

9 スーパーコンピュータ「京」が拓く コンピュータ創薬の未来

中津井雅彦^{1,3} 奥野恭史^{2,3,4}

1 神戸大学大学院工学研究科 2 京都大学大学院医学研究科 3 理化学研究所計算科学研究機構 4 先端医療振興財団先端医療センター研究所

医薬品の開発は、10年以上の長い年月と、500億円以上の巨額の費用がかかると言われている。またこの十数年、製薬業界では、世界的規模で、新薬の承認数が横ばい状態（20品目程度／年）であるのに対し、研究開発費が増え続けるという深刻な問題に直面している。この医薬品開発の高コスト化は、製薬会社に対する直接の経済的負担になっているばかりでなく、超少子高齢化社会を迎える日本の医療費に対しても間接的影響を与えている。このことから、医薬品開発の効率化をはかり、開発コストを下げることは、製薬業界のみならず、日本の医療費問題にとっても重要な課題となっている。このような背景から、近年、医薬品開発を効率化する有力なアプローチとして、計算機シミュレーションによる創薬に大きな期待が寄せられている。しかしながら、現状の創薬計算技術にはいまだ残された課題が多く、新薬創出を強力に加速する技術革新が待ち望まれてきた。筆者らは、製薬企業、IT企業とともにコンソーシアムを設立し、スパコン「京」を用いた創薬応用計算に取り組んできた。この約1年半の取り組みの中で見えてきた、モンスターマシン「京」の驚異的パフォーマンスとそれによる創薬計算の展望を紹介する。

KBDD コンソーシアムの目標

創薬計算技術は、医薬品開発プロセスの「化合物ライブラリから疾患原因タンパク質に結合し制御する化合物を探索するスクリーニング段階」と、「発見した候補化合物の活性向上と副作用回避を目指して化学構造を変換する最適化段階」において広く用

いられてきた。たとえば、計算機上でタンパク質と化合物の3次元モデルが結合するかどうかを評価するドッキング計算プログラムを用いて多数の化合物の結合活性を仮想的に評価するバーチャルスクリーニングや、量子化学的に見積もられたエネルギー関数を用いて結合親和性を推定することで活性化化合物を最適化する試みなどがなされてきた。しかしながら、現状の創薬計算技術は計算速度や予測精度の面において満足できる性能ではなく、次に挙げる2点が永年の課題として残されていた。

【課題1】医薬品候補化合物の探索では、膨大な化合物候補（10の60乗以上の化合物数）と多数の創薬標的タンパク質候補との莫大な組合せ数の相互作用評価（生理活性の予測）を行うことが理論上必要であるが、創薬現場では計算速度の問題から数百万程度の組合せの相互作用評価しか行われていない。

【課題2】医薬品候補化合物の探索と最適化での現状の創薬計算技術の予測精度は、平均5%程度であり、タンパク質と化合物との結合親和性を正確かつロバストに予測できる計算技術が創薬現場では確立されていない。

これら2つの課題の克服には、いずれも大規模計算が必須であり、計算機パワーの問題が現場利用の大きな障壁となっていた。そこで筆者らは、製薬企業19社、IT企業2社、アカデミア5機関からなる産官学コンソーシアム（名称：K supercomputer based drug discovery (KBDD) コンソーシアム）を設立し、「京」の圧倒的なマシンパワーを用いることで、この2つの根本的課題の克服に挑戦してきた（図-1）。特に、これらの課題を克服するための計算が実現できたとしても、結果を得るために日常

業務を逸脱した複雑な操作を必要としたり、数カ月もの計算時間を要したりするようでは、産業利用では何の役にも立たない。このような観点から、創薬現場で「京」を実践的に利用できる計算フローの構築も我々の目標とした。

