02 医療におけるビッグデータ利活用
―精神神経系疾患の診断系の開発を中心として―

応用 一般

石井一夫（東京農工大学）

医療におけるビッグデータの最近の状況

米国の状況

BD2K

医療分野では，医療情報などを活用して医療に還元する動きは従来から存在した．しかし，高速自動シーケンサーが実用化した 2005 年頃からヒトゲノム解析にかかるコストが急速に減少し，大量のゲノムデータが産生されるようになった（図 -1）．このため，個人別にゲノム解読を行い疾患の診断や治療に活用する個別化医療が現実のものとなってきている．

ビッグデータは従来型データベース管理ツールやデータ処理ソフトで処理できないほど，大量で（Volume），多種類で（Variety），高頻度に変化する（Velocity）データであるが，生命科学分野ではゲノム配列データほどそれに似つかわしいものはないと思われる．

したがって医療分野でビッグデータ利活用が話題になるのは必然と思われたが，究極の個人データである医療データを，ビッグデータ利活用という視点で語られることを医療関係者や生命科学研究者が嫌ったのか，その動きは鈍かった．しかし，ここ数カ月で（2014 年前半），次世代シーケンサーのデータを活用した臨床研究や，医療におけるビッグデータ，ハイパフォーマンスコンピューティング（HPC）の利活用に関するセミナや研究会が増えてきているように感じられる．

これは，昨年（2013 年），アメリカ国立衛生研究所（NIH）に，医療分野におけるビッグデータの利活用を図るために設立された BD2K（NIH Big Data to Knowledge）イニシアチブなどが，ワークショップやセミナなどで情報発信を始めたことが 1 つの発端であると考えられる．

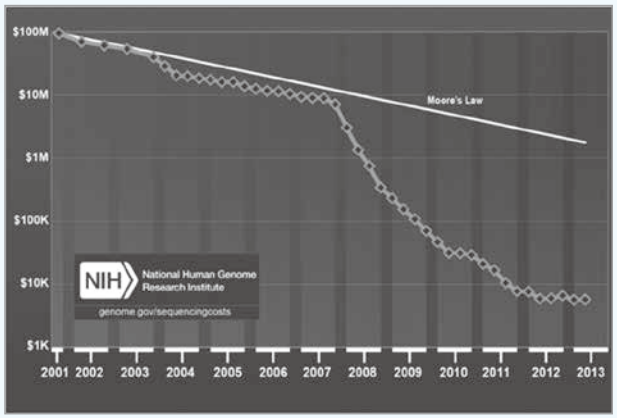

図 -1　ヒトゲノム解読のコスト減少（National Human Genome Research Institute の記事より）

eMERGE

米国では，電子カルテ（EMR）とゲノム情報を統合して，個別化医療に応用しようという動きが活発であり，代表的なものに米国立ヒトゲノム研究所（NHGRI：National Human Genome Research Institute）の資金提供を受けている eMERGE（Electronic Medical Records and Genomics）ネットワークがある（図 -2）．この組織により，2007 年から 2011 年にかけて，ノースウエスタン大学，メイヨークリニック，ワシントン大学 Group Health Cooperative，ヴァンダービルト大学および，Marshfield クリニックの 5 つの機関で EMR による表現型情報の整備が行われてきた．2011 年からは，さらにペンシルバニア子供病院などが加わり，EMR へのゲノム情報（変異情報）の統合が実施されている．

これらの医療機関のほか，ベイラー医科大学，ハーバード大学などでゲノム情報を利用した個別化医療が実施されている．対象とする疾患は，遺伝性疾患やがんが圧倒的に多く，ゲノムに基づく創薬的な

アプローチも試みられている．

たとえば，eMERGEネットワークに参加しているヴァンダービルト大学病院は，PREDICT（Pharmaco-genomic Resource for Enhanced Decisions in Care and Treatment）と呼ばれる薬剤代謝酵素CYPの遺伝的背景に基づいた薬剤投与予測システムを稼働させている．

図-2　eMERGEネットワークに参加している組織（http://emerge.mc.vanderbilt.edu/about-emerge-network より）

