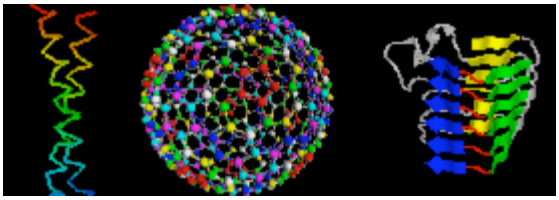
MIT
Department of Mathematics
& The Theory of
Computation Group
At CSAIL



Theory of Computation/Bioinformatics Seminar

Speaker: Paul Vitanyi, CWI and the University of Amsterdam
Title: Clustering by compression --- the Similarity metric
Date: Wednesday, 2 June 2004 ***PLEASE NOTE DAY, TIME & PLACE
Time & Location:
Refreshments: 3:45 pm in the Theory of Computation Lab at MIT's Building 32, Stata Center Room G-575
Talk: 4:00 pm the Theory of Computation Lab at MIT's Building 32, Stata Center, Room G-575
URL: http://www-math.mit.edu/compbiosem/

Abstract:

A new class of metrics appropriate for measuring effective similarity relations between sequences, say one type of similarity per metric, is studied. We propose a new ``normalized information distance", based on the noncomputable notion of Kolmogorov complexity, and show that it minorizes every metric in the class (that is, it is universal in that it discovers all effective similarities). We demonstrate that it too is a metric and takes values in [0,1]; hence it may be called the similarity metric. This is a theory foundation for a new general practical tool. We give two distinctive applications in widely divergent areas (the experiments by necessity use just computable approximations to the target notions). First, we computationally compare whole mitochondrial genomes and infer their evolutionary history. This results in a first completely automatic computed whole mitochondrial phylogeny tree. Secondly, we give fully automatically computed language tree of 52 different language based on translated versions of the ``Universal Declaration of Human Rights".

We also present a fully automatic method for music classification, based only on compression of strings that represent the music pieces. The method uses no background knowledge about music whatsoever: it is completely general and can, without change, be used in different areas like linguistic classification and genomics. It is based on an ideal theory of the information content in individual objects (Kolmogorov complexity), information distance, and a universal similarity metric. Experiments show that the method distinguishes reasonably well between various musical genres and can even cluster pieces by composer. This generated some interest in the popular press, see http://www.cwi.nl/~paulv

The method is implemented and available as public software, and is robust under choice of different compressors. To substantiate our claims of universality and robustness, we report evidence of successful application in areas as diverse as genomics, virology, languages, literature, music, handwritten digits, astronomy, and combinations of objects from completely different domains, using statistical, dictionary, and block sorting compressors. In genomics we presented new evidence for major questions in Mammalian evolution, based on whole-mitochondrial genomic analysis: the Eutherian orders and the Marsupionta hypothesis against the Theria hypothesis. Joint work with Ming Li, Bin Ma, Rudi Cilibrasi, Ronald de Wolf, Xin Chen, Xin Li.

The seminar is co-hosted by Professor Peter Clote of Boston College's Biology and Computer Science Departments and MIT Professor of Applied Math Bonnie Berger. Professor Berger is also affiliated with CSAIL & HST.

Massachusetts Institute
of Technology
77 Massachusetts Avenue
Cambridge, MA 02139

For General Questions, please contact kvdickey@mit.edu