Université Louis Pasteur
4, rue Blaise Pascal
67000 Strasbourg

DESS C.C.I.
Compétence Complémentaire en Informatique
2001-2002

Développement d'un logiciel bioinformatique d'analyse de groupes de protéomes et de génomes : Application aux bactéries pathogènes impliquées dans les diarrhées humaines

Jean Muller

Responsable de stage : Olivier Poch

Laboratoire de Biologie et de Génomique Structurales de l'IGBMC
1, rue Laurent Fries BP 10142
67404 ILLKIRCH CEDEX

Le travail présenté dans ce rapport a été réalisé au sein du Laboratoire de Biologie et de Génétique Structurales dirigé par Dino Moras à l'Institut de Génétique et de Biologie Moléculaire et Cellulaire.

Je tiens à remercier Olivier Poch, mon maître de stage, pour m'avoir fait confiance et permis d'intégrer une équipe dynamique. Merci aussi pour son enthousiasme quotidien et ses longues réflexions qui m'ont ouvert l'esprit sur des concepts biologiques insoupçonnés.

Merci à Dino Moras et Jean Claude Thierry pour m'avoir permis d'effectuer ce stage au sein du laboratoire, me faisant bénéficier d'un environnement scientifique riche.

Un énorme merci à Raymond Ripp pour m'avoir guidé à travers son GScope et appris les joies de la programmation en Tcl dans la bonne humeur.

Et merci à toute l'équipe : Odile Lecompte, Frédéric Chalmel, Serge Uge, Annaïck Carles, Yann Brelivet, Julie Thompson, Véronique Prigent, Arnaud Muller, Nicolas Wicker, Luc Moulinier, Frédéric Plewniak et Laurent Bianchetti, pour leur accueil lors de mon arrivée, leur amitié et pour tout le temps qu'ils ont passé sans compter à me former, à répondre à mes questions et à me supporter.

Sommaire

I. Introduction

Depuis bientôt 25 ans, la génomique ou l'étude des génomes s'est considérablement enrichie. En effet avec l'aboutissement de nombreux programmes de séquençage, la qualité (séquences complètes et non plus parcellaires) et la quantité des séquences présentes dans les banques nucléiques et protéiques ne font que croître. Les chercheurs disposent maintenant de plus de 90 génomes complets, du plus petit (ex : *Mycoplasma genitalium* 580 070 paires de bases) au plus gros (ex : *Homo Sapiens* $3e^{9}$ paires de bases).

Ceci laisse entrevoir des perspectives intéressantes quant au développement immédiat et à plus long terme de cette discipline et ses applications variées.

On parle d'ailleurs de véritable révolution génomique. Cette période est génératrice de nouvelles idées, de nouveaux concepts qui permettent d'entrevoir le travail non plus sur un seul génome, mais de combiner les approches et les raisonnements sur plusieurs génomes simultanément.

Des outils de traitement massif, et automatiques, sont apparus, permettant de supprimer les tâches redondantes et d'apporter de nouvelles approches. Dans ce contexte, l'informatique appliquée à la biologie trouve tout son sens et montre alors toute sa puissance. La compréhension et la maîtrise des problèmes et méthodologies à la fois biologiques et bioinformatique sont alors indispensables pour pouvoir implémenter les solutions les plus efficaces et les mieux adaptées aux problèmes posés.

Les premiers sujets de telles études sont bien sûr orientés vers les organismes modèles majeures comme l'Homme (*Homo sapiens*), la souris (*Mus musculus*), le riz (*Oryza sativa*), la levure (*Saccharomyces cerevisiae*) ou les bactéries pathogènes.

Par l'analyse fonctionnelle (organisation, évolution) des génomes, les études de présence et d'absence de protéines (les génomes sont complets) et la phylogénie, la génomique va nous apporter un certain nombre de réponses aux problèmes biologiques actuels comme par exemple des maladies encore incurables. Appliquer à des organismes dits « dangereux » ou pathogènes ces analyses permettront en outre de déterminer des cibles thérapeutiques ou des mécanismes permettant d'empêcher certaines maladies de se développer.

La disponibilité croissante du nombre de génomes complets nous permet d'envisager et de réaliser des analyses différentielles entre ces génomes en terme de gain ou de perte de gènes.

J'ai donc été chargé de développer un outils capable de réaliser une étude comparative des génomes d'une série de bactéries pathogènes (DiaBac, Diarrheal Bacteria) ayant pour caractéristique commune, comme leur nom l'indique, de provoquer une diarrhée. Cette étude a pour but ambitieux de déterminer, par exemple, les protéines communes responsables, directement ou indirectement, de ce pouvoir pathogène. Elle prendra la forme d'un bilan de Présence/Absence appliquée sur les protéomes des bactéries DiaBac par rapport aux génomes complets disponibles.

II. Matériels et Méthodes

1. Environnement

a) Le laboratoire

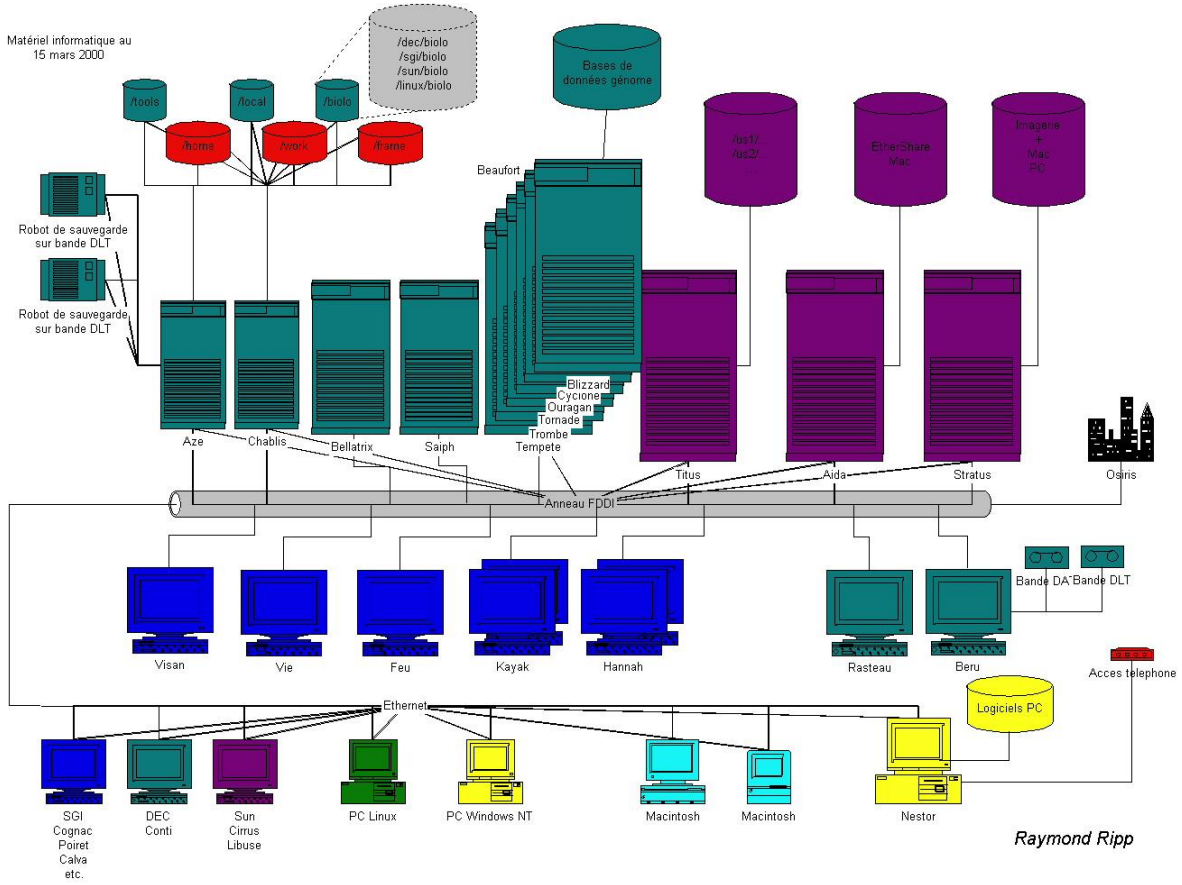
Le travail présenté dans ce rapport a été réalisé dans l'équipe de bioinformatique dirigée par Olivier Poch au sein du Laboratoire de Biologie et Génomique Structurales dirigé par Dino Moras à l'Institut de Génétique et de Biologie Moléculaire et Cellulaire (IGBMC Illkirch-Graffenstaden Strasbourg).

L'IGBMC est un institut de recherche fondamentale qui se consacre à l'étude des génomes d'eucaryotes supérieurs et au contrôle de l'expression génétique au cours du développement embryonnaire et de la différenciation normale et pathologique. L'institut est à la fois un laboratoire propre du Centre National de la Recherche Scientifique (CNRS), une unité de recherche de l'Institut National de la Santé et de la Recherche Médicale (INSERM) et un centre de recherche de la Faculté de Médecine de l'Université Louis Pasteur (ULP).

L'équipe de bioinformatique travaille sur l'analyse des génomes et des gènes en intégrant des données de séquences, de structures et des données bibliographiques. Elle s'attache tant à l'étude de familles impliquées dans la transcription que dans le développement de logiciels pour l'analyse, la gestion et la visualisation des informations autour des gènes et des génomes. Au sein de l'équipe ou Groupe de Bioinformatique et Génomique (BioIG), la plupart des personnes ont une formation soit à dominante biologique et structuraliste ou à dominante informatique. Ces compétences permettent à l'équipe grâce à une synergie et une cohésion forte de développer des concepts et des logiciels novateurs, appliqués à la biologie.

b) L'environnement informatique

Le système informatique de l'IGBMC est représenté par le schéma suivant :



Le réseau fonctionne en 100 Mbits/s sur un étage et en 1 Gbits/s entre les étages de l'institut.

Le parc informatique est composé pour les postes de travail de 300 MACs, de 100 PCs et d'une quinzaine de stations graphiques Silicon (SGI) dédiées à la modélisation moléculaire. Les différents serveurs disponibles tournent tous sous un système Unix et sont de fabrication SUN ou Compaq. La sauvegarde des données est réalisée par une machine Compaq connecté à 3 robots de sauvegarde pour une capacité totale de 3 To environ.

Les calculs de génomique sont exécutés sur le cluster Compaq baptisé Beaufort. Il est constitué de 6 machines de quadri processeurs (Alpha ev67) disposant chacune de 4 Go de mémoire vive, reliées entre elles par Memory Chanel. Ces machines fonctionnent sous le système Tru64UNIX et disposent d'un espace disque global de 500 Go.

L'importance de la gestion du système informatique est primordiale car de nombreuses cultures informatiques se croisent au sein de l'institut (PC, Mac, systèmes UNIX…) et doivent interagir.

2. Les banques de données

Les banques de données sont de plus en plus abondamment utilisées dans les études biologiques et génomiques. L'importance et la taille des données, qui y sont contenues, ne font que croître de jour en jour. Cette information et les programmes disponibles pour les traitées sont régulièrement remaniés. C'est un fait auquel il faut pouvoir s'adapter quand on veut y puiser des informations et dont il faut tenir compte lors de l'analyse et la compréhension de ces données.

Il existe de très grands centres serveurs tels que le NCBI (aux Etats-Unis), le DDBJ (au Japon) et l'EMBL (en europe) offrant l'accès libre et gratuit aux données et aux programmes nécessaires à leurs tratement. Par ailleurs de multiples banques spécialisées sont accessibles comme, par exemple, FlyBase (spécialisée dans les séquences de mouches comme *Drosophila melanogaster*) ou encore la banque de séquences des récepteurs nucléaires.

a) Les banques généralistes de séquences

Il existe trois banques nucléiques l'EMBL, Genbank et la DDBJ. Elles ont la particularité, suite à des accords les liant, d'être chacune le miroir de l'autre. Elles sont bien sûr disponibles via Internet.

-EMBL: (European Molecular Biology Laboratory)
La base de données de nucléotides de l'EMBL est la première base de données de séquences nucléotidiques en Europe. La version actuelle contient, au 20 juillet 2002, 17 226 421 entrées.
http://www.embl.org/

-GenBank :(NCBI, Etats-Unis)
GenBank est une collection de séquences génétiques fournies par des laboratoires ou provenant de projets de séquençage à grande échelle. La version actuelle contient, au 20 juillet 2002, 17 755 878 entrées.
http://www.ncbi.nih.gov/

-DDBJ: (DNA Data Bank of Japan)
Basé à Mishima au Japon c'est la banque équivalente à l'EMBL et à Genbank qui est en charge de la collecte des séquences pour la zone asiatique. La version actuelle contient, au 20 juillet 2002, 17 260 693 entrées.
http://www.ddbj.nig.ac.jp/

Les banques de données protéiques dont voici quelques exemples :

-SwissProt :
SwissProt est une base de données de séquences protéiques annotées créée en 1986. La version actuelle contient, au 20 juillet 2002, 111298 entrées.
http://www.ebi.ac.uk/ebi_docs/swissprot_db/swisshome.html

-SpTrEMBL :
Cette banque est constituée à partir de la banque nucléique de l'EMBL traduite en séquence protéique. Ces séquences une fois annotées seront intégrées dans la banque SwissProt.

-**PDB** (Protein Data Bank) :
La PDB est la principale banque de structures disponibles sur le Web. Elle renferme au 30 juillet 2002 prés de 18800 structures de macromolécules biologiques déterminées expérimentalement par rayons X ou par Résonance Magnétique Nucléaire.
http://www.rcsb.org/pdb/

-Des génomes complets :
96 organismes, dont le génome a été complètement séquencé, sont disponibles. On compte à ce jour 16 génomes d'archaea, 68 de bactérie et 12 d'eucaryote.

Les différentes banques utilisées à travers cette étude, lors de recherche de séquences sont Swissprot, SpTrEMBL et la PDB pour les séquences protéiques. La banque locale des génomes complets a permis de réaliser des recherches sur des séquences nucléiques. GenBank a permis de récupérer les séquences des génomes complets finis et disponibles pour la communauté scientifique.

b) L'accès aux informations contenues dans les banques

L'accès aux données est possible sur les sites Web de chacunes des banques par accès direct, par mot clé ou par recherche d'homologie de séquence. Ces données relatives à une séquence sont la séquence elle-même bien sûr mais aussi les informations qui lui sont liées, telles que, par exemple, l'organisme, la fonction si elle est connue, les références bibliographiques et les liens vers d'autres bases de données. (cf ANNEXE 1)

SRS est un des programmes les plus utilisés et permet un accès par les champs textuels d'une entrée dans une banque alors que BLAST utilise le champs séquence.

