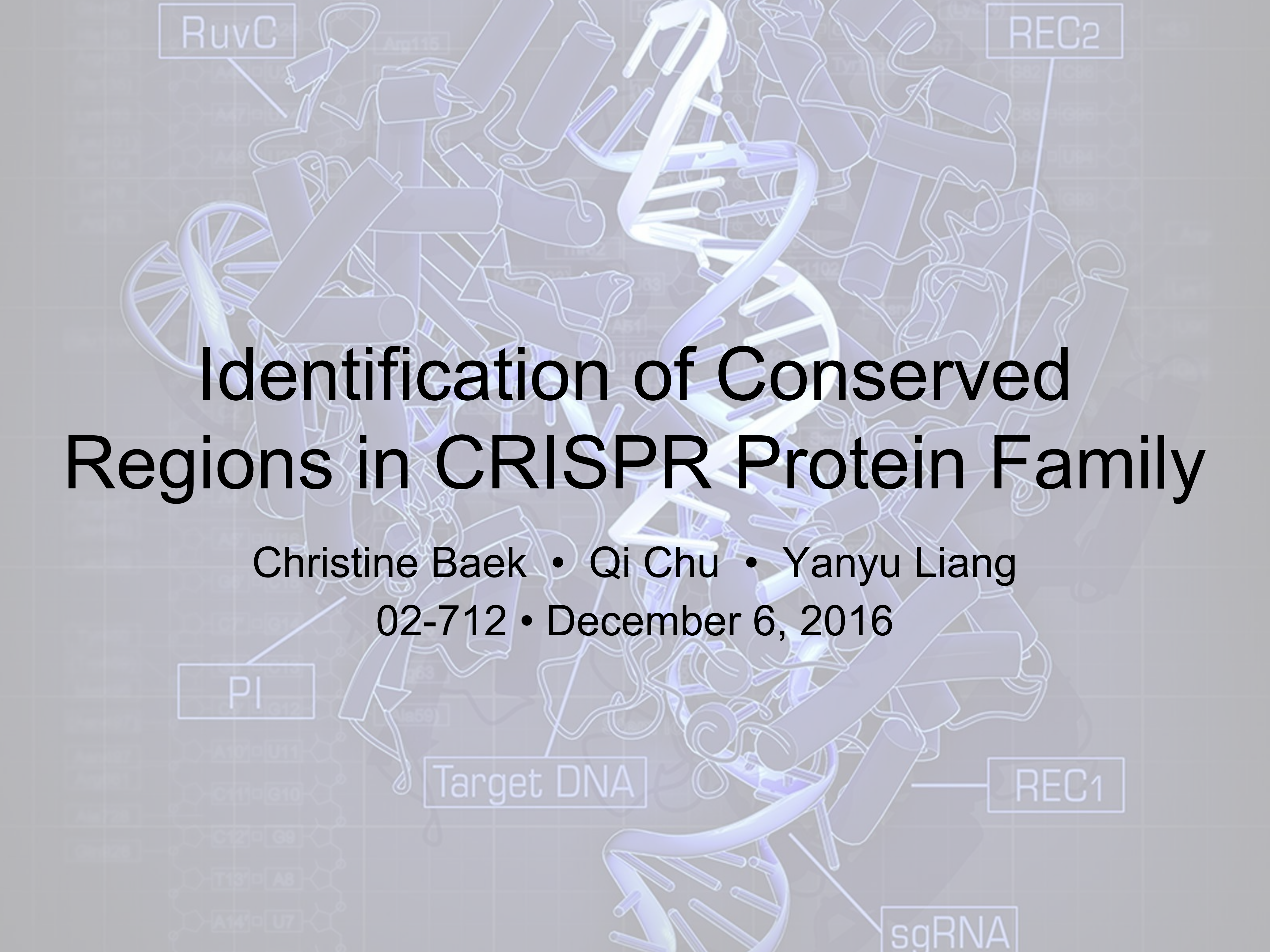


Identification of Conserved Regions in CRISPR Protein Family

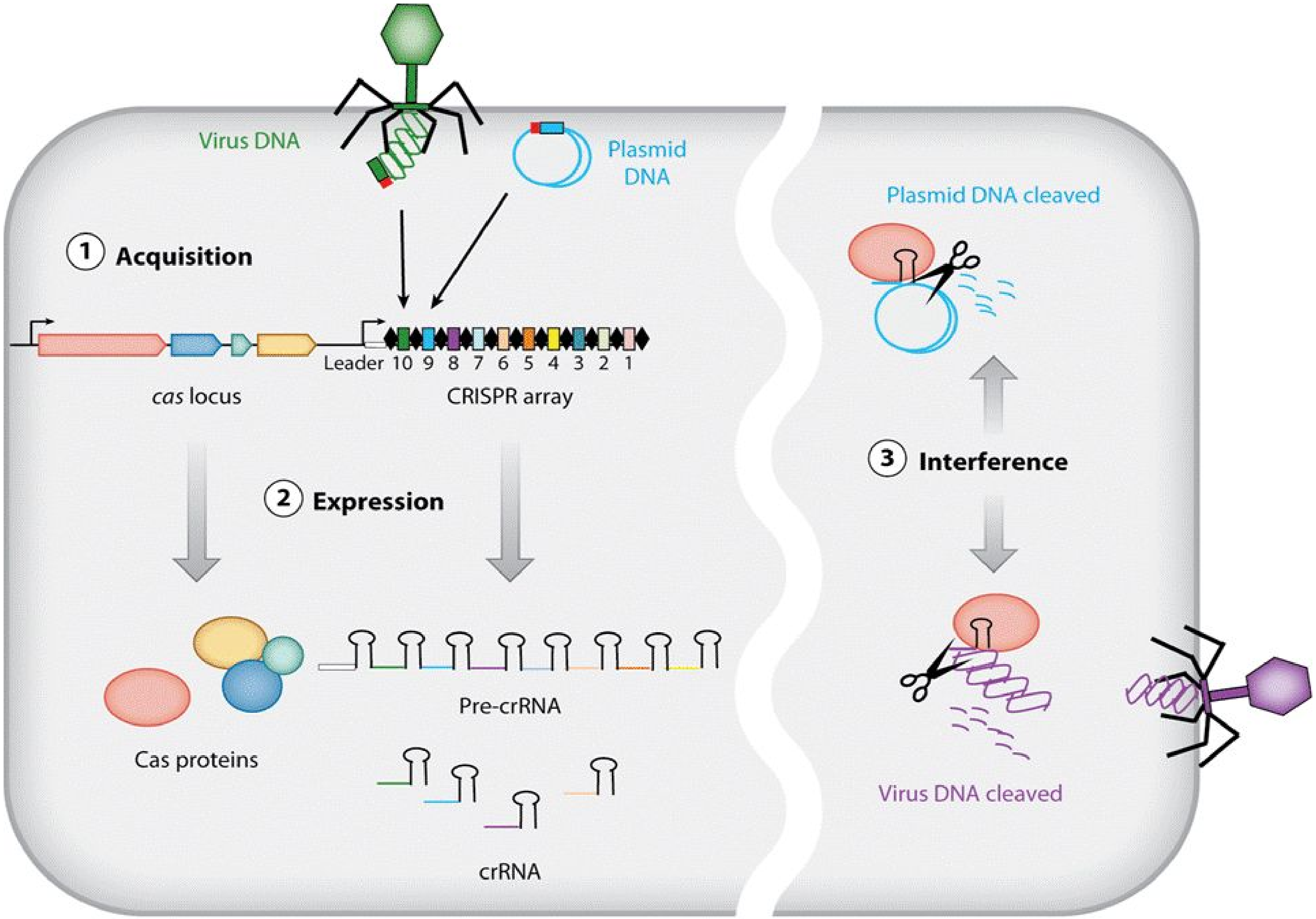
Christine Baek • Qi Chu • Yanyu Liang

02-712 • December 6, 2016

Introduction

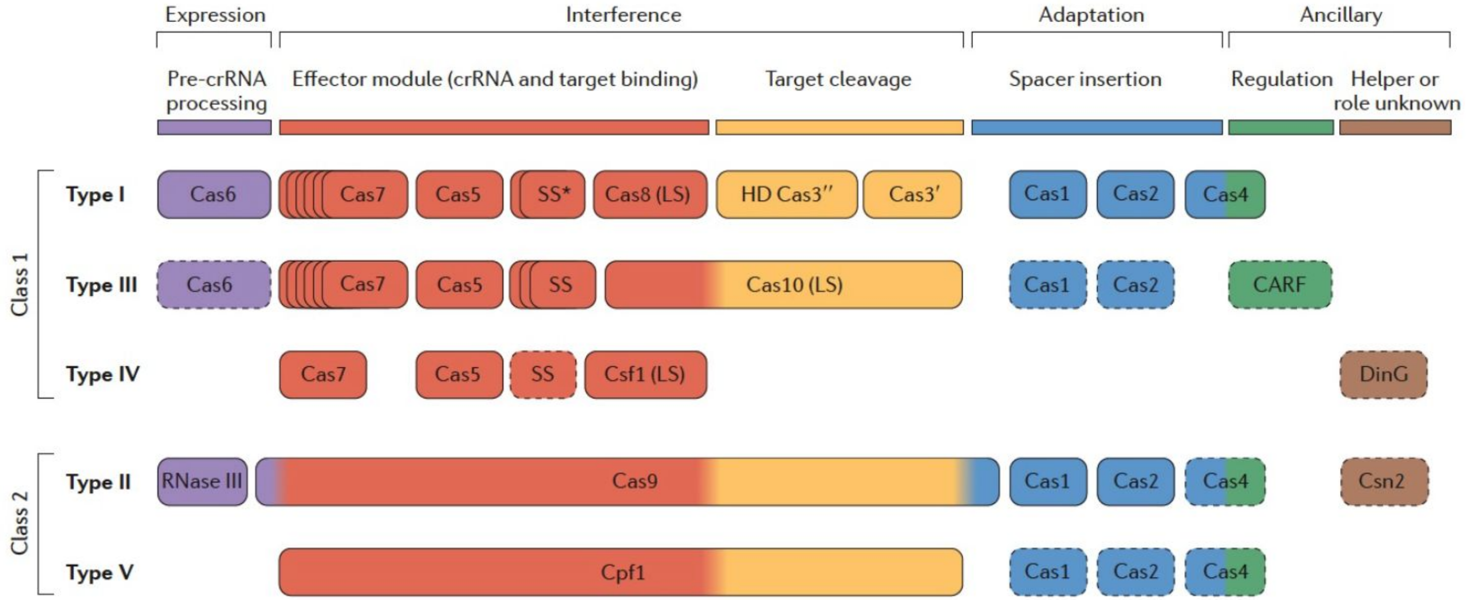
- CRISPR/*Cas* is virtually used by all kinds of bioengineering applications that benefit from sequence-level recognition
    - Genome Editing
        - In vitro (for transfection and genome editing)
        - In vivo (for studies and therapeutics)
    - Localization studies
