Hypothesis

In silico analysis of the tRNA:m1A58 methyltransferase family: homology-based fold prediction and identification of new members from Eubacteria and Archaea

Janusz M. Bujnicki*

Bioinformatics Laboratory, International Institute of Molecular and Cell Biology, ul. ks. Trojdena 4, 02-109 Warsaw, Poland

Received 14 August 2001; revised 20 September 2001; accepted 20 September 2001

First published online 5 October 2001

Edited by Gunnar von Heijne

Abstract The amino acid sequences of Gcd10p and Gcd14p, the two subunits of the tRNA:(1-methyladenosine-58; m^1A58) methyltransferase (MTase) of *Saccharomyces cerevisiae*, have been analyzed using iterative sequence database searches and fold recognition programs. The results suggest that the 'catalytic' Gcd14p and 'substrate binding' Gcd10p are related to each other and to a group of prokaryotic open reading frames, which were previously annotated as hypothetical protein isoaspartate MTases in sequence databases. It is predicted that the prokaryotic proteins are genuine tRNA:m^1A MTases based on similarity of their predicted active site to the Gcd14p family. In addition to the MTase domain, an additional domain was identified in the N-terminus of all these proteins that may be involved in interaction with tRNA. These results suggest that the eukaryotic tRNA:m^1A58 MTase is a product of gene duplication and divergent evolution of a possibly homodimeric prokaryotic enzyme.

Key words: Methyltransferase; 1-Methyladenosine; RNA modification; Molecular evolution; Protein structure prediction

1. Introduction

The majority of structurally characterized *S*-adenosylmethionine (AdoMet)-dependent methyltransferases (MTases) adopt a very similar three-dimensional fold (comprehensively reviewed in [1]). Nevertheless, apart from several motifs delineating the AdoMet-binding face of the common catalytic domain, MTases show minimal sequence conservation between families and therefore are difficult to analyze using conventional searches alone [1–3]. In distinct families the catalytic domain is usually fused to unrelated auxiliary domains that participate in target recognition and binding. Thus, even if the sequence similarities help to pinpoint the likely substrate of a given enzyme (e.g. a nucleic acid base, a protein or some low molecular weight (MW) molecule), the precise prediction of reaction specificity is usually not possible, unless a sufficiently closely related homolog of known function is found [3].

This problem is particularly evident for MTases modifying nucleic acids, especially RNA. A large number of distinct enzymes are present in the cell, catalyzing similar reactions but in different classes of RNA or at different locations in an RNA molecule. The methylated nucleotides are believed to play key roles in the function of the ribonucleoprotein particles in vivo, influencing processes such as maturation of various pre-RNAs, stabilizing the assembly and transport of ribosomes and spliceosomes, and modulating splicing and protein synthesis (reviewed in [4]). However, only a very limited number of RNA MTases have been identified and characterized to date. To aid in characterization of novel RNA-modifying enzymes, we and others have been searching sequence databases and carrying out phylogenetic inference and, in certain cases, molecular modeling. This effort resulted in the identification of candidates for novel RNA MTases and correlation of these sequence data with known enzymatic activities [3,5–11]. Still, many RNA MTase families remain unidentified.

