A-019

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

Material Design System Using Computational Science and Data Mining

林　亮子†　　水関　博志‡
Ryoko Hayashi　　Hiroshi Mizuseki

1．はじめに

近年では GPGPU やマルチコア計算機，グリッド／クラウドコンピューティング環境が容易に利用できるようになってきている．これらは小規模～中規模の計算資源とみることができ，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きており[1][2]，それらを利用して誰でも容易に大量の材料設計計算を行うことができる．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．しかし，材料設計関連分野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．

一般的なデータ処理技術に目を向けると，近年ではいわゆるデータマイニング技術が成熟してきており，主要な手法が誰でも容易に利用できる環境が整ってきている[3]．そこで著者らは近年，計算結果の自動データ処理に基づく材料設計を目標として，研究開発を行っている[4]．本稿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，直鎖炭化水素の構造最適化ジョブの性能評価を用いて，本研究が扱える分子サイズと計算モデルを議論する．第４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５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

2.1.　システムの開発方針

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．

2.2.　システムの概要と問題点

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．

2.3.　データ処理

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．

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．

†金沢工業大学, Kanazawa Institute of Technology

‡Korea Institute of Science and Technology, KIST

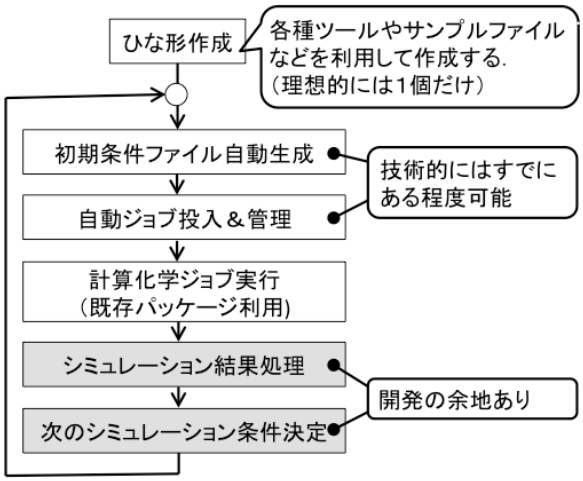

図 1. 分子設計システムの概要図
Figure 1. An illustrated image for molecular design system

３． 取り扱い可能な分子サイズの評価

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．文献[2]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）

これらは文献[2]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

表１　性能評価を行った計算機環境の諸元
Table 1　Computational environment for the performance evaluation.

機種名	Mac Pro 4
OS	Mac OS X 10.6.8
CPU	Quad-Core Intel Xeon×1
周波数	2.66GHz
コア数	4
メモリ	8GB (DDR3 ECC,2GB,1066MHz×4)
キャッシュ	二次キャッシュ（コア単位）256KB, 三次キャッシュ 8MB

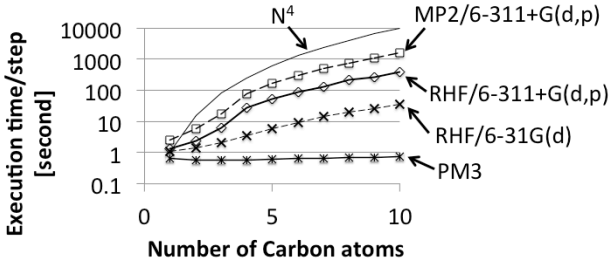

図 2. 直鎖炭化水素における炭素原子個数と１ステップの実行時間の関係
Figure 2. Relationships between the numbers of carbon atoms with execution time per step using alkane single chain

れていくので，おおむね実行時間も短くなる．

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．

４．決定木を用いた分子構造の分類

4.1.　実験条件

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

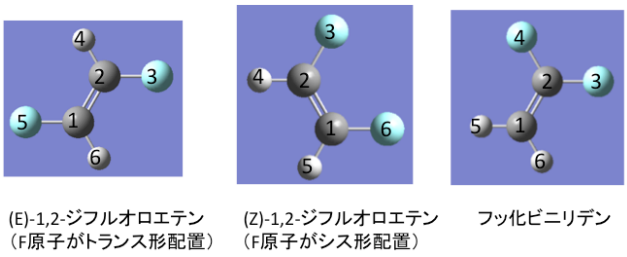

図 3. $C_2H_2F_2$ 分子の 3 種類の異性体
Figure 3. Three isomers of $C_2H_2F_2$ molecule

```
#T MP2/6-31G(d) Opt Test geom=connectivity   ←計算内容を指定するキーワード
C2H2F2 pattern 3   ←コメント行

0 1
C          0.00000000   0.00000000   0.00000000
C         -1.17094448   0.00000000  -0.68224340
H         -1.17094448   0.00000000  -1.75224340
H         -2.09864940   0.00000000  -0.14907965
F          0.00000000   0.00000000   1.07000000
F          0.92770492   0.00000000  -0.53316375
                                    ←各原子の種類と位置座標

1 2 2.0 5 1.0 6 1.0
2 3 1.0 4 1.0
3
4
5
6
                                    ←結合している原子と結合の種類
```