具体的には、「課題1」に対しては、独自に開発する高速な化合物探索計算法「Chemical Genomics-based Virtual Screening 法 (CGBVS 法)」¹⁾を「京」に実装し、化合物とタンパク質の大規模相互作用空間の超高速探索を試みた。CGBVS 法は大規模相互作用データを機械学習することでバーチャルスクリーニングを可能にしたものであり、「ビッグデータ創薬」の先駆的取り組みとして位置付けられる。また、「課題2」に対しては、化合物とタンパク質との結合自由エネルギーの高精度な推定が期待されてきた「Massively Parallel Computation of Absolute binding Free Energy (MP-CAFEE 法)」²⁾を「京」に実装・チューニングし、化合物とタンパク質との結合自由エネルギーの予測を行い、その性能評価を行った。MP-CAFEE 法はアンサンブル型の分子動力学 (MD) 計算を通じて、タンパク質と化合物の結合自由エネルギーを算出するものであり、「シミュレーション創薬」の本格的取り組みと言える。

なお、「ビッグデータ創薬」や「シミュレーション創薬」といった呼び名は、まだ広く浸透した呼び方ではないが、近年注目される科学的手法である「データ科学」と「シミュレーション科学」を用いた創薬という意味を込めて命名したものである。すなわち、ビッグデータ創薬とは、日々蓄積されつづける膨大な生命科学データから創薬に有用な新知見を発見するデータ科学であり、シミュレーション創薬とは、実際の実験を行う代わりに計算機上で薬物作用の模擬実験を行うシミュレーション科学と定義できる。以下では、これらビッグデータ創薬とシミュレーション創薬の我々のコンソーシアムでの成果を紹介する。



申請主体 (事務局): NPO法人パイオグリッドセンター関西
研究代表 京都大学医学研究科 奥野恭史

製薬関連企業 (19社):

アスビオファーマ、杏林製薬、エーザイ、小野薬品工業、科研製薬、キッセイ薬品工業、参天製薬、塩野義製薬、大日本住友製薬、田辺三菱製薬、日本新薬、協和発酵キリン、持田製薬、千寿製薬、大正製薬、帝人ファーマ、東レ、日産化学工業、マルホ

IT企業 (2社): 京都コンステラ・テクノロジーズ、三井情報

大学等: 京都大学医学研究科、産業技術総合研究所、理研AICS、先端医療振興財団、都市活力研究所

- ・ コンピュータ創薬の根本課題に挑戦
- ・ 次世代の計算創薬「ビッグデータ創薬」と「シミュレーション創薬」の開拓
- ・ 製薬会社による現場利用に耐え得る計算フロー (計算精度と計算時間) の構築
- ・ 我が国のコンピュータ創薬の中心拠点形成

図-1 KBDD コンソーシアム (K supercomputer based drug discovery コンソーシアム)

スパコン「京」によるビッグデータ創薬

スパコン「京」は2011年6月および11月にLINPACK ベンチマークのTOP500において世界1位を見事獲得した。世界一と言われてもなかなかピンとこないものであるが、スパコン「京」の計算速度は1秒間に1京 (10の16乗) 回の計算 (足し算や掛け算) ができる速度10ペタフロップス、たとえば世界の全人口70億人が1秒間に1回計算しても17日かかる計算量を「京」は1秒間で終了するスピードをほこる。

では、「京」で化合物の理論上の総数10の60乗個 (実際はこれ以上のバリエーションがある) を計算することを考えよう。仮に1つの化合物の活性を計算で評価するのに1秒間かかるとしたら、「京」は1秒間に10の16乗個の化合物を計算できることになるが、10の60乗個の化合物を計算するには「京」をもってしても実に10の36乗年を要するものと推計される。10の36乗年後にそもそも地球が存在しているかも不明であるが、化合物のバリエーション10の60乗以上というのは我々の想像をはるかに超えた規模であることが分かるであろう。つまり創薬計算はビッグデータ解析そのものであり、「京」の

