Published online 8 December 2016

Nucleic Acids Research, 2017, Vol. 45, No. 7 ***e47***
doi: 10.1093/nar/gkw1236

Physico-chemical fingerprinting of RNA genes

Ankita Singh[1], Akhilesh Mishra[1,2], Ali Khosravi[3], Garima Khandelwal[4] and B. Jayaram[1,2,5,*]

[1]Supercomputing Facility for Bioinformatics & Computational Biology, Indian Institute of Technology, Hauz Khas, New Delhi-110016, India, [2]Kusuma School of Biological Sciences, Indian Institute of Technology, Hauz Khas, New Delhi-110016, India, [3]Ale-Taha Institute of Higher Education, Tehran, Iran, [4]Cancer Research UK Manchester Institute, The University of Manchester, Wilmslow Road, Manchester M20 4BX, UK and [5]Department of Chemistry, Indian Institute of Technology, Hauz Khas, New Delhi-110016, India

Received November 09, 2016; Editorial Decision November 23, 2016; Accepted November 29, 2016

ABSTRACT

We advance here a novel concept for characterizing different classes of RNA genes on the basis of physico-chemical properties of DNA sequences. As knowledge-based approaches could yield unsatisfactory outcomes due to limitations of training on available experimental data sets, alternative approaches that utilize properties intrinsic to DNA are needed to supplement training based methods and to eventually provide molecular insights into genome organization. Based on a comprehensive series of molecular dynamics simulations of Ascona B-DNA consortium, we extracted hydrogen bonding, stacking and solvation energies of all combinations of DNA sequences at the dinucleotide level and calculated these properties for different types of RNA genes. Considering ∼7.3 million mRNA, 255 524 tRNA, 40 649 rRNA (different subunits) and 5250 miRNA, 3747 snRNA, gene sequences from 9282 complete genome chromosomes of all prokaryotes and eukaryotes available at NCBI, we observed that physico-chemical properties of different functional units on genomic DNA differ in their signatures.

INTRODUCTION

Genome annotation, the task of identifying protein coding mRNA genes, non-coding RNA (tRNA, miRNA, snRNA, rRNA etc.) genes, promoters/regulatory switches etc. has been receiving extensive attention since 1997 with the first report of sequencing of a complete genome (1). Over the years, various computational methods have displayed a potential for fast and accurate characterization of genes (2–10). Majority of these methods are knowledge-based and involve sophisticated statistical and mathematical techniques for training and prediction (11–16). Although, such innovations form the mainstay today, these are influenced by sparse experimental data available for training thus limiting their performance (17–20), making them genome dependent (21) and opaque to molecular interpretations. As a consequence, improvements to these approaches primarily depend on enlarging the training data sets (22).

An alternative approach to solve this complex challenge is based on the hypothesis that different functional units on genomic DNA differ in their physico-chemical properties, which, in principle, can be extracted from atomic models of DNA (13,23–29). Presently, there exists a wide hiatus between data-driven statistical tools for genome annotation and molecular simulation based atomic level descriptions of oligonucleotides (30–41). Though understanding the language of DNA at a molecular level is compelling, there will always be a need for both the approaches. The challenge for atomic models lies in extracting from the exponentially growing genomic data, the hidden physico-chemical signatures of function.

Of particular interest is the evolving knowledge of non-coding RNAs. Recent findings implicate a role for ncRNAs in gene regulation, dosage compensation, genomic imprinting, cell differentiation, organogenesis, development, metabolism, homeostasis and disease (38–47).

The present work encompasses elucidation of physico-chemical fingerprints for different functional units in prokaryotic and eukaryotic genomes on the basis of atomic level descriptions of oligonucleotides derived from molecular simulations.

