Project Name: Searching Pathogenic Motif-ome with special references to *Mycobacterium tuberculosis* and *Campylobacter jejuni*

Project Members:

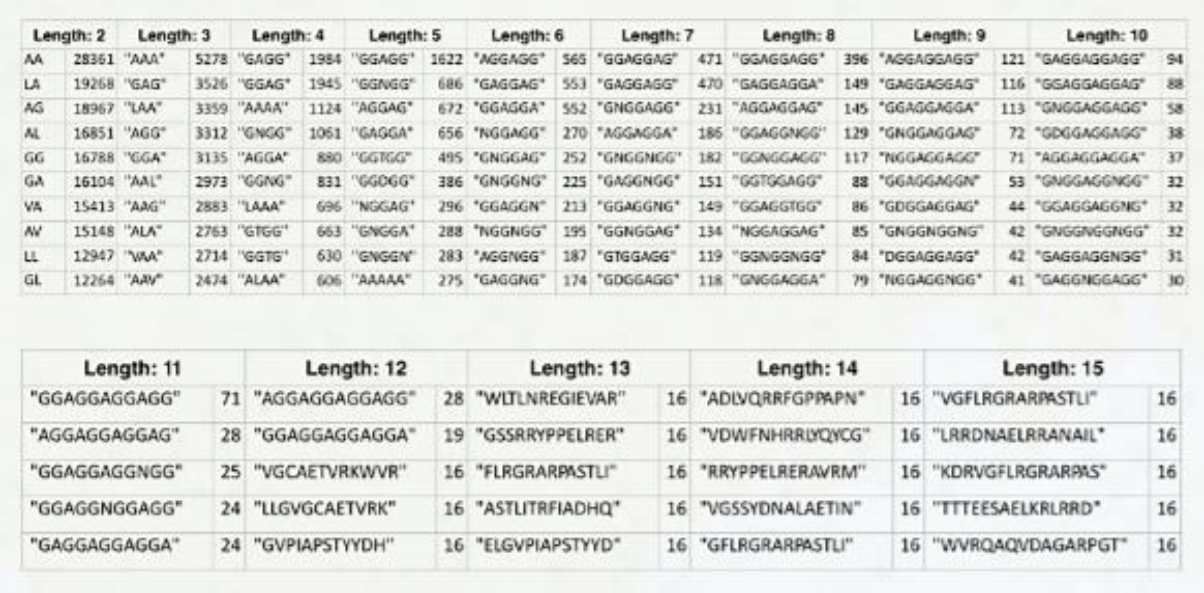
Mentor: Dr. Asani Bhaduri

Students: Ankit Pathak, Vani Singh

Other Team Member: Dr. Abhijeet Parmar

Project Abstract:

Pathogenic bacteria have been plaguing mankind since time immemorial. Although several genome and proteome of pathogenic microbes have been unearthed, the analyses are mostly dependent on a few model organisms. We have tried to find novel and hitherto unreported motif in several pathogen proteome utilizing brute-force and expectation-maximization algorithm. The protein sequence stretches of length 2 to 20 in pathogenic proteomes were scrutinised for repetition and subjected to analysis for finding common and unique repeat sequences subtracting the repeats from those observed in common model organisms like *Escherichia coli*, *Bacillus subtilis*. This process resulted in collection of several unique repeats which could be represented as unique protein motifs. Based on the functional, positional and/or structural similarities between the proteins that possess these unique repeats we are proposing several novel motifs. We are pursuing a special subset of these hypothetical motifs unique in *Mycobacterium tuberculosis* and *Campylobacter jejuni*.



