システムがんNewsletter No.１１

２０１４年５月

システム的統合理解に基づくがんの先端的診断、治療、予防法の開発

文部科学省科学研究費補助金 新学術領域研究（複合領域：4201）

公募研究の紹介

第二世代モチーフ解析法に基づく がん細胞に特異的な転写制御経路の発見

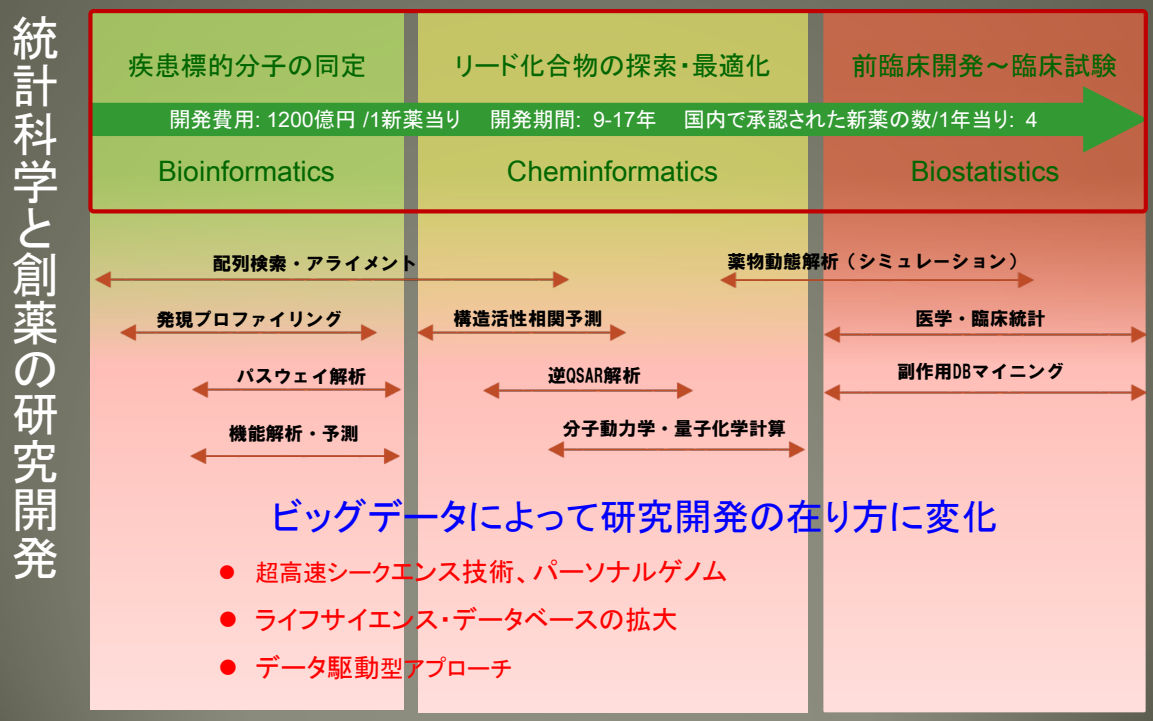
吉田 亮

システム研究機構 統計数理研究所 准教授

私は統計屋です。ベイズ統計学と機械学習の方法を用いてライフサイエンス分野の様々な問題に取り組んでいます。細胞・分子イメージングの画像情報処理、神経科学、量子化学計算による分子設計など、統計科学独自の視点からユニークな応用研究を戦略的に開拓していくことが私の行動原理です。本稿では、統計科学にもとづく医療創薬の研究開発を俯瞰しながら、「システムがん」の研究活動を紹介します。

現在の新薬の研究開発の主な流れは、(1)疾患標的分子の同定、(2)標的分子の働きを活性あるいは阻害するリード化合物の探索・最適化、(3)前臨床・臨床試験という三つのプロセスで構成されています。統計による揺らぎはありますが、1新薬当たりの研究開発費は約1200億円、開発期間は10年と言われています。統計科学は個々のプロセスに対して、バイオインフォマティクス、

