Incremental algorithms for large homologous gene families

Jean-François Dufayard

Manolo Gouy

François Rechenmann

HELiX
BIOINFORMATICS

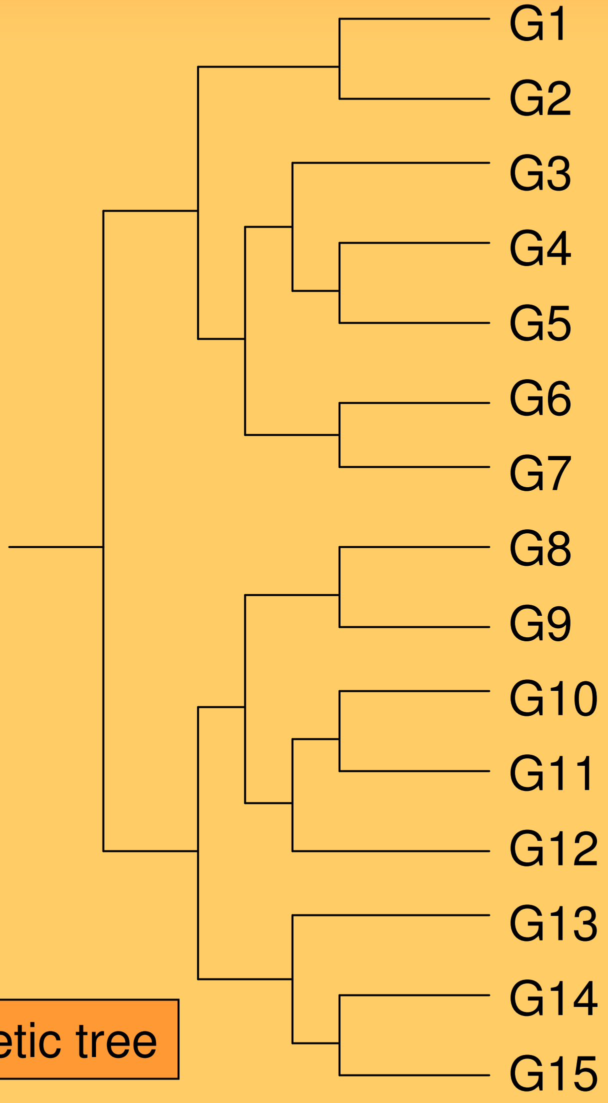
UMR 5558
Biométrie, biologie
évolutive.

INRIA
RHÔNE-ALPES

Problematic (1/3)

Phylogenic analysis

```
L-QGNDLLEDSPYEPVNSRLSDIFRAVPFISDVFQQVEHI-SKGNNCLDAAKACNLDDTC
L-QGNDLLEDSPYEPVNSRLSDIFRAVPFISDVFQQVEHI-SKGNNCLDAAKACNLDDTC
LTEGEEFYEASPYEPVTSRLSDIFRLASIFSGTGADP-VVSAKSNHCLDAAKACNLNDNC
LTEGEEFYEASPYEPVTSRLSDIFRLASIFSGTGADP-VVSAKSNHCLDAAKACNLNDNC
L-QGNDLLEDSPYEPVNSRLSDIFRLAPIVS-----VEPVLSKGNNCLDAAKACNLNDTC
LAEGEEFYEASPYEPITSRLSDIFRLASIFSGM--DP-ATNSKSNHCLDAAKACNLNDNC
LTEGEEFYEASPYEPVTSRLSDIFRLASIFSGTGTDP-AVSTKSNHCLDAAKACNLNDNC
L-QGNDLLEDSPYEPVNSRLSDIFRAVPFIS-----VEHI-SKGNNCLDAAKACNLDDTC
L-QGNDLLEDSPYEPVNSRLSDIFRVVPFIS-----VEHI-PKGNNCLDAAKACNLDDIC
L-QGNDLLEDSPYEPVNSRLSDIFRVVPFISDVFQQVEHI-PKGNNCLDAAKACNLDDIC
L-QGNDLLEDSPYEPVNSRLSDIFRAVPFIS-----VEHI-SKGNNCLDAAKACNLDDTC
LTEGEEFYEASPYEPVTSRLSDIFRLASIFSGTGTDP-AVSTKSNHCLDAAKACNLNDNC
LMEGMNVLESSPYEPFIRGF-DYVRLASITAGSENEVTQV----NRCLDAAKACNVDEMC
---GEEFYEASPYEPVTSRLSDIFRLASIFSGTGADP-VVSAESNHCLDAAKACNLNDNC
---G----------------------------TGADP-VVSAESNHCLDAAKACNLNDNC
```

