Avances en Proteómica

Davor Pavisic - Gonzalo Argote - Soraya Ordoñez - Oscar Antezana - Ariel Cary
Erick Antezana - Reynaldo Vargas

Instituto de Investigación en Informática Aplicada (IIIA)
Universidad Católica Boliviana, Regional Cochabamba.
E-mail: dp@ ucbcba.edu.bo

Resumen

Este artículo intenta brindar una introducción a la bioinformática, sus aplicaciones en la biología molecular y sus implicaciones en la medicina. Para esto, primeramente se define brevemente a los genes y su función en el proceso de formación de proteínas; estos dos términos ayudan a definir a la proteómica como el estudio de las proteínas expresadas por el genoma. El éxito de la proteómica depende de herramientas que permitan determinar la secuencia de aminoácidos que componen a una proteína. En este artículo se describen algunas técnicas bioinformáticas que actualmente se utilizan para la determinación de estas secuencias. La determinación de secuencias permite establecer una relación entre las proteínas y los genes que las codifican, conocimiento que puede servir para muchos propósitos diferentes. Entre los más importantes es establecer el papel que juegan los genes durante la enfermedad.

1. Introducción

La publicación de la teoría de Darwin sobre la evolución de las especies en 1859 y los postulados de Mendel sobre la herencia en 1866 se convirtieron en los primeros pasos dados en el área de la genética. El descubrimiento de los genes como el elemento básico para la construcción de seres vivos convierte a la biología, y más particularmenete a la biología molecular, en uno de los focos de atención más importantes para gran parte de los centros de investigación, tanto públicos como privados.

En Mayo de este año, el *Human Genome Project* y la empresa Celera Genomics, anunciaron la finalización del primer borrador del genoma humano: aproximadamente 3 gigabytes de pares de bases A, C, T y G que podrían, fácilmente, llenar más de 2000 diskettes de computadora. Pero este evento es tan solo el umbral del portal que se abre ante un universo completamente nuevo y vasto. Actualmente los científicos del área están generando bases de datos gigantescas que contienen detalles sobre dónde y en qué tejidos del cuerpo diferentes genes se "activan", las formas de las proteínas que los genes codifican, cómo iteractuan las proteínas unas con otras y el papel que juegan durante la enfermedad. La nueva disciplina de la

bioinformática, una fusión entre las ciencias de la computación y la biología, intenta darle sentido a esta inmensa cantidad de información que, día a día aumenta exponencialmente.

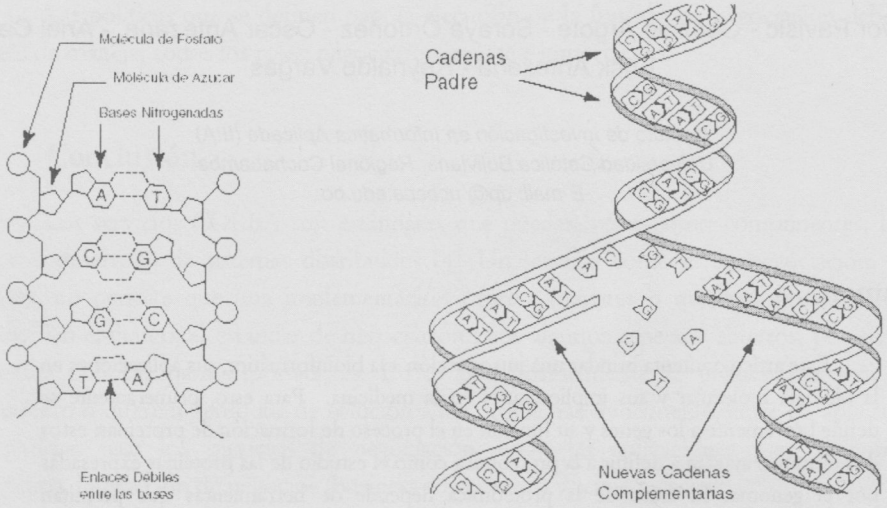

Figura A Figura B

Figura 1. (A) Las cuatro bases nitrogenadas de ADN. La adenina (A) se asocia con la timina (T) mientras que la citosina (C) se asocia con la guanina (G). Las cadenas se mantienen unidas por puentes de hidrógeno. La secuencia de bases en un gen contiene la información necesaria para la síntesis de proteínas. (B) Replicación: cada hilo simple se convierte en una plantilla para la síntesis de un nuevo hilo complementario y cada molécula hija forma una copia exacta de la molécula padre.

El juego completo de instrucciones para replicar un organismo se encuentra en el *genoma*. El genoma contiene el plano maestro para todas las estructuras celulares y actividades que desarrollará un organismo o célula durante su vida. Dentro del núcleo de cada una de las millones de células humanas se encuentra el genoma humano. Éste consiste de moléculas de ADN enroscadas, muy compactas y de proteínas asociadas que se organizan en estructuras llamadas *cromosomas*. Para cada organismo, desde simples bacterias hasta organismos sorprendentemente complejos como los seres humanos, los componentes de estos delgados hilos de ADN codifican toda la información necesaria para crear y mantener la vida. Para comprender cómo el ADN desempeña su función, es necesario un cierto conocimiento de su estructura y organización.

1.1. El ADN

Una molécula de ADN consiste de dos cadenas en forma de doble hélice entrelazada de gran longitud. Cada cadena es un arreglo lineal de cuatro unidades similares que se repiten (*nucleótidos*). Cada nucleótido está compuesto de un azúcar, un fosfato y una base nitrogenada. Las cadenas de ADN se mantienen juntas por medio de enlaces débiles de hidrógeno entre sus bases. Cuatro tipos de bases diferentes se encuentran en el ADN: *adenina* (A), *timina* (T), *guanina* (G) y *citosina* (C) (Figura 1.A). El orden particular del arreglo de bases a lo largo de esta columna es llamado secuencia de ADN. La secuencia especifica con exactitud las instrucciones genéticas que se requieren para crear y mantener con vida a un organismo determinado con sus rasgos únicos y particulares. El tamaño del genoma se mide generalmente por el número de pares de bases; el genoma humano contiene aproximadamente tres mil millones de pares [1,2].

