Pacific Symposium on Biocomputing 5:597-598 (2000)

Applications of Information Theory to Biology

T. Gregory Dewey
Keck Graduate Institute of Applied Life Sciences
535 Watson Drive, Claremont CA 91711, USA

Hanspeter Herzel
Innovationskolleg Theoretische Biologie, HU Berlin,
Invalidenstrasse 43, D-10115, Berlin, Germany

Information theory offers a number of fertile applications to biology. These applications range from statistical inference to foundational issues. There are a number of statistical analysis tools that can be considered information theoretical techniques. Such techniques have been the topic of PSB session tracks in 1998 and 1999. The two main tools are algorithmic complexity (including both MML and MDL) and maximum entropy. Applications of MML include sequence searching and alignment, phylogenetic trees, structural classes in proteins, protein potentials, calcium and neural spike dynamics and DNA structure. Maximum entropy methods have appeared in nucleotide sequence analysis (Cosmi et al., 1990), protein dynamics (Steinbach, 1996), peptide structure (Zal et al., 1996) and drug absorption (Charter, 1991).

Foundational aspects of information theory are particularly relevant in several areas of molecular biology. Well-studied applications are the recognition of DNA binding sites (Schneider 1986), multiple alignment (Altschul 1991) or gene-finding using a linguistic approach (Dong & Searls 1994). Calcium oscillations and genetic networks have also been studied with information-theoretic tools. In all these cases, the information content of the system or phenomena is of intrinsic interest in its own right.

In the present symposium, we see three applications of information theory that involve some aspect of sequence analysis. In all three cases, fundamental information-theoretical properties of the problem of interest are used to develop analyses of practical importance. The work of Grosse, Buldyrev, Stanley, Holste and Herzel investigates the average mutual information (AMI) content of coding and non-coding regions of DNA. The AMI provides a tool for elucidating species-independent patterns that differ between coding and non-coding regions. Histograms of this novel coding measure are virtually identical for various taxonomic classes. Since the AMI algorithm requires no species-dependent training, it can be applied easily to newly sequenced genomes. Algorithms based on AMI are competitive with conventional algorithms for identifying protein-coding regions. In the paper by Dewey, the evolution of the Shannon information entropy of sequence populations in *in vitro* selection-amplification protocols is investigated. It is seen

that for simple experimental designs, the Shannon entropy is a Lyapounov function of the evolving, dynamical system. As such, it can be used to assess the dynamical stability of such systems and reveals problems associated with the optimization of in vitro evolutionary systems. The work by Kshischo and Laessig presents a statistical theory of probabilistic sequence alignment. This method is a generalization of information-theoretical approaches and makes an analogy with the "thermodynamic" partition function at finite temperature. Finite-temperature alignments can be used to characterize the significance of an alignment and the reliability of its single element pairs. This results in improved accuracy of the resulting alignments.

This current era is seeing the generation of an enormous quantity of data by high-throughput technologies. The resulting problems and databases invite the application of information theory. Increasingly, there is a need for new methods of statistical inference that are suited to the both the large databases and the types of data that are seen in modern biology. In addition to the applications in this session, a number of other papers deal with information theoretical issues. These include work is such diverse fields as genome expression analysis and natural language processing. Information theory cannot be regarded as specific subfield of biology but it can provide a powerful methodological framework for attacking certain types of problems. We anticipate continued use of such methodology in complex problems of sequence and map alignment, motif identification and cluster analysis.

References

S.F. Altschul, J. Mol. Biol. 219, 555-565, 1991.

M. K. Charter & S.F. Gull, J. Pharmacokin. Biopharm. 19, 197, 1991.

C. Cosmi, V. Cuoma & M. Ragosta, J. Theor. Biol. 147, 423, 1990.

S. Dong & D.B. Searls, Genomics 23, 540-551, 1994.

T.D. Schneider, G.D. Stormo, L. Gold & A. Ehrenfeucht, J. Mol. Biol. 188, 415-431, 1986.

P.J. Steinbach, Biophys. J. 70, 1521-1528, 1996.

Zal, Franck, Lallier, H. Francois & A. Toulmond, J. Biol. Chem.271, 8875 (1996).