Профили выравниваний

План

- Белки: гомология и сходство последовательностей
- Профиль семейства доменов: надо учитывать возможность вставок/делеций
- Паттерн и PROSITE
- PSSM и psi-BLAST
- Pftools (PROSITE, myHits) и **HHM-профили** (Pfam)
- Как интерпретировать результат поиска (**ROC-кривая**)

1. Паттерны для поиска в базах последовательностей

Prosite (http://prosite.expasy.org/), fuzzpro и fuzznuc в EMBOSS

ПРОПУСТИТЬ

Паттерн для цинкового пальца

Prosite

Паттерн для цинкового пальца типа C2H2:

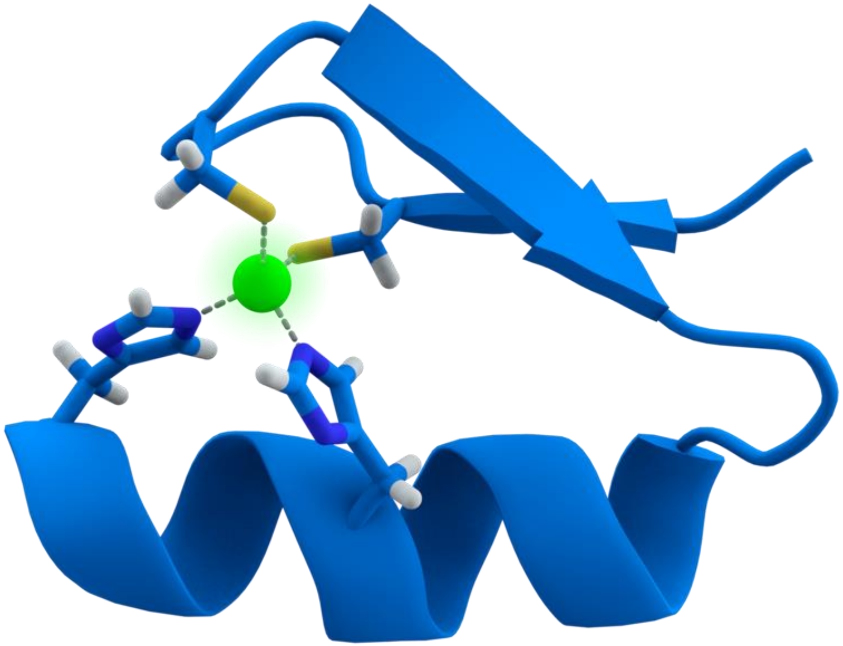
C-x(2,4)-C-x(3)-[LIVMFYWC]-x(8)-H-x(3,5)-H

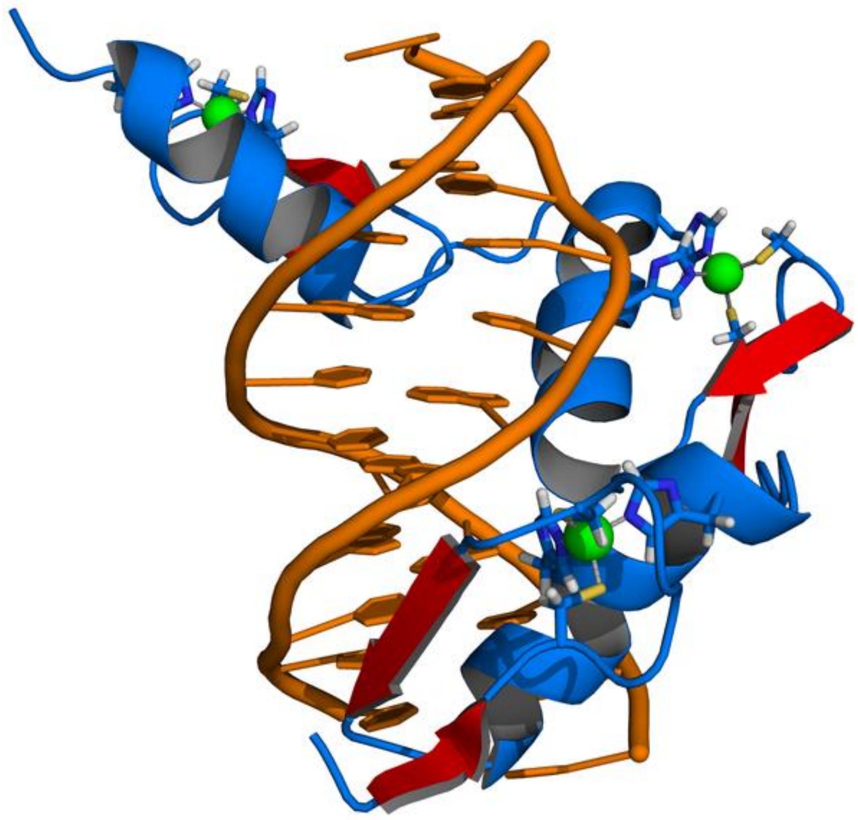
- [a-zAZ] –все возможные аминокислоты в данной позиции
- X(2,4) – любая аминокислота от 2 до 4 раз
- X(3) – любая аминокислота ровно 3 раза
- {P} – любая аминокислота, кроме пролина

