

Spaghetti Code, Soupy Logic, and the Expression of Genes

Jim Kent

*Baskin School of Engineering
University of California, Santa Cruz*

Abstract

The human genome could be characterized as 3 billion bases of the most convoluted spaghetti code in existence. While for the most part the genome is not something we engineers would want to emulate in our own products, there are some lessons to be learned from it. More importantly understanding the genome in all its warts, convolutions, and spots of brilliance will lead to medical advances as fundamental as the understanding of infectious disease. Of peculiar interest to many of us who straddle the disciplines of software and biology is the study of gene expression. How does a cell decide which of its 25,000 genes to use at any given time? There are multiple regulatory networks which decide the usage patterns of a gene. The most fundamental of these networks - the interaction between transcription factor proteins and the DNA that they bind to - resembles in

many ways a neural network implemented in soup. Studying and characterizing this soupy logic is by no means easy, but progress is being made on many fronts. Comparative genomics helps separate functional DNA from the relics of intracellular parasites and evolutionary dead ends that make up more than 90% of the human genome. Advances in mRNA sequencing have made it easier to locate the true transcription start site, which is a hot spot of regulatory activity. DNA microarrays allow us to measure in parallel the expression patterns of entire genomes. CHIP/Chip techniques can physically map the binding sites of transcription factors genome-wide. Developing software to evaluate, integrate, store, and display this data is a complex, but most worthwhile challenge. This talk will review some of the software developed both at UCSC and elsewhere to address this challenge.