Identification of Conserved Regions in CRISPR Protein Family

02-712 Final Project

November 27, 2016

Christine Baek christib@andrew.cmu.edu
Qi Chu qchu@andrew.cmu.edu
Yanyu Liang yanyul@andrew.cmu.edu
Department of Computational Biology, Carnegie Mellon University

Abstract

The abstract goes here.

1 Introduction

Project is about the evolution and relationship of various *Cas* (CRISPR-associated) proteins. CRISPR system works by base-pair recognition of foreign genetic material and subsequent nuclease activity on the non-self genome. This has been adopted for the purpose of genetic engineering, and its cleavage based on base-pairing (as opposed to protein-DNA recognition of ZNF or TALENs) provide improved accuracy and reduced costs (no protein engineering involved). CRISPR is as diverse as the species that utilize them, but ultimately have the same function. While *Cas9* (isolated from *Streptococcus pyogenes*) currently the protein of choice for such applications due to smallest number of involved components, it would be beneficial to study the other Cas proteins as well since *s.pyogenes Cas9* is limited in terms of PAM (Protospacer Adjacent Motif), of `-NGG` or to be used in conjunction with *s.pyogenes Cas9*.

Cas9 currently is predominantly used for genome engineering because it's only 1 protein that you have to worry about, and was the first CRISPR protein that was successfully adapted for genome engineering. Now that we know more about the CRISPR system and somewhat better understanding of what each domain does, it may be beneficial to explore using some of the other CRISPR proteins to achieve same goal. Cas9's big size (1400AA residues) has been source of concern for therapeutic applications. Also, using different Cas proteins potentially offer greater choice of PAM sequence. By identifying conserved motifs between different CRISPR proteins, we can hopefully identify regions of corresponding activity in those CRISPR proteins, and compare to *Cas9* which has been studied in greater detail in terms of structure [9] or function [5] compared to other *Cas* proteins.

In this paper, we employ 3 different approaches to identify motifs, or conserved regions of significance in terms of *Cas* function for better understanding of the *Cas* protein family and mechanism of each

component.

1.1 CRISPR/Cas

CRISPR(Clustered regularly interspaced short palindromic repeats) is a microbial adaptive immune system. While bacteria and archaea utilize CRISPR system to store foreign genetic material to distinguish self vs non-self, this system has been adopted and exploited by scientists since 2012 genome engineering tool, as discussed in [6] and [7]. CRISPR initially began as next-generation tool to replace ZNFs and TALENs, it has since then been modified for non-genome engineering purposes such as CRISPRi [8]. There also have been attempts to reduce off-target effects by modifying the nuclease domain [10] .

Naturally in bacteria or archaea, CRISPR proteins have distinct roles in the three phases of CRISPR system as follows :

1. Acquisition : foreign genetic material enters microbe, which is cut by CRISPR protein and inserted into CRISPR array. This fragment is now a *spacer* separated by *repeats*, hence the name
2. Expression : CRISPR array, which include multiple spacers separated by repeats, is expressed as a single RNA. This is then cleaved into individual units known as *crRNA*, which contain single spacer. *crRNA* forms complex with one or more CRISPR proteins (depending on the CRISPR system)
3. Interference : upon recognition of specific foreign genetic material via base-pairing with the spacer in crRNA, CRISPR protein in complex with the spacer cleaves the foreign genetic material.

1.2 Past Approaches

Functionally related regions can be clustered by evidences in experimental data. As summarized in Figure. 2. Previous work has found conserved regions on the sequence level using sequence alignment and structural information [**?**] inside each sub-type of the *Cas* system but not across the whole *Cas* family.

1.3 Goal of Paper

In this paper, we would like to answer the question that whether or not the proteins in Cas family share some sequence level similarity. And more specifically, as Cas9 is a multi-domain protein with multiple functions and each function can be achieved by other single-function protein in Cas family, we would like to explore if we can map such functional domain similarity on the basis of primary sequence similarity between Cas9 and other cas proteins. Informally, we tend to solve the following problem:

- **Input**: Two sets of sequences $C_1 = \{p_1, ..., p_m\}$ and $C_2 = \{q_1, ..., q_n\}$
- **Output**: A set of regions $R = \{r_1, ..., r_k\}$ such that r_i *occurs* and is *representative* in both C_1 and C_2

