

***Searching for Molecular Solutions* – Additional Material**

CHAPTER 5

These Files contain additional material relevant to **Chapter 5** of *Searching for Molecular Solutions*. The page numbers of the book pertaining to each section are shown in the Table below, the corresponding page number for this file, and the title of each relevant section.

Contents:

~ Book Reference Page Number	Page Number in this File	Section	
		No.	Title
152	2	A6	Rubisco Enzyme
185	8	A7	Genetic Code Choice

Section A6: ***Rubisco Enzyme***

Relevant to the section 'Better Enzymes', beginning on p. 152 of *Searching for Molecular Solutions*.

A Rubric on Rubisco

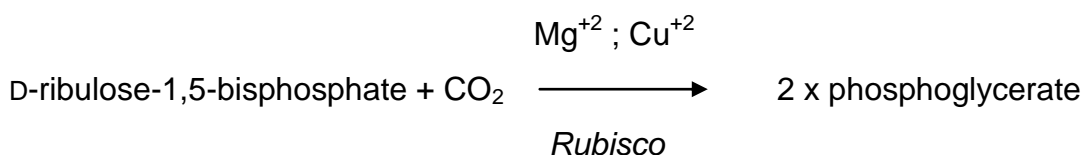
As an adjunct to the subsection of Chapter 5 on improvement of enzyme catalysis, we can look at a specific example in a little more detail. And what better case study than the most important enzyme on Earth? In the opinion of many, ribulose-1,5-bisphosphate carboxylase / oxygenase deserves this accolade, since it is the pivotal enzyme of photosynthesis, upon which a huge portion of the biosphere is fundamentally dependent ♥^{4,5}. The somewhat long-winded descriptor for this enzyme has inevitably been abbreviated, generating 'Rubisco' *. All photosynthesis on this planet depends on the process which this enzyme catalyzes *, the fixation of carbon dioxide through reaction with the doubly phosphorylated derivative of the pentose sugar ribulose:

♥ Or at least the part of biosphere which is most familiar and most important to us.

'Chemolithotrophic' bacteria fix inorganic carbon using chemical energy sources (such as sulfur oxidation¹) and do not require light. Such organisms form the basis of ecosystems (such as deep-ocean 'black-smoker' communities) which are independent of photosynthesis^{2,3}.

* The term sounds a bit like a company, but this is coincidental. However, applications of successfully-engineered variants of Rubisco will certainly have considerable commercial potential.

^ Note that though this step is a fundamental component of the overall photosynthetic reaction, it does not directly use the energy of light (photons). The ribulose-1,5-bisphosphate substrate for Rubisco is derived through other enzymatic reactions requiring ATP and NADPH cofactors produced during the light-dependent photosynthetic steps⁶.



This is the ‘carboxylase’ aspect of Rubisco, but it can also catalyze reaction of the same ribulose bisphosphate with oxygen, and is therefore also an oxygenase (giving us the full enzymatic title). Photosynthesizing organisms (and thus organisms expressing this enzyme) are usually divided into prokaryotic (cyanobacteria) and eukaryotic (algae and higher plant) categories, but in a very real sense all photosynthesis is of prokaryotic origin. The chloroplast photosynthetic factories of plants are derived from cyanobacterial endosymbionts in eukaryotic cells, in an ancient and extremely important case of ‘horizontal transfer’, as noted in Chapter 3. (For convenience and clarity, though, we will continue to use the conventional prokaryotic / eukaryotic labels in reference to Rubisco enzymes from different sources).

An interesting and much-noted feature of Rubisco is its apparent low efficiency ⁴, with a turn-over number (k_{cat}) of $\sim <10 \text{ s}^{-1}$ for variants of this enzyme from a wide range of photosynthetic organisms ⁷. Also, its co-activity as an oxygenase acts as a drain on the efficiency of the photosynthetic process as a whole ⁴. Somewhat ironically, the combination of its essential nature and ‘sluggish’ performance as a biocatalyst has resulted in a large investment in its synthesis by photosynthetic organisms, to the extent that it is credited as being the most abundant protein on this planet ⁸. The importance of Rubisco for plant growth and the apparent sizable scope for its improvement has prompted intensive research efforts towards generating more efficient versions of this pivotal enzyme, including the application of directed evolutionary approaches ^{5,9}.