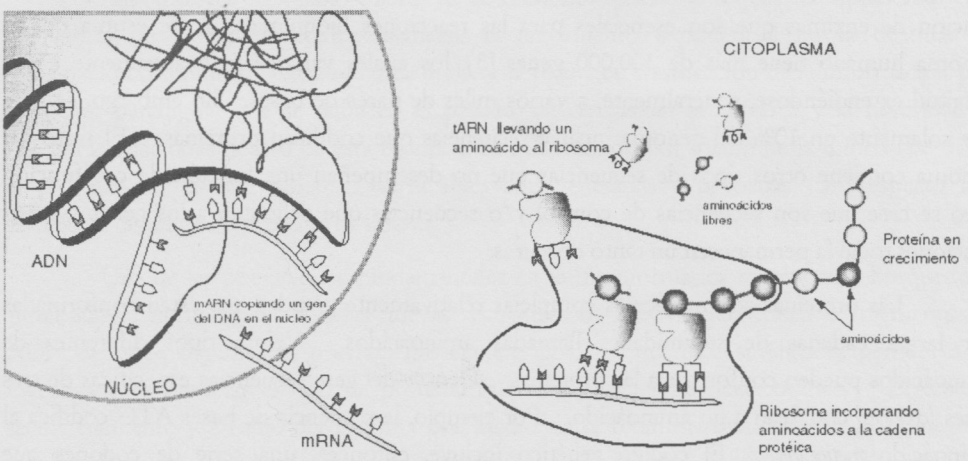

Figura 2. Cuando los genes son expresados, la información genética (secuencia base) del ADN es copiada a una molécula mensajera mARN. Luego, las moléculas de mARN abandonan el núcleo de la célula y entran al citoplasma donde tripletas de bases (codones) especifican qué aminoácidos particulares se deben enlazar para fabricar una proteína indiviudal. Este proceso se lleva a cabo en los ribosomas que leen el código genético del mARN y, tomando los aminoácidos de los tARNs, los adjuntan a la proteína que se está formando.

Cada vez que una célula se divide en dos células hijas, el genoma completo debe ser duplicado. Para los humanos y otros organismos complejos, esta duplicación ocurre en el núcleo durante la división celular. En este proceso, la molécula de ADN se desenrrolla y los enlaces débiles entre los pares de bases se rompen, permitiendo que las dos cadenas que la

componen se separen. Cada cadena dirige la síntesis de una nueva cadena complementaria donde los nucleótidos libres se unen con sus bases complementarias. Las estrictas reglas de formación de pares de bases indican que la adenina se juntará solamente con la timina formando un par A-T; y la citosina con la guanina formando un par C-G. Por lo tanto, cada célula hija recibe una cadena nueva y una cadena vieja de ADN (Figura 1.B). Este mecanismo de reproducción minimiza la incidencia de errores (mutaciones) que pueden afectar fuertemente al organismo resultante y a sus descendientes.

1.2. Genes y Proteínas

Cada molécula de ADN contiene muchos miles de genes: las unidades básicas tanto físicas como funcionales de la herencia. Un gen es una secuencia específica de bases de nucleótidos que contiene la información requerida para construir una proteína. Las proteínas, a su vez, proveen los componentes estructurales de las células y los tejidos o desempeñan la función de enzimas que son esenciales para las reacciones bioquímicas. Se estima que el genoma humano tiene más de 100,000 genes [3], los cuales varían significativamente en su longitud extendiéndose, generalmente, a varios miles de pares de bases. Sin embargo, se sabe que solamente un 10% del genoma incluye secuencias que codifican proteínas. El resto del genoma contiene otros tipos de secuencias que no desempeñan una función de codificación pero se cree que son secuencias de control y/o secuencias que delimitan a los genes y cuyas funciones todavía permanecen un tanto obscuras.

Las proteínas son moléculas complejas relativamente grandes que están conformadas por largas cadenas de subunidades llamadas aminoácidos. Veinte tipos diferentes de aminoácidos pueden conformar a las proteínas. Dentro del gen, secuencias específicas de tres bases (*codones)* codifican a un aminoácido. Por ejemplo, la secuencia de bases ATG codifica al aminoácido *metionina.* El código genético incluye, entonces, una serie de codones que especifican qué aminoácidos son necesarios para formar una proteína específica.

Las instrucciones codificadas para la generación de proteínas que se encuentran en un gen son transmitidas indirectamente vía un ácido ribonucleico mensajero (mARN), una molécula intermediaria de ácido ribonucleíco en forma de cadena simple. Para que la información de un solo gen pueda ser expresada, una cadena complementaria de mARN es generada a partir de la plantilla de ADN original en el núcleo; este proceso es llamado *transcripción.* El mARN generado sale del núcleo al citoplasma celular donde, a su vez, sirve como plantilla para la síntesis de proteínas. Esta síntesis se lleva a cabo en los ribosomas donde cada codón es *traducido* a un aminoácido. Las proteínas van generandose a medida que los aminoácidos son adheridos a la cadena proteica (ver Figura 2). La proteína, generalmente, sufre modificaciones estructurales durante su traducción o una vez que es liberada del ribosoma. Existen literalmente docenas de diferentes tipos de modificaciones que puede sufrir una proteína; cada una de estas puede influir en las características y funciones de la proteína lo que incluye un factor adicional de estudio a la secuencia de la proteína: la estructura.

