Scheer, H. S., & Katz, J. J. (1975) in *Porphyrins and Metalloporphyrins* (Smith, K. S., Ed.) pp 339-524, Elsevier, New York.

Sheard, B., Yamane, T., & Shulman, R. G. (1970) *J. Mol. Biol. 53*, 25-48.

Shulman, R. G., Wüthrich, K., Yamane, T., Antonini, E., & Brunori, M. (1969) *Proc. Natl. Acad. Sci. U.S.A. 63*, 623-628.

Smith, K. M., Simpson, D. J., & Snow, K. M. (1986) *J. Am. Chem. Soc. 108*, 6834-6835.

Timkovich, R., & Vavra, M. R. (1985) *Biochemistry 24*, 5189-5196.

Traylor, T. G., & Traylor, P. S. (1982) *Annu. Rev. Biophys. Bioeng. 11*, 105-127.

Wright, K. A., & Boxer, S. G. (1981) *Biochemistry 20*, 7546-7556.

Prediction of a Common Structural Domain in Aminoacyl-tRNA Synthetases through Use of a New Pattern-Directed Inference System[†]

Teresa A. Webster,* Richard H. Lathrop, and Temple F. Smith

Molecular Biology Computer Research Resource, Dana-Farber Cancer Institute, Harvard School of Public Health, Boston, Massachusetts 02115, and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139

Received January 7, 1987; Revised Manuscript Received May 13, 1987

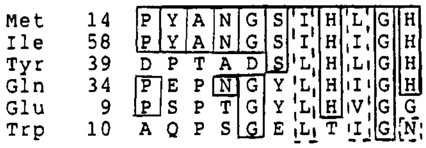
ABSTRACT: The aminoacyl-tRNA synthetases are united by a common function with little evidence of a common structural relationship. Outside of an 11 amino acid stretch called the "signature sequence", no global primary sequence similarity exists. The signature sequence matches 4-11 amino acids in several aminoacyl-tRNA synthetases. High-resolution X-ray data are available for two of these enzymes, revealing that their signature sequence regions are small segments of a common mononucleotide binding foldlike structure. A new methodology for the analysis of dissimilar primary sequences supports the expectation that all of the signature sequence regions form a common structure. In our analysis, two complex pattern descriptors were constructed to describe the synthetase mononucleotide binding fold. These were compared to primary sequences annotated with predicted secondary structures and hydropathy profiles. Regions in 8 out of 12 (67%) heterologous aminoacyl-tRNA synthetase groups (where each group is specific for the same amino acid) match the first descriptor, and 7 of these (58%) also match the second descriptor. In contrast, only 4 regions in a set of 54 control proteins (7.4%) match the first descriptor, and only 2 regions (3.7%) match both. Alignment of these 8 regions to the descriptor (1) positions all known signature sequence regions as the first loop of a mononucleotide binding foldlike structure, (2) extends the previous alignments by another 40-odd amino acids, and (3) identifies potential sites in 3 out of 6 heterologous aminoacyl-tRNA synthetases with no previous alignments. Potential sites are also proposed for two additional heterologous synthetases on the basis of matches to less specific descriptors.

Aminoacyl-tRNA synthetases share a common function, which is to attach an amino acid to its cognate tRNA in protein biosynthesis. Despite this, they have little to unite them as common protein structures (Schimmel & Söll, 1979; Schimmel, 1987). Their quaternary structures vary from α, α_2, and α_4 to $\alpha_2\beta_2$, and the individual subunits, which contain a complete set of substrate sites, range in size from 300 to 1000 amino acids. Twenty-two aminoacyl-tRNA synthetase sequences have been generated from three bacterial species and from *Saccharomyces cerevisiae* [reviewed by Schimmel (1987)]. These form 12 heterologous groups of synthetases, each specific for the same amino acid. Throughout this paper, synthetases in the same group are usually referred to as one type of 12 heterologous synthetases. Synthetases within the same group share primary sequence similarities of high statistical significance and are therefore believed to be homologous. However, computer searches for sequence similarities between pairs of synthetases from different groups have not revealed any extended regions of similarity (Hountondji et al., 1986a,b; Schimmel, 1987).

Four lines of evidence suggest that common structures may exist among heterologous synthetases. One, high-resolution X-ray structures are available for *Bacillus stearothermophilus* Tyr-tRNA synthetase and a fragment of *Escherichia coli* Met-tRNA synthetase. These reveal the existence of a common structure found in many nucleotide binding proteins: the Rossman mononucleotide binding fold (MNBF) (Rossman et al., 1975; Zewler et al., 1982; Blow et al., 1983). From cocrystal structures of the nucleotide-bound enzymes and from site-specific mutagenesis studies, the function of this domain is implied to be the binding of ATP and the aminoacyl adenylate [reviewed by Blow and Brick (1985)]. Two, the Ile-, Glu-, Gln-, and Trp-tRNA synthetases, whose structures are unknown, all contain "signature sequence" regions (Figure 1) which match sequences within the Tyr- and Met-tRNA synthetase MNBF-like structures (Barker & Winter, 1982; Webster et al., 1984; Myers & Tzagoloff, 1985; Breton et al.,

[†] This work was supported in part by NIH Grant RR02275 and by DARPA under ONR Contract N0014-80-C-0505.

* Address correspondence to this author at the Molecular Biology Computer Research Resource, Dana-Farber Cancer Institute.



Met	14	P	Y	A	N	G	S	I	H	L	G	H
Ile	58	P	Y	A	N	G	S	I	H	I	G	H
Tyr	39	D	P	T	A	D	S	L	H	L	G	H
Gln	34	P	E	P	N	G	Y	L	H	I	G	H
Glu	9	P	S	P	T	G	Y	L	H	V	G	G
Trp	10	A	Q	P	S	G	E	L	T	I	G	N

FIGURE 1: Alignment of "signature sequence" regions of *E. coli* aminoacyl-tRNA synthetases. *Solid boxes* indicate identities, and *dashed boxes* indicate conservative substitutions. *Numbers* indicate the first amino acid position. The signature sequence is defined as a match of 11 amino acids (10 identities and 1 conservative change) between *E. coli* Ile-tRNA synthetase and *E. coli* Met-tRNA synthetase (Webster et al., 1984).

1986). Three, affinity labeling studies have led to identification of lysine residues at the CCA binding site of tRNA in *E. coli* Met- and Tyr-tRNA synthetase (Hountondji et al., 1985, 1986a,b). Sequence similarities of 4–8 amino acids have been found between this 11 amino acid "Lys-335" region of Met-tRNA synthetase and the Ile-, Trp-, Tyr-, Gln-, Phe-, and Ala-tRNA synthetases (Hountondji et al., 1986a). Four, deletion studies of the larger synthetases suggest that variability in subunit lengths is not caused by differences in the sizes of the catalytic domains but results from variable fusions of extra polypeptide domains to a catalytic core (Cassio et al., 1971; Waye et al., 1983; Jasin et al., 1983; Schimmel et al., 1984).

In this study, we tested the expectation that all signature sequence regions are part of a common MNBF, by using a methodology which employs a new pattern-directed inference system. It is designed to identify probable common structures among dissimilar primary sequences by allowing search descriptors to include predicted secondary structure elements. In order to confirm its validity, the method was first applied to the set of NAD and FAD binding proteins which were reviewed by Birktoft and Banaszak (1984) and are known to have a common structural motif. Subsequent application to aminoacyl-tRNA synthetases, one, identifies the signature sequence region as the first loop of a potential MNBF structure, two, extends the proposed alignment approximately another 40 amino acids, and three, identifies potential MNBF regions in 3 heterologous synthetases which do not contain an identifiable signature sequence.

