Prédiction de sites d'interaction des protéines par analyse d'arbres phylogénétiques

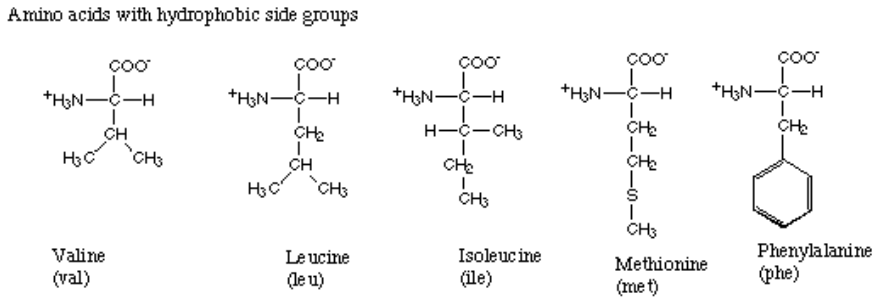
Stéfan Engelen

Génomique Analytique, INSERM U511

Université Pierre et Marie Curie

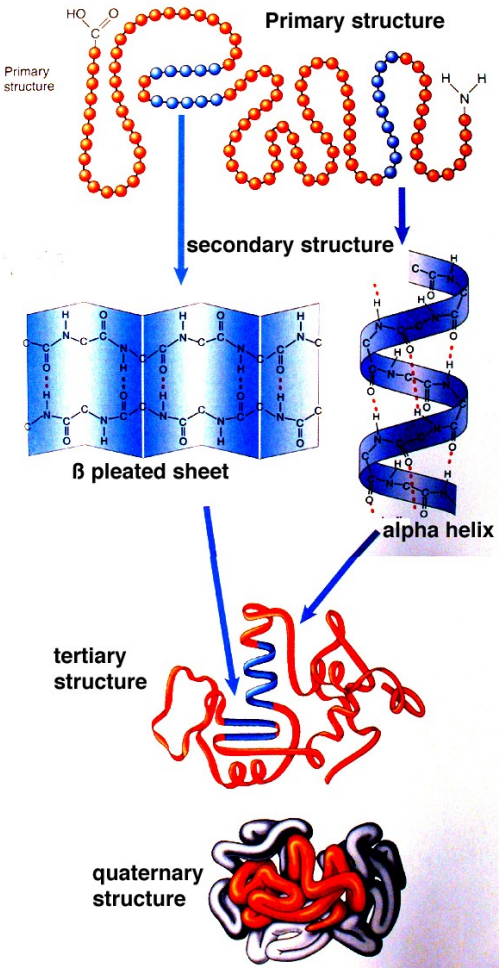
Les protéines

- Suite linéaire d'acides aminés représentés par des lettres … *LNSVEFSSFECPSARGFHM…*
- 20 acides aminés différents



- Longueur typique aux alentours de 300 AA, intervalle de 100 à 5000 AA
- Responsables de la plupart des fonctions d'une cellule :
  - transport de molécules (transporteur)
  - coupure de molécules, protéines (enzyme) …

Structure des protéines



Chaîne d'acides-aminés (1D)

Chaîne d'éléments structuraux réguliers (2D)

Structure 3D d'une chaîne d'acides-aminés.

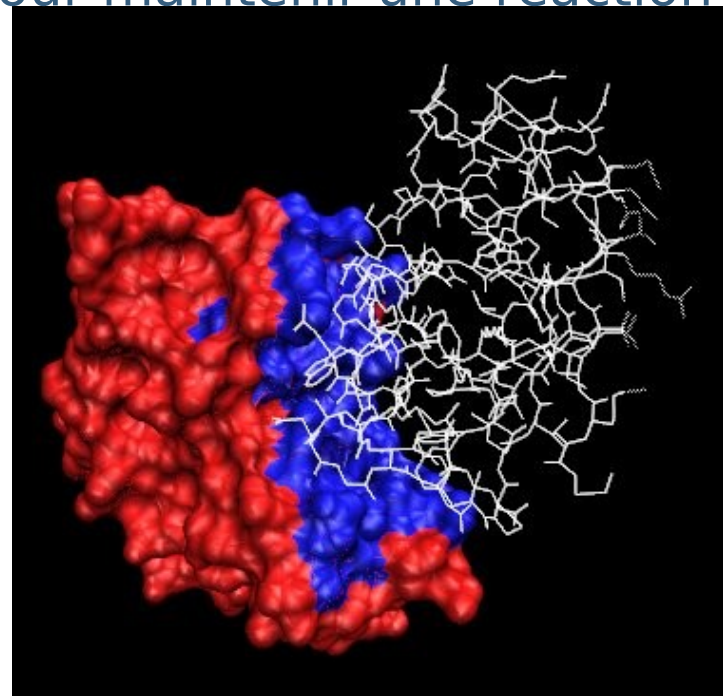
Structure 3D de plusieurs chaînes d'acides-aminés.

Structure des protéines

- Une protéine se replie dans une structure unique, qui est dépendante seulement de la séquence (C. Anfinsen, 1973).
- Protéines homologues :
  - Séquences d'organismes différents ayant divergées au cours de l'évolution à partir d'un ancêtre commun par substitution, délétion ou insertion d'AA
  - Les structures des protéines homologues sont plus conservées au niveau de la structure 3D que des séquences
  - Fonction quasi identique
- Cœur (AA non accessibles) des protéines homologues assez conservé

Les protéines

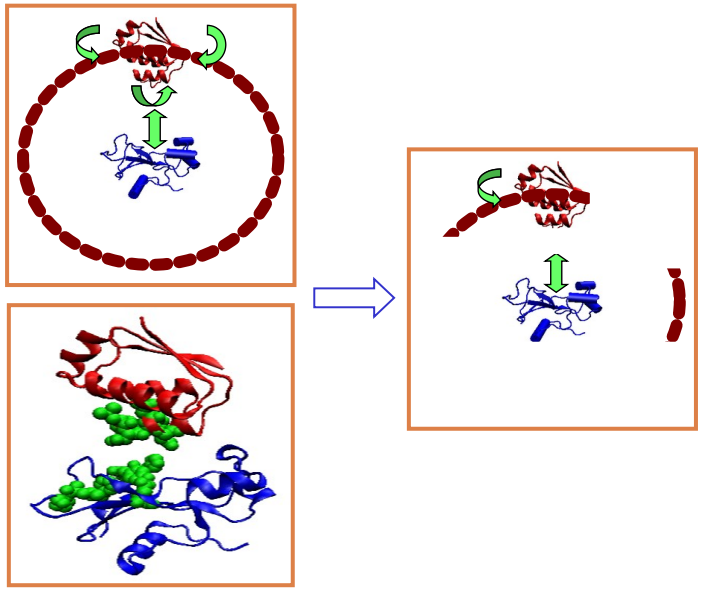
- Surface moins conservée avec des régions fonctionnelles conservées
  - En structure pour maintenir des interactions (emboîtement) avec d'autres composants moléculaires
  - En séquence pour maintenir une réaction moléculaire particulière



Projet global

Intégration des données d'évolution
JET au docking moléculaire (MAXDO)

MAXDO seul :
150 protéines
22500 interactions
1 machine :14 siècles
Grille WCG : 7 mois
1000 à 5000 internautes)

