***Н.А. Балашенко**, м.н.с. ГНУ «Институт генетики и цитологии НАН Беларуси», магистрант ГУО «Институт подготовки научных кадров НАН Беларуси»*

Информационные технологии в генетике

В статье описывается и группируется по направлениям опыт использования информационных технологий в генетике. В настоящее время в генетике в рамках реализации крупномасштабных проектов получено колоссальное количество данных, требующих обработки. Единственный способ обработки этих данных – привлечение средств информационных технологий.

На сегодняшний день развитие генетических исследований невозможно представить без использования информационных технологий, которые задействованы практически на каждом этапе изучения генетического кода, начиная от поиска литературы, заканчивая анализом огромных объемов информации. Каждый год появляются новые специализированные программные продукты, облегчающие работу ученых, а также предоставляющие в их распоряжение новый инструментарий, что способствует появлению новых видов и технологий исследований. Информационные ресурсы помогают не только рационализировать работу ученых, они также применяются для решения амбициозных задач, которые показались бы научной фантастикой еще двадцать лет назад. Скорость получения данных в генетике на сегодняшний день превышает скорость обработки этих данных. Количество и разнообразие информации, получаемой в результате исследований, делает информационные технологии незаменимыми в генетических исследованиях. Такой подход хорош тем, что компьютеры могут обрабатывать большие объемы данных о последовательности ДНК, генетической экспрессии, аминокислотной последовательности и т.д. Ниже будет рассмотрен ряд наиболее комплексных проблем в области исследований генетических основ клеточной трансформации, которые решаются самыми передовыми средствами информационных технологий.

Как уже было сказано выше, в сфере генетики существует перевес в сторону накопления данных. Это рождает потребность в осуществлении анализа данных большого объема. В англоязычной литературе даже появился термин «the Big Data», под которым обычно подразумеваются способы решения данной проблемы. Основной массив данных, требующий привлечения средств анализа данных большого объема, относится к геномике. Данные, полученные при полногеномном секвенировании, представляют собой набор последовательностей нуклеотидов разной длины – «обрывки генетического кода». Эти последовательности отражают участки генома, но не представляют собой последовательность целиком, как она существует в клетках (так как современные технологии позволяют считывать лишь участки нуклеотидных последовательностей и не позволяют считывать сразу всю последовательность). Полученные последовательности перекрываются, что дает возможность выстроить их в один ряд, однако количество этих перекрывающихся последовательностей таково, что обработка этих данных с помощью программных средств, устанавливаемых на персональный компьютер, весьма затруднительна. Для решения таких задач привлекаются средства облачных технологий, т.е. с помощью конфигурируемых вычислительных ресурсов, доступных через сеть Интернет.

Похожая проблема возникает и при осуществлении РНК-секвенирования, осуществляемого для изучения работы генома. Однако последовательности, которые необходимо восстановить, более короткие. Но после обработки этой информации возникает новая проблема. Работа генов осуществляется путем их взаимодействия между собой и с окружающей средой. В результате образуются сложные генные сети, которые очень сложно анализировать.

Облачные технологии при обработке и анализе последовательностей ДНК

Облачные технологии – это концепция сетевого доступа к устройствам для хранения информации, программным приложениям для ее обработки и иным средствам работы с информацией. В таблице 1 представлены примеры ресурсов, использующих данную концепцию.

Таблица 1

Примеры компаний и учреждений, которые предоставляют возможности для создания, хранения, анализа, визуализации экспериментальных и клинических данных

Компания\ Институт	**Предоставляемые возможности**	**Веб-сайт**
Appistry	Доступ к высокопроизводительной большой самоорганизующейся площадке для хранения и обработки клинической информации Appistry	www.appistry.com
BGI	Доступ к BGI – вычислительной платформе, представляющей собой интегрированный сервис, состоящий из универсального программного обеспечения, предназначенного для обработки крупных объемов данных в биоинформатике	www.genomics.cn/en
CLC Bio	CLC Bio имеет платформу, оптимизированную для лучшей производительности. CLC Bio используют собственные алгоритмы, основанные на опубликованных методах, успешно ускоряющих процесс обработки данных	www.clcbio.com
GNS Healthcare	GNS Healthcare разработала масштабный подход к решению проблемы больших объемов данных, которые могли бы быть применены по отрасли здравоохранения	www.gnshealthcare.com
Foundation Medicine	Foundation Medicine находится на переднем крае полногеномных исследований в области онкологии. Является первой платформой для разработки диагностических методов в онкологии, объединения клинические данные и большие аналитические возможности	www.foundationmedicine.com

