Multiple Sequence Alignment

Matej Ferenc

Jakub Repický

Why to use MSA?

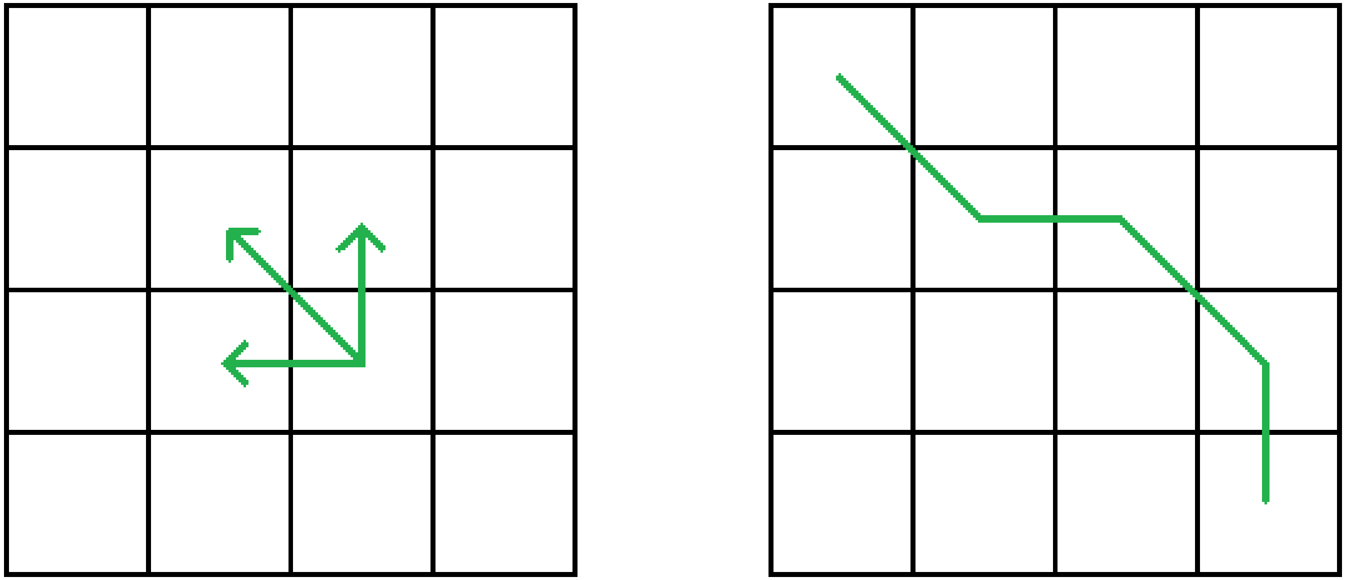
- To identify common conserved sequential motives and assess probability of their functional importance
- To obtain information about evolutionary relationships and history
- To construct phylogenetic trees

Pairwise Alignment

- Aligns two sequences
- We use dynamic programming
- Can be computed in O(nm)
- Parameters: gap penalties, substitution matrix
- We fill the matrix, always using maximum of three previously computed values:

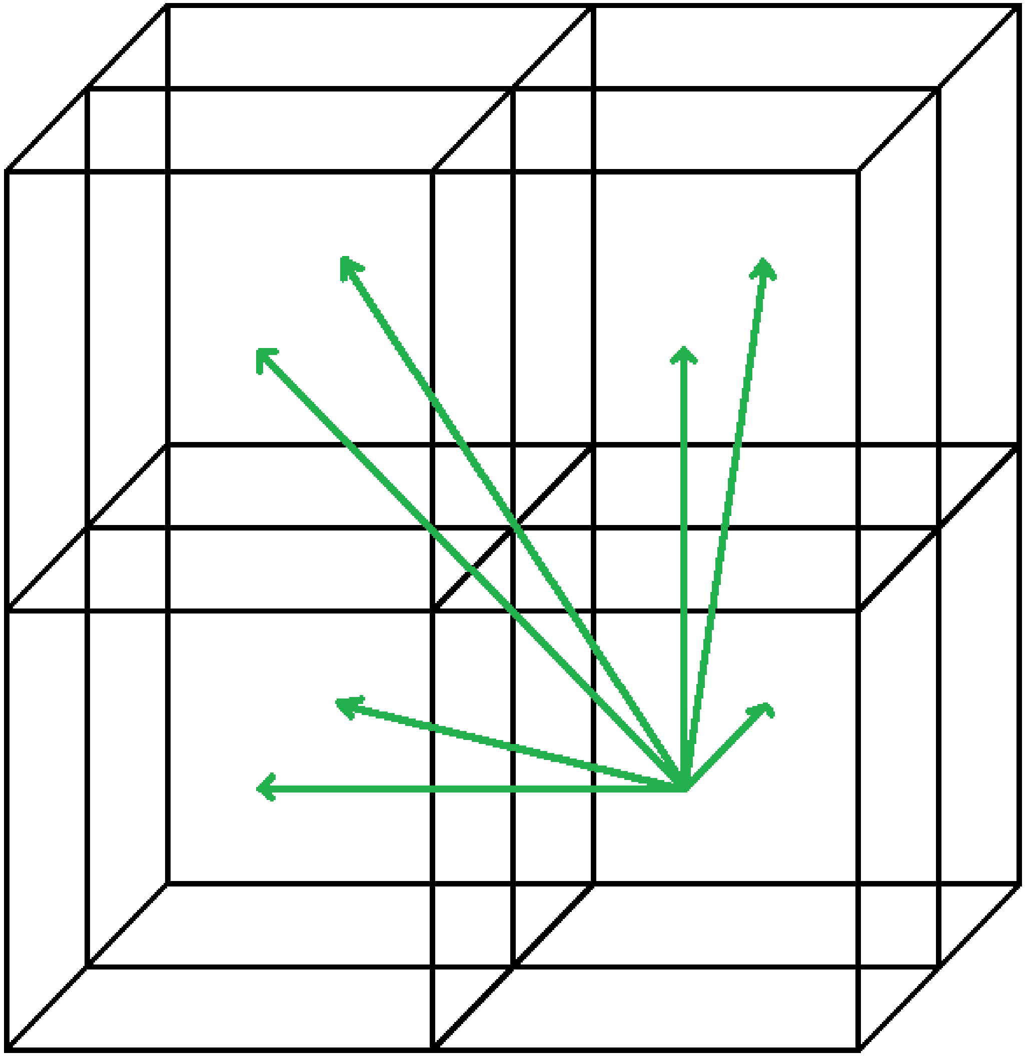
$$f_{i,j} = \max\{f_{i-1,j-1} + s(a_i, b_j),\ f_{i-1,j} + gap,\ f_{i,j-1} + gap\}$$

