BIOINFORMATICS APPLICATIONS NOTE Vol. 18 no. 6 2002
Pages 886–887

Synonymous–non-synonymous mutation rates between sequences containing ambiguous nucleotides (Syn-SCAN)

Matthew J. Gonzales[1], Jonathan M. Dugan[2] and Robert W. Shafer[1,*]

[1]Division of Infectious Diseases and [2]Stanford Medical Informatics, Stanford University Medical Center, 300 Pasteur Drive, Room S-156, Stanford, CA 94305, USA

Received on August 23, 2001; revised on November 22, 2001; January 7, 2002; accepted on January 9, 2002

ABSTRACT

Summary: Direct PCR sequencing on genetic material containing allelic mixtures results in sequences containing ambiguous nucleotides. Because codons exhibiting allelic mixtures present evidence of evolutionary pressure, it is important to include this information in the assessment of codon synonymy. We developed a program, 'Synonymous–Nonsynonymous Mutation Rates between Sequences Containing Ambiguous Nucleotides' (Syn-SCAN), that calculates synonymous and non-synonymous substitution rates using a model that includes allelic mixtures.

Availability: Syn-SCAN is implemented on the web and can be downloaded from http://hivdb.stanford.edu.

Contact: rshafer@cmgm.stanford.edu

Highly polymorphic RNA viruses such as human immunodeficiency virus type 1 (HIV-1) and hepatitis C exist within individuals as a quasispecies of heterogeneous yet closely related genomes (Martell *et al.*, 1992; Coffin, 1995). Although clonal virus sequencing can determine the genetic sequence for individual members of a virus quasispecies, direct-PCR 'population-based' sequencing is increasingly used because of its ability to detect nucleotide mixtures and lower cost. When direct PCR sequencing is done on genetic material containing allelic mixtures, the resulting sequence contains ambiguous nucleotides, such as R (A/G) and M (A/C).

Nucleotide substitutions that cause an amino acid change are non-synonymous; those that do not are synonymous. The ratio of non-synonymous to synonymous substitutions in a protein-coding gene reflects the relative influence of positive selection and neutral evolution. Several methods have been developed to estimate the numbers of synonymous and non-synonymous substitutions between two sequences and programs based on these methods are used often (e.g. MEGA (Kumar *et al.*, 2000), SNAP (Korber, 2000)). These programs, however, ignore codons with allelic mixtures.

Because codons with ambiguous nucleotides caused by allelic mixtures are likely to be undergoing more rapid evolution than codons without mixtures, we developed a program, Syn-SCAN, that calculates synonymous and non-synonymous substitution rates using a model that includes genetic mixtures. In this model, a virus population containing a single nucleotide (e.g. A) at a position is evolutionarily closer to a population containing a mixture of A and a second nucleotide (e.g. A/G = R) than to a population containing a different nucleotide (G). Such partial differences often indicate that the virus population within an individual is changing, particularly when the second nucleotide has emerged during selective antiretroviral drug pressure (Wei *et al.*, 1995).

Syn-SCAN requires that input sequences are multiply aligned and positioned in the appropriate reading frame. The numbers of potential synonymous (S) and non-synonymous (N) substitutions per sequence are calculated by iterating through each codon in a sequence using a hash table with the number of potential synonymous substitutions for each of the 64 non-ambiguous codons (Figure 1a). Codons containing ambiguous nucleotides are broken down into their component mixtures and S and N are determined by averaging the potential for synonymous and non-synonymous substitutions for each component.

The numbers of synonymous (S_d) and non-synonymous (N_d) differences between two sequences are calculated by iterating through each pair of aligned codons in two sequences. When differences between codons lacking ambiguous nucleotides are encountered, the extent of synonymy is determined using the hash table with the number of synonymous and non-synonymous changes between any two codons (Figure 1b). When differences

*To whom correspondence should be addressed.

(a)

Codon	Syn Potential
AAA	1
AAT	1
AAC	1
AAG	1
.	.
.	.
ATA	2
ATC	2
ATG	0
ATT	2
.	.
.	.
GGA	3
GGC	3
GGG	3
GGT	3
.	.
.	.
CTA	4
CTC	3
CTG	3
CTT	3

(b)

Codon	Codon	No. Syn Changes	No. Nonsyn Changes
AAA	AAA	0	0
AAA	AAC	0	1
AAA	AAG	1	0
AAA	AAT	0	1
.	.	.	.
.	.	.	.
.	.	.	.
CCC	ACA	2	2
CCC	CCA	1	0
CCC	GCA	2	2
.	.	.	.
.	.	.	.
.	.	.	.
CCC	GGA	3	9
CCC	GGC	0	4
.	.	.	.
.	.	.	.
.	.	.	.
GGG	AAA	3	9
GGG	AAG	0	4

