Progressive alignment: Scoring scheme

- Scoring scheme is arguably the most influential component of the progressive algorithm
- Matrix-based algorithms
  - ClustalW, MUSCLE, Kalign
  - Use a substitution matrix to assess the cost of matching two symbols or two profiled columns
  - Once a gap, always a gap
- Consistency-based schemes
  - T-Coffee, Dialign
  - Compile a collection of pairwise global and local alignments (primary library) and to use this collection as a position-specific substitution matrix

Consistency-based approaches

- T-Coffee
  - M-Coffee & 3D-Coffee (Expresso)
- Principle
  - Primary library
  - Library extension

T-Coffee: Primary library

Input sequences

```
SeqA  GARFIELD THE LAST FAT CAT
SeqB  GARFIELD THE FAST CAT
SeqC  GARFIELD THE VERY FAST CAT
SeqD  THE FAT CAT
```

Primary library: collection of global/local pairwise alignments

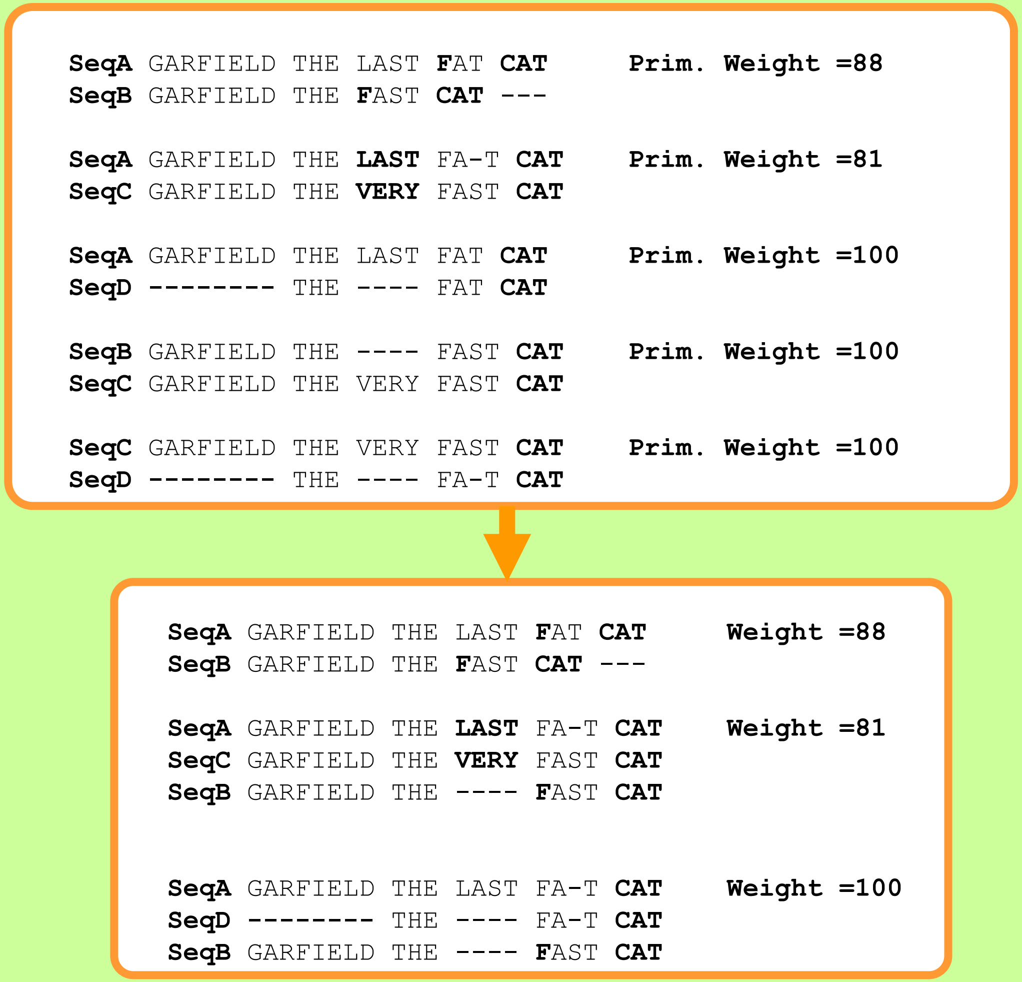
```
SeqA  GARFIELD THE LAST FAT CAT
SeqB  GARFIELD THE FAST CAT

SeqA  GARFIELD THE LAST  FA-T CAT
SeqC  GARFIELD THE VERY FAST CAT

SeqA  GARFIELD THE LAST FAT CAT
SeqD  - - - - - - - - -  THE - - - -  FAT CAT
```

```
SeqB  GARFIELD THE  - - - -  FAST CAT
SeqC  GARFIELD THE VERY FAST CAT

SeqB  GARFIELD THE FAST CAT
SeqD   - - - - - - - - - THE  FA-T CAT

SeqC  GARFIELD THE VERY FAST CAT
SeqD   - - - - - - - - - THE  - - - -  FA-T  CAT
```

