Multiple Sequence Alignment (I)

Jianlin Cheng, PhD

School of Electrical Engineering and Computer Science
University of Central Florida

2006

What's Multiple Sequence Alignment?

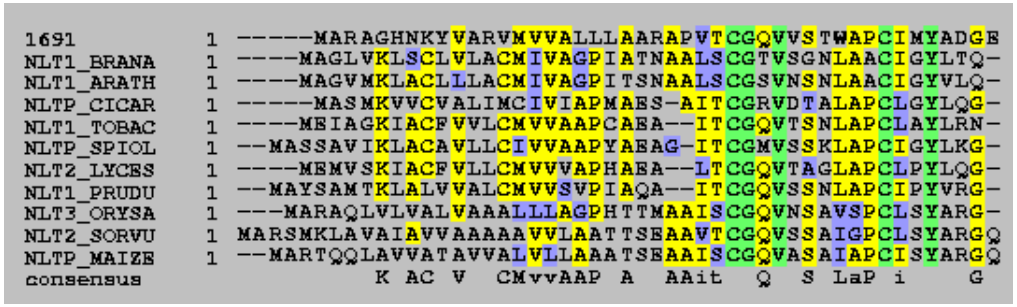
- A generalization of pairwise sequence alignment
- Align multiple sequences (e.g., a protein family) to maximize alignment score
- Biological goal is to identify evolutionary relationships, structures, and functions

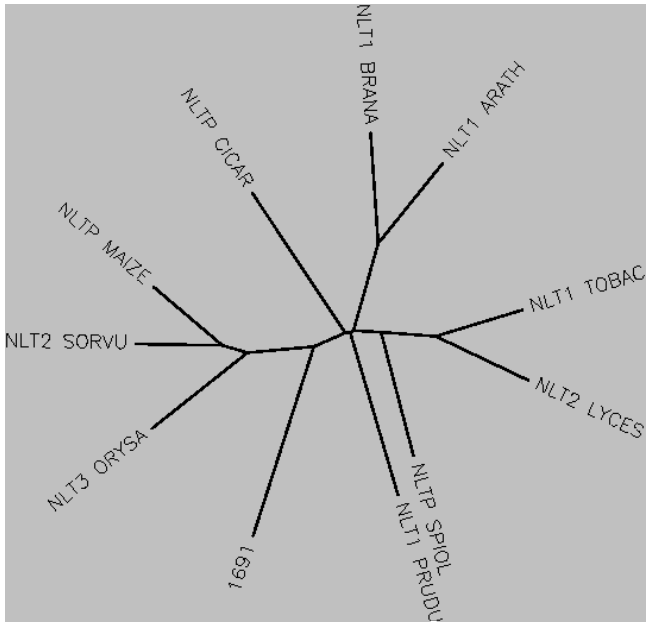
Why is MSA important?

- Construct phylogenetic tree
- Construct protein profile for database search and structure prediction
- Elucidate evolutionary conservation
- Identify DNA binding /regulatory sites
- Identify structural and functional important sites of proteins
- Improve alignment quality

