



Systematic investigation of sequence and structural motifs that recognize ATP



Ke Chen^{a,*}, Dacheng Wang^a, Lukasz Kurgan^b

^aSchool of Computer Science and Software Engineering, Tianjin Polytechnic University, Tianjin 300387, China

^bDepartment of Electrical and Computer Engineering, 2nd floor, ECERF (9107 116 Street), University of Alberta, Edmonton, AB T6G 2V4, Canada

ARTICLE INFO

Article history:

Received 16 December 2014

Received in revised form 5 April 2015

Accepted 18 April 2015

Available online 20 April 2015

Keywords:

ATP

Binding site

Sequence motif

Structural motif

ABSTRACT

Interaction between ATP, a multifunctional and ubiquitous nucleotide, and proteins initializes phosphorylation, polypeptide synthesis and ATP hydrolysis which supplies energy for metabolism. However, current knowledge concerning the mechanisms through which ATP is recognized by proteins is incomplete, scattered, and inaccurate. We systemically investigate sequence and structural motifs of proteins that recognize ATP. We identified three novel motifs and refined the known *p*-loop and class II aminoacyl-tRNA synthetase motifs. The five motifs define five distinct ATP–protein interaction modes which concern over 5% of known protein structures. We demonstrate that although these motifs share a common GXG tripeptide they recognize ATP through different functional groups. The *p*-loop motif recognizes ATP through phosphates, class II aminoacyl-tRNA synthetase motif targets adenosine and the other three motifs recognize both phosphates and adenosine. We show that some motifs are shared by different enzyme types. Statistical tests demonstrate that the five sequence motifs are significantly associated with the nucleotide binding proteins. Large-scale test on PDB reveals that about 98% of proteins that include one of the structural motifs are confirmed to bind ATP.

© 2015 Elsevier Ltd. All rights reserved.

1. Introduction

Adenosine-5'-triphosphate (ATP) is a multifunctional nucleotide that plays an important role in energy metabolism, signaling, and replication and transcription of DNA. ATP is primarily responsible for providing energy for enzymes and a multitude of cellular processes including biosynthetic reactions, motility and cell division (Kadi et al., 2007; Asenjo et al., 2006; Hu et al., 2002). Currently, the Protein Data Bank (PDB) (Berman et al., 2000) contains 6134 protein entries that are annotated as “ATP binding”. These proteins have diverse tertiary structure, molecular function and they encompass a number of different enzyme types including kinases, synthetases, hydrolases and ATPases. In spite of the ubiquity of ATP binding and the fact that ATP binding sites are regarded as valuable drug targets for antibacterial and anti-cancer chemotherapy (Maxwell and Lawson, 2003; Rock et al., 2007; Metlitskaya et al., 2006), no comprehensive attempts were made to categorize and describe similarities in the ATP binding.

The sequence/structural motifs are becoming increasingly important in the analysis of protein's function. The PROSITE (Hulo et al., 2008) database stores the sequence patterns, signatures, and

profiles for specific protein families. However, some of the sequence signatures and profiles are generated computationally and they lack annotations of the associated biological functions. Since sequence patterns are usually summarized within a specific protein family, motifs shared by several superfamilies would not be included in the PROSITE. Over the last two decades, significant efforts were put into characterization of sequence motifs associated with specific families or specific molecular function. For instance, sequence motifs have been defined for subclasses of membrane-bound chemokines (Bazan et al., 1997); LXXLL was defined as the binding motif that facilitates the interaction of different proteins with nuclear receptors (Heery et al., 1997); A short amino acid (AA) segment of the monoubiquitinated endocytic proteins Eps15 and Eps15R was found to be indispensable for monoubiquitination (Polo et al., 2002); Dozens of sequence motifs were generated for serine, threonine and tyrosine phosphorylation sites (Schwartz and Gygi, 2005). Existing works have also explored sequence and structural similarity for some of the ATP-binding proteins. The sequential *p*-loop, GXXXXGKS(T), motif which interacts with ATP and its analogs was described over a decade ago (Walker et al., 1982; Saraste et al., 1990; Kobayashi and Go, 1997a) and more recently sequence motifs in kinases and class II aminoacyl-tRNA synthetase (AARS) were investigated (Barker and Dayhoff, 1982; Brenner, 1987; Scheeff and Bourne, 2005; Eriani et al., 1990). A few structural motifs were also

* Corresponding author. Tel.: +86 13752339078.

E-mail address: kchen1.tjpu@hotmail.com (K. Chen).

discussed for phosphate and adenosine groups of ATP, however these motifs are incomplete, they lack characterization of sequence patterns and were generated from proteins with high sequence identity (Denessiouk et al., 2001; Brakoulias and Jackson, 2004).

We use a comprehensive dataset of diverse ATP–protein complexes to find and characterize, both at the sequence and structure levels, five generic binding motifs that facilitate the interaction between ATP and proteins. Although we found that the motifs share some similarities, i.e., GXG tripeptide appears in all motifs, they recognize ATP in different ways. The *p*-loop motif targets the phosphates of ATP, the class II AARS motif interacts with the adenosine of ATP, while the other three motifs bind to both phosphates and adenosine. We refined the known form of the *p*-loop motif and we show that some of the motifs that define different classes of AARSs overlap with motifs responsible for ATP binding. Three motifs including class II AARS, protein kinase and actin-like ATPase are associated with specific protein families and biological processes while the *p*-loop and strand-sandwich motifs are involved in a diverse range of processes. We applied the five motifs to predict ATP-binding proteins by scanning the sequences and structures deposited in PDB. Evaluation using gene ontology annotation (GOA) (Barrell et al., 2009) shows that close to 98% of the predictions are confirmed as the ATP-binding proteins.

2. Materials and methods

2.1. Data preparation

A set of all available 413 ATP–protein complexes was extracted from PDB as of December 16, 2013. The corresponding protein sequences were filtered with CD-hit (Huang et al., 2010) to reduce the pairwise sequence identity in this set to less than 40%. The resulting 137 ATP–protein complexes were used to derive the binding modes.

2.2. Definition of protein–ATP interaction

Hydrogen bonds were calculated with HBPLUS (McDonald and Thornton, 1994). The method finds all proximal donor (D) and acceptor (A) atom pairs that satisfy specified geometrical criteria for the formation of the bond. Theoretical hydrogen atom (H) positions of both protein and ligand are calculated with REDUCE program (Word et al., 1999).

A non-hydrogen atom A_1 of a protein and a non-hydrogen atom A_2 of ATP form van der Waals contact if the distance d between these two atoms satisfies

$$d < \text{vdW}(A_1) + \text{vdW}(A_2) + 0.5 \text{ \AA}$$

where $\text{vdW}(A_i)$ is the vdW radii of A_i and where these two atoms do not form a hydrogen bond. This definition is consistent with a recent study concerning investigation of atomic level patterns in protein–small ligand interaction (Chen and Kurgan, 2009).

2.3. Feature-based representation of the ATP–protein interaction

ATP contains 31 non-hydrogen atoms. The interactions for each of these atoms, denoted as A_i , $i = 1, 2, \dots, 31$, are described using 120 dimensional feature vector. Consequently, we compute $31 \times 120 = 3720$ features for each ATP–protein complex. The features encode coordinates of non-hydrogen atoms of a given protein that are within 3.5 Å from A_i ; the threshold value is consistent with a previous study (Chen and Kurgan, 2009). If no such atoms are found then the 120 dimensions for A_i are set as 0. Otherwise, the selected atoms are categorized into 40 groups, denoted as G_j , $j = 1, 2, \dots, 40$, where each group is described by three features that represent three coordinates of the spatial

position of its representative atom. Each group concerns one of the 20 AA types and two types of atoms, i.e., nitrogen and oxygen that have strong propensity to form hydrogen bonds and the remaining atoms. If none of the selected atoms belongs to G_j , then the corresponding three feature values for are set to 0. Otherwise, the atom with the minimal distance to A_i is selected as the representative atom for G_j . The coordinates of the representative atom are normalized with respect to the ATP molecule. For each of the 40 groups, A_i and two atoms of ATP that are covalently connected to A_i are set as the reference points. We perform one translational and three rotational transformations such that A_i is located at the origin and other two atoms are on the y -axis and xy -plane.

