Patterns of nucleotide and amino acid substitution

Introduction

So I've just suggested that the neutral theory of molecular evolution explains quite a bit, but it also ignores quite a bit.[1] The derivations we did assumed that all substitutions are equally likely to occur, because they are selectively neutral. That isn't plausible. We need look no further than sickle cell anemia to see an example of a protein polymorphism in which a single amino acid difference has a very large effect on fitness. Even reasoning from first principles we can see that it doesn't make much sense to think that all nucleotide substitutions are created equal. Just as it's unlikely that you'll improve the performance of your car if you pick up a sledgehammer, open its hood, close your eyes, and hit something inside, so it's unlikely that picking a random amino acid in a protein and substituting it with a different one will improve the function of the protein.[2]

The genetic code

Of course, not all nucleotide sequence substitutions lead to amino acid substitutions in protein-coding genes. There is redundancy in the genetic code. Table 1 is a list of the codons in the universal genetic code.[3] Notice that there are only two amino acids, methionine and tryptophan, that have a single codon. All the rest have at least two. Serine, arginine, and leucine have six.

Moreover, most of the redundancy is in the third position, where we can distinguish 2-fold from 4-fold redundant sites (Table 2). 2-fold redundant sites are those at which either one

[1] I won't make my bikini joke. I'll save that for when we get to quantitative genetics. Nonetheless, the "pure" version of the neutral theory of molecular evolution makes a *lot* of simplifying assumptions.

[2] Obviously it happens sometimes. If it didn't, there wouldn't be any adaptive evolution. It's just that, on average, mutations are more likely to decrease fitness than to increase it.

[3] By the way, the "universal" genetic code is not universal. There are at least eight, but all of them have similar redundancy properties.

Codon	Amino Acid	Codon	Amino Acid	Codon	Amino Acid	Codon	Amino Acid
UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys
UUC	Phe	UCC	Ser	UAC	Tyr	UGC	Cys
UUA	Leu	UCA	Ser	UAA	Stop	UGA	Stop
UUG	Leu	UCG	Ser	UAG	Stop	UGG	Trp
CUU	Leu	CCU	Pro	CAU	His	CGU	Arg
CUC	Leu	CCC	Pro	CAC	His	CGC	Arg
CUA	Leu	CCA	Pro	CAA	Gln	CGA	Arg
CUG	Leu	CCG	Pro	CAG	Gln	CGG	Arg
AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser
AUC	Ile	ACC	Thr	AAC	Asn	AGC	Ser
AUA	Ile	ACA	Thr	AAA	Lys	AGA	Arg
AUG	Met	ACG	Thr	AAG	Lys	AGG	Arg
GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly
GUC	Val	GCC	Ala	GAC	Asp	GGC	Gly
GUA	Val	GCA	Ala	GAA	Glu	GGA	Gly
GUG	Val	GCG	Ala	GAG	Glu	GGG	Gly

Table 1: The universal genetic code.

Codon	Amino Acid	Redundancy
CCU	Pro	4-fold
CCC		
CCA		
CCG		
AAU	Asn	2-fold
AAC		
AAA	Lys	2-fold
AAG		

Table 2: Examples of 4-fold and 2-fold redundancy in the 3rd position of the universal genetic code.

of two nucleotides can be present in a codon for a single amino acid. 4-fold redundant sites are those at which any of the four nucleotides can be present in a codon for a single amino acid. In some cases there is redundancy in the first codon position, e.g, both AGA and CGA are codons for arginine. Thus, many nucleotide substitutions at third positions do not lead to amino acid substitutions, and some nucleotide substitutions at first positions do not lead to amino acid substitutions. But every nucleotide substitution at a second codon position leads to an amino acid substitution. Nucleotide substitutions that do not lead to amino acid substitutions are referred to as *synonymous substitutions*, because the codons involved are synonymous, i.e., code for the same amino acid. Nucleotide substitutions that do lead to amino acid substituions are *non-synonymous substitutions*.

Rates of synonymous and non-synonymous substitution

By using a modification of the simple Jukes-Cantor model we encountered before, it is possible make separate estimates of the number of synonymous substitutions and of the number of non-synonymous substitutions that have occurred since two sequences diverged from a common ancestor. If we combine an estimate of the *number* of differences with an estimate of the *time of divergence* we can estimate the rates of synonymous and non-synonymous substitution (number/time). Table 3 shows some representative estimates for the rates of synonymous and non-synonymous substitution in different genes studied in mammals.

Two very important observations emerge after you've looked at this table for awhile. The

Locus	Non-synonymous rate	Synonymous rate
Histone		
H4	0.00	3.94
H2	0.00	4.52
Ribosomal proteins		
S17	0.06	2.69
S14	0.02	2.16
Hemoglobins & myoglobin		
α-globin	0.56	4.38
β-globin	0.78	2.58
Myoglobin	0.57	4.10
Interferons		
γ	3.06	5.50
$\alpha 1$	1.47	3.24
$\beta 1$	2.38	5.33

Table 3: Representative rates of synonymous and non-synonymous substitution in mammalian genes (from [11]). Rates are expressed as the number of substitutions per 10^9 years.

first won't come as any shock. The rate of non-synonymous substitution is generally lower than the rate of synonymous substitution. This is a result of my "sledgehammer principle." Mutations that change the amino acid sequence of a protein are more likely to reduce that protein's functionality than to increase it. As a result, they are likely to lower the fitness of individuals carrying them, and they will have a lower probability of being fixed than those mutations that do not change the amino acid sequence.

