

DA-2

森下 真一 (東京大学)

大規模生物データ処理のための並列データベース



次世代ゲノムシーケンサーの出力はハードディスクの進化を超える

ゲノム配列を解読するための次世代シーケンサー (Next-generation sequencer: NGS) の出力は、ハードディスクの容量増大を上回るペースで進化している。2010年現在、1ドル当たり100万塩基 (ACGTからなる文字列) を出力することができるようになり (図1)、生物の謎、遺伝子の機能、病気の原因の解明など、生物・医学を含む様々な分野で活用されるようになってきた。また、微生物のゲノム配列を解読することで、環境、バイオ燃料への応用なども進んできている。

さらに、データ量だけではなく、次世代シーケンサーがゲノムを解読するスピードも飛躍的に進化している。1日当りに出力できる塩基数は、2002年の200万塩基 (20MB/day) から、2010年では250億塩基 (25GB/day, Illumina HiSeq2000シーケンサーによる) にもなり、今後さらに1時間で1000億塩基 (100GB/hour, 2.4TB/day) に達するシーケンサーも登場する見込みとなっている。中国の北京ゲノム研究所 (BGI) では、既に128台のHiSeq 2000シーケンサーを導入しており、その1日当たりの最大出力は、およそ3兆塩基 (3TB/day) にも及んでいる。

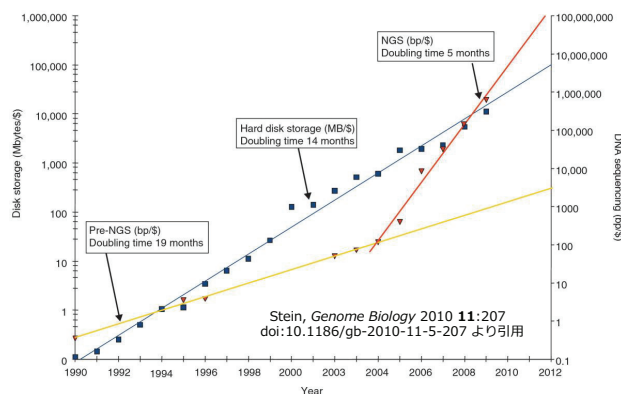
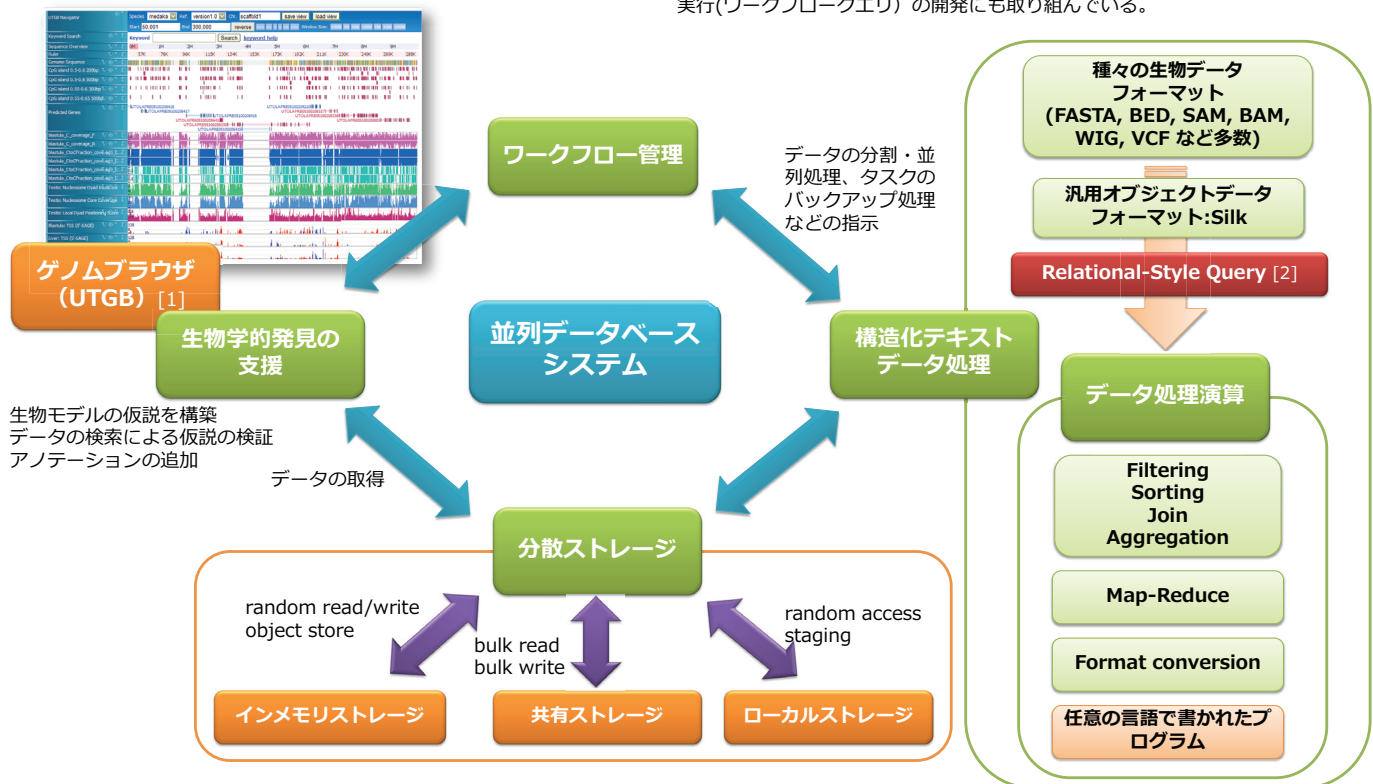


図1. 次世代シーケンサー (NGS) の登場によって、1ドル当たりで出力できる塩基数 (bp: base pair) の増加傾向は、ハードディスクの1ドル当たりのバイト容量の増加傾向を上回った。



関連文献

- [1] UTGB Toolkit for Personalized Genome Browsers. Taro L. Saito, Jun Yoshimura, Shin Sasaki, Budrul Ahsan, Atsushi Sasaki, Reginaldo Kuroshu and Shinichi Morishita. *Bioinformatics* (June 2009).
 [2] Relational-Style XML Query. Taro L. Saito, Shinichi Morishita. *ACM SIGMOD 2008*.