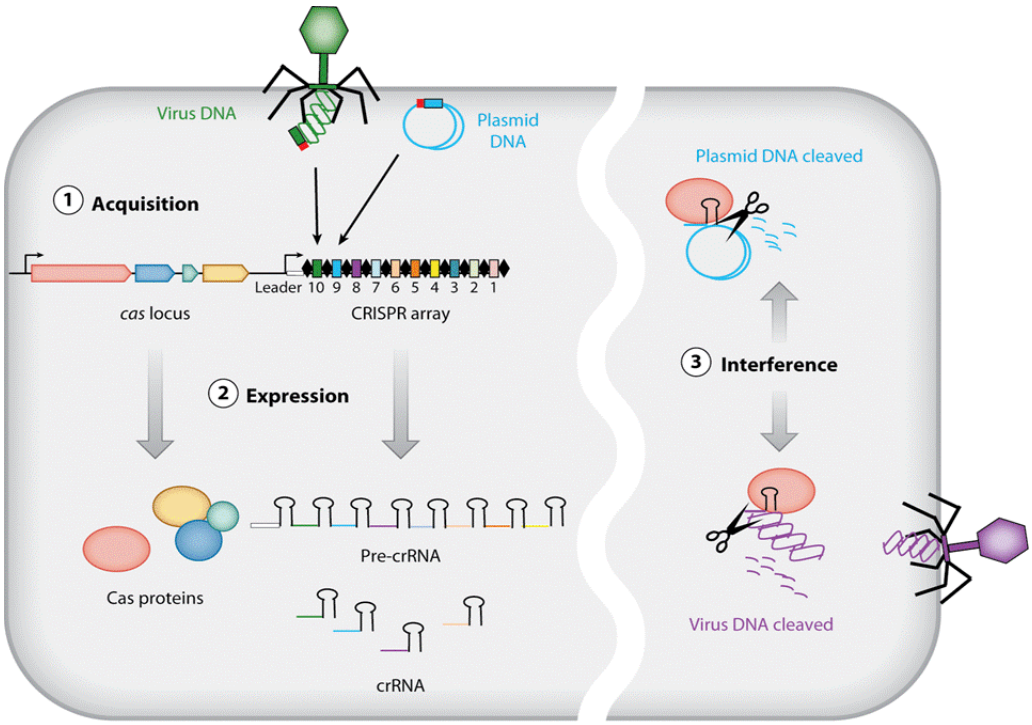

Figure 1: Overview of CRISPR proteins and their function [4]

Here *occurs* and *representative* can be explained in different ways under different strategies. In the following section we will discuss the two strategies we proposed to solve this problem.

1.4 Approaches in This Paper

In this paper, we propose the following two strategies: i) alignment; ii) motif finding. First of all, our problem is naturally a multiple sequence alignment problem. Notice that other Cas proteins is like a substring of Cas9, then to recognize such local similarity, both semi-global alignment and local alignment is suitable in this case. And here *occurs* and *representative* mean that r_i is optimal in alignment.

Besides profile-based local alignment or semi-global alignment techniques, an alternative approach is to make use of motif information. Motif finding problem is defined as to find representative pattern in a collection of sequences. Following this idea, we can first find motif in C_1 and perform pattern recognition in C_2. The motif found in C_1 carries the *representative* signature of C_1 and the recognition in C_2 tests whether it meets the requirement to be *occur* in both C_1 and C_2. Furthermore, for motif analysis, we propose two widely used methods: i) Gibbs sampling; ii) Hidden Markov Models.

2 Methods

All code and output files are available on `https://github.com/cookie223/CAS_project`.

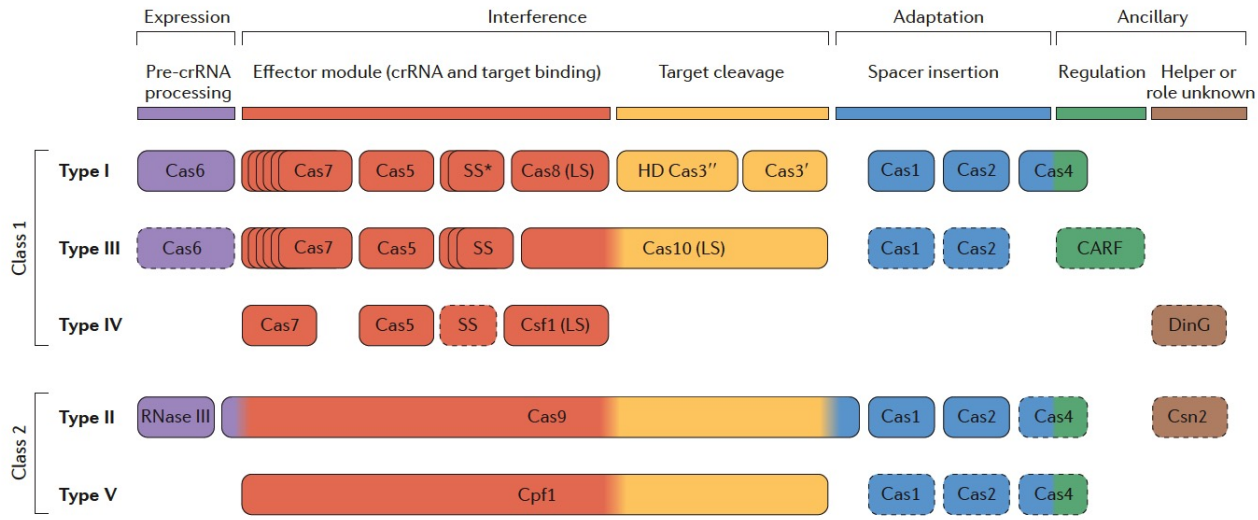

Figure 2: Conserved building blocks of *Cas* family proteins [**?**]

Any reference to files in this report indicate filepath based on root of the repository.

In this paper, we use 3 different approaches, each with different strengths and limits as for discovering patterns and information from multiple related sequences. Each method has its own section which discusses overview of algorithm/model, pros and cons of given model, detailed protocol and parameters, and analysis performed.

Protein sequences were used (as opposed to DNA), to uncover preservation of *Cas* protein's functional motifs. Protein sequence analysis much more appropriate for such purpose than DNA sequence especially for distantly related sequences.

2.1 Data Retrieval

Gene sequences for *Cas* family including *Cas1* through *Cas10* were searched and downloaded from NCBI. For each gene, a variety of species were selected to have all the sequences of the proteins in the *Cas* system of one species and also maintain a certain level of variety. The selection is also subject to the availability in NCBI. For domain-specific profile HMM, domain markers were fectched from EMBL-EBI (http://www.ebi.ac.uk).

