Multiple Sequence Alignment

Lectures 19 – Nov 30, 2011
CSE 527 Computational Biology, Fall 2011

Instructor: Su-In Lee
TA: Christopher Miles

Monday & Wednesday 12:00-1:20
Johnson Hall (JHN) 022

1

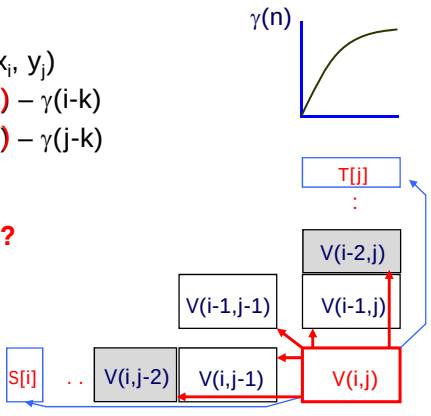
Revisit: General Gap Dynamic Programming

Iteration:

$$V(i, j) = \max \begin{cases} V(i-1, j-1) + s(x_i, y_j) \\ \max_{k=0\ldots i-1} V(k, j) - \gamma(i-k) \\ \max_{k=0\ldots j-1} V(i, k) - \gamma(j-k) \end{cases}$$

Previously...

Is this correct?

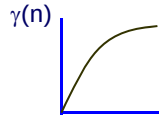


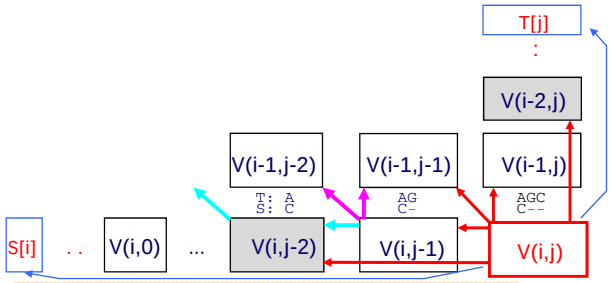
2

Revisit: General Gap Dynamic Programming

Iteration:

$$V(i, j) = \max \begin{cases} V(i-1, j-1) + s(x_i, y_j) \\ \max_{k=0\ldots i-1} V(k,j) - \gamma(i-k) \\ \max_{k=0\ldots j-1} V(i,k) - \gamma(j-k) \end{cases}$$





Is **max [V(i,j-2)- γ(2), V(i,j-1) - γ(1)]** still correct?

V(i,j-1) = V(i,j-2) - γ(1)

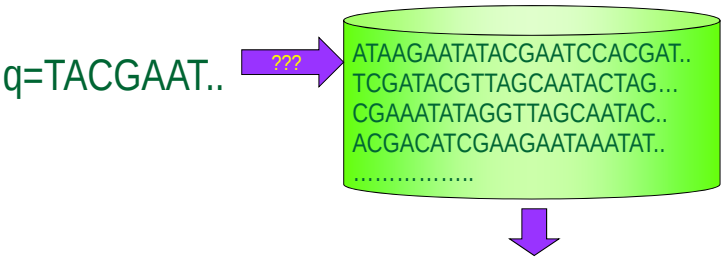
Then, **V(i,j-1) - γ(1) = V(i,j-2) - 2 γ(1) < V(i,j-2) - γ(2)**

3

Database Search

The problems:

- Dynamic programming: prohibitively complex
- Exact matching: prohibitively mismatch-sensitive



acACGAATa**TACGAAT**ccACGA-T..
_tACGAAT-**TACGAAT**-tACGAaT__

4

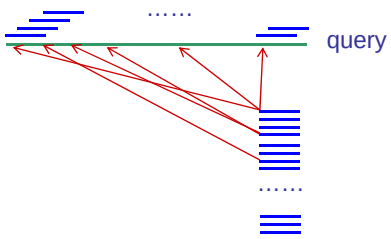
BLAST — Original Version

Dictionary:

All words of length k (~11)

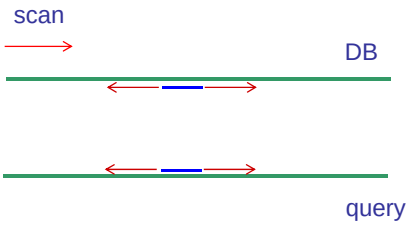
Match k-mers btw query and DB seqs

Alignment initiated between words of alignment score ≥ T (high scoring pairs)



