Modelos de interacção genética de dois genes em fenótipos binários complexos: nova modelação estatística e aplicação de metodologia bayesiana

Nuno Sepúlveda
Instituto Gulbenkian de Ciência
Carlos Daniel Paulino
Instituto Superior Técnico, Departamento de Matemática e CEMAT
Carlos Penha-Gonçalves
Instituto Gulbenkian de Ciência

Resumo: Em trabalhos anteriores foram propostos diversos modelos estatísticos para a penetrância de forma a inferir a interacção de dois genes dialélicos na construção de fenótipos binários complexos: modelos de acção independente, modelos de inibição e modelos de número mínimo de alelos. Estes modelos baseiam-se numa decomposição da penetrância através da abordagem por penetrâncias alélicas, que permitiu a inclusão dos conceitos mendelianos de dominância e recessividade alélica na sua modelação. Pretende-se aqui dar a conhecer os avanços mais recentes na parte da modelação da interacção genética, apresentando uma nova decomposição da penetrância e uma nova formulação matemática da dominância e da recessividade. Aplicam-se ainda ferramentas bayesianas para o ajustamento dos modelos de interacção genética a dados experimentais com recurso ao método de amostragem de Gibbs. Toda a metodologia é exemplificada num conjunto de dados de um estudo da susceptibilidade da malária cerebral em ratinhos.

Palavras–chave: fenótipos binários complexos, interacção genética, abordagem por penetrâncias alélicas, análise bayesiana, método de amostragem de Gibbs.

Abstract: In previous papers we proposed several models for penetrance to infer about genetic interaction between two diallelic genes in complex binary traits: independent action models, inhibition models, and minimal allele models. These models are based on a decomposition of penetrance through the allelic penetrance approach, which allows the inclusion of the Mendelian concepts of dominance and recessiveness in the modelling. Here we show recent advances in the genetic interaction modelling, namely presenting a new decomposition of penetrance and a new mathematical formulation of dominance and recessiveness. We apply Bayesian ideas to fit the models to experimental data. With this purpose we estimate parameters via Gibbs sampling. We exemplify all the methodology in a dataset taken from a experimental genetic study of cerebral malaria in mice.

Keywords: complex binary traits, genetic interaction, allelic penetrance approach, Bayesian analysis, Gibbs sampling.

1 Introdução

Fenótipos binários complexos são características biológicas classificadas em duas categorias (*e.g.*, presença ou ausência de uma doença), que mostram frequentemente um padrão complexo de hereditariedade. Actualmente acredita-se que tais fenótipos são o resultado de uma rede intrincada de interacções genéticas e ambientais (Griffiths *et al.*, 2000), o que faz com que indivíduos com o mesmo genótipo possam ou não manifestar o mesmo fenótipo. Para lidar com esta complexidade, existe o conceito de penetrância, que é a probabilidade de um indivíduo manifestar o fenótipo de interesse dado o seu genótipo.

A avaliação estatística da interacção genética em fenótipos binários complexos está actualmente restrita ao ajustamento de modelos lineares generalizados para a penetrância (revistos em Cordell *et al.*, 2001, e Cordell, 2002). Porém, esta análise fornece interpretação genética reduzida (Cordell *et al.*, 2001). Para colmatar esta deficiência interpretativa, tem-se vindo a propor diversos modelos para o caso simples da interacção de dois genes dialélicos em contextos de cruzamentos genéticos experimentais (Sepúlveda, 2004; Sepúlveda *et al.*, 2004a,b). Estes novos modelos baseiam-se em mecanismos especiais de acção genética, que incorporam os conceitos mendelianos de dominância e recessividade alélica através da abordagem por penetrâncias alélicas (APA).

Neste trabalho pretende-se dar a conhecer os novos avanços na modelação da interacção genética, nomeadamente, na formulação da APA. Para além disso, dá-se aqui especial ênfase à metodologia bayesiana para a comparação, selecção e estimação de modelos, o que contrasta com o que foi feito anteriormente, onde se adoptou a metodologia de máxima verosimilhança via algoritmo EM (Sepúlveda *et al.*, 2004a,b). Com esse fim, aplica-se o método de amostragem de Gibbs com recurso aos pacotes estatísticos WinBUGS (Spiegelhalter *et al.*, 2003) e BOA (Smith, 2003). Uma descrição mais detalhada sobre este trabalho pode ser encontrada em Sepúlveda (2004).

A secção 2 apresenta a nova formulação da APA, enquanto que a secção 3 descreve a sua aplicação à modelação de interacção genética. A secção 4 fornece as ferramentas inferenciais bayesianas usadas neste trabalho. A secção 5 exemplifica a aplicação de toda a metodologia num conjunto de dados referente ao controlo genético da malária cerebral em ratinhos (Bagot *et al.*, 2002). Para finalizar, a secção 6 dedica-se à discussão de todo o conteúdo deste trabalho.

2 Abordagem por penetrâncias alélicas

Desde os primórdios da Genética, o fenómeno da penetrância tem sido recorrentemente observado tanto em fenótipos de animais como de plantas, sendo usualmente justificado pela presença de múltiplos genes e de efeitos ambientais. Contudo, a genética experimental oferece a possibilidade de observar variabilidade fenotípica mesmo quando (i) há apenas um gene a controlar o fenótipo, (ii) o resto do genoma está essencialmente fixo, e (iii) as condições ambientais

estão sob um rigoroso controlo experimental. Estas observações sugerem uma propriedade intrínseca da penetrância de índole estocástica associada à própria expressão do genótipo ao nível do fenótipo.

