

Структурная биология белка

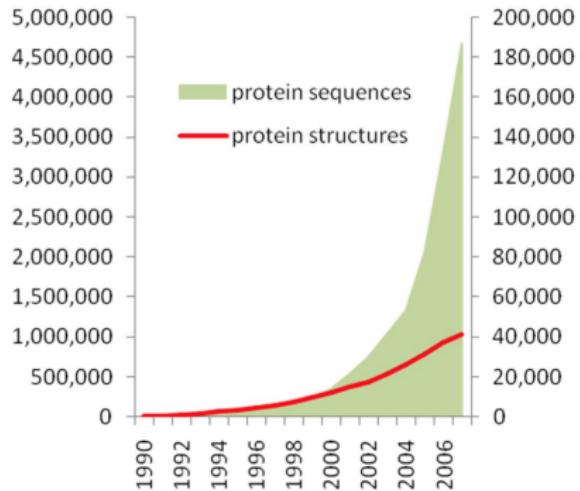
Лекция 1: Введение в протеомику и сравнение белков

Павел Яковлев

СПбАУ НОЦНТ РАН

9 сентября 2014

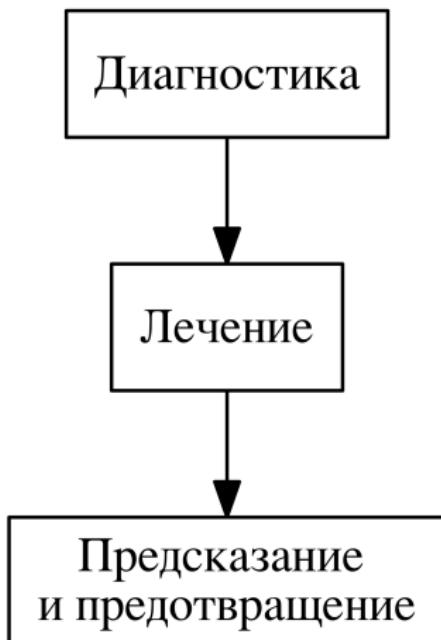
Мотивация



Вычислительная протеомика:

- Биология
- Математика
- Информатика
- Физика
- Химия

Bioinformatics ∈ Health Science



Геномика:

- определение мутаций и перестроек;
- определение бактериального состава;
- поиск вирусных фрагментов;
- генотерапия.

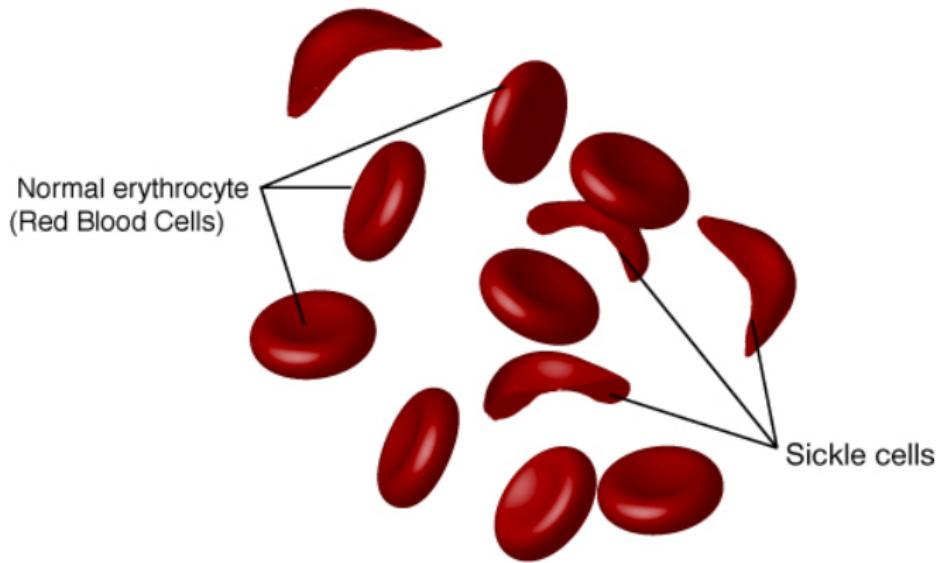
Протеомика:

- определение нарушений функций белка и белковых взаимодействий;
- создание превентивных и терапевтических лекарственных средств.

Серповидно-клеточная анемия

6Glu>Val
↓

| | |
|-----|--|
| HbA | ACACCA <u>T</u> GGTG <u>C</u> ATCT <u>G</u> ACT <u>C</u> CT <u>G</u> AGGGAG <u>A</u> AGTC <u>T</u> G <u>C</u> CG <u>T</u> T <u>A</u> CT <u>G</u> |
| HbS | ACACCA <u>T</u> GGTG <u>C</u> AC <u>C</u> CT <u>G</u> ACT <u>C</u> CT <u>G</u> T <u>G</u> GAG <u>A</u> AGTC <u>T</u> G <u>C</u> CG <u>T</u> T <u>A</u> CT <u>G</u> |

