Journal of Biomolecular Structure & Dynamics, ISSN 0739-1102
Volume 24, Issue Number 2, (2006)

Tracing Ancient mRNA Hairpins

http://www.jbsdonline.com

Idan Gabdank[1,*]
Danny Barash[1,2]
Edward N. Trifonov[2]

[1]Department of Computer Science
Ben Gurion University of the Negev
P.O.B 653
Be'er Sheva 84105, Israel
[2]Genome Diversity Center
Institute of Evolution
University of Haifa
Haifa 31905, Israel

Abstract

From recent developments of the early evolution theory it follows that the earliest mRNAs were short (~20 nt) (G+C)-rich polynucleotides. These short sequences could form hairpins, which would be of high evolutionary advantage because of stability and uniqueness of their conformations. Due to mutations accumulated during billions of years of evolution, the speculated earliest hairpins would largely lose the initial complementarities. Some of the original complementary base-to-base contacts, however, may have survived. Computational analysis of modern prokaryotic mRNA sequences reveals excess population of the expected short range complementarities. The derived earliest mRNA hairpin size fully corresponds to the predicted size of ancient coding duplexes. The repertoire of the surviving hairpins traced in modern mRNA confirms duplex structure of the earliest mRNA, suggested by the early molecular evolution theory.

Key words: Evolution of triplet code; Ancient binary alphabet of proteins; mRNA hairpins; Repertoire of ancient mRNA; Earliest mini-genes; Earliest mRNA; and Computational sequence analysis.

Introduction

The notion of gradual change from simple Life forms to more advanced ones is at the very basis of evolutionary theory. Remarkable progress of Life Sciences in the last half-century added to this view a molecular dimension by demonstrating evolutionary relationships amongst genomes and genes. Inevitably, it comes to the earliest, molecularly simplest events, to the very origin of Life.

Already in 1953, experiments conducted by S. Miller (1-3) demonstrated that in conditions imitating primordial atmosphere, the chemically simple amino acids can be synthesized abiotically. The early Life may have started largely with what the environment would provide, omitting some biosyntheses. Darwinian retrospective would suggest then that the earliest Life forms could content with a limited repertoire of genes and proteins, which, in their turn, did not have to be as large as today (4). Respective ancient nucleotide and amino acid sequences or some of their original features may have survived and, perhaps, could be traced within present-day sequences. A challenging goal would be an attempt to reconstruct these early sequences and the earliest biomolecular events from those traces, if found.

One such reconstruction is the recently derived consensus chronology of amino acids and evolutionary chart of codons (5, 6). Several important predictions are suggested by the evolutionary chart, and confirmed by analysis of modern sequence data (7-9). In particular, the size of the earliest mRNA is predicted and estimated – 18 to 21 bases encoding peptides of 6-7 residues (8). In a different development, a search for universally conserved (omnipresent) protein sequence motifs also points to that size of 6 to 9 amino acid residues (10 and manuscript in preparation).

[*]Phone: (972-8)647-2714
Fax: (972-8)647-7650
Email: gabdank@cs.bgu.ac.il

Would it be possible to reconstruct further characteristics of the earliest mRNA, such as their sequence and structure? In particular, one would expect these small molecules to have had a secondary hairpin structure. The structure would provide stability and better survivability to the molecules.

In this work the analysis of modern mRNA sequences is performed in search for traces of the ancient hairpins. Despite billions of years of mutational changes accumulated in the sequences, certain features are found, indeed, presumably originally belonging to the earliest hairpins. The earliest hairpin mRNAs are estimated to involve about 18-21 bases, in close agreement with the above estimates of the sizes of the earliest mini-genes.

Results and Discussion

What May Have Retained in the Sequences?

The reconstruction of evolutionary history of the triplet code (5, 6) holds that the first two codons to appear were GGC, for glycine, and GCC, for alanine, making a complementary pair. The complementarity, first pointed to by Eigen and Schuster (11), is one of the key principles of the reconstruction. Presumably, the earliest mRNA molecules existed in form of duplexes, such that each of the strands served as mRNA. The hypothesis on the initial coding in both of the complementary strands has been recently strongly supported by detecting significant excess of open reading frames in the "non-coding" strands of DNA genes (12). In the mixture of the earliest mRNAs, therefore, each mRNA would be expected to have its complementary counterpart, in equal amount.

