FEBS Letters 426 (1998) 86–92 FEBS 20090

A transcription frame-based analysis of the genomic DNA sequence of a hyper-thermophilic archaeon for the identification of genes, pseudo-genes and operon structures

Jörg M. Suckow[a], Naoki Amano[a], Yuko Ohfuku[a,b,c], Jun Kakinuma[a,d], Hideaki Koike[a], Masashi Suzuki[a,d,*]

[a] *AIST-NIBHT CREST Centre of Structural Biology, Higashi 1-1, Tsukuba 305-0046, Japan*
[b] *Doctoral Program in Agricultural Sciences, University of Tsukuba, Tennohdai 1-1-1, Tsukuba 305-0006, Japan*
[c] *National Institute of Technology and Evaluation, Nishihara 2-49-10, Shibuya, Tokyo 151, Japan*
[d] *Graduate School of Human and Environmental Sciences, University of Tokyo, Komaba 3-8-1, Meguro, Tokyo 153, Japan*

Received 13 February 1998; revised version received 10 March 1998

Abstract An algorithm for identifying transcription units, independently regulated genes and operons, and pseudo-genes that are not expected to be expressed, has been developed by combining a system for predicting transcription and translation signals, and a system for scoring the triplet periodicity in ORF candidates. By using the algorithm, the 1.09 Mb sequence that covers approximately 60% of the genome of *Pyrococcus* sp. OT3 has been analyzed. The identified ORFs show the expected biological and physical characteristics, while the rejected ORF candidates do not. Frequent use of operon structures for transcription, and gene duplication followed by mutation or termination of the duplicated genes, are discussed.

Key words: Transcription; Genome; Nucleotide triplet periodicity; Operon; Pseudo-gene; Archaeobacterium

1. Introduction

Recently, the whole DNA sequences of a dozen genomes have been determined [1–12]. Three of them are that of archaeobacteria ([3,7,8], see also review [13]). Since a systematic collection of archaeobacterial cDNA sequences is not currently available, identification of the genes in these genomes needs to be carried out on a theoretical basis.

It seems scientific to identify open reading frames (ORFs) by predicting signals for transcription and translation, and to apply the same criteria throughout the genomic DNA sequences. Such an analysis will lead to the identification of not only individual protein genes but also operon structures and pseudo-genes. Here pseudo-genes are defined as ORF candidates that are homologous to protein genes known outside the genome or paralogous to ORFs in the same genome, but lack the signal for transcription or translation; note that the definition is essentially the same as that in molecular biology, i.e. 'dead' genes that were functional in the past (thus, these are similar to other genes still functioning) but are not expressed any longer (because these do not have signals). On the other hand misidentification of pseudo-genes as genes and failure in the identification of operon structures are problematic for the understanding of the transcription network. To our knowledge no signal-based analysis has been carried out with any archaeobacterial genomic DNA sequence, and no theoretical identification of pseudo-genes or operon structures has been reported (see review [14] for a discussion of the difficulties).

As a step toward our final goal of understanding the whole transcription network of archaeobacteria, we have developed an algorithm which is able to predict transcription and translation signals, and have applied it consistently on the DNA sequence that covers approximately 60% of the whole genome of a hyper-thermophilic archaeon, *Pyrococcus* sp. OT3, in order to identify protein genes, operons and pseudo-genes. Since not many biological observations are available, in order to evaluate our identification the biological and physical characteristics of the identified ORFs are compared with those of the rejected ORF candidates.

A very small number, 37, of pseudo-genes have been identified which possess all the characteristics expected of an ORF except for the transcription or translation signal, while most of the ORF candidates which have homologous genes outside the genome have been identified as ORFs, populating 61% of the total of 1111 identified ORFs. Thus, even if we had failed to identify signals for those ORFs identified as pseudo-genes, the success rate still appears to be high.

The strain whose DNA sequence is analyzed in this paper was first deposited in the Japan Collection of Microorganisms (JCM) by Dr. Masuchi (University of Tokyo) with the name *Pyrococcus shinkaii* OT3. The name was wrongly cited as *Pyrococcus shinkai* (note the missing 'i' at the end) in reports [15,16]. After the deposited strain was withdrawn, two new strains isolated from the same source were deposited with the names *Pyrococcus horikoshii* OT3 (JCM9974) and *Pyrococcus horikoshii* JA1 (JCM9975), respectively. The relation of the two strains to *P. shinkaii* OT3 remains unclear. With the above confusion and without a detailed description of *P. shinkaii* or *P. horikoshii* having been published, we felt it better to describe the strain as *Pyrococcus* sp. OT3.

