Journal of Theoretical Biology 432 (2017) 100–108

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/jtbi

Joint parameter estimation in the QTL mapping of ordinal traits

Xiaona Sheng[a], Yihong Qiu[b], Ying Zhou[b,*], Wensheng Zhu[c,*]

[a] School of Information Engineering, Harbin University, Harbin 150086, China
[b] School of Mathematical Sciences, Heilongjiang University, Harbin 150080, China
[c] School of Mathematics and Statistics,Northeast Normal University, Changchun 130024, China

ARTICLE INFO

Article history:
Received 26 December 2016
Revised 9 July 2017
Accepted 5 August 2017
Available online 12 August 2017

Keywords:
Cumulative logistic regression model
EM algorithm
Multiple-interval mapping
Ordinal traits

ABSTRACT

With the rapid development of statistical genetics, the deep researches of ordinal traits have been gradually emphasized. The data of these traits bear relatively less information than those of continuous phenotypes, therefore it is more complex to map the quantitative trait loci (QTL) of ordinal traits. In this paper, the multiple-interval mapping method is considered in the genetic mapping of ordinal traits. By combining threshold model and statistical model, we build a cumulative logistic regression model to express the relationship between the ordinal data and the QTL genotypes. In order to make the interval mapping more straightforward, we treat the recombination rates as unknown parameters, and then simultaneously obtain the estimates of QTL positions, QTL effects and threshold parameters via the EM algorithm. We perform simulation experiments to investigate and compare the proposed method. We also present a real example to test the reasonableness of the considered model and estimate both model parameters and QTL parameters. Both results of simulations and example show that the method we proposed is reasonable and effective.

1. Introduction

With the appearance of more and more research results in recent years, it was found that as the unit of genetic functions, genes are not completely distributed on chromosomes at random. Some genes that control a certain trait usually have similar structures and related functions, and they are distributed in a certain region of a chromosome in the form of gene cluster. Over the years, people have provided lots of statistical methods on QTL mapping analysis. Performing interval mapping of gene loci has become an important tool in the studies of statistical genetics (Haley and Knott, 1992; Jansen, 1993; Kao et al., 1999; Lander and Botstein, 1989; Zeng, 1994). However, most of these statistical methods are fit for continuous traits.

In some studies of QTL mapping, there exists a special kind of traits which are ordinal. These traits include several ordinal classifications. For example, when studying psychological phenomenon, we usually measure attitude or preference according to several types (strongly disagree, disagree, neutral, support, and strongly support). Also, sometimes people are more willing to convert a continuous variable to an ordinal categorical variable, such as when analyzing drug use and health of cocaine users. Researchers can divide their mental depression into four categories (slight depression, low depression, moderate depression, and high depression) by psychology indicators. These response variables are usually coded by a sequence of integers, e.g., 1, 2, 3, 4, etc., and the categories have explicit sort from low to high, but the gap between neighboring categories is unknown. Thus, an ordinal trait is different from to a quantitative trait or a qualitative trait. It is usually divided into the classification of quantitative traits, but sometime it is also regarded as a qualitative trait to a certain extent because quantitative change can cause a qualitative one. On phenotypic scale, most ordinal traits do not follow normal distribution, and they usually present as binomial or multinomial states.

So far, there have been some researches on QTL mapping of ordinal traits, because these researches usually have certain biological significance or economic value. In modern animal breeding, people pay more and more attention to ordinal traits. One reason is that many important economic traits belong to ordinal ones, moreover, many examples have shown that conducting researches on ordinal traits can meet many actual needs. Hackett and Weller (1995) proposed a method to perform genetic mapping of ordinal traits by generalized linear models. Rao and Xu (1998) applied generalized linear model to map ordinal traits in four-way cross populations. Within the framework of generalized linear model, a method based on threshold model of mapping QTL of resistance (ordinal) traits in livestock was simulated, and the efficiency of the method was compared with that of gen-

* Correspondence authors.
E-mail addresses: yzhou@aliyun.com (Y. Zhou), wszhu@nenu.edu.cn (W. Zhu).

http://dx.doi.org/10.1016/j.jtbi.2017.08.007
0022-5193/

eral linear methods (Yin et al., 2005). Xu and Xu (2006) proposed a multivariate model to analyze ordinal traits with the EM algorithm to estimate parameters. Using a unified mixture generalized linear model, Chen and Liu (2009) proposed a multiple-interval mapping method in their hybridization experiments. Recently, Feng et al. (2013) applied an efficient hierarchical generalized linear mixed model to map QTL of ordinal traits in crop cultivars.

Besides, Rao and Xia (2000) used mixture threshold model in their mapping of ordinal traits. Spyrides-Cunha et al. (2000) discussed the application of proportional odds model when dealing with ordinal data. Rao and Xia (2001) proposed a novel statistical model aiming at structural heterogeneity problem and promoted the research progress of ordinal traits. Based on a threshold model, Li et al. (2006) proposed a multiple-interval mapping method about ordinal trait. Yi et al. (2007) proposed an ordinal Probit model to map ordinal traits with epistasis. Liu et al. (2009) chose the maximum likelihood interval-mapping method to conduct linkage analysis for gene loci of ordinal traits. Based on the method of composite interval mapping, Jia and Liu (2011) identified the QTL for resistance to rice blast.

Statistical methods about QTL detection and ordinal-trait mapping have been proposed and continuously improved, so that the efficiency and accuracy of QTL detection have been gradually improved. However, the detecting accuracy in mapping genetic loci of ordinal traits and the implementation of algorithm still need further research and exploration. In this study, we proposed a new strategy for mapping ordinal traits in the framework of multiple-interval mapping. Through simulation experiments and example analysis, we validate that the new method is qualified to map QTL of ordinal traits and has its advantages.

