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:

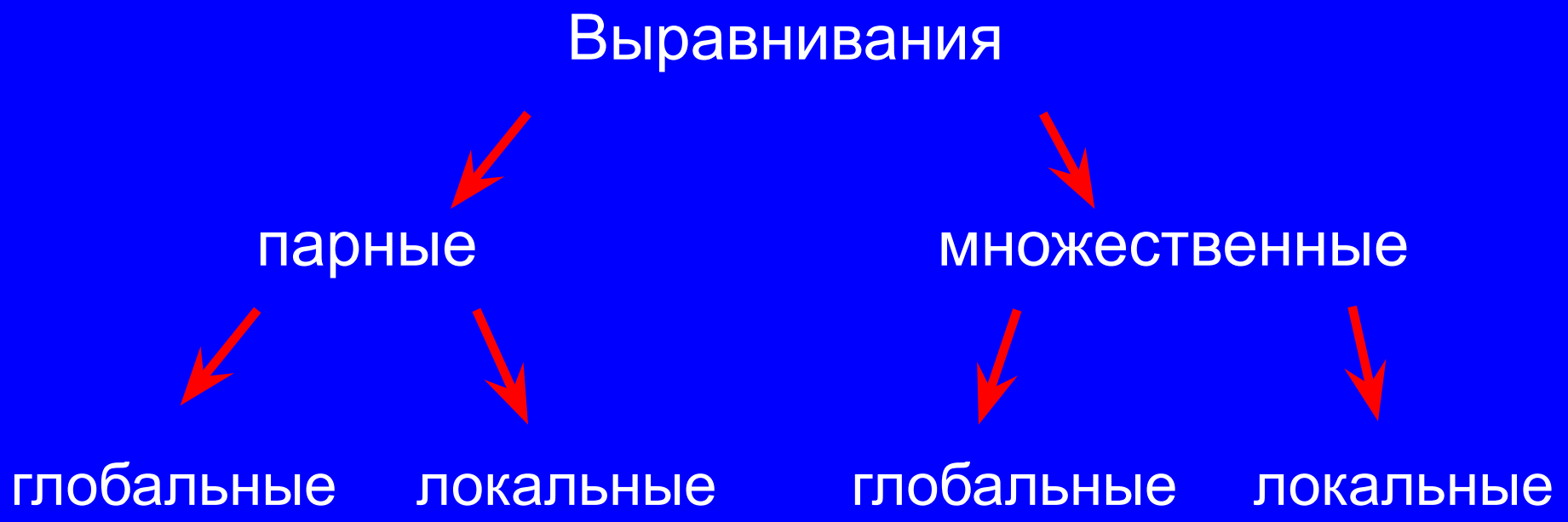
- ✔ Гомологичные остатки один под другим
- ✔ Остатки в одинаковом пространственном положении один под другим
- ✔ Остатки, имеющие одинаковую функциональную нагрузку, один под другим
- ✔ Одинаковые или похожие остатки один под другим

Какое выравнивание интереснее?

```
                                  *          20
XYLR_ECOLI  :  GYPSLQYFYSVFKKAYDTTPKEYR  :  24
XYLR_HAEIN  :  GYPSIQYFYSVFKKEFEMTPKEFR  :  24
```

```
                                  *          20
ADIY_ECOLI  :  GYNSTSYFISVFKDFYGMTPLHYV  :  24
APPY_ECOLI  :  GYNSTSYFICAFKDYYGVTPSHYF  :  24
CELD_ECOLI  :  GYSSPSLFIKTFKKLTSFTPKSYR  :  24
CFAD_ECOLI  :  GISSASYFIRVFNKHYGVTPKQFF  :  24
ENVY_ECOLI  :  GYSSTSYFISVFKAFYGLTPLNYL  :  24
FAPR_ECOLI  :  GYTSVSYFIKTFKEYYGVTPKKFE  :  24
MELR_ECOLI  :  GFRSSSRFYSTFGKYVGMSPQQYR  :  24
RHAS_ECOLI  :  GFSDSNHFSTLFRREFNWSPRDIR  :  24
ROB_ECOLI/  :  RFDSQQTFTRAFKKQFAQTPALYR  :  24
TETD_ECOLI  :  QFDSQQSFTRRFKYIFKVTPSYYR  :  24
XYLR_ECOLI  :  GYPSLQYFYSVFKKAYDTTPKEYR  :  24
XYLR_HAEIN  :  GYPSIQYFYSVFKKEFEMTPKEFR  :  24
```

Какие бывают выравнивания?