メイヨークリニックは，乳がん患者のゲノム配列解析を行いテーラーメードのがん化学療法を実施するBEAUTY（The Breast Cancer Genome Guided Therapy Study）プロジェクトを実施している．

今後，コンソーシアムを形成している比較的大きな医療機関だけでなく，小さな病院でもビッグデータの利活用が大きく進むことが期待される．

■ バイオバンクと国内の状況

バイオバンク

日本国内におけるゲノム解析と医療情報が結びついた動きとしては，特定地域においてがんなどの検体を収集し，医療情報とゲノム情報を統合して集積するバイオバンクを設置し地域医療に貢献しようという試みがいくつかなされている．代表的なものに東北メディカル・メガバンクなどがある．

ゲノム科学におけるビッグデータ分析

■ ゲノムビッグデータ分析の実際

増え続ける情報に対して

実際のデータ分析において，大量データを処理する際に，メモリ不足等の理由で，データベースやデータ分析ソフトが有効に機能しないという場面が頻発するようになっている．このためにファイルを分割して処理したり，より高スペックの解析サーバを導入したりして対応する．定型的な方法はなく，筆者らはいろいろな方法を組み合わせて対応しているが大きく分けると以下の4つの方法を用いている[1)〜3)]．

(1) モンテカルロ法によるシミュレーション[4)]

いわゆる乱数に基づいて，大量データからデータを一部抽出し，元のデータのパラメータを推定する．データの抽出方法により，ブートストラップ（復元抽出），ジャックナイフ（非復元抽出），マルコフ連鎖モンテカルロ法（特定の規則に基づく乱数発生）などがある．

特に，新規の機器を購入したりする必要がなく，すぐに実施できるので，大量データの要約パラメータ（平均や標準偏差，分布など）を知りたいときには頻用する．

(2) 大容量メモリ，メニーコアサーバによるHPCアプローチ

理化学研究所や，遺伝学研究所，統計数理研究所，北海道大学など各拠点となる情報基盤センター大型計算機システムの共同利用により，スパコンを使用する．特にヘビーユーザである場合は，共同利用施

設では効率的，有効的な利用は困難であるので，自前の施設構築を検討するほうが有効かもしれない．

（3）Hadoop を用いた並列分散処理

分散ファイルシステム，分散処理システムとしての選択肢の 1 つに Hadoop の利用がある．特に個人レベル，一研究室レベルでのクラスタによる分散ファイル環境は，経済的にもスペース的にも困難が伴うので，AWS（アマゾン ウェブ サービス）などのパブリッククラウドサービスを用いることが有効であると考えられる．

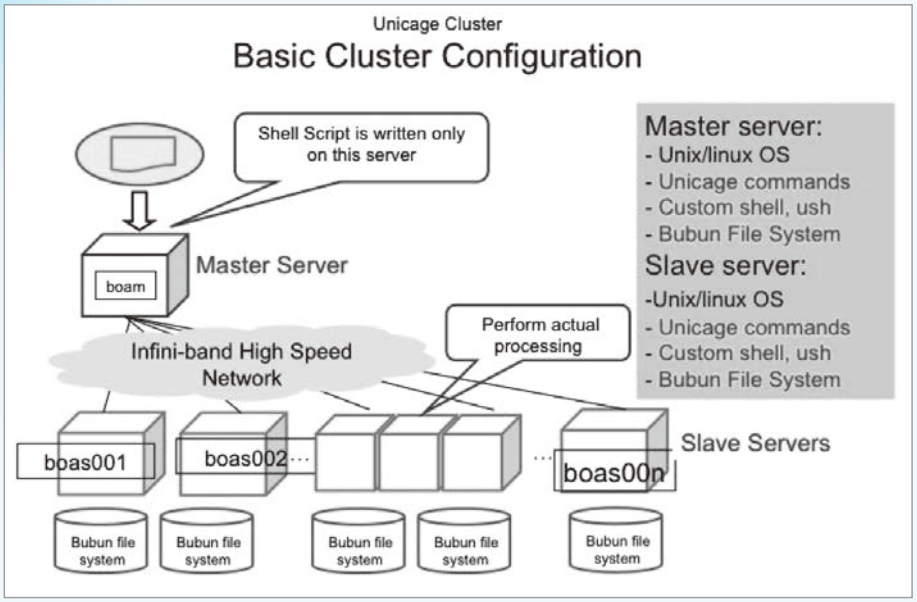

図 -3　シェルスクリプトを用いた分散ファイル・分散処理システム uspBOA（MIT（マサチューセッツ工科大学）セミナ「CRIBB（Computing Resources in Boston and Beyond）」（2013）発表資料より）

