This paper (written spring 2016) is the sum of my research for my two semesters working at the Columbia Fernandez Protein Biophysics Lab, which involved building a pipeline to recognize the CnaB motif (a certain sequence of amino acids) in proteins. See https://github.com/kpet123/String-Search-Related-Bioinformatics-Programs for the code for this pipeline. I also built a phylogeny tree to show possible evolutionary paths for the CnaB motif. This research was presented at the Museum of Natural History.

Identification of Putative CnaA and CnaB Adhesion Domains and Investigating their Evolution

Kaitlin Pet, Dan Echelman, Julio Fernandez

Abstract

Adhesion is very important for bacterial survival because they often exist in mechanically challenging environments. Bacteria have evolved many methods to adhere to surfaces, including both protein and polysaccharide extensions to help them attach (Hori 2010). This paper will focus on protein adhesins, specifically the CnaA and CnaB protein subunits. Because these two subunits contain intramolecular isopeptide bonds, they have sections that are mechanically inextensible when subjected to nanonewton-scale forces, which could hold an evolutionary advantage in adhering to surfaces. (Echelman 2016). Proteins containing these two subunits have thus far been found in the surface adhesion proteins of pathogenic Gram-positive bacteria, and thus studying these attachments could yield the creation of therapeutic applications (Greene 2014). We sought to characterize the evolutionary range of these two subunits by using bioinformatics methods to do a blanket search for homologous sequences then test if those sequences had CnaA/CnaB-specific sequence motifs. Though we are still in the process of characterizing a CnaA-specific motif, we have created a library of putative CnaBs from across the tree of life, including many pathogens with previously uncharacterized CnaBs. We have also refined our experimental system to allow for higher-throughput screening of putative CnaBs/CnaAs by simplifying the process of synthesizing proteins that can be tested in single molecule force spectroscopy for presence of the formed isopeptide bonds characterizing those subunits.

Introduction

One of the most important reasons that bacterial species are such a ubiquitous part of our planet is their ability to adhere strongly to a variety of substrates. These substrates range from biofilms (bacteria adhering to each other) to teeth and other human tissues. (Hori 2010). Because environments such as the human body are mechanically harsh (Echelman 2016), bacteria have developed long, protein based adhesion molecules in order allow them to both pass the energy barrier involved in sticking to a surface and adhere to that surface in the face of mechanical perturbations (Hori 2010). By investigating these adhesion molecules from a mechanical perspective, we can get a better idea of the morphology of earliest colonizers of this earth and open the possibility of developing methods to prevent undesirable adherence of bacteria on surfaces.

The CnaA and CnaB subunits are two characterized subunits found in Gram positive bacteria surface adhesion proteins. When subject to mechanical force, proteins with these subunits show a very interesting property: intramolecular isopeptide bonds that act as a “mechanical short” when the subunit is being pulled. Isopeptide bonds are covalent bonds distinct from the peptide bonds forming the protein backbone. They are observed to form between a lysine and

an aspartic acid residue and their assumed function is to both stabilize the adhesion molecule (Wang 2013) and protect domains behind the covalent bond from being exposed to the high mechanical force characteristic of the bacteria's living environment. (Echelman 2016) The difference between these two domains lies in the position of the isopeptide bond. In CnaBs, the lysine and aspartic acid are at opposite ends of the domain, causing the entire domain to be inextensible behind the isopeptide bond. In CnaAs, on the other hand, there is a good chunk of domain not behind the mechanical short. Referred to as an "isopeptide delimited loop" (IDL), the suggested function of this subunit is to act as a shock absorber when bacteria are suddenly pulled with high force (Echelman 2016). Adhesion proteins containing these subunits usually contain a combination of the two; SpaA, for example, consists of a CnaB closely followed by a CnaA. (See "List of solved CnaA and CnaB proteins" in supplemental) The goal of this project was to a) use characteristic amino acid motifs of CnaA and CnaB amino acid sequences to generate a library of putative CnaAs and CnaBs, then build those proteins to confirm the presence of an isopeptide bond, and b) to use those sequences to create a phylogenic tree.

Materials and Methods

BioInformatics

Analysis was performed using custom code written using python (especially the Biopython Package (Cock, 2009)) and BASH (UNIX). Program code and detailed explanations are available https://github.com/kpet123.

