ATP-binding site as a further application of neural networks to Residue Level Prediction

Shandar Ahmad and Zulfiqar Ahmad

Abstract— Similar neural network models based on single sequence and evolutionary profiles of residues have been successfully used in the past for predicting secondary structure, solvent accessibility, protein-, DNA- and carbohydrate- binding sites. ATP is a ubiquitous ligand in all living-systems, involved in most biological functions requiring energy and charge transfer. Prediction of ATP-binding site from single sequences and their evolutionary profiles at a high throughput rate can be used at genomic level as well as quick clues for site-directed mutagenesis experiments. We have developed a method for such predictions to demonstrate yet another application of sequence-base prediction algorithms using neural networks. This method can achieve 81% sensitivity and 69% specificity which are mutually adjustable in a wide range on a three-fold cross-validation data set.

I. Introduction

Adenosine 5'-triphosphate (ATP) is the universal energy currency, paramount for all forms of life from bacteria to mankind. A typical 70 kg human with relatively sedentary lifestyle generates around 2.0 million kg of ATP from ADP and Pi in a 75-year lifespan [1,2].

Both the synthesis and hydrolysis of ATP at a molecular level are of great importance for understanding the enzymatic mechanism and for drug design. ATP binding is associated directly or indirectly with many diseases including a class of severely debilitating diseases known collectively as mitochondrial myopathies.

Molecular modulation of the ATP binding sites for the purpose of better catalytic activity is a routine in many laboratories [3-6]. Incremental introduction of charged residues in catalytic sites of E. coli ATP synthase was shown to improve the catalytic activity by ten times [4,] while negatively charged residues were found to abrogate the Pi binding in the same catalytic site [7, 8]. Knowledge of specific amino acids involved in ATP binding plays crucial role in this direction. One of the easiest ways is to have sequence based information on ATP binding and hence a prediction method using only sequence inputs will be very useful.

Many structure and functional properties such as binding sites defined at a single residue level have been recently predicted from single amino acid sequence and their evolutionary profiles with varied degrees of accuracy [9-12], using neural networks and other machine learning algorithms. ATP-binding sites, thus happen to be another case of residue-level function of proteins, for which a neural network can be successfully applied. Here we make first attempt to achieve this. Our prediction method can achieve an adjustable sensitivity and specificity of prediction such the average of these scores called net prediction reaches 75%.

II. Material and Methods

Data sets:

There were 340 entries in Protein Data Bank (PDB) as updated in October 2007, which have ATP ligand in their coordinate file. When protein chains were separated 983 proteins were obtained from them, many of which are redundant and solved at a poor resolution. 546 protein chains were obtained at a resolution cutoff set at 2.5Å, of which 499 had 30 or more residues in their primary sequence. These protein chains were clustered using BLASTCLUST program at 25% similarity cutoff and 140 clusters of proteins each representing a unique family/ group were obtained. Many of these protein chains did not contact ATP at all, so chains with less than 10 residues in contact with ATP were removed, leaving a final set of 65 protein chains meeting all the above criterion of selection. List of PDB codes for the selected proteins including chain name as the fifth letter in the code, is shown in Table 1.

Table 1. List of representative ATP-binding protein chains.

PDB	Name	Class	Family	Architecture	Source
1A82A	Dethiobiotin Synthetase	$\alpha\beta$	P-loop containing	3-layer $\alpha\beta\alpha$ Sandwich	E. coli
1B8AA	Aspartyl-tRNA Synthetase	β	Nucleic acid binding	β Barrel	P.kodakaraensis
1CSNA	Casein Kinase-1	$\alpha\beta$	Protein Kinase	2-layer Sandwich	S. pombe
1DV2A	Biotin Caboxylase Mutant E288K	$\alpha\beta$	Rudiment single	3-layer $\alpha\beta\alpha$ hybrid motif sandwich	E. coli
1DY3A	Hydroxymethylpterin	$\alpha\beta$	HPPK pyrophosphokinase	2-layer sandwich	E. coli
1E24A	Lysyl-tRNA Synthetase	β	Nucleic acid binding	β Barrel	E. coli
1E2QA	Thymidylate Kinase	$\alpha\beta$	P-loop containing	3-layer $\alpha\beta$ sandwich	H. sapiens

Manuscript received on December 14, 2007.