2. Materials and methods

2.1. Nucleotide sequence and databases

The nucleotide sequence analyzed in this study was determined by another group at the National Institute of Technology and Evaluation with the help of our institute, NIBHT (DDBJ accession numbers AB 9464–9506), and has been published (http://www.bio.nite.go.jp/).

The nucleotide numbering scheme used in this study for the whole nucleotide sequence and for each ORF candidate, and details of our analysis are shown in the 23 January 1998 version of our database, ARCHAIC (http://www.aist.go.jp/RIODB/archaic/).

*Corresponding author (a). Fax: (81) (298) 54-6041.

PII S0014-5793(98)00323-8

2.2. Prediction for transcription and translation signals

Archaeobacteria use TATA box binding protein (TBP) and transcription factor B (TFB) for the initiation of transcription [17,18]. For the prediction of the TFB-TBP binding sites, a scoring system was made for all possible combinations of 11 bases. At each position each of the four types of base was given a score between 0 and 100, essentially depending on the frequency of the base at the position in a collection of sequences that were manually identified as putative signal sequences, and that were confirmed as sharing a consensus sequence, being statistically different from a collection of random sequences. Thus, the scores given to each position were independent of each other. The sum of the scores for the 11 positions was used as the score of the sequence, which should be a value between 0 and 1100. The best sequence, AAAAGTTTATA, was given the highest score of 715.9, while the poorest one, GCCCCGCCCCC, received a score of 66.8. Eleven base sequences that were given scores higher than the threshold value, and were followed by possible start codons, ATG, GTG or TTG, with insertions of 21–131 bases were predicted to be transcription signals. The threshold value was increased stepwise so that signals followed by a larger insertion required a higher score: 539 or higher for an insertion of 21–28 bases, 544.5 for 29–45 bases, 577.5 for 46–53 bases, and 594 for 54–131 bases.

Another scoring system was made in order to predict translation initiation signals for all combinations of nine bases in essentially the same way as that made for the transcription signals. The Shine-Dalgarno sequence [17,18], AGGAGGTGA, was given the highest score of 614.5, while the poorest one, CCCCCACCC, received a score of 35.8. Nine base sequences that were given scores of 369 or higher, and were followed by possible start codons with insertions of 2–9 bases, were predicted to be translation signals.

The possibility of the use of CTG and ATT as a start codon has been reported [1–12]. In this study, an ORF, ot0358394, was identified with CTG, and two ORFs, ot0285088 and ot0144789, with ATT. These genes fulfilled all the requirements and had high homology to known ORFs starting with normal start codons. Another ORF, ot0796073, was identified with AAA for the same reason. Transcription of ot0285088 in *Pyrococcus* OT3 has been confirmed experimentally (Kawashima et al., unpublished).

2.3. Analysis of the RYY periodicity in ORF candidates

A scoring system was made in order to evaluate the triplet periodicity [19–21] in ORF candidates. Purine (R) and pyrimidine (Y) bases at the first position in the coding phase were given scores of 67.92 and 32.08, respectively; at the second position, 48.37 and 51.63, respectively, and at the third position, 53.60 and 46.40, respectively. The value averaged through each ORF candidate was used as its RYY score. In this system the highest score is given to RYR instead of RYY. This was necessary to take into consideration that, in general, ORFs possessed more R bases than Y bases.

2.4. Identification of RNA genes

Ribosomal 16S and 23S RNA genes, and the archaeobacterial 7S RNA gene were identified by homology search using FASTA version 3.0 [22] in the GCG package version 9.0 (Genetics Computer Group) to reference RNA gene sequences of other archaeobacteria collected in databases (Ribosomal Database Project [23], GenBank [24], EMBL database [25] and DDBJ [26]). The tRNA genes were identified by using the algorithm tRNA scan-SE [27].

2.5. Homology search for translated amino acid sequences and identification of paralogues

Nucleotide sequences were translated to amino acid sequences using the 'Translate' program in the GCG package. Homology search of the translated protein sequences to known protein sequences was carried out using the program FASTA version 3.0 [22], and the database OWL (s-ind2.dl.ac.ak in pub/database/owl). The criteria chosen for homology identification were: number of amino acid residues in the domain shared by the original and reference protein sequences 50 or larger, ratio of the domain to the whole length of the reference protein 0.7 or higher, and identity inside the domain 25% or higher. The search for genes or pseudo-genes paralogous to each other was carried out essentially in the same way as above, using a collection of sequences of ORF candidates identified in OT3. The length of the shorter of the two gene or pseudo-gene products was used for calculating the ratio of the domain shared by the two.

