Annexe I : Genstyle, un outil en ligne pour l'analyse des signatures génomiques

I.1 Motivation du projet

L'analyse des séquences d'ADN par le biais de la signature génomique intéresse de plus en plus de scientifiques. Durant mon travail de thèse, j'ai été assez souvent sollicité pour mettre en place une approche par nos méthodes sur diverses questions biologiques. Pourtant, dans la plupart des cas, l'utilisation des signatures génomiques ne requiert que peu de connaissances préalables. La mise à disposition d'outils adaptés peut donc permettre à un grand nombre de chercheurs de comparer rapidement leurs séquences du point de vue de la signature génomique. C'est dans cet esprit que nous avons développé Genstyle

I.2 Présentation de Genstyle

Genstyle est disponible gratuitement en ligne sur le site http://genstyle.imed.jussieu.fr/. C'est un ensemble d'outils associés à une base de données recensant l'ensemble des séquences disponibles sur Genbank. L'ensemble est organisé autour d'un espace de travail (baptisé workspace) qui donne accès aux trois boîtes à outils□

- Les outils de collecte des séquences□(sequence collector toolbox)□

Ces outils sont destinés à importer les séquences d'intérêt dans l'espace de travail. L'utilisateur peut soit importer ses propres séquences, soit les choisir dans la base de données (à partir du nom de l'espèce, de la taxonomie, ou du numéro d'accession Genbank).

- Les outils de tri des séquences de l'espace de travail (sequence filter toolbox)□

Pour un travail donné, l'utilisateur peut n'être intéressé que par un sous-ensemble des séquences de l'espace de travail. Des outils de sélection automatique des séquences sont disponibles pour faciliter leur tri. Il est possible de choisir les séquences sur la base de leur longueur, suivant leur statut (ARN / ADN mitochondrial / Chloroplaste / ADN et divers). Les résultats des sélections sont modifiables par l'utilisateur.

- Les outils d'analyse des séquences□(sequence analyser toolbox)□

  Plusieurs outils sont proposés pour l'analyse des séquences choisies. Pour chaque séquence, il est possible de construire la signature génomique, et de la comparer à celles de l'ensemble des espèces représentées dans Genbank (la comparaison se fait à l'aide de la distance euclidienne). Les signatures de l'ensemble des séquences sélectionnées peuvent être affichées simultanément. Enfin, une «□carte□» des séquences est proposée. Cette carte exprime approximativement les distances entre signatures, elle est construite par analyse en composantes principales.

Genstyle permet ainsi à n'importe quel utilisateur de produire en quelques minutes des résultats simples tels que□

- L'identification de l'espèce d'origine de courts fragments d'ADN□ Genstyle donne la possibilité de rechercher parmi de nombreuses espèces celles dont la signature ressemble le plus à celle d'une séquence d'intérêt. Dans le cas où la signature de l'organisme d'origine est disponible dans Genstyle, il est probable que celle-ci sera la plus proche. Cette fonctionnalité peut également être avantageusement utilisée sur un transfert horizontal supposé□dans ce cas, Genstyle peut proposer des donneurs possibles.
- La comparaison des styles d'écriture des espèces au moyen de distances entre signatures. A partir des distances entre signatures d'espèces, il est possible de construire un arbre taxonomique.

- L'analyse des liens unissant les séquences d'intérêt du point de vue de la signature génomique (par analyse en composantes principales).

La recherche de fragments originaux au sein d'un génome (par comparaison de plusieurs séquences issues de la même espèce). Ces fragments originaux étant probablement des ARN ou des transferts horizontaux (Dufraigne, 2004; Dufraigne, 2005).
A tout moment, l'utilisateur a la possibilité de sauvegarder son espace de travail pour des recherches ultérieures.