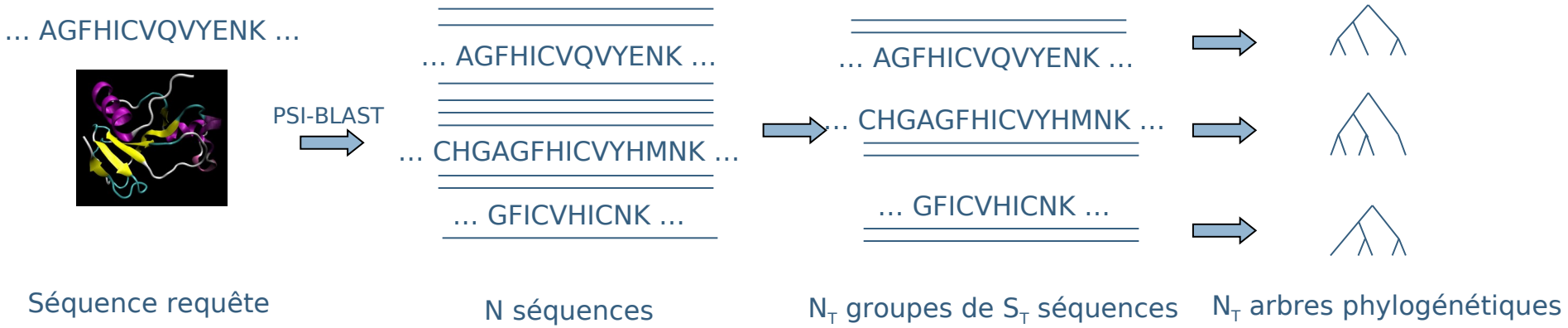


JET + MAXDO
4000 protéines
16 000 000
interactions

Réduction de l'espace des calculs
passage à une échelle plus grande possible

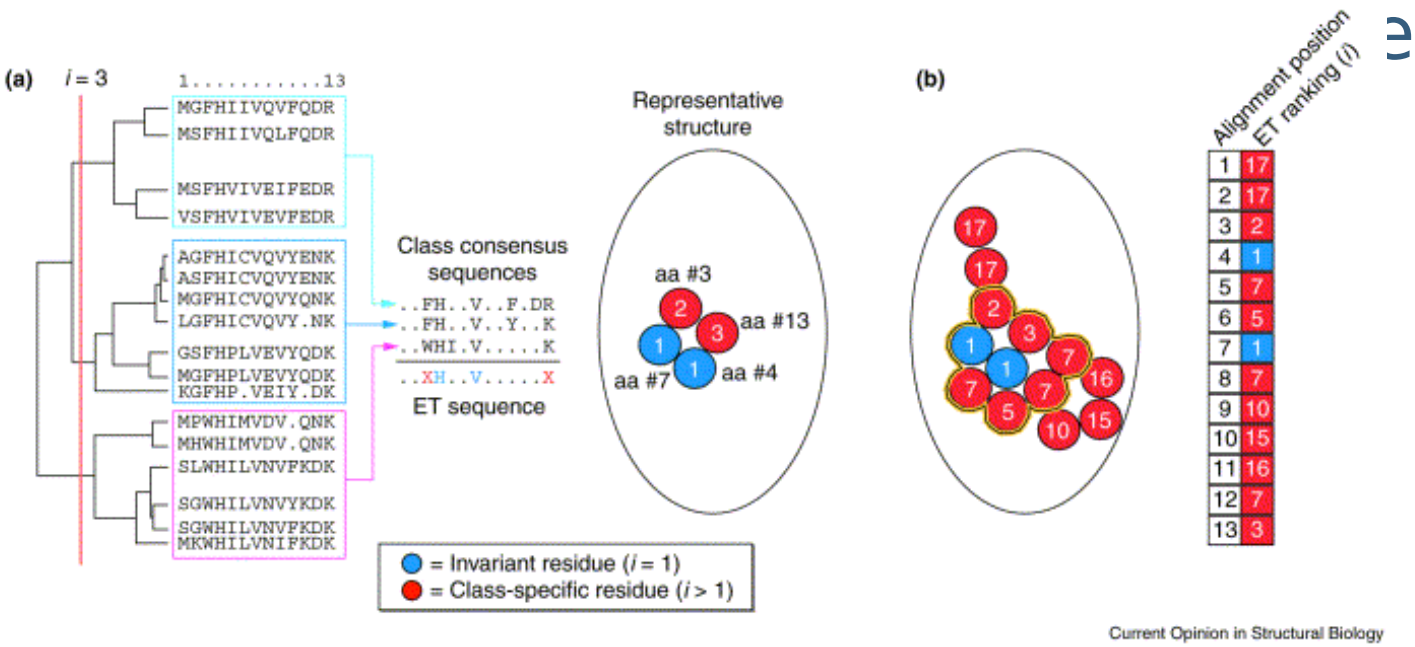
JET : Joint Evolutionary trees

- Séquence requête possédant une structure 3D connue
- Recherche des homologues : PSI-BLAST
- Échantillonnage aléatoire des séquences
  - N_T groupes de S_T séquences
- N_T Alignement multiple : CLUSTALW
- Construction de N_T arbres phylogénétiques : NJ (Neighbor Joining)



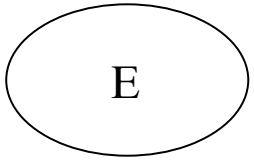
JET : Joint Evolutionary trees

- Évaluation de la conservation des résidus (acides aminés) au sein de chaque arbre : calcul des traces
- Clusterisation des résidus les plus c e
  re



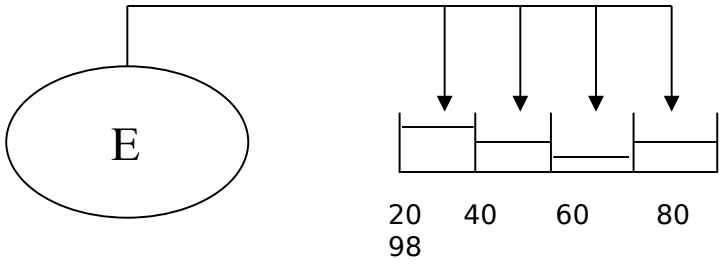
Échantillonnage aléatoire des séquences

- Motivation : ensemble E de séquences obtenues par PSI-BLAST
  - Répartition non uniforme en terme d'identité
  - Familles de séquences sous ou sur représentées
  - Nombre de séquences grand



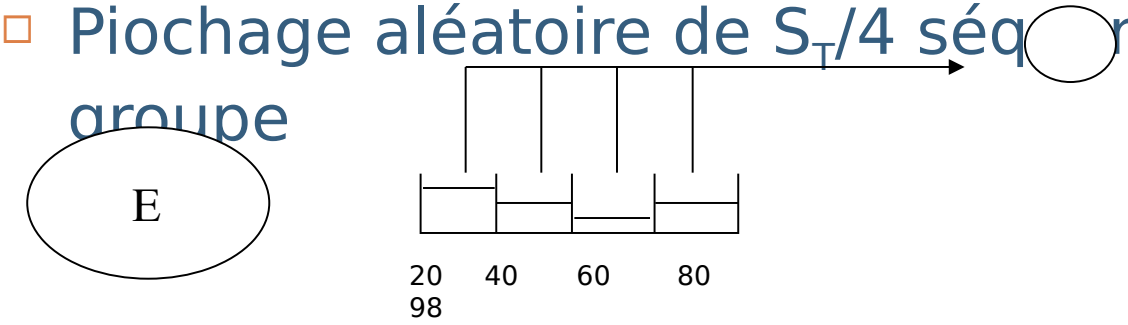
Échantillonnage aléatoire des séquences

- Motivation : ensemble E de séquences obtenues par PSI-BLAST
  - Répartition non uniforme en terme d'identité
  - Familles de séquences sous ou sur représentées
  - Nombre de séquences grand
- Répartition des séquences de E dans 4 groupes d'identité par rapport à la séquence référence (20-40, 40-60, 60-80, 80-98)



