

# Adaptive basis of codon usage in the haploid moss *Physcomitrella patens*

HK Stenøien

Plant Ecology/Department of Ecology and Evolution, Evolutionary Biology Centre, Uppsala University, Villav. 14, Uppsala SE-752 36, Sweden

Patterns of codon usage bias were studied in the moss model species *Physcomitrella patens*. A total of 92 nuclear, protein coding genes were employed, and estimated levels of gene expression were tested for association with two measures of codon usage bias and other variables hypothesized to be associated with gene expression. Codon bias was found to be positively associated both with estimated levels of gene expression and GC content in the coding parts of studied genes. However, GC content in noncoding parts,

that is, introns and 5' and 3' untranslated regions (UTRs), was not associated with estimated levels of gene expression. It is argued that codon bias is not shaped by mutational bias, but rather by weak natural selection for translational efficiency in *P. patens*. The possible role of life history characteristics in shaping patterns of codon usage in this species is discussed.

*Heredity* (2005) 94, 87–93. doi:10.1038/sj.hdy.6800547  
Published online 13 October 2004

**Keywords:** *Arabidopsis*; bryophyte; evolution; gene expression; selection; translation

## Introduction

The extent and origin of codon usage bias is an important indicator of the forces shaping genome evolution in living organisms. Several studies of codon bias have revealed differences in how evolutionary forces act among and within genomes. Codon bias is generally thought to be the result of the balance between mutation and weak selection on synonymous codons (Li, 1987; Bulmer, 1991; Akashi and Eyre-Walker, 1998; Duret and Mouchiroud, 1999). Differences in the codon bias between genes in the same organism is attributed to mutational bias and the variability in guanine + cytosine (GC) content throughout the genome (Ikemura, 1985), partly due to the dispersion of large isochores homogeneous for GC content (Bernardi *et al*, 1985). In many organisms, codon usage in highly expressed genes is shown to be dependent on the relative proportions of isoaccepting tRNAs. Thus, the degree of codon bias for individual genes is associated with the level of translation in many unicellular (Ikemura, 1981, 1985; Gouy and Gautier, 1982; Andersson and Kurland, 1990; Sharp and Matassi, 1994; Kanaya *et al*, 1999) and multicellular organisms (Fennoy and Bailey-Serres, 1993; Akashi, 1994; Moriyama and Powell, 1997; Akashi and Eyre-Walker, 1998; Duret and Mouchiroud, 1999; Duret, 2000; Kanaya *et al*, 2001; Musto *et al*, 2001). In contrast, mutation pressure has been shown to be the dominant factor shaping both codon usage bias and base composition in, for example, several mammalian genomes (Wolfe *et al*, 1989; Sharp *et al*, 1993; Francino and Ochman, 1999). The

lack of translational selection in mammals and some *Drosophila* species has been explained by their relatively small effective population sizes, meaning that genetic drift will dominate the evolutionary dynamics of mutations that only differ marginally in fitness (Shields *et al*, 1988; Sharp *et al*, 1995; Akashi, 1997; Jenkins and Holmes, 2003). Also, other factors are associated with codon usage including gene length (Comeron *et al*, 1999; Duret and Mouchiroud, 1999; Coghlan and Wolfe, 2000), recombination levels across genomes (Kliman and Hey, 1993; Hey and Kliman, 2002), gene density (Hey and Kliman, 2002) and gene structure (Comeron and Kreitman, 2002).

Even though genome-wide patterns of adaptive codon bias have been convincingly demonstrated in a wide range of phyla, relatively few higher plant and animal species have been studied. In plants, studies of selection on codon usage have been undertaken in *Zea mays* (Fennoy and Bailey-Serres, 1993), *Nicotiana tabacum* and the liverwort *Marchantia polymorpha* (Morton, 1993, 1994, 1998), and *Arabidopsis thaliana* (Duret and Mouchiroud, 1999). Studies of more species are needed in order to determine how widespread adaptive codon usage is in plants.

*Physcomitrella patens* (Hedw.) Bruch & Schimp has been established as an important model species of primitive moss plants. Major progress has been made in recent years in analysing the molecular biology and physiology of this species (eg Markmann-Mulisch *et al*, 2002; Bezanilla *et al*, 2003). A substantial breakthrough was achieved with the demonstration that homologous recombination occurs in its nuclear DNA at frequencies which are orders of magnitude greater than those known from higher plants (Schaefer and Zryd, 1997). This makes *Physcomitrella* the only land plant with an efficient system for reverse genetics (Strepp *et al*, 1998), that is, introduction of precisely predetermined DNA sequence alterations into cultured organisms to study their phenotypic effects. The gene targeting efficiency in *P.*

*patens* can be compared with that observed in *Saccharomyces cerevisiae*, making *P. patens* the new 'green yeast' (Schaefer, 2001). The gene target efficiency, together with several other aspects of the *P. patens* biology (eg small size, short life cycle, easy to cultivate, basal phylogenetic position and thus ideal for comparative studies), have led different research groups to advocate *P. patens* as the next plant to have its genome completely sequenced (Pryer *et al*, 2002; Rensing *et al*, 2002; Schaefer, 2002). In August 2004, scientists from The United States, England, Japan and Germany were successful in receiving approval for sequencing the *Physcomitrella* genome.

