2015 年 10 月 16, 17 日
東京大学医科学研究所講堂

文科省科研費基盤 A (15H01678)
「大規模複雑データの理論と方法論の総合的研究」
研究代表者: 青嶋 誠（筑波大学）
シンポジウム

生命科学データ解析の方法論と健康科学への応用

開催責任者
井元清哉（東京大学医科学研究所）

内容・目的

超高齢化社会を迎える我が国にとって必須の課題である健康寿命を延ばすためには、ゲノム、およびゲノム由来データ、感染症などの定点観測データ、レセプト情報・特定健診等情報等の時間軸を有する我が国に特有の優れた健康・医療に関する大規模データを、数理モデリングとシミュレーションにより個々人の疾患の統合予知・予防に活用することが期待されています。本シンポジウムにおいては、DNA、RNA、エピゲノムなどゲノム関連データに対する新たな解析手法の提案、上述の健康・医療に関する大規模データ解析、およびそれらを用いた疾患の予知・予防を目指した研究開発に関わる講演を頂きます。

【プログラム】

10 月 16 日（金）

13:00-13:50 【特別講演】
岡田 随象（東京医科歯科大学大学院医歯学総合研究科）
「ビッグデータ時代の遺伝統計学と疾患病態解明・新規創薬への展望」

14:00-14:40
荒木 由布子（静岡大学情報学部情報社会学科）、岡田 栄作（浜松医科大学医学部健康社会医学講座）、近藤 克則（千葉大学予防医学センター環境健康学研究部門）、尾島 俊之（浜松医科大学医学部健康社会医学講座）
「10 年間追跡調査に基づく高齢者の要介護認定リスク因子の探索的検討」

14:40-15:20
日笠 幸一郎（京都大学大学院医学研究科附属ゲノム医学センター）
「大規模ゲノムコホート事業と日本人遺伝子多型データベース」

15:30-16:10
西野 穣（名古屋大学大学院、CREST）、高地 雄太（理化学研究所、CREST）、野間 久史（統計数理研究所、CREST）、重水 大智（理化学研究所、CREST）、森園 隆（理化学研究所、CREST）、角田 達彦（理化学研究所、CREST）、松井 茂之（名古屋大学大学院、CREST）
「複雑疾患の GWAS データへの階層混合モデリングと解析例」

16:10-16:50
植木 優夫（久留米大学バイオ統計センター）、田宮 元（東北大学東北メディカル・メガバンク機構）
「GWAS データに基づく円滑閾値型推定方程式による遺伝的予測」

16:50-17:30
新井田 厚司（東京大学医科学研究所ヘルスインテリジェンスセンター）
「がんの進化シミュレーションによる腫瘍内不均一性生成原理の探索」

18:00-20:00　　懇親会（近代医科学記念館）

10 月 17 日（土）

10:00-10:40
斎藤 正也（統計数理研究所データ同化研究開発センター）
「大規模感染症シミュレーションの近年の動向」

10:40-11:20
矢原耕史（久留米大学バイオ統計センター、東大院・新領域、College of Medicine, Swansea Univ.)、Xavier Didelot (Dept. Infectious Diseases Epi., Imperial College London)、M. Azim Ansari (Dept. Statistics, Univ. Oxford)、Samuel K. Sheppard (College of Medicine, Swansea Univ.)、Daniel Falush (Dept. Evol. Genetics, Max Planck Institute)
「病原細菌の種内多数のゲノムデータから組換えのホット領域を推定する手法の開発 」

11:40-12:20
鈴木 譲（大阪大学大学院理学研究科）
「グラフィカルモデルによる遺伝子の多重選択のアプローチ」

12:20-13:00
松井 秀俊（九州大学大学院数理学研究院）
「スパース正則化に基づく経時測定データの判別と遺伝子データ解析への応用」

演題：ビッグデータ時代の遺伝統計学と疾患病態解明・新規創薬への展望
Statistical Genetics in the Big Data era and its contribution to disease biology and drug discovery.

氏名：岡田　随象

所属：東京医科歯科大学 大学院医歯学総合研究科 テニュアトラック講師

要旨：遺伝統計学とは、生物における遺伝情報と形質情報との結びつきを、統計解析を通じて明らかにする研究分野である。ヒトゲノム配列が解読されてから 10 年以上が経過し、マイクロアレイや次世代シークエンサー技術に代表されるゲノム解読技術の著しい進歩は、数千人～数十万人規模のサンプルを対象とした大規模ヒトゲノム解析を現実のものとしている。遺伝統計学は、大規模ヒトゲノム解析を通じたヒト疾患感受性遺伝子の同定だけでなく、多彩な生物学的・医学的データベースとの横断的解析を通じて新たな疾患病態の解明や新規創薬に貢献できる学問分野としても注目を集めている。近年では、特定の疾患の組み合わせにおける合併率の変化など、疾患疫学研究が指摘した疑問点の解決に大規模ゲノムデータを用いるアプローチも始まっている。一方で、既に到来しつつある「ビッグデータ時代」において、生命情報科学に携わる個々の研究者が進むべき方向性については、模索が続いているのも現状である。本セミナーでは、我々が行ってきた遺伝統計解析の成果を紹介すると共に、これからのヒトゲノム解析の展望について述べたい。

10 年間追跡調査に基づく高齢者の要介護認定リスク因子の探索的検討

荒木 由布子　静岡大学情報学部 情報社会学科
岡田 栄作　浜松医科大学 医学部 健康社会医学講座
近藤 克則　千葉大学 予防医学センター環境健康学研究部門
尾島 俊之　浜松医科大学 医学部 健康社会医学講座

厚生労働省の調査では，2015 年 3 月時点で要支援・要介護の認定を受けた人の数は今年度になり初めて 600 万人を突破し，これは国民の約 20 人に 1 人が要支援・要介護認定者であることを示している．認定者は増加の一途で，2025 年には団塊の世代が 75 歳以上になるため増加ペースはさらに上がることが予想される．要介護認定の増加は介護給付金の増加を招き，これ以上の負担増は国家の財政を圧迫することにつながり現役世代を逼迫するため，要介護認定者の健康状態までに健康状態を損なう事をどのように防げばよいか，より健康寿命を延ばすにはどうすればよいのか，の知見を得ることが緊急の課題となっている．

要介護認定のリスクについて，厚生労働省は「運動器の機能向上」，「栄養改善」，「口腔機能の向上」，「閉じこもり予防・支援」，「うつ予防・支援」，「認知症予防・支援」の 6 個の強化すべき分野を設定している２）．先行研究においても要介護リスク要因については研究が蓄積されており，例えば平井らは３），要介護認定をエンドポイントとしたコホート研究により要介護のリスク要因を検討した．また，要介護が疑わしい対象者をいち早くスクリーニングするために，介護保険制度の二次予防事業の対象者の把握（要介護認定が疑わしい対象者）には，25 項目の基本チェックリストを用いている４）が，この基本チェックリストによる要介護認定の予測精度を検証した報告は少ない．妥当性が検証された論文でも 1 年間という短期の要介護認定発生者を追跡した研究５）であって，10 年間という長期にわたり追跡調査し，要介護認定のリスク因子を検証する研究は本研究が初の試みである．

本研究では要介護認定のリスクと考えられる約 300 項目の中から影響力の強い変数を絞り込み，要介護認定に至るリスクモデルを構築する事が目的である．本発表はそのためのパイロットスタディであり，予後モデル構築後は，リスク因子間の構造を解明し，その機序を明らかにする．

本発表ではパイロットスタディとして愛知県知多地域を対象地域とし，2003 年 10 月の調査に回答し，65 歳以上で追跡可能な約 15000 サンプルを 3436 日間追跡調査したデータの解析について報告する．調査開始から要介護 2 以上発生までの時間をエンドポイントとし，要介護 2 以上発生の状態となる候補予後因子は AGES2003 基本調査票に基づく社会的要因，心理的要因，生活習慣などの計 331 因子である．このような大規模データの長期間追跡データは欠測値を多く含み，通常の統計解析に頻繁に用いられるリストワイズ除去ではデー