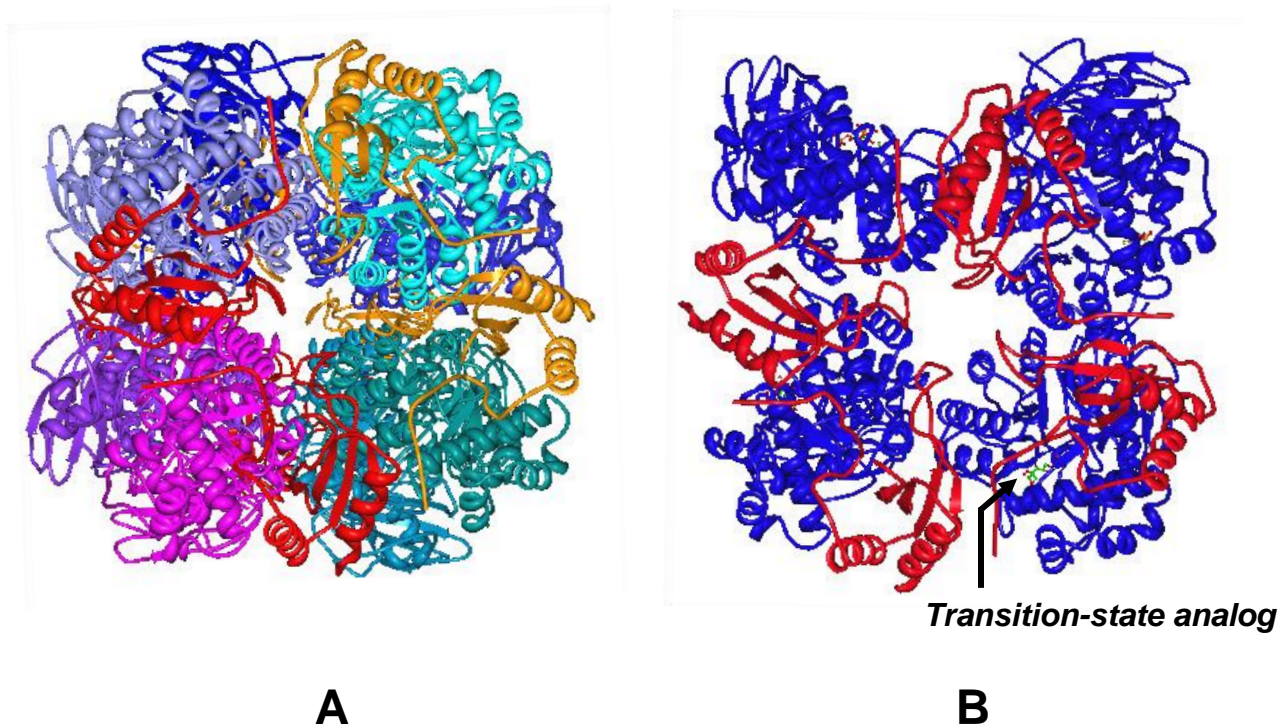


Fig. 5A6.1

Structure of spinach 'Form I' Rubisco, showing Large (L; 55kD) and Small (S; 15kD) subunits. **A.** L8S8 complex; each L subunit shown with a separate color; S subunits red or orange. Total molecular weight of complex is 560 kDa. **B.** Half complex (L4S4), L subunits blue; S subunits red, position of the transition-state analog used for the structural determination¹⁰ within one subunit shown by arrow. Source of structures: [Protein Data Bank](#)¹¹; [1UPM](#) (A)¹² and (B) [8RUC](#)¹⁰. Images generated with Protein Workshop¹³.

Rubisco enzymes from prokaryotes and eukaryotes have certain structural differences, but the large subunits of each (which are most important for catalysis)⁴ may derive from a common ancestor involved in the methionine salvage pathway^{14,15}. Though evolutionary divergences in the Rubisco large subunit have occurred in response to selective pressures, they are highly

conserved in phylogeny¹⁶. Land plant Rubisco (as represented by spinach) consists of a large complex of 8 large and 8 small subunits (Fig. 5A6.1). Genes encoding protein components of photosynthetic pathways may be constrained for progressive evolutionary change owing to their high degrees of connectivity¹⁷, and Rubisco is included in this assessment. Rubisco genes appear to have been widely naturally distributed by lateral (horizontal) transfer processes, (especially among prokaryotes^{18,19}), so much of the existing diversity for Rubisco evolution may have already been sampled by natural recombinational processes.

