福島県立医科大学 学術機関リポジトリ

FUKUSHIMA MEDICAL UNIVERSITY

| | |
|---|---|
| Title | NIRF研究会報告「情報と確率の等価性原理」：学術活動 |
| Author(s) | 森, 努 |
| Citation | 福島県立医科大学看護学部紀要. 25: 31-41 |
| Issue Date | 2023-03 |
| URL | http://ir.fmu.ac.jp/dspace/handle/123456789/1983 |
| Rights | © 2023 福島県立医科大学看護学部 |
| DOI | |
| Text Version | publisher |

This document is downloaded at: 2024-04-25T18:57:55Z

Fukushima Medical University

# 学　術　活　動

## NIRF 研究会報告「情報と確率の等価性原理」

森　　努（看護学部生命科学部門）

### はじめに

NIRF 研究会は，クロマチン蛋白 NIRF ／ UHRF2の機能解析を目標に始められた[1)]。従来の分子生物学に従えば，分子機能や病態関連性を実験で追求すべきところである。だが構造解析から，NIRF は多数の細胞周期制御因子をはじめとする，広範なタンパク質と相互作用を示す「ネットワーク中心性」によって特徴付けられることが明らかとなった[2, 3)]。そこで私たちは，実験的手法に依存する通常の遺伝子機能解析ではなく，数理的手法に基づくシステム生物学的解析を指向することとした。この数理的手法では，解析対象を選ばない普遍性が得られるというメリットも生じる。

以上の判断に基づいて開発を進めた結果，すべての遺伝子を対象とし，実験に依存しない機能解析法である *ab initio* 遺伝子軌道法を開発することができただけでなく，その理論背景を保証する「情報と確率の等価性原理」をはじめ，多くの定理を発見・証明することができた。この経緯は昨年度の研究会報告[4)] および公開論文[5)] に概略を記したので，ご参照頂ければ幸いである。

### 遺伝子情報のシステム的認識

遺伝情報の複雑さを思い起こして欲しい。各遺伝子の機能および遺伝子間の相互作用を一般的に表現することは，元来甚だ困難と言わざるを得ない。40億年の進化過程で獲得した遺伝子情報は莫大であり，遺伝子ごとに全く異なる構造・機能を発揮するばかりか，発現様式も極めて多彩だからである。そうした複雑な遺伝子同士の相互関係は，より広範な情報交換を産みだす源泉であり，精細な相互作用に基づいて深淵な情報処理が行われている。結果的にこれらの遺伝情報がコードする中枢神経系・免疫系は，人知の到底及ばぬほど精緻に進化した大規模ネットワークシステムであり，宇宙規模とも言われるほど莫大な情報の統合処理が実現している。

翻って，個々の遺伝子は大規模情報処理システムを構成する情報処理装置としての部品的役目を担うのだから，旧来の分子生物学で扱い得る分子機能を超えて，現時点では予測も付かない複雑な生理機能にまで貢献しているのは当然であろう。特に NIRF のようなネットワーク中心性を持つ遺伝子が扱う情報の範囲は，ゲノム全体が処理する宇宙規模の情報の一部であり，「小宇宙」という規模表現が適切かと思われる。

こうした莫大かつ予見不可能な遺伝子機能を計算で求めるためには，解析技術に最大の一般性が要求されることは明らかだろう。故に，私たちの情報生物学においては，如何に精巧な構造機能を持つ遺伝子であれ，如何に巨大かつ緻密な情報交換であれ，一意に表現できる数理的手法を開発することが必要となる。換言すれば，私たちが開発すべき解析技術は大きな処理能力と広範な普遍性を持たねばならず，限られた一部の遺伝子だけに適用されるものであってはならない。敢えて言おう。遺伝子のみならず，森羅万象の森羅万象に対する関わりを明らかにしたい。その為には，宇宙に有る全ての「存在」を抽象的に表現し，それらの相互関係を統一的に記述できることが必要である。

それでは私たちが表現すべき「存在」とは何だろうか。それは「確率」であると考える。現代物理学においては，すべての現象は確率的に記述され，あらゆる物体は存在確率で表現される。さらに確率論・統計学の教えるところによれば，存在同士の「関係」もまた確率的に表現される。これら既存の物理学と確率統計学に対して，私たちの選択は「情報」を用いることであった。以下に示すように情報は確率と等価であるから，すべての存在も，それらの相互関係も，等しく情報で表現することが可能となるだろう。実際，開発に至った方法論は，新規ながら一般性があり，遺伝子のみならず人間も含めた，情報を持つすべての因子同士の相互作用を記述可能である。さらに「情報と確率の等価性原理」は単純な数式でありながら，確率統計学と情報理論という，従来ほぼ独立に発展してきた二つの数理体系を結びつける。情報と確率の普遍性を考えれば，「等価性原理」は宇宙の森羅万象の間で交わされる相互関係を一意に表現するための，方法論的基礎を与えるものと期待される。

## 遺伝子情報の数理的特性

さて，遺伝子機能を特定する目的で遺伝子の発現する「情報」に注目するにせよ，その方法論を開発する手掛かりや，方法論を適用すべき対象は何処に求めれば良いのだろうか。そもそも遺伝子が発現する情報は，無限に近い高次元に及ぶ複雑極まりない代物である。こうした超高次元の巨大情報に対して，正面から情報理論を適用した例は過去に例がない。そこで私は，遺伝子の発現する情報の進化過程を，情報理論と集団遺伝学を組み合わせて追跡することにより，遺伝情報のエッセンスを描出することに挑戦した[4, 5]。得られた結論を要約すると，個々の遺伝子が単独で発揮する情報量よりも，遺伝子間相互作用で産み出される多次元情報量の方が，圧倒的に大きいことが判明した。すなわち，

$$H(N_i) \gg H(M_i) \quad (1)$$

これは，遺伝子 $G_i$ がコードするネットワーク構造に基づくネットワーク情報量 $H(N_i)$ は，タンパク質など遺伝子産物の分子構造として発現する分子情報量 $H(M_i)$ を遙かに凌駕することを示している。

一見単純なこの数式(1)が，ワトソン＝クリックによるDNA二重らせん発見以降の生物学の常識を覆す，パラダイムシフトを意味することにお気付きだろうか。従来，遺伝子はヌクレオチド配列によりタンパクの分子構造をコードし，それによって生理機能を発揮し，恒常性維持に関わるとされてきた。この旧来のコンセプトが誤りなのではない。だが，大いに不十分なのである。上式により，遺伝子がコードするのは，目に見える分子構造だけではなく，目に見えないネットワーク構造も含まれることになる。しかも後者の方が圧倒的に多量の情報をコードするのだ。

