This is a reprint of
**Lecturas Matemáticas**
*Volumen* **25** *(2004), páginas 211–218*

# Implementación de servicios de análisis usando tecnología *Open-Source*

Alexander García
University of Queensland, Australia

**Lecturas Matemáticas**
*Volumen* **25** *(2004), páginas 211–218*

# Implementación de servicios de análisis usando tecnología *Open-Source*

Alexander García Castro
University of Queensland, Australia

Abstract. Open–source technology for biochemistry and molecular biology applications.

*Key words and phrases.* Open-source technology, biochemistry, molecular biology, bioinformatics.

*1991 Mathematics Subject Classification.* Primary: 92C40. Secondary: 92D20

Resumen. Tecnología Open-source para aplicaciones en la bioquímica y la biología molecular.

Dentro de las labores cotidianas de muchos laboratorios, independientemente de su tamaño, se requiere del uso de herramientas informáticas para acceder o para analizar los datos provenientes de diferentes experimentos. Este escrito se enfoca en aquellos procesos que dentro de la bioquímica o la biología molecular requieren esta clase de infraestructura informática. En él se abordarán y describirán dos actividades principales: aquellas labores directamente relacionadas con el almacenamiento y acceso a la información y aquellas que están mas íntimamente relacionadas con el análisis de datos.

A medida en que se vayan desarrollando los dos aspectos antes mencionados será mas claro para el lector cómo se complementan mutuamente.

La implementación de ambientes de análisis en biología molecular y bioquímica requiere un claro entendimiento de los recursos que se pretenden facilitar, además de las operaciones que sobre estos ambientes se van a llevar a cabo. Estas y otras consideraciones las discutiremos a lo largo del presente artículo.

Una parte de la bioinformática esta relacionada con la provisión de herramientas para el análisis y el almacenamiento de información biológica. Las diferentes herramientas de análisis disponibles en este campo son heterogéneas, se presentan en una alta diversidad de lenguajes en cuanto a su implementación, carecen de diseño centrado en el usuario (en su mayoría son herramientas que deben usarse a través de una línea de comandos sobre sistemas operativos tipo UNIX), no permiten la fácil e inmediata generación de *workflows*, presentan una solución a un problema particular y aunque algunos algoritmos de análisis requieren del acceso a bancos de datos (BLAST), la disponibilidad de estos es independiente de la instalación de la herramienta.

Las diferentes implementaciones de los algoritmos de análisis están disponibles sobre Internet. En algunos casos estos servicios son asequibles a través de un navegador (pues proveen una interfaz gráfica, GUI). En la mayoría de los casos es posible tener acceso al código fuente de las implementaciones. Igual situación se presenta con relación a la disponibilidad de bancos de datos, pues en general es posible por vía de FTP (*File Transfer Protocol*) contar con copias locales.

Una pregunta recurrente por parte de algunos usuarios se refiere a la disponibilidad de las herramientas: si están disponibles ¿por qué molestarme en tenerlas localmente?

Si bien es cierto que estas herramientas están disponibles a través de Internet también es cierto que la capacidad de análisis disponible sobre servidores de acceso público es limitada, no sólo por razones técnicas sino también por razones de seguridad. Debe también quedar claro que la mayoría de los análisis en biociencias no se limitan a la aplicación de un solo algoritmo sobre un conjunto de datos (*input*) sino que comprenden el encadenamiento secuencial de varios algoritmos de análisis sobre uno o más conjuntos de datos iniciales (*workflows*). Llevar a cabo un

alineamiento en sí mismo no ayuda mucho y hacerlo con pocos datos es poco ilustrativo si se tiene acceso a genomas completos.

Las diferentes preguntas que son posibles sobre los bancos de datos por vía de interfaces WEB son poco expresivas en términos de la riqueza semántica que la información biológica tiene. Además de que son pocas las herramientas para la manipulación de la información que este tipo de interfaces provee. Un ejemplo que ilustra esta situación es la poca capacidad que actualmente se tiene para relacionar las diferentes dimensiones de información que las estructuras proteicas tienen: estructura primaria, secundaria, terciaria y cuaternaria. Es así como tiene sentido tener estas herramientas y los bancos de datos localmente, pues, además de potenciar el acceso a la información, se genera un *know-how* en varios niveles (técnicos y científicos), se abren puertas para proyectos relativos a la implementación o simple mejora de las herramientas existentes.

