

# Graduate Computational Genomics

02-710 / 10-810 & MSCBIO2070

## Multiple Sequence Analysis

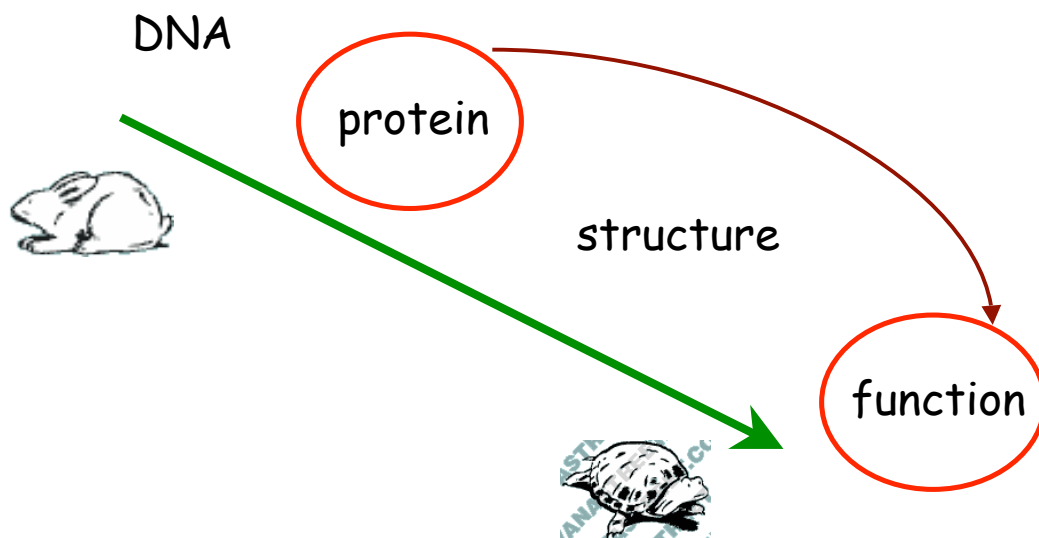
Takis Benos

Lecture #4a, January 17, 2008



Reading: Durbin *et al.* "Biological Sequence Analysis", Ch. 6

## Paces of evolution



# Outline of multiple sequence alignment



- Background
- Scoring of a MSA
- Algorithms
  - Feng-Doolittle
  - Barton-Sternberg
  - CLUSTALW
  - Genetic algorithms
  - Profile HMMs

Benos 02-710/MSCBIO2070 17-JAN-2008

3

## Background (cntd)



### Pfam (curated)

```
CYB_ASCSU  HFNGASLFFIFLYLHFLKGLF...FMSY..RLKK..VWVS
CYB6_MARPO HRWSASMMVLMMILHIFRVYL...TGGFKKPREL..TWVT
CYB_TRYBB  HICFTSLLYLLLYIHIFKSITLIILFDTH..IL...VWFI
*          :*::  ::  :*::

```

Benos 02-710/MSCBIO2070 17-JAN-2008

4



## MSA scoring

---

- Complete probabilistic model:  
Impractical (very complex; not enough data).
- Simplifying assumptions:
  1. Individual columns are statistically independent.
  2. Residues *within* the column can be considered independent (*i.e.*, information on phylogeny is ignored).

$$\text{Score}(\text{alignment}) = \text{Score}(\text{gaps}) + \sum_i \text{Score}(\text{col}_i)$$



## MSA scoring (cntd)

---

- Example:
  - Aligning  $N$  sequences of length  $L$  requires  $(2L)^{N-2}$  pairwise comparisons.
  - You have 15 sequences, 50 a.a. long.
  - Your computer needs 1 sec for each pairwise comparison.
  - How many sequences you'll align until the end of our sun? (*i.e.* approx. 5 billion years)



## MSA scoring (cntd)

### Method-1: minimum entropy

Assuming independence *between* as well as *within* columns

$$P(\text{col}_i) = \prod_{\alpha} p_{i\alpha}^{C_{i\alpha}}$$

Now we define:

$$\text{Score}(\text{col}_i) = -\log P(\text{col}_i) = -\sum_a C_{i\alpha} \cdot \log p_{i\alpha} = -\sum_a C_{i\alpha} \cdot \log \frac{C_{i\alpha}}{N(i)}$$

