E. coli translation initiation factor IF2 – an extremely conserved protein. Comparative sequence analysis of the *infB* gene in clinical isolates of *E. coli*

Søren A. de A. Steffensen, Ane B. Poulsen[1], Kim K. Mortensen, Hans U. Sperling-Petersen*

Department of Biostructural Chemistry, Institute of Molecular and Structural Biology, Aarhus University, Gustav Wieds Vej 10, DK-8000 Aarhus C, Denmark

Received 17 October 1997

Abstract The functionally uncharacterised N-terminal of translation initiation factor IF2 has been found to be extremely variable when comparing different bacterial species. In order to study the intraspecies variability of IF2 the 2670 basepairs nucleotide sequence of the *infB* gene (encoding IF2) was determined in 10 clinical isolates of *E. coli*. The N-terminal domains (I, II and III) were completely conserved indicating a specific function of this region of IF2. Only one polymorphic position was found in the deduced 890 amino acid sequence. This Gln/Gly490 is located within the central GTP/GDP-binding domain IV of IF2. The results are further evidence that IF2 from *E. coli* has reached a highly defined level of structural and functional development.

Key words: Translation initiation; Initiation factor IF2; *infB*; Biodiversity; Population genetics; *Escherichia coli*

1. Introduction

The initiation of mRNA translation in prokaryotes is promoted by 3 proteins, initiation factors IF1, IF2 and IF3. Of these, IF2 is the largest (molecular mass 97.3 kDa in *E. coli*). IF2 interacts with at least 3 components during the initiation: GTP, fMet-tRNA$_f^{Met}$ and ribosomes. Through these interactions IF2 initiates the binding of fMet-tRNA$_f^{Met}$ to the 70S ribosome in a process involving hydrolysis of GTP to GDP. In *E. coli* IF2 is expressed in three natural forms: IF2α, IF2β and IF2γ. IF2α, consisting of 890 amino acid residues, is the largest form. IF2β and IF2γ result from two internal in-frame initiation sites in the *infB* gene encoding IF2, and are thus identical to IF2α at the C-terminal and lack the N-terminal 157 and 164 amino acid residues respectively [1]. A six domain structural model has been proposed for this multi-functional factor [2] in which domain I constitutes the difference between IF2α and IF2β. The GTP-binding domain is located within domain IV and the C-terminal domain VI has been shown to be involved in the tRNA binding [3]. Domain V probably supports the functions of domains IV and VI. No specific activity has been assigned to domains I, II and III. The nucleotide sequence of *infB* was originally determined by Grunberg-Manago and coworkers [4] and recently an identical *infB* sequence of *E. coli* K-12 was obtained as part of the *E. coli* Genome Project [5]. The *infB* sequence has been obtained from a number of other different bacterial species. These sequences, accessible from the nucleotide databases, are homologous at the 3′-end/C-terminal, but vary considerably in the 5′-end/N-terminal in both length and composition. Of these other sequenced species only *Bacillus subtilis* has multiple forms of IF2, i.e. both an α- and a β-form [6]. In our investigations of the population genetics of *infB* in *E. coli*, we have sequenced the complete *infB* gene (2670 bp) in 10 clinical isolates and obtained partial sequences of additionally 36 clinical and animal *E. coli* isolates. The clinical isolates have previously been characterised for their activity of the outer membrane protease OmpT [7]. Through the work described here, it was our aim to gain additional clues to the function of the N-terminal domains (I, II and III) of IF2α, and to point out additional important sites in the protein.

2. Materials and methods

40 wild-type *E. coli* strains were isolated from patient samples of urine, blood and wounds at the Department of Clinical Microbiology, Aalborg Hospital, Aalborg, Denmark and named numerically from EcoAU9301 to EcoAU9340. Another 6 *E. coli* strains were isolated from animal (elephant, pig, chicken and bull) feaces samples from Thailand and Spain and named EcoAU9501, EcoAU9502, EcoAU9503, EcoAU9510, EcoAU9511 and EcoAU9514. The *infB* sequence of *E. coli* strain K-12 was used as reference for detecting variations. The K-12 *infB* sequence was also used for designing primers for PCR and sequencing. In this article we assign base number one to the adenine of the initiating ATG codon of IF2α.

