MAFFT: Multiple Sequence Alignment using Fast Fourier Transform

Intro

2-Step Procedure

- Homology Identification using FFT
- Alignment Scoring/Selection

Faster computation due to…

- Approx. O(NlogN) homology detection
- Simpler-to-compute scoring function

Defining the Signal

For Amino Acid Sequences

2-dimensional signal [volume, polarity]

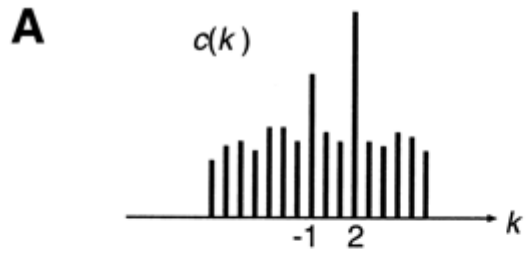
For Nucleotide Sequences

4-dimensional signal [A,T,G,C frequencies]

DFT

Transformation of signal from time to frequency space

$$X(k)=\sum_{n=0}^{N-1} x(n)\, e^{\frac{-i2\pi nk}{N}}, k=0 \ldots N-1$$



FFT

DFT in regular form takes O(N^2)

$$X(k)=\sum_{n=0}^{N-1} x(n)\, e^{\frac{-i2\pi nk}{N}}, k=0 \ldots N-1$$

FFTs compute same values in O(NlogN)

FFT: Cooley-Tukey

Recursive division of sequence into 2 sections

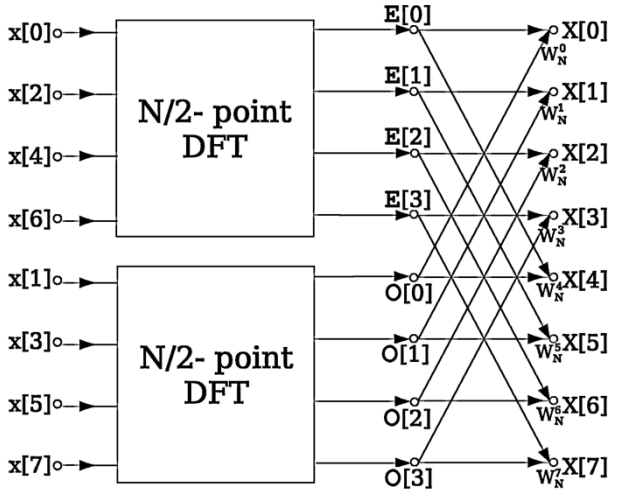
$$X_k = \sum_{m=0}^{N/2-1} x_{2m} e^{-\frac{2\pi i}{N}(2m)k} + \sum_{m=0}^{N/2-1} x_{2m+1} e^{-\frac{2\pi i}{N}(2m+1)k}$$

$$X_k = \underbrace{\sum_{m=0}^{N/2-1} x_{2m} e^{-\frac{2\pi i}{N/2}mk}}_{\text{DFT of even-indexed part of } x_m} + e^{-\frac{2\pi i}{N}k} \underbrace{\sum_{m=0}^{N/2-1} x_{2m+1} e^{-\frac{2\pi i}{N/2}mk}}_{\text{DFT of odd-indexed part of } x_m} = E_k + e^{-\frac{2\pi i}{N}k} O_k.$$