## MSA scoring (cntd)

### Method-2: sum of pairs (SP)

BLOSUM

$$SP(i) = \text{Score}(\text{col}_i) = \sum_l \sum_{k < l} \text{score}_i(k, l)$$

- There is no probabilistic justification for SP method
- Evolutionary events are overcounted

## MSA scoring: SP

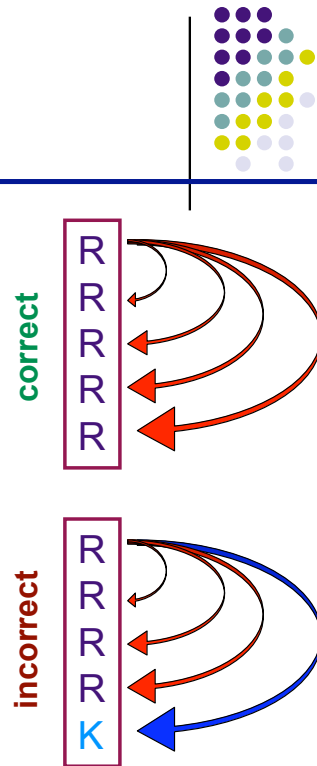
Problem. Compare...

$N$  sequences with *Arg* at position  $i$ .

- BLOSUM62:  $Sc(Arg, Arg) = 5$
- $SP(corr) = 5 \times N(N - 1) / 2$

$N-1$  sequences with *Arg*, one with *Lys*.

- BLOSUM62:  $Sc(Arg, Lys) = 2$
- $SP(incorr) = SP(corr) - 3 \times (N - 1)$
- $(SP(corr) - SP(incorr)) / SP(corr) = 6 / 5N !!$



Benos 02-710/MSCBIO2070 17-JAN-2008

9

## Progressive algorithms

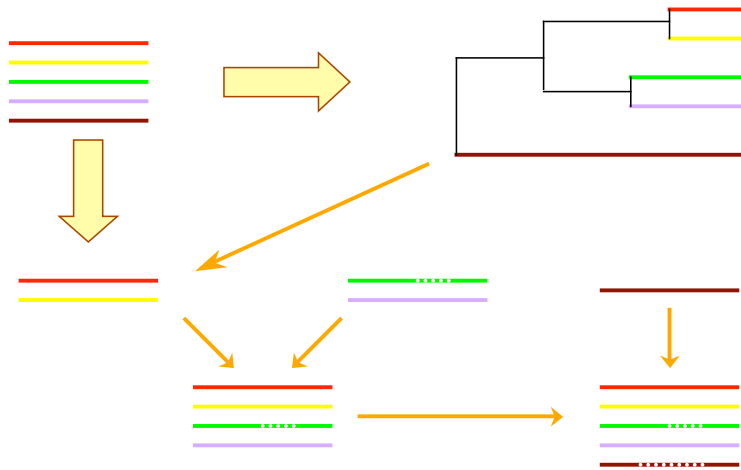
• General idea:

- Build a "guide tree" (quick 'n' dirty) with approximate sequence distances
- Align the two closest sequences according to some scoring scheme; fix their alignment.
- Continue with next sequence and/or alignment, until all sequences are aligned.

Benos 02-710/MSCBIO2070 17-JAN-2008

10

## Progressive algorithms (cntd)



Benos 02-710/MSCBIO2070 17-JAN-2008

11

## MSA methods: *Feng-Doolittle*



[Feng & Doolittle, 1987]:

- Calculate a "distance matrix", using all pairwise scores.
- Construct a *guide tree* from this distance matrix.
- Starting from the first node added to the guide tree, align the child nodes.
- Repeat for other nodes in the order they were added to the tree.
- Each new sequence is added after compared to *every* sequence in the current alignment.
- When an (sub)alignment is added, there is an all-to-all comparison.

