Лекция 3. Выравнивание биологических последовательностей (5 марта).

1.1 Введение, общая биология.

1.1.1 Клетка.

Биоинформатика — методология анализа результатов молекулярной биологии, а молекулы «живут» в клетках. Клетки можно классифицировать по типу организации:

- прокариоты:
  - бактерии,
  - археи;
- эукариоты:
  - животные клетки,
  - растительные клетки,
  - грибные клетки,
  - клетки протистов.

Так, например, бактериальная клетка окружена фосфолипидной мембраной (или двумя) и клеточной стенкой. В мембрану встроены различные белки. Цитоплазма — раствор малых, средних и крупных молекул в воде.

Клетка животного тоже окружена мембраной, а её цитоплазма содержит многочисленные органеллы, некоторые из которых имеют собственные мембраны.

1.1.2 Основные процессы молекулярной биологии.

- транскрипция — «переписывание» ДНК на РНК,
- трансляция — «перевод» РНК в белок,
- репликация — удвоение ДНК с сохранением информации,
- репарация — исправление ошибок в тексте, записанном на ДНК,
- регуляция — определение, каким генам экспрессироваться в данный момент.

1.1.3 Геном.

Определение. Геном — совокупность всего генетического материала, который находится в клетке организма и определяет его характеристики. Он содержит биологическую информацию, необходимую для построения и поддержания организма.

Определение. Ген — участок генома, транскрибирующийся в функциональную РНК.

Определение. Экспрессия гена — производство окончательного продукта (белка или РНК) на основе гена.

Молекулы ДНК — носители генома.

Химическая формула ДНК однозначно определяется последовательностью оснований: аденин (A), гуанин (G), тимин (T), цитозин (C).

ДНК представляется в виде двух цепей. Они соединены водородными связями между комплементарными основаниями, образуя двойную спираль.

Правило комплементарности — аденин (A) всегда соединяется с тимином (T), гуанин (G) всегда соединяется с цитозином (C) (в РНК вместо тимина — урацил (U)).

Принцип антипараллельности цепей — противоположная направленность двух нитей двойной спирали ДНК: одна нить имеет направление от $5'$ к $3'$, другая — от $3'$ к $5'$[1]. Последовательность ДНК записывается от $5'$ к $3'$.

Определение. Репликация ДНК — это процесс копирования ДНК перед делением клетки, при котором двойная спираль ДНК расплетается, и каждая цепь служит матрицей для синтеза новой комплементарной цепи. В результате образуются две идентичные молекулы ДНК, каждая из которых состоит из одной старой (материнской) и одной новой (дочерней) цепи.

1.1.4 Эволюционная биология.

Определение. Мутация (точечная) — точечное изменение в последовательности ДНК. Они бывают двух видов: точечные замены и инсерции / делеции (вставка / потеря участка ДНК, отличить инсерцию от делеции невозможно).

Определение. Математическое ожидание — среднее значение случайной величины. В дискретном (конечном) случае: пусть X — случайная величина, принимающая значения $x_1, x_2, \ldots, x_n$. Пусть также вероятности этих значений равны $p_1, p_2, \ldots, p_n$ соответственно; для любых i, $p_i \geq 0; p_1 + p_2 + \cdots + p_n = 1$. Тогда математиеское ожидание $\mathbb{E}(X)$ равно

$$\mathbb{E}(X) = \sum_{i=1}^{n} x_i p_i = x_1 p_1 + x_2 p_2 + \ldots x_n p_n.$$

Определение. Нейтральная мутация — мутация, не влияющая на матожидание числа потомков. Такие мутации не влияют на приспособленность организма, а их судьба определяется генетическим дрейфом, а не отбором.

Определение. Генетический дрейф — случайное колебание частоты нейтрального полиморфизма. Математическая модель такого процесса называется «случайным блужданием».

Каждому варианту генома можно сопоставить его «приспособленность» f = матожидание числа потомков организма с таким геномом (через какой-то фиксированный промежуток времени). В подавляющем большинстве случаев новая мутация порождает либо нейтральный вариант ($f = 1$), либо вредный ($f < 1$). Вероятность закрепиться для новой нейтральной мутации очень мала, но не равна нулю. Вредный вариант тоже начинает «блуждать», но вероятность «шага вверх» оказывается меньше вероятности «шага вниз». Это очень сильно уменьшает вероятность закрепления — тем сильнее, чем меньше f, и тем сильнее, чем больше популяция.

Определение. Стабилизирующий (отрицательный) отбор — явление невозможности закрепления вредной мутации.

[1] Обозначения $5'$ и $3'$ возникли из химической структуры дезоксирибозы и нумерации её углеродов и работ ранних биохимиков.

Если вдруг $f > 1$, то вероятность закрепления мутации вырастает во много раз.

Определение. Положительный отбор — процесс закрепления полезных мутаций.

Полезных мутаций так мало, потому что большинство возможных полезных мутаций уже закрепились. Обычно, полезные мутации начинают появляться в заметном количестве только при изменении условий жизни организмов. Так, например, при появлении нового источника пищи или новой опасности или попадании части популяции в другой климат.

