Vol.2014-HPC-145 No.27
2014/7/29

第一原理計算による分子の物理化学データベース構築

中田 真秀[1,a)]

概要：化学において第一原理計算を行えば、多くの場合実験結果と同等または凌駕するような計算結果が得られるため、分子の網羅的な計算に基づいたデータベースを作成することには大きな意味がある。我々はアメリカ国立衛生研究所が主催する The PubChem Project (http://pubchem.ncbi.nlm.nih.gov/) にある世界最大規模の分子データを元にして、第一原理計算を行った結果を The PubChemQC Project (http://pubchemqc.riken.jp/) として 2013/12/21 から公開している。このプロジェクトでは、データベース作成、そしてそれから得られるビッグデータの解析から、完全なる in silico での新規機能分子の創出、提案を行うシステム構築を目指している。2014/6/30 の時点で 116,869 分子の計算が終了、結果の公開をしていることを報告する。

1. はじめに

化学の研究の重要なものに「ある物性、機能を持った分子を設計したい」というものがある。逆に個々の分子を特定すれば、第一原理計算を行えば、多くの場合実験結果と同等または凌駕するような計算結果が得られる。

化学は Schrödinger 方程式に従うため、量子化学は原理的には分子設計できる能力がある [1]。にもかかわらず残念ながら実験に取って代わるほどとはなっておらず、実験結果の説明程度にとどまっていることが多い。この理由には、化学的に妥当な計算を行うには、経験がないと難しい、計算量が大変多いなどがあるが、もっと大きな理由としては、化学的直感とよばれる人間の思考をある程度模倣できるようなシステムがないということがある。

化学的直感は、非常に高度な思考ではあると思われる。しかし、それのコンピュータによる再現の目標をそのすべてではなく、今までの知識を前提とした若干の延長に置くとすると、そこまでは困難ではないのではないと考えられる。しかも、それだけでも人間一人の知識量には限界があるため意味があると考える。これらを意識した研究に、近年、機械学習分野で deep learning の研究があり、非常に盛んである。成功例としては、画像認識 [2] [3] や音声認識 [4] などがあげられる。もちろん化学への応用も始まっている [5]。

機械学習における精度を上げるためには多くのデータが必要となるが、化学におけるパブリックな分子データベースには統一された物理量 (たとえば双極子モーメント、イオン化エネルギー、励起エネルギーなど) が掲載されているものは殆ど無く、最も大規模なものでも NIST Chemistry WebBook 程度である [6]。たとえば、赤外スペクトルは 16000 分子以上、UV 可視光スペクトルは 1600 分子実験値などが提示されている。当然ながら分子の性質を調べる実験にはコストが大変かかる。もしこれを理論計算で行うならばならばコンピュータの利用コストのみで実験を行う必要もなく、圧倒的に安全、高速かつ経済的である。

分子データベースとして世界最大規模かつ無料で利用できるものとしては PubChem プロジェクトがある [7]。これは NIH の分子ライブラリロードマップイニシアチブの構成部分であり、小さな分子の生物学的な活性について網羅的に情報を提供しようとしている。このプロジェクトは Pubchem Substance, Pubchem BioAssay, Pubchem Compound の 3 つがあるが PubChem compound は標準化し、意味があり、重複のない分子構造を記録してる。例えば、多くの製薬企業が自社のカタログなどから抽出した化合物を提供している。他には炭素、窒素、酸素、硫黄、塩素、水素のみ、13 原子からなる分子をすべて網羅的に数え上げた GDB-13 [8] や、人間によって精選された、薬理活性を示すと思われる生理活性を持った分子データベースである CheEMBL [9] などがある。

今回は PubChem Compound に掲載されている分子の第一原理計算を行うことにした。理由は、PubChem は世界最大規模の収録数を誇っていること、他の ChemEMBL や他のデータベースからの分子も入ってきていることが挙げられる。また、GDB-13 は元素の種類が足りないこと、