I.3 Historique du projet

Une applet Java était disponible en ligne depuis 2000 permettant de rechercher les plus proches voisins de la signature d'une séquence fournie par l'utilisateur parmi une base de 12 000 signatures (ce travail a entre autres été mené par Philippe Renaud-Goud).
En 2002, John Lecointe (étudiant de DESS) a également travaillé à la mise en place d'une méthode rapide pour la de recherche des plus proches voisins grâce à une organisation préalable d'une base de 50 000 signatures. En 2003, Eric Oeuillet (étudiant de DEA) a mis à la disposition de l'équipe la base de 50 000 signatures en intranet. Il a également continué les recherches entreprises par John Lecointe.
En 2004, Matthieu Massin est arrivé dans notre équipe pour un stage de fin d'étude d'ingénieur en informatique (ESIEA) et j'ai participé à son encadrement. Il est resté un peu plus d'un an. C'est avec son arrivée que le projet Genstyle a pris toute son ampleur. Durant tout ce temps, j'ai travaillé sur ce projet en collaboration avec Matthieu : mon apport était surtout destiné à anticiper les besoins des utilisateurs, tandis que Matthieu était en charge des contraintes liées à l'implémentation. Nous avons donc créé ensemble la version finale de Genstyle. Une partie du travail a été déléguée. En effet, la même année, Matthieu Moratille (IUT d'informatique, Montreuil) a développé une nouvelle version de l'applet Java, Lies Dermoune (école d'ingénieur en informatique ESIEA) a mis en place la procédure d'import des séquences dans l'espace de travail, Caroline Devic (étudiante en DEA d'informatique médicale, Paris VI) et Nathalie Destrubé (ingénieur de l'EPF), ont conçu et implémenté les procédures optimisées pour la recherche des plus proches voisins d'une signature. L'ensemble des personnes travaillant sur le projet Genstyle a étroitement collaboré avec Matthieu Massin et moi pour que leurs apports puissent facilement êtres intégrés au sein du projet.

I.4 Implémentation

Plusieurs langages informatiques ont été utilisés pour l'implémentation de Genstyle :

- Le HTML (dialogue avec l'utilisateur)
- Le PHP/MySQL (interrogation des bases) en effet, Genstyle utilise pour son fonctionnement plusieurs bases de données, dont une base de séquences génomiques miroir de Genbank (baptisée « Genstyle companion database »).
- Le C pour les calculs matriciels (calcul des signatures, des distances et de l'analyse en composantes principales).
- fLe JAVA pour l'applet permettant de calculer la signature d'une séquence puis sa comparaison à une base de donnée de signatures d'espèces.

I.5 Organisation des données

Le fonctionnement de Genstyle nécessite en permanence le classement d'informations et l'utilisation de données. Les données sont stockées dans des tables relationnelles. Deux types de données sont nécessaires : les données biologiques et les données relatives aux utilisateurs. Les données biologiques consistent essentiellement en séquences génomiques auxquelles sont rattachées des informations associées. La signature génomique étant une caractéristique relative à l'espèce, nous avons pris le parti d'organiser les séquences par espèce (figure I.1).

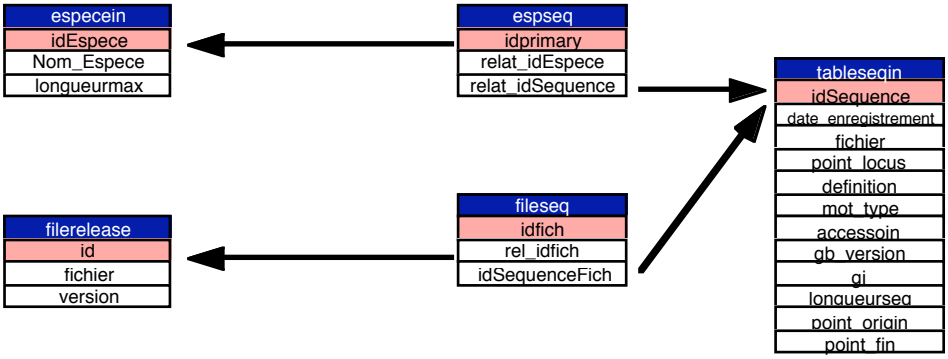

Figure I□/□Base de données sous-jacente à Genstyle. En bleu□le nom de la table, en rouge, la clé principale.