2. La Bioinformática

La Bioinformática es el área de la ciencia en la cual la biología molecular, las ciencias de la computación y las tecnologías de información se juntan en una sola disciplina. Los objetivos principales del área son crear y madurar nuevas perspectivas globales, a partir de las cuales principios biológicos unificadores pueden ser discernidos.

Existen tres áreas fundamentales en la bioinformática: el desarrollo de nuevos algoritmos y técnicas estadísticas con las que se puedan encontrar y cuantificar relaciones entre elementos de bases de datos muy grandes; el análisis y la interpretación de varios tipos de datos incluyendo secuencias de nucleótidos y aminoácidos, dominios de proteínas y estructuras proteicas; y, finalmente, el desarrollo e implementación de herramientas que permitan administrar y acceder de forma eficiente a diferentes tipos de información.

Los motivos por los cuales es interesante aplicar enfoques computacionales a problemas de la bioinformática incluyen: el crecimiento exponencial en la cantidad de información disponible; la necesidad de unificar la forma de clasificación de la información para evitar un paradigma que se daba en el pasado: *un científico-un gen/proteína* y la necesidad de encontrar relaciones dentro de esta montaña de información por medio de la minería de datos, el proceso por el cual hipótesis verificables son generadas y probadas con respecto a la estructura de un gen o proteína de interés.

Una de las operaciones fundamentales en la bioinformática involucra la búsqueda de similitudes u homologías entre las nuevas secuencias de ADN y secuencias o fragmentos de ADN de otros organismos previamente encontrados. Encontrar similitudes permite a los investigadores predecir el tipo de proteína que una nueva secuencia codifica. El área de secuenciamiento y estudio del genoma por medio de la bioinformática adoptó el nombre de *genómica* mientras que el estudio de las proteínas expresadas por el gen adoptó el nombre de *proteómica*. La genómica, que estudia la decodificación de las reglas que el genoma de cada organismo contiene, permite determinar cómo los aminoácidos serán encadenados para formar proteínas. Esta codificación se extrae directamente de los cromosomas de las células del individuo que se desea estudiar y es una característica prácticamente fija. Sin embargo, esta determinación no dice mucho sobre la estructura final o función de las proteínas; una vez que una proteína sale de la "línea ensambladora" de gen-a-proteína, ésta es alterada a medida que toma su forma final: carbohidratos, fosfatos, sulfatos y otros residuos se adhieren a ella dándole su forma final y alterando su masa y función. El proteoma (las proteínas expresadas por el genoma), a diferencia del genoma, no es un conjunto de características estáticas. Al contrario, el proteoma cambia dependiendo del estado de desarrollo del organismo, del tejido al que pertenece o incluso de acuerdo a las condiciones del medio en las que se encuentra el organismo [4].

2.1. Proteómica

La proteómica trata de subsanar el vacío de conocimiento que existe entre el ADN y sus productos finales. El objetivo final de la proteómica es la deducción de la estructura y las interacciones de todas las proteínas en una célula dada. La comparación de mapas proteicos de células sanas con los mapas de células enfermas podría permitir a los investigadores entender los cambios que ocurren en las señales que generan las células y los procesos metabólicos que se llevan a cabo. Por ejemplo, las empresas farmacéuticas podrían diseñar nuevas pruebas de diagnóstico e identificar nuevos y mejores receptores para nuevos fármacos.

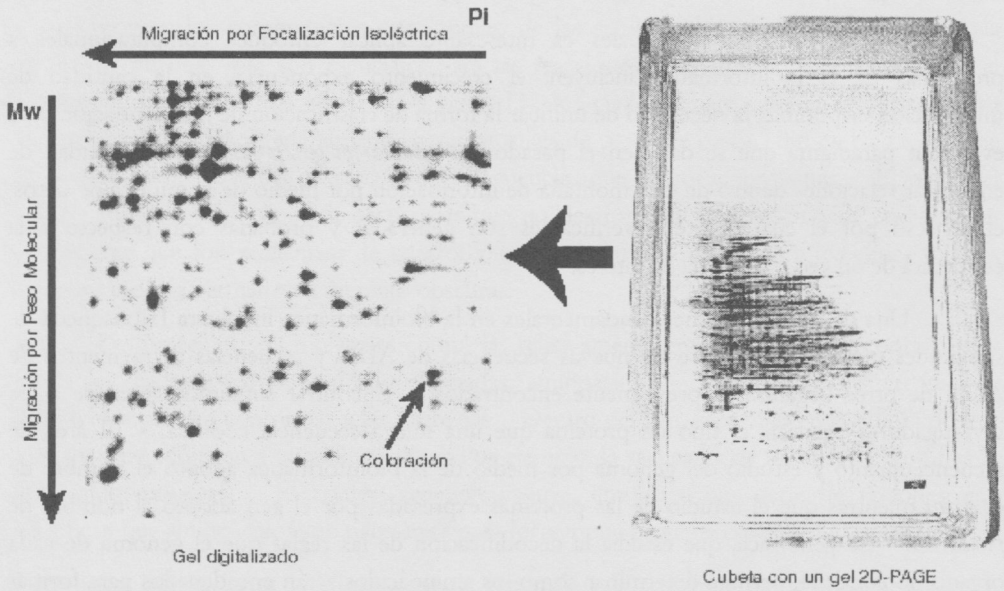

Figura 3. Proceso 2D-PAGE. En esta figura se observa un gel con el mapa proteico obtenido mediante esta técnica. Las proteínas en el gel han sido separadas en dos dimensiones: punto isoeléctrico y peso molecular. Luego se realizó un proceso de coloración para poder resaltar el mapa proteico resultante.