（次ページへ続く）

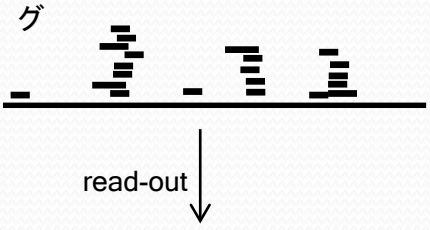


ケムインフォマティクス、バイオスタティスティクスという呼称のもと、テクノロジーと人材供給の両面において非常に重要な役割を担っています。近年世界的に加速するライフサイエンス・データの大規模化と多様化は、医療創薬の研究開発の在り方に大きな変化を促しています。超高速シークエンス技術の普及により世界中の研究拠点でペタバイト級のゲノムデータが驚異的な速度で生成されています。リード化合物の探索では、数百万件の化合物データにもとづく候補化合物の活性予測、さらにコンピュータシミュレーションにもとづく低分子化合物の分子設計など、様々な局面で統計科学のテクノロジーが威力を発揮します。ライフサイエンスの諸分野では、革新的な計測技術の出現によりデータ駆動型アプローチへの期待が高まっています。しかしながら、同時にデータの大規模化による弊害が顕在化してきました。これは統計屋として非常に残念なことですが、あまりに急速な計測技術の進歩に対して、データ解析手法の整備が追いついていけないという状況がここ数年続いています。

ここでその一例として、本グループの「システムがん」の研究を紹介します。超高速シークエンス技術がデータ量の爆発的増大をもたらしたことで、これまでのデータ解析手法に抜本的な見直しが求められているというお話しです。モチーフ探索、すなわちDNAの塩基配列に埋め込まれた短い保存配列を発見する問題は、バイオインフォマティクス創世期からの研究対象であり、これまでに非常に多くの解析方法が提案されてきました。しかしながら、近年のデータの大規模化に際し、既存の方法とそのソフトウェアは機能を果たせなくなってきました。既存法の設計概念には、長さ10^3bp、配列数のオーダで10^2程度のデータサイズしか想定されておらず、計算量と検出力の両面でデータ量の増加に全くスケーリングしません。例えばChIP-seq法によるプロモータ解析では、測定値のピーク検出により転写調節因子の結合部位の候補として約10^4 本の断片配列を絞り込み、これにモチーフ探索を適用して転写調節因子の認識配列を同定します。既存の方法では、この問題を解くことができません。そこで世界中のバイオイ

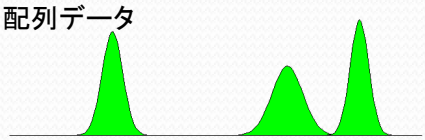
ChIP-seq解析による転写調節因子の予測

1. ChIP-seq解析: 免疫沈降で回収したDNA断片の配列をシークエンシング

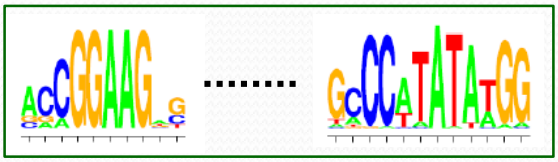


2. 転写調節因子の結合部位を絞り込む

10,000～60,000本、長さ200～500bpの配列データ

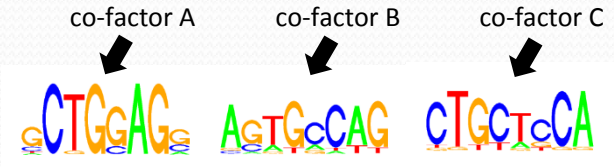


3. モチーフ探索を実行し、保存パターンを網羅的に検出



4. 転写因子データベースの類似検索

転写共役因子のアノテーション



ンフォマティシャンが競って新しい解析技術を開発しています。本グループには、ベイズ統計学、文字列解析、モンテカルロ計算の最先端技術を有する研究者が参画しています。ミッションは、世界最高性能のモチーフ探索アルゴリズムを完成させ、それをがんゲノムの研究で実用化することです。

近年「ビッグデータ」という言葉が巷に飛び交っています。ライフサイエンスの世界では、これに「医療」や「ゲノム」という枕詞が付けられます。しかしながら、統計科学の研究者の中には、そこに目新しいコンセプトを見出せず、ブームと距離を置こうとするものも少なくありません。大量のデータから有用な法則性を見出し、理論が確立されていない複雑な現象を理解・予測することは統計科学がずっとやってきたことです。そもそもデータがビッグでなければ、統計科学は必要ありません。近年の多分一過性のブームのおかげで、統計科学に注目が集まることは非常に嬉しいことです。私のところにくる共同研究のオファー数はここ数年で格段に増え、とても忙しくなりました。数年後には統計科学の進歩がライフサイエンスに追いつき、がん研究の世界でビッグデータなんて言葉を使うものが誰もいなくなり、穏やかな研究生活に戻れることを心から願っています。

