Algorithmes incrémentaux pour la gestion de grandes familles de séquences homologues

Jean-François Dufayard
Guy Perrière
Manolo Gouy
François Rechenmann

HELIX
BIOINFORMATICS

UMR 5558
Biométrie, biologie
évolutive.

INRIA
RHÔNE-ALPES

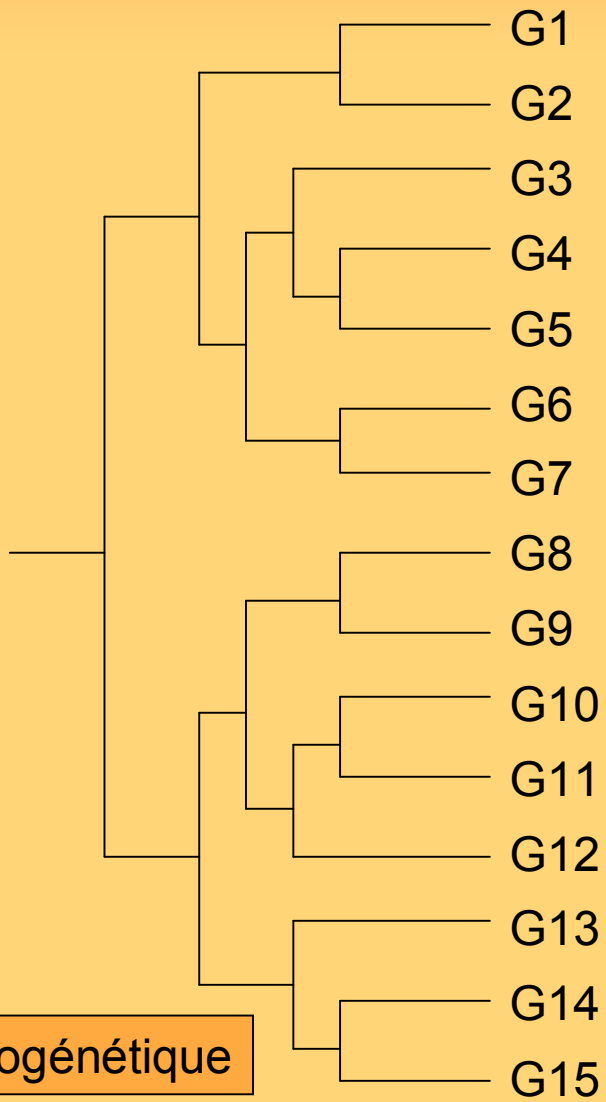
Problématique (1/2)

Génomique comparative

Analyses phylogénétiques

```
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
```

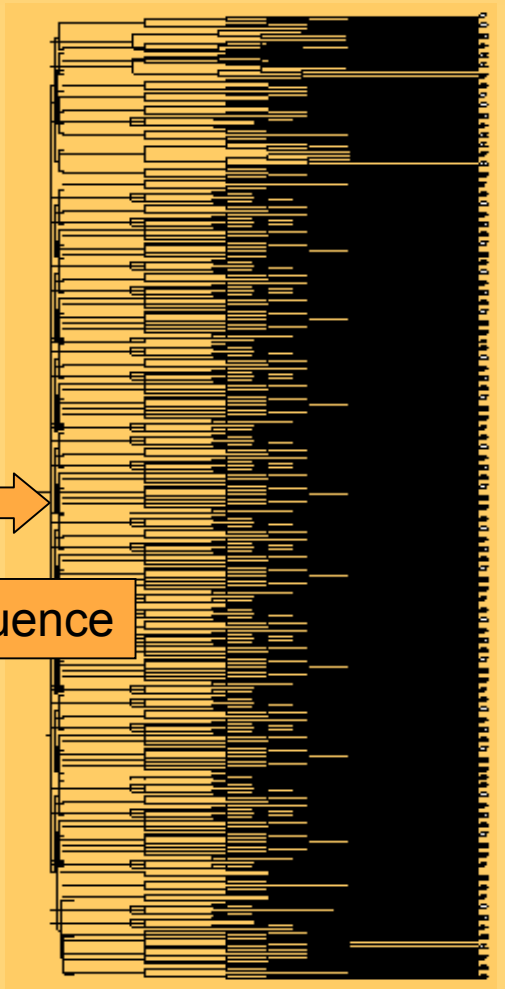
Alignement multiple



Arbre phylogénétique

Problématique (2/2)

