Biochimie 81 (1999), 235–244

Disordered C-terminal domain of tyrosyl-tRNA synthetase: Secondary structure prediction

Lutz Jermutus*, Valérie Guez, Hugues Bedouelle**

Groupe d'Ingénierie des Protéines (CNRS URA 1129), Unité de Biochimie Cellulaire, Institut Pasteur, 28, rue du Docteur-Roux, 75724 Paris cedex 15, France

(Received 2 May 1998; accepted 7 January 1999)

Abstract — The C-terminal domain (residues 320–419) of tyrosyl-tRNA synthetase (TyrRS) from *Bacillus stearothermophilus* is disordered in the crystal structure and involved in the binding of the anticodon arm of $tRNA^{Tyr}$. The sequences of 11 TyrRSs of prokaryotic or mitochondrial origins were aligned and the alignment showed the existence of conserved residues in the sequences of the C-terminal domains. A consensus could be deduced from the application of five programs of secondary structure prediction to the 11 sequences of the query set. These results suggested that the sequences of the C-terminal domains determined a precise and conserved secondary structure. They predicted that the C-terminal domain would have a mixed fold (α/β or $\alpha+\beta$), with the α-helices in the first half of the sequence and the β-strands mainly in its second half. Several programs of fold recognition from sequence alone, by threading onto known structures, were applied but none of them identified a type of fold that would be common to the different sequences of the query set. Therefore, the fold of the C-terminal, anticodon binding domain might be novel. © Société française de biochimie et biologie moléculaire / Elsevier, Paris

aminoacyl-tRNA synthetase / tyrosyl-tRNA synthetase / structural disorder / structural prediction / threading

1. Introduction

The aminoacyl-tRNA synthetases are the enzymes that translate the genetic code in vivo since they attach the amino acids to the corresponding transfer-RNAs. They are constructed in a modular way and divided into two structural classes. They contain a catalytic domain, whose fold is common to the elements of a class, and a domain of binding to the anticodon, whose fold varies between the elements of a class [1]. The crystal structure of tyrosyl-tRNA synthetase from *Bacillus stearothermophilus* (Bst-TyrRS) has been determined at high resolution [2, 3]. Each subunit of Bst-TyrRS comprises two domains in the crystal structure, an N-terminal domain (residues 1 to 319) which has a well resolved structure, and a C-terminal domain (residues 320–419) which appears disordered and for which it has not been possible to trace the polypeptide chain in the electron density map. The N-terminal domain contains the interface of dimerization and the sites of binding for tyrosine, tyrosyl-adenylate and the acceptor arm of $tRNA^{Tyr}$ [3–5]. The C-terminal domain contains the binding site for the anticodon arm of $tRNA^{Tyr}$ [5]. Thus, the C-terminal domain of Bst-TyrRS constitutes an interesting case of association between a structural disorder and a function. Recently, we have shown that the isolated form of the C-terminal domain has secondary structure, is compact and unfolds through a cooperative transition [6]. We then undertook two different approaches to gain additional structural information. We wanted to know whether the C-terminal domain has a fluctuating conformation, is in an incomplete state of folding like the molten globular state, or has a well defined three-dimensional structure. We wanted to understand the cause of its disorder and the relations between this disorder and the function of $tRNA^{Tyr}$ recognition. One approach is experimental and based on spectral methods (fluorescence, circular dichroism, nuclear magnetic resonance). The other approach is based on structural predictions. Here, we report the results of this second approach.

The prediction of the structure of proteins from their sequence finds its rationale in Anfinsen et al.'s experiments, showing that the information for folding and structure is included within the amino acid sequence [7]. Most methods of structural prediction are based on the assumption of a hierarchical mechanism of protein folding, in which elements of secondary structure are first formed by local interactions, then assembled to form the tertiary structure. Recent research on protein folding has shown that this assumption is incorrect and that folding

* Present address: Biochemisches Institut, Universitaet Zurich, Winterthurerstr. 190, 8057 Zurich, Switzerland
** Correspondence and reprints
Abbreviations: aaRS, aminoacyl-tRNA synthetase where aa is the amino acid in the three- or one-letter code; CD, circular dichroism; 2D, secondary structure; 3D, tertiary structure; PDB, Protein Data Bank; SE, standard error; TyrRS or YRS, tyrosyl-tRNA synthetase.

occurs through a mechanism of simultaneous nucleation and condensation [8]. So, most algorithms of secondary structure prediction (2D prediction) scan the protein sequence linearly, under a window of a fixed size, even though distal interactions can influence secondary structure. As the formation of an α-helix is based mainly on residues close in primary sequence and as the formation of a β-sheet requires the interaction of residues distant in primary sequence, the prediction of α-helices may be favored for certain sequences by the linear algorithms of 2D prediction [9, 10]. Several solutions exist to circumvent this problem. A first approach consists in using a variety of different prediction algorithms [11]. A second approach consists in using a set of homologous sequences, as diverse as possible, as an input for prediction [12, 13]. The rationale of this second approach is to use the evolutionary information. Statistics and thresholds can be used with these two approaches to decide whether there exists a consensus of prediction. The prediction is considered meaningful only if such a consensus can be derived. A third approach consists in predicting the tertiary structure directly from sequence, by threading it onto known three-dimensional structures [14, 15]. We used these different approaches to predict the structure of the C-terminal domain of Bst-TyrRS. We found that we could derive a consensus of 2D prediction from a statistical evaluation of the individual predictions. We validated this consensus with other structural criteria like the existence of N- or C-caps, the amino acid composition of the predicted structural elements [16] or their hydrophobic moment [17]. We did not obtain a clear prediction of fold for the C-terminal domain by the methods of threading. Therefore, it may have a novel fold which is not present in the libraries of known structures.