Ancient Hairpins

To search for intact ancient hairpins in modern sequences would be naive. Structural studies of natural mRNA do not reveal any conspicuous stem/loop size or any major difference with random sequence RNA (13-15). All what could remain of the initial complementarities after massive mutational changes is, perhaps, elevated frequencies of complementary matches of individual bases at short distances corresponding to the size of the ancient hairpins. The codon evolution chart suggests an important clue as to which of three codon positions may at least partially survive and still show the hypothetical ancient complementary correspondence. The GGC glycine triplets in one strand of the early mRNA duplex correspond to GCC alanine triplets in the complementary strand. The history of the coding triplets according to the evolutionary chart involves all possible changes in the first and third positions of the triplets, while the middle positions are rather conserved allowing only for G to A and C to U transitions (5, 6). As a result, the original glycine triplets turn into anything with purine (R) in the middle, while alanine triplets, into middle pyrimidine (Y) triplets. Hence, two independent families of codons descending from original glycine and alanine codons, and two families of amino acids, with purine in the central codon position, *G* family: G, D, S, E, R, Q, N, K, C, H, Y, and W, or with central pyrimidine, *A* family: A, V, T, S, P, L, I, F, and M. Notably, the *G* alphabet consists largely of polar and charged residues, whereas the *A* alphabet consists of primarily hydrophobic residues.

According to the codon evolution chart and in agreement with higher frequency of R to R and Y to Y transition mutations, during the mRNA duplex stage of evolution the amino acid changes would have to be conservative so that the replacements would confine to the same kind, of *A* or *G* family, keeping the central R or Y in the mRNA codons unchanged. Remarkably, the amino acid replacements in modern protein sequences, time-wise far away from the early mRNA duplex stage in the evolution of the code, are still conservative in the above sense (9). Indeed, majority of statistically significant substitutions in the PAM-120 matrix (16) are of

the conservative type. Similar effect is observed when the data are taken from the BLOSSUM substitution matrix (17). The conservation of the family type during amino acid replacements is illustrated by Figure 1A where the substitution matrices are shown in rearranged form such that the amino acids encoded by xRx (*G*-type) are placed next to one another, as well as the amino acids encoded by xYx codons (*A*-type). The original full matrices PAM and BLOSSUM after such rearrangement transform in two-box matrices, where essentially all *A*-type to *A*-type substitutions concentrate in one box, while *G* to *G* substitutions concentrate in the second box. If similar control exercise is performed with groups of amino acids encoded by Rxx and Yxx codons, respectively (Figure 1B), the two-box feature disappears. This demonstrates that, indeed, even during amino acid replacements in modern proteins purines in xRx codons are conserved, as well as pyrimidines in the xYx codons. A lesson for our ancient hairpin search problem is that the "complementary" (11) contacts R·Y and Y·R between central bases of ancient codons in the mini-mRNA have better chance to still retain in extant mRNA sequences.

