

Computational Approaches Towards Human Genome Annotation

Mark Gerstein

Molecular Biophysics and Biochemistry
Yale University

A central problem for 21st century science will be the analysis and understanding of the human genome. My talk will be concerned with topics within this area, in particular annotating pseudogenes (protein fossils) in the genome. I will discuss a comprehensive pseudogene identification pipeline and storage database we have built. This has enabled use to identify >10K pseudogenes in the human and mouse genomes and analyze their distribution with respect to age, protein family, chromosomal location. One interesting finding is the large number of ribosomal pseudogenes in the human genome, with 80 functional ribosomal proteins giving rise to ~2,000 ribosomal protein pseudogenes.

I will try to inter-relate our studies on pseudogenes with those on tiling arrays, which enable one to comprehensively probe the activity of intergenic regions. At the end I will bring these together, trying to assess the transcriptional activity of pseudogenes.

Throughout I will try to introduce some of the computational algorithms and approaches that are required for genome annotation and tiling arrays -- i.e. the construction of annotation pipelines, developing algorithms for optimal tiling, and refining approaches for scoring microarrays.

<http://bioinfo.mbb.yale.edu>

<http://pseudogene.org>