

Algorithmes incrémentaux pour la gestion de grandes familles de séquences homologues

Jean-François Dufayard
Guy Perrière
Manolo Gouy
François Rechenmann



Problématique (1/2)

Génomique comparative

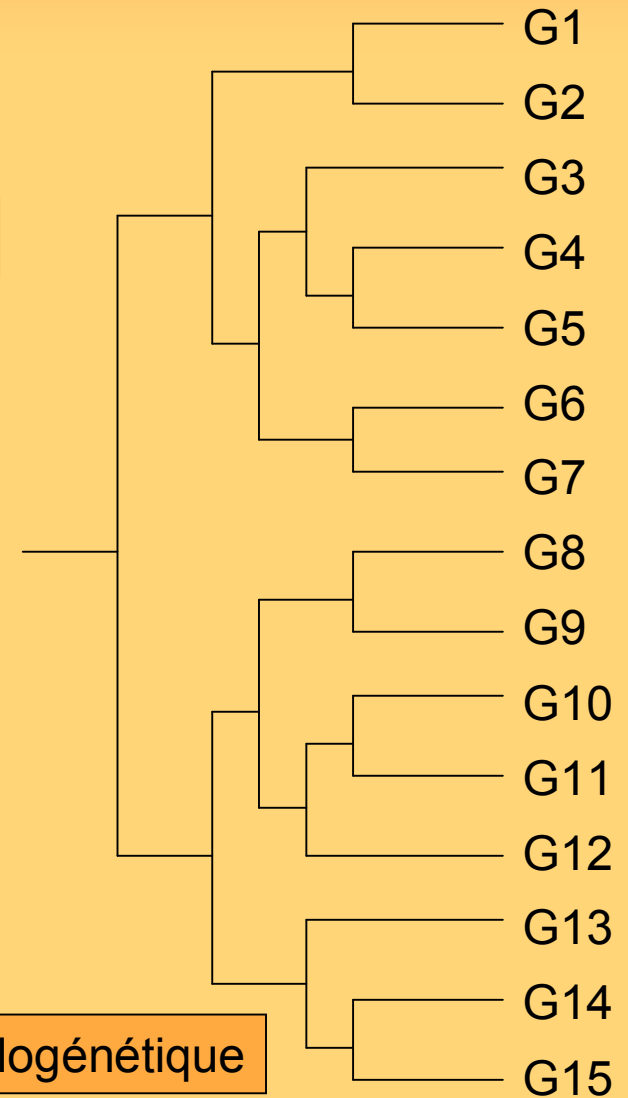


Analyses phylogénétiques

```
L-QGNDLLEDSPYEPVNSRLSDIFRAVPFISDVFQQVEHI-SKGNNCLDAAKACNLDDTC
L-QGNDLLEDSPYEPVNSRLSDIFRAVPFISDVFQQVEHI-SKGNNCLDAAKACNLDDTC
LTEGEEFYEASPYEPVTSRLSDIFRLASIFSGTGADP-VVSAKSNHCLDAAKACNLNDNC
LTEGEEFYEASPYEPVTSRLSDIFRLASIFSGTGADP-VVSAKSNHCLDAAKACNLNDNC
L-QGNDLLEDSPYEPVNSRLSDIFRLAPIVS-----VEPVLSKGNNCLDAAKACNLNDTC
LAEGEEFYEASPYEPITSRLSDIFRLASIFSGM--DP-ATNSKSNHCLDAAKACNLNDNC
LTEGEEFYEASPYEPVTSRLSDIFRLASIFSGTGTDPAVSTKSNHCLDAAKACNLNDNC
L-QGNDLLEDSPYEPVNSRLSDIFRAVPFIS-----VEHI-SKGNNCLDAAKACNLDDTC
L-QGNDLLEDSPYEPVNSRLSDIFRVVPFIS-----VEHI-PKGNNCLDAAKACNLDDIC
L-QGNDLLEDSPYEPVNSRLSDIFRVVPFISDVFQQVEHI-PKGNNCLDAAKACNLDDIC
L-QGNDLLEDSPYEPVNSRLSDIFRAVPFIS-----VEHI-SKGNNCLDAAKACNLDDTC
LTEGEEFYEASPYEPVTSRLSDIFRLASIFSGTGTDPAVSTKSNHCLDAAKACNLNDNC
LMEGMNVLESPYEPFIRGF-DYVRLASITAGSENEVTQV----NRCLDAAKACNVDEM
---GEEFYEASPYEPVTSRLSDIFRLASIFSGTGADP-VVSAESNHCLDAAKACNLNDNC
---G-----TGADP-VVSAESNHCLDAAKACNLNDNC
```