図 4. Gaussian09 の入力ファイル例
Figure 4. An example of input file for Gaussian09

存在しないので，構造の数値的な扱いが容易である．$C_2H_2F_2$ 分子は 3 種類の異性体を持つが，それらの原子配置を図 3 に示す．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．

なお，図 3 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．3 つの異性体に共通して，炭素原子には番号 1 と 2 がついており，1 個のフッ素原子には 3 がついている．しかし 3 番目以降は図中で半時計回りに番号が付されているために，番号と原子の種類の対応は，異性体ごとに異なる．

実験の手順を示す．まず図 3 のような構造を Gaussian09 の支援ソフトである GaussView5 を用いて手動で作成し，保存すると，Gaussian09 で使用する入力ファイルができる．入力ファイルの一例を図 4. に示す．図 4. に示すように，入力ファイルは計算内容を指定する行，計算したい原子の種類と位置座標，結合している原子と結合の種類をテキストで記述したもので，エディタ等で編集することも可能である．図 4.のような内容のファイルを入力に用いて Gaussian09 を実行して構造最適化を行い，結果データから必要部分を抽出したものを分類に使用する．

異性体分類に使用するデータを今回は 2 種類用意した．

表 2. Z マトリックスのデータ例
Table 2. Data example for Z matrix

isomer	R (1,2)	R (1,3)	R (1,4)	R (2,5)	R (2,6)	A (2,1,3)	A (2,1,4)	A (3,1,4)	A (1,2,5)	A (1,2,6)	A (5,2,6)
trans	1.300	1.360	1.065	1.360	1.065	120.4	124.5	115.1	120.4	124.5	115.1
cis	1.301	1.358	1.065	1.065	1.358	123.4	122.0	114.6	122.0	123.4	114.6
vinylidene	1.298	1.335	1.335	1.067	1.067	125.3	125.3	109.4	120.0	120.0	120.0

概要を以下に述べる．

データ1. 原子の番号付けを GaussView5 まかせとし，Z マトリックスを分類に使用する．使用したデータの一例を表 2.に示す．Z マトリックスは化学で分子構造を数値化するのによく用いる形式で，結合している原子間で，2 原子間結合距離と 3 原子がなす角度，4 原子がつくる 2 面角を含む．今回は平面上に存在する分子であるため，2 面角は実際には用いない．表 2. 中の R(1,2)から R(2,6)が 2 原子間結合距離で，A(2,1,3)から A(5,2,6)が 3 原子がなす角度である．R(1,2)は 1 番目の原子と 2 番目の原子での原子間距離であり，A(2,1,3)は 2 番目，1 番目，3 番目の原子の並びが作る角度である．原子が作る角度を陽に扱えることが，このデータ形式の利点である．一方表 2 では，原子の種類をデータに一切含まないことに注意が必要である．

データ2. 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．使用したデータの一例を表 3. に示す．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．これは一旦データ 1 と同様に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．表 3. に示すように，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表 3 の量を抽出できる．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．

今回はデータ 1 とデータ 2 のいずれも，「どの異性体か」をデータに含めており，異性体を分類する規則を決定することを目標とする．今回は異性体を分類する決定木の作成を試み，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． 入力データは RHF/3-21g, RHF/6-31G(d), MP2/6-31G(d) の 3 種類の計算方法と基底関数の組み合わせで表 2. と表 3. の量を用意して使用した．

4.2. 3 種類の異性体の決定木

最初に，データ 1 を用いて得られた決定木を図 5 に示す．これは R の出力の木に，図 3 の分子構造を相当する部分に付して作成し，分類規則の意味を図 5 中に吹き出しで記入した．mvpart は 2 進分岐する決定木を作成するため，分類規則は 2 つ得られる．

図 5. で得られた分岐規則の妥当性を議論する．トランス型ジフルオロエテンが他の異性体から分岐する規則は，原子 2 と原子 5 の距離すなわち R(2,5)が他の異性体よりも大

表 3. 原子間距離行列のデータ例
Table 3. Data example for atomic distance matrix.

