IBM

Blue Gene/L によるタンパク質モデリング

スーパーコンピューターでのモデリングとデータ可視化による現実世界での科学の進歩

Blue Gene®/L スーパーコンピューターは、最先端を行く計算処理能力と複雑なデータの可視化ツールを科学者たちに提供します。科学者たちは、それぞれの専門分野の第一線で活躍するためにこの技術を必要としています。この記事では、分子生物学者がタンパク質の複雑な分子構造についてよりよく理解する手段として、コンピューターでタンパク質のフォールディングおよびミスフォールディングのシミュレーションを行う上で、どのようにこの技術を利用しているかを学びます。

Chris Ward は、ケンブリッジ大学でエンジニアリングの学位を取得した後、1982年に英国ハースレイにある IBM UK Development Laboratories に入社しました。これまでディスク・ファイルから商標付きミドルウェアに至るまで、数多くの IBM 製品の開発に取り組んできました。IBM WebSphere Software および IBM Lotus Software が現在の IBM の顧客にとって貴重であるように、彼は IBM の未来の顧客に役立つ技術に携わる幸運に恵まれています。

Ruhong Zhou は、Computational Biology Center/IBM Thomas J. Watson Research Center の Research Staff Scientist であると同時に、ケンブリッジ大学化学学部の非常勤教授でもあります。彼は1997年、コロンビア大学で Bruce Berne とともに博士号を取得しました。2 年半の間、Richard Friesner (コロンビア大学) と William Jorgensen (イエール大学) で分極可能な力場とタンパク質リガンドの結合メカニズムを研究した後、2000年に IBM Research 入社しました。彼が書いた出版物は 80 にのぼり、特許も 7 つ取得しています。また、主要なコンファレンスや大学で数々の招待講演を行った他、コンピューターを利用した生物学、化学、生物物理学における複数のコンファレンスで議長を務めた経験もあります。1997年にコロンビア大学の Hammett Award、1995年に American Chemical Society on Computational Chemistry の DEC Award、そして 2005年と 2008年に IBM の Outstanding Technical Achievement Award を受賞しました。現在、彼が興味を持っている研究には、コンピューターを利用した生物学とバイオインフォマティクスのための革新的手法とアルゴリズムの開発、そしてタンパク質フォールディング、リガンドと受容体の結合、タンパク質構造予測のための大規模なシミュレーションがあります。

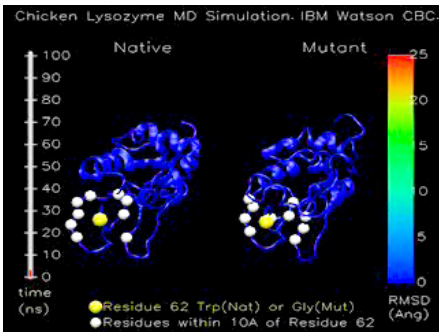
2009年 6月 09日

2001年に IBM の科学研究員たちは新しいサーバー・ファミリーの設計に着手しました。これらのサーバーは、現在、IBM System Blue Gene® という名前で販売されていますが、初めて利用できるようになったのは 2004年のことです。最初に登場したのは Blue Gene/L (この記事の本題です)、次に Blue Gene®/P と続きました。

Blue Gene スーパーコンピューター・ファミリーは、標準的なプログラミング環境でウルトラ・スケールの性能を発揮するように設計されていると同時に、消費電力、冷却性能、フロア·スペースに関しても優れた効率性を実現するようになっています。Blue Gene は数多くの大学、政府、そして民間の研究所で利用され、電波天文学、タンパク質フォールディング、気候研究、宇宙論、そして薬剤開発の分野でコンピューターによる研究が進められています。Blue Gene システムは科学の研究方法に桁違いの大きな変化をもたらしています。それは、このシステムが既存の複雑なモデルに代わるモデルを設計および実行する上で、よりコスト効率に優れたツールになるからです。