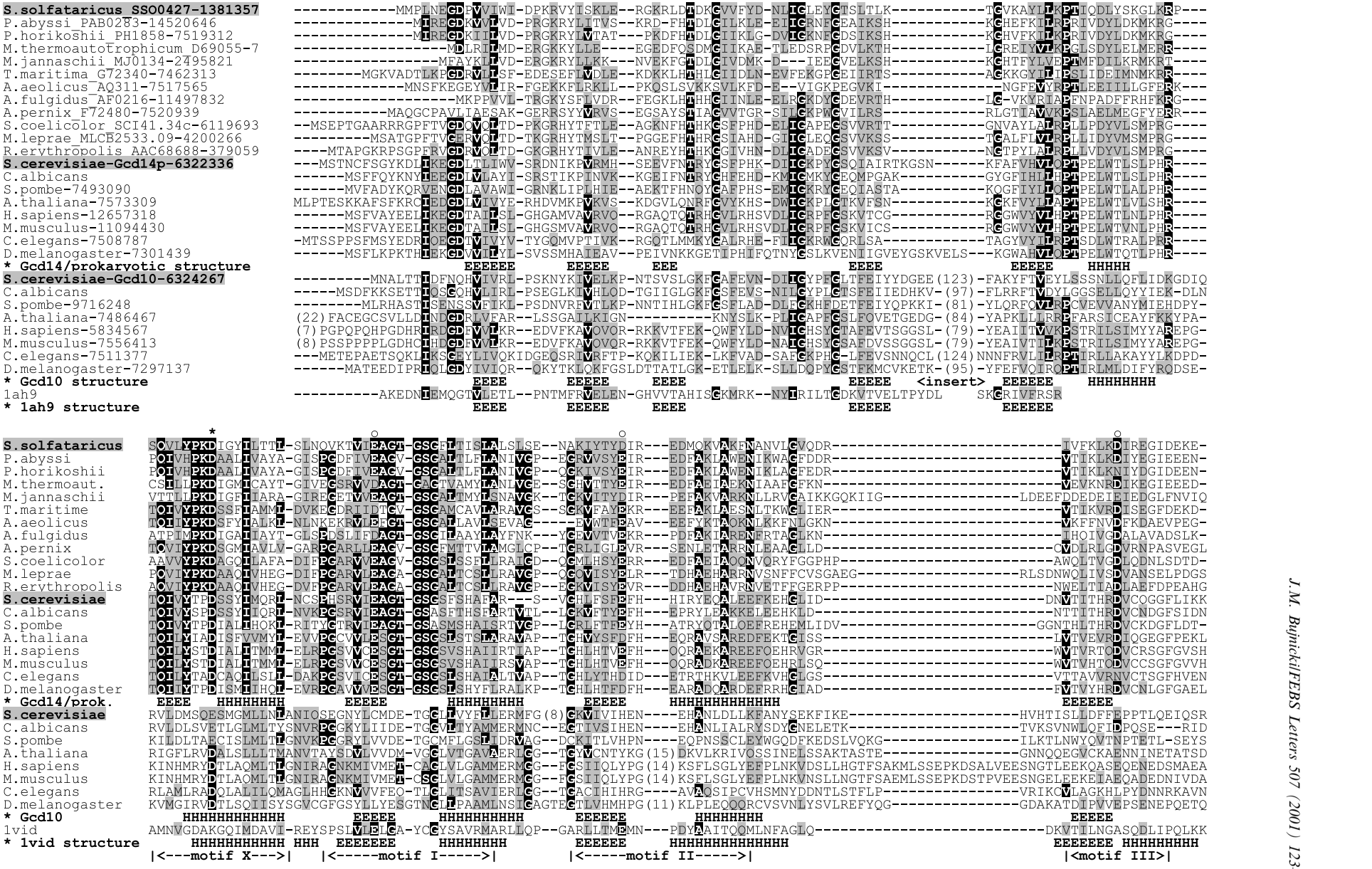
The occurrence of 1-methyladenosine (m^1A) at position 58 of the TΨC loop has been reported in tRNAs from all three domains of life, suggesting an evolutionarily conserved role in tRNA structure or function [12]. In the yeast *Saccharomyces cerevisiae*, the two-subunit tRNA:m^1A58 MTase Gcd10p/Gcd14p has been found to be essential for cell growth under normal conditions [13]. It has been shown that Gcd14p is required for the catalytic activity and AdoMet binding, in agreement with the presence of the typical AdoMet-binding motifs, while Gcd10p is essential for tRNA binding [14]. In tRNA, m^1A has been also found at position 9 (in eukaryotic mitochondria), 14 (in Eukaryota) and 22 (in few prokaryotes) (reviewed in [4]), but the genes encoding the MTases responsible for these modifications have not been identified to date.