We perform feature selection to find the most relevant coordinates for the ATP–protein interaction. The 3720 features correspond to 1240 sets of normalized coordinates. The normalized coordinates are spaced between 2 Å and 3.5 Å from the origin; the corresponding distribution is given in Fig. 1 in Supplement. For each set of coordinates, we consider a set X of its non-zero values across different structures where a given set is regarded as non-zero if at least one of its coordinates does not equal 0. The cardinality of X quantifies the number of atoms from different structures which are encoded in the coordinate set. We investigate whether the atoms from X are clustered together or spaced randomly, when compared to an average distribution of coordinates in our dataset. The randomly spaced coordinate sets, which correspond to atoms for which no similar spatial arrangement is found, are discarded. We compute average Euclidian distance D_1 based on the distances for all pairs of atoms from X and we compare that value to an average distance D_2 of a set of random coordinates that follow the distribution from Fig. 1 in Supplement and which are sampled 10000 times. Given that smaller D_1 values correspond to a more compact distribution of the atoms, we regard X as clustered (non-random) if D_1 is smaller than 95% of the 10000 D_2 values. As a result, total of 192 features that correspond to 64 coordinate sets are retained for the subsequent clustering.

2.4. Clustering

A single linkage method, which has been widely applied in clustering of protein and DNA/RNA sequences, e.g., in BLASTClust (Altschul et al., 1997) and CluStr (Petryszak et al., 2005), was used to cluster the 137 structures encoded using the 192 features. This is a hierarchical clustering method in which the distance between two clusters is defined as the distance between the two closest elements in these clusters. We used fifth level in the dendrogram generated by this clustering method to cluster the binding sites. This level includes 5 clusters with at least 4 proteins, which were used to derive the binding modes, while the remaining clusters are smaller and include 2 or fewer proteins. Four out of the five large clusters are confined to a single protein superfamily and the corresponding structural analysis reveals that these clusters have similar ATP binding sites. The sixth level further subdivides these clusters and thus the corresponding modes would be less generic. On the other hand, the fourth level includes 2 large clusters with proteins that belong to 3 superfamilies for which we could not derive unique structural motifs. The dendrogram of the clustering result is demonstrated in Fig. 4 in Supplement.

2.5. Structural motif

Each cluster includes multiple structures that share similar spatial arrangement of binding residues; see Fig. 1A. The structures of the binding sites were superimposed by performing one translational and three rotational transformations such that the C5 atom of ATP is at the origin point, the C4 and N7 atoms of ATP are

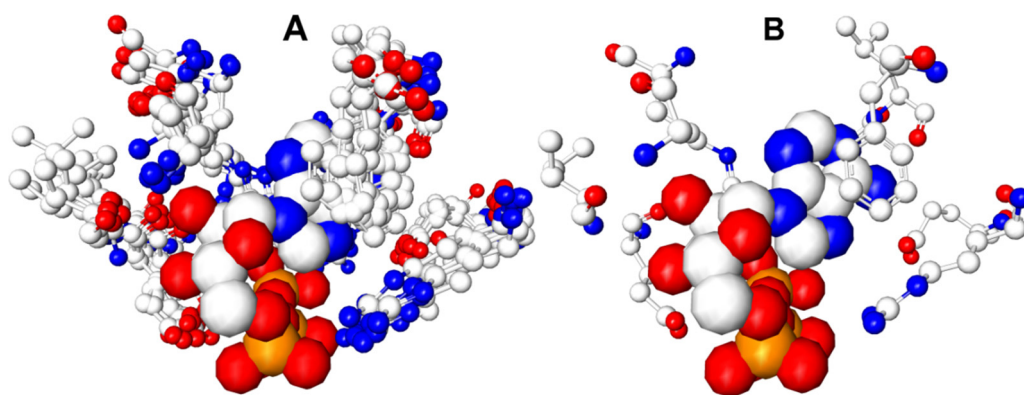


Fig. 1. (A) Superimposed cluster of ATP-binding site structures that belong to the “class II aminoacyl- tRNA synthetase” binding mode. (B) Structural template (structural motif) of the “class II aminoacyl- tRNA synthetase” binding mode defined by selecting median structure for each binding residue from the cluster shown in Fig.1A. ATP is shown in space-fill format and the binding residues are using ball and stick format.

on y -axis and on xy -plane, respectively. The structural motif is defined using median structures of individual residues in the structures from a given cluster; see Fig. 1B. The median structure for a given binding residue is selected as the set of coordinates that gives minimal value of distance to the coordinates of the remaining corresponding residues in the cluster

$$D_i = \sum_{i \neq j} \text{RMSD}(S_i, S_j)$$

where $i = 1, 2, \dots, x$ is a structure index, and S_1, S_2, \dots, S_x denote normalized coordinates of this residue in x structures in the cluster.

2.6. Statistical test of the significance of the sequence motifs

The sequence motifs were searched against a set of protein sequences that were filtered at 40% sequence similarity using CD-Hit²³ from the dataset of sequences extracted from PDB as described in the “prediction of ATP-binding proteins” section. Each sequence was annotated as “nucleotide-binding” or “not nucleotide-binding” based on the GOA (Barrell et al., 2009). The statistical test is based on a method that was used by Senes and colleagues to verify significance of sequence motifs in transmembrane helices (Senes et al., 2000). The total number of protein sequences in the set is denoted as n_1 and the number of

sequences annotated as “nucleotide-binding” in the entire set is denoted as n_2 . For a given sequence motif X , the number of protein sequences that contain motif X is denoted as n_3 . The motif X is regarded as associated with the nucleotide binding proteins if its occurrence in n_2 “nucleotide-binding” proteins is significantly higher than its occurrence Y in n_2 randomly selected sequences in the entire dataset. Assuming the sequences in the set are independent (sequences that share more than 40% similarity were removed), Y satisfies binomial distribution. The probability of Y being equivalent to or greater than n_3 , $\text{Pr}(Y \geq n_3)$, is used to evaluate the significance. Since some positions in the motifs are not invariant but rather they are conserved in majority of the aligned sequences (see Table 1 in the main document and Table 1 Supplement), a protein sequence is regarded to contain motif X if a segment of the sequence is identical to X or if it differs from X at at most one position.

2.7. Prediction of ATP-binding proteins

The prediction method for the p -loop binding mode differs from the prediction for the other four modes. We use both the sequential and the structural motifs for the p -loop mode, while for the remaining modes we use only the structural motifs.

Table 1

Protein chains that include the p -loop binding mode. The binding segment is shaded and the most frequent AA type at a given position is shown in bold.