CnaB Library Creation

CnaB with known crystal structure were found to possess a clearly-characterized motif, henceforth known as the "ebox motif", illustrated in Figure 1B. The position of each ebox motif in the 23 crystal structures was first deduced (See Figure 1A for visual), then each solved CnaB was submitted to BLAST search to recover homologs. All homologs were then tested for presence of the ebox motif in a corresponding position to their alignment (Figure S1 for visual aid). All discovered homologs fitting that criteria were inputted into a CnaB library text file. See Pattern.py and makeLibrary_oo.py for code.

CnaA Motif Search

Several potential CnaA motifs, some found in (Kang 2007) were tested on the set of all crystalized CnaAs but no matches were found. Testing motifs against potentially smaller evolutionary units was attempted, i.e. searching in groups based on the number of CnaAs each protein contained, but at this point no motifs have been found. See Future Directions for how we will proceed in the CnaA motif search.

Primer Design

A program was written to create PCR primers. Input is the protein sequence then DNA sequence in FASTA format. Output is every sequence under 40bp that ended with one guanine, each potential primer's AT content, and predicted melting temperature of the primer, which was calculated by summing AT content*2 and CG content*4. See makeprimer.py for code. This will be used when creating primers testing of putative library proteins.

Conversion to Object Oriented Code

To allow for modularity and quick modifications based on system changes increased, certain programs such as the library-generator (in makeLibrary.py) and motif searchers (in Pattern.py) were modified to object-oriented code so existing methods work for more types of data and different situations. "Object-Oriented Programming" is a way of programming such that one

creates 'objects' (units of code) with changeable parameters that serve a particular type of function. For example, the code searching the CnaB motif could initially search only for that particular motif sequence, and was implemented by testing if the sequence of that specific motif matched any point in the CnaB homolog sequence being searched. However, as focus shifted toward finding a motif characteristic of the CnaA subunit, a need arose to quickly test many possible motif sequences on the set of known CnaAs. Therefore, a general "Motif" object was created that would take a flexible motif as a parameter and find instances of that motif in a query sequence. The motif was flexible in 2 ways. First of all, notating 'x' within the motif would allow any amino acid to be recognized as part of the motif, thereby allowing the use of gapped motifs such as LPXTG. Second, it allowed a certain position in the motif to be occupied by more than one amino acid. For example, the CnaB motif's first position was observed to either by phenylalanine or tyrosine, so the first "Motif" position was capable of encompassing two possibilities.

Molecular Biology

A plasmid construct containing four I27s (an immunoglobin domain that shows a distinctive and well-characterized pattern when exposed to single-molecule force spectroscopy) with restriction sites was created (see Figure 2 up to step 2). XL10 *E. Coli* cells were used to replicate plasmids. Further processing of this plasmid will be outlined in "Future Directions"

Results

CnaB Library Creation

BLAST searching known CnaB sequences with a very low homology cutoff (e-value of 20) and maximizing hits to 10,000 per query, 17,659 homologies were matched and of those, 10944 had corresponding eboxes in their alignment. We did not worry as much about the low level of homology because anything without a matching ebox would be automatically discarded. In fact, we were most interested in proteins with low homology because there was a desire to discover "edge cases" of CnaBs and cover the whole spectrum of the existence of this subunit.

Before our search, CnaBs have only been characterized in *Firmicutes* and *Actinobacteria*, both Gram-positive bacteria phyla. However, our library contains putative CnaBs from a much wider range or organisms, including pathogenic Gram-negative bacteria like *Chlamydia* and bacteria with usual surface composition like *Mycoplasma*, which has a three-layered cell membrane but no cell wall (Domermuth, 1964). Non-pathogenic Gram-positive bacteria, such as *Lactobacillus Brevis,* were also found to have putative CnaBs. (See Figure 3 for tree of all phyla contained in library)

A few especially interesting interesting hits were CnaBs not in bacteria. These included *Methanobrevibacter ruminantium*, an Archea, and *Trichuris trichiura,* a nematode. It is possible the *Trichuris* hit resulted from experimental error; when BLASTED, the *Trichuris* protein had 99% homology to a protein from the Gram-positive bacteria *Enterococcus faecium*. Nonetheless, the huge range of organisms containing putative CnaBs is a testament to the potential usefulness of intramolecular isopeptide bonds to create inextensible protein domains.