Паттерны (fingerprints) для белков и средства поиска по паттерну есть в ProSite и пакете EMBOSS

ПРОПУСТИТЬ

Цинковые пальцы C2H2





2. PSSM – аналог PWM для белков

Psi-BLAST – итерактивный вариант BLAST, использующий блоки множественного выравнивания и поиск по PSSM

PSSM – то же, что PWM

- PSSM, или Position-Specific Scoring Matrix, строится по блоку – выравниванию без вставок/делеций
- Используется в программе PSI-BLAST (и MEME)
  - по последовательностям из списка находок, отмеченным для очередной итерации, строится выравнивание
  - в выравнивании находятся блоки
  - по блоку строится PSSM
  - по всем PSSM ведется поиск; веса разных PSSM в одной банковской последовательности суммируются
  - получается, что PSI-BLAST разрешает участки переменной длины между находками PSSM, но никак их не использует при вычислении веса

3. Профили выравниваний и поиск по ним в базах последовательностях

HMM = Hidden Markov Model

Технология HMM реализована в пакете HMMER. Он включен в EMBOSS.

В БД Prosite реализована аналогичная, но не идентичная, технология Pftools

Профили

- На вход подается выравнивание с инделями
- По нему строится т.н. профиль HMM (Hidden Markov Model)
- Профиль HMM можно выровнять с последовательностью и получить вес выравнивания. Локальное и глобальное выравнивание.
- Профиль калибруется по случайному банку для нормализации веса и расчета E-value
- При наличии множества последовательностей, про которые известен ответ – есть в них домен или нет, - можно уточнить порог нормализованного веса для находки
- С помощью профиля в базе последовательностей (Uniprot) находятся участки с весом больше порога, следовательно, белки, содержащие домен.
- Важное отличие профиля от PWM:
профиль может быть построен по выравниванию с инделями

НММ Профиль. Немножко теории

- По выравниванию создается автомат для генерации последовательностей
  - Этот автомат умеет генерировать случайные последовательности конечной (<u>но не фиксированной!</u>) длины
  - Он настроен так, чтобы создавать последовательности, “похожие” на выравнивание, с бóльшей вероятностью
- Для каждой входной последовательности можно (т.е. существуют алгоритмы) определить вероятность её сгенерировать этим автоматом.
- Если эта вероятность превышает порог, то последовательность считается соответствующей профилю.

Автомат выглядит так:

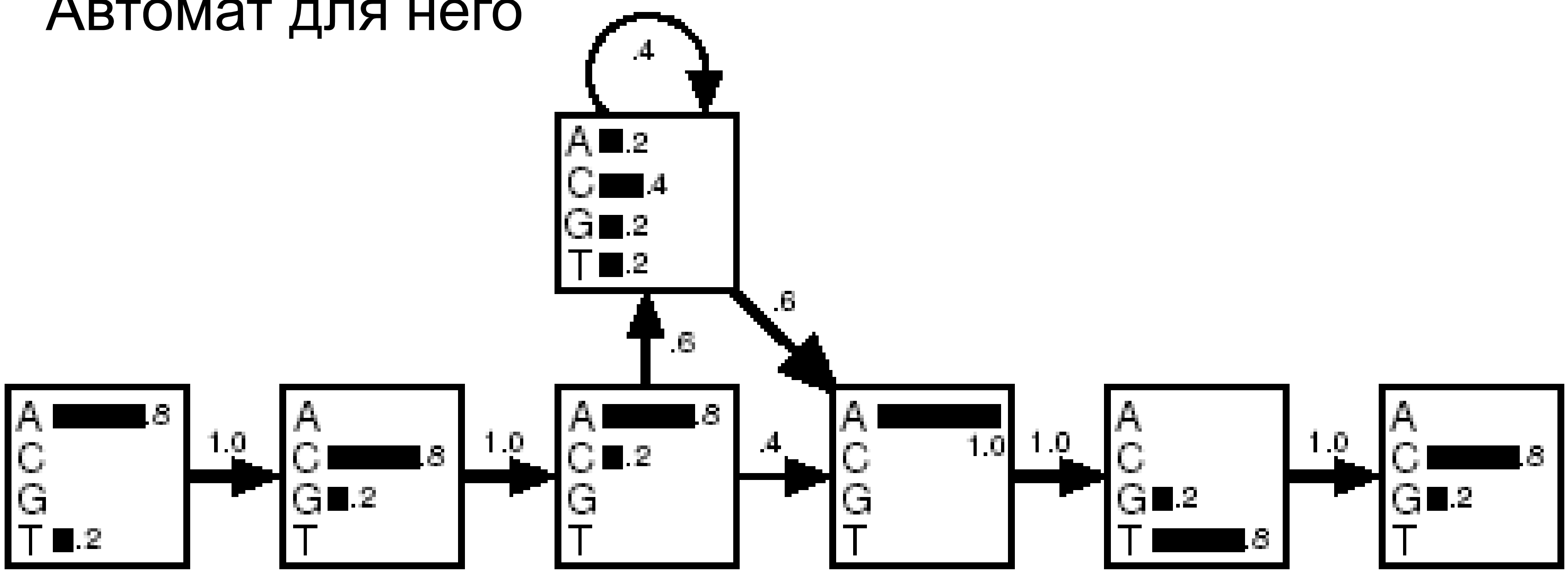
Выравнивание

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

Вероятности в квадратиках называются *эмиссионными вероятностями*

Вероятности на стрелочках - *вероятностями перехода*

Автомат для него



from Krogh, “Computational Methods in molecular biology, pages 45-63, Elsevier, 1998.