The *infB* gene was amplified by PCR performed directly on cell colonies picked from an agar plate. Different primer sets were applied. If necessary the PCR products were purified from excess primers and nucleotides using the method described in [8] before serving as template for cyclesequencing. PCR was performed with Thermoprime Plus DNA polymerase (Advanced Biotechnologies, UK) in the buffer supplied. Cyclesequencing was performed with 5′-flourescein labeled primers using the ThermoSequenase fluorescent labeled primer cyclesequencing Kit (Amersham Life Science, UK). By this approach, cloning steps were avoided thereby eliminating the risk of polymerase errors. The sequencing reactions were analysed on an A.L.F. DNA Sequencer (Pharmacia Biotech, Sweden). The same strand was sequenced multiple times and in the case of amino acid changes and poorly resolved regions both strands were sequenced.

3. Results and discussion

3.1. Nucleotide polymorphism

The translation initiation factor IF2 is known to be one of the most important proteins in *E. coli*. We chose to sequence the gene for IF2, *infB*, in a number of natural *E. coli* isolates for several reasons. The amino acid sequences of most proteins which have been studied in different isolates have shown considerable variation. This seemed a priori to be unlikely for

*Corresponding author. Fax: +45 (86) 18 28 12.
E-mail: husp@kemi.aau.dk

[1] Present address: Statens Serum Institut, Artillerivej 5, DK-2300 København S, Denmark.

0014-5793/97/$17.00 © 1997 Federation of European Biochemical Societies. All rights reserved.
PII S0014-5793(97)01472-5

IF2, since several specific amino acid mutations were found to be lethal for the cell [9]. The amino acid sequence of IF2 from several bacterial species is known and a characteristic feature has appeared. While the C-terminal two third of the *E. coli* molecule seems to be homologous in all strains, the N-terminal part is highly variable. We have found that IF2 in numerous *E. coli* isolates have identical sizes when analysed by SDS-PAGE (data not shown). However these results gave no information about conservation or not of the amino acid sequence and no such data are known.

In the following, the *infB* sequence of *E. coli* strain K-12 [5] will be used as reference, and sequence polymorphisms will be referred to as variations as compared to the reference sequence. The sequence data are presented as variations only (Fig. 1).

In 10 *infB* sequences from EcoAU9301 to EcoAU9310 no deletions or insertions were observed in the 2670 basepair coding sequence. A total of 66 polymorphic positions were found, of which the 3 nucleotides of codon number 490 lead to the only amino acid variation found: Gln/Gly490. The rest of the polymorphic positions were due to synonymous variations. Three of the 10 sequences, EcoAU9303, EcoAU9304 and EcoAU9305, showed no variations at all, and EcoAU9308 showed only a single variation. Strains EcoAU9306 and EcoAU9310 were identical, displaying 40 variations. EcoAU9302, EcoAU9307 and EcoAU9309 only differed from each other in a total of 4 positions, but differed from the reference sequence in 38 to 41 positions. Strain EcoAU9301 showed 30 polymorphic positions, of which 12 were restricted to this strain, or 18% of the total number of polymorphic positions. Of the 66 polymorphic positions only one (position 2109) showed more than two possible nucleotides. Sequence divergence between pairs of strains was up to 1.7% of nucleotide positions, with an average of 1.0%, and an average of 0.06% of amino acid positions.

3.2. Recombinational patterns in infB

From Fig. 1 it is seen that a cluster of variations is situated between position 1368 and 1638 in EcoAU9306, EcoAU9310, EcoAU9302, EcoAU9307 and EcoAU9309. This cluster, which includes the substituted codon leading to the amino acid change, is shared by 5 of the 10 strains. The relatively high frequency of variations in the limited region from position 1368 to 1638 (10% of the gene holds approximately 40% of the variations in these isolates), shows that this segment probably has been introduced by a homologous recombinational contact with a more distantly related cell. The described cluster is also part of a so-called 'mosaic structure', which reveals that at least one other recombination certainly has taken place. The mosaic structure consists of the 'underlined' and 'italic' variations in Fig. 1. Both donor lineages and the recipient lineage are represented, as will be described here. It is seen that the lineage represented by EcoAU9301 has at some time during evolution 'donated' the 'italic' variations to the lineage represented by the group of EcoAU9302, EcoAU9307 and EcoAU9309, resulting in the lineage represented by EcoAU9306. EcoAU9306 is thus composed of both 'underlined' and 'italic' variations together with variations introduced both before and after the recombination.

