Indian J. Phys. **79**(**6**), 559-562 (2005)

Rapid Communication
AN INTERNATIONAL JOURNAL
IJP

Algorithm for pattern recognition in nano-sized archaea

Jayprokas Chakrabarti*, Satyabrata Sahoo, Bibekanand Mallick, Smarajit Das and Zhumur Ghosh
Computational Biology Group (CBG), Department of Theoretical Physics,
Indian Association for the Cultivation of Science, Kolkata-700 032, India

E-mail : tpjc@iacs.res.in

Received 1 April 2005, accepted 29 April 2005

Abstract : Hidden patterns abound in genome sequences. Sophisticated mathematical algorithms spot them. As of now, several powerful tools exist for identification of transfer-RNA genes from genomes. These sometimes fail to identify when introns are at noncanonical sites. We discuss our approach to this problem of identification and apply it to the genome of *Nanoarchaeum equitans*. Using our algorithm, we identify the four tRNA genes that were missed by the present standard tRNA search programs in *N. equitans.* The recent split-tRNA hypothesis [*Nature* **433**, 537 (2005)] identified the missing ones. However, our solutions are different. We argue the case in favour of our solutions.

Keywords : tRNA, split-tRNA hypothesis, tRNA search program, *Nanoarchaeum equitans.*

PACS Nos. : 87.14.-g, 87.15.-v, 89.20.-a, 89.75.-k

1. Introduction

Many sophisticated nonlinear algorithms [1] exist for pattern formation and recognition [2,3]. For instance, there are several computational approaches to detect transfer-RNA (tRNA) genes from a genome [4]. These tRNA genes have characteristic pattern over the genomes. To identify these on the sequences, there are algorithms. Notable amongst these are tRNAScan-SE [5] and ARAGORN [6]. Most of these tRNA search programs key on primary sequence patterns and/or secondary structures specific to tRNAs. Quite a few loopholes exist. These have to do with the inability of existing routines to identify tRNA genes with noncanonical introns in them. These are unusually located introns in tRNA genes (tDNAs). The standard (canonical) introns are located between bases 37 and 38 in tDNA. The noncanonical introns are the ones located elsewhere [7]. Identification of tDNAs harbouring these noncanonical introns is the subject of this paper. Some of the tRNA genes are either misidentified or missed by existing search algorithms. In this work, we discuss some of these misidentified and non-identified tRNA genes in the nano-sized *Nanoarchaeum equitans* Kin4-M (*N. equitans* for short) by our in-house algorithm.

N.equitans (NC_005213) belonging to the novel archaeal phylum 'Nanoarchaeota'[8], so far seem to have the smallest genome of all known cellular life forms. *N. equitans* is a hyperthermophile. This is the most compact, with 95% of the DNA predicted to encode proteins or stable RNAs. It is the smallest genome resembling an intermediate between smallest living organism like *Mycoplasma genitalium* and big viruses like pox virus. Many symbiotic or parasitic bacteria have small cells and reduced genomes but within archaea *N. equitans* is the first reported archaea to have such characteristics. Again, its extreme living conditions correlate to early environmental conditions suggesting that 'Nanoarchaeota' are a primitive form of microbial life.

The primary tRNA sequence changes to secondary cloverleaf structure [9]. The secondary structure of tRNA has : (i) Acceptor or A-arm. In this, 5′ and 3′ ends of tRNA are base paired into a stem of 7 bps (ii) DHU or D-Arm. Structurally a stem-loop, D-Arm frequently

*Corresponding Author

contains the modified base dihydrouracil. (iii) Anticodon or AC-arm, made of a stem and a loop containing the anticodon. The canonical structure of AC-loop is essential for interactions with ribosomal A and P sites during protein synthesis [10]. At 5′ end of this loop is a pyrimidine base at 32, followed by an invariant U at 33. The anticodon triplet, at 34, 35, 36 is in the exposed loop region. (iv) An Extra Arm, or V-Arm. This arm is not always present. It is of variable length and is largely responsible for the variation in length of tRNAs. The classification of tRNAs into types I and II, depends on length of V-arm [11]. (v) T-ψ-C Arm or T-arm : This arm has conserved sequence of three ribonucleotides : ribothymidine, pseudouridine and cytosine. T-arm has stem-loop secondary structure and (vi) tRNA terminates with CCA at 3′ end. In case CCA is absent in tDNA, it is added during maturation to tRNA.

