ISSN 0121–1129

Experiencias de desarrollo en tareas de procesamiento y gestión de datos en bioinformática

Development experiences in processing and data management tasks on bioinformatics

Fecha de recepción: 5 de agosto de 2011
Fecha de aprobación: 20 de octubre de 2011

Alejandro Hadad*, Franco Simonetti**,
Luisina Pocay**, Walter Elias**

Resumen

En el presente documento se describen brevemente algunas experiencias de desarrollo relacionadas a tareas de procesamiento y gestión de datos para el área de bioinformática. En dicha área una de sus características es que los mismos suelen tener alta dimensionalidad, siendo esta una de las características que dificultan cualquier proceso de reconocimiento de patrones. Por otro lado en lo que respecta a tareas de gestión de datos, el profesional que trabaja en ámbitos de bioinformática usualmente necesita disponer de suficientes variantes en cuanto al manejo de diferentes formatos de almacenamiento y métodos de procesamiento, dada la naturaleza investigativa de su profesión que le requiere continuamente realizar ensayos in silico de manera no estandarizada. Este perfil de trabajo requiere herramientas con suficiente flexibilidad a fin de dar

Abstract

This document briefly describes some development experiences related to processing tasks and data management for the bioinformatics area. In this area one of its characteristics is that they tend to have high dimensionality, this is one of the characteristics which hinder the pattern recognition. On the other hand in regard to data management tasks, professional working in the bioinformatics fields usually needs to have sufficient variations in the management of different storage formats and processing methods, given the investigative nature of their profession that requires continuous testing in silico on a non-standardized way. This job profile requires tools with sufficient flexibility to support those tasks. Considering these two aspects in this work are shown first experience with regard to dimensional reduction strategies. On the other hand

* PhD (C) en Ingeniería, Mención en Sistemas de Información, UTN-FRSF, Argentina. Biongeniero, Profesor Investigador en la Facultad de Ingeniería, Universidad Nacional de Entre Ríos. Docente en Facultad de Ciencia y Tecnología, Universidad Autónoma de Entre Ríos. Oro Verde, Entre Ríos, Argentina.
** Licenciados en Bioinformática, Facultad de Ingeniería, Universidad Nacional de Entre Ríos, Argentina.

soporte a dichas tareas. Teniendo en cuenta estos dos aspectos en este trabajo se muestran en primer lugar una experiencia en relación a estrategias para la reducción dimensional. Por otro lado se muestra una experiencia en el diseño y desarrollo de un sistema flexible basado en pipeless y pipelines.

Palabras clave: Procesamiento, Gestión de datos, Bioinformática.

is an experience in designing and developing a flexible system based on pipeless and pipelines.

Keywords: Processing, Data Management, Bioinformatics.

I. Introducción

El creciente poder de procesamiento y sofisticación de las herramientas y técnicas analíticas ha dado como resultado, para la gestión y procesamiento de datos, la creación de estructuras y programas que proporcionan almacenamiento, funcionalidad y receptividad a las consultas, que van más allá de las posibilidades de las bases de datos destinadas a transacciones. A este poder en progresivo aumento se le ha unido una gran demanda para mejorar el rendimiento del acceso a datos que tienen las bases de datos. Muchos usuarios tan solo necesitan acceso de lectura a los datos, pero requieren un acceso muy rápido a un gran volumen de datos que puedan descargarse cómodamente en su computador personal. A menudo, esos datos proceden de varias bases de datos. Dado que muchos análisis realizados son concurrentes y predecibles, los vendedores de software y el personal de mantenimiento de sistemas han comenzado a diseñar sistemas para realizar estas funciones.

El área bioinformática se inscribe dentro de esta tendencia, pero presenta algunas particularidades; desde el punto de vista del procesamiento de datos, una de sus características es que estos suelen tener alta dimensionalidad, característica que dificulta cualquier proceso de reconocimiento de patrones. Por otro lado, en lo que respecta a tareas de gestión de datos, el profesional que trabaja en ámbitos de bioinformática usualmente necesita disponer de suficientes variantes en cuanto al manejo de diferentes formatos de almacenamiento y métodos de procesamiento, dada la naturaleza investigativa de su profesión. Este perfil de trabajo requiere herramientas con suficiente flexibilidad, a fin de dar soporte a dichas tareas.

