Seminario Bioinformática: The ten most wanted solutions in protein bioinformatics. Problem 2

Jorge Hernán Victoria Moreno.

Grupo DESTINO
Pontificia Universidad Javeriana Cali

Periodo 2008-2

Introducción

- Las proteínas tienen características o propiedades que pueden ser observadas desde su secuencia.
  - Elementos estructurales de las proteínas.
  - Sitios funcionales.
  - Compartimientos celulares donde reside.
  - Sitios que pueden ser modificados. Etapa "Posttranslacional".

Especificación del problema

- Predecir características de una secuencia de proteínas.
- Objetivos:
  - Extraer reglas de conjuntos de proteínas conocidas que compartan una característica común y aplicarlas a un conjunto de casos desconocidos.
  - Inferir una o más reglas de un conjunto de proteínas de entrenamiento que compartan una propiedad dada.
- Si la regla es lo suficientemente general, puede ser usada para predecir la presencia de la propiedad analizada en otras proteínas.

Como abordar el problema

- Patrones Deterministas
  - Presencia de conjuntos de aminoácidos.
- Metodos Estocásticos
  - Dada una secuencia o una subsecuencia, cuál es la probabilidad de que pertenezca a un conjunto de secuencias de entrenamiento positivas.

Patrones Deterministas

- Un patrón de una secuencia se asocia a una propiedad o característica de la proteina.
- Ejemplo:
  - Proteasa: Enzima que produce el Virus de la Hepatitis C (VHC) o el Virus de Inmonudeficiencia Humana (VIH).
  - Usa una molécula de Agua (Hidrolasa)
    A-B + H2O $\rightarrow$ A-OH + B-H
  - Proteína NS3 del VHC: Es una Proteasa en donde el aminoácido que cataliza se llama Serine.
  - Patrón de las proteasas:
    [D,E] - S - G - [G,S]
    D = Aspartic Acid.
    E = Glutamic Acid.
    S = Serine.
    G = Glycine.

Patrones Deterministas(2)

- Un patrón se puede definir como.

$$P = p_1, ..., p_n, p_i \in \Sigma \quad (1)$$

- $\sum$ = Alfabeto de los 20 aminoácidos.
- Ejemplo:
  D - X(1,4) - [L,I] - X - [D,E]
- Entonces se redefine un patrón como:

$$P = p_1, ..., p_n, p_i \in \Sigma \cup X, dondeX = \{x(n_1, n_2) \in N\} \quad (2)$$

  y x es cualquier simbolo.
- { L,I } = Cualquier simbolo menos L o I.

Para Obtener los patrones...

- Primera forma.
  - Se enumera un conjunto de posibles patrones.
  - Se calcula que tan bien lo patrones se ajustan a las proteínas con base a una función fitness definida previamente.
  - Seleccionar los patrones de mayor fitness.
  - Combinarlos entre si para sacar patrones mas óptimos.
- Segunda forma.
  - Hacer alineamiento de secuencias. (Problema 1).
- IMPORTANTE: Se sabe previamente que las proteinas tienen una propiedad común.

Patrones Estocásticos

- La probabilidad de que un aminoácido se encuentre en una determinada posición.
- Perfiles:
  - Dado un alineamiento multiple, calcular la probabilidad de que cada aminoácido se encuentre en una posición dada.
  - Cuantas veces aparece un aminoácido en determinada posición dividido entre el número de secuencias.
- Modelos Ocultos de Markov.
- Redes neuronales Artificiales.
- Perfiles + redes neuronales artificiales.

Especificación y sensibilidad de una función de predicción

- Como saber si el método de predicción que se está utilizando converge a resultados satisfactorios.
- Al evaluar muestras de test:
  - F_+ Número de casos identificados positivos.
  - F_- Número de casos identificados negativos.
  - F_n Número de casos falsos negativos.
  - F_p Número de casos falsos positivos.
- Coeficiente de correlación.