[1] 理化学研究所 情報基盤センター
351-0198 埼玉県和光市広沢 2-1
[a)] maho@riken.jp

網羅的だが存在しない分子も多く含んでいること、の理由で採用しなかった。

PubChem には 5000 万以上の分子が登録されており、また、理論化学計算は非常に時間がかかるため網羅的に計算結果を提示するというのは現実的ではない側面もある。例えば、1 日 1 万分子計算できたとして 5000 日、14 年かかる。現在の我々の持っているリソースでは、1 日 1000 分子から 10000 分子の間であり、実現はかなり難しい。しかし今後の量子化学のアルゴリズムの進展、および特に計算機の発展を考えると不可能ではないと考える。

このプロジェクトにおける分子情報の提供は以下の点で非常に有益であると考えられる。

- 量子化学計算により、実験値と比較しうる非常に精密な結果を提供できる。
- 物性値の横断的な検索による、今まで不可能であった圧倒的に効率的かつ低コストな物質探索。
- 機械学習による分子計算の精度向上が可能になると考えられる。

これらを踏まえ、PubChemQC プロジェクトを行っている [10]。計算の大まかな流れについては以下のようになっている。

- Pubchem compound から分子の情報をダウンロード。
- 初期分子構造の生成、つづいて半経験的分子軌道法である PM3 を用いての構造最適化を行う。
- さらに構造最適化を Hartree-Fock 計算を STO-6G 基底関数で行い、密度汎関数法 (B3LYP 汎関数) を用いて 6-31G(d) 基底関数を用い、構造最適化を行う。
- 最後に先ほど得られた分子の構造に対し、エネルギーの低い順に 10 個励起状態計算を時間依存密度汎関数法に 6-31G(d)+基底関数を用いて行う。
- 計算が終了したものから http://pubchemqc.riken.jp/ へアップロード。

2. 関連研究

これまでのデータベース構築には非経験的分子軌道法を用いた計算が多かった。PubChem3D[13] は PubChem プロジェクトのものであるが、構造最適化は MMFF94s 力場を用いて 50 の水素原子以下、15 の回転可能な結合、H, C, N, O, F, Si, P, S, Cl, Br, および I 原子のみ含んだ分子について行ってある。これは精度が第一原理計算ではないため、精度は格段に劣る。また、NIST の Computational Chemistry Comparison and Benchmark DataBase[14] は 1591 分子について、350,000 種類以上の計算が掲載されている。基底関数の選択や計算手法は様々である。我々の研究は基底関数、計算手法は結果のコンシステンシー (分子毎に手法が違うと比較が困難になる) や大量分子の計算に耐えうる、ということから構造最適化には 6-31G(d) 基底関数、励起状態には 6-31G(d)+基底関数を用い、計算手法は密度汎関数法に B3LYP 汎関数を用い、約 12 万分子の結果という大量の結果を提示できている。

3. 計算の進め方

量子化学計算における分子によらないほぼ一貫したパラメータ探索、自動計算システムなどの開発に主眼においた。分子の数が多く、一つ一つの計算は依存しない計算である (いわゆる Embarrassingly parallel な計算である)。サブセクションで各々計算技術の概要を示す。

3.1 PubChem Compound からの分子の情報収集

分子の情報は PubChem プロジェクトの Compound から FTP サイトから取得した [11]。2014/6/27 原稿執筆時には、51,011,180 分子あった [12]。日々更新されているため、取得した日によって内容は更新されていくはずではあるが、データの更新と我々の計算が一致しているか、乖離しているかは簡単にチェックできるため、今回は厳密にチェックはしなかった。FTP サイトでは約 25,000 分子を単位として、いくつかのフォーマットで、3000 ファイルほど提供されている。今回はフォーマットとしては SDF (structure-data file) を使った。PubChem Compound の SDF には

