

Множественные выравнивания

- ✓ Зачем все это нужно?
- ✓ Глобальные множественные выравнивания – основы алгоритма, программы
- ✓ Где искать на Web?
- ✓ Можно ли редактировать множественное выравнивание?
- ✓ Локальные множественные выравнивания

Что такое множественное выравнивание?

Несколько гомологичных последовательностей, написанных друг под другом оптимальным способом:

- ✓ Гомологичные остатки один под другим
- ✓ Остатки в одинаковом пространственном положении один под другим
- ✓ Остатки, имеющие одинаковую функциональную нагрузку, один под другим
- ✓ Одинаковые или похожие остатки один под другим

Какое выравнивание интереснее?

```

                                *                20
XYLR_ECOLI : GYPSLQYFYSVFKKAYDTTPKEYR : 24
XYLR_HAEI N : GYPSI QYFYSVFKKEFEMTPKEFR : 24
```

```

                                *                20
ADI Y_ECOLI : GYNSTSYFI SVFKDFYGMPLHYV : 24
APPY_ECOLI : GYNSTSYFI CAFKDYYGVTPSHYF : 24
CELD_ECOLI : GYSSPSLFI KTFKKLTSFTPKSYR : 24
CFAD_ECOLI : GISSASYFI RVFNKHYGVTPKQFF : 24
ENVY_ECOLI : GYSSSTSYFI SVFKAFYGLTPLNYL : 24
FAPR_ECOLI : GYTSVSYFI KTFKEYYGVTPKKFE : 24
MELR_ECOLI : CFRSSSRFYSTFGKYVGMSPPQQYR : 24
RHAS_ECOLI : CFSDSNHFFSTLFRREFNWSPRDI R : 24
ROB_ECOLI / : RFDSQQTFTRAFKKQFAQTPALYR : 24
TETD_ECOLI : QFDSQQSFTRRFKYI FKVTPSYR : 24
XYLR_ECOLI : GYPSLQYFYSVFKKAYDTTPKEYR : 24
XYLR_HAEI N : GYPSI QYFYSVFKKEFEMTPKEFR : 24
```

Какие бывают выравнивания?

Выравнивания

парные

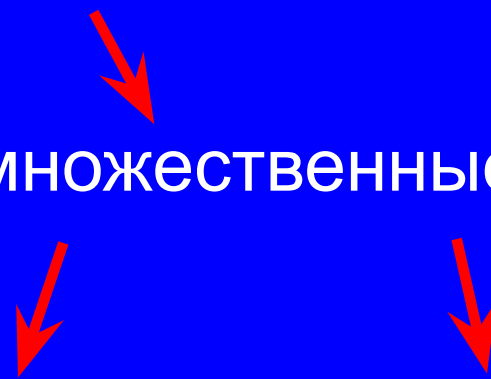
множественные

глобальные

локальные

глобальные

локальные



Зачем нужно множественное выравнивание?

- ✓ Перенос аннотации
- ✓ Предсказание функции каждого остатка (например, выявление остатков, составляющих активный центр фермента)
- ✓ Моделирование 3D – структуры
- ✓ Реконструкция эволюционной истории последовательности (филогения)
- ✓ Выявление паттерна функциональных семейств и сигналов в ДНК
- ✓ Построение доменных профайлов
- ✓ Аккуратный дизайн праймеров для PCR анализа

Как выбрать последовательности для множественного выравнивания?

- ✓ Выравнивайте белки, а не ДНК, если есть выбор
- ✓ Последовательностей лучше много, но не слишком (~ 10-15)
- ✓ В выборке лучше избегать:
 - слишком похожих последовательностей (>90% id)
 - слишком разных последовательностей (<30% id с большинством)
 - неполных последовательностей (фрагментов)
 - тандемных повторов

Изучая новую последовательность

- ✓ Выборка на основе BLAST
- ✓ Подробно охарактеризованные последовательности - аннотация
- ✓ Совсем неохарактеризованные (hypothetical proteins) – достаточный уровень разнообразия
- ✓ Выравнивание по всей длине
- ✓ e-value – 10^{-40} – 10^{-6}
- ✓ Избегать partial sequences

Подготовка выборки

BLAST => сохранить все последовательности
разом в FASTA формате или сразу на
выравнивание