A APA tem como objectivo modelar o comportamento de um gene dialélico num cruzamento experimental entre duas linhas puras. A ideia básica consiste em decompor a penetrância de um genótipo em componentes interna e externa, em que a primeira refere-se exclusivamente à expressão do genótipo em favor do fenótipo de interesse, enquanto que a segunda modela a acção dos factores externos ao genótipo (i.e., resto do genoma e/ou factores ambientais).

O fenótipo de interesse pode ser herdado de duas maneiras: (i) através da própria expressão do genótipo, ou (ii) na ausência da expressão do genótipo, através da expressão de factores externos. Assim, uma decomposição da penetrância de um genótipo i pode ser feita da seguinte forma

$$\pi_i = \pi_i^{int} + (1 - \pi_i^{int})\pi_{ext}, \tag{1}$$

onde π_i^{int} é a chamada penetrância interna do genótipo i e π_{ext} é a probabilidade "média" de os factores externos exprimirem o fenótipo de interesse (penetrância externa) quando considerados todos os genes e os seus respectivos genótipos no resto do genoma e quando contabilizados todos os efeitos possíveis de cariz ambiental ao nível da expressão do fenótipo.

Divide-se ainda a penetrância interna nas contribuições dos alelos que compõem o genótipo. Neste contexto, a expressão de um alelo de um genótipo ao nível do fenótipo é vista como estocástica, podendo ser relacionada com uma prova de Bernoulli com probabilidade de sucesso dada pela probabilidade de um alelo ser expresso ao nível do fenótipo, a *penetrância alélica*. Por simplicidade matemática, as expressões alélicas de um genótipo são consideradas independentes entre si.

Considere-se um gene com alelos a e b. O modelo de dominância é aqui condicionado à expressão de pelo menos um alelo dominante ao nível do fenótipo, enquanto que o de recessividade à expressão de pelo menos um alelo recessivo quando o alelo dominante não o está a fazer. Assim, se o alelo a está a conferir o fenótipo de uma forma dominante, as penetrâncias genotípicas internas são dadas por

$$\pi_i^{int} = \begin{cases} \theta_a^2 + 2\theta_a\,(1-\theta_a)\,, & \text{se } i = aa \\ \theta_a, & \text{se } i = ab \\ 0, & \text{se } i = bb \end{cases} \tag{2}$$

onde θ_a é a penetrância do alelo a. No caso do alelo a ser recessivo, as penetrâncias genotípicas internas são

$$\pi_i^{int} = \begin{cases} \theta_a^2 + 2\theta_a\,(1-\theta_a)\,, & \text{se } i = aa \\ \theta_a(1-\theta_b), & \text{se } i = ab \\ 0, & \text{se } i = bb \end{cases} \tag{3}$$

onde θ_b é a penetrância do alelo b. Note-se que a recessividade é modelada à custa de dois parâmetros, θ_a e θ_b. Ora, para alguns modelos de interacção

genética que incorporam este conceito, existe uma situação de sobreparametrização ou de um modelo saturado para dados de retrocruzamento. Para solucionar este problema, faz-se uma simplificação da equação (3) assumindo que o alelo dominante está sempre a ser expresso (mas não conferindo o fenótipo por hipótese), $\theta_b = 1$, o que conduz às penetrâncias genotípicas internas

$$\pi_i^{int} = \begin{cases} \theta_a^2 + 2\theta_a\left(1-\theta_a\right), & \text{se } i = aa \\ 0, & \text{se } i = ab \\ 0, & \text{se } i = bb \end{cases}. \tag{4}$$

3 Modelação da interacção genética

3.1 Modelos de acção independente

Tal como o seu nome sugere, os modelos de acção independente (MAI) consideram que o fenótipo de interesse pode ser adquirido pela expressão independente de cada gene. Nos MAI especifica-se a natureza genética dos alelos conferidores de fenótipo (dominantes ou recessivos).

Para derivar as penetrâncias genotípicas desta classe de modelos, usam-se os argumentos da APA. Portanto, a penetrância genotípica é dividida nas suas componentes interna e externa. Ora, o facto de se considerar expressões independentes dos alelos conferidores de cada gene implica que a penetrância genotípica interna satisfaz a relação probabilística de união de dois eventos independentes (referentes às expressões de cada gene), i.e.,

$$\pi_{ij}^{int} = \pi_i^{int} + \pi_j^{int} - \pi_i^{int}\pi_j^{int}, \tag{5}$$

onde π_i^{int} e π_j^{int} são as penetrâncias internas do genótipo i do gene 1 e do genótipo j do gene 2, respectivamente. Se um alelo conferidor de um gene for dominante, a penetrância genotípica marginal interna segue a equação (2). Caso o alelo conferidor seja recessivo, então a penetrância genotípica marginal interna exprime-se pela equação (4).

Por ambos os genes terem alelos conferidores, a inclusão da componente externa no modelo é feita da mesma maneira que no caso de um só gene (veja-se equação (1)), mas agora estendida para o caso digénico. Assim, a penetrância de um genótipo combinado genérico (i, j) é dada por

$$\pi_{ij} = \pi_{ij}^{int} + \left(1 - \pi_{ij}^{int}\right)\pi_{ext}, \tag{6}$$

onde π_{ij}^{int} é dado por (5) e π_{ext} é a penetrância externa.