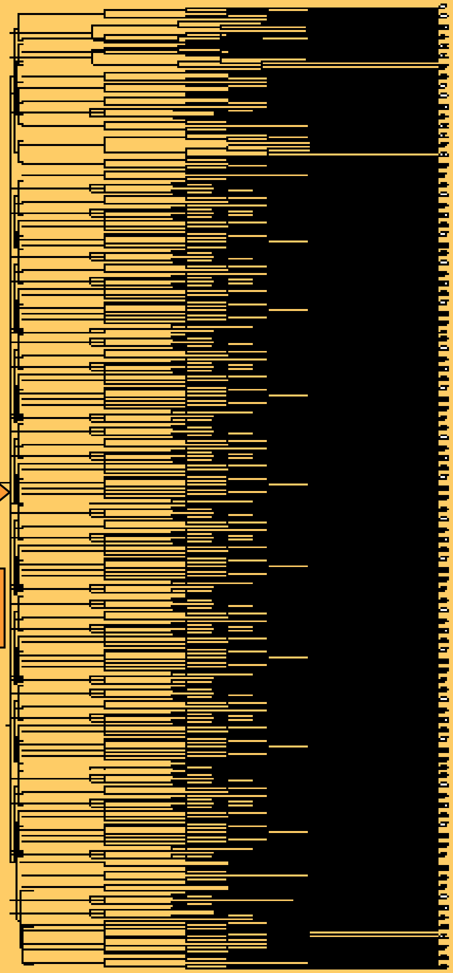


Multiple alignment

Phylogenetic tree

Problematic (2/3)

```
L-QGNDLLEDSPYEPVNSRLSDIFRAVPFISDVFQQVEHI-SKGNNCLDAAKACNLDDTC
L-QGNDLLEDSPYEPVNSRLSDIFRAVPFISDVFQQVEHI-SKGNNCLDAAKACNLDDTC
LTEGEEFYEASPYEPVTSRLSDIFRLASIFSGTGADP-VVSAKSNHCLDAAKACNLNDNC
LTEGEEFYEASPYEPVTSRLSDIFRLASIFSGTGADP-VVSAKSNHCLDAAKACNLNDNC
L-QGNDLLEDSPYEPVNSRLSDIFRLAPIVS-----VEPVLSKGNNCLDAAKACNLNDTC
LAEGEEFYEASPYEPITSRLSDIFRLASIFSGM--DP-ATNSKSNHCLDAAKACNLNDNC
LTEGEEFYEASPYEPVTSRLSDIFRLASIFSGTGTDP-AVSTKSNHCLDAAKACNLNDNC
L-QGNDLLEDSPYEPVNSRLSDIFRAVPFIS-----VEHI-SKGNNCLDAAKACNLDDTC
L-QGNDLLEDSPYEPVNSRLSDIFRVVPFIS-----VEHI-PKGNNCLDAAKACNLDDIC
L-QGNDLLEDSPYEPVNSRLSDIFRVVPFISDVFQQVEHI-PKGNNCLDAAKACNLDDIC
L-QGNDLLEDSPYEPVNSRLSDIFRAVPFIS-----VEHI-SKGNNCLDAAKACNLDDTC
LTEGEEFYEASPYEPVTSRLSDIFRLASIFSGTGTDP-AVSTKSNHCLDAAKACNLNDNC
LMEGMNVLESSPYEPFIRGF-DYVRLASITAGSENEVTQV----NRCLDAAKACNVDEMC
---GEEFYEASPYEPVTSRLSDIFRLASIFSGTGADP-VVSAESNHCLDAAKACNLNDNC
---G----------------------------TGADP-VVSAESNHCLDAAKACNLNDNC
L-QGNDLLEDSPYEPVNSRLSDIFRAVPFISDVFQQVEHI-SKGNNCLDAAKACNLDDTC
L-QGNDLLEDSPYEPVNSRLSDIFRAVPFISDVFQQVEHI-SKGNNCLDAAKACNLDDTC
LTEGEEFYEASPYEPVTSRLSDIFRLASIFSGTGADP-VVSAKSNHCLDAAKACNLNDNC
LTEGEEFYEASPYEPVTSRLSDIFRLASIFSGTGADP-VVSAKSNHCLDAAKACNLN
L-QGNDLLEDSPYEPVNSRLSDIFRLAPIVS-----VEPVLSKGNNCLDAAKACNLNDTC
LAEGEEFYEASPYEPITSRLSDIFRLASIFSGM--DP-ATNSKSNHCLDAAKACNLNDNC
LTEGEEFYEASPYEPVTSRLSDIFRLASIFSGTGTDP-AVSTKSNHCLDAAKA
L-QGNDLLEDSPYEPVNSRLSDIFRAVPFIS-----VEHI-SKGNNCLDAAKA
L-QGNDLLEDSPYEPVNSRLSDIFRVVPFIS-----VEHI-PKGNNCLDAAKACNLDDIC