タの多くが失われてしまう上，バイアスが生じる．このため，欠測値の発生したシステムを考慮し，適切に対応する必要がある 9)，10)．本研究ではこの問題を回避するため，多重代入法により欠測値を補定した 11)，12)． 要介護認定２のレベルまで健康状態が悪化する予後因子の探索を行うため，男女別にスパース性を仮定したCox回帰モデルを用い6)，7)， パラメータ推定と変数選択にはElastic Net を用いた 6)．研究結果である予後モデルの信頼性が重要となるが，交差検証法により内的妥当性を考慮し，C 統計量を用いて予後モデルの予測の正確性を評価した 8)．また，本発表では対象地域を上記に限っているが，外的妥当性の検証のため今後異なる地域のコホートに予後モデルを適用していく．

男女別に解析した結果のリスク因子の詳細は当日報告するが，今回の結果で予測因子と結論付いた変数は，短期間調査でそれぞれの因子を別々にみたときに影響が大きいとされたものと同じものが多く含まれた妥当な変数であった．今回の，多くの変数から構成される長期的な追跡調査データに基づき予後モデルの信頼性や再現性に注意して絞り込まれた男女別のリスク因子は貴重であり，今後のより精度の高い介護認定２のスクリーニング指標の開発に貢献できると考えられる．今後の課題は，リスク因子間の構造を解明していくことでリスク因子の機序を明らかにし，健康寿命の増長に役立てる事を目指す．

【参考文献】

１）厚生労働省.介護保険事業状況報告（暫定）（平成 27 年 4 月分）
http://www.mhlw.go.jp/topics/kaigo/osirase/jigyo/m15/1504.html

２）「介護予防のための生活機能評価に関するマニュアル」分担研究班（主任研究者　鈴木隆雄）．介護予防のための生活機能評価に関するマニュアル（改訂版）．2009 年．
http://www.mhlw.go.jp/topics/2009/05/dl/tp0501-1c.pdf

３）平井寛，近藤克則，尾島俊之，村田千代栄．地域在住高齢者の要介護認定のリスク要因の検討 ：AGES プロジェクト 3 年間の追跡研究．日本公衆衛生雑誌 56(8)，501-512，2009.

４）「介護予防のための生活機能評価に関するマニュアル」分担研究班（主任研究者　鈴木隆雄）．介護予防のための生活機能評価に関するマニュアル．2009 年．
http://www.mhlw.go.jp/topics/2009/05/dl/tp0501-1c.pdf

５）遠又靖丈，寶澤篤，大森（松田）芳，永井雅人，菅原由美，新田明美，栗山進一，辻一郎． 1 年後の要介護認定発生に対する基本チェックリストの予測妥当性の検証． 日本公衆衛生雑誌． 58(1)： 3-12， 2011.

６）Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. J.R.Stat.Soc.Ser.B, 67, 301-320.

7) Tibshirani R. (1997). The lasso method for variable selection in the Cox model. Stat. Med. 16, 385-395.

8) Yuan, Y. et al. (2014). Assessing the clinical utility of cancer genomic and proteomic data across tumor types. Nature Biotechnology 32, 644 - 652.

9) 岩崎学. (2002).「不完全データの統計解析」，エコノミスト社.

10) 星野崇宏. (2009).『調査観察データの統計科学：因果推論・選択バイアス・データ融合』，岩波書店.

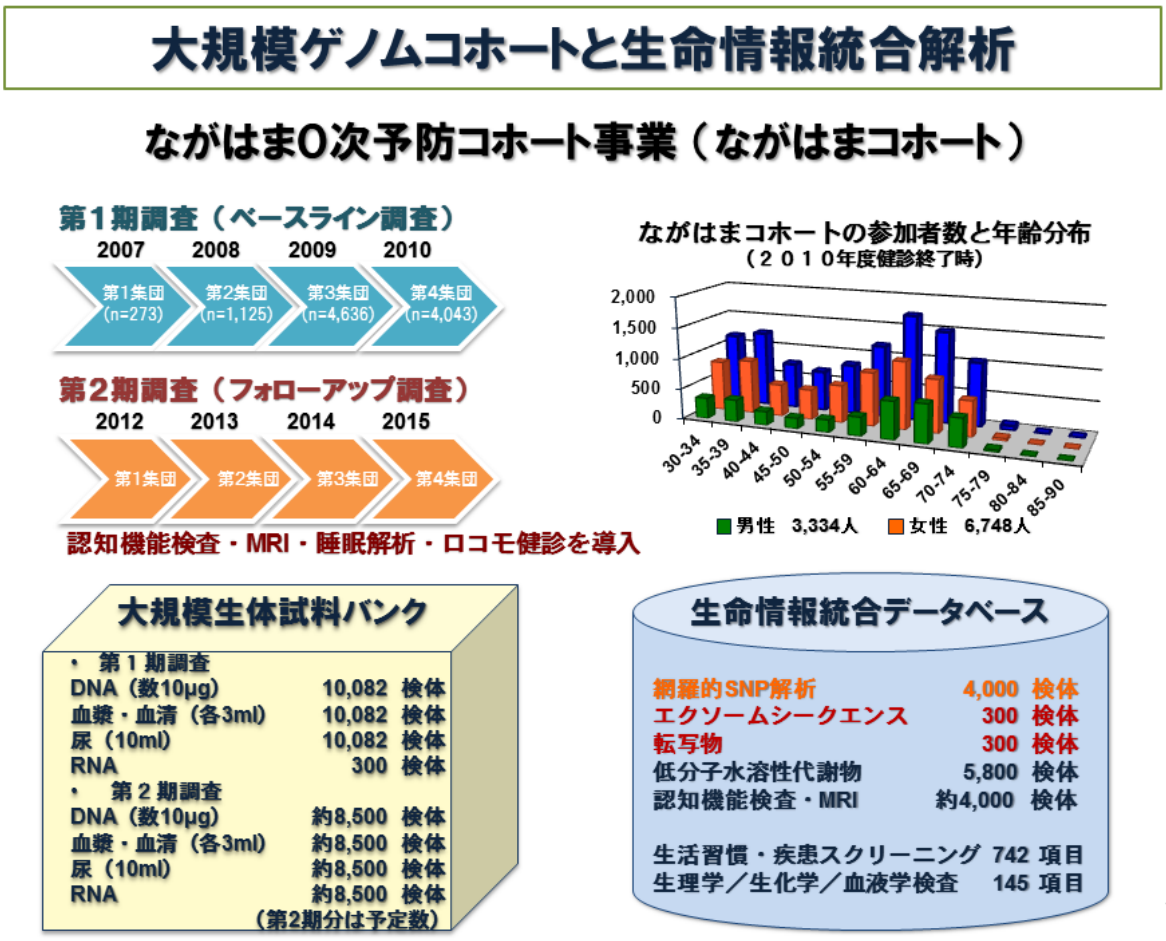
11) White IR and Royston P. (2009). Imputing missing covariate values for the Cox model. Stat Med. 28, 1982-1998.

12) Honaker J, King G and Blackwell M. (2011). Amelia II: A program for missing data. Journal of Statistical Software 45(7), 1-47.