    - Much more
- All of above is based on a single *Cas* protein, out of many possible candidates : *Cas9*
    - Initially chosen for its relative simplicity (single protein required for application)
- Much of current publication is focused on applications of *Cas9*, rather than the CRISPR/*Cas* system itself
    - Size of *Cas9*, as well as its PAM can be limiting for applications
- We attempt 3 different modeling approaches to identify conserved regions and motifs in different CRISPR/*Cas* protein, to broaden potential *Cas* proteins that may be used for future applications (and cuz, science)

CRISPR/Cas9


Virus DNA
Plasmid DNA
1 Acquisition
Leader 10 9 8 7 6 5 4 3 2 1
cas locus
CRISPR array
2 Expression
Cas proteins
Pre-crRNA
crRNA
Plasmid DNA cleaved
3 Interference
Virus DNA cleaved

Goal of Project

- To identify conserved regions shared by Cas9 and other Cas proteins



Three Methods for Motif Identification

- Pairwise Sequence Alignment
- Gibbs Sampling
- Profile HMM

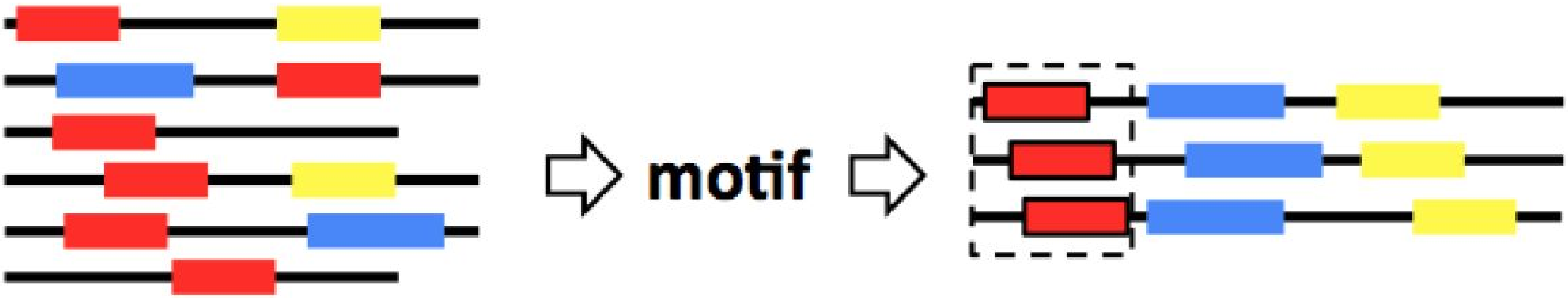
Data (Input)

