ゲノム創薬・創発フォーラム 第４回シンポジウム

「デジタル技術とマスデータが拓く新たな創薬研究」

2020 年 9 月 23 日 (水) 13:00 – 18:00

開催趣旨

ゲノム創薬・創発フォーラムはヒトゲノム解明が進みつつあった 1998 年に発足したゲノム創薬フォーラムに源流をもちます。2013 年には創薬だけでなく様々な医療分野への展開を目指したゲノム創薬・医療フォーラムとなり、2019 年より、異なる分野の専門家の議論によるイノベーションを誘発したいという思いが 「創発」 という言葉に込められ、新たにゲノム創薬・創発フォーラムとして発足しました。

今回は、会員の皆さまからリクエストの多かった「マスデータ解析」を題材に、「テキストデータ」、「実験、測定データ」および「画像データ」の３つに分け、それぞれアカデミアと産業界の先生方に理論から社会実装について幅広くご講演頂きます。

マスデータはその膨大さから人間の直観では全体像を把握しづらく、デジタル技術などの適用が必須となり、大きな成果を得るためにはそういった技術を「効果的に組み合わせる」ことが必要です。技術的知識を背景に色々な掛け算を暗算し、それが「解決すべき事」という課題と直観的に結びついたときイノベーションが生まれます。当たり前の事は全て試されてしまう、こういう時代だからこそ暗算を通じて培った人間の直観が大きな意味を持ちます。個々の講演について興味をもって聞いていただきたいのはもちろんですが、是非、「今どこまで手が届くのか」、「課題解決につながる掛け算の実例」という面にも興味をもって聞いていただきたいと希望します。

一見、まったく別物のように見えるメルカリや Uber のようなサービスも「マッチング技術」×「個人への与信技術」＝「活用されていない時間や資源の有効活用」という式で表されると気づけば、身の回りにはその式が威力を発揮する分野がたくさん見えてきます。講演で見つけた式をご自身の分野に適用する新たな試みや、アカデミアや産業界の参加者を巻き込んで式自体をつくりだすという「創発」が起こることを期待しています。