2.2 Sequence Alignment using Dynamic Programming

Semiglobal ailgnment using Needleman-Wunsch [2] and Local Alignment using Smith-Waterman Algorithm [3] were implemented for pairwise sequence comparison. Code available at `https://github.com/cookie223`

2.2.1 Model & Algorithm Overview

Semiglobal alignment and local alignment were performed on various *Cas* sequences. Because certain families of *Cas* proteins are composed of multiple genes, it is impossible to do a global alignment,

or multiple sequence alignment with all *Cas* sequences. Instead, *s. pyogenes Cas9* was used as a reference sequence, to which all other *Cas* sequence was aligned to in pairwise sequence alignment.

2.2.2 Pros and Cons

Because alignment were done against *s. pyogenes Cas9* rather than a progressive alignment, for *Cas* sequences highly divergent from *s. pyogenes Cas9* may be aligned to an inaccurate site. Each alignment is guaranteed to return the global maximum or the most optimal alignment with given parameters, which may or may not be the actual corresponding motif. Also, this method assumes site independence, and does not discriminate conserved regions (which motifs would likely be part of) as opposed to fast-evolving regions.

2.2.3 Protocol

Following parameters were used for sequence alignment :

- Scoring Matrix = BLOSUM62 (BLASTP default)
- Affine Gap Penalty = -10 (BLASTP default is 11)
- Gap Extension Penalty = -1 (BLASTP default)
- End Gap Penalty (for Semi-Global only) = -3

Gap opening / affine gap penalty was slightly lowered to relax requirement for opening gap, as this experiment is for identifying local regions rather than strict sequence search.

2.2.4 Analysis

After each *Cas* sequence was aligned to *s. pyogenes Cas9*, it was then analyzed for the following values :

- Start position (row, col) of traceback
- End position (row, col) of traceback
- Alignment score (based on parameters discussed in section 2.2
- Average Score per base : alignment score / number of bases in alignment
- % Sequence Aligned : length of alignment / length of query (non-reference) sequence

This was done for both semi-global and local alignment outputs.

2.3 Gibbs Sampling

2.3.1 Model & Algorithm Overview

Gibbs sampling approach is based on position-specific scoring matrix [12], or PSSM, is one of the ways to model a motif. Suppose we are working on sequence set with Σ as alphabet and the length

of the motif is w. Then PSSM is a $|\Sigma|$-by-w matrix with entry:

$$S_c(A) = \log \frac{\Pr(A \text{ is at } c\text{th column}|\text{motif})}{\Pr(A \text{ is at } c\text{th column}|\text{background})}$$

, where $S_c(A), A \in \Sigma$ is the score of alphabet A appearing at cth position in the motif and the score is a log odds ratio.

Under PSSM setup, the motif is ungapped with a fixed length and each position is scored independently to each other. PSSM provides a way to parameterize motif and, furthermore, the motif finding problem can be cast as an optimization problem as follow:

$$\max_{S,o} \sum_{i=1}^{n} \sum_{c=1}^{w} S_c(q_i[o_i + c]) \tag{1}$$

, where $q_i[x]$ is the xth character of ith sequence in the collection and o_i indicates the starting site of the motif for ith sequence. With various size of the sequence set, this problem is NP-hard and [13] has proposed a Gibbs sampling approach to solve it.

Note that with known o, solving for S is reduced to a maximum likelihood estimation, which is trivial to solve, so the core of this problem is to find the optimal $o^\star$. Let S^o denote the scoring matrix induced by o and $f_S(o)$ denote the objective in Equation (1) (with S as scoring matrix), then we can encode the probability distribution of o according to $f(o)$:

$$\Pr(o) \propto e^{f_{S^o}(o)}$$

Gibbs sampler can be used to sample from this distribution with transition probability as follow:

$$\begin{aligned}
&\log q(o_1, ..., o_i', ..., o_k | o_1, ..., o_i, ..., o_k) \\
=&\log \frac{1}{k} \Pr(o_1, ..., o_i', ..., o_k | o_1, ..., o_i, ..., o_k) \qquad (2) \\
=&S^{(o_1,...,o_i',...,o_k)}(o_i') \sum_{j \neq i}^{k} S^{(o_1,...,o_i',...,o_k)}(o_j) + const \\
\approx&S^{o_{-i}}(o_i') \sum_{j \neq i}^{k} S^{o_{-i}}(o_j) + const \\
=&S^{o_{-i}}(o_i') + const \qquad (3)
\end{aligned}$$

, where $S^{o_{-i}}$ is the scoring matrix derived from o without o_i and $S(o_i)$ is the score of o_i based on S. With this setup, [13] proposed Algorithm 1.

2.3.2 Pros and Cons

Gibbs sampling approach is easy to implement and it converges to optimal solution as the running time goes to infinity. Besides, PSSM representation of a motif is easy to understand and visualize. While, the downside of this approach is that it cannot bear gap. If the sequence collection shares a gapped motif, PSSM based Gibbs sampler will fail to recognize it. Furthermore, the length of the motif and the number of steps are two hyper-parameters which should be specified by the user. In practice, it takes extra computing time to search for a suitable width and to stop early or later is instance specific, and has no general strategy to follow.