S. A. (Corresponding Author) is with the National Institute of Biomedical Innovation, Saito-Asagi, Ibaraki, Osaka, Japan phone: 81-72-641-9848 ; e-mail: shandar@ netasa.org).

Z. A. Author, Department of Biological Sciences, East Tennessee State University, Johnson City TN 37614, USA

978-1-4244-1821-3/08/$25.00©2008 IEEE

1E8XA	Phosphoinositide 3-kinase	α	ARP repeat PI3K	Roll	S. scrofa
1EE1A	NH3 dependent NAD+ Synthetase	αβ	Adenine nucleotide α hydrolases –like	3-layer αβα sandwich	B. subtilis
1F2UA	RAD-50 ABC-ATPase	αβ	P-loop containing	NA	P. furiosus
1F9AA	Hypothetical Protein (MJ0541)	αβ	Nucleotidylyl transferase	3-layer αβα sandwich	M. jannaschii
1FMWA	Myosin II	β	Myosin S1 Fragment	NA	D. discoideum
1G64A	Adenosyltransferase	αβ	P-loop Containing	transferase sandwich	S. typhimurium
1G64B	Adenosyltransferase	αβ	P-loop Containing	3-layer αβα sandwich	S. typhimurium
1GN8A	Phophopantetheine Adenosyltransferase	αβ	Nucleotidylyl transferase	3-layer αβα sandwich	E. coli
1GZ3A	NAD-dependent Malic enzyme	αβ	NAD (P)-binding	Rossamann fold	H. sapiens
1I7LA	Synapsin II	αβ	Pre ATP Grasp domain	2-layer sandwich	R. norvegicus
1II0B	Arsenite-Translocating ATPase	αβ	P-loop containing sandwich	3-layer αβα	E. coli
1J1ZA	Argininosuccinate Stythetase	αβ	Adenine nucleotide α hydrolases –like	3-layer αβα sandwich	T. thermophilus
1J7K	DNA helicase RUVB Mutant (P216G)	NA	NA	NA	T. maritima
1JJVA	Dephospho-COA Kinase	αβ	P-loop containing	3-layer αβα sandwich	H. influenzae
1KJ9B	Glycinamide ribo Nucleotide transformylase	β	Rudiment single hybrid motif	3-layer αβα sandwich	E. coli
1KO5A	Gluconate Kinase	αβ	P-loop containing	3-layer αβα sandwich	E. coli
1KP3A	Argininosuccinate Stythetase	αβ	Adenine nucleotide α hydrolases –like	3-layer αβα sandwich	E. coli
1KP8A	GroEL	NA	NA	NA	E. coli
1KVKA	Mevalonate kinase	αβ	Ribosomal protein S5 domain-like	2-layer sandwich	R. norvegicus
1MB9B	β-Lactam Synthetase	αβ	Adenine nucleotide	4-layer α hydrolases –like sandwich	S. clavuligerus
1MJHB	Hypothetical Protein (MJ0577)	NA	NA	NA	M. jannaschii