Los biólogos moleculares han estado tratando de desentrañar la configuración proteica de las células por décadas; sin embargo, del mismo modo que con genómica, que trata de identificar a los genes, el éxito de proteómica depende de la habilidad para desarrollar técnicas que puedan identificar rápidamente el tipo, la cantidad y las actividades de miles de proteínas en una célula. Esta información puede fácilmente revolucionar el campo de la medicina. Por

ejemplo, analizando una muestra del tejido de un paciente, sería posible detectar variaciones en las concentraciones normales de ciertas proteínas y diagnosticar enfermedades mucho antes que los síntomas se presenten. Small Molecule Therapeutics, una empresa de New Jersey (USA), ha desarrollado una técnica para comprender la función de una proteína. Su técnica primero encuentra dos proteínas que iteractúan entre sí, luego crea fragmentos de una de las proteínas. Algunos de estos fragmentos pueden bloquear toda iteracción futura con la proteína intacta. Los científicos pueden deducir cómo se alteran las funciones de una célula a causa de esta inhibición. La compañía ha utilizado esta técnica para detectar inhibidores para la proteína RAS que puede producir cáncer [5].

3. Identificación de Proteínas

En los proyectos de proteoma, los cuales apuntan a identificar y caracterizar todas las proteínas expresadas por un organismo o tejido, la identificación de proteínas es una actividad central [3]. La identificación hace que una proteína sea descrita de acuerdo a la secuencia de aminoácidos que la conforman. Es posible, de este modo, establecer una relación entre esas secuencias de aminoácidos con los genes que las codifican y, por lo tanto, relacionar los genomas con los proteomas. Una vez establecida la secuencia (o parte de la secuencia), la identificación puede ser realizada comparando la secuencia obtenida con bases de datos enormes que contienen proteínas y sus secuencias [6]. La identificación es también el primer paso para los estudios de las modificaciones co y post-traducción que sufren las proteínas. Adicionalmente, por medio de la identificación es posible llegar a deducir la función de las proteínas dentro el organismo o tejido.

3.1. El Problema de la Idenficación de Proteínas

Antes del proceso de identificación, es necesario desnaturalizar la proteína (Figura 3). La desnaturalización se encarga de romper todos aquellos enlaces no-covalentes, así como enlaces entre sus aminoácidos y cualquier grupo fosfático o con el entorno (enlaces débiles). Estos enlaces son los que se encargan de dar su forma tridimensional a la proteína. Al romperlos, se modifica la estructura de la proteína de tal manera que su nueva forma sea la de una cadena lineal. La desnaturalización es necesaria para poder separar todas las proteínas de una muestra.

Actualmente, la mejor forma de separar (y visualizar) las proteínas de una muestra es a través del uso de la técnica de Electroforesis Bidimensional en geles de polyacrilamida (Two-Dimensional Polyacrilamide Gel Electrophoresis o 2D-PAGE). 2D-PAGE posibilita generar, a partir de una muestra de un fluído o de un tejido, un mapa de proteínas ordenadas espacialmente en dos dimensiones. Completada su separación mediante la técnica 2D-PAGE, cada una de las proteínas presentes en la muestra puede ser aislada físicamente. Luego, es posible proceder a un análisis detallado para poder a) identificar directamente a la proteína de

acuerdo a sus atributos en el gel o, b) determinar la secuencia de aminoácidos que la compone para luego identificar a la proteína.

3.2. 2D-PAGE

Una vez depositada una muestra biológica en el gel, se genera el mapa proteico en dos etapas que corresponden a cada una de las dos dimensiones mencionadas: En la primer etapa, la focalización isoeléctrica (IEF), las proteínas son separadas en un gradiente de pH hasta que alcanzan una posición estacionaria donde su carga eléctrica neta es cero (punto isoeléctrio, *pI*)[1]. Las proteínas separadas por IEF son separadas luego ortogonalmente por electroforesis en la presencia de sulfato dodecil de sodio (SDS-PAGE). En la segunda etapa, las proteínas ligadas a SDS son separadas en el gel de poliacrilamida de acuerdo a su peso molecular, *Mw*, aplicando una diferencia de potencial perpendicular al gradiente de pH. El peso molecular de una proteína se ve afectado por la masa total de la misma. Adicionalmente, se efectúa un proceso de coloración del gel para poner en evidencia a las proteínas. La técnica más utilizada para la coloración es la de *silver staining* [7]. El resultado de este proceso es un gel en el que las proteínas que se encuentran en la muestra están separadas bidimensionalmente de acuerdo a su punto isoeléctrico y su peso molecular. Cada proteína es, por lo tanto, un punto (generalmente referido como *spot*) en el gel con varios atributos asociados (ver Figura 3).

La introducción de los nuevos geles IPG (Immbolised pH gradient) eliminaron varios problemas de inestibilidad de gradiente y de pobre capacidad de carga de los geles. Éstos están disponibles comercialmente en una amplia variedad de intervalos de pH, por lo que se pueden hacer estudios específicos de proteínas en rangos determinados de pH. Esta disponibilidad comercial hace que distintos geles obtenidos por electroforesis puedan ser intercambiados y los resultados comparados entre laboratorios. También, con la ayuda de software especializado, mapas producidos por diferentes laboratorios pueden ser directamente comparados. Esto ha llevado a que varios grupos de investigación hagan disponibles sus mapas de proteínas en Internet, como es el caso de SWISS-2DPAGE disponible en el servidor ExPASy en Ginebra Suiza [6].

3.3. Identificación de proteínas utilizando su Punto Isoeléctrico y Peso Molecular

Una forma de identificar a las proteínas de una imagen 2D PAGE es mediante el uso de un mapa 2D PAGE de referencia o gel maestro (*master gel*) [8]. El *master gel* es una imagen 2D PAGE representativa de un conjunto de geles y producida a partir de una muestra biológica específica. Alternativamente, el master gel puede ser producido sintéticamente mediante la

[1]El punto isoeléctrico de una proteína desnaturalizada es un parámetro que está determinado por la composición de aminoácidos, las terminaciones N y C de los aminoácidos, y cualquier modificación post-traducción que haya sufrido la proteína.

unión ponderada de un grupo de geles. En estos mapas, cada *spot*, correspondiente a una proteína, está vinculado por un número de acceso (AC) a una bases de datos proteica y posee un vínculo a información textual relacionada a las proteínas del mapa 2D PAGE.