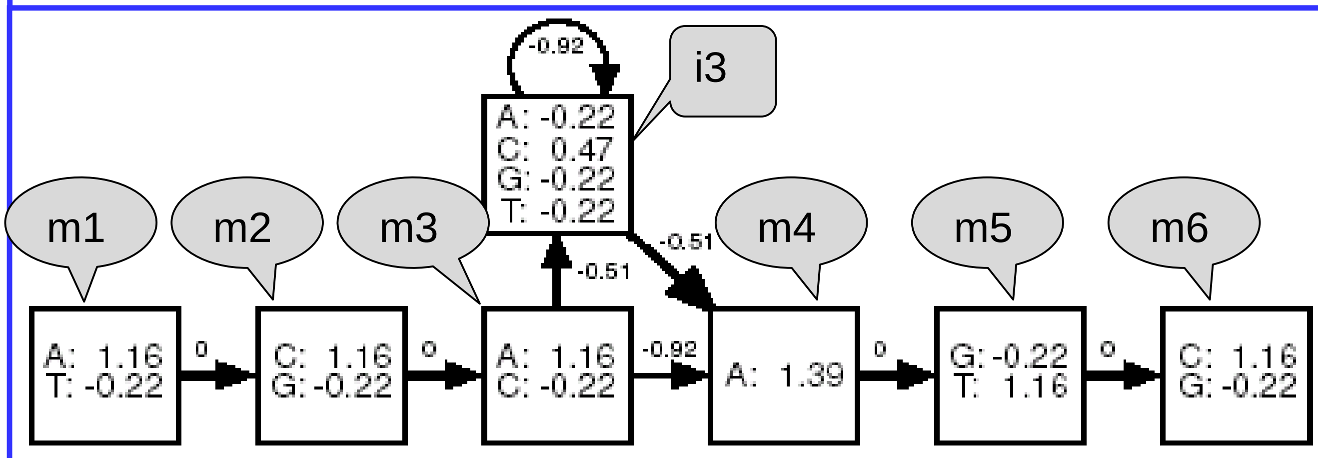
Логарифм отношения правдоподобия, log-odds

- Вероятность буквы в данной позиции следует сравнивать с базовой – по геному - частотой этой буквы.
- Пусть базовые частоты всех букв одинаковы и, следовательно, равны 0.25
- Отношение правдоподобия для буквы А в первой позиции примера равно 0.8/0.25 = 3.2
- Удобнее взять логарифм – чтобы складывать, а не умножать:

  log-odds = ln 3.2 = 1.16
- Log-odds >> 0 – за то, что буква А не случайно похожа на колонку выравнивания
- Log-odds $\approx$ 0 – за то, что буква А соответствует случайному выбору
- Log-odds << 0 – за то, что буква А избегается в колонке выравнивания

Определим вес данного выравнивания последовательности ACACATC с профилем

	Sequence	Probability ×100	Log odds
Consensus	A C A C - - A T C	4.7	6.7
Original sequences	A C A - - - A T G	3.3	4.9
	T C A A C T A T C	0.0075	3.0
	A C A C - - A G C	1.2	5.3
	A G A - - - A T C	3.3	4.9
	A C C G - - A T C	0.59	4.6
Exceptional	T G C T - - A G G	0.0023	-0.97



$$
\begin{aligned}
\text{log-odds}(\mathbf{ACACATC}) &= 1.16+0+1.16+0+1.16-0.51+ \\
&\quad 0.47-0.51+1.39+01.16+0+1.16 \\
&= 6.64.
\end{aligned}
$$

Мы нашли

- вес ACACATC = 6.64
- … и выравнивание относительно профиля:

