Grupo de Bioinformática
Depto. de Genética - FMRP/USP

Faculdade de Medicina 55 anos
USP - RIBEIRÃO PRETO - SP

Programas de Alinhamento

Departamento de Genética FMRP- USP

Alynne Oya Chiromatzo
alynne@lgmb.fmrp.usp.br

Grupo de Bioinformática
Depto. de Genética - FMRP/USP

Faculdade de Medicina 55 anos
USP - RIBEIRÃO PRETO - SP

Sumário

- Introdução
- Programas de alinhamento para buscas em base de dados
  - Fasta
  - Blast
- Programa para alinhamento múltiplo de seqüências
  - Clustal

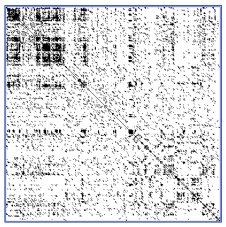
Introdução

Grupo de Bioinformática
Depto. de Genética - FMRP/USP

Faculdade de Medicina 55 anos
USP - RIBEIRÃO PRETO - SP

- Alinhamento de pares de seqüências

Programação dinâmica



Matriz de pontos

MATRIZ	GAP	A	A	G
N-W	0	1	2	3
GAP 0	0	-10	-20	-30
A 1	-10	0	-10	-20
G 2	-20	-10	1	-10
T 3	-30	-20	-11	-2
A 4	-40	-30	-20	-12
C 5	-50	-40	-30	-21

A = AGTAC
B = A--AG

Alinhamento Global
(Needleman & Wunsch)

		G	A	A	T	T	C	A	G	T	A
	0	0	0	0	0	0	0	0	0	0	0
G	0	5	1	0	0	0	0	0	5	1	0	0
G	0	5	2	0	0	0	0	0	5	2	0	0
A	0	1	10	7	3	0	0	5	1	2	0	5
T	0	0	6	7	12	8	4	1	2	6	8	4
C	0	0	2	3	8	9	13	9	5	2	4	5
G	0	5	1	0	4	5	9	10	14	10	6	2
A	0	1	10	6	2	1	5	14	10	11	7	11

```
GAATTCAG   GAATTCAG
||||||||   ||||||||
GGA-TC-G   GGAT-C-G

GAATTC-A   GAATTC-A
||||||||   ||||||||
GGA-TCGA   GGAT-CGA
```

Alinhamento Local
(Smith & Waterman)

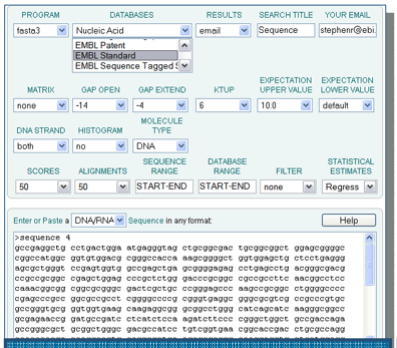
Introdução

Grupo de Bioinformática
Depto. de Genética - FMRP/USP

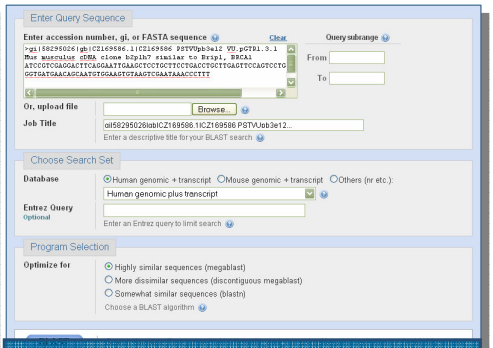
Faculdade de Medicina 55 anos
USP - RIBEIRÃO PRETO - SP

- Alinhamento de pares de seqüências

Método de busca por palavras



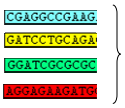
FASTA
(Fast Alignment Search Tool All)



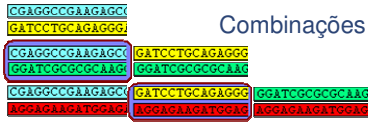
BLAST
(Basic Local Alignment and Search Tool)