MATERIALS AND METHODS

Data description

The genomic data for the present study is compiled from the National Centre of Biotechnology Information (NCBI) repository. All types of mRNA, tRNA, miRNA snRNA and rRNA (different subunits), sequences except plasmid, bacteriophage, BAC/YAC, cosmid, clone and viral sequences are considered in this study. The genome fasta file (.fna for prokaryotes and .fa for eukaryotes) and corresponding genbank (.gbk for prokaryotes and .gb for eukaryotes) files are downloaded from the NCBI's ftp site

*To whom correspondence should be addressed. Tel: +91 11 2659 1505; Fax: +91 11 2658 2037; Email: bjayaram@chemistry.iitd.ac.in

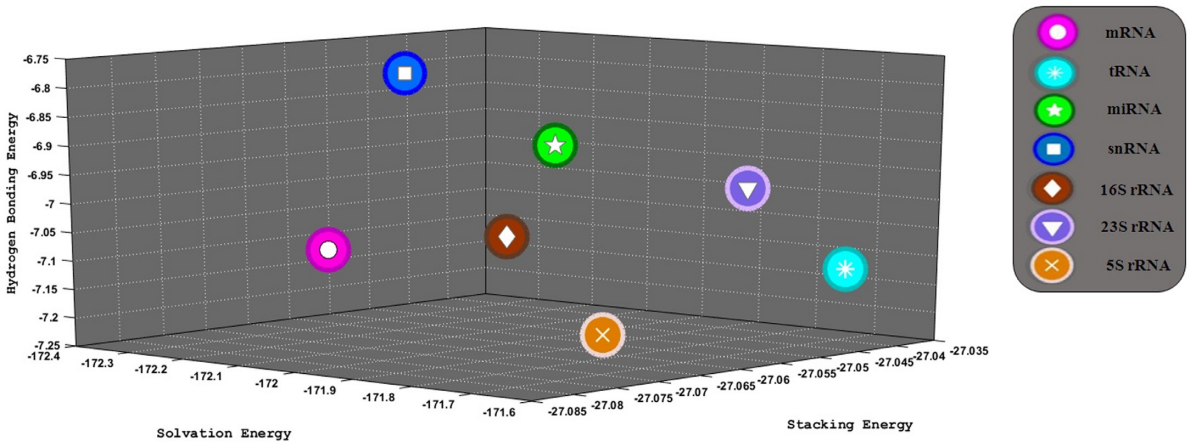

Figure 1. Differentiation of RNA genes obtained on the basis of physico-chemical properties. Average values of hydrogen bonding energy per bp (kcal/mol), stacking energy per bp (kcal/mol) and solvation energy per bp (kcal/mol), of ∼7.6 million RNA genes comprising ∼7.3 million mRNA (magenta, circle), 255 524 tRNA (cyan, star), 5250 miRNA (green, pentagon), 3747 snRNA (blue, square), 13 997 16S rRNA (brown, diamond), 13 745 23S rRNA (purple, triangle) and 12 907 5S rRNA (orange, cross) genes are shown for 9282 prokaryotic and eukaryotic genomes.

Table 1. A self-consistent set of molecular dynamics derived hydrogen bond, stacking and solvation energies (kcal/mol) for double helical dinucleotide steps

Dinucleotide	Hydrogen bond	Stacking energy	Solvation
AA	-5.44	-26.71	-171.84
AC	-7.14	-27.73	-171.11
AG	-6.27	-26.89	-174.93
AT	-5.35	-27.20	-173.70
CA	-7.01	-27.15	-179.01
CC	-8.48	-26.28	-166.76
CG	-8.05	-27.93	-176.88
CT	-6.27	-26.89	-174.93
GA	-7.80	-26.78	-167.60
GC	-8.72	-28.13	-165.58
GG	-8.48	-26.28	-166.76
GT	-7.14	-27.73	-171.11
TA	-5.83	-26.90	-174.35
TC	-7.80	-26.78	-167.60
TG	-7.01	-27.15	-179.01
TT	-5.44	-26.71	-171.84

