課題名（タイトル）：

NMR メタボロミクス用理論化学シフトデータベースの構築

利用者氏名：○近山英輔*、伊藤研悟*
所属：*環境資源科学研究センター統合メタボロミクス研究グループ環境代謝分析研究チーム

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

NMR メタボロミクスは、NMR スペクトルを解析して、多数の代謝化合物を網羅的に検出する。そこではプロファイリング、定量、そして化学シフトデータベースを用いた、複数物質の組成の一斉アノテートを行う。NMR メタボロミクスの適用範囲は広く、実際に、微生物、酵母、昆虫、動物、哺乳類、植物、ヒト、病気診断、農産物、海藻、食品、医薬品、バイオマス、環境の評価など様々な研究に応用されている。化合物を検出する際に重要になる化学シフトデータベースに関しては、潜在的に数十万種ある膨大な代謝化合物の中で、標準データを計測できる割合はわずかである。残りのほとんどを量子化学計算または情報科学的な計算による理論化学シフトで補完することにより、NMR メタボロミクス用化学シフトデータベースを構成できるが、現在化学シフトの理論計算値の精度は、混合物の NMR 計測スペクトルの中で異なる代謝化合物を区別できるほど十分ではない。また、分子の配座、あるいはその集合である構造アンサンブルが観測される実験値化学シフトに影響を与えるが、その影響が、個々の研究に必要な理論化学シフト計算とどれほど関連しているかはあまり自明ではない。

環境資源科学研究センター環境代謝分析研究チームでは、NMR で世界最高度のメタボロミクス解析プラットフォームを構築してきた。そこでは代謝物の標準化合物による化学シフトデータベースが重要な要素となっており、現在 800 代謝物以上の実験的計測データが蓄積されている。しかし、代謝物総体は例えば植物で 20 万種以上あると言われており、その全ての化合物を入手することはできない。また、例え入手できたとしても、その全てを NMR で計測することは不可能である。従って、計算機を用いて大量の代謝化合物についてその化学シフト予測値を計算し、データベース化することは、予測値の精度を犠牲にしても、未知試料の候補代謝物同定時に非常に有用となる。RICC には Gaussian が利用可能となっており、Gaussian では NMR 化学シフトの理論値が複数の量子化学計算アルゴリズムで計算可能となっていることから、本課題の最終目的として、量子化学計算による大量の化合物の化学シフトデータベース構築を設定している。

量子化学計算による理論化学シフトと分子構造の関係性は、分子の構造が緻密に反映される利点により、構造推定などの研究にも用いられてきたが、大量の代謝化合物に対する第一原理計算による理論化学シフトデータベースの構築をする際には、これが仇になり、大量の代謝化合物の 1 個 1 個についてそれぞれどのような立体配座を選択するかで、化学シフトの計算精度に大きな影響が出る。このことは、単一分子内のみを対象にした数々の理論化学シフト計算の研究で示唆されている。また、様々な配座によるゆらぎが、遮蔽定数に影響を与えた結果についての分子動力学やモンテカルロ法を用いた調査もある。グルコースを対象にしたそのような研究では、最も化学シフトに重要なものはジオメトリかもしれないと示された。また別の C70 を対象にした研究では、13C の計算に対し、多くの理論レベルについて計算している。分子動力学の結果も含め、ジオメトリの最も大きな化学シフトへの影響を結論している。分子力場計算法やモンテカルロ法による配座検索が構造アンサンブルを考慮した理論化学シフトの精密計算に寄与できるという報告もあり、このように、個々の分子の詳細な研究の結果からも、構造アンサンブルの影響下で大量の代謝化合物にシステマティックに適用する場合の配慮が示唆されている。さらに、基本的な HF と 6-31G、DFT、2 次摂動論である MP2、aug-cc-pVDZ 基底などのどの理論レベルを使用すべきか、構造は 1 点計算か、構造アンサンブルにすべきか、溶媒を入れるべきか、実験値との補正法、などの問題に対する大量の分子を対象にした一斉計算に対する統一的な回答はまだ与えられていない。その中で最も重要な項目は、構造アンサンブルの影響を考慮しながら、精度と計算時間の妥協点を探し、大量分子に一斉適用できるシステマティックな理論的化学シフト計算手法を見出すことである。このような、理論化学シフトの

