Functional Site Identification
Problem 6
The Ten Most Wanted Solutions in Bioinformatics

Diana Hermith

Seminario de Investigación
Grupo Destino
Pontificia Universidad Javeriana, Cali
2008

Agenda

- **Introducción**

- **Metodologías Propuestas:**

  - **Genómica Estructural**
  - **Superposición Estructural**
  - **Clasificación Estructural de Proteínas**
  - **Detección del Sitio Activo**

INTRODUCCIÓN

Introducción

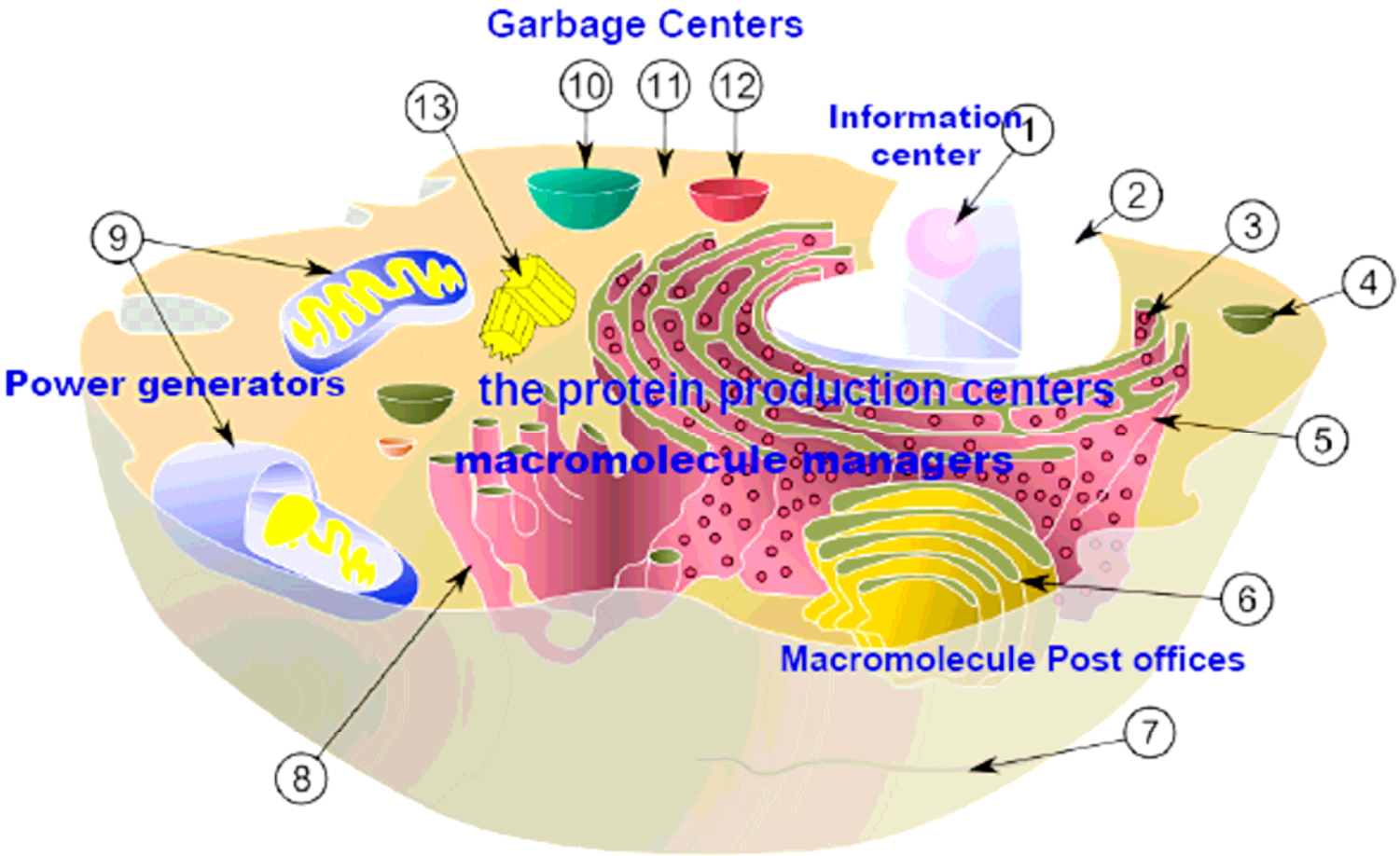
Las proteínas intervienen en todas las funciones vitales de los organismos

Podemos categorizar sus funciones en:

- **Enzimáticas**
- **Hormonales**
- **Transporte**
- **Defensa**
- **Movimiento**
- **Reserva**
- **Transmisión de señales**
- **Reconocimiento de señales**
- **Estructurales**
- **Regulación**

Biochemistry, 3rd Edition, Mathews, van Holde, Ahern. Mc Graw Hill

Cell consists of systems



(1) nucleolus (2) nucleus (3) ribosome (4) vesicle (5) rough endoplasmic reticulum (ER) (6) Golgi apparatus Cytoskeleton (8) smooth ER (9) mitochondria (10) vacuole (11) cytoplasm (12) lysosome (13) centrioles

American Chemical Society, Jun Xu, Ph. D., January 24, 2008, San Diego

Introducción

Término	Definición
Homólogo	Surgieron de una proteína ancestral en común, y su relación evolutiva es evidente por similitudes en la secuencia, la estructura y/o en la función.
Análogo	Son similares de alguna forma, pero no hay evidencias de ancestro común. Análogos estructurales comparten el mismo plegamiento, y análogos funcionales la misma función.
Ortólogos	Son genes equivalentes en diferentes especies que surgieron de un ancestro común por especiación.
Parálogos	Surgieron por duplicación de genes dentro de un genoma, y tienen funciones diferentes, pero generalmente relacionadas.
Residuos funcionales	Incluyen residuos de unión, en contacto con sustrato y cofactor, y residuos catalíticos que intervienen en el mecanismo enzimático.

Molecular Biology: an Information Science

- Central Dogma of Molecular Biology

```
DNA
 -> RNA
  -> Protein

   -> Phenotype
    -> DNA
```

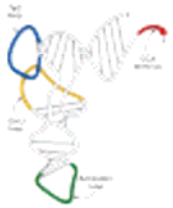
- Molecules
  - ◊ Sequence, Structure, Function
- Processes
  - ◊ Mechanism, Specificity, Regulation

- Central Paradigm for Bioinformatics

```
Genomic Sequence Information
 -> mRNA (level)
  -> Protein Sequence
   -> Protein Structure
    -> Protein Function
     -> Phenotype
```

- Large Amounts of Information
  - ◊ Standardized
  - ◊ Statistical

(idea from D Brutlag, Stanford, graphics from S Strobel)



•Most cellular functions are performed or facilitated by proteins.
•Primary biocatalyst
•Cofactor transport/storage
•Mechanical motion/support
•Immune protection
•Control of growth/differentiation

•Genetic material

•Information transfer (mRNA)
•Protein synthesis (tRNA/mRNA)
•Some catalytic activity

Introducción

La conservación evolutiva de un patrón en una secuencia es indicativo de la presión selectiva, y, porque la selección natural actúa sobre la función, los residuos son conservados, ya sea directamente responsables de la función o para el mantenimiento de la estructura que permite el correcto posicionamiento de los residuos funcionalmente importantes.

Introducción

En este problema, se pretende abordar el problema de cómo detectar los residuos funcionalmente relevantes cuando se conoce la estructura tridimensional de la proteína

Dificultad del problema:

- La definición de “función” de una proteína es ambigua (aspectos bioquímicos, biológicos, fisiológico, clínico, etc.).
- Función es todo aquello que le sucede o se hace a través de una proteína.
- La caracterización de las proteínas o “anotaciones” almacenadas en las bases de datos no tienen un lenguaje estandarizado, y en consecuencia son difícil de interpretar por los sistemas de cómputo.
- Generalmente las anotaciones están en lenguaje natural con cometarios particulares.

METODOLOGIAS PROPUESTAS

Metodologías Propuestas

- **Genómica Estructural**
- Superposición Estructural
- Clasificación Estructural de Proteínas
- Detección del Sitio Activo

Genómica Estructural

La cristalografía no puede resolver un gran porcentaje de las estructuras de las proteínas en los años inmediatos.

Genómica estructural es una iniciativa para resolver un gran porcentaje mediante modelaje por homología. Existen actualmente 9 centros/consorcios de genómica estructural en USA, además de los proyectos comerciales y centros de otros países, hasta un total de >20 (>70 instituciones).

1. Identificar para las cuales **no existe un modelo experimental.**
2. Escoger algunos miembros de cada familia como "**objetivos**".
    a. El Protein Data Bank mantiene un registro de objetivos.
    b. En febrero de 2004, se había registrado aproximadamente 50.000 objetivos
3. Resolver un objetivo para cada familia mediante **cristalografía de alta resolución**, proporcionando un nuevo modelo.
    a. Este es el **cuello de botella**.

