Vol.2013-MPS-96 No.26
Vol.2013-BIO-36 No.26
2013/12/12

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

林亮子[†1]　水関博志[†2]

計算化学や計算物理のシミュレーションにおいては大量のデータが生成されるが，その利用技術にはまだ開発の余地がある．特に最近では小規模～中規模の計算性能がデスクトップ PC やクラウドサービスなどで大量かつ安価に利用できるため，既存の逐次処理パッケージプログラムを並行して大量に実行することが可能である．しかし，その結果データを活用するには自動処理が必須である．そこで近年筆者らは多数のシミュレーションを分散実行し，得られた大量の結果データに対してデータマイニングを利用した自動データ処理を行い，分子を設計するシステムを研究している．本稿ではこのシステムの最近の進行状況を報告する．

A Molecule-based Material Design System Using Distributed Processing and Data Processing

RYOKO HAYASHI[†1]　HIROSHI MIZUSEKI[†2]

Since computational chemistry or computational physics simulations output large-scale data, the technology utilize them have not be full-grown. Especially, the mass of computational power from small-scale to middle-scale became reasonable recently so that existent serial package programs are easy to run as distributed jobs. However, the auto-processing for the mass of data from such distributed many simulations is necessary. Therefore we have studied a molecule-based material design system based on many distributed simulations and auto-data-processing technology including data mining. This manuscript reports the recent research progress about our system.

1. はじめに

近年では計算機が高速化しており，また，計算化学や計算物理のシミュレーションプログラムも成熟してきたので[1][2]，それらを利用して誰でも容易に大量のシミュレーションを行うことができる．しかし，シミュレーション結果は大量の数値データであり，有効利用するためにはデータの自動処理が必要であるが，シミュレーションにおけるデータの自動処理技術は，まだ開発の余地がある．一方で最近ではデータマイニング技術が成熟してきており，データマイニングの主要な手法が容易に誰でも利用できる環境が整ってきている[3]．

著者らは以上の状況を考慮して，分散処理を用いた分子設計システムに関する研究を行っている[4]．本研究ではまず分子数個レベルで原子の総数が数十個以内で構成する材料を想定する．本稿ではデータマイニング技術を用いてシミュレーションの結果得られる分子構造の分類を試みた結果を報告する．

本稿の構成は次の通りである．第２章では本研究が目的とするシステムの概要を紹介する．第３章では，決定木を用いて分子の異性体の分類を試みた結果を紹介する．第４章では本稿の結果をまとめ，今後の課題を述べる．

2. 分子設計システムの概要

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 のように，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．さらに，使用する原子など，自動生成したい範囲を示すのに必要な条件を設定すると，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．これはひな形中で変更したい箇所に文字列処理を行い，原子位置座標や原子記号を置換することで実装できる．さらに，生成したファイルを用いて自動的にジョブを投入し，終了したかどうかを管理する．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．本研究では，シミュレーションには広く使われる量子化学パッケージの一つである Gaussian09 を使用する．

図 1 において，シミュレーションを実行するまでの手順に技術的に大きな問題はない．一方でシミュレーション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．本研究では主に構造最適化を扱うが，シミュレーションの結果として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，シミュレーションの結果ファイルを用いて，得られた構造を認識する必要がある．

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，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．

†1 金沢工業大学
Kanazawa Institute of Technology
†2 Korea Institute of Science and Technology
Korea Institute of Science and Technology

Vol.2013-MPS-96 No.26
Vol.2013-BIO-36 No.26
2013/12/12

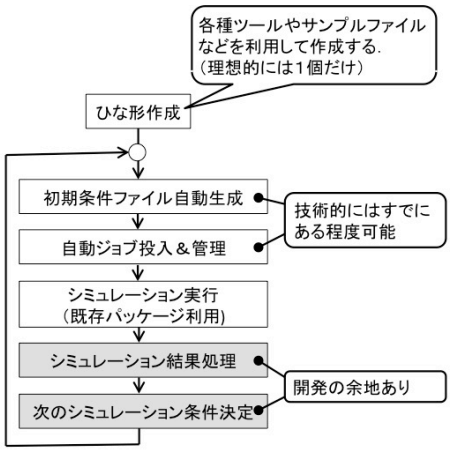

図 1　分子設計システムの概要図

Figure 1 Illustrated image for molecular design system.