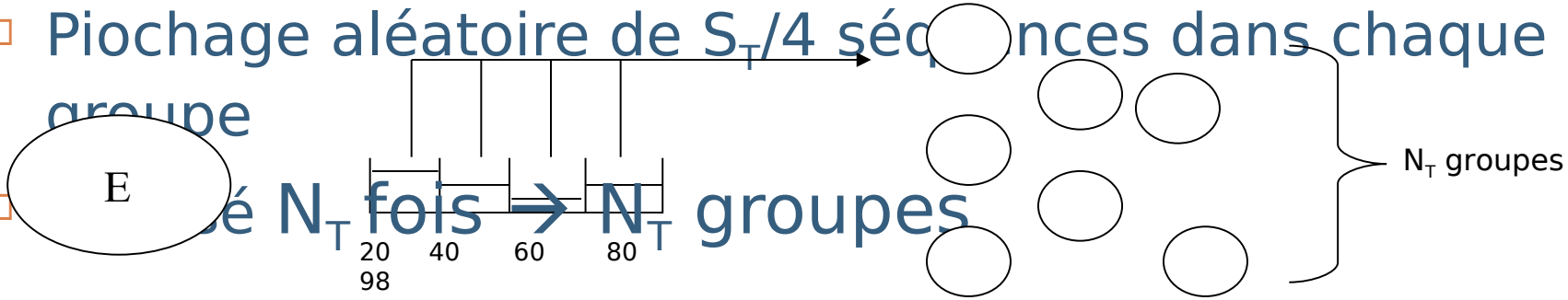
Échantillonnage aléatoire des séquences

- Motivation : ensemble E de S_N séquences obtenues par PSI-BLAST
  - Répartition non uniforme en terme d'identité
  - Familles de séquences sous ou sur représentées
  - Nombre de séquences grand
- Répartition des séquences de E dans 4 groupes d'identité par rapport à la séquence référence (20-40, 40-60, 60-80, 80-98)
- Piochage aléatoire de $S_T/4$ séquences dans chaque groupe



Échantillonnage aléatoire des séquences

- Motivation : ensemble $\mathbb{E}$ de S_N séquences obtenues par PSI-BLAST
  - Répartition non uniforme en terme d'identité
  - Familles de séquences sous ou sur représentées
  - Nombre de séquences grand
- Répartition des séquences de $\mathbb{E}$ dans 4 groupes d'identité par rapport à la séquence référence (20-40, 40-60, 60-80, 80-98)
- Piochage aléatoire de $S_T/4$ séquences dans chaque groupe
- Répété N_T fois → N_T groupes

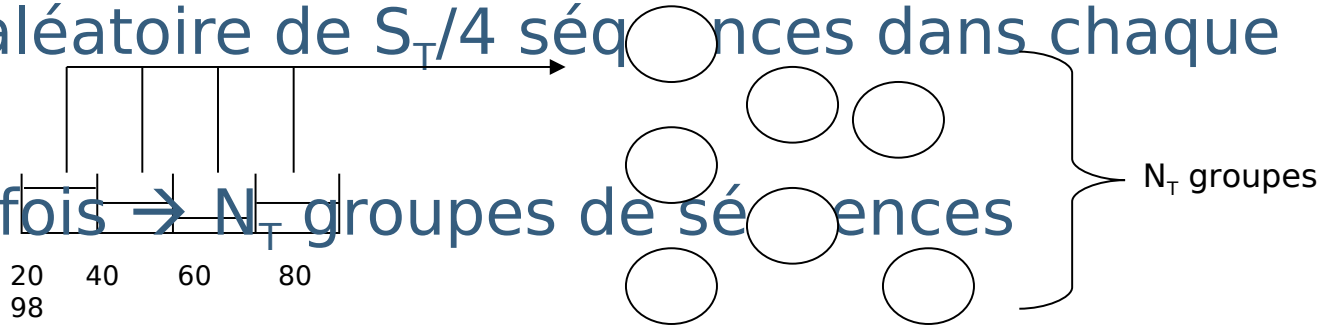


Échantillonnage aléatoire des séquences

- Motivation : ensemble E de S_N séquences obtenues par PSI-BLAST
  - Répartition non uniforme en terme d'identité
  - Familles de séquences sous ou sur représentées
  - Nombre de séquences grand
- Répartition des séquences de E dans 4 groupes d'identité par rapport à la séquence référence (20-40, 40-60, 60-80, 80-98)
- Piochage aléatoire de $S_T/4$ séquences dans chaque groupe

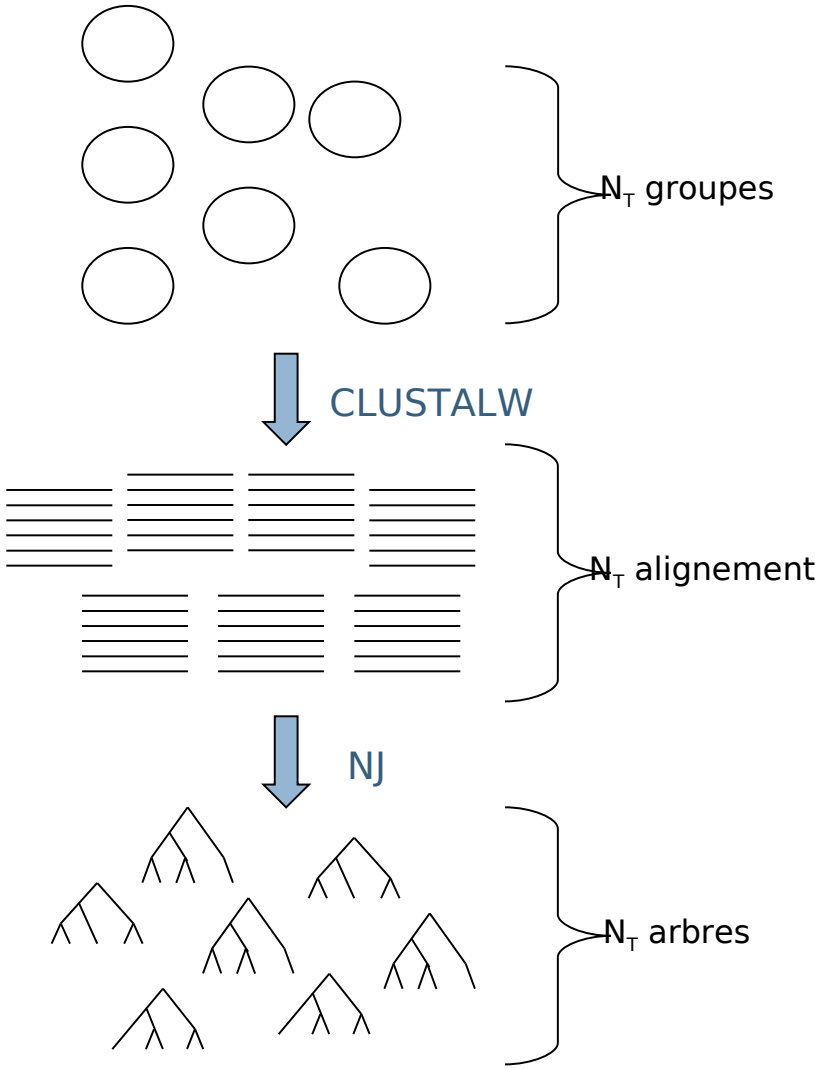
$$C_{N_i}^{\frac{S_T}{4}} \geq 2 \times N_T$$

