

厚生労働科学研究費補助金
(政策科学総合研究事業 (臨床研究等 ICT 基盤構築・人工知能実装研究事業))
分担研究報告書

課題名 : 新薬創出を加速する症例データベースの構築・拡充/創薬ターゲット推定アルゴリズムの開発に関する研究
研究分担者名 : 高村 大也
国立研究開発法人産業技術総合研究所・人工知能研究センター・研究チーム長

研究要旨

基礎生物学系文献からの分子、病態の知識抽出については、構築した言語リソースの評価を進め、またそれを用いて、エンティティ、関係及びイベント抽出モデルの開発及びツール化を行った。また、論文から情報を抽出するキュレーション技術の開発を行った。また、疾患ネットワークの構築について、その方法の開発及び、視覚化及びツール化を進めた。
さらに、基礎生物学データベースの情報と抽出情報との融合を目指し、当研究所で開発が続けられている PoSSuM などの蛋白質に関するデータベースを利用した検索・推論基盤の構築に向けた準備を実施した。

A. 研究目的

本研究プロジェクトは、論文などの技術文献から、生命系のエンティティ (蛋白質、疾患など) やイベント、及びそれらの間の関係を、自動的に抽出する技術を開発することを目的とする。さらに、それを用いて実際に知識抽出を行い、抽出した知識を、エンティティをノード、イベントをエッジとし、関係を持つノード同士を結合した、巨大な高次ネットワーク構造 (疾患ネットワーク) で表す。また、既存の蛋白質データベースやオントロジーの情報などを合わせて、蛋白質の配列比較や化合物の構造比較に関するバイオインフォマティクス/ケモインフォマティクスの技術と自然言語処理の技術を統合して、蛋白質などのエンティティ間の類似度も測定可能にする。疾患ネットワークと、エンティティ間類似度を利用することで、柔軟かつ広範囲にエンティティ間の相互作用を推論する技術を開発する。

より具体的には、次の 3 つの課題を解く。

- ① 基礎生物学系文献からの知識抽出
- ② 基礎生物学情報と抽出情報との融合
- ③ 知識の表現方法及び柔軟な推論技術

B. 研究方法**① 基礎生物学系文献からの知識抽出 :**

前年度までに、特発性肺線維症に関する文献要旨データ (120件分) に対して、エンティティ (12タイプ : Disorder (疾患名)、Cell (細胞種)、Pharmacological substance (薬剤名)、GDPs (遺伝子、遺伝子産物) など)、イベント (14タイプ : Regulation (制御)、Gene expression (遺伝子発現) など)、エンティティ間リレーション (4タイプ : Subject-Disorder (患者-疾患名) など) のマニュアルアノテーション及びノーマリゼーションによって構築したコーパスを学習・評価データセットに用いて、深層学習によるエンティティ抽出・リンキングモデル、関係抽出モデル、イベント抽出モデルの開発の継続及び学習と評価を行った。

これらの開発した深層学習モデル手法の一般公開と利用促進を図ることを目的に、ウェブツール化を行なった。さらに、②につながるよう、知識抽出モデルの出力から疾患ネットワークを構築するためのルールを策定する。

② 知識の表現方法及び柔軟な推論技術の開発 :

①で開発した知識抽出モデルによる計算結果から、エンティティをノード、イベントをエッジとし、疾患ネットワークを実際に視覚化して表現する方法、知識グラフに利用するためのデータの形式化、そのツールの開発を行った。

③ 基礎生物学情報と抽出情報との融合 :

新規標的候補などの同定に向けた文献情報と基礎生物学情報とのさらなる融合を目指し、当研究所で継続的に開発されている PoSSuM などの蛋白質に関するデータベースを利用し、知識抽出及び推論技術により構築された疾患ネットワークの検証及び拡充 (新規情報の追加など) を可能とする検索・推論基盤の構築に資する。近年、蛋白質立体構造決定技術の向上により利用可能な構造情報が急増している。これらの全ての情報を利用するため新規サーバーを導入し、昨年度より PoSSuM データベースの更新作業を行っている。本年度は PoSSuM データベースの更新にも着手する。これと同時に、検索・推論基盤構築に向けたケーススタディとして、肺線維症などの発症や治療に関わる免疫システムの論文調査を行い、抽出データを基に検索・推論基盤の構築を目指す。

