Bioinformation
www.bioinformation.net

open access
Database

HNHDb: A database on pattern based classification of HNH domains reveals functional relevance of sequence patterns and domain associations.

Alaguraj Veluchamy*, Sujitha Mary, Vishal Acharya, Preeti Mehta, Taru Deva, Sankaran Krishnaswamy

Center of Excellence in Bioinformatics, School of biotechnology, Madurai Kamaraj University, Madurai, India; Alaguraj Veluchamy – Email: alaguraj@mkustrbioinfo.in; Phone: +91 452 2459141; Fax: +91 452 2459105; *Corresponding author

Received June 24, 2009; Accepted July 26, 2009; Published September 06, 2009

Abstract:
The HNH Database is a collection and sequence-based classification of HNH domain proteins. The database contains about 1913 HNH domain containing proteins, and is classified into 10 subsets based on the sequence pattern. Each of these subsets has unique signature sequences. We have shown a correlation between the subset combination and their domain association and function. Functional divergence of this domain may be due to the combination of these conserved patterns and the large variations in the non-conserved regions. HNHDb is freely available at http://bicmku.in:8081/hnh.

Keywords: HNH domains; HNHDb; sequence classification

Background:
Decomposing each protein into modular domains and each domain into subclasses is a basic prerequisite for accurate functional classification of protein molecules. The protein sequence classification is also helpful in organizing huge data produced by large-scale genome sequencing projects. The domain HNHc (SMART id: SM00507) is a conserved domain of around 50 amino acids, characterized by the presence of central conserved Asp/His residue flanked by conserved His (N-terminal) and His/Asp/Glu (C-terminal) residues at some distance. HNH domains are found among homing endonucleases, inteins, Group I and Group II introns, as well as free standing ORFs in viruses, archaebacteria, eubacteria and eukaryote, showing a polyphyletic relationship and are associated with a range of DNA binding proteins, performing a variety of binding and cutting functions **[1-3]**. They are involved in a variety of cellular activities including bacterial toxicity, homing functions in group I and II introns and inteins, recombination, developmentally controlled DNA rearrangement, phage packaging and as restriction endonuclease **[4-6]**. HNH homing endonucleases are members of the five families of homing endonucleases including LAGLIDADG, GIY-YIG, HNH, HIS-CYS box and cyanobacterial intron homing endonucleases. Among these five families, HNH family is evolutionarily more related to His-Cys box nucleases. Structurally, HNH-motif family is one among the six families of His-Me finger endonucleases and has a topology similar to that of a zinc finger motif **[7]** or treble clef motif **[8]**, containing two ß-strands and one α-helix linked together by a divalent metal ion often referred to as ßßα-Me motif. Although variations in sequence are common among the members of the His-Me finger endonuclease superfamily, they share a structural similarity over some particular residues in the active site region **[9]** and many of the type-II restriction endonucleases are found to belong to HNH fold **[10, 11]**. The HNH catalytic motif are highly adaptable and shows slight configurational modifications depending on the enzyme and the substrate **[12, 13]** and they are found to differ in their activity against double stranded DNA, single stranded DNA and single stranded RNA. The goal of this classification is to address the growing need to corroborate and integrate data by delineating characteristic subsequences using a regular expression type method. Since few of the HNH proteins are of known function such a classification database will help functional and structural analysis.

Methodology:
Datasets:
We used HNH domain sequence family derived from SMART database.