(ftp://ftp.ncbi.nih.gov/genomes/). The data set comprised of 4143 completely annotated eukaryotic (protozoa, fungi, plant, invertebrates and vertebrates) genomes and 5139 completely annotated prokaryotic (archaea and bacteria) genomes. Coordinates which formed partial gene sequences and genes categorized as hypothetical, predicted, probable, putative products or peptides are avoided for maintaining reliability of the data. Gene sequences that contain any base other than A, T, G and C are also filtered out along with genes classified with unknown class type. This presented a data set of ∼7.3 million mRNA, ∼0.25 million tRNA and 5250 miRNA, 3747 snRNA sequences, 40 649 rRNA (5S, 5S-like,16S, 16S-like, 23S, 23S-like subunits) for the current study. Complete RNA data set and computed parameters are provided in public domain at http://www.scfbio-iitd.res.in/software/data_RNA.jsp.

Methodology

In pursuit of exploring the physico-chemical information hidden in DNA sequences, we have considered here, three properties viz. hydrogen bonding energy (per base pair (bp)), stacking energy (per bp) and solvation energy (per bp). Based on a comprehensive series of molecular dynamics simulations of the Ascona B-DNA consortium on all possible tetra-nucleotide combinations (30–34), we first extracted the hydrogen bonding, stacking and solvation energies of all combinations of DNA sequences at the trinucleotide levels by averaging their total occurrences in all the possible tetra-nucleotides. These trinucleotide energy values are further mapped into dinucleotide values. Details of the mapping procedure (13) are provided in the supplementary information of previous work (25). Dinucleotide frequencies values are given in supplementary S1 here. The hydrogen bonding, stacking and solvation energy values for each dinucleotide derived from molecular dynamics simulations (35–37) are provided in Table 1. The methodology for the calculation of physico-chemical properties of gene sequences is presented below.

Hydrogen bonding energy. Hydrogen bonding energy (kcal/mol) is calculated from the dinucleotide hydrogen bond energies (Table 1), by moving one base at a time, thus,

Table 2. Calculated averages (Avg) and standard deviations (SD) of hydrogen bonding (HB), stacking and solvation energies (in kcal/mol) for each RNA gene

Type	Avg. HB energy	SD for HB energy	Avg. stacking energy	SD for stacking energy	Avg. solvation energy	SD for solvation energy	Sample size (Number of genes)
mRNA	-7.04	0.354	-27.08	0.076	-171.98	0.423	7 295 415
tRNA	-7.12	0.323	-27.04	0.066	-171.68	0.468	255 524
miRNA	-6.95	0.348	-27.04	0.110	-172.22	0.776	5250
snRNA	-6.81	0.184	-27.05	0.051	-172.35	0.338	3747
16S rRNA	-7.04	0.224	-27.06	0.030	-171.94	0.231	13 997