Polymer name	PDBcode:(chain)	Binding segment	SCOP classification	EC number
Phosphoenolpyruvate carboxykinase	1AYL:A	G L S G T G K T T L S T	c.91.1.1	4.1.1.49
Thymidylate kinase	1E2Q:A	G V D R A G K S T Q S R	c.37.1.1	2.7.4.9
Shikimate kinase	2IYW:A	G L P G S G K S T I G R	c.37.1.2	2.7.1.71
Antigen peptide transporter 1	2IXG:A	G P N G S G K S T V A L	Unavailable	Unavailable
Protein (chloramphenicol phosphotransferase)	1QHJ:A	G G S S A G K S G I V R	c.37.1.3	2.7.1.-
Gluconate kinase	1K05:A	G V S G S G K S A V A S	c.37.1.17	2.7.1.12
Myosin ii heavy chain	1FMW:A	G E S G A G K T E N T K	Unavailable	Unavailable
Hsdr	2W00:A	H T T G S G K T L T S F	Unavailable	3.1.21.3
Arsenical pump-driving atpase	1H0:A	G K G G V G K T T M A A	c.37.1.10	3.6.3.16
Dephospho-coa kinase	1JJV:A	G G I G S G K T T I A N	c.37.1.1	2.7.1.24
Psp operon transcriptional activator	2C96:A	G E R G T G K E L I A S	NA	Unavailable
ATP synthase subunit alpha	2R9V:A	G D R Q T G K T A I A I	Unavailable	3.6.3.14
Ced-4	2A5Y:B	G R A G S G K S V I A S	Unavailable	Unavailable
Nitrogenase iron protein 1	2C8V:A	G K G I G K S T T T T Q	c.37.1.10	1.18.6.1
Cystic fibrosis transmembrane conductance regulator	1XML:C	G S T G A G K T S L L M	c.37.1.12	3.6.3.49
DNA mismatch repair protein muts	1W7A:B	G P N M G G K S T Y M R	c.37.1.12	Unavailable
Twitching motility protein pilt	2EWW:A	G P T G S G K S T T I A	Unavailable	Unavailable
TT1252 protein	1UF9:C	G N I G S G K S T V A A	c.37.1.1	Unavailable
DNA packaging protein gp17	200H:A	L S R Q L G K T T V A I	Unavailable	Unavailable

First, a protein sequence of a given PDB entry is compared with the sequential GXXGKG(T)T motif. Protein segments that have at least four residues in common with the sequential motif are kept (the alignment between X and any amino acid is not counted); we assume no deletions or insertions in the alignment. Next, the RMSD value is computed between the structures of the main chain of the selected segments and the main chain of the corresponding structural motif. The structures with RMSD < 0.5 Å are predicted as the ATP-binding sites, and the corresponding proteins are predicted as ATP-binding using the *p*-loop mode.

The structural motifs for the remaining 4 modes are defined using the median structures for each binding residue which is shaded in Table 1 in the main document and in Table 1 in Supplement, respectively. We predict a given protein structure as ATP-binding if it contains four residues that share similar spatial arrangement of their side chains with the side chains of residues in the structural motif. The similarity concerns two conditions: (1) the corresponding residues are of the same AA type; and (2) the RMSD between the structure of the four residues in the input protein and the motif < 1 Å.

3. Results

Five sequence/structural motifs were extracted from a dataset of 137 ATP–protein complexes (details concerning the dataset preparation were given in Supplement). For each complex, a set of features are calculated to represent the ATP–protein interaction. The non-significant features are removed and the remaining features are used for clustering. The clustering yields five clusters where each cluster contains at least four proteins which interact with ATP in a similar fashion (details concerning feature representation of ATP–protein binding, feature selection and clustering were given in Supplement). Each cluster defines an ATP–protein interaction mode and is associated with different protein families. Sequence/structural motifs are extracted for each cluster (mode). Sequence motifs are found by sequence alignment. The structural motifs are defined as a collection of spatial positions of binding residues that share similar side chain arrangement in majority of the structures in a given cluster as described in Supplement. Here we discuss the sequence conservation and similarity of the spatial arrangement of binding residues for each of the five motifs.

3.1. The *p*-loop motif

This motif was originally introduced by Walker and colleagues as GXXXXGKS(T) (Walker et al., 1982). In later studies, Kobayashi and Go (1997a) found a similar spatial arrangement of the backbones and side chains of several residues that are involved in binding ATP/ADP analogues in proteins with different folds. Although the *p*-loop motif is already known, it defines the most common way in which proteins recognize ATP and our analysis results in a slight refinement of its form.

Total of 19 out of 137 proteins interact with ATP through this motif (mode), see Table 1. These structures are annotated with structural classification from SCOP (Andrejeva et al., 2008) and enzyme commission (EC) numbers. The sequence identity between any pair of these proteins is below 25%. They interact with ATP mainly through a segment of six residues which are shaded in Table 1. The alignment of the common 19 binding segments reveals that the central two residues, Gly and Lys are invariant. Residues on three other positions are also relatively conserved, i.e., the first, fifth and last positions have the probability of 74% to be Gly 95% to be Ser or Thr (Ser and Thr share similar physicochemical properties), and 58% to be Thr respectively. The segment is extended by three residues on both sides to investigate sequence

conservation of the adjacent residues. Only the first residue in the extended segment is relatively conserved and has a probability of 89% to be Gly. The alignment suggests that *p*-loop binding mode is characterized by the gXXgXGKs(t) t sequential motif where X denotes a wild card, AAs in the bracket is substitutable with the preceding AA, a letter in upper/lower case indicates an invariant residue / residue conserved in majority of the aligned segments, respectively.

The 6-residue segment mainly binds to the β -phosphate and it does not interact with the adenosine group. Atomic view of the hydrogen bonds and van der Waals contacts between the *p*-loop segment and ATP is shown in Table 2. After superimposing all 19 structures into the same coordinate system, we found that these structures share similar secondary structure arrangement in the vicinity of the *p*-loop segment (i.e., the *p*-loop motif is always preceded by a beta sheet and followed by a helix, see Table 2) and similar spatial arrangement of the side chains of the *p*-loop residues. This suggests that the tertiary structure of the *p*-loop segment is also conserved.

The *p*-loop segment interacts with ATP through the phosphates. Therefore, it binds not only to ATP/ADP and their analogs but it also interacts with other ligands that contain phosphates or sulfates. For instance, a phosphate group (PO_4^{3-}) was observed to interact with the *p*-loop motif in chain C of vitamin B12 import ATP-binding protein btuD (PDB code: 2QJ9), see Fig. 2A. The corresponding binding segment GPNGAGKST matches all six conserved residues of the sequential *p*-loop motif gXXgXGKs(t) t. A sulfate group (SO_4^{2-}) interacts with the *p*-loop motif in chain A of cell division protein ftsY (PDB code: 2Q9A (Reyes et al., 2007)), see Fig. 2B. The binding segment GVNGVGKTT also matches all six conserved residues with the sequential motif of the *p*-loop mode. This demonstrates that the *p*-loop motif binds not only to ATP and its analogs, but it also interacts with phosphate and sulfate groups, and ligands that include these groups. The *p*-loop binding site is not limited to the ATP, but rather it should recognize a diverse range of nucleotides, all of which contain between one and three phosphates.

3.2. Class II aminoacyl-tRNA synthetase motifs

Prior works reveal three sequence motifs which were used to partition the AARSs into classes (Eriani et al., 1990). The first motif is summarized as $g\varphi XX\varphi XXP\varphi\varphi$ where X stands for any residue, φ stands for hydrophobic residues, and a letter in upper/lower case indicates an invariant/conserved in majority of the aligned segments residue, respectively. The second motif includes two sequence segments, F(YH)RXE(D) segment followed 4–12 residues and the R(H)XXXFXD(E) segment. The third motif is $\lambda X\varphi g\varphi g\varphi eR\varphi\varphi\varphi\varphi$ where λ stands for small amino acids including P, G, S and T. We show that these motifs are essential for AARS–ATP interaction.

