様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成２５年４月１０日現在

機関番号：１２６０１
研究種目：基盤研究（B）
研究期間：２０１０～２０１２
課題番号：２２３０００９５
研究課題名（和文）　時空間情報を利用した統計遺伝学モデルの開発

研究課題名（英文）　Development of models in statistical genetics that incorporate spatio-temporal information

研究代表者
岸野　洋久（Hirohisa Kishino）
東京大学・大学院農学生命科学研究科・教授
研究者番号：００１４１９８７

研究成果の概要（和文）：
統計遺伝学の諸手法の検出力を高めるために、本研究では種々の時空間情報を利用したモデルの開発を行った。分子進化では、タンパク質にかかる多様化圧の空間集積性をイジングモデルで表現し、インフルエンザ HA において宿主細胞の受容体との結合部位に多様化圧を検出した。集団遺伝では、熱帯熱マラリアの遺伝的多様度がアフリカから離れるにつれ減衰するパターンを飛び石モデルで説明し、伝搬過程を推定した。量的遺伝では、遺伝子ネットワークの情報を取り入れ、自閉症の症例対照研究の低頻度 CNV のデータで、CNV の横切る遺伝子数に関する空間スキャン統計量により疾患関連部分グラフを抽出した。

研究成果の概要（英文）：
To improve the power of methods in statistical genetics, we developed statistical models that utilize the information of spatio-temporal structures. By developing the Ising model prior of the spatial aggregation of the dn/ds ratio, our model of molecular evolution detected the diversifying selection acting on the receptor-binding region of influenza HA. Based on the observation of the negative correlation between within-population genetic diversity of malaria and geographic distance from Sub-Saharan Africa, we modelled the expansion of malaria out of Africa by considering a one-dimensional stepping stone and estimated the rate of colonization and the local carrying capacities. By developing a circular scan statistics that measure difference in the number of genes intersected by rare CNVs between the cases and controls, we extracted disease-associated gene clusters from within a whole gene pathway.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２０１０年度	5,300,000	1,590,000	6,890,000
２０１１年度	4,000,000	1,200,000	5,200,000
２０１２年度	4,400,000	1,320,000	5,720,000
年度			
年度			
総　計	13,700,000	4,110,000	17,810,000

研究分野：統計科学
科研費の分科・細目：情報学・統計科学
キーワード：統計遺伝学，時空間情報，タンパク質の適応進化，立体構造，階層ベイズ，系統的多様性，発現情報，遺伝子ネットワーク

１．研究開始当初の背景

21 世紀になり、発現プロファイル、SNP (1 塩基多型)のデータに加え、タンパク質の立体構造、遺伝子ネットワークなどの実験データが急速に整備されている。これにより、分子レベルの遺伝現象と表現型の変化を包括的に解析する素地が整いつつある。しかし、研究開発当初において、統計遺伝学の理論はこれらの豊富な情報を余すところなく取り込んでいるとは言えない状況にあった。

２．研究の目的

本研究では、統計遺伝学で現在重要視されている課題、タンパク質の適応進化、分子系統地理、量的遺伝を取り上げ、階層ベイズモデルおよびこれと共通するアイデアの分析枠組みで時空間構造に関する付加情報を取り込み、これまでにない高い検出力でこれらの課題に答えることを目的とした。

第一の課題では、コード領域の塩基配列の進化の履歴を推測し、変異のパターンを見ることにより多様化圧の強さを推定することができる。実際には、免疫系の強い多様化圧のかかるウイルスのタンパク質においても、多様化圧は一様に働くのではなく、宿主受容体との結合部位、あるいはその隣接した領域に限定して働くことが想定される。そこで、階層ベイズモデルにより PDB に登録された立体構造の情報を分析に取り入れ、多様化圧のかかる領域を検出する方法を開発することを目的とした。第二の課題では、分集団構造と遺伝的多様度の地理的空間パターンを合体過程でモデリングし、集団の時空間伝搬のパターンを推定することを目的とした。第三の課題では、遺伝子ネットワークの情報を隣接行列から得られるノード間距離の事前分布の形で利用し、発現あるいはマーカー遺伝子型と表現型との間の関連を検出する際の多重性に伴う検出力の低下を克服することを目的とした。