2. Materials and methods

2.1. Sequences

We used the following TyrRS sequences of prokaryotic or mitochondrial origins: from *Thiobacillus ferrooxidans* (TfoYRS) [18], *Bacillus caldotenax* [19], *Bacillus stearothermophilus* (BstYRS) [20], the products of the *tyrS* gene (BsuSYRS) [21] and of the *tyrZ* gene (BsuZYRS) [22] from *Bacillus subtilis*; from *Escherichia coli* (EcoYRS) [23], *Haemophilus influenzae* and *Mycoplasma genitalium* (HinfYRS and MgenYRS; both from the TIGR database at http://www.tigr.org/tdb), *Mycobacterium leprae* (MlepYRS; from the MIPS database at http://www.mips.biochem.mpg.de), *Mycobacterium smegmatis* (MsmeYRS; partial C-terminal sequence) [24], *Mycobacterium tuberculosis* (MtubYRS) [25], *Saccharomyces cerivisiae* mitochondria (ScmiYRS; J.E. Hill and A.A. Tzagoloff; SWISSPROT data base, accession number P48527), *Podospora anserina* mitochondria [26] and *Neurospora crassa* mitochondria [27]. The published sequence of the last protein contains several errors; the correct sequence is stored in GenBank under accession number M17118 X52285 (NEUTYRSM).

2.2. Alignment

The 11 TyrRS sequences were aligned with the PILEUP program [28] which is included in the GCG-package (Genetics Computer Group, Inc., Madison, USA) [29]. The alignment was slightly improved by hand. The C-terminal domain of each TyrRS was defined by comparison with Bst-TyrRS (see *Results*). The input for the programs of prediction that use a sliding window of sequence, included 20 residues upstream of the C-terminal domain. The input included only the C-terminal domain, without extension, in the other cases (the threading programs included). The PHYLIP program, version 3.5c (Phylogeny Inference Package; J. Felsenstein, University of Washington; http://evolution.genetics.washington.edu/phylip.html), was used to build an evolutionary tree of the sequences.

2.3. Structural class predictions

Several methods exist to predict the structural class of a protein from its sequence, i.e., α, β, α + β (mixed fold with dominantly antiparallel β-strands), α/β (mixed fold with dominantly parallel β-strands). We used the algorithm of Genfa et al. [30], which is so far the most accurate. In this algorithm, each protein is represented by a vector in a 20-dimensional space, whose coordinates are the frequencies of occurrence of the 20 amino acids. The algorithm consists in calculating the distances between the protein of unknown structure and each of four average proteins, representative of the four structural classes. The smallest distance corresponds to the most related structural class. The mean distance and associated standard error (SE) for the 11 TyrRS sequences of the query set (strictly restricted to the C-terminal domain) were calculated for each structural class. We also used the SSCP program (http://www.embl-heidelberg.de) [31, 32]. The mean contents and SE in each structural state (α-helix, β-strand, random coil) for the 11 sequences of the query set were calculated as above.

2.4. Secondary structure (2D) predictions

We used several methods of 2D prediction. Whether an individual sequence or the multiple sequence alignment was taken as input is indicated for each method. The predictions were restricted to three types of secondary structure, as defined in the DSSP program [33]: H for the helical structures α, 3_{10} and π; E for extended β-strands and C for turns, bends, coils, β-loops and non-periodic structures.

2.5. 2D-predictions based on neural networks

The PHD method (http://www.embl-heidelberg.de) [34, 35] uses three levels of neural networks and the evolutionary information contained in a multiple sequence alignment. The third level of neural networks, called jury decision, combines the predictions of all the previous levels. The iterative method of Chandonia and Karplus (C&K) uses a pair of algorithms which are based on neural networks and predict the class of tertiary structure and the secondary structure of proteins. Each algorithm improves the accuracy of the prediction by using the information provided by the other [36]. This method of prediction was kindly run for us by Dr. J.M. Chandonia. The input for both methods was the alignment of the 11 TyrRS sequences of the query set. Both methods give the probabilities for helical, extended or coil structures at each position. The prediction was marked with small letters, h, e or c, if the reliability was lower than 5 for the PHD method, or if the probability was lower than 0.5 for the C&K method.

2.6. 2D-predictions based on homologous sequences

The SSPRED algorithm (http://www.embl-heidelberg.de) [37] is based on the analysis of the amino acid substitutions at each position in a multiple alignment of sequences. These substitutions are compared with matrices of residue exchanges which have been calculated for each of the three states of secondary structure (helix, strand, coil) from a training set, derived from protein families. Only the filtered prediction was used for comparison with other methods. The strength of the prediction at each position can be deduced from the changes between the unfiltered (Pred SS) and filtered (Clean SS) predictions. Filtering consists in adjusting the predictions inside or on the borders of the elements of secondary structure to their environment. Altered predictions were marked with small letters.

2.7. 2D predictions based on amino acid propensities

We applied the GOR II program [38] and the Chou-Fasman (C&F) program [39], which are supplied with the Homology software (BIOSYM Inc., San Diego, USA). The programs were run on a window of 17 residues, as recommended [12]. The numerical probability for each position of each sequence and for each type of secondary structure (helix, strand, loop, coil for GOR II, and helix, strand, turn for C&F) were extracted. The mean probability and SE were then calculated for each position of the alignment and each type of secondary structure. In studies published by others, the secondary structure with the highest probability is generally assigned at each position, but this method of decision overpredicts the helical state [12]. We developed a new method of decision to determine whether more than one type of secondary structure is probable at a given position, and assign predictions. For simplicity, the predictions of GOR II for coil and loop were treated as two independent predictions for coil and the prediction of C&F for turn as a prediction for coil. A single prediction was assigned if no other mean fell within the ± 2SE interval of the highest mean. Two predictions were assigned if another mean was inside the ± 2SE interval of the highest mean (prediction of the highest mean figuring on top of the other). No prediction was assigned at a position (empty space) if all mean values were within the ± 2SE interval of the highest mean.

2.8. Consensus of 2D prediction

We derived a consensus from the five methods of 2D-prediction above, by using the following criteria. All the prediction methods were treated equally. The assignment for an extended β-strand followed less stringent criteria than the one for a helix, as the later tends to be overpredicted (see *Introduction*). We assigned a secondary structure to a segment of sequence only if at least three adjacent residues had identical predictions. The consensus predictions were summarized as follows. A strong prediction of a neural network method or of the SSPRED method (a capital letter), or a single prediction of the GOR II or C&F method, gave one count; a weak prediction (a small letter), or a double prediction, respectively, gave half a count. A prediction for a helix needed ≥ 3 counts while a prediction for an extended β-strand was possible with ≥ 2 counts.

