平成２５年７月２２日

科学技術振興機構（ＪＳＴ）
Tel：03-5214-8404（広報課）

産業技術総合研究所
Tel：029-862-6216（報道室）

東京工業大学
Tel：03-5734-2975（広報センター）

理化学研究所
Tel：048-467-9272（広報室）

# ビッグデータから新たな科学的発見をもたらす統計手法を開発

**ポイント**

- ビッグデータからの科学的発見のためには、正確な検定値（Ｐ値）の算出が必要。
- 超高速アルゴリズムを用いた新たな統計検定手法を開発し、発見力を大幅に改善した。
- 物理学、医学、化学など全ての実験科学において世界中での広い利用が期待される。

ＪＳＴ 課題達成型基礎研究の一環として、産業技術総合研究所 生命情報工学研究センターの津田 宏治　主任研究員（ＪＳＴ ＥＲＡＴＯ「湊離散構造処理系プロジェクト」グループリーダー）、東京工業大学 大学院情報理工学研究科 計算工学専攻の瀬々 潤 准教授、理化学研究所 統合生命医科学研究センターの岡田 眞里子　チームリーダーらは、従来に比べて格段に高い精度で誤発見の確率を示す検定値（Ｐ値[注1)]）を計算するアルゴリズム（手順）を開発しました。

自然科学で得られるデータ量は増加の一途をたどり、これらを有効に解析できる方法が望まれています。しかし、従来の統計検定手法は観測できる対象が増えれば増えるほど、発見の基準を厳しくしなくてはなりません。その結果、観測対象が増えたのに、科学的発見が減るという奇妙な現象「ビッグデータのパラドックス」が起きる場合があります。特に、複合的な組み合せ因子に対して極めて保守的な検定値（Ｐ値）を出すことが多く、有意義な実験結果が不当に低く評価されることがありました。

本研究グループでは、超高速アルゴリズム[注2)]の技法を用いて、従来法より、格段に精度の高いＰ値を算出する新手法を開発しました。この手法を、乳がん細胞株の増殖・分化に関与している転写因子の研究に利用したところ、既存の遺伝子発現データから新たな組み合わせ因子を発見することに成功しました。

開発した手法を用いれば、これまで見過ごされてきた組み合わせ因子の発見が可能になります。本成果は、物理学、医学、化学など、全ての実験科学に貢献するものであり、今後世界中で広く利用されることが期待されます。

本研究成果は、米国科学雑誌「米国科学アカデミー紀要（ＰＮＡＳ）」のオンライン速報版で２０１３年７月２２日（米国東部時間）の週に公開されます。

本成果は、以下の事業・研究領域・研究課題によって得られました。
戦略的創造研究推進事業　ＥＲＡＴＯ型研究
研究プロジェクト：「湊離散構造処理系プロジェクト」
研 究 総 括：湊 真一（北海道大学 大学院情報科学研究科　教授）
グループリーダー：津田 宏治（産業技術総合研究所 生命情報工学研究センター　主任研究員）
研 究 期 間：平成２１年度～平成２６年度
上記研究課題では、超高速アルゴリズムを用いて、実問題を短時間に効率よく処理する技術基盤の構築を目指します。

**＜研究の背景と経緯＞**

自然科学では新しい現象を見つけたとき、系のゆらぎや観測のあいまいさを考慮した上で、その結果の信頼性を担保する必要があります。科学データの解析において、この信頼性担保には、統計検定が欠かせません（図１）。統計検定では、誤発見の確率を示す検定値（Ｐ値）が計算され、あるしきい値（一般には、０．０５）以下の場合にのみ、信頼しうる科学的発見として認められ、論文に記すことができます。

