

Algorithmes incrémentaux pour la gestion de grandes familles de séquences homologues

Jean-François Dufayard
Guy Perrière
Manolo Gouy
François Rechenmann



Problématique (1/2)

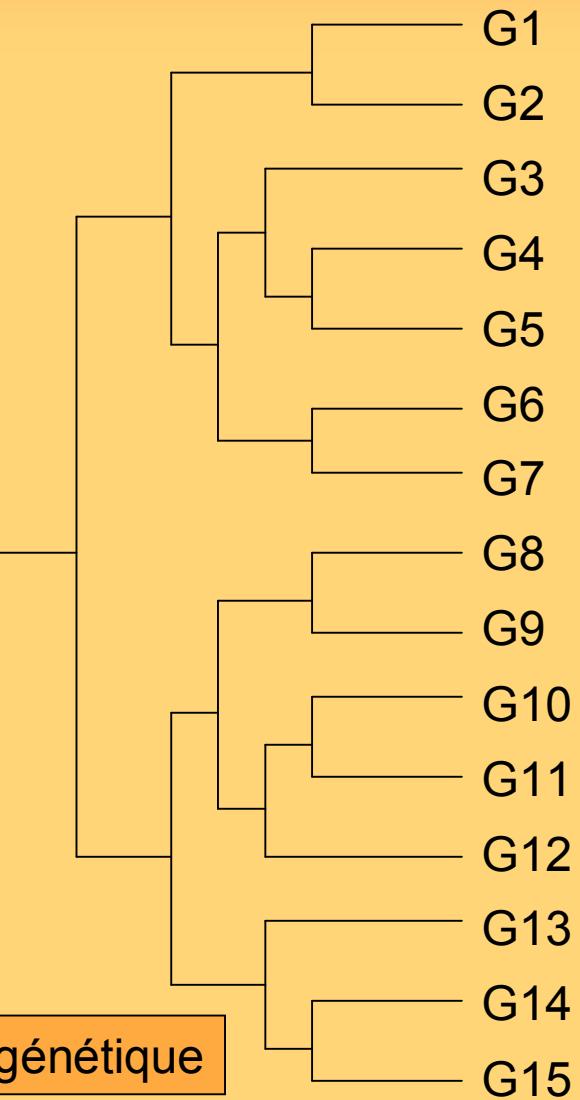
Génomique comparative

Analyses phylogénétiques

```
L-QGNDLLEDSPYEPVNSRLSDIFRAVPFISDVFQQVEHI-SKGNNCLDAAKACNLDDTC  
L-QGNDLLEDSPYEPVNSRLSDIFRAVPFISDVFQQVEHI-SKGNNCLDAAKACNLDDTC  
LTEGEEFYEA  
SPYE  
PVTSRLSDIFRLASIFSGTGADP-VVSAKS  
SNHCLDAAKACNLNDNC  
LTEGEEFYEA  
SPYE  
PVTSRLSDIFRLASIFSGTGADP-VVSAKS  
SNHCLDAAKACNLNDNC  
L-QGNDLLEDSPYEPVNSRLSDIFRLAPIVS-----VEPVLSKGNNCLDAAKACNLNDTC  
LAEGEEFYEA  
SPYE  
PITSRLSDIFRLASIFSGM--DP-ATNSKS  
SNHCLDAAKACNLNDNC  
LTEGEEFYEA  
SPYE  
PVTSRLSDIFRLASIFSGTGTD  
P-AVSTKS  
SNHCLDAAKACNLNDNC  
L-QGNDLLEDSPYEPVNSRLSDIFRAVPFIS-----VEHI-SKGNNCLDAAKACNLDDTC  
L-QGNDLLEDSPYEPVNSRLSDIFRVVPFIS-----VEHI-PKGNNCLDAAKACNLDDIC  
L-QGNDLLEDSPYEPVNSRLSDIFRVVPFISDVFQQVEHI-PKGNNCLDAAKACNLDDIC  
L-QGNDLLEDSPYEPVNSRLSDIFRAVPFIS-----VEHI-SKGNNCLDAAKACNLDDTC  
LTEGEEFYEA  
SPYE  
PVTSRLSDIFRLASIFSGTGTD  
P-AVSTKS  
SNHCLDAAKACNLNDNC  
LMEGMVLESSSPYEPFIRGF-DYVRLASITAGSENEVTQV---NRCLDAAKACNVDEMC  
---GEEFYEA  
SPYE  
PVTSRLSDIFRLASIFSGTGADP-VVSAES  
SNHCLDAAKACNLNDNC  
---TGADP-VVSAES  
SNHCLDAAKACNLNDNC
```

