Cytosine Deamination and Evolution

Posted on [February 22, 2009](https://designmatrix.wordpress.com/2009/02/22/cytosine-deamination-and-evolution/) | [Leave a comment](https://designmatrix.wordpress.com/2009/02/22/cytosine-deamination-and-evolution/#respond)

*It has been argued that no engineer would have used cytosine as part of the genetic material because of its predisposition for deamination. But it’s exactly this predisposition that might cause an engineer of evolution to include it.*

Life itself appears to have been designed to minimize errors. The universal nature of the proof-reading/repair machinery, the optimized genetic code, and the G/C:A/T parity code all converge on this point. Yet despite this design logic, there is the interesting fact that cytosine is especially prone to deamination, where the removal of its exocyclic amino group converts it into uracil (a base normally found in RNA). Uracil does not exist in DNA, thus it can be effectively detected and removed by repair enzymes. However, if not detected and repaired, it can base pair with adenine, meaning that it would specify adenine during DNA replication. In a subsequent round of replication, the adenine in turn would specify thymine. The bottom line is that spontaneous deamination of cytosine can lead to a base substitution known as a transition, where C is replaced by T (and G is replaced by A on the other strand of DNA). We might expect such mutations to be quite common, as the rate constant for cytosine deamination at 37 degree C in single stranded DNA translates into a half-life for any specific cytosine of about 200 years. In fact, such high rates of deamination led researchers Poole et. al to complain of “confounded cytosine!” [1]

We would thus seem to have two contradictory lines of evidence. On one hand, there is the growing list of evidence to support the hypothesis that error correction was an important principle guiding the design of life. Yet the incorporation of cytosine works against such efforts, given its predisposition to spark a mutation. In fact, Poole et al. go so far as to argue, “Any engineer would have replaced cytosine, but evolution is a tinkerer not an engineer.” From a design perspective, how might these contrary dynamics be reconciled? That is, given the emphasis on error correction, why would an engineer include cytosine?

The first possible explanation may center around necessity. Put simply, there is no better choice of four different nucleotides. This can be seen simply by considering that the exocyclic amino group prone to removal also participates in the hydrogen-bonding of GC pairs. In other words, it is part of the parity code (it is the first ‘1’ in the four-bit ‘100,1’ code for C).

A second possible explanation was that cytosine was chosen *because* of its predisposition to undergo deamination. This explanation may also intersect with the hypothesis of necessity, as a good designer often finds ways to turn a “design problem” into an opportunity. In this case, let me propose that cytosine, far from being something any engineer would replace, may actually have played an instrumental role in the front-loading of evolution. Put simply, C-to-T transitions, as a function of deamination, may have posed a form of “direction” on evolution.

One way to appreciate this is to consider the mutagenic effects of deamination among the four nucleotides used by DNA. A single-stranded DNA molecule with 2 million bases will experience a single deamination event involving cytosine every 2.8 hours (at pH 7.4 and 37 degree C). In contrast, it would take 140 hours for an adenine to experience deamination. Given that guanine deaminates at rates similar to adenine, and thymine lacks an exocyclic base, and thus experiences no deamination, we can see that the simple process of deamination would strongly favor cytosine as a target (figure 1). In other words, targeting for deamination is clearly not random.

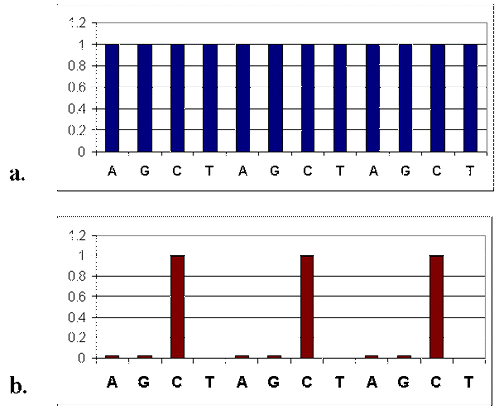


Figure 1. Likelihood of any nucleotide being targeted for deamination along AGCTAGCTAGCT strand of DNA. X-axis shows the hypothetical DNA sequence. Y-axis is normalized to cytosine deamination events per 2.8 hours in ssDNA at 37 degree C and pH 7.4. 1a shows hypothetical events where each nucleotide is equally prone to mutations. 1b shows same sequence experiencing measured rates of deamination. G deamination events were assumed equivalent to A.

