ELSEVIER

Gene 294 (2002) 25–33

GENE
AN INTERNATIONAL JOURNAL ON GENES AND GENOMES

www.elsevier.com/locate/gene

A systematic investigation identifies a significant number of probable pseudogenes in the *Escherichia coli* genome

Keiichi Homma[a,b], Satoshi Fukuchi[a,b], Takeshi Kawabata[a,1], Motonori Ota[a,2], Ken Nishikawa[a,*]

[a]*Laboratory of Gene-Product Informatics, Center for Information Biology–DNA Data Bank of Japan, National Institute of Genetics,*
1111 Yata, Mishima, Shizuoka, 411-8540, Japan
[b]*Japan Science and Technology Corporation, 1-8, Honcho 4-chome, Kawaguchi City, Saitama, 332-0012, Japan*

Received 13 February 2002; received in revised form 31 May 2002; accepted 11 June 2002
Received by T. Gojobori

Abstract

Pseudogenes are open reading frames (ORFs) encoding dysfunctional proteins with high homology to known protein-coding genes. Although pseudogenes were reported to exist in the genomes of many eukaryotes and bacteria, no systematic search for pseudogenes in the *Escherichia coli* genome has been carried out. Genome comparisons of *E. coli* strains K-12 and O157 revealed that many protein-coding sequences have prematurely terminated orthologs encoding unstable proteins. To systematically screen for pseudogenes, we selected ORFs generated by premature termination of the orthologous protein-coding genes and subsequently excluded those possibly arising from sequence errors. Lastly we eliminated those with close homologs in this and other species, as these shortened ORFs may actually have functions. The process produced 95 and 101 pseudogene candidates in K-12 and O157, respectively. The assigned three-dimensional structures suggest that most of the encoded proteins cannot fold properly and thus are dysfunctional, indicating that they are probably pseudogenes. Therefore, the existence of a significant number of probable pseudogenes in *E. coli* is predicted, awaiting experimental verification. Most of them were found to be genes with paralogs or horizontally transferred genes or both. We suggest that pseudogenes constitute a small fraction of the genomes of free-living bacteria in general, reflecting the faster elimination than production of pseudogenes. © 2002 Elsevier Science B.V. All rights reserved.

Keywords: Three-dimensional structure; Structure prediction; Gram-negative bacteria; Position-specific iterated basic local alignment search tool; Horizontal transfer

1. Introduction

In recent years much attention has been paid to pseudogenes. Pseudogenes account for a significant proportion of eukaryotic genomes (Gonçalves et al., 2000; Harrison et al., 2001), and occupy 27% of the genome of the obligate parasitic bacterium *Mycobacterium leprae* (Cole et al., 2001). A number of pseudogenes were found not only in the genomes of other obligate parasites (Casjens et al., 2000; Andersson et al., 1998; Ogata et al., 2001), but also in those of free-living proteobacteria, including *Salmonella enterica*, which is phylogenetically very close to *Escherichia coli* (Parkhill et al., 2000a,b, 2001a,b). These investigations suggest that *E. coli* may also have pseudogenes. Indeed, some truncated genes have already been reported in this bacterium (e.g. Brinkkötter et al., 2000). However, no methodical exploration of pseudogenes in this model organism has been undertaken thus far.

The processed pseudogenes found in mammalian genomes are characterized by the presence of direct repeats immediately 5′ and 3′ to the pseudogenes, the absence of intervening sequences, and the presence of a poly A tract at the 3′ end (Vanin, 1985). The absence of such features in other pseudogenes makes it crucial to verify the lack of functions of the products. Papers reporting systematic inquiries of pseudogenes in bacteria use the existence of one or more mutations that would ablate expression as the major criterion for identifying pseudogenes. The difficulty of experimentally proving the lack of function of a protein

---

Abbreviations: BLAST, basic local alignment search tool; ORF, open reading frame; PDB, Protein Data Bank; PSI-BLAST, position-specific iterated BLAST

* Corresponding author. Tel.: +81-559-81-6859; fax: +81-559-81-6889.
E-mail address: knishika@genes.nig.ac.jp (K. Nishikawa).

[1] Present address: Graduate School of Information Science, Nara Institute of Science and Technology, 8916-5 Takayama, Ikoma, Nara, 630-0101, Japan.

[2] Present address: Global Scientific Information and Computing Center, Tokyo Institute of Technology, O-okayama, Meguro, Tokyo, 152-8550, Japan.

0378-1119/02/$ - see front matter © 2002 Elsevier Science B.V. All rights reserved.
PII: S0378-1119(02)00794-1

makes it desirable to devise computational methods to clearly identify pseudogenes.

Through a systematic investigation of the whole genome sequences of the *E. coli* strains K-12 and O157, which diverged around 4.5 million years ago (Reid et al., 2000), we identified a considerable number of pseudogene candidates in both strains. This kind of methodical research can uncover probable pseudogenes that sequence information alone cannot identify, and thus will assist experimental efforts to identify pseudogenes.

2. Materials and methods

We used the K-12 sequences of MG1655 with 4289 ORFs (Blattner et al., 1997) and W3110 (Itoh et al., 1996; Yamamoto et al., 1997; http://ecoli.aist-nara.ac.jp/), 99.976% of which agree after the exclusion of IS elements and prophage-related sequences. Additional W3110 data were obtained by personal communication with T. Horiuchi. We also made use of the O157:H7 sequences of EDL933 (Perna et al., 2001) and Sakai (Hayashi et al., 2001), whose agreement is estimated at 99.972%, as above. (We consider two sequences to be in agreement at a position if the two aligned nucleotides are consistent within ambiguities, e.g. we regard that one sequence with G agrees with another with S at the position, because S stands for either G or C.) For the initial screening of pseudogene candidates, we used the K-12 sequence of MG1655 and the O157:H7 sequence of EDL933. The GTOP database (Kawabata et al., 2002; http://spock.genes.nig.ac.jp/~genome/) mainly uses PSI-BLAST (Altschul et al., 1997; Kawabata et al., 2000) to predict proteins' three-dimensional structures, because this program detects weaker sequence homologies than BLAST does (Altschul et al., 1997). In the GTOP version of February 2002, BLAST predicted the structures of 31% of the *E. coli* ORF products, while PSI-BLAST assigned structures to 46%. By individual examination we deemed a protein segment stable if its alignment covered most of a SCOP domain (Murzin et al., 1995) and otherwise classified it as unstable. In general, all of the aligned segments encompassing more than 75% of the SCOP domains were appraised as stable, while those covering less than 65% were categorized as unstable without exception. A set of proteins whose ortholog consists of multiple SCOP domains was judged to be stable only if all of the domains were present and stable. The same procedure with CATH domains (Orengo et al., 1997) produced identical results.

