



システム癌
新次元

がんシステムの新次元俯瞰と攻略

文部科学省科学研究費補助金 新学術領域研究

(研究領域提案型)(複合領域：4701)

Newsletter

No. 19



CONQUERING CANCER THROUGH NEO-
DIMENSIONAL SYSTEMS UNDERSTANDING

2019年4月

公募研究

データ同化及び深層学習を利用した
発癌予測モデルの構築と実装

公募研究 A01-9-18 研究代表者

北海道大学大学院医学研究院社会医学分野 教授 西浦 博



がん患者発生をコンピュータ内で再現したい

これまで様々な研究を通じて癌の発生メカニズムが理解されてきました。癌細胞は遺伝子の異常が起こることで発生しますが、喫煙や感染症、遺伝的リスクなど様々な要因によってその異常が誘発されます。蓄積された研究の成果として、リスク要因と特定部位の癌の間に一对一の関係があるものも少なくありません。例えば、肺がんの最も良く知られているリスク因子は喫煙です。また、子宮頸がんはヒトパピローマウイルスの感染に引き続いて発生します。それは集団レベルでの患者データと健常者データの比較に基づく疫学研究によっても明らかにされてきましたし、また、分子生物学的な研究を通じても癌発生リスクの特定とその生物学的なメカニズムの解明が行われています。

しかし、そういったリスク因子の特定は予防活動や早期治療などには繋がりがつつあるものの、がん発生の予測には十分に使われてきませんでした。どのくらいの喫煙をどれくらいの期間だけ続けた者が何名いるのか情報がわかれば、本来的には未来の肺がん患者数はわかるはずです。その理由の1つとして、がん発生の自然史

(生物学的な癌発生までのプロセス)について十分に捉えたような予測の枠組みができていないことが挙げられます。

例えば、がんの誘発待ち時間(induction period)と呼ばれるものが典型例です。これを理解するために喫煙と肺がんの関係を考えてみましょう。今日まで、日本における喫煙率は大幅に減少しました。戦後約20年間の男性では8割を超える男性が喫煙者である頃もありましたが、現在は2割を下回るまでに至っています。このような劇的な喫煙率の減少は、将来的には肺がんの罹患率や死亡率を劇的に下げるであろうことが期待されています。特に喫煙に密接に関係していると言われる肺の小細胞がんや扁平上皮がんの罹患率は極端に下がる可能性が高いです。なのに、現在の肺がんの罹患率や死亡率は減少どころか上昇する傾向にあります。もちろん、高齢化に伴って人口中で肺がんのハイリスク者が増えている影響もあると思いますが、根本的な疑問として、なぜすぐに肺がんが減り始めないのでしょうか。それは、単純に時点での喫煙率ではなくて、一定期間を継続して喫煙することによって、一定期間が起っており、それまでに何十年という時間がかかるから、という事実のためだと考えられます。つまり、当然のこ

とではあるのですが、喫煙を開始してすぐに肺がん罹患するヒトは極めて稀であり、何年も何年も持続して暴露をした末に遺伝子異常が起こるのです。これを予測する立場から考える必要があります。すなわち、リスク因子の変動による未来の癌を予測するには、がんの自然史やその他の人口動態など集団生物学的なプロセスを十分に理解した上で、その過程を直接的にモデル化した上でコンピュータ内で確実に再現し、それを実行に移す必要があるのです。

言い換えれば、ヒトが生まれてから年齢を重ねて発癌を経験するプロセスとは数理的に記述可能なシステムそのものです。私が公募研究として本領域で実施している研究では、時刻あるいは年齢と共に変動する発癌リスクに暴露されつつ加齢し、その過程の一部のイベントとして癌が発生することを理解しつつ、その上で数理的に発癌過程を捉えたモデルを構築して、妥当性と信頼性に優れた疫学的ながん発生の予測を実現することです。プロトタイプモデルが完成すれば、ヒトの遺伝多型情報や生活習慣の詳細も加味したテイラーメイドでの癌リスク定量化を実現するマイクロシミュレーションも夢ではないと考えています。