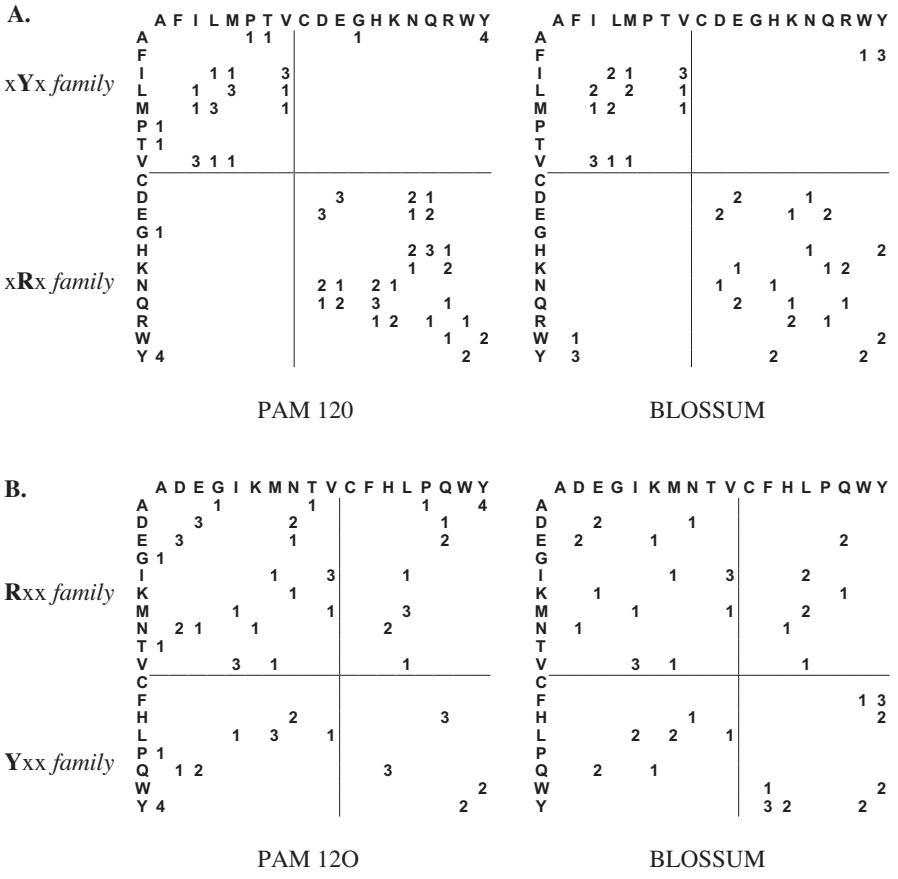

Figure 1: PAM-120 (*left*) and BLOSSUM (*right*) substitution matrices rearranged in (**A**) amino acid families *G* (xRx codons) and *A* (xYx codons), and (**B**) amino acid families corresponding to Rxx codons, and Yxx codons.

In the Figure 2 an example of the hypothetical earliest duplex mRNA $(GGC)_4$ $(GCC)_3$·$(GGC)_3(GCC)_4$ (here of the length 21 base pairs) is shown together with respective two mRNA hairpins with matching codon frames. We are interested, first of all, in the sequence distances between the middle codon positions (in gray). The longest distance between complementary middle G and C in the hairpin of 21 bases as in the Figure 2 is 18 bases (G and C of the two codons at the ends of the chains). Other such distances are 12 and 6 bases. In a similar hairpin of 18 bases (not shown) the complementary distances would be 15 and 9 bases. Thus, the ancient mRNA hairpins of 18 to 21 bases with still conserved complementary R and Y in the middle codon positions would bias the R to Y distances in modern mRNA. An excess of the short distances in the region up to 18 bases is expected in the respective distance histograms. No excess should be observed beyond 18 bases in the illustration case. Generally, the upper limit of the excess region would correspond to the size of the hypothetical earliest mRNA. Eight phylogenetically

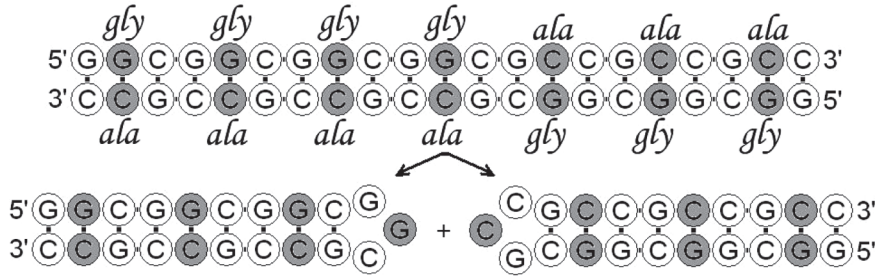

Figure 2: An example of a hypothetical ancient mRNA duplex along with its disassembly products: two hairpins. Middle positions of the triplets which are mostly conserved are in gray.

diverse prokaryotic and archaeaic organisms were taken for the distance analysis: *Archaeoglobus fulgidus*, *Bacillus subtilis*, *Escherichia coli*, *Fusobacterium nucleatum*, *Gloeobacter violaceus*, *Methanosarcina acetivorans*, *Staphylococcus aureus*, and *Sulfolobus solfataricus*. Only middle positions of mRNA codons were considered, and the complementary R to Y distances were scored.

