Функциональные характеристики BioUML

Область применения

Платформа BioUML предназначена для обработки и анализа данных в области биоинформатики, системной биологии, и биомедицины.

Краткое описание возможностей

BioUML (Biological Universal Modeling Language; https://www.biouml.org) — интегрированная система для анализа данных системной биологии, биоинформатики и биомедицины, которая позволяет осуществлять обработку широкого спектра качественных и количественных данных, сборку и аннотирование геномов, визуальное моделирование и построение иерархических биологических моделей.

Платформа BioUML интегрирована с репозиториями git, где пользователи могут хранить свои модели и другие данные. Платформа BioUML обладает расширенными возможностями для анализа данных и визуализации биомедицинских данных, в частности:

1) любые программы и ядра Jupyter можно подключить к ней с помощью технологии Docker;

2) интегрирована с Galaxy и Galaxy ToolShed;

3) обеспечивает двустороннюю интеграцию с языками программирования R и Python (Jupyter ноутбуки): сценарии могут выполняться на ее веб-страницах, а ее функции могут вызываться из скриптов;

4) с помощью плагинной архитектуры можно добавлять специализированные просмотрщики и редакторы. Например, таким образом интегрируются мощные браузеры генома, а также средства просмотра молекулярной трехмерной структуры;

5) поддерживает анализ данных с использованием сценариев (собственный формат, Galaxy, CWL, nextFlow). Так же, как и родительская система, Платформа BioUML поддерживает новую ветвь развития аналогичных систем — u-science — универсальную научную платформу, которую можно настроить под конкретные исследовательские требования.

BioUML использует СУБД MySQL Community Server версии 5.7 и выше и разработана с использованием инструментальных средств языка программирования Java на основе дистрибутива виртуальной машины GraalVM 21 версии и более поздних. При этом может быть использован Java Development Kit версий 8 или 11, а также OpenJDK.

Для самостоятельной сборки и установки на компьютер пользователя необходимо использовать https://gitverse.ru/biouml/BioUML/.

Если BioUML планируется разворачивать на сервере для использования несколькими и более пользователями, то рекомендуется ориентироваться на данные ниже:

- Процессор уровня Intel(R) Xeon(R) CPU E5-2660 и выше с количеством ядер 32 и более;
- Количество оперативной памяти 64 GB и более;
- 2 TB дискового пространства на быстрых носителях для Docker образов и контейнеров;
- 4 TB дискового пространства для репозитория пользовательских проектов из расчета 100 активных проектов. При наличии большего числа проектов дисковое пространство должно быть пропорционально увеличено.

Задачи, функции и назначение BioUML

Ниже приведен перевод части текста из статьи автора разработки и его группы (Kolpakov F, Akberdin I, Kashapov T, Kiselev L, Kolmykov S, Kondrakhin Y, Kutumova E, Mandrik N, Pintus S, Ryabova A, Sharipov R, Yevshin I, Kel A. BioUML: an integrated environment for systems biology and collaborative analysis of biomedical data. Nucleic Acids Res. 2019 Jul 2;47(W1):W225-W233; https://pmc.ncbi.nlm.nih.gov/articles/pmid/31131402/), в которой дано описание возможностей BioUML.

BioUML поддерживает работу с данными основных мировых стандартов, используемые в системной биологии:

- **SBML:** Systems Biology Markup Language (13) служит для формального описания математических моделей. BioUML поддерживает все версии SBML, от l1v2 до последней l3v2, включая пакеты расширений 'fbc' (14) и 'comp' (15).
- **SBGN:** Системная биологическая графическая нотация (6) используется для визуального описания элементов модели (комплексов, отсеков, типов молекул, реакций и т. д.). BioUML полностью поддерживает диаграммы описания процессов SBGN и использует их для визуального представления моделей SBML. BioUML также поддерживает язык разметки XML SBGN-ML (https://github.com/sbgn/sbgn/wiki/SBGN_ML), который облегчает обмен диаграммами SBGN между инструментами.
- **Antimony**: Формат текста, понятный человеку, который поддерживает большинство функций SBML (16). В BioUML он автоматически преобразуется в диаграммы SBML в нотации SBGN. BioUML поддерживает импорт и экспорт в формат antimony.

