

New types of conserved sequence domains in DNA binding regions of homing endonucleases

Einat Sitbon and Shmuel Pietrokovski

Molecular Genetics department, Weizmann Institute of Science, POBox 26, Rehovot 76100, Israel.

Address correspondence to Shmuel Pietrokovski <pietro@bioinfo.weizmann.ac.il>

Abstract

DNA-binding protein domains are highly diverse in sequence and structure. Currently, there is no general way to identify such domains from protein sequence. Homing endonucleases specifically bind very long DNA regions. Only a few DNA binding domains from these proteins have been studied in detail. We identified four new types of conserved sequence domains in homing endonucleases and related proteins. The conserved domains are 14 to 50 amino acids long, appearing in various combinations with each other and with other conserved domains. One domain includes a motif previously shown by structure determination as a sequence-specific DNA binding helix and two other domains are similar by sequence to helix-turn-helix DNA binding domains and conserved regions of DNA binding proteins. Modular occurrence, presence in known and putative DNA binding protein regions and similarity to DNA binding motifs identify the new domains as nuclease-associated modular DNA binding domains (NUMODs). The highly modular nature of these domains and their concurrent appearance in various homing endonucleases suggest they act together to create highly specific DNA binding.

Keywords: DNA binding protein domains, homing endonuclease, sequence motifs, domain modularity.

Proteins bind DNA in a sequence specific manner using various types of domains. These DNA binding domains often occur in different combinations with each other and with domains of other functions. This enables diverse binding specificities in different contexts. Modularity of the DNA binding domains also facilitates evolutionary changes in binding specificity and context. Better understanding of DNA binding modules will be useful in identifying DNA binding proteins, as well as in designing such proteins [e.g. 5]

Homing endonucleases are rare-cutting enzymes that mediate the horizontal transfer of different types of genetic elements to unoccupied integration points. This is done by highly specific cleaving, or nicking, of very long (12-40 bp) DNA recognition sequences. Homing endonucleases are encoded in many group I and II introns, and in inteins. Four families of homing endonucleases are known, each characterized by several conserved short sequence motifs of its nuclease domain [3]. The DNA binding activity and regions of homing endonucleases are not characterized as well as their nuclease activity and regions. Here we present analysis of the DNA binding domains of different homing endonucleases. We identified several motifs that appear in different homing endonucleases families in a modular fashion. Sequence similarity, context analysis and protein structure data show our motifs to be new types and variants of DNA binding domains.

We started the analysis by searching for conserved motifs in the sequences of GIY-YIG and HNH homing endonucleases and other proteins with these nuclease domains. In all cases the nuclease domains were excluded, restricting the search to the putative DNA binding domains. Four conserved regions were found and named NUMOD1 to 4 for NUclease-associated MOdular Domains. NUMOD1 to 3 each consist of a single ungapped motif, 14 to 34 amino acids long, and NUMOD4 is made up of three closely spaced ungapped motifs (figure 1). Following their identification block multiple sequence alignments of the domains were used to search sequence and block databases for other members and for related conserved regions.

All NUMODs are modular, appearing in different combinations with each other and with different nuclease domains (figure 2). NUMOD1 occurs in single or tandem repeats in HNH, GIY-YIG and LAGLI-DADG homing endonucleases. The motif is similar to a conserved region of the bacterial sigma54-activator DNA-binding proteins and its carboxy-terminal 15 amino acids are also similar to the amino-terminal helix

of HTH DNA binding motifs (figure 3). In HTH motifs this helix is responsible for non-sequence specific interactions with DNA [19]. NUMOD2 occurs in HNH and GIY-YIG homing endonucleases (figures 2 and 4). It is significantly similar across its almost its entire length to HTH DNA binding domains across all their conserved region (figures 3 and 5). Most of these HTH domains are classified in the SCOP database [13] in the super family of "Winged helix" DNA-binding domain. We identify NUMOD2 motif at probable homing endonucleases in fungal mitochondria and bacteriophages. NUMOD1 and NUMOD2 sequence motifs are included together in SMART [12] domain 00497.5 (IENR1 intronic repeat). Our sequence to sequence and block to block comparisons find the two motifs clearly distinct. However, parts of both motifs are similar to the N' helix of the HTH DNA binding domain (figure 3). NUMOD3 occurs in single and tandem repeats in GIY-YIG proteins, including the DNA binding domain of I-TevI homing endonuclease (figures 2 and 4). The structure of this domain was solved together with its DNA substrate [17]. NUMOD3 corresponds to a beta-turn loop helix sub-domain of the structure. This region binds the DNA in a sequence specific manner with the helix inserting and distorting the minor groove of the DNA [17]. The I-TevI DNA binding domain is made up of two additional sub-domains connected by loops. The domain forms a unique extended structure that wraps its DNA substrate (figure 6). NUMOD4 is made up of three conserved motifs that together with short unconserved connecting regions span 48 to 50 amino acids. The domain occurs in uncharacterized ORFs and in putative HNH endonucleases of various bacteriophages (figure 2).

Our analysis of homing endonuclease protein sequences identified four new sequence domains that appear in a similar modular manner. One of these domains (NUMOD3) includes a region shown by structure determination as a minor groove binding helix [17]. Two other domains (NUMOD1 and 2) are similar by sequence to HTH DNA binding domains. NUMOD4 too is suggested to be a DNA binding domain by its modular presence in the substrate binding regions of homing endonucleases.