1NGFA	HSP-70	αβ	Actin-like ATPase domain	2-layer sandwich	B. taurus
1NSFA	N-ethylmaleimide Sensitive factor	αβ	P-loop containing	3-layer αβα sandwich	C. griseus
1OS1A	Phosphoenol pyruvate Carboxykinse	αβ	PEP Carboxykinse like	3-layer αβα sandwich	E. coli
1QHXA	Chloramphenicol Phosphotransferse	αβ	P-loop containing	3-layer αβα sandwich	S. venezuelae
1R8BA	t-RNA nucleotidyl transferase	α	PAP/OAS1 substrate binding domain	\ NA	A. fulgidus
1S9JA	Mitogen activated Protein kinase 1	αβ	Protein kinase	2-layer sandwich	H. sapiens
1SU2A	Nudix Hydrolyase DR1025	αβ	Nudix	αβ complex	D. radiodurans
1SVMC	T-antigen Helicase	αβ	P-loop containing	3-layer αβα sandwich	S. virus 40
1TIDA	anti-sigma factor SpoIIA	αβ	ATPase domain HSP90	2-layer sandwich	B. sstearothermophilu
1U5RA	TAO2 kinase Domain	αβ	Protein kinase	NA	R. norvegicus
1W7AA	DNA mismatch Repair protein muts	α	DNA repair protein MutS	NA	Synthetic construct
1X01A	Biotin protein Ligase	NA	NA	NA	P. horikoshii
1XDNA	RNA editing Ligase 1	αβ	DNA ligase	2-layer sandwich	T. brucei
1XEXA	SMC ATPase	αβ	P-loop containing	NA	P. furiosus
1XNGB	NH3 dependent NAD+ Synthetase	NA	NA	NA	H. pylori
1Y8QD	Sumo E1 Activating enzyme	αβ	Ubiquitin- like	3- layer αβα sandwich	H. sapiens
1YAGA	Yeast actin Human gelsolin	αβ	Actin- like ATPase domain	2-layer sandwich	S. cerevisiae
1YUNA	Nicotinate-nucleotide Adenylyltransferase	NA	NA	NA	P. aeruginosa
1Z0SA	NAD kinase	αβ	NA	3- layer αβα sandwich	A. fulgidus
1ZAOA	Rio2 kinase	NA	NA	NA	A. fulgidus
2A84A	Pantothenate synthetase	αβ	NA	3- layer αβα sandwich	M. tuberculosis
2AQXA	Inositol 1, 4, 5-Tri Phosphate 3-kinase B	NA	NA	NA	M. musculus
2ARUA	Lipoate-Protein ligase A	NA	NA	NA	T. acidophilum
2BEKA	Segregation protein	NA	NA	NA	T. thermophilus
2BU2A	Pyruvate dehydrogenase kinase 2	NA	NA	NA	H. sapiens
2C8VA	Nitrogenase Iron protein	αβ	NA	3- layer αβα sandwich	A. vinelandii
2FAQA	LigD polymerase Domain	NA	NA	NA	P. aeruginosa
2HMUA	Yua protein RCK domain	αβ	NA	3- layer αβα sandwich	B. subtilis
2IVPA	UP1 protein	NA	NA	NA	P. abyssi
2J3MB	Prolyl-tRNA Synthetasae	NA	NA	NA	E. faecalis

