MICROARRAYS

Introducción

La Bioinformática tradicionalmente se ha encargado del análisis computacional de grandes conjuntos de datos biológicos moleculares, hasta 1995 el interés estaba centrado en las estructuras de las proteínas, volcándose a partir de ésta fecha en la secuenciación del genoma, y ha sido a partir de 1998 cuando se ha producido una gran cantidad de datos proporcionados por las nuevas tecnologías, que permiten experimentos en particular sobre la expresión genética, que sólo están limitados por el coste y la imaginación de los investigadores.

Una de las tecnologías que permite obtener información sobre la expresión del genoma en un sentido amplio son los microarrays de ADN complementario.

Una vez que se dispone totalmente o parcialmente de la secuencias del genoma de un ser vivo, el siguiente paso es entender su funcionamiento, y dentro de éste objetivo general se plantean objetivos más concretos entre los que se encuentran:

* Cual es la función de los distintos genes, y en qué procesos participan.
* Cómo se regulan los genes, cómo interaccionan los genes con sus productos, cuáles son las redes de interacción.
* Cómo los niveles de expresión se diferencian en los distintos tipos de células y en diferentes estados, cómo cambia el nivel de expresión cuando aparece una determinada enfermedad, o las células están sometidas a un tratamiento específico.

Para alcanzar éstos objetivos, uno de los primeros aspectos que se debe conocer, es la cantidad de gen transcrito en distintos tejidos bajo condiciones variadas. Aunque el mARN no es el último producto de un gen, la transcripción es el primer paso en la regulación de los genes y los niveles de transcripción son necesarios para entender las redes de regulación de los genes. Además la medición del nivel de mARN es muchísimo más económica que medir el nivel de proteínas, y obviamente está relacionado.

La capacidad de manejar la expresión de los genes al nivel de transcripción, es posible gracias a la tecnología de los microarrays de cADN.

Un microarray es un portaobjetos de vidrio o silicio, sobre el cual hebras simples de moléculas de ADN, se asocian a localizaciones fijas llamadas spots. Puede haber decenas de miles de spots sobre un microarray, cada una de ellas relacionada con un gen, aspecto que sin esta tecnología estaría limitada a sólo 5 o 6 genes simultáneos. Los microarrays explotan la posibilidad de la unión preferencial de secuencias complementarias de hebras simples de ácidos nucleicos.

Una de las formas más populares que se usan, es comparar la cantidad de mARN en dos muestras (ejemplo y control). Los mARN de ambas muestras se convierten en cADN (ADN donde se han eliminado los intrones, en teoría regiones que no codifican información), que es más estable, a continuación se tintan con etiquetas fluorescentes, p.e.j. una roja y otra verde, posteriormente se enjuagan en el microarray, de forma que las secuencias de genes que hay en el substrato hibridizan (se unen) a las secuencias complementarias que hay en los spots.

Para medir la cantidad relativa de ARN hibridizado, el array se excita mediante láser, de forma que los spots tenderán a ser rojos o verdes, dependiendo del tipo de muestra que más se haya hibridizado, amarillo en el caso de que las cantidades sean similares, y negro si no ha ocurrido ninguna hibridación. A partir de los colores, se pueden inferir el nivel de expresión en las dos muestras.

Midiendo los niveles de expresión de los genes de un organismo bajo condiciones distintas, en diferentes estados de desarrollo, y en diferentes tejidos, podemos construir los perfiles de expresión de los genes, que caracterizan el comportamiento dinámico de cada gen en el genoma.

Los perfiles de expresión aludidos vendrán representados por una matriz, donde las filas representan los genes del tejido ejemplo y las columnas las distintas situaciones (tipos de tejido, estados de desarrollo, tratamientos etc). Cada celdilla contendrá un número que caracteriza el nivel de expresión del gen concreto en un estado determinado (equivalente a la cantidad de proteína que dicho gen sintetizaría).

