Identification of promoter sequences using archaeal genomic sequences

By Naoki AMANO, Kuniaki TSUJI, and Masashi SUZUKI^{†)}

National Institute of Advanced Industrial Science and Technology (AIST)

AIST Tsukuba Center 6-10, 1-1-1, Higashi, Tsukuba, Ibaraki 305-8566

(Communicated by Masanori OTSUKA, M. J. A., May 12, 2003)

Abstract: Using the genomic sequence of the hyperthermophilic archaeon *Pyrococcus* sp. OT3, 381 operons and 644 independently regulated genes have been theoretically identified on the basis of the presence or absence of transcription and translation signals upstream of open reading frames, and that of the triplet phasing for combining the purines and pyrimidines inside. Positioned upstream, regions of at shortest 50-60 bps appear to function as promoters for specifically regulating transcription of these 1,025 units.

Key words: Gene regulation; Shine-Dalgarno sequence; TATA box; transcription; translation.

Introduction. The genomic sequences of over 100 organisms have been already determined. Identification of genes using nucleotide sequences is difficult, and so many of the determined genomic sequences are left without being fully analyzed. Theoretical identification of genes is a process of selecting the likeliest candidates from the entire population of open reading frames (ORFs) found in a genome, i.e. runs of triplets able to code amino acid residues, sandwiched between possible start and stop codons. The difficulty here is that the majority of these ORFs are, in fact, not true genes but artifacts.

Most of a bacterial genome is used for coding genes, and two neighboring genes do not generally overlap to a large degree. Thus, each base essentially becomes part of a single gene only. However, at most 6 ORFs can be formed over a base, three using the different coding phases along a DNA strand, and the other three along the complementary strand. Thus, the ratio of the number of false ORFs to that of real genes can be as high as 5. The ratio is even higher for genomes of eukaryotes, where regions not coding genes increase with the genome size. In the human genome over 90% are non-coding regions, where formal ORFs can be formed.

It is also difficult to identify start codons of genes uniquely; the largest possible form of an ORF might not be the correct one. Not only ATG, which is used for cod-

ing Met, but also GTG for Val and TTG for Leu are used as start codons, and a number of these triplets are generally found within each ORF. Having these problems, it is difficult to even make simple statistics of the frequency of 20 amino acid residues in proteins or to make statistics of codon usage.

In order to improve this situation we have invented a method of theoretically identifying genes using archaeal genomic sequences¹⁾⁻⁴⁾; see also a conceptualization of this method.⁵⁾ In this paper some details of the gene identification³⁾ using the complete genomic sequence of the hyper-thermophilic archaeon *Pyrococcus* sp. OT3 (DDBJ accession numbers AB9464-9531) are described; they are slightly different from those used for analyzing an incomplete version of the genomic sequence¹⁾ and those used for a preliminary analysis of the complete version.²⁾ This identification was found to be easier than that for the meso-thermophile *Thermoplasma volcanium*, partly because of clearer signal sequences designed for interactions with proteins and ribosomal RNAs at temperatures of 100 °C or higher.⁴⁾ Even the number of formal ORFs found in the genome, and therefore, that of the false candidates are smaller. Details of the gene identification of *T. volcanium* will be described in another paper.

Strategy for identifying genes. Candidates of genes were defined as 5,363 ORFs found in the genomic sequence of *P. OT3* able to code 50 or more amino acid residues each. The gene-likelihood of these ORFs was characterized by measuring, in succession, the closeness

^{†)} Correspondence to: M. Suzuki at AIST.

Fig. 1. Two types of genes, A and B, identified (a) by the scheme (b), using the genomic sequence of *P. OT3*. The ORFs were screened by examining the presence or absence of the "triplet" phasing inside, followed by examinations of translation ("SD") and transcription ("TATA") signals in their upstream regions. "Distance" was another examination carried out by measuring the distances to the nearest genes positioned upstream (see text).

of three types of characteristics of these ORFs to similar characteristics expected for true genes (Fig. 1b); presence of transcription and translation signals upstream and triplet phasing of positioning purines (Rs) and pyrimidines (Ys) inside; the nature of these characteristics will be explained in the following sections.