The aim of the present study was to quantify patterns of codon usage in this important model plant. Associations between gene expression (estimated by mRNA transcript abundance), codon bias, GC content and gene length in a number of studied genes were tested using multiple regression analysis. The results indicate that levels of gene expression are associated with codon bias and GC content in *P. patens*. The association between codon bias and estimated gene expression seems not to be caused by mutational bias, but rather by natural selection for translational efficiency.

## Materials and methods

*P. patens* sequences were downloaded from GenBank release 131.0 and EMBL release 72 by using the NCBI Entrez retrieval systems (<http://www.ncbi.nlm.nih.gov/Entrez/>). Only nuclear genes, being fully annotated and protein coding were retrieved, excluding working drafts and expressed sequence tags (ESTs). Similar sequences were removed from the data set using the software Dambe version 4.1.15 (Xia and Xie, 2001). The final data set consisted of 92 *P. patens* sequences (the complete data set can be accessed at <http://www.vaxtbio.uu.se/resfold/stenoien/physcomitrella/>).

Gene expression was estimated as the number of hits when comparing a sequence with publicly available ESTs for *P. patens* (see Duret and Mouchiroud, 1999; Marais and Duret, 2001; Hey and Kliman, 2002; Miyasaka, 2002). All available cDNA libraries in release 101502 of the NCBI dbEST database (Boguski *et al*, 1993) were used, comprising 50 250 *P. patens* ESTs. Each sequence was compared with the available cDNA libraries using MegaBlast (Zhang *et al*, 2000). Both strands were compared for each sequence, and segments within query sequences showing low compositional complexity were masked. Blastn alignments showing at least 95% identities were counted as sequence match when significance level was set to  $E = 0.001$ .

Codon bias was estimated as the frequency of optimal codons (Fop, Ikemura, 1981). Optimal codons are defined as triplets, which relative to their synonyms occur significantly more often in highly than in lowly expressed sequences. Multivariate correspondence analysis was performed to identify putative optimal codons, that is, codons occurring more often in highly than in lowly biased genes in *P. patens*. A  $\chi^2$  test was performed in order to test whether codon usage differed among highly and lowly expressed genes. Here, two subsets of genes were drawn from the data set, a set of genes with low expression (ie nine genes only matching one EST) and a set of genes with high expression (ie six genes matching more than 50 ESTs), representing 4095 and 1759 codons,

respectively. A  $\chi^2$  contingency analysis was performed to compare codon usage frequencies between the two categories of genes. Codon usage bias was also estimated by the effective number of codons (ENC) summary statistics (Wright, 1990). This measure does not require knowledge of optimal codons and is inversely proportional to the extent of nonuniform codon usage. Thus, ENC ranges from 20, indicating maximum codon bias, to 61, indicating equal usage of all synonymous codons. Multivariate analysis and codon bias estimations were carried out by using the CodonW software (<http://www.molbiol.ox.ac.uk/cu/>, J. Peden, unpublished). GC content was measured in coding and noncoding regions, that is, GC content overall in coding regions, GC content in third codon positions (N3) and GC content in untranslated regions of a gene (5' untranslated regions (UTRs), 3' UTRs and introns). The various measures of GC content were calculated with DnaSP version 3.53 (Rozas and Rozas, 1999).

Two stepwise multiple regressions were performed, one for each measure of codon bias. In both regression analyses, the estimated gene expression was defined as dependent variable, and codon bias (Fop and ENC), gene length, GC content overall in coding region and GC content in N3 as independent variables. The significance of the regression models was tested with analysis of variance (ANOVA). Kolmogorov–Smirnov and Shapiro–Wilks tests on studentized residuals revealed departure from normality, and the dependent variables were therefore ln-transformed. Studentized residuals for the regression analyses were plotted against the various independent variables in order to check for linearity between dependent and independent variables. Observed and predicted values of gene expression for the various genes were plotted for the regression analyses in order to check whether the variance of residuals was constant over the entire range of predicted values. Cook's distances were computed to measure how regression coefficients changed when single genes were eliminated from the analyses. Multicollinearity was investigated with the tolerance statistic, that is, the proportion of variability of a given independent variable not explained by its linear relationships with other independent variables in the model. Student's *t*-test and Pearson's analyses were performed to further study the associations among the variables.

To test for associations between transcript abundance and substitution rates, putative *A. thaliana* genes orthologous to the studied *P. patens* genes were identified through tBlastx searches (Altschul *et al*, 1990). Each gene pair was aligned by the Align algorithm (Myers and Miller, 1989) as implemented in the Biology Workbench 3.2 web station (<http://workbench.sdsc.edu/>) and by visual inspection. Substitution patterns were investigated between aligned sequence pairs. Thus, the number of substitutions per synonymous site ( $K_s$ ) was estimated for aligned sequence pairs using the K-Estimator ver. 6.0 software (Comeron, 1999). Kimura's two-parameter model was used to correct for multiple hits (Kimura, 1980).