この記事では、タンパク質モデリング関連のプロジェクトの 1 つによって達成された進歩を紹介します。図 1 に、Blue Gene/L の威力によって現在どの程度の計算処理が可能になっているかを示します。この初期構造は、リゾチームの結晶構造から始まっています (ソースについては、「参考文献」を参照)。

図 **1.** 生細胞内部での **10** マイクロ秒にわたるリゾチームの変化の一部 **(**動画をご覧ください**)**



プロテオミクス: タンパク質の活動

生体高分子であるタンパク質は、生命体には不可欠の成分であり、細胞内でのあらゆるプロセスに関与します。タンパク質の多くは生化学反応を触媒する酵素です。なかには細胞シグナリングと免疫反応に関与するタンパク質もありますが、大部分のタンパク質は筋肉と細胞骨格に対する構造的および機械的機能を持ちます。以下の 2 つの例から、タンパク質がいかに広範囲にわたり、いかに重要であるかがわかります。

あるタンパク質は血液の「赤み」に関与します。このタンパク質が、肺から体のその他すべての部分に酸素を運びます。

ツタウルシに含まれる毒に対する人体の反応に関与するタンパク質もあります。この毒は、刺激性は極めて高いながらも、通常は害がありません。

地球上の生物に関わるタンパク質には数十万もの種類があります。プロテオミクスとは、タンパク質がどのように機能し、相互作用するか、そしてタンパク質の多様性と特殊性が私たちを取り巻く生命体のなかでどのように展開しているかについての研究のことです。この記事ではタンパク質とは何か、タンパク質はどのようにして形成されるのか、そしてタンパク質が存在している系にどのようにタンパク質は影響を及ぼしているのか、について簡単に説明します。

DNA は、あらゆる動植物のすべての細胞に含まれる、情報を格納する構成要素です。DNA には、**A**、**C**、**T**、**G** (DNA に含まれるアデニン、シトシン、チミン、グアニンのこと。RNA ではチミンはウラシルに置き換えられます) と呼ばれる化学成分 (ヌクレオチド) の配列として情報が格納されます。これらの成分は遠目から見ると非常に似通っているため、DNA のどの部分を見ても、全体的な形状はすべて同じように見えます。この形状とは、お馴染みのワトソン·クリック·モデルの二重らせん構造です。

DNA の情報を読み出す際には、DNA のらせん構造が解かれ、DNA の内部パターンの表現によって RNA と呼ばれる別の分子が形成されます。セメント状の漆喰に鍵を押し付けて鍵のイメージを漆喰に刻み込むようにするわけです。次にこの RNA 分子は、リボソーム (多目的工場のように振る舞うタンパク質) に青写真として提示されます。するとリボソームが A/C/T/G コードを 3 つ一組で読み取って、64 文字の「アルファベット」を引き出せるようにします。

これらの「文字」のうち、20 文字はタンパク質を構成するアミノ酸に相当するものです。アミノ酸は、主に私たちが摂取する食物から取り込まれます (ヒトに必要なアミノ酸のなかには体内で合成できないものもあります。この「必須」アミノ酸は、食物から摂取しなければなりません)。それぞれのアミノ酸には「頭 (head)」と「しっぽ (tail)」があります。リボソームは 20 文字のそれぞれに対応するアミノ酸を見つけて、アミノ酸を頭からしっぽの順に (head-to-tail 型で) 連結していきます (その他の「文字」は、連結の開始と終了のタイミングを指示します)。このようにアミノ酸が連結された結果、新たなタンパク

質の分子が作り出されます。その構造は、使用した DNA のセクションに刻み込まれたコードに厳密に従っています。

タンパク質の分子の原子間にある圧力と張力、細胞に含まれるわずかに塩分を含んだ水との相互作用、そして熱とも呼べる不規則振動によって、タンパク質の分子は特有の形状に「折りたたまれる」ことになります。

タンパク質の分子は極めて安定しており、なかには何百年もの間変化せず、それを造った元の生物には持ちこたえられない何百度もの温度にも耐えられるものもあります。これらのタンパク質の分子は、強力な化学物質や高圧、高温、低温によって変性するか、あるいは別の生命体の餌食となるまで、ほとんどその状態を変えません。