Pairwise Alignment



- For each cell compute maximum of three neighbouring cells

3-D Alignment



- For each cell compute maximum of seven neighbouring "cubes"

3-D Alignment

$$f_{i,\,j,\,k} = \max \left\{ \begin{array}{ll} f_{i-1,\,j-1,\,k-1} & + \, s(a_i,b_j,c_k) \\ f_{i-1,\,j-1,\,k} & + \, s(a_i,b_j,\_) \\ f_{i-1,\,j,\,k-1} & + \, s(a_i,\_,c_k) \\ f_{i,\,j-1,\,k-1} & + \, s(\_,b_j,c_k) \\ f_{i-1,\,j,\,k} & + \, s(a_i,\_,\_) \\ f_{i,\,j-1,\,k} & + \, s(\_,b_j,\_) \\ f_{i,\,j,\,k-1} & + \, s(\_,\_,c_k) \end{array} \right\}$$

Where s is a 3-dimensional substitution matrix

k-D Alignment

- Assume we want to align k sequences, each n symbols long. We need to fill a k-dimensional array, thus running time is O(n^k).
- Because of exponential running time, we don't usually use k-dimensional multiple alignment
- Although this can be improved by Carrilo-Lipman Heuristic which sets a bound of the score of alignment so that not all regions of the dynamic programming lattice have to be explored

Back to pairwise Alignment

- Can we align more than two sequences using only pairwise alignment?
- Idea: assume two aligned sequences, we will call it a profile. We can easily extend the pairwise alignment to work with profiles

```
                                ATGAGATC
ATAGTTC + ATGAGATC  =           AT-AGTTC

                                ATGAGATC
ATGAGATC                        AT-AGTTC
AT-AGTTC  +  TTGAGTC     =      TTGAGT-C
```

Aligning Profiles

		A A	T T	G _	A A	G G	A T	T T	C C
T									
G									
A									
G									
T									
A									
C									

- The algorithm works similarly, but computing the substitution value is a little different

$$S(a_i, b_j) = \sum_x (P_i(x) \sum_y (P_j(y) \cdot s(x, y)))$$

Progressive Multiple Alignment

- In what order should we add sequences to the profile?
- Generally, a tree model is preferred as it is biologically most relevant. First align most similar sequences and then add them to the rest of the sequences.
- We will need a similarity matrix

Similarity matrix

	s1	s2	s3	s4	s5	s6
s1	-	-	-	-	-	-
s2	0.17	-	-	-	-	-
s3	0.59	0.60	-	-	-	-
s4	0.59	0.59	0.13	-	-	-
s5	0.77	0.77	0.75	0.75	-	-
s6	0.81	0.82	0.73	0.74	0.80	-

- At each step we combine two most similar clusters.
- Similarity of two clusters A and B is defined as an average of similarities of pairs of sequences in A and B

