Available online at www.sciencedirect.com
ScienceDirect

Journal of Theoretical Biology 248 (2007) 745–753

Journal of Theoretical Biology

www.elsevier.com/locate/yjtbi

Calculation of folding energies of single-stranded nucleic acid sequences: Conceptual issues

Donald R. Forsyke*

Department of Biochemistry, Queen's University, Kingston, Ontario, Canada K7L3N6

Received 30 May 2007; received in revised form 5 July 2007; accepted 9 July 2007
Available online 18 July 2007

Abstract

The stability of a folded single-stranded nucleic acid depends on the composition and order of its constituent bases and may be assessed by taking into account the pairing energies of its constituent dinucleotides. To assess the possible biological significance of a computed structure, Maizel and coworkers in the 1980s compared the energy of folding of a natural single-stranded RNA sequence with the energies of several versions of the same sequence produced by shuffling base order. However, in the 2000s many took as self-evident the view that shuffling at the mononucleotide level (single bases) was conceptual wrong and should be replaced by shuffling at the level of dinucleotides (retaining pairs of adjacent bases). Folding energies then became indistinguishable from those of corresponding shuffled sequences and doubt was cast on the importance of secondary structures. Nevertheless, some continued productively to employ the single base shuffling approach, the justification for which is the topic of this paper. Because dinucleotide pairing energies are needed to calculate structure, it does not follow that shuffling should not disrupt dinucleotides. Base shuffling allows determination of the relative contributions of base composition and base order to total folding energy. The potential for secondary structure arises from pressures acting at both DNA and RNA levels, and is abundant throughout genomes—with a probable primary role in recombination. Within a gene the potential can often be accommodated, and base order and composition work together (values have the same negative sign) in contributing to total folding energy. But sometimes protein-coding pressure on base order conflicts with the pressure for secondary structure and the values have opposite signs. Total folding energy can be deemed of potential biological significance when the average of several readings is significantly less than zero.

Keywords: Base composition; Base order; Base shuffling; Dinucleotide; Folding energetics; Gene identification; MicroRNAs; Non-coding RNAs; Recombination; Secondary structure

1. Introduction

Apart from their encoding of proteins, single stranded nucleic acids have numerous other roles that involve their adoption of higher order structures (Forsdyke, 2006). Interest in computational approaches to the determination of nucleic acid structure has increased with the recognition of an abundance of non-coding RNAs (ncRNAs) in genomes, other than the well characterized ribosomal and transfer RNAs (rRNAs and tRNAs). However, the quest to identify and to establish genomic locations for ncRNAs through their structures is hindered by serious conceptual problems.

The folding of a nucleic acid is hierarchical and sequential—the primary sequence determines secondary structure, which, in turn, determines higher ordered structure (Tinoco and Bustamante, 1999). A computer method for displaying the potential of successive segments of single-stranded nucleic acids to fold into secondary structures was developed by Le and Maizel (1989) for RNA, and was extended to DNA (Forsdyke, 1995a, b; Heximer et al., 1996). The difference between the folding energy value of a natural segment, and the mean of the folding energy values of several versions of the same segment generated by randomly shuffling base order, produced a metric ("segment score" or "FORS-D value")

*Tel.: +1 613 533 2980; fax: +1 613 533 2490.
E-mail address: forsdyke@queensu.ca

0022-5193/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2007.07.008

that could be related to functional aspects of the segment. However, the growing use of the method (specifically a study by Seffens and Digby in 1999), was brought into question by Workman and Krogh (1999), who considered that shuffling at the mononucleotide level (single bases) was conceptually wrong and should be replaced by shuffling at the level of dinucleotides (retaining pairs of adjacent bases). The observation that this dinucleotide shuffling approach failed to demonstrate that tRNAs had folding energies distinguishable from their shuffled versions, was dismissed as revealing that "the method is not always sensitive enough to discriminate between random sequences and RNA with a known secondary structure."