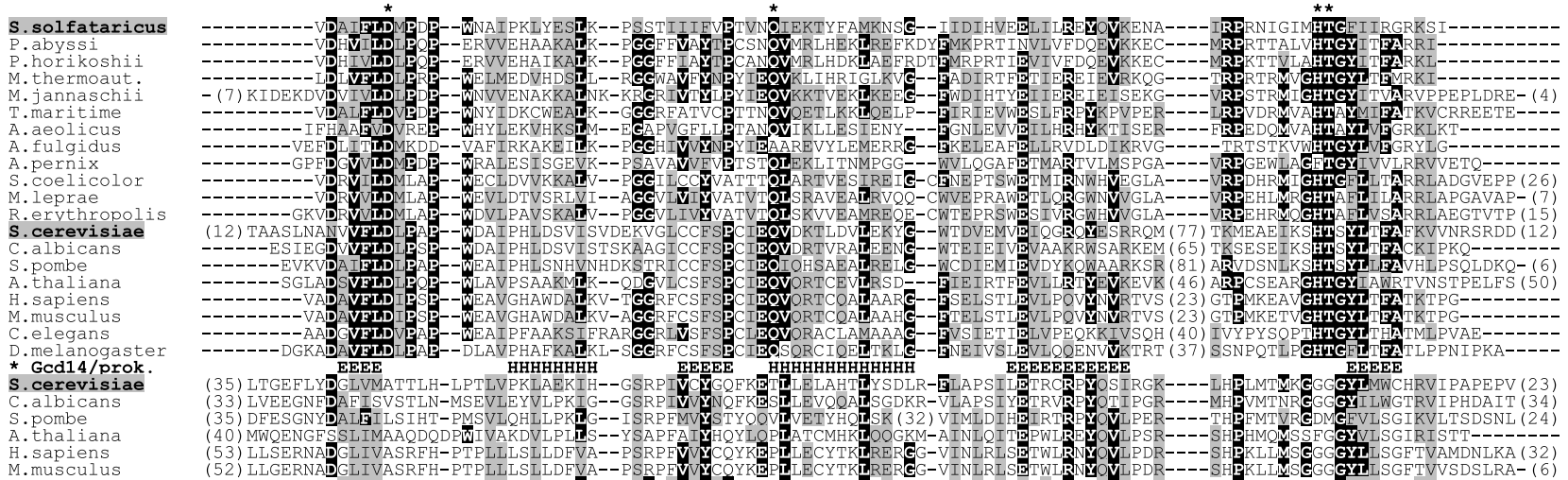
2. Sequence analysis and structure prediction

As a part of a systematic in silico analysis of the MTase superfamily, the non-redundant database at the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov) was searched using sequences of Gcd10p and Gcd14p as queries. Seeding the PSI-BLAST program [15] with Gcd10p (with the expectation (*e*) value profile inclusion threshold of 10^{-3}) retrieved no homologs with significant *e*-values besides its obvious eukaryotic orthologs. The first non-Gcd10 hit was the open reading frame SSO0427 from *Sulfolobus solfataricus*, reported with the *e*-value of 0.061 at first detection. However,

*Fax: (48)-22-668 5288.
E-mail address: info@bioinfo.pl (J.M. Bujnicki).

PII: S0014-5793(01)02962-3


A
S.solfataricus_SSO0427-1381357
P.abyssi_PAB0283-14520646
P.horikoshii_PH1858-7519312
M.thermoautotrophicum_D69055-7
M.jannaschii_MJ0134-2495821
T.maritima_G72340-7462313
A.aeolicus_AQ311-7517565
A.fulgidus_AF0216-11497832
A.pernix_F72480-7520939
S.coelicolor_SCI41.34c-6119693
M.leprae_MLCB2533.09-4200266
R.erythropolis_AAC68688-379059
S.cerevisiae-Gcd14p-6322336
C.albicans
S.pombe-7493090
A.thaliana-7573309
H.sapiens-12657318
M.musculus-11094430
C.elegans-7508787
D.melanogaster-7301439
* Gcd14/prokaryotic structure
S.cerevisiae-Gcd10-6324267
C.albicans
S.pombe-9716248
A.thaliana-7486467
H.sapiens-5834567
M.musculus-7556413
C.elegans-7511377
D.melanogaster-7297137
* Gcd10 structure
1ah9
* 1ah9 structure
S.solfataricus
P.abyssi
P.horikoshii
M.thermoaut.
M.jannaschii
T.maritime
A.aeolicus
A.fulgidus
A.pernix
S.coelicolor
M.leprae
R.erythropolis
S.cerevisiae
C.albicans
S.pombe
A.thaliana
H.sapiens
M.musculus
C.elegans
D.melanogaster
* Gcd14/prok.
S.cerevisiae
C.albicans
S.pombe
A.thaliana
H.sapiens
M.musculus
C.elegans
D.melanogaster
* Gcd10
1vid
* 1vid structure
|<---motif X--->|
|<-----motif I----->|
|<-----motif II----->|
|<motif III>|



