4ao Development and Application of Novel Pattern Discovery Techniques for Problems in Biochemical Engineering

Kyle L. Jensen and Gregory Stephanopoulos

Here I present a summary of my doctoral thesis work on the development an application of novel pattern discovery techniques for the analysis of diverse biomolecular data. A unifying theme of this work is the use of sequence-level motifs to inform and drive experiments, particularly for directed evolution and rational design. I will discuss two topics in detail:

- The development of GeMoDA (Generic Motif Discovery Algorithm), a novel algorithm for discovering repeated patterns in real-valued, sequential data without requiring alignment or pattern enumeration. GeMoDA is generic in the sense that it is applicable to almost any type of sequential data, e.g. DNA/protein sequences, protein structures, and time series data. Given a user-defined notion of similarity and identity, GeMoDA deterministically enumerates all repeated patterns of a minimum length and support. GeMoDA's output motifs can be modeled using any number of motif representations including position-weight matrices, hidden Markov models, and regular expressions. We show applications of GeMoDA in three domains: protein and DNA sequence motif discovery, and the discovery of repeated substructure discovery in protein structures. Unlike other algorithms for these two applications---viz. Gibbs sampler, MEME, Dali, and others---GeMoDA's output is guaranteed to be both optimal and exhaustive.
- The rational construction of a highly active Pl-lambda promoter variant using a simple, bootstrapping statistical method for parsing out the phenotypic contribution of a single mutation from clones which, when sequenced each is shown to contain a multitude of mutations and varied phenotypes. The method assumes that, given N phenotypic classes, mutations that do not effect the phenotype should partition between the N classes based on a multinomial distribution. Here we show that deviations from this distribution are indicative of a link between specific mutations and phenotype. As a proof-of-principle, we detail the construction of a highly active Pl-lambda promoter variant. Our results show that our statistical method can predict the effect of individual mutations. Furthermore, combinations of the influential sites produced promoter variants with activities exceeding those produced by random mutagenesis. We suggest that this method may be useful for expediting directed evolution experiments. In effect, our method allows for small forays into sequence space to be be translated into larger steps in phenotype space; or, a more directed directed evolution.

In addition, I will discuss a few research projects that are tangentially related to the theme of my thesis work including 1) the design of novel machine learning tools for the recognition of HIV protease substrate sites; 2) the use of bioinformatics methods for mapping the intellectual property landscape of the human genome; and 3) the role of intellectual property in chemical engineering graduate education. I will be presenting most of these topics in more detail individually at this year's conference.