SCIENTIFIC REPORTS

OPEN

SUBJECT AREAS:
SEQUENCE ANNOTATION
DATA MINING

Received
17 March 2014

Accepted
18 August 2014

Published
25 September 2014

Correspondence and requests for materials should be addressed to C.L. (carmay@gate.sinica.edu.tw)

Sequence-motif Detection of NAD(P)-binding Proteins: Discovery of a Unique Antibacterial Drug Target

Yun Hao Hua[1], Chih Yuan Wu[1], Karen Sargsyan[1] & Carmay Lim[1,2]

[1]Institute of Biomedical Sciences, Academia Sinica, Taipei 115, Taiwan, [2]Department of Chemistry, National Tsing Hua University, Hsinchu 300, Taiwan.

Many enzymes use nicotinamide adenine dinucleotide or nicotinamide adenine dinucleotide phosphate (NAD(P)) as essential coenzymes. These enzymes often do not share significant sequence identity and cannot be easily detected by sequence homology. Previously, we determined all distinct locally conserved pyrophosphate-binding structures (3d motifs) from NAD(P)-bound protein structures, from which 1d sequence motifs were derived. Here, we aim to establish the precision of these 3d and 1d motifs to annotate NAD(P)-binding proteins. We show that the pyrophosphate-binding 3d motifs are characteristic of NAD(P)-binding proteins, as they are rarely found in nonNAD(P)-binding proteins. Furthermore, several 1d motifs could distinguish between proteins that bind only NAD and those that bind only NADP. They could also distinguish between NAD(P)-binding proteins from nonNAD(P)-binding ones. Interestingly, one of the pyrophosphate-binding 3d and corresponding 1d motifs was found only in enoyl-acyl carrier protein reductases, which are enzymes essential for bacterial fatty acid biosynthesis. This unique 3d motif serves as an attractive novel drug target, as it is conserved across many bacterial species and is not found in human proteins.

Nicotinamide adenine dinucleotide (NAD) and nicotinamide adenine dinucleotide phosphate (NADP), collectively referred to as NAD(P), are important coenzymes widely used in biochemical processes of living cells. Among the ~0.54 million sequences in the June 2013 update of the UniProtKB/Swiss-Prot database[1], ~5.4% proteins are annotated as binding NAD(P). The NAD(P)-binding enzymes are involved in catalyzing redox or nonredox reactions. Many of these enzymes are therapeutic drug targets; e.g., the ADP-ribosylating toxins[2] and polyADP-ribose polymerases[3,4]. However, NAD(P)-binding enzymes often do not share significant sequence identity and cannot be easily detected by sequence homology. Hence, 1d sequence motifs characteristic of NAD(P)-binding enzymes would be useful in predicting if a protein binds NAD(P)[5–11].

A few consensus sequences have been proposed for Rossmann-fold NAD(P)-binding proteins. Rossmann et al.[12] first found the phosphate-binding sequence **G**-X_{1-2}-**G**-X-X-**G** from an alignment of the sequences of dogfish lactate dehydrogenase, pig, lobster, and yeast glyceraldehyde-3-phosphate dehydrogenase, horse liver alcohol dehydrogenase, and bovine glutamate dehydrogenase. However, this phosphate-binding motif is relatively short and exceptions to this motif have been found[13], so it would not be a reliable signature for Rossmann-fold NAD(P)-binding proteins. Subsequently, Kleiger and Eisenberg[14] found **G**-X-X-X-[G/A] motifs following the phosphate-binding **G**-X_{1-2}-**G**-X-X-**G** motif in flavin adenine dinucleotide (FAD) and NAD(P)-binding Rossmann folds. They proposed an extended **G**-X_{1-2}-**G**-X-X-**G**-X-X-X-[G/A] motif as an indicator of Rossmann folds that bind FAD or NAD(P). However, recent analyses[11] showed that the fourth residue after the third conserved Gly in the **G**-X_{1-2}-**G**-X-X-**G** motif is not a conserved Gly or Ala but is variable. Using geometric matching to cluster phosphate-binding sites of Rossmann-fold proteins with similar 3D structure, Brakoulias and Jackson[15] found a variant of the **G**-X-**G**-X-X-**G** motif, namely, **G**-X-**G**-X-V-**G**, and a new **G**-X-X-X-**G**-I-**G** motif. Because 1d motifs with similar structures depend on the cofactor type (NAD or NADP) and on the side chain orientations[11], 1d motifs derived without consideration of the cofactor type and/or local similarity of both backbone and side chain structures would not be reliable in annotating protein function from sequence alone.

In our previous work[11], we presented a strategy to identify 1d motifs from a set of NAD(P)-binding proteins sharing little sequence identity, but having in common a locally conserved structure (3d motif) for a certain function. We found twelve distinct locally conserved structures for binding NAD(P) pyrophosphates consisting of a β-strand, followed by a turn/loop and a phosphate-binding α-helix. These pyrophosphate-binding βα