- 分子の水素をのぞいた三次元構造
- IUPAC 名
- InChI, SMILES 名
- 分子量

などが記述されている。三次元構造は PubChem 3D プロジェクト [13] によって最適化されたものではあるが今回は用いていない。

3.2 SMILES 表記

分子の初期座標生成には SMILES 表記から行った。SMILES は Simplified Molecular-Input Line-Entry System の頭文字をとったもので、ASCII 文字を用い、一次元的に、分子の構造を表現する方法である [15]。以下、SMILES 表記について非常に簡単に説明する。原子は B, C, N, O, P, S, F, Cl, Br, I はそのまま表記する。他はブラケットでくくる (たとえば金は [Au]) 基本的に水素は必要な数だけ存在しているとし、省略する。省略できない場合は陽に書く。化学結合は、単結合の場合省略する。二重結合は “=” で書く。従って二酸化炭素は O=C=O となり、三重結合は “#” と書く (窒素分子は N#N となる)。結合の分岐は CC(=O)C など括弧でくくる。ベンゼンのような環化合物は C1=CC=CC=C1 と書き、この二つの “1” が環の閉じている部分をさす。

SMILES 表記は一分子についていくつも存在することがあるため、ユニークな対応を求めるようなアルゴリズムが研究されている。その中で特に重要なのが Universal SMILES [16] である。これを用いると PubChem 分子の 100 万分子

程度について 99.77% の分子について SMILES 表記がユニークに定まったとあった。さらに OpenBABEL[17] にも実装されているため手軽に利用することができる。

また、分子については、鏡に映した構造と元の構造が一致しないような、不斉構造が重要となる場合がある。その情報を一部含んだ Isomeric SMILES からの構造を計算に用いた。

3.3 量子化学計算の詳細

PubChem の SDF を 1 単位に分解し、分子量でソートした。中には化合物の混合物が含まれているため、それは除外した。これは SMILES 表記に "." が含まれている場合なので機械的に除去できる。例えば薬理分子を安定化させるために塩化水素を入れている場合がそうである。またフェロセンのような金属錯体も除外される場合がある。

そして量子化学計算での方法論および基底関数の都合上、都合上 H, He, Li, Be, B, C, N, O, F, Ne, Na, Mg, Al, Si, P, S, Cl, Ar, K, Ca, Sc, Ti, V, Cr, Mn, Te, Co, Ni, Cu および Zn を含んだ化合物のみを計算している。

それをふまえ一分子ずつ、Isomeric SMILES 表記を切り出し、OpenBABEL により分子を 3 次元化した ("–gen3d -addH")。それを初期構造として PM3 法により構造最適化を行う。そこで求まった構造を用い、Hartree-Fock 法で STO-6G 基底関数を用いて構造最適化を行う。最終的には密度汎関数法に B3LYP 汎関数、6-31G(d) 基底関数を用いて構造最適化を行った。最後の密度汎関数法での構造最適化を行ったところは実際は三段階踏んでいる。最初に FireFly[19] でまず構造最適化を行い、それに続いて GAMESS[18] で構造最適化を行っている。理由は FireFly のほうが高速ではあるが、計算精度が若干劣るため、途中経過のみを用い、最終計算はより信頼性の高いパッケージである GAMESS を用いることにしたためである。三段階目は最終構造を入力とした計算を行い、構造最適化を一度で終了させ、真に極値であること容易に示すためである。

合計して 6 回も計算を行っている。そもそも SMILES 表記を用いた初期構造から直接、密度汎関数法に B3LYP 汎関数、6-31G(d) 基底関数を用いた計算を行えばよいとも考えられなくもないが、それでは構造が収束しない分子がでてくる。段階的に計算のクオリティをあげてゆく方法を採った。

