Sistemas Inteligentes para el análisis de datos en biología

Universidad Pontificia de Salamanca

Manuel Martín-Merino

Email: mmartinmac@upsa.es

Contenido

1. Introducción a la bioinformática
2. Técnicas avanzadas de clasificación: Aplicaciones en bioinformática
3. Técnicas no supervisadas para el análisis de datos en bioinformática

Introducción a la bioinformática

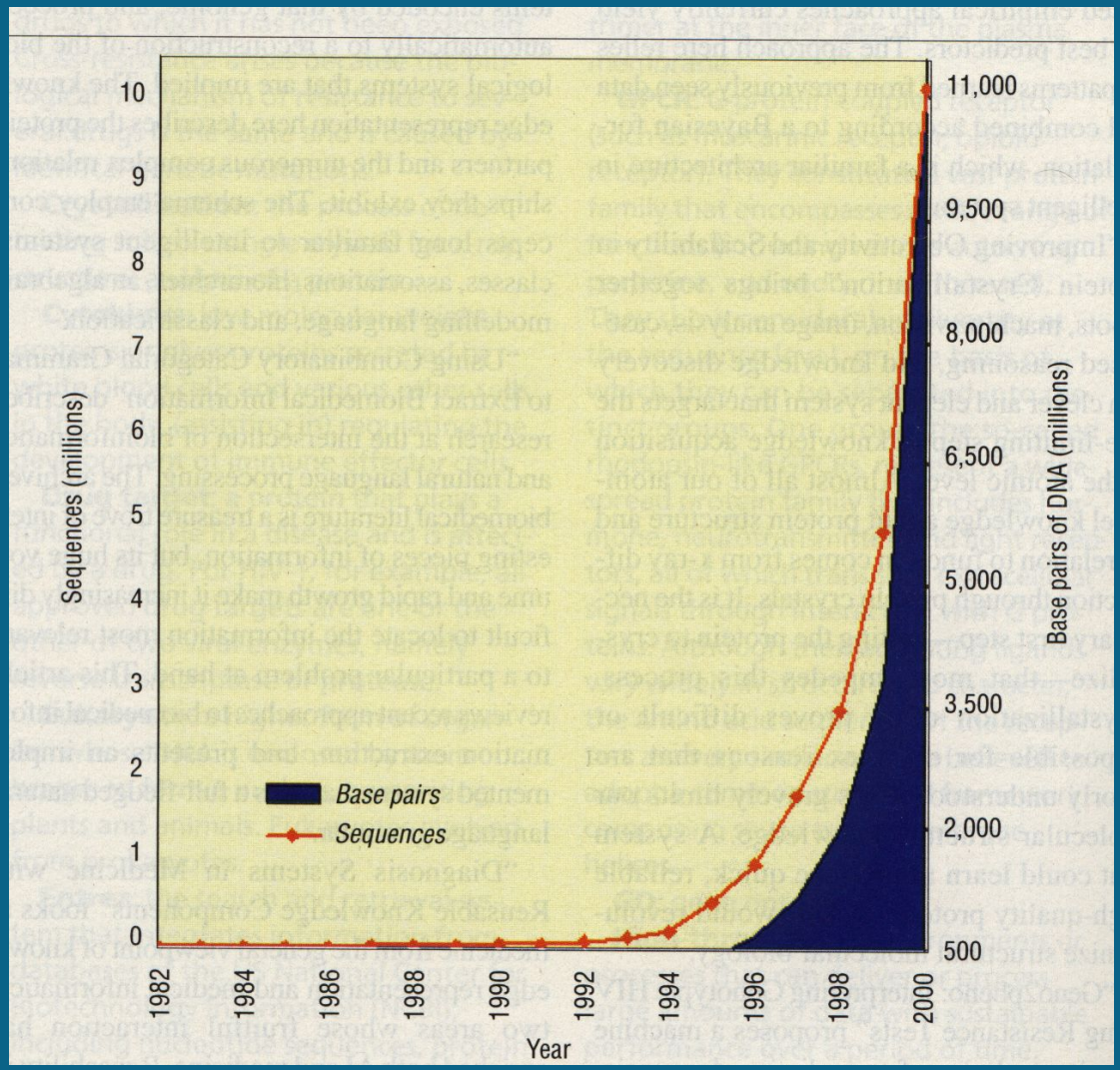
Contenido

1. Motivación
2. Conceptos básicos de biología molecular
3. Retos para la IA. en biología
4. Datos generados por la biología molecular
5. Discusión: conclusiones

Motivación (I)

- La primera mitad del XX ha supuesto un gran avance para la física, las matemáticas y la química. La segunda mitad ha sido protagonizada por la biología y la informática.
- La biología ha generado cantidades masivas de datos debido a una nueva generación de tecnologías.
- Los biólogos no pueden analizar dichos datos debido a su complejidad y cantidad. Se necesitan técnicas inteligentes de análisis.