Our analysis shows that 11 out of 137 proteins interact with ATP based on the class II aminoacyl-tRNA synthetase mode (motifs), see Table 1 in Supplement. These proteins belong to class II AARS and include lysyl-, prolyl-, aspartyl-, glycyl, histidyl-, threonyl- and seryl-tRNA synthetases. The sequence similarities of any pair of these chains are below 35%. These AARSs include three sequence segments with nine positions at which residues interact with ATP in majority of the 11 structures, see Table 3A. Among the nine interacting residues, two Arg residues (at the first and ninth positions) are invariant while the remaining seven residues are not fully conserved. We extended the analysis to 3 positions away from the binding residues to inspect the sequence conservation and found two sequence motifs. The first motif includes fRXe segment followed 5–8 residues and the h(r)XXef segment, and it corresponds to the first of the three segments. The sequence

Table 2

Overview of the five binding modes. The secondary structure arrangement of the binding site is shown using cartoon representation where helices are colored in red, strands in yellow, coils in blue. The atomic view of the ATP-binding residues interaction shows ATP using the “stick” form and binding residues in the “ball and stick” form. Oxygen is colored in red, nitrogen in blue, carbon in white and phosphorus in yellow. Distances between selected pairs of spatially close atoms are shown in green. For the sequential motif, X denotes a wild card, ϕ indicates hydrophobic residues, λ stands for small amino acids including P, G, S and T, AA in the bracket is substitutable with the preceding AA, and a letter in upper/lower case indicates an invariant residue / residue conserved in majority of the aligned segments, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

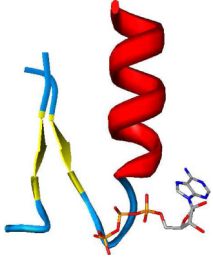
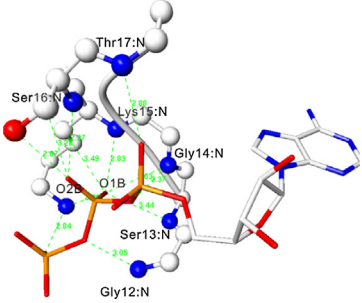
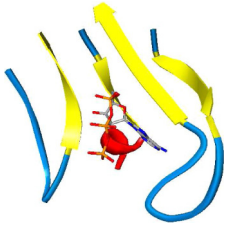
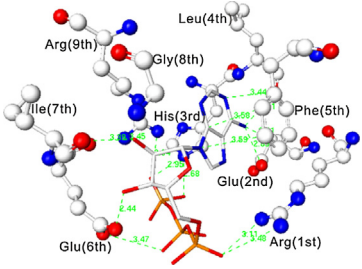
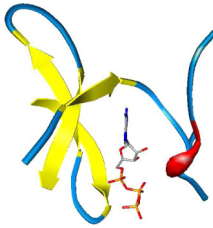
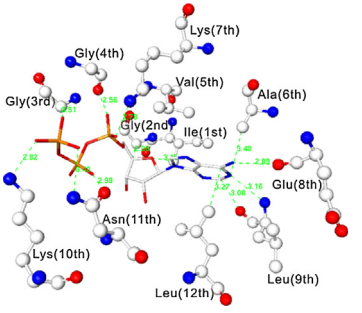

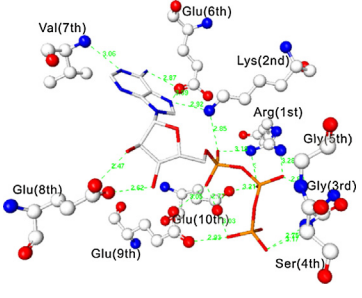
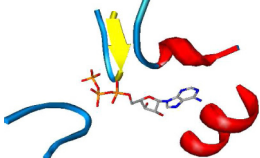
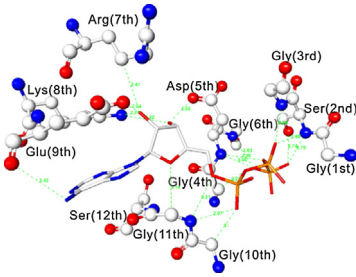
Binding mode	Secondary structure arrangement of the binding site	Atomic view of the ATP-binding site interaction	Sequential motif	Chemical group recognized by motif
<i>p</i> -Loop			gXXgXGKS(t) t	Phosphates
Class II aminoacyl-tRNA synthetase			rRXe h(r)XXef gXgXXR	Adenosine
Protein kinase			I(L)GXGXfgXv ϕ A ϕ K lvme rD ϕ kpXNXI	Phosphates and adenosine
Strand sandwich			p ϕ ϕ V(I) KXXXXXgXgveXXXXXXeXXXXe	Phosphates and adenosine
Actin-like ATPase domain			DnGT(S) gXXkXG dsGdGv E(R) XiKE(R) VLsGGS(T)T	Phosphates and adenosine

Table 2 (Continued)

Binding mode	Secondary structure arrangement of the binding site	Atomic view of the ATP-binding site interaction	Sequential motif	Chemical group recognized by motif
				

conservation in the second segment is too low to define a motif. The second motif is based on the last segment and it takes a form of gXgXXR. We observe that our first and second motifs are consistent, with a few refinements, with the second and third motifs in the study by Eriani et al. (1990), respectively. When compared to Eriani's motifs, our motifs are more compact and we could not group the residues based on the hydrophobicity, likely because our source sequences were more diverse and produced more substitutions at some positions that were assumed by Eriani et al. (1990) to be conserved. The previous study has shown that these motifs differentiate between class II and class I AARSs. Our work indicates that two of these motifs, upon some refinement, are responsible for binding of ATP to class II AARSs.

The 11 structures share similar spatial arrangement of the side chains of the 9 binding residues. The binding site consists of 4 anti-parallel strands, 2 loops and 1 helix. A set of median structures of each structurally conserved binding residue (Fig. 1B) among its multiple structures (Fig. 1A) is assumed as the structural motif. This motif, together with analogous structural motifs for the other four modes, is used as a template for prediction of ATP-binding proteins.

3.3. Protein kinase motifs

Two sequential motifs have been previously proposed for kinases by Brenner (Brenner, 1987) and Baker (Baker and Dayhoff, 1982), respectively. The first motif is formulated as L(IV)H(Y)XDF (ILMVY)XXXNXF(ILMV)F(ILMV) and the second takes form of LGXGXFGXV, which is generalized as L(IV)GXGXF(Y)GXL(IV) in a later study (Bairoch and Claverie, 1988). These motifs were used to discriminate the kinases among other proteins. We note that these motifs do not occur in all kinases, i.e., the first motif fails for cAMP- and cGMP-dependent kinases, protein kinases C and the *abl*, *fes*, *fps*, *ros* and *trk* oncogene products.

We found that 9 out of 137 proteins interact with ATP through common motifs in which all 9 proteins are kinases, see Table 1 in Supplement. The sequence identity of any pair of their chains is below 25%. They include four separate sequence segments where residues at twelve positions interact with ATP in majority of the protein kinase structures (see Table 3B). Atomic view of the hydrogen bonds and van der Waals contacts formed between the binding residues and ATP is shown in Table 2. Among the interacting residues, four residues including Gly at the third

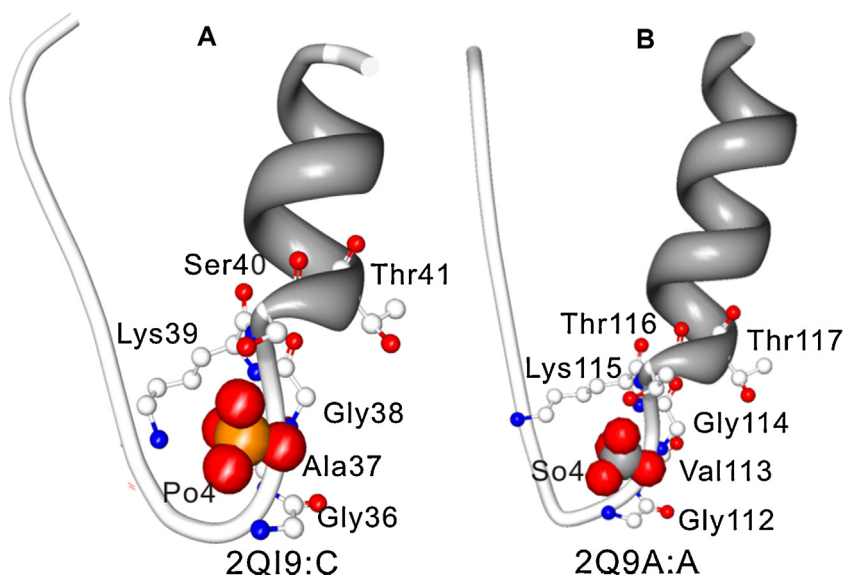


Fig. 2. (A) Phosphate bound to chain C of vitamin B12 import ATP-binding protein btuD (PDB code: 2QI9). The binding segment between Gly36 and Thr41 shares five residues with *p*-loop motif GSGKS(T)T. (B) Sulfate bound to chain A of cell division protein ftsY (PDB code: 2Q9A). The binding segment between Gly112 and Thr117 shares five residues with the *p*-loop motif.