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!

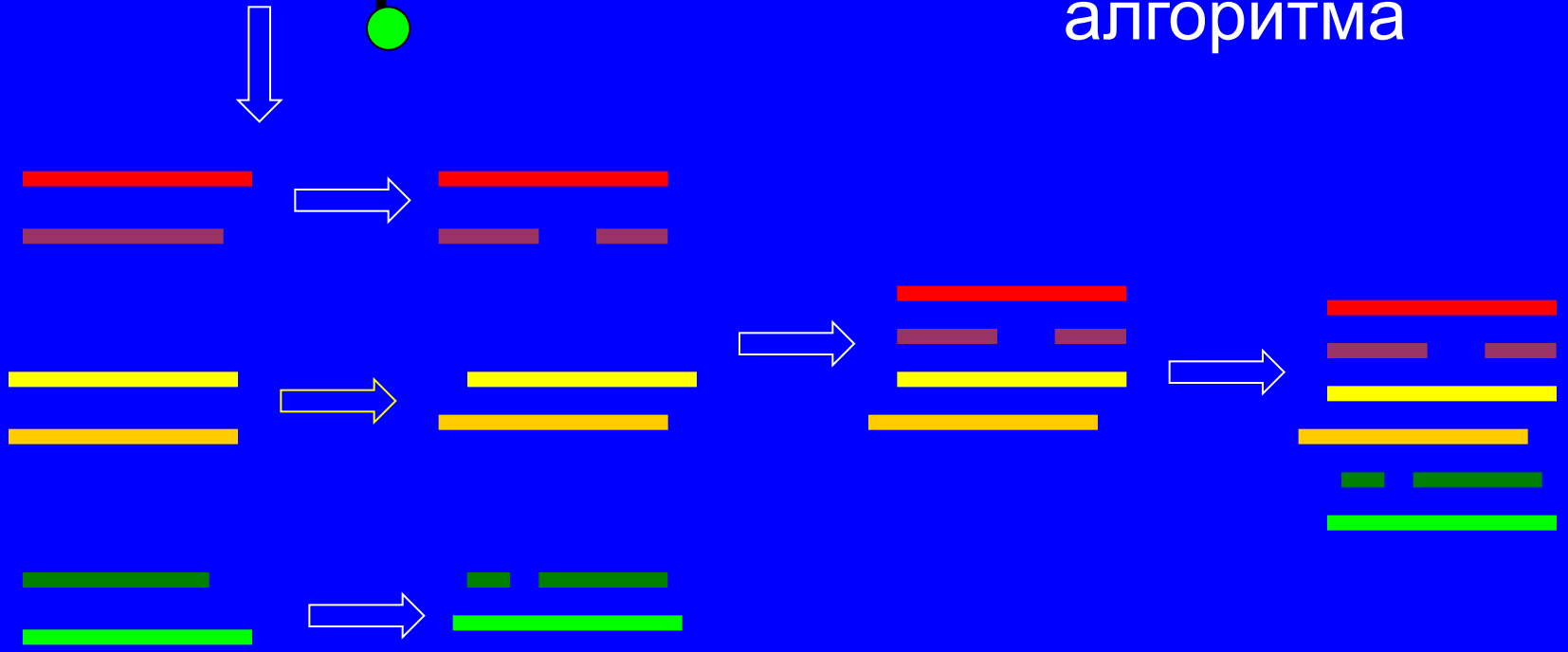
Как можно строить глобальное множественное выравнивание?

