

Поиск по сходству последовательностей

Нуклеотидный BLAST

Ваня Русинов

Парные выравнивания (повторение)

ОСНОВЫ

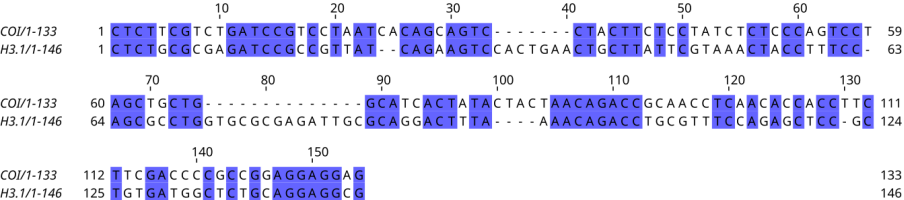
- ▶ Что такое парное выравнивание?
- ▶ ... эволюционное выравнивание?
- ▶ ... оптимальное парное выравнивание?
- ▶ Применяется ли неоптимальное парное выравнивание?
- ▶ Чем различаются глобальное и локальное парные выравнивания?
- ▶ Что нужно для построения парного выравнивания?

Параметры

- ▶ Что такое матрица весов замен?
- ▶ Сколько чисел в нуклеотидной матрице?
- ▶ Какие бывают штрафы за индели?

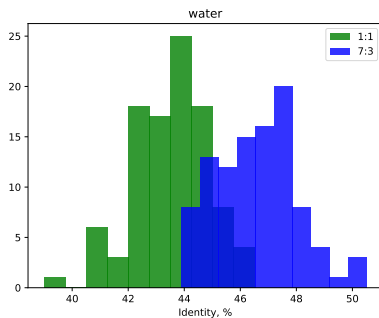
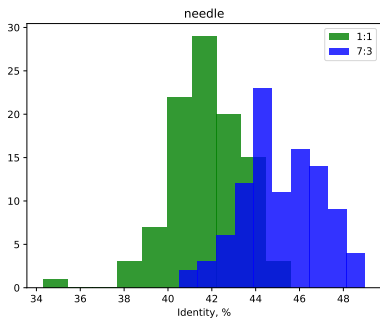
Identity (пример)

Какой процент identity можно считать свидетельством гомологии?



Identity

Парное выравнивание с параметрами по умолчанию на случайных последовательностях.



NCBI BLAST

Основы алгоритма

0. Подготовка базы (хеширование)
1. Разбивка query на "слова"
2. Поиск совпадений на одной диагонали и небольшом расстоянии
3. Построение выравнивания

Score (вес) и bit-score (вес в битах)

- ▶ Вес зависит от матрицы весов замен и штрафов за гэпы
- ▶ Bit-score – нормированный вес, не зависит от параметров вычисления веса
- ▶ По bit-score можно оценить случайность находки: если bit-score равен 30, то надо перебрать 2^{30} пар случайных последовательностей, чтобы получить одно выравнивание с таким или большим весом

$$S' = \frac{\lambda S - \ln(K)}{\ln(2)},$$

где S – вес выравнивания, λ и K – константы, зависящие от параметров вычисления веса

E-value

E-value – математическое ожидание количества находок с таким же или **большим весом в случайном банке того же размера**

(все важные слова выделены жирным)

$$E = mn \cdot 2^{-S'}$$

На самом деле, эта формула работает только для выравниваний без гэпов 😞

P-value

P-value – оценка вероятности получить хотя бы одно выравнивание с таким же или большим весом случайно

$$P = 1 - e^{-E}$$

Если $E < 0.01$, $P \approx E$

Виды BLAST

Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance. [Learn more](#)

NEWS

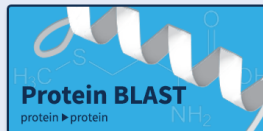
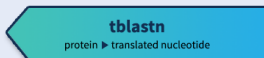
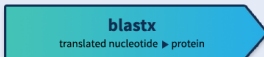
End of updates for BLAST+ version 4 databases (dbv4)

Start moving to the new version 5 databases!

Fri, 27 Sep 2019 16:00:00 EST

[More BLAST news...](#)

Web BLAST



BLAST Genomes

Search

[Human](#)

[Mouse](#)

[Rat](#)

[Microbes](#)

Разные алгоритмы blastn

megablast оптимизирован для поиска очень близких гомологов

blastn для поиска сколько-нибудь похожих последовательностей

discontiguous megablast хеширует длинные слова, но не требует полного совпадения

Разные алгоритмы blastn

Алгоритм	Длина слова по умолчанию	Возможные значения
megablast	28	16, 20, ..., 32, 48, 64, 128, 256
discontiguous megablast	11, разрывное слово (101101100101101101)	11 и 12, можно менять длину и тип шаблона
blastn	11	7, 11, 15

Белки или гены

Задача: найти гомологов для некоторой кодирующей нуклеотидной последовательности.

Белки или гены

Задача: найти гомологов для некоторой кодирующей нуклеотидной последовательности.

Кодирующие последовательности нужно транслировать перед поиском!

