様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　３０　年　　６　月　２８　日現在

機関番号：　８２６２６
研究種目：　若手研究(B)
研究期間：　2013～2017
課題番号：　２５７３０１５４

研究課題名（和文）機械学習を利用した反応材料分布と環境エネルギー条件の推 定法構築

研究課題名（英文）Estimation of material distributions and energy conditions for chemical reactions using machine learning

研究代表者

城　真範（Shiro, Masanori）

国立研究開発法人産業技術総合研究所・情報・人間工学領域・研究グループ付

研究者番号：９０３５７２４４

交付決定額（研究期間全体）：（直接経費）　　3,300,000円

研究成果の概要（和文）：化学反応のシミュレーションには通常量子力学の計算が必要で、低分子であっても、多くの計算リソースを必要とする。本研究では、これを解決するため、関係する分子の量とエネルギーを統計分布で表現し、モンテカルロ法でシミュレーションする方法を提案、一部実装した。提案方法は粒子の状態を分布で代用するため、分子に特異的な構造に依存するような高分子の反応には向かないが、モンテカルロ法であるため計算機の能力に応じた正確さの結果を得ることができる。このため一般のパソコンで粗い結果を得てから大型計算機で正確な結果を効率よく探索可能である。
プログラムは広く国民に無償利用していただけるよう公開準備をしているところである。

研究成果の概要（英文）：Usually, quantum-mechanical calculations with heavy resources are necessary for the simulation of the chemical reactions even if for a small molecule. In this study, we expressed quantity of molecules and each of the energy of molecules with statistical distributions, and proposed a method by Monte Carlo simlation as the substitute for the quantum-mechanical based. We implemented our method partly used standard C++ language. Our method is not suitable for the reactions depends on the molecular structures in many of high polymers, but can get the most suitable result by the adjusted accuracy of the computer ability. We are now preparing to be able to use the program widely.

研究分野：　非線形時系列解析、機械学習

キーワード：　化学反応　シミュレーション　モンテカルロ法　無機分子　逆問題　機械学習　低分子　生命発生

１．研究開始当初の背景

従来より、化学反応の計算機的手法は大きく分けて、量子力学を使うものと古典力学を使うものに分けられる。前者では分子の波動関数を適切な有限個の基底で表現し、固有値問題を解くことで化学反応をシミュレートし、後者は分子を剛体球あるいはその集合と見なして、剛体球同士をぶつけ、ある確率で原子同士の組み換えを起こすことで化学反応をシミュレートする。前者は高い信頼度で結果を提供する反面、原子個数に対して 3 乗程度のオーダの計算量が必要で、典型的にはピコ秒のオーダの計算を完了するために何日あるいは何週間もの計算が必要である。後者は近似の程度によって計算時間を現実的な量にできる反面、水溶液中など極性分子が関係する結果では特に信頼性に難があるといわれている。こうした現状下で、疎視化の方法が様々に研究され、代表的には平均場近似が挙げられる。これは反応に関係する数個の分子のみを正確に記述し、その他の分子の影響をポテンシャル場として扱うものである。固体物理や素粒子物理の領域において多数の顕著な成果が報告されているが、水溶液中の反応では極性や水素結合の影響を取り入れる必要があり、結果の妥当性には慎重な議論が必要である。

これらの計算を使って、実際にある目的の化合物をデザインしようとする場合、実際には候補となる原料物質と環境条件を様々に変えて多数のシミュレーションを繰り返す必要がある。原料物質が推定しにくい場合では探索空間が広がるため、極めて解きにくい問題となっていた。特に薬剤など複雑な化合物をシミュレーションによってデザインすることは難しく、もっぱら計算機の進歩に頼る大規模計算が解決方法の主流と見なされがちであった。

２．研究の目的

本研究の目的とするところは、反応後の生成物質を与えて反応前の原料物質を推定する逆問題を最初から念頭に置き、化学反応を伴う分子シミュレーションにおいて、量子力学的取扱いを回避し、計算量を劇的に低減することである。そのため、化学反応を独立な多数の段階に分割し、シミュレーションを何度も繰り返さなくても、有効な反応経路の概略を提供できる方法を提案・開発することである。

本研究では、反応前の原料物質を推定するために、化学反応のネットワークを確率分布のネットワークに置き換え、適切な仮定のもとで、ベイジアンネットワークの確率推論問題として機械学習の枠組みで処理することを想定する。すなわちネットワークの下流側（反応生成物側）から上流側（反応原料物側）に向かってベイズの定理を用いて逆に分布を推定してゆくことで化学反応の終端物質から原料物質に向かって逆シミュレーションを可能とする。このためには、出力結果は確率分布でなくてはいけなく、またそれらの分布を適切に設定するために、前提となる各段階のシミュレーションは可能な限り高速でなくてはいけない。