a

Number of nucleotides analyzed : 1,090,796
(covering approximately 60% the whole genome)
A/T content : 57.8%
Number of ORFs (protein genes) identified : 1,111
Number of RNA genes identified : 25
Number of protein pseudo-genes identified : 37
Number of operons identified : 197
Number of genes identified in operons : 498
Average number of genes / operon : 2.5
Average number of ORFs (protein genes) / 1kb : 0.94
Average number of residues / protein : 269

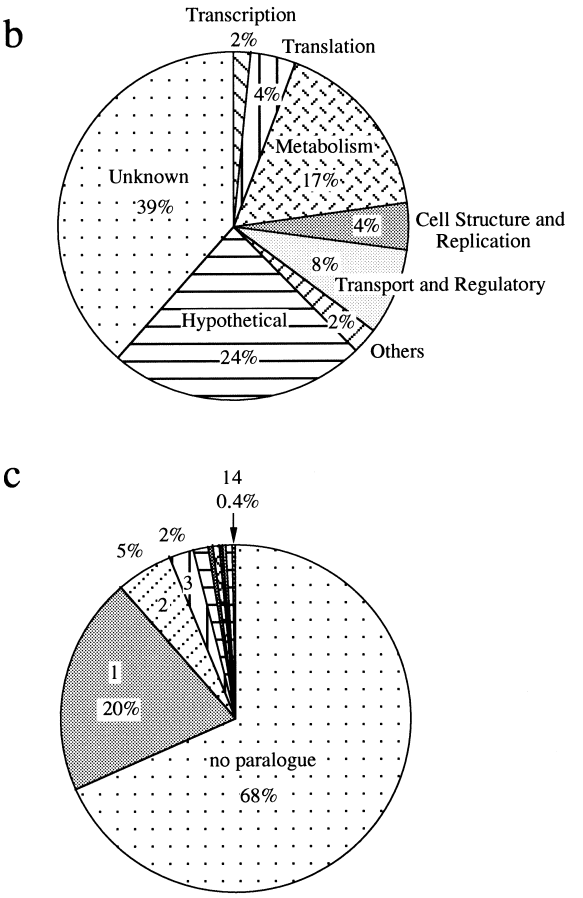

Fig. 1. Summary of the identification. a: Statistics of the identified genes, operons and pseudo-genes. One of the operons, ot867165, is composed of three RNA genes. This operon was identified manually, and not by the algorithm reported. Another operon, ot1009261, has one RNA gene. b: ORFs (protein genes) classified by their functional characteristics. The 'unknown' category is that for proteins which have no significant homology to known proteins, while the 'hypothetical' category is that for proteins which have significant homology only to known proteins whose function is unidentified. c: ORFs (protein genes) classified by the number of genes and pseudo-genes paralogous to them, i.e. 1, 2–14.

3. Results and discussion

3.1. General strategy for the identification of transcription units

A section in the nucleotide sequence which had 150 or more

bases, starting with one of the possible start codon sequences (see Section 2.2) and ending with one of the possible stop codon sequences, TGA, TAG and TAA, was nominated as a candidate for an ORF.

In the first step, ORFs were identified solely on the basis of the prediction of transcription and translation signals for the ORF candidates. The ORF candidates that had both transcription and translation signals were classified as ORFs. When an identified ORF was followed by another ORF candidate which had a translation signal but no transcription signal, and when the first base in the ORF candidate was positioned inside the region covering the base positioned 79 bases upstream of the stop codon end of the preceding ORF to the base positioned 57 bases downstream of the end, the ORF candidate downstream was also classified as an ORF, thus forming an operon. When another ORF candidate was found further downstream and fulfilled what was required as with the second ORF, the operon was extended to include a third ORF. This process was repeated until no such ORF candidates were found further downstream. When an identified ORF was not followed by another ORF candidate which fulfilled the requirements for operon formation, it was classified as an independently regulated gene.

The 1194 ORFs identified in the first step were further examined in terms of the tendency to create triplet periodicity that has been observed with ORFs of many species [19–21] and 83 were rejected. This will be described in detail in Section 3.3. Thus, all the processes of the identification of operons and protein genes (Fig. 1a) were totally independent of the possible biological function of the genes identified. For the identification of 25 RNA genes, a homology search to known RNA genes was carried out. Only the transcription signal was taken into consideration for the identification of RNA genes.