Alignment:

Ungapped extensions until score below statistical threshold



Output:

All local alignments with score > statistical threshold

BLAST — Original Version

Example:

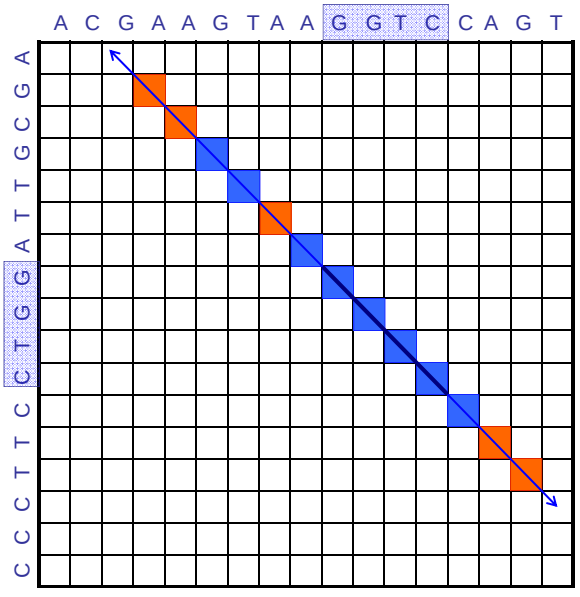
k = 4,
T = 4

The matching word GGTC initiates an alignment

Extension to the left and right with no gaps until alignment falls < 50%

Output:

GTAAGGTCC
GTTAGGTCC



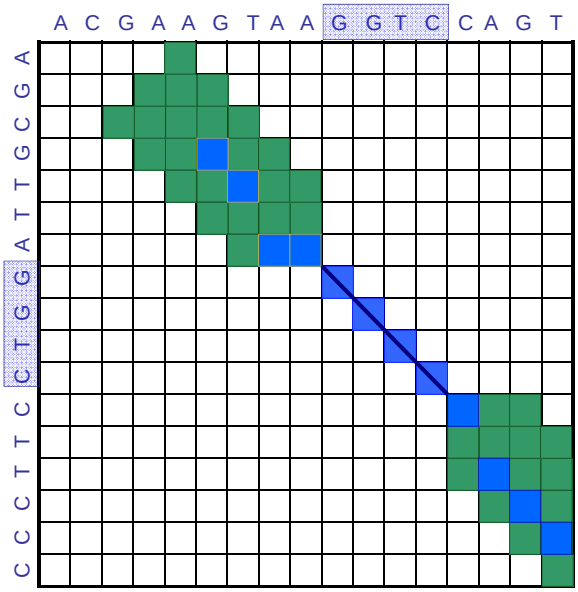
Gapped BLAST

Added features:

- Pairs of words can initiate alignment
- Extensions with gaps in a band around anchor

Output:

```
GTAAGGTCCAGT
GTTAGGTC-AGT
```



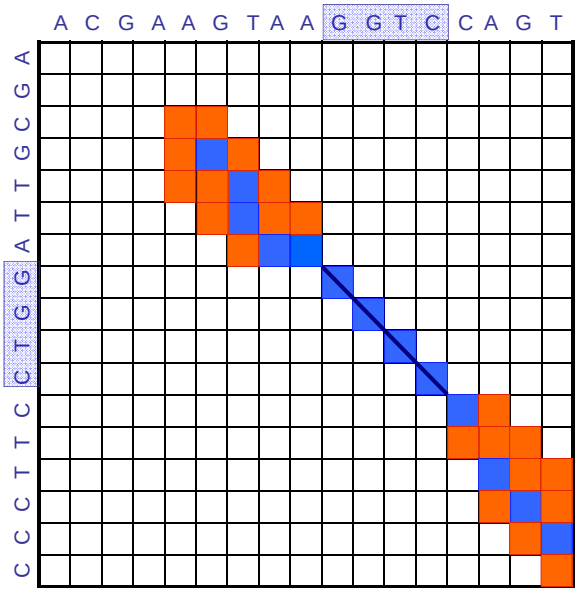
Gapped BLAST

Added features:

