MASSA: Análisis masivo de asociación mediante 'simulated annealing'

Miguel Pérez-Enciso[1,2]

[1] Institut Català de Recerca i Estudis Avançats, Lluis Companys 23, Barcelona 08010
[2] Departament de Ciència Animal i dels Aliments, Facultat de Veterinària, Universitat Autònoma de Barcelona, Bellaterra, 08193

INTRODUCCION

Uno de los temas de investigación más activos en la actualidad en Genética, especialmente humana, consiste en la identificación de las mutaciones causales implicadas en la variabilidad de los caracteres cuantitativos, llamados también 'complejos'. Para ello, necesitamos de métodos estadísticos adecuados que extraigan el máximo de información de la cantidad de datos de polimorfismos que se están generando. En la actualidad, ya se han propuesto una serie de métodos, más o menos potentes y con más o menos restricciones, que intentan resolver esta cuestión. La mayoría de estos métodos están pensados para caracteres binarios (Nelson et al., 2001; Pociot et al., 2004; Ritchie et al., 2001).

En este trabajo, presentamos los fundamentos teóricos de un nuevo método que ofrece gran flexibilidad en la modelización, sin prácticamente limitaciones en el número de individuos o de marcadores. El método presenta propiedades óptimas para caracteres con distribución continua, pero puede utilizarse también con caracteres binarios. El programa MASSA implementa este método y estará disponible próximamente.

MATERIAL Y METODOS

Teoría

Supóngase que n individuos han sido genotipados para un número s de marcadores (bialélicos o no), s puede ser (mucho) más grande que n, pero el número máximo de haplotipos nunca excederá 2n. Supongamos que los haplotipos son conocidos o se pueden estimar. Un modelo general explicativo de los datos es el modelo mixto

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zg} + \mathbf{e} = \mathbf{X}\ \mathbf{b} + \mathbf{\Sigma}\ \mathbf{Z_j g_j} + \mathbf{e}, \qquad (1)$$

donde **y** contiene lo fenotipos, **X** y **Z** son matrices de incidencia, **b** es un vector con efectos fijos, y **g** contiene los efectos genéticos que pueden descomponerse en una serie de loci. En lo sucesivo, entenderemos por locus una serie de marcadores que actúan conjuntamente, es decir, si hay diversas posiciones que presentan epistasia las incluimos dentro del mismo 'locus'. Por tanto, la aditividad entre los distintos loci está garantizada por construcción del modelo. El objetivo es encontrar el modelo que mejor explica los datos con un compromiso entre mejor ajuste y mínimo número de parámetros (loci). La varianza de los valores genéticos es

$$\text{Var}(\mathbf{g}) = \sum_{j=1}^{Q} \mathbf{Z_j G_j Z_j'} \sigma_j^2$$

donde $\mathbf{Z_j G_j Z_j'}$ es la matriz de varianzas covarianzas entre dos individuos para el locus en cuestión. Dado que pensamos en loci causales, la matriz **G** es diagonal, dos haplotipos o son idénticos o no lo son. Al tratar los haplotipos como aleatorios,

disminuimos el problema de un exceso de alelos cuando los haplotipos incluyan muchos marcadores.

Un modelo viene determinado inequívocamente por un vector de tamaño s; en cada posición, 0 por convención indica que ese marcador es neutro, y un número j, 1=j=k indica que forma parte del locus j. Es obvio que es imposible enumerar el número total de modelos aun con un número de marcadores relativamente pequeño. Por tanto, es necesario recurrir a técnicas de maximización estocásticas. Aquí hemos empleado el simulated annealing (Kirkpatrick et al., 1983). En esta técnica, se acepta una nueva solución con probabilidad

$$P_{accept} = Min\{1, exp[-(A_{new}-A_c)/T] \},$$

donde Anew (Ac) es el valor de la función asociado a la nueva solución o a la actual y T es la 'temperatura' actual del sistema, ésta se reduce por un factor (0.9 en nuestro caso) cada cierto número de iteraciones (500-1000). Como función a minimizar utilizamos el Criterio de Información de Akaike (AIC) o el Criterio de Información Bayesiano (BIC), que vienen dados por AIC = -2 log l + 2*k, y BIC = -2 logl + log(n)*k, donde k es el número de loci y n, el de observaciones. La verosimulitud se calculó mediante el programa Qxpak (Pérez-Enciso and Misztal, 2004). Es importante recordar que estamos comparando modelos que no son necesariamente jerárquicos, por lo que un test de cociente de verosimilitudes no es válido. Para el funcionamiento del SA, necesitamos criterios para proponer nuevas soluciones a testar en cada iteración. Utilizamos una serie de movimientos:

- *In*: se inserta un nuevo locus.
- *Out*: se elimina un locus.
- *Join*: se fusionan dos loci (se crea epistasia) .
- *Split*: un locus se divide en dos.
- *Glue*: dos loci consecutivos se fusionan, así como las posiciones intermedias.
- *Grow*: se añaden maracdores nuevos a un locus existente.
- *Shave*: se eliminan marcadores de un locus.
- *Shift*: se corre la posición del locus.

