

Using Self-Similar Geometric Structures to Represent Letter-Sequence-Indexed Statistical Summaries from Gene Regulation and Peptide Docking Studies

Daniel B. Carr
dcarr@gmu.edu

Abstract

The paper addresses the challenge of representing statistics that are indexed by sequences of letters in a way that has the potential of revealing structure in the space of all combinations. The approach develops coordinate systems based on simple geometric structures: tetrahedrons in the case of 4 nucleotides and icosahedron face centers in the case of 20 amino acids. The paper demonstrates two self-similar coordinate generating mechanisms that help to provide cognitive accessibility, self-similarity at different scales and at the same scale. The coordinate systems directly represent short sequences of say 6 nucleotides or 3 amino acids and extend to longer sequences by connecting points with line segments. Variations can modify the space to produce simpler appearance. New visualization software will illustrate applications to gene regulation and peptide docking studies.