３．研究の方法

(1) タンパク質の適応進化：拡張イジングモデルによる多様化圧の領域推定

遺伝子配列を比較することによってタンパク質の進化の歴史を推定することができる。塩基配列上につらなる DNA の 3 つ組（コドン）はタンパク質配列を作るアミノ酸に変換されるが、コドン・アミノ酸の対応表の冗長性のため、塩基置換にはアミノ酸を変えるもの（非同義置換）と変えないもの（同義置換）がある。非同義置換率 d_N と同義置換率 d_S の比 $\omega = d_N / d_S$ が異常に大きい場合は、アミノ酸を積極的に変異させるような多様化圧が働いていることが疑われる。

ここでは多様化圧がかかる領域を感度よく推定することが目的である。分子進化がマルコフ過程に従うと仮定し、第 h コドンの i から j に変異する推移率を

$$q_{ij}^{(h)} = \begin{cases} 0 & 2\text{ つ以上の置換を伴う変異} \\ \pi^{(j)} & \text{トランスバージョン型同義置換} \\ \kappa\pi^{(j)} & \text{トランジション型同義置換} \\ \omega_h\pi^{(j)} & \text{トランスバージョン型非同義置換} \\ \omega_h\kappa\pi^{(j)} & \text{トランジション型非同義置換} \end{cases}$$

とモデル化する。分子系統樹 T に対して、配列 $\mathbf{X} = (\mathbf{X}_1, ..., \mathbf{X}_n)$ の尤度は

$$L = \prod_{h=1}^{n} f_T(\mathbf{x}_h)$$

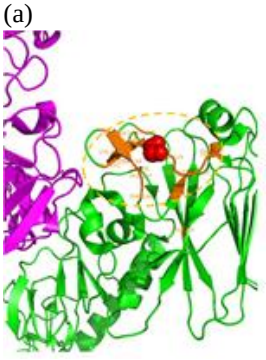
$$= \prod_{h=1}^{n} \left[\sum_{Z_{h0}} \pi_{Z_{h0}} \prod_{v_l \in V(T) \setminus v_0} \sum_{Z_{hv_l}} P_{Z_{anc(v_l)} Z_{v_l}} \left(t_{anc(v_l), v_l} \mid \omega_h \right) \right]$$

と、節を結ぶ枝における推移確率と平衡確率を乗じることにより得られる。$\omega = d_N / d_S$ については、純化圧($\omega_1 < 1$), 中立 ($\omega_2 \sim 1$), 多様化圧 ($\omega_3 > 1$)の 3 状態にカテゴライズし、拡張イジングモデル

$$P(s_1, ..., s_n) \propto \exp\left(\lambda \left(\sum_{h<h'} \left(\delta_{s_h s_{h'}} - \frac{1}{3} \right) \exp(-\alpha r_{hh'}) \right) \right)$$

により多様化圧の空間集積性の事前分布を表現する。図 1 に見られるように、タンパク質の畳み込みにより、空間的に集積している領域も一次配列上は飛び火する。立体構造の情報を取り込むことにより、初めて空間集積性の事前情報を分析に取り入れることが可能となる。周辺尤度を最大化させることにより、超パラメータ λ , α を推定する。

(a)



(b)

QDLPGNDNSTATLCLGHHAVPNGTLVKTITDDQIE
VTNATELVQSSSTGKICNNPHRILDGIDCTLIDAL
LGDPHCDVFQNETWDLFVERSKAFSNC PYDVPDY
ASLRSLVASSGTLEFITEGFTWTGV TONGGSNACK
RGP SSGFF SRL NWL KSGSTYPVLNVTMPNNDNFD
KLYIWGIHHPSTNQEQTSLYVQASGRVTVSTRRSQ
QTIIPNIGSRPWVRG SSRISIYWTIVKPGDVLVI
NSNGNLIAP GYFKMRTGKSSIMRSDAPIDTCISE
CITPNGSIPNDKPFQNVNKITYGACPKYVKQNTLK
LATGMRNVPEKQT

図 1　抗体（ピンク色）に結合するインフルエンザ HA タンパク(緑色)。(a) 宿主受容体との結合部位に位置するアミノ酸(赤)から 10 Å 以内の近傍（橙色）。(b) アミノ酸配列上で(a)に示された 10 Å 近傍に対応するアミノ酸(白色)。

(2) 分子系統地理：飛び石合体過程モデルと ABC による集団の時空間伝搬の推定

遺伝的多様性と分集団構造は、現在の集団の交配様式と過去の集団の履歴の時空間パターンが反映されている。集団が何らかの要因でボトルネックを受けると、遺伝的多様度も減少する。その後集団内に突然変異が蓄積するにつれ、次第に多様度が復元してくる。そこでしばしば、遺伝的多様度の地理的な空間分布を、それら分集団のボトルネック後の時間の空間分布の言葉に焼き直すことにより、集団の拡散と伝搬の歴史を推定することが可能となる。