Apart from its 'connectivity' in photosynthetic circuitry and the corresponding issue of its inherent evolvability, the engineering of Rubisco presents other significant challenges. In land plants, the large subunits of this enzyme are encoded by chloroplast genomes, but the small subunits are nuclear, and special transport and chaperoning processes are vital for the correct assembly into the final protein complexes⁴. Owing to these dictates, land plant Rubisco has not been successfully assembled in *E. coli*, and expression studies on Rubisco in the latter bacterium (including directed evolution attempts) have accordingly used Rubisco enzymes from prokaryotes with more tractable quaternary structures[▼]. Another complication arises from the need to at least preserve the original carbon dioxide specificity of modified versions of Rubisco when attempting to enhance catalysis, since it is the net photosynthetic turnover which is the ultimate pay-off. (A faster enzyme with increased oxygen binding in lieu of carbon dioxide will not be useful).

As if these difficulties were not enough, there may be an even more fundamental issue with the general aim of Rubisco improvement. It has been proposed that Rubisco enzymes are in essence already doing more or less the best job

▼ The 'Form II' Rubisco enzymes of some cyanobacteria (such as *Rhodospirillum rubrum*) are homodimers of the large subunits and correspondingly simpler to assemble, but Rubisco from *Synechococcus* has a comparable L8S8 structure to land plants, but can still be assembled in *E. coli* under the aegis of host chaperones²⁰.

possible under the circumstances. Distinguishing carbon dioxide from oxygen by binding an appropriate transition state intermediate may be a difficult balancing act, which may in the end require compromise between substrate specificity and catalytic efficiency⁷. Despite evidence for continuing evolutionary ‘fine-tuning’ of Rubisco sequences¹⁶, their wide phylogenetic conservation is consistent with an enzyme with limited catalytic opportunity to move in sequence space. Many prokaryotic and eukaryotic photosynthetic organisms possess mechanisms for concentrating carbon dioxide for processing by Rubisco²¹⁻²³. The existence of such concentration processes could be interpreted as supporting the case that this enzyme in isolation is inherently limited in its evolvable selectivity for carbon dioxide in an oxygen-rich atmosphere[▼].

Given the importance of Rubisco in plant growth and thus agriculture, the potential difficulties with its improvement have not deterred protein engineers. It has been noted that natural Rubisco enzymes do show considerable variation in their specificity factors for carbon dioxide and oxygen ($[k_{\text{cat}}/K_m] \text{ CO}_2 / [k_{\text{cat}}/K_m] \text{ O}_2$), suggesting that this is amenable to modification at least to an extent⁴. Some organisms possess multiple forms of Rubisco which appear to be deployed as needed under changing carbon dioxide levels²⁴, suggesting that a single broad-spectrum enzyme with suitable efficiency is not readily naturally evolvable (if possible at all). Directed evolution of a cyanobacterial Rubisco in *E. coli*²⁵ has led to modest improvement in catalytic efficiency without sacrificing carbon dioxide specificity²⁰. It has been suggested that the artificial *E. coli* environment (an engineered ‘horizontal transfer event’) may allow a greater exploration of functional sequence space for Rubisco than is possible in its natural settings²⁰. Yet to be ultimately useful, an artificially-evolved Rubisco must function *in situ* within the photosynthetic machinery of higher plants⁵, which may raise additional difficulties owing to Rubisco’s protein-protein connectivity¹⁷. Some of Rubisco’s

▼ It is accepted that over geological time the $\text{CO}_2:\text{O}_2$ ratio has dramatically fallen (largely under the influence of photosynthesis itself), and that CO_2 –concentrating mechanisms have evolved in response to this environmental change^{22,23}.

regulatory interaction partners may nonetheless be useful targets for enhancement themselves, such as the Rubisco activation protein ⁴, or possibly Rubisco-specific chaperones ²⁶. Alternatively, cellular carbon dioxide concentrating mechanisms may be improvable, and offer indirect benefit to Rubisco efficiency ²⁷.

It remains to be seen as to how far Rubisco can be artificially improved by either rational approaches or directed evolution, but this system is fascinating and highly challenging from multiple viewpoints. Rubisco offers hurdles at the expression, assembly, and interactive systems levels, and will doubtless continue to provide many useful lessons in all these areas. But it remains possible that real photosynthetic improvement is severely constrained with natural Rubisco enzymes and their associated photosynthetic processes, by fundamental limitations of protein-based enzymes themselves. Even if this is ultimately confirmed, it simply raises the challenge even higher, to a more sophisticated level of molecular engineering or evolution. For example, the performance of Rubisco may benefit from judicious insertions of unnatural amino acids, through engineered genetic code extensions; noted in Chapter 5 of *Searching for Molecular Solutions*. Design with a systems biological view will no doubt be important if an improved Rubisco variant is to be functional in its *in situ* context. The frequently-acknowledged need to continually address the substrate preferences of Rubisco leads to the general subject of the evolutionary alteration of substrate specificity, also considered within Chapter 5.