NUMOD3 corresponds to part of a sub-domain in the DNA binding region of I-TevI [17]. This region has a non-globular structure composed of several sub-domains that are independently stable. We identify NUMOD3 motifs in different sequence contexts, often appearing with other NUMOD motifs and as tandem repeats. Typically the motifs are closely positioned to each other (figure 4). This suggests that NUMOD motifs are part of additional non-globular DNA binding regions.

Methods that characterize and search for short ungapped sequence regions facilitated our analysis of modular sequence domains. Typical sequence search methods search for longer and possibly gapped regions or sequential occurrence of several short regions [1, 14]. Block sequence analysis can confidently identify short (<15 aa) regions regardless of their context due to its use of multiple sequence information in constructing a search matrix of the aligned regions [7]. Beyond block to sequence searches are block to block comparison and multiple block alignment methods [11, 15]. These methods can identify genuine sequence relations undetected by sequence to sequence and multiple-alignment to sequence procedures [11].

Together with expanding the known repertoire of DNA binding domains our finding may be useful for designing new DNA binding proteins. The relatively short length and high modularity of the domains we identified suggest they can be used in new combinations and contexts. Most of these new domains are found in intronic homing endonucleases that probably target their intron flanks [3]. This greatly assists the identification of the specific DNA sequences recognized by each domain. Identifying targets for specific domains will allow rational use, and perhaps even modification, of the domains for binding DNA.

Appendix

NUMOD 1	width = 34			ORF regions
name	gi	start		(for nucleotide sequences)
A287R_pbcv1	1181450	204	SKIVYQYDL DGM YIDKFRS CREAGRSLGKGHKYI	
A315L_pbcv1	1181478	193	AKCVYQFDMNGNFIQS FQTVAEAAEYLGKVHGGI	
A651L_pbcv1	2447115	175	AKKVYQYDMDGKYIGWFDSCEEAARHLEKSDGSD	
COBil_chleu-mt	2865254	253	SKPLGLYDTNNQLLKEFESITEASMYFKCDRKKI	
COBil_chlel-mt	2193888	243	PAPINLVDS DGKILASFKSI SAAAKHYGGCRKHL	
COBil_neucr-mt	13116	260	TRPVVLYNLNRTIYGKYSTILEAANA INCNEKTI	
COBil_sacdo-mt	13617	230	SLPLYLYNMNNE LICSFESRKVA AHL LGCNDRTI	691-1698
COBi2_podan-mt	1334531	247	TRPVVLYNLNGTVYGEYSTIL DAAKSINCDEKTI	
COIi14_podan-mt	1334547	371	IKAVFVYDINRKF I GKYDGVTD AQKALNLSHSTI	
COIi4_agrae-mt	2738528	333	SLAVFVYVKIKKF ICKYEGVTKAQEALKINHSII	
Y03E_BPT4	141153	128	GKPIYQYDLNGNFIRKYRCITDAAEDMSYSCSTS	
		212	NVPVFQYD TTGKLLRVFPRIKDA AVSVKGCMSNI	
o296_CVK2	2190279	230	SKRVYQYDS DGN YIRSFGSYREAGRSIGKSHSYI	147500-148249
o360_allm-mt	459018	165	SVPIFAYDYNGEF IGSYDSIDKASKALNVISKTI	
		235	AELVFAYDIDGNL I GKYSSGGEASKALGVSTSSI	
		304	AKATFVYDS DNNFVGEFGSVRAAAKALXXXXXXXX	
SAG1876_stra5	22538014	124	CKPVEQFTLEGEFINTFDSIKSASMTGISSQRI	
URFli_neucr-mt	83808	250	PLGVGIYDLEDNL I LKFSNNVELAKYLGISKVTV	
o6_bp1p31	9885252	118	SKAVEAYDNQGNFVMKFRSKQEAERHGYFSSAVV	
o233_triire-mt	18640459	172	GLEVEITDLETNTITVYSSIREAAKFLNSDIKTL	
e11_BPbIL170	9630614	114	FKKVIQLDLNDNVLNEFESMVQAEQETGVSRRNI	
o292_Rh136-mt	15147258	236	KKAINVYDKNNNLLYNFESITETALKLNI PKSSI	
Q0255_yeast-mt	6226541	417	NVGVFVYDLNNTLIMTFTGYRPAATYFNCSKHEI	
NUMOD 2	width=19			ORF regions
name	gi	start		(for nucleotide sequences)
ND1i1_neucr-mt	14129	268	NNVELAKYLGISKVTVGKY	
ND1i1_podco-mt	1743352	269	NNVELAKYLDISRVTVGKY	623-1537
e20_BPbIL1	3282299	106	SFSDLAKYVGVSHQSVSRN	
en_BPphiE	495456	155	GNSSIWKALNKGSV LASGY	
ND132_cocpo	contig132	209	NNTELANYLNISKVTVGKY	TIGR contig132 2405-2127