- **SedML**: язык описания разметки экспериментов моделирования (17) описывает шаги моделирования модели и облегчает воспроизводимость экспериментов моделирования. В BioUML он транслируется в рабочие процессы, что позволяет анализировать и моделировать математические модели и данные биоинформатики.
- Однако многие модели требуют некоторых функций, которые отсутствуют в вышеупомянутых стандартах. В этих случаях стандарт SBML предоставляет механизмы расширения через элементы XML <notes> и <annotation>. Используя эти расширения, BioUML хранит всю дополнительную информацию о моделях (например, атрибуты представления диаграммы и макет).
- SBGN был разработан независимо от SBML, поэтому он не определяет визуальные синтаксисы для событий, функций, назначений и других математических элементов. Чтобы решить эту проблему, мы расширили диаграммы процессов SBGN дополнительными глифами для представления и использования их в наших собственных нотациях. Подробную информацию о типах моделей и их визуальных представлениях можно найти по адресу http://wiki.biouml.org/index.php/Diagram_type .
- **Механизм моделирования**: BioUML автоматически генерирует программный код, который используется для моделирования поведения анализируемой модели. В настоящее время BioUML генерирует высокооптимизированный код Java и использует собственные современные механизмы моделирования. Для каждой диаграммы он предоставляет список доступных механизмов. Например, сеть реакций может быть смоделирована как система ОДУ или как стохастическая модель типа Джиллеспи. Выбранный механизм моделирования предоставляет список доступных решателей. Доступные решатели ОДУ включают JVODE, который представляет собой пакет CVODE, перенесенный с C на Java и разработанный в Ливерморской национальной лаборатории имени Лоуренса (18). Он использует многошаговый метод Адамса-Моултона и обратный дифференциальный алгоритм, решатель RADAU5 (19), а также классические алгоритмы (Эйлера, Дорманда-Принса (20)). Механизм стохастического моделирования предоставляет точные методы Джиллеспи (21) и Гибсона-Брука (22), а также методы приближения.

Преобразование диаграммы в имитируемое состояние выбранным движком симуляции является предпосылкой для симуляции. Таким образом, иерархическая диаграмма может быть преобразована в обычную «плоскую» диаграмму с реакциями и сущностями.

Диаграмма на основе агентов может быть частично сглажена, где все поддиаграммы одного типа могут быть преобразованы в один объединенный агент.

Есть несколько других более простых препроцессоров. Например, ограничения SBML преобразуются в дискретные события, тем самым останавливая моделирование при нарушении ограничения. Кроме того, быстрые реакции преобразуются в алгебраические уравнения, а булевы выражения преобразуются в числовые выражения и т. д.

Другие движки моделирования:

- Гемодинамика: специально разработана для решения задач PDE, описывающих кровоток в артериях.
- Популяция: решает задачи NLME с использованием библиотеки R.
- Динамический FBA: динамически запускает анализ баланса потоков одновременно с моделированием ODE.

Модульное моделирование

При модульном подходе исследуемая система рассматривается как набор взаимосвязанных подсистем. Каждая подсистема может рассматриваться и моделироваться независимо. Интеграция этих моделей (или модулей) приводит к более сложной модели всей системы. Модули могут использовать различные математические формализмы и шкалы. Их можно создавать, проверять и улучшать независимо, и их можно рассматривать как заменяемые части. Модули предоставляют явные интерфейсы, через которые их можно подключать, не раскрывая их внутреннюю структуру пользователю. Мы рассматриваем модули как математические модели; их интерфейсами являются переменные и постоянные параметры. Например, значение переменной в одной модели может быть постоянным, в то время как в другой модели оно динамически изменяется. Численные вычисления выполняются двумя способами:

i. Уплощение: модульная модель может быть преобразована в немодульную модель путем агрегирования всех элементов всех модулей с автоматическим разрешением установленных связей между переменными (23).
ii. Агентное моделирование: Каждый модуль моделируется независимо с помощью собственного симулятора и формализма. Планировщик координирует их взаимодействие, отправляя и получая числовые значения подключенных переменных (24).

