科学研究費助成事業　　研究成果報告書

平成　２６　年　　６　月　２７　日現在

機関番号：　３３３０２
研究種目：　基盤研究(C)
研究期間：　2011～2013
課題番号：　２３５００１３８

研究課題名（和文）データマイニングを用いた材料設計シミュレーション基盤システム

研究課題名（英文）Simulation-based material design system using data mining

研究代表者
林　亮子（Hayashi, Ryoko）

金沢工業大学・工学部・講師

研究者番号：３０３０３３３２

交付決定額（研究期間全体）：（直接経費）　　3,300,000円、（間接経費）　　990,000円

研究成果の概要（和文）：近年ではGPGPUやマルチコア計算機のように小規模～中規模の計算資源が大量かつ安価である．また，計算化学や計算物理のプログラムも成熟している．一方，計算結果の実体は大量の数値データであり，多数の計算結果を有効利用するためにはデータの自動処理が必要である．
本課題では基礎的な性能評価を行って取り扱い可能な系サイズを検討し，さらにデータマイニング技術を用いて結果データ中の分子構造の分類を試みた．その結果，炭素原子１０個程度までであれば大量のシミュレーション実行が可能であることがわかった．さらに，試験的に分子の構造の自動分類を行い，人間が分子の構造を分類するのと同様のルールを自動抽出できた．

研究成果の概要（英文）：Recently, GPGPU or multi-core computers are very familiar as the small-scale of middle-scale computing environment. Many package programs for computational chemistry and computational physics are also full-grown. On the other hand, a result of such package programs is the mass of numerical data so that we require auto-processing of the data for utilizing many computational results. This theme examined a fundamental performance evaluation in order to discuss suitable system size. Furthermore, this theme tried classifying molecular structure in a computational result via data mining technology. As the result, we found that we can run massively many computational jobs for less than ten carbon atoms. Moreover, auto-classification for molecular structure can find out as the same rule as the human do it.

研究分野：　情報学

科研費の分科・細目：　メディア情報学・データベース

キーワード：　データマイニング　ナノテクノロジー　分子　シミュレーション

１．研究開始当初の背景

近年の電子計算機の発達に伴い，材料設計シミュレーションが盛んになり，多くの材料設計プログラムが開発されてきている．しかし，材料設計プログラムを用いた材料設計では，トライアンドエラーのプロセスが必要であり，複数の初期条件を用いた複数のシミュレーション結果を統計処理することで初めて真の結果が得られる．これまでのシミュレーションを用いた材料設計プロセスでは，材料設計分野の研究者がすべてのプロセスに介在して行われてきたが，人間の処理能力には限界があるので，人力によるシミュレーション結果の処理が研究のボトルネック化している．一方で情報科学分野における情報システム技術は近年発達が目覚ましく，文字列処理や自動ファイル操作技術が発達し，さらにはデータマイニング技術が成熟してきているので，シミュレーション結果を計算機が自動判定する土壌ができている．

研究代表者は，これまで材料設計シミュレーションの高速化および並列化の研究を行い，材料設計分野の研究者とも共同研究を行っている．その過程で，シミュレーション結果の自動処理の必要性に気づき，近年は関連した研究を行っている．

材料設計シミュレーションにおける結果処理のボトルネック問題は知られつつあり，本課題の目標に近い，米国で実施されているNSF の Cyber-Enabled Discovery and Innovation (CDI)プログラムでは関連するテーマは５，６件ある．日本においても，ガラス物質や合金などを扱う何人かの研究者が自動データ処理を利用した材料設計に注目しているが，シミュレーション技術と自動データ処理技術を組み合わせて材料設計を行おうとする研究はまだ少ない．

２．研究の目的

本研究では，データマイニングを材料設計シミュレーションに組み合わせた材料マイニングシステムの構築を試みる．これは，シミュレーションを階層化し，精度は低いが短時間で終了するシミュレーションをクラスタ／グリッドやクラウド計算機で網羅的に実行し，その結果にデータマイニングを適用してさらに有望そうな計算条件を発見し，半自動的にあるいは自動的にシミュレーション実行する仕組みである．この考えの背景にあるのは，これまでの少数のシミュレーションをピンポイント的に行って人力で処理するという研究手法から，データマイニング技術を活用して多数のシミュレーションを網羅的に行う研究手法へのパラダイムシフトである．

