



Skew in CG content near the transcription start site in *Arabidopsis thaliana*

Tatiana Tatarinova, Vyacheslav Brover, Maxim Troukhan and Nickolai Alexandrov*

Ceres, Inc, 3007 Malibu Canyon Road, Malibu, CA, 90265, USA

Received on January 6, 2003; accepted on February 20, 2003

ABSTRACT

We have discovered a novel statistical feature of *Arabidopsis thaliana* genome that remarkably correlates with a position of transcription start site—CG skew peak. We hypothesize that the phenomenon can be explained by the higher mutability of unprotected cytosines.

Keywords: CG skew, promoter, transcription, mutation, full-length cDNA

Contact: nicka@ceres-inc.com

We have observed an intriguing statistical feature of the *Arabidopsis* genome—a peak in the CG skew at the transcription start site (TSS). This analysis was enabled by the combination of two independent large-scale genomic projects: The *Arabidopsis* Genome Initiative (2000) and the Ceres plant full-length cDNA sequencing project. We have extracted 8000 promoter and 5' UTR sequences by aligning the 5' ends of full-length cDNAs with the *A.thaliana* genomic DNA using BLASTN (Davuluri *et al.*, 2000). For every position we calculated the CG skew (Fig. 1) and demonstrated that it has a maximum at the TSS (position 0 on the plot).

Our further studies showed that promoters from other organisms retrieved from the Eukaryotic Promoter Database (Grigoriev, 1999) also show a peak at the TSS (data not shown).

Prior publications (Myllykallio *et al.*, 2000; Freeman *et al.*, 1998; Grigoriev, 1998; Périer *et al.*, 2000; Grigoriev, 1999; Picardeau *et al.*, 2000; Karlin and Mrázek, 1998) associated a peak in the cumulative CG skew curve with the location of the replication origin. Codon usage differences were shown to be correlated with genes expression levels (Gentles and Karlin, 2001; Kochetov *et al.*, 1999; Davuluri *et al.*, 2000; Karlin and Mrázek, 2000) and with chromatin structure (Ioshikhes *et al.*, 1999). A number of researchers (Beletskii and Bhagwat, 1996; Freeman *et al.*, 1998; Grigoriev, 1998; Kochetov *et al.*, 1999) linked the DNA strand asymmetry to transcription-coupled effects. These ideas can be extended to explain the CG skew peak

at the TSS. Cytosine deamination is likely to be enhanced by DNA transcription (Beletskii *et al.*, 2000). The length of time a mutable cytosine base spends in an unpaired state determines the probability of its mutation to thymine (Beletskii and Bhagwat, 1996; Grigoriev, 1998). Upon binding of the transcription complex, the DNA unwinds at the TSS producing a transcription bubble of about 15–20 nucleotides in length. The bubble makes both strands prone to C–T mutations, but the RNA polymerase preferentially protects nucleotides on the non-transcribed strand (Schmitz and Galas, 1979). Hence C–T mutations occur more frequently on the transcribed strand, and consequently the non-transcribed strand becomes guanine-poor. This explains the increase in CG skew at the TSS. However, during transcription the transcribed strand becomes protected by the nascent RNA (Beletskii *et al.*, 2000; Beletskii and Bhagwat, 1996). Thus cytosine deamination is suppressed in the transcribed strand, but is still allowed in the non-transcribed strand. Therefore we should (and we do!) observe a decrease in CG skew in the non-transcribed strand.

REFERENCES

- The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, **408**, 796–815.
- Beletskii,A., Grigoriev,A., Joyce,S. and Bhagwat,A.S. (2000) Mutations induced by bacteriophage T7 RNA polymerase and their effects on the composition of the T7 genome. *J. Mol. Biol.*, **300**, 1057–1065.
- Beletskii,A. and Bhagwat,A. (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. Genetics
- Myllykallio,H., Lopez,P., López-García,P., Heilig,R., Saurin,W., Zivanovic,Y., Philippe,H. and Forterre,P. (2000) Bacterial mode of replication with eukaryotic-like machinery in a hyperthermophilic archaeon. *Science*, **288**, 2212–2215.
- Freeman,J.M., Plasterer,T.N., Smith,T.F. and Mohr,S.C. (1998) Patterns of genome organization in bacteria. *Science*, **279**, 1827.
- Grigoriev,A. (1998) Analyzing genomes with cumulative skew diagrams. *Nucleic Acid Res.*, **26**, 2286.