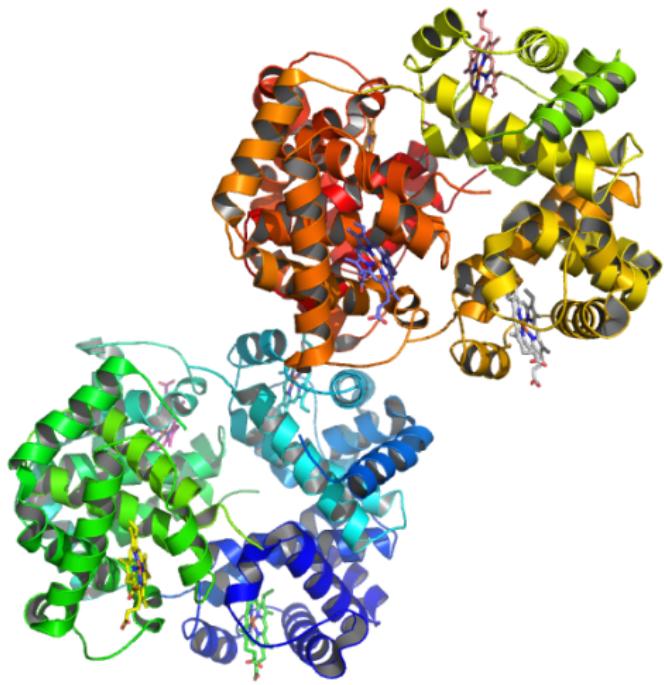


Серповидно-клеточная анемия

May 21, 2014 – 29,901,117 Homo Sapiens SNPs (dbSNP build 141)

Связанные мутации:

- Приапизм:
 - rs2249358
 - rs211239
- Инсульт:
 - rs11853426
 - rs267196
 - rs408505
 - rs3917733
 - rs284875
 - rs989554
- Язвы ног:
 - rs736839

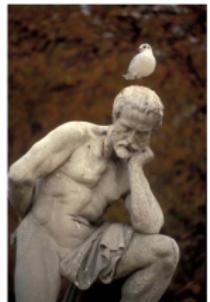
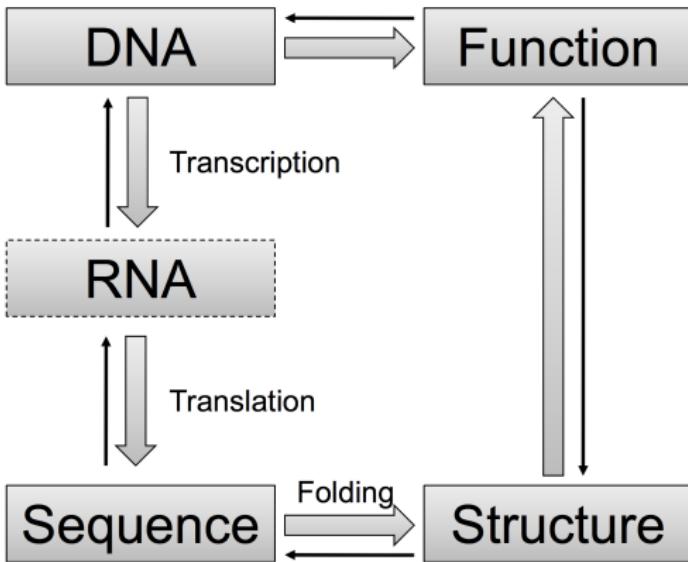
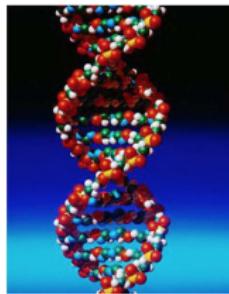


План курса

Лекции:

- ① Введение в вычислительную протеомику.
- ② Вторичные структуры белка.
- ③ Петли во вторичной структуре.
- ④ Построение третичных структур.
- ⑤ Сравнение структур белков: структурное выравнивание.
- ⑥ Докинг: моделирование белковых взаимодействий.
- ⑦ Задача дизайна белков.

ДНК -> ... -> Функция



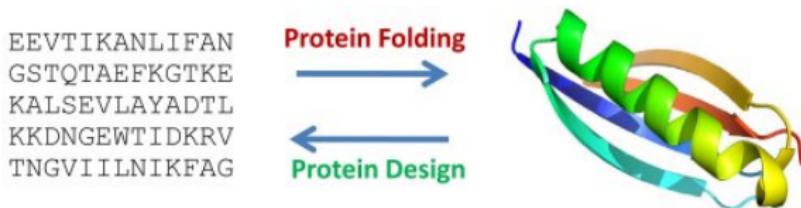
Вопросы протеомики

Annotation: Какие замены приводят к нарушению функции белка? Влияют ли разные замены по-разному?

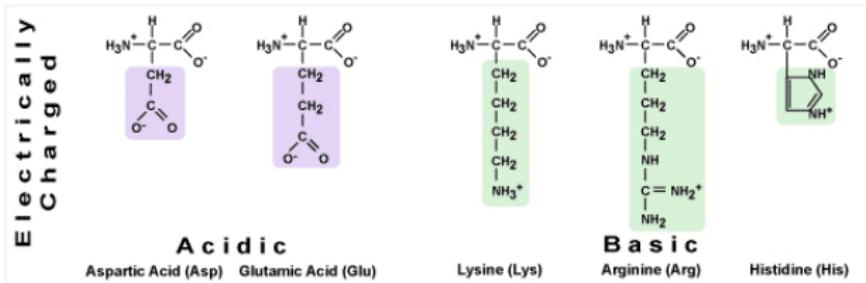
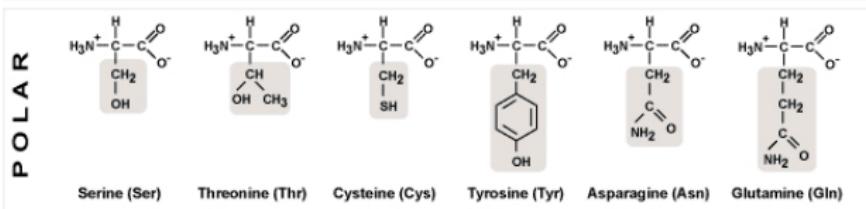
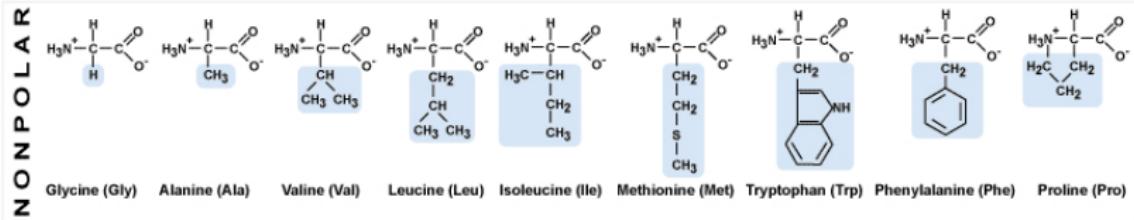
Folding: Как устроен белок?