Does this physical bias exert itself in living things? Yes. Table I shows the relative mutation rates involving each of the four nucleotides measured in 16 different mammalian pseudogenes [2]. Since pseudogenes have no known function, we can eliminate selection as a contributing factor for any pattern we see.

Table I. Relative substitution rates.

ORIGINAL NUCLEOTIDE

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Mutant Nucleotide | A | T | C | G |
| A | – | 4.4 | 6.5 | 20.7 |
| T | 4.7 | – | 21.0 | 7.2 |
| C | 5.0 | 8.2 | – | 5.3 |
| G | 9.4 | 3.3 | 4.2 | – |

Clearly, the C-T transitions (with their coupled G-A transitions) are the most frequently occurring base substitutions. In figure 2, the average substitution rate for each of the four nucleotides is shown as it would be experienced along a hypothetical sequence of ATCGGATGAATCGATCT (red bars). Note that when CG-TA transitions are removed (blue bars), no bias appears and the relative mutation rate would look more like figure 1a.

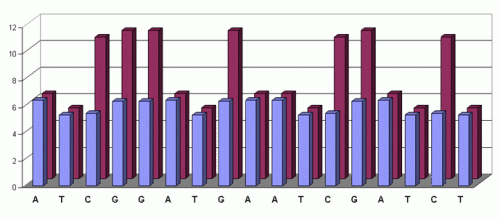


Figure 2. Average substitution rates for each of the four nucleotides with (red) C-T transitions and without (blue) C-T transitions.

One might note that the relative frequency of C-T transitions is not as dramatic as one might expect from figure 1 (where cytosines experience approximately 50 deaminations for every one experienced by adenine). This reflects the activity of the DNA repair enzyme, uracil glycosylase, which detects and eliminates uracil residues inserted into DNA. The possible teleological significance of this enzyme will be discussed below.

This bias for C-T transitions has also been documented in several studies. For example, 31 spontaneous mutations from the APRT locus from Chinese hamster ovary cells were analyzed [3]. Of these, 27 were single base substitutions, where 22/27 were C-T transitions. A similar bias was also seen in an extensive survey of a bacterial gene, where 86% of the base substitutions among 85 spontaneous mutations in *lacI* were C-T transitions. [4]

Given the inherent bias toward C-T transitions, we can safely conclude that base substitution mutations are not random among the four nucleotides G, C, T, and A. From a teleological perspective, this provides a mechanism by which a designer could impose some form of quasi-direction on evolution. My working hypothesis is that the original cells were designed and such design entailed a degree of front-loading, in which certain evolutionary trajectories were made more likely. The initial designed state would translate as a front-loaded state given that evolution borrows rather than invents *de novo*. That is, evolution could “unpack” buried designs. The bias toward C-T transitions could thus be used to exploit this front-loaded state. For example, protein X could be designed to fulfill function A. But buried in the sequence of X is the potential for function B (that is, function B is nearby in sequence space relative to sequence X/function A). A gene duplication, followed by exposure to the C-T mutational “stream,” could unlock function B, if the original sequence X was specifically designed to be unlocked by such mutations.

It is difficult to test such a hypothesis, as the unfolding of this front-loaded state may have been limited such that it did not reach very far from the original state. Nevertheless, the fact that the genetic code has been essentially frozen since its design may provide generic clues to the type of evolution that was/is built into life. Put simply, given that the genetic code itself was designed to minimize deleterious mutations, perhaps it was likewise designed to exploit the evolutionary potential of C-T transitions.

There are two primary points of focus where C-T transitions may exert their effect. The first revolves around genomic organization, as a drift in the C-to-T direction may play a role in modifying the genome’s system architecture. Such large scale genomic changes may even influence the proteaome [5]. In future essays, I shall speculate along this point of focus.