Introdução

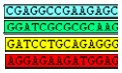
- Alinhamento múltiplo de seqüências



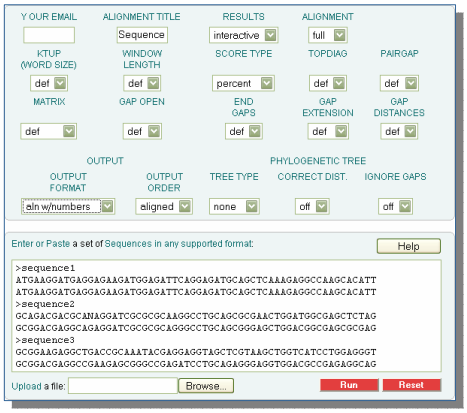
Seqüências de entrada



Combinações



Alinhamento final



Programas de alinhamento para buscas em base de dados

- FASTA

1985 FASTP
Alinhamento de seqüências de proteínas

1988 FASTA
Pacote de softwares para
alinhamento de seqüências de ADN e proteínas

Dr. William R. Pearson
Bioquímico
Prof.de Bioquímica e
Ciência da Computação
Universidade de Virgínia

Ms. David J. Lipman
Biólogo
Diretor do NCBI
National Center for
Biotechnology Information

Programas de alinhamento para buscas em base de dados

- FASTA – Web

http://www.ebi.ac.uk/fasta33/

Programas de alinhamento para buscas em base de dados

- FASTA – Web

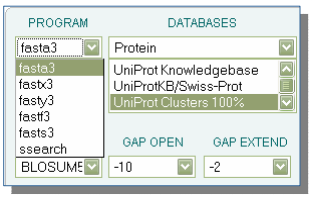
fasta3: procura seqüências similares em uma biblioteca de proteínas ou de seqüências de DNA

fastx/y3: compara uma seqüência de DNA com um banco de proteínas nos dois sentidos

fastf3: compara uma mistura de peptídeos com um banco de dados de proteínas

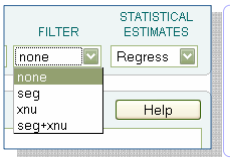
fasts3: compara peptídeos que estão ligadas com um BD de proteínas

ssearch3: procura por seqüências similares em um biblioteca de proteína ou de seqüências de DNA



Programas de alinhamento para buscas em base de dados

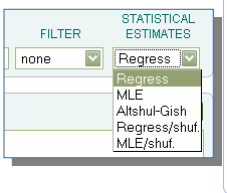
- FASTA – Web



seg: mascara regiões de baixa complexidade

xnu: mascara regiões contendo repetições internas de curta periodicidade

seg + xnu: combinação dos anteriores



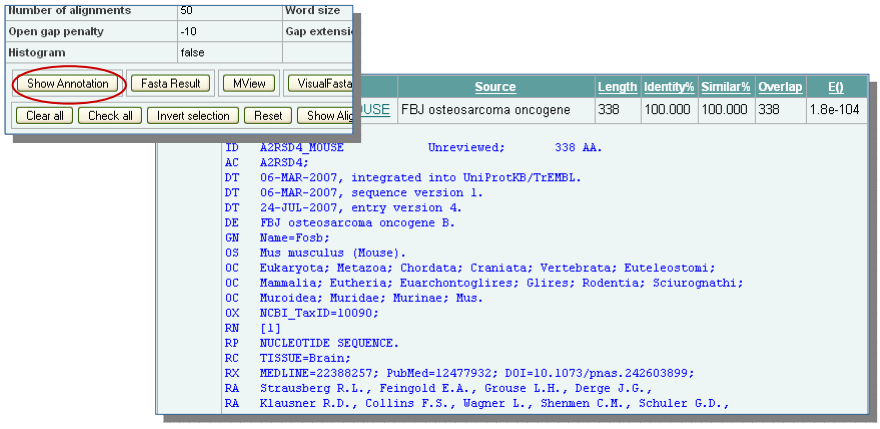
Regress: usa regressão ponderada da pontuação média contra o tamanho das seqüências da biblioteca