Docking: Как белки взаимодействуют между собой?

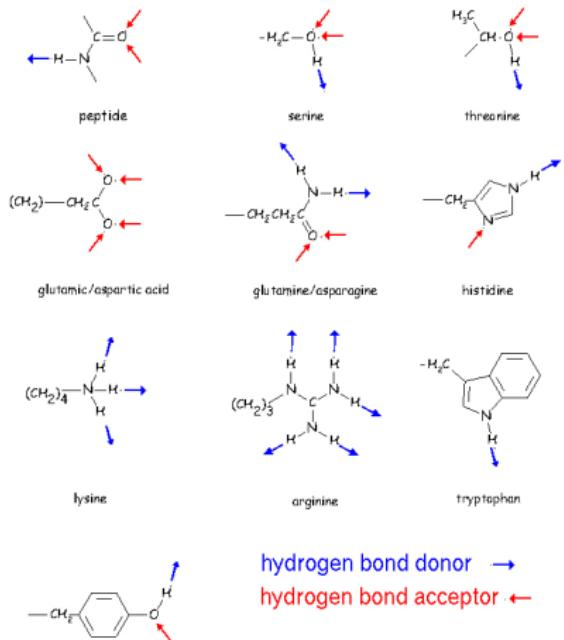
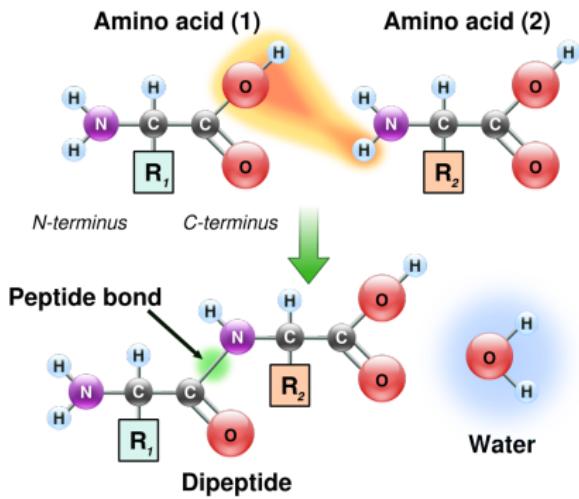
Design (inverse folding): Как создать белок с нужной функцией?



АМИНОКИСЛОТЫ



Связи аминокислот

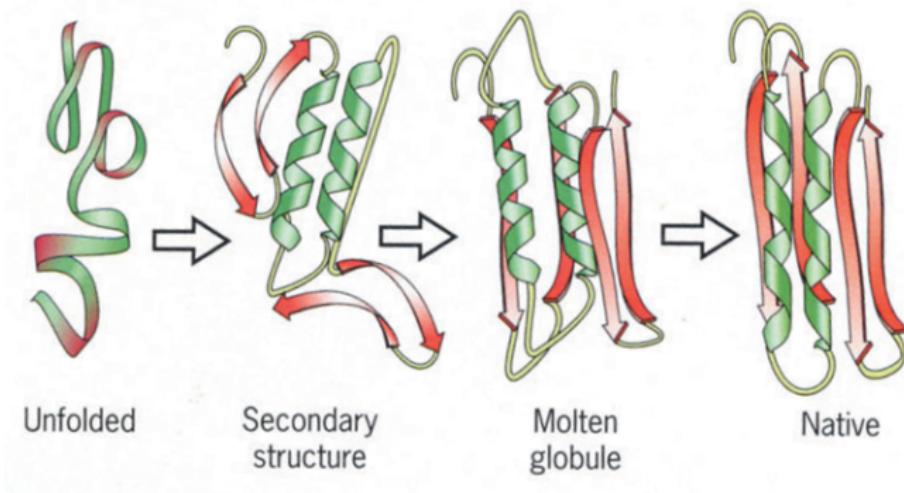


Парадокс Левинталя

Промежуток времени, за который полипептид приходит к своему скрученному состоянию, на много порядков меньше, чем если бы полипептид просто перебирал все возможные конфигурации.

Пример: для цепи из 100 аминокислот возможное количество конформаций 10^{100} . При переборе со скоростью 10^{13} конформаций в секунду поиск занял бы порядка 10^{80} лет.

Этапы фолдинга



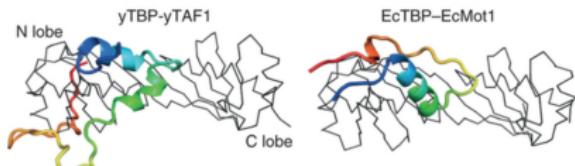
Анализ схожести последовательностей

Образование третичной структуры из локальных независимых фрагментов – элементов вторичной структуры – указывает на образование схожих структур схожими протяженными аминокислотными последовательностями.

| | | |
|--------|---|--|
| yTAF1 | 19 23 26 27 | |
| | T N L A N E D E A Y E - - A I F G G E - F G S L E I G S Y I G G | |
| EcMot1 | 129 123 120 118 | |
| | S G A K F D I D S L F N A E V H E - L K A I K R V M R | |

Метрики между строками:

- расстояние Хэмминга;
- расстояние Левенштейна;
- выравнивание с использованием матриц.