Оценка параметров

BioUML предоставляет несколько стохастических и детерминированных методов глобальной оптимизации (25), включая стратегию стохастической ранжирующей эволюции (26), оптимизацию роя частиц (27), клеточные генетические алгоритмы (28)

и другие. Мы добились значительного ускорения этих методов с помощью параллельных вычислений. Алгоритмы могут использовать экспериментальные данные в форме течения времени или стационарного состояния с точными или относительными значениями. BioUML также поддерживает многоэкспериментальную оценку параметров. Подробное сравнение с другим программным обеспечением можно найти в (25).

Анализ модели

Мы реализовали ряд методов анализа и редукции моделей, в том числе:

- Анализ идентифицируемости позволяет сделать вывод о том, насколько хорошо параметры модели аппроксимируются на основе количества и качества экспериментальных данных (29 , 30).
- Поиск линейных, мономолекулярных и псевдомономолекулярных реакций (31).
- Анализ квазиустойчивого состояния (32).
- Анализ чувствительности устойчивого состояния модели (33).
- Анализ метаболического контроля количественно определяет, как потоки и концентрации видов зависят от параметров системы (34).
- Стехиометрический анализ выводит линейные зависимости между скоростями потоков и производными концентрации реагентов (31).
- Анализ сохранения массы разлагает стехиометрическую матрицу на произведение ее линейно независимых строк и матрицы связей (35).

АНАЛИЗ БИОМЕДИЦИНСКИХ ДАННЫХ

Для обработки и анализа омикс-данных и других биомедицинских данных мы интегрировали лучшие платформы в соответствующих областях — R/Bioconductor (36) и Galaxy — в платформу BioUML и разработали более 300 собственных методов анализа (http://wiki.biouml.org/index.php/Category:Analyses).

- **Интеграция с R.** BioUML имеет двунаправленную интеграцию с R. Скрипты R можно использовать в BioUML четырьмя способами: (i) Пользователь может создавать, редактировать и выполнять скрипты R в панели документов BioUML. Редактор поддерживает подсветку синтаксиса; (ii) Панель «Скрипт» позволяет пользователю вводить и выполнять команды R; (iii) Скрипты R могут быть строительными блоками рабочего процесса BioUML; и (iv) Существует ряд инструментов анализа Java, которые предоставляют удобный интерфейс для настройки параметров анализа с последующей генерацией соответствующего скрипта R. Для выполнения скрипта R сервер BioUML вызывает R. Текстовый вывод

отображается на вкладке «Вывод». Графические результаты (графики, дендрограммы и т. д.) отображаются на отдельных страницах.

Чтобы получить доступ к серверу BioUML изнутри R, мы разработали пакет rbiouml (https://cran.r-project.org/package=rbiouml). Пакет содержит функции для получения данных из репозитория BioUML, импорта/экспорта данных, запуска анализов и рабочих процессов, а также управления очередью выполнения.

- **JavaScript API** . Пользователь может использовать JavaScript (документ, консоль, строительный блок в рабочем процессе) аналогично скриптам R. API предоставляет функции для получения данных из репозитория BioUML, импорта/экспорта данных, запуска инструментов анализа и рабочих процессов, а также обеспечивает подробный доступ к сложным объектам BioUML (например, моделям). В отличие от скриптов R, JavaScript выполняется внутри сервера BioUML.
- **Интеграция с Galaxy** . Платформа Galaxy предоставляет явные описания (файл XML инструмента Galaxy) параметров для тысяч биологических инструментов, в основном инструментов командной строки. BioUML расширяет синтаксис конфигурации инструмента Galaxy, что обеспечивает более тесное взаимодействие между системами Galaxy и BioUML (http://wiki.biouml.org/index.php/Creating_Galaxy_tool).
- BioUML может читать эти XML-файлы и генерировать формы, в которых пользователь может указывать значения для соответствующих параметров инструментов, интегрированных в Galaxy.
- **Рабочие процессы** . Для воспроизводимых исследований аналитические инструменты могут быть объединены в рабочие процессы. BioUML предоставляет мощный редактор для визуального построения рабочих процессов, а движок для выполнения рабочих процессов находится на сервере или в облаке.