スーパーコンピュータ「京」の利用

ような超高速マシンを開発するだけではなく、膨大なデータを高速処理できるビッグデータ解析技術の開発が必須である。

このことから、KBDD コンソーシアムではビッグデータ創薬技術として、並列化による高速化が期待される化合物探索計算法 CGBVS 法を「京」に実装することにした。CGBVS 法は、既知のタンパク質と化合物の相互作用情報を機械学習して相互作用の有・無の 2 クラス判別器を構築することにより、予測対象のタンパク質（アミノ酸一次配列）と化合物（化学構造）の組合せに対して相互作用の有・無を判定する計算法である（図-2）。CGBVS 法の学習機械には、情報科学分野のさまざまな予測問題においてすぐれた応用実績を持つサポートベクタマシン（SVM）が用いられており、いったん、予測モデルを構築した後は、単純な行列演算を繰り返すだけで予測対象のエントリー数をスケールアップできる数理的枠組みを有している。そこで我々は、8 万

個以上のノード（プロセッサ）を並列につないだ大規模並列システムである「京」に CGBVS 法を実装し、膨大な化合物とタンパク質の組合せの相互作用予測計算を大規模並列化することで超高速化を試みた。

具体的には、「京」に CGBVS 法を実装し、プログラムのスケーラビリティおよび計算速度の評価を行った（図-3）。評価は、タンパク質 233 種と化合物 100 万種のすべての組合せである 2 億 3,300 万ペアの相互作用の有無について、並列数を変えながら行った。その結果、9,600 ノード（76,800 コア）の大規模並列計算において、計算速度の高速化傾向が観測され、CGBVS 法は並列化効率に非常に優れたプログラムであることが確認できた。また、「京」の 9,600 ノード並列で 2 億 3,300 万ペアを計算するのに 1,921 秒かった。このことから、「京」の全ノード 88,128 個をフルに用いて CGBVS 計算を

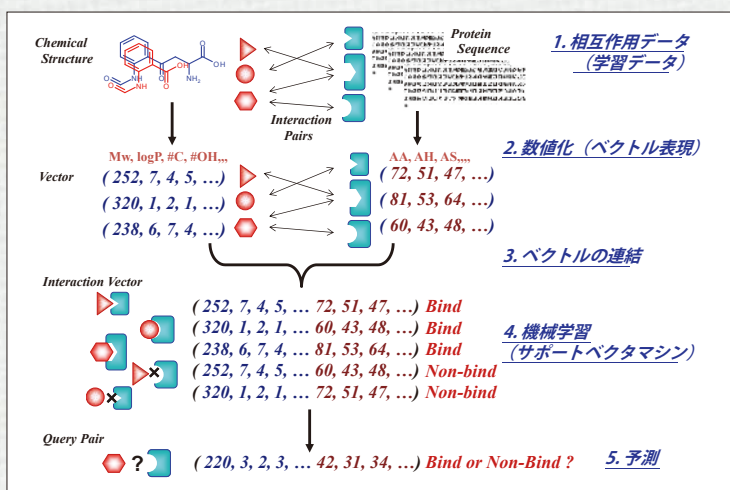


図-2 CGBVS アルゴリズムの概要

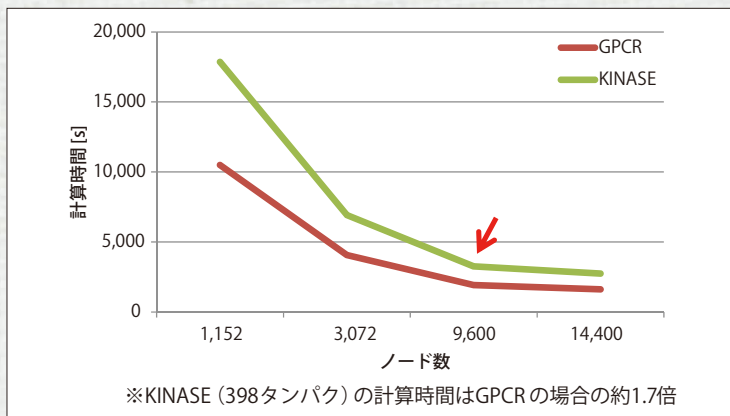


図-3 「京」における CGBVS のスケーラビリティ

行った場合、1 秒あたり 100 万個（10 の 6 乗個）、1 日あたり 864 億個（8.64 × 10 の 10 乗）の化合物ータンパク質ペアを計算できることになる。

