T-Coffee: A Novel Method for Fast and Accurate Multiple Sequence Alignment

Beth Crompton

http://www.worth1000.com/entries/247500/247882kPRo_w.png

Multiple Alignment is Hard

- General empirical models fail at <30% sequence identity
- Progressive Alignment is too greedy
- Simultaneous alignment is very resource intensive

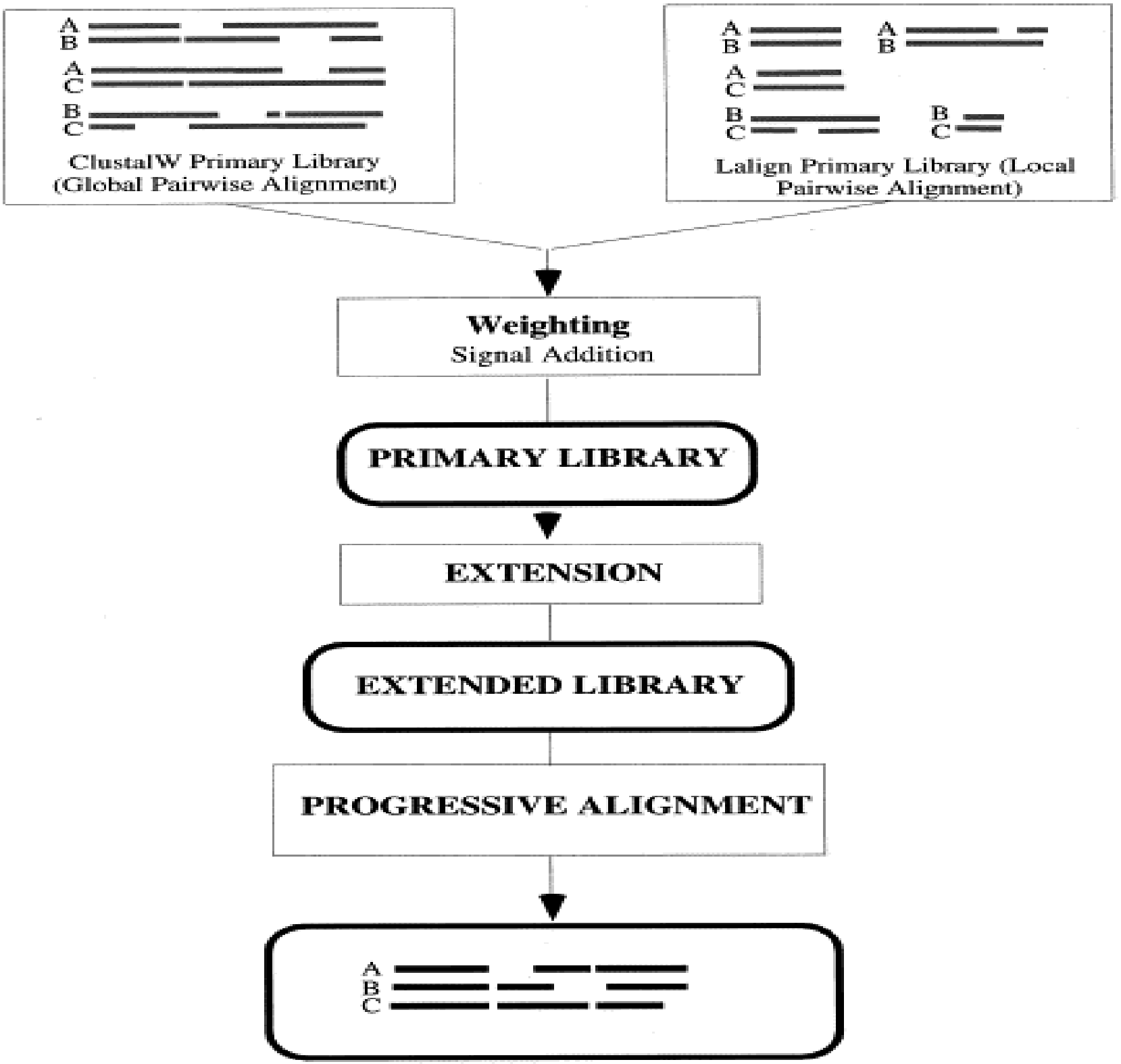
T-Coffee is the solution to all your problems

http://www.saynotocrack.com/wp-content/uploads/2007/07/evil-latte.jpg

Tree-based Consistency Objective Function For AlignmEnt Evaluation

http://farm1.static.flickr.com/120/284971917_d60dcce9a0.jpg?v=0

- Generates multiple alignments from heterogeneous data source in reasonable runtime
- Optimizes the multiple alignments to fit the pairwise alignments


