「科学的発見・社会的課題解決に向けた各分野のビッグデータ利活用推進のための次世代アプリケーション技術の創出・高度化」
平成 25 年度採択研究代表者

H26 年度
実績報告書

船津　公人

東京大学大学院工学系研究科
教授

医薬品創薬から製造までのビッグデータからの知識創出基盤の確立

§1. 研究実施体制

(1)「船津」グループ
① 研究代表者:船津　公人 （東京大学大学院工学系研究科、教授）
② 研究項目
・製造プラントにおける大規模運転データベースの管理手法の開発
・プラント運転モニタリングのための自動的モデル構築システムの開発
・運転監視・プロセス制御のための知識抽出
・反応の評価手法の開発

(2)「奥野」グループ
① 主たる共同研究者:奥野恭史 （京都大学大学院医学研究科、教授）
② 研究項目
・ケミカル情報とバイオ情報の統合化と高速処理を可能にするデータ構造とアルゴリズムの開発
・ケミカル情報とバイオ情報の相互作用ビッグデータ解析を可能にする数理的モデルの開発

(3)「泰地」グループ
① 主たる共同研究者:泰地　真弘人 （(独)理化学研究所 生命システム研究センター、副センター長）
② 研究項目
・仮想大規模ライブラリの拡充

・超大規模ライブラリからの有用情報検索技術の開発
・超大規模仮想ライブラリのコンテンツ可視化技術の開発

§2. 研究実施の概要

(1)「船津」グループ

産業プラントにおいては測定困難なプロセス変数をリアルタイムに推定する手法としてソフトセンサーが広く使用されている。予測的なソフトセンサーを運用するため、前年度にデータベースの管理用の指標 database monitoring index (DMI) および DMI を用いたデータベース管理手法を開発した。しかし、データベースに外れ値が混入するとソフトセンサーの精度が低下してしまう。今年度は最初のデータベースから自動的に外れ値を検出する手法を考案した。本手法を用いてプラントのデータ解析を行った結果、提案手法により適切に外れ値を検出可能でありソフトセンサーの構築に適切なデータベースが得られることを確認した。

上述のシステムにより処理されたデータベースを用いて、自動的にソフトセンサーモデルを構築するシステムを開発した。モデルを構築する際は事前に決定すべきパラメータが存在するため予測的なモデルを構築可能なパラメータの組合せを高速かつ自動的に決定する方法を開発した。

前年度に立てた運転データを活用したプロセス管理の方針に基づき、ソフトセンサーを利用した効率的なプロセス制御手法を開発した。プロセスの動特性を考慮して運転データから目的変数 y と説明変数 X の間でモデル y=f(X)を構築することで、所望の y の値を達成する X の制御方法を出力できる。continuous stirred tank reactor (CSTR) のシミュレータを用いたケーススタディを行い、観測可能な外乱および非観測外乱が存在する状況下においても、y の設定値変更に対して提案手法が有効であることを示した。

泰地グループで構築中の大規模仮想ライブラリの探索により薬物候補構造を取得できるが、同時に提示されるその合成経路情報の妥当性を評価するためには、当該反応の遷移状態(TS)の有無を確かめる必要がある。その目的のため、TS 情報を有する遷移状態ライブラリ(QMLB)の作成と、QMLB から反応情報を自動で選択するデータベースシステム(遷移状態データベース、TSDB)の開発を行った。

(2)「奥野」グループ

奥野グループは、「大量のタンパク質 対 化合物情報からの創薬指針の抽出」を担当している。平成 26 年度では、開発項目1)ケミカル情報とバイオ情報の統合化と高速処理を可能にするデータ構造とアルゴリズムの開発と、2)ケミカル情報とバイオ情報の相互作用ビッグデータ解析を可能にする数理的モデルの開発を行った。具体的には、米国 NIH/NCBI が公開している世界最大の生物活性データベース PubChem/Bioassay について NoSQL 型の MongoDB 形式を採用し、2,133,668 化合物における生物活性レコード 132,942,821 のデータ量のデータベースを構築した。さらには、化合物とタンパク質の相互作用データのビッグデータ化に対応できる新たな機械学習法として、Deep Learning(深層学習)に基づく手法の開発を行った。

(3)「泰地」グループ

大規模仮想化合物ライブラリの高度化を目的として次の三項目、仮想大規模ライブラリの拡充、超大規模ライブラリからの有用情報検索技術の開発、超大規模仮想ライブラリのコンテンツ可視化

技術の開発、について研究を実施している。今年度は各項目について以下の研究を行った。

仮想大規模ライブラリの拡充については、種となる化学構造に反応トランスフォームを適用して新規構造を得る手法により、約 31 億件にのぼる化学構造を創出した。この内重複する構造の排除を行うことで約 17.5 億件の非冗長な構造ライブラリが得られた。更に付随する合成経路情報として順合成 3 千 6 百万件と逆合成 15 億件が得られた。化学構造創出エンジンの改良についてはオープンソースの RDKit を利活用し、既存エンジンの特徴を後に組込む方針での開発を行った。改良エンジンを用いたテスト生成を行い、平成 27 年度内のエンジン改良基本作業終了に向けた更なる改善点の検討を行った。

超大規模ライブラリからの有用情報検索技術の開発については、超大規模の格納情報から利用者が合理的な時間で目的とする薬物候補情報を絞り込むことが可能な検索技術が必要となるため、分子量をはじめとした物性値条件に加え、部分構造および Tanimoto 係数に基づく類似度による検索機能の強化を行った。導入した高速ストレージ上に単一の基本データベースを再構築し、事前に計算可能な物性値やフィンガープリント情報をも格納しておくことで条件検索を高速化した。

また、超大規模仮想ライブラリのコンテンツ可視化技術の開発については、Deep Learning の応用可能性を中心に基礎的な検討を続けている。

・代表的な原著論文

1) Matheus de Souza Escobar, Hiromasa Kaneko, Kimito Funatsu, Combined Generative Topographic Mapping and Graph Theory Unsupervised Approach for Non-linear Fault Identification, AIChE Journal, accepted. DOI: 10.1002/aic.14748
2) Hiromasa Kaneko and Kimito Funatsu, Fast Optimization of Hyperparameters for Support Vector Regression Models with Highly Predictive Ability, Chemometrics and Intelligent Laboratory Systems, 142, 64-69, 2015. DOI:10.1016/j.chemolab.2015.01.001
3) 木村一平, 金子弘昌, 船津公人, ソフトセンサーとその逆解析を利用した新規フィードフォワード制御手法の開発, 化学工学論文集, 41(1), 29-37. 2015. DOI: 10.1252/kakoronbunshu.41.29