Cooley-Tukey

DFT displays periodicity

$$E_{k+\frac{N}{2}} = E_k$$
$$O_{k+\frac{N}{2}} = O_k$$



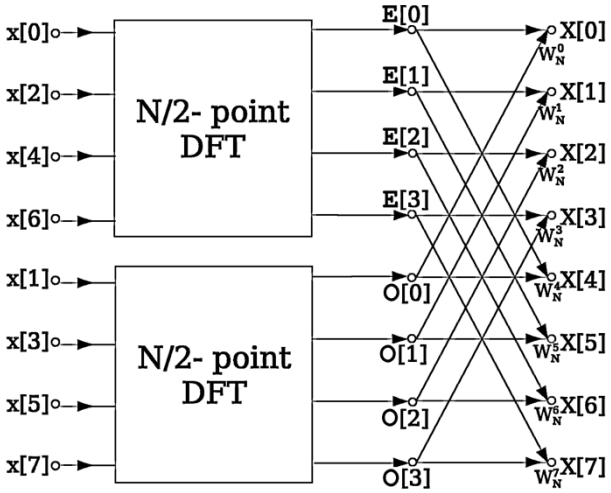
$$X_k = \begin{cases} E_k + e^{-\frac{2\pi i}{N}k} O_k & \text{for } 0 \le k < N/2 \\ E_{k-N/2} + e^{-\frac{2\pi i}{N}k} O_{k-N/2} & \text{for } N/2 \le k < N. \end{cases}$$

$$\begin{aligned} e^{\frac{-2\pi i}{N}(k+N/2)} &= e^{\frac{-2\pi ik}{N} - \pi i} \\ &= e^{-\pi i} e^{\frac{-2\pi ik}{N}} \\ &= -e^{\frac{-2\pi ik}{N}} \end{aligned}$$

Cooley-Tukey

$$\begin{aligned} X_k &= E_k + e^{-\frac{2\pi i}{N}k} O_k \\ X_{k+\frac{N}{2}} &= E_k - e^{-\frac{2\pi i}{N}k} O_k \end{aligned}$$

$X_{0,\ldots,N-1} \leftarrow$ **ditfft2**(x, N, s):
    if N = 1 then
        $X_0 \leftarrow x_0$
    else
        $X_{0,\ldots,N/2-1} \leftarrow$ **ditfft2**(x, $N/2$, $2s$)
        $X_{N/2,\ldots,N-1} \leftarrow$ **ditfft2**(x+s, $N/2$, $2s$)
        for k = 0 to $N/2-1$
            t $\leftarrow X_k$
            $X_k \leftarrow$ t + exp($-2\pi i\ k/N$) $X_{k+N/2}$
            $X_{k+N/2} \leftarrow$ t − exp($-2\pi i\ k/N$) $X_{k+N/2}$
        endfor
    endif



