Eur. J. Biochem. *191*, 347–358 (1990)

Recognition of different nucleotide-binding sites in primary structures using a property-pattern approach

Peer BORK and Christian GRUNWALD

Department of Biomathematics, Central Institute of Molecular Biology, Academy of Sciences of German Democratic Republic, Berlin-Buch, German Democratic Republic

(Received March 2, 1990) – EJB 90 0326

Consensus sequence patterns for β-α-β folds binding FAD, NAD and GTP were constructed on the basis of 11 steric and physicochemical properties. These property patterns permit detection and distinction of the respective nucleotide-binding sites on the basis of amino acid sequence analysis alone. The SWISS-PROT database (release 9) was screened with the three calculated patterns, and nucleotide-binding sites identified are presented. They correspond to existing structure data (if known). For the detected sequence segments we are able to predict the β-α-β motif as well as the respective binding sites. For some of the proteins so detected a nucleotide-binding capacity has not previously been reported.

Many mononucleotide- and dinucleotide-binding proteins possess a common structural motif although the underlying primary structures vary greatly. Often two neighboring β strands and an antiparallel interconnecting α helix participate in the binding of the ribose moiety of different nucleotides [1]. The evolutionary history of this β-α-β motif is stil unclear [2]. Whereas for the NAD-binding domain the existence of a common ancestor NAD-binding protein is assumed [3], a convergent development is more likely in the case of ATP-binding sites [4]. Apart from this obvious evolutionary interest, a study of the sequence relationship is of practical importance as well. Nucleotide binding is involved in many metabolic reactions and in central regulatory mechanisms of the cell. Furthermore, the nucleotide-binding properties of several oncogene products have stimulated the interest in the relationship between primary structure and the β-α-β motif. Many papers have therefore dealt with recognition and prediction of nucleotide-binding sites. In most cases such studies have focussed on preservation of amino acid patterns in the nucleotide-ribose-binding loop [5–9] between βa and αb (for nomenclature see [1]). Wierenga and Hol [10] included hydrophobic properties of the interacting secondary structures into their predictions. This hydropathy pattern has been further refined [11–13]. But steric and other structural properties of amino acids may also contribute to the functional and structural features of the β-α-β motif.

The recognition and prediction of nucleotide-binding sites is a suitable application for our pattern-search algorithm based on steric and physicochemical properties [14] (for a review of pattern-search methods see [15]). Our program PAT [14] tested on many examples [16] calculates consensus patterns of steric and physicochemical properties from a master set of aligned sequence segments and subsequently screens protein-sequence databases. By including known interactions within the β-α-β folds and the functional requirements of nucleotide binding it has been possible to discriminate the different binding sites of FAD, NAD and GTP on the basis of information of the primary structure alone.

Correspondence to P. Bork, Academy of Sciences of German Democratic Republic, Central Institute of Molecular Biology, Department of Biomathematics, Robert-Rössle-Straße 10, Berlin, GDR 1115

Enzymes. Alcohol dehydrogenase (EC 1.1.1.1); glucose dehydrogenase (EC 1.1.1.47); glutathione reductase (EC 1.6.4.2); lipoamide dehydrogenase (EC 1.8.1.4); malate dehydrogenase (EC 1.1.1.37); mercuric reductase (EC 1.16.1.1); NADH dehydrogenase (EC 1.6.99.3); phosphoribosyl aminoimidazole carboxylase (EC 4.1.1.21); ribitol dehydrogenase (EC 1.1.1.56); tryptophan 2-monooxygenase (EC 1.13.12.3); UDP-glucose epimerase (EC 5.1.3.2).

MATERIALS AND METHODS

Each amino acid is represented by a vector of steric and physicochemical properties, either present or absent. The property set proposed by Taylor [17] was taken from Zvelebil et al. [18]. Because of some special features, glycine as well as proline have each been treated as a 'property'. The vector of each amino acid contains 11 properties (Fig. 1). There are also vectors standing for an undefined amino acid and for a gap. The program PAT analyzes the master set, position by position, and evaluates whether a considered property is present 'always' (1) or 'never' (0) or 'sometimes' (dot) (Fig. 2). This leads to vectors of 1's, 0's and dots characterizing the property pattern in each sequence position. These vectors are compatible with sets of amino acids (Fig. 2) according to their property vectors. It may seem to be a complicated way to calculate a pattern, but for predictive use it has the advantage, that similar amino acids which are not in the master set nevertheless belong to the property pattern.

Database screening

The basis of database search may be the different sets of amino acids (for this task many programs exist) or the property pattern itself (Fig. 2). Because of additional weighting possibilities (not used here) PAT directly compares the property pattern with all possible subsequences of all entries in the database. Any amino acid is called compatible with the pattern, if its property vector contains all properties that have to be 'always' present and no property that has to be 'never'

	ASR	Hyd	Pos	Neg	Pol	Chr	Sml	Tin	Ali	Aro	Pro	Gly
1	Ala A	1	0	0	0	0	1	1	0	0	0	0
2	Cys C	1	0	0	0	0	1	0	0	0	0	0
3	Asp D	0	0	1	1	1	1	0	0	0	0	0
4	Glu E	0	0	1	1	1	0	0	0	0	0	0
5	Phe F	1	0	0	0	0	0	0	0	1	0	0
6	Gly G	1	0	0	0	0	1	1	0	0	0	1
7	His H	1	1	0	1	1	0	0	0	1	0	0
8	Ile I	1	0	0	0	0	0	0	1	0	0	0
9	Lys K	1	1	0	1	1	0	0	0	0	0	0
10	Leu L	1	0	0	0	0	0	0	1	0	0	0
11	Met M	1	0	0	0	0	0	0	0	0	0	0
12	Asn N	0	0	0	1	0	1	0	0	0	0	0
13	Pro P	0	0	0	0	0	1	0	0	0	1	0
14	Gln Q	0	0	0	1	0	0	0	0	0	0	0
15	Arg R	0	1	0	1	1	0	0	0	0	0	0
16	Ser S	0	0	0	1	0	1	1	0	0	0	0
17	Thr T	1	0	0	1	0	1	0	0	0	0	0
18	Val V	1	0	0	0	0	1	0	1	0	0	0
19	Trp W	1	0	0	1	0	0	0	0	1	0	0
20	Tyr Y	1	0	0	1	0	0	0	0	1	0	0
21	Asx B	0	0	0	1	0	0	0	0	0	0	0
22	? X	1	1	1	1	1	1	1	1	1	1	1
23	Glx Z	0	0	0	1	0	0	0	0	0	0	0

Fig. 1. *The properties of amino acids used in this approach.* Hyd, hydrophobic; Pos, positive charge; Neg, negative charge; Pol, polar; Chr, charge; Ali, aliphathic; Sml, small; Tin, tiny; Aro, aromatic; Pro, proline; Gly, glycine. An amino acid either possesses the property (1) or not (0) (for details see [14, 18])

present. 'Dot' places in the pattern are considered to be neutral. If an amino acid is not compatible with the property pattern, a mismatch is recorded. The number of allowed mismatches can be specified. For details including the search for several motifs see [14].

Construction of property patterns for nucleotide-binding sites

Several general properties of the nucleotide-binding β-α-β folds were included in the calculation of all property patterns. (a) In the nucleotide-ribose-binding loop between the βa strand and the αb helix a pattern of tiny amino acids is typical (see the region of the two conserved glycines in Fig. 2), because more bulky side chains would interfere with the effect of the αb dipole moment upon the ligand [19, 20] and would fail to

```
structure            bbbb    aaaaaaaaaaaa      bbbb
remark                    ! !             ***
hydrophobic          .11111.1..11.....11...0.1...1.1.
pos.charge           .00000000000.000.0....0.0..0000.
neg.charge           .00000000000.000.00.0..000.0000.
polar                1.0000.0..0......0.1...1.....0.1
charge               .00000000000.000.0......0..00001
small                .....1111.1...........1.........
tiny                 00..0111..10.......0...0..0...00
aliphatic            01..1000..0........0..000.....10
aromatic             .0000000000.000..0.0..0.00.00000
proline              00000000.0000000000000.000.00.00
glycine              000001.1...000..00.00.0000000.00
corresponding
permissible          DIAAIGAGAAACAAAAAAADAADHAACAAAID
amino acids:         ELCCL G CCGICCCCCCCECCNKCCDCCCLE
                     HVIIV S GG LDIGFDIFKFDPNMIEIIGVK
                     KWLL    NI MGLIGELGNHESQTKFLLIWR
                     NYMM    PL TIMLIFMHQIF-R-LHMMLY
                     Q VV    SM VKNMLHVIRKG T MINTM
                     R       TN WLQNMI K LH - NKQVP
                     T        Q YMSQNK L MI   QLSWV
                              S  NTSQL M NK   RMTY
                              T  QVTSM T QL   SNV
                              V  RWVTN V RM   TPW
                              W  SYWVQ W SN   WQY
                              Y  T YWR Y TQ   YR
                                 V  YS  VR    T
                                 W  T   WS    V
                                 Y  V   YT    W
                                    W   V     Y
                                    Y   W
                                        Y
```

