

# Simulation of Protein Families in the Twilight Zone

Cory L. Strope<sup>1</sup>, Stephen D. Scott<sup>2</sup>, Etsuko N. Moriyama<sup>3</sup>

**Keywords:** indels, twilight zone, sequence simulation

## 1 Introduction.

When the sequence identity of proteins drops below 30% identity alignment methods often fail to produce reliable alignments. This so-called “Twilight Zone” of protein sequence similarity is currently the focus of many bioinformatics research, with the majority investigating the structural properties of the proteins. However, this approach suffers from two disadvantages:

- (1) sequence similarity searching methods based on alignments are in general fast and widely used, whereas methods using structural information are slower than sequence based methods, and
- (2) protein structural information takes a great deal of time to resolve through experimental methods, such as Nuclear Magnetic Resonance and protein crystallography, so that structural databases hold structural information for only 1 out of every 100 sequenced proteins. Their family representation is also skewed based on how successfully experimental methods can be applied (*e.g.* not many transmembrane proteins have their structures solved).

For these reasons, it is still desirable to find twilight-zone proteins through sequence information only. Many efforts have been done to find these remotely similar proteins by: using sequence similarity searching through the use of different combinations of gap opening and extension costs, using different protein scoring matrices, or incorporating secondary structure information in reconstructing alignments. Regardless, the performance of any method needs to be statistically validated before being accepted. One way of validation is objective analysis of available methods when they are applied against remotely similar sequences with known homologous relationships. Currently, the procurement of such objective analysis for the methods to identify twilight-zone level similarities is hindered by inadequate numbers or unreliability of representative datasets. In reality, such analysis can be done only with appropriately simulated data.

## 2 Software and files.

Our method, indel-PSeq-Gen, simulates dynamic protein evolution through the introduction of empirical models of insertion and deletion (indel) events, as well as amino acid substitutions, that occur during sequence evolution. We incorporated the indel model into the sequence simulation package PSeq-Gen [1]. The PERL script Hybridizer incorporates the creation of multidomain sequence families into our simulation, with the ability to specify different evolutionary rates to different domains, thus allowing for the different degrees of

---

<sup>1</sup>Dept. of Computer Science and Engineering, University of Nebraska, Lincoln, NE 68588-0115  
E-mail: [cstrope@cse.unl.edu](mailto:cstrope@cse.unl.edu)

<sup>2</sup>Dept. of Computer Science and Engineering, University of Nebraska, Lincoln, NE 68588-0115,  
E-mail: [sscott@cse.unl.edu](mailto:sscott@cse.unl.edu)

<sup>3</sup>School of Biological Sciences, Plant Science Initiative, University of Nebraska, Lincoln, NE 68588-0660, [emoriyama2@unlnotes.unl.edu](mailto:emoriyama2@unlnotes.unl.edu)

conservation of structurally and functionally distinct regions of the proteins. Finally, our method outputs the “true” multiple alignment of the simulated sequences, which is necessary for alignment methods to be compared for their performance. Figure 1 gives the flow diagram of the protein family simulation.

### 3 Figures

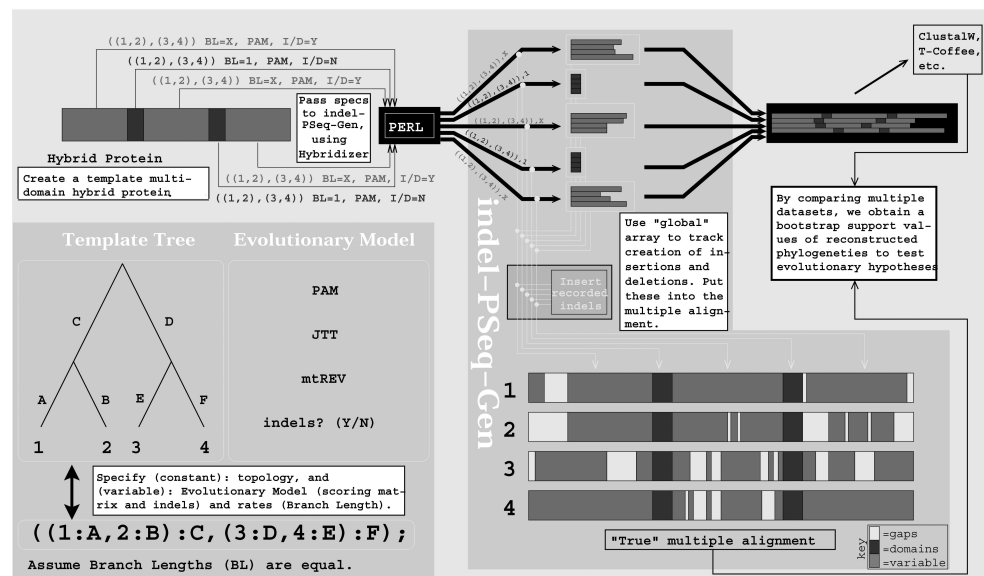


Figure 1: Flow diagram for the creation of hybrid protein families. A template topology with branch lengths, a protein scoring matrix, and the choice of including indels are supplied by the user. The PERL script *Hybridizer* then calls *indel-PSeq-Gen* for each of the domains, and outputs the raw sequences, which can then be multiply aligned. *Hybridizer* also collects the multiple alignments created by *indel-PSeq-Gen* and assembles them in order, and outputs the “true” multiple alignment of the protein sequence family.

### 4 References and bibliography.

#### References

- [1] Grassly, N., Rambaut, A. and Adachi, J. 1997. PSeq-Gen: An application for the monte carlo simulation of protein sequence evolution along phylogenetic trees. *Bioinformatics*, 13:559-560.