###
#
MSKCC Document Delivery Services
#
Thursday, November 17, 2005
#
##

Request ID: DDS36179
User: Gangi-Dino, Rita
Location: MSK
Requested on: 11/16/2005
Needed by: 11/21/2005
Journal Title: Proteins
ISSN:
Article Author(s): Bork, P, Holm, L, Koonin, E
Article Title: The cytidyltransferase superfamily: identification of nucleotide binding site and fold prediction
Year: 1995
Volume: 22
Issue:
Pages: 259- 266
PMID:

PROTEINS: Structure, Function, and Genetics 22:259–266 (1995)

The Cytidylyltransferase Superfamily: Identification of the Nucleotide-Binding Site and Fold Prediction

Peer Bork,[1,2] Liisa Holm,[1] Eugene V. Koonin,[3] and Chris Sander[1]
[1]*EMBL, 69012 Heidelberg,* [2]*Max-Delbrück Center for Molecular Medicine, 13125 Berlin-Buch, Germany, and*
[3]*National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894*

ABSTRACT **The crystal structure of glycerol-3-phosphate cytidylyltransferase from *B. subtilis* (TagD) is about to be solved. Here, we report a testable structure prediction based on the identification by sequence analysis of a superfamily of functionally diverse but structurally similar nucleotide-binding enzymes. We predict that TagD is a member of this family. The most conserved region in this superfamily resembles the ATP-binding HiGH motif of class I aminoacyl-tRNA synthetases. The predicted secondary structure of cytidylyltransferase and its homologues is compatible with the α/β topography of the class I aminoacyl-tRNA synthetases. The hypothesis of similarity of fold is strengthened by sequence–structure alignment and 3D model building using the known structure of tyrosyl tRNA synthetase as template. The proposed 3D model of TagD is plausible both structurally, with a well packed hydrophobic core, and functionally, as the most conserved residues cluster around the putative nucleotide binding site. If correct, the model would imply a very ancient evolutionary link between class I tRNA synthetases and the novel cytidylyltransferase superfamily.**

Key words: **homology search, phosphodiesterases, sequence analysis, structure prediction, threading**

INTRODUCTION

Cytidylyltransferase was entered into, though not solved in time for, the protein structure prediction contest organized by J. Moult et al. (Asilomar, CA, December 4–8, 1994) as a target sequence with no homology to known structures. Current ab initio structure prediction methods are quite successful at the level of secondary structure but are less reliable for tertiary structure.[28] Our approach was to pursue remote sequence relationships which in many cases can lead to rather precise structural and functional predictions.[3]

It has been suggested that all natural proteins fall into a limited set of evolutionarily and thereby structurally related families.[8,25] The crystal structures of myoglobin and hemoglobin were the first example to indicate the extreme sequence variability which can be supported on a common structural framework. Thirty years later, there are now numerous structurally characterized families in which the sequence signals of common ancestry are washed out, except perhaps for key structural or functional residues around the active site.[19] In the present study, detailed sequence analysis demonstrated the existence of a new, large superfamily of putative α–β phosphodiesterases including the TagD cytidylyltransferase and revealed an apparent link between this superfamily and class I aminoacyl-tRNA synthetases whose structure is known. The hypothesis of homology is supported by sequence conservation of the putative NMP-binding site, agreement between the predicted (cytidylyltransferases) and observed (tRNA synthetases) secondary structures, evaluation of three-dimensional packing, and functional considerations.

SEQUENCE ANALYSIS

The previously reported similarity of the *Bacillus subtilus* TagD sequence[23] to eukaryotic choline cytidylyltransferases (CTPTs) was confirmed by Blastp[1] database searches. In addition, eukaryotic CTPTs are similar to the N-terminus of the yeast protein Muq1[20] and to a segment of the biochemically uncharacterized Aut protein which is required for autotrophic growth of *Alcaligenes eutrophus*.[14]

The family was expanded further by focusing on the most conserved region (Fig. 1) by use of property patterns (Propat[27]), profile (Profilesearch[16]), and motif search tools (MoST).[33] For example, analysis of the Blastp output using the CAP (Consistent Alignment Parser) program[33] identified segments similar to the conserved N-terminal motif (Fig. 1) in the *E. coli* proteins KdtB, PanC, NadR, and YaaC. The resulting conserved alignment block was converted into a position-dependent weight matrix with

Received December 8, 1994; revision accepted February 14, 1995.

Address reprint requests to the Protein Design Group, EMBL, Meyerhofstrasse 1, D-69012 Heidelberg, Germany.

Abbreviations: TagD, glycerol-3-phosphate cytidylyltransferase; NMP, nucleotide phosphate; NMN, nicotinamide nucleotide, CTPT, choline cytidylyltransferase.

which iterative database searches were carried out using the MoST program. With a ratio of the expected to the observed number of selected sequence segments of 0.01 as the cut-off for the MoST search, the process converges, after two iterations, at a unique set of proteins with no obvious false positives. This semiautomatic procedure thus allowed the identification of several other proteins of this emerging superfamily including a second motif in Muq1 (suggesting a domain duplication; Fig. 1, Table 1).

When the sequences of the proteins in this superfamily were compared using the MACAW program,[31] statistically highly significant similarity was detected only for the N-terminal motif (Fig. 1). For example, when closely related sequences were excluded from the analysis to avoid an overestimate of significance,[31] the probability of detecting this motif by chance in seven distantly related sequences (yeast CTPT and Muq1, *E. coli* YaaC, KdtB, PanC, CitC, and NadR) was as low as about 10^{-8}. All these observations underscore that the detected motif can serve as a signature for the new superfamily.