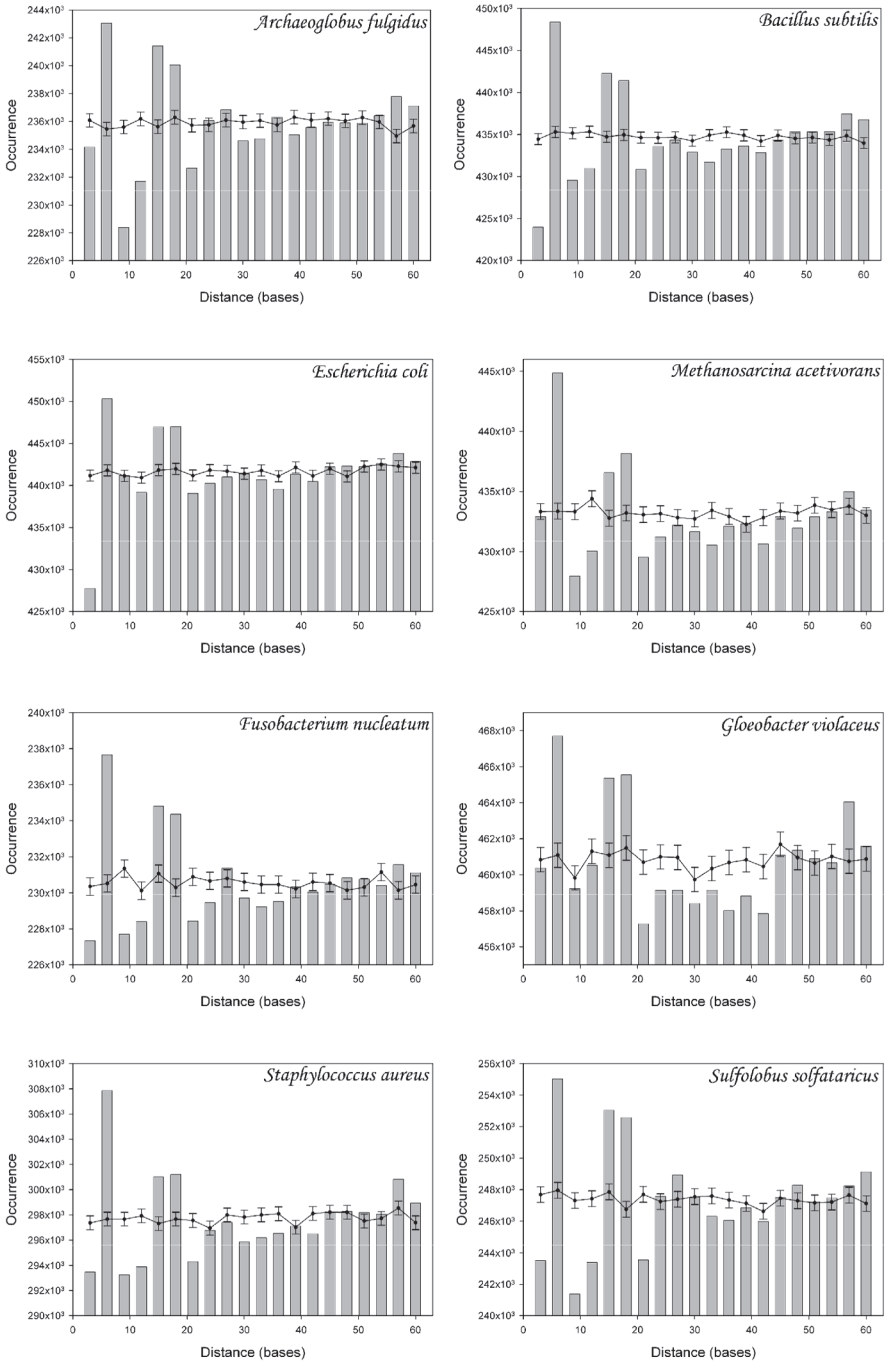

Figure 3: Complementary distance histograms. Middle codon position R to Y distances are measured and scored using mRNA of *Archaeoglobus fulgidus*, *Bacillus subtilis*, *Escherichia coli*, *Fusobacterium nucleatum*, *Gloeobacter violaceus*, *Methanosarcina acetivorans*, *Staphylococcus aureus*, and *Sulfolobus solfataricus*.

