プロジェクト名：　データ中心ケミストリ

プロジェクトディレクター：　佐藤 寛子　准教授（国立情報学研究所）

〔1〕 研究計画・研究内容について

(1) 目的・目標

現在までに存在が確認されている化学物質は約 7000 万種類であり、年間数十万～百万種オーダーで増え続けている。しかし、理論的には存在しうるが、いまだ人類が手にしていない化学物質種の数はこれを遥かに凌駕することが明らかとされつつある。データ中心ケミストリプロジェクトでは、この革新をもたらす「埋蔵分子」を理論的に探索・発掘し、これらを供給する化学反応経路と、エネルギーや電子状態に関する物理化学パラメータとともにデータベース化し、研究・教育機関や一般から利用可能なウェブシステムを構築することを目的とする。データケミストリの視点から、自己発展型の異分野融合研究資源を創出し、物質科学に関する種々の問題解決のための情報資源を整備したい。化学物質が反応経路によってネットワーク状に複雑に連結されたデータから有用な知識を導きだすための基盤技術を開発し、科学の諸問題の解決のためのデータケミストリの新展開を目指す。さらに、生命科学データベースとの統合による､ケミカル–ライフサイエンスの協調的発展の新しい方法論を提供することも視野に入れる。

(2) 必要性・重要性（緊急性）

化学物質の情報は、アメリカ化学会の一部局：ケミカルアブストラクト（CAS: Chemical Abstracts Service）によって網羅的に編纂されている。現在までに知られている化学物質は約 7000 万種類であり、年間数十万～百万種のオーダーで増え続けている。化学物質の文献や分子構造・分光学・物性データや、供給方法は、化学物質を取扱う幅広い産業や学術分野にとって不可欠な情報である。

化学物質の可能性と化学反応性は、量子力学の理論を化学に応用した量子化学によって原理的に解く（予測する）ことが可能である。化学物質とその供給方法は、ポテンシャル曲面上の極小点（反応物・生成物）と鞍点（遷移状態）を結ぶ化学反応経路を探索することによって得られる。これを自動的に探索することは不可能であると考えられていたが、2004 年に大野公一（東北大学名誉教授・量子化学探索研究所長、本事業の共同研究者）・前田理（当時東北大学、現北海道大学）らにより発表された画期的な手法により、従来の常識が破られた。本手法は、GRRM（Global Reaction Route Map）法とよばれ、ポテンシャル曲面を超球面探索法により網羅的に探索する。現在までに種々の化学物質に適用され、従来知られていた化学物質や反応経路を遥かに凌駕する可能性が発掘されることが明らかとなってきた。

本プロジェクトでは、これらの革新をもたらす「埋蔵分子」を発掘・探索し、研究情報基盤として提供するとともに、化学物質ビッグデータの科学の諸問題の解決のための手法を開発する。このような理論化学に立脚した化学物質のビッグデータは、まだ世界のどこにも存在していない。化学物質ビッグデータを利用したデータケミストリによる科学の諸問題解決への取組みも未踏の領域である。今後重要性が増すと期待されるマテリアルズインフォマティクスのデータ基盤となる重要な柱の 1 つであり、データ創出アルゴリズム（GRRM）とともに、日本の科学・技術により、日本が主導となり、世界に先駆けて緊急に本研究基盤を整備することが重要である。

(3) 期待される成果等（学問的効果、社会的効果、改善効果等）

ケミカルアブストラクト（CAS）により編纂されている化学物質データは文献を情報ソースとしたものであり､対象とされているのは天然物質からの抽出や人工合成によって存在が確認された物質である。

これに対して、本プロジェクトで発掘する化学物質データには、これらの既知物質に加えて、理論的には存在しうるが、まだ実験的に存在が確認されていない新規物質、すなわち「埋蔵分子」も多く含まれ、機能性物質の創成に結びつく大きなポテンシャルを秘めている。その応用範囲は、医薬品、農薬、食品、新エネルギー資源を含む、新規機能性物質とその供給方法について研究する幅広い分野に渡ると想定される。実際に、本化学物質データのデータベース化への製薬企業等の産業界からの期待は大きく、学術的にも高い関心が寄せられている。また、ライフサイエンスデータと連携することで、代謝機構などの生体内化学反応の理論的解明に大きく寄与することが期待される。加えて、本システムでは、分子構造が多段階の化学反応ステップを経て変化する様子をムービーで観察することができる。これらは教育・一般からウェブを介してアクセスできるものとする計画であり、中学・高等学校における理科教育や大学での教養・専門教育に貢献できるものとなると期待される。