Length: 2		Length: 3		Length: 4		Length: 5		Length: 6		Length: 7		Length: 8		Length: 9		Length: 10	
AA	28361	"AAA"	5278	"GAGG"	1984	"GGAGG"	1622	"AGGAGG"	565	"GGAGGAG"	471	"GGAGGAGG"	396	"AGGAGGAGG"	121	"GAGGAGGAGG"	94
LA	19268	"GAG"	3526	"GGAG"	1945	"GGNGG"	686	"GAGGAG"	553	"GAGGAGG"	470	"GAGGAGGA"	149	"GAGGAGGAG"	116	"GGAGGAGGAG"	88
AG	18967	"LAA"	3359	"AAAA"	1124	"AGGAG"	672	"GGAGGA"	552	"GNGGAGG"	231	"AGGAGGAG"	145	"GGAGGAGGA"	113	"GNGGAGGAGG"	58
AL	16851	"AGG"	3312	"GNGG"	1061	"GAGGA"	656	"NGGAGG"	270	"AGGAGGA"	186	"GGAGGNGG"	129	"GNGGAGGAG"	72	"GDGGAGGAGG"	38
GG	16788	"GGA"	3135	"AGGA"	880	"GGTGG"	495	"GNGGAG"	252	"GNGGNGG"	182	"GGNGGAGG"	117	"NGGAGGAGG"	71	"AGGAGGAGGA"	37
GA	16104	"AAL"	2973	"GGNG"	831	"GGDGG"	386	"GNGGNG"	225	"GAGGNGG"	151	"GGTGGAGG"	88	"GGAGGAGGN"	53	"GNGGAGGNGG"	32
VA	15413	"AAG"	2883	"LAAA"	696	"NGGAG"	296	"GGAGGN"	213	"GGAGGNG"	149	"GGAGGTGG"	86	"GDGGAGGAG"	44	"GGAGGAGGNG"	32
AV	15148	"ALA"	2763	"GTGG"	663	"GNGGA"	288	"NGGNGG"	195	"GGNGGAG"	134	"NGGAGGAG"	85	"GNGGNGGNG"	42	"GNGGNGGNGG"	32
LL	12947	"VAA"	2714	"GGTG"	630	"GNGGN"	283	"AGGNGG"	187	"GTGGAGG"	119	"GGNGGNGG"	84	"DGGAGGAGG"	42	"GAGGAGGNGG"	31
GL	12264	"AAV"	2474	"ALAA"	606	"AAAAA"	275	"GAGGNG"	174	"GDGGAGG"	118	"GNGGAGGA"	79	"NGGAGGNGG"	41	"GAGGNGGAGG"	30

Length: 11		Length: 12		Length: 13		Length: 14		Length: 15	
"GGAGGAGGAGG"	71	"AGGAGGAGGAGG"	28	"WLTLNREGIEVAR"	16	"ADLVQRRFGPPAPN"	16	"VGFLRGRARPASTLI"	16
"AGGAGGAGGAG"	28	"GGAGGAGGAGGA"	19	"GSSRRYPPELRER"	16	"VDWFNHRRLYQYCG"	16	"LRRDNAELRRANAIL"	16
"GGAGGAGGNGG"	25	"VGCAETVRKWVR"	16	"FLRGRARPASTLI"	16	"RRYPPELRERAVRM"	16	"KDRVGFLRGRARPAS"	16
"GGAGGNGGAGG"	24	"LLGVGCAETVRK"	16	"ASTLITRFIADHQ"	16	"VGSSYDNALAETIN"	16	"TTTEESAELKRLRRD"	16
"GAGGAGGAGGA"	24	"GVPIAPSTYYDH"	16	"ELGVPIAPSTYYD"	16	"GFLRGRARPASTLI"	16	"WVRQAQVDAGARPGT"	16

Figure 1. Putative motifs and corresponding number of occurrence in *M. tuberculosis*

>*M.tuberculosis* Rv3611
VAIANPAEPGAAGRHHQPRGDRKPRAWRQCGPQNGPRRSQAITPEPGAAGRHHQPRGDRK
PRAWRQCGPQNGPRRSQAITPEPGAAGRHHQPRGDRKPRAWRQCGPQNGPRRSQAITPEP
GAAGRHHQPRGDRKPRAWRQCGPQNGPRRSQAITPEPGAAGRHHQPRGDRKPRAWRQCGP
QNGPRRSQAITPEPGAAGRHWLDQRPVVPDGVGKSDS

Figure 2. A 19 amino acid long stretch repeated thrice in a M. tuberculosis protein

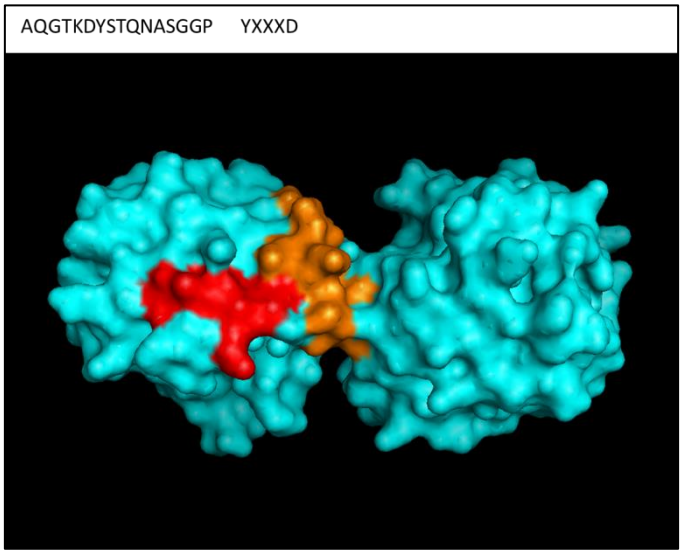


Figure 3. Two motif stretches shown in 3-D protein structure