観測できる対象（例：ＤＮＡの変異）が増えると、誤発見の確率も高くなります。誤発見を避けるには、対象数が増えれば増えるほど、発見の基準を厳しくしなくてはなりません。一般的な多重検定法[注３）]では、Ｐ値に大きな補正係数を掛けて（補正Ｐ値）、それでも０．０５以下の場合のみ発見とみなします（図２）。最もシンプルでよく用いられるボンフェローニ法[注４）]では、ｎ個の対象があれば、Ｐ値にｎを掛けて補正し、それでも０．０５以内であれば、発見として認めます。その結果、観測対象が増えたのに、科学的発見が減るという奇妙な現象「ビッグデータのパラドックス」が起きる場合があります。

特に、複合的な組み合わせ因子を考えると（図３）、対象数ｎが爆発的に大きくなるため、ほぼ発見は不可能となってしまいます。このため、細胞のｉＰＳ化を引き起こす４つの転写因子などに見られる組み合わせ因子を、データから見つけ出すことは困難でした。

**＜研究の内容＞**

本研究では、従来よりも格段に正確な補正Ｐ値を計算できるアルゴリズムＬＡＭＰ（Ｌｉｍｉｔｌｅｓｓ－Ａｒｉｔｙ Ｍｕｌｔｉｐｌｅ ｔｅｓｔｉｎｇ Ｐｒｏｃｅｄｕｒｅ、無限次数多重検定法）を開発しました。ＬＡＭＰでは、出現頻度の低い組み合わせは誤発見率を変化させないという数理的性質に注目し、超高速アルゴリズムを用いて無為な出現頻度の低い組み合わせを特定し取り除くことによって、補正係数を大幅に削減しています。またＬＡＭＰでは通常のボンフェローニ法と比べて、統計的な検定の精度を保ったままで、補正係数を十分に低くすることができます。この手法を用いて、ヒトの乳がん細胞株の遺伝子発現データを再解析したところ、これまで見過ごされてきた、最大８個の転写因子の組み合わせが乳がん細胞の増殖に関与していることを発見できました。

出現頻度の低い組み合わせが誤発見率を変化させないという事実は、１９９０年に米国のタローネによって明らかになっていましたが、アルゴリズムを用いて、それらを実際に数えあげて、生命科学データに適用したのは世界初です。生命科学で広く用いられているＦＤＲ[注５）]による方法では、誤発見率については妥協することで、発見力を高めていますが、この手法ではそのような妥協をせず、アルゴリズムのみによって発見力を大幅に高めることに成功しました。

**＜今後の展開＞**

本成果により、転写因子の組み合わせ効果の研究をはじめ、複数の遺伝子が原因となっている疾患の同定や多数の部位が関わる脳の高次機能の解明など、複合要因に起因する現象の解明が加速されることが期待されます。さらに、複数の薬剤を組み合わせた創薬、多数の項目からなるアンケートの分析など、広く自然科学から社会科学分野の実験結果の評価に影響を及ぼすと考えられます。

