UNIVERSITY OF TORONTO

1

Multiple sequence alignments: goals

Evolutionary analyses
- identify homology
- build phylogenies
- test evolutionary models

Functional analyses
- identify conserved regions
- identify protein families

Structural analyses
- identify sequence co-variation
- homology modeling

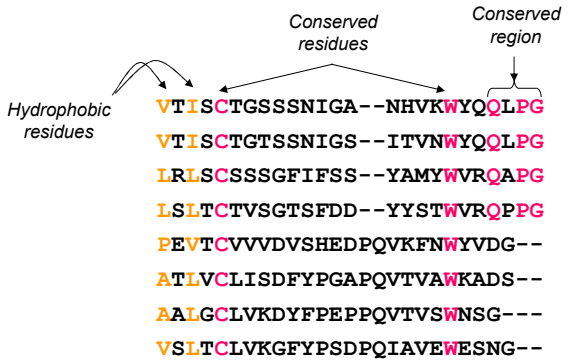
Practical application
- identify conserved primer binding sites
- design of mutagenesis experiments
- mutant analysis

Multiple sequence alignment: evolutionary history

```
VTISCTGSSSNIGA--NHVKWYQQLPG
VTISCTGTSSNIGS--ITVNWYQQLPG
LRLSCSSSGFIFSS--YAMYWVRQAPG
LSLTCTVSGTSFDD--YYSTWVRQPPG
PEVTCVVVDVSHEDPQVKFNWYVDG--
ATLVCLISDFYPGAPQVTVAWKADS--
AALGCLVKDYFPEPPQVTVSWNSG---
VSLTCLVKGFYPSDPQIAVEWESNG--
```

- MSA Columns = Homology
- Identification of homologous residues greatly facilitated by multiple comparisons

Multiple sequence alignment: structure / function

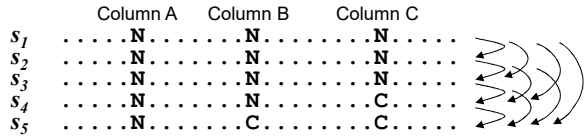


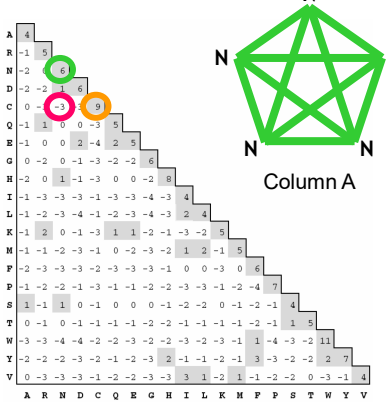
- MSA Columns = Homology
- Identification of homologous residues greatly facilitated by multiple comparisons
- Homology selectively maintained due to structural or functional constraints
- MSAs identify *conserved* or *structurally equivalent* residues and regions

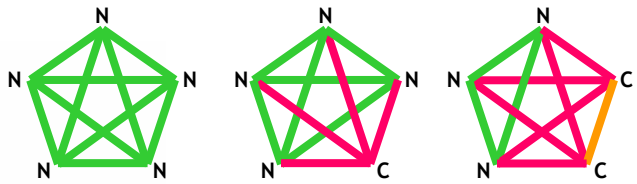
Scoring a multiple sequence alignment: sum-of-pairs (SP) scoring

- Standard MSA scoring method
- SP is a column-by-column cost/weight function
- SP scored using a substitution matrix (e.g PAM or BLOSUM)
- MSAs *maximize* total alignment score by *maximizing* each column SP score
- Assumes column independence

Multiple sequence alignment: sum-of-pairs (SP) scoring







Alignmt	Score	Column A	Column B	Column C
N – N	6	10	6	3
N – C	-3	0	4	6
C – C	9	0	0	1
		60	24	9

Multiple Sequence Alignment algorithms

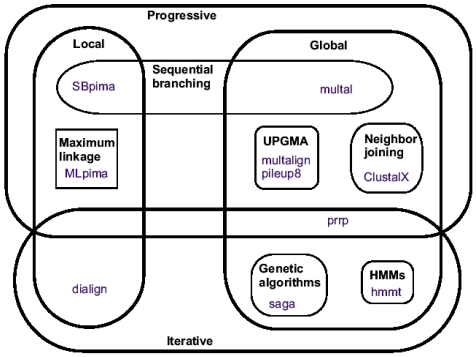
Multidimensional dynamic programming. If we have *N* sequences of length *L*, we need ~10 x L^N nsec to calculate a dynamic programming matrix, or 32 thousand years for 10 seqs of 100 residues! Not so practical...

