**Aplicación web para descarga y manejo local en Bases de datos MySql de datos bacterianos.**

*Kenneth Brenes Gatgens, A20869, II semestre 2013. Bases de Datos Biológicas, profesor: Allan Orozco*

**Introducción**

A continuación se describe el desarrollo de una herramienta web enfocada en tecnologías de código abierto que básicamente indexa la información de bacterias proveniente del repositorio FTP abierto provisto por el National Center for Biotechnology Information.

La herramienta está en etapa *beta* y presenta funcionalidad básica. Permite buscar por nombre a una bacteria específica y una vez encontrada, baja la información a una base de datos local. La idea es que esta sea la base de una aplicación robusta para manipular y general búsquedas de cadenas-patrones en genes, entre muchos otros usos que podría tener.

**Marco Teórico**

Se quiere facilitar el trabajo de los investigadores biotecnólogos ofreciendo una herramienta que haga más rápida la manipulación de grandes volúmenes de datos.

El objeto de estudio son las fuentes de datos bacterianas provistas por el NCBI. La idea es que estos datos sean aún más manipulables, dado que estando en línea, en archivos planos, y en servidores remotos se dificulta el tratamiento de las mismos.

La forma en que el NCBI ofrece sus datos es ordenada, pero para ciertas tareas específicas como cruces, comparaciones, o simplemente lecturas de secuencias se dificulta el camino por el simple hecho de no tener de forma centralizada la información mínima básica de una especie en particular. Entonces, aunque se cuenta con la información, se necesita tiempo del investigador para identificar y conjuntar los datos, y es esto lo que se desea ahorrar por parte de la aplicación propuesta.

El aporte de una interfaz simple, limpia, y apegada a los estándares de desarrollo web vendría a facilitar esta tarea de investigación en particular, y aunque en la etapa *beta* se contemplan sólo el análisis de bacterias, con simples implementaciones futuras se podría analizar estructuras de datos relacionadas con otros organismos. Se quiere una aplicación mantenible y que pueda crecer con esfuerzo colaborativo.

1

Además la idea de hacer un diseño multiplataforma web(móviles y tabletas) era básica. Por este motivo, la interfaz es responsiva y se adapta al tamaño de la pantalla en que esté siendo visto de forma elegante.

**Metodología**

La aplicación está creada usando tecnología LAMP (Linux-Apache-MySQL- PHP) lo que no agrega costos sobre las herramientas de desarrollo.

A nivel de código, no usa un Framework particular aunque si se trata de separar de forma ordenada e intuitiva en copas tipo MVC(Model-View-Controller) para facilitar la comprensión, mantenimiento y futuros desarrollos. Además sigue el esquema de documentación PHP Doc para estar apegado al estándar. Los nombres de las funciones, variables, clases y documentación en general está en inglés para facilitar una comprensión más universal, y en general para seguir modelos de implementación más comunes.

En Front-end se utiliza *Bootstrap* como framework. Se escoge este por ofrecer gran cantidad de herramientas que hacen una implementación ligera de HTML y CSS, es multiplataforma y responsivo.

La base de datos tiene arquitectura SQL, y usa como motor MySQL Se presentan por ahora sólo tres tablas: *Genes, Sequences* y *Species,* todas con una integridad referencial básica. Más adelante se explicará propiamente sobre los campos e información relevante usada para el diseño.

La interfaz cuenta con dos páginas:

1. ***Inicial.*** Presenta un campo de entrada de datos, donde el sistema tiene la capacidad de usar un autocompletar a la hora de filtrar e ingresar los datos. Esto con el fin de facilitar el deletreo de palabras que tienen cierta dificultad de memorizar y(o digitar) como en este caso nombres científicos de bacterias. Por ahora muestra sólo un botón: ‘”Guardar Localmente” que dispara el evento de análisis de datos y guarda en la base de datos.

Se ve de la siguiente forma:

2

2. **Resultados.** Ventana que despliega los resultados, una vez que han sido guardados en la base de datos local. Separa los datos por especie, secuencia y genes. Tiene la opción de ir a la página *Inicial* para ejecutar otras búsquedas. Se ve de la siguiente forma:

3

Resolución en Ipad:

4

Resolución en Iphone:

El código de la aplicación en su totalidad es de libre descarga en github, y el URL de enlace es:

*https://github.com/kgatjens/bio\_ftp.git*

Ahí se puede encontrar un dumb de la base de datos, y el código versionado completo.