-SRS (Sequence Retrieval System):

SRS a été développé depuis 1990 dans les grands centres de bioinformatique (EMBL, EBI…) et enveloppe à ce jour environ 450 banques de données (publiques ou privées) répartie sur 35 sites Web dans 24 pays. Il assure l'indexation des informations contenues dans les champs des différentes banques et permet par des références croisées et liens dits transitifs d'atteindre toutes les banques. Une version est ainsi installée à l'IGBMC pour les banques de données disponibles en local (SwissProt, SPTrembl, Genbank…..). Cette disponibilité locale

permet d'une part une gestion individualisée, puisque selon les besoins l'intégralité d'une banque n'est pas toujours nécessaire. D'autre part, le travail des chercheurs en est grandement facilité et accéléré (cf ANNEXE 2).

-BLAST: (Basic Local Alignment Search Tool)

La constitution des banques de séquences permet aux chercheurs de comparer leurs séquences à celles de la banque choisie. Leur but est de trouver une séquence similaire (homologue ou non homologue).

Des séquences homologues descendent d'un ancêtre commun, elles sont le résultat de la duplication d'un gène. Selon le moment où cette duplication est apparue, soit avant ou après la spéciation de cet ancêtre, on parlera respectivement d'orthologue (homologue dans un autre organisme) ou de paralogue (homologue dans le même organisme). On peut ainsi comparer des séquences d'ADN (enchaînement d'acides nucléiques) ou des séquences protéiques (enchaînement d'acides aminés) à des séquences présentes dans les banques nucléiques et/ou protéiques.

L'intérêt de telles recherches réside dans le fait de pouvoir assigner, par comparaison aux séquences déjà identifiées, des fonctions à des parties de protéines ou des protéines complètes.

Les protocoles existants déterminent les meilleurs alignements locaux de séquences deux à deux issues d'une banque de séquences en leur attribuant un score.

Le programme BLAST [Altschul *et al*., 1990] est une série de logiciels (cf Tableau 1) qui permettent de rechercher dans les banques de données (nucléiques et/ou protéiques) des séquences homologues à une séquence test appelée « Query ». Il détermine ainsi des zones similaires de taille maximale, ces zones sont appelées HSP (High-scoring Segment Pair).

Un score est ainsi calculé pour chaque séquence à partir de la longueur et la composition de la séquence, de la taille de la banque et de la matrice de score utilisée, c'est l'« Expect Value » ou « E » qui permet de classer les séquences homologues entre elles. Un « E » significatif doit être le plus petit possible, car il reflète le nombre d'alignements qui atteindraient le même score par hasard. C'est pourquoi la limite significative biologiquement parlant, est fixée à 10^{-3}.

Le fichier de sortie d'un Blast se compose d'une liste de séquences de protéines ou « Hits » classée de façon croissante à partir de la séquence dont l'« E » est le plus faible. On peut ainsi obtenir les alignements des HSPs de la séquence « Query » avec chacune des cibles trouvées dans la ou les banques sélectionnées (cf ANNEXE 3).

3. L'alignement multiple

L'alignement multiple permet de comparer plusieurs séquences en même temps, sur toute leur longueur. On parle de MACS (Maximum Alignement of Complete Sequences) qui permet d'exploiter le maximum d'information disponible dans une famille de protéines en fournissant une étude suivant l'évolution et les évènements l'ayant perturbée.

C'est un outil précieux et essentiel pour l'étude d'une famille biologique. Il occupe un rôle central dans l'ère post-génomique en permettant :

- Identifier des gènes, assigner des fonctions
- Corriger et valider les séquences
- Étudier la phylogénie
- Décrypter l'organisation modulaire des protéines
- Révéler des contraintes évolutives sur certains acides aminés.
- Rechercher des protéines homologues dans les bases de données

L'alignement multiple est un pilier de la modélisation moléculaire

4. Programmes développés au laboratoire

Le laboratoire a mis en place une colonne de programmes destinée à l'analyse complète de séquences à partir d'un Blast. Cette colonne vertébrale ou « Pipe-Align » s'articule autour des programmes Ballast, DbClustal, NorMD, Secator, et Ordali. Une plate-forme de génomique a également été développée au travers de GScope qui a, en plus de ses particularités propres, une vocation intégrative de toute cette démarche (cf Figure 1).

Ballast [Plewniak *et al*., 2000]**:**
C'est un programme traitant les sorties de Blast et permettant de déterminer une liste de motifs localement conservés (LMS : Local Maximum Segment) le long de la séquence.
(http://igbmc.u-strasbg.fr:8080/ballast.html)

DbClustal [Thompson *et al*., 2000]**:**
C'est un programme d'alignement multiple de séquences protéiques globales à partir des alignements les plus significatifs d'une recherche BLAST. Il est basé sur l'algorithme de ClustalW tout en intégrant la spécificité d'alignement locaux obtenus par Blast et son post traitement (Ballast). Ballast fournit à DbClustal au travers des LMS, des points d'ancrage entre les différentes séquences à aligner. DbClustal permet donc de réunir les avantages des alignements locaux et globaux pour fournir le meilleur alignement possible tout en gardant une rapidité de calcul importante.
(http://igbmc.u-strasbg.fr:8080/DbClustal/dbclustal.html)

NorMD [Thompson *et al*., 2001]**:**
C'est une fonction objective qui permet de fournir un score, reflet de la qualité d'un alignement multiple. Il contribue donc à l'amélioration des alignements multiples et à leur validation.
(http://igbmc.u-strasbg.fr:8080/NorMD/normd.html)

Secator [Wicker *et al*., 2001]**:**
C'est un programme de regroupement de séquences à partir d'un alignement multiple de séquences complètes. Il est basé sur la méthode de classification hiérarchique ascendante. A partir d'un alignement de séquences, il effectue un arbre phylogénétique et regroupe les séquences homologues d'un sous-arbre en calculant les valeurs de dissimilarité sur chaque nœud. Les groupes ainsi constitués partagent des domaines ou des régions conservées au sein de leurs séquences.

DPC [Wicker *et al*.in press]**:**
Density of Points Clustering est un programme de clustering qui détermine automatiquement le nombre de groupes d'un ensemble de données. Pour cela il permet de classer des points en fonction de leur densité de répartition. Il utilise un algorithme original partant de l'idée qu'un cluster doit être divisé en deux à partir du moment où entre ses deux clusters la rareté des points est plus importante comparé à la densité des points des clusters voisins. Ceci est appliqué autant de fois que c'est nécessaire.
Il a été développé dans le but d'organiser la quantité importante de données issues de l'analyses différentielles des transcriptomes.

Ordali :
C'est un logiciel exploitant les alignements multiples qui permet de déterminer des résidus conservés (invariants), appelés « points Ordali » dans des groupes de séquences. Ordali permet de représenter graphiquement l'alignement et les « points Ordali », ainsi que de basculer à tout moment sur la structure tridimensionnelle pour les visualiser en 3D.

5. GScope

GScope est un programme permettant de créer une banque de données pour une ou plusieurs séquences de protéines ou pour un génome complet. De plus, GScope comprend tout un ensemble de routines permettant de visualiser, d'analyser et de gérer les informations présentes dans la banque de donnée d'un projet.

Lors de la création d'une base de données GScope (cf Figure 2) à partir d'un génome, les protéines ou ORFs (« Open Reading Frame », cadres ouverts de lecture) du génome seront prédites à partir de l'ADN correspondant au génome en faisant appel à un programme extérieur (Glimmer) puis représentées graphiquement dans une fenêtre appelé « Board » (cf ANNEXE 4 et 5).

Ce dernier permet à l'utilisateur de se déplacer naturellement au milieu de son génome et lui apporte une facilité de traitement et de présentation de l'information disponible. D'autres éléments génétiques, tels que les tRNAs (ARN transfert) ou rRNA (ARN ribosomaux) seront également prédits au moyen de programmes extérieurs.

Une fois cette première étape réalisée, les protéines prédites vont être utilisées pour réaliser automatiquement des recherches de séquences dans les banques. Ces recherches vont constituer la banque de donnée regroupant les résultats des recherches BlastP dans les banques peptidiques (SwissProt, SpTrEMBL et PDB), les résultats des recherches tBlastN contres les banques des génomes complets et des BlastXs sont faits pour les régions

intergéniques.. Finalement, des alignements multiples des séquences similaires détectées dans les banques sont réalisés pour chaque protéines et rendus disponibles

L'ensemble de ces données déduites des programmes intégrés (Ballast, DbClustal, DPC…) ainsi qu'un certain nombre d'analyse automatique sur les résultats obtenus vont constituer la banque de données GScope caractéristique d'un génome.

Dans le cadre de ce projet nous avons utilisé un certain nombre des fonctionnalités de Gscope. Celles-ci sont indiquées en rouge sur la Figure 2

a) Présence/Absence

La Présence/Absence telle qu'elle est définie dans GScope répertorie les protéines présentes ou non dans l'alignement multiple final des meilleurs hits du BlastP. à raison d'un pourcentage d'identité supérieure ou égal à 20% qui correspond à la définition de l'homologie.

b) HDA (Homolog Detection Agreement)

Le bilan HDA est basé sur l'analyse des résultats de deux recherches Blasts effectués par GScope à savoir tBlastN et BlastP et permet de définir des bilans de Présence/Absence. En effet, les recherches BlastP à partir d'une protéine permettent de détecter la présence d'une protéine similaire dans le protéome d'un organisme tandis que la détection dans une recherche tBlastN indique la présence d'une protéine similaire potentielle dans le génome d'un organisme. Une différence entre les deux (classiquement une présence dans le tBlastN et une absence dans le BlastP) permet de mettre en évidence des erreurs éventuelles du programme GLIMMER aboutissant à des protéines non prédites et donc absentes dans le protéome de l'organisme considéré.

Cette procédure HDA permet donc de savoir si la protéine étudiée possède une protéine similaire dans un autre organisme ou pas. Cette fonction est accessible dans GScope à travers une coloration des Orfs du génome.

Code couleur des différentes catégories possibles : HDA

Vert :	Présents dans tBlastN (dans le génome) et dans BlastP (dans le protéome)
Rouge :	Absents dans tBlastN (dans le génome) et dans BlastP (dans le protéome)
Orange :	Variable
Jaune :	Présent dans tBlastN et absent dans BlastP (=protéine non créée)

On peut choisir au moyen d'un menu radio les organismes à inclure, à rejeter et à ignorer de la coloration.

c) Croisement des informations (café des sciences)

Le projet que nous voulons réaliser nécessite de traiter des informations liées à différents génomes et de pouvoir interroger et utiliser des procédures de GScope sur des génomes différents. Cependant, une fois la base de donnée crée, en exécutant GScope, nous ne pouvons accéder aux informations que d'un seul génome..

Pour gérer cela un stagiaire du Dess C.C.I. a implémenté un serveur appelé « café des sciences » à partir duquel on peut lancer des requêtes à plusieurs GScope distincts. Celles-ci sont lancées à un « animateur de café » qui gère et interroge ses « savants » spécialistes dans un domaine (par exemple pour un génome en particulier). On peut ainsi exécuter des procédures externes dirigées vers d'autres génomes répertoriés par GScope.

d) Validation du codon initiateur (start codon)

Ce programme utilisé dans la validation des séquences protéiques permet grâce à l'utilisation et l'intégration de l'alignement multiple obtenu par DbClustal et du 'clustering' effectué par Secator qui regroupe les séquences les plus proches entre elles, de tester la validité du codon initiateur (le premier) d'une protéine P. Le programme compare les séquences du groupe Secator de la protéine P pour détecter la présence d'une extension dans la protéine P. Si une extension existe au début de la protéine P (partie appelée N-terminale), alors le programme vérifie si un autre codon initiateur est présent dans la protéine P à une position équivalente ou proche de celle observée dans les autres séquences. Si tel est le cas, un autre codon initiateur est proposé pour la protéine P (cf figure 3).

6. Choix du langage de programmation : Tcl/Tk

Tcl est un langage de commandes interprété (mais compilé à la volée), multi plate-forme, puissant. Les programmes développés peuvent en effet être utilisés sous la majorité des UNIX, Linux, Windows et MacOS sans aucun changement du programme.

Les variables employées sont non typées, et considérées comme des chaînes de caractères ou des listes (ensemble d'éléments ordonnés pouvant également être des listes). Il permet de manipuler des chaînes de caractères, des tableaux qui permettent l'adressage associatif (par le contenu) et des expressions régulières très facilement.

Par ailleurs l'extension graphique Tk est particulièrement complète et facile d'utilisation, permettant ainsi de créer des interfaces graphiques rapidement.

Ce langage faisant également parti de ceux les plus utilisés au laboratoire et notamment pour GScope, il a donc paru naturel de l'adopter pour le projet.

III. Le Projet

1. Pourquoi ? Sur Qui ?

Mon projet portait sur la détermination des protéines communes, s'il en existe, impliquées dans les diarrhées d'origines bactériennes. Pour cela on se propose de réaliser une étude comparative des protéomes de bactéries impliquées dans ce processus. Cette étude revêtant un côté redondant et inaccessible sans un traitement informatique, j'ai donc développé un protocole et les outils informatiques nécessaires pour la réaliser.

Les bactéries pathogènes que nous avons intégrées au projet sont les suivantes :

DiaBac
Available complete genomes of BACteria that cause DIArrhea

Organisme	Souche	Nom dans GScope
Campylobacter jejuni	NCTC 1116	CJEJ
Escherichia coli	K12- MG1655	ECOL
Escherichia coli	O157:H7	ECOH
Helicobacter pylori	26695	HPYL
Helicobacter pylori	J99	HPYJ
Staphylococcus aureus	N315	SAUR
Salmonella enterica	CT18	SENT
Salmonella typhimurium	NCTC 1116	STYM
Vibrio cholerae	serotype O1, Biotype El Tor, N16961	VCHO

Elles ont été choisies selon trois critères à savoir i) provoquer une diarrhée, ii) se développer dans le système digestif de l'Homme et iii) avoir un génome disponible complètement séquençé.