L-QGNDLLEDSPYEPVNSRLSDIFRVVPFISDVFQQVEHI-PKGNNCLDAAKACNLDDIC
L-QGNDLLEDSPYEPVNSRLSDIFRAVPFIS-----VEHI-SKGNNCLDAAKACNLDDTC
LTEGEEFYEASPYEPVTSRLSDIFRLASIFSGTGTDP-AVSTKSNHCLDAAKACNLNDNC
LMEGMNVLESSPYEPFIRGF-DYVRLASITAGSENEVTQV----NRCLDAAKACNVDEMC
---GEEFYEASPYEPVTSRLSDIFRLASIFSGTGADP-VVSAESNHCLDAAKACNLNDNC
---G----------------------------TGADP-VVSAESNHCLDAAKACNLNDNC
L-QGNDLLEDSPYEPVNSRLSDIFRAVPFISDVFQQVEHI-SKGNNCLDAAKACNLDDTC
L-QGNDLLEDSPYEPVNSRLSDIFRAVPFISDVFQQVEHI-SKGNNCLDAAKACNLDDTC
LTEGEEFYEASPYEPVTSRLSDIFRLASIFSGTGADP-VVSAKSNHCLDAAKACNLNDNC
LTEGEEFYEASPYEPVTSRLSDIFRLASIFSGTGADP-VVSAKSNHCLDAAKACNLNDNC
L-QGNDLLEDSPYEPVNSRLSDIFRLAPIVS-----VEPVLSKGNNCLDAAKACNLNDTC
LAEGEEFYEASPYEPITSRLSDIFRLASIFSGM--DP-ATNSKSNHCLDAAKACNLNDNC
LTEGEEFYEASPYEPVTSRLSDIFRLASIFSGTGTDP-AVSTKSNHCLDAAKACNLNDNC
```



New sequence

Incremental algorithms

Problematic (3/3)

Main goal: add sequences in a phylogenetic tree and its alignment

- Avoid redondant calculations
- Preserve the quality of the tree and the alignment

Application: *European Small Subunit Ribosomal RNA database*

- Family of 35 000 sequences (~1500 nucleotids per sequence)
- A tree and an alignment is computed for 10 000 sequences, **manually**.
- New sequences frequently added to the database.

Adding a new sequence, main problems

- How to find the location of the new sequence in the alignment ?

- How to find the location of the new sequence in the tree ?

- How to insert the new sequence in the alignment, knowing its location ?

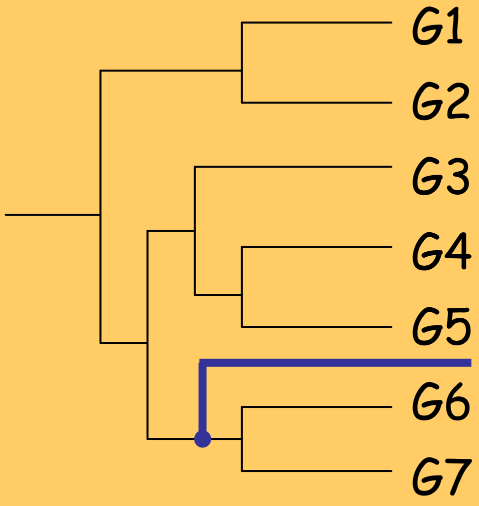
- How to insert the new sequence in the tree, knowing its location ?