There are four steps to the method (Figure 2): One is analysis of common structural motifs shared by the X-ray structures to construct an initial pattern descriptor. Two, the primary sequences of two sets of proteins—one united by the common function and the other a representative control set—are annotated with predicted secondary structures. Three, the pattern-directed inference system ARIADNE (Lathrop et al., 1987)[1] carries out the search and identification of matches between the descriptor and the two sets of proteins. Four, the current descriptor is iteratively refined to increase its ability to discriminate between the functional and control sets.

There are three components to this method. The first is the descriptor vocabulary. The second is a procedure for constructing and optimizing the descriptor. The third is the pattern-directed inference system [previously described in Lathrop et al. (1987);[1] see Materials and Methods], which searches for occurrences of the descriptor within the annotated protein sequences.

The descriptor vocabulary currently consists of primary sequence elements, "higher-order" elements, and "spacer" elements. Primary sequence elements can be a specific amino acid or a class. Higher order elements are combinations of primary sequence elements and/or secondary structure elements that can be inferred from the primary sequence by a variety of empirical prediction schemes. These schemes suffer from accuracies of only 50–70% and include underprediction, overprediction, and boundary errors (Schultz & Schirmer, 1979; Kabsch & Sander, 1983), and so the vocabulary is constructed to include a tolerance for error as described below.

An initial descriptor is constructed on the basis of analysis of X-ray structures of common structural motifs and available biochemical and genetic data. This contains only the minimum primary and secondary structure elements which appear essential. The complexity of the initial descriptor is iteratively increased or decreased to maximize its discrimination between the functionally related proteins and a representative control set.

Much of the utility of the method arises from the inclusion of complex, higher order elements. Previous pattern-matching methods have utilized complex combinations of primary sequence elements, which in some cases represent secondary structure elements (Cohen et al., 1986; Taylor, 1986; Wierenga et al., 1986; Bashford et al., 1987; Gascuel & Danchin, 1987). Our method differs in that it directly utilizes secondary structure elements generated by empirical prediction schemes. This increases the sensitivity of the descriptors so that, for example, we identify potential MNBF secondary structure elements in synthetases, whereas, a primary sequence pattern for a similar $\beta\alpha\beta$-fold (Wierenga et al., 1986) does not identify any potential sites in synthetases.

There are several factors which may influence the accuracy of descriptors containing these higher order elements. One, a tolerance for error is built into the descriptor vocabulary. Each element's significance in the pattern is allowed to vary by weighting its contribution to an overall "similarity score" (Materials and Methods). A tolerance for secondary structure boundary errors is included in two ways. A primary sequence element can be combined with a predicted secondary structure element within a range of positions, and predicted secondary structure and higher order elements can be flanked by "spacer elements". Spacer elements, "$X_{i,j}$", are amino acid strings of any type within a length range i–j. Two, a minimal descriptor complexity is required to reduce the background of false positive matches (sites which do not in fact fold into the descriptor structure) in the control set. This controls secondary structure overprediction to a certain extent because, the more complex the descriptor, the more unlikely it becomes that the required sequence of primary and secondary structure elements will occur by chance.

A final factor which may influence descriptor accuracy is our optimization of secondary structure prediction. Our PRSTRC (Ralph et al., 1987) implementation of the Chou/Fasman (Chou & Fasman, 1978) prediction scheme exploits the fact that different protein structural classes have different biases for optimal prediction (Garnier, 1978), by giving the user control over several parameter values. We adjusted these so that the α-helices and β-sheets of the descriptor structure (Figure 4A) were predicted with 100% accuracy, and those of the entire N-terminal $\alpha\beta$ domains (Blow et al., 1983) of the Met- and Tyr-tRNA synthetases were predicted with 75% accuracy (Materials and Methods). These optimized parameters may improve the structure prediction accuracy for the synthetases with unknown three-dimensional structure. Independent predictions from more than one prediction scheme were combined. Hydrophobic moment (described below), used to predict α-helices, is completely independent, and average

[1] In this system, when the rule pattern is recognized, the action is to construct a higher order structure. Therefore, this expert system differs in that it does not employ the more heuristic "rule of thumb" knowledge typically associated with rule-based expert systems.

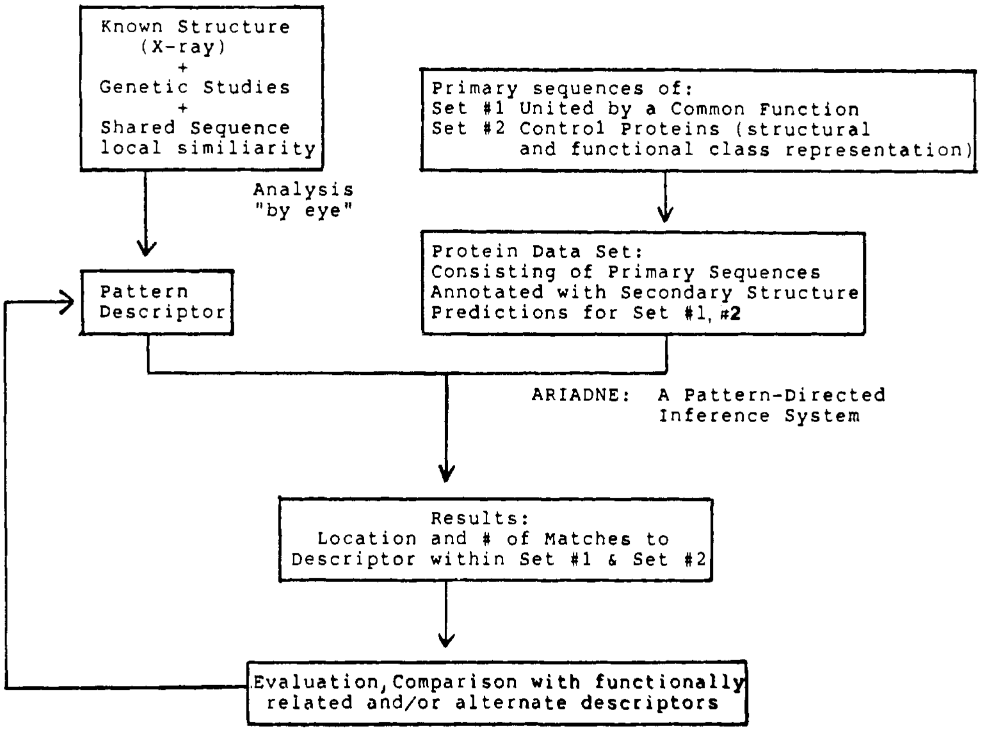

Known Structure
(X-ray)
+
Genetic Studies
+
Shared Sequence
local similiarity
Analysis
"by eye"
Pattern
Descriptor
Primary sequences of:
Set #1 United by a Common Function
Set #2 Control Proteins (structural
and functional class representation)
Protein Data Set:
Consisting of Primary Sequences
Annotated with Secondary Structure
Predictions for Set #1, #2
ARIADNE: A Pattern-Directed
Inference System
Results:
Location and # of Matches to
Descriptor within Set #1 & Set #2
Evaluation, Comparison with functionally
related and/or alternate descriptors

FIGURE 2: Flow chart of the methodology which utilizes the pattern-directed inference system.

hydropathy, used to predict β-strands, is partially independent of the Chou/Fasman predictions. Finally, when the same region is predicted to fold into more than one structure, each prediction is separately included in the search data set. This factor substantially reduces underprediction.