The view of Workman and Krogh won wide support (Rivas and Eddy, 2000; Katz and Burge, 2003; Clote et al., 2005). Yet the alleged Le-Maizel house-of-cards did not tumble down over night. Some continued productively employing the individual base shuffling approach (Le et al., 2001, 2002, 2003; Forsdyke, 2002; Xue and Forsdyke, 2003; Washietl and Hofacker, 2004; Zhang et al., 2005a, b). Indeed, "good results" with the "simple model" appeared to justify "neglect" of dinucleotide shuffling (Washietl et al., 2005). Others, while seeing "no clear solution to the dilemma," suggested relaxation of "the constraint that every dinucleotide count … be preserved" (Babak et al., 2007). I here discuss various aspects of nucleic acid folding in the hope of shedding some light on the controversy. It seems that, although there are problems with the original formulation of Le and Maizel (1989), there are even more with that of Workman and Krogh (1999). Furthermore, few seem to have thought in terms of a folding pressure arising primarily at the DNA level rather than at the level of the RNA transcribed from that DNA.

2. RNA and DNA folding

Transcribed RNA is synthesized sequentially beginning at the 5′ end and terminating at the 3′ end. In the crowded intracellular environment, with protein concentrations around 300 mg/ml, the folding of the 5′ end of RNA should begin prior to the synthesis of (and hence without necessary reference to the structure of) the 3′ end (Forsdyke, 2006). The final structure would then be partly determined by this sequential mode of synthesis and partly by interaction with other cellular components, including RNA chaperones (Cristofari and Darlix, 2002). Approaches to determining the final operational structure of an RNA have included the computer-assisted folding of the entire sequence, or of sequential sections of that sequence, the latter approach being more likely to reflect a sequential mode of assembly during transcription. That the structure was likely to be of biological importance was obvious in the case of some non-protein-coding RNAs (tRNAs, rRNAs), but was not obvious for mRNAs. Yet, the potential for structure of mRNAs and the possibility that this might require accommodation to their protein-coding role has long been recognized (Ball, 1973; Forsdyke and Mortimer, 2000), and has gathered much support (Meyer and Miklós, 2005; Shabalina et al., 2006).

Segments of single-stranded DNA are potentially extrudable from duplex DNA, especially when it has been subjected to negative supercoiling (Murchie et al., 1992; Krueger et al., 2006) and contains palindrome-like repeats (McMurray, 1999; Kogo et al., 2007). That this DNA property is widely and abundantly distributed along molecules, and is of general occurrence, is suggested by (i) the approximate equifrequencies of complementary oligonucleotides (e.g. CAT and AUG) throughout the genomes of many species ("Chargaff's second parity rule;" Forsdyke and Mortimer, 2000), (ii) direct measurements of folding potential (Forsdyke, 1995c; Heximer et al., 1996), and (iii) association with recombination (Zhang et al., 2005a, b). The duplex strand "unpairing" model of Crick (1971) postulated that such extrusion would occur during meiotic recombination (Forsdyke, 2007a). Many genomic translocations involve recombinations between sequences of similar base composition and distinctive propensities for secondary structure (Gotter et al., 2004). The propensity for such secondary structure would be conserved if it bestowed advantages either at the level of the conventional phenotype (natural selection) or of the genome phenotype (physiological "reprotypic" selection; Forsdyke, 2001, 2006). If not disadvantageous, conservation could also occur by chance isolation in small founder populations (random drift).

3. Principles of secondary structure calculation

Calculations of the secondary structures of single-stranded nucleic acids take into account the energetics both of the stems (which usually contribute to stability) and of various loops and bulges (which usually decrease stability). Despite many complexities such calculated structures have proved valuable guides to the corresponding higher ordered structures, and hence to potential biological functions (Zuker, 2000; Mathews, 2006). The distribution of folding potential along a nucleic acid may be evaluated by calculating stability values (folding free energies) for consecutive windows (segments of uniform length) along the sequence. As far as the nucleic acid is concerned, the values arrived at depend—indeed, can only depend—on what bases are present in a window (base composition) and how they are ordered (base order). Although each element of a secondary structure (stems, loops, bulges) has to be considered separately, happily, decomposing a window sequence into its constituent overlapping dinucleotides (for each of which the energetics of base pairing with its complement is known) usually suffices to determine stem energetics, and it is not necessary to take into account higher order oligonucleotides (Borer et al., 1974; SantaLucia et al., 1996). Thus a trinucleotide (3 bases) can be decomposed into two overlapping "nearest neighbour" dinucleotides (each of 2 bases that overlaps the other by 1 base). Similarly, a tetranucleotide (4 bases) can