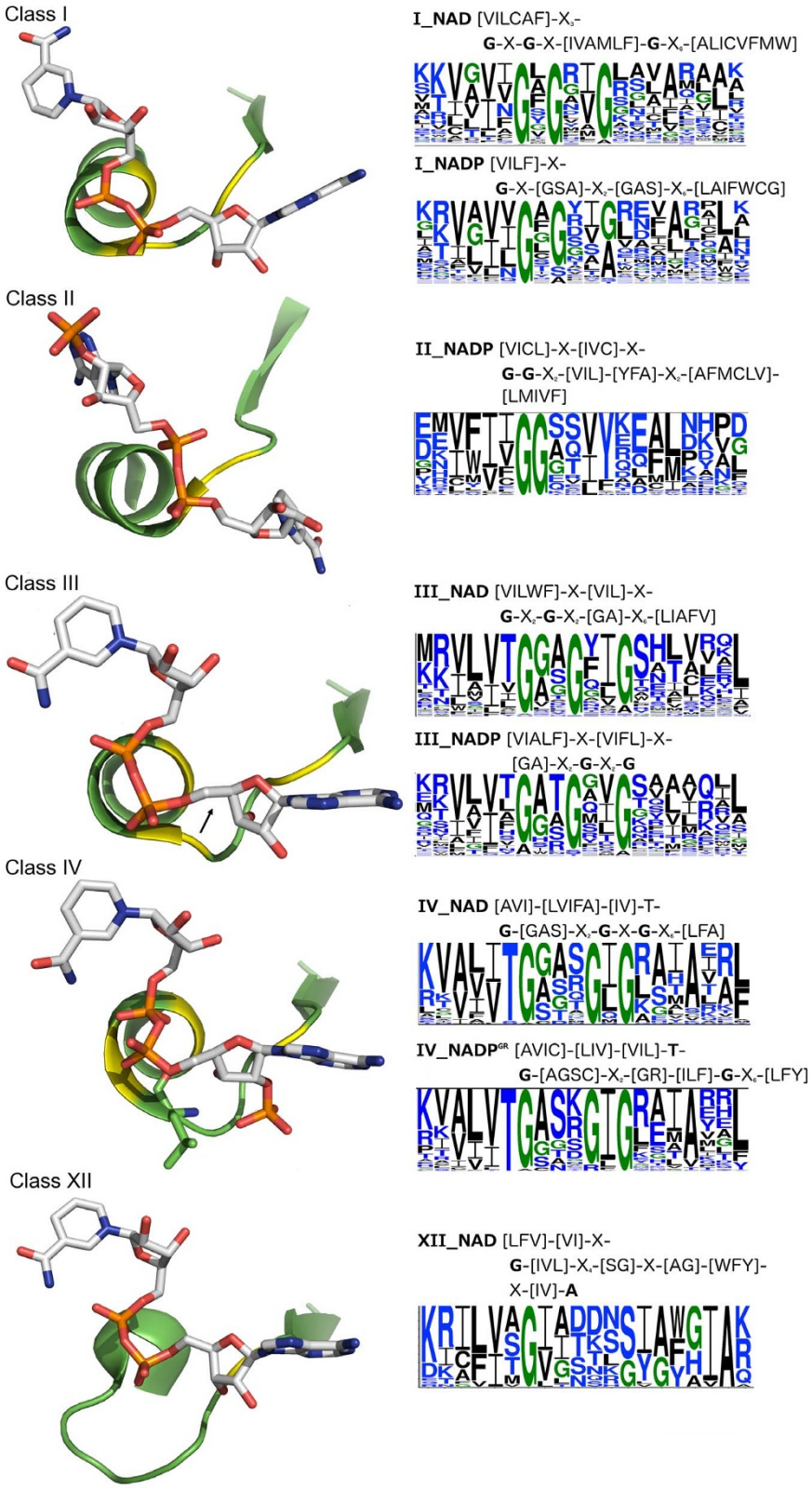

Figure 1 | **Derivation of 1d-motifs from distinct 3d-motifs.** (Left) The distinct locally conserved pyrophosphate-binding βα structures derived from NAD(P)-binding domains where the total number of βα structures exceeds 25. The βα structure is in green with the regions containing conserved glycines highlighted in yellow, while NAD(P) is shown in stick format. The class III and IV structures share a common backbone conformation but exhibit different side chain orientations: in the class IV structure (1zk4-A), the Leu side chain is shown in stick, but the corresponding side chain in the class III structure (1sby-A), indicated by the black arrow, point in an opposite direction. (Right) Sequence logos derived from aligning the same-length sequences comprising the distinct pyrophosphate-binding βα structures and corresponding 1d motif. Glycine is shown in green, polar (S, T, Y, N, Q, H, K, R, D, E) residues in blue, and nonpolar (A, V, L, I, P, W, F, C, M) residues in black.

Table 1 | Description of data sets employed

Redundant Dataset	# of Proteins	Description of dataset
3d-NAD(P)	1,096	Protein structures with NAD(P) bound
3d-FAD	348	Protein structures with FAD bound
3d-PO4	10,292	Protein structures with ≥1 phosphate group bound, excluding NAD(P) but including FAD
3d-nonPO4	33,514	Protein structures with no bound NAD(P) or phosphate
1d-NAD(P)	24,516	Sequences of NAD(P)-binding proteins
1d-NAD	15,340	Sequences of proteins that bind only NAD
1d-NADP	6,722	Sequences of proteins that bind only NADP
1d-nonNAD(P)	402,353	Sequences of nonNAD(P)-binding proteins
1d-FAD	949	Sequences of FAD-binding proteins
1d-PO4	131,165	Sequences in 1d-nonNAD(P) that bind ≥1 phosphate, including FAD-binding protein sequences
1d-nonPO4	271,188	Sequences in 1d-nonNAD(P) that do not bind phosphate