B

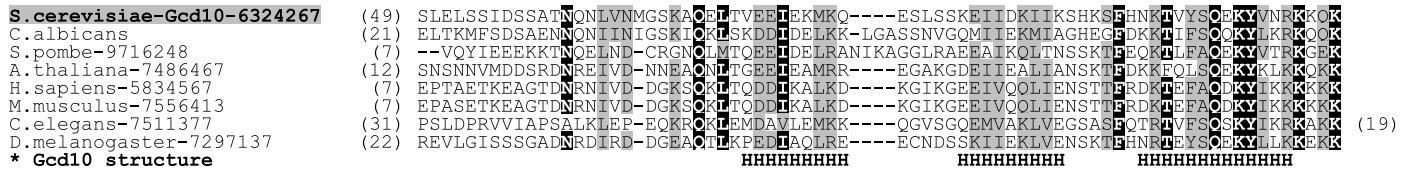


Fig. 1. A: Conserved sequence blocks in the alignment of the tRNA:m1A58 MTase family. B: The small domain present only in the Gcd10p subunit. Conserved motifs are labeled according to [1]. Highly conserved residues are shown on the black background, the residues with invariant physicochemical character (hydrophobic, small etc.) are shown on the gray background. The residues predicted to bind the cofactor and participate in catalysis in the 'active' subunits are indicated by 'o' and '*', respectively. The predicted secondary structure is shown separately for the sequences of proteins from the Gcd14p and Gcd10p subfamilies. The sequences of proteins suggested by threading as structural templates are shown below the alignment with the respective Protein Data Bank accession numbers: the translation initiation factor IF1 (1ah9) representing the OB-fold and the catechol *O*-MTase (1vid) representing the MTase fold.

using Gcd14p as a query with the same parameters resulted in a multitude of hits to various MTases. Using a stringent *e*-value threshold of 10^{-20} resulted in retrieval of eukaryotic orthologs of Gcd14p and a family of uncharacterized MTases from Eubacteria and Archaea that are annotated as putative protein isoaspartyl MTases (PIMT). These sequences showed significant *e*-values on first detection (for instance 10^{-15} for the *S. solfataricus* SSO0427) and retained the conservation pattern typical of the Gcd14p proteins (Fig. 1) rather than that of the PIMT family [16]. In addition to this, reciprocal searches initiated with the newly retrieved prokaryotic sequences detected Gcd14p with higher *e*-values (10^{-24}) than the bona fide PIMT family members (10^{-12}).

The weak link provided by the SSO0427 sequence suggested that the Gcd10p and Gcd14p families might be related. Therefore, to identify the homologs of Gcd10p among structurally similar proteins, its sequence was submitted to the Structure Prediction Meta Server [17,18] (http://bioinfo.pl/meta/). Threading results revealed that the Gcd10p sequence is compatible with the MTase fold, according to the first hits reported by FFAS [19] (high score 14.14) and other servers (albeit with low scores). Querying the Meta Server with the sequence of *Candida albicans* Gcd10p resulted in even higher scores reported by FFAS (19.38), mGenTHREADER [20] (100%), and INBGU [21] (15.5%). The presence of putative MTase motifs in Gcd14p has been reported previously [14] and its similarity to MTases was confirmed by threading (scores: FFAS-28.9, mGenTHREADER-100%, INBGU-43.5). The presence of the common motifs in Gcd10p and Gcd14p sequences was confirmed using the Gibbs sampling procedure [22]; motifs VI, VIII and the carboxy-proximal part of motif I were detected with probability of occurring by chance $<10^{-12}$). A good correlation of secondary structure elements predicted independently for the Gcd10p and Gcd14p families was discovered using PSIPRED [23] (Fig. 1). Taken together, these findings strongly suggest that the two subunits of the eukaryotic tRNA:m^{1}A58 MTase evolved from a common ancestor, presumably from a duplicated prokaryotic-type gene.

