Journal of Industrial Microbiology & Biotechnology (2001) 27, 163–169
© 2001 Nature Publishing Group 1367-5435/01 $17.00

www.nature.com/jim

Cross-genomic analysis of the translational systems of various organisms

K Fujita[1], T Horie[1] and K Isono[2]

[1]*Graduate School of Science and Technology, Faculty of Science, Kobe University, Kobe, Japan;* [2]*Department of Biology, Faculty of Science, Kobe University, Kobe, Japan*

We have characterized the genes encoding ribosomal proteins (r-proteins) as well as other translation-related factors of 15 eubacteria and four archaebacteria, and the genes for the mitochondrial r-proteins of *Saccharomyces cerevisiae* by using the complete genomic nucleotide sequence data of these organisms. In eubacteria, including two species of *Mycoplasma*, the operon structure of the r-protein genes is well conserved, while their relative orientation and chromosomal location are quite divergent. The operon structure of the r-protein genes in archaebacteria, on the other hand, is quite different from eubacteria and also among themselves. In addition, many archaebacterial r-proteins show similarity to rat cytoplasmic r-proteins. Nonetheless, characteristic features of several genes encoding proteins of functional importance are well conserved throughout the bacterial species including archaebacteria, as well as in *S. cerevisiae*. We searched for the genes encoding mitochondrial r-proteins in yeast by combining informatics and genetic experiments. Furthermore, we characterized some of the r-proteins genes by exchanging portions between *Escherichia coli* and *S. cerevisiae* and performed functional analysis of some of the genes from different evolutionary points of view. Our work may be extended towards phylogenetic analysis of organisms producing secondary metabolites of various sorts. *Journal of Industrial Microbiology & Biotechnology* (2001) **27,** 163–169.

Keywords: genome data; ribosomal proteins; phylogenetic relationship; mitochondrial ribosomes

Introduction

Establishment of the genomic nucleotide sequences of about 20 organisms has made it possible to perform extensive cross-genomic comparison of various biological structures of interest. We have been analyzing the structure and function of the ribosomal proteins (r-proteins) and their genes in several model organisms, in particular *Escherichia coli* and *Saccharomyces cerevisiae*. Since the ribosome is an essential subcellular organelle and composed of RNA and a large number of proteins, systematic analysis of its components is expected to reveal clues as to how individual components are interrelated with each other and to what extent their structural and functional relations are conserved during the course of evolution. Consequently, we took advantage of the complete genomic nucleotide sequence data and used them for the analysis of the ribosome and r-proteins from these points of view.

There are many structural entities that play roles in the translation of genetic messages within the cell. Of them, the ribosome is the pivotal structure on which key steps of the decoding genetic messages take place. Two ribosomal subunits of unequal size occur in all organisms. They contain RNA and more than 70 different protein molecules, and the actual steps of translation, namely decoding the genetic messages and simultaneous transpeptidation (amino acid polymerization) reactions, occur in the cavity created and protected by the two ribosomal subunits. Because of the high degree of functional importance in the translation of genetic messages, interaction between ribosomal subunits and their individual components must have been highly elaborated during the course of evolution. A mutation in one of the components will affect its local conformation, thereby changing its interaction with other components, and consequently the function of the whole ribosome may be altered. Mutations such as resistance to streptomycin and other antibiotics as well as those leading to temperature-sensitive assembly of the ribosomal subunits are examples of this kind. However, at the same time, it should be noted that mutants of *E. coli* which apparently lacked a few r-proteins have been reported [11]. Moreover, the ribosomal components of bacteria such as *E. coli* and *Bacillus stearothermophilus* that are evolutionary rather than distantly related are interchangeable at least when analyzed *in vitro* [32].

Since in *E. coli* all r-proteins and their genes have been extensively characterized, and since *in vitro* reconstitution of RNA and r-proteins into an active ribosomal subunit is possible, it would be interesting to analyze as to what extent we might be able to correlate the evolutionary conservation and structural importance of individual r-proteins in *E. coli* and related bacteria. Furthermore, comparative studies of r-proteins genes at the genomic level would clarify whether any one or more r-proteins are, either partly or totally, dispensable or not, and if so, what would be a prerequisite for that to occur.

Earlier, it was reported that the mitochondrial ribosome (mito-ribosome) of *S. cerevisiae* apparently contained more proteins than its *E. coli* counterpart [21]. It appears a little strange in view of the fact that the mito-ribosome is engaged in the translation of only a limited number of messages encoded in the mitochondrial genome. Moreover, all, except one, proteins in the yeast mito-ribosome are encoded by nuclear genes, while the RNA components are transcribed from the mitochondrial genome. This poses another interesting problem concerning the informational interaction between the nuclear and mitochondrial genes with respect to their cooperation with each other in the synthesis of mito-ribosomes. For unequivocal identification of r-proteins, isolation of individual

Correspondence: Dr K Isono, Department of Biology, Faculty of Science, Kobe University, Rokkodai, Kobe 657-8501, Japan

Received 21 September 1999; accepted 22 September 2000

proteins followed by their amino acid analysis is essential. Therefore, we purified and characterized as many yeast mito-ribosomal proteins (mito-r-proteins) as possible by using various methods. Based upon the data thus obtained, we performed systematic analysis of the yeast genome for the presence of the genes encoding likely mito-r-proteins, as will be reported below.

