

データマイニング技術を用いた $C_2H_2F_2$ 分子の異性体自動分類

林 亮子[†]

[†]: 金沢工業大学 工学部 情報工学科 〒924-0838 石川県白山市八束穂 3-1

近年では計算機が高速化し、大量の計算資源が利用できるため、計算化学や計算物理のシミュレーションを用いた材料設計が注目されている。しかし、大量の材料設計シミュレーション結果は大量の数値データであってデータの自動処理が必要であり、大量のデータから知識発見を行うデータマイニング技術^[1]によって材料設計技術がより発達する可能性がある。筆者らは近年、多数のシミュレーションの自動実行とデータマイニングを組み合わせた材料設計システムに関する研究を行っており、本ポスターでは、シミュレーション結果データを用いて分子構造の自動分類を試みた結果を報告する。

今回は $C_2H_2F_2$ 分子を扱う。 $C_2H_2F_2$ 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、フッ素の位置関係が異なる 3 種類の異性体を持つことが知られている。そこで、 $C_2H_2F_2$ 分子の 3 種類の異性体を初期条件に用いて量子化学パッケージ Gaussian09^[2]によって構造最適化した結果データを用意し、得られた構造の分類を試みる。データ処理には、統計処理プログラミング言語 R^[3]を用いる。R はオープンソースでユーザがパッケージを開発可能であり、主要なデータマイニング手法がすでにパッケージ化されているので、それらを利用する。

Gaussian09 の出力ファイルは大量のデータを含むが、構造を表すデータとして Z マトリックスと原子間距離行列が含まれている。Z マトリックスは分子構造を表すのに良く使われる形式で、原子間の相対距離の一部とボンド角や二面角などの分子構造を直接示す量が含まれるが、原子を識別する番号の適切な取り扱いが必要であり、データマイニングを適用するのは一段難しい。そこで、原子の識別番号がデータに及ぼす影響が小さいと考えられる原子間距離行列を用いて異性体分類を試みる。結果は当日紹介する予定である。関連分野の方々のご意見を戴けると幸いである。

References

- [1] 「データマイニング入門」, 豊田秀樹, 東京図書, (2008).
- [2] M. J. Frisch, et. al, Gaussian 03M, Revision E.01, Gaussian, Inc., Wallingford CT, (2004).
- [3] “R: A language and environment for statistical computing“, R Development Core Team, R Foundation for Statistical Computing, Vienna, (2009).

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。