FUNCTIONAL IMPLICATIONS

The basic features of the proteins in the new superfamily are indicated in Table 1. Those with known enzymatic activity (Table 1) cleave the α–β phosphodiester bond in either CTP (TagD, CTPT) or ATP (PanC, CitC). These observations, together with the data on operon organization,[9] suggest that the specificity of *E. coli* KdtB (the gene is located next to 3-deoxy-D-mannooctulosonic acid transferase) and of the putative orthologues from *Synechococcus* and *Mycobacterium* is that of a 3-deoxy-D-mannooctulosonic acid cytidylyltransferase (Table 1). Previously, another *E. coli* protein, KdsB, has been shown to possess this activity,[15] but it belongs to a distinct group of nucleotidyltransferases (E. V. K., unpublished observations). The observations may help in clarifying the precise roles of KdsB and KdtB in bacterial lipopolysaccharide biosynthesis.

NadR is a repressor controlling the transcription of NAD biosynthesis genes, probably utilizing NAD as a corepressor; NadR also may be directly involved in NMN transport.[13,37,38] The possibility that NadR could have evolved from an NTP-utilizing enzyme is particularly intriguing given the sequence and structural similarity between repressors of sugar and nucleotide metabolism operons and sugar-binding proteins (ref. 24 and references therein). Derivation of transcription regulators from metabolic enzymes and binding proteins may have occurred more than once in evolution.

Muq1 is involved in mRNA splicing but its biochemical activity remains unknown.[20] Given the fact that many splicing steps are ATP-dependent, ATPase activity or at least ATP binding seems plausible for Muq1.

For the better characterized proteins of the new superfamily, a common denominator appears to be nucleotide binding. Therefore it is likely that the highly conserved motif, typically located near the N-terminus (Fig. 1), represents a nucleotide-binding site.

SIMILARITY TO THE HiGH MOTIF OF CLASS I tRNA SYNTHETASES

Although the motif corresponding to residues 3–26 of TagD is specific to the new superfamily reported here, there is a surprising resemblance between the putative nucleotide binding site signature and the ATP-binding HiGH motif of class I aminoacyl-tRNA synthetases (Fig. 1; for review see ref. 10). Both protein families have the motif located near the N-terminus of the domain. In the crystal structures of aminoacyl-tRNA synthetases the HiGH motif is in the first loop of a β–α–β unit. The PHD neural network method[28] predicts a β–α–β secondary structure for the conserved motif in the TagD superfamily with the HiGH-like signature in the first loop. This hints at a possible structural analogy with class I aminoacyl-tRNA synthetases which extends to functional similarity, as proteins in both families possess both α–β phosphodiesterase and NMP transferase activities (Table 1).

To test these hypotheses, we asked whether or not TagD and its homologues are likely to adopt a fold similar to that of class I tRNA synthetases also in the neighborhood of the putative HiGH-like motif. This was an arduous undertaking, as proteins within the TagD superfamily already are so diverse that standard alignment programs do not align them reasonably over the full length of the sequences, and as no automated method was available to jointly align the two superfamilies in sequence-structure threading. However, after several rounds of multiple sequence alignment for secondary structure prediction, visual inspection, and threading for 3D model building, we obtained a plausible structural model of the full-length TagD protein. Below, we describe each step of this cyclic process separately, and finish with a discussion of the final model.