The results are shown in Figure 3 for each genome separately. The distance histograms show a very clear excess of complementary base pairs at distances 6, 15, and 18. No elevation is seen at distances 9 and 12 bases. The results, thus, confirm the expected sharp drop of occurrences at the upper limit 18 bases, in exact accordance with the value suggested by the size of the ancient mRNA, as estimated earlier (8). The low occurrences of the complementary distances 9 and 12 bases, are caused, perhaps, by a sequence bias due to amphipathic alpha-helices that have charged and polar residues preferentially on one side of the helix, while hydrophobic residues are on the other side. This condition leads to a natural pressure to avoid occurrence of complementary R and Y bases of middle codon positions at distances 9 and 12 bases that corresponds to 3 and 4 amino acid residues – closest to one turn of alpha helix (3.6 residues). Ten to eleven base periodicity in prokaryotic sequences (18) may also be linked to the observed effect, in synergy with alpha-helices (19). One may also argue that all the ups and downs in the histograms of the Figure 3, including the sharp drop at position 18, are somehow due to the alpha-helical periodicity. This is not the case, however, as the comparison of the combined histogram (Figure 4A) and its filtered version (Figure 4B) demonstrate. To eliminate the apparent oscillation with the period 3.6 triplets in the original histogram, it is smoothed twice by a running average of 9 bases (averaging 3 bins). As the Figure 4B shows, the drop at 18 bases is seen after the filtering as well, while the alpha-helical oscillation is gone.

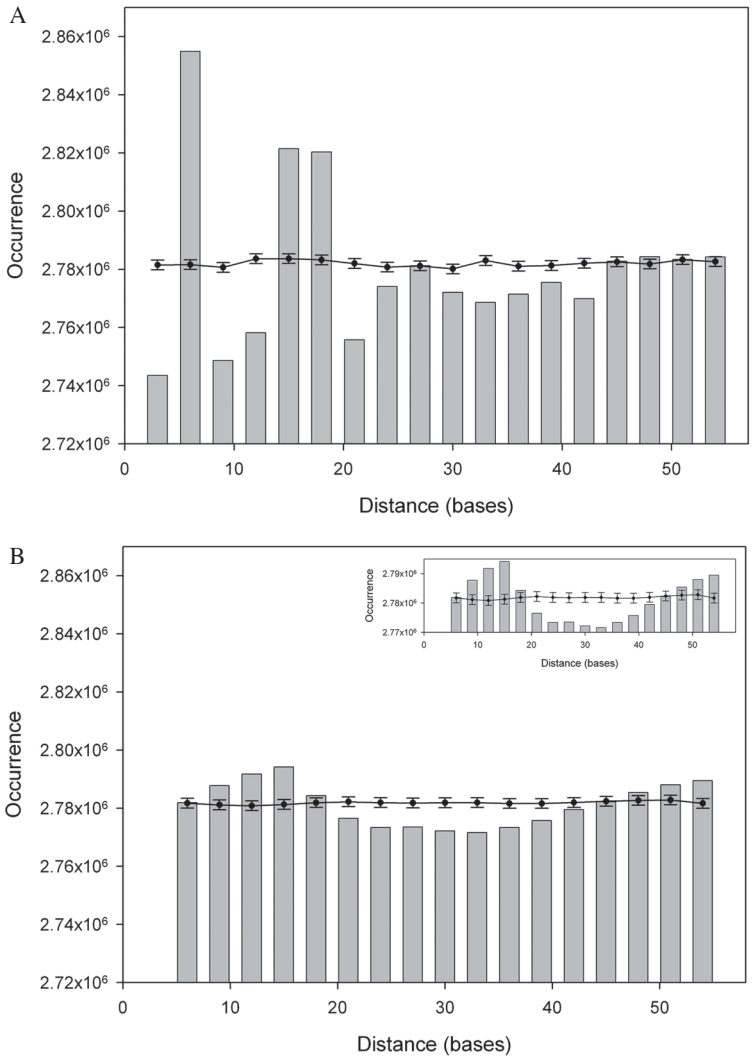

Figure 4: (**A**) Combined histogram of the complementarity Y to R occurrences in middle codon positions detected in eight phylogenetically diverse prokaryotic and archaeaic organisms (see Figure 3). (**B**) Filtration of the alpha-helical oscillation. The (4A) chart was smoothed by a running average of 3 bins. The preferred distances in the region up to 18 codons are also shown in the insert.