MLE: usa a estimativa de máxima parcimônia para λ

Regress/MLE huf: estima os parâmetros estatísticos a partir de cópias aleatórias de cada biblioteca de seqüências

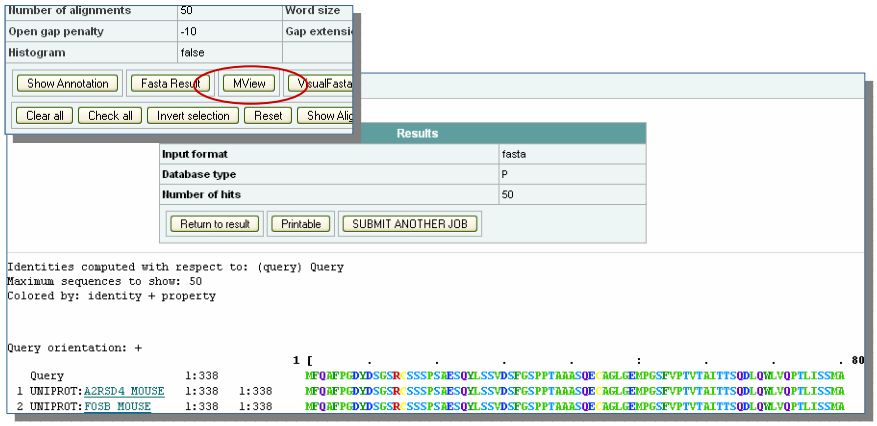
Programas de alinhamento para buscas em base de dados

- FASTA – Web



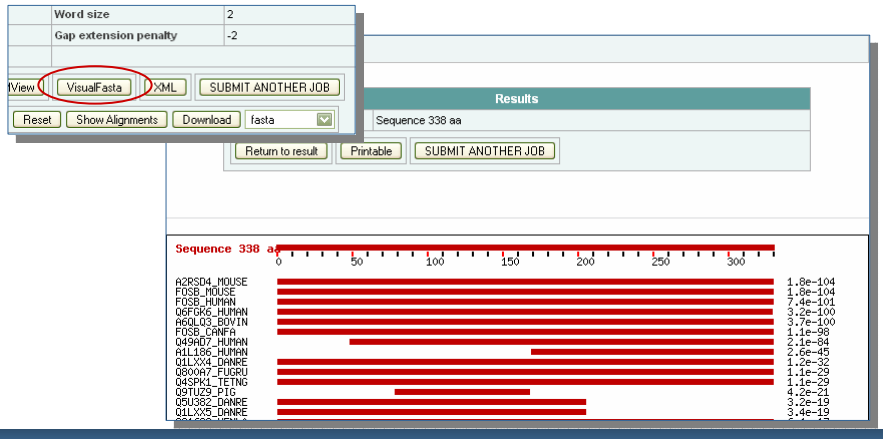
Programas de alinhamento para buscas em base de dados

- FASTA – Web



Programas de alinhamento para buscas em base de dados

- FASTA – Web



Programas de alinhamento para buscas em base de dados

Grupo de Bioinformática
Depto. de Genética - FMRP/USP

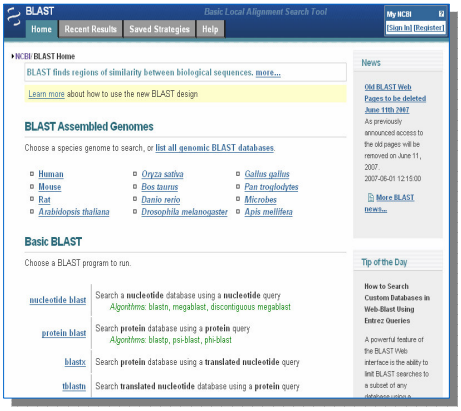
Faculdade de Medicina 55 anos
USP - RIBEIRÃO PRETO - SP