- Réalisé N_T fois → N_T groupes de séquences



Alignement multiple et construction d'arbres phylogénétiques

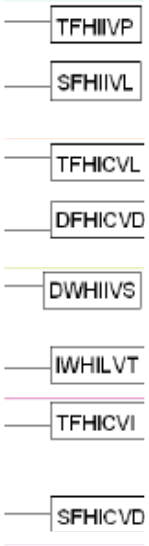
- Alignement multiple de chaque groupe de séquences avec CLUSTALW
  - N_T alignement de S_T séquences
  - N_T matrices de distances
- Construction d'arbres par Neighbor Joining (NJ) à partir des matrices de distances
  - N_T arbres métriques



Évaluation de la conservation des résidus : calcul des séquences consensus et backtraces

- Séquences consensus des feuilles = séquences de l'alignement

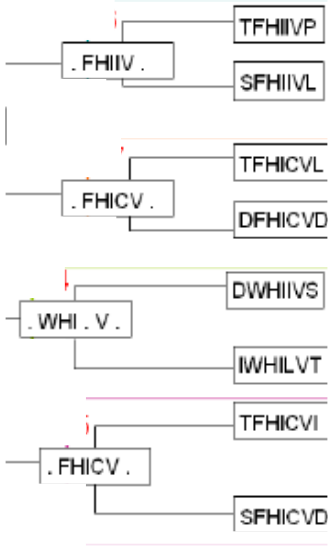
Arbre des séquences consensus



Évaluation de la conservation des résidus : calcul des séquences consensus et backtraces

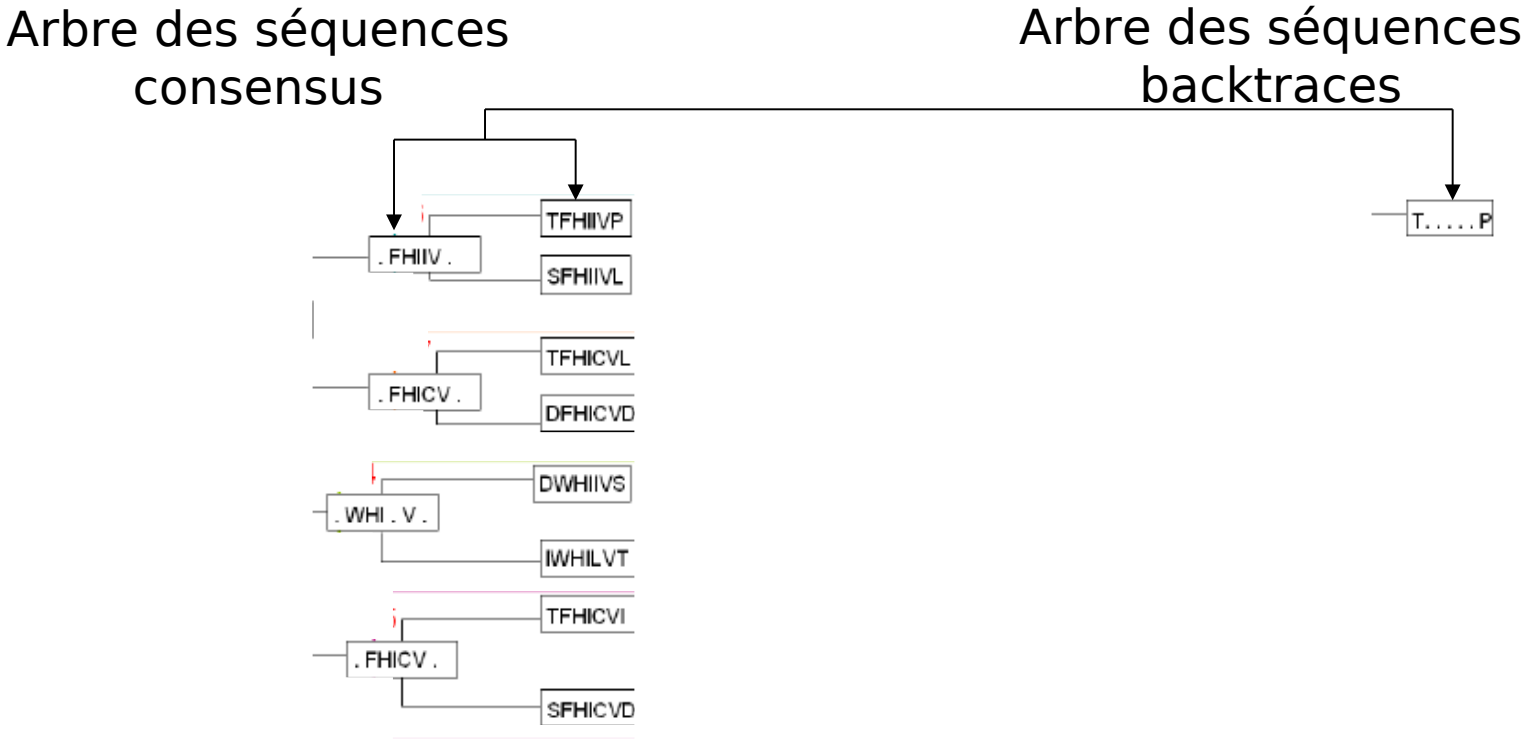
- consensus(n) = consensus(fils_g(n)) ∩ consensus(fils_d(n))

Arbre des séquences consensus



Évaluation de la conservation des résidus : calcul des séquences consensus et backtraces

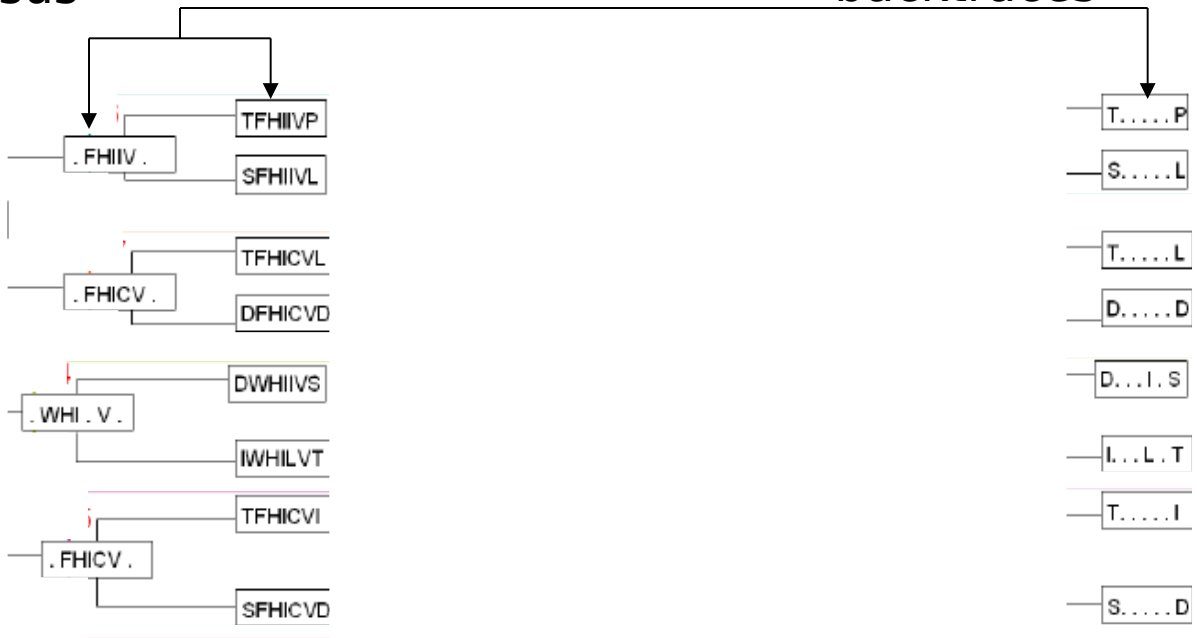
- backtrace(n) = consensus(n) – consensus (père(n))



