Journal of Animal and Veterinary Advances 11 (10): 1622-1626, 2012
ISSN: 1680-5593

Codon-Substitution Simulation Models for Detecting Molecular Adaptive Evolution

Lihua Fan, Min Tao, Shujun Wang and Chengyu Hu
College of Life Science and Food Engineering, Nanchang University,
330031 Nanchang, China

Abstract: Branch-Site and Site-Specific Models are statistical simulation methods for detecting molecular adaptive evolution at individual among specific lineages. The models not only indicate whether genes of phylogeny are under positive selection or not but also can identify the codon sites which are under the positive selection promoting gene divergence and polymorphism. For protein-coding gene, detecting the positive selection sites is important to understand the structure and function of proteins. Interferon is a very important cytokine in innate immune system. Interferon genes present evolutional polymorphism when the innate immune factors are under high selective pressure. Using the computer models to simulate and analyze the molecular adaptive evolution of interferon genes, it is found that interferon genes are under the positive selection and the positive selection sites existing in the interferon gene sequences are also located. These results show that Branch-Site Models and Site-Specific Models can detect molecular adaptive evolution of innate immune genes which are highly divergent.

Key words: Sequence, models, cytokin, protien-coding, gene, phylogeny

INTRODUCTION

For protein-coding genes, the codons under selective pressure have three evolvement ways: the negative selection, neutral evolution and positive selection (Goldman and Yang, 1994). The neutral evolution sites don't suffer the selection effect because basically their displacements don't affect the protein's structure and function. The negative selection sites are important to maintain the natural structure of protein and would influence natural function of protein when these sites are replaced, so the sites have a low codon substitution rate. However, the positive selection can endue a protein with a new structure or function that is in favor of individual subsistence and reproduce and the sites have a high codon substitution rate. Thus, the detecting and identifying of positive selection sites is important to understand the structure and function of proteins (Yuan and Duan, 2003).

The Interferon (IFN) is a very important cytokine in innate immune system of host (Medzhitov and Janeway, 1998) models to analyze the rmolecular adaptive evolution of mammalian interferon gene families' Open Reading Frames (ORF). Interferon has the function of anti-virus activate cellular antivirus and defense the infection of diversiform homologous and heterogenous viruses. Besides, interferon can regulate other physiological processes of physical cells such as cell growth and differentiation, cell apoptosis and physical and cellular immune reaction (Stark *et al*., 1998; Sen, 2001). The corporate evolution of hosts and viruses under the high selective pressure would result in interferon genes and produce present polymorphism and evolvement. Thus, interferon genes and their coding proteins are excellent materials and candidate markers for investigating non-synonymous substitution and selection effect of genes under the selective pressure (Murphy, 1993).

Now Darwin's Theory of Evolution by Natural Selection is generally accepted by biologists but the importance of natural selection in molecular evolution has ever been a matter of debate for a long time (Sharp, 1997). The neutral theory states that most observed molecular variation both polymorphism within species and divergence between species is due to random fixation of selectively neutral mutations (Yang and Bielawski, 2000). Based on this theory, a series of statistical models to detect molecular adaptive evolution had been proposed. By using maximum likelihood models, Kao and Lee (2002) found that the difference in phosphoglucose isomerases of hagfish, zebrafish, gray mullet, toad and snake could well reflect principle of their molecular evolvement and species evolution. Xin *et al*. (2005) analyzed the nucleotide acid molecular evolution of uncoding region

Corresponding Author: Lihua Fan, College of Life Science and Food Engineering, Nanchang University, 330031 Nanchang, China

and packing protein coding region of HIV-1 and obtained well-forecasted effect. However, there are few researches on molecular adaptive evolution of immune genes or immune interrelated genes among species, due to their biggest divergence. Using traditional statistical models to analyze the immune or immune interrelated genes could misestimate the data and then bring many difficulties to the forecast and analysis.

Branch-Site and Site-Specific Models are statistical simulation methods used to detect selection effect for the single codon site along gene sequences (Yang and Bielawski, 2000; Yang and Nielsen, 2002). The models not only calculate the non-synonymous substitution rate (dN) and synonymous substitution rate (dS) at the codon level but also the transition/transversion ratio of codon-substitution and estimate the codon frequency from the analyzed sequences and don't fix the substitution rate to simply a numeral. Thus, Branch-Site and Site-Specific Models could be applied to detect the molecular adaptation of immune genes or immune-related genes. Researchers used these two models to analyze the molecular adaptation of mammalian interferon gene families' Open Reading Frames (ORF). The results show that both *IFN-α* and *IFN-γ* genes are suffered from positive selection effect and identify the positive selection sites in the gene sequences. Branch-Site Simulation Models can detect well the molecular adaptation of innate immune genes which have the biggest evolutional divarication in certain area.