- Pairs of words can initiate alignment
- Nearby alignments are merged
- Extensions with gaps until score < T below best score so far

Output:

```
GTAAGGTCCAGT
GTTAGGTC-AGT
```



Example

Query: **gattacaccccgattacaccccgattaca** (29 letters)

[2 mins]

Database: All GenBank+EMBL+DDBJ+PDB sequences (but no EST, STS, GSS, or phase 0, 1 or 2 HTGS sequences) 1,726,556 sequences; **8,074,398,388 total letters**

>gi|28570323|gb|AC108906.9| Oryza sativa chromosome 3 BAC OSJNBa0087C10 genomic sequence, complete sequence Length = 144487 **Score = 34.2 bits** (17), Expect = 4.5 **Identities = 20/21 (95%)** Strand = Plus / Plus

```
Query:           4  tacaccccgattacaccccga 24
                    ||||||| ||||||||||||||
Sbjct:      125138 tacacccagattacaccccga 125158
```

Score = 34.2 bits (17),

Expect = 4.5 Identities = 20/21 (95%) Strand = Plus / Plus

```
Query:           4 tacaccccgattacaccccga 24
                   ||||||| ||||||||||||||
Sbjct:      125104 tacacccagattacaccccga 125124
```

>gi|28173089|gb|AC104321.7| Oryza sativa chromosome 3 BAC OSJNBa0052F07 genomic sequence, complete sequence Length = 139823 **Score = 34.2 bits** (17), Expect = 4.5 **Identities = 20/21 (95%)** Strand = Plus / Plus

```
Query:           4 tacaccccgattacaccccga 24
                   ||||||| ||||||||||||||
Sbjct:        3891 tacacccagattacaccccga 3911
```


Example

Query: Human atoh enhancer, 179 letters

[1.5 min]

Result: 57 blast hits

1.	gi|7677270|gb|AF218259.1|AF218259 Homo sapiens ATOH1 enhanc...	355	1e-95
2.	gi|22779500|gb|AC091158.11| Mus musculus Strain C57BL6/J ch...	264	4e-68
3.	gi|7677269|gb|AF218258.1|AF218258 Mus musculus Atoh1 enhanc...	256	9e-66
4.	gi|28875397|gb|AF467292.1| Gallus gallus CATH1 (CATH1) gene...	78	5e-12
5.	gi|27550980|emb|AL807792.6| Zebrafish DNA sequence from clo...	54	7e-05
6.	gi|22002129|gb|AC092389.4| Oryza sativa chromosome 10 BAC O...	44	0.068
7.	gi|22094122|ref|NM_013676.1| Mus musculus suppressor of Ty ...	42	0.27
8.	gi|13938031|gb|BC007132.1| Mus musculus, Similar to suppres...	42	0.27

gi|7677269|gb|AF218258.1|AF218258 Mus musculus Atoh1 enhancer sequence Length = 1517

Score = 256 bits (129), Expect = 9e-66 Identities = 167/177 (94%),

Gaps = 2/177 (1%) Strand = Plus / Plus

```
Query:    3 tgacaatagagggtctggcagaggctcctggccgcggtgcggagcgtctggagcggagca 62
            ||||||||||||| ||||||||||||||||||| ||||||||||||||||||||||||||
Sbjct: 1144 tgacaatagaggggctggcagaggctcctggccccggtgcggagcgtctggagcggagca 1203

Query:   63 cgcgctgtcagctggtgagcgcactctcctttcaggcagctccccggggagctgtgcggc 122
            ||||||||||||||||||||||||| ||||||||| ||||||||||||||| |||||
Sbjct: 1204 cgcgctgtcagctggtgagcgcactc-gctttcaggccgctccccggggagctgagcggc 1262

Query:  123 cacatttaacaccatcatcacccctccccggcctcctcaacctcggcctcctcctcg 179
            ||||||||||||| || ||| ||||||||||||||||||| ||||||||||||||
Sbjct: 1263 cacatttaacaccgtcgtca-ccctccccggcctcctcaacatcggcctcctcctcg 1318
```

Outline

- Review: database search
  - BLAST
- Multiple sequence alignment
  - Progressive multiple alignment methods (fast and simple)
    - PileUp, Clustal
  - Iterative methods (slow but accurate)
    - Muscle
  - Consistency-based method (slow but accurate)
    - T-coffee, ProbCons

