生命基盤ソフトウェア開発・高度化チーム成果報告

MD コアソフトウェア
および
ソフトウェア高度化

泰地　真弘人
次世代計算科学研究開発プログラム
生命体基盤ソフトウェア開発・高度化チーム チームリーダー

発表者紹介

1992 年 3 月　東京大学理学系研究科物理学専攻博士課程修了
1992 年 4 月　東京大学教養学部助手
1997 年 1 月　統計数理研究所助教授
2002 年 4 月　理化学研究所ゲノム科学総合研究センターチームリーダー
2008 年 4 月　理化学研究所基幹研究所グループディレクター
2011 年 4 月　理化学研究所生命システム研究センター生命モデリングコアコア長

研究分野

高性能計算・計算機アーキテクチャ

MD コアソフトウェアおよびソフトウェア高度化

泰地　真弘人
次世代計算科学研究開発プログラム
生命体基盤ソフトウェア開発・高度化チーム チームリーダー

1. 目的

京コンピュータは、80，000 プロセッサ、640，000 コア以上を有する大規模な超並列計算機です。このような規模の計算機を使いこなすためには、並列処理、特に京コンピュータに通暁した研究者による支援体制が必要です。そのために、分子動力学計算コアソフトウェアの開発などの独自開発を通じた人員育成を行うと同時に、他チームで開発されたソフトウェアの性能評価・開発支援を行っています。さらに、開発容易化のためのミドルウェア開発・可視化ソフトウェア開発を行い、産業利用に向けた大規模仮想化合物ライブラリの開発を進めています。

2. 現時点での成果

2.1 分子動力学コアソフトウェア

大規模に並列化された分子動力学ソフトウェア開発を、高度化人員の育成も兼ねて行っています。京コンピュータ上での大規模並列化を達成し、ピーク性能 3.5PFLOPS 相当の京コンピュータを用いて 1.315PFLOPS の性能を達成しました。このときの実行効率は 37%であり、高い性能を達成しました。

2.2 ソフトウェア高度化

ソフトウェアの性能評価・チューニング支援を行っています。第一走者アプリケーションを中心に、性能評価報告を行いました。チューニング支援については、特に神戸拠点における京に向けての開発支援を中心に、開発実施本部との窓口・京に向けたチューニング外注の窓口機能を担っています。

2.3 可視化ソフトウェア

大規模データを扱える並列可視化システム ISL-LSV を開発しました。GPU を用いた高速レンダリングで 100GB のデータに対して 100GPU 並列を用いて数 fps の描画速度で表示し，基本的な結果表示を行えます．また，リモート環境での可視化にも対応しています.

2.4 並列処理ミドルウェア

ソフトウェア開発容易化のためのミドルウェア Sphere を開発しました。連続体シミュレーションを中心に、ソルバ群の組み合わせにより容易に並列化されたソフトウェア開発を可能にします。本ソフトウェアは臓器全身スケール研究チームで開発されている ZZ-EFSI/ZZ-HIFU、細胞スケール研究チームで開発されている RICS で用いられています。

2.5 大規模仮想化合物ライブラリ

反応データベースから抽出された反応前後の反応部位構造変化トランスフォームを活用し、標的構造に対する前駆体構造を反応スキームとして提示できるシステムを連続運用することによって、合成ルートを付与した大規模バーチャルライブラリ構築システムを開発しました。このバーチャルライブラリに含まれる化合物は、それぞれの合成前駆体構造との関連を保持したツリー構造を形成し、入手可能な化合物に至る合成経路の提示や、複数の合成経路コスト評価の可能性を有しており、新規の薬剤開発に有効であると期待されます。

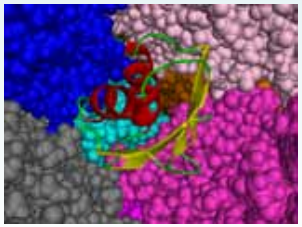
2. プロジェクト終了時の達成目標

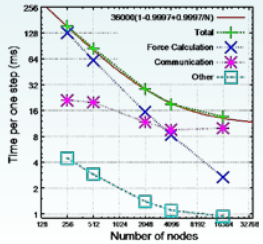
プロジェクト終了時に、より多くのアプリケーションが京コンピュータ上で十分な性能を発揮していることが我々の目標です。分子動力学コアソフトウェアでは、強スケーリング性能の向上に取り組み、50 粒子/core 程度までのスケーリングの達成を目指します。また、創薬応用に向けた 10 億化合物規模の大規模化合物ライブラリを完成させます。

以上

MDコアソフトウェア およびソフトウェア高度化

理化学研究所
次世代計算科学研究開発プログラム
生命体基盤ソフトウェア開発・高度化チーム
チームリーダー
泰地　真弘人





謝辞・注意事項

- 京での計算に関しては京速コンピュータ京の試験利用、および本年3月 での特別運用での結果です。また、PCクラスタでの性能計測に関しては理化学研究所情報基盤センターのRICCを使用しています。
- 京の開発は現在進行中であり、ここで示す性能は暫定値です。

ISLiM成果報告会2011

背景・目的

- 背景
  - 京=64万コア：大規模並列化の要請
  - ソフトウェア開発の困難・高性能計算技術の必要性
  - 京利用上のノウハウの蓄積
- 目的
  - 各チームのソフトウェアの高度化
  - 需要の多いコアソフトウェアの開発
    MDコアソフトウェア

