情報生命科学特別講義　第 7 回 2011 年 11 月 18 日（金）

Fast search algorithms applied to bioinformatics
高速検索技術のバイオインフォマティクスへの応用

Kouji Tsuda / 津田宏治

Due to rapid increase of biological data, simple data analysis such as clustering is often hard due to excessive computational cost. In this lecture, I introduce a fast method for all pairs similarity search called multiple sorting. In this method, a data point is first converted to a short string called sketches and similar pairs are found by iteration of masked sorting. As applications, I report hierarchical clustering of more than 10 million short reads, and the similarity-based analysis of more than 1 million ligand binding sites.

近年、情報爆発に伴い、科学研究におけるデータの量が急激に増加している。そのため、クラスタリング等の基本的な処理が、計算量超過のため実行できない事態となっている。本講義では、距離の近いデータ点のペアを全て発見する問題（全ペア類似度検索問題）に焦点をあて、そのための高速アルゴリズムである複合ソート法を紹介する。この方法では、まずデータ点を、スケッチと呼ばれる文字列に変換してから、マスク付きのソーティングを用いて高速に似たペアを発見することができる。応用としては、1000万点を超えるDNAリードの階層的クラスタリングと、100万点を超えるリガンド結合サイトの解析を紹介する。