Through this process of screening the ORFs, genes were identified, and at the same time classified to two types, A and B (Fig. 1b). Upstream of a type A gene expected was a transcription signal followed by a translation signal (Fig. 1a). Upstream of a type B gene, instead of a transcription signal, another gene needed to be closely positioned along the same strand; in our

Fig. 2. Frequencies of purines (Rs) and pyrimidines (Ys) in reference to the upstream and downstream ends of transcription units, operons and independent genes, identified.

identification, with its downstream end being positioned at -35 to $+122$, when position 0 was defined as the first base of the type B gene. Here 72% of type B genes follow preceding genes positioned within their ± 10 bps upstream, and the one distanced 122 bases is a single and isolated example. In this way, transcription units were classified to those of single type A genes independently regulated, and those composed of A and B genes, i.e. operons (Fig. 1a).

Our method is a process of extracting common characteristics shared by a smaller population, thereby selecting examples of the same quality by searching through a larger population. Thus the standard characteristics expected for true genes (i.e. the signals and the triplet phasing) were identified using a selection of ORFs identified as orthologues of genes essential for the survival of organisms (e.g. RNA polymerase and tRNA synthetases). It is safe to assume that these ORFs are real genes. With these sequences being well conserved among organisms, their start codons were determined relatively easily by multiple sequence alignments.

Triplet phasing. Before further describing the process of gene identification, some characteristics shared by the genes finally identified are discussed in order to show the reliability of this method. In Fig. 2 the frequencies of purines (A and G bases) and pyrimidines (T and C bases) inside and outside the 1,025 transcrip-

Fig. 3. The triplet phasing inside genes. (a) Distribution of scores given to ORFs by calculating matches of the frequencies of the A/T/G/C bases in the three coding phases inside each ORF with similar frequencies averaged for the initial set of genes. (b) The frequencies of the bases in the three phases averaged for the final set of genes identified.

tion units identified, i.e. independent genes and operons, are shown in reference to the predicted start codons of the genes positioned most upstream and stop codons of

the genes positioned most downstream. Flat 50% lines could have emerged, if blocks of sequences were to be collected randomly from the genomic sequence. In contrast, regularities and patterns can be identified in Fig. 2, suggesting that a particular type of blocks sharing the same characteristics were selected. Generally positions and sequences of signals for terminating transcription are less conserved and thus were not used for our gene identification. Nevertheless, though broader and less clear than transcription and translation signals upstream, peaks of Y bases were found downstream of the units.

Inside the genes identified was a zigzag with the periodicity of 3 bases. This triplet periodicity is associated with the use of the particular code for relating triplets and amino acid residues, and the frequency of the 20 amino acid residues in proteins. However, it is not clear whether the latter two factors happen to have created the triplet periodicity, or the triplet periodicity is important, thereby having affected the latter two; see refs 1) and 5), and the papers cited therein. In the first of the three phases, the frequency of R was found to be the highest (Fig. 3b).

For each ORF, a score, $\sum F_{nN}^C \times F_{nN}^S$ for $n=1-3$ and $N=A, T, G, C$, was calculated where, e.g., F_{1A}^C was the frequency (%) of A in phase 1 in the candidate, while, e.g., F_{1A}^S was a similar frequency averaged for genes in the initial set. Distribution of these scores (Fig. 3a) was bimodal. The majority with lower values being false ORFs, and the minority of ~1,800 ORFs was expected to be real genes. The distribution of 1,788 genes finally identified is overlaid in Fig. 3a, which corresponds well to that of the secondary peak in the initial distribution, confirming the overall precision of our gene identification.

Transcription and translation signals. The standard transcription signal used by archaea is the TATA-box (or box A), TTTATATA, following a run of A bases (box B); the TATA-box being recognized by TATA-box binding protein (TBP), and box B by transcription factor B. The standard translation signal is the Shine-Dalgarno (SD) sequence, AGGAGGTGA, which is complementary to the downstream end of 16S rRNA. Sequences closest to these standard sequences were searched by scanning the upstream region of each ORF and by calculating scores; $\sum F_{nN}^C \times F_{nN}^S$ for $N=A/T/G/C$ and $n=1-12$ for the transcription signal and 1-9 for the translation signal, where, e.g., F_{1A}^C was the frequency (%) in the candidate of A at position 1, while, e.g., F_{1A}^S was a similar frequency averaged for genes in the initial set.