モデル推定の一手段として機械学習を利用する

別の事例として、胃がんがHelicobacter pyloriという細菌感染に誘発されることに注目しましょう。今日までに抗菌薬を利用したH.pylori除菌療法が確立されつつあり、ひとことと比べれば胃潰瘍や胃がんの患者が減り始めているのではないかとささやかれています。ピロリ菌という発癌の暴露自体を除去あるいは軽減しようという対策の論拠は論理的に極めて明快ですし、皆さんも「胃がんが防げるなら」と考えて副作用がゼロではない除菌療法をガマンして実施する決断をされることだと思います。今後ピロリ菌の暴露によって引き起こされる癌は長期的に減少することは間違いないでしょう。

問題はその胃がん減少を如何にして確実に予測するかです。図をご覧ください。これは状態空間モデルと呼ばれる統計学的なモデリング手法を活用して胃がんの発生を予測する方法を模式図にし

たものです。ある時刻でピロリ菌の感染が発生しますが、一部は感染が不顕性のまま胃に何も病変を起こさずに済むとされています。残りの者は、前癌病変とよばれる急性胃炎や胃十二指腸潰瘍などを経験します。いわゆる「胃がある」という経験の最初の病気に相当します。それに引き続いて慢性萎縮性胃炎が発生します。胃酸を抑える薬を内服したり、胃粘膜を保護する薬を内服したりして治療を繰り返します。それが何十年と引き続くと、最終的には胃がんが発生します。このプロセスはコンパートメント型モデルと呼ばれます。それぞれの進行に要する遷移率は文献的に既知な場合が多く、十分にこのような単純なモデルを利用しても胃がん予測が実装可能です。同モデルを微分方程式系で書き、それを解くことによって積分方程式を用いて時刻あたりの胃がん発生数の予測された数をモデル化することができます。これまでに癌登録データや癌死亡者数に関する過去のデータがありますから、

それに適合させることによって不明パラメータの推定値が得られるのです。

予測を本当に実施する際の推定では計算能力の優れたコンピュータを利用します。様々な地域や性・年齢などの属性別で複数の観察データにモデルを適合しないといけないためです。上述の状態空間モデルではベイズ推定の仲間の1つである従来の粒子フィルタを用いたデータ同化を実施しています。また、むやみに粒子フィルタを利用した推定を実装することが困難なので、機械学習のテクニックの1つであるパラレルテンパリングという技術を利用します。

まずは、明確な暴露に引き続いて起こる肺がんや胃がんを対象にモデル化と予測を成功させようと努力しています。妥当性と信頼性の比較検討を行い、方法的基盤を盤石なものにして、より精度の高い予測の実現に向けて邁進していきたいと考えています。

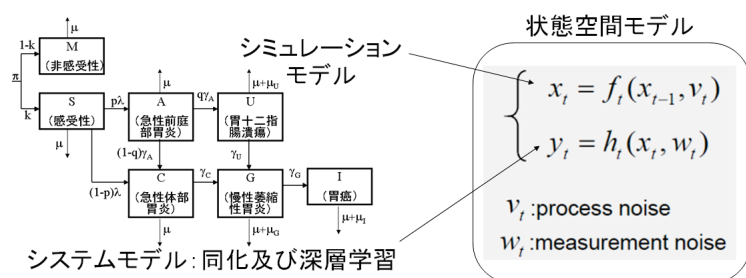


図: 状態空間モデルを活用した胃がん予測研究の模式図

公募研究

がんシステムの新次元理解に向けた プロテオゲノムビッグデータ解析基盤の構築

公募研究 A02-1-18 研究代表者

名古屋大学大学院医学系研究科 准教授 松井 佑介



データ科学駆動型のがん研究

現代のがん研究では、各病態において分子レベルでの生体メカニズムがどのように関わっているのかを明らかにすることが一つの重要な課題となっています。私たちの研究グループは「データ科学」と呼ばれる、統計科学、情報科学、そして生命科学を融合させたデータ駆動型アプローチによってがんの発症、進展、治療標的の同定に向けた技術開発や研究に取り組んでいます(図1)。