En determinados casos la matriz de perfiles contendrá la cantidad relativa de gen expresado de la muestrea ejemplo con respecto a la muestra de control, y en otros casos, contendrá la cantidad absoluta, dependiendo del caso, el tratamiento será distinto. Lo más normal, es que los datos representen cantidades relativas.

En el futuro con la construcción de bases de datos con éste tipo de matrices, ayudará a entender la regulación de los genes, los caminos metabólicos y de señales, los mecanismos genéticos de las enfermedades, las respuestas al tratamiento de drogas, etc. Así, si la sobrexpresión de un grupo de genes aparece en un tejido canceroso, es decir que determinado grupo de genes produce mayor cantidad de mARN en el tejido cancerígeno (muestra) que en el tejido normal (control), podremos investigar qué otras condiciones afectan a éstos genes, qué otros genes tienen perfiles de expresión similares, las características de dichos genes, la incidencia que tiene determinado tipo de droga sobre la enfermedad etc.

En primer lugar, los datos necesitan ser preprocesados, siendo las etapas usuales en el preprocesamiento de los datos:

1.- Cambio de escala
2.- Eliminación de genes duplicados
3.- Manejo de datos perdidos
4.- Eliminación de datos planos
5.- Normalización y estandarización.

Manejo de datos perdidos
En algunos casos puede que no aparezcan datos en determinados ensayos (columnas) por diversas razones, en algunos casos los datos perdidos estarán representados por espacios en blanco y en otros por caracteres no numéricos. Aunque es un problema abierto se utilizan algunas de las siguientes alternativas para su resolución:

La alternativa más drástica es la de la eliminación de los genes de los que no se disponga suficiente información, para ello se fija un % por debajo del cual el gen será eliminado, el % relaciona el número de datos que debería tener el gen, con el número de datos que realmente conocemos.

Para sustituir los datos que aún faltan, existen varias alternativas: Sustituir los datos perdidos por: I) por 1 en el caso de datos relativos, o por cero en el caso de valores absolutos, ya que esto supone una solución balanceada con respecto a la sobre-expresión y la sobre-represión. II) por la media de la fila. III) por la mediana de la fila IV) Usar el método de los k vecinos más cercanos (siendo k un parámetro escogido por el usuario) utilizando para ello cualquier medida de distancia (correlación de Pearson, distancia euclídea o minimización de la varianza) aunque basta con la euclídea, ya que el problema que podía aparecer con los outliers (valores aislados), se ve reducido si antes se ha realizado una transformación logarítmica. Este método es el más utilizado, siempre que se disponga de bastantes genes completos, y el número de datos perdidos en cada gen sea pequeño, normalmente se necesitan disponer de 1000 datos completos, para poder estimar 5 datos perdidos.

PRACTICA-3

La presente práctica está encaminada a clasificar los datos de Microarray, con objeto de realizar un pronóstico de supervivencia de pacientes con un tipo de cáncer de linfoma DLBCL, utilizando redes neuronales feedforward, y la detección de genes que tengan mayor influencia en la enfermedad, información que posteriormente será utilizada por el experto para analizar más a fondo la enfermedad.

La práctica constará de dos partes: I) Clasificación semi-guiada. II) Clasificación libre. En ambos casos se proporcionarán dos tipos de datos:

I.-1 La hoja de cálculo en formato Exel "Datos_Microarrays_Cancer" contiene datos de Microarrays de 4026 genes (filas) de 96 pacientes. En la presente práctica solo estamos interesados en los ejemplos de pacientes con cáncer de linfoma DLCL, (40 pacientes) además del identificador de los genes que aparece en la segunda columna. El primer ejemplo empieza en la columna 7 con el encabezamiento DLCL-0042. Los datos se proporcionan preprocesados, salvo la fase de la sustitución de datos perdidos, que deberá ser ejecutada previamente a la clasificación. Utilizar el método de los k vecinos más cercanos (tomar k = 15). La función en Matlab es knnimpute.