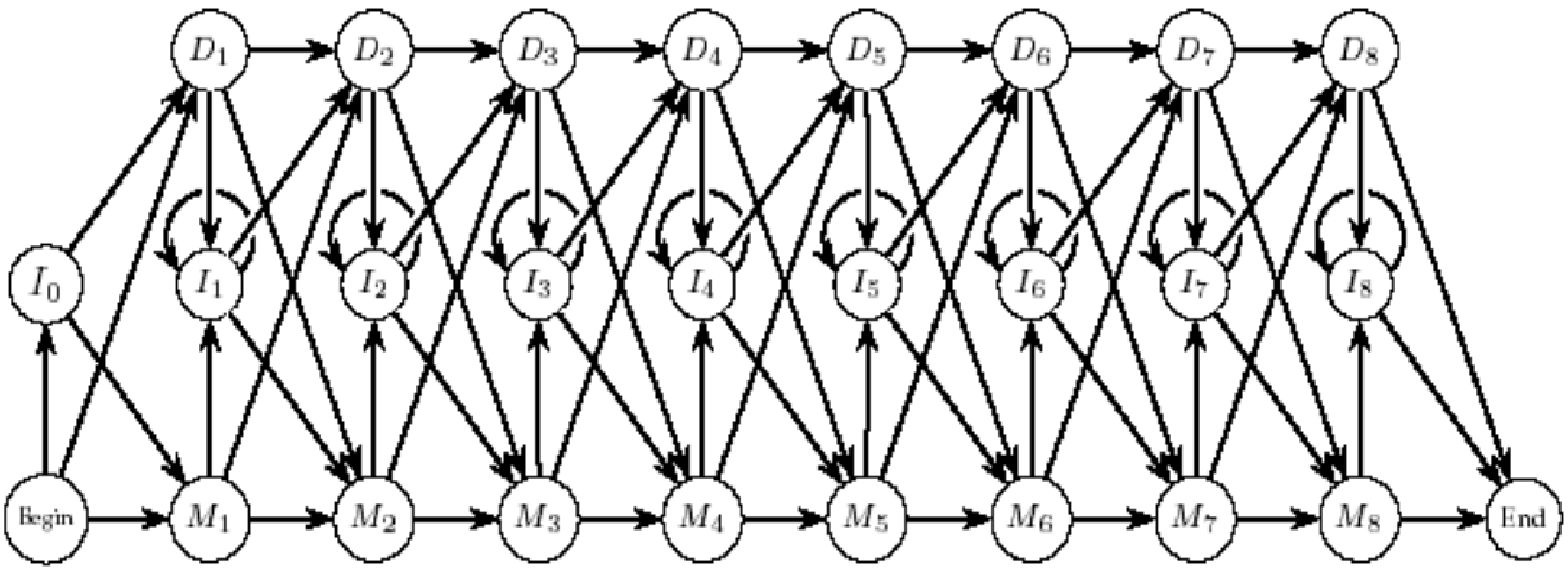
A	**C**	**A**	**C**	**A**	**T**	**C**
m1	**m2**	**m3**	**i3**	**m4**	**m5**	**m6**

Задачу нахождения лучшего по весу выравнивания входной последовательности и HMM профиля решает **<u>алгоритм Viterbi</u>**

Более сложная ситуация

- Возможны вставки (i) в любом месте
- Возможны делеции (d) в любом месте
- Разрешены все возможные переходы между вершинами b (begin), m(match), i(insertion), d(delelion), e(end):
  - b => m1, b => d1, b => i1
  - m => следующую m, m => i, m=>d, m => e
  - i => i, i => m, i => d, i => e
  - d => d, d => m, d => i, d => e

Граф HMM для выравнивания, в котором восемь колонок без гэпов



Из презентации безымянного сотрудника ИППИ)

HMM профиль, построенный HMMer'ом

log-odds(эмиссионных вероятностей для **m**)

log(вероятностей переходов

log-odds(эмиссионных вероятностей для **i)**

	A	C	D	E	F	G	H	I	K	L	M
	m->m	**m->i**	**m->d**	**i->m**	**i->I**	**d->m**	**d->d**	**b->m**	**m->e**		
	-126	*	-3585								
1	**-3610**	**-3114**	**-6053**	**-5506**	**2082**	**-5684**	**-4554**	**1759**	**-5277**	**2345**	**-632**
-	**-149**	**-500**	**233**	**43**	**-381**	**399**	**106**	**-626**	**210**	**-466**	**-720**
-	**-6**	**-8606**	**-9649**	**-894**	**-1115**	**-701**	**-1378**	**-126**	*****		
2	604	2386	-4230	-3967	-3020	-2605	-3120	685	-3662	-2921	-2216
-	-149	-500	233	43	-381	399	106	-626	210	-466	-720
-	-6	-8606	-9649	-894	-1115	-701	-1378	*	*		
3	**595**	**-2622**	**-4509**	**-4862**	**-5190**	**3595**	**-4388**	**-5082**	**-4974**	**-5307**	**-4405**
-	**-149**	**-500**	**233**	**43**	**-381**	**399**	**106**	**-626**	**210**	**-466**	**-720**
-	**-6**	**-8606**	**-9649**	**-894**	**-1115**	**-701**	**-1378**	*****	*****		
4	-4592	-3891	-6106	-6010	4096	-5830	-2943	-1896	-5700	1283	-1205
-	-149	-500	233	43	-381	399	106	-626	210	-466	-720
-	-6	-8606	-9649	-894	-1115	-701	-1378	*	*		
5	**403**	**-1180**	**-3654**	**-3023**	**2363**	**-2897**	**-1771**	**922**	**-2629**	**268**	**-383**
-	**-149**	**-500**	**233**	**43**	**-381**	**399**	**106**	**-626**	**210**	**-466**	**-720**
-	**-6**	**-8606**	**-9649**	**-894**	**-1115**	**-701**	**-1378**	*****	*****		
6	-3348	-5115	3925	-1340	-5451	-3081	-2608	-5586	-3075	-5406	-4883
-	-149	-500	233	43	-381	399	106	-626	210	-466	-720
-	-6	-8606	-9649	-894	-1115	-701	-1378	*	*		
7	**2841**	**-2218**	**-4381**	**-4396**	**-4354**	**1529**	**-3793**	**-4064**	**-4191**	**-4344**	**1956**
-	**-149**	**-500**	**233**	**43**	**-381**	**399**	**106**	**-626**	**210**	**-466**	**-720**
-	**-6**	**-8606**	**-9649**	**-894**	**-1115**	**-701**	**-1378**	*****	*****		