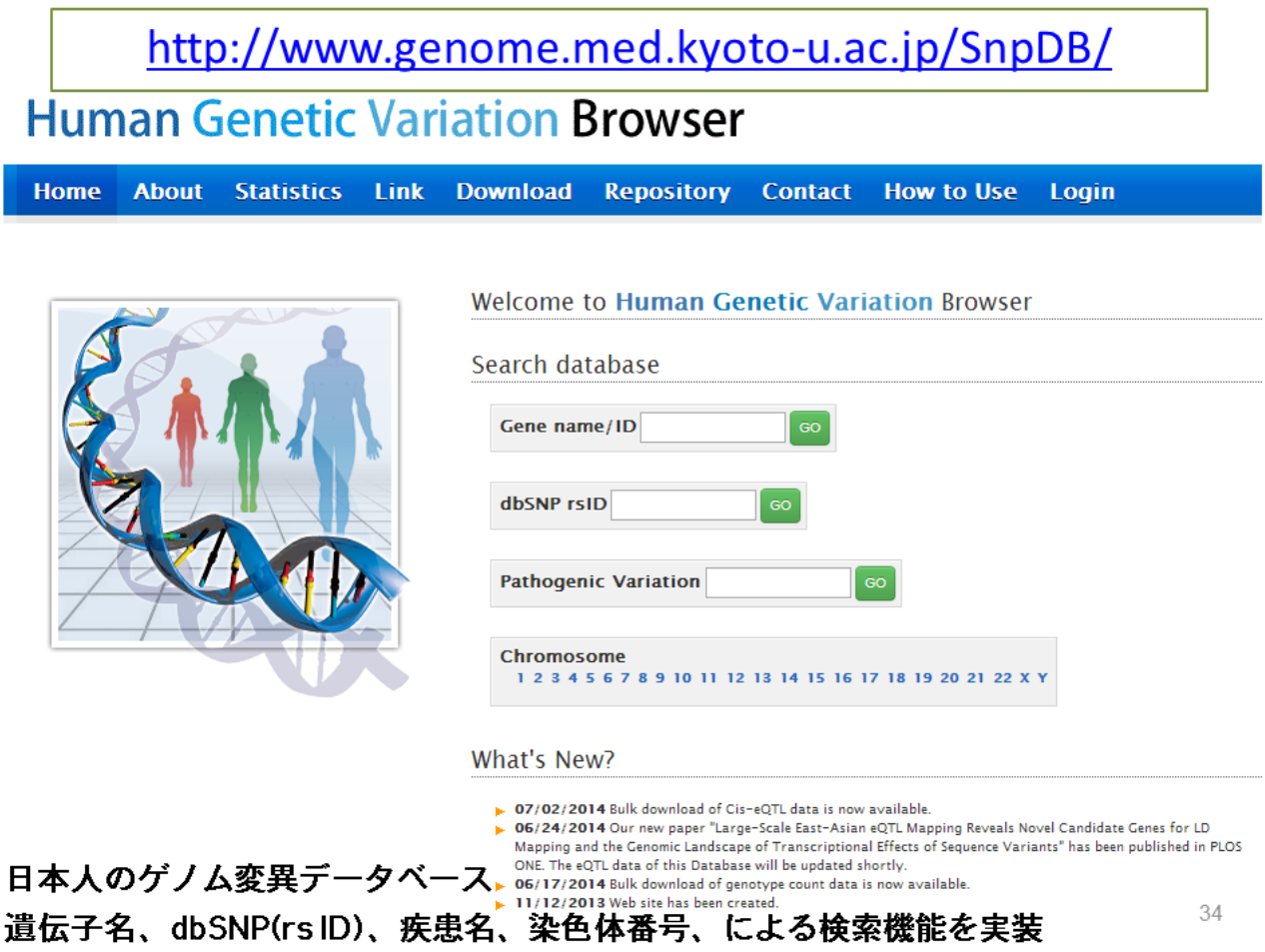
大規模ゲノムコホート事業と日本人遺伝子多型データベース

京都大学大学院医学研究科附属ゲノム医学センター
特定講師・日笠　幸一郎

臨床症状が現れる前の発症前診断、究極の医療とも言うべき発症前治療を実現化するためには、大規模集団を対象に長期の追跡をおこない、継続的に得られる生体分子の詳細な分析・解析に基盤を置きつつ質の高い疾病罹患情報や環境・生活習慣情報と統合した解析が不可欠である。このような、健康を遺伝子や細胞レベルに限らず、「分子を通して身体全体で見る」新たな予防医学のアプローチは、従来の疫学研究とは全く異なるもので、多くの研究分野や産業界の協力で推進すべき学際的研究である。京都大学ゲノム医学センターでは、ヒト生命情報統合研究のモデルケースとして、2005 年より滋賀県長浜市で地域住民を対象にゲノムコホート研究を進め、10,082 名の参加者について、環境・生活習慣調査、生化学・血液学検査などの情報や、白血球 RNA の発現解析、代謝物の網羅的解析、SNP アレイを用いたゲノム多型解析、Exome シークエンシングを実施している。これらの解析事例、及び、情報を統一的に管理・運用・公開するために構築しているデータベースの概要について紹介し、ヒト生命情報統合解析の戦略について議論した。



http://www.genome.med.kyoto-u.ac.jp/SnpDB/



日本人のゲノム変異データベース
遺伝子名、dbSNP(rs ID)、疾患名、染色体番号、による検索機能を実装

本データベースには、日本人遺伝子リファレンス情報（3,248 人の健常者の SNP 頻度情報、1,208 検体のエクソーム解析情報）、稀少難治性疾患の遺伝子変異情報（13 疾患、320 変異）が集約され、公開から 3 ヶ月で約 72 万回のアクセスを得ており、疾患ゲノム解析への関心の高さ、含まれるゲノム情報の有用性が証明された。本データベースが難病研究の中核として機能することは疑いなく、遺伝子リファレンス情報・疾患遺伝子情報を手がかりとして当該疾患領域の医療・研究が加速され、難病研究領域の全体的なレベルアップに繋がるとともに、希少難治性疾患の遺伝子診断における標準化が進むなど、臨床に直結する基盤情報が提供される。今後、様々な疾患でシークエンスデータに基づく関連解析が増加すると予測されるが、そのような研究の基盤として、利用価値が高い。
本データベースの利活用により、疾患の原因変異究明のプロセスが飛躍的に向上することが期待される。将来に向けて、本事業で得られた成果やバイオリソースなどを、研究者コミュニティが広く活用できるような仕組みをさらに発展させることが重要である。今後は、日本人ゲノム変異データベースの検体数を増やすとともに、難病の遺伝子変異の登録を難病研究班にうながし、新たなデータの蓄積、機能の向上、維持管理などを継続的におこない、より充実したデータベースとすることが望まれる。本事業によって達成された大規模研究成果のデータベース化と公開により、難病研究、ゲノム医学研究が大きく発展する礎になれば幸いである。

複雑疾患の GWAS データへの階層混合モデリングと解析例

西野穣 [1,4], 高地雄太 [2,4], 野間久史 [3,4], 重水大智 [2,4], 森園隆 [2,4], 角田達彦 [2,4], 松井茂之 [1,4]
[1] 名古屋大学大学院, [2] 理化学研究所, [3] 統計数理研究所, [4] 科学技術振興機構 CREST

1. 背景と目的

リウマチ、糖尿病、統合失調症などの多くの精神疾患は複雑疾患と呼ばれ、それらの発症には多くの環境因子と遺伝因子が寄与していると考えられている。ヒトゲノムの一部の塩基対では、複数の塩基タイプが確認され、これを SNP（一塩基多型）という。GWAS（Genome-Wide Association Study）は、数十万個以上の SNP の個人の型（遺伝子型）を一度に決定できる SNP チップを用い、ゲノムワイドに SNP と形質との関連を一挙に調べる方法である。これまで、GWAS によって多くの疾患関連 SNP が同定されてきた。しかし、多くの複雑疾患において、有意な SNP だけでは家系分析から推定される遺伝的寄与度（Heritability）のごく一部しか説明できない「Missing heritability」と呼ばれる現象が見られることが分かってきた[1]。一方、チップ上の全 SNP を同時に用いて推定される遺伝的寄与度（SNP heritability）は、Heritability の相当な割合を占める事が知られている[2]。この事は、SNP チップ上には、有意水準には達しない小さな効果を持つ多数の関連 SNP が存在する事を示唆している。そこで本研究では、階層混合モデルに基づいて、全 SNP に存在する関連 SNP の割合と効果サイズの分布を推定し、その下で小さな効果を持つ SNP の有効な評価を可能とする解析法を確立する目的とする。

2. 階層混合モデル

目的変数としてケース（患者）を$z = 1$、コントロール（正常）を$z = 0$、説明変数として各 SNP の遺伝子型 AA: $x = 0$, Aa: $x = 1$, aa: $x = 2$としたロジスティック回帰をm個の SNP 毎に行い、j番目の SNP についての対数オッズ比の最尤推定値 $Y_j = \hat{\beta}_j$とその分散 $\hat{V}_{\hat{\beta}_j}$を得たとする。いま、j番目の SNP の確率分布に対して次の階層混合モデルを仮定する：

$$f_j(y_j) \;=\; \pi f_{0j}(y_j) \;\;+\;\; (1-\pi) f_{1j}(y_j)$$

ここで、f_{0j}は null（関連なし）のコンポーネントに対応し$y_j \sim N(0, \hat{V}_{\hat{\beta}_j})$ と指定する。f_{1j}は、non-null（関連あり）のコンポーネントに対応し$y_j|\beta_j \sim N(\beta_j, \hat{V}_{\hat{\beta}_j}),\ \beta_j \sim g_1$ という階層構造を指定する。効果サイズの分布g_1 は指定しない。 πとg_1の推定は、EM アルゴリズムにより行い（経験ベイズ推定）[3]、これらの推定によって効果サイズの全体像を捉える事が可能となる。そして、これらの遺伝的構造（πとg_1）の下で得られた各 SNP に関する効果サイズのベイズ的推定値は、モンテカルロシミュレーションによって、効果サイズの全領域に渡りほぼ不偏である事が分かった。