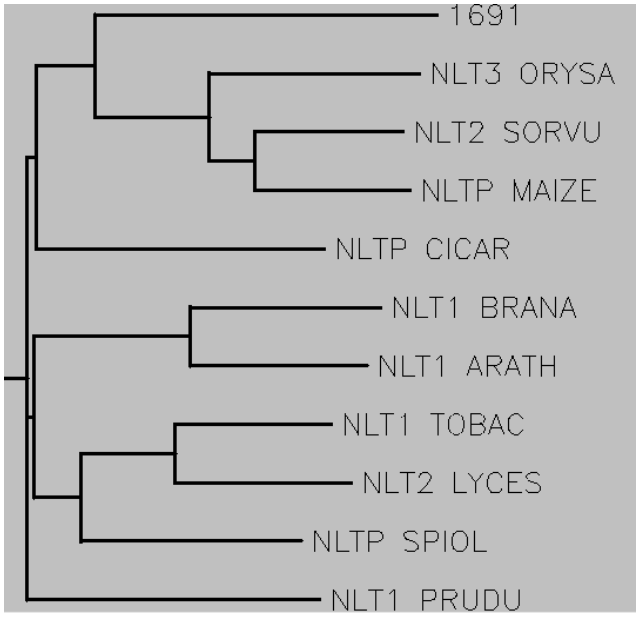
Example 1: plylogenetic tree

MSA





Phylip unrooted tree

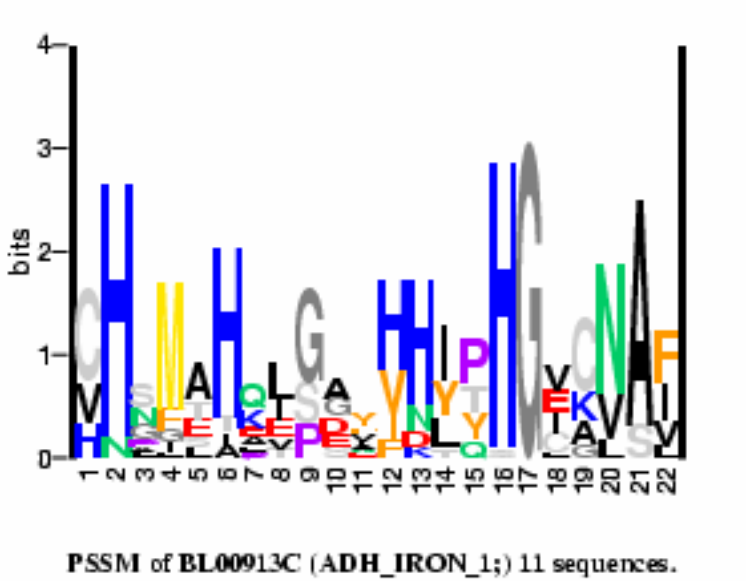


Phylip rooted tree

Example 2: Functional Important Residues / Site

```
FUCO_ECOLI VHGMAHPLGAFYNTPHGVANAI
GLDA_BACST HNGFTALEGEIHHLTHGEKVAF
GLDA_ECOLI VHNGLTAIPDAHHYYHGEKVAF
MEDH_BACMT VHSISHQVGGVYKLQHGICNSV
ADH1_CLOAB CHSMAHKTGAVFHIPHGCANAI
ADHE_ECOLI CHSMAHKLGSQFHIPHGLANAL
ADH2_ZYMMO VHAMAHQLGGYYNLPHGVCNAV
ADH4_YEAST VHALAHQLGGFYHLPHGVCNAV
ADHA_CLOAB CHPMEHELSAYYDITHGVGLAI
ADHB_CLOAB VHLMEHELSAYYDITHGVGLAI
```

iron containing alcohol dehydrogenases



PSSM of BL00913C (ADH_IRON_1;) 11 sequences.

conserved histidines probably bind the ferrous ion(s) required for these enzymes activity

Source: http://bioinformatics.weizmann.ac.il/~pietro/Making_and_using_protein_MA/

Scoring Function of MSA

- Amino acid level: PAM and BLOSUM
- Score of a column: sum of the scores of all residue pairs in the column.
- Score of an MSA: sum of the scores of all columns.

Example: Scoring Function

```
FUCO_ECOLI VHGMAHPLGAFYNTPHGVANAI
GLDA_BACST HNGFTALEGEIHHLTHGEKVAF
GLDA_ECOLI VHNGLTAIPDAHHYYHGEKVAF
MEDH_BACMT VHSISHQVGGVYKLQHGICNSV
```

BLOSUM62

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	9																			
S	-1	4																		
T	-1	1	5																	
P	-3	-1	-1	7																
A	0	1	0	-1	4															
G	-3	0	-2	-2	0	6														
N	-3	1	0	-2	-2	0	6													
D	-3	0	-1	-1	-2	-1	1	6												
E	-4	0	-1	-1	-1	-2	0	2	5											
Q	-3	0	-1	-1	-1	-2	0	0	2	5										
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	8									
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5								
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5							
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5						
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4					
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4				
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4			
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6		
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7	
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11

Column 3:

Residue pairs: (M,F), (M,G), (M,I), (F,G), (F,I), (G,I)

Scores: 0, -3, 1, -3, 0, -4

Score of column 3 is -9

For m sequences, number of residue pairs of each column: (m-1)*m / 2. So what's time complexity?

DP for Multiple Sequence Alignment

To align prefixes:
P[1..i], Q[1..j], R[1..k]

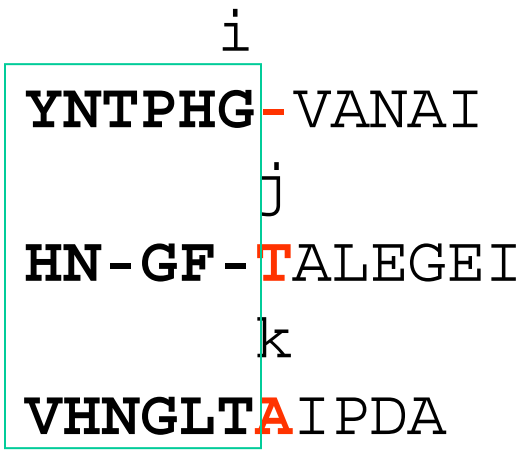
Consider $2^3 - 1 = 7$ situations

(1) Given P[1..i-1],Q[1..j-1],R[1..k-1], match P[i], Q[j], and R[k]

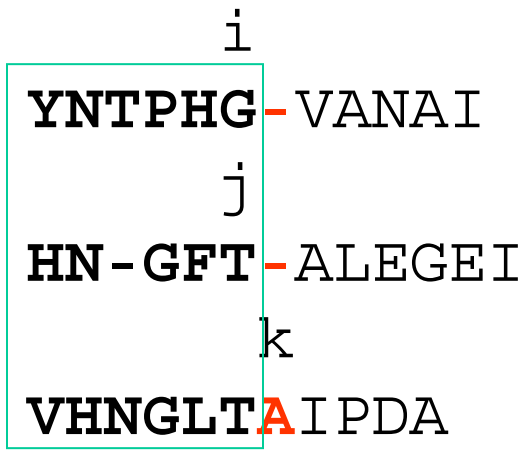
```
      i
YNTPHGVANAI
     j
HNGFTALEGEI
      k
VHNGLTAIPDA
```

```
          i
YNTP-HGVANAI
          j
HN-GF-TALEGEI
          k
VHNGLTAIPDA
```

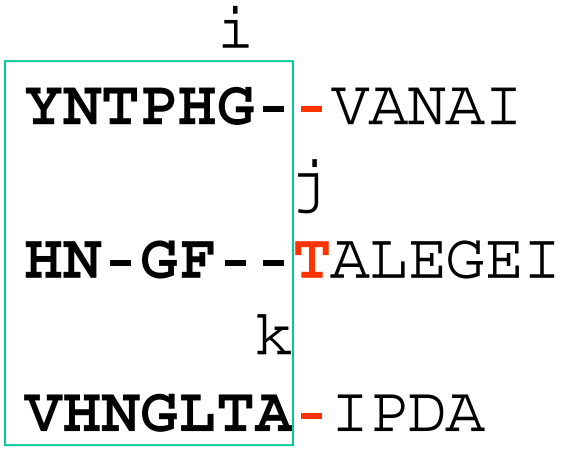
(2) Given P[1..i],Q[1..j-1],R[1..k-1], match gap, Q[j], and R[k]