Fig. 2. *Property pattern for FAD-binding sites.* In the remark line the sign '!' means that at this position the amino acid is obligatory (mismatch forbidden) and '*' marks positions where deletions are allowed. The specified properties occur 'always' (1), 'never' (0) or 'sometimes' (.) in the respective positions. In the lower part of the figure all those amino acids have been inserted which are compatible with the pattern in the pertinent position. For example aspartate (D) is compatible at the first place of the pattern, because it is polar and neither tiny, nor aliphatic, nor glycine, nor proline. The dot places are not evaluated, because they stand for non-obligatory properties. The longer the 'tail' of permissible amino acids, the less the corresponding position is specified by the property vector. Amino acids compatible with the pattern may or may not have been present in the master set; pos, positive; neg., negative

meet steric requirements (see e.g. [21]). (b) The two interacting β strands are highly hydrophobic [22], because they are located in the core of a β sheet, or are otherwise shielded against the protein surface (positions in both β strands are therefore mostly hydrophobic in Fig. 2). (c) In order to interact with the two β strands, the α helix must have a hydrophobic moment [23] which implies a particular hydrophobicity pattern in the sequence [24] (see the α helix in Fig. 2). (d) The side chain of the last residue in the second β strand forms a stable bond with the nucleotide-ribose moiety [12] and therefore has to be polar (last position in Fig. 2). (e) The loop between the αb helix and the βb strand is turned away from the binding site, and its length can be expected to vary (in the pattern such positions are marked by '*'). The relationship outlined under (a) – (e) have been included and appear in corresponding positions in the consensus property patterns (Figs 2 – 4 and Appendix).

```
structure          bbbbb     aaaaaaaaaaaa          bbbbb
remark             *    ! *!              *********
hydrophob.         .1111.1..1.11..1...1.............1.1.0
pos.charged        .000000000.00..00.00...0..........0.000
neg.charged        00000000000000.00..00.0...........00001
polar              100.0.0..0.00......0.............0.0.1
charged            .000000000.00..00..0.............0.001
small              ......1.11..1.......................
tiny               .00.0.1.11..1.......................0
aliphatic          .1..1.0.00..0.........0.............0
aromatic           000.0.0.00.00..0..0..0000.........0000.0
proline            00000.00000000.00000000...........000000
glycine            000.001..1.0.0.00..00.0.0.........000.00
```

Fig. 3. *Property pattern of NAD-binding sites.* The symbols are the same as in Fig. 2. Most conserved are the glycines in the ADP-ribose binding loop (marked with '!'). The second loop without a role in nucleotide binding may vary to a much greater extent (many insertions may occur). Hydrophob., hydrophobic

Property pattern of FAD-binding sites

The pattern of FAD-binding sites is based on an alignment of five known FAD-binding sites using simularities *with* glutathione reductase, whose structure has been determined (see Table 2 in [6]). Some possible insertions in the region between the helix and the second β strand (βb) were taken into account. In a second step some closely related enzymes from other species were included. PAT calculated the consensus pattern in Fig. 2.

Property pattern of NAD-binding sites

The structure of several NAD-dependent dehydrogenases is known (see Protein Structure Data Bank [25]). Since the primary structures of closely related enzymes from other species is available, a good initial alignment of 30 binding sites could be established. An insertion was permitted (Fig. 3) within the well-known GXGXXG/A amino acid pattern of the nucleotide-ribose-binding loop (cf. alcohol dehydrogenases of yeast). NADP-binding sites do not always form a β-α-β fold. For instance, only one of the two NADP-binding sites of dihydrofolate reductase possesses a β-α-β-like structure. Because of this uncertainty, no separate pattern was constructed for NADP.

Property pattern of GTP-binding sites

The structure of some GTP binding sites is known [26 – 28]. The specific nucleotide-ribose-binding loop exhibits a G/

```
structure          bbbb      aaaaaaaaaaaaa
remark                    !!                          !            !!
hydrophob.         .11111....11.111..1.............  ..1.1101.1......  1101....
pos.charged        .00000..0.010000..0....0.0....0.  ..0..00000...0..  0001.0..
neg.charged        000000.0.00000000000.......0....  0.0.00.0000....0  0000....
polar              1.0.00....011.....0.............  ......1..01.....  0011.1..
charged            .00000....010000..0.............  ..0..0.000......  0001....
small              .....1....1011....0.............  ......11.1......  ..10....
tiny               00.001....10..00..0.0...........  ..0.000..10.....  ..00.0..
aliphatic          0...100.0.0000..................  ......0.000.0...  ..00.0..
aromatic           0.0.00.00.000000...000..00....00  ......0000.00.00  .000000.
proline            00000000000000000000000..0.0.00  00000000.0000000  0000000.
glycine            00.00....0100000.00000.......0..  .0000000.1000000  ..000000
```

Fig. 4. *Property pattern of GTP-binding sites.* The symbols are the same as in Fig. 2. The first sequence segment forms a β-α motif (the second β-strand does not necessarily lie in the segment), the second sequence segment is a magnesium-binding site, and the third a turn specific for guanidine fixation

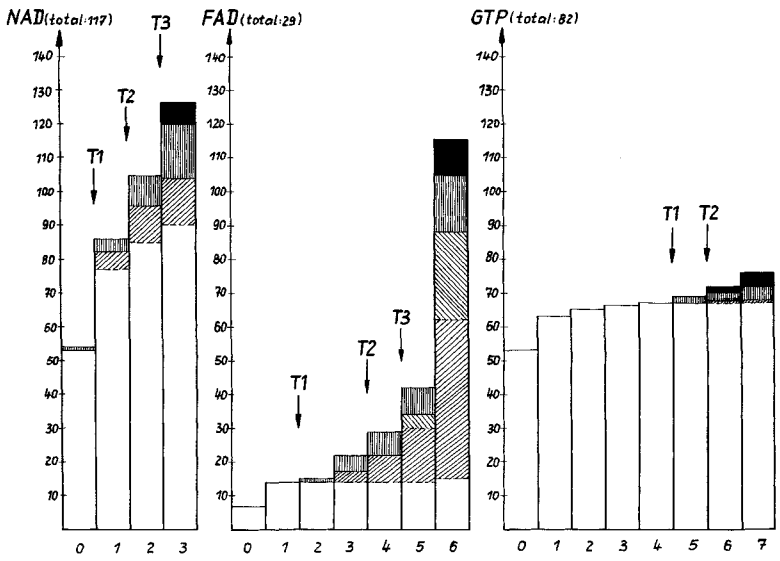

Fig. 5. *A cumulative representation of the number of detected sequence segments, depending on the number of permissible mismatches.* The 'total' number of respective binding sites denoted in the SWISS-PROT is also shown. □ Binding sites as looked for. ▨ Binding sites with similar function to those looked for (for example, a FAD-binding site detected by the NAD pattern or a ATP-binding site detected by the GTP-binding pattern). ▧ Nucleotide-binding β-α-β motifs in general. ▥ Proteins whose nucleotide-binding capacity has not been reported so far. A β-α-β motif is conceivable. ■ Sequence sections with different folding patterns, without nucleotide binding. The thresholds (T1 – T3) were chosen to allow searches with different degrees of relatedness to the sites looked for. The most stringent criterion (threshold 1, up to T1 mismatches) was found to exclude any type of binding site other than that looked for. Threshold 2 (up to T2 mismatches) was found to detect already similar functions (for instance dinucleotides) and the β-α-β fold. Threshold 3 (up to T3 mismatches) was found to tolerate additional β-α-β motifs performing binding of even less related nucleotides. For all thresholds no sequence section with known different folding pattern was found

AXXXXGK pattern [29]. Since this is not yet sufficient for a distinctive recognition, we included also the conserved magnesium-binding site and a turn region specific for guanine fixation [27 – 29] into the pattern calculation (Fig. 4). The master set containing 25 sequences was taken from Table 1 of [30].

Tubulins seem to have a different GTP-binding mode. The conserved guanine-ribose-binding loop contains the consensus pattern GGGXGXG [7]. Such binding sites have been excluded from the master set.

