Regroupement et Visualisation de Protéines par les architectures de Domaines

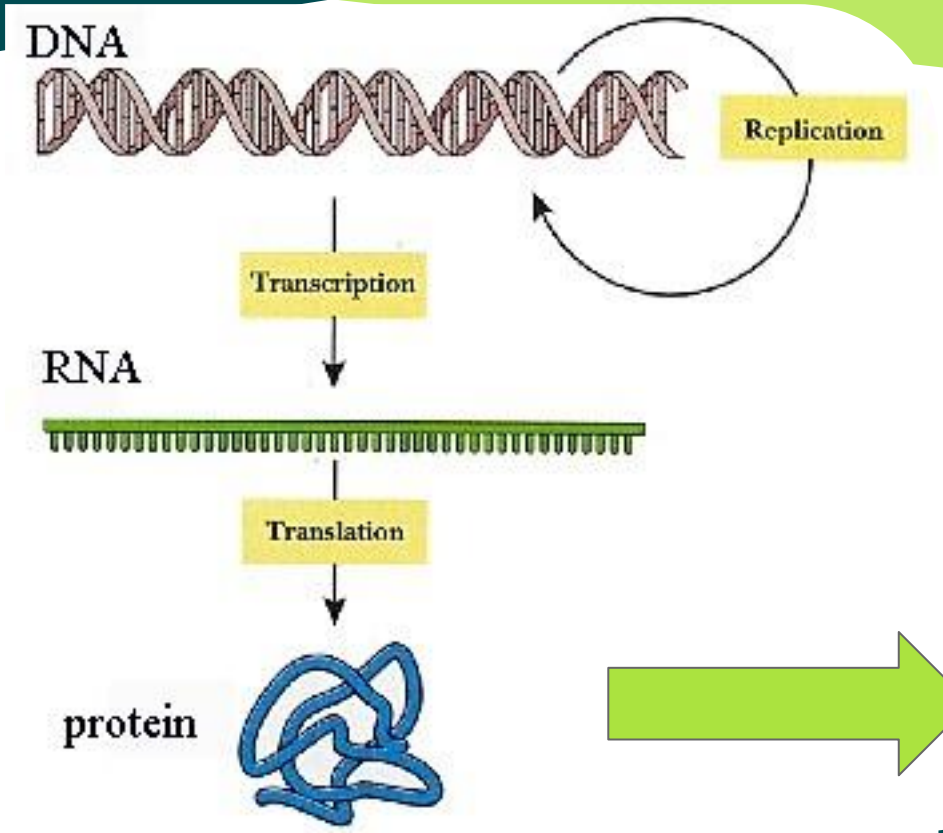
Chadi Jaouadi Paul Saïghi

Encadré par : Juliana Silva Bernardes

Sommaire

- I) Problématique
- II) Objectifs
- III) Regroupement hiérarchique
- IV) Visualisation graphique de protéines
- V) Résultats
- VI) Conclusion et Evolution

Problématique



Exemple de protéines

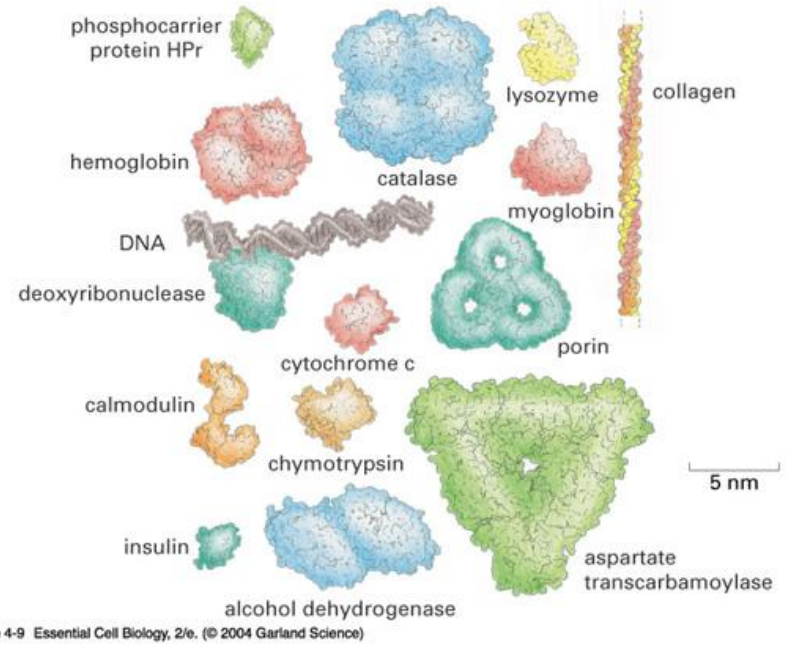
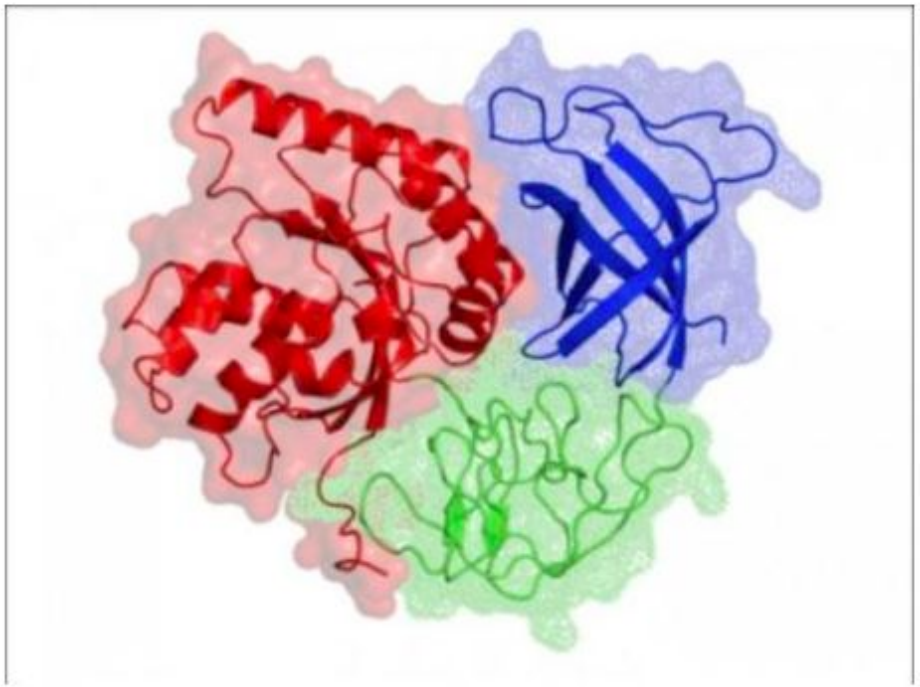