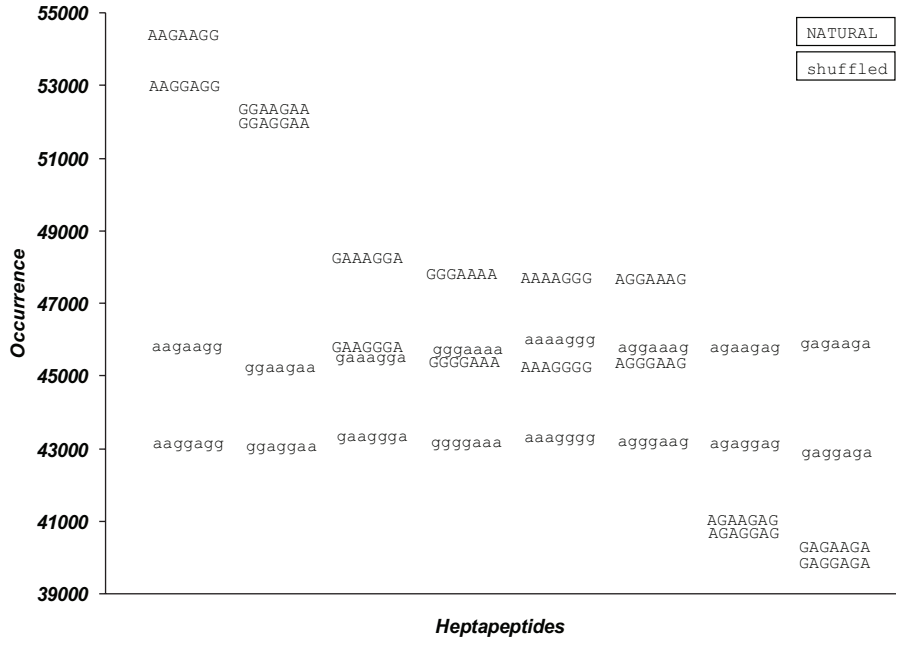

Figure 5: Repertoire of ancient heptapeptides as extracted from modern mRNA sequences. Peptide sequences were rewritten in binary form, A-type residues (for xYx codons) and G-type (for xRx codons). Natural sequences are shown in uppercase letters and shuffled sequences are shown in lowercase letters. Overall composition of the proteomes: *A* 51.85%, *G* 48.15%. This difference causes slight excess of peptides containing four *A* residues over peptides with four G residues.

The surviving complementarity of R and Y bases in the middle of the codons of present-day mRNA sequences appear to represent traces of ancient small mRNA molecules. If respective 21 base hairpin fragments are extracted from the modern mRNA, they may to some degree represent the original repertoire of the ancient mRNA hairpins and of the peptides encoded by the mini-mRNAs. For that purpose one can take simply protein sequences and rewrite them in binary form, with *G*-type residues (encoded by xRx triplets) and *A*-type residues (encoded by xYx triplets). There are a total of 16 possible 7-residue long "complementary" binary amino acid sequences corresponding to the 21 bases long RNA sequences. For example, heptapeptide *AGAGGAG* would correspond to the mRNA sequence xYxxRxxYxxRxxRxxYxxRx where the first three middle codon positions Y-R-Y are complementary to the last three: R-Y-R. In the Figure 5 the 16 "hairpin" heptapeptides extracted from the same eight prokaryotic proteomes are displayed, sorted in descending order of their occurrence. Two important features are revealed. First, the natural occurrences of the heptapeptides are very much different from their occurrences in respective randomized (shuffled) proteomes. The most frequent ones are *AAGAAGG* and *AAGGAGG*, while the least frequent are *GAGAAGA* and *GAGGAGA*. Note, that the binary amino acid compositions of these pairs are the same. Second, the heptapeptides, complementary to one another, appear with approximately the same frequency. If one is frequent (rare), the another one is frequent (rare) as well. This fully corresponds to the prediction (see above) on the basis of the duplex nature of the earliest mRNAs – two complementary mRNAs in one duplex. Similar analysis of hexapeptides revealed the following descending order: AGGAAG, GAAGGA, AAAGGG, GGGAAA, GGAGAA, AAGAGG, AGAGAG, GAGAGA. Since hexapeptides are self complementary the complementary parity rule could not be displayed by the hexapeptides. Interestingly, the complementary tripeptides GAG, AGA are avoided in heptapeptides as well. Perhaps simple alternation of A and G is avoided in both cases.