Выбор матрицы расстояний

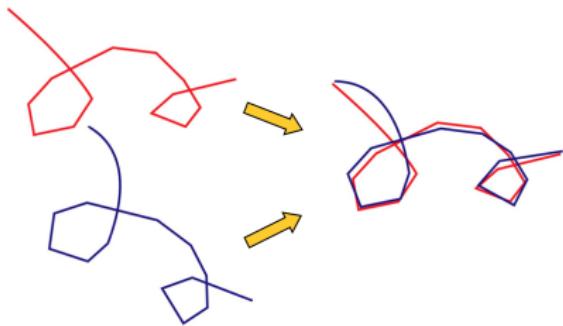


Матрицы для выравнивания:

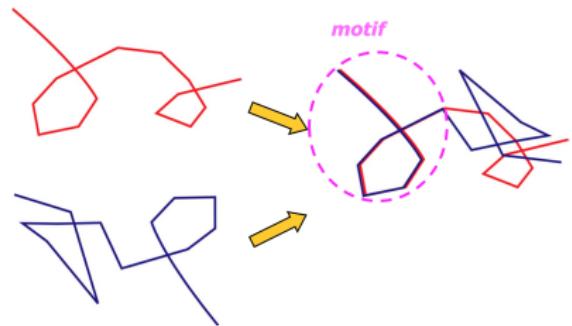
- BLOSUM – матрица родственных белков
Основа: близкие гомологи
- PAM – эволюционная матрица
Основа: теоретическая эволюция с установленной степенью мутации

Парные выравнивания: выбор вида

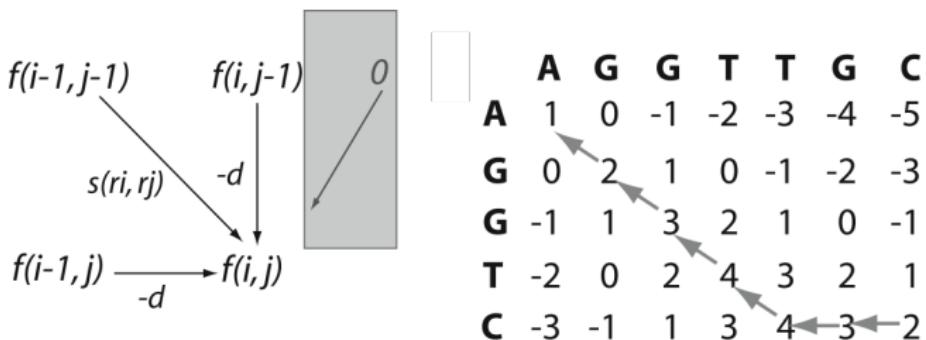
Глобальное:



Локальное:



Парные выравнивания: динамика



| A | G | G | T | T | G | C |
|---|---|---|---|---|---|---|
| A | G | G | T | - | - | C |

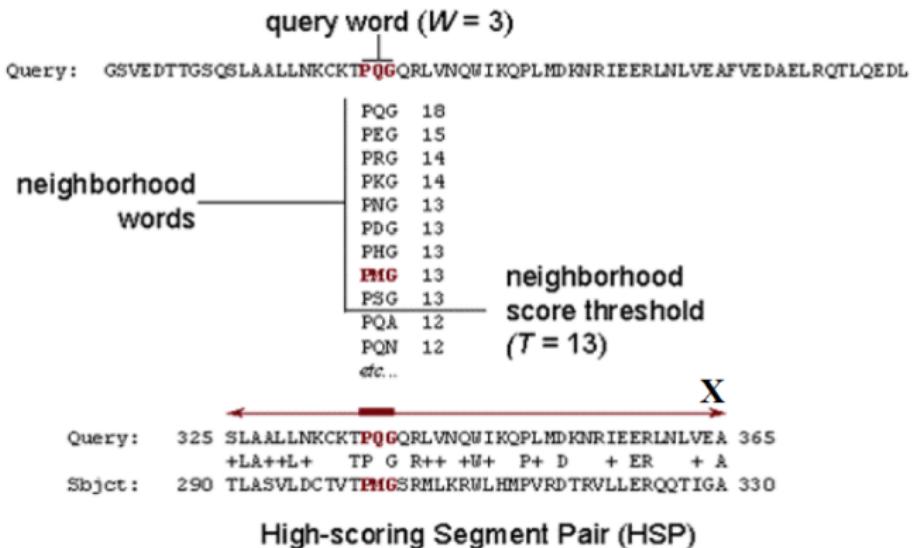
Парные выравнивания: BLAST I

Идея: Несмотря на снижение сходства родственных последовательностей при их расхождении с течением времени, мы можем рассчитывать обнаружить короткие участки высокого сходства, не затронутые мутациями.

Парные выравнивания: BLAST II

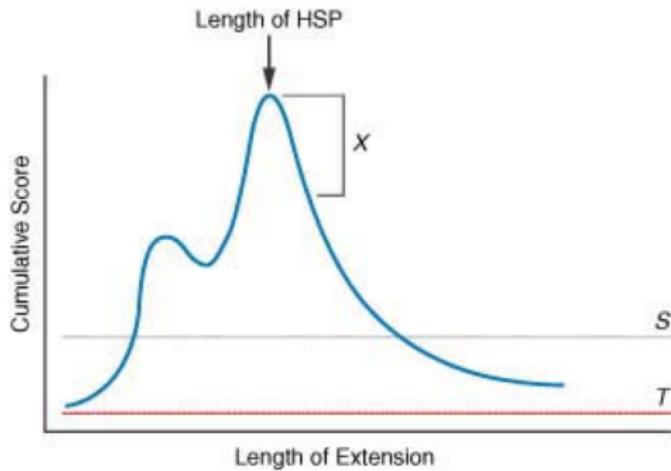
- ① Для каждого слова длины W в искомой последовательности составляется список схожих слов, вес выравнивания которых выше определенного порога T .
- ② По заранее построенной хэш-таблице ищем последовательности в базе данных, имеющие точное вхождение данных слов.
- ③ Расширяем выравнивание вправо и влево от найденных "затравок", используя алгоритм динамического программирования; ищем продленные слова, обладающие максимальным сходством с запросом (high-scoring segment pairs, HSP).