The second observation is more subtle. Rates of non-synonymous substitution vary by more than two orders of magnitude: 0.02 substitutions per nucleotide per billion years in ribosomal protein S14 to 3.06 substitutions per nucleotide per billion years in γ-interferon, while rates of synonymous substitution vary only by a factor of two (2.16 in ribosomal protein S14 to 5.50 in γ interferons. If synonymous substitutions are neutral, as they probably are to a first approximation,[4] then the rate of synonymous substitution should equal the mutation rate. Thus, the rate of synonymous substitution should be approximately the same at every locus, which is roughly what we observe. But proteins differ in the degree to which

[4]We'll see that they may not be completely neutral a little later, but at least it's reasonable to believe that the intensity of selection to which they are subject is less than that to which non-synonymous substitutions are subject.

their physiological function affects the performance and fitness of the organisms that carry them. Some, like histones and ribosomal proteins, are intimately involved with chromatin or translation of messenger RNA into protein. It's easy to imagine that just about any change in the amino acid sequence of such proteins will have a detrimental effect on its function. Others, like interferons, are involved in responses to viral or bacterial pathogens. It's easy to imagine not only that the selection on these proteins might be less intense, but that some amino acid substitutions might actually be favored by natural selection because they enhance resistance to certain strains of pathogens. Thus, the probability that a non-synonymous substitution will be fixed is likely to vary substantially among genes, just as we observe.

Revising the neutral theory

So we've now produced empirical evidence that many mutations are *not* neutral. Does this mean that we throw the neutral theory of molecular evolution away? Hardly. We need only modify it a little to accomodate these new observations.

- *Most non-synonymous substitutions are deleterious.* We can actually generalize this assertion a bit and say that most mutations that affect function are deleterious. After all, organisms have been evolving for about 3.5 billion years. Wouldn't you expect their cellular machinery to work pretty well by now?

- *Most molecular variability found in natural populations is selectively neutral.* If most function-altering mutations are deleterious, it follows that we are unlikely to find much variation in populations for such mutations. Selection will quickly eliminate them.

- *Natural selection is primarily purifying.* Although natural selection for variants that improve function is ultimately the source of adaptation, even at the molecular level, most of the time selection is simply eliminating variants that are less fit than the norm, not promoting the fixation of new variants that increase fitness.

- *Alleles enhancing fitness are rapidly incorporated.*[5] They do not remain polymorphic for long, so we aren't likely to find them when they're polymorphic.

As we'll see, even these revisions aren't entirely sufficient, but what we do from here on out is more to provide refinements and clarifications than to undertake wholesale revisions.

[5]To be more precise I should have written *Alleles enhancing fitness are rapidly incorporated,* **when they are not lost quickly as a result of genetic drift**.

References

[1] The International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449(7164):851–861, 2007.

[2] J C Fay and C.-I. Wu. Hitchhiking under positive Darwinian selection. *Genetics*, 155:1405–1413, 2000.

[3] Y X Fu. Statistical properties of segregating sites. *Theoretical Population Biology*, 48:172–197, 1995.

[4] Y.-X. Fu. Statistical tests of neutrality of mutations against population growth, hitchhiking, and background selection. *Genetics*, 147:915–925, 1997.

[5] Feng Guo, Dipak K Dey, and Kent E Holsinger. A Bayesian hierarchical model for analysis of SNP diversity in multilocus, multipopulation samples. *Journal of the American Statistical Association*, 104(485):142–154, March 2009.

[6] J L Hubby and R C Lewontin. A molecular approach to the study of genic heterozygosity in natural populations. I. The number of alleles at different loci in *Drosophila pseudoobscura*. *Genetics*, 54:577–594, 1966.

[7] M Kreitman. Nucleotide polymorphism at the alcohol dehydrogenase locus of *Drosophila melanogaster*. *Nature*, 304:412–417, 1983.

[8] M Kreitman and M Aguadé. Excess polymorphism at the alcohol dehydrogenase locus in *Drosophila melanogaster*. *Genetics*, 114:93–110, 1986.

[9] M Kreitman and R R Hudson. Inferring the evolutionary history of the *Adh* and *Adh-dup* loci in *Drosophila melanogaster* from patterns of polymorphism and divergence. *Genetics*, 127:565–582, 1991.

[10] R C Lewontin and J L Hubby. A molecular approach to the study of genic heterozygosity in natural populations. II. Amount of variation and degree of heterozygosity in natural populations of *Drosophila pseudoobscura*. *Genetics*, 54:595–609, 1966.

[11] W.-H. Li. *Molecular Evolution*. Sinauer Associates, Sunderland, MA, 1997.

[12] F Tajima. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, 123:585–595, 1989.

[13] K Zeng, Y.-X. Fu, S Shi, and C.-I. Wu. Statistical tests for detecting positive selection by utilizing high-frequency variants. *Genetics*, 174:1431–1439, 2006.

Creative Commons License