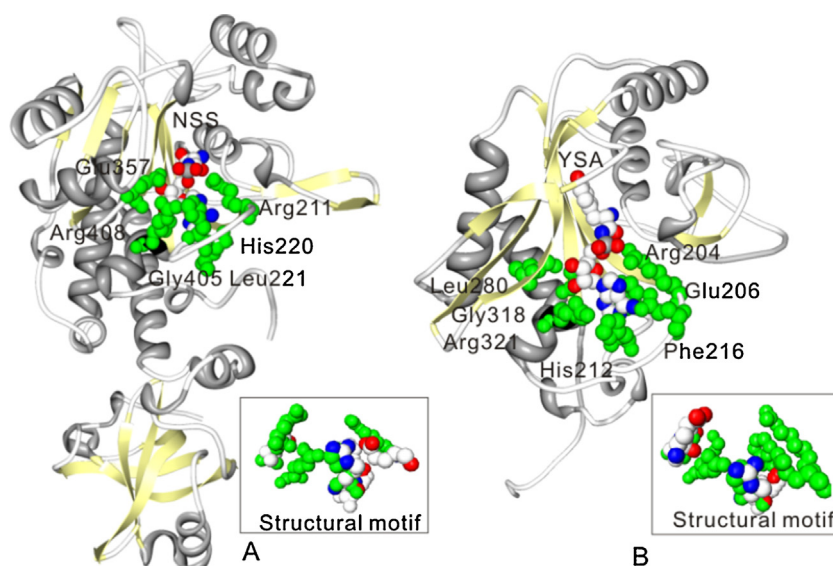


Fig. 3. (A) Structure of asparaginyl- tRNA synthetase (PDB code: 1 × 55). (B) Structure of phenylalanyl- tRNA synthetase (PDB code: 2ALY). The structural motifs used to predict these proteins as ATP-binding are shown in the same spatial orientation in boxes located below the corresponding protein structures. Residues that match between the protein structure and the structural motif are shown in green. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

position, Ala at the sixth position, Lys at the seventh position and Asn at the eleventh position are fully conserved. We extended each of these segments by 3 positions on both sides to investigate sequence conservation. Sequence alignment reveals that the sequential motifs for the four binding segments are I(L)GXGXfgXv, φ A φ K, lvme and rD φ kpXNXI, respectively, where X denotes a wild card and φ indicates hydrophobic residues. Our first motif is consistent with Baker's LGXGXFGXV motif with a slight modification concerning conservation of a few residues at the C-terminus. This indicates that Baker's motif defines the kinases based on their binding to ATP. The remaining three motifs are unique to this study. The four motifs form the ATP-binding site and are shared by a number of protein kinases, although they are not specific to all kinases, i.e., some kinases interact with ATP through the *p*-loop motif.

The 9 structures also share similar arrangement of the side chains of several binding residues. Similar to class II aminoacyl-tRNA synthetase binding mode, the ATP-binding site of protein kinase also consist of 4 anti-parallel strands and 1 helix. However, the key residues and the sequence motifs of these two modes are different, see Table 3A and B. The spatial orientation of the superimposed side chains of virtually all twelve binding residues is similar; they are shaded in Table 3B. The median structures of the superimposed binding residues define the structural motif for this binding mode.

3.4. Strand sandwich motifs

This binding mode covers 7 ATP-interacting proteins; see Table 1 in Supplement. They include carboxylase, synthetase, formyltransferase and ATPase subunit. Although these proteins differ in their cellular functions, they share similar secondary structure arrangement at the ATP binding site. The binding site mainly consists of two layers of strands with ATP located in between, see Table 2. Except the chain A of two carboxylases (PDB code: 1DV2 (Thoden et al., 2000) and 3BG5 (Xiang and Tong, 2008)) that share 31% sequence identity, the pairwise sequence similarity for the remaining protein chains is below 20%.

Previous studies have demonstrated that three ATP-dependent enzymes with different folds have similar spatial arrangement at

their ATP-binding sites which consists of two anti-parallel β -sheets (Hibi et al., 1996; Kobayashi and Go, 1997b). The EKWL segment establishes hydrogen bonds and hydrophobic interaction with cofactors and the authors also show that Met154, Met259 and Leu269 are crucial for the interaction between these proteins and ATP.

We cover a wider range of protein families and define two sequence motifs and several positions that characterize proteins that interact with ATP based on the strand-sandwich binding mode. These protein chains bind to ATP through four segments and ten positions (see Table 3C) for majority of the seven complexes. The Lys residue at the second position and Glu at the sixth position are invariant. Alignment of the binding segments reveals two sequential motifs, one for the second segment, p $\varphi\varphi$ V(I)KXXXXXgXg, and the other for the fourth segment, veXXXXX-XeXXXXe. Although some binding residues of the remaining first and third segments are relatively conserved, i.e, Lys(Arg) at the first position, Glu and Ile(Leu Val) at the sixth and seventh positions, respectively, each of these segments contains no more than two conserved residues. Therefore, we did not define sequence motifs for these two segments. The previously discovered EKWL segment corresponds to the third segment in the alignment, see Table 3C. The Glu and Leu residues are among the ten positions that define the ATP interaction for this mode. This segment is not converted into a sequence motif due to the low conservation of the adjacent residues. However, these two residues are included in the structural motif and used for prediction of ATP-binding proteins. The Met154, Met259 and Leu269 residues suggested by Kobayashi and Go (1997b) differ from the key interacting residues identified in this study, which mostly include hydrophilic residues that favor formation of hydrogen bonds.

Similarly to the other binding modes, the side chains of the binding residues of the "strand sandwich" proteins are structurally conserved. A structural motif is defined as the collection of median structures of the superimposed binding residues.

3.5. Actin-like ATPase motifs

Sequential motifs are also found for 4 actin-related proteins that interact with ATP, see Table 1 in Supplement. Three polymers

include either actin or actin-related proteins and the fourth one is 70 kilodalton heat shock proteins (Hsp70). The pairwise sequence similarity of these chains is below 35%. These proteins interact with ATP through four segments at twelve positions (see Table 3D) for at least three out of the four structures. Atomic view of the contacts formed between protein and ATP is shown in Table 2. The binding sequence segments are relatively conserved and we propose the corresponding four sequential motifs that include DnGT(S) gXXkXG, dsGdGv, E(R) XiKE(R), and VLsGGs(T)T.

The ATP-binding site of this mode consists of 2 helices and 2 loops. The spatial arrangement of the binding residues of the four proteins is also similar. The superimposed structures show similarity in the orientation of the side chains of virtually all twelve binding residues. A structural motif is generated for prediction of ATP-binding proteins.

3.6. The statistical significance of the sequence motifs

Statistical tests were performed to investigate the significance of the association between the motifs and the nucleotide binding. The significance was measured by the probability of Y being equivalent to or greater than n_3 , where Y is the occurrence of a motif X in n_2 randomly selected protein sequences and n_3 is the actual occurrence of motif X in n_2 nucleotide-binding proteins. The probabilities were calculated for each individual sequence motif and the combination of all motifs that belong to the same interaction mode; see Table 2 in Supplement. The $\Pr(Y \geq n_3)$ values of five individual motifs (GXXGXGKS(T)T, I(L)GXGXFQXV, RD ϕ KPXNXL, VEXXXXXXEXXXE, and E(R)XiKE(R)) are below 10^{-10} , which indicates that they are significantly more likely to occur in the nucleotide binding proteins when compared with a generic set of protein chains. These motifs correspond to four different binding modes, the p -loop, protein kinase, strand-sandwich and actin-like ATPase modes. Since in all but the p -loop binding mode the ATP-protein binding spans multiple motifs, we investigate the significance of each set of sequence motifs that define a given binding mode. When using the multiple motifs in each mode together, the probabilities equal 3.9×10^{-11} , 5.2×10^{-43} , 5.2×10^{-18} and 8.8×10^{-21} for the class II aminoacyl-tRNA synthetase, the protein kinase, the strand-sandwich and the actin-like ATPase modes, respectively. This suggests that the five sets of motifs are significantly associated with the nucleotide binding proteins.