Algorithm 1 Gibbs sampler for motif finding

Input: sequence set q, motif width w
Output: set of starting points o

initialize o
$o^\star \leftarrow o$
while forever **do**
    pick i from $1, ..., |q|$ with uniform distribution
    update $S \leftarrow S^{o_{-i}}$
    compute transition probability $p(o_i') \leftarrow \exp\{S(o_i')\}$
    update $o_i \sim \text{Multinomial}(p, 1)$
    update $o^\star$ if there is any improvement
end while

2.3.3 Protocol

Following steps were applied to the analysis:

1. Implement a Gibbs sampler described in Algorithm 1
2. Train a set of motifs with various widths using Cas5 and Cas7
3. Score Cas9 sequences based on learned motifs

The widths used started from 10 to up to 190 and the size of interval was 10. To retrieve not only the most representative motif but other sub-optimal ones, for every width we repeated the Gibbs sampling procedure five times and recorded the best output for every run. For every Markov chain, we ran 600 steps because it was sufficient to reach a suboptimal with 600 steps for all instances in analysis.

2.3.4 Analysis

To show our Gibbs sampler converge, we track the objective value along the Markov chain. And to show our motif finding based strategy works, we performed the proposed procedure on simulated sequences and globin sequences given in Hmmer [11] where two sets of sequences share similar primary sequence patterns and checked whether our approach could recover such similarity.

2.4 Domain-specific profile HMM

2.4.1 Model & Algorithm Overview

To find out whether a sequence of amino acid belongs some domain, we can build a model of the domain and try to match the sequence of the model. Profile Hidden Markov Model is one of the models we can build to figure out whether a sequence contains the domain. The model of profile HMM is shown in Figure 3.

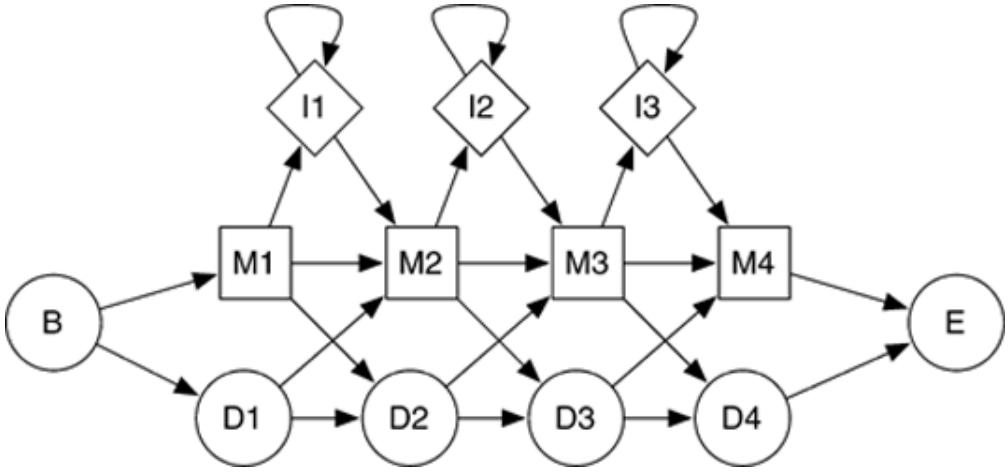

Figure 3: Profile Hidden Markov Model [11]

2.4.2 Pros and Cons

- Pros
  1. We can leverage the abundant prior knowledge of Cas9 domain markers by building profile HMM using the alignment of the domains rather than the full sequence.
  2. Compared with pair-wise sequence alignment, HMM can find more cases of distantly related sequences.
  3. As shown in the figure, HMM can model insertions and deletions.
- Cons
  1. To leverage the prior knowledge of domain markers, those markers need to be fetched separately from other data source rather than learned by the algorithm.
  2. The number of parameters is very large and they need to be optimized.

2.4.3 Protocol

a. A set of *Cas9* or *Cas5* sequences and their domain markers are fetched from EMBL-EBI (http://www.ebi.ac.uk).

b. Sequences of each of the shared domains are subtracted from the full *Cas9* or *Cas5* sequences.

c. For each domain, multiple sequences from different *Cas9* or *Cas5* are aligned by Clustal Omega (http://www.clustal.org/).

d. A profile HMM is built on the multiple sequence alignment for each domain by Hmmer (http://hmmer.org/) [11].

e. Search for matches using the profile HMM in the sequences of all previously downloaded *Cas* family proteins.

2.4.4 Analysis

For method testing, a set of globin sequences given in Hmmer [11] is used. Since there are *Cas9* and *Cas5* in the previously downloaded *Cas* family proteins sequences, they also act as positive control since the method should be able to find the domains in these proteins.
For output analysis, HMM is able to give the probability of given sequence emitted from the underlying domain profile HMM.

3 Results

Each of the three approaches for identifying motifs of *Cas* proteins and the resulting data are presented below.