1. Obtenga **milígramos** de la proteína **pura, soluble**:

a. **Aproximadamente 60% de las secuencias se expresan bien.**

b. **Solo la mitad (procariotas) o la cuarta parte (eucariotas) de éstas son solubles.**

2. Obtenga cristales de alta calidad:

a. **Generalmente se prueban cientos de condiciones de cristalización diferentes.**

b. **Los cristales tienen que estar sueltos (sin pegarse unos con otros), suficientemente grandes, y preferentemente ni en forma de agujas u hojas.**

c. **Los cristales con gran número de copias en la unidad asimétrica son problemáticos.**

d. **Aproximadamente la mitad de las proteínas solubles que se expresan de forma eficiente cristalizan, pero únicamente un tercio de estos cristales son útiles.**

b. Inferir la función a partir de estructuras similares de función conocida, para confirmarlo bioquímicamente.

c. Realizar modelaje por homología a todos los miembros de cada familia, utilizando los nuevos modelos.

Metodologías Propuestas

- Genómica Estructural
- **Superposición Estructural**
- Clasificación Estructural de Proteínas
- Detección del Sitio Activo

Root Mean Square Deviation

We previously introduced the definition of root mean square deviation:

$$rmsd = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left[(x_i - x_i')^2 + (y_i - y_i')^2 + (z_i - z_i')^2\right]}$$

where (x_i, y_i, z_i) and (x_i', y_i', z_i') are the coordinates of the atoms that we want to superimpose to each other.

If the correspondence between the pairs of atoms we want to superimpose is known, we can easily measure their rmsd and also calculate how to optimally superimpose them. We apply the rigid-body translation $T = (T_x, T_y, T_z)$ and rotation $R = (R_x, R_y, R_z)$ to one of the proteins that minimize the rmsd between the given set of atom pairs:

$$rmsd(T, R) = \min_{T,R} \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left[(x_i - R_x x_i' + T_x)^2 + (y_i - R_y y_i' + T_y)^2 + (z_i - R_z z_i' + T_z)^2\right]}$$

Superposición Estructural

Cuando se comparan dos estructuras de proteínas con secuencias idénticas,
como una estructura y un modelo o dos determinaciones estructurales
de la misma proteína en diferentes condiciones, podemos superponer cada par
equivalente de átomos, pero el resultado puede ser engañoso. Si una región de la
modelo es incorrectamente posicionado con respecto al resto de la estructura, la
rmsd pueden ser muy elevados y una superposición puede dejar de destacar la similitud
de las demás regiones de la estructura.

Superposición Estructural

Debemos decidir si preferimos tener más puntos equivalentes a expensas de un menor rmsd y cuánta "Calidad" estamos dispuestos a perder para lograr una superposición que incluya más átomos de nuestras proteínas.

La óptima superposición de dos estructuras de proteínas sólo puede definirse si decidimos o bien el número mínimo de los átomos que queremos superponer uno sobre el otro; o el valor máximo de rmsd que estamos dispuestos a aceptar.

Matrices de Distancia

En una matriz de distancia, cada columna y cada fila de la matriz representa un aminoácido de una proteína, y cada celda contiene la distancia intermolecular entre los aminoácidos de la fila y los aminoácidos de la columna. La distancia puede calcularse entre Cαs o Cβs, en función de la aplicación específica.

Matrices de Distancia

Proteínas con similares estructuras tridimensionales, tienen conjuntos similares de distancia interresiduos, y por lo tanto, similares matrices de distancia.

Si comparten una semejanza estructural local, esta característica se refleja en la presencia de un patrón similar en la matriz de distancia, y la detección de submatrices similares en las dos matrices pueden proporcionar el punto de partida para la superposición estructural.

Esta estrategia se utiliza en el muy popular DALI método. http://es.wikipedia.org/wiki/Alineamiento_estructural

Metodologías Propuestas

- Genómica Estructural
- Superposición Estructural
- **Clasificación Estructural de Proteínas**
- Detección del Sitio Activo

Dominios

Las proteínas mas grandes tienen dominios reconocibles, pensados como unidades evolutivas, unidades estructurales compactas, que se pueden plegar independientemente y pueden formar proteínas multimodulares e incluso, un dominio insertarse en otro.

Todos los esquemas de clasificación de proteínas apuntan a dividirlas según los dominios, pero la información estructural es muy incompleta (a la mayoría de las proteínas no se les ha determinado la estructura).