Dada una imagen 2D PAGE sin procesar, la identificación de proteínas se realiza de la siguiente manera. Primero, se realiza un análisis de la imagen para detectar y cuantificar los *spots* del mapa. Segundo, se realiza una comparación con un *master gel*, emparejando los *spots* detectados con los correspondientes del *master gel*. Para realizar el análisis y emparejamiento de geles, existen sistemas informáticos comerciales como, por ejemplo, MELANIE-II que fue parcialmente desarrollado en el I.I.I.A.[2] [9]. Si para un spot dado, existe su correspondiente par en el master gel, éste puede ser fácilmente asociado a una proteína conocida. Finalmente, el nombre de la proteína e información adicional puede ser obtenida consultando a los servidores de bases de datos de proteínas con el número de acceso del *master gel* [8].

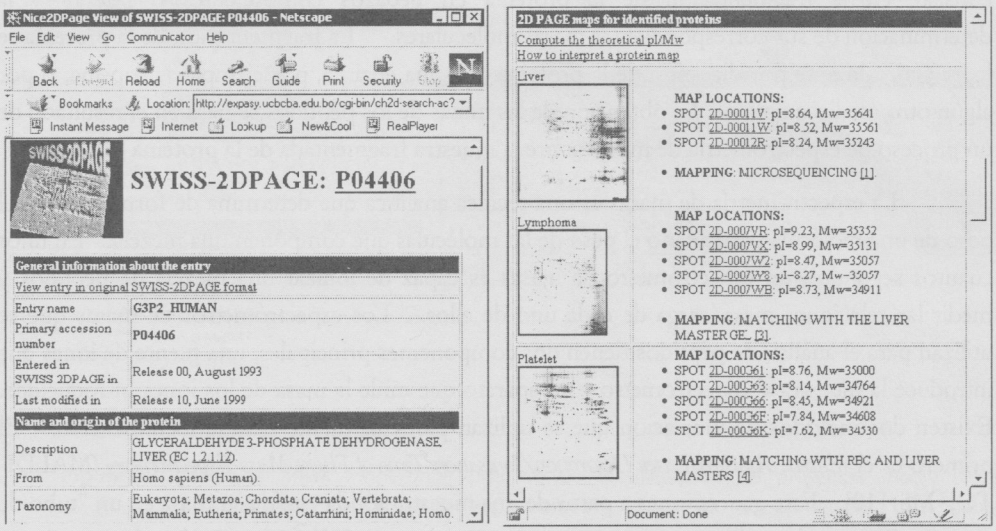

Figura 4. Esta base de datos contiene mapas proteicos que muestran las posiciones de las proteínas, así como también sus posiciones teóricas. Además, puede ser accedida remotamente mediante Internet a través del servidor ExPASy. Luego de realizar la consulta, dado un número de acceso a la base de datos, SWISS-2DPAGE muestra el nombre de la proteína, la posición de la proteína en un mapa 2D PAGE, su descripción, una lista de mapas en los que se ha identificado la proteína y otra información específica.

[2] Instituto de Investigación en Informática Aplicada, Universidad Católica Boliviana.

El almacenamiento de estos mapas de referencia involucra grandes volúmenes de información. Por esto, es necesaria la utilización de bases de datos de proteínas que integren esta información y sean accesibles a los investigadores de todo el mundo (Figura 4). Por tanto, estas bases de datos permiten compartir e integrar información concerniente a las proteínas contenida en imágenes 2D PAGE, así como también información relacionada: secuencias de proteínas, composición de aminoácidos, etc. Actualmente, existen varias bases de datos proteicas accesibles mediante Internet [6]. Cada una de estas bases de datos almacena información de acuerdo a un interés específico.

3.4. Identificación de proteínas utilizando su Peptide Mass Fingerprint

Alternativamente, después de su separación mediante la técnica 2D PAGE, la proteína de interés es aislada y físicamente extraida para luego determinar algunos de sus atributos experimentalmente utilizando el análisis llamado *peptide mass fingerprint* (PMF). Este método consiste en la fragmentación de la proteína en pedazos componentes o *péptidos* y la determinación de sus correspondientes masas moleculares. La fragmentación de la proteína se logra generalmente por digestión de la proteína con una enzima, por ejemplo la tripsina, o por algún otro medio químico. La obtención de las masas de los péptidos se basa en el resultado de un proceso de espectrometría de masas sobre la muestra fragmentada de la proteína [3].