2.9. Controls of 2D predictions

The methods in this section were used to probe and refine the consensus of prediction. We used the algorithm of Eisenberg et al. [17] to calculate the profile of amphipathy for each sequence of the query set and detect periodicities in its hydrophobicity. The hydrophobic moment μ was calculated with an angle d of 100°, modulating an α-helix, and a window of 11 residues as previously described [12]. A mean value $<\mu>$ and SE were calculated for each position of the sequence alignment. We derived the following decision criteria from the initial description of the method [17]: a segment of sequence was predicted in α-helix if at least four residues within the segment had $<\mu>$ - 2SE ≥ 2. The ETH method [40, 41] predicts whether a residue belongs to the surface, interior or active site of a protein or is a parsing residue by analyzing how and to which extent the amino acids at a given position in a multiple alignment vary or are conserved. The periodicity of this tertiary structure information is then used to recognize the secondary structure elements. The sequences were divided into four subfamilies for this method of prediction, on the basis of the evolutionary tree of the TyrRSs (see *Results*). The segments that were

predicted in helix, were screened for stability determinants [16]. The sequences were screened for possible segments of extended conformation with the SSP program in single sequence mode (http://kiwi.imgen.bcm.tmc.edu:8088/search-launcher/Launcher.html) [42, 43]. The corresponding algorithm takes into account the amino acid composition of the internal parts of potentially helical or extended segments, but also their terminal and adjacent regions to predict the elements of secondary structure.

2.10. Threading

The TOPITS (http://www.embl-heidelberg.de) [44], 123D (http://cartan.gmd.de/nick/ run123D.html) [45] and 3D-profile (http://www.doe-mbi.ucla.edu/people/frsvr/frsvr.html) [46] programs were accessible via WWW. THREADER 2 (http://globin.bio.warwick.ac.uk/threader/) [47] was run at the Institut Pasteur (Paris, France). All programs were run in single sequence mode. TOPITS and THREADER 2 were also run with the consensus of 2D-prediction as additional input. ProFIT [48] was run either in single sequence mode or with the alignment of the 10 sequences of the query set (MsmeYRS excluded), independently of any 2D prediction.

3. Results

3.1. Sequences of the query set and their alignment

We took 10 tyrosyl-tRNA synthetases (TyrRS) of prokaryotic origin and the mitochondrial TyrRS from *Saccharomyces cerevisiae* to constitute the query set. The mitochondrial TyrRSs from *Neurospora crassa* and *Podospora anserina* not only charge $tRNA^{Tyr}$ but are also involved in the splicing of group I introns [26, 27]. Their C-terminal domains include additional segments which could be involved in this splicing activity and which make the alignment of their sequences with those of other TyrRSs only possible if long insertions and deletions are accepted (not shown). TyrRS from *Bacillus caldotenax*, which has 98% identity with TyrRS from *B.stearothermophilus*, was excluded to maximize the content of the query set in evolutionary information. The 11 TyrRS sequences of the query set were aligned with the PILEUP program *(figure 1)*. The C-terminal domain of each TyrRS was defined as the C-terminal sequence of the protein that began with the residue homologous to Glu320 of Bst-TyrRS. The phylogenetic tree of the C-terminal domains was calculated from the alignment of *figure 1* (restricted to residues 320 and following) with the PHYLIP program. The TyrRSs were divided into four subfamilies on the basis of this tree. Subfamily 1 included ScmiYRS; subfamily 2, BsuSYRS, BstYRS and EcoYRS; subfamily 3, MlepYRS, MtubYRS and MsmeYRS; and subfamily 4, BsuZYRS, HinfYRS, TfoYRS and MgenYRS.

3.2. Structural class predictions

We applied the method of Genfa et al. [30] to predict the structural class of the C-terminal domain of TyrRS. The mean distances and associated standard errors between the 11 TyrRSs of the query set (strictly restricted to the C-terminal domain) and the four structural classes were the following: 10.2 ± 0.8 for the α-class, 11.5 ± 0.5 for the β-class, 9.4 ± 0.6 for the $\alpha + \beta$ class, and 9.3 ± 0.6 for the α/β class. Thus, this method classified the C-terminal domain into the mixed classes but we could not distinguish between the $\alpha + \beta$ and α/β classes. The two versions of the SSCP program [31, 32] also classified the C-terminal domain in the mixed classes. The method of Genfa et al. predicted a closer relationship of the C-terminal domain to the all helical class than to the all extended class and the SSCP method predicted a higher content in α-helices than in β-sheets for each sequence of the query set.

3.3. Secondary structure predictions

We predicted the secondary structure of the C-terminal domain of TyrRS by a combination of methods based on neural networks, matrices of residue exchanges and propensities of amino acids. The predictions were limited to three states: helical, extended β-strand and coil (*Materials and methods*). The outcome of each method and the consensus of prediction are given in *figure 1*. This consensus resulted from the application of five different methods to 11 homologous sequences, and allowed us to divide the C-terminal domain into four helical, four extended and seven coil segments of secondary structure. We then critically analyzed each of these segments with independent structural criteria. Favorable and unfavorable characteristics of each segment are listed below. The numbering of the positions corresponds to the Bst-TyrRS sequence.

3.3.1. Helix α1

In the crystal structure of Bst-TyrRS, helix H5' (residues 309–318) is partially disordered [3]. Our prediction localized the N-terminus of this helix correctly and predicted its C-terminus at position 321 rather than 318. Helix H5' may not be fully visible in the electron density map and our prediction may have identified its real C-terminus. This segment of sequence contained a well-defined local maximum of the hydrophobic moment *(figure 2)* and an abundance in alanine, which is the most stabilizing residue at internal helical positions. These two observations were consistent with the prediction in helix. However, no clear C- or N-cap could be identified [16].