３．研究の方法

本研究は主に計算化学を扱うが，現在の計算化学プログラムは非常に高機能であり，計算化学の複数の専門家が何年もかけて開発する．実際にそのようにして開発されてきた複数のプログラムが現在パッケージ化されている．著者らは計算化学プログラムの開発よりもその応用に力点を置くため，本研究では既存の計算化学パッケージを利用することとし，システム全体の構築に注力する．

本研究では最初のステップとして，計算化学プログラム Gaussian09 を使用する．Gaussian09 は最も有名な量子化学パッケージの一つであり，有償ではあるが研究機関が保有する計算サーバ上で利用可能であることが多く，PC 版は研究者個人の予算規模でも購入可能である．また解説書も日本語版を含め充実していて，比較的利用しやすい．

Gaussian09 では，計算を開始する初期条件を計算本体とは別の初期条件ファイルとして作成し，実行時に読み込ませる．そのため本研究では初期条件ファイルを用いる実行方法を前提とする．多くの計算化学や計算物理のパッケージプログラムでも同様の実行方法であるため，本研究の内容は，ファイル形式を合わせることで他の多くのプログラムにも応用可能である．

本研究が目標とする分子設計システムの概要を図１に示す．図１に示すように，ユーザは各種ツールやサンプルファイルを利用して，初期条件ファイルのひな形を作成する．次に，その設定に乱数を組み合わせてシステムは初期条件ファイルを複数個自動生成する．これはひな形中で文字列処理を行い，原子の位置座標などを置換することで実装できる．そして，生成した初期条件ファイルを用いて自動的にジョブを投入し，実行状況を管理する．これは，スクリプトプログラムなどを用いて技術的に十分可能である．

図１において，ジョブを実行するまでの手順には，技術的に大きな問題はない．一方でジョブ実行後に生成するのは大量の数値データであり，その処理には多くの課題がある．応用上は設計した分子が安定に存在する必要があるため，本研究では主に構造最適化を扱うが，構造最適化ジョブの結果得られた構造が初期条件で設定した構造と同一である保証はないため，ジョブの結果得られた構造を認識する必要がある．分子には，構成する原子の種類と個数が同じでも異なる構造を持つ異性体が存在するため，計算機を用いた分子構造の認識は大きな問題である．

本研究で主に問題となるのは，シミュレーションの結果データ処理である．Gaussian09 の出力データは途中経過を含む膨大な文字列と数値のデータであるため，必要な部分を切り出して処理する必要がある．また，得られる数値データはゆらぎを持った連続値であるため，単純なしきい値処理ではうまく扱うことができない．また適切に分類する手法も明確でない場合も多い．そのため，データ処理では近年発達の著しいデータマイニング技術を利用することが有効であると考えられる．

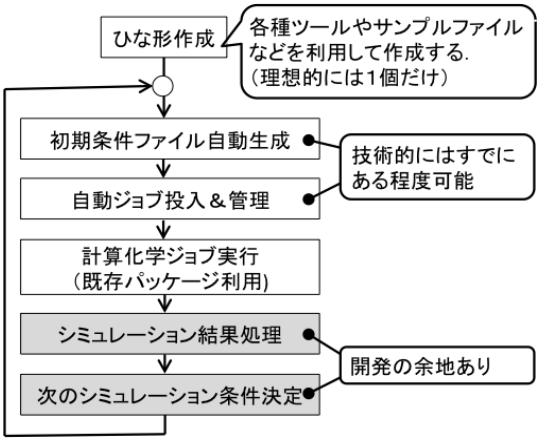

図 1. 分子設計システムの概要図

本研究では，データ処理に統計解析環境 R を利用する．R はオープンソースの統計処理用プログラミング言語であり，通常の統計処理機能が充実している．さらに，数多くのデータマイニング手法が既にパッケージ化されており，誰でも無料で利用することができる．さらに必要であれば独自のパッケージを開発することも可能であるため，本研究での利用に適している．

