ProSA: una interfaz de Web basada en Perl para el análisis de secuencias

Mauricio Herrera Cuadra
arareko@campus.iztacala.unam.mx

1. Introducción

1.1 El análisis de secuencias y su importancia

La información funcional y hereditaria de un organismo se encuentra almacenada en moléculas de DNA, RNA y proteínas, todas estas macromoléculas son cadenas lineales compuestas de moléculas más pequeñas. Estas macromoléculas son ensambladas a partir de un alfabeto fijo de compuestos químicos bien conocidos: el DNA está formado por cuatro desoxirribonucleótidos, el RNA está formado por cuatro ribonucleótidos y las proteínas están formadas por 20 aminoácidos. Debido a que estas macromoléculas son cadenas lineales de compuestos definidos, pueden ser representadas como secuencias de símbolos. Estas secuencias pueden ser entonces comparadas para encontrar similitudes que sugieran que las moléculas están relacionadas por su forma o función.

Es importante recordar que una secuencia biológica (DNA, RNA o proteína) posee una función química, pero cuando esta es reducida a un código de letras sencillas funciona también como una etiqueta única, casi como un código de barras. Desde el punto de vista de la tecnología de la información, la información de las secuencias es invaluable. La etiqueta de la secuencia puede ser aplicada a un gen, su producto, su función, su rol en el metabolismo celular, etc.

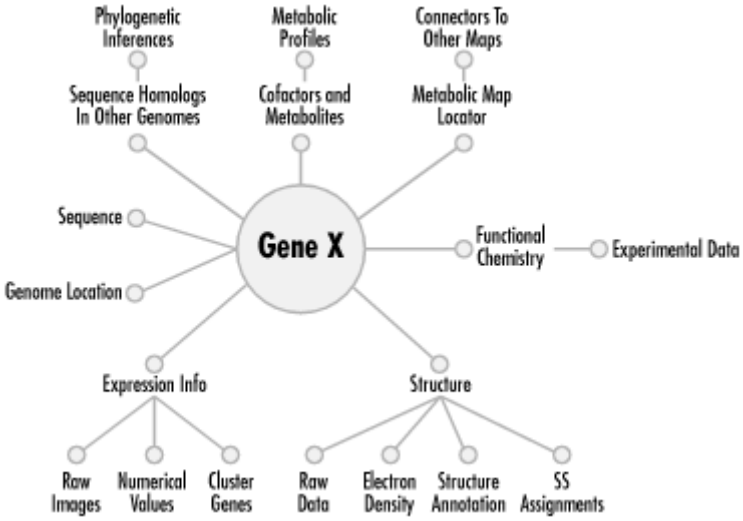

Figura 1. Posible información asociada a un solo gen

Sin embargo, la cuestión más importante acerca de estas etiquetas, es que no solamente identifican un gen particular; también contienen patrones biológicamente significativos, que permiten comparar diferentes etiquetas, conectar información, y hacer inferencias. Así que no solamente las etiquetas pueden conectar toda la información acerca de un gen, éstas pueden servir para conectar información sobre genes que son ligera o drásticamente diferentes en su secuencia.

Los datos de las secuencias de genes son el más abundante tipo de información, y existe un gran conjunto de métodos y herramientas computacionales que pueden ayudar a analizar los patrones contenidos en dicha información. La comparación de secuencias de genes, o análisis de secuencias biológicas, es uno de los procesos utilizados para comprender la evolución de las secuencias. Es una disciplina importante dentro de la biología computacional y la bioinformática.

1.2 La investigación biológica a través de la World Wide Web

La Internet ha cambiado completamente la forma en que los científicos buscan e intercambian información. La información que antes era comunicada en papel ahora es digitalizada y distribuida a partir de bases de datos centralizadas, las revistas ahora son publicadas "en línea", y casi cualquier grupo de investigación posee un Sitio Web que ofrece de todo, desde publicaciones hasta descargas de software y servicios automatizados de procesamiento de datos.

Los científicos utilizan los servicios Web en Internet para la mayoría de los análisis de datos hoy en día. Esto es debido a su accesibilidad, interfaz simple de documentos y formularios, y frecuentemente servicios gratuitos que proveen muchas herramientas de análisis y bases de datos actualizadas. Aún cuando las interfaces de Web para los análisis de biología molecular no siempre son la mejor opción, si éstas son capaces de realizar el trabajo, son preferibles a un programa ejecutándose bajo algún sistema operativo específico.