Кроме точечных мутаций, бывают крупные перестройки генома. В частности, случаются дупликации генов. В большинстве случаев одна из копий накапливает мутации (нейтральные, поскольку есть вторая копия!) и превращается в нефункциональный псевдоген. Но изредка вторая копия приобретает новую функцию! Так возникают паралогичные гены. Примеры паралогов у человека: девять гемоглобинов, красный, зелёный и синий опсины. У человека подавляющее большинство генов имеет паралоги.

1.2 Выравнивание биологических последовательностей.

Рассмотрим последовательности миоглобинов человека, мыши и быка:

```
MYG_HUMAN MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASE 60
MYG_MOUSE MGLSDGEWQLVLNVWGKVEADLAGHGQEVLIGLFKTHPETLDKFDKFKNLKSEEDMKGSE 60
MYG_BOVIN MGLSDGEWQLVLNAWGKVEADVAGHGQEVLIRLFTGHPETLEKFDKFKHLKTEAEMKASE 60
          **************.*******: ******** **. *****:******:**:* :**.**

MYG_HUMAN DLKKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHKIPVKYLEFISECIIQVLQSKH 120
MYG_MOUSE DLKKHGCTVLTALGTILKKKGQHAAEIQPLAQSHATKHKIPVKYLEFISEIIIEVLKKRH 120
MYG_BOVIN DLKKHGNTVLTALGGILKKKGHHEAEVKHLAESHANKHKIPVKYLEFISDAIIHVLHAKH 120
          ****** ******* ******:* **:: **:***.*************: **.**: :*

MYG_HUMAN PGDFGADAQGAMNKALELFRKDMASNYKELGFQG 154
MYG_MOUSE SGDFGADAQGAMSKALELFRNDIAAKYKELGFQG 154
MYG_BOVIN PSDFGADAQAAMSKALELFRNDMAAQYKVLGFHG 154
           .*******.**.*******:*:*::** ***:*
```

Видно, что большинство букв совпадает, но некоторые различаются. Это последовательности гомологичных белков, что означает, что эти белки произошли от общего предка. За время, прошедшее от существования общего предка, некоторые буквы менялись, но большинство остались неизменными. Заметим также, что здесь последовательности имеют одну длину.

Определение. Выравнивание — метод сравнения последовательностей разной длины.

Рассмотрим теперь последовательности миоглобинов человека и рыбы: у них разная длина.

```
MYG_HUMAN MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASE 60
MYG_DANRE ----MADHDLVLKCWGAVEADYAANGGEVLNRLFKEYPDTLKLFPKFSGISQGD-LAGSP 55
              .: :***: ** ****  .:* *** **** :*:**: * **. :.. * : .*

MYG_HUMAN DLKKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHKIPVKYLEFISECIIQVLQSKH 120
MYG_DANRE AVAAHGATVLKKLGELLKAKGDHAALLKPLANTHANIHKVALNNFRLITEVLVKVMAEKA 115
           :  ******. ** :** **.* * :****::**. **: :: :.:*:* :::*: .*

MYG_HUMAN PGDFGADAQGAMNKALELFRKDMASNYKELGFQG 154
MYG_DANRE --GLDAAGQGALRRVMDAVIGDIDGYYKEIGFAG 147
            .:.* .***:.:.:: .  *: . ***:** *
```

Гэпы (прочерки в последовательности) показывают, что для данного аминокислотного остатка нет гомологичного в другом белке. Биологическая причина — инсерции и делеции, закрепившиеся в эволюции.

Выделяют программы локального и глобального выравнивания. Программа глобального выравнивания только расставляет гэпы, выравнивая последовательности по всей длине.

Программа локального выравнивания выбирает в каждой последовательности по участку, а затем выравнивает между собой эти участки.

Глобальное выравнивание используется, когда есть уверенность в том, что последовательности гомологичны. Если последовательности содержат только гомологичные участки или если про последовательности неизвестно ничего, то используется локальное выравнивание.

Примеры алгоритмов выравнивания:

- парное глобальное выравнивание (global alignment of two sequences) — алгоритм Нидлмана–Вунша (Needleman&Wunsch); вот ссылка на визуализацию работы алгоритма,
- парное локальное выравнивание (local alignment of two sequences) — алгоритм Смита–Уотермена (Smith&Waterman),
- множественное выравнивание (multiple alignment), когда последовательностей больше двух, — ClustalW, Muscle, MAFFT, T-Coffee и пр.

Программы выравнивания действуют формально и выдадут выравнивание, даже если на вход им подать негомологичные белки. Смысла такое выравнивание иметь не будет.

Даже если белки гомологичны, в выравнивании, выданном программой, не обязательно будут сопоставлены гомологичные аминокислотные остатки или нуклеотиды. Алгоритмы выравнивания сделаны так, чтобы максимизировать долю правильных сопоставлений. Но невозможно придумать такой алгоритм, который бы выдавал биологически правильный ответ всегда. Выравнивание, выданное программой, — это всего лишь реконструкция причины (эволюции) по последствиям (т.е. современным последовательностям).

Выравнивание необходимо, чтобы

- предсказать функцию белка на основании установленной гомологии с экспериментально изученными белками,
- выяснить, какие аминокислотные остатки консервативны (следовательно, важны для функции белка),
- оценить время расхождения последовательностей, а если последовательностей много, то реконструировать их эволюционную историю.