がんと次世代シーケンサー

2000年代頃開発された次世代シーケンサー(Next Generation Sequencer; NGS)と呼ばれる、一度に大量の塩基配列

(ATGC配列情報)を読み取ることのできる技術の台頭によって、細胞中のATGCという4つの文字列(塩基対)からなる30億塩基対のDNA情報(ゲノム)が取得することが可能となり、またそれらを応用することによって、遺伝子発現を制御する因子(エピゲノム)や、ゲノムの一部が転写されて産み出される転写物(トランスクリプトーム)といった分子レベルでの膨大な生命情報の取得が可能となり、セントラルドグマにおけるがん発症に関わる数多くの分子機構が世界中の研究者によって明らかにされてきました。

次世代シーケンサー×質量分析器から 迫るがん研究

一方で、トランスクリプトームはさらに翻訳

されることによりタンパク質が生成され、さらにリン酸化などの化学修飾が伴うことで、生体内ではさまざまな機能が制御、調節、維持されています。しかし、これらを分子網羅的なレベル(ゲノムワイド)に調べるためには、NGSとは原理が異なる質量分析器と呼ばれるテクノロジーをベースとしたプロテオーム情報の取得が必要でした。これまでの研究ではNGSによるゲノムやトランスクリプトーム情報と質量分析器によるゲノムワイドなプロテオーム情報を積極的に結びつける研究(プロテオゲノミクス)は十分とは言えない状況でした。

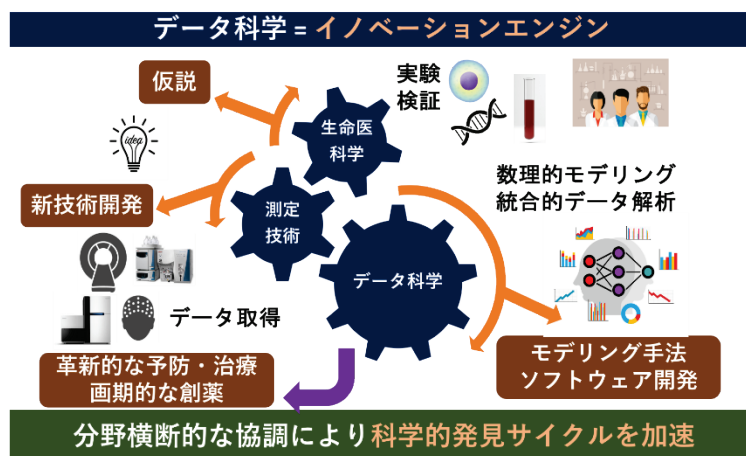


図1: データ科学駆動型アプローチの概要

近年、液体クロマトグラフィー質量分析法 (Liquid Chromatography / Mass Spectrometry; LC-MS/MS) の飛躍的進歩により、NGSを元に解明されてきたがん病態に関わる分子メカニズムの破綻が、実際のタンパク質レベルの生命現象にどのような影響を与えているのかを網羅的に測定することが技術的に可能となったことで、世界中で多数の大規模臨床プロジェクトが立ち上がっており、がんのプロテオゲノム研究は新しいフェーズに入っています。

データ科学によるプロテオゲノミクスへの挑戦

しかし、NGSとLC-MS/MSという原理的に

異なる測定技術から生み出されるデータを解析することは、異なる情報解析技術が必要であることを意味します。そのため、異種の測定プラットフォームを情報工学的にシームレスに繋ぐ技術開発が必要であることは勿論のこと(図2)、さらに生命医科学的な知見から想定される仮説をプロテオゲノムのレベルで立証するためのデータ科学的なモデリング技術が必要となります(図3)。

私たちの研究グループでは、NGSにより得られる分子情報をLC-MS/MSによるタンパク質同定へ組み込むための技術開発や、ゲノム異常がタンパク質の実際の働きにどのような影響を与えるかを推定する