La mayoría de los usuarios encuentran problemas al utilizar programas para el análisis de secuencias. No solamente son difíciles de aprender debido a los parámetros, sintaxis y semántica, sino a que muchos son diferentes. Debido a esto, los programadores se han dedicado a construir interfaces de Web que simplifiquen el aprendizaje y utilización de dichos programas, un claro ejemplo de dicha tendencia son interfaces como: *Virtual PCR* y *WebPHYLIP*. Inclusive se han desarrollado sistemas avanzados, tales como: *Pise*, que permiten la generación de interfaces de Web a partir de programas de biología molecular más sencillos.

1.3 La base de datos PROSITE

Un motivo es una región o porción de una secuencia de proteína que posee una estructura específica y es funcionalmente significativa. Las familias de proteínas a menudo son caracterizadas mediante uno o más de tales motivos. La detección de motivos en proteínas es un problema importante puesto que los motivos portan y regulan varias funciones, y la presencia de motivos específicos puede ayudar a clasificar una proteína.

PROSITE es una colección de descriptores de motivos dedicada a la identificación de familias de proteínas y dominios. Los descriptores de motivos utilizados en PROSITE son patrones o perfiles, los cuales han sido derivados a partir de alineamientos múltiples de secuencias homólogas. Esto proporciona a estos descriptores de motivos la notable ventaja de identificar relaciones distantes entre secuencias que hubieran pasado inadvertidas mediante alineamiento simple de secuencias. Los patrones y perfiles poseen tanto ventajas como desventajas, los cuales definen su área de aplicación.

PROSITE es un método para determinar cual es la función de proteínas no caracterizadas que han sido traducidas de secuencias de cDNA o DNA genómico. Esta base de datos está elaborada de tal forma que con herramientas computacionales apropiadas, pueda ser rápido y factible el identificar a qué familia conocida de proteínas (si la hay) pertenece una nueva secuencia. En algunos casos, la secuencia de una proteína desconocida se encuentra lejanamente relacionada con cualquier proteína de estructura conocida para poder detectar su semejanza por medio de alineamiento de secuencias completas. Sin embargo, puede ser identificada por la presencia en su secuencia de un bloque particular de tipos de residuos, diversamente conocidos como patrones, motivos, firmas, o huellas digitales. Estos motivos

sobresalen debido a los requerimientos particulares en la estructura de regiones específicas de una proteína, los cuales pudieran ser importantes, por ejemplo, por sus propiedades de anclaje, o por su actividad enzimática.

PROSITE se encuentra disponible como una serie de archivos de texto que proveen los datos, además de documentación. El sitio de PROSITE (*http://www.expasy.org/prosite/*) está provisto de una interfaz de usuario que permite indagar en la base de datos y examinar la documentación. La base de datos también puede obtenerse para instalación local a través del sitio FTP de PROSITE. Su utilización es gratuita para usuarios no comerciales.

1.4 Perl y su aplicación en Bioinformática

Una gran parte de la Biología Computacional consiste de tareas frecuentes de procesamiento de textos, tales como la manipulación de cadenas, concordancia de expresiones regulares, traducción de archivos, e interconversión de formato de datos. Por consiguiente, muchos desarrolladores en la comunidad bioinformática hacen uso extenso del lenguaje de programación Perl, el cual sobresale en dichas tareas.

Perl es popular entre los biólogos debido a su carácter práctico. La información biológica en las computadoras tiende a estar organizada en archivos de texto o en bases de datos relacionales. Cualquiera de estas fuentes de datos es fácil de manejar con programas en Perl. Perl se ha convertido en una especie de fenómeno en el área, puesto que muchos biólogos lo encuentran como un lenguaje fácil de aprender que posee muchas de las herramientas que ellos necesitan: en particular su soporte para el procesamiento de textos y expresiones regulares lo hacen adecuado para tareas complejas de traducción de textos (comunes en bioinformática).

Perl ha madurado de un simple lenguaje de “script” a un poderoso ambiente de programación tanto para el estilo procedimental como para el orientado a objetos. Mientras que sigue siendo utilizado para crear programas simples “desechables”, también se utiliza para diseñar aplicaciones complejas, modulares, bien documentadas y mantenibles. La facilidad de utilización de Perl para una variedad de tareas, tanto de alto nivel como para programación de CGI, es inigualable.

Un ejemplo sobresaliente del papel que ha jugado Perl en bioinformática, es cuando permitió a los científicos del Proyecto Genoma Humano el intercambiar datos y comparar los resultados que se estaban produciendo en 2 diferentes centros de secuenciamiento.

2. Justificación

El análisis de secuencias es una de las metodologías más utilizadas en bioinformática y recientemente en biología molecular, por lo que es importante el desarrollo de herramientas computacionales adecuadas y eficientes para llevar a cabo el trabajo.