A título ilustrativo, considere-se um gene com alelos a_1 e b_1, e outro com alelos a_2 e b_2. Sejam a_1 e a_2 os alelos conferidores de cada gene, em que o primeiro alelo é dominante e o segundo recessivo. Este modelo está doravante denotado por MAI(D(a_1)/R(a_2)). Assim, as penetrâncias internas π_i^{int} e π_j^{int} são dadas por (2) com a penetrância alélica θ_{a_1} e (4) com a penetrância alélica

Tabela 1: Matriz de penetrâncias do MAI(D(a_1)/R(a_2)) para os genótipos de uma geração F_2 de um intercruzamento.

Genótipos	a_2a_2	a_2b_2	b_2b_2
a_1a_1	$\pi_{a_1a_1/a_2a_2}$	$\pi_{a_1a_1}$	$\pi_{a_1a_1}$
a_1b_1	$\pi_{a_1b_1/a_2a_2}$	$\pi_{a_1b_1}$	$\pi_{a_1b_1}$
b_1b_1	$\pi_{a_2a_2}$	π_{ext}	π_{ext}

θ_{a_2}, respectivamente. A matriz de penetrâncias para os genótipos de uma geração F_2 de um intercruzamento segue a estrutura paramétrica apresentada na tabela 1, em que

$$\begin{aligned}
\pi_{a_1a_1/a_2a_2} &= \pi^{int}_{a_1a_1} + \pi^{int}_{a_2a_2} - \pi^{int}_{a_1a_1}\pi^{int}_{a_2a_2} + \left(1-\pi^{int}_{a_1a_1}\right)\left(1-\pi^{int}_{a_2a_2}\right)\pi_{ext},\\
\pi_{a_1b_1/a_2a_2} &= \pi^{int}_{a_1b_1} + \pi^{int}_{a_2a_2} - \pi^{int}_{a_1b_1}\pi^{int}_{a_2a_2} + \left(1-\pi^{int}_{a_1b_1}\right)\left(1-\pi^{int}_{a_2a_2}\right)\pi_{ext}\\
\pi_{a_1a_1} &= \pi^{int}_{a_1a_1} + \left(1-\pi^{int}_{a_1a_1}\right)\pi_{ext},\\
\pi_{a_1b_1} &= \pi^{int}_{a_1b_1} + \left(1-\pi^{int}_{a_1b_1}\right)\pi_{ext},\\
\pi_{a_2a_2} &= \pi^{int}_{a_2a_2} + \left(1-\pi^{int}_{a_2a_2}\right)\pi_{ext},
\end{aligned}$$

onde $\pi^{int}_{a_1a_1} = \theta^2_{a_1} + 2\theta_{a_1}\left(1-\theta_{a_1}\right)$, $\pi^{int}_{a_1b_1} = \theta_{a_1}$ e $\pi^{int}_{a_2a_2} = \theta^2_{a_2} + 2\theta_{a_2}\left(1-\theta_{a_2}\right)$.

3.2 Modelos de inibição

Os modelos de inibição (MI) postulam um mecanismo de expressão onde existe um alelo conferidor num gene e um alelo inibidor num outro gene. O alelo conferidor visa produzir o fenótipo de interesse através da sua expressão alélica, enquanto que o alelo inibidor tem apenas o papel de inibição da expressão do alelo conferidor. Note-se que a ausência de expressão do alelo inibidor não consegue produzir *per se* o fenótipo de interesse. Os alelos conferidor e inibidor podem ter uma natureza genética dominante ou recessiva.

Neste cenário, o fenótipo de interesse só é herdado quando o alelo conferidor está a exprimir o fenótipo na ausência de acção de inibição por parte do alelo inibidor. Assim, assumindo que alelo conferidor está no gene 1 e o alelo inibidor no gene 2, a penetrância interna de um genótipo combinado (i,j) satisfaz

$$\pi^{int}_{ij} = \pi^{int}_{i}\left(1-\phi^{int}_{j}\right), \tag{7}$$

onde π^{int}_{i} é a penetrância interna do genótipo i do gene 1 referente à expressão do fenótipo de interesse e ϕ^{int}_{j} é a penetrância interna do genótipo j do gene 2 com respeito à acção de inibição da expressão do gene 1. Tal como no caso dos MAI, a natureza dominante ou recessiva dos alelos de cada gene é introduzida no modelo através da especificação das penetrâncias genotípicas marginais internas π^{int}_{j} e ϕ^{int}_{i} em termos das equações (2) e (4).

Tabela 2: Matriz de penetrâncias do MI($\mathrm{R}_c(a_1)/\mathrm{R}_i(b_2)$) para os genótipos de uma geração F_2 de um intercruzamento.

Genótipos	a_2a_2	a_2b_2	b_2b_2
a_1a_1	$\pi_{a_1a_1}$	$\pi_{a_1a_1}$	$\pi_{a_1a_1/b_2b_2}$
a_1b_1	π_{ext}	π_{ext}	π_{ext}
b_1b_1	π_{ext}	π_{ext}	π_{ext}

Como se assume que os factores externos só contribuem para o fenótipo de interesse quando não há expressão do genótipo combinado para esse fenótipo e como existe apenas um alelo conferidor, a penetrância do genótipo combinado (i, j) tem de obedecer à seguinte fórmula

$$\pi_{ij} = \pi_i^{int}\left(1-\phi_j^{int}\right) + \left(1-\pi_i^{int}\right)\left(1-\phi_j^{int}\right)\pi_{ext}, \tag{8}$$

a qual difere da equação (6) para os MAI.