Для нормализации веса и вычисления E-value находок проводят калибровку HMM профиля на множестве случайных последовательностей

Профиль pftools для C2H2 из Prosite

```
/GENERAL_SPEC: ALPHABET='ABCDEFGHIKLMNPQRSTVWYZ'; LENGTH=28;
/DISJOINT: DEFINITION=PROTECT; N1=3; N2=26;
/NORMALIZATION: MODE=1; FUNCTION=LINEAR; R1=-0.6689; R2=0.02078310; TEXT='-LogE';
/CUT_OFF: LEVEL=0; SCORE=441; N_SCORE=8.5; MODE=1; TEXT='!';
/CUT_OFF: LEVEL=-1; SCORE=344; N_SCORE=6.5; MODE=1; TEXT='?';
/DEFAULT: D=-20; I=-20; B1=-50; E1=-50; MI=-105; MD=-105; IM=-105; DM=-105;

             A   B   C   D   E   F   G   H   I   K   L   M   N   P   Q   R   S   T   V   W   Y   Z
/I:        B1=0; BI=-105; BD=-105;
.............
/M: SY='C'; M=-10,-20,118,-30,-30,-20,-30,-30,-30,-30,-20,-20,-20,-40,-30,-30,-10,-10,-10,-50,-30,-30;
/M: SY='E'; M= -5,  3,-24,  3,  6,-22,-11, -6,-20,  1,-21,-14,  4, -1,  1, -3,  5,  2,-18,-29,-15,  3;
/I:        I=-12; MI=0; MD=-30; IM=0; DM=-30;
/M: SY='E'; M= -9, -2,-26,  1, 14,-18,-17, -4,-13, -1,-11, -8, -5,-12,  4, -5, -5, -8,-12,-24, -9,  8;
/M: SY='C'; M=-10,-20,119,-30,-30,-20,-30,-30,-30,-30,-20,-20,-20,-40,-30,-30,-10,-10,-10,-50,-29,-30;
/M: SY='G'; M= -3, -1,-28, -1, -7,-28, 36,-11,-33,-11,-27,-18,  4,-15,-10,-12,  1,-13,-27,-24,-23, -9;
/M: SY='K'; M=-10, -2,-28, -3,  8,-25,-19, -7,-26, 36,-24, -8, -1,-12, 10, 27, -9, -9,-18,-19, -8,  8;
/M: SY='A'; M=  8, -7, -9,-11, -7,-17, -7,-14,-16, -6,-16,-11, -4,-15, -6, -5,  8,  4, -7,-27,-15, -7;
/M: SY='F'; M=-19,-29,-19,-37,-28, 71,-29,-17,  0,-28,  9,  0,-20,-30,-36,-19,-19, -9, -1,  9, 31,-28;
................
/M: SY='H'; M=-20,  0,-30,  0,  0,-20,-20, 99,-30,-10,-20,  0, 10,-20, 10,  0,-10,-20,-30,-30, 20,  0;
/M: SY='Q'; M=-10,-10,-25,-12,  1,-16,-22, -2, -6,  1, -3,  6, -9,-17, 13,  3, -9, -8, -9,-19, -4,  6;
/M: SY='R'; M=-13, -8,-26, -9,  0,-19,-19, -4,-21, 20,-16, -6, -2,-17,  6, 35, -8, -7,-14,-21, -9,  0;
/I:        I=-12; MI=0; MD=-29; IM=0; DM=-29;
/M: SY='V'; M= -3,-16,-17,-21,-17, -6,-25,-20, 11,-15,  2,  3,-12,-18,-14,-14, -2,  9, 13,-25, -7,-17;
/M: SY='H'; M=-20,  0,-30,  0,  0,-20,-20, 97,-30,-10,-20,  0, 10,-20, 10,  0,-10,-20,-30,-30, 19,  0;
...................
/I:        E1=0;
```

C-x(2,4)-C-x(3)-[LIVMFYWC]-x(8)-H-x(3,5)-H

Профиль Pftools

- Используется в Prosite, БД MyHits. Доступен как stand alone пакет
- Не использует HMM – менее обоснован теоретически
- Зато веса букв понятны, поэтому профиль можно редактировать вручную
- Есть конверторы профиль Pftools <=> Профиль HMM