The second design principle that may be involved focuses on the small-scale coding effects of C-T transitions. I count 37 different codons that contain cytosine. A single C-T transition expands this set to 48 codons (since some of the original codons had two or more C’s). Of the 48 substitutions, 17 are silent (not changing the amino acid sequence). Among the 31 changes, 4 introduce premature stop codons (nonsense mutations) and 27 result in amino acid changes (missense mutations). The codon pool containing at least one cytosine codes for ten of the twenty amino acids, while the pool resulting from the 27 C-T transitions contains nine amino acids, *where the mutant pool is almost twice as hydrophobic as the original* (Table II).

Table II. Amino acids coded by C-containing codons before and after C-T transitions. Hydrophobicity values are from [6] and reported in parentheses.

|  |  |
| --- | --- |
| C-codon amino acids | C-T codon amino acids |
| Leu (.943) | Phe (1.0) |
| Ser (.359) | Leu (.943) |
| Pro (.711) | Ser (.359) |
| Thr (.450) | Ile (.943) |
| Tyr (.880) | Met (.738) |
| His (.165) | Tyr (.88) |
| Glu (.043) | Cys (.68) |
| Gln (.251) | Trp (.878) |
| Arg (.000) | Val (.825) |
| Ala (.616) |  |
| AVG. – (.433) | AVG. – (.835) |

If the amino acids are weighted to reflect how often they are coded for (due to the degeneracy of the code), the average hydrophobicity value increases from 0.442 to 0.806. Since 19 of the 23 missense mutations would involve such a change, this indicates that 83% of the peaks in the hypothetical sequence from Figure 2 would be sampling from a hydrophobic pool. Even more remarkable than the near doubling of the pool’s hydrophobicity is the fact that *each amino acid substitution*, with the exception of Pro -> Ser, is accompanied with an increase in hydrophobicity (Figure 3).

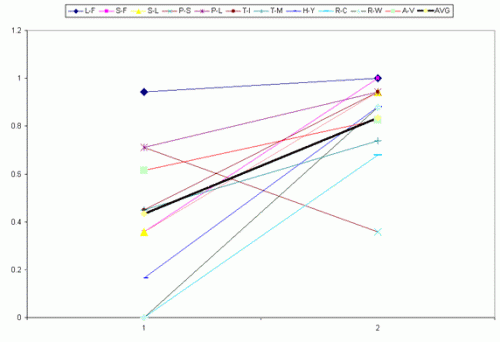


Figure 3. Changes in amino acid hydrophobicity resulting from C-T transitions on the coding strand. 1 represents the C-containing containing codons, while 2 represents the mutant codons. The black line in the center represents the average change.

Note from Figure 3 that the original hydrophobic values are spread from a range of 0 to 0.943. If we exclude the Pro -> Ser substitution (see below), the new pool is much more clustered, ranging from 0.68 to 1.0. In fact, if we arrange the 20 amino acids in terms of increasing hydrophobicity, it becomes clear that the C-T transitions are sampling almost exclusively from the most hydrophobic amino acids (Figure 4).

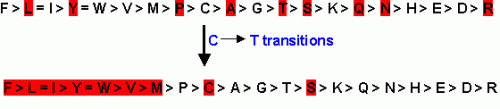


Figure 4. Amino acids arranged according to hydrophobicity [6]. Amino acids highlighted in red are coded for by C-containing condons (top row) and C-T transition effect codons (bottom row).

Given that the genetic code has been optimized to prevent deleterious mutations, it is quite noteworthy that the subset of mutations stemming from C-T transitions (as a consequence of deamination of C) result in some fairly radical substitutions. If we employ the Gonnet Pam250 matrix, only 4 of the 27 substitutions involve switches among the “strong” group of amino acids (L -> F; H -> Y) and another 8 occur among the “weak” group (A -> V; P ->S). If we include the P->S substitutions and consider the other 15 substitutions, and apply the Chou & Fasman rules used to predict protein structure, it is apparent that the substitutions swap from a pool that is largely indifferent to secondary structure replacing it with a pool that is often involved in alpha-helix and beta-strand formation (Figures 5, 6).