Com o propósito de exemplificar os resultados derivados acima, considere-se um MI em que o gene 1 tem um alelo conferidor recessivo a_1 e outro gene tem um alelo inibidor recessivo b_2. Este modelo fica representado por MI($\mathrm{R}_c(a_1)/\mathrm{R}_i(b_2)$), onde os índices c e i denotam um alelo conferidor e inibidor, respectivamente. Como ambos os genes são recessivos, as penetrâncias internas de cada gene são dadas pela equação (4), mas com diferentes parâmetros. Por aplicação da equação (8), chega-se à matriz de penetrâncias da tabela 2 relativa a um intercruzamento com as seguintes fórmulas

$$\begin{aligned}
\pi_{a_1a_1} &= \theta_{a_1}^2 + 2\theta_{a_1}(1-\theta_{a_1}) + (1-\theta_{a_1})^2\pi_{ext},\\
\pi_{a_1a_1/b_2b_2} &= \left[\theta_{a_1}^2 + 2\theta_{a_1}(1-\theta_{a_1})\right](1-\theta_{b_2})^2 + (1-\theta_{a_1})^2(1-\theta_{b_2})^2\pi_{ext}.
\end{aligned}$$

3.3 Modelos de número mínimo de alelos

Os modelos de número mínimo de alelos (MNMA) traduzem um mecanismo em que o fenótipo de interesse é conferido sempre que o número de alelos conferidores de um genótipo combinado a serem expressos simultaneamente ultrapassa um determinado patamar, que define a ordem dos MNMA.

Para facilitar a descrição probabilística dos MNMA, os genótipos combinados são identificados em termos do número de alelos conferidores em cada gene. Assim, as penetrâncias "observável" e interna de um genótipo combinado com x_i alelos conferidores no gene $i = 1, 2$ são representadas por $\pi_{x_1x_2}$ e $\pi_{x_1x_2}^{int}$, respectivamente. A penetrância do alelo conferidor a_i do gene i está denotada por γ_{a_i}.

Sejam Y_1 e Y_2 as variáveis aleatórias que indicam o número de alelos conferidores a serem expressos em cada gene com distribuições binomiais com um

Tabela 3: Matriz de penetrâncias do $\text{MNMA}_2(a_1/a_2)$ para os genótipos de uma geração F_2 de um intercruzamento.

Genótipos	a_2a_2	a_2b_2	b_2b_2
a_1a_1	π_{22}	π_{21}	π_{20}
a_1b_1	π_{12}	π_{11}	π_{ext}
b_1b_1	π_{02}	π_{ext}	π_{ext}

número de provas x_i e probabilidade de sucesso $\gamma_{a_i}, i = 1, 2$. Assumindo independência entre Y_1 e Y_2, a função de probabilidade do número total de alelos conferidores de um genótipo combinado (x_1, x_2) a serem expressos simultaneamente, $Y = Y_1 + Y_2$, é dada por

$$P\left[Y = y|(x_1, x_2)\right] = \sum_{l=0}^{\min(x_1,y)} P\left[Y_1 = l|(x_1, x_2)\right] P\left[Y_2 = y - l|(x_1, x_2)\right], \quad (9)$$

onde

$$P\left[Y_i = y_i|(x_1, x_2)\right] = \left(\begin{array}{c} x_i \\ y_i \end{array} \right) \gamma_{a_i}^{y_i} \left(1 - \gamma_{a_i}\right)^{x_i - y_i}. \quad (10)$$

Assim, um MNMA de ordem $k = 1, \ldots, 4$ possui a seguinte penetrância interna

$$\pi_{x_1x_2}^{int} = P\left[Y \geq k|(x_1, x_2)\right] = \sum_{y=k}^{x_1+x_2} P\left[Y = y|(x_1, x_2)\right], \quad (11)$$

onde $P\left[Y = y|(x_1, x_2)\right]$ segue a equação (9). Note-se que $\pi_{x_1x_2}^{int} = 0$ quando $k > x_1 + x_2$. Como os dois genes podem conferir o fenótipo, os factores externos são incluídos no modelo através da equação (6), tal como se fez nos MAI.

A título ilustrativo, a tabela 3 apresenta a estrutura da penetrância relativa a um intercruzamento para o MNMA de ordem 2 com alelos conferidores a_1 e a_2 (denotado por $\text{MNMA}_2(a_1/a_2)$), onde, para $x_1 + x_2 \geq 2 \wedge x_1 > 0 \wedge x_2 > 0$,

$$\pi_{x_1x_2} = \sum_{y=2}^{x_1+x_2} P\left[Y = y|(x_1, x_2)\right] + \pi_{ext} \sum_{y=0}^{1} P\left[Y = y|(x_1, x_2)\right]$$

com $P\left[Y = y|(x_1, x_2)\right]$ determinado por (9) com parâmetros θ_{a_1} e θ_{a_2}, e

$$\begin{aligned} \pi_{20} &= \gamma_{a_1}^2 + \left(1 - \gamma_{a_1}^2\right) \pi_{ext}, \\ \pi_{02} &= \gamma_{a_2}^2 + \left(1 - \gamma_{a_2}^2\right) \pi_{ext}. \end{aligned}$$

4 Metodologia inferencial

Os dados de cruzamentos genéticos são tipicamente representados por tabelas de contingências $I \times J \times 2$, onde I e J são os números de genótipos de cada gene numa geração F_2. Por exemplo, nos intercruzamentos e nos retrocruzamentos tem-se $I = J = 3$ e $I = J = 2$, respectivamente. Como modelo amostral, vai-se supor sem perda de generalidade um produto de distribuições binomiais, uma distribuição binomial por cada genótipo combinado dos dois genes.