3.　実データへの適用例

岡田[4]は、リウマチの感受性変異を同定するために、ヨーロッパ人とアジア人の合計 22 の GWAS 研究のメタアナリシスを実施した。このうちの 4 つのアジア人 GWAS（4873 ケース、17642 コントロール）に関するメタアナリシスの要約データ（各 SNP のオッズ比の最尤推定値およびその信頼区間）を上記の$\hat{\beta}_j$(対数オッズ比)とその分散 $\hat{V}_{\hat{\beta}_j}$に変換し、階層混合モデルを適用した。その結果、関連 SNP の割合 $(1-\pi)$を 2.83 %、g_1を Figure 1 のとおり推定した（リスクアリルの平均オッズ比はわずか約 1.04 と非常に小さい）。 また、π とg_1 の推定に基づいたβ_jの絶対値の事後平均を用いることにより、リウマチでは MAF (Minor Allele Frequency)と効果サイズはあまり関係がないことも確認できた（Figure 2)。ただし、GWAS データを用いているため希な変異（rare variants）については言及できない。

Figure 1: g_1の推定

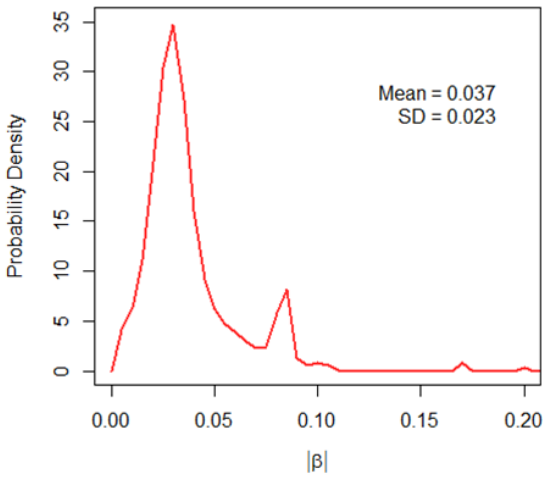
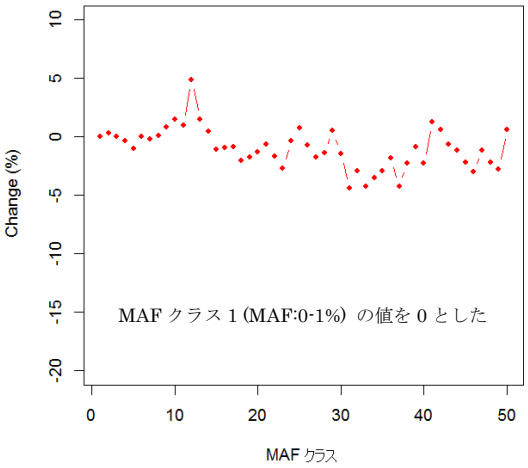


Figure 2: βの絶対値の事後平均の変化(%)



[参考文献]

1. Manolio, T. a et al. Finding the missing heritability of complex diseases. Nature 461, 747–753 (2009).
2. Lee, S. H., Wray, N. R., Goddard, M. E. & Visscher, P. M. Estimating missing heritability for disease from genome-wide association studies. Am. J. Hum. Genet. 88, 294–305 (2011).
3. Matsui, S. & Noma, H. Estimating effect sizes of differentially expressed genes for power and sample-size assessments in microarray experiments. Biometrics 67, 1225–1235 (2011).
4. Okada, Y. et al. Genetics of rheumatoid arthritis contributes to biology and drug discovery. Nature 506, 376–381 (2014).

GWASデータに基づく円滑閾値型推定方程式による遺伝的予測

久留米大学 バイオ統計センター　植木 優夫
東北大学 東北メディカル・メガバンク機構　田宮 元

1. はじめに

ゲノムワイド単一塩基多型（SNP）データを用いたヒト形質の遺伝的要因を探る，いわゆるゲノムワイド関連研究（GWAS) は，近年，益々活発に研究が進んでいる．個別化医療の実現に向けては，SNP-GWAS の結果から疾病発症リスクや量的形質値を予測可能な数学モデルが必要となる．ただし，標準的な SNP-GWAS データは 50〜100 万個以上の SNPs が含まれる超高次元データであり，$p \gg n$ 問題と呼ばれる統計学上の困難が存在する．したがって，高精度な予測モデルの構築は非常にチャレンジングな問題となっている．本研究では，円滑閾値型推定方程式 (Ueki 2009) を用いた新たな遺伝的予測法を提案する．本手法によって，大規模 SNP-GWAS データにおいても高速に高精度な予測モデルを構築することが可能となる．

2. 周辺関連性に基づく円滑閾値型推定方程式

SNP-GWAS データにおいて，形質値（疾患の有無などの二値形質，あるいは臨床値などの量的形質）と遺伝子の間の関連性を発見するための標準的な手法は，各 SNP ひとつづつと形質間の関連性を一変量回帰モデルを通じて調べる方法である．すなわち，p 個の SNP があり，n 人の形質値 $y = (y_1, \ldots, y_n)^T$ と j 番目の SNP $X_j = (X_{1j}, \ldots, X_{nj})^T$ （$X_{ij} \in \{0, 1, 2\}$）に対し，周辺回帰モデル

$$y = \beta_{j,0} + X_j \beta_{j,1} + \epsilon,$$

において帰無仮説 $H_0 : \beta_{j,1} = 0$ を検定する方法である．疾患の有無のような二値形質では，線形回帰モデルの代わりにロジスティック回帰モデルが用いられる．このように周辺的な関連性に基づけば，$p \gg n$ 条件下でも実行可能となる．偽陽性を制御するために，検定の多重性を考慮し，P 値がゲノムワイド有意水準 5×10^{-8} 以下を示した SNP を有意な SNP と認めることが多い．これは新規遺伝子を発見するために偽陽性率の制御を重要視した戦略であるが，予測モデルの構築においては，予測性能の向上が目標となる．Purcell *et al.* (2009) は，5×10^{-8} よりも大きな P 値のカットオフ値を用いて，この P 値カットオフ以下の P 値を与えた SNP を予測モデルに用いる遺伝子スコア法を提案している．彼らは，クロスバリデーションを用いて，様々な P 値カットオフ値の候補の中から最適な予測性能を与えるカットオフを選んだ．Purcell らの手法は周辺回帰による回帰係数の線型結合によって予測スコアを作成するため，SNP 間の相関（すなわち連鎖不平衡）が考慮されていない．また，データに関する不連続性による予測性能の悪化が懸念される（Breiman 1996）.

ここで，SNP 間の相関を重回帰モデル $y = X\beta + \epsilon$ によって考慮することを考える．$X = (X_1, \ldots, X_p)$ は p 個の SNP からなる $n \times p$ デザイン行列であり，β は各 SNP に対応する回帰係数ベクトルである．いま，j 番目の SNP の周辺関連性に対応するカイ 2 乗統計量 $T_j(y, X)$ と，ある P 値カットオフに対応する閾値 $t > 0$ により，添字集合 $A = \{j : T_j(y, X) > t\}$ に対して，

$$\hat{\beta}_A = (X_A^T X_A)^{-1} X_A^T y, \quad \hat{\beta}_{A^c} = 0,$$

によって回帰係数 $\hat{\beta}$ を求めれば，$p \gg n$ であっても計算可能である．ただし，データに関する不連続性が，$T_j(y, X) > t$ という SNP のフィルタリングに存在している．より具体的には，$\hat{D} = \mathrm{diag}(\hat{D}_1, \ldots, \hat{D}_p)$，$\hat{D}_j = 1\{T_j(y, X) > t\}$ とおくことで，回帰係数 $\hat{\beta}$ は

$$(I_p - \hat{D})\{X^T(X\hat{\beta} - y)\} + \hat{D}\hat{\beta} = 0,$$

の解として解釈でき，データに関する不連続性は指示関数 $\hat{D}_j$ に存在していることがわかる．ただし，I_p は p 次元単位行列である．指示関数を連続な関数で置き換えることで，データに関する不連続性を取り除くことが可能となる．我々は以下の円滑閾値型方程式を用いることを提案する．

$$(I_p - \check{D})\{X^T(X\check{\beta} - y)\} + \tau\check{D}\check{\beta} = 0,$$

ここで，$\check{D} = \mathrm{diag}(\check{D}_1, \ldots, \check{D}_p)$，$\check{D}_j = \min[1, \{t/T_j(y, X)\}^{\frac{1+\gamma}{2}}]$，$\tau$，$\gamma$ は，所与の定数である．$T_j(y, X) \leq t$ であれば $\check{D}_j = 1$ となり，$\check{\beta}_j = 0$（スパース解），他方で，$T_j(y, X) > t$ であれば $\check{D}_j < 1$ となり，$\check{\beta}_j \neq 0$ を得る．さらに，SNP フィルタリングの連続化によって一般化自由度 (Ye 1998) が陽に求まり，Stein の不偏リスク推定（SURE）によるモデル選択から最適な t を決定することが可能となる．すなわち，以下の C_p 規準を最小化する t を用いることで，クロスバリデーションに頼らない，高速かつ高精度な予測モデルを構築できる：

$$C = \sum_{i=1}^{n}(y_i - X_i\check{\beta})^2 + 2\sigma^2\mathrm{GDF}.$$

ここで，GDF は一般化自由度である．数値実験および実データ適用により，提案手法が実際に優れた性能を示すことが確認された．

参考文献

[1] Breiman L. (1996). Heuristics of instability and stabilization in model selection. *Ann Statist* **24**: 2350–83.