なお、新型コロナウイルス感染症の予防対策からネットでの開催を基本に準備を進める予定ですので、どうぞ奮ってご参加下さい。

オーガナイザー：
大阪大学大学院医学系研究科教授 岡田 随象
塩野義製薬株式会社 創薬疾患研究所 前川 和彦

第４回シンポジウムプログラム

日時：2020 年９月２３日（水）13-18 時
場所：ネット開催
主要テーマ：デジタル技術とマスデータが拓く新たな創薬研究

13:00-13:05「代表挨拶」 東京理科大学生命医科学研究所 松島 網治 先生

13:05-13:10 「開催趣旨」 大阪大学大学院医学系研究科 岡田 随象 先生

塩野義製薬株式会社 創薬疾患研究所 前川 和彦 先生

13:10-13:55「言語処理技術による創薬・医療への支援」

産業技術総合研究所人工知能研究センター 高村 大也 先生

13:55-14:40「データの FAIR 化によるテキストデータの有効活用：COVID-19 創薬への適用」

SciBite 社 Patrick Rummler 先生

14:40-15:25「メタボロームデータ取得方法の現状と課題」

九州大学 生体防御医学研究所附属トランスオミクス医学研究センター 馬場 健史 先生

休憩

15:35-16:20「大規模なメタボローム解析に向けて」

ヒューマン・メタボローム・テクノロジーズ株式会社 山本 博之 先生

16:20-17:05「大規模イメージングデータを用いた生命システムのデータ駆動モデリング」

理化学研究所生命機能科学研究センター 大浪 修一 先生

17:05-17:50「医用画像データと画像解析技術による知見抽出」

キヤノンメディカル社 坂口 卓弥 先生

17:50-18:00 全体質疑

言語処理技術による創薬・医療への支援

産業技術総合研究所人工知能研究センター
知識情報研究チーム長
高村 大也

創薬・医療分野において、莫大な量の文献が日々生み出されている。これは人間が読める量をはるかに超えており、そこに表された知識や情報は十分に有効活用できていない。このような問題の解決策の一つとして考えられるのが言語処理技術である。言語処理技術とは、コンピュータ(人工知能)を用い人間の言語を自動的に処理する技術であり、言語理解と言語生成に大別される。
言語理解の技術を用いることで、創薬・医療分野の大量の文献を自動的に読解し、そこから知識を獲得することが可能になる。例えば、遺伝子やタンパク質間の相互作用を収集し、パスウェイを自動構築するなどの応用が考えられる。一方、言語生成は、数値データや画像データから、そこに含まれている内容を文章という形で表現する技術である。医療画像の読影レポートの自動作成などへの応用が期待されている。
本講演では、自然言語処理の基本的な考え方について説明しつつ、いくつかの研究成果やプロジェクト、また COVID-19 関連文献の解析などの応用例を紹介し、言語処理技術による創薬・医療分野への支援の可能性について議論する。

データの FAIR 化によるテキストデータの有効活用：COVID-19 創薬への適用

Findable, Accessible, Interoperable and Reusable Data
~ Empower Data with FAIR Principles and Fight Against COVID-19 ~

SciBite 社
Patrick Rummler

Overview Presentation

- **Introduction SciBite**
- **Example: FAIR Data with VOCabs**
- **Demo: The power of FAIR Data**
- **Applying VOCabs in the space of COVID-19**
- **Ingesting FAIR Data into a knowledge graph**
- **Apply a machine learning model**
- **Make predictions in the space of COVID-19**

Abstract

COVID19 has presented the scientific community with a challenge of unprecedented scale and urgency. For SciBite, with our focus on text analytics and FAIR Data, it has also presented a unique challenge. How does one go about using text analytics to make inferences about a disease which is only mentioned in a handful of papers? Here, we describe one approach we explored to identify drug repurposing candidates for this novel disease.

We began with the "CORD19" dataset - a collection of papers relating to assorted coronaviruses released by the Allen Institute for AI. As is often the case when approaching a text analytics problem, the first step was to use named entity recognition to identify key concepts mentioned in the text: genes, species, drugs, indications, and so on. Each one of these entity types is captured using a 'vocabulary (VOCab)' within our named entity recognition system. Within each VOCab, entities are given IDs, preferred names and sets of carefully curated synonyms. The severity of the situation also warranted creating new, bespoke VOCabs to capture COVID19 proteins (and their

human targets) and coronavirus strains.

The next step was to capture how the entities stood in relation to each other. The simplest way to do this was by counting cooccurrences of entities within sentences. So, for every pair of entities that occurs together at least once, we had a count. We could then use these counts alongside counts of how often the entities occurred independently to get a mutual information score for each pair of entities. Put simply, we scored how important the relationship between each entity pair was. At this point, having calculated weighted edges between different entities, we had effectively created a graph.

A graph contains relationships (or edges) between entities (or nodes). However, with COVID19 being such a new phenomenon, there were very few edges connected to it in our graph. What we needed to do was to infer edges that were likely based on the sparse information that did exist for COVID19. To do this we used a machine learning model which would take a pair of nodes as inputs and predict whether a relationship existed between them. After training the model, it was able to do this with a high degree of accuracy. We then took a subset of drugs that had passed phase 4 trials and fed the model with pairs consisting of one of these drugs plus COVID19. The result was then a machine learning model (AI) based prediction as to whether a specific drug likely had some relationship with COVID19.