Teniendo en cuenta estos dos aspectos, en este trabajo se exponen, en primer lugar, una experiencia en relación con estrategias para la reducción dimensional, y, en segundo lugar, una experiencia en el diseño y desarrollo de un sistema flexible basado en pipeless y pipelines.

II. Elementos de trabajo y metodologías

A. Selección de características

El objetivo de este desarrollo es analizar el espacio de representación de los modelos de la actividad durante los procesos de transcripción de la proteína p53 (activas vs. inactiva) basado en datos de simulaciones biofísicas. Los datos utilizados (Fig. 1) para las experiencias son de disponibilidad pública y fueron etiquetados mediante determinaciones in vivo [1,2,3].

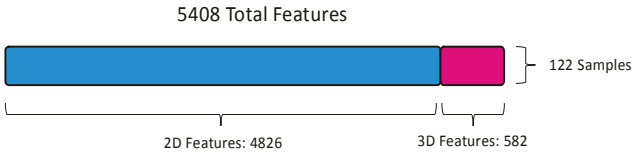

Fig. 1. Características 2D y 3D y número de muestras

En este trabajo se realizó un análisis de las características del set de datos con el fin de reducir la dimensionalidad del espacio de trabajo, y para extraer información sobre el comportamiento de estas variables. El set de datos elegido está compuesto por características derivadas de modelados atómicos por homología y simulaciones de dinámica molecular. Estas características se corresponden con mapas de propiedades de superficie (2D) y mapas estructurales de distancia (3D).

La metodología de selección de variables consistió en la utilización de algoritmos de teoría de la información, cuyos resultados fueron rankeados, y seleccionada una porción de dicho *ranking*. Las variables finales para los sets de entrenamiento fueron escogidas teniendo en cuenta la combinación de los métodos de selección previamente utilizados (Fig. 2).

Se utilizó un clasificador SVM con núcleo gaussiano, al igual que en trabajos anteriores, a fin de comparar únicamente el efecto de la selección de variables.

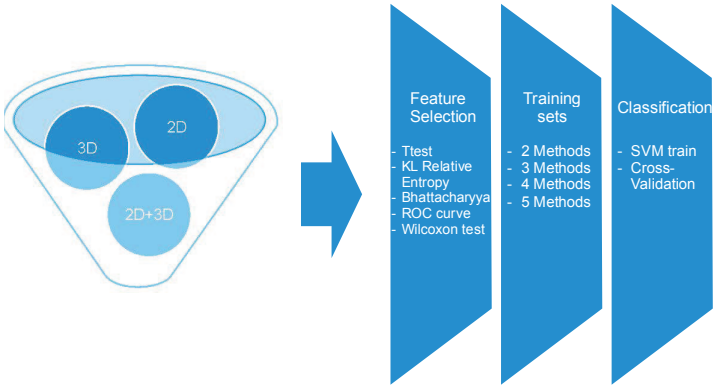

Fig. 2. Ensayos realizados para distintos conjuntos de características y métodos de selección

B. Gestión de datos

Pipeless

Un pipeless puede considerarse como un agente autónomo, experto en la realización de una tarea específica; recibe una o más entradas que pueden provenir del propio usuario o de otros pipeless. Las funciones programadas del sistema en lenguaje PHP tienen la capacidad de interpretar los datos ingresados al pipeless y producir una cadena de eventos que derivan en la salida deseada. Puede considerarse a un pipeless desde la línea de comandos de un programa con sus parámetros y modificadores, hasta una sentencia de consulta SQL.

Pipeline

Los pipelines constituyen una conexión secuencial de pipeless, en la que la salida de un pipeless corresponde a la entrada del siguiente. El pipeless de inicio puede contener varias entradas (por ejemplo, una secuencia múltiple de nucleótidos). Se ha dejado la posibilidad para que el sistema pueda ser extendido a una estructura de pipelines tipo arbórea, que permite la ejecución de varios pipelines tomando salidas múltiples.