MATERIALS AND METHODS

Models and parameters: Branch-Site Models include two types of models: Model A and B. The models considered there are different ω rates among the codon sites and the sequences have four kinds of codon sites. In Model A, fix ω_0 = 0 and ω_1 = 1 whereas in Model B they are free parameters b the sequences data (Yang *et al*., 1998). The Model B has 5 free parameters (P_0, P_1, ω_0, ω_1, ω_2) and ω_2>1 when the foreground branch under positive selection (Table 1). However, the Model M3 (one of the Site-Specific Models divides the sites into two classes, the ω_0 and ω_1. Thus, the Model M3 only has three free parameters (P_0, ω_0, ω_1) which is two less than its extension Model B and has ω_1>1 when the foreground branch under positive selection. In this study, researchers used Model M3 and its extension Model B to analyze the open reading frame codons of interferon genes. In addition, the M0 Model which has only one free parameter has been used as a reference. The estimated parameters in the three models such as the transition/transversion ratio, proportion P_0 and P_1 and ω ratio by used Maximum Likelihood (ML) approach. After parameters estimated by ML are obtained an empirical Bayes Method will be applied to infer which site is be the most likely form (Yang and Bielawski, 2000).

Table 1: Parameters in the branch-site models

Site classes	Proportion	Background	Foreground
0	P_0	ω_0	ω_0
1	P_1	ω_1	ω_1
2	$P_1 = (1 - P_0 - P_1)\ P_0/(P_0+P_1)$	ω_0	ω_2
3	$P_1 = (1 -P_0 - P_1)\ P_1/(P_0+P_1)$	ω_1	ω_2

In Model A fix $\omega_0 = 0$ and $\omega_1 = 1$ whereas in Model B both ω_0 and ω_1 are free parameters

Interferon gene sequences: Both mammalian *IFN-α* and *IFN-α* gene sequences were downloaded from the GenBank (gene names and gene register numbers see the phylogenic tree).

Analytic methods: Researchers constructed the phylogeny of interferon genes and then endued the branch under positive selection with a priori value. For example, in the analysis about *IFN-α* gene family, researchers are interested in testing whether positive selection has occurred along the lineages. For convenience, researchers refer to branches for which we test positive selection as the foreground branches and all others the background branches (Yang *et al*., 1998).

RESULTS AND DISCUSSION

Analysis to positive selection sites of mammalian IFN-α

Phylogeny of mammalian IFN-α: Researchers constructed phylogeny of mammalian IFN-α using public reported gene sequences. They are 12 Human (Hs) IFN-α sequences, 10 Mouse (Mm) IFN-α sequences, 4 dog (Cf) IFN-α sequences, 2 cat (Fc) IFN-α sequences, 1 cattle (Bt) IFN-α sequence and 1 pig (Ss) IFN-α sequence. Figure 1 shows that gene tree of IFN-α reflects the evolutionary relations among species.

Analysis to evolvement of mammalian IFN-α: Compared with Model B and Model M3, the LRT statistic is 24 l = 2 [(−5674.92)−(−5684.11)] = 18.38 with df = 2 and $p<0.01$. So, Model B fits the mammalian IFN-α data better than model M3 (in 1% level). Model M3 didn't find out remarkable positive selection sites, yet Model B obtained positive selection sites 130 Q and 132 L with posterior probability >95% (Table 2). Besides the two sites, there are eight sites which posterior probabilities are <90% but >50%. They are 60 G and 155 R (p>0.90), 12A, 33R, 42V, 110N, 179A and 183A (p>0.50) and the eight sites could be considered as positive selection sites in certain extent.

