共同利用（産業利用トライアルユース：先端研究施設共用促進事業『みんなのスパコン』TSUBAME によるペタスケールへの飛翔） 成果報告書 平成24年度 課題種別

利用課題名 リガンドベースの仮想スクリーニングシステムの大規模システムによる実用実験
英文： Benchmarking of ligand-based virtual screening system on distributed systems

利用課題責任者 若林 登
Noboru Wakabayashi

所属 株式会社ヒューリンクス
Hulinks Inc.
http://www.hulinks.co.jp/

邦文抄録（300 字程度）
リガンドベースの仮想スクリーニングシステムは並列化が有効であるか、また、有効である場合、どの程度の高速化を実現できるかを測定した。ベンチマークの結果、111 万化合物のデータベースの少数のコアを使った理論上の計算時間を、並列化による実測で 4%以下にまで短縮することができた。本評価により、仮想スクリーニングの並列化による、創薬プロセスの短縮を実現する可能性が見えた。

英文抄録（100 words 程度）
We did benchmarking test of Ligand-based compound virtual screening system in order to confirm that parallel processing contributes to minimize computing time. We achieved 96% less calculation time with 1.11 million molecules than the calculation time of theoretically calculated by fewer numbers of cores. We can predict to reduce drug development process by large scale parallel computing.

Keywords: 仮想スクリーニング、並列化

背景と目的

創薬の現場では、実際に実験室で化合物を合成する前に、計算機を使用してその化合物の振る舞いをシミュレートをして候補を絞り込み、少数の候補化合物を実際に合成して創薬を進めることが多い。この時、数百万の化合物のデータベースの中から、高速に、かつ要求される精度で抽出できることが開発競争に置ける鍵となる。

ベンチマーク実験に使用したシステムは英国 Cresset BioMolecular Discovery Ltd. 製の blazeV10 である。これはリガンドの 3 次元のフィールド（化合物の静電、形状などを表現した「場」）を、データベース内のすべての化合物のフィールドと照合し、その結果をスコアリングして、条件に高く一致するものからソートするものである。

概要

本試験では、約 111 万の化合物が登録されているデータベースを、計算エンジンとは別に、ノードを固定し、リレーショナルデータベースシステム（RDBMS）として設置した。このデータベースに対して、指定されたリガンドを使い、指定されたコアの数に従ってデータベースを分割してその結果を Web インターフェイスを通して返すものである。RDBMS には MySQL を使用した。

少ないコア数では、使用規定時間内に計算が終了しないため、化合物の数を 10%にして処理時間を計測し、それを 10 倍したものを理論上の計算時間と仮定した。

1 ノード（12 コア）での理論上の計算時間、48.5 時間が、80 ノード（480 コア）では 1.5 時間で計算が完了するようになった。

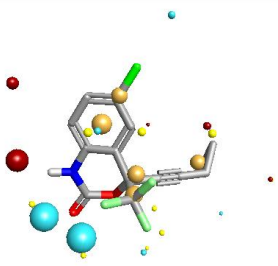

図1：試験で使用したリガンド（物質名：Efavirenz；非ヌ

クレオシド系逆転写酵素阻害薬）のフィールド

結果および考察

まず、理論上の計算時間を見積もるため、2つの基本試験を行った。

試験1：同一コア数でノードを分散することによる計算時間への影響

試験2：同一コア数、同一ノードで対象化合物数を増加させた時の変化

これは、予想計算時間が使用規定時間である 24 時間を超えることが予想されるため、理論上の計算時間の算出が目的である。

試験1の結果

ノード数	コア構成	コア数	化合物数	計算時間
1	12	12	1.1 万	0:29:51
2	6+6	12	1.1 万	0:27:16
3	4+4+4	12	1.1 万	0:27:15
4	3+3+3+3	12	1.1 万	0:27:15

試験2の結果

ノード数	コア構成	コア数	化合物数	計算時間
10	12	120	1.1 万	0:06:16
10	12	120	5.5 万	0:18:17
10	12	120	11.1 万	0:34:17
10	12	120	27.7 万	1:24:26
10	12	120	55.5 万	2:52:48
10	12	120	111 万	6:04:52

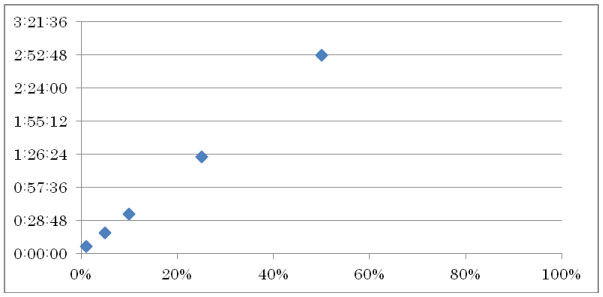

図2：同一条件下での化合物数と計算時間

試験1の結果により、ノード分散による影響は小さいことが分かる。

X 軸を 111 万化合物を 100%として計算時間をプロットすると、ほぼ線形に計算時間が増加していることがわかる。

これらの結果から、1 ノード 12 コアで 111 万化合物をスクリーニングすると、60 時間（10 ノードで 6 時間より）かかると推測した。

これを踏まえたうえで、111 万化合物に対して、並列度を増し（コア数を増加）、計算時間を測定した。

ノード数	コア構成	コア数	化合物数	計算時間
10	12	120	111 万	6:04:52
20	12	240	111 万	3:04:49
30	12	360	111 万	2:07:33
40	12	480	111 万	1:42:30
[参考]80	6	480	111 万	1:36:24
50	12	600	111 万	1:49:35
100	12	1200	111 万	2:15:52
200	12	2400	111 万	3:03:20

