GENE CONVERSION AND THE INFINITE-SITES MODEL

Stanley A. Sawyer
Washington University

Abstract

The infinite-sites model assumes that there have been only simple point mutations since the common ancestor of a sample of aligned DNA sequences, and in particular no recombination or gene conversion. A statistical test for detecting a history of gene conversion within a sample of aligned DNA sequences is applied to various data sets. The results show strong evidence for multiple intragenic conversion events at two loci in the bacterium *E. coli* and in a repeated family in maize. The data suggest that the rate of these short conversion events may be much larger than the neutral mutation rate per base. Thus the basic assumptions of the infinite-sites model may often be violated in natural data sets. The same test applied to a gene in an RNA virus and to a sample of mitochondrial DNA did not show gene conversion.

Introduction. The basic genetic material or DNA of plants and animals can be thought of as a set of strings of letters from an alphabet {T, C, A, G}, with one string for each chromosome. Thus to a geneticist a mouse is the same as a tomato, since both have about the same amount of DNA. A gene or genetic locus is a segment of a chromosome that affects some trait. Experiments beginning in the middle 1960's showed that many enzymes were unexpectedly polymorphic; i.e., their genetic loci had genes of many different types. A protracted and sometimes heated debate arose about the selective significance of these polymorphisms. Some biologists felt that most of these observed enzyme polymorphisms were selectively neutral or nearly neutral, while others asserted that different enzymes were unlikely to be selectively equivalent, and so various forms of balancing selection must be involved (Kimura, 1968; Wills, 1973; see also Hartl, 1989; Hartl and Sawyer, 1991). In an attempt to resolve this controversy, Ewens (1972) derived a formula for the expected distribution of genes among allelic classes in a random sample from a population that is selectively neutral at that genetic locus. The infinite-alleles model in all of its various forms can be considered an extension of Ewens' (1972) work. The vast number of recent papers on the infinite-alleles model by Donnelly, Ethier, Griffiths, Hoppe, Hudson, Kaplan, Kurtz, Tavare, Watterson, and others shows the health and vigor of current mathematical research in this field.

The infinite-alleles model assumes an equilibrium selectively-neutral random-mating population subject to mutation, where each new mutant gene is of a type that is entirely new to the population. In 1972, most data on enzyme polymorphisms came from protein electrophoresis, which allows genes to be distinguished but gives no further information about them. Since 1983, DNA sequences of genes at polymorphic loci have become available. Foreshadowing this experimental advance, Watterson (1975) considered the infinite-alleles model where the alleles are DNA sequences with the assumption that each new mutation changes a single base (i.e., one of the letters T,C,A,G) at a previously undisturbed site. This model is now known as the infinite-sites model. An important feature of the infinite-sites model is that the pedigree of a sample can be read off from the DNA sequences if the sequence of the common ancestor is known (Felsenstein 1982, p380). One can often use "outgroups" to find information about the DNA sequence of the common ancestor. For example, if DNA sequences are known for seven squirrels and a hedgehog at a particular genetic locus, and if the squirrel sequences are polymorphic at a site for which one of the bases T,C,A,G also occurs in the hedgehog at that site, then the base in the hedgehog is probably in the common ancestor of both species. However, the assumption that all mutations since the common ancestor of the sample were simple point mutations at distinct sites is crucial. One of the purposes of this article is to show that this assumption may often be violated.

One way in which a multiple-base mutation can occur is if a recombination or crossover event occurs within a gene. This produces a gene composed of parts of the two parental genes, and is likely to be new to the population if the parental genes were distinct or if the partial genes were misaligned. Humans have about one crossover event per chromosome pair for each offspring, and the chromosomes are on the average about 10^8 bp (base pairs) long. The point mutation rate in higher animals is about 5×10^{-9} per site per year (Li *et al.* 1985), or about 10^{-7} per generation per site in humans. Thus perhaps about 10% of mutations in humans are due to crossover events within genes. However, the distribution of crossover points is highly nonuniform, and the ratio for particular genes may be either much less or much more than 10% .