A especificação bayesiana *a priori* começa por assumir uma independência entre os parâmetros de cada modelo. Pela inovação conceptual das penetrâncias alélicas e externa, parece razoável a adopção de distribuições não-informativas numa óptica de aplicação do princípio de Bayes-Laplace, i.e., distribuições i.i.d. Uniformes no intervalo (0,1).

A elevada complexidade algébrica das penetrâncias genotípicas dos modelos faz transparecer a necessidade de recorrer a métodos expeditos para o cálculo das respectivas distribuições *a posteriori*. Assim, faz-se uso do muito em voga método de amostragem de Gibbs disponível no WinBUGS (Spiegelhalter *et al.*, 2003). Sepúlveda (2004) prova que as distribuições condicionais completas são log-côncavas, o que viabiliza a aplicação do método de rejeição adaptativa proposto por Gilks (1992) para a simulação de valores das respectivas distribuições condicionais completas. A análise de convergência dos valores simulados foi feita no BOA (Smith, 2003).

Para a comparação e selecção de modelos, existem diversas ferramentas bayesianas (*vide* Paulino *et al.*, 2003a), com maior ou menor justificação teórica no edifício metodológico bayesiano e de menor ou maior facilidade de implementação prática. Neste trabalho recorre-se às seguintes medidas: (a) a probabilidade preditiva *a priori* (PPP), (b) a soma dos logaritmos das ordenadas preditivas condicionais (SLNCPO), (c) a média *a posteriori* da função paramétrica de Pearson e (d) a medida DIC. Os pormenores do seu cálculo podem ser encontrados em Sepúlveda (2004).

Devido ao elevado número de modelos de acção genética passíveis de se poderem ajustar aos dados, desenvolveu-se uma estratégia de selecção de modelos, dividida em três fases, que fosse possível de ser executada em tempo real: (i) numa primeira fase estabelece-se o conjunto inicial de modelos através de uma avaliação empírica dos efeitos dos alelos de cada gene em relação ao fenótipo de interesse; (ii) numa segunda fase calcula-se a medida DIC e a média *a posteriori* da função paramétrica de Pearson, escolhendo os modelos que tiverem simultaneamente os menores valores dessas duas medidas; (3) numa última fase comparam-se os modelos através das PPP e da SLNCPO, seleccionando aqueles que tiverem os maiores valores dessas medidas.

Em termos de estimação paramétrica, calculam-se, essencialmente, estimativas das penetrâncias alélicas, externas e genotípicas. Nesta fase faz-se uma análise exploratória dos valores simulados das distribuições *a posteriori* dos parâmetros (e das suas funções paramétricas de interesse), com o cálculo de medidas típicas (*e.g.*, média, mediana e desvio-padrão), intervalos de credibili-

Tabela 4: Dados da malária cerebral em ratinhos, onde s_i e r_i representam os alelos herdados das estirpes susceptível e resistente no locus $i = 1, 2$, respectivamente. A penetrância refere-se ao fenótipo de susceptibilidade.

Genótipos		Fenótipo		
Locus 1	Locus 2	Susc.	Resist.	Penet.
s_1s_1	s_2s_2	35	10	0.78
	r_2s_2	25	23	0.56
r_1s_1	s_2s_2	27	21	0.52
	r_2s_2	9	40	0.18

dade HPD através do método proposto por Chen e Shao (1999), etc. No caso da estimação simultânea de c penetrâncias genotípicas, aplica-se o método de Bonferroni para a determinação de uma região de credibilidade conjunta a $\gamma \times 100\%$, calculando o produto cartesiano dos intervalos de credibilidade HPD individuais a $\gamma^{1/c} \times 100\%$ para cada penetrância genotípica.

5 Aplicação

Bagot *et al.* (2002) descrevem um estudo sobre o controlo genético da malária cerebral em ratinhos. O delineamento experimental consistiu num retrocruzamento entre duas estirpes de ratinhos, uma susceptível e outra resistente à doença, onde a primeira geração foi cruzada com a estirpe parental susceptível. Os dados da tabela 4 reportam-se aos genótipos e aos respectivos fenótipos dos indivíduos da geração F_2 nos dois *loci* mais associados à doença. Pretende-se agora aplicar toda a metodologia apresentada de forma inferir sobre a acção mais plausível dos dois *loci* na construção do fenótipo de susceptibilidade.

A observação atenta da tabela 4 mostra que os alelos derivados da estirpe susceptível em ambos os *loci* tendem a aumentar a probabilidade de um ratinho ser susceptível à doença. Usando esta observação como um crivo inicial para a selecção de modelos, escolhem-se MNMA e MAI com alelos conferidores derivados da estirpe susceptível, e MI que tenham um alelo conferidor proveniente da estirpe susceptível num dos locus e um alelo inibidor derivado da outra estirpe noutro gene. Deste conjunto excluem-se todos os modelos que não estejam parametrizados pelas penetrâncias dos alelos de cada *locus* para os dados em questão, pois estes modelos não esclarecem acerca da acção dos dois *loci* (*e.g.*, $\mathrm{MI}(\mathrm{D}_c(s_1)/\mathrm{R}_i(r_2))$). Assim, a segunda etapa de comparação e de selecção de modelos abrange com um conjunto de 12 modelos: $\mathrm{MNMA}_k(s_1/s_2)$, $k = 1, 2, 3, 4$, os quatro tipos de MAI com alelos conferidores s_1 e s_2, $\mathrm{MI}(\mathrm{D}_c(s_1)/\mathrm{D}_i(r_2))$, $\mathrm{MI}(\mathrm{D}_i(r_1)/\mathrm{D}_c(s_2))$, $\mathrm{MI}(\mathrm{D}_i(r_1)/\mathrm{R}_c(s_2))$ e $\mathrm{MI}(\mathrm{R}_c(s_1)/\mathrm{D}_c(r_2))$.