¿Cómo contar con un ambiente de análisis que sea escalable, de bajo costo, mantenible, y sobre todo útil para apoyar la investigación que se lleva a cabo? Es pertinente aquí una aclaración. Si bien es cierto que la bioinformática tiene un área que pretende facilitar el acceso a la información y a la generación de herramientas de análisis, es muy claro que estas labores son guiadas por las necesidades propias de la investigación. Idealmente el bioinformático estará involucrado en varios frentes dentro de un centro de investigación. Idealmente estará en capacidad de apoyar las labores de sus colegas brindándoles otras perspectivas propias de quien maneja en razón de su labor una visión más macro. En la práctica no es fácil contar con esta clase de bioinformático, y no es tampoco fácil introducir la bioinformática en las labores propias de los *wet-biologists*. Se ha comprobado que éste es un trabajo largo y muchas veces frustrante.

Entendiendo de antemano que el simple hecho de contar con un bioinformático y con el ambiente computacional no garantiza que de un momento a otro aquellos investigadores del laboratorio vayan inmediatamente a calcular condiciones de PCR por vía del *software* o a estar versados en los diferentes contenidos de los bancos de datos, empezaremos por describir un posible ambiente de análisis que puede ser alcanzado por etapas dependiendo de las necesidades.

Consideremos un escenario básico. Acceso a bancos de datos y a un conjunto de herramientas de análisis (*suite*). Existe un paquete, de fácil implementación local, que permite mantener cualquier banco de datos en biología molecular. Es el SRS (*Sequence Retrieval System*). El SRS brinda una interfaz única de consulta al usuario que, además, es independiente de la base de datos. Para los casos en los cuales se desee presentar una base de datos creada *in-house*, SRS también puede usarse. Consultas básicas, así como algunas de nivel medio de complejidad, pueden hacerse usando el lenguaje de consultas propio del SRS. La actualización de los bancos de datos puede hacerse de manera automática si se cuenta con la coacción que lo permita. En este sentido, una conexión por vía de cable MODEM es suficiente, dependiendo del tamaño de las actualizaciones. En general, para mantener copias del EMBL, GenBank, GO (*Genome Ontology*), SwissProt y MedLine se recomienda en estos casos seguir procesos manuales. Sin embargo, si se distribuye el proceso es posible ejecutarlo de manera automática. Con relación a la disponibilidad de herramientas de análisis el EBI ofrece una completa suite de análisis, la EMBOSS (*European Molecular Biology Open Software Suite*).

Existen diferentes interfaces gráficas para el uso de EMBOSS; la selección depende enteramente de las necesidades locales. PISE (*Pasteur Institute Software Enviroment*, http://www.pasteur.fr/recherche/unites/sis/Pise/), por un lado, ofrece la generación automática de interfaces para GCG y EMBOSS (http://www.emboss.org/), GPIPE (http://kun.homelinux.com/Pise/5.a/gpipe.html) permite la generación de *workflows* para todos los paquetes de análisis cuyas interfaces sean manejadas por vía de PISE. W2H (http://www.w2h.dkfz-heidelberg.de). Por otro lado, no automatiza la generación de interfaces de la misma manera en que lo hace PISE, pero igualmente su instalación por defecto provee capacidades de interfaz gráfica para los mismos paquetes que PISE. En ambos casos los requerimientos a nivel de cliente son mínimos: basta con contar con un navegador. En relación a las capacidades del servidor para aquellas instalaciones de SRS, la limitante está más bien dada por la capacidad de almacenamiento que por la velocidad del procesador o memoria RAM. El mismo servidor puede usarse para almacenar los bancos