The ortholog of a K-12 gene (as identified by Blattner et al., 1997; Perna et al., 2001) is defined to be the homologous O157 ORF with synteny having sequence identity over 80% in the encoded amino acid residues, and the opposite case is similarly determined. (In actuality the identity between orthologs exceeded 95% in over 80% of the cases examined.) Synteny is defined to exist between genes A and B of different strains if there are two or more ORFs in the immediate vicinity of gene A that have homology to ORFs in the neighborhood of gene B. Paralogs and homologs were searched by BLAST in the SWISS-PROT database (release 39.24, http://www.expasy.ch/sprot/) with the cut-off *E*-value set at 10^{-3}. Homologs of comparable lengths satisfy all of the following: (i) they are not annotated as hypothetical, putative, or fragmental in the SWISS-PROT database; (ii) their lengths differ by less than 50 amino acid residues from that of the query protein; and (iii) after alignment, their N- and C-termini fall within 50 amino acid residues of the query. We introduced the last condition to eliminate homologs of similar lengths whose homologous regions relative to the N-termini of the proteins significantly differ from that of the query protein, as happens when a homologous protein consists of domains A and B and the query is composed of domains B and C, for example.

3. Results and discussion

Comparing the two genomes, we noticed that many genes encoding stable proteins have shorter orthologs whose products are apparently unstable. The example presented in Fig. 1A shows that *Z4353* encodes a stable protein in O157, while the proteins encoded by the two contiguous orthologs in K-12 become stable only when combined. The *Z4353* product is aligned to nearly the entire structure of the SCOP domain c.69.1.9, while the two K-12 proteins are aligned to two fragments of the same domain. (We refer to groups like the two K-12 genes as 'sets' in the following.) The SCOP domain is the evolutionarily conserved unit of protein structure (Murzin et al., 1995) and can be considered as a folding unit, as any significant deletions of a SCOP domain render it unfoldable and consequently dysfunctional (Poon et al., 1991). Therefore, even if these ORFs are expressed, the products are dysfunctional. In addition, neither of the products of the K-12 genes has annotated homologs of comparable lengths in the SWISS-PROT database, while that of the O157 ortholog has some. Most functional genes have some homologs of similar lengths in other genomes, both because mutational events that produce functional genes often occur independently in different species and because functional genes tend to be evolutionarily conserved, while pseudogenes are not. These observations strongly suggest that the K-12 genes are without function and therefore are probably pseudogenes. We subsequently undertook a systematic screening for probable pseudogenes in the two genomes by first searching for orthologs of unequal lengths.

3.1. Systematic search for pseudogenes in E. coli

Using the GTOP database (Kawabata et al., 2002) constructed by our group, in which the results of extensive data analyses of all of the proteins in 70 species are currently presented, we identified orthologs in the genomes of *E. coli* strains K-12 and O157. We then selected those with muta-