3.3.2. Coil c1

Several insertions and deletions of residues in this segment, and the presence of conserved Pro, Gly and Ser

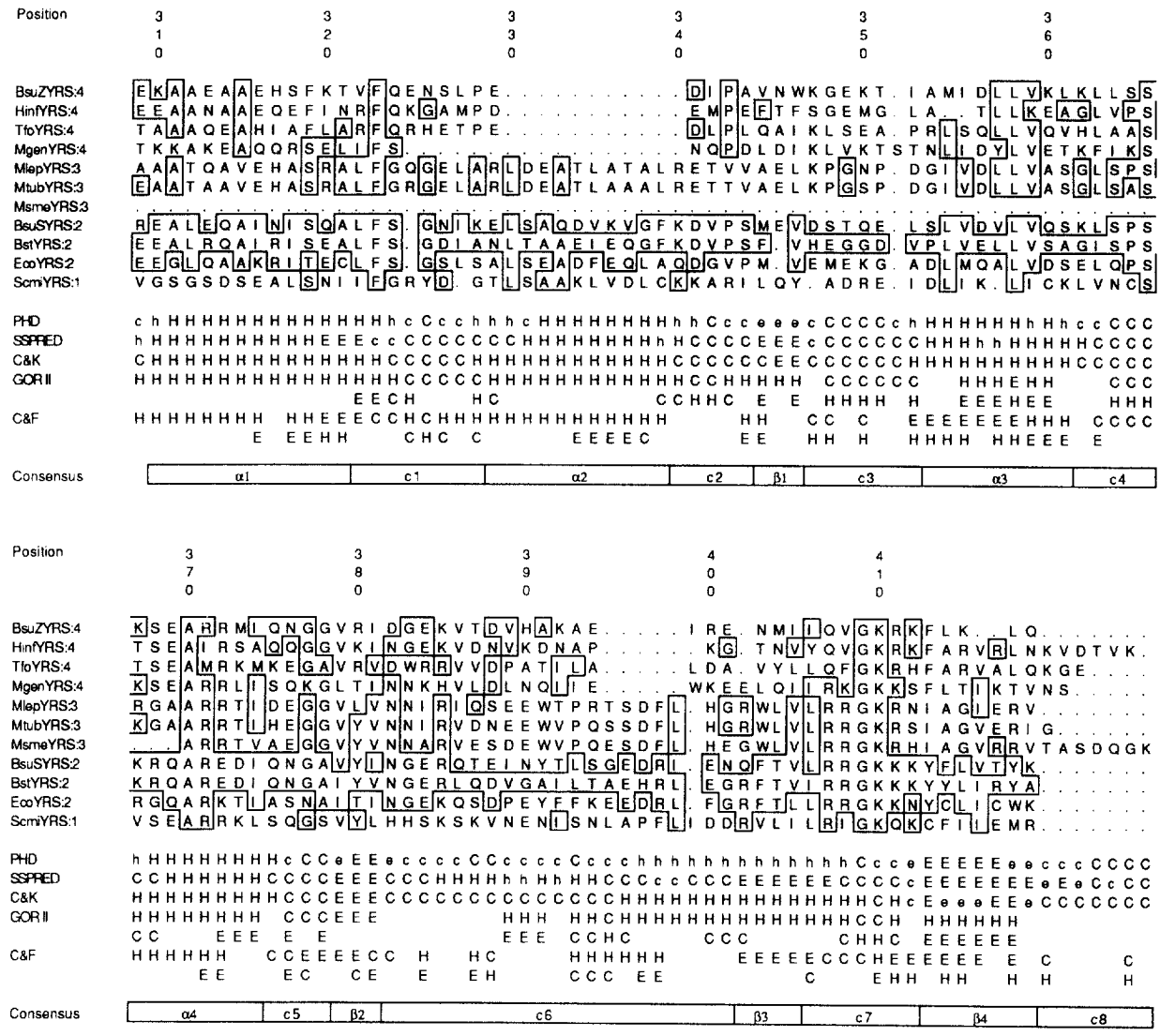

Figure 1. Multiple sequence alignment and 2D-predictions for the C-terminal domains of TyrRSs. The residues are numbered according to the sequence of Bst-TyrRS. The abbreviations for the organisms of origin are defined in *Materials and methods*. The number after the name of each TyrRS refers to one of four subfamilies, defined in *Results*. The residues that are identical to the residue of Bst-TyrRS are boxed at each position of the alignment. The methods of 2D-prediction (PHD, SSPRED, C&K, GORII and C&F), their outputs (H, h or α for helix; E, e or β for extended β-strand; C or c for coil) and the definition of the consensus are described in *Materials and methods*.

residues were consistent with the prediction in coil. The ETH method indeed defines Pro, Gly and to a lesser extent Asn, Asp and Ser as parsing residues [41]. This segment could be flexible and responsible for the observed disorder of the crystal structure. The fully conserved Phe at position 323 could contact the bound tRNA, given the tendency of aromatic residues to be involved in molecular recognition [49], and this contact could rigidify the segment in the complex.

3.3.3. Helix α2

An N-cap was present at position 331, with Asp, Ser or Thr residues, and a C-cap was present at position 340, with Arg or Lys residues in all the organisms except