ATP-binding sites are considerably more flexible in their structure than GTP-binding sites. In numerous primary structures of ATP-binding proteins we could determine only one glycine as a common essential constituent of the ribose-binding loop. A universal pattern for distinctive recognition of ATP-binding sites has not been found. For instance, some protein superfamilies contain the so-called Walker motif [5], the protein kinases have a motif similar to dinucleotide-binding proteins [30 – 32], at least the cation-dependent ATPases show a different pattern in their nucleotide-ribose-binding loop [33], and within the tRNA-synthetase family this loop exists in many variants [9].

RESULTS

The three property patterns for FAD, NAD, GTP (Figs 2 – 4) were used to search matching sequences in the SWISS-PROT database ([30], release 9.0). Sequence segments showing a match were studied in detail to determine whether the segment was really an expected site as predicted by the pattern. Fig. 5 shows the statistics of predictions. Permitting some mismatches functionally and structurally, related regions were also recognized.

By defining optimal thresholds (Fig. 5) we may separate 'wanted' from 'unwanted' samples of segments. The quality of prediction may be assessed by the number of false positives (i.e. sites predicted for a pattern, but in fact not belonging to that class) and by the number of false negatives (i.e. sites not found by the pattern).

FAD-binding sites

SWISS-PROT keyword lines list 29 proteins as 'Fas-binding'; we found 14 of them with threshold T1 (0 or one mismatch), see Fig. 5. A less stringent search (up to five mismatches) did not identify additional FAD-binding proteins (Fig. 6) but recognized NAD-binding sites (see threshold T2 in Fig. 5) and other β-α-β motifs. In most of the FAD-binding proteins not detected FAD is covalently bound. They may have a different structure.

Binding sites for FAD and NAD(P) are well distinguished even when they coexist in one molecule. They are apparently closely related by evolutionary descent, as are those of e.g. glutathione reductase (GHSR$HUMAN, GHSR$ECOLI), lipoamide dehydrogenase (POD3$PSEPU, PYD3$ECOLI) and NADH dehydrogenase (DUNA$ECOLI) (see Fig. 6 and Appendix I).

NAD-binding sites

The database includes 117 NAD-binding proteins (excluding membrane-resident proteins of the electron-transport chains, being apparently unrelated to our master set of NAD-binding proteins). The pattern search detects 90 of them, together with a number of NADP-binding sites (with a β-α-β fold). With less than three mismatches only a small number

code	protein (broad class)	pos	mis	[aB]	[bA]	[bB]	nucl.
ALOX$HANPO =	alcohol oxidase	8	0	DIIVVGGGSTGCCIAGRLANLDDQNLTVALIE			FAD
DHNA$ECOLI =	NADH dehydrogenase	171	0	NIAIVGGGATGVELSAELHNAV-KQLHSYGYK			(FAD?)
MERA$NEUCR =	mercuric reductase	99	0	HIAVIGSGGAAMAAALKAVEQG---ARVTLIE			FAD
MERA$PSEAE =	:	100	0	QVAVIGSGGAAMAAALKAVEQG---AQVTLIE			FAD
MERA$SHIFL =	:	99	0	HIAVIGSGGAAMAAALKAVEQG---ARVTLIE			FAD
PHHY$PSEFL =	p-hydroxybenzoate hydroxylase	4	0	QVAIIGAGPSGLLLGQLLHKAG---IDNVILE			FAD
PYD3$ECOLI =	lipoamide dehydrogenase	7	0	QVVVLGAGPAGYSAAFRCADLG---LETVIVE			FAD
ADRO$BOVIN =	NADPH:adrenodoxin oxidoreductase	40	1	QICVVGSGPAGFYTAQHLLKHHSR-AHVdIYE			FAD
FRDA$ECOLI =	fumarate reductase	7	1	DLAIVGAGGAGLRAAIAAAQANPNAKIALIsK			FAD
GSHR$ECOLI =	gluthathione reductase	6	1	DYIAIGGGSGGIASINRAAmYG---QKCALIE			FAD
GSHR$HUMAN =	:	22	1	DYLVIGGGSGGLASArRAAELG---ARAAVVE			FAD
MERA$STAAU =	mercuric reductase	87	1	DLLIIGSGGAAFSAAIKAnENG---AKVAMVE			FAD
OXDA$PIG =	D-amino acid oxidase	2	1	RVVVIGAGVIGLSTALCIHERY-H-SVLqPLD			FAD
POD3$PSEPU =	lipoamide dehydrogenase	8	1	tLLIIGGGPGGYVAAIRAGQLG---IPTVLVE			FAD
TR2M$PSESY =	tryptophan 2-monooxygenase	40	2	RVAIVGAGiSGLVAATELLRAG--vKDVVLYE			? (1)
BEN4$PSEPU =	benzene degratation system	145	3	RLLIVGGGlIGCEVATTArKLG---LsVTILE			? (2)
DHLO$AGRT4 =	lysopine dehydrogenase	3	3	KVAILGAGNVALTLAGDLARRL--gQVSsIWa			NAD
DHSO$SHEEP =	sorbitol dehydrogenase	172	3	KVLVcGAGPIGLVNLLAAKaMG--AAQVVVtD			NAD
P49$STRLI =	P49 protein	3	3	DaVVVGAGPNGLTAAVELARRG---fPVAVfE			? (3)
TDH$ECOLI =	threonine 3-dehydrogenase	166	3	DVLVsGAGPIGIMAAAVAKhVG--ARNVVItD			NAD
YNI2$METTH =	hypothetical nif protein	2	3	KIVVVGGGTSGLLSALALeKeG---hDVLVLE			? (4)
Y21K$ECOLI =	hypothetical 21K protein	8	3	DVIIIGGGhAGtEAAMAAARMG---QQTLLLt			? (5)
GDHA$NEUCR =	glutamate dehydrogenase	221	4	RVALsGSGNVAqYAALKLIELG--ATVVsLsD			NADP
GSHR$HUMAN =	gluthathione reductase	189	4	RsVIVGAGyIAVEMAGILsaLG--SKTSLMIR			NAD(P)
LDH$BACST =	lactate dehydrogenase	8	4	RVVVIGAGfVGaSYVFALMNQG-iAdEIVLID			NAD
PNTA$ECOLI =	pyr. nucleotide transhydrogenase	166	4	KVMVIGAGVAGLAAIGAAnsLG---AIVrAfD			NAD
POD3$PSEPU =	lipoamide dehydrogenase	174	4	HLVVVGGGyIGLELGIAyrKLG---AQVsVVE			NAD
PUR6$ECOLI =	phos.rib.aminoimidazol carboxylase	64	4	QVIIaGAGGAAhLpGMIAAKTL---VPVLGVp			? (6)
TR2M$AGRT4 =	tryptophan 2-monooxygenase	238	4	KVAVIGAGiSGLVVANELLhAG--vdDVTIYE			? (7)

Fig. 6. *Screening result for the FAD pattern.* First column, SWISS-PROT codes; second column, position of the sequence segment; third column, number of mismatches, fourth column, alignment of detected sequence segments as one letter code (amino acids not compatible with the pattern are in lower case; fifth column, nucleotide bound (for those in parentheses binding is not certain). The β strands are denoted by [βa] or [βb], the αb helix by [αb]. Remarks. (1) and (7), tryptophan 2-monooxygenases were also detected with the NAD pattern (see Appendix I). (2) This enzyme is involved in benzene degradation, a dinucleotide as hydrogen acceptor is conceivable. (3), (4) and (5) Hypothetical proteins. (6) This phosphoribosyl aminoimidazole carboxylase is involved in purine synthesis

of false positives was found (Fig. 5 and Apendix I). They all correspond to a β-α-β motif.

GTP-binding sites

Apart from tubulins, a total of 82 proteins are labeled as 'GTP binding' in the database. By comparison with our property pattern we found 67 of them (Fig. 5, for details see Appendix II). Nearly all of the remaining 15 sequences belong to the protein family of the so-called 'negative factors'.

The pattern G/AXXXXGK expected for the guanine-ribose-binding loop [29] is confirmed by our results. Already the β-α-β motif excludes nearly all other functionally unrelated sequence segments. In this step of database search many ATP-binding β-α-β folds were also detected. The second and third motif of the pattern (Fig. 4) discriminated clearly the otherwise very similar binding sites for GTP from those for ATP. Other GTP-binding proteins, not yet contained in SWISS-PROT (release 9.0), do match the property pattern. In proteins without a GTP-binding property at least five mismatches were noted (thresholds T1 in Fig. 5).

DISCUSSION

Our strategy is based on the obligatory presence or absence of amino acid properties (Fig. 1) at a position defined by a set of aligned master specimens. This is a plausible and simple heuristic rule wich does not require complicated score calculations.