Figure 4-9 Essential Cell Biology, 2/e. (© 2004 Garland Science)

Domaines



Architectures de domaines

Identifiant Protéine	Taille	Domaine1	Confiance	Début	Fin	Domaine2	Confiance	Début	Fin
Coscinodiscus	1300	PF00385	1.18e-14	39	76	PF00595	5.17e-09	305	342
Agaricus_bisporus_XP	2500	PF00145	9.16e-120	272	599	PF01485	0.000158	696	729
Amphiprora_paludosa	765	PF00145	8.62e-64	346	647				
Arabidopsis_thaliana1	497	PF00145	5.78e-115	19	377				
Arabidopsis_thaliana2	870	PF00145	3.79e-10	306	491	PF00145	3.88e-43	502	622
Aspergillus_arachidicola	1234	PF01426	3.73e-3	138	250	PF00145	3.07e-80	321	591
Aspergillus_niger	611	PF01426	2.83e-31	132	240	PF00145	5.62e-87	306	513

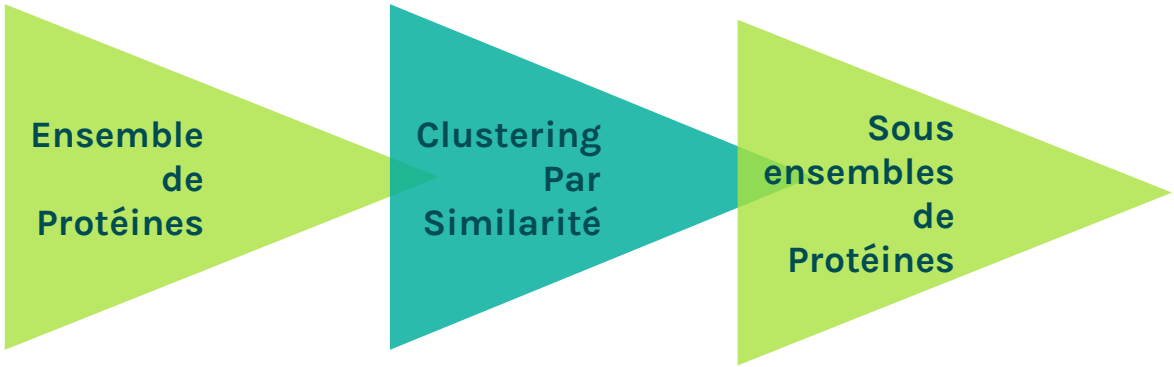
Table 1: Identification de domaines

Objectifs

- Représentations textuelles non adaptées
- Le Biologiste a besoin d'une visualisation graphique interactive
- Nous avons:
  - Implémenté **un algorithme de regroupement** pour les protéines ayant des architectures de domaines similaires
  - Développé **un site web**

Regroupement hiérarchique

- Clustering



- Cluster = un sous-ensemble composé de protéines ayant toutes des caractéristiques communes

Regroupement hiérarchique : Algorithme

- Entrée: [P1, P2, P3] et N = nombre de clusters
- Construit une **Matrice comparaison inter-protéine**
- Créer un Cluster par Protéine : { G1[P1],G2[P2],G3[P3] }
- Boucler
  - Mettre à jour la **Matrice comparaison inter-cluster**
  - Fusionner les deux clusters les plus similaires
- Arrêter si nombre de clusters est égal à N