Парные выравнивания: BLAST III



Парные выравнивания: BLAST IV

Прекращаем расширение выравнивания, если падение суммарного веса выравнивания от точки последнего максимума достигнет заранее установленного порога X .



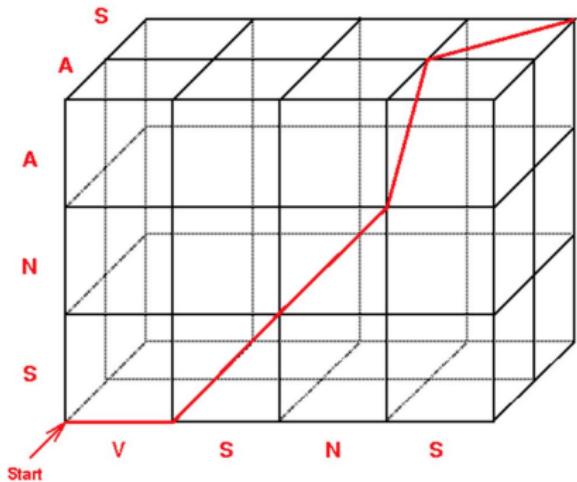
Множественные выравнивания

Зачем?

- Аннотация белков.
- Выявление активных центров гомологов.
- Связь протяженных фрагментов со структурой.
- Построение профилей доменов.

```
DAMMfly2R_ : MYLPERTEHQKIERLY-----DSNRVN-----AEPGQGL---  
DCP1fly2R_ : -----MTD-----ECVTRNYGVGIRSPNGSENRGS-FIMADNTDAK-----GCTPESIVVGG  
DRICEfly3R : MDATNNGESADQVGIRVGN-----PEQPNDHTDALGSV-GSGGAGSSGLVAGSSHFY-----GSGAIGQLANG  
DECAYfly3R : MDDTDIFSLFGQKNKHK-----KDKADATKIA-----HTPTSEL---  
DRONCfly3L : MQPPELEIGMPFKRHREHIRKLNILNVEWNTYERLAMECVQQGILTVQMLRNTQDLNGK-FPNMDEKDVREQHRRLLLKITQRGPTAYNLLINA  
STRICAfly2 : MGWWSKSETDRSQPSQELVAQDPRTRVQTTSAAETTTNTAVQNSSTITDNKKQTVTFI-TTRQTVTHTQRALITETTTRRTPSQABLEALFAKI  
DREDDPAfly : MSASAIYRPFPKVKHFCIFFIAMAGSNLLIHLDTIDQNDLIYVERDMNFAQKVGLCFL-LYGDDHSDATYILQKLLAMTRSDFPQS DLLIFAK  
DREDDPBfly : MSASAIYRPFPKVKHFCIFFIAMAGSNLLIHLDTIDQNDLIYVERDMNFAQKVGLCFL-LYGD DHSDATYILQKLLAMTRSDFPQS DLLIFAK  
DREDDPCfly : MSASAIYRPFPKVKHFCIFFIAMAGSNLLIHLDTIDQNDLIYVERDMNFAQKVGLCFL-LYGD DHSDATYILQKLLAMTRSDFPQS DLLIFAK
```

MSA: динамика



$$d_{i,j,k} = \begin{cases} d_{i-1,j-1,k-1} + S(x_i, x_j, x_k) \\ d_{i,j-1,k-1} + S(-, x_j, x_k) \\ d_{i-1,j,k-1} + S(x_i, -, x_k) \\ d_{i-1,j-1,k} + S(x_i, x_j, -) \\ d_{i,j,k-1} + S(-, -, x_k) \\ d_{i,j-1,k} + S(-, x_j, -) \\ d_{i-1,j,k} + S(x_i, -, -) \end{cases}$$

Ячеек в матрице: $L_1 \cdot L_2 \cdot \dots \cdot L_N = O(L_{max}^N)$

Вариантов для ячейки: $O(2^N)$

Сложность:

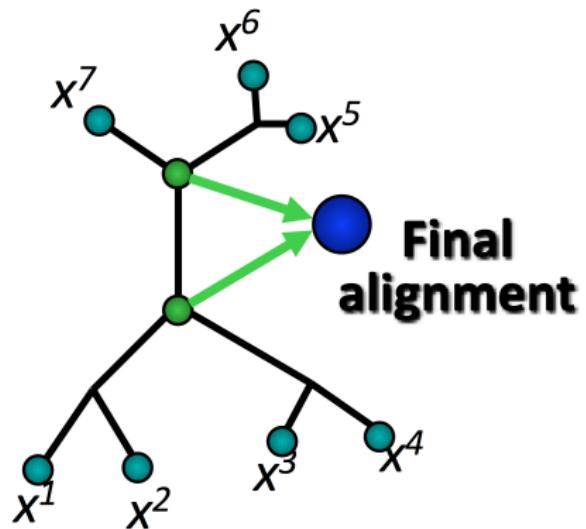
- Временная: $O(2^N \cdot L^N)$
- Пространственная: $O(L^N)$