MAFFT FFT Usage

Signal value in “frequency” domain is correlation at offset

“Frequency” is the sequence offset

$$c_v(k) = \sum_{1 \le n \le N, 1 \le n+k \le M} \hat{v}_1(n)\hat{v}_2(n+k),$$

$$c_p(k) = \sum_{1 \le n \le N, 1 \le n+k \le M} \hat{p}_1(n)\hat{p}_2(n+k)$$

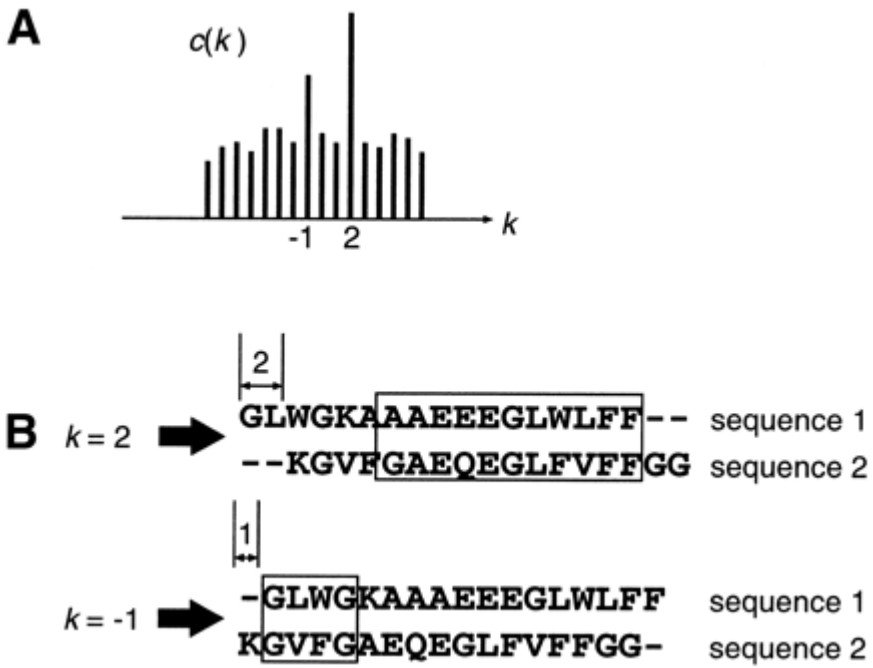
Finding Homologies

Original Computation

Slide box over all offsets

With FFT

Only look at offsets with large score



Generalization to Multiple Sequences

$$\hat{v}_{\text{group1}}(n) = \sum_{i \in \text{group1}} w_i \cdot \hat{v}_i(n),$$

$$\hat{p}_{\text{group1}}(n) = \sum_{i \in \text{group1}} w_i \cdot \hat{p}_i(n).$$

Alignment: Scoring System

$M\square_{ab} = [(M_{ab} - \Sigma_a f_a M_{aa})/(\Sigma_a f_a M_{aa} - \Sigma_{a,b} f_a f_b M_{ab})] + S^a$

f_a is frequency of a

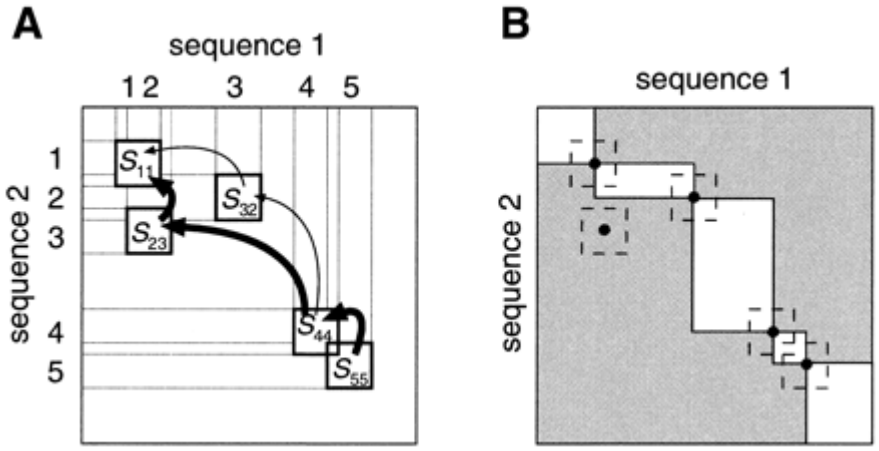
S^a is a predetermined gap extension penalty

$$H(i,j) = \sum_{n \in \text{group1}, m \in \text{group2}} w_n w_m \hat{M}_{A(n,\,i)B(m,\,j)},$$

Alignment

Can jump between homologies

Less computation than NW



$G_1(i, x) = S^{op} \cdot \{1 - [g_1{}^{start}(x) + g_1{}^{end}(i)]/2\}$

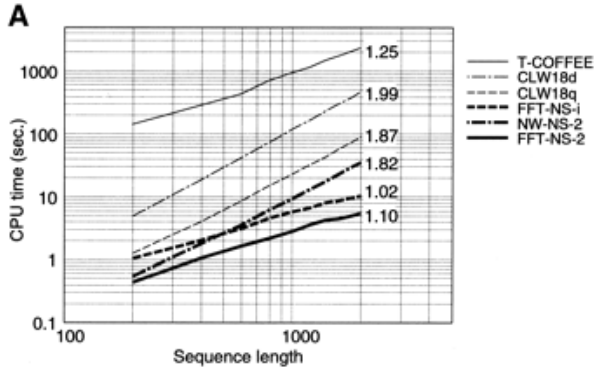
$$g_1^{start}(x) = \sum_{m \in \text{group1}} w_m \cdot a_m(x) \cdot z_m(x+1)$$