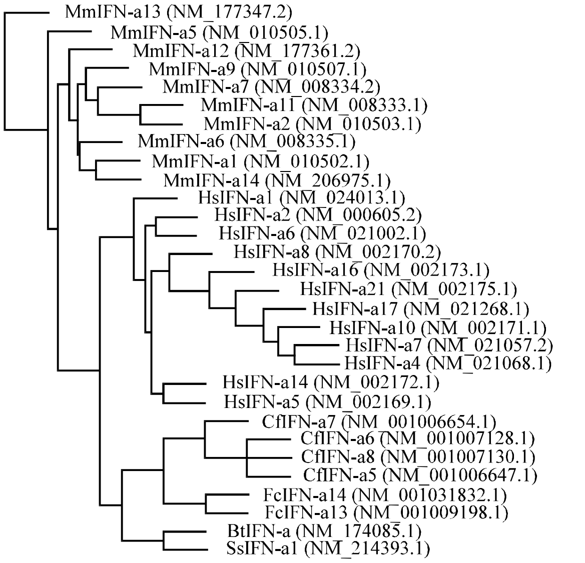

Fig. 1: Phylogeny of mammalian IFN-α. Mm IFN-α 6 is foreground branch for under the positive selection

Table 2: Parameters estimate of mammalian IFG-α gene data

Model	P	L	Estimates of parameters	Positively selected sites
M0:one-ratio	1	-5741.33	$\omega = 0.577$	None
M3:discrete (κ = 2)	3	-5684.11	$P_0 = 0.476$ $P_1 = 0.524$ $\omega_0 = 0.214$, $\omega_1 = 0.966$	None
Model B	5	-5674.92	$P_0 = 0.087$, $P_1 = 0.346$ $P_2+P_3 = 0.567$, $\omega_0 = 0.267$, $\omega_1 = 0.796$, $\omega_2 = 2.081$	130Q, 132L, p>0.95

P is the number of free parameters for the ω parameters indicating positive selection are presented in bold-type. Those in parentheses are presented for clarity only but are not free parameters for example, under M3, P_0, ω_0 and ω_1 are free parameters, $P_1 = 1-P_0$. Estimates of κ range from 2.41-2.43 among models. Sites potentially under positive selection are identified using the mouse Mm IFN-α 6 sequence as the reference

Not only the foreground branch (mouse *IFN-α 6* gene) of mammalian IFN-α suffered from positive selection effect but also the background branch (such as human IFN-α 1). The results account for interferon suffered great strong positive selective pressure during the evolutional process.

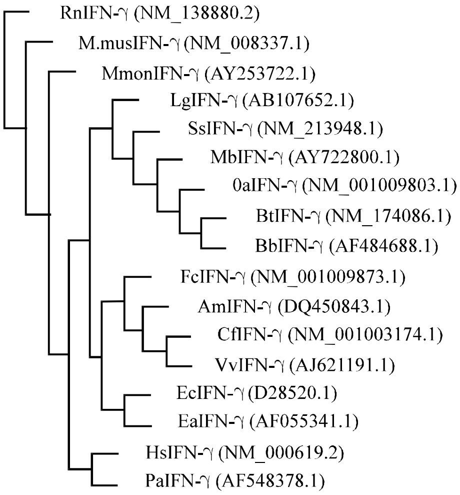

Fig. 2: Phylogeny of mammalian IFN-γ. Hs IFN-γ is foreground branch for under the positive selection

Table 3: Parameters estimate of mammalian *IFN-γ* gene data

Model	P	L	Estimates of parameters	Positively selected sites
M0:one-ratio	1	-3531.57	$\omega = 0.481$	None
M3:discrete (κ = 2)	3	-3491.62	$P_0 = 0.685$ $P_1 = 0.314$ $\omega_0 = 0.257$, $\omega_1 = 1.198$	See text
Model B	5	-3485.72	$P_0 = 0.084$, $P_1 = 0.167$ $P_2+P_3 = 0.749$, $\omega_0 = 0.071$, $\omega_1 = 0.479$, $\omega_2 = 1.828$	14I, 102V 110K, 113D, p>0.95 88s p>0.99

P is the number of free parameters for the ω ratios. Estimates of κ range from 2.79-2.90 among models. Sites potentially under positive selection are identified using the human Hs IFN-γ sequence as the reference

Analysis to positive selection sites of mammalian IFN-γ

Phylogeny of mammalian IFN-γ: Interferon γ is the only member of type interferon also known as immune interferon. Researchers searched 17 interferon *γ* gene sequences from the GenBank. They are Human (Hs), baboon (Pa), mule (Ea), horse (Ec), fox (Vv), dog (Cf), panda (Am), cat (Fc), buffalo (Bb), cattle (Bt), sheep (Oa), musk deer (Mb), pig (Ss), guanaco (Lg), guinea pig (Mmon), mouse (Mm) and rat (Rn) interferon *α* gene sequences. Researchers constructed phylogeny of mammalian IFN-γ using these public reported sequences (Fig. 2). Figure 2 shows that gene tree of IFN-γ reflects the homologous and evolutional relations among species.