Location of the new sequence (1/2)

- How to find the location of the new sequence in the alignment ?

⇕

- How to find the location of the new sequence in the tree ?



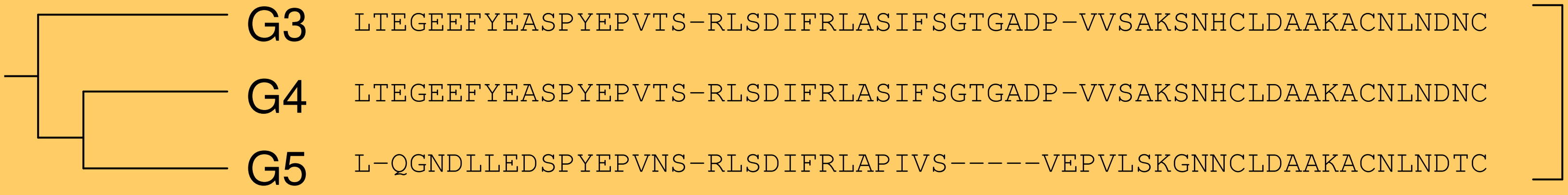
```
L-QGNDLLEDSPYEPVNSRLSDIFRAVPFISDVFQQVEHI-SKGNNCLDAAKACNLDDTC
L-QGNDLLEDSPYEPVNSRLSDIFRAVPFISDVFQQVEHI-SKGNNCLDAAKACNLDDTC
LTEGEEFYEASPYEPVTSRLSDIFRLASIFSGTGADP-VVSAKSNHCLDAAKACNLNDNC
LTEGEEFYEASPYEPVTSRLSDIFRLASIFSGTGADP-VVSAKSNHCLDAAKACNLNDNC
L-QGNDLLEDSPYEPVNSRLSDIFRLAPIVS-----VEPVLSKGNNCLDAAKACNLNDTC
LTEGEEFYEASPYEPVTSRLSDIFRLASIFSGTGTDP-AVSTKSNHCLDAAKACNLNDNC
LAEGEEFYEASPYEPITSRLSDIFRLASIFSGM--DP-ATNSKSNHCLDAAKACNLNDNC
LTEGEEFYEASPYEPVTSRLSDIFRLASIFSGTGTDP-AVSTKSNHCLDAAKACNLNDNC
```

Location of the new sequence (2/2)

If S and the outgroup are grouped:

- •Remove independant gaps in the founded block, and in S.
- •Align S and the founded block.
- •Restart the research in the founded block.

Stop when the research doesn't bring an amelioration.



```
G3   LTEGEEFYEASPYEPVTS-RLSDIFRLASIFSGTGADP-VVSAKSNHCLDAAKACNLNDNC
G4   LTEGEEFYEASPYEPVTS-RLSDIFRLASIFSGTGADP-VVSAKSNHCLDAAKACNLNDNC
G5   L-QGNDLLEDSPYEPVNS-RLSDIFRLAPIVS-----VEPVLSKGNNCLDAAKACNLNDTC

S    LTEGEEFYEA-------SSRLSDIFRLASIFSGTGTDP-AVSTKSNHCLDAAKACNLNDNC
```