それでは何故，これほど重要な事柄が今日まで知られて来なかったのだろうか？これはおそらく単純な理由に因ると思われる。分子構造は物質的なものであり，目に見える形で捉えやすい。しかも分子情報の遺伝単位はACGTのヌクレオチドと確定している。反対に，遺伝子間相互作用は数理的なものであり，容易に目に見えないだけでなく，単位となる遺伝情報が特定されて来なかった。私たちの「情報と確率の等価性原理」は，従来知られなかった遺伝情報単位を特定するものである。

生物学における重要性に加えて，式(1)は，従来の分子生物学が手段としてきた分子機能の解析では，莫大な遺伝子情報のごく一部しか解明することができないことを示唆している。反対に，数理解析によりネットワーク構造を解析すれば，実験を行わずして遺伝子機能の大部分を求めることが出来るだろう。さらに言えば，もし私たちが開発した数理解析法によって遺伝子機能を正しく特定することに成功するならば，ネットワーク情報の優越性ならびに遺伝情報の進化過程等々，私たちが考察したプロセスの妥当性を示す証拠にもなる。その際に得られる知見は，様々な角度から生物学に質的転換をもたらすことに繋がると期待される。

さらに，式(1)は生物学の質的前進を意味するに留まらない。ヒトゲノム計画が終了して20年経過したが，いまだ原因不明で治療法の無い疾患が数多く残されている。その主な理由は，染色体上の遺伝子の多くで，機能が依然不明であることに由来している。ところが従来の“wet”な実験手法では，個々の遺伝子機能の決定までに10年以上の歳月と，大きな人的・設備投資ならびに巨額の資金を必要とする。これに対して，前式に基づく数理解析法が実現できれば，実験操作の不要な機能解析法を開発でき，必要とされる負担を大きく軽減できるだろう。遺伝子機能解析は分単位のごく短時間の計算で完了するため，時間とコストを大幅に逓減できるからである。そればかりか，原理的に遺伝子産物の分子構造を問わないため，タンパク質をコードするprotein-coding geneも，タンパク質をコードしないnon-coding RNA geneも，同一の計算法で解析が可能である。さらに，抗がん剤開発において重要な免疫チェックポイント遺伝子の同定など，著しく実験困難な遺伝子についても機能計算を可能とする[5]。以上に挙げた事柄は，いずれも数理的手段による遺伝子機能解析法を開発することの重要性を示している。こうして開発された技術を，私たちは“*ab initio* 遺伝子軌道法”と命名した[5]。

## Cancer Informaticsの基礎

前段の考察を踏まえ，私は膨大な遺伝子間相互作用の総体を，計算機上で再現する作業に取り掛かった。この目的で私は公共データベースを利用することとした。TCGA（The Cancer Genome Atlas）というアメリカNIHの癌データベースである[6]。個々の実験施設で実行可能な実験の量とサンプル数には初めから限界があり，遺伝子間相互作用を包括的に解析するためのデータを得るには規模が小さすぎる。それに対して，米国が国家予算を投入して構築した大規模データベースが公開する未解析データは膨大であり，重大な情報が豊富に潜在している。だがこれまでは，量を質に変換する方法論を提供できるような，充分な基礎理論が存在しなかった。そこで私は以下に述べるように，正常組織と癌組織に関して，それらの相違点と共通点の２つの観点から基礎理論を立

ち上げるとともに，効率的なデータ利用を可能とする新規戦略を展開することとした。ご注意頂きたいのは，私が，通常の cancer informatics で行われているような，遺伝子と癌との関係を解明する目的のみで TCGA を利用した訳ではないことである。

癌はゲノム全体に及ぶ広範な異常を伴う疾患であり，多数の遺伝子に多彩な変化が生じる。それら多くの遺伝子変化が発癌と進展とに関わり，幾つかの共通する変化（増殖・生存・不死化・血管新生・浸潤転移など）が引き起こされる結果として[7, 8]，癌組織に特徴的な形態学的･病態生理学的な表現型が形成される。しかしながら，癌と正常細胞の間には，相違点ばかりが存在している訳ではない。過去に明らかにされてきた正常遺伝子の生理的機能のほとんどは，癌細胞を用いた実験で解明されてきたことを忘れてはならない。つまり，ほとんどの遺伝子について，正常細胞中と癌細胞中での性質に大きな共通点が存在することになる。だから TCGA が包含するデータの中には，癌を特徴付ける情報のみが含まれるのではなく，正常細胞中の正常遺伝子が発揮する機能に関する情報も多量に含まれるのである。

癌細胞でありながら正常遺伝子の機能が大きく反映される理由は，癌発生に進化的メカニズムが介在するからである。恒常性の維持された発癌前の段階においては，遺伝子変異が引き起こす適応度 fitness の変化は一般的には僅かであり，生物進化と同様に「ほぼ中立」な遺伝子変化が大部分を占める[9]。この「ほぼ中立」な遺伝子変化が優先的に生じる理由は，ドラマチックな変化が生じることよりも，むしろ正常機能が損なわれないことの方が，発癌前細胞の生き残りには大切だからである。なお発癌後には大規模な突然変異が次々に生じるようになり，大きな適応度を逐次獲得して激しい細胞増殖が引き起こされる。しかしこの段階においても，発癌前に生じた変化の多くは存続すると考えられている。

以上述べた理由から，癌遺伝子データベースが公開するデータの中には，正常組織と癌組織に関する相違点と共通点の双方に関する情報が存在している。そこで私は TCGA データベースを基にして，二つの informatics アルゴリズムを作成することとした。一つ目は，癌細胞を特徴付ける遺伝子を抽出するためのアルゴリズムで，STAIC（A Strategic Tool for Cancer Genome Analysis）と命名した。多くの既存データベース閲覧プログラムが癌遺伝子の変異プロファイル解析を目指しているのとは違い，STAIC は癌遺伝子それ自体の発見を効率化する目的で作成されたものである。実際，重要な癌関連遺伝子を同定することに成功しているが[5]，これはまた別の機会に述べたい。二つ目のアルゴリズムは，遺伝子の正常機能を算出する目的で作成された“*ab initio* 遺伝子軌道法”である。これは後述する数理解析により遺伝子間相互作用を算出し，遺伝子のネットワーク機能を特定するものである。

## 遺伝子情報のデータベース解析

私たちは情報と確率に焦点を当て，遺伝子間相互作用を解析する数理手段の開発に乗り出した。まず考慮すべき事柄は，①個々の遺伝子が発現する多次元情報量の評価と，②それら多次元情報因子同士が形成する，超高次元相互情報量の解析法を開拓することである。