Analysis to evolvement of mammalian IFN-γ: Compared with Model B and M3, the LRT statistics 2l = 2 [(−3485.72) −(−3491.62)] = 11.8 with df = 2 and p<0.01. So, Model B fits the mammalian IFN-γ data better than Model M3 (in 1% level) (Table 3). The following remarkable positive selection sites are identified by Model M3: 6Y, 26P, 50T, 83F, 109K, 112R, 146A,157M, 158L (p>0.95); 14I, 36K, 42H, 43S, 46A, 88S, 102V,110K, 113D, 116E (p>0.99). Whereas, the positive selection sites calculated by Model B are 14I, 102V, 110K, 113D,116E (p>0.95); 88S (p>0.99). In the result of Model B, all the sites identified by Model M3 have ω>1 and their posterior probabilities are >0.95, so these sites aren't remarkable positive selection sites.

Although, Model B fits the mammalian IFN-γ data better than Model M3 in the statistical aspect, Model B not always has better suitable advantage than Model M3 in the biological truth because of the high selective pressure. Under this situation, the sites identified by Model M3 could be treated as the positive selection sites of mammalian IFN-γ.

Predicting the positive selection sites could provide theoretical reference for further research on gene function. The forecast of positive selection sites in the interferon genes has some theory guidance meaning for understanding the structure and function of interferon protein such as on the artificial mutation research of interferon and evolvement analysis including polymorphism within species and divergence among species. Previous studies had usually employed a pairwise approach, calculating synonymous (dS) and non-synonymous (dN) rates between two sequences by averaging over all codons in the gene and over the period that separates the sequences and used the measure ω (ωdN = dS) to indicate selective pressure at the protein level. A ω ratio greater than one means that non-synonymous mutations offer fitness advantages and are fixed in the population at a higher rate than synonymous mutations. Positive selection can thus be detected by identifying cases where ω>1. But many amino acids in a functional protein might be largely invariable (with ω close to 0) because of strong structural and functional constraints; the average dN has commonly little power in detecting positive selection (Yang and Nielsen, 2002).

The Site-Specific Models allow the ω ratio to vary among sites but not among lineages. Positive selection is detected at individual sites only if the average dN over all lineages is higher than the average dS. If adaptive evolution occurs at a few time points and affects a few amino acids, the models might lack power in detecting positive selection (Yang and Nielsen, 2002; Crandall *et al*., 1999). Branch-Site Models as a class of simulation models have more power to analyze the coding protein DNA sequences by the maximum likelihood approach. The Branch-Site Models allow the ω ratio to vary among both sites and lineages and so could improve the power of the Likelihood Ratio Test (LRT) to detect positive selection along pre-specified lineages. A major use of the models is to analyze the evolution of gene families or molecular adaptation of genes.

CONCLUSION

In this study, the results showed that interferon genes sufffered a strong positive selection effect. For instance, many remarkable positive selection sites were found out in the mammalian IFN-α, especially in the *IFN-γ* gene sequences. These results account for interferon genes under the molecular adaptation evolution caused by the evolutional pressure. The positive Darwin selection effect of interferon genes is rmuch stronger than common housekeeping genes (Hurst and Smith, 1999). On the other hand, compared with Model M3 and B could not always detect more positive selection sites (Table 3). That may be caused by that different models used different probability criterions. For example, Yang, etc. used the probability criterion of p>0.5 as they analyzed the angiosperm phytochrome gene and considered some site is a positive selection site if its posterior probability >0.5 (Yang and Nielsen, 2002). In this study, researchers used model M3 to analyze human *IFN-α* gene family and obtained more significant positive selection sites than Model B, the sites were 103D, 177L and 184R (p>0.95). However, in Model B parameters of the three codons were 103D (p = 0.932; ω = 2.338), 177L (p = 0.945; ω = 2.366) and 184R (p = 0.947; ω = 2.370). Although, their posterior probability is <0.95, it could still be considered that they were positive selection sites when the probability criterion decreased to 90%.