図 2 は集団が次第に分布域を拡大していく様子を表現している。ある分集団が速度 r で膨張し、環境収容量に達すると d だけ離れた隣接地点に伝搬して行く。分集団間には m の移住がある。ある地点に伝搬してきた当初は、分集団は環境収容量の c だけの割合であったとすると、生息域拡大の速度は $-dr/\log c$ となる。

こうした集団の履歴に伴い遺伝的多様度の空間的不均質性が生まれる様子は、合体過程による遺伝子の系図をモデリングすることにより定量化することができる。ただし、その尤度を陽な形で書き下すことはできない。そこで近似ベイズ計算(ABC: Approximate Bayesian Computation)により、パラメータをベイズ推定する。事前分布から生成されたパラメータ値に基づき合体過程のシミュレーションを行い、遺伝的多様度の空間分布を計算する。観測値と近いものを採用することにより、パラメータの事後分布を得る。

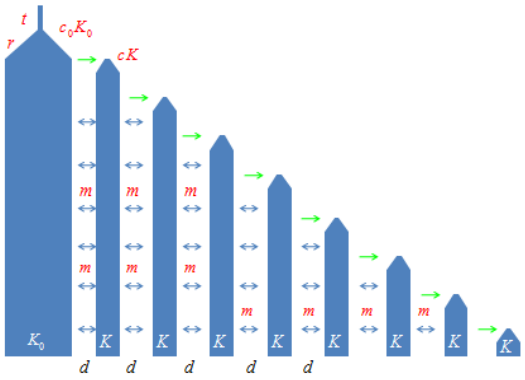


図 2　飛び石モデルと生息域拡大。

(3) 量的遺伝：遺伝子ネットワーク上のスキャン統計量による部分グラフの抽出

SNP や発現情報の解析では、多重性を補正すると p 値の閾値を極めて厳しく設定する必要がある。ところが多くの場合、それに見合うだけの大標本を得ることは困難である。この問題を克服するために、SNP や発現プロファイルを個別に扱うのではなく、機能的にまとめられる遺伝子集合のリストを考え、リストの中から有意な遺伝子集合を選択する方法などが提案されている。

ここではリストに限定することなく、タンパク質間相互作用や先行研究の蓄積である文献の共起表現などを統合して構築された遺伝子ネットワークの情報を低頻度 CNV (rare Copy Number Variation) と疾患の関連を調べる量的遺伝の解析に取り入れる。グラフの隣接行列から計算されるノード間の距離に基づき、空間スキャン統計量を構築する。低頻度 CNV が横切る遺伝子のうち円内に位置する遺伝子の割合をケース群コントロール群で対比し、p 値を最小にする円を探索する（図 3)。

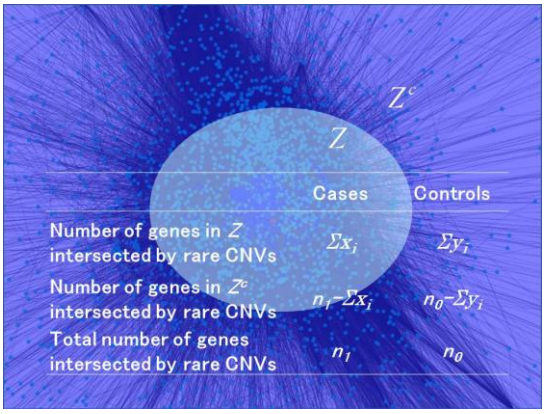


図 3　遺伝子ネットワークにおける低頻度の CNV が横切る遺伝子数のスキャン統計。

４．研究成果

(1) .拡張イジングモデルによるインフルエンザ HA にかかる多様化圧の領域推定

非同義置換率 d_N と同義置換率 d_S の比 $\omega = d_N/d_S$ に関する立体構造上の空間的集積性の事前分布を導入し、ベイズ因子の比較から空間相関のレンジを推定した。40 年にわたるインフルエンザ HA の配列を解析したところ、宿主細胞の受容体との結合部位に多様化圧がかかっていること、配列のみの解析は擬陽性を多く拾うことが分かった。