NUMOD 3 name	width=14 gi	start	ORF regions (for nucleotide sequences)	
A287R_pbcv1	1181450	161	FGKHHDEETKKKMS	
		178	KGKQLTEETKKKIS	
		137	FGKQLSEETKKKMS	
A315L_pbcv1	1181478	166	YGKKHTEESLKKQS	
		142	YGKPQKEEVKSKIS	
		118	YNKHHSEESKKKIS	
A495R_pbcv1	1620166	149	KGKIVSKETKKKMS	
		132	KGKIVSKDSKKKMS	
		115	LGKTHSEELYKKKMS	
A651L_pbcv1	2447115	148	IGMTHTEESKKKIS	
AT6i1_podan-mt	83822	112	SGWRHSEATIESMR	
COBi1_tripa-mt	732979	162	YGVKHTTEETKAAMR	
COBi1_chleu-mt	2865254	184	LGKHTTEETKNFMK	
COBi1_chlel-mt	2193888	174	LGKHTVETRNKMK	
COBi1b_podan-mt	578862	138	LGKHTTEEDKEKMR	
COBi2_podan-mt	1334531	184	FGYKHTVIDRQKMK	
COIi1i1_podan-mt	1334559	192	YGRTHSEETKALMA	
		164	LGHKLTEETKAKIS	
		233	LGRKHSEETLLKMS	
		216	FGKTHSEKTKELMG	
COIi14_podan-mt	1334547	228	LGKHTTEETKKLLS	
COIi2_chlel-mt	4379168	207	IGFKHSEETKRLLS	
		227	RGKPLSDETKTKLS	
		251	FGKKHSEEFKAWLS	
COIi4_agrae-mt	2738528	191	LGKHTKETKELLS	
ND1i2_allm-mt	2147548	175	AGYKHTDEAKAKMS	
		199	FGKSHTDEAKAKIS	
		223	FGKSHTDEAKAKIS	
ND1i4b_podan-mt	1334568	164	EGYKHTDEAKLKML	
		189	YGKKHTEQTLKLLS	
		211	YGKHLSEETKKKIS	
ND1i1_neucr-mt	14129	183	LGKHTTEEARLKMV	
		208	FGKHTTEALGLIS	
		230	FGKKHSEATKASMS	
SEGA_BPT4	417766	135	LGKKQSEETKAKRK	
		101	IGYRVSSSETKEKIS	
SEGC_BPT4	2506234	111	YGKSHSRETRLKIS	
SEGD_BPT4	20141632	115	TGVKQSDETIAKRV	
TEV1_BPT4	6094464	178	FNHKKSDITKSKIS	
o211a_allm-mt	2144206	145	LGKHTTEETKAKIS	
		122	FGVKASDETKAKMS	
		186	LGKPLSDEIRAKMS	
		169	FGKHTTEEAKAKIS	
BmoI_bacmo	12958588	171	KGKKHSEESKTKLS	1121-1918
		143	FGRKHTETTKLKIS	
		198	YGKTHSDEFKTYMS	
ND1i2_podca-mt	1495718	218	FGRKHSEVTKDKIS	1754-2629
o245_cvk2	2190279	158	YGKNHSKETKKKQS	836-3
o378_Rh136	15147263	200	AGYTFSDETKLLKMS	
o233_trire-mt	18640459	122	TGRKHSDEVKDLMS	
o639_bacsp	15211880	129	FGRRHSEKTKMIIS	
COBi_canal-mt	12539616	186	KGYKHSEESMAKMK	24384-25343
o732_bacsp	15211875	150	DGRKHTTEETKRKMS	
o212_Rh136	15147290	114	MGKKHSLETRIKMS	

NUMOD 4 width = 19,9,12

							ORF regions (for nucleotide sequences)
name	gi	start	A	start	B	start	C
IHmuII_BPSP82	1085761	13	YQISDNGDIFSLKSNRVLK	38	NGYIYIHLT	51	KKAFTIHRLVAL
O36_1_BPSP	1090456	15	YEVSTLGRVVRKKEGGRIK	38	RGYMLRWLY	51	KKDWYVHRLVAL
O38_BPPLLH	945381	14	YEVSDLGRVRSYATGKCAY	41	DGYSHIALR	54	AYEFRLNRLVAA
YG31_BPSP1	465641	13	YQVSNTGEVYSIKSGKTLK	37	DGYHRIGLF	50	GKTFQVHRLVAI
e11_BPbIL170	3282308	16	YEVSNLGVVNIKSGRIK	40	NGYLMHQLC	53	KKNLFLHRIIAT
e37_BPbIL170	3282283	11	YIILSNGEVWVKIHKNHYRK	37	NGYWNVSIN	46	NKATLLHKVITR
en_BPphiE	495456	13	YRISDNGDIFSLKSNKVLK	37	NGYTYIHLT	50	KKSFTIHRLVAL
o193_bpt5	15420	15	YLISPYGEVYSTKSNKLLT	39	AGYPFVTFY	52	NVSIVLHRLHAR 3919-4497
o194_bpt5	15420	12	YLVNEAGDVFSTFTNKVLS	35	DGYPAVKLQ	48	QTSVLIHRIISH 2291-2872
o41_BPr1t	1353558	11	YSINENGMVRNDNTEHIKQ	36	NGYLIVDLY	49	SEKVPIHRLVAE
yosQ_bacs	2634396	11	WWITETGVII SKKLKPRK	35	HGYEMIGYT	49	TQNYLVHRLVAK
O202_bacan	21397577	27	YTISSMGRVKSTNKNSGKS	55	SDYLVVNL	65	KKTHYVHRLVAM
o48_bp1A2	22296570	20	YEVSNKGRVKSlyTGKILH	44	TGYLYASMV	57	HF'SKSVHRLVAQ
1461_spyM18	19748602	11	YSINENGVVRNDITGRIKK	36	NGYLIVDLY	49	SEKVPIHRLVAE
0606_lacga	23002766	18	YQVSNLGRVKSlyKNTKIL	43	RGYQYVMFF	56	YKHFLVHRLVAQ