A more important type of non-simple-point mutation is *gene conversion* within a gene. Gene conversion is any process that copies the bases from one segment of DNA to another segment of the same length, and can occur by a variety of mechanisms. The source and target DNA segments can be on the same chromosome, or else they can be on different chromosomes in the same individual or in different individuals. Table 1 illustrates a gene conversion event where the source and target DNA segments are 15bp long. Note that only the 4 bases that are flagged actually change in the target sequence. In observed data, this event would appear as a cluster of four point mutations. Since the new bases at these sites are the same as in the source sequence, the mutations are not independent. If the source and target segment happened to be identical, no change would occur. Gene mutations that are definitely due to gene conversion have been identified in humans (Slightom *et al.* 1980), mice (Weiss *et al.* 1983), and in many other creatures as well. Nagylaki (1988), Ohta (1986), Watterson (1989), and Griffiths and Watterson (1990) are a partial list of recent theoretical papers on gene conversion.

Table 1: An Illustration of Gene Conversion

```
······ctttaagcgcat GCAGAAGGCTTTAACcgaatgatcggt······
                      ↓ ↓   ↓ ↓
······attgtgtttatt GCACATGGATCTAAC tagtgccggtga······
```

Many mechanisms for gene conversion depend on a combination of spontaneous DNA hybridization and mismatch repair. For example, the two DNA strands in each of the two DNA segments in Table 1 may accidentally separate, and then one strand from each of the two segments may temporarily bind together (or "hybridize") by the same chemical forces that normally bind the two strands in a single DNA segment. Enzymes that normally correct mismatches between the two strands in normal DNA may then force the second strand to be identical with the first. In general, the mismatch repairs may not all be in the same direction, and not all mismatches may be repaired, but Table 1 may occur. If the two mismatched strands then separate and re-attach with their original strands, and if the mismatch repair enzymes copy the introduced changes to the other DNA strand in the target DNA segment, then the changes illustrated in Table 1 will result.

Genes usually have an "upstream" control region followed by one or more "coding regions" and perhaps a trailing "downstream" control region. Triples of bases in coding regions form *codons* that specify unique amino acids. Proteins and enzymes are composed of chains of amino acids. Since there are 64 possible codons and only 20 amino acids, the mapping cannot be one-to-one, with non-uniqueness typically in the third position of codons. Table 2 gives the first 36 bases (and the resulting amino acids) in the coding region of a typical gene for the enzyme 6-phosphogluconate dehydrogenase (abbreviated "*gnd*") in the bacterium *Escherichia coli.* Note that the amino acid glycine is represented in Table 2 by two different codons (GGC and GGT), and phenylalanine by two codons TTT and TTC. Since TTC and TTT are the only codons for phenylalanine, the third positions of these codons are called *two-fold degenerate* or *two-fold silent sites* . Since all four codons GGx code for glycine, the third positions of those codons are called *four-fold degenerate* or *four-fold silent*. There is only one three-fold degenerate amino acid, isoleucine (Ile). We implicitly assume that silent changes in DNA are selectively neutral, since the same protein is encoded. Although selection can act in various ways on silent DNA differences, the average effect of selection on configurations at silent sites appears to be small (Sawyer *et al.* 1987; Hartl and Sawyer, 1990).

Table 2: The Initial Codons and Amino Acids in *gnd*

GCA	GAA	GGC	TTT	AAC	TTC	ATT	GGT	ACC	GGT	GTT	TCC. .
Ala	Glu	Gly	Phe	Asn	Phe	Ile	Gly	Thr	Gly	Val	Ser. . . .

We next describe a statistical test for a pattern of gene conversion in a sample of DNA sequences that looks for nonrandom pairwise concordances at silent sites. We then describe some data sets for which this test is extremely significant for apparent within-locus gene conversion.