Discussion/ Future Directions

CnaA Motif recognition

As mentioned earlier, a CnaA motif has been elusive. Since the CnaAs have a specific mechanical phenotype (they contain both a) an extensible section and b) an inextensible section due to an intramolecular isopeptide bond) it is likely there are specific protein sequences other

than the observed isopeptide residues K and N that contribute to that structure. At this point, there are two possible avenues to continue investigating in regards to ways of identifying putative CnaAs.

The first possibility is a further refining of the direct-motif-search approach that can be done by making modifications to the currently existent “Motif” program to allow to create more general motif. This could be implemented by a) interfacing with Biopython’s amino acid side chain type recognition (e.g. aliphatic, hydrophilic, etc.) and b) utilizing consensus motifs instead of absolute ones (e.g. a match would be recognized if a high percentage of amino acids fit a certain motif).

The second possibility is looking through the current CnaB library for CnaA subunits. Since CnaA and CnaBs usually exist in conjunction with each other, it would be very simple to find putative CnaAs in this way. A putative CnaA was already discovered when looking at the Methanobrevibacter DNA, as a section of the protein showed high homology to the FimA CnaA and had catalytic residues in the same location on an alignment. [data not shown]

Building Proteins to Test for Isopeptide Bonds

When expressing proteins to be tested using single-molecule force spectroscopy, the usual method of synthesis is a multi-step process. By creating a plasmid with a Bgl2 restriction enzyme cut site flanked on both sides by two I27s, we have reduced this workload to one step- ligating in DNA of the protein-of-interest flanked with Bam and Bgl2 cut sites (Bam and Bgl sticky ends can join together and form a Bsty cut site). Using this method, we can more quickly express putative CnaBs to test for the presence of an isopeptide and confirm whether CnaBs do in fact have such a large spread in the tree of life.

Acknowledgements

Thanks to everyone in the Fernandez Lab for creating such a warm and kind environment. Especially thanks to Dan Echelman for teaching me all the lab techniques and taking the time to make sure I actually knew what I was doing.

Statement on what work was not done by Kaitlin-Discovery of CnaB motif, compilation of solved proteins with CnaA and CnaB domains, and developing the higher-throughput method of expressing proteins to be analyzed by single molecule force spectroscopy was done by Dan Echelman.

References

Cock, Peter J. A., et al.(2009). “Biopython: freely available Python tools for computational molecular biology and bioinformatics”. Bioinformatics **25**(11): 1422–1423 . doi:10.1093/bioinformatics/btp163

Domermuth, C., et al. (1964). "Ultrastructure of Mycoplasma Species." Journal of Bacteriology **88**: 727-744.

Echelman, D., et al. (2016). "CnaA domains in bacterial pili are efficient dissipators of large mechanical shocks." Proceedings of the National Academy of Sciences **113**(9): 2490-2495.

Hori, K. and S. Matsumoto (2010). "Bacterial Adhesion: From Mechanism to Control." Biochemical Engineering Journal **48**(3): 424-434.

Kang, H., et al. (2007). "Stabilizing Isopeptide Bonds Revealed in Gram-Positive Bacterial Pilus Structure " Science **318**(5856): 1625-1628.

Madden T. The BLAST Sequence Analysis Tool. 2002 Oct 9 [Updated 2003 Aug 13]. In: McEntyre J, Ostell J, editors. The NCBI Handbook [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2002-. Chapter 16. Available from: http://www.ncbi.nlm.nih.gov/books/NBK21097/

Sievers F., et al. (2011) “Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega” Molecular Systems Biology 7 Article number: 539 doi:10.1038/msb.2011.75

Wang, B., et al. (2013). "Isopeptide bonds mechanically stabilize spy0128 in bacterial pili." Biophysics Journal **104**(9): 2051-2057.

Zhang, X. H., et al. (2003). "Sequence Information for the Splicing of Human Pre-mRNA Identified by Support Vector Machine Classification." Genome Research **13**(12): 2637-2650.