A tabela 5 apresenta os valores da média *a posteriori* da função paramétrica de Pearson e da medida DIC para cada um dos modelos descritos acima.

Tabela 5: Comparação e selecção dos modelos para os dados da malária cerebral.

Modelo	E(Pearson)	DIC	SLNCPO	PPP
MAI($D(s_1)/D(s_2)$)	16.660	36.262	—	—
MAI($D(s_1)/R(s_2)$)	7.449	26.335	-118.587	6.06×10^{-8}
MAI($R(s_1)/D(s_2)$)	9.215	28.249	—	—
MAI($R(s_1)/R(s_2)$)	3.210	22.643	-116.424	1.28×10^{-6}
MI($D_c(s_1)/D_i(r_2)$)	7.917	26.762	-118.546	1.45×10^{-7}
MI($D_i(r_1)/D_c(s_2)$)	9.219	28.132	—	—
MI($D_i(r_1)/R_c(s_2)$)	15.140	34.635	—	—
MI($R_c(s_1)/D_i(r_2)$)	18.050	37.734	—	—
$MNMA_1(s_1/s_2)$	16.660	36.262	—	—
$MNMA_2(s_1/s_2)$	6.090	25.266	-117.806	2.21×10^{-7}
$MNMA_3(s_1/s_2)$	3.038	22.035	-116.243	1.41×10^{-6}
$MNMA_4(s_1/s_2)$	19.660	37.734	—	—

Destacam-se os seguintes cinco modelos por apresentarem valores baixos em ambas as medidas: MAI($D(s_1)/R(s_2)$), MAI($R(s_1)/R(s_2)$), MI($D_c(s_1)/D_i(r_2)$), $MNMA_2(s_1/s_2)$ e $MNMA_3(s_1/s_2)$. Para refinar este conjunto de modelos, calculam-se as estimativas da PPP e da SLNCPO (veja-se, novamente, tabela 5). Ora, estas medidas distinguem, claramente, os modelos MAI($R(s_1)/R(s_2)$) e $MNMA_3(s_1/s_2)$ dos restantes. Por apresentarem também os melhores resultados para a média *a posteriori* da função paramétrica de Pearson e para a medida DIC, estes dois modelos são considerados como os que melhor se adequam aos dados.

Por estes resultados, há evidências para dois tipos distintos de acção entre os dois *loci* no controlo da susceptibilidade à SMC: (1) cada *locus* é recessivo com respeito ao alelo herdado da estirpe susceptível e suficiente para causar doença, actuando de uma forma autónoma entre si, ou (2) os *loci* constroem conjuntamente o fenótipo de interesse, sentindo-se o seu efeito quando há pelo menos três alelos herdados da estirpe susceptível a serem expressos de uma forma simultânea, independentemente de onde esses alelos estão localizados no genótipo combinado.

Após um diagnóstico apropriado de convergência (vejam-se detalhes em Sepúlveda, 2004), o método de amostragem de Gibbs conduziu às estimativas *a posteriori* apresentadas na tabela 6 para os parâmetros dos dois "melhores" modelos. Note-se que o grau de credibilidade individual dos intervalos HPD para as penetrâncias genotípicas foi estabelecido em 98.7%, de forma a garantir um grau de credibilidade global de pelo menos 95% para a respectiva região de credibilidade.

As estimativas pontuais bayesianas para a expressão dos alelos diferem de um modelo para o outro. Este facto não é de estranhar, uma vez que as estruturas paramétricas dos dois modelos são algo diferentes. Contudo, as estimativas da

Tabela 6: Principais estimativas *a posteriori* dos parâmetros e das suas funções paramétricas de interesse para os dois "melhores" modelos dos dados da malária cerebral. [a]IC HPD a 95% e [b]IC HPD a 98.7%.

MAI(R(s_1)/R(s_2))					
Parâmetro	Média	Mediana	Desvio Padrão	IC HPD	
θ_{s_1}	0.243	0.278	0.057	0.125	0.348^a
θ_{s_2}	0.278	0.278	0.056	0.167	0.388^a
π_{ext}	0.195	0.191	0.053	0.102	0.306^a
$\pi_{s_1s_1/s_2s_2}$	0.759	0.761	0.042	0.649	0.855^b
$\pi_{s_1s_1/s_2r_2}$	0.537	0.539	0.065	0.379	0.691^b
$\pi_{s_1r_1/s_2s_2}$	0.579	0.581	0.062	0.418	0.727^b
$\pi_{s_1r_1/s_2r_2}$	0.195	0.191	0.055	0.077	0.335^b
MNMA$_3$(s_1/s_2)					
Parâmetro	Média	Mediana	Desvio Padrão	IC HPD	
γ_{s_1}	0.702	0.704	0.111	0.485	0.915^a
γ_{s_2}	0.776	0.781	0.111	0.584	1.000^a
π_{ext}	0.211	0.208	0.057	0.105	0.322^a
$\pi_{s_1s_1/s_2s_2}$	0.782	0.785	0.042	0.663	0.871^b
$\pi_{s_1s_1/s_2r_2}$	0.512	0.512	0.058	0.374	0.651^b
$\pi_{s_1r_1/s_2r_2}$	0.542	0.543	0.056	0.397	0.670^b
$\pi_{s_1r_1/s_2r_2}$	0.211	0.208	0.057	0.087	0.358^b

