Statistics in Molecular Biology
IMS Lecture Notes – Monograph Series (1999) Volume 33

CORRELATED MUTATIONS IN MODELS OF PROTEIN SEQUENCES: PHYLOGENETIC AND STRUCTURAL EFFECTS

BY ALAN S. LAPEDES[1], BERTRAND G. GIRAUD, LONCHANG LIU AND GARY D. STORMO

Los Alamos National Laboratory, Santa Fe Institute, Service Physique Théorique, DSM, C.E.N. Saclay and University of Colorado

Covariation analysis of sets of aligned sequences for RNA molecules is relatively successful in elucidating RNA secondary structure, as well as some aspects of tertiary structure [Gutell et al. (1992)]. Covariation analysis of sets of aligned sequences for protein molecules is successful in certain instances in elucidating certain structural and functional links [Korber et al. (1993)], but in general, pairs of sites displaying highly covarying mutations in protein sequences do not necessarily correspond to sites that are spatially close in the protein structure [Gobel et al. (1994), Clarke (1995), Shindyalov et al. (1994), Thomas et al. (1996), Taylor & Hatrick (1994), Neher (1994)]. In this paper we identify two reasons why naive use of covariation analysis for protein sequences fails to reliably indicate sequence positions that are spatially proximate. The first reason involves the bias introduced in calculation of covariation measures due to the fact that biological sequences are generally related by a non-trivial phylogenetic tree. We present a null-model approach to solve this problem. The second reason involves linked chains of covariation which can result in pairs of sites displaying significant covariation even though they are not spatially proximate. We present a maximum entropy solution to this classic problem of "causation versus correlation". The methodologies are validated in simulation.

1. Introduction. Analysis of sets of aligned sequences, such as RNA or protein sequences, is a common procedure in bioinformatic analysis. Various methods have been developed to describe aligned sequences: "consensus" sequences which are determined by the most conserved symbol in each sequence position; "profiles" [Gribskov et al. (1987)] which represent the probability distribution of symbols in each position, and can also include inserts and deletes with fixed position independent penalties; and "hidden Markov models" [Krogh et al. (1994)], which represent single site probability distributions as well as position dependent probability distributions for insertions and deletions. Correlation analysis extends such methods to consideration of the probability distribution for pairs of symbols in all possible pairs of positions in the sequence. "Mutual information", a measure of correlation for discrete symbols [Cover &

[1] Research supported by the Department of Energy under contract W-7405-ENG-36
AMS 1991 *subject classifications.* Primary 62F03 ; secondary 62P10, 92D20.
Key words and phrases. Correlated mutations, phylogeny, interaction, structure from sequence.