Our study utilized two descriptors which contain the same secondary structure elements predicted by two different schemes. A third "composite" descriptor required that secondary structure elements be predicted by both schemes. One scheme uses the Chou/Fasman pseudoprobability constants (Chou & Fasman, 1978) and was carried out by the program PRSTRC (Ralph et al., 1986); the other uses hydropathy values (Kyte & Doolittle, 1982) and is described under Materials and Methods. Hydropathy values imply protein structure because of the following relationships: one, α-helices that lie at the protein surfaces tend to have a hydrophobic and hydrophilic face (Eisenberg et al., 1982) and consequently a large "hydrophobic moment" (Eisenberg et al., 1984); two, β-sheets, which frequently occur in the interior of globular proteins, tend to be particularly rich in hydrophobic residues (Chou & Fasman, 1978).

Materials and Methods

Hydropathy-Based Secondary Structure Predictions. Peaks in hydropathy profiles, "H-peak" elements, were assigned as potential β-strands. A hydropathy profile was generated by determining the Gaussian-weighted average hydropathy, $H(n)$, of a short segment and attaching the value to the central amino acid, n, in the segment. $H(n)$ is calculated as follows:

$$H(n) = \frac{\sum_{i=-w}^{w} e^{-(i/s)^2} Y(n+i)}{\sum_{i=-w}^{w} e^{-(i/s)^2}} \qquad (1)$$

where Y is the numerical hydropathy value from the scale of Kyte and Doolittle (1982) of the $(n+i)$th residue. We set w (the window) and s (the smoothing factor) equal to 4 and 2, respectively.

Peaks in hydrophobic moment profiles, "P-peak" elements, were assigned as potential α-helices. Hydrophobic moment (Eisenberg et al., 1984) is a value which reflects the degree to which the hydropathy profile of a periodic protein structure alternates between being hydrophilic and hydrophobic. A hydrophobic moment profile was generated by determining the hydrophobic moment, $P(n)$, for a segment and attaching the value to the central amino acid, n, in the segment. $P(n)$ is calculated as follows:

$$P(n) = \left\{\left[\sum_{i=-w}^{w} \sin\left(\frac{2\pi i}{p}\right) Y(n+i)\right]^2 + \left[\sum_{i=-w}^{w} \cos\left(\frac{2\pi i}{p}\right) Y(n+i)\right]^2\right\}^{1/2} \qquad (2)$$

We set p (the period) to 3.6 and w to 3. The profile was then smoothed by computing Gaussian-weighted averages for segments using eq 1 with $P(n+i)$ substituted for $Y(n+i)$, w equal to 15, and s equal to 9.

In both cases, the size of the Gaussian was optimized for peaks which corresponded to the nucleotide binding fold surface helices and buried β-strands of *B. stearothermophilus* Tyr-tRNA synthetase.

Chou/Fasman-Based Secondary Structure Predictions. Secondary structure predictions were carried out by PRSTRC, a modified Chou–Fasman analysis. The following user input values [defined in Ralph et al. (1986)] were used for aminoacyl-tRNA synthetases: α former = 1.12; α threshold = 1.08; α cutoff = 1.00; β former = 1.23; β cutoff = 0.97; minimum turn value = 0.50. These parameter values optimized predictions of the *B. stearothermophilus* Tyr-tRNA synthetase secondary structure elements [see Figure 2 of Winter et al. (1983)] and *E. coli* Met-tRNA synthetase (S. Brunie, personal communication) as described above.

Construction of Control Set. Fifty-four proteins were selected for the control set in order to create a set which structurally and functionally represents known proteins. Two proteins were selected from each of the 27 functional groups of the PIR/NBRF data base (George et al., 1986), one with a

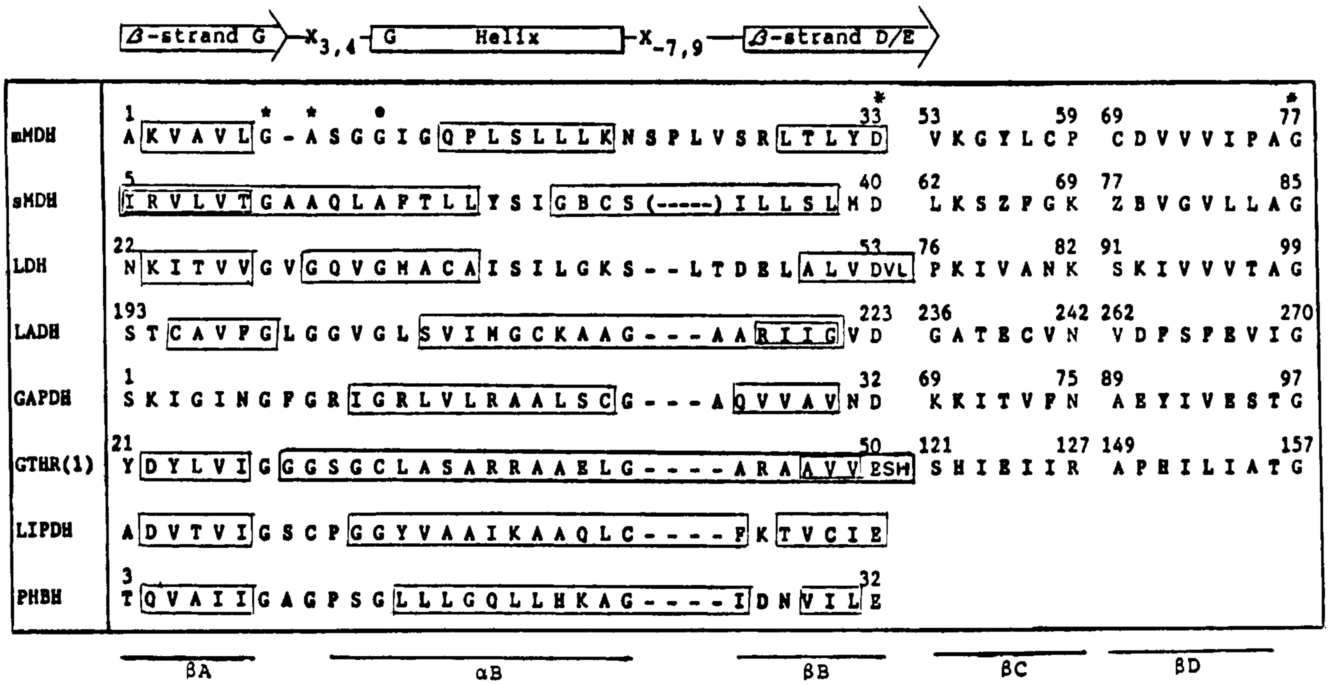

β-strand G
X3,4
G Helix
X-7,9
β-strand D/E
mMDH 1 * * • 33 53 59 69 77
A KVAVL G-ASGGIG QPLSLLLK NSPLVSR LTLYD VKGYLCP CDVVVIPAG
sMDH 5 40 62 69 77 85
IRVLVT GAAQLAPTLL YSI GBCS(-----)ILLSL HD LKSZPGK ZBVGVLLAG
LDH 22 53 76 82 91 99
N KITVV GV GQVGMACA ISILGKS--LTDBL ALVDVL PKIVANK SKIVVVTAG
LADH 193 223 236 242 262 270
ST CAVFG LGGVGL SVIMGCKAAG---AA RIIG VD GATBCVN VDFSPBVIG
GAPDH 1 32 69 75 89 97
SKIGINGFGR IGRLVLRAALSC G---A QVVAV ND KKITVFN ABYIVBSTG
GTHR(1) 21 50 121 127 149 157
Y DYLVI G GGSGCLASARRAABLG----ARA AVVESH SHIBIIR APHILIATG
LIPDH
A DVTVI GSCP GGYVAAIKAAQLC----F K TVCIE
PHBH 3 32
T QVAII GAGPSG LLLGQLLHKAG----I DN VILE
βA
αB
βB
βC
βD