3.2. ORF candidates with signals and with no signal

For the evaluation of the gene identification, 1194 ORF candidates with signals (OCSs), namely the ORFs identified as independently regulated and inside operon structures in the first step, were compared with 2113 ORF candidates with no signal (OCNSs) that were rejected. The OCSs, if translated, would encode proteins of 269.4 amino acid residues on average, the value being much larger than that of OCNSs at 78.7 (Fig. 2a). The OCSs had, in general, higher homology values to the known proteins than OCNSs; the average FASTA z-score value of OCSs was 506.2, while that of OCNSs was 135.7 (Fig. 2b). It is possible, but unlikely, that a section of a

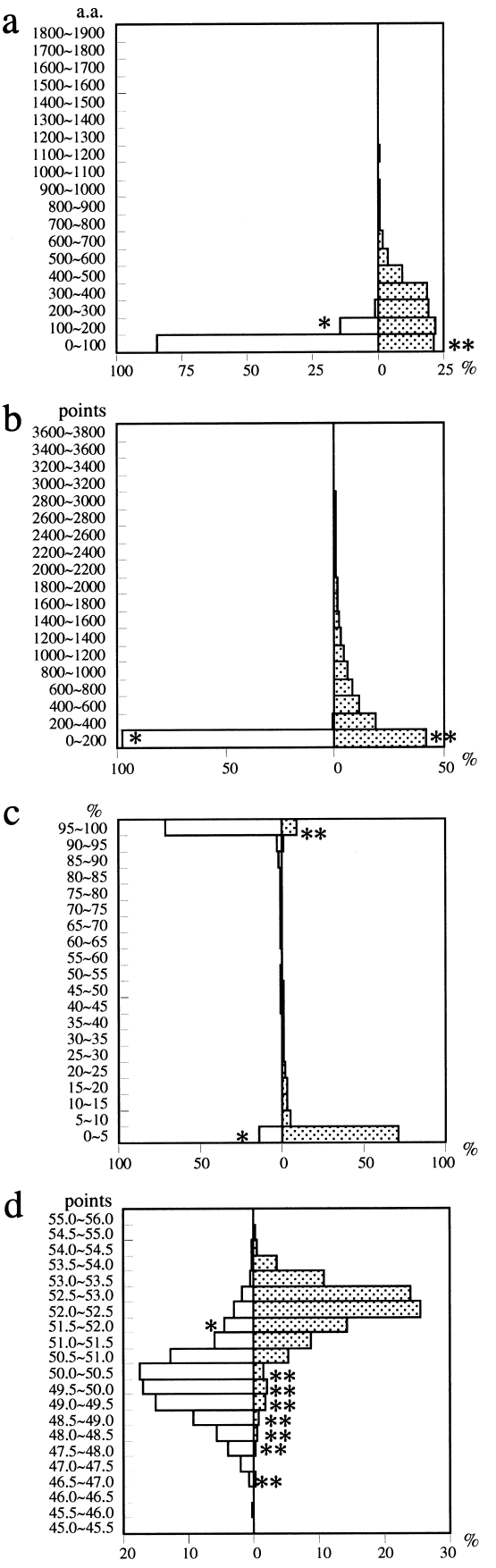

Fig. 2. Differences between ORF candidates with signals (OCSs), shown to the right, and ORF candidates with no signal (OCNSs), shown to the left in amino acid residue numbers, if translated (a), in the highest FASTA z-score homology values to the known protein genes (b), in the percentage of the lengths overlapped by other candidates (c), and in the tendency to form RYY periodicity (d, see Section 2 for the RYY score). In order to compensate the difference in the entry numbers, 1194 of OCSs and 2113 of OCNSs, the abscissae are normalized by the total entry number of each group. A single asterisk indicates the statistical class in the population to which the majority of the pseudo-genes belong (see text). A double asterisk in a–c indicates the statistical class to which the majority of the OCNSs which have RYY score values smaller than 50.5 belong, and in d indicates all the classes that have such an entry.