3.1 Sequence Alignment using Dynamic Programming

Table 1: Semi-Global Alignment Output

Gene Name	Alignment Score	Average Score / base	% Sequence aligned	Traceback Start	Traceback End
Cas10_Mtuberculosis	234	0.212148685	1	[1365][810]	[330][1]
Cas10_Phorikoshii	335	0.249813572	1	[1368][762]	[41][1]
Cas10_Ssolfataricus	364	0.27063197	1	[1322][1046]	[62][1]
Cas10_Tvolcanium	343	0.305976806	1	[1203][776]	[146][1]
Cas1_DvsH_plasmid	160	0.282186949	1	[1297][344]	[734][1]
Cas1_Gvaginalis	163	0.332653061	1	[1367][321]	[886][1]
Cas1_K-12	105	0.243055556	0.911764706	[428][306]	[1][28]
Cas1_Tdenticola	176	0.387665198	1	[519][291]	[79][1]
Cas2_K-12	65	0.5	1	[528][95]	[400][1]
Cas2_Lsalivarius	64	0.444444444	1	[617][102]	[476][1]
Cas2_StB20-like	68	0.586206897	1	[217][88]	[102][1]
Cas3_DvsH_plasmid	152	0.145315488	1	[1059][703]	[32][1]
Cas3_Gvaginalis	260	0.229681979	0.858199753	[1119][811]	[1][116]
Cas3_K-12	233	0.185805423	1	[1254][889]	[29][1]
Cas4_K-12	171	0.341317365	1	[1323][364]	[833][1]
Cas4_Ssolfataricus	87	0.294915254	1	[679][203]	[387][1]
Cas4_TtenaxKra	57	0.230769231	1	[573][191]	[330][1]
Cas5_Gvaginalis	140	0.309050773	1	[1185][292]	[733][1]
Cas5_K-12	79	0.302681992	0.8	[261][225]	[1][46]
Cas6_Cbotulinum	158	0.478787879	1	[740][230]	[414][1]
Cas6_Hvolcanii_plasmid	99	0.25848564	1	[1309][273]	[929][1]
Cas6_Ssolfataricus	118	0.280952381	1	[1117][288]	[701][1]
Cas7_Hvolcanii_plasmid	164	0.316602317	1	[899][341]	[390][1]
Cas7_K-12	171	0.341317365	1	[1323][364]	[833][1]
Cas7_Ssolfataricus	147	0.267272727	1	[950][312]	[404][1]
Cas8_LsV	136	0.265625	1	[1289][314]	[783][1]
Cas8_Pdistasois	237	0.295511222	1	[1342][573]	[559][1]
Cas8_Pgingivalis	212	0.280794702	1	[1340][500]	[609][1]
Cas9_Bthermosphacta	1726	1.249818972	1	[1365][1301]	[47][1]
Cas9_Cindologenes	476	0.282157676	1	[1366][1444]	[3][1]
Cas9_Cochracea	375	0.276344878	0.650315347	[1322][1427]	[1][500]
Cas9_Hpullorum	180	0.27820711	1	[643][345]	[4][1]
Cas9_Hpullorum_2	330	0.391459075	1	[1347][703]	[549][1]
Cas9_Kkingae	520	0.370634355	0.997172479	[1341][1061]	[1][4]
Cas9_Movipneumoniae	519	0.340998686	0.988216811	[1349][1273]	[1][16]
Cas9_Nlactamica	505	0.353889278	0.994459834	[1368][1083]	[1][7]
Cas9_Pacidlactici	1786	1.209207854	0.999267399	[1366][1365]	[1][2]
Cas9_Pnultocida	510	0.365068003	0.997161779	[1348][1057]	[1][4]
Cas9_Ranatipestifer	391	0.239143731	1	[1366][1406]	[3][1]
Cas9_Sgallolyticus	4529	3.274765004	0.999271137	[1368][1372]	[1][2]
Cas9_Smoniliformis	514	0.327597196	1	[1368][1260]	[3][1]
Cas9_Spaucimobilis	415	0.290616246	1	[1362][1091]	[3][1]

Method as discussed in section 2.2. Visualization of data is available in Figure 4

Table 2: Local Alignment Output

Gene Name	Alignment Score	Average Score / base	% Sequence aligned	Traceback Start	Traceback End
Cas10_Mtuberculosis	317	0.239064857	0.988888889	[1345][808]	[46][8]
Cas10_Phorikoshii	313	0.246650906	0.874015748	[1368][705]	[105][40]
Cas10_Ssolfataricus	432	0.316020483	0.864244742	[1365][929]	[16][26]
Cas10_Tvolcanium	362	0.312878133	0.923969072	[1312][749]	[165][33]
Cas1_DvsH_plasmid	152	0.290630975	0.959302326	[1237][342]	[728][13]
Cas1_Gvaginalis	120	0.270880361	0.99376947	[1332][320]	[904][2]
Cas1_K-12	107	0.29558011	0.666666667	[1340][263]	[983][60]
Cas1_Tdenticola	124	0.294536817	0.965635739	[1096][287]	[687][7]
Cas2_K-12	43	0.651515152	0.663157895	[1119][86]	[1055][24]
Cas2_Lsalivarius	77	0.611111111	0.794117647	[621][101]	[496][21]
Cas2_StB20-like	72	1.028571429	0.568181818	[738][87]	[669][38]
Cas3_DvsH_plasmid	210	0.195712954	0.944523471	[1099][683]	[48][20]
Cas3_Gvaginalis	261	0.244382022	0.937114673	[1317][790]	[272][31]
Cas3_K-12	248	0.245787909	0.750281215	[1232][873]	[265][207]
Cas4_K-12	177	0.353293413	0.848901099	[1355][325]	[856][17]
Cas4_Ssolfataricus	100	0.304878049	0.827586207	[1308][180]	[981][13]
Cas4_TtenaxKra	60	0.810810811	0.277486911	[883][175]	[811][123]
Cas5_Gvaginalis	159	0.42513369	0.863013699	[797][274]	[424][23]
Cas5_K-12	77	0.292775665	0.68	[493][184]	[233][32]
Cas6_Cbotulinum	132	0.371830986	0.82173913	[453][215]	[99][27]
Cas6_Hvolcanii_plasmid	93	0.322916667	0.648351648	[755][225]	[468][49]
Cas6_Ssolfataricus	121	0.292978208	0.954861111	[1208][286]	[799][12]
Cas7_Hvolcanii_plasmid	177	0.5	0.692082111	[416][297]	[63][62]
Cas7_K-12	177	0.353293413	0.848901099	[1355][325]	[856][17]
Cas7_Ssolfataricus	139	0.445512821	0.769230769	[1144][311]	[837][72]
Cas8_LsV	168	0.305454545	0.984076433	[592][313]	[43][5]
Cas8_Pdistasois	222	0.308333333	0.848167539	[772][572]	[65][87]
Cas8_Pgingivalis	192	0.309677419	0.666	[690][498]	[75][166]
Cas9_Bthermosphacta	1736	1.260711692	0.996156802	[1361][1296]	[47][1]
Cas9_Cindologenes	658	0.443396226	0.810249307	[1366][1170]	[3][1]
Cas9_Cochracea	493	0.361172161	0.62789068	[1352][896]	[3][1]
Cas9_Hpullorum	174	0.280645161	0.985507246	[616][342]	[6][3]
Cas9_Hpullorum_2	351	0.445997459	0.928876245	[1345][701]	[614][49]
Cas9_Kkingae	528	0.372881356	0.97737983	[1368][1042]	[3][6]
Cas9_Movipneumoniae	620	0.417508418	0.865671642	[1357][1118]	[2][17]
Cas9_Nlactamica	580	0.420594634	0.874422899	[1359][955]	[3][9]
Cas9_Pacidlactici	1793	1.214769648	0.998534799	[1365][1364]	[1][2]
Cas9_Pnultocida	577	0.416907514	0.904446547	[1361][961]	[3][6]
Cas9_Ranatipestifer	529	0.354795439	0.826458037	[1368][1162]	[3][1]
Cas9_Sgallolyticus	4540	3.289855072	0.997084548	[1366][1369]	[1][2]
Cas9_Smoniliformis	543	0.378133705	0.843650794	[1368][1063]	[3][1]
Cas9_Spaucimobilis	404	0.289191124	0.973418882	[1332][1063]	[4][2]