MDコアソフトウェア　概要・アプローチ

- 研究開発コードの概要
  - 古典分子動力学計算
  - 京コンピュータ
    - 大規模並列
    - MPI/OpenMP　ハイブリッド並列
    - SIMD
    - 3(+3)Dトーラスネットワーク(TOFU)
  - 再利用性
- アプローチ
  - 空間分割
  - PME以外の遠距離相互作用計算
  - C++
  - 通信パターン

現在までの研究開発成果

- 機能
  - 空間分割
    - セルインデックス法
  - 結合
    - AMBER力場
    - （CHARM力場）
  - 短距離相互作用計算
    - ペアリスト方式
  - 遠距離相互作用
    - PME
    - （FMM, MultiGrid, STGPME）
- 性能
  - 1.315 PFlops、 効率 37.16%
  - 27,648ノード、180,881,424原子

京は現在開発中であり、本性能は暫定値です。

現在までの研究開発成果

- 問題点
  - SIMD
    - 真率が低いループ内IF文
  - 開発途上のコンパイラ(C++)
    - 最適化の制約
  - 同時通信の競合
    - ノード当り最大342ノードとのSend/Receive

現在までの研究開発成果

- 対策
  - 真率が低いループ内IF文
    - ペアリスト方式
      - IF文が真になる可能性が高い原子ペアのリストを作る
  - 最適化の制約
    - 計算カーネル部分を開発が進んでいたFortranで書き換えた。
    - 最適化が行なわれるデータ構造へ変更した。
  - 同時通信の競合
    - トーラスの各軸別に逐次通信

現在までの研究開発成果

- Weak scaling (418,707 atom / 64 node)

# of node	64	256	512	4096	13824	27648
# of atom	418,707	1,674,828	3,349,656	26,797,248	90,440,712	180,881,424
Total	152.84	153.66	153.34	153.88	153.92	154.29
Force Calculation	128.21	128.21	128.36	128.34	128.06	128.15
Communication	20.19	20.99	20.50	20.87	21.21	21.55
Other	4.45	4.47	4.48	4.67	4.64	4.60

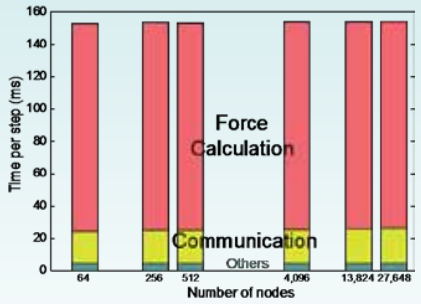
- Strong scaling (1,674,828 atom)

# of Node	256	512	2048	4096	16384
# of Import cell	936	768	504	446	342
# of Communication node	26	44	124	174	342
Total	153.66	84.63	28.57	18.98	13.58
Force Calculation	128.21	61.78	15.42	8.33	2.68
Communication	20.99	19.96	11.75	9.55	9.96
Other	4.47	2.89	1.40	1.10	0.94

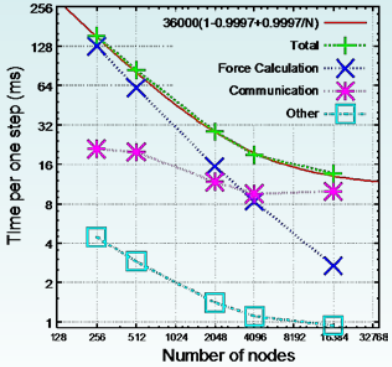
京は現在開発中であり、本性能は暫定値です。

現在までの研究開発成果

- Weak scaling



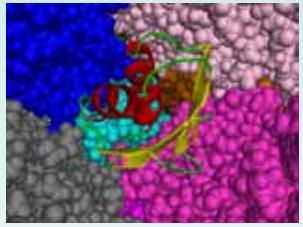
- Strong scaling



京は現在開発中であり、本性能は暫定値です。

現在までの研究開発成果

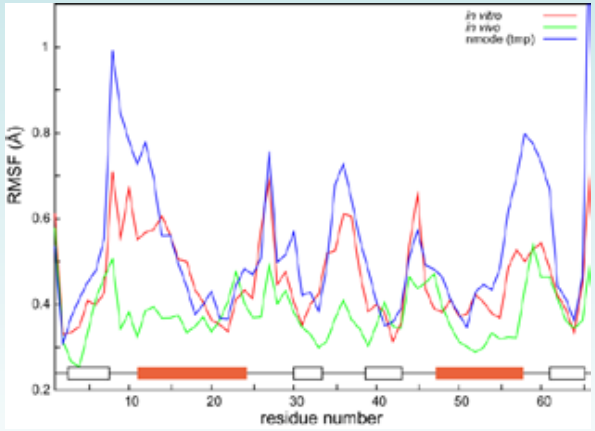
- 計算事例
  - Proteins in a Crowded Environment
  - 8 TTHA + 64 ovalbumin + 316,952 water
    - 1,344,488 atom, 240 Å
    - タンパク質が30%　細胞内環境
  - 4 TTHA + 137,977 water
    - 418,707 atom, 160 Å



  - TTHA1718 1007 atom
  - Ovalbumin 5998 atom

[1] Rivas, G., Ferrone, F. and Herzfeld, J. Life in a crowded world. *EMBO reports*, 5, 1 (Jan 2004), 23-27.
[2] Chebotareva, N. A., Kurganov, B. I. and Livanova, N. B. Biochemical effects of molecular crowding. *Biochemistry. Biokhimiia*, 69, 11 (Nov 2004), 1239-1251.
[3] Chebotareva, N. A., Andreeva, I. E., Makeeva, V. F., Livanova, N. B. and Kurganov, B. I. Effect of molecular crowding on self-association of phosphorylase kinase and its interaction with phosphorylase b and glycogen. *J. Mol. Recognit.*, 172004), 426-432.