I.-2 Un fichero "DLCL_Supervivencia.txt" donde aparece el indicador de supervivencia a lo largo de 10.8 años, para cada uno de los pacientes que se desea considerar.

I) Clasificación semi-guiada

Debido a los dos problemas más destacados en el análisis de los datos inherente al tipo de información proporcionado por los Microarrays (escaso número de ejemplos, y elevado número de variables), se utilizará una aproximación progresiva en la solución del problema, poniendo de manifiesto la potencia de las RN.

Fase a)

Se entrenarán 4 RN, cada una de ellas con conjuntos de entrenamiento y validación (30 ejemplos) y de test (los 10 ejemplos restantes), de forma que los conjuntos de test sean distintos para cada una de las 4 RN.

Cada ejemplo se representará mediante 3 neuronas: La primera neurona indicará el signo del nivel de expresión de los genes: 1 (negativo), 0 (positivo). La segunda y tercera neurona cuantificará el nivel de expresión en valores absolutos: 00 ($0 <= x <= 0.5$), 01 ($0.5 < x <= 2.0$), 11 ($2.0 < x <= 9$)

En esta fase, la capa de entrada tendrá 4026*3 neuronas, una capa intermedia de 100 neuronas, y una neurona en la capa de salida, que indica la supervivencia mediante 1 ó 0.

Las mejores RNs se obtendrán mediante el % de acierto medio sobre los conjuntos de test de los 10+10+10+10 = 40 elementos.

Resumiendo:

I.- Se diseña una RN, se escogen aleatoriamente 30 ejemplos de entrenamiento y validación, y el resto de ejemplos (10), se usan de test. Se halla el % de acierto con los datos de test, teniendo en cuenta que la salida hay que redondearla a uno ó a cero

II.- La misma topología de RN se usa con 30 ejemplos de entrenamiento y validación, y el resto de ejemplos (10), se usan de test. Hay que tener en cuenta que estos 10 ejemplos de test, deben ser diferentes de los escogidos anteriormente. Se halla el % de acierto con los datos de test.

III.- La misma topología de RN se usa con 30 ejemplos de entrenamiento y validación, y el resto de ejemplos (10), se usan de test. Hay que tener en cuenta que estos 10 ejemplos de test, deben ser diferentes de los escogidos anteriormente. Se halla el % de acierto con los datos de test.

IV.- La misma topología de RN se usa con 30 ejemplos de entrenamiento y validación, y el resto de ejemplos (10), se usan de test. Hay que tener en cuenta que estos 10 ejemplos de test, deben ser diferentes de los escogidos anteriormente. Se halla el % de acierto con los datos de test.

Se halla la media de los % de los 4 casos anteriores.

Todo lo anterior debe de repetirse para obtener la mejor topología y el mejor % de acierto.

La RN propuesta en la fase a) es bastante grande, por lo que el entrenamiento, o será muy lento, o necesitará mucha memoria, dependiendo de la función de entrenamiento; en este sentido si el ordenador utilizado carga demasiados programas en memoria, es probable que ni llegue a funcionar. Si alguno tiene dificultades en el sentido indicado, puede pasar a la siguiente fase.

Fase b)
Esta fase será prácticamente idéntica a la fase anterior, salvo la codificación, y tendrá por objeto conjuntamente con la fase c), obtener los genes que más influyen en el pronóstico de supervivencia. En esta fase, cada ejemplo se representará solo por 2 neuronas: (10) para niveles de expresión superiores al control, es decir positivos, y (01) para niveles inferiores, valores negativos. En este caso la RN tendrá 4026*2 neuronas en la capa de entrada, 67 en la capa intermedia y 1 en la capa de salida.
La elección de las mejores 4 RNs es muy importante, ya que tendrá mucha influencia en el resultado final.