Hoy en día es posible realizar muchos de estos análisis mediante herramientas en Internet que facilitan la utilización y aprendizaje de estas metodologías, esto es a través de interfaces sencillas para los usuarios nuevos y al mismo tiempo poderosas para los usuarios avanzados. La mayoría de estas herramientas utilizan el lenguaje de programación Perl debido a su alta eficiencia para el procesamiento de textos y desarrollo de aplicaciones Web.

La base de datos PROSITE es una de las más conocidas y utilizadas para la identificación de dominios funcionales en secuencias de proteínas. Existen algunas herramientas que realizan búsquedas dentro de esta base de datos, desafortunadamente estas búsquedas están limitadas a que el usuario introduzca

secuencias de proteína únicamente, excluyendo la posibilidad de hacer búsquedas a partir de secuencias de nucleótidos, las cuales son más usuales de obtener en los experimentos de laboratorio.

Resulta necesario desarrollar una aplicación que permita a los usuarios realizar búsquedas en la base de datos a partir de ambos tipos de secuencias (nucleótidos y/o proteína), para así poder obtener un mayor beneficio tanto de la base de datos como de las secuencias obtenidas en el laboratorio.

3. Objetivos

3.1 Objetivo General

Desarrollar una interfaz de Web para el análisis de secuencias de nucleótidos y/o aminoácidos que utilice la base de datos PROSITE para la búsqueda de dominios proteínicos conocidos en ellas.

3.2 Objetivos Particulares

- Elaborar un diseño de software adecuado para facilitar su administración y utilización.
- Utilizar para su implementación el lenguaje de programación Perl y evaluar su eficiencia.
- Evaluar la facilidad de utilización de la interfaz y de interpretación de sus resultados.
- Determinar la veracidad de los resultados proporcionados por la aplicación.

4. Metodología

4.1 Diseño de ProSA

ProSA es una abreviatura para "Protein Sequence Analyzer". La idea general de esta aplicación es que sea una interfaz de Web que permita a los usuarios realizar análisis de secuencias de nucleótidos y/o aminoácidos mediante búsquedas en la base de datos PROSITE, con la finalidad de poder predecir a que posible familia de proteínas pertenece una secuencia obtenida en el laboratorio.

Para llevar a cabo la implementación de esta aplicación, se requirió el planteamiento del siguiente diseño, el cual consta de 3 partes principales:

1. Un script que actualice regularmente la base de datos PROSITE.
2. Una interfaz de Web para el usuario.
3. Una aplicación CGI que realice el análisis e imprima los resultados al usuario.

4.2 Implementación de ProSA

Una vez elaborado el diseño de la aplicación, se procedió a su implementación para la cual se utilizó una PC con las siguientes características:

- Sistema operativo: FreeBSD 4.8
- Interprete de Perl 5.8
- Servidor de Web: Apache 1.3.27

La aplicación se escribió en un editor de texto con apoyo de la bibliografía más conocida sobre Perl, expresiones regulares y la reciente bibliografía sobre Perl para bioinformática.

Cabe mencionar que la aplicación pudo haber sido desarrollada en cualquier otro sistema operativo (Linux, Mac OS ó Windows) que contara con un intérprete de Perl y un servidor de Web. La elección tomada para este trabajo fue por decisión del autor debido a varias razones, dentro de las cuales destacan:

1. El robusto soporte de FreeBSD para el desarrollo de aplicaciones Web.
2. La facilidad de administración y seguridad de Apache.
3. FreeBSD, Perl y Apache son proyectos de software libre que cuentan con una gran cantidad de documentación, soporte y usuarios, además de que son gratuitos.

4.3 Obtención de secuencias para análisis con ProSA

Para comprobar la eficiencia de la aplicación, se hizo una selección en GenBank (*http://www.ncbi.nlm.nih.gov/Genbank/*) de 10 secuencias con tablas de traducción diferentes. La siguiente tabla describe brevemente cada una de estas secuencias:

Tabla 1. Secuencias seleccionadas para análisis con ProSA

Secuencia	Número(s) de Acceso	Longitud (nucleótidos)	Tabla de Traducción
Virus de inmunodeficiencia humana 1, genoma completo	NC_001802	9181	Standard (1)
DNA mitocondrial de *Chelonia mydas*, secuencia completa	NC_000886	16497	Vertebrate Mitochondrial (2)
Mitocondria de *Saccharomyces cerevisiae*, genoma completo	NC_001224	85779	Yeast Mitochondrial (3)
Mitocondria de *Acanthamoeba castellanii*, genoma completo	U12386	41591	Mold Mitochondrial, Protozoan Mitochondrial, Coelenterate Mitochondrial, Mycoplasma, Spiroplasma (4)
Mitocondria de *Drosophila melanogaster*, genoma completo	NC_001709	19517	Invertebrate Mitochondrial (5)
mRNA para hemoglobina de *Tetrahymena pyriformis*	D13920	587	Ciliate Nuclear, Dasycladacean Nuclear, Hexamita Nuclear (6)
mRNA macronuclear para protein-cinasa nuclear putativa de *Euplotes octocarinatus* (gen npk 1)	AJ249683	1322	Euplotid Nuclear (10)
Gen potenciador de la infectividad a macrófagos (mip) de *Legionella lytica* cepa LLAP-9, cds parciales	AF148986	598	Bacterial and Plant Plastid (11)
Mitocondria de *Scenedesmus obliquus*, genoma completo	X17375 AJ271733 AJ272528 AJ277429 AJ400708	42781	*Scenedesmus obliquus* mitochondrial (22)
DNA mitocondrial de *Thraustochytrium aureum*, genoma parcial	AF288091	31570	*Thraustochytrium* mitochondrial code (23)