References

1. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, *et al.*: "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." *Nucleic Acids Res* , **25**:3389-3402 (1997).
2. Bailey TL, Elkan C: "Fitting a mixture model by expectation maximization to discover motifs in biopolymers" In *Second International Conference on Intelligent Systems for Molecular Biology*. Edited AAI Press, Menlo Park, California, 1994: 28-36.
3. Belfort M, Roberts RJ: "Homing endonucleases: keeping the house in order" *Nucleic Acids Res* , **25**:3379-88 (1997).
4. Burger G, Werner S: "The mitochondrial URF1 gene in *Neurospora crassa* has an intron that contains a novel type of URF." *J Mol Biol* , **186**:231-242 (1985).
5. Choo Y, Isalan M: "Advances in zinc finger engineering" *Curr Opin Struct Biol* , **10**:411-6. (2000).
6. DeLano WL. "The PyMOL Molecular Graphics System" *Journal ed^eds* (2002).
7. Henikoff JG GE, Taylor N, Pietrokovski S, Henikoff S: "Using the Blocks Database to Recognize Functional Domains" In *Current Protocols in Bioinformatics*. Edited by Andreas D. Baxevanis DBD. New York, John Wiley & Sons, 2002.
8. Henikoff S, Henikoff JG, Alford WJ, Pietrokovski S: "Automated construction and graphical presentation of protein blocks from unaligned sequences" *Gene* , **163**:GC17-26 (1995).
9. Koonin EV, Tatusov RL, Rudd KE: "Sequence similarity analysis of *Escherichia coli* proteins: functional and evolutionary implications." *Proc Natl Acad Sci USA* , **92**:11921-11925 (1995).
10. Kowalski JC, Belfort M, Stapleton MA, Holpert M, Dansereau JT, Pietrokovski S, *et al.*: "Configuration of the catalytic GIY-YIG domain of intron endonuclease I- Tev I: coincidence of computational and molecular findings." *Nucleic Acids Res* , **27**:2115-2125 (1999).
11. Kunin V, Chan B, Sitbon E, Lithwick G, Pietrokovski S: "Consistency analysis of similarity between multiple alignments: prediction of protein function and fold structure from analysis of local sequence motifs" *J Mol Biol* , **307**:939-49. (2001).
12. Letunic I, Goodstadt L, Dickens NJ, Doerks T, Schultz J, Mott R, *et al.*: "Recent improvements to the SMART domain-based sequence annotation resource" *Nucleic Acids Res* , **30**:242-4. (2002).
13. Lo Conte L, Ailey B, Hubbard T, Brenner S, Murzin A, Chothia C: "SCOP: a structural classification of proteins database." *Nucleic Acids Res* , **28**:257-259 (2000).
14. Park J, Karplus K, Barrett C, Hughey R, Haussler D, Hubbard T, *et al.*: "Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods" *J Mol Biol* , **284**:1201-1210 (1998).
15. Pietrokovski S: "Searching databases of conserved sequence regions by aligning protein multiple-alignments" *Nucleic Acids Res* , **24**:3836-45 (1996).
16. Schuler GD, Altschul SF, Lipman DJ: "A workbench for multiple alignment construction and analysis" *Proteins* , **9**:180-90 (1991).
17. Van Roey P, Waddling CA, Fox KM, Belfort M, Derbyshire V: "Intertwined structure of the DNA-binding domain of intron endonuclease I-TevI with its substrate" *Embo J* , **20**:3631-7. (2001).
18. Wallace JC, Henikoff S: "PATMAT: a searching and extraction program for sequence, pattern and block queries and databases" *Comput Appl Biosci* , **8**:249-54 (1992).
19. Wintjens R, Rooman M: "Structural classification of HTH DNA-binding domains and protein-DNA interaction modes" *J Mol Biol* , **262**:294-313. (1996).

Figure legends

Figure 1, NUMOD sequence motifs. Motifs are shown as sequence logos where the height of amino acids is proportional to their conservation in each position [Henikoff, 1995 #109]. NUMOD were identified by grouping sequence regions from proteins with GIY or HNH nuclease domains using the BLAST program after removing the nuclease regions themselves. Blocks were identified in each group using different local multiple alignment methods: BlockMaker [8], MACAW [16], and MEME program [2]. After the initial blocks were built, they were used to search the NCBI non-redundant protein and nucleotide databases and unfinished genomes data iteratively using BLIMPS [18] and MULTIMAT [8] programs. New identified motifs were added to the blocks and searches were repeated until convergence.

Figure 2, Modular occurrence of NUMOD motifs. Examples of different motif arrangements in various proteins. Protein names are in the swissprot format style (gene_organism) with “-mt” specifying mitochondrial encoded genes. Sequence accessions and alignments are given in the appendix. NUMOD1-4 are labeled by corresponding numbers, and nuclease domains are labeled by their name: GIY, HNH and LAGLI-DADG.

Figure 3, similarity of NUMOD1 and NUMOD2 to other conserved motifs. Block to block alignment of Sigma54 activator motifs with NUMOD1 (top, Z-score 15.8), NUMOD1 with HTH motifs (middle, Z-score 7.0), and NUMOD2 with HTH motifs (bottom, Z-score 7.1). Arrows indicate aligned positions for each pair. Sigma54 activators motif block includes motifs from 50 such sequences and the HTH block is a composite block of 609 HTH motifs from the ecmot database [9, 15].