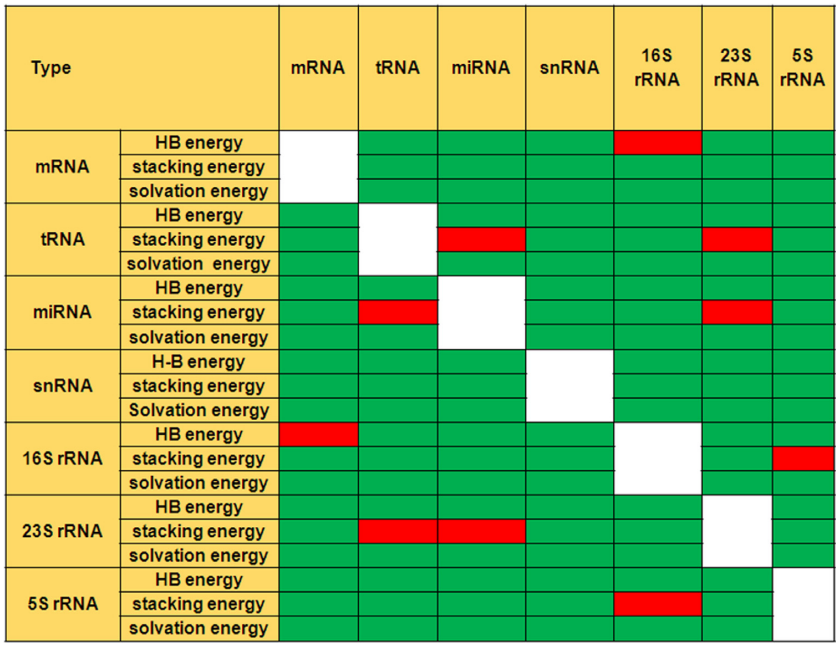

Figure 2. Results of two sample *t*-test for comparing means of each physico-chemical property considered (hydrogen bonding, stacking and solvation energies) for each pair of RNAs. Green represents that the separation in the mean values is statistically significant and red represents that it is statistically insignificant.

considering all the 'N – 1' dinucleotide steps in a sequence of length 'N'. As shown in Equation (1), the total hydrogen bonding energy is then divided by the number of dinucleotides (22).

$$\text{Hydrogen bonding energy (per bp)} = \frac{\text{Total hydrogen bond energy}}{\text{No. of dinucleotides}} \quad (1)$$

Similarly, stacking energy and solvation energies of all RNA genes are calculated from the dinucleotide stacking and solvation energies (Table 1) and the nucleotide sequences.

RESULTS AND DISCUSSION

Here, we investigate the possibility of the presence of physico-chemical signatures in genomic DNA that can convey the functional role of genic sequences. In this pursuit, we have calculated the average hydrogen bonding, stacking and solvation energies of mRNA, tRNA, miRNA, snRNA and rRNA genes. The average value of each property together with the standard deviation for each type of RNA gene is presented in Table 2.

The average values of the physico-chemical properties in Table 2 are plotted on a 3D graph (Figure 1) where the three coordinates represent hydrogen bonding energy, stacking energy and solvation energy for all the major classes of coding mRNAs and non-coding RNAs (tRNA, miRNA, snRNA and rRNA). Different orientations of Figure 1 are provided in Supplementary Figures S1 and S2. It is seen clearly from Figure 1 that genes of all major classes of RNA have different physico-chemical signatures.

To assess the statistical significance ($P < 0.05$) of the separation in average values of the physico-chemical properties of different RNA genes seen in Figure 1 and Table 2, we have performed a two-sided Student's *t*-test (Supplementary S1) between all the possible pairs for all the three properties. The results are shown in Figure 2. The green color represents that the difference in the mean values of the reference physico-chemical property is statistically significant, while red color represents that the separation of mean values is statistically insignificant. As seen clearly, the mean values are well separated for all the RNA genes in at least two properties if not all. More specifically, solvation energy is a significant parameter for separation of all RNA pairs, while stacking energy separations are significant for

all RNA pairs except miRNA from tRNA, 23s rRNA from tRNA, 23s rRNA from miRNA and 16s rRNA from 5s rRNA. Further, hydrogen bond energy is also a very significant parameter for separation of all RNA types except for 16s rRNA from mRNA. To further validate these results we have also performed Welch's t-test (Supplementary S2), which is an adaption of Student's *t*-test. The results are similar. Essentially the differences in averages noticeable in Figure 1 are statistically significant.

CONCLUSION

An analysis of over 9282 prokaryotic and eukaryotic genomes comprising ∼7.6 million genes clearly points to the existence of physico-chemical fingerprints of the functional destiny of DNA sequences (Figure 1). Earlier studies (48,49), have implicated base sequence dependent shape and electrostatic potential in the grooves of DNA, as well as DNA curvature and bendability (50) in molecular function. The present study provides us with additional information imprinted in the DNA sequences, suggestive of a plausible physico-chemical property based mechanism of read-out of DNA function.

AVAILABILITY

Complete RNA data set and computed parameters are provided in public domain at http://www.scfbio-iitd.res.in/software/data_RNA.jsp.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank the Ascona B-DNA Consortium for providing the molecular dynamics simulation data, Prof. D. L. Beveridge and Dr Surjit Dixit for helpful comments. The authors thank Ms. Varsha Singh and Ms. Kritika Karri for their participation during the initial stages of the project.