[2] Purcell SM, Wray NR, Stone JL, Visscher PM, O 'Donovan MC, Sullivan PF, Sklar P, International Schizophrenia Consortium. (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**: 748–52.

[3] Ueki M. (2009). A note on automatic variable selection using smooth-threshold estimating equation. *Biometrika* **96**: 1005–11.

[4] Ye J. (1998). On measuring and correcting the effects of data mining and model selection. *J Am Statist Assoc* **93**: 120–31.

がんの進化シミュレーションによる腫瘍内不均一性生成原理の探索

新井田厚司
東京大学医科学研究所ヘルスインテリジェンスセンター健康医療計算科学分野

概　　要：がんは細胞のゲノムに変異が蓄積し増殖能力が高いものが進化的に選択された結果生じる。この進化の過程で様々なクローンが生み出され一つの腫瘍内においてゲノムレベルの不均一性を生み出していると考えられている。演者は九大別府病院との共同研究で一人の患者からの大腸がんの複数の部位から得たDNAをexome sequencingすることにより大腸がんに広汎な腫瘍内不均一性が存在するのを見出した。また他の癌腫についても同様の解析により腫瘍内不均一性の報告がなされている。しかしながら、それを生み出す原理の探求についての試みはほとんどなされていない。この目的のために演者は腫瘍内不均一性を再現する、がんの進化シミュレーションモデル、Branching evolutionary Process (BEP) Modelを構築した。また本研究ではスーパーコンピュータ「京」を利用して膨大な組み合わせのパラメーターセットでBEP modelによるがんの進化シミュレーションを行うことにより、実験データと同様の高い腫瘍内不均一性が生み出される条件の網羅的探索をおこなった。その結果、高い遺伝子変異率、がん幹細胞の存在を仮定すると高い腫瘍内不均一性が再現できることを見出した。更にシミュレーション結果から細胞の増殖に寄与するドライバー遺伝子は進化の初期に獲得され全てのがん細胞に共有されている一方で、不均一性を生み出している変異の大部分は細胞の増殖速度に影響を与えない中立変異であることが示唆された。以上、本研究により腫瘍内不均一性を生み出している進化原理の一端が大腸がんゲノム解析とスーパーコンピュータを用いたがんの進化シミュレーションにより明らかにされた。

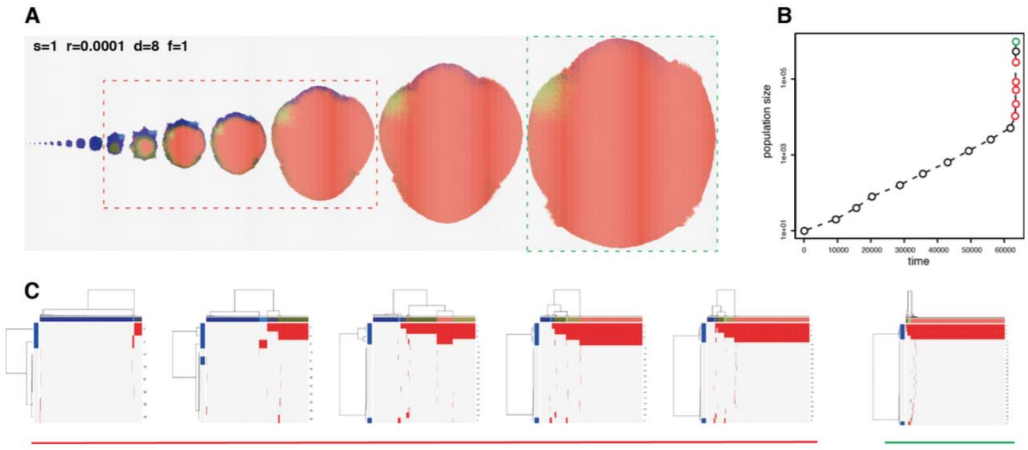


図 1 ：Clonal selective sweep. (A) Snap shots of growing tumors in a simulation with indicated parameter sets. Differently colored cell populations represent each clone. (B) A growth curve of the simulated tumor. The snap shots were obtained at each plotted point. (C) Mutation profiles of the simulated tumor during growth (indicated by red dashed rectangles, plotted points and under lines) and the end point (indicated by green dashed rectangles, plotted points and under lines). Colored bars on the columns represent each clone.

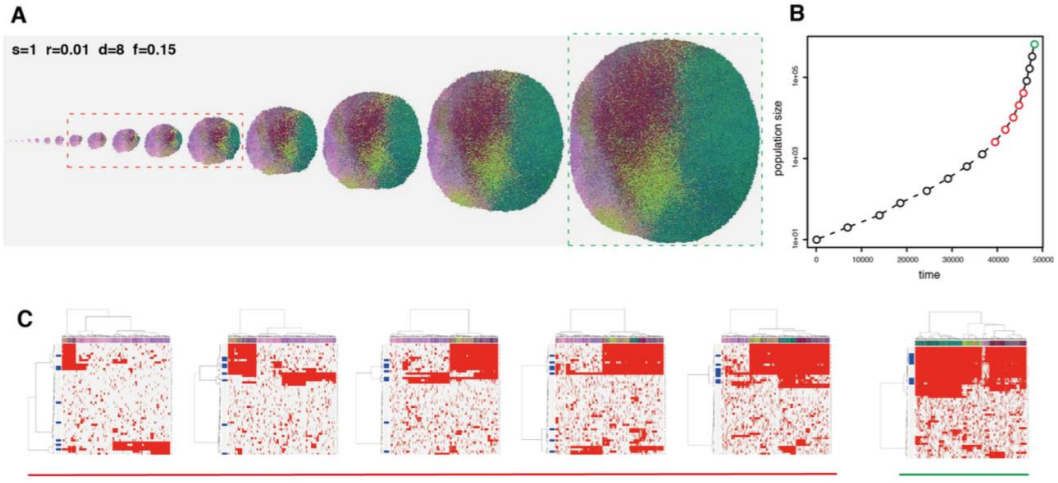


図 2 ：Neutral evolution generating fractal heterogeneity. (A), (B), (C) は図 1 と同様

大規模感染症シミュレーションの近年の動向

斎藤　正也　（統計数理研究所データ同化研究開発センター）

人類は感染症の脅威に持続的に曝されている． 1918 年に流行した A (H1N1)型新型インフルエンザ・ウイルス(通称・スペインかぜ)は全世界に大きな被害をもたらし，近年では 2009 年に A (H1N1)型の新型ウイルスが世界的流行を起こした. また, 2013 年には A(H7N9)型の新型トリウイルスが発見され，香港で局地的流行を引き起こした．幸いなことに，2009 年新型は季節性インフルエンザと同程度の毒力で，2013 年新型は安定したヒト・ヒト間感染力を獲得しておらず，局地的流行に留まっている．また，インフルエンザ以外にも 2014 年に西アフリカにおけるエボラウイルス病の流行，2015 年の韓国の病院内で伝染が起った中東呼吸器症候などの致死性の高い感染症に遭遇している．