続いて、参画製薬企業の希望標的タンパク質をエントリーに含む GPCR（G protein-coupled receptor, G タンパク質共役受容体）233 種とキナーゼ 398 種の計 631 種のタンパク質を標的として、米国 NCBI（National Center for Biotechnology Information）が公開する PubChem データベース³⁾の化合物 3,000 万種とのすべての組合せ 189.3 億ペアの CGBVS 法による相互作用予測を行った。189.3 億ペアのタンパク質と化合物の相互作用ペアは、筆者の知る限り世界最大規模の予測であり、たとえば、筆者の研究室で保有する計算機 16 ノード（128 コア）を用いて CGBVS 計算を行うのに約 2 年（762 日）を必要とし、創薬現場のタイムスケールとしては計算することが非現実な規模である。これに対

し、「京」の8万ノードを一気に利用できたとする、たった5時間45分で計算が完了するものと見積もられ、「京」の計算機パワーは現場利用における計算規模を桁違いにスケールアップする効果を有していることが分かる。なお、189.3億ペアの相互作用予測結果のデータはすでに参画する製薬会社に提供され、参画企業11社（2013年度参加企業）のうち、実際に実験などをして医薬品開発をスタート

させている企業は7社あり、3社は計算結果を用いた医薬品開発の準備中である。

また、CGBVS法によるタンパク質と化合物との相互作用の予測結果と実験値との比較を行ったところ、正確度（Accuracy）約79%と非常に高い正答率であることも示された⁴⁾。図-4は、相互作用（キナーゼ388種と化合物500種）の予測結果と実験結果との比較を視覚的に示した一例である。左右のヒートマップのまだら模様が類似していることから、CGBVS法による予測結果は、実験で得られる相互作用パターンをうまく再現できていることが分かる。

スパコン「京」によるシミュレーション創薬

コンピュータを使って薬を分子設計することは、創薬分野では今や広く認知された当たり前の技術である。実際、コンピュータ上にタンパク質と化合物の立体構造モデルを再現し、結合状態を自動でモデリングするドッキング計算のきれいな映像を見た読者も少なくないであろう。しかしながら、この計算結果の予測精度は平均で5%程度であり、お世辞にも高いと言えるものではない。ドッキング計算法の予測精度が低迷する原因は種々考えられるが、主要因の1つに物理的な計算機パワー不足が挙げられる。すなわち十分な計算機パワーがない中で現実的な計算時間で結果を得るためには、近似的な計算を