現在までの研究開発成果

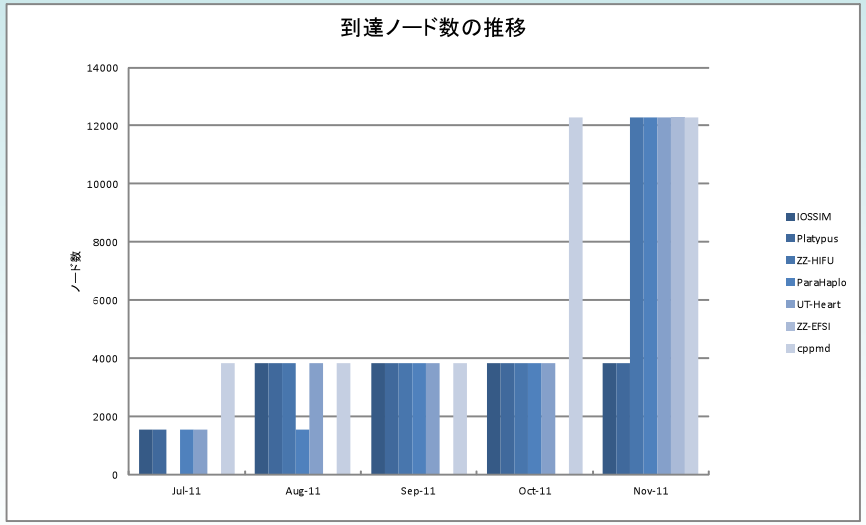


- Conformational fluctuations of TTHA for in vivo (viv) and in vitro systems (vit1, vit2). The green, red, and blue curves indicate the root mean square fluctuation (RMSF) of TTHA for in vivo and in vitro systems. The abscissa axis is the residue number of TTHA and vertical axis is RMSF value (in Angstrom). Red and white boxes indicate alpha-helices and beta-sheets, respectively.

プロジェクト終了時の達成目標

- 800万原子（ノードあたり100原子）で1step /10ms
  - 2 fs / step で 0.7 ns / day
- 5億原子（ノードあたり6000原子）で3PFlops
  - 160^3 nm^3

ISLiMコードの高度化支援



ISLiMコードの高度化支援

移植
- RICC→FX1→京
- クロス環境への対応
- 富士通コンパイラへの対応
- 多段階ビルド

単体チューニング
- SIMD化
- ソフトウェアパイプライニング
- スレッド化 OpenMP
- ループ最適化

並列チューニング
- 通信パターン最適化
  - 直接、XYZ間接
- 通信時間隠蔽
- スレッド同期の削減
- データ分割の見直し

その他
- アルゴリズム変更
  - FFT
- 分散入出力方式
- メモリ使用量削減

ISLiMコードの高度化支援

	移植	単体チューニング	並列チューニング	その他
分子	○ 富士通コンパイラ対応(Cafemol)	○ SIMD, Pipeline (Cafemol)		
細胞				
臓器全身	○ Vsphere(ZZ-EFSI)	○	○	
脳	○ ビルド方法の構築 (IOSSIM) 富士通コンパイラ対応(NEST)	○ thread化 (NEST) SIMD化(IOSSIM)		○ メモリ消費削減 (NEST)
データ解析融合		○ SIMD, Pipeline (LiSDAS)		○ 分散I/O (Parahaplo) , FFT (MegaDock)
高度化	○	○	○	

アプリケーションミドルウェア SPHERE と 大規模データ可視化 LSV

野田　茂穂
独立行政法人　理化学研究所
次世代計算科学研究開発プログラム

発表者紹介

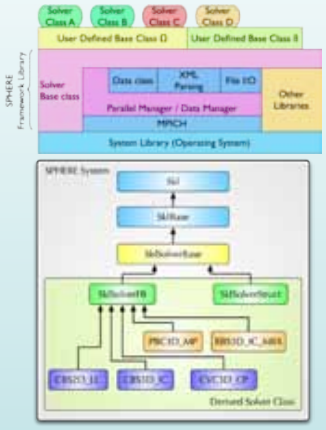
1991 年 3 月　信州大学工学部機械工学科卒業
1991 年 4 月　株式会社富士通長野システムエンジニアリング入社
2009 年 4 月　独立行政法人理化学研究所　情報基盤センター

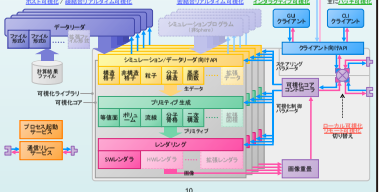
研究分野

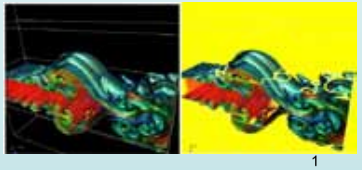
CFD、バイオエンジニアリング、並列プログラム、可視化システム

SPHERE (Skelton for Physical and Engineering Research)
LSV (Large Scale Visualization)