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



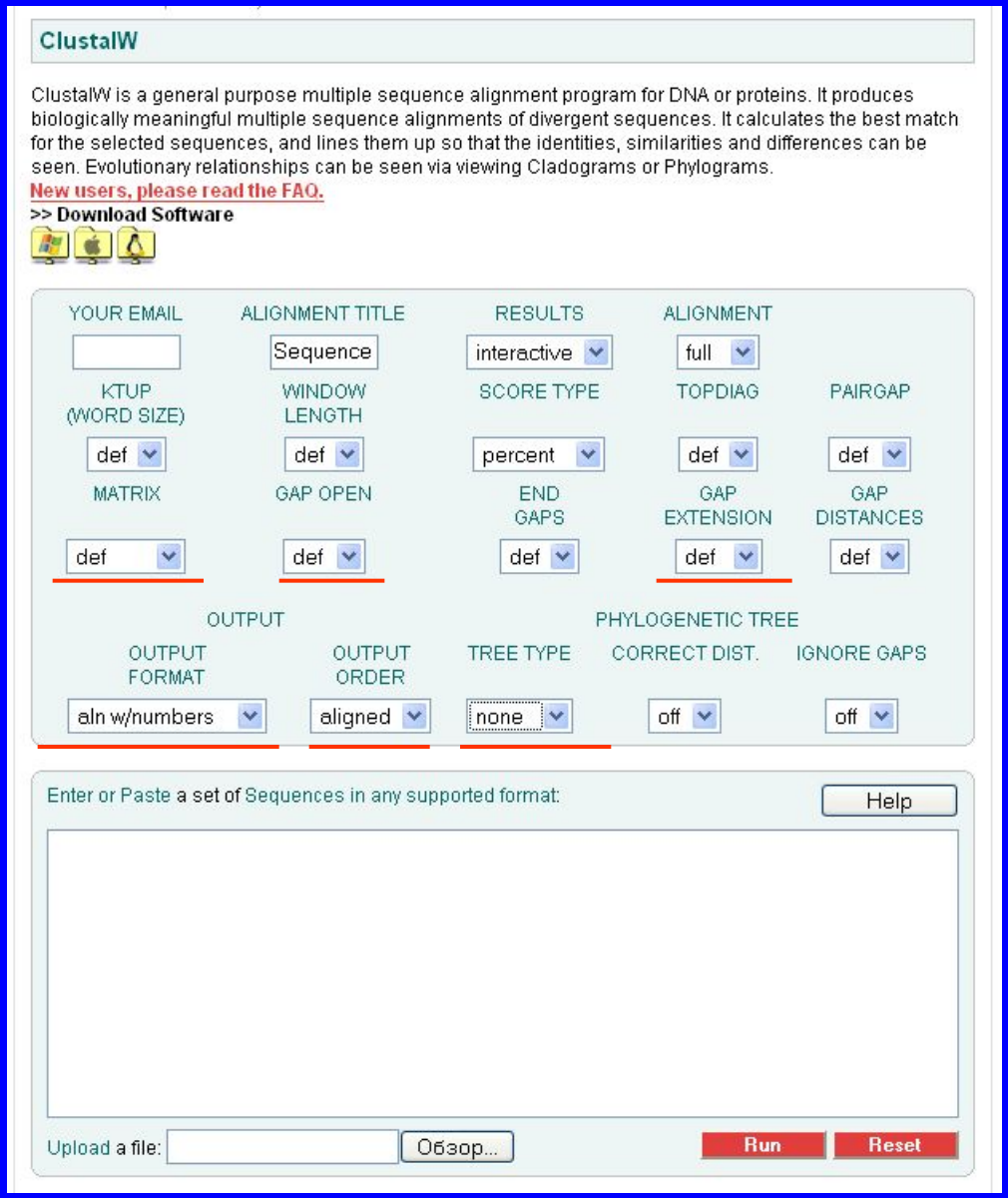
Очевидные недостатки:

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)

