

# Alignements multiples

Celine Brochier

[celine.brochier@up.univ-mrs.fr](mailto:celine.brochier@up.univ-mrs.fr)

<http://194.57.197.233:800/>

04.91.10.64.75

# Intérêt des alignements multiples

- Caractérisation des familles multigéniques
- Démontrer l'homologie entre les nouvelles séquences et des familles de séquences préexistantes
- Prédiction de structures secondaires et tertiaires
- Suggestion d'oligonucléotides pour fabriquer des amorces
- Phylogénies

# Alignement de deux séquences

- Programmation dynamique
  - Algorithme de Needleman et Wunsch (1970)
  - Algorithme de Waterman (1976)
- Alignement mathématiquement optimal
  - Table de scores pour les matches et mismatches entre tous les acides aminés ou les nucléotides :
    - Matrices PAM250, BLOSUM62
    - Définition des pénalités pour l'insertion de gaps de tailles différentes

# Alignement de n séquences simultanément

- L'accroissement du nombre de séquences dans les banques de données oblige le développement et l'utilisation de méthodes d'alignements multiples
- Généralisation de la méthode d'alignement de 2 séquences à l'alignement de n séquences simultanément
  - ⇒ Impossible si on traite un nombre de séquences  $>$  à 10
  - ⇒ Besoin d'utiliser des heuristiques  $\Leftrightarrow$  ALIGNEMENT MULTIPLE PROGRESSIF

# L'homologie, base théorique de l'alignement multiple

- Les séquences homologues sont reliées d'un point de vue évolutif
- Idée = construire progressivement un alignement, à partir de séries de séquences (ou de groupes de séquences) alignées deux à deux, suivant un ordre de branchement donné par un arbre phylogénétique
  - Alignement des séquences les plus proches d'un point de vue phylogénétique
  - Intégration progressive des séquences un peu plus éloignées
- Approche suffisamment rapide pour permettre la construction d'alignements contenant un grand nombre de séquences

# Algorithme de CLUSTAL W

- Alignement de toutes les paires de séquences deux à deux par l'algorithme de Needleman et Wunsh
- Construction d'une matrice de distances d'après la divergence mesurée entre chaque paire de séquences
- Calcul d'un arbre guide à partir de la matrice de distances
- Alignement progressif des séquences suivant l'ordre de branchement donné par l'arbre

# Exemple

- Alignement de 7 séquences de globines:
  - Hémoglobine  $\beta$  Humaine (Hbb\_H)
  - Hémoglobine  $\alpha$  Humaine (Hba\_H)
  - Hémoglobine  $\beta$  Cheval (Hbb\_C)
  - Hémoglobine  $\alpha$  Cheval (Hba\_C)
  - Myoglobine de cétacé *Physeter catodon* (Myo)
  - Hémoglobine V de lamproie *Petromyzon marinus* (Glb5)
  - Leghémoglobine II de Lupin (Lgb)

# Alignement des séquences 2 à 2 et construction de la matrice de distances

- Alignement des séquences 2 à 2 par programmation dynamique (algorithme de Needleman et Wunsh) connaissant une matrice de similarité et les pénalité dues aux gaps (ouverture et extension)
- Score = nombre d'identités / nb de résidus comparés (excluant les gaps)
- % de divergence =  $1 - \text{score}$
- Remarque : le calcul du score ne tient pas compte des substitutions multiples, mais on peut utiliser des modèles d'évolution comme Kimura ou JC pour en tenir compte