遺伝子は莫大かつ多彩な情報を，緻密な時間的・空間的プログラムに従う様式で発現する。これを情報理論で扱いたいのだが，単純な一方向の遺伝情報ではないから，多次元情報理論の適用が必須となる。だが，Shannon によって創始された情報理論は，高次元の情報源を扱う方向には，まったく進歩を見せていない。当然ながら，遺伝情報のような無限次元に近い超高次元情報量と，それらの複合体である実質無限次元の相互情報量を扱う方法には，誰も手を付けて来なかった。そこで私は，コンピュータプログラムで用いられる線形インデックスを用いて，それらを無限次元確率変数および，無限次元相互情報量として扱う手段を開発した。詳細は論文[5]をご覧頂くことにして，ここでは概要のみを述べる。

TCGA データベースは30種類以上の癌種に渡って，総計１万例以上の癌組織から全遺伝子情報を収集し，mRNA, CNV, mutation, DNA methylation, histone methylation 等のデータに加えて，治療経過・組織像・臨床経過などを完全匿名化した上で公開している。このとき，公開された個々の遺伝子データは，充分な症例数が得られることを前提として，統計学的な扱いができるだろう。つまり，多次元中心極限定理により，各遺伝子が示す離散的あるいは連続的な状態データに関して，その実現確率の期待値を充分正確に算出することが可能となる。

私たちはデータベースから得られた，遺伝子の各状態データを線形インデックスで区別した上で，それら各状態の実現確率を，後述の informatics 戦略により評価した。実際用いたインデックス数は，各遺伝子につき329である。このとき２遺伝子間の相互作用計算に用いられる総インデックス数は，$329^2 = 108{,}241$となる。私たちは，全ての癌種の全患者に由来する全遺伝子のデータを用いて，全遺伝子が全遺伝との間で交換する情報量を定量する方針を採用した。実行した評価計算回数は，一遺伝子につき64,944,600,000回（約650億回）である。この大規模データ解析における定量的な情報評価法として導かれた informatics 戦略が，「情報と確率の等価性原理」である。

## 遺伝子間相互作用と情報理論

「情報と確率の等価性原理」は，TCGA データの評価方法を最適化する過程で見いだされた。遺伝子間相互作用を確率的な現象と捉えるとき，遺伝子間で行われている情報交換の大きさを正確に表現する方法は無いだろうか？様々な統計学的指標を試した結果，Fisher 正確確率 $P_F$ の対数を取った $\mathbf{-log}p_F$ が，遺伝子間相互作用の生物学的特徴を忠実に表現可能であることが見出された。驚いたのは，$-\log p_F$ を用いて IPA（Ingenuity Pathway Analysis）解析を実行すると，通常なら時間を掛けて行う，困難な実験でしか見出せないような複雑な遺伝子機能が，実験を行わずとも容易く計算できてしまうことであった。これは一体どうしたことだろう？私は長い時間を掛けて計算プログラムの改良を続け，数多くの遺伝子間相互作用について試行した結果から，$-\log p_F$ が何らかの重要な生物学的意味を持つに違いないと確信するに至った。だが，遺伝子情報の状態確率に関する研究は，まったく行われていなかった。長い時間を費やして数学の定理を探す生活がここから始まった。

それまで数学と一切無縁だった私が，食事中も就寝中も自動車運転中も，新規の数式を求め続けることになったのである。そして２年ほど経ったある夏の朝，$-\log p_F$ が未知の情報量を表現することに気がついた。一般に確率 $p$ の対数が情報を意味することを，ふと思い出したからである。しかし $p_F$ が意味するのは１事象の生起確率ではなく，２因子間の関連性を示す有意確率である。だから $-\log p_F$ が表現するのは通常の情報ではなく，遺伝子間の相互作用の大きさを規定する情報に違いない。結論として，後述の通り，$-\log p_F$ は遺伝子間で共有される**相互情報量 mutual information（MI）**と等価であることが見出された。ここで MI の意味合いを整理するため，Shannon が創始した情報理論についておさらいをさせて頂きたい。

宇宙に存在するすべての因子は情報を持っており，私は情報因子 informaton と呼ぶことにした。情報とは何であろうか。自己情報量の定義は，$I = -\log p$ であり，これは情報が確率 $p$ の別表現であることを意味している。この $I$ は事象の起こりにくさ，すなわち希少価値を示す指標である。さらに情報因子の持つ自己情報量 $I$ の総量を考え，$H = -\Sigma p\log p$ と定義したものが情報エントロピーであり，これは情報量とも呼ばれる。一方，情報因子同士で共有される情報の大きさを表現するのが相互情報量 MI であり，２つの確率変数 $X$，$Y$ の間で共有される $MI$ は，

$$MI(X;Y) = H(X)+H(Y)-H(X,Y) \quad (2)$$

で記述される。ここで $H(X, Y)$ は結合エントロピーと呼ばれ，２つの確率変数 $X$，$Y$ の取る状態の組み合わせで形成されるエントロピーを表す。一方，$MI(X ; Y)$ は $X$，$Y$ 間に共有された情報の大きさを示している。ベン図（Fig.1）において，$MI$ は交わり部分の面積に対応するエントロピーである。

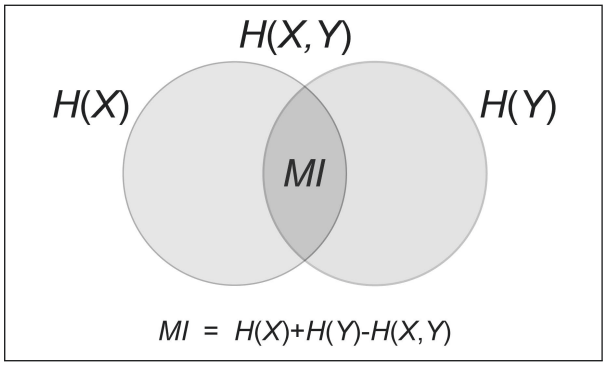


**Fig.1　相互情報量を示すベン図。左側と右側の円は，それぞれ確率変数 $X$ と $Y$ のエントロピー $H(X)$ と $H(Y)$ をあらわす。左右両方を合わせた全体は結合エントロピー $H(X, Y)$，左右の交差部は相互情報量 $MI$ をあらわす。**