Method as discussed in section 2.2. Visualization of data is available in Figure 5

3.2 Gibbs Sampling

3.2.1 Proof of concept

3.2.2 Motif finding in Cas5 and Cas7

3.2.3 Pattern recognition in Cas9

3.3 Domain-specific profile HMM

3.3.1 Protocol testing

For method quality control purpose, some protocol was applied to globins4.fasta (a file of 4 sequences of globins generated from a tutorial file of HMMer). We built the profile Hmm and save it in golbins4.hmm. Using this profile HMM, we searched globins45.fa (a tutorial file of HMMer) and the result is saved into golbins.search. The summary of found domains is in the header part of the file as shown in Figure. 6. An example alignment of the found domain is shown in Figure. 7. We can see that a domain is found in HBB_MANSP. The profile HMM model can also be read from the alignment. Full result is available in golbins.search file.

3.3.2 Finding domains of *Cas9*

After subtracting and aligning all the shared domains of *Cas9* marked in EMBL-EBI. We built profile HMMs from these multi-sequence alignments and save them into data/domains_Cas9/IPR032239.hmm, etc. Then we use these models to find the domains of *Cas9* in all *Cas* family protein sequences, hoping to find similar domains in other proteins in different subtypes. The summary of one of the search results can be seen in Figure. 8. The profile HMM can find similar domains in other *Cas9* proteins, which is as expected. But all the models failed to find any similar domains in other *Cas* family proteins other than *Cas9*.

3.4 Finding domains of *Cas5*

We also tried the other direction. That is to use domain markers of *Cas5*, which is a component of Type I system, and to find similar domains in *Cas9*. The protocol is the same and one of the summaries of the results can be found in Figure 9. The result is similar, the model is able to find some similar domains in other *Cas5* proteins but not other *Cas* family proteins.

4 Conclusion

. . .

TODO: we need to do this one together

TODO: please include any other resources or papers you referenced

References

[1] K. S. Makarova, et al., *An updated evolutionary classification of CRISPR-Cas systems* http://dx.doi.org/10.1038/nrmicro3569, 28 September 2015

[2] Saul B. Needleman, Christian D. Wunsch, *A general method applicable to the search for similarities in the amino acid sequence of two proteins* http://www.sciencedirect.com/science/article/pii/0022283670900574, 28 March 1970

[3] Smith, Temple F., Waterman, Michael S., *Identification of Common Molecular Subsequences* Journal of Molecular Biology. 147: 195-197. doi:10.1016/0022-2836(81)90087-5. PMID 7265238., 1981

[4] Devaki Bhaya,Michelle Davison, and Rodolphe Barrangou, *CRISPR-Cas Systems in Bacteria and Archaea: Versatile Small RNAs for Adaptive Defense and Regulation* Annual Review of Genetics Vol. 45: 273-297 (Volume publication date December 2011) DOI: 10.1146/annurev-genet-110410-132430

[5] Mali, P., Esvelt, K. M., & Church, G. M. (2013). Cas9 as a versatile tool for engineering biology. Nature Methods, 10(10), 957?963. doi:10.1038/nmeth.2649

[6] Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E., *A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity* Science. 2012 Aug 17;337(6096):816-21. doi: 10.1126/science.1225829. Epub 2012 Jun 28.

