Опыт быстрого развертывания облачной вычислительной среды для обработки научных данных*

А.А. Московский[1], В.А. Миронов[1], А.Т. Брехов[1], М.Г. Хренова[1], И.В. Поляков[1], Д.А. Фирсов[1], А.В. Немухин[1], Д.В. Подгайный[2], О.И. Стрельцова[2]

МГУ имени М.В. Ломоносова, Химический факультет[1],
Объединенный Институт Ядерных Исследований, ЛИТ[2]

Во многих областях науки по мере накопления экспериментальных данных и результатов моделирования становятся все более актуальны проблемы их хранения, эффективной обработки и обмена данными в междисциплинарных исследованиях. Одним из примеров является изучение механизмов химических и фотофизических процессов в белковых системах. С развитием технологий сверхбыстрой рентгеновской спектроскопии становится возможным исследовать структуры промежуточных состояний и интермедиатов реакций с высоким временным разрешением. Полученные экспериментальные данные для каждой реакции являются четырехмерными: это серии трехмерных карт электронной плотности для различных интервалов во времени по ходу протекания химического процесса. Сравнение полученных данных с результатами моделирования помогает уточнить экспериментальные структуры и установить механизм процесса. В настоящий момент количество экспериментальных данных по сверхбыстрой рентгеновской спектроскопии и результатов их моделирования не очень велико и подобный анализ может быть проделан вручную. Однако ожидается значительный рост их количества в ближайшие годы, и актуальной становится проблема их автоматического анализа.

Одной из важных проблем, возникающих при накоплении экспериментальных и расчетных данных по электронной плотности для реакций, является проблема их хранения. Помимо того, что данные могут занимать большой объем, необходимо обеспечить надежность их хранения и постоянную их доступность. В то же время, желательно, чтобы способ хранения данных не сказывался на эффективности доступа к ним. Одним из перспективных способов решения данных проблем является использование облачных технологий. В описываемой архитектуре для хранения и обработки данных по электронным плотностям реакций, результаты экспериментов и моделирования хранятся в распределенном хранилище, построенном на базе обычных серверов и даже персональных компьютеров и виртуальных машин. Надежность и доступность обеспечивается частичной репликацией данных между узлами. Для ускорения обработки данных предполагается запускать программы анализа локально на тех узлах, которые хранят копии необходимых файлов. Управление операциями обеспечивается с помощью системы оркестрации Kubernetes.

За короткий срок была развернута территориально-распределенная информационная среда, позволяющая гибкую настройку политики доступа к вычислительным ресурсам и массиву данных. Важным аспектом работы среды является поддержка сервисов хранения достаточно большого объема данных, а также возможность проведения расчетов с использованием высокопроизводительных вычислительных средств, таких как кластеры. Мы полагаем, что в будущем среда будет востребована для решения широкого круга научных проблем. В первую очередь, это сопоставление результатов теоретических расчетов трехмерного распределения электронной плотности интермедиатов ферментативных реакций с экспериментальными данными рентгеноструктурного анализа. Также, с помощью данного сервиса можно будет проводить поиск схожих распределений электронной плотности в базе данных не только по их атрибутам, но и по сформированным индексам. Потенциальной областью применения являются также методические исследования – сравнение точности и эффективности методов теоретического исследования ферментативных процессов, что позволит пользователям выбирать оптимальный метод для изучения разных классов белковых систем.

* Работа выполнена при поддержке Российского Фонда Фундаментальных Исследований (проект №18-29-13006).