情報理論の最大の要点は，情報の「意味」を度外視していることである。「情報理論」という名前から，情報の意味を問う学問と思われがちだが，真逆である。情報の持つ意味を捨象し，生起確率のみに注目して確立されたのが情報理論だからである。確率だけに基づいて学問体系を築きあげたことが情報理論の一般性の源泉であり，文字通り情報を持つ森羅万象に適用される理論体系を構築する足掛かりとなった。だから遺伝子情報に関して情報理論を適用する場合，遺伝子が発現する各状態の生物学的意味ではなく，各々の状態の実現確率（状態確率）のみを扱うことになる。意味を扱わないので，リン酸化・タンパク分解・転写制御・エピゲノム制御・複合体形成…その他諸々の機能も，あらゆる遺伝子情報が等価に評価されることになる。同様に，遺伝子間相互作用も確率のみで評価され，如何なる形式で情報伝達がなされるかは一切考慮されない。さらに言えば，情報因子が遺伝子であるか否かも情報理論は問題としないし，交換される情報が生物学的なものであれ，電子工学的なものであれ，物理学的なものであれ，すべて等しい表現が可能である。この高度な抽象性こそが情報理論の汎用性なのであるが，現象を扱う生物学者にとっては，親しみにくさを感じる部分かも知れない。

ともあれ，こうした情報理論の持つ抽象性を踏まえた上で，TCGA データベース解析が明らかにする事柄を考察しよう。遺伝子間の相互依存性を示す有意確率が相互情報量 MI と等価である事実，それらが生物学的な機能を記述可能であるという事実は，遺伝子同士が**情報と確率**という**共通言語**で会話していることを意味している。この会話の成立に際して，どの様な情報交換手段を用いたかは全く問題外であることに留意されたい。しかし私たちは，本来は意味を持たない「情報」に注目して計算を実行することで，逆説的ながら，遺伝情報が発揮する生物学的機能という「意味」を解釈することに成功した[5)]。すなわち，通常の実験に基づく分子機能解析では捉えられない遺伝子機能が，次に述べる数理解析で抽出可能となるのである。

## 情報と確率の等価性原理

情報理論と確率統計学は，ともに高度発達した数理体系であり，それぞれ幅広い領域で応用されている。両体系の完成ぶりを見れば，それらに新たな進歩が生じる余地など全く無さそうに思える。だが以下に示す通り，双方の学問体系を結ぶべき領域に空白が存在していた。

それぞれの数理体系において，２つの確率変数の相互依存性を示す指標に関して，

① 確率統計学での指標である有意確率（$p$ 値）は，サンプルサイズに依存して大きく変動するため，定量的ではなかった。

② 情報理論での指標である相互情報量 MI については，確率統計学的な性質が不明だった。

確率変数間の相互依存性は，両学問体系における中心課題である。同時に，この相互依存性は元来同じものを観察するのだから，有意確率 $p$ 値と相互情報量 MI の間には何らかの関連性があって然るべきだろう。それにも関わらず，これらの指標はまったく別個に扱われており，それらの相互関係は知られていない。相互依存性を示す２つの重要な指標同士の関係が不明確であることは，情報理論と確率統計学の両体系にまたがって存在する，本質的な未解決課題であると言えよう。この観点から見れば，上記①と②が意味することは，$p$ 値と MI が，ともに絶対的・統一的な指標とは言えないことである。

私たちは遺伝子間の相互作用解析を進める過程で，この未解決問題を発見するとともに，問題を解決する手段を見出した。以下にその導出過程を概説する。

### １）確率論における等価性原理

はじめに確率論を応用して，２つの確率変数の間に情報交換の生じる確率を求めよう。

① 最大エントロピー原理：２つの離散型確率変数 $X$, $Y$ に最大エントロピー原理を適用する。この原理は，確率変数が従う確率分布が不明な場合に，「不確かさ」の指標であるエントロピーが最大になるような分布を想定するのが適切と考えるものである。これにより $X$, $Y$ がともに一様分布すると仮定する。

② 統計学における等確率の原理：確率変数 $X$ が一様分布に従うとき，$X$ が取り得る各状態はどれも等しい確率で生起する。これは物理学の１分野である統計力学における「等確率の原理」に対応するもので，物理学と確率統計学の近縁さを示している。一様分布する $Y$ についても同様である。このとき $X$ と $Y$ の取り得る状態数（＝場合の数）をそれぞれ $W_X$，$W_Y$ とすると，それらの情報エントロピー $H(X)$，$H(Y)$ は，自然対数を用いて，

$$H(X) = \log W_X, \quad H(Y) = \log W_Y \tag{3}$$

で表現される。これらは統計力学において，熱力学的エントロピー $S$ を表すボルツマンの原理 $S = k_B \log W$ に対応する（$k_B$ はボルツマン定数，$W$ は微視的状態数）。

③ さらに２つの確率変数 $X$, $Y$ を合わせた同時確率変数（$X$, $Y$）にも最大エントロピー原理を適用し，一様分布を仮定する。このとき結合エントロピー $H(X, Y)$ は，同時確率変数の状態数 $W_{XY}$ を用いて，

$$H(X,Y) = \log W_{XY} \tag{4}$$

と表現できる。

④ 確率変数 $X$ と $Y$ は一様分布であるから，生起確率 $p_X$ と $p_Y$ はそれぞれの状態数に比例し，

$$\begin{aligned} p_X &\propto W_X = \exp[H(X)], \\ p_Y &\propto W_Y = \exp[H(Y)] \end{aligned} \tag{5}$$

が成立する。同様に，同時確率変数（$X$, $Y$）の生起確率 $p_{XY}$ については，

$$p_{XY} \propto W_{XY} = \exp[H(X,Y)] \tag{6}$$

が成り立つ。

⑤ 確率論における情報と確率の等価性原理：$X$ と $Y$ が独立のときの状態確率を１とするとき，２者間で共有される相互情報量の大きさが $MI$ となる確率 $p_{MI}$ は，

$$\begin{aligned} p_{MI} &= \frac{W_{XY}}{W_X \times W_Y} = \frac{\exp[H(X,Y)]}{\exp[H(X)+H(Y)]} \\ &= \frac{\exp[H(X)+H(Y)-MI]}{\exp[H(X)+H(Y)]} = e^{-MI} \end{aligned} \tag{7}$$

となる。したがって，

$$p_{MI} = e^{-MI}, \quad MI = -\log p_{MI} \tag{8}$$

が成り立つ。この関係式は相互情報量 $MI$ が，情報が交換される確率 $p_{MI}$ の別表現であることを示しており，これに私たちは**「情報と確率の等価性原理」**と名付けた。また式(8)は，先述した自己情報量 I の実現確率 $p_I$ について成立する

$$p_I = e^{-I}, \quad I = -\log p_I \tag{9}$$

と同型である。つまり情報量とその実現確率が，互いの指数関数と対数関数で表現されることは，単一確率変数の自己情報量でも，２つの確率変数の間で共有される相互情報量でも，等しく成立することになる。