経時測定データの解析

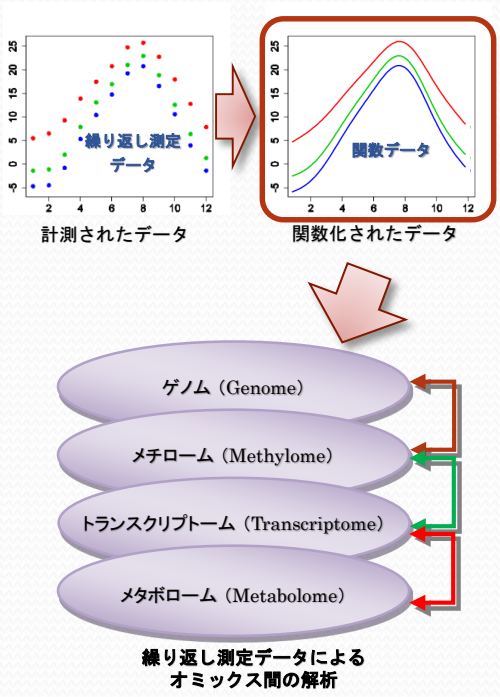
松井 秀俊

九州大学 数理科学研究院 助教

近年の計算機の性能や計測・測定技術の急速な進展によって、大量かつ複雑な形式をもつデータが容易に得られるようになりました。生命科学の分野もその例に漏れず、日々大量のデータが計測されています。私が専門としている統計科学では、このようなデータから可能な限り精度よく情報を抽出するための理論や方法論についての研究が行われています。

計測データの一例として、時間経過に伴う変動を調べるため、複数の時点で繰り返して計測値を取得するという形式のデータがあります。このような形式のデータは「繰り返し測定データ」や「パネルデータ」などと呼ばれており、私が研究対象としているデータの形式の一つです。繰り返し測定データは、その構造から解析が困難になる場合があります。例えば、個体ごとに計測時点が異なったり、計測の失敗などにより計測時点数が異なったりする場合などです。このようなデータに対して古典的な多変量解析手法を直接適用しても、適切な結果を得ることができません。

私は、このような形式のデータの構造を捉え、適切に解析するための方法について研究しています。そのための方法の一つに、関数データ解析と呼ばれる方法があります。これは、繰り返し測定データを時間の関数とみなして関数化処理し、得られた関数集合を改めてデータとして扱おう、というものです。データを関数化することで、上で述べた計測時点数の不一致などの問題点を解消することができます。そして、関数データ集合に対して、主成分分析や回帰分析といった多変量解析手法を適用するためのモデリング手法について現在研究しています。



繰り返し測定データによるオミックス間の解析

私は、このような統計的モデリング手法を、生命分子に対する網羅的解析、すなわちオミックス解析に適用することで、複雑な構造を内包した生命活動のシステムを明らかにすることを目標としています。特に、オミックス間の解析の一つとして、転写産物と代謝産物の生成の関係を統計モデルで表すことを目指しています。転写産物と代謝産物の間には非常に複雑な経路があることが知られているため、様々なアプローチからモデルを特定する必要があると考えています。