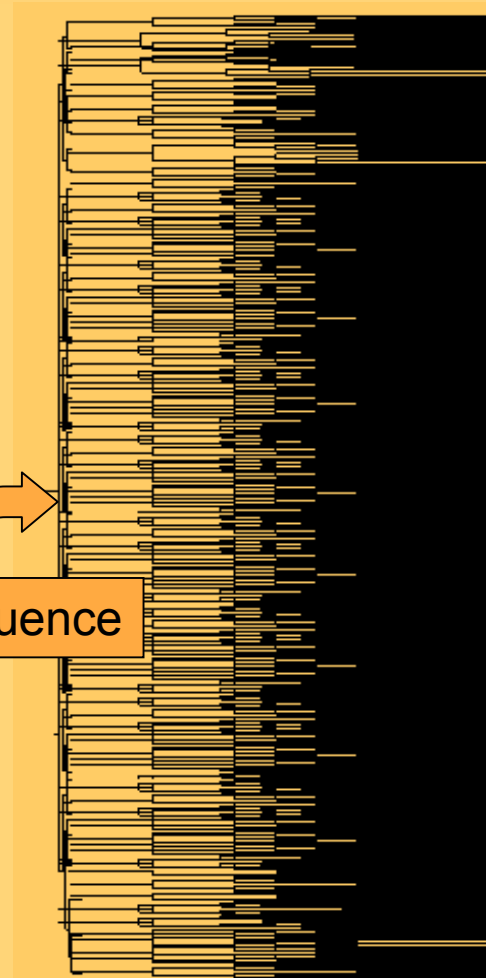
Alignement multiple

Arbre phylogénétique



Problématique (2/2)

L-QGNDLLEDSPYEPVNSRSLSDIFRAVPPFISDVFQQVEHI-SKGNNCLDAAKACNLDDTC
L-QGNDLLEDSPYEPVNSRSLSDIFRAVPPFISDVFQQVEHI-SKGNNCLDAAKACNLDDTC
L-TEGEEFYEASPYEPVTSRSLSDIFRLASIFSGTGADP-VVSAKSNHCLDAAKACNLNDNC
L-TEGEEFYEASPYEPVTSRSLSDIFRLASIFSGTGADP-VVSAKSNHCLDAAKACNLNDNC
L-QGNDLLEDSPYEPVNSRSLSDIFRLAPIVS-----VEPVL SKGNNCLDAAKACNLNDTC
L-AEGEEFYEASPYEPITSRSLSDIFRLASIFSGM--DP-ATNSKSNHCLDAAKACNLNDNC
L-TEGEEFYEASPYEPVTSRSLSDIFRLASIFSGTGADP-AVSTKSNHCLDAAKACNLNDNC
L-QGNDLLEDSPYEPVNSRSLSDIFRAVPPFIS-----VEHI-SKGNNCLDAAKACNLDDTC
L-QGNDLLEDSPYEPVNSRSLSDIFRVVPPFIS-----VEHI-PKGNNCLDAAKACNLDDIC
L-QGNDLLEDSPYEPVNSRSLSDIFRVVPPFISDVFQQVEHI-PKGNNCLDAAKACNLDDIC
L-QGNDLLEDSPYEPVNSRSLSDIFRAVPPFIS-----VEHI-SKGNNCLDAAKACNLDDTC
L-TEGEEFYEASPYEPVTSRSLSDIFRLASIFSGTGADP-AVSTKSNHCLDAAKACNLNDNC
L-MEGMNVLESSPYEPFIRGF-DYVRLASITAGSENEVTQV----NRCLDAAKACNVDEMC
---G-----TGADP-VVSAESNHCLDAAKACNLNDNC
L-QGNDLLEDSPYEPVNSRSLSDIFRAVPPFISDVFQQVEHI-SKGNNCLDAAKACNLDDTC
L-QGNDLLEDSPYEPVNSRSLSDIFRAVPPFISDVFQQVEHI-SKGNNCLDAAKACNLDDTC
L-TEGEEFYEASPYEPVTSRSLSDIFRLASIFSGTGADP-VVSAKSNHCLDAAKACNLNDNC
L-TEGEEFYEASPYEPVTSRSLSDIFRLASIFSGTGADP-VVSAKSNHCLDAAKACNLNDNC
L-QGNDLLEDSPYEPVNSRSLSDIFRLAPIVS-----VEPVL SKGNNCLDAAKACNLNDTC
L-AEGEEFYEASPYEPITSRSLSDIFRLASIFSGM--DP-ATNSKSNHCLDAAKACNLNDNC
L-TEGEEFYEASPYEPVTSRSLSDIFRLASIFSGTGADP-AVSTKSNHCLDAAKACNLNDNC
L-QGNDLLEDSPYEPVNSRSLSDIFRAVPPFIS-----VEHI-SKGNNCLDAAKACNLDDTC
L-QGNDLLEDSPYEPVNSRSLSDIFRVVPPFIS-----VEHI-PKGNNCLDAAKACNLDDIC
L-QGNDLLEDSPYEPVNSRSLSDIFRVVPPFISDVFQQVEHI-PKGNNCLDAAKACNLDDIC
L-QGNDLLEDSPYEPVNSRSLSDIFRAVPPFIS-----VEHI-SKGNNCLDAAKACNLDDTC
L-TEGEEFYEASPYEPVTSRSLSDIFRLASIFSGTGADP-AVSTKSNHCLDAAKACNLNDNC
L-MEGMNVLESSPYEPFIRGF-DYVRLASITAGSENEVTQV----NRCLDAAKACNVDEMC
---G-----TGADP-VVSAESNHCLDAAKACNLNDNC
L-QGNDLLEDSPYEPVNSRSLSDIFRAVPPFISDVFQQVEHI-SKGNNCLDAAKACNLDDTC
L-QGNDLLEDSPYEPVNSRSLSDIFRAVPPFISDVFQQVEHI-SKGNNCLDAAKACNLDDTC
L-TEGEEFYEASPYEPVTSRSLSDIFRLASIFSGTGADP-VVSAKSNHCLDAAKACNLNDNC
L-TEGEEFYEASPYEPVTSRSLSDIFRLASIFSGTGADP-VVSAKSNHCLDAAKACNLNDNC
L-QGNDLLEDSPYEPVNSRSLSDIFRLAPIVS-----VEPVL SKGNNCLDAAKACNLNDTC