研究は、応用上の波及効果もある。例えば、環境代謝分析研究チームで研究対象としているバイオマス研究に応用可能である。海藻類バイオマスの産業的価値は極めて高いと考えられるが、その効能について評価されていないものは多数存在し、海藻類の構成成分を網羅的かつ高感度に検出可能な計測技術高度化が望まれている。このような計測技術高度化において、安定同位体標識技術を利用した NMR 計測により網羅的な検出が可能であるが、実験化学シフトデータベースや既往論文に報告のない海藻類の未知成分を帰属することは容易ではなく、従来の帰属手法に頼らない理論化学シフト計算に基づく帰属方法の発展が期待されている。そこで我々は、海藻類の 13C 標識サンプルを用いて、多様な NMR 計測法により代謝混合物群の網羅的な分析を行うと共に、それらの結合情報に対してグラフ理論に基づいた解析を行い、また、量子化学計算により推定構造の妥当性も検証した。

２．　具体的な利用内容、計算方法

昨年度までに 29 種の代謝化合物構造に対し、Gaussian を用いた理論化学シフト計算と GROMACS を用いた分子動力学（MD）計算によるシステマティックな評価を終了している。今年度は、代謝化合物をさらに増加させる手法を検討し、さらに 29 種の上記結果において、構造アンサンブルに対する主成分分析と構造アンサンブルー化学シフトの相互相関解析を検討した。29 種化合物のシステマティックな評価結果については、論文にまとめている最中である。また、昨年度と同様に、理論化学シフト計算の応用として海藻類の未知成分の解析手法の研究も発展させた。

具体的に、MD 計算と量子化学計算の手順を復習しておくと、RICC の計算リソースを用いて、各々の化合物で 10 ns の古典定温定積 MD 計算により生成された 100 構造に対し、HF/B3LYP(DFT)/MP2 のメソッドと 6-31G/aug-cc-pVDZ の基底セットによる理論レベルを用いた第一原理量子化学計算により遮蔽定数を求めている。各化合物は、化合物ファイルで与えられている初期座標から、エネルギー最小化を真空中で行い、1 ns の平衡化の後、10 ns の MD シミュレーションにかけられた。MD に用いた力場は PubChem にて取得した sdf ファイルを Open Babel を用いて Mol2 形式に変換し，acpype を用いて GROMACS 用 GAFF 力場を生成した。MD は倍精度の GROMACS 4.0.5 で修正 Wolf 法を静電力計算に用い、真空中、298.15 K の定温シミュレーションを Nose-Hoover 法で行った。100 ps 毎に各化合物の分子構造をサンプルし、各化合物につき計 100 構造を得ている。MD で得られた各構造の座標についての遮蔽定数の第一原理計算は、RICC 上の Gaussian09 で HF、B3LYP、MP2 の 3 種の理論レベルと 6-31G、aug-cc-pVDZ の 2 種の基底セットについて計算している。故に 1 理論レベル、1 基底セットにつき 100 構造の各化合物の分子内原子の遮蔽定数を得ている。計 2900 構造である。29 化合物は、in-house で作成している化学シフトデータベースの中から、分子量 200 以下の化合物を全て選択している。今年度は、実験値化学シフトの帰属を CMC-assist ソフトウェアで確認する追加作業などを行った。また、理論化学シフトの ANOVA 解析も行った。

海藻類の未知成分の帰属法の研究については、方法論の確からしさを検証するにあたり、混合試料の既知の 12 成分および標準物質を対象に Gaussian を使用して構造最適化および遮蔽定数・スピン結合定数を算出し、実測値との比較を行った。この時の計算レベルは B3LYP/6-311++G**であり、真空中および SCRF 理論に基づいた溶媒効果を取り入れて計算を行った。しかし、連続誘電体近似では、溶媒を分子として取り扱っていないため、溶媒和構造が得られない。実際に溶媒分子を考慮して計算することにより精度が向上すると考え、量子力学（QM）/分子力場計算（MM）法を用いて計算を行った。また、水溶媒中では、分子がイオン化していることが考えられるため、イオン化を考慮した計算も行った。QM/MM 法およびイオン化を考慮した計算には混合試料中のギ酸を対象に行った。QM/MM 法では、溶質を B3LYP/6-311++G**レベルで、溶媒（水分子 20 個）を UFF 力場で計算を行った。また、安定構造の存在確率による化学シフトの補正を行うためエネルギー計算を行った。イオン化していないギ酸は安定構造が 2 つしか存在しないと考えられたため MD による計算は行わず、2 つの安定構造についてエネルギー計算を行い、ボルツマン分布により存在確率を算出後、加重平均により化学シフトを補正した。