Рабочие процессы BioUML могут включать следующие типы компонентов:

- Метод анализа: Метод для анализа с указанными входами/выходами и параметрами. Это может быть метод BioUML, инструмент Galaxy или Java-оболочка для функций R.
- Скрипт анализа: скрипт R или код JavaScript, методы R.
- Параметр анализа: Подмножество параметров, которые пользователь должен указать для запуска рабочего процесса.
- Выражение анализа: используется для установки и соединения входных и выходных параметров анализа в рабочем процессе.

- Цикл: Подмножество шагов рабочего процесса, которые будут выполняться повторно. Циклы могут выполнять итерации по элементам папки, по столбцам таблицы, по диапазонам целых чисел и по массивам элементов. Подробнее см. http://wiki.biouml.org/index.php/Workflow .

ВИЗУАЛИЗАЦИЯ ПУТИ

Редактор/просмотрщик диаграмм BioUML может использоваться не только для визуального моделирования, но и для визуализации различных биологических путей. Для этой цели сервер BioUML содержит следующие базы данных: Reactome (37), PantherDB (38) и Biomodels (https://www.ebi.ac.uk/biomodels/). Можно загружать собственные пути в следующих форматах: BioPAX, Antimony, SBGN-ML, SBML и Cytoscape CX (39).

BioUML использует несколько алгоритмов для автоматической компоновки визуальных диаграмм, включая иерархическую, направленную силу, жадную и сетевую компоновки (40).

Данные экспериментов по «омике» (транскриптомике, протеомике, метаболомике) можно сопоставить с различными биологическими путями и визуализировать, выделив соответствующие узлы на диаграмме (http://wiki.biouml.org/index.php/Expression_mapping).

Интегрированный браузер генома

BioUML предоставляет полностью интегрированный браузер генома (41), который поддерживает большинство функций, доступных в других современных браузерах генома, и включает в себя полный набор инструментов визуализации результатов обработки данных, который широко используется для визуализации информации из базы данных GTRD (42).

Совместные воспроизводимые исследования

Пользовательские данные (таблицы, диаграммы и т. д.) в BioUML организованы в проекты. Администратор (создатель) проекта может приглашать других пользователей к участию в проекте и управлять их разрешениями. Регистрация пользователей и управление правами доступа осуществляется через центральную систему аутентификации и авторизации (https://bio-store.org). Все действия пользователей в проекте, включая выполненные анализы и скрипты, отслеживаются в журнале проекта.

BioUML предоставляет функциональность совместного редактирования. Числовые модели, пути и рабочие процессы могут одновременно изменяться несколькими исследователями, а изменения мгновенно отображаются на экранах всех пользователей, а встроенная

функция чата облегчает координацию и совместную работу пользователей. Система также поддерживает контроль версий и возможность возврата к предыдущим версиям.

СЛУЧАИ ИСПОЛЬЗОВАНИЯ

От виртуальной ячейки к виртуальному пациенту

Видение BioUML заключается в предоставлении вычислительной платформы для создания виртуальных клеток, виртуальных физиологических людей и виртуальных пациентов. Мы создали две базы данных на сервере BioUML, которые демонстрируют нашу работу в этом направлении с использованием платформы BioUML.