Estas secuencias fueron almacenadas en archivos con formato GenBank. Para el análisis de cada secuencia, se copiaron las líneas correspondientes a los nucleótidos y se introdujeron en la interfaz de Web, seleccionando el tipo de secuencia y tabla de traducción correspondiente. Se utilizaron como clientes los navegadores Web: Mozilla, Konqueror e Internet Explorer para evaluar el formato de la interfaz y sus resultados.

5. Resultados

5.1 Implementación de ProSA

El resultado de la implementación consistió en una serie de archivos jerarquizados de acuerdo a su función dentro de la aplicación. La siguiente figura resume el sistema de archivos obtenido:

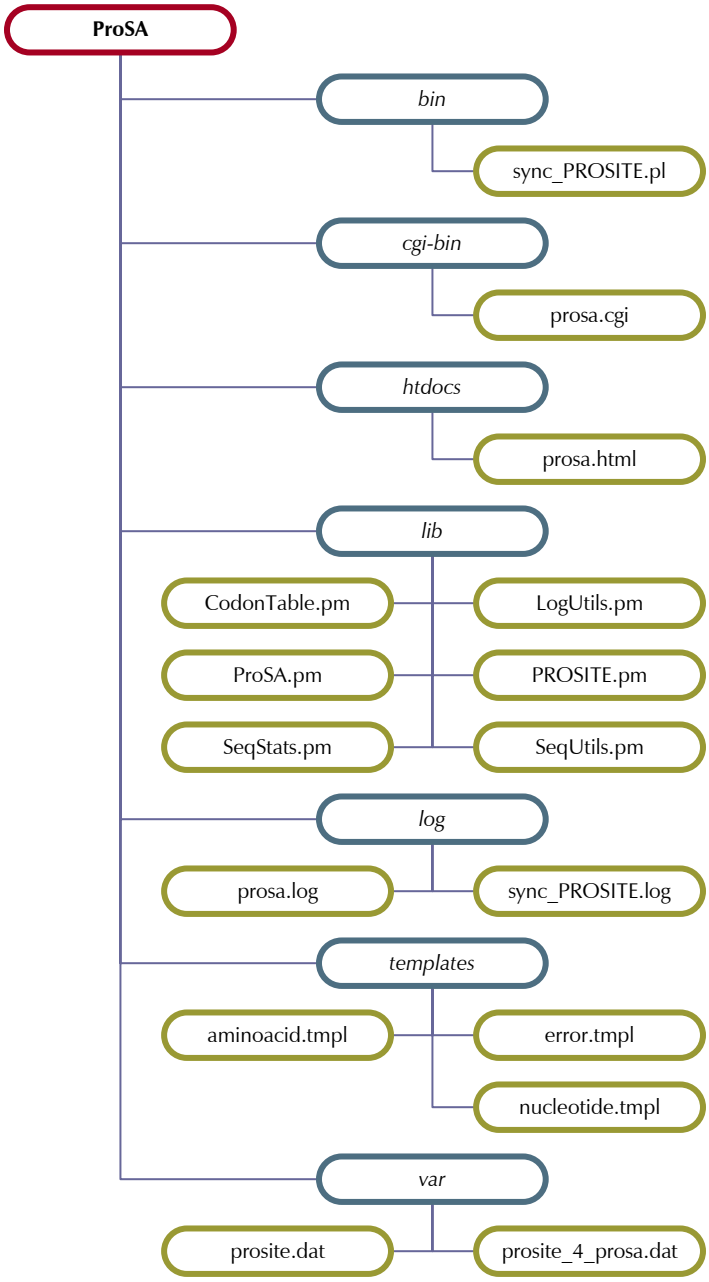

Figura 2. Jerarquía de archivos de ProSA

Este sistema de archivos se compone de directorios organizados de manera similar a la convención de los sistemas Unix. Esta elección es comúnmente utilizada para el desarrollo de aplicaciones, lo cual permite organizar los archivos de manera sencilla y eficiente. La siguiente tabla describe el contenido de los directorios en la presente implementación:

Tabla 2. Contenido de los directorios de ProSA

Directorio	Contenido
bin	Script para actualizar la base de datos PROSITE
cgi-bin	Aplicación CGI que realiza el análisis
htdocs	Interfaz de Web para el usuario
lib	Módulos para el script de actualización y la aplicación CGI
log	Bitácoras de actividad
templates	Plantillas HTML para el despliegue de los resultados de la aplicación CGI
var	Archivos de la base de datos PROSITE

Para lograr algunas partes de la implementación, se recurrió a los siguientes módulos de Perl existentes: `CGI`, `Date::Calc`, `HTML::Template` y `LWP::Simple`. El primero forma parte de la distribución estándar de Perl, mientras que los demás fueron instalados de CPAN (*http://www.cpan.org/*). Todos los módulos contenidos en el directorio `lib` fueron elaborados por el autor para satisfacer las necesidades particulares de la implementación.

5.2 Ejemplo de utilización de ProSA

A continuación se describirá paso a paso y con ayuda de figuras la realización de un análisis con ProSA.

5.2.1 Introducción de una secuencia y selección de parámetros

- Introducir la secuencia que se desea analizar en la caja de texto más grande que aparece en el formulario HTML (**Paso 1**).
- Seleccionar un tipo de secuencia para el análisis (**Paso 2**).
- Seleccionar un Código Genético para la traducción en caso de que la secuencia sea de nucleótidos (**Paso 3**).
- Hacer clic en el botón *"Analizar"* (**Paso 4**).