2J9EB	GLNK1 (MJ0059) Hypothetical	NA	NA	NA	M. jannaschii
2OGXA	Molybdenum storage protein	NA	NA	NA	A. vinelandii
2Q0DA	Uridylyl Transferase	NA	NA	NA	T. brucei
2QRDG	Adenylate Sensor	NA	NA	NA	S. pombe
2QXLB	Yeast Hsp110	NA	NA	NA	S. cerevisiae
2YWWB	Aspartate Carbamoyltransferase	NA	NA	NA	M. jannaschii
2Z02A	phosphoribosylamino XX	NA	Imidazolesuccinocarboxamide Synthase	NA	M. jannaschii

Definition of binding site:

A residue is defined to be ATP-binding if any of its atoms comes within a cutoff distance from an ATP atom. For all the calculations presented in this work, cutoff is taken as 3.5Å. Distance calculations were done using in-house program, developed by one of us. This definition is similar to our other works and is found to work better in those examples [13].

Propensity:

General residue preference to bind ATP molecules can be estimated from the statistics of their contacts and calculating their propensity scores [13]. ATP-binding propensity (or simply propensity) of a residue type (e.g. Ala) has been defined as the proportion of ATP-binding residues of that type relative to overall proportion of ATP-binding residues. Thus, the propensity P(i) for residue i is given as

$$P(i) = [N(ib)/N(i)] / [N(0b)/N(0)] \qquad (1)$$

Where N(ib) is the number of binding residues of type i, and N(i) is the total number of residues of this type. N(0b) refers to binding residues of all type and N(0) is the total number of residues considered. Error bars were estimated by creating random lists of proteins from the data set, pooling binding and non-binding data for a list at a time, calculating propensity for each list and finding their mean and standard deviation values.

Neural network:

A multi-layered feed-forward neural network, similar to our previous works has been used. Input layer consists of rows from evolutionary profile matrix i.e. Position Specific Scoring Matrix (PSSM) corresponding to a target residue and its neighbors. PSSM was generate using 3 iterations of PSI BLAST program [14]. No transformation of these numerical values was carried out. Number of neighbors was systematically changed from one to four. Four-neighbor information raised the number of trainable parameters to 212 weights and 108 biases. Due to a limited amount of available data higher number of neighbors was not tried. Network consisted of one hidden layer with two units and transfer functions were arctan and sigmoid for hidden and output layer respectively. Training was performed by optimizing coefficient of correlation between desired binding state (0 or 1) and predicted neural network output for each pattern (residue) using steepest descent method. Training by steepest descent was preferred because error cannot be explicitly defined in a positive performance score like correlation coefficient and therefore could not be back propagated. Prediction scheme is shown in Figure 1.

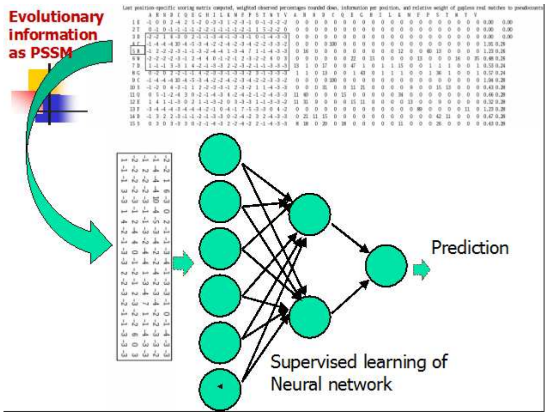
Evolutionary information as PSSM
Prediction
Supervised learning of Neural network

Figure 1: Prediction of ATP Binding site using PSSM rows as neural network input.

.

Performance measurement:

Performance score throughout the training was the Matthews correlation of coefficient defined as a ratio of covariance between predicted and desired values of binding state to the product of variances in each of the two.

III. RESULTS AND DISCUSSION

Residue propensity scores

Prior to embarking upon a neural network based prediction, we computed the propensity scores of all 20 types of amino acid residues in our database of ATP-binding proteins. To have a fair idea of propensities ATP-binding residues were compared with other propensity scores, from our published works on DNA-binding, sugar-binding and binding to other ligands [9,13] (Cys and Val have been left out due to a small number in contact with all ligands). Certain observations are made from these graphs as follows:

1. Smallest amino acid residue Gly has the highest propensity (2.1) to interact with ATP, with His and Ser coming close second (1.8) and third (1.8) respectively. Large propensity of Gly suggests a non-specific nature of ATP interactions and importance of backbone interactions. Backbone accessibility of residues to ATP molecules may be an important factor determining interaction between these molecules.

2. Positively charged residues Arg (prop=1.5) and Lys (prop=1.7) also have a high propensity, indicating that positively charged residues may have a greater possibility to interact with ATP. It may be noted that DNA-binding proteins, well known to interact through their Arg and Lys residues do not show a high propensity for His, which is observed in ATP-binding sites. This anomaly i. e. a similarity with Arg and Lys and difference from His residues may be either due to different oxidation states of His in the two cases or due to structural requirements of DNA, which may be absent in case of ATP, the later being a small molecules.

3. ATP shows smallest propensity for Trp (prop=0.5) residues in contrast to sugars (prop> 3.0), other ligands (prop> 2.0) and DNA (prop close to 1, almost neutral). This could be partly because Trp is typically a buried residue and only those Trp side chains which can stack over Adenine ring will likely interact with ATP. This is however a tentative conclusion and is being further examined.