FUNDING

Department of Biotechnology, Govt. of India [to SCFBio]. Funding for open access charge: Department of Biotechnology, Govt. of India.
Conflict of interest statement. None declared.

REFERENCES

1. Fleischmann,R.D., Adams,M.D., White,O., Clayton,R.A., Kirkness,E.F., Kerlavage,A.R., Bult,C.J., Tomb,J.F., Dougherty,B.A., Merrick,J.M. *et al.* (1995) Whole-genome random sequencing and assembly of Haemophilus influenza. *Science*, **269**, 496–512.
2. Gelfand,M.S., Mironov,A.A. and Pevzner,P.A. (1996) Gene recognition via spliced sequence alignment. *Proc. Natl. Acad. Sci. U.S.A.*, **93**, 9061–9066.
3. Meyer,I.M. and Durbin,R. (2004) Gene structure conservation aids similarity based gene prediction. *Nucleic Acids Res.*, **32**, 776–783.
4. Mignone,F., Grillo,G., Liuni,S. and Pesole,G. (2003) Computational identication of protein coding potential of conserved sequence tags through cross-species evolutionary analysis. *Nucleic Acids Res.*, **31**, 4639–4645.
5. Korf,I., Flicek,P., Duan,D. and Brent,M.R. (2001) Integrating genomic homology into gene structure prediction. *Bioinformatics*, **17**, S140–S148.
6. Keller,O., Kollmar,M., Stanke,M. and Waack,S. (2011) A novel hybrid gene prediction method employing protein multiple sequence alignments. *Bioinformatics*, **27**, 757–763.
7. Guigó,R., Knudsen,S., Drake,N. and Smith,T. (1992) Prediction of gene structure. *J. Mol. Biol.*, **226**, 141–157.
8. Uberbacher,E.C., Hyatt,D. and Shah,M. (2004) GrailEXP and genome analysis pipeline for genome annotation. *Curr. Protoc. Hum. Genet.*, doi:10.1002/0471142905.hg0605s39.
9. Cantarel,B.L., Korf,I., Robb,S.M., Parra,G., Ross,E., Moore,B., Holt,C., Sánchez Alvarado,A. and Yandell,M. (2008) MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res.*, **18**, 188–196.
10. Yeh,R.F., Lim,L.P. and Burge,C.B. (2001) Computational inference of homologous gene structures in the human genome. *Genome Res.*, **11**, 803–816.
11. Mathé,C., Sagot,M.F., Schiex,T. and Rouzé,P. (2002) Current methods of gene prediction, their strengths and weaknesses. *Nucleic Acids Res.*, **30**, 4103–4117.
12. Bandyopadhyay,S., Maulik,U. and Roy,D. (2008) Gene identification: classical and computational intelligence approaches. *IEEE Trans. Syst. Man. Cybern. C Appl. Rev.*, **38**, 55–68.
13. Singhal,P., Jayaram,B., Dixit,S.B. and Beveridge,D.L. (2008) Prokaryotic gene finding based on physicochemical characteristics of codons calculated from molecular dynamics simulations. *Biophys. J.*, **94**, 4173–4183.