Benos 02-710/MSCBIO2070 17-JAN-2008

12

## MSA methods: *Barton-Sternberg*



[Barton & Sternberg, 1987]:

- Find the two sequences with the highest pairwise score; build a profile.
- Find the sequence that is closest to this profile; align it to it.
- Repeat until all sequences have been aligned to a single profile.
- Remove seq-1 and re-align it to the profile; calculate the new score.
- Repeat with seq-2, etc.
- Repeat the procedure a fixed number of times, or until convergence occurs (i.e. score doesn't change).

Benos 02-710/MSCBIO2070 17-JAN-2008

13

## MSA methods: *CLUSTALW*



[Thompson, Higgins & Gibson, 1994]:

- Similar to Feng-Doolittle.
- Builds profiles and aligns the profiles.
- Tree.
  - Build on Kimura's model with NJ algorithm.
  - Can be re-adjusted on the fly
- Heuristics.
  - Sequences are weighted to compensate for biased representation
  - Various scoring matrices, depending on the expected similarity
  - Various gap & mismatch penalties, depending on the position of the alignment

Benos 02-710/MSCBIO2070 17-JAN-2008

14

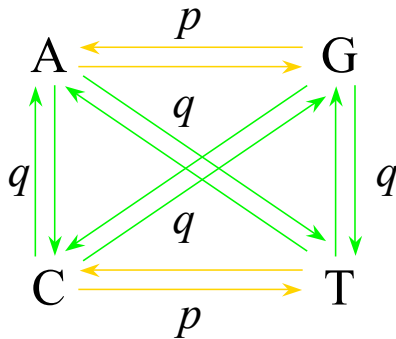


## Nucleic acid distances (cntd)

Kimura's 2-parameter model (1980):

$$D_{K2p} = -\frac{1}{2} \cdot \ln(1 - 2P - 2Q) - \frac{1}{4} \cdot \ln(1 - 2Q)$$

$$\text{Var}(D_{K2p}) = \frac{c_1^2 P + c_3^2 Q - (c_1 P + c_3 Q)^2}{N}$$



$$c_1 = 1/(1 - 2P - Q)$$

$$c_2 = 1/(1 - 2Q)$$

$$c_3 = (c_1 + c_2)/2$$

Source:  
<http://helix.biology.mcmaster.ca/721/distance/distance.html>.

Benos 02-710/MSCBIO2070 17-JAN-2008

15

## Building trees: UPGMA algorithm

Unweighted Pair Group Method with Arithmetic mean (UPGMA):

- Assign each node to its own cluster.
- The distance between two clusters is the average distance between all pairs.
- Join the two closest clusters,  $i$  and  $j$ .
- Add new node at  $D_{ij} = d_{ij}/2$ .
- Recalculate distances.
- Repeat until two clusters remain.

*Assumption: molecular clock*



Benos 02-710/MSCBIO2070 17-JAN-2008

16

# Building trees: Neighbour-joining



## Neighbour-joining (NJ):

- Similar to UPGMA (bottom up clustering)
- No molecular clock assumption.
- The new node is added in distance  $D_{ij} = d_{ij} - r_i - r_j$ .
  - In that way the topology with the overall smallest total length is preferred

*Requirement: additive distance metric*

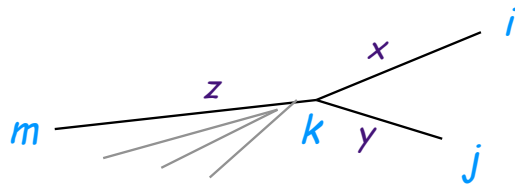
## NJ algorithm



- **Initialization:**
  - Both  $T$  (set of all clusters) and  $L$  (set of leaf nodes) initialized to the total number of sequences.
- **Iteration:**
  1. Pick nodes  $i$  and  $j$  for which  $D_{ij}$  (Kimura for CLUSTALW) is minimal
  2. Define new node  $k$  and set  $d_{km} = 0.5 (d_{im} + d_{jm} - d_{ij})$  for all  $m$
  3. Remove  $i$  and  $j$  from  $L$  and add  $k$  to  $T$ .
  4. Lengths of  $k$ :  $d_{ik} = 0.5 (d_{ij} + r_i - r_j)$  and  $d_{jk} = d_{ij} - d_{ik}$
$$r_i = \frac{1}{|L|-2} \sum_{k \in L} d_{ik} \quad r_j = \frac{1}{|L|-2} \sum_{k \in L} d_{jk}$$
- **Termination:** When leaves are exhausted.