El procesamiento de grandes volúmenes de información provenientes de datos de secuenciación de diferentes especies propone desafíos importantes a las ciencias de la computación y a la ingeniería. La optimización de los algoritmos resulta tan importante como la optimización de los recursos computacionales y de la estructura de comunicación asociada. Una de las dificultades más comunes para este tipo de tareas es que hay que esperar la finalización en la ejecución de una aplicación para cambiarle el formato a los resultados con una estructura consistente para las entradas de otra aplicación.

La estrategia de pipelines expone una solución que permite desencadenar una serie de procesos controlados realizando incluso pasos intermedios de formateo de datos según los requerimientos de cada pipeless en particular; esto optimiza tanto el uso de recursos informáticos como el tiempo, disminuyendo críticamente el tiempo muerto entre ejecución y ejecución [4,5,6].

Desde el punto de vista biológico se relevaron tres situaciones, o casos representativos, que fueron tenidas en cuenta durante el proceso de diseño. Uno de los casos fue realizar una comparación fenotípica; para ello, el biólogo (usuario) abre el sistema y elige comparación fenotípica; luego, dentro de los fenotipos, selecciona una propiedad en particular, por ejemplo, patogenicidad; realiza una observación de la distribución de patogenicidad entre las distintas cepas, y obtiene estadísticos de la distribución conjunta de variables fenotípicas (presencia o no de un gen codificante para una proteína de pared celular propia de bacterias patógenas). Enseguida elige crear una categoría de comparación, en este caso, patogenicidad;

esto implica que se van a buscar diferencias entre los organismos patógenos y no patógenos. Elige dentro de lo genotípico-genómico una de las siguientes características: genómicas agregadas sencillas, como tamaño de los genomas, número de genes, contenido de GC, uso de codones, uso de aminoácidos y distribución de ARNt; presencia de algún gen en particular; presencia de alguna vía metabólica; estructura de vías metabólicas, es decir, quienes tienen los mismos genes y quienes distintos; clases de proteínas. Como último paso, realiza comparaciones, visualiza los resultados y selecciona algunos datos para continuar con los análisis, guardando la selección con alguna etiqueta.

Otro caso considerado estuvo orientado a caracterizar funciones en un contexto evolutivo (integración). Para ello, el usuario, abre el sistema y selecciona las secuencias que desea (con blastP); lleva a cabo un alineamiento de secuencias (con ClustalX, T-Coee, Muscle), y visualiza los alineamientos (con GeneDoc). Luego realiza un análisis filogenético (con Mega4) y determina así las estructuras secundarias, 3D, dominios (con IterProScan, Swiss-Model). Identifica las interacciones sistemáticas de proteínas dentro del organismo seleccionado (UVCluster), realiza un análisis y visualiza los resultados (con TreeView). Selecciona los datos para continuar con los análisis y guarda la selección con alguna etiqueta.

El último caso presenta como objetivos analizar un genoma bacteriano y buscar genes candidatos potenciales para el desarrollo de vacunas. El análisis in silico de genomas completos tiene la potencialidad de proveernos con las bases para un entendimiento global de la genética, la bioquímica, la fisiología y la patogénesis de un microorganismo dado. Como el número de genomas secuenciados se incrementa anualmente, es posible ahora comparar un grupo significativo de secuencias genómicas entre bacterias relacionadas evolutivamente; en particular, el análisis de variabilidad genética entre un patógeno y especies no patogénicas relacionadas. Este análisis genera rápidamente una colección de genes potencialmente responsables de la adquisición de virulencia, y, por tanto, con implicaciones prácticas en el diseño de vacunas.

El biólogo abre el sistema y selecciona secuencias; elige los genes e identifica los que codifican para proteínas y la asignación de la función probable a cada una de las secuencias. Es decir, que el biólogo realiza una anotación del genoma; a través de Blast, anota los genes identificados, buscando secuencias homólogas en bases de datos; realiza la localización celular del producto de los genes anotados. El biólogo lleva a cabo esta actividad para identificar los candidatos. Finalmente, gracias a que existe un gran número de microorganismos con sus genomas secuenciados -incluso dentro de una misma especie-, el biólogo verifica si los genes candidatos detectados están conservados. Para ello localiza las secuencias ortólogas en los genomas relacionados, realiza un alineamiento de secuencias y determina el grado de conservación de cada candidato.

