Seminario Bioinformática: The ten most wanted solutions in protein bioinformatics. Problem 1

Gloria Inés Alvarez V.

Departamento de Ciencias e Ingeniería de la Computación
Pontificia Universidad Javeriana Cali

Periodo 2008-2

Introducción

- El objetivo de la bioinformática es asistir a la biología experimental en asignar función o sugerir hipótesis funcionales para todas las proteinas.
- Es más fácil y rápido obtener una secuencia de nucleótidos que codifican una proteina, que secuenciar químicamente la proteina misma.
- El material genético no contiene sólo genes, también hay:
  - Regiones regulatorias
  - Regiones no codificantes de función desconocida
  - Repeticiones largas y cortas
  - Pseudogenes
  - Retropseudogenes
  - Regiones microsatélite, minisatélite y satélite
  - Transposons and retrotransposons
  - Vestigios virales
  - Intrones

Introducción

Este libro no aborda el problema de detectar genes, que es un inmenso reto que la genómica le pone a la bioinformática. En su lugar, nos concentramos en técnicas que se pueden aplicar a derivar conocimiento funcional de una proteina, una vez se conoce su secuencia completa de aminoácidos.

Estructura de las proteinas

- La función de una proteina depende de su forma tridimensional.
- Una proteina es un polímero especial que, dadas ciertas condiciones de pH, temperatura y fuerza iónica, asume una y sólo una estructura tridimensional específica.
- El resto de este capítulo se dedica a la estructura de proteinas que permanecen en un ambiente polar, qué es un ambiente polar?
- La secuencia de aminoácidos de una proteina constituye su estructura primaria.
- La estructura secundaria se compone de regiones α y β.
- Una proteina formada por más de una cadena de aminoácidos tiene estructura cuaternaria.

Estructura de las proteinas

- Una proteina está formada por dos lóbulos o dominios.
- Los dominios no son idénticos, pero tienen una semejanza topológica y pueden describirse como estructuras de tres capas: dos regiones externas α (hélices) y una parte central en forma de hoja β.
- Los mecanismos celulares pueden modificar químicamente la estructura primaria de una proteina después que esta ha sido sintetizada. Los cambios pueden ser temporales o permanentes.

Problema 1: Alineamiento de secuencias de proteinas

- Las proteinas se van modificando por aparición de nuevos aminoácidos. La nueva versión puede ser viable o no, es decir, puede seguir realizando correctamente su función o no, y puede hasta volverse común si confiere alguna ventaja selectiva al individuo.
- En algunos casos no es posible realizar modificación de algunos aminoácidos porque se pierde la función.
- Al ubicar los aminoácidos inamovibles por función, se puede establecer esa función en otras cadenas que lo contengan y que hayan evolucionado de la cadena conocida.

Problema 1: Alineamiento de secuencias de proteinas

- Dadas dos secuencias de proteinas, calcular la probabilidad de que ellas tengan un ancestro común, es decir, que sean homólogas.
- Dadas dos secuencias de proteinas que son homólogas, identificar todos los pares de aminoácidos de las dos proteinas que derivan de un mismo aminoácido en el ancestro común.
- Los dos anteriores se pueden juntar planteando el problema así: dadas dos secuencias de aminoácidos, identificar la correspondencia entre pares de aminoácidos que maximizan la probabilidad de que ellos deriven de un ancestro común y calcular la probabilidad de que tal ancestro exista.
- Se supone que la correspondencia óptima es la que minimiza el número de mutaciones necesarias para convertir una cadena en la otra.

Ortología

Supongamos el gen Ga que codifica la proteina Pa, la cual realiza una función Fa.

- Ortología: suponer que se da un proceso de especiación y después de ocurrir mutaciones las dos especies tienen genes Ga' y Ga'' que codifican respectivamente las proteinas Pa' y Pa'', dichas proteinas siguen siendo compatibles con la función Fa. Pa' y Pa'' son ortólogas, comparten aminoácidos y realizan la misma función.

Paralogía

Supongamos el gen Ga que codifica la proteina Pa, la cual realiza una función Fa.