Имена последовательностей:

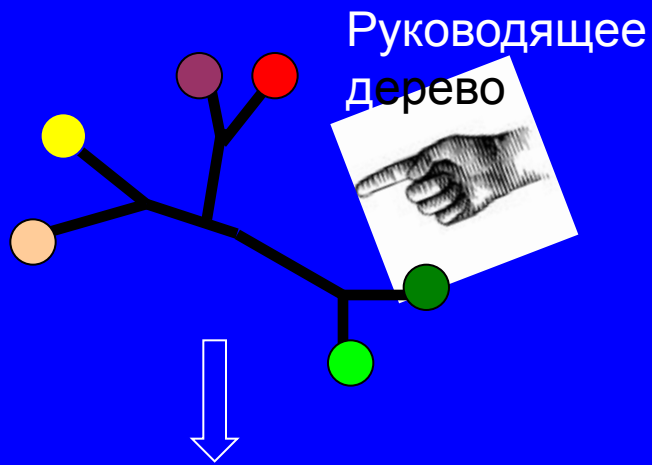
- ✓ не более 15 символов
- ✓ без пробелов
- ✓ как можно меньше служебных символов —
можно “_”
- ✓ нельзя использовать одинаковых имен!

Как можно строить глобальное множественное выравнивание?

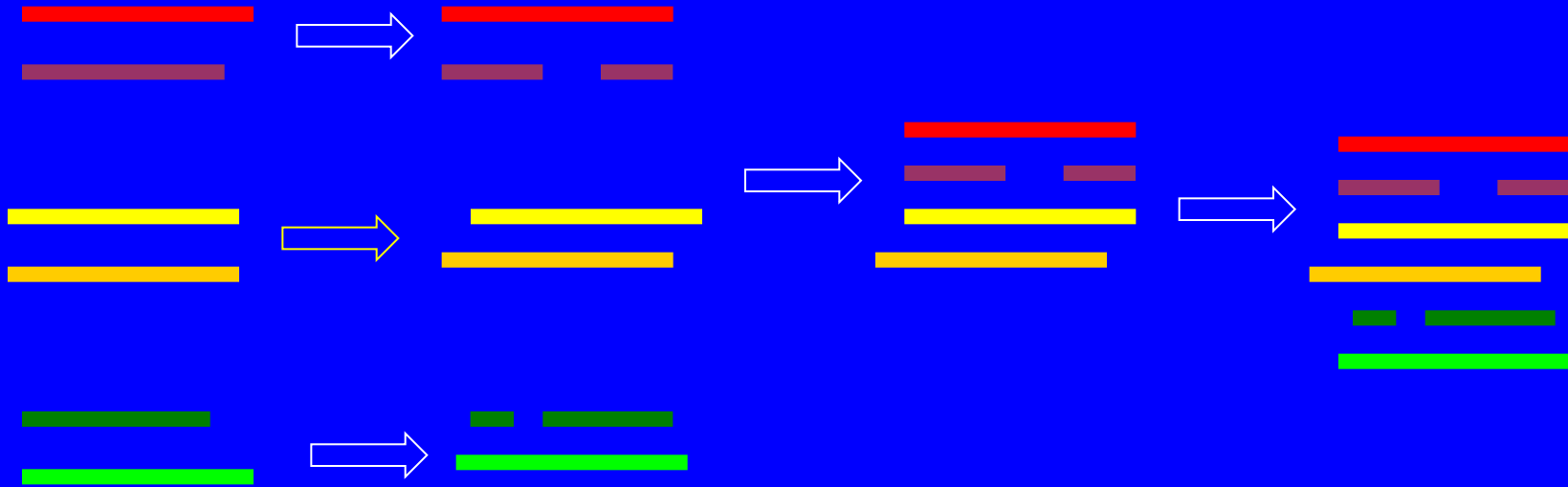
Можно пытаться строить точно также, как и парное – слева направо, максимизируя вес выравнивания по столбцам (алгоритм Нидельмана –Вунша)

Построение множественного выравнивания N последовательностей

$$t = L^N !!!$$



Алгоритм ClustalW – пример эвристического прогрессивного алгоритма



Очевидные недостатки:

- 1) Результат зависит от порядка выравниваний;
- 2) «один раз гэп – всегда гэп»

Современные методы построения множественного выравнивания (MSA, multiple sequence alignment):

- ✓ Алгоритм ClustalW (реализации **ClustalX**, **emma** из EMBOSS) – до сих пор самый популярный, но уже устаревший метод (на Web – например, <http://www.ebi.ac.uk/Tools/clustalw/index.html>)
- ✓ **Muscle** – быстрее и немного точнее, самый новый и довольно модный (http://phylogenomics.berkeley.edu/cgi-bin/muscle/input_muscle.py)
- ✓ **T-COFFEE** – заметно точнее, но существенно медленнее (http://www.igs.cnrs-mrs.fr/Tcoffee/tcoffee_cgi/index.cgi)


Использование ClustalW

ClustalW

ClustalW is a general purpose multiple sequence alignment program for DNA or proteins. It produces biologically meaningful multiple sequence alignments of divergent sequences. It calculates the best match for the selected sequences, and lines them up so that the identities, similarities and differences can be seen. Evolutionary relationships can be seen via viewing Cladograms or Phylograms.

[New users, please read the FAQ.](#)

>> [Download Software](#)



YOUR EMAIL	ALIGNMENT TITLE	RESULTS	ALIGNMENT	
<input type="text"/>	<input type="text" value="Sequence"/>	<input type="text" value="interactive"/>	<input type="text" value="full"/>	
KTUP (WORD SIZE)	WINDOW LENGTH	SCORE TYPE	TOPDIAG	PAIRGAP
<input type="text" value="def"/>	<input type="text" value="def"/>	<input type="text" value="percent"/>	<input type="text" value="def"/>	<input type="text" value="def"/>
MATRIX	GAP OPEN	END GAPS	GAP EXTENSION	GAP DISTANCES
<input type="text" value="def"/>	<input type="text" value="def"/>	<input type="text" value="def"/>	<input type="text" value="def"/>	<input type="text" value="def"/>

OUTPUT		PHYLOGENETIC TREE		
OUTPUT FORMAT	OUTPUT ORDER	TREE TYPE	CORRECT DIST.	IGNORE GAPS
<input type="text" value="aln w/numbers"/>	<input type="text" value="aligned"/>	<input type="text" value="none"/>	<input type="text" value="off"/>	<input type="text" value="off"/>

Enter or Paste a set of Sequences in any supported format:

Upload a file:

Какие output-форматы бывают

- ✓ Post-script, pdf, html – только графика
- ✓ FASTA – последовательности отдельно, но с пробелами (PIR – аналогично)
- ✓ MSF (ALN, Phylip, Selex ...) – наглядно.
Сверху – описание выборки: программа, название последовательностей, их длина, вес в выравнивании; потом само выравнивание блоками по 60 остатков

Перевод форматов: READSEQ

(<http://www-bimas.cit.nih.gov/molbio/readseq/>)

WWW READSEQ Sequence Conversion

Function: Converts input DNA/AA sequence to specified format (Input format is determined automatically).

As of Feb. 3, 2006, Sequence Conversion uses the JAVA version of the READSEQ program (Readseq version 2.1.22 (02-May-2005), READSEQ is maintained at the [IUbio Archive](#) site at University of Indiana.

Pearson/Fasta **Format for the output** (Use this [form](#) for 'Pretty' format)

Additional formatting options:

Mixed Case Lower Case Upper Case

Remove Gap Symbols ('-')


Please enter or paste sequence[s] to be converted (most [formats](#) accepted):

Credits: WWW implementation by [BIMAS Staff](#)

Аналогично: SEQCHECK

ClustalW - output

ClustalW Results

Results of search	
Number of sequences	5
Alignment score	18810
Sequence format	Pearson
Sequence type	aa
ClustalW version	1.83
JalView	Start Jalview 
Output file	clustalw-20071126-12304771_output
Alignment file	clustalw-20071126-12304771aln
Guide tree file	clustalw-20071126-12304771.dnd
Your input file	clustalw-20071126-12304771.input

[SUBMIT ANOTHER JOB](#)