- ▶ белковые последовательности более консервативные
- ▶ можно учитывать биохимические свойства аминокислот
- ▶ автоматически учитывается рамка считывания

Задачи

- ▶ Какому виду принадлежит секвенированный фрагмент 16S РНК из метагенома?

Задачи

- ▶ Какому виду принадлежит секвенированный фрагмент 16S РНК из метагенома? **blastn**

Задачи

- ▶ Какому виду принадлежит секвенированный фрагмент 16S РНК из метагенома? **blastn**
- ▶ Как найти координаты секвенированного регуляторного участка для организма с известным геномом?

Задачи

- ▶ Какому виду принадлежит секвенированный фрагмент 16S РНК из метагенома? **blastn**
- ▶ Как найти координаты секвенированного регуляторного участка для организма с известным геномом? **megablast**

Задачи

- ▶ Какому виду принадлежит секвенированный фрагмент 16S РНК из метагенома? **blastn**
- ▶ Как найти координаты секвенированного регуляторного участка для организма с известным геномом? **megablast**
- ▶ Как узнать, какие полиморфизмы содержатся в секвенированном участке гена?

Задачи

- ▶ Какому виду принадлежит секвенированный фрагмент 16S РНК из метабенома? **blastn**
- ▶ Как найти координаты секвенированного регуляторного участка для организма с известным геномом? **megablast**
- ▶ Как узнать, какие полиморфизмы содержатся в секвенированном участке гена? **blastx**

Задачи

- ▶ Какому виду принадлежит секвенированный фрагмент 16S РНК из метабенома? **blastn**
- ▶ Как найти координаты секвенированного регуляторного участка для организма с известным геномом? **megablast**
- ▶ Как узнать, какие полиморфизмы содержатся в секвенированном участке гена? **blastx**
- ▶ Есть ли теломераза у новосеквенированного эукариота?

Задачи

- ▶ Какому виду принадлежит секвенированный фрагмент 16S РНК из метабенома? **blastn**
- ▶ Как найти координаты секвенированного регуляторного участка для организма с известным геномом? **megablast**
- ▶ Как узнать, какие полиморфизмы содержатся в секвенированном участке гена? **blastx**
- ▶ Есть ли теломераза у новосеквенированного эукариота? **tblastn**

Задачи

- ▶ Какому виду принадлежит секвенированный фрагмент 16S РНК из метагенома? **blastn**
- ▶ Как найти координаты секвенированного регуляторного участка для организма с известным геномом? **megablast**
- ▶ Как узнать, какие полиморфизмы содержатся в секвенированном участке гена? **blastx**
- ▶ Есть ли теломераза у новосеквенированного эукариота? **tblastn**
- ▶ Насколько похожи геномы двух сравнительно близких вирусов?

Задачи

- ▶ Какому виду принадлежит секвенированный фрагмент 16S РНК из метабенома? **blastn**
- ▶ Как найти координаты секвенированного регуляторного участка для организма с известным геномом? **megablast**
- ▶ Как узнать, какие полиморфизмы содержатся в секвенированном участке гена? **blastx**
- ▶ Есть ли теломераза у новосеквенированного эукариота? **tblastn**
- ▶ Насколько похожи геномы двух сравнительно близких вирусов? **tblastx**

Задачи

- ▶ Какому виду принадлежит секвенированный фрагмент 16S РНК из метабенома? **blastn**
- ▶ Как найти координаты секвенированного регуляторного участка для организма с известным геномом? **megablast**
- ▶ Как узнать, какие полиморфизмы содержатся в секвенированном участке гена? **blastx**
- ▶ Есть ли теломераза у новосеквенированного эукариота? **tblastn**
- ▶ Насколько похожи геномы двух сравнительно близких вирусов? **tblastx**
- ▶ Какие глобальные перестройки генома произошли после разделения двух родственных видов бактерий?

Задачи

- ▶ Какому виду принадлежит секвенированный фрагмент 16S РНК из метабенома? **blastn**
- ▶ Как найти координаты секвенированного регуляторного участка для организма с известным геномом? **megablast**
- ▶ Как узнать, какие полиморфизмы содержатся в секвенированном участке гена? **blastx**
- ▶ Есть ли теломераза у новосеквенированного эукариота? **tblastn**
- ▶ Насколько похожи геномы двух сравнительно близких вирусов? **tblastx**
- ▶ Какие глобальные перестройки генома произошли после разделения двух родственных видов бактерий? **blast2seq**

Веб-интерфейс BLAST

По ссылкам есть pdf с картинками и описанием всяких галочек, окошек, etc Их я советую хотя бы пролистать (там в сумме всего 12 страниц).

https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs

Getting Started

- [Guide to BLAST home and search pages](#)
- [Blast report description](#)
- [Blast topics](#)

About BLAST

- [Frequently Asked Questions](#)
- [NCBI Handbook: BLAST](#)
- [The Statistics of Sequence Similarity Scores](#)
- [BLAST glossary](#)
- [References](#)
- [Blast+ Command Line Applications User Manual](#)
- [BLAST News directory](#)