Our results document satisfactory identification and discrimination of nucleotide-binding sites. A stringent search (up to threshold T1 in Fig. 5) lists only sites looked for (with a risk of false negatives, i.e. overlooked sites). A more relaxed criterion eliminates nearly all false negatives (at the risk of false positives, i.e. the related NAD-binding sites appear during a search for the FAD motif), whereas the loosest criterion finds nucleotide-binding β-α-β folds in general, and excludes structures to a satisfactory degree (threshold T3 in Fig. 5).

Thus mononucleotide- and dinucleotide-binding sites could be clearly distinguished. ATP-binding sites are not found by a property-pattern search for GTP-binding motifs, showing the importance of the second and third motif. The FAD pattern does not recognize the closely related NAD-bindng sites, even when both occur in one protein. Strong position-dependent differences between the two sites were not found, but the FAD motif has a more stringent set of properties (only relatively few 'dots'; compare Figs 2 and 3). In the FAD-binding site no insertion was observed in the adenine-ribose-binding loop and at least five of the six amino acids in this region have to be tiny. Flavin, being more bulky than the nicotinamide of NAD, may restrict the variability of the FAD-binding site. More sequences are needed to verify this conclusion.

The adenine-ribose-binding loop between the βa strand and the αb helix also represents the most stringent criterion in the NAD-binding sites (Fig. 3). Tiny amino acids (t) were found at the only permissible insertion locus of the binding loop (GXtGXXG/A). In this case, nearly all the amino acids of the loop have to be tiny (GttGtXG/A). Ribitol dehydrogenase and glucose dehydrogenase contain a similar pattern in their potential nucleotide-ribose-binding loop: GttXGXG [34], as do the GTP-binding tubuline.

Only some NADP-binding sites were detected by the NAD pattern. We suspect that the additional phosphate group of NADP increases the flexibility of the motif. Thus, in the NADP-binding site of several mercuric reductases the second glycine in the adenine-ribose-binding loop, presupposed to be essential in NAD-binding sites (in Fig. 3 both glycines are marked by '!'), is replaced by histidine. The third important glycine of the NAD motif is replaced by alanine in NADP-binding sites [35, 36], but specific NAD-bindig sites may likewise have an alanine, as do certain FAD-binding motifs, like mercuric reductaes. Nevertheless, this alanine seems to be one of the key residues in the distinction between NAD- and NADP-binding sites [36].

Distinction of binding sites as described here is impossible with patterns in which only one amino acid appears in one position [37], even by allowing for conservative exchange. Obviously, it is the conservation of obligatory steric and physicochemical properties to which our method is tailored. The nucleotide-binding sites found with the property patterns are in accordance with structural data (where available).

Furthermore, we are able to predict nucleotide-binding sites in hitherto unstudied sequences with the mismatch threshold chosen (Fig. 5). We have found a number of binding sites that were not labeled in the database, i.e. a NAD-binding site on an enzyme involved in benzene degradation (BEN4$-PSEPU). UDP-glucose epimerase (GALX$SACCA), detected without deviation from the NAD pattern is not labeled, but indeed has NAD as a cofactor. We were further compelled to conclude that one dinucleotide-binding site is present in tryptophan 2-monooxygenases (TR2M$PSESY, TR2M$-AGRT4) and in the hypothetical protein 21K (Y21K$ECOLI), because they were detected by both dinucleotide patterns. For the hypothetical P49 protein (P49$STRLI), the hypothetical *nif* protein (YNI2$METTH), phosphoribosyl aminoimidazole carboxylase (PUR6$-ECOLI) and the pentafunctional aromatic polypeptide (ARO1$ASPNI) we can predict nucleotide-binding sites. (The exact positions and the respective sequence segments are aligned in Fig. 6 and in the Appendix.)

Our method is also useful for sequenced proteins known to bind nucleotide because we are able to predict the exact position of their β-α-β topology (if present). Therefore this determination of functional sites implies prediction of topological elements (in this case the β-α-β motif).

We conclude that NAD, FAD- and GTP-binding sites forming a β-α-β motif can be detected and classified by our method. We are present studying the same task for other functional sites or domains conserved in their topology.

We thank Prof. J. G. Reich for helpful discussions and critical reading of the manuscript. This work was supported by Prof. W. Pfeil and Dr W. Schöpp.

REFERENCES

1. Rossmann, M. G., Liljas, A., Bränden, C.-I. & Banaszak, L. I. (1975) in *The enzymes*, 3rd edn (Boyer, P., ed.) vol. 11, pp. 61–102, Academic Press, New York.
2. McLachlan, A. D. (1987) *Cold Spring Harbor Symp. Quant. Biol. 52*, 411–420.
3. Rossmann, M. G., Moras, D. & Olsen, K. W. (1974) *Nature (Lond.) 250*, 194–199.
4. Fothergill-Gilmore, L. A. (1986) in *Multidomain protein structure and evolution* (Hardie, D. G. & Coggins, J. R. eds) pp. 85–174, Elsevier Press.
5. Walker, J. E., Saraste, M., Runswick, W. J. & Gay, N. J. (1982) *EMBO J. 1*, 945–951.

6. Rice, D. W., Schulz, G. E. & Guest, J. R. (1984) *J. Mol. Biol. 174*, 483–496.
7. Möller, W. & Amons, R. (1985) *FEBS Lett. 186*, 1–7.
8. Fry, D. C., Kuby, S. A. & Mildvan, A. S. (1986) *Proc. Natl Acad. Sci. USA 83*, 907–911.
9. Webster, T. A., Lathrop, R. H. & Smith, T. F. (1987) *Biochemistry 26*, 6950–6957.
10. Wierenga, R. K. & Hol, W. G. J. (1983) *Nature (Lond.) 302*, 842–844.
11. Sternberg, M. J. E. & Taylor, W. R. (1984) *FEBS Lett. 175*, 387–392.
12. Wierenga, R. K., DeMayer, M. C. H. & Hol, W. G. J. (1985) *Biochemistry 24*, 1346–1357.
13. Wierenga, R. K. & Hol, W. G. J. (1986) *J. Mol. Biol. 187*, 101–107.
14. Bork, P. & Grunwald, C. (1989) *Stud. Biophys. 129*, 231–240.
15. Taylor, W. R. (1988) *Protein Eng. 2*, 77–86.
16. Bork, P. (1989) *FEBS Lett. 257*, 191–195.
17. Taylor, W. R. (1986) *J. Mol. Biol. 188*, 233–258.
18. Zvelebil, M. J., Barton, G. J., Taylor, W. R. & Sternberg, M. J. E. (1987) *J. Mol. Biol. 195*, 957–961.
19. Hol, W. G. J., Van Duijnen, P. T. & Berendsen, H. J. C. (1978) *Nature (Lond.) 273*, 443–446.
20. Hol, W. G. J. (1985) *Prog. Biophys. Mol. Biol. 45*, 149–195.
21. Edwards, M. S., Sternberg, M. J. E. & Thornton, J. M. (1987) *Protein Eng. 1*, 173–181.
22. Lim, V. I. (1974) *J. Mol. Biol. 88*, 857–872.
23. Sweet, R. M. & Eisenberg, D. (1983) *J. Mol. Biol. 171*, 479–488.
24. Palu, J. & Puigdomenech, P. (1974) *J. Mol. Biol. 88*, 475–469.
25. Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Brice, M. D., Rogers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977) *J. Mol. Biol. 112*, 535–542.
26. LaCour, T. F. M., Nyborg, J., Thirup, S. & Clark, B. F. C. (1985) *EMBO J. 4*, 2385–2388.
27. Jurnak, F. (1985) *Science 230*, 32–36.
28. Pai, E. F., Kabsch, W., Krengel, U., Holmes, K. C., John, J. & Wittinghofer, A. (1989) *Nature (Lond.) 341*, 209–214.
29. Dever, T. E., Glynias, M. J. & Merrick, W. C. (1987) *Proc. Natl Acad. Sci. USA 84*, 1814–1818.
30. Bairoch, A. & Claverie, J. M. (1988) *Nature (Lond.) 331*, 22.
31. Hanks, S. K., Quinn, A. M. & Hunter, T. (1988) *Science 241*, 42–52.
32. Leader, D. P. (1988) *Nature (Lond.) 333*, 308.
33. Taylor, W. R. & Green, N. M. (1989) *Eur. J. Biochem. 179*, 241–248.
34. Jörnvall, H., van Bahr-Lindström, H., Jung, K. V., Ulmer, W. & Fröschl, M. (1984) *FEBS Lett. 165*, 190–196.
35. Hanukoglu, I. & Gutfinger, T. (1989) *Eur. J. Biochem. 180*, 479–484.
36. Scrutton, N. S., Berry, A. & Perham, R. N. (1990) *Nature (Lond.) 343*, 38–42.
37. Argos, P. & Lebermann, R. (1985) *Eur. J. Biochem. 152*, 651–656.
38. Joh, T., Takeshima, H., Tsuzuki, T., Shimada, K., Tanase, S. & Morino, Y. (1987) *Biochemistry 26*, 2515–2520.