(倫理面への配慮)

本研究は、人及び動物を研究対象としていない研究であるため、倫理面の問題はないと判断した。

C. 研究結果

前年までに開発された言語リソースを用い、エンティティ抽出モデル、エンティティリンキングモデル（ノーマリゼーションモデル）、関係抽出モデル、イベント抽出モデルの開発の継続、及び学習と評価を行った。これらのモデルはAPIツール化し、本研究予算にて購入した所内設置の計算サーバ上に実装した。また、上記モデルで抽出した情報から、疾患ネットワークの構築のルールを開発し、それを視覚化して表現するシステムを構築した。また、論文から情報を抽出するキュレーション技術の開発を継続した。

さらに、当研究所で開発された蛋白質の基質結合（候補）部位に関するデータベース PoSSuM を例として、検索・推論基盤の構築に向けた準備（PoSSuM の更新）を実施した。PDB データベースに 2020 年 9 月時点で登録されていた約 16 万件の蛋白質の立体構造及びアノテーションデータに対する、基質結合（候補）部位の類似性検索等の計算及びウェブサイトの更新作業を完了した。また、本プロジェクトで対象疾患の発症や治療に関わると予測された蛋白質の学術論文調査について、調査結果の再抽出及び抽出結果の再フォーマットを実施した。

D. 考察

疾患ネットワーク構築ルールについては、さらなる検討が必要である。またキュレーション技術についても継続しての研究開発が必要である。開発したAPIについては、セキュリティ関係の手続きを経て、医薬基盤・健康・栄養研究所にて構築が進められているオープンプラットフォーム「峰」を含めた外部から利用してもらえるように設定する予定である。また、PoSSuM と TargeMine との接続について今後さらなる検討を要すると考えられる。

E. 結論

情報抽出技術の開発、検索・推論基盤の構築のそれぞれについて、予定していた開発目標を達成した。

F. 研究発表

1. 論文発表
なし
2. 学会発表
なし

G. 知的財産権の出願・登録状況（予定を含む。）

1. 特許取得
なし
2. 実用新案登録
なし
3. その他
なし

厚生労働科学研究費補助金
(政策科学総合研究事業 (臨床研究等 ICT 基盤構築・人工知能実装研究事業))
分担研究報告書

課題名 : 新薬創出を加速する症例データベースの構築・拡充/創薬ターゲット推定アルゴリズムの開発

研究分担者名 : 西村 紳一郎

国立大学法人北海道大学大学院先端生命科学研究院 先端融合科学研究部門
 新薬探索研究分野 教授

研究要旨

本研究チームは「血中エクソソームの由来臓器を糖鎖解析により推定する基盤技術」という研究テーマを分担している。令和3年度は前年度完了したマウス主要臓器（脳・肝臓・肺・骨格筋）と大腿骨、腎臓、膵臓、脾臓、心臓、胃、腸（大腸及び小腸）、皮膚、甲状腺、卵巣、子宮、睾丸及び血清（エクソソームを含む）の定量的糖鎖プロファイリングによって得られた世界初の「マウス組織・臓器糖鎖アトラス」の発信とそのプレセンスの向上、さらにヒト組織糖鎖アトラスへの展開を企図した。

A. 研究目的

昨年度完了したマウスの主要臓器と組織の定量的糖鎖プロファイルデータについて様々な角度から解析を行い、臓器特異的な「糖衣（糖鎖パターン）」が存在することを証明する。また、多様な糖鎖パターンの解析において機械学習による評価法の有効性を明らかにする。

