Paquetage de logiciels scientifiques de spectrométrie de masse pour Debian

Filippo Rusconi <filippo.rusconi@u-psud.fr | lopippo@debian.org>
Laboratoire de Chimie physique — UMR 8000 CNRS — Université Paris Sud
Orsay, France

Introduction

La spectrométrie de masse a révolutionné la manière de regarder les objets biochimiques, comme les protéines, les sucres et les acides nucléiques. Indirectement, la spectrométrie de masse permet ainsi d'avoir un aperçu moléculaire d'organes entiers et s'est ainsi progressivement rendue indispensable en biologie.

Mais la spectrométrie de masse est une discipline relativement récente, comparée à la génétique ou à la génomique. C'est la raison pour laquelle la familiarité des usagers de la spectrométrie de masse avec les systèmes informatiques anciens (UNIX) et le logiciel libre (GNU/Linux) est assez limitée. Ce n'est que depuis quelques années que l'on ressent un frémissement d'intérêt pour les plateformes de développement logiciel libres (comme Ubuntu, par exemple), mais cet intérêt et encore aujourd'hui fondé quasi-uniquement sur la gratuité de la plateforme.

Les mentalités doivent évoluer pour donner libre cours à la publication du code source des logiciels développés selon une licence libre, et cela de manière à mutualiser les efforts de développement.

Dans ce contexte, nous avons le projet de rendre disponibles le plus largement possible les logiciels libres développés pour la spectrométrie de masse. Notre plateforme de développement est la distribution Debian GNU/Linux.

Matériels & méthodes

- Chaque logiciel est développé par son/ses auteur/s respectif/s. Le langage de développement est choisi par chaque projet et n'intervient pas dans le choix de faire ou pas un paquetage pour ce logiciel ;
- Plusieurs développeurs participent au projet de paquetage et le travail collaboratif a lieu sur la plateforme du projet Debian (`alioth.debian.org`) ;
- Tous les développeurs du projet utilisent la distribution Debian GNU/Linux (`http://www.debian.org`) ;
- Le logiciel *massXpert* est développé en C++ avec les librairies Qt ;

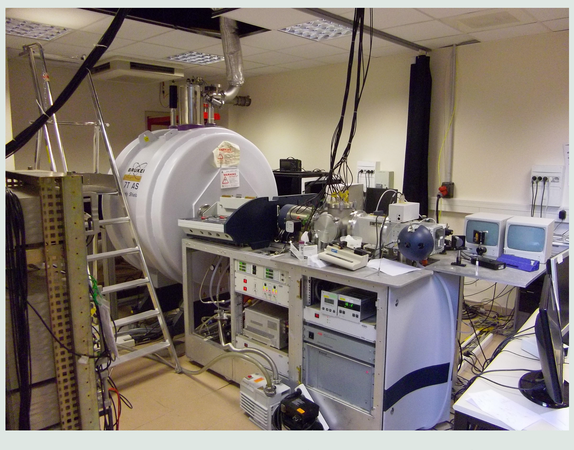
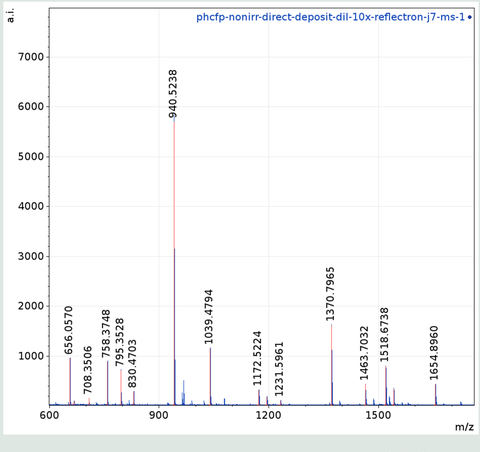


Photo : Vincent Steinmetz (U-Psud, Orsay)

⇒



Spectre de masse : $i = f(\frac{m}{z})$

⇒

Emploi de la spectrométrie de masse en recherches fondamentale et appliquée

- Caractérisation de protéines, de glucides et d'acides nucléiques (ADN, ARN) → étude de la fonction et de la régulation des polymères biologiques ;
- Développement de nouveaux médicaments : suivi des effets d'une molécule sur une culture cellulaire.

