論文の内容の要旨

生産・環境生物学専攻

平成 28 年度博士課程進学

氏名　田中凌慧

指導教員名　岩田洋佳

論文題目　ゲノミック予測を用いた育種の効率化・最適化に関する理論的研究

2050 年までに地球の人口は 90 億人を突破すると試算されており、持続可能な食料供給を実現するには育種を加速する技術の開発は不可欠である。ゲノミック予測とは、ゲノムワイドマーカー遺伝子型をもとに表現型値を予測することをいい、この予測値をもとに選抜を行ことをゲノミック選抜という。ゲノミック選抜は、育種において大きなコストを要している表現型評価を、予測モデルを用いて大幅に省略でき、育種の効率化に貢献する重要技術と考えられている。本研究では、ゲノミック予測を用いた育種の効率化・最適化を目的とした新規手法の開発・予測モデルの評価を行なった。

1. 能動学習に基づくゲノミック予測モデルの効率的構築

ゲノミック選抜は、ある系統を選抜するか淘汰するかの二値分類問題として捉えることもできる。このとき、予測モデルの分類精度は選抜効率に直結する重要な因子であり、できる限り高い

分類精度を実現する予測モデルを構築することが望ましい。一般に、ゲノミック予測の精度は、モデル構築に用いる系統に依存することが知られており、同じ系統数でも、どの系統を用いるかを適切に選択することで、高精度な予測モデルを構築できると考えられる。無作為抽出された訓練データを受動的に用いて予測モデルを学習するのではなく、既存のデータと予測モデルを踏まえて能動的に訓練データを選択し学習する方法は能動学習とよばれ、機械学習の分野を中心に開発が進み、創薬等の分野では既に実用化されている。能動学習は幅広い研究分野でその有効性が確かめられているが、ゲノミック予測で扱う高次元かつノイズの大きいデータでも機能するかは未知数であった。

そこで、本研究では、能動学習をゲノミック予測に応用することで、予測精度を効率よく向上できるかをシミュレーションを用いて検証した。遺伝子型の選抜・淘汰を分類問題としてあつかい、サポートベクトルマシンを用いて分類モデルを構築した。能動学習には、最も標準的なアルゴリズム uncertainty sampling を採用した。４つの実データセットを用いて検証した結果、延べ 22 形質のうち 17 形質で能動学習により有意に分類精度が向上し、3 形質で低下した。以上の結果から、能動学習がゲノミック予測の訓練データ選択法として有用であることが示された。

2. ベイズ最適化に基づく優良系統の効率的発見

遺伝資源の持つ多様な有用変異を活用することは育種における重要課題である。しかし、現状では、遺伝資源の利用はコア・コレクションに含まれるごく一部の系統に限られている。ゲノミック予測を用いれば、一部の系統の表現型とマーカー遺伝子型から構築された予測モデルをもとに、マーカー遺伝子型が取得された全ての系統について遺伝子型値を予測できる。一部系統の表現型評価-モデル更新-未試験系統の選抜、というサイクルを繰り返すことで、マーカー遺伝子型をもつ全ての遺伝資源系統を対象に、有用系統の探索を行うことができる。

本研究では、ゲノミック予測を用いた優良遺伝資源系統の探索を black-box 最適化の枠組みで定式化するとともに、ベイズ最適化とよばれる最適化アルゴリズムに基づき、期待改善量という新たな選抜基準をもとに選抜を行う方法を提案した。期待改善量は予測モデルの不確実性を考慮した選抜基準であり、不確実性が高い、すなわち、非常に劣った系統である可能性もあるが、既存の系統よりも優れた系統である可能性も高い系統を優先的に選ぶ戦略を与える。実データを用いたシミュレーションによる検証の結果、期待改善量に基づく選抜戦略は、通常のゲノミック予測で行われる選抜戦略に比べて、平均で 30%ほど少ない試験系統数で、遺伝資源内の優良系統を発見できた。通常の選抜戦略は、単純に予測値の大きな系統から順に選抜するが、その場合、一部の系統だけで構築された予測モデルを盲信してしまう。期待改善量に基づく選抜戦略は、不確実性を考慮することにより、予測モデルの誤りを適宜修正しつつ、既存の系統を上回る系統を選抜できると考えられた。この結果から、期待改善量に基づく選抜戦略により、ゲノミック予測を用いた優良遺伝資源系統の探索をさらに加速できると期待される。

3. ゲノミック予測における多環境試験デザインの最適化

多環境試験は、遺伝子型と環境の交互作用（GxE; genotype-by-environment interaction）に関する情報を得るために必須である。しかし、多数の系統を用いた大規模な多環境試験を行うには大きな金銭的・労力的に大きなコストが必要であり、通常は、主要な系統に絞って多環境試験を実施し、それらの系統についてのみ GxE を評価する。しかし、ゲノミック予測を多環境の表現型データに拡張する（多環境ゲノミック予測）ことで、一部の表現型データをもとに、試験しなかった表現型を補完することも可能である。つまり、ゲノム情報によって系統間の類似性が定義されていれば、必ずしも一部の系統を選んで多環境試験を行う必要はなく、それぞれの環境で異なる系統を試験しても、GxE に関する知見を得ることができる。