The multiple sequence alignment revealed a conserved extension to the N-terminus of the common MTase domain. Threading analysis (scores: INBGU-11.8, FFAS-5.42) as well as secondary structure prediction suggests that the sequence and structure of this region is compatible with the OB-fold, characteristic for domains often found in proteins interacting with nucleic acids [24,25]. Interestingly, there is an additional small domain inserted into the N-terminal domain in the Gcd10p family (Fig. 1B). In many RNA MTase families, the predicted RNA-binding domains are found N-terminally of the catalytic domain [3,8–10,26–29]. A strong, analogous hypothesis is that the predicted N-terminal domain of tRNA:m^{1}A58 MTases is involved in target recognition by these enzymes. It will be interesting to determine what role is played by the non-identical N-terminal domains in each of the subunits of the heterodimeric eukaryotic enzyme and by the additional domain in the Gcd10p subunit.

3. Prediction of functionally important residues

Predictions based on tertiary fold data and multisequence alignment indicate that invariant or nearly invariant residues E^{116}, E^{139}, and D^{168} (numbering for *S. cerevisiae* Gcd14p) are involved in cofactor binding, while residues D^{98}, D^{203}, Q^{235}, H^{354}, and T^{355} are involved in catalysis. These residues are conserved in the Gcd14p family and in prokaryotic proteins, strongly suggesting that the latter are not paralogs of PIMT (which uses different residues for catalysis [16]), but tRNA:m^{1}A58 MTases. However, they are absent from Gcd10p, which is consistent with the role of Gcd14p being the catalytic subunit.

4. Conclusions

I predict that the two-subunit yeast tRNA:m^{1}A58 MTase evolved from a duplicated, possibly dimeric enzyme, in which one of the subunits lost the ability to bind tRNA on its own and in the other the catalytic site degenerated. The hypothesis of the dimeric structure of prokaryotic enzymes is in good agreement with the experimental estimation of the MW for tRNA:m^{1}A58 MTases from *Thermus flavus* [30] and *Thermus thermophilus* [31] (78 kDa and 60 kDa, respectively). No sequences of Gcd14p homologs are available from these organisms, however the calculated MW for all sequences of predicted prokaryotic tRNA:m^{1}A58 MTase analyzed in this work are in the range of 27.8 kDa (*Aeropyrum pernix*) to 35.2 kDa (*Rhodococcus erythropolis*).

Analogous evolutionary scenarios have been suggested for eukaryotic and prokaryotic tRNA splicing enzymes and tRNA adenosine deaminases. It is believed that the heterotetrameric yeast Sen nuclease evolved from the homotetrameric archaeal EndA enzyme, in which the structurally identical subunits have non-equivalent roles in tRNA binding and catalysis [32,33]. Similarly, the heterodimeric yeast Tad2p/Tad3p adenosine-34 deaminase has a single counterpart in *Escherichia coli* [34]. It seems that replacement of homomultimeric tRNA modification enzymes with heteromultimers encoded by duplicated and diverged genes is a common evolutionary mechanism in eukaryotic organisms. Since the tRNA molecule is asymmetric, the use of different subunits in the multimeric enzyme probably allows more flexible optimization of the binding surface and provides efficient means to circumvent the problem of concerted amino acid substitutions in structurally identical, but functionally non-equivalent subunits of the homomultimeric proteins. Although the detailed modeling of mechanism of substrate binding and catalysis or the possible mode of subunit interaction in tRNA:m^{1}A58 MTases is beyond the limits of the present study, I hope that this analysis will facilitate further structural, functional, and phylogenetic studies of this protein family and its relationship to other MTases.

Acknowledgements: I would like to thank Richard Leach for comments on the manuscript. This work was supported by the Polish State Committee for Scientific Research (KBN, Grant 8T11-F01019).

References

[1] Fauman, E.B., Blumenthal, R.M. and Cheng, X. (1999) in: S-Adenosylmethionine-dependent methyltransferases structures and functions (Cheng, X. and Blumenthal, R.M., Eds.), pp. 1–38, World Scientific, Singapore.
[2] Bujnicki, J.M. (1999) Silico Biol., 1, 1–8, http://www.bioinfo.de/isb/1999-01/0016/.
[3] Bujnicki, J.M. (2001) Sequence-structure-function relationships in evolution of nucleic acid modification enzymes, Ph.D. thesis, Warsaw University, Warsaw.