３． 結果

29 種代謝化合物の理論化学シフトの評価では、真空中の 10 ns の定温定積分子動力学の結果から連続的に時間的に 100 ps 毎に等間隔でサンプリングした 101 個の配座を用いている。それぞれの配座に対し、6 種の理論レベルによる第一原理量子化学計算を適用し、分子内原子の遮蔽定数を求め、化学シフトへの検量線（Table 1、実際に使用した値よりも大きく有効桁数を落としてある。6-31G 基底の結果のみ表示）により、に変換したものである。各々の代謝化合物における理論化学シフトの時間変化を見ると、分子内の原子間力に従って決定論的に運動しているにもかかわらず、原子間の非線形相互作用により、単純な周期運動を行っていないように見える。例えば Ethanol のある水素原子では 4000 ps 付近にある鋭いピーク、5-(2-Hydroxyethyl)-4-methylthiazole のある炭素原子では 800, 3000, 4200 ps 付近にある鋭いピークなどはその影響であろう。当然、6 種の理論レベルのどれにおいても立体構造との相関が高く、グラフの形が一致していた。変換前の遮蔽定数のグラフ上下を反転し、検量線の係数をかけてオフセットを与えると変換後の化学シフトを得る。変換前の遮蔽定数では理論レベル間でオフセットがずれるという現象が起こるが、化学シフトではそれが解消される。5-(2-Hydroxyethyl)-4-methylthiazole のある炭素原子の遮蔽定数では、化学シフトに変換するときの MP2 のオフセットの差異が他の原子より大きかった。

Table 1 検量線

	^{1}H (ppm)	r_{H}	^{13}C (ppm)	r_{C}
RHF	-0.9*x*+29	-0.98	-0.8*x*+188	-0.99
B3LYP	-1.0*x*+33	-0.98	-1.0*x* +200	-0.99
MP2	-1.0*x*+34	-0.98	-1.2*x* +234	-0.99

今回の 1H の理論化学シフト計算の 1 例として、理論レベルにほとんど依存せずに良い精度を達成した、エチレングリコール分子の 6-31G 基底／RHF での結果を示す（Table 2）。実験値と計算値（Average、構造アンサンブルの平均値）との差は 4 つの水素原子に対し、それぞれ-0.04, -0.12, 0.09, 0.05 ppm であり、高精度を達成している。しかしながら、きわめて精度の悪い

Table 2 エチレングリコールの 1H 理論化学シフト

Atom No.	Exper-imental (ppm)	Computed (ppm)			
		Average	Deviation	Maximum	Minimum
1	3.65	3.61	0.71	5.58	2.11
2	3.65	3.53	0.63	5.17	2.37
3	3.65	3.56	0.61	5.00	2.13
4	3.65	3.70	0.65	5.18	1.93

例として、シトシンの例も述べる（Table 3）。これらも 6-31G 基底／RHF での結果であるが、シトシンの場合は理論レベルが上がるにつれ、わずかな改善が見られたが、本質的ではなく、他の多くの化合物に比べ異質の誤差がある。誤差は、-1.04, 1.68 ppm であった。この異質の誤差は、窒素原子を含む複素芳香環で、隣り合わせの H-C-C-H に特有の現象の可能性が考えられた。

Table 3 シトシンの 1H 理論化学シフト

Atom No.	Exper-imental (ppm)	Computed (ppm)			
		Average	Deviation	Maximum	Minimum
1	5.98	4.94	0.49	6.06	3.67
2	7.48	9.16	0.43	10.62	7.92

今回、29 種代謝化合物の基本評価終了を受け、大量の代謝化合物解析へ向けたスクリプト開発を試験的に行った。その結果、PubChem で sdf ファイルのテキストデータが連続的に連なった 1 ファイルを取得できることが分かった。459 化合物分の 1 ファイルを取得し、それを Linux コマンドの csplit コマンドで 1 化合物毎の sdf ファイルへ分割した。これを RICC、および Linux マシン上で acpype にて GROMACS 計算用のセットアップスクリプトを開発し、実行した結果、結果に食い違いが生ずることがわかった。Linux では、459 個のうち 355 個の分子で成功した。失敗の主原因は原子間の衝突による力場の不整合のようであり、問題点が明確化した。