MSA: прогрессивное выравнивание

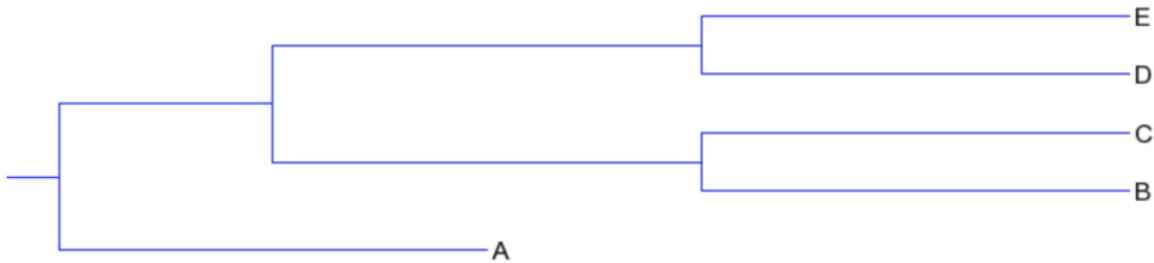
Feng and Doolittle 1987, Higgins and Sharp 1988

Алгоритм:

- 1 построить все попарные выравнивания между последовательностями для получения матрицы расстояний;
- 2 построить приближенное филогенетическое дерево одним из методов иерархической кластеризации;
- 3 осуществить последовательное профильное выравнивание от более похожих последовательностей к менее похожим.



Иерархическая кластеризация



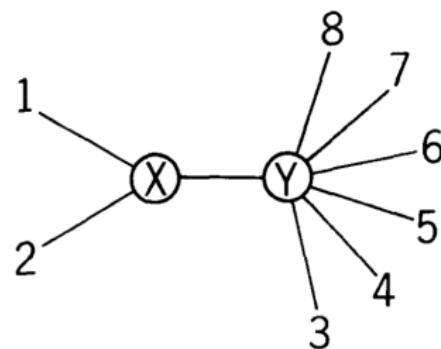
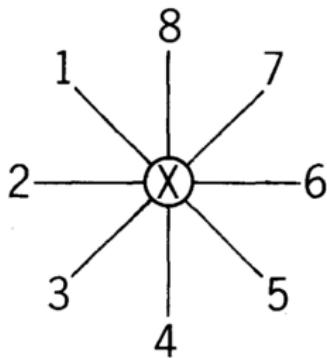
Идея: все элементы множества имеют различные попарные расстояния друг с другом. Требуется упорядочить их в некоторую структуру, топология которой зависит от заданных расстояний. Иными словами, имеется матрица всех попарных расстояний, по которой требуется построить дерево с сохранением всех расстояний. **Проблема:** как пересчитывать расстояния от сформированного дерева до оставшихся последовательностей или других поддеревьев?

Neighbor Joining I

Naruya Saitou and Masatoshi Nei, 1987

Алгоритм:

- Начинаем с дерева с топологией “звезды”.



Neighbor Joining II

Алгоритм:

- Посчитать матрицу Q по имеющимся дистанциям:

$$Q_{i,j} = (n - 2) \cdot d_{i,j} - \sum_{k=1}^n d_{i,k} - \sum_{k=1}^n d_{j,k}$$

Neighbor Joining III

- Объединить два ближайших элемента f и g (в соответствии с матрицей Q), добавив общего родителя u .
Расстояние до u :

$$\delta_{f,u} = \frac{1}{2}d_{f,g} + \frac{1}{2(n-2)}\left(\sum_{k=1}^n d_{f,k} - \sum_{k=1}^n d_{g,k}\right)$$

$$\delta_{g,u} = d_{f,g} - \delta_{f,u}$$

Расстояния от u до оставшихся элементов:

$$d_{u,k} = \frac{1}{2}(d_{f,k} + d_{g,k} - d_{f,g})$$

Neighbor Joining IV

Сложность:

- $(n - 3)$ итерации;
- на каждом шаге пересчитывается \mathbf{Q} размера $O(N^2)$

Итого: $O(n^3)$

Достоинства:

- скорость, можно выравнивать большие объемы данных и проводить bootstrap;
- субоптимальное решение;
- статистически консистентный.

Недостатки:

- результат – неукорененное дерево;
- жадный алгоритм, может выдать неправильную топологию;
- проверяет только одно дерево.

WPGMA и UPGMA

PGMA – Pair-Group Method Using Arithmetic Averages

Расстояние до родителя:

$$D_{(u,v),u} = D_{(u,v),v} = \frac{1}{2} D_{u,v}$$

Weighted:

$$D_{(u,v),w} = \frac{D_{u,w} + D_{v,w}}{2}$$

Unweighted:

$$D_{(u,v),w} = \frac{|u| \cdot D_{u,w} + |v| \cdot D_{v,w}}{|u| + |v|}$$

Выравнивание по профилю

Варианты выравнивания:

- последовательность на последовательность;
- последовательность на профиль;
- профиль на профиль.

Выравнивания осуществляются стандартными алгоритмами выравнивания с учетом изменения алфавита и матрицы выравнивания.

Position-Specific Scoring Matrix I

Вес профиля для АК a в позиции p :

$$M(p, a) = \sum_{b \in AA} f(p, b) \cdot s(a, b)$$

$f(p, b)$ – частота АК b в позиции p

$s(a, b)$ – score выравнивания a на b (BLOSUM, PAM, etc)

Аналогично считается score для gap.