Évaluation de la conservation des résidus : calcul des séquences consensus et backtraces

Arbre des séquences consensus

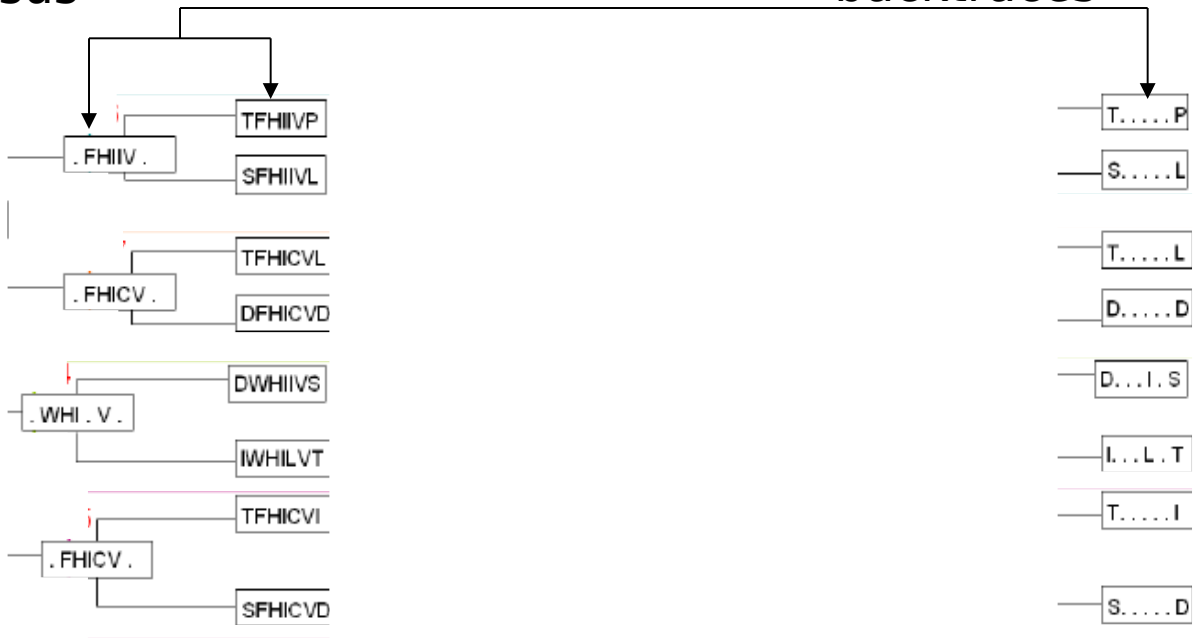
Arbre des séquences backtraces



Évaluation de la conservation des résidus : calcul des séquences consensus et backtraces

Arbre des séquences consensus

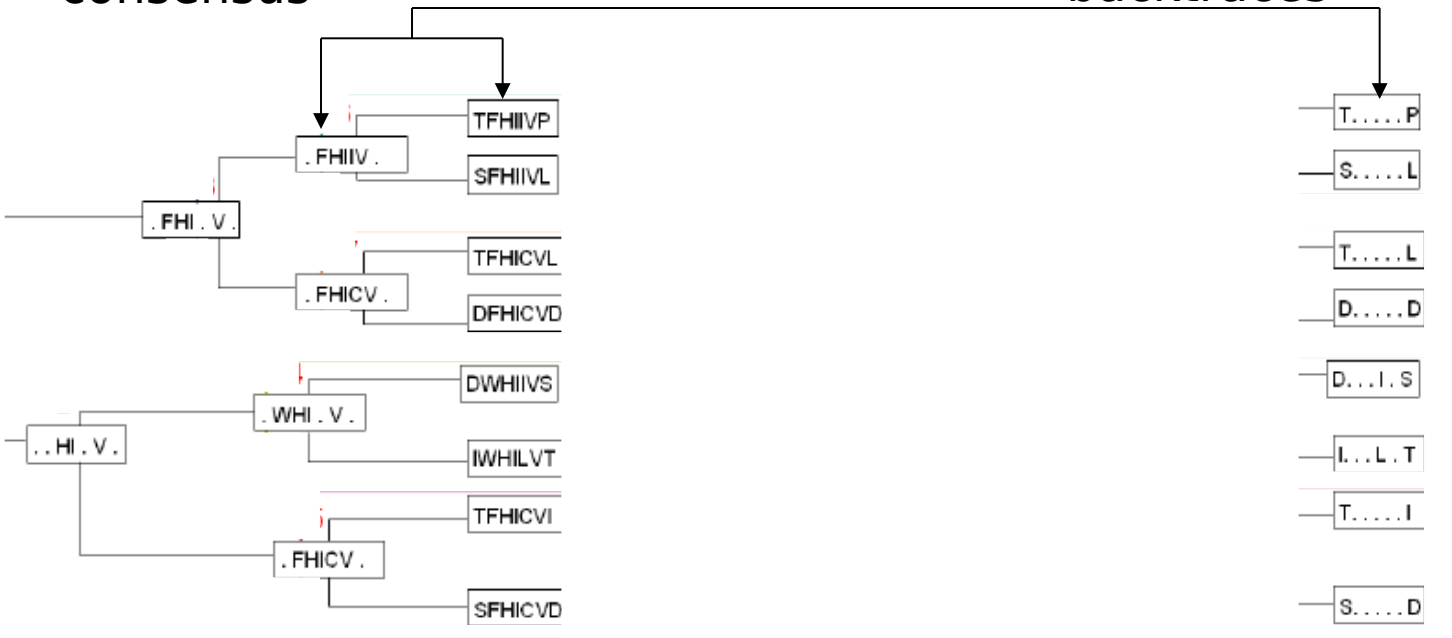
Arbre des séquences backtraces



Évaluation de la conservation des résidus : calcul des séquences consensus et backtraces