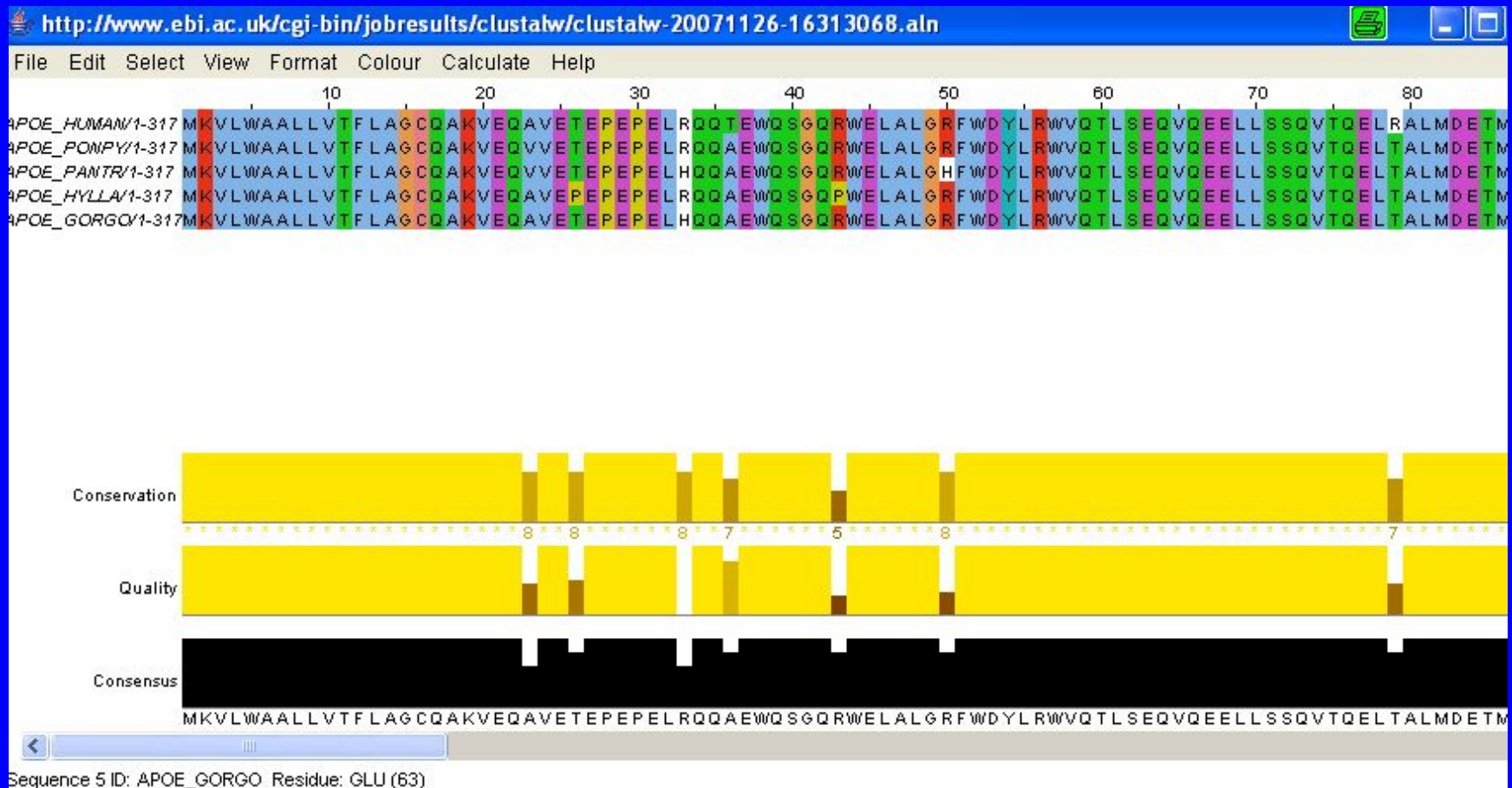
To save a result file right-click the file link in the above table and choose "Save Target As".
If you cannot see the JalView button, reload the page and check your browser settings to enable Java Applets.