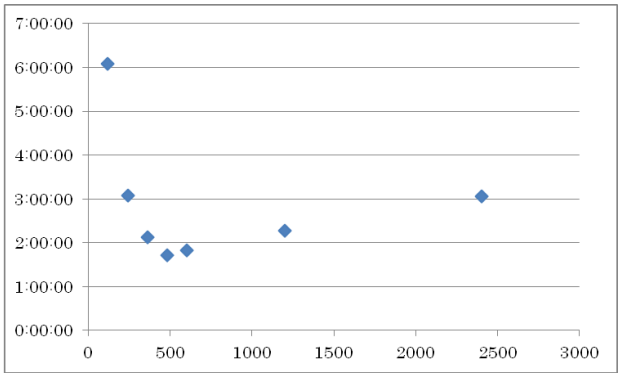

図3：計測結果（2400 コアまで）（X: コア数／Y: 時間）

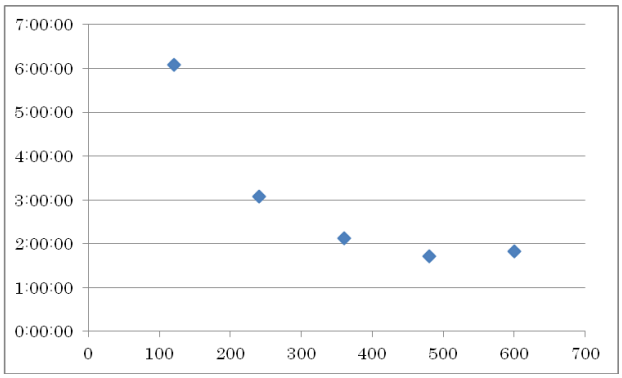

図4：計測結果拡大図（600 コアまで）（X: コア数／Y: 時間）

以上の結果より、現在のソフトウェアの仕様では一定レベルまでの並列化により、大幅な計算時間の短縮が図れることが分かった。

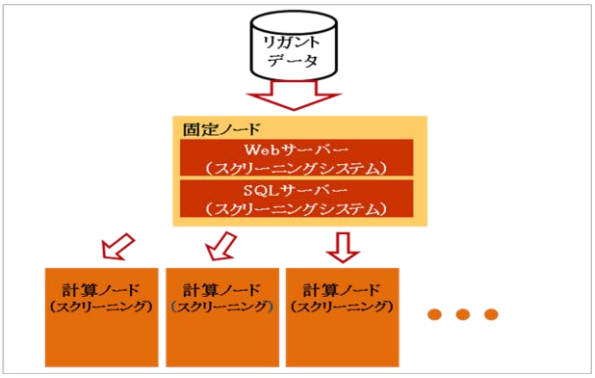

図5：システム概念図

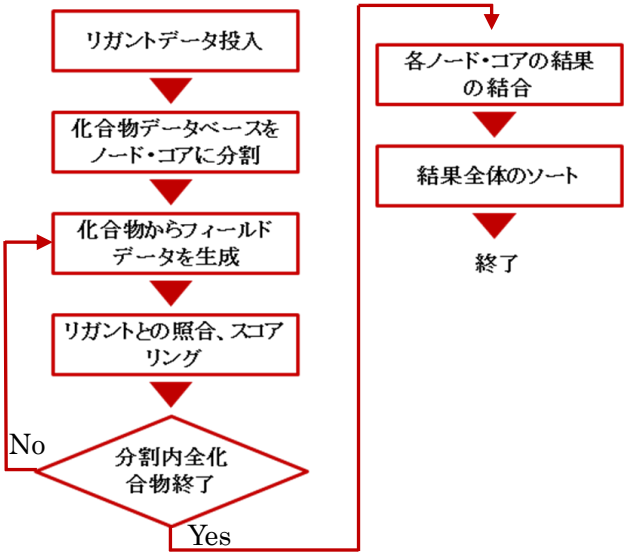

図6：処理フロー

まとめ、今後の課題

最終の結果の 50 ノード、600 コアの計算時間、あるいはそれ以上の並列化の結果は、40 ノード、480 コアより処理に時間がかかっている。これは本システムにおける、処理の分割、処理の結合とそのソートに伴うオーバーヘッドが徐々に増すことを示している。

これは、対象とする化合物数がさらに大規模（例えば、化合物のデータベースを 500 万化合物にする等）になれば、このオーバーヘッドの影響が発生するノード数、コア数は多くなると予想するが、今回の試験では行わないものとする。

本試験により、創薬プロセスの短縮に並列化は有効であり、スーパーコンピューターの機能が貢献することを示している。