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-ATNSKSNHCLDAAKA
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



Nouvelle séquence

Algorithmes incrémentaux

Méthodes (1/2)

Deux méthodes de référence:

- *Neighbor-Joining* (avec distances évolutives 2-Kimura, ou Poisson)
- Méthode progressive d'alignement (utilisation d'un arbre guide)

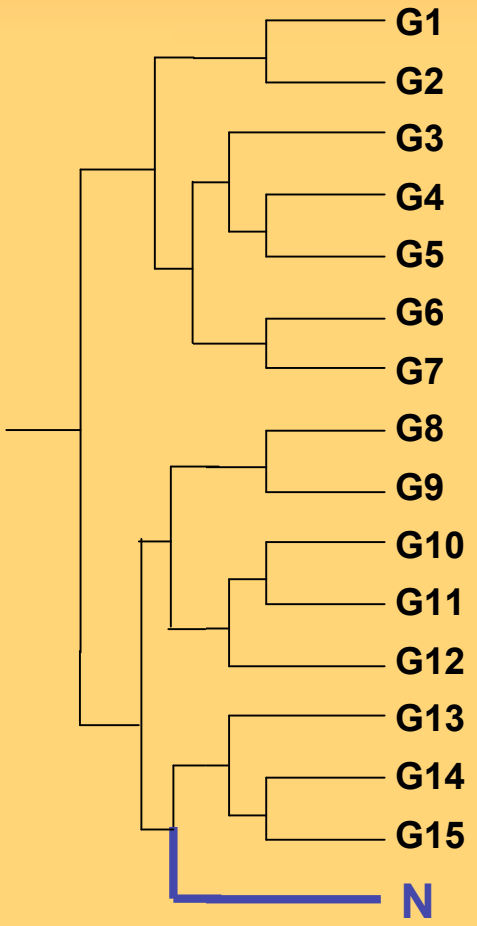
-« Un alignement de bonne qualité nécessite un arbre guide proche de l'arbre phylogénétique de la famille. »

-« Un arbre phylogénétique de bonne qualité nécessite un alignement multiple de bonne qualité. »

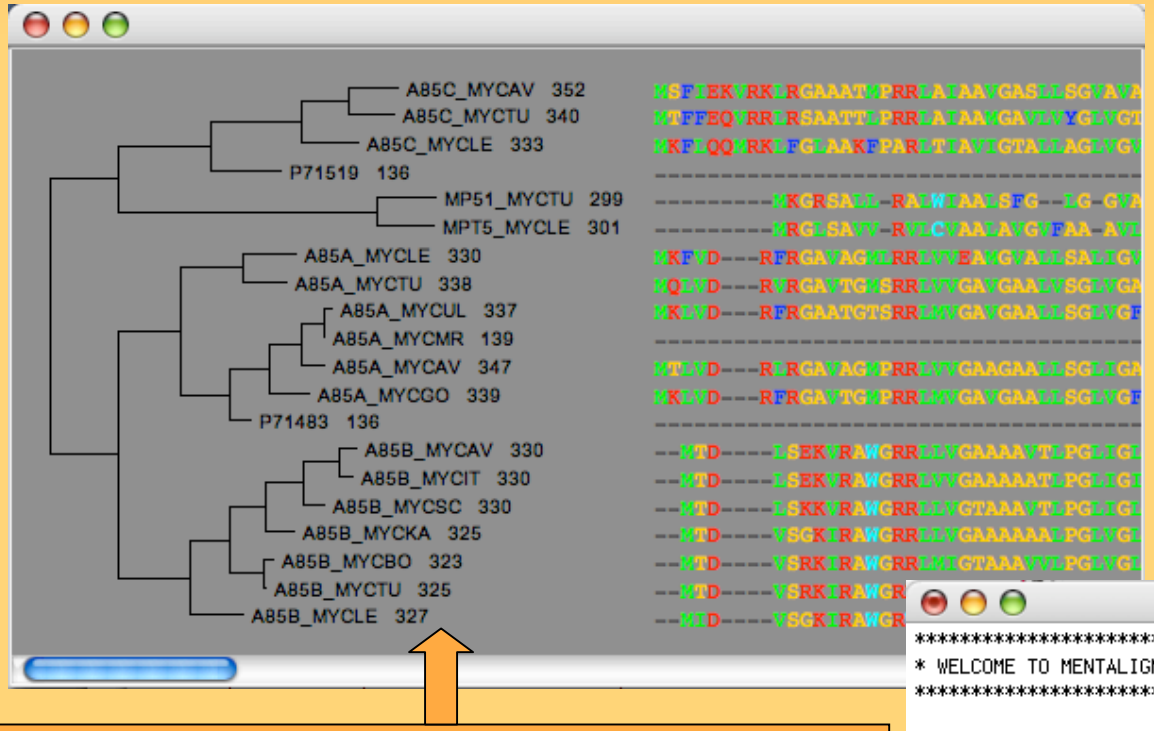
Traiter l'arbre phylogénétique et l'alignement en même temps.