- Paralogía: suponer que la porción del gen que contiene Ga se duplica, dando orígen a una copia idéntica Gb que en principio codifica una proteina Pb idéntica a Pa, al ocurrir un proceso de especiación y mutaciones posteriores se tienen dos genes Ga', Gb' y Ga'', Gb'' tales que: $Ga' y Ga''$ siguen codificando una proteina que realiza la función Fa, sin embargo, las copias $Gb' y Gb''$ habrÃ¡n mutado sin la restricción de seguir cumpliendo la función Fa y de hecho pueden haber adquirido la propiedad de realizar otra función distinta. A pesar de eso si se comparan Ga', Gb' se encontrará que comparten muchos aminoácidos, ellas son parálogas.

Familias de Proteinas

Cuando dos proteinas empiezan a diverger, acumulan cambios hasta el punto que pueden llegar a parecerse tanto como lo hacen dos secuencias no relacionadas entre sí. Sin embargo, detectar una relación lejana es supremamente importante porque:

- Esto incrementa el número de proteinas con las que se puede realizar inferencia funcional.
- Facilita la detección de regiones funcionalmente importantes.
- Puede revelar encadenamientos evolutivos inesperados entre organismos, lo que puede ayudar a entender la forma como se desarrolla la vida.

Para detectar estas relaciones remotas se usa el hecho que la homología es transitiva. AsÃ se construyen familias de proteinas relacionadas evolutivamente entre sí.

Matrices de Similaridad

Un aminoácido no puede ser reemplazado por cualquier otro durante el proceso de mutación. Ellos deben tener propiedades químicas y eléctricas similares. Se debe estimar la probabilidad de que un amino
acido reemplace a otro. Esto se hace experimentalmente y se reportan en tablas llamadas de similaridad o de sustitución.

	A	R	N	D	C	O	E	C	H	I	L	K	M	F	P	S	T	W	Y	V
A	2	-2	0	0	-2	0	0	1	-1	-1	-2	-1	-1	-3	1	1	1	-6	-3	0
R	-2	6	0	-1	-4	1	-1	-3	2	-2	-3	3	0	-4	0	0	-1	2	-4	-2
N	0	0	2	2	-4	1	1	0	2	-2	-3	1	-2	-3	0	1	0	-4	-2	-2
D	0	-1	2	4	-5	2	3	1	1	-2	-4	0	-3	-6	-1	0	0	-7	-4	-2
C	-2	-4	-4	-5	12	-5	-5	-3	-3	-2	-6	-5	-5	-4	-3	0	-2	-8	0	-2
Q	0	1	1	2	-5	4	2	-1	3	-2	-2	1	-1	-5	0	-1	-1	-5	-4	-2
E	0	-1	1	3	-5	2	4	0	1	-2	-3	0	-2	-5	-1	0	0	-7	-4	-2
G	1	-3	0	1	-3	-1	0	5	-2	-3	-4	-2	-3	-5	0	1	0	-7	-5	-1
H	-1	2	2	1	-3	3	1	-2	6	-2	-2	0	-2	-2	0	-1	-1	-3	0	-2
I	-1	-2	-2	-2	-2	-2	-2	-3	-2	5	2	-2	2	1	-2	-1	0	-5	-1	4
L	-2	-3	-3	-4	-6	-2	-3	-4	-2	2	6	-3	4	2	-3	-3	-2	-2	-1	2
K	-1	3	1	0	-5	1	0	-2	0	-2	-3	5	0	-5	-1	0	0	-3	-4	-2
M	-1	0	-2	-3	-5	-1	-2	-3	-2	2	4	0	6	0	-2	-2	-1	-4	-2	2
F	-3	-4	-3	-6	-4	-5	-5	-5	-2	1	2	-5	0	9	-5	-3	-3	0	7	-1
P	1	0	0	-1	-3	0	-1	0	0	-2	-3	-1	-2	-5	6	1	0	-6	-5	-1
S	1	0	1	0	0	-1	0	1	-1	-1	-3	0	-2	-3	1	2	1	-2	-3	-1
T	1	-1	0	0	-2	-1	0	0	-1	0	-2	0	-1	-3	0	1	3	-5	-3	0
W	-6	2	-4	-7	-8	-5	-7	-7	-3	-5	-2	-3	-4	0	-6	-2	-5	17	0	-6
Y	-3	-4	-2	-4	0	-4	-4	-5	0	-1	-1	-4	-2	7	-5	-3	-3	0	10	-2
V	0	-2	-2	-2	-2	-2	-2	-1	-2	4	2	-2	2	-1	-1	-1	0	-6	-2	4