＜参考図＞

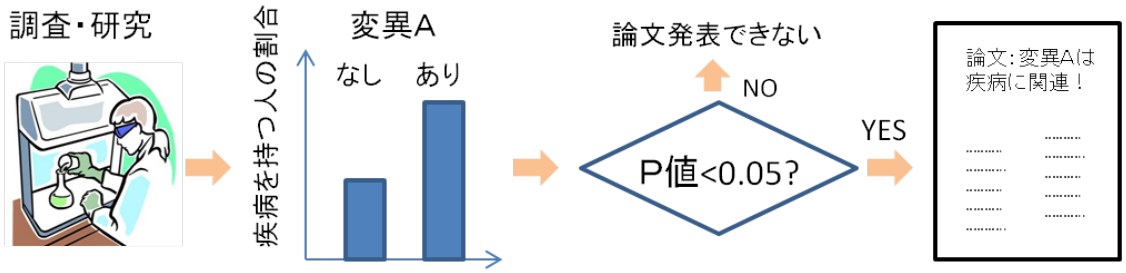


**図１　科学における統計検定の役割**

データからある結果を主張する際には、信頼性を評価するため統計検定を行わなくてはならない。ほとんどの科学雑誌では、Ｐ値のない結果を出版することはできない。

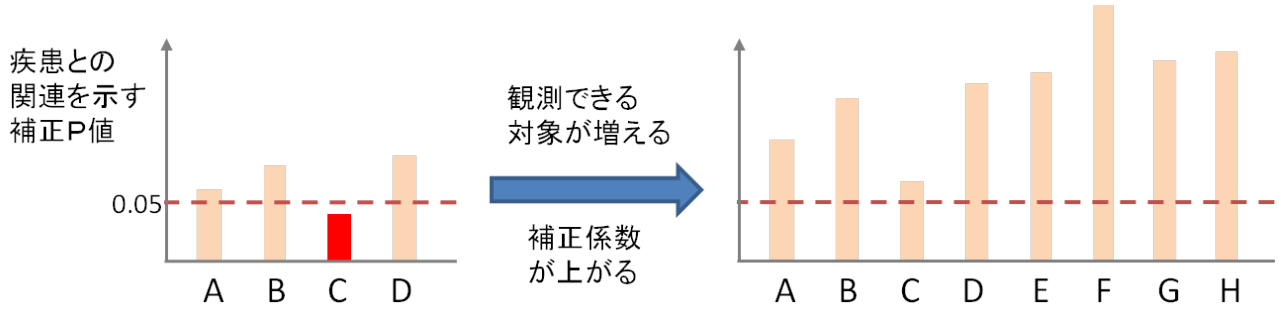


**図２　ビッグデータのパラドックス**

多重検定補正によって観測対象が増加しても、科学的発見につながらないことがある。左図では、対象数が少ないため、対象Ｃの補正Ｐ値は０．０５以下であり、発見として認められるが、８個に増加すると、補正係数が上がり補正Ｐ値が０．０５を越えてしまい発見として認められない。