- BLAST

1990 BLAST

- Alinha seqüências de ADN ou proteínas com as seqüências da base de dados
- Mais rápido que o FASTA

http://www.ncbi.nlm.nih.gov/BLAST/

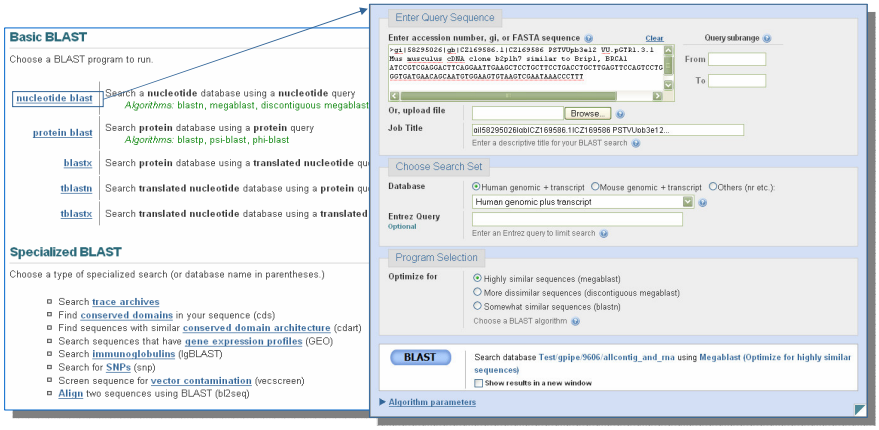


Programas de alinhamento para buscas em base de dados

Grupo de Bioinformática
Depto. de Genética - FMRP/USP

Faculdade de Medicina 55 anos
USP - RIBEIRÃO PRETO - SP

- BLAST – Web

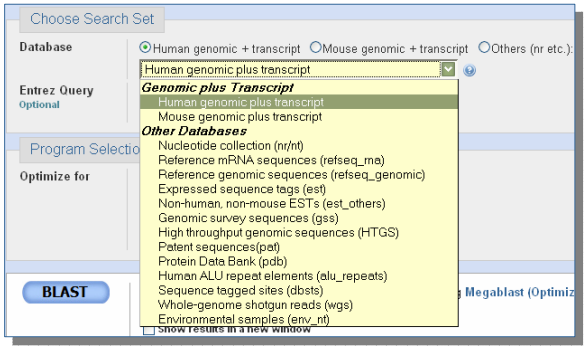


Programas de alinhamento para buscas em base de dados

Grupo de Bioinformática Depto. de Genética - FMRP/USP

Faculdade de Medicina 55 anos USP - RIBEIRÃO PRETO - SP

- BLAST – Web



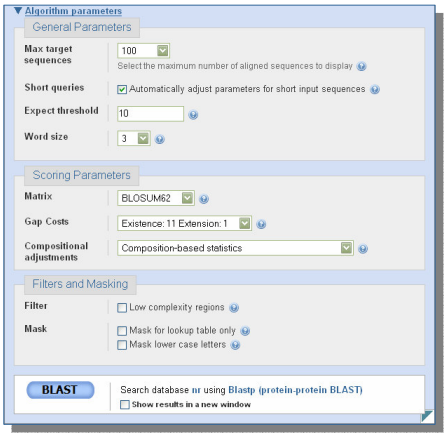
refseq: banco de dados de seqüências não redundantes
gss: seqüências de origem genômica e não por cDNA
HTGS: *High Throughput Genomic Sequences* são dados de seqüências genômicas não terminadas

Programas de alinhamento para buscas em base de dados

Grupo de Bioinformática Depto. de Genética - FMRP/USP

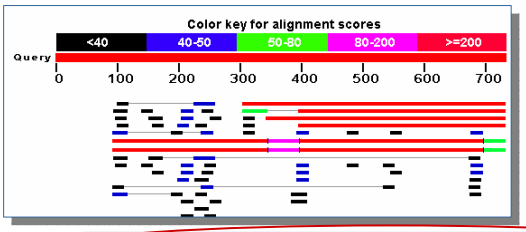
Faculdade de Medicina 55 anos USP - RIBEIRÃO PRETO - SP