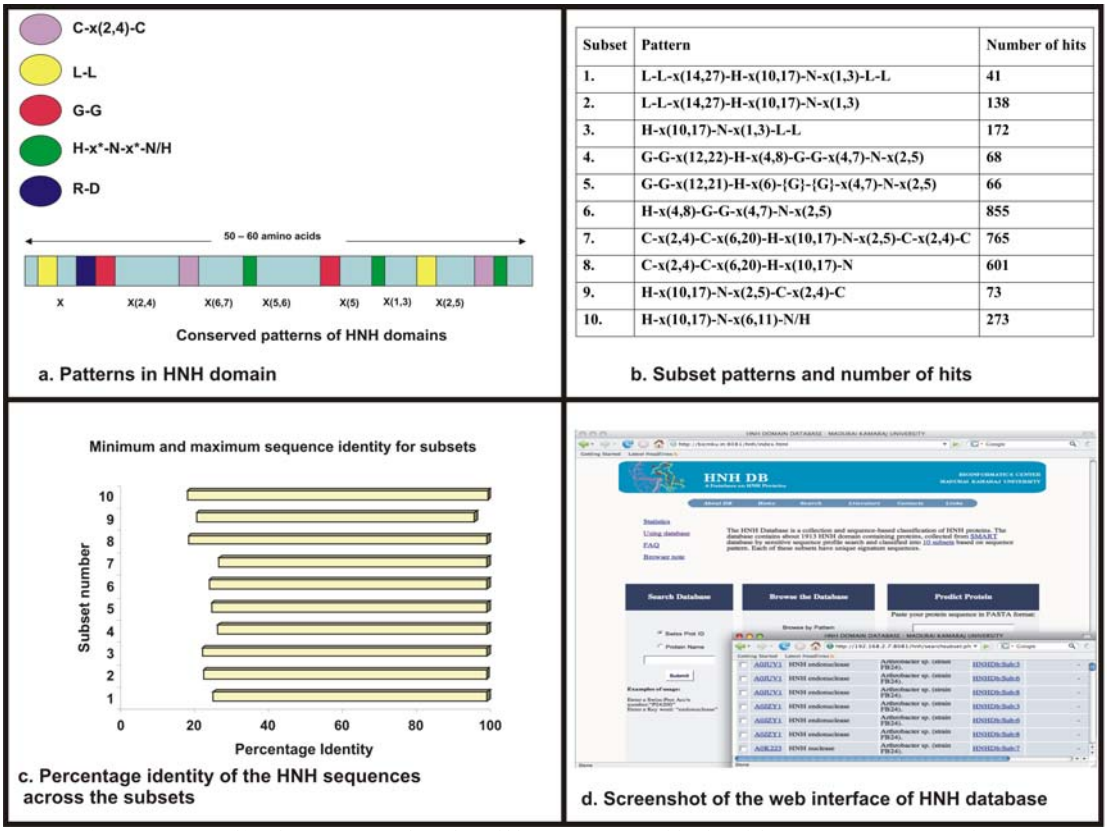
HNH Protein sub-classification:
The sub-classification of HNH proteins involves two steps: (1) Generating a PROSITE pattern from unaligned HNH domain sequences with the PRATT program **[14]** and the resulting multiple patterns are used to generate weight based matrices **[15]**. (2) The WAPAM was used to search against sequences in SMART **[16]** database and cluster them into subsets. A total of 2483 HNH domain sequences obtained from the SMART database were cross-checked with Swissprot. The redundant sequences were removed and obsolete entries were deleted, forming set of 2143 sequences. These subset members along with other features were inlaid in planned database architecture.

Database Design:
The HNH database is implemented in MySQL (v 4.1.12) (RDBMS) with PHP (v 4.3.5) as a front-end tool. Perl scripts are used to generate the PROSITE patterns and construction of this database. HNHDb is well linked to other databases like Swiss-Prot by its accession number, PDB by its Id. HNH protein sequences can be queried using accession numbers, PROSITE pattern or a key word search of protein name/function. ClustalW and Mview are integrated so that user can select sequences and align. The subsets are named HNHDb:Sub:1 to HNHDb:Sub:10.

Discussion:
About 90% (1913 HNH domain proteins into 10 subsets) of the available HNH proteins are classified and the database covers 100% of all HNH proteins in Pfam **[17]**. Each subclass has a particular set of defined conserved patterns. Few functional classes were derived, which have combinations of these patterns. The whole HNH domain appears to be a mere combination of these patterns. Their functional variation could possibly result from the presence or absence of these significant regions.