構造アンサンブルの主成分分析の結果では、Altis らの 2 面角主成分分析という手法、及びデカルト座標の手法を用い、その結果、両者はおおよそ整合関係にあるようであり、サンプリング空間の可視化に成功した。

我々の 29 種化合物評価の結果は、これまでに行われてきた 1 分子についての研究結果とほぼ同じ方向を示

している。我々は、DFT の精度よりも大きな全体的な広範囲についての、構造ゆらぎと理論レベルの差異に注目して検討を行ったが、Bagno らは、グルコースに注目し、特に DFT の精度に近い 1H で 3.3-3.8 ppm の 0.5 ppm 程度の小さな領域、の精密化についても詳細な検討をおこなっている。Bagno らはグルコースで最も化学シフト計算の重要なのものはジオメトリかもしれないと言っており、我々の結果は完全に同意する。また Kaminsky らは、C70 の 13C の計算に対し、多くの理論レベルについて計算している。彼らは MD も評価しており、最も大きな化学シフトへの影響を結論している。今回我々の結果でも、構造アンサンブルに対し、一般に注意する必要があることが分かった。Bagno らが示したように、より精密化が必要な場合は、量子化学計算で構造最適化をしてから化学シフト計算するとよいだろう。また、彼らは溶媒の効果を入れることが分子内結合などのジオメトリ情報が改善できることを示している。Kaminsky らの行ったように理論レベルや溶媒の効果を詳細に検討することも精度の向上に有効である。また MD による十分な構造サンプリングが期待できない場合、MM でのサーチやモンテカルロ法での構造サーチが改良に期待できる。

海藻類の未知成分の帰属法の研究については、算出されたギ酸の 1H および 13C の化学シフトを実測値と比較した結果、SCRF 理論に基づいた溶媒効果を取り入れた計算が良い近似解を得られた。真空中および QM/MM 法では計算結果があまり変わらなかった。イオン化を考慮したものとしていないものでは、化学シフトが大きく変化する結果となった。今回はほとんどがギ酸を用いての計算方法の選定であったため、他の分子での誤差が分からなかった。多くの分子の結果を得ることで回帰分析、重回帰分析、パターン認識によるスケーリングファクターを算出することで、より良い近似解が得られることが予想される。また、今回の計算に用いた標準物質は TMS であったが、実際には DSS を用いているため、これによる誤差要因も考えられた。DSS の遮蔽定数を計算する必要がある。混合試料中における他分子においても計算を行い、誤差の範囲が大きい場合には、複数の分子について同様の計算を行い、統計解析によりスケーリングファクターを算出する必要と、最終目標としては、海藻成分の未知シグナルの理論的帰属を試みることが求められる

４．　まとめ

ほとんどの 1H と 13C への配座のゆらぎは取り得る範囲に対して相当の大きさの範囲であり、ほとんどの 1H とかなりの 13C に対し、メソッドや基底セットによる差異よりも非常に大きいため、構造を平均化することで、計算値は精密化できるかもしれない。1H では、基底セットの改良効果が見られたが、13C ではあまり変わらなかった。一点計算での高精度理論化学シフトは構造の選択を慎重にすべきであり、今回の計算では、計算量が軽量な DFT の代わりに計算量の重い MP2 を使用する利点はほとんど見られなかった。昨年度までの 29 種代謝化合物の結果を論文にまとめていくにあたり、構造アンサンブルの詳しい解析を追加した。

５．　今後の計画・展望

今後の展望としては、今回の結果での未調査点である。今回の計算では、構造揺らぎと第一原理計算の理論レベルに注目してそこから得られる結論を得た。他の様々な影響は無視している中では、水分子の影響がある。しかし最近の研究ではこれらは構造の影響に比較すると小さいと評価されているが、Brancato らの研究では、170 に対して大きなずれを観測している。我々の結果では、Cytosine の 11H や 3C、Indole 9C、Coniferyl aldehyde の 10C など、水和の直接影響が少なそうな箇所で実験値との大きなずれが観測されたことは水影響が間接的であることを示唆するかもしれない。量子の MD からの平均構造に関して、エネルギー最小構造との間の差として化学シフトの影響が観測されている。その他では、全部の原子を見ていない、29 化合物しか見ていない、MD の時間などの不明点があるが、これも含み、様々な影響について広範囲にシステマティックに十分に精査された研究はまだない。これらの影響は当然あるだろうが、しかし今回の結論は、平均として、覆らないだろう。

６．　利用がなかった場合の理由

利用あり