Figure 4, Domain organization of *Neurospora crassa* mitochondrial ND1 intron 1 homing endonuclease [4]. The GIY endonuclease domain [10] is boxed and the NUMOD motifs are marked in gray and labeled.

Figure 5, Consistent alignment of NUMOD2 and HTH DNA binding motifs. A, each block is shown as a box labeled with its name. Lines connecting the blocks indicate similar blocks (Z scores 5.6 to 8.1) with bold lines being extremely high scores (Z-scores greater than 8.1). B, schematic multiple alignment of NUMOD2 and HTH blocks. Blocks are shown as lines where each character corresponds to an alignment position. Double lines show the region aligned with all other blocks in this group. The LAMA [15] and CYRCA [11] programs were used to identify similarity to protein families by comparing the NUMOD2 block to the Blocks+ database version 13.0. Block accessions are: LuxR bacterial regulatory proteins- IPB000792, ArsR bacterial regulatory proteins - IPB001845B, LacI bacterial regulatory proteins - IPB000843A, Crp bacterial regulatory proteins - IPB001808B, IS30 transposases - IPB001598A, Site-specific recombinases - IPB001822D.

Figure 6, Structure of a NUMOD3 motif. A, overview of I-TevI DNA binding region structure (pdb code 1i3j) [17] with sub-domains labeled and NUMOD3 motif in red. The motif sequence is FNHKHSDITKSKIS, located at positions 178-184. B, close-up of the NUMOD3 motif. The backbone is shown in red, and the side chains are colored in CPK scheme. Residues are named and numbered, except Ile 190, which is hidden behind other residues. Both figures were created by the PyMol program [6].