これらに関連して，統計力学で扱う正準集団における熱力学的エントロピー $S$ も指数分布することが知られている。これは物理学における重要な原理であり，カノニカル（正準）分布と呼ばれている。情報理論と統計力学の近縁性は以前から指摘されているところであるが，また一つ，新たな関連性を示唆する式が得られた訳である。私たちは(8)および(9)に表現される情報量と実現確率の関係の一般性に鑑み，これら情報量の指数関数に基づいた確率分布形式を**情報カノニカル分布**（infocanonical distribution）と命名した。

⑥　情報理論に基づく有意確率：式(8)の $p_{MI}$ は，離散型確率変数 $X$ と $Y$ の間で共有される相互情報量が $MI$ となる確率質量関数を表している。しかし $X$ と $Y$ の標本空間が充分に大きい場合，$p_{MI} = e^{-MI}$ を連続化・正規化して確率密度関数

$$f_{MI} = e^{-MI} \tag{10}$$

を得ることができる。このときエントロピー $H(X), H(Y)$ が無限大に近づく条件下で $MI \to \infty$ の極限を考えることにより，$X$ と $Y$ の有意確率は，

$$p\text{-value} = \int_{MI}^{\infty} e^{-MI} dMI = e^{-MI} \tag{11}$$

となる。つまり $p_{MI} = e^{-MI}$ は大きさ $MI$ の相互情報量が実現する確率だけでなく，有意確率，すなわち，$MI$ 以上の大きさの相互情報量が偶然発生する確率をも表現する。有意水準を $p$-value ＝0.05とすれば，そのとき $MI$ ＝2.9957≒3.00である。したがって $MI$ ≥3.00ならば情報交換が偶然ではなく，$X$，$Y$ 間に有意な相互依存性があると言える。

以上の通り，**確率論における「情報と確率の等価性原理」**は，２確率変数間の相互情報量の大きさ $MI$ とその実現確率 $p_{MI}$，およびその有意確率 $p$-value について，これら３者の等価性を示すものである。結論として，確率論の領域においては，確率統計学の指標である有意確率と，情報理論の指標である相互情報量とを統一することに成功した。

**２）統計学における等価性原理**

次に，確率論での議論にひき続いて，等価性原理が統計学でどの様に観察されるかを考察しよう。

①　情報交換の１回発生確率：確率変数 $X$，$Y$ の相互依存性の程度を見積もるために，２者間で生じる情報交換を繰り返し観測する試行を考える。具体的には，$X$ と $Y$ の状態を繰り返して観察し，観察終了後に集計する行為が該当する。ここで，情報交換が発生する場合には，１回あたり大きさ $MI$ の相互情報量が生じるものと仮定すると，情報交換の生じる確率は１回あたり $p_{MI} = e^{-MI}$ である。

②　情報交換の多数回発生確率：試行を $N$ 回繰り返すとき，$X$，$Y$ 間で交換される相互情報量の総量の期待値は $N \cdot MI$ であり，その実現確率 $p_{N \cdot MI}$ との関係は，

$$p_{N \cdot MI} = e^{-N \cdot MI}, \quad MI = -\frac{1}{N} \log p_{N \cdot MI} \tag{12}$$

である。式(12)は $N$ の大きさに関わらず成立する。

③　情報理論に基づく有意確率：式(12)の $p_{N \cdot MI}$ は，離散型確率変数 $X$ と $Y$ の間で共有される相互情報量の総量が $N \cdot MI$ となる確率質量関数を表している。しかし $X$ と $Y$ の標本空間が充分に大きい場合は，$p_{N \cdot MI} = e^{-N \cdot MI}$ を連続化・正規化して確率密度関数

$$f_{N \cdot MI} = N e^{-N \cdot MI} \tag{13}$$

を得ることができる。このときエントロピー $H(X), H(Y)$ が無限大に近づく条件下で $MI \to \infty$ の極限を考えることにより，$X$ と $Y$ の有意確率は，

$$p\text{-value} = \int_{MI}^{\infty} N e^{-N \cdot MI} dMI = e^{-N \cdot MI} \tag{14}$$

となる。つまり $p_{N \cdot MI} = e^{-N \cdot MI}$ は，総量 $N \cdot MI$ の相互情報量が実現する確率だけでなく，有意確率，すなわち，$MI$ 以上の大きさの相互情報量が偶然発生する確率をも表現する。有意水準を $p$-value ＝0.05とすれば，そのとき $MI$ ＝2.9957／$N$ ≒3.00／$N$ である。したがって $MI$ ≥3.00／$N$ ならば情報交換が偶然ではなく，$X$，$Y$ 間に有意な相互依存性があると言える。

| X / Y | 1 | 2 | · | · | *i* | · | · | *m*-1 | *m* | |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | | | | | | | | | $b_1$ |
| 2 | | | | | | | | | | $b_2$ |
| · | | | | | | | | | | · |
| · | | | | | | | | | | · |
| *j* | | | | | $x_{ij}$ | | | | | $b_j$ |
| · | | | | | | | | | | · |
| · | | | | | | | | | | · |
| *n*-1 | | | | | | | | | | $b_{n-1}$ |
| *n* | | | | | | | | | | $b_n$ |
| | $a_1$ | $a_2$ | · | · | $a_i$ | · | · | $a_{m-1}$ | $a_m$ | *N* |

**Fig.2　*m*×*n* 分割表。**確率変数 *X*，*Y* がそれぞれ $X_1$–$X_m$，$Y_1$–$Y_n$ の値を取るときの分割表を示す。$X_{ij}$ は同時度数，$a_1$–$a_m$ および $b_1$–$b_n$ は周辺度数，*N* はサンプルサイズ。

④　従来統計学による分析：次に，上記試行の結果を *m*×*n* 分割表（Fig.2）に割り当てて，従来の統計学による分析を行う。ある分割表の各セルの同時度数を $x_{ij}$，周辺度数を $a_i$，$b_j$ とするとき，その分割表に対応した結果の得られる確率 $p_C$ は，

$$p_C = \frac{\prod_{i=1}^{m} a_i! \prod_{j=1}^{n} b_j!}{N! \prod_{i,j} x_{i,j}!} \tag{15}$$