Section A7: **Genetic Code Choice**

This area is relevant to the sections of *Searching for Molecular Solutions* which consider genetic code expansions (pp. 174-186). It can be considered as an after-thought for the last subsection, 'Unnatural Rewards and Limits' on p. 185.

Origins and Nature of the Genetic Code

There is in fact a fundamental issue hiding in the background to all work with novel genetic codes which space did not permit considering in *Searching for Molecular Solutions*. Why is the natural genetic code constituted as it is? Is the genetic code arbitrary? If not, what factors have determined its present arrangement? When codon reassignment is a real scientific issue, it is very pertinent to look at these questions in some detail, although of course people have pondered them ever since the genetic code was deciphered.

Pick a Code – Any Code?

If the genetic code is considered from purely a logical point of view, there is no reason to conclude that it is anything but an arbitrary rendering of protein sequence information into a digital code carried by nucleic acids²⁸. This is consistent with the early view of the code as a 'frozen accident' which is too deeply entrenched in life's fundamental activities to replace²⁹. Yet this interpretation does not stand up to abundant analyses of the real biological genetic code arrangement. A central issue is that the code itself has evolved, and this proposition alone would suggest that the code is not just one alternative randomly plucked from all other possibilities.

Reasonably simple inspection can show that the code is unquestionably non-random, since amino acids with the same codon usually only vary at third 'wobble' positions, and amino acids with similar hydrophilicities have related codon patterns [▼]. Codons can also be grouped by a measure of polarity for their encoded amino acids ('polarity requirement'³⁰⁻³²), and other non-random patterns in the genetic code have been noted ³³. These deviations from randomness have been frequently ascribed to selective forces channeling the code into a form which has the highest likelihood of minimizing the effects of errors, either at the level of mutations or mistranslations ^{*} during protein synthesis ^{30,35,36}. The rationale for this is simply that a code will be selectively favored where point mutations (or single-base codon : anticodon 'misreads' during translation) tend to substitute amino acids with similar properties, over a code which has a purely random codon assignment. Most analyses and models suggest that the code is optimized in this regard but not at a global perfect optimum ³³⁻³⁵. This has been interpreted as consistent with an early phase of code evolution which tended to become fixed with the arise of organismal complexity and increasing likelihood that any further changes would have global negative impact on an organism's fitness. As we will see shortly, though, even some complex cellular organisms have undergone certain coding changes, and a genetic code which is 'universal' in all aspects does not exist.

[▼] If a codon has an A in the second position then it encodes an amino acid with polar / hydrophilic character (for example, GAC for aspartic acid); if a U in the second position then a hydrophobic amino acid is encoded (eg. UUC for phenylalanine ³⁰. Note, though, that the second codon position alone is not sufficient to define the hydropathic status of encoded amino acids. (For example, all second position codons with A encode polar / hydrophilic amino acids, but not all hydrophilic amino acids have A at the second codon positions, as with hydrophilic arginine specified by CGN codons).

^{*} Note that 'translational robustness' (noted in Chapter 2 of *Searching for Molecular Solutions*) reflects the ability of a protein to tolerate errors in translation which inevitably occur; this is quite distinguishable from code optimization to minimize the effects of translational errors throughout an organism's proteome ³⁴.

Three major classes of theories for the origin of the genetic code have in fact been put forward by numerous workers³⁰. 'Adaptive' theories in general include the above error-minimization proposal, but can be generalized to include any function for which a selective advantage can be postulated. The 'historical' or 'coevolutionary' theory proposes that a simple proto-code was expanded through recruitment of new amino acids through biosynthetic pathways from precursor amino acids already in the primordial code. The modified amino acids were proposed to 'take over' codons from their precursors. (For example, hydroxylation of phenylalanine to tyrosine would by this scheme lead to the new assignment of codons to tyrosine (UAU / UAC) related to those for phenylalanine (UUU/UUC). The coevolutionary theory has been extensively criticized on several grounds, including the accuracy of proposed pathway linkages³⁷. 'Chemical' theories propose that direct molecular interactions between RNA molecules and amino acids led to specific coding assignments. In this view, the chemical 'fit' (or stereochemistry) between specific RNA codon triplets and their cognate amino acids was a precursor to coding development. By the strongest interpretation of this viewpoint, the genetic code by this proposal is anything but arbitrary, and in effect is virtually pre-set to take the form that it has. The basic chemical aspects of the genetic code can also be invoked to set certain probable limits. For example, no amino acids are specified solely by a pyrimidine in the third codon position, and this may be an inherent coding limitation³⁷. (Thus NNC and NNT codons are always specified by the first two codon positions. AAC and AAT are both asparagine; CAC and CAT are both histidine, etc.). In contrast, purines (A or G) at the third position can specify a codon (thus ATA and ATG are separate codons for isoleucine and methionine respectively).