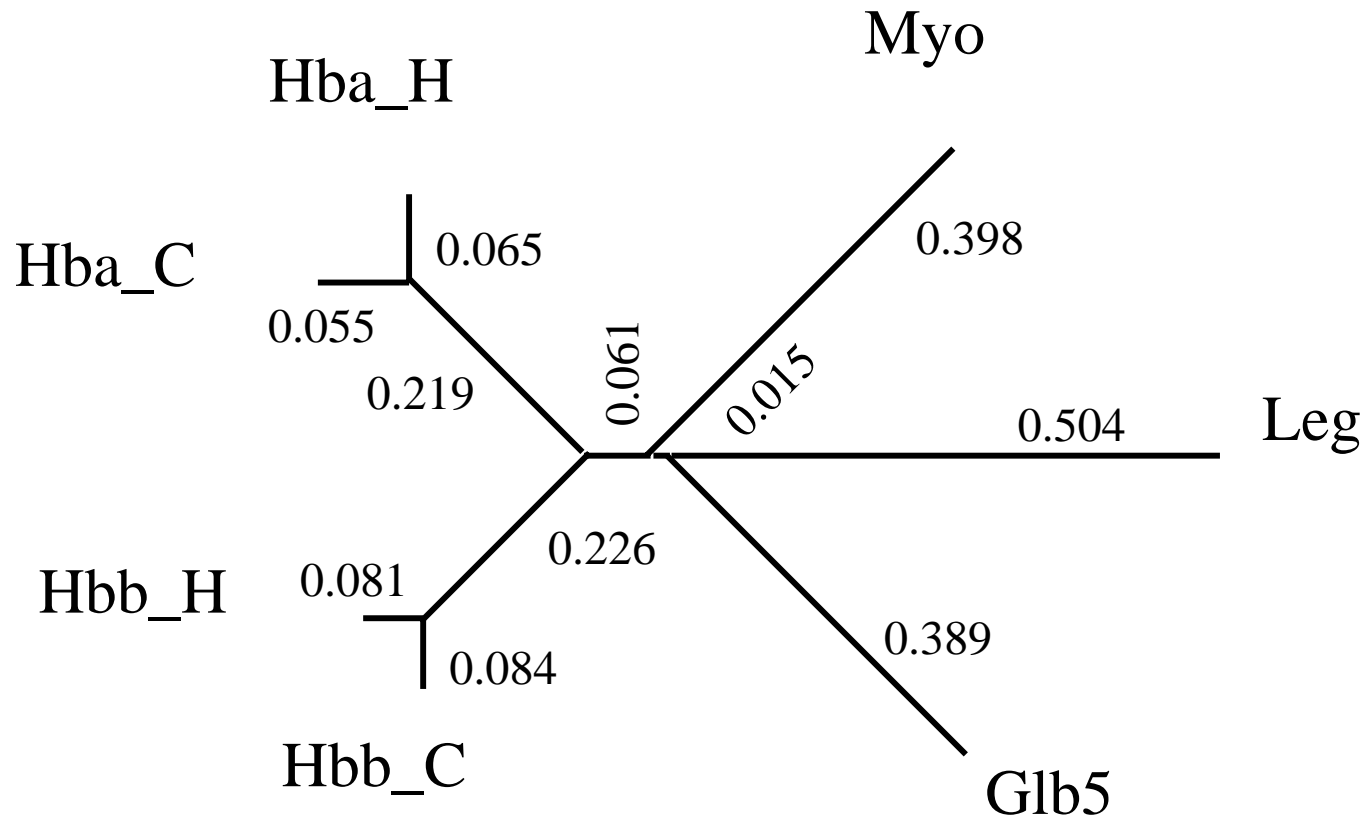
# Alignement des séquences 2 à 2 et construction de la matrice de distances

|       | Hbb_B | Hbb_C | Hba_H | Hba_C | Myo  | Glb5 | Lgb |
|-------|-------|-------|-------|-------|------|------|-----|
| Hbb_H | -     |       |       |       |      |      |     |
| Hbb_C | 0.17  | -     |       |       |      |      |     |
| Hba_H | 0.59  | 0.60  | -     |       |      |      |     |
| Hba_C | 0.59  | 0.59  | 0.13  | -     |      |      |     |
| Myo   | 0.77  | 0.77  | 0.75  | 0.75  | -    |      |     |
| Glb5  | 0.81  | 0.82  | 0.73  | 0.74  | 0.80 | -    |     |
| Lgb   | 0.87  | 0.86  | 0.86  | 0.88  | 0.93 | 0.90 | -   |

