To solve the progressive-alignment problems

- Incorporate more information to reduce early errors:
  - Structural alignment (e.g., Expresso, PROMALS3D, TM-Coffee, PRALINE)
  - Profile alignment (e.g., PRALINE, PSI-Coffee, PROMAPS3D)
- To avoid the greedy-algorithm problem:
  - Iterative refinement to search the global maxima
    → A good objective function is required (e.g., MUSCLE, MAFFT, ProbCons)
- Global (or local) only alignment problem
  - Combine both methods (e.g., T-Coffee)
- More accurate insertion/deletion placement
  - Phylogeny aware gap-placement
    - Phylogeny aware gap-placement
      - PROK, WebPROK:
        - Available in EBI Tools website
          - http://www.ebi.ac.uk/Tools/msa/

More multiple alignment methods

- Incorporate profiles, TM, secondary/3D structures
  - PROK, WebPROK:
    - http://www.ebi.ac.uk/Tools/msa/
- Phylogeny aware gap-placement
  - PROK, WebPROK:
    - http://www.ebi.ac.uk/Tools/msa/
- Hidden Markov model (HMM) or pair-HMM based
  - PROK, WebPROK:
    - Available in EBI Tools website
      - http://www.ebi.ac.uk/Tools/msa/
MUSCLE (Edgar 2004)

http://www.drive5.com/muscle/
http://www.ebi.ac.uk/Tools/msa/muscle/

1. Draft progressive alignment:
   - K-mer distance & UPGMA

2. Improved progressive alignment:
   - Kimura protein distance
   - Tree comparison (branching orders are changed or not)
   - Iteration until the tree stays the same

3. Iterative refinement
   - The tree is partitioned
   - Profiles are obtained from each subtree
   - Profile alignment
   - Iteration based on SP score

MAFFT (Katoh et al. 2005)

http://mafft.cbrc.jp/alignment/software/index.html
http://www.ebi.ac.uk/Tools/msa/mafft/

1. First progressive alignment: FFT-NS-1
   - 6-tuples distance & UPGMA
   - Fast Fourier transform (FFT) is used to detect highly similar segments
   - Segment to segment dynamic programming

2. Improved progressive alignment: FFT-NS-2
   - A better distance matrix from FFT-NS-1 tree

3. Iterative refinement: FFT-NS-i
   - The tree dependent restricted partitioning
   - Group-to-group alignment
   - Iteration based on the weighted SP score

Instead of FFT, full dynamic programming can be used: NW-NS-[12i] [after version 5.0]

MAFFT

http://mafft.cbrc.jp/alignment/software/eval/accuracies.html

See also Katoh et al. (2017) [A new review article]

PRALINE (Simossis, Kleinjung & Heringa 2005)

http://www.ibi.vu.nl/programs/pralinewww/

Table 2. BAIBASE Q scores on subsets

<table>
<thead>
<tr>
<th>Method</th>
<th>Ref1</th>
<th>Ref2</th>
<th>Ref3</th>
<th>Ref4</th>
<th>Ref5</th>
</tr>
</thead>
<tbody>
<tr>
<td>MUSCLE</td>
<td>0.887</td>
<td>0.935</td>
<td>0.823</td>
<td>0.876</td>
<td>0.968</td>
</tr>
<tr>
<td>MUSCLE-p</td>
<td>0.871</td>
<td>0.928</td>
<td>0.813</td>
<td>0.857</td>
<td>0.974</td>
</tr>
<tr>
<td>T-Coffee</td>
<td>0.866</td>
<td>0.934</td>
<td>0.787</td>
<td>0.917</td>
<td>0.957</td>
</tr>
<tr>
<td>NWNSi</td>
<td>0.867</td>
<td>0.923</td>
<td>0.787</td>
<td>0.904</td>
<td>0.963</td>
</tr>
<tr>
<td>CLUSTALW</td>
<td>0.861</td>
<td>0.932</td>
<td>0.751</td>
<td>0.823</td>
<td>0.859</td>
</tr>
<tr>
<td>FFPNSi</td>
<td>0.838</td>
<td>0.908</td>
<td>0.708</td>
<td>0.793</td>
<td>0.975</td>
</tr>
</tbody>
</table>