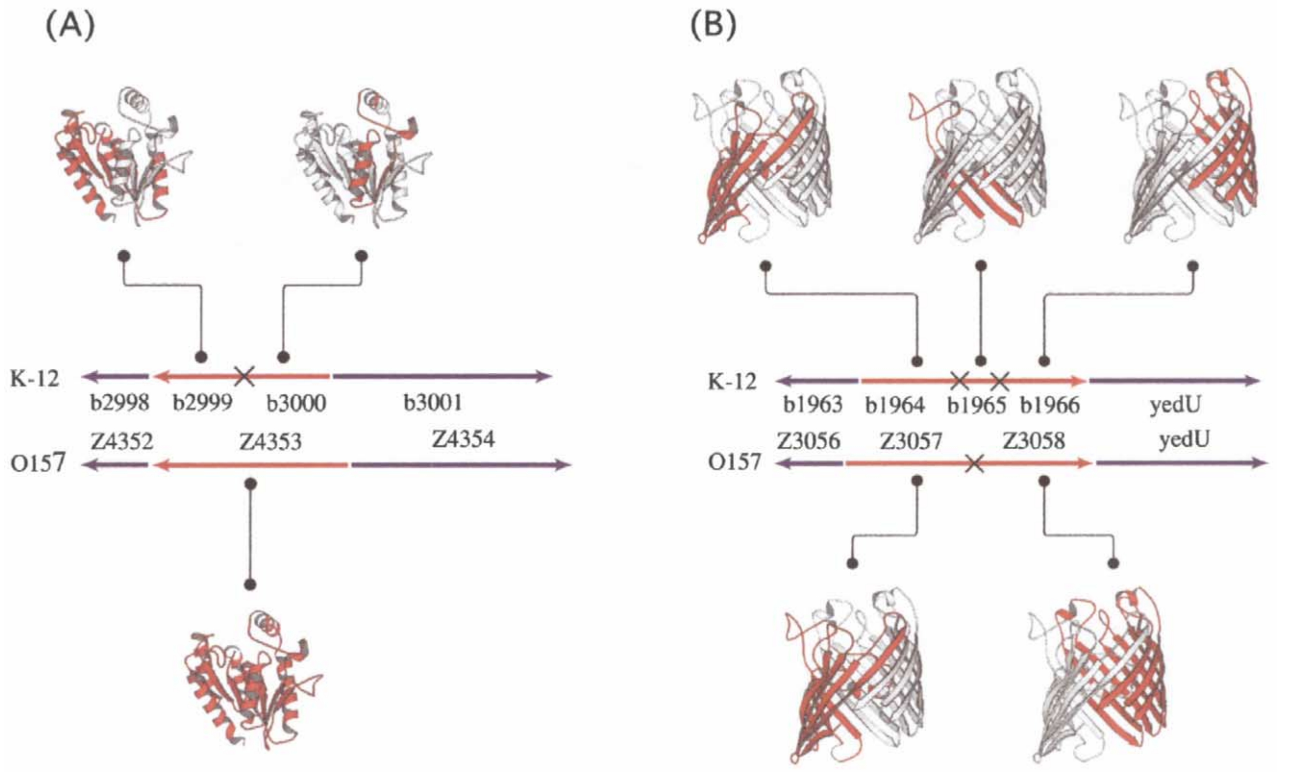

Fig. 1. Three-dimensional structures predicted by homology of proteins encoded by likely pseudogenes and their orthologs, and a schematic representation of the ORFs in the vicinity. The structures to which products of orthologous genes are aligned are shown, with the aligned sections colored red. Arrows in the middle represent some ORFs in the neighborhood. (A) The proteins encoded by K-12:*b3000* and *b2999* respectively have 98 and 100% amino acid identity with the O157:*Z4353* product, and all of them are aligned to 1din (dienelactone hydrolase) in the Protein Data Bank (PDB). The neighboring ORFs of the two strains displayed are also orthologous to each other. (B) *b1964* and *b1966* in K-12 encode proteins that are 91 and 86% identical to the *Z3057* and *Z3058* products in O157, respectively. All of the red-shaded proteins are aligned to the same structure, namely that of porin (PDB ID: 1osm). Both the upstream and downstream ORFs in K-12 shown are also orthologs of the corresponding ORFs in O157.

tions that result in shortening of more than 50 amino acid residues from their orthologs in O157 and considered them as initial pseudogene candidates. To the group we added ORFs contiguous to the candidate ORFs corresponding to the remaining regions of the O157 orthologs, e.g. *b2650*, which appears in the final list of K-12 pseudogene candidates (row 37, Table 1). Sometimes two or three ORFs in K-12 were found to be orthologous to two ORFs of different lengths in O157 (for example, see row 11 in Table 1). In such cases, all of them were considered to be pseudogene candidates.

To exclude those pseudogene candidates arising from DNA sequence errors, we exploited the fortunate availability of two independently determined sequences for most of the K-12 genome (Blattner et al., 1997; Itoh et al., 1996; Yamamoto et al., 1997) and for the entire O157 genome (Perna et al., 2001; Hayashi et al., 2001). From the list of 112 K-12 ORFs, we dropped five whose independently determined sequences of the same strain disagreed at sites critical for causing length differences with the O157 orthologs. Although some of the discrepancies may be attributable to real differences between isolates of each strain, we decided to err on the side of caution and excluded all of those as ‘possibly’ arising from sequence errors. In this and the following step, we discarded the entire set of ORFs if one of the ORFs was judged to be possibly functional, because the remaining members have a high chance of being functional. Next, we removed 12 that have annotated homologs of comparable lengths in SWISS-PROT. This is an essential step to screen out shortened but functional genes, as explained above. We thus obtained 95 ORFs as pseudogene candidates in K-12 (Table 1). The identical procedure with the two strains reversed yielded 101 ORFs in O157 as pseudogene candidates (Table 2), after the elimination of 13 and 23 ORFs in the first and the second screening steps from an initial list of 137 ORFs. The final lists include several sets of split genes that are considered to produce fused proteins by the SWISS-PROT database, e.g. *yigW_1* and *yigW_2*; *molR_1*, *molR_2*, and *molR_3*; *yifM_1* and *yifM_2* (respectively in rows 21, 33, and 52 in Table 1). We consider it likely that most of these are pseudogenes, because the two independently determined sequences are in agreement in these regions and because our further analysis lends credence to the idea (see Section 3.2).