Materials and methods

Comparison of r-proteins

The genomic nucleic acid sequence data of organisms listed in Table 1 were retrieved either from the DDBJ/EMBL/GenBank nucleic acid databases (genomes section) or from the World Wide Web server at TIGR (http://www.tigr.org/) and/or from the individual organism databases (references to the organisms listed are given in the footnote to Table 2). The reference amino acid sequence data were obtained from the Swiss-Prot (release 37.0). To perform systematic comparison of individual r-proteins, we mostly used the FASTA program [33]. For this purpose, the genomic nucleotide sequence of each organism was first translated into amino acid sequences in six reading frames and then subjected to FASTA analysis against the amino acid sequence data of the r-proteins and other translation-related factors of *E. coli* and *B. stearothermophilus* [39]. The FASTA scores obtained were subsequently converted into "degrees of conservation" by normalizing them with the corresponding "self-examination" data of *E. coli* [19].

Analysis of the yeast mito-r-proteins

Data for the yeast mito-r-proteins as well as their genes, collected by Kitakawa and Isono [28] and Graack and Wittmann-Liebold [21], were taken as controls for the analysis of the yeast genome by GeneMark [5]. A total of 55 genes listed in Table 2 were used for the construction of GeneMark matrices of orders two through four that are specific for yeast mito-r-proteins according to the procedure described earlier [24]. Of the 55 genes used, 27 encode proteins showing similarity to r-proteins of *E. coli* and other organisms as indicated. GeneMark analysis was then performed using the mito-r-protein gene matrices (others, three and four) along with the yeast matrices of the same orders that were retrieved from the GeneMark WWW server at E-mail: http://genemark.biology.gatech.edu/GeneMarck. The GeneMark data obtained were subsequently classified and ORFs encoding proteins longer than 30 amino acid residues with GeneMark scores higher than 0.6 were selected for further analysis.

Table 1 Microorganisms used in this work

Organism	Size (Mb)	Number of ORFs
My. genitalium	0.58	470
My. pneumoniae	0.82	679
Borrelia burgdorferi	0.91	843
C. trachomatis	1.04	894
R. prowazekii	1.11	834
Treponema pallidum	1.13	1041
C. pneumoniae	1.23	1052
A. aeolicus	1.55	1512
He. pylori strain J99	1.64	1495
He. pylori	1.67	1590
H. influenzae	1.83	1743
Synechocystis sp.	3.57	3168
B. subtilis	4.21	4100
Mycobacterium tuberculosis	4.41	3924
E. coli	4.67	4288
M. jannaschii	1.67	1738
Pyrococcus horikoshii OT3	1.73	2061
Methanobacterium thermoautotrophicum	1.75	1855
Archaeoglobus flugidus	2.18	2436
S. cerevisiae	12.07	5885

Results and discussion

Phylogenetic distance of organisms measured by comparison with E. coli r-proteins en masse

First, we performed extensive comparisons of r-protein genes of organisms listed in Table 1 with those of *E. coli*. A total of 55 r-protein genes listed in Table 2 were used for this purpose. Similarly, the genes encoding other transcription/translation-related factors such as initiation factors, elongation factors, peptide chain release factors, r-protein modifying enzymes, RNA polymerase subunits, *etc.*, were analyzed (data not shown). The nucleotide sequence, along with its reverse complementary sequence for each organism, was cut into segments of 55,000 nucleotides, allowing terminal 5000 nucleotides to overlap with the neighboring segments and then translated in six reading frames. They were then subjected to extensive FASTA analysis with the *E. coli* protein sequences. The results were manually inspected to evaluate the compared sequences so that even if overall FASTA scores were low, the data were taken for further analysis if the region of similarity spread widely along the ORF/gene translations. The raw FASTA scores thus obtained were subsequently converted into what we termed "degrees of conservation" by normalizing the scores with those obtained by self-examination of the corresponding *E. coli* proteins.

Results are summarized in Table 3. It is readily obvious that *Hemophilus influenzae* is the closest relative of *E. coli* as far as the r-protein genes are concerned. There are several noticeable differences in the data thus obtained from those obtained by comparing the ribosomal RNA sequences alone. First of all, although *Aquifex aeolicus* [13] was said to be placed closest to the branch point of eubacteria and archaebacteria by the rRNA-based calculation [9] as discussed by Pennisi [34], it seems much closer to *E. coli* as presented in Table 3. Indeed, Deckert *et al.* [13] pointed out that the *A. aeolicus* proteins deduced from the nucleotide sequence data that are involved in translation (including r-proteins) are more similar to *E. coli* than to *Methanococcus jannashii* [8]. We believe that our data are more appropriate for the evaluation of the phylogenetic relationships of organisms than comparing just gene of either RNA or protein even if the gene used is of prime importance as in the case of ribosomal RNA.

Another point that should be noted is the "distance" measured in our way which suggests that, despite the kingdom barrier, the kinship of *S. cerevisiae* mitochondria with *E. coli* is closer than that of archaebacteria. Our calculation suggests that all of the eubacteria analyzed, including the two *Chlamydias* and the two *Mycoplasmas*, are closely related with *E.coli* as far as their translation-related proteins are concerned. Furthermore, the phylogenetic distances of the four species of archaebacteria from *E. coli* are very large. Apparently, the divergence in the translational systems of these archaebacteria is much greater than expected from the r-RNA-based phylogenetic tree. To evaluate the validity of our calculation, we need to perform similar

Table 2 Degree of conservation of r-proteins of the organisms listed in Table 1[a]