FIGURE 3: Alignment of the descriptor with dinucleotide binding domains (Birktoft & Banaszak, 1984; Figure 4, reprinted with permission). The descriptor of the first $\beta\alpha\beta$-fold of the dinucleotide binding domain (described in text) is at the top of the figure. The amino acid regions predicted to fold into β-strand A (βA), α-helix B (αB), and β-strand B (βB) are enclosed in the first, second, and third *box*, respectively, of each sequence. The secondary structure assignments from the X-ray data are indicated by *solid bars* below the figure. Abbreviations for enzyme names are as follows: mMDH, mitochondrial malate dehydrogenase; sMDH, cytosolic malate dehydrogenase; LDH, lactate dehydrogenase; LADH, liver alcohol dehydrogenase; GAPDH, glyceraldehyde-3-phosphate dehydrogenase; GTHR, glutathione reductase; LIPDH, lipoamide dehydrogenase; PHBH, *p*-hydroxybenzoate hydroxylase. Asterisks indicate the positions of conserved amino acids.

known structure and one with an unknown structure. Among the set of 27 proteins with known structures, an attempt was made to select evenly from the 4 structural classes defined by J. Richardson (Richardson, 1981): α, α/β, β, and small and disulfide rich. This test set is thus intentionally a nonrandom collection drawn to uniformly sample the space of known protein types. However, selection within each functional type and structural class was at random. The PIR/NBRF locus names of this set are HRTHBD, MYSLG, LZBPT4, FXCLEX, XNCHDC, R5EC7, L2HUMC, CCBN, IHKREV, CSBO, ISCHT, KIBYG, TISY, CVJB, JGECA, RCBPL, VCTNS, NTSR3C, IPPG, VBHU, KLBOI, YKPG, CPBOG2, PSPGA, UHHU2, QXBO6L, ACHUA1, NFRT1, BHTLD, O4RBP2, MCSW, HSSF2M, APBPML, RKEGL, HLMSAB, C3MSAT, IVMSA1, AXPG, R3EC1, UBCHA, LQBP37, QXHU6M, WMEC15, DECHE, BMTD, FJEC, FOFV1R, KIRBCM, QRHULD, TXEC, RGBYA2, FSFB, YVBPMS, and SYECCS.

The above set was used for the synthetase descriptor without filtering of nucleotide binding proteins, since the structure of the synthetase MNBF is considerably more complex than that of other nucleotide binding classes (Results). The synthetase descriptor incorporates these differences. The control set for the NAD/FAD binding descriptor was filtered to remove the nine nucleotide binding proteins, since the structures of the first $\alpha\beta$ unit of known ATP, GTP, and dinucleotide binding classes (Results) are usually quite similar.

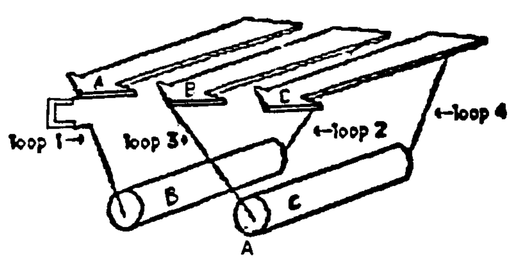
ARIADNE. A hierarchical, pattern-directed inference system, ARIADNE [described fully in Lathrop et al. (1987)],[1] is used to search multiple paths through the network of multiple secondary structure predictions. Direct representation of higher order protein structures as pattern elements is allowed. ARIADNE finds optimal alignments of the descriptor after allowing gaps and insertions.

In this implementation, "matches" were regions which gave positive "similarity scores" [described in Lathrop et al. (1987)][1] with the descriptor. Match scores were 1 for amino acids, "H-peak" elements, and "P-peak" elements; and the value of the averaged pseudoprobability constants for PRSTRC-predicted secondary structure elements. Mismatches were not allowed for any elements except the spacer elements. The mismatch score for these was $-(1+r/3)$ where r is the number of amino acids outside of the specified range.

RESULTS

NAD and FAD Binding Proteins. We calibrated the method against a set of known structures, the "dinucleotide binding domains". The dinucleotide binding domain exists in a class of enzymes called oxidoreductases. X-ray structures of seven unrelated oxidoreductases show that they share a common structural motif of four parallel β-strands and at least one α-helix which have the same connectivity and topology (Birktoft & Banaszak, 1984). Six unrelated NAD and FAD binding domains have been aligned on the basis of conformational similarities (sMDH, LDH, LADH, GAPDH, GTHR, and PHBH) and two by homology (mMDH and LIPDH). This alignment shows that only 5 out of approximately 95 structurally equivalent amino acids are conserved (Birktoft & Banaszak, 1984) (Figure 3).

We constructed a single descriptor which correctly locates the dinucleotide binding domain in six of these eight enzymes (75%) when the initial input is only primary sequence. The descriptor (top of Figure 3) contains three higher order elements: (1) a predicted β-strand with a Gly within one residue of the C-terminus; (2) a predicted α-helix with a Gly within four residues of the N-terminus; and (3) a predicted β-strand with an Asp or Glu within two residues of the C-terminus. One boundary of each higher order element was set at the position of the included primary sequence element and the other at the terminus of the predicted secondary structure. The three primary sequence elements play important functional and structural roles in the dinucleotide binding domain (Birktoft & Banaszak, 1984). Between the higher order elements are spacer elements: $X_{3,4}$ and $X_{-7,9}$. The -7 represents an allowed



A

βA	LOOP 1	αB	LOOP 2	βB	Descriptor
β-strand	$X_{0,4}$-turn-$X_{2,6}$-	G[HNQ] Helix----	$X_{0,11}$-	β-strand	1
H-Peak -	$X_{2,8}$-turn-$X_{2,6}$-	G[HNQ] $X_{0,7}$-P-Peak	$X_{11,21}$-	H-peaK	2

B

FIGURE 4: Mononucleotide binding foldlike structure and descriptors. (A) Schematic drawing of mononucleotide binding foldlike structure (Zelwer et al., 1982; Blow et al., 1983, 1985). The β-strands are represented by *arrows*, and the α-helices are represented by *cylinders*. The *bracket* in loop 1 represents a turn. (B) Descriptors 1 and 2 (described in the text).

overlap of seven amino acids between the predicted C-terminus of element 2 and the N-terminus of element 3.