3.2. Characteristics of the pseudogene candidates

We pinpointed the mutations in the pseudogene candi-

Table 1
Pseudogene candidates in *E. coli* K-12

	K-12 ORF(s)	Disruption(s)[a]	Structure(s)[b]	O157 ortholog(s) (amino acid residues)
1	*eaeH*[c]	C-del	− , S	*eaeH* (1417)
2	*yeeP*	N-del	− , S	*Z1650* (290)
3	*yfaH*	N-del	− , S	*Z3495* (292)
4	*yjbI*[c]	− 1, +1	− , S	*yjbI* (526)
5	*ykgA*[c]	C-del	$\underline{S}$, S	*ykgA* (296)
6	*yaiT*[c], *yaiU*[c]	+ IS	$\underline{S}$, S	*Z0469* (980)
7	*ybcY*[c]	− 2	$\underline{S}$, S	*Z1925* (222)
8	*b1369*	N-del, +IS	$\underline{S}$, S	*Z3310* (199)
9	*gapC_1*, *gapC_2*	1T, −1	$\underline{S}$, S	*gapC* (333)
10	*b1472*	C-del	$\underline{S}$, S	*Z2239* (366)
11	*b1964*, *b1965*, *b1966*	1T, +1, −10, Ins, −9	$\underline{S}$, S	*Z3057* (224), *Z3058* (191)
12	*b1995*, *b1998*	+ IS, C-del	$\underline{S}$, S	*Z3159* (719)
13	*b2654*, *b2655*[d], *b2656*[d]	1T, +1, 1T, −1T	$\underline{S}$, S	*Z3954* (412)
14	*b2657*, *b2658*	1T	$\underline{S}$, S	*Z3955* (335)
15	*b3000*, *b2999*	− 2, −1	$\underline{S}$, S	*Z4353* (308)
16	*agaA*[c]	N-del	$\underline{S}$, S	*Z4489* (377)
17	*yhfO*[c], *yhfP*[c]	+ 1, −1T	$\underline{S}$, S	*Z4733* (275)
18	*glvG*[c]	M-del	$\underline{S}$, S	*Z5177* (440)
19	*ilvG_1*[c]	− 2	$\underline{S}$, S	*ilvG* (548)
20	*yigL*[c]	3 × (−2), 6 × (−1), −4	$\underline{S}$, S	*Z5347* (305)
21	*yigW_1*[c], *yigW_2*[c]	+ 1	$\underline{S}$, S	*tatD* (264)
22	*b1720*, *b1721*	+ 1	*S*, S	*Z2749* (632)
23	*b2228*, *b2227*	1T	S, S	*Z3481* (1534)
24	*glvC*[c], *glvB*[c]	1T	S, S	*Z5178* (540)
25	*b0263*[c]	N-del	− , −	*afuB* (692)
26	*b0499*[c]	N-del	− , −	*Z0651* (1398)
27	*ybfD*	1T, M-del	− , −	*Z0700* (285)
28	*b1016*, *b1017*	− 1	− , −	*Z1519* (279)
29	*ydaW*	C-del	− , −	*Z2395* (270)
30	*b1458*, *b1459*	+ 1, 1T, −5, C-del	− , −	*Z0275* (390)
31	*b1568*	C-del	− , −	*Z2047* (183)
32	*b1980*, *b1979*	1T	− , −	*Z3137* (350)
33	*molR_1*, *molR_2*, *molR_3*	Ins, M-del, Ins, +1	− , −	*molR_C* (426), *molR_D* (579)
34	*yehQ*	− 3, 1T, C-del	− , −	*yehQ* (745)
35	*yfcU*, *b2337*	+ 3, 1T, −1T	− , −	*Z3600* (879)
36	*b2353*	N-del	− , −	*Z0316* (197)
37	*b2650*, *b2649*	1T, −1T	− , −	*Z3950* (449)
38	*b2680*, *b2681*	− 7	− , −	*Z3982* (394)
39	*b2863*, *b2862*	+ 5, C-del	− , −	*Z4201* (201), *Z4200* (342)
40	*ygfQ*, *ygfR*	− 1	− , −	*Z4223* (455)
41	*ygiR*, *b3015*	− 1	− , −	*Z4370* (739)
42	*yhaN*[c], *yhaM*[c]	+ 1	− , −	*Z4462* (436)
43	*agaW*[c]	C-del	− , −	*Z4486* (259)
44	*yhdR*[c], *yhdP*[c]	+ 1	− , −	*Z4604* (1266)
45	*gntU_1*[c], *gntU_2*[c]	− 1	− , −	*gntU* (446)
46	*yhiL*[c], *yhiK*[c]	− S, −1, −3	− , −	*Z4888* (557)
47	*yhiS*[c]	C-del	− , −	*Z4907* (371)
48	*yibJ*[c]	1T	− , −	*yibJ* (312)
49	*yidX*[c]	Ins, +1	− , −	*Z5187* (325)
50	*kup*[c]	+ 1, −1, −1	− , −	*kup* (622)
51	*yifN*[c], *b3776*[c]	− 1	− , −	*Z5287* (163)
52	*yifM_1*[c], *yifM_2*[c]	+ 1	− , −	*Z5304* (359)
53	*yigE*[c], *b3814*[c]	− 1	− , −	*Z5332* (254)
54	*trkH*[c]	− 1	− , −	*trkH* (483)
55	*b3913*[c], *b3914*[c]	− 1	− , −	*yiiO* (167)
56	*yjbL*[c], *yjbM*[c]	Ins, −1	− , −	*Z5646*
57	*phnE*[c], *b4103*[c]	+ 8	− , −	*phnE* (276)
58	*ytfI*[c]	− 1	− , −	*Z5826* (312)
59	*yjiP*[c], *yjiQ*[c]	1T, +5, +6	− , −	*Z5940* (260)

[a] N-del, C-del, and M-del respectively denote deletions of the N- and C-terminus and a middle section exceeding 10 bp, while $+n$ and $-n$ indicate an insertion and a deletion of n base pairs, respectively. 1T and −1T respectively signify the creation and loss of a stop codon by base substitution, whereas −S

dates that disrupt the coding sequences (Table 1 for K-12 and Table 2 for O157), an example of which is presented in Fig. 2. Some genes received base substitutions that created stop codons in the middle, while others got deletions or insertions that altered the reading frames, leading to premature termination. Many genes were multiply disrupted (e.g. row 20 in Table 1), like the pseudogenes in *M. leprae* (Cole et al., 2001), and deletions occurred more frequently than insertions (113 vs. 39). In some instances, a base substitution destroyed the termination codon, resulting in an extension of the ORF until the next stop codon was encountered (for example, see row 17 in Table 1). DNA sequence errors that escaped detection even by the comparison of two available sequence sets are very unlikely to account for the many cases with large deletions or insertions or with multiple disruptions. Admittedly, there are mechanisms that produce a continuous protein molecule from divided ORFs: UGA read-through (Engelberg-Kulka and Schoulaker-Schwarz, 1996), suppression of a frameshift mutation (Ibid.), and RNA editing. However, the first two mechanisms are rare in bacteria and are even more unlikely to operate in cases where multiple disruptions exist, while the last one has not been reported in bacteria. These mechanisms are thus unlikely to be operative in many of the pseudogene candidates.

