論文の内容の要旨

生産・環境生物学専攻

平成 24 年度博士課程入学

氏名　小野木章雄

指導教員名　岩田洋佳

論文題目

効率的育種のための全ゲノム情報を活用する統計遺伝学的手法に関する研究

育種とは遺伝的な改良を通じて望ましい性質を持つ品種を作出することを指す。近代的の進展は統計学と密接に関わってきた。特に Best linear unbiased predictor（BLUP）法による育種価の予測や、DNA マーカーを用いた連鎖地図の作成、量的形質遺伝子座（quantitative trait locus、QTL）のマッピングといった統計的手法は、動植物の育種に大きな貢献を果たしてきた。近年分子生物学の発展により、高密度の DNA マーカーが多数の個体・系統で得られるようになっている。そのためそれらを利用した表現型あるいは育種価の予測、つまりゲノムワイド予測や、QTL のマッピング、つまりゲノムワイド関連解析といった統計的手法が育種でも大きな注目を集めている。そこで本研究では高密度かつ多量のマーカー情報の育種への活用に向けて、必要と考えられる統計的手法の開発、改善、及び検証をイネやウシなどから得られた実データあるいはシミュレーションデータをもとに行った。

これまで育種で用いられる統計的手法は、それを実装した多くのソフトウェアやプログラムを通じて様々な種や集団で試され、その性質や有用性が評価されてきた。ゲノムワイド予測や関連解析についても多くの手法が提案されているが、それらが使いやすく安定して働くソフトウェアとして提供されることは、その有用性を知る上で重要となる。そこで本研究ではまずゲノムワイド予測及び関連解析のためのソフトウェア「VIGoR（variational Bayesian inference for genome-wide regression）」を開発した。これらの解析では全マーカーを表現型値あるいは育種価に回帰する手法（ゲノムワイド回帰）がよく用いられるが、VIGoR では主用な 6

つのベイズ回帰手法について、一般に用いられるマルコフ連鎖モンテカルロ法（Markov chain Monte Carlo、MCMC）より高速な変分ベイズ法を用いて解を求める。このため MCMC に基づく手法より、多数のマーカーあるいは形質からなるデータセットへの適用が可能となる。VIGoR は Linux 及び Mac 向けのコマンドラインプログラム、及びプラットフォームを問わない統計処理言語 R のパッケージ両方の形で提供される。いずれの形式においても、ハイパーパラメータの最適化や予測能力検証のための交差検証を自動で行うことができる。またコマンドラインプログラムは独自の入出力ファイル形式を持つが、他のゲノムワイド関連解析ソフトウェアの入力ファイルや、マーカー遺伝子型補完用ソフトウェアの出力ファイルを入力ファイルとして用いることもできる。以上の仕様により VIGoR は、ベイズに基づくゲノムワイド回帰手法の育種での有用性評価に貢献可能なソフトウェアであると考えられる。

ゲノムワイド回帰においては多数のマーカー遺伝子型を説明変数として用いるため、説明変数の数（p）が被説明変数（個体・系統）の数（n）よりはるかに多い状況で、その効果をどのように学習するか、いわゆる“large p, small n”と呼ばれる問題がある。これまでにベイズ回帰手法を含め様々な手法が提案されてきたが、それらの性質や適用範囲は明確になっていない。本研究ではアジア栽培イネ（*Oryza sativa* L.）において、実データ及びシミュレーションデータを用いて 8 つ予測手法とそれらの平均を用いる合計 9 つの手法を比較しその適用範囲を探索した。実データ（8 形質）を用いた比較では、最も正確な予測は 3 形質ではランダムフォレスト、2 形質では reproducing kernel Hilbert space 回帰（RKHS）、1 形質では genomic BLUP（GBLUP）で、2 形質では全手法の平均で得られた。シミュレーションデータは実データのマーカー遺伝子型をもとに、QTL の数、遺伝率、エピスタシスの有無、学習セットのサイズ、及び連鎖不平衡（linkage disequilibrium、LD）の長さを考慮した合計 150 通りのシナリオに沿って作成した。各シナリオで予測の正確さについての手法間順位を比較することにより、各予測手法の持つ性質を推測することができた。Lasso や elastic net などの変数選択手法や、GBLUP、RKHS、ランダムフォレストはそれぞれ明確な適用範囲がある一方、Bayesian lasso などのベイズ回帰手法や全手法平均は広い適用範囲を持っていた。以上の結果はゲノムワイド予測のための手法選択においてより詳細な知見を与えるとともに、アジア栽培イネの育種に限らず、LD の程度が強い比較的小さな集団におけるゲノムワイド予測にも有用な知見を与えると考えられる。