最後に最適化された構造を用い、時間依存密度汎関数法で、6-31G(d)+基底関数を用い、励起状態をエネルギーの低い順から 10 個ずつ求めた。

計算結果のファイルおよび入力ファイルを http://pubchemqc.riken.jp にアップロードした。

3.4 計算機について

計算アルゴリズムは、CPU に多くの負担がかかり、メモリ量、メモリバンド幅、ストレージのスピードにはあまり影響されないダイレクト法を主に用いた。

現在利用中の計算機は

(1) 理化学研究所情報基盤センター RICC (CPU: Intel Xeon 5570 (2.93GHz))
(2) 東京大学情報基盤センター FX10 スーパーコンピュータシステム (CPU: SPARC64 IXfx 1.848 GHz)
(3) 理化学研究所 Quest クラスタ (CPU: Core2 L7400 (1.50GHz)) 500 台
(4) Intel Core(TM) i7 920 @ 2.67GHz 4 コア 1 台
(5) Intel Xeon(R) X3470 @ 2.93GHz 2 コア 1 台
(6) Intel Xeon(R) E5-2650 @ 2.00GHz 8 コア 2 台
(7) Intel Xeon(R) E5-2670 @ 2.60GHz 16 コア 3 台
(8) Intel Xeon(R) E5-2680 @ 2.70GHz 16 コア 1 台
(9) Intel Xeon(R) E5-2603 @ 1.80GHz 8 コア 1 台
(10)Intel Core(TM) i7-3770K @ 3.5GHz 4 コア 1 台
(11)Intel Xeon(R) X5355 @ 2.66GHz 8 コア 1 台

となっている。

3.5 計算結果について

基底関数、計算方法、構成原子とその分子構造に依存することと、実験値との比較を網羅的には行えないため、計算精度については正確なことはいえないが、電子構造としては、Hartree-Fock 法よりよく、MP2 法程度に求まっていると考えられる。分子構造は、必ずしもエネルギー最小の構造ではなく、SMILES から OpenBABEL によって出された初期構造や、途中に使った計算手法に依存するが、現在多くの研究で実質的に分子構造が求められないものに対してこれに近いやり方が使われているので、もし非常に大きく違う場合はその都度再計算することになると思われる。

計算時間は計算機、分子の性質 (特に分子量) によって大きく変わるため単純な比較はできないためおおざっぱになるが、分子量が 500 以下ならば、RICC を用いた場合 (CPU: Intel Xeon 5570 (2.93GHz) x 2)4 コア、12 時間程度で終わった。それ以上の分子量を持った分子は構造最適化にかかる時間が多くなってくるため、計算は行えていない。一番時間がかかるのは、密度汎関数法で構造最適化をする部分および時間依存密度汎関数法で励起状態を求める部分であった。他の計算にかかる時間はほとんど無視できる。また、一つの SDF には最大 25000 分子存在可能だが、実際は 20000 分子程度、混合物をのぞくと 18000 分子程度であるため、2〜3 週間程度で一つの SDF を処理できることになる。尚、正確な時間測定は困難かつ意味が無い。一貫した計算パラメーターを確立するのには半年程度の時間を要したこと、スクリプトの調整、バグ修正などで測定の意味、時間が大きく変わること、2014 年に入っていったん計算をやり直したこと、そのときに途中結果を利用して計算短縮を図ったこと、分子によって大きく計算時間がかわ

ること、などがあげられる。

4. まとめと今後の展望

PubChemQC プロジェクト http://pubchem.riken.jp/の現在の概要を述べた。NIH の PubChem Compound に登録されている 5000 万分子のうち約 12 万分子の基底状態の最適化された分子構造およびその時の電子構造、そしてこの分子構造を用いた励起状態の電子構造が第一原理量子化学計算プログラムパッケージ GAMESS および FireFly によって計算、公開されている。分子の計算には密度汎関数法 B3LYP 汎関数、基底関数には 6-31G(d) および 6-31G(d)+ を用いた。

結果は入出力ファイルとして提示しているため、結果の再現および再利用は容易である。現在の計算リソースでは 1 日 1000 分子から 10000 分子の計算が可能である。

今後は化合物検索ができるような Web サイト構築、さらなる分子構造、および物性の提供 (振動構造および NMR ケミカルシフト熱力学的諸量、溶媒効果を含んだ計算、励起状態での構造最適化など)、それらからの化合物の特性の予想、実験の支援などを行えるようにしたい。

謝辞 理化学研究所情報基盤センターの黒川原佳さん、姫野龍太郎先生、インテル株式会社堀越将司さんには計算リソースの提供、助言、メンテナンスなどをしていただいた。またこの研究の一部は東京大学東京大学情報基盤センタースーパーコンピューティング部門の「若手・女性利用」制度を利用した。本研究の計算結果の一部は、RIKEN Integrated Cluster of Clusters (RICC) システムを利用して得られた。