東京大学　小野謙二
理化学研究所　野田茂穂

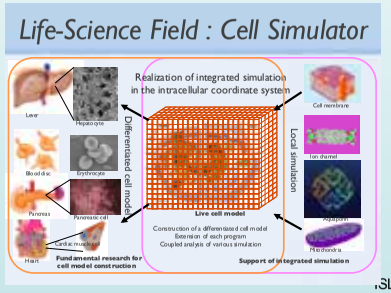


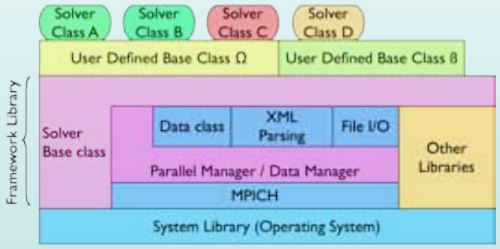




背景・目的(SPHERE)

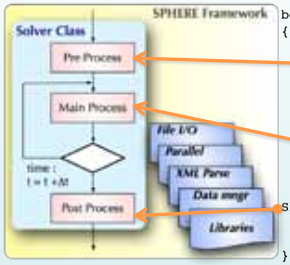
- 背景(シミュレーションプログラムは。。。)
  - シミュレーション対象がMulti-physics/Multi-scale化する事によりコードが複雑化
  - HPCアーキテクチャーはMulti-core化、ヘテロジーニアス化など複雑で多様化
  - 新しいアルゴリズムを用いてHPCの恩恵を得るには、物理と計算機の両方のスキルが必要
- 目的
  - アルゴリズム研究開発に専念できる環境を提供
  - 多様化するアーキテクチャーのローカライズを吸収





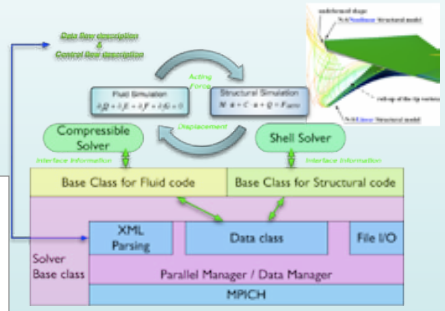
概要・アプローチ

Code Skeleton



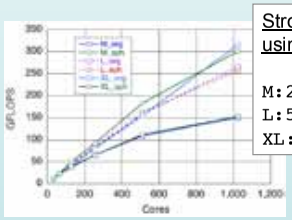

```
bool SklSolverBase::SklSolverExec()
{
  int init_ret =
    SklSolverInitialize();
  ...
  int loop_ret =
    SklMainLoop();
  ...
SklSolverPost();
  return true;
}
```

Coupling Solver Control on SPHERE



Strong scaling on RSCC using Poisson Solver

M:256x128x128
L:512x256x256
XL:1024x512x512



現在までの開発成果と今後

- Chip: Intel, AMD, SparcⅧfx, SX, PowerPC
- on RICC : gnu compiler, Intel compiler, Fujitsu compiler, Fujitsu MPI, mpich, OpenMPI, OpenMP
- on K : Fujitsu compiler, Fujitsu MPI, OpenMP
- on BG/L : …

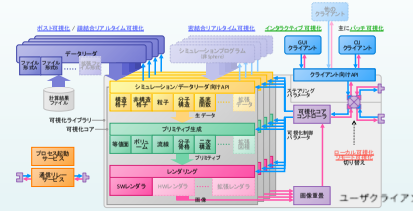
- ISLiMでの適用アプリケーション : 構造格子ステンシル計算
  - ZZ-EFSI
  - ZZ-HIFU
  - RICS

- 今後。。。 :
  - これまでは、ツール開発の視点でベースとなる環境の開発を遂行
  - これまでの開発を元に、非構造格子ステンシル計算、粒子系への拡張、演算ライブラリの組込みなど、まだ研究と開発を両輪とした成果を組み込み提供していく

背景・目的(LSV)

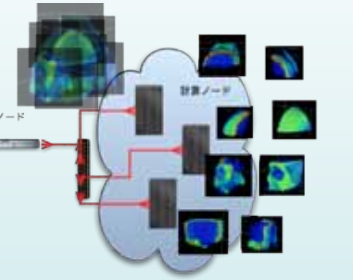
- 背景(可視化)
  - 可視化の目的はデータの理解促進。重要なのはインタラクティブ性と品質。
  - 大規模な計算は大規模なデータを創出し、その理解に可視化は欠かせない。
  - 従来のアプリケーションでは大規模データを可視化し、データを理解する事が困難。
- 目的
  - データの増大に対してスケーラブルな可視化システムの開発
  - インタラクティブ性(5fps)、UI、分散データ、GPU有無、リモートアクセス等への対応