Alignement multiple

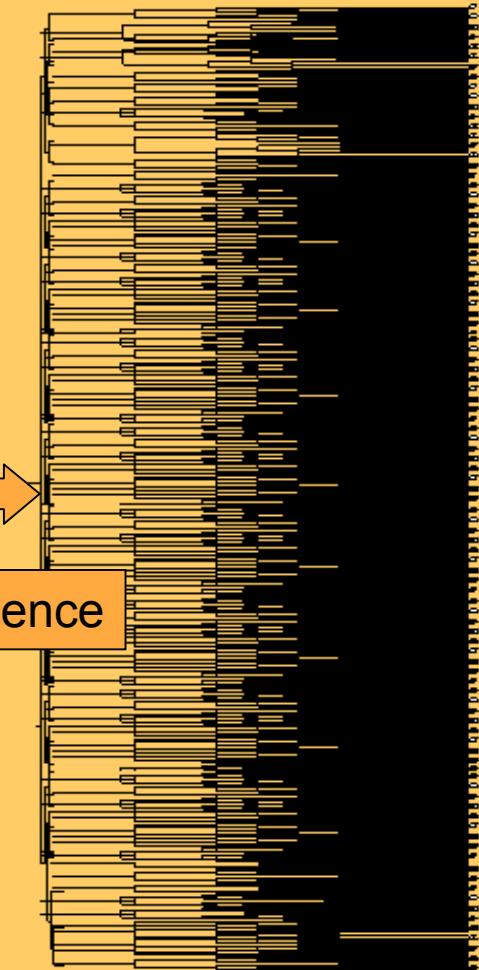
Arbre phylogénétique



Problématique (2/2)

