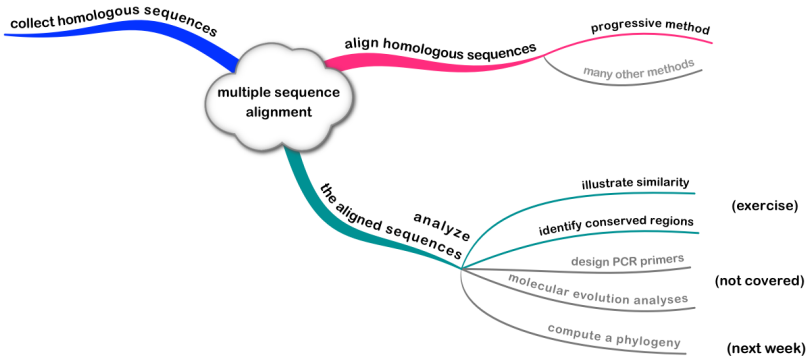
Bioinformatik

Stefanie Hartmann

Wintersemester 2019 / 2020, Universität Potsdam

Multiple sequence alignments
Nov 22, 2019

Overview



Multiple sequence alignment (MSA)

input:

```
ENKSSVLFYGGGAIIAVWLSSIVVKAVDSVPMFGVLPNILELVGLGYSGWFVYR
EDKYAVTAIGVAALVGLWTAIGAIKAIDRLPMFGINGLLPGVLELVGIGYTGWFTYR
EDLFALAGIGFAGIAALWASINLVEVIDKLPVLPLLFELIGILVAWLFIYQ
QVTPTLVLYGSGALVVLWVLSAIDSIPLVPKVLELIGTGYSIWFTSR
ENSYTALVSALVTIWISSIVVSALDSVPLVPQVMEVVGLGFTVWFTSR
KETSTFVMYGSGAFIAGWILSAVVSAIDSIPLFPKILQIVGLGYTIWFSTR
EDKPTFLLYSGGAVVALWLTTVVVGAINSVPLLPKILELVGLGYTGWFVYR
```

MSA algorithm

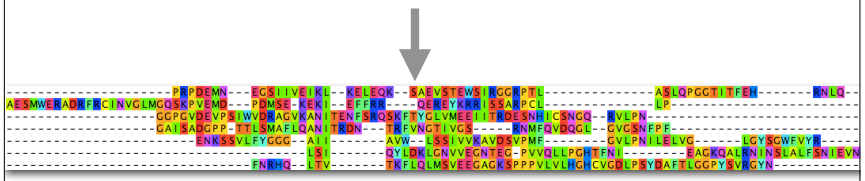
output:

```
EDLFALAGIGFAGIAALWASINLVEVIDKLPV------LPLLFELIGILVAWL
ENSYTA---LVSALVTIWISSIVVSALDSVPL------VPQVMEVVGLGFTVW
EDKYAVTAIGVAALVGLWTAIGAIKAIDRLPMFGINGLLPGVLELVGIGYTGW
ENKSSVLFYGGGAIIAVWLSSIVVKAVDSVPMFG---VLPNILELVGLGYSGW
EDKPTFLLYSGGAVVALWLTTVVVGAINSVPL------LPKILELVGLGYTGW
QVTPTLVLYGSGALVVLWVLS----AIDSIPL------VPKVLELIGTGYSIW
KETSTFVMYGSGAFIAGWILSAVVSAIDSIPL------FPKILQIVGLGYTIW
```


MSA challenges (I)

- collecting homologous input sequences
  - input = non-homologous sequences:
    an alignment will be computed!
  - mathematically correct vs. biologically correct