La pathogénicité de ces bactéries est liée à des ensembles de gènes connus et regroupés dans les génomes sous la forme d'îlots, ce sont les « îlots pathogènes ». On connaît entres autres le VPI (*Vibrio cholera* pathogenic island), les SAPIs (*Staphylococcus aureus* pathogenic island), les cag pathogenic island (*Helicobacter pylori*) et les SPIs (*Salmonella* pathogenic island) et sont souvent spécifiques de chaque organisme (cf figure 4).

Cette étude vise donc les protéines directement responsables de la diarrhée, souvent localisées dans les îlots pathogènes, et les protéines indirectement responsables qui peuvent être présentes dans plus d'organismes.

En effet ces dernières possèdent des fonctions, susceptibles d'être partager par des organismes qui ne sont pas forcément pathogènes, telles que la faculté de coloniser un milieu (ex : l'intestin), la survie à des pH extrêmes, un complexe de sécrétion des protéines (telles que des toxines), la faculté d'échapper aux défenses du système immunitaire ou une vitesse de croissance élevée, etc.

Le module informatique devra automatiser un certain nombre de nouvelles analyses, être réutilisable pour d'autres projets et permettre de poser d'autres questions tout en s'intégrant dans la plate-forme du laboratoire, à savoir GScope.

Déroulement :
La première étape consiste à effectuer l'analyse des génomes DiaBac. Leurs bases de données GScope doivent être mise à jour ou simplement créées (comme pour *Staphylococcus aureus* par exemple).
La deuxième étape consiste en la validation des protéomes et les bilans de Présence/Absence proprement dits.
Enfin la dernière étape comprendra l'analyse des résultats obtenus.

2. Mise à jour des génomes

Pour que le projet puisse débuter sur une base saine, il est préférable que les éléments de travail soient tous au même niveau de mise à jour.

Les banques de données évoluent tellement rapidement, qu'une comparaison entre génomes, pour lesquels les recherches dans les banques de séquences ont été faites à plusieurs mois d'intervalle, risque d'être mal interprété. Il faut par exemple que les recherches dans les banques de séquences (Blast) soient relancées. C'est pourquoi le plus gros du travail au début fut de remettre à jour les bases de données GScope de nos génomes.

Cette réactualisation requiert beaucoup de temps et d'attention. En effet, il faut environ trois journées complètes et sans rencontrer de problèmes pour réaliser la base de données GScope d'un seul génome. La mise en route de neuf génomes en simultanée demande de ce fait encore plus d'attention afin que toutes les données soient intègres et complètes.

Pendant la mise à jour des génomes et lié à la concurrence des neuf génomes, il m'a paru nécessaire de disposer d'un outil qui me permettrait de savoir à quel niveau du processus de création de la base GScope je me trouvais. Cet outil n'existant pas, j'ai donc développé à partir d'un schéma des dépendances fonctionnelles des données, un module intégré à GScope pour tester de façon automatique l'état d'avancement d'un projet. Ce programme nommé « Inventory Check » vérifie les différentes étapes de la mise en place d'une base de données GScope. Ainsi il teste l'existence de différents fichiers banques liés à cette base (par exemple les résultat des recherche par BlastP et tBlastN), leurs analyses, des fichiers de configuration nécessaires au bon déroulement des procédures. En visualisant par l'intermédiaire du panneau de contrôle (cf ANNEXE 6) ces points cruciaux, « Inventory Check » signale des fichiers manquants et indispensables, des liens et permet de mettre en avant les choses restant à faire. Toujours à partir de son interface « Inventory Check » permet par un simple clic, d'exécuter les procédures manquantes. Ce module permettra en outre l'automatisation « presque » complète de GScope.

Ceci m'a permis à la fois de comprendre comment fonctionnait GScope au plus profond de ses procédures et en même de temps de voir évoluer un tel programme jour après jour. Ce travail a nécessité l'adaptation du code aux nouveaux génomes lancés, la correction de bugs existant. Ce fut notamment l'occasion de mettre à plat certains principes de GScope.

3. Validation et études des protéomes

Dans le but de réaliser un bilan de Présence/Absence sur les protéomes de bactérie il se pose deux problèmes. Les programmes de prédiction automatiques des protéines d'un génome, comme Glimmer par exemple, peuvent d'une part, ne pas réussir à créer des protéines dont le cadre ouvert de lecture est pourtant présent dans le génome et d'autre part, créer des protéines trop longues ou trop courtes. Ces défaillances aboutissent à des protéomes ayant des protéines 'ratées' qui devront être créées ou des protéines mal définies qui devront être révisées. Ceci nécessite donc un traitement préalable.

De plus, les protéines trop longues induisent dans les recherches Blast des régions supplémentaires qui peuvent fausser leurs résultats et peuvent masquer d'autres protéines. Pour remédier à ce type de problème, j'ai utilisé le programme « StartCodonValidation », pour lequel ce travail à servi de cas test. J'ai créé à partir du génome de *Bacillus Subtilis,* un protéome artificiellement 'rallongé'. Pour ce faire, toutes les protéines ayant un codon initiateur alternatif précédent le codon initiateur défini dans la banque ont été rallongées à concurrence de 100 acides aminés au maximum. Le choix du génome de *Bacillus Subtilis* s'est justifié par l'existence de nombreuses protéines définies expérimentalement. On devrait donc, en appliquant sur son protéome étendu, retrouver les protéines originales. Ainsi sur les 1140 protéines étendues, notre programme a correctement corrigé 735 protéines et a détecté 148 autres protéines comme étant allongées mais en prédisant un codon initiateur incorrect. Ce programme ayant démontré ses capacités, je l'ai appliqué aux protéomes de DiaBac et les résultats sont représentés dans le Tableau 2 ci-contre

J'ai ainsi pu répertorier les protéines trop longues et les modifier. Ceci a concerné environ 4% de chaque protéome ce qui représente, par exemple pour Saur, 70 protéines.

Les protéomes DiaBac sont donc relativement « propres » et reflète le travail important effectué, par les chercheurs, sur ce type de bactérie.

Les protéines 'ratées' dans un protéome ou trop courtes peuvent être en partie traitées par l'étude des régions intergéniques présentes entre deux 'ORFs' (TROU dans GScope). Ces régions sont susceptibles de contenir des protéines trop petites pour être détectées par les programmes de prédiction ou des extrémités de protéines ratées.

Les recherches BlastX réalisées sur ces régions permettent à partir de la séquence nucléique du génome de révéler des séquences de protéines éventuelles (complètes ou non). Seules ont été prises en compte les recherches BlastX ayant révélé des séquences potentielles ayant un score significatif. Les résultats sont présentés dans le Tableau 3. On a ainsi pu, par exemple pour Hpyl, créer cinq protéines supplémentaires et en modifier 20 autres.

Une fois les bases GScope des génomes DiaBac mises à jour et leurs protéomes validés, nous avons pu réaliser les bilans de Présence/Absence des protéines par rapport à l'ensemble des génomes et protéomes connus (58 dans le cadre de cette étude).

4. Profil de Présence/Absence des protéines DiaBac

A partir des protéomes validés, nous voulons détecter si des familles de protéines homologues de DiaBac sont présentes dans l'ensemble des autres organismes complets disponibles.

Protocole HDA :

Afin de réaliser ceci nous disposons du protocole HDA (Homologue Detection Agreement) qui permet, à partir d'un seul génome, de caractériser la présence ou l'absence de protéines homologues dans l'ensemble des 58 génomes complets. Le HDA intègre les recherches dans le BlastP (présence dans le protéome prédit) et le tBlastN (présence dans le génome).

La présence dans les deux simultanément sera la démonstration la plus probante de la présence d'une protéine. A l'inverse, les protéines non déterminées lors de la prédiction du protéome d'un génome vont apparaître dans notre analyse comme des protéines présentes dans le tBlastN mais absentes dans le BlastP. Ces protéines seront alors créées automatiquement et intégrées dans le protéome du génome considéré.

Protocole X-HDA :

Mon projet couplant l'étude de plusieurs protéomes DiaBac, il nous a fallu développer et implémenter un nouveau protocole dans GScope afin de réaliser un bilan complet qui intègre les profils de présence/absence à partir des protéomes des neuf bactéries DiaBac et non plus d'un seul comme c'est le cas dans HDA. Ce protocole a été appelé X-HDA pour Crossed-linked Homologue Detection Agreement. Le X-HDA s'appuie également sur les recherches BlastP et tBlastN pour pouvoir classer les familles de protéines et va constituer le niveau d'intégration supérieur nécessaire pour déterminer l'existence de protéines éventuellement impliquées dans les phénomènes liés aux diarrhées d'origine bactérienne.

Pour cela, plusieurs problèmes sont apparus. Il fallait dans un premier temps constituer des familles de protéines homologues dans les organismes DiaBac puis rechercher et calculer les profils de présence de ces familles de protéines dans tous les génomes.

La constitution des familles de protéines se fait par deux recherches successives d'homologues parmi tous les Blasts des protéines DiaBac. En effet il arrive souvent qu'une seule protéine ne détecte pas l'ensemble de ses homologues. Pour palier à ce problème, nous avons constitué une première liste des numéros d'accès des homologues DiaBac déterminée à partir d'un organisme DiaBac, puis à partir de ces numéros d'accès nous avons été recherchés dans leur blasts respectifs si d'autres homologues DiaBac étaient présents (voir figure 5). On obtient ainsi une liste plus complète contenant tous les numéros d'accès des protéines homologues dans DiaBac. Cette liste constitue notre famille de protéines homologues qui sera utilisée pour réaliser notre bilan (BlastP et tBlastN).

D'un point de vue informatique, les fichiers de sortie blast ne contenant que les numéros d'accès (cf annexe), il nous a fallu rechercher pour chaque numéro d'accès, son organisme correspondant dans le cadre de requête de type SRS puis implémenter toute une série de tableaux et de listes définissant son appartenance à un protéome DiaBac et son ordre dans le blast. De plus, nous avons systématiquement intégré dans nos tableaux quel était l'organisme DiaBac responsable de l'introduction d'une protéine dans la liste finale des homologues d'une famille. Cette information, que nous n'avons pas utilisée dans le cadre de cette étude, devrait permettre de vérifier le degré d'homogénéité d'une famille. Ces recherches dans les Blasts d'autres génomes ont nécessité l'utilisation intensive

du « café des sciences » et donc le couplage effectif des génomes dans GScope au moyen de procédures permettant d'activer automatiquement plusieurs projets GScope simultanément.

Par la suite, pour chaque famille de protéines de DiaBac, nous avons calculé les pourcentages de Présence/Absence, à partir des deux types de Blasts, pour chaque organisme. On peut noter que les valeurs obtenues nous permettent également d'inverser notre regard et d'obtenir, si on le désire, un bilan individuel pour chacun des organismes ayant un génome complet. En combinant ces données, chaque bilan particulier d'un organisme par rapport à une famille pourra rentrer dans une catégorie définie par le X-HDA.

Code couleur des différentes catégories du X-HDA :

Vert :	Présent dans 100% des tBlastNs (dans le génome) et des BlastPs (dans le protéome).
Rouge :	Absent dans 100% des tBlastNs (dans le génome) et des BlastPs (dans le protéome).
Bleu :	Présent dans >50% des tBlastNs et >50% des BlastPs
Orange :	Présent dans <50% des tBlastNs et <50% des BlastPs
Jaune :	Au moins présent dans un tBlastN et toujours absent dans BlastP (=protéine non créée)

Ce bilan permet, comme le HDA, de mettre en avant des protéines 'ratées' dans un protéome prédit, mais que nous avons détectées dans son génome. La catégorie correspondante est la catégorie dite des bilans Jaunes. Cette fois-ci de part la nature du bilan X-HDA, elles sont définies pour plusieurs organismes et non plus pour un seul comme dans le HDA « simple ».
(ANNEXE 7 et 8)
La fonction HDA intégrée dans GScope, est accessible par une coloration des protéines sur le « board » du génome. Désormais et puisque le bilan X-HDA s'applique à plusieurs organismes en même temps, il a donc fallu mettre au point et implémenter une interface particulière pour pouvoir le visualiser et y travailler.

La constitution des familles, l'intégration du X-HDA et l'interface graphique de présentation et de travail du bilan X-HDA ont nécessité un développement informatique, que j'ai implémenté. Ces procédures sont intégrées dans GScope à tous les niveaux et sont accessibles facilement au travers des boutons sur la vitrine ou bien encore sur le board de GScope.
La création des familles de protéines homologues a été testée et validée sur plusieurs génomes de départ. Le calcul est relativement long mais il est lancé en automatique sur tout un génome. La conception et la réflexion faite sur les procédures permettent de les tester individuellement sur une seule protéine ou en global. Un souci particulier a été apporté lors de son implémentation, afin de conserver à chaque étape de son déroulement des informations capitales, pour pouvoir le relancer sans perte d'information et donc de temps.
(ANNEXE 4 et 5

5. Regroupement ou 'Clustering'

Ce type de bilan croisé génère beaucoup de données. On dispose d'un bilan X-HDA, pour autant de familles que de protéines dans le génome de départ (par exemple Vcho équivaut à 3799 familles) et pour les 58 génomes complets testés. L'analyse des données de ces bilans X-HDA nécessite donc, de par leur quantité, un traitement supplémentaire. Ce traitement prendra la forme d'un regroupement des données ou 'clustering'. Ce 'clustering' réalisé avec l'aide du programme DPC (Wicker and al in press) ou Density of Point Clustering vise à réduire la quantité des données en les rassemblant par groupe de signification proche. Il a été calculé sur les valeurs du bilan à savoir le pourcentage de présence dans le BlastP. Chaque famille de protéines est donc regroupée en fonction de la proximité des profils de présence/absence des protéines homologues présentes dans l'ensemble des génomes complets.

Les clusters pourront comprendre des familles de protéines dont le profil de présence ne sera pas strictement identique. Cela permet d'atténuer les erreurs de détermination de Présence/Absence (liées aux limitations de détection du programme Blast) et d'apporter une flexibilité accrue plus à même de rendre compte de la grande flexibilité de perte ou de gain de protéines observé dans les études de protéomes bactériens [Ettema *et al*., 2001].

Cette étape a exigé l'implémentation de procédures informatiques pour pouvoir adapter les données des bilans X-HDA au format d'entrée du programme DPC.
(ANNEXE 9)

Nous avons ainsi obtenu pour Vcho 167 clusters contenant de 5 à 819 familles. Ce dernier cluster comprend toutes les familles de protéines qui ne sont présentes que dans le génome de *Vibrio choléra*.