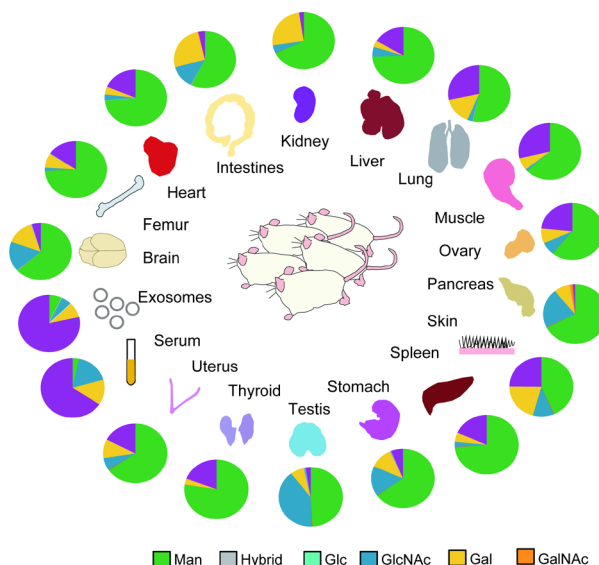


図1 マウスグライコムアトラスの対象とした組織

(倫理面への配慮)

動物実験に際しては、動物愛護管理法及び関係基準等を遵守し、動物福祉の基本概念である3R: Replacement, Reduction, Refinementの原則及び動物実験倫理に基づいて適正に実施した。

C. 研究結果

各臓器・組織、血清・エクソソームの糖鎖プロファイルを種々の方法で比較・解析した結果を図2～図6に示した。

これらの結果から、それぞれの組織・臓器及び血清・エクソソームによって糖鎖プロファイルは大きく異なっていること、さらに個々の臓器・組織に特異的に発現しているユニークな構造の糖鎖も存在することも初めて明らかとなった。例えば、脳と肺においては、他の臓器あるいは培養細胞などには見られない特徴的な糖鎖が数種類存在していた。また、このような組織・臓器ごとに特徴的な糖鎖構造のプロファイ