In order to investigate whether this sample of *P. patens* genes represented the whole genome, a similar analysis was conducted on the orthologous *A. thaliana* genes. Gene expression levels, codon biases and GC content measures were estimated for the homologous *A. thaliana* genes in a similar way to the *P. patens* genes (see above),

except that Fop was calculated from previously determined optimal codons (Chiapello *et al* (1998); the *A. thaliana* data set can be accessed at <http://www.vaxtbio.uu.se/resfold/stenoien/physcomitrella/>). Statistical analysis on the *A. thaliana* genes was performed similarly to the procedures employed on the *P. patens* genes.

Statistical analyses were conducted using the SPSS version 11.0.0 software (SPSS Inc., Chicago, USA).

## Results

The two codon bias measures Fop and ENC are negatively correlated ( $-0.213$ ,  $P=0.043$ ), since ENC is inversely proportional to the degree of codon bias. The subsets of lowly and highly expressed genes have a significantly different codon usage pattern according to a  $\chi^2$  test ( $P<0.0001$ ).

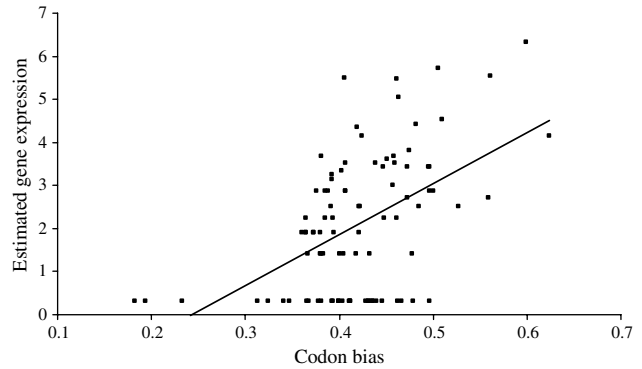
The number of EST hits was ln-transformed prior to analyses. For both multiple regression analyses, the studentized residuals were normally distributed according to both Kolmogorov-Smirnov (with Lilliefors' significance correction) and Shapiro-Wilks tests. Studentized residuals for both regression analyses were plotted against the various independent variables in order to check for linearity between dependent and independent variables. The various scatterplots show a random distribution of residuals around 0 on the *y*-axis, indicating a linear relationship between the various variables (data not shown). Observed and predicted values of transcript abundance for the various genes were plotted for the two regression analyses, and scatterplots indicate constant variance of residuals (data not shown). Both regression models are significant according to an ANOVA test ( $P<0.001$ ; see Table 1). Codon usage bias and GC content overall are significantly associated with the levels of transcript abundance, both when Fop and ENC are employed in the tests (Figure 1).

Cook's distances were obtained for both regression analyses. Cook's distance for a gene depends on both the residual and leverage values, that is, how far the values for a case are from the means of all the independent variables. Cook's distances greater than 1 are considered to deserve some scrutiny (Norusis, 2000). No Cook's distances were found to be larger than 0.06 in the two regression analyses, indicating that no single gene

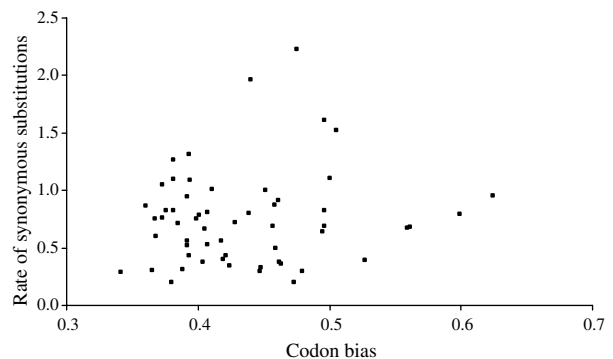
**Table 1** Results of the two multiple regression analyses on variables associated with estimated levels of gene expression in 92 *P. patens* genes

Codon usage bias measure	Included variables in model	Standardized beta coefficients	t	P-value
Fop	GC content overall	0.317	2.632	0.010
	Codon bias	0.298	2.468	0.015
ENC	GC content overall	0.477	5.181	<0.001
	Codon bias	-0.185	-2.010	0.047

mRNA transcript abundance is treated as dependent variable, and codon bias measured by Fop and ENC, GC content overall, GC content in N3 position and gene length are treated as independent variables. Only independent variables included in the regression models are presented.



**Figure 1** Codon usage bias in studied *P. patens* genes, as measured by the Fop, plotted against gene expression as estimated by mRNA transcript abundance (ln-transformed). There is a significant positive association.

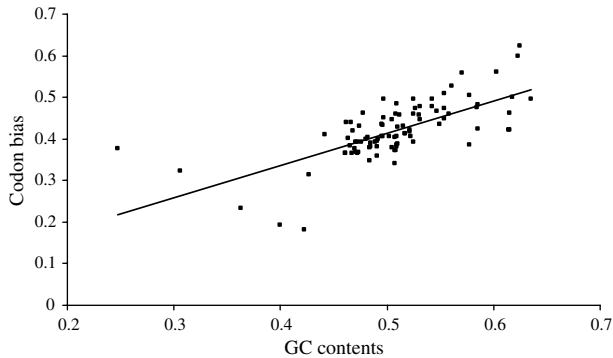


**Figure 2** Rate of  $K_s$  between homologous *P. patens* and *A. thaliana* genes plotted against codon usage bias, as measured by the Fop in *P. patens* genes. No significant trend is observed.