Pseudogenes by definition are dysfunctional and therefore cannot be essential genes. Thus, we checked the final pseudogene candidates in K-12 against the lists of essential genes in the databases of the *E. coli* Genome Project at the University of Wisconsin-Madison (http://www.genome.wisc.edu) and of Profiling of *E. coli* Chromosome (http://www.shigen.nig.ac.jp/ecoli/pec). As of May 2002, none of them were classified as essential genes, although as much as 5.3% of the genes in the whole genome were categorized as such. Moreover, 82% of the final pseudogene candidates in K-12 are hypothetical genes (i.e. those with names starting with either y or b), while merely 53% of the genes were so codified in the whole genome. Additionally, the same holds true for the K-12 orthologs of the O157 pseudogene candidates: none of them falls in the essential category and as much as 77% of them are hypothetical genes. These findings suggest that neither the pseudogene candidates nor their orthologs play important roles in *E. coli* cells.

We then consulted GTOP again to find the prediction of the three-dimensional structure by the homology of the final pseudogene candidates. An assigned three-dimensional structure is absent in six sets of pseudogene candidates in K-12 and O157, while some structure is assigned to the orthologs (rows with ' − , S' in Tables 1 and 2). Furthermore, out of the 45 sets of pseudogene candidates with assigned structures (denoted 'S, S'), 41 were judged to be unstable (the first character underlined in the same column), while one (the first character italicized, row 22 in Table 1) was unclassifiable. Examples of such unstable structures together with the arrangements of the corresponding ORFs on the chromosomes can be found in Fig. 1. In the case presented in Fig. 1B, a combination of three K-12 ORFs and another of two O157 genes both encode a porin-like structure, while all of the individual ORF products are evidently unstable. The *agaA* gene in K-12 (row 16, Table 1) offers another illuminating example: the product of its O157 ortholog, *Z4489*, is almost entirely aligned to urease (PDB ID: 4ubp) by PSI-BLAST, while the *agaA* product is aligned to a small fragment of the same structure and appears unstable. Moreover, the *agaA* gene has no homologs of comparable lengths in other species. These observations strongly suggest that the *agaA* gene is dysfunctional, in agreement with the conclusion reached by an experimental investigation (Brinkkötter et al., 2000).

In sum, 47 out of a total of 51 sets of genes whose products can be structurally studied were deemed to encode unstable proteins and therefore are probable pseudogenes (rows 1–21 in Table 1 and rows 1–26 in Table 2). At present, the three-dimensional structures cannot be assigned to the products of the rest of the candidates, i.e. those in rows 25–59 and rows 28–64 in Tables 1 and 2, respectively, because the homologous structures are accidentally missing in the currently available data. When the accumulation of more structural data results in the assignment of three-dimensional structures, they are likely to show the same tendency as those with structural assignments at present. That is, most of them will encode unstable proteins and thus are pseudogenes. Therefore, slightly less than 100 pseudogenes are likely to be present in each strain of *E. coli*.

3.3. Density of pseudogenes

Although most of the final candidates are probably pseudogenes, the number of pseudogenes in each strain is presumably an underestimate. For example, pseudogenes with orthologs that were lost in the other strain and those with substitutions leading to functional impairment could not be detected by the procedure described in Section 3.1. It is therefore conceivable that somewhat more than 100 pseudogenes exist in each of the two strains of *E. coli.*

---

represents a start codon loss by the same mechanism. +IS and Ins stand for the introduction of an insertion sequence element and some other sequence over 10 bp in the middle, respectively.

[b] The first character in the structure column represents the presence (S) or absence (−) of an assigned three-dimensional structure in the pseudogene or the set of pseudogenes, while the second character shows the same in the ortholog(s). The underlined structures were judged to be unstable, while the italicized one could not be assessed.

[c] Only the K-12:MG1655 sequence is available for these ORFs. A few of these ORFs may be eliminated after the completion of the W3110 sequence.

[d] According to the sequence of K-12:W3110 (Yamamoto et al., 1997), one ORF encompassing b2655 and b2656 exists, as the stop codon of b2655 is absent, resulting in two K-12 ORFs corresponding to Z3954.