Le ‘clustering’ effectué sur la valeur du BlastP du bilan permet effectivement de regrouper des familles de protéines selon un profil relativement identique. Nous avons comparé les groupes obtenus par la nouvelle méthode et les groupes issus de la fonction Présence/Absence de GScope. Ceci nous a montré que les groupes définis avec DPC, à partir des bilans X-HDA, sont relativement cohérents du point de vue fonctionnel et qu’ils admettaient bien une relative plasticité favorisant le regroupement de profils non strictement identiques.

La dernière étape restant à faire est l’analyse des résultats expérimentaux de tous les clusters pour comptabiliser les protéines cibles et donner les conclusion biologiques adaptées aux organismes que nous avons étudiés.

6. Résultats majeurs

D’un point de vue informatique, mon travail sur le projet a permis d’ajouter des fonctionnalités supplémentaires à GScope notamment :

Inventory Check
Create Familly,
X-HDA

De plus, les fichiers de description des résultats des recherches Blast contiennent à présent, en plus des numéros d‘accès et des scores Blast (« l’Expect Value »), l’organisme associé et son appartenance éventuelle à un projet regroupant plusieurs génomes. Ceci accélérera énormément la création de familles.

L’utilisation intensive du café des sciences pour ce protocole a servi de cas-test pour déterminer les bugs encore présents. Notamment, lorsque plusieurs processus étaient exécutés en parallèle. La gestion interne de ces processus ‘entrants’ utilisait des appels à des variables globales accessibles par tous. Aucune sécurité ne permettait d’assurer la pérennité de ces variables, ceci avait pour conséquences l’arrêt simple du programme lors de partages par des processus parallèles. Nous avons traqué et corrigé ce bug afin de pouvoir tout simplement continuer le projet.

D’un point de vue biologique bien que bons nombres des données restent à étudier. On peut être confiant sur les résultats d’un tel protocole. Un premier regard sur les bilans permet de rapprocher certains organismes de DiaBac entre eux. On peut ainsi regrouper par exemple Vcho, Ecol/Ecoh, Sent/Stym d’un côté et Saur, Cjej, Hpyl/HpyJ de l’autre. On trouve plus souvent des protéines présentes en même temps dans un des deux groupes plutôt que dans les deux simultanément. Ces rapprochements ne sont pas seulement dus à leurs proximités phylogénétiques puisque Saur est assez éloigné des trois autres. Ce côté devra être approfondi pour confirmer ou infirmer cette constatation.

La mise en évidence des protéines ‘ratées’ dans les protéomes est primordiale. Grâce à la catégorie des jaunes, on a pu caractériser plus de 200 protéines ‘ratées’ dans d’autres génomes que DiaBac. Ces protéines seront créées et intégrées dans les génomes de leurs organismes respectifs. Certaines de ces protéines ‘ratées’ sont détectées à partir de familles de protéines contenues dans les îlots pathogènes. Elles peuvent donc contribuer à la pathogénicité d’autres organismes chez qui elles n’étaient pas encore identifiées. Ces protéines devront être étudiées plus avant.

D’autre part, le ‘clustering’ a permis de mettre en avant des groupes de protéines intéressants. En orientant la recherche en partant de clusters contenant certaines protéines pathogènes on a pu remontrer la présence des îlots pathogènes.

Le bilan montre cependant quelques limites liées aux sorties de Blast. Certaines protéines vont aboutir à un nombre très important de protéines similaires ayant des « Expect Value » significatifs. Ceci a pour effet de reléguer d’autres protéines similaires, susceptibles de rentrer dans nos critères, à des rangs trop élevés pour qu’ils soient détectés par GScope et comptabilisés dans nos calculs. Cette perte d’information se voit dans la catégorie des « Jaune » qui nécessiteront une attention particulière. Cependant c’est une chose qu’il faut garder à l’esprit mais qui n’est pas gênante.

Bien entendu, il faut encore apprendre à analyser plus précisément ces résultats expérimentaux pour pouvoir réellement tirer des conclusions. C’est une démarche qui demande du temps et une certaine habitude. Les outils pour le faire sont, en tous cas, à présent disponibles.

IV. Bilan

Ce stage de fin de Dess CCI a été une formidable expérience qui m'a permis de mettre en pratique les compétences informatiques acquises lors cette formation. L'implémentation de solutions à des problématiques qui touchent ma première compétence à savoir la Biologie m'ont bien entendu réjouit et faisait parti des objectifs majeurs que je m'étais fixés lors de la recherche du stage.

J'ai souvent entendu parler de bioinformatique mais la comprendre, l'utiliser et y participer ma beaucoup enthousiasmé.

Le sujet de stage proposé par Olivier Poch est passionnant avec des débouchées intéressantes. Les outils développés sont fonctionnels et commencent à donner des résultats, c'est donc une énorme satisfaction supplémentaire pour ma part.

De plus l'intégration dans une équipe si dynamique et si intéressante m'a réconcilié avec le monde de la recherche et me laisse entrevoir d'autres débouchés.

Au final ce stage m'a permis de renforcer mon goût pour l'informatique, les problèmes biologiques et me permettra de me lancer à fond dans la bioinformatique, une nouvelle passion. Je suis donc plus que satisfait de ces quatre derniers mois passés à l'IGBMC.

Bibliographie

F.Plewniak, J.D.Thompson and O.Poch ***Ballast: Blast post-processing based on locally conserved segments*** *Bioinformatics*, 2000, Vol. 16, No 9**,** 750-759

Thompson J.D., Plewniak F., Thierry J.-C. and Poch O. ***DbClustal :rapid and reliable global multiple alignments of protein sequences detected by database searches*** Nucleic Acid Res. 2000, Vol.28, No 15 2919-2926

Thompson JD, Plewniak F, Ripp R, Thierry JC, Poch O. ***Multiple Sequence Alignment Objective Function*** *Journal of Molecular Biology*, 2001, Vol.314, No 4 937-951

Pearson WR, Lipman DJ. ***Improved tools for biological sequence comparison (Fasta).*** Proc Natl Acad Sci U S A. 1988 Apr; 85(8): 2444-8.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. ***Basic local alignment search tool***. J Mol Biol. 1990 Oct 5; 215(3): 403-10.

Wicker N, Perrin GR, Thierry JC, Poch O. ***Secator: a program for inferring protein subfamilies from phylogenetic trees***. Mol Biol Evol. 2001 Aug; 18(8):1435-41

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. ***Gapped BLAST and PSI-BLAST: a new generation of protein database search programs***. Nucleic Acids Res. 1997 Sep 1; 25(17): 3389-402. Review

Ettema T, van der Oost J, Huynen M. ***Modularity in the gain and losses of genes applications for functions prediction***. Trends Genet. 2001 Sep 17(9): 487-7.

J. Parkhill*, B. W. Wren[2], K. Mungall*, J. M. Ketley[3], C. Churcher*, D. Basham*, T. Chillingworth*, R. M. Davies*, T. Feltwell*, S. Holroyd*, K. Jagels*, A. V. Karlyshev[2], S. Moule*, M. J. Pallenß, C. W. Pennk, M. A. Quail*, M-A. Rajandream*, K. M. Rutherford*, A. H. M. van Vliet S. Whitehead* & B. G. Barrell*. ***The genome sequence of the food-borne pathogen Campylobacter jejuni reveals hypervariable sequences***. Nature 2000 Feb 10, Vol 403: 685-88

Nicole T. Perna*[2], Guy Plunkett III[3], Valerie Burland[3], Bob Mau[3], Jeremy D. Glasner[3], Debra J. Rose[3], George F. Mayhew[3], Peter S. Evans[3], Jason Gregor[3], Heather A. Kirkpatrick[3], Gyorgy Posfaiß, Jeremiah Hackett[3], Sara Klink[3], Adam Boutin[3], Ying Shao[3], Leslie Miller[3], Erik J. Grotbeck[3], N. Wayne Davis[3], Alex Limk, Eileen T. Dimalantak, Konstantinos D. Potamousis[3]k, Jennifer Apodaca[3]k, Thomas S. Anantharaman, Jieyi Lin#, Galex Yen*, David C. Schwartz*[3]k, Rodney A. WelchI & Frederick R. Blattner*[3]. ***Genome sequence of Enterohaemorrhagic Escherichia coli O157:H7***
Nature 2001 Jan 25, Vol 409 529-34

Jean-F. Tomb, Owen White, Anthony R. Kerlavage, Rebecca A. Clayton, Granger G. Sutton, Robert D. Fleischmann, Karen A. Ketchum, Hans Peter Klenk, Steven Gill, Brian A. Dougherty, Karen Nelson, John Quackenbush, Lixin Zhou, Ewen F. Kirkness, Scott Peterson, Brendan Loftus, Delwood Richardson, Robert Dodson, Hanif G. Khalak, Anna Glodek, Keith McKenney, Lisa M. Fitzegerald, Norman Lee, Mark D. Adams, Erin K. Hickey, Douglas E. Berg, Jeanine D. Gocayne, Teresa R. Utterback, Jeremy D. Peterson, Jenny M. Kelley, Matthew D. Cotton, Janice M. Weidman, Claire Fujii, Cheryl Bowman, Larry Watthey, Erik Wallin, William S. Hayes, Mark Borodovsky, Peter D. Karp, Hamilton O. Smith, Claire M. Fraser & J. Craig Venter. ***The complete genome sequence of the gastric pathogen Helicobacter Pylori*** Nature 1997 Sep 25, Vol 388 539–47

Richard A. Alm, Lo-See L. Ling, Donald T. Moir, Benjamin L. King, Eric D. Brown, Peter C. Doig, Douglas R. Smith, Brian Noonan, Braydon C. Guild, Boudewijn L. deJonge, Gilles Carmel, Peter J. Tummino, Anthony Caruso, Maria Uria-Nickelsen, Debra M. Mills, Cameron Ives, Rene Gibson, David Merberg, Scott D. Mills, Qin Jiang, Diane E. Taylor, Gerald F. Vovis & Trevor J. Trust. ***Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen Helicobacter pylori*** Nature 1999 Feb 25, Vol 397 176-180

Makoto Kuroda, Toshiko Ohta, Ikuo Uchiyama, Tadashi Baba, Harumi Yuzawa, Ichizo Kobayashi, Longzhu Cui, Akio Oguchi, Ken-ichi Aoki, Yoshimi Nagai, JianQi Lian, Teruyo Ito, Mutsumi Kanamori, Hiroyuki Matsumaru, Atsushi Maruyama, Hiroyuki Murakami, Akira Hosoyama, Yoko Mizutani-Ui, Noriko K Takahashi, Toshihiko Sawano, Ryu-ichi Inoue, Chikara Kaito, Kazuhisa Sekimizu, Hideki Hirakawa, Satoru Kuhara, Susumu Goto, Junko Yabuzaki, Minoru Kanehisa, Atsushi Yamashita, Kenshiro Oshima, Keiko Furuya, Chie Yoshino, Tadayoshi Shiba, Masahira Hattori, Naotake Ogasawara, Hideo Hayashi, Keiichi Hiramatsu. ***Whole genome sequencing of meticillin-resistant Staphylococcus aureus*** The Lancet 2001 April 21, Vol 357 1225-40

J. Parkhill*, G. Dougan², K. D. James*, N. R. Thomson*, D. Pickard², J. Wain², C. Churcher*, K. L. Mungall*, S. D. Bentley*, M. T. G. Holden*, M. Sebaihia*, S. Baker*, D. Basham*, K. Brooks*, T. Chillingworth*, P. Connerton², A. Cronin*, P. Davis*, R. M. Davies*, L. Dowd*, N. White³, J. Farrar³, T. Feltwell*, N. Hamlin*, A. Haque², T. T. Hienß, S. Holroyd*, K. Jagels*, A. Kroghk, T. S. Larsenk, S. Leather*, S. Moule*, P. O¬ 'Gaora², C. Parryß, M. Quail*, K. Rutherford*, M. Simmonds*, J. Skelton*, K. Stevens*, S. Whitehead* & B. G. Barrell*. ***Complete genome sequence of amultiple drug resistant Salmonella enterica serovar Typhi CT18*** Nature 2001 Oct 25, Vol 413 848-52

Michael McClelland*, Kenneth E. Sanderson², John Spieth³, Sandra W. Clifton³, Phil Latreille³, Laura Courtney³, Steffen Porwollik*, Johar Ali³, Mike Dante³, Feiyu Du³, Shunfang Hou³, Dan Layman³, Shawn Leonard³, Christine Nguyen³, Kelsi Scott³, Andrea Holmes³, Neenu Grewal³, Elizabeth Mulvaney³, Ellen Ryan³, Hui Sun³, Liliana Floreaßk, Webb Millerß, Tamberlyn Stoneking³, Michael Nhan³, Robert Waterston³ & Richard K. Wilson³***. Complete genome sequence of Salmonella enterica serovar Typhimurium LT2*** Nature 2001 Oct 25, Vol 413 852-56

John F. Heidelberg*, Jonathan A. Eisen*, William C. Nelson*, Rebecca A. Clayton, Michelle L. Gwinn*, Robert J. Dodson*, Daniel H. Haft*, Erin K. Hickey*, Jeremy D. Peterson*, Lowell Umayam*, Steven R. Gill*, Karen E. Nelson*, Timothy D. Read*, Herve¬ Tettelin*, Delwood Richardson*, Maria D. Ermolaeva*, Jessica Vamathevan*, Steven Bass*, Haiying Qin*, Ioana Dragoi*, Patrick Sellers*, Lisa McDonald*, Teresa Utterback*, Robert D. Fleishmann*, William C. Nierman*, Owen White *, Steven L. Salzberg*, Hamilton O. Smith*², Rita R. Colwell³, John J. Mekalanosß, J. Craig Venter*² & Claire M. Fraser*. ***DNA sequence of both chromosomes of the cholera pathogen Vibrio cholerae*** Nature 2000 Aug 3, Vol 406 477-84

GOLD: Genomes Online Database is a World Wide Web resource for comprehensive access to information regarding complete and ongoing genome projects around the world.
http://wit.integratedgenomics.com/GOLD/

Glossaire

ADN
Abréviation d'Acide DésoxyriboNucléique. Macromolécule servant de support de l'information génétique chez la plupart des êtres vivants. Sa structure est universelle, seule la longueur de la molécule variant selon les espèces. L'ADN est constitué de deux brins complémentaires arrangés en double hélice, c'est le constituant des génomes et des chromosomes.
Un brin d'ADN est constitué d'une chaîne plus ou moins longue de désoxyribonucléotides (nucléotides comportant un sucre, le désoxyribose, et une base azotée choisie parmi la thymine T, la cytosine C, l'adénine A et la guanine G). Chaque base est appariée, sur le brin complémentaire, à une autre par des liaisons hydrogène, suivant deux couples possibles : A--T et G--C.