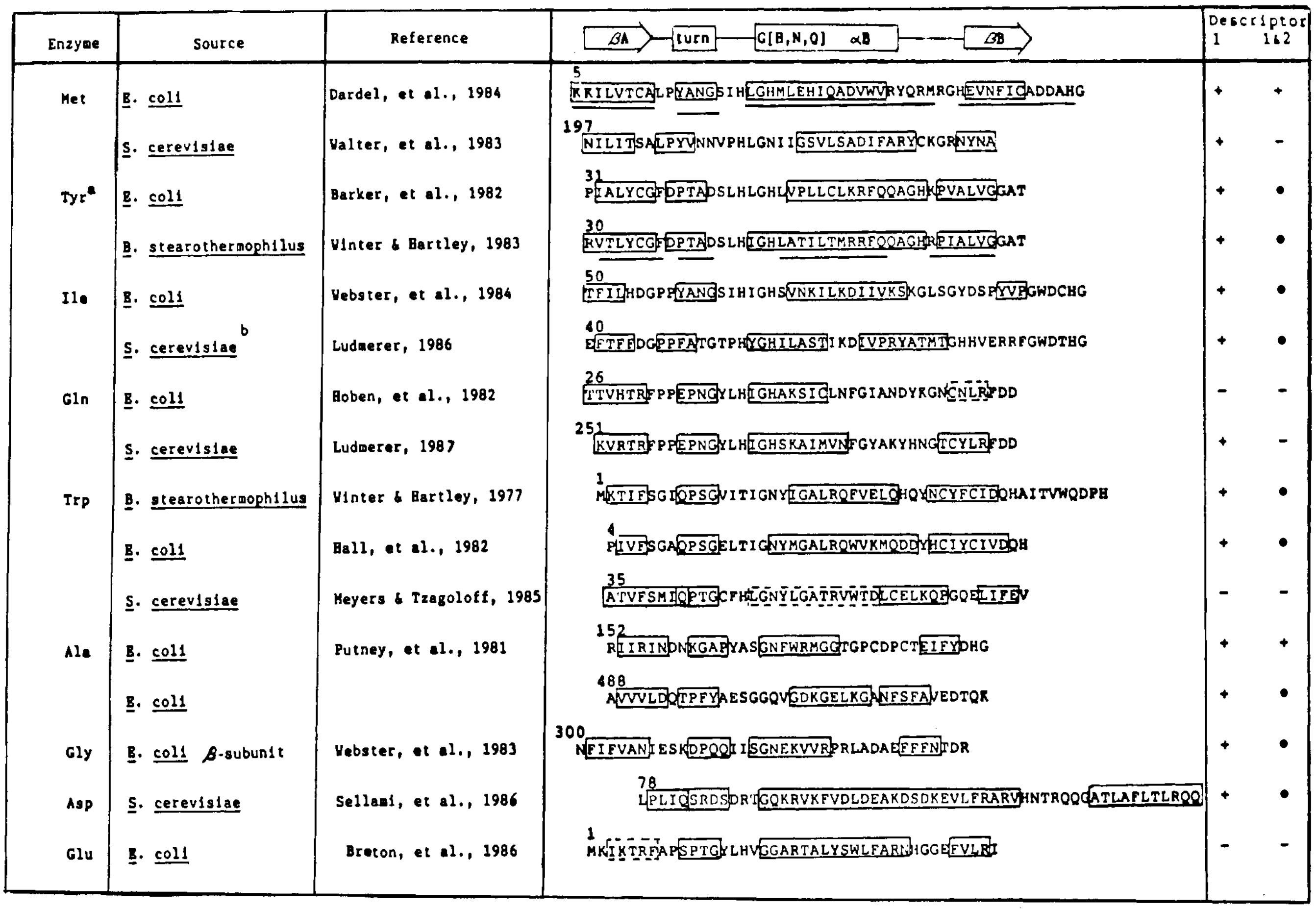
The dinucleotide binding sites of sMDH and GAPDH (Figure 3) do not match this descriptor. In sMDH, the α-helix N-terminal Gly has been replaced by Ala-16, and the N-terminus of the predicted helix is more than four residues beyond this Ala. β-Strand A is not predicted in GAPDH. This shows that missed identification can result if the descriptor is too specific and/or if the secondary structure predictions are inaccurate. One false positive match occurred within amino acids 198–235 in LDH in addition to the six correct matches. In this region, the α-helix which corresponds to α-helix B in the descriptor structure was erroneously predicted. Since confirmation of X-ray structures will not usually be available, estimation of the extent of false positive background can be determined only by comparison to the set of control proteins. Only 2 out of 45 (4.4%) control proteins (Materials and Methods) match the descriptor.

Aminoacyl-tRNA Synthetases. The two well-defined aminoacyl-tRNA synthetase X-ray structures share a common structure which resembles the Rossman mononucleotide binding fold (MNBF) (shown schematically in Figure 4A). Loop 1 of this structure contains the signature sequence. Two descriptors (Figure 4B) were constructed to describe this MNBF. They contain higher order elements which occur in the first $\beta\alpha\beta$-fold formed by β-strand A, loop 1, α-helix B, loop 2, and β-strand B. The four higher order elements of the first descriptor (top row of Figure 4B) contain secondary structure elements predicted by Chou/Fasman pseudoprobabilities (Materials and Methods). These elements are (1) a predicted β-strand, (2) a predicted β-turn, (3) a predicted α-helix with the following dipeptide—a Gly followed by either His, Asn, or Gln within four residues of the N-terminus, and (4) a predicted β-strand. The higher order elements of the second descriptor (bottom row of Figure 4B) are similar except that the predicted α-helix (P-peak) and β-strands (H-peaks) are based on hydropathy values (Materials and Methods).

Construction of descriptors 1 and 2 with the dipeptide G[HNQ] as an essential element was obtained after partial descriptor optimization and is based on the following information. The sequence H[bulky hydrophobe]GH (amino acids 45–48 in *B. stearothermophilus* Tyr-tRNA synthetase) is the most conserved portion of the signature sequence region (Figure 1). The Gly residue occupies an analogous position in a common nucleotide binding $\alpha\beta$-unit, which is characterized by a loop connecting a β-strand at the carboxyl end of a β-sheet to an approximately antiparallel α-helix (Schultz & Shirmer, 1979). This $\alpha\beta$-unit—which is found in GTP binding EF-Tu (la Cour et al., 1985), ATP binding adenylate kinase (Fry et al., 1986), and dinucleotide binding oxidoreductases (Rossman et al., 1975; Birkloft & Banaszak, 1984)—differs from the two known synthetase $\alpha\beta$-units in that the loop is much shorter and changes direction much more abruptly (Blow & Brick, 1985). However all $\alpha\beta$-units have in common a glycine found at the C-terminus of the loop close to and/or at the beginning of the helix (the third asterisk in Figure 3). Occasionally, an alanine is found in this position (Birkloft & Banaszak, 1984) and is tolerated in the genetically engineered Tyr-tRNA synthetase (Brown et al., 1986). In oxidoreductases, a larger side chain would prevent compact folding of the $\alpha\beta$-unit (Birkloft & Banaszak, 1984; Wierenga et al., 1985). Since this glycine is the only highly conserved primary sequence element found across all nucleotide binding classes, it was included in all descriptors as an essential element.

In the optimization process, we tested descriptors with residues corresponding to the *B. stearothermophilus* Tyr-tRNA synthetase His-45 position, which plays an apparent catalytic role (Leatherbarrow et al., 1985), and the His-48 position, which makes an apparent H-bond contact with the ATP (Lowe et al., 1985). We did not test restrictions on the type of residue allowed in the "bulky hydrophobe" position since it does not have an apparent essential role. We found that a version of descriptor 1 which contains the peptide [HNQ]XG rather than G[HNQ] matches a much smaller percentage of potential sites in the synthetase set relative to the control set. Also, a descriptor containing the dipeptide G[HNQKRC] rather than G[HNQ] is much less specific. The latter is consistent with mutagenesis experiments which show that replacing His-48 with Asn or Gln yields active Tyr-tRNA synthetase, whereas replacement with Lys abolishes all measurable activity (Lowe et al., 1985). Thus, partial optimization resulted in our inclusion of the dipeptide G-[HNQ] in descriptors 1 and 2.