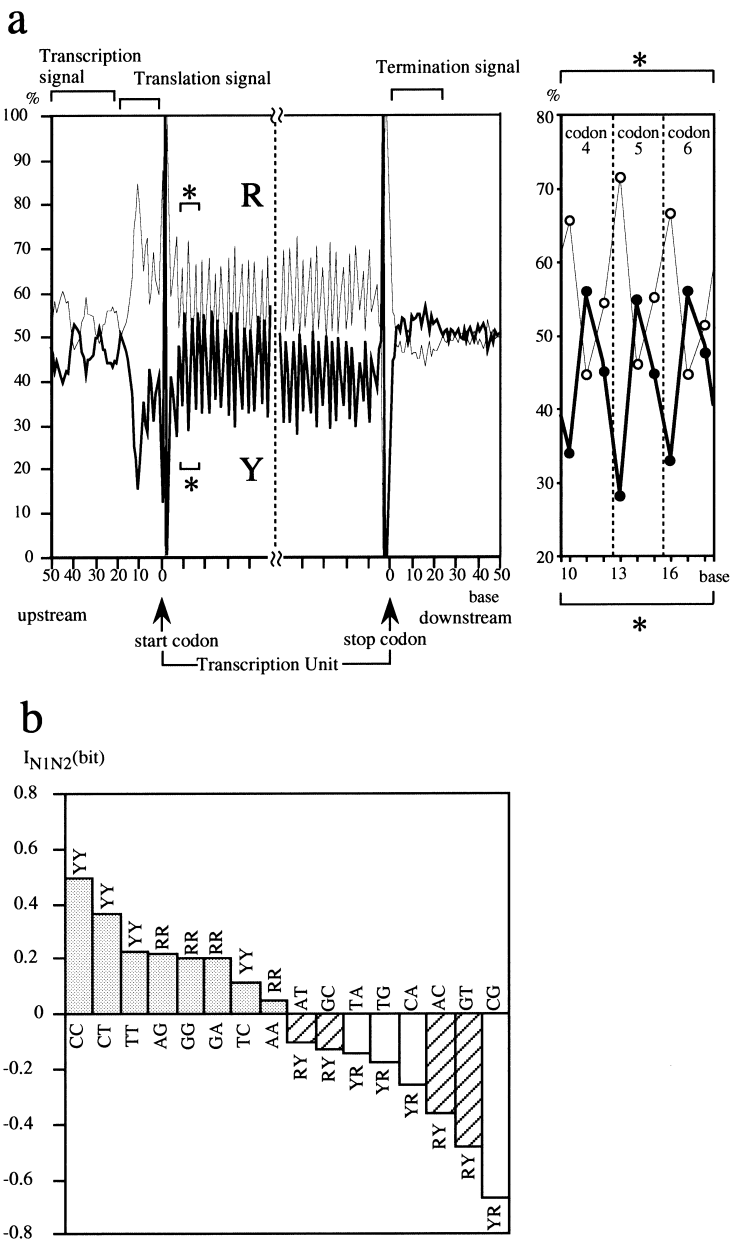

Fig. 3. Arrangement of different types of bases in the ORFs. a: Frequency of pyrimidines (Y, bold line) and purines (R, thinnner line) averaged at each position in reference to the transcription units. An enlargement of the part marked with the asterisk is shown on the right. b: The frequency of a combination of a type N1 base and a type N2 base in the N1N2 sequence evaluated in terms of the negative entropy, I_{N1N2}, defined as $\log_2 F_{N1N2} - \log_2 F_{N1} - \log_2 F_{N2}$ along the lines proposed earlier [32]. Positive values of the index, I_{N1N2}, indicate that the steps are found more frequently than expected from the frequency of the four bases, while negative values are less frequent than expected. Different types of bars are used depending on the Y/R types of the sequences: YY and RR, RY, and YR. Note that all the YY and RR have positive values, while RY and YR have negative values.

large number of nucleotide bases is shared by two or more genes. Most of the OCSs were not overlapped by other ORF candidates to a high degree, the part overlapped by other candidates being 14.8% on average, while OCNSs overlapped to a high degree, 81.2% on average (Fig. 2c).

In summary, OCNSs are small in size, have low homology to the known protein genes, and are found overlapped by other candidates to a high degree, and thus are unlikely candidates for an ORF, while OCSs are better candidates. Thus our identification of ORFs in the first step appears to be successful.

3.3. RYY periodicity in ORFs

The 1194 ORFs identified in the first step had more purine (R) bases (57%) than pyrimidine (Y) bases (43%) on average (Fig. 3a, note that this plot was made by using the 1111 ORFs which passed further examination in the second step instead of using all the 1194 ORFs identified in the first step, but no major change was made to the plot by the exclusion of the 83). The R/Y ratio was found to increase towards the ends of transcription units, and to become the highest around 20 bases at the ends (Fig. 3a). Purine bases were more frequent than pyrimidine bases, but to a lesser extent in the