Au total, cinq tables sont nécessaires au fonctionnement de Genstyle□

- La table «□especein□. Cette table regroupe les informations sur les espèces. Elle est importante, en effet les signatures génomiques étant spécifiques de l'espèce, les séquences sont organisées par espèce. A chaque espèce sont associés son numéro d'identification, le nom et la longueur de la plus grande séquence disponible (ce qui donne une idée de la fiabilité de la meilleure signature disponible pour l'espèce).
- La table «□ableseqin□. Cette table regroupe les informations sur l'ensemble des séquences génomiques disponibles (numéro d'identification, type de séquence (ARN / ADN mitochondrial / Chloroplaste / ADN et divers), longueur, numéro d'accession Genbank, …). Cette base miroir de Genbank est cruciale dans le fonctionnement de Genstyle.
- La table «□espseq□ qui lie les séquences et leur espèce d'appartenance.
- La table «□ilerelease□ contient l'ensemble des fichiers Genbank. C'est dans cette table que sont stockées les séquences décrites dans la table «□ableseqin□.
- La table «□ileseq□ permet la liaison des tables «□ilerealease□ et «□ableseqin□.

I.6 Genstyle aujourd'hui

Genstyle est maintenant opérationnel et disponible au public à l'adresse http://genstyle.imed.jussieu.fr/. La version finale a été testée par des scientifiques (en particulier des biologistes), de façon à s'assurer que l'utilisation de Genstyle est intuitive, et permet d'obtenir les résultats qui intéressent les utilisateurs potentiels.

Genstyle a maintenant une portée nationale (Lespinats, 2005) et internationale (Fertil, 2005). Nous avons d'ailleurs depuis quelques mois des retours positifs sur ce travail (France, Allemagne, USA, etc …). Il est donc important de poursuivre la promotion de cet outil. Il serait également bon de mettre en place une analyse de la fréquentation du site pour évaluer son impact.

Annexe II : Liste de génomes utilisés

II.1 Liste de 43 procaryotes

Aeropyrum pernix
Agrobacterium tumefaciens C58
Aquifex aeolicus
Archaeoglobus fulgidus
Bacillus subtilis
Bacillus halodurans
Borrelia burgdorferi
Brucella melitensis
Buchnera sp APS
Campylobacter jejunii
Chlamydia muridarum
Clostridium acetobutylicum 824
Corynebacterium glutamicum
Deinococcus radiodurans
Escherichia coli
Halobacterium sp NRC-1
Helicobacter 26695pylori
Lactococcus lactis
Listeria innocua
M pneumoniae
Mesorhizobium loti
Methanococcus jannaschii
Mycoplasma genitalium
N MC58meningitidis
Nostoc sp. PCC 7120
Pasteurella multocida PM70
Pseudomonas aeruginosa PA01
Pyrobaculum aerophilum
Pyrococcus abyssi
S typhimurium
Sinorhizobium meliloti 1021
Staphylococcus aureus N315
Streptococcus pneumoniae
Sulfolobus solfataricus
Synechocystis sp PCC6803
Thermoplasma acidophilum
Thermotoga maritima
Treponema pallidum
Vibrio cholerae
Xylella fastidiosa
Yersinia pestis
haemophilus influenzae
ureaplasma urealyticum

II.2 Liste de 80 génomes complets

Organisme	Domaine du vivant	longueur du génome (en paires de bases)
Aeropyrum pernix	archéobactérie	1670334
Agrobacterium tumefaciens	bactérie	4916363
Aquifex aeolicus	bactérie	1551335
Archaeoglobus fulgidus	archéobactérie	2178400
Arabidopsis Thaliana (chr2)	eucaryote	19646744
Bacillus subtilis	bactérie	4214814
Bacillus halodurans	bactérie	4202353
Borrelia burgdorferi	bactérie	910681
Brucella melitensis	bactérie	3294921
Buchnera sp APS	bactérie	640681
Caenorhabditis elegans	eucaryote	2387915
Campylobacter jejunii	bactérie	1641480
Caulobacter crescentus	bactérie	4016947
Chlamydia muridarum	bactérie	997516
Chlamydia pneumoniae AR39	bactérie	1229784