4. Интерпретация результатов поиска по профилю

Профиль

- Служит для предсказания принадлежности последовательности семейству
- Оценивает числом – E-value (или нормализованным весом) – сходство последовательности и профиля
- Чтобы получить предсказание необходимо выбрать порог E или веса T: **E < e** (=0.001) (или **T > t** (= 10 ?)) <=> последовательность принадлежит семейству
- Проверку профиля и выбор порога следует выполнять на множестве последовательностей с известным ответом (ROC-кривая); если, конечно, такие есть)
- Часто между последовательностями “точно, принадлежит” и “точно, не из семейства” есть “серая зона”, зона неопределенности
- Скачек веса как один из признаков для выбора порога

HMMer search параметры

- -E 0.1 (порог E-value находки)
- -T 20   (порог веса находки)

Проверка профиля на множестве последовательностей <u>с известным ответом</u> про каждую последовательность

- Выберем порог **t**
- Тогда <u>предсказывается,</u> что находка
  - принадлежит семейству, если ее вес **T >= t**
  - не принадлежит, если **T < t**
- (аналогично для E-value)

Таблица проверки предсказания

		Заболевание		
		Присутствует (Positive)	Отсутствует (Negative)	
Тест	Положительный (Positive)	True Positive, TP	False Positive, FP	TP+FP
	Отрицательный (Negative)	False Negative, FN	True Negative, TN	FN+TN
		TP+FN	FP+TN	

Характеристики предсказания

Чувствительность (sensitivity): доля позитивных результатов теста в группе больных пациентов

Специфичность (specificity): доля негативных результатов теста в группе здоровых пациентов

Учёные люди знают еще много параметров, которые можно извлечь из таблицы 2x2 (справа)

sensitivity or true positive rate (TPR)

eqv. with hit rate, recall

$$TPR = TP/P = TP/(TP + FN)$$

specificity (SPC) or true negative rate (TNR)

$$SPC = TN/N = TN/(TN + FP)$$

precision or positive predictive value (PPV)

$$PPV = TP/(TP + FP)$$

negative predictive value (NPV)

$$NPV = TN/(TN + FN)$$

fall-out or false positive rate (FPR)

$$FPR = FP/N = FP/(FP + TN)$$

false negative rate (FNR)

$$FNR = FN/(FN + TP) = 1 - TPR$$

false discovery rate (FDR)

$$FDR = FP/(FP + TP) = 1 - PPV$$

---

accuracy (ACC)

$$ACC = (TP + TN)/(P + N)$$

F1 score

is the harmonic mean of precision and sensitivity

$$F1 = 2TP/(2TP + FP + FN)$$

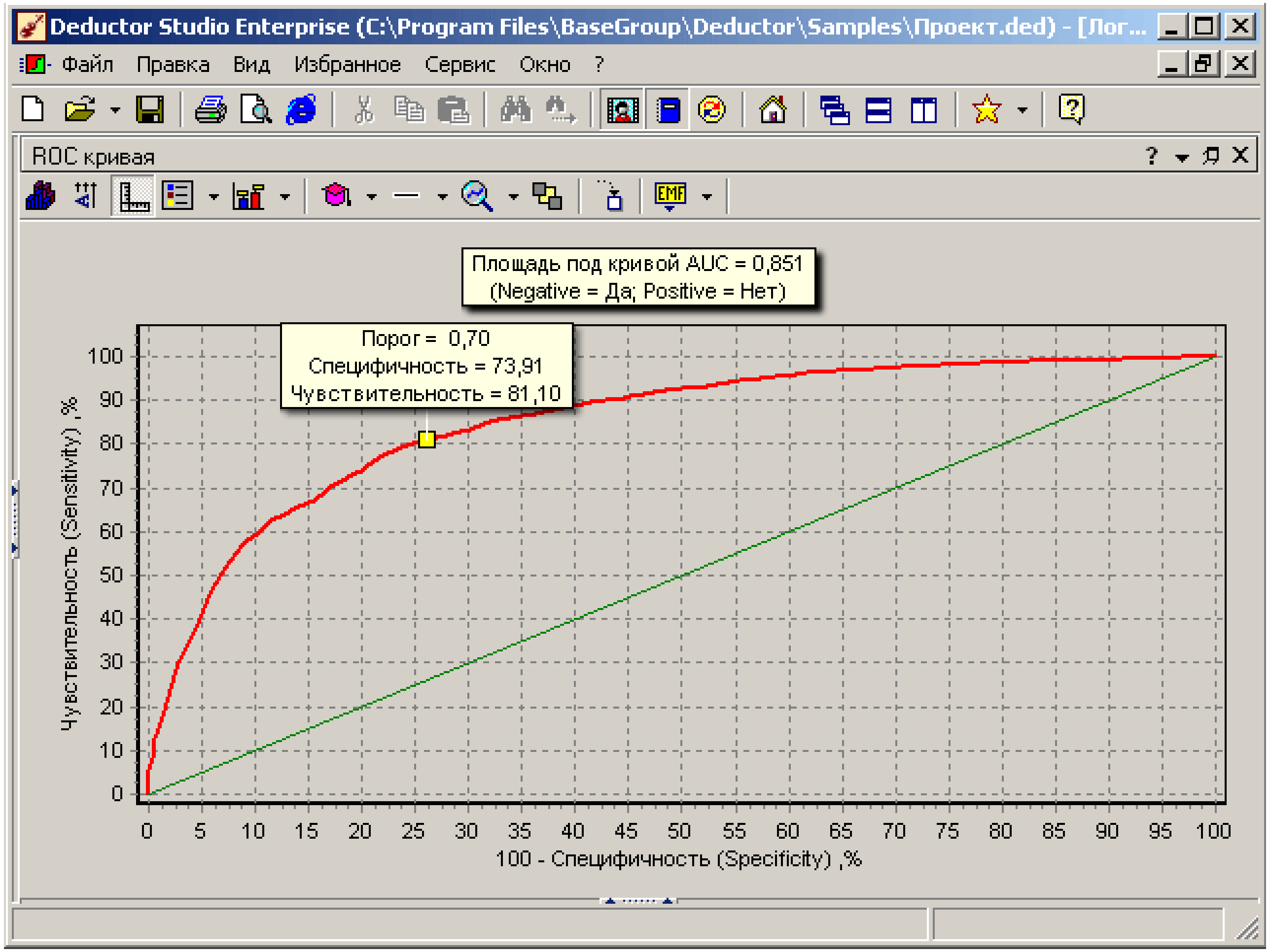
Matthews correlation coefficient (MCC)

