RANK DISTANCE SHEDS LIGHT ON GENOME EVOLUTION

JOAO MEIDANIS AND LEONID CHINDELEVITCH

In this talk we discuss a representation of biological genomes as square, symmetric, orthogonal, 0-1 matrices. In this representation, the human genome, with its nearly 20,000 genes, would generate a 40,000 x 40,000 matrix.

It turns out that the rank distance applied to two genome matrices has biological significance: it is related to the smallest number of basic rearrangement mutations, such as reversals, translocations, transpositions (these with weight 2), etc. that explain the differences between the two genomes. Therefore, closer genomes will produce small rank distances.

Given k genomes, their evolutionary history can be described by a phylogenetic tree. Such trees can be estimated given the pairwise distances, but to reconstruct potential ancestors in the trees it is often necessary to solve the median problem: given 3 genomes A, B, and C, find a fourth genome M that minimizes $d(A, M) + d(B, M) + d(C, M)$. For the rank distance and genome matrices, this problem is NP-hard (Sarkis, personal communication). However, we present here fast algorithms that solve this problem exactly for orthogonal matrices, with practical experiments using both real and simulated data.

The first algorithm is based on a decomposition of $\mathbb{R}^n$ into a direct sum of linear subspaces related to $\ker(A - B)$, $\ker(B - C)$, and $\ker(C - A)$. The idea is that a candidate median should agree with A or B in $\ker(A - B)$, with B or C in $\ker(B - C)$, and with C or A in $\ker(C - A)$. In the complementary space, where A, B, and C do not agree with one another, we set the candidate median to mimic I, the identity matrix. The candidate matrix thus constructed is called M_I, and can be shown to be always a median. We present an $O(n^3)$ algorithm to compute M_I.

The second algorithm uses a "walk towards the median" paradigm. Starting from any of the input matrices, say, B, the algorithm produces rank-1 "steps", which are rank-1 matrices that, added to B, decrease its distance to both A and C simultaneously. It can be shown that such steps always exist for orthogonal matrices, and can be found in polynomial time. The algorithm stops when no more improvement can be done, which is equivalent to saying that B is between A and C in terms of the rank distance (the triangle inequality becomes an equality). We have an $O(n^4)$ algorithm implementing this idea, and recently devised a clever, $O(n^3)$ scheme that works for orthogonal, symmetric matrices.

Extensions for genomes with unequal content, abundant in practice, will be discussed as well.

University of Campinas, Campinas, Brazil
E-mail address: meidanis@ic.unicamp.br

Simon Fraser University, Burnaby, Canada
E-mail address: leonid@sfu.ca

Key words and phrases. Genome rearrangements, Breakpoints, Symmetric matrices.