Matrices de Similaridad

Se estima la probabilidad usando el criterio de máxima verosimilitud:

- Sea f_{ij} las veces que un par de aminoácidos i y j ocurren apareados en secuencias relacionadas evolutivamente.
- Sea f_i el número de veces que el aminoácido i aparece en esa familia de proteinas y sea f_j el número de veces que aparece el aminoácido j.
- $\frac{f_{ij}}{f_i f_j}$ es un estimativo de la verosimilitud de que i y j se hayan reemplazado el uno al otro durante la evolución.

Penalizaciones Indel

- Las matrices sirven para modelar estadísticamente las mutaciones, pero hay otros fenómenos poco comunes que también transforman las cadenas pero que no quedan registrados en las matrices por su escasa frecuencia de ocurrencia, ellos son las inserciones y borrados de minoácidos (llamadas indels o gaps). Estos eventos deben ser penalizados en un modelo para reflejar su baja probabilidad de ocurrencia.
- Las proteinas tienen una estructura compacta que dificulta los gaps, además la inserción o borrado de un aminoácido puede ser desestabilizante para la proteina.

Penalizaciones Indel

Algunas formas de modelar la penalización son:

- Asociar un valor fijo de penalización a cada inserción o borrado. Esto no es exacto porque los indels tienden a ocurrir en ciertas regiones de la proteina no en cualquier sitio.
- En el modelo lineal se penaliza cada gap con un valor w_0 que se aumenta en una cantidad $w_e < w_0$ por cada incremento en la longitud del gap.
- Un modelo alternativo penaliza cada aminoácido en un gap con un valor menor que el anterior. Por ejemplo, si al anterior se le penalizó con w_0, al siguiente se le penaliza con $log(w_0)$.

Alineamiento Local vs. Global

Definición

El alineamiento entre dos proteinas es global cuando pretende encontrar la correspondencia óptima entre todos lso aminoácidos de ambas secuencias.

Definición

El alineamiento entre dos proteinas es local cuando pretende detectar regiones de similitud entre las dos secuencias.

Alineamiento Global de Dos secuencias de Proteinas

- Se considera que el mejor alineamiento es el que requiere el menor número de sustituciones, inserciones y eliminaciones.
- El problema de alineamiento se puede reducir al problema de encontrar un alineamiento entre dos cadenas de símbolos que maximice un puntaje calculado con base en la matriz de similitudes de aminoácidos y un valor de penalización para los gaps. El alineamiento obtenido tiene interés biológico siempre y cuando los valores usados para calcular el puntaje correspondan a la realidad biológica de las cadenas de proteinas.

Alineamiento Global de Dos secuencias de Proteinas

El algoritmo de Needleman y Wunsch realiza el alineamiento eficientemente mediante programación dinámica.

$$F(i,j) = max \begin{cases} F(i-1,j-1) + \sigma(s_i, t_i) \\ F(i-1,j) - w_i \\ F(i,j-1) - w_j \end{cases}$$

Alineamiento Local de Dos Secuencias de Proteinas

Se puede hacer con una modificación del algoritmo anterior, debida a Smith y Waterman. La principal diferencia es que no se admiten puntajes negativos, porque un buen alineamiento local rara vez generar gaps o sustituciones:

$$F(i,j) = max \begin{cases} 0 \\ F(i-1,j-1) + \sigma(s_i, t_i) \\ F(i-1,j) - w_i \\ F(i,j-1) - w_j \end{cases}$$

Alineamiento de Múltiples secuencias

El problema consiste en encontrar un alineamiento entre N secuencias que maximice el puntaje total asignado. Hay varias alternativas de solución:

- SP-score: calcula el puntaje sumando los puntajes obtenidos al alinear cada pareja de secuencias en el conjunto. Parte de una suposición incorrecta, que es la independencia estadśitica de las secuencias.
- Extender el algoritmo de Needleman y Wunsch lleva a una solución de complejidad exponencial. En realidad, se sabe que el problema es NP.
- Cálculo progresivo: Se alinean dos secuencias, el resultado de ello con una tercera, y así sucesivamente. Es sensible al orden de alineamiento.