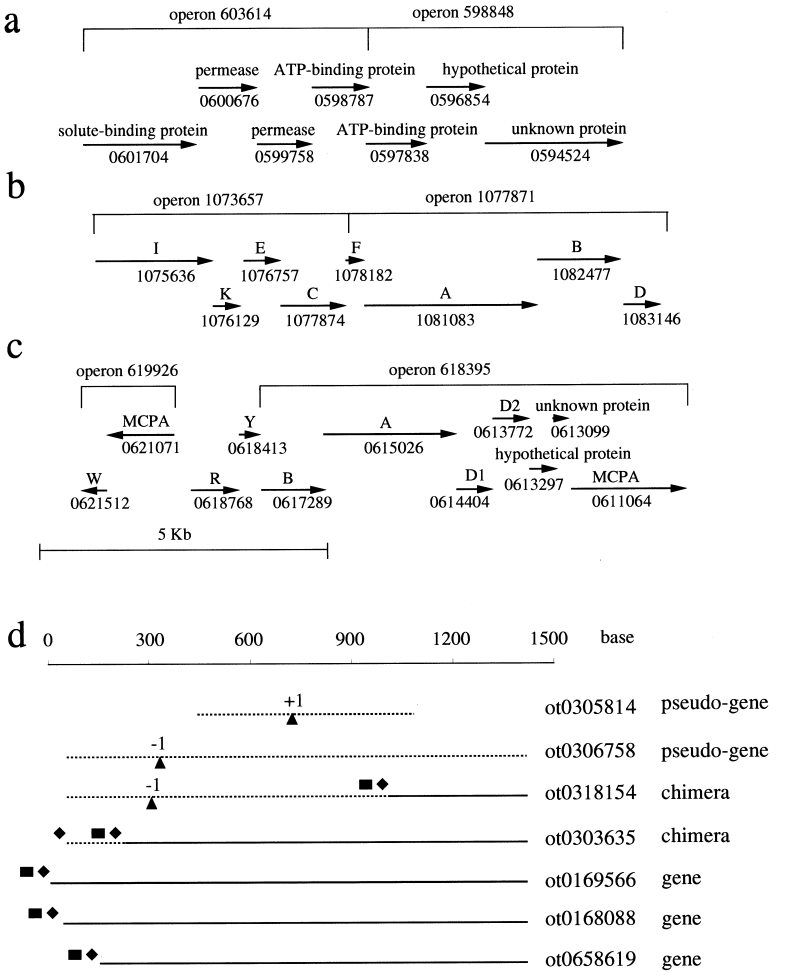

Fig. 4. Examples of identified operons and pseudo-genes. a–c: Operons of ABC transporter genes (a), H^+ transport ATP synthetase genes (b) and chemotaxis-related genes (c). Note that in each subfigure the genes are grouped into two operons. Operons and genes are labelled with the identification codes used in the ARCHAIC database. A scale for the length of 5000 bases is shown. d: A family of genes (indicated by bold lines) and pseudo-genes (indicated by broken lines) paralogous to each other (paralogue family 95 in ARCHAIC). Members ot0318154 and ot0303635 are pseudo-eu chimeras. Symbols are used for indicating the transcription (■) and translation (◆) signals. Numbers +1 and −1 indicate frame-shifts by insertion and deletion, respectively, of one base.

transcription signals and to a larger extent in the translation signals, while pyrimidines were frequent in the termination signals.

The R/Y plot showed distinctive sawtooth-like characteristics reflecting a triplet periodicity inside the transcription units. On average, R was frequent at the first position in the codons, and Y at the second position (Fig. 3a). Third positions were occupied more by R bases, but, when the average R/Y ratio in the ORFs was taken into consideration, it became clear that Y bases were more frequent than expected (data not shown).

A tendency to create triplet periodicity was not observed outside the transcription units, and it was terminated at both ends by other types of triplets. RYR is positioned at the majority of the start codons, and YRR is positioned at all the stop codons. The structural characteristics of DNA are determined mainly by the arrangement of purines and pyrimidines [21,28,29]. Thus, it seems natural to expect that differences in the physical characteristics of transcription units and the other regions are used in some way for the identification of the coding regions by the transcription machinery. Since the translation signals showed a distinctive pattern in terms of the R/Y combination, they might also be used indirectly as signals at the process of transcription.

It has been reported that the triplet periodicity can be explained by the triplet codon table, particular amino acid frequencies in proteins, and particular codon usage for the same amino acid residues [30]. However, it is still possible that the requirement for the RYY periodicity in ORFs affected the formation of the triplet codon table etc. during the process of evolution.

On average, OCSs had a RYY score (see Section 2) of 52.1 ± 1.1, which is significantly higher than the average RYY of OCNSs of 49.9 ± 1.3 (Fig. 2d, note that the difference between the two average values, 2.2, is approximately twice the S.D. values inside the groups). Eighty-three OCSs had RYY scores lower than 50.5, and thus deviated from the average by 1.45 S.D. or more (indicated by double asterisks in Fig. 2d). Seventy-four of them, if translated, would code 100 or smaller numbers of amino acid residues, 81 of them had z-scores smaller than 200, and 51 of them had an overlap ratio higher than 90% (indicated by double asterisks in Fig. 2a–c). These are the same characteristics as those of OCNSs. Thus, it was decided that only the other OCSs remained

identified as ORFs, while these 83 OCSs were rejected. Since RYY periodicity is associated with codon usage, the consideration of the RYY score is equivalent to that of the codon usage for the identification of ORFs, which is routinely carried out.