[7] Ran, F Ann and Hsu, Patrick D and Wright, Jason and Agarwala, Vineeta and Scott, David A and Zhang, Feng, *Genome engineering using the CRISPR-Cas9 system* Nat. Protocols(2013) http://dx.doi.org/10.1038/nprot.2013.143

[8] Qi LS, Larson MH, Gilbert LA, Doudna JA, Weissman JS, Arkin AP, Lim WA., *Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression* Cell. 2013 Feb 28;152(5):1173-83. doi: 10.1016/j.cell.2013.02.022.

[9] Hiroshi Nishimasu, F. Ann Ran, Patrick D. Hsu, Silvana Konermann, Soraya I. Shehata, Naoshi Dohmae, Ryuichiro Ishitani, Feng Zhang, Osamu Nureki *Crystal Structure of Cas9 in Complex with Guide RNA and Target DNA* http://www.cell.com/cell/pdf/S0092-8674(14)00156-1.pdf February 13, 2014

[10] Ran FA, Hsu PD, Lin C-Y, et al. *Double nicking by RNA-guided CRISPR Cas9 for enhanced genome editing specificity.* Cell. 2013;154(6):1380-1389. doi:10.1016/j.cell.2013.08.021.

[11] HMMER 3.1b2 (February 2015); http://hmmer.org/

[12] G. D. Stormo, T. D. Schneider, L. Gold, and A. Ehrenfeucht, "Use of the 'perceptron'algorithm to distinguish translational initiation sites in e. coli," *Nucleic Acids Research*, vol. 10, no. 9, pp. 2997–3011, 1982.

[13] C. E. Lawrence, S. F. Altschul, M. S. Boguski, J. S. Liu, A. F. Neuwald, J. C. Wootton, *et al.*, "Detecting subtle sequence signals: a gibbs sampling strategy for multiple alignment," *SCIENCE-NEW YORK THEN WASHINGTON-*, vol. 262, pp. 208–208, 1993.

Figure 4: Pairwise sequence alignment of various *Cas* protein sequences against *s. pyogenes Cas9* protein sequence. Green bars show coverage of semi-global alignment of individual sequence against *s. pyogenes Cas9*. Purple bars show coverage of local alignment of individual sequence against *s. pyogenes Cas9*. Darker color indicates higher average score per base, and therefore higher sequence similarity. Each grey marker represents 10 amino acid residues

Figure 5: Pairwise sequence alignment of various *Cas* protein sequences against *s. pyogenes Cas9* protein sequence. Green bars show coverage of semi-global alignment of individual sequence against *s. pyogenes Cas9*. Purple bars show coverage of local alignment of individual sequence against *s. pyogenes Cas9*. Darker color indicates higher average score per base, and therefore higher sequence similarity. Each grey marker represents 10 amino acid residues

```
# hmmsearch :: search profile(s) against a sequence database
# HMMER 3.1b2 (February 2015); http://hmmer.org/
# Copyright (C) 2015 Howard Hughes Medical Institute.
# Freely distributed under the GNU General Public License (GPLv3).
# - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
# query HMM file:                  globins4.hmm
# target sequence database:        globins45.fa
# - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

Query:       globins4  [M=149]
Scores for complete sequences (score includes all domains):
   --- full sequence ---   --- best 1 domain ---    -#dom-
    E-value  score  bias    E-value  score  bias    exp  N  Sequence    Description
    -------  ------ -----    -------  ------ -----   ---- --  --------    -----------
    8.7e-67  215.6   2.9    9.7e-67  215.4   2.9    1.0  1  MYG_ESCGI
    1.1e-65  211.9   0.1    1.3e-65  211.8   0.1    1.0  1  HBB_MANSP
    7.4e-65  209.3   0.2    8.2e-65  209.2   0.2    1.0  1  HBB_CALAR
    5.5e-64  206.5   1.2    6.1e-64  206.3   1.2    1.0  1  MYG_HORSE
    2.8e-63  204.2   0.1    3.1e-63  204.1   0.1    1.0  1  HBB_URSMA
    9.9e-63  202.4   0.5    1.1e-62  202.3   0.5    1.0  1  HBB_RABIT
    2.6e-62  201.1   1.3    2.8e-62  200.9   1.3    1.0  1  HBA_PONPY
      2e-61  198.2   1.1    2.2e-61  198.1   1.1    1.0  1  HBB_SPECI
      1e-60  195.9   1.7    1.1e-60  195.8   1.7    1.0  1  MYG_LYCPI
    1.1e-60  195.8   0.3    1.2e-60  195.7   0.3    1.0  1  MYG_PROGU
    1.4e-60  195.5   0.7    1.5e-60  195.3   0.7    1.0  1  HBB_SPETO
```

Figure 6: profile HMM search results summary in globins.search

```
>> HBB_MANSP
   #    score  bias  c-Evalue  i-Evalue hmmfrom  hmm to    alifrom  ali to    envfrom  env to     acc
 ---   ------ ----- --------- --------- ------- -------    ------- -------    ------- -------    ----
   1 !  211.8   0.1   1.3e-65   1.3e-65       1     149 []       1     146 []       1     146 [] 0.99

  Alignments for each domain:
  == domain 1  score: 211.8 bits;  conditional E-value: 1.3e-65
   globins4   1 vvLseaektkvkavWakveadveesGadiLvrlfkstPatqefFekFkdLstedelkksadvkkHgkkvldAlsdalakldekleaklkdLselHakklk 100
                v+L+++ekt+v+++W+kv  +v+e+G+++L rl++++P+tq+fF++F+dLs +d++++++vk+Hgkkvl+A+sd+l++ld +l++++++LselH++kl+
  HBB_MANSP   1 VHLTPEEKTAVTTLWGKV--NVDEVGGEALGRLLVVYPWTQRFFDSFGDLSSPDAVMGNPKVKAHGKKVLGAFSDGLNHLD-NLKGTFAQLSELHCDKLH 97
                69***************..*************************************************.***************** PP

   globins4 101 vdpkyfkllsevlvdvlaarlpkeftadvqaaleKllalvakllaskYk 149
                vdp++fkll++vlv+vla++++keft++vqaa++K++a va++la+kY+
  HBB_MANSP  98 VDPENFKLLGNVLVCVLAHHFGKEFTPQVQAAYQKVVAGVANALAHKYH 146
                **********************************************7 PP
```

