平成 28 年度　利用報告書

課題名（タイトル）：

量子化学による分子の巨大データベース構築

利用者氏名：　〇中田 真秀*　島崎 智実**

所属：*情報基盤センター 和光ユニット

**** 計算科学研究機構** 量子系分子科学研究チーム

１．　本課題の研究の背景、目的、関係するプロジェクトとの関係

大規模量子化学計算データベースを構築する。数多くの化学物質に対して幅広い特性を実験によって求めることは困難な場合が多い。毒性等を持つ分子が存在することや、信頼性の高い実験を多数の分子に行うには非常にコストが掛かるためである。本研究プロジェクトでは、量子化学計算を用いることによって、数多くの分子に対するデータベースを作成する。量子化学計算を用いれば、非常に高精度で分子の特性を求めることが出来る。例えば、典型的な B3LYP/6-31G*計算での幾何構造での誤差は結合長 0.1A 程度、結合角数度程度である。本研究では、世界最大規模の分子データベースである NIH の PubChem プロジェクトからの分子情報を取得し、RICC 上で量子化学計算を行う。そして、計算結果を蓄えるデータベースを構築する。本プロジェクトでは、このように構築したデータベース(http://pubchemqc.riken.jp/)を基礎として、データマイニングや機械学習などのデータサイエンス技術を導入する。そして、これらの要素技術を連携させることによって、材料探索法の確立を目指す。本研究では、特に有機電子デバイス材料開発への応用を試みる。

２．　具体的な利用内容、計算方法

分子情報を InChI および SMILES 形式で PubChem Compounds データベースから取得する。それらを用い、初期座標を生成、半経験的分子軌道法などを経て、最終的には分子の幾何構造を B3LYP/6-31G* レベルで最適化する。それから HOMO-LUMO ギャップ、励起状態、双極子モーメント等、電子デバイスに有用な計算も行う。計算は GAMESS, SMASH などの量子化学計算プログラムパッケージを用いた。

有機電子デバイスなどに応用する際は、候補分子を多数必要とする。しかし、PubChem Compounds データベースに登録されている分子だけでは充分でないことが予想される。従って、PubChem Compounds の計算結果を学習させ、より簡単な量子化学計算の方法を模索する必要がある。今年度は、サポートベクターマシン(SVM)を用いた機械学習により、分子の性質予測を行った。ここで、HOMO-LUMO ギャップおよび励起エネルギーをターゲット特性とした

３．　結果

昨年度に引き続き、pubchemqc プロジェクトは新たに 100 万分子程度の計算を終了した。これまでの計算を合算すると、基底状態計算では 400 万分子、励起状態計算では 300 万分子のデータを持つデータベースを構築した。

SVM による分子の励起エネルギー予想は下の表にまとめる。およそ 0.3eV 以内の誤差で励起エネルギーが予測することが出来る。これは有機デバイスだけでなく化学反応などにも応用できる程度の精度である。

	narrow dataset	wide dataset
SVM-RBF	0.283	0.351
SVM-poly2	0.314	0.412
SVM-poly3	0.368	0.510

４．　今後の計画・展望

さらなる分子の計算を行い、データベースの拡充を図る。そして、SVM だけでなく、深層学習等も検討する。そして、分子の基底状態の幾何構造推定、励起エネルギー予測、HOMO-LUMO ギャップ予測を行う。また、これらの技術に基づいたデータベース検索システムを構築中であり、有機電子デバイスに対して適用を試みる。