Table 2
Pseudogene candidates in *E. coli* O157

	O157 ORF(s)	Disruption(s)[a]	Structure(s)[b]	K-12 ortholog(s) (amino acid residues)
1	*Z2291*	1T, 1T	−, S	*rimL* (179)
2	*Z4950*, *Z4949*	1T	−, S	*yhjQ* (242)
3	*Z0666*, *Z0667*	1T	S, S	*ybbX* (453)
4	*Z1046*, *Z1045*	−1	S, S	*ybiW* (810)
5	*Z1051*, *Z1052*	1T	S, S	*ybiK* (321)
6	*ycbF*	−1	S, S	*ycbF* (245)
7	*Z1352*	C-del	S, S	*b1554* (177)
8	*Z1972*	−3, 1T, −1, C-del	S, S	*b1202* (955)
9	*Z2071*	−2, −3, C-del	S, S	*b1554* (177)
10	*Z2084*	N-del	S, S	*b1579* (398)
11	*Z2251*	1T	S, S	*b1462* (205)
12	*Z2333*, *Z2334*	M-del	S, S	*b1377* (377)
13	*Z2474*, *Z2473*	1T	S, S	*b1310* (430)
14	*Z2476*, *ycjM*	−1, −1	S, S	*ycjM* (568)
15	*intP_2*	N-del, C-del	S, S	*b1579* (398)
16	*uidA_2*, *uidA_1*	+2	S, S	*uidA* (603)
17	*Z3057*, *Z3058*	−1T, −1, +10, M-del, +9	S, S	*b1964* (171), *b1965* (69), *b1966* (134)
18	*Z3531*	M-del, −1, −1	S, S	*yfbL* (325)
19	*Z4080*, *Z4079*	1T	S, S	*ygcQ* (297)
20	*Z4104*, *ygcY*	1T	S, S	*ybcY* (446)
21	*Z4258*	1T	S, S	*ygfI* (303)
22	*agaI_1*, *agaI_2*	1T	S, S	*agaI*[c] (251)
23	*Z4501*, *Z4504*	−1, −2, −3, M-del, +IS, −1, −4	S, S	*yraK*[c](363)
24	*Z4734*, *Z4735*	1T	S, S	*yhfQ*[c] (261)
25	*Z4945*, *yhjL*	+1, −10	S, S	*yhjL*[c] (1166)
26	*yjiE*	1T	S, S	*yjiE*[c] (303)
27	*Z2220*, *Z2221*	−10	S, S	*b1489* (807)
28	*Z0272*	N-del, 1T, −2, −1, −3	−, −	*rhsE* (682)
29	*Z0274*	−1	−, −	*ydcD* (160)
30	*yafM*	1T	−, −	*yafM* (165)
31	*Z0388*	1T	−, −	*ykgH* (222)
32	*Z0406*	+10	−, −	*yahC* (165)
33	*Z0828*, *ybfM*	−1	−, −	*ybfM* (468)
34	*Z0856*, *Z0854*	1T, C-del	−, −	*ybfD* (253)
35	*ybgQ*, *Z0870*	1T, +3	−, −	*ybgQ* (818)
36	*Z1038*	−S	−, −	*b0816* (89)
37	*Z1288*, *Z1289*	−1	−, −	*ycbS* (866)
38	*Z1419*, *Z1420*	−3, +3, +1	−, −	*yccE* (418)
39	*yceE*	1T	−, −	*yceE* (408)
40	*treA*	M-del	−, −	*treA* (565)
41	*Z2176*, *Z2175*	1T	−, −	*b1527* (371)
42	*hipA*	N-del	−, −	*hipA* (440)
43	*Z2254*, *Z2253*	1T, Ins, +1, −1	−, −	*ydcC* (378)
44	*Z2428*, *Z2431*	Altenative start, −6, 1T	−, −	*ydaH* (510)
45	*Z2542*	M-del, M-del	−, −	*yciQ* (631)
46	*Z2619*, *Z2618*	−1	−, −	*uidC* (417)
47	*Z2795*	−S	−, −	*b1762* (387)
48	*Z2836*	1T	−, −	*yeaP* (384)
49	*Z3144*, *Z3143*	−S, 1T	−, −	*yeeO* (547)
50	*molR_C*, *molR_D*	M-del, Ins, M-del, −1	−, −	*molR_1* (274), *molR_2* (645), *molR_3* (333)
51	*Z3292*, *Z3293*	1T	−, −	*yehM* (759)
52	*Z3480*, *Z3479*	1T	−, −	*b2226* (549)
53	*yfaA*	−1	−, −	*yfaA* (578)
54	*Z3627*	N-del, 1T, −6	−, −	*dsdX* (445)
55	*Z3634*, *yfdE*	−4	−, −	*yfdE* (394)
56	*srlA_1*, *srlA_2*	−1	−, −	*srlA* (187)
57	*srlE_1*, *srlE_2*	−2	−, −	*srlE* (319)
58	*Z4201*, *Z4200*	−5, Ins	−, −	*b2863* (278), *b2862* (99)
59	*Z4367*, *yqhG*	−4, −3	−, −	*yqhG* (309)
60	*agaC*	1T	−, −	*agaC*[c] (267)
61	*Z5154*, *Z5153*	−1	−, −	*yicO*[c] (470)
62	*rhaT*	1T	−, −	*rhaT*[c] (344)

Table 2 (*continued*)

	O157 ORF(s)	Disruption(s)[a]	Structure(s)[b]	K-12 ortholog(s) (amino acid residues)
63	*frwC_1*, *frwC_2*	1T	−, −	*frwC*[c] (359)
64	*yjgG_1*, *yjgG_2*	1T	−, −	*yjgG*[c] (110)

[a] N-del, C-del, and M-del respectively denote deletions of the N- and C-terminus and a middle section exceeding 10 bp, while +*n* and −*n* indicate an insertion and a deletion of *n* base pairs, respectively. 1T and −1T respectively signify the creation and loss of a stop codon by base substitution, whereas −S represents a start codon loss by the same mechanism. +IS and Ins stand for the introduction of an IS element and some other sequence over 10 bp in the middle, respectively.

[b] The first character in the structure column represents the presence (S) or absence (−) of an assigned three-dimensional structure in the pseudogene or the set of pseudogenes, while the second character shows the same in the ortholog(s). The underlined structures were judged to be unstable, while the italicized one could not be assessed.

[c] Only the K-12:MG1655 sequence is available for these ORFs.

Considering that there are 4288 protein-coding genes in K-12 (Blattner et al., 1997) pseudogenes nevertheless seem to be sparsely distributed in the *E. coli* genome. The pseudogene fraction (defined to be the ratio of the number of pseudogenes to the number of all gene-like entities including pseudogenes) is approximately 2% in this bacterium. The pseudogene fractions in other genomes of free-living bacteria are also small (Parkhill et al., 2000a,b, 2001a,b), never exceeding 5%. Pseudogene elimination from the genomes of free-living bacteria must be much faster than pseudogene production. By contrast, some, if not all, obligate parasites contain a considerably higher fraction of pseudogenes, as illustrated by *M. leprae* and *Borrelia burgdorferi*, in which the fractions of pseudogenes are 41 and 24%, respectively (Cole et al., 2001; Casjens et al., 2000). The ratio of production to elimination of pseudogenes must be higher in these obligate parasites than in free-living bacteria. In obligate parasites, functional redundancy with the host organisms may cause faster pseudogene generation and weaker selection pressure may result in slower pseudogene removal.

3.4. Origins of the probable pseudogenes

How did these probable pseudogenes arise? As demonstrated in the case of *Buchnera* (Silva et al., 2001), it is probable that many non-advantageous genes become pseudogenes before getting eliminated. Retrotransposition in mammals and duplication of genomic DNA in eukaryotes in general are known to give rise to pseudogenes (Mighell et al., 2000). Not surprisingly, we found no evidence of the former in *E. coli*. By contrast, 59% of the probable pseudogenes were found to have paralogs of unequal lengths. Since it is unclear whether these paralogs were generated by gene duplication, the mechanism should be restated: some become pseudogenes because they were made functionally redundant by the presence of paralogs, whatever their origins are.