タンパク質が時間、温度、そして周囲の分子によって、どのような形状にどのように変化するか次第で、そのタンパク質の分子が持つ役割 (酸素を運ぶ役割であるか、ツタウルシに対するアレルギー反応を起こす役割であるか、あるいはその他の分子規模で発生し得ることを行う役割であるか) が決まります。

図 2 に、DNA のお馴染みの球棒モデルを示します (これは立体イメージです。イメージのソースについては、「参考文献」を参照してください)。

図 **2. DNA** の球棒モデル

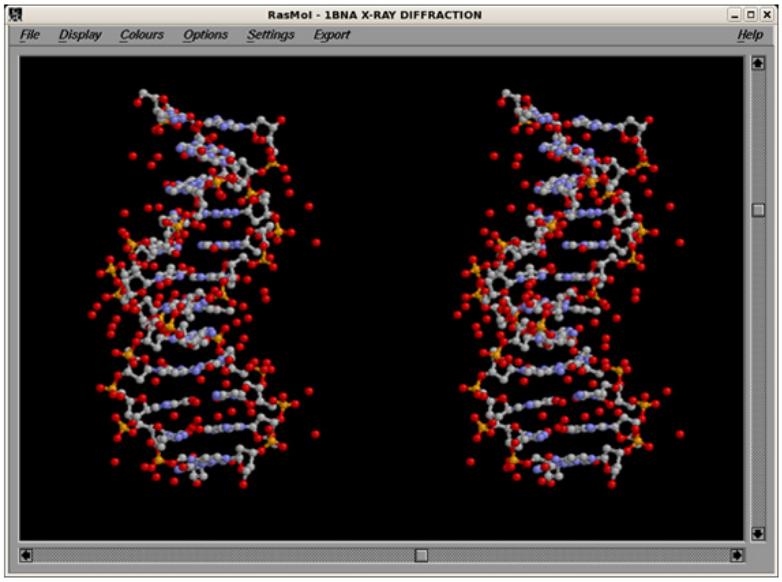
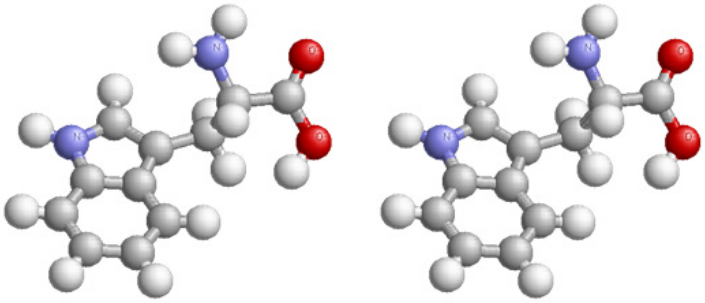


図 3 は、20 の標準アミノ酸のうちの 1 つ、トリプトファンです (これは立体イメージです。イメージのソースについては、「参考文献」を参照してください)。

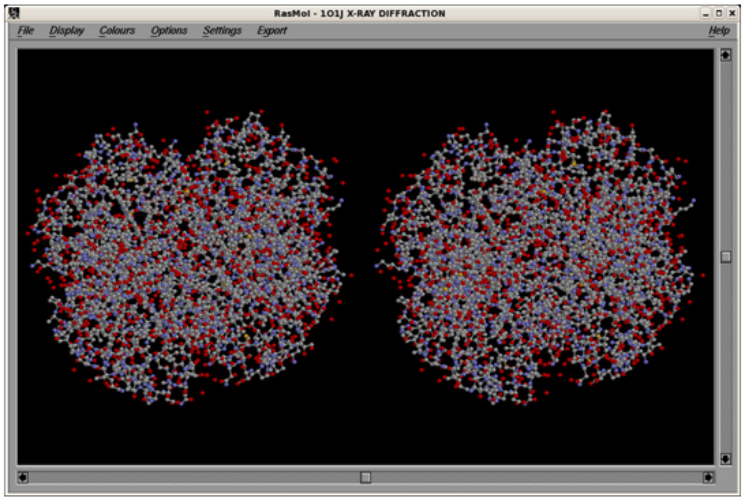
図 **3. 20** の標準アミノ酸のうちの **1** つ、トリプトファン



アミノ酸は、1 つの分子のヒドロキシ基 O-H (図 3 の右側) が分離され、別の分子の N (図 3 の上部) から H が切り離されてから、この 2 つの分子が結合されることによって、タンパク質に合成されます。分離されて残った H-O-H は水分子となります。あらゆるアミノ酸には、こうした特徴を持つ原子団があります (図 3 の右上)。