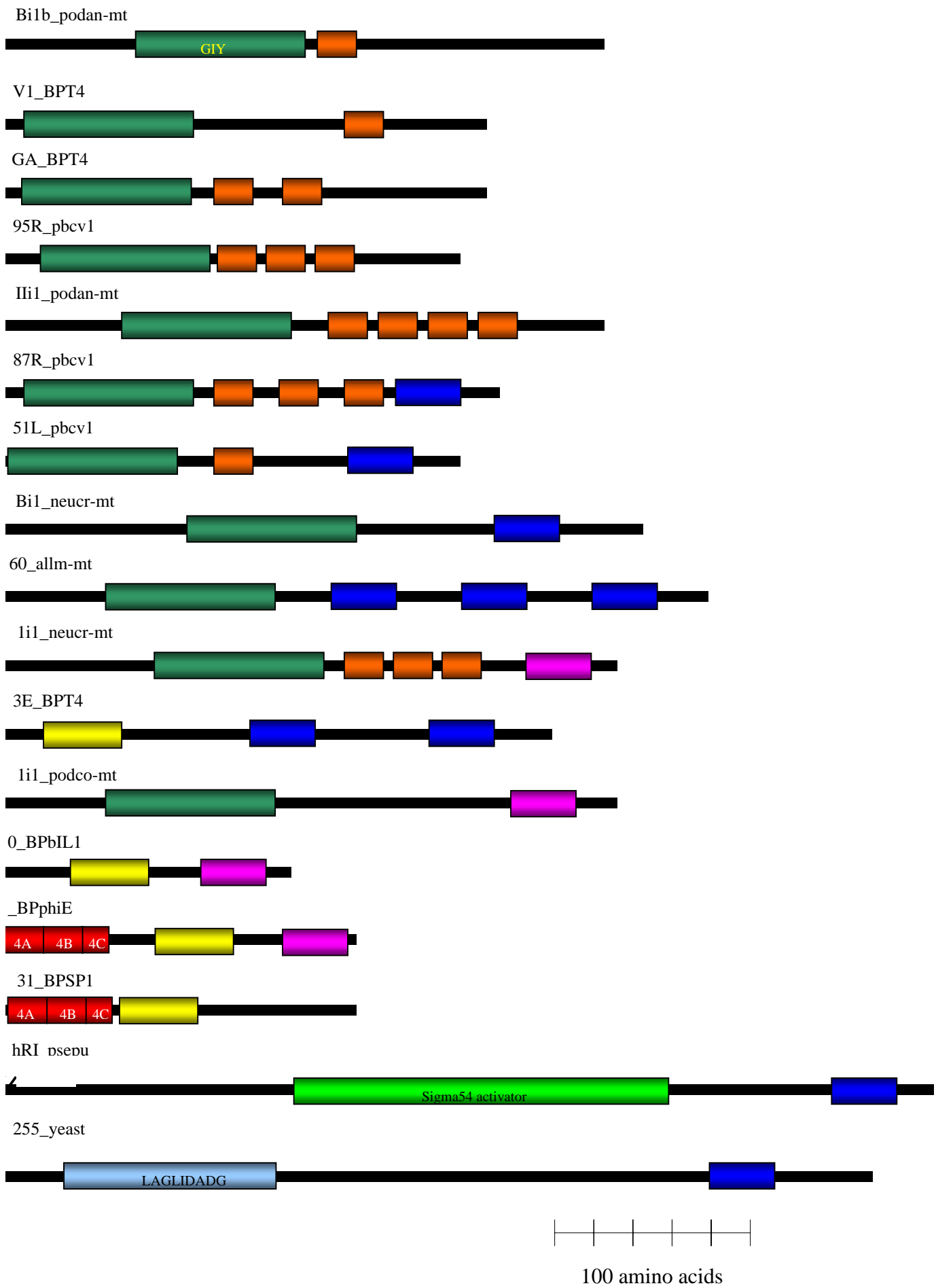


Figure 2

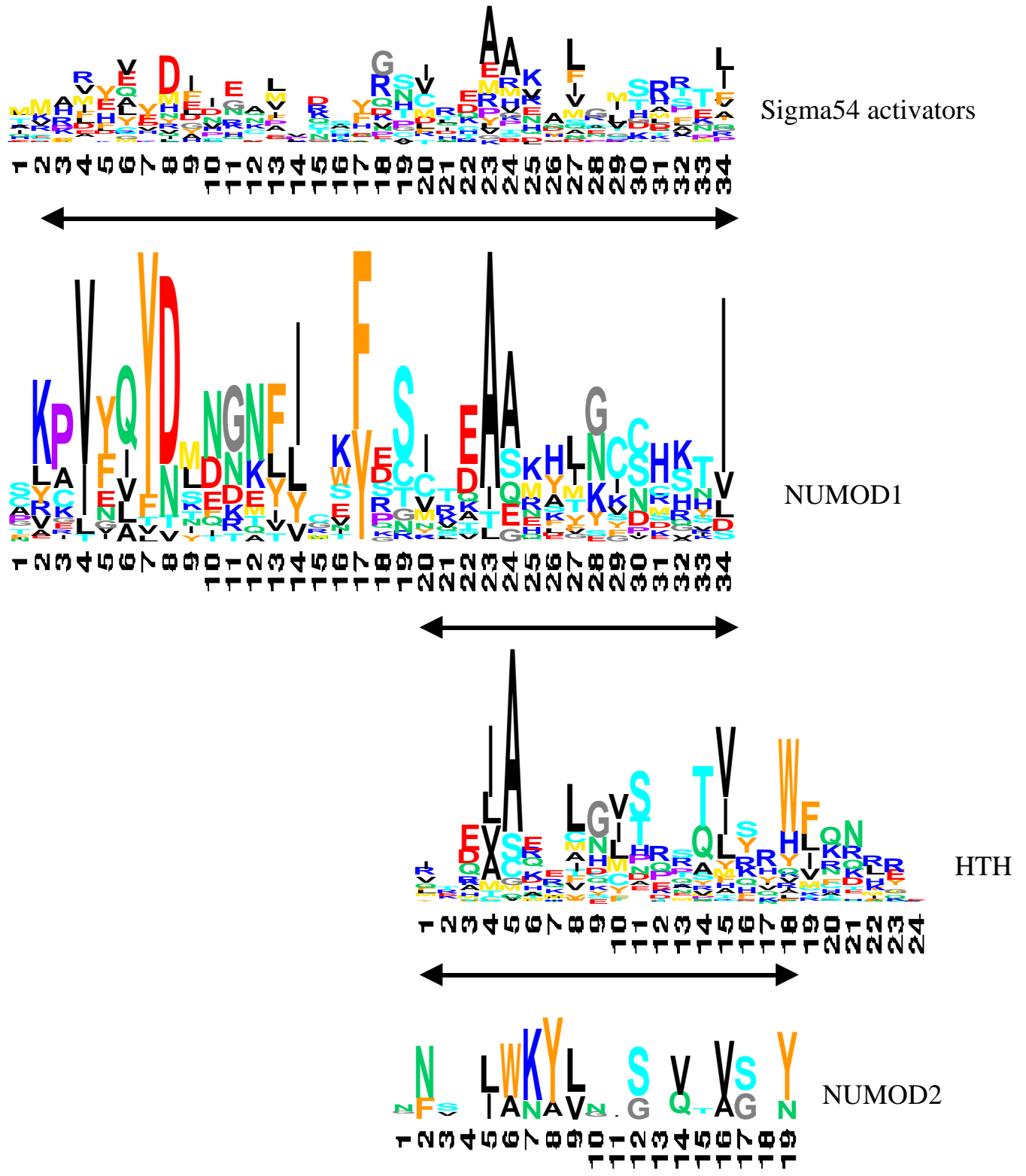
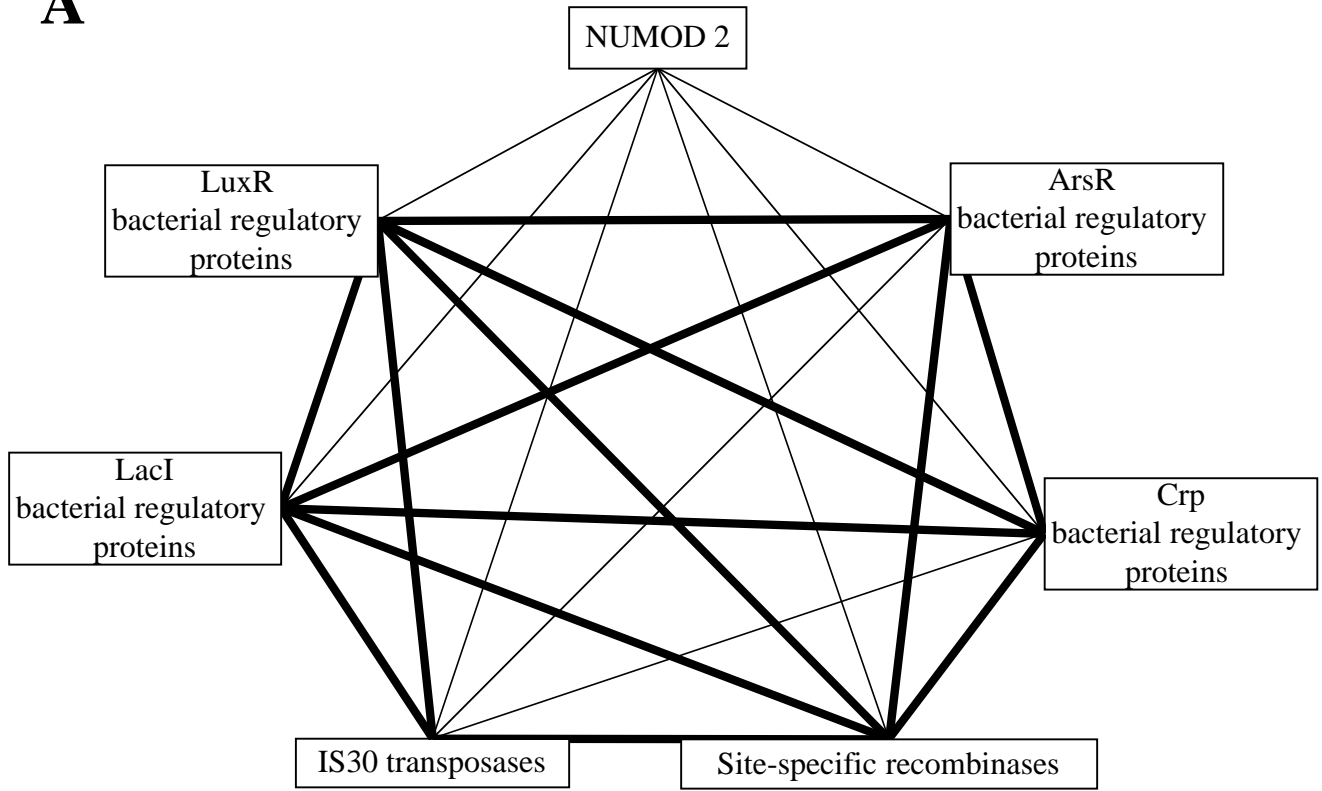


