Identification of Conserved Regions in CRISPR protein family

November 27, 2016

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

Abstract

The abstract goes here.

1 Introduction

TODO: Christine : brief intro

1.1 CRISPR/Cas

TODO: Christine writes about CRISPR bg Subsection text here.

1.2 Past Approaches

TODO: Qi : summarize the HMMer approach?

1.3 Approaches in This Paper

TODO: everyone?

1.4 Goal of Paper

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

2 Methods

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

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

TODO: Qi : talk about source of data

2.2 Sequence Alignment using Dynamic Programming

Semiglobal ailgnment using Needleman-Wunsch [3] and Local Alignment using Smith-Waterman Algorithm [4] were implemented for pairwise sequence comparison.

code available at `https://github.com/cookie223/CAS_project/tree/master/dp`

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

talk about overview of what the method does (method itself, not in detail of how you used it)

2.3.2 Pros and Cons

of using the method - what is it capable of, what are the limitations ?

2.3.3 Protocol

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

2.3.4 Analysis

Discuss METHOD for analysis, not the actual result/analysis itself.

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 1.

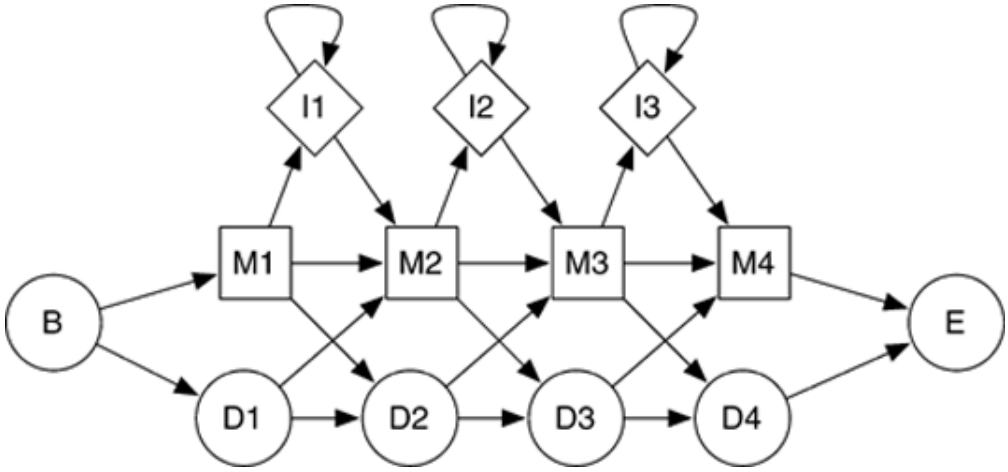

Figure 1: Profile Hidden Markov Model [5]

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/) [5].

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 [5] 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 2

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 3

3.2 Gibbs Sampling

f

i
l
l
.
.
.

3.3 HMM

f

i
l
l
.
.
.

4 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, et al., *An updated evolutionary classification of CRISPR-Cas systems* http://dx.doi.org/10.1038/nrmicro3569, 28 September 2015

[3] 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

[4] 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

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

Figure 2: 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 3: 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