- BLAST – Web



Programas de alinhamento para buscas em base de dados

- BLAST – Web



Legend for links to other resources: U UniGene E GEO G Gene S Structure M Map Viewer

Sequences producing significant alignments:
(Click headers to sort columns)

Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
Transcripts							
NM_002287.3	Homo sapiens leukocyte-associated immunoglobulin-like receptor	803	803	58%	0.0	99%	UEGM
NM_021706.2	Homo sapiens leukocyte-associated immunoglobulin-like receptor	624	704	51%	3e-176	100%	UEGM
NM_002288.3	Homo sapiens leukocyte-associated immunoglobulin-like receptor	478	478	53%	5e-132	90%	UEGM
NM_021270.2	Homo sapiens leukocyte-associated immunoglobulin-like receptor	383	383	46%	2e-103	89%	UEGM
XM_001125918.1	PREDICTED: Homo sapiens hypothetical protein LOC727872 (LOC	40.1	40.1	2%	3.5	100%	GM

Programas de alinhamento para buscas em base de dados

- BLAST – Local

```
$ blastall -p blastn -d hs_est -i entrada.fa -o out.entrada
```

-p Nome do programa

blastp: seq. de proteína contra BD de seq. de proteína

blastn: seq. de nucleotídeos contra banco de nucleotídeos

blastx: seq. de nucleotídeos nos 6 quadros de leitura contra BD de seq. de proteína

tblastn: seq. de proteína contra BD de nucleotídeos dinamicamente traduzidos nos 6 quadros de leitura

tblastx: seq. de nucleotídeos contra um BD de nucleotídeos traduzidos

-d Nome do banco de dados

Programas de alinhamento para buscas em base de dados

- BLAST – Local

```
BLASTP 2.2.2 [Jan-08-2002]

Reference: Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schaffer,
Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997),
"Gapped BLAST and PSI-BLAST: a new generation of protein database search
programs",  Nucleic Acids Res. 25:3389-3402.

Query= gi|730028|sp|P40692|MLH1_HUMAN MUTL PROTEIN HOMOLOG 1 (DNA
MISMATCH REPAIR PROTEIN MLH1)
         (756 letters)

Database: sp35
           69,113 sequences; 25,083,768 total letters|

Searching.....................................................done

                                                                  Score     E
Sequences producing significant alignments:                       (bits)  Value

gi|604369|sp|P40692|MLH1_HUMAN MUTL PROTEIN HOMOLOG 1 (DNA MISMA...  1293  0.0
gi|825572|sp|P38920|MLH1_YEAST MUTL PROTEIN HOMOLOG 1 (DNA MISMA...   437  e-122
gi|1573016|sp|P44494|MUTL_HAEIN DNA MISMATCH REPAIR PROTEIN MUTL.     190  4e-48
gi|304915|sp|P23367|MUTL_ECOLI DNA MISMATCH REPAIR PROTEIN MUTL.      190  5e-48

Query: 128 ASYSDGKLKAPPKPCAGNQGTQITVEDLFYNIATRRKALKNPSEEYGKILEVVGRYSVHN 187
            SY++GK+    PKP AG  GT I VEDLF+NI +R +AL++ ++EY KIL+VVGRY++H+
Sbjct: 125 VSYAEGKMLESPKPVAGKDGTTILVEDLFFNIPSRLRALRSHNDEYSKILDVVGRYAIHS 184

Query: 188 AGISFSVKKQGETVADVRTLPNASTVDNIRSIFGNAVSRELIEI---GCEDKTLAFKMNG 244
               I FS KK G++    +   P+ +  D IR++F  +V+  LI       ED  L   ++G
Sbjct: 185 KDIGFSCKKFGDSNYSLSVKPSYTVQDRIRTVFNKSVASNLITFHISKVEDLNLE-SVDG 243
```