生命科学の研究に携わって間もない私ですが、これまでに身に着けた統計学に関する知識や技術を生命システムの解明に少しでも貢献できればと考えています。

ゲノム領域特異的ヒストン修飾の変更技術による新たながんエピゲノムのシステム理解

永瀬 浩喜

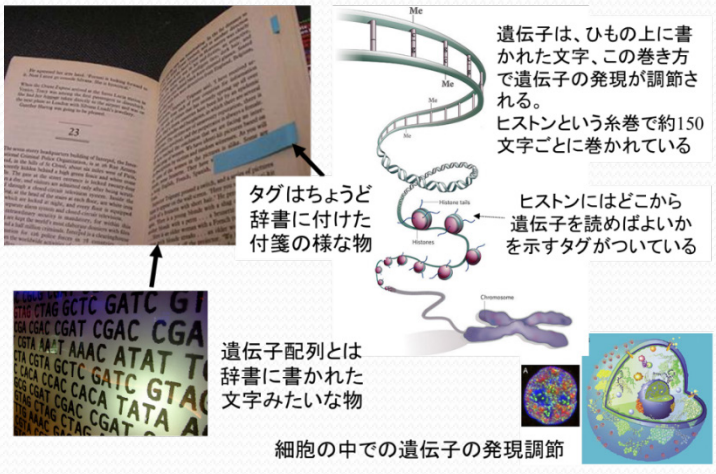
千葉県がんセンター研究所 所長

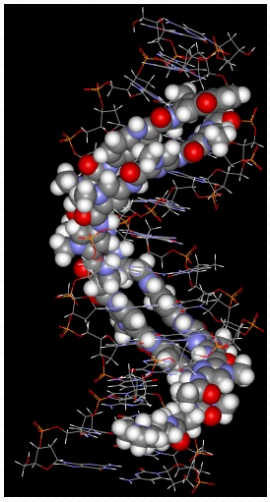
人のからだは、生まれ持った体質と生まれた後の環境因子、生活習慣や食べ物、周囲の汚染物質等により影響を受け、病気の原因になったり、体調の変化を生むことがあります。がんに関しても、様々な物質が細胞に影響をあたえ、がん化が進行していくことで生じるものと考えられています。この様々な物質の影響は遺伝子に直接変化を与える事もあるのですが、直接変化を起こすよりも間接的に遺伝子の出かた、いつどのようなときに遺伝子を出すべきかという調節の機構に影響を与える事の方が多いことが解ってきました。

これは下図に示すようにエピジェネティクスと呼ばれています。ちょうどDNAは糸のようなものなのですが、糸の上に文字が書かれたヒトの設計図となっているわけです。この糸の巻き方で遺伝子の出かたが調節されるわけです。ちょうど辞書を読むときに目次や付箋、横に着いた色分けや順序割などで場所が解りやすくして、読むべき場所を探し当てるようなことに似ています。このタグ（付箋）の位置が変えられると間違えて読んでしまうわけです。色々な環境因子はこの付箋の位置を変えてしまうことがあります。そして病気やがんに繋がっていくわけです。

我々の研究室では、このDNAの巻き方を変える薬を開発しました。辞書の文字を検索して見つけて、その所に付箋を置くような薬です。右下の図に、その薬の構造を解析したものを示します。図の格子状で示しているDNAは2重らせん構造を取っていますが、この螺旋には2つの溝（幅が狭い溝と広い溝）が生じます。この幅の狭い方の溝に入り込み一定の遺伝子の配列を見つけて結合する化合物（右下の図で円形のものの連なりで示してあります）を合成できるようにしました。この化合物は、決められた遺伝子の場

DNA配列以外の遺伝子の調節機構（エピジェネティクス）





（前ページからの続き）

所に糸の巻き方を変えるお薬を運ぶ役目を果たしてくれます。このことで遺伝子の発現を変えるエピジェネティックな変化がどのようにヒトの細胞に影響を与えているのかを研究したり、色々な環境物質が変化させたこの遺伝子の調節機構を元に戻すことが出来ないかと研究を進めています。最終的にはこの薬でがんなどの病気を治すことができるのではないかと考えています。

このくすりを本当に人に使用するためには、ヒトの遺伝子の配列を知ったうえで薬をデザインしなければなりません。人の遺伝子の配列は一人ひとり少しずつ違いますし、遺伝子は30億の文字の配列からなります。この配列を調べてどのようにデザインをすればよいかを確認するためにこのシステムがんのプロジェクトでスーパーコンピューターを利用して計算することでくすりをデザインする方法を研究させていただいています。

このくすりを創る技術は、文部科学省の補助で、京都大学との共同で開発され特許化されています（特許第4873510 号）。また京都大学ではこの薬で、遺伝子の発現を調節することで、マウスの皮膚の細胞からiPS細胞を作りだすことにも成功しています。

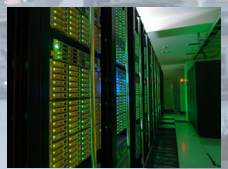
この全く新しい薬による技術手法によってがんエピゲノム（がんのエピジェネティックな変化）によるがん化のメカニズムを解析し、がんの新しい治療薬が創れると確信し、研究を進めています。これらの研究から世界初の新たな成果が日本で生み出され、社会、医療へ多大なる貢献を来すと期待しています。

Information

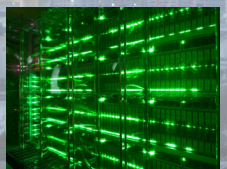
Human Genome Center

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
http://supcom.hgc.jp

システムがん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

文部科学省新学術領域研究
システムがん
システム的統合理解に基づくがんの先端的診断、治療、予防法の開発

オフィシャルWebサイト
http://systemscancer.hgc.jp

システムがんの研究内容、構成研究者情報、発表論文、研究集会などの詳しい情報はこのWebサイトでご覧ください。

twitter @SystemsCancer

システムがんの研究成果、世界のがん研究、研究拠点、学会情報など最先端な話題をつぶやいています。

新学術領域研究「システムがん」ニュースレター No.11
発行日★平成26年5月19 日（初版）
発行★システム的統合理解に基づくがんの先端的診断、治療、予防法の開発
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

編集★サトウアユ