AWS においては，クラウド上で AMI（Amazon Machine Image）を用いてデータ解析環境を構築し，Amazon EMR（Elastic MapReduce）を用いて容易に分散処理環境を構築できる．

これらの環境を利用して，Contrail，Coudbrush（以上，ゲノム配列データアセンブリソフト），CrossBow，Myrna（以上，ゲノム配列データマッピングソフト），NCBI BLAST（相同性検索マッピングソフト），R/BioConductor（生物学統計データ解析ソフト）などのオープンソースソフトウェアが利用できる．

Hadoop の活用法については本特集の水丸の稿，パブリッククラウドサービスについては，本特集の吉荒の稿も参照いただきたい．

（4）Hadoop を用いないシェルスクリプトによる並列分散処理

シェルスクリプトは，比較的簡単に記述でき，メニーコアの HPC や，コンピュータクラスタで容易に分散ファイルや分散処理システムを構築できる．

筆者らは，本特集の當仲との共同研究により，シェルスクリプトによる分散ファイルシステム uspBOA によりヒトゲノムの次世代シーケンサーデータのクオリティチェックのリアルタイム処理を達成した（図 -3）．Hadoop に比べて容易に構築できるシステムとして興味深いものであると考える．その他の手法については當仲の稿をご覧いただきたい．

ゲノム科学におけるビッグデータ分析の応用事例として―精神神経系疾患―

精神神経系疾患を巡る現状

患者数の推移

近年の国内経済状況の長期悪化，および社会状況の変化により，精神神経系疾患に罹患する患者数が急激に増えている．2011 年の第 19 回社会保障審議会医療部会資料の「患者調査」に基づく資料によると，1999 年から患者数は急激に増加しており，2008 年にはその患者数は，がん患者の 2 倍以上で，糖尿病よりも多くなっている（図 -4）．このなかでも，1996 年には 43.3 万人であったうつ病の気分障害の総患者数は，2008 年には 104.1 万人と 12 年間で 2.4 倍に増加した．「患者調査」は，医療機関に受診している患者数の統計データであるが，うつ病患者の医療機関への受診率は低いことが分かっており，実際にはこれより多くの患者がいることが推測される．

国内の自殺者数は，1998 年以降，継続して毎年

3万人を超えており同資料においても3万1千人で，がん，心疾患，脳血管疾患，肺炎，老衰，不慮の事故に続き，死因の第7位となっており，社会的な問題となっている．

自殺の大きな要因としてあげられるのは，うつ病などの精神疾患との因果関係である．Thomas E. Ellis らによれば，自殺既遂者の95%は何らかの精神疾患を患っていて，その大半が治療可能だったという研究結果もある．日本の警視庁発表ではうつ病による自殺が27.6%と最も多い．