The average Q score for each method on each BAIBASE subset is shown. Ref1 is the largest subset with 85 test sets, comprising almost 80% of the database. Other subsets are smaller. For example, Ref4 and Ref5 have 12 alignments each, and there are large variances in the individual scores from which the averages are computed. In our opinion, it is not possible to draw meaningful conclusions about the relative performance of different methods on these subsets.
ProbCons (Do et al. 2005)

http://probcons.stanford.edu/

- ProbCons: Probabilistic Consistency-based Multiple Alignment
  1. Initial alignment:
     - Generate posterior probability matrices for each sequence pair based on a hidden Markov model (pair HMM)
  2. Consistency transformation:
     - Update matrices incorporating alignment consistency
  3. Progressive alignment using a guide tree
  4. Iterative refinement
     - Random partitioning and realignments

Clustal Ω http://www.clustal.org/omega/

- Progressive alignment following the guide tree
- Features a fast method for making “guide trees”
  - calculates only distances to n references (mBed method)
  - scalable for very large datasets
- Alignment is done using HAlign (a profile hidden Markov model alignment)
  - highly accurate alignment
- Simple iterative refinement
  - Alignment is converted to HMM
  - Realign input sequences against the HMM

Sievers et al. (2011) Molecular Systems Biology 7: 539
Sievers et al. (2018) Protein Science 27: 135
**PnpProbs**

https://github.com/ytye/PnpProbs

Progressive alignment for "normally" related sequences & non-progressive probabilistic alignment for "distantly" related sequences

**Table 1** Average SP and TC scores on DDBJ

<table>
<thead>
<tr>
<th></th>
<th>20%</th>
<th>25%</th>
<th>30%</th>
<th>35%</th>
<th>40%</th>
<th>45%</th>
<th>50%</th>
<th>55%</th>
<th>60%</th>
<th>65%</th>
<th>70%</th>
<th>75%</th>
<th>80%</th>
<th>85%</th>
<th>90%</th>
<th>95%</th>
<th>100%</th>
</tr>
</thead>
<tbody>
<tr>
<td>PnpProbs</td>
<td>80.61</td>
<td>82.02</td>
<td>84.56</td>
<td>87.09</td>
<td>89.64</td>
<td>92.19</td>
<td>94.74</td>
<td>97.29</td>
<td>99.84</td>
<td>92.39</td>
<td>94.94</td>
<td>97.49</td>
<td>99.04</td>
<td>100.59</td>
<td>102.14</td>
<td>103.69</td>
<td>105.24</td>
</tr>
<tr>
<td>ClustalW</td>
<td>79.21</td>
<td>81.66</td>
<td>84.21</td>
<td>86.76</td>
<td>89.31</td>
<td>91.86</td>
<td>94.41</td>
<td>96.96</td>
<td>99.51</td>
<td>92.06</td>
<td>94.61</td>
<td>97.16</td>
<td>99.71</td>
<td>102.26</td>
<td>104.81</td>
<td>107.36</td>
<td>109.91</td>
</tr>
<tr>
<td>MAFFT</td>
<td>77.86</td>
<td>80.31</td>
<td>82.86</td>
<td>85.41</td>
<td>87.96</td>
<td>90.51</td>
<td>93.06</td>
<td>95.61</td>
<td>98.16</td>
<td>90.71</td>
<td>93.26</td>
<td>95.81</td>
<td>98.36</td>
<td>100.91</td>
<td>103.46</td>
<td>106.01</td>
<td>108.56</td>
</tr>
</tbody>
</table>

**PRANK, WebPRANK**

Mind the gaps: Progress in progressive alignment

D. G. Higgins*, G. Blackshields, and J. W. Holm
Conway Institute, University College Dublin, Dublin, Dublin 4, Ireland

"CLUSTALW attempts to compensate by using an elaborate scoring scheme to encourage gaps to end up on top of each other... results in alignments that are very "block-like"...