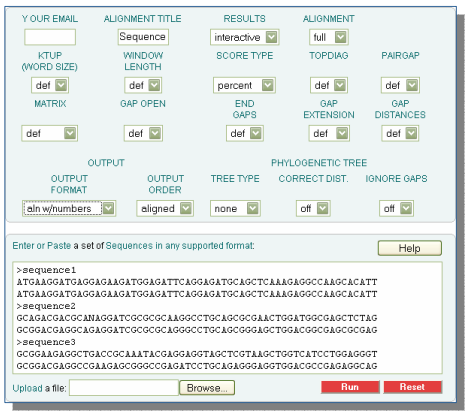

```
SMATCH REPAIR PROTEIN MLH1).

= 48/781 (6%)

LIQIQDNGTG 67
++QI DNG+G
VLQITDNGSG 64
KTADGKCAYR 127
K  + +CA+R
KVKEDRCAWR 124
```


Programas para alinhamento múltiplo de seqüências

- Clustal

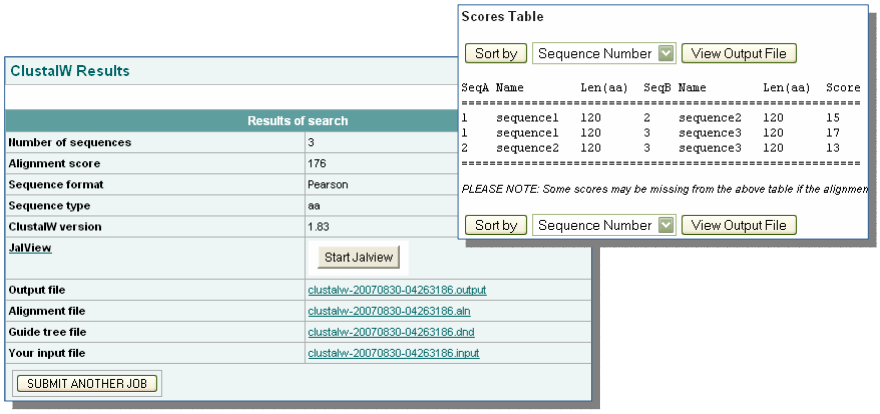
1994 Clustal
- Clustal: atribui pesos iguais a todas seqüências
- ClustalW: atribui diferentes pesos às seqüências
- ClustalX: proporciona uma interface gráfica para o ClustalW

http://www.ebi.ac.uk/Tools/clustalw/



Programas para alinhamento múltiplo de seqüências

- Clustal – Web



ClustalW Results

Results of search	
Number of sequences	3
Alignment score	176
Sequence format	Pearson
Sequence type	aa
ClustalW version	1.83
JalView	Start Jalview
Output file	clustalw-20070830-04263186.output
Alignment file	clustalw-20070830-04263186.aln
Guide tree file	clustalw-20070830-04263186.dnd
Your input file	clustalw-20070830-04263186.input

SUBMIT ANOTHER JOB

Scores Table

Sort by | Sequence Number | View Output File

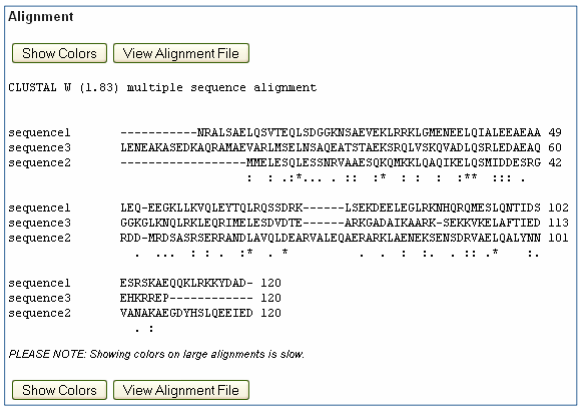
SeqA	Name	Len(aa)	SeqB	Name	Len(aa)	Score
1	sequence1	120	2	sequence2	120	15
1	sequence1	120	3	sequence3	120	17
2	sequence2	120	3	sequence3	120	13