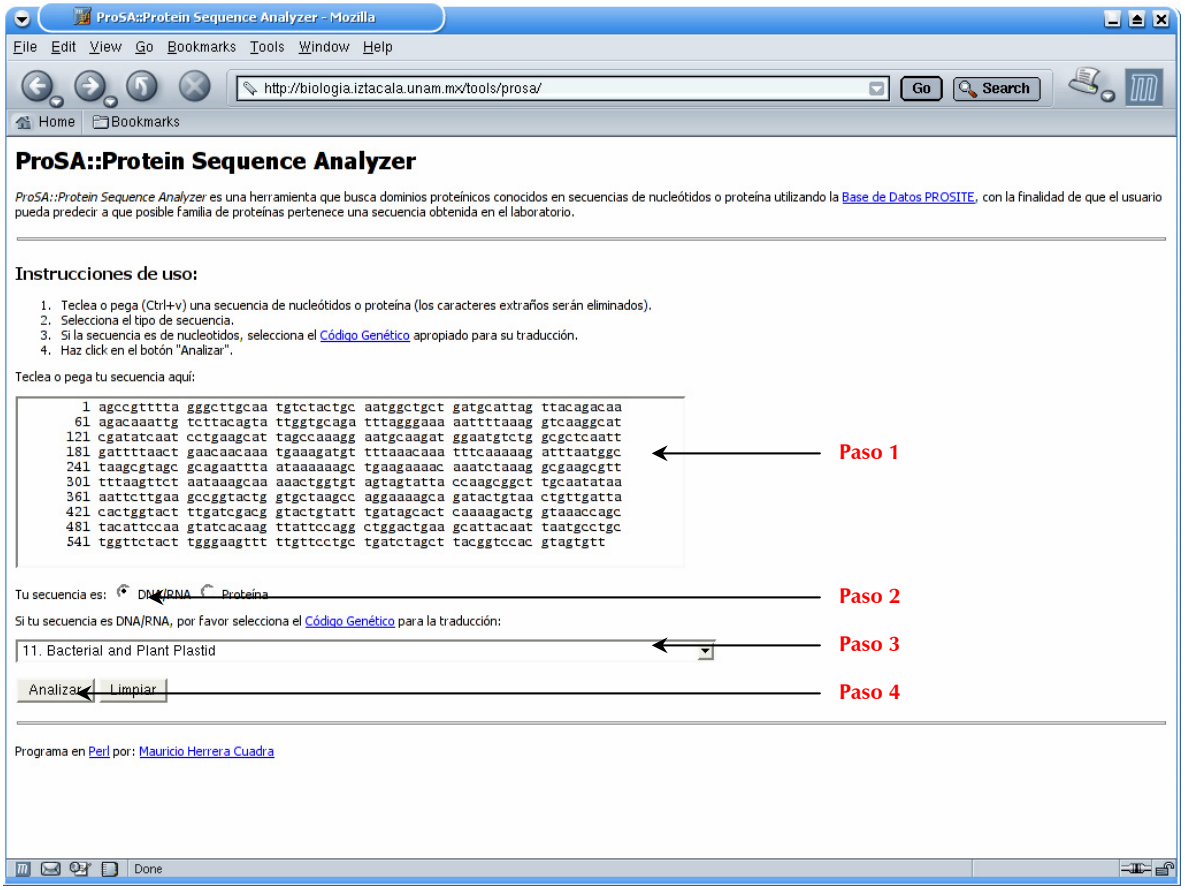

Figura 3. Ejemplo de llenado del formulario de ProSA

5.2.2 Despliegue de los resultados del análisis

En la siguiente figura se muestran los resultados del análisis de la secuencia introducida. Se pueden observar datos como: secuencia introducida, duración del análisis, secuencia de mRNA utilizada para la traducción, número de nucleótidos, peso molecular, número y porcentaje de Adeninas, Uracilos, Guaninas, Citosinas, porcentaje de Adenina-Uracilo y Guanina-Citosina, Código Genético utilizado para la traducción, secuencia del 1er marco de lectura, número de subsecuencias, longitud y peso molecular de cada una de ellas, y los datos más importantes obtenidos con esta aplicación: los patrones de PROSITE que se localizaron en cada subsecuencia, así como su posición, fragmento concordante, patrón de búsqueda y expresión regular equivalente.