From Fig. 1 it is seen that 'italic' variations in EcoAU9306 are interrupted by two 'underlined' variations, separating the variation in position 2652 from the other 'italic' variations. This can be explained in three ways: (i) the single 'italic' variation in position 2652 has arisen independently in EcoAU9301 and EcoAU9306, (ii) it has been introduced by a second recombinational event between the same two lineages, or (iii) two segments have been incorporated simultaneously in the same recombinational contact. The first two possibilities are more or less statistical improbable, for which reason the latter possibility is in favour. Mechanistically, the latter suggestion of segmental recombination is interesting because the two known main modes of genetic exchange in bacteria is conjugation and transduction, which involves molecules of ≥100 kb [10,11]. It has recently been shown in transduction experiments in *E. coli*, that these large DNA regions indeed can be exchanged as a string of smaller segments,

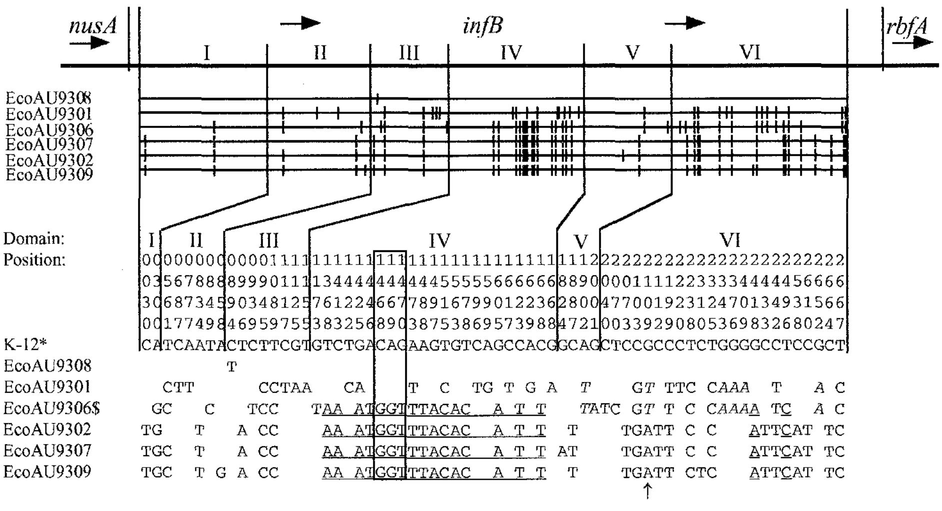

Fig. 1. The results of sequencing *infB* in 10 clinical *E. coli* isolates are shown as substitutions of the reference sequence from *E. coli* strain K-12. At the top the physical positions in *infB* of the polymorphic positions are indicated and at the bottom the actual variations are listed. A position with two different substitutions is indicated with an arrow. The boxed sequence shows codon no. 490 which specifies Gln (CAG) or Gly (GGT). The underlined and italic variations are participants of a mosaic structure. *: EcoAU9303, EcoAU9304 and EcoAU9305 are identical to K-12. $: EcoAU9310 is identical to EcoAU9306.

probably due to restriction modification of the entrant molecule [11]. Our observations are thus in support of this mechanism.

3.3. Analysis of the central region of infB encoding the GTP-binding domain

The observation of an amino acid substitution in the G-domain incited us to perform further sequencing of the recombinational region harbouring the amino acid substitution (position 1368 to 1638, data not shown) in additionally 36 *E. coli* isolates. Thirty of these isolates were, as described in Section 2, clinical isolates from Denmark, as were the 10 isolates in which the complete *infB* gene was sequenced. The remaining 6 isolates were from faeces samples of wild living animal sources from Thailand and Spain. According to the nucleotide sequences the total of 47 isolates, including the laboratory strain K-12, could be divided into three discrete groups. Group I consisted of 1907 isolates with sequences identical or similar to the reference strain K-12. Group II consisted of 10 isolates with sequences similar to EcoAU9301. In group I and II the main part of the sequences were diversified by single variations. Group III consisted of 20 isolates, which all had a variational pattern identical to the one found between position 1368 and 1638 in EcoAU9302, EcoAU9306, EcoAU9307 and EcoAU9309 (Fig. 1). Thus they also contained the amino acid substitution. The fact that all the sequences of group III were identical, indicates that this gene segment has been incorporated recently into *E. coli infB*, as it has not (yet) retained any additional single variations. Furthermore, this segment has spread to almost half of the investigated sample, indicating it to be evolutionarily successful.

Geographically, it was found that the 6 isolates from Thailand and Spain were grouped together with the 0 danish isolates, distributed with 5 in group I and 1 in group III. Although the data is limited, this shows that the same *E. coli* clones can be expected to be distributed world wide.