"...there may be a price for this prettiness and detachment from phylogenetic reality. CLUSTALW (and other programs) may be guilty of "overalignment", that is where sequences that should not go together are forced into neat-looking blocks. These overaligned regions may be neat looking but misleading."

"There is an understandable tendency for users of multiple alignment software to want their residues neatly aligned in blocks and columns. This is fine when such blocks are biologically accurate as will happen in parts of protein alignments. In cases where insertions or deletions have happened in a less organized manner, as will happen in many noncoding DNA sequences and in less organized parts of protein sequences, such block-like alignments may be biologically meaningless. Perhaps we need to reeducate our eyes to see beauty in what actually happened rather than what looks nice on paper."
BAli-Phy: Statistical coestimation method

http://www.bali-phy.org; Nute et al. (2018, Systematic Biology)

Alignment trimming/filtering

- **Gblocks** Talavera and Castresana (2007), Systematic Biology 56: 564-577
  - Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis
    - http://molevol.cmima.csic.es/castresana/Gblocks.html
    - http://molevol.cmima.csic.es/castresana/Gblocks_server.html
    - http://phylogenome.lirmm.fr/phylo.cgi/one_task.cgi?task_type=gblocks
    - (included in Phylogeny.fr)

- **trimAl** Capella-Gutierrez et al. (2009), Bioinformatics 25: 1972-1973
  - A tool for automated alignment trimming in large-scale phylogenetic analysis
    - http://trimal.cgenomics.org/trimal
    - http://phylemon2.bioinfo.cipf.es (included in Phylemon2)

See also TCS paper by Chang et al. (2014), Mol. Biol. Evol. 31: 1625-1637.

GUIDANCE2: Guide-tree based alignment confidence

http://guidance.tau.ac.il/index.html; Penn et al. (2010); Iba et al. (2015)

Alignment trimming/filtering: Gblocks

- **Gblocks** Talavera and Castresana (2007), Systematic Biology 56: 564-577
  - Used by Gblocks (relaxed)
  - Used by Gblocks (stringent)

Reflects the robustness of an alignment to perturbations introduced by uncertain (bootstrapped) guide trees, varied gap open penalties, and co-optimal alignments.
GUIDANCE2: Guide-tree based alignment confidence

http://guidance.tau.ac.il/index.html

Many alternative MSAs are better than base MSA
Not a single improved MSA; But can be used to improve phylogeny

SuperMSA: base MSA concatenated with alternative MSA

Multiple alignment quality

<table>
<thead>
<tr>
<th>Program</th>
<th>SP score</th>
<th>TC score</th>
</tr>
</thead>
<tbody>
<tr>
<td>bali_score</td>
<td></td>
<td></td>
</tr>
<tr>
<td>qscore</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Veralign</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T-coffee</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MUMSA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GUIDANCE2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AlignStat</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sequence logo</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SuiteMSA</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

BIOS477/877 L16 - 32

Benchmark alignment database: BAiLBASE

http://www.lbgi.fr/bailbase/

Thompson et al. (1999) Nucleic Acid Res. 27: 2682-90

9 reference alignment sets

- Reference 1: equidistant sequences with various levels of conservation
- Reference 2: families aligned with a highly divergent "orphan" sequence
- Reference 3: subfamilies with <25% residue identity between groups
- Reference 4: sequences with N/C-terminal extension
- Reference 5: internal insertions
- Reference 6: internal repeat patterns, inversions, transmembrane regions, etc.
- Reference 7: linear motifs
- Reference 8: mixed

BIOS477/877 L16 - 33

Multiple alignment quality

- Sum of pairs score (SPS) and column score (CS):
  - SPS=Proportion of correctly aligned AA pairs
  - CS=Proportion of correctly aligned columns

BIOS477/877 L16 - 34

Other benchmark alignment database

- HOMSTRAD: Homologous Structure Alignment Database
  http://migueschlab.org/homstrad/
- PREFAB: Protein Reference Alignment Benchmark
  http://www.drive5.com/muscle/prefab.html (MUSCLE website)
- SABmark: Sequence and Structure Alignment Benchmark
  http://bioinformatics.vub.ac.be/databases/databases.html