We, thus, may conclude that the extant prokaryotic mRNA sequences still contain detectable traces of ancient mRNA hairpins. Their size is 18-21 nucleotides, in agreement with other estimates. The complementary parity of the hairpin mRNAs confirms prediction on the duplex nature of early messengers, with both strands coding early mini-proteins.

Methods

Complete sets of mRNA sequences of *Archaeoglobus fulgidus*, *Bacillus subtilis*, *Escherichia coli*, *Fusobacterium nucleatum*, *Gloeobacter violaceus*, *Methanosarcina acetivorans*, *Staphylococcus aureus*, and *Sulfolobus solfataricus* were taken from the TIGR web server. Sequences with unusual start codons and uncertain nucleotides were discarded.

Complementarities were scored up to a distance of 60 nucleotides. The 60 last nucleotides of the mRNA sequences were not analyzed to eliminate end effect, excess of short distances. The shuffling for all mRNA sequences was performed by swapping each nucleotide with an arbitrarily chosen one within a range of 20 triplets. An original program "Scanner" in Java programming language was written specifically for this study. The program can be obtained by request. It uses mRNA sequences in FASTA format as an input and generates data for the histogram of complementary distances detected in the sequences.

Acknowledgment

The research was supported by a grant from the Israel USA binational science foundation BSF 2003291. The authors would like to thank Dr. V. Zhurkin for a fruitful discussion.

References and Footnotes

1. S. L. Miller. *Science 117*, 528-529 (1953).
2. S. L. Miller, H. C. Urey. *Science 130*, 245-251 (1959).
3. S. L. Miller. *Cold Spr. Harb. Symp. Quant. Biol. 52*, 17-27 (1987).
4. E. N. Trifonov. *Origin Life Evol. Biosph.* In press (2006).
5. E. N. Trifonov. *Gene 261*, 139-151 (2000).
6. E. N. Trifonov. *J. Biomol. Struct. Dyn. 22*, 1-11 (2004).
7. E. N. Trifonov. *Ann. NY Acad. Sci. 870*, 330-338 (1999).
8. E. N. Trifonov, A. Kirzhner, V. M. Kirzhner, I. N. Berezovsky. *J. Mol. Evol. 53*, 394-401 (2001).
9. E. N. Trifonov. *Discovering Biomolecular Mechanisms with Computational Biology.* Ed. F. Eisenhaber. Landes Bioscience, Georgetown, in press (2005).
10. Y. Sobolevsky, E. N. Trifonov. *J. Mol. Evol. 61*, 591-596 (2005).
11. M. Eigen, P. Schuster. *Naturwissenschaften 65*, 341-369 (1978).
12. W. L. Duax, R. Huether, V. Z. Pletnev, D. Langs, A. Addlagatta, S. Connare, L. Habegger, J. Gill. *Proteins 61*, 900-906 (2005).
13. V. G. Tumanyan, L. E. Sotnikova, A. V. Kholopov. *Doklady Biochemistry 166*, 63-66 (1966).
14. D. R. Groebe, O. C. Uhlenbeck. *Nucleic Acids Res. 16*, 11725-11735 (1988).
15. W. Fontana, D. A. M. Konings, P. F. Stadler, P. Schuster. *Biopolymers 33*, 1389-1404 (1993).
16. S. F. Altschul. *J. Mol. Biol. 219*, 555-565 (1991).
17. S. Henikoff, J. G. Henikoff. *Proc. Natl. Acad. Sci. USA 89*, 10915-10919 (1992).
18. H. Herzel, O. Weiss, E. N. Trifonov. *Bioinformatics 15*, 187-193 (1999).
19. V. B. Zhurkin. *Nucleic Acids Res. 9*, 1963-1971 (1981).

Date Received: May 10, 2006

Communicated by the Editor Ramaswamy H. Sarma