3.4. Non-synonymous substitution

The only amino acid variation detected when deducing the amino acid sequences from the nucleotide sequences was residue 490: glutamine to glycine. For this residue the entire codon (position 1468–1470) was changed (CAG to GGT). The amino acid substitution as compared to K-12 was found in 20 of the 47 strains sequenced, which is 43% of the isolates analysed. The amino acid residue 490 is located in the GTP-binding domain of IF2, which makes it possible to compare with not only sequences of *infB* in other bacteria but also with the homologous sequences of GTP-binding domains of other translation factors: EF-Tu, EF-G and RF-3. Position 490 corresponds to position 126 in EF-Tu from *E. coli*, which holds an interspeciesly conserved glycine. Gly126 in *E. coli* EF-Tu is positioned in the λ7-loop and outside the GTP-binding pocket. Nevertheless, mutation of this residue to an alanine reduced the in vitro difference in affinities for GDP over GTP 40 times. Gly126 is therefore thought to have some structural importance for the transition between the GDP and GTP conformations of EF-Tu [12]. However, the corresponding positions in IF2, EF-G and RF-3 hold various residues in the different sequenced species, as determined from aligned sequences accessible from databases. In IF2 though, glycine is often found at this position. The indication we found, that IF2 from *E. coli* successfully has adapted a glycine in the corresponding position, supports that IF2 will undergo a structural transition upon GTP hydrolysis similar to the one in EF-Tu.

3.5. 3.4. The intraspecific conservation of infB/IF2

In general, the IF2 molecule is very conserved within the different *E. coli* strains, and both the internal β- and γ-initiation sites were found to be conserved too. There was only one amino acid change in 5 of 11 gene products of 890 amino acid residues, corresponding to an average pairwise difference between sequences of 0.06%. This is approximately 2, 5, 5 and 20 times lower than has been reported for the *gapA*, *putB*, *mdh* and *gnd* gene products respectively in other intraspecific studies of *E. coli* [13–16]. These genes respectively encode glyceraldehyde-3-phosphate dehydrogenase, proline permease, malate dehydrogenase and 6-phosphogluconate dehydrogenase. At the nucleotide level the same tendency is observed, as the average pairwise difference between *infB* sequences was 1.0% of the number of positions. This is 0.2, 2.4, 1.1 and 4.3 times lower than reported for the *gapA*, *putB*, *mdh* and *gnd* genes respectively [13–16]. Overall *infB* holds very few non-synonymous variations per polymorphic position, which indicates a high evolutionary pressure on the initiation factor IF2. The 5′-end of the *infB* gene is extremely conserved. The part of the gene encoding the first two domains (nucleotide residues 1 to 867) holds 8 polymorphic positions i.e. 0.9% of the positions, whereas the rest of the gene holds 3.2% polymorphic positions. This might indicate an evolutionary pressure, not only at the protein level, but also at the DNA/mRNA level.

3.6. N-terminal function

As a key factor in the translation mechanism, IF2 is expected to be under a strong selective constraint for amino acid substitutions at least in the domains important for the binding and hydrolysis of GTP, recognition of fMet-tRNA$_f^{Met}$ and binding to ribosome, that is domains IV, V and VI.

As described we find no amino acid variations in the N-terminal domains I, II and III of IF2 from the 10 strains sequenced. We found this result very surprising considering the extreme interspecies variability of the known N-terminal sequences. Furthermore, it has been shown that *E. coli* cells harbouring a truncated form of IF2 lacking domains I, II and III proliferate as good as wild-type cells at temperatures above 40°C, indicating that the functions of the first three domains of IF2 are not strictly essential to the cell [17]. However, it has also been shown that cells expressing only IF2α or IF2β/γ are not viable at 30°C [18]. This indicates that the cells do need both IF2α and IF2β/γ under non-optimal conditions. It could therefore be argued that the N-terminal region must play an important role for the cell in some part of the life cycle. The two-habitat aspect of *E. coli*, with a part of the life cycle outside the intestinal environment, might be very important in explaining the role of the N-terminal domains. It is possible that these domains have no critical function when the cell experiences optimal growth conditions, while the functions of the domains are only important under harsh conditions in the extraintestinal environment. We are currently sequencing *infB* from other enteric species to gain additional information of a possible correlation of the N-terminal function and the life cycle.