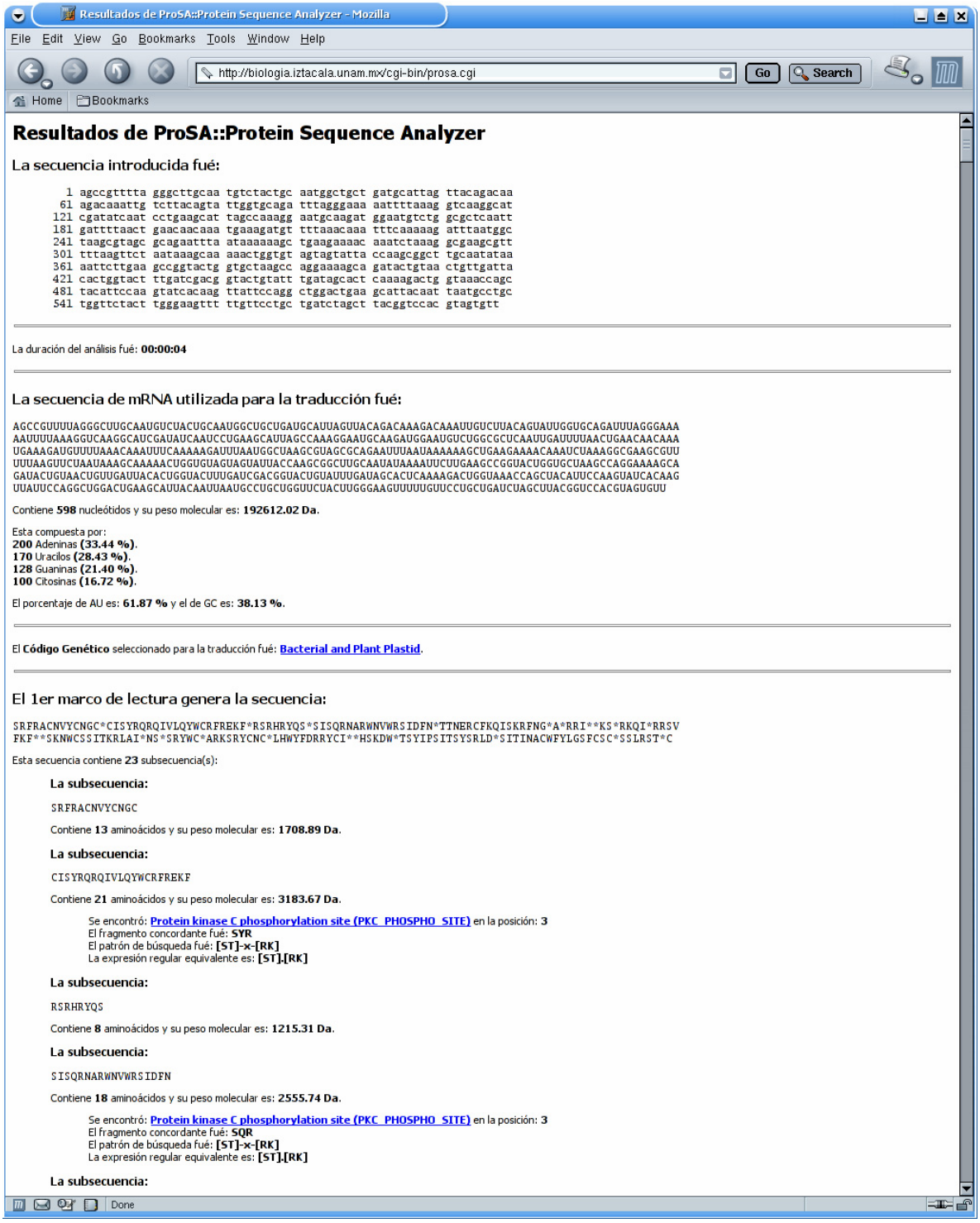

Figura 4. Página con los resultados del análisis realizado por ProSA

Nuevamente se aprovecharon las ventajas del lenguaje HTML para incluir vínculos en la página de resultados del análisis. El primero es hacia el sitio del NCBI donde se encuentra la información referente a los Códigos Genéticos. Para cada análisis de secuencias de nucleótidos éste vínculo se estará apuntando hacia la sección correspondiente al Código Genético seleccionado por el usuario. Los

vínculos restantes son hacia la documentación correspondiente a cada patrón de PROSITE encontrado en las subsecuencias de proteína.

5.3 Análisis con ProSA

Una vez realizado el análisis de las secuencias seleccionadas en GenBank, se procedió a la interpretación de los resultados obtenidos por la aplicación CGI. Los resultados completos fueron almacenados en archivos HTML. Se utilizó como nombre para cada archivo el número de acceso correspondiente en GenBank para cada secuencia.

6. Discusión

El análisis de secuencias es una de las metodologías más utilizadas en bioinformática. El desarrollo de herramientas computacionales adecuadas y eficientes permite auxiliar a los biólogos en la búsqueda del significado y función de las secuencias contenidas en los genes. El poder realizar estos análisis a través de Internet permite el acercamiento de todo tipo de usuarios, los cuales no requieren de conocimientos en programación para utilizar las aplicaciones.

ProSA es una aplicación con un diseño adecuado, pues éste permite un fácil mantenimiento y administración de la aplicación, además de que su instalación y configuración no es complicada. Otra de las principales ventajas de su diseño es que se pueden reutilizar todos sus módulos y funciones. Perl es un lenguaje que fomenta estas prácticas a través de su esquema de programación modular. La implementación puede utilizarse como base para el desarrollo de futuras aplicaciones, o extenderse para realizar búsquedas utilizando otras bases de datos u otro tipo de secuencias.

El resultado de esta implementación está orientado a computadoras con características similares (Linux, *BSD, Mac OS X y demás tipos de Unix), mas no por eso se encuentra limitado para su utilización en otras plataformas (Mac OS < 9, Windows) después de una configuración apropiada.

La utilización de Perl como lenguaje para desarrollar la implementación fue eficiente, esto se debió a que Perl es un lenguaje de programación suficientemente maduro, que cuenta con un gran soporte para la búsqueda de patrones y el desarrollo de aplicaciones Web. La gran cantidad de módulos disponibles a través de CPAN permite la rápida implementación de casi cualquier diseño de software.

A pesar de la existencia del proyecto BioPerl, no se utilizó para esta implementación ninguno de los módulos ahí existentes para la manipulación de secuencias, esto se debió a que -desde el punto de vista del autor- el proyecto BioPerl no es aún lo suficientemente maduro como para elaborar aplicaciones que se utilizarán dentro de un ambiente de producción (como es denominado en informática).

BioPerl funciona bien para aplicaciones desarrolladas dentro de un laboratorio, donde los usuarios son los mismos investigadores que elaboran herramientas para simplificar tareas diarias, mas no para usuarios finales que probablemente no puedan corregir los errores en las dependencias que ocasiona el continuo desarrollo y evolución de los módulos del proyecto.

Tomando en cuenta que con la tecnología de secuenciación actual solo es posible obtener secuencias de hasta 500 nucleótidos por experimento, y que los tiempos obtenidos para los análisis con ProSA fueron bastante cortos (2 seg. para 587 nucleótidos), podemos decir que la aplicación es bastante eficiente para analizar secuencias obtenidas mediante estos experimentos.

Esto no significa que ProSA esté limitado para analizar secuencias mayores, por el contrario, la evidencia de que tal análisis es posible se presenta en los resultados aquí descritos (15:04 para 85779 nucleótidos).

El único inconveniente será que para obtener una secuencia de tal longitud se necesitarán decenas de experimentos de secuenciación, además de un largo proceso previo de ensamblaje de la secuencia.

En el caso de que se contara con dicha secuencia, el análisis sería demasiado lento para llevarse a cabo por Internet. En el presente trabajo se realizaron modificaciones en la configuración de Apache, ajustando la variable `Timeout` a 1200 segundos (20 minutos), con lo cual el análisis de las secuencias grandes pudo llevarse a cabo. Esto se logró a costa de que la aplicación consumió una gran cantidad de memoria, debido al inmenso volumen de datos que fue generado, además de que el procesador era ocupado en un alto porcentaje por el proceso correspondiente. Esta aproximación no es la más adecuada, ya que en el caso de una máquina con pocos recursos (poca memoria y/o procesador lento), estos dos factores podrían comprometer su rendimiento, sobre todo si ésta funciona como servidor de Web.