Chlamydia pneumoniae CWL029	bactérie	1229940
Chlamydia trachomatis	bactérie	1042519
Chlamydophila pneumoniae J138	bactérie	1226564
Clostridium acetobutylicum 824	bactérie	3940880
Clostridium perfringens	bactérie	3031430
Corynebacterium glutamicum	bactérie	3309400
Deinococcus radiodurans	bactérie	3257648
Drosophila melanogaster (Chr3)	eucaryote	27509812
Escherichia coli	bactérie	4638690
Escherichia coli O157-H7	bactérie	5468733
Escherichia coli O157-H7_1	bactérie	5498450
H37RV Mycobacterium tuberculosis	bactérie	4411529
haemophilus influenzae	bactérie	1830021
Halobacterium sp NRC-1	archéobactérie	2014239
Helicobacter 26695pylori	bactérie	1668040
Helicobacter J99pylori	bactérie	1643740
Human (Chr 22)	eucaryote	33476901
Lactococcus lactis	bactérie	2365589
Leishmania major CHR1	eucaryote	268984
Listeria innocua	bactérie	3011208
Listeria monocytogenes	bactérie	2944528
Mus musculus	eucaryote	9122522
M pneumoniae	bactérie	816394
M thermoautotrophicum	archéobactérie	1751377
Mesorhizobium loti	bactérie	7036074
Methanococcus jannaschii	archéobactérie	1664957
Mycobac tuberculosis CDC1551	bactérie	4403661
Mycobacterium leprae TN	bactérie	3268203
Mycobacterium tuberculosis	bactérie	4411529
Mycoplasma genitalium	bactérie	580073
Mycoplasma_pulmonis	bactérie	963879
N MC58meningitidis	bactérie	2272351
N Z2491meningitidis	bactérie	2184406
Nostoc sp. PCC 7120	bactérie	6413771
Pasteurella multocida PM70	bactérie	2257487
Plasmodium falciparum CHR3	eucaryote	1060085
Pseudomonas aeruginosa PA01	bactérie	6264403
Pyrobaculum aerophilum	archéobactérie	4444860
Pyrococcus abyssi	archéobactérie	1765118
Pyrococcus furiosus	archéobactérie	1908253
Pyrococcus horikoshii	archéobactérie	1738505
Ralstonia solanacearum	bactérie	3716413
Rhodobacter capsulatus	bactérie	1315108
Rickettsia conorii Malish 7	bactérie	1268755
Rickettsia prowazekii	bactérie	1111523
S typhimurium	bactérie	4857432
Salmonella Typhi CT18	bactérie	4809037
Sinorhizobium meliloti 1021	bactérie	3654135
Staphylococcus aureus N315	bactérie	2813641
Staphylococcus_aureus_Mu50	bactérie	2878134
Streptococcus pneumoniae	bactérie	2160837
Streptococcus pneumoniae R6	bactérie	2038615

Streptococcus pyogenes	bactérie	1852441
Sulfolobus solfataricus	archéobactérie	2992245
Sulfolobus tokodaii	archéobactérie	2694765
Synechocystis sp PCC6803	bactérie	3573470
Thermoplasma acidophilum	archéobactérie	1564906
Thermoplasma volcanium	archéobactérie	1584854
Thermotoga maritima	bactérie	1860725
Treponema pallidum	bactérie	1137944
ureaplasma urealyticum	bactérie	750665
Vibrio cholerae	bactérie	4033427
Xylella fastidiosa	bactérie	2679305
Yeast genome	eucaryote	11935987
Yersinia pestis	bactérie	4653728