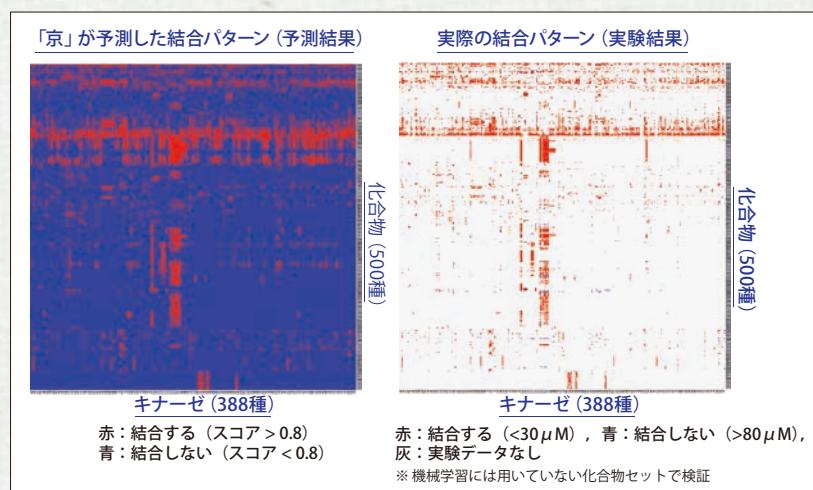


図-4 「京」によるCGBVSの予測結果と実際の結合パターンの比較

行わなければならない、そのために予測精度が損なわれているのである。たとえば、通常のドッキング計算では、計算負荷を軽減するためタンパク質の動きはほとんど固定した状態（化合物の構造変化のみを許容）で計算され、さらに水分子の動きも計算に考慮せず溶媒効果を反映できない現状にある。スパコン「京」の登場は、これらの計算機パワーに起因する問題を解消し、タンパク質の動きや溶媒効果を十分に考慮した精密な結合シミュレーションが実時間でできる可能性を切り拓いたものと言える。

そこでKBDDコンソーシアムでは、分子動力学シミュレーションにより結合自由エネルギーを高精度に予測できることで定評があるMP-CAFEE法を独自に「京」に実装し、その性能評価とプログラム整備を行ってきた。MP-CAFEE法は富士通研究所の藤谷ら（現東京大学）が考案した結合自由エネルギーを算出する方法で、自由エネルギー摂動法を非平衡統計理論に基づいて拡張した方法である。分子動力学シミュレーションに基づく結合自由エネルギーの計算では、タンパク質・化合物の結合状態と、非結合状態との間で、自由エネルギーの変化を計算する。これら二状態間での自由エネルギー差が、タンパク質と化合物が結合する際の結合自由エネルギー、すなわち結合強度に相当する。MP-CAFEE法では、タンパク質と化合物との相互作用ポテンシャルに強制的に荷重係数をかけることで、複合体状態から単独分子状態の遷移過程を複数のこま切りの

スーパーコンピュータ「京」の利用

状態遷移としてモデル化（32 個の状態遷移のモデル）し、各状態遷移でのアンサンブルから、トータルの自由エネルギーを算出する方法を採用している（図-5）。また、MP-CAFEE 法では、この 32 個のこま切れの状態遷移過程の計算（非平衡計算）に先だって、十分に平衡化した定常状態の複合体構造を得る必要があり、我々は比較的長時間（50 ナノ秒）の平衡化計算を 5 つの初速度で実行し、適切な初期構造を選択している。続いて、選択した平衡化構造を初期構造として、6 つの初速度から非平衡計算（32 個の状態遷移）を 2 ナノ秒実行し、最終的に 192 個（ $= 6 \times 32$ ）のアンサンブルをとるという方法を採用している。このような大規模アンサンブルシミュレーションにより、初期状態への依存を排除し、ロバストな結合自由エネルギー計算を可能としているが、その一方で大規模アンサンブルシミュレーションであるがゆえに、計算量が膨大となり、通常の計算機環境で MP-CAFEE 法を実行することは現実的ではなかった。そのため MP-CAFEE 法の実行とその計算報告はこれまでに藤谷らによるものしかなかったことから、同法の真の性能や汎用性において第三者による検証が求められていた。

そこで我々は、3 種類のキナーゼタンパク質と骨格の異なる化合物 30 個（CHK1:6 化合物, CDK2:12 化合物, ERK2:12 化合物）について、MP-CAFEE 法による結合自由エネルギー値の算出を行い、実験値との比較や必要な計算時間の算定を行った。

そこで我々は、3 種類のキナーゼタンパク質と骨格の異なる化合物 30 個（CHK1:6 化合物, CDK2:12 化合物, ERK2:12 化合物）について、MP-CAFEE 法による結合自由エネルギー値の算出を行い、実験値との比較や必要な計算時間の算定を行った。

図-6 は CHK1 タンパク質と 6 種類の化合物との「京」による予測値（結合自由エネルギー値）と実験値（ K_i 値）のそれぞれを横軸と縦軸にとりプロットしたものであり、予測値と実験値の相対値が高い相関を示していることが分かる。また、別のタンパク質 CDK2 や ERK2 においても、MP-CAFEE 法によって予測された結合自由エネルギー