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. »

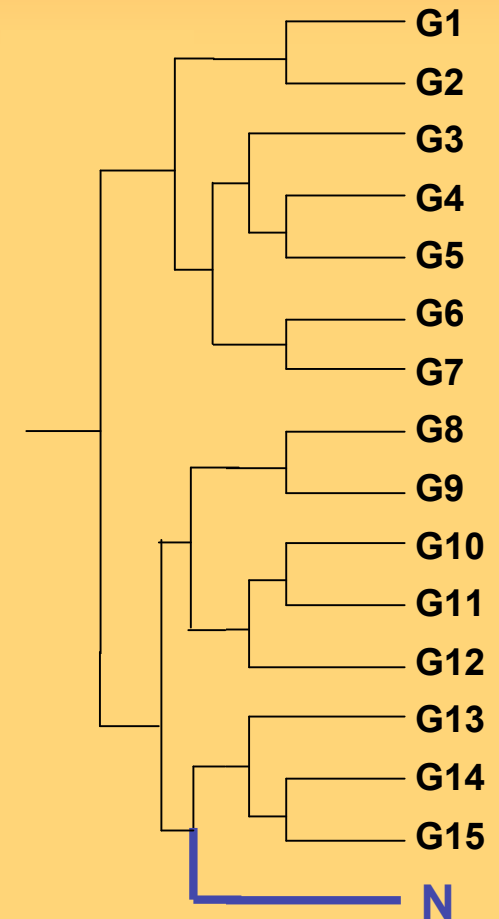
-« Un arbre phylogénétique de bonne qualité nécessite un alignement multiple de bonne qualité. »