Statistical Tests for Gene Conversion. Given an aligned sample of DNA sequences, a gene conversion event will tend to produce a segment in which two sequences are identical. If the segment is long, and if there has been no subsequent mutations in the segment, the set of polymorphic sites that distinguish the two sequences may have a highly nonuniform distribution. Under various assumptions one can assign a conservative P-value for this apparent gene conversion by classical probabilistic arguments (Stephens 1985). However, if the segments undergoing gene conversion are short, or if they once were long but have since been altered by mutation, then individual gene conversion events will not be detectable by this approach. The procedure we describe below is a permutation test that is designed to detect the result of gene conversion, while not necessarily being able to detect specific occurrences (Sawyer 1989).

In general, permutation tests have three basic components. The first is a null hypothesis H_0, in common with all statistical tests. In our case, this is the assumption that there has been no recombination or gene conversion since the common ancestor of the sample. The distribution of bases at silent polymorphic sites would then be determined by independent neutral mutation acting on the same pedigree at all sites. The test we consider below will depend only on the silent polymorphic sites and their relative order along the sequenced region. It is useful to think of this data as an $n \times s$ matrix of bases T, C A, G, where n is the number of aligned sequences and s is the number of silent polymorphic sites. Given H_0, the columns in this matrix have a statistical distribution depending only on the ancestral base and the degeneracy k at a particular site. If we relabel the bases at each site as 1, ..., k where the base relabeled as i forms the i^{th} most common nucleotide at that site, the columns at silent sites will have a distribution given H_0 depending only on the degeneracy k.

The second component of permutation tests is a class of permutations G of the data preserving H_0. More precisely, we assume that the data are the result of observations of random variables whose joint distribution given H_0 is not changed by the permutations in G. The class G in our case will consist of all permutations of the columns of the $n \times s$ matrix that permute silent polymorphic sites of the same degeneracy among themselves. (We treat silent polymorphic sites of irregular degeneracy as a single class for simplicity.) If H_0 is correct, the observed data set should appear as a typical member of the class of permuted data sets corresponding to the permutations in G. If the observed data set is atypical, we would suspect hidden structure inconsistent with H_0. The class G is intended to break up a pattern of short pairwise-conserved segments that may not be immediately apparent.

The third component of a permutation test is a "score"; i.e., a statistic applied to possible data sets that should be small if H_0 is correct, but large if H_0 fails in some way that we wish to detect. The P-value of the permutation test is defined as the probability that the data set generated by a randomly chosen permutation from G will have a score larger than or equal to the observed score. Since G will usually be too large for the P-value to be calculated exactly, we will approximate the P-value by using 10,000 randomly chosen permutations from G.

The score we choose is as follows. If two of the n sequences are compared, they will differ in a set of $d \leq s$ silent polymorphic sites. These d sites partition the array of silent polymorphic sites into $d + 1$ segments, which we will call the *condensed fragments* determined by this partition ("condensed" because we have thrown out the non-silent-polymorphic sites). Each condensed fragment corresponds to a segment from the sequenced region in which the two sequences are identical, except within amino acid polymorphic codon positions, and which is bounded either by two of the d discordant silent polymorphic sites or by one of the discordant sites and one end of the sequenced region. If x_i is the length of the i^{th} condensed fragment, then $\Sigma x_i = s - d$. Each of the $n(n - 1)/2$ pairs of sequences defines a set of $d_k + 1$ condensed fragments, where $1 \leq k \leq n(n -1)/2$. The score we will use is the sum of the squares of the condensed fragment lengths (abbreviated SSCF), which is defined as the double sum $\Sigma\Sigma\ x_i^2$ with summation over the $n(n - 1)/2$ pairs of sequences in the outer sum, and over the $d_k + 1$ condensed fragments determined by the k^{th} sequence pair in the inner sum. Thus if $d_k = d$ for all k, the sum defining SSCF would have $(d + 1)n(n - 1)/2$ terms. The score SSCF is one of four similar scores defined in Sawyer (1989).