メタボロームデータ取得方法の現状と課題

九州大学生体防御医学研究所
トランスオミクス医学研究センターメタボロミクス分野教授
馬場 健史

メタボロミクス（メタボローム解析）は俯瞰的視点から代謝とそれに関連する生体分子を広く見出せることから，近年その利用価値が高まっている．これまでに各種クロマトグラフと質量分析計を組み合わせたメタボローム分析手法が開発され，それらを用いた応用研究が盛んに行われている．現在用いられているメタボローム分析方法には使用する装置や解析対象の分析条件ごとに様々な手法が存在し，それぞれの研究者が選択した分析方法を用いてメタボローム解析を実施している．また，試料調製方法やデータ解析方法についても使用する分析方法により異なってくるため，使用する手法により異なるメタボロームデータが取得されているのが現状である．さらに，試料調製，機器分析，データ解析のそれぞれのプロセスにおいて，技術開発やバリデーションが不十分なところがあり，現在においても様々な技術的課題を抱えている．なかでも，定量値，すなわち代謝物の濃度値を取得については，早期に解決が望まれている重要な課題である．メタボローム分析法のほとんどが比較対象とするサンプル間のピークエリア（ピーク高さ）比較に基づく相対比較であり，サンプル内の各代謝物の濃度値は算出できていない．質量分析計は，感度，選択性が高く，化合物情報を得られることから，バラエティーに富んだ代謝物の一斉分析を行うメタボローム分析において非常に有用な検出器であるが，一方で共溶出が頻発するメタボローム分析においてはマトリックス効果による影響を受けるため一般的な外部検量線を用いた定量方法を使用することができない．また，質量分析においては安定同位体ラベル化体を用いた絶対定量値の取得が可能であるが，解析対象化合物の全ての標準品を入手することができず，メタボローム解析における絶対定量は事実上困難な状況にある．近年では，メタボロームと他のオミクスデータを統合して解析するトランスオミクス解析が試みられてきている．生体内における代謝変動を理解するためには，表現型や他のオミクスデータと対応させながら解析する必要があり，このためには代謝ネットワークを構成する変数である化合物濃度と酵素濃度・速度定数などのデータを必要とする．代謝物濃度の絶対定量値が取得できない現状では，メタボローム解析結果を多階層オミクスデータと対応させながら生理学的・生化学的考察を深めるトランスオミクス研究を実施することは困難である．また，異なる施設間での異なる分析装置や異なる分析手法で取得したデータを直接比較できないことも複数機関で取得したデータの統合解析の大きな障壁となっている．

本講演では，メタボロームデータの取得における現状および課題について言及するとともに，定量値の取得方法の開発を含む高品質のメタボロームデータ取得に向けた様々な取り組みについても紹介し，その重要性や今後の課題について議論したい．

大規模なメタボローム解析に向けて

ヒューマン・メタボローム・テクノロジーズ株式会社
情報解析部長
山本 博之

メタボローム解析は、これまで比較的少数検体を用いた基礎研究において利用されることが多かったが、近年ヒトの臨床研究において数百から数千検体のサンプルを対象にしたコホート研究などで利用され始めており、大規模な試験におけるメタボローム解析のニーズが増してきている。
質量分析を用いたメタボローム解析では、装置間差や測定中に何らかのトラブルで測定が中断した際に感度が変化する等の問題があり、データの統合を前提として数百から数千のサンプルのメタボローム解析を行う際には、いくつかの工夫が必要となる。多検体のメタボローム解析における問題点とその解決法の代表的なものとして、2011 年の Dunn らによるガスクロマトグラフィー-質量分析計と液体クロマトグラフィ-質量分析計を用いた血液サンプルでの報告 (Dunn et al., Nature Protocols, 6, 1060-1083 (2011))、2013 年の Want らによる組織サンプルでの報告 (Want et al., Nature Protocols, 8, 17-32 (2013)) がある。
弊社では、キャピラリー電気泳動-質量分析計(CE-MS)によるメタボローム解析を用いた受託解析サービスを提供しているが、CE-MS を用いた大規模なサンプルを対象としたメタボローム解析を用いたコホート研究として、2012 年から山形県鶴岡市において慶應義塾大学先端生命科学研究所を中心に行われている「鶴岡みらい健康調査」がある。その 1 つの成果として、Harada らは Coefficient of Variation (CV)を基準に CE-MS と他のプラットフォームで測定された共通の代謝物における再現性が、同等かそれ以上だと結論付けている(Harada et al., PLoS One, 13(1), e0191230 (2018))。
本講演では、これまでに報告されている大規模なメタボローム解析における問題点と解決策を紹介すると共に、CE-MS を用いた大規模なメタボローム解析における弊社での取り組みを紹介する。最後に創薬の応用への可能性についても議論したい。