База данных **виртуальных ячеек** включает три проекта:

i. Модульная модель апоптоза (23) является наиболее подробной современной моделью апоптоза. Модель разделена на 13 модулей, которые включают 280 видов (белки, их комплексы, модификации, такие как различные формы одной и той же молекулы, и преобразования, например фосфорилирование) и 372 реакции, применяющие действие масс, а также кинетику Михаэлиса-Ментен с 459 параметрами.

ii. Сигнальные пути CD95 и NF-κB (43): При идентификации параметров на основе экспериментальных данных для линий клеток человека мы столкнулись с проблемой переобучения модели. Для решения этой проблемы мы использовали технику редукции модели, которая позволила нам получить валидный набор параметров, подкрепленный достаточным количеством экспериментальных данных для модулей, связанных с сигнальными путями CD95 и NF-κB.

iii. Комплексная модель клетки *Mycoplasma genitalium* (44): Модель состоит из 28 подмоделей, использующих различные математические формализмы (ODE, стохастический, FBA). Первоначально эта модель была реализована в MATLAB. В 2016 году несколько исследовательских групп попытались воссоздать эту модель с использованием стандартов SBML и SBGN (45). Полностью завершены были только две подмодели — цитокинез и полимеризация FtsZ.

База данных **Virtual human** включает в себя ряд модульных моделей, описывающих физиологию человека, в том числе классическую модель кровообращения (46), модель работы сердца и кровотока (47), комплексную модель кровотока по 55 крупнейшим артериям человеческого тела (48) и модели, ориентированные на регуляцию объема крови (включая почки) (49 , 50).

Антигипертензивные препараты : это база данных фармакокинетических (ФК) и фармакодинамических (ФД) моделей антигипертензивных препаратов из разных групп, включая алискирен, лозартан, амлодипин, эналаприл, бисопролол и гидрохлоротиазид.

Комплексная модель : Эта база данных объединяет физиологические модели с моделями PK/PD для создания так называемых «виртуальных пациентов». Они могут быть созданы в различных формах с использованием различных частей человеческих физиологических моделей, с различными фокусами на подсистемах в зависимости от целей исследования.

Виртуальная мышца (51): Это подробная кинетическая модель, описывающая как облегченный, так и пассивный транспорт метаболитов между мышечными тканями и кровеносными сосудами, а также метаболические процессы в клеточных компартментах (цитозоле и митохондриях). Мы перестроили эту модель как модульную модель, которая стала примером многоуровневой модели, принимая во внимание клеточные компартменты и организацию тканей.

База данных ГТРД

База данных GTRD демонстрирует, как платформа BioUML может быть использована для создания веб-интерфейса для доступа к базе данных. Мы разработали специальную перспективу GTRD (42 , 52), которая обеспечивает просмотр, отображение информации, расширенные возможности поиска и интеграцию браузера генома и информации из базы данных Ensembl (структуры генов, повторы и т. д.) для визуализации данных GTRD.

Сценарии (workflow) как кулинарная книга для анализа данных омики

Каждый сценарий можно рассматривать как готовый рецепт для конкретного анализа соответствующих омиксных данных. Ученому нужно только импортировать данные, выбрать соответствующий рецепт, указать входные/выходные данные и нажать кнопку «Запустить». Платформа автоматически проанализирует данные. Это была ключевая идея платформы geneXplain (http://genexplain.com/genexplain-platform/) (53), которая теперь предоставляет сотни рабочих процессов для анализа различных типов омиксных данных (микрочипы, транскриптомика, протеомика, метаболомика и т. д.). Платформа geneXpain является ветвью дерева BioUML с фокусом на коммерческое применение. Она включает в себя такие коммерческие базы данных, как TRANSFAC ® (факторы транскрипции и их сайты связывания в геноме; 54), TRANSPATH ® (сеть передачи сигнала в эукариотических клетках; 55) и HumanPSD ® (биомаркеры заболеваний, лекарства и клинические испытания; 56). Платформа geneXplain содержит несколько собственных сложных методов анализа промоторов и путей, таких как Match™ (57) для идентификации участков связывания факторов транскрипции, CMA (Composite Module Analysis; 58) для

идентификации составных регуляторных модулей в промоторах и энхансерах, инструменты для поиска главных регуляторов (59) в сетях и другие инструменты.