penetrância externa são muito semelhantes em ambos os modelos. Em relação às penetrâncias genotípicas, as médias e as medianas *a posteriori* estão bastante próximas dos respectivos valores observados tanto no MAI(R(s_1)/R(s_2)) como no MNMA$_3$(s_1/s_2). Para além disso, as respectivas regiões de credibilidade incluem as penetrâncias empíricas, praticamente, no seu centróide. Por fim, a figura 1 mostra que as densidades *a priori* da penetrância do genótipo s_1s_1/s_2s_2 para ambos os modelos diferem, substancialmente, das suas contrapartidas *a posteriori*, o que indica uma forte actualização do conhecimento *a priori* através dos dados. Esta observação pode ser estendida para as restantes penetrâncias genotípicas (gráficos não apresentados).

6 Discussão

A modelação da dominância e da recessividade num gene dialélico tem vindo a merecer atenção em fenótipos binários complexos (Sepúlveda *et al.*, 2004a,b). A dominância foi desde logo fácil de conceber, o que contrastou com a situação de recessividade. Sepúlveda *et al.* (2004a) referem-se à recessividade na ausência de expressão do alelo dominante, que está na linha de pensamento de não haver recessividade mas apenas dominância incompleta (Griffiths *et al.*, 2000). Ora,

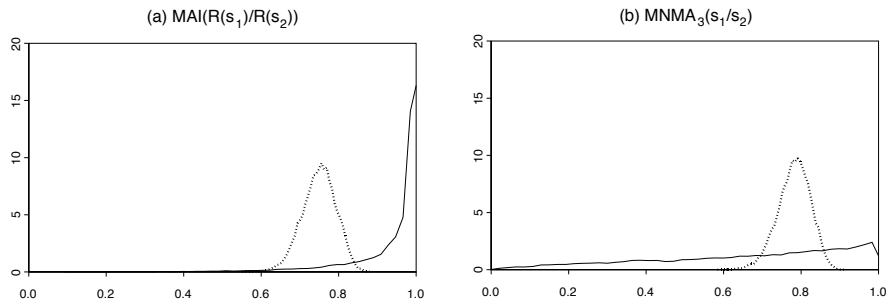

Figura 1: Densidades *a priori* (a cheio) e *a posteriori* (a ponteado) da penetrância do genótipo s_1s_1/s_2s_2 induzidas pelos MAI(R(s_1)/R(s_2) e MNMA$_3$(s_1/s_2).

esta definição não atribui um papel activo ao alelo recessivo na modelação, o que parece contrariar a expressão estocástica dos alelos tanto ao nível molecular como celular. Sepúlveda *et al.* (2004b) resolveram este problema definindo recessividade quando há dois alelos recessivos a exprimirem-se simultaneamente. Porém, esta definição continuou a não atribuir importância ao alelo recessivo na modelação da penetrância do genótipo heterozigótico. Neste trabalho conseguiu-se ultrapassar essa deficiência definindo recessividade quando o alelo dominante não está exprimir o seu fenótipo e o alelo recessivo o está a fazer. Como mostra a substituição da equação (3) na equação (1), a recessividade assim definida é tomada como o modelo saturado para o caso da modelação da acção de um único gene. Contudo, há que referir que, para evitar eventuais problemas de sobreparametrização nos modelos de interacção genética, assumiu-se que o alelo dominante se exprime sempre, cancelando o potencial efeito estocástico do alelo recessivo na penetrância do genótipo heterozigótico (veja-se a equação (4)).

Duas decomposições distintas da penetrância foram experimentadas anteriormente. Em Sepúlveda *et al.* (2004a), a penetrância de um genótipo i foi decomposta em

$$\pi_i = (1 - \pi_{ext})\pi_i^{int} + (1 - \pi_i^{int})\pi_{ext}, \tag{12}$$

onde π_{ext} e $1-\pi_{ext}$ são as probabilidades dos factores externos estarem a favorecer ou a suprimirem o fenótipo de interesse, respectivamente. Ora, a equação acima toma uma expressão matemática idêntica à da probabilidade de má classificação em dados binários (veja-se, por exemplo, Paulino *et al.*, 2003b), que exibe falta de identificabilidade. Para evitar este problema, Sepúlveda *et al.* (2004b) usaram a mesma decomposição apresentada neste trabalho (veja-se a equação (1)). Contudo, esta não inclui uma potencial acção de inibição da componente interna da penetrância por parte dos factores externos. De facto, o caso mais geral seria considerar

$$\pi_i = \pi_{ext}^{*}\pi_i^{int} + (1 - \pi_i^{int})\pi_{ext}, \tag{13}$$

onde π^*_{ext} denota a probabilidade de os factores externos não estarem a suprimir a expressão do gene em causa. No entanto, a sua transposição para a modelação de interacção genética conduziria a um problema de sobreparametrização para dados de retrocruzamentos, o que limitaria a sua aplicação na prática.