３．研究の方法

本研究の提案する方法は平均場近似の一種であり、反応に直接関与するする分子だけに着目することで計算量を減らし、さらに周囲の分子分布や分子の形状効果などもすべてノンパラメトリックな確率分布として扱うことで、極性や幾何構造による影響等を統一的に扱うものである。また関係する分子運動や外部刺激もすべて単一の環境エネルギー分布として扱うことで分子の反応特性を確率分布のみで記述する。確率分布自体を疎視化あるいは精緻化すれば、計算時間と精度を適切に調整することが可能である。

従来これに似た方法は Gillespie のアルゴリズムとして知られていた。遺伝子発現ネットワークにおけるタンパク質相互作用などに応用されているが、試験管の実験で平衡状態に達した化学反応から、結合と解離定数を速度定数として計測できることが前提となっている。本研究が想定する化学反応は高温・高圧環境下での未知の反応も含むため、速度定数の実測を前提とはできない。そこで、本研究では粗視化をより一層押し進め、速度定数をハイパーパラメータとして、それ自体を学習対象とすることとした。

研究は二段階に分かれる。第一段階は、反応の順方向を新しい方法でシミュレーションし、必要な着目分子が周囲のエネルギー分布と原料分布によってどのような物質に変化するかを確率分布の形で与えることである。本研究の手法では、常にある一群の分子に着目し（これを着目分子群と呼ぶ）、その周囲に存在する分子の分布（存在分布）、周囲のエネルギーの分布（エネルギー分布）、着目分子の内部にある有限個の反応点（化学反応に関係する原子の結合部分）にエネルギーを分配するための分布（分配分布）、各反応点が選択される分布（選択分布）の 4 つの確率分布が想定される。あるステップにおいて、存在分布に沿って反応に関係する分子を乱数により抽出し、分子内の反応点とその点に与えられるエネルギー値を分配分布と選択分布に従って抽出する。さらにエネルギー分布から、そのステップにおける反応エネルギーを乱数により抽出する。この反応エネルギーは例えば、ボルツマン分布として表現される周囲の分子の熱運動や、環境外からの電磁波（紫外線、可視光線、赤外線）、超音波による低周波刺激、リンの酸化-還元反応に伴う活性などをまとめて表現している。高速化のために計算は古典的に取扱い、抽出された反応エネルギーが、抽出された原子ペアの組み換えに必要なエネルギーよりも大きければ反応が実現し、小さければその反応は

起こらないとする。選択分布は最も単純なモデルでは一様分布であると仮定され、シミュレーションのステップ毎に、反応点から等確率で一つの反応点が選ばれる。さらに存在分布に従って反応相手の分子が選ばれ、さらに選ばれた分子の各反応点の中から選択分布によって一つの反応ポイントが選ばれる。反応ポイントの結合エネルギーと、サンプルとして得られたエネルギー値と反応点での分配分布の値の積を比べ、前者が小さければ結合の組み替えが起こるとし、着目している分子は分解や結合を伴って新しい物質に変化する。

同時に存在分布の中で新しい物質の割合が少しだけ増加する。そこでエルゴード性を仮定し、このような反応ステップを、周囲の物質分子の分布と着目分子の履歴から得られた分布が（ある範囲で）一致するまで繰り返す。一致すれば、一つの順方向シミュレーションを終了し、反応が平衡に達したとみなす。4 つの分布の様々な組み合わせにおいてこの順問題を解いておけば、それらを組み合わせることで、大規模な化学反応を高効率でシミュレーションできることになる。

より精緻化したモデルでは、水素結合、極性、触媒による影響を反映し、選択分布が一様ではなくなる。本研究のモデルにおいて触媒（酵素を含む）の効果は反応点の中から特定の部位だけにエネルギーを配分することに対応する。既知の知見とシミュレーションの結果を照合することで、選択分布、分配分布、エネルギー分布を調整し、実際の系に近い平衡状態を試行的に得る。このようにして様々な物質の反応について 4 つの確率分布の組を学習し、仮に分布の一部が未知の場合でも、ある範囲の精確さで平衡状態が推定できるようにする。

第二段階は確率分布同士のネットワークを構成し、いくつかの反応後物質を与えて確率推論によって逆問題を解けるようにする。しかしながらこれは将来的な課題を含んでおり、必ずしも本研究の主たるターゲットではない。

第二段階では第一段階で得られた個々の反応系の細部に踏み込まず、化学反応を単に確率分布同士の関係として解釈し、確率分布同士のネットワークを考える。例えばブドウ糖がある割合で存在することを最終状態とするならば、適当な段数の反応ネットワークを設定し、その各段に事前分布として適当な確率分布を設定し、最終状態のブドウ糖割合から各反応による確率分布の事後状態を逆推定してゆくことで、原料物質の分布と環境エネルギーを推定する。これは機械学習の枠組みにおいてはベイジアンネットワークの確率推論を実行することに他ならない。実際にはまずエタノール等、簡単で反応のよく知られた生成物質を与えて実際の原料物質の分布とエネルギー分布を推定する。さらに数種類の簡単なアミノ酸（アラニン、グリシンなど）、 ATP、ブドウ糖、脂質二重膜分子等で推定可能になるようシミュレータの改良を行う。これらの試験物質は溶液中の物性と、結晶解析で得られた物性とが異なる場合が多いためデータベースを参照して対応する必要がある。