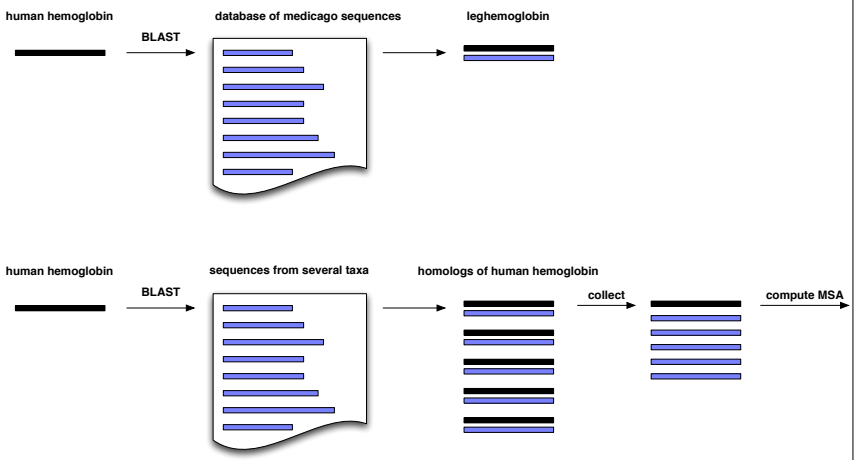
```
ENKSSVLFYGGGAIIAVWLSSIVVKAVDSVPMFGVLPNILELVGLGYSGWFVYR
GGPGVDEVPSIWVDRAGVKANITENFSRQSKFTYGLVMEEIITRDESNHICSNGQRVLPN
LSIQYLDKLGNVVEGNTEGPVVQLLPGHTFNIEAGKQALRNINSLALFSNIEVN
FNRHQLTVTKFLQLMSVEEGAGKSPPPVLVLHGHCVGDLPSYDAFTLGGPYSVRGYN
PRPDEMNEGSIIVEIKLKELEQKSAEVSTEWSIRGGRPTLASLQPGGTITFEHRNLQ
GAISADGPPTTLSMAFLQANITRDNTRFVNGTIVGSRNMFQVDQGLGVGSNFPF
AESMWERADRFRCINVGLMGQSKPVEMDPDMSEKEKIEFFRRQEREYKRRISSARPCLLP
```



MSA challenges (I)

- collecting homologous input sequences
  - input = non-homologous sequences:
    an alignment will be computed!
  - mathematically correct vs. biologically correct
- covered methods compute global multiple alignments
  - need globally homologous input sequences

Collecting homologous sequences



Collecting homologous sequences

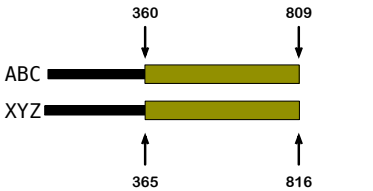
Query: sequence ID: ABC, length: 809 aa

Good BLAST hit: sequence ID: XYZ, length: 816 aa

```
sequence ID: seqXYZ    Length: 451

Score          Expect   Identities      Positives       Gaps
723 bits(1865) 0.0      335/451(74%)    389/451(86%)    1/451(0%)

Query  360  LSIQYDKLGNVVEGNTEPTQDELPKQLLPGHTFNIEAGKQALRNINSLALFSNIEVN
            ++ Q+ DKLGN VEGNT+ P++ RELP+QL PGH FNI AGKQAL+NINSLALFSNIEVN
Sbjct  365  VAYQFDKLGNFVEGNTQIPIDRELPQQLRPGHIFNIGAGKQALKNINSLALFSNIEVN

            PRPDEMNEGSIIVEIKLKELEQKSAEVSTEWSIVPGRGGRPTLASLQPGGTITFEHRNLQ
            PRPDE  EG I+VEIKLKELE KSAEVSTEWSIVPGR GRPTLAS+QPGGT++FEHRN+
            PRPDETKEGGIVVEIKLKELEPKSAEVSTEWSIVPGRQGRPTLASIQPGGTVSFEHRNIY

            GLNRSLTGSVTTSNFLNPQDDLAFKMEYAHPYLDGVDN-PRNRTLRVSCFNSRKLSPVFT
            GLNRS+ GSVT+SN LNPQDDL+FK+EY HPYLDGVD+  +NRT + SCFN+RKLSPVF
            GLNRSIVGSVTSSNLLNPQDDLSFKLEYVHPYLDGVDDRNKNRTFKTSCFNTRKLSPVFV

            GGPGVDEVPSIWVDRAGVKANITENFSRQSKFTYGLVMEEIITRDESNHTCSNGQRVLPN
             GP +DE P +WVDR G KANITE+F+RQSKFTYGLV+EEI TRDE+  IC++G R +P+
            AGPMMDEAPPVWVDRVGFKANITESFTRQSKFTYGLVVEEITTRDETYSICTHGSRAMPS

            GAISADGPPTTLSGTGIDRMAFLQANITRDNTRFVNGTIVGSRNMFQVDQGLGVGSNFPF
            G +S DGPPTTLSGTG+DRMAFLQANITRDNT FVNG ++G R +FQ+DQGL+GS  F
            GGLSMDGPPTTLSGTGVDRMAFLQANITRDNTEFVNGAVIGDRCIFQLDQGLGIGSKMFL

            FNRHQLTVTKFLQLMSVEEGAGKSPPPVLVLHGHYGGCVGDLPSYDAFTLGGPYSVRGYN
            FNRHQLT+TKF+ L + E+G GK  P VLVLHGHY GCVGDLPSYDAFTLGGPYSVRGY
            FNRHQLTLTKFINLNNQEKGVGKPLPAVLVLHGHYAGCVGDLPSYDAFTLGGPYSVRGYG

            MGEIGAARNILELAAEIRIPIKGTHVYAFAEHGTDLGSSKDVKGNFTVVYRRMGQGSSYG
            MGE+GA+RN+LE+A E+R+P+K T+VY FAEHGTDLGSSKDVKGNFT  +RR+G GSSYG
            MGELGASRNVLEVAGEVRVPVKNTYVYGFAEHGTDLGSSKDVKGNPTEFFRRVGHGSSYG

            AGMKLGLVRAEYAVDHNSGTGAVFFRTGERF  809
             G+KLGLVR EY VDHN+G G VFFRTGERF
            VGVKLGLVRGEYIVDHNTGAGTVFFRTGERF  816
```

```
Query  360  LSIQ
            ++ Q
Sbjct  365  VAYQ
```

```
FGERF  809
FGERF
FGERF  816
```