SECONDARY STRUCTURE PREDICTION

An initial multiple alignment of the TagD superfamily was constructed using the ClustalW program.[34] In addition to the putative nucleotide-binding site near the N-terminus, the alignment of these distantly related proteins was anchored on a few relatively conserved blocks, for example, a serine- and threonine-rich segment followed by a clearly defined hydrophobicity pattern at the C-terminus (Fig. 1). The mid-part of the family alignment was manually improved on later iterations in an attempt to optimize conserved physical properties of amino acids at a given position. The PHD neural network method[28] predicted an alternating β/α secondary structure

```
                     nucleotide-binding
                  =======================
                   B1              A1               B2
2D predict:     L BBBBb   L   aaAAAAAAAAAAA LLL bBBBBBB LLL

KDTB_ECOLI    2 QKRAIYPGTFDPITNGHIDIVTRATQMFD--HVILAIAASPSK      P23875
KDTB/SYNCO    2 VLNAIYPGSFDPITFGHLDIIERGCRLFD--QVYVAVLRNPNK      L19521
KDTB/MYCTU    1 MTGAVCPGSFDPVTLGHVDIFERAAAQFD--EVVVAILVNPAK      U00024
KDTB/MYCCA    1 MKIAIYPGSFNPFHKGHLNILKKAILLFD--KVYVVVSKNVNK      U15110
CTPT_RAT     76 PVRVYADGIFDLFHSSHARALMQAKNLFP--NTYLIVGVCCDE      P19836
CTPT_YEAST  103 PIRIYADGVFDLFHLGHMKQLEQCKKAFP--NVTLIVGVPSDK      P13259
MUQ1_YEASTN   7 PDKVWIDGCFDFTHHGHAGAILQARRTVSKENGKLFCGVHTDE      P33412
TAGD_BACSU    1 MKKVITYGTFDLLHWGHIKLLERAKQLG----DYLVVAISTDE      P27623
AUT/ALCEU    28 RPLVFTNGVFDILHRGHATYLAQARALG----ASLVVGVNSDA      U07639
MUQ1_YEASTC 197 EDCVYVDGDFDLFHMGDIDQLRKLKMDLHP-DKKLIVGITTSD...   P33412
YAAC_PSEFL   15 RGCVATIGNFDGVHRGHQAILARLRERAVELGVPSCVVIFEPQ...   P22990
YAAC_ECOLI   16 EGCVLTIGNFDGVHRGHRALLQGLQEEGRKRNLPVMVMLFEPQ...   P08391
PANC_ECOLI   21 GKRVALVPTMGNLHDGHMKLVDEAKARAD---VVVVSIFVNPM...   P31663
PANC/YEAST   59 RETIGFVPTMGCLHSGHASLISQSVKENT---YTVVSIFVNPS...   Z38059
NADR_SALTY   63 KNIGVVFGKFYPLHTGHIYLIQRACSQVD--ELHIIMGYDDTR...   P27278
CITC/KLEPN  145 RKIGAIVMNANPFTLGHRWLVEQAASQCD--WLHLFVVKEDAS...   X79817

consensus:         hhh G Ft hH GHh hh thtt h    hhhhhhh ttt
exceptions:                  T  D
tRNA synth:               t hHhGHh
exceptions:                  T  N

                                     A2          B3          A3      B4
2D predict:     LL     LLLL  L AAAAAAAAAAAAAAa  bBBBLLLLL aAAAAAAa aaAAA

KDTB_ECOLI   43 ----------KPMFTLEERVALAQQATAHLGNVEVVG-FSSDLANFARNQHATVLI
KDTB/SYNCO   43 ----------QPMFSVQERLEQIAKAIAHLPNAQVDS-FEEGLVNYARQRQAGAIL
KDTB/MYCTU   42 ----------TGMFDLDERIAMVKESTTHLPNLRVQV--GHGLVDFVRSCGMTAIV
KDTB/MYCCA   42 ----------SLDPDLQSRVENIKNLIKDFSNVEIII-NENKLTTIAKELNACFII
CTPT_RAT    117 ---LTHNFKGFTVMNENERYDAVQHCR-YVDEVVVRAPWTLTPEFLAEHR-IDFVA
CTPT_YEAST  144 ---ITHKLKGLTVLTDKQRCETLTHCR-WVDEVVVPAPWCVTPEFLLEHK-IDHVA
MUQ1_YEASTN  50 ---DIQHNKGTPVMNSSERYEHTRSNR-WCSEVVVEAPYVTDPNWMDKYQ-CQYVV
TAGD_BACSU   40 ---FNLQKQKKAYHSYEHRKLILETIR-YVDEVIIPKNWEQKKQDIIDHN-IDVFV
AUT/ALCEU    67 SVKMLGKGDDRPLNHESDRMALLAALE-SVDLVAAMF-REQTPVELIRLVRPDIYV

consensus:                    t ttRh  htth   ht h h      th  hht   h hhh

                                   A4             B5             A5
2D predict:           LL     aaAAAAAAAAAAaLLL bbBBBBBB L LL   aAAAAAaa L

KDTB_ECOLI   88 RGLRAVADFE----YEMQLAHMNRHLMPELE-SVFLMPSKEWSFISSSLVKEVARH
KDTB/SYNCO   88 RGLRVLSDFE----LELQMANTNKTLASDLE-TVFLTTSTEYSFLSSSLVKEVARF
KDTB/MYCTU   86 KGLRTGTDFE----YELQMAQMNKHIAGVD--TFFVATAPRYSFVSSSLAKEVAML
KDTB/MYCCA   87 RGLRSQADFE----YEIKYYDGFKSLDPNIEVVYFISDYDKRSL-SSTILREIEFY
CTPT_RAT    168 H-----DDIPYSSAGSDDVYKHIKEA------GMFAPTQRTEGISTSDIITRIVRD
CTPT_YEAST  195 H-----DDIPYVSADSDDIYKPIKEM------GKFLTTQRTNGVSTSDIITKIIRD
MUQ1_YEASTN 101 HG----DDIT-IDANGEECYKLVKEM------GRFKVVKRTYGVSTTEIIHRILTK
TAGD_BACSU   91 MG----DDWE-------GKFDFLKDQ------CEVVYLPRTEGISTTKIKEEIAGL
AUT/ALCEU   121 KG----GDYD------IDTLEETRLVRSWG--GQAYAIPFLHDRSTTKLLTRVRQG

consensus:            tDht       thht  +          hhhh    th oothhtthh
```

Fig. 1. Multiple alignment of the cytidylyltransferase superfamily. First column, protein codes (those containing an underscore were taken from the SWISS-PROT database[2]); second column, sequence number at the beginning of the segment within the respective proteins; last column, database accession numbers. Top line, secondary structure predictions using the profile neural network program PHD[28]: A(a) and B(b) are α-helix and β-strand, respectively. Capital letters indicate an expected accuracy of at least 82%, lowercase letters indicate a lower expected accuracy. L denotes confidently assigned loop segments; blanks denote loops predicted at lower expected accuracy. Consensus line: h denotes hydrophobic positions, t indicates polar or turn-like positions, bold capitals highlight amino acids that are conserved in at least 90% of the sequences. The bottom line shows the conserved motif in class I aminoacyl-tRNA synthetases. The predicted nucleotide-binding site is overlined, with the boundaries of the sequence motif determined using the MACAW program. NadR from *E. coli* and several mammalian CTPTs are omitted as they are more than 90% identical with representatives in the alignment. As the bottom 7 sequences are difficult to align confidently except for the N-terminal motif, only a partial alignment is shown. The only discrepancy between the PhD prediction and our model (strand B4) is underlined.