III. Resultados

A. Selección de características

Los mejores rendimientos fueron obtenidos con la combinación de 4 métodos de selección de características en el subconjunto de características 2D con el menor número de vectores de soporte (100). Para otras combinaciones de características se obtuvieron desempeños similares, pero con casi todos las muestras como vectores de soporte.

En el set de características definido por nuestros 4 métodos de selección sobre las características 2D, sobre un ranking del 10% del total de ellas, se observó que se utilizaba la menor cantidad de vectores de soporte para encontrar el hiperplano separador; sin embargo, en general se necesitaba un gran porcentaje de vectores de soporte para la mayoría de los clasificadores (80% to 90%). Esto puede significar que el clasificador SVM falla en distinguir características comunes de cada clase, porque no son adecuadas y, por lo tanto, sobreajustan los datos de entrada. El resultado será el que clasificará correctamente aquellos ejemplos que son muy similares a aquellos pertenecientes al set de entrenamiento. En este caso, si la variabilidad de los datos fuera mayor, la tasa de error aumentaría.

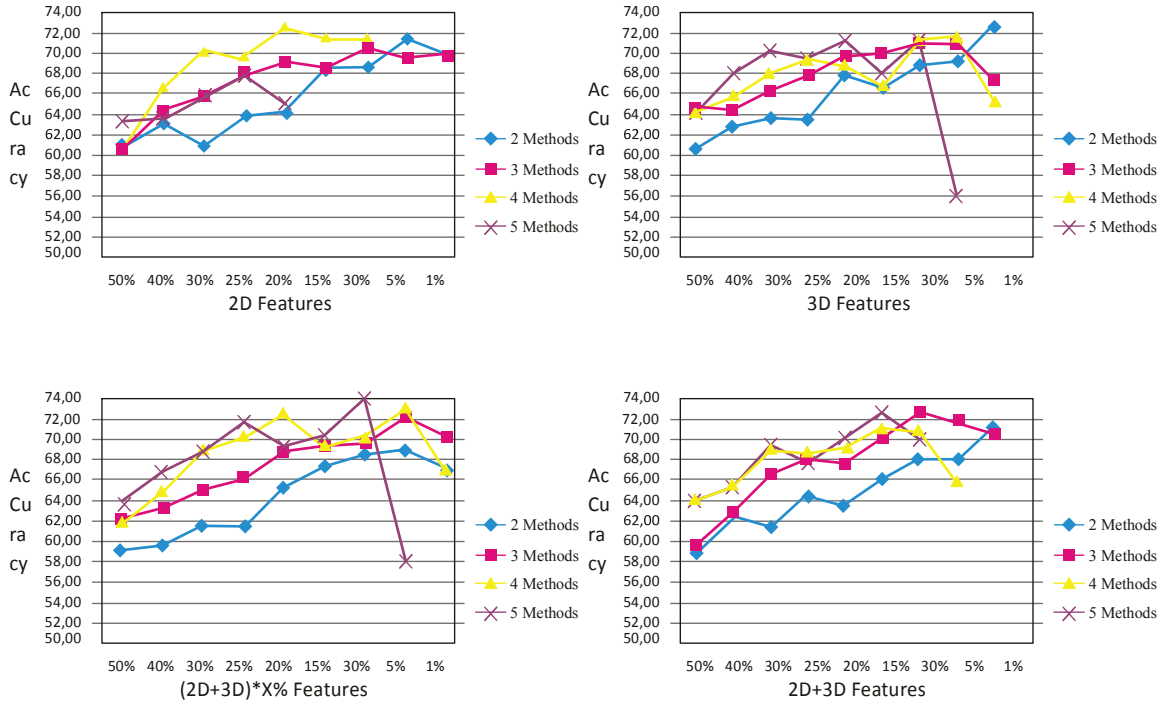

Fig. 3. Performances para todas las combinaciones ensayadas

B. Gestión de procesos y datos

Básicamente, este es un proyecto de desarrollo donde se conjugan una primera etapa de revisión de la bibliografía existente y un diseño y desarrollo del software básico. Implementada la metodología de arquitectura de software de pipelines (para los casos considerados) que provee potencialidad adicional al sistema, se logro una mayor interacción y flexibilidad al momento de realizar el análisis correspondiente. La base de datos (Fig. 4) cuenta con una estructura relativamente simple y estable, básicamente con información de anotación genómica y fenotípica (propiedades composicionales e información eco-fisiológica disponible).