La espectrometría de masas es una técnica analítica que determina de forma precisa el peso de una molécula particular o el peso de las moléculas que componen una mezcla. En unos cuantos segundos, un espectrómetro de masas es capaz de ionizar una mezcla de péptidos y medir las relaciones masa/carga de cada uno de ellos. Los espectrómetros de masas que se utilizan para el análisis de péptidos tienen dos componentes principales: una fuente de iones que introduce la muestra al espectrómetro y un aparato que mide la masa de los iones introducidos. Existen dos métodos de ionización que se utilizan particularmente para la identificación. El primero es el *Matrix Assisted Laser Desorption/Ionisation Time of Flight Mass Espectrometry* (MALDI-TOF MS) [10]. Éste genera iones a partir de una muestra sólida y mide su masa en un "tubo de vuelo". El segundo es *Electrospray Ionisation Mass Espectrometry* (ESI MS) [10], el cual genera iones a partir de una muestra líquida y mide su masa en un aparato que puede también ser un detector de tiempo de vuelo. Si bien ambos métodos producen resultados similares, aparentemente MALDI-TOF MS está emergiendo como la mejor alternativa para el análisis de proteínas [3]. En el MALDI-TOF MS la muestra es depositada sobre una sonda por co-cristalización con una *matriz* (ácido aromático que es fácilmente soluble y absorbe la luz del láser utilizado para la ionización) y luego es introducida a la cámara de ionización que se encuentra al vacío. La ionización es inducida por pulsos cortos de un láser enfocado en la muestra. En las configuraciones más típicas, las fuentes de iones de MALDI están acopladas a un analizador de masas que mide el *tiempo de vuelo*. La relación masa a carga (m/z) de iones formados por este método de ionización son luego medidos en el analizador de masas. Conociendo la carga de los iones medidos, es fácil deducir su masa. Con este método, es posible lograr exactitudes de 0.01% al analizar polipéptidos con masas de hasta 30-40 kDa [10]. La espectrometría de masas da como resultado un espectro donde el eje horizontal corresponde a la relación masa/carga de

la partícula medida y el eje vertical corresponde al número de partículas medidas que tengan esa relación particular.

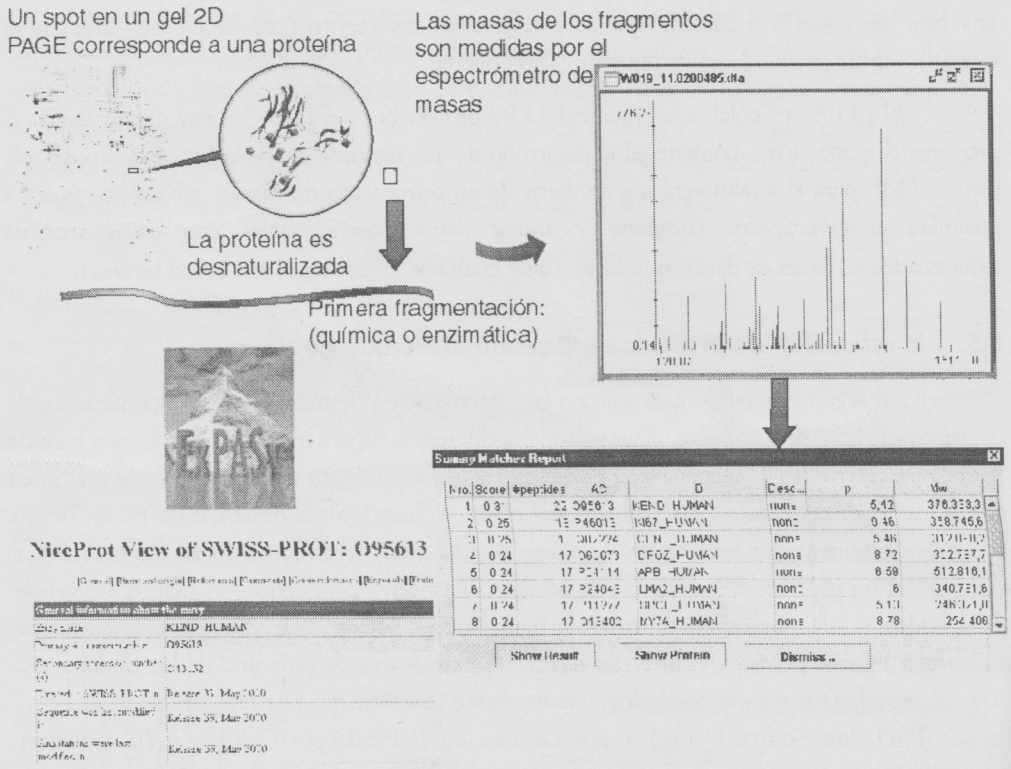

Figura 5. *Peptide Mass Fingerprint,* proceso de identificación de proteínas mediante la comparación de datos de espectrometría de masas experimentales con datos teóricos contenidos en bases de datos.

A pesar que la interpretación de los espectros generados por esta técnica puede ser complicada, estos datos proveen información importante para: primero, determinar si éstos corresponden a una proteína conocida y/o modificada; y segundo, para determinar si más de una proteína es co-migradora en una banda de un spot en un gel. Esta información junto con el uso de recursos como las bases de datos de secuencias de proteínas forman parte del corazón de la identificación de proteínas basadas en MS y es lograda mediante el uso de la informática.

El procedimiento de identificación consiste en comparar el espectro experimental con espectros teóricos. Esto se logra calculando todas las posibles masas que pueden ser generadas fragmentando todas las secuencias en una cierta proteína contenida en una Base de Datos. Los resultados son desplegados como una lista de proteínas candidatas ordenada de acuerdo a un criterio de bondad (Figura 5). Esta técnica, sin embargo, presenta algunas falencias: el

programa no puede determinar a ciencia cierta si una identificación es correcta; además, para lograr una identificación, es necesario que la proteína buscada se encuentre en la base de datos y es deseable, por lo tanto, contar con la mayor cantidad de registros en la base. Además, existe también la posibilidad que se encuentren masas de péptidos similares en diferentes elementos de una base de datos y a medida que se incrementan los registros de esta base, también se incrementan las posibilidades de tener un falso positivo.

El I.I.I.A., en colaboración con el S.I.B.[3] de Ginebra y el H.U.G.[4], está desarrollando el proyecto *Biograph.* Este consiste en el desarrollo de una herramienta de software de apoyo a la técnica PMF para el análisis gráfico de datos de espectrometría de masas. Esta herramienta posibilita al investigador comparar y anotar datos experimentales con datos teóricos almacenados en bases de datos mundiales y a las cuales se accede mediante la red Internet.