Position-Specific Scoring Matrix II

| POS | PROBE | CONSENSUS | PROFILE | | | | | | | | | | | | | | | | | | | | |
|-----|---------|-----------|---------|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|-----|----|-----|
| | | | A | C | D | E | F | G | H | I | K | L | M | N | P | Q | R | S | T | V | W | Y | +/- |
| 1 | E G V L | V | 3 | -2 | 3 | 4 | 0 | 4 | -1 | 3 | -1 | 4 | 4 | 1 | 1 | 1 | -2 | 1 | 2 | 6 | -6 | -2 | 9 |
| 2 | L L S P | L | 2 | -2 | -2 | -1 | 3 | 0 | -1 | 3 | -1 | 6 | 5 | -1 | 3 | 0 | -1 | 3 | 1 | 4 | 1 | -1 | 9 |
| 3 | V V V V | V | 2 | 2 | -2 | -2 | 2 | 2 | -3 | 11 | -2 | 8 | 6 | -2 | 1 | -2 | -2 | 0 | 2 | 15 | -9 | -1 | 9 |
| 4 | K E A T | A | 6 | -2 | 5 | 6 | -5 | 4 | 1 | 0 | 5 | -2 | 0 | 3 | 3 | 3 | 1 | 3 | 6 | 0 | -6 | -4 | 9 |
| 5 | A P L P | A | 6 | -1 | 0 | 1 | -2 | 2 | 0 | 1 | 0 | 2 | 2 | 0 | 8 | 2 | 0 | 2 | 2 | 3 | -5 | -4 | 9 |
| 6 | G G G G | G | 7 | 1 | 7 | 5 | -6 | 15 | -1 | -3 | 0 | -4 | -3 | 4 | 3 | 2 | -3 | 6 | 4 | 2 | -11 | -7 | 9 |
| 7 | S S Q E | D | 4 | -1 | 7 | 7 | -6 | 7 | 2 | -2 | 2 | -3 | -2 | 4 | 3 | 6 | 1 | 6 | 2 | -1 | -6 | -5 | 9 |
| 8 | S S T P | S | 4 | 4 | 2 | 2 | -4 | 4 | -1 | 0 | 2 | -3 | -2 | 2 | 7 | 0 | 1 | 10 | 6 | 0 | -2 | -4 | 9 |
| 9 | V L V A | V | 5 | 0 | -1 | -1 | 3 | 1 | -2 | 7 | -2 | 7 | 6 | -1 | 1 | -1 | -3 | 0 | 2 | 10 | -5 | -1 | 9 |
| 10 | K R R S | R | 0 | -1 | 1 | 1 | -5 | 0 | 2 | -2 | 8 | -3 | 1 | 3 | 3 | 3 | 10 | 5 | 1 | -2 | 7 | -5 | 9 |
| 11 | M L I I | I | 0 | -2 | -3 | -2 | 7 | -3 | -3 | 11 | -1 | 11 | 10 | -2 | -2 | -1 | -2 | -2 | 1 | 9 | -3 | 1 | 9 |
| 12 | S S T S | S | 4 | 6 | 2 | 2 | -3 | 5 | -1 | 0 | 2 | -3 | -2 | 3 | 4 | -1 | 1 | 12 | 6 | 0 | 0 | -4 | 9 |
| 13 | C C C C | C | 3 | 15 | -5 | -5 | -1 | 2 | -1 | 3 | -5 | -8 | -6 | -3 | 1 | -6 | -3 | 7 | 3 | 3 | -13 | 10 | 9 |
| 14 | K S Q R | K | 1 | -2 | 3 | 3 | -6 | 1 | 3 | -2 | 7 | -3 | 0 | 3 | 3 | 5 | 7 | 4 | 1 | -2 | 2 | -5 | 9 |
| 15 | A A G S | A | 10 | 3 | 4 | 3 | -5 | 8 | -1 | -1 | 1 | -2 | -1 | 3 | 4 | 1 | -2 | 7 | 4 | 2 | -6 | -4 | 9 |
| 16 | T S D S | S | 4 | 3 | 5 | 4 | -5 | 6 | 0 | 0 | 2 | -3 | -2 | 4 | 3 | 1 | 1 | 9 | 6 | 0 | -3 | -4 | 9 |
| 17 | G G S Q | G | 5 | 1 | 6 | 5 | -6 | 9 | 1 | -2 | 1 | -3 | -2 | 4 | 3 | 4 | 0 | 6 | 3 | 0 | -6 | -6 | 9 |
| 18 | Y F L S | F | -1 | 2 | -4 | -3 | 9 | -3 | 0 | 4 | -3 | 6 | 3 | -1 | -3 | -3 | -3 | 1 | -1 | 2 | 7 | 7 | 9 |
| 19 | T T R L | T | 1 | -2 | 0 | 1 | 0 | 0 | 0 | 2 | 2 | 2 | 3 | 1 | 1 | 1 | 3 | 1 | 7 | 2 | 1 | -2 | 9 |
| 20 | F F . L | F | -2 | -3 | -6 | -4 | 10 | -4 | -1 | 6 | -4 | 9 | 6 | -3 | -4 | -4 | -3 | -2 | -1 | 3 | 7 | 8 | 4 |
| 21 | S S . D | S | 3 | 2 | 5 | 4 | -4 | 5 | 0 | -1 | 2 | -3 | -2 | 4 | 3 | 1 | 1 | 8 | 2 | -1 | -2 | -3 | 4 |
| 22 | S S . S | S | 2 | 3 | 1 | 1 | -2 | 3 | -1 | 0 | 1 | -2 | -1 | 2 | 2 | 0 | 1 | 8 | 2 | 0 | 1 | -2 | 4 |
| 23 | . . . G | G | 2 | 0 | 2 | 1 | -2 | 4 | 0 | 0 | 0 | -1 | -1 | 1 | 1 | 1 | -1 | 2 | 1 | 1 | -3 | -2 | 4 |
| 24 | . . . D | D | 1 | -1 | 4 | 3 | -2 | 2 | 1 | 0 | 1 | -1 | -1 | 2 | 1 | 2 | 0 | 1 | 1 | 0 | -3 | -1 | 4 |
| 25 | . . . G | G | 2 | 0 | 2 | 1 | -2 | 4 | 0 | 0 | 0 | -1 | -1 | 1 | 1 | 1 | -1 | 2 | 1 | 1 | -3 | -2 | 4 |
| 26 | . A G N | A | 6 | 0 | 4 | 3 | -4 | 6 | 1 | -1 | 1 | -2 | -1 | 5 | 2 | -2 | -1 | 3 | 3 | 1 | -5 | -3 | 4 |
| 27 | Y N Y T | Y | 0 | 5 | 0 | -1 | 5 | -1 | 2 | 1 | -1 | 0 | -1 | 4 | -3 | -2 | -2 | 0 | 3 | 0 | 3 | 6 | 4 |
| 28 | E D D Y | D | 2 | -2 | 9 | 8 | -3 | 3 | 4 | -1 | 1 | -3 | -2 | 5 | -1 | 4 | -1 | 1 | 1 | -1 | -6 | 0 | 9 |
| 29 | L M A L | L | 3 | -5 | -3 | -1 | 6 | -1 | -2 | 6 | -1 | 10 | 10 | -2 | 0 | 0 | -2 | -1 | 0 | 6 | -1 | 0 | 9 |
| 30 | Y N A W | N | 4 | 1 | 3 | 2 | 0 | 2 | 3 | -1 | 1 | -1 | -1 | 8 | 0 | 1 | -1 | 2 | 1 | -1 | -1 | 2 | 9 |
| * | * | | | | | | | | | | | | | | | | | | | | | | |
| * | * | | | | | | | | | | | | | | | | | | | | | | |
| * | * | | | | | | | | | | | | | | | | | | | | | | |
| 48 | S G N S | S | 4 | 3 | 5 | 3 | -4 | 7 | 0 | -2 | 2 | -4 | -3 | 6 | 3 | 1 | 0 | 10 | 3 | 0 | -2 | -4 | 9 |
| 49 | S S N Y | S | 2 | 5 | 2 | 1 | 1 | 2 | 1 | 0 | 1 | -2 | -2 | 5 | 1 | -1 | 0 | 8 | 1 | -1 | 3 | 1 | 9 |