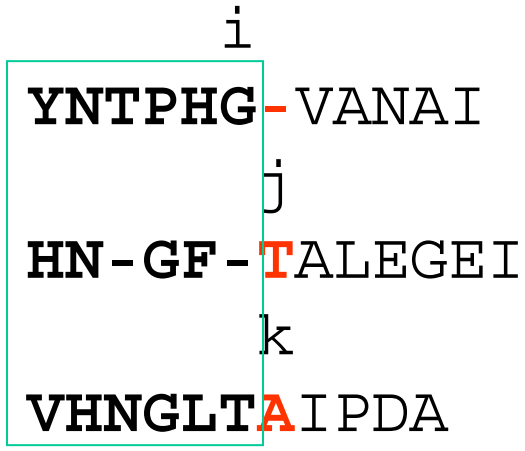
(3) Given P[1..i],Q[1..j],R[1..k-1], match gap, gap, and R[k]



(4) Given P[1..i],Q[1..j-1],R[1..k], match gap, Q[j], and gap



(5) Given P[1..i],Q[1..j-1],R[1..k-1], match gap, Q[j], and R[k]



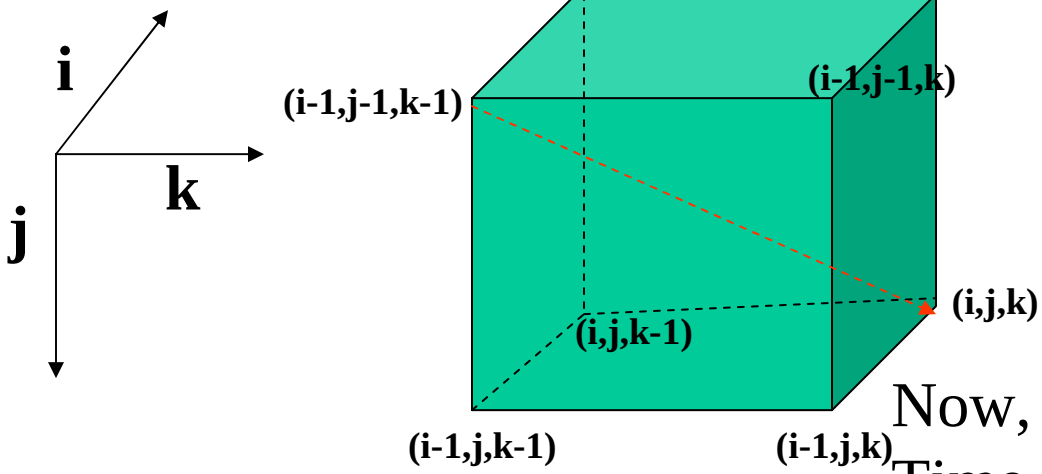
(6)Given P[1..i-1],Q[1..j],R[1..k-1], match P[i], gap, and R[k]

```
          i
YNTP-HGVANAI
       j
HN-GFT-ALEGEI
        k
VHNGLTAIPDA
```

(7)Given P[1..i-1],Q[1..j-1],R[1..k], match P[i], Q[j], and gap

```
          i
YNTP-H-GVANAI
         j
HN-GF--T-ALEGEI
        k
VHNGLTA-IPDA
```

Cubic Matrix



Now, fill a 3D matrix. Time = n^3. Each Step has $2^3 - 1$ paths.

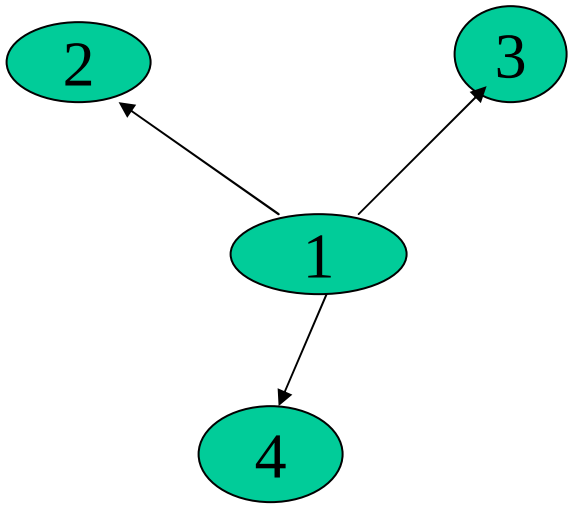
Now, generalized to m sequences Time Complexity = $n^m * 2^m * m^2$

Heuristic MSA Approaches

- DP can’t be used for more than three sequences
- Star approach
- Progressive approach
- Iterative approach
- Hidden Markov Model

Star Approach

Use one sequence of your interest or closest to all other sequences as seed. Align the seed to other sequences and generate a MSA from pairwise alignments.