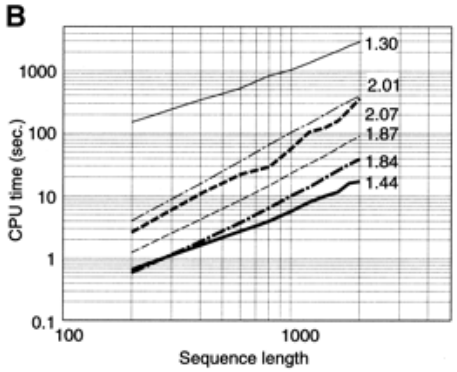
$$g_1^{end}(i) = \sum_{m \in \text{group1}} w_m \cdot z_m(i-1) \cdot a_m(i),$$

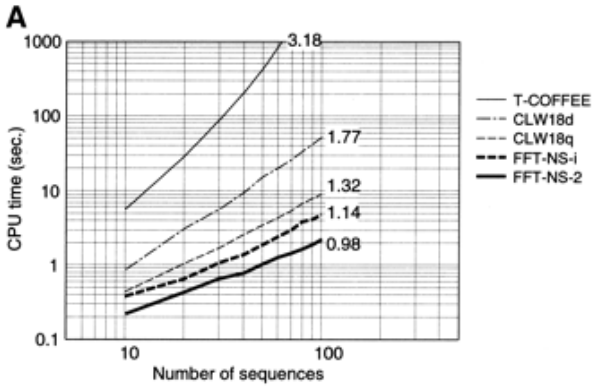
Comparisons

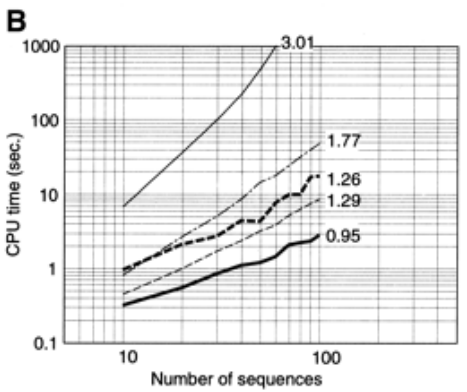
- MAFFT
  - FFT-NS-1
  - FFT-NS-2
  - FFT-NS-i
- T-COFFEE
- CLUSTALW
  - 1.82d
  - 1.82q

Runtime

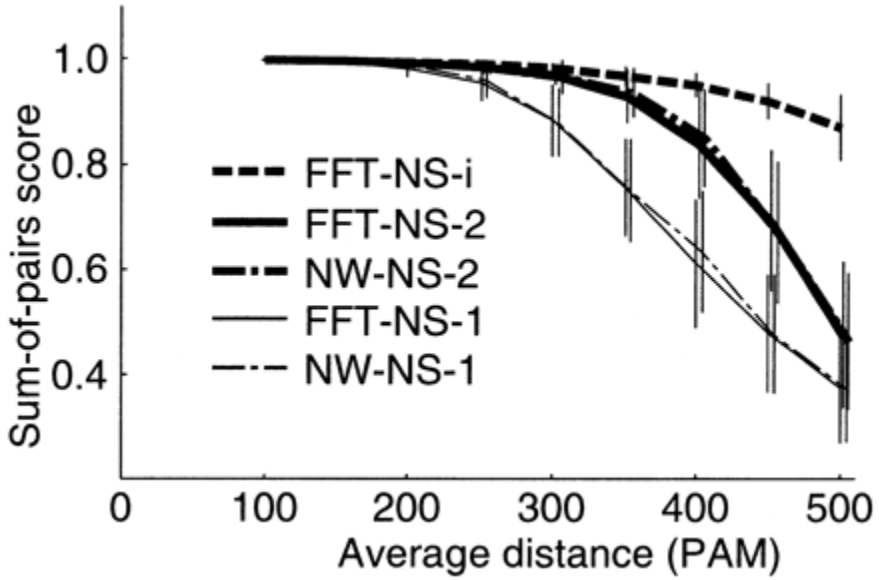

A
1000
100
10
1
0.1
CPU time (sec.)
100
1000
Sequence length
1.25
1.99
1.87
1.82
1.02
1.10
T-COFFEE
CLW18d
CLW18q
FFT-NS-i
NW-NS-2
FFT-NS-2


B
1000
100
10
1
0.1
CPU time (sec.)
100
1000
Sequence length
1.30
2.01
2.07
1.87
1.84
1.44