Types of multiple sequence alignments

local multiple sequence alignment

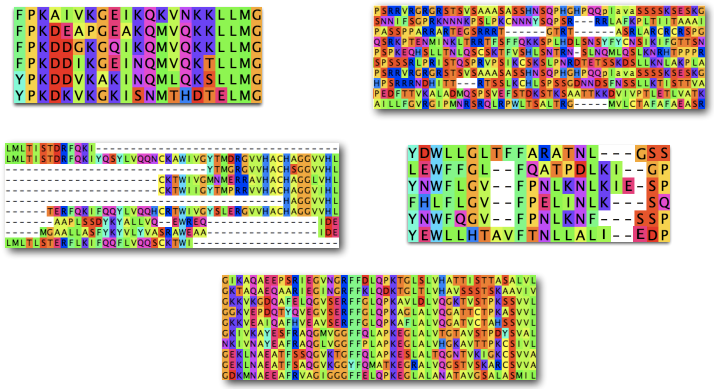
- contains aligned (and possibly unaligned) regions
- software: Dialign, HMMer

global multiple sequence alignment

- the software tries to align the sequences end to end
- most MSA software

MSA challenges (II)

- biological
  - criterion for accuracy?



MSA challenges (II)

- biological
  - criterion for accuracy?
  - reconcile multiple pw alignments into a msa

3 sequences

1 ACTG
2 ACG
3 ATG

optimal pairwise alignments

1 ACTG
2 AC–G

1 ACTG
3 A–TG

2 ACG
3 ATG

optimal multiple alignment

1 ACTG
2 AC–G
3 A–TG

MSA challenges (II)

- biological
  - criterion for accuracy?
  - reconcile multiple pw alignments into a msa
- computational
  - (mathematical) accuracy: no fast solution exists, all approaches use heuristics

Computing MSAs

algorithmic approaches

- many different types of heuristics exist
- progressive alignment is most frequently used

implementations (programs):

- Align-m, AMAP, BlastAlign, ClustalW, ClustalX, ClustalO, DCA, DIALIGN-2, HMMER, ITERALIGN, Kalign, MACAW, MAFFT, Match-Box, MAVID, MSA, Multalin, MULTIALIN, MUSCA, MUSCLE, Nomad, PCMA, PileUp, POA, PRALINE, Prank, ProAlign, ProbCons, PRRP, PSAlign, SAGA, SAM, SAM-T99, T-Coffee, ...

