NIH grant proposal

Douglas Myers-Turnbull, Robin Betz, Yunsup Jung

Things to do

2: Write a little more here.

4: Finish

Contents

Specific aims

We propose a large-scale bioinformatics study to identify the effects of synonymous codon usage on protein structure. We intend to address causal relationships rather than statistical associations by developing a mathematically and statistically rigorous framework, which we will use to address a number of hypotheses. TODO: Write a little more here. [Robin]

Aim 1: Develop a comprehensive, verifiable, and rigorous framework to elucidate relationships between codon usage bias and protein structure.

We will develop a mathematical and computational framework that will allow the robust detection of relationships between codon usage and variables describing protein structure, even when the overall statistical correlation between the two variables is low.

Aim 2: Identify particular effects of synonymous mutations on protein structure.

We will use the framework established in aim 1 to develop a computational pipeline to detect proteins whose structures are affected by differential codon usage. We hypothesize that many of these structural changes can cause protein misfolding, which may be clinically important.

Aim 3: Elucidate general mechanisms that underlie differential codon usage.

We intend to use the same framework and pipeline described in the previous aims to investigate our hypotheses that codon usage bias is causally related to protein domain organization, secondary structure, knotting, folding environment, and structural complexity. Secondarily, we wish to establish codon usage bias as an important biological mechanism.

Significance

Existing literature

Correlation with expression

It is known that codon usage bias correlates with expression levels [14, 18]. There are strong indications that abundance of isoaccepting tRNA molecules is a causal factor of this differential expression [20, 27, 29, 35]. Proteins with high expression levels contain greater levels of frequently used codons, and frequently used codons are associated with higher levels of corresponding tRNAs. Thus there is believed to be a positive selection for a small, constrained set of codon–tRNA combinations, and concentrations of codons and tRNAs are related in a positive feedback cycle, where bias in either causes a positive selection for bias in the other. Therefore, statistically significant violations of this general trend are interesting because they indicate the presence of other selection biases. Such selection biases may be at the heart of important mechanisms of protein expression, some which are probably currently unknown.

Terminology

Because of strong correlation described above, we call infrequently used codons *slow*, and frequently used codons *fast*, except in cases where we address this correlation directly. Furthermore, we consider sequences containing a large proportion of fast codons to have high *codon usage bias*, and sequences containing either a moderate or low proportion to be *unbiased*, even though the bias of sequences with low proportion still deviate from the statistical mean.

Importance in translational dynamics

There is also substantial evidence that codon usage bias is also fundamentally linked to translation dynamics [3, 6, 7, 13, 23, 25]. Factors including mRNA secondary structure [2, 9, 34], mRNA stability [17], codon–tRNA affinity [20], and translational errors [12, 27, 41] have been suggested. In addition, studies have shown that codon usage bias is strongly correlated to ribosomal traffic [6, 8, 21]. Particularly, was demonstrated that, for efficient translation, overall codon usage should be skewed in favor of fast codons, and there should be a gradient in which slow codons are prevalent at the beginning of the transcript but rare toward the middle and end. Mitarai and Pedersen [25] used a simple computational model to show that the introduction of even a single slow codon near the end of the transcript can cause ribosomal traffic jams that drastically decrease translation rate, and presumably also expression. This is believed to be due to ribosomal queueing, a ribosome translating an mRNA transcript interacts physically with the translation process of another ribosome on the same transcript. This prevents both ribosomes from proceeding to subsequent codons.

Effects on protein structure

Because slow codons can cause pauses in translation, it has been suggested that synonymous codon usage can influence protein folding, leading to different folded states for transcripts with differing synonymous codon usage [6, 10, 22, 38]. Several studies have found that codon usage bias is correlated to protein structure [1, 4, 15, 31], including protein secondary structure [15, 28, 36] and domain organization [16, 28].

The first known effect of synonymous mutations on protein structure came from Crombie et al. [10], who replaced 10 consecutive slow codons with fast codons in the E. coli TRP3 protein. Doing so reduced the native enzymatic activity by 1.5-fold. Still more surprising was the finding by **?**] that a silent polymorphism in the mammalian multi-drug-resistant gene MDR1 altered its folded state and decreased its substrate specificity, without affecting its expression. A recent study by Zhou et al. [40] found a similar result for a natural mutation in the clinically important circadian rhythm protein FRQ. These studies hint to the existance of more examples of such effects. Furthermore, we argue that additional such cases are likely to be clinically important.