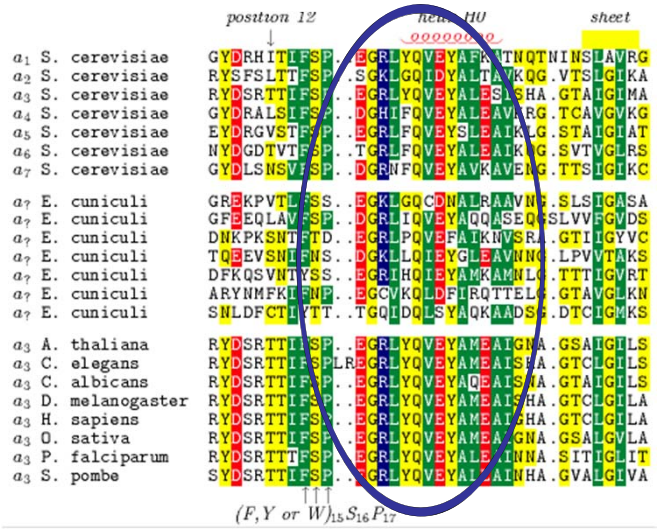
11

Why Multiple Alignment?

- Alignment of sequences allows us to examine homologous regions
  - Two proteins with regions of high sequence similarity are likely to perform the same function
  - Conserved regions point to structural similarity

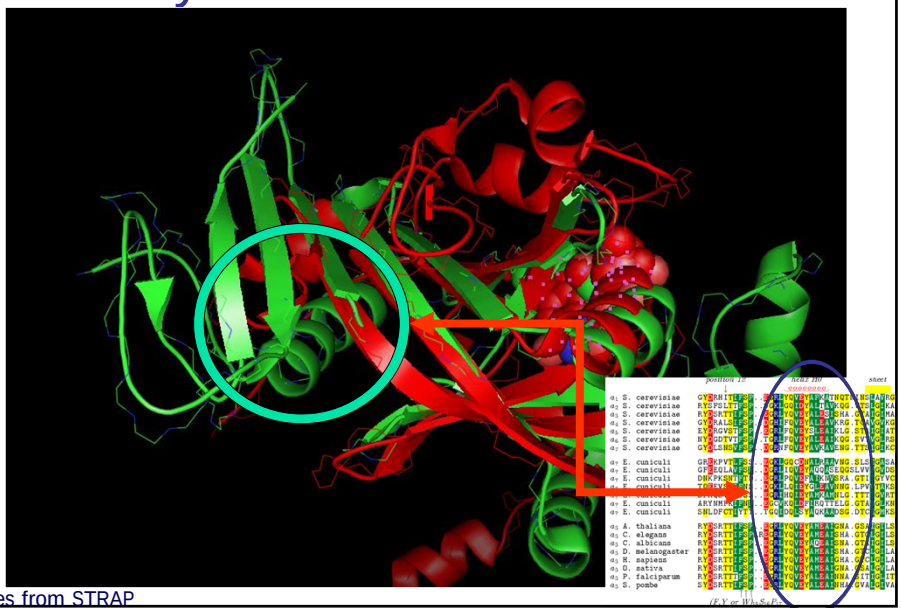
12

Multiple Sequence Alignment



Images from STRAP

Aligned Regions Represent Spatial Similarity



Images from STRAP

Why Multiple Alignment?

- Alignment of sequences allows us to examine homologous regions
  - ~~Two~~ proteins with regions of high sequence similarity are likely to perform the same function
  - Conserved regions point to structural similarity
- Evolutionary history can be inferred from similarity
  - Aligned residues ~~pair~~ should have evolved form the same ancestral residue
- We need to generalize the pairwise sequence alignment method to >2 input sequences.
  → **Multiple sequence alignment methods**

15

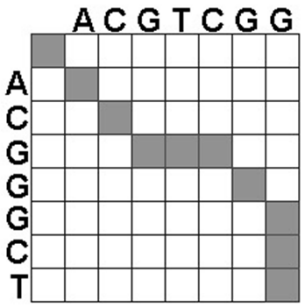
Recap on Alignments

- Classic pairwise sequence alignment
  - Dynamic programming approaches
  - Use affine gap penalty to more accurately model evolutionary events