Progressive MSA

- Profile methods – **Clustal**
- Iterative methods

Local MSA

- **DIALIGN**



Thompson (1999) NAR 27:2682. *Other methods have been developed since the publication of this graphic, but it does give a nice overview about how to think about classifying the different methods.*

Almost all MSA techniques are heuristics.

Progressive Alignment

Concept

- Any 2 sequences can be aligned accurately and rapidly via dynamic programming
- Once alignment is made, equivalent to any other sequence
- aligned set of sequences = **profile**
- Pairs of profiles, or profiles and sequences can be aligned accurately and rapidly via dynamic programming
- Progressively align more distantly related profiles and sequences

No separation of alignment scoring and alignment optimization

No optimization of a global scoring function

Progressive Alignments
profile methods

2 profiles can be aligned without disturbing the alignment of either individual profile

Given 2 profiles:

```
ACGTA        AAGTAA
AC-TA        TCG-AA
```

To align these profiles without disturbing their internal alignments, we can:

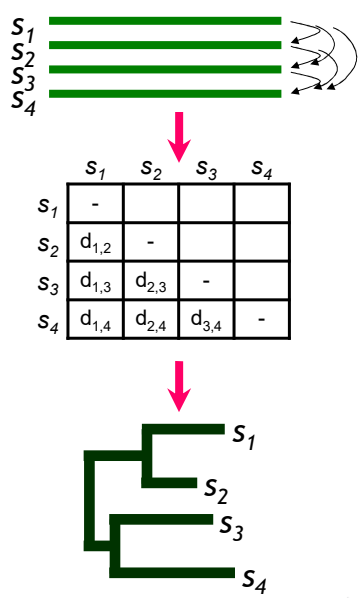
```
 ACGT  A        ACGT  A        ACGTA  -
 AC-T  A        AC-T  A        AC-TA  -
AAGTA  A       AAGTAA  -       AAGTA  A
TCG-A  A       TCG-AA  -       TCG-A  A
```

Equivalent to dynamic programming with 2 sequences

Once a gap, always a gap!

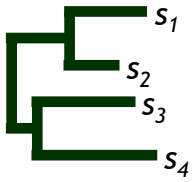
Clustal
distances and guide tree

1. *Compute all pairwise global alignments*
   - Use fast *k*-tuple or slow dynamic programming to compute global pairwise alignments
   - Affine gap penalties
2. *Calculate pairwise distances*
3. *Use distance matrix to calculate guide tree*
   - Neighbour-Joining
   - midpoint rooted

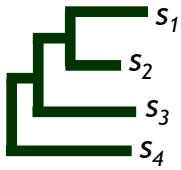


Clustal alignment

4. *Determine order of alignment based on guide tree*



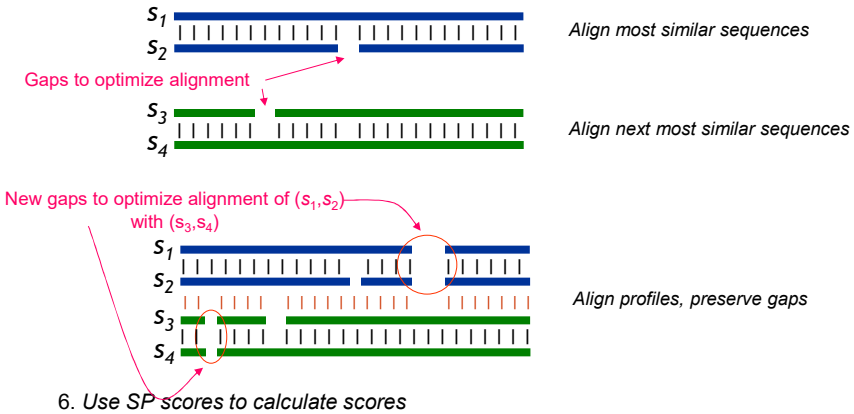
Align s_1 to s_2
Align s_3 to s_4
Align s_1:s_2 to s_3:s_4