Table I lists the number of heterologous synthetase and control protein matches to the first, second, and composite descriptor. A MNBF descriptor should match on the order of only one site per synthetase, which is the usual binding stoichiometry for ATP and the aminoacyl adenylate (Schimmel & Soll, 1979). The first descriptor is quite specific. Of the 12 heterologous synthetases, 8 (67%) contain sites which match the descriptor. In contrast, only 4 of the 54 control



Enzyme	Source	Reference	βA turn G[B,N,Q] αB βB	Descriptor 1	Descriptor 1&2
Met	E. coli	Dardel, et al., 1984	5 KKILVTCALPYANGSIHLGHMLEHIQADVWVRYQRMRGHEVNFICADDAHG	+	+
	S. cerevisiae	Walter, et al., 1983	197 NILITSALPYVNNVPHLGNIIGSVLSADIFARYCKGRNYNA	+	-
Tyr[a]	E. coli	Barker, et al., 1982	31 PIALYCGFDPTADSLHLGHLVPLLCLKRFQQAGHKPVALVGGAT	+	●
	B. stearothermophilus	Winter & Bartley, 1983	30 RVTLYCGFDPTADSLHIGHLATILTMRRFQQAGHRPIALVGGAT	+	●
Ile	E. coli	Webster, et al., 1984	50 TFILHDGPPYANGSIHIGHSVNKILKDIIVKSKGLSGYDSPYVPGWDCHG	+	●
	S. cerevisiae[b]	Ludmerer, 1986	40 EFTFFDGPPFATGTPHYGHILASTIKDIVPRYATMTGHHVERRFGWDTHG	+	●
Gln	E. coli	Hoben, et al., 1982	26 TTVHTRFPPEPNGYLHIGHAKSICLNFGIANDYKGNCNLRFDD	-	-
	S. cerevisiae	Ludmerer, 1987	251 KVRTRFPPEPNGYLHIGHSKAIMVNFGYAKYHNGTCYLRFDD	+	-
Trp	B. stearothermophilus	Winter & Bartley, 1977	1 MKTIFSGIQPSGVITIGNYIGALRQFVELQHQYNCYFCIDQHAITVWQDPH	+	●
	E. coli	Hall, et al., 1982	4 PIVFSGAQPSGELTIGNYMGALRQWVKMQDDYHCIYCIVDQH	+	●
	S. cerevisiae	Meyers & Tzagoloff, 1985	35 ATVFSMIQPTGCFHLGNYLGATRVWTDLCELKQPGQELIFEV	-	-
Ala	E. coli	Putney, et al., 1981	152 RIIRINDNKGAPYASGNFWRMGGTGPCDPCTEIFYDHG	+	+
	E. coli		488 AVVVLDQTPFYAESGGQVGDKGELKGANFSFAVEDTQK	+	●
Gly	E. coli β-subunit	Webster, et al., 1983	300 NFIFVANIESKDPQQIISGNEKVVRPRLADAEFFFNTDR	+	●
Asp	S. cerevisiae	Sellami, et al., 1986	78 LPLIQSRDSDRTGQKRVKFVDLDEAKDSDKEVLFRARVHNTRQQGATLAFLTLRQQ	+	●
Glu	E. coli	Breton, et al., 1986	1 MKIKTRFAPSPTGYLHVGGARTALYSWLFARNHGGEFVLRI	-	-

FIGURE 5: Proposed alignment of aminoacyl-tRNA synthetase sequences with the first βαβ-fold of a mononucleotide bindinglike structure. The regions predicted to fold into βA, the turn, αB, and βB are enclosed in the first, second, third, and fourth *solid box*, respectively, of each sequence. Sequences which match the composite descriptor and/or descriptor 1 are indicated with a "+" to the right of the figure. Secondary structure assignments for the *B. stearothermophilus* Tyr-tRNA synthetase and *E. coli* Met-tRNA synthetase X-ray structures are indicated with an *underline*. The α-helix and β-strand assignments for the tyrosyl enzyme were taken from Table I of Blow et al. (1983), and those for the methionyl enzyme were from S. Brunie (personal communication). The turn assignments were taken from the stereoscopic drawing in Figure 1A of Blow et al. (1983). *Dashed boxes* indicate weakly predicted secondary structure elements, using the following changed PRSTRC parameters (see Materials and Methods): β former lowered to 1.20 for *E. coli* Gln-tRNA synthetase, α cutoff lowered to 0.94 for *S. cerevisiae* Trp-tRNA synthetase, and β former lowered to 1.17 for Glu-tRNA synthetase. The two β-strands enclosed in dashed boxes are predicted by the unmodified Chou and Fasman prediction scheme (a cluster of three β formers out of five residues is a predicted β-strand region). Footnote *a*, The alignment with *Bacillus caldotenax* Tyr-tRNA synthetase (Jones et al., 1986) is not included since it is 99% homologous with the enzyme from *B. stearothermophilus*. Footnote *b*, The incomplete sequence contains the first 140 N-terminal amino acids.

proteins (7.4%) match the descriptor. Only one synthetase, Ala-tRNA synthetase, contains two sites. The second descriptor overpredicts the descriptor structure since 58% of the control proteins match this descriptor and all synthetases contain multiple matching sites. The composite descriptor is the most specific. Of the 12 heterologous synthetases, 7 (58%) compared to only 2 of the 54 control proteins (3.7%) contain sites which match this descriptor. Although the synthetase primary sequences (average length 640 amino acids) tend to be longer than the proteins in this control set (average length 250 amino acids), the composite descriptor matches 10 sites per 10 000 synthetase amino acids compared to only 1.5 sites per 10 000 control protein amino acids.

Figure 5 lists all probable candidates for MNBF regions within the 12 analyzed heterologous aminoacyl-tRNA synthetases. These include the nine regions in eight synthetases (one region in Met-, Tyr-, Ile-, Gln-, Trp-, Gly-, and Asp-tRNA synthetases and two regions in the Ala-tRNA synthetase) which match the first descriptor. Homologous sequences from species variants are listed for each proposed MNBF region of the Met-, Tyr-, Ile-, Gln-, and Trp-tRNA synthetases. These sequences all match the first descriptor except those for the *E. coli* Gln-tRNA synthetase and *S. cerevisiae* Trp-tRNA synthetase proposed MNBF regions. These two sequences are each missing only a single predicted secondary structure element. Each of these elements is in fact weakly predicted (dashed boxes in Figure 5) but is outside of parameter ranges used in this study. No matches to the first descriptor occurred within any of the three Thr-tRNA synthetases (Mayaux et al., 1983; Pape et al., 1985; Pape & Tzagoloff, 1985), the two His-tRNA synthetases (Freedman et al., 1985; Natsoulis et al., 1986), the one Phe-tRNA synthetase (Mechulam et al., 1985), and the one Glu-tRNA synthetase (Breton et al., 1986). An alignment to the descriptor structure is proposed for Glu-tRNA synthetase (Figure 5) based on a homology to Gln-tRNA synthetase (Breton et al., 1986; Discussion).