de datos y para proveer acceso a las herramientas de análisis. En casos en los cuales se considere hacer análisis a gran escala de genomas completos es importante contar con un *cluster* de servidores. Considerando que la mayoría de las arquitecturas de *clusters* disponibles asumen que los equipos tendrán esa única destinación (no serán computadores de escritorio), es conveniente la generación de una tecnología que permita la utilización de equipos de escritorio para despachar desde ellos tareas y trabajos en ambientes *grid* (redes de procesadores prestando un servicio). La idea es dar un completo uso a la capacidad de cómputo del equipo de escritorio; por ejemplo, esta tecnología deber permitir llevar a cabo complejos análisis durante las noches, fines de semana y festivos. Las tareas se despachan, se inician, se paran cuando el equipo esté usándose (igualmente al detectar un movimiento del *mouse*) y se retoman las tareas en los puntos en los que se dejaron en el momento en que haya disponibilidad para hacerlo. Se trata de llevar mas allá el mismo paradigma que los proyectos SETI usan para computación masiva. Mac ofrece en este sentido una interesante solución; ésta tiene sin embargo el inconveniente de estar disponible solo para sistemas operativos Mac OX 10. Sin embargo, es interesante analizar soluciones novedosas para sistemas operativos tipo Windows y Linux, por ser estos los que mayoritariamente se usan en los equipos de escritorio.

La tecnología para implementar un ambiente como el anteriormente descrito no es costosa. En todos los casos es posible usar Linux como OS (sistema operativo). Los servidores requeridos pueden basarse en tecnología Intel. El costo estimado para un *cluster* de 10 servidores (cada equipo con 2 gigas en RAM y 80 Gigas en capacidad de almacenamiento hacen un total de 20 Gigas de RAM y 800 gigas de almacenamiento), no excede los 10 millones de pesos. Una capacidad de cómputo como la que hemos descrito es alcanzable paulatinamente en la medida en que siempre es posible añadir nuevos equipos al *cluster*, además de, claro está, de existir la posibilidad de incrementar las capacidades de los equipos del *cluster*. Un *cluster* como el descrito permitiría mantener bancos de datos locales, herramientas de análisis y dar cierta capacidad para ejecutar operaciones de análisis relativos a genómica comparativa. No se trata de un *cluster* para hacer computación a gran escala.

El manejo de la información interna, aquella generada por el laboratorio, es también posible llevarlo a cabo sobre herramientas *open-source*. La sistematización de datos en el ámbito de los laboratorios, LIMS, es un área de activa investigación por los retos tecnológicos que plantea. La publicación de esta información sobre Internet supone una unificación de procesos; ésta es posible de llevar a cabo y es igualmente un área activa de investigación y desarrollo. Consideremos un escenario en el cual investigadores en un laboratorio que cuente con un sistema como el descrito anteriormente deseen publicar una base de datos sobre Internet, es decir, hacerla públicamente disponible para otros investigadores. La opción de usar SRS como sistema de publicación es clara, pues garantiza igualmente la inmediata disponibilidad de esa nueva información sobre una red de servidores SRS que podrían incluir esa nueva base de datos sobre sus implementaciones. Es importante aclarar que SRS no es un LIMS, y por tanto el desarrollo del LIMS idealmente debería considerar su conectividad con la plataforma de manejo de bancos de datos.

Existe igualmente la tecnología *open-source* que permite el manejo ordenado de la publicación institucional por vía de la WEB. La estandarización del proceso de publicación organizacional sobre INTERNET puede lograrse mediante el uso de ZOPE (www.zope.org). ZOPE es un CMS (*Content Management System*) y básicamente proporciona un marco de trabajo fácil de usar donde la información se actualiza sin necesidad de tener más conocimientos que el manejo de un procesador de texto. De esta manera se simplifica el proceso de mantener un sitio corporativo y proveer un servicio bioinformático, basado en *software open-source*. Es importante aclarar que si bien la infraestructura de software no tiene costo directo si lo tiene el *know-how* asociado a esta clase de soluciones.