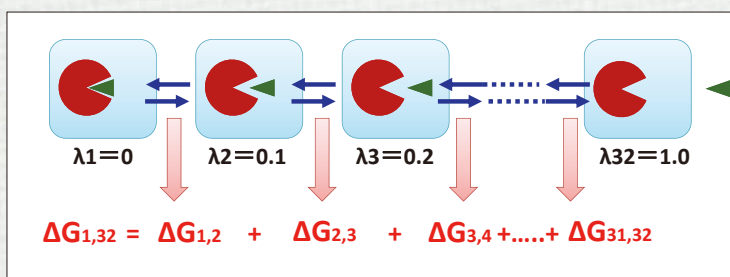


図-5 MP-CAFEE 法におけるタンパク質-化合物複合体状態とタンパク質単独状態との間の自由エネルギー算出方法

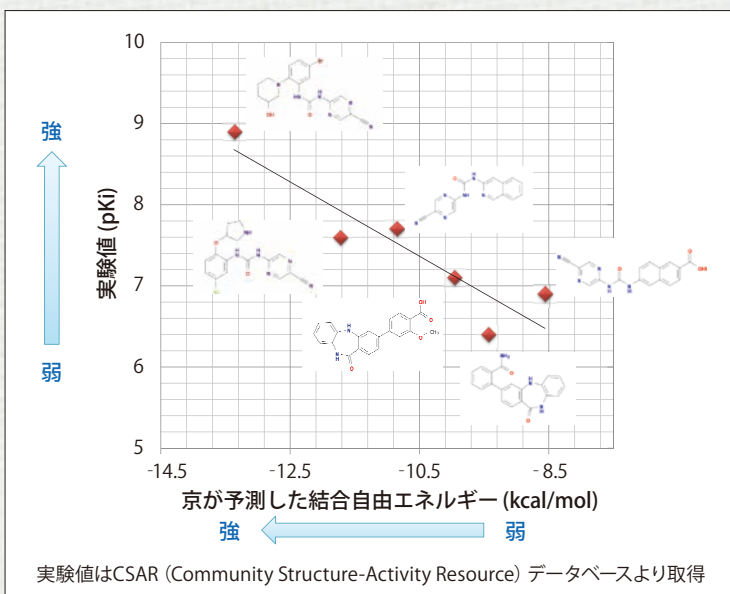


図-6 MP-CAFEE 法による結合自由エネルギー予測結果と実験値⁵⁾との比較 (CHK1)

値と実験値との間で良好な相関が得られたことから、MP-CAFEE 法はタンパク質や化合物の種類によらず安定に結合親和性の相対強度を予測する性能を有するものと考えられる。

また、MP-CAFEE 法の最大の弱点である計算時間については、まず、「京」における並列計算のスケラビリティについて調査した。MP-CAFEE 法の計算は、前述のとおり、非平衡過程の初期構造を選択するための平衡化計算（5 通り）と、結合状態から非結合状態への遷移過程を計算する非平衡計算（32 の遷移過程 $\times$ 6 通りの初期ベクトル = 192 通り）とに分かれる。実際のシミュレーション計算では、ドメイン分割による並列化を行うため、系の規模に応じて分割数に限界があり、すなわち一単位の計算についての並列数に限度がある。たとえば、CHK1 の平衡化計算では 104 並列（「京」13 ノード）で、また結合状態から非結合状態の遷移過程における非

平衡過程の分子動力学計算では24並列（「京」3ノード）で、一単位の計算は性能向上が頭打ちになる。しかしながら、MP-CAFEE法は、“Massively Parallel”という名の通り、遷移状態や初速度ベクトルを変えながら多数の分子動力学シミュレーションを実行するのが特徴の1つであり、これらの計算は互いに独立であるため、容易に並列化できる。KBDDコンソーシアムでも、平衡化計算の5通り、および非平衡過程の192通りの計算は、それぞれ計算性能へのペナルティなしに並列化できることを確認した。計算時間については、前述のCHK1, CDK2, ERK2と化合物の合計15ペアの結合自由エネルギーをMP-CAFEE法によって計算するのに、筆者らの研究室の計算機環境では約20年かかるものと推計されたが、「京」の10%程度のノードを用いることで、1週間程度で結果の取得が可能であることが示された。すなわち、「京」の約8万ノードすべてを持っていれば、同時に約150セットのタンパク質・化合物間の結合強度計算が可能である。

