Minimal Entropy Probability Paths Between Genome Families

Calvin Ahlbrandt[1], Gary Benson[2] and William Casey[3]

Abstract

Use the motivation that nature might efficiently carry out mutations of genome snippets in a manner such that the increase in entropy involved in making the change is as small as possible. We characterize genome snippets by listing a probability vector in 4 dimensions, where the components of the probability vector are the probability of occurrence of each of the bases A, C, G and T. Then a genome family would be defined by all sequences, regardless of length, which have the same probability vector. Given two families with probability vectors $\mathbf{a}$ and $\mathbf{b}$, we define a distance function based as the infimum of path integrals of the entropy function $H(p)$ over all admissible paths $p(t)$, $0 \leq t \leq 1$, with $p(t)$ a probability vector such that $p(0) = \mathbf{a}$ and $p(1) = \mathbf{b}$. If the probability paths $p(t)$ are parameterized as $y(s)$ in terms of arc length s and the optimal path is smooth with arc length L, then smooth and "rich" optimal probability paths may be numerically estimated by iterating Newton's method on solutions of a two point boundary value problem, with unknown distance L between the abcissas, for the Euler–Lagrange equations resulting from a multiplier rule for the constrained optimization problem. Matlab code for these numerical methods is provided which works only for "rich" optimal probability vectors. These methods motivate a definition of an elementary distance function which is easier and faster to calculate, works on non–rich vectors, does not involve variational theory and does not involve differential equations, but is a better approximation of the minimal entropy path distance than the distance $||\mathbf{b} - \mathbf{a}||_2$.

[1] *Department of Mathematics, University of Missouri, Columbia, MO 65211–0001, calvin@math.missouri.edu*

[2] *Department of Biomathematical Sciences, Mount Sinai School of Medicine,1 Gustave L. Levy Place,New York. NY 10029,benson@ecology.biomath.mssm.edu*

[3] *Courant Institute, New York University, 251 Mercer St, NYC, NY-10012, wcasey@cims.nyu.edu*