4. Acidic residues Asp and Glu (prop= 1.3 and 0.8 respectively) behave quite differently with a large difference in their propensity, suggesting that their interaction with ATP is not determined by their acidic nature- in which they are similar- but other structural or chemical considerations may be at play.

5. Other hydrophobic residues such as Ile, Leu, Pro, Gln, and Thr show almost similar propensities in most ligands except Asn, which shows N-linked contacts with sugars and therefore has large propensity for sugar-binding residues.

6. Overall variation between propensities of residues is less in ATP compared with other ligands including DNA-binding sites. Higher propensity of Gly, supported by this observation, only indicates that there are strong backbone interactions of residues with ATP, which allow all accessible atoms of the proteins to interact and side chain atoms confer additional specificity to these backbone interactions.

Role of neighbors

As stated above, propensity scores of 20 amino acid residues in ATP-binding regions are less diverse than their counterparts in other ligand-binding sites. However, ATP-binding to proteins typically occurs at small specific regions as we observed from their complexed structures, which means ATP molecules are well recognized by proteins in their specific sites. How do the ATP molecules find their partner locations on proteins? Two possibilities remain: one is that ATP-binding occurs in cooperation with several residues coming together to create a right environment within a protein sequence and second is the possibility that the local and three-dimensional structure of the protein forms exactly the right kind of geometry to support ATP molecule. In this work, we are trying to predict ATP-binding sites from sequence only and therefore structure aspect will not be covered here. For the amino acid sequence, we may obtain information about the sequence and evolutionary context of each residue by constructing their alignment profile and using them for prediction with a non-linear model (neural network in our case).

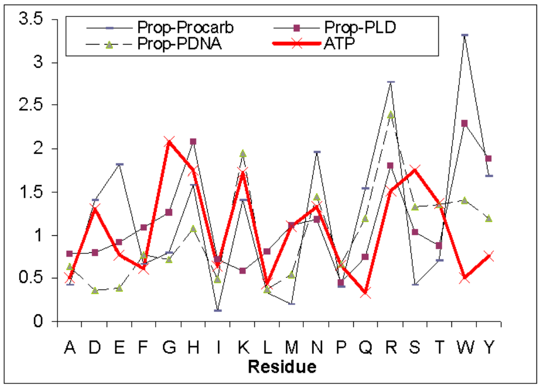

Figure 2: Propensity scores of residues in ATP-binding sites compared with DNA-, Carbohydrate and other Ligand-binding sites. PLD, PDNA and Procarb stand for all ligands in PLD database, DNA and carbohydrate binding.

We trained a neural network with single hidden layer of two nodes to develop a prediction system for ATP binding sites. Neural network uses alignment profile of the target residue and its four N-terminal and C-terminal residues to describe its environment. Training the neural network on 2/3rd data and using the remaining 1/3rd data for evaluating prediction performance, we obtained three sets of performance scores. Typical prediction is in terms of a real number between 0 and 1, which can be transformed to class prediction of two states (binding and non-binding) by taking different thresholds such that the neural network outputs higher than that represent a binding target residues and lower represent a non-binding residue. By changing the threshold value, sensitivity and specificity of prediction may be modulated, such that a higher cutoff will pick up fewer binding sites but each with a higher confidence, whereas lower thresholds will result in more false binding sites. Thus results from prediction using different thresholds may be plotted as a graph-called ROC- between sensitivity and specificity and the best performance comes when both of these scores are at the highest. This is measured by taking the average of the two. It may be noted that taking the average of sensitivity and specificity ensures that the reported scores give equal weight to the prediction of both positive and negative class. Similar sores such as F1 score, a harmonic mean of sensitivity and precision could also serve the same purpose but we retained current scores partly to be consistent with our other works on similar problems, and in part as we observed that F1 score did not enhance our performance evaluation strategy. Figure 3 shows a typical ROC graph for our prediction system. As shown in the figure best performance was reached at a high sensitivity of 81% at which neural network could give a specificity of 69%. Thus a large number of ATP-binding residues could be identified by this method. Further work is in progress to examine specific roles of neighbors and making the predictions available through a web server.