In fact, if the strong version of the chemical-origin code theory was correct, it would predict that only one code is probable, or indeed perhaps even possible. Early observations seemed to support the notion of a truly universal code, but this is now known to be an inaccurate picture (as we consider further below). Yet known coding deviations are not complete re-inventions of the standard genetic

code but only modifications affecting a few amino acids³⁸, so the chemical theory was not refutable on these grounds alone. In recent times, a very interesting way of putting RNA-amino acid interactions to experimental testing has emerged. The technology for selecting functional RNA species *in vitro* (aptamers; as detailed in the Chapter 6 of *Searching for Molecular Solutions*) allows one to search through a vast number of RNA sequences to find those which bind to an amino acid of interest. If codon (or anticodon) triplets are relevant to direct RNA-peptide binding, one might expect to find that specific amino acid-binding RNA aptamers have triplet base sequences matching the codon or anticodon for the same amino acid. Sets of aptamers binding different amino acids have been duly isolated and sequenced. In the case of arginine, a significant number of specific binding aptamers possessed base triplets corresponding to arginine codons at a rate far beyond chance expectations^{39,40}. A number of other, but not all, amino acids also show such associations between their binding aptamers and codons or anticodons⁴⁰. It has been pointed out that theories of code origin and evolution are not necessarily mutually exclusive^{30,40}, and a chemical origin for some codon assignments does not rule out different origins for other amino acid codons.

A difficulty for adaptive code theories lies in the level of code plasticity versus its permanent fixation. If any change in the genetic code is lethal to an organism, how can it evolve? Independently-arising proto-codes during the early phase of the origin life could compete with each other until the best versions predominated, but once a 'universal' form was fixed in the Last Universal Common Ancestor of all extant life on Earth, change might grind to a halt. Or so it was believed until evidence to the contrary started to come in. the genomes with mitochondria (the energy-producing organelles for almost all eukaryotes) have a modified code, as do a number of fungi and protozoa³⁸. It was noted in Chapter 5 of *Searching for Molecular Solutions* that stop codons can be naturally reassigned to the 'additional' amino acids selenocysteine and pyrrolysine, and some natural alterations to normal amino acid coding also

involve stop codons. Some changes, though, involve natural reassignments of pre-existing codons, proving that it is not impossible for sense codons to change 'meaning' in complex eukaryotes. The mechanism for such coding switches is unclear, although some ingenious suggestions have been put forward ³⁸. Some experimental evidence indicates that the development of codon ambiguity (a precursor to coding switch) could offer selective advantages under some conditions ⁴¹.

Be that as it may, we can now better consider the natural genetic code and its variation in the light of modern unnatural coding reassignments. As also noted in Chapter 5, the charging of tRNAs by cognate aminoacyl-tRNA synthetases clearly has no inherent chemical limitation to the natural 20 amino acids. Nor is the ribosomal protein synthesis machinery bounded only by the natural amino acids (although some limitations exist). As continued progress is made with *in vivo* incorporation of unnatural amino acids, the desire for increasing levels of multiple foreign amino acid insertions within single proteins is likely to grow commensurately. In turn, this will necessitate the reassignment of an increasing number of codons, requiring sophisticated engineering of whole genomes.

With this point in mind, we can take codon reassignment to a logical end-point and conduct a thought-experiment. If the genetic code was completely arbitrary, it might be proposed that an entirely new code could be applied to an entire genome of an organism, as long as each reassigned codon was accompanied by systematic changes to corresponding tRNAs and aminoacyl-tRNA synthetases ²⁸. If this was magically done in one fell swoop, then one might expect the organism to continue along happily, since although its genome has been recoded, its expressed proteome would be exactly the same. One might also concede that the new global code is inferior to the natural code in a number of ways (such as error minimization, as we have considered above), but could not the radically-revised organism survive in a coddled laboratory environment when

free from competition with its natural counterparts? In other words, yes, the natural genetic code is certainly non-arbitrary, but could not an alternative work at all?

Alas, things are not so simple. As well as coding sequences for protein, there are many DNA sequences which act as control elements for the organization, expression, and regulation of the genome. These include (but are not limited to) target sequences for regulatory DNA binding proteins, splice signals, enhancers, and many sequences transcribed into regulatory RNAs. If all such control sequences were exclusively located outside of protein coding sequences, then the 'global code switch' thought experiment becomes more feasible, but in practice control elements can be found within coding tracts themselves[▼]. By making a global code change, such internal control signals would be spuriously altered, with potentially fatal global effects.