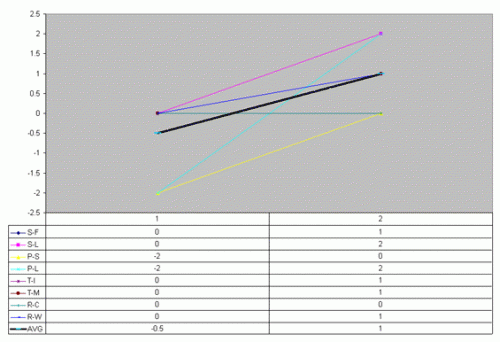


Figure 5. The effect of C-T transitions on alpha-helix formation. Numerical values were assigned to residues based on their frequency of occurrence of each amino acid in alpha helices (as per Chou & Fasman). 2= strong former; 1 = former; 0 = indifferent; -1 = breaker; -2 = strong breaker. Change in avg values shown in dark black line.

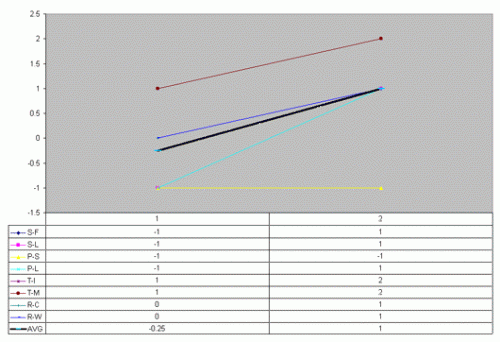
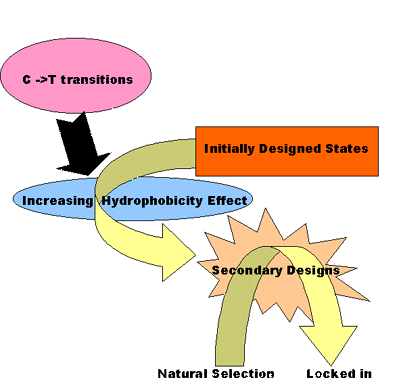


Figure 6. The effect of C-T transitions on beta-sheet formation. Numerical values were assigned to residues based on their frequency of occurrence of each amino acid in beta sheets (as per Chou & Fasman). 2= strong former; 1 = former; 0 = indifferent; -1 = breaker; -2 = strong breaker. Change in avg values shown in dark black line.

As can be seen from figure 5, these substitutions take from a pool that either breaks alpha-helices or is indifferent to their formation and replaces them with alpha-helix formers. In fact, the average score increases from -0.5 to 1.0. Likewise, figure 6 shows an original pool that either breaks or is indifferent to beta-sheets that is transformed into a pool that largely forms beta-strands.

Since hydrophobic interactions play a large role in protein folding and structure, the effects illustrated in figure 3 suggest C-T transitions may play a significant role in protein evolution. What’s more, figures 5 and 6 suggest C-T transitions may also result in both alpha helix and beta sheet elongation/formation. This raises the intriguing possibility that the genetic code was not only designed to minimize deleterious mutations, but that this design objective was balanced against a seemingly contrary objective to evolve new proteins through what I will call the Increasing Hydrophobicity Effect (IHE). This might also explain why serine, rather than proline, is included in the amino acid pool generated by C-T transitions (figure 4 and the sole exception from figure 3). Serine is present as a consequence of losing proline. This substitution removes a strong helix breaker and replaces it with a residue that is indifferent to helix formation. And proline is not added to the mix so as to not increase the chances that an existing helix is broken.