大規模イメージングデータを用いた生命システムのデータ駆動モデリング

理化学研究所生命機能科学研究センター
発生動態研究チーム チームリーダー
大浪 修一

顕微鏡ライブイメージング技術の近年の著しい発展により、分子から細胞、器官、個体までの様々なスケールで、生命現象に関する４次元（3 次元＋時間）の動画像データの取得が可能となった。細胞や器官、個体などの４次元の動画像データに画像認識技術を適用することにより、細胞の位置や形態、遺伝子発現量の変化などの生命現象の動態を定量的かつ高スループットに計測することが可能である。このようにして取得した大規模な画像データ・生命動態データの活用により、近年、動的システムとしての生命の理解を目指すイメージングデータ駆動型の生命科学研究が可能になってきている。本講演では、イメージングデータ駆動型の新しい生命科学研究の展開を、我々の研究成果を具体例に紹介する。

我々は、顕微鏡ライブイメージング技術と画像認識技術を融合して、線虫 C. elegans 胚の細胞核の４次元の分裂動態を自動的に計測する装置を世界に先駆けて開発した。この装置を用いて、遺伝子ノックダウン胚の細胞核分裂動態を計測した大規模データコレクションを構築し、イメージングデータ駆動型解析の研究開発を行っている。これまでに、表現型発現の揺らぎを利用して、発生プロセスの中で発現する様々な形質間の因果関係を導出する方法や、表現型発現の相関の外れ値を利用して、形質間の因果関係の分子機構に関与する遺伝子群を導出する方法等の開発に成功している。また、これらの形質発現因果関係ネットワークとオミックス解析より得られた分子ネットワークを統合解析するための可視化解析システムも開発している。

これまでに開発したデータ駆動型解析を哺乳類胚やオルガノイドに適用するために、現在我々は光シート顕微鏡を基盤とした専用ライブイメージングシステムの開発を行っている。当システムを利用した哺乳類胚・オルガノイドを標的としたイメージングデータ駆動型解析により、創薬研究に新しい可能性を切り開くことを期待している。

本講演では更に、イメージングデータ駆動型の新しい生命科学の推進を目的に世界規模で進んでいるバイオイメージデータのデータ共有連携や、機械学習を活用した細胞状態の推定、Hi-C データ解析等、我々のその他の研究の最新の成果も紹介し、動的システムとしての生命の理解を目指す生命科学のデータ駆動解析の今後の展望について議論したい。

医用画像データと画像解析技術による知見抽出

キヤノンメディカルシステムズ研究開発センター
臨床アプリ研究部長
坂口　卓弥

レントゲン写真、超音波動画、MRI 画像などの医用画像は長らく、定性的に経験則により、「画像所見」を「読影」するために使われてきていた。

近年のデジタル技術の発達はこれらの定量化を促進し、例えば疾患形状測定や性状変化量検出などを自動で高速に処理できるようになってきており、医用画像は定量的バイオマーカーとしての役割に重きを置かれるように変化してきている。

さらに最近は、画像を診療判断に役立てる試みとして、画像から「知見」を抽出する画像解析技術開発が盛んにおこなわれている。

本講演ではこうした技術動向の事例を共有させていただき、創薬をご専門とする皆様方と新しい知見を探究できることを期待する。