様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２８　年　　６　月　１５　日現在

機関番号：　１３３０２
研究種目：　若手研究(B)
研究期間：　2012～2015
課題番号：　２４７００１４５

研究課題名（和文）機械学習を用いた水溶液にひそむ動的秩序の抽出

研究課題名（英文）Extraction of dynamical hidden order of water by mining simulation data

研究代表者

ＤＡＭ　Ｈｉｅｕ Ｃｈｉ（DAM, Hieu Chi）

北陸先端科学技術大学院大学・知識科学研究科・准教授

研究者番号：７０３９７２３０

交付決定額（研究期間全体）：（直接経費）　　2,500,000円

研究成果の概要（和文）：ミクロなスケールでみた水分子の振る舞いには，ランダムに見えて幅広い多様性がある．それは水が，小さな分子量（M=18）に対して，比較的大きな静電気双極子モーメントを持ち，お互いに“粘りつく”ように緊密な相互作用をして存在しているからである．また複雑なエネルギーランドスケープを持ち，様々な局所構造をとる．周りとの相互作用が大きい水分子の動きは，その相互作用を通じて周りの環境に関する情報を取り入れている．この“振る舞い”を，データマイニングするにあたり，自然言語処理的な“文脈”の解釈を行う可能性を念頭に置いて，解析系を構築した．

研究成果の概要（英文）：Water molecules with their electrostatic dipole moments and characteristic hydrogen bond network have tight interaction to each other as well as to proteins in solutions. The water molecules are moving under the interactions with the surrounding water molecules and information about the local chemical environment is implicitly included in their dynamical behaviors. By applying data mining techniques to the simulation data of protein solution, we have constructed an analysis system for analyzing the dynamical behavior of water molecules for extracting the hidden dynamical structure of protein solution.

研究分野：　計算物理学

キーワード：　データマイニング　動的振る舞い　水素結合ネットワーク

１．研究開始当初の背景

タンパク質の機能発現はタンパク質自身や周辺環境のわずかな差異が決定的な役割を果たし，そのメカニズム解明には分子レベルでの詳細な解析が不可欠である．特に溶媒水溶液の影響は非常に大きく，近年ではタンパク質表面付近の水(水和水)の緩和構造や，水溶液系全体の水素結合ネットワーク構造からタンパク質の構造・機能への多大な寄与が明らかになってきた．しかし，水溶液の本質的因子はネットワーク構造の”ゆらぎ”に潜んでおり，微視的解析により分子の局所構造を分断的に捉えるのみではその全貌は分からず，”タンパク質-溶液”一体系として全体を俯瞰し，”ゆらぎ”とタンパク質機能発現との関連性を理解する事が重要である．

この様な分子レベルでのダイナミクスを追跡する手段として計算機上で微視的な運動方程式を解き，全原子の位置・速度情報を特定出来る分子動力学(MD)シミュレーションは極めて有効な解析手法である．近年では 1 万コア以上の超巨大並列計算機や GPGPU, Grid 計算など計算機技術の発展により，巨大なタンパク質複合体構造系の再現が可能になり，フォールディング機構の解明や創薬への応用など活発な発展が見られる．巨大系シミュレーションでは再現される現象は緻密であるが，そのデータ量とデータの複雑性から人間の直感に頼った解析は困難で，その中に内包されている水素結合ネットワークや”ゆらぎ”情報を採掘する事は難しい．水溶液の理解には内包データから自然現象を採掘する革新的な解析手法の開発が不可欠である．

２．研究の目的

本研究の目的は情報科学理論分野で発展目覚ましいデータマイニング(MD)手法と理論物理分野で広く用いられている分子動力学(MD)シミュレーションを融合(DM/MD 融合法)した巨大データ解析の為の方法論を提案し，タンパク質水溶液系における水分子の振るまいを分子レベルで分類する事である．本手法では MD シミュレーションのデータは物理方程式の解という観点を越え，より広義の「データ集合」と見なし，主観や経験に頼らずパターン認識技術を駆使する事で全水分子個々の振る舞いを分類する．研究対象はタンパク質とし，その水和水･バルク水の別を分類，さらにタンパク質間相互作用に関与する分子集団の存在とその特徴を明らかにする．

以上の目的をうけて，１)「手法の開発および水溶液に対する適用； MD とデータマイニングによる水分子の振る舞いの解明」，および今回研究を進める過程で新しい視点を得たことによる手法の改良２)「マイニング手法の改良；教師有り学習と教師無し学習の融合」，の二部に分けて報告する．