Limitations of existing approaches

In general, the bioinformatics studies by Adzhubei et al. [1], Biro [4], Gu et al. [15], Saunders and Deane [31] controlled for very few variables and were therefore able to identify only a few clear correlations (most notably with protein secondary structure). However, we hypothesize that these results were negative because the effect of codon usage bias on structure is a weak signal: the effects on most protein structure are minor or nonexistent for most proteins. However, the weakness of the signal does not belie the presence of general mechanisms behind the effects; this is evidence because, as shown by Crombie et al. [10], Zhou et al. [40**?**], differential codon usage can dramatically affect the folding of certain proteins. Therefore, we further hypothesize that general mechanisms exist even if few general correlations do, and that controlling for more variables will allow us to elucidate general mechanisms.

Further applications

In addition to clinically significant synonymous mutations and other differential codon usage, our data will have impact on additional applications. It has been recently shown that codon usage bias can be a pivotal factor in de novo protein design [19]. Although it is know that designed transcripts should contain mostly fast codons, and that there should be a gradient of codon bias along the transcript sequence, potential unwanted effects of synoymous codon usage on a protein product is not generally considered as part of protein design. We therefore note that the discovery of general mechanisms may be important to this application.

Assessing the impact of codon usage on protein structure has implications for protein folding analyses, especially if correlation is found between fast or slow codons and domain or interdomain regions. Prediction of protein structure given amino acid sequence is one of the foremost problems in biochemistry, however known determinants of structure are few [5] and the predictions made by current computational models frequently fall short of native conformations [11, 33]. Finding a relationship between codon bias and protein structure would provide considerable additional predictive power to such models.

Innovation

The previous studies by Adzhubei et al. [1], Biro [4], Gu et al. [15], Saunders and Deane [31] examining protein structure in the context of codon usage bias have been constrained to examiniations of statistical correlations. Although they and many other studies [] have suggested that differential codon usage may influence protein structure, such effects have only been demonstrated in vivo by Zhou et al. [40**?**].

No studies attempting to investigate causal relationships between codon usage bias and protein structure have been thus far published in peer-reviewed journals, and, to our knowledge, no such investigations have been performed. Therefore, we conclude that our proposal is strictly unique in this endevor.

In addition, the previous studies were partly unsuccessful in establishing even statistical correlations. The investigation by Saunders and Deane [31], which is the most recent, interrogated differences in codon usage within helices, strands, and coils by applying both Mantel–Haenszel statistics and a χ^2-test. They found few significant differences between the three secondary structural types, but did find a significant decrease in codon usage bias near the transitions between secondary structural elements. They also investigated the hypothesis that slow codons are frequent around domain boundaries; the results in this case were negative. However, these studies relied on simple statistical techniques that lack the power necessary to find statistical associations for weaker signals.

TODO: Finish

Approach

Aim 1

We primarily aim to investigate the relationship between secondary structural elements (SSEs) and slow codons, with the aim of finding biases towards slow codon enrichment in certain elements. The null hypothesis for this situation is that there is no meaningful relationship between frequency of slow codons and domain features. However, we hypothesize that such a relationship does exist, as previous research indicates the introduction or removal of slow codons has a dramatic effect on protein expression.

Mathematical model

Bias towards fast or slow codons in mRNA sequence will be established using the codon adaptation index (CAI) described by Sharp and Li [32]. Codons are assigned weights depending on the frequency of their corresponding tRNA in the appropriate organism:

$$w_{ij} = \frac{X_{ij}}{X_{i^*}} \tag{1}$$

The codon usage bias in a particular region is then the geometric mean of the codons that comprise it:

$$CAI = (\Pi_{k=1}^{L} w_k)^{\frac{1}{L}} \tag{2}$$

$$= \exp(\frac{1}{L}\Sigma_{k=1}^{L} \log w_k) \tag{3}$$

Pipeline

Domain information will be obtained from the Structural Classification of Proteins (SCOP) database [26]. Coding mRNA sequences for the protein will be obtained from the NCBI Reference Sequence Database (RefSeq) [30]. Sequences will be correlated to structures by means of Structure Integration with Function, Taxonomy, and Sequence (SIFTS) database [37].

Benchmarking

Many proteins will be examined over a number of species. While our preliminary analysis has been performed in yeast, we hope to analyze over 10,000 human proteins in hopes of maximizing sample size.