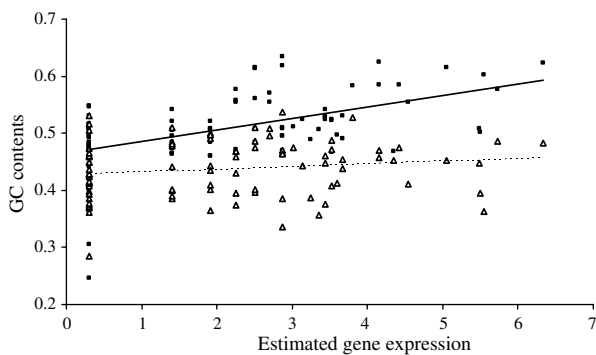
changes the results to any major degree if removed from the analyses. Tolerance levels are high for all significant independent variables ( $>0.5$ ), indicating no multicollinearity in the data sets.

Orthologous *A. thaliana* sequences alignments were attempted, and 57 alignments were generated. These 57 sequence pairs were subjected to quantification of  $K_s$ . There were no significant correlations between  $K_s$  or any measure of codon bias or estimated gene expression for the 57 studied sequence pairs (Figure 2).

Most optimal codons in *A. thaliana* end in G or C (Chiapello *et al*, 1998; Duret and Mouchiroud, 1999), and this trend also seems to be present in *P. patens*. There is a strong positive correlation between GC content in N3 and Fop (Pearson's coefficient 0.541,  $P<0.001$ ). The same positive correlation is observed between GC content overall and Fop (Pearson's coefficient 0.688,  $P<0.001$ , Figure 3). Both GC levels in third codon positions and GC levels overall increase with expression levels (Pearson's correlation coefficients 0.46 and 0.41, respectively,  $P<0.001$ ). There is no significant increase in GC content in untranslated 5'- and 3'-flanking UTRs or introns with increasing gene expression. Figure 4 shows gene expression plotted against GC contents overall in coding and noncoding regions of the studied genes. There is on average an 8% higher GC content in coding



**Figure 3** GC content overall of studied *P. patens* genes plotted against codon bias, as measured by Fop. There is a significant positive association.



**Figure 4** Scatterplot of GC content in coding regions (black squares) and GC content in introns, 5' and 3' UTRs (white triangles) of studied *P. patens* genes plotted against estimated levels of gene expression. Gene expression is estimated as abundance of homologous ESTs to a given gene in public databases, and the scale of the  $x$ -axis is logarithmic. The solid trend line is the regression of GC content in coding regions on EST estimates, while the stippled trend line is the regression of GC content in noncoding regions on EST estimates. There is a higher GC content in coding versus noncoding regions of studied genes, and GC content in coding regions increases with gene expression levels.

versus noncoding regions of the studied genes ( $P < 0.001$ ), and this discrepancy is even higher in highly biased genes, because of the positive correlation between GC content in coding regions and codon bias.

There is a negative correlation between gene length and codon bias in the studied genes for ENC (Pearson's correlation 0.251,  $P = 0.017$ ), but not for Fop. A similar negative correlation between codon bias and gene length has been reported in other multicellular organisms (Moriyama and Powell, 1997; Duret and Mouchiroud, 1999). A number of hypotheses have been advocated to explain this correlation, but up to now this relationship remains largely unexplained (Akashi, 2001).

For *A. thaliana* sequences, multiple regression analyses were performed with transcript abundance as dependent variable, and gene length, codon bias (Fop), GC content overall and GC content in N3 as independent variables. All together 59 homologous *A. thaliana* genes were employed. After ln transformation, the studentized residuals are normally distributed according to Kolmogorov-Smirnov (with Lilliefors' significance correction) and Shapiro-Wilks tests. The regression model is

significant according to an ANOVA test ( $P < 0.001$ ). Codon bias is also significantly positively associated with transcript abundance levels for *A. thaliana* genes ( $P < 0.001$ ,  $R^2 = 0.236$ , standardized beta coefficient 0.486 and  $t = 4.199$ ). No other variable is significantly associated with estimated transcript abundance in the *A. thaliana* genes.

## Discussion

Duret and Mouchiroud (1999) were the first to use sequence matches to EST libraries in their study of relationships between codon usage and mRNA abundance in *A. thaliana*, *Drosophila melanogaster* and *Caenorhabditis elegans*. As these authors point out, mRNA abundance estimates may be error prone because of biases in the tissues sampled, biases in cloning of mRNAs and the normalization of cDNA libraries prior to sequencing, that is, adjustment toward uniform concentrations of cDNAs from different genes, causing an underestimation of highly expressed genes. Nevertheless, broad-scale positive associations between mRNA abundance and codon bias are in concordance with results from experimental studies (Shields *et al*, 1988; Stenico *et al*, 1994; Chiapello *et al*, 1998; Akashi, 2001). Moreover, the imprecise estimates of gene expression and underestimation of highly expressed genes yield conservative tests of statistical association. Associations between, for example, codon bias and gene expression may therefore be even stronger than what is estimated in the present study, because at least some *P. patens* cDNA libraries contained in the NCBI dbEST database are normalized. Another source of bias is that expression measures that require mRNA to be maintained for some time (eg EST and SAGE) are likely prone to a GC bias, since GC-rich sequences tend to decay slower than AT-rich ones (Margulies *et al*, 2001). It is difficult to say how serious bias this might represent. It is, in any case, appropriate to do multiple regression analyses in studies based on EST matching in order to separate the effects of the various variables and exclude spurious associations. Low statistical power due to normalization, GC bias or other biases in the cDNA libraries may explain the relatively low  $R^2$  values in the present regression analyses.