PLEASE NOTE: Some scores may be missing from the above table if the alignmen

Sort by | Sequence Number | View Output File

Programas para alinhamento múltiplo de seqüências

- Clustal – Web



Alignment

Show Colors | View Alignment File

```
CLUSTAL W (1.83) multiple sequence alignment


sequence1       -----------NRALSAELQSVTEQLSDGGKNSAEVEKLRRKLGMENEELQIALEEAEAA 49
sequence3       LENEAKASEDKAQRAMAEVARLMSELNSAQEATSTAEKSRQLVSKQVADLQSRLEDAEAQ 60
sequence2       ------------------MMELESQLESSNRVAAESQKQMKKLQAQIKELQSMIDDESRG 42
                                   :  : .:*... . ::  :*  : :  :  :** ::: .

sequence1       LEQ-EEGKLLKVQLEYTQLRQSSDRK------LSEKDEELEGLRKNHQRQMESLQNTIDS 102
sequence3       GGKGLKNQLRKLEQRIMELESDVDTE------ARKGADAIKAARK-SEKKVKELAFTIED 113
sequence2       RDD-MRDSASRSERRANDLAVQLDEARVALEQAERARKLAENEKSENSDRVAELQALYNN 101
                  .  ...  : : .  :*  . *          .  .  :  :.  . :: .*    :.

sequence1       ESRSKAEQQKLRKKYDAD- 120
sequence3       EHKRREP------------ 120
sequence2       VANAKAEGDYHSLQEEIED 120
                  . :

PLEASE NOTE: Showing colors on large alignments is slow.
```

Show Colors | View Alignment File

* Idêntico
. semi conservada
: conservada

Programas para alinhamento múltiplo de seqüências

Grupo de Bioinformática
Depto. de Genética - FMRP/USP

Faculdade de Medicina 55 anos
USP - RIBEIRÃO PRETO - SP

- Clustal – Local

ftp://ftp.ebi.ac.uk/pub/software/unix/clustalw/

```
$ wget ftp://ftp.ebi.ac.uk/pub/software/unix/clustalw/
clustalw1.83.linux.tar.gz
$ tar -zxvf clustalw1.83.linux.tar.gz
```

Programas para alinhamento múltiplo de seqüências

Grupo de Bioinformática
Depto. de Genética - FMRP/USP

Faculdade de Medicina 55 anos
USP - RIBEIRÃO PRETO - SP

- Clustal – Local

```
clustalw -entrada.fa -align -tree
```

-infile	Arquivo de entrada
-align	Faz o alinhamento múltiplo
-bootstrap(=n)	Faz o Bootstrap em uma árvore Neighbour Join NJ (n= número de bootstraps; def. = 1000).
-convert formato	Converte as seqüências de entrada em um diferente
-help ou -check	Mostra uma visão geral sobre os parâmetros
-options	Lista os parâmetros da linha de comando
-tree	Calcula a árvore neighbour join

Programas para alinhamento múltiplo de seqüências

Grupo de Bioinformática
Depto. de Genética - FMRP/USP

Faculdade de Medicina 55 anos
USP - RIBEIRÃO PRETO - SP

- Clustal – Local

PIR

```
>P1;CRAB_ANAPL
ALPHA CRYSTALLIN B CHAIN (ALPHA(B)-CRYSTALLIN).
  MDITIHNPLI RRPLFSWLAP SRIFDQIFGE HLQESELLPA SPSLSPFLMR
  SPIFRMPSWL ETGLSEMRLE KDKFSVNLDV KHFSPEELKV KVLGDMVEIH
  GKHEERQDEH GFIAREFNRK YRIPADVDPL TITSSLSLDG VLTVSAPRKQ
  SDVPERSIPI TREEKPAIAG AQRK*
```

1) '>' <cod_desc>; <seq_id>
2) Descrição textual
3) Seqüência e '*' marcando o final