Knome	Knome анализирует информацию о секвенировании генома с помощью программного обеспечения, позволяющего одновременно изучить и сравнить многие гены, генные сети, и геномы, а также интегрировать другие формы молекулярных и немолекулярных данных	www.knome.com
NextBio	Система передачи данных NextBio позволяет пользователям систематически интегрировать и интерпретировать полученные в других лабораториях и собственные молекулярногенетические данные о модельных организмах и клиническую информацию об отдельных пациентах	www.nextbio.com

Проблема построения и анализа генных сетей в генетических исследованиях и пути ее решения средствами современных информационных технологий

Для изучения взаимодействий генов было разработано множество методических подходов, таких как CRISPR и скрининг shRNA. Анализ данных, полученных в результате использования подходов CRISPR, представляет собой построение генных сетей, отражающих взаимодействия генов в клетке.

Для анализа данных о взаимодействии генов был разработан программный продукт NEST.

Веб-приложения и исходный код NEST находятся в свободном доступе по электронному адресу http://nest.dfci.harvard.edu.

Исходный код NEST также дополнительно доступен по адресу https://github.com/foreverdream2/NEST/releases.

Моделирование в генетических исследованиях

Основными подходами к моделированию клеточной трансформации являются моделирование динамики клеточной популяции, инициирования и прогрессии опухолей, методы построения филогенетических деревьев для моделирования отношений между клеточными субклонами и вероятностные графические модели для описания зависимостей между мутациями. Для анализа динамики популяций клеток может быть использована эволюционная теория, что позволит сделать вывод об этапах развития, например, опухоли в соответствии с молекулярными данными. Эволюционное моделирование помогает понять, как опухоли возникают, а также спрогнозировать ход развития заболевания и исход медицинских вмешательств.

Метод компьютерной видеомикроскопии живых клеток

Компьютерное микрофотографирование или видеомикроскопия (time-lapse microscopy) – это метод многократного фиксирования изображений микрообъектов, которые получают с помощью микроскопа в сочетании с фото- или видеокамерой, которые в свою очередь передают изображения на компьютер. С помощью специализированного программного обеспечения из набора фотографий создается видеоролик, отражающий процессы, происходящие в клеточной популяции. Компьютерная видеомикроскопия живых соматических клеток позволяет автоматизировать получение и анализ экспериментальных данных о динамике ряда процессов в клеточных популяциях с помощью специального программного обеспечения. Данный метод позволяет наблюдать процессы, происходящие на уровне микрообъектов. Этот экспериментальный подход позволяет изучать динамику различных клеточных событий, что является преимуществом по сравнению с экспериментами с использованием фиксированных цитологических препаратов. Компьютерная видеомикроскопия живых клеток находит все большее применение в разработке клеточных технологий. Выполняемый с помощью данного метода анализ индивидуальных клеток и их клоновых потомств во

времени позволяет изучать пролиферацию и клеточную гибель (апоптоз), а также эпигенетические процессы при онкотрансформации. Компьютерная видеомикроскопия также используется для изучения действия противоопухолевых препаратов на раковые клетки. В отличие от математического моделирования динамики клеточной популяции компьютерная видеомикроскопия позволяет наблюдать процесс напрямую, однако математическое моделирование важно для понимания особенностей динамики роста популяций онкоклеток, а также для прогнозирования в онкологии.

Для более детальной обработки изображений существует специализированное программное обеспечение – как закрытое, так и открытое. Наиболее распространенными пакетами бесплатного программного обеспечения являются «CellProfiler», «ImageJ», «Fiji», позволяющие измерять различные параметры клеток.

Базы данных в генетических исследованиях

Базы данных в рассматриваемой предметной области содержат данные, полученные при исследовании онкологии и онкогенетики. Информация может быть представлена в виде данных, прошедших первичную обработку, также некоторые ресурсы предлагают средства для онлайн-анализа данных и последующей загрузки результатов.

