Identification of Conserved Regions in CRISPR protein family

Christine Baek[1] christib@andrew.cmu.edu
Qi Chu[1] qchu@andrew.cmu.edu
Yanyu Liang[1] yanyul@andrew.cmu.edu

[1]Computational Biology, Carnegie Mellon University

Abstract

The abstract goes here.

I. Introduction

TODO: Christine : brief intro

A. CRISPR/Cas

TODO: Christine writes about CRISPR bg Subsection text here.

B. Past Approaches

Functionally related regions can be clustered by evidences in experimental data. As summarized in Figure. 1. Previous work has found conserved regions on the sequence level using sequence alignment and structural information [5] inside each sub-type of the *Cas* system but not across the whole *Cas* family.

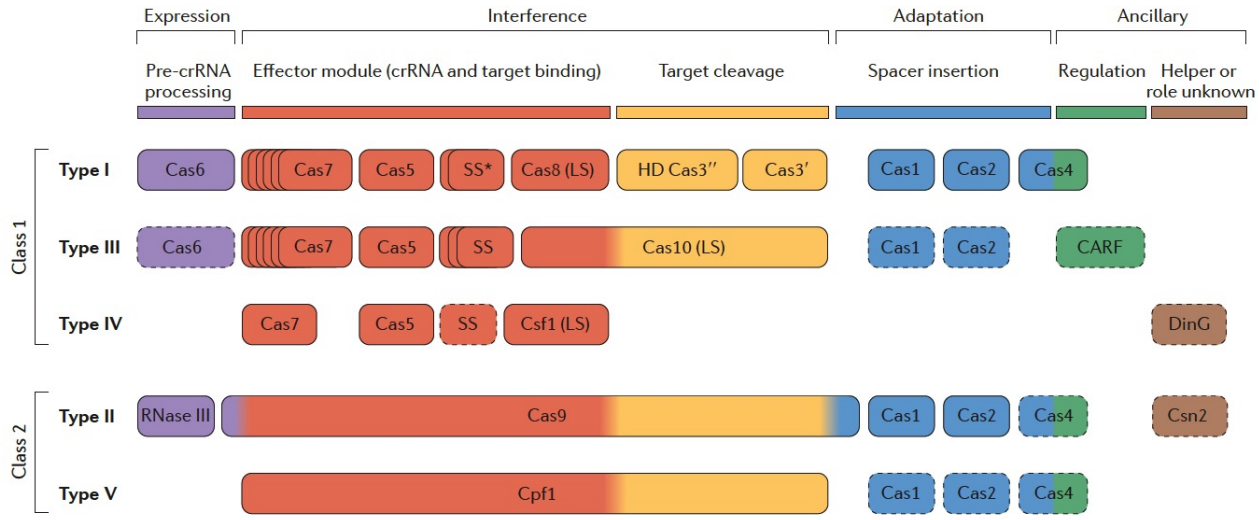

Fig. 1. Conserved building blocks of *Cas* family proteins [4]

C. Approaches in This Paper

TODO: everyone?

D. Goal of Paper

TODO: Yanyu : short discussion of our goal in this paper

II. Methods

We used 3 differenr approaches .. blahblahblah. All used the same .fa sequence, etc. etc. talk about the data itself here, and why we used 3 different methods. **TODO:** Christine : fill in this section, and why protein sequence

A. Data Retrieval

TODO: Qi : talk about source of data

B. Sequence Alignment using Dynamic Programming

1) Model & Algorithm Overview: talk about overview of what the method does (method itself, not in detail of how you used it)

2) Pros and Cons: of using the method - what is it capable of, what are the limitations ?

3) Protocol: implementation details - justifications for decisions you made when you ran the experiment, parameters, etc.

4) Analysis: Discuss METHOD for analysis, not the actual result/analysis itself.

C. Gibbs Sampling

1) Model & Algorithm Overview: talk about overview of what the method does (method itself, not in detail of how you used it)

2) Pros and Cons: of using the method - what is it capable of, what are the limitations ?

3) Protocol: implementation details - justifications for decisions you made when you ran the experiment, parameters, etc.

4) Analysis: Discuss METHOD for analysis, not the actual result/analysis itself.

D. Domain-specific profile HMM

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 2.

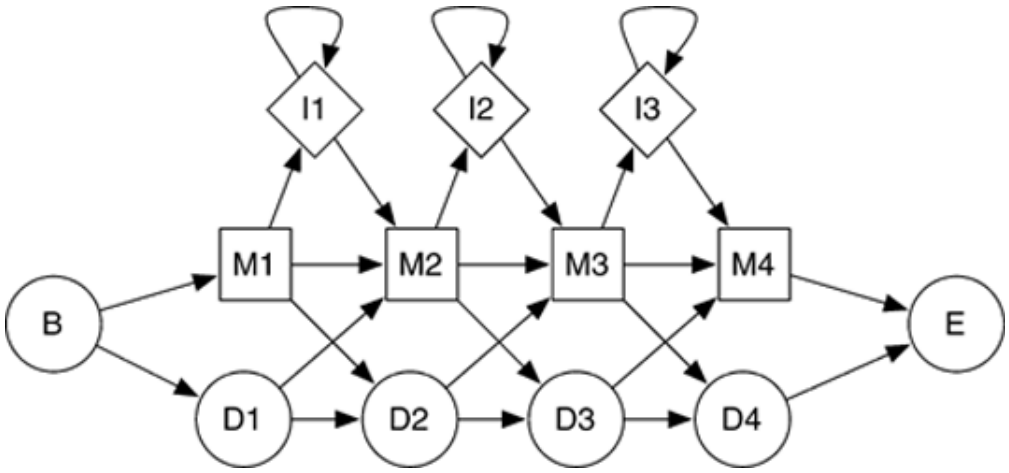

Fig. 2. Profile Hidden Markov Model [3]

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.

3) Protocol:

1) A set of *Cas9* or *Cas5* sequences and their domain markers are fetched from EMBL-EBI (http://www.ebi.ac.uk).
2) Sequences of each of the shared domains are subtracted from the full *Cas9* or *Cas5* sequences.
3) For each domain, multiple sequences from different *Cas9* or *Cas5* are aligned by Clustal Omega (http://www.clustal.org/).
4) A profile HMM is built on the multiple sequence alignment for each domain by Hmmer (http://hmmer.org/) [3].
5) Search for matches using the profile HMM in the sequences of all previously downloaded *Cas* family proteins.

4) Analysis: For method testing, a set of globin sequences given in Hmmer [3] 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.

III. Results

Individual results for each
TODO: put in appropriate figures, and any analysis for each

A. Sequence Alignment using Dynamic Programming

B. Gibbs Sampling

C. HMM

IV. Conclusion

TODO: we need to do this one together
TODO: please include any other resources or papers you referenced

References

[1] H. Kopka and P. W. Daly, *A Guide to LaTeX*, 3rd ed. Harlow, England: Addison-Wesley, 1999.
[2] K. S. Makarova, Y.I. Wolf, O.S. Alkhnbashi, E.V. Koonin et al., *An updated evolutionary classification of CRISPR-Cas systems* http://dx.doi.org/10.1038/nrmicro3569
[3] HMMER 3.1b2 (February 2015); http://hmmer.org/
[4] Makarova, Kira S., et al. "An updated evolutionary classification of CRISPR-Cas systems." Nature Reviews Microbiology (2015).
[5] Makarova, Kira S et al. ?Unification of Cas Protein Families and a Simple Scenario for the Origin and Evolution of CRISPR-Cas Systems.? Biology Direct 6 (2011): 38. PMC. Web. 26 Nov. 2016.