Section 16. Applications

The Invited Speaker, Bonnie Berger, was not able to attend the Congress.

Recent Developments
in Computational Gene Recognition

Serafim Batzoglou, Bonnie Berger, Daniel J. Kleitman,
Eric S. Lander, and Lior Pachter

Abstract. We survey recent mathematical and computational work in the field of gene recognition, focusing on the techniques that have been developed to tackle the problem of identifying protein coding regions in genes. We also present a new approach to gene recognition which is based on a variety of tools we have developed.

1 Introduction

1.1 What do you do with 100KB of human genomic DNA?

Recent advances in DNA sequencing technology have led to rapid progress in the Human Genome Project. Within a few years, the entire human genome will be sequenced. The rapid accumulation of data has opened up new possibilities for biologists, while at the same time unprecedented computational challenges have emerged due to the mass of data. The questions of what to do with all the new information, how to store it, retrieve it, and analyze it, have only begun to be tackled by researchers [11]. These problems are distinguished from classical problems in biology, in that their solution requires an understanding not only of biology, but also of mathematics and computer science. Of the many problems, it is clear that the following tasks are of importance:

- Finding genes in large regions of DNA.

- Identifying protein coding regions within these genes.

- Understanding the function of the proteins encoded by the genes.

The important third problem, namely understanding the function of a newly sequenced gene, requires the solution of the second problem, identification of critical subregions which code for protein. Protein coding regions have different statistical characteristics from noncoding regions, and it is primarily this feature which