structures, labeled I,…,XII, are present in nearly three-quarters of the NAD(P)-binding domains in the Protein Data Bank (PDB)[16]. Sequence motifs were then derived from class I, II, III, IV, and XII structures, but not from the other 3d motifs, which do not have enough sequences (≤14) to generate statistically significant 1d motifs. The same-length sequences from NAD and NADP-bound structures comprising each pyrophosphate-binding structural class in Fig. 1 were aligned separately. For example, out of 105 structures with the class IV 3d motif, 45 contain NAD and 60 contain NADP; alignment of the 45 sequences from the NAD-bound structures with the class IV 3d motif yielded [AVI]-[LVIFA]-[IV]-**T**-**G**-[GAS]-X_2-**G**-X-**G**-X_6-[LFA], whereas alignment of the 60 sequences from the NADP-bound structures comprising the same 3d motif yielded [AVIC]-[LIV]-[VIL]-**T**-**G**-[AGSC]-X_2-[GR]-[ILF]-**G**-X_6-[LFF]. The consensus NAD(P)-binding sequences derived from the 3d motifs in Fig. 1 appear to be statistically significant, as they are found in ≤1.2% of randomized sequences (see Supplementary Table S1), except for the NADP-binding consensus sequences corresponding to structural class III (~3.6%) and class I (14%)[11]. However, the randomized sequences are not real biological sequences, therefore the potential of these NAD(P) 1d motifs to annotate NAD(P)-binding proteins remains unclear.

In this work, we address the following questions: (1) How often do the distinct pyrophosphate-binding 3d motifs in Fig. 1 occur in nonNAD(P)-binding proteins? (2) Since the 1d motifs in Fig. 1 were derived from either NAD or NADP-bound structures, can they distinguish between proteins that bind only NAD and those that bind only NADP? (3) Can the NAD(P) pyrophosphate-binding 1d motifs distinguish between NAD(P)-binding proteins and nonNAD(P)-binding ones? In particular, can they differentiate proteins that bind FAD, which is similar to NAD and also has a pyrophosphate group? Notably, we are interested in the precision (fraction of correctly predicted NAD(P)-binding proteins) of the motifs in Fig. 1. To address these questions, we created four datasets of 3d structures and seven datasets of 1d sequences (see Table 1). The results show that the 3d motifs in Fig. 1 are statistically significant, as they are rarely found in 3d structures of nonNAD(P)-binding proteins. Several 1d motifs could correctly distinguish between proteins that bind only NAD and those that bind only NADP. Furthermore, 1d motifs derived from class II, IV, and XII 3d motifs can be used to distinguish NAD(P)-binding proteins from nonNAD(P)-binding ones.

Table 2 | Frequency distribution of the NAD(P) pyrophosphate-binding 3d motifs in the PDB

	% frequency of structural class *j* in 3d dataset[a]				% PPV of 3d-NAD(P) vs.		
Class *j*	NAD(P)	FAD	PO_4	$nonPO_4$	FAD	PO_4	$nonPO_4$
I	24.6	36.2	2.6	0.2	68	51	80
II	2.2	0.3	0.05	0.01	96	83	89
III	12.9	0	0.1	0.04	100	92	92
IV	11.3	0	0	0	100	100	100
XII	1.6	0	0	0	100	100	100

[a]The number of structures in the given dataset containing the 3d motif belonging to class *j* divided by the total number of structures/proteins in the given dataset, multiplied by 100.

Results

Four pyrophosphate-binding 3d motifs are characteristic of NAD(P)-binding proteins. To assess if the distinct pyrophosphate-binding 3d motifs in Fig. 1 are characteristic of NAD(P)-binding proteins, we computed the occurrence frequency of a 3d motif corresponding to structural class *j* in ≤3.5 Å protein structures containing (1) NAD(P), (2) FAD, (3) phosphate-containing ligands including FAD, and (4) no NAD(P), FAD, or phosphate groups. For each of these 3d motifs, the percentage occurrence frequency in the NAD(P)-binding proteins is significantly greater than that in the NADP(P)-free proteins, except the class I 3d motif, which appears more often in FAD-binding proteins than in NAD(P)-binding ones (see Table 2). All the pyrophosphate-binding 3d motifs except the class I motif can distinguish NAD(P)-binding proteins from nonNAD(P)-binding proteins with positive predictive values (PPVs) ≥ 83%. Interestingly, the class IV and XII 3d motifs seem to be unique to NAD(P)-binding proteins, as they were not found in any of the NAD(P)-free structures. The class III 3d motif, which has a similar backbone structure as the class IV motif but different side chain orientations (see Fig. 1), is not found in any of the FAD structures and rarely in the other NAD(P)-free structures (PPV ~ 92%). The class I 3d motif, which occurs most frequently in NAD(P)-binding proteins, can differentiate NAD(P)-binding proteins from nonphosphate-binding ones (PPV ~ 80%), but not from proteins that bind phosphate-containing ligands (PPV ~ 51%).

