Genomics and Bioinformatics MBB 452a / 752a
Final Project 2001

Codon usage and the prediction of gene expressivity

Michael Seringhaus
Yale University
Department of Molecular Biophysics and Biochemistry
michael.seringhaus@yale.edu

Due December 13, 2001

Instructor: Mark Gerstein
Assistants: Dov Greenbaum, Jeff Sabina

Overview of codon usage

Most genes in a genome or genome type share the same coding strategy, or preference among synonymous codons, distinct from the pattern of frequency variation in amino acid selection [1]. This preference was first proposed as a strategy to regulate gene expression in 1981, when Grantham *et al.* suggested selection at the third base position achieves regulation of mRNA expressivity [2], and Ikemura proposed that preferred codons correlate with the abundance of complementary tRNAs and serve to optimize the translational system in *E.coli* [3,4]. The three termination codons are not recognized by tRNAs, but nonetheless show preference in their synonymous selection [5]. A strong correlation was quickly confirmed between codon composition and mRNA expressivity, even within the same operon – codon choice was said to influence peptide elongation rate and protein yield, thus affecting gene expression [6].

While for many organisms, a strong correlation can be shown between codon preference and gene expressivity, specific codon preferences – and pool sizes of corresponding tRNA – vary between species and genome. Yeast and *E. coli*, for instance, both show strong preferences for a certain subset of their available codons; there is little overlap between these preferred subsets when comparing the two organisms, though both correlate well with tRNA levels in the respective species [7].

Can codon usage be used to predict gene expression?

The first report that gene expressivity could successfully be predicted by codon usage patterns came in 1986, when Sharp *et al.* used clustering to predict expression of 110 Yeast genes [8]. The Codon Adaptation Index (CAI), introduced shortly after, is a measure of directional synonymous-usage codon bias that can be used to predict gene expression, using a reference set of highly expressed genes to score relative preference of individual codons [9,10]. A prominent experimental approach came in 1987, when researchers at Genentech undertook systematic replacement of 39% of all codons in the highly expressed yeast gene PGK1, changing them to rare codons; protein expression dropped tenfold, while mRNA expression dropped only threefold, demonstrating that rare codons do have a regulatory effect on gene expression independent of mRNA levels [11]. This lent support to the general notion that gene expression could be predicted by codon usage patterns.

The caveat to such predictive methods is that certain organisms (*Streptomyces*, *Pseudomonas*) show little or no correlation between codon usage and expressivity; while these organisms show codon bias, it relates more to GC-content or other factors [12,13]. Some researchers have gone so far as to deny any relationship between codon usage and expression whatsoever, proposing that codon preference is instead indicative of a global system to maximize the effectiveness of translational machinery [14]. Rogue theories aside, it has been repeatedly and consistently shown that a strong correlation exists between codon usage and gene expression in a variety of organisms – whether this represents causality or mere correlation remains open to debate, but this relationship has been used successfully in a predictive capacity for over a decade.

Computer / Statistical Applications

Given that a correlation exists between codon usage and gene expression in most model organisms studied to date, it is unsurprising that researchers have focused on refining the CAI and developing computer applications to assist in codon bias evaluation and prediction of gene expression: deriving predictive information from entire genomes at a time is, after all, the perfect application of computer technology to biology. Such work has covered the various methods of counting codon usage [15,16], context analysis of a codon's environment from a broad statistical standpoint [17,31], and examination of bias along the length of a gene sequence [18]. Computer programs were available for codon counting and bias index determination as early as 1992 [19,20], and more recently, hidden Markov model approaches have been applied to codon usage data to predict chimerism in proteins [21]. Codon usage has also been used to predict localization of eukaryotic ribosomal proteins and tRNA synthetases [22]. GenBank sequence data was first used in codon usage tabulations in 1991 [30].

The harvest of new information from the application of these techniques has been extensive and intriguing. For example, computer analyses determined that the CG dinucleotide is found almost exclusively in the eight rarest codons in yeast and primates [15] – this has recently been suggested to correlate with transcriptional regulation via methylation of CG dinucleotides, a phenomenon seen in mammalian genomes [23]. Codon bias has been shown to be less extreme near the beginning of a genetic sequence [18]. Comeron and Aguadé found through computer simulation that methods of tabulating global codon usage are subject to bias, depending on the lengths of the gene sequences used to make the predictions [16].

With the recent flood of sequence and genomic expression data from SAGE and microarray experimentation, the door is open to more comprehensive study of codon usage and its correlation to gene expression. Specifically, study of codon usage patterns offers an interesting and potentially powerful tool to explore the puzzling differences between mRNA and protein expression levels – since regulation of expressivity by codon preference should occur only at the translational level, these analyses may bridge the gap, enabling predictions of true protein expression from observed mRNA levels. Such an approach, if successful, could be very useful in predicting highly expressed or even essential subsets of genes in new genomes [24].