Matrice comparaison inter-protéine

- Similarité entre deux Protéines
  Distance de Damerau Levenshtein :
  - Transposition
  - Suppression
  - Insertion
  - Substitution

- L'Architecture des Domaines
  comme critère de similarité

P1: D1

P2: D1-D2

P3: D1-D2-D3

	P1	P2	P3
P1	0	1	2
P2	1	0	1
P3	2	1	0

Créer un Cluster par Protéine

G1[P1] G2[P2] G3[P3]

- On rentre dans la Boucle

10

	G1	G2	G3
G1	0	1	2
G2	1	0	1
G3	2	1	0

Matrice distance de la liste de clusters actuelle

Fusion des deux Clusters les plus similaires

G1[P1] **G2[P2]** **G3[P3]**

	G1	G2	G3
G1	0	1	2
G2	1	0	1
G3	2	1	0

Matrice distance de la liste de clusters actuelle

Fusion des 2 Clusters les plus similaires

G2[P1, P2]
G3[P3]

Nouvelle matrice

G2[P1, P2] G3[P3]

	G2	G3
G2	0	1.5
G3	1.5	0

Matrice distance de la nouvelle liste De Clusters

Matrice de distance inter-cluster

- **Critère de Lien Moyen**

Pour toute (prot1, prot2) appartenant à (G1,G2)

Similarité(G1,G2) =

MOY(similaritéProt(prot1,prot2))

= [similaritéProt(P1,P3)+simlaritéProt(P2,P3)] / 2

= [2 + 1] / 2

= 1.5

	P1	P2	P3
P1	0	1	2
P2	1	0	1
P3	2	1	0

*Matrice distance **inter-protéine***

G2[P1, P2] **G3[P3]**

	G2	G3
G2	0	1.5
G3	1.5	0

*Matrice distance **inter-cluster***

Arrêter si nombre de cluster est égal à N

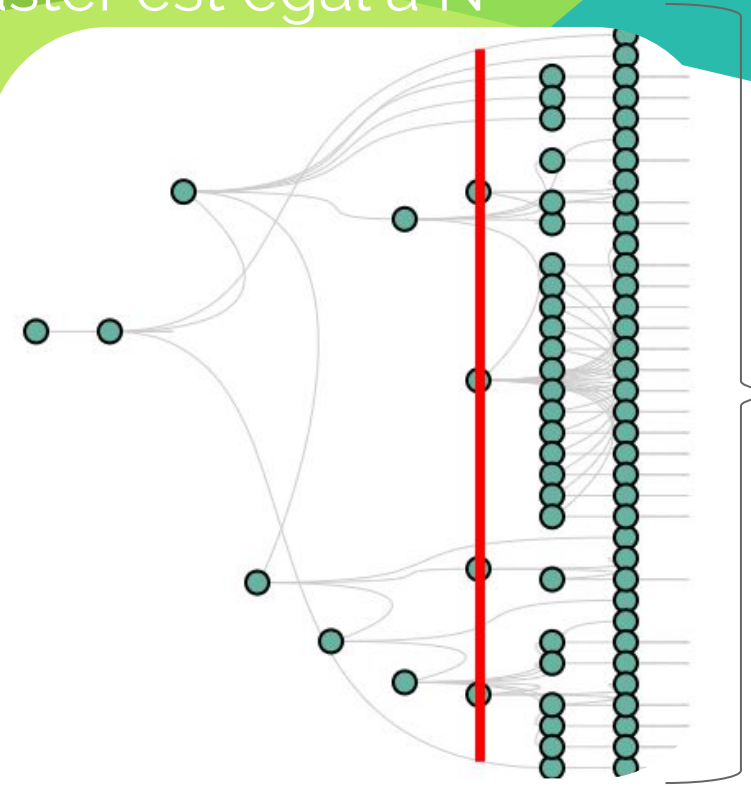
L'utilisateur peut choisir le nombre de cluster opportun à partir d'un Arbre Dendrogram

Entrée Initiale:

Une trentaine
De Protéiens

Arrêter si nombre de cluster est égal à N

L'utilisateur peut choisir le nombre de cluster opportun à partir d'un Arbre Dendrogram

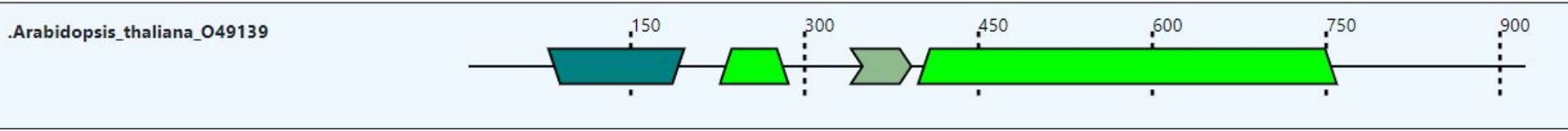


Entrée Initiale:

Une trentaine
De Protéiens

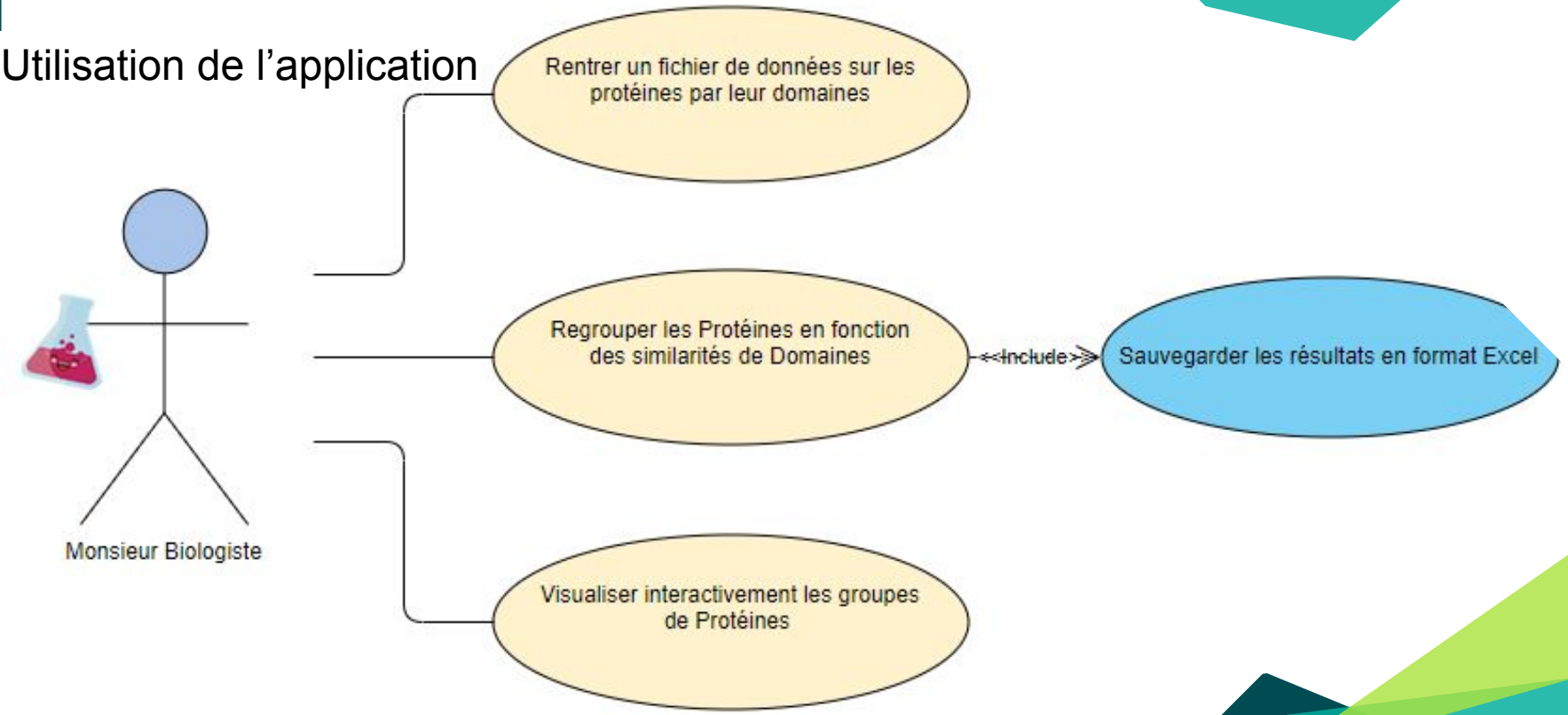
Visualisation Graphique d'une Protéine

- Distinguer les Domaines

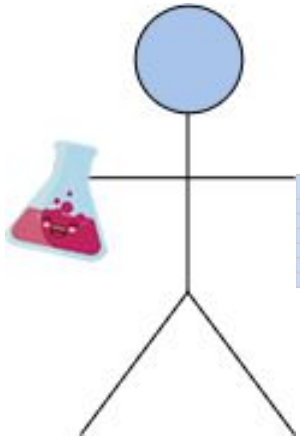


Resultats

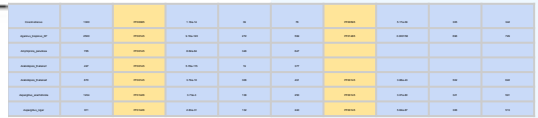
Cas d'Utilisation de l'application



Interface Utilisateur