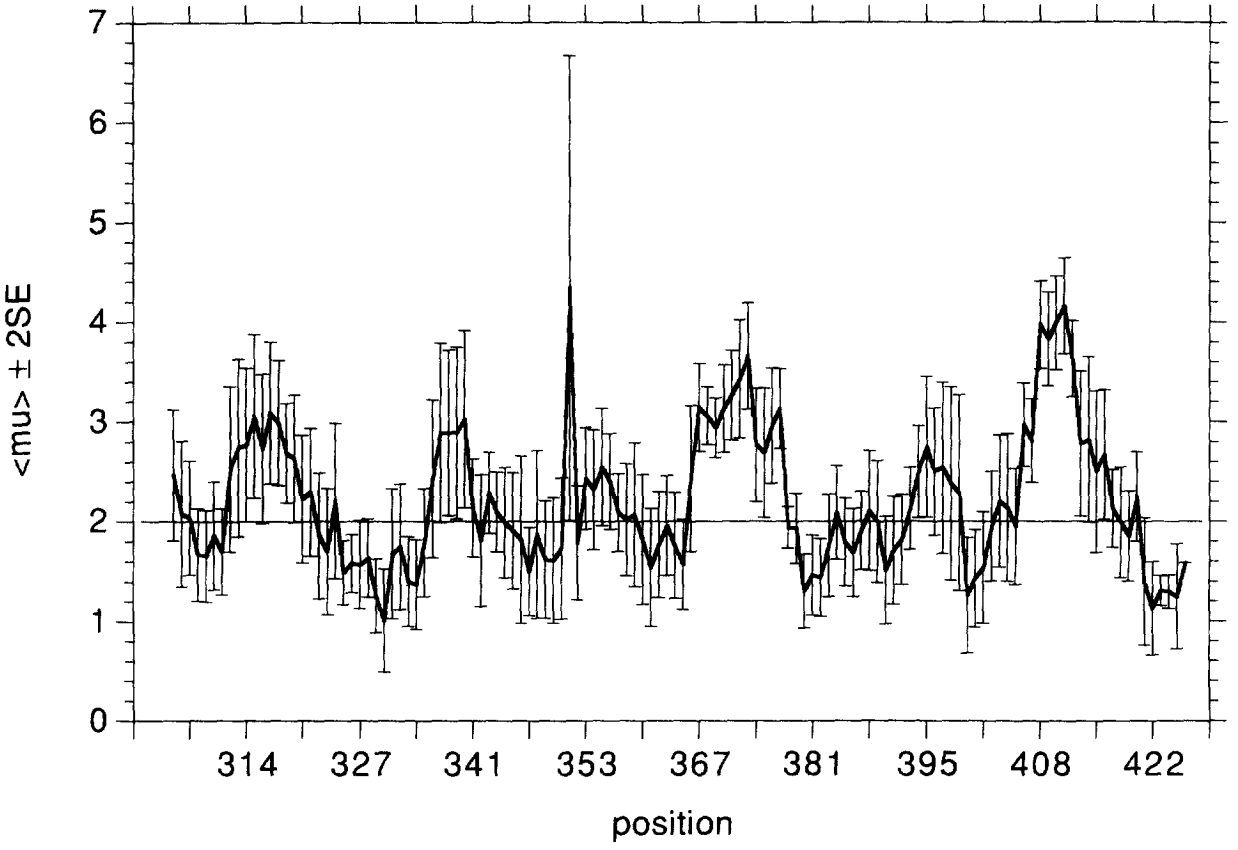

Figure 2. Mean amphipathy profile of the C-terminal domains of the TyrRSs. The hydrophobic moment μ was calculated at each position of each TyrRS sequence as described [17]. <mu>, mean value of μ at each position of the multiple sequence alignment of *figure 1*; SE, associated standard error. The positions are numbered according to the sequence of Bst-TyrRS (see *figure 1*).

E. coli. Residues 337–340 corresponded to a local maximum of the hydrophobic moment *(figure 2)*. This indicated that an entire helix was inserted where coil c1 seemed to form a structural shortcut in the TyrRSs of subfamily 4.

3.3.4. Coil c2

The presence of Pro, Asp and Ser residues in this segment was consistent with its prediction in coil.

3.3.5. Strand β1

This segment contained a surface-indicating residue and two flanking interior-indicating residues for subfamilies 1, 3 and 4, according to the ETH method. This pattern was compatible with the prediction in β-strand.

3.3.6. Coil c3

The occurrence of insertions and deletions, and the high content in Gly and Pro residues were consistent with the prediction.

3.3.7. Helix α3

This segment contained a repetition, starting at position 353, of one surface-indicating residue followed by two or three interior-indicating residues for all subfamilies, according to the ETH method. This pattern suggested a mostly buried amphiphilic helix. However, this segment contained only a low local maximum of the hydrophobic moment *(figure 2)* and no conserved N- or C-cap could be detected. Part of the segment could be in β-strand as it contained the most prominent hydrophobic cluster of the C-terminal domain and as the SSP program predicted this type of secondary structure for seven sequences of the query set (data not shown) [43].

3.3.8. Coil c4

The presence of Pro and Ser residues was consistent with the prediction.

3.3.9. Helix α4

This segment contained a high local maximum of the hydrophobic moment *(figure 2)*. A clear N-cap was visible at position 369. We did not detect any C-cap. The conserved hydrophobic residues at positions 370 and 374 were compatible with a regular α-helix.

3.3.10. Coil c5

This segment was characterized by a high frequency of Gly residues at positions 377 and 378. The length of this segment suggested a loop between two elements of secondary structure.

3.3.11. Strand β2

This segment contained one surface-indicating residue at position 380 and two flanking interior-indicating residues, according to the ETH method. This pattern was compatible with an amphiphilic β-strand. Furthermore, the SSP method predicted an extended structure for all the sequences (not shown).

3.3.12. Coil c6

In the N-terminal part of this segment (positions 382–394), the prediction in coil was consistent with the high frequency of hydrophilic and surface-indicating residues. In its C-terminal part (positions 395–401), it was consistent with the presence of deletions and insertions in all the sequences. The hydrophobic moment had a local maximum in this segment but did not fulfil the conditions for a helical assignment *(figure 2)* (*Materials and methods*). Interestingly, the same sequences had deletions in segment α2 and in segment c6. This observation suggested that these two regions could be close in the tertiary structure of the C-terminal domain.

3.3.13. Strand β3

This segment mainly contained interior-indicating residues, except for the N-terminal position. However, the segment was too small for an internal helix, which would require at least 8 residues [41].

3.3.14. Coil c7

The conserved Gly residue was consistent with a loop between two elements of secondary structure. Five other positions showed fairly conserved surface-indicating residues, so that the ETH method also predicted a loop. The segment contained a local maximum of the hydrophobic moment and fulfilled the conditions for a helical assignment *(figure 2)*. However, the maximum of the hydrophobic moment was due to the high conservation of Arg and Lys residues, which have the highest hydrophobicity values, and to their occurrence with a periodicity of three.

3.3.15. Strand β4

The ETH method gave the patterns iiiisis, iisissi and iisisis, (i, interior-indicating residue; s, surface-indicating residue) respectively for subfamilies 2, 3 and 4, in this segment. All three patterns were compatible with a β-strand prediction.