Progressive alignment = ...?

Достоинства:

- скорость: $O(L^2 N^2)$;
- соответствие биологической природе.

Недостатки:

- качество зависит от порядка добавления в выравнивания;
- однажды допущенные ошибки остаются в выравнивании;
- очень плохо работает на сильно удаленных последовательностях;
- нет возможности оценить качество выравнивания без знания истинного выравнивания.

| | | | | | |
|-----------|---|-----|-----|-----|-----|
| Hbb_Human | 1 | - | | | |
| Hbb_Horse | 2 | .17 | - | | |
| Hba_Human | 3 | .59 | .60 | - | |
| Hba_Horse | 4 | .59 | .59 | .13 | - |
| Myg_Whale | 5 | .77 | .77 | .75 | .75 |



alpha-helices

| | | | | | |
|---|---------------------------|---|--|--|--|
| 1 | PEEKSAVTALWGKVNV--VDEVGG | 2 | | | |
| 2 | GEEKAAVLALWDKVN--EEEVGG | 3 | | | |
| 3 | PADKTNVKAAANGKVGAGHAGEYGA | 4 | | | |
| 4 | AADKTNVKAAWSKVGGHAGEYGA | 5 | | | |
| 5 | EHEWQLVLHVWAKVEADVAGHGQ | | | | |

Улучшения

- применение bootstrap при построении деревьев;
- использование альтернативных функций оценки профиля;
- применения предварительного локального выравнивания участков и объединение профилей;
- использование НММ для слияния соседних столбцов и уменьшения количества пропусков;
- эвристики!

Эволюция профилей

Минимизация взаимной энтропии:

Идея: использовать для каждой позиции наиболее подходящую модель истанций между АК. В позиции p при находящейся в ней АК b перебираются различные значения x (в соответствии с моделью РАМ), минимизирующие функцию:

$$H = - \sum_{a \in AA} f_a \cdot \ln p_a$$

f_a – наблюдаемая частота a

p_a – предсказанная частота a при условии, что b – предшествующая аминокислота, а x – использующийся при оценке дистанции вес.

Эвристики

- Назначить разный вес последовательностям для разной глубины прогрессивного выравнивания.
- Менять матрицы выравнивания на разных этапах в зависимости от удаленности последовательностей.
- Уменьшить в ранних выравниваниях вес gap.
- Использовать зависимую стоимость гэпа от заменяемой аминокислоты:
 - стоимость gap уменьшается для гидрофобных аминокислот;
 - стоимость gap в потенциальных петлях больше, чем в потенциальных складках.

Сравнение качества

Вес выравнивания:

$$S(m) = \sum_i S(m_i)$$

Вес колонки:

$$S(m_i) = - \sum_a p_{i,a} \log p_{i,a}$$

где

$$p_{i,a} = \frac{c_{i,a}}{\sum_b c_{i,b}}$$

$c_{i,a}$ – количество остатков **a** в позиции **i**

Домашнее задание

Реализовать прогрессивное выравнивание с выбором метода построения дерева: Neighbor Joining, UPGMA, WPGMA.

Улучшение выравнивания опционально.

Вход: FASTA-файл, матрица выравнивания (txt from NCBI FTP), тип кластеризации.

Выход: FASTA-файл с гэпами.

Репозиторий: <https://github.com/bioinf/proteomics2014>

Срок: 4 недели (7 октября)

yakovlev@biocad.ru