Scores Table

Sort by

SeqA Name	Len(aa)	SeqB Name	Len(aa)	Score
1 APOE_HUMAN	317	2 APOE_PONPY	317	264
1 APOE_HUMAN	317	3 APOE_PANTR	317	264
1 APOE_HUMAN	317	4 APOE_GORGO	317	264
1 APOE_HUMAN	317	5 APOE_HYLLA	317	256
2 APOE_PONPY	317	2 APOE_PONPY	317	271

JaView – редактирование выравниваний



http://www.ebi.ac.uk/cgi-bin/jobresults/clustalw/clustalw-20071126-16313068.aln

File Edit Select View Format Colour Calculate Help

10 20 30 40 50 60 70 80

APOE_HUMAN/1-317 MKVLWAALLVTF LAGCQAKVEQAVET EPEPELRQQTEWQSGQRWELALGRFWDYLRWVQTLSEQVQEELLSSQVTQELRALMDETM

APOE_PONPY/1-317 MKVLWAALLVTF LAGCQAKVEQVVE T EPEPELRQQAEWQSGQRWELALGRFWDYLRWVQTLSEQVQEELLSSQVTQELTALMDETM

APOE_PANTR/1-317 MKVLWAALLVTF LAGCQAKVEQVVE T EPEPELRQQAEWQSGQRWELALGRFWDYLRWVQTLSEQVQEELLSSQVTQELTALMDETM

APOE_HYLLA/1-317 MKVLWAALLVTF LAGCQAKVEQAVET EPEPELRQQAEWQSGQPWELALGRFWDYLRWVQTLSEQVQEELLSSQVTQELTALMDETM

APOE_GORGO/1-317 MKVLWAALLVTF LAGCQAKVEQAVET EPEPELRQQAEWQSGQRWELALGRFWDYLRWVQTLSEQVQEELLSSQVTQELTALMDETM

Conservation

Quality

Consensus

MKVLWAALLVTF LAGCQAKVEQAVET EPEPELRQQAEWQSGQRWELALGRFWDYLRWVQTLSEQVQEELLSSQVTQELTALMDETM

Sequence 5 ID: APOE_GORGO Residue: GLU (63)

Другие программы для редактирования выравниваний (stand-alone):
GeneDoc; CINEMA; Seaview; Belvu; Bioedit; DCSE
Список - <http://bioweb.pasteur.fr/cgi-bin/seqanal/review-edital.pl>

TCoffee

- ✓ Построение множественных выравниваний
- ✓ Оценка достоверности существующего выравнивания
- ✓ Использование 3-D структуры при построении выравнивания
- ✓ Сравнение и комбинирование выравниваний

TCoffee

[HOME](#) | [references](#) | [help](#) | 

TCoffee

A collection of tools for Computing, Evaluating and Manipulating Multiple Alignments of DNA, Protein Sequences and Structures

Mirror sites:       

ALIGNMENT				
TCOFFEE	<input type="button" value="Regular"/>	<input type="button" value="Advanced"/>	cite	?
EXPRESSO(3DCoffee)	<input type="button" value="Regular"/>	<input type="button" value="Advanced"/>	cite	?
MCOFFEE	<input type="button" value="Regular"/>	<input type="button" value="Advanced"/>	cite	?
COMBINE	<input type="button" value="Regular"/>	<input type="button" value="Advanced"/>	cite	?
EVALUATION				
CORE	<input type="button" value="Regular"/>	<input type="button" value="Advanced"/>	cite	?
iRMSD-APDB	<input type="button" value="Regular"/>	<input type="button" value="Advanced"/>	cite	?
PROCESSING				
PROTOGENE	<input type="button" value="Regular"/>	<input type="button" value="Advanced"/>	cite	?

Mirror sites:       

Выход – файлы [clustalw_aln](#) Выход – файлы [clustalw_aln](#), [fasta_aln](#) Выход – файлы [clustalw_aln](#), [fasta_aln](#), [phyloip](#) Выход – файлы [clustalw_aln](#), [fasta_aln](#).

Как использовать Toffee для других целей

- Множественное выравнивание на основе 3D-структуры (Expresso): надо заменить 1 или более имен в FASTA формате последовательностей на PDB-идентификатор соответствующей структуры. Тест – “Template file” (число структур). Если не в PDB – “Advanced”
- Alignment evaluation – готовое выравнивание на вход. На выходе – раскрашенное выравнивание (score.html, score.pdf): каждый столбец покрашен в соответствии с качеством – красный/оранжевый/желтый - хорошо

Как “читать” множественное выравнивание?