参考文献

[1] P. A. M. Dirac, Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character, Vol. 123, No. 792 (6 April 1929).
[2] Q. V. Le, M. Ranzato, R. Monga, M. Devin, K Chen, G. S. Corrado, J. Dean, A. Y. Ng, “Building High-level Features Using Large Scale Unsupervised Learning”, ICML, 2012.
[3] A. Krizhevsky, I. Sutskever, G. H. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks”, Advances in Neural Information Processing Systems 25 (NIPS 2012).
[4] F. Seide, G. Li and D. Yu. Conversational Speech Transcription Using Context-Dependent Deep Neural Network, in INTERSPEECH, pp. 437-440 (2011).
[5] M. Rupp, A. Tkatchenko, K.-R. Müller, O. A. von Lilienfeld: Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning, Physical Review Letters, 108(5):058301, 2012, G. Montavon, M. Rupp, V. Gobre, A. Vazquez-Mayagoitia, K. Hansen, A. Tkatchenko, K.-R. Müller, O.A. von Lilienfeld, Machine Learning of Molecular Electronic Properties in Chemical Compound Space, New Journal of Physics, 2013, to appear.
[6] NIST Chemistry WebBook 入手先 ⟨http://webbook.nist.gov/chemistry/⟩ (2014.06.30).
[7] Blton E, Wang Y, Thiessen PA, Bryant SH. PubChem: Integrated Platform of Small Molecules and Biological Activities. Chapter 12 IN Annual Reports in Computational Chemistry, Volume 4, American Chemical Society, Washington, DC, 2008 Apr.
[8] Blum L. C.; Reymond J.-L. J. Am. Chem. Soc.,2009, 131 (25), 8732-8733.
[9] ChEMBL: a large-scale bioactivity database for drug discovery” Nucleic Acids Research. Volume 40. Issue D1. Pages D1100-D1107. 2011.
[10] The PubChemQC Project 入手先 ⟨http://pubchemqc.riken.jp/⟩ (2014.06.30).
[11] 入手先 ⟨ftp://ftp.ncbi.nlm.nih.gov/pubchem/Compound/CURRENT Full/SDF/⟩ (2014.06.30).
[12] PubChem Compound 入手先 ⟨https://www.ncbi.nlm.nih.gov/pccompound?term=all[filt]⟩ (2014.06.30).
[13] E. E. Bolton, J. Chen, S. Kim, L. Han, S. He, W. Shi, V. Simonyan, Y. Sun, P. A Thiessen, J. Wang, B. Yu, J. Zhang and S. H Bryant PubChem3D: a new resource for scientists, Journal of Cheminformatics (2011), 3:32.
[14] NIST Computational Chemistry Comparison and Benchmark Database NIST Standard Reference Database Number 101 Release 16a, August 2013, Editor: Russell D. Johnson III 入手先 ⟨http://cccbdb.nist.gov/⟩ (2014.06.30).
[15] Anderson, E.; Veith, G. D.; Weininger, D. (1987). SMILES: A line notation and computerized interpreter for chemical structures. Duluth, MN: U.S. EPA, Environmental Research Laboratory-Duluth. Report No. EPA/600/M-87/021.
[16] Noel M O ’ Boyle, Towards a Universal SMILES representation - A standard method to generate canonical SMILES based on the InChI, Journal of Cheminformatics 2012, 4:22.
[17] N. M O’Boyle, M. Banck, C A James, C. Morley, T. Vandermeersch, and G R Hutchison, Open Babel: An open chemical toolbox, J. Cheminf. 2011, 3:33.
[18] “Advances in electronic structure theory: GAMESS a decade later” M.S.Gordon, M.W.Schmidt pp. 1167-1189, in “Theory and Applications of Computational Chemistry: the first forty years” C.E.Dykstra, G.Frenking, K.S.Kim, G.E.Scuseria (editors), Elsevier, Amsterdam, 2005.
[19] Alex A. Granovsky, Firefly version 8.0, 入手先 ⟨http://classic.chem.msu.su/gran/firefly/index.html⟩ (2014.06.30).
[20] Stewart, James J. P. (1989). “Optimization of parameters for semiempirical methods I. Method”. J. Comput. Chem. 10 (2): 209.