16

Sequence Alignment Revisited

- Two sequences of length L requires O(L^2) space and O(L^2) time.



$$V(i,j) = \max \begin{cases} V(i\text{-}1,j\text{-}1) + \sigma(S[i],T[j]) \\ V(i\text{-}1,j) + \sigma(S[i], \text{-}) \\ V(i,j\text{-}1) + \sigma(\text{-}, T[j]) \end{cases},$$

17

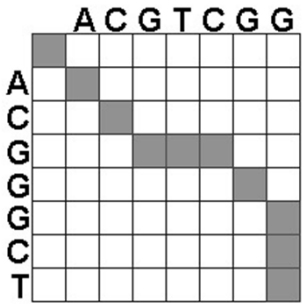
Recap on Alignments

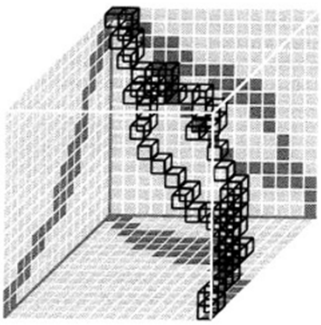
- Classic pairwise sequence alignment
  - Dynamic programming approaches
  - Use affine gap penalty to more accurately model evolutionary events
- Multiple sequence alignment approaches
  - Why is the classic pairwise alignment not extendable to multiple sequences?

18

Sequence Alignment Revisited

- Two sequences of length N requires $O(L^2)$ space and $O(L^2)$ time.





- Three sequences of length N requires $O(L^3)$ space and $O(L^3)$ time.

19

Image from Durbin et al

Sequence Alignment Revisited

- Two sequences of length L require $O(L^2)$ space and $O(L^2)$ time
- Three sequences of length L require $O(L^3)$ space and $O(L^3)$ time

Four sequences?

N sequences?

Generally time is $O(L^N)$

20

Run-time for the Calculations

- Let's assume that we have N sequences of length L
  - Time complexity is $O(L^N)$.
  - Assume this computation takes $10^{(2N-4)}$ seconds.

Two sequences take 1 second
Three sequences take 10 seconds

In our example they had N=12 sequences
$10^{2x12-4} = 10^{20}$ seconds
3 trillion years!!

21

Solutions

- Heuristic approaches to multiple sequence alignment
  - Progressive multiple alignment methods (fast and simple)
    - PileUp, Clustal
  - Iterative multiple alignment methods (slow but accurate)
    - Muscle
  - Consistency-based multiple alignment methods
    - T-coffee, ProbCons

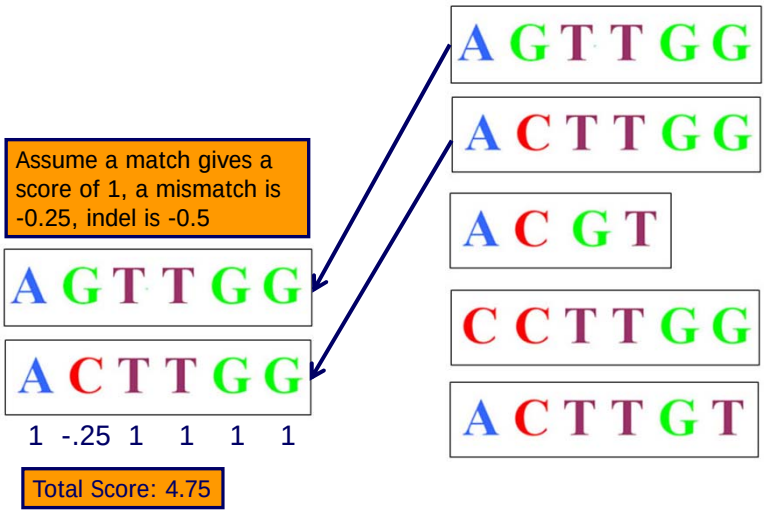
22

Progressive Multiple Alignment

- The most practical and widely used method in multiple sequence alignment is the hierarchical extensions of pairwise alignment methods.
- Key idea: multiple alignments is achieved by successive application of pairwise methods.