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

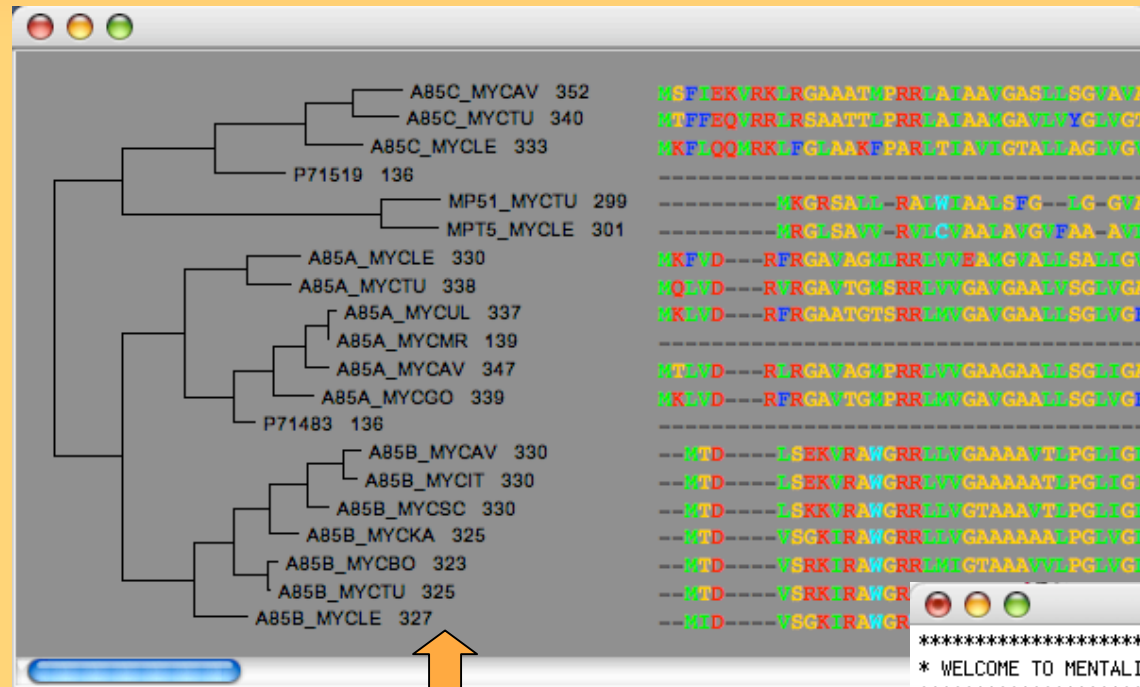


Méthodes (2/2)

L-QGNDLLEDSPYEPVNSRLSDIFRAVVPFISDVFQQVEHI-SKGNNCLDAAKACNLDDTC
L-QGNDLLEDSPYEPVNSRLSDIFRAVVPFISDVFQQVEHI-SKGNNCLDAAKACNLDDTC
LTEGEEFYEASPYEPVTSRLSDIFRLASIFSGTGADP-VVSAKSNHCLDAAKACNLNDNC
LTEGEEFYEASPYEPVTSRLSDIFRLASIFSGTGADP-VVSAKSNHCLDAAKACNLNDNC
L-QGNDLLEDSPYEPVNSRLSDIFRLAPIVS-----VEPVL SKGNNCLDAAKACNLNDTC
LAEGEEFYEASPYEPITSRLSDIFRLASIFSGM--DP-ATNSKSNHCLDAAKACNLNDNC
LTEGEEFYEASPYEPVTSRLSDIFRLASIFSGTGTDPA-VSTKSNHCLDAAKACNLNDNC
L-QGNDLLEDSPYEPVNSRLSDIFRAVVPFIS-----VEHI-SKGNNCLDAAKACNLDDTC
L-QGNDLLEDSPYEPVNSRLSDIFRVVPFIS-----VEHI-PKGNNCLDAAKACNLDDIC
L-QGNDLLEDSPYEPVNSRLSDIFRVVPFISDVFQQVEHI-PKGNNCLDAAKACNLDDIC
L-QGNDLLEDSPYEPVNSRLSDIFRAVVPFIS-----VEHI-SKGNNCLDAAKACNLDDTC
LTEGEEFYEASPYEPVTSRLSDIFRLASIFSGTGTDPA-VSTKSNHCLDAAKACNLNDNC
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.

```
Terminal — java — 80x22
*****
* WELCOME TO MENTALIGN V1.0 *
*****

What task do you want to perform ?
 1 - See sequences.
 2 - See a tree.
 3 - See a tree and its sequences.
 4 - Align sequences from scratch.
 5 - Add sequences to an alignment and a tree.
 6 - Compute a tree for a given alignment.
 7 - Translate alignment file into fasta file.
 8 - Clean nucleic alignment file.
 9 - Clean proteic alignment file.

Make your choice: 4

*****
* ALIGN SEQUENCES FROM SCRATCH. *
*****

Input fasta file: |
```

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 Ribosomiaux
 - Cytochromes B
 - ...
- Gérer de grands recueils de familles de séquences homologues:
 - HOVERGEN
 - HOBACGEN
 - ...

ARTICLE N°64