E.co prot	Size (a.a)	H.in	B.sb	My.tb	R.pr	Syn	B.br	T.pl	H.py99	H.py	A.ae	Ch.tr	Ch.pn	M.gn	M.pn	S.ce mit	M.jn	P.ho	A.fl	M.th	S.ce cyt-1	S.ce cyt-2
S1	558	81	26	31	50	14	29	36	31	31	29	50	49	–	–	–	7	–	–	–	7	–
S2	242	86	59	55	52	56	53	58	52	52	56	49	47	32	33	20	–	9	11	8	–	–
S3	234	91	63	60	53	59	49	50	64	63	61	57	57	41	40	–	26	22	21	21	–	–
S4	207	94	52	46	35	42	43	41	56	56	49	37	37	38	39	–	14	12	12	–	–	–
S5	168	95	63	62	52	50	54	58	53	52	56	50	52	49	49	24	26	22	23	25	23	–
S6	132	76	33	34	34	27	–	–	–	31	29	21	22	17	18	–	–	–	–	–	–	–
S7K	180	82	61	54	52	53	52	59	61	60	50	56	56	52	49	25	–	–	–	20	–	–
S8	131	90	56	58	45	59	42	48	41	41	21	43	44	48	50	–	26	27	23	26	–	–
S9	131	89	56	52	55	46	56	46	50	49	39	50	46	50	51	49	–	25	28	29	–	–
S10	104	98	74	66	63	71	65	67	70	70	64	74	73	47	48	29	37	39	41	33	29	–
S11	130	96	70	65	56	66	65	62	58	58	65	56	56	44	50	–	29	32	33	35	27	27
S12	125	98	64	76	56	82	79	74	78	78	79	76	79	64	64	57	15	20	18	–	–	–
S13	118	84	71	67	54	59	65	62	58	61	64	55	56	63	65	30	23	20	19	19	17	19
S14	102	94	39	47	45	49	27	25	25	24	28	52	53	23	22	31	–	–	–	–	–	–
S15	90	85	70	63	62	59	62	56	59	59	60	59	59	48	47	41	–	24	–	–	–	–
S16	74	83	48	–	60	33	52	53	59	53	–	–	–	32	–	36	–	–	–	–	–	–
S17	85	85	51	51	46	50	50	46	42	42	62	40	37	32	–	–	34	–	30	35	32	32
S18	76	94	63	44	50	48	58	51	45	45	42	56	56	47	47	–	–	–	–	–	–	–
S19	93	93	76	74	68	74	59	57	67	67	53	64	63	65	65	36	29	39	29	31	23	–
S20	87	81	40	41	33	26	24	28	29	29	38	30	–	20	–	–	–	–	–	–	–	–
S21	72	85	47	–	–	40	42	40	44	45	40	39	40	–	–	–	–	–	–	–	–	–
S22	46	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
L1	235	89	58	54	58	57	49	45	62	63	53	60	60	50	51	21	21	24	24	27	–	–
L2	274	92	64	61	62	55	60	59	54	53	49	56	57	47	48	19	27	26	23	29	22	22
L3	210	91	52	55	51	53	50	49	25	26	50	45	47	42	43	40	16	14	16	15	–	–
L4	202	88	47	33	30	39	33	29	28	28	26	24	24	29	29	–	–	–	–	–	–	–
L5	180	94	71	68	67	69	68	65	57	55	70	56	58	64	67	–	25	26	20	23	19	19
L6	178	84	54	52	46	51	44	45	45	46	40	48	46	42	44	31	20	–	–	26	15	–
L7/L12	122	72	52	54	57	33	34	55	69	30	70	45	43	18	19	–	–	–	–	–	–	–
L9	150	81	29	42	38	27	27	32	33	–	37	25	24	21	21	–	–	–	–	–	–	–
L10	166	91	44	34	29	69	57	28	–	65	22	–	–	38	39	40	34	11	38	–	–	–
L11	143	86	95	69	57	62	64	63	67	68	65	61	62	45	46	35	15	35	–	34	–	–
L13	143	93	61	59	50	66	58	58	55	55	63	56	56	48	50	36	–	20	20	–	–	–
L14	124	91	62	74	72	60	64	67	71	69	59	67	65	57	57	–	24	28	21	27	21	21
L15	145	89	49	34	32	42	37	39	36	36	38	35	36	38	40	34	16	–	–	16	–	–
L16	137	93	67	56	59	71	61	60	67	67	56	62	62	49	51	33	–	–	–	–	–	–
L17	128	93	44	46	60	45	43	47	46	45	50	29	28	31	29	33	–	–	–	–	–	–
L18	118	89	51	45	41	52	32	39	27	28	46	36	37	29	29	–	–	–	17	19	–	–
L19	116	94	62	58	52	62	60	59	52	52	48	50	52	51	50	–	–	–	14	–	–	–
L20	119	97	70	65	65	65	55	48	61	61	63	55	52	55	59	–	–	–	–	–	–	–
L21	104	82	52	44	45	41	41	45	46	47	44	45	46	31	31	–	–	–	–	–	–	–
L22	111	95	60	53	50	47	38	38	33	33	36	46	44	45	51	–	24	24	–	22	–	–
L23	101	75	27	27	35	35	33	35	28	28	38	–	–	–	–	–	27	27	21	22	24	–
L24	105	84	49	41	49	–	45	44	28	29	37	23	27	31	–	–	16	17	18	–	20	21
L25	97	70	–	35	44	–	16	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
L27	86	87	53	63	56	65	60	52	62	62	56	51	51	48	49	50	–	–	–	–	–	–
L28	79	91	28	48	33	38	–	–	27	–	–	40	39	–	–	–	–	–	–	–	–	–
L29	64	81	48	46	41	–	–	–	41	41	35	28	–	40	43	–	42	–	31	–	–	–
L30	60	82	48	52	47	–	32	48	–	–	–	–	–	–	–	31	–	–	–	–	–	–
L31	71	82	55	62	15	38	31	67	44	44	41	27	27	41	39	–	–	–	–	–	–	–
L32	58	81	–	–	47	–	27	–	–	–	–	29	27	42	40	–	–	–	–	–	–	–
L33	56	85	42	59	56	33	52	41	41	41	40	50	51	34	33	32	–	–	–	–	27	–
L34	47	89	72	61	74	54	78	73	66	66	53	64	63	72	70	53	–	23	–	–	–	–
L35	66	86	48	40	30	31	43	50	33	33	37	43	44	37	35	21	–	–	–	–	–	–
L36	39	89	69	65	34	82	69	74	83	83	63	–	–	73	77	64	–	–	–	–	–	–
Total	7219	4726	2866	2696	2601	2505	2487	2465	2428	2420	2365	2267	2208	2060	1977	951	573	564	551	539	306	161