Sobre esta base de datos se desarrolla un marco de análisis que permita combinar en forma flexible herramientas bioinformáticas disponibles, tales como suite del NCBI, EMBOSS y programas de alineamiento de secuencias, filogenéticos, entre otros.

Los diferentes pipelines definidos por los usuarios, así como un conjunto inicial predeterminado de ellos a partir de las consideraciones sobre los análisis más usuales en el campo, pueden ser grabados para ser usados o redefinidos tantas veces como sea necesario.

Una base de datos independiente puede ser definida por los usuarios de acuerdo con sus necesidades de análisis específicas, permitiendo guardar la información relevante y usarla en nuevas comparaciones, con información obtenida tanto a partir del análisis de datos ya almacenados o datos completamente diferentes (por ejemplo, epidemiológicos, bioquímicos, etc.).

La interfaz web permite el acceso y ejecución de los pipelines por esta vía en una arquitectura cliente-servidor (Fig. 5).

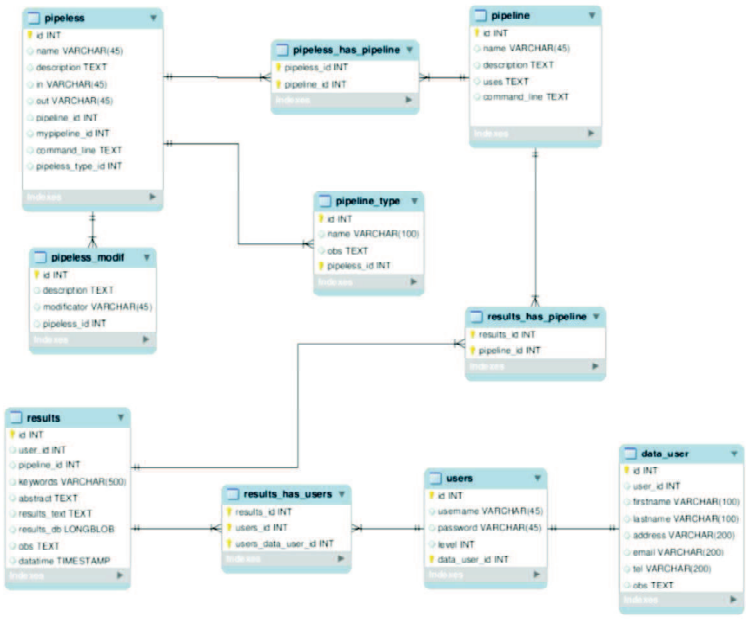

Fig. 4. Base de datos de pipelines

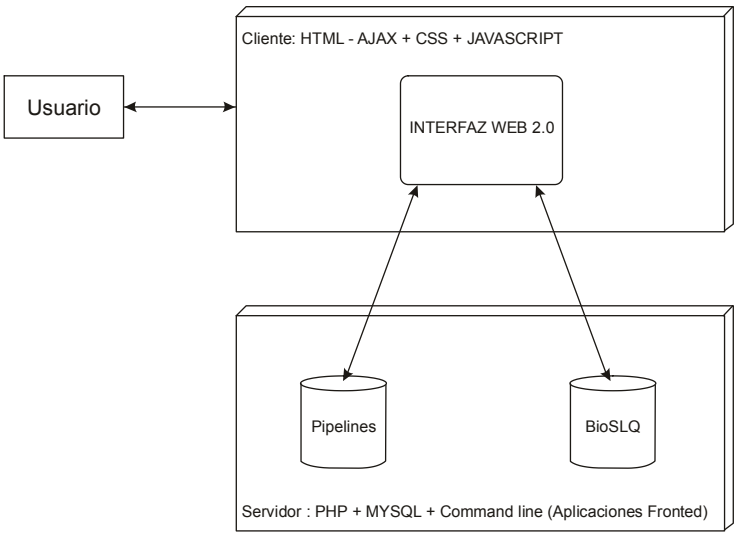

Fig. 5. Arquitectura de dos capas

Este sistema de gestión está destinado a grupos de investigación que cuenten con la necesidad de una herramienta que les permita realizar el análisis comparativo entre una secuencia genómica aportada o existente en la base de datos y otras secuencias del mismo género, para de esa forma dar respuesta a las preguntas biológicas.