(Used by PSI-BLAST)

Select a seed

1. VHGMAHPLAFYNTHGVANAI
2. HNGFTEGEIHHLTHGEKVF
3. VHNGLAIPDAHYHGEKVAF
4. VHSHQVGGVYKLQHGICNSV

Pairwise

```
1. VHGMAHPLAFYNT-HGVANAI
2. HNGFT-E-GEIHHLTHGEKVF

1. VHGMAHPL-AFYNTHGVANAI
3. VHNGL-AIPDA-HYHGEKVAF

1. VHGMAHPL-AFYNT-HGVANAI
4. VHS--HQVGGVYKLQHGICNSV
```

Do combination step by step using seed as an anchor

```
1. VHGMAHPLAFYNT-HGVANAI
2. HNGFT-E-GEIHHLTHGEKVF

1. VHGMAHPL-AFYNTHGVANAI
3. VHNGL-AIPDA-HYHGEKVAF
```

```
1. VHGMAHPL-AFYNT-HGVANAI
2. HNGFT-E--GEIHHLTHGEKVF
3. VHNGL-AIPDA-HY-HGEKVAF

1. VHGMAHPL-AFYNT-HGVANAI
4. VHS--HQVGGVYKLQHGICNSV
```

```
1. VHGMAHPL-AFYNT-HGVANAI
2. HNGFT-E--GEIHHLTHGEKVF
3. VHNGL-AIPDA-HY-HGEKVAF
4. VHS--HQVGGVYKLQHGICNSV
```

Progressive Approach

- Basic idea: align similar sequences first

  (1) construct a coarse phylogenetic tree

  (2) use the tree as guide tree to align sequences progressively.
- Representative methods: CLUSTALW and T-COFEE

Simplified Example of Progressive MSA

```
1. VHGMAHPLAFYNTHGVANAI
2. HNGFTEGEIHHLTHGEKVF
3. VHNGLAIPDAHYHGEKVAF
4. VHSHQVGGVYKLQHGICNSV
```

Align sequences pairwisely to construct a similarity matrix

	1	2	3	4
1	10	5	6	8
2	5	9	7	4
3	6	7	9	4
4	8	4	4	10

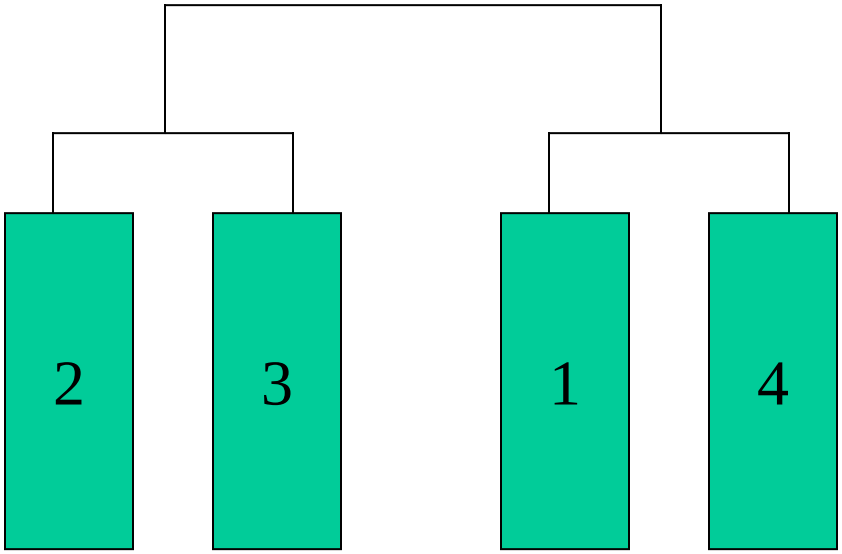
Step 1

Convert similarity matrix to distance matrix if necessary

$(S_{aa} + S_{bb})/2 - S_{ab}$

	1	2	3	4
1	0	4.5	3.5	2
2	4.5	0	2	5.5
3	3.5	2	0	5.5
4	2	5.5	5.5	0

Step 2



Construct a guide tree from distance/similarity matrix

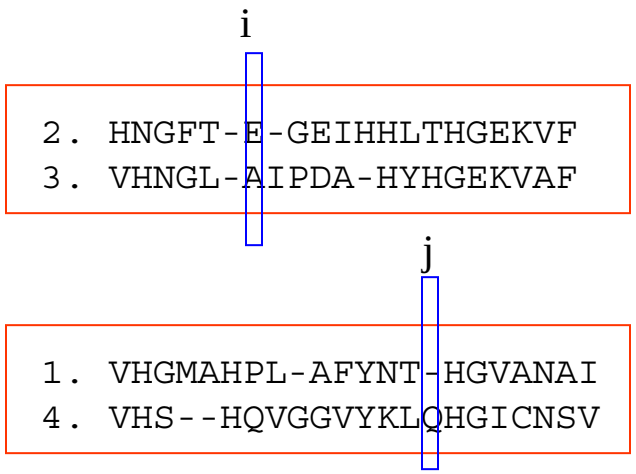
Step 3

Step 4

Align sequence progressively according to guide tree.

- Align 2 and 3
- Align 1 and 4
- Align group of 2-3 and group of 1-4

Question:
How to align a group of sequences against another group of sequences?



Score(i,j) = S(E,-)+S(E,Q)+S(A,-)+S(A,Q)

Align one group with another group:

- Treat each group as a single generalized sequence consisting of columns
- Use the same dynamic programming algorithm
- Once add one gap into a position in one group, all sequences in the group is added the gap
- Score(Col_i, Col_j) is the sum score of all residue pairs across column i and column j)

Some Progressive MSA Tools

- **CLUSTALW**(X) (neighbor-joining algorithm, similarity matrix)
- PILEUP (UPGMA algorithm for guide tree)
- MULTALIGN(UPGMA algorithm)
- T-COFFEE (improved over CLUSTALW, but slower)
- 3D-COFFEE (use structure information)

Iterative Approach

- Progressive approach is a heuristic approach. Optimal MSA is not guaranteed.
- Alignments in previous steps are fixed. Error propagation.
- Iterative approach tries to adjust alignments according to an objective function to improve alignments.