で与えられる。超幾何分布確率 $p_H$ は，周辺度数が固定された分割表において，観測された結果が得られる確率 $p_C$ である。なお，$p_H$ は確率質量関数である。一方，Fisher 正確確率 $p_F$ は，周辺度数が固定された分割表において，観測された結果，もしくはそれよりも起こりにくい結果が得られる確率 $p_C$ の総和であり，

$$p_F = \sum_{p_C \le p_H} p_C \tag{16}$$

で表現される。この $p_F$ は有意確率を表している。

⑤　統計学における情報と確率の等価性原理：サンプルサイズ *N* が限りなく大きい場合，相互情報量 *MI* と超幾何分布確率 $p_H$，Fisher 正確確率 $p_F$ に関し，以下の関係式が成立する。

$$MI = -\frac{1}{N}\log p_H = -\frac{1}{N}\log p_F \tag{17}$$

さらに，式(17)が式(12)の右式と漸近的に等しいことは，解

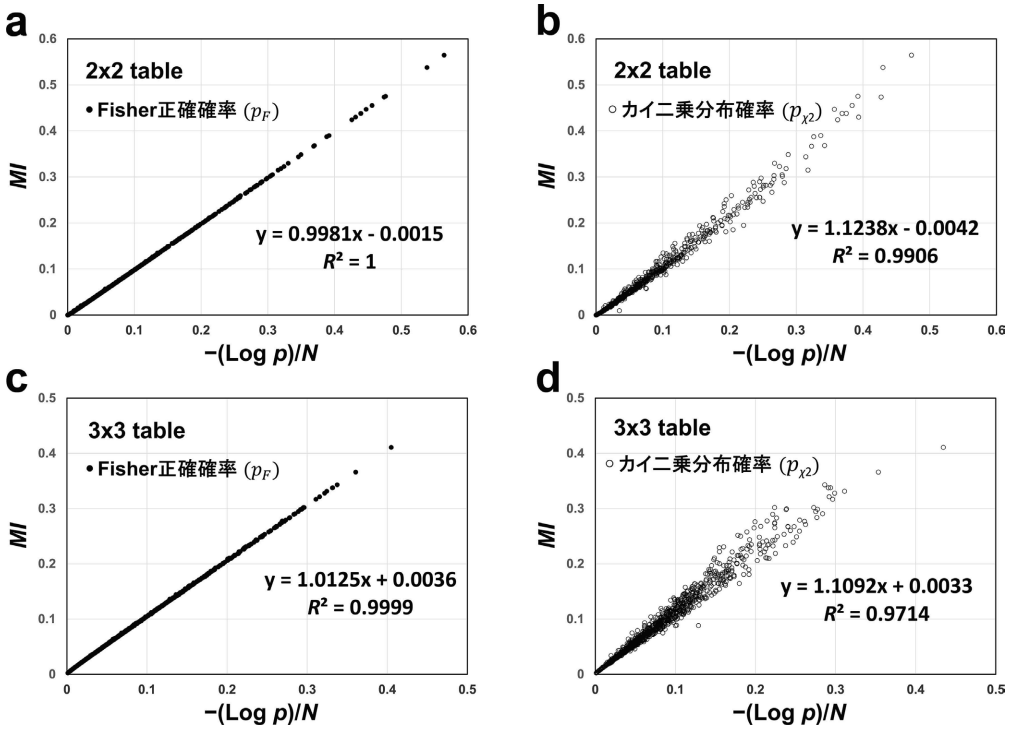


**Fig.3　分割表における相互情報量 *MI* と有意確率の関係。**乱数で作成した2x2分割表（**a**，**b**）と3x3分割表（**c**，**d**）における *MI* に対し，フィッシャー正確確率（●）とカイ二乗分布確率（○）の対数 ÷ サンプルサイズ *N*（=1,000）をプロットした。

析的に証明できる[5)]。この結果，十分に大きなサンプルサイズが確保される条件下では，遺伝子間の Fisher 正確確率 $p_F$ から導かれた -log$p_F$ は，遺伝子間の相互情報量 *MI* に比例することが示された（Fig.3a，3c）。

以上において，式⑿と⒁は，相互情報量の大きさ $N \cdot MI$ と，情報交換の確率質量関数 $p_{N \cdot MI}$，および有意確率 *p*-value の等価性を示している。さらに式（17）は，$N \cdot MI$ と統計学における確率質量関数 $p_H$，および有意確率 $p_F$ の，漸近的等価性を示している。そこでこれらの等価性を総称して，**統計学における「情報と確率の等価性原理」**と呼称する。結論として，統計学の領域においても，確率統計学の指標である有意確率と，情報理論の指標である相互情報量を統一することに成功したことになる。

## 等価性原理による遺伝子解析

情報理論と確率統計学は互いに近接した学問領域でありながら，それぞれが独自に進歩してきた。Fisher が正確確率検定を公表したのが1922年[10)]，Shannon が情報理論を発表したのが1948年[11)] であるが，それ以後，２つの数理体系の統一が達成されたことはない。これに対して，私たちの「情報と確率の等価性原理」は，70年ぶりに両学問体系の統合を導く新定理である。この定理を適用することによって，莫大な高次元データから成る多数のデータセットで構成された TCGA データベースの数理解析が可能となり，相互情報量 MI に基づく遺伝子間相互作用を正確に定量できるようになった。以下に，この原理がもたらす効能を列挙する。

### １）確率統計学での有意確率（*p* 値）の情報理論的解釈を可能とすること

これまでは，確率変数同士の相互依存性をあらわす *p* 値について，情報理論的な意味付けが為されたことはない。しかし得られた数式は，分割表で求めた有意確率である Fisher 正確確率 $p_F$ の意味するものが，確率変数間に共有される相互情報量 MI と漸近的に等価であることを示した。サンプルサイズ $N$ が等しい場合，より小さな $p_F$ は，より大きな情報交換の存在を意味する。先述の通り，informaton である遺伝子同士の相互作用は確率論的な現象であり，等価性原理によってその大きさが定量可能となった。

### ２）相互情報量 MI のサンプルサイズ安定性により，正確なデータ比較が可能となること

Fisher 正確確率 $p_F$ の欠点として，サンプルサイズ $N$ の変動に対して不安定であることが挙げられる。そのため，$N$ の異なるデータセットの間で，$p_F$ を正確に比較検討することは難しい。それに対して，相互情報量 MI は $N$ の変化に対して安定であるため，データセット間での比較検討が容易である。その結果，TCGA のように多種類の癌データセットから得られた計算結果同士を，定量的に比較することが可能となる。

### ３）相互情報量 MI の統計学的有意性が評価可能となること

従来は相互情報量 MI の確率統計学的な性質は知られて来なかった。これに対して，等価性原理を応用することにより，得られた *MI* に対して伝統的な確率統計学的手法を適用することが可能となった。そこから得られるメリットのひとつは，*MI* の有意確率が把握できることである。