Appendix I. *A sample of sequence segments, detected by a search for NAD-binding sites*

The symbols are the same as in Fig. 6. All sequence segments, detected with up to two mismatches are shown. Dinucleotide-binding motifs containing three mismatches are also shown. Other proteins that do not have nucleotide-binding sites were also found with three mismatches. (1) Enzyme related to benzene degradation. The existence of a NAD-binding site is probable. (2) Cytosolic malate dehydrogenase of pig [38] was included to support the hypothesis of the variable loop 2 length. (3) Threonine dehydrates may have a different fold. (4) and (5) Tryptophan 2-monooxygenase were also found with the FAD pattern (Fig. 4). A nucleotide-binding site is probable, because the pathway is cofactor dependent. (7) Contact site A protein precursor is unlikely to have a nucleotide-binding site as it is localised on the surface of cell membranes. (6), (8), (9) and (10) Pentafunctional aromatic polypeptide (it has at least one ATP-binding site), α-amylase inhibitor, sporulation protein and hypothetical protein 21K may indeed have a hitherto undescribed nucleotide-binding site

code	protein (broad class)	pos	mis	[aB]	[bA]	[bB]	nucl
ADH$ARATH	= alcohol dehydrogenase	197	0	S-VAIFGL-GAVGLGAAEGARIAGA	--------SRIIGVD		NAD
ADH$DROMA	= :	8	0	NVIFVAGL-GGIGLDTSKELVKRDL	--------KNLVILD		NAD
ADH$DROME	= :	8	0	NVIFVAGL-GGIGLDTSKELLKRDL	--------KNLVILD		NAD
ADH$DROOR	= :	8	0	NVIFVAGL-GGIGLDTSKELVKRDL	--------KNLVILD		NAD
ADH$DROPS	= :	6	0	NVVFVAGL-GGIGLDTSRELVKRNL	--------KNLVILD		NAD
ADH$DROSI	= :	8	0	NVIFVAGL-GGIGLDTSKELLKRDL	--------KNLVILD		NAD
ADH$SCHPO	= :	175	0	W-ICIPGAGGGLGHLAVQYAKAMAM	---------RVVAID		NAD
ADH1$DROMU	= :	6	0	NIIFVAGL-GGIGFDTSREIVKSGP	--------KNLVILD		NAD
ADH1$HORVU	= :	197	0	T-VAIFGL-GAVGLAAAEGARIAGA	--------SRIIGVD		NAD
ADH1$MAIZE	= :	197	0	T-VAVFGL-GAVGLAAAEGARIAGA	--------SRIIGVD		NAD
ADH1$YEAST	= :	172	0	W-VAISGAAGGLGSLAVQYAKAMGY	---------RVLGID		NAD
ADH2$DROMU	= :	6	0	NIIFVAGL-GGIGFDTSREIVKSGP	--------KNLVILD		NAD
ADH2$MAIZE	= :	197	0	T-VAIFGL-GAVGLAAMEGARLAGA	--------SRIIGVD		NAD
ADH3$YEAST	= :	200	0	W-VAISGAAGGLGSLAVQYATAMGY	---------RVLGID		NAD

ADHA$HUMAN =	:	194	0	T-CAVFGL-GGVGLSAIMGCKAAGA--------ARIIAVD	NAD
ADHA$MOUSE =	:	194	0	T-CAVFGL-GGVGLSVIIGCKAAGA--------ARIIAVD	NAD
ADHB$HUMAN =	:	194	0	T-CAVFGL-GGVGLSAVMGCKAAGA--------ARIIAVD	NAD
ADHE$HORSE =	:	194	0	T-CAVFGL-GGVGLSVIMGCKAAGA--------ARIIGVD	NAD
ADHG$HUMAN =	:	194	0	T-CAVFGL-GGVGLSVVMGCKAAGA--------ARIIAVD	NAD
ADHP$HUMAN =	:	200	0	T-CAVFGL-GGVGLSAVMGCKAAGA--------SRIIGID	NAD
ADHS$HORSE =	:	194	0	T-CAVFGL-GGVGLSVIMGCKAAGA--------ARIIGVD	NAD
BEN4$PSEPU = benzene degradation system		145	0	R-LLIVGG-GLIGCEVATTARKLGL---------SVTILE	? (1)
ECH$RAT = enoyl-CoA hydrolase		297	0	S-VGVLGL-GTMGRGIAISFARVGI---------SVVAVE	NAD
G3P$BACST = glyceraldehyde-3-P dehydrogenase		3	0	K-VGINGF-GRIGRNVFRAALKNPD-------IEVVAVND	NAD
G3P$CAEEL =	:	5	0	N-VGINGF-GRIGRLVLRAAVEKDT-------VQVVAVND	NAD
G3P$CHICK =	:	2	0	K-VGVNGF-GRIGRLVTRAAVLSGK-------VQVVAIND	NAD
G3P$ECOLI =	:	4	0	K-VGINGF-GRIGRIVFRAAQKRSD-------IEIVAIND	NAD
G3P$RAT =	:	2	0	K-VGVNGF-GRIGRLVTRAAFSCDK-------VDIVAIND	NAD
G3P$THEAQ =	:	2	0	K-VGINGF-GRIGRQVFRILHSRGV--------EVALIND	NAD
G3P1$DROME =	:	3	0	K-IGINGF-GRIGRLVLRAAIDKGA--------SVVAVND	NAD
G3P2$DROME =	:	3	0	K-IGINGF-GRIGRLVLRAAIDKGA--------NVVAVND	NAD
G3P2$HUMAN =	:	4	0	K-VGVNGF-GRIGRLVTRAAFNSGK-------VDIVAIND	NAD
G3P2$YEAST =	:	2	0	R-VAINGF-GRIGRLVMRIALSRPN-------VEVVALND	NAD
G3PC$MAIZE =	:	5	0	K-IGINGF-GRIGRLVARVALQSED-------VELVAVND	NAD
G3PC$SINAL =	:	6	0	K-IGINGF-GRIGRLVARVILQRND-------VELVAVND	NAD
GALX$SACCA = UDP-glucose epimerase		12	0	KIVLVTGGAGYIGSHTVVELIENGY---------DCVVAD	(NAD?)
LDH$BACME = lactate dehydrogenase		11	0	K-VAVIGT-GFVGSSYAFSMVNQGI-------ANELVLID	NAD
LDH$BACST =	:	8	0	R-VVVIGA-GFVGASYVFALMNQGI-------ADEIVLID	NAD
LDH$THECA =	:	2	0	K-VGIVGS-GMVGSATAYALALLGV-------AREVVLVD	NAD
LDHH$CHICK =	:	21	0	K-ITVVGV-GQVGMACAISILGKGL-------CDELALVD	NAD
LDHH$HUMAN =	:	22	0	K-ITVVGV-GQVGMACAISILGKSL-------ADELALVD	NAD
LDHH$PIG =	:	22	0	K-ITVVGV-GQVGMACAISILGKSL-------TDELALVD	NAD
LDHM$HUMAN =	:	21	0	K-ITVVGV-GAVGMACAISILMKDL-------ADELALVD	NAD
LDHM$MOUSE =	:	21	0	K-ITVVGV-GAVGMACAISILMKDL-------ADELALVD	NAD
LDHM$PIG =	:	21	0	K-ITVVGV-GAVGMACAISILMKEL-------ADEIALVD	NAD
LDHM$RAT =	:	22	0	K-ITVVGV-GAVGMACAISILMKDL-------ADELALVD	NAD
LDHM$SQUAC =	:	22	0	K-ITVVGV-GAVGMACAISILMKDL-------ADEVALVD	NAD
LDHX$HUMAN =	:	21	0	K-ITIVGT-GAVGMACAISILLKDL-------ADELALVD	NAD
LDHX$MOUSE =	:	21	0	K-ITVVGV-GNVGMACAISILLKGL-------ADELALVD	NAD
MDHM$MOUSE = malate dehydrogenase		26	0	K-VAVLGASGGIGQPLSLLLKNSPL-------VSRLTLYD	NAD
MDHM$PIG =	:	2	0	K-VAVLGASGGIGQPLSLLLKNSPL-------VSRLTLYD	NAD
MDHM$RAT =	:	26	0	K-VAVLGASGGIGQPLSLLLKNSPL-------VSRLTLYD	NAD
PNTA$ECOLI = pyridine transhydrogenase		166	0	K-VMVIGA-GVAGLAAIGAANSLGA---------IVRAFD	NAD(P)
ADH1$ASPNI = alcohol dehydrogenase		173	1	T-VAIVGAGGGLGSLAqQYAKAMGI---------RVVAVD	NAD
ADH2$YEAST =	:	173	1	W-aAISGAAGGLGSLAVQYAKAMGY---------RVLGID	NAD
ADH3$ASPNI =	:	175	1	T-VAIVGAGGGLGSLAqQYAKAMGL---------RTIAID	NAD
DAPB$ECOLI = dihydrodipicolinare reductase		7	1	R-VAIAGAGGRMGRQLIQAALALEg-------VQLGAALE	NAD(P)
DHNO$AGRT7 = nopaline dehydrogenase		23	1	T-VGVLGS-GHAGTALAAWFASRhV------PTALWAPAD	NAD(P)
DHSO$SHEEP = sorbitol dehydrogenase		172	1	K-VLVCGA-GPIGLVnLLAAKAMGA--------AQVVVTD	NAD

Appendix I.

