

A New Method for Calculating Molecular Genetic Similarity

Huijiang Gao¹, Runqing Yang², Wenzhong Zhao¹, Yuchun Pan²

1 School of Animal Science, Northeast Agriculture University, Harbin, Heilongjiang 150030, China

2 School of Agriculture and Biology, Shanghai Jiaotong University, Shanghai 201101, China

Abstract: Nei's genetic similarity is a similar coefficient for describing the difference of two binary variables, and it hasn't a connection with the relationship between individuals. According to the definition of the relationship coefficient, a new formula of genetic similarity is put forward as

$$r_{A(x,y)} = \frac{2N_{xy}^2}{N_x N_y \sqrt{(1+F_x)(1+F_y)}} \text{ or } r_{A(x,y)} = \frac{N_{xy}^2}{N_x N_y}. \text{ An example confirms that this calculating}$$

formula of genetic similarity is significantly better than Nei's on judging relationship between individuals. [Nature and Science. 2005;3(1):71-74].

Key Words: genetic similarity; relationship coefficient; verification

Introduction

The method for estimating genetic distance according to the polymorphism of genetic productions, as isozyme, blood type and leukocyte antigens, is increasingly replaced by DNA polymorphism. The most common methods for testing the DNA polymorphism include restrictive fragment length polymorphism (RFLP), variable number of tandem repeats (VNTR) and random amplified polymorphic DNA (RAPD), etc^[1]. It has been verified in many experience that the genetic purity in breed and genetic diversity between breeds can be determined effectively by proper statistical method according to the fingerprinting atlas of individual DNA. At present, the genetic similarity between individuals is scaled by Nei's formula, and the individuals' relationship is estimated from it. But this index is a kind of distance or similarity coefficient for describing the difference of two binary variables, and it has no certain connections with the relationship between individuals. So a new calculating formula of genetic similarity is preserved, according to the definition of relationship coefficient. An example proved that it is significantly better than Nei's on judging relationship between individuals.

1 Genetic Similarity

One can get an electrophoretic atlas by using

present molecular mark, and comparing them binately. Using 1-1 denotes having no polymorphism, namely, monomorphism, if both two parallel samples with the same molecular weight have all bands, and 1-0 denotes having polymorphism if only one sample. For estimating the genetic purity in a breed and comparing the difference between breeds, Nei defined the genetic similarity of any two individuals in 1979 as:

$$S = \frac{2N_{xy}}{N_x + N_y} \quad (1)$$

Here, N_{xy} is the sharing bands of two individuals, x and y; N_x and N_y represent the individual bands of x and y, respectively. The mean S in a breed reflects the similarity or different degree of DNA fingerprinting atlas accusing in this breed. When the mean genetic similarity matrix of multi-breeds is calculated, the tree derivation of relationship can be made through cluster analysis.

In fact, formula (1) is a kind of distance coefficient for describing the difference or similarity degree of two binary variables. As an index of relationship between individuals, the relationship coefficient is the frequency that two individuals are of the same genes from a common ancestor. According to the mark results, the common genes in both x and y has N_{xy} , genes in x has

N_x , in y has N_y , so the frequency that x has common gene is N_{xy}/N_x , that y has common gene is

N_{xy}/N_y , then the inbreeding coefficient of the progenies of two individuals, namely, the frequency that the progenies are of same gene from patents can be calculated by [5]:

$$F_{xy} = \frac{N_{xy}}{N_x} \times \frac{N_{xy}}{N_y} = \frac{N_{xy}^2}{N_x N_y} \quad (2)$$

Supposing an inbreeding coefficient of x and y is F_x and F_y respectively, then the relationship coefficient between x and y is:

$$r_{A(x,y)} = \frac{2F_{xy}}{\sqrt{(1+F_x)(1+F_y)}} \quad (3)$$

If x and y are not inbreeding individuals, that is $F_x = F_y = 0$, then formula (3) may be as follow:

$$r_{A(x,y)} = 2F_{xy} = \frac{2N_{xy}^2}{N_x N_y} \quad (4)$$

If x and y are inbreeding individuals, formula (2) and (3) show that F_{xy} may be bigger when F_x and F_y are bigger. If F_x and F_y are ignored, the relationship coefficient estimated by formula (4) may be over 1. So for assuring $0 \leq r_{A(x,y)} \leq 1$ and allowing for comparability of the genetic similarity between individuals when F_x and F_y are still unknown, formula (4) can be simplified as:

$$r_{A(x,y)} = \frac{N_{xy}^2}{N_x N_y} \quad (5)$$

Since $N_x + N_y \geq 2\sqrt{N_x N_y}$, so $\frac{2N_{xy}}{N_x + N_y} \leq$

$$\frac{N_{xy}}{\sqrt{N_x N_y}} = \sqrt{r_{A(x,y)}}, \quad \text{and } S \leq \sqrt{r_{A(x,y)}}.$$

When $N_x = N_y = N_{xy}$, namely, the bands of both individual are all the same, then $S = \sqrt{r_{A(x,y)}}$.

2 Example and Verification

Analyze the nucleus DNA genetics and variation of 9 Min pigs with a known relationship RAPD and ISSR respectively. After selecting 12 primers with polymorphism and separating amplification productions by electrophoresis, the fingerprinting atlas of RAPD including 97 amplification segments can be obtained. And 64 amplification segments can be tested by 8 ISSR primers and can obtain accordingly fingerprinting atlas of ISSR. To calculate the genetic similarities of any two individuals by formula (1) and (5), take them as the estimation of relationship coefficients between accordingly individuals and put the results in Tables 1 and 2.

