様式 C-19

科学研究費補助金研究成果報告書

平成２２年　５月２０日現在

研究種目：特定領域研究
研究期間：2006～2009
課題番号：18079012
研究課題名（和文）　個別化医療の実現に向けた情報統計力学的理論構築及び手法開発

研究課題名（英文）　Development of theory and procedures for personalized medicine using statistical mechanisms of information processing

研究代表者
井上　真郷（INOUE Masato）
早稲田大学・理工学術院・准教授
研究者番号：70376953

研究成果の概要（和文）：
一人ひとりの遺伝情報を活用して，診断や予後，副作用の予測精度をより高める個別化医療に必須の haplotype 推定手法について研究した．これは一塩基多型(SNP)データより，被験者の遺伝情報を父・母由来に分離する情報処理技術である．結果，従来法よりも大規模なデータを高精度で推定できる斬新なアルゴリズムを開発し，フリーソフトウエアとして公開した．他にも，情報統計力学的手法を用いた高速な近似推定手法を開発した．

研究成果の概要（英文）：
We investigated the haplotype inference method, which is an indispensable information processing technique for personalized medicine providing high quality diagnoses or medical treatments based on individual genetic information. This method involves dividing individual genetic information into two groups, one for paternal origins and the other for maternal origins. Our main achievement is the development of a novel highest quality haplotype inference method. The developed software is available for free through our website. We also developed more powerful but less accurate inference methods by using statistical mechanical techniques.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２００６年度	2,000,000	0	2,000,000
２００７年度	2,600,000	0	2,600,000
２００８年度	2,600,000	0	2,600,000
２００９年度	2,500,000	0	2,500,000
年度			
総　計	9,700,000	0	9,700,000

研究分野：総合領域
科研費の分科・細目：感性情報学・ソフトコンピューティング・確率的情報処理
キーワード：情報統計力学，確率的情報処理，個別化医療，haplotype 推定，遺伝子-疾患関連解析

１．研究開始当初の背景

医学データは一般にモデル化が難しく，また被験者数が限定されることから，大標本に基づく統計的推論手法が応用されにくい分野が残っていた．このような情報処理技術が充実されれば，より精度の悪い安価なデータに基づいて，信頼できる診療予測を立てる事が可能となり，個別化医療(オーダーメイド医療)の情報処理的側面におけるボトルネックが改善できるのではないかと期待されていた．

個別化医療とは，ヒトにおける遺伝情報と臨床像との関連を発見し，この関連に基づいて，個人の遺伝情報からどのような診療結果となるかを高い精度で予測し，よりよい診療をしようというものである．

ヒトの遺伝情報の多様性は専ら一塩基多型(SNP)によると言われ，これはゲノムの約0.1%を占めるため，ヒトゲノムが 30 億塩基対，ヒトが二倍体であることから，一人につき約 600 万ビットの情報を持っている．これに対し，余程の大規模調査で無い限り，患者に由来する疾病データはせいぜい数百人規模でしか集まらない．このような歪なデータに対して，どのような推定手法が適切かつ信頼に足るのか，また，その推定精度は何がしかの条件を境に統計力学的な相転移現象が見られるのではないか，と思われた．

２．研究の目的

個別化医療の情報処理的側面で生ずるであろうボトルネックの解消に向けた基礎的な研究を目的とした．具体的には以下の通りである．

(1) 集団の一塩基多型(SNP)データから個々人の haplotype を推定する新たな手法を開発する．EM アルゴリズムを用いる標準的な推定手法は，扱える SNP 部位数が 25 程度と少なかった．この数字は扱う被験者の数や，欠損値の割合，計算機の性能に依存するが，指数として効くため，例え数人の小規模データをスパコンで解析したとしても，40 部位程度が限度となる．この問題は，様々な近似を用いた推定手法が提案されているが，上手く高次元データを扱うことで，近似を用いずにより多くの SNP 部位を扱えるような推定手法を開発することを目的とした．また，推定結果について，最適な haplotype が複数存在することがあり，強磁性相，スピングラス相の相境界等の関連を研究することを目的とした．また，研究途中で SNP データに加えて，コピー数多型(CNV)データについても haplotype 推定の重要性が高まってきたため，このデータも扱えるようになることを目的に加えた．