この場合にも、能動学習により試験するべき系統を選んだ場合と同様に、各環境で試験すべき系統を適切に選ぶことにより、得られる予測モデルの精度が向上する可能性がある。本研究では、ゲノミック予測のモデル構築のためにどの系統を用いるか、という訓練集団の最適化のために提案された予測誤差分散（PEV; prediction error variance）および決定係数（CD; coefficient of determination）を多環境におけるゲノミック予測にも拡張し、どの系統をどこで試験すべきか、という多環境試験のデザインの最適化に用いた。PEV や CD を多環境試験のデザインの最適化に用いる場合には、遺伝子型値の環境間相関や、対象形質の遺伝率を超パラメータとして事前に設定する必要がある。本研究では、この超パラメータの設定が、PEV や CD によって得られる多環境試験のデザインを大きく左右することを明らかにした。例えば、環境間相関が低い場合には、すべての候補系統を満遍なく試験するようなデザインが選ばれ、逆に、環境間相関が高い場合には、一部の代表的な系統を複数の環境で試験するようなデザインが選ばれた。また、多環境試験のデザインを PEV や CD によって最適化する場合には、これら超パラメータを妥当な値に設定する必要があることを明らかにした。例えば、表現型値に GxE の影響がほとんどなく表現型値の環境間相関が 0.9 を超えるような場合に、環境間相関を 0.25 と設定してしまうと、PEV や CD を用いることにより、予測精度が逆に悪化した。しかし、真の状態と大きく異なる設定をしない限り、PEV や CD による最適多環境試験デザインを用いて予測精度を改善できることがわかった。

4. ゲノミック予測に基づく交配後代の分離予測に関するシミュレーション研究

ゲノミック予測は、ある個体の持つマーカー遺伝子型をもとに、その個体の遺伝子型値を予測する手法である。そして、個体や系統を、表現型値ではなく予測値をもとに選抜するのが基本的な利用法である。しかし、両親のマーカー遺伝子型とマーカー間の組換え価が与えられれば、その後代個体のマーカー遺伝子型をシミュレーションによって生成できる。こうして生成される仮想のマーカー遺伝子型にゲノミック予測を適用すれば、後代個体のもつ遺伝子型値の平均値や分散の予測値を得ることができる。これら予測値は育種家が交配組合せを選定する際に有益な情報

となる。

後代の分離予測はゲノミック予測の重要な活用手段であるにも関わらず、分離予測の精度（つまり、後代遺伝子型値の平均値や分散の予測精度）については十分な検討がなされてこなかった。本研究では、ベイズリッジ回帰（BRR; Bayesian ridge regression）およびBayesAとよばれる２つの代表的な予測モデルについて、後代分離の予測精度に注目したモデル比較を行なった。

まず、数少ない先行研究のほとんどが、後代分散を厳密に正確な方法で計算していないことを明らかにした。具体的には、訓練データから構築された予測モデルには常に不確実性がともなうため、後代遺伝子型値の分散を予測する場合には、その不確実性を考慮しなければならない。考慮しない場合には、後代分散が大きく過小予測される可能性があることを理論式の導出により明らかにした。さらに、BRRとBayesAが、後代平均の予測についてはほとんど同程度の予測精度を与えるにも関わらず、後代分散の予測については、BayesAのほうがはるかに優れた予測精度を与えることを明らかにした。BRRによって予測された後代分散は、ほとんどの交配組み合わせで似通った値を示す縮小予測になる傾向が、BayesAに比べて強く見られた。すなわち、交配組み合わせ間で後代遺伝子型値の分散の大小を予測・比較したい場合には、BayesAを使用することが強く推奨される。BayesAに類似するBayesBが、別の観点からも、交配による世代の変化に頑健な予測ができることが先行研究で示されており、本研究の結果と合わせて、BRRに対するBayesAの優位性が示された。

ゲノミック予測は、育種家によりともすると主観的に行われる育種を客観化するための強力な道具であり、本研究で扱った能動学習やベイズ最適化は、その道具を合理的に運用するための手法である。本研究で提案した手法により、いくつかの単純化された条件のもとでは、ゲノミック予測の優れた運用法が与えられることがわかった。しかし、実際の植物育種は非常に複雑であり、本研究はそのごく一部を切り取って最適化したものにすぎない。持続的な食料供給の実現に向けて、本研究のようなアプローチをもとにした育種のモデル化と最適化を、さらに推し進めることが望まれる。