Arbre des séquences consensus

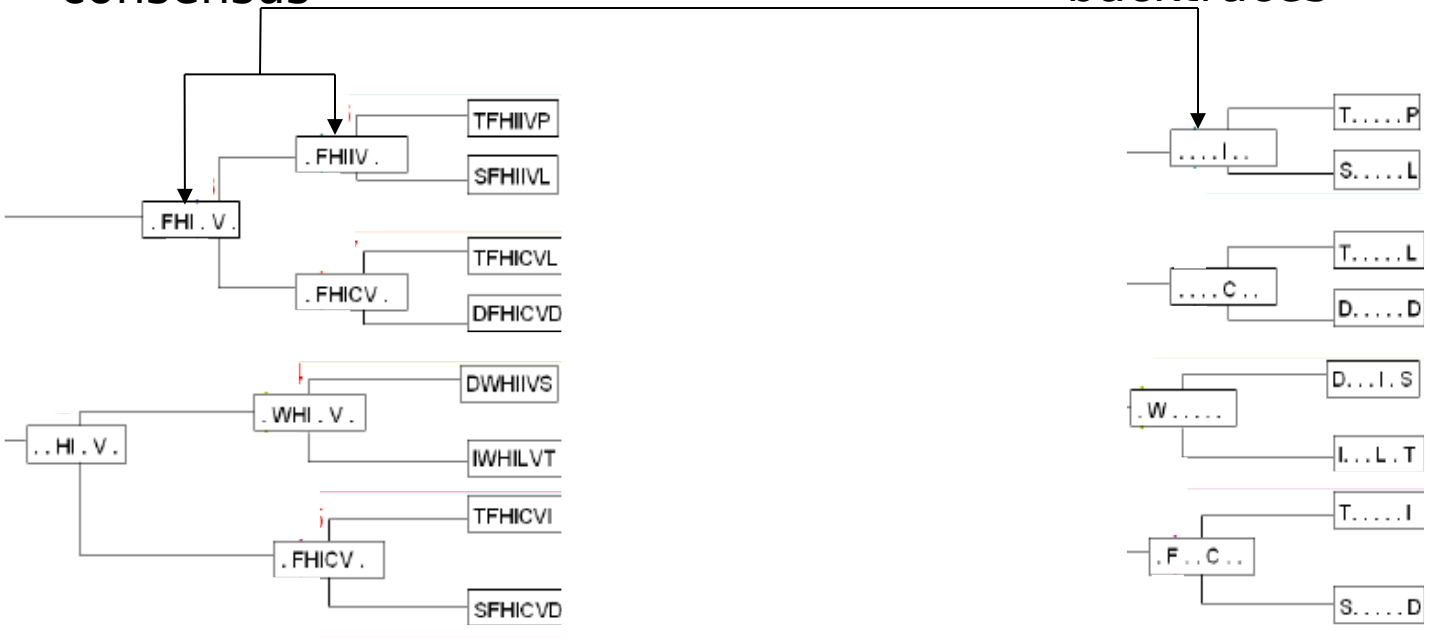
Arbre des séquences backtraces



Évaluation de la conservation des résidus : calcul des séquences consensus et backtraces

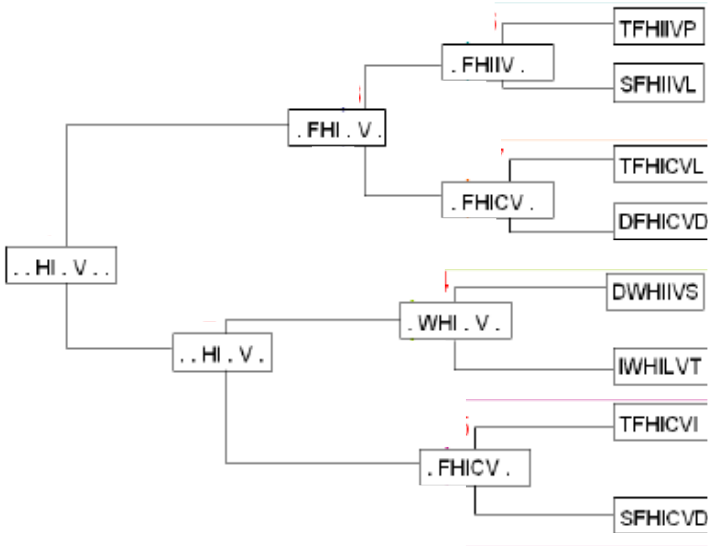
Arbre des séquences consensus

Arbre des séquences backtraces

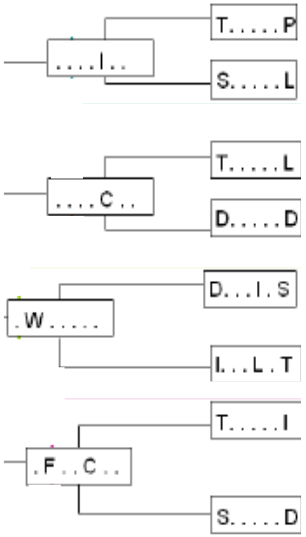


Évaluation de la conservation des résidus : calcul des séquences consensus et backtraces