ル情報が得られたことによって混合物であるエクソソームの糖鎖構造プロファイルを構成する臓器・組織の割合などを推定することも可能であることが示唆されている。

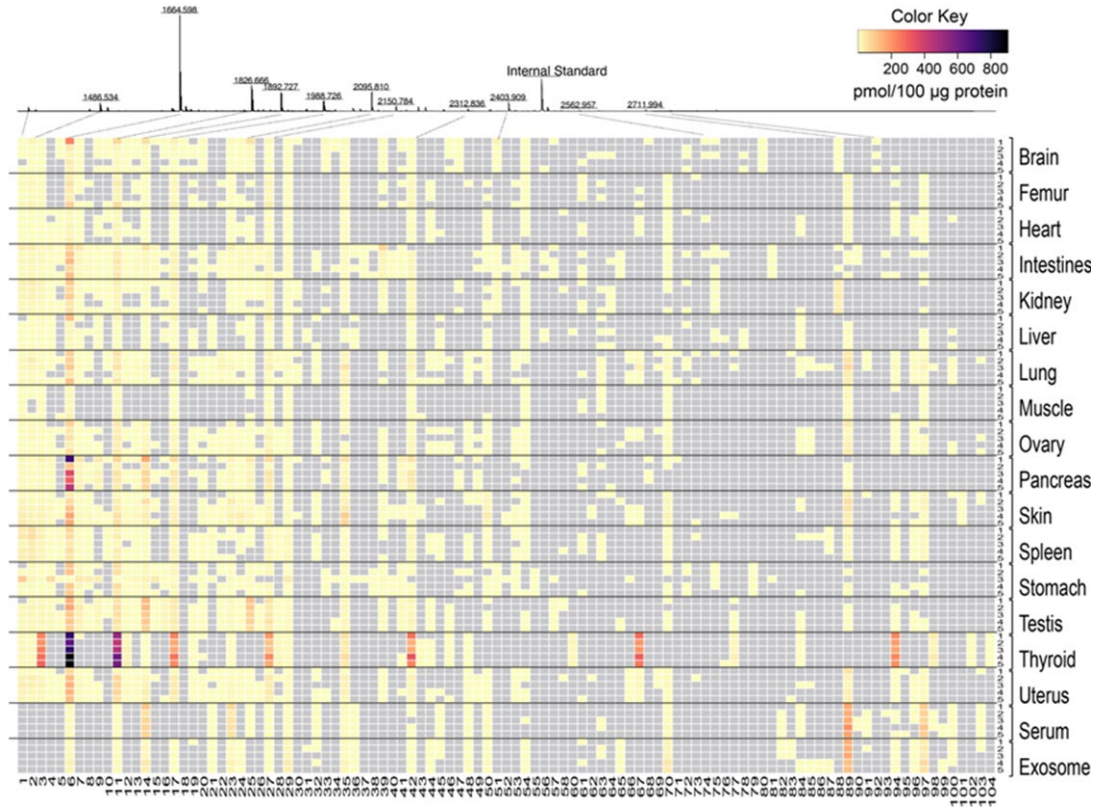


図2 マウス組織グリコームの糖鎖パターンの全体像を示すヒートマップ

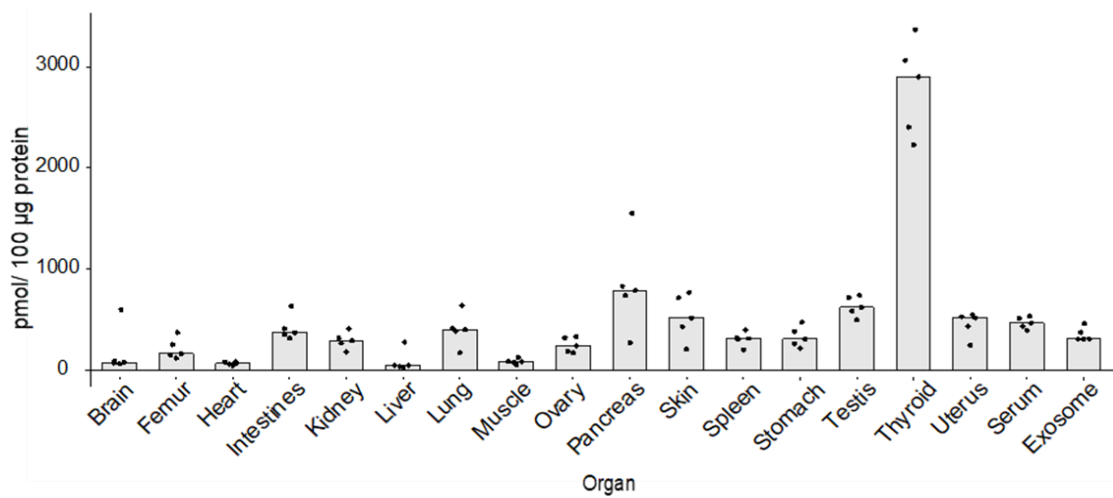


図3 各組織の全糖鎖発現量の比較(ピコモル/タンパク質100マイクログラム)