Clasificación de dominios proteicos

CATH: Los dominios son asignados por *herencia* (para proteínas con secuencia muy similar a otra que ya esté en la base de datos) o por *consenso* (si en tres algoritmos diferentes se deduce la misma estructura) o *manualmente* (si los algoritmos no aportan la misma respuesta).

Clasificación de dominios proteicos

SCOP: Las asignaciones son hechas manualmente y las proteínas pertenecen a un dominio sólo si uno de los dominios existe independiente en la base de datos. La organización es jerárquica: clase, plegado, súperfamilia y familia.

Clasificación de dominios proteicos

DALI y VAST: Son totalmente automatizados. Las proteínas se consideran de la misma familia de homólogas si cumplen una de estas condiciones:

- evidencia de similitud de secuencia significativa.
- evidencia de similitud de estructura significativa.
- evidencia de similitud funcional, ubicación y configuración del sitio activo.

Metodologías Propuestas

- Genómica Estructural
- Superposición Estructural
- Clasificación Estructural de Proteínas
- **Detección del Sitio Activo**

Detección del Sitio Activo

Tal vez la forma más general para definir regiones funcionales en las proteínas es cuando estas se unen a otras moléculas, a los sustratos en los sitios activos de enzimas,
y con otras macromoléculas como en anticuerpos y proteínas reguladoras.

¿Podemos
detectar sitios activos sobre la base del conocimiento de la estructura tridimensional?

Detección del Sitio Activo

La secuencia genética cambia, cambiando la estructura tridimensional, y por lo tanto la función.

Posibles Camino a nuevas funciones

Nuevo gen – nueva proteína – nueva función

Moonlighting: el mismo producto génico puede tener distintas funciones en distintos ambientes, la función es dependiente del pH, localización celular, ligandos disponibles, etc.

Mutaciones: mutaciones locales que llevan a una actividad distinta, pero relacionada con la original.

Detección del Sitio Activo

Oligomerización: nuevas funciones con oligómeros idénticos, relacionados, o muy diferentes entre sí.

Duplicación: nuevo producto génico con nueva función.

Construcción modular (mix and match), nuevas actividades al unirse dos genes.

Actividad catalítica y especificidad de sustrato

La actividad catalítica y la especificidad por el sustrato caracterizan las propiedades funcionales de las enzimas. Las enzimas primitivas probablemente tuvieron un amplio rango de actividades y/o especificidad por el sustrato, y se fueron especializando al incorporar en la evolución grupos funcionales adicionales.

Especificidad de sustrato

Está directamente determinada por la naturaleza de los residuos de unión, aprovechándose las propiedades de los 20 aminoácidos. Incluso la especificidad puede ser rediseñada en laboratorio por mutaciones puntuales.

Especificidad de sustrato

Los loops superficiales generalmente contienen los determinantes estructurales para la especificidad de sustrato, y esto facilita la rápida divergencia evolutiva y la adaptación, ya que puede variar la estructura del loop sin afectar el plegamiento global de la proteína.

Especificidad de sustrato

La naturaleza química de los residuos involucrados en la interacción no es necesariamente el único determinante de la especificidad, la mutación de algunos residuos lejanos le dan forma al sitio activo para la complementariedad del sustrato.

Mecanismo catalítico

Algunas familias enzimáticas presentan gran divergencia en sus actividad catalítica, como también en su especificidad por el sustrato. El estudio detallado de su estructura y sus características bioquímicas revelaron que las diferentes actividades y sus reacciones asociadas compartían cofactor, nucleófilo o mecanismo de reacción.

Mecanismo catalítico

Con el descubrimiento de proteínas homólogas que catalizan un paso bioquímico común en los contextos de diferentes reacciones globales, se vió que los residuos esenciales para dicha actividad se conservan durante la evolución, mientras que los grupos funcionales adicionales determinan el destino del intermediario de reacción, como la especificidad por el sustrato.

Los problemas que afectan a nuestra capacidad para detectar los sitios son funcionales, en la práctica, son causados por la extrema versatilidad de las estructuras de las proteínas.

Su capacidad para finamente modular su actividad, es el resultado de su flexibilidad y capacidad para asumir ventaja del medio ambiente, que es lo que se interpone en el camino de muchos métodos automáticos.

Promising Avenues

- A combination of heuristic and chemical knowledge is required.

- Searching all the proteins of known structure for similar clusters of amino acids to detect common subsites potentially involved in function is Computationally prohibitive.

- We must use our chemical knowledge to reduce the number of patterns to be compared, and we can do so by only considering residues that are likely to be involved in chemical reactions because of their chemicophysical properties.

- We can use evolutionary information to discard nonconserved residues.