L-QGNDLLEDSPYEPVNSRLSDIFRAVPFISDVFQQVEHI-SKGNNCLDAAKACNLDDTC
L-QGNDLLEDSPYEPVNSRLSDIFRAVPFISDVFQQVEHI-SKGNNCLDAAKACNLDDTC
LTEGEFFYEASPYEPVTSRLSDIFRLASIFSGTGADP-VVSAKSNHCLDAAKACNLNDNC
LTEGEFFYEASPYEPVTSRLSDIFRLASIFSGTGADP-VVSAKSNHCLDAAKACNLNDNC
L-QGNDLLEDSPYEPVNSRLSDIFRLAPIVS-----VEPVLSKGNNCLDAAKACNLNDTC
LAEGEFFYEASPYEPITSRLSDIFRLASIFSGM--DP-ATNSKNHCLDAAKACNLNDNC
LTEGEFFYEASPYEPVTSRLSDIFRLASIFSGTGADP-AVSTSKSNHCLDAAKACNLNDNC
L-QGNDLLEDSPYEPVNSRLSDIFRAVPFIS-----VEHI-SKGNNCLDAAKACNLDDTC
L-QGNDLLEDSPYEPVNSRLSDIFRVVPFIS-----VEHI-PKGNNCLDAAKACNLDDIC
L-QGNDLLEDSPYEPVNSRLSDIFRVVPFISDVFQQVEHI-PKGNNCLDAAKACNLDDIC
L-QGNDLLEDSPYEPVNSRLSDIFRAVPFIS-----VEHI-SKGNNCLDAAKACNLDDTC
LTEGEFFYEASPYEPVTSRLSDIFRLASIFSGTGADP-AVSTSKSNHCLDAAKACNLNDNC
LMEGMNVLESSPYEPFIRGF-DYVRLASITAGSENEVTQV---NRCLDAAKACNVDEMC
---G-----TGADP-VVSAESNHCLDAAKACNLNDNC
L-QGNDLLEDSPYEPVNSRLSDIFRAVPFISDVFQQVEHI-SKGNNCLDAAKACNLDDTC
L-QGNDLLEDSPYEPVNSRLSDIFRAVPFISDVFQQVEHI-SKGNNCLDAAKACNLDDTC
LTEGEFFYEASPYEPVTSRLSDIFRLASIFSGTGADP-VVSAKSNHCLDAAKACNLNDNC
LTEGEFFYEASPYEPVTSRLSDIFRLASIFSGTGADP-VVSAKSNHCLDAAKACNLNDNC
L-QGNDLLEDSPYEPVNSRLSDIFRLAPIVS-----VEPVLSKGNNCLDAAKACNLNDTC
LAEGEFFYEASPYEPITSRLSDIFRLASIFSGM--DP-ATNSKNHCLDAAKACNLNDNC
LTEGEFFYEASPYEPVTSRLSDIFRLASIFSGTGADP-AVSTSKSNHCLDAAKACNLNDNC
L-QGNDLLEDSPYEPVNSRLSDIFRAVPFIS-----VEHI-SKGNNCLDAAKACNLDDTC
L-QGNDLLEDSPYEPVNSRLSDIFRVVPFIS-----VEHI-PKGNNCLDAAKACNLDDIC
L-QGNDLLEDSPYEPVNSRLSDIFRVVPFISDVFQQVEHI-PKGNNCLDAAKACNLDDIC
L-QGNDLLEDSPYEPVNSRLSDIFRAVPFIS-----VEHI-SKGNNCLDAAKACNLDDTC
LTEGEFFYEASPYEPVTSRLSDIFRLASIFSGTGADP-AVSTSKSNHCLDAAKACNLNDNC
LMEGMNVLESSPYEPFIRGF-DYVRLASITAGSENEVTQV---NRCLDAAKACNVDEMC
---G-----TGADP-VVSAESNHCLDAAKACNLNDNC
L-QGNDLLEDSPYEPVNSRLSDIFRAVPFISDVFQQVEHI-SKGNNCLDAAKACNLDDTC
L-QGNDLLEDSPYEPVNSRLSDIFRAVPFISDVFQQVEHI-SKGNNCLDAAKACNLDDTC
LTEGEFFYEASPYEPVTSRLSDIFRLASIFSGTGADP-VVSAKSNHCLDAAKACNLNDNC
LTEGEFFYEASPYEPVTSRLSDIFRLASIFSGTGADP-VVSAKSNHCLDAAKACNLNDNC
L-QGNDLLEDSPYEPVNSRLSDIFRLAPIVS-----VEPVLSKGNNCLDAAKACNLNDTC

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)

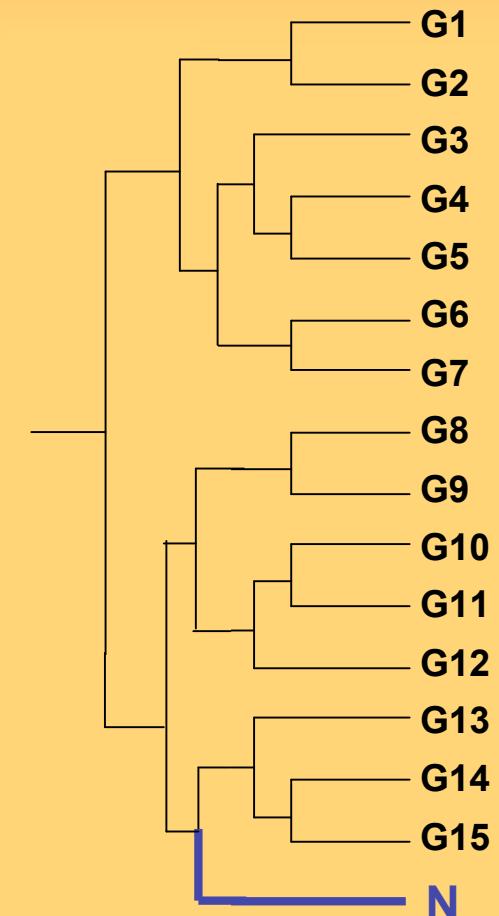
-« Un alignement de bonne qualité nécessite un arbre guide proche de l'arbre phylogénétique de la famille. »