3.7. Prediction of ATP-binding proteins

The sequential and structural motifs of the five modes are compared against known protein sequences and structures from PDB to predict potential ATP-binding proteins. 55283 structures were extracted on Jan 19, 2014 from the PDB, and among them 1225, 92, 1098, 80 and 152 proteins are predicted as the ATP-binding proteins for the consecutive five binding modes, respectively. The predicted proteins were evaluated against GOA (Barrell et al., 2009) to investigate predictive performance. Among the proteins that are predicted to interact using the first mode, 736 are annotated as ATP-binding, 445 as nucleotide binding or interacting with ATP analogs, 21 are annotated with a molecular function which is not nucleotide binding, and 23 have no annotation. Therefore, 96.4% of the predictions are confirmed to interact with ATP or its analogs. The success rates for the proteins predicted to use the second, third, fourth and fifth modes equal 97.8%, 99.7%, 93.8% and 100%, respectively. Total of 2593 out of 2647 predictions are confirmed as ATP-binding proteins which corresponds to 98% success rate. Over 90% of the predictions for the second and third modes belong to tRNA-synthetases and protein kinases, respectively.

Fig. 2 in Supplement shows that a substantial fraction of predicted protein chain shares low sequence similarity with the proteins used to derive the corresponding structural/sequence motifs. More specifically, 78% of the predictions for the first mode have less than 20% sequence similarity with any of the 19 proteins used to define the p -loop motif. This is expected since the corresponding sequential motif consists of a single segment which could be found in proteins from a variety of folds. Similarly, low, <20%, identity is characteristic to around 44% of the predictions for the second mode. For the third binding mode, 59% of predictions share 20–40% sequence similarity and additional 17% share less than 20% similarity, which suggests that different types of kinases may share low sequence identity. The two remaining modes include a smaller fraction of about 25% and 12% of predictions that share below 20% sequence identity with sequences used to derive the motifs.

The predictions concerning the “class II aminoacyl-tRNA synthetase” mode serve as a case study. AARS catalyzes the esterification of a specific amino acid or its precursor to one of all its compatible cognate tRNAs to form an aminoacyl-tRNA. The aminoacyl-tRNA is crucial for translation of the genetic code (nucleotide sequence of mRNA) into the AA sequence (polypeptide chain) (Cusack, 1997). The ATP-AARS interaction initiates this series of molecular recognitions. The structural motif for this mode is generated from eleven structures which include lysyl-, prolyl-, aspartyl-, glycy-, histidyl-, threonyl-, pyrrolysine- and seryl-tRNA synthetases. Besides these eight types, our predictions also include asparaginyl- and phenylalanyl- tRNA synthetases. The sequences that belong to these two types of synthetases share less than 20% similarity to any sequence used to derive the corresponding motif. We selected 1×55 (Iwasaki et al., 2006) and 2ALY (Kotik-Kogan et al., 2005) as representative structures for the asparaginyl- and phenylalanyl- tRNA synthetases, respectively. Six residues in 1×55 including Arg211, His220, Leu221, Glu357, Gly405 and Arg408, shown in green in protein structure in Fig. 3A, share similar spatial arrangement with the corresponding structural motif shown in the box. The six residues are at the binding site of the 5'-O-[N-(L-asparaginyl) sulfamoyl]adenosine (NSS), and they interact with the adenosine group of NSS. This suggests that this protein likely binds to ATP that also includes adenosine group, which is confirmed based on the GOA annotation. Similarly, seven residues of 2ALY including Arg204, Glu206, His212, Phe216, Leu280, Gly318 and Arg 321, which were identified using the structural motif, see Fig. 3B, interact with the adenosine group of 5'-O-[N-(L-tyrosyl) sulfamoyl] adenosine (YSA).

The results demonstrate that the predicted ATP-binding proteins include both close homologues of the structure used to derive the corresponding binding motifs as well as proteins that have highly dissimilar sequences and which undertake different molecular functions.

3.8. The role of metal ions in protein-ATP interactions

Metal ions are generally believed to play an important role in mediating interactions between nucleotides and proteins. To this end, we have analyzed the role of metal ions for each of the binding modes. Mg^{2+} is frequently observed at the p -loop motifs, i.e., 13 out of 19 protein-ATP interactions that employ the p -loop binding mode are mediated by Mg^{2+} . Interestingly, only Mg^{2+} , none of other metal ions, is involved in the interaction between ATP and the p -loop motifs. Mg^{2+} generally interacts with β - and γ -phosphates of ATP. For the class II aminoacyl-tRNA synthetase mode, Mg^{2+} and Mn^{2+} are involved in 3 ATP-protein interactions respectively, which occupy $6/11 = 55.6\%$ of the structures in this mode. Typically, 3 metal ions of the same type interact with the β - and γ -phosphates of ATP. Metal ions are also frequently observed at the protein kinase motifs,

i.e., 7 out of 9 protein–ATP interactions that employ the protein kinase binding mode are mediated by Mg^{2+} or Mn^{2+} . The metal ions generally interact with two phosphates simultaneously, which is likely to stabilize the interaction between ATP and proteins. For the strand sandwich mode, metal ions are involved in 2 out of 7 structures. 2 Mg^{2+} interact with phosphates of ATP for each of the binding sites. As metal ions are not observed for majority of the binding sites of this mode, metal ions may not play a crucial role as in other modes. For the actin-like ATPase mode, Ca^{2+} are observed in two binding sites while metal ions are not involved in the other two ATP–protein interactions. The impact of metal ions for the interaction between ATP and the actin-like ATPase motif is hard to be concluded as this mode only includes 4 structures.

4. Discussion

We summarized five classes of sequence motifs that are crucial for the ATP interaction. The motifs differ, between the classes, in their primary sequence, secondary structure arrangement and three dimensional arrangements. This suggests that ATP recognizes proteins in multiple ways (modes). The *p*-loop motif recognizes ATP through the phosphates. Using the case study, we demonstrate that the class II AARS interacts not only with ATP and its analogs but also with NSS and YSA. These three ligands share identical adenosine group, see Fig. 3 in Supplement, which is linked to a sulfate in NSS and YSA, and to a phosphate in ATP; the remaining structure of these ligands differs. Binding of the three ligands to the same class II AARS binding site suggests that the binding site likely recognizes the ligands through the adenosine. This is supported by two recent studies in which another two adenosine-containing compounds were observed to bind to AARS at the ATP-binding site (Rock et al., 2007; Metlitskaya et al., 2006). The other three binding modes establish a number of contacts with both adenosine and phosphate groups. They include multiple sequence segments where some of the segments, i.e., glycine-rich segments, mainly interact with the phosphates and some other segments bind to the adenosine. This suggests that although these binding modes recognize both adenosine and phosphate groups, they include short motifs which target specific group types. This observation is supported by a recently released structure of serine/threonine-protein kinase haspin (PDB code: 3IQ7), which contains IGEGVFGVEV, VAIK and RDLHWGNVL segments that are similar to the protein kinase motif. The structure includes a phosphate and adenosine-like ligand, (2R,3R,4S,5R)-2-(4-amino-5-iodo-7H-pyrrolo[2,3-D]pyrimidin-7-yl)-5-(hydroxymethyl) tetra hydrofuran-3,4-diol (5ID), at the site defined by these segments. Adenosine also binds to the ATP-binding site of *a.fulgidus* Rio1 serine protein kinase (Hvorup et al., 2007) (PDB code: 1ZTF). These structures demonstrate that the phosphate and adenosine groups could individually bind to the protein kinase motifs.