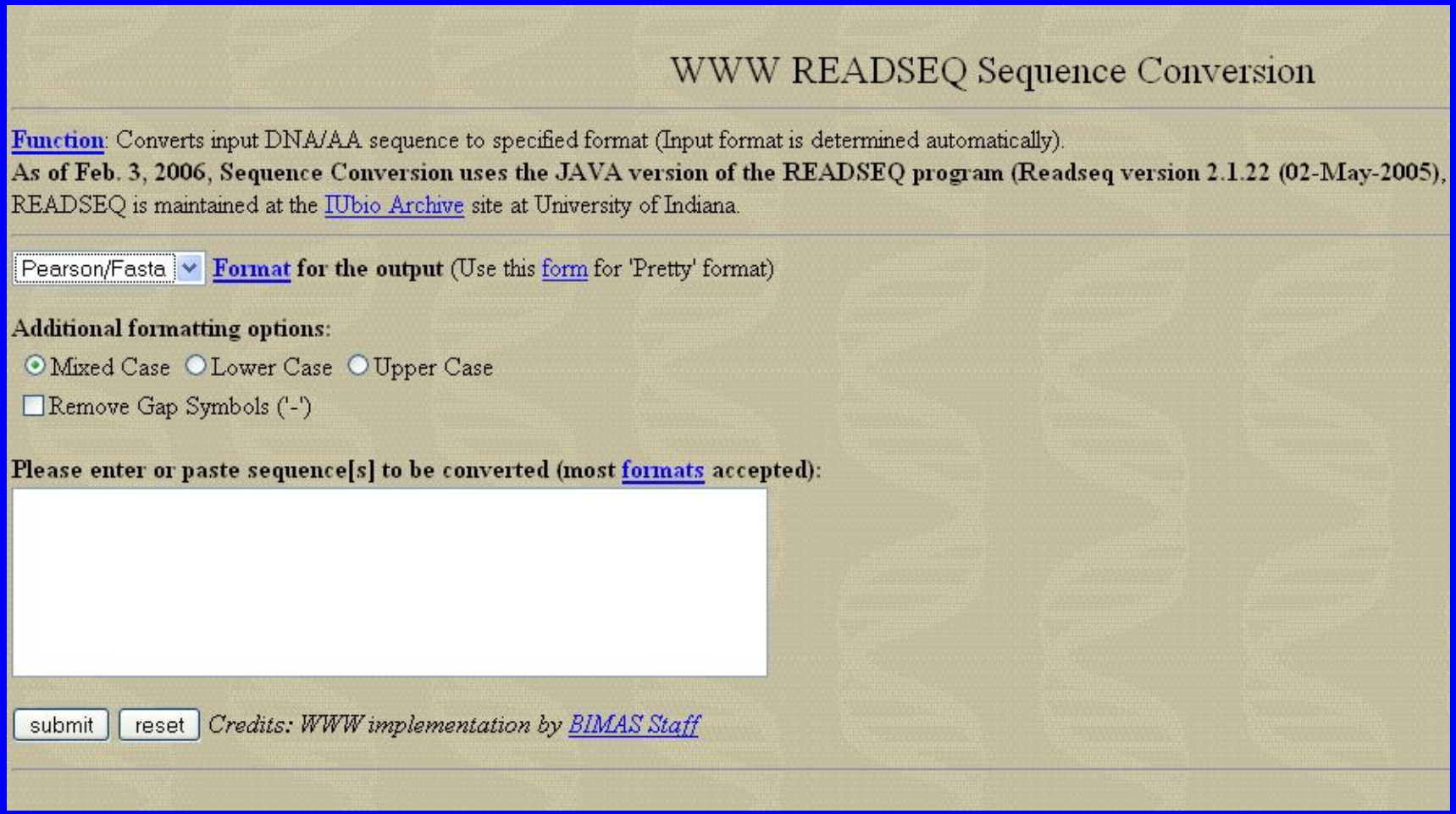
Использование ClustalW



Какие output-форматы бывают

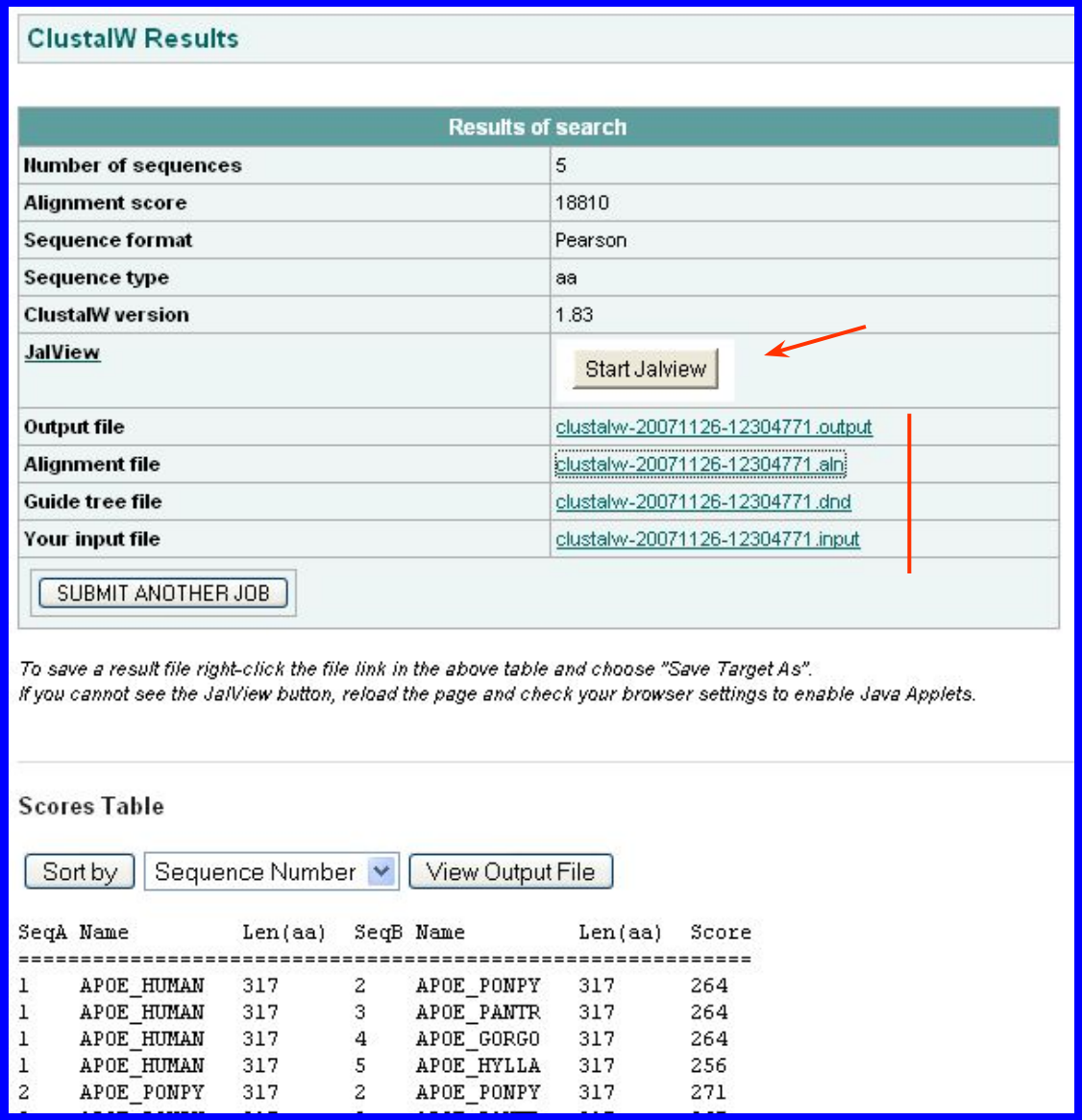
- ✔ Post-script, pdf, html – только графика
- ✔ FASTA – последовательности отдельно, но с пробелами (PIR – аналогично)
- ✔ MSF (ALN, Phylip, Selex …) – наглядно. Сверху – описание выборки: программа, название последовательностей, их длина, вес в выравнивании; потом само выравнивание блоками по 60 остатков

Перевод форматов: READSEQ (http://www-bimas.cit.nih.gov/molbio/readseq/)



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-20071126-12304771.aln
Guide tree file	clustalw-20071126-12304771.dnd
Your input file	clustalw-20071126-12304771.input