To illustrate this, let's consider a simple prokaryotic example. The bacteriophage lambda of *E. coli* encodes a protein (termed 'O' by the alphabetical gene nomenclature for this phage) which is critical for its replication. The details of this need not concern us here, but the relevant point is that O is a DNA-binding protein whose target sequences ('iterons') reside within its own coding sequence. We then decide in our thought-experiment to arbitrarily change the entire genetic code for this phage (and necessarily also its *E. coli* host) by permuting second codon positions bearing A or C (A becomes C, and vice versa, such that all codons with A or C in the second position become switched over). For example, the normal TCC and TAC codons for serine and tyrosine respectively at the DNA level are swapped by this operation to TAC (serine) and TCC (tyrosine) (Fig. 5A7.1 below). The entire tRNA / aminoacyl-tRNA synthetase machinery has

▼ An intriguing observation in this regard is that the natural genetic code is almost optimal for carrying additional 'second-code' sequences⁴².

undergone corresponding alterations (thought experiments are convenient for some things ♥), so all proteins should be made just as before. Is all well? Many problems might ensue, but as far as the phage is concerned, it's a disaster. The same protein O is made, but owing to the sequence changes engendered to bring in the new genetic code, the target 'iteron' DNA binding sites within its own coding sequence have vanished (Fig. 5A7.1 below). Lambda cannot then replicate its own genome.

This is not to say that engineering an entire organism with a completely altered code would be impossible, but it would require a design effort rather greater than a 'simple' code switch. More and more, questions of this nature will be plumbed as codon reassignments for unnatural amino acid incorporation into proteins proceed. The development of codon reassignment technology is thus of great benefit for addressing many significant scientific issues not previously accessible to experimentation.

♥ Assuming TAA, TGA, and TAG stop codons were kept constant, this operation would also change the relative numbers of degenerate codons for amino acids. Again using the serine / tyrosine example, tyrosine would have the same number of codons as before (its 'switched' codons corresponding to TCC and TCT), but serine would have four codons instead of its normal six (the 'switched' TAC and TAT, plus its normal AGC and AGT).

O Protein DNA sequence

ATGACAAATACAGCAAAAATACTCAACTTCGGCAGAGGTAAC TTTGCCGGACAGGAGCGTAATGTGGCAGATC
 TCGATGATGGTTACGCCAGACTATCAATATGCTGCTTGAGGCTTATTCGGGCGCAGATCTGACCAAGCGACA
 GTTTAAAGTGCTGCTTGCCATTCTGCGTAAACCTATGGGTGGAATAAACCAATGGACAGAATCACCGATTCTC
 AACTAGCGAGATTACAAAGTTACCTGTCAAACGGTGCAATGAAGCCAAGTTAGAACTCGTCAGAATGAATATT
 ATCAAGCAGCAAGGCGGCATGTTTGACCAATAAAAAACATCTCAGAATGGTGCATCCTCAAAACGAGGGAA
 AATCCCTAAAACGAGGGATAAAACA CCCTCAAATTGGGGGATTCTATCCTCAAAACAGGGGGACAAA
 AGACACTATTACAAAAGAAAAAGAAAAGATTATTCGTCAGAGAATTCTGGCGAATCCTCTGACCAGCCAGAAA
 ÉÉ ..**TGA**

Normal Code, O protein sequence

O Protein DNA binding sites boxed

I	P	Q	N	E	G	K	S	P	K	T	R	D	K	T
É	ATC CCT CAA AAC GAG GGA					AA	A TCC CCT AAA ACG AGG GAT					T	AAA	ACA
S	L	K	L	G	D	C	Y	P	S	K	Q	G	D	
TCC CTC AAA TTG GGG GAT					TGC	TAT	T CCC TCA AAA CAG GGG GAC					É	.	

Altered Code, O protein sequence

I	P	Q	N	E	G	K	S	P	K	T	R	D	K	T
ATC	CAT	CCA	ACC	GCG	GGA	ACA	TAC	CAT	ACA	AAG	AGG	GCT	ACA	AAA
S	L	K	L	G	D	C	Y	P	S	K	Q	G	D	
TAC	CTC	ACA	TTG	GGG	GCT	TGC	TCT	CAC	TAA	ACA	CCG	GGG	GCC	

Fig. 5A7.1

Fig. 5A7.1. Effect of code change on a DNA binding protein which recognizes sequences within its own protein coding sequence. Coding DNA sequence for Lambda O is shown at the top (initiation ATG and stop TGA codons shown in blue), with its internal binding sites ('iterons' ⁴³) boxed in light yellow. A partial translated O protein sequence is shown in the region of the iteron binding sites. If the code is altered with an A / C permutation at the second position for each codon, the resulting sequence is shown at the bottom, with altered codons in red. In this case, the new code specifies the same protein sequence, but the internal DNA target site for the same protein is destroyed.

