NIH grant proposal

Douglas Myers-Turnbull, Robin Betz, Yunsup Jung

Things to do

7: Fill in when run finishes [Douglas]

7: Finish this paragraph [Douglas]

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.

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 [16, 20]. There are strong indications that abundance of isoaccepting tRNA molecules is a causal factor of this differential expression [25, 33, 36, 44]. 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 [4, 8, 9, 15, 30, 31]. Factors including mRNA secondary structure [3, 11, 43], mRNA stability [19], codon–tRNA affinity [25], and translational errors [14, 33, 51] have been suggested. In addition, studies have shown that codon usage bias is strongly correlated to ribosomal traffic [8, 10, 26]. 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 [31] 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 [8, 12, 27, 48]. Several studies have found that codon usage bias is correlated to protein structure [1, 6, 17, 38], including protein secondary structure [17, 35, 38, 45] and domain organization [18, 35].

The first known effect of synonymous mutations on protein structure came from Crombie et al. [12], 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. [50] 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 [6], Gu et al. [17], Saunders and Deane [38] 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. [12], Zhou et al. [50**?**], 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 [21]. 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 [7] and the predictions made by current computational models frequently fall short of native conformations [13, 40]. 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 [6], Gu et al. [17], Saunders and Deane [38] 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. [50**?**].

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 [38], 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.

Immense progress has been made in both the quality and quantity of information contained within bioinformatics databases within the past few years, making analyses of codon usage across thousands or more proteins possible. Examining larger sample sizes enables the study of family or activity based codon bias as well as allowing for a more robust statistical analysis when compared with previous studies where fewer than 300 proteins were examined.

Approach

Aim 1

Determination of causation

Ultimately, we are interested in causation related to codon usage bias. The analyses described above are independently useful, both for novel findings and verification of previous results. While we argue that such

results are interesting, biologically relavent, and warrant further investigation, they may indicate only indirect correlations. It is entirely possible for one variable to account for another in whole or in part; for example, domain boundaries may be enriched for slow codons only because they contain more strands than helices, or that solvent accessibility around domain boundaries may be the more fundamental explanation.

Data sets

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.

We will construct 5–10 species-dependent data sets, covering eukaryotes, invertibrates, and Homo sapiens. Using more than 1 data set will allow us to control for inter-species variation. We will necessarily limit the data sets to contain only genes annotated with structures in the Protein Data Bank (PDB) [5].

The content of databases that permit public deposition of entries, such as the NCBI and the PDB, are significantly biased toward sequences and structures that are of experimental interest. To control for such bias, we will cluster the genes by alignments using Basic Local Alignment Tool (BLAST) [] pairwise, then clustering by 40% sequence identity to remove homologs.

Coding mRNA sequences for the protein will be obtained from the NCBI Reference Sequence Database (RefSeq) [37]. Sequences will be correlated to structures by means of Structure Integration with Function, Taxonomy, and Sequence (SIFTS) database [46].

Quantification of bias

Bias towards fast or slow codons in mRNA sequence will be established using the codon adaptation index (CAI) described by Sharp and Li [39]. 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:

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

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

Hypotheses

We are currently interested in several hypotheses and open questions:

1. **Secondary structure** Are particular secondary structural elements (SSEs) enriched for slow codons? Are the transitions between SSEs, in agreement with Saunders and Deane [38], enriched for slow codons?

   To investigate this hypothesis, we will use Define Secondary Structure of Proteins (DSSP) [23, 24] to identify SSEs. Although DSSP is currently outperformed in accuracy by sophisticated machine learning

methods, the use of machine learning algorithms such as PROTEUS presents significant pitfalls in performing statistical analysis because decisions by the classifier depend on the training data, and individual classification decisions are unstable subject to arbitrary bounds. Because our methodology depends so heavily on such analysis to improve overall sensitivity, we consider this loss of statistical rigor unacceptable.

We will first attempt to verify the results of Saunders and Deane [38] by duplicating their approach. Specifically, we will calculate the distribution of codon usage bias within helices, strands, and coils.We will apply a Mantel–Haenszel and a χ^2-test to determine significance between the categories. Secondly, we select 2–5 large data sets (at least 2,000 genes each), and expand the categories to also include, at a minimum, 3_{10}-helices, β-helices, and π-helices. We will also distinguish between β-sheets and isolated β-strands. We will then determine significance using the same statistical techniques.

To address the second part of this question, we will calculate codon usage bias around the transitions between SSEs, using windows of 0, 1, 2, 4, and 6 residues.