SUBMIT ANOTHER JOB

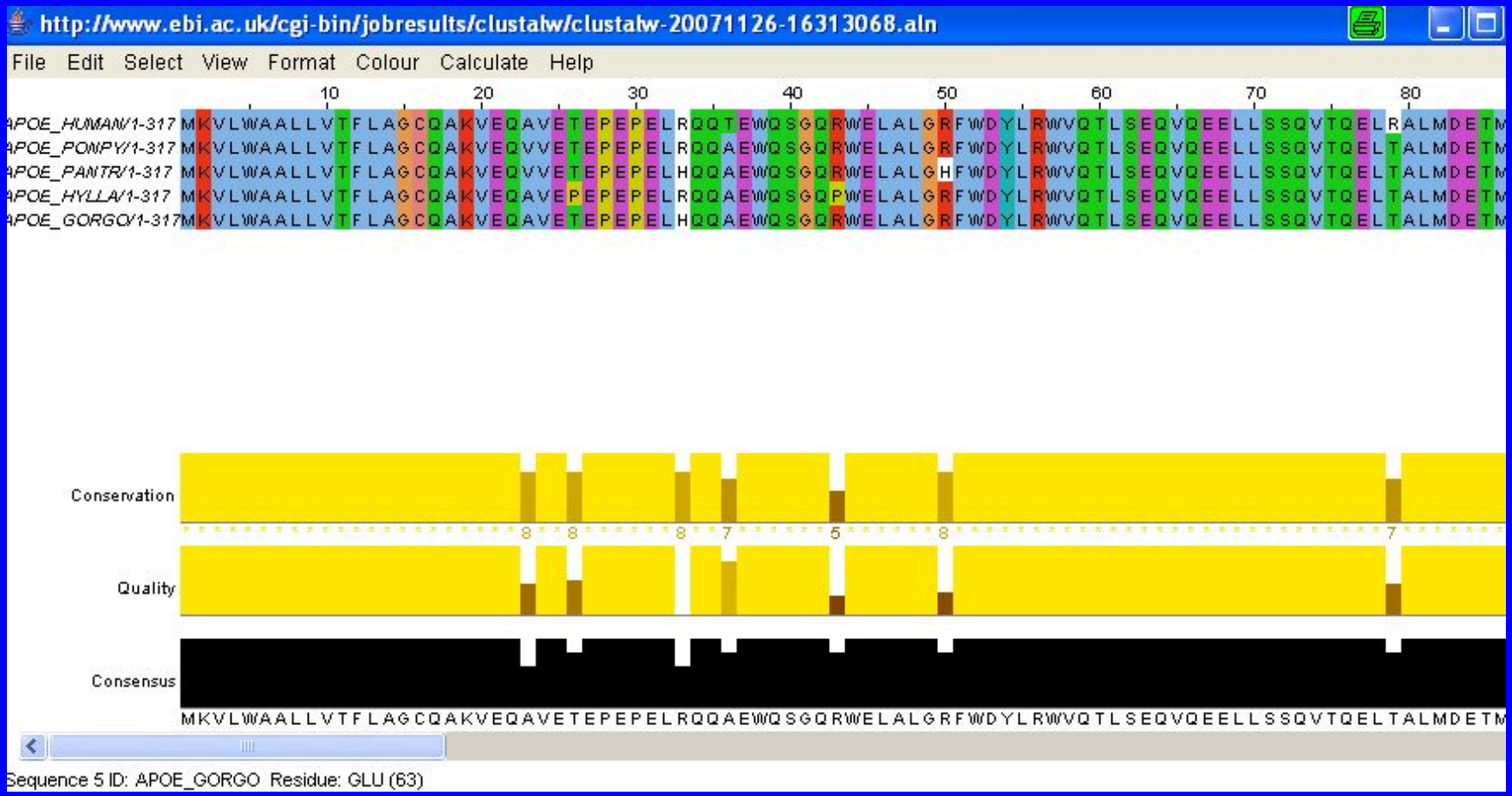
To save a result file right-click the file link in the above table and choose "Save Target As".
If you cannot see the JalView button, reload the page and check your browser settings to enable Java Applets.