Iterative Methods

- PRRP: optimize a progressive alignment by iteratively dividing the sequences into two groups and realigning them using a group-to-group alignment algorithm
- SAGA: use a genetic algorithm
- Iterative approach often improves alignment accuracy at the expense of computation time

Hidden Markov Model Approach

- MUSCLE (http://www.drive5.com/muscle/, Robert Edgar)
- State of Art Multiple Sequence Alignment
- Online service: http://phylogenomics.berkeley.edu/cgi-bin/muscle/input_muscle.py
- Not discussed here, but in Advanced Bioinformatics (Spring, 2007)

A Case Study of Using CLUSTALW

HIV-1 Capsid Protein

VHQAISPRTLNAWVKVVEEKAFSPEVIPMFSALSEGATPQDLNTMLNTVGGHQA
AMQMLKETINEEAAEWDRVHPVHAGPIAPGQMREPRGSDIAGTTSTLQEQIGW
MTNNPPIPVGEIYKRWIILGLNKIVRMYSPTSILDIRQGPKEPFRDYVDRFYKTLR
AEQASQEVKNWMTETLLVQNANPDCKTILKALGPAATLEEMMTACQ

Search Database

POL-HV1BR

VHQAISPRTLNAWVKVVEEKAFSPEVIPMFSALSEGATPQDLNTMLNTVGGHQAAMQMLK
ETINEEAAEWDRVHPVHAGPIAPGQMREPRGSDIAGTTSTLQEQIGWMTNNPPIPVGEIY
KRWIILGLNKIVRMYSPTSILDIRQGPKEPFRDYVDRFYKTLRAEQASQEVKNWMTETLL
VQNANPDCKTILKALGPAATLEEMMTACQ

AAN23474

VHQAISPRTLNAWVKVIEEKAFSPEVIPMFSALSEGATPQDLNTMLNTVGGHQAAMQMLK
ETINEEAAEWDRIHPVHAGPIAPGQMREPRGSDIAGTTSTLQEQIGWMTSNPPIPVGEIY
KRWIILGLNKIVRMYSPVSILDIRQGPKEPFRDYVDRFYKTLRAEQASQDVKNWMTETLL
VQNANPDCKTILKALGPAATLEEMMTACQ

AAN73808

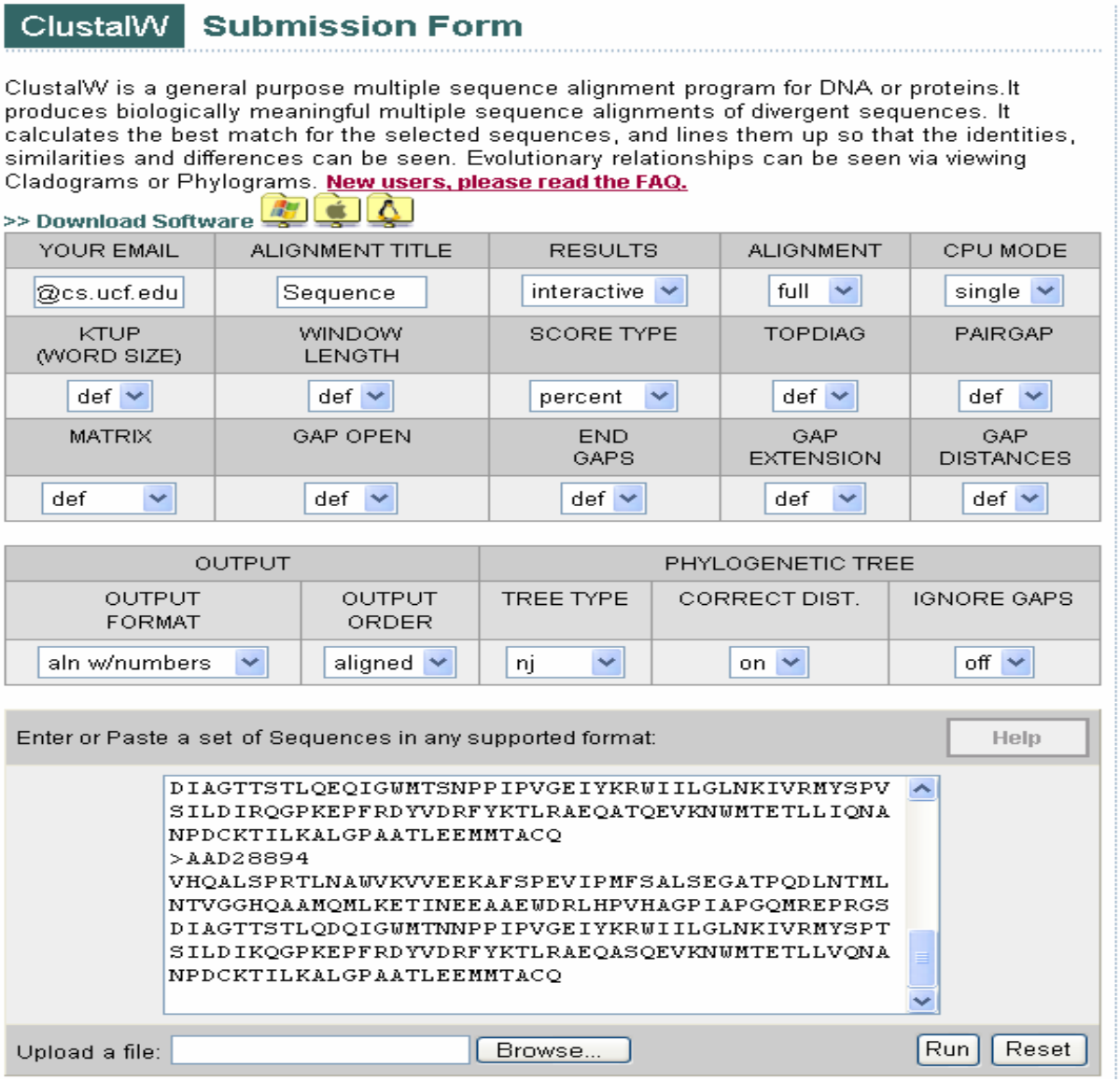
VHQAISPRTLNAWVKVIEEKAFSPEVIPMFSALSEGATPQDLNTMLNTVGGHQAAMQMLK
ETINEEAAEWDRLHPVHAGPVAPGQMRDPRGSDIAGTTSTLQEQIGWMTSNPPIPVGEIY
KRWIILGLNKIVRMYSPVSILDIRQGPKEPFRDYVDRFYKTLRAEQATQEVKNWMTETLL
IQNANPDCKTILKALGPAATLEEMMTACQ