A
B
A
C
B
C
ClustalW Primary Library
(Global Pairwise Alignment)
A
B
A
B
A
C
B
C
B
C
Lalign Primary Library (Local
Pairwise Alignment)
Weighting
Signal Addition
PRIMARY LIBRARY
EXTENSION
EXTENDED LIBRARY
PROGRESSIVE ALIGNMENT
A
B
C

Weighted Primary Library

http://www.emmitsburg.net/humor/pictures/2007/Fw%20Coffee%20art.jpg

Align pairs using ClustalW & Lalign

Each aligned pair is a constraint

Assign weights to each constraint using Sequence Identity method

Weight represents correctness of constraints

Extending the library

Problem: fitting a set of weighted constraints into a multiple alignment is NP-complete

Solution: heuristics!

Weights

Already indicate similarity between two sequences

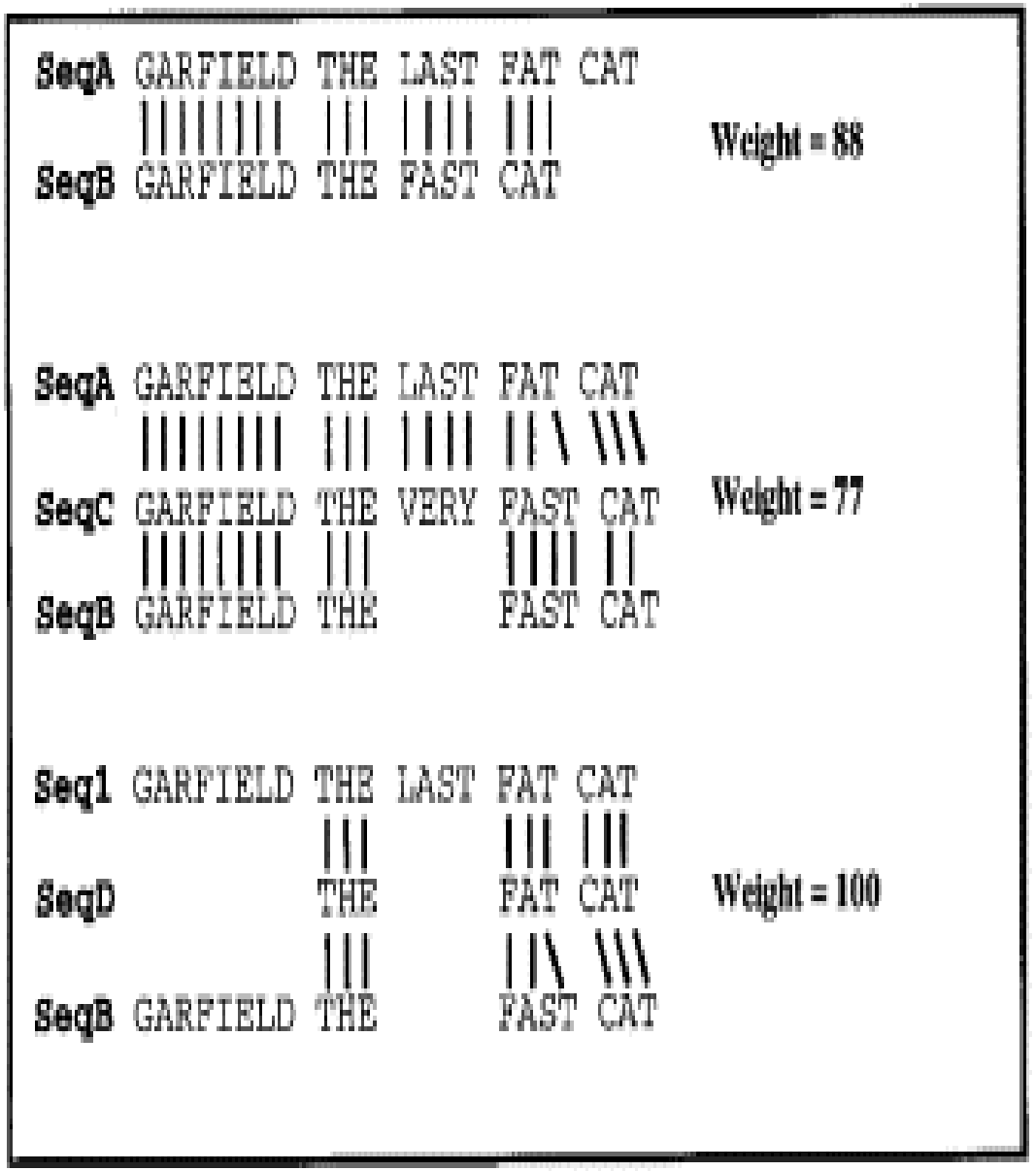
Should indicate consistency with all other sequences involving those residues