Figure 1

A

```
tr|Q6NF81|Q6NF81_CORDI            DGKAVLSGIHLGTLQLESNVMKYTDAW---AGKGTEFCLVETATASGYELLPKPVIVKLE
sp|P18481|TEE6_STRP6              ----ISTQVSSGKYKIKELKA--------------PK---------GYSLNTETYEITAN
tr|Q81D71|Q81D71_BACCR_SECOND     NGVIKWSNIPYGDYQIFETKA--------------PTYTKEDGTKTSYQLLKDPIDVKIS
tr|Q84A41|Q84A41_STRAG            DGIITITGLKEGTYYLVEKKA--------------PL---------GYNLLDNSQKVILG
tr|B3FNT1|B3FNT1_STREE_FIRST      KIQKATEDIFSG------------VAY----GHAGEYVYDVAEAKTGWQAITKNGKTI--
sp|P18477|FM1_ACTVI_FIRST         GDVVKAGALKSTTVQKITTGANGLASFTDAQTEVGAYLVSETRTPDKVIPAEDFVVTLPM
tr|Q6NK05|Q6NK05_CORDI_FIRST      ADA---KGHETSTTKEVETSGNGTAVFD--NLDLGIYLVEETKAPDGIVTGAPFIVSIPM
tr|B3FNT1|B3FNT1_STREE_SECOND     GETFTVEQLP----------------------AGSKYTVTET-GVAGYTDSSIYTT---N
sp|Q53654|CNA_STAAU_FIRST         K--Y---DEG----------------------KKIEYTVTED-HVKDYTTDI--------
sp|Q53654|CNA_STAAU_SECOND        E--K---AKG----------------------QQVKYTVEELTKVKGYTTHVDNND---M
tr|A7KT39|A7KT39_STREE_FIRST      PQTFKLSGAMPATAMKKLTEA-EGAKFNTANLPAAKYKIYEIHSLSTYVGEDGATLT---
tr|A7KT39|A7KT39_STREE_SECOND     KNTVTVNGLDKN----------------------TEYKFVER-SIKGYSADYQEI-----
tr|B9UQT9|B9UQT9_STRAG_FIRST      SAKATAATSFKHTFEN-L-------------DNAKTYRVIE--RVSGYAPEYVS------
tr|Q8G9G1|Q8G9G1_STRPY            DGQVKDFYLMPG----------------------KYTFVETAAPDGYEVATAI----TF
tr|A7KT39|A7KT39_STREE_THIRD      QGRFEITGLLAG----------------------TYYLEETKQPAGYALLTSRQ---KF
tr|B9UQT9|B9UQT9_STRAG_SECOND     KGQFEITGLTEG----------------------QYSLEETQAPTGYAKLSGDV---SF
tr|O68212|O68212_ACTNA            QGAINVKGLFISDSID-GANR-D----NQKDATARCYVLVETKAPAGYVLPAGDG---AV
sp|P18477|FM1_ACTVI_SECOND        QGTVEINYLRANDYVN-GAKK-D----QLTDE--DYYCLVETKAPEGYNLQADPL---PF
tr|Q6NK05|Q6NK05_CORDI_SECOND     DGTFTIDGLHVTDFED-GKEA-A----P----ATKKFCLKETKAPAGYALPDPNVTEIEF
tr|Q8E0S8|Q8E0S8_STRA5            SGHIRISGLIHG----------------------DYVLKEIETQSGYQIGQAET---AV
tr|Q8E0S9|Q8E0S9_STRA5            DGTFEIKGLAYAVD----ANA-E----G----TAVTYKLKETKAPEGYVIPDKEI---EF
tr|Q81D71|Q81D71_BACCR_FIRST      NGHIRVQGLEYG----------------------EYYFQETKAPKGYVIDPTKR---EF
```