(4) 独創性・新規性等

本プロジェクトで開発する化学物質データベースは、従来技術と比較して、埋蔵分子の発掘方法、蓄積・供給方法、科学研究への活用のいずれにおいても高い新規性をもつ。 従来の文献に基づくデータベースは科学研究にとって不可欠なデータソースであるが、実験によって観測された既知反応が手入力により編集されるため、新規性は低く、入力ミス等によるデータ精度の低下が起こりやすい。また量子化学の理論に基づく化学反応経路データベースの研究は国内外でわずかにみられるが、いずれもごく小規模であり、また、これらのデータソースを統計解析やデータマイニングなどの数理・情報的な手法により科学研究に活用する発想も見られない。一般に、量子化学の理論にもとづく化学反応経路探索は試行錯誤により実施され、反応物から生成物への単一の経路を探索するだけでも数ヶ月～数年かかり、時には見つからないこともあるが本データベースに蓄積された類似反応を直接またはデータマイニングを通じて利用することで、数時間～数日で確実に反応経路を見つけることが可能となると期待される。さらに、従来にない画期的な特徴として、本データベースでは別の化学反応が起こる可能性も網羅されているので、副生成物を最小限にする最適化など、化学合成研究や工業化学における化学合成プロセス設計を合理的かつ効率的に行い、化学製品の品質を向上させることが可能となる。

(5) これまでの取り組み内容の概要及び実績

本プロジェクトに関する研究として、以下の準備研究を行ってきている。

まず、(I)データベース設計、(II)インタラクティビティ向上のための処理系の整備、(III)ウェブコミュニティ環境の開発を目的としたユーザのプライマリ利用シーンの抽出とインタラクションフローの構築、(IV)データ登録・管理システムの設計と、化学反応経路探索を実施する GRRM (Global Reaction Route Map) プログラムの最新版との適合化を行った。さらに、探索された化学反応経路データを可視化するソフトウェアとこれと連動する検索アルゴリズムの開発と実装に着手した。GRRM 開発者らの並列分散処理型の GRRM との連携方法について議論も継続して実施している。システムの要は、大規模データの効率的な検索・可視化と、データ発掘（化学反応経路探索）の効率化にあるが、前者については、上記(I),(II),(III)により、必要な要素技術の抽出と今後の開発に向けた準備を行うことができた。また、後者については、上記(IV)により、密接に連携して進めることを確認し、今後の開発の見通しを得ている。

なお、本事業と関連した研究課題に対して、以下の研究費の支援を得ている。

- 国立情報学研究所公募型共同研究費（平成 21～25 年度）
- 住友財団基礎科学研究助成（平成 23 年 11 月～24 年 10 月）
- 平成 24 年度情報システム研究機構長裁量経費・データ中心科学リサーチコモンズ基盤整備に

向けた先導的研究・事業（平成 24 年度）
- 科学研究費補助金挑戦的萌芽研究（平成 25～27 年度）

(6) 国内外における関連分野の学術研究の動向

ケモインフォマティクス、データケミストリの分野では、統計処理を行うための十分な量のデータを得ることが求められることから、実験データやパラメータを用いた経験的手法により計算されるデータを用いることが現在でも主流である。一方で、計算機の高速化と量子化学計算手法の発展により、量子化学にもとづくより精度の高いデータを用いたデータマイニングを分子設計や合成設計に利用する分野に高い関心がもたれ始めている。国内では、ケモインフォマティクスに加えて、マテリアルズインフォマティクスといった新語も聞かれるようになり、量子化学を専門とする研究者がインフォマティクスに参画しようとする動きが見られる。欧州では、COST（European Cooperation in Science and Technology）の共同プロジェクト等で、種々の量子化学計算ソフトウェアを統合したデータフローを開発し量子化学計算データを蓄積・活用しようとする動きが見られる。量子化学の理論に基づく化学反応経路データベースの研究は国内外でわずかにみられるが（ミネソタ大学・D.G. Truhlar らのライブラリ、山口大学・堀憲次らの TSDB、台湾国立中正大学・W.P. Hu らのデータベース等）、いずれもごく小規模である。また、これらのデータソースを統計解析やデータマイニングなどの数理・情報的な手法により科学研究に活用する報告も見られず、量子化学に基づく大規模な化学反応経路データ基盤に関する研究は、本プロジェクトを除いて現段階で皆無であると言ってよい。