Mr Biologiste

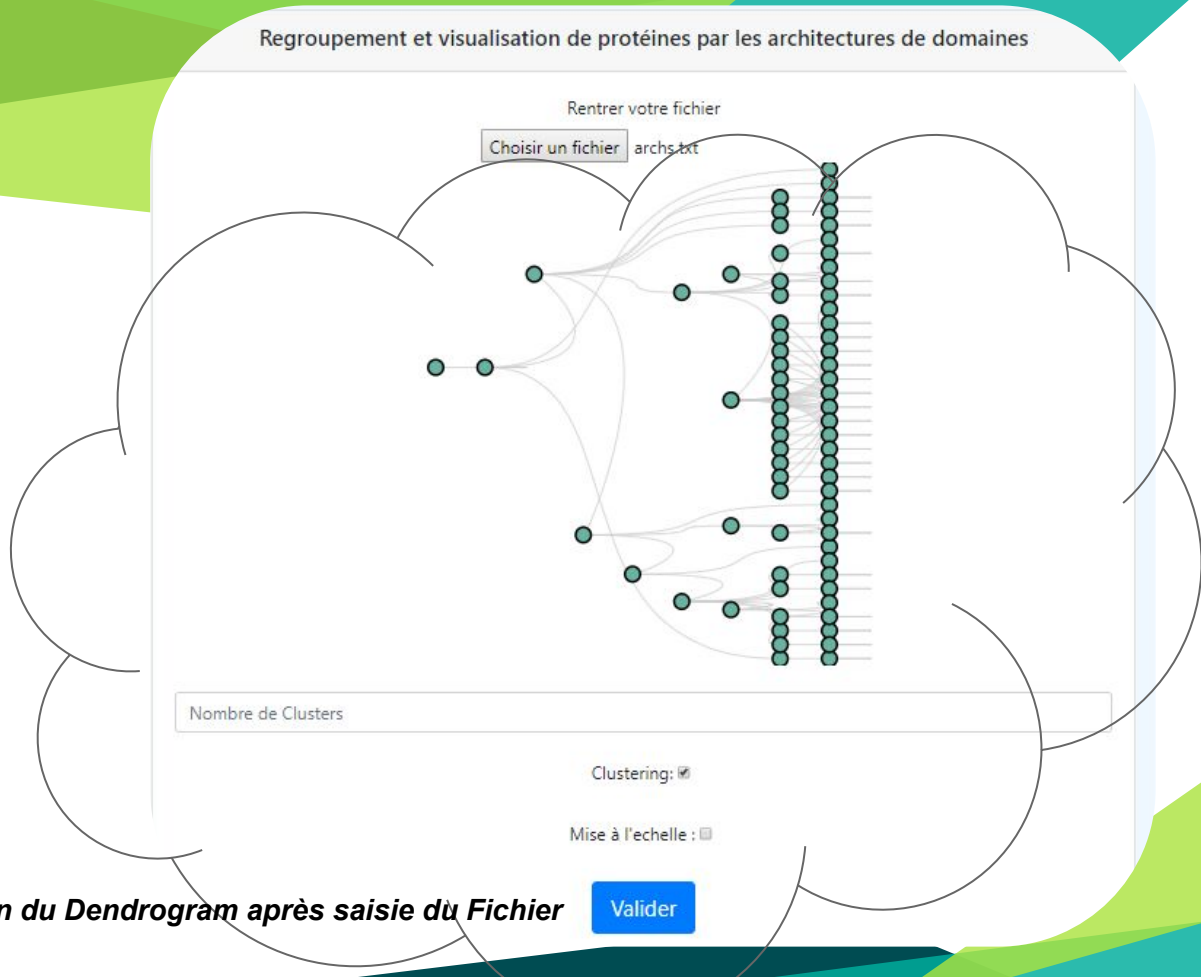


Fichier Données de Protéines
Par architecture de domaines

Capture Ecran 2: Formulaire de saisie de fichier

Dendrogram

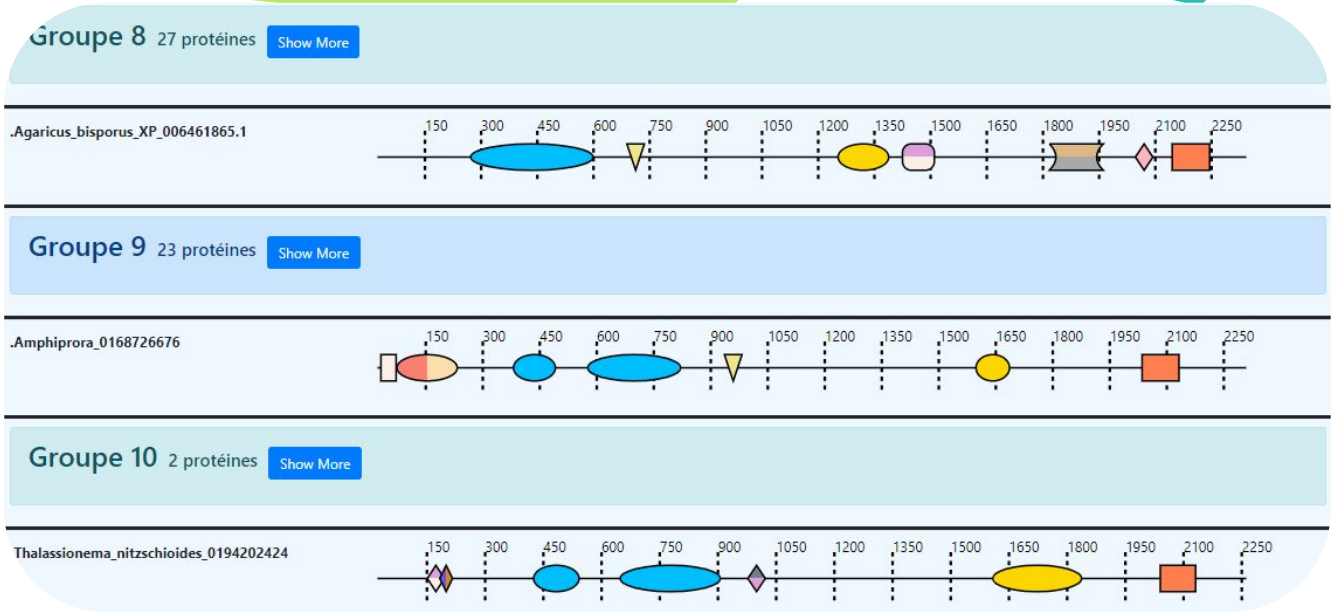
Choisir le nombre de cluster opportun à partir d'un Arbre Dendrogram