Align s_1 to s_2
Align s_3 to s_1:s_2
Align s_4 to s_1:s_2:s_3

Clustal alignment

5. *Perform alignments*
   - sequence-sequence, sequence-profile, or profile-profile
   - existing gaps fixed in a profile
   - new gaps must be inserted into all sequences in a profile



Align most similar sequences

Align next most similar sequences

Align profiles, preserve gaps

6. *Use SP scores to calculate scores*

Clustal
potential problems

Clustal is a “greedy” algorithm

- makes the best immediate solution (local choice) in hopes of finding the best overall (global) solution
- choices are made regardless of later consequences
- early mistakes get propagated throughout the rest of the alignment

	Alignment 1	Alignment 2	Alignment 3
Inital Alignment	ACTTA AGT-A	ACTTA AG-TA	ACTTA A-GTA

new sequence ACGTA

Clustal
potential problems

Clustal is a ‘greedy’ algorithm

- finds best immediate solution, regardless of later consequences
- early mistakes get propagated throughout the rest of the alignment

	Alignment 1	Alignment 2	Alignment 3
Inital Alignment	ACTTA AGT-A	ACTTA AG-TA	ACTTA A-GTA
Later Alignment	ACTTA AGT-A ACGTA	ACTTA AG-TA ACGTA	ACTTA A-GTA ACGTA

optimal

Clustal
substitution matrices

Clustal uses dynamic substitution matrices

- distances among sequences determines the substitution matrix
- distances based on guide tree

Sequence Identity	Matrix
80% – 100%	Blosum80
60% – 80%	Blosum60
40% - 60%	Blosum45
< 30%	Blosum30

Clustal
sequence weights and gap penalties

Clustal weights sequences to reduce biases introduced by evolutionary history – more divergent sequences carry more weight

Gap opening penalty (OP)

- Decrease for more divergent sequences
- Increase for sequences of same length

Extension penalty (EP)

- Varies depending on differences in length

Position-specific gap penalties

- OP and EP lowered if gap already exists at a given position
- Otherwise, OP increased if gaps nearby

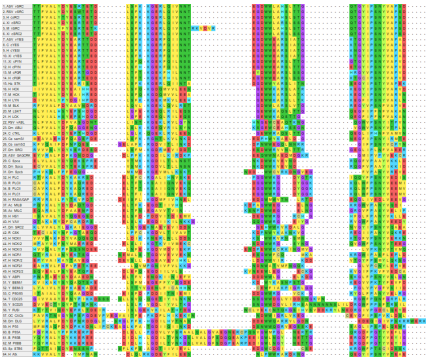
Residue-specific gap penalties

- OP decreased within run of hydrophobic residues, e.g. presence of a loop structure
- Otherwise, multiply OP by residue specific values

AA	Penalty	AA	Penalty
Q	1.13	M	1.29
C	1.13	N	0.63
D	0.96	P	0.74
E	1.31	Q	1.07
F	1.20	R	0.72
G	0.61	S	0.76
H	1.00	T	0.89
I	1.32	V	1.25
K	0.96	W	1.00
L	1.21	Y	1.23

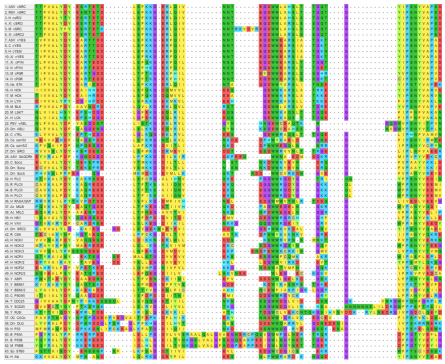
Clustal parameters

best parameters



Pairwise: BLOSUM, w. gap open penalty = 4, gap extension penalty = 0
MSA: BLOSUM, gap open penalty = 12, gap extension penalty = 0.1

default parameters



Pairwise: Gonnet, w. gap open penalty = 10, gap extension penalty = 0.1
MSA: Gonnet, gap open penalty = 10, gap extension penalty = 0.2

SH3 domain alignments after Yuan et al. 1999 MULTICLUSTAL: a systematic method for surveying Clustal W alignment parameters. Bioinformatics 15: 862-3.