ADN complémentaire (ADNc ou cDNA)
ADN simple brin synthétisé à partir d'un brin d'ARN : il est obtenu après une réaction de transcription inverse d'un ARN mature et représente ainsi la copie de l'ARN. En biologie moléculaire, cette synthèse permet d'obtenir des copies d'ARN messager sous forme d'ADN : l'ADNc offre l'avantage d'être plus stable que la molécule d'ARNm et de pouvoir être stocké, copié et séquencé.

ARN
Abréviation de Acide RiboNucléique. C'est un acide nucléique à structure proche de celle de l'ADN qui est monobrin et moins stable. Le sucre présent dans les nucléotides de l'ARN est le ribose ; les quatre bases azotées sont l'adénine A, la cytosine C, la guanine G, l'uracile U remplaçant la thymine T présente dans l'ADN, dans les appariements avec A.
On rencontre différents types d'ARN dans les cellules (ARN messager ou mRNA, ARN de transfert ou tRNA, ARN ribosomal ou rRNA).
Voir Transcription, Traduction.

ARN messager (ARNm ou mRNA)
Molécule d'ARN, produite par la cellule à partir de l'ADN lors de la transcription.. La séquence des bases azotées portées par les nucléotides d'un ARNm porte le code nécessaire à la synthèse d'une protéine.

BLAST
Basic Local Alignment Search Tool

bp
Unité de mesure de la taille d'une séquence d'ADN en paires de bases. kb : kilo paires de bases = 1000 bp.

cDNA
Voir ADN complémentaire.

DNA
cf. ADN

Enzyme
Protéine produite par un être vivant pour catalyser des réactions biochimiques spécifiques, dans des conditions compatibles avec la vie.

EST
Expressed Sequence Tag. Une séquence EST est une étiquette (fragment d'une extrémité) d'un ADNc. Une séquence EST est donc une séquence de 100 à 150 nucléotides d'ADNc correspondant à une des extrémités d'un ARNm.

Eucaryote
Etre vivant dont le matériel génétique de chaque cellule est enfermé dans un noyau limité par une double membrane (champignons, levures, animaux, végétaux). L'existence de ce noyau cellulaire donne son nom aux acides nucléiques comme l'ADN et l'ARN.
Voir Procaryote.

Exon

Partie d'un gène eucaryote qui contient une séquence codante et qui est susceptible d'être conservée dans l'ARN lors de l'épissage.
Voir Intron.

Gène
Le gène correspond à un fragment de la molécule d'ADN, une séquence de nucléotides, qui comprend un promoteur de transcription suivi d'une séquence codant pour un ARN. Cet ARN peut avoir une fonction biologique ou coder pour une protéine.

Gènes orthologues
Gènes d'espèces différentes, divergence suite à un événement de spéciation conservent souvent la même fonction.

Gènes paralogues
Gènes d'une même espèce, divergence par duplication dans le même génome.

Génome
Totalité du matériel génétique chromosomique d'un organisme.

Homologie
Se dit si les séquences descendent d'un ancêtre commun.

Intron
Partie d'un gène excisée de l'ARN lors de la maturation en ARN messager (épissage). Les introns n'existent que chez les eucaryotes.
Voir Exon.

MACS
Maximum Alignement of Complete Sequences

mRNA
Voir ARN messager.

Nucléotide
Constituant élémentaire des acides nucléiques (ADN ou ARN), composé d'une base azotée (adénine A, guanine G, cytosine C, ou thymine T dans l'ADN ou uracile U dans l'ARN), associée à un ou plusieurs phosphates, et à un sucre (ribose dans l'ARN ou à un désoxyribose dans l'ADN).

ORFs
Open Reading Frame ou cadre ouvert de lecture
La phase ouverte (ORF, Open Reading Frame) est la région de l'ADN qui sépare deux codons STOP (donc potentiellement codante). Dans celle-ci, une séquence codante (CDS, région traduite en protéine) commence par un codon INITIATEUR (ATG), se termine par le codon STOP (TAA par ex) et est précédée d'un site de liaison aux ribosomes (RBS). Il y a six cadres de lecture possible dans la séquence car l'ADN est lu par codon c'est à dire une succession de trois bases.

Orthologue :
Voir Gènes orthologues

Paralogue :
Voir Gènes paralogues

Phylogénie
Figure en forme d'arbre traduisant les relations de parenté entre des organismes ou des molécules.

Procaryote
Se dit des cellules dépourvues de noyau cellulaire. Les êtres vivants procaryotes sont généralement unicellulaires comme les bactéries, les cyanobactéries, les archéobactéries. Leur matériel génétique est de l'ADN circulaire diffus dans le cytoplasme de la cellule.
Voir Eucaryote.

Protéine
Macromolécule organique composée essentiellement d'acides aminés reliés par la liaison peptidique. Seuls vingt acides aminés entrent dans la composition des protéines naturelles. Les protéines interviennent dans toutes les

réactions biochimiques des organismes, notamment grâce à leur structure spatiale. Une protéine est l'expression d'un gène qui permet sa synthèse au sein des cellules, au cours du processus de traduction des ARN messagers.

Protéome
Le protéome est le "complément protéique total du génome", c'est à dire l'ensemble des protéines exprimé par le génome d'une espèce donnée.

Ribosome
Complexe protéique intervenant dans la synthèse des protéines. Il effectue, avec l'aide des ARN de transfert (ARNt), la traduction en protéines des ARNm.

RNA
Voir ARN.

Séquence peptidique
Ordre des acides aminés sur la chaîne d'acides aminés formant une protéine. Chaque acide aminé est représenté par une lettre (V pour valine, L pour leucine…).

Séquence nucléique
Ordre des bases sur la chaîne linéaire de nucléotides formant un acide nucléique (ADN ou ARN) : chaque nucléotide étant représenté par l'initiale de la base qui le constitue (T pour thymine, C pour cytosine, A pour adénine et G pour guanine).

Shine-Dalgarno
Séquence consensus de fixation du ribosome sur le mRNA.

Similarité
Elle indique une ressemblance entre des séquences qui peut être mesurer en pourcentage d'identité de résidus.

Spéciation
Formation, au cours de l'évolution biologique, d'espèces distinctes, génétiquement isolées les unes par rapport aux autres.

Traduction
Processus au cours duquel le message génétique des ARN messagers est traduit en la séquence d'acides aminés codée par cet ARN. Le code pour chaque acide aminé successif est une combinaison de trois nucléotides (triplet ou codon) qui sont déchiffrés les uns après les autres au niveau du ribosome. A la fin de la traduction on obtient une protéine.

Transcription
Transfert de l'information génétique d'un gène, depuis une molécule d'ADN vers une molécule d'ARN.

Transcriptome
Ensemble des ARN messagers transcrits à partir du génome.

Annexe 1

Exemple d'une entrée de la banque SwissProt avec la séquence P54954.

Champs Texte

Numéro d'identification →
Accession number →
Dates →
Descriptions →
Organisme →
Classification →
Références bibliographiques →
Commentaires →
Liens vers les autres banques →

```
ID   YXEO_BACSU     STANDARD;      PRT;   249 AA.
AC   P54954;
DT   01-OCT-1996 (Rel. 34, Created)
DT   01-OCT-1996 (Rel. 34, Last sequence update)
DT   15-JUN-2002 (Rel. 41, Last annotation update)
DE   Probable amino-acid ABC transporter ATP-binding protein yxeO.
GN   YXEO OR LP9G.
OS   Bacillus subtilis.
OC   Bacteria; Firmicutes; Bacillales; Bacillaceae; Bacillus.
OX   NCBI_TaxID=1423;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   STRAIN=168;
RX   MEDLINE=98044033; PubMed=9384377;
RT   "The complete genome sequence of the Gram-positive bacterium Bacillus
RT   subtilis.";
RL   Nature 390:249-256(1997).
CC   -!- FUNCTION: PROBABLY PART OF A BINDING-PROTEIN-DEPENDENT
CC       TRANSPORT SYSTEM FOR AN AMINO ACID. PROBABLY RESPONSIBLE
CC       FOR ENERGY COUPLING
CC       TO THE TRANSPORT SYSTEM.
CC   -!- SUBCELLULAR LOCATION: MEMBRANE-ASSOCIATED (POTENTIAL).
CC   -!- SIMILARITY: BELONGS TO THE ABC TRANSPORTER FAMILY.
DR   EMBL; D45912; BAA08331.1; -.
DR   EMBL; Z99124; CAB15984.1; -.
DR   SubtiList; BG11891; yxeO.
DR   InterPro; IPR003593; AAA_ATPase.
DR   InterPro; IPR003439; ABC_transportr.
DR   Pfam; PF00005; ABC_tran; 1.
DR   ProDom; PD000006; ABC_transportr; 1.
DR   SMART; SM00382; AAA; 1.
DR   PROSITE; PS00211; ABC_TRANSPORTER; 1.
KW   Hypothetical protein; ATP-binding; Transport; Membrane;
KW   Complete proteome.
```

Champs Séquence

Séquence →

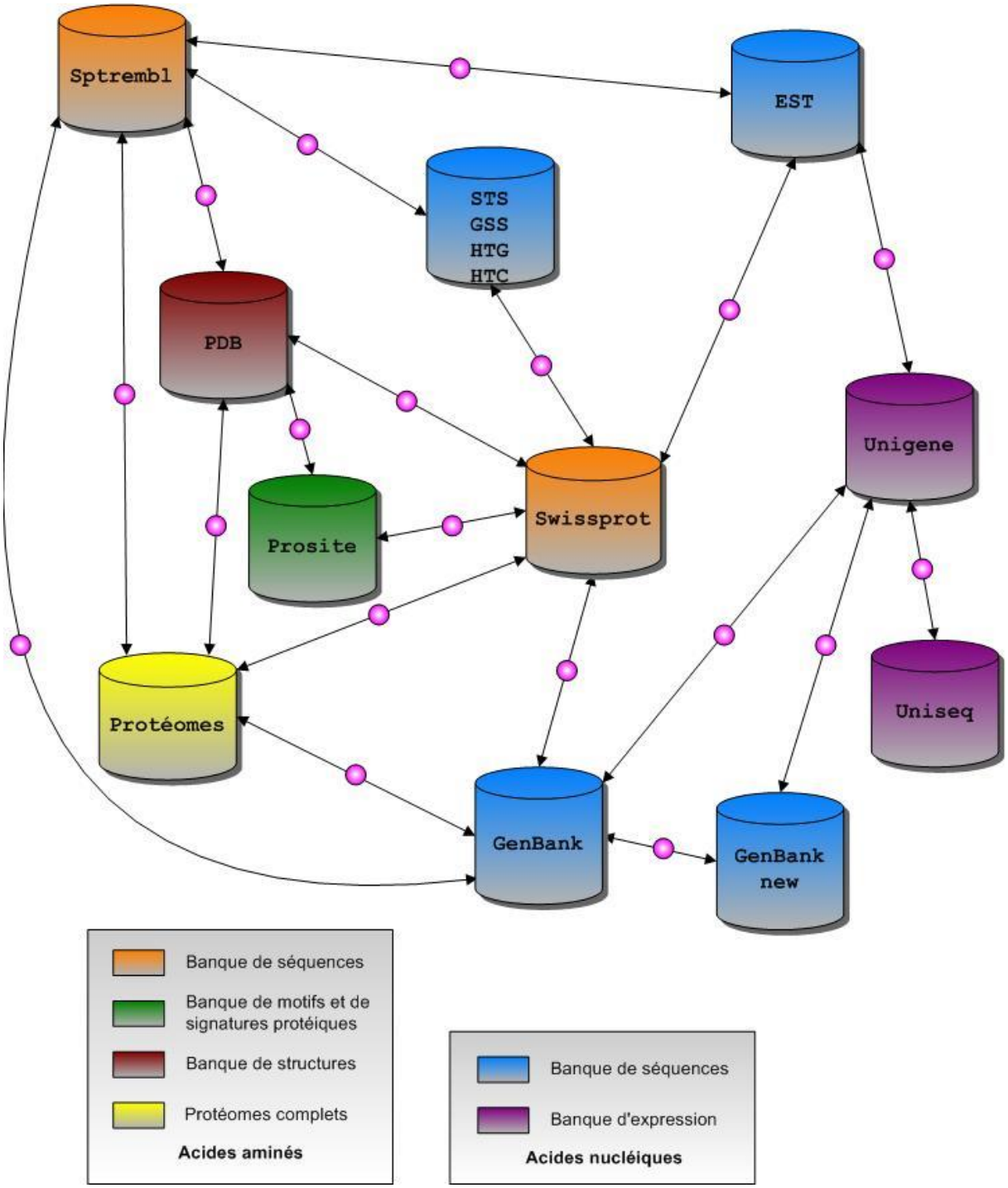
```
FT   NP_BIND      34     41       ATP (POTENTIAL).
SQ   SEQUENCE   249 AA;  27742 MW;  A63886EDE69AB80B CRC64;
     1  MITVKNIRKA  FKDLVVLDGI  DLEVKRGEVV  AIIGPSGSGK     STLLRCLNLL
    51  ERPDQGLIEI  GEAKLNAEKF  TRKEAHRLRQ  QTAMVFQNYN  LFKNKTALQN
   101  ITEALIVAQH  KPRDEAKRIG  MEILKQVGLE  HKADSYPITM   SGGQQQRIGI
   151  ARALAVNPHA ILLDEPTSAL  DPELVTGVLQ  VIKSIAEKQT     TMIIVTHEMA
```

Chaque entrée de SwissProt est composée de lignes commençant par deux caractères, ce sont les champs. Ils indiquent le contenu de la ligne, exemple :