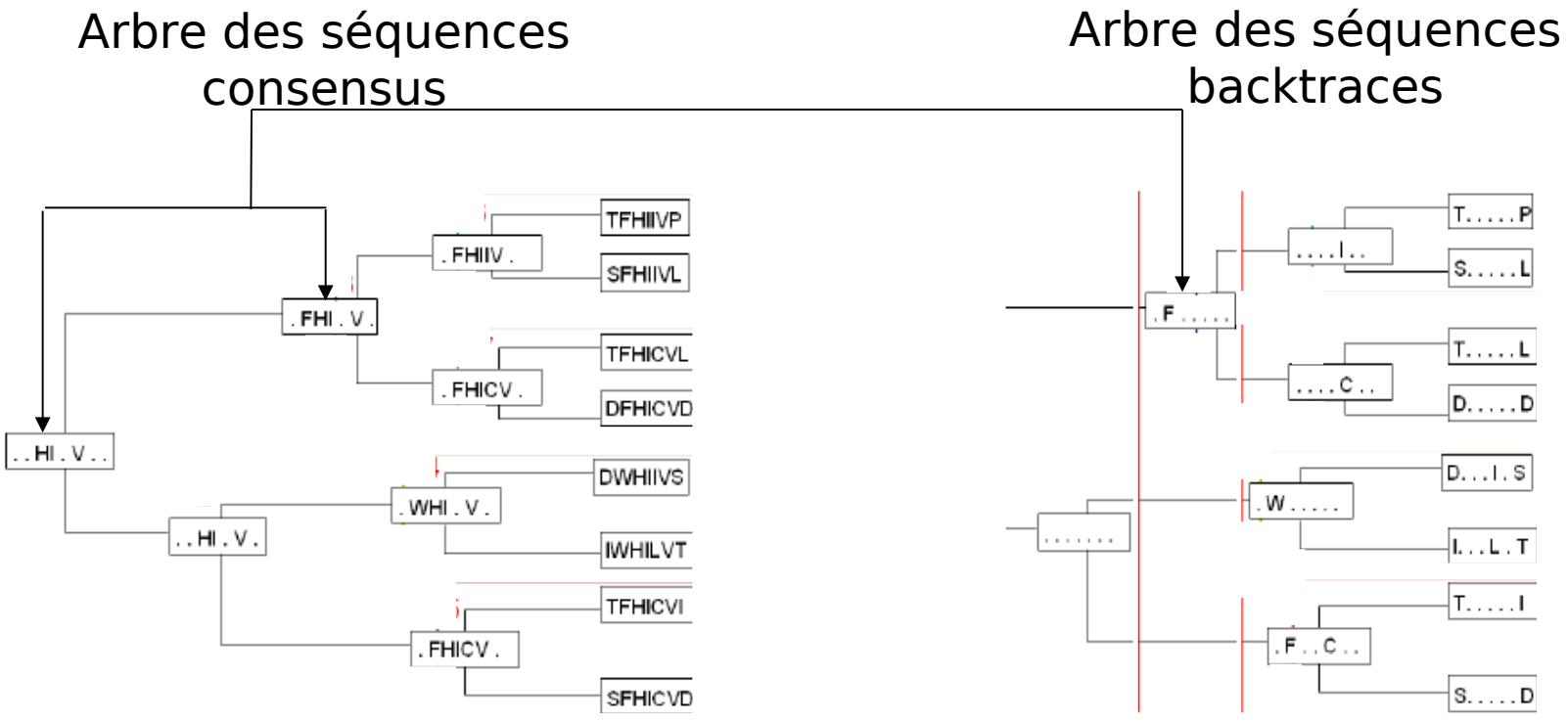
Arbre des séquences consensus



Arbre des séquences backtraces

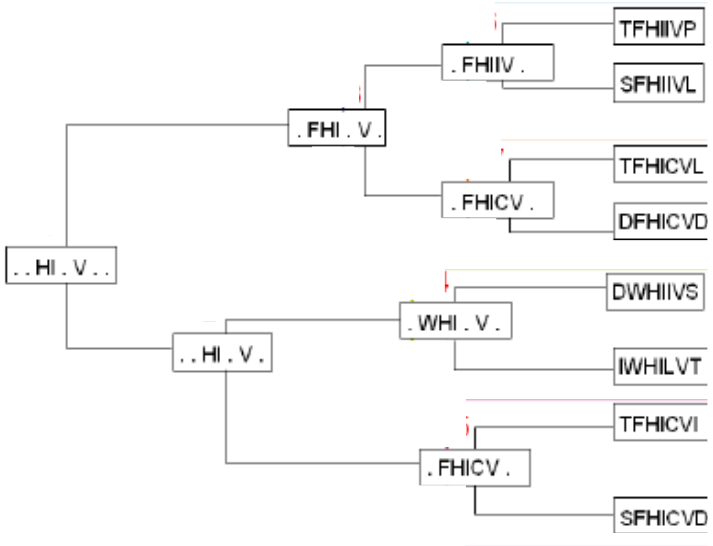


Évaluation de la conservation des résidus : calcul des séquences consensus et backtraces

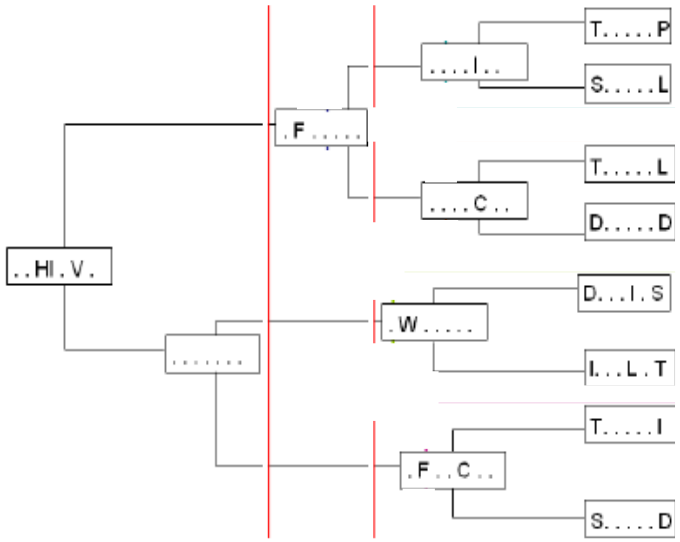


Évaluation de la conservation des résidus : calcul des séquences consensus et backtraces

Arbre des séquences consensus

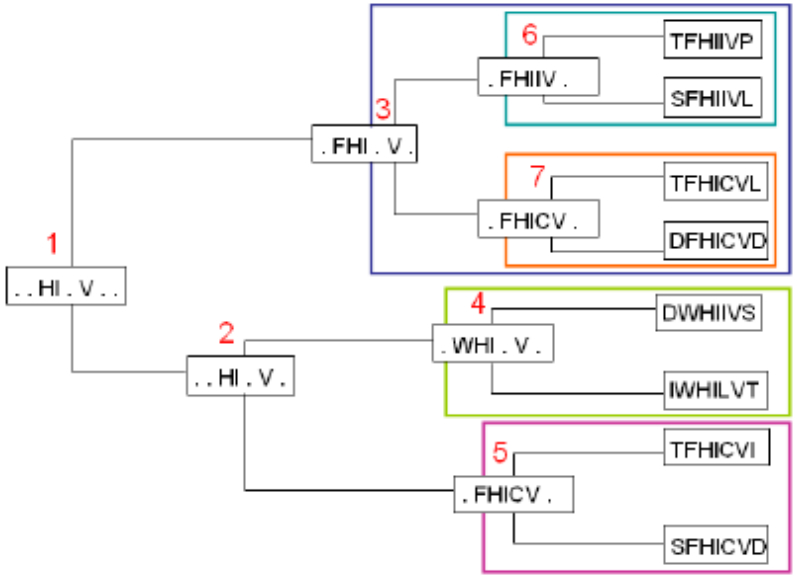


Arbre des séquences backtraces



Évaluation de la conservation des résidus : calcul des traces

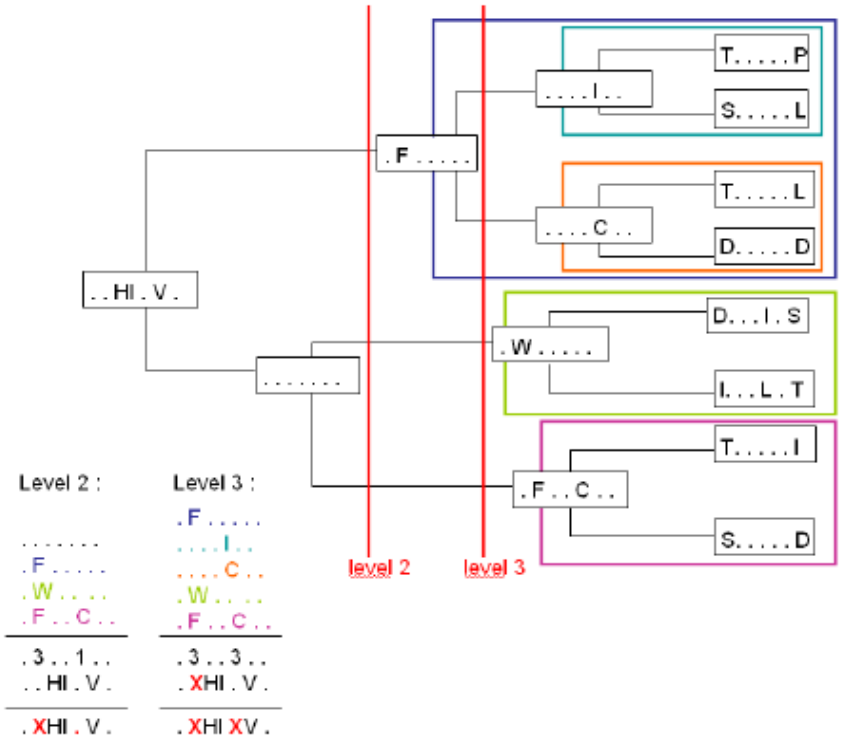
- Notion de rang pour les nœud internes :
  - Rang(racine)=1
  - Rang(i)=n si pour tout nœud j tel que $d_{racine,j} < d_{racine,i}$ on a Rang(j) < n et au moins un nœud j tel que Rang(j) = n-1