Four pyrophosphate-binding 1d motifs can distinguish between NAD- and NADP binding proteins. Some of the 3d motifs in Fig. 1 appear to be NAD or NADP-specific; e.g., the class II 3d motif was found only in NADP-bound structures, while the class XII 3d motif was found predominantly in NAD-bound structures. Furthermore, the pyrophosphate-binding 1d motifs were derived from NAD and NADP-bound protein structures separately[11] (see Fig. 1). To determine if the pyrophosphate-binding 1d motifs can distinguish between NAD- and NADP-binding proteins, the % occurrence frequencies of the 1d motifs in the 1d-NAD and 1d-NADP datasets and PPVs were computed (see Table 3). Four of the 1d motifs can distinguish between NAD and NADP-binding proteins with PPVs ≥ 76%. Remarkably, the II_NADP motif derived from class II NADP-bound protein structures was not found in any of the NAD-binding proteins, whereas the XII_NAD motif derived from class XII NAD-bound protein structures was not found in the 1d-NADP dataset.

Table 3 | Precision of the 1d motifs to distinguish between NAD- and NADP-binding proteins

1d motif	Consensus sequence	NAD[a]	NADP[a]	%PPV
I_NAD	[VILCAF]-X_3-**G**-X-**G**-X-[IVAMLF]-**G**-X_6-[ALICVFMW]	18.5	6.5	82[b]
I_NADP	[VILF]-X-**G**-X-[GSA]-X_2-[GAS]-X_6-[LAIFWCG]	9.3	22.6	61[c]
II_NADP	[VICL]-X-[IVC]-X-**G**-**G**-X_2-[VIL]-[YFA]-X_2-[AFMCLV]-[LMIVF]	0	0.4	100[c]
III_NAD	[VILFW]-X-[VIL]-X-**G**-X_2-**G**-X_2-[GA]-X_6-[LIAFV]	2.4	7.5	34[b]
III_NADP	[VILFA]-X-[VILF]-X-[GA]-X_2-**G**-X_2-**G**	2.8	3.9	47[c]
IV_NAD	[AVI]-[LVIFA]-[IV]-**T**-**G**-[GAS]-X_2-**G**-X-**G**-X_6-[LFA]	0.5	2.3	26[b]
IV_NADP	[AVIC]-[LIV]-[VIL]-**T**-**G**-[AGSC]-X_2-[GR]-[ILF]-**G**-X_6-[LFY]	0.4	1.8	76[c]
XII_NAD	[LFV]-[VI]-X-**G**-[IVL]-X_4-[SG]-X-[AG]-[WFY]-X-[IV]-A	0.06	0	100[b]

[a]The number of protein sequences in the given dataset matching the 1d motif divided by the total number of sequences in the given dataset, multiplied by 100.
[b]The number of true positives is the number of NAD-binding sequences matching a 1d motif derived from NAD-bound structures, whereas the number of false positives is the number of NADP-binding sequences matching the same 1d motif.
[c]The number of true positives is the number of NADP-binding sequences matching a 1d motif derived from NADP-bound structures, whereas the number of false positives is the number of NAD-binding sequences matching the same 1d motif.

In contrast, the 1d motifs derived from class I and class III NADP-bound protein structures (I_NADP and III_NADP) as well as those derived from class III and class IV NAD-bound protein structures (III_NAD and IV_NAD) cannot distinguish between NADP- and NAD-binding proteins. The difference in specificity of the I_NAD and I_NADP motifs indicates that the presence of hydrophobic residues either four residues before the first conserved glycine (VILCAF) or preceding the third conserved glycine (IVAMLF), and/or the strict conservation of all three glycines, appear to be characteristic features of proteins with the class I 3d motif that bind only NAD. Along the same vein, the difference in specificity of the IV_NADP and IV_NAD motifs indicates that the allowance of arginine at the position of the second conserved glycine followed by hydrophobic residues; i.e., [GR]-[ILF], seems to be a signature of proteins with class IV 3d motif that bind only NADP.

1d motifs can distinguish between NAD(P)-binding and FAD-binding proteins. Since FAD is most similar to NAD, do the 1d motifs in Fig. 1 also bind the FAD pyrophosphate group? To answer this question, the 1d motifs were tested on the 1d-FAD dataset, which contains sequences from the UniProtKB/Swiss-Prot June 2013 database[1] with the ligand keyword FAD. Interestingly, although the pyrophosphate group is common to both FAD and NAD(P), the 1d motifs in Fig. 1 appear to recognize specifically the NAD(P) pyrophosphate with PPVs $\geq$ 96%, except for the 1d_NADP motif where the PPV is 84%. Notably, the 1d motifs derived from the class II, IV, and XII 3d motifs were not found in the 1d-FAD dataset.

1d motifs derived from class II, IV, and XII 3d motifs can distinguish between NAD(P)- and nonNAD(P)-binding proteins. To determine if the 1d motifs derived from NAD(P)-bound protein structures can distinguish between NAD(P) and nonNAD(P)-binding proteins, the % occurrence frequencies of the 1d motifs in the 1d-NAD(P), 1d-PO_4 (which include FAD-binding sequences), 1d-nonPO_4, and 1d-nonNAD(P) datasets were computed. Sequences in the 1d-PO_4 and 1d-nonPO_4 datasets comprise the 1d-nonNAD(P) dataset. The results in Table 4 show that although the number of NAD(P)-binding proteins is an order of magnitude less than the number of nonNAD(P)-binding proteins, the % occurrence frequencies of the 1d motifs in the 1d-NAD(P) dataset are significantly greater than those in the 1d-PO_4 or 1d-nonPO_4 dataset. Like the class IV and XII 3d motifs, the IV_NAD, IV_NADP, and XII_NAD motifs seem to be unique to NAD(P)-binding proteins, as they were not found in any of the nonNAD(P)-binding protein sequences. The II_NADP motif was also not found in nonphosphate-binding proteins (PPV = 100%) but do occur in phosphate-binding proteins, yielding a lower PPV of 79%. Like the II_NADP motif, the I_NAD and III_NAD motifs occur more often in the 1d-PO_4 dataset than in the 1d-nonPO_4 one, hence they can discern nonphosphate-binding proteins from NAD(P)-binding ones with a PPV of 84 and 78%, respectively.