４．研究成果

これまでにプログラミング言語 Perl を用いて，初期条件ファイル自動生成の中核部分を開発した．これは，文字列処理を用いて出力ファイルから最適化後の位置座標を抽出し，乱数を利用して元の位置座標周辺に数値を変化させて，初期条件設定に用いるひな形ファイル中の座標指定部分で自動置換するプログラムである．本課題は文字列処理を活用する必要があるが，Perl は文字列処理に適したプログラミング言語であり，さらに Perl のプログラム中から Gaussian や R プログラムを実行することも可能であるため，本課題のシステム全体を Perl で実装する予定である．

本研究では個々のシミュレーションは数分程度以内を前提としている．その程度の実行時間で扱える分子サイズを確認する．「電子構造論による化学の探求」，Foresman and Frisch，田崎　健三訳，ガウシアン社，(1998).では Gaussian09 を使用する演習問題が掲載されており，第２章の上級演習 2.7 は直鎖炭化水素で炭素原子が２個～１０個の場合のシングルポイントエネルギー計算である．この演習用の入力ファイル中に含まれる原子の初期座標をそのまま利用して，計算内容を構造最適化に変更し，計算モデルを変更して実行時間を計測する．

実験条件を述べる．使用した計算機の諸元を表１に示す．　今回は日常の開発を想定して，通常の文書作成などにも使用する Mac OS のデスクトップ PC を用いている．計算モデルは，今回は計算方法と基底関数の組み合わせを４種類使用する．これらは Gaussian09 での指定方法を用いると次のように表される．

１．MP2/6-311+G(d,p) （Møller-Plasset 摂動法）
２．RHF/6-311+G(d,p)（Hartree-Fock 法）
３．RHF/6-31G(d)　（同上）
４．PM3　（半経験的方法）

これらは上記の文献中の演習問題で用いられている組み合わせであり，直鎖炭化水素は化学的に特異な物質ではないので，おおむね問題のない条件と考えられる．上記の１．が今回使用する中で最も計算モデルが複雑であり，一般に実行時間が長い．そして，列挙した順番に計算モデルが簡略化されていくので，おおむね実行時間も短くなる．

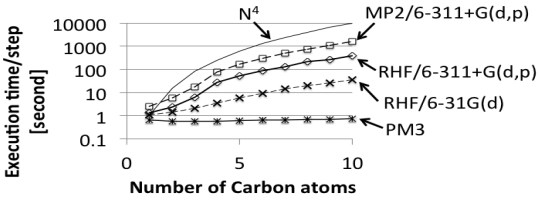

図 2. 直鎖炭化水素における炭素原子個数と１ステップの実行時間の関係

Gaussian09 における構造最適化では，４つの収束条件を満たすまでポテンシャルエネルギー面上での探索を繰り返す．今回用いたいずれの計算条件でも，炭素原子１０個以内の場合，繰り返し回数は最大５回であった．本稿では，Gaussian09 の出力ファイル中の CPU time を計測結果とし，CPU time を繰り返し回数で割った，１ステップあたりの実行時間を示す．

炭素原子１０個までの直鎖炭化水素を用いて，実行時間を調査した結果を図２に示す．図２はステップあたりの実行時間であるため，構造最適化に要する実行時間は図２中の時間にステップ数を乗算する必要がある．ステップ数は最大５であるため，利用可能なシミュレーションの実行時間の上限を 1000 秒とすると，１ステップあたり 200 秒になる．すなわち図２で 200 秒程度以下を満たす領域が現在扱える条件である．図２において 200 秒程度で計算できるのは，電子相関を考慮する MP2 の場合炭素原子５個，RHF/6-311+G(d,p)で炭素原子８個である．これらの手法の時間計算量は原子個数を N とすると $O(N^4)$程度と言われており，逐次処理でこれらの手法を粗放的に実行するのは難しい．図 2.には参考のため，N^4の曲線も記入した．

電子相関を含む方法のうち最も簡素なものに相当する RHF/6-31G(d)は炭素原子 10 個を 100 秒以下で計算可能であって，粗放的シミュレーションに利用できる．半経験的手法である PM3 は他の３つよりも緩やかに実行時間が増加するので，炭素原子１０個でも１ステップが１秒以下であり，より大きな分子の構造最適化を扱う粗放的シミュレーション

が可能であることがわかる．今回使用した計算機環境は通常業務に使用するものであり，現在購入可能な計算機は一般にこれより高速であるため，今後はより短時間でシミュレーションを実行可能であると考えられる．

