JACS
JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

Communication

pubs.acs.org/JACS

DNA–Water Interactions Distinguish Messenger RNA Genes from Transfer RNA Genes

Garima Khandelwal[†,‡] and B. Jayaram*[,†,‡,§]

[†]Department of Chemistry, [‡]Supercomputing Facility for Bioinformatics and Computational Biology, and [§]Kusuma School of Biological Sciences, Indian Institute of Technology Delhi, Hauz Khas, New Delhi-110016, India

Supporting Information

ABSTRACT: Physicochemical properties of DNA sequences as a guide to developing insights into genome organization has received little attention. Here, we utilize the energetics of DNA to further advance the knowledge on its language at a molecular level. Specifically, we ask the question whether physicochemical properties of different functional units on genomes differ. We extract intramolecular and solvation energies of different DNA base pair steps from a comprehensive set of molecular dynamics simulations. We then investigate the solvation behavior of DNA sequences coding for mRNAs and tRNAs. Distinguishing mRNA genes from tRNA genes is a tricky problem in genome annotation without assumptions on length of DNA and secondary structure of the product of transcription. We find that solvation energetics of DNA behaves as an extremely efficient property in discriminating 2 063 537 genes coding for mRNAs from 56 251 genes coding for tRNAs in all (∼1500) completely sequenced prokaryotic genomes.

DNA sequences work through their structure, dynamics, and thermodynamics in carrying out their molecular functions. One of them is the conversion of genome to transcriptome, that is, DNA to RNA, which involves unpackaging of DNA, recognition by regulatory regions, unwinding of DNA, and strand separation, all of which would require a consideration of inter-base pair (stacking) and intra-base pair (hydrogen bonding) interactions, DNA–solvent interactions, structural flexibility of DNA sequences, and so forth. Since physical measurements on individual elements are typically not feasible, methods to predict properties for DNA sequences of any length and composition are required, which may include experimentally determined properties for prototype sequences[1,2] or indices obtained heuristically or calculated from theoretical models involving molecular simulations.[3−8] Our previous explorations in this regard yielded compelling clues to the differential behavior of energetics of functionally different DNA sequences.[9−12] The problem of identifying tRNA genes has been difficult particularly without secondary structural inputs. Most of the computational methods for gene prediction are designed to locate protein coding regions, that is, mRNA genes, and depend invariably on some form of database training.[13−17] Physicochemical properties inherent to DNA sequences as a tool to develop insights into genome organization have not been harnessed. We report here a systematic investigation of the energetics of DNA sequences coding for over 2 million mRNA sequences and over 50 000 tRNA sequences. We note that solvation energies show an elegant separation of DNA sequences coding for mRNAs from those for tRNAs.

Structure and dynamics of DNA oligonucleotides of all possible tetra-nucleotide combinations have been characterized recently by the Ascona B-DNA Consortium[3−5] in a series of comprehensive state of the art molecular dynamics simulations. These simulations have enabled us to extract solvation energies of all 16 dinucleotide steps (Table 1) to probe the molecular thermodynamic signatures of various functional units on genomes within an additive framework. The proximity criterion is used to assign water molecules to each distinct trinucleotide in the double helix trajectory and the interaction energies are computed.[18] These trinucleotide values are further mapped onto dinucleotide solvation values (details of the mapping are provided in the Supporting Information).

Table 1. Molecular Dynamics Derived Solvation Energies of Dinucleotides

dinucleotides	solvation energy (kcal/mol)
AA	−171.84
AC	−171.11
AG	−174.93
AT	−173.70
CA	−179.01
CC	−166.76
CG	−176.88
CT	−174.93
GA	−167.60
GC	−165.58
GG	−166.76
GT	−171.11
TA	−174.35
TC	−167.60
TG	−179.01
TT	−171.84

The genomic data for the present study has been compiled from the National Centre of Biotechnology Information (NCBI, URL: http://www.ncbi.nlm.nih.gov/) repository. The genome sequences were downloaded from the NCBI ftp site