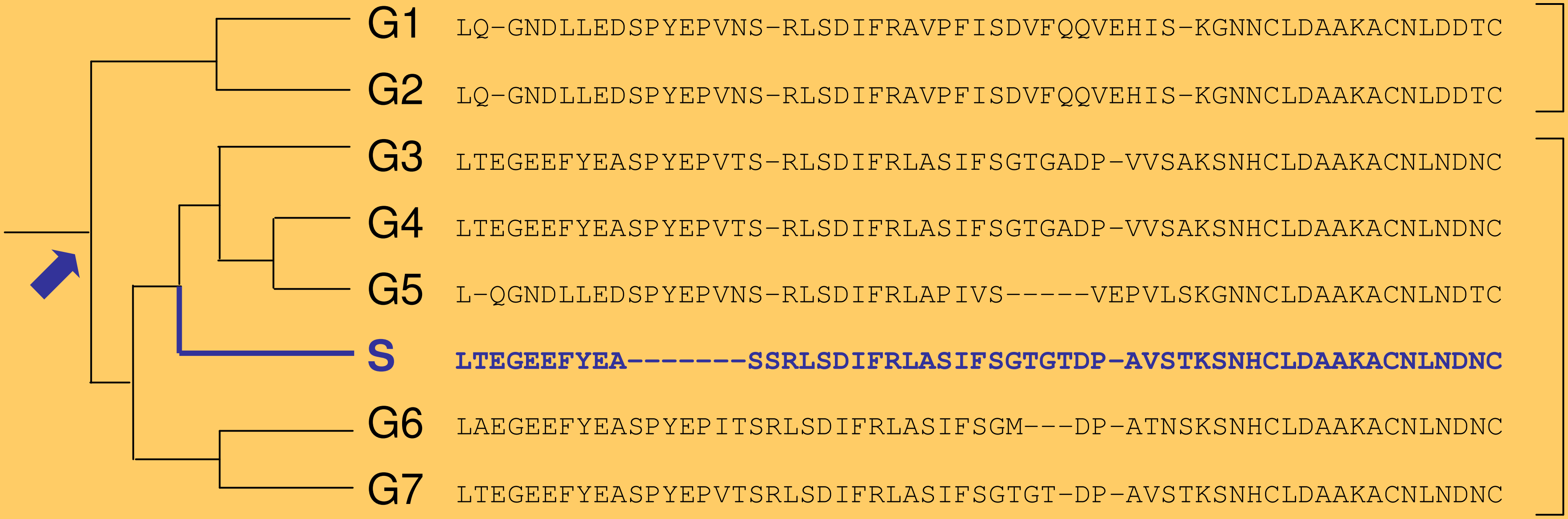
Adding the new sequence to the alignment

- How to insert the new sequence in the alignment, knowing its location ?

⇕

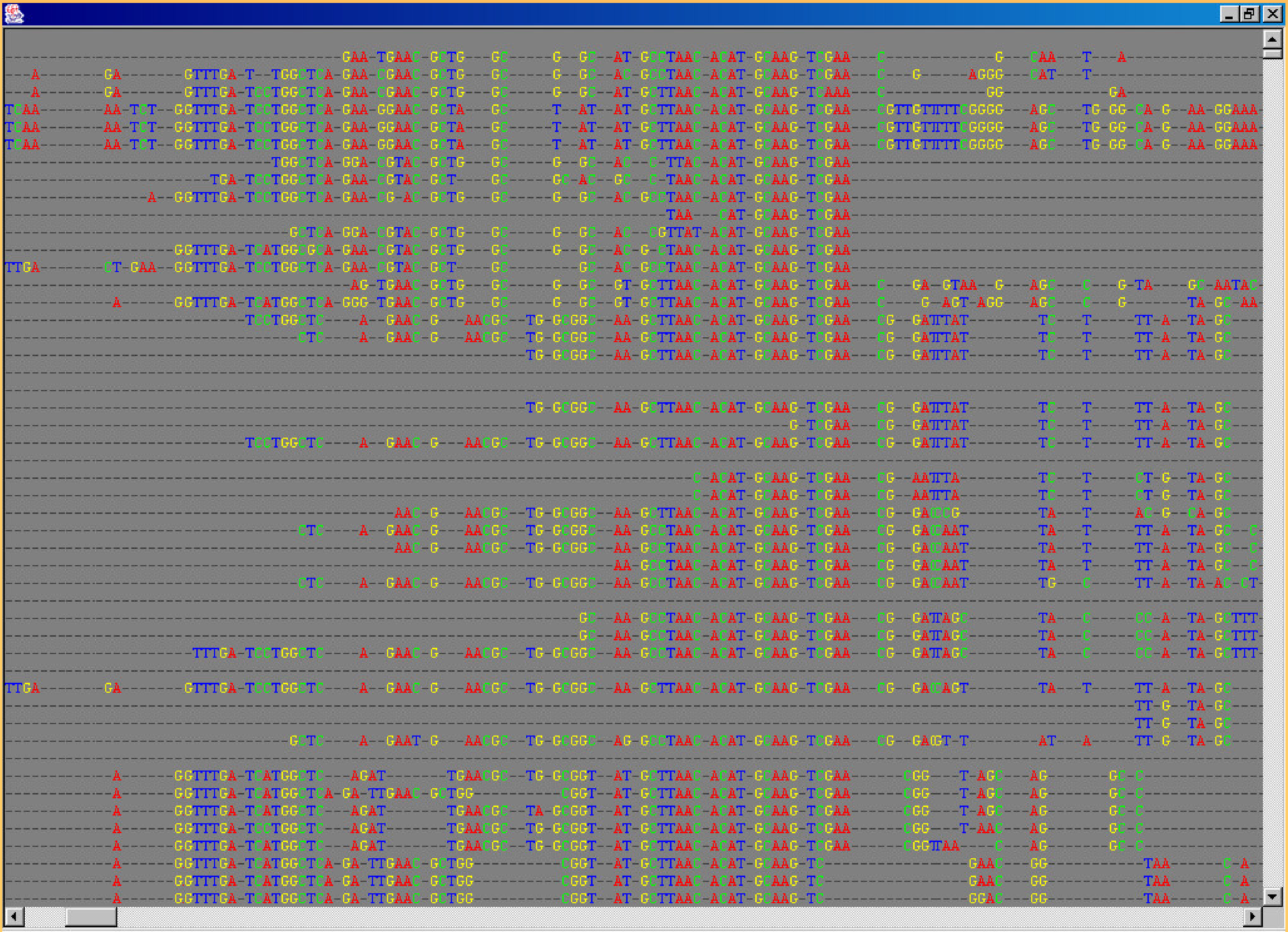
- What would be the alignment, if entirely recomputed with S, using the *progressive multiple alignment* method ?