Iterative Multiple Sequence Alignment Methods

Progressive Alignment methods

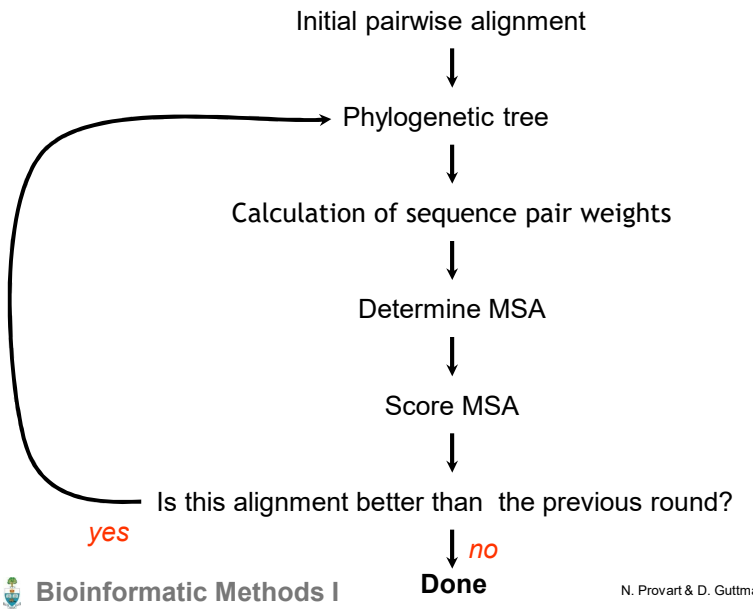
- Major problem – propagation of errors in the initial alignment throughout the MSA.

Iterative methods correct for this by repeatedly realigning subgroups and then realigning these subgroups into the global alignment.

Selection of groups can be based upon

- order of sequences on a phylogenetic tree
- separation of the sequence from the rest
- random sampling

Iterative Multiple Sequence Alignment Methods



Local Multiple Sequence Alignment
DIALIGN

Comparison of sequence segment pairs – not single residues

Segment pairs = "**diagonals**"

- diagonal = gap-free pair of sequences of equal length
- mismatches allowed
- must be consistent
- weight assigned based on the alignment quality

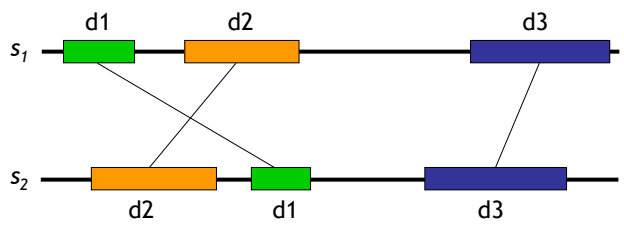
Alignments constructed by connecting consistent diagonals

- no gap penalty

Morgenstern, B. (1999) DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment. Bioinformatics 15(3): 211-218.

