Utilisation, développement et déploiement de logiciels libres pour la protéomique

Olivier Langella[1], Benoît Valot[2], Michel Zivy[1].

[1]CNRS, [2]INRA, Plateforme d'Analyse Protéomique de Paris Sud-Ouest,
UMR de Génétique Végétale, 91190 Gif-sur-Yvette, France.

http://pappso.inra.fr/bioinfo

Liste de diffusion : pappso-tools@listes.inra.fr

Contexte

Les progrès techniques en spectromètrie de masse (augmentation de la précision et des fréquences d'acquisition) ouvrent la voie à de nouvelles applications (Quantification MS, Métaprotéomique, De Novo).
Ces évolutions nécessitent une adaptation des outils utilisés en protéomique pour pouvoir stocker efficacement ces masses de données, mais aussi pour pouvoir les traiter.

Nos objectifs

- Apporter des solutions techniques pérennes et évolutives pour assurer le stockage et l'exploitation des données
- Maîtriser la chaine d'outils utilisée pour pouvoir l'adapter à nos besoins
- Développer si nécessaire de nouveaux outils
- Mettre à disposition et mutualiser les développements

Moyens

Les logiciels libres fournissent des solutions pérennes à nos problématiques systèmes et logicielles. Par essence, ils sont hautement adaptables et modifiables au gré des besoins. Une communauté importante dans le domaine de la spectrométrie de masse est en cours d'émergence (poster de **F. Rusconi**).
PAPPSO a donc fait le choix du logiciel libre pour répondre à ses besoins scientifiques. Cela nous a amené à les utiliser, mais aussi à proposer nos développements sous forme de logiciels libres.
Nous présentons ici quelques logiciels ou ressources clés de notre infrastructure, ainsi que les ressources développées en interne que nous diffusons par l'intermédiaire de SourceSup (poster de **S. Médard**) et du projet PLUME (poster de **D. Rousse**).

Développement de logiciels

Pour nos développements, nous réutilisons autant que possible le code libre déjà disponible dans le domaine de la protéomique (ObiWarp, X! Tandem...).

Nous utilisons aussi des bibliothèques généralistes dont certaines sont peut-être méconnues :
- ALGLIB (C++) : analyses de données, statistiques
- CImg (C++) : traitement d'images
- jOOQ (Java) : SQL orienté objet (bon remplaçant d'Hibernate)
- jQuery (JavaScript) : interfaces web
- MascotDatFile (Java) : parsing de fichiers Mascot

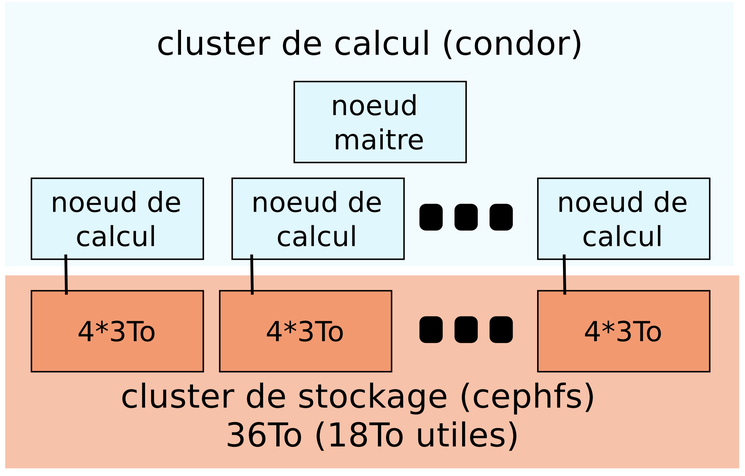
Nos logiciels

- MassChroQ (C++, Qt) : quantification de peptides par spectromètrie de masse
- Beads (C++, Qt) : détection de spots sur images de gels
- X!TandemPipeline (Java, Swt) : manipulation de résultats d'identifications MS
- PROTICdb (Java, REST Cxf) : base de données de protéomique.

Systèmes

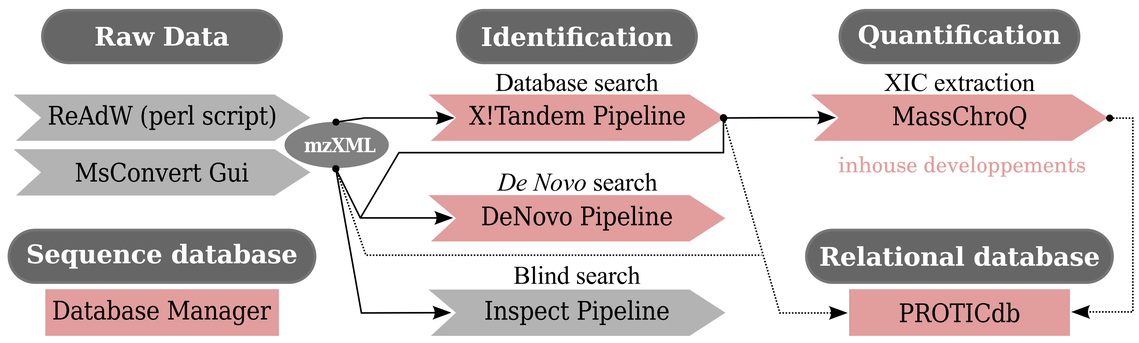
- Debian GNU/Linux, pour nos serveurs et postes de bureautique (stable, déploiement et mises à jour automatiques...).
- Ceph, système de stockage distribué.
- HTCondor, pour la répartition des calculs dans notre cluster.