3.7. Conclusion

We have studied the sequence of *infB*/IF2 within a broad sample material of the *E. coli* population. The results indicate that the N-terminal region of initiation factor IF2 is important to the cell since the presence of the initiation sites for the three natural forms of IF2 is conserved in the 10 strains analysed, and no amino acid substitutions or deletions were found in the N-terminal region. This further supports our hypothesis that the domains I, II and III have a differentiated importance for the cell in specific parts of the life cycle. It was shown that IF2 is an extremely conserved protein in *E. coli*, with only one amino acid substitution detected in the 890 residue protein. We found that the amino acid substitution most likely is attributable to a recent recombinational replacement of a segment of *infB*. The nature of the substitution, Gln90 to Gly, further correlates the functions of the G-domains of IF2 and EF-Tu, as an interspeciesly conserved Gly is found at the corresponding position in EF-Tu.

The absence of other amino acid variations together with other researchers observations that most amino acid substitutions studied were lethal for the cell, show the crucial importance of IF2 for the *E. coli* cell and indicate that this protein has reached a highly defined level of structural and functional development.

In our ongoing research, we are aiming at crystallising intact IF2 as well as isolated IF2 domains for more detailed structural studies. The function of the IF2 N-terminal part is still a puzzle. We attempt to study this aspect by approaching natural life conditions for *E. coli* in our studies. This together with structural data hopefully soon will shed light on some – so far unknown – function of initiation factor IF2 in *E. coli*.

Acknowledgements: The authors want to thank Birgitte Korsager and Lis Hyttel for their help in the isolation of the strains used and Jotun Hein for valuable discussions. This work was supported by Familien Hede Nielsens Fond, the Danish Biotechnological Research and Development Programme (Grant no. 9520036-28807) and the Human Capital and Mobility Programme of the European Community (Contract ERBCHRXCT9-0529).

References

[1] Nyengaard, N.R., Mortensen, K.K., Lassen, S.F., Hershey, J.W.B. and Sperling-Petersen, H.U. (1991) Biochem. Biophys. Res. Commun. 181, 1572–1579.

[2] Sperling-Petersen, H.U. and Mortensen, K.K. (1990) Protein Eng. 3, 343–344.

[3] Severini, M., Choli, T., Teana, A.L. and Gualerzi, C.O. (1992) FEBS Lett. 297, 226–228.

[4] Sacerdot, C., Dessen, P., Hershey, J.W.B., Plumbridge, J.A. and Grunberg-Manago, M. (1984) Proc. Natl. Acad. Sci. USA 81, 7787–7791.

[5] Blattner, F.R., Plunkett III, G., Mayhew, G.F., Perna, N.T. and Glasner, F.D. (1997) Acc. no. ECAE000397.

[6] Shazand, K., Tucker, J., Chiang, R., Stansmore, K., Sperling-Petersen, H.U., Grunberg-Manago, M., Rabinowich, J.C. and Leighton, T. (1990) J. Bacteriol. 172, 2675–2687.

[7] Steffensen, S., Poulsen, A.B., Mortensen, K.K., Korsager, B. and Sperling-Petersen, H.U. (1994) Biochem. Mol. Biol. Int. 34, 1245–1251.

[8] Hansen, N.J.V., Kristensen, P., Lykke, J., Mortensen, K.K. and Clark, B.F.C. (1995) Biochem. Mol. Biol. Int. 35, 461–465.

[9] Laalami, S., Timofev, A.V., Putzer, H., Leautey, J. and Grunberg-Manago, M. (1994) Mol. Microbiol. 11, 293–302.

[10] Sternberg, N. (1990) Proc. Natl. Acad. Sci. USA 87, 103–107.

[11] McKane, M. and Milkman, R. (1995) Genetics 139, 35–43.

[12] Knudsen, C.R., Kjærsgård, I.V.H., Wiborg, O. and Clark, B.F.C. (1995) Eur. J. Biochem. 228, 176–183.

[13] Nelson, K., Whittam, T.S. and Selander, R.K. (1991) Proc. Natl. Acad. Sci. USA 88, 6667–6671.

[14] Nelson, K. and Selander, R.K. (1992) J. Bacteriol. 174, 6886–6895.

[15] Boyd, E.F., Nelson, K., Wang, F.-S., Whittam, T.S. and Selander, R.K. (1994) Proc. Natl. Acad. Sci. USA 91, 1280–1284.

[16] Bisercic, M., Feutrier, J.Y. and Reeves, P.R. (1991) J. Bacteriol. 173, 3894–3900.

[17] Laalami, S., Putzer, H., Plumbridge, J.A. and Grunberg-Manago, M. (1991) J. Mol. Biol. 220, 335–349.

[18] Laalami, S., Sacerdot, C., Vachon, G., Mortensen, K., Sperling-Petersen, H.U., Cenatiempo, Y. and Grunberg-Manago, M. (1991) Biochimie 73, 1557–1566.