Árboles Filogenéticos

- Son estructuras que representan las relaciones evolutivas verdaderas entre un conjunto de proteinas relacionadas entre sí.
- Se puede construir como un árbol en el que los nodos son secuencias y los arcos representan la distancia ó no similitud entre ellas.
- Hay varios métodos para construir una aproximación a este árbol, pero para usarla en alineamiento múltiple no se necesita construir una aproximación muy exacta.

Protocolo de Feng y Doolittle

- Calcular las distancias entre cada par de secuencias usando para ello el árbol aproximado que sólo indicará el orden en el cual realizar los alineamientos
- Alinear cada par de secuencias, de alineamiento y secuencia o de alineamientos empezando por las hojas del árbol, de modos que los más similares se alinean primero y luego los menos similares.

Problema: al adicionar nuevas secuencias no se modifica el alineamiento de las anteriores. Solución volver a calcular el orden para el nuevo conjunto y volver a alinear todo.

Perfiles

Dado un alineamiento múltiple, uno puede derivar la probabilidad de que cada aminoácido se encuentre en una posición dada:

- Se cuenta el número de veces que el aminoácido aparece en esa posición y se divide por el número de secuencias disponibles.
- Cuando aparece una nueva secuencia, se puede calcular la probabilidad de sus aminoácidos en sus respectivas columnas y multiplicar esos valores para saber la probabilidad de que la cadena encaje en el perfil.

Modelos Ocultos de Markov

Este modelo probabilístico permite responder las siguientes preguntas:

1. Dado un conjunto de secuencias de aminoácidos homólogas, cuales son las probabilidades de transición y emisión en los estados que mejor describen esta familia?
2. Dada una nueva secuencia de aminoácidos, cual es la probabilidad de que la secuencia haya sido generada por el HMM de una familia de proteinas?
3. Dada una secuencia de aminoácidos, cual es la secuencia de pasos que con mayor probabilidad hace encajar la secuencia en la familia? (un paso es: asociar dos aminoácidos, hacer una inserción o hacer un borrado)

Búsqueda en una Base de Datos

Todas las secuencias de proteinas conocidas y sus anotaciones funcionales se van almacenando en bases de datos biológicas. Cuando se obtiene una nueva proteina, ella se puede intentar alinear con todas las de la base de datos con el fin de encontrar eventuales nuevas relaciones evolutivas.

Desde el punto de vista computacional, esto significa alinear una consulta con una gran cantidad de secuencias y ordernarlas de acuerdo al puntaje de las alineaciones. El método debe ser rápido y escalable.

Búsqueda en una Base de Datos

Desde el punto de vista biológico, el problema es mucho más complejo:

- Si se desea inferir la estructura de la secuencia consultada, se deben obtener alineamientos realistas con proteinas homólogas.
- Si se desea inferir la función, es necesario distinguir entre cadenas parálogas y ortólogas, distinguiendo qué tanto del material de una proteina ortóloga puede transferirse a la proteina de consulta y también hay que considerar si las anotaciones funcionales de la proteina en la base de datos realmente pueden transferirse a la proteina consultada.

El problema de la búsqueda en bases de datos suele resolverse mediante heurísticos.

Heurísticos de Búsqueda en Bases de Datos

- FASTA (Fast-All): realiza un enfoque multipasos. Primero encunetra secuencias cortas exactas compartidas, luego extrae secuencias con un alto número de secuencias cortas exactas compartidas que pueden ser parte del mismo alineamiento, finalmente alinea las secuencias seleccionadas con la secuencia de consulta haciendo programación dinámica.
- Blast: es otro paquete buscador. Parte de secuencias cortas exactas compartidas que luego intenta extender en ambas direcciones hasta que obtenga un puntaje máximo.

Heurísticos de Búsqueda en Bases de Datos

- Psi-Blast: primero hace la búsqueda al estilo de Blast, recolecta y alinea las secuencias que podrían ser homólogas de la proteina consultada y las usa para construir un perfil. Luego usa el perfil para calcular la probabilidad de que una nueva secuencia pertenezca al alineamiento. Posteriormente usa el perfil para encontrar nuevas secuencias en la base de dato que lo cumplan, las adiciona al grupo, recalcula el perfil y así sucesivamente hasta que no pueda adicionar nada nuevo.