研　究　報　告　書

「仮説世界と物理世界の相互浸透モデリングによる知の創生」

研究タイプ：通常型
研究期間：　平成 20 年 10 月～平成 26 年 3 月
研 究 者：　大羽　成征

1.　研究のねらい

本研究課題の対象は、観測データに基づいて世界を理解するために行われる統計的モデリングの方法論全般である。個々具体的な問題をワークベンチとしながら、抽象的なフレームワークとしての新機軸を打ち出すことをねらう。

観測に基づいてモデルを現実の物理世界に近付けることを目的とするとき、その評価は近未来予測値の誤差で行われる。例えば、データ同化と呼ばれる確率的モデリングの方法論がその典型例であり、密度・精度の限られた観測のもとで気象予測成果を上げるなどの効果が知られている。一方で、モデリングの主要目的が、科学的インパクトの大きな「発見」を行うことである場合（典型的にはシステム生物学・医学の分野におけるモデリング）には、その評価基準は統計的仮説の検定精度（第一種・第二種過誤）である。主要目的が異なればそれに応じてモデリング方法論も異なってしかるべきである。

そこで、本研究課題では、科学的インパクトのある発見と統計的検定精度を目的とする場合の物理世界モデリングの方法論を考える。まず複雑な構造を持った仮説がある場合の構造利用の方法をさぐる。これを「仮説世界ベースモデリング」と呼ぶ。次に物理世界モデルからのフィードバックによって仮説の構造が変化する場合を想定した方法を探る。これを「相互浸透モデリング」と呼ぶ。

またチャレンジングな実問題におけるモデリング研究のケーススタディを行うなかで本コンセプトの有効性を実証し，仮説世界と物理世界との間の相互浸透の様々な様式を見いだし，さらには科学的な「知」の創出過程に関する新しい知見に至ることをねらう。

2.　研究成果

（1）概要

理論的成果の第一は、多重検定におけるオンデマンド統計量の構成方法[1]の提案である。多重検定とは、多数の帰無仮説のもとで統計的仮説検定を同時に行う問題のことであり、仮説世界ベースモデリングの理論的基盤である。複数検定の根拠となる観測データが共通の物理世界を対象にしているとき、これを利用した検出力向上の余地がある。物理世界の確率的観測モデルの十分統計量を用いて検定統計量デザインすることで、検出力が最大化できることを示した。

多重検定問題では、検定対象となる複数仮説が、背景に同一の物理モデルを共有している場合が多い。そこで、物理モデルに含まれる未知モデルパラメタの推定結果を、仮説検定の問題設定にフィードバックさせることが合理的である。物理世界を構成する、未知変数と未知パラメタと、仮説世界を構成する隠れ変数とを同時にベイズ推定する相互浸透型の構造を与えることで、その外側で行う多重検定の検出力を改善することができることを示した。

実問題応用として、以下のような成果を得た。

遺伝子発現量に基づく癌悪性度関連遺伝子の検出力を大幅に改善できることを示した[2]。これは多次元のオンデマンド特徴量を用いることで、癌に関係ない遺伝子発現変動の情報を利用できるようになったことによる。

マウス胚の体節形成システムを調べる共同研究において、ベイズ的階層モデルに基づく検定統計量をデザインし、遺伝子ノックアウトの有無に基づく相違の検出力向上に貢献した[3]。

血中に溶け出したがん細胞由来 DNA を次世代シーケンサによって測定することによる新しい侵襲性の低い肺癌診断法を開発する共同研究において、検出精度を高めた[4]。

神経細胞のカルシウムイメージング時系列データから、ベイズ的系列モデルに基づくオンデマンド統計量によって、蛍光量時系列に見られる神経スパイクの検出精度を改善した。また、こうして得られた複数神経細胞のスパイクデータから、細胞間の機能的結合（グレンジャー因果）を検出した。このさい、オンデマンド統計量に基づく経験ベイズ検定を使用することで、FDR(false discovery rate)を制御しつつ検出力を改善できることを示した[投稿中]。

（2）詳細

研究テーマ A「仮説世界ベースモデリングおよび相互浸透モデリングの理論的基盤の確立」

仮説世界ベースモデリングの理論的基盤として、多重検定の性能を最適化するオンデマンド統計量の構成方法[1]を提案した。多重検定とは、多数の統計的仮説検定を同時に行う問題の総称である。統計的仮説検定の目的の第一は、帰無仮説が成立しているにもかかわらず、対立仮説を誤って採択してしまう「第一種過誤」の確率を、予め定めた値（有意水準）未満に抑えること。目的の第二は、第一の目的が果たされている保証のもとで、なお帰無仮説を棄却してインパクトのある発見を主張できる確率「検出力」を高めることである。多重検定でもこれらの目的に違いは無いが、(a) 第一種過誤を定義するさいに多重検定特有の補正を必要とすることと、(b) 検出力を高めるために、検定で用いる情報に共通部分があることを利用した工夫の余地がある。(a)を満たしつつ(b)を実現するための方法として、最適統計量の理論(Storey ら, 2006)が知られていたが、真の物理世界モデルに対応する尤度関数のパラメタが完全に既知であるとする実用上非現実的な仮定が必要とされており、近似の精度は悪かった。オンデマンド特徴量の理論[1]は、尤度関数の十分統計量を検定統計量として用い、経験ベイズ検定を併用することで、共通情報を最大限に用いて検出力を最大化できることを示した。