AC pour Accession Number ou numéro d'accès

Annexe 2

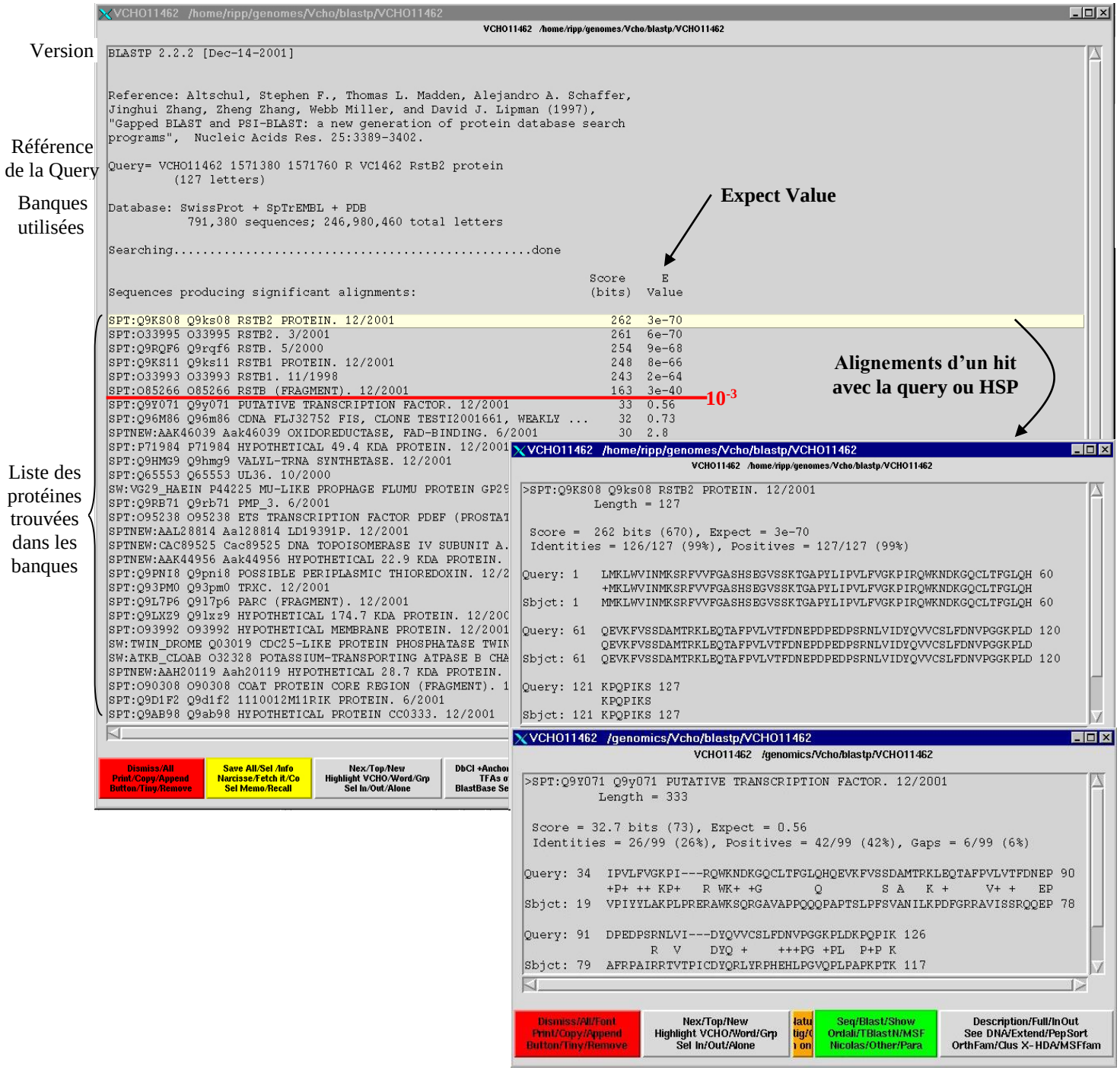
Schéma relationnel des différentes bases de données de SRS



Annexe 3

Une sortie BlastP

Résultat de la recherche par BlastP de l'Orf VCHO11462 dans les banques SwissProt, PDB et SpTrEMBL



En rouge la limite significative de 10^{-3} est indiquée. On peut voir deux alignements particuliers, le meilleur hit de la recherche et le premier hit sous la limite significative. On constate bien que le second alignement n'est plus significatif.

Annexe 4

La vitrine qui permet d'accéder aux principales fonctions de GScope

Version et génome utilisé

La fonction « Inventory Check » qui permet de tester l'avancement d'une base de données Gscope

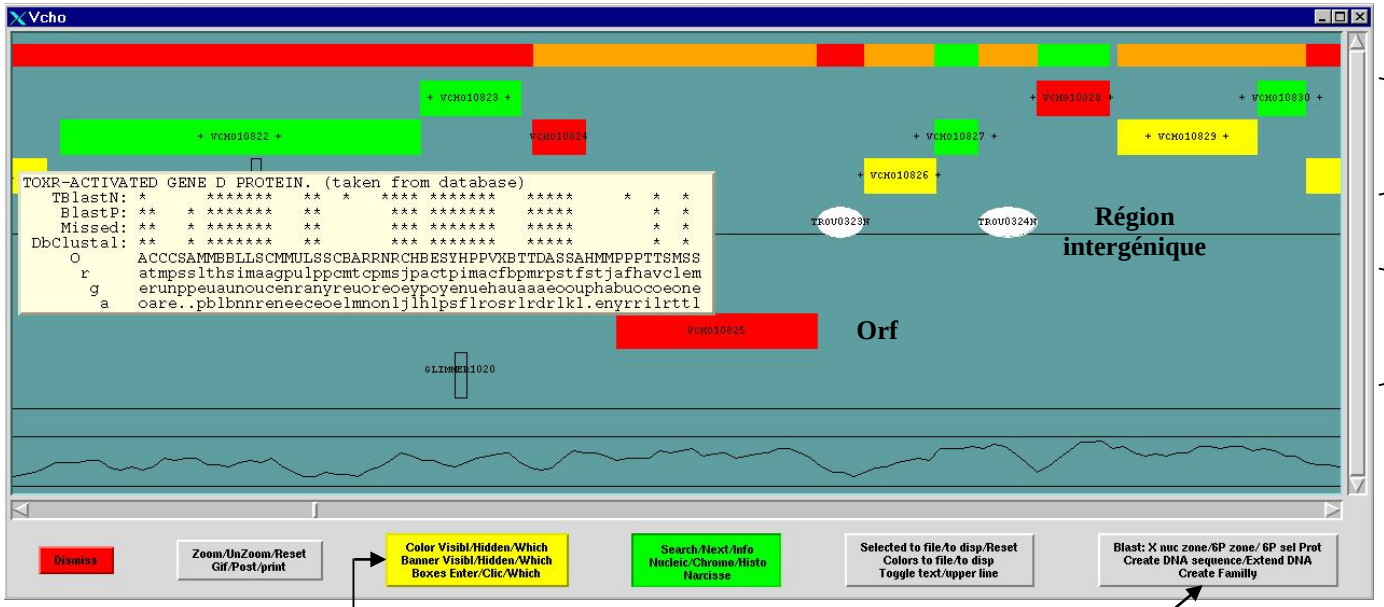
Pour accéder au board

La fonction X-HDA qui permet de visualiser un ou plusieurs bilans et dans créer

Pour quitter Gscope

Annexe 5

Le « board » qui permet de naviguer sur un génome et d'accéder aux données et informations disponibles



3 cadres de lecture du premier brin d'ADN

3 cadres de lecture du second brin

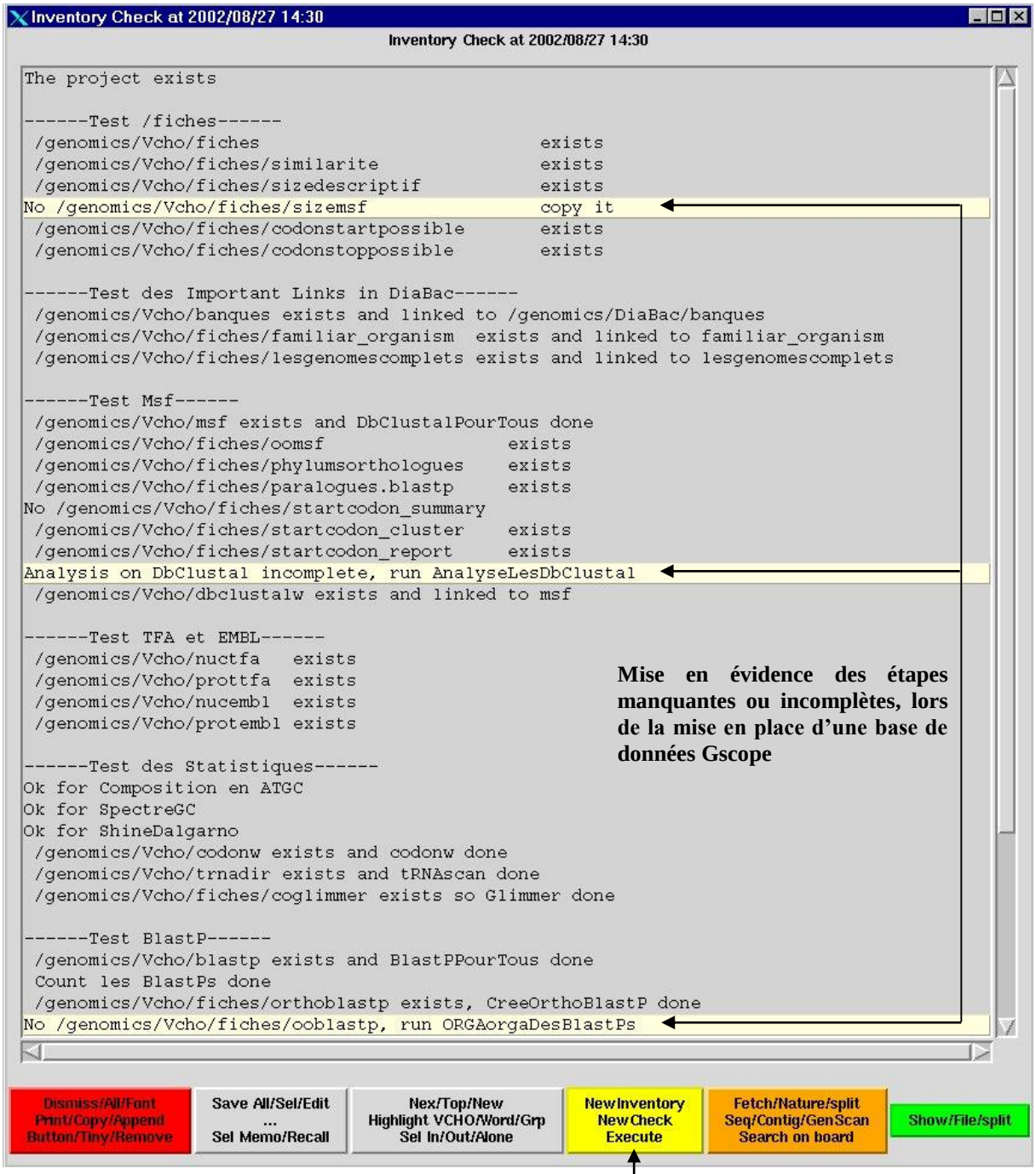
% G+C

Fonctions disponibles en cliquant sur :
Mouse1, 2 ou3
Shift+Mouse1, 2 ou3
Control+Mouse1, 2 ou3

Fonction « Create Family » disponible par Control+Mouse1. Elle permet de créer les familles de protéines homologues

Annexe 6

Panneau de contrôle de « Inventory Check »

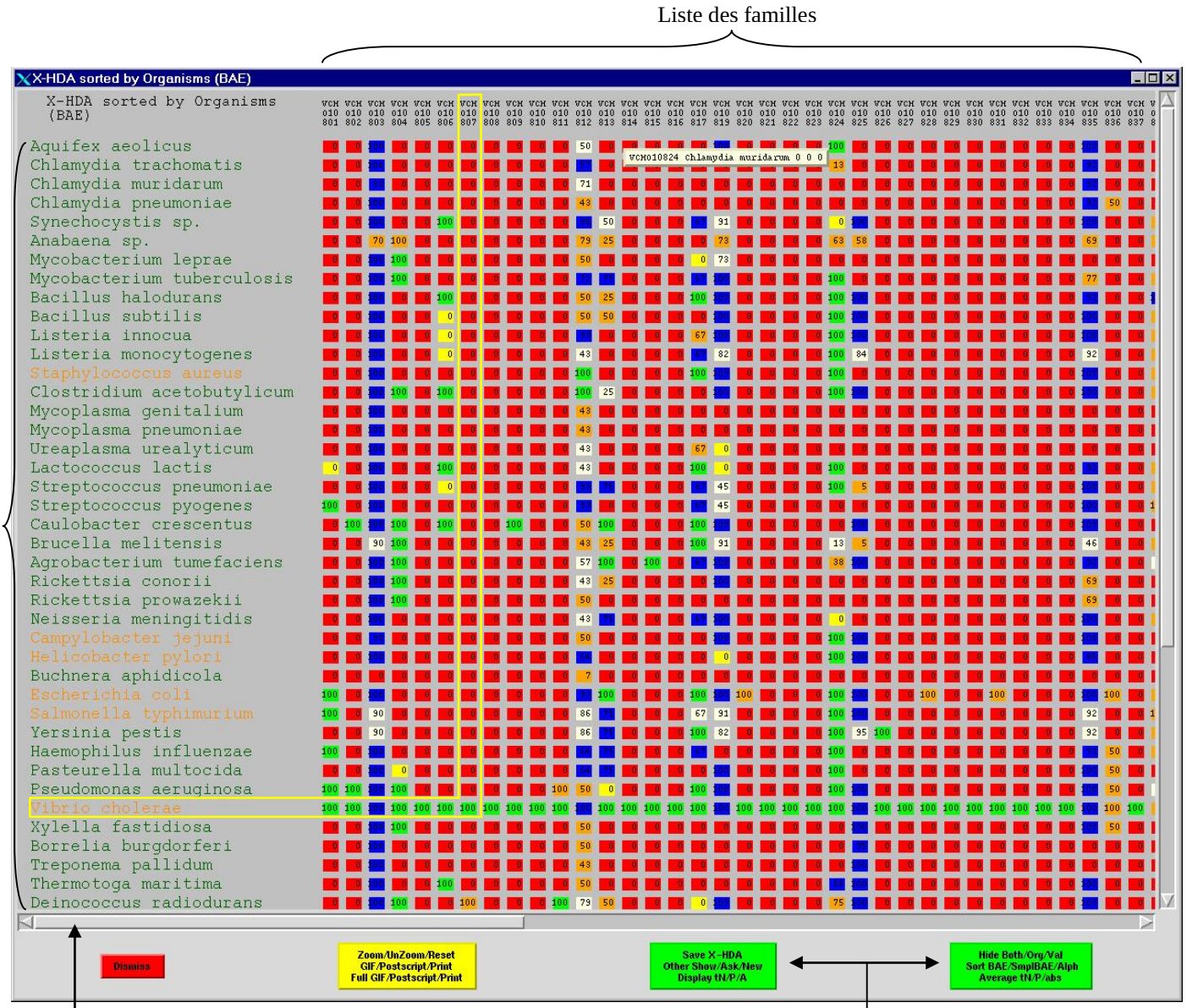


Mise en évidence des étapes manquantes ou incomplètes, lors de la mise en place d'une base de données Gscope

Pour exécuter les procédures illuminées : Control+Mouse1

Annexe 7

Un bilan de quelques familles triées dans l'ordre Bactérie-Archae-Eucaryote :



La coloration des organismes (58 au total) est Orange pour les organismes DiaBacs, Vert pour les Bactéries, Bleu pour les Archaea et enfin Rouge pour les Eucaryotes.