Summary

The trend initially elucidated by Toshimichi Ikemura from merely several dozen gene sequences has stood the test of time; over two decades later, we can demonstrate even more clearly the correlation between codon preference, tRNA pool levels and gene expressivity. Codon usage is widely accepted to correlate with protein expression in many organisms (recently, *C. elegans* was shown to adhere to this trend [29]), and this correlation can theoretically be exploited to predict gene expression from sequence, given a valid CAI and other organism-specific parameters.

Computer techniques have been employed, both to assemble indexes like the CAI, and to probe the characteristics of these correlations. With the advent of genomic data, we must consider carefully the effectiveness of computer predictions, since given enough data, algorithms can find nearly any pattern we search for. This is reflected in recent work, where codon bias has been shown to correlate to nearly any biological trend that researchers have sought (protein expression, mRNA concentration, amino acid conservation, G-C content, secondary structure, protein localization, hydrophobicity, protein length) [12,13,22,25,26,27,28]. Computer techniques can help us formulate relevant questions, but to pose these effectively we should still engage in experimental approaches with *in vivo* systems. Surprisingly little experimentation has been done in this area, presumably because of the significant work required to undertake large-scale codon-replacement or tRNA-anticodon acceptor mutation. However, armed with fully sequenced and annotated genomes, powerful methods of gene modification and replacement, microarray / SAGE approaches to mRNA expression characterization and protein quantification techniques, the stage is set for significant experimental analysis of codon bias. This review of the study of codon usage highlights what I consider to be the main forte of bioinformatics – analyzing trends and generating predictions from masses of data, which then lay the groundwork for further discovery by experimentation.

References Cited

1. Grantham R, Gautier C, Gouy M. Codon frequencies in 119 individual genes confirm consistent choices of degenerate bases according to genome type. *Nucleic Acids Res* 1980 May 10; 8(9):1893-912.

2. Grantham R, Gautier C, Gouy M, Jacobzone M, Mercier R. Codon catalog usage is a genome strategy modulated for gene expressivity. *Nucleic Acids Res* 1981 Jan 10;9(1):r43-74.

3. Ikemura T. Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes. *J Mol Biol* 1981 Feb 15;146(1):1-21.

4. Ikemura T. Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the E. coli translational system. *J Mol Biol* 1981 Sep 25;151(3):389-409.

5. Bienz M, Kubli E, Kohli J, deHenau S, Huez G, Marbaox G, Grosjean H. Usage of the three termination codons in a single eukaryotic cell, the Xenopus laevis oocyte. *Nucleic Acids Res* 1981 Aug 11;9(15):3835-50.

6. Gouy M, Gautier C. Codon usage in bacteria: correlation with gene expressivity. *Nucleic Acids Res* 1982 Nov 25;10(22):7055-74.

7. Ikemura, T. Codon usage and tRNA content in unicellular and multicellular organisms. *Mol Biol Evol* 1985; 2(1):13-14.

8. Sharp PM, Tuohy TM, Mosurski KR. Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. *Nucleic Acids Res* 1986 Jul 11;14(13):5125-43.

9. Sharp PM, Li WH. An evolutionary perspective on synonymous codon usage in unicellular organisms. *J Mol Evol* 1986;24(1-2):28-38.

10. Sharp PM, Li WH. The Codon Adaptation Index – a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res* 1987 Feb 11;15(3):1281-95.

11. Hoekema A, Kastelein RA, Vasser M, de Boer HA. Codon replacement in the PGK1 gene of Saccharomyces cerevisiae: experimental approach to study the role of biased codon usage in gene expression. *Mol Cell Biol* 1987 Aug;7(8):2914-24.

12. West SE, Iglewski BH. Codon usage in Pseudomonas aeruginosa. *Nucleic Acids Res* 1988 Oct 11;16(19):9323-35.

13. Wright F, Bibb MJ. Codon usage in the G+C-rich Streptomyces genome. *Gene* 1992 Apr 1;113(1):55-65.

14. Andersson SG, Kurland CG. Codon preferences in free-living microorganisms. *Microbiol Rev* 1990 Jun;54(2):198-210.

15. Zhang SP, Zubay G, Goldman E. Low-usage codons in Escherichia coli, yeast, fruit fly and primates. *Gene* 1991 Aug 30;105(1):61-72.

16. Comeron JM, Aguadé M. An evaluation of measures of synonymous codon usage bias. *J Mol Evol* 1998;47:268-274.

17. Bulmer M. The effect of context on synonymous codon usage in genes with low codon usage bias. *Nucleic Acids Res* 1990 May 25;18(10):2869-73.