A
1000
100
10
1
0.1
CPU time (sec.)
10
100
Number of sequences
3.18
1.77
1.32
1.14
0.98
T-COFFEE
CLW18d
CLW18q
FFT-NS-i
FFT-NS-2


B
1000
100
10
1
0.1
CPU time (sec.)
10
100
Number of sequences
3.01
1.77
1.26
1.29
0.95

Sum-of-Pairs


Sum-of-pairs score
1.0
0.8
0.6
0.4
FFT-NS-i
FFT-NS-2
NW-NS-2
FFT-NS-1
NW-NS-1
0
100
200
300
400
500
Average distance (PAM)

BaliBASE Benchmarks

Method	CPU time (s)	Cat. 1	Cat. 2	Cat. 3	Cat. 4	Cat. 5	Average1	Average2
Progressive methods								
PIMA	1116	0.825/0.737	0.751/0.127	0.525/0.262	0.700/0.480	0.788/0.555	0.772/0.558	0.718/0.432
CLW18d	2202	0.871/0.792	0.856/0.329	0.754/0.490	0.745/0.417	0.852/0.617	0.844/0.639	0.816/0.529
CLW18q	1657	0.871/0.790	0.859/0.334	0.763/0.473	0.728/0.402	0.887/0.709	0.847/0.644	0.824/0.542
NW-AP-2	250	0.842/0.746	0.833/0.268	0.770/0.443	0.703/0.311	0.851/0.667	0.821/0.593	0.800/0.487
NW-NS-2	243	0.849/0.761	0.844/0.334	0.779/0.486	0.797/0.532	0.951/0.826	0.845/0.652	0.844/0.588
FFT-NS-2	227	0.849/0.761	0.844/0.334	0.779/0.486	0.797/0.532	0.951/0.826	0.845/0.652	0.844/0.588
Iterative refinement methods and T-COFFEE								
DIALIGN2-1	18132	0.792/0.681	0.814/0.219	0.673/0.327	0.818/0.615	0.938/0.840	0.801/0.584	0.807/0.536
PRRP	9782	0.871/0.793	0.860/0.354	0.823/0.569	0.663/0.275	0.885/0.742	0.845/0.646	0.820/0.547
T-COFFEE	12065	0.876/0.797	0.856/0.343	0.777/0.497	0.811/0.555	0.961/0.901	0.865/0.683	0.856/0.619
FFT-NS-i	1466	0.864/0.787	0.853/0.363	0.789/0.518	0.799/0.534	0.956/0.835	0.857/0.675	0.852/0.607
Number of alignments		82	23	12	15	12	144	–

LSU rRNA

Method	CPU time (s)	Sum-of-pairs score	Column score
72 sequences × 1305–5183 sites			
CLW18d	1998	0.692	–
CLW18q	600.2	0.597	–
NW-AP-2	197.0	0.796	–
NW-NS-2	205.2	0.770	–
FFT-NS-2	73.39	0.769	–
FFT-NS-i	251.8	0.781	–
59 sequences × 2810–5183 sites			
T-COFFEE	35 860	0.806	0.559
CLW18d	1523	0.754	0.411
CLW18q	395.6	0.643	0.315
NW-AP-2	153.7	0.823	0.482
NW-NS-2	159.8	0.793	0.463
FFT-NS-2	51.09	0.794	0.468
FFT-NS-i	181.7	0.817	0.552

RNA Polymerase Sequences

Method	CPU time (s)	Number of correctly aligned blocks
76 sequences × 1182–2890 sites		
CLW18d	675.5	10
CLW18q	159.4	10
NW-AP-2	54.95	8
NW-NS-2	59.30	11
FFT-NS-2	18.15	11
FFT-NS-i	173.1	11
24 sequences × 1206–2890 sites		
T-COFFEE	745.3	11
CLW18d	100.1	9
CLW18q	50.78	9
NW-AP-2	20.79	10
NW-NS-2	22.77	11
FFT-NS-2	7.150	11
FFT-NS-i	46.00	11

Questions