このような潜在的な感染症の流行に対して対策を事前に用意しておくことは重要である．そのとき，計算機上での流行シミュレーションが有望視されている．感染症の数理モデルには多様なバリエーションが存在する．ここでは，インフルエンザなど罹患によって免疫が獲得される疾患を流行の 1 シーズンに限って記述することに制限して，個人の病態が感受性 (S)，感染力獲得 (I)，回復後免疫獲得 (R)へと進行すると仮定する．感染力獲得の待ち時間を考慮して，S と I の間に曝露/潜伏 (E) を入れることもある．このように病態を表す状態を導入すると，各状態にいる個人の人数が時間とともにどう変化するかを与える常微分方程式系（SIR モデルと呼ばれる）を書き下すことができる：

$$\frac{dS}{dt} = -\lambda IS, \quad \frac{dI}{dt} = \lambda IS - \gamma I, \quad \frac{dR}{dt} = \gamma I.$$

この式は Δt だけ時間が経過する間に，感染者(I)と感受性者(S)が接触することで $\lambda I S \Delta t$ 人の新規感染者が発生するいっぽうで，I 人のうちの $\gamma I \Delta t$ が回復によって感染者集団から取り除かれることをあらわしている．SIR モデルには強い仮定が置かれている．集団内での人の接触が一様であることと感染期間が指数分布に従うことである．これらの仮定は現実的ではないが，特定の状況下では大集団内での感染動向はよく捉えることが知られており，SIR モデルおよびその派生系が広く疫学で用いられている．

SIR モデルの拡張には 2 通りの方向がある．ひとつは，伝染の成立を確率過程に置き換えることである．例えば，S から I に移る人数が二項分布 Bin ($n = S$, $p = \lambda I \Delta t$)に従うとすることが考えられる． もうひとつの拡張は，人と人の接触の仕方に居住地，年齢，性別などによる違いを取り入れることである. これらの属性で添え字づけた変数 S_i, I_i, R_i を用意し，これらの変数の数量変化を記述する SIR モデルと類似の方程式系によって実現できる．本項で扱う個人ベースモデルは，両方向の拡張を行ったときの最も複雑な場合と考えられる．すなわち，本来の SIR モデルが病態や居住地域・性別ごとの人数の変化を記述するのに対し，個人ベースシミュレーションでは確率的に進行する病態が個人毎に追跡される．ただし，どの程度個別性をシミュレーションに取り入れるかは目的に応じて変わる．Riley

(2007)によると，単純なものから複雑なものへ 4 つのタイプに分類される．

- パッチ型 – 対象地域を複数のサブ地域に分けて，それぞれの地区内で感受性者は等しい感染力に曝されると仮定する．
- 距離型 – 感受性者は感染者からの距離に反比例する感染力に曝される．
- グループ型 – それぞれのグループ内で，感受性者は等しい感染力に曝される．また，適当なグループ間の相互作用を導入することで，雑踏での伝染などを表現する．
- ネットワーク型 – 家族関係や友人関係など可能な接触をグラフ(ネットワーク)として表現する．

本発表ではインフルエンザを対象した個人ベースシミュレーションの利用事例を紹介した．特に，大規模シミュレーションをもとに政策提言を行った草分け的研究として，東南アジア一円での新型インフルエンザ封じ込め可能性の Ferguson らによる試算と，パーソントリップ調査を感染症流行に応用した研究として八島らの研究をとりあげた．

参考文献

- Riley, Large-Scale spatial-transmission models of infectious Disease, Science 316, 1298 (2007).
- Ferguson et al., Strategies for containing an emerging influenza pandemic in Southeast Asia, Nature 437(8), 209—214 (2005).
- Moser et al. An outbreak of influenza aboard a commercial airliner. American Journal of Epidemiology 110, 1-6 (1979).
- Cauchemez et al., A Bayesian MCMC approach to study transmission of influenza: application to household longitudinal data. Statistics in Medicine 23, 3469-3487 (2004).
- Yashima & Sasaki, Epidemic Process over the Commute Network in a Metropolitan Area. PLoS ONE 9(6) (2014): e98518. doi:10.1371/journal.pone.0098518

病原細菌の種内多数のゲノムデータから
組換えのホット領域を推定する手法の開発

○矢原耕史[1,2,5]、Xavier Didelot[3]、M. Azim Ansari[4]、Samuel K. Sheppard[5]、Daniel Falush[6]

[1]久留米大・バイオ統計; [2]東大院・新領域; [3]Dept. Infectious Diseases Epi., Imperial College London; [4]Dept. Statistics, Univ. Oxford; [5]College of Medicine, Swansea Univ.; [5]Dept. Evol. Genetics, Max Planck Institute

1. 背景

突然変異と組換えは、ゲノムに多様性を生み出す、生物の適応進化の源である。突然変異率はゲノム内の特定の領域で高いことが知られ、その疾患との関係が注目されている（Michaelson, 2012, Cell)。一方、組換えは、突然変異よりも検出が難しい。実験によってゲノム上の特定の領域に生じる組換えの回数を測定することは可能だが、自然界での組換えの痕跡の検出は別問題であり、その回数をゲノムに沿って推定すること自体が未解決の難問である。今世紀になって、種内多数の全ゲノム配列を比較するというアプローチが可能になり、それによってゲノム全域に渡って組換え率を推定し、特に組換え率が局所的に上昇している領域を推定することが、重要な課題になっている（Hinch 2011, Nature; Auton 2012, Science)。

この観点に基づき、演者はこれまで、病原細菌におけるゲノムの組換えに注目した研究に取り組んできた。まず、全生物の中で集団平均組換え率が最も高いとされるヘリコバクター・ピロリ菌を用い、ゲノム配列の比較に基づいて全遺伝子について組換えの回数を推定し、ほぼ全ての遺伝子で組み換えが生じていることを示した（Yahara 2012, *Genome. Biol. Evol.*)。さらに、ヒト集団遺伝学で開発された隠れマルコフモデルであるインシリコ染色体ペインティング法（図 1）（Lawson 2012, *PLoS Genetics*)）を応用し、 組換えの痕跡（ある個体の DNA 断片が別個体のゲノムに入り込んだモザイク構造）をゲノム全域に渡って明らかにした（Yahara 2013, *Mol. Biol. Evol.*)。

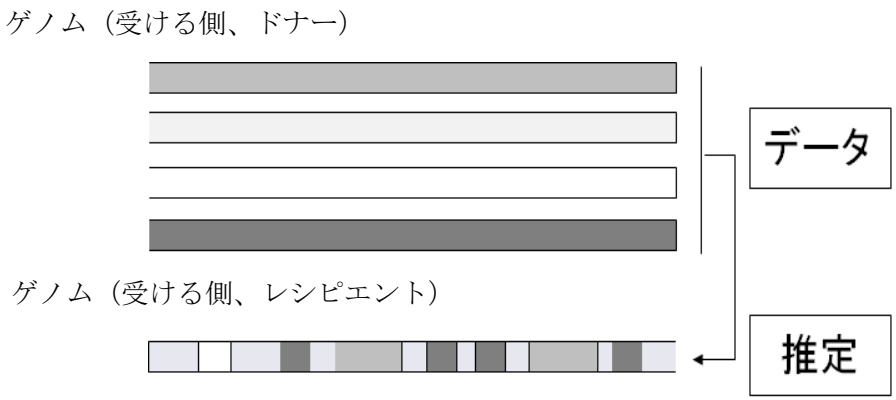


図 1: インシリコ染色体ペインティング法
（組換えの痕跡としてのモザイク構造を推定）

2. 内容

しかし、染色体ペインティング法は、ゲノム上のある領域に生じた最近 1 回の組換えしか検出できず、過去の履歴を考慮していない。そのため、組換えの生じた回数や、組換えの「ホット領域」（組換えが繰り返し生じた結果、個体間を高頻度で移動したように見える領域）を推定することは出来ない。組換えの「ホット領域」の例としては、髄膜炎菌の外膜タンパク質の遺伝子が知られ、それが菌のヒトへの感染との環境適応に重要な役割を果たすことが知られているが、そうした領域がゲノム内にどれだけ存在するのかは、一部の種（大腸菌）を除いて未解明のままだった。そこで本研究では、病原細菌のゲノムに沿って 1 塩基レベルで組換え率の指標を計算し、「ホット領域」を推定する新たな手法を開発した。