Investigation

Codon usage bias will be examined on domain boundary regions and non-boundary regions, and the overall CAI value will be compared across these regions. If a significant difference in CAI across secondary structure features is found, the hypothesis will be supported.

Aim 2

Aim 3

Folding study

As an auxillary study, we will investigate the effects of synonymous mutations in detail for select cases using MD-based folding software. Although de novo folding is still largely unsolved, and de novo folding algorithms are still in their infancy, such an investigation may still reveal insight for some cases that is unavailable through other means. [39]

Bibliography & References Cited

[1] a a Adzhubei, I a Adzhubei, I a Krasheninnikov, and S Neidle. Non-random usage of 'degenerate' codons is related to protein three-dimensional structure. *FEBS letters*, 399(1-2):78–82, December 1996. ISSN 0014-5793. URL `http://www.ncbi.nlm.nih.gov/pubmed/8980124`.

[2] S B Baim, D F Pietras, D C Eustice, and F Sherman. A mutation allowing an mRNA secondary structure diminishes translation of Saccharomyces cerevisiae iso-1-cytochrome c. *Molecular and cellular biology*, 5 (8):1839–46, August 1985. ISSN 0270-7306. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=366899&tool=pmcentrez&rendertype=abstract.

[3] Kajetan Bentele, Paul Saffert, Robert Rauscher, Zoya Ignatova, and Nils Blüthgen. Efficient translation initiation dictates codon usage at gene start. *Molecular systems biology*, 9(675):675, January 2013. ISSN 1744-4292. doi: 10.1038/msb.2013.32. URL http://www.ncbi.nlm.nih.gov/pubmed/23774758.

[4] Jan Charles Biro. Indications that "codon boundaries" are physico-chemically defined and that protein-folding information is contained in the redundant exon bases. *Theoretical biology & medical modelling*, 3: 28, January 2006. ISSN 1742-4682. doi: 10.1186/1742-4682-3-28. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1560374&tool=pmcentrez&rendertype=abstract.

[5] Joseph D Bryngelson, Jose Nelson Onuchic, Nicholas D Socci, and Peter G Wolynes. Funnels, pathways, and the energy landscape of protein folding: a synthesis. *Proteins: Structure, Function, and Bioinformatics*, 21(3):167–195, 1995.

[6] J Ross Buchan and Ian Stansfield. Halting a cellular production line: responses to ribosomal pausing during translation. *Biology of the cell / under the auspices of the European Cell Biology Organization*, 99(9):475–87, September 2007. ISSN 1768-322X. doi: 10.1042/BC20070037. URL http://www.ncbi.nlm.nih.gov/pubmed/17696878.

[7] Gina Cannarozzi, Gina Cannarrozzi, Nicol N Schraudolph, Mahamadou Faty, Peter von Rohr, Markus T Friberg, Alexander C Roth, Pedro Gonnet, Gaston Gonnet, and Yves Barral. A role for codon order in translation dynamics. *Cell*, 141(2):355–67, April 2010. ISSN 1097-4172. doi: 10.1016/j.cell.2010.02.036. URL http://www.ncbi.nlm.nih.gov/pubmed/20403329.

[8] Gina Cannarozzi, Nicol N Schraudolph, Mahamadou Faty, Peter Von Rohr, Markus T Friberg, Alexander C Roth, Pedro Gonnet, Gaston Gonnet, and Yves Barral. Theory A Role for Codon Order in Translation Dynamics. *Cell*, 141(2):355–367, 2010. ISSN 0092-8674. doi: 10.1016/j.cell.2010.02.036. URL http://dx.doi.org/10.1016/j.cell.2010.02.036.

[9] J V Chamary and Laurence D Hurst. Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals. *Genome biology*, 6(9):R75, January 2005. ISSN 1465-6914. doi: 10.1186/gb-2005-6-9-r75. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1242210&tool=pmcentrez&rendertype=abstract.

[10] T Crombie, J P Boyle, J R Coggins, and a J Brown. The folding of the bifunctional TRP3 protein in yeast is influenced by a translational pause which lies in a region of structural divergence with Escherichia coli indoleglycerol-phosphate synthase. *European journal of biochemistry / FEBS*, 226(2):657–64, December 1994. ISSN 0014-2956. URL http://www.ncbi.nlm.nih.gov/pubmed/8001582.

[11] Rhiju Das. Four small puzzles that rosetta doesn't solve. *PLoS One*, 6(5):e20044, 2011.