Progressive alignment

1. compute a pairwise distance matrix

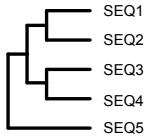
```
>SEQ1
EDRENMLCLSCYLHVTRKPLQSIKTAYLYFSVQSTGGKVG
>SEQ2
EDAENFLCLSCLLHVTGKPLQSIKTRYLYFSVQSTGGKVG
>SEQ3
EDFENMLGVHCLSCYLHVTRKILQSIKTAYLYFSIQSTLGKIG
>SEQ4
EDRENMLGVHCLGCNLHVTRKPLQSIKTAGGYFSIQSTLGKIG
>SEQ5
DDRDNMLGVLCLSNYLHVTAKPLQSTKTAYMYFSVGNRQSTPGKVG
```

	SEQ1	SEQ2	SEQ3	SEQ4	SEQ5
SEQ1	0.0				
SEQ2	12.5	0.0			
SEQ3	20.0	27.5	0.0		
SEQ4	25.0	32.5	14.0	0.0	
SEQ5	27.5	35.0	32.6	34.9	0.0

Progressive alignment

1. compute a pairwise distance matrix
2. use distance values to compute a guide tree

	SEQ1	SEQ2	SEQ3	SEQ4	SEQ5
SEQ1	0.0				
SEQ2	12.5	0.0			
SEQ3	20.0	27.5	0.0		
SEQ4	25.0	32.5	14.0	0.0	
SEQ5	27.5	35.0	32.6	34.9	0.0



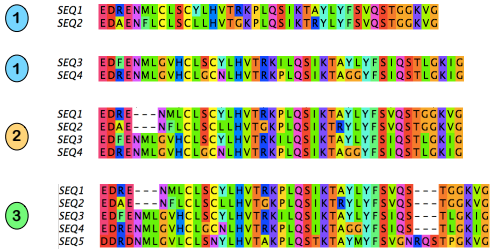
Progressive alignment

1. compute a pairwise distance matrix
2. use distance values to compute a guide tree
3. align sequences based on the guide tree
   - start with the most similar sequences
   - progressively add more distant sequences
   - once computed, subalignments are “frozen”

alignment

1. sequence – sequence
2. alignment – alignment
3. sequence – alignment

Progressive alignment



alignment

1. sequence – sequence
2. alignment – alignment
3. sequence – alignment

Growing the multiple sequence alignment

3 how to align a sequence to an alignment?

- compare sequence with all sequences in the group
- highest scoring pairwise alignment determines how sequence will be aligned to the group

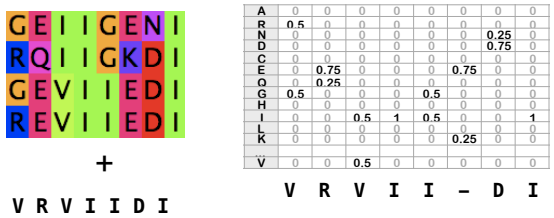
2 how to align an alignment to an alignment?

- compare all sequence pairs between the groups
- best pairwise alignment determines the alignment of the two groups

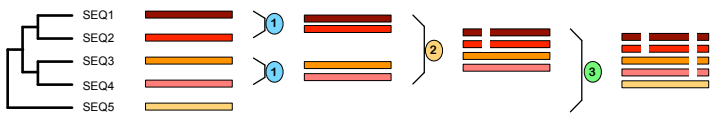
Growing the multiple sequence alignment

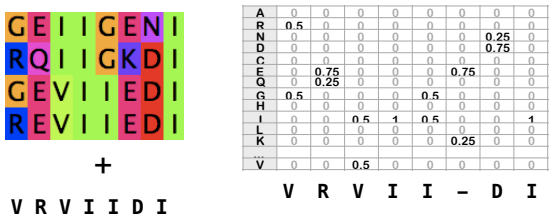
better: use profile alignments, not pairwise alignments

- a profile:
  - table: 20 rows, one for each amino acid, as many columns as the alignment has (or transposed)
  - profile columns contain information about how conserved the corresponding alignment column is

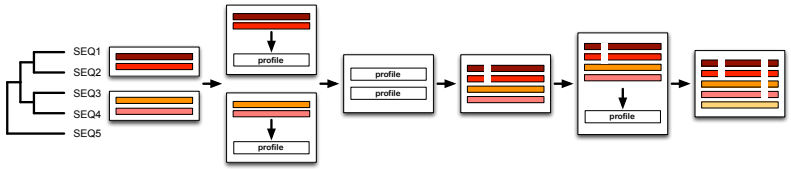


Growing the multiple sequence alignment





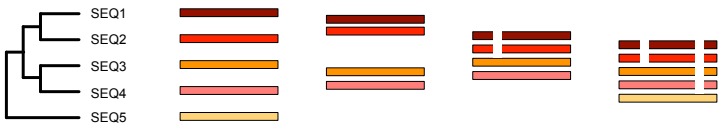
Progressive alignment using profiles



- align sequences 1 and 2, 3 and 4
- compute profiles for subalignments 1-2, 3-4
- align profile 1-2 to profile 3-4, discard profiles, result: subalignment 1-2-3-4
- compute profile for subalignment 1-2-3-4
- align 5 to profile 1-2-3-4, discard profile
- only report final alignment 1-2-3-4-5