AAD28894

VHQALSPRTLNAWVKVVEEKAFSPEVIPMFSALSEGATPQDLNTMLNTVGGHQAAMQMLK
ETINEEAAEWDRLHPVHAGPIAPGQMREPRGSDIAGTTSTLQDQIGWMTNNPPIPVGEIY
KRWIILGLNKIVRMYSPTSILDIKQGPKEPFRDYVDRFYKTLRAEQASQEVKNWMTETLL
VQNANPDCKTILKALGPAATLEEMMTACQ

ClustalW

- Web service:
  http://www.ebi.ac.uk/clustalw/



ClustalW Submission Form

ClustalW is a general purpose multiple sequence alignment program for DNA or proteins.It produces biologically meaningful multiple sequence alignments of divergent sequences. It calculates the best match for the selected sequences, and lines them up so that the identities, similarities and differences can be seen. Evolutionary relationships can be seen via viewing Cladograms or Phylograms. **New users, please read the FAQ.**

>> Download Software

YOUR EMAIL	ALIGNMENT TITLE	RESULTS	ALIGNMENT	CPU MODE
@cs.ucf.edu	Sequence	interactive	full	single
KTUP (WORD SIZE)	WINDOW LENGTH	SCORE TYPE	TOPDIAG	PAIRGAP
def	def	percent	def	def
MATRIX	GAP OPEN	END GAPS	GAP EXTENSION	GAP DISTANCES
def	def	def	def	def

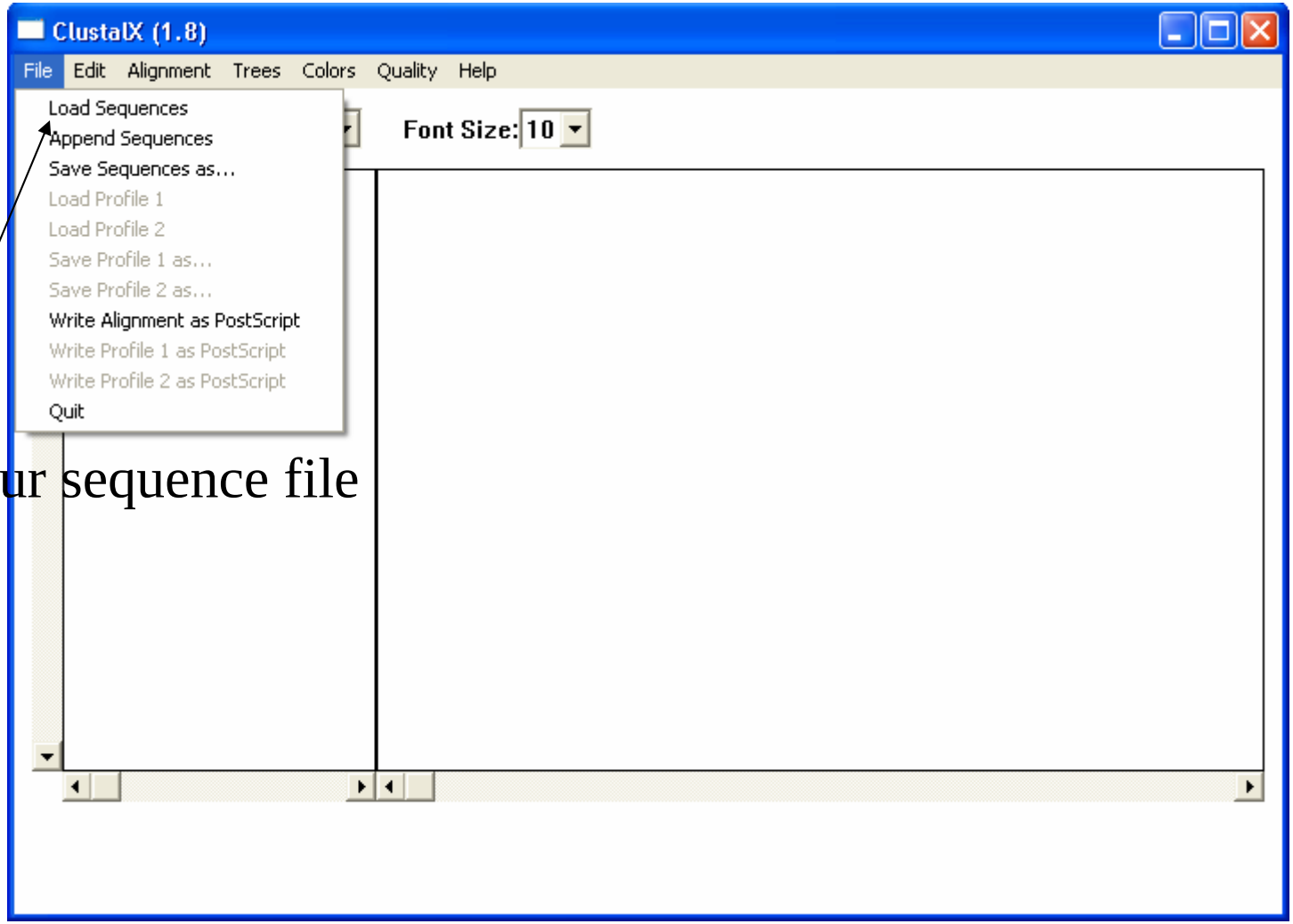
OUTPUT		PHYLOGENETIC TREE		
OUTPUT FORMAT	OUTPUT ORDER	TREE TYPE	CORRECT DIST.	IGNORE GAPS
aln w/numbers	aligned	nj	on	off