The attachment of amino acid to their corresponding tRNA is catalyzed by aminoacyl-tRNA synthetase (AARS) [12,13]. Accurate acylation of tRNA depends on two factors : a set of nucleotides in tRNA molecule (identity elements) responsible for proper identification by AARS [14] and competition between different synthetases for tRNAs [15]. Tertiary L-shape of tRNA facilitates its identification by AARS for aminoacylation. L-shape comes about through the interaction between D-arm and T-arm. There are a few key features that maintain the L-shape of tRNA [16]. These interactions include Watson-Crick base pairing, Hoogsteen base pairing, and triple-helical base pairing. It is generally accepted that the major interactions maintaining the L-shape occur at the corner of the molecule where D- and T-loops meet. This region, called DT [17], contains several elements, including the reverse-Hoogsteen bp U54:A58 and C55-mediated U-turn in T-loop, the inter-loop bps G18:C55 and G19:C56 and stack of four mutually intercalated purine bases A58–G18–R57–G19. This intra-loop U54:A58 is stacked on G53.C61 at the end of T stem and forces the two bases at positions 59 and 60 to loop out, forming a characteristic T-loop of 5 bases instead of 7. This characteristic T loop conformation is important for recognition by elongation factors.

The genome of *N. equitans* consists of a single, circular chromosome of 490, 885 base pairs (bp). It has an average G+C content of 31.6% [18]. Presumably because of this small genome, this archaea has an unusually high gene density, and stable RNA sequences, together covering 95% of the genome. 38 tRNA genes are reported and cross-checked using standard routines (tRNAScan-SE and ARAGORN) include an unusual second copy of $tRNA^{Ser}$(CGA). However, four tRNA genes (for glutamate, histidine, tryptophan and initiator methionine) remained unidentified in the genome. This is due to their unusual sequence or structure. We identify them now using our in-house-developed software. Recently, these missing tRNAs were identified using a new split-tRNA hypothesis. However, our solutions are different. We argue the case in favour of our solutions.

2. Methodology

The entire genome is obtained from NCBI (http://www.ncbi.nlm.nih.gov), accession no. NC_005213. Raw tDNA sequences are found by searching the different motifs present in the consensus sequence of different tDNAs of archaea. At first, we adopted the standard cloverleaf model for studying the secondary structure of predicted tDNAs of *N. equitans*. In doing so, we got some false positives and a few tDNAs were missed out. We then imposed constraints, unique to archaeal tDNA. A regular cloverleaf structure was searched in tRNA genes of the genome of *N. equitans* by adopting archaeal tDNA features. The constraints of lengths of stems of regular tDNA, acceptor arm, D-arm, anticodon arm and T-arm are 7, 4, 5 and 5bp respectively. That aside parameters and constraints used in the search for cloverleaf tDNAs are : (a) T8 (except Y8 in *M. kandleri*), G18, R19, R53, Y55, and A58 are considered as conserved bases for archaea. (b) the lengths of introns and V-arm are allowed from 6 to 121 and up to 21 respectively; (c) positions optionally occupied in D-loop are 17, 17a, 20a and 20b; (d) canonical and noncanonical introns may or may not be present. Keeping these constraints, we were able to extract 38 tDNAs. After getting the tDNAs, we ran the standard routines to check for the secondary structure. We developed consensus tRNA sequences for archaea and measured homology with tRNA of *N. equitans* as a further check.

3. Results and discussion

The recent algorithm [19] for five split tDNAs in *N. equitans* is new. It locates missing $tRNA^{Trp}$, $tRNA^{iMet}$, $tRNA^{Glu}$ and $tRNA^{His}$. But the split $tRNA^{Trp}$(CCA) solution is anomalous; the $tRNA^{iMet}$ solution [19] lacks cognition elements for aminoacylation. In view therefore, we present here alternate non-split composite solutions for $tRNA^{Trp}$, $tRNA^{iMet}$, $tRNA^{Glu}$ and $tRNA^{His}$.

Earlier [8], tRNA genes in *N. equitans* were exhaustively explored. The remarkable algorithms

tRNAScan-SE [5] and ARAGORN [6] located all tRNAs except $tRNA^{Trp}$, $tRNA^{iMet}$, $tRNA^{Glu}$ and $tRNA^{His}$. The new algorithm of Randau *et al* [19] locates these missing ones.

However, the $tRNA^{Trp}$(CCA) reported [19] is anomalous : (i) There is GG preceding the anticodon. We studied all archaeal $tRNA^{Trp}$(CCA) and found this to be an exception. U33 is known [20] to contribute to tRNA-ribosomal binding. Its absence is puzzling. (ii) Further, archaeal $tRNA^{Trp}$(CCA) always have discriminator base A73. This discriminator A73 is of modest preference for aminoacylation [21]. Randau *et al* [19] have C73. Again, the 73rd discriminator base of archaeal $tRNA^{iMet}$(CAU) is always A73. But, $tRNA^{iMet}$(CAU) solution [19] is anomalous, it has U73.