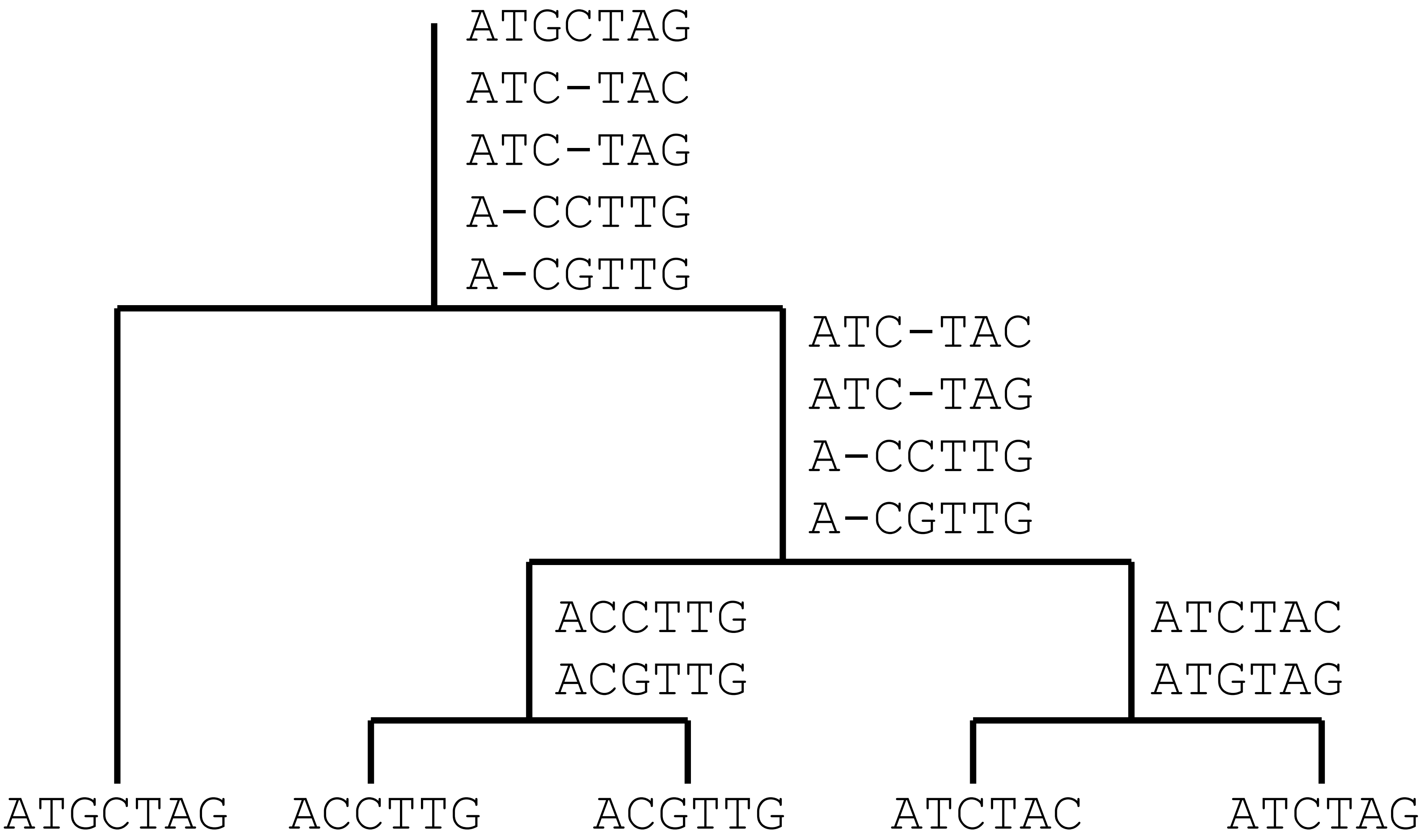
$$S(A,B)=\frac{1}{|A|\cdot|B|}\sum_{x\in A}\sum_{y\in B}s(x,y)$$

- This method is called Unweighted Pair Group Method with Arithmetic mean (UPGMA)

Dendrograms

- Are created by methods like UPGMA or Neighbour-joining.
- Concern an evolutionary distance of sequences
- Also called Guide Trees