*To whom correspondence should be addressed.

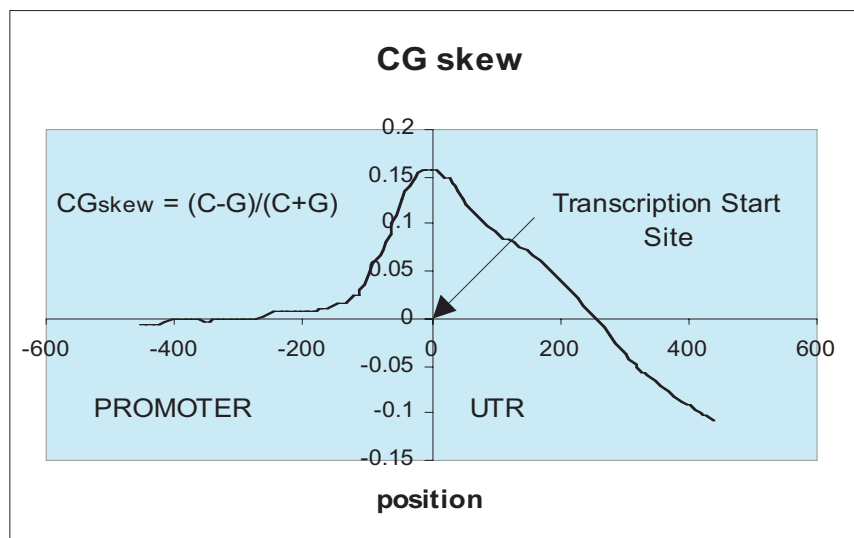


Fig. 1. Analysis of 8000 *A.thaliana* sequences near the transcription start site. CG skew $\equiv (C-G)/(C+G)$ values were computed with a sliding window of 100 bases in the non-transcribed strand, where C is the total number of cytosines and G is the total number of guanines for all sequences in the window.

Grigoriev,A. (1998) Genome arithmetic. *Science*, **281**, 1923–1998.

Francino,M.P. and Oshman,H. (1997) *Trends Genet.*, **13**, 240–245.

Schmitz,A. and Galas,D.J. (1979) The interaction of RNA polymerase and lac repressor with the lac control region. *Nucleic Acid Res.*, **6**, 111–137.

P erier,R.C., Praz,V., Junier,T., Bonnard,C. and Bucher,P. (2000) The Eukaryotic Promoter Database (EPD). *Nucleic Acids Res.*, **28**, 302–303.

Gentles,A.J. and Karlin,S. (2001) Genome-scale compositional comparisons in eukaryotes. *Genome Res.*, **11**, 540–546.

Grigoriev,A. (1999) Strand-specific compositional asymmetries in double-stranded DNA viruses. *Virus Res.*, **60**, 1–19.

Kochetov,A.V., Ponomarenko,M.P., Frolov,A.S., Kisselev,L.L. and Kolchanov,N.A. (1999) Prediction of eukaryotic mRNA translational properties. *Bioinformatics*, **15**, 704–712.

Altschul,S.F., Madden,T.L., Sch affer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.

Nucleic Acids Res., **25**, 3389–3402.

Picardeau,M., Lobry,J.R. and Hinnebusch,B.J. (2000) Analyzing DNA strand compositional asymmetry to identify candidate replication origins of *Borrelia burgdorferi* linear and circular plasmids. *Genome Res.*, **10**, 1594–1604.

Davuluri,R.V., Suzuki,Y., Sugano,S. and Zhang,M.Q. (2000) CART classification of human 5' UTR sequences. *Genome Res.*, **10**, 1807–1816.

Karlin,S. and Mr azek,J. (2000) Predicted Highly Expressed Genes of Diverse Prokaryotic Genomes. *J. Bacteriol.*, **182**, 5238–5250.

Karlin,S. and Mr azek,J. (1998) Strand Compositional Asymmetry in bacterial and large viral genomes. *Proc. Natl Acad. Sci. USA*, **95**, 3720–3725. Evolution.

Ioshikhes,I., Trifonov,E.N. and Zhang,M.Q. (1999) Periodical distribution of transcription factor sites in promoter regions and connection with chromatin structure. *Proc. Natl Acad. Sci. USA*, **96**, 2891–2895.