Méthodes (2/2)

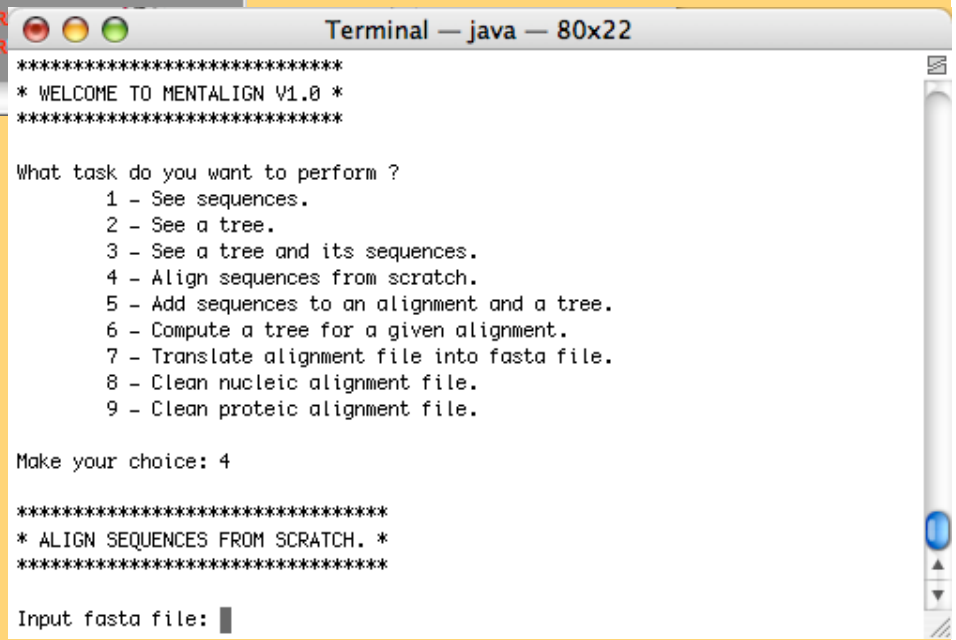
```
L-QGNDLLEDSPYEPVNSRLSDIFRAVPFISDVFQQVEHI-SKGNNCLDAAKACNLDDTC
L-QGNDLLEDSPYEPVNSRLSDIFRAVPFISDVFQQVEHI-SKGNNCLDAAKACNLDDTC
LTEGEEFYEASPYEPVTSRLSDIFRLASIFSGTGADP-VVSAKSNHCLDAAKACNLNDNC
LTEGEEFYEASPYEPVTSRLSDIFRLASIFSGTGADP-VVSAKSNHCLDAAKACNLNDNC
L-QGNDLLEDSPYEPVNSRLSDIFRLAPIVS-----VEPVLSKGNNCLDAAKACNLNDTC
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
LTEGEEFYEASPYEPVT------FRL---FSGTGTDPA-VSTKSNHCLDAAKACNLNDNC
```



Implémentation dans « Mentalign »



Famille des antigènes 85-A, 85-B et 85-C, issue de la banque HOBACGEN.



Conclusion

On dispose d'algorithmes incrémentaux capables d'ajouter une ou plusieurs séquences à un arbre et un alignement.

Applications potentielles:

- Gérer de grandes familles de séquences homologues:
  - ARN Ribosomaux
  - Cytochromes B
  - …
- Gérer de grands recueils de familles de séquences homologues:
  - HOVERGEN
  - HOBACGEN
  - …

ARTICLE N°64