Motivación (II)



ADN

Es una cadena de moléculas compuestas por nucleótidos enlazados linealmente. Existen cuatro tipos de nucleótidos A(adenine), C(cytosine), G(guanine) y T(thynine). Cadenas moleculares de sólo algunos nucleótidos se llaman oligonucleótidos.

Los nucleótidos A-T y G-C son complementarios. Pueden formar enlaces de hidrógeno estables.

ADN complementario (cADN) es una secuencia complementaria de ADN leída en dirección inversa.

Cadenas complementarias pueden enlazarse fuertemente formando una estructura de doble hélice. Este proceso se llama hibridación.

mARN

Es una cadena de nucleótidos similar al ADN aunque con propiedades químicas ligeramente diferentes. Están compuestas por una única cadena y pueden formar estructuras tridimensionales.

Gen es una subcadena del genoma responsable de la producción de moléculas de ARN.
En el proceso de expresión de los genes el ARN es sintetizado como una secuencia complementaria de una parte del gen.

mARN Moléculas de ARN que controlan la producción de proteínas y sus características.

Proteinas (I)

Son polímeros compuestos por aminoácidos. Hay 20 tipos diferentes. Su secuencia constituye la estructura primaria proteínas.

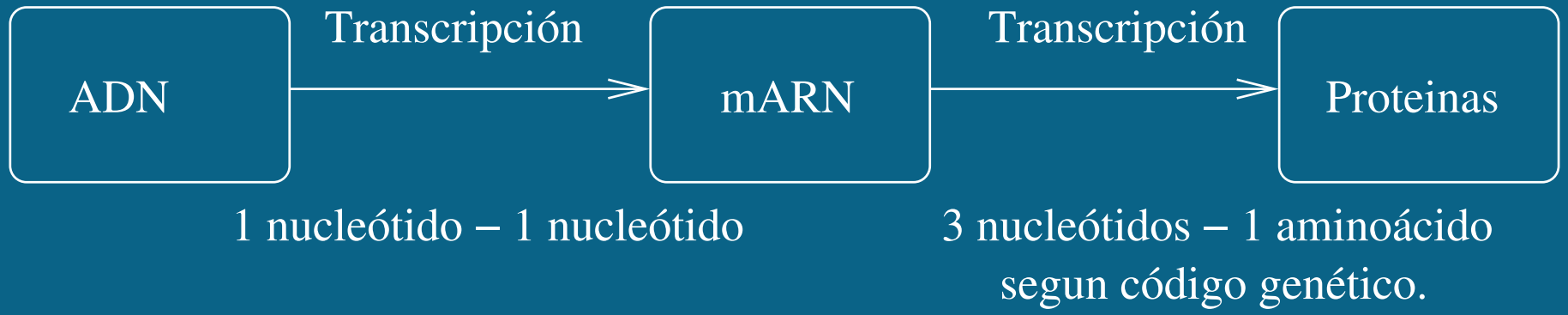
Cada proteina se pliega en una estructura tridimensional $\rightarrow$ estructura terciaria, determinada fundamentalmente por la primaria.

Las subestructuras tridimensionales más importantes son las hélices α y los hilos $\beta \rightarrow$ estructura secundaria de las proteínas.

La estructura de la proteina plegada permite saber qué aminoácidos hay en su superficie y por tanto sus funciones

Proteinas (II)

Expresión de genes en proteinas



- Transcripción: Se genera una copia del ARN complementaria al ADN, mARN.
- Transcripción: Los ribosomas sintetizan proteínas especificadas por grupos de 3 nucleótidos consecutivos.

Contenido

1. Motivación
2. Conceptos básicos de biología molecular
3. Retos para la IA. en biología
4. Datos generados por la biología molecular
5. Discusión: conclusiones

Identificación estructura 3D proteinas

El método más utilizado se basa en los siguientes pasos [5]:

- Purificación proteina.
- Cristalización: Fijar los parámetros para una cristalización correcta es complejo.
- Difracción por rayos X
- Analisis de la imagen

Problema: Clasificar las imágenes del cristal, y visualizar loas resultados de los experimentos de cristalización proteinas.