G3P$HOMAM	= glyceraldehyde-3-P dehydrogenase	2	1	K-IGIdGF-GRIGRLVLRAALSCGA--------QVVAVND	NAD
G3P$PIG	= :	2	1	K-VGVdGF-GRIGRLVTRAAFNSGK--------VDIVAIND	NAD
G3P$ZYGRO	= :	3	1	N-VsVNGF-GRIGRLVTRIAISRKD--------INLVAIND	NAD
G3P1$HUMAN	= :	4	1	K-VGVdGF-GRIGRLVTRAAFNSGK--------VDIVAIND	NAD
G3P3$YEAST	= :	2	1	R-IAINGF-GRIGRLVLRLALQRKD--------IEVVAVbD	NAD
G3PA$TOBAC	= :	59	1	K-VAINGF-GRIGRNFLRCWHGRKD-----SpLDVIAIND	NAD
G3PB$TOBAC	= :	56	1	K-VAINGF-GRIGRNFLRCWHGRKD-----SpLDVVVVND	NAD
GPDA$DROVI	= glycerol-3-P dehydrogenase	5	1	N-VCIVGS-GNnGSAIAKIVGANAAALPEFEERVTMFVYE	NAD
GPDA$RABIT	= :	4	1	K-VCIVGS-GdWGSAIAKIVGGNAAQLAQFDPRVTMWVFE	NAD
HCDH$PIG	= hydroxyacyl acid dehydrogenase	17	1	h-VTVIGG-GLMGAGIAQVAAATGH---------TVVLVD	NAD
LDH$LACCA	= lactate dehydrogenase	10	1	K-VILVGD-GAVGSSYAFAMVLQGI--------AQEIglVD	NAD
LDHM$CHICK	= :	21	1	K-IsVVGV-GAVGMACAISILMKDL--------ADELTLVD	NAD
MDH$ECOLI	= malate dehydrogenase	2	1	K-VAVLGAAGGIGQALALLLKTQlP------SGSELSLYD	NAD
(2)	= :	5	1	R-VLVTGAAGQIAYSLLYSIGNgSV--FGKDQPIILVLLD	NAD
MERA$PSEAE	= mercuric dehydrogenase	100	1	Q-VAVIGS-GGAAMAAALkAVEQGA---------QVTLIE	FAD
:	= :	252	1	R-LAVIGS-GYIAAELGQMFHNLGT---------EVTLMq	NADP
OXDA$PIG	= D-amino acid oxidase	2	1	R-VVVIGA-GVIGLSTALCIHERyH--------SVLQPLD	FAD
POD3$PSEPU	= lipoamide dehydrogenase	8	1	T-LLIIGG-GPGGYVAAIrAGQLGI---------PTVLVE	FAD
:	= :	174	1	h-LVVVGG-GYIGLELGIAYRKLGA---------QVSVVE	NAD
PYD3$ECOLI	= :	7	1	Q-VVVLGA-GPAGYSAAFrCADLGL---------ETVIVE	FAD
:	= :	176	1	R-LLVMGG-GIIGLEMGTVYHALGS---------QIdVVE	NAD
SERA$ECOLI	= phosphoglycerate dehydrogenase	153	1	K-LGIIGY-GHIGTQLGILAeSLGM---------YVYFYD	NAD
TDH$ECOLI	= threonine dehydrogenase	166	1	d-VLVSGA-GPIGIMAAAVAKHVGA--------RNVVITD	NAD
THDH$ECOLI	= threonine dehydratase	182	1	RVFVpVGG-GGLAAGVAVLIKQLMP--------QIKVIAVE	? (3)
TR2M$AGRT4	= tryptophane 2-monooxygenase	238	1	K-VAVIGA-GIsGLVVANELLHAGV--------DDVTIYE	? (4)
TR2M$PSESY	= :	40	1	R-VAIVGA-GIsGLVAATELLRAGV--------KDVVLYE	? (5)
ARO1$ASPNI	= aromatic polypeptide	1418	2	S-aLVVGG-GGTARAAIYALHNMGY--------SPIYIVgE	? (6)
CSA$DICDI	= contact site A protein	36	2	Y-ITITGT-GFTGTPVVTIgGQTCDP--------VIVANt	? (7)
DDH$CORGL	= diaminopimelate dehydrogenase	5	2	R-VAIVGY-GNLGRSVeKLIAKQPD--------MDLVGIf	NADP
DHGL$BOVIN	= glutamate dehydrogenase	246	2	T-FAVQGF-GNVGLHsMRYLHRFGA--------KCVAVgE	NAD(P)
DHGL$CHICK	= :	249	2	T-FAVQGF-GNVGLHsMRYLHRFGA--------KCVAVgE	NAD(P)
DHGL$HUMAN	= :	303	2	T-FVVQGF-GNVGLHsMRYLHRFGA--------KCIAVgE	NAD(P)
DHNA$ECOLI	= NADH dehydrogenase	170	2	N-IAIVGG-GATGVELSAELHNAvK--QLHSYGYKGLtNE	(NAD?)
DHOM$CORGL	= homoserine dehydrogenase	19	2	g-IALLGF-GTVGTEVMRLMTEYGD----ELAHRIGgPLE	NAD
DHSA$BACSU	= succinate dehydrogenase	5	2	S-IIVVGG-GLAGLMATIkAAESGM---------AVKLFs	NADP
FRDA$ECOLI	= fumarate reductase	7	2	d-LAIVGA-GGAGLRAAIAAAQANP--------NAKIALIs	FAD
GDHA$NEUCR	= glutamate dehydrogenase	221	2	R-VALSGS-GNVAQYAALkLIELGA--------TVVSLsD	NADP
GSHR$ECOLI	= glutathione reductase	169	2	R-VAVVGA-GYIAVELAGVINGLGA--------KThLFVr	NAD(P)
IAA$STRGS	= alpha-amylase inhibitor	47	2	dILTFPGY-GTrGNEVLGAVLCATD---------GSALPVD	? (8)
MERA$NEUCR	= mercuric reductase	99	2	h-IAVIGS-GGAAMAAALkAVEQGA---------RVTLIE	FAD
MERA$SHIFL	= :	99	2	h-IAVIGS-GGAAMAAALkAVEQGA---------RVTLIE	FAD
MERA$STAAU	= :	87	2	d-LLIIGS-GGAAFSAAIkANENGA---------KVAMVE	FAD
PHHY$PSEFL	= p-hydroxybenzoate hydroxylase	4	2	Q-VAIIGA-GPsGLLLGQLLHKAGI---------DnVILE	FAD
PROC$ECOLI	= pyrroline-carboxylate reductase	4	2	K-IGFIGC-GNMGKAILGGLIASGQ--------VLPgQIWv	NAD(P)
SP5F$BACSU	= sporulation protein	117	2	f-LGLIGLSGFVGLVLSApYRIKRI--------TSYLNPWE	? (9)

Appendix I.

```
TYRA$ECOLI = prephenate dehydrogenase          99   2   RpVVIVGGGGQMGRLFeKMLTLSGY---------QVRILE     NAD
Y21K$ECOLI = hypothetical protein               8   2   d-VIIIGG-GHAGTEAAMAAARMGQ--------QTLLLTh     ? (10)
AK1H$ECOLI = homoserine dehydrogenase         466   3   e-VFVIGV-GGVGgALLEQLKRQQS-------WLKnGHID     NAD
AK2H$ECOLI =        :                         458   3   g-LVLFGk-GNIGSRWLELFAREQS------TLSArTGFE     NAD
DHNA$ECOLI = NADH dehydrogenase                 5   3   K-IVIVGG-GAGGLEMATQLGHKlg----RKKKAKITLVD     (FAD?)
GDHA$ECOLI = glutamate dehydrogenase          234   3   R-VsVSGS-GNVAQYAIEkAMEFGA--------RVITAsD     NADP
GDHA$YEAST =        :                         219   3   R-VTISGS-GNVAQYAALkVIELGg--------TVVSLsD     NADP
GSHR$HUMAN = glutathione reductase            189   3   R-sVIVGA-GYIAVEMAGILSALGS--------KTsLMIr     NADP
GUAB$ECOLI = IMP dehydrogenase                316   3   S-aVkVGI-GPGsICTTRIVTGVGVP-------QITAVAD     NAD
LDH$STRCR  = lactate dehydrogenase              7   3   K-VILVGD-GAVGSAYAILddEHAV--------LPVSVFq     NAD
```