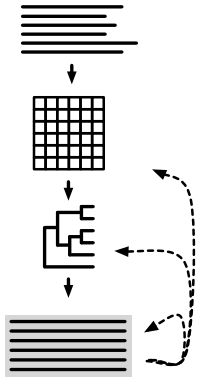
Progressive alignment: major weakness

- distant sequences present problems
- long insertions or deletions present problems
- once they are introduced, alignment errors cannot be corrected



Iterative refinement

1. compute an initial alignment
2. modify it
   - re-align a single sequence or re-align several sequences
   - evaluate the new alignment
   - repeat (n times, or until no further improvement is observed)
3. keep the best alignment



- slower than purely progressive algorithms

Sum-of-pairs score

- sum of scores of all <u>induced</u> pairwise alignments
- uses a substitution scoring matrix

identity: 1, mismatch: -1, gap: -2

```
seq1 ATTCACAGT
seq2 AT-CACGGT
seq3 ATTCAC-GT
```

score = 4+6+3 = 13

```
seq1 ATTCACAGT
seq2 AT-CACGGT
```

score = 4

```
seq1 ATTCACAGT
seq3 ATTCAC-GT
```

score = 6

```
seq2 AT-CACGGT
seq3 ATTCAC-GT
```

score = 3

Progressive alignment algorithm

implemented in the software ClustalW, ClustalX, ClustalO

- optional iterative refinement: groups of sequences are realigned
- position-specific gap penalties
  - increased gap penalties in flanking regions of a gap
  - lower gap penalty in regions already containing a gap
- residue-specific gap penalties
  - increased gap penalties in hydrophobic protein regions
- if alignment score for a sequence is low, it will be added later
- use of four different substitution matrices
- and many more

Progressive alignment & improvements

many more improvements
(many different programs)

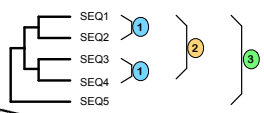
very fast heuristics
(ClustalO)

adjusted scoring
matrices & gaps

iterative
refinement

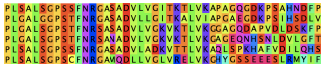
```
>SEQ1
EDRENMLCLSCYLHVTRKPLQSIKTAYLYFSVQSTGGKVG
>SEQ2
EDAENFLCLSCLLHVTGKPLQSIKTRYLYFSVQSTGGKVG
>SEQ3
EDFENMLGVHCLSCYLHVTRKILQSIKTAYLYFSIQSTLGKIG
>SEQ4
EDRENMLGVHCLGCNLHVTRKPLQSIKTAGGYFSIQSTLGKIG
>SEQ5
DDRDNMLGVLCLSNYLHVTAKPLQSTKTAYMYFSVGNRQSTPGKVG
```

	SEQ1	SEQ2	SEQ3	SEQ4	SEQ5
SEQ1	0.0				
SEQ2	12.5	0.0			
SEQ3	20.0	27.5	0.0		
SEQ4	25.0	32.5	14.0	0.0	
SEQ5	27.5	35.0	32.6	34.9	0.0



Now that you've got a MSA...

- design of PCR-primers
- evolutionary analysis of genes/organisms
- analysis of the alignment,
  characterization of conserved regions



  - amino acid (base) frequency
  - physico-chemical properties of amino acids (bases), substitution matrix
  - gaps
  - evaluation per alignment, per column/region

Clustal Conserved positions

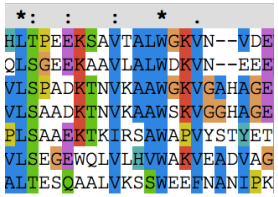
* positions which have a single, fully conserved residue.

: one of the following 'strong' groups is fully conserved:

STA	NEQK	NHQK	NDEQ	QHRK
MILV	MILF	HY	FYW	

. one of the following 'weaker' groups is fully conserved:

CSA	ATV	SAG	STNK
STPA	SGND	SNDEQK	NDEQHK
NEQHRK	FVLIM	HFY	



Clustal X Colour Scheme

Each residue in the alignment is assigned a colour if the amino acid profile of the alignment at that position meets some minimum criteria specific for the residue type.