Application of 3d and 1d motifs in human proteome annotation. All the 3d motifs in Fig. 1 (except class I), which could distinguish between NAD(P)- and nonphosphate-binding proteins with $\geq$90% PPV (see Table 2) were used to predict NAD(P)-binding proteins in human structures from the June 2013 release of the PDB[16]. Interestingly, the class XII 3d motif was not found in any human protein structure. The class II, III and IV 3d motifs were found in 41 human proteins, whose structures indeed contain NAD(P), confirming all the predictions.

The 1d motifs derived from the class IV and XII 3d motifs, which could distinguish between NAD(P)- and nonNAD(P)-binding proteins with 100% PPV (see Table 4), were used to predict NAD(P)-binding proteins in human sequences from the June 2013 UniProtKB/Swiss-Prot database[1]. Like the class XII 3d motif, the XII_NAD motif was not found in any human protein sequence

Table 4 | Precision of the 1d motifs to distinguish between NAD(P)-binding and nonNAD(P)-binding proteins

	% frequency of 1d motif in 1d dataset[a]					% PPV of 1d-NAD(P) vs.			
1d motif	NAD(P)	FAD	PO_4	nonPO_4	nonNAD(P)	FAD	PO_4	nonPO_4	nonNAD(P)
I_NAD	13.2	13.8	0.7	0.2	0.4	96	78	84	68
I_NADP	12.4	61.0	3.5	1.8	2.4	84	40	38	24
II_NADP	0.1	0	0.005	0	0.002	100	79	100	79
III_NAD	4.8	1.7	0.4	0.1	0.2	99	68	78	57
III_NADP	3.1	2.2	0.8	0.3	0.5	97	41	47	28
IV_NAD	1.2	0	0	0	0	100	100	100	100
IV_NADP	1.0	0	0	0	0	100	100	100	100
XII_NAD	0.07	0	0	0	0	100	100	100	100

[a]The number of protein sequences in the given dataset matching the 1d motif divided by the total number of sequences in the given dataset, multiplied by 100.