A penetrância externa foi assumida ser idêntica ao longo de todos os genótipos combinados de dois genes. Isso pode não ser verdade, nomeadamente, em situações em que existem outros genes no fundo genético em ligação com aqueles em estudo. Neste cenário, há que fazer ou $\pi_{ext,ij}$, ou $\pi_{ext,i}$ ou $\pi_{ext,j}$, dependendo da natureza de ligação dos genes do fundo genético com aqueles em estudo. Contudo, cair-se-ia em situações de sobreparametrização ou de falta de identificabilidade, o que não seria desejável. Portanto, a suposição de $\pi_{ext}, \forall i, j$ é imposta por motivos pragmáticos de realização de inferências.

Por tudo o que já foi dito, a escolha da decomposição da penetrância adoptada é uma espécie de compromisso entre a sua generalidade e os eventuais problemas de falta de identificabilidade ou de sobreparametrização na modelação da interacção genética.

Toda a metodologia bayesiana apresentada foi aplicada a um conjunto de dados referentes à interacção de dois *loci* na manifestação de susceptibilidade à malária cerebral em ratinhos. Os resultados mostraram evidências para dois tipos distintos de acção genética: (i) uma acção independente entre os dois *loci* com alelos conferidores de cada locus herdados da estirpe susceptível e ambos recessivos; e (ii) uma acção de cooperação entre *loci* em que a susceptibilidade está condicionada à expressão de pelo menos três alelos conferidores no genótipo combinado dos *loci*. Assim, é possível recomendar a realização de futuras experiências. Sugere-se, então, a construção de uma linha de ratinhos congénicos para cada um dos *locus*, em que os alelos da estirpe susceptível são inseridos num fundo genético da estirpe resistente. Caso não se observe o fenótipo de susceptibilidade nessas duas linhas, elimina-se o modelo de acção independente, passando-se a avaliar o modelo de número mínimo de alelos de ordem 3 através da inserção de pelo menos 3 alelos derivados da estirpe susceptível nos dois *loci* num fundo genético resistente.

Agradecimentos

Ao Jorge Carneiro, Rui Gardner e Tiago Paixão do Instituto Gulbenkian de Ciência (IGC) pelos seus valiosos comentários no decurso deste trabalho. Ao IGC pelo apoio financeiro fornecido na realização deste trabalho.

Referências

[1] Bagot, S., Campino, S., Penha-Gonçalves, C., Pied, S., Cazenave, P. e Holmberg, D. (2002). Identification of two cerebral malaria resistance loci using an inbred wild-derived mouse strain. *Proceedings of the National Academy of Sciences*, Vol. 99, p. 9919-9923.

[2] Chen, M.-H. e Shao, Q.-M. (1999). Monte Carlo estimation of Bayesian credible and HPD intervals. *Journal of Computational and Graphical Statistics*, Vol. 8, p. 69-92.

[3] Cordell, H. (2002). Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Human Molecular Genetics*, Vol. 11, p. 2463-2468.

[4] Cordell, H., Todd, J., Hill, N., Lord, C., Lyons, P., Peterson, L., Wicker, L. e Clayton, D. (2001). Statistical modeling of interlocus interactions in a complex disease: rejection of the multiplicative model of epistasis in type I diabetes. *Genetics*, Vol. 158, p. 357-367.

[5] Gilks, W. (1992). Derivative-free adaptive rejection sampling for Gibbs sampling. Em *Bayesian Statistics 4* (Bernardo, J. M., Berger, J. O., Dawid, A. P. e Smith, A. F., eds.), p. 641-665. Oxford University Press.

[6] Griffiths, A., Miller, J., Suzuki, D., Lewontin, R. e Gelbart (2000). *An Introduction to Genetic Analysis* (seventh edition). W. H. Freeman.

[7] Paulino, C. D., Amaral Turkman, M. A. e Murteira, B. (2003a). *Estatística Bayesiana*. Fundação Calouste Gulbenkian.

[8] Paulino, C. D., Soares, P. e Neuhaus, J. (2003b). Binomial regression with misclassification. *Biometrics*, Vol. 59, p. 670-675.

[9] Sepúlveda, N. (2004). *Modelos Estatísticos Para a Acção Conjunta de Dois Loci em Fenótipos Binários Complexos*. Tese de Mestrado, Instituto Superior Técnico, Universidade Técnica de Lisboa.

[10] Sepúlveda, N., Paulino, C. D. e Penha-Gonçalves, C. (2004a). Modelos de interacção genética: uma abordagem por penetrâncias alélicas. Em *A Estatística com Acaso e Necessidade* (Rodrigues, P. M. M., Rebelo, E. L. e Rosado, F., eds.), p. 735-746. Edições SPE.

[11] Sepúlveda, N., Paulino, C. D. e Penha-Gonçalves, C. (2004b). Statistical models for the joint action of two loci in complex binary traits. *Preprint* 4/2004, Departamento de Matemática, Instituto Superior Técnico, Universidade Técnica de Lisboa.

[12] Smith, B. (2003). *Bayesian Output Analysis Program (BOA) Version 1.0 User's Manual*. Department of Biostatistics, College of Public Health, University of Iowa.

[13] Spiegelhalter, D., Thomas, A., Best, N. e Lunn, D. (2003). *WinBUGS User Manual, Version 1.4*. MRC Biostatistics Unit, Institute of Public Health & Department of Epidemiology and Public Health, Imperial College School of Medicine.