[4] Grosjean, H. and Benne, R. (1998) Modification and editing of RNA, ASM Press, Washington, DC.
[5] Gustafsson, C., Reid, R., Greene, P.J. and Santi, D.V. (1996) Nucleic Acids Res. 24, 3756–3762.
[6] Reid, R., Greene, P. and Santi, D.V. (1999) Nucleic Acids Res. 27, 3138–3145.
[7] Motorin, Y. and Grosjean, H. (1999) RNA 5, 1105–1118.
[8] Bujnicki, J.M. (2000) FASEB J. 14, 2365–2368.
[9] Bujnicki, J.M. and Rychlewski, L. (2000) Acta Microbiol. Pol. 49, 253–260.
[10] Bujnicki, J.M. and Rychlewski, L. (2001) Acta Microbiol. Pol. 50, 7–17.
[11] Bujnicki, J.M., Feder, M., Radlinska, M. and Rychlewski, L. (2001) BMC Bioinform. 2, 2.
[12] Bjork, G.R., Ericson, J.U., Gustafsson, C.E., Hagervall, T.G., Jonsson, Y.H. and Wikstrom, P.M. (1987) Annu. Rev. Biochem 56, 263–287.
[13] Anderson, J., Phan, L., Cuesta, R., Carlson, B.A., Pak, M., Asano, K., Bjork, G.R., Tamame, M. and Hinnebusch, A.G. (1998) Genes Dev. 12, 3650–3662.
[14] Anderson, J., Phan, L. and Hinnebusch, A.G. (2000) Proc. Natl. Acad. Sci. USA 97, 5173–5178.
[15] Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Nucleic Acids Res. 25, 3389–3402.
[16] Skinner, M.M., Puvathingal, J.M., Walter, R.L. and Friedman, A.M. (2000) Struct. Fold. Des. 8, 1189–1201.
[17] Bujnicki, J.M., Elofsson, A., Fischer, D. and Rychlewski, L. (2001) Bioinformatics 17, 750–751.
[18] Lundstrom, Rychlewski, L., Bujnicki, J.M. and Elofsson, A. (2001) Protein Sci, in press.
[19] Rychlewski, L., Jaroszewski, L., Li, W. and Godzik, A. (2000) Protein Sci. 9, 232–241.
[20] Jones, D.T. (1999) J. Mol. Biol. 287, 797–815.
[21] Fischer, D. (2000) Pac. Symp. Biocomput., 119–130.
[22] Neuwald, A.F., Liu, J.S., Lipman, D.J. and Lawrence, C.E. (1997) Nucleic Acids Res. 25, 1665–1677.
[23] McGuffin, L.J., Bryson, K. and Jones, D.T. (2000) Bioinformatics 16, 404–405.
[24] Murzin, A.G. (1993) EMBO J. 12, 861–867.
[25] Draper, D.E. and Reynaldo, L.P. (1999) Nucleic Acids Res. 27, 381–388.
[26] Aravind, L. and Koonin, E.V. (1999) J. Mol. Evol. 48, 291–302.
[27] Anantharaman, V., Koonin, E.V. and Aravind, L. (2001) FEMS Microbiol. Lett. 197, 215–221.
[28] Aravind, L. and Koonin, E.V. (2001) Trends Biochem. Sci. 26, 215–217.
[29] Bujnicki, J.M., Blumenthal, R.M. and Rychlewski, L. (2001) J. Mol. Microbiol. Biotechnol., in press.
[30] Morozov, I.A., Gambaryan, A.S., Lvova, T.N., Nedospasov, A.A. and Venkstern, T.V. (1982) Eur. J. Biochem. 129, 429–436.
[31] Yamazaki, N., Hori, H., Ozawa, K., Nakanishi, S., Ueda, T., Kumagai, I., Watanabe, K. and Nishikawa, K. (1994) Biosci. Biotechnol. Biochem. 58, 1128–1133.
[32] Li, H., Trotta, C.R. and Abelson, J. (1998) Science 280, 279–284.
[33] Bujnicki, J.M. and Rychlewski, L. (2000) FEBS Lett. 486, 328–329.
[34] Gerber, A.P. and Keller, W. (2001) Trends Biochem. Sci. 26, 376–384.