(2) 前項の研究を進める過程で，haplotype の生成モデルを構築することが重要という結論に達した．ヒトの遺伝子は，少なくとも 100 世代に亘って交配，変異，自然選択を繰り返しており，これを数学的に扱い易い式で簡潔に記述することは一般に容易ではない．ここではその端緒として，集団の一塩基多型(SNP)データから組換 hot spot(haplotype block の境界)を推定することを目的に加えた．具体的には，組換 hot spot とは，減数分裂時に父母由来の相同染色体が交叉，混じり合って新しい染色体を構成する際に，交叉の起き易い染色体上の場所のことを指す．組換 hot spot の存在割合は 10^{-4}〜10^{-5} 塩基対，交叉の起こる頻度は減数分裂 1 回当たり 10^{-4}〜10^{-3} 程度と言われるが，量的なものであり明確な基準は無い．この問題は古くから存在するが，未だに綺麗には解かれていない．良く用いられる指標として r^2 や D' などがあるが，染色体上のある場所一個所について複数の値が出てきてしまうため，その複数の値をどのように総合的に判断すればよいかが定まっておらず，主観的な判断に委ねられていた．このような問題を解決し，各部位について単一の指標値(交叉率)を SNP データから求められるような手法を開発することを目的とした．

(3) 疾病・治療効果等の臨床データとの関連を類型化して発見し，診療予測に有用な SNP 座位を特定する理論・手法を構築・開発することを目的とした．特に，300 万個所ある SNP の内のどの SNP が特定の疾患に関係するのか，ということを統計的に正しく判断できる遺伝子-疾患関連解析手法の開発を目的とした．

(4) 上記以外に，先駆的な確率的情報処理課題についても研究を行い，この分野での様々な解法，アプローチが上記課題に応用できないか，その可能性を探ることを目的とした．

３．研究の方法

(1) haplotype 推定に関しては，既存の EM アルゴリズムを用いた手法を近似を導入することなく，より大規模なデータに適用可能となるようなアルゴリズムを開発し，C#言語による実装を行った．この際，計算量についてはアルゴリズムのどの部分が重要なのか，どのような枝刈りが有効なのか，といったことについては理論計算のみでは予測しきれない面が多分にあり，様々なアイデアを実装しては計算時間を計測する，という試行錯誤を繰り返す必要があった．また，実データ，人工データを用いて検証を行い，更にその結果をアルゴリズムの改良へと繋げた．特に実データには約 1%の欠損値が含まれており，この存在が計算時間を大きく悪化させることが分かり，部分的に欠損値処理に特化した高効率のアルゴリズムを開発，追加した．

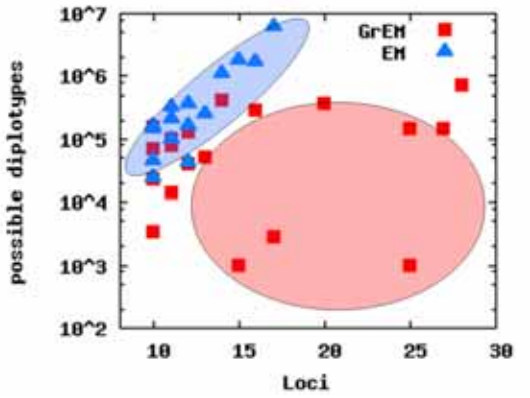
(2) 組換 hot spot 推定に関しては，既存手法の検証から始めた．この際のポイントは，統計的モデルが実際の遺伝とどの程度整合性がとれているか，また，モデルをデータに基づき選択する際の基準は妥当か，また，最適なモデルが現実的な計算量で求めることが可能か，という点である．次に，独自のモデルを開発した．具体的には，交流の乏しい遺伝集団が複数存在する，というモデルを構築した．次に MDL 基準によるモデル選択基準を導入して，最適モデルを求める手法を開発して検証した．この結果，最適解が安定して求まらないことが分かったため，EM アルゴリズムと遺伝的アルゴリズムの二つの最適化手法の組合せ，といったハイブリッド手法を開発して更に検証を行った．

また，上記手法とは別に，マクロの統計量に基づいて交叉率を推定する手法を開発し，検証を行った．この手法は，遺伝モデルを単純化して，ロバストな推定が行えるよう設計した．