There is a positive association between gene expression levels, as measured by transcript abundance, and both GC levels overall and codon usage bias in *P. patens* genes. Codons frequently employed in biased genes mostly encode G or C in N3, possibly indicating correlated effects of GC content and codon bias on transcript abundance. However, the high tolerance levels of the significant variables ( $> 0.5$ ) suggest that multicollinearity is not a problem in the regression models, and that GC content and codon bias have separate effects on transcript abundance variability.

The positive association between codon usage bias and transcript abundance may reflect mutational bias, or coadaptation between codon usage and tRNA abundance optimizing the efficiency of protein synthesis. Such selection for translational efficiency has been proposed to explain codon bias in several multicellular eukaryotes (eg Akashi, 1994; Moriyama and Powell, 1997; Akashi and Eyre-Walker, 1998; Duret and Mouchiroud, 1999; Duret, 2000; Kanaya *et al*, 2001). Morton (1993)

found that highly expressed chloroplast genes in the liverwort *M. polymorpha* tended to be biased in codon composition, and selection for codon usage was demonstrated in this species. However, in *P. patens* the mutational bias hypothesis cannot be totally rejected merely based on a transcript abundance/codon bias correlation. This is because there may be an association between transcript abundance and mutational patterns (Duret and Mouchiroud, 1999), as seen in, for example, *Escherichia coli* genes where an overall increase in the frequency of C-to-T mutations has been observed with increasing mRNA transcript levels (Beletskii and Bhagwat, 1996).

If selection affects synonymous codon usage, then the intensity of selection should be positively correlated with the transcript abundance level. Owing to this stronger purifying selection, lower rates of synonymous substitutions are expected in highly expressed genes compared with weakly expressed genes. It has been shown both in bacteria and *Drosophila* that synonymous substitution rates are lower in genes with a strong codon usage bias (highly expressed) than in other genes (Sharp and Li, 1987, 1989; Shields *et al.*, 1988; Powell and Moriyama, 1997). Therefore, one way of testing whether codon usage is selectively constrained in *P. patens* is to test whether transcript abundance levels are associated with number of substitutions per site at synonymous sites ( $K_s$ ) between *P. patens* and some other species. Unfortunately, little sequence data are available from close relatives of *P. patens*. Nishiyama *et al.* (2003) have found that at least two-thirds of *A. thaliana* genes may have homologues in *P. patens* gametophytes, and putative *A. thaliana* homologues have been identified in the present study. However, a number of these gene pairs have been difficult to align at the nucleotide level because of many substitutions and indels accumulated since the splitting of these two lineages. For 57 more-or-less convincingly aligned gene pairs, estimated  $K_s$  values do not significantly correlate with measures of codon bias and the trend even seems to be slightly positive (Figure 2). This may either indicate that codon bias is not the result of weak selection in *P. patens* or that the statistical power is too low to detect an association between  $K_s$  and codon bias measures. The latter explanation cannot be ruled out because of the difficulties in aligning homologous sequence pairs between these distantly related plants.

It is possible that highly expressed genes reside in genomic areas rich in specific residues, as observed in other organisms (eg Francino and Ochman, 1999). If specific types of mutations occur more often in highly compared to lowly expressed genes, then biased nucleotide composition can result from purely nonadaptive evolutionary forces, possibly explaining the lack of association between codon bias and  $K_s$ . However, if mutational bias is causing biased GC content at synonymous sites in highly expressed genes, it should affect coding and noncoding sites in a similar way. That is, if highly expressed genes tend to reside in GC-rich genomic regions, then the levels of GC elevation should be the same in exons and noncoding regions (ie introns and 5' and 3' UTRs). In this study, the GC content clearly increases with levels of transcript abundance in the studied genes, since the optimal codons mostly end in G and C. However, the GC level in noncoding regions is significantly lower than in coding regions. Furthermore,

there is no significant increase in GC content in noncoding regions with increased transcript abundance levels (Figure 4). If highly expressed genes do not reside in GC-rich genomic regions, then another possible explanation for the observed results is that nucleotide bias is due to a positive correlation between mutation and transcription rates (Beletskii and Bhagwat, 1996). However, GC levels are also significantly higher in exons compared to introns in the studied *P. patens* genes (*t*-test,  $P < 0.001$ ). Furthermore, there is no significant increase in C-to-T mutations with increasing mRNA transcript abundance in the present data set (data not shown) as opposed to observations in, for example, *E. coli* (Beletskii and Bhagwat, 1996). Therefore, the correlation between codon bias and transcript abundance in *P. patens* seems not to be due to any form of mutational bias. One may therefore conclude that there is an adaptive basis for codon usage in a range of different genes in *P. patens*. Adaptive codon bias also seems present in at least three studies of vascular plants, that is, *Z. mays* (Fennoy and Bailey-Serres, 1993), *N. tabacum* (Morton, 1993) and *A. thaliana* (Duret and Mouchiroud, 1999), together with the bryophyte *M. polymorpha* (Morton, 1993).