Predicción estructura secundaria proteinas (I)

La predicción de la estructura de las proteinas mediante rayos X o resonancia magnética es un proceso costoso y laborioso.

En los últimos años se han utilizado redes neuronales para predecir la estructura secundaria y tridimensional de las proteinas [1] con los mejores resultados.

- A partir de la cadena de aminoácidos se predice la probabilidad de aparición de las estructuras hélices α e hilos β. Se trata de un problema de clasificación.
- Se identifica la topología de dichas estructuras.
- Se predice la estructura tridimensional de las proteinas.

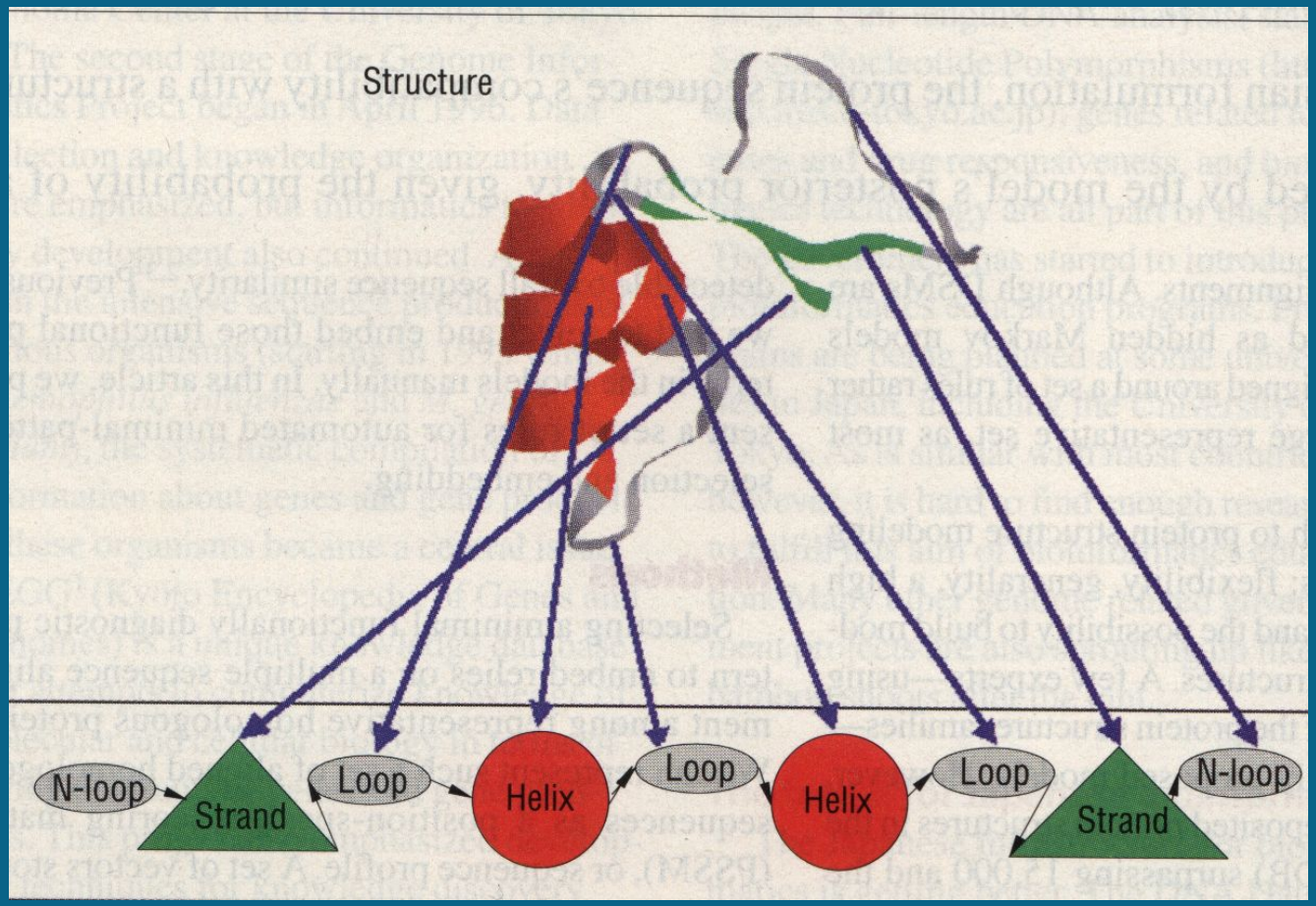


Figura 1: Elementos básicos de la estructura secundaria proteina.