Received: March 2, 2012
Published: May 1, 2012

   dx.doi.org/10.1021/ja3020956 | J. Am. Chem. Soc. 2012, 134, 8814–8816

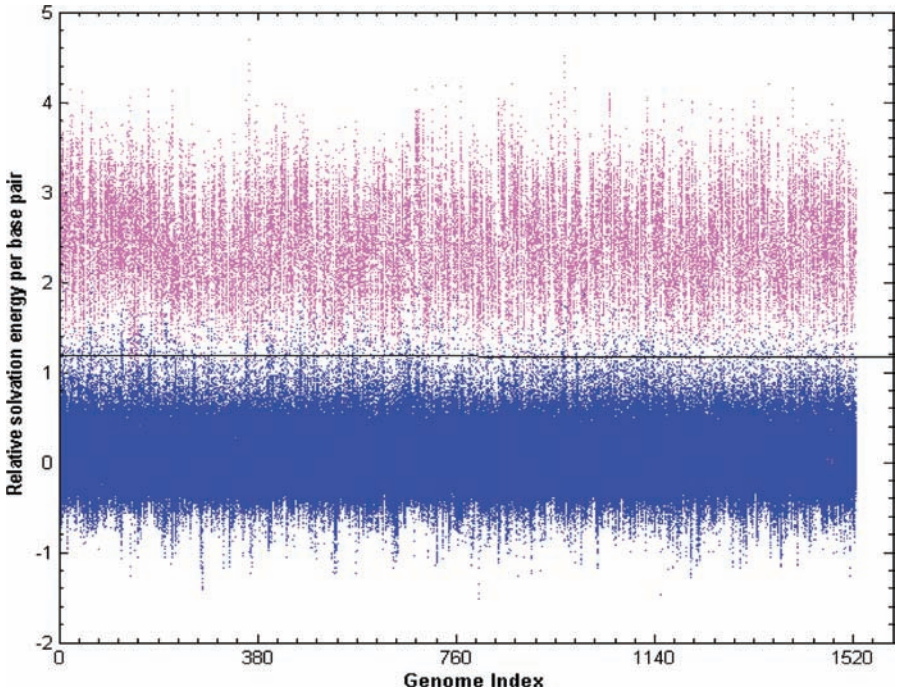

Figure 1. Relative solvation energy per base pair of DNA sequences coding for 2 063 537 mRNAs (blue) and 56 251 tRNAs (pink) from 1531 genomic sequences. The *X*-axis denotes the index of the genome while the *Y*-axis depicts the solvation energy of the sequence relative to the average for that genome. A threshold value of 1.2 (solid black line) separates 99.75% tRNA genes from 99.83% mRNA genes.

(ftp://ftp.ncbi.nih.gov/genomes/Bacteria/). The gene coordinates for protein coding and tRNA genes were extracted from the GenBank file (.gbk). The protein coding genes which formed hypothetical, probable, or putative products or peptides were removed and the remaining gene coordinates were used to extract the gene sequences from the fasta file (.fna). These gene sequences were then used to calculate the solvation energies from the dinucleotide solvation energies provided in Table 1, by moving one base at a time, thus, considering all $(N - 1)$ dinucleotide steps in a sequence of length '*N*'.

To illustrate further, for a sequence of 15 base pairs such as:

$$5'\ \mathrm{ATGCACGAGCTAACC}\ 3'$$
$$3'\ \mathrm{TACGTGCTCGATTGG}\ 5'$$

there are 14 dinucleotide steps: AT + TG + GC + CA + AC + CG + GA + AG + GC + CT + TA + AA + AC + CC. The total solvation energy for this sequence is computed by adding the 14 individual dinucleotide solvation energies with reference to data given in Table 1, which equals to −2412.39 kcal/mol. The solvation energy per base pair is then computed by dividing the total solvation energy by the length of the sequence as given below:

$$\text{Solvation energy per base pair} = \frac{\text{Total solvation energy}}{\text{Length of the sequence}}$$

which for the above 15 bp sequence is −160.83 kcal/mol.

The average solvation energy per base pair for each genome is then calculated by summing up the solvation energies of all the mRNA genes of a genome and then dividing by the total number of genes as shown below:

$$\text{Average solvation energy per basepair} = \frac{\sum \text{Solvation energy per base pair of each mRNA gene}}{\text{Total number of mRNA genes}}$$

This average solvation energy is then used to estimate the relative solvation energies of all genes (mRNA as well as tRNA) of a genome by subtracting the average solvation energy from the absolute solvation energy per base pair of each gene calculated as above. Stated alternatively, the relative energies are obtained by taking the average value of mRNA genes of each genome as the reference for that particular genome. This is repeated for all the mRNA and tRNA genes in the genomic and plasmid sequences available at NCBI.

Relative solvation energies per base pair of all available mRNA genes and tRNA genes from the sequence data of 1046 prokaryotic organisms and 485 plasmid sequences are depicted in Figure 1. Each point in the figure refers to a single complete gene sequence extracted from the NCBI data. The figure shows a clean separation of mRNA genes from tRNA genes over all the genomes considered. The solvation energy distribution is tight and more importantly nonoverlapping with very few exceptions. More specifically, a threshold value of 1.2 separates 56 111 (99.75%) tRNA gene sequences (lying above this threshold), from 2 060 064 (99.83%) mRNA gene sequences (lying below this threshold).

That tRNA genes are less-well solvated as compared to those of mRNAs is also indicated by Figure 1. Noting that tRNAs form highly stabilized secondary and tertiary structures with multiple intramolecular interactions and less solvation, while mRNAs do not, one might surmise that the imprints of the destiny of the nucleotide sequences are built evolutionarily into DNA sequences. Simply stated, DNA displays very clear physicochemical fingerprints for its sequences with different functions. That DNA is stabilized by aqueous environment and that water is implicated in DNA−protein interactions have been known for some time.[19−24] The present finding in this backdrop is both surprising and revealing and provides a clue to distinguishing DNA regions coding for mRNAs from those

dx.doi.org/10.1021/ja3020956 | *J. Am. Chem. Soc.* 2012, 134, 8814−8816

coding for tRNAs without the need for database driven algorithms and resources.