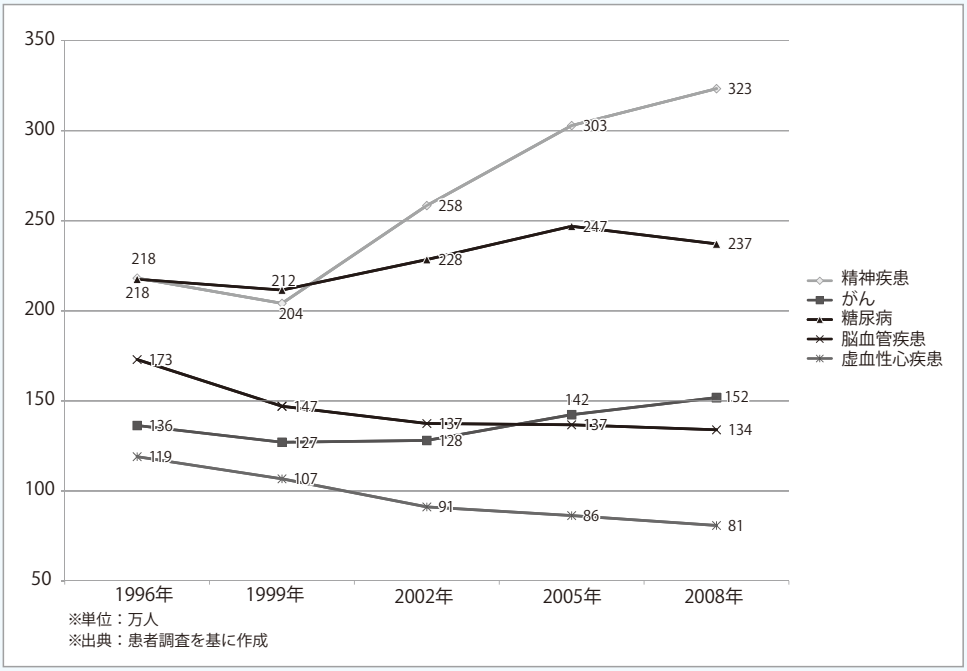

図-4　疾病別の医療機関にかかっている患者数の年次推移資料：患者調査より

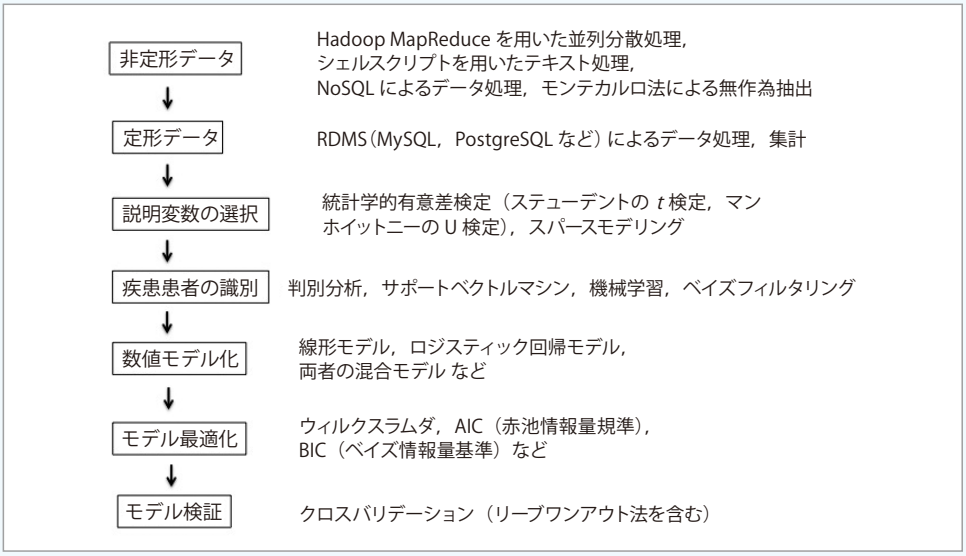

図-5　データ分析のワークフロー

精神神経系疾患の診断の状況

一方で，うつ病および総合失調症などの精神神経系疾患で最も広く用いられる診断基準は，アメリカ精神医学会による精神障害の診断と統計の手引き改訂4版（DSM-IV-TR）と，世界保健機関の疾病および関連保健問題の国際統計分類（ICD-10）である．うつ病および総合失調症の原因はいまだ不明であり，その診断は，上記診断基準により，もっぱら医師による患者からの愁訴に基づく問診により実施されてきた．科学的エビデンスによる診断法の存在しない，いわゆる"アンメットメディカルニーズ"（未充足の医療ニーズ）の疾患である．

精神神経系疾患診断系構築のためのデータマイニング

筆者は，精神神経系疾患の診断系を確立することを目標に，徳島大学大学院ヘルスバイオサイエンス研究部神経情報医学部門との共同研究により精神神経系疾患のゲノムレベルのデータを用いたデータ分析を開始した．図-5にデータ分析の大まかなワークフローを示す．

たとえば次世代シーケンサーデータは，大量のテキストデータであるので，比較的少量のデータであれば簡単に処理できるが，数十ギガバイトを超えるデータである場合は，従来の解析システムではデータ処理が困難になる一種の非定形データと呼ぶこと

