Введение в биоинформатику

Программные средства и базы данных

Жулин Игорь Борисович и
центр биоинформатики ПСПбГМУ им. И.П.Павлова

Виктор Садовничий: ректор МГУ, президент Союза ректоров

Будущее инженерной науки будет связано с робототехникой, биоинженерией и биоинформатикой, а также физико-химическим конструированием

Челябинск, совете Союза ректоров, посвященном созданию новой концепции инженерного образования. 26.09.2014

Что такое биоинформатика?

Википедия (англ):

Биоинформатика — это междисциплинарная область исследований и разработки методов и программных средств для изучения биологических данных.

Зачем она нужна?

- Иногда по существу вопроса
- Иногда для экономии времени
- Иногда в дополнение

Какие узкие места были у биоинформатики при ее возникновении?

National Science Foundation (США):

Узкие места в биоинформатике:

- необходимость обучения биологов владению передовыми вычислительным средствами,
- Необходимость приглашения программистов в эту развивающуюся область знаний,
- ограниченная доступность баз данных биологической информации,
- потребность в более эффективных и интеллектуальных поисковых системах для этих баз данных.

Цели биоинформатики

- Конструирование и сопровождение биологических баз данных.
- Разработка программного обеспечения для анализа последовательностей, структур и функций.
- Применение или разработка подходов к пониманию биологических данных.

Зачем углубляться в изучение средств и баз данных по биоинформатике?

Важно иметь представление о базовых концепциях и алгоритмах в биоинформатике (исследования *in silico*),

также как необходимо понимать базовые основы и химический базис молекулярной биологии, генетики, биохимии если вы занимаетесь лабораторными экспериментами (исследования *in vitro*).

Что такое алгоритм?

Алгоритм – это набор инструкций, которые необходимо выполнить для того, чтобы решить задачу.

Что такое база данных?

- Коллекция взаимосвязанных элементов данных
  - *таблиц (например, генов, организмов, последовательностей и т. д.)*
  - *столбцов (полей)*
  - *строк (записей)*

Биологические базы данных

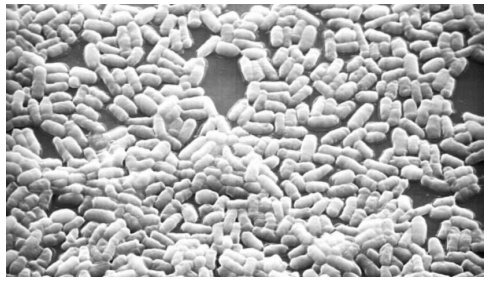
- Сколько их?
  - Разные по размеру, качеству, полноте, степени интереса
- Что представляет из себя «хорошая» база данных?

Базы данных

GenBank	www.ncbi.nlm.nih.gov	последовательности нуклеотидов
Ensembl	www.ensembl.org	человеческие, мышиные и др. геномы
PubMed	www.ncbi.nlm.nih.gov	научные публикации
NR	www.ncbi.nlm.nih.gov	белковые последовательности
SWISS-PROT	www.expasy.ch	белковые последовательности
InterPro	www.ebi.ac.uk	белковые домены
OMIM	www.ncbi.nlm.nih.gov	генетические заболевания
Enzymes	www.chem.qmul.ac.uk	ферменты
PDB	www.rcsb.org/pdb/	белковые структуры
KEGG	www.genome.ad.jp	метаболические пути

Специализированные базы данных

- **PomBase** компиляция данных, посвященных организму *Schizosaccharomyces pombe*



- **Wormpep** предсказанные белки по проекту секвенирования Caenorhabditis elegans (C. elegans).

- **Mistdb** база данных передачи сигналов в микробиологии

Вебсайт или веб-приложение?

- Статический – Интерактивный контент
- Чисто информационный – Программное средство

Веб-серверы и облачные вычисления

Преимущества:

- Удаленная инфраструктура, возможно, имеет большую вычислительную мощность чем ваша.
- Обновления данных и изменения функциональности происходят онлайн.

Недостатки:

- Вы используете чужой компьютер с ограниченными возможностями администрирования и программирования.
- Вы (вероятно) имеете урезанный выбор опций или мощности.
- Взаимодействие с внешними сетями может значительно замедлить выполнение задачи.

Серверы хороши для разовой проверки какой-то идеи, при критической необходимости в вычислительных сверхмощностях, а также в период интенсивно обновляемых разработчиком данных и функций.