The sequences and positions of transcription signals identified 21-130 bases upstream of genes in the initial

Fig. 4. The frequencies of the A/T/G/C bases at positions in the transcription signals identified with the initial (a) and final (b) sets of genes, and in the translation signals identified with initial (c) and final (d) sets.

set and those in the final set were compared (Fig. 4a, b). The two sets were similar to each other, showing the homogeneity between the genes from which the information was extracted and those identified by using the

information. In both sets the consensus sequence was AAagTTTATAAAA, where capital letters were used for the bases conserved in over 50% of the examples, positioned most frequently 30-40 bps upstream of the identified

Fig. 5. The positions of translation signals relative to the start codons of the first genes in the transcription units in the initial (a) and final (b) sets, and those of translation signals relative to the start codons of genes in the initial (c) and final (d) sets.

transcription units (Fig. 5a, b).

The initial and final sets of translation signals identified 0-9 bps upstream of genes were also similar (Fig. 4c, d). Although the same consensus sequence

aGgAGGTGA was found most frequently at the positions 5 bases upstream of the identified genes in both sets (Fig. 5c, d), this frequency was found to decrease from ~40% in the initial set to ~30% in the final set, increasing

Fig. 6. The lengths of regions upstream of transcription units finally identified. Regions of type I are sandwiched between pairs of genes running in the same directions, and those of type II are sandwiched between pairs running in the opposite directions. The two distributions tail beyond 400 bps towards 1,000 bps, and 43 entries of type I and 23 entries of type II in these tails are not shown.

examples with smaller n numbers. This slight difference might have originated in misidentification of some of the start codon positions by a few bases each. For example, when two or more possible start codons were closely positioned and associated with the same SD sequences; e.g. GTG and ATG in [GGAGGTGA]NN[GTG][ATG], the ones positioned most upstream, i.e. GTG, were selected, thereby reducing the n numbers.

Identified transcription units and their promoters. Because of the multiple parameters used for the gene identification, it was difficult to uniquely select a particular set of threshold values. Here we depend on the characteristics associated with the identified genes in order to show the reliability of the selected set of parameter values (Figs. 2-5). Altogether

1,788 protein genes were identified, producing the gene density of 1.05/Kbps. Of these 1,144 genes were grouped into 381 operons, with the average number of genes in each operon, 3.0, and the other 644 were classified as those independently regulated, thereby forming 1,025 transcription units.

Regions upstream of these transcription units were classified into two subgroups (Fig. 6), 447 positioned between pairs of units running in the same directions, thus containing promoters only of the units downstream (type I in Fig. 6), and 268 sandwiched between pairs of units running in the opposite directions, thus containing promoters of two units each (type II in Fig. 6). The other 43 units positioned downstream of RNA genes were not included in these statistics. For type I regions, the length ~50 bps was found to be the most frequent, while that of ~110 bps was the most frequent for type II regions. Thus, it is likely that for each transcription unit the upstream region of at shortest 50-60 bps is used as the promoter, with the TATA-box being positioned around its center. Here, we take the different lengths of types I and II as another indication of the overall precision of our identification of genes, their types and their start codons.

Acknowledgements. This work was supported by the Core Research for Evolutional Science and Technology (CREST) program of the Japan Science and Technology Corporation (JST), and the Mitsubishi Foundation. We thank Dr. Lester Clowney for his critical reading of the manuscripts.

References

- 1) Suckow, J. M., Amano, N., Ohfuku, Y., Kakinuma, J., Koike, H., and Suzuki, M. (1998) FEBS Lett. **426**, 86-92.
- 2) Ohfuku, Y., Koike, H., Azuma, Y., Kawashima, T., Kudo, N., Amano, N., Kakinuma, J., Suckow, J., and Suzuki, M. (1998) Proc. Japan Acad. **74B**, 90-95.
- 3) Koike, H., Amano, N., Tateno, M., Ohfuku, Y., Suckow, J. M., and Suzuki, M. (1999) Proc. Japan Acad. **75B**, 37-42.
- 4) Kawashima, T., Amano, N., Koike, H., Makino, S., Higuchi, S. *et al.* (2000) Proc. Natl. Acad. Sci. U. S. A. **97**, 14257-14262.
- 5) Suzuki, M. (1999) Proc. Japan Acad. **75B**, 81-86.
- 6) Amano, N., Ohfuku, Y., and Suzuki, M. (1997) Biol. Chem. **378**, 1397-1404.