(1)手法の開発および水溶液に対する適用；MD とデータマイニングによる水分子の振る舞いの解明

ミクロなスケールでみた水分子の振る舞いには，ランダムに見えて幅広い多様性がある．それは水が，小さな分子量（M=18）に対して，比較的大きな静電気双極子モーメントを持ち，お互いに“粘りつく”ように緊密な相互作用をして存在しているからである．「くの字」に折れ曲がった分子構造により水素結合による多くの種類の配位構造をもつこともあいまって，全体として複雑なエネルギーランドスケープを持ち，したがって水分子集団のとりうる局所構造は多様である．

周辺環境との相互作用が大きい水分子の動きは，その相互作用を通じて周りの環境に関する情報を取り入れている．すなわち，水の振る舞い自体に周辺環境の情報が埋め込まれていることが期待される．この期待により，水分子の個々の振る舞いに着目してマイ

ニングを行うことで周辺環境に関する知見まで得られることを目指した.

この“水分子の振る舞い”を，データマイニングするにあたり，自然言語処理的な“文脈”の解釈を行う可能性を念頭に置いて，解析系を構築した.

(2) 新しい視点によるマイニング手法の改良；教師有り学習と教師無し学習の融合

さて，水分子の振る舞いに関するマイニングを進めた過程において，より包括的な材料科学に関するデータマイニングの手法の開発を行うことが出来た．蛋白質水溶液中の水分子の動力学からのデータマイニングは，蛋白質というヘテロな環境，それに付随するヘテロな水分子の環境をもつ系である事から，予測というよりは，データのクラスタリングの比重が高く，より高度なマイニング系を検討し，構築するには不利である．そこで我々は，より単純なシリコン液体の系を用い，各物理量がどのような構造を持つのかを検討し，データマイニングによって何が予測され，コントロールすることのできる対象となるのかを検討した.

予測・学習において，用意されたデータから意味のある結果を得るには，なんらかの形で意味のある結果に繋がる情報をデータに含めておく必要がある．しかしながらマイニング以前に目的の情報を確実に用意するのは多くの場合困難である．したがって，あらかじめできうる限り多くの種類のデータを対象に含める必要がある.

しかしながら，このように多数のパラメータを採用したときに，全オブジェクトに対して一気に学習・予測を進めようとすると，モデルが複雑であることが原因で，限られた条件・情報による収束的な解が得られにくいという欠点がある．もちろんマイニングには，特徴量の選択や情報の縮約などに既知の手法が多数あり，データに含まれる目的以外の情報を排除することは可能である．一方その適用についてのストラテジーは解くべき問題につよく依存し，一般的な手法として確立されたものではなかった.

今回この問題を根本的に解決するアプローチを得たのでそれを紹介する.

３．研究の方法

(1)手法の開発および水溶液に対する適用；MD とデータマイニングによる水分子の振る舞いの解明

研究は以下のように多段階的なプロセスによって行った.

1. 水分子の分子動力学計算(MD)
2. トラジェクトリを取り出し，特徴量を抽出する.
3. 特徴空間を構築して，その上でクラスタリングを行う.
4. クラス化された水分子の振る舞いに関して，物理化学的な性質を再現して，その意味を反映する.

プロセス 3 における特徴量の抽出は以下のように行った．まずタンパク質などの表面にある水分子の結合サイトを想定し，類似の振る舞いがバルク空間においても存在すると仮定する．これにより蛋白質表面からの距離に無関係な水分子の仮想的な“サイト”を，水分子座標のガウス分布によって定義した．全ての水分子の全てのトラジェクトリ時間領域に対して，混合分布モデル（Gaussian Mixture Model）をもちいてクラスタリングを行った．MD 計算をおこなった全空間に仮想的に設定された水和“サイト”は，それぞれ 3 次元のガウス分布であり，そのパラメータ群をベクトルデータ化し “水分子の振る舞い”をしめす特徴データベクトルを得た．次にデータから特徴量の候補を抽出し，特徴空間を構築した．理想的な特徴空間を得るために，その後の試行錯誤的・再帰的にアップデ

ートを行った．最後に特徴空間上の全てのデータ点（水分子の振る舞いをクラスタ化した特徴量の集合）を，ふたたび特徴空間上でクラスタリングを行い，特徴空間上で表現されたガウス混合分布からなる複数のクラスを得た．

各クラスに属する特徴空間における点は，特徴量によって分割された“水分子の振る舞い”であり，実際にトラジェクトリの部分に対応しているので，そこから様々な物理化学量を計算できる．すなわち，各“水分子の振る舞いクラス”に特徴的な物理化学量を議論することが出来る．

(2) 新しい視点によるマイニング手法の改良；教師有り学習と教師無し学習の融合

いま一般的なモデルについて考える．学習・予測の際に，（結果的に）解空間で局所的に関連のあるパラメータ・オブジェクトを集め，そのグループ毎に単純なモデルを構築できるとする．こういった際に，よく使われるのが線形モデルである本節(1)で使われたGaussian Mixture Modelも 線形モデルの一種である．ここで，その複数の単純モデルによる線形結合を考えよう．この場合の各単純モデルは，unsupervised learning でもってその解を独立に推察することができるとする．