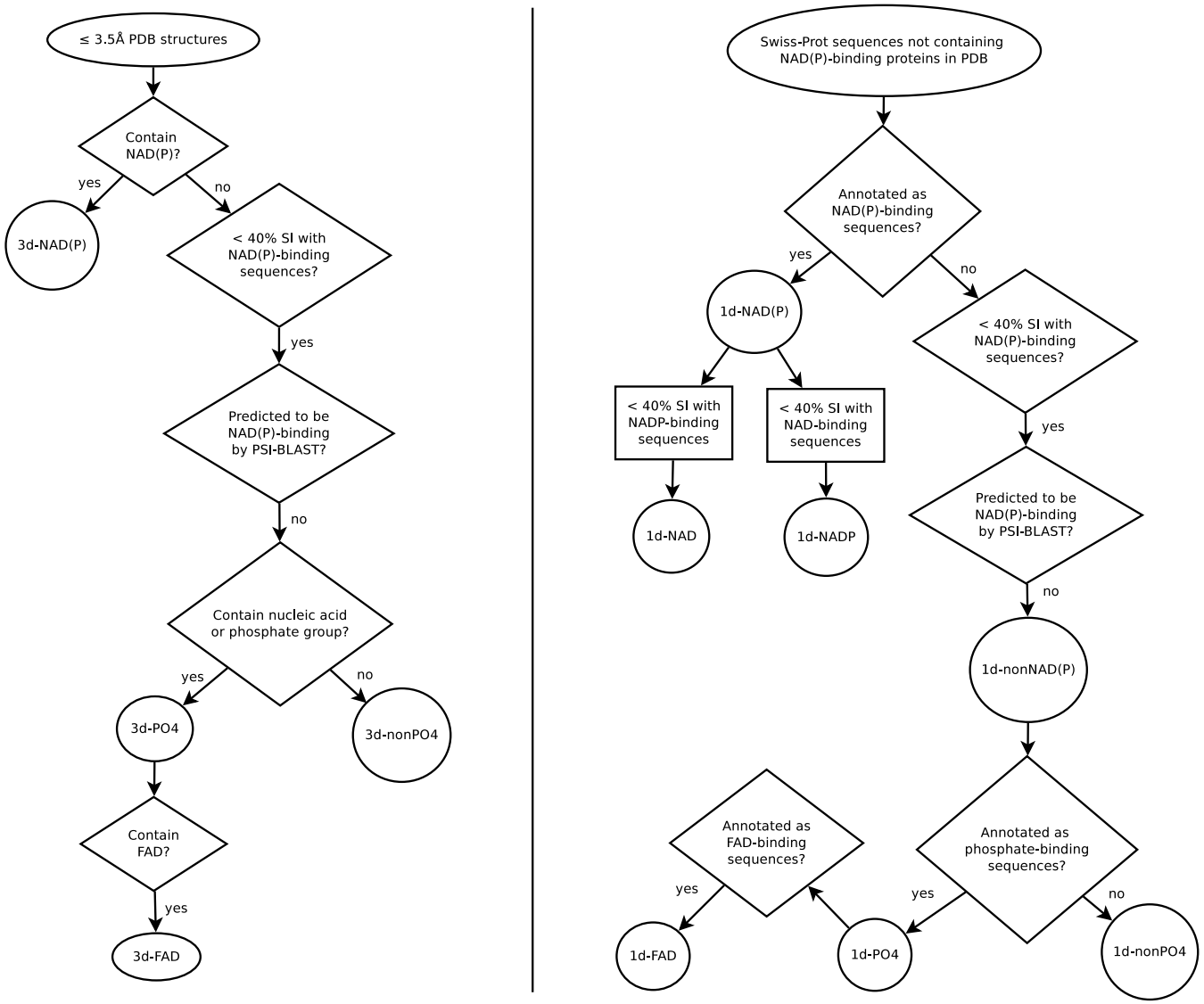

Figure 2 | Flowchart of protocol for generating 3d datasets and 1d datasets. See text in Methods for a description of the process used to generate the four 3d datasets (left), and seven 1d datasets (right). SI denotes sequence identity.

(see Discussion). The IV_NAD and IV_NADP motifs predicted 25 and 21 NAD(P)-binding proteins, respectively, out of which two are novel (accession numbers Q8N5I4 and Q96LJ7). The II_NADP 1d motif, which can discern NAD(P)-binding proteins from nonphosphate-binding ones with 100% PPV, predicted two NAD(P)-binding human sequences, one of which is novel (accession number Q9GZT4).

Discussion

This work has shown that the distinct locally conserved structures employed by NAD(P)-binding proteins for the same function; viz., binding the pyrophosphate, rarely occur in other proteins, especially those do not bind phosphate-containing ligands. Given a novel structure of a protein with unknown function, the 3d motifs in Fig. 1 could help to not only identify a NAD(P)-binding protein, but also suggest the pyrophosphate-binding site. This could in turn help to dock the cofactor to the protein. Given a novel sequence with little homology to existing sequences, 1d motifs derived from class IV and XII 3d motifs, which are not found in any nonNAD(P)-binding proteins, can be used to annotate NAD(P)-binding proteins, whereas the II_NADP motif, which was not found in nonphosphate-binding proteins, can distinguish between NAD(P)- and nonphosphate-binding proteins. These 1d motifs predicted three novel NAD(P)-binding human sequences.

This work has also shown the usefulness of the motifs by revealing a novel drug target region with unique sequence and structural characteristics: The locally conserved class XII phosphate-binding structure and sequence are found only in bacterial enoyl-acyl carrier protein reductases (EC 1.3.1.9/1.3.1.10), which are key enzymes of the type II fatty acid synthesis system. Because new antibiotics are urgently needed for multidrug-resistant bacteria and the function of enoyl-acyl carrier protein reductase is essential for the bacterial survival[17], the class XII 3d motif serves as an attractive novel drug target region since it is conserved across many bacterial species and is not found in any human proteins.

Methods

Dataset of NAD(P)-bound protein structures. A set of redundant NAD(P)-binding protein structures was created by searching the June 2013 release of the PDB[16] for ≤3.5 Å X-ray structures of proteins bound to oxidized or reduced NAD(P). If a NAD(P)-binding protein has multiple structures, then the highest resolution structure was chosen. If the structure contains multiple subunits, only one representative conformation was included. This generated 1,096 NAD(P)-binding proteins in the 3d-NAD(P) dataset (Fig. 2, left).