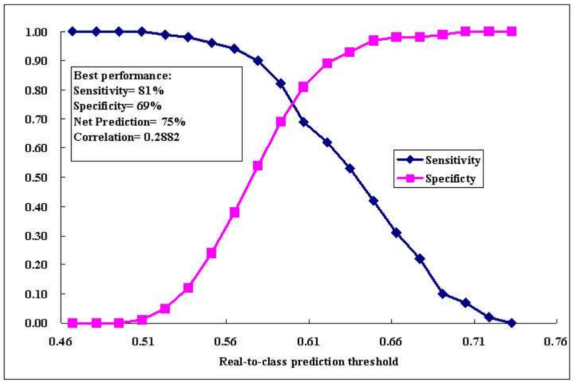

Figure 3: Sensitivity versus specificity of ATP-binding prediction using PSSM rows of target residue and four neighbors. Total amino acid composition of the protein is also used as a descriptor.

IV. Conclusions

ATP-binding sites can be predicted from evolutionary profiles of proteins at target locations and sequence neighbors. Best performance scores reach near 75% average of sensitivity and specificity, mutual balance between these scores can be adjusted from the real-valued prediction output.

References

[1] Senior, A.E., Nadanaciva, S., and Weber, J. (2002) The molecular mechanism of ATP synthesis by F1Fo-ATP synthase. Biochim. Biophys. Acta 1553, 188-211.

[2] Weber, J., and Senior, A.E. (2003) ATP synthesis driven by proton transport in F1Fo- ATP synthase. FEBS Lett. 545, 61-70.

[3] Ahmad, Z., and Senior, A.E. (2005) Identification of phosphate binding residues of Escherichia coli ATP synthase. J. Bioenerg. Biomembr. 37, 437-440.

[4] Ahmad, Z., and Senior, A.E. (2005) Modulation of charge in the phosphate- binding site of Escherichia coli ATP synthase. J. Biol. Chem. 280, 27981-27989.

[5] Ahmad, Z., and Senior, A.E. (2004) Mutagenesisn of residue βArg -246 in the phosphate- binding subdomain of catalytic sites of Escherichia coli F1-ATPase. J. Biol. Chem. 279, 31505-31513.

[6] Ahmad, Z., and Senior, A.E. (2005) Involvement of ATP synthase residues αArg -376, βArg -182, and βLys -155 in Pi binding FEBS Lett. 579, 523-528.

[7] Ahmad, Z., and Senior, A.E. (2004) Role of βAsn -243 in the phosphate- binding subdomain of catalytic sites of Escherichia coli F1-ATPase. J. Biol. Chem. 279, 4607-46064.

[8] Laura, B., Grindstaff J. and Ahmad Z. (unpublished results)

[9] Ahmad S, Sarai A. PSSM-based prediction of DNA binding sites in proteins. BMC Bioinformatics. 2005, 6:33.

[10] Tjong H, Zhou HX, DISPLAR: an accurate method for predicting DNA-binding sites on protein surfaces. Nucleic Acids Res. 2007;35(5):1465-77

[11] Hwang S, Gou Z, Kuznetsov IB, DP-Bind: a web server for sequence-based prediction of DNA-binding residues in DNA-binding proteins,Bioinformatics.2007 Mar 1;23(5):634-6.

[12] Ofran Y, Mysore V, Rost B., Prediction of DNA-binding residues from sequence. Bioinformatics. 2007 Jul 1;23(13):i347-53.

[13] Malik A and Ahmad S, Sequence and structural features of carbohydrate binding in proteins and assessment of predictability using a neural network, BMC Structural Biology 2007, 7:1

[14] Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ., Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res. 1997 Sep 1;25(17):3389-402.