Algoritmos de Estimacion de Distribuciones en Problemas Bioinformáticos

Jose A. Lozano

Intelligent Systems Group

Universidad del País Vasco

http://www.sc.ehu.es/isg/

Composición del Grupo

- 4 profesores doctores
- 2 profesores no doctores
- 8 estudiantes de doctorado
- profesores y estudiantes visitantes

Actividad Investigadora

- Modelos Gráficos Probabilísticos: Redes Bayesianas, Gaussianas, etc.
- Clasificación. Reconocimiento de Patrones
- Computación Evolutiva
- Aplicaciones

Gene selection in microarrays

- Dado un problema de clasificación a partir de datos provenientes de microarrays, hallar el conjunto de genes que ”mejor” predicen la clase
- Tres aproximaciones clásicas: filter, wrapper, filter-wrapper
- En nuestro caso:
  - Técnicas: Naive-Bayes + UMDA, con leaving-one-out
  - Conjuntos de datos: Colon (Alon et al. 1999) (62 casos, 2000 genes), Leucemia (Golub et al. 1999) (72 casos, 7129 genes).

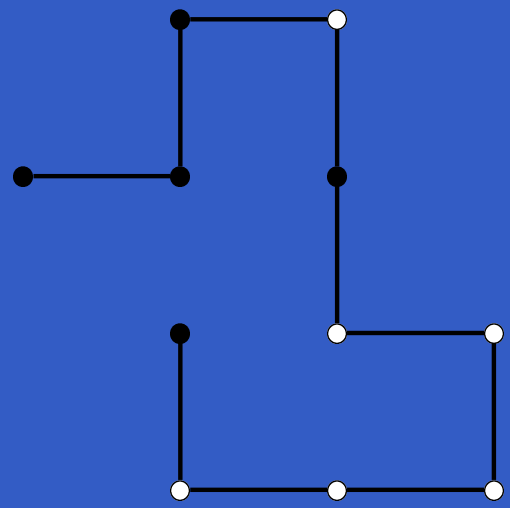
Gene selection in microarrays

- Problema del lupus eritematoso sistémico (LES) y el síndrome antifosfolipídico (SAF)
- Microarray: 40 casos, 20 LES, 10 SAF, 10 sanos, 8808 genes
- Tres discretizaciones diferentes: equal width, equal frequency, Fayyad & Irani
- Aplicar un CFS
- Intersección de los CFS
- Búsqueda de genes similares

Protein folding problem

- Se trata de obtener la estructura de plegado de las proteinas: esta estructura determina sus propiedades
- Es un probema muy complejo, por lo tanto se utilizan modelos simplificados
- HP-model: dos tipos de residuos, hidrophobic (H) and hidrophilic or polar (P)
- La función objetivo depende de las interacciones entre residuos no consecutivos
- Dos tipos de modelos 2D y 3D

Protein folding problem



Solución óptima para la secuencia
HHHPHPPPPPH

Protein folding problem

- Codificación: direccional
- Aproximación basada en EDAs: tres modelos diferentes:

$$p_{MK}(\mathbf{x}) = p(x_1, \ldots, x_{k+1}) \prod_{i=k+2}^{n} p(x_i \mid x_{i-1}, \ldots, x_{i-k})$$

$$p_{Tree}(\mathbf{x}) = \prod_{i=1}^{n} p(x_i | pa(x_i))$$

$$p_{MT}(\mathbf{x}) = \sum_{j=1}^{m} \lambda_j p^j_{Tree}(\mathbf{x})$$

Multiple Sequence Alignment Problem

- Dado un conjunto de secuencias, se trata de alinearlas globalmente de la mejor forma posible
- Importante en muchos campos de la bioinformática: detección de estructura de las proteinas, detección de genes, etc.
- Ámpliamente tratado en la bibliografía: ClustalW, T-Coffee, SAGA, ...

Multiple Sequence Alignment Problem

Ejemplo:

S_1 :	M	P	Q	I	L	L	L	V		M	P	Q	I	L	L	L	V
S_2 :	M	L	R	L	L					M	L	R	–	L	L	–	–
S_3 :	M	K	I	L	L	L				M	–	K	I	L	L	L	–

Multiple Sequence Alignment Problem

- Difícil encontrar funciones objetivo que cumplan criterios biológicos
- La función objetivo más comúnmente utilizada es la: "weighted sum-of-pairs with affine gap penalities"
- Para funciones objetivo concretas existen algoritmos de programación dinámica que resuelven el problema de forma óptima

RNA folding problem

- El plegamiento de RNA determina parcialmente las funcionalidades del mismo: producción de proteinas
- El RNA esá compuesto de cuatro nucleótidos: A, C, G, U
- El plegamiento se produce mediante la unión de pares de bases canónicas: GC, AU, GU

RNA folding problem

- Se computan todas las posibles helices que se pueden formar H
- Se realiza una búsqueda de un subconjunto S de H de tal forma que cada hélice no comparta nucleótidos
- Minimizar la energia de la estructura:
  $E(GC) = E(CG) = -3$
  $E(AU) = E(UA) = -2$
  $E(GU) = E(UG) = -1$

Problemas diversos

- Clustering de genes en microarrays
- Redes de regulación genéticas
- Problema de "splicing"

Comentarios finales

- ¿Qué papel jugamos los informáticos y cuál juegan los biólogos?
- ¿Qué implicaciones tiene introducirnos en el campo de la bioinformática?
- Apuesta fuerte