(c)

Nucleotide	Nucleotide	Score
A	A	0
A	C	1
A	G	1
A	T	1
A	R	0.5
A	Y	1
A	W	0.5
A	S	1
A	M	0.5
A	K	1
.	.	.
.	.	.
R	A	0.5
R	C	1
R	G	0.5
R	T	1
R	R	0
R	Y	1
R	W	0.75
R	S	0.75
R	M	0.75
R	K	0.75

Fig. 1. Three data structures used by Syn-SCAN. Table 1 has 64 entries containing the number of potential synonymous substitutions for each of the non-ambiguous codons. Table 2 has 4096 entries containing the number of synonymous and non-synonymous changes between any two codons. Table 3 has 225 entries containing nucleotide distance scores between each of the ambiguous and non-ambiguous nucleotides. The contents of Table 3 are modified at run-time based on user defined preferences. syn, synonymous, nonsyn, non-synonymous.

between codons with ambiguous nucleotides are encountered, the nucleotide substitution matrix containing both ambiguous and unambiguous nucleotides (Figure 1c) is used to modify the extent of synonymy obtained from the hash table in Figure 1b.

The proportion of synonymous (p_S) substitutions per sequence comparison is obtained by dividing S_d by the number of potential synonymous sites (S). The proportion of non-synonymous (p_N) substitutions per sequence comparison is obtained by dividing N_d by the number of potential non-synonymous sites (N). The synonymous (d_S) and non-synonymous distances (d_N) are calculated by applying the Jukes–Cantor correction for back-mutation. The program output contains each of the distance measurements and text files containing matrices of d_S and d_N values in a format suitable for analysis by phylogenetic programs. Syn-SCAN is written in Perl and runs in Windows and Unix environments.

Syn-SCAN generates a nucleotide substitution matrix at run-time based on a user-selected weighting scheme. The default weighting assigns a distance between two ambiguous nucleotides and between an ambiguous and non-ambiguous nucleotide that is proportional to the extent of ambiguity (1- to 4-fold) of each of the nucleotides and inversely proportional to the number of shared nucleotides (i.e. R and M share one nucleotide, A). This weighting scheme is recommended because it accounts for the fact that when mixtures are present, a change at a nucleotide position may result from a change in the proportion of two competing populations rather than from a new mutation. To examine the results that would be generated by other programs that calculate synonymous–non-synonymous mutation rates, users have the option of ignoring partial differences.

There are two online implementations of Syn-SCAN. The first accepts sequences of any protein-coding gene. The second accepts paired HIV-1 sequences tested for drug resistance. Sample data sets, as well as other published sequence data sets (Condra *et al.*, 1996; Bacheler *et al.*, 2000) indicate that mutations selected during antiretroviral drug therapy proceed through an intermediate stage in which both wildtype and mutant residues are present. Syn-SCAN provides genetic distance estimates that take this intermediate stage into account making the program a unique tool for quantitative studies of intra-host virus evolution.

REFERENCES

Bacheler,L.T. and Anton,E.D. *et al.* (2000) Human immunodeficiency virus type 1 mutations selected in patients failing efavirenz combination therapy. *Antimicrob Agents Chemother*, **44**, 2475–2484.

Coffin,J.M. (1995) HIV population dynamics in vivo: implications for genetic variation, pathogenesis, and therapy. *Science*, **267**, 483–489.

Condra,J.H. and Holder,D.J. *et al.* (1996) Genetic correlates of in vivo viral resistance to indinavir, a human immunodeficiency virus type 1 protease inhibitor. *J. Virol.*, **70**, 8270–8276.

Korber,B. (2000) HIV signature and sequence variation analysis. In Rodrigo,A.G. and Learn,G.H. (eds), *Computational and Evolutionary Analysis of HIV Molecular Sequences*. Kluwer, Dordrecht, pp. 55–72.

Kumar,S. *et al.* (2000) *MEGA: Molecular Evolutionary Genetics Analysis, ver 2*, Pennsylvania State University, University Park and Arizona State University, Tempe.

Martell,M. *et al.* (1992) Hepatitis C virus (HCV) circulates as a population of different but closely related genomes: quasispecies nature of HCV genome distribution. *J. Virol.*, **66**, 3225–3229.

Wei,X. *et al.* (1995) Viral dynamics in human immunodeficiency virus type 1 infection. *Nature*, **373**, 117–122.