Identification and classification of Cas proteins by novel HMM

Introduction

Project is about the evolution and relationship of various Cas (CRISPR-associated) proteins. CRISPR is the adaptive immune system of bacteria and archaea. 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 Cas9 is limited in terms of PAM (Protospacer Adjacent Motif), of `-NGG` or to be used in conjunction with Cas9.

Goals

There has been previous attempt to identify different domains of various Cas proteins, such as in [1]. We plan to develop a HMM for identification of the domains as well as highly conserved regions, and cross-validate our results with the published work.

Work Plan

1. Collect sequence data in Cas family
2. Generate simulated sequences to test algorithm
3. Implement HMM and test its performance on simulated data
4. Organize and clean up Cas sequences based on mutliple sequence alignment result
5. Phylogenetic analysis of Cas sequences
6. Feed Cas sequences into HMM algorithm to learn interesting structure for each domain
7. Predict Cas 9 domain structure based on learned HMMs and analyze results

References

[1] An updated evolutionary classification of CRISPR-Cas systems
`http://www.nature.com/nrmicro/journal/v13/n11/full/nrmicro3569.html`