Table 1. Genetic similarities between 9 individuals of Min pigs by RAPD

Pigs	1	2	3	4	5	6	7	8	9
1		0.6642	0.9321	0.7426	0.8726	0.6970	0.6321	0.7901	0.8875
2	0.5868		0.6254	0.8058	0.8546	0.7315	0.6532	0.9568	0.8198
3	0.8867	0.6831		0.6825	0.9458	0.8846	0.9114	0.8974	0.6142
4	0.6717	0.6501	0.4744		0.8456	0.7745	0.9271	0.6452	0.8954
5	0.8232	0.7825	0.9015	0.8032		0.6983	0.9019	0.7249	0.9358
6	0.6228	0.7674	0.8484	0.7376	0.7712		0.9238	0.9010	0.6888
7	0.6185	0.5532	0.8470	0.8603	0.8233	0.8714		0.6247	0.7542
8	0.6285	0.9352	0.8245	0.5716	0.5511	0.8501	0.4542		0.9215
9	0.7888	0.6821	0.8521	0.7488	0.8965	0.5488	0.7232	0.8683	

Table 2. Genetic similarities between 9 individuals of Min pigs by ISSR

Pigs	1	2	3	4	5	6	7	8	9
1		0.7412	0.9218	0.8125	0.8612	0.7158	0.8412	0.9123	0.5486
2	0.6369		0.9123	0.5486	0.5649	0.7315	0.7456	0.9546	0.9218
3	0.8642	0.8443		0.8821	0.7864	0.7489	0.8787	0.9215	0.9141
4	0.7835	0.8321	0.5512		0.8541	0.8742	0.7154	0.5478	0.9147
5	0.7982	0.8343	0.6972	0.7777		0.9412	0.9123	0.8326	0.8519
6	0.8625	0.7446	0.8446	0.7974	0.9124		0.9356	0.9159	0.8745
7	0.6064	0.6733	0.7878	0.5512	0.8344	0.8882		0.6415	0.7542
8	0.7332	0.9357	0.8623	0.5488	0.7232	0.8611	0.6484		0.7058
9	0.8514	0.8505	0.8483	0.8476	0.7505	0.8500	0.6676	0.8634	

To evaluate the relationship and difference between the genetic similarities different formulas in Tables 1 and 2 and the relationship coefficient between accordingly individuals in Table 3 by rank correlation and mean absolute error. It can be obtained by calculating the rank correlation between the S in Table 1 and the relationship coefficient between accordingly individuals in Table 3 is 0.1907 ($P > 0.05$), their mean absolute error is 0.4113; the rank correlation between the $r_{A(x,y)}$ in Table 1 and the relationship coefficient between accordingly individuals in Table 3 is 0.4821 ($P < 0.01$), the mean absolute error is 0.3486; the rank

correlation between the S in Table 2 and the relationship coefficient between accordingly individuals in Table 3 is 0.3411 ($0.01 < P < 0.05$), the mean absolute error is 0.4290; the rank correlation between the $r_{A(x,y)}$ in Table 2 and the relationship coefficient between accordingly individuals in Table 3 is 0.6093 ($P < 0.01$), the mean absolute error is 0.3486. It shows that $r_{A(x,y)}$ by RAPD and ISSR are both significantly better than Nei's on judging relationship between individuals.

Table 3. The relationship and inbreeding coefficients of 9 individuals

Pigs	1	2	3	4	5	6	7	8	9
1	0.0557	0.3586	0.3462	0.4143	0.2415	0.5432	0.3259	0.1598	0.4158
2		0.0625	0.5872	0.2548	0.4872	0.7214	0.3541	0.6841	0.2147
3			0.0901	0.0320	0.1478	0.3214	0.4325	0.6741	0.5874
4				0.1462	0.4857	0.4754	0.0576	0.2174	0.3214
5					0.0876	0.5417	0.3541	0.0147	0.4874
6						0.0434	0.6487	0.6547	0.2514
7							0.1066	0.1147	0.3258
8								0.0543	0.6841
9									0.0885

Limited by outlay it only allows for a few animals and primers in verifying the validity of the new formula, but doesn't study the effect of the number of animals and primers on the validity of the new formula. It needs further research on this point.

Correspondence to:

Runqing Yang

School of Agriculture and Biology

Shanghai Jiaotong University

Shanghai 201101, China

E-mail: runqingyang@sjtu.edu.cn

References

- [1] Chao Yongxin, et al. The protection of animal variety [M] Beijing: Agricultural Science Press of China. 1995:111-24.
- [2] Lynch M. The similarity index and DNA fingerprinting[J]. Mol Bio Evol 1990;7:478-80.
- [3] Zhang Raoting, Fang Kaitiai. Introduction of multivariate statistical analysis [M] Beijing: Science Press. 1999:393-401.
- [4] Bai Liping. The fingerprinting analysis of random amplification polymorphic DNA of pigs [Master thesis]. Haerbin: The Animal Science Department of Northeast Agriculture University. 1996.
- [5] Sheng Zhilian, Wu Changxin. Quantitative genetics [M] Beijing: Agriculture Press of China. 1995:28-4