ゲノムワイド予測はマーカー遺伝子型から表現型予測を行うため、それに基づく選抜、つまりゲノミックセレクションでは、表現型評価を行わずに選抜が可能となり改良速度の向上などが期待できる。しかしゲ

ノムワイド予測では環境情報を用いないため、未試験の環境下での植物の表現型予測は難しい。一方で環境情報を用いた表現型予測の方法としては作物モデルと呼ばれる数理モデルが知られるが、このモデルでは遺伝情報を考慮しないため新しい遺伝子型（品種・系統）に対する予測を行うことができない。そこで本研究では作物モデルとゲノムワイド予測モデルを統合した新たなモデルを提案する。具体的にはイネ出穂期を予測する developmental rate（DVR）モデルのパラメータにゲノム情報と結びつけたベイズモデルを複数開発した。これらのモデルを緯度において多様な環境下で栽培されたコシヒカリとカサラスの戻し交雑自殖系統のデータに適用した。手法の予測能力は 1 環境抜き、1 遺伝子型抜き、1 環境と遺伝子型の組み合わせ抜きの 3 通りの交差検証を用いて評価した。また比較として DVR モデルでパラメータを推定後、そのパラメータに対してゲノムワイド予測を行うことで新たな遺伝子型に対する予測を行う 2 段階法も行った。その結果、DVR モデルのパラメータをベイズ回帰手法（extended Bayesian lasso）を用いて全ゲノムワイドマーカーに回帰し、DVR モデルパラメータとそれに対するマーカー効果を同時に推定する提案モデルが、全ての交差検証において最も正確な予測を与えた。この結果は環境情報も利用した新たなゲノムワイド予測モデルの可能性を示すものであり今後の進展が望まれる。

ゲノムワイド予測が新たな育種手法として注目を集める一方で、動植物ともにこれまで続けられてきた育種手法も存在する。それら既存の育種システムを大きく変更することなくゲノムワイド予測の利点を取り込むことが望ましい。肉牛など動物育種においては後代検定を行い血縁情報に基づく BLUP 法で選抜候補の育種価を推定することが広く行われている。この既存のシステムにゲノム情報を取り込む 1 つの手法として、血縁情報とマーカー遺伝子型情報の双方に基づく BLUP 法（single-step GBLUP、ssGBLUP）が提案されている。日本固有の肉牛である黒毛和種においても近年徐々にマーカー遺伝子型情報が蓄積されてきているが、その育種システムにどのようにゲノム情報を取り入れるかは未だ検証されていない。そこで本研究では黒毛和種において ssGBLUP 法を適用し、その予測能力の検証を行った。その結果、肥育牛の表現型においては検証した 3 枝肉形質全てにおいて、ssGBLUP 法は BLUP 法より予測の正確さが高かった。またどのような個体のマーカー遺伝子型を優先的に決定するべきかを調べるため、既にマーカー遺伝子型を持つ個体を複数のサブセットに分割し、各セットを除いた場合に予測の正確さに与える影響を検証した。その結果、後代を多く持つ種雄牛や表現型記録を持つ肥育牛のマーカー情報が大きな影響を持つことが分かり、これらの個体について優先的にマーカー遺伝子型を収集することで、効率的に予測能力を向上できることが示された。以上の結果はゲノム情報

の活用が始まったばかりの黒毛和種の今後の育種に有用な知見となると考えられる。

育種は統計学的な手法に支えられ発展してきたため、統計的手法の発展は特定の作物や品種に限らず育種分野全体に恩恵をもたらすことができる。そのため本研究で得られた知見は実際にデータ解析を行ったイネやウシに限らず、様々な種や集団において有用な知見となり得る。高密度マーカーや全塩基配列に代表される入手可能なゲノム情報は今後も増加していくと予想されるため、ゲノムワイド予測及び関連解析のための効率的な統計的手法の重要性は、育種においても増していくであろう。さらに現在分子生物学の進展により、様々な分子、つまり RNA やタンパク質、代謝産物等について多量の情報が得られるようになってきた。これらの情報は生物の発生や発達、及び環境への応答に対するより詳細な知見をもたらすと考えられる。そのため今後の育種にはゲノムや表現型だけでなく、これらオミックスと呼ばれる情報や環境情報を利用するような統計的手法が必要となると考えられる。本研究で示されたゲノムワイド予測と作物モデルを組み合わせた手法の優位性は、複数の情報を統合するモデリングがその活用のために有用であることを示唆している。