COMPLEXITY AND INFORMATION-THEORETIC APPROACHES TO BIOLOGY

David L. Dowe
Central Inductive Agency, Department of Computer Science,
Monash University, Clayton, Vic. 3168, Australia
e-mail: dld@cs.monash.edu.au

Klaus Prank
Department of Clinical Endocrinology, Medical School Hannover,
Carl-Neuberg-Str. 1, D-30625 Hannover, Germany
e-mail: ndxdpran@rrzn-serv.de

Notions of information and/or complexity have been applied to the analysis of a broad spectrum of biologically relevant problems, such as protein structure prediction, DNA motif discovery and compression, and neural spike trains[a].

These notions have been studied by Kolmogorov (1965) and Chaitin (1966), with Solomonoff (1964), Wallace (1968) and Rissanen (1978) applying these to problems of statistical and inferential learning and to prediction. The methods of Solomonoff, Wallace and Rissanen have respectively come to be known as Algorithmic Probability (ALP), Minimum Message Length (MML) and Minimum Description Length (MDL). All these methods relate to and can be thought of in terms of Shannon's information theory. These approaches were discussed in some detail at the recent Information, Statistics and Induction in Science (ISIS) conference (Melbourne, Australia, August 1996, World Scientific. Conference and Program Chair: David Dowe). As discussed in papers by Wallace, Rissanen and others, information-theoretic inference leads to statistically consistent and invariant estimation methods[b].

An MDL/MML perspective has been suggested by a number of authors in the context of approximating unknown functions with some parametric approximation scheme (such as a neural network). The designated measure to optimize under this scheme combines an estimate of the cost of misfit with an estimate of the cost of describing the parametric approximation (Akaike 1973; Rissanen 1978; Barron and Barron, 1988; Wallace and Boulton, 1968).

[a] We thank all colleagues who submitted to or reviewed for this stream. DLD was supported by Australian Research Council (ARC) Large Grant No. A49602504, and K.P. was supported by a travel grant from the Deutsche Forschungsgemeinschaft (Pr 333/8-1).

[b] It was D. Dowe's interest in information theory for these forms of inductive inference and also (as in the call for papers) in probabilistic prediction, and K. Prank's interest in biological information processing (e.g. neural spike trains) that led to this conference stream involving these related themes.

In this track Schmidt uses an information-theoretic framework to estimate the probability of the score of gapped alignments which can be applied to amino acid substitution matrices. Steeg and Pham present an information-theoretic approach to the finding of higher-order correlations in protein databases which overcomes limitations of previous methods. The notion of MML is used by Edgoose *et al.* in a study of the classification of protein structure using a sequential model of dihedral angles. A widely used information-theoretic DNA compression scheme as proposed by Milosavljevic and Jurka (1993) is shown to be inefficient and more efficient coding schemes (some of which run in linear time) are discussed by Powell *et al.*. On a different theme, Saitou addresses the ubiquitous phenomenon of self-assembling automata using a model of conformational self-assembly.

In the last couple of years it became apparent that pulsatile activity is an ubiquitous phenomenon in biological systems for the transmission and processing of information in the temporal and spatial domain as discussed across a wide range of disciplines at the recent International Conference on Biological Information Processing (ICOBIP '97, see ref. 1). This pulsatile mode of information transfer has the advantage to achieve superior signal-to-noise ratios by frequency modulation (FM) coding and to avoid desensitization to substained signals. Examples include the regulation of hormone secretion by the frequency of intracellular Ca^{2+} ($[Ca^{2+}]_i$) oscillations, and the differentiation of neurons in response to specific patterns of $[Ca^{2+}]_i$ oscillations. The frequency of signalling is proportional to the time-lag of the effect on a log-log scale across many orders of magnitude as demonstrated for action potential coding (milliseconds) to endocrine coding (hours to days). In contrast to most engineered communications systems which use well-defined frequencies for the transmission of information, many biological signalling systems use fluctuating signals. The information-theoretic work of Strong *et al.* on the quantification of information transmission in neural spike trains and of Prank *et al.* for the information transfer from extracellular hormonal stimuli to $[Ca^{2+}]_i$ spike trains is based on the characteristic of time-varying stimuli. The question on the temporal scale of information processing is addressed by Paluš *et al.* using an algorithm to estimate coarse-grained entropy rates from $[Ca^{2+}]_i$ spike trains. In contrast to the filters used in the studies of Prank *et al.* and Strong *et al.*, Bousquet *et al.* ask the question of whether the hippocampus as part of the nervous system works like an adaptive Kalman filter.

References

1. N.C. Spitzer and T.J. Sejnowski, *Science* **277**, 1060 (1997).