3. 決定木を用いた分子構造の分類

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３種類の異性体を持つが，それらの原子配置を図 2 に示す．３種類の異性体間の主な違いは，フッ素原子の位置関係である．なお，図 2 中で原子に記入した番号は，Gaussian09 の入力ファイル中で各原子に自動的につけられた番号である．

実験条件を示す．まず図 2 のような構造を Gaussian09 の支援ソフトである GaussView5 で作成し，Gaussian09 を用いて構造最適化を行い，結果データから原子間の距離行列を抽出して分類に使用する．なお，今回使用したデータでは，図 2(a)と図 2(b)で同じ種類の原子には同じ番号が付いており，いずれも２個の炭素原子には１と３が付いているが，図 2(c)では原子の番号付けが異なっていて２個の炭素原子には１と４が付いていることに注意が必要である．

今回は，分子構造を分類するために距離行列を入力として決定木を作成した．決定木作成には，R のパッケージ mvpart を用いた．このパッケージはジニ係数を計算して，２進分岐する決定木を作成する．得られた決定木を図 3 に示す．これは R の出力であるが，そのままではわかりにくいので，分類規則の意味を図 3 中に吹き出しで記入した．

図 3 で得られた分岐規則の妥当性を議論する．(a と b)と c の間の分類規則は「原子３と原子１間距離<1.192 なら異性体 c である」というものである．しかし，a と b では原子１と原子３のどちらも炭素原子であるが，c では原子３は水素原子であるため，番号付けをしている対象の原子の種類が異なることが影響している可能性がある．一方，a と b の分岐規則では原子４と原子２の距離で分岐している．原子４と原子２はいずれも水素原子であって， $C_2H_2F_2$ 分子を扱う場合，水素原子間距離の違いで分岐することとフッ素原子間距離の違いで分岐することは論理的に等価であるため，人間が目で見てこれらの分子を分類する場合に近い規則で分岐したと考えられる．

(a) E-1,2-ジフルオロエテン　(b) Z-1,2-ジフルオロエテン　(c)フッ化ビニリデン

図 2　$C_2H_2F_2$ 分子の３種類の異性体

Figure 2 3 isomers of $C_2H_2F_2$ molecule.

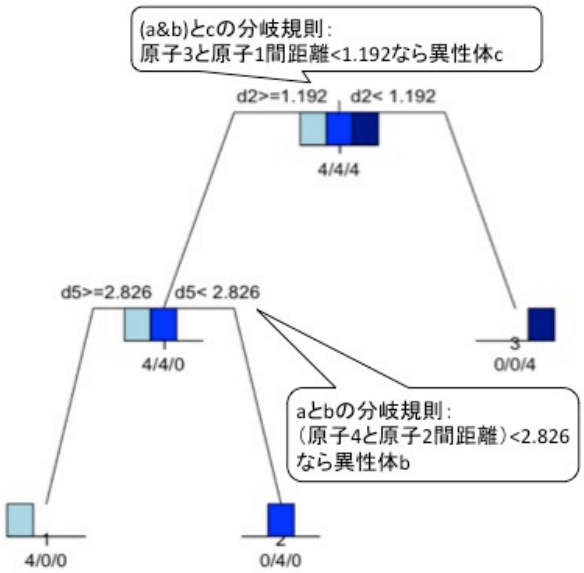

図 3　$C_2H_2F_2$ 分子を３種類の異性体に分類する決定木

Figure 3 Decision tree to classify a $C_2H_2F_2$ molecule to 3 isomers.

4. おわりに

本稿ではシミュレーション結果中の分子構造の自動分類を試みたが，原子の番号の付け方に依存して，不適切な分類ルールを作成してしまう可能性がわかった．この問題を解決するには，入力データに原子記号を用いることや，適切に分類できるように原子の番号を付けることが考えられる．今後の課題として検討していきたい．

謝辞　本研究の一部は科学研究費補助金基盤(C)課題番号 23500138 による．関係各位に感謝する．

参考文献

1) M. J. Frisch, et. al, Gaussian 03M, Revision E.01, Gaussian, Inc., Wallingford CT, (2004).
2) 「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).
3) 「データマイニング入門」, 豊田　秀樹編著, 東京図書, (2008).
4) “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", 林 亮子, 水関 博志, 第３６回情報化学討論会講演要旨集, pp.58-59, (2013).