[a] Values indicating the "degree of conservation" were calculated as described in Materials and Methods. Abbreviations for organism names are: E.co, *E. coli* [4]; H.in, *H. influenzae* [15]; B.sb, *B. subtilis* [31]; M.tb, *Myc. tuberculosis* [10]; R.pr, *R. prowazekii* [2]; Syn, *Synechocystis* sp. [26]; B.br, *Bo. burgdorferi* [17]; T.pl, *T. pallidum* [18]; H.py99, *He. pylori* J99 [1]; H.py, *H. pylori* [37]; A.ae, *A. aeolicus* [13]; Ch.tr, *C. trachomatis* [36]; Ch.pn, *C. pneumoniae* [25]; M.gn, *My. genitalium* [16]; M.pm, *My. pneumoniae* [22]; S.ce, *S. cerevisiae* [20]; M.jn, *M. jannaschii* [8]; P.ho, *P. horikoshii* [27]; A.fl, *Ar. fulgidus* [30]; M.th, *Me. thermoautotrophicum* [35].

systematic comparison using the r-protein genes of at least one of the archaebacteria. If a reciprocal analysis based on such information yields a phylogenetic "distance" between *E. coli* and the archaebacterium in question in a manner comparable to the one presented in Table 3, then the validity of our calculation will be greatly strengthened. To do so, we need to have the amino acid sequence data of the r-proteins of that archaebacterial species so as to establish unequivocally the genes encoding respective

Table 3 Phylogenetic distances measured by all r-proteins[a]

H. influenzae	4726
B. subtilis	2866
Myc. tuberculosis	2696
R. prowazekii	2601
Synechocystis sp.	2505
Bo. burgdorferi	2487
T. pallidum	2465
He. pylori J99	2428
He. pylori	2420
A. aeolicus	2365
C. trachomatis	2267
C. pneumoniae	2208
My. genitalium	2060
My. pneumoniae	1977
S. cerevisiae	951
M. jannaschii	573
P. horikoshii	564
Ar. fulgidus	551
Me. thermoautotrophicum	539

[a]The values are cumulative "degree of conservation" data shown in Table 2.

r-proteins. However, such a task is rather difficult and seems not very practical. It is known that, e.g., in *M. jannashii*, there are many putative r-proteins which show similarity to rat cytoplasmic r-proteins [19]. However, no proof has been established as to whether they actually function as r-proteins in *M. jannashii*.

Earlier, Andersson *et al.* [2] described the results of their comparative analysis of r-proteins encoded in bacterial, mitochondrial and chloroplast genomes. They chose r-proteins S2, S3, S7, S8, S9, S10, S11, S12, S13, S14, S19, L5, L6 and L16 for their comparison. Their results are distinctly different from ours: e.g., they assigned *Rickettsia prowazekii* at a more distant place than *Helicobacter pylori* towards the mitochondrial genomes, thereby implicating the phylogenetic proximity of *R. prowazekii* with mitochondria. However, they included only 14 r-proteins (mostly from the small subunit) for their calculation to make it possible to perform direct comparison with mitochondria and chloroplasts. We confirmed their results by using the same 14 r-proteins for calculation. The two *Chlamydia* species and the two *He. pylori* strains became much closer to *E. coli*, while *R. prowazekii* and *A. aeolicus* were farthest from *E. coli* except for the two *Mycoplasma* species (Table 4). It is apparent from our data that the inclusion of all r-protein genes, especially the genes that have disappeared from the mitochondrial and chloroplast genomes during the course of evolution, is important for the estimation of phylogenetic relationship of organisms.

Another point that should be noted is that protein S1, which is the largest of all r-proteins of *E. coli* and behaves like a factor as manifested by its involvement in the Qβ phage replicase, is at least structurally well conserved among the organisms examined as shown in Table 2. However, the function of its homologs in other organisms might be different from that of *E. coli* S1, as discussed by Danchin [12]. Earlier, we experimentally proved that its functional homolog was absent from the ribosome of *B. stearothermophilus* [23], a closer relative of *B. subtilis*, despite the fact that the genome of *B. subtilis* contained the gene encoding a structural homolog of *E. coli* S1. Clearly, additional experiments are necessary to elucidate the function of S1 homologs in organisms other than *E. coli,* including *B. subtilis.* During the course of analyses presented in Table 2, we also noticed that some of the small r-proteins such as L34 and L35, whose precise functions are not well established, are well conserved in the organisms analyzed. Why they are rather highly conserved and what roles they play in the ribosome remain to be investigated.