Figure 3

1 MIKKMKS^{YIN} MNSTV^{TTLGA} NYLVHTGYFN TISRLKVCTI ASYRYYSTSK
51 SDSQSSDLPP VPIFTINNLN NKDSIKSSRI LLKDKG^{GIYS FINTVNNNQY}
101 ^{IGSAKDFYLR LNEHLENKKS NIALQKAF^{TK} YGLDKFIFCI YEYFTYESKI}
151 ^{ISHKALTDLE TSYINRFNFD NLYNFKAIAT} ^{NUMOD3} SSLGYKHTEE ARLKMVDYYK
201 ^{NUMOD3} DKNNHPM^{FGK THTEEALGLI SKPGELNPMF} ^{NUMOD3} GKKHSEATKA SMSEKKNKYP
251 LGVGIYDLED ^{NUMOD2} NLILKFSNNV ELAKYLGISK VTVGKYLNSG LVYNKTYRFK
301 PIQD

Figure 4

A



B

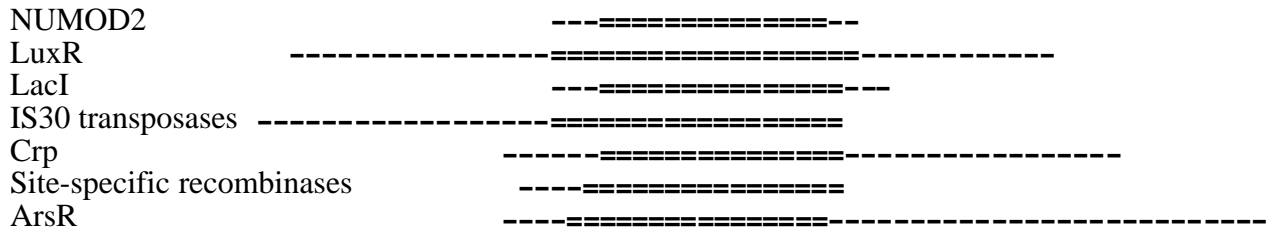
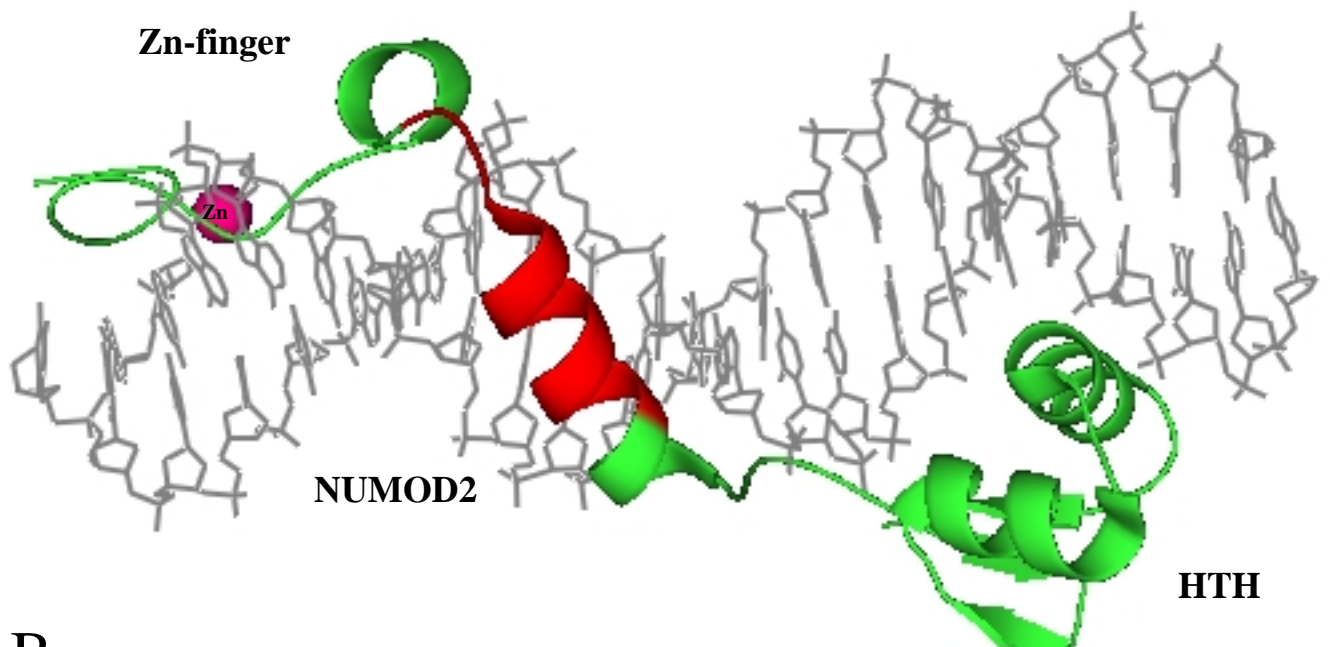