TABLE I. New Superfamily of Nucleotide-Binding Proteins

Proteins*	New	Activity/function	Reaction/ligand/cofactor	References
TAGD_BACSU		Glycerol-3-phosphate cytidyltransferase; techoic acid biosynthesis, cell wall biogenesis	CTP + SN-glycerol 3 phosphate = CDP-glycerol +PP	Mauel et al.[23]
CTPT_RAT, CTPT_YEAST[†]		Choline phosphate cytidyl transferase; phosphatidyl choline biosynthesis, membrane biogenesis	CTP + choline phosphate = CDP-choline +PP	Tsukagoshi et al.[35]; Kalmar et al.[21]
MUQ1_YEAST (Nt), MUQ1_YEAST (Ct)	*	Involved in mRNA splicing	?	Horowitz and Abelson[21]
KDTB_ECOLI, KDTB/SYNCO, KDTB/MYCTU	*	Putative 3-deoxy-mannooctulosonate cytidylyl-transferase (CMP-KDO synthetase); involved in bacterial cell wall lipopolysaccharide biosynthesis	CTP + 3-deoxy-D-manno-octulosonate = CMP-3-deoxy-D-manno-octulosonate +PP	Clementz and Raetz[9]
PANC_ECOLI, PANC/YEAST	*	Pantoate β-alanine ligase; panthothenate biosynthesis	ATP + pantoate + β-alanine = AMP + pantothenate +PP	
CITC/KLEPN	*	Acetate:SH-citrate lyase ligase; citrate assimilation	ATP + acetate +HS-R-ACP = AMP +acetyl-S-CoA + PP	Bott and Dimroth[4]
YAAC_ECOLI, YAAC_PSEFL	*	?	?	Yura et al.[36]
NADR_ECOLI, NADR_SALTY	*	Transcription repressors, also involved in NMN transport	NMN?, NAD?	Zhu et al.[38]
AUT/ALCEU		Involved in autotrophic growth	?	Freter and Bowien[14]
More than 100 known sequences		Class I tRNA synthetases	ATP + aa + tRNA(aa) = AMP + aa − tRNA(aa) + PP	Review: Delarue and Moras[10]

*Codes as in Figure 1; new: similarity has not been described before.

[†]Very similar orthologues have been sequenced from mouse and hamster.

from alignments including only more closely related subgroups, for which a correct overall alignment is unambiguous, and this pattern persisted when the combined alignment of distantly related members of the TagD superfamily was used as input to the neural network program. The final multiple alignment and corresponding secondary structure prediction are shown in Figure 1.

TOPOGRAPHY PREDICTION

The loops between the predicted secondary structure elements are relatively short suggesting a parallel β-sheet topology where the connecting helices run in a direction opposite to the strands. If the HiGH motif is indeed present, then strands 1 and 2 should be adjacent in a β–α–β supersecondary structure unit. These strands have hydrophobic stretches of several residues which suggest that they should be in a sheet in the protein interior. The relatively polar strands 3 and 5 are candidates for sheet edges. The substrate binding site in α/β enzymes is invariably found at the point where loops wind in opposite direction.[7] By this rule, strand 1 should be at a sheet switch point as the nucleotide binding site was identified in the following loop. Taken together, these considerations fit well with the topography of class I aminoacyl-tRNA synthetases (Fig. 2), a slight variation of the classic nucleotide binding fold, and rule out many other alternatives.

THREADING

Columnwise sequence alignment exploits the evolutionary conservation of functionally and structurally important residue positions but has its lim-