(3) 臨床データと関連のある SNP 座位の発見については，正確な p 値を計算するための理論構築を試みた．従来のアプローチでは，個々の SNP について独立性検定を行い，多数の SNP についての結果を Bonferroni 法や Holm 法などの多重検定法を用いて補正する，ということを行う．しかし，これらの方法は SNP 数が多数の場合には過剰に保守的であり，p 値が大きく出過ぎてしまうという欠点があった．正しく p 値を求めるためには厳密計算を行えば良いのだが，これを単純なアルゴリズムとして実装すると，計算量が被験者数に対して指数的に増加してしまい，忽ち破綻する．この問題に対して，同じエネルギーを持つミクロ状態は纏めて扱う，という統計力学的なアプローチを基に，正確な p 値を求めるアルゴリズムを開発した．

(4) 他の先駆的な確率的情報処理課題については，上記課題と並行して積極的に領域内での共同研究を行い，解法の応用の可能性を探った．具体的には，画像処理，誤り訂正符号などの課題を扱った．

４．研究成果



(1) 開発した haplotype 推定手法は，既存 8 手法と比して推定精度が最良であった．また，扱える一塩基多型(SNP)座位数も，パソコンレベルで 25 程度から 40 程度へと拡張することができた．図は，計算量の削減度合いを表わしているが，既存手法の EM だと，横軸で SNP 数が増加するにつれて，縦軸の計算量が指数的に増加しているのが分かる．一方，考案手法 GrEM では，SNP 数が増加しても計算量には殆ど影響していないことが分かる．計算量がばらついているのは，解析したデータセット（点の一つ一つ）によって本質的な複雑度が異なったためと思われる．

また，SNP データに加えてコピー数多型(CNV)データも扱えるよう拡張できた．この拡張が比較的順調に行ったのは，SNP データの時に推定アルゴリズムを集合演算という形で抽象化していたことが大きいと思われる．実装したアルゴリズムはフリーソフトウエアとして公開した．この推定法は，最適解をある条件の下で全て求めることが可能のため，強磁性-スピングラス相転移などの観点から，今後の研究に役立つと期待される．

また，情報統計力学的な手法を応用して，精度は劣るが計算量が SNP 座位数，被験者数に比例する高速推定手法を開発した．この手法によれば，haplotype 推定問題をスピングラスモデルの基底状態探索問題に近似的に置き換えることができ，今後相図の作成等が可能ではないかと思われた．

(2) 二種類の組換 hot spot 推定手法を開発した．前者に関しては，最適化手法に様々なヒューリスティクスを導入し，問題規模が小さい場合は交叉率をほぼ正しく推定できるような推定手法を開発することができた．但し，問題規模が大きくなると解が安定して求まらなくなるという問題が残った．これについては，今後モデルの改善など，様々な観点から改良を行う必要があると思われる．

後者に関しては，人工データに対して比較的安定した解を求めることができた．また，理論的な考察を更に進めたところ，求めている解が情報学的に何を意味しているのかを Kullback-Leibler 距離を用いて明確に説明することができた．同時に，この解が交叉率に関する粗い近似解であることが分かり，今後，更に精度の良い手法を開発する際の有力な手掛かりとなった．

(3) 臨床データと関連のある SNP 座位の発見については，Fisher 正確確率検定の多重検定版の理論を構築した．また，これを解くための具体的なアルゴリズムを開発した．この方法に従えば，正確な p 値を従来手法よりは格段に少ない計算量で求めることができると思われた．一方，このアルゴリズムは実装が難しく，これにはもう暫くの時間が必要であった．

(4) 他の確率的情報処理課題については，画像処理，誤り訂正符号等について新たな手法開発等がなされた．また，この際に得た知見については，例えば変分 Bayes 法や，Bowman-Levin 法などの最適化手法は，本研究の(1)や(2)で最適解を求める際に応用することができ，有用であった．

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 16 件）

[1] Masashi Kato, Qian Ji Gao, Hiroshi Chigira, Hiroyuki Shindo, Masato Inoue
A haplotype inference method based on sparsely connected multi-body Ising model
Journal of Physics: Conference Series, (in press). (査読有)

[2] Hiroyuki Shindo, Hiroshi Chigira, Tomoyo Nagaoka, Naoyuki Kamatani, Masato Inoue
Grouping preprocess for haplotype inference from SNP and CNV data
Journal of Physics: Conference Series, 197, 012009, 2009. (査読有)