Dendrograms - example



ATGCTAG
ATC-TAC
ATC-TAG
A-CCTTG
A-CGTTG

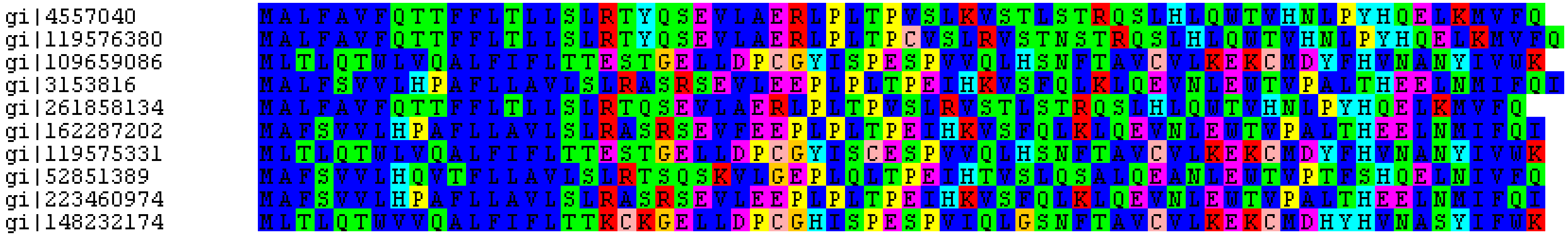
ATC-TAC
ATC-TAG
A-CCTTG
A-CGTTG

ACCTTG
ACGTTG

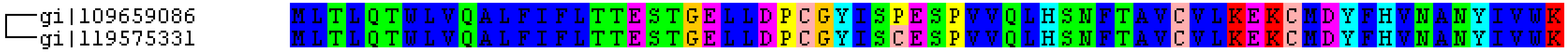
ATCTAC
ATGTAG

ATGCTAG ACCTTG ACGTTG ATCTAC ATCTAG

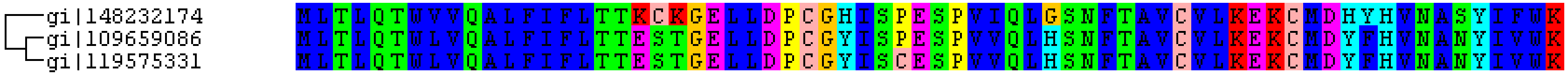
Example



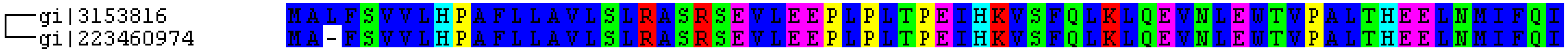
```
gi|4557040     MALFAVFQTTFFLTLLSLRTYQSEVLAERLPLTPVSLKVSTLSTRQSLHLQWTVHNLPYHQELKMVFQ
gi|119576380   MALFAVFQTTFFLTLLSLRTYQSEVLAERLPLTPCVSLRVSTNSTRQSLHLQWTVHNLPYHQELKMVFQ
gi|109659086   MLTLQTWLVQALFIFLTTESTGELLDPCGYISPESPVVQLHSNFTAVCVLKEKCMDYFHVNANYIVWK
gi|3153816     MALFSVVLHPAFLLAVLSLRASRSEVLEEPLPLTPEIHKVSFQLKLQEVNLEWTVPALTHEELNMIFQI
gi|261858134   MALFAVFQTTFFLTLLSLRTQSEVLAERLPLTPVSLRVSTLSTRQSLHLQWTVHNLPYHQELKMVFQ
gi|162287202   MAFSVVLHPAFLLAVLSLRASRSEVFEEPLPLTPEIHKVSFQLKLQEVNLEWTVPALTHEELNMIFQI
gi|119575331   MLTLQTWLVQALFIFLTTESTGELLDPCGYISCESPVVQLHSNFTAVCVLKEKCMDYFHVNANYIVWK
gi|52851389    MAFSVVLHQVTFLLAVLSLRTSQSKVLGEPLQLTPEIHTVSLQSALQEANLEWTVPTFSHQELNIVFQ
gi|223460974   MAFSVVLHPAFLLAVLSLRASRSEVLEEPLPLTPEIHKVSFQLKLQEVNLEWTVPALTHEELNMIFQI
gi|148232174   MLTLQTWVVQALFIFLTTKCKGELLDPCGHISPESPVIQLGSNFTAVCVLKEKCMDHYHVNASYIFWK
```



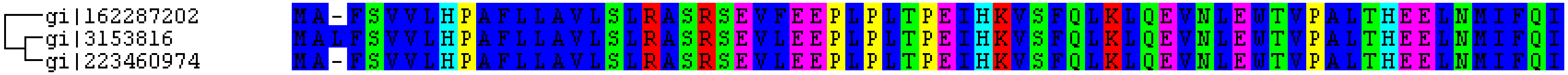
```
gi|109659086   MLTLQTWLVQALFIFLTTESTGELLDPCGYISPESPVVQLHSNFTAVCVLKEKCMDYFHVNANYIVWK
gi|119575331   MLTLQTWLVQALFIFLTTESTGELLDPCGYISCESPVVQLHSNFTAVCVLKEKCMDYFHVNANYIVWK
```



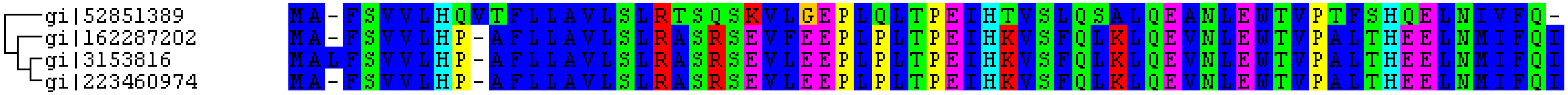
```
gi|148232174   MLTLQTWVVQALFIFLTTKCKGELLDPCGHISPESPVIQLGSNFTAVCVLKEKCMDHYHVNASYIFWK
gi|109659086   MLTLQTWLVQALFIFLTTESTGELLDPCGYISPESPVVQLHSNFTAVCVLKEKCMDYFHVNANYIVWK
gi|119575331   MLTLQTWLVQALFIFLTTESTGELLDPCGYISCESPVVQLHSNFTAVCVLKEKCMDYFHVNANYIVWK
```



```
gi|3153816     MALFSVVLHPAFLLAVLSLRASRSEVLEEPLPLTPEIHKVSFQLKLQEVNLEWTVPALTHEELNMIFQI
gi|223460974   MA-FSVVLHPAFLLAVLSLRASRSEVLEEPLPLTPEIHKVSFQLKLQEVNLEWTVPALTHEELNMIFQI
```