3.4. Fold prediction

We tried several programs of sequence threading to propose a fold for the C-terminal domain. The query sequences were strictly restricted to the C-terminal domain with these programs. Two of them, TOPITS [44] and THREADER 2 [47], allow one to give custom-designed predictions of secondary structure as input, along with the sequence of interest. Therefore, we performed threading with each of the 11 sequences of the query set and the consensus of 2D prediction as inputs *(figure 1)*. The reliability index of the 2D-predictions (between 0 and 1, with 1 indicating the highest degree of confidence) was set to 0.2 for segments β1, α3, β3 and c7 and to 0.8 for all the other ones. With TOPITS, the 5 template structures that were top-ranking for each query sequence, were common to the 11 sequences of the query set. However, the corresponding Z-scores were low (around 2), the alignments between the template and query sequences comprised long insertions and deletions, the template structures did not correspond to single protein domains, and they had different folds. These five template proteins were of mixed folds: aspartyl-tRNA-synthetase (PDB identifier, 1asz), α1 antichymotrypsin (2ach), horse leucocyte elastase (1hle), modified α1 antichymotrypsin (7api) and mannose-binding protein A (1rtm). Aspartyl-tRNA-synthetase [50] ranked first for several sequences. The aligned region (residues 286 to 423) was in the α/β catalytic domain but was not a distinct domain by itself. With THREADER 2, the results were very inconsistent, i.e., the different sequences threaded to very different folds. The alignments to the template sequences were very poor even when the Z-scores were high (> 3.5).

We also performed threading independently of 2D predictions with the 3D-profile [46], THREADER 2 and ProFIT [48] programs. With 3D-profile and THREADER 2, each query sequence was threaded to a different template structure and often to a different fold. With ProFIT, five of the sequences and the multiple sequence alignment threaded on OB-folds (oligonucleotide/oligosaccharide binding fold) [51]. However, neither the OB-fold nor any other fold consistently gave high Z-scores and sequence alignments of good quality when the query sequence varied. We therefore concluded that ProFIT did not give a clear and reliable prediction of fold for the C-terminal domain.

4. Discussion

4.1. Conservation of sequence and structure features in the C-terminal domain

We could align the sequences of the C-terminal domains for the 11 TyrRSs of the query set and the alignment

showed the existence of highly conserved residues *(figure 1)*. These results strongly suggested that the C-terminal domains had conserved functional or structural features. They were consistent with and extended previous data showing that six positively charged residues of the C-terminal domain from Bst-TyrRS, which are involved in the recognition of $tRNA^{Tyr}$, are conserved in the prokaryotic and mitochondrial TyrRSs (see below) [5, 25, 52]. We applied two different algorithms to each sequence of the query set, to predict the structural class of the C-terminal domain. Both algorithms predicted a mixed fold, containing α-helices and β-sheets, with a predominance of α-helices. We applied five different methods of secondary structure (2D) prediction to the sequences of the query set or to their alignment. The different methods gave very close predictions for the C-terminal domain, from which we could deduce a consensus of 2D-prediction and identify several segments of secondary structure *(figure 1)*. This consensus was compatible with general rules of protein structure and with the predictions of structural class. Therefore the multiple alignment, the structural class predictions and the 2D-predictions together strongly suggested that the sequences of the C-terminal domains determined a precise and conserved secondary structure. They also suggested that a structural element (the putative helix α2) in the hinge region between the N- and C-terminal domains, was deleted in the TyrRSs of subfamily 4. Among the six positively charged residues of the C-terminal domain from Bst-TyrRS that are involved in the recognition of $tRNA^{Tyr}$ [5], Arg368 and Arg371 were located in helix-α4 and separated by one turn of helix, whereas Arg407, Arg408, Lys410 and Lys411 were located in a loop, between strands β3 and β4.

4.2. The C-terminal domain may have a novel fold

None of the methods of threading gave a tertiary structure (3D) prediction of the C-terminal domain that was clear and common to all the TyrRSs of the query set. Several reasons could explain this failure. The C-terminal domain could be in an unfinished state of folding and the use of threading methods in this case would be inappropriate since the scoring functions for these methods have generally been derived from sets of proteins with a well defined three-dimensional structure. If the C-terminal domain has a well defined structure, its fold could be novel and absent from the data banks. The structure of the C-terminal domain of the TyrRSs and the closest structure in the data banks could include non-homologous regions that were important in size or number. Or the programs could be unable to extract the corresponding fold from the data banks. Several threading programs (TOPITS and THREADER 2 when run with sequence as the only input, 3D-profile, 123D) have a built-in method of 2D-prediction. These methods of 2D-prediction could be biased towards α-helices (see *Introduction*) and the programs that incorporate them, could be unable to thread a sequence on structures that are in extended conformation.

4.3. Comparison with data of circular dichroism

The content of the isolated C-terminal domain of Bst-TyrRS in secondary structure elements has been predicted from its far-UV CD spectra. These predictions gave about 16% of residues in α-helices, 27% in β-strands, and thus a mixed fold with a predominance of β-strands [6]. The prediction of structural class of the C-terminal domain from the 11 sequences of the query set (*Results*) also gave a mixed fold, but with a predominance of α-helices. The percentages of residues in α-helix and in β-strand, calculated from the consensus of 2D-prediction *(figure 1)*, were respectively higher (30%) and lower (16%) than the percentages calculated from the CD spectra. These differences could be explained by the fact that the CD data have been collected for wavelengthes above 190 nm, and that a reliable prediction of the content in α-helix but not of the content in β-strand can be calculated from this region of the spectra [53]. However, it is also possible that our predictions of structural class and of secondary structure from sequence alone are imprecise, or that the isolated C-terminal domain does not adopt a fully native structure, or that the binding of $tRNA^{Tyr}$ is necessary to induce the predicted conformation.

5. Concluding remarks

The apparent disorder of the C-terminal domain in the crystal structure of Bst-TyrRS [3] could have several causes. The C-terminal domain could be in an unfinished folding state in the absence of tRNA, e.g., in a molten globular state, and its conformation could fluctuate permanently in the crystals. It could have a single and defined structure but adopt several orientations with respect to the N-terminal domain in the crystals, due to the flexibility of the polypeptide that links them. Finally, it could adopt several different stable structures, which would coexist in the crystals. The results of our 2D-predictions and of our protein engineering experiments did not favor the third cause. They could not distinguish between the first two causes but nevertheless suggested that the C-terminal domain adopts a well defined structure in some conditions. Segments c1 and α2 of our consensus of 2D-prediction had sequences that were more variable than the remainder of the C-terminal domain *(figure 1)*. This variability was compatible with some flexibility of the linking polypeptide between the N- and C-terminal domains. The apparent disorder of the C-terminal domain of Bst-TyrRS could thus have the same cause as the disorders of the N-terminal α-helical arms of SerRS and PheRS from *Thermus thermophilus,* which stretch out into the solvent. In the crystal structure of the free SerRS, the orientation of