Analysis of the five binding modes reveals that Gly occurs 14 times among the 49 residues that are found to interact with ATP. The Gly residues are usually clustered together in the sequence, i.e., the binding motifs of the first, second, third and fourth modes include GXG tripeptide and the fifth mode incorporates GTG and GDG tripeptides. This suggests that ATP tends to bind to glycine-rich regions. The glycine residues that form hydrogen bonds with the phosphates of ATP are crucial for ATP-binding. For instance, a mutation of Gly residues at ATP-binding motifs of heat shock protein 90 (Hsp90) was shown to diminish ATP-binding (Grenert et al., 1997). Another study reported that mutations from Gly to Ser at the glycine-rich loop in the ATP-binding site of protein kinase result in a substantial decrease in enzymatic activity and thermal stability (Hemmer et al., 1997).

Three of the motifs are associated with a specific chemical reaction and/or biological function. The class II AARS motif is

involved in the formation of aminoacyl-adenylates which are used for translation of nucleotide sequence to AA sequence. The protein kinase motif is associated with phosphorylation by which the γ -phosphate is transferred to another molecule. The actin-like ATPase motif is specific for ATP hydrolysis that either leads to the conformational changes of Hsp70 or supplies the energy for actin activity (Flaherty et al., 1994; Nolen et al., 2004). In contrast, the *p*-loop and the strand sandwich motifs are involved in a diverse range of chemical reactions and biological processes. The *p*-loop motif is directly involved in phosphorylation of gluconate, shikimate and thymidylate. Through the interaction with the ATP, conformational changes in the *p*-loop and adjacent residues are crucial for the activity of phosphoenolpyruvate carboxykinase, type I restriction-modification enzyme HSDR (ATP hydrolysis), and nitrogenase (Lapkouski et al., 2009; Sen et al., 2006). ATP-binding also leads to significant conformational changes in the structure of the strand sandwich motif. These changes are important for the activity of the carboxybiotin-carboxyl-carrier protein and the N^5 -Carboxyaminoimidazole ribonucleotide synthetase, and for the formylation of the glycylamide ribonucleotide (Thoden et al., 2008).

We inspected the topology of the overall structure of the domains that contain the proposed motifs. Using the SCOP (Andreeva et al., 2008) hierarchy we observe that the domains from each of the four binding modes, except the *p*-loop, belong to the same corresponding superfamily, see Table 1 in Supplement. The *p*-loop motif is found in both “*p*-loop containing nucleoside triphosphate hydrolases” and “PEP carboxykinase-like” superfamilies, see Table 1. Overall, the results show that proteins that share similar binding mode are usually confined to the same superfamily, although their sequences may share low similarity. This observation is consistent with a recently study by Kinjo and Nakamura (2009), who determined approximately 3000 well-defined structural motifs for ligand-binding sites and found that majority of these motifs concern a single family or superfamily.

We analyzed the enzyme commission (EC) numbers, which determine enzyme-catalyzed reactions, for proteins that share the same ATP-binding motif. Two motifs, the *p*-loop and strand sandwich, are shared by different types of enzymes. The *p*-loop motif is found in oxidoreductases, transferases, hydrolases and lyases, and the strand sandwich motifs appear in transferases, lyases and ligases. The class II AARS motifs only occur in ligases forming aminoacyl-tRNA designated as EC 6.1.1. The protein kinase motifs appear exclusively in protein kinases designated as EC 2.7. The actin-like ATPase domain proteins could not be analyzed since they lack the EC annotations. This result demonstrates that different types of enzymes may share similar binding sites for some ligands like ATP. Most importantly, even though some proteins share similar global structural topology and similar binding site for ATP, our analysis indicates that they may participate in different catalytic reactions.

The five modes proposed in this work define major interaction types that are shared by hundreds of proteins in the case of the *p*-loop and the protein kinase modes and dozens of proteins for each of the remaining three modes. The 5 motifs cover 36.5%, i.e., 50 out of the 137, of the considered protein–ATP complexes. A similar rate was observed among the predictions. Given that the entire PDB includes 6469 ATP-binding and 8413 nucleotides-binding proteins, as annotated in GOA, predictions using the five binding motifs generated 2134 ATP-binding and 2637 nucleotide-binding proteins, which corresponds to 33% and 31.3% coverage, respectively.

Funding

This research was supported by the National Natural Science Foundation of China (Grant no. 11201334), Science and Technology

Commission of Tianjin Municipality (Grant no. 12JCYBJC31900) to KC.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at <http://dx.doi.org/10.1016/j.compbiolchem.2015.04.008>.