この手法の土台となっているのは、前述の「インシリコ・染色体ペインティング法」である。これは、N 個体のゲノム全域の 1 塩基多型とそのポジションを入力データとし、ある個体（レシピエント）のゲノム上の各塩基について、それが残り N-1 個体（ドナー）のどれに由来するかの確率（組換えの確率、コピー確率）を推定する。ここで、各塩基におけるドナーからレシピエントへのコピー確率を、N x N の行列として表現し、それがゲノム全体の平均からどれだけ乖離しているのかという観点から、塩基あたりの組換え率と強く相関する新たな統計量を考案した。病原細菌はクローン増殖するため、データの中にクローンが含まれていると、クローンが組換えのドナーとして推定されてしまうが、この点を解決するアルゴリズムを開発した。また、組換えのホット領域推定に関する bootstrap support value を開発した。さらに、入力データの中に欠損値が含まれる場合への対処法も開発した。

この手法を、まずシミュレーションデータに適用し、その感度・特異度を評価した。次に、大腸菌の 27 本の完全ゲノムデータに適用し、既知の組換えのホット領域が推定できることを確認した。さらに、人獣共通感染性細菌種の 200 本の不完全ゲノムデータ（次世代シーケンサデータ）に適用し、新たな組換えのホット領域（図 3）を明らかにした。これにより、100 本以上のゲノムの塩基配列データから組換えのホット領域を明らかにすることが、初めて可能となった。

そしてこの手法を、計算機クラスター上での並列計算によって高速な計算を可能とし、メモリ使用率を低く抑え、シンプルに使用可能なソフトウェア（https://github.com/bioprojects/orderedPainting）として実装し、一般公開した（Yahara 2014, *Mol. Biol. Evol*; Yahara in press, *Mol. Biol. Evol*; 矢原 2015, 化学と生物「今日の話題」）。

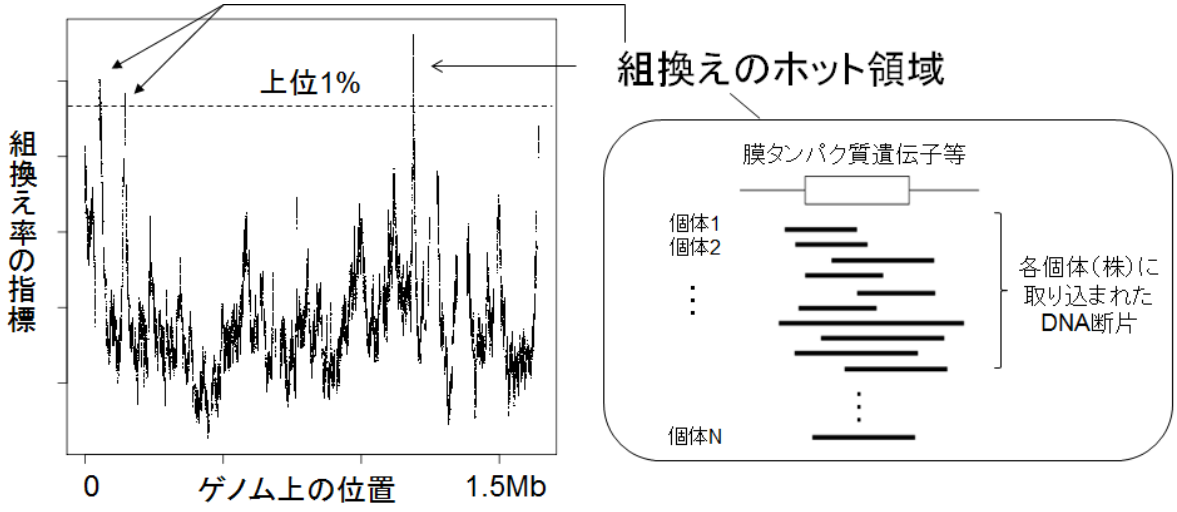


図 2: ゲノム全域に渡る組換え率の変化と、組換えのホット領域

グラフィカルモデルによる遺伝子の多重選択のアプローチ

鈴木 譲
大阪大学大学院理学研究科

Abstract

This paper proposes an estimator of mutual information for both discrete and continuous variables, and apply it to the Chow-Liu algorithm to find a forest that expresses probabilistic relations among them. The state of arts assumes the ANOVA model to estimate the mutual information when one variable is discrete and the other is Gaussian. As a result, it is hard to obtain the maximum likelihood of three connected vertices such that the center is Gaussian and the other two are discrete, so that the state of arts restricts the class to the forest such that no Gaussian node is between discrete nodes. The proposed method executes in a general setting without any assumption: preparing several histograms, computing the mutual information values, and choosing the maximum value. We prove how many histograms should be prepared and prove that the estimated mutual information is no larger than zero if and only if the variables are independent for a large sample size. Finally, we apply the proposed method to genome analysis, and in the experiments using gene expression and SNP (single nucleotide polymorphism) data, we demonstrate that the proposed method successfully captures the causality among them whereas the state of arts fails because the gene expression and SNP nodes are separated.

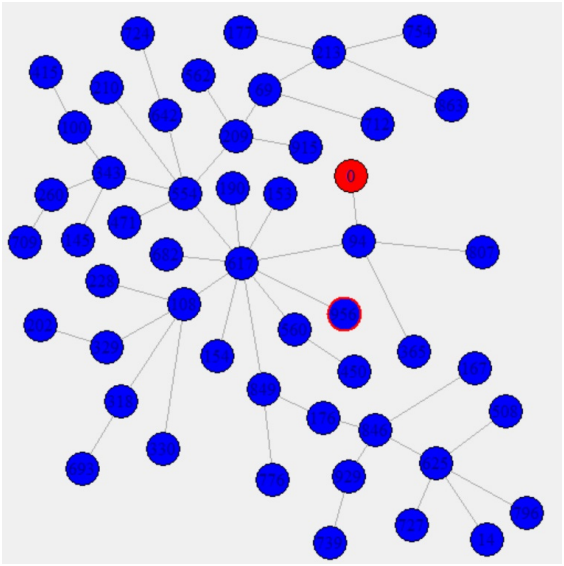


Figure 1: The forest consisting of expression data of the top 50 genes and the class, marked by red. The class node is connected only to one gene A.202580_x_at (94). .

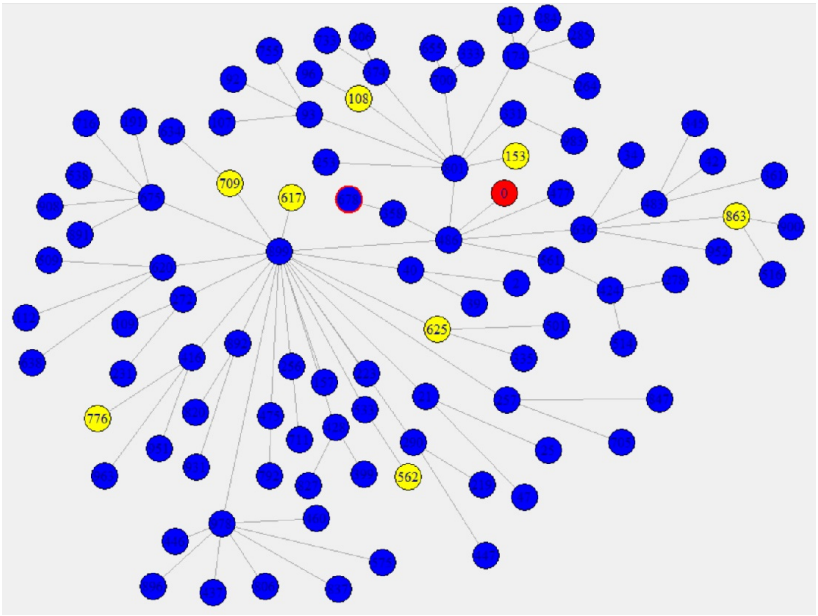


Figure 2: The forest consisting of expression data of all the 1000 genes and the class (the subgraph consisting of genes within diameter four from the class node is shown, and (the class and top 50 genes are marked by "red" and "yellow", respectively)). The class node is connected only to A.215303_at (486). .