Preliminary data: The research by Gu et al. [17], Oresic et al. [35], Saunders and Deane [38], Thanaraj and Argos [45] all found statistically significant correlations between codon usage bias and protein secondary structure.

2. **Domain boundaries** Are domain boundaries enriched for slow codons? To investigate this question, we will use domain classification by the Structural Classification of Proteins version 2 (SCOP2) [2, 32]. As a manually curated database, SCOP2 is very reliable for domain assignment. Although we admit that its coverage is limited (167,547 domains in 59,514 protein structures, about 62% of the PDB), we argue that will still result in sufficiently large data sets.

   Using the intersection of SCOP and the data sets described above, we will calculate codon usage bias near and apart from domain boundaries, using windows of 0, 2, 5, 8, and 12 residues from the domain boundary positions. We will then apply a Mantel–Haenszel test to determine significance.

   Preliminary data: The research by Gu et al. [18], Oresic et al. [35] showed an enrichment, though Saunders and Deane [38] found no statistically significant difference. Because all three studies used different methods, the work by Saunders et al. did not supersed the results by Gu et al. and Oresic et al; rather, there is still evidence that a signifant correlation exists.

3. **Structural complexity and knotting**

   Due to the complexity of the energy funnels for the folding of many proteins [7, 34], we hypothesize that proteins with complex tertiary structures are enriched for slow codons. In particular, we hypothesize that residues with more interactions with other residues in a protein are more likely to be encoded for by slow codons, and that proteins with more self-interactions—that is, have more interactions to overcome during the folding process—are enriched for slow codons.

   We also conjecture that knotted proteins are enriched for slow codons, and that codon usage bias is inversely correlated with the complexity of the knot. To investigate these hypotheses, we will identify knots using the existing algorithms by Lai et al. [28], Virnau et al. [47]. Although neither algorithm can detect all knots, we argue that this is sufficient for our purposes.

   Knots are classified up to isomorphism by the Jones polynomial [22], a knot invariant that revolutionized knot theory in part because it is easy to compute. It is defined in terms of an arbitrary projection D of a knot K by:

$$f_D(A) = (-A^3)^{-w(D)}\langle D\rangle \tag{4}$$

   where $w(D)$ is the writhe of the diagram D, and $\langle D\rangle$ is its Kauffman bracket.

The result of the computation is a Laurent polynomial that uniquely identifies the knot. Moreover, the number of terms in the Laurent polynomial and its coefficients can be used to define a complexity $\xi(K)$ of the knot. We hypothesize that the knot complexity ξ is associated with the CAI of a gene.

Preliminary data: Using the same data set of 437 S. cerevisiae genes, we investigated the correlation between codon usage bias and two simple measures of structural complexity:

a) Sequence length

b) The average squared distance between all residues:

$$\frac{1}{n^2}\sqrt{\sum_{R\in A}\sum_{S\in A}||R-S||_2^2} \tag{5}$$

where A is a single protein, and $||R-S||_2$ is the two-norm of the difference between the vectors corresponding to residues R and S.

TODO: Fill in when run finishes [Douglas]

In the case of structural complexity, we are interested in statistical associations that are potentially nonlinear. For this purpose we will use the Hilbert–Schmidt Independence Criterion (HSIC) [**?**] to identify statistical associations, both linear and nonlinear.

The HSIC is a statistical test defined using a generalization of the Frobenius norm for linear operators called

The Hilbert–Schmidt (HS) norm is an operator norm defined for an arbitrary operator $C : G \to F$:

: for an operator $C : G \to F$, where G and F are required only to be orthonormal bases of separable Hilbert spaces.

$$||C||_{HS}^2 = \sum_{i,j}\langle C\nu_i, \mu_j\rangle_F^2 \tag{6}$$

where ν and μ are orthonormal bases of F and G. The HS norm is extremely similar to the Frobenius norm for matrices. However, the generalization permits its use for arbitrary operators rather than only linear operators (specifically, the only requisite is that F and G are Hilbert spaces). A cross-covariance operator C_{xy} is then defined, and the HSIC is defined as:

$$HSIC(p_{xy}, F, G) := ||C_{x,y}||_{HS}^2 \tag{7}$$

where p is a probability measure. [41] The HSIC is a measure of statistical dependence: the higher the HSIC between F and G, the greater the dependence between F and G. Moreover, the HSIC has been proved to be a measure of arbitrary statistical dependence Song et al. [42]. Thus the advantage of this method is that it permits the determination of nonlinear associations without the use of a kernel trick or artificial regularization. This should allow us to identify potential relationships between protein complexity and codon usage bias in a manner that is sensitive and statistically robust, and that does not depend on an arbitrary selection of kernels or regularization terms.