Progressive Multiple Alignment

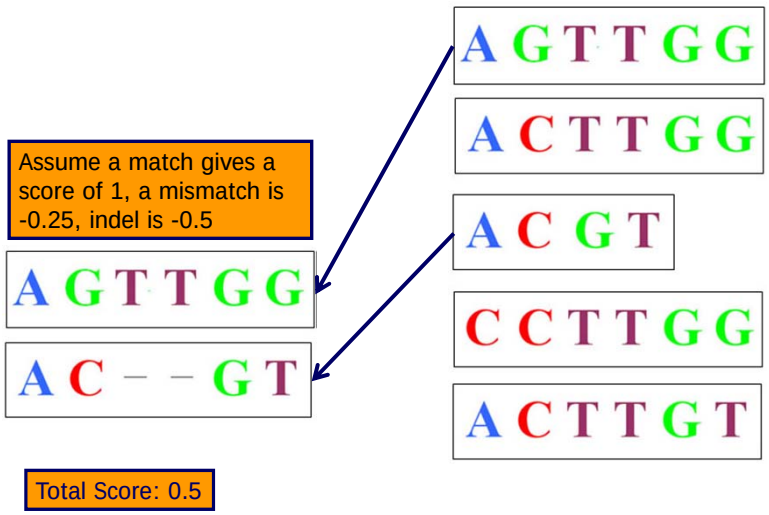
- Perform pairwise alignments for all sequences



24

Progressive Multiple Alignment

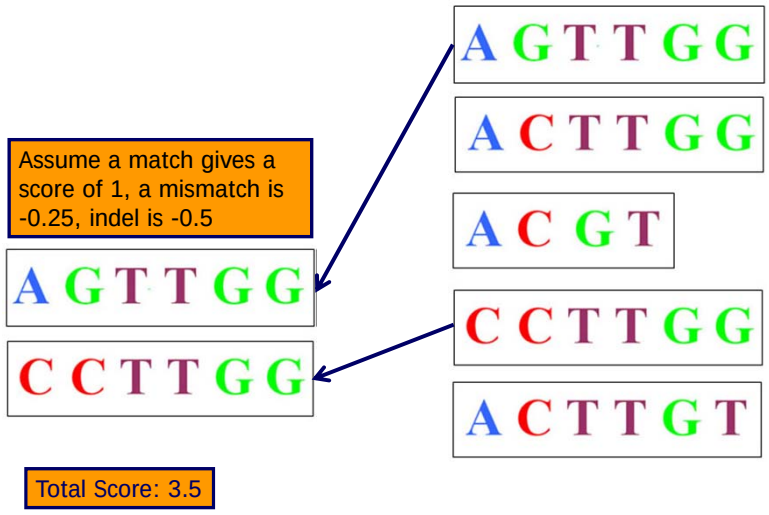
- Perform pairwise alignments for all sequences



25

Progressive Multiple Alignment

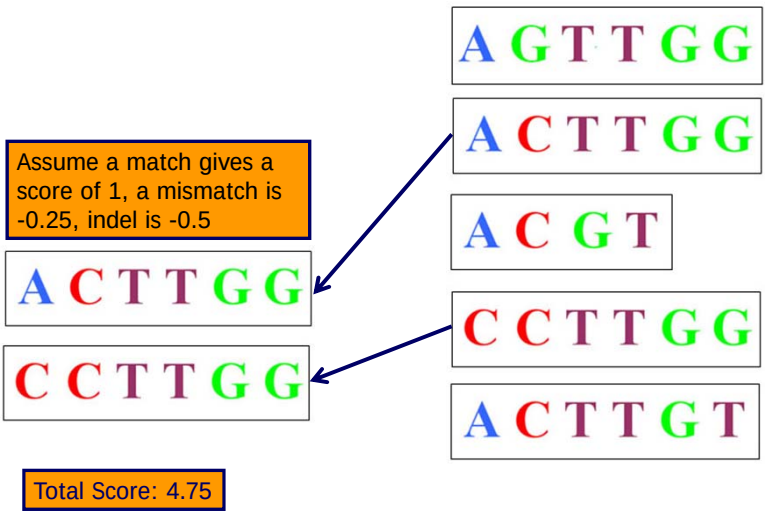
- Perform pairwise alignments for all sequences