Architecture matérielle



- Matériel standard peu couteux
- Architecture évolutive (ajout de puissance de calcul ou de stockage en fonction des besoins)

Logiciels de protéomique



- Basé uniquement sur des logiciels libres (respect des standards, faibles coûts, pérennité)
- Intégration d'outils existants
- Développement de nouveaux outils

Compilation, déploiement

Nos logiciels utilisent "cmake" pour la compilation. Nous fournissons des versions Windows de nos logiciels par cross compilation (mingw, Qt).
Les logiciels en Java sont disponibles sur notre site via Java Web Start.
Enfin, tous nos logiciels sont packagés pour Debian (dépôt apt disponible, compatible Ubuntu) et nous travaillons avec F. Rusconi pour leur intégration au projet Debian via le groupe de travail "Debichem".

Vrai OpenSource

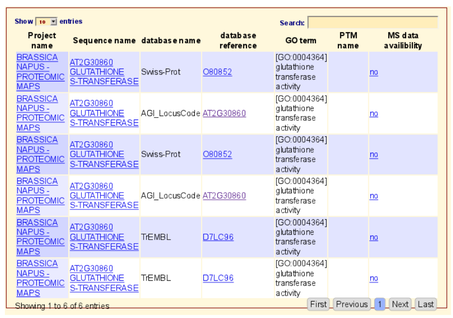
Tous nos développements sont disponibles par l'intermédiaire de **SourceSup** (dépôts git), en direct (pas de développement confidentiel en interne), sous licence GPLv3. Toutes les contributions sont les bienvenues.

Nos bibliothèques

Nous concevons nos logiciels de manière modulaire. Ceci nous permet de partager énormément de code entre nos différents projets. Dès que possible, nous essayons de grouper nos modules dans des bibliothèques indépendantes, documentées et facilement réutilisables.

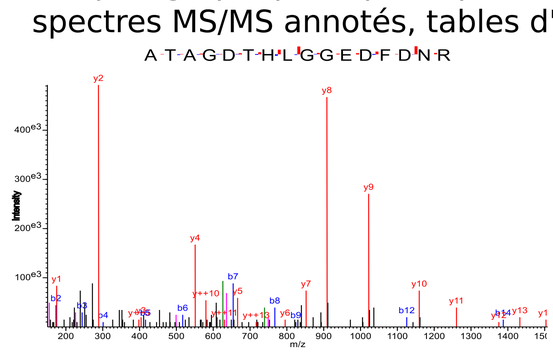
X2DBI (serveur Java Jetty + interface HTML, jQuery)

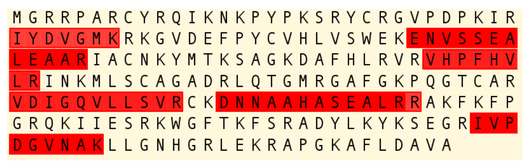
Création d'interface graphique de requêtes pour bases de données SQL.
A partir d'un format XML simple décrivant la structure de votre base de données, X2DBI génère une interface graphique simple permettant de construire des requêtes et de les exécuter. X2DBI propose l'autocomplétion des champs, la visualisation des résultats, l'export des données en ODS (LibOdsStream) ou en CSV.
A tester sur http://moulon.inra.fr/protic/public (bouton "Query").



LibProteomicSvg (Java)

- Manipulation de documents HTML5, XHTML et SVG sous formes d'objets.
- Objets graphiques spécifiques à la protéomique : couverture peptidique, spectres MS/MS annotés, tables d'ions...





LibOdsStream (Java ou C++)

Lecture et écriture de fichiers ODS (Open Document Spreadsheet) en flux (faible empreinte mémoire).
Avantages par rapport au texte tabulé :
- lisible et modifiable dans un tableur (LibreOffice ou Excel).
- typage des cellules (chiffres, dates, textes).
- liens hypertextes cliquables.
- pas de problème avec les caractères spéciaux ou tables de caractères.
- pas de problème de séparateurs (colonnes, fin de lignes, échappements).
- tailles des fichiers minimes.

This work was partially supported by IBiSA

SourceSup hosts the projects

GPL Free Software — Free as in Freedom