It might be questioned as to how representative a sample of 92 genes are of the complete *P. patens* genome. Biased sampling of genes may result in observed patterns, which are not necessarily informative on how evolutionary forces act on the genome as a whole. However, two features indicate that the studied genes may be relatively representative. Firstly, measures of Cook's distances do not indicate that outliers are influencing the results to any extent, so that removal of specific genes will not alter the overall results of this analysis. Secondly, 59 homologous *A. thaliana* genes were identified, and these genes were subjected to the same analyses as the *P. patens* genes. The result of the multiple regression shows that codon bias is explaining variability levels in transcript abundance for this small set of genes. This confirms the results of genomic surveys of *A. thaliana* genes by Duret and Mouchiroud (1999). Thus, given that 59 *A. thaliana* genes show patterns of variability in concordance with patterns obtained in large-scale studies, then it might be that 92 *P. patens* genes homologous to these *A. thaliana* genes are representative of codon usage patterns in large parts of the *P. patens* genome.

The lack of translational selection in mammals and *Drosophila* has been explained by low effective population sizes in studied species (Shields *et al.*, 1988; Sharp *et al.*, 1995; Akashi, 1997; Jenkins and Holmes, 2003). In species containing substantial genetic drift, stochastic allele frequency changes will determine the evolutionary fate of mutations that differ only marginally in fitness. For instance, in a comparison of two *Drosophila* species, Akashi (1997) found that *D. simulans* had both larger effective population size and more codon bias in studied genes than the closely related *D. melanogaster*. On the other hand, in the annual, selfing flowering plant *A. thaliana*, where drift is highly efficient in shaping genetic structuring (Abbott and Gomes, 1989; Innan *et al.*, 1997; Clauss *et al.*, 2002), selective constraints on codon usage have been shown to be associated with levels of gene expression (Duret and Mouchiroud, 1999). *P. patens* is an annual, haploid opportunist living in open, disturbed habitats (Nyholm, 1956). It lacks an efficient spore

spreading system (Reski, 1999), implying low migration capabilities, it is bisexual and self-fertile, and together these life history characteristics may indicate a relatively strong influence of founder events and genetic drift in shaping genetic structure of *P. patens* populations. Still, weak selection for codon usage seem to operate at the genomic level, implying either that the overall effective population size is larger than expected in this species or that genetic drift do not significantly influence levels of weak molecular selection on codon usage, as it seems to be the case in *A. thaliana*. If this latter is the case, then this seeming paradox is analogous to the problem of explaining the high within population genetic variability in molecular markers (eg Wyatt *et al*, 1989; Stenøien and Sæstad, 1999; Stenøien and Flatberg, 2000; Wilson and Provan, 2003), as well as quantitative traits (Stenøien *et al*, 1997) found in many primitive plants that are expected to experience substantial genetic drift.

## Acknowledgements

I thank Kent Holsinger and two anonymous reviewers for valuable comments on the manuscript. This study was partly financed by the Swedish Research Council (Grant no. 621-2002-5896).

