

配列アラインメントプログラム MAFFT と その構造への拡張 — MAFFTash

Kazutaka Katoh (CBRC, AIST)

Hiroyuki Toh (CBRC, AIST)

Outline

- About MAFFT
- Structure-Sequence alignment - MAFFTash
 - ASH part
 - MAFFT part
- Benchmark without reference alignment
- Discussion

Multiple **sequence** alignment methods

ClustalW (1994, 2007)



A lot of improvements in
accuracy, speed and scalability

PRRN, PRIME (1996, 2006-)

T-Coffee (2000-)

MAFFT (2002-)

Kalign (2005-)

ProbCons (2005-)

MUMMALS, PROMALS (2006-)

PicXAA (2010-)

MAFFT - multiple **sequence** alignment program

"Among the nine programs tested, the iterative approach available in **Mafft** (L-INS-i) and **ProbCons** were **consistently the most accurate**, with **Mafft** being the faster of the two." Nuin et al. (BMC Bioinformatics, 2006)

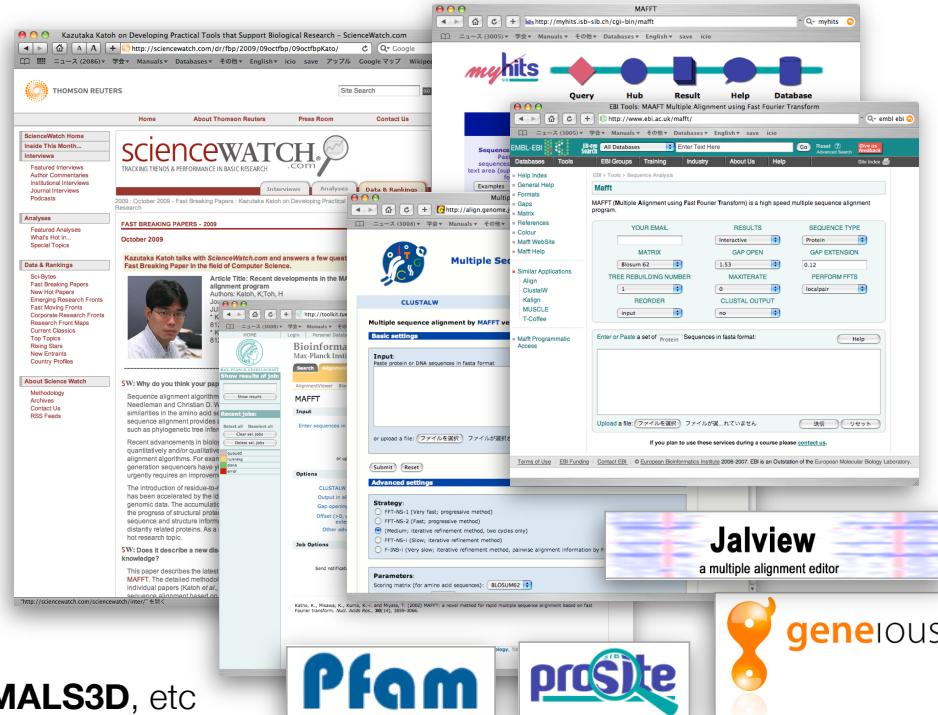
"The **Mafft** strategy L-INS-i **outperforms the other methods**, although the difference between the **Probcons**, **TCoffee** and **Muscle** is mostly insignificant." Ahola et al. (BMC Bioinformatics, 2006)

"**MAFFT version 5** [25] with the option 'G-INS-i' **ranks first** throughout all test-sets." Wilm et al. (Algorithms for Molecular Biology, 2006)

"We suggest **MAFFT** as a good candidate program for initial alignments, particularly where little is known about the data." Golubchik et al. (Molecular Biology and Evolution, 2007)

"Multiple sequence alignment software, such as **MAFFT** [28] and **Muscle** [29] are fast and accurate to the point where **manual curation of alignments is rarely employed** for Pfam seed alignments." Sammut et al. (Briefings in Bioinformatics, 2008)

Our paper (Katoh & Toh 2008) on MAFFT6 was named as a **Fast Breaking Paper** (Oct., 2009) and a **New Hot Paper** (Nov., 2009) in Thomson-Reuters' ScienceWatch.