- ✓ Хорошее выравнивание – высоко-консервативные блоки, перемежающиеся блоками с инсерциями/делециями
- ✓ ДНК – консервативные “островки”
- ✓ Качество – score, локально важно
- ✓ “consensus” – строка с символами “*”, “.”, “.” – консервативный, похожие по размеру и гидропатичности, похожие по размеру ИЛИ гидропатичности, соответственно

Если консервативны только отдельные столбцы

- ✓ W, Y, F – консервативное гидрофобное ядро, стабилизирующая роль в ядре. Если и мутируют, то между собой
- ✓ G, P – фланкируют бета-стренды и альфа-спирали
- ✓ C – участвует в образовании дисульфидных мостиков – одинаковое расстояние между
- ✓ H, S – каталитические центры протеаз
- ✓ K, R, D, E – заряженные аминокислоты, участвуют в связывании лигандов
- ✓ L – редко консервативны. Формируют leucine zipper – белок-белковые взаимодействия

Локальное множественное выравнивание – постановка задачи

Ряд последовательностей, в каждой из которых есть интересное слово (либо точно, либо с небольшим количеством замен) известной длины

=> Найти и описать это слово

Идея. Будем искать перепредставленное слово. Стартуем со всех слов в выравнивании, ищем лучшее его представление в каждой из последовательностей и потом уточняем по полученному профайлу

Как это выглядит

<i>dnaN</i>	ACATTATCCGTTAGGAGGATAAAAATG
<i>gyrA</i>	GTGATACTTCA G GGAGGTTTTTTAATG
<i>serS</i>	TCAATAAAAAAAGGAGT T GTTTCGCATG
<i>bofA</i>	CAAGCGAAGGAGA T TGAGAAGATTTCATG
<i>csfB</i>	GCTAACTGTAC C GGAGGTGGAGAAGATG
<i>xpaC</i>	ATAGACACAGGAGT T CGATTATCTCATG
<i>metS</i>	ACATTCTGATTAGGAGGTTTCAAGATG
<i>gcaD</i>	AAAAGGGATAT T GGAGGCCAATAAATG
<i>spoVC</i>	TATGTGACTAA G GGAGGATTCGCCATG
<i>ftsH</i>	GCTTACTGTGGGAGGAGGTAAGGAATG
<i>pabB</i>	AAAGAAAAT A GAGGAATGATACAAATG
<i>rplJ</i>	CAAGAATCTACAGGAGGTGTAACCATG
<i>tufA</i>	AAAGCTCTTAAGGAGGATTTTAGAATG
<i>rpsJ</i>	TGTAGGCGAAAAGGAGGGAAAATAATG
<i>rpoA</i>	CGTTTTGAAGGAGGGTTTTAAGTAATG
<i>rplM</i>	AGATCATTTAGGAGGGGAAATTCAATG
Cons	tacataaaggagggtttaa a aat

Gibbs sampler

Let's A be a signal (set of sites), and $I(A)$ be its information content.