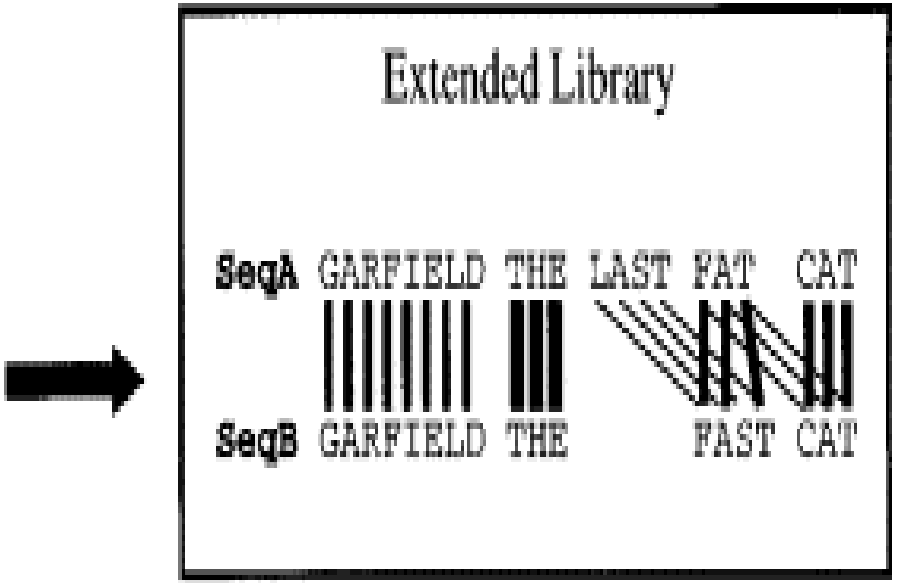
Weight of a pair becomes sum of all weights found examining all triplets involving that pair