が可能である．これらは必要に応じて分散処理などが必要になってくる．

通常のデータベースや，統計解析環境の R で処理できるレベルまで，データを小さくすることができれば統計的な解析が可能になる．

(1) 説明変数の選択

すでに述べたように次世代シーケンサーや，マイクロアレイによる網羅的発現解析や網羅的ゲノムメチル化解析を用いた場合，発現変動遺伝子や，ゲノム DNA メチル化部位など，大量の説明変数が得られる．

これらの説明変数は，負の二項分布またはポアソン分布をすることが知られている．したがって通常のステューデント *t* 検定を用いずに，マンホイットニーの U 検定などのノンパラメトリック検定を用いて選択することが多い．

(2) 識別方法と説明変数の最適化

複数の説明変数による疾患，患者の識別には，多変量解析（重回帰分析，判別分析，クラスタ分析），サポートベクトルマシン，機械学習（自己組織化マップ（SOM）など），ベイズフィルタリング，ランダムフォレストなどの利用が考えられる．

説明変数の選択は単純に，マンホイットニーの U 検定などの *p* 値などだけでは不十分で，いろいろな検討が必要である．検討方法には，総当たり法，最良優先探索，焼きなまし法，遺伝的アルゴリズムなどがある．

これらの説明変数の選択には統計学的な方法だけでなく，説明変数の重み付けや，因果関係の解釈などのために，生物学的データの解釈は必須である．候補となるマーカを染色体上にマッピングし，その妥当性を検証するなど，常に生物学的評価を意識しないと正確な判断を見誤ることもあり得る．データ分析には単に情報処理能力や統計処理能力だけでなく，対象分野の専門知識がどうしても必要である．

(3) 得られた数理モデルの検証

診断法の正確性の評価は，感度および特異度が用いられる．感度は，陽性と判定されたもののうちの真の陽性の割合で，特異度は，陰性と判定されたもののうちの真の陰性の割合であり，両者が 100%に近ければ近いほど，優れた診断系であると言える．

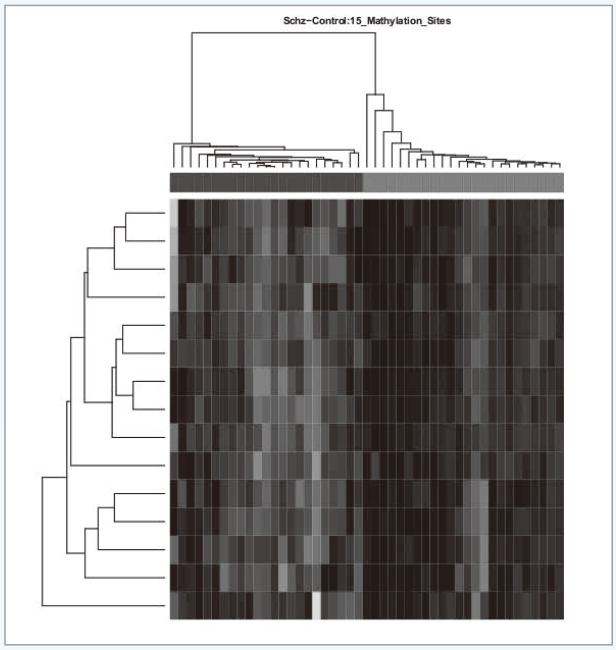

図 -6　分析結果のヒートマップによる視覚化例

最終的に得られた数理モデルは，クロスバリデーション（交差検定法）などにより検証する．

(4) データ分析の実際

解析データは，マイクロアレイや次世代シーケンサーなど大量のゲノムデータを用い，たとえば，今回用いる次世代シーケンサーデータでは，百数十人分のデータはテラバイト級となる．このため，理化学研究所の「京」互換スーパーコンピュータ SCLS，AWS 上のクラウドコンピュータクラスタなどの高パフォーマンスコンピュータ環境を適宜用いてデータ解析を行う．