DIALIGN

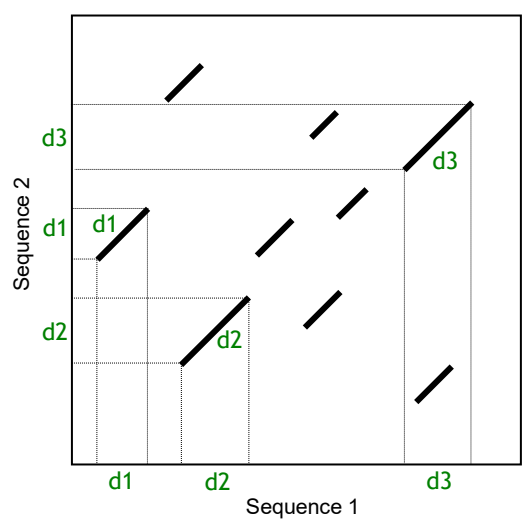
Consistent sets of diagonals



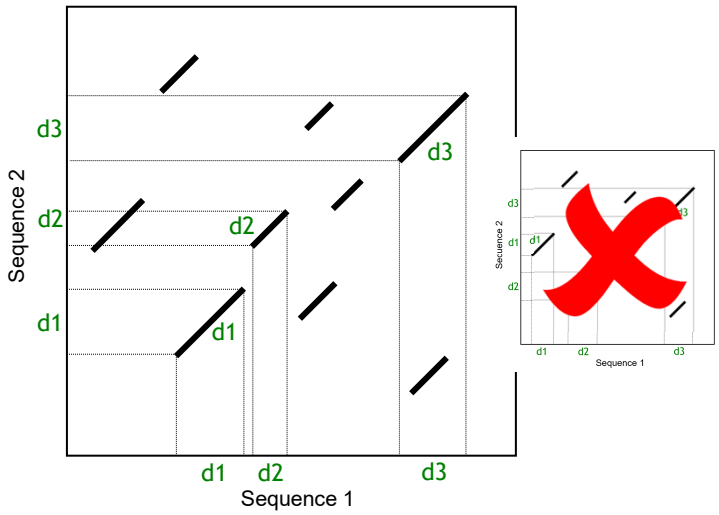
consistent: d1+d3, d2+d3
inconsistent: d1+d2

Which set of diagonals has greater significance?

DIALIGN



DIALIGN



DIALIGN

Overlap Weights

= weight of diagonal + degree of overlap with other diagonals

- weight of diagonal

  w(D) = -log $P(l_D, s_D)$

  s_D = sum of individual similarity values of residue pairs

  protein alignments : BLOSUM62

  DNA alignments : match=1, mismatch=0

  l_D = length of diagonal

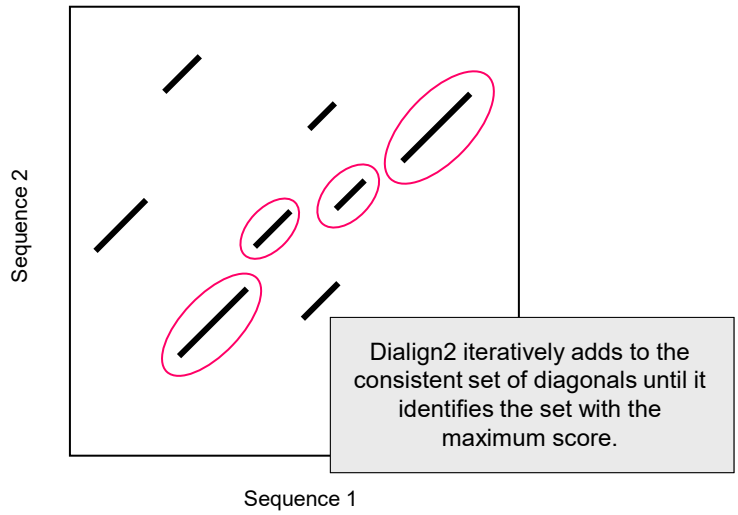
  $P(l_D, s_D)$ = prob of finding a diagonal of length l_D with a score $\geq s_D$ when comparing 2 random sequences of same length

- overlap weight favours motifs occurring on multiple sequences

Morgenstern, B. (1999) DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment. Bioinformatics 15(3): 211-218.