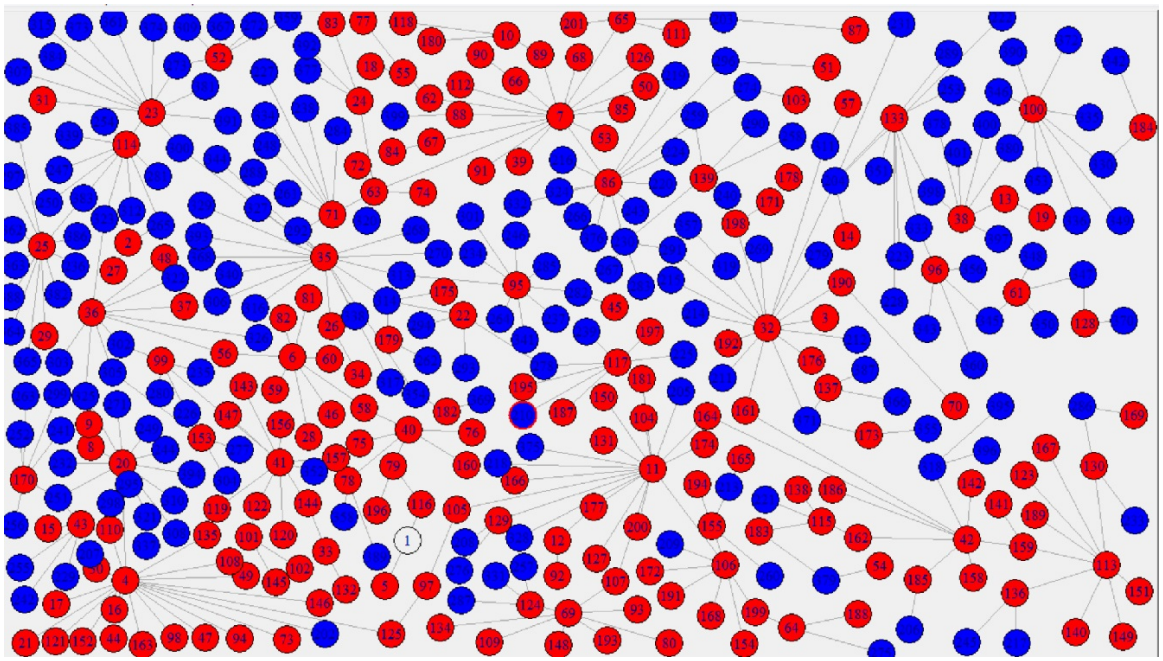


Figure 3: The forest for the 200 gene expressions, 200 SNPs, and male/female information: the discrete and continuous nodes are separated, and express the causality among the 401 variables (the gene expression and SNPs are marked by blue and red, respectively).

スパース正則化に基づく経時測定データの判別と遺伝子データ解析への応用

九州大学大学院数理学研究院　松井秀俊

1　概要

回帰モデルに含まれるパラメータ推定において，L_1 ノルムを含む制約を課したスパース正則化法は，推定量に安定性を与えると同時に変数選択の役割も担う方法として注目を集め，理論，応用の両側面から幅広い適用例が報告されている（例えば，Hastie et al., 2015）．本報告では，複数の個体に対して経時的に観測，測定されたデータを関数化処理し，得られた関数化データ集合（Ramsay and Silverman, 2005）に対してスパース正則化を適用し，変数を選択する方法について述べる．

関数データに基づくロジスティック回帰モデルにスパース正則化を適用することで，パラメータの推定と同時に，判別に寄与している変数の選択を試みる．制約としては，elastic net (Zou and Hastie, 2005) 型および sparse group lasso (Friedman et al., 2010) 型に基づく二種類の制約を紹介し，それぞれがもたらす効果について紹介する．モデルの推定については，正則化最尤法の枠組みで推定法を紹介し，それに伴う制約の形状を紹介する．そして，これらの手法を，それぞれ多発性硬化症に関する遺伝子発現データの解析と，イースト遺伝子発現データ解析へ適用した結果について報告する．

2　関数ロジスティック回帰モデル

いま，関数データとして与えられた n 個の p 変量説明変数 $x_{ij}(t)$ $(i = 1, \ldots, n,\ j = 1, \ldots, p)$ と，L (≥ 3) 群のラベルを表す二値変数 y_{il} $(l = 1, \ldots, L-1)$ が得られたとする．このとき，関数ロジスティック回帰モデルは，データが観測された下での l 群への判別確率 $\pi_l(\boldsymbol{x}_i; \boldsymbol{b})$ を用いて次で与えられる．

$$\log\left\{\frac{\pi_l(\boldsymbol{x}_i;\boldsymbol{b})}{\pi_L(\boldsymbol{x}_i;\boldsymbol{b})}\right\} = \beta_{0l} + \sum_{j=1}^{p}\int x_{ij}(t)\beta_{jl}(t)dt. \quad (l = 1, \ldots, L-1) \tag{1}$$

ここで β_{0l} は定数項，$\beta_{jl}(t)$ は係数関数で，$\boldsymbol{b}$ はモデルに含まれる未知パラメータベクトルとする．説明変数および係数関数が基底関数展開によって表現できるという仮定を置くことで，モデル (1) は次式のように，ベクトルと行列に基づく回帰モデルの形で表現できる．

$$\log\left\{\frac{\pi_l(\boldsymbol{x}_i;\boldsymbol{b})}{\pi_L(\boldsymbol{x}_i;\boldsymbol{b})}\right\} = \sum_{j=1}^{p} Z_j \boldsymbol{b}_{jl}. \tag{2}$$

ただし Z_j は説明変数から構成される既知の行列，$\boldsymbol{b}_{jl}$ はパラメータベクトルである．

3　モデル推定

関数ロジスティック回帰モデル (2) の係数パラメータ $\boldsymbol{b}$ を正則化最尤法，すなわち次で定義される正則化対数尤度関数の最大化により推定する．

$$\ell_\lambda(\boldsymbol{b}) = \sum_{i=1}^{n} \log f(\boldsymbol{y}_i|\boldsymbol{x}_i;\boldsymbol{b}) - nP_{\lambda,\alpha}(\boldsymbol{b}).$$

ただし $P_{\lambda,\alpha}(\boldsymbol{b})$ はパラメータに対する制約関数で，制約そのものの強さを規定する正則化パラメータ $\lambda > 0$ と，追加の調整パラメータ $\alpha \in [0,1]$ である．これに L_1 ノルムを含む関数を仮定することで，いくつかのパラメータを 0 と推定できる．本報告では，次の二種類の制約と，これらがもたらす効果について紹介する．

$$P_{\lambda,\alpha}(\boldsymbol{b}) = \frac{1}{2}(1-\alpha)\sum_{j=1}^{p}\lambda_j\sum_{l=1}^{L-1}\|\boldsymbol{b}_{jl}\|_2^2 + n\alpha\sum_{j=1}^{p}\lambda_j\left\{\sum_{l=1}^{L-1}\|\boldsymbol{b}_{jl}\|_2^2\right\}^{\frac{1}{2}}$$

$$P_{\lambda,\alpha}(\boldsymbol{b}) = n(1-\alpha)\sum_{j=1}^{p}\lambda_j\left\{\sum_{l=1}^{L-1}\|\boldsymbol{b}_{jl}\|_2^2\right\}^{1/2} + n\alpha\sum_{j=1}^{p}\lambda_j\sum_{l=1}^{L-1}\|\boldsymbol{b}_{jl}\|_2.$$

ただし $\lambda_j = \sqrt{M_j}\lambda$，$M_j$ はパラメータベクトル $\boldsymbol{b}_{jl}$ の次元とする．いずれの制約についても，解析的に推定量を導出することは困難であるため，反復的に推定量を導出する方法について紹介する．

4　適用例

提案した手法を，経時測定された二種類の遺伝子発現データへ適用する．4.1 節，4.2 節でそれぞれ，3.1 節，3.2 節で述べた制約を適用したデータおよび推定方法について紹介する．結果の詳細については当日報告する．なお，4.1 節の結果の詳細については Kayano et al. (2015) で述べられている．

参考文献

Friedman, J., Hastie, T., and Tibshirani, R. (2010), A note on the group lasso and a sparse group lasso, *arXiv preprint arXiv:1001.0376*.

Hastie, T., Tibshirani, R., and Wainwright, M. (2015), *Statistical Learning with Sparsity: The Lasso and Generalization*, Boca Raton: Chapman & Hall/CRC.

Kayano, M., Matsui, H., Yamaguchi, R., Imoto, S., and Miyano, S. (2015), Gene set differential analysis of time course expression profiles via sparse estimation in functional logistic model with application to timedependent biomarker detection, *Biostatistics*, To appear.

Ramsay, J. and Silverman, B. (2005), *Functional data analysis 2nd ed.*, New York: Springer.

Zou, H. and Hastie, T. (2005), Regularization and variable selection via the elastic net, *J. Roy. Statist. Soc. Ser. B*, 67, 301–320.