■ ASSOCIATED CONTENT

Supporting Information

Mapping procedure of dinucleotides along with molecular dynamics derived trinucleotide solvation energies. Separate figures of relative solvation energy per base pair for mRNA genes and tRNA genes and both of them together. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

bjayaram@chemistry.iitd.ac.in

Author Contributions

Both authors have contributed equally.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors thank the Ascona B-DNA Consortium for making the molecular dynamics simulation data available, Prof. D. L. Beveridge and Dr. Surjit Dixit for helpful comments and NCBI for free access to genomic information. Program support to the Supercomputing Facility for Bioinformatics & Computationally Biology, IIT Delhi from the Department of Biotechnology, Government of India is gratefully acknowledged. G.K. is a recipient of DBT-JRF.

■ REFERENCES

(1) SantaLucia, J., Jr.; Hicks, D. *Annu. Rev. Biophys. Biomol. Struct.* **2004**, *33*, 415.

(2) Rangannan, V.; Bansal, M. *Bioinformatics* **2010**, *26*, 3043.

(3) Beveridge, D. L.; Barreiro, G.; Byun, K. S.; Case, D. A.; Cheatham, T. E., III; Dixit, S. B.; Giudice, E.; Lankas, F.; Lavery, R.; Maddocks, J. H.; Osman, R.; Seibert, E.; Sklenar, H.; Stoll, G.; Thayer, K. M.; Varnai, P.; Young, M. A. *Biophys. J.* **2004**, *87*, 3799.

(4) Dixit, S. B.; Beveridge, D. L.; Case, D. A.; Cheatham, T. E., III; Giudice, E.; Lankas, F.; Lavery, R.; Maddocks, J. H.; Osman, R.; Sklenar, H.; Thayer, K. M.; Varnai, P. *Biophys. J.* **2005**, *89*, 3721.

(5) Lavery, R.; Zakrzewska, K.; Beveridge, D. L.; Bishop, T. C.; Case, D. A.; Cheatham, T., III; Dixit, S.; Jayaram, B.; Lankas, F.; Laughton, C.; Maddocks, J. H.; Michon, A.; Osman, R.; Orozco, M.; Perez, A.; Singh, T.; Spackova, N.; Sponer, J. *Nucleic Acids Res.* **2009**, *38*, 299.

(6) Stolz, R.; Bishop, T. C. *Nucleic Acids Res.* **2010**, *38*, W254.

(7) Lankas, F.; Gonzalez, O.; Heffler, L. M.; Stoll, G.; Moakher, M.; Maddocks, J. H. *Phys. Chem. Chem. Phys.* **2009**, *11*, 10565.

(8) Lafontaine, I.; Lavery, R. *Biophys. J.* **2000**, *79*, 680.

(9) Dutta, S.; Singhal, P.; Agrawal, P.; Tomer, R.; Kritee; Khurana, E.; Jayaram, B. *J. Chem. Inf. Model.* **2006**, *46*, 78.

(10) Singhal, P.; Jayaram, B.; Dixit, S. B.; Beveridge, D. L. *Biophys, J.* **2008**, *94*, 4173.

(11) Jayaram, B. *J. Mol. Evol.* **1997**, *45*, 704.

(12) Khandelwal, G.; Jayaram, B. *PLoS One* **2010**, *5*, e12433.

(13) Rivas, E.; Eddy, S. R. *Bioinformatics* **2000**, *16*, 583.

(14) Rivas, E.; Eddy, S. R. *BMC Bioinf.* **2001**, *2*, 8.

(15) Tran, T. T.; Zhou, F.; Marshburn, S.; Stead, M.; Kushner, S. R.; Xu, Y. *Bioinformatics* **2009**, *25*, 2897.

(16) Laslett, D.; Canback, B. *Nucleic Acids Res.* **2004**, *32*, 11.

(17) Médigue, C.; Moszer, I. *Res. Microbiol.* **2007**, *158*, 724.

(18) Mezei, M.; Beveridge, D. L. In *Methods in Enzymology*; Packer, L. Ed.; Academic Press: Orlando, FL, **1986**; Vol. *127*, pp 22−47.

(19) Berman, H. M. *Curr. Opin. Struct. Biol.* **1994**, *4*, 345.

(20) Westhoff, E.; Beveridge, D. L. Hydration of Nucleic Acids. In *Water Science Reviews*; Franks, F., Ed.; Cambridge Univ. Press: Cambridge, U.K., 1990; Vol. *5*, pp 24−136.

(21) Elcock, A. H.; McCammon, J. A. *J. Am. Chem. Soc.* **1995**, *117*, 10161.

(22) Spyrakis, F.; Cozzini, P.; Bertoli, C.; Marabotti, A.; Kellogg, G. E.; Mozzarelli, A. *BMC Struct. Biol.* **2007**, *7*, 4.

(23) Jayaram, B.; Jain, T. *Annu. Rev. Biophys. Biomol. Struct.* **2004**, *33*, 343.

(24) Reddy, C. K.; Das, A.; Jayaram, B. *J. Mol. Biol.* **2001**, *314*, 619.