T-Coffee and Concistency…

```
SeqA GARFIELD THE LAST FAT CAT     Prim. Weight =88
SeqB GARFIELD THE FAST CAT ---

SeqA GARFIELD THE LAST FA-T CAT    Prim. Weight =81
SeqC GARFIELD THE VERY FAST CAT

SeqA GARFIELD THE LAST FAT CAT     Prim. Weight =100
SeqD -------- THE ---- FAT CAT

SeqB GARFIELD THE ---- FAST CAT    Prim. Weight =100
SeqC GARFIELD THE VERY FAST CAT

SeqC GARFIELD THE VERY FAST CAT    Prim. Weight =100
SeqD -------- THE ---- FA-T CAT
```

T-Coffee and Concistency…



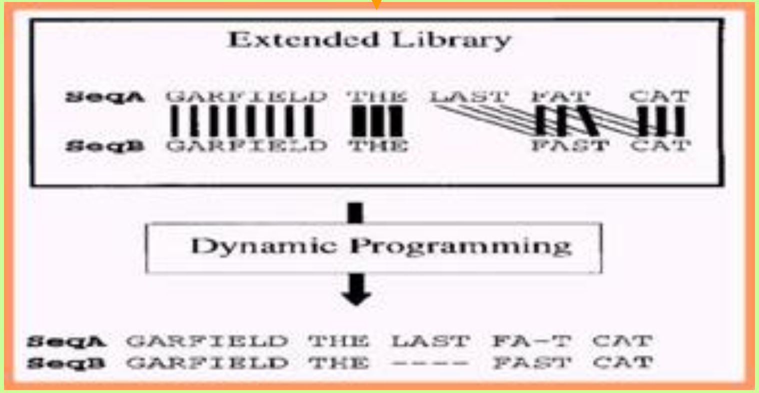
```
SeqA GARFIELD THE LAST FAT CAT     Prim. Weight =88
SeqB GARFIELD THE FAST CAT ---

SeqA GARFIELD THE LAST FA-T CAT    Prim. Weight =81
SeqC GARFIELD THE VERY FAST CAT

SeqA GARFIELD THE LAST FAT CAT     Prim. Weight =100
SeqD -------- THE ---- FAT CAT

SeqB GARFIELD THE ---- FAST CAT    Prim. Weight =100
SeqC GARFIELD THE VERY FAST CAT

SeqC GARFIELD THE VERY FAST CAT    Prim. Weight =100
SeqD -------- THE ---- FA-T CAT
```

```
SeqA GARFIELD THE LAST FAT CAT     Weight =88
SeqB GARFIELD THE FAST CAT ---

SeqA GARFIELD THE LAST FA-T CAT    Weight =81
SeqC GARFIELD THE VERY FAST CAT
SeqB GARFIELD THE ---- FAST CAT


SeqA GARFIELD THE LAST FA-T CAT    Weight =100
SeqD -------- THE ---- FA-T CAT
SeqB GARFIELD THE ---- FAST CAT
```

T-Coffee and Concistency…

Extended library: new pairwise alignment (AB), (AC), (AD), (BC), (BD) and (CD)

```
SeqA GARFIELD THE LAST FAT CAT     Weight =88
SeqB GARFIELD THE FAST CAT ---

SeqA GARFIELD THE LAST FA-T CAT    Weight =81
SeqC GARFIELD THE VERY FAST CAT
SeqB GARFIELD THE ---- FAST CAT


SeqA GARFIELD THE LAST FA-T CAT    Weight =100
SeqD -------- THE ---- FA-T CAT
SeqB GARFIELD THE ---- FAST CAT
```