モデリングする技術の開発を進めており、NGSデータにLC-MS/MSによる新しい次元のデータを取り込んだデータ科学駆動型アプローチの基盤構築を通じて、がん発症および進展に関わる分子機序解明に向けた生命医科学研究に貢献してこうと考えています。

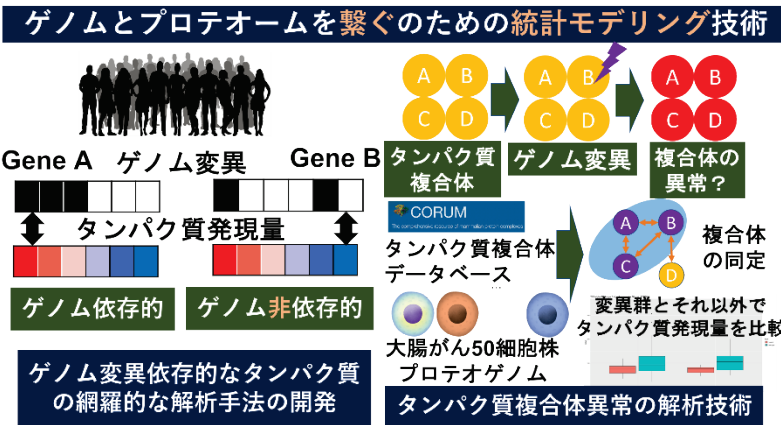


図2: NGSとLC-MS/MSを「繋ぐ」情報解析基盤の構築

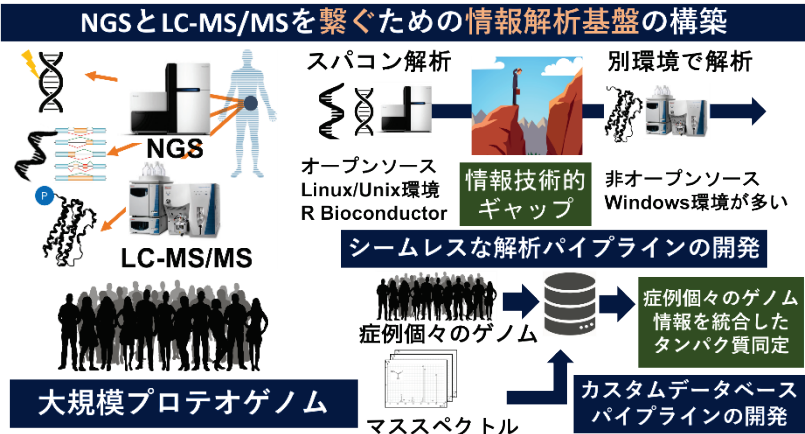


図3: ゲノムとプロテオームを「繋ぐ」ための統計的モデリング手法

Information

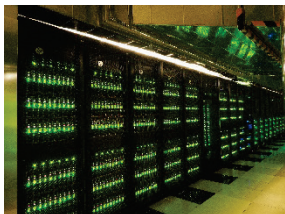


Human Genome Center

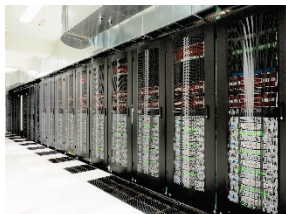
東京大学医科学研究所

ヒトゲノム解析センタースーパーコンピュータ

システム癌新次元で用いられるデータ解析ソフトウェアは、ヒトゲノム解析センターのスーパーコンピュータで動いています。スーパーコンピュータはどなたでもご利用になれます(有償)。



大容量のディスク装置



高密度の計算機

supcom.hgc.jp



システム癌
新次元

<http://neosystemscancer.hgc.jp/>

新学術領域研究「システム癌新次元」ニュースレター No. 19

発行日★平成31年4月22日(初版)

発行★がんシステムの新次元俯瞰と攻略

領域代表者★宮野 悟

- 東京大学医科学研究所 ヒトゲノム解析センター DNA情報解析分野
- 〒108-8639 東京都港区白金台4-6-1
- TEL: 03-5449-5615 FAX: 03-5449-5442
- E-mail: miyanolab-jimu@edelweiss.hgc.jp