

STANFORD UNIVERSITY  
DEPARTMENT OF STATISTICS  
DEPARTMENT SEMINAR

4:15 p.m., Tuesday, May 2, 2000  
Sequoia Hall Rm. 200  
(Cookies at 3:45 p.m. in 1st Floor Lounge)

*Prof. Minping Qian*  
*Department of Probability & Statistics*  
*School of Math, Peking University*

**Promoter finding and linguistic**

According Fickett and Wasserman [1] " TFBSs stand out clearly against a non-conserved background" , but the problem of promoter prediction based on over-representation of TFBSs remains unsolved. The difficulty is that there are many TFBS-like patterns in non-promoters that mimic real TFBSs and lead to many false positives in promoter prediction. Since we have more than 2000 TFBSs, there will be more 2,000,000 pairs and 2 billion of triples of TFBSs. We introduce the relative entropy and a method for mining in data for the biological significant words—the over-represented words, and regulations for combination of words into phrases by hunting the over-represented TFBS pairs from non-conservative almost random back ground, is proposed. Our statistical results suggest that TFBSs come in well-defined pairs and triples, etc. In this way a stochastic regular grammar would hopefully be obtained.