Capture Ecran 3: Apparition du Dendrogram après saisie du Fichier

Visualisation de Clusters

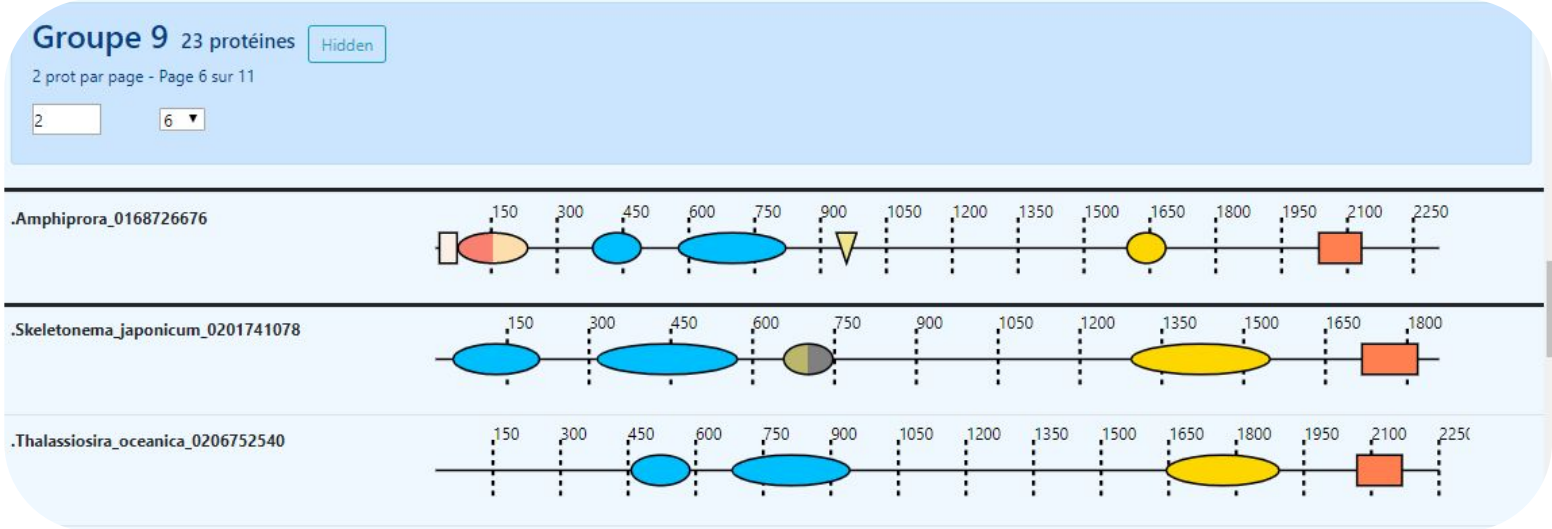
Mr Biologiste



Capture Ecran 4 : Regroupement de Proteines

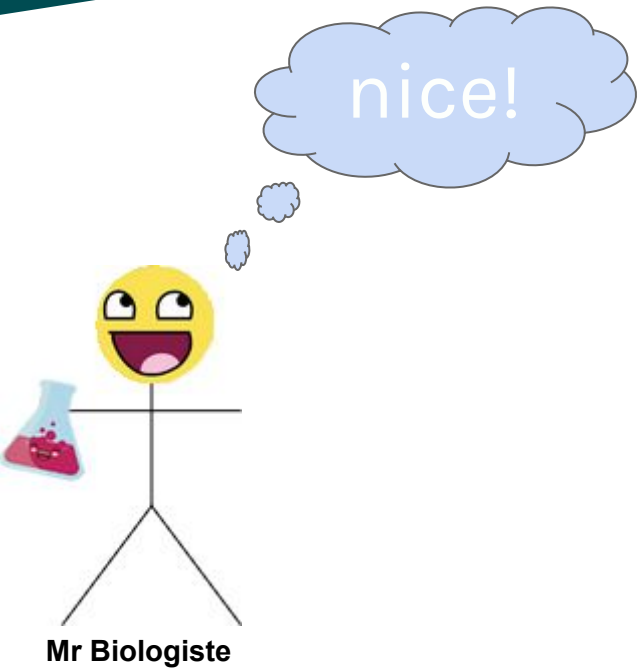
Visualisation Clusters - Show More

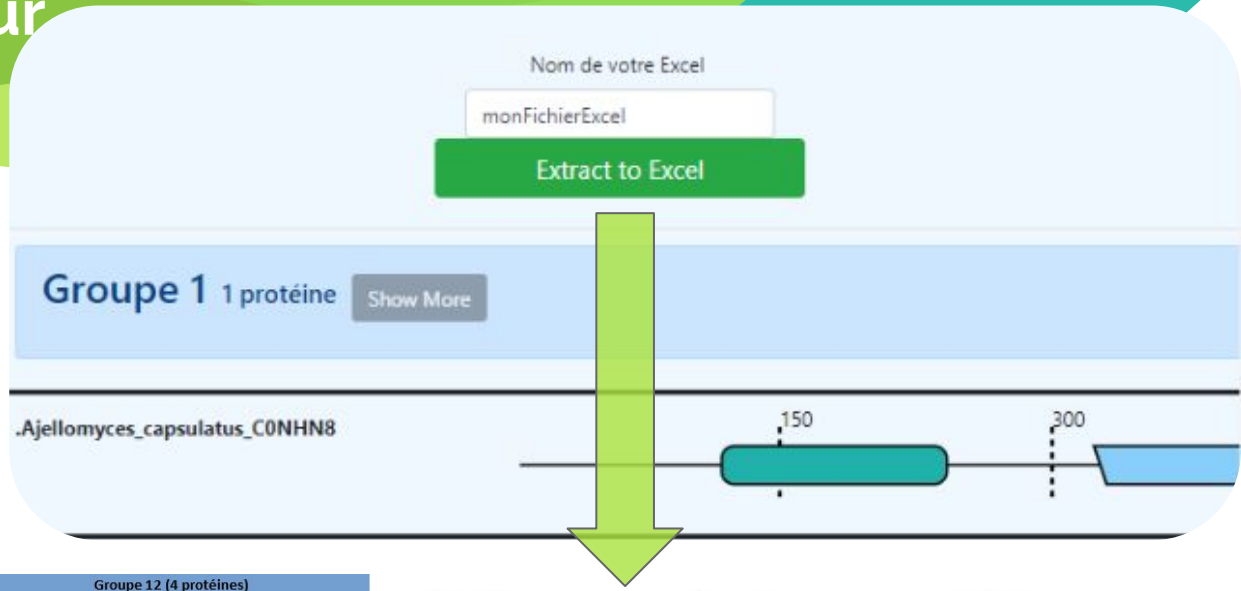
Mr Biologiste



Capture Ecran 5 : Regroupement de Proteines - Show More Groupe 9

Sauvegarde en Tableur





Groupe 12 (4 protéines)										
Danio_rerio_Q588C7	1409	PF00307	7,91E-35	16	113	PF00855	1,79E-17	790	848	PF00628	8,47E-07	1066	1116
Homo_sapiens_Q9UBC3_	851	PF00855	2,93E-31	222	295	PF00628	1,93E-06	478	530	PF00145	1,67E-19	576	760
Homo_sapiens_Q9Y6K1_	973	PF10310	1,68E-25	1	172	PF00855	4,28E-29	290	364	PF00628	9,1E-06	536	586
Mus_musculus_O88508	988	PF10310	1,7E-22	1	168	PF00855	3,73E-29	286	360	PF00628	8,74E-06	532	582
Groupe 13 (10 protéines)										
Ascobolus_immersus_O42731	1426	PF12047	5,4E-42	93	248	PF01426	6,6E-22	432	575	PF01426	1,11E-09	638	744
Colletotrichum_gloeosporioides_L2GAQ2	1130	PF12047	8,28E-34	173	266	PF01426	1,68E-33	450	539	PF01426	1,26E-07	564	637
Cordyceps_militaris_G3JGU4	1147	PF12047	6,81E-44	158	267	PF01426	6,78E-35	436	527	PF01426	1,07E-06	554	656