今回，このような条件を満たすモデルと，学習・予測すべき具体的な対象については，第一原理 MD からの atomic potential の学習・予測という問題を提唱する．

ここでは問題は，第一原理 MD による原子配置をふくむ一連の量子化学データから，いかにして特徴ある原子の状態を記述し学習・予測するかにある．これらを解く過程は以下のようになり，それぞれの手法を構築する必要がある．

① local structure の表現
② 関数の学習
③ 第一原理MD計算からの出力total energyからatomic energyを算出・分配する

まず上でも述べたように，原子群の局所構造からの情報を獲得するために，RDF(radial distribution function) を対象原子のある近傍についていくつかの原子に関して定義する．これによって各原子の環境をベクトル化できる．(①の解決)

次に②関数の学習に関する手法を開発する．MDでは，真空界面の液体シリコンを用いたが，この場合，バルクの状態と界面の状態で電子状態や（化学結合の）結合状態が異なる．このように異なった状態においては，atomic energy も異なる．これらのおおまかに2つのグループに関して各物理量をどのようにわけるかという問題が生じる．

一方,total energyは計算するのは容易で，第一原理 MD の結果をそのまま参照すれば良い．原子の状態（Chemical Environment）を次元削減やPCA でもって表現し，そこから特徴空間上で表現したのが図1である．これは

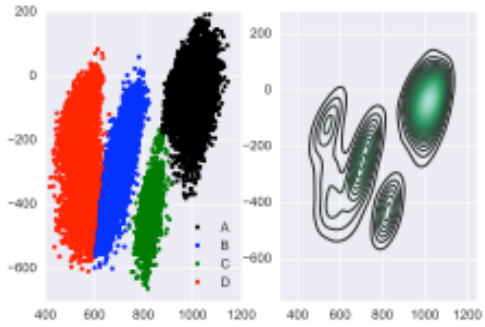

図 1. 混合分布モデルによる原子環境の分類

基本的にGMMになっており，簡単にグループ分けができる．

さて，このような状態で，原子の状態の描写（デスクリプター）から，トータルエネルギーを予測するためには，どうすればいいのであろうか．我々はここで，図1のガウス混合分布をそのまま予測式，学習式の定式化に導入し，全体のプレディクターを，混合ガウス分布の線形結合で表現する手法を考案した．これによると，全体のパラメータを用い

た式で，局所的な解をすべて満足するような予測式がたてられるのである．

これは考えてみれば当たり前な話で，GMM の有限な値をとる範囲では，有界なパラメータ値が保証されている一方で，それ以外の領域では，そのパラメータは GMM の中心値から離れることで限りなく小さくなり，予測式のなかでの寄与は小さくなる．このことにより，GMM の近傍では，ローカルな予測式の解が保証され，離れたところでは寄与が小さくなるような条件が得られた．

以上の手法によって，全体の解を満足し，局所解も満足するような，予測式をたてることができた．

４．研究成果

(1)手法の開発および水溶液に対する適用；MD とデータマイニングによる水分子の振る舞いの解明

我々は，バルク水と水溶性タンパク質の溶液を対象に，以上の解析を行い，タンパク質水和水に特徴的な水分子の振る舞いクラスを得た．それによると，第一水和圏内に分布し，非常に長い緩和時間を持ち，タンパク質の疎水面と親水面において動的性質に違いが見られた．また，タンパク質溶液に特徴的な別のクラスでは，非常に速い緩和をもつ一群が存在し，それぞれのクラス間でのダイナミクス（時空相関関数）から水和層において特徴的な振る舞いが観察された．同様の解析を孤立タンパク質のみならずタンパク質間相互作用に関与する水分子にも行い，物理化学量の相関を得た．

水分子の“振る舞い”に関する特徴空間上のクラスタリングによって，水の振る舞い自体を分離することが可能となり，かつそれらに基づいて物理化学的な量を算出し，物理化学的な特性空間で特徴を同定する事が可能となった．また水分子の振る舞いが異なるクラス間で，属する水分子による相関を考察することが出来るようになった．そのひとつが，第一水和水として安定的に存在するタンパク質水和サイトであり，もうひとつが第一水和圏にありながら安定水和サイトの近傍で相互作用する速い拡散定数を示す振る舞いの水分子である．これらの複雑な動的構造は，これまでみられないものである．今後さまざまな物理化学量と関連づけることで，これらの動的な振る舞いから，「バイオロジカル・ウォーター」に関する動的な知見を得るきっかけとなりうる点で注目に値する．