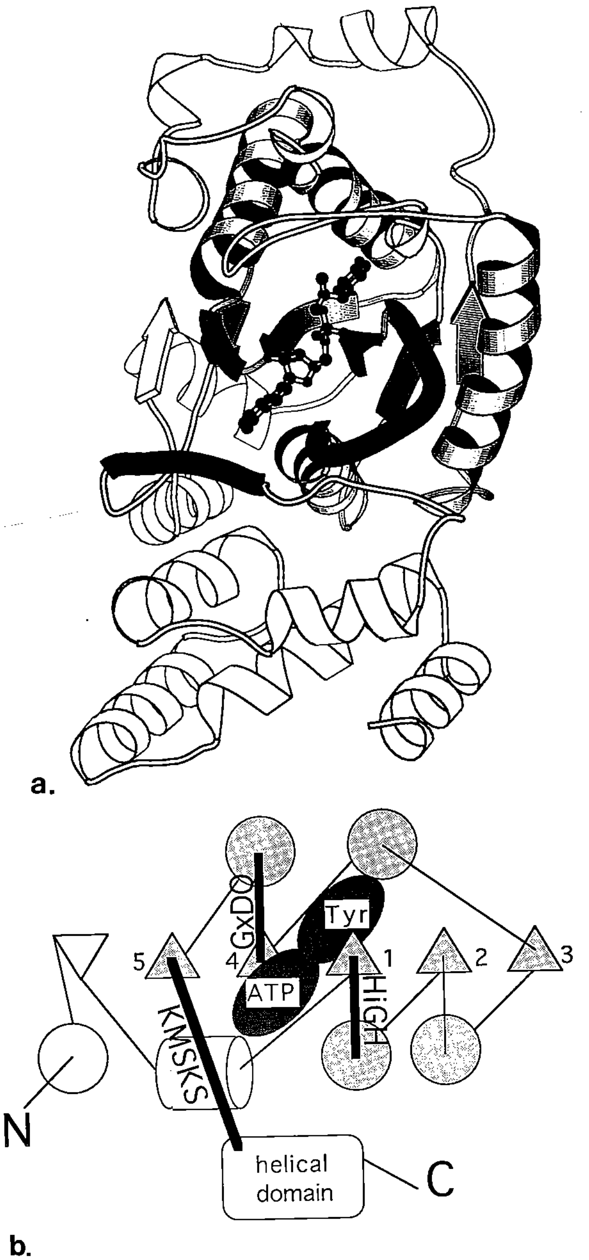

Fig. 2. Ribbon (**a**)[22] and topology (**b**) diagrams of a class I aminoacyl-tRNA synthetase in complex with tyrosyl-AMP.[5] The three loops shown as thick black lines form the ATP binding site in tRNA synthetase which are characterized by the HiGH and KMSKS sequence motifs. Tyrosyl- and a number of other class I aminoacyl-tRNA synthetases also have a conserved GxDQ motif involved in nucleotide-binding that is located in the loop following strand 4.[11] Light shading shows the region corresponding to the TagD sequence. TagD (129 residues) is much shorter than tyrosyl-tRNA synthetase (317 residues), but is comparable in length with other doubly wound α/β domains.

itations when the aligned sequences are very diverged. In threading, one attempts to take into account physical interactions which are cooperative in nature and involve residues that may be distant along the sequence. One of the clearest asymmetries in globular protein structures is that in the distribution of hydrophobic and polar residues in the protein interior and on the surface, respectively. This effect was used in optimizing the alignment of TagD onto the tyrosyl tRNA synthetase template structure, in the following semiautomatic iterative procedure.

The program MaxSprout[17] was used to build rough three-dimensional models of TagD, given an alignment to the template structure. The models were constrained so that the backbone coordinates of the template were kept fixed and only side chain orientations were optimized in rotamer space. In the initial sequence-structure alignment, any identified sequence motifs (HiGH, and see below) and predicted/actual secondary structure elements were matched. The corresponding 3D model was then evaluated using atomic solvation preference.[18] The sequence-structure alignment was then optimized by exhaustively examining the effect of shifts in the alignment on the next-generation 3D model and accepting favorable shifts.

Difficulties encountered in the process of generating good quality 3D models also gave feedback on the multiple sequence alignment and secondary structure prediction steps. In addition, solvation preference analysis strongly supports our identification of the hydrophobic block corresponding to residues 86–91 of TagD as beta strand (B4 in Fig. 1) rather than as a continuation of helix A3 as predicted by PHD (Figs. 1 and 3). The predicted fifth helix was modeled on the loop segment following B5 to preserve reasonable chain connectivity.

DESCRIPTION OF THE 3D MODEL OF TagD

The final sequence-structure alignment and solvation preference profile of the TagD model* are shown in Figure 3. The 3D model includes all 129 residues of the TagD sequence. Deletions relative to the tRNA synthetase template structure have been made in loops between points of close spatial proximity. The model contains a pronounced hydrophobic core as is typical in globular proteins. We see no unresolvable atomic clashes so the model should provide a reasonable starting point for further refinement by energy minimization. Most important, the model implies a detailed conservation of the nucleotide-binding site between the TagD superfamily and class I aminoacyl-tRNA synthetases.

The nucleotide binding site of class I aminoacyl-tRNA synthetases is formed by three loops (Fig. 2). These loops are characterized by the HiGH motif in loop B1–A1, the KMSKS motif (consensus sequence)

*The model coordinates are available via anonymous FTP from ftp.embl-heidelberg.de (Internet address) in the directory/pub/databases/protein_extras/models.