References:

1. Kodama, Y. & Watanabe, K. Isolation and characterization of a sulfur-oxidizing chemolithotroph growing on crude oil under anaerobic conditions. *Appl Environ Microbiol* **69**, 107-12 (2003).
2. Kelley, D. S. et al. A serpentinite-hosted ecosystem: the Lost City hydrothermal field. *Science* **307**, 1428-34 (2005).
3. Brazelton, W. J., Schrenk, M. O., Kelley, D. S. & Baross, J. A. Methane- and sulfur-metabolizing microbial communities dominate the Lost City hydrothermal field ecosystem. *Appl Environ Microbiol* **72**, 6257-70 (2006).
4. Spreitzer, R. J. & Salvucci, M. E. Rubisco: structure, regulatory interactions, and possibilities for a better enzyme. *Annu Rev Plant Biol* **53**, 449-75 (2002).
5. Griffiths, H. Plant biology: designs on Rubisco. *Nature* **441**, 940-1 (2006).
6. Nelson, N. & Yocum, C. F. Structure and function of photosystems I and II. *Annu Rev Plant Biol* **57**, 521-65 (2006).
7. Tcherkez, G. G., Farquhar, G. D. & Andrews, T. J. Despite slow catalysis and confused substrate specificity, all ribulose biphosphate carboxylases may be nearly perfectly optimized. *Proc Natl Acad Sci U S A* **103**, 7246-51 (2006).
8. Ellis, R. J. The most abundant protein in the world. *Trends Biochem Sci* **4**, 241-244 (1979).
9. Normile, D. Agricultural research. Consortium aims to supercharge rice photosynthesis. *Science* **313**, 423 (2006).
10. Andersson, I. Large structures at high resolution: the 1.6 Å crystal structure of spinach ribulose-1,5-bisphosphate carboxylase/oxygenase complexed with 2-carboxyarabinitol bisphosphate. *J Mol Biol* **259**, 160-74 (1996).
11. Berman, H., Henrick, K. & Nakamura, H. Announcing the worldwide Protein Data Bank. *Nat Struct Biol* **10**, 980 (2003).
12. Karkehabadi, S., Taylor, T. C. & Andersson, I. Calcium supports loop closure but not catalysis in Rubisco. *J Mol Biol* **334**, 65-73 (2003).
13. Moreland, J. L., Gramada, A., Buzko, O. V., Zhang, Q. & Bourne, P. E. The Molecular Biology Toolkit (MBT): a modular platform for developing molecular visualization applications. *BMC Bioinformatics* **6**, 21 (2005).
14. Ashida, H. et al. A functional link between RuBisCO-like protein of *Bacillus* and photosynthetic RuBisCO. *Science* **302**, 286-90 (2003).