(2) 新しい視点によるマイニング手法の改良；教師有り学習と教師無し学習の融合

実際に，この手法による予測精度の向上を示した．シンプルな線形結合モデルの方法の場合と，混合分布による線形結合モデルの場合とでは，予測精度が 3 倍（バラツキが 1/3）程度向上している．これは，局所解と広域解を同時に満たすことのできた予測式による効果であることを確認した．

(3) 全体のまとめ

今回の研究計画を実行するにあたって，１）タンパク質溶液中の水分子の動的な構造が同定され水溶性タンパク質およびタンパク質間の相互作用に関与すると推察される水分子の振る舞いに関していくつかの候補を得た．２）溶液のダイナミクスにおいて，学習・予測における一般的なフレームワークを確立し，かつ混合分布のアイデアを予測式そのものに取り入れる事によって，多数のパラメータを採用しながら予測精度を大きく向上させることが可能となった．

これら成果はそれぞれ，物理過程をマイニングするにあたって物理量を特徴空間に落とし込む際の新たな手法，および，その後の学習・予測の過程における一般性をもった新たなアプローチの開発と位置づけられる．今後，この手法を用いて様々な材料を対象にすることで，大きな展開が期待される．

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９件）

1. Anh Tuan Nguyen, Van Thanh Nguyen, Thi Tuan Anh Pham, Viet Thang Do, Huy Sinh Nguyen and **Hieu Chi Dam**, "*Correlation between charge transfer and exchange coupling in carbon-based magnetic materials*", AIP Advances **5**, 107109 (2015). (査読有り)
2. T. Kawasaki, V.C. Nguyen, L.M. Nguyen, T.B. Ho, and **H.C. Dam**, "*Partially Clustered Linear Regression towards Improvement of Clustering Performance*". In Proceedings of ACIS 2014, pp. 110–113, Nha Trang (2014). (Corresponding author) (査読有り)
3. ***H.C. Dam**, T.L. Pham, T.B. Ho, A.T. Nguyen, and V.C. Nguyen, "*Data mining for materials design: A computational study of single molecule magnet*". The Journal of Chemical Physics, 140, 044101 (2014). (査読有り)
4. T.H. Nguyen, K. Umemoto, **H.C. Dam**, T.V.D. Dang, "The role of coordinators in value co-creation process in software offshoring: A Knowledge Management Perspective", International Journal of Knowledge and Science, 5, 2, 1-18 (2014). (査読有り)
5. Ho Tu Bao, Taewijit Siriwon,. Ho Quang Bach, **H.C. Dam**, "*Progressive Trends in Knowledge and System-based Science for Service Innovation*" (Chapter 7). Big Data and Service Science, 127-144, IGI Global (2013). (査読有り)
6. T.C. Nguyen, H. Mizuta, T.C. Bach, N. Otsuka, **H.C. Dam**, "*Ab-initio calculations of electronic properties and quantum transport in U-shaped Graphene nanoribbons*", International Journal of Computational Materials Science and Engineering, World Scientific., Vol. 01, Issue. 03, 1250030, 11 pages (2013). (Corresponding author). (査読有り)
7. T.L. Pham, A. Sugiyama, T. Masuda, T. Shimoda, N. Otsuka, **H.C. Dam**, "*Ab-initio study of intermolecular interaction and structure of liquid cyclopentasilan*", Chemical Physics, Vol. 400, 59, 6 pages (2012). (Corresponding author) (査読有り)
8. H. Kameda, J. Li, **H.C. Dam**, A. Sugiyama, K. Higashimine, T. Uruga, H. Tanida, K. Kato, T. Kaneda, T. Miyasako, E. Tokumitsu, T. Mitani, T. Shimoda, "Crystallization of lead zirconate titanate without passing through pyrochlore by new solution process", J. Eur. Ceram. Soc, 32, 1667-1680 (2012). (査読有り)
9. H. Mizuta, Z. Moktadir, S. A. Boden, N. Kalhor, S. Hang, M. E. Schmidt, N. T. Cuong, **H.C. Dam**, N. Otsuka, M. Muruagnathan, Y. Tsuchiya, H. Chong, H. N. Rutt and D. M. Bagnall , "Fabrication and ab initio study of downscaled graphene nanoelectronic devices", Proc. SPIE, 8462, 846206, (2012). (査読有り)

〔学会発表〕（計０件）

〔図書〕（計０件）

〔産業財産権〕
○出願状況（計０件）

○取得状況（計０件）

〔その他〕
ホームページ等

６．研究組織

(1)研究代表者　Dam Hieu Chi (Dam Hieu Chi)
北陸先端科学技術大学院大学・知識科学研究科・准教授
研究者番号：70397230

(2)研究分担者
研究者番号：

(3)連携研究者
研究者番号：