すでに述べたように，本研究ではシミュレーションの結果として得られる分子構造が初期条件と異なる可能性があるため，結果データ中の分子構造を認識する必要がある．そこで本稿では，$C_2H_2F_2$ 分子を用いて試験的に分子構造を分類した結果を示す．$C_2H_2F_2$ 分子はエチレン分子 C_2H_4 において水素を２個フッ素に置換したものと考えることができ，エチレンと同様に１つの平面上に６個の原子が存在する構造を持つことが知られている．すると，立体異性体が存在しないので，構造の数値的な扱いが容易である．３種類の異性体間の主な違いは，２個のフッ素原子の位置関係であるため，視覚的にも異性体を分類でき，分類結果の確認が容易である．

Gaussian09 では扱う原子に識別番号を付けて扱うが，今回は原子の番号付けを原子番号に対応づけ，原子間距離行列を異性体分類に使用する．原子間距離行列は，分子中の全ての原子の組み合わせで２原子間距離を計算し，行列状に並べたものである．今回，２個の炭素原子(C)には番号１と２を振り，水素原子(H)には３と４，フッ素原子(F)には５と６を振る．これは一旦入力ファイルを作成した後に，手動で各原子の初期位置座標を行ごと入れ替えたものを入力ファイルに用いると作成できる．今回用いるのは原子間距離行列のうち，同種原子の原子間距離，そして C-H 距離，C-F 距離，H-F 距離それぞれの最大値および最小値である．原子の番号付けを原子番号に対応づけた状態では，原子間距離行列から機械的に必要なデータを抽出できる．このデータでは，原子の種類を原子の番号として間接的に含めることができる．

今回用意したデータを用いて得られた決定木を図 3.に示す．図３は，R の出力に異性体の構造図と規則の意味を説明する吹き出しを付して作成した．図 3.ではまずフッ化ビニリデンが分岐しており，その分岐規則は２つの H 間距離である．ジフルオロエテンがシス形かトランス形かの分岐でも H 間距離を用いて分岐している．元のデータを見ると，C 間距離は３種類の異性体に大きな違いはないが，H で最大の距離は最小の距離の 1.66 倍であり，明らかに差がある．F も 1.63 倍異なるが H よりは差が小さい．異種原子での原子間距離はここまで顕著な差はない．そのため，H の原子間距離が分岐規則に採用されたと考えられる．得られた分岐規則は分子中の原子の数値的な配置とよく適合するものと考えられる．

本研究では粗放的に大量のシミュレーションを実行することを目標とし，Gaussian09 の標準的な計算方法と基底関数の組み合わせにおける実行時間を調べた．その結果，電子相関を考慮した基礎的な計算では，炭素原子１０個程度までならデスクトップ PC でも扱えることがわかった．さらに，決定木を用いて計算結果得られる分子構造の自動分類を試みた．その結果，おおむね妥当と考えられる分岐規則を自動的に得ることができた．

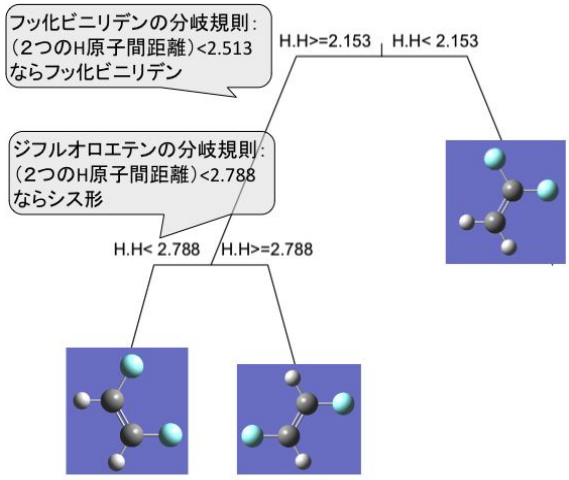

図 3. 原子間距離行列を用いて $C_2H_2F_2$ 分子を３種類の異性体に分類する決定木

今後の課題は計算結果の自動分類をさらに検討していくことである．より複雑な分子では，異性体も複雑かつ多種類になるため，そのような分子でも適切に分類できる手法を検討する．また，現段階ではシステムを部分的に実装しており，システム全体として機能するに至っていないので，一通り実装を終えることも今後の課題である．