〔2〕 研究計画（平成 25～30 年度の研究計画を記入）

(1) 全体計画

現在までに存在が確認されている化学物質は約 7000 万種であり、年間数十万～百万種オーダーで増え続けている。しかし、理論的には存在しうるが、いまだ人類が手にしていない化学物質種の数はこれを遥かに凌駕することが明らかなっている。本事業では、この革新をもたらす「埋蔵分子」を理論的に探索・発掘し、これらを供給する化学反応経路を、分子のポテンシャルエネルギーや電子状態に関する物理化学的パラメータとともにデータベース化し、研究・教育機関・一般から利用できるウェブシステムを構築する。データケミストリの視点から、自己発展型の異分野研究資源を創出し、専門的な研究利用から教育機関、一般まで、多様な形で幅広く共有できるウェブコミュニティを形成する。化学物質が反応経路によってネットワーク状に複雑に連結されたデータから有用な知識を導き出す基盤技術を開発する。生命科学データベースとの統合化による、ケミカルーライフサイエンスの協調的発展の新しい方法論を提供する。情報学、統計科学、生命情報学、物理化学から開発チームを形成し研究を効果的に推進する。

(2) 各年度の計画

期間内に、化学物質データの探索・蓄積を自動的に行うデータベースシステムを確立し、研究者から一般まで段階的に公開する。事業計画期間後は、データ創出を行う共同プロジェクトの枠を国内外に拡大し、大規模に化学物質データの探索を行い、研究基盤として供給・運用を行う。

平成 25 年度

平成 25 年度は、平成 24 年度に「平成 24 年度機構長裁量経費　データ中心リサーチコモンズ基盤整備に向けた先導的研究・事業」の支援を受けて実施したシステム設計に沿って、専門的な研究者が利用する際に必要となる可視化機能の詳細について抽出し、実装を行う。化学反応経路探索を行う GRRM

プログラムでは、量子化学に基づく分子ポテンシャルエネルギー曲面に沿って可能な化学反応経路が網羅的に探索される。ここから得られる反応経路は複数の安定構造とこれらの間に位置する最もエネルギーの高い状態（遷移状態）を結ぶネットワーク構造をもつ（各ノードは安定構造と遷移構造，および解離チャネル)。網羅的に探索が行われることで、同一の組成式について、大量のノードから構成される化学反応経路ネットワークが得られる。例えば、C_6H_6 に対しては、安定構造だけでも 5000 種を超えるノードが生成される（計算継続中)。この大規模な化学反応経路ネットワークから、解析を行う研究者が必要な情報を取り出すことができるインタラクティブな可視化ソフトウェアを開発することがポイントである。

平成 26 年度

年度の初頭に昨年度までに開発した研究者向けの可視化ソフトウェアを、まずはユーザからのフィードバックを得ることを目的として、一般に無償公開する。本ソフトウェアの開発言語としては Smalltalk を用いているが、平成 26 年度は、これを Java 等のより汎用的な開発言語へ移植する。これと並行して、可視化システムへの新規機能の実装と、検索アルゴリズムの開発と実装を行う。これらをまとめてパッケージ化し、平成 26 年度末を目処にオープンソースとして公開する。

平成 27 年度

データベースシステムの開発を中心に進める。ユーザが個別にデータを登録する機能に加えて、分散型の化学反応経路探索プログラム（GRRM プログラム。連携研究機関で開発されるもの）を本システムと連携させ、化学物質データ探索–登録–管理を自動的に行うことのできる、自己発展型のデータベースの開発を開始する。一方で、化学物質データ探索より得られる化学反応経路ネットワークデータを、データマイニングや統計解析、モデリング手法により分類・解析し、新規物質の発見や物性予測などに繋がる新規データケミストリ手法の開発を開始する。