## NJ algorithm (cntd)



Set distance  $km$  (for every  $m$ ):

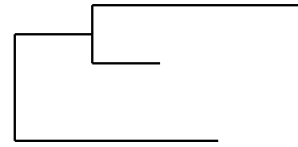
$$d_{km} = 0.5 (d_{im} + d_{jm} - d_{ij})$$

Distance  $ik$ :

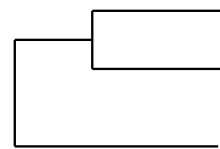
$$d_{ik} = 0.5 (d_{ij} + r_i - r_j)$$

$$r_i = \frac{1}{|L|-2} \sum_{m \in L} d_{im} \quad r_j = \frac{1}{|L|-2} \sum_{m \in L} d_{jm}$$

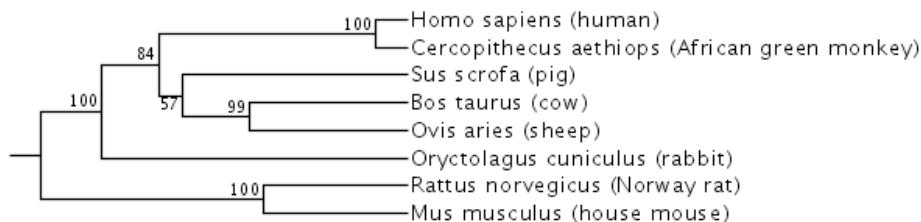
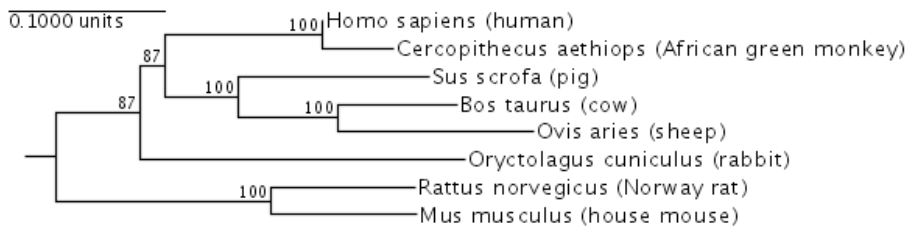
NJ



UPGMA



## NJ vs. UPGMA



Source: <http://www.clcbio.com/>



## Comments

- Unlike pairwise alignments, multiple alignment methods **are not** guaranteed to find the optimal alignments.

Benos 02-710/MSCBIO2070 17-JAN-2008

21

## Multiple alignments: general (cntd)



### Pfam (curated)

```
CYB_ASCSU  HFNGASLFFIFLYLHFLFKGLF...FMSY..RLKK..VWVS
CYB6_MARPO HRWSASMMVLMMILHIFRVYL...TGGFKKPREL..TWVT
CYB_TRYBB  HICFTSLLYLLLYIHIFKSITLIILFDTH..IL...VWFI
*          :*::  ::  :*:*:
          .*.  .*
```

### CLUSTALW (automatic)

```
CYB_ASCSU  HFNGASLFFIFLYLHFLFKGLFFMSYR--LKVWVS
CYB6_MARPO HRWSASMMVLMMILHIFRVYLTGGFKKPREL TWVT
CYB_TRYBB  HICFTSLLYLLLYIHIFKSITLIILFDTHIIVWFI
*          :*::  ::  :*:*:
          .*.  .*
```

Benos 02-710/MSCBIO2070 17-JAN-2008

22



## Comments

---

- Unlike pairwise alignments, multiple alignment methods **are not** guaranteed to find the optimal alignments.
- Multiple alignments are used to calculate profiles characteristic for protein families.
- The profiles can be used to identify new (distant) members of these families.



## Acknowledgements

---

Theory and examples from the following:

- R. Durbin, S. Eddy, A. Krogh, G. Mitchison, "Biological Sequence Analysis", 1998, Cambridge University Press
- Li & Graur "Fundamentals of Molecular Evolution", 1991, Sinauer Associates
- <http://www.sbc.su.se/~per/molbioinfo2001/>