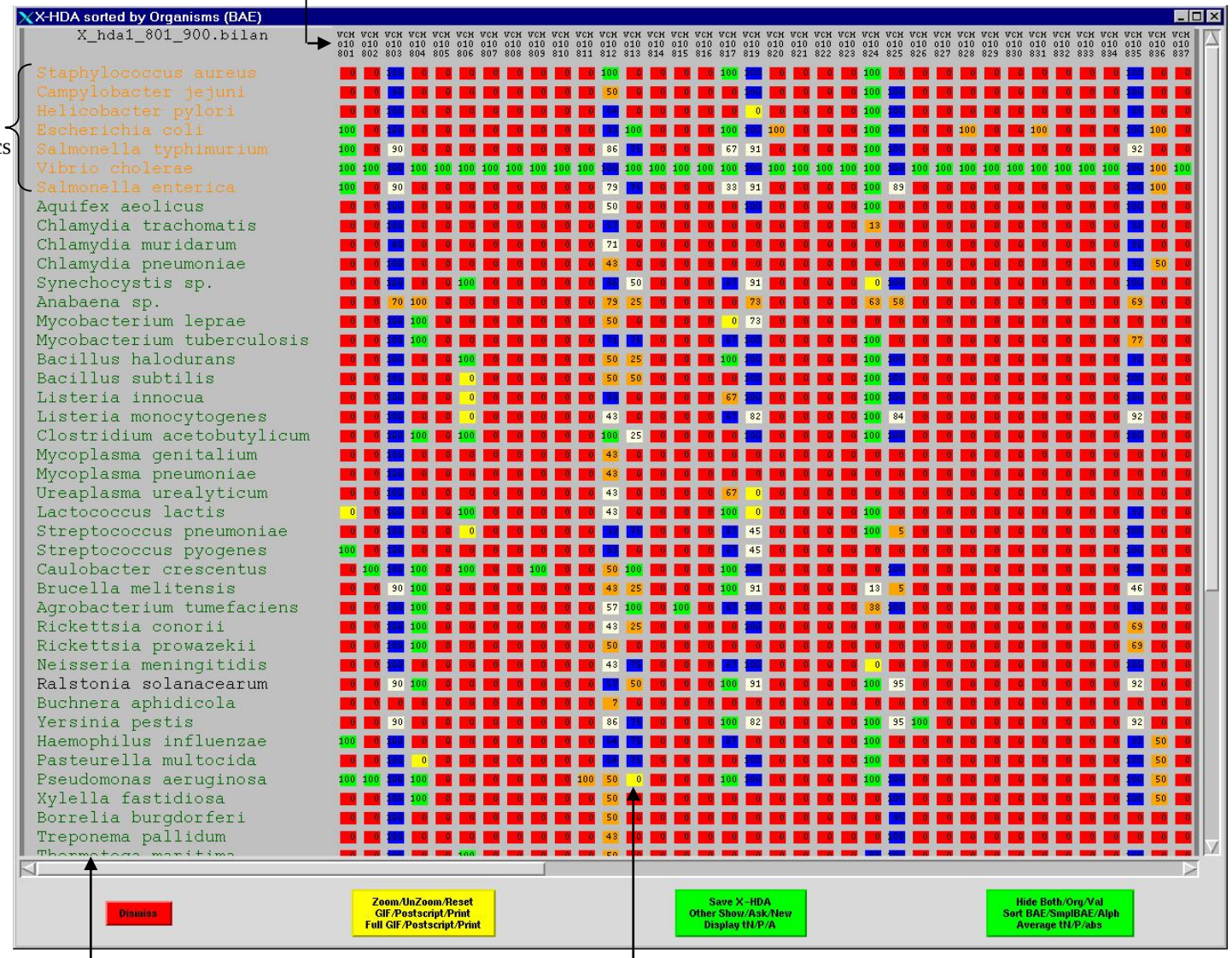
Les différentes fonctionnalités disponibles sur les boutons (sauvegarde, trie par organismes, montrer les différentes valeurs…..).

Encadré en jaune, le bilan particulier de la famille VCHO10807 pour Vibrio Cholera. Elle est présente dans tous les Blasts de la famille, elle est donc colorée en vert. La valeur affichée içi est le % de présence dans le BlastP.

Annexe 8

Le même bilan trié dans l'ordre DiaBac-Bactérie-Archae-Eucaryote :

Elément actif :
Mouse1 positionne sur le board
Mouse2 Membres de la famille
Mouse3 Cluster de la famille
Shift+Mouse1,2 pour trier sur l'une des deux valeurs du bilan



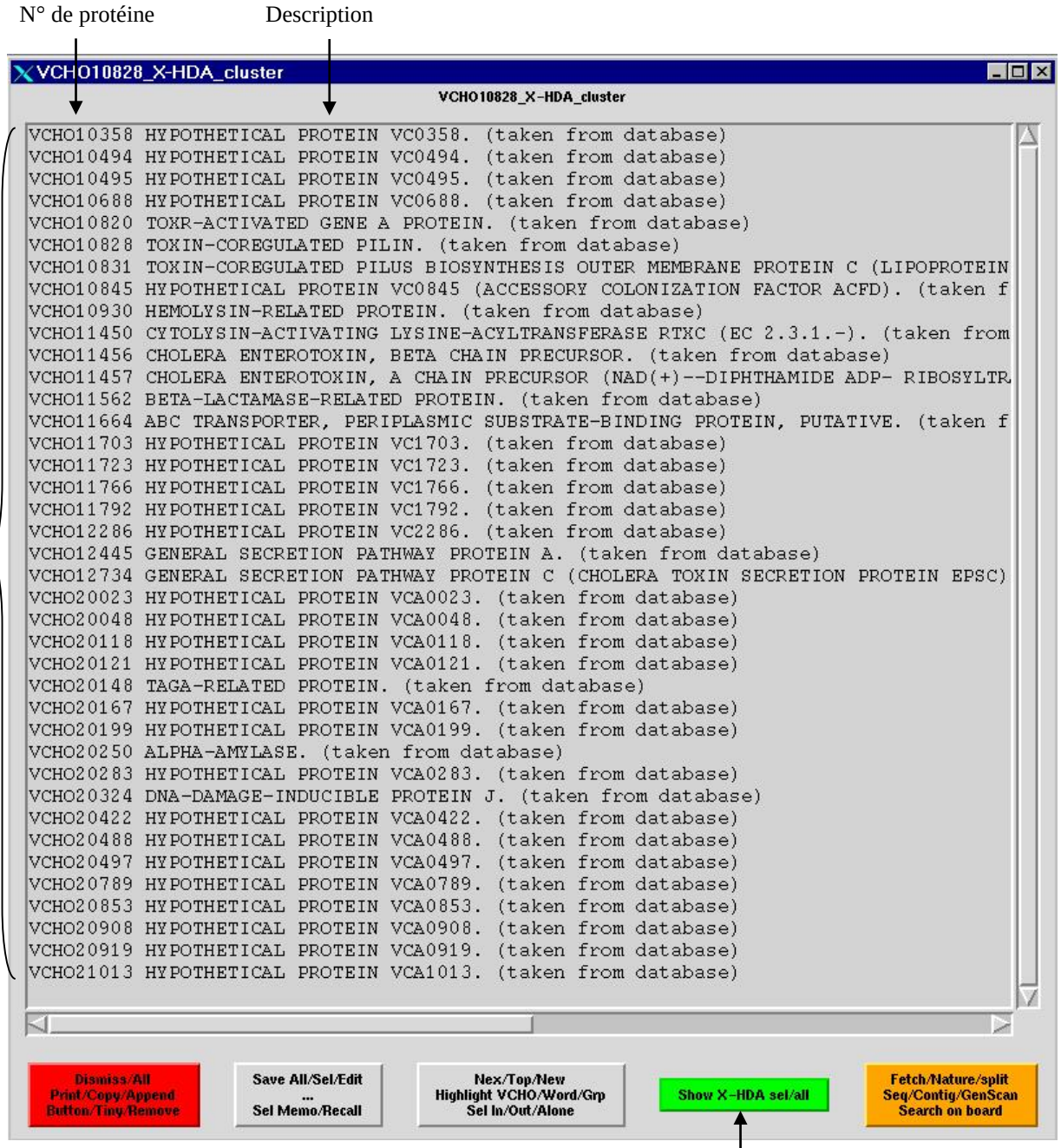
Les DiaBacs

Protéine de la famille de l'Orf VCHO10813 ratée chez Pseudomonas aeruginosa

Elément actif :
Shift+Mouse1,2 pour trier sur l'une des deux valeurs du bilan

Annexe 9

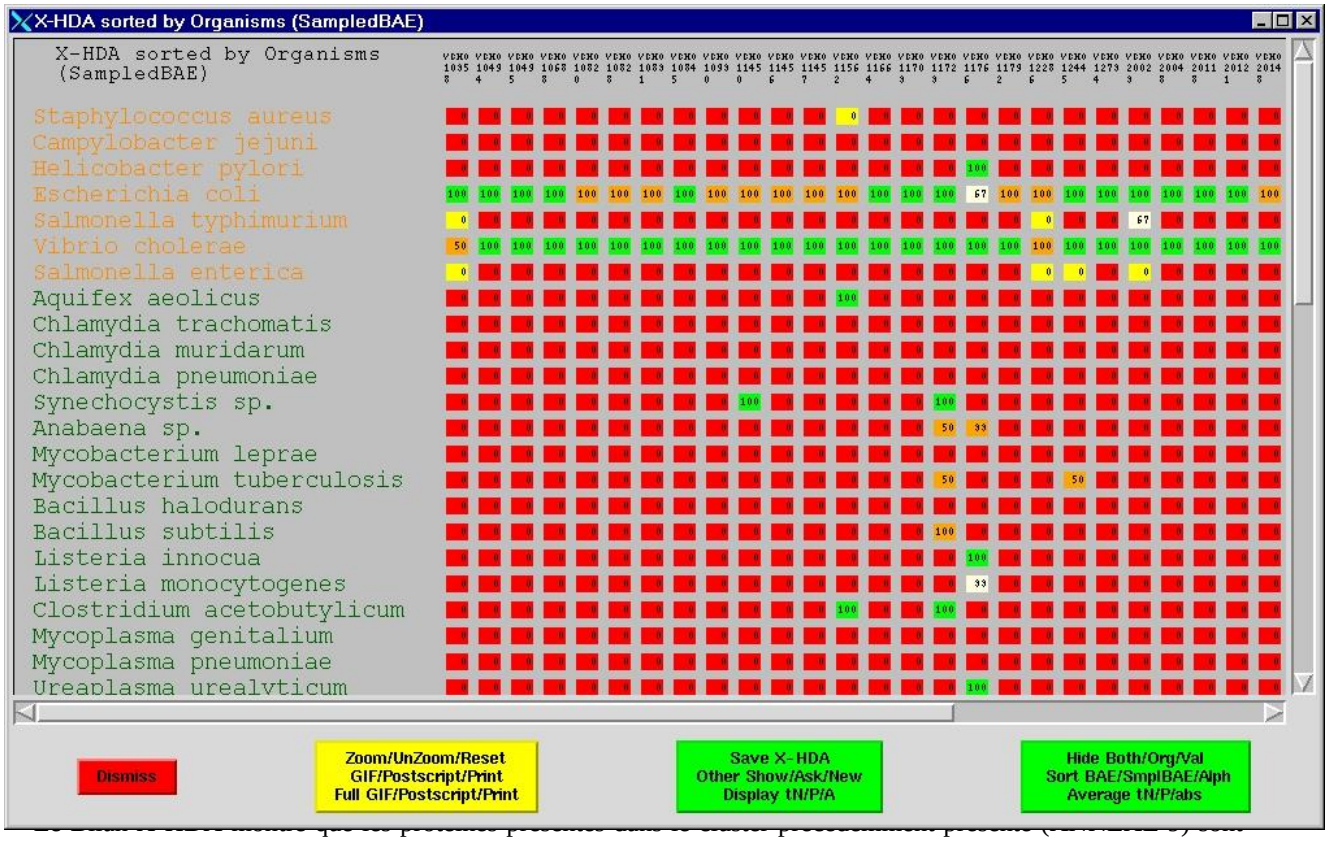
Exemple de constitution d'un cluster obtenu à partir des bilans X-HDA



On voit dans ce cluster la présence de protéines constitutives de l'îlot pathogène de Vcho (VCHO10820-11457) ainsi que des protéines sans fonctions connues (« Hypothetical Protein »). Ces protéines n'appartiennent pas aux îlots pathogènes mais peuvent être potentiellement impliquées dans la diarrhée.

Annexe 10

Bilan X-HDA du cluster présenté en Annexe 9



liées à deux organismes en particulier, à savoir Vibrio choléra et Escherichia coli. Dans le cadre de la pathogénicité des bactéries, c'est un point intéressant qu'il faut encore étudier et développer.