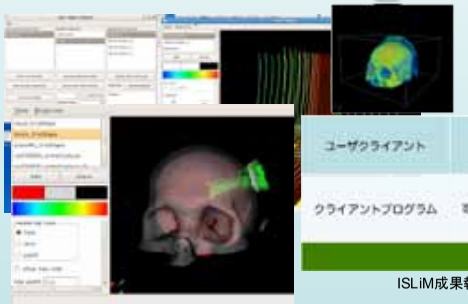
概要・アプローチ

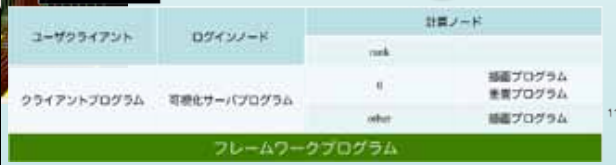


論理設計

LSVシステム概要

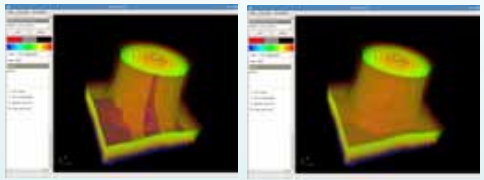




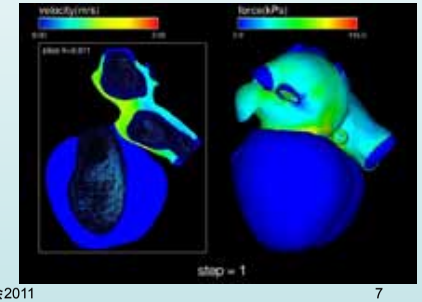


現在までの開発成果と今後

- 重畳処理の性能とクオリティの確保が課題:クオリティレベルをユーザが選択可能



- GPGPUクラスタでのハードウェア並列レンダリング:100ノードで100GBデータを数fps。
- ソフトウェアレンダリングライブラリの適用による、スパコン本体可視化の実現。(RICC)
- 今後:「京」へのLSVの移植及び可視化ライブラリの拡充



まとめ

- SPHERE :
  - シミュレーションアルゴリズム開発者が容易に並列プログラムを構築するツール
  - SPHERE利用による多少のの並列性能低下よりも生産性の向上のメリットが大
  - ISLiMの臓器全身チーム、細胞チームをターゲットとして開発を行い、今後は他のチームでも適用できる物を目指す
- LSV :
  - これからの大規模データに対応する可視化ソフトウェアとして開発
  - 分散データ対応や高速なレンダリングの視点から、並列可視化システムとして設計
  - ハードウェアレンダリングだけでなくソフトウェアレンダリング機構を持つ事により、スパコン本体を用いた大規模な可視化処理にも対応
  - レイトレースによる高品位レンダリングや先進的な可視化アルゴリズムを実装

※「京」は現在開発中であり、性能や動作状況は現状の値です。
※京での実行に関しては、京速コンピュータ京の試験利用での結果です。
※GPGPUクラスタ及びPCクラスタは理化学研究所情報基盤センターのRICCを利用しています。

創薬プラットフォーム
－大規模バーチャルライブラリの開発－

船津公人
東京大学大学院工学系研究科
化学システム工学専攻　教授

発表者紹介

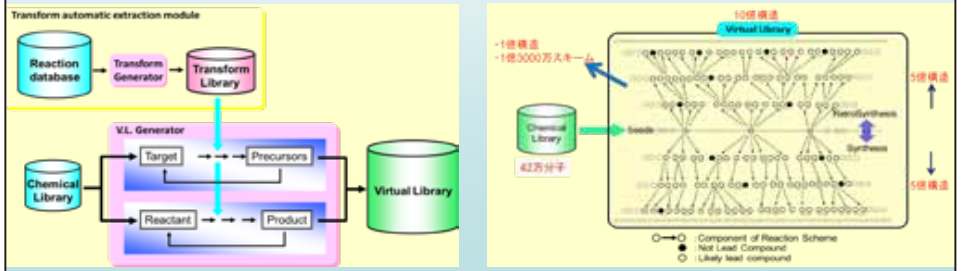
1983 年 3 月　九州大学大学院理学研究科化学専攻博士課程修了
1984 年 3 月　豊橋技術科学大学工学部物質工学系　助手
1988 年 3 月　豊橋技術科学大学工学部知識情報工学系　助手
1992 年 4 月　豊橋技術科学大学工学部知識情報工学系　助教授
2004 年 4 月　東京大学大学院工学系研究科　教授

研究分野

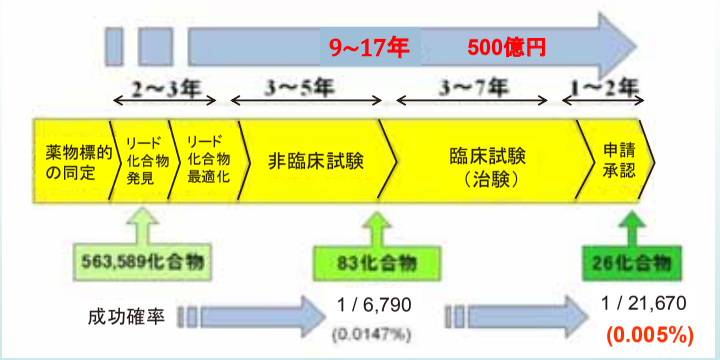
ケモインフォマティックス（化学情報学）
関連する研究分野としてプロセスシステム工学、有機合成化学、分析化学、触媒化学、創薬化学など

発表創薬プラットフォーム
－大規模バーチャルライブラリの開発－

東京大学
大学院工学系研究科化学システム工学専攻
教授
船津公人



新薬開発の特質 （03年～07年）



てきすとぶっく 製薬産業2009

- R&Dリスクが高い
  長期、投資額、成功確率・・・
- 漏れの懸念
  たまたま手に入った化合物の中から選択している・・

化合物ライブラリの現状

■実物

・米国

大学	化合物数
Broad Institute	＞250,000
UCSF	＞150,000
Stanford	＞130,000
Kalamazoo Valley Community College	＞100,000
UCLA	＞ 75,000