26

Progressive Multiple Alignment

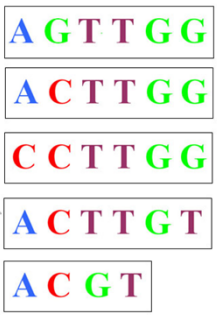
- Perform pairwise alignments for all sequences



27

Progressive Multiple Alignment

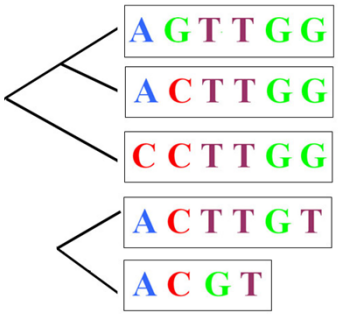
- Create guide tree from pairwise alignments



28

Progressive Multiple Alignment

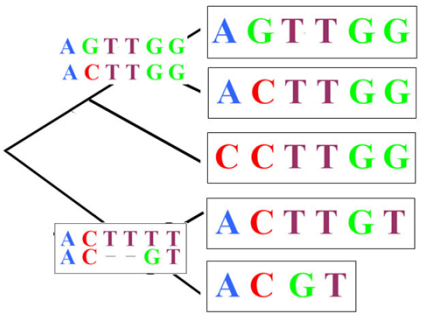
- Create guide tree from pairwise alignments
- Use tree to build multiple sequence alignment



29

Progressive Multiple Alignment

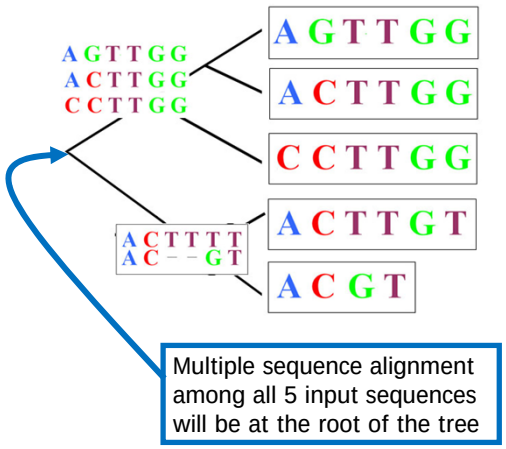
- Create guide tree from pairwise alignments
- Use tree to build multiple sequence alignment
  - Align most similar sequences first (give the most reliable alignments)
  - Align the profile to the next closest sequence



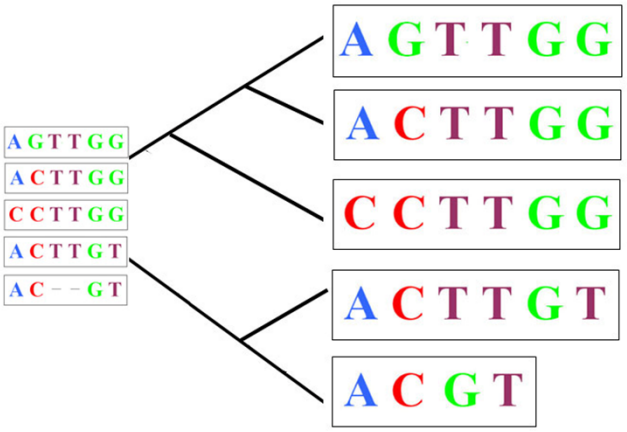
30

Progressive Multiple Alignment

- Create guide tree from pairwise alignments
- Use tree to build multiple sequence alignment
  - Align most similar sequences first (give the most reliable alignments)
  - Align the profile to the next closest sequence



Progressive Multiple Alignment



32

The PileUp Algorithm

- First, PileUp calculates approximate pairwise similarity scores between all sequences to be aligned, and they are clustered into a dendrogram (tree structure).
- Then the most similar pairs of sequences are aligned.
- Averages (similar to consensus sequences) are calculated for the aligned pairs.
- New sequences and clusters of sequences are added one by one, according to the branching order in the dendrogram.

33

PileUp website – http://hku.hk/bruhk/gcgdoc/pileup.html

Choosing Sequences for PileUp

- As far as possible, try to align sequences of similar length.
- PileUp can align sequences of up to 5000 residues, with 2000 gaps (total 7000 characters).
- PileUp is a good program only for similar (close) sequences.