[12] D Allan Drummond and Claus O Wilke. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell*, 134(2):341–52, July 2008. ISSN 1097-4172. doi: 10.1016/j.cell.2008.05.042. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2696314&tool=pmcentrez&rendertype=abstract.

[13] Kurt Fredrick and Michael Ibba. How the sequence of a gene can tune its translation. *Cell*, 141(2):227–9, April 2010. ISSN 1097-4172. doi: 10.1016/j.cell.2010.03.033. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2866089&tool=pmcentrez&rendertype=abstract.

[14] Regina M Goetz and Anders Fuglsang. Correlation of codon bias measures with mRNA levels: analysis of transcriptome data from Escherichia coli. *Biochemical and biophysical research communications*, 327(1): 4–7, February 2005. ISSN 0006-291X. doi: 10.1016/j.bbrc.2004.11.134. URL http://www.ncbi.nlm.nih.gov/pubmed/15629421.

[15] Wanjun Gu, Tong Zhou, Jianmin Ma, Xiao Sun, and Zuhong Lu. Folding type specific secondary structure propensities of synonymous codons. *..., IEEE Transactions on*, 2(3):150–157, 2003. URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1229599.

[16] Wanjun Gu, Tong Zhou, Jianmin Ma, Xiao Sun, and Zuhong Lu. The relationship between synonymous codon usage and protein structure in Escherichia coli and Homo sapiens. *Bio Systems*, 73(2):89–97, March 2004. ISSN 0303-2647. doi: 10.1016/j.biosystems.2003.10.001. URL http://www.ncbi.nlm.nih.gov/pubmed/15013221.

[17] Wanjun Gu, Tong Zhou, and Claus O Wilke. A universal trend of reduced mRNA stability near the translation-initiation site in prokaryotes and eukaryotes. *PLoS computational biology*, 6(2):e1000664, February 2010. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1000664. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2816680&tool=pmcentrez&rendertype=abstract.

[18] Claes Gustafsson, Sridhar Govindarajan, and Jeremy Minshull. Codon bias and heterologous protein expression. *Trends in biotechnology*, 22(7):346–53, July 2004. ISSN 0167-7799. doi: 10.1016/j.tibtech.2004.04.006. URL http://www.ncbi.nlm.nih.gov/pubmed/15245907.

[19] Claes Gustafsson, Sridhar Govindarajan, and Jeremy Minshull. Codon bias and heterologous protein expression. *Trends in biotechnology*, 22(7):346–353, 2004.

[20] Stefan Klumpp, Jiajia Dong, and Terence Hwa. On ribosome load, codon bias and protein abundance. *PloS one*, 7(11):e48542, January 2012. ISSN 1932-6203. doi: 10.1371/journal.pone.0048542. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3492488&tool=pmcentrez&rendertype=abstract.

[21] a a Komar, T Lesnik, and C Reiss. Synonymous codon substitutions affect ribosome traffic and protein folding during in vitro translation. *FEBS letters*, 462(3):387–91, December 1999. ISSN 0014-5793. URL http://www.ncbi.nlm.nih.gov/pubmed/10622731.

[22] Anton a Komar. A pause for thought along the co-translational folding pathway. *Trends in biochemical sciences*, 34(1):16–24, January 2009. ISSN 0968-0004. doi: 10.1016/j.tibs.2008.10.002. URL http://www.ncbi.nlm.nih.gov/pubmed/18996013.

[23] Monica Marin. Folding at the rhythm of the rare codon beat. *Biotechnology journal*, 3(8):1047–57, August 2008. ISSN 1860-7314. doi: 10.1002/biot.200800089. URL http://www.ncbi.nlm.nih.gov/pubmed/18624343.

[24] Andrew C R Martin. Mapping PDB chains to UniProtKB entries. *Bioinformatics (Oxford, England)*, 21(23): 4297–301, December 2005. ISSN 1367-4803. doi: 10.1093/bioinformatics/bti694. URL http://www.ncbi.nlm.nih.gov/pubmed/16188924.

[25] Namiko Mitarai and Steen Pedersen. Control of ribosome traffic by position-dependent choice of synonymous codons. *Physical biology*, 10(5):056011, October 2013. ISSN 1478-3975. doi: 10.1088/1478-3975/10/5/056011. URL http://www.ncbi.nlm.nih.gov/pubmed/24104350.