・韓国：ナショナルバンク
17万化合物

・東京大学： 2010年3月
196,970化合物

■DB

REAL Database
1500万化合物

Drug-like selection
400万化合物

purchasable
1700万化合物

5530万化合物

■VL

everything
2800万化合物

purchasable
1700万化合物

drug-like
1000万化合物

化合物ライブラリの問題点

■規模と多様性

創薬の研究対象となる化合物
C、N、O、S からなる30原子まで[1] : 10^{60}個

スクリーニング用実在化合物[2] : 5.5×10^{07}個

$$\frac{1}{10^{53}}$$

全体のごく僅かな部分からヒット確率を上げるには、
全体空間からできるだけ万遍なくサンプルを揃えるべき・・・

■入手可能性

・メガファーマが保有する化合物(1990年代) : 10^{06}個

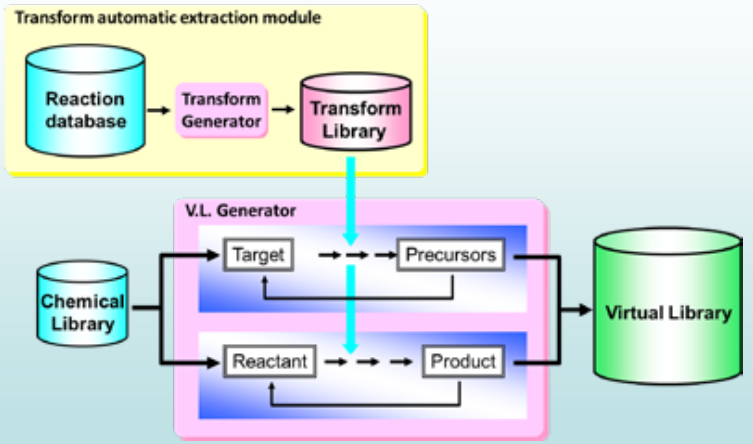
・バーチャルライブラリ
バーチャルスクリーニングで高スコアを得た化合物であっても
合成検討にコストがかかる