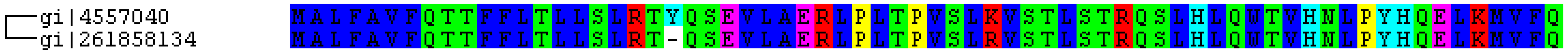


```
gi|162287202   MA-FSVVLHPAFLLAVLSLRASRSEVFEEPLPLTPEIHKVSFQLKLQEVNLEWTVPALTHEELNMIFQI
gi|3153816     MALFSVVLHPAFLLAVLSLRASRSEVLEEPLPLTPEIHKVSFQLKLQEVNLEWTVPALTHEELNMIFQI
gi|223460974   MA-FSVVLHPAFLLAVLSLRASRSEVLEEPLPLTPEIHKVSFQLKLQEVNLEWTVPALTHEELNMIFQI
```

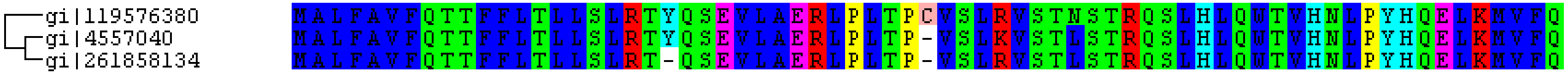


```
gi|52851389    MA-FSVVLHQVTFLLAVLSLRTSQSKVLGEPLQLTPEIHTVSLQSALQEANLEWTVPTFSHQELNIVFQ-
gi|162287202   MA-FSVVLHP-AFLLAVLSLRASRSEVFEEPLPLTPEIHKVSFQLKLQEVNLEWTVPALTHEELNMIFQI
gi|3153816     MALFSVVLHP-AFLLAVLSLRASRSEVLEEPLPLTPEIHKVSFQLKLQEVNLEWTVPALTHEELNMIFQI
gi|223460974   MA-FSVVLHP-AFLLAVLSLRASRSEVLEEPLPLTPEIHKVSFQLKLQEVNLEWTVPALTHEELNMIFQI
```