Una de las soluciones más adecuadas sería modificar el diseño de la aplicación orientándolo hacia un modelo de cómputo distribuido, en el que una computadora central se encargue de recibir las secuencias y se repartan fragmentos a otras maquinas para que realicen partes del análisis y devuelvan los resultados para su ensamble y despliegue al usuario. Este diseño podría reducir considerablemente el tiempo para la realización del análisis de una secuencia grande, además de que no agotaría los recursos de la máquina que funciona como servidor puesto que el trabajo estaría repartido.

ProSA es una aplicación sencilla para los usuarios nuevos y al mismo tiempo poderosa para los usuarios avanzados. La interfaz de Web resulta bastante fácil de utilizar, ya que cuenta con descripciones sobre la aplicación y su uso, así como vínculos a sitios que proporcionan información relacionada con los datos que se utilizan para el análisis.

El despliegue de los resultados se encuentra en un formato atractivo y legible para su interpretación. Una posible mejora sobre su aspecto podría ser el resaltar directamente en las secuencias de proteína los patrones de PROSITE encontrados. Se podrían consensar códigos de color para patrones que presenten actividad biológica semejante, tal como hacen los programas de visualización molecular, por ejemplo: *RasMol*.

Los resultados proporcionados por la aplicación son plenamente confiables, puesto que todos los patrones de PROSITE encontrados en los análisis coincidieron con la actividad correspondiente a los genes publicados en GenBank.

Ocurrieron algunas discrepancias en cuanto a los marcos de lectura donde se encontraron muchos de los patrones de actividad biológica significativa, las cuales son justificables debido a que la mayoría de las secuencias corresponden a organismos eucariotes. Es importante recordar que la estructura de los genes eucariotes es más complicada que la de los genes procariotes. Al contrario de los genes procariotes, los genes eucariotes se encuentran a menudo fragmentados en piezas que son ensambladas antes de la traducción. En eucariotes, el mRNA es procesado antes de ser traducido. Existen dos tipos de procesamiento: el corte y la poli-adenilación. El corte une las secuencias codificantes y elimina los elementos intermedios. Las secuencias que terminan dentro del mRNA maduro son llamadas exones, y las intermedias son llamadas intrones. En la poli-adenilación se añaden 50 o más nucleótidos de adenina al final del mRNA.

ProSA es una aplicación que traduce las secuencias de nucleótidos tal y como son proporcionadas por el usuario. El único pre-procesamiento que lleva a cabo es el de eliminar los caracteres que no pertenezcan al tipo de secuencia seleccionado para el análisis, pero carece de métodos para la distinción de intrones y exones, por lo que la detección de secuencias no codificantes dependerá total y absolutamente de la habilidad y experiencia del usuario para la interpretación de los resultados.

Sin embargo, con la ayuda de ProSA será relativamente sencillo encontrar genes en genomas procariotes. Esto se debe a que los genes procariotes son un poco más simples. Estos contienen un promotor que determina cuando un gen debe ser transcrito y una región codificante que contiene la secuencia de DNA para una proteína. En estos genomas será muy raro encontrar marcos de lectura abiertos (ORF) largos, esto se debe a que es mas probable encontrar codones STOP cada 21 tripletes aproximadamente (ya que existen 3 codones STOP de un total de 64 combinaciones de tripletes). Por ejemplo, será realmente poco probable encontrar un ORF que tenga una longitud de 900 nucleótidos (en promedio, las proteínas poseen una longitud de 300 aminoácidos); aunque si sucediera, sería una clara señal de que el ORF codifica para una proteína real. Por supuesto, algunos genes codifican para proteínas pequeñas, y encontrar éstos será un poco más difícil.

Por tales motivos, la interpretación de resultados de los análisis de secuencias pertenecientes a organismos eucariotes puede resultar un poco más compleja, sin embargo no es imposible, clara evidencia de esto es el presente trabajo.

La aplicación final se instaló como una herramienta en el Sitio Web de la Carrera de Biología (*http://biologia.iztacala.unam.mx/tools/prosa/*). Si se desea obtener una copia gratuita del software para instalación y uso local en otros servidores será necesario contactar al autor.

En conclusión, ProSA resulta una herramienta innovadora, puesto que no existía una aplicación que realizara búsquedas en la base de datos PROSITE a partir de secuencias de nucleótidos. Con el desarrollo de esta aplicación se ha podido resolver este problema, son los usuarios finales los que deberán juzgar los beneficios aquí planteados.