The table below gives these criteria as clauses: **{+X%,xx,y}**, where X is the minimum percentage presence for any of the xx (or y) residue types.

Clustal X Default Colouring

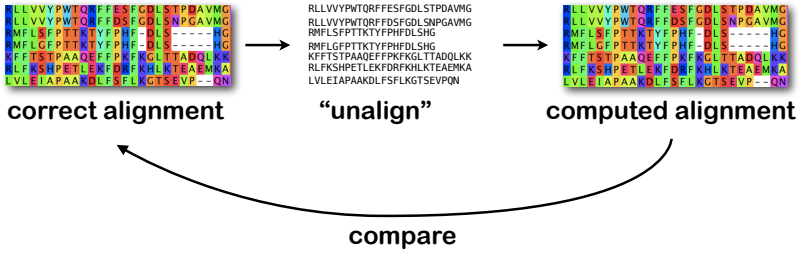
Residue at position	Applied Colour	{ Threshhold, Residue group }
A,I,L,M,F,W,V	BLUE	{+60%, WLVIMAFCHP}
R,K	RED	{+60%,KR},{+80%, K,R,Q}
N	GREEN	{+50%, N}, {+85%, N,Y}
C	BLUE	{+60%, WLVIMAFCHP}
C	PINK	{100%, C}
Q	GREEN	{+60%,KR},{+50%,QE},{+85%,Q,E,K,R}
E	MAGENTA	{+60%,KR},{+50%,QE},{+85%,E,Q,D}
D	MAGENTA	{+60%,KR}, {+85%, K,R,Q}, {+50%,ED}
G	ORANGE	{+0%, G}
H,Y	CYAN	{+60%, WLVIMAFCHP}, {+85%, W,Y,A,C,P,Q,F,H,I,L,M,V}
P	YELLOW	{+0%, P}
S,T	GREEN	{+60%, WLVIMAFCHP}, {+50%, TS}, {+85%,S,T}

MSA is a difficult problem!

- MSA is a hard problem, biologically and computationally
- no program or parameter set works for all sequence families
- refinement of alignments is often necessary (incl. exclusion of sequences/regions)
- many different challenges:
  - degree of sequence identity
  - insertions, deletions
  - presence of repeats, rearrangements
  - very long sequences
  - lots of sequences
  - local or global homology
  - etc

Evaluation of new methods

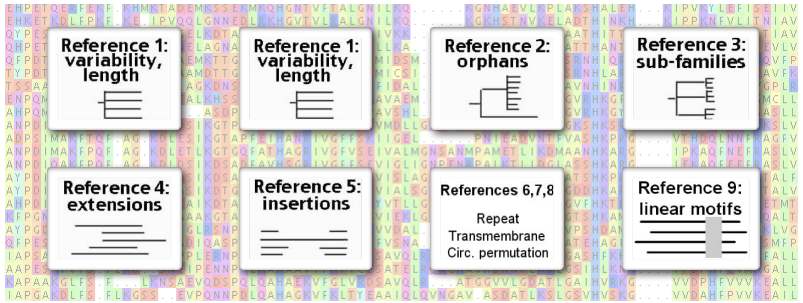
- simulated reference alignments
- manually curated reference alignments



Balibase: reference sequence alignments

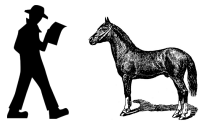
sets of sequences for which there are

- 3D structures
- corresponding high-quality multiple sequence alignments



Today's exercise

1. the data: seven globin homologs
   - hemoglobin beta: human_b, horse_b
   - hemoglobin alpha: human_a, horse_a
   - lamprey_g
   - myoglobin: whale_m
   - leghemoglobin: lupine_l





2. computing a multiple alignment (and guide tree)
   - ClustalW (linux terminal)
3. viewing a multiple alignment
   - read and evaluate a MSA (linux terminal, Jalview)

Key terms and concepts

- purpose & application of MSAs
- considerations for using BLAST to collect sequences for a MSA
- computing MSA: challenges
- progressive alignment
  - steps
  - profiles
  - advantages, disadvantages
  - iterative refinement
- sum of pairs score
- development & evaluation of new MSA methods