図 4 は、ヘモグロビン・タンパク質を視覚化した様子です (これは立体イメージです。イメージのソースについては、「参考文献」を参照してください)。

図 **4.** ヘモグロビン・タンパク質



ヘモグロビンは 4 つのサブユニットからなる合計 574 のアミノ酸分子です。鉄原子を使って合成されるヘモグロビン (鉄原子がどのようにタンパク質に合成されるかについては、この記事では説明しません) は、血流に酸素を送ります。酸素の運搬システムは鉄分子だけでも可能ですが、ヘモグロビンの構造が提供するタンパク質の「ケージ」によって遙かに効率的になります。このイメージを立体ビューアーで表示すると、3D による原子構造を抽出することができます。しかしこれよりも複雑になってきた場合、何が行われているかを可視化するには別の方法が必要になってきます。

商業的および学究的動機

タンパク質をはじめとする生物の基本的な構成要素について理解が深まるにつれ、次第に医薬品の設計と公衆衛生の保護に進歩がもたらされるようになってきました。そんななか、現在話題となっているのは、タンパク質のミスフォールディングとアグリゲーションで

野生型とは何か
野生型 (wildtype。「wild type」、「wild-type」と表記されることもあります) とは、生命体、遺伝子、系統、特性が自然に形作る典型的な型のことです。表現型

す。タンパク質が意図された形状ではない形状に折りたたまれることにより、異なる特性を持つ不活性タンパク質となる場合がよくありますが、このことが、アルツハイマー、クロイツフェルト・ヤコブ病、牛海綿状脳症 (狂牛病)、ハンチントン病、パーキンソン病、嚢胞性線維症やその他のアミロイド症、などの神経変性病の発症原因となる可能性があります。

(目に見える形で現れる生命体の特性。通常は遺伝子と環境要因の表現) について言う場合、野生型は自然個体群における最も一般的な形質を特徴付けます。また、遺伝子型 (目に見えない遺伝的構造) について言う場合は、遺伝子座のそれぞれで、野生型表現型にするために必要な対立遺伝子を定義します。野生型は優性または劣性のどちらでもありません。野生型の反義語としてふさわしいのは、突然変異型です。

これらの重大な疾患に対する治療の研究では、タンパク質の分子が有益な構造から別の折りたたみ構造に変異する原因を理解するための研究が盛んに行われています。ケンブリッジ大学の Chris Dobson とその共同研究者たちが行った最近の実験 (「参考文献」にリンクを記載) では、アミロイドと原線維は従来のベータ・アミロイド・ペプチドからだけでなく、適切な条件が揃えばほとんどすべてのタンパク質 (リゾチームなど) からも形成可能であることが証明されています。事実、リゾチーム・タンパク質での単一突然変異 (W62A) が、タンパク質を野生型 (囲み記事を参照) に比べて遙かに不安定な状態にすることがあります。さらにこれは、「長期疎水性相互作用」という鍵が失われることにより、タンパク質のミスフォールディングが発生して尿素溶液にアミロイドが形成されることもあります。

科学者たちはまだ、この単一の W62 がフォールディング・プロセス中に長期疎水性相互作用に大きく作用し、推定では核形成部位から機能上の理由で表面化する仕組みを解明していません。この仕組みが解明されると、単一突然変異の作用、そしてタンパク質のミスフォールディングとアグリゲーションに関連する前述の疾患の背後にあるメカニズムについて、より深く理解するための貴重な機会がもたらされます。