  ➢ Recompute block alignments from current node to the root.



```
G1  LQ-GNDLLEDSPYEPVNS-RLSDIFRAVPFISDVFQQVEHIS-KGNNCLDAAKACNLDDTC
G2  LQ-GNDLLEDSPYEPVNS-RLSDIFRAVPFISDVFQQVEHIS-KGNNCLDAAKACNLDDTC
G3  LTEGEEFYEASPYEPVTS-RLSDIFRLASIFSGTGADP-VVSAKSNHCLDAAKACNLNDNC
G4  LTEGEEFYEASPYEPVTS-RLSDIFRLASIFSGTGADP-VVSAKSNHCLDAAKACNLNDNC
G5  L-QGNDLLEDSPYEPVNS-RLSDIFRLAPIVS-----VEPVLSKGNNCLDAAKACNLNDTC
S   LTEGEEFYEA-------SSRLSDIFRLASIFSGTGTDP-AVSTKSNHCLDAAKACNLNDNC
G6  LAEGEEFYEASPYEPITSRLSDIFRLASIFSGM---DP-ATNSKSNHCLDAAKACNLNDNC
G7  LTEGEEFYEASPYEPVTSRLSDIFRLASIFSGTGT-DP-AVSTKSNHCLDAAKACNLNDNC
```

