На основе интегрированных генотипических и фенотипических данных разрабатываются модели геномной селекции, позволяющие предсказывать характеристики еще не испытанных в поле линий и гибридов. Разрабатываются модели для предсказания гетерозисного эффекта при скрещивании конкретных родительских линий, что позволяет оптимизировать подбор пар для гибридизации. Также ведется работа над моделями, учитывающими взаимодействие генотип-среда, что помогает подбирать или создавать сорта, адаптированные к конкретным регионам выращивания.

Выводы: Таким образом, разработанная биоинформатическая система представляет собой мощную, масштабируемую и гибкую инфраструктуру. Она обеспечивает автоматизированную обработку больших объемов данных и применение передовых методов анализа. Система успешно протестирована на ряде хозяйственно ценных культур, среди которых соя, горох, кукуруза, сахарная свёкла, подсолнечник, капуста и рис. Внедрение этой платформы позволяет существенно повысить эффективность и скорость создания новых сортов и гибридов растений с заданными характеристиками.