Based upon the data described above, it seems possible to extend the biochemical and genetic data obtained with *E. coli* r-proteins to other organisms, including those producing useful secondary metabolites, if they are within a reasonable phylogenetic distance from *E. coli*. Since the ribosomes and other translation-related factors are essential in synthesizing all cellular components including secondary metabolites of various sorts, such an approach might be useful in searching for bacteria that are phylogenetically related to those listed in Table 3. All 14 eubacterial species from *H. influenzae* down to *Mycoplasma pneumoniae* listed in Table 3 can possibly be analyzed in this way, although perhaps it might not be so easy to do so with organisms showing lower scores, such as *Chlamydia trachomatis*, *C. pneumoniae*, *My. genitalium* and *My. pneumoniae*. However, at least at the moment, we have no means to perform such an analysis with the four archaebacteria, since they are phylogenetically too widely distant from *E. coli*.

Search for possible mito-r-protein genes in yeast

Previously, we reported many mito-r-proteins that we isolated and characterized mainly from the large subunit of the yeast mitoribosome [28]. A total of 60 mito-r-proteins have been identified in our studies and the work reported by others as listed in Table 5. They are largely basic proteins harboring pI of 10 or higher [21]. Of the 60 mito-r-proteins, 28 show similarity to other r-proteins, especially to those of *E. coli*. However, the remaining 32 proteins do not show an appreciable degree of similarity to any known protein from yeast or other origins. They are interpreted to have been recruited from other sources during the course of evolution. Since the yeast mito-ribosome appears to contain as many as 80 proteins [21], the list is not complete, especially for proteins of the small subunit.

Table 4 Phylogenetic distances measured by 14 r-proteins[a]

H. influenzae	1272
B. subtilis	880
Synechocystis sp.	865
Myc. tuberculosis	848
C. trachomatis	797
H. pylori J99	794
C. pneumoniae	793
H. pylori	791
Bo. burgdorferi	785
T. pallidum	776
R. prowazekii	771
A. aeolicus	744
My. pneumoniae	699
My. genitalium	684
S. cerevisiae	341
Me. thermoautotrophicum	269
P. horikoshii	259
Ar. fulgidus	243
M. jannaschii	230

[a]The values are cumulative "degree of conservation" data shown in Table 2. Only the 14 mito-r-protein genes analyzed by Andersson *et al.* [2] for their phylogenetic estimation as listed in the text were used for calculation.

Table 5 Summary of mito-r-proteins of *S. cerevisiae*

Protein	ORF	Gene	Chromosome	Length	Homolog[b]	Essential?
Large subunit proteins						
YmL2	YNL005c	*MRP7*	14	371	L27	yes
YmL3	YMR024w	*MRPL3*	13	390		–[a]
YmL4	YLR439w	*MRPL4*	12	319		yes
YmL5/7	YDR237w	*MRPL7*	4	292	L5	–
YmL6	YML025c	*YML6*	13	286	L4	–
YmL8	YJL063c	*MRPL8*	10	238	L17/S13	yes
YmL9	YGR220c	*MRPL9*	7	269	L3	yes
YmL10	YNL284c	*MRPL10*	14	272	L15	–
YmL11	YDL202w	*MRPL11*	4	249	L10	–
YmL13	YKR006c	*MRPL13*	11	275		no
YmL14	YMR193w	*MRPL14*	13	258	L28	–
YmL15	YLR312wa	*MRPL15*	12	253		–
YmL16	YHR147c	*MRPL6*	8	214	L6	yes
YmL17	YNL252c	*MRPL17*	14	281		yes[d]
YmL18	YNL284c	*MRPL10*	14	272	L15	–
YmL19	YNL185c	*MRPL19*	14	158	L11	–
YmL20	YKR085c	*MRPL20*	11	195		yes
YmL23	YOR150w	*MRPL23*	15	164	L13	–
YmL24	YMR193w	*MRPL14*	13	258	L28	–
YmL25	YGR076c	*YMR26*	7	156		yes
YmL27	YBR282w	*MRPL27*	2	146		yes
YmL28	YDR462w	*MRPL28*	4	147		–
YmL30	YNL252c	*MRPL17*	14	281		yes[d]
YmL31	YKL138c	*MRPL31*	11	131		yes
YmL32	YCR003w	*MRPL32*	3	183		–
YmL33	YMR286w	*MRPL33*	13	99	L30/L16	yes
YmL34	YKL170w	*MRPL38*	11	138	L14	–
YmL35	YDR322w	*MRPL35*	4	367		–
YmL36	YBR122c	*MRPL36*	2	196		–
YmL37	YBR268w	*MRPL37*	2	105		–
YmL38	YKL170w	*MRPL38*	11	138	L14	–
YmL39	YML009c	*MRPL39*	13	70	L33	–
YmL40	YPL173w	*MRPL40*	16	297	S4*	–
YmL41	YDR405w	*MRP20*	4	263	L23	yes
YmL44	YMR225c	*MRPL44*	13	98		–
YmL45[c]						
YmL47	YBL038w	*RML16*	2	232	L16	yes
YmL49	YJL096w	*MRPL49*	10	224		yes[d]
–	YEL050c	*RML2*	5	393	L2	yes
–	YKL167c	*MRP49*	11	137		no[e]
–	YDR115w	–	4	105	L34	–
–	YDR116c	–	4	285	L1	–
–	YGL068w	–	7	194	L12	–
–	YPL183wa	–	16	93	L36	–
Small subunit proteins						
YMS2	YHR075c	*MRPS2*	8	400		–
YMS16	YKL003c	*MRP17*	11	131		yes
YMS18	YNL306w	*MRPS18*	14	217	S11	–
YMS-A	YGR084c	*MRP13*	7	324		no
YMS-T	YDL045wa	*MRP10*	4	95		yes
–	Q0140	*varl*	mt	396		yes
–	YDR347w	*MRP1*	4	321		yes
–	YPR166c	*MRP2*	16	115	S14	yes
–	YHL004w	*MRP4*	8	394	S2	yes
–	YBR251w	*MRPS5*	2	307	S5	no[d]
–	YBR146w	*MRPS9*	2	278	S9	yes
–	YNR036c	*MRPS12*	14	153	S12	yes
–	YBL090w	*MRP21*	2	177	S21	yes
–	YDR337w	*MRPS28*	4	286	S15	yes
–	YPL118w	*MRP51*	16	344		yes
–	YNL137c	*NAM9*	14	485	S4	yes
–	YOR158w	*PET123*	15	318		yes
–	YPL013c	*LPA4*	16	121	S16/S24	–
–	YDR041w	–	4	203	S10	–
–	YJR113c	–	10	247	S7	–
–	YMR188c	–	13	237	S17	–
–	YNL081c	–	14	143	S13	–
–	YNR037c	–	14	91	S19	–
Subunit unknown						
–	YFR049w	*YMR31*	6	123		–