前述のとおり、KBDDコンソーシアムではこれまで、CHK1, CDK2, ERK2といったキナーゼについて、MP-CAFEE法による結合強度の予測精度を検証してきた。現在は、他の種類のタンパク質・化合物セットとして、プロテアーゼ（uPA: 7化合物）やGPCR（AA2AR: 9化合物, ADRB2: 3化合物）についても検証を重ねている段階である。また、種々のタンパク質・化合物セットを用いた予測精度の検証を通じて、プレ処理である平衡化計算によって得られる膨大な数の構造データ（2,500万構造×5通り＝1億2,500万構造）から、いかに結合状態から非結合状態の遷移過程のモデリングのための初期構造を選択するかが、結合強度の予測精度にとって重要であることが分かってきた。KBDDコンソーシアムでは、現在、この非平衡過程の初期構造（平衡化状態の複合体構造）を適切に選択するための基準についての詳細な検討を行っている。複数の条件設定下でMP-CAFEE法の計算過程を実行し、そのすべての結果を現実的な時間で得ることが可能な「京」の圧倒的な計算パフォーマンスがあってこそ、このような検討が可能となる。

ポスト「京」への期待

上述の通り、標的タンパク質に対するリード化合物の探索において189.3億ペアを対象とする世界最大規模のバーチャルスクリーニングや、精密な分子動力学シミュレーションによる複数のタンパク質・化合物セットの結合自由エネルギー予測を現実的な時間で実現するなど、「京」のパフォーマンスは、これまで年単位の計算時間を要するために諦めていた計算テーマの実時間実行を可能にする驚異的なものである。一般に現在の汎用計算機は15年前のスーパーコンピュータの速度であると言われていることから、今、我々は「京」を通じて15年先の創薬計算分野の世界を垣間見ていることになる。しかしながら、150個のタンパク質と化合物との結合自由エネルギーを精密に計算するのに「京」の全ノードを用いても1週間かかることから分かるように、創薬計算分野にとってはまだまだ高速なマシンが必要であり、ポスト「京」、次世代スーパーコンピュータの開発において期待するものである。最後に、当該プロジェクトの遂行にあたり、KBDDコンソーシアムの参画メンバ、理化学研究所計算科学研究機構、高度情報科学技術研究機構、最先端次世代研究開発支援プログラムに深く感謝申し上げる。

参考文献

- 1) Yabuuchi, H., et al.: *Mol. Syst. Biol.*, 7, 472 (2011).
- 2) Fujitani, H., et al.: *Phys. Rev. E*, 79, 021914 (2009).
- 3) PubChem データベースの Web サイト, <http://pubchem.ncbi.nlm.nih.gov/>
- 4) Brown J. B., et al.: Compound-protein Interaction Prediction within Chemogenomics: Theoretical Concepts, Practical Usage, and Future Directions, *Mol. Inform.*, 32(11-12), 906 (2013).
- 5) CSAR データベースの Web サイト, <http://www.csardock.org/>

(2014年3月3日受付)

中津井雅彦 nakatsui@port.kobe-u.ac.jp

1980年生まれ。2008年博士（システム生命科学）九州大学大学院システム生命科学府。産業技術総合研究所情報工学研究センター、京都大学大学院薬学研究科、高機能遺伝子デザイン技術研究組合を経て、2013年より現職。

奥野恭史 okuno@pharm.kyoto-u.ac.jp

1970年生まれ。1993年京都大学薬学部卒業、同大学院薬学研究科進学、2000年博士（薬学）取得。同大化学研究所バイオインフォマティクスセンター特任助手、同大薬学研究科特任助手および准教授を経て、2008年同大薬学研究科システム創薬科学寄附講座教授。2014年4月より現職。