第 10 回 2013 年 12 月 20 日（金）14:50 ~ 16:30

Next-generation data searching algorithms for future life science

これからの生命科学を切り拓く次世代検索アルゴリズム

Tetsuo Shibuya / 渋谷　哲朗

Human Genome Center (HGC)
University of Tokyo
東京大学医科学研究所　ヒトゲノム解析センター

In the big-data era, we need to develop more and more sophisticated methods for database searching. We will discuss a new algorithm design paradigm called SMAD (statistical model-based algorithm design) that may enable faster search over such large-scale databases. We will introduce faster searching / indexing algorithms based on the SMAD on protein 3-D structure databases. Searching for protein structure-function relationships represents a fundamental approach for determining the function of proteins with unknown functions. Our new indexing algorithm based on the SMAD enables queries in O(m + N/m^{0.5}) time, after O(N log N) preprocessing, where N is the database size and m is the query length. It is about 2 to 50 times faster than the previous practically best-known O(N) algorithm, which was also proposed by us, even if we include the preprocessing time. It is almost 20-1000 times faster than the naive comparison algorithm,
Keywords: indexing data structure, succinct data structure, statistical model-based algorithm design

ビッグデータ時代には、大規模データベースを検索する洗練されたアルゴリズムの整備がこれまでよりもさらに増して必要とされている。ここでは、統計的モデルを考慮する新しいアルゴリズム設計パラダイムＳＭＡＤを用いて、そのような大規模データベース検索を高速化する試みについて紹介する。我々はこのＳＭＡＤを用いて、新たな３次元立体構造検索のアルゴリズムは、O(N log N)のデータベース前処理をすることによって、O(m + N/m^{0.5}) 時間の高速検索が可能となる（N:データベースサイズ、m:クエリーサイズ）。これは、旧来の最高速のアルゴリズムよりも２～５０倍、現在でも最もよく使われる単純なアルゴリズムと比べれば２０倍～１０００ 倍もの高速化を精度の犠牲なしに実現している。
キーワード: 索引データ構造，簡潔データ構造，統計モデルに基づくアルゴリズム設計