- *Cas1* through *Cas10*
- Various species for each
- NCBI
- Amino Acid sequences used
    - Much more useful for identifying conserved regions in distantly related, divergent sequences

Method - Sequence Alignment

- Overview : compare 2 sequences, and identify which regions they are the most similar (and likely to have common ancestors)
- Alignment Methods
  - Semi-Global Alignment (Needleman-Wunsch)
    - Maps entire query sequence to reference sequence, with assumption that query sequence is shorter than the reference (begins later, ends earlier)
  - Local Alignment (Smith-Waterman)
    - Identifies only the highest similar regions between the query and reference sequence
- Parameters
  - BLAST default parameters (BLOSUM62, end gap, affine gap, and gap extension penalty)
- Considers biophysical properties of amino acids into account
  - Discrimination between different types of substitution, such as I→L vs I→W

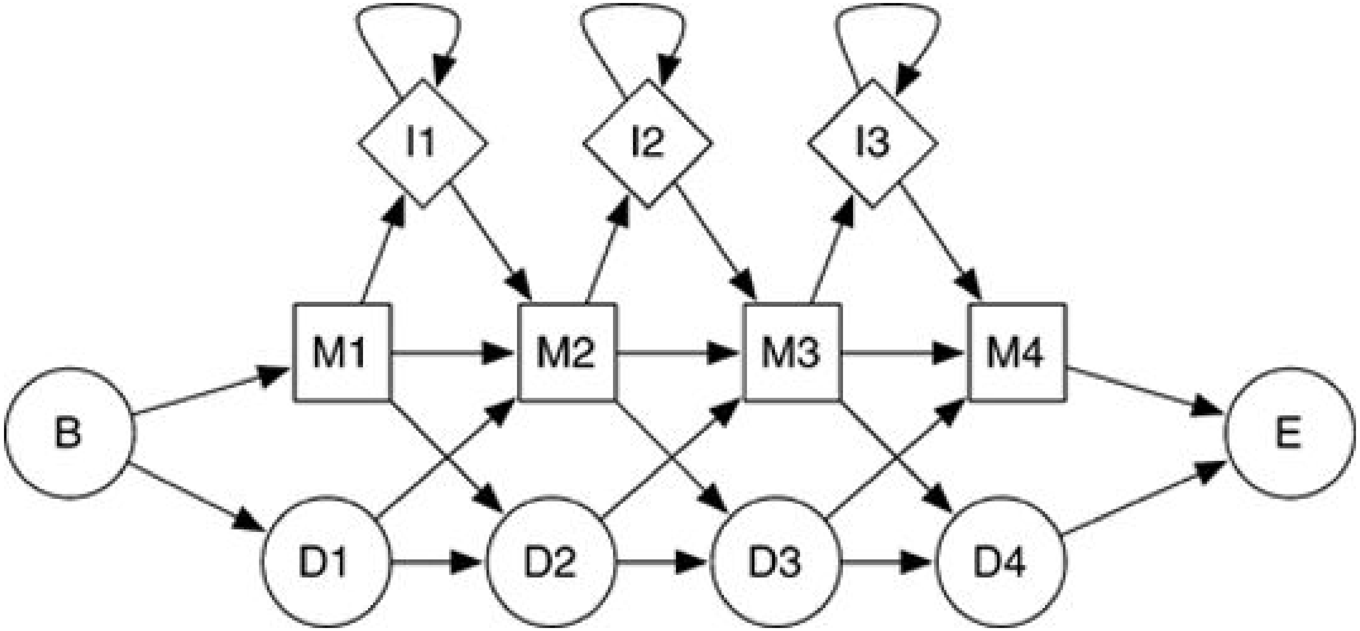
Method - Gibbs Sampling



- Method overview
  - Step 1: find motif in a sequence collection
  - Step 2: Recognize the motif in another collection
  - If the recognition succeeds, it helps to locate the conserved regions between the two sequence collections