Blue Gene/L 技術は、このような疾患の研究を行うための強力な手段となります。それはこの技術によって、よりコスト効果の高い (そしてより素早い) 方法でタンパク質のフォールディングとミスフォールディングの影響をモデル化できるからです

モデリングの対象について

図 1 は、単一突然変異によるリゾチーム・タンパク質のミスフォールディング・シーケンスの一部を可視化した動画からキャプチャーしたものです。リゾチームは人間の免疫系に含まれるタンパク質で、適切に機能していれば、体内に侵入してきた細菌の細胞壁に穴を開けて細菌を死滅させます。

単一突然変異 (DNA での異なるシーケンス) は、リボソームがリゾチーム分子を形成するときに異なるアミノ酸を使用する原因となります。理論としては、この異なるアミノ酸によって、リゾチームが折りたたまれる形状に影響が及ぼされ、その結果リゾチーム分子が異なった形状になり、リゾチーム分子が細菌の細胞壁に穴を開ける効果にも違いが出てきます。この変化を理解することにより、突然変異による細菌性疾患からの回復を支援する医薬品、あるいはその他の形での治療を設計できる可能性があります。

研究の一環として、私たちは 1 つのリゾチーム分子に含まれるすべての原子の位置と速度に加え、約 10,000 の水分子と尿素分子 (このシミュレーションは、実験を模倣して 8 モルの尿素溶液で行われています) についても原子の位置と速度をコンピューターのメモリーに保管しています。原子間の力をモデリングする方法は多数ありますが、私たちが使用しているのは結合力を対象とした球とばねによるモデルを、(電荷を帯びた原子間の静電力を対象とした) 逆二乗法則モデルと、(至近距離にありながらも共有結

合していない原子を対象とした) 求引/反発モデルとを使ってアレンジした方法です。このモデルは時系列で実行されます。時間ステップごとに、各原子での力を計算した後、ニュートンの第二法則に従って速度と位置を更新します。

各時間ステップ (約 1000 兆分の 1 秒という極めて小さい単位) では、計算対象となる力が原則として何億もあります。この莫大な計算量、そしてシミュレーションは対象とする運動をモデリングできるだけの長い時間 (数マイクロ秒) 実行できなければならないことから、作成方法がわかっている最大のコンピューターをもってしてでも、この手法が現実的になったのはつい最近のことです。私たちが行っている方法とこれに代わる手法についての詳細は、「参考文献」に記載されている「Destruction of long-range interactions by a single mutation in lysozyme」のリンクを参照してください。

ラボの装備

ニューヨーク州ヨークタウンにある IBM Watson Research Lab には、BlueGene/L サーバーのラックが 20 台あります。それぞれのラック内には 1,024 の PowerPC® デュアル・コア・マイクロプロセッサー・チップが収容されており、各マイクロプロセッサーには 512MB の RAM が接続されています。さらに、ラック内の 64 のチップごとにもう 1 台、1Gbps の Ethernet リンクに接続されたマイクロプロセッサーが用意されています。これらの 320 の Ethernet リンクは、標準 Ethernet スイッチ経由でディスク、テープ、言語コンパイラー、そしてジョブ管理ソフトウェアを備えた標準 IBM Power Systems マシンに接続されています。

このリゾチームのモデリング作業では数ヶ月にわたり、BlueGene/L プロセッサーのラックを平均 4 台使用して 10 マイクロ秒を超える分子動力学データを集めました。このアプリケーションは、シミュレーション対象のすべての原子の位置と速度を定期的に書き出します (上述の合成動画は、この一連の情報を部分的に使用して作成されました)。そのため、シミュレーションの実行を再開する必要がある場合には常に、位置と速度の適切なセットを再ロードすることができます。再開が必要となるのは、計画されたシャットダウンの後や、不測の障害がマシンに発生した後、あるいは科学的に興味深いモデル・イベントを時間ステップの粒度を変えて再生する場合などです。