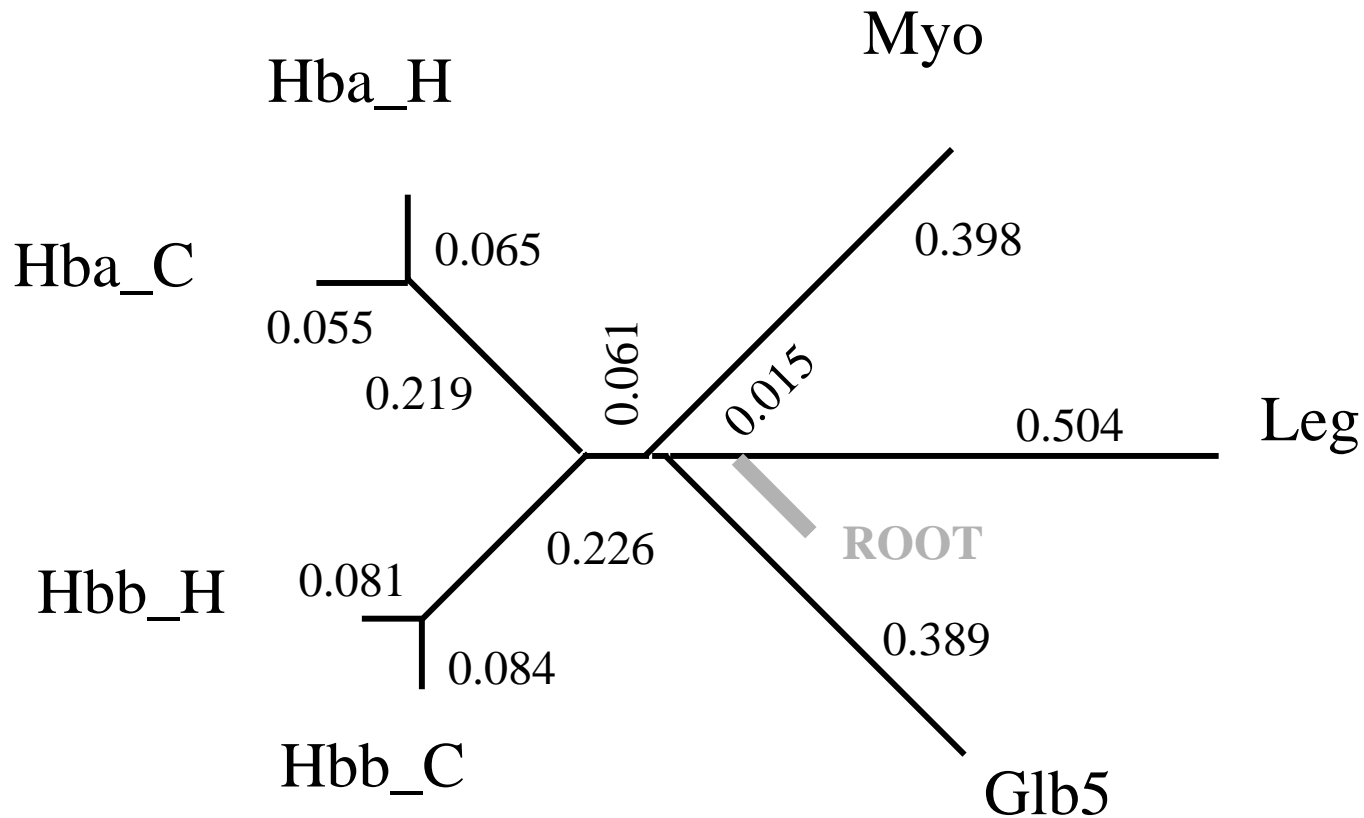
# Construction de l'arbre guide

- Arbre phylogénétique non raciné construit par la méthode du Neighbor-Joining à partir de la matrice de distances calculée précédemment
  - Longueur des branches  $\Leftrightarrow$  proportionnelle à la divergence estimée
  - Racine placée au « poids moyen »  $\Leftrightarrow$  Longueur des branches d'un côté de la racine = longueur des branches de l'autre côté