3.5. *Idenficación de proteínas utilizando el Método* de Novo

Un segundo método que utiliza a la espectrometría de masas para la identificación de proteínas es el método *de Novo.* Este método, a diferencia del anterior, no necesita una base de datos para determinar la secuencia de aminoácidos que conforman a un péptido y se realiza en dos etapas bien definidas. En una primera etapa (primer MS) se obtienen las masas de los pétidos ya fragmentados previamente; en una seguna etapa (segundo MS) se fragmenta por segunda vez un péptido seleccionado y se obtiene su espectro pemitiendo un análisis a nivel de aminoácidos. Algunos espectrómetros de masa pueden seleccionar y fragmentar de forma automática algunos péptidos y medir las relaciones masa/carga de cada uno de los fragmentos. Esta segunda fragmentación puede lograrse mediante dos técnicas: por Descomposición Post-Fuente (PSD Post Source Decay) ó, por Disociación Inducida por Colisión (CID Collision Induced Dissociation).

En la primera técnica, la muestra es bombardeada con un haz de luz láser de más alto poder que lo normal para dar mayor energía e ionizar a la muestra, la cuál se fragmenta al momento de ingresar al tubo del espectrómetro. En la segunda técnica, la muestra también es bombardeada con un haz de luz láser, pero la fragmentación se la realiza en el interior del tubo de vuelo mediante la colisión de la muestra con un gas inerte. Este método es llamado MS/MS porque consiste en la aplicación doble del proceso de espectrometría de masas sobre una muestra y consiste en elegir electrónicamente en el espectrómetro uno de los péptidos que sea de interés. Este péptido lleva el nombre de ión padre. Posteriormente el ión padre es reabsorbido en el espectrómetro donde se realiza un CID o PSD para fragmentarlo nuevamente y obtener pedazos que, a su vez, son medidos por el espectrómetro de masas. El resultado del proceso de doble de fragmentación (MS/MS) es también un espectro donde el eje de absisas corresponde a las masas de los fragmentos medidos y el eje de ordenadas corresponde al número de iones medidos para la masa correspondiente. El problema de secuenciamiento

[3] Swiss Institute of Bioinformatics, Ginebra - Suiza.

[4] Hospital Universitario de Ginebra - Suiza.

consiste en derivar la secuencia de aminoácidos que conforman al péptido analizado dado sus espectros MS/MS. De esta manera, es posible deducir la cadena de aminoácidos que componen al ión padre. Para un proceso de fragmentación ideal y para un espectrómetro de masas ideal, la secuencia de un péptido puede ser determinada simplemente comparando la diferencia de masas de iones consecutivos de un espectro con las masas teóricas de los aminoácidos (Figura 6). Esta situación ideal ocurriría si el proceso de fragmentación pudiese ser controlado de tal modo que cada péptido estuviese cortado entre dos aminoácidos consecutivos y una única carga fuese retenida solamente en el pedazo con terminal-N. Sin embargo, en la práctica, el proceso de fragmentación en los espectrómetros de masas está lejos de ser ideal. Como resultado, el secuenciamiento de péptidos continua siendo un problema abierto [11].

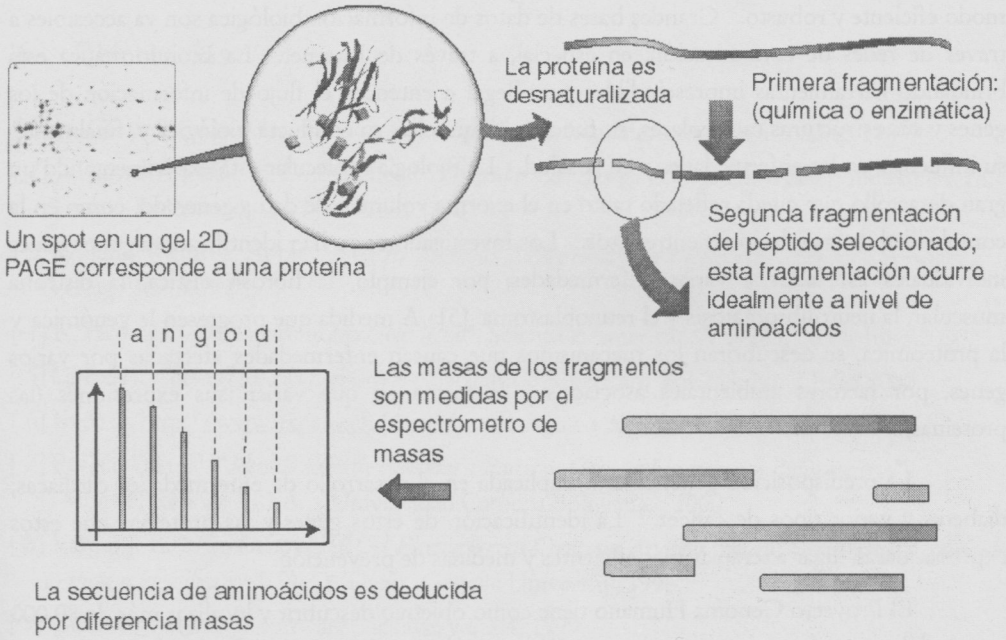

Figura 6. Determinación de la secuencia de aminoácidos via el proceso *De Novo* (MS-MS).

La ventaja de este método, a diferencia de los mencionados previamente, radica en que su forma operativa no depende de una base de datos y la determinación de secuencias de aminoácidos de las muestras analizadas puede lograrse sin tener datos históricos del péptido o proteína en cuestión.

El I.I.I.A., en colaboración con la Universidad de Ginebra, el H.U.G. y la empresa Geneva Bioinformatics, está investigando nuevos algoritmos para el secuenciamiento automático de los espectros generados por MS/MS. Resultados preliminares muestran que, pese a la complejidad del problema, es posible automatizar este proceso [7].