DIALIGN

Optimal alignment = collection of consistent diagonals with maximum sum of overlap weights



DIALIGN

```
I A V L F A E D
| | | | | |\ \
L A V I F G S
    \ \ \   \ \
W D D V T F D A E
```

consistent diagonals

```
I A - V L F - A E d

L A - V I F - G s -

w d d V T F d A E -
```

alignment

DIALIGN

- Especially well suited to detect local similarities in otherwise unrelated sequences.
- Pairwise as well as multiple alignments can be performed.
- Will align nucleotide sequences based on their amino acid alignments

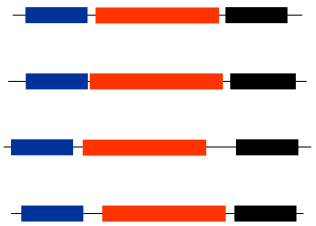
DIALIGN output

```
At_NP_200220     438   ITGPNTGGKT ICLKSVGLAA MMAKSGLYVL AT-ESARIPW FDNIYADIGD
O.sativa_NP_     431   ITGPNTGGKT ISLKTVGLAS LMAKIGLYIL AS-EPVKIPW FNAVYADIGD
Synechococcu     357   ITGPNTGGKT VTLKSIGLAS LMARAGLLLP CA-GMPTLPW CAQVLADIGD
Prochlorococ     375   ITGPNTGGKT VTLKSVGLAL LMARAGLLLP CT-GSPSMPW CAQVLADIGD
NostocZP_001     113   ITGPNTGGKT VTLKTLGLAA LMAKVGLFVP AR-EPVEIPW FDKVLADIGD
Prochlorococ     352   ITGPNTGGKT VTLKSIGLAI LMTKLGLFLP CV-GEPSLPW CNQVLADIGD
Staph_aureus     334   ITGPNTGGKT VTLKTLGLII VMAQSGLLIP TL-DGSQLSV FKNVYCDIGD
B.anthracis_     333   ITGPNTGGKT VTLKTVGICV LMAQSGLHIP VM-DESEICV FKNIFADIGD
B.halodurans     333   ITGPNTGGKT VTLKTIGLLT LMAQSGLHVP AE-EESELAV FKHVFADIGD
O.sativa_CAE     442   ISGPNTGGKT ATMKTLGLAS LMSKAGMFFP AK-GTPRLPW FDQVLADIGD
At_NP_176687     400   ISGPNTGGKT ALLKTLGLLS LMSKSGMYLP AK-NCPRLPW FDLILADIGD
At_E96674        400   ISGPNTGGKT ALLKTLGLLS LMSKSGMYLP AK-NCPRLPW FDLILADIGD
B.halodurans     328   ITGPNTGGKT VTLKTVGLLT LMAQTGLMIP AgpKGS-LHV FQKLFVDIGD
At_NP_189075     670   LTGPNGGGKS SLLRSICAAA LLGISGLMVP AE-SAC-IPH FDSIMLHMKS
At_BAB02932      664   LTGPNGGGKS SLLRSICAAA LLGISGLMVP AE-SAC-IPH FDSIMLHMKS
At_AAO49798      766   LTGPNGGGKS SLLRSICAAA LLGISGLMVP AE-SAC-IPH FDSIMLHMKS
B.cereus_NP_     599   ITGPNMSGKS TYMRQLALVT VMSQIGCFVP AT-EAV-LPV FDQIFTRIGA
B.halodurans     603   ITGPNMAGKS TYMRQLALIA IMGQIGCFVP AD-EAR-LPI FDQVFTRIGA

                       8888888888 8888888889 9888888887 5403332444 5556666666

At_NP_200220     487   EQSLLQSLST FSGHLKQISE ILSHST---S ---------- --RSLVLLDE
O.sativa_NP_     480   EQSLTQSLST FSGHLKQIGA IRAWST---S ---------- --QSLVLLDE
Synechococcu     406   EQSLQQSLST FSGHIKRIGR ILEALQSGPS ---------- --PALVLLDE
Prochlorococ     424   EQSLQQNLST FSGHVKRIGC ILEALNEGPG ---------- --PALVLLDE
NostocZP_001     162   EQSLQQSLST FSGHIRRISR ILEALGpes gsedgekemp npQ--VLLDE
Prochlorococ     401   EQSLQQNLST FSGHVVRIIR ILDAIAIRGS ---------- --SILLDE
Staph_aureus     383   EQSIEQSLST FSSHMTNIVE ILKHAD---K ---------- --HSLVLFDE
B.anthracis_     382   EQSIEQSLST FSSHMVNIVD ILEKAD---F ---------- --ESLVLFDE
B.halodurans     382   EQSIEQSLST FSSHMTNIVD ILGKVD---H ---------- --ESLVLFDE
O.sativa_CAE     491   HQSLEHSLST FSGHISRLRK IVQVVS---K ---------- --DSLVLIDE
At_NP_176687     449   PQSLEQSLST FSGHISRIRQ ILDIAS---E ---------- --NSLVLLDE
At_E96674        449   PQSLEQSLST FSGHISRIRQ ILDIAS---E ---------- --NSLVLLDE
B.halodurans     377   GQSIEQNLST FSSRLTNIIH ILERAD---D ---------- --QTLVLVDE
At_NP_189075     718   YDSPVDGKSS FQVEMSEIRS IVSQAT---S ---------- --RSLVLIDE
At_BAB02932      712   YDSPVDGKSS FQVEMSEIRS IVSQAT---S ---------- --RSLVLIDE
At_AAO49798      814   YDSPVDGKSS FQVEMSEIRS IVSQAT---S ---------- --RSLVLIDE
B.cereus_NP_     647   ADDLISGQST FMVEMLEAKN AIANAS---E ---------- --RSLILFDE
B.halodurans     651   ADDLASGQST FMVEMLETKY ALQKAT---Q ---------- --NSLILLDE

                       6666666666 6666655545 5543330003 0000000000 0036677677
```