Figure 5

A



B

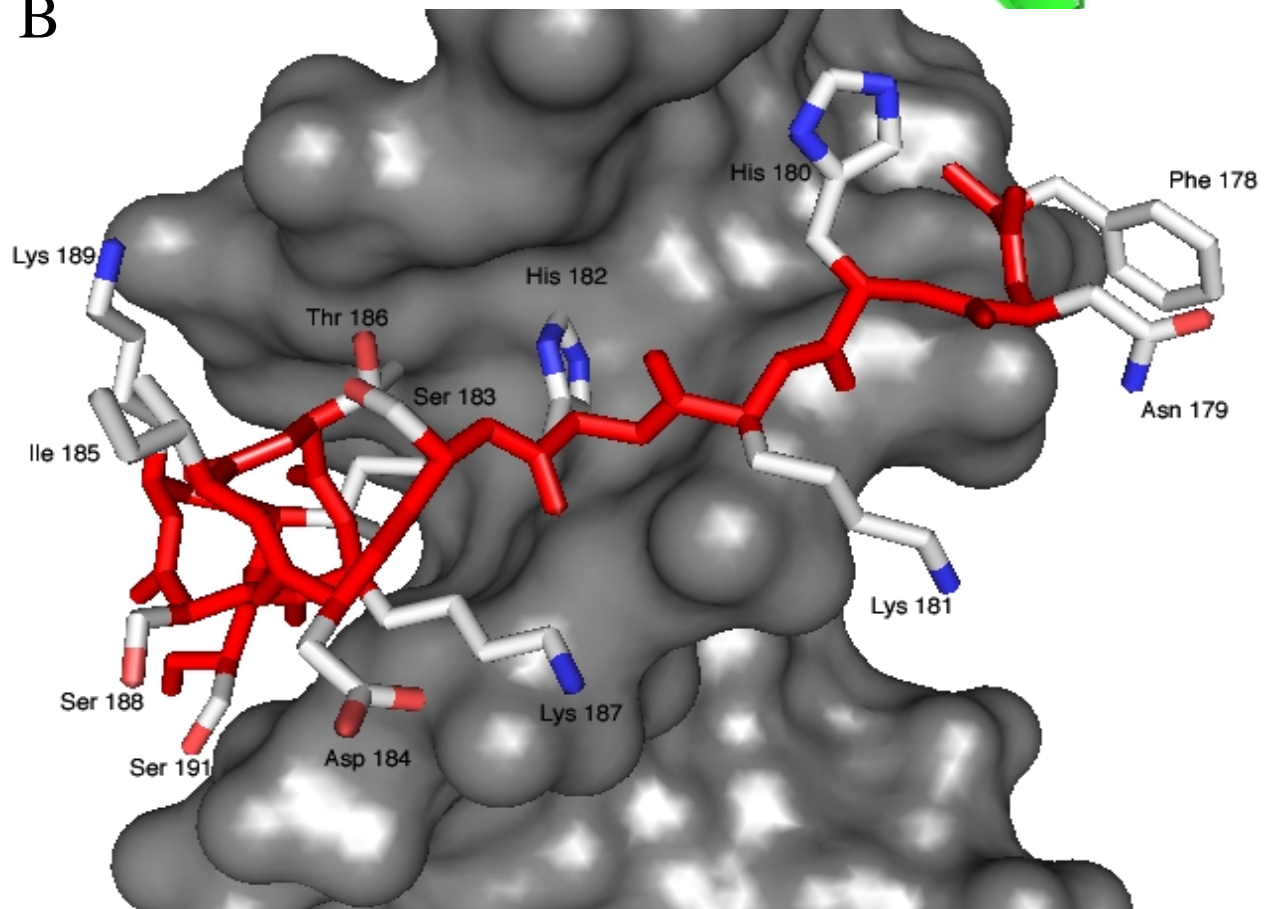


Figure 6