-« 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
LTEGEEFYEAASPYPEVTSRLSDIFRLASIFSGTGADP-VVSAKSNHCLDAAKACNLNDNC
LTEGEEFYEAASPYPEVTSRLSDIFRLASIFSGTGADP-VVSAKSNHCLDAAKACNLNDNC
L-QGNDLLEDSPYEPVNSRLSDIFRLAPIVS-----VEPVLSKGNNCLDAAKACNLNDTC
LAEGEEFYEAASPYPEITSRLSDIFRLASIFSGM--DP-ATNSKSNHCLDAAKACNLNDNC
LTEGEEFYEAASPYPEVTSRLSDIFRLASIFSGTGTDP-AVSTKSNHCLDAAKACNLNDNC
L-QGNDLLEDSPYEPVNSRLSDIFRAVPFIS-----VEHI-SKGNNCLDAAKACNLDDTC
L-QGNDLLEDSPYEPVNSRLSDIFRVVPFIS-----VEHI-PKGNNCLDAAKACNLDDIC
L-QGNDLLEDSPYEPVNSRLSDIFRVVPFISDVFQQVEHI-PKGNNCLDAAKACNLDDIC
L-QGNDLLEDSPYEPVNSRLSDIFRAVPFIS-----VEHI-SKGNNCLDAAKACNLDDTC
LTEGEEFYEAASPYPEVTSRLSDIFRLASIFSGTGTDP-AVSTKSNHCLDAAKACNLNDNC
LMEGMNVLESSPYEPFIRGF-DYVRLASITAGSENEVTQV----NRCLDAAKACNVDEMC
---GEEFYEAASPYPEVTSRLSDIFRLASIFSGTGADP-VVSAESNHCLDAAKACNLNDNC
---G-----TGADP-VVSAESNHCLDAAKACNLNDNC
LTEGEEFYEAASPYPEVTFRL---FSLGTDPA-VSTKSNHCLDAAKACNLNDNC



Implémentation dans « Malign »

The screenshot shows the Mentalign software interface. On the left, a phylogenetic tree displays a family of sequences from the HOBACGEN database. The tree is rooted and branches into three main clades. Sequences include A85C_MYCAV (352), A85C_MYCTU (340), A85C_MYCLE (333), P71519 (136), MP51_MYCTU (299), MPT5_MYCLE (301), A85A_MYCLE (330), A85A_MYCTU (338), A85A_MYCUL (337), A85A_MYCMR (139), A85A_MYCAV (347), A85A_MYCCO (339), P71483 (136), A85B_MYCAV (330), A85B_MYCIT (330), A85B_MYCSC (330), A85B_MYCKA (325), A85B_MYCBO (323), A85B_MYCTU (325), and A85B_MYCLE (327). An orange arrow points from a text box below to the tree. On the right, a sequence alignment is shown with color-coded amino acids. Below the alignment, a terminal window titled "Terminal — java — 80x22" displays the software's command-line interface, including task selection options and a choice of 4, followed by an input field for a FASTA file.

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

```
*****  
* 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 protein 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 Ribosomaux
 - Cytochromes B
 - ...
- Gérer de grands recueils de familles de séquences homologues:
 - HOVERGEN
 - HOBACGEN
 - ...

ARTICLE N°64