# Identification of Conserved Regions in CRISPR protein family

Christine Baek Computational Biology
Carnegie Mellon University
baek@cmu.edu Qi Chu Computational Biology
Carnegie Mellon University
qchu@andrew.cmu.edu Yanyu Liang Computational Biology
Carnegie Mellon University
yanyul@andrew.cmu.edu

**Abstract**

The abstract goes here.

## I. Introduction

This demo file is intended to serve as a "starter file" for IEEE conference papers produced under LaTeX using IEEEtran.cls version 1.8b and later. I wish you the best of success.

mds
August 26, 2015

### A. Subsection Heading Here

Subsection text here.

*1) Subsubsection Heading Here:* Subsubsection text here.

## 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.

### A. 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.

### B. 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.

### C. 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 **??**.

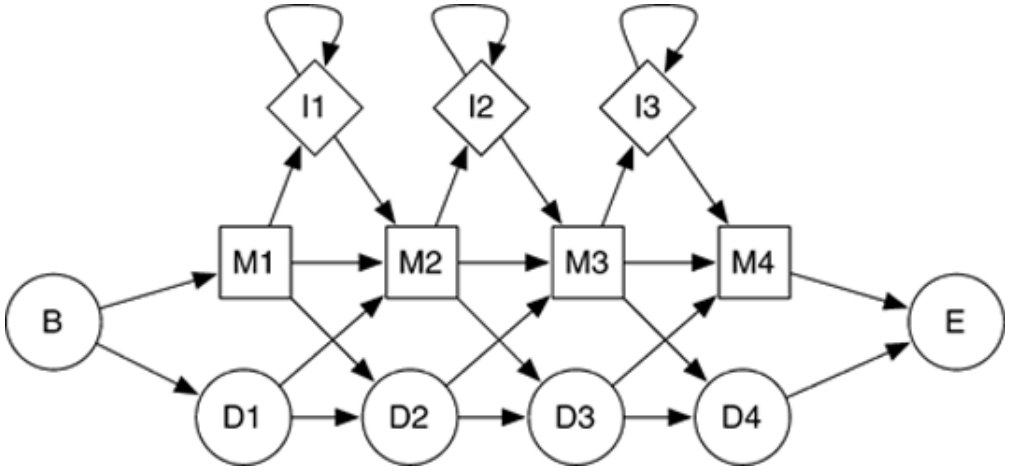


Fig. 1. Profile Hidden Markov Model

*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 a
  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:* label=$\alpha$

1) A set of Cas9 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 sequences.
3) For each domain, multiple sequences from different Cas9 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/).
5) Search for matches

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

## III. Results

## IV. Conclusion

The conclusion goes here.

## Acknowledgment

The authors would like to thank...

## 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/