The approach of Song et al. [42] for feature selection and optimization of the HSIC was shown to have good results, and we will abide by this procedure.

Analysis of accuracy

Many proteins will be examined over a number of species. While our preliminary analysis has been performed in approximately 500 yeast proteins, we hope to analyze over 10,000 human proteins in hopes of maximizing sample size.

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. [29]

Aim 2

We also aim to examine existing proteins that have been demonstrated to be affected by the removal of slow codons, and characterize in what circumstances a protein could be reasonably assumed to require slow codons for normal function.

We will run our analysis on species-independent protein families that contain at least one protein shown experimentally to require slow codons. We will complement our structural analysis with a phylogenetic examination of conserved regions to see if the bias towards slow codons is strictly evolutionarily enforced or if it is an isolated development in some species.

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. [49]

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] a. Andreeva, D. Howorth, C. Chothia, E. Kulesha, and a. G. Murzin. SCOP2 prototype: a new approach to protein structure mining. *Nucleic Acids Research*, pages 1–5, November 2013. ISSN 0305-1048. doi: $10.1093/\mathrm{nar/gkt1242}$. URL http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkt1242.

[3] 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.

[4] 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: $10.1038/\mathrm{msb.2013.32}$. URL http://www.ncbi.nlm.nih.gov/pubmed/23774758.

[5] H M Berman, T Battistuz, T N Bhat, W F Bluhm, P E Bourne, K Burkhardt, Z Feng, G L Gilliland, L Iype, S Jain, P Fagan, J Marvin, D Padilla, V Ravichandran, B Schneider, N Thanki, H Weissig, J D Westbrook, and C Zardecki. The Protein Data Bank. *Acta, Crystallogr. D. Biol. Crystallogr.*, 58:899–907, May 2002.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

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

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

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

[22] VFR Jones. The Jones polynomial. *preprint*, pages 1–21, 2005. URL http://www.math.berkeley.edu/~vfr/jones.pdf.

[23] Robbie P Joosten, Tim a H te Beek, Elmar Krieger, Maarten L Hekkelman, Rob W W Hooft, Reinhard Schneider, Chris Sander, and Gert Vriend. A series of PDB related databases for everyday needs. *Nucleic acids research*, 39(Database issue):D411–9, January 2011. ISSN 1362-4962. doi: 10.1093/nar/gkq1105. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3013697&tool=pmcentrez&rendertype=abstract.

[24] W Kabsch and C Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–637, December 1983. ISSN 0006-3525. doi: 10.1002/bip.360221211. URL http://www.ncbi.nlm.nih.gov/pubmed/6667333.

[25] 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.

[26] 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.

[27] 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.

[28] Yan-Long Lai, Chih-Chieh Chen, and Jenn-Kang Hwang. pKNOT v.2: the protein KNOT web server. *Nucleic acids research*, 40(Web Server issue):W228–31, July 2012. ISSN 1362-4962. doi: 10.1093/nar/gks592. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3394322&tool=pmcentrez&rendertype=abstract.

[29] Wei Liu, Anuj Srivastava, and Jinfeng Zhang. A mathematical framework for protein structure comparison. *PLoS Comput Biol*, 7(2):e1001075, 2011.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] José Nelson JN Onuchic and Peter G PG Peter G Wolynes. Theory of protein folding. *Current opinion in structural biology*, 14(1):70–5, February 2004. ISSN 0959-440X. doi: 10.1016/j.sbi.2004.01.009. URL http://www.ncbi.nlm.nih.gov/pubmed/15102452http://www.sciencedirect.com/science/article/pii/S0959440X04000107.

[35] 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.

[36] 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.

[37] 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.

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

[39] 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.

[40] 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.

[41] Le Song, Alex Smola, and Arthur Gretton. Feature selection via dependence maximization. *The Journal of Machine ...*, 13:1393–1434, 2012. URL http://dl.acm.org/citation.cfm?id=2343691.

[42] Le Song, Alex Smola, and Arthur Gretton. Feature selection via dependence maximization. *The Journal of Machine ...*, 13:1393–1434, 2012. URL http://dl.acm.org/citation.cfm?id=2343691.

[43] 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.

[44] 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.

[45] 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.

[46] 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.

[47] Peter Virnau, Leonid a Mirny, and Mehran Kardar. Intricate knots in proteins: Function and evolution. *PLoS computational biology*, 2(9):e122, September 2006. ISSN 1553-7358. doi: 10.1371/journal.pcbi.0020122. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1570178&tool=pmcentrez&rendertype=abstract.

[48] 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.

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

[50] 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.

[51] 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.