DISCUSSION

The analysis of eight NAD and FAD binding enzymes demonstrates that, despite a lack of primary sequence similarity, it is possible to construct a single descriptor which correctly locates six of eight (75%) common nucleotide binding structures based only on information derived from the primary structure. Only one of the seven regions located by the descriptor is a false positive, and this is located in the single protein that has two matches to the descriptor. This analysis is also a test case for the indirect method of estimating the

Table I: Number of Occurrences of Positive Matches with Descriptors in the Set of Aminoacyl-tRNA Synthetases (synth) and the Set of Control Proteins (cnt)[a]

matches/ protein	Chou/Fasman descriptor 1		hydropathy descriptor 2		composite descriptor	
	synth	cnt	synth	cnt	synth	cnt
0	4	49	0	22	5	51
1	7	4	0	16	6	2
2	1	0	4	7	1	0
3	0	0	3	4	0	0
4	0	0	2	1	0	0
5	0	0	2	0	0	0
6	0	0	1	1	0	0
7	0	0	0	0	0	0
8	0	0	0	2	0	0
total	9/12	4/53	41/12	68/53	8/12	2/53

[a] The values listed under each descriptor are the number of proteins with the specified number of multiple matches. For example (see values in the first column), four proteins in the synth set have 0 matches to descriptor 1, seven proteins each have one match, and one protein has two matches to descriptor 1. The total (matches/protein) = (the sum of all matches in all proteins in the set)/(the total number of proteins in the set).

false positive background, which is by comparison to the percentage of matched control proteins. For the NAD/FAD binding enzymes, it validates the intuition that when the proportion of functionally related protein which are matched (75% in this case) is much greater than the proportion of matched control proteins (4.4%) the majority of matches (86% in this case) are true positives.

The synthetase regions (shown in Figure 5) which match the first descriptor are likely MNBF regions based both on statistics and by their agreement with available primary sequence and structural alignments. Descriptor 1 matches 67% of the heterologous synthetases and only 7.4% of the control proteins. The χ^2 statistic (74 with one degree of freedom) indicates that these differences are significant at $p < 0.001$. As an additional control on the descriptor specificity, the dipeptide G[HNQ] was reversed to [HNQ]G. This results in complete loss of discrimination between the synthetase set (2 heterologous groups matched/12 = 17%) and the control set (6 proteins matched/54 = 11%) with $p = 0.75$. Descriptor 1 uniquely matches the first $\beta\alpha\beta$-unit of the known MNBF regions of the Tyr- and Met-tRNA synthetases (underlined in Figure 5) and also the signature sequence regions of Gln-, Trp-, and Ile-tRNA synthetase. In all alignments, the signature sequence region matches loop 1 of the descriptor structure, as expected. Thus, this analysis supports the expectation that this sequence is diagnostic of MNBF regions, extends the alignment another approximately 40 amino acids, and identifies potential secondary structure elements.

Matches are found for three enzymes, the Gly-, Asp-, and Ala-tRNA synthetases, for which no previous evidence for a MNBF existed. Genetic studies with Ala-tRNA synthetase show that only the N-terminal 385 amino acids are required for aminoacyl adenylate synthesis (Jasin et al., 1983) and collaborate the match at amino acids 152–187. The second Ala-tRNA synthetase match is most likely a false positive; however, the ATP binding stoichiometry has not been determined for this enzyme (Schimmel & Soll, 1979).

The remaining three potential true positive matches do not contain a common residue in the position which aligns with the *B. stearothermophilus* Tyr-tRNA synthetase His-45 (Figure 5). Thus, they and the Trp-tRNA synthetase appear to be lacking a residue which has a known catalytic role (Leatherbarrow et al., 1985). In addition, the three sites contain a Ser or Thr in the following position, which is usually occupied by a bulky hydrophobic residue (Figure 1). Although the descriptor is quite specific; the false positive background is not zero, and some or all of these matches are possibly erroneous. However, it is also possible that these enzymes could contain a class or classes of sites, which are less similar to those of the Tyr- and Met-tRNA synthetases. If so, the roles of the above two residues may be substituted by other residues in the three-dimensional structures.

No regions in the Thr-, Phe-, Glu-, and His-tRNA synthetases match the first descriptor. MNBF structures could exist in these enzymes and not be identified because existing secondary structure elements are not predicted and/or the structures in these enzymes do not contain all the descriptor elements. Both are possibly the case for the *E. coli* Glu-tRNA synthetase. A potential MNBF site has been located on the basis of sequence similarity between it and *E. coli* Gln-tRNA synthetase (Brenton et al., 1986; Figure 5). However, in this site, the potential β-strand A is only weakly predicted (dashed box, Figure 5), and a Gly (Gly-19) is found in place of the His, Asn, or Gln element of the descriptor. It is possible that this enzyme tolerates the loss of an H-bond contact in this position, as does the genetically engineered Tyr-tRNA synthetase (Lowe et al., 1986).

We attempted to identify possible sites in the remaining Thr-, His-, and Phe-tRNA synthetase by constructing a much less specific version of descriptor 1. This was allowed to match weakly predicted secondary structure elements and contains only the Gly primary sequence element. Although this descriptor has a high false positive background, a site is proposed for the Thr-tRNA synthetase group because a match occurs at the same homologous site within the three enzyme forms. These matches are located at amino acids 250–295 in the *E. coli* enzyme, amino acids 336–381 in the *S. cerevisiae* cytoplasmic enzyme, and amino acids 47–99 in *S. cerevisiae* mitochondrial enzyme. We could not identify probable sites in Phe-tRNA synthetase and His-tRNA synthetase. Although it is recognized that this may be caused by general limitations in secondary structure predictions, it is also possible the sites in these enzymes may differ in some fundamental way from those in other synthetases.

This study has demonstrated the specificity of partially optimized descriptors for potential synthetase MNBF structures. These sites are presented as plausible models which can be subjected to experimental validation. Future work will be directed at automating all components of the method and thus increasing the ability to generate optimal descriptors.

Acknowledgments

We thank Dr. Paul Schimmel for his advice and helpful comments. We also thank Dr. Simone Brunie for providing us with unpublished structural information on *E. coli* Met-tRNA synthetase.

References

Barker, D. G., & Winter, G. (1982) *FEBS Lett. 145*, 191–193.

Barker, D. G., Ebel, J.-P., & Bruton, C. J. (1982) *Eur. J. Biochem. 127*, 449–457.

Bashford, D., Chothia, C., & Lesk, A. M. (1987) *J. Mol. Biol. 196*, 199–216.

Birkloft, J. J., & Banaszak, L. J. (1984) in *Peptide and Protein Reviews* (Hearn, M. T. W., Ed.) Vol. 4, pp 1–47, Marcel Dekker, New York.

Blow, D. M., & Brick, P. (1985) in *Biological Macromolecules and Assemblies 2: Nucleic Acids and Interactive Proteins*

(Jurnak, F. A., & McPherson, A., Eds.) pp 442–468, Wiley, New York.

Blow, D. M., Bhat, T. N., Metcalfe, A., Risler, J. L., Brunie, S., & Zewler, C. (1983) *J. Mol. Biol. 171*, 571–576.

Bradley, M. K., Smith, T. F., Lathrop, R. H., Livingston, D. M., & Webster, T. A. (1987) *Proc. Natl. Acad. Sci. U.S.A. 84*, 4026–4030.

Breton, R., Sanfacon, H., Papyannopoulous, I., Biemann, K., & Lapointe, J. (1986) *J. Biol. Chem. 261*, 10610–10617.

Brown, K. A., Vrielink, A., & Blow, D. M. (1986) *Trans. Biochem. Soc. 14*, 1228–1229.

Cassio, D., & Waller, J. P. (1971) *Eur. J. Biochem. 20*, 283–300.

Chothia, C. (1984) *Annu. Rev. Biochem. 53*, 537–572.

Chou, P. Y., & Fasman, G. D. (1978) *Annu. Rev. Biochem. 47*, 251–276.