## References

- Abbott RJ, Gomes MF (1989). Population genetic structure and outcrossing rate of *Arabidopsis thaliana* (L.) Heynh. *Heredity* **62**: 411–418.
- Akashi H (1994). Synonymous codon usage in *Drosophila melanogaster*: natural selection and translational accuracy. *Genetics* **136**: 927–935.
- Akashi H (1997). Codon bias evolution in *Drosophila*. Population genetics of mutation-selection drift. *Gene* **205**: 269–278.
- Akashi H (2001). Gene expression and molecular evolution. *Curr Opin Genet Dev* **11**: 660–666.
- Akashi H, Eyre-Walker AC (1998). Translational selection and molecular evolution. *Curr Opin Genet Dev* **8**: 688–693.
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990). Basic local alignment search tool. *J Mol Biol* **215**: 403–410.
- Andersson SGE, Kurland GC (1990). Codon preferences in free-living microorganisms. *Microbiol Rev* **54**: 198–210.
- Beletskii A, Bhagwat AS (1996). Transcription-induced mutations: increase in C to T mutations in the nontranscribed strand during transcription in *Escherichia coli*. *Proc Natl Acad Sci USA* **93**: 13919–13924.
- Bernardi G, Olofsson B, Filipinski J, Zerial M, Salinas J, Cuny G *et al* (1985). The mosaic genome of warm-blooded vertebrates. *Science* **228**: 953–958.
- Bezani M, Pan A, Quatrano RS (2003). RNA interference in the moss *Physcomitrella patens*. *Plant Physiol* **133**: 470–474.
- Boguski MS, Lowe TM, Tolstoshev CM (1993). dbEST – database for ‘expressed sequence tags’. *Nat Genet* **4**: 332–333.
- Bulmer M (1991). The selection-mutation-drift theory of synonymous codon usage. *Genetics* **129**: 897–907.
- Chiapello H, Fisacek F, Caboche M, Henaut A (1998). Codon usage and gene function are related in sequences of *Arabidopsis thaliana*. *Gene* **209**: GC1–GC38.
- Clauss MJ, Cobban H, Mitchell-Olds T (2002). Cross-species microsatellite markers for elucidating population genetic structure in *Arabidopsis* and *Arabis* (Brassicaceae). *Mol Ecol* **11**: 591–601.
- Coghlan A, Wolfe KH (2000). Relationship of codon bias to mRNA concentration and protein length in *Saccharomyces cerevisiae*. *Yeast* **16**: 1131–1145.
- Comeron JM (1999). K-Estimator: calculation of the number of nucleotide substitutions per site and the confidence interval. *Bioinformatics* **15**: 763–764.
- Comeron JM, Kreitman M (2002). Population, evolutionary and genomic consequences of interference selection. *Genetics* **161**: 389–410.
- Comeron JM, Kreitman M, Aguad M (1999). Natural selection on synonymous sites is correlated with gene length and recombination in *Drosophila*. *Genetics* **151**: 239–249.
- Duret L (2000). tRNA gene number and codon usage in the *C. elegans* genome are co-adapted for optimal translation of highly expressed genes. *Trends Genet* **16**: 287–289.
- Duret L, Mouchiroud D (1999). Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis*, *Drosophila*, and *Arabidopsis*. *Proc Natl Acad Sci USA* **96**: 4482–4487.
- Fennoy SL, Bailey-Serres J (1993). Synonymous codon usage in *Zea mays* L. nuclear genes is varied by levels of C and G-ending codons. *Nucleic Acids Res* **21**: 5294–5300.
- Francino HP, Ochman H (1999). Isochores result from mutation not selection. *Nature* **400**: 30–31.
- Gouy M, Gautier C (1982). Codon usage in bacteria: correlation with gene expressivity. *Nucleic Acids Res* **10**: 7055–7074.
- Hey J, Kliman RM (2002). Interactions between natural selection, recombination and gene density in the genes of *Drosophila*. *Genetics* **160**: 595–608.
- Ikemura T (1981). Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes. *J Mol Biol* **146**: 1–21.
- Ikemura T (1985). Codon usage and tRNA content in unicellular and multicellular organisms. *Mol Biol Evol* **2**: 13–34.
- Innan H, Terauchi R, Miyashita NT (1997). Microsatellite polymorphism in natural populations of the wild plant *Arabidopsis thaliana*. *Genetics* **146**: 1441–1452.
- Jenkins GM, Holmes EC (2003). The extent of codon usage bias in human RNA viruses and its evolutionary origin. *Virus Res* **92**: 1–7.
- Kanaya S, Yamada Y, Kudo Y, Ikemura T (1999). 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* **238**: 143–155.
- Kanaya S, Yamada Y, Kinouchi M, Kudo Y, Ikemura T (2001). Codon usage and tRNA genes in eukaryotes: correlation of codon usage diversity with translation efficiency and with CG-dinucleotide usage as assessed by multivariate analysis. *J Mol Evol* **53**: 290–298.
- Kimura M (1980). A simple method for estimating evolutionary rate of base substitution through comparative studies of nucleotide sequences. *J Mol Evol* **16**: 111–120.
- Kliman RM, Hey J (1993). Reduced natural selection associated with low recombination in *Drosophila melanogaster*. *Mol Biol Evol* **10**: 1239–1258.
- Li WH (1987). Models of nearly neutral mutations with particular implications for nonrandom usage of synonymous codons. *J Mol Evol* **24**: 337–345.
- Marais G, Duret L (2001). Synonymous codon usage, accuracy of translation, and gene length in *Caenorhabditis elegans*. *J Mol Evol* **52**: 275–280.
- Margulies EH, Kardia SL, Innis JW (2001). Identification and prevention of a GC content bias in SAGE libraries. *Nucleic Acids Res* **29**: E60.
- Markmann-Mulisch U, Hadi MZ, Koepchen K, Alonso JC, Russo VE, Schell J *et al* (2002). The organization of *Physcomitrella patens* RAD51 genes is unique among eukaryotic organisms. *Proc Natl Acad Sci USA* **99**: 2959–2964.
- Miyasaka H (2002). Translation initiation AUG context varies with codon usage bias and gene length in *Drosophila melanogaster*. *J Mol Evol* **55**: 52–64.
- Moriyama EN, Powell JR (1997). Codon usage bias and tRNA abundance in *Drosophila*. *J Mol Evol* **45**: 514–523.