18. Bulmer M. Codon usage and intragenic position. *J Theor Biol* 1988 Jul 8;133(1):67-71.

19. Lloyd AT, Sharp PM. CODONS: a microcomputer program for codon usage analysis. *J Hered* 1992 May-Jun;83(3):239-40.

20. Wang TT, Cheng WC, Lee BH. A simple program to calculate codon bias index. *Mol Biotechnol* 1998 Oct;10(2):103-6.

21. Hunter L, Zeeberg B. Identifying chimerism in proteins using hidden Markov models of codon usage. *Proc Int Conf Intell Syst Mol Biol* 1997;5:153-6.

22. Chiapello H, Ollivier E, Landes-Devauchelle C, Nitschke P, Risler JL. Codon usage as a tool to predict the cellular location of eukaryotic ribosomal proteins and aminoacyl-tRNA synthetases. *Nucleic Acids Res* 1999 Jul 15;27(14):2848-51.

23. Kanaya S, Yamada Y, Kinouchi M, Kudo Y, Ikemura T. Codon usage and tRNA genes in eukaryotes: correlation of codon usage diversity with translational efficiency and with CG-dinucleotide usage as assessed by multivariate analysis. *J Mol Evol* 2001;53:290-298.

24. Seringhaus M, Jansen R, Gerstein M. Unpublished results.

25. de Miranda AB, Alvarez-Valin F, Jabbari K, Degrave WM, Bernardi G. Gene expression, amino acid conservation, and hydrophobicity are the main factors shaping codon preferences in Mycobacterium tuberculosis and Mycobacterium leprae. *J Mol Evol* 2000;50:45-55.

26. Chuisano ML, Alvarez-Valin F, Di Guilio M, D'Onofrio G, Ammirato G, Colonna G, Bernardi G. Second codon positions of genes and the secondary structures of proteins. Relationships and implications for the origin of the genetic code. *Gene* 2000; 261:63-69.

27. Coghlan A, Wolfe KH. Relationship of codon bias to mRNA concentration and protein length in Saccharomyces cerevisiae. *Yeast* 2001;16:1131-1145.

28. Gygi SP, Rochon Y, Franza R, Aebersold R. Correlation between protein and mRNA abundance in yeast. *Mol Cell Biol* 1999 Mar;19(3):1720-1730.

29. Duret L. tRNA gene number and codon usage in the C. elegans genome are co-adapted for optimal translation of highly expressed genes. *Trends Genet* 2000 July;16(7)287-289.

30. Wada K, Wada Y, Doi H, Ishibashi F, Gojobori T, Ikemura T. Codon usage tabulated from the GenBank genetic sequence data. *Nucleic Acids Res* 1991 Apr 25;19 Suppl:1981-86.

31. Buckingham RH. Codon context. *Experientia* 1990 Dec 1;46(11-12):1126-33.

Other References

Andersson SGE, Kurland CG. An extreme codon preference strategy: codon reassignment. *Mol Biol Evol.* 1991;8(4):530-544.

Bulmer M. Coevolution of codon usage and transfer RNA abundance. *Nature*. 1987 Feb 19-25;325(6106):728-30.

Diaz-Lazcoz Y, Henaut A, Vigier P, Risler JL. Differential codon usage for conserved amino acids: evidence that the serine codons TCN were primordial. *J Mol Biol*. 1995 Jul 7;250(2):123-7.

Forsburg SL. Codon usage table for Schizosaccharomyces pombe. *Yeast*. 1994 Aug;10(8):1045-7.

Kanaya S, Yamada Y, Kudo Y, Ikemura T. Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of Bacillus subtilis tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis. *Gene.* 1999 Sep 30;238(1):143-55.

Kim CH, Oh Y, Lee TH. Codon optimization for high-level expression of human erythropoietin (EPO) in mammalian cells. *Gene.* 1997 Oct 15;199(1-2):293-301.

Kunisawa T. Synonymous codon preferences in bacteriophage T4: a distinctive use of transfer RNAs from T4 and from its host Escherichia coli. *J Theor Biol* 1992 Dec 7;159(3):287-98.

Li H, Luo L. The relation between codon usage, base correlation and gene expression level in Escherichia coli and yeast. J Theor Biol. 1996 Jul 21;181(2):111-24.

Lloyd AT, Sharp PM. Synonymous codon usage in Kluyveromyces lactis. Yeast. 1993 Nov;9(11):1219-28.

Lloyd AT, Sharp PM. Evolution of codon usage patterns: the extent and nature of divergence between Candida albicans and Saccharomyces cerevisiae. Nucleic Acids Res. 1992 Oct 25;20(20):5289-95.

Miyasaka H. The positive relationship between codon usage bias and translation initiation AUG context in Saccharomyces cerevisiae. Yeast. 1999 Jun 15;15(8):633-7.

Sharp PM, Cowe E. Synonymous codon usage in Saccharomyces cerevisiae. Yeast. 1991 Oct;7(7):657-78.

Sharp PM, Stenico M, Peden JF, Lloyd AT. Codon usage: mutational bias, translational selection, or both? Biochem Soc Trans. 1993 Nov;21(4):835-41.