Enter or Paste a set of Sequences in any supported format: Help

```
DIAGTTSTLQEQIGWMTSNPPIPVGEIYKRWIILGLNKIVRMYSPV
SILDIRQGPKEPFRDYVDRFYKTLRAEQATQEVKNWMTETLLIQNA
NPDCKTILKALGPAATLEEMMTACQ
>AAD28894
VHQALSPRTLNAWVKVVEEKAFSPEVIPMFSALSEGATPQDLNTML
NTVGGHQAAMQMLKETINEEAAEWDRLHPVHAGPIAPGQMREPRGS
DIAGTTSTLQDQIGWMTNNPPIPVGEIYKRWIILGLNKIVRMYSPT
SILDIKQGPKEPFRDYVDRFYKTLRAEQASQEVKNWMTETLLVQNA
NPDCKTILKALGPAATLEEMMTACQ
```

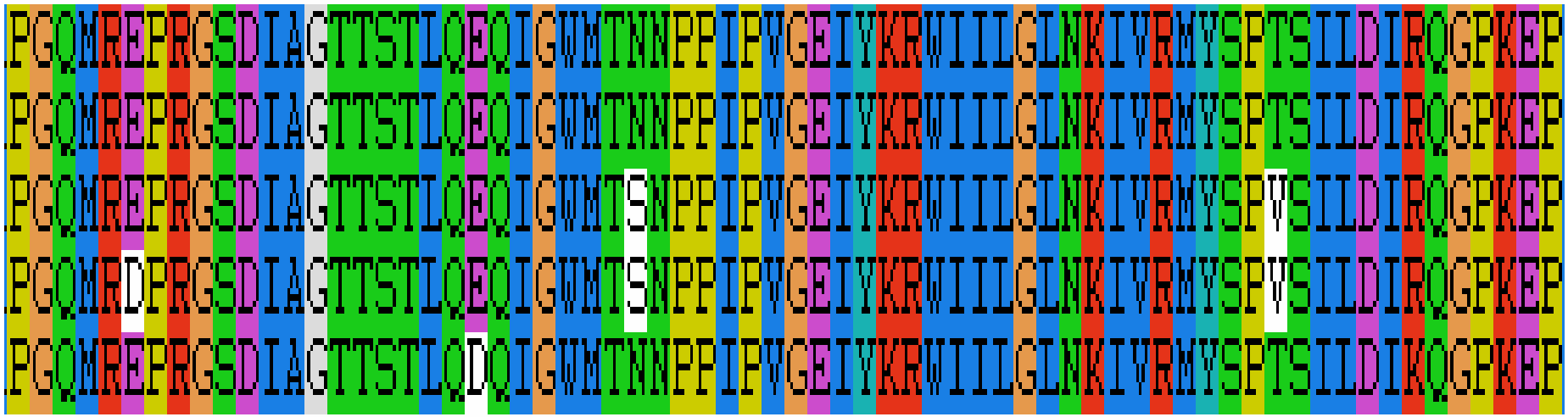