５．主な発表論文等

〔雑誌論文〕（計２件）

① 林　亮子，水関　博志，``分散処理とデータ処理技術を利用した分子設計システム''，情報処理学会研究報告，査読無，Vol．2013-MPS-96，No．26，(2013)．

② 林　亮子，水関　博志，「計算科学とデータマイニングを用いた材料設計システム」，FIT2014 第 13 回情報科学技術フォーラム，査読無，掲載予定，(2014)．

〔学会発表〕（計１３件）

① 林 亮子，水関博志，"データマイニング技術を用いたシミュレーション結果自動分類の試み"，ナノ学会　第１２回大会，京都大学，京都府宇治市，２０１４年５月２２日．

② 林 亮子，「分類木を用いた分子構造の自動分類の試み」，第５４回人工知能学会分子生物情報研究会，北陸先端科学技術大学院大学，石川県能美市，２０１４年３月２０日．

③ 林 亮子，"データマイニング技術を用いた C2H2F2 分子の異性体自動分類"，２

０１４年ハイパフォーマンスコンピューティングと計算科学シンポジウム HPCS2014, 一橋大学, 東京都千代田区, ２０１４年1月7日.

④ 林 亮子, 水関 博志, "粗放的シミュレーション実行に基づく分子設計支援システムの試み", 第３６回情報化学討論会,筑波大学, 茨城県つくば市, ２０１３年１１月8日.

⑤ 林　亮子, 水関　博志, ”IT と分子シミュレーションを用いた分子設計システム”, ナノ学会第１１回大会, 東京工業大学, 東京都目黒区, ２０１３年6月6日.

⑥ 林　亮子, ”シミュレーションを用いた材料設計システムにおけるケモインフォマティクスの応用可能性”, 2013 年ハイパフォーマンスコンピューティングと計算科学シンポジウム HPCS2013, 東京工業大学, 東京都目黒区, ２０１３年1月１５日.

⑦ "マルチコアCPU上における古典MDの MPI プログラム", 平成２４年度北陸地区学生による研究発表会, 亀山 侑弥, 桜井　翔, 林　亮子, 福井工業高等専門学校, 福井県鯖江市, ２０１３年3月9日.

⑧ "GPGPU を用いた流体シミュレーションの高速化", 平成２４年度北陸地区学生による研究発表会, 藤井　浩貴, 古山　彰一, 林　亮子, 福井工業高等専門学校, 福井県鯖江市, ２０１３年3月9日.

⑨ "MPIを用いた並列古典MDプログラムのマルチコア上での性能評価（リンクセル法）", 桜井　翔, 亀山　侑弥, 林　亮子, 第２６回分子シミュレーション討論会, 九州大学, 福岡県福岡市, ２０１２年１１月２６日.

⑩ "MPIを用いた並列古典MDプログラムのマルチコア上での性能評価（Direct N^2 法）", 亀山　侑弥, 桜井　翔, 林　亮子, 第２６回分子シミュレーション討論会, 九州大学, 福岡県福岡市, ２０１２年１１月２６日.

⑪ "GPGPU を用いたダムブレイク問題の流体シミュレーション", 平成２４年度電気関係学会北陸支部連合大会, 藤井　浩貴, 古山　彰一, 林　亮子, 富山県立大学, 富山県射水市, ２０１２年9月1日.

⑫ 林　亮子, ”量子化学シミュレーション結果からの規則抽出の試み”, 第１０回先進的計算基盤システムシンポジウム SACSIS2012, 神戸国際会議場, 兵庫県神戸市, ２０１２年5月１６日.

⑬ 林　亮子, ”材料設計シミュレーションにおける規則抽出の試み”, 2012 年ハイパフォーマンスコンピューティングと計算科学シンポジウム HPCS2012, 名古屋大学, 愛知県名古屋市, ２０１２年1月２４日.

６．研究組織

(1)研究代表者

林　亮子　　（Ryoko Hayashi）

金沢工業大学・工学部・講師

研究者番号：30303332

(2)研究分担者

（研究分担者は置いていない）

(3)連携研究者

水関　博志（Hiroshi Mizuseki）

東北大学・金属材料研究所・准教授

（現在, Korea Institute of Science and Technology, Center for Computational Science）

研究者番号：00271966（東北大学当時）