the helical arm (residues 24–100) with respect to the catalytic domain is different in the two subunits of the dimer [54]. Moreover, in the structure of the 1:1 complex between SerRS and $tRNA^{Ser}$, the helical arm that interacts with the tRNA, shows a third orientation and the one that does not interact with the tRNA, is partially absent (residues 34-86) from the electron density [55]. Residues 1–84 of the α-subunit of PheRS are absent from the electron density of the free enzyme, whereas they form a helical arm, similar to the SerRS one, in the structure of the complex with $tRNA^{Phe}$ [56]. Nevertheless, there are differences between the C-terminal domain of Bst-TyrRS, which belong to the class I of aminoacyl-tRNA synthetases, and the helical arms of SerRS and PheRS, which belong to class II. The CD data and our secondary-structure predictions show that the C-terminal domain of TyrRS has not a coiled-coil structure. Moreover, it specifically recognizes the anticodon of $tRNA^{Tyr}$ whereas the helical arms of SerRS and PheRS interact with the D and TC loops of the cognate tRNAs.

Our results of sequence alignment and 2D-predictions will help us to further dissect the relations between the apparent disorder of the C-terminal domain of Bst-TyrRS, its structure and its function by mutagenesis experiments. Once the structure of the C-terminal domain for the TyrRS from one organism will be solved, our results will help to extend this structure to the TyrRSs from other organisms.

Acknowledgments

We are grateful to J.M. Chandonia and M. Karplus (Harvard University) for running their program of prediction on our sequence alignment and to Francisco S. Domingues and Manfred Sippl (University of Salzburg) for running the ProFIT program on the TyrRS sequences.

References

[1] Schimmel P., Giege R., Moras D., Yokoyama S., An operational RNA code for amino acids and possible relationship to genetic code, Proc. Natl. Acad. Sci. USA 90 (1993) 8763–8768.

[2] Brick P., Blow D.M., Crystal structure of a deletion mutant of a tyrosyl-tRNA synthetase complexed with tyrosine, J. Mol. Biol. 194 (1987) 287–297.

[3] Brick P., Bhat T.N., Blow D.M., Structure of tyrosyl-tRNA synthetase refined at 2.3 Å resolution, J. Mol. Biol. 208 (1989) 83–98.

[4] Fersht A.R., Dissection of the structure and activity of the tyrosyl-tRNA synthetase by site-directed mutagenesis, Biochemistry 26 (1987) 8031–8037.

[5] Bedouelle H., Winter G., A model of synthetase/transfer RNA interaction as deduced by protein engineering, Nature 320 (1986) 371–373.

[6] Guez-Ivanier V., Bedouelle H., Disordered C-terminal domain of tyrosyl transfer-RNA synthetase: evidence for a folded state, J. Mol. Biol. 255 (1996) 110–120.

[7] Anfinsen C.B., Haber E., Sela M., White F.H., The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain, Proc. Natl. Acad. Sci. USA 47 (1961) 1309–1314.

[8] Fersht A.R., Nucleation mechanism in protein folding, Curr. Opin Struct. Biol. 7 (1997) 3–9.

[9] Sawyer L., Holt C., The secondary structure of milk proteins and their biological function, J. Dairy Sci. 76 (1993) 3062–3078.

[10] Wang Z.X., Assessing the accuracy of protein secondary structure, Nature Struct. Biol. 1 (1994) 145–146.

[11] Bazan J.F., Helical fold prediction for the cyclin box, Proteins Struct. Funct. Genet. 24 (1996) 1–17.

[12] Niermann T., Kirschner K., Use of homologous sequences to improve protein secondary structure prediction, Methods Enzymol. 202 (1991) 45–59.

[13] Francesco V.D., Garnier J., Munson P.J., Improving protein secondary structure prediction with aligned homologous sequences, Protein Sci. 5 (1996) 106–113.

[14] Jones D.T., Progress in protein structure prediction, Curr. Opin Struct. Biol. 7 (1997) 377–387.

[15] Finkelstein A.V., Protein structure: what is it possible to predict now, Curr. Opin Struct. Biol. 7 (1997) 60–71.

[16] Fersht A.R., Serrano L., Principles of protein stability derived from engineering experiments, Curr. Opin. Struct. Biol. 3 (1993) 75–83.

[17] Eisenberg D., Weiss R.M., Terwilliger T.C., The hydrophobic moment detects periodicity in protein hydrophobicity, Proc. Natl. Acad. Sci. USA 81 (1984) 140–144.

[18] Salazar O., Sagredo B., Jedlicki E., Soll D., Weygand-Durasevic I., Orellana O., *Thiobacillus ferrooxidans* tyrosyl-tRNA synthetase functions in vivo in *Escherichia coli*, J. Bacteriol. 176 (1994) 4409–4415.

[19] Jones M.D., Lowe D.M., Borgford T., Fersht A.R., Natural variation of tyrosyl-tRNA synthetase and comparison with engineered mutants, Biochemistry 25 (1986) 1887–1891.

[20] Winter G., Koch G.L.E., Hartley B.S., Barker D.G., The amino acid sequence of the tyrosyl-tRNA synthetase from *B. stearothermophilus*, Eur. J. Biochem. 132 (1983) 383–387.

[21] Henkin T.M., Glass B.L., Grundy F.J., Analysis of the *Bacillus subtilis tyrS* gene: conservation of a regulatory sequence in multiple tRNA synthetase genes, J. Bacteriol. 174 (1992) 1299–1306.

[22] Glaser P., Kunst F., Débarbouillé M., Vertès A., Danchin A., Dedonder R., A gene encoding a tyrosine tRNA synthetase is located near *sacS* in *Bacillus subtilis*, DNA Seq. 1 (1991) 251–261.

[23] Barker D.G., Bruton C.J., Winter G., The tyrosyl-tRNA synthetase from *Escherichia coli*, FEBS Lett. 150 (1982) 419–423.

[24] Predich M., Doukhan L., Nair G., Smith I., Characterization of RNA polymerase and two sigma-factor genes from *Mycobacterium smegmatis*, Mol. Microbiol. 15 (1995) 355–366.

[25] Nair S., de Pouplana L.R., Houman F., Avruch A., Shen X., Schimmel P., Species-specific tRNA recognition in relation to tRNA synthetase contact residues, J. Mol. Biol. 269 (1997) 1–9.