La estructura de archivos es la siguiente:

5

**Análisis de Datos**

La indexación de los datos se hace partiendo de este folder: *ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria ,* el campo de entrada del nombre de la bacteria a nivel de cliente(navegador) va a contener todos los nombres de los directorios del FTP, estos nombres a fin de cuentas, son los que serán tomados como nombre de las bacterias. El sistema está configurado para empezar al análisis usando esa ruta de FTP, pero es un parámetro modificable en una línea en caso se quisiera cambiar el tipo de organismos.

Una vez seleccionada la bacteria en particular, y después de presionar el botón para el guardado local, inicia el análisis de los archivos respectivos de donde se sacará la información.

Los archivos manipulados por bacteria seleccionada son:

* \*.GeneMark-2.5m
* \*.ptt
* \*.GeneMarkHMM-2.6r

Todos los anteriores tienen la información requerida a nivel interno. Se crea una lógica para analizar sintácticamente y extraer solo los información necesaria. Con fines ilustrativos solo se están extrayendo **diez** genes por secuencia, con el fin de no hacer el procesamiento lento. Pero este cantidad es fácilmente modificable en un parámetro, e inclusive se puede no poner límite para que ingrese todas las secuencais del FTP.

Los campos extraidos son:

|  |  |  |
| --- | --- | --- |
| **Tabla: Genes** | **Tabla: Sequences** | **Tabla: Species** |
| SpeciesNo | SpeciesNo | SpeciesNo |
| SequenceNo | SequenceNo | SpeciesID |
| GeneNo | SequenceID | SpeciesUid |
| GenePID | SequenceDesc | Finished |
| id | SequenceLength |  |
| GeneName |  |  |
| GeneSynonym |  |  |
| GeneCode |  |  |
| GeneCOG |  |  |
| GeneProduct |  |  |
| GeneRNA |  |  |
| GeneGC |  |  |
| GeneStrand |  |  |
| GeneStart |  |  |
| GeneEnd |  |  |
| GeneLength |  |  |
| GeneKey |  |  |

6

Se crea la relación entre estas tablas de forma que una secuencia puede tener varios genes. Un dump de la base de datos puede ser descargado en directorio ***/dump\_db/*** dentro del repositorio versionado.

**Conclusiones, Alcances y Limitaciones**

El sistema está abierto a futuros desarrollos, y la idea es complementar trabajo colaborativo para llegar a darle robustez y ofrecer nuevas funcionalidades.

Por ahora no hace manejo de excepciones, y está escueto en mensajes informativos y de error, pero esto se ve como un desarrollo pendiente.

El ser una plataforma implementada con tecnologías de código abierto, la hace de fácil inversión, y manipulación. Fue pensando así en un inicio para que el desarrollo sea extendible, y las excusas de costos por parte de interesados y o empresas no sea una traba.

La aplicación no ha tenido un proceso de calidad ni de pruebas adecuado, por lo tanto puede presentar errores en tiempo de ejecución que no han sido manejados previamente por ser casos no probados. Si bien la estructura de archivos es muy similar, pueden haber casos con algunas bacterias en donde el análisis de los archivos sea diferente y no se comporte conforme el patrón(no se probaron todos los casos) y esto pueden generar que el procesamiento pueda ser lento o no dar resultados.

El diseño por ser responsivo, no limita la cantidad de dispositivos y plataformas que pueden ser usadas. Este es un factor de gran importancia por la gran trascendencia de el uso de aparatos móviles, y no debía ser una limitación para el desarrollo en tiempos actuales.

**Referencias**

* Copyright © 2001-2013 The PHP Group http://php.net/docs.php
* Wikipedia® Patrones de Diseño, 9 de Diciembre 2013

http://en.wikipedia.org/wiki/Model%E2%80%93view%E2%80%93controller

* Ubuntu documentation, 28 de marzo 2013 https://help.ubuntu.com/community/ApacheMySQLPHP
* Bootstrap, v3.0.3, http://getbootstrap.com/getting-started/
* MySQL © 2013, Oracle Corporationhttp://www.mysql.com/products/enterprise/

7

National Center for Biotechnology Information, 28 de octubre 2009 ftp://ftp.ncbi.nih.gov/genomes/Bacteria/

8