(5) 分析結果の視覚化

分析結果は，グラフやヒートマップで視覚化する（図 -6）．

精神神経系疾患診断系構築の実際

筆者らは，上記に述べた方法を用いて，うつ病，総合失調症，双極性感情障害などの精神神経系疾患において，良好な識別能力を有する診断系の構築に成功している．従来では，問診による経験的判断でしか，診断がつかなかった精神神経系疾患においてエビデンスに基づく診断を導入する糸口が見えてき

ている．詳細は関連の臨床系の雑誌に報告するのでそちらを参照してほしい[5),6)]．

医療分野のビッグデータ利活用の問題点に関する考察と思い

以上，医療分野におけるビッグデータ利活用の現状とデータ分析の実際について走り走り述べてきた．ビッグデータ活用の実際の課題として，Hadoop などのビッグデータプラットフォームの導入や，データベースの構築，データアナリティクス・データマイニング手法の確立などがあるが，実際のデータ分析は，単に情報技術や，統計技術の集積だけでは困難で，その生物現象そのものの理解や背景知識が不可欠になってくる．

医療分野におけるビッグデータ利活用は，個人情報を扱うということと，産業上の大きな利益が得られるということから，プライバシー問題や知的財産権の問題も重要である．このことから医療分野におけるビッグデータ利活用において，課題はまだまだ残されている．

よく指摘されることに，データベースの構築やインフラはできたけれど，それを実際に分析して活用できる人材が不足しているという点がある．たとえば eMERGE の PREDICT のような人口知能的なシステムの構築などは，目を見張るものがある．しかし，データ分析は，単に情報技術や統計技術の集積だけでは困難で，専門知識がどうしても欠かせないことから，各分野の専門家にデータアナリティクスを浸透させていく人材育成が必要だと考える．

ビッグデータに期待することは，イノベーションの推進である．データマイニングによって得られる知識の発見により新しい産業が生まれ，先端技術を社会に還元し，技術革新により世界を変えていく．そのようなサイクルが進むことが期待されることから，ビッグデータを単なるバズワードに終わらせるわけにはいかないと考える．

くしくも，1998 年に Larry Page と Sergey Brin が Google を設立して以来，大規模な検索エンジンの構築のためのツールとして用いてきた BigTable や Google File System，MapReduce などの製品群が，今日のビッグデータ利活用の源流と考えると，このイノベーションは一時的なはやりで終わることなく，今後も世界を変えていく原動力になり続けるであろうと信じる．

参考文献

1）石井一夫：ゲノム科学におけるビッグデータ分析・大規模データマイニング，BIOINDUSTORY, 31, 6, pp.67-73（2014）．
2）石井一夫：解説：医療，農学，環境分野におけるビッグデータ解析，生物工学会誌，92, 2, pp.92-93（2014）．
3）石井一夫，佐藤　暁，古崎利紀，有江　力，寺岡　徹：ゲノム科学におけるビッグデータ・データマイニング，日本統計学会誌，43, 1, pp.90-111（2013）．
4）Rizzo, M.（著），石井一夫，村田真樹（共訳）：R による計算機統計学，オーム社（2011）．
5）Fuchikami, M., Morinobu, S., Segawa, M., Okamoto, Y., Yamawaki, S., Ozaki, N., Inoue, T., Kusumi, I., Koyama, T., Tsuchiyama, K. and Terao, T. : DNA Methylation Profiles of the Brain-Derived Neurotrophic Factor（BDNF）Gene as a Potent Diagnostic Biomarker in Major Depression, PLoS One. 6, 8, e23881（2011）．
6）Kinoshita, M., Numata, S., Tajima, A., Shimodera, S., Ono, S., Imamura, A., Iga, J., Watanabe, S., Kikuchi, K., Kubo, H., Nakataki, M., Sumitani, S., Imoto, I., Okazaki, Y. and Ohmori, T. : DNA Methylation Signatures of Peripheral Leukocytes in Schizophrenia., Neuromolecular Med. 15, 1, pp.95-101（2013）．

（2014 年 6 月 18 日受付）

■ **石井一夫（正会員）**　kishii@cc.tuat.ac.jp

東京農工大学農学府特任教授．徳島大学大学院医学研究科博士課程修了後，理化学研究所ゲノム科学総合研究センター，ノースウエスタン大学医学部などを経て現職．専門：ゲノム科学，計算機統計学，データマイニング．