The gene sequences analyzed in this study are open reading frame sequences that code the interferon proteins and they directly reflect selective pressure suffered by the interferon. The analysis didn't involve uncoding regions such as untranslation regions, promoter, introns, etc. Later studies indicate the uncoding regions of functional genes also exist selective effect (Larizza *et al*., 2002; Michael *et al*., 2004) reported that 5'-terminal uncoding region of CTGF exist positive selection sites which have significant posterior probability (Xin *et al*., 2005). So, if analyzinge the complete interferon gene sequences further, several new positive selection sites would be identified that reflect characteristic and principle of interferon evolvement. IFN-β secreted from the fibroblast and compared with IFN-α has 40% homologous degree. In this study we also used Model M3 and B to analyze molecular evolvement of six IFN-β sequences downloaded from GenBank and obtained two more different results (Model M3 identified more positive selection sites than Model B). It appears to show that codon substitution simulation models would well in effect when analyzinge a lot of isogenous sequences but might produce error (make mistake) when analyzinge few sequences.

The codon-substitution simulation models based on ω ratio >1 are strict to analyze positive selection along gene sequences, so sometimes they fail to detect molecular adaptation. The same failure occurs when the number of samples is small. In addition, codon substitution models seem no impact on positive selection effect of uncoding sequences which can regulate gene transcription. As the localizations of the codon-substitution simulation models on application, it is necessary to develop a new kind of simulation model based on current models to detect molecular adaptation.

ACKNOWLEDGEMENTS

This research was supported by the grants from the National Natural Science Foundation of China (30560116) and the Jiangxi Provincial Department of Education Plan (20070042).

REFERENCES

Crandall, K.A., C.R. Kelsey, H. Imamichi, H.C. Lane and N.P. Salzman, 1999. Parallel evolution of drug resistance in HIV: Failure of nonsyn-onymous/ synonymous substitution rate ratio to detect positive selection. Mol. Biol. Evol., 16: 372-382.

Goldman, N. and Z.A. Yang, 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol. Biol. Evol., 11: 725-736.

Hurst, L.D. and N.G. Smith, 1999. Do essential genes evolve slowly. Curr. Biol., 9: 747-750.

Kao, H.W. and S.C. Lee, 2002. Phosphoglucose isomerases of Hagsh, Zebrash, Gray Mullet, Toad, and Snake, with reference to the evolution of the genes in vertebrates. Mol. Biol. Evol., 19: 367-374.

Larizza, A., W. Makalowski, G. Pesole and C. Saccone, 2002. Evolutionary dynamics of mammalian mRNA untraslated regions by comparative analysis of orthologous humans, artiodactyls and rodent gene pairs. Compt. Chem., 26: 479-490.

Medzhitov, R. and C.A. Jr. Janeway, 1998. An ancient system of host defence. Curr. Opin. Immunol., 10: 12-15.

Michael, H., S. Fang and C.L. Wu, 2004. Inference of positive and negative selection on the 5' regulatory regions of *Drosophila genes*. Mol. Biol. Evol., 21: 374-383.

Murphy, P.M., 1993. Molecular mimicry and the generation of host defense protein diversity. Cell, 72: 823-826.

Sen, G.C., 2001. Viruses and interferons. Ann. Rev. Microbiol., 55: 255-281.

Sharp, P.M., 1997. In search of molecular darwinism. Nature, 385: 111-112.

Stark, G.R., I.M. Kerr, B.R. Williams, R.H. Silverman and R.D. Schreiber, 1998. How cells respond to interferons. Annu. Rev. Biochem., 67: 227-264.

Xin, L., C. Hong and W. Wen, 2005. A new method for detecting natural selection at the level of nucleotide sites. Zool. Res., 26: 225-229.

Yang, Z. and R. Nielsen, 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specic lineages. Mol. Biol. Evol., 19: 908-917.

Yang, Z., R. Nielsen and M. Hasegawa, 1998. Models of amino acid sub-stitution and applications to mitochondrial protein evolution. Mol. Biol. Evol., 15: 1600-1611.

Yang, Z.H. and J.P. Bielawski, 2000. Statistical methods for detecting molecular adaptation. Trends Ecol. Evol., 15: 496-503.

Yuan, Z. and C.Z. Duan, 2003. Recent progress of sequences analysis methods in molecular evolutionary biology. Chin. Bull. Bot., 20: 462-468.