Microarrays ADN (I)

- El cDNA microarray [6, 8] está compuesto por una malla de puntos sobre una lámina de cristal. Cada punto contiene una secuencia de DNA llamada *probe*, específica de un determinado gen.
- En cada experimento, dos muestras (una de control) son transcritas a partir del mARN de las células. Las muestras son marcadas con (Cy3 y Cy5) que tienen longitudes de onda en el verde y rojo.
- Las muestras de cDNA son hibridadas con el microarray. Si una secuencia de cDNA es complementaria de la secuencia de DNA asignada a un probe, se hibridará y será detectada por la fluorescencia.

- La intensidad luminosa de cada punto es leida por un escaner laser.
- La imagen se preprocesa, elimina ruido, normaliza, identifican los puntos etc.

Microarrays ADN (II)

La medida de la expresión de los genes (mARN) permitirá:

- Identificar genes que tienen funciones similares. Problema de cluster [4].
- Identificación de ciertas enfermedades como el cancer. Problema de clasificación [3].

Anotación función genes

Cuando se identifica una secuencia nueva de nucleótidos, es necesario saber si codifica una proteina y además:

- La familia de la proteina.
- Enfermedades con las que está relacionada.
- Estructura tridimensional ....

Necesitamos algoritmos que organicen automáticamente la información contenida en las bases de proteinas (cluster y clasificación) [7].
Sistemas inteligentes que recuperen información relacionada con la familia de proteinas identificada [7].

Resistencia virus a drogas

Estudio de la resistencia del virus VIH a diversas drogas en función de fragmentos de ADN [2].

- Se secuencian las partes relevantes del gen que codifica el objetivo viral de la droga.
- Se clasifica el virus como resistente o no resisetente a la droga.

Problema: Clasificación de la secuencia de ADN del virus en función de experimentos realizados anteriormente.

Ontologías en medlines

Existe una gran cantidad de artículos, e información relacionada con diferentes aspectos de la genómica.

Problema:

- Crear ontologías que identifique un vocabulario común y estructurado que permita clasificar la información disponible.
- Sistemas inteligentes que permitan extraer relaciones gen-gen, proteina-gen, gen-droga etc.
- Sistemas que permitan manejar datos heterogeneos.

Datos generados por la biología molecular

Los datos generados por la biología molecular tienen características peculiares.

Tipo datos	**Representación**
Secuencias	
• ADN	cadena de nucleótidos $\{A,C,G,T\}$
• mRNA	cadena de ribonucleótidos $\{A,C,G,U\}$
• Proteínas	cadena de aminoácidos (20)
Expresión/ localización	
• Expresión de genes	vectores, matrices reales
• Expresión de proteinas	vectores, matrices reales
• Localización proteínas	categórica
Textos científicos	
• Textos: Artículos, abstracts etc.	datos textuales (ingles)

Conclusiones: Discusión

La biología dispone de herramientas experimentales muy potentes que generan gran cantidad de datos.

Necesitamos técnicas capaces de extraer automáticamente la información contenida en dichos datos.

- Técnicas de clasificación.
- Técnicas de cluster y análisis no supervisado de datos multivariantes.

Referencias

[1] P. Baldi and G. Pollastri. A machine learning strategy for protein analysis. *IEEE Intelligent Systems*, 17(2):28–35, March/April 2002.

[2] N. Beerenwinkel, T. Lengauer, J. Selbig, B. Schmidt, H. Walter, K. Korn, R. Kaiser, and D. Hoffman. Geno2pheno: Interpreting genotypic hiv drug resistance tests. *IEEE Intelligent Systems*, 16(6):35–41, November/December 2001.

[3] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46:389–422, 2002.

[4] D. Jiang, C. Tang, and A. Zhang. Cluster analysis for gene

expression data: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 16(11), November 2004.

[5] I. Jurisica, P. Rogers, J. I. Glasgow, R. J. Collins, R. Wolfley, J. R. Luft, and T. DeTitta. Improving objectivity and scalability in protein crystallization. *IEEE Intelligent Systems*, 16(6):26–34, November/December 2001.

[6] A. W.-C. Liew, H. Yan, and M. Yan. Pattern recognition techniques for the emerging field of bioinformatics: A review. *Pattern Recognition*, 38:2055–2073, 2005.

[7] J. Reich, A. Mitchell, C. Goble, and T. Attwood. Toward more intelligent annotation tools: A prototype. *IEEE Intelligent Systems*, 16(6):42–51, November/December 2001.

[8] T. Speed. *Statistical Analysis of Gene Expression Microarray Data*. Chapman & Hall/CRC Press, New York, 2003.