Prototype



```
--------------------------------------GAA-TGAAC-GCTG---GC-----G--GC--AT-GCCTAAC-ACAT-GCAAG-TCGAA---C------------G--CAA---T---A----------------
---A-------GA-------GTTTGA-T--TGGCTCA-GAA-CGAAC-GCTG---GC-----G--GC--AC-GCCTAAC-ACAT-GCAAG-TCGAA---C---G-----AGGG--CAT---T--------------------
---A-------GA-------GTTTGA-TCCTGGCTCA-GAA-CGAAC-GCTG---GC-----G--GC--AT-GCTTAAC-ACAT-GCAAG-TCAAA---C-----------GG-----------GA----------------
TCAA-------AA-TCT--GGTTTGA-TCCTGGCTCA-GAA-GGAAC-GCTA---GC-----T--AT--AT-GCTTAAC-ACAT-GCAAG-TCGAA---CGTTGTTTTCGGGG--AGC---TG-GG-CA-G--AA-GGAAA-
TCAA-------AA-TCT--GGTTTGA-TCCTGGCTCA-GAA-GGAAC-GCTA---GC-----T--AT--AT-GCTTAAC-ACAT-GCAAG-TCGAA---CGTTGTTTTCGGGG--AGC---TG-GG-CA-G--AA-GGAAA-
TCAA-------AA-TCT--GGTTTGA-TCCTGGCTCA-GAA-GGAAC-GCTA---GC-----T--AT--AT-GCTTAAC-ACAT-GCAAG-TCGAA---CGTTGTTTTCGGGG--AGC---TG-GG-CA-G--AA-GGAAA-
------------------------------TGGCTCA-GGA-CGTAC-GCTG---GC-----G--GC--AC--C-TTAC-ACAT-GCAAG-TCGAA----------------------------------------------
-----------------------TGA-TCCTGGCTCA-GAA-CGTAC-GCT----GC-----GC-AC--GC--C-TAAC-ACAT-GCAAG-TCGAA----------------------------------------------
----------------A--GGTTTGA-TCCTGGCTCA-GAA-CG-AC-GCTG---GC-----G--GC--AC-GCCTAAC-ACAT-GCAAG-TCGAA----------------------------------------------
---------------------------------------------------------------------------TAA---CAT-GCAAG-TCGAA----------------------------------------------
--------------------------------GCTCA-GGA-CGTAC-GCTG---GC-----G--GC--AC--CGTTAT-ACAT-GCAAG-TCGAA----------------------------------------------
-------------------GGTTTGA-TCATGGCGCA-GAA-CGTAC-GCTG---GC-----G--GC--AC-G-CTAAC-ACAT-GCAAG-TCGAA----------------------------------------------
TTGA-------CT-GAA--GGTTTGA-TCCTGGCTCA-GAA-CGTAC-GCT----GC--------GC--AC-GCCTAAC-ACAT-GCAAG-TCGAA----------------------------------------------
---------------------------------------AG-TGAAC-GCTG---GC-----G--GC--GT-GCTTAAC-ACAT-GCAAG-TCGAA---C---GA-GTAA--G--AGC---C---G-TA----GC-AATAC-
------------A------GGTTTGA-TCATGGCTCA-GGG-TGAAC-GCTG---GC-----G--GC--GT-GCTTAAC-ACAT-GCAAG-TCGAA---C----G-AGT-AGG--AGC---C---G-------TA-GC-AA-
---------------------------TCCTGGCTC----A--GAAC-G---AACGC--TG-GCGGC--AA-GCTTAAC-ACAT-GCAAG-TCGAA---CG--GATTAT-------TC---T-----TT-A--TA-GC----
---------------------------------CTC----A--GAAC-G---AACGC--TG-GCGGC--AA-GCTTAAC-ACAT-GCAAG-TCGAA---CG--GATTAT-------TC---T-----TT-A--TA-GC----
-----------------------------------------------------------TG-GCGGC--AA-GCTTAAC-ACAT-GCAAG-TCGAA---CG--GATTAT-------TC---T-----TT-A--TA-GC----
----------------------------------------------------------------------------------------------------------------------------------------------
----------------------------------------------------------------------------------------------------------------------------------------------
-----------------------------------------------------------TG-GCGGC--AA-GCTTAAC-ACAT-GCAAG-TCGAA---CG--GATTAT-------TC---T-----TT-A--TA-GC----
-----------------------------------------------------------------------------------------G-TCGAA---CG--GATTAT-------TC---T-----TT-A--TA-GC----
---------------------------TCCTGGCTC----A--GAAC-G---AACGC--TG-GCGGC--AA-GCTTAAC-ACAT-GCAAG-TCGAA---CG--GATTAT-------TC---T-----TT-A--TA-GC----
----------------------------------------------------------------------------------------------------------------------------------------------
------------------------------------------------------------------------------C-ACAT-GCAAG-TCGAA---CG--AATTA--------TC---T-----CT-G--TA-GC----
------------------------------------------------------------------------------C-ACAT-GCAAG-TCGAA---CG--AATTA--------TC---T-----CT-G--TA-GC----