Scores Table

Sort by | Sequence Number | View Output File

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

JalView – редактирование выравниваний

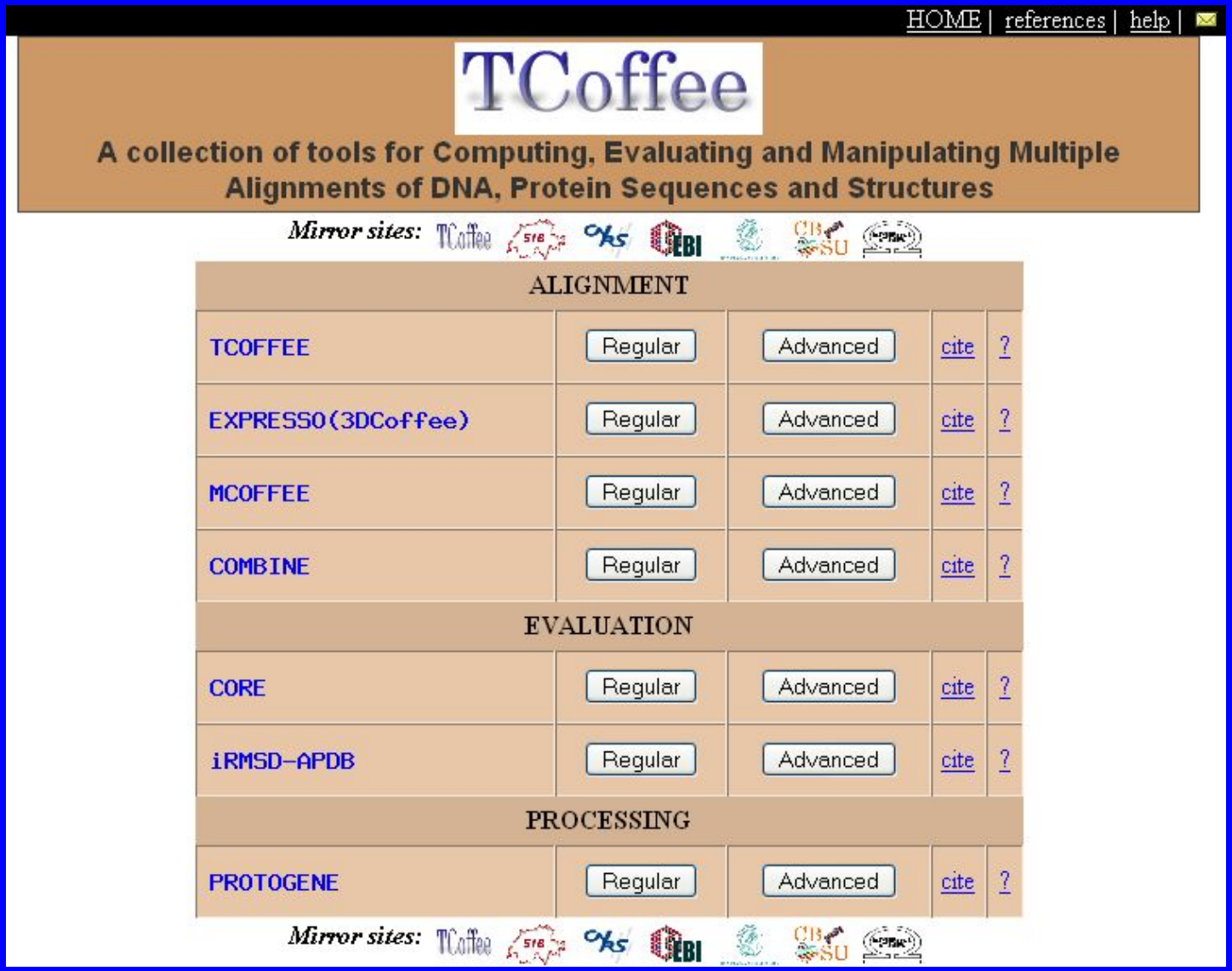


Другие программы для редактирования выравниваний (stand-alone):
GeneDoc; CINEMA; Seaview; Belvu; Bioedit; DCSE
Список - http://bioweb.pasteur.fr/cgi-bin/seqanal/review-edital.pl