unaligned regions

relative alignment score

Multiple Sequence Alignment

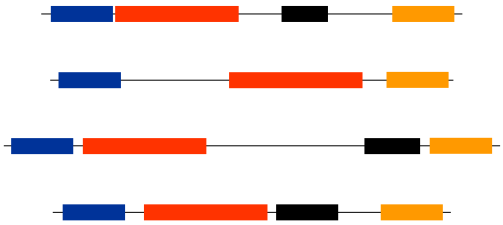


Sequences are related over their entire length

- Global MSA : Clustal
- MUSCLE* also good, and faster than Clustal for larger sets of sequences

  * see Edgar et al. (2004); http://dx.doi.org/0.1093/nar/gkh340

Multiple Sequence Alignment

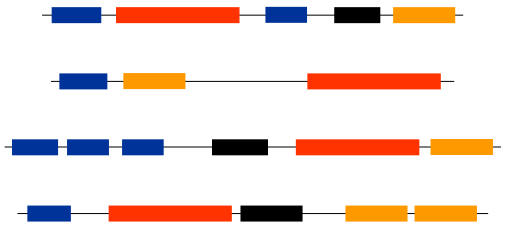


Sequences share conserved blocks separated by large insertions of unrelated material. Blocks are generally in the same order, but may not always present.

- Local MSA : DIALIGN
- T-Coffee* also useful (generates libraries of both global and local pairwise MSAs to start the progressive alignment process)

  * see Notredame *et al.* (2000); http://dx.doi.org/10.1006/jmbi.2000.4042

Multiple Sequence Alignment



Sequences contain a non-consistent set of conserved blocks.

- Motif-based MSA
- MAFFT will automatically adjust alignment algorithm to suit input sequences (similar or divergent)

MSA Hints – DNA vs. protein

- Proteins are easier to align than DNA.
- Therefore, if your DNA sequences are too divergent try aligning their amino acid translation, and then translating the sequence back to DNA

```
ATGATGGGGAGTCCCCTCGTT
ATGGGCGCCCCTATTGTG
```

unaligned DNA

↓

```
MMGSTIV
MGSTIV
```

unaligned translated protein

↓

```
MMGSTIV
M-GSTIV
```

aligned translated protein

↓

```
 M  M  G  S  T  I  V
ATGATGGGGAGTCCCCTCGTT
ATG---GGCGCCCCTATTGTG
 M  -  G  S  T  I  V
```

aligned DNA

MSA Warning

- MSA algorithms assume that sequences are homologous.
- MSA programs will align anything and all sequences, even if they are not homologous.
- If it looks wrong it probably is wrong!