2. Theory and method

In QTL mapping, a necessary step is to find an appropriate model to link trait values with QTL genotypes. For continuous traits, statistical models, such as regression model can be directly used in multiple-interval mapping (MIM). But for ordinal traits, these models are not suitable to use directly. Currently, threshold model has been discussed by more and more researchers. In this model, it is assumed that there is an underlying unobservable response variable for the observable ordinal trait, called liability, which may be continuous, and when it reaches a certain threshold value, the classification type can be determined. Thus, if we want to build a model to describe the relationship between ordinal data and QTL genotypes, firstly we should connect ordinal data with continuous liability via the threshold model, and then link the liability to QTL genotypes by another suitable statistical model.

2.1. Threshold model

Assume that there are c observable values for an ordinal-scaled trait in a trial, coded as 1, 2,..., c. Let Y denote the trait value of an individual, which takes value from {1, 2, ..., c}, and ξ denote the underlying liability of the individual. The two kinds of variables can be linked by the following threshold model:

$$\alpha_{q-1} < \xi \le \alpha_q \Leftrightarrow Y = q;\ \ q = 1, 2, \ldots, c;\ \ \alpha_0 = -\infty;\ \ \alpha_c = \infty, \tag{1}$$

where α_r $(r = 0, 1, 2, \ldots, c)$ is a set of fixed (unknown) values in ascending order, called thresholds.

2.2. Statistical model

In this paper, according to the particularity of ordinal trait, we choose a cumulative logistic regression model as statistical model. We assume that the trait we discuss is determined by m_0 diallelic QTL which are located in m marker intervals $(m_0 \le m)$, and each interval contains at most one QTL. If a backcross (BC) population is considered, the basic model of cumulative logistic regression model is

$$\xi = \mu + \sum_{j=1}^{m} \beta_j G_Q^j + \sum_{r<t\le m} \delta_{rt} G_Q^r G_Q^t + \varepsilon, \tag{2}$$

where ξ is the liability, μ denotes the overall mean of genotypic values, β_j represents the main effect of the QTL in the jth marker interval, and G_Q^j is the corresponding QTL genotype in the jth interval. We let $G_Q^j = 0$, if there does not exist QTL within the jth interval of an individual; otherwise, G_Q^j can be expressed as

$$G_Q^j = \begin{cases} 1, & for\ Q_jQ_j, \\ 2, & for\ Q_jq_j. \end{cases}$$

δ_{rt} $(r, t = 1, 2, \ldots, m)$ denotes the epistatic effect of the QTL within the rth interval and the tth intervals, and $G_Q^r G_Q^t$ stands for the interaction between the corresponding loci. ε is the error term of the model.

Denote the cumulative probabilities of c ordinal classifications by γ_1, γ_2, ..., γ_c, where $\gamma_c = 1$. Assume ε follows the standard logistic distribution, so $\xi - (\mu + \sum_{j=1}^{m} \beta_j G_Q^j + \sum_{r<t\le m} \delta_{rt} G_Q^r G_Q^t)$ follows the standard logistic distribution. According to the threshold model (1), we have

$$\gamma_k = P(Y \le k|G_Q) = P(\xi \le \alpha_k|G_Q) = \frac{\exp\left[\alpha_k - \left(\mu + \sum_{j=1}^{m} \beta_j G_Q^j + \sum_{r<t\le m} \delta_{rt} G_Q^r G_Q^t\right)\right]}{1 + \exp\left[\alpha_k - \left(\mu + \sum_{j=1}^{m} \beta_j G_Q^j + \sum_{r<t\le m} \delta_{rt} G_Q^r G_Q^t\right)\right]}. \tag{3}$$

Eq. (3) can be written as a generalized linear model

$$\ln \frac{\gamma_k}{1-\gamma_k} = \alpha_k - \mu - \sum_{j=1}^{m} \beta_j G_Q^j - \sum_{r<t\le m} \delta_{rt} G_Q^r G_Q^t. \tag{4}$$

For an F_2 population, we need to modify the statistical model (2) as

$$\xi = \mu + \sum_{j=1}^{m} a_j x_j + \sum_{j=1}^{m} d_j u_j + \sum_{r<t} (aa)_{rt} x_r x_t + \sum_{r\ne t} (ad)_{rt} x_r u_t + \sum_{r<t} (dd)_{rt} u_r u_t + \varepsilon, \tag{5}$$

where ξ, μ and ε have same explanations with those in model (2); a_j and d_j are respectively the additive effect and the dominant effect of the QTL in the jth interval in the F_2 design; $(aa)_{rt}$, $(ad)_{rt}$, $(dd)_{rt}$ $(r, t = 1, 2, \ldots, m)$ are respectively additive × additive, additive × dominant, dominant × dominant epistatic effects between QTL in the rth interval and the tth interval; x_j and u_j are the corresponding genotype variables for the additive and dominant effects. Define $x_j = u_j = 0$, if there is no QTL in the jth interval for an individual; otherwise,

$$x_j = \begin{cases} 1, & for\ Q_jQ_j, \\ 0, & for\ Q_jq_j, \\ -1, & for\ q_jq_j, \end{cases} \quad \text{and} \quad u_j = \begin{cases} -1/2, & for\ Q_jQ_j, \\ 1/2, & for\ Q_jq_j, \\ -1/2, & for\ q_jq_j, \end{cases}$$

are defined.

Similarly, by combining the threshold model (1), model (5) can be rewritten as the following generalized linear model