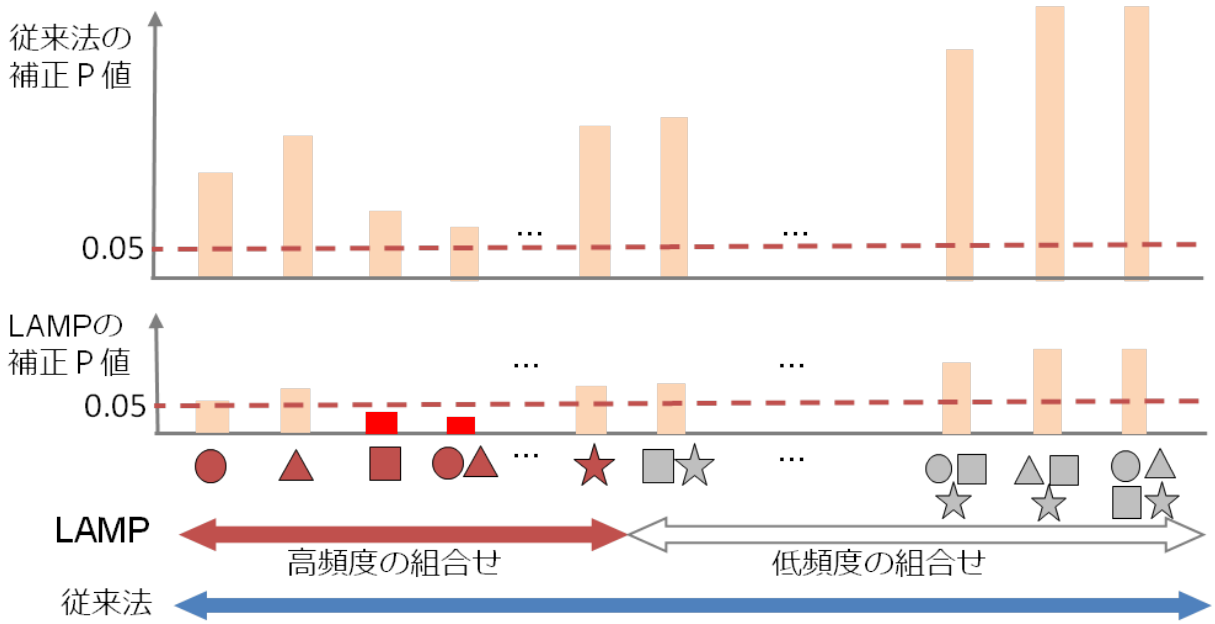


**図３　ＬＡＭＰによる組み合わせ因子発見**

従来のボンフェローニ法では、全ての組み合わせ因子の数を補正係数として用いるのに対し、ＬＡＭＰでは、高頻度の組み合わせのみを数え上げることによって、補正係数を正当なレベルまで引き下げることができる。それにより、赤色で示した組み合わせ因子は、発見として認められる。頻度のしきい値は、アルゴリズムによって自動的に決定される。

**＜用語解説＞**

注１）　Ｐ値

　データから発見された事柄が誤りである確率のことです。ＡはＢを引き起こすという事柄に関するＰ値は、ＡとＢが完全に独立であるという仮定（帰無仮説）のもとで、得られた観測データより極端なものが得られる確率として計算されます。

注２）　超高速アルゴリズム

　コンピューターによって、膨大な組み合わせの数え上げなどの複雑な計算を超高速に実行する演算手順のことです。湊離散構造処理系プロジェクトでは、超高速アルゴリズムの技法を研究開発しており､例えば電力網などのシステム検証や最適化､データマイニング、知識発見などを含む分野横断的かつ大規模な実問題を高速に処理するための技術基盤を構築しています。

注３）　多重検定法

　複数の対象に対して同時に検定を行う場合には、各々の誤発見の確率を抑えるだけでは不十分です。例えば、各々の誤発見率が５％でも、１０個の対象がある場合には、一回でも誤発見が起きる確率（Ｆａｍｉｌｙ－ｗｉｓｅ　Ｅｒｒｏｒ　Ｒａｔｅ）は、最大１０倍の５０％にもなります。多重検定法では、Ｆａｍｉｌｙ－ｗｉｓｅ　Ｅｒｒｏｒ　Ｒａｔｅが５％以内に収まるよう、Ｐ値に補正係数を掛けて調整します。

注４）　ボンフェローニ法

　ボンフェローニ法は、最もシンプルでよく用いられる多重検定法です。この方法では、ｎ個の対象があれば、Ｐ値にｎを掛けて補正し、それでも５％以内であれば、発見として認めます。その結果、Ｆａｍｉｌｙ－ｗｉｓｅ　Ｅｒｒｏｒ　Ｒａｔｅを必ず５％以下に抑えることができます。

注５）　ＦＤＲ

　Ｆａｌｓｅ　Ｄｉｓｃｏｖｅｒｙ　Ｒａｔｅの略です。Ｆａｍｉｌｙ－ｗｉｓｅ　Ｅｒｒｏｒ　Ｒａｔｅが、誤発見が一回でも起きる確率を指すのに対し、ＦＤＲは、発見された対象のうち誤っているものの割合を指します。Ｆａｍｉｌｙ－ｗｉｓｅ　Ｅｒｒｏｒ　Ｒａｔｅは５％以下でなくても、ＦＤＲを５％以下に抑えればよいとするのが、ＦＤＲに基づく多重検定法で、生命科学で広く用いられています。

**＜論文タイトル＞**

“Statistical Significance of Combinatorial Regulations”
（組み合わせ制御の統計的有意性）

**＜お問い合わせ先＞**

瀬々 潤（セセ ジュン）
東京工業大学 大学院情報理工学研究科 計算工学専攻
〒152-8550 東京都目黒区大岡山２－１２－１
Tel／Fax：03-5734-3526
E-mail：sesejun@cs.titech.ac.jp