(2) 飛び石モデルと ABC によるマラリア伝搬速度の推定

世界中からサンプリングされた 519 の熱帯熱マラリアについて、2 つのハウスキーピング遺伝子 serca と adsl を解析したところ、アフリカからアジア、オセアニアへと遠ざかるにつれ、遺伝的多様度が減少する様子が観察された。すなわち、マラリアの侵入以降突然変異が集団に蓄積するが、東に行くほど経過

時間が短い。サブサハラを起源とするマラリアが東へ伝搬する様子を、飛び石モデルで表現した。遺伝子の合体過程をシミュレートする近似ベイズ計算により局所集団の成長速度と環境収容量、伝搬速度を推定した。

(3) 遺伝子ネットワークにおける疾患関連低頻度 CNV の集積性

自閉症のケースコントロール研究における低頻度 CNV を円形スキャンで分析したところ、癌関連遺伝子の他にユビキチン関連の遺伝子群が抽出された。シミュレーションを通して提案手法が優れた感度および特異度を持つことが示された。

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 19 件）

① Bizinoto, M. C., Yabe, S., Leal, É., Kishino, H., de Oliveira Martins, L., de Lima, M., L., Morais, E., R., Diaz, R. S., and Janini, L. M. (2013). Codon pairs of the HIV-1 *vif* gene correlate with CD4+ T cell count. *BMC Infectious Diseases*. **13**: 173. (査読有) DOI: 10.1186/1471-2334-13-173

② Kitada, S., Fujikake, C., Asakura, Y., Yuki, H., Nakajima, K., Vargas, K. M., Kawashima, S., Hamasaki, K., and Kishino, H. (2013). Molecular and morphological evidence of hybridization between native *Ruditapes philippinarum* and the introduced *Ruditapes* form in Japan. *Conservation Genetics*. in press. (査読有) DOI: 10.1007/s10592-013-0467-x

③ Ishibashi, K., Mawatari, N., Miyashita, S., Kishino, H., Meshi, T., and Ishikawa, M. (2012). Coevolution and hierarchical interactions of *Tomato mosaic virus* and the resistance gene *Tm-1*. *PLoS Pathogen*. **8(10)**: e1002975. (査読有) DOI: 10.1371/journal.ppat.1002975

④ Mollah,M. M. H., Mollah, M. N. H., and Kishino, H. (2012). *β*-empirical Bayes inference and model diagnosis of microarray data. *BMC Bioinformatics*. **13**: 135. (査読有) DOI: 10.1186/1471-2105-13-135

⑤ Nishiyama, T., Kishino, H., Suzuki, S., Ando, R., Niimura, H., Uemura, H., Horita, M., Ohnaka, K., Kuriyama, N., Mastuo, K., Guang, Y., Wakai, K., Hamajima, N., and Tanaka, H. (2012). Detailed analysis of Japanese population substructure with a focus on the Southwest Islands of Japan. *PLoS ONE*. **7 (4)**: e35000. (査読有) DOI: 10.1371/journal.pone.0035000

⑥ Mariadassou, M., Bar-Hen, A., and Kishino, H. (2012). Taxon Influence Index: assessing taxon-induced incongruities in phylogenetic inference. *Systematic Biology*. **61**: 337–345. (査読有) DOI: 10.1093/sysbio/syr129

⑦ Nishiyama, T., Takahashi, K., Tango, T., Pinto, D., Scherer, S. W., Takami, S., and Kishino, H. (2011). A scan statistic to extract causal gene clusters from case-control genome-wide rare CNV data. *BMC Bioinformatics*. **12**: 205. (査読有) DOI: 10.1186/1471-2105-12-205

⑧ Watabe, T. and Kishino, H. (2010). Structural considerations in the fitness landscape of a virus. *Molecular Biology and Evolution*. **27**: 1782-1791. (査読有) DOI: 10.1093/molbev/msq056

⑨ Koyano, H. and Kishino, H. (2010). Quantifying biodiversity and asymptotics for a sequence of random strings. *Physical Review E*. **81**: 061912. (査読有) DOI: 10.1103/PhysRevE.81.061912

⑩ Tanabe, K., Mita, T., Jombart, T., Eriksson, A., Palacpac, N., Ranford-Cartwright, L., Sawai, H., Sakihama, N., Horibe, S., Ohmae, H., Nakamura, M., Ferreira, M. U., Escalante, A. A., Prugnolle, F., Björkman, A., Färnert, A., Akira Kaneko, A., Horii, T., Manica, A., Kishino, H., and Balloux, F. (2010). *Plasmodium falciparum* accompanied the human expansion out of Africa. *Current Biology*. **20**: 1283-1289. (査読有) DOI: 10.1016/j.cub.2010.05.053