情報理論の指標である相互情報量 MI の応用は，電気通信から画像処理まで広い範囲に渡る。だがこれまでは，得られた *MI* の信頼性を把握する手段が欠けていた。もともと MI には「サンプルサイズ数 $N$ に関わらず安定」という長所があるのだが，その一方で，算出された *MI* の価値を評価するには $N$ の重みを正当に考慮しなければならないはずである。しかし Shannon 情報理論には *MI* の有意性を評価する手段が無く，これが医学生物学領域で情報理論の適用が限られてきた理由のひとつであろう。

これに反して，私たちは等価性原理 $p_{N \cdot MI} = e^{-N \cdot MI}$ を適用することで，*MI* とその有意確率を同時に知ることができるようになった。その結果，TCGA のように極めて多くのデータから算出された *MI* の持つ統計学的信頼性を，明確に示すことが可能となる。

### ４）相互情報量 MI を正確に求める目的でメタ解析が利用可能となること

メタ解析は，複数のデータセットで得られた分析結果を統合することで，より精度の高い解析を行う統計学的方法である。これまでの情報理論においては，複数の相互情報量 MI を統合することは不可能だった。しかし上記３）に記した通り，MI を確率に変換することによって，統計学的なメタ解析が実行可能となる。その結果，極めて高い有意水準で *MI* を特定できるようになった。実際，*ab initio* 計算で得られた遺伝子間 *MI* の *p* 値は，10の－100乗から10の－300乗にも達し，巨大な有意性を示すものである。

### ５）多次元確率変数同士の有意確率が算出可能になること

Fisher 正確確率 $p_F$ は，$m \times n$ 分割表のサイズが大きくなったり，サンプルサイズ $N$ が大きくなったりすると，

実質的に計算不能となる。ところが遺伝子間相互作用の解析では，かなり大きな分割表を用いて，多次元確率変数同士の相互依存性を計算する必要があり，この目的で $p_F$ を用いることは甚だ困難と言える。これと反対に相互情報量 MI であれば，原理的には多次元相互情報量の算出が可能である。さらに私たちは，超高次元相互情報量であっても効率的に計算できる方法を開発した[5)]。その結果，サンプルサイズ $N$ が大きい場合で，しかも高次元分割表の場合であっても，有意確率が算出できるようになった。

これに加えて，等価性原理で得られる有意確率が，従来標準的に用いられてきたカイ二乗分布で得られる有意確率をかなり上廻る正確さを示す（Fig.3）ことも，大きな長所の一つとして挙げておきたい。

**6）情報距離と機能的近接度が算出可能になること**

情報理論を用いると，確率変数間の情報メトリック $r$（information metric）を以下のように表現できる。

$$r = H(X)+H(Y)-2MI(X;Y) \quad (18)$$

この情報メトリック $r$ は数学の距離の公理を厳密に満たし，informaton 間の情報距離を表現すると言って良い。私たちは今回開発された数理技術によって MI を定量し，情報距離 $r$ を計測する手がかりを得た。一般的には，遺伝子間の MI が大きいことは遺伝子間の情報距離 $r$ が近いことを意味する。そのため，MI の大きさを *ab initio* 計算で定量し，間接的に $r$ を比較することにより，遺伝子同士の機能的近接度を導くことが可能である[5)]。

**7）等価性原理から *ab initio* 遺伝子軌道法へ**

以上１）－６）を統合することで，遺伝子間相互作用の正確な導出が可能となった。遺伝子間の Fisher 正確確率 $p_F$ から得られた $-\log p_F$ は，遺伝子間相互情報量 MI の別表現である。この MI が遺伝子の生物学的機能を反映するのであるから，遺伝子同士が MI という情報単位で結ばれたネットワークとして機能することに他ならない。すなわち，分子情報量 $H(M_i)$ を規定する情報単位がヌクレオチドであるのに対して，ネットワーク情報量 $H(N_i)$ の基本単位は相互情報量 MI であると結論できる。

この際，莫大なサンプル数を持つデータベースを用いた数理解析を行うことで，遺伝子間 MI の集積としてのネットワーク情報量が正確に算定できることになる。すなわち，遺伝子 $G_i$，$G_j$ 間の２遺伝子ネットワーク（two-gene network）の情報量を $H(N_{ij})$ とすると，遺伝子 $G_i$ がコードするネットワーク情報量は，

$$H(N_i) = \sum_{j \neq i} H(N_{ij}) \quad (19)$$

と表現できる。このネットワーク情報量の算出に基づいて，実験に依存せず，高速に遺伝子機能を特定する方法が“*ab initio* 遺伝子軌道法”である[5)]。

現実に超高次元の遺伝子間相互情報量の算出を行うためには，等価性原理の他にも多くの数理学的技術進歩と高度な informatics プログラムの開発を必要とし，私たちが特許（**特許第6820621号**）を取得するまでには数年を要した。論文[5)]に記載したとおり，従来の実験技術では容易に発見できなかった重要な疾患関連遺伝子を見出すことに成功している。こうした実際の解析例を含めた報告は，また機会を改めて行うこととしたい。

## 等価性原理の展開

本稿に述べたように，私は遺伝子間相互作用の検討を出発点として，情報理論と確率統計学を結びつける定理を発見するに至った。遺伝子機能解析に役立てる実用性は達成できたが，さらなる特性も明らかになりつつあり，以下に列挙したい。

**1）情報理論の優越性**

良く知られる通り，Fisher 正確確率検定を初めとする従来の統計学においては，分割表の周辺度数を固定して生起確率を計算するのが常套手段であった。だがこれとは逆に，本稿で述べた等価性原理は，確率変数が一様分布すると仮定する「等確率の原理」に立脚している。この手法は「最大エントロピー原理」に基づいており，ベイズ理論を応用したものである。この意味では，「情報と確率の等価性原理」は，伝統的な確率統計学とは相反する出発地点を持つと言える。ところが興味深いことに，伝統的統計学の所産である超幾何分布確率 $p_H$ も Fisher 正確確率 $p_F$ も，サンプルサイズ $N \rightarrow \infty$ の極限において，それらの対数の符号を変えたものが相互情報量 MI と等しくなる。すなわち，サンプルサイズを増やすことは，確率変数の一様分布を仮定した場合と同等の結末となる。

ではこの知見は，「情報と確率の等価性原理は，既存統計学の近似である」ことを意味するのであろうか？答えは「No」である。何故なら，$N$ の増加に伴って $-\log p_H$ や $-\log p_F$ が $N \cdot MI$ に近づくのであって，$p_{N \cdot MI} = e^{-N \cdot MI}$ が $p_H$ や $p_F$ に近づくのではないからである。この２つは同値ではない（簡単な計算で示されることであるから，興味がおありの方は是非試して頂きたい）。つまり「既存統計学が，等価性原理の近似」なのであり，「情報理