Figure 7: profile HMM domain alignment in globins.search

```
qchu@Qis-MacBook-Pro domains_Cas9 (master)*$ cat IPR003615_result.txt
# hmmsearch :: search profile(s) against a sequence database
# HMMER 3.1b2 (February 2015); http://hmmer.org/
# Copyright (C) 2015 Howard Hughes Medical Institute.
# Freely distributed under the GNU General Public License (GPLv3).
# - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
# query HMM file:                  IPR003615.hmm
# target sequence database:        ../../../genes/complete_amino_acids.fa
# - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

Query:       IPR003615  [M=50]
Scores for complete sequences (score includes all domains):
   --- full sequence ---   --- best 1 domain ---    -#dom-
    E-value  score  bias    E-value  score  bias    exp  N  Sequence                    Description
    ------- ------ -----    ------- ------ -----   ---- --  --------                    -----------
    7.4e-31   97.3   4.4      2e-30   95.9   4.4    1.8  1  Cas9_Sgallolyticus.fasta_1
    1.1e-25   80.7   3.5    2.7e-25   79.5   3.5    1.7  1  Cas9_Pacidlactici.fasta_1
      2e-23   73.5   1.0    5.7e-23   72.1   1.0    1.9  1  Cas9_Bthermosphacta.fasta_1
    5.7e-19   59.3   0.1    1.8e-18   57.6   0.1    2.0  1  Cas9_Cindologenes.fasta_1
    8.9e-16   49.0   0.0    2.2e-15   47.8   0.0    1.7  1  Cas9_Pnultocida.fasta_1
    6.7e-15   46.2   0.2    6.2e-14   43.1   0.2    2.9  1  Cas9_Smoniliformis.fasta_1
    2.1e-14   44.6   0.1    5.6e-14   43.3   0.1    1.8  1  Cas9_Hpullorum_2.fasta_1
      3e-14   44.2   0.0    6.1e-14   43.2   0.0    1.6  1  Cas9_Nlactamica.fasta_1
    4.9e-14   43.5   0.0    9.4e-14   42.6   0.0    1.5  1  Cas9_Spaucimobilis.fasta_1
    2.8e-13   41.0   0.1    2.8e-13   41.0   0.1    2.5  2  Cas9_Ranatipestifer.fasta_1
    4.3e-13   40.5   0.0    1.1e-12   39.2   0.0    1.8  1  Cas9_Kkingae.fasta_1
    7.7e-13   39.6   3.3    7.7e-13   39.6   3.3    2.5  1  Cas9_Movipneumoniae.fasta_1
    8.8e-11   33.1   0.1    8.9e-10   29.8   0.0    2.9  2  Cas9_Cochracea.fasta_1
```

Figure 8: One example result for searching for *Cas9* in all *Cas* family proteins

```
# hmmsearch :: search profile(s) against a sequence database
# HMMER 3.1b2 (February 2015); http://hmmer.org/
# Copyright (C) 2015 Howard Hughes Medical Institute.
# Freely distributed under the GNU General Public License (GPLv3).
# - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
# query HMM file:                  IPR013422.hmm
# target sequence database:        ../../../genes/complete_amino_acids.fa
# - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

Query:       IPR013422  [M=40]
Scores for complete sequences (score includes all domains):
   --- full sequence ---   --- best 1 domain ---    -#dom-
    E-value  score  bias    E-value  score  bias    exp  N  Sequence               Description
    ------- ------ -----    ------- ------ -----   ---- --  --------               -----------
    1.2e-15   49.0   0.3    2.5e-15   47.9   0.3    1.6  1  Cas5_K-12.fasta_1
    5.5e-10   30.8   0.0    8.8e-10   30.2   0.0    1.3  1  Cas5_Gvaginalis.fasta_1


Domain annotation for each sequence (and alignments):
>> Cas5_K-12.fasta_1
   #    score  bias  c-Evalue  i-Evalue hmmfrom  hmm to    alifrom  ali to    envfrom  env to     acc
 ---   ------ ----- --------- --------- ------- -------    ------- -------    ------- -------    ----
   1 !   47.9   0.3   1.2e-16   2.5e-15       1      40 []       5      44 ..       5      44 .. 0.98

  Alignments for each domain:
  == domain 1  score: 47.9 bits;  conditional E-value: 1.2e-16
          IPR013422  1 lllelfaplaswrkPsasqersSyplPpPStilGaLaAil 40
                       l+l+l++p+++w++P + ++r++ ++P++S++lG+L+A+l
  Cas5_K-12.fasta_1  5 LILRLAGPMQAWGQPTFEGTRPTGRFPTRSGLLGLLGACL 44
                       79**********************************97 PP
```

Figure 9: One example result for searching for *Cas9* in all *Cas* family proteins