⑪ 渡部輝明，岸野洋久 (2013). ウイルスタンパク質変異にかかる多様化圧の空間分布. 統計数理. 第60 巻第2号: 305–316. (査読有)

⑫ 渡部輝明，岸野洋久 (2012). タンパク質適応進化の時空間モデル. 統計数理. 第60 巻第1号: 27–36. (査読有)

〔学会発表〕（計 34 件）

① Kishino, H.（代表）, Watabe, T., Nakamichi, R., and Kitada, S. Spatiotemporal Modeling to Measure the Effects of Mutations and Selection Pressures. The 59th World Statistics Congress（招待講演）. 2013年8月25日. 香港、中華人民共和国

② Nakamichi, R. （代表）, Kishino, H., and Kitada, S. Inference of direct effect and module structure of transcriptome behind phenotype via graphical modelling. International Biometric Conference. 2012年8月30日. 神戸国際会議場（兵庫県）

③ 渡部輝明（代表），岸野洋久. 階層ベイズ

モデルによるタンパク質にかかる多様化圧の時空間集積性の推定. 日本進化学会. 2012年08月21日. 首都大学東京（東京都）

④ Mollah, M. M. H.（代表）, Mollah, M. H. N., and Kishino, H. Detection of irregular patterns of gene expression and diagnose the model based on β-weight. 日本計量生物学会. 2012年5月25日. 統計数理研究所（東京都）

⑤ Nishiyama, T.（代表）, Hamajima, N., Suzuki, S., Kishino, H., Japan Multi-institutional Collaborative Cohort. Japanese population structure estimated from the Japanese Multi-institutional Collaborative Cohort data. American Society of Human Genetics. 2011年10月13日. モントリオール(カナダ)

⑥ Koyano, H.（代表） and Kishino, H. Measuring alpha diversity of microbial communities. Evolutionary Biology Meeting at Marseilles. 2011年9月29日. マルセイユ(フランス)

⑦ Mollah, M. M.（代表）, Mollah, N. H., 岸野洋久. Beta-divergence approach to detect abnormal expression profiles. 日本統計関連学会. 2011年9月6日. 九州大学(福岡県)

⑧ Watabe, T. and Kishino, H.（代表）. Spatio-temporal modeling of protein adaptive evolution. Brazilian Congress of Genetics. 2011年8月31日. Águas de Lindóia, Brazil (ブラジル)

⑨ Watabe, T.（代表） and Kishino, H. Structural considerations in the fitness landscape of a virus. Society for Molecular Biology and Evolution. 2011年7月27日. 京都大学(京都府)

⑩ Nishiyama, T.（代表）, Takahashi, K., Tango, T., Takami, S., Kishino H. Scan statistics for pathway-based genomewide association study. American Society of Human Genetics. 2010年11月3日. ワシントン(米国)

⑪ 渡部輝明（代表）, 岸野洋久. タンパク質にかかる多様化圧の空間分布とベイズ推定. 日本統計関連学会. 2010年9月7日. 早稲田大学（東京都）

⑫ 渡部輝明（代表）, 岸野洋久. タンパク質共進化と分子進化速度の時空間変動. 日本蛋白質科学会. 2010年6月16日. 札幌コンベンションセンター（北海道）

⑬ 岸野洋久（代表）, Thorne, J. L., de Oliveira Martins, L., 渡部輝明. ベイズ統計手法の応用の現状と展望. 日本計量生物学会. 2010年5月22日. 統計数理研究所（東京都）

〔図書〕（計 1 件）

① 繁桝算男,岸野洋久,大森裕浩（監訳）(2011). ベイズ統計分析ハンドブック (Bayesian Thinking: Modeling and Computation, edited by D. K. Dey and C. R. Rao). 朝倉書店. 総ページ数: 1047

６．研究組織

(1) 研究代表者

岸野　洋久（Hirohisa Kishino）
大学院農学生命科学研究科・教授
研究者番号：00141987

(2) 研究分担者

渡部　輝明（Teruaki Watabe）
高知大学・医学情報センター・講師
研究者番号：90325415

西山　毅（Takeshi Nishiyama）
高知大学・医学情報センター・講師
研究者番号：40571518