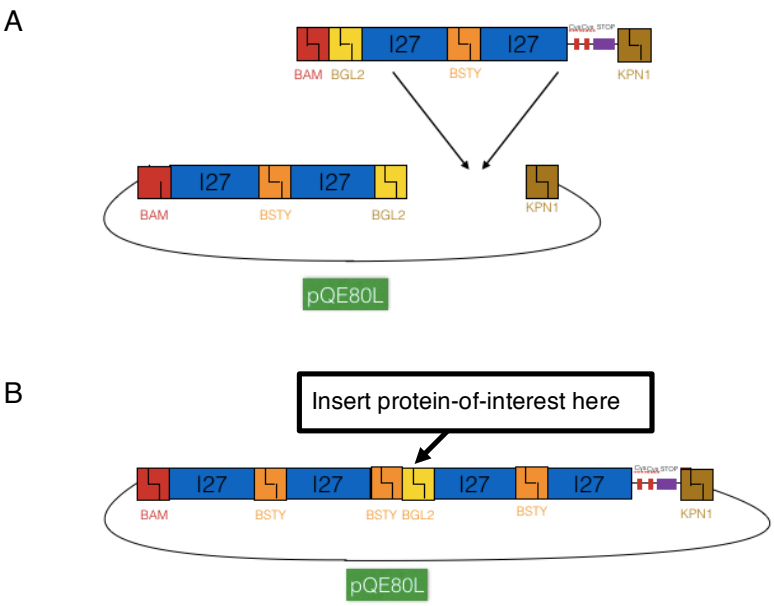
B

CnaB Motif

F	X	I	X	E
Y		L		
		V		
		F		

Figure 1. A: Clustal Alignment of solved CnaBs with ebox motif highlighted in yellow. B: Ebox motif, where each column is a list of the amino acids possible at that position and ‘x’ represents any amino acid.

Figure 2



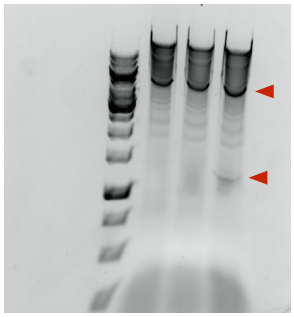


Figure 2. A: Insert of two additional I27 domains with restriction sites into pQE80L plasmid with two I27s already. B: Completed construct. BAM and Bgl2 sites combine to form a Bsty site. Protein of interest will be inserted at the Bgl site. C: Gel showing plasmid in (B) digested with Bam and Kpn1. Top red arrow points to plasmid, and bottom area is construct with 4 I27s

Figure 3

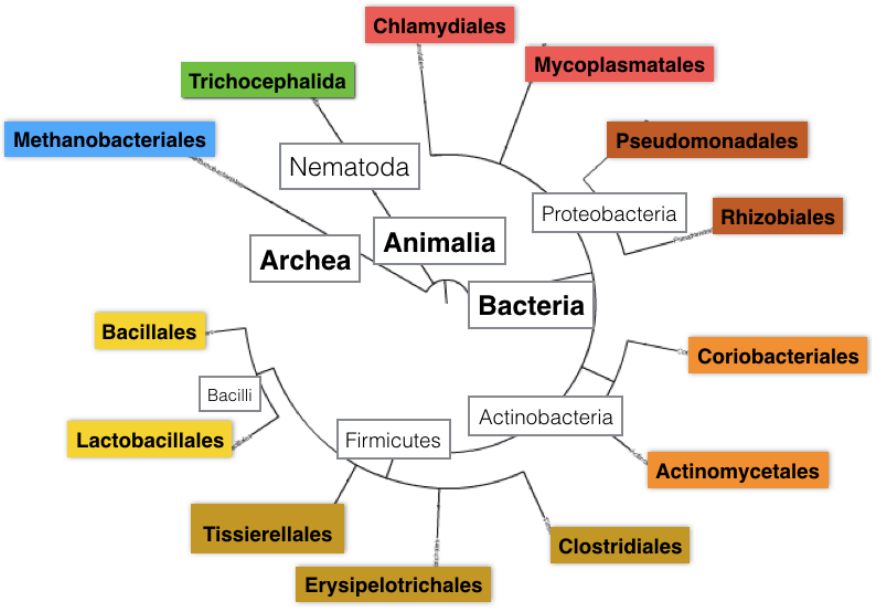


Figure 3: Phylogenic tree of all phyla represented in the CnaB library. Though more than half the phyla are, like known CnaBs, from *Actinobacteria* or *Fimicutes*, the other half includes Gram-positive bacteria, animals, and archea.