# Construction de l'arbre guide

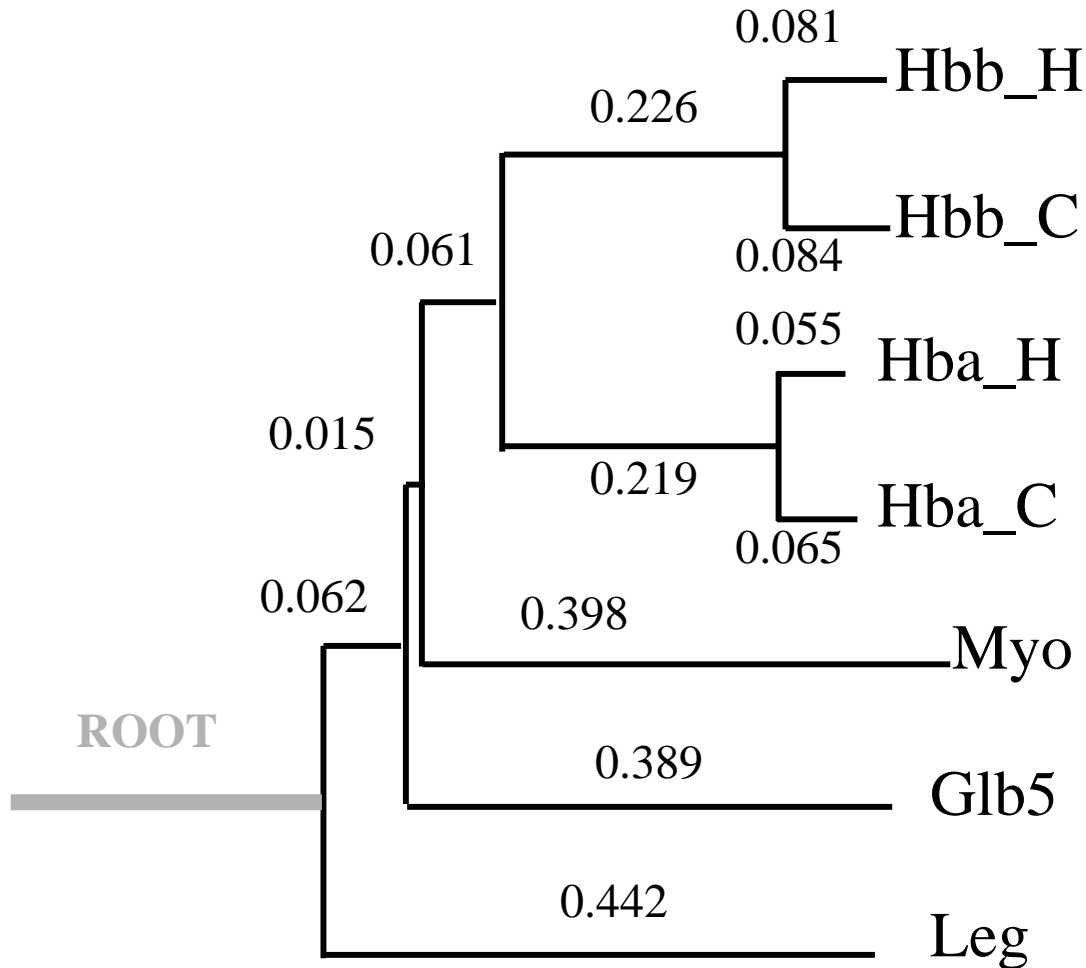


# Placement de la racine



Positionnement de la racine au poids moyen (point à partir duquel les longueurs moyennes des branches de chaque côté du nœud sont égales)