Cohen, F. E., Abarbanel, R. M., Kuntz, I. D., & Fletterick, R. J. (1986) *Biochemistry 25*, 266–275.

Creighton, T. E. (1983) in *Proteins*: *Structure and Molecular Properties*, pp 252–265, W. H. Freeman, New York.

Dardel, F., Fayat, G., & Blanquet, S. (1984) *J. Bacteriol. 160*, 1115–1122.

Eisenberg, D., Weiss, R. M., & Terwilliger, T. C. (1982) *Nature* (*London*) *299*, 371–374.

Eisenberg, D., Weiss, R. M., & Terwilliger, T. C. (1984) *Proc. Natl. Acad. Sci. U.S.A. 81*, 140–144.

Freedman, R., Gibson, B., Donavan, D., Biemann, K., Eisenbeis, S., Parker, J., & Schimmel, P. (1985) *J. Biol. Chem. 260*, 10063–10068.

Fry, D. C., Kuby, S. A., & Mildvan, A. S. (1986) *Proc. Natl. Acad. Sci. U.S.A. 83*, 907–911.

Garnier, J., Osguthorpe, D. J., & Robson, B. (1978) *J. Mol. Biol. 120*, 97–120.

Gascuel, O., & Danchin, A. (1987) *J. Mol. Evol.* (in press).

George, D. G., Barker, W. C., & Hunt, L. T. (1986) *Nucleic Acids Res. 14*, 11–15.

Hall, C. V., van Cleemput, M., Muench, K. H., & Yanofsky, C. (1982) *J. Biol. Chem. 257*, 6132–6136.

Hoben, P., Royal, N., Cheung, A., Fumiaki, Y., Biemann, K., & Söll, D. (1982) *J. Biol. Chem. 257*, 11644–11650.

Hountondji, C., Blanquet, S., & Lederer, F. (1985) *Biochemistry 24*, 1175–1180.

Hountondji, C., Dessen, P., & Blanquet, S. (1986a) *Biochimie 68*, 1071–1078.

Hountondji, C., Lederer, F., Dessen, P., & Blanquet, S. (1986b) *Biochemistry 25*, 16–21.

Jasin, M., Regan, L., & Schimmel, P. R. (1983) *Nature* (*London*) *306*, 441–447.

Jones, M. D., Lowe, D. M., Borgford, T., & Fersht, A. R. (1986) *Biochemistry 25*, 1887–1891.

Kabsch, W., & Sander, C. (1983) *FEBS Lett. 155*, 179–182.

Kyte, J., & Doolittle, R. F. (1982) *J. Mol. Biol. 157*, 105–132.

laCour, T. F. M., Nyborg, J., Thirup, S., & Clark, B. F. C. (1985) *EMBO J. 4*, 2385–2388.

Lathrop, R. H., Webster, T. A., & Smith, T. F. (1987) *Commun. ACM* (in press).

Leatherbarrow, R. J., Fersht, A. R., & Winter, G. (1985) *Proc. Natl. Acad. Sci. U.S.A. 82*, 7840–7844.

Lowe, D., Fersht, A. R., & Wilkinson, A. J. (1985) *Biochemistry 24*, 5106–5109.

Ludmerer, S. (1986) Ph.D. Thesis, Massachusetts Institute of Technology.

Ludmerer, S., & Schimmel, P. (1987) *J. Biol. Chem. 262*, 10801–10806.

Mayaux, J.-F., Fayat, G., Fromant, M., Springer, M., Grunberg-Manago, M., & Blanquet, S. (1983) *Proc. Natl. Acad. Sci. U.S.A. 80*, 6152–6156.

Mechulam, Y., Fayat, G., & Blanquet, S. (1985) *J. Bacteriol. 163*, 787–791.

Myers, A. M., & Tzagoloff, A. (1985) *J. Biol. Chem. 260*, 15371–15377.

Natsoulis, G., Hilger, F., & Fink, G. R. (1986) *Cell* (*Cambridge, Mass.*) *46*, 235–243.

Pape, L. K., & Tzagoloff, A. (1985) *Nucleic Acids Res. 13*, 6171–6183.

Pape, L. K., Koerner, T. J., & Tzagoloff, A. (1985) *J. Biol. Chem. 260*, 15362–15369.

Putney, S. D., Royal, N. J., DeVegvar, H. N., Herlihy, W. C., Biemann, K., & Schimmel, P. (1981) *Science* (*Washington, D.C.*) *213*, 1497–1501.

Ralph, W. W., Webster, T. A., & Smith, T. F. (1987) *Comput. Appl. Biosci.* (in press).

Richardson, J. S. (1977) *Nature* (*London*) *268*, 495–500.

Richardson, J. S. (1981) *Adv. Protein Chem. 34*, 167–339.

Rossmann, M. G., Lipias, A., Branden, C. I., & Banaszak, L. J. (1975) *Enzymes*, *3rd Ed. 9*, 61–102.

Schimmel, P. R. (1987) *Annu. Rev. Biochem. 56*, 125–158.

Schimmel, P. R., & Söl, D. (1979) *Annu. Rev. Biochem. 48*, 601–648.

Schimmel, P., Jason, M., & Regan, L. (1984) *Fed. Proc., Fed. Am. Soc. Exp. Biol. 43*, 2987–2990.

Schultz, G. E., & Schirmer, R. H. (1979) in *Principles of Protein Structure*, Springer-Verlag, New York.

Sellami, M., Chatton, B., Fasiolo, F., Dirheimer, G., Ebel, J.-P., & Gangloff, J. (1986) *Nucleic Acids Res. 14*, 1657–1666.

Sternberg, M. J. E., & Thorton, J. M. (1976) *J. Mol. Biol. 105*, 367–382.

Taylor, W. R. (1986) *J. Mol. Biol. 188*, 233–258.

Walter, P., Gangloff, J., Bonnet, J., Boulanger, Y., Ebel, J.-P., & Fasiolo, F. (1983) *Proc. Natl. Acad. Sci. U.S.A. 80*, 2437–2441.

Waterman, M. S. (1984) *Bull. Math. Biol. 46*, 473–500.

Wayne, M. M. Y., Winter, G., Wilkinson, A. R., & Fersht, A. R. (1983) *EMBO J. 2*, 1827–1829.

Weber, P. G., & Salemme, F. R. (1980) *Nature* (*London*) *287*, 82–84.

Webster, T. A., Gibson, B. W., Keng, T., Biemann, K., & Schimmel, P. (1983) *J. Biol. Chem. 258*, 10637–10641.

Webster, T. A., Tsai, H., Kula, M., Mackie, G. A., & Schimmel, P. R. (1984) *Science* (*Washington, D.C.*) *226*, 1315–1317.

Wierenga, R. K., DeMaeyer, M. C. H., & Hol, W. G. J. (1985) *Biochemistry 24*, 1346–1357.

Wierenga, R. K., Terpstra, P., & Hol, W. G. L. (1986) *J. Mol. Biol. 187*, 101–107.

Winston, P. H. (1984) *Artificial Intelligence*, *2nd ed.*, pp 113–114, Addison-Wesley, Reading, MA.

Winter, G. P., & Hartley, B. S. (1977) *FEBS Lett. 80*, 340–342.

Winter, G., Koch, G. L. F., Hartley, B. S., & Barker, D. G. (1983) *Eur. J. Biochem. 132*, 383–387.

Zewler, C., Risler, J. L., & Brunie, S. (1982) *J. Mol. Biol. 155*, 63–81.