相互浸透型モデリング構想の射程は広いため、まずは簡単な原型モデルを提案した。多重検定問題では、検定対象となる複数の仮説が、背景に同一の物理モデルを共有している場合が多い。そこで、物理モデルに含まれる未知モデルパラメタの推定結果を、仮説検定の問題設定にフィードバックさせることが合理的である。これを、図 1 のような隠れ変数モデルで表現し、相互浸透型モデル原型と呼ぶことにした。物理世界を構成する、未知変数 X と未知パラメタΘ、仮説世界を構成する隠れ変数 B とを同時にベイズ推定することで、その外側で行う多重検定 h の検出力を改善できることを示した。

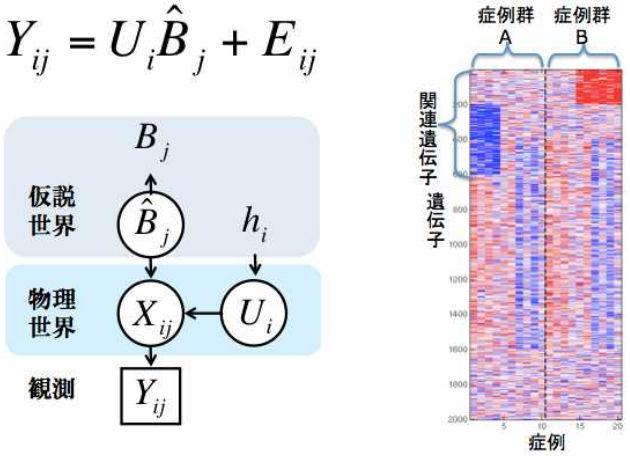

図 1: 相互浸透型の単純な原型（左）とこれが想定している多重検定問題（右）

研究テーマ B「仮説世界ベースモデリングと相互浸透モデリングの実世界問題への応用」

[2] 因子統計量に基づく 2 標本平均差検定の検出力改善

マイクロアレイなどの方法による遺伝子発現量の網羅的測定に基づいて癌悪性度関連遺伝子を見つけたい。このとき、悪性症例と良性症例の間で各遺伝子の発現量の平均値が異なることを、それぞれの遺伝子で検定する多重検定を適用するのが標準的な方法である。しかし、厳密な統制に基づく 2 標本検定と違い、臨床症例群間の比較では、癌悪性度以外の様々な特徴量を背景に含むため、比較対照群毎の平均値で説明できない様々な背景変動因子の影響を含んでおり、これに基づく強い検定間相関が想定される。図 1（右）はその典型を示した人工データである。

本研究では、こうした背景変動因子を陽に取り入れた多次元の因子統計量を検定統計量として用い、経験ベイズ検定法を適用した。これを用いた検定により、背景変動因子を顕著に含むデータにおける関連遺伝子検出力が顕著に改善された（図 2）。これは成果[1]のオンデマンド統計量デザインポリシーに従った応用の成果である。

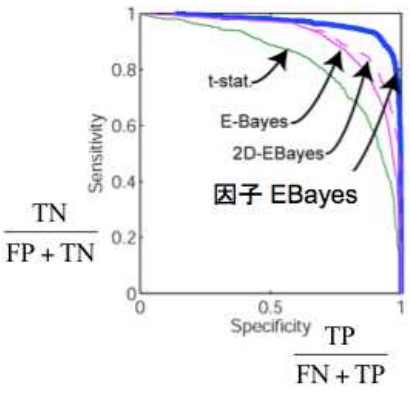

図 2 背景変動因子が存在するとき、オンデマンド統計量に基づく経験ベイズ法（因子 EBayes）による関連遺伝子検出力は、他の方法による検定によるものよりも顕著に高い。

[3] マウス胚の体節形成システムにおける遺伝子ノックアウトの有無に基づく相違の検出力向上

マウスの胚は、その発生時に生化学反応システムが規則正しいリズムを刻むことによって、規則正しい体節構造を作っていると考えられており、そのシステムにおいて Nrarp 遺伝子が大事な役割を果たしていると考えられている。これを検証するために、Nrarp 遺伝子をノックアウトしたマウスとそうでないマウスの間で、特定タイミングにおける体節数の平均に有意な差があることを示した。このさい、ベイズ的階層モデルに従う特別な統計的検定法をデザインして用いた。

マウス胚の体節数データは、複数母胎それぞれ複数の胎児について計測したものである。受精後の時間経過を揃えたとしても、マウスの胚の成長度は個体間である程度ばらつく。また個体間のばらつきは、母胎にも依存する。そこで、母胎を同じくする胎児の中で Nrarp 遺伝子ノックアウトの有無による体節数の差を比較するべく、ベイズ的階層モデルを立て、これに基づく検定統計量をデザインすることで検出力を高めた。なお、あえてこうした構造を無視した単純な 2 標本検定を同データに適用した場合には、有意差は無かった。