Appendix II. *List of identified GTP-binding sites containing the three motifs: guanine-ribose-binding fold, magnesium-binding site and a turn specific for guanine fixation.*

The symbols are the same as in Fig. 6 or Appendix I. For the first motif four mismatches were allowed, for both other motifs only two. The program system PAT lists only proteins with sequence sections that match all three motifs. In the cases marked by '?' the existence of a mononucleotide-binding site is not known to the authors

codes	proteins (broad class)	guanine ribose binding motif pos. m [beta] [alpha]	Mg-binding site pos. m	guanine fixation pos. m [turn]	nucl.
EF1A$ARTSA	= elongation factor	8 0 NIVVIGHVDSGKSTTTGHLIYKCGGIDKRTIE	84 0 YYVTIIDAPGHRDFIK	150 0 GVNKMDST	GTP
EF1A$DROME	= :	9 0 NIVVIGHVDSGKSTTTGHLIYKCGGIDKRTIE	85 0 YYVTIIDAPGHRDFIK	151 0 GVNKMDSS	GTP
EF1A$HUMAN	= :	9 0 NIVVIGHVDSGKSTTTGHLIYKCGGIDKRTIE	85 0 YYVTIIDAPGHRDFIK	151 0 GVNKMDST	GTP
EF1A$RHIRA	= :	9 0 NVVVIGHVDSGKSTTTGHLIYKCGGIDKRTIE	85 0 YNVTVIDAPGHRDFIK	151 0 AINKMDTT	GTP
EF1A$YEAST	= :	9 0 NVVVIGHVDSGKSTTTGHLIYKCGGIDKRTIE	85 0 YQVTVIDAPGHRDFIK	151 0 AVNKMDSV	GTP
EF1B$DROME	= :	9 0 NIVVIGHVDSGKSTTTGHLIYKCGGIDKRTIE	85 0 YYVTIIDAPGHRDFIK	151 0 GVNKMDST	GTP
EFTU$ECOLI	= :	14 0 NVGTIGHVDHGKTTLTAAITTVLAKTYGGAAR	75 0 RHYAHVDCPGHADYVK	134 0 FLNKCDMV	GTP
EFTU$EUGGR	= :	14 0 NIGTIGHVDHGKTTLTAAITMALAATGNSKAK	75 0 RHYAHVDCPGHADYVK	134 0 FLNKEDQV	GTP
EFTU$METVA	= :	9 0 NVAFIGHVDAGKSTTVGRLLLDGGAIDPQLIV	85 0 YEVTIVDCPGHRDFIK	147 0 AVNKMDTV	GTP
EFTU$THETH	= :	14 0 NVGTIGHVDHGKTTLTAALTYVAAAENPNVEV	76 0 RHYSHVDCPGHADYIK	135 0 FMNKVDMV	GTP
GBA1$YEAST	= G-protein	43 0 KLLLLGAGESGKSTVLKQLKLLHQGGFSHQER	313 0 SKFKVLDAGGQRSERK	386 0 FLNKIDLF	GTP
GBAO$BOVIN	= :	35 0 KLLLLGAGESGKSTIVKQMKIIHEDGFSGEDV	195 0 LHFRLFDVGGQRSERK	268 0 FLNKKDLF	GTP
GBAS$BOVIN	= :	42 0 RLLLLGAGESGKSTIVKQMRILHVNGFNGEGG	217 0 VNFHMFDVGGQRDERR	290 0 FLNKQDLL	GTP
GBAS$HUMAN	= :	42 0 RLLLLGAGESGKSTIVKQMRILHVNGFNGEGG	217 0 VNFHMFDVGGQRDERR	290 0 FLNKQDLL	GTP
GBAS$MOUSE	= :	40 0 RLLLLGAGESGKSTIVKQMRILHVNGFNGDEK	200 0 VNFHMFDVGGQRDERR	273 0 FLNKQDLL	GTP
GBAS$RAT	= :	42 0 RLLLLGAGESGKSTIVKQMRILHVNGFNGEGG	217 0 VNFHMFDVGGQRDERR	290 0 FLNKQDLL	GTP
GBI1$BOVIN	= :	35 0 KLLLLGAGESGKSTIVKQMKIIHEAGYSEEEC	194 0 LHFKMFDVGGQRSERK	267 0 FLNKKDLF	GTP
GBI2$HUMAN	= :	35 0 KLLLLGAGESGKSTIVKQMKIIHEDGYSEEEC	195 0 LHFKMFDVGGQRSERK	268 0 FLNKKDLF	GTP
GBI2$MOUSE	= :	35 0 KLLLLGAGESGKSTIVKQMKIIHEDGYSEEEC	195 0 LHFKMFDVGGQRSERK	268 0 FLNKKDLF	GTP
GBI2$RAT	= :	35 0 KLLLLGAGESGKSTIVKQMKIIHEDGYSEEEC	195 0 LHFKMFDVGGQRSERK	268 0 FLNKKDLF	GTP
GBI3$HUMAN	= :	35 0 KLLLLGAGESGKSTIVKQMKIIHEDGYSEDEC	194 0 LYFKMFDVGGQRSERK	267 0 FLNKKDLF	GTP
GBI3$RAT	= :	35 0 KLLLLGAGESGKSTIVKQMKIIHEDGYSEDEC	194 0 LYFKMFDVGGQRSERK	267 0 FLNKKDLF	GTP
GBT1$BOVIN	= :	31 0 KLLLLGAGESGKSTIVKQMKIIHQDGYSLEEC	190 0 LNFRMFDVGGQRSERK	263 0 FLNKKDVF	GTP
RAB1$RAT	= ras related oncogene products	13 0 KLLLIGDSGVGKSCLLLRFADDTYTESYISTI	60 0 IKLQIWDTAGQERFRT	122 0 VGNKCDLI	GTP

```
RAB2$HUMAN =       :                    8 0 KYIIIGDTGVGKSCLLLQFTDKRFQPVHDLTI   55 0 IKLQIWDTAGQESFRS  117 0 IGNKSDLE   GTP
RAB2$RAT   =       :                    8 0 KYIIIGDTGVGKSCLLLQFTDKRFQPVHDLTM   55 0 IKLQIWDTAGQESFRS  117 0 IGNKSDLE   GTP
RAB4$RAT   =       :                   10 0 KFLVIGNAGTGKSCLLHQFIEKKFKDDSNHTI   57 0 VKLQIWDTAGQERFRS  119 0 CGNKKDLD   GTP
RAL$CALJA  =       :                   16 0 KVIMVGSGGVGKSALTLQFMYDEFVEDYEPTK   62 0 VQIDILDTAGQEDYAA  125 0 VGNKSDLE   GTP
RAS$CARAU  =       :                    5 0 KLVVVGAGGVGKSALTIQLIQNHFVDEYDPTI   51 0 CLLDILDTAGQEEYSA  114 0 VGNKCDLP   GTP
RAS$DICDI  =       :                    5 0 KLVIVGGGGVGKSALTIQLIQNHFIDEYDPTI   51 0 CLLDILDTAGQEEYSA  114 0 VGNKADLD   GTP
RAS$SCHPO  =       :                   10 0 KLVVVGDGGVGKSALTIQLIQSHFVDEYDPTI   56 0 AVLDLLDTAGQEEYSA  119 0 VANKCDLE   GTP
RAS1$YEAST =       :                   12 0 KIVVVGGGGVGKSALTIQFIQSYFVDEYDPTI   58 0 SILDILDTAGQEEYSA  121 0 VGNKLDLE   GTP
RAS2$DROME =       :                    7 0 KLVVVGGGGVGKSAITIQFIQSYFVTDYDPTI   56 0 IFYLVLDTAGQEEFSA  119 0 VGNKCDLK   GTP
RAS2$YEAST =       :                   12 0 KLVVVGGGGVGKSALTIQLTQSHFVDEYDPTI   58 0 SILDILDTAGQEEYSA  121 0 VGNKSDLE   GTP
RAS3$DROME =       :                    5 0 KIVVLGSGGVGKSALTVQFVQCIFVEKYDPTI   51 0 CMLEIVNTAGTEQFTA  113 0 VGNKCDLE   GTP
RASH$CHICK =       :                    5 0 KLVVVGAGGVGKSALTIQLIQNHFVDEYDPTI   51 0 CLLDILDTAGQEEYSA  114 0 VGNKCDLP   GTP
RASH$HUMAN =       :                    5 0 KLVVVGAGGVGKSALTIQLIQNHFVDEYDPTI   51 0 CLLDILDTAGQEEYSA  114 0 VGNKCDLA   GTP
RASH$MSV   =       :                    5 0 KLVVVGAKGVGKSALTIQLIQNHFVDEYDPTI   51 0 CLLDILDTAGQEEYSA  114 0 VGNKCDLA   GTP
RASH$MSVHA =       :                   57 0 KLVVVGARGVGKSALTIQLIQNHFVDEYDPTI  103 0 CLLDILDTTGQEEYSA  166 0 VGNKCDLA   GTP
RASH$RRASV =       :                   64 0 KLVVVGARGVGKSALTIQLIQNHFVDEYDPTI  110 0 CLLDILDTAGQEEYSA  173 0 VGNKCDLA   GTP
RASK$HUMAN =       :                    5 0 KLVVVGAGGVGKSALTIQLIQNHFVDEYDPTI   51 0 CLLDILDTAGQEEYSA  114 0 VGNKCDLP   GTP
RASK$MOUSE =       :                    5 0 KLVVVGAGGVGKSALTIQLIQNHFVDEYDPTI   51 0 CLLDILDTAGQEEYSA  114 0 VGNKCDLP   GTP
RASK$MSVKI =       :                    5 0 KLVVVGASGVGKSALTIQLIQNHFVDEYDPTI   51 0 CLLDILDTTGQEEYSA  114 0 VGNKCDLP   GTP
RASL$HUMAN =       :                    5 0 KLVVVGAGGVGKSALTIQLIQNHFVDEYDPTI   51 0 CLLDILDTAGQEEYSA  114 0 VGNKCDLP   GTP
RASL$MOUSE =       :                    5 0 KLVVVGAGGVGKSALTIQLIQNHFVDEYDPTI   51 0 CLLDILDTAGQEEYSA  114 0 VGNKCDLP   GTP
RASN$HUMAN =       :                    5 0 KLVVVGAGGVGKSALTIQLIQNHFVDEYDPTI   51 0 CLLDILDTAGQEEYSA  114 0 VGNKCDLP   GTP
RASN$MOUSE =       :                    5 0 KLVVVGAGGVGKSALTIQLIQNHFVDEYDPTI   51 0 CLLDILDTAGQEEYSA  114 0 VGNKCDLP   GTP
RHO$APLCA  =       :                    7 0 KLVIVGDGACGKTCLLIVFSKDQFPEVYVPTV   53 0 VELALWDTAGQEDYDR  115 0 VGNKKDLR   GTP
RHO1$HUMAN =       :                    7 0 KLVIVGDGACGKTCLLIVFSKDQFPEVYVPTV   53 0 VELALWDTAGQEDYDR  115 0 VGNKKDLR   GTP
RHO6$HUMAN =       :                    7 0 KLVVVGDGACGKTCLLIVFSKDEFPEVYVPTV   53 0 VELALWDTAGQEDYDR  115 0 VANKKDLR   GTP
RHO9$HUMAN =       :                    7 0 KLVIVGDGACGKTCLLIVFSKDQFPEVYVPTV   53 0 VELALWDTAGQEDYDR  115 0 VGNKKDLR   GTP
SEC4$YEAST =       :                   22 0 KILLIGDSGVGKSCLLVRFVEDKFNPSFITTI   69 0 VKLQLWDTAGQERFRT  131 0 VGNKSDME   GTP
YPT1$YEAST =       :                   10 0 KLLLIGNSGVGKSCLLLRFSDDTYTNDYISTI   57 0 VKLQIWDTAGQERFRT  119 0 VGNKCDLK   GTP


EFG$ECOLI  = elongation factor         12 1 NIGIsAHIDAGKTTTTERILFYTGVNHKIGEV   82 0 HRINIIDTPGHVDFTI  140 0 FVNKMDRM   GTP
EFTU$YEAST =       :                   50 1 NIGTIGHVDHGKTTLTAAITKTLAAKGGANfL  111 0 RHYSHVDCPGHADYIK  170 0 FVNKVDTI   GTP
GBT2$BOVIN = G-protein                 35 1 KLLLLGAGESGKSTIVKQMKIIHQDGYSpEEC  194 0 LNFRMFDVGGQRSERK  267 0 FLNKKDLF   GTP
GST1$YEAST =       :                  262 1 SLIFMGHVDAGKSTMgGNLLYLTGSVDkRTIE  338 0 RRYTILDAPGHKMYVS  404 0 VVNKMDDP   GTP
IF2$BACST  = initiation factor         246 1 vVTIMGHVDHGKTTLLDAIRHSKVTEQEAGGI  291 0 KKITFLDTPGHEAFTT  349 0 AINKMDKP   GTP
IF2$ECOLI  =       :                   395 1 vVTIMGHVDHGKTSLLDYIRSTKVASKEAGGI  440 0 GMITFLDTPGHAAFTS  498 0 AVNKIDKP   GTP
LEPA$ECOLI = LEPA protein                6 1 NFsIIAHIDHGKSTLSDRIIQICGGLSDREME   71 0 YQLNFIDTPGHVDFSY  129 0 VLNKIDLP   GTP
RAB3$RAT   = ras related oncogene products  24 1 KILIIGNSSVGKTSFLFRyADDSFTPAFVSTV   71 0 IKLQIWDTAGQERYRT  133 0 VGNKCDME   GTP
RAS1$DROME =       :                    5 1 KLVVVGpGGVGKSALTIQLIQNHFVDEYDPTI   51 0 CLLDILDTAGQEEYSA  114 0 AGNKCDLA   GTP
SUF1$YEAST = SUF12 supressor protein  262 1 SLIFMGHVDAGKSTMgGNLLYLTGSVDKRTIE  338 0 RRYTILDAPGHKMYVS  404 0 VVNKMDDP   GTP


EF2$MESAU  = elongation factor         21 1 NMsVIAHVDHGKSTLTDSLVCKAGIIASARAG   98 1 FLINLIDsPGHVDFSS  156 0 MMNKMDRA   GTP
PPCK$RAT   = PEP carboxykinase        232 1 gSGYGGNSLLGKKCFALRIASRLAKEEGWLAE  312 0 IAWMKFDAQGNLRAIN  386 1 WkNKEWRP   GTP


PPCK$CHICK =       :                  232 1 gSGYGGNSLLGKKCFALRIASRIAKEEGWLAE  312 1 IAWMKFDeLGNLRAIN  386 1 WkNKDWTP   GTP
```