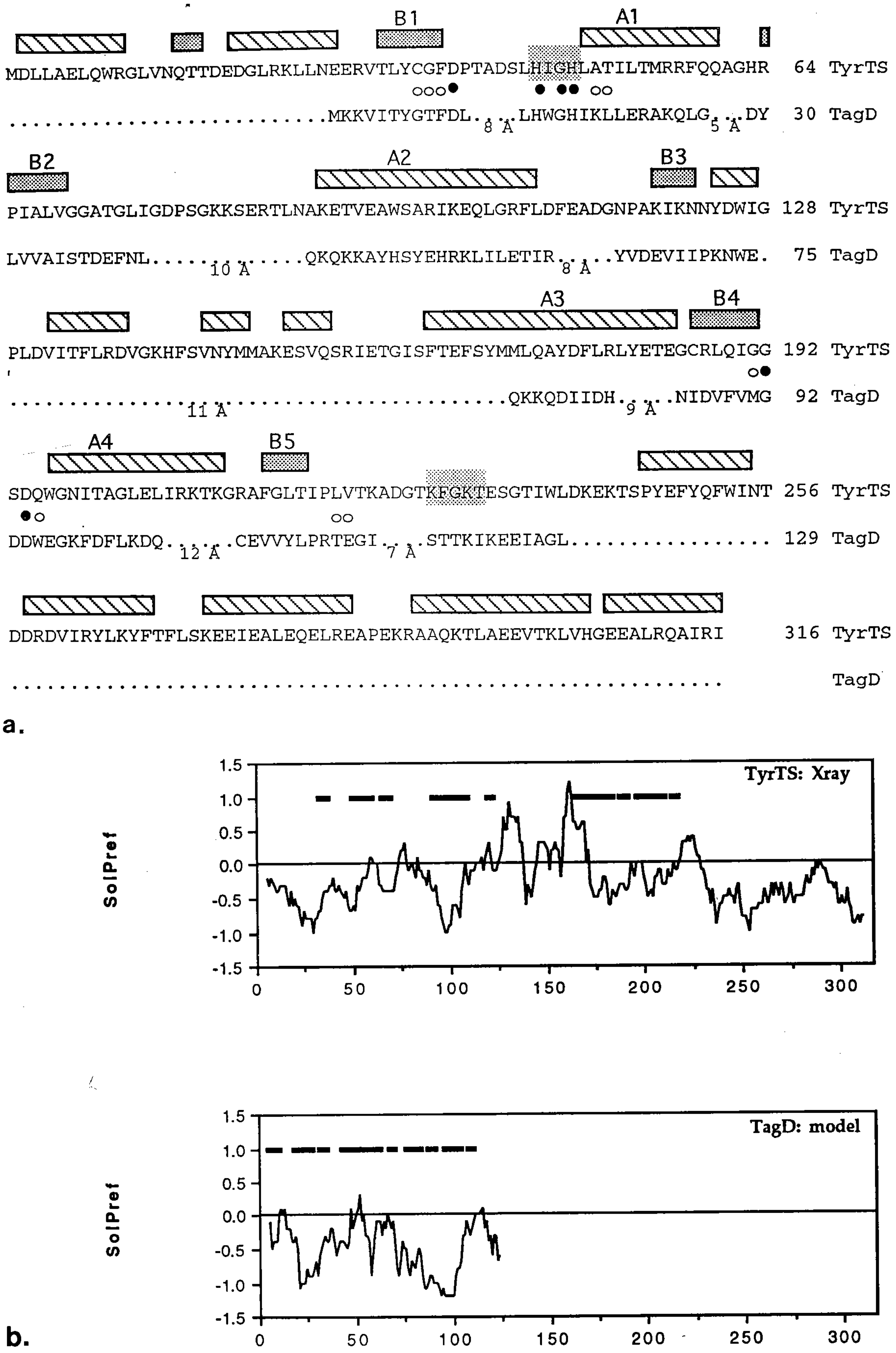
B1
A1
MDLLAELQWRGLVNQTTDEDGLRKLLNEERVTLYCGFDPTADSLHIGHLATILTMRRFQQAGHR 64 TyrTS
○○○● ● ●● ○○
..........................MKKVITYGTFDL. . . .LHWGHIKLLERAKQLG. . . .DY 30 TagD
8 Å
5 Å
B2
A2
B3
PIALVGGATGLIGDPSGKKSERTLNAKETVEAWSARIKEQLGRFLDFEADGNPAKIKNNYDWIG 128 TyrTS
LVVAISTDEFNL...............QKQKKAYHSYEHRKLILETIR. . . . .YVDEVIIPKNWE. 75 TagD
10 Å
8 Å
A3
B4
PLDVITFLRDVGKHFSVNYMMAKESVQSRIETGISFTEFSYMMLQAYDFLRLYETEGCRLQIGG 192 TyrTS
○●
....QKKQDIIDH. . . . .NIDVFVMG 92 TagD
11 Å
9 Å
A4
B5
SDQWGNITAGLELIRKTKGRAFGLTIPLVTKADGTKFGKTESGTIWLDKEKTSPYEFYQFWINT 256 TyrTS
●○ ○○
DDWEGKFDFLKDQ. . . . .CEVVYLPRTEGI. . . .STTKIKEEIAGL................. 129 TagD
12 Å
7 Å
DDRDVIRYLKYFTFLSKEEIEALEQELREAPEKRAAQKTLAEEVTKLVHGEEALRQAIRI 316 TyrTS
.... TagD
a.
TyrTS: Xray
SolPref
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
0
50
100
150
200
250
300
TagD: model
SolPref
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
0
50
100
150
200
250
300
b.

Fig. 3.

in the loop between B5 and the C-terminal domain (for review see ref. 10), and an invariant aspartic acid residue in the B4–A4 loop, at least in the subfamily of tyrosyl- and tryptophanyl-tRNA synthetases.[11] In our structural model, the strongest sequence conservation in the TagD superfamily maps exactly to these motifs.

An ATP molecule can be accommodated in the putative nucleotide binding site in the TagD model and can make similar interactions with the protein as in tyrosyl-tRNA synthetase. The ATP binding pocket is lined by 15 residues in the crystal structure of tyrosyl tRNA synthetase in complex with tyrosyl adenylate.[5] As many as six of these are either invariant or strongly conserved in both the TagD superfamily and class I aminoacyl-tRNA synthetases (black bullets in Fig. 3). In the known structure of tyrosyl tRNA synthetase, they are involved in binding the ribose (H-bonds from main chain of G192, side chains of D194 and H48) and phosphate groups of the nucleotide (H-bond from main chain of D38). Moreover, if ATP/CTP bind in similar orientations, it is most likely that the tyrosyl binding pocket will be used for glycerol-3-phosphate binding in TagD.