$$\frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Sources: Fawcett (2006) and Powers (2011).[4][1]

Как выбрать порог? ROC-кривая

(англ. *receiver operating characteristic, операционная характеристика приёмника*)

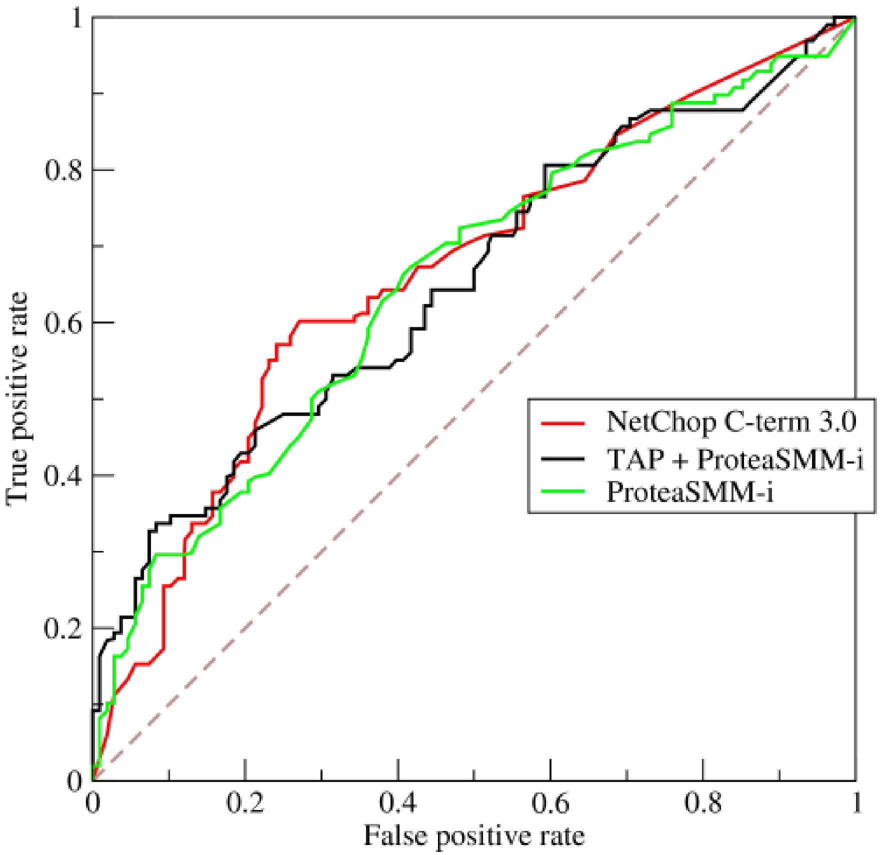


ROC-кривая

(англ. *receiver operating characteristic*, *операционная характеристика приёмника*)

- Строится в том случае, когда предсказание основано на вычислении числа, например, нормализованного веса находки профиля
- Предсказание должно быть проверено на данных с ИЗВЕСТНЫМ ОТВЕТОМ.
- Удобна для выбора порога t: если нормализованный вес больше t, то предсказываем принадлежность семейству
- Также используют для сравнения разных правил предсказания (площадь под кривой)
- Следует помнить, что ROC-кривая имеет смысл только при разумных значениях порога; разумность определяется задачей