isomer	C–C	H–H	F–F	C–H max	C–H min	C–F max	C–F min	H–F max	H–F min
trans	1.300	3.060	3.559	2.096	1.065	2.308	1.360	2.608	2.053
cis	1.301	2.430	2.795	2.072	1.065	2.341	1.358	3.313	2.045
vinylidene	1.298	1.848	2.179	2.051	1.067	2.338	1.335	3.290	2.607

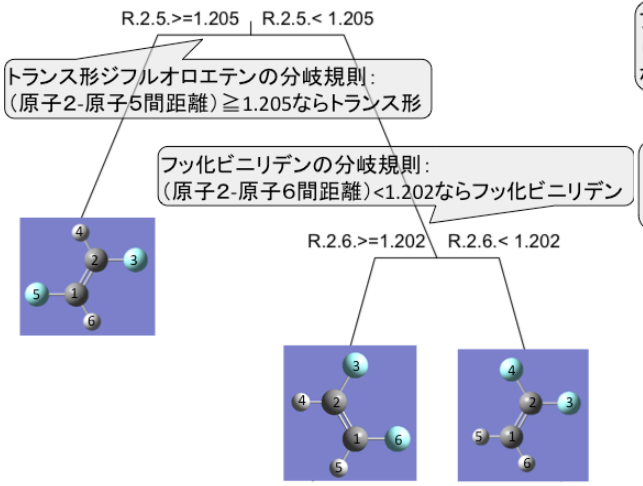

図 5. Z マトリックスを用いて $C_2H_2F_2$ 分子を
３種類の異性体に分類する決定木
Figure 5. A decision tree to classify a $C_2H_2F_2$ molecule to
3 isomers using Z matrix

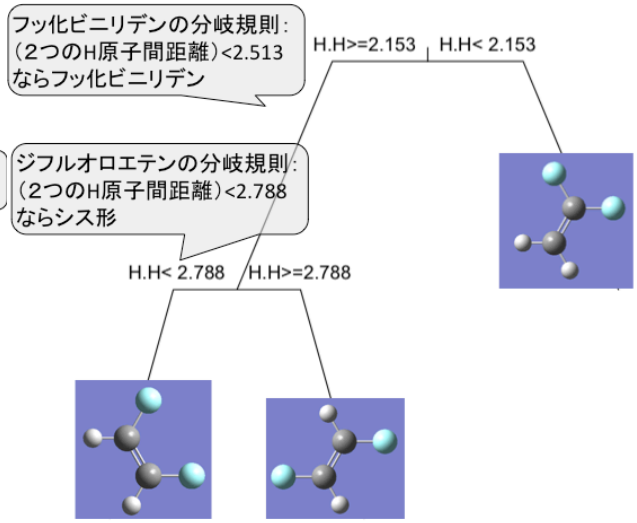

図 6. 原子間距離行列を用いて $C_2H_2F_2$ 分子を
３種類の異性体に分類する決定木
Figure 6. A decision tree to classify a $C_2H_2F_2$ molecule to
3 isomers using distance matrix among atoms

きいことであるが，原子２と５はトランス形ジフルオロエテンでは C と F であり，他２つの異性体では C と H である．そのため，間接的に H と F の違いで分岐していることとなる．そして，フッ化ビニリデンとシス形ジフルオロエテンの分岐規則は原子２と原子６の距離であるが，これも原子６が H か F の違いで分岐していることとなる．

表 2. 中の数値を，分岐規則に採用されたものとそれ以外で比較する．R(1,2)および R(1,3)は異性体間でほとんど差がない．３種類の異性体で共通して R(1,2)は C 同士，R(1,3)は C と F の原子間距離であって，ほとんど同じである．全て同じ原子のトランス形ジフルオロエテンの R(2,5)は他の種類の異性体の 1.3 倍程度であり，明らかに差がある．R(1,4)もフッ化ビニリデンが他の異性体の 1.3 倍程度で明らかに差があるが，R(2,5)よりもわずかに差が小さいので，R(1,4)を用いた分岐規則は採用されなかったと考えられる．R(2,6)は R(2,5)よりも異性体間の差は小さいが R(1,4)よりは大きかったので，２つ目の分岐規則として採用されたと考えられる．３体角は原子間距離ほど顕著な差がなく，最大でも 10%程度の差であったため，分岐規則には採用されなったと考えられる．

次に，データ２を用いて得られた決定木を図 6. に示す．図 6. も図 5. と同様に，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．データ２では，原子の種類を間接的にデータ中に含んでいる．図 6. 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．表 3. 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．

いずれのデータでも，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．化学的，実験的な知見との比較を今後検討したい．

５． おわりに

本稿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．

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．

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]"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, 水関　博志，情報処理学会研究報告，Vol. 2013-MPS-96, No. 26, (2013).