The loop carrying the KMSKS sequence motif of class I aminoacyl-tRNA synthetases moves towards the active site during the catalytic cycle.[12,26,30] The terminal position (serine/threonine) of the KMSKS motif is essential in stabilizing the transition state of tyrosyl-tRNA synthetases.[12] The corresponding region in the cytidylyltransferase superfamily succeeding the B5 strand shows an accumulation of serines and threonines. Given the similarities of the two other loops in the nucleotide binding site, we speculate that this serine/threonine-rich block (STTKI) is functionally related to the KMSKS motif.

There are some interesting residue substitutions at the proposed nucleotide binding site between the TagD model and tyrosyl-tRNA synthetase. W95 in TagD is positioned to make an equivalent hydrogen bond to the ribose ring as does Q195 in the known structure. The large volume change due to the mutation of G191 in the crystal structure of tyrosyl-tRNA synthetase for M91 in TagD is accommodated by exclusion of a buried crystal water molecule. Finally, we note that the second position in the HiGH motif is buried and occupied by hydrophobic residues in tRNA synthetases but exposed in the TagD superfamily so that polar residues become acceptable.

Fig. 3. (**a**) Result of threading the TagD sequence through the known structure of tyrosyl tRNA synthetase (TyrTS),[5] optimizing atomic solvation preference of the implied 3D model of TagD. Secondary structure elements in TyrTS are shown above the sequence (helix: stippled boxes, strands: gray). Elements of the structural core are labeled as in Figure 2. The HiGH and KMSKS (KFGKT) motifs are shaded in the TyrTS sequence. Residues that line the ATP binding pocket are indicated by circles between the sequences; filled circles denote strong conservation in *both* superfamilies. In addition, the flexible loop with the KFGKT motif moves during catalysis towards the ATP binding site and stabilizes the transition state.[12,26,30] Distances between neighboring C^α atoms across gaps are shown below the gaps. As a trans peptide bond places adjacent C^α atoms at a distance of 3.8 Å, the C^α atom of a chain terminal residue is in principle free to move up to 8 Å so the gaps could be closed in the TagD model with minor structural adjustments. (**b**) Atomic solvation preference profiles of the X-ray structure of tyrosyl tRNA synthetase (top) and of TagD model (bottom). The preference is here plotted as a pseudoenergy: the lower the better. The bars show the positions of the secondary structure elements B1, A1, B2, A2, B3, A3, B4, A4 and B5. The profiles are averaged over a window of eleven residues. Qualitatively, the profile for the TagD model is as favorable as that for the crystal structure of tyrosyl-tRNA synthetase. The two highest positive peaks in the profile for TyrTS correspond to the dimer interface[6] and thus correctly represent unsatisfied interactions.

CONCLUSION

Several independent lines of argument support the identification of the nucleotide-binding site and the fold prediction for TagD and its sequence relatives. This work shows that sophisticated sequence analysis in combination with sequence-structure alignment can be a productive approach to the prediction of three-dimensional protein structure.

REFERENCES

1. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., Lipman, D. J. Basic local alignment search tool. J. Mol. Biol. 215: 403–410, 1990.
2. Bairoch, A., Boeckmann, B. The SWISS-PROT protein sequence databank: Current status. Nucl. Acids Res. 22: 3578–3580, 1994.
3. Bork, P., Ouzounis, C., Sander, C. From genome sequences to protein function. Curr. Opin. Struct. Biol. 4:393–403, 1994.
4. Bott, M., Dimroth, P. *Klebsiella pneumoniae* genes for citrate lyase and citrate lyase ligase: Localization, sequencing, and expression Mol. Microb. 14:347–356, 1994.
5. Brick, P., Bhat, T.N., Blow, D.M. Structure of the tyrosyl-tRNA synthetase refined at 2.3 Ångströms resolution. Interaction of the enzyme with the tyrosyl adenylate intermediate. J. Mol. Biol. 208:83–98, 1989.
6. Brick, P., Blow, D.M. Crystal structure of a deletion mutant of a tyrosyl-tRNA synthetase complexed with tyrosine. J. Mol. Biol. 194:287–297, 1987.
7. Brändén, C.-I. Relation between structure and function of α/β proteins. Q. Rev. Biophys. 13:317–338, 1980.
8. Chothia, C. One thousand families for the molecular biologist. Nature (London) 357:543–544, 1992.
9. Clementz, T., Raetz, C.R.H. A gene coding for 3-deoxy-D-manno-octulosonic-acid transferase in *E. coli.* J. Biol. Chem. 266:9687–9696, 1991.
10. Delarue, M., Moras, D. The aminoacyl-tRNA synthetase family: modules at work. BioEssays 15:675–687, 1993.
11. Doublié, S., Bricogne, G., Gilmore, C., Carter, C.W., Jr. Tryptophanyl-tRNA synthetase crystal structure reveals an unexpected homology to tyrosyl-tRNA synthetase. Structure 3:17–31, 1995.
12. First, E.A., Fersht, A.R. Involvement of threonine 234 in catalysis of tyrosyl adenylate formation by tyrosyl-tRNA synthetase. Biochemistry 32:13644–13650, 1993.
13. Foster, J. W., Park, Y. K., Fenger, T., Spector, M. P. Regulation of NAD metabolism in *Salmonella typhimurium:* Molecular sequence analysis of the bifunctional nadR regulator and the nadA-pnuC operon. J. Bacteriol. 172:4187–4196, 1990.
14. Freter, A., Bowien, B. Identification of a novel gene aut, involved in autotrophic growth of *A. eutrophus.* J. Bacteriol. 176:5401–5408, 1994.
15. Goldman, R. C., Bolling, T. J., Kohlbrenner, W. E., Kim, Y., Fox, J. L. Primary structure of CTP:CMP-3-deoxy-D-manno-octulosonate cytidylyltransferase (CMP-KDO syn-