В качестве примера можно привести средство обработки информации MAGI для базы данных TCGA. Этот продукт позволяет осуществлять анализ генетических последовательностей, мутаций. Это веб-приложение с открытым исходным кодом. MAGI позволяет пользователям искать, визуализировать и комментировать большой набор данных, включая данные из проекта «The Cancer Genome Atlas». В дополнение MAGI также позволяет исследователям, загружать данные, полученные самостоятельно и сравнить результаты их обработки с уже имеющимися.

В таблице 2 приведен список баз данных, которые служат специально для исследований генетических основ онкотрансформации. Этот список не содержит баз данных, применяемых во многих предметных областях.

Таблица 2

Специализированные базы данных в области молекулярной генетики рака

Название базы данных	Организация-правообладатель	Вид данных с точки зрения онкогенетики	Объект исследования
The BioExpress® Oncology Suite	Ocimum Bio Solutions, США	Экспрессия генов	Клеточные линии и ткани *Homo sapiens*, *Mus musculus*, *Rattus norvegicus*
Oncomine	Compendia Bioscience, Inc., США	Экспрессия генов	Клеточные линии и ткани *Homo sapiens*
OncoLand	Omicsoft Corporation, США	Колличество копий генов, мутации, метилирование, фосфорилирование, экспрессия генов, микроРНК, белков	Клеточные линии и ткани *Homo sapiens*, *Rattus norvegicus*, *Mus musculus*
ClinicalTrials.gov	National Institutes of Health, США	Разное	Ткани *Homo sapiens*
Project Data Sphere	The CEO Life Sciences Consortium, США	Разное	Ткани *Homo sapiens*
Mouse Retrovirus Tagged Cancer Gene Database	Institute of Molecular and Cell Biology, Сингапур	Мутации	Ткани *Mus musculus*
International Cancer Genome Consortium	–	Мутации	Ткани *Homo sapiens*
Catalogue Of Somatic Mutations In Cancer (COSMIC)	Wellcome Trust Sanger Institute, Великобритания	Мутации	Ткани *Homo sapiens*
Network of Cancer Genes	King's College London, Великобритания	Мутации	Ткани *Homo sapiens*

Oncoreveal	Bopazisi University, Турция	Экспрессия генов	Ткани *Homo sapiens*
cBio Cancer Genomics Portal	Memorial Sloan-Kettering Cancer Center, США	Колличество копий генов, мутации, метилирование, фосфорилирование, экспрессия генов, микроРНК, белков	Ткани *Homo sapiens*
The Cancer Genome Atlas (TCGA)	National Cancer Institute, США	Колличество копий генов, мутации, метилирование, экспрессия генов, микроРНК, белков	Ткани *Homo sapiens*
Mouse Tumor Biology Database	The Jackson Laboratory, США	Колличество копий генов, мутации, экспрессия генов	Ткани *Mus musculus*
Integrative Oncogenomics Cancer Browser (IntOGen)	Universitat Pompeu Fabra, Испания	Колличество копий генов, мутации, экспрессия генов	Ткани *Homo sapiens*
OncoDB.HCC	Academia Sinica, Тайвань	Колличество копий генов, экспрессия генов	Ткани *Homo sapiens*, *Mus musculus*, *Rattus norvegicus*
Progenetix	Universitat Zurich, Швейцария	Колличество копий генов	Ткани *Homo sapiens*
CancerResource	University Medicine Berlin, Германия	Чувствительность к препаратам различных типов опухолей	Ткани *Homo sapiens*
Roche Cancer Genome Database (RCGDB)	Roche Diagnostics, Penzberg, Германия	Количество копий генов, экспрессия генов	Ткани *Homo sapiens*, *Mus musculus*, *Rattus norvegicus*

Информационно-поисковые системы в сфере генетики

Информационно-поисковые системы в области генетики связаны с базами данных, созданных при выполнении крупномасштабных международных проектов. Наличие этих поисковых систем облегчает работу исследователей и позволяет находить информацию о последовательностях РНК, ДНК, результатах клинических исследований, которую невозможно было бы получить любым другим способом. Большая часть этой информации находится в открытом доступе.