Weight of non-existent residues becomes 0

Library extension-continued

c)Extended Library for seq1 and seq2

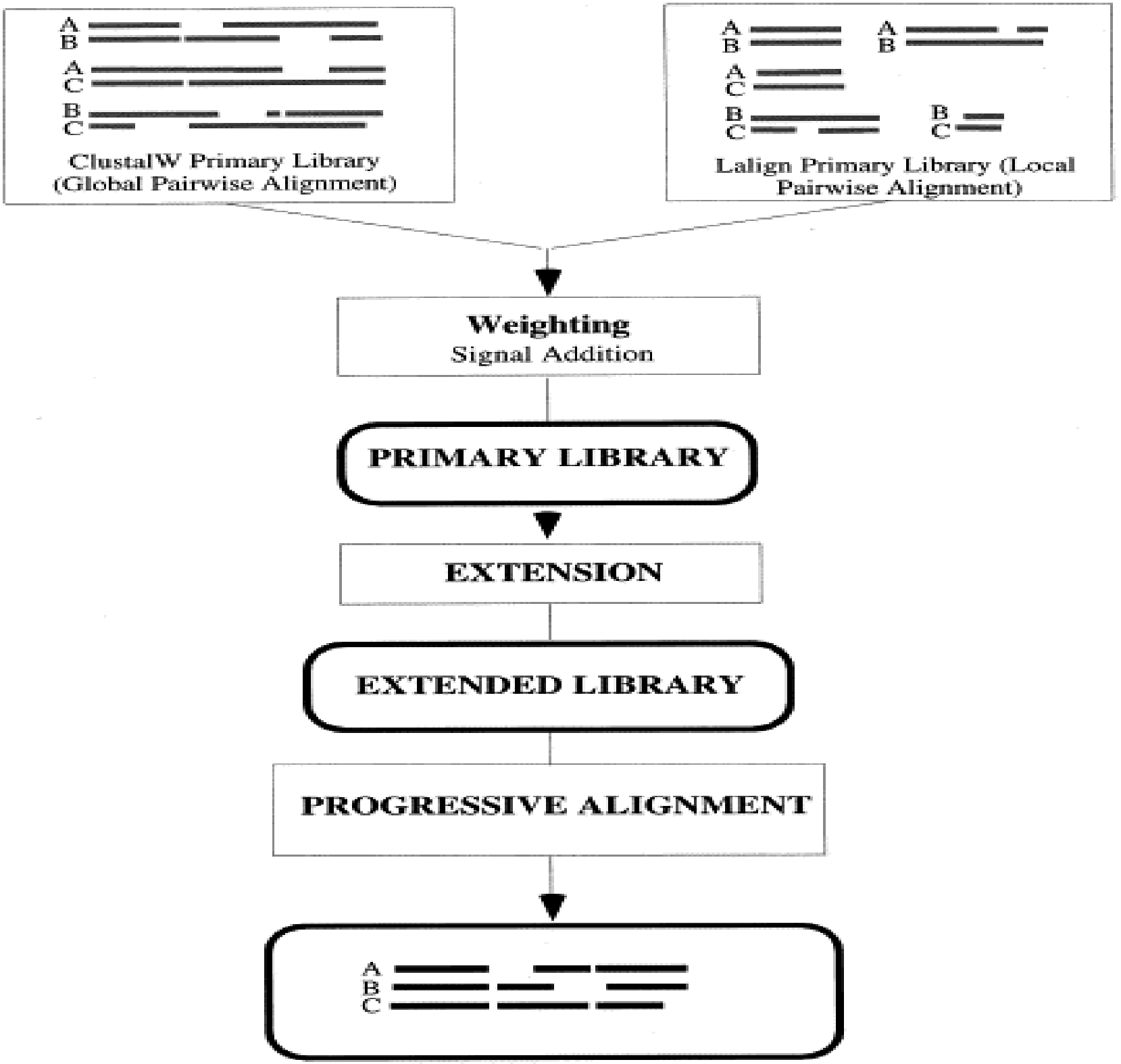




Dynamic Programming

SeqA GARFIELD THE LAST FA-T CAT

SeqB GARFIELD THE ---- FAST CAT


A
B
A
C
B
C
ClustalW Primary Library
(Global Pairwise Alignment)
A
B
A
B
A
C
B
C
B
C
Lalign Primary Library (Local
Pairwise Alignment)
Weighting
Signal Addition
PRIMARY LIBRARY
EXTENSION
EXTENDED LIBRARY
PROGRESSIVE ALIGNMENT
A
B
C