thetase) from *Escherichia coli.* J. Biol. Chem. 261:15831–15835, 1986.

16. Gribskov, M., McLachlan, A. D., Eisenberg, D. Profile analysis: Detection of distantly related proteins. Proc. Natl. Acad. Sci. U.S.A. 84:4355–4358, 1987.
17. Holm, L., Sander, C. Fast and simple Monte Carlo algorithm for side chain optimization in proteins: Application to model building by homology. Proteins 14:213–223, 1992.
18. Holm, L., Sander, C. Evaluation of protein models by atomic solvation preference. J. Mol. Biol. 225:93–105, 1992.
19. Holm, L., Sander, C. Searching protein structure databases has come of age. Proteins 19:165–173, 1994.
20. Horowitz, D. S., Abelson, J. M. A small U5 ribonucleotide particle protein involved only in the second step of RNA-splicing in *S. cerevisiae.* Mol. Cell. Biol. 13:2959–2970, 1993.
21. Kalmar, G. B., Kay, R. J., Lachance, A., Aebersold, R., Cornell, R. B. Cloning and expression of rat liver CTP: phosphocholine cytidylyltransferase: An amphiphatic protein that controls phosphatidylcholine synthesis. Proc. Natl. Acad. Sci. U.S.A. 87:6029–6033, 1990.
22. Kraulis, P. MOLSCRIPT: A program to produce both detailed and schematic plots of protein structures. J. Appl. Crystallogr. 24:946–950, 1991.
23. Mauel, C., Young, M., Karawata, D. Genes concerned with synthesis of poly (glycerol phosphate), the essential teichoic acid in *B. subtilis* strain 168, are organized in two divergent transcription units. J. Gen. Microbiol. 37:929–941, 1991.
24. Nichols, J. C., Vyas, N. K., Quiocho, F. A., Matthews, K. S. Model of lactose repressor core based on alignment with sugar-binding proteins is concordant with genetic and chemical data. J. Biol. Chem. 268:17602–17612, 1993.
25. Orengo, C. A., Jones, D. T., Thornton, J. M. Protein superfamilies and domain superfolds. Nature (London) 372:631–634, 1994.
26. Perona, J. J., Rould, M. A., Steitz, T. A. Structural basis for transfer RNA aminoacylation by *E. coli* glutaminyl-tRNA synthetase. Biochemistry 32:8758–8771, 1993.
27. Rohde, K., Bork, P. A fast, sensitive pattern-matching approach for protein sequences. Comput. Appl. Biosci. 9:183–189, 1993.
28. Rost, B and Sander, C. Combining evolutionary information and neural networks to predict secondary structure. Proteins 19:55–72, 1994.
29. Sander, C., Schneider, R. Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins 9:56–68, 1991.
30. Schmitt, E., Meinnel, T., Blanquet, S., Mechulan, Y. Methionyl-tRNA synthetase needs an intact and mobile KMSKS motif in catalysis of methionyl adenylate formation. J. Mol. Biol. 242:566–577, 1994.
31. Schuler, G. D., Altschul, S. F., Lipman, D. J. A workbench for multiple sequence alignment construction and analysis. Proteins 9:180–190, 1991.
32. Schulz, G. E. Binding of nucleotides by proteins. Curr. Opin. Struct. Biol. 2:61–67, 1992.
33. Tatusov, R., Altschul, S. F., Koonin, E. V. Detection of conserved segments in proteins: Iterative scanning of sequence databases with alignment blocks. Proc. Natl. Acad. Sci. U.S.A. 91:12091–12095, 1994.
34. Thompson, J. D., Higgins, D. G., Gibson, T. J. CLUSTALW: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, Nucl. Acids Res. 22:4673–4680, 1994.
35. Tsukagoshi, Y., Nikawa, J. I., Yamashita, S. Molecular cloning and characterisation of the gene encoding cholinphosphate cytidylyltransferase in *S. cerevisiae.* Eur. J. Biochem. 169:477–486, 1987.
36. Yura, T., Mori, H., Nagai, H., Nagata, T., Tsukihara, A., Fujita, N., Isono, K., Mizobuchi, K., Nakata, A. Systematic sequencing of the *E. coli* genome: Analysis of the 0–2.4 min region. Nucl. Acids Res. 20:3305–3308, 1992.
37. Zhu, N., Roth, J. R. The nadI region of *S. typhimurium* encodes a bifunctional regulatory protein, J. Bacteriol. 173:1302–1310, 1991.
38. Zhu, N., Olivera, B. M., Roth, J. R. Activity of the nicotinamide nucleotide transport system in *Salmonella typhimurium.* J. Bacteriol. 173:1311–1320, 1991.