4. Perspectivas Futuras

La investigación en genética molecular está revolucionando a pasos gigantescos la práctica médica y la investigación biológica. La práctica médica se verá radicalmente alterada cuando se combine la información genética con las nuevas tecnologías clínicas basadas en el diagnóstico del ADN. Se vislumbra el nacimiento de la "Medicina Molecular" que se caracterizará, no sólo por el tratamiento de los síntomas de las enfermedades, sino por la búsqueda de las causas últimas de éstas. Se desarrollarán nuevas técnicas de diagnóstico, más rápidas y más fiables que facilitarán la prevención de las enfermedades. Se desarrollarán fármacos más específicos y efectivos e incluso se practicará la terapia génetica.

Las herramientas informáticas son cruciales para almacenar e interpretar datos de un modo eficiente y robusto. Grandes bases de datos de información biológica son ya accesibles a través de redes de comunicación, en especial, a través de Internet. La Bioinformática está brindando herramientas imprescindibles para llegar a entender el flujo de información de los genes y sus estructuras moleculares, su función bioquímica, su conducta biológica y, finalmente, su influencia en las enfermedades y en la salud. La Biología Molecular está experimentando un gran desarrollo que queda reflejado tanto en el enorme volumen de datos generado, como en la complejidad de las relaciones entre éstos. Los investigadores ya han identificado algunos genes individuales asociados a varias enfermedades; por ejemplo, la fibrosis cística, la distrofia muscular, la neurofibromatosis y el retinoblastoma [5]. A medida que progresen la genómica y la proteómica, se descubrirán los mecanismos que causan enfermedades afectadas por varios genes, por factores ambientales asociados y la forma en que varían sus expresiones (las proteínas).

La predisposición genética está implicada en el desarrollo de enfermedades cardíacas, diabetes y varios tipos de cáncer. La identificación de estos genes y las proteínas que estos expresan darán lugar a terapias más eficientes y medidas de prevención.

El Proyecto Genoma Humano tiene como objetivo descubrir y localizar más de 80.000 genes del genoma y poner toda esta información a disposición de la comunidad de investigadores en Biomedicina. El estudio del genoma también contribuye al entendimiento del desarrollo embrionario y el envejecimiento humano. La investigación genética encontrará múltiples aplicaciones en las industrias farmaceúticas y alimenticias. También, en Salud Pública se observan esfuerzos encaminados a la diseminación de información genética, la formación de los profesionales de la Salud y el desarrollo de políticas que incorporen el conocimiento genético en la práctica epidemiológica.

Entre los impactos potenciales de las tecnologías de adquisición y gestión de la información genética, aplicadas en la investigación biomédica, se pueden citar: medicina preventiva, medicamentos personalizados por genotipo, procesamiento paralelo masivo de información genética, diagnóstico en las postas médicas, sistemas de estudio epidemiológico-genéticos, etc.

El I.I.I.A., a través de los convenios de colaboración con el S.I.B. y el H.U.G., está investigando tecnologías de punta y desarrollando algoritmos que apoyan directamente al progreso de la bioinformática, la biología molecular y la medicina.

5. Agradecimientos

Este trabajo está apoyado por el Instituto Suizo de Bioinformática, el Laboratorio Central de Química Clínica del Hospital Universitario de Ginebra y Geneva Bioinformatics SA (Genebio). Agradecimientos especiales para el Dr. Ron Appel, el Prof. Denis Hoschtrasser, el Dr. Robin Grass y el Dr. Manfred Heller.

6. Referencias

[1] H. Curtis "*Biología*" 4ta. Edición Editorial Medica Panamericana Buenos Aires 1983.

[2] J. Darnell, H. Lodish, D. Baltimore "*Mollecular Cell Biology*" 2nd. Edition Scientific American Books - Freeman & Co. New York 1990.

[3] Wilkins, Williams, Appel. "*Proteome Research: New Frontiers in Functional Genomics*". Springer - Verlag. Berlin 1997.

[4] K. Howard. "*The Bioinfomatics Gold Rush*", Scientific American, Vol 283, No 1. Julio 2000.

[5] C. Ezzell. "*Beyond the Human Genome*", Scientific American, Vol 283, No 1 Julio 2000.

[6] Expasy http://www.expasy.ch (Accedido en Agosto 2000)

[7] Pavisic D et. al. "*Identificación de Proteínas a Partir de Geles de Electroforesis Bidimensional*". Technical Report: I.I.I.A.. Universidad Católica Boliviana. 2000.

[8] Vargas J. R. "*Two-Dimensional Gel Electrophoresis Computer Analysis System: From Image Acquisition to Protein Identification*". Ph.D. Thesis, Geneva University, 1996.

[9] GenBio, Melanie-II software package, ver 2.4, 1999.

[10] S. Patterson, R. Aebersold. "*Mass Spectrometric Approaches for the Identification of Gel-Separated Proteins*". Electrophoresis 16, 1995.

[11] V. Dancik, T. Addona, K. Clauser, J. Vath, Pavel Pevzner. "*De Novo Peptide Sequencing via Tandem Mass Epectrometry*" - Journal of Computational Biology, Vol 6, No. 3/4, 1999.

[12] The Human Genome Project "*To Know Ourselfs*" U.S. Department of Energy and The Human Genome Project. 1996.

[13] Appel R.D., Sanchez J-C., Bairoch A., Golaz O., Miu M., Vargas R., Hochstrasser D. "*SWISS-2DPAGE: A Database of Two-dimensional Gel Electrophoresis Images*". Electrophoresis 14, 1232-1238, 1993.

[14] Appel R.D., Bairoch A., Hochstrasser D.F. "A New Generation of Information Retrieval Tools for Biologists: The Example of the ExPASy WWW Server". Trends Biochem. Sci. 19:258-260, 1994.