[5]次世代 DNA シーケンサによる低浸襲肺癌診断法開発と、そのための相互浸透型モデリング

分子標的薬イレッサは EGFR 遺伝子に変異のある肺がん患者に選択的に投与されることで初めて治療効果が期待できる抗がん剤であるが、がん組織採取による遺伝子変異の検査は患者の負担が大きい。そこで血液検査など低浸襲検査による代替が可能であれば、医療に対する貢献は大きい。本研究では、次世代 DNA シーケンサによる DNA カウントを用いた血液中残渣 DNA に含まれる、ごく微量の腫瘍由来 DNA から EGFR 変異の検出法を開発した。

大羽は、次世代シーケンサ特有のノイズ要因に基づく帰無仮説モデリングを行い、保守的な検出基準をデザインする箇所で本研究に貢献した。

3. 今後の展開

オンデマンド統計量および相互浸透型モデルに基づく多重検定は、多くの要因が複雑に絡み合う現実世界の観測から知識創生してゆくための基盤技術になる。相互浸透型モデルについては、原型を発展させ、より複雑な基礎方程式を含むような物理世界モデルとドッキングした場合の挙動を調べてゆきたい。

4. 評価

(1)自己評価

仮説世界ベースモデリングの構想についてはオンデマンド統計量のアイディアを与え、相互浸透型モデリングの構想については原型モデルの姿を明確にするところまでを、さきがけ研究期間に達成し、また多くの実世界問題への応用例を示せたことに満足している。

一方で、複雑な構造をもつ物理世界を背景として持つ問題(ヒト身体運動や、ヒト脳機能)におけるモデリング研究については、まだ形になっていない。当初構想の中にある新しい世界観に、社会的インパクトを伴わせるためには、こうした難しい問題をスッキリと解くデモンストレーションを重ねてゆくことが必要である。

（2）研究総括評価（本研究課題について、研究期間中に実施された、年2回の領域会議での評価フィードバックを踏まえつつ、以下の通り、事後評価を行った）。

仮説世界と物理世界の相互浸透モデリングというユニークな視点に立ち、個別仮説ベースの仮説検定とベイズ的な事前知識(物理モデル)を融合させようという課題である。5年間で医療分野等の実際のフィールドで使われる伝統的な検定論と、発見的なベイズ的アプローチの間を埋める実際的かつ統一的な理論が得られることを期待していた。

理論的成果として、相互浸透型の構造を与えることで、その外側で行う多重検定の検出力を改善することができることを示した。また、実問題応用として、遺伝子発現量に基づく癌悪性度関連遺伝子の検出力の大幅改善、ベイズ的階層モデルに基づく検定統計量のデザインにより、遺伝子ノックアウトの有無に基づく相違の検出力向上等、医療分野での検出力向上に大きく寄与している。

当初、さきがけ期間中に取り掛かりたいとしていた複雑な構造をもつ物理世界を背景として持つ、ヒトの身体運動や脳機能等のモデリング研究については、手がつかなかったとのことであるが、相互浸透型モデリングの原型モデルの姿を明確にし、多くの実世界問題への応用例を示したことを高く評価したい。

5. 主な研究成果リスト

（1）論文（原著論文）発表

1. Oba, S. and Ishii, S. Optimal Sufficient Statistics for Parametric and Non-Parametric Multiple Simultaneous Hypothesis Testing. The International Journal of Biostatistics. (2009), 5, 1, Article 20.
2. Oba, S. and Ishii, S. Differential gene detection incorporating common expression pattern. Journal of Physics, Conference Series. (2010), 197(012007).
3. Kim, W., Matsui, T., Yamao, M., Ishibashi, M., Tamada, K., Takumi, T., Kohno, K., Oba, S., Ishii, S., Sakumura, Y., and Bessho, Y. The period of the somite segmentation clock is sensitive to Notch activity. Molecular Biology of Cell. (2011), 22(18), 3541-3549.
4. Takahashi, N., Oba, S., Yukinawa, N., Ujita, S., Mizunuma, M., Matsuki, N., Ishii, S., and Ikegaya, Y. High-speed multineuron calcium imaging using Nipkow-type confocal microscopy. Current Protocols in Neuroscience, (2011) 2:Unit2.14.
5. Kukita, Y., Uchida, J., Oba, S., Nishino, K., Kumagai, T., Taniguchi, K., Okuyama, T., Imamura, F., and Kato K. Quantitative identification of mutant alleles derived from lung cancer in plasma cell-free DNA via anomaly detection using deep sequencing data. PLoS ONE. (2013), 8(11): e81468.

（2）特許出願

研究期間累積件数：0 件

（3）その他の成果（主要な学会発表、受賞、著作物、プレスリリース等）

1. Kouno, M., Nakae, K., Oba, S., and Ishii, S. (2012) Microscopic image restoration based on

tensor factorization of rotated patches. International Symosium on Artificial Life and Robotics, 902-905.

2. Aki, S., Oba, S., Nakae, K., and Ishii, S. (2012) A sparse random method to estimate neuronal structure from spike sequence. International Symosium on Artificial Life and Robotics, 718-721.