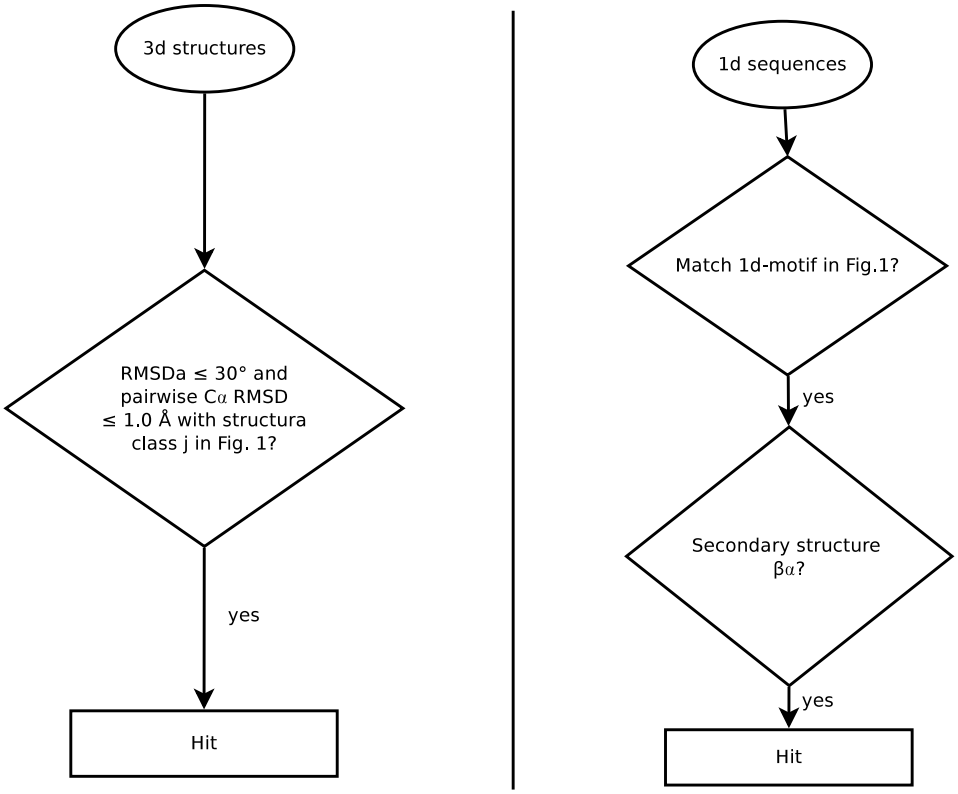

Figure 3 | Flowchart of process for determining hits. A hit was recorded if (left) the 3d structure and one of the 3d motifs in Fig. 1 shared RMSDa $\leq 30^{\circ}$ and pairwise Cα RMSD $\leq$ 1.0 Å, or (right) the 1d sequence matched one of the 1d motifs in Fig. 1 and the matched segment has a βα structure.

Datasets of NAD(P)-binding sequences. All NAD(P)-binding sequences were extracted from the manually curated UniProtKB/Swiss-Prot June 2013 database[1] by searching for the ligand keyword NAD or NADP. They were compared to those in the PDB and identical sequences were removed. This yielded a set of 24,516 NAD(P)-binding sequences (1d-NAD(P) dataset). To create a set of protein sequences that bind only NAD (1d-NAD) and another set of sequences that bind only NADP (1d-NADP), the annotated NAD-binding and NADP-binding sequences in the 1d-NAD(P) dataset were compared. Those sharing $\geq$ 40% sequence identity were removed, as such sequences may bind both NAD and NADP. This yielded 15,340 NAD-binding and 6,722 NADP-binding sequences (Fig. 2, right).

Dataset of NAD(P)-free protein structures. To obtain NAD(P)-free protein structures, the sequences of all proteins with $\leq$3.5 Å PDB structures were compared with the NAD(P)-binding sequences using CD-HIT-2D[18]. Those sharing $\geq$ 40% sequence identity were removed, as these structures might be similar to the NAD(P)-bound protein structures so their sequences might bind NAD(P). Sequences predicted by PSI-BLAST[19] to be NAD(P)-binding with an E-value $<$ 0.005 were also removed. The remaining NAD(P)-free protein structures were divided into two groups: (i) those containing nucleic acids or cofactors with phosphate groups and (ii) those without any bound phosphate. The first group contained 10,292 NAD(P)-free structures with phosphate-containing ligands (3d-PO_4 dataset), while the second group comprised 33,514 NAD(P)-free structures with no phosphate groups (3d-nonPO_4 dataset) (Fig. 2, left). From the 3d-PO_4 dataset, 348 structures that contained FAD were extracted to generate the 3d-FAD dataset.

Dataset of NAD(P)-free sequences. To determine how well the 1d motifs can distinguish NAD(P)-binding proteins from nonNAD(P)-binding ones, three 1d datasets for nonNAD(P)-binding proteins were created. All NAD(P)-binding sequences in the 1d-NAD(P) dataset were removed from the June 2013 UniProtKB/Swiss-Prot sequences, yielding 427,592 putative non-NAD(P)-binding sequences. If the latter shared $\geq$ 40% sequence identity with the NAD(P)-binding sequences or were predicted by PSI-BLAST to be NAD(P)-binding with an E-value $<$ 0.005, they were removed. This yielded 402,353 non-NAD(P)-binding sequences (1d-nonNAD(P) dataset), out of which 131,165 are annotated to bind nucleic acids or cofactors with phosphate groups (1d-PO_4 dataset), while the remaining 271,188 sequences are assumed not to bind to phosphate groups (1d-nonPO_4 dataset) (Fig. 2, right). A subset of 949 FAD-binding sequences (1d-FAD dataset) was extracted from the 1d-PO_4 dataset using the ligand keyword FAD in the UniProtKB/Swiss-Prot database[1].

Secondary structure prediction. Since the 1d motifs were derived from locally conserved βα structures (see Fig. 1), secondary structures were assigned to the sequences in the 1d-NAD(P) and 1d-nonNAD(P) datasets as follows: First, sequences that share $\geq$40% sequence identity were grouped together[18] and the longest sequence in a group was chosen as the representative one. This yielded 2,377 NAD(P)-binding and 78,656 nonNAD(P)-binding nonredundant sequences. Next, Porter 4.0[20] was used to predict the secondary structure of each nonredundant sequence. Sequences that share $\geq$40% sequence identity were aligned using Clustal Omega 1.20[21] and assigned the predicted secondary structures of the nonredundant sequence. A hit was recorded if a 1d sequence matched one of the 1d motifs in Fig. 1 and the matched segment has a βα structure (Fig. 3, right).