[26] Kämper U., Kück U., Cherniack A.D., Lambowitz A.M., The mitochondrial tyrosyl-tRNA synthetase of *Podospora anserina* is a bifunctional enzyme active in protein synthesis and RNA splicing, Mol. Cell. Biol. 12 (1992) 499–511.

[27] Atkins R.A., Lambowitz A.M., A protein required for splicing group I introns in *Neurospora* mitochondria is mitochondrial tyrosyl-tRNA synthetase or a derivative thereof, Cell 50 (1987) 331–345.

[28] Feng D.F., Doolittle R.F., Progressive sequence alignment as a prerequisite to correct phylogenic trees, J. Mol. Evol. 25 (1987) 351–360.

[29] Devereux J., Haeberli P., Smithies O., A comprehensive set of sequence analysis programs for the VAX, Nucleic Acids Res. 12 (1984) 387–395.

[30] Genfa Z., Xinhua X., Chun-Ting Z., A weighting method for predicting protein structural class from amino acid composition, Eur. J. Biochem. 210 (1992) 747–749.

[31] Eisenhaber F., Imperiale F., Argos P., Frommel C., Prediction of secondary structural content of proteins from their amino acid composition alone. I. New analytic vector decomposition methods, Proteins Struct. Funct. Genet. 25 (1996) 157–168.
[32] Eisenhaber F., Frommel C., Argos P., Prediction of secondary structural content of proteins from their amino acid composition alone. II. The paradox with secondary structural class, Proteins Struct. Funct. Genet. 25 (1996) 169–179.
[33] Kabsch W., Sander C., Dictionary of protein secondary structure: Pattern of recognition of hydrogen-bonded and geometrical features, Biopolymers 22 (1983) 2577–2637.
[34] Rost B., Sander C., Combining evolutionary information and neural networks to predict protein secondary structure, Proteins Struct. Funct. Genet. 19 (1994) 55–77.
[35] Rost B., Sander C., Prediction of protein secondary structure at better than 70% accuracy, J. Mol. Biol. 232 (1994) 584–599.
[36] Chandonia J.M., Karplus M., Neural networks for secondary structure and structural class predictions, Protein Sci. 4 (1995) 275–285.
[37] Mehta P., Heringa J., Argos P., A simple and fast approach to prediction of protein secondary structure from multiply aligned sequences with accuracy above 70%, Protein Sci. 4 (1995) 2517–2525.
[38] Gibrat J.F., Garnier J., Robson B., Further developments of protein secondary structure prediction using information theory, J. Mol. Biol. 198 (1987) 425–443.
[39] Chou P.Y., Fasman G.D., Empirical predictions of protein conformation, Annu. Rev. Biochem. 47 (1978) 251–276.
[40] Benner S.A., Predicting the conformation of proteins from sequences. Progress and future progress, J. Mol. Recogn. 8 (1995) 9–28.
[41] Benner S.A., Badcoe I., Cohen M.A., Gerloff D.L., *Bona fide* prediction of aspects of protein conformation, J. Mol. Biol. 235 (1994) 926–958.
[42] Solovyev V.V., Salamov A.A., Predicting α-helix and β-strand segments of globular proteins, CABIOS 10 (1994) 661–669.
[43] Salamov A.A., Solovyev V.V., Protein secondary structure prediction using local alignments, J. Mol. Biol. 268 (1997) 31–36.
[44] Rost B., Schneider R., Sander C., Protein fold recognition by prediction-based threading, J. Mol. Biol. 270 (1997) 471–480.
[45] Alexandrov N.N., Nussinov R., Zimmer R.M., Fast protein fold recognition via sequence to structure alignment and contact capacity potentials, in: Hunter L., Klein T.E. (Eds.), Pacific. Symp. Biocomputing '96, World Scientific Press, Singapore, 1996, pp. 53–72.
[46] Fischer D., Eisenberg D., Fold recognition using sequence–derived predictions, Protein Sci. 5 (1996) 947–955.
[47] Jones D.T., Miller R.T., Thornton J.M., Successful protein fold recognition by optimal sequence threading validated by rigorous blind testing, Proteins Struct. Funct. Genet. 23 (1995) 387–397.
[48] Flöckner H., Braxenthaler M., Lackner P., Jaritz P., Ortner M., Sippl M., Progress in fold recognition, Proteins Struct. Funct. Genet. 23 (1995) 376–386.
[49] Mian I.S., Bradwell A.R., Olson A.J., Structure, function and properties of antibody binding sites, J. Mol. Biol. 217 (1991) 133–151.
[50] Ruff M., Krishnaswamy S., Boeglin M., Poterszman A., Mitschler A., Podjarny A., Rees B., Thierry J.C., Moras D., Class II aminoacyl transfer RNA synthetases: crystal structure of yeast aspartyl-tRNA synthetase complexed with tRNA (Asp), Science 252 (1991) 1682–1689.
[51] Murzin A.G., OB (oligonucleotide/oligosaccharide binding)-fold: common structural and functional solution for non-homologous sequences, EMBO J. 12 (1993) 861–867.
[52] Bedouelle H., Guez-Ivanier V., Nageotte R., Discrimination between transfer-RNAs by tyrosyl-tRNA synthetase, Biochimie 75 (1993) 1099–1108.
[53] Johnson W.C Jr., Protein secondary structure and circular dichroism: a practical guide, Proteins Struct. Funct. Genet. 7 (1990) 205–214.
[54] Fujinaga M., Berthet-Colominas C., Yaremchuk A.D., Tukalo M.A., Cusack S., Refined crystal structure of the seryl-tRNA synthetase from *Thermus thermophilus* at 2.5 Å resolution, J. Mol. Biol. 234 (1993) 222–233.
[55] Biou V., Yaremchuk A., Tukalo M., Cusack S., The 2.9 Å crystal structure of *T. thermophilus* seryl-tRNA synthetase complexed with $tRNA^{Ser}$, Science 263 (1994) 1404–1410.
[56] Goldgur Y., Mosyak L., Reshetnikova L., Ankilova V., Lavrik O., Khodyreva S., Safro M., The crystal structure of phenylalanyl-tRNA synthetase from *Thermus thermophilus* complexed with cognate $tRNA^{Phe}$, Structure 5 (1997) 59–68.