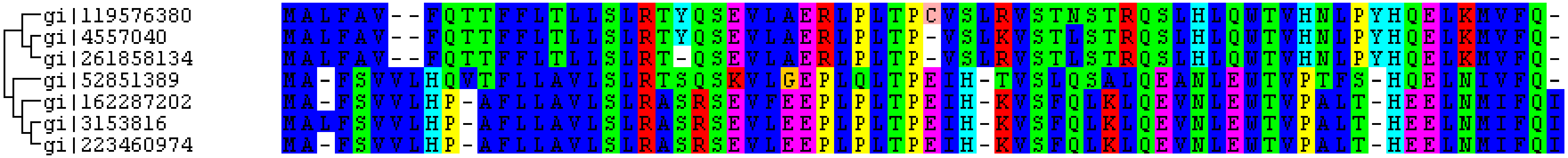
Progressive Multiple Alignment



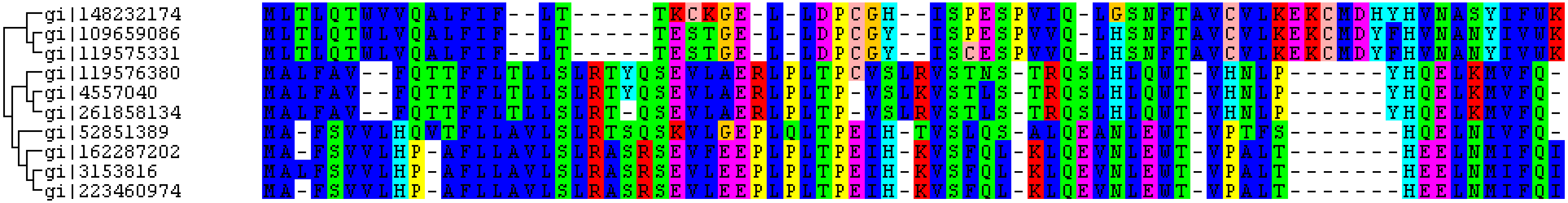
```
gi|4557040     MALFAVFQTTFFLTLLSLRTYQSEVLAERLPLTPVSLKVSTLSTRQSLHLQWTVHNLPYHQELKMVFQ
gi|261858134   MALFAVFQTTFFLTLLSLRT-QSEVLAERLPLTPVSLRVSTLSTRQSLHLQWTVHNLPYHQELKMVFQ
```



```
gi|119576380   MALFAVFQTTFFLTLLSLRTYQSEVLAERLPLTPCVSLRVSTNSTRQSLHLQWTVHNLPYHQELKMVFQ
gi|4557040     MALFAVFQTTFFLTLLSLRTYQSEVLAERLPLTP-VSLKVSTLSTRQSLHLQWTVHNLPYHQELKMVFQ
gi|261858134   MALFAVFQTTFFLTLLSLRT-QSEVLAERLPLTP-VSLRVSTLSTRQSLHLQWTVHNLPYHQELKMVFQ
```



```
gi|119576380   MALFAV--FQTTFFLTLLSLRTYQSEVLAERLPLTPCVSLRVSTNSTRQSLHLQWTVHNLPYHQELKMVFQ-
gi|4557040     MALFAV--FQTTFFLTLLSLRTYQSEVLAERLPLTP-VSLKVSTLSTRQSLHLQWTVHNLPYHQELKMVFQ-
gi|261858134   MALFAV--FQTTFFLTLLSLRT-QSEVLAERLPLTP-VSLRVSTLSTRQSLHLQWTVHNLPYHQELKMVFQ-
gi|52851389    MA-FSVVLHQVTFLLAVLSLRTSQSKVLGEPLQLTPEIH-TVSLQSALQEANLEWTVPTFS-HQELNIVFQ-
gi|162287202   MA-FSVVLHP-AFLLAVLSLRASRSEVFEEPLPLTPEIH-KVSFQLKLQEVNLEWTVPALT-HEELNMIFQI
gi|3153816     MALFSVVLHP-AFLLAVLSLRASRSEVLEEPLPLTPEIH-KVSFQLKLQEVNLEWTVPALT-HEELNMIFQI
gi|223460974   MA-FSVVLHP-AFLLAVLSLRASRSEVLEEPLPLTPEIH-KVSFQLKLQEVNLEWTVPALT-HEELNMIFQI
```



```
gi|148232174   MLTLQTWVVQALFIF--LT-----TKCKGE-L-LDPCGH--ISPESPVIQ-LGSNFTAVCVLKEKCMDHYHVNASYIFWK
gi|109659086   MLTLQTWLVQALFIF--LT-----TESTGE-L-LDPCGY--ISPESPVVQ-LHSNFTAVCVLKEKCMDYFHVNANYIVWK
gi|119575331   MLTLQTWLVQALFIF--LT-----TESTGE-L-LDPCGY--ISCESPVVQ-LHSNFTAVCVLKEKCMDYFHVNANYIVWK
gi|119576380   MALFAV--FQTTFFLTLLSLRTYQSEVLAERLPLTPCVSLRVSTNS-TRQSLHLQWT-VHNLP------YHQELKMVFQ-
gi|4557040     MALFAV--FQTTFFLTLLSLRTYQSEVLAERLPLTP-VSLKVSTLS-TRQSLHLQWT-VHNLP------YHQELKMVFQ-
gi|261858134   MALFAV--FQTTFFLTLLSLRT-QSEVLAERLPLTP-VSLRVSTLS-TRQSLHLQWT-VHNLP------YHQELKMVFQ-
gi|52851389    MA-FSVVLHQVTFLLAVLSLRTSQSKVLGEPLQLTPEIH-TVSLQS-ALQEANLEWT-VPTFS-------HQELNIVFQ-
gi|162287202   MA-FSVVLHP-AFLLAVLSLRASRSEVFEEPLPLTPEIH-KVSFQL-KLQEVNLEWT-VPALT-------HEELNMIFQI
gi|3153816     MALFSVVLHP-AFLLAVLSLRASRSEVLEEPLPLTPEIH-KVSFQL-KLQEVNLEWT-VPALT-------HEELNMIFQI
gi|223460974   MA-FSVVLHP-AFLLAVLSLRASRSEVLEEPLPLTPEIH-KVSFQL-KLQEVNLEWT-VPALT-------HEELNMIFQI
```