モデルの実行

このアプリケーションは、MPICH ジョブ・サブミッションと同様のメカニズムによって Blue Gene/L ノードで起動されます (MPICH は無料で入手できる、MPI (Message Passing Interface) の移植可能な実装です。リンクについては「参考文献」を参照)。アプリケーションには、クラスター内の各プロセッサーによって POSIX ファイルシステム環境が提供されます。データは、アプリケーションが読み取れるように IBM GPFS (General Parallel File System) ファイルシステムにセットアップすることができます。アプリケーションが結果を書き込む際には、結果は同じく GPFS ファイルシステムに送られて外部で使用できるようになります。

このような時系列モデリング・アプリケーションの場合には通常、アプリケーションがファイルシステムから初期条件を読み取った後、モデル状態の定期的な「スナップショット」をファイルシステムに書き込みます。

このアプリケーションがもたらす成果

この動画は、これまで一度も見ることのできなかった世界を垣間見させてくれます。当然、これが真実を表しているかどうかはわかりません。科学者というものは常に、モデルが表すものを現実の世界で見

えるものと比較しなければならないからです。現実でのリゾチームのミスフォールディングを目にすることはまだ夢の話で、「固定的な」立体配座の一部を「見る」ことでさえも、サンプルを用意して、これらのサンプルを電子顕微鏡で詳細に調べるか、あるいは多数のリゾチーム分子を結晶化させてから X 線回折スペクトロスコピーを使用することになります。しかし、これらの実験的手法では、タンパク質の運動についての実態まではわからないのが通常です。

したがって、疾患関連のミスフォールディングに関与する分子の運動と重要な変化を詳細に調べるには、現在の大規模なシミュレーションが唯一の手段となります。うまくいけば、ミスフォールディングの調査を可能にするこの技術が、既成概念の枠を破り、アミロイド症の研究における最先端技術をさらに進歩させることになるでしょう。この技術はまた、次世代の科学者がこうした種類の問題を解決するための訓練にも使用することができます。その際、この新しい方法は、この種の研究を行う際の主要な手段として使用されます。

今後の予測

実際のところ、私たちは将来を予測しようと思うほど厚かましくはありません。しかし敢えて予測するならば、Blue Gene コンピューティングは引き続き発展の道を辿っていくと思います (私たちが使用しているのはバージョン L ですが、市販されている Blue Gene/P では 1 チップあたりのプロセッサー数が 4 基、Ethernet が 10Gbps にアップグレードされている他、多数の改善が加えられています)。さらに計算能力を駆使する演算を行うコスト、そして規模も速度も上回るストレージのコスト (どちらも、この記事で説明しているデータ可視化タスクに大きく関与するコスト) は、これからも下がっていくはずです。また、下がっていかなければなりません。なぜなら公的研究にとっても、企業が製品を市場に出すためにも、科学者が取り組まなければならない高度なモデリングには世界の何倍分もの価値があるためです。

ここで説明したリゾチーム・モデルは、コンピューターを利用した生物学という新しい分野について表面的にほんの少しかじっただけに過ぎません。一般公開されている Protein Data Bank (リンクについては「参考文献」を参照) に構造が記載されているタンパク質は 50,000 種類を超えています。分析すると薬学的に有益な可能性がある化合物は数百万とあり、タンパク質とその欠陥に関係することがわかっているヒトの疾患も数多くあります。さらに、この規模でのモデリングによるメリットがある他の数えきれないほどの研究分野については、この記事では触れてもいません。Blue Gene の仕事はまだ始まったばかりです。

参考文献

学ぶために

Blue Gene/L に関する研究の概要が、IBM Blue Gene プロジェクト・ページに記載されています。その他の Blue Gene ソリューションのコンポーネントおよびリソースには以下のものがあります。

- IBM Blue Gene/P ソリューション・ページ
- IBM General Parallel File System
- IBM XL C/C++ Advanced Edition for Blue Gene コンパイラーに関するすべての情報
- Blue Gene 技術に関する IBM Redbooks
- 著者の 1 人が使用した Blue Gene の写真