IV. Conclusiones

En relación con el análisis de extracción de características se observaron diferencias en el rango del 20% en el desempeño de los clasificadores con los diferentes grupos de características, lo cual es comparable con lo reportado en trabajos que utilizan el mismo data set.

Dado el gran número de vectores de soporte utilizados por el clasificador, y los porcentajes alrededor del 70% reportados en trabajos de referencia, puede ser que el set de características elegido alcance para caracterizar los datos; sin embargo, por la forma en que están distribuidos, será más difícil delimitar una clase de otra y se necesitarán más ejemplos para encontrar el hiperplano separador.

En cuanto al sistema de gestión de datos, la estrategia de pipelines permitió abordar esto de forma efectiva, generando una nueva forma de abordaje a problemas complejos, especialmente en lo relacionado a la genómica comparativa. La posibilidad de ampliar el uso de esta herramienta en grupos de investigación del ámbito público o privado es una opción que vale la pena considerar. Debido a la naturaleza genérica del desarrollo, la creación de nuevos pipelines está limitada solo por la capacidad del usuario para generar los flujos de trabajo adecuados que permitan su ejecución; esto promueve su utilización en otros ámbitos de la genómica, proteómica o disciplinas similares.

En el futuro será importante tener en cuenta que estas aplicaciones suelen encontrarse con importantes problemas de escalabilidad, disponibilidad, seguridad e integración. Para solventar estos problemas es posible que sea conveniente pasar a una arquitectura de 3 capas: una capa para guardar los datos (base de datos), una capa para centralizar la lógica y, por último, la interfaz gráfica de usuario; si establecemos una separación entre la capa de interfaz gráfica (cliente), replicada en cada uno de los entornos de usuario, y la capa modelo, quedaría centralizada en un servidor de aplicaciones. De esta manera, la centralización de los aspectos de seguridad y transaccionalidad sería responsabilidad del modelo, manteniendo la sencillez de los clientes, no replicando la lógica. Con este enfoque es posible distribuir el procesamiento de manera tal que el sistema radique en un medio físico con su base de datos, los datos sean procesados en un sistema de alto rendimiento y la interfaz sea ejecutada en la PC del usuario

Referencias

[1] S. A. Danziger, R. Baronio, L. Ho, L. Hall, K. Salmon, G.W. Hatfield, P. Kaiser, and R. H. Lathrop. «Predicting Positive p53 Cancer Rescue Regions Using Most Informative Positive (MIP) Active Learning». *PLOS Computational Biology*, 5(9), e1000498, 2009.

[2] S. A. Danziger, J. Zeng, Y. Wang, R. K. Brachmann, and R. H. Lathrop. «Choosing where to look next in a mutation sequence space: Active Learning of informative p53 cancer rescue mutants». *Bioinformatics*, 23(13), 104-114, 2007.

[3] S. A. Danziger, S. J. Swamidass, J. Zeng, L. R. Dearth, Q. Lu, J. H. Chen, J. Cheng, V. P. Hoang, H. Saigo, R. Luo, P. Baldi, R. K. Brachmann, and R. H. Lathrop. «Functional census of mutation sequence spaces: the example of p53 cancer rescue mutants». *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM,* 3, 114-125, 2006.

[4] N. Rego, H. Naya, G. Lamolle, F. Alvarez-Valin. «Evolutionary and comparative genomics of Leptospira». *RECIIS* 1(2 Supl): 321-328, 2008.

[5] M. Picardeau *et al.* «*f* Genome Sequence of the Saprophyte Leptospira biexa Provides Insights into the Evolution of Leptospira and the Pathogenesis of Leptospirosis». *PLoS ONE* 3(2): e1607.doi:10.1371 journal.pone.0001607.

[6] D. Frank. «*f*XplorSeq: A software environment for integrated management and phylogenetic analysis of metagenomic sequence data». BMC Bioinformatics, 9: 420 doi:10.1186/1471-2105-9-420, 2008.