References

- Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 17, 3389–3402.
- Andreeva, A., Howorth, D., Chandonia, J.M., Brenner, S.E., Hubbard, T.J., Chothia, C., Murzin, A.G., 2008. Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res.* 36, D419–425.
- Asenjo, A.B., Weinberg, Y., Sosa, H., 2006. Nucleotide binding and hydrolysis induces a disorder-order transition in the kinesin neck-linker region. *Nat. Struct. Mol. Biol.* 13, 648–654.
- Bairoch, A., Claverie, J.M., 1988. Sequence patterns in protein kinases. *Nature* 331, 22.
- Barker, W.C., Dayhoff, M.O., 1982. Viral src gene products are related to the catalytic chain of mammalian cAMP-dependent protein kinase. *Proc. Natl. Acad. Sci. U. S. A.* 79, 2836–2839.
- Barrell, D., Dimmer, E., Huntley, R.P., Binns, D., O'Donovan, C., Apweiler, R., 2009. The GOA database in 2009 – an integrated gene ontology annotation resource. *Nucleic Acids Res.* 37, D396–D403.
- Bazan, J.F., Bacon, K.B., Hardiman, G., Wang, W., Soo, K., Rossi, D., Greaves, D.R., Zlotnik, A., Schall, T.J., 1997. A new class of membrane-bound chemokine with a CX3C motif. *Nature* 385, 640–644.
- Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E., 2000. The protein data bank. *Nucleic Acids Res.* 28, 235–242.
- Brakoulias, A., Jackson, R.M., 2004. Towards a structural classification of phosphate binding sites in protein-nucleotide complexes: an automated all-against-all structural comparison using geometric matching. *Proteins* 56, 250–260.
- Brenner, S., 1987. Phosphotransferase sequence homology. *Nature* 329, 21.
- Chen, K., Kurgan, L., 2009. Investigation of atomic level patterns in protein-small ligand interactions. *PLoS ONE* 4, e4473.
- Cusack, S., 1997. Aminoacyl-tRNA synthetases. *Curr. Opin. Struct. Biol.* 7, 881–889.
- Denessiouk, K.A., Rantanen, V.V., Johnson, M.S., 2001. Adenosine recognition: a motif present in ATP-, CoA-, NAD-, NADP-, and FAD-dependent proteins. *Proteins* 44, 282–291.
- Eriani, G., Delarue, M., Poch, O., Gangloff, J., Moras, D., 1990. Partition of tRNA synthetases into two classes based on mutually exclusive sets of sequence motifs. *Nature* 347, 203–206.
- Flaherty, K.M., Wilbanks, S.M., DeLuca-Flaherty, C., McKay, D.B., 1994. Structural basis of the 70-kilodalton heat shock cognate protein ATP hydrolytic activity: II. Structure of the active site with ADP or ATP bound to wild type and mutant ATPase fragment. *J. Biol. Chem.* 269, 12899–12907.
- Grenert, J.P., Sullivan, W.P., Fadden, P., Haystead, T.A., Clark, J., Mimnaugh, E., Krutzsch, H., Ochel, H.J., Schulte, T.W., Sausville, E., et al., 1997. The amino-terminal domain of heat shock protein 90 (hsp90) that binds geldanamycin is an ATP/ADP switch domain that regulates hsp90 conformation. *J. Biol. Chem.* 272, 23843–23850.
- Heery, D.M., Kalkhoven, E., Hoare, S., Parker, M.G., 1997. A signature motif in transcriptional co-activators mediates binding to nuclear receptors. *Nature* 387, 733–736.
- Hemmer, W., McGlone, M., Tsigelny, I., Taylor, S.S., 1997. Role of the glycine triad in the ATP-binding site of cAMP-dependent protein kinase. *J. Biol. Chem.* 272, 16946–16954.
- Hibi, T., Nishioka, T., Kato, H., Tanizawa, K., Fukui, T., Katsube, Y., Oda, J., 1996. Structure of the multifunctional loops in the nonclassical ATP-binding fold of glutathione synthetase. *Nat. Struct. Biol.* 3, 16–18.
- Hu, Z.L., Gogol, E.P., Lutkenhaus, J., 2002. Dynamic assembly of MinD on phospholipid vesicles regulated by ATP and MinE. *Proc. Natl. Acad. Sci. U. S. A.* 99, 6761–6766.
- Huang, Y., Niu, B., Gao, Y., Fu, L., Li, W., 2010. CD-HIT suite: a web server for clustering and comparing biological sequences. *Bioinformatics* 26, 680–682.
- Hulo, N., Bairoch, A., Bulliard, V., Cerutti, L., Cuče, B.A., de Castro, E., Lachaize, C., Langendijk-Genevaux, P.S., Sigrist, C.J., 2008. The 20 years of PROSITE. *Nucleic Acids Res.* 36, D245–9.
- Hvorup, R.N., Goetz, B.A., Niederer, M., Hollenstein, K., Perozo, E., Locher, K.P., 2007. Asymmetry in the structure of the ABC transporter-binding protein complex BtuCD-BtuF. *Science* 317, 1387–1390.
- Iwasaki, W., Sekine, S., Kuroishi, C., Kuramitsu, S., Shirouzu, M., Yokoyama, S., 2006. Structural basis of the water-assisted asparagine recognition by asparaginyl-tRNA synthetase. *J. Mol. Biol.* 360, 329–342.
- Kadi, N., Oves-Costales, D., Barona-Gomez, F., Challis, G.L., 2007. A new family of ATP-dependent oligomerization-macrocyclization biocatalysts. *Nat. Chem. Biol.* 3, 652–656.
- Kinjo, A.R., Nakamura, H., 2009. Comprehensive structural classification of ligand-binding motifs in proteins. *Structure* 17, 234–246.
- Kobayashi, N., Go, N., 1997a. ATP binding proteins with different folds share a common ATP-binding structural motif. *Nat. Struct. Biol.* 4, 6–7.
- Kobayashi, N., Go, N., 1997b. A method to search for similar protein local structures at ligand binding sites and its application to adenine recognition. *Eur. Biophys. J.* 26, 135–144.
- Kotik-Kogan, O., Moor, N., Tworowski, D., Saffro, M., 2005. Structural basis for discrimination of γ -phenylalanine from γ -tyrosine by phenylalanyl-tRNA synthetase. *Structure* 13, 1799–1807.
- Lapkouski, M., Panjigar, S., Janscak, P., Smatanova, I.K., Carey, J., Ettrich, R., Csefalvay, E., 2009. Structure of the motor subunit of type I restriction-modification complex EcoR124I. *Nat. Struct. Mol. Biol.* 16, 94–95.
- Maxwell, A., Lawson, D.M., 2003. The ATP-binding site of type II topoisomerases as a target for antibacterial drugs. *Curr. Top Med. Chem.* 3, 283–303.
- McDonald, I.K., Thornton, J.M., 1994. Satisfying hydrogen bonding potential in proteins. *J. Mol. Biol.* 238, 777–793.
- Metlitskaya, A., Kazakov, T., Kommer, A., Pavlova, O., Praetorius-Ibba, M., Ibba, M., Krashennnikov, I., Kolb, V., Khmel, I., Severinov, K., 2006. Aspartyl-tRNA synthetase is the target of peptide nucleotide antibiotic Microcin C. *J. Biol. Chem.* 281, 18033–18042.
- Nolen, B.J., Littlefield, R.S., Pollard, T.D., 2004. Crystal structures of actin-related protein 2/3 complex with bound ATP or ADP. *Proc. Natl. Acad. Sci. U. S. A.* 101, 15627–15632.
- Petryszak, R., Kretschmann, E., Wieser, D., Apweiler, R., 2005. The predictive power of the CluSTR database. *Bioinformatics* 21, 3604–3609.
- Polo, S., Sigismund, S., Faretta, M., Guidi, M., Capua, M.R., Bossi, G., Chen, H., De Camilli, P., Di-Fiore, P.P., 2002. A single motif responsible for ubiquitin recognition and monoubiquitination in endocytic proteins. *Nature* 416, 451–455.
- Reyes, C.L., Rutenber, E., Walter, P., Stroud, R.M., 2007. X-ray structures of the signal recognition particle receptor reveal targeting cycle intermediates. *PLoS ONE* 2, e607.
- Rock, F.L., Mao, W., Yaremchuk, A., Tukalo, M., Crépin, T., Zhou, H., Zhang, Y.K., Hernandez, V., Akama, T., Baker, S.J., et al., 2007. An antifungal agent inhibits an aminoacyl-tRNA synthetase by trapping tRNA in the editing site. *Science* 316, 1759–1761.
- Saraste, M., Sibbald, P.R., Wittinghofer, A., 1990. The p -loop – a common motif in ATP-binding and GTP-binding proteins. *Trends Biochem. Sci.* 15, 430–434.
- Scheeff, E.D., Bourne, P.E., 2005. Structural evolution of the protein kinase-like superfamily. *PLoS Comput. Biol.* 1, e49.
- Schwartz, D., Gygi, S.P., 2005. An iterative statistical approach to the identification of protein phosphorylation motifs from large-scale data sets. *Nat. Biotechnol.* 23, 1391–1398.
- Sen, S., Krishnakumar, A., McClelland, J., Johnson, M.K., Seefeldt, L.C., Szilagy, R.K., Peters, J.W., 2006. Insights into the role of nucleotide-dependent conformational change in nitrogenase catalysis: structural characterization of the nitrogenase Fe protein Leu127 deletion variant with bound MgATP. *J. Inorg. Biochem.* 100, 1041–1052.
- Senes, A., Gerstein, M., Engelman, D.M., 2000. Statistical analysis of amino acid patterns in transmembrane helices: the GxxxG motif occurs frequently and in association with beta-branched residues at neighboring positions. *J. Mol. Biol.* 296, 921–936.
- Thoden, J.B., Blanchard, C.Z., Holden, H.M., Waldrop, G.L., 2000. Movement of the biotin carboxylase B-domain as a result of ATP-binding. *J. Biol. Chem.* 275, 16183–16190.
- Thoden, J.B., Holden, H.M., Firestone, S.M., 2008. Structural analysis of the active site geometry of N^5 -carboxyaminoimidazole ribonucleotide synthetase from *Escherichia coli*. *Biochemistry* 47, 13346–13353.
- Walker, J.E., Saraste, M., Runswick, M.J., Gay, N.J., 1982. Distantly related sequences in the α - and β -subunits of ATP synthase, myosin, kinases and other ATP-requiring enzymes and a common nucleotide binding fold. *EMBO J.* 1, 945–951.
- Word, J.M., Lovell, S.C., Richardson, J.S., Richardson, D.C., 1999. Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. *J. Mol. Biol.* 285, 1735–1747.
- Xiang, S., Tong, L., 2008. Crystal structures of human and *Staphylococcus aureus* pyruvate carboxylase and molecular insights into the carboxyltransfer reaction. *Nat. Struct. Mol. Biol.* 15, 295–302.