Сайты специализированных журналов

Существует ряд журналов, публикующих исследования в области генетики (таблица 3). Кроме сайтов самих журналов для поиска статей можно использовать онлайн базы данных, такие как http://www.sciencedirect.com. Одной из самых удобных систем по поиску статей в области генетики являются базы данных NCBI (The National Center for Biotechnology, электронный адрес – http://www.ncbi.nlm.nih.gov). Особенно удобна база данных PMC (http://www.ncbi.nlm.nih.gov/pmc), содержащая статьи, находящиеся в открытом доступе.

Таблица 3

Примеры специализированных журналов, публикующих результаты генетических исследований

Название журнала	Сайт журнала
Cancer genetics	http://www.cancergeneticsjournal.org/
American Journal of Human Genetics	http://www.cell.com/ajhg/home/
Cell	http://www.cell.com/
BMC Genetics	http://www.biomedcentral.com/bmcgenet/
BMC Medical Genetics	http://www.biomedcentral.com/bmcmedgenet/
Genetics	http://www.genetics.org/
International Journal of Molecular Epidemiology and Genetics	http://www.ijmeg.org/
Molecular Cytogenetics	http://www.molecularcytogenetics.org/
Clinical Epigenetics	http://www.clinicalepigeneticsjournal.com/
Nature	http://www.nature.com/index.html

Заключение

В последние годы произошел невероятный рывок в использовании информационных технологий в сфере генетических исследований. Благодаря быстрому развитию молекулярной генетики изменилось наше понимание процессов, происходящих в клетке. Это приблизило нас к решению многих стоящих перед современной генетикой проблем. В рамках крупномасштабных проектов получено колоссальное количество данных, требующих обработки. Единственный способ обработки этих данных – привлечение средств информационных технологий. Современная генетика немыслима без информационных технологий.

Литература

1. Квитко, О.В. Разработка методов компьютерной видеомикроскопии живых клеток для медицинской трансплантологии, биотехнологии животных и токсикологии / О.В. Квитко, И.И. Конева, Я.И. Шейко, В.Д. Трусова, С.Н. Шевцова, Н.А. Балашенко, А.С. Сапун, С.Е. Дромашко // Молекулярная и прикладная генетика. – 2009. – Т.10. – С.89-100.

2. Николайчик Е.А., Валентович Л.Н. SQ – компьютерная программа для редактирования и анализа биологических последовательностей // Труды Белорусского государственного университета. Физиологические, биохимические и молекулярные основы функционирования биосистем – 2010. – Вып. 5. ч.1. – С.154-162.

3. Beerenwinkel, N. Cancer Evolution: Mathematical Models and Computational Inferenc / N. Beerenwinkel, R. F. Schwarz, M. Gerstung, F. L. Markowetz // Syst. Biol. – 2015. – Vol.64. – P.e1-e25.

4. Costa, F.F. Big Data in Genomics: Challenges and Solutions / F.F. Costa // Laboratory Journal. – 2012. – Vol.11. – P.2-4.

5. Etani, N. Database application model and its service for drug discovery in Model-driven architecture / N. Etani // Journal of Big Data. – 2015. – Vol. 2. – P.1-17.

6. Jiang, P. Network analysis of gene essentiality in functional genomics experiments / P. Jiang, H. Wang, W. Li, C. Zang, B. Li // Genome Biol. – 2015. – Vol. 16. – P.1-10.

7. Herland, M. A review of data mining using big data in health informatics / M. Herland, T. M. Khoshgoftaar, R. Wald // Journal Of Big Data. – 2014. – Vol. 1. – P.1-35.

8. Kunkel, T.A. 2000. DNA replication fidelity / T.A.Kunkel, K.Bebenek // Annu. Rev. Biochem. – Vol. 69. – P.497-529.

9. Toga A.W. Sharing big biomedical data / A.W. Toga, I.D. Dinov // Journal of Big Data. – 2015. – Vol.2 – P.1-12.

10. Weiss, D.G. Videomicroscopy. Light microscopy in biology. A practical approach. / D.G. Weiss, W.Maile, R.A. Wick. – Oxford, 2007. – 32 p.

N.A. Balashenko

Information Technologies in Genetics

The article describes and groups on directions the experience of information technologies in genetics. Currently a huge amount of data received in genetics as part of large-scale projects, and it requires processing. The only way to handle these data is use of information technology.

Статья поступила 27.04.2016