平成 28 年度

可視化ソフトウェアとデータベースを統合したウェブシステムを開発する。研究者向けに加えて、教育・一般からの利用も想定したデータ検索・インタフェース機能の開発を開始する。研究者向けのデータ登録・検索インタフェース機能の拡充を行うとともに、自己発展型のデータベースの開発を継続して実施する。平成 29 年度より、自己発展型データベースシステムを利用して、大規模な化学物質データ探索の実施を開始する。データケミストリ手法の開発を継続して実施し、平成 30 年度から本事業終了後にかけて、実際の科学の問題への応用へと発展させる。

平成 29 年度

自己発展型データベースシステムを利用して、大規模な化学物質データ探索の実施を開始する。システム全体の機能拡充を進めながら、データケミストリ手法の開発を継続して実施する。

平成 30 年度

教育・一般利用も含めたウェブシステムを公開する。平成 30 年度から本事業終了後にかけて、実際の科学の問題への応用へと発展させる。公開されるウェブシステム・データベースの運用・保守は、科学技術振興機構等の公的機関に委託し、研究成果の社会への還元と、継続利用を実施していけるようにしたいと考えている。

〔3〕 研究推進・実施体制

・研究代表者

〔国立情報学研究所〕　　佐藤寛子

・共同研究者

〔国立情報学研究所〕　　宇野毅明

〔統計数理研究所〕　　吉田　亮、中野純司
〔東北大学、量子化学探索研究所、国立情報学研究所〕　　大野公一
〔京都大学〕　　中小路久美代
〔東京大学〕　　有田正規、岩田　覚
〔京都産業大学〕　　青木　淳
〔スイス連邦工科大学〕　　ハンス・ペーター　ルーティー

〔4〕 研究の進捗状況

本プロジェクトは平成 26 年度よりリサーチコモンズ事業に正式参画するが、平成 25 年度はこれに先立った種々の準備研究を実施し、本格的な参画に備えることを目的とした。具体的には理論的に探索される化学反応経路ネットワークデータを可視化する機能を開発することを目指した。本可視化ソフトウェアは、化学反応経路探索プログラムから出力されるテキストタイプの化学反応経路ネットワークデータを可視化するものである。探索結果を研究者が専門的な観点から解析できるインタラクションデザインと実装を行う。可視化にあたっては，検索アルゴリズムとの連携も鑑みながら、インタラクションデザインの観点から可視化ソフトウェアの設計を行う。また、上記の他の共同研究者とも適宜議論を行いながらソフトウェアの開発を進める。実装は作業委託により効率的に実施する。以上を目的として、研究を実施した。

〔5〕 研究成果物

① 知見・成果物・知的財産権等

② 成果発表等
＜論文発表＞
〔学術論文〕
〔データベース〕
〔著書等〕
〔解説・総説〕
〔その他〕
＜会議発表等＞
〔招待講演〕
◎国際会議

1. 佐藤寛子，“QM-Based Data Chemistry: Chemoinformatics Meets Quantum Chemistry” Asian International Symposium, 94th Annual Meeting of The Chemical Society of Japan，名古屋, 2014 年 3 月 29 日.

〔一般講演〕
◎国内会議

1. 佐藤寛子，Stefano Borini，小田朋宏，中小路久美代，大野公一，”理論化学―データケミストリー：超球面探索法より得られる化学反応経路データの蓄積と活用“，第 7 回分子科学討論会，京都，2013 年 9 月 24 日.
2. 佐藤寛子，小田朋宏，中小路久美代，大野公一，”化学反応経路の全面探索の可視化とデータマイニングによる発見への取組み“，化学反応経路探索のニューフロンティア 2013，京都，2013 年 9 月 28 日.
3. 佐藤寛子，小田朋宏，中小路久美代，宇野毅明，田中宏明，岩田覚，大野公一，“「埋蔵分子」発掘プロジ

ェクト : 化学反応経路マップのインタラクティブ可視化に向けて", インタラクション 2014, 東京, 2014 年 2 月 28 日.

＜受　賞＞

③　その他の成果発表