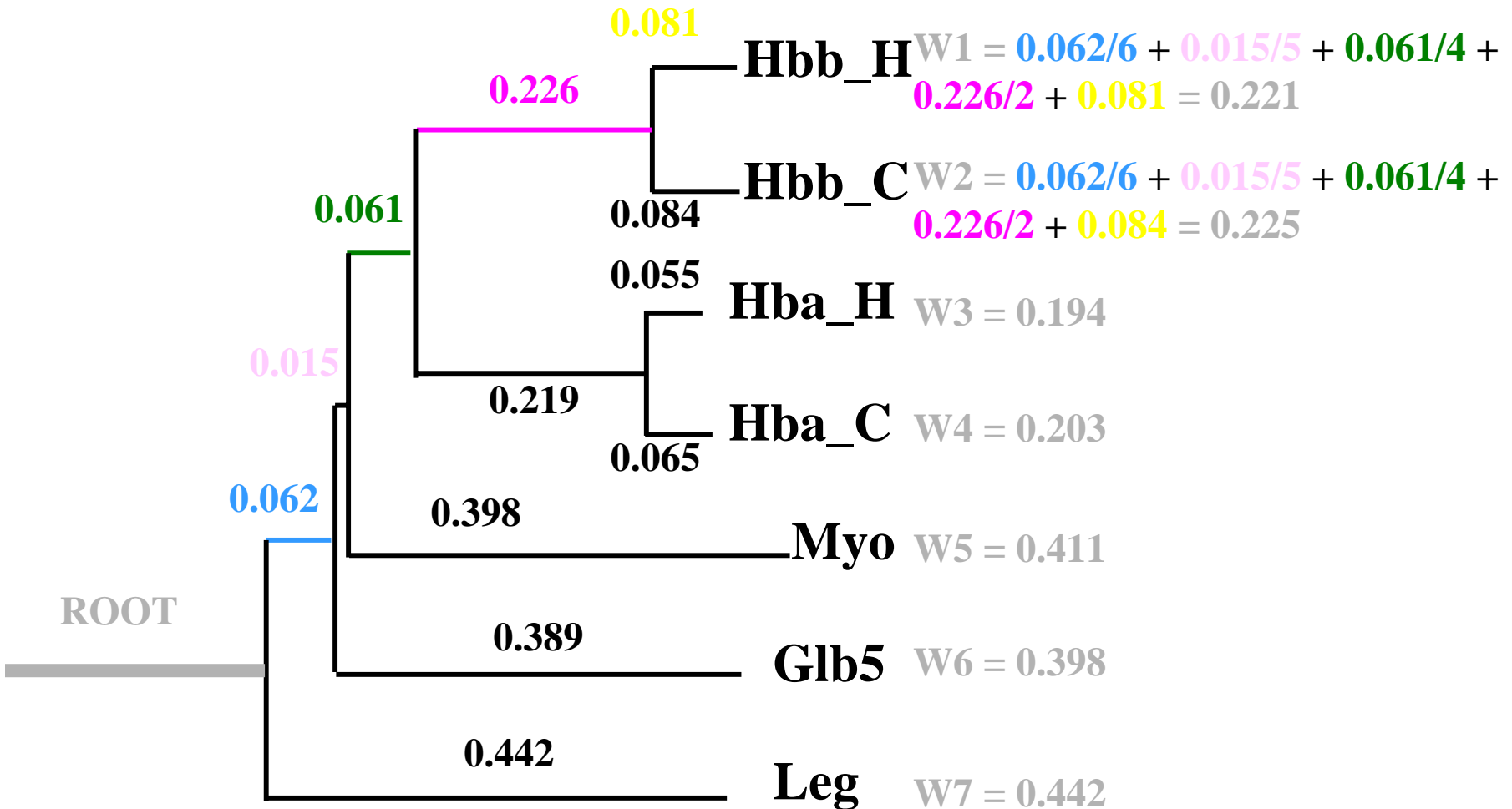
# Arbre guide raciné



# Pondération des séquences

- Principe : attribuer un poids à chaque branche de l'arbre
  - => Dépend de la taille de la branche et du nombre de taxa partageant cette branche (redondance de l'information)
    - => longueur de la branche / nombre de taxa partageant cette branche
- Poids d'une séquence =  $\Sigma$  des longueurs des branches pondérées de la racine au taxon considéré

# Pondération des séquences



# Alignement progressif

- Principe : utiliser une série de paires d'alignements pour aligner des groupes de séquences de plus en plus larges, en respectant l'ordre de branchement dans l'arbre guide (des feuilles vers la racine)



# Alignement progressif

- Dans l'exemple des globines, on aligne dans l'ordre:
  - Les  $\beta$  globines humaines et de cheval
  - Les  $\alpha$  globines humaines et de cheval
  - Les  $\alpha$  et  $\beta$  hémoglobines
  - Les  $\alpha$ ,  $\beta$  hémoglobines et la myoglobine
  - Les hémoglobines, myoglobine et l'hémoglobine de lamproie
  - La leghémoglobine avec toutes les autres