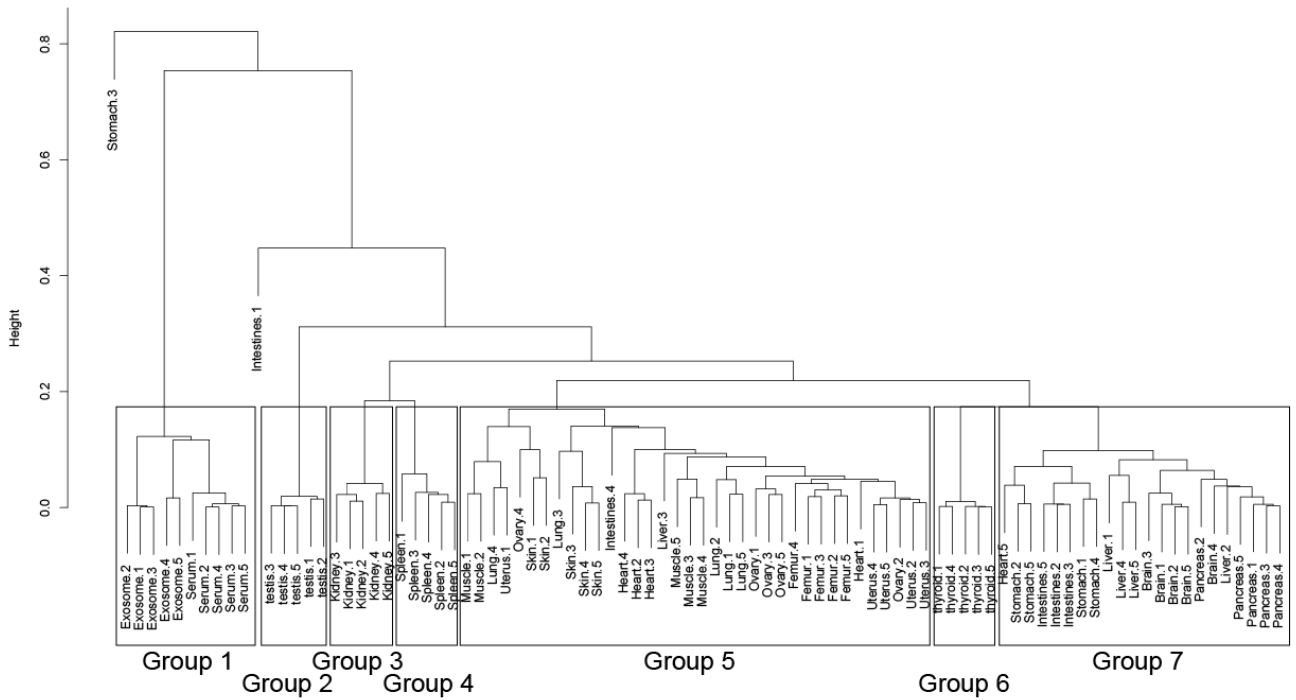


図4 各組織の糖鎖パターンのクラスター解析

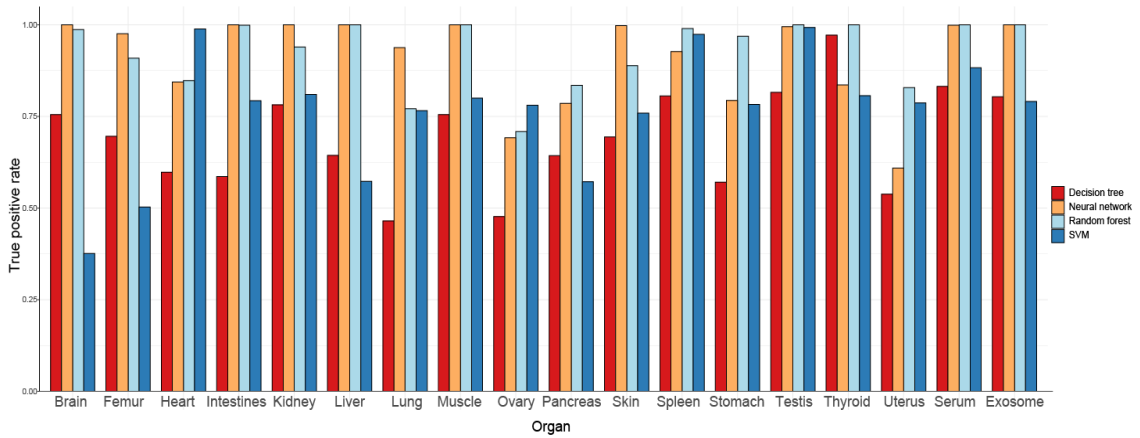


図5 各組織の糖鎖パターンを4つのアルゴリズムにより機械学習により予測させた

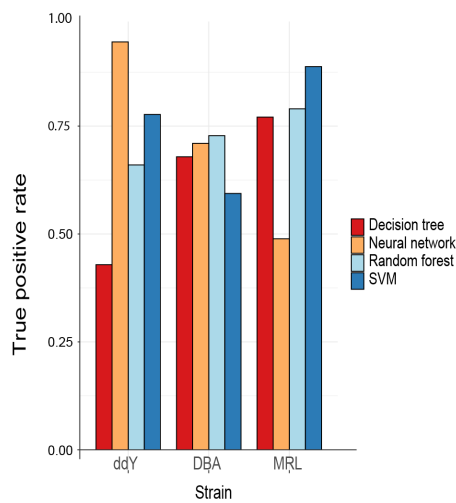


図6 3種のマウスをAIにより区別

D. 考察

糖鎖（構造）プロファイルはそれぞれの臓器（細胞）及び組織間で大きく異なっていることが改めて証明された。一方、血液から抽出・精製されるエクソソームの糖鎖プロファイルは由来する臓器・組織の糖鎖プロファイルの総和として観測されることを意味している。その際に各臓器・組織に由来するエクソソームの重みづけは個々の臓器や組織の定量的な糖鎖プロファイルとの相関（紐づけ）から予測可能であると考えられる。

この結果はマウス血液中のエクソソーム糖鎖プロファイルと今回取得した個々の臓器・組織糖鎖プロファイルの比較からエクソソームの由来臓器・組織の割合を推定することができることを意味している。言い換えると、何らかの疾患により血液中のエクソソーム糖鎖プロファイルに変化が見られた場合に、それがどの臓器・組織の異常・変調によるのかがAIなどにより容易に検出できることを示唆している（例えば図5や図6参照）。

来年度に向けて、正常マウスの臓器・組織糖鎖アトラスを健常ヒト臓器・組織糖鎖アトラスに拡張する予定である。その結果をもとに、患者血清のエクソソームの糖鎖プロファイルからの新しい診断技術の研究開発に展開する。

E. 結論

特定の組織・臓器や細胞から放出されるエクソソームの血中濃度は主要な臓器や組織の糖鎖プロファイルデータベースとの相関を解析することによって推定できることが様々な解析手法によって証明された。すなわち、正常マウスの臓器・組織糖鎖アトラスを健常ヒト臓器・組織糖鎖アトラスに拡張すれば患者血清のエクソソームの糖鎖プロファイルからの新しい診断技術の研究開発に展開できる可能性が高い。