Для регулярной ежедневной работы желательно получить и установить программное обеспечение локально

Советы для использования удаленных серверов в научных исследованиях

- Записывайте все:
  - имя и/или адрес сервера, названия баз данных и версии.
  - дату
  - идентификационные номера последательностей
  - параметры (настройки, умолчания, параметры запросов)
- Сохраняйте результаты
- Установите локально если знаете, что будете использовать в будущем

«Кейс с инструментами» специалиста-биоинформатика

Предсказание генов

- GRAIL (Xgrail, JavaGrail, etc.) *Gene Relationships Across Implicated Loci*
- Geneid
- Netgene
- GenMark
- Fexon, Hexon
- GENSCAN
- Xpound
- Genefinder

Выравнивание последовательностей

- Выравнивание двух последовательностей
- Одновременное выравнивание нескольких последовательностей

Выравнивание пар последовательностей

- **SIM** (только белки) – поиск k-лучших непересекающихся выравниваний (ExPASy, Швейцария)
- **ALIGN** – оптимальное глобальное *выравнивание без сокращений* (EERIE, Франция)
- **LALIGN** – вычисление N-лучших локальных выравниваний (EERIE)
- **LFASTA** – поиск локальных совпадений, демонстрирущих локальные выравнивания (EERIE)
- **BLAST 2** – локальное выравнивание с использованием BLAST (NCBI, США)
- **LAP2** – локальное выравнивание ДНК–белок LAP2 (MTU, США)

Взаимное выравнивание нескольких последовательностей

- ClustalW
- MAFFT
- T-Coffee
- MUSCLE

Поиск совпадений

- BLAST
- **blastp,** запрос — последовательность аминокислот к базе белковых последовательностей.
- **blastn** запрос — последовательность нуклеотидов к базе известных последовательностей нуклеотидов.
- **blastx** запрос — *последовательность нуклеотодов и продукты ее трансляции (обе нити) к базе белковых* последовательностей.
- **tblastn** запрос — последовательность аминокислот к базе последовательностей нуклеотидов, динамически транслируемых на всех рамках считывания (обеих нитях).
- **tblastx** запрос — все трансляции последовательности нуклеотидов к динамически вычисляемым трансляциям базы данных нуклеотидных последовательностей.

Предсказание структуры белка

- Ab initio («с нуля»): основанное на минимизации энергии
- Распознавание фолдинга: последовательность -> вторичная структура, затем выравнивание вторичных структур со вторичными структурами соответствующих белков, и т.д.
- Статистическое: основанное на «скрытых образцах (hidden patterns)»; схожие шаблонные образцы -> сходная структура.

Предсказание вторичной структуры белка

- **Coils** – предсказания суперспиральных регионов
- **nnPredict** – использует двухуровневую нейронную сеть
- **PSSP / SSP** – сегментно-ориентированные предсказания
- **PSSP / NNSSP** – предсказание методом ближайшего соседа
- **SAPS** – статистический анализ белковых последовательностей
- **Paircoil** – поедсказания суперспиральных регионы по корреляциям парных остатков
- **Protein Hydrophilicity /Hydrophobicity**
- **SOPM** – самооптимизирующийся метод предсказания

Предсказание функции белка

- Homology
- Sequence
- Structure
- Genomic context
- Co-expression

Филогенетический анализ

- Построение эволюционных деревьев на основе дивиргенций возникших в связянных последовательностях
  - Phyml
  - raxML
  - FastTree
  - Protdist and Neighbor-joining
  - …

- Как построить бактериальное дерево жизни?
- Как лучше всего построить дерево жизни, включающее человека, шимпанзе, лошадь и крысу?

Предсказание эффекта мутаций

- PolyPhen
- SIFT
- SNAP
- PROVEAN

…

Какая информация используется для решения о том, является мутация разрушительной или нет?

Take home messages

- Задачи биоинформатики очень похожи на лабораторные эксперименты:
  - Мы должны понимать концепции, используя программные средства и базы данных
- Это замечательно выполнять эксперименты в рамках компьютерной геномики, если вы четко понимаете что делаете.
- Разработка алгоритма (или подхода) требует опыта и знаний в соответствующих областях.
- Если вы в лаборатории компьютерной биологии, то ведение журналов – необходимый навык.

Вопрос

- Как мы используем избыточность в последовательностях?