[a] The – symbol indicates either "not found" or "experiments not done".
[b] The names indicate homologous *E. coli* r-proteins except for S4* which is a protein identified in potato [6].
[c] The previous identification of YmL45 and ORF YGL125w was mistaken (see text for details).
[d] We newly confirmed that these proteins are essential for mitochondrial function.
[e] Reported by Fearon and Mason [14].

To search for the genes encoding the remaining 20 or so r-proteins, we analyzed the genomic nucleotide sequence data of *S. cerevisiae* by the computer program, GeneMark, as described before [24]. For this purpose, new matrices of orders two through five for GeneMark analysis were first prepared using the cumulative nucleotide sequence data of mito-r-protein genes listed in Table 5. They were then used to survey the nucleotide sequence data of individual chromosomes for the occurrence of likely mito-r-protein genes. In addition, we surveyed the yeast chromosomes with matrices prepared for average yeast genes as controls. The GeneMark scores thus obtained were then compared between the corresponding results with mito-r-protein matrices and yeast matrices of the same orders. ORFs encoding proteins of less than 300 amino acid residues that showed GeneMark scores of 0.6 or higher with the mito-r-protein matrices but lower scores with the yeast matrices were selected, translated and subjected to FASTA analysis against the Swiss-Prot database. Many of the mito-r-proteins that we had previously characterized could be identified in this way, as expected. At the same time, genes encoding proteins, such as heat shock proteins, some of the G-proteins and translation initiation factors were also given high scores in the GeneMark analysis with the mito-r-protein matrices. In addition, the genes for proteins classified as "hypothetical" by the yeast genome analysis were selected.

The genes encoding the last category of proteins described above are of potential interest. Some of them might encode new mito-r-proteins that are hitherto unknown. However, there is no obvious way to analyze the functions of these genes/ORFs other than performing gene disruption and/or intracellular localization of their products, both of which require time and intensive attention. As an alternative way, we search for their homologs in the genome of *Caenorhabditis elegans* [38], hoping that at least some of them could be identified if their homologs in *C. elegans* have been established to exist or further classified as mito-r-proteins. However, we were unable to identify any of them in this way. Obviously, the phylogenetic distance between *S. cerevisiae* and *C. elegans* is not close enough for this type of analysis. Experiments are currently underway to perform disruption of some of the ORFs of unknown function and to analyze them.

Specific analysis of several mito-r-proteins

In addition to searching for new mito-r-protein genes in the *S. cerevisiae* genome as described above, we performed experiments to characterize some of the yeast mito-r-proteins in detail. As