Applications of the SSCF Test. Results for a number of DNA sequence data sets are shown in Table 3. The first three numerical columns in Table 3 are the number n of aligned sequences, the number s of silent polymorphic sites, and the length of the sequenced region in bases. The first two rows are for two enzymes in *Escherichia coli* (*phoA* is an abbreviation for alkaline phosphatase). The observed score is highly significant for both *E. coli* genetic loci. Table 3 also gives the distance between the observed score and the mean of the permuted scores in terms of the standard deviation of the permuted scores. While this distribution is not normal, the fact that the observed value of SSCF for *gnd* is 7.21 standard deviations above the mean is consistent with the fact that the largest of 10,000 simulated values of SSCF was still well below the observed value. An analysis of the distribution of uncondensed fragment lengths shows a significant excess of fragments of length 70-200bp for *gnd* ($P < 10^{-3}$), and of 190bp and longer for *phoA* ($P \approx 0.03$). While SSCF was highly significant for both *gnd* and *phoA*, only the largest uncondensed fragment was significant by itself for *gnd*, and no fragments were significant for *phoA*, using a related test that applies to single fragments (Sawyer 1989). These observations suggest a pervasive pattern of intragenic recombination within both the *gnd* and *phoA* loci (see also DuBose *et al*, 1988, for *phoA*). None of the six largest uncondensed fragments in *gnd* were bounded by an endpoint of the sequenced region, which suggests a pattern of either short gene conversion events or else large gene conversion events punctuated by later mutation.

The test was nonsignificant when applied to a simulated *gnd* data set with mutation only. When 80 simulated gene conversion events for segments of 50-200bp were added to the simulated *gnd* data set, the results approximated those of the observed *gnd* data set (Table 3). This corresponds to a rate of gene conversion about 20 times that of point mutation per base. The SSCF test was also nonsignificant when applied to 13 strains of human influenza A virus. The human influenza data was not expected to show gene conversion because (1) the rapid mutation rates of RNA retroviruses such as influenza may

Table 3: SSCF Test Results for Gene Conversion

	Number of				(Obs.-Mean)[a]
Data Set	Seqs.	Polys.	Bases	P-value	/ S. D.
E. coli (*gnd* locus[b])	7	81	768	0	7.21
E. coli (*phoA* locus[c])	8	61	1413	0	5.33
Sim. *gnd*, mut. only[d]	7	78	768	0.2719	0.54
Sim. *gnd*, mutation and gene conversion[d]	7	83	768	0	7.46
Influenza type A[e]	13	52	678	0.6295	-0.38
Cytochrome c[f]	9	*52*	*82*	0.0966	1.36
Human globins[g]	3	36	1592	0	12.58
Maize heterochromatin[h]	11	36	180	0.0003	4.71
Primate mitochondria[i]	5	283	896	0.9625	-1.61

[a] Mean and S.D. are the mean and standard deviation of the scores for 10,000 random permutations of the data.

[b] Sawyer *et al.* (1987) [c] DuBose *et al.* (1988) [d] Sawyer (1989) [e] Buonagurio et al. *(*1986*)* [f] Sneath *et al.* (1975) [g] Slightom *et al.* (1980) [h] Dennis and Peacock (1984) [i] Brown *et al.* (1982)

obliterate any evidence for gene conversion, and (2) the virus strains were gathered at different times over a 50 year period. All polymorphic sites can be permuted (and used as potential discordant sites) instead of silent polymorphic sites, with similar results in most cases. However, when this test is used with all polymorphic sites, a significant result may be due to parallel selection or some other selective effect rather than gene conversion. When applied to nine bacterial protein sequences of 82 amino acid bases each, the result approached significance ($P \approx 0.097$; see Table 3). Since these bacteria are thought to be too distantly related for gene conversion to occur, the near-significance is most likely due to parallel evolution in portions of the sequenced region. The human globin entry in Table 3 relates to a single apparent gene conversion between two loci on the same human chromosome.

Dennis and Peacock (1984) have DNA sequences for 11 repeats of a 180bp segment in knob heterochromatin, which is a family with up to 10^6 tandem repeats in maize. Since these repeats are adjacent and similar, between-repeat gene conversion is expected by spontaneous hybridization. The SSCF test behaves as expected (Table 3). Gene conversion between mitochondria in different animals is thought to be quite rare. Consistent with this popular wisdom, the SSCF test detects no gene conversion in a sample of mitochondrial DNA sequences from seven primates (Table 3). See Sawyer (1989) and Hartl and Sawyer (1990) for more examples and for more biological detail about the examples in Table 3.