Method - HMM

- profile-HMM
- Built for each marked domain in Cas5/9



source:http://hmmer.org/

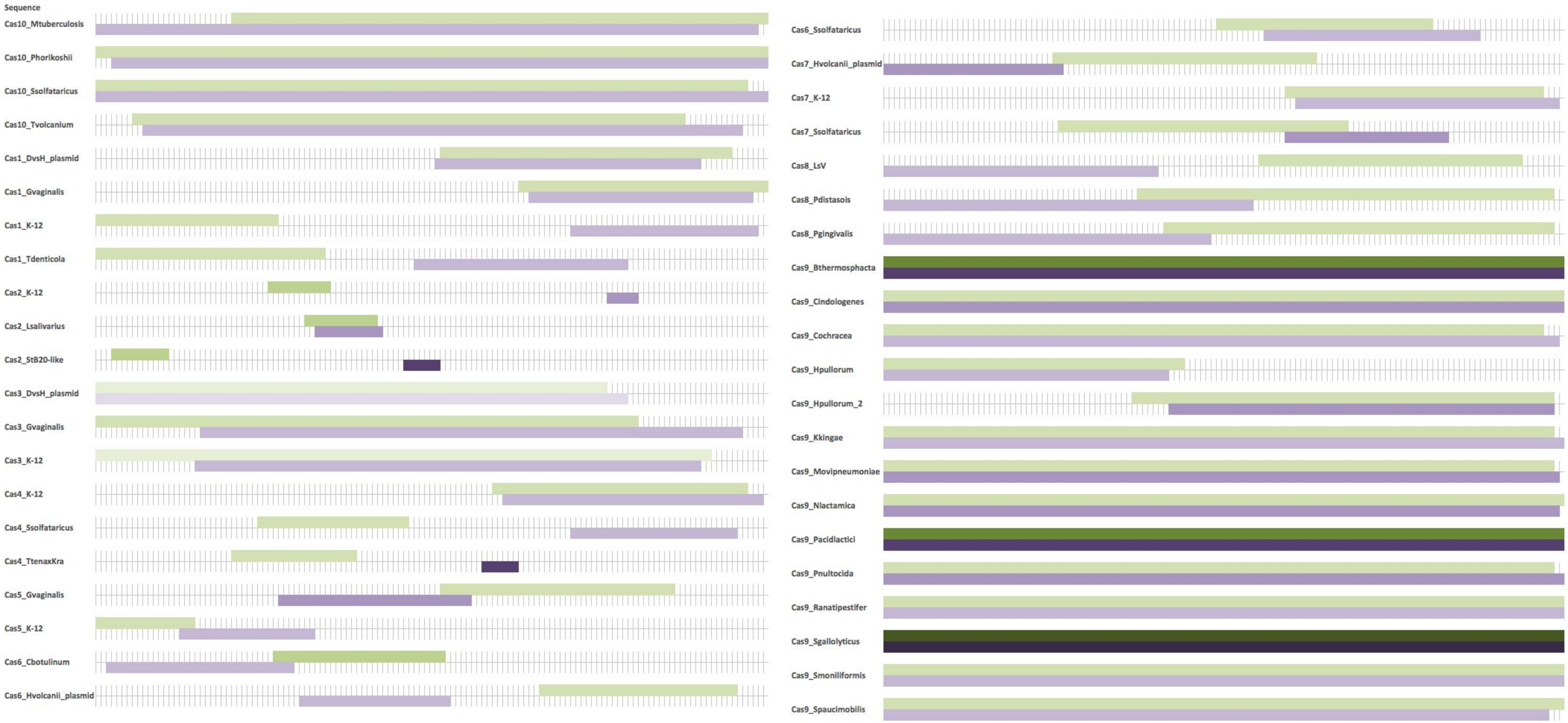
Results

Results - Sequence Alignment

Individual *Cas* sequences mapped to *Cas9*

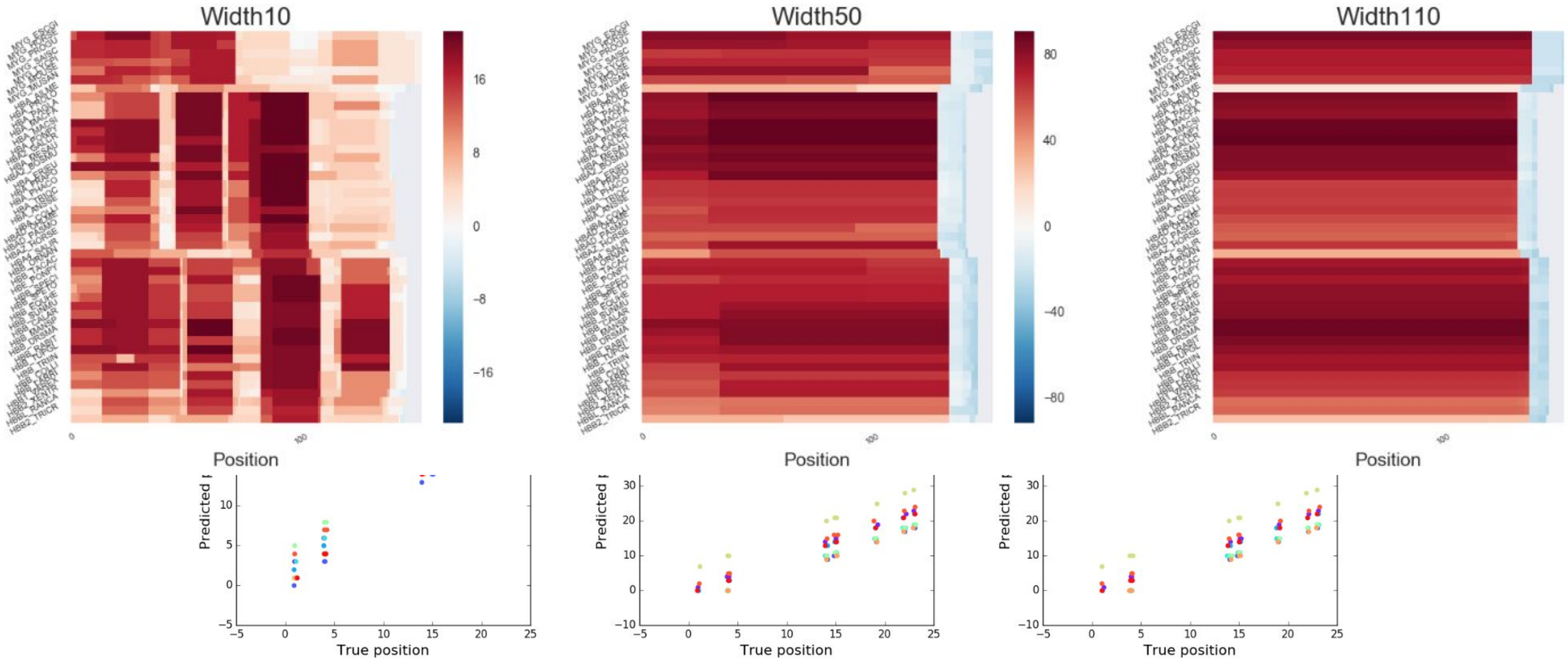
Darker color means higher sequence similarity (higher avg score per position)