- Morton BR (1993). Chloroplast DNA codon use: evidence for selection at the psb A locus based on tRNA availability. *J Mol Evol* **37**: 273–280.
- Morton BR (1994). Codon use and the rate of divergence of land plant chloroplast genes. *Mol Biol Evol* **11**: 231–238.
- Morton BR (1998). Selection on the codon bias of chloroplast and cyanelle genes in different plant and algal lineages. *J Mol Evol* **46**: 449–459.
- Musto H, Cruveiller S, D'Onofrio G, Romero H, Bernardi G (2001). Translational selection on codon usage in *Xenopus laevis*. *Mol Biol Evol* **18**: 1703–1707.
- Myers EW, Miller W (1989). Optimal alignments in linear space. *CABIOS* **4**: 11–17.
- Nishiyama T, Fujita T, Shin-I T, Seki M, Nishide H, Uchiyama I et al (2003). Comparative genomics of *Physcomitrella patens* gametophytic transcriptome and *Arabidopsis thaliana*: implication for land plant evolution. *Proc Natl Acad Sci USA* **100**: 8007–8012.
- Norusis MJ (2000). *SPSS 10.0 Guide to data analysis*. Prentice-Hall: Englewood Cliff, NJ.
- Nyholm E (1956). *Illustrated Moss Flora of Fennoscandia, II Musci, Fasc. 2*. Swedish Natural Science Research Council: Kungälv.
- Powell JR, Moriyama EN (1997). Evolution of codon usage bias in *Drosophila*. *Proc Natl Acad Sci USA* **94**: 7784–7790.
- Pryer KM, Schneider H, Zimmer EA, Ann Banks J (2002). Deciding among green plants for whole genome studies. *Trends Plant Sci* **7**: 550–554.
- Rensing SA, Rombauts S, Van de Peer Y, Reski R (2002). Moss transcriptome and beyond. *Trends Plant Sci* **7**: 535–538.
- Reski R (1999). Molecular genetics of *Physcomitrella*. *Planta* **208**: 301–309.
- Rozas J, Rozas R 1999. DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. *Bioinformatics* **15**: 174–175.
- Schaefer DG (2001). Gene targeting in *Physcomitrella patens*. *Curr Opin Plant Biol* **4**: 143–150.
- Schaefer DG (2002). A new moss genetics: targeted mutagenesis in *Physcomitrella patens*. *Annu Rev Plant Biol* **53**: 477–501.
- Schaefer DG, Zryd JP (1997). Efficient gene targeting in the moss *Physcomitrella patens*. *Plant J* **11**: 1195–1206.
- Sharp PM, Li WH (1987). The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. *Mol Biol Evol* **4**: 223–230.
- Sharp PM, Li WH (1989). On the rate of DNA-sequence evolution in *Drosophila*. *J Mol Evol* **28**: 398–402.
- Sharp PM, Matassi G (1994). Codon usage and genome evolution. *Curr Opin Genet Dev* **4**: 851–860.
- Sharp PM, Stenico M, Peden JF, Lloyd AT (1993). Codon usage: mutational bias, translational selection, or both? *Biochem Soc Trans* **21**: 835–841.
- Sharp PM, Averof M, Lloyd AT, Matassi G, Peden JF (1995). DNA sequence evolution: the sounds of silence. *Philos Trans R Soc Lond Ser B* **349**: 241–247.
- Shields DC, Sharp PM, Higgins DG, Wright F (1988). Silent sites in *Drosophila* genes are not neutral: evidence of selection among synonymous codons. *Mol Biol Evol* **5**: 704–716.
- Stenico M, Lloyd AT, Sharp PM (1994). Codon usage in *Caenorhabditis elegans*: delineation of translational selection and mutational bias. *Nucleic Acids Res* **22**: 2437–2446.
- Stenøien H, Bakken S, Flatberg KI (1997). Phenotypic variation in the *Sphagnum recurvum* complex: a cultivation experiment. *J Bryol* **19**: 731–750.
- Stenøien HK, Sæstad SM (1999). Genetic structure in three haploid peat mosses (*Sphagnum*). *Heredity* **82**: 391–400.
- Stenøien HK, Flatberg KI (2000). Genetic variability in the rare Norwegian peat moss *Sphagnum troendelagicum*. *Bryologist* **103**: 794–801.
- Strepp R, Scholz S, Kruse S, Speth V, Reski R (1998). Plant nuclear gene knockout reveals a role in plastid division for the homolog of the bacterial cell division protein FtsZ, an ancestral tubulin. *Proc Natl Acad Sci USA* **95**: 4368–4373.
- Wilson PJ, Provan J (2003). Effect of habitat fragmentation on levels and patterns of genetic diversity in natural populations of the peat moss *Polytrichum commune*. *Proc R Soc Lond Ser B* **270**: 881–886.
- Wolfe K, Sharp PM, Li WH (1989). Mutation rates differ among regions of the mammalian genome. *Nature* **337**: 283–285.
- Wright F (1990). The 'effective number of codons' used in a gene. *Gene* **87**: 23–29.
- Wyatt R, Odrzykoski IJ, Stoneburner A (1989). High levels of genetic variability in the haploid moss *Plagiomnium ciliare*. *Evolution* **43**: 1085–1096.
- Xia X, Xie Z (2001). DAMBE: data analysis in molecular biology and evolution. *J Hered* **92**: 371–373.
- Zhang Z, Schwartz S, Wagner L, Miller W (2000). A greedy algorithm for aligning DNA sequences. *J Comput Biol* **7**: 203–214.