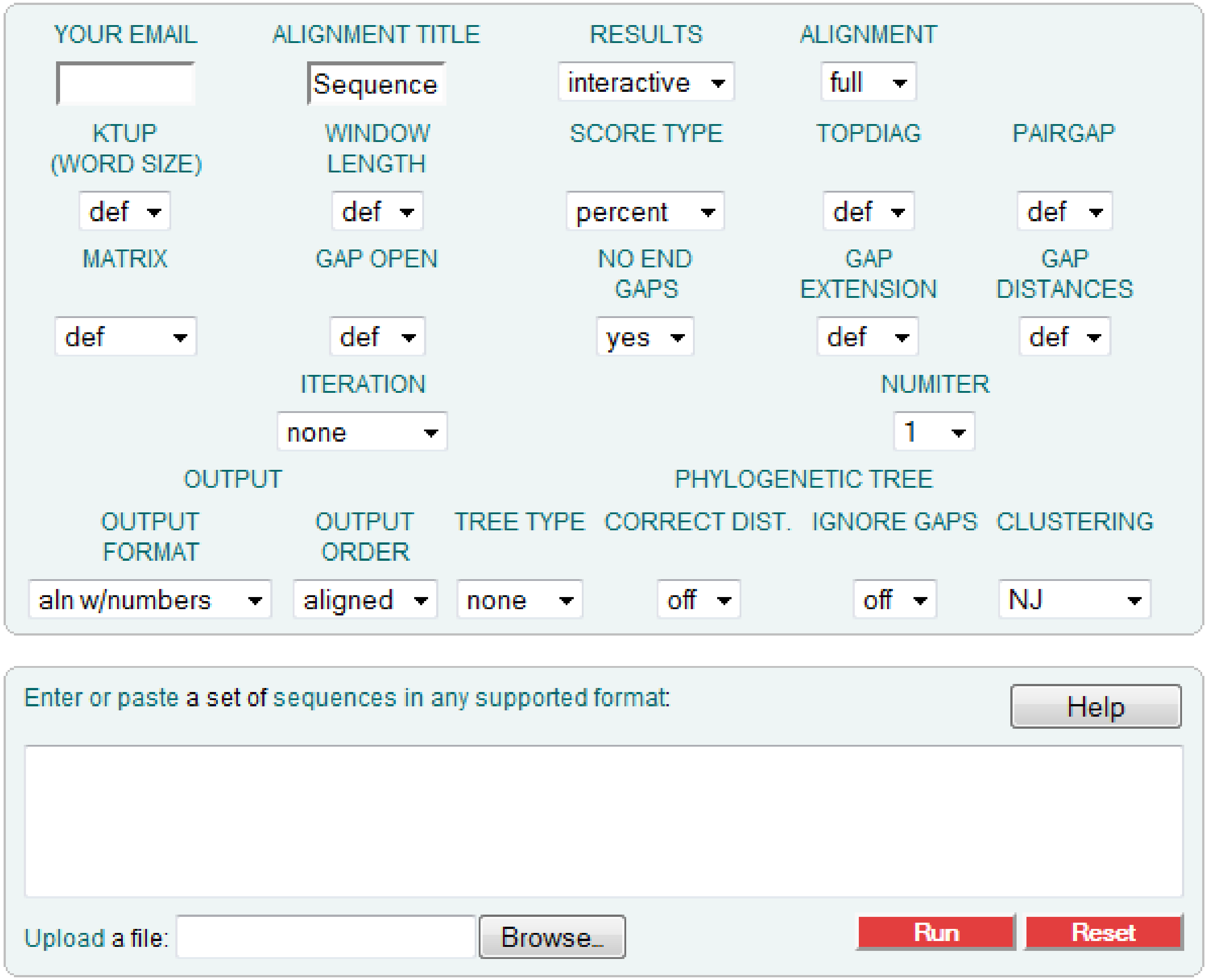
Progressive Multiple Alignment

- Once we have inserted a gap into sequence, it stays there
- Therefore we have to build strong initial alignments
- Clustal, T-Coffee

ClustalW

- Distance Matrix (Pairwise Alignments)
- Guide Tree
- Progressive Alignment
- Gap Open Penalty, Gap Extension Penalty
  - Similarity of sequences
  - Lengths of sequences
  - "GOP->(GOP+log(MIN(N,M))) * (average residue mismatch score) * (percent identity scaling factor)"
  - "GEP -> GEP*(1.0+|log(N/M)|)"
- 80-100%: PAM20, 60-80%: PAM60, 40-60%: PAM120, 0-40%: PAM350.
- 80-100%: BLOSUM80, 60-80%: BLOSUM62, 30-60%: BLOSUM45, 0-30%: BLOSUM30

ClustalW



YOUR EMAIL	ALIGNMENT TITLE	RESULTS	ALIGNMENT	
	Sequence	interactive	full	
KTUP (WORD SIZE)	WINDOW LENGTH	SCORE TYPE	TOPDIAG	PAIRGAP
def	def	percent	def	def
MATRIX	GAP OPEN	NO END GAPS	GAP EXTENSION	GAP DISTANCES
def	def	yes	def	def
	ITERATION		NUMITER	
	none		1	

OUTPUT		PHYLOGENETIC TREE			
OUTPUT FORMAT	OUTPUT ORDER	TREE TYPE	CORRECT DIST.	IGNORE GAPS	CLUSTERING
aln w/numbers	aligned	none	off	off	NJ

Enter or paste a set of sequences in any supported format:

Upload a file:

Iterative Multiple Alignment

- When constructing alignment, it realigns sequences already aligned
- Variety of methods exists
- For example: after the alignment is done, remove a sequence and add it to the alignment again
- MUSCLE (multiple sequence comparison by log-expectation)

Other methods

- Many other methods have been used to align more sequences
- Hidden Markov Models, Motif finding, Genetic algorithms

Comparing Alignments

- How to find out which alignment is better?
- How do we mathematically define “better”?
- Sum of Pairs Score:

$$\texttt{SP}\begin{pmatrix}\texttt{ATC-TAC}\\ \texttt{ATC-TAG}\\ \texttt{A-CCTTG}\\ \texttt{A-CGTTG}\end{pmatrix} = \begin{array}{l}\texttt{SP(AAAA)} + \texttt{SP(TT--)} + \\ \texttt{SP(CCCC)} + \texttt{SP(--CG)} + \\ \texttt{SP(TTTT)} + \texttt{SP(TTAA)} + \\ \texttt{SP(CGGG)}\end{array}$$

$$\begin{array}{l}\texttt{SP(--CG)} = \texttt{s(-,-)} + \texttt{s(-,C)} + \texttt{s(-,G)} + \texttt{s(-,C)} \\ \qquad\qquad\qquad + \texttt{s(-,G)} + \texttt{s(C,G)}\end{array}$$