Purple : local alignment, Green = semi-global alignment



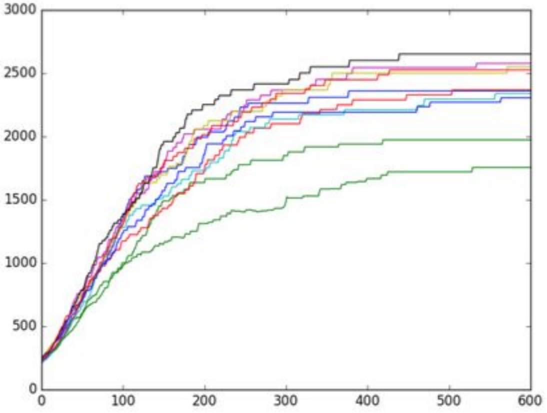
Results - Gibbs Sampling

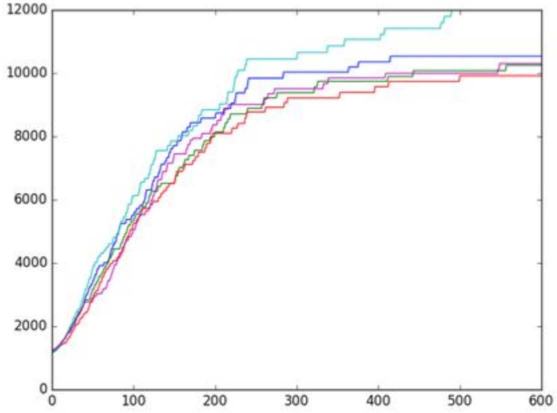
- Simulated data & globin data (positive control)
  - Robust to motif width
  - Capable to recover the position of the motif