[3] Ken-ichi Tamura, Miho Komiya, Masato Inoue, Yoshiyuki Kabashima
Decoding algorithm of low-density parity-check codes based on Bowman-Levin approximation
New Generation Computing, 27, 347-363, 2009. (査読有)

[4] Hiroyuki Shindo, Hiroshi Chigira, Junji Tanaka, Naoyuki Kamatani, Masato Inoue
Grouping preprocess to accurately extend application of EM algorithm to haplotype inference
Journal of Human Genetics, 53(8), 747-756, 2008. (査読有)

[5] Satohiro Tajima, Masato Inoue, Masato Okada
Bayesian-Optimal Image Reconstruction for Translational-Symmetric Filters
Journal of the Physical Society of Japan, 77(5), 054803, 2008. (査読有)

[6] Shinpei Hara, Yuta Akira, Eisuke Ishii, Masato Inoue, Masato Okada
LDPC decoding dynamics from a PCA viewpoint
Interdisciplinary Information Sciences, 13(1), 43-48, 2007. (査読有)

[7] Masato Inoue, Koji Hukushima, Masato Okada
Analysis method combining Monte Carlo simulation and principal component analysis -- application to Sourlas code
Journal of the Physical Society of Japan, 75(8), 084003, 2006. (査読有)

〔学会発表〕（計 20 件）

[1] Masashi Kato, Qian Ji Gao, Hiroshi Chigira, Hiroyuki Shindo, Masato Inoue
A haplotype inference method based on sparsely connected multi-body Ising model
International Workshop on Statistical-Mechanical Informatics 2010 (IW-SMI2010), 2010/3/8, Kyoto.

[2] 塩塚 丁二郎，永田 賢二，岡田 真人，井上 真郷
階層パターンを持つ自己相関型連想記憶モデルの PCA による解析
第 12 回情報論的学習理論ワークショップ (IBIS 2009), P042, 2009/10/20, 福岡.

[3] 荒木 佑季，永田 賢二，岡田 真人，井上 真郷
混合 Bernoulli 分布に基づく変分 Bayes 法による連想記憶モデルの解析
第 12 回情報論的学習理論ワークショップ (IBIS 2009), P047, 2009/10/19, 福岡.

[4] Hiroyuki Shindo, Hiroshi Chigira, Tomoyo Nagaoka, Naoyuki Kamatani, Masato Inoue
Grouping preprocess for haplotype inference from SNP and CNV data
International Workshop on Statistical-Mechanical Informatics 2009 (IW-SMI2009), 2009/9/13, Kyoto.

[5] Hiroyuki Shindo, Hiroshi Chigira, Tomoyo Nagaoka, Naoyuki Kamatani, Masato Inoue
Haplotype inference from SNP and CNV data assuming Hardy-Weinberg equilibrium
The 1st International Symposium on Bio-medical Science and Engineering, 2009/7/18, Tokyo.

[6] 田村 健一，小宮 美穂，井上 真郷，樺島 祥介
Bowman-Levin 法を用いた LDPC 復号
第 11 回情報論的学習理論ワークショップ (IBIS2008), 2008/10/29, 仙台.

[7] 進藤 裕之，千明 裕，田中 順治，鎌谷 直之，井上 真郷
haplotype 推定における指数関数的計算量を削減するためのグルーピング前処理
日本人類遺伝学会第 52 回大会，2007/9/14，東京．

[8] 橋口 友美，井上 真郷，岡田 真人
畳み込みフィルタ出力を考慮した画像修復
日本物理学会 2007 年春季大会，2007/3/20，鹿児島．

〔図書〕（計１件）
[1] Masato Inoue, Shin Ishii, Yoshiyuki Kabashima, Masato Okada (editors)
International Workshop on Statistical-mechanical Informatics 2009 (IW-SMI 2009)
Journal of Physics: Conference Series 197, 2009, pp. 244.

〔その他〕
ホームページ等
ソフトウエア公開
HaploBorder (SNP データより haplotype を推定するフリーソフトウエア)
http://www.eb.waseda.ac.jp/m_inoue/downloads/

６．研究組織
(1)研究代表者
井上 真郷（INOUE MASATO）
早稲田大学・理工学術院・准教授
研究者番号：70376953

(2)研究分担者
なし

(3)連携研究者
なし