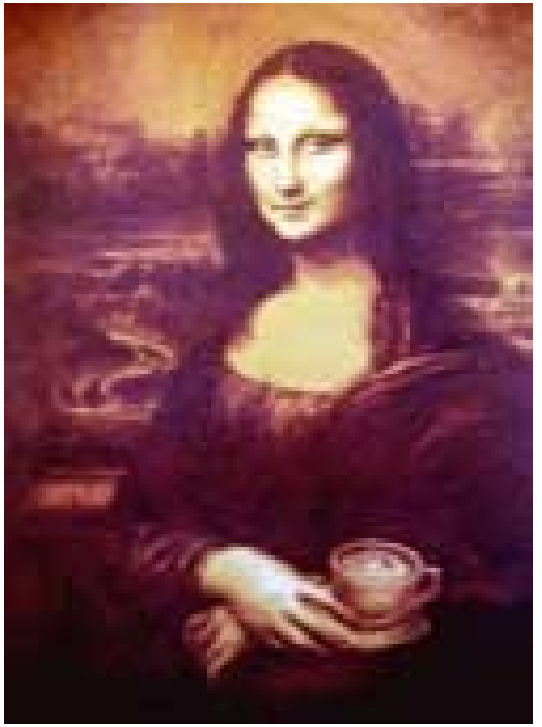
Progressive Alignment

Pair-wise alignments

distance matrices

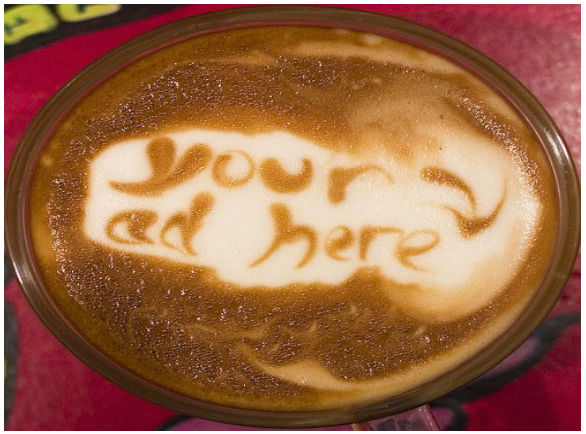
guide tree

multiple alignment



http://www.monalisamania.com/graphics/art/monalatte.jpg

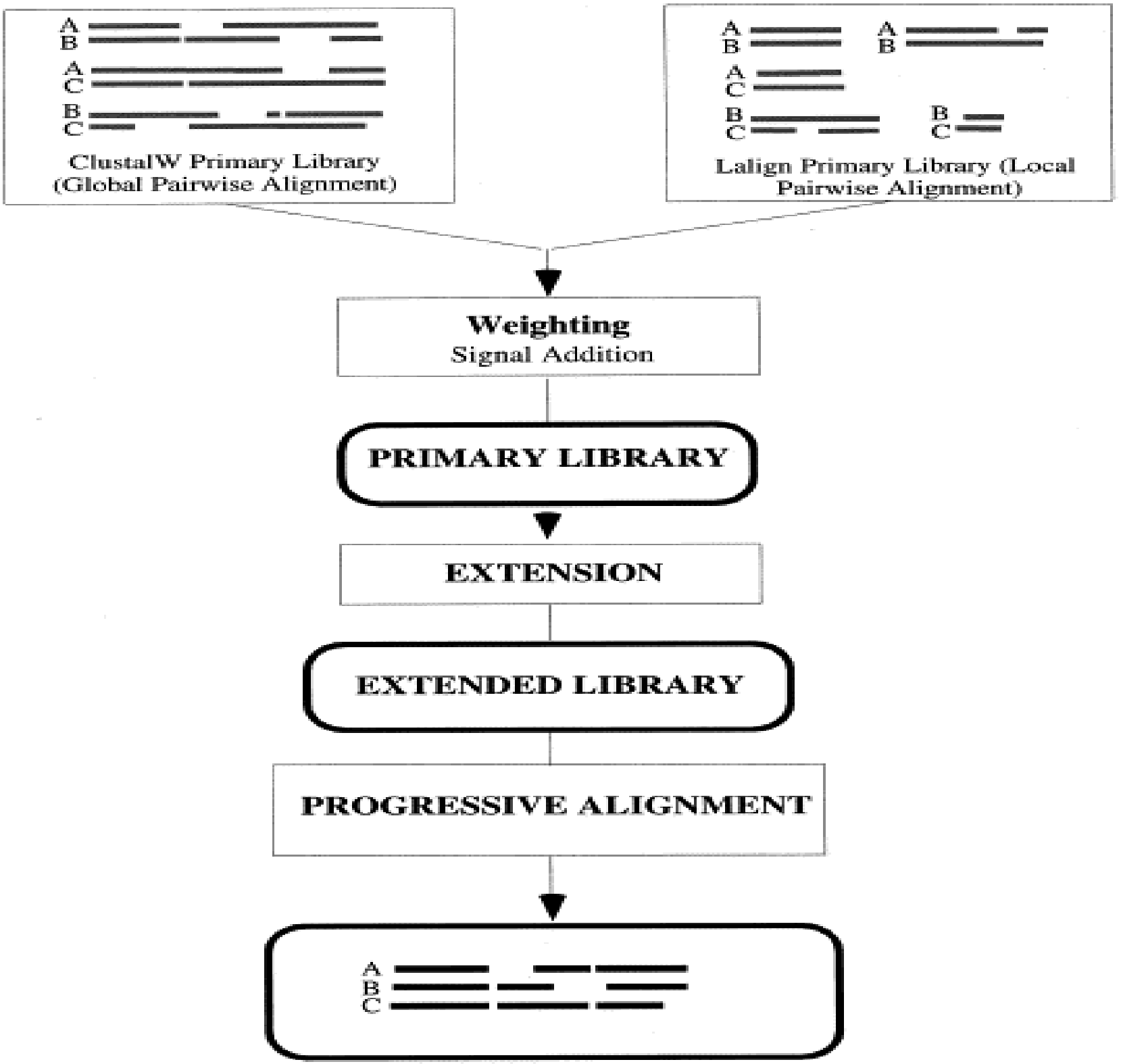
http://www.woostercollective.com/2007/04/01/coffeeart4.jpg