目的

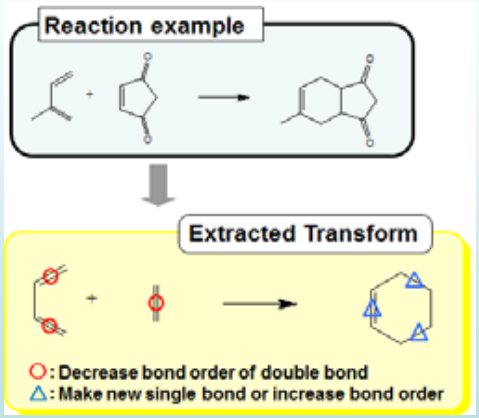
■リード構造探索のための
大規模な化合物仮想ライブラリ（VL）の構築

■VLに含まれる仮想化合物の多様性の確保、および
仮想化合物の合成ルート情報の提供

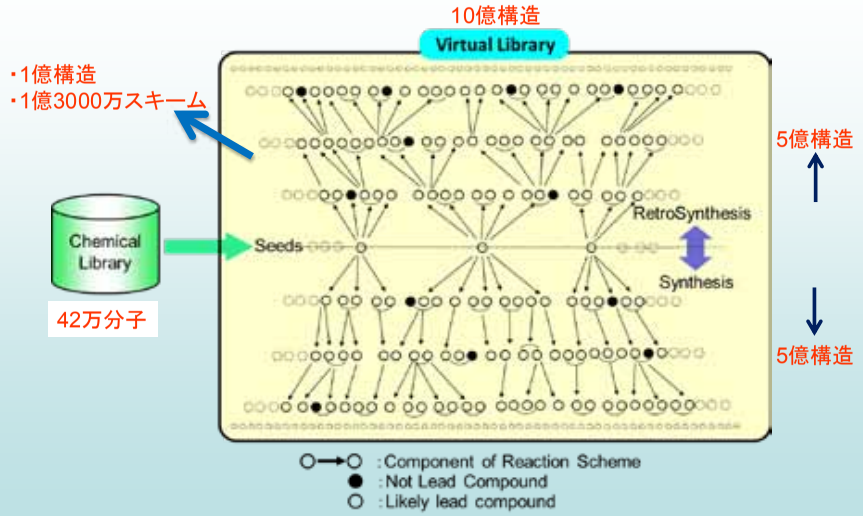
システム概要



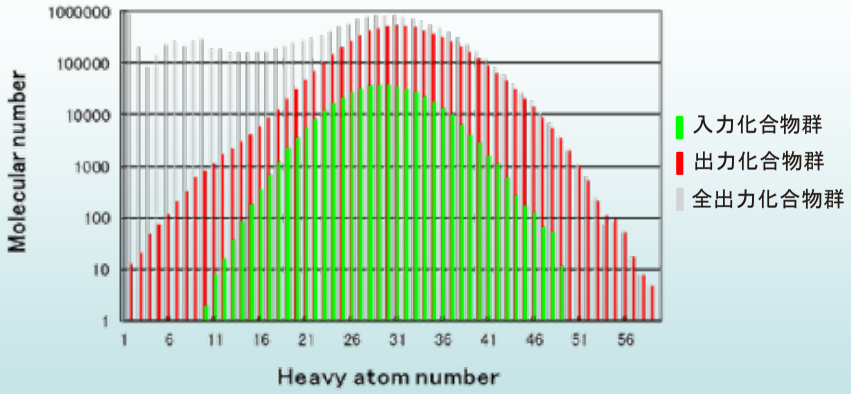
Transformの抽出例



バーチャルライブラリの構成と目標



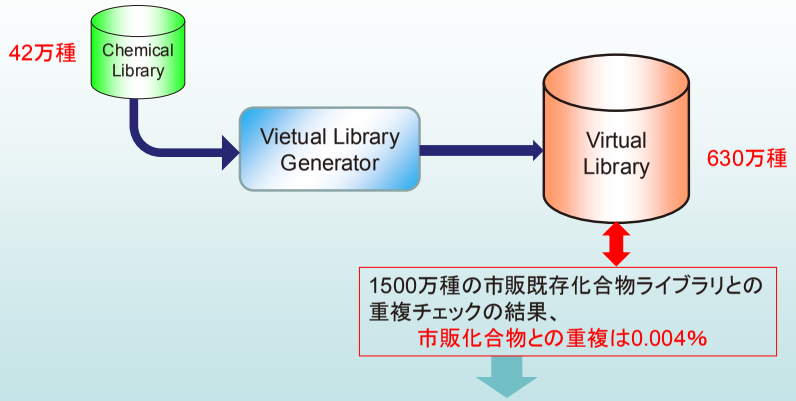
出力化合物の重複率



出力化合物の重複率： 58%
42万種の入力構造群から1500万の化合物を発生し、ユニークな構造は630万種

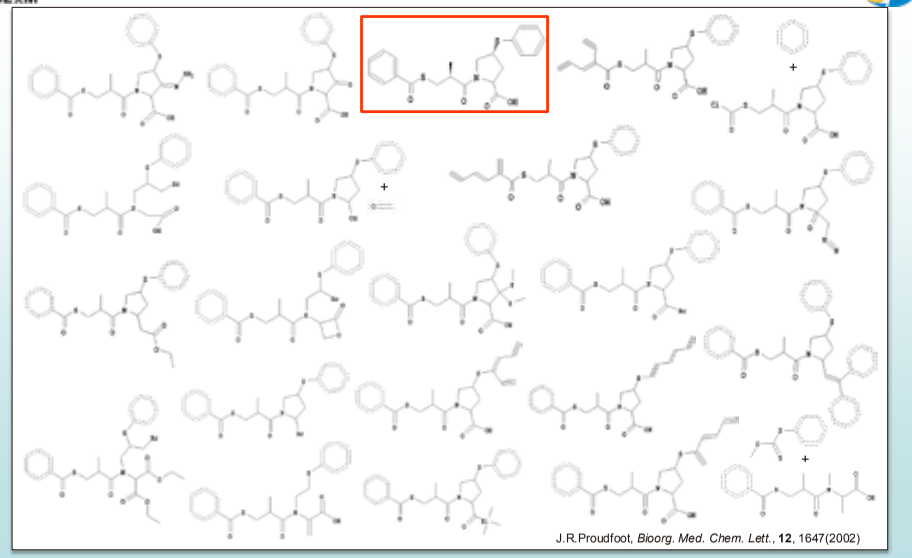
"Life Chemicals (Virtual)" subset (# 27) – 490709 entries

発生化合物群の新規性



当システムがライブラリに出力する仮想化合物は
ほとんどが新規化合物

構造出力例



Lipinski's rule of 5, rule of 3

	rule of 5	rule of 3
分子量	＜500	＜300
水素結合供与体数	≦5	≦3
水素結合受容体数	≦10	≦3
分配係数計算値（cLogP）	≦5	≦3
回転可能結合数	≦10	≦3
極性表面積（Å^2）	≦140	≦60

・Rule of 5 ： 経口吸収性の高い構造特性
・Rule of 3 ： リード化合物 (フラグメント化合物) として相応しい構造特性

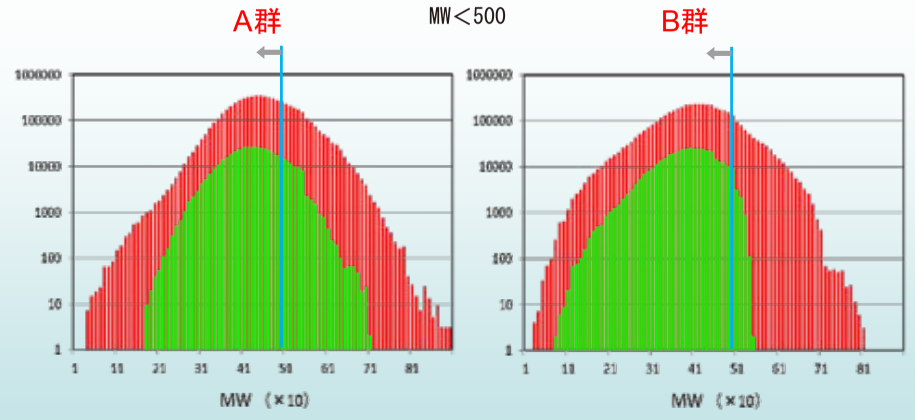
[1] Lipinski, C., et. al., *Adv. Drug. Deliv. Rev.*, **23**, 3(1997)
[2] Veber et al.,*J. Med. Chem.* **45**, 2615-2623 (2002)

Lipinski's rule of 5　どのくらい満たす？

	rule of 5	input	output
分子量	＜500	85.6%	75.3%
水素結合供与体数	≦5	100.0%	99.9%
水素結合受容体数	≦10	99.9%	99.7%
分配係数計算値（cLogP）	≦5	79.9%	79.9%
回転可能結合数	≦10	98.2%	93.0%
極性表面積（Å^2）	≦140	99.0%	96.4%

"Life Chemicals (Virtual)" subset (# 27)

