Interpreting the human genome sequence, using stochastic grammars

Richard Durbin
The Sanger Centre
Wellcome Trust Genome Campus
Hinxton
Cambridge CB10 1SA
UK

`rd@sanger.ac.uk`

Abstract

The 3 billion base pair sequence of the human genome is now available, and attention is focusing on annotating it to extract biological meaning. I will discuss what we have obtained, and the methods that are being used to analyse biological sequences. In particular I will discuss approaches using stochastic grammars analogous to those used in computational linguistics, both for gene finding and protein family classification.