Évaluation de la conservation des résidus : calcul des traces

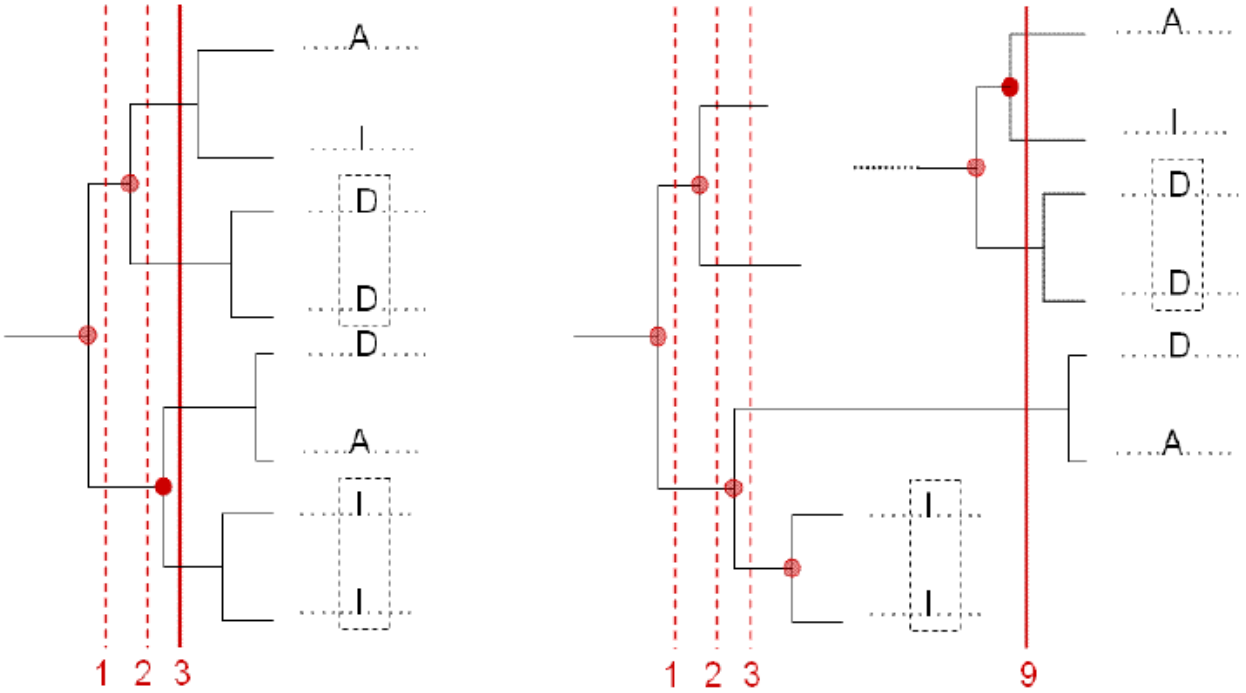
- Soit un noeud x de rang n, on coupe l'arbre aux positions correspondant à la distance d(racine,x). Si un résidu est backtrace dans au moins 2 des sous arbres résultant alors il est trace de niveau n.

Motivation : Pemet de récupérer les conservations locales de l'arbre



Évaluation de la conservation des résidus : calcul des traces

- Comparaison avec trace de ET (Evolutionary trace, O.Lichtarge)



Évaluation de la conservation des résidus : calcul des traces

- Un score de conservation d_j est calculé sur l'ensemble des arbres pour chaque résidus de la séquence

- Plus les résidus sont conservés, plus la t

$$d_j = \frac{1}{M} \sum_{t=1}^{M_j} \left(\frac{(N_t - l_j^t)}{N_t} \right)$$

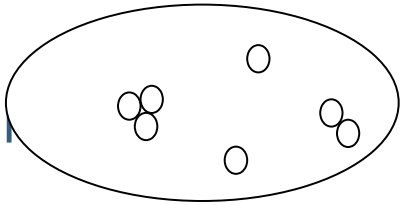
l_j^t : niveau de trace du résidu r_j dans l'arbre t
M : nombre d'arbres
M_j : nombres d'arbres ou le résidu r_j est trace
N_t : niveau maximal de l'arbre t

Clusterisation des résidus traces : motivations

- Les résidus à l'interface entre deux protéines forment des patches sur la surface des protéines
  - Clusterisation des résidus de surface
- Les résidus aux interfaces des protéines sont plus conservés que les autres résidus de surface
  - Clusterisation des résidus montrant une trace significative
- Seulement 39% des résidus d'une interface montrent une conservation significative
  - Clusterisation des résidus selon la trace pour former une graine que l'on étend ensuite
- Les résidus les plus conservés sont presque toujours à l'interface
  - Clusterisation des résidus par trace décroissante

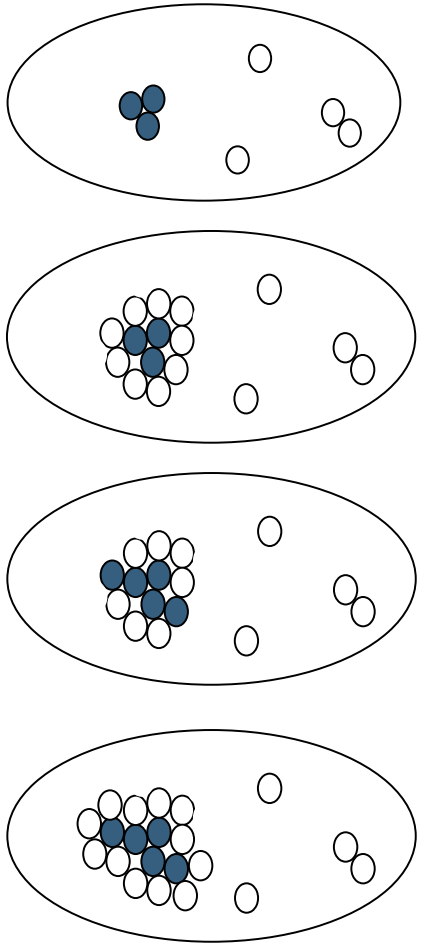
Clusterisation des résidus : algorithme

- Étape 1: Trier par trace décroissante les résidus de surface et de trace > seuil_residu
- Étape 2 (création graine): Pour chaque résidu i dans l'ordre du tri
  - Elargissement d'un cluster si :
    - résidu assez proche du cluster (5A)
    - $d_{cluster}$ après ajout du résidu > seuil_grai
  - Création d'un nouveau cluster si:
    - le résidu ne clusterise pas avec un cluster
    - d_i > seuil_graine



Clusterisation des résidus : algorithme

- Étape 3 : Sélection des graines de taille > seuil_taille
- Étape 4 : Collecte des résidus voisins aux graines
  - Si pas de voisins FIN
- Étape 5 : Extension des graines :
  - ajout des résidus voisins dans l'ordre du tri
  - Respect des règles de l'étape 2 en remplaçant seuil_graine par seuil_cluster
  - Retour à l'étape 4



Clusterisation des résidus : seuils

- Les résidus clusterisent différemment et plus ou moins bien en fonction de la structure considérée
  - Seuil_taille calculé par génération aléatoire de clusters sur la structure considérée = variable en fonction de la structure de la protéine
- La distribution des traces peut varier (protéines très conservées ou inversement)
  - Seuil_residu, seuil_graine et seuil_clusters fixé avec des niveaux de confiance sur la distribution des valeurs de traces

JET : Résultats