Paracoccidioides_lutzii_C1H2T7	1274	PF12047	3,53E-47	189	296	PF01426	2,38E-31	458	547	PF01426	1,38E-09	569	656
Pseudogymnoascus_verrucosus_A0A1B8GCG9	1305	PF12047	1,2E-36	164	258	PF01426	3,88E-29	455	544	PF01426	1,99E-06	582	680
Purpureocillium_lilacinum_A0A179HTB7	1150	PF12047	1,34E-40	168	276	PF01426	5,13E-35	448	537	PF01426	1,14E-05	608	665
Rhizophagus_irregularis_A0A1C9IHL2	1373	PF12047	1,46E-30	208	352	PF01426	9,9E-30	514	626	PF01426	4,53E-20	683	825
Grosmannia_clavigera_F0XN91	1160	PF12047	1,06E-39	90	197	PF01426	1,55E-33	359	456	PF00145	4,9E-108	656	1056
Pochonia_chlamydosporia_A0A179FKN0	990	PF12047	4,21E-38	16	112	PF01426	4,84E-36	290	383	PF00145	1,5E-107	545	940
Neurospora_crassa_Q3Y3Z1	1296	PF12047	8,94E-45	218	324	PF01426	1,82E-48	575	679	PF00567	8,37E-05	666	721
Groupe 14 (6 protéines)										
Bos_taurus_Q24K09	1626	PF06464	7,61E-45	17	106	PF12877	1,51E-18	124	328	PF12047	5,4E-69	397	531
Homo_sapiens_P26358	1614	PF06464	8,47E-45	16	105	PF12877	1,7E-24	125	354	PF12047	4,86E-69	399	533
Ratus_norvegicus_Q9Z330	1619	PF06464	7,54E-45	16	106	PF12877	9,7E-10	93	344	PF12047	5,53E-65	405	539