Programmes	Séquence appât	Banque	Comparaison	Exemples d'utilisation
BLASTP	Protéine	Protéines	Niveau protéique	Recherche de protéines homologues
BLASTN	ADN	ADN	Niveau nucléique	Recherche de RNA structuraux, de séquences répétées, éléments régulateurs
TBLASTN	Protéine	ADN (traduit en séquence protéique dans les 6 cadres de lecture *)	Niveau protéique	Localiser un gène sur un génome, rechercher des similitudes entre une protéine et une séquence génomique pas ou mal annotée
BLASTX	ADN (traduit en séquence protéique dans les 6 cadres de lecture)	Protéines	Niveau protéique	Trouver les phases de lecture dans une séquence codante
TBLASTX	ADN (traduit en séquence protéique dans les 6 cadres de lecture)	ADN (traduit en séquence protéique dans les 6 cadres de lecture)	Niveau protéique	Combine les avantages de TBLASTN et BLASTX mais la recherche est plus longue

Tableau 1 : Présentation des différentes possibilités des logiciels BLAST

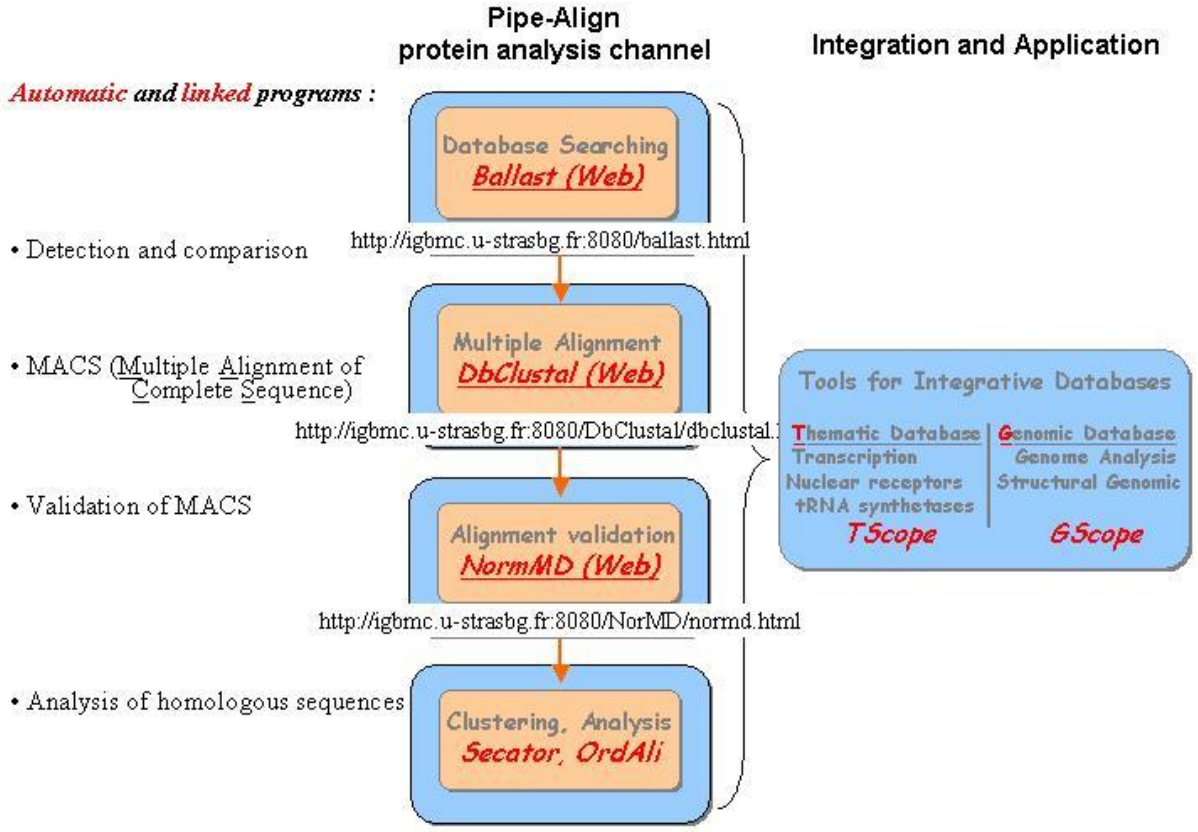
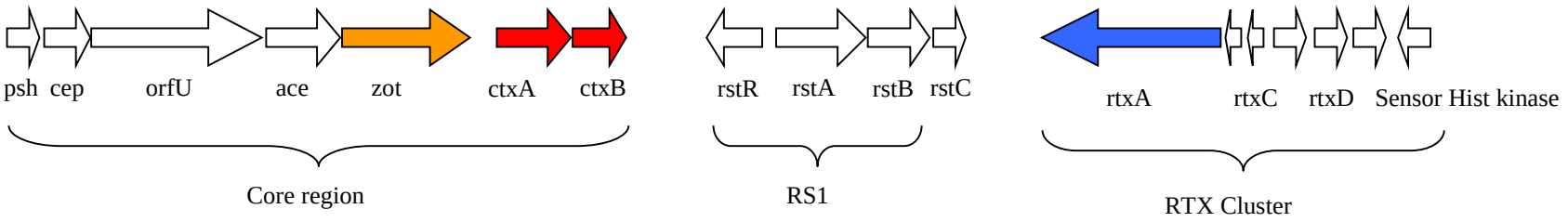


Figure 1 : Enchaînements des programmes de la colonne d'analyse de séquence développée dans le laboratoire de Bioinformatique.

Figure 4 : Un des îlots pathogènes majeur de Vibrio Choléra

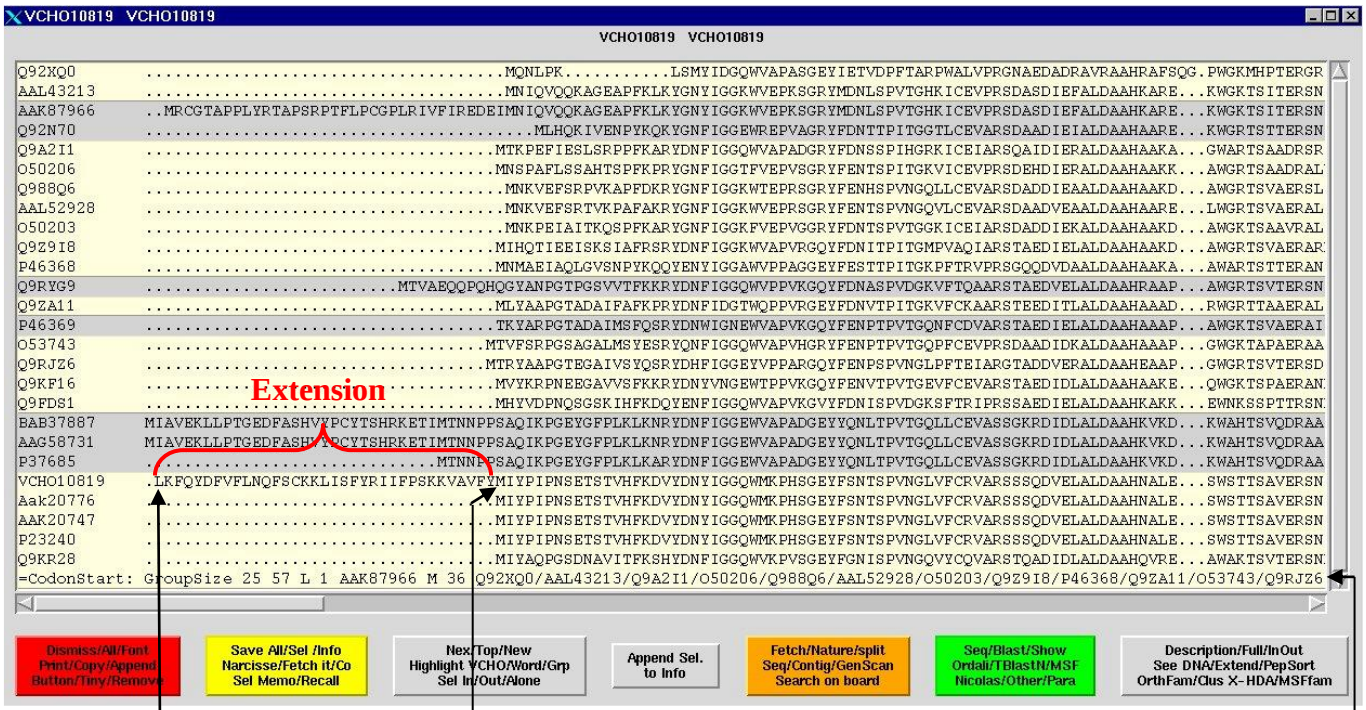


Les flèches représentent les ORFs ainsi leur sens dans le génome. Les noms en minuscule sont les noms des gènes associés à ces ORFs.

Cet îlot fait 62 679 paires de bases. Il contient deux regroupements de gènes codant pour des toxines directement responsables de la pathogénicité de Vibrio Choléra. En particulier ctxA at ctxB qui sont les gènes qui codent pour la toxine du choléra et rtxA un autre toxine. Les autres gènes présents sont des gènes de régulation.

Figure 3 :Exemple d'utilisation de « Start codon validation » pour proposer un autre codon initiateur.

Dans ce cas l'Orf de Vcho, VCHO10819, possède une extension en début de protéine par rapport aux autres séquences proches. De plus un codon initiateur, la méthionine (M), est situé à une position proche de celui des autres séquences. Le programme suggère ainsi que cette protéine est trop longue.et qu'il faut la coupée en M36.



Ancien codon initiateur

Nouveau codon initiateur proposé

Liste des codons initiateurs alternatifs et des Access des séquences qui permettent de les proposer

Organismes	Taille des protéomes (Nb Orfs)	Start codon validé			Start codon alternatif			Séquences exclues de l'analyse
			% du protéome	% des testés		% du protéome	% des testés	
H.pylori	1566	1200	77	95	64	4	5	302
H.pylori J99	1491	1225	82	98	21	1	2	245
V.cholera	3828	2151	56	82	488	13	18	1189
C.jejuni	1654	1182	71	98	27	2	2	445
S.aureus	2594	1949	75	97	70	3	3	575
S.typhimurium	4451	3884	87	96	145	3	4	422
S.enterica	4600	3822	83	97	134	3	3	644
E.Coli K12	4289	3476	81	90	404	9	10	409
E.Coli O157	5349	4428	83	91	429	8	9	492

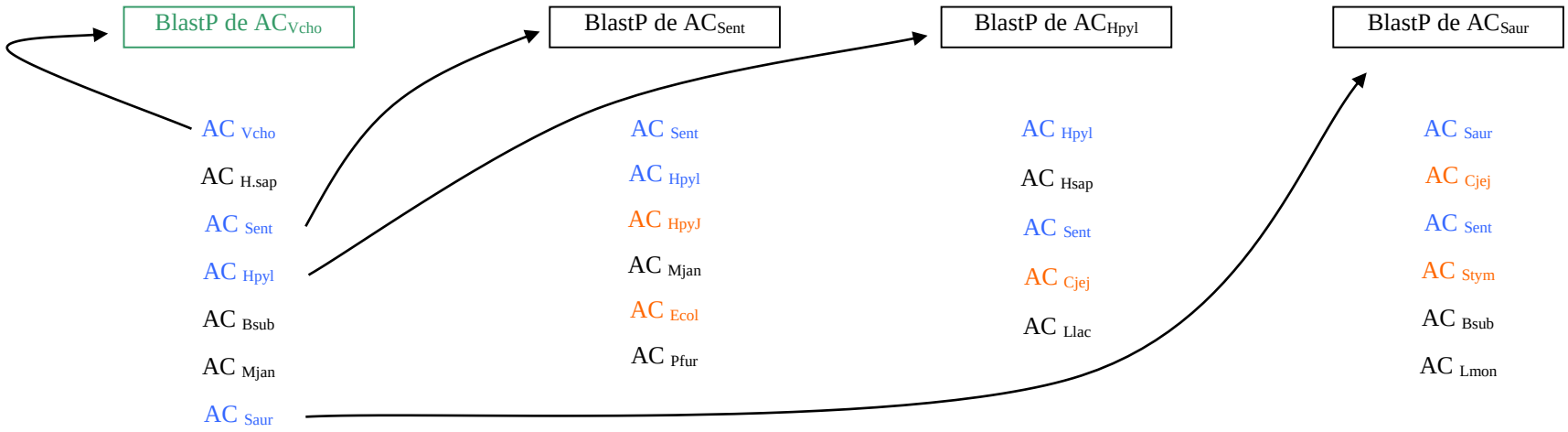
Tableau 2 : Résultat de l'analyse du codon initiateur avec le programme Start codon validation

Ce tableau contient les statistiques de la validation du codon initiateur sur les séquences des protéines de chacun des génomes. Il faut noter que toutes les protéines n'ont pu être testées car par exemple le nombre de séquences présentes dans l'alignement multiple était trop faible pour être significatif.

	Total des Trous	Nb de recherches BlastX comportant des protéines significatives		Nb Protéines Ratées	Nb Protéines mal définies
Organismes			%		
C.jejuni	309	9	2,9	0	1
H.pylori	502	83	16,5	4	43
H.pylori J99	489	58	11,9	5	20
S.aureus	1428	158	11,1	19	31
V.cholera	1667	284	17,0	1	51
Moyenne		118,4	11,9	5,8	29,2

Tableau 3 : Résultat de l'analyse des régions intergéniques

Figure 5 : Principe de la création des familles de protéines homologues dans DiaBac



Le blastP de départ est représenté en **Vert**. Les numéros d'accès sont représentés par les deux lettres AC associé à son organisme. La première liste est constituée des numéros d'accès de couleur **Bleue**. La deuxième liste, complète, est composée des numéros d'accès de la première liste (**Bleue**)et de ceux de couleur **Orange.**

Dans notre cas, la famille sera constituée des numéros d'accès suivant :

AC Vcho, AC Sent, AC Hpyl, AC HpyJ, AC Saur, AC Ecol , AC Cjej, AC Stym

Les Access des autres organismes Mjan (*M.janashii*), Hsap (*H.Sapiens*), Pfur (*P.furiosous*), Llac (*L.lactis*), Bsub (*B.subtilis*) et Lmon (*L.monocytogenes*) ne sont pas retenus.

Figure 2 : Schéma des dépendances fonctionnelles des données générées par GScope