La probabilidad de que ocurra cada uno de los movimientos varía. Por ejemplo, la probabilidad de incrementar el número de loci aumenta si el número existente es bajo y viceversa, la probabilidad de que un locus sea eliminado es mayor cuanto menor sea la heredabilidad asociada.

Simulación

Se utilizaron los genotipos simulados correspondientes a 5 Mb presentados en (Lin et al., 2004), y que consisten en diversas poblaciones de entre 1500 y 3000 individuos, con 500 SNPs genotipados por individuo. Elejimos diversas formas para simular el fenotipo asociado a cada individuo. En todos los casos, se eligieron dos SNPs al azar entre las posiciones 1 y 20, y 480 y 500, uno por cada intervalo. Las situaciones fueron:

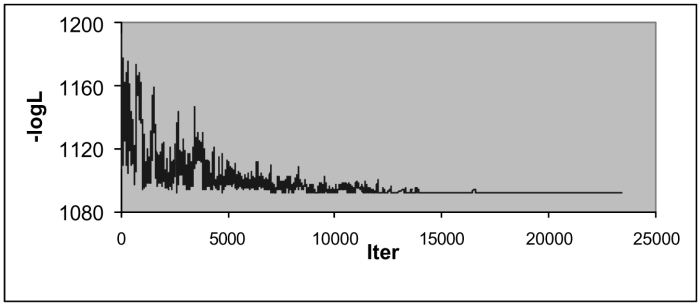
- Fenotipo continuo, no interacción: se simularon dos loci con efecto aditivo.
- Fenotipo continuo, epistasia decreciente: se simularon dos loci con epistasia decreciente (Crow and Kimura, 1970), tabla 4.1.3a.
- Fenotipo continuo, no interacción: se simularon dos loci con epistasia adaptativa (Carlborg and Haley, 2004).

- Fenotipo binario, no interacción. Se simuló un fenotipo como en el primer caso y, aplicando un umbral, se clasificaron los fenotipos en 1 y 2 de forma que el 50% fueran de cada tipo, aproximadamente.

Los valores asignados a cada genotipo se hicieron de forma que fueran el mismo a cada SNP y que la heredabilidad del carácter fuera 0.3 en todos los casos.

RESULTADOS Y DISCUSION

Los resultados iniciales son muy prometedores y se presentarán en la comunicación. A pesar de que el método está diseñado para caracteres continuos, en todos los casos analizados se identificaron los SNPs causales. La Figura 1 muestra cómo el AIC disminuye a medida que el programa avanza. También se discutirán las ventajas y limitaciones de este método.



Agradecimientos

Gracias a Jesús Fernández por educarme en el 'simulated annealing' y a Dave Cutler por los datos simulados y el programa hap2. Trabajo financiado por el MEC, proyecto AGL2004-00103.

REFERENCIAS

Carlborg, O., and C. S. Haley. 2004. Epistasis: Too often neglected in complex trait studies? Nat Rev Genet 5: 618-625.

Crow, J. F., and M. Kimura. 1970. An introduction to population genetics theory. Harper & Row, New York.

Kirkpatrick, S., C. D. J. Gerlatt, and M. P. Vecchi. 1983. Optimization by simulated annealing. Science 220: 671-680.

Lin, S., A. Chakravarti, and D. J. Cutler. 2004. Exhaustive allelic transmission disequilibrium tests as a new approach to genome-wide association studies. Nat Genet.

Nelson, M. R., S. L. Kardia, R. E. Ferrell, and C. F. Sing. 2001. A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation. Genome Res 11: 458-470.

Pérez-Enciso, M., and I. Misztal. 2004. Qxpak: A versatile mixed model application for genetical genomics and qtl analyses. Bioinformatics 20: 2792-2798.

Pociot, F., A. E. Karlsen, C. B. Pedersen, M. Aalund, and J. Nerup. 2004. Novel analytical methods applied to type 1 diabetes genome-scan data. Am J Hum Genet 74: 647-660.

Ritchie, M. D. et al. 2001. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. Am J Hum Genet 69: 138-147.