Analizar datos requiere un ambiente donde las diferentes operaciones, herramientas tanto de manipulación como de análisis de información, además de los filtros requeridos, estén disponibles de manera coherente. Las operaciones de análisis requieren en muchos casos un previo tamizaje de la información extraída de diferentes bancos de datos. Muchas operaciones básicas de bases de datos deben llevarse a cabo previamente a la etapa de análisis. Una vez se llegue al uso de herramientas de análisis,

diferentes operaciones de manipulación de la información siguen siendo necesarias, particularmente en labores propias de anotación de genomas. Ambas labores se entrecruzan; los límites, como se ha dicho antes, no son claros; pero sí es clara la necesidad de presentar un ambiente unificado sobre el cual el investigador pueda llevar a cabo sus experimentos.

¿Cómo se complementan mutuamente aquellas actividades de análisis y aquellas relativas a operaciones sobre bancos de datos? Hasta ahora hemos presentado un panorama no específico en el cual se han mencionado herramientas y bancos de datos. Un ejemplo en el cual se requiere una interacción entre ambas actividades es la localización de mutaciones puntuales. Esta clase de experimentos se inician con una proteína y residuos conocidos de interés. La primera pregunta es aquella relativa a la disponibilidad de la estructura proteica; esta actividad implica consultas sobre una o varias fuentes de información. Estas consultas pueden llevarse a cabo utilizando las interfaces que los diferentes servicios proveen; de contarse con los datos sobre un repositorio local este proceso podría ser automatizado y los resultados serían más exactos. En caso de no contarse con la estructura para una proteína o proteínas en particular, se deberá iniciar una búsqueda de estructuras similares. Esta búsqueda implica el uso de BLAST; el proceso podría soportarse si se contara con un ambiente en el cual los resultados pudieran ser manipulados (contextualizados). En el diseño de un experimento de detección de mutaciones puntuales los pasos que hemos descrito antes son solo algunos dentro de varios otros que conforman el flujo de trabajo.

¿Cuáles son algunos proyectos viables dentro del contexto colombiano en el área de la bioinformática? Existe una clara carencia de recursos, pero de igual manera existe una clara disponibilidad de material humano, y se tiene un *know-how* básico que permitiría eventualmente el desarrollo de proyectos de alto impacto en bioinformática. Antes de responder a esta pregunta conviene tener una definición operativa de bioinformática. Ella podría ser la siguiente: se trata de una disciplina que pretende brindar herramientas para la manipulación, el análisis y el manejo de información proveniente de proyectos genoma. Dentro de este contexto existen muchos problemas interesantes que implican, en nuestro concepto, más investigación tecnológica orientada al desarrollo de

*software* que a la investigación básica. La integración de la información en bioinformática es un área de gran interés, pues busca dar herramientas que permitan una fácil y rápida integración de diferentes tipos de datos en la medida en que ellos estén disponibles; el problema es que debido al alto nivel de interdependencia e interrelación de la información biológica, pequeñas variaciones afectan ostensiblemente el sistema como tal. Se requieren igualmente ambientes de *software* sobre los cuales las diferentes implementaciones de algoritmos puedan integrarse de manera tal que puedan usarlos investigadores con mínimos conocimientos de informática. La investigación en la integración de bancos de datos en bioinformática ha evolucionado de la simple integración de datos a la integración de información con el fin de extraer conocimiento por vía del uso de ontologías y técnicas de *procesamiento natural de lenguaje* (NLP por sus siglas en inglés).

¿Cómo tomar los requerimientos para desarrollar herramientas en biociencias? Las investigaciones en HCI (*Human Computer Interaction*) pretenden dar luces en este sentido: ¿cómo evaluar las interfaces que actualmente se tienen para determinar si estas efectivamente presentan metáforas entendibles para los usuarios? Por otro lado el simple hecho de levantar requerimientos para posteriores diseños en campos tan altamente fragmentados es de por sí un área de investigación (similar al diseño de tecnología para niños). ¿Es posible generar ambientes sobre los cuales se puedan modelar experimentos *in sillico*? Esta clase de investigaciones no es costosa, y sí es de alto impacto. La información está disponible sobre Internet y en su mayoría no es de acceso restringido.

(Recibido en mayo de 2004)

Alexander García Castro
Institute for Molecular Bioscience
The University of Queensland, Brisbane, Australia
*e-mail:* a.garcia@imb.uq.edu.au