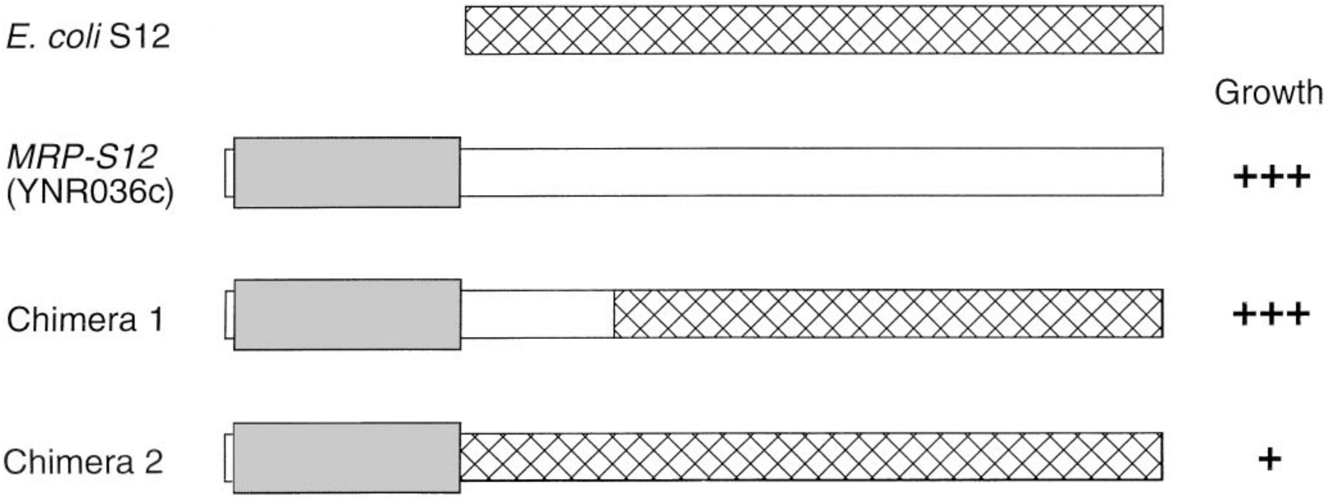

Figure 1 Functional similarity of r-protein S12 of *E. coli* and its yeast mitochondrial homolog. Two types of fusion genes chimeric between *E. coli rpsL* and yeast *MRP-S12* were constructed. The hatched area indicates the region derived from *rpsL*, and gray and white boxes the MTS and the remaining region of *MRP-S12*, respectively. Chimera 1 contains the highly conserved C-terminal half derived from *rpsL*, while chimera 2 contains the entire *rpsL* except for the MTS. + + + , normal growth; + , slow growth.

shown in Table 2, protein S12 is one of the most conserved r-proteins in all organisms analyzed. It is the target of streptomycin resistance in *E. coli* and many mutants resistant to the drug are known. There are four sites within S12 protein that are altered in such mutants [3]. *S. cerevisiae* has been found to possess a homolog of this protein not only in the mito-ribosome but also in the cytoplasmic ribosome, although we failed to identify the cytoplasmic homolog in our initial survey (Table 2).

The structure of the protein deducted from the ORF termed YNR036c is shown in Figure 1. There is a stretch in its N-terminal region which is considered to be a matrix-targeting signal (MTS). The C-terminal half of the amino acid sequence of protein S12 and its homologs is particularly highly conserved as indicated, including the yeast cytoplasmic homolog, RPS28, which has been reported to be involved in resistance to the antibiotic paromomycin [3]. All sites altered in streptomycin-resistant mutants are located within this region. A disruptant of the ORF YNR036c, which we named *MRP-S12*, was unable to grow on a glycerol-containing medium and its colonies were *petite* (respiration-deficient) on a glucose-containing agar plate. We constructed plasmids encoding chimeric proteins by fusing the *E. coli* S12 (*rpsL*) gene with *MRP-S12* and expressed the individual chimeric genes in the disruptant. Replacement of the highly conserved C-terminal half of the *MRP-S12* gene with the corresponding *E. coli rpsL* gene did not appreciably alter growth, while another chimera in which the region except for the MTS was completely replaced with the *E. coli rpsL* grew very slowly as indicated. These results indicate that the basic function of the *MRP-S12* gene and its homologs resides in the C-terminal highly conserved region, while the organism-specific function is expressed in the less-conserved region of the protein. Whether this conclusion can be generalized or not remains to be analyzed further with other r-proteins.

As mentioned earlier, we have identified a total of 13 new mito-r-proteins in the genomic sequence of *S. cerevisiae* [29]. We have chosen three of the newly identified ORFs/genes, YNL252c, YGL125w and YJL096w, and characterized them further. Of the three, YGL125w was found to encode methylenetetrahydrofalate reductase which was not associated with mito-ribosomes (Kishida *et al.*, unpublished results) and hence, our previous assignment of the sequenced peptide of the sequence MdlaYEASLaQ with the peptide MDRMYEASLPQ that is encoded by YGL125w was most probably mistaken. A disruptant of the ORF YJL096w, which we assigned to encode mito-r-protein YmL49, was very similar to the disruptants of other mito-r-protein genes: inability to grow on glycerol and *petite* colony formation on glucose. Furthermore, the protein encoded by a YJL096w derivative to which an HSV tag was attached resides in the mitochondrial fraction (data not shown). Therefore, we concluded that protein YmL49 is indeed a mito-r-protein and its gene corresponds to ORF YJL096w. This protein does not show an appreciable degree of similarity to any known r-protein in the public databases.

Concluding remarks

The results described above clearly indicate that for the unequivocal identification of r-proteins, purification and biochemical characterization are essential. We believe that this is true for other genes as well. From the genomic sequence data alone, it is often very difficult, if not impossible, to assign a function to an ORF/gene even if we perform various sophisticated ciber analyses. Indeed, examples of mis-assignments have been discussed by Brenner [7] with respect to the annotation of the genomic sequence data of *My. genitalium*, which is considered to contain an essential set of genes to sustain life. The situation would become serious if the organism in question is phylogenetically only remotely related to any of the experimentally well-studied model organisms such as *E. coli* and *S. cerevisiae*.

References

1 Alm RA, LS Ling, DT Moir, *et al.* 1999. Genomic sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori*. *Nature* 397: 176–180.

2 Andersson SG, A Zomorodipour, JO Andersson, *et al.* 1998. The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. *Nature* 396: 133–140.

3 Anthony RA and SW Liebman. 1995. Alterations in ribosomal protein RPS28 can diversely affect translational accuracy in *Saccharomyces cerevisiae*. *Genetics* 140: 1247–1258.

4 Blattner FR, G Plunkett III, CA Bloch, *et al.* 1997. The complete genome sequence of *Escherichia coli* K-12. *Science* 277: 1453–1474.

5 Borodovsky M and JD McIninch. 1993. GenMark: parallel gene recognition for both DNA strands. *Comput Chem* 17: 123–133.

6 Braun HP, M Emmermann, H Mentzel and UK Schmitz. 1994. Primary structure and expression of a gene encoding the cytosolic ribosomal protein S4 from potato. *Biochim Biophys Acta* 1218: 435–438.

7 Brenner SE. 1999. Errors in genome annotation. *Trends Genet* 15: 132–133.

8 Bult CJ, O White, GJ Olsen, *et al.* 1996. Complete genome sequence of the methanogenic archaeon, *Methanococcus jannashii*. *Science* 273: 1058–1073.