Las redes entrenadas que mejor funcionen deben de ser grabadas, ya que serán utilizadas en la fase c)

Fase c)
Con objeto de diferenciar los genes más representativos, se aplicará el siguiente proceso: A partir de la fase anterior, de los donantes clasificados correctamente en los conjuntos de test, se escoge un número "adecuado" de ejemplos que hayan proporcionado los mejores resultados en el test, (alrededor de 12 (Valor orientativo)). Los "mejores resultados" se interpretan, en el sentido de que si la salida para un determinado paciente se esperaba que fuera 1, y la salida ha sido 0.97, será mejor que otro paciente del que se esperaba que la salida fuera un 1, y la obtenida ha sido 0.95

A continuación, tomando como base estos ejemplos diferenciados, trataremos de buscar los genes más representativos, es decir con mayor influencia. Para cada uno de los ejemplos escogidos y la RN entrenada obtenida en el apartado b) asociada al ejemplo, realizamos la siguiente operación:

Para cada gen producimos una perturbación (sustituyendo el valor de la neurona (1 0) por (0.85 0), o bien, la neurona (0 1) por (0 0.85)), a continuación obtendremos la salida utilizando la RN entrenada asociada al ejemplo en el apartado b), y anotaremos la variación producida en la salida de la red con respecto al valor obtenido sin la perturbación. De esta forma estaremos evaluando la importancia de cada gen en la salida de la red.

Ordenamos para cada ejemplo los genes en función de las variaciones (de mayor variación a menor variación), quedándonos con el 25% de los genes que sean responsables de una mayor variación.

Finalmente se escogen aquellos genes que aparezcan en al menos 4 de los (aproximadamente 12) pacientes o ejemplos anteriores.

Como resultado de todas estas operaciones se obtendrán alrededor de 34 genes (valor orientativo), que se supone que son los que tienen más peso en la clasificación.

Fase d)
Con los 34 genes anteriores se entrenan 10 RN, cada una de ellas con 36 conjuntos de entrenamiento y validación por una parte, y 4 conjuntos de test, de forma que todos los ejemplos en algún momento sean considerados de test. En este caso se deben hacer el entrenamiento sin codificar las entradas, es decir (0.34, 0.59). La clasificación debiera de ser "casi" correcta, lo que llevaría a pensar, que la mayor parte de los 34 genes, si no todos, están involucrados en la enfermedad considerada. El % de acierto se computará sobre los $4 + 4 + \ldots + 4 + 4 = 40$ ejemplos.

Hay que hacer notar que diversos experimentos pueden dar diferentes genes diferenciados, aunque con algunos de ellos comunes.

II) Clasificación libre.

Esta parte de la práctica consistirá en aplicar uno o varios métodos de selección de características (de genes), y aplicar la fase d), con objeto de validar los genes escogidos.

A modo de sugerencia, algunos métodos aplicables pueden ser: SFS, la medida de distancia LS_bound, Enfriamiento Simulado, ' genevarfilter', 'genelowvalfilter', 'geneentropyfilter'.

Con respecto a SFS existen algunas versiones muy sofisticadas que tardan más que la original.

Entregar:

Software
1.- El programa fuente para procesar los datos de la Hoja de cálculo, que debe llamarse "Procesar"
2.- El fichero obtenido con los 40 ejemplos, después de aplicar "Procesar".
1.- El programa fuente para la eliminación de datos perdidos, que se deberá llamar "DatosPerdidos".
2.- El fichero preprocesado con los 40 ejemplos.
3.- El resto de programas debidamente documentados.
4.- Las redes entrenadas de la fase d)

Documentación en formato Word o PDF
1.- Presentación detallada de todos los procesos seguidos para las diferentes clasificaciones, así como los resultados en cada caso **(% de aciertos sobre los 40 casos** (imprescindible para superar la práctica), número e identificador (segunda columna) de los genes diferenciados, etc.).
2.- Documentación **totalmente detallada** del método de selección de características seguido.

La Fecha de entrega de ésta práctica y la siguiente será hasta el 6 de Junio de 2014