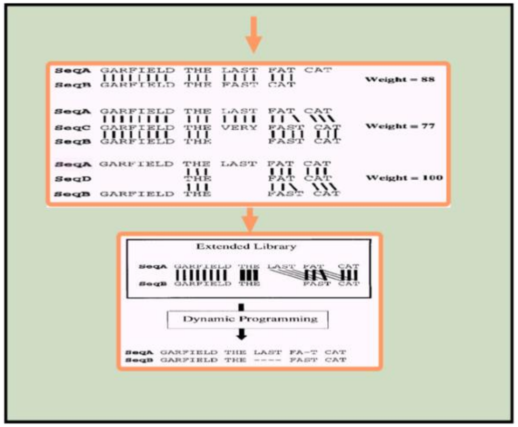
Extended Library

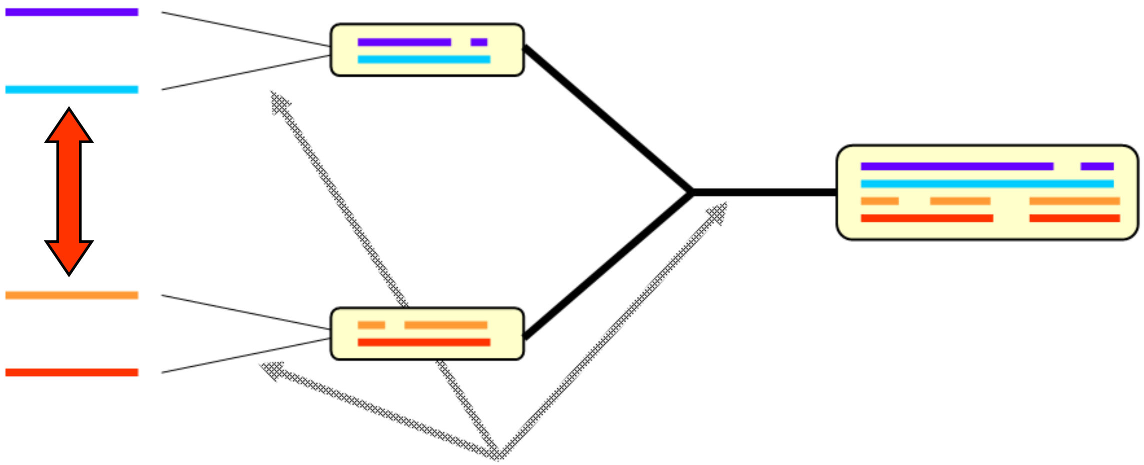
```
SeqA GARFIELD THE LAST FAT  CAT
SeqB GARFIELD THE      FAST CAT
```

Dynamic Programming

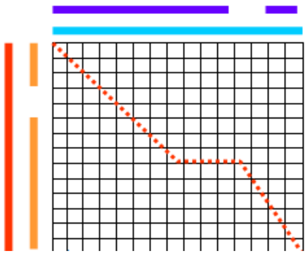
```
SeqA GARFIELD THE LAST FA-T CAT
SeqB GARFIELD THE ---- FAST CAT
```

T-Coffee and Concistency…


SeqA GARFIELD THE LAST FAT CAT
SeqB GARFIELD THE FAST CAT
Weight = 88
SeqA GARFIELD THE LAST FAT CAT
SeqC GARFIELD THE VERY FAST CAT
SeqB GARFIELD THE FAST CAT
Weight = 77
SeqA GARFIELD THE LAST FAT CAT
SeqD THE FAT CAT
SeqB GARFIELD THE FAST CAT
Weight = 100
Extended Library
SeqA GARFIELD THE LAST FAT CAT
SeqB GARFIELD THE FAST CAT
Dynamic Programming
SeqA GARFIELD THE LAST FA-T CAT
SeqB GARFIELD THE ---- FAST CAT



Dynamic Programming Using A Substitution Matrix



T-Coffee uses progressive strategy to derive multiple alignment

- Guide tree
- First align the closest two sequences (DP using the weights derived from the extended library)
- Align two “alignments” (using the weights from the extended library -- average over each column)
- No additional parameters (gaps etc)
  - The substitution values (weights) are derived from extended library which already considered gaps
  - High scoring segments (consistent segments) enhanced by the data set to the point that they are insensitive to the gap penalties

MUSCLE: a tool for fast MSA

- Initial progressive alignment followed by horizontal refinement (stochastic search for a maximum objective score
  - Step 1: draft progressive (using k-mer counting for fast computation of pairwise distance; tree building using UPGMA or NJ)
  - Step 2: Improved progressive to improve the tree and builds a new progressive alignment according to this tree (can be iterated).
  - Step 3: Refinement using tree-dependent restricted partitioning (each edge is deleted from the tree to divide the sequences into two disjoint subsets, from each a profile is built; the profile-profile alignment is computed, and if the score improves, retain the new alignment).
- Ref: MUSCLE: a multiple sequence alignment method with reduced time and space complexity; BMC Bioinformatics 2004, 5:113

Multiple alignment: History

1975 Sankoff
Formulated multiple alignment problem and gave DP solution
1988 Carrillo-Lipman
Branch and Bound approach for MSA
1990 Feng-Doolittle
Progressive alignment
1994 Thompson-Higgins-Gibson-ClustalW
Most popular multiple alignment program
1998 DIALIGN (*Segment-based multiple alignment)*
2000 T-coffee *(consensus-based)*
2004 MUSCLE
2005 ProbCons (uses Bayesian consistency)
2006 M-Coffee (consensus meta-approach)
2006 Expresso (3D-Coffee; use structural template)
2007 PROMALS (profile-profile alignment)

Summary & references

- "A majority of studies indicate that consistency-based methods are more accurate than their matrix-based counterparts, although they typically require an amount of CPU time N times higher than simpler methods (N being the number of sequences)"
- http://tcoffee.vital-it.ch/cgi-bin/Tcoffee/tcoffee_cgi/index.cgi
- Recent evolutions of multiple sequence alignment algorithms. 2007, 3(8):e123
- Issues in bioinformatics benchmarking: the case study of multiple sequence alignment. Nucleic Acids Res. 2010 Jul 1 Chapter 6