In the absence of conclusive aminoacylation experiment and the anomalies listed above, we reanalyzed the missing tRNAs for Nanoarchaea. In the split-tRNA hypothesis [19], the structures (5-primed end split at 37 followed by invert-repeat element, 3-primed end preceeded by invert-repeat element *etc.*) of tDNA-Glu/His are similar to tDNA-Trp/iMet. If tRNA-Trp/iMet are anomalous, how functional are tRNA-Glu/His? Are there other solutions? From the classic work [22] (and the references therein) on tRNA, it is known that archaeal tRNA harbour noncanonical introns. Canonical introns are located between bases 37 and 38 of tRNA; noncanonical introns occur elsewhere. We looked for the possibility that tDNA-Trp/iMet/Glu/His have noncanonical introns. We found composite solutions that do not suffer from the anomalies above. These solutions are :

$tRNA^{Trp}$ gene 151992-152078

5'-TAGAAAAATT**TTTAAA**TATCTATCTATTGCAATCTC ***GGG GCCGTAGCTCAGCCAGGCAGAGCGGCGGATTCGAAGCCGAA GCTCCAGACCCGTAGGTCGGGGGTTCGAATCCCCCCGGCCC CA*** -3'

$tRNA^{Ala}$ gene 36249-36429

5'-TCGTTAATTCTTACAGTAACATT**TATAAA**TGGTTTTGTTAT AACTTACTA***CGCGGGGTGGGGCAGCCTGGAGTGCCCGGGGGG CTGATATCCCCCTGGCCGCCTTTT***TTCATATTTAATGGACCCG CCGGGATTCGAACCCGGGGCCTCCGCCTTGCGAGGGCGGCGT CCTACCGCTGGA***CTACGGG***CCC***GGTTTC***GATTTTAGATACAAA ATAAATACATTTTT ***TGTAA***-3'

$tRNA^{Glu}$(CTC) gene 151992-152078

5'-AATT**TTTAAA**TATCTATCTATTGCAATCTC ***GGGGCCGTAG CTCAGCCAGGCAGAGCGGCGGATT***CGAAGCCGAAG ***CTCCAG ACCCGTAGGTCGGGGGTTCGAATCCCCCCGGCCCCA***-3'

$tRNA^{His}$(ATG) gene 327362-327625

5'-ATAATT**TTTAAA**TCGTTTCTTTTATTCTATTG ***GCCGCCGTAGCT CAGCGGTCAGAGCGCCCCGGCT***CATAGCATGGGC ***TATGA***GCTCTGAC CCGAAAGGGGATGATCTCGGGGGCTCTTATGCCCCCTCGTGAGAAA ***CCGGGAGGTCGCGGGTTCGAATCCCCGCGGCGG***CATCACAATTTTT ATATAAACCTAA***CC***-3'

Here, we have marked tRNAs in bold *italics*, introns in normal font within the gene sequence, the conserved archaeal Box A promoter-elements [23] in larger font present ahead of the gene. We found the right secondary structures for all these tRNAs, and the bulge-helix-bulge (BHB) motifs. Note, for instance, the following important features of this $tRNA^{Trp}$(CCA) : U8, A14, A21, U33, G18:U55, G19:C56, U54:A58 and G30:C40, the anticodon CCA at 34, 35, 36, and finally A73. These bases/base-pairs are conserved in all $tRNA^{Trp}$(CCA) in archaea. tRNAScan-SE identifies bases 151992 to 152081 as $tRNA^{Ser}$(CGA). Note there is another $tRNA^{Ser}$(CGA) between 486337 and 486426. The one between 151992 and 152081 is unlikely to be $tRNA^{Ser}$(CGA) : none of the conserved bases/base pairs of archaeal $tRNA^{Ser}$ *viz.* G1:C72, G18:U55, G19:C56, U54:A58, G26:U44, G53:U61, U33, G73 appear. Again, the Variable-arm is absent. It is known [24] that G73 and Variable-arm contain identity elements for Ser-RS.

From our study of 22 fully sequenced archaea, the 73rd discriminator base of $tRNA^{iMet}$(CAU) is A73. Our $tRNA^{iMet}$(CAU) has A73. It shares all features of archaeal $tRNA^{iMet}$(CAU).