34

PileUp considerations

- PileUp does global multiple alignment, and therefore is good for a group of similar sequences.
- PileUp will fail to find the best local region of similarity (such as a shared motif) among distant related sequences.
- PileUp always aligns all of the sequences you specified in the input file, even if they are not related.
  - The alignment can be degraded if some of the sequences are only distantly related.

35

PileUp Considerations

- Since the alignment is calculated on a progressive basis, the order of the initial sequences can affect the final alignment.

- PileUp parameters: 2 gap penalties (gap insert and gap extend) and an amino acid comparison matrix (e.g. BLOSUM62).
- PileUp will refuse to align sequences that require too many gaps or mismatches.
- PileUp will take quite a while to align more than about 10 sequences

36

CLUSTAL

- Clustal is a stand-alone (i.e. not integrated into GCG*) multiple alignment program that is superior in some respects to PileUp
- Works by progressive alignment: it aligns a pair of sequences then aligns the next one onto the first pair
- Most closely related sequences are aligned first, and then additional sequences and groups of sequences are added, guided by the initial alignments
- Uses alignment scores to produce a phylogenetic tree

* GCG (Genetic Computer Group) is a package of sequence analysis program

CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. Thompson, J.D., Higgins, D.G. and Gibson, T.J. Nucleic Acids Research 22, 4673-4680 (1994).

CLUSTAL

- Aligns the sequences sequentially, guided by the phylogenetic relationships indicated by the tree
- Is available with a great web interface: http://www.ebi.ac.uk/clustalw/
- Also available in Biology Workbench

38

Comparison

- Main differences between PileUp and Clustal:
  - The metric used to compare the sequences for the initial "guide tree" uses a full global, optimal alignment in PileUp instead of the fast, approximate ones in Clustal. This makes PileUp much slower for the comparison of long sequences. In principle, the distances calculated from PileUP will be more sensitive than ours, but in practice it will not make much difference, except in difficult cases.
  - During the multiple alignment, terminal gaps are penalised in Clustal but not in PileUp. This will make the PileUp alignments better when the sequences are of very different lengths (has no effect if there are no large terminal gaps).

39

Multiple Alignment tools on the Web

- There are a variety of multiple alignment tools available for free on the web.
- Clustal is available from a number of sites (with a variety of restrictions)
- Other algorithms are available too

40

Outline

- Review: database search
  - BLAST
- Multiple sequence alignment
  - Progressive multiple alignment methods (fast and simple)
    - PileUp, Clustal
  - Iterative methods (slow but accurate)
    - Muscle
  - Consistency-based method (slow but accurate)
    - T-coffee, ProbCons

41

What We've Covered So Far

Lecture 1: Course logistics, short intro to molecular biology, example project topics [PPT] [PDF]

Lecture 2: Introduction to Bayesian networks for computational biology [PPT] [PDF]

Lecture 3: Maximum Likelihood Estimation, Expectation Maximization [PPT] [PDF]

Lecture 4: Genetic basics, QTL mapping, Association studies [PPT] [PDF]

Lecture 5: QTL mapping, haplotypes [PPT] [PDF]

Lecture 6: Haplotype reconstruction [PPT] [PDF]

Lecture 7: Disease association studies [PPT] [PDF]

Lecture 8: Linkage analysis [PPT] [PDF]

Lecture 9: Inferring transcriptional regulatory networks I [PDF] [PPT]

Lecture 10: Inferring transcriptional regulatory networks II [PDF] [PPT]

Lecture 11: Advanced topics in inferring regulatory networks [PDF]

Lecture 12: Regulatory motif finding I [PDF]

Lecture 13: Regulatory motif finding II [PDF]

Lecture 14: Inferring the signaling networks I [PDF]

Lecture 15: Inferring the signaling networks II [PDF]

Lecture 16: Sequence alignment [PDF]

Lecture 17: Scoring alignments [PDF]

Lecture 18: Local sequence alignment and heuristic local aligners [PDF]

Lecture 19: Multiple sequence alignment I [PDF]

Lecture 20: Multiple sequence alignment II [PDF]

ML basics (Bayesian networks, MLE, EM)

Genetics (association studies, phasing, linkage analysis)

Systems biology (gene regulation, gene interaction)

Sequence analysis

42