14. Zhu,W., Lomsadze,A. and Borodovsky,M. (2010) Ab initio gene identification in metagenomic sequences. *Nucleic Acids Res.*, **38**, e132.
15. Goel,N., Singh,S. and Aseri,T.C. (2013) A comparative analysis of soft computing techniques for gene prediction. *Anal. Biochem.*, **438**, 14–21.
16. Soh,J, Gordon,P.M.K. and Sensen,C.W. (2012) *Genome Annotation*, Chapman & Hall/CRC, Boca Raton.
17. Libbrecht,M.W. and Stafford,W. (2015) Noble machine learning applications in genetics and genomics. *Nat. Rev. Genet.*, **16**, 321–332.
18. Zickmann,F. and Renard,B.Y. (2015). IPred - integrating ab initio and evidence based gene predictions to improve prediction accuracy. *BMC Genomics*, **16**, 134–142.
19. Drăgan,M.A., Moghul,I., Priyam,A., Bustos,C. and Wurm,Y. (2016). GeneValidator: identify problems with protein-coding gene predictions. *Bioinformatics*, **32**, 1559–1561.
20. Stephs,Z.D., Lee,S.Y., Faghri,F., Campbell,R.H., Zhai,C., Efron,M.J. and Robinson,G.E. (2015). Big Data: Astronomical or Genomical? *PLoS Biol.*, **13**, e1002195.
21. Land,M., Hauser,L., Jun,S.R., Nookaew,I., Leuze,M.R., Ahn,T.H., Karpinets,T., Lund,O., Kora,G., Wassenaar,T. *et al.* (2015) Insights from 20 years of bacterial genome sequencing. *Funct. Integr. Genomics*, **15**, 141–161.
22. Muir,P., Li,S., Lou,S., Wang,D., Spakowicz,D.J., Salichos,L., Zhang,J., Weinstock,G.M., Isaacs,F., Rozowsky,J. *et al.* (2016) The real cost of sequencing: scaling computation to keep pace with data generation. *Genome Biol.*, **17**, 53–61.
23. Korf,I. (2004) Gene finding in novel Genomes. *BMC Bioinformatics*, **5**, 59–67.
24. Dutta,S., Singhal,P., Agrawal,P., Tomer,R., Kritee,K., Khurana,E. and Jayaram,B. (2006) A physico-chemical model for analyzing DNA sequences. *J. Chem. Inf. Model*, **46**, 78–85.
25. Khandelwal,G. and Jayaram,B. (2012) DNA-water interactions distinguish messenger RNA genes from transfer RNA genes. *J. Am. Chem. Soc.*, **134**, 8814–8816.
26. Khandelwal,G., Gupta,J. and Jayaram,B. (2012) DNA energetics based analyses suggest additional genes in prokaryotes. *J. Biosci.*, **37**, 433–444.
27. Khandelwal,G. and Jayaram,B. (2010) A Phenomenological model for predicting melting temperatures of DNA sequences. *PLoS One*, **5**, e12433.
28. Kanhere,A. and Bansal,M. (2005) A novel method for prokaryotic promoter prediction based on DNA stability. *BMC Bioinformatics*, **6**, 1–10.
29. Lafontaine,I. and Lavery,R. (2000) Optimization of nucleic acid sequences. *Biophys. J.*, **79**, 680–685.