15. Ashida, H., Danchin, A. & Yokota, A. Was photosynthetic RuBisCO recruited by acquisitive evolution from RuBisCO-like proteins involved in sulfur metabolism? *Res Microbiol* **156**, 611-8 (2005).
16. Kapralov, M. V. & Filatov, D. A. Widespread positive selection in the photosynthetic Rubisco enzyme. *BMC Evol Biol* **7**, 73 (2007).
17. Shi, T., Bibby, T. S., Jiang, L., Irwin, A. J. & Falkowski, P. G. Protein interactions limit the rate of evolution of photosynthetic genes in cyanobacteria. *Mol Biol Evol* **22**, 2179-89 (2005).
18. Delwiche, C. F. & Palmer, J. D. Rampant horizontal transfer and duplication of rubisco genes in eubacteria and plastids. *Mol Biol Evol* **13**, 873-82 (1996).
19. Uchino, Y. & Yokota, A. "Green-like" and "red-like" RubisCO cbbL genes in Rhodobacter azotoformans. *Mol Biol Evol* **20**, 821-30 (2003).
20. Greene, D. N., Whitney, S. M. & Matsumura, I. Artificially evolved Synechococcus PCC6301 Rubisco variants exhibit improvements in folding and catalytic efficiency. *Biochem J* **404**, 517-24 (2007).
21. Fukuzawa, H. et al. Ccm1, a regulatory gene controlling the induction of a carbon-concentrating mechanism in Chlamydomonas reinhardtii by sensing CO₂ availability. *Proc Natl Acad Sci U S A* **98**, 5347-52 (2001).
22. Badger, M. R. & Price, G. D. CO₂ concentrating mechanisms in cyanobacteria: molecular components, their diversity and evolution. *J Exp Bot* **54**, 609-22 (2003).
23. Giordano, M., Beardall, J. & Raven, J. A. CO₂ concentrating mechanisms in algae: mechanisms, environmental modulation, and evolution. *Annu Rev Plant Biol* **56**, 99-131 (2005).
24. Yoshizawa, Y., Toyoda, K., Arai, H., Ishii, M. & Igarashi, Y. CO₂-responsive expression and gene organization of three ribulose-1,5-bisphosphate carboxylase/oxygenase enzymes and carboxysomes in Hydrogenovibrio marinus strain MH-110. *J Bacteriol* **186**, 5685-91 (2004).
25. Parikh, M. R., Greene, D. N., Woods, K. K. & Matsumura, I. Directed evolution of RuBisCO hypermorphs through genetic selection in engineered E.coli. *Protein Eng Des Sel* **19**, 113-9 (2006).
26. Saschenbrecker, S. et al. Structure and function of RbcX, an assembly chaperone for hexadecameric Rubisco. *Cell* **129**, 1189-200 (2007).
27. Parry, M. A., Madgwick, P. J., Carvalho, J. F. C. & Andralojc, P. J. Prospects for increasing photosynthesis by overcoming the limitations of Rubisco. *J Agricult Sci* **145**, 31-43 (2007).
28. Hofstadter, D. R. *Metamagical Themas: Questing for the Essence of Mind and Pattern* (Basic Books, New York., 1985).

29. Crick, F. H. The origin of the genetic code. *J Mol Biol* **38**, 367-79 (1968).
30. Knight, R. D., Freeland, S. J. & Landweber, L. F. Selection, history and chemistry: the three faces of the genetic code. *Trends Biochem Sci* **24**, 241-7 (1999).
31. Woese, C. R., Dugre, D. H., Saxinger, W. C. & Dugre, S. A. The molecular basis for the genetic code. *Proc Natl Acad Sci U S A* **55**, 966-74 (1966).
32. Woese, C. R., Olsen, G. J., Ibba, M. & Soll, D. Aminoacyl-tRNA synthetases, the genetic code, and the evolutionary process. *Microbiol Mol Biol Rev* **64**, 202-36 (2000).
33. Ardell, D. H. & Sella, G. No accident: genetic codes freeze in error-correcting patterns of the standard genetic code. *Philos Trans R Soc Lond B Biol Sci* **357**, 1625-42 (2002).
34. Novozhilov, A. S., Wolf, Y. I. & Koonin, E. V. Evolution of the genetic code: partial optimization of a random code for robustness to translation error in a rugged fitness landscape. *Biol Direct* **2**, 24 (2007).
35. Freeland, S. J. & Hurst, L. D. The genetic code is one in a million. *J Mol Evol* **47**, 238-48 (1998).
36. Freeland, S. J., Wu, T. & Keulmann, N. The case for an error minimizing standard genetic code. *Orig Life Evol Biosph* **33**, 457-77 (2003).
37. Ronneberg, T. A., Landweber, L. F. & Freeland, S. J. Testing a biosynthetic theory of the genetic code: fact or artifact? *Proc Natl Acad Sci U S A* **97**, 13690-5 (2000).
38. Knight, R. D., Freeland, S. J. & Landweber, L. F. Rewiring the keyboard: evolvability of the genetic code. *Nat Rev Genet* **2**, 49-58 (2001).
39. Knight, R. D. & Landweber, L. F. Guilt by association: the arginine case revisited. *Rna* **6**, 499-510 (2000).
40. Yarus, M., Caporaso, J. G. & Knight, R. Origins of the genetic code: the escaped triplet theory. *Annu Rev Biochem* **74**, 179-98 (2005).
41. Pezo, V. et al. Artificially ambiguous genetic code confers growth yield advantage. *Proc Natl Acad Sci U S A* **101**, 8593-7 (2004).
42. Itzkovitz, S. & Alon, U. The genetic code is nearly optimal for allowing additional information within protein-coding sequences. *Genome Res* **17**, 405-12 (2007).
43. Zahn, K. & Blattner, F. R. Binding and bending of the lambda replication origin by the phage O protein. *Embo J* **4**, 3605-16 (1985).