# Alignement progressif

- A chaque étape on utilise un algorithme de programmation dynamique :
  - La matrice de substitution est choisie en fonction de la divergence entre les groupes de séquences qu'on cherche à aligner:
    - 80% < identité < 100% => PAM20
    - 60% < identité < 80% => PAM60
    - 40% < identité < 60% => PAM120
    - 0% < identité < 40% => PAM350
  - Pénalité pour ouverture (GOP) et extension (GEP) des gaps variable selon les séquences et les positions considérées :
    - $GOP = ( GOP + \text{Log}(\min(N,M)) ) \times \text{moyenne des scores des substitutions dans la matrice de substitution utilisée} \times \% \text{ d'identité des séquences} \Leftrightarrow \uparrow$  si les séquences sont très divergentes
    - $GEP = GEP \times (1.0 + |\text{Log}(N/M)| ) \Leftrightarrow$  limite l'introduction de trop grands INDEL dans la plus petite séquence
    - $GOP / 3$  dans les régions hydrophiles (D,E,G,K,N,Q,P,R,S)
    - Si la région présente un INDEL le GOP est  $\uparrow$  :  $GP = GOP \times ( 2 + (( 8 - \text{distance de l'INDEL}) \times 2) / 8 )$
    - Le GOP varie suivant l'a.a. présent

# Alignement progressif

- Calcul du score à une position = moyenne des scores obtenus par toutes les comparaisons 2 à 2 des séquences de chaque groupe pondérés par le poids de chaque séquence

# Alignement progressif

- Exemple: on cherche à aligner un groupe de 4 séquences (déjà alignées) avec un groupe de 2 séquences (déjà alignées)

Calcul du score:

|   |                     |                       |
|---|---------------------|-----------------------|
| 1 | PEEKSAV <b>T</b> AL | M(T, V) x w1 x w5 +   |
| 2 | GEEKA <b>V</b> LAL  | M(T, I) x w1 x w6 +   |
| 3 | PADKTN <b>V</b> KAA | M(L, V) x w2 x w5 +   |
| 4 | AADKTN <b>V</b> KAA | M(L, I) x w2 x w6 +   |
|   |                     | M(K, V) x w3 x w5 +   |
| 5 | EGEWQ <b>L</b> VLHV | M(K, I) x w3 x w6 +   |
| 6 | AAEK <b>T</b> KIRSA | M(K, V) x w4 x w5 +   |
|   |                     | M(K, I) x w4 x w6 / 8 |

Score associé à la comparaison d'un gap = 0  $\Leftrightarrow$  plus mauvais score possible