Upload a file: Browse... Run Reset

ClustalX with GUI

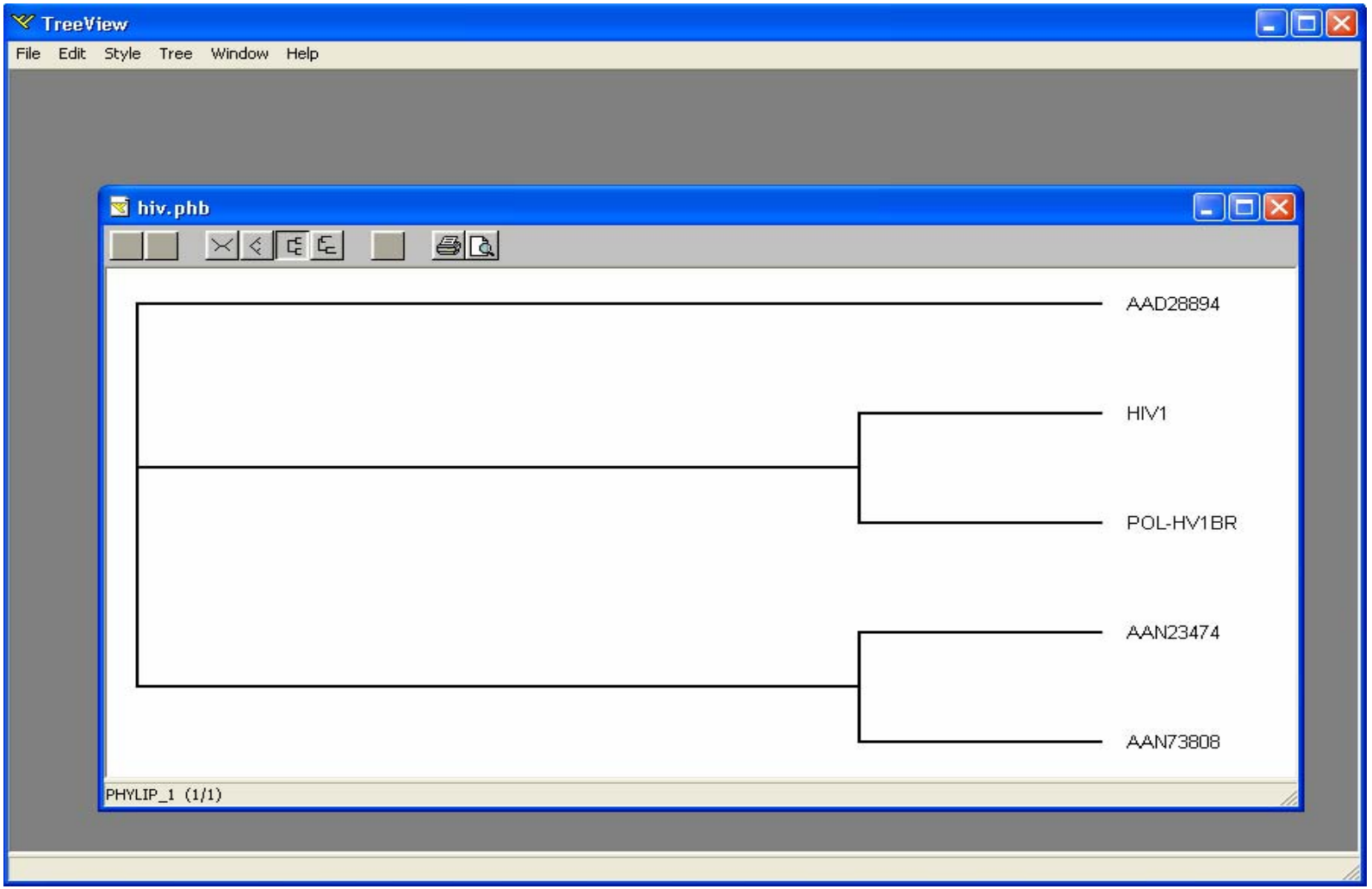


Load your sequence file

http://bioinformatics.unc.edu/software/opensource/index.htm

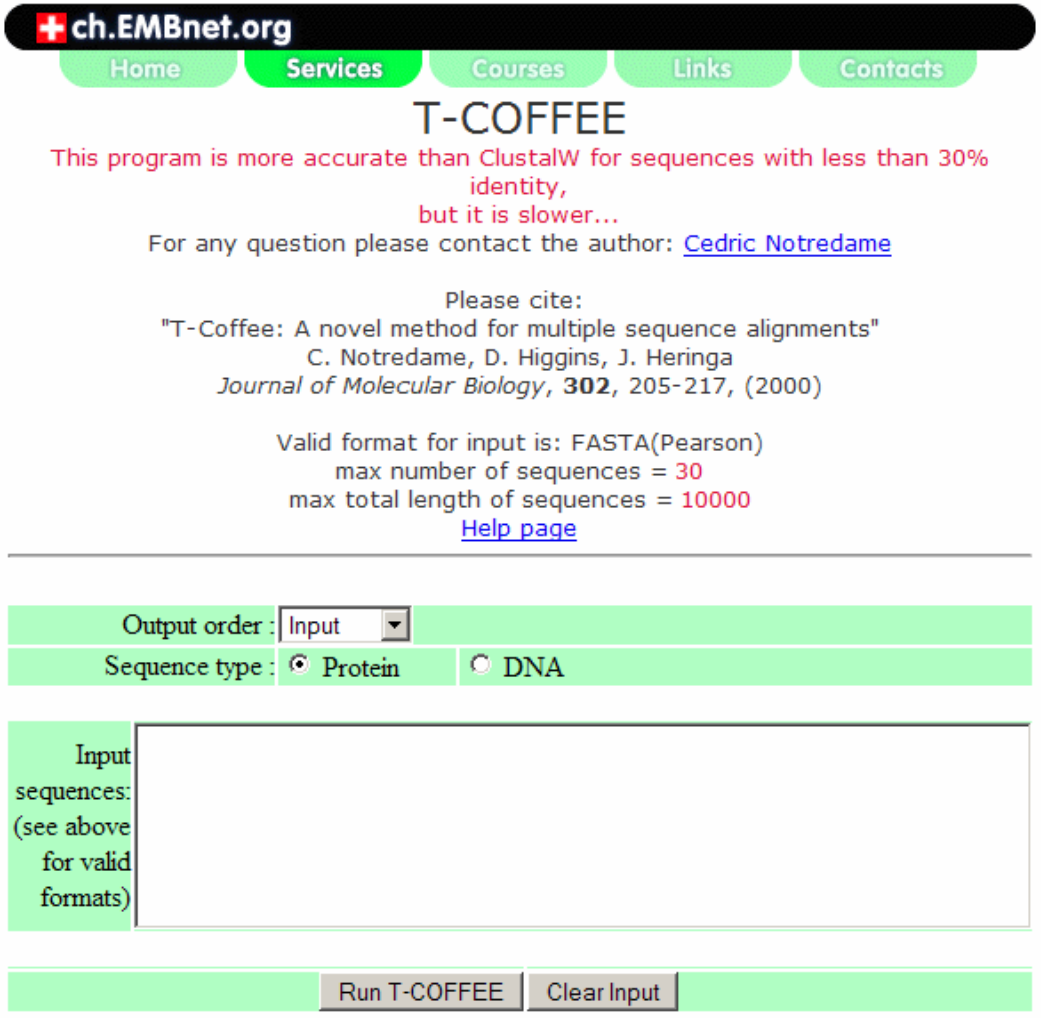


Visualized MSA (conservation and variation)



Distance Based Pylogenetic Tree (visualized by treeview: http://taxonomy.zoology.gla.ac.uk/rod/treeview.html)

T-Coffee



Address: http://www.ch.embnet.org/software/TCoffee.html