G. 研究発表

1. 論文発表

- 1) Koide R., Hirane N., Kambe D., Yokoi Y., Otaki M., Nishimura S-I., “Antiadhesive nanosome elicits role of glycocalyx of tumor cell-derived exosomes in the organotropic cancer metastasis” *Biomaterials* 280, 121314 (2022)

2. 学会発表

- 1) 西村紳一郎、「Antibodies recognizing dynamic neoepitopes generated by cancer-specific truncated immature O-glycosylation」、環太平洋国際科学会議 2021/ The 2021 International Chemical Congress of Pacific Basin Societies / Pacifichem2021 (オンライン開催)、2021年12月、招待講演
- 2) 西村紳一郎、「動的エピトープ理論—翻訳後糖鎖修飾が惹起するタンパク質の脆弱性 / The dynamic epitope theory-Vulnerability of proteins induced by posttranslational glycosylation」、第70回高分子討論会 (オンライン開催)、2021年9月、招待講演
- 3) 西村紳一郎、「ヒト組織由来エクソソームの糖鎖プロファイリングによる新しいバイオマーカーの探索」、官民研究開発投資拡大プログラム (PRISM) 令和二年度成果報告会 (オンライン開催)、2021年7月

F. 知的財産権の出願・登録状況（予定を含む。）

4. 特許取得

なし

5. 実用新案登録

なし

6. その他

なし