分布の異なる種構造群の活用検討



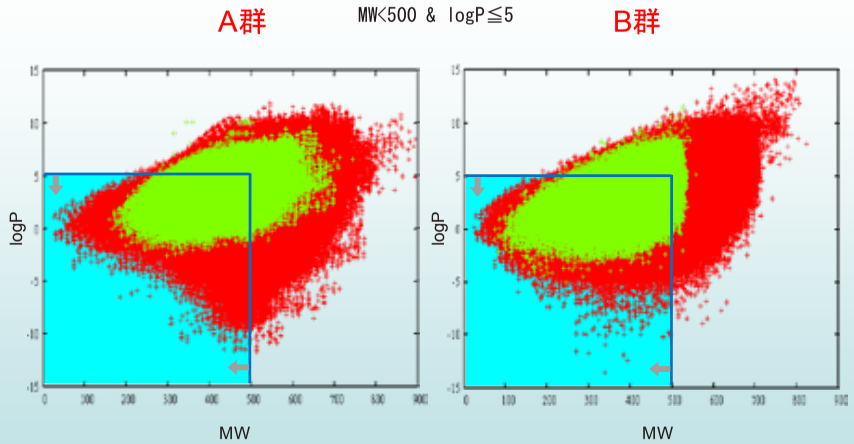
入力化合物群A　85.6%
出力化合物群　75.3%

入力化合物群B　98.5%
出力化合物群　88.1%

"Life Chemicals (Virtual)" subset (# 27) – 490709 entries

"Life Chemicals" subset (# 124) - 412431 entries

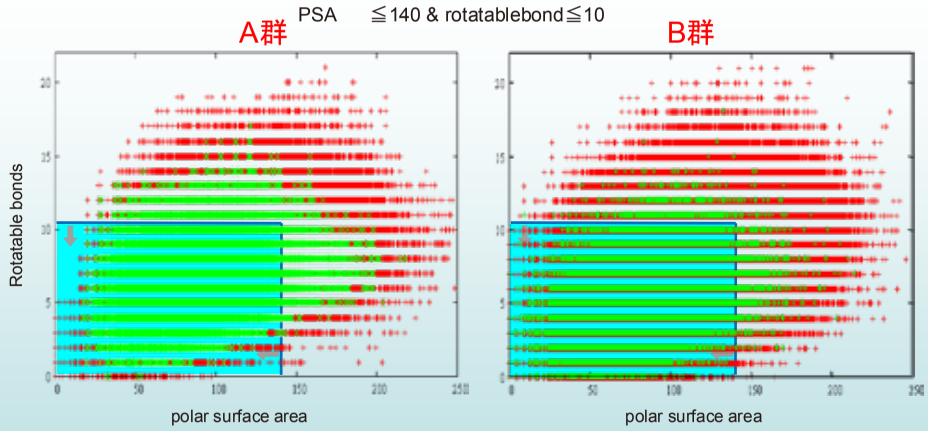
入力化合物群の特性分布の影響



× 入力化合物群 A　70.7%
＋ 出力化合物群　64.8%

× 入力化合物群 B　88.3%
＋ 出力化合物群　79.3%

極性表面積と回転可能結合数の関係



× 入力化合物群 A　97.3%
＋ 出力化合物群　89.9%

× 入力化合物群 B　98.9%
＋ 出力化合物群　91.0%

入力化合物群の特性分布の影響

	rule of 5	入力化合物群 A input	入力化合物群 A output	入力化合物群 B input	入力化合物群 B output
・分子量	＜500	85.6%	75.3%	98.5%	88.1%
・水素結合供与体数	≦5	100.0%	99.9%	99.9%	99.9%
・水素結合受容体数	≦10	99.9%	99.7%	99.9%	99.6%
・分配係数計算値（cLogP）	≦5	79.9%	79.9%	89.3%	87.0%
・回転可能結合数	≦10	98.2%	93.0%	99.5%	95.0%
・極性表面積（Å^2）	≦140	99.0%	96.4%	99.4%	95.1%
・分子量 VS. 分配係数計算値		70.7%	64.8%	88.3%	79.3%
・極性表面積 VS. 回転可能結合数		97.3%	89.9%	98.9%	91.0%
・水素結合供与体数 VS. 水素結合受容体数		99.9%	99.7%	99.9%	99.6%

出力構造の特徴分布は、入力構造群の特徴の影響を受ける
→ 適切な種構造群を活用すれば、適切な出力構造群が得られる可能性が増す

まとめ

化合物群を種とし、反応DBより抽出したTransformを活用して
合成経路情報を含む大規模バーチャルライブラリを構築した。

創生した化合物は、入力構造群の特徴を引き継ぎ、
薬となりうる化合物を種とすると、薬となりうる構造を出力する傾向を認めた

■新規性:
市販化合物ライブラリとの対比で 重複0.004%、 新規性の確保を確認済み
→ 規模を拡大して確認する

■分布: （ Lipinski's ruleの指標によるチェック）
新規な化合物構造からなる出力構造群は、
入力構造群の分布を若干拡張するが、 特性を反映している
→ 多様性の評価を進める

■規模:
入力化学構造群に対して、
順合成方向、逆合成方向にそれぞれ3段階、
各方向で5億件の化学構造による
総10億化合物からなるバーチャルライブラリを構築し、ライブラリの特性評価を進める

（謝辞）
本資料集に記載されている「京」での計算は、2011 年 3 月の「京」の特別運用およびその後の試験利用によって行われたものです。

また、本資料集に記載されている「京」を使用した測定値は、開発整備中の「京」による、測定時点での数値です。