In the ORFs, the YY and RR steps were found more frequently than expected from the A/T/G/C content, while the YR and RY steps showed a tendency to be avoided (Fig. 3b). Since the conformation of the YY/RR steps is closest to the standard B-conformation [29], the immediate conclusion is that in OT3 deviation of the local DNA conformation from the standard B-conformation is minimized. Similar tendencies, but to lesser degrees, are generally associated with other genomic DNA sequences [21,31]. Thus, the YY/RR regularity is more specific to the OT3 strain in comparison with the RYY periodicity found more generally (Suzuki et al., unpublished). Since the optimum growth temperature of strain OT3 is close to 100°C, the YY/RR regularity might be associated with the possible thermal stability of the genomic DNA molecule in some way.

3.4. Operon structures and pseudo-genes

About a quarter of the identified operons had two or more protein genes whose function was identified by homology search to known proteins. In most of these operons the genes grouped together had functions that were expected to be biologically associated (see some examples shown in Fig. 4a–c). The current system might be under-estimating the region covered by each operon in some cases, since, for example, eight clustering genes for subunits of H^+ transporting ATP synthetase, ot1075636–1083146, were divided into two operons but not assembled into a single operon (Fig. 4b). Alternatively, an operon might not be able to continue for too many bases and such a gene cluster might indeed be assembled into two separate transcription units.

Many of the identified ORFs had paralogous genes inside the genome. In addition, 37 OCNSs that had RYY score values as high as that expected for an ORF were found to be paralogous to some of the ORFs identified in OT3, or homologous to the protein genes known outside the genome (the statistical class in the population to which the majority of the 37 OCNSs belong is indicated by a single asterisk in Fig. 2). The overlap ratio of the majority of the 37 OCNSs was small, and thus they were essentially positioned in unoccupied space. If translated, they would encode numbers of amino acid residues larger than that expected of an average OCNS. Thus, these OCNSs were defined as pseudo-genes, and were separated from the other OCNSs.

Altogether, the identified protein genes and pseudo-genes comprised 150–200 families of paralogues. In a family shown in Fig. 4d, three protein genes, ot0169566, ot0168088 and ot0658619, were associated with two pseudo-genes, ot0305814 and ot0306758, and chimera genes, ot0318154 and ot0303635. The pseudo-genes, ot0305814 and ot0306758, had no translation or transcription signal upstream and had frame-shifts. Thus, it is highly unlikely that these are functional. The chimera genes, ot0318154 and ot030365, were identified as a shorter gene but their upstream regions had clear homology to the longer genes.

In the OT3 sequence, approximately one third of the identified ORFs had at least one paralogue (Fig. 1c); on average, 0.52 paralogous genes or pseudo-genes are expected for each ORF. The number is expected to increase when the rest of the genomic DNA sequence is determined, since an ORF which has no paralogue in the analyzed sequence may have one in the rest of the genome.

The majority, 61%, of the identified ORFs had homology to known protein genes whose function has been either identified or not (Fig. 1b). Thus, it is slightly puzzling that the majority of the pseudo-genes did not show high z-score values (Fig. 2b). Many pseudo-genes are not homologous to genes identified outside the genome but are paralogous to genes inside. Further analysis is necessary for a better understanding of gene duplication followed by mutation of the duplicated gene, which is expected to be one of the major driving forces in the evolution of the genome.

Acknowledgements: We thank Drs. Steven Brenner and Tetsuji Yada for their help at an early stage of the work. This work was supported by the Core Research for Evolutional Science and Technology (CREST) program of the Japan Science and Technology Corporation. J.M.S. is a postdoctoral fellow financially supported by the Human Frontier Science Program (HFSP).