Phylip

```
 5 176
cox2_leita   MAFILSFWMI FLLDSVIVLL SFVCFVCVWI CALLFSTVLL VSKLNNIYCT
cox2_crifa   MAFILSFWMI FLIDAVIVLL SFVCFVCIWI CSLFFSSFLL VSKINNVYCT
cox2_bsalt   MSFIISFWML FLIDSLIVLL SGAIFVCIWI CSLFFLCILF ICKLDYIFCS
cox2_trybb   MSFILTFWMI FLMDSIIVLI SFSIFLSVWI CALIIATVLT VTKINNIYCT
cox2_tborr   MLFFINQLLL LLVDTFVILE IFSLFVCVFI IVMYILFINY NIFLKNINVY
```

1) Num. de espécies e tam. da seqüência
2) Nome da seqüência e seqüência separa em blocos de10 caracteres

P1 Proteína (completa)
F1 Proteína (fragmento)
DL DNA (linear)

DC DNA (circular)
RL RNA (linear)
RC RNA (circular)

Aln W

```
CLUSTAL W (1.82) multiple sequence ali

FOSB_MOUSE      MFQAFPGDYDSGSRCSSSPSAESQYLSSVDSFGSPPTAAASQECAGLGEMPGSFVPTVTA 60
FOSB_HUMAN      MFQAFPGDYDSGSRCSSSPSAESQYLSSVDSFGSPPTAAASQECAGLGEMPGSFVPTVTA 60
                ************************************************************

FOSB_MOUSE      ITTSQDLQWLVQPTLISSMAQSQGQPLASQPPAVDPYDMPGTSYSTPGLSAYSTGGASGS 120
FOSB_HUMAN      ITTSQDLQWLVQPTLISSMAQSQGQPLASQPPVVDPYDMPGTSYSTPGMSGYSSGGASGS 120
                ***********************************.***************:*.**:******
```

TPGLSAYST
TPGMSGYSS
***:*.**:

* Idêntico
. semi conservada
: conservada

Programas para alinhamento múltiplo de seqüências

Grupo de Bioinformática
Depto. de Genética - FMRP/USP

Faculdade de Medicina 55 anos
USP - RIBEIRÃO PRETO - SP

- Clustal – Local

CGC

```
   MSF:  510  Type: P    Check:  7735  ..

Name: ACHE_BOVIN oo  Len:  510  Check:  9842  Weight:  16.0
Name: ACHE_HUMAN oo  Len:  510  Check:  8553  Weight:  17.8
Name: ACHE_MOUSE oo  Len:  510  Check:  229  Weight:  13.5
Name: ACHE_RAT oo  Len:  510  Check:  8410  Weight:  14.2
Name: ACHE_XENLA oo  Len:  510  Check:  2703  Weight:  99.3

//

ACHE_BOVIN      MARALLGALL MLQLLGRGEG KMEESELVHY LFVTVDPGKR PV
ACHE_HUMAN      MARAPLGVLL LLGLLGRGVG KMEELSLVHH LFNYDPGSR PV
ACHE_MOUSE      MARALLGALL LLTLTGRSQG KMEELSLVRH LFNYLPECH PV
ACHE_RAT        MTRALLGTLL LLALTGPSQG KMEELSLVRH LFNYDPECH PV
ACHE_XENLA      MSGVRILSL LILHENLAS ESEESRLISH LFTSVDQKAR PS
```

Tamanho, tipo e checksum interno

GDE

```
{
name "Short name for sequence"
longname "Long (more descriptive) name for sequence"
sequence-ID "Unique ID number"
creation-date "mm/dd/yy hh:mm:ss"
direction [-1|1]
strandedness [1|2]
type [DNA|RNA||PROTEIN|TEXT|MASK]
offset (-999999,999999)
group-ID (0,999)
creator "Author's name"
descrip "Verbose description"
comments "Lines of comments that can be fairly arbitrary text about a
sequence. Return characters are allowed, but no internal double quotes
or brace characters. Remember to close with a double quote"
sequence "gctagctagctagctagctcttagctgtagtcgtagctgatgctagct
gatgctagctagctagctgatcgatgctagctgatcgtagctgacg
gactgatgctagctagctagctagctgtctagtgtcgtagtgcttattgc"
}
```

Seqüência de *gaps* do início da seqüência para que alinhe corretamente com outras