TCoffee

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

TCoffee



Выход – файлы clustalw_alnВыход – файлы clustalw_aln, fasta_alnВыход – файлы clustalw_aln, fasta_aln, phylipВыход – файлы clustalw_aln, fasta_aln,

Как использовать TCoffee 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

```
dnaN      ACATTATCCGTTAGGAGGATAAAAATG
gyrA       GTGATACTTCAGGGAGGTTTTTTAATG
serS       TCAATAAAAAAAGGAGTGTTTCGCATG
bofA           CAAGCGAAGGAGATGAGAAGATTCATG
csfB        GCTAACTGTACGGAGGTGGAGAAGATG
xpaC          ATAGACACAGGAGTCGATTATCTCATG
metS       ACATTCTGATTAGGAGGTTTCAAGATG
gcaD       AAAAGGGATATTGGAGGCCAATAAATG
spoVC      TATGTGACTAAGGGAGGATTCGCCATG
ftsH      GCTTACTGTGGGAGGAGGTAAGGAATG
pabB           AAAGAAAATAGAGGAATGATACAAATG
rplJ       CAAGAATCTACAGGAGGTGTAACCATG
tufA        AAAGCTCTTAAGGAGGATTTTAGAATG
rpsJ       TGTAGGCGAAAAGGAGGGAAAATAATG
rpoA          CGTTTTGAAGGAGGGTTTTAAGTAATG
rplM         AGATCATTTAGGAGGGGAAATTCAATG
Cons           tacataaaggaggtttaaaaat
```

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

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

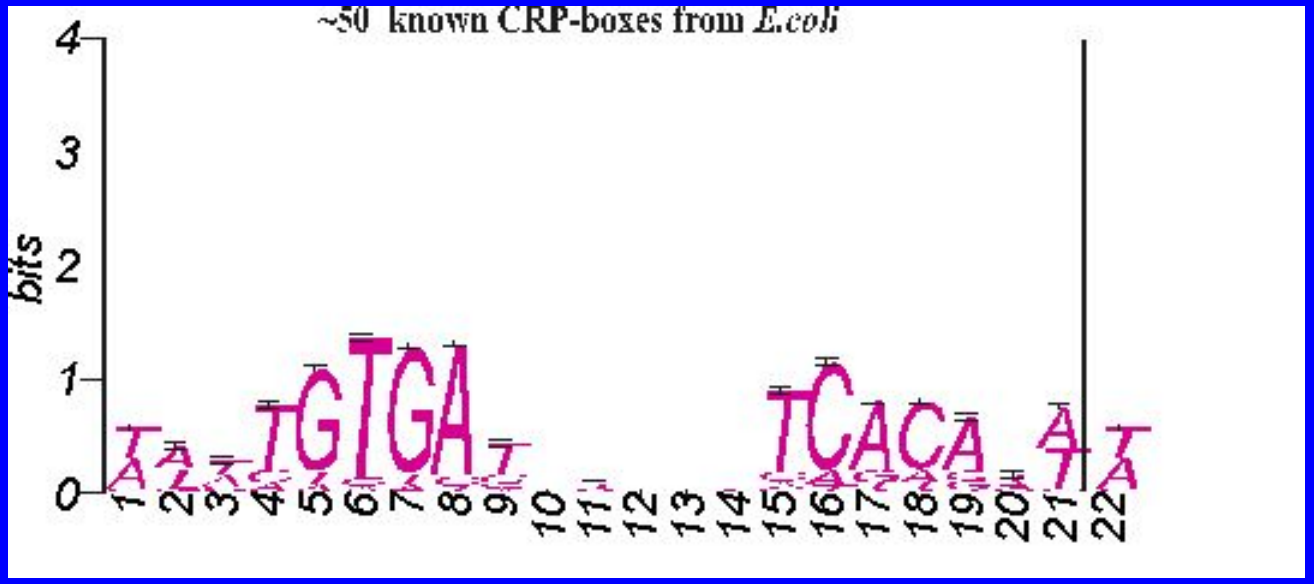
For each candidate site the total time of occupation is computed.

(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*~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

j	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 $I = \Sigma_j \Sigma_b f(b,j)[\log f(b,j) / p(b)]$

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