Such considerations allow for a further provocative hypothesis: a set of originally designed proteins may have been designed such that they could exploit the effects of C-T transitions, in essence channeling original designs in a direction to increase the chances that a buried, secondary design is extracted. And such events may have been tied to front-loading. (Figure 7) For example, proteins that played essential roles in the evolution of multicelluarity may have been spawned from the IHE. This scenario portrays at least some evolutionary events as an interaction between four dynamics: random mutation, mutation rigged by natural law, intelligent design, and natural selection. This hypothesis also makes a prediction that can be tested. For example, if the multicellular state was front-loaded with life’s design, we would expect to find conserved, multicellular-specific proteins have crucial FLIYWVMCS residues that can explained by C-T transitions relative to their ancestral state.  
Figure 7. The C-to-T transitions can be likened to a stream, constantly pushing amino acid content toward a more hydrophobic state. Given the context provided by the originally designed state (continually reflected by evolution’s tendency to borrow from pre-existing states), buried secondary designs may be unmasked on a somewhat regular basis. If the act of unmasking occurs in an appropriate environment, selection will lock the secondary design into the biosphere.

There are two other further considerations that may also be part of the picture. The rate of cytosine deamination in double-stranded DNA is less than 1% of that found in single-stranded DNA. Given that DNA can exist in a transient single-stranded state during transcription, this suggests that the IHE may also be targeted to genes that are actively transcribed. In fact, experimental evidence has shown there is a distinct bias for C-T transitions in the coding strand as a consequence of transcription. [7] Thus, a specified set of amino acids is being thrown at every transcribed gene. It’s as if life’s proteins are being forced to evolve.

Secondly, it is important to remember that most cytosine deamination events are corrected ultimately through the action of the DNA repair enzyme, uracil glycosylase. We can thus view the C-T transitions as a flux that is controlled by the activity of this enzyme. This creates the potential for both positive and negative feedback. For example, there may be residues important to uracil glycosylase activity that can be lost by C-T transitions. If this were to happen, a positive-feedback loop might set up resulting in the accelerated evolution of transcribed genes. Also, perhaps cytidine deaminase may function in a fashion similar to its role in somatic hypermutation. On the other hand, there may exist triggers that increase uracil glycosylase activity, thus decreasing the C-T flux. For example, might there be cases where “sick” cells, hindered by C-T transition mutations involving any number of pathways, activate a regulatory pathway that increases glycosylase activity?

From a design perspective, this model of evolution need not exert its effects in a ubiquitous fashion. Instead, perhaps only key events in life’s evolution were significantly helped by this “directed evolution.” However, there is an unfortunate caveat worth mentioning. The ability for C-T transition to unmask front-loaded states may have long ceased to exist. Such a dynamic may have been crucial to some early events in evolution, yet given these states may have dissipated (the proximal objectives were reached), current mutations may no longer reflect any detectable*design bias*. In such a case, the current predominance of C-T transitions (involved in some disease states) may simply be a *vestige of design*.

Although the above model is mostly speculative, and has yet to include other forms of mutation, it can serve as one platform for further design investigations. While some argue that life would be boring if evolution worked by engineering, I think the possibilities highlighted above indicate otherwise – evolution may be smarter and more sophisticated than we have yet to appreciate.

REFERENCES

1. Poole A, Penny D, Sjoberg BM. 2001. Confounded cytosine! Tinkering and the evolution of DNA. *Nat Rev Mol Cell Biol* Feb;2(2):147-51

2. Cited in Nei, M. 1987. *Molecular Evolutionary Genetics*. Columbia University Press; New York, p. 28

3. PJ de Jong, AJ Grosovsky, and BW Glickman. 1988. Spectrum of spontaneous mutation at the APRT locus of Chinese hamster ovary cells: an analysis at the DNA sequence level. *PNAS* 85(10): 3399-503.

4.Yatagai F, Glickman BW. 1990. Specificity of spontaneous mutation in the lacI gene cloned into bacteriophage M13. *Mutat Res* 1990 Jan;243(1):21-8

5.<http://www.socgenmicrobiol.org.uk/QUA/049906.pdf>

6.<http://psyche.uthct.edu/shaun/SBlack/aagrease.html>

7 A. Beletskii and Ashok S. Bhagwat. 1996. Transcription-induced mutations: Increase in C to T mutations in the nontranscribed strand during transcription in *Escherichia coli*. *PNAS.*93: 13919-13924

[This was originally written around 2002).