第2回 2012年10月19日（金）15:40 ~ 16:30

Pairwise sequence alignment by probabilistic models: from simple, to non-linear, to giga-scale

Martin Frith / マーティン・フリス
Computational Biology Research Center (CBRC)
National Institute of Advanced Industrial Science and Technology (AIST)
産業技術総合研究所　生命情報工学研究センター

Pairwise sequence alignment is arguably the most fundamental task in computational biology. It is possible to view alignment in terms of probabilistic models, and this viewpoint leads to a surprising richness of useful applications. I will start by reviewing the basics: alignment score matrices as log likelihood ratios between two models, and the pair-HMM interpretation of gapped alignment. Among other things, this viewpoint enables us to estimate the reliability of every part of an alignment. We can build on this viewpoint to align DNA reads to a genome accurately. In particular, it straightforwardly allows us to estimate the confidence (e.g. 97%) that each alignment is to the true orthologous locus of the read. Building further, we can accurately align paired sequences (common in recent DNA sequencing), or perform "split alignment" of sequences against a genome. Split alignment means that different parts of one query sequence may align to different parts of the genome: this is relevant for mRNA or cancer DNA. In conclusion, the probabilistic model viewpoint provides an almost-mechanical way to extend alignment in all kinds of useful directions.

Keywords: alignment, genome, LAST, Bayesian, orthology