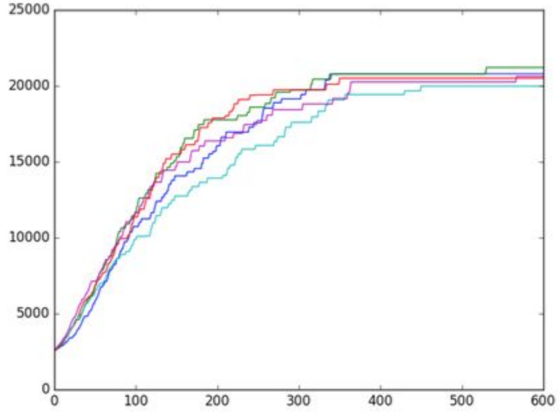


Results - Gibbs Sampling

- Cas 5 vs. Cas 9 - training & recognition



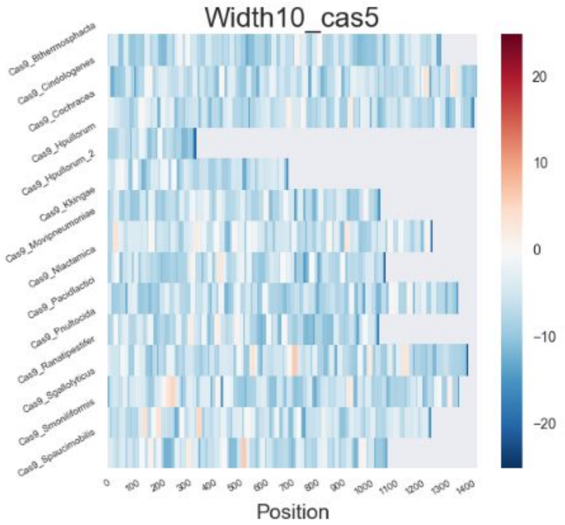


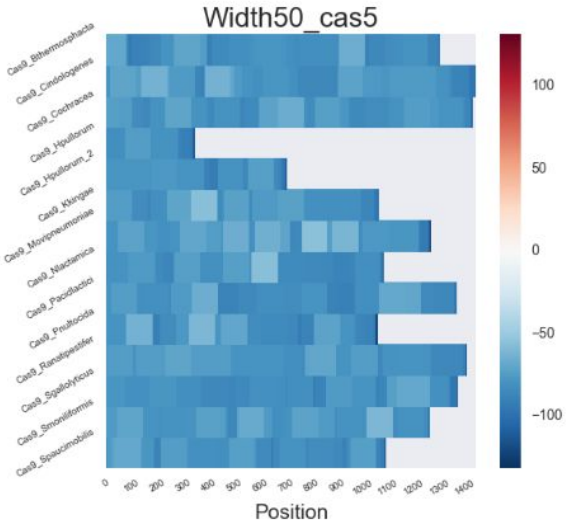


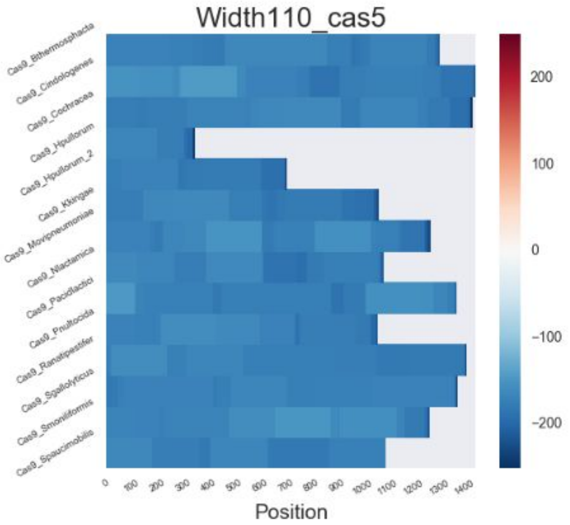
$\text{Width} = 10$

$\text{Width} = 50$

$\text{Width} = 110$



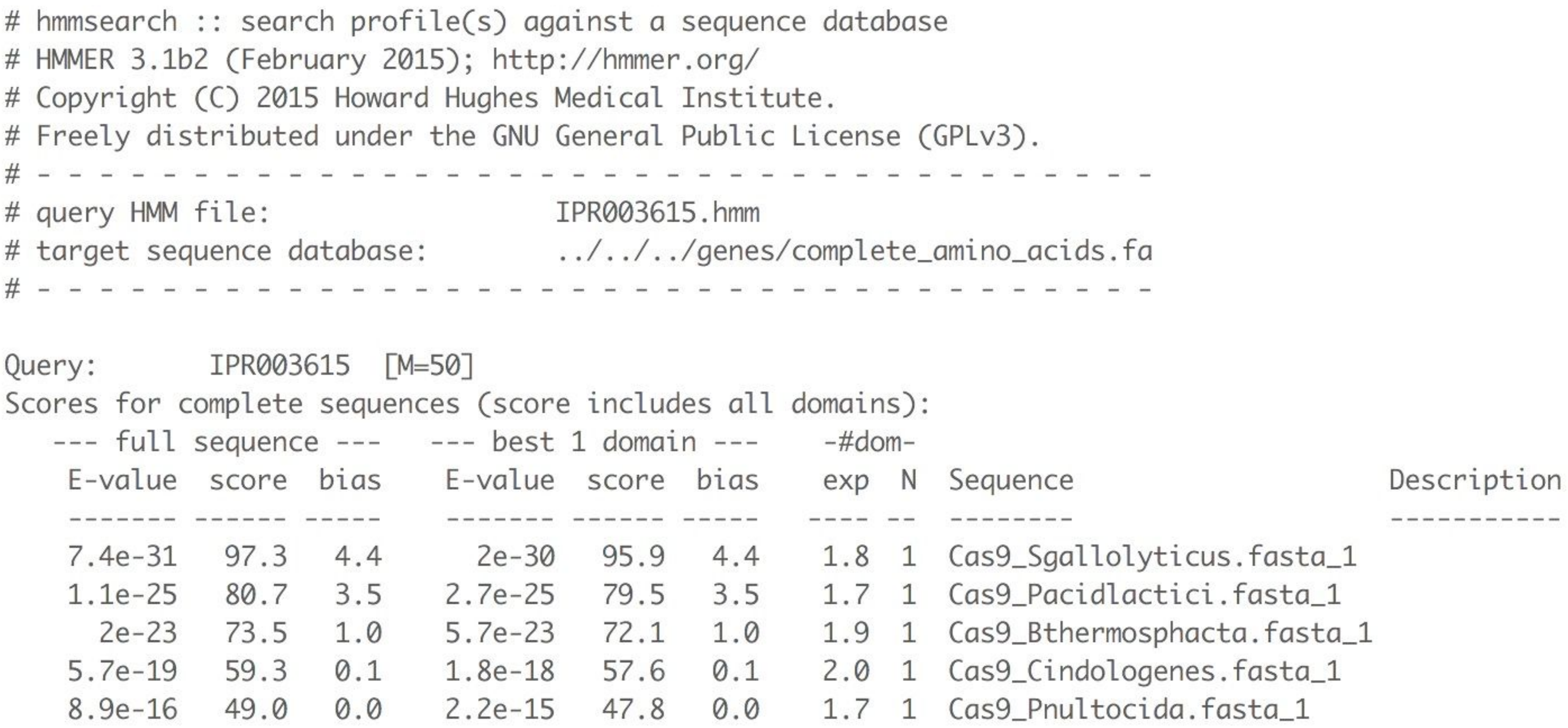




Results - HMM

Search for domain IPR003615 (from *Cas9*) in all collected *Cas* family protein

- Can find similar domains but only in other *Cas9*



```
# 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
```

Overall Results

- Sequence alignment performs well on evolutionarily close sequence, but performs poorly on highly divergent or short sequences
- Gibbs sampling approach is not capable to find conserved region between *Cas5/7* and *Cas9*, even if it is sensitive enough in positive control (simulated & globin)
- Profile HMM can find conserved domains in the same protein as the source in different species but cannot find similar domain in different proteins in the *Cas* family

The fact that this protein family is highly divergent through a long evolutionary timeframe (prokaryote), with limited data affected performance of each model

Conclusion

- More data is needed
- Input sequences represent billions of years in evolutionary distance (compared to Globin protein family of eukaryote)
- Some success with relatively closely related sequences, but divergent sequences showed limited success in identification of conserved regions
- CRISPR/Cas is far too divergent for a single method, especially if verified on eukaryotic evolutionary relationship, to work with direct application
- Future directions
    - Reduce size of alphabet
    - Add structural information into motif model
    - Obtain and test with more data
    - Develop models specifically designed for distantly related/divergent protein families

Questions