

It's All In The Bits: Improved Database Searching with Better Bits

Andrew Smellie

Harold E. Helson

Cambridgesoft Inc., Cambridge, USA

Traditionally, substructure searches in databases have been performed by first reducing the topological representation of the molecule into an encoded representation in a bit string, where each bit in the string codes for the presence of one or more substructures. In its simplest form, a molecule is decomposed into fragments which are hashed into an enormous bitstring. Various techniques are used to reduce the length of the bitstring so that it is tractable to store it in a computer. In this paper, we describe a technique of generating a reduced length bitstring whilst attempting to preserve the maximal amount of information.

Additionally, we introduce a novel data structure that takes particular advantage of the way distances are computed in a bitstring space (i.e. the tanimoto coefficient) to greatly accelerate nearest neighbor searching and similarity calculations in those spaces.

Examples will be shown that demonstrate the screening effectiveness with the modified bitstring by comparison with traditional methods. Using the improved bitstring, it will be shown that search speeds are greatly enhanced.