At each step a new site is selected in one sequence with probability

$$P \sim \exp [I(A_{\text{new}})]$$

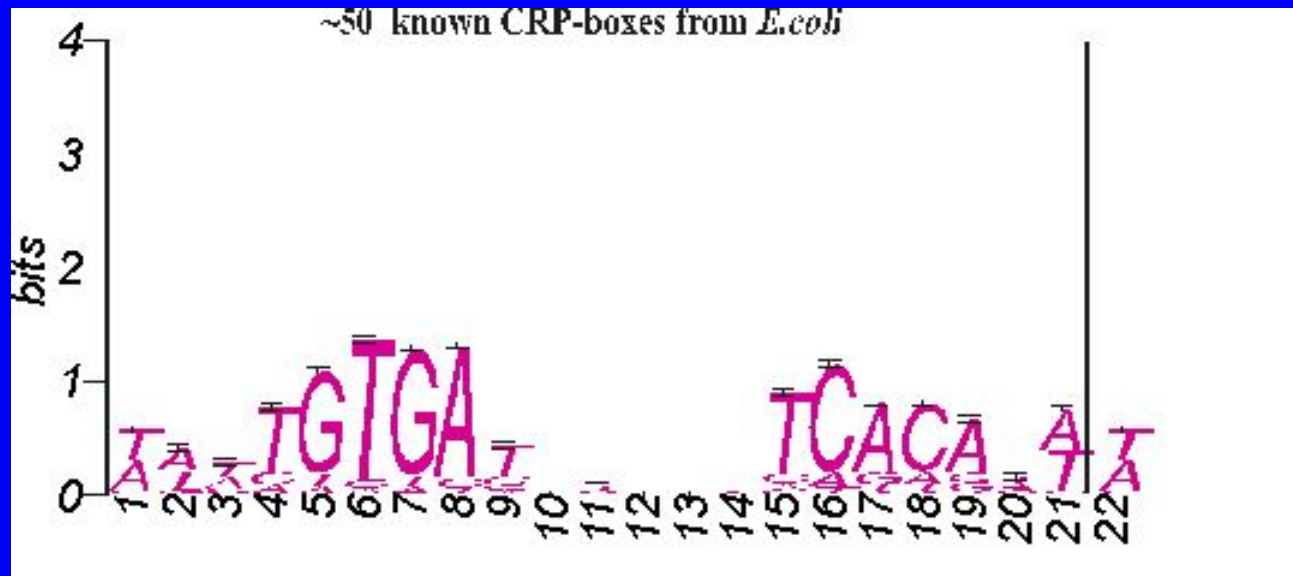
For each candidate site the total time of occupation is computed.

(Note that the signal changes all the time)

Соответствующие программы

Название программы	Адрес(а)
Gibbs Sampler	http://bioweb.pasteur.fr/seqanal/interfaces/gibbs-simple.html http://bayesweb.wadsworth.org/gibbs/gibbs.html/
Pratt	http://www.ebi.ac.uk/pratt/
eMotif	http://motif.stanford.edu/distributions/emotif/
MEME	http://meme.sdsc.edu/meme/meme.html
TEIRESIAS	http://cbcsrv.watson.ibm.com/Tspd.html
Bioprospector	http://robotics.stanford.edu/~xsliu/BioProspector/
Improbizer	http://www.soe.ucsc.edu/~kent/improbizer/improbizer.html
BLOCK-Maker	http://blocks.fhcrc.org/blocks/blockmkr/make_blocks.html

Представление результатов таких программ – Logos



Программы построения –

<http://www-lmmb.ncifcrf.gov/~toms/sequencelogo.html>;

<http://www.cbs.dtu.dk/~gorodkin/appl/plogo.html>

Greedy algorithms (MEME)

Find a signal among all k -words (assuming that we know the length signal).

For all k -words it's too time-consuming ($k \sim 16$). So initially we consider only k -words that were present in the fragments.

For each k -word construct a matrix of "sites": alignment of best "copies" of the k -word from every sequence fragment.

Select the best k -word. What is the measure for comparison of matrices? Information content!

Greedy algorithms. Cont'd

- Select the k-word with maximal information content

Problem. We considered only k-words from our sequences => may select not the signal (the consensus word), but only its best representative in our sample

Solution. For each k-word from the sample construct PWM and reconstruct the frequency matrix based on it. Repeat until stabilization of the matrix. Use the consensus of this matrix.

Limitation of greedy algorithms

- Started from k-words in our sequences and increase the information content at each step => find a local (not global) maximum of the functional.
- We need an alternative algorithm that will not be “greedy”!

Frequency matrix

<i>j</i>	a	C	G	m	A	A	A	C	G	t	T	T	k	C	k	T
A	6	0	0	2	9	9	8	0	0	1	0	0	0	0	0	0
C	1	8	0	7	0	0	1	9	0	0	0	0	0	9	1	0
G	1	1	9	0	0	0	0	0	9	1	1	0	5	0	5	0
T	1	0	0	0	0	0	0	0	0	7	8	9	4	0	3	9

Information content $I = \sum_j \sum_b f(b,j) [\log f(b,j) / p(b)]$

$$W(b,j) = \ln(N(b,j)+0.5) - 0.25 \sum_i \ln(N(i,j)+0.5)$$