４．研究成果

初年度は、主に文献調査と、関係する専門領域の研究者とディスカッションを多く行った。特に研究会等で積極的に発表し、広く興味を持ってくれた研究者との議論を行い、設計システムの問題点と改良方法、利用可能な既存技術、ライブラリなどを横断的に調べた。その結果、本提案システムは十分に実現可能であるものの、類似の先行研究が多数存在し、さらにその問題点を克服することで、新しい方向性の意義を提案できることが分かった。

具体的には、申請時の方法では反応に関係する分子が同時に 3 つ以上である場合（最も簡単な場合は水素の燃焼など）をシミュレートできなかったが、分子が遊離してイオンになる過程(反応の素過程)も一つの化学反応として解釈することで、これを克服できることが分かった。また、プロトタイプの作成により、着目分子以外をすべて平均場と見なすと、探索空間に対して反応履歴の時系列が不足し収束しにくいことも分かった。そこで、分子を種類と運動エネルギーで区分し保持する方法を考案した。もう一つの展開法として、着目を単一分子ではなく、ある程度の分子団として見ることで反応の結果分子が分解したり結合したりしても全体を一つの塊として見る方法を考えた。

二年目は、主に逆問題を解くためのプログラム開発に注力した。特に順問題よりも逆問題を解くことに力点をおいて、仮定された問題についてのいくつかのサブルーチン開発を行った。すなわち、世の中にすでにある化学反応のデータを使って、目的の原材料と環境を得ることに重点を置いた。しかしながら既知の化学実験データの開示が難しく、共同研究契約にかなりの時間を要することが分かった。このため引き続き、研究会等で積極的に発表し、広く興味を持ってくれた研究者との議論を行い、データ提供に結びつけることを模索した。さらに先行研究を実装し、その問題点を実際に洗い出した。研究の進展に伴い、申請時のアイデアからは変容し、反応の素過程に踏み込む方法を進めた。

三年目および最終年度については、引き続き、他機関を含めた幅広い研究者と積極的に議論し、実装とその改良を行った。反応時の計算負荷を減らすため、分子基単位の計算において新しい分子が生成されたとき、乖離しうる全パターンをメモリ上に持つという、やや強引な設計とならざるえなく、パターンは指数的に増大するので効率の悪さが問題となった。さらにベンゼン環をはじめとする環

状分子の扱いが想定外に難しく、結合点を与えるだけで適切な分解を高速に計算できる仕組みを検討した。
　量子計算をするべき化学反応を粗視化によって回避することが本課題のテーマの一つであるので、成果を論文として公刊するには正確と考えられる量子計算あるいは実際の化学実験との結果を比較する必要があるが、その点でデータ入手の壁にぶつかった。本研究がテストケースとして行っている簡易な計算は量子計算における主要なテーマでないため、比較のためのデータが見つけられず、より積極的に他機関の研究者との連携が不可欠であるという認識を新たにした。実装上、様々なテスト用の擬似的なデータを（半自動的に）生成する必要があり、そのためのシステム構築を行い、学会にて発表した。
　なお、期間満了後も引き続き論文化とプログラム公開までを進める予定である。

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計１件）
○城 真範、センシティブ情報を回避する疑似データの作成器、信学技報、査読無、115巻351号 pp.37-40.

〔学会発表〕（計７件）
① 城 真範、有機化学反応シミュレーションの高度化について、細胞を作る研究会、2013
② 城 真範,牧野 貴樹,合原 一幸、Proposal of a New Method in Chemical Simulation、International Symposium on Innovative Mathematical Modelling, 3rd.、2013年
③ 城 真範、こういう分子シミュレーションって何がダメなんでしょうか？、生物物理若手の会、2013年
④ 城 真範、機械学習手法による化学反応のシミュレーション、東京大学生命化学シンポジウム、2013年
⑤ 城 真範、低分子における時間逆方向シミュレーションの可能性、「細胞を創る」研究会7.0、2014年
⑥ 城 真範, 赤穂 昭太郎、化学反応シミュレーションの結果から最適反応温度を求めるための予備研究、第17回情報論的学習理論ワークショップ、2014年
⑦ 城 真範、センシティブ情報を回避する疑似データの作成器、電子通信情報学会SITE研究会、2015年

〔図書〕
なし

〔産業財産権〕
なし

〔その他〕
なし

６．研究組織
(1)研究代表者
城 真範（SHIRO,Masanori）
産業技術総合研究所・情報・人間工学領域・主任研究員
研究者番号：９０３５７２４４

(2)研究分担者
なし

(3)研究協力者
なし