9 Burggraf S, GJ Olsen, KO Stetter and CR Woese. 1992. A phylogenetic analysis of *Aquifex pyrophilus*. *Syst Appl Microbiol* 15: 353–356.

10 Cole ST, R Brosch, J Parkhill, *et al.* 1998. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* 393: 537–544.

11 Dabbs ER. 1987. Aspects of ribosomal structure and function, as revealed by mutants lacking individual ribosomal proteins. *Mol Gen (Life Sci Adv)* 6: 61–66.

12 Danchin A. 1997. Comparison between the *Escherichia coli* and *Bacillus subtilis* genomes suggests that a major function of polynucleotide phosphorylase is to synthesize CDP. *DNA Res* 4: 9–18.

13 Deckert G, PV Warren, T Gaasterland, *et al.* 1998. The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus*. *Nature* 392: 353–358.

14 Fearon K and TL Mason. 1992. Structure and function of MRP20 and MRP49, the nuclear genes for two proteins of the 54 S subunit of the yeast mitochondrial ribosome. *J Biol Chem* 267: 5162–5170.

15 Fleischmann RD, MD Adams, O White, *et al.* 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 269: 496–512.

16 Fraser CM, JD Gocayne, O White, *et al.* 1995. The minimal gene complement of *Mycoplasma genitalium*. *Science* 270: 397–403.

17 Fraser CM, S Casjens, WM Huang, *et al.* 1997. Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*. *Nature* 390: 580–586.

18 Fraser CM, SJ Norris, GM Weinstock, *et al.* 1998. Complete genome sequence of *Treponema pallidum*, the syphilis spirochete. *Science* 281: 375–388.

19 Fujita K, T Baba and K Isono. 1998. Genomic analysis of the gene encoding ribosomal proteins in eight eubacterial species and *Saccharomyces cerevisiae*. In: Genome Informatics. Universal Academy Press, Yebis, Tokyo, Japan, 1998, pp. 3–12.

20 Goffeau A, *et al.* 1997. The yeast genome directory. *Nature* 387 (suppl).

21 Graack HR and B Wittmann-Liebold. 1998. Mitochondrial ribosomal proteins (MRPs) of yeast. *Biochem J* 329: 433–488.

22 Himmelreich R, H Hilbert, H Plagens, *et al.* 1996. Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. *Nucleic Acids Res* 24: 4420–4449.

23 Isono K and S Isono 1976. Lack of ribosomal protein S1 in *Bacillus stearothermophilus*. *Proc Natl Acad Sci USA* 73: 767–770.

24 Isono K, JD McIninch and M Borodovsky. 1994. Characteristic features of the nucleotide sequences of yeast mitochondrial ribosomal protein genes as analyzed by computer program, GeneMark. *DNA Res* 1: 263–269.

25 Kalman S, W Mitchell, R Marathe, *et al.* 1999. Comparative genomes of *Chlamydia pneumoniae* and *C. trachomatis*. DDBJ/EMBL/GenBank accession no. AE001363.

26 Kaneko T, S Sato, H Kotani, *et al.* 1996. Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC6803: II. Sequence determination of the entire genome and assignment of potential protein-coding regions. *DNA Res* 3: 185–209.

27 Kawarabayasi Y, M Sawada, H Horikawa, *et al.* 1998. Complete sequence and gene organization of the genome of a hyper-thermophilic archaebacterium, *Pyrococcus horikoshii* OT3. *DNA Res* 5: 55–76.

28 Kitakawa M and K Isono. 1991. The mitochondrial ribosomes. *Biochemie* 73: 813–825.

29 Kitakawa M, HR Graack, L Grohmann, *et al.* 1997. Identification and characterization of the genes for mitochondrial ribosomal proteins of *Saccharomyces cerevisiae*. *Eur J Biochem* 245: 449–456.

30 Klenk HP, RA Clayton, JF Tomb, *et al.* 1997. The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*. *Nature* 390: 364–370.

31 Kunst F, N Ogasawara, I Moszer, *et al.* 1997. The complete genome sequence of the Gram-positive bacterium *Bacillus subtilis*. *Nature* 390: 249–256.

32 Nomura M and VA Erdmann. 1970. Reconstitution of 50S ribosomal subunits from dissociated molecular components. *Nature* 228: 744–748.

33 Pearson WR and DJ Lipman. 1988. Improved tools for biological sequence comparison. *Proc Natl Acad Sci USA* 85: 2444–2448.

34 Pennisi E. 1998. Genome data shake tree of life. *Science* 280: 672–674.

35 Smith DR, LA Doucette-Stamm, C Deloughery, *et al.* 1997. Complete genome sequence of *Methanobacterium thermoautotrophicum* ΔH: functional analysis and comparative genomics. *J Bacteriol* 179: 7135–7155.

36 Stephens RS, S Kalman, C Lammel, *et al.* 1998. Genome sequence of an obligate intracellular pathogen of humans: *Chlamydia trachomatis*. *Science* 282: 754–759.

37 Tomb JF, O White, AR Kerlavage, *et al.* 1997. The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature* 388: 539–547.

38 The *C. elegans* Sequencing Consortium. 1998. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. 282: 2012–2018.

39 Wittmann-Liebold B, AKE Köpke, E Arndt, *et al.* 1986. Sequence comparison and evolution of ribosomal proteins and their genes. In: The Ribosome: Structure, Function and Evolution. *Am Soc Microbiol*, Washington, DC, pp. 598–616.