Appendix II.

```
ERA$ECOLI  = ERA gene product              10 2 fIAIVGRPNVGKSTLLNKLLgQKISITSRKAQ   56 2 YQAIYVDTPGlhNESK  122 0 AVNKVDNV   GTP

POLG$WNV   = polyprotein                 1407 3 gLMFAAFVISGKSTdMWIeRTADITWESDAEI  122 2 GAVTLsNfGGKVMMTV   71 0 GVNKQTAM    ?
                                           . .                                    . .                 1713 1 AINKRlRT

ITAV$HUMAN = vitronectin receptor         912 4 iVCqVGRLDRGKSAILYVkSLLWTeTFMNKEN   97 2 CQpIeFDATGNRDYAK  938 0 FMNKENQN    ?
NIFH$METVO = nitrogenase                    3 3 KFCIyGKGGIGKSTnVGNMAAALAeDGKKVLV  101 2 TAVDMLDrLGvYDQLK  234 2 IANKfrEL   ATP
                                           . .                                    . .                  246 1 YeNKKTTI
POLG$ENMTV = polyprotein                 1307 3 vVVLrGAAGQGKSvTSQIIAQSVSKMAFGRQS 1794 2 GSaIICNVNGKKAVYg  473 2 TnNKRYpY    ?
                                           . .                                    . .                 2203 1 pANKTTTF
TALA$MPOV3 = large T antigen              570 3 NILFrGpVNSGKTgLAAALISLLGGKSLNINC  617 2 FVVCFeDVkGQIALNK  629 1 ALNKQlQP  (no?)
TALA$POVMA =       :                      568 3 NILFrGpVNSGKTgLAAALISLLGGKSLNINC  615 2 FVVCFeDVkGQIALNK  627 1 ALNKQlQP  (no?)
```