[26] A G Murzin, S E Brenner, T J P Hubbard, and C Chothia. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, 247:536–540, 1995.

[27] Hamed Shateri Najafabadi, Jean Lehmann, and Mohammad Omidi. Error minimization explains the codon usage of highly expressed genes in Escherichia coli. *Gene*, 387(1-2):150–5, January 2007. ISSN 0378-1119. doi: 10.1016/j.gene.2006.09.004. URL http://www.ncbi.nlm.nih.gov/pubmed/17097242.

[28] Matej Oresic, Michael Dehn, Daniel Korenblum, and David Shalloway. Tracing specific synonymous codon-secondary structure correlations through evolution. *Journal of molecular evolution*, 56(4):473–84, April 2003. ISSN 0022-2844. doi: 10.1007/s00239-002-2418-x. URL http://www.ncbi.nlm.nih.gov/pubmed/12664167.

[29] Joshua B Plotkin and Grzegorz Kudla. Synonymous but not the same: the causes and consequences of codon bias. *Nature reviews. Genetics*, 12(1):32–42, January 2011. ISSN 1471-0064. doi: 10.1038/nrg2899. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3074964&tool=pmcentrez&rendertype=abstract.

[30] Kim D Pruitt, Tatiana Tatusova, and Donna R Maglott. Ncbi reference sequences (refseq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic acids research*, 35(suppl 1): D61–D65, 2007.

[31] Rhodri Saunders and Charlotte M Deane. Synonymous codon usage influences the local protein structure observed. *Nucleic acids research*, 38(19):6719–6728, 2010.

[32] Paul M Sharp and Wen-Hsiung Li. The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic acids research*, 15(3):1281–1295, 1987.

[33] M Scott Shell, S Banu Ozkan, Vincent Voelz, Guohong Albert Wu, and Ken A Dill. Blind test of physics-based prediction of protein structures. *Biophysical journal*, 96(3):917–924, 2009.

[34] Nina Stoletzki. Conflicting selection pressures on synonymous codon use in yeast suggest selection on mRNA secondary structures. *BMC evolutionary biology*, 8:224, January 2008. ISSN 1471-2148. doi: 10.1186/1471-2148-8-224. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2533328&tool=pmcentrez&rendertype=abstract.

[35] S Tavaré and B Song. Codon preference and primary sequence structure in protein-coding regions. *Bulletin of mathematical biology*, 51(1):95–115, January 1989. ISSN 0092-8240. URL http://www.ncbi.nlm.nih.gov/pubmed/2706404.

[36] T a Thanaraj and P Argos. Protein secondary structural types are differentially coded on messenger RNA. *Protein science : a publication of the Protein Society*, 5(10):1973–83, October 1996. ISSN 0961-8368. doi: 10.1002/pro.5560051003. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2143259&tool=pmcentrez&rendertype=abstract.

[37] Sameer Velankar, José M Dana, Julius Jacobsen, Glen van Ginkel, Paul J Gane, Jie Luo, Thomas J Oldfield, Claire ODonovan, Maria-Jesus Martin, and Gerard J Kleywegt. Sifts: Structure integration with function, taxonomy and sequences resource. *Nucleic acids research*, 41(D1):D483–D489, 2013.

[38] Gong Zhang, Magdalena Hubalewska, and Zoya Ignatova. Transient ribosomal attenuation coordinates protein synthesis and co-translational folding. *Nature structural & molecular biology*, 16(3):274–80, March 2009. ISSN 1545-9985. doi: 10.1038/nsmb.1554. URL http://www.ncbi.nlm.nih.gov/pubmed/19198590.

[39] Yang Zhang. I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics*, 9:40, 2008.

[40] Mian Zhou, Jinhu Guo, Joonseok Cha, Michael Chae, She Chen, Jose M Barral, Matthew S Sachs, and Yi Liu. Non-optimal codon usage affects expression, structure and function of

clock protein FRQ. *Nature*, 495(7439):111–5, March 2013. ISSN 1476-4687. doi: 10.1038/nature11833. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3629845&tool=pmcentrez&rendertype=abstract.

[41] Tong Zhou, Mason Weems, and Claus O Wilke. Translationally optimal codons associate with structurally sensitive sites in proteins. *Molecular biology and evolution*, 26(7):1571–80, July 2009. ISSN 1537-1719. doi: 10.1093/molbev/msp070. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2734146&tool=pmcentrez&rendertype=abstract.