The HNH domain sequence can be characteristically represented in a PROSITE pattern as: [L]-[L]-x-[R]-[D]-G-G-x(2,4)-C-x(2,4)-C-x(6,7)-[D]-H-x(5,6)-G-G-x(5)-N-x(1,3)-[L]-[L]-x(2,5)-C-x(2,4)-C-[NH]. Most of the HNH domain sequences contain repetition of the few dyads i.e. -G-G- or –L-L- or -C-x(2,4)-C-. This could possibly make the difference in the function or different DNA-binding strategies. The sequences which do not have these characteristic patterns may be due to mis-annotation or the group may be different altogether. Apart from H-N-H residues, there are three dyads, which by combinations i.e. either by presence or absence, forms particular pattern.

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 4(2): 80-83 (2009)



Subset	Pattern	Number of hits
1.	L-L-x(14,27)-H-x(10,17)-N-x(1,3)-L-L	41
2.	L-L-x(14,27)-H-x(10,17)-N-x(1,3)	138
3.	H-x(10,17)-N-x(1,3)-L-L	172
4.	G-G-x(12,22)-H-x(4,8)-G-G-x(4,7)-N-x(2,5)	68
5.	G-G-x(12,21)-H-x(6)-{G}-{G}-x(4,7)-N-x(2,5)	66
6.	H-x(4,8)-G-G-x(4,7)-N-x(2,5)	855
7.	C-x(2,4)-C-x(6,20)-H-x(10,17)-N-x(2,5)-C-x(2,4)-C	765
8.	C-x(2,4)-C-x(6,20)-H-x(10,17)-N	601
9.	H-x(10,17)-N-x(2,5)-C-x(2,4)-C	73
10.	H-x(10,17)-N-x(6,11)-N/H	273

Figure 1: (a) HNH pattern; (b) Subset patterns and number of hits; (c) Percentage identity of the HNH sequence across the subsets; (d) Screenshot of the web interface of the HNH Database.

Defining features:
Common protein signatures **(Figure1a)** are: (1) A L-L leucine dyad at the beginning of the domain; (2) An R-D, immediately following leucine dyad; (3) Presence of Cysteine double dyad one before the first H and the other between N and H. This cysteine dyad is found to be delimiting the domain boundary; (4) Presence of GG double dyad, one before first H and the other between H and N of the HNH. (5) H-x(*)-N-y(*)-N/H. The unusual property is that the defining features in the patterns occur twice, one at the N-terminal side and the other at the C-terminal, wrapping around the HNH pattern active residues.

Subset description:
The seven distinct features seen in HNH sub-classification by Mehta et al **[9]** is diffused among 10 subsets adding some more features to the HNH proteins. The most prominent patterns **(Figure 1b)** are G-G and -C-x(2,4)-C-, which marks the boundary of the domain at the N and C-terminal. The two Cysteine dyads are present at both ends of the domain sequence, with 30 amino acids gap. If the domain sequence is long then the mid portion is increased and the dyads remain at the ends. These pattern can then be characteristically represented as –C-x(2,4)-C-x(30)-C-x(2,4)-C-. The glycine dyad at the C-terminal is more conserved than the one at the N-terminal. Each subset has particular number of hits and their similarity is in large variation, although a common pattern is found **(Figure 1c)**.

Functional annotation:
We observe that a particular functional class has a specific set of protein signature combination **(Table 1 in supplementary material)**. Most of the functional classes are assigned based on the associated domain. Although the DNA binding activity is invariable to the HNH domain, the biological function may vary due to the associated domain. These functional classes ensure that the derived patterns are highly unlikely to have emerged by chance. PDB structures of few members of functional classes are known. These can help in homology modeling of the members of the clusters whose structures are not known.