RCSB Protein Data Bank (PDB) は生体高分子の研究を目的としたアーカイブ

で、実験的に決定されたタンパク質、核酸、および複合体の構造に関する情報が保管されています。Educational Resources には、Molecule of the Month などの素晴らしいリソースもあります。

図 1 のソース・データは PDB の 1.33 A、structure of tetragonal hen egg white lysozyme ニワトリ卵白リゾチームの正方晶の構造です。

図 2 のソース・データは、PDB の B-DNA 12 量体、構造および動力学です。

図 3 は、ニューヨーク州立大学でホストされている MathMol ライブラリーの好意により掲載しています。

図 4 のソース・データは、PDB の脱酸素化ヘモグロビン (A-GLY-C:V1M,L29F,H58Q; B,D:V1M,L106W) です。

Chris Dobson 教授のグループ Web サイトに、詳細な分子生物学研究へのリンクが掲載されています。

「Destruction of long-range interactions by a single mutation in lysozyme」(R. Zhou、M. Eleftheriou、A. Royyuru、B. J. Berne 共著、Proc. Natl. Acad. Sci.、2007年) で、この記事で紹介したシミュレーションで使用されているモデリング手法を詳細に説明しています。

「Parallel implementation of the replica exchange molecular dynamics algorithm on Blue Gene/L」(M. Eleftheriou、A. Rayshubski、J. W. Pitera、B. G. Fitch、R. Zhou、R. S. Germain 共著、IEEE、2006年) では、この記事のシミュレーションで使用されている数学的手法のいくつかについて説明しています。

MPICH の次の段階である MPICH2 は、高性能で高範囲にわたって移植可能な (そして無料の) MPI (Message Passing Interface) の実装です。

Argonne Leadership Computing Facility では、計算科学コミュニティーに Blue Gene/P の処理時間を提供する提携プログラムを用意しています。

デモ用モデリング・アプリケーションは、世界各国の IBM Innovation Centers から入手できます。

「高性能 Linux クラスタリング」は、Linux によるハイパフォーマンス・コンピューティングの背景を説明している 2 回の連載です。第 1 回 (developerWorks、2005年9月) では、HPC の基本、使用可能なクラスターのタイプ、クラスター構成を選ぶ理由、そして HPC における Linux の役割について述べています。MPI を使用した並行プログラミングについて検討する第 2 回 (developerWorks、2005年10月) では、クラスターの管理とベンチマークを取り上げ、オープン・ソース・ソフトウェアを使用して Linux クラスターをセットアップする方法を紹介します。

「Port Fortran applications」(developerWorks、2009年4月) は、さまざまなハイパフォーマンス・コンピューティング・システム間で Fortran アプリケーションを移植する際の障害を克服する上で参考になります。

developerWorks Linux ゾーンに豊富に揃った Linux 開発者向けの資料を調べてください。記事とチュートリアルの人気ランキングも要チェックです。

developerWorks に掲載されているすべての「Linux のヒント」シリーズの記事と Linux チュートリアルを参照してください。

developerWorks の Technical events and webcasts で最新情報を入手してください。

製品や技術を入手するために

Open Discovery は Fedora Core をベースに、AFL (Academic Free license) によってライセンス交付されるバイオインフォマティクス・ソフトウェア・ツールの Live Linux ディストリビューションです。配列分析から分子力学に至るまでのタスクに対応できるこのツールは、DVD または USB ストレージ・キーからブートすることが可能で、データの永続化を特徴としています。インドのチェンマイにある SRM 大学 Ramapuram 校のバイオインフォマティクス学部に深く感謝します。

この記事で説明したアプリケーションに統合されているツールには、3D Fast Fourier Transform Library for Blue Gene/L、そして著者 Chris Ward の Custom Math Functions for High Performance Computing があります。

developerWorks から直接ダウンロードできる IBM ソフトウェアの試用版を使用して、Linux で次の開発プロジェクトを構築してください。

議論するために

My developerWorks コミュニティーに加わってください。自分個人のプロファイルとカスタム・ホーム・ページを作成して、自分の興味に合わせて developerWorks をカスタマイズしたり、他の developerWorks ユーザーと対話したりすることができます。