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. and Merrick, J.M. et al. (1995) Science 269, 496–512.
[2] Fraser, C.M., Gocayne, J.D., White, O., Adams, M.D., Clayton, R.A., Fleischmann, R.D., Bult, C.J., Kerlavage, A.R., Sutton, G. and Kelly, J.M. et al. (1995) Science 270, 397–403.
[3] Bult, C.J., White, O., Olsen, G.J., Zhou, L., Fleischmann, R.D., Sutton, G.G., Blake, J.A., FirzGerald, L.M., Clayton, R.A. and Gocayne, J.D. et al. (1996) Science 273, 1058–1073.
[4] Kaneko, T., Sato, S., Kotani, H., Tanaka, A., Asamizu, E., Nakamura, Y., Miyajima, N., Hirosawa, M., Sugiura, M. and Sasamoto, S. et al. (1996) DNA Res. 3, 109–136.
[5] Himmerlreich, R., Hilbert, H., Rlagens, H., Pirki, E., Li, B.-C. and Herrmann, R. (1996) Nucleic Acids Res. 24, 4420–4449.
[6] Blattner, F.R., Plunkett III, G., Bloch, C.A., Perna, N.T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J.D., Robe, C.K. and Mayhew, G.F. et al. (1997) Science 277, 1453–1462.
[7] Klenk, H.-P., Clayton, R.A., Tomb, J.-F., White, O., Nelson, K.E., Ketchum, K.A., Dodson, R.J., Gwinn, M., Hickey, E.K. and Peterson, J.D. et al. (1997) Nature 390, 364–370.
[8] Smith, D.R., Doucette-Stamm, L.A., Deloughery, C., Lee, H., Dubois, J., Aldredge, T., Bashirzadeh, R., Blakely, D., Cook, R. and Gilbert, K. et al. (1997) J. Bacteriol. 179, 7135–7155.
[9] Tomb, J.-F., White, O., Kerlavage, A.R., Clayton, R.A., Sutton, G.G., Fleischmann, R.D., Ketchum, K.A., Klenk, H.P., Gill, S. and Dougherty, B.A. et al. (1997) Nature 388, 539–547.
[10] Kunst, F., Ogasawara, N., Moszer, I., Albertini, A.M., Alloni, G., Azevedo, V., Bertero, M.G., Bessières, P., Bolotin, A. and Borchert, S. et al. (1997) Nature 390, 249–256.
[11] Goffeau, A., Aert, R., Agostini-Carbone, M.L., Ahmed, A., Aigle, M., Alberghina, L., Albermann, K., Albers, M., Aldea, M. and Alexandraki, D. et al. (1997) Nature 387, (Suppl.) 5–105.
[12] Fraser, C.M., Casjens, S., Huang, W.M., Sutton, G.G., Clayton, R., Lathigra, R., White, O., Ketchum, K.A., Dodson, R. and Hickey, E.K. et al. (1997) Nature 390, 580–586.
[13] Olsen, G.T. and Woese, C.R. (1997) Cell 89, 991–994.
[14] Ficket, J.W. (1996) Trends Genet. 12, 316–320.
[15] Swinbanks, D. (1995) Nature 374, 583–583.
[16] Barker, S. (1996) Nature 381, 455–455.
[17] Brown, J.W., Daniels, C.J. and Reeve, J.N. (1989) CRC Crit. Rev. Microbiol. 16, 287–337.
[18] Zilling, W., Palm, P., Reiter, W.-D., Gropp, F., Pühler, G. and Klenk, H.-P. (1988) Eur J. Biochem. 173, 473–482.
[19] Fickett, J.W. (1982) Nucleic Acids Res. 10, 5303–5318.
[20] Tsonis, A.A., Elsnor, J.B. and Tsonis, P.A. (1991) J. Theor. Biol. 151, 323–331.

[21] Amano, N., Ohfuku, Y. and Suzuki, M. (1997) Biol. Chem. 378, 1397–1404.
[22] Pearson, W.R. and Lipman, D.J. (1988) Proc. Natl. Acad. Sci. USA 85, 2444–2448.
[23] Maidak, B.L., Olsen, G.J., Larsen, N., Overbeek, R., McCaughey, M.J. and Woese, C.R. (1997) Nucleic Acids Res. 25, 109–111.
[24] Benson, D.A., Boguski, M.S., Lipman, D.J. and Ostell, J. (1997) Nucleic Acids Res. 25, 1–6.
[25] Stoesser, G., Sterk, P., Tuli, M.A., Stoehr, P.J. and Cameron, G.N. (1997) Nucleic Acids Res. 25, 7–13.
[26] Tateno, Y. and Gojobori, T. (1997) Nucleic Acids Res. 25, 14–17.
[27] Lowe, T.M. and Eddy, S.R. (1997) Nucleic Acids Res. 25, 955–964.
[28] Suzuki, M. and Yagi, N. (1995) Nucleic Acids Res. 23, 2083–2091.
[29] Suzuki, M., Amano, N., Kakinuma, J. and Tateno, M. (1997) J. Mol. Biol. 274, 421–435.
[30] Staden, R. (1984) Nucleic Acids Res. 12, 551–567.
[31] Nussinov, R. (1984) J. Mol. Evol. 20, 111–119.
[32] Gatlin, L.L. (1968) J. Theor. Biol. 18, 181–194.