Further, residues of subset 1 to subset 9 may be part of SDRs (specificity determining residues) and subset 10 is active residues or functionally conserved residues (FCRs). The variation in the SDRs may leads to divergence in activity. Regular expression based methods are better at differentiating SDRs and FCRs. DATABASE ACCESS: The database **(Figure 1d)** and associated information files are freely accessible at the URL http://bicmku.in:8081/hnh.

Conclusion:
The relationship between the cellular role of a protein, conserved subsequence in a domain and the type of domain association is established in HNH domain containing sequences. The promiscuity of the HNH domain sequences, conservation of SDRs and their organism wise distribution among the subsets, suggests horizontal transfer and co-evolution.

Acknowledgement:
DBT, GOI for COE in Bioinformatics facility

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 4(2): 80-83 (2009)

References:
[1] JM Bujnicki *et al., Trends Biochem. Sci.* **26**:9 (2001) [PMID: 11165501].
[2] AE Gorbalenya, *Protein Sci.* **3**:1117 (1994) [PMID: 7920259].
[3] L Aravind *et al., Nucleic. Acids. Res.* **28**:3417 (2000) [PMID: 10982859].
[4] JZ Dalgaard *et al., Nucleic. Acids. Res.* **25**:4626 (1997) [PMID: 9358175].
[5] DR Edgell, *Curr.Biol.* **19**: R115 (2009) [PMID: 19211047].
[6] RP Bonocora, DA Shub, *Curr. Biol.* **19**:223 (2009) [PMID: 19200727].
[7] MJ Sui *et al., Protein. Sci.* **11**:2947 (2002) [PMID:12441392].
[8] NV Grishin, *Nucleic. Acids. Res.* **29**:1703 (2001) [PMID: 11292843].
[9] P Mehta *et al., Protein. Sci.* **13**:295 (2004) [PMID: 14691243].
[10] E Kriukiene *et al., Biochim. Biophys. Acta.* **1751**:194 (2005) [PMID: 16024301].
[11] M Saravanan *et al., Nucleic. Acids. Res.* **32**:6129 (2004) [PMID: 15562004].
[12] BS Chevalier, BL Stoddard, *Nucleic. Acids. Res.* **29**: 3757 (2001) [PMID: 11557808].
[13] EA Galburt, BL Stoddard, *Biochemistry* **41**:13851 (2002) [PMID: 12437341].
[14] I Jonassen *et al., Protein. Sci.* **4**:1587 (1995) [PMID: 8520485].
[15] G Stéphane *et al., Parallel Computing* **31**:73 (2005).
[16] I Letunic *et al., Nucleic. Acids. Res.* **32**:D142 (2004) [PMID: 14681379].
[17] RD Finn *et al., Nucleic. Acids. Res.* **36**: D281 (2008) [PMID: 18039703].

Edited by P. Kangueane
Citation: Veluchamy *et al.*, Bioinformation 4(2): 80-83 (2009)

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 4(2): 80-83 (2009)

Supplementary material:

Table 1: Functional pattern combination

S.I	Associated domains	Function	LL at start	LL bet N and H	RD before first C-x(2)-C	C-x(2)-C at N-terminal	C-x(2)-C at C-terminal	GG after N-terminal C-x(2)-C	GG between First H and N	H-x-N-y-N/H
1.	RVT_1	Reverse transcription	*	*	*	*	-	-	-	*
2.	NUMOD4/ NUMOD1	DNA binding homing endonucleases	-	-	-	-	-	-	-	*
3.	Pyocin/ Cloacin	Bacteriocins	-	-	-	-	-	-	*	*
4.	DUF222/ DUF262	Unknown	-	-	-	*	-	-	*	*
5.	AP2	Plant defense-Transcriptional control	-	-	-	-	-	-	*	*
6.	MutS_1	DNA mismatch repair	-	-	-	*	*	-	-	*
7.	ResIII	Type-III restriction enzyme	-	-	-	*	*	-	*	*
8.	Helicase_C	helicase	-	-	-	-	*	-	*	*

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 4(2): 80-83 (2009)