Conclusion, Evolution

- Importance pour le Biologiste
- Evolution : Exploiter d’autres algorithmes de Regroupement

Merci !

Des questions?

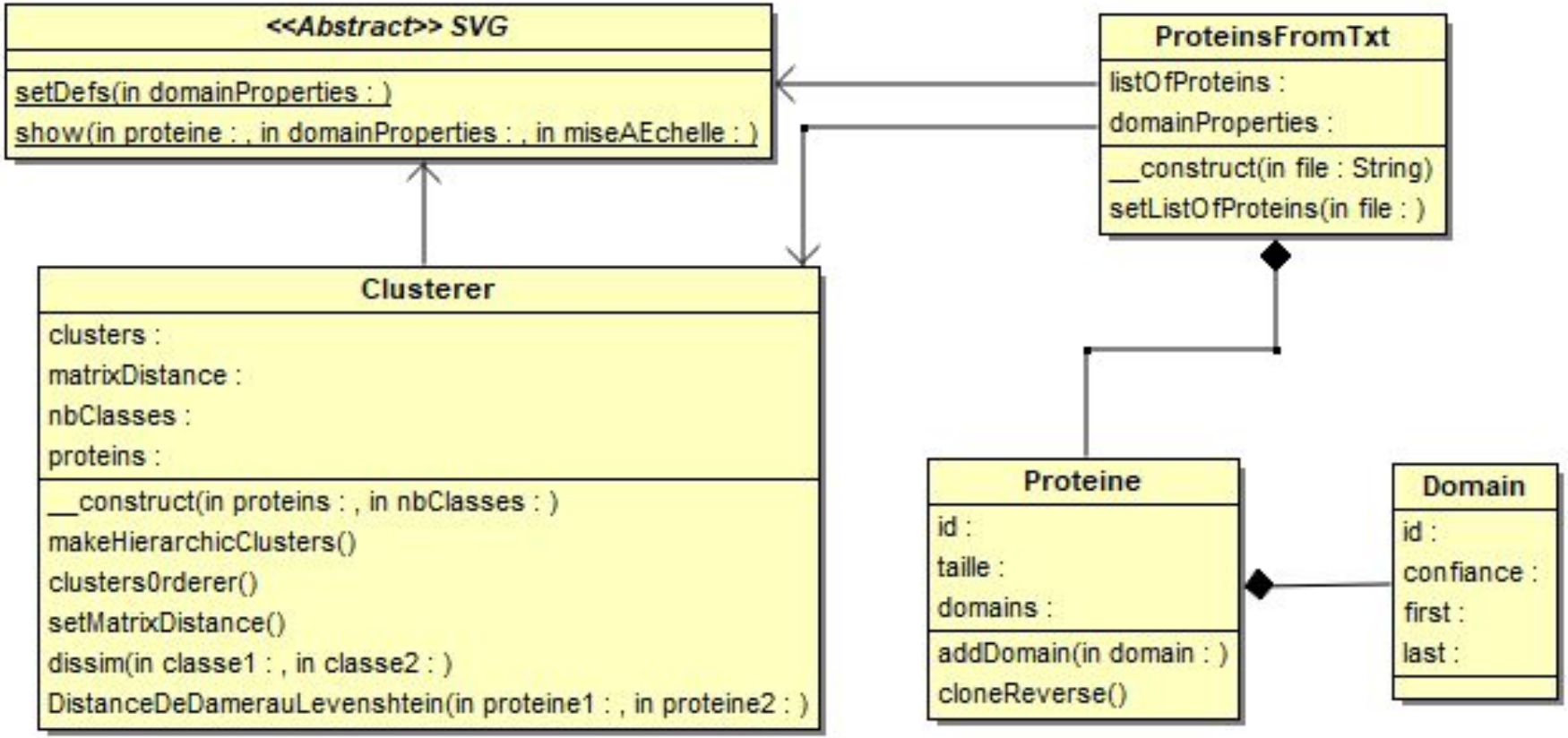


Figure 1: Diagramme de Classes UML