Used in: **SATCHMO-JS**, **SATé**, **PROMALS3D**, etc

Extensions of MAFFT

- **Protein structure-sequence alignment**

- RNA alignment
- Large-scale alignment
- Fast classification of sequences
- Parallel processing
- GUI
- User-friendly web service
- etc..

Outline

- About MAFFT
- Structure-Sequence alignment - MAFFTash
 - ASH part
 - MAFFT part
- Benchmark without reference alignment
- Discussion

Structure information is greatly helpful

Sequence alignment is:

Not applicable to distantly related proteins

Structure alignment is:

Applicable to more distantly related proteins

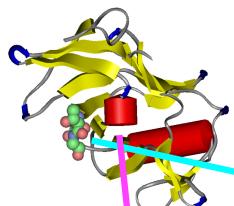
Because:

Generally, protein structures are evolutionarily
more conserved than sequences

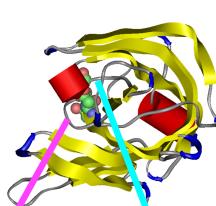
Example

```
>> 0: 2aprB.pdb  
>> 1: 1hih.pdb  
>> 2: 2apra.pdb
```

2apr C-terminal domain



2apr N-terminal domain



Active sites

1hih HIV protease

Conserved β strand

YDSTKFKGSLLTVPI--DNSRGWWGITVDRATVGTSTVASSFDG**IILDGT**TLLILPN---
P-----QIT----LW-QRPLVT---IKIGG--Q--LKEA**LLDTGA**DDTV--LE--
AG----VGTVPMTDYGNDIEYYGQ---VTIGTPGK--KFNL**DFDTGS**SDLWIASTLC
GGG EEE -- EEEEEEEEEE LEEEEEE EEE ---
----- - EE - EEE EEEEEEE E-E --
----- EEE EEE EEEEE- EEE E - EEEEEEE EEEE

----- N-T - AASVARAYGASDNG - DGTYTISCDTSAFK ----- P
----- EM - S - LPGRWKPKMIGGIG-GFIKVRQYDQ -
TNCGSGQTKYDPNQSSTY - QADGR-TWSISYGDGSSASGILAKDN - VNLGGLLIK
----- H-H - HHHHHHH - EEE -
----- EEEEEEEEEE - EEEEEEE EEE -
----- E-EEEEEE EEEEEEEEEE EEE EEE

LVFSINGASFQVSPDSLVFEEFQG - QCIAGFGYGN --- WGF-**AIIIG** -
ILIEICGHKA - IGTIVLV-GPT - PVN-**IIG** -
G - QTIELAKRE-AASFASGPNDG**LLG**LGFDTITTVRGV
EEEEEE EEEEEEE HHHH EEE - EEE - EEE -
EEEEEE EEE - EEEEEE - EEE - EE -
----- EEEEEEEEEE EHHHH EEEE GGG

----- DTFL-KN - N - YVVFN - QGV-PEVQIAPVAE
----- RNLLTQIGC - TLNF -
KTPMDNLISQGLISRPIFGVYLKGAKNGGGGEYIFGG -
----- HHG-GE - E - EEEE - EEEE
----- HHHH - EEEE EEEE EEEE -
HHHHHHH EEEE EEEE EEEE -

Integration of sequence and structure information

However, available 3D structural information is sparse
in comparison with sequence information



We need alignment methods that integrate 3D structural
information and amino acid sequences.



1. More reliable prediction of functional residues in a protein
2. Evolutionary analyses of distantly related proteins
3. Improvement in accuracy of homology modeling
4. Improvement in sensitivity and specificity of database search using profiles

Existing tools

- 3D-COFFEE (O'Sullivan *et al.* 2004)
 - Sequence alignment: T-COFFEE
 - Structural alignment: SAP, LSQman
 - Threading: Fugue
- PROMALS3D (Pei *et al.* 2008)
 - Sequence alignment: MUMMALS, MAFFT
 - Structural alignment: FAST, TMalign
 - Secondary structure prediction: PSIPRED

Collaborators

- Daron M. Standley (Osaka Univ.)
- Huy Dinh (Osaka Univ.)
- Haruki Nakamura (Osaka Univ.)
- Kentaro Tomii (CBRC, AIST)