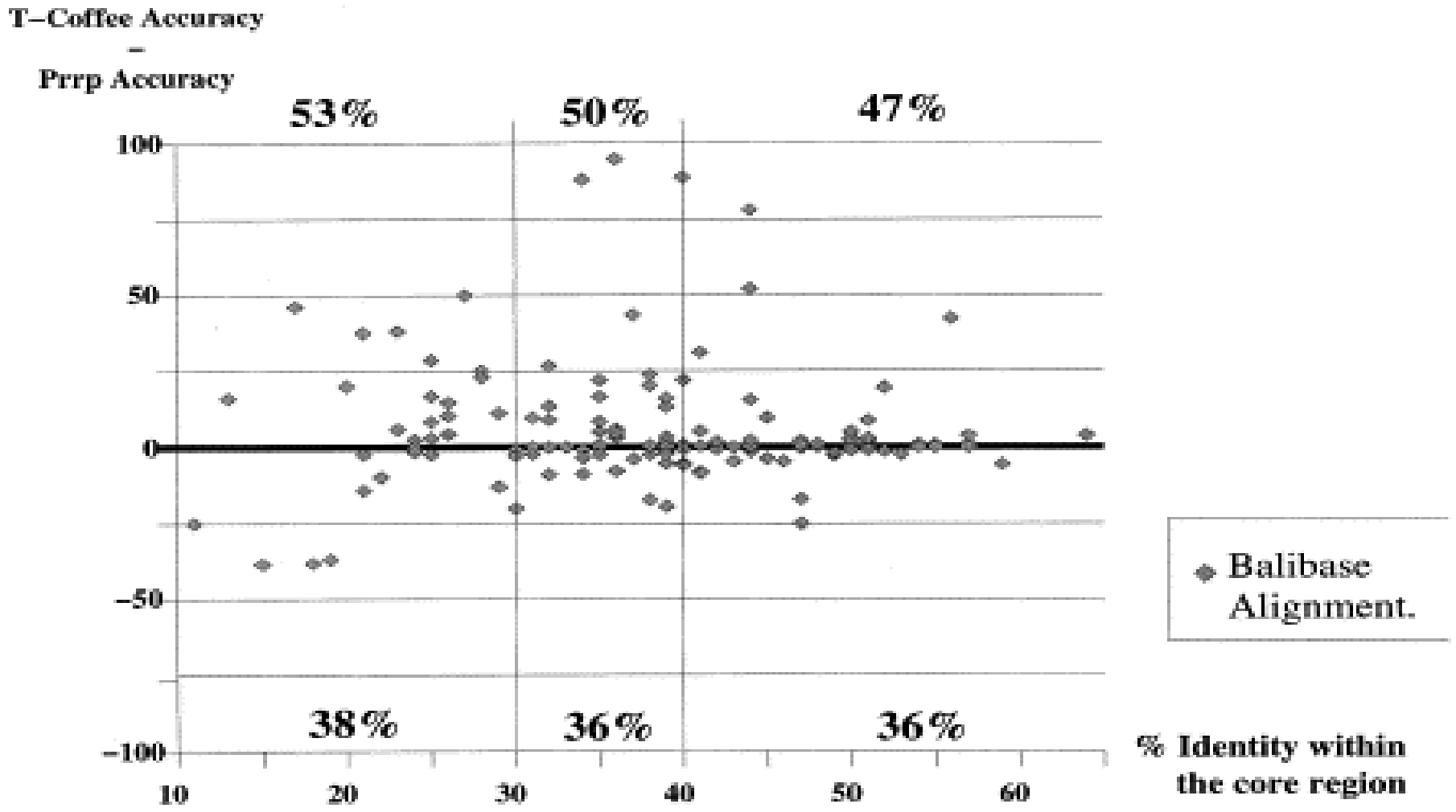
Differences

No need for scoring matrices or gap penalties!

Pairwise alignments are guided toward consistency with multiple alignment anyway.


A
B
A
C
B
C
ClustalW Primary Library
(Global Pairwise Alignment)
A
B
A
B
A
C
B
C
B
C
Lalign Primary Library (Local
Pairwise Alignment)
Weighting
Signal Addition
PRIMARY LIBRARY
EXTENSION
EXTENDED LIBRARY
PROGRESSIVE ALIGNMENT
A
B
C

But is it biologically accurate?



Why T-Coffee is worth the effort

Method	Cat1 (81)	Cat2 (23)	Cat3 (4)	Cat4 (12)	Cat5 (11)	Total1 (141)	Total2 (141)	Significance
Dialign	71.0	25.2	35.1	74.7	80.4	61.5	57.3	11.3^{a}
ClustalW	78.5	32.2	42.5	65.7	74.3	66.4	58.6	26.2^{a}
Prrp	78.6	32.5	50.2	51.1	82.7	66.4	59.0	36.9^{a}
T-Coffee	80.7	37.3	52.9	83.2	88.7	72.1	68.7	

Library extension makes all the difference

Particularly useful for Kinases

```
                                                        NBS
g11a_orysa lshfkllkklgcgdigsvylsels---gtesyfamKVMDKas-------
kp68_human gmdfkeieligsggfgqvfkakhr---idgktyviKRVKYnn-------
gcn2       --tlkrlnfsgqgafgqvvkarna---ldsryyaiKKIRNte-------