--------------------------------------------AAC-G---AACGC--TG-GCGGC--AA-GCTTAAC-ACAT-GCAAG-TCGAA---CG--GACCG--------TA---T-----AC-G--CA-GC----
---------------------------------CTC----A--GAAC-G---AACGC--TG-GCGGC--AA-GCCTAAC-ACAT-GCAAG-TCGAA---CG--GACAAT-------TA---T-----TT-A--TA-GC--C-
--------------------------------------------AAC-G---AACGC--TG-GCGGC--AA-GCCTAAC-ACAT-GCAAG-TCGAA---CG--GACAAT-------TA---T-----TT-A--TA-GC--C-
---------------------------------------------------------------------AA-GCCTAAC-ACAT-GCAAG-TCGAA---CG--GACAAT-------TA---T-----TT-A--TA-GC--C-
---------------------------------CTC----A--GAAC-G---AACGC--TG-GCGGC--AA-GCCTAAC-ACAT-GCAAG-TCGAA---CG--GACAAT-------TG---C-----TT-A--TA-AC-CT-
----------------------------------------------------------------------------------------------------------------------------------------------
-----------------------------------------------------------------GC--AA-GCCTAAC-ACAT-GCAAG-TCGAA---CG--GATAGC-------TA---C-----CC-A--TA-GCTTT-
-----------------------------------------------------------------GC--AA-GCCTAAC-ACAT-GCAAG-TCGAA---CG--GATAGC-------TA---C-----CC-A--TA-GCTTT-
---------------------TTTGA-TCCTGGCTC----A--GAAC-G---AACGC--TG-GCGGC--AA-GCCTAAC-ACAT-GCAAG-TCGAA---CG--GATAGC-------TA---C-----CC-A--TA-GCTTT-
----------------------------------------------------------------------------------------------------------------------------------------------
TTGA-------GA-------GTTTGA-TCCTGGCTC----A--GAAC-G---AACGC--TG-GCGGC--AA-GCTTAAC-ACAT-GCAAG-TCGAA---CG--GACAGT-------TA---T-----TT-A--TA-GC----
-------------------------------------------------------------------------------------------------------------------------------TT-G--TA-GC----
-------------------------------------------------------------------------------------------------------------------------------TT-G--TA-GC----
--------------------------------GCTC----A--GAAT-G---AACGC--TG-GCGGC--AG-GCCTAAC-ACAT-GCAAG-TCGAA---CG--GACGT-T------AT---A-----TT-G--TA-GC----
----------------------------------------------------------------------------------------------------------------------------------------------
------------A------GGTTTGA-TCATGGCTC---AGAT-------TGAACGC--TG-GCGGT--AT-GCTTAAC-ACAT-GCAAG-TCGAA------CGG---T-AGC---AG-------GC-C-------------
------------A------GGTTTGA-TCATGGCTCA-GA-TTGAAC-GCTGG----------CGGT--AT-GCTTAAC-ACAT-GCAAG-TCGAA------CGG---T-AGC---AG-------GC-C-------------
------------A------GGTTTGA-TCATGGCTC---AGAT-------TGAACGC--TA-GCGGT--AT-GCTTAAC-ACAT-GCAAG-TCGAA------CGG---T-AGC---AG-------GC-C-------------
------------A------GGTTTGA-TCCTGGCTC---AGAT-------TGAACGC--TG-GCGGT--AT-GCTTAAC-ACAT-GCAAG-TCGAA------CGG---T-AAC---AG-------GC-C-------------
------------A------GGTTTGA-TCATGGCTC---AGAT-------TGAACGC--TG-GCGGT--AT-GCTTAAC-ACAT-GCAAG-TCGAA------CGGTAA----C---AG-------GC-C-------------
------------A------GGTTTGA-TCATGGCTCA-GA-TTGAAC-GCTGG----------CGGT--AT-GCTTAAC-ACAT-GCAAG-TC----------------GAAC---GG----------TAA-------C-A-
------------A------GGTTTGA-TCATGGCTCA-GA-TTGAAC-GCTGG----------CGGT--AT-GCTTAAC-ACAT-GCAAG-TC----------------GAAC---GG----------TAA-------C-A-
------------A------GGTTTGA-TCATGGCTCA-GA-TTGAAC-GCTGG----------CGGT--AT-GCTTAAC-ACAT-GCAAG-TC----------------GGAC---GG----------TAA-------C-A-
```

Discussion / Improvements

- Placement method: greedy algorithm. Errors are definitive, and generate other errors.
  - ➢ Local recomputing algorithm currently in test.

- Alignment of large blocks of sequences (several thousands) too complex.
  - ➢ Find a small set of representative sequences is a difficult problematic.

- Simplify the algorithm for simple cases.
  - ➢ Identify simple cases.
  - ➢ Preserve the alignment quality.