References

[1] Brown, W., E. Prager, A. Wang, and A. Wilson (1982). Mitochondrial DNA sequences of primates: tempo and mode of evolution. *J. Mol. Evol.* **18**, 225-239.

[2] Buonagurio, D., S. Nakada, J. Parvin, M. Krystal, P. Palese, and W. Fitch (1986). Evolution of human influenza A viruses over 50 years: rapid, uniform rate of change in NS gene. *Science* **232**, 980-982.

[3] Dennis, E. and W. Peacock (1984). Knob heterochromatin homology in maize and its relatives. *J. Mol. Evol.* **20**, 341-350.

[4] DuBose, R., D. Dykhuizen, and D. Hartl (1988). Genetic exchange among natural isolates of bacteria: recombination within *the phoA* locus of *Escherichia coli. Proc. Nat. Acad. Sci. USA* **85**, 7036-7040.

[5] Ewens, W. (1972). The sampling theory of selectively neutral alleles. *Theoret. Population Biol.* **3,** 87-112.

[6] Felsenstein, J. (1982). Numerical methods for inferring evolutionary trees. *Quarterly Review of Biology* **57**, p379-404.

[7] Griffiths, R. and G. Watterson (1990). The number of alleles in multigene families. To appear in *Theoret. Population Biol.*, **37**, p 110-123.

[8] Hartl, D. (1989). Evolving theories of enzyme evolution. *Genetics* **122**, 1-6.

[9] Hartl, D. and S. Sawyer (1991). Inference of selection and recombination from nucleotide sequence data. In press, *Jour. Evol. Biol.*

[10] Kimura, M. (1968). Evolutionary rate at the molecular level. *Nature* **217**, 624-626.

[11] Li, W.-H., C.-I. Wu, and C.-C. Luo (1985). A new method of estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Mol. Biol. Evo*l. **2**, 150-174.

[12] Nagylaki, T. (1988). Gene conversion, linkage, and the evolution of multigene families. *Genetics* **120**, 291-301.

[13] Ohta, T. (1986). Actual number of alleles contained in a multigene family. *Genet. Res.* **48**, 119-123.

[14] Sawyer, S., D. Dykhuizen, and D. Hartl (1987). Confidence interval for the number of selectively neutral amino acid polymorphisms. *Proc. Nat. Acad. Sci. USA* **84**, 6225-6228.

[15] Sawyer, S. (1989). Statistical tests for detecting gene conversion. *Mol. Biol. Evol.* **6**, 526-538.

[16] Slightom, J., A. Blechl, and O. Smithies (1980). Human fetal G_γ and A_γ globin genes: complete nucleotide sequences suggest that DNA can be exchanged between these duplicated genes. *Cell* **21**, 627-638.

[17] Sneath, P., M. Sackin, and R. Ambler (1975). Detecting evolutionary incompatibilities from protein sequences. *Systematic Zoology* **24**, 311-332.

[18] Stephens, J. (1985). Statistical methods of DNA sequence analysis: detection of intragenic recombination or gene conversion. *Mol. Biol. Evol.* **2**, 539-556.

[19] Watterson, G. (1975). On the number of segregating sites in genetical models without recombination. *Theoret. Population Biol.* **7**, 256-276.

[20] Watterson, G. (1989). Allele frequencies in multigene families. II. Coalescent approach. *Theoret. Population Biol.* **35**, 161-179.

[21] Weiss, E., L. Golden, R. Zakut, A. Mellor, K. Fahrner, S. Kvist, and R. Flavell (1983). The DNA sequence of the $H\text{-}2K^b$ gene: evidence for gene conversion as a mechanism for the generation of polymorphism in histocompatibility antigens. *EMBO J.* **2**, 453-462.

[22] Wills, C. (1973). In defense of naive pan-selectionism. *Amer. Naturalist* **107**, 23-34.