What accounts for the rest? Since horizontal transfer of genes is a widespread process not only in obligate parasites (Andersson and Andersson, 1999; Wolf et al., 1999) but also in free-living bacteria like *E. coli* (Blattner et al., 1997; Lawrence and Ochman, 1997, 1998), a mechanism to delete non-advantageous genes to maintain a compact genome must exist. As in the case of genes with paralogs, we consider it likely that many of the acquired genes become pseudogenes during disintegration. Interestingly, 47% of the probable pseudogenes in K-12 were judged to be acquired genes (Ibid.). As only 18% of the K-12 genes were classified as horizontally transferred genes, this at first glance means that horizontal transfer genes are (0.47/0.18)/(0.53/0.82), or approximately 4.0 times more likely than their native coun-

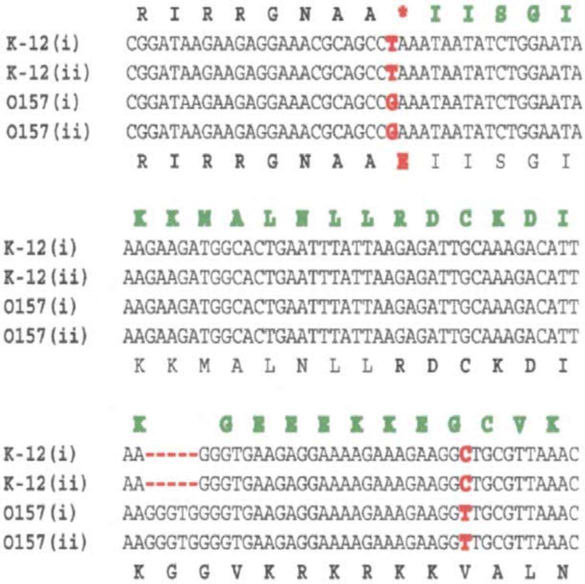
R I R R G N A A * I I S G I
K-12(i) CGGATAAGAAGAGGAAACGCAGCCTAAATAATATCTGGAATA
K-12(ii) CGGATAAGAAGAGGAAACGCAGCCTAAATAATATCTGGAATA
O157(i) CGGATAAGAAGAGGAAACGCAGCCGAAATAATATCTGGAATA
O157(ii) CGGATAAGAAGAGGAAACGCAGCCGAAATAATATCTGGAATA
R I R R G N A A E I I S G I
K K M A L N L L R D C K D I
K-12(i) AAGAAGATGGCACTGAATTTATTAAGAGATTGCAAAGACATT
K-12(ii) AAGAAGATGGCACTGAATTTATTAAGAGATTGCAAAGACATT
O157(i) AAGAAGATGGCACTGAATTTATTAAGAGATTGCAAAGACATT
O157(ii) AAGAAGATGGCACTGAATTTATTAAGAGATTGCAAAGACATT
K K M A L N L L R D C K D I
K G E E E K K E G C V K
K-12(i) AA-----GGGTGAAGAGGAAAAGAAAGAAGGCTGCGTTAAAC
K-12(ii) AA-----GGGTGAAGAGGAAAAGAAAGAAGGCTGCGTTAAAC
O157(i) AAGGGTGGGGTGAAGAGGAAAAGAAAGAAGGTTGCGTTAAAC
O157(ii) AAGGGTGGGGTGAAGAGGAAAAGAAAGAAGGTTGCGTTAAAC
K G G V K R K R K K V A L N

Fig. 2. Sequence alignment of pseudogene candidates and their orthologs. A continuous section of an alignment of a probable K-12 pseudogene, *b1459*, and its O157 ortholog, *Z0275*, is presented. (They are entered in row 30 of Table 1; *b1458* is located far upstream and therefore is not shown in the figure.) Independently determined sequences of the K-12 derivatives MG1655 (Blattner et al., 1997) (i) and W3110 (Yamamoto et al., 1997) (ii) and those of the O157:H7 isolates EDL933 (Perna et al., 2001) (i) and Sakai (Hayashi et al., 2001) (ii) are placed in the middle, with the reading frames in K-12 and O157 added above and below, respectively. The bases and amino acid residues that differ between the two strains are rubricated, with the introduced stop codon represented by a star (1T in Table 1), while the 5 bp deletion (−5 in the table) is symbolized by red dashes. A hypothetical reading frame in K-12 after the stop codon is shown in green.

terparts to become pseudogenes. We observe, however, that our method fails to detect pseudogenes of foreign origin that got transferred after the branching of the two strains. Since an accurate recounting would uncover more pseudogenes of foreign origin, the actual likelihood of horizontally transferred genes becoming pseudogenes must be even higher. Although the original list of horizontally transferred genes was later found to contain some errors (Koski et al., 2001), it is still plausible that such genes are more prone to become pseudogenes than native genes. The two mechanisms are not exclusive: 38% of the probable K-12 pseudogenes with paralogs were also classified as horizontally transferred genes, presumably reflecting the tendency of acquired genes that are less advantageous than the paralogs to become pseudogenes.

3.5. Conclusions

The *E. coli* genome contains a significant number of probable pseudogenes. We hope that theoretical investigations like this one will help narrow down the focus of experimental approaches to identify pseudogenes, as experimental verification of the absence of translation, transcription, or functionality is laborious. We propose that non-advantageous genes, consisting of a large number of horizontally transferred genes as well as native ones, often become pseudogenes before being eliminated altogether. We suggest that the rapid elimination of pseudogenes, as compared to their rate of production, results in a small fraction of genomes detected as pseudogenes in free-living bacteria in general.

Acknowledgements

We express our gratitude to A. Nishimura and Y. Yamazaki for vigorous discussions, T. Horiuchi for communication of unpublished data, J. Kato for critical comments, and H. Toh for erudite advice. This work was supported in part by postdoctoral fellowships to K.H. and S.F. in the ACT-JST program of Japan Science Technology, Corp. and a grant-in-aid from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

References

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402.