Les fabricants

- Les fabricants de machines luttent pour le contrôle des données acquises par les laboratoires et donc pour le contrôle des usagers « *vendor lock-in* » ;
- Les formats propriétaires (compression des données) « enferment » des teraoctets de donnéees sensibles de protéomique ;
- Chaque logiciel devrait disposer d'une fonction facilement accessible d'export des données dans un format standard ouvert.

Les utilisateurs

De l'ancienne culture :

- Années 70 : données au format texte simples et faciles à traiter ;
- Plateformes de l'époque : PDP-11, VAX/VMS, DEC-ALPHA, UNIX...

De l'âge moderne :

- Années 2000 : données complexes hiérarchisées (*full scan, ms/ms*) ;
- Plateformes MS-Windows NT4, Millennium, 2000...

Ces différences historiques/culturelles ont impacté le comportement associé à l'emploi des solutions logicielles

Debian comme distribution pour la spectrométrie de masse doit satisfaire deux types différents de publics

Le massiste et/ou l'utilisateur final :

- Outils de stockage/gestion des données ;
- Interfaçages faciles avec les bases de données ;
- Applications spécifiques de traitement des données ;
- Interrogation bio-centrique des données ;
- Outils de rendu graphique des données ;

Le développeur :

- Logiciels d'interfaçage avec les bases de données (*web services*) ;
- Écriture de code en C/C++, Python, Java, Ruby...
- Systèmes de construction de logiciels (CMake, Autotools, BJam...
- Outils de génération de documentation (sgml/xml, TeXinfo, LaTeX, Sphinx...)

Trois types de logiciels

Outils logiciels de création de workflow :

`openms` (5)
`libpwiz` (3)
`tandem-mass` (1)

Librairies/exécutables de process de données (quantification, conversion de format) :

`python-mzml` (2)
`r-cran-readbrukerflexdata` (1)
`r-cran-maldiquant` (1)
`r-cran-base64enc` (1)

Logiciels pour l'utilisateur final :

`polyxmass` (2)
`massxpert` (3)
`lutefisk` (1)
`mmass` (2)

Une suite logicielle massive :

- C++ ; LGPL 2.1 ; CMake ;
- 2 libraries (fondamentale and Interf. graph.) ;
- 114 binaires ;
- Documentation importante ;
- Répertoire `contrib` avec dépendances externes ;
- Décisions basées sur la variable `DESTDIR` discutables ;
- Interactions utiles avec les membres du projet ;
- Build-Depends : debhelper (>=7.0.50), dpkg-dev (>= 1.16.1), quilt (>= 0.60-2), cmake (>=2.6.3), libxerces-c-dev (>= 3.1.1), libgsl0-dev (>= 1.15+dfsg), libboost1.49-dev, libboost-iostreams1.49-dev, libboost-date-time1.49-dev, libboost-math1.49-dev, seqan-dev (>= 1.3.1), libsvm-dev (>= 3.12), libglpk-dev (>= 4.45), zlib1g-dev (>= 1.2.7), libbz2-dev (>= 1.0.6), cppcheck (>= 1.54), libqt4-dev (>= 4.8.2), libqt4-opengl-dev (>= 4.8.2), libqtwebkit-dev (>= 2.2.1), coinor-libcoinutils-dev (>= 2.6.4), imagemagick, doxygen (>= 1.8.1.2), texlive-extra-utils, texlive-latex-extra, latex-xcolor, texlive-font-utils, ghostscript, texlive-fonts-recommended

Plateforme Debian : bilan

POUR :

- Beaucoup de logiciels déjà empaquetés ;
- Debian est la base de beaucoup de dérivés ;
- Infrastructure Debian Pure blends ;
- Puissante infrastructure de développement partagé (Alioth) ;
- Robustesse de Debian.

CONTRE :

- Empaquetage de logiciels sophistiqué et complexe ;
- Distribution intimidante pour les novices, en particulier pour le biochimiste non informaticien (améliorations sensibles avec le temps).

Difficultés :

- Logiciels Java livrés avec beaucoup de librairies binaires ;
- Logiciels non-libres mais fort utiles ;
- Bases de données de substances non libres.

Conclusions :

- Projet au long cours ;
- Nécessité de constituer une communauté ;
- Merci à Olivier Langella et Sebastian Gibb.