# Alignement progressif

```
gi | 122615 | sp | P02023 | HBB_HUMAN      -----MVHLTPEEKSAVTALWGKVN--VDEVGGEALGRLLVVYPWTQR
gi | 70401 | pir | | HBHO                  -----VQLSGEEKA AVLALWDKVN--EEEVGGEALGRLLVVYPWTQR
gi | 122412 | sp | P01922 | HBA_HUMAN      -----MVLSPADKTNVKA AWGKVG AHAGEYGA EALERMFLSFPTTKT
gi | 2144717 | pir | | HAHO                -----MVLSAADKTNVKA AWKVG GHAGEYGA EALERMFLGFPTTKT
gi | 127687 | sp | P02185 | MYG_PHYCA   -----VLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLE
gi | 121233 | sp | P02208 | GLB5_PETMA  PIVDTGSVAPLSAAEKT KIRSAWAPVYSTYETSGVDILVKFFTSTPAAQE
gi | 126238 | sp | P02240 | LGB2_LUPLU  -----GALTESQAALVKSSWEEFNANIPKHTRHFFILVLEIAPA AKD
                                     * : : * . : . : * :
gi | 122615 | sp | P02023 | HBB_HUMAN      FFESFGDLSTPD AVMGNPKVKAHGKKVLGAFSDGLAHLDN-----LKGTF
gi | 70401 | pir | | HBHO                  FFDSFGDLSNPGAVMGNPKVKAHGKKVLHSFGEGVHHLDN-----LKGTF
gi | 122412 | sp | P01922 | HBA_HUMAN      YFPHF-DLS-----HGSAQVKGHGKKVADALTNVAHVDD-----MPNAL
gi | 2144717 | pir | | HAHO                YFPHF-DLS-----HGSAQVKAHGKKVGDALTLAVGHLDD-----LPGAL
gi | 127687 | sp | P02185 | MYG_PHYCA   KFDRFKHLKTEAEMKASEDLKKHGVTVLTALGAILKKGH-----HEAEL
gi | 121233 | sp | P02208 | GLB5_PETMA  FFPKFKGLTTADQLKKSADVRWHAERI INAVNDAVASMDDT--EKMSMKL
gi | 126238 | sp | P02240 | LGB2_LUPLU  LFSFLKGTSEVP--QNNPELQAHAGKVFKLVEAAIQLQVTGVVVTDATL
                                     * : . . . : * . : . :
gi | 122615 | sp | P02023 | HBB_HUMAN      ATLSELHCDKLHVDPENFRL LGNVLVCVLAHFFGKEFTPPVQAAYQKVVA
gi | 70401 | pir | | HBHO                  AALSELHCDKLHVDPENFRL LGNVLVVVLARHFGKDFTPELQASYQKVVA
gi | 122412 | sp | P01922 | HBA_HUMAN      SALSDLHAHAKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLA
gi | 2144717 | pir | | HAHO                SNLSDLHAHAKLRVDPVNFKLLSHCLLSTLAVHLPNDFTPAVHASLDKFLS
gi | 127687 | sp | P02185 | MYG_PHYCA   KPLAQSHATKHKIPIKYLEFISEAIIHVLHSRHPGDFGADAQGAMNKALE
gi | 121233 | sp | P02208 | GLB5_PETMA  RDLSGKHAKSFQVDPQYFKVLA AVIADTVAAG-----DAGFEKLMS
gi | 126238 | sp | P02240 | LGB2_LUPLU  KNLGSVHVSKGVAD-AHFPVVK EAILKTIKEVVGAKWSEELNSAWTIAYD
                                     * . * . : . : : . : . . .
gi | 122615 | sp | P02023 | HBB_HUMAN      GVANALAHKYH-----
gi | 70401 | pir | | HBHO                  GVANALAHKYH-----
gi | 122412 | sp | P01922 | HBA_HUMAN      SVSTVLTSKYR-----
gi | 2144717 | pir | | HAHO                SVSTVLTSKYR-----
gi | 127687 | sp | P02185 | MYG_PHYCA   LFRKDIAAKYKELGYQG
gi | 121233 | sp | P02208 | GLB5_PETMA  MICILLRSAY-----
gi | 126238 | sp | P02240 | LGB2_LUPLU  ELAIVIKKEMNDAA---
```

# Qualité des alignements obtenus

- Alignement des séquences proches très rapide et fiable
  - Donne des informations importantes sur la variabilité des positions et la position des gaps
    - La position des gaps introduits pendant la phase précoce de l'alignement ne sont pas changés lors de l'alignement de séquences plus divergentes
  - Sert de canevas pour l'alignement de séquences plus divergentes

# Qualité des alignements obtenus

- Séquences très divergentes < 25-30% d'identité
- Stratégie adoptée par la méthode des alignements progressifs
  - Ordre d'ajout des taxa est primordial ( $\Leftrightarrow$  qualité de l'arbre initial)
  - Ajout progressif des séquences  $\Leftrightarrow$  les erreurs survenant dans les premières étapes ne seront pas corrigées lors de l'ajout de nouvelles séquences
  - Choix des paramètres d'alignement (matrice de substitution, pénalités d'ouverture et de fermeture des gaps doivent être adaptés)
- Aucune garantie que la solution finale soit proche de la solution exacte

# Ressources

- ClustalW en ligne
  - [http://www.infobiogen.fr/services/analyseseq/cgi-bin/clustalw\\_in.pl](http://www.infobiogen.fr/services/analyseseq/cgi-bin/clustalw_in.pl)
- ClustalW download (unix, mac, dos...)
  - <http://www.es.embnet.org/Services/ftp/software/ebi/>