Structural similarity definition. To determine whether a PDB structure contained any of the distinct pyrophosphate-binding 3d motifs in Fig. 1, we used two similarity measures: (1) the root-mean-square deviation of Cα atoms (RMSD) and (2) the root-mean-square deviation of dihedral angles (RMSDa). First, a 12-residue sliding window was used to scan each protein in the 3d datasets (see above). Each 12-residue segment, described by a vector of backbone ϕ and ψ dihedral angles $V_1(\phi_1, \psi_1, \ldots, \phi_{12}, \psi_{12})$, was superimposed upon the central 12 residues of each distinct pyrophosphate-binding structure, described by the vector $V_2(\phi_1, \psi_1, \ldots, \phi_{12}, \psi_{12})$. The RMSDa was computed according to:

$$\mathrm{RMSDa}\ (V_1, V_2) = \sqrt{\sum_{i=1}^{11} [\psi_i(V_1) - \psi_i(V_2)]^2 + [\phi_{i+1}(V_1) - \phi_{i+1}(V_2)]^2 / 22} \quad (1)$$

The PDB structure containing $V_1(\phi_1, \psi_1, \ldots, \phi_{12}, \psi_{12})$, was considered to possess a given pyrophosphate-binding structure in Fig. 1 if the RMSDa was $\leq 30^{\circ}$ and the pairwise Cα RMSD was $\leq$1.0 Å (Fig. 3, left).

1. Magrane, M. & the, Uniprot & Consortium. UniProt Knowledgebase: a hub of integrated protein data. *Database* **2011**, bar009 (2011).
2. Yates, S. P., Jorgensen, R., Andersen, G. & Merrill, A. R. Stealth and mimicry by deadly bacterial toxins. *Trends Biochem. Sci.* **31**, 123–133 (2006).
3. Peralta-Leal, A. *et al.* PARP inhibitors: New partners in the therapy of cancer and inflammatory diseases. *Free Radic. Biol. Med.* **47**, 13–26 (2009).
4. Kirkland, J. B. Poly ADP-ribose polymerase-1 and health. *Exp. Biol. Med.* **235**, 561–568 (2010).
5. Watson, J. D., Laskowski, R. A. & Thornton, J. M. Predicting protein function from sequence and structural data. *Curr. Opin. Struct. Biol.* **15**, 275–284 (2005).

6. Sigrist, C. J. *et al*. PROSITE: a documented database using patterns and profiles as motif descriptors. *Brief. Bioinform.* **3**, 265–274 (2002).
7. Mathura, V. S., Schein, C. H. & Braun, W. Identifying property based sequence motifs in protein families and superfamilies: application to DNase-1 related endonucleases. *Proteins* **19**, 1381–1390 (2003).
8. Hulo, N. *et al*. The PROSITE database. *Nucleic Acids Res.* **34**, D227–D230 (2006).
9. Wass, M. N. & Sternberg, M. J. E. ConFunc–functional annotation in the twilight zone *Bioinformatics* **24**, 798–806 (2008).
10. Wu, C. Y., Chen, Y. C. & Lim, C. A structural-alphabet-based strategy for finding structural motifs across protein families. *Nucleic Acids Res., 38*, e150 (2010).
11. Wu, C. Y., Hwa, Y.-H., Chen, Y. C. & Lim, C. Hidden Relationship between conserved residues and locally conserved phosphate-binding structures in NAD(P)-binding Proteins. *J. Phys. Chem. B* **116**, 5644–5652 (2012).
12. Rossmann, M. G., Liljas, A., Branden, C. I. & Banaszak, L. T. Evolutionary and structural relationships among dehydrogenases. *The Enzymes* **11**, 61–102 (1975).
13. Bellamacina, C. R. The nicotinamide dinucleotide binding motif: a comparison of nucleotide binding proteins. *Faseb J.* **10**, 1257–1269 (1996).
14. Kleiger, G. & Eisenberg, D. GXXXG and GXXXA Motifs Stabilize FAD and NAD(P)-binding Rossmann Folds Through Cα–HO Hydrogen Bonds and van der Waals Interactions. *J. Mol. Biol.* **323**, 69–76 (2002).
15. Brakoulias, A. & Jackson, R. M. 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 (2004).
16. Berman, H. M. *et al*. The Protein Data Bank. *Acta Crystallogr. Sect. D-Biol. Crystallogr.* **58**, 899–907 (2002).
17. Lu, X., Huang, K. & You, Q. Enoyl acyl carrier protein reductase inhibitors: a patent review (2006–2010). *Expert Opin. Ther. Patents* **21**, 1007–1022 (2011).
18. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).
19. Altschul, S. F. *et al*. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
20. Mirabello, C. & Pollastri, G. Porter, PaleAle 4.0: high-accuracy prediction of protein secondary structure and relative solvent accessibility. *Bioinformatics* **29**, 2056–2058 (2013).
21. Sievers, F. *et al*. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7**, 539 (2011).

Acknowledgments

We thank Peggy Chiu for helpful discussion. This work was supported by the Human Frontier Science Program, Academia Sinica, and the National Science Council, Taiwan.

Author contributions

Y.H.H. and C.Y.W. performed the research. K.S. helped with statistical analyses. Y.H.H. prepared figure and tables. C.L. designed the project and wrote the manuscript text. All authors reviewed the manuscript.

Additional information

Supplementary information accompanies this paper at http://www.nature.com/scientificreports

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Hua, Y.H., Wu, C.Y., Sargsyan, K. & Lim, C. Sequence-motif Detection of NAD(P)-binding Proteins: Discovery of a Unique Antibacterial Drug Target. *Sci. Rep.* **4**, 6471; DOI:10.1038/srep06471 (2014).