論が統計学よりも正確である」が正しい解答と言える。この情報理論の優越性を意味する所見が，等価性原理のもたらす大きなパラダイム変革のひとつであろう。

医学生物学・看護学領域では，より小さな $p$ 値，すなわち明確な統計的有意性を得る目的で症例数を多数確保する試みは，ごくありふれた研究方針である。しかしながら，小さな $p$ 値を得ることの真の意味は，精度の高い相互情報量 MI の確定値を得ることにあると言えよう。

### 2）統計力学への応用

統計力学は物理学の主領域の一つである。私たちが用いた「等確率の原理」は，統計力学の基本原理として知られるものであり，当然のことながら，今回得られた知見を物理学に応用するのは必然的な流れであろう。事実，等価性原理 $p_{MI} = e^{-MI}$ を統計力学に適用することで，重大知見が得られる。それは，２つの物理学的 informaton（量子など）に共有される相互情報量 MI について，「大きな $MI$ ほど実現しにくい」ことである。言い換えれば，「MI は自発的に減少する傾向を持つ」ことを意味する。式(2)より，これは「合成系の結合エントロピー $H$（$X$，$Y$）は自発的に大きくなる傾向を持つ」ことを意味しており，エントロピー増大を意味する「熱力学の第二法則」に沿う結論といえる。

しかし現実には，物理系においても $MI > 0$ が普遍的に観察される。さらに生物系も含めて観察すれば，遺伝子間相互作用のように，$MI > 0$ となる事態があまねく発生していることが判る。こうした状況は，「MI の自発的減少」とは相反する「力」が存在することを示唆すると言える。私が公開した論文[5]は，この点に重点を置いて記述したものである。

### 3）人間関係論への応用

人間の存在様式を確率変数で表現する場合，人間同士の相互関係は相互情報量 MI で表現できる。ここで注意が必要なのは，遺伝子情報が人間をコードするのであるから，人間の複雑さが遺伝子の複雑さを上廻ることは無い，という真実である。こう断言できる理由は，昨年の研究会報告[1]に詳しく述べたので，ご参照頂きたい。

さて，私たちは等価性原理を含め，遺伝子機能を記述する一連の数理手段を開発したが，そのエッセンスは，遺伝子より単純な存在である人間とその相互関係を表現する際にも適用される，と考えるのは自然であろう。公開された論文[5]においては，人間の間に働く「統計情報力学[4]的な」力と，その相互関係の時間発展によってもたらされる人類の歴史に言及した。遺伝子間相互作用を知るものは，人間同士の相互関係をも善く識ることになるのである。

## 最後に

本年度に公開した論文[5]は，３ヶ月余りで400件のダウンロードを記録した。この紀要が出版される頃に1,000件を超す勢いであり，好評を博していることは望外の喜びである。本稿で取り上げた「情報と確率の等価性原理」は，この論文の40分の１程度の内容を概説したもので，確率統計学と情報理論の統合を指向した。しかし情報と確率とは，その普遍性にこそ最大の特質があるのだから，広く宇宙の森羅万象に対して情報理論を適用することが可能であろう。だから論文では物理学の統一を主題に据え，医学生物学と看護学まで統合する試みを述べた。

「情報と確率の等価性原理」は，合成系のエントロピー増大をもたらす意味で，自然な法則であるように見える。だが，現実には相互情報量 MI が増大し，合成系のエントロピーを減少させる場合も多いことを忘れるべきではない。そこには未発見の物理法則が存在するだろうからである。論文では，この議論を立証するために数多くの事例を挙げた。完成には莫大な労力と時間を要したが，新規の疾患遺伝子の特定，２つの特許（**特許第6820621号，特許第7114112号**）の成立を含め，数多くの新知見を世に問うことができた。共著の先生方には，この場をお借りして心から感謝を申し上げたい。

## 文　献

1）Mori, T., Li, Y., Hata, H., et al.: NIRF, a novel RING finger protein, is involved in cell-cycle regulation, Biochem. Biophys. Res Commun, 296（3）, 530-536, 2002. DOI: https://doi.org/ 10.1016/S0006-291X（02）00890-2

2）Mori, T., Ikeda, D. D., Fukushima, T., et al.: NIRF constitutes a nodal point in the cell cycle network and is a candidate tumor suppressor, Cell Cycle, 10（19）, 3284-3299, 2011. DOI: https://doi.org/ 10.4161/cc.10.19.17176

3）Mori, T., Ikeda, D. D., Yamaguchi, Y., et al.: NIRF/UHRF2 occupies a central position in the cell cycle network and allows coupling with the epigenetic landscape, FEBS Lett, 586（11）, 1570-1583, 2012. DOI: https://doi.org/10.1016/j.febslet.2012.04.038

4）森　努：NIRF 研究会報告「情報生物学への招待」，福島県立医科大学看護学部紀要，24，21-28，2022.

5）Mori, T., Kawamura, T., Ikeda, D.D., et al.: Influential Force: from Higgs to the Ab Initio Genetic Orbital Method, Jxiv, 2022. DOI: https://doi.org/10.51094/jxiv.156

6）Tomczak, K., Czerwińsa, P. and Wiznerowicz, M., Review The Cancer Genome Atlas（TCGA）: an immeasurable source of

knowledge. Contemporary Oncology/Współczesna Onkologia, 2015（1）, 68-77, 2015. DOI: https://doi.org/10.5114/wo.2014.47136

7）Hanahan, D. and Weinberg, R. A., Hallmarks of cancer: the next generation, Cell, 144, 646-674, 2011. DOI: https://doi.org/10.1016/j.cell.2011.02.013

8）Hanahan, D., Hallmarks of cancer: new dimensions, Cancer Discovery, 12（1）, 31-46, 2022. DOI: https://doi.org/10.1158/2159-8290.CD-21-1059

9）Cannataro, V. L. and Jeffrey P. T., Neutral theory and the somatic evolution of cancer, Molecular biology and evolution, 35（6）, 1308-1315, 2018. DOI: https://doi.org/10.1093/molbev/msy079

10）Fisher, R. A., On the interpretation of $\chi 2$ from contingency tables, and the calculation of P, Journal of the royal statistical society, 85（1）, 87-94, 1922. DOI: https://doi.org/10.2307/2340521

11）Shannon, C. E., A mathematical theory of communication. The Bell system technical journal, 27（3）, 379-423, 1948. DOI: https://doi.org/10.1002/j.1538-7305.1948.tb01338.x