Andersson, J.O., Andersson, S.G., 1999. Insights into the evolutionary process of genome degradation. Curr. Opin. Genet. Dev. 9, 664–671.

Andersson, S.G., Zomorodipour, A., Andersson, J.O., Sicheritz-Pontén, T., Alsmark, U.C.M., Podowski, R.M., Näslund, A.K., Eriksson, A., Winkler, H.H., Kurland, C.G., 1998. The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. Nature 396, 133–140.

Blattner, F.R., et al., 1997. The complete genome sequence of *Escherichia coli* K-12. Science 277, 1453–1474.

Brinkkötter, A., Klöß, H., Alpert, C.-A., Langeler, J.W., 2000. Pathways for the utilization of N-acetyl-galactosamine and galactosamine in *Escherichia coli*. Mol. Microbiol. 37, 125–135.

Casjens, S., et al., 2000. A bacterial genome in flux: the twelve linear and nine circular extrachromosomal DNAs in an infectious isolate of the Lyme disease spirochete *Borrelia burgdorferi*. Mol. Microbiol. 35, 490–516.

Cole, S.T., et al., 2001. Massive gene decay in the leprosy bacillus. Nature 409, 1007–1011.

Engelberg-Kulka, H., Schoulaker-Schwarz, R., 1996. Suppression of termination codons. In: Neidhardt, F.C. (Ed.). 2nd Edition. *Escherichia coli* and *Salmonella*: Cellular and Molecular Biology, 1. ASM Press, Washington, DC, pp. 909–921.

Gonçalves, I., Duret, L., Mouchiroud, D., 2000. Nature and structure of human genes that generate retropseudogenes. Genome Res. 10, 672–678.

Harrison, P.M., Echols, N., Gerstein, M.B., 2001. Digging for dead genes: an analysis of the characteristics of the pseudogene population in the *Caenorhabditis elegans* genome. Nucleic Acids Res. 29, 818–830.

Hayashi, T., et al., 2001. Complete genome sequence of enterohemorrhagic *Escherichia coli* O157:H7 and genomic comparison with a laboratory strain K-12. DNA Res. 8, 11–22.

Itoh, T., et al., 1996. A 460-kb DNA sequence of the *Escherichia coli* K-12 genome corresponding to the 40.1–50.0 min region on the linkage map. DNA Res. 3, 379–392.

Kawabata, T., Arisaka, F., Nishikawa, K., 2000. Structural/functional assignment of unknown bacteriophage T4 proteins by iterative database searches. Gene 259, 223–233.

Kawabata, T., Fukuchi, S., Homma, K., Ota, M., Araki, J., Ichiyoshi, N., Nishikawa, K., 2002. GTOP: a database of protein structures predicted from genome sequences. Nucleic Acids Res. 30, 294–298.

Koski, L.B., Morton, R.A., Golding, G.B., 2001. Codon bias and base composition are poor indicators of horizontally transferred genes. Mol. Biol. Evol. 18, 404–412.

Lawrence, J.G., Ochman, H., 1997. Amelioration of bacterial genomes: rates of change and exchange. J. Mol. Evol. 44, 383–397.

Lawrence, J.G., Ochman, H., 1998. Molecular archaeology of the *Escherichia coli* genome. Proc. Natl. Acad. Sci. USA 95, 9413–9417.

Mighell, A.J., Smith, N.R., Robinson, P.A., Markham, A.F., 2000. Vertebrate pseudogenes. FEBS Lett. 468, 109–114.

Murzin, A.G., Brenner, S.E., Hubbard, T., Chothia, C., 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol. 247, 536–540.

Ogata, H., et al., 2001. Mechanisms of evolution in *Rickettsia conorii* and *R. prowazekii*. Science 293, 2093–2098.

Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B., Thornton, J.M., 1997. CATH – a hierarchic classification of protein domain structures. Structure 5, 1093–1108.

Parkhill, J., et al., 2000a. The genome sequence of the food-borne pathogen *Campylobacter jejuni* reveals hypervariable sequences. Nature 403, 665–668.

Parkhill, J., et al., 2000b. Complete DNA sequence of a serogroup A strain of *Neisseria meningitidis* Z2491. Nature 404, 502–506.

Parkhill, J., et al., 2001a. Genome sequence of *Yersinia pestis*, the causative agent of plague. Nature 413, 523–527.

Parkhill, J., et al., 2001b. Complete genome sequence of a multiple drug resistant *Salmonella enterica* serovar Typhi CT18. Nature 413, 848–852.

Perna, N.T., et al., 2001. Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. Nature 409, 529–533.

Poon, D., Schroeder, S., Wang, C.K., Yamamoto, T., Horikoshi, M., Roeder, R.G., Weil, P.A., 1991. The conserved carboxy-terminal domain of *Saccharomyces cerevisiae* TFIID is sufficient to support normal cell growth. Mol. Cell. Biol. 11, 4809–4821.

Reid, S.D., Herbelin, C.J., Bumbaugh, A.C., Selander, R.K., Whittam, T.S., 2000. Parallel evolution of virulence in pathogenic *Escherichia coli*. Nature 406, 64–67.

Silva, F.J., Latorre, A., Moya, A., 2001. Genome size reduction through multiple events of gene disintegration in *Buchnera* APS. Trends Genet. 17, 615–618.

Vanin, E.F., 1985. Processed pseudogenes: characteristics and evolution. Annu. Rev. Genet. 19, 253–272.

Wolf, Y.I., Aravind, L., Koonin, E.V., 1999. Rickettsiae and Chlamydiae: evidence of horizontal gene transfer and gene exchange. Trends Genet. 15, 173–175.

Yamamoto, Y., et al., 1997. Construction of a contiguous 874-kb sequence of the *Escherichia coli* K-12 genome corresponding to 50.0–68.8 min on the linkage map and analysis of its sequence features. DNA Res. 4, 91–113.