Remarkably, our $tRNA^{Glu}$(CUC) and $tRNA^{Trp}$(CCA) overlap with one another. Note that the $tDNA^{Glu}$(CUC) has a noncanonical intron at 33. $tDNA^{Trp}$(CCA) has a noncanonical intron at 30. *N. equitans* has the smallest genome known. Noncanonical introns here compactify two tDNAs. Interestingly, this compactification is at work for $tRNA^{His}$ as well.

Codon usage study of histidine in 22 archaea reveals the ratio of the number of CAU-codon to CAC-codon to be anomalously high in *N.equitans.* Amongst archaea *N. equitans* is special in this respect. For $tRNA^{His}$ ATG is the likely anticodon. This is precisely what we found : $tDNA^{His}$(ATG) lying between 327362 and 327520. It has two noncanonical introns located between 32/33 and 71/72 of 13 and 25 bases respectively. In addition to these, there is the canonical intron of 53 bases. Remarkably again, this $tDNA^{His}$(ATG) overlaps with $tDNA^{eMet}$(CAU), located between 327362 and 327500. $tDNA^{eMet}$(CAT) has a canonical intron of 66 bases.

Randau *et al*'s split-tRNA solutions are new. Splitting decompactifies the genome. Further, some of the split solutions are anomalous. Our solutions have overlapping composite tRNA genes [25]. tRNA genes are woven together by introns. They appear just suited for *N. equitans* that has the smallest genome.

References

[1] J D Murray *Mathematical Biology : An Introduction* (3rd edn.) (Berlin : Springer Verlag) (2002)

[2] M C Cross and P Hohenberg *Rev. Mod. Phys*. **65** 851 (1993)

[3] A J Koch and H Meinhardt *Rev. Mod. Phys*. **66** 1481 (1994)

[4] N el-Mabrouk and F Lisacek *J. Mol. Biol*. **264** 46 (1996)

[5] T M Lowe and S R Eddy *Nucleic Acids Research* **25** 955 (1997)

[6] D Laslett and B Canback *Nucleic Acids Research* **32** 11(2004)

[7] M Sugita, L Luo, M Ohta and H Itadani *DNA Research* **2** 71 (1995)

[8] E Waters, M J Hohn, I Ahel, D E Graham, M D Adams, M Barnstead, K Y Beeson, L Bibbs, R Bolanos, M Keller, K Kretz, X Lin, E Mathur, J Ni, M Podar, T Richardson G G Sutton, M Simon, D Söll, K O Stetter, J M Short and M Noordewier *Proc. Natl. Acad. Sci*. **100** 12984 (2003)

[9] N Kanjo and H Inokuchi *DNA Research* **6** 71 (1999)

[10] D Thirumalail, V Ashwin and J K Bhattacharjee *Phys. Rev. Lett*. **77** 5385 (1996)

[11] R Giege, M Sissler and C Florentz *Nucleic Acids Research* **26** 5017 (1998)

[12] M Ibba and D Soll *Annu. Rev. Biochem*. **69** 617 (2000)

[13] M Ibba and D Soll *EMBO Reports* **2** 382 (2001)

[14] R Giege, J D Puglisi and C Florentz *Nucleic Acid Research* **45** 129 (1993)

[15] J M Sherman, M J Rogers and D Soll *Nucleic Acids Research*. **20** 2847 (1992)

[16] S V Steinberg, F Leclerc and R Cedergren *J. Mol. Biol*. **266** 269 (1997)

[17] E I Zagryadskaya, N Kotlova and S V Steinberg *J. Mol. Biol*. (in press)

[18] S Chattopadhyay, S Sahoo, W A Kanner and J Chakrabarti *J. Comp. Funct. Genom.* **4** 56 (2003)

[19] L Randau, R Munch, M J Hohn, D Jahn and D Söll *Nature* **433** 537 (2005)

[20] S S Ashraf, G Ansari, R Guenther, E Sochacka, A Malkiewicz and P F Agris *RNA* **5**(4) 503(1999)

[21] Q Guo, Q Gong, Ka-Lok Tong, B Vestergaard, A Costa, J Desgres, M Wong, H Grosjean, G Zhu, J T Wong and H Xue *J. Biol. Chem*. **277** 14343 (2002)

[22] C Marck and H Grosjean *RNA* **9** 1516 (2003)

[23] J R Palmer and C J Daniels *J. Bact.* **177** 1844 (1995)

[24] R Giege, M Sissler and C Florentz *Nucleic Acids Research* **26** 5017 (1998)

[25] A Reichert, U Rothbauer and M Mörl *J. Biol. Chem* **273** 31977 (1998)