Comparing Alignments

- Entropy:

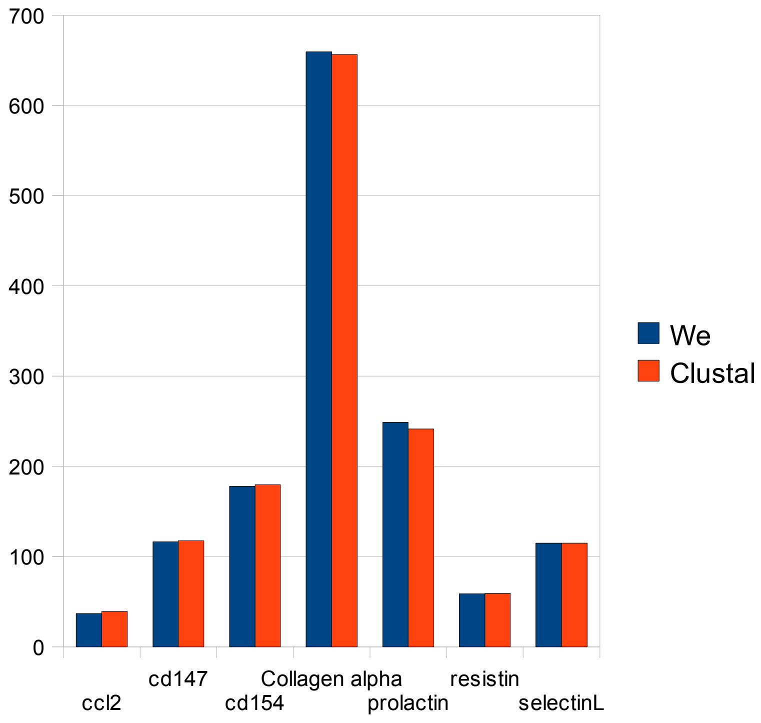
$$Entropy = \sum_{all\ columns} \sum_{x \in Alphabet} p_x \cdot \log(p_x)$$

- Alignments with lower entropy are better

Comparing Alignments

- Comparing our own method with Clustal using Entropy objective function

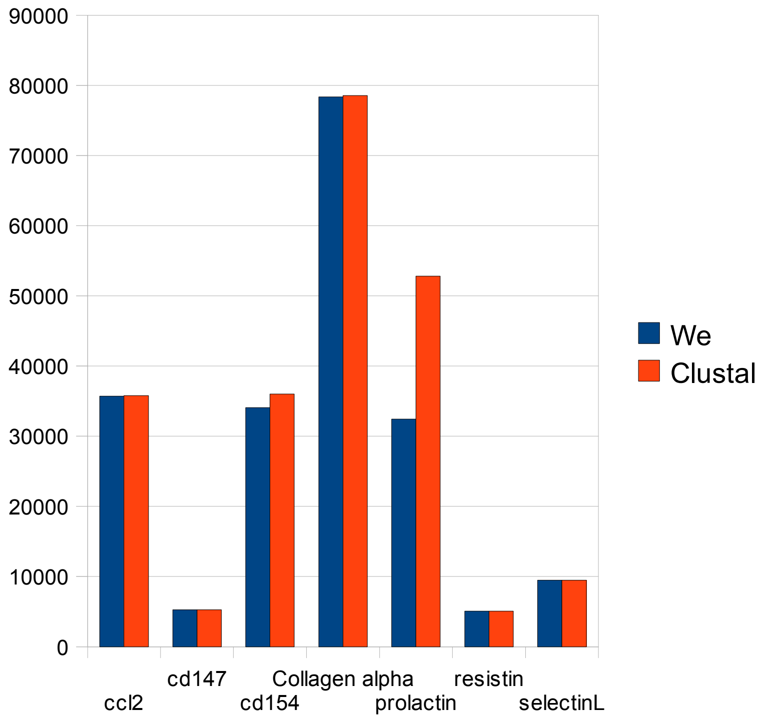
Protein	We	Clustal
ccl2	36.89	39.17
cd147	116.43	117.39
cd154	177.89	179.49
Collagen alpha	659.50	656.31
prolactin	248.77	241.32
resistin	58.65	59.40
selectinL	114.95	114.95



Comparing Alignments

- Comparing our own method with Clustal using Sum-of-Pairs objective function (Blosum62)

Protein	We	Clustal
ccl2	35686	35782
cd147	5279	5255
cd154	34064	36011
Collagen alpha	78360	78534
prolactin	32432	52804
resistin	5064	5057
selectinL	9481	9481


ccl2
cd147
cd154
Collagen alpha
prolactin
resistin
selectinL
We
Clustal