BIOS477/877 L16 - 35

BIOS477/877 L16 - 36
Sequence similarities and search methods

- **Midnight Zone**: Sequence comparisons fail to detect any structural similarities.
- **Twilight Zone**: More sensitive methods are required (patterns, profiles, profile HMMs).
- **Automatic alignment methods**: Robust (e.g., BLAST).
- **~10% or lower**: Sequence comparisons fail to detect any structural similarities.
- **~20 ~ 30%**: More sensitive methods are required.
- **~40% or higher**: Sequence similarities and search methods.

Alignments → patterns → functions

- **Query sequence (protein or nucleotide)**
- **Search result**: High similarity
- **Multiple alignment**: Low similarity

**How to identify conserved patterns**

- **Conserved pattern**
- **Higher functional constraint**: Functionally important
- **Majority-Rule Consensus Sequence**
- **Pattern**
  - Conserved pattern including only identical sites
  - Conserved pattern with more flexibility
  - Regular expression
- **Consensus Sequence**
  - Residues probably not important for functions are included

---

** Alignments → patterns → functions **

- **Sequence alignment**: BLAST (pairwise local alignment)
- **Search result**: High similarity
- **Multiple alignment**: Low similarity

**How to identify conserved patterns**

- **Conserved pattern**
- **Higher functional constraint**: Functionally important
- **Majority-Rule Consensus Sequence**
- **Pattern**
  - Conserved pattern including only identical sites
  - Conserved pattern with more flexibility
  - Regular expression
- **Consensus Sequence**
  - Residues probably not important for functions are included

---

**How to identify conserved patterns**

- **Conserved pattern**
- **Higher functional constraint**: Functionally important
- **Majority-Rule Consensus Sequence**
- **Pattern**
  - Conserved pattern including only identical sites
  - Conserved pattern with more flexibility
  - Regular expression
- **Consensus Sequence**
  - Residues probably not important for functions are included
PROSITE Pattern Database

Database of protein domains, families and functional sites

- consists of biologically significant sites, patterns, and profiles
  http://prosite.expasy.org/

• PROSITE pattern syntax is described in:
  http://prosite.expasy.org/prosuser.html#meth1

PROSITE: PS00237

[GSTALIVMFYWC] [GSTANCPDE] {EDPKRH} [LIVMNQGA] x(2) [LIVMFY]

PRINTS Database

- PRINTS: a compendium of protein fingerprints
- A fingerprint is a group of conserved motifs used to characterize a protein family
  http://www.bioinf.manchester.ac.uk/dbbrowser/PRINTS/index.php
  (Last updated in 2012)

PRINTS entry: 5HT1ARCEPTR

5HT1ARCEPTR1
GGQNNNTAGDEPFGDG
GGQNNNTAGDEPFGDG
GGQNNNTAPAPFPGDG

5HT1ARCEPTR2
GNNVTSISKVLQVSQVS
GNNVTSISKVLQVSQVS
GNNVTSISKVLQVSQVS

5HT1ARCEPTR3
RTPEDRSDPDACTISK
RTPEDRSDPDACTISK
RTPEDRSDPDACTISK

5HT1ARCEPTR4
FRIRKTVKKVEKKGAGTSLG
FRIRKTVKKVEKKGAGTSLG
FRIRKTVKKVEKKGAGTSLG

5HT1ARCEPTR5
WRRCAENRAVGTPCTNG
WRRCAENRAVGTPCTNG
WRLGVESKAGGALCANG

5HT1ARCEPTR6
AVRQGDDEATLEVIEVHRVG
AVRQGDDEATLEVIEVHRVG
AVRQGDDEATLEVIEVHRVG

5HT1ARCEPTR7
APACLERKNERNAEAK
APACLERKNERNAEAK
APASFERKNERNAEAK

7 motifs

Conserved R

Only short regions can be represented in regular expression patterns