Пример сравнения



ROC-кривые трёх методов предсказания эпитопов

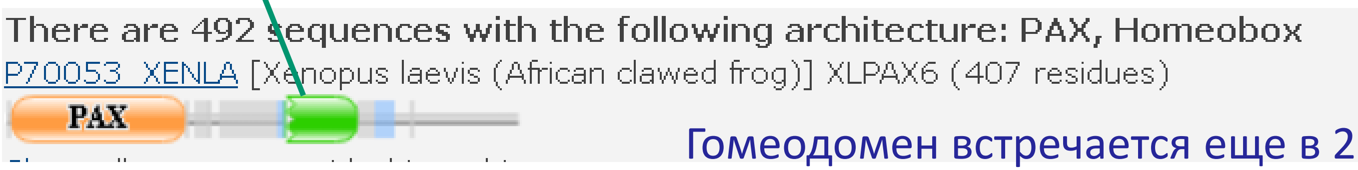
Построение ROC кривой

- Результаты поиска отсортировать по убыванию по нормализованному весу (первая колонке выдачи pfsearch) , добавить заголовки столбцов
- Добавить столбец с отметкой правильных находок буквой “Y”; используйте vlookup (ВПР)
- На отдельном листе сделать две колонки: 1 – специфичность (ось X) и чувствительность (ось Y)
- Написать формулы для расчета значений оси X и Y; формула в i-й строке считает, что первые i находок предсказываются принадлежащими семейству, а все – остальные – не принадлежащими семейству. Используйте команду countif (счётесли)

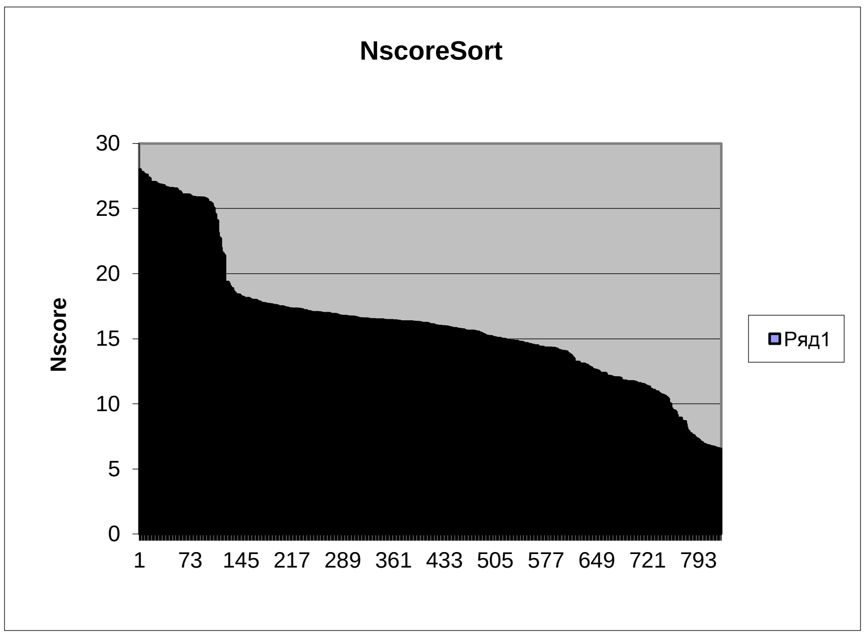
Ступенька нормализованного веса

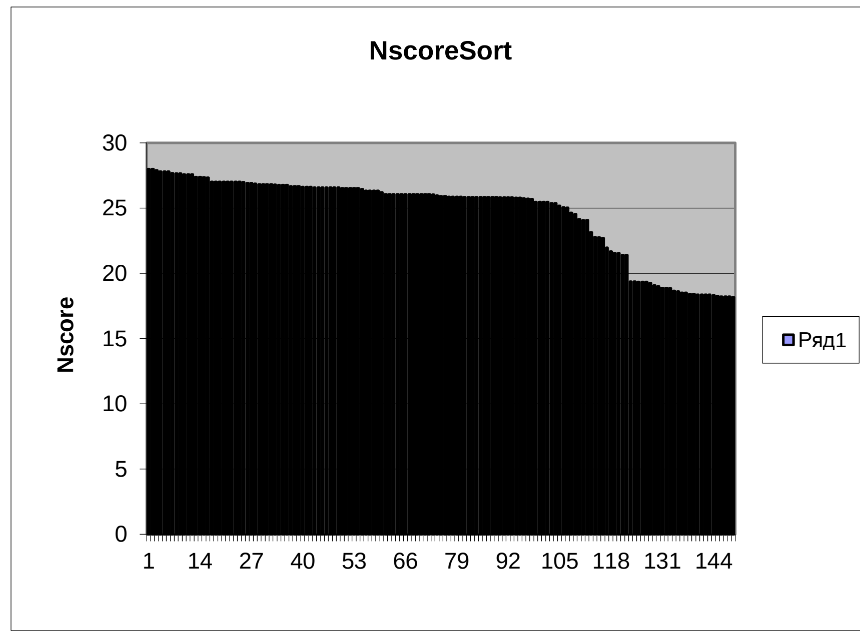
Пример: Paired-like homeodomain family

Гомеодомен



Гомеодомен встречается еще в 289 архитектурах





КОНЕЦ

Порог Nscore = 21