30. McCaskill,J.S. (1990) The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, **29**, 1105–1119.
31. Lavery,R., Zakrzewska,K., Beveridge,D., Bishop,T.C., Case,D.A., Cheatham,T., Dixit,S., Jayaram,B., Lankas,F., Laughton,C. *et al.* (2009) A systematic molecular dynamics study of nearest-neighbor effects on base pair and base pair step conformations and fluctuations in B-DNA. *Nucleic Acid Res.*, **38**, 299–313.
32. Peyrard,M., Dauxois,T., Hoyet,H. and Willis,C.R. (1993) Biomolecular dynamics of DNA: statistical mechanics and dynamical model. *Physica D*, **68**, 104–115.
33. Passi,M., Maddocks,J.H., Beveridge,D., Bishop,T.C., Case,D.A., Cheatham,T., Dans,P.D., Jayaram,B., Lankas,F. *et al.* (2014) {mu}ABC: a systematic microsecond molecular dynamics study of tetranucleotide sequence effects in B-DNA. *Nucleic Acids Res.*, **42**, 12272–12283.
34. SantaLucia,J. Jr (1998) A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc. Natl. Acad. Sci. U.S.A.*, **95**, 1460–1465.
35. Beveridge,D.L., Barreiro,G., Byun,K.S., Case,D.A., Cheatham,T.E. 3rd, Dixit,S.B., Giudice,E., Lankas,F., Lavery,R., Maddocks,J.H. *et al.* (2004) Molecular dynamics simulations of the 136 unique tetranucleotide sequences of DNA oligonucleotides. I. Research design and results on d (CpG) steps. *Biophys. J.*, **87**, 3799–3813.
36. Dixit,S.B., Beveridge,D.L., Case,D.A., Cheatham,T.E. 3rd, Giudice,E., Lankas,F., Lavery,R., Maddocks,J.H., Osman,R., Sklenar,H. *et al.* (2005) Molecular dynamics simulations of the 136 unique tetranucleotide sequences of DNA oligonucleotides II: Sequence context effects on the dynamical structures of the 10 unique dinucleotide steps. *Biophys. J.*, **89**, 3721–3740.
37. Beveridge,D.L., Cheatham,T.E. III and Mezei,M. (2012) The ABCs of molecular dynamics simulations on B-DNA, circa 2012. *J. Biosci.*, **37**, 379–397.
38. Gebetsberger,J. and Polacek,N. (2014) Slicing tRNAs to boost functional ncRNA diversity. *RNA Biol.*, **10**, 1798–1806.
39. Ding,Q., Zhu,H., Zhang,B., Soriano,A., Burns,R. and Markesbery,W.R. (2012) Increased 5S rRNA oxidation in Alzheimer's disease. *J. Alzheimer's Dis.*, **29**, 201–209.
40. Vilotti,S., Codrich,M., Dal Ferro,M., Pinto,M., Ferrer,I., Collavin,L., Gustincich,S. and Zucchelli,S. (2012) Parkinson's disease DJ-1 L166P alters rRNA biogenesis by exclusion of TTRAP from the nucleolus and sequestration into cytoplasmic aggregates via TRAF6. *PLoS One*, **27**, e35051.
41. Levinger,L., Mörl,M. and Florentz,C. (2004) Mitochondrial tRNA 3′ end metabolism and human disease. *Nucleic Acids Res.*, **32**, 5430–5441.
42. Jia,Z., Wang,X., Qin,Y., Xue,L., Jiang,P., Meng,Y., Shi,S., Wang,Y., Qin Mo,J and Guan,M.X. (2013) Coronary heart disease is associated with a mutation in mitochondrial tRNA. *Hum. Mol. Genet.*, **15**, 4064–4073.
43. Washietl,S., Will,S., Hendrix,D.A., Goff,L.A., Rinn,J.L., Berger,B. and Kellis,M. (2012) Computational analysis of noncoding RNAs. *Wiley Interdiscip. Rev. RNA.*, **3**, 759–778.
44. Gutschner,T. and Diederichs,S. (2012) The hallmarks of cancer: a long non-coding RNA point of view. *RNA Biol.*, **9**, 703–719.
45. Malumbres,M. (2013) miRNAs and cancer: an epigenetics view. *Mol. Aspects Med.*, **34**, 863–874.
46. Tammen,S.A., Friso,S. and Choi,S.W. (2013) Epigenetics: the link between nature and nurture. *Mol. Aspects Med.*, **34**, 753–764.
47. Zhou,G., Shi,X., Zhang,J., Wu,S. and Zhao,J. (2013) MicroRNAs in osteosarcoma: from biological players to clinical contributors, a review. *J. Int. Med. Res.*, **41**, 1–12.
48. Mauro,V.P. and Edelman,G.M. (1997) rRNA-like sequences occur in diverse primary transcripts: implications for the control of gene expression. *Proc. Natl. Acad. Sci. U.S.A.*, **94**, 422–427.
49. Rohs,R., West,S.M., Sosinsky,A., Liu,P., Mann,R.S. and Honig,B. (2009) The role of DNA shape in protein-DNA recognition. *Nature*, **461**, 1248–1253.
50. Bansal,M., Kumar,A. and Yella,V.R. (2014) Role of DNA sequence based structural features of promoters in transcription initiation and gene expression. *Curr. Opin. Struct. Biol.*, **25**, 77–85.