st11_yeast pknwlkgacigsgsfgsvylgmna---htgelmavKQVEIknnnigvpt
kin3_yeast rseyqvleeigrgsfgsvrkvihi---ptkkllvrKDIKYgh-------
nima_emeni adkyevlekigcgsfgiirkvkrk---sdgfilcrKEINYik-------
kin1_yeast lgdwefvetvgagsmgkvklakhr---ytnevcavKIVNRat----kaf
kcc1_yeast kkkyvfgktlgagtfgvvrqaknt---etgedvavKILIKka-------
ks62_human psqfellkvlgqgsfgkvflvkkisgsdarqlyamKVLKKat-------
kpc1_yeast ldnfvllkvlgkgnfgkvilsksk---ntdrlcaiKVLKKdn-------
ypk2_yeast iddfdllkvigkgsfgkvmqvrkk---dtgkiyalKALRKay-------
krac_dicdi vadfellnlvgkgsfgkviqvrkk---dtgevyamKVLSKkh-------
kgp2_drome ltdlrviatlgvggfgrvelvqtn--gdssrsfalkqmkksq-------
kapa_mouse ldqfdriktlgtgsfgrvmlvkhk---esgnhyamKILDKqk-------
kdca_drome lenyitravlgngsfgtvmlvrek---sgknyyaaKMMSKed-------
ark1_human mndfsvhriigrggfgevygcrkr---dtgkmyamKCLDKkr-------
dmk_human  rddfeilkvigrgafsevavvkmk---qtgqvyamKIMNKwd-------
dbf2_yeast nrdfemitqvgqggygqvylarkk---dtkevcalKILNKkl-------
pim1_human esqyqvgpllgsggfgsvysgirv---sdnlpvaiKHVEKdr-------
```

Efficiency

Runs in quadratic time
$O(N^2L^2)+O(N^3L) + O(N^3) + O(NL^2)$

2X Slower than ClustalW
(more overhead)