KITE: Exploiting Kernel Methods to Identify Genetic Interactions from High-dimensional QTL Data

Xiang Fu
xfu@bu.edu
Boston University Faculty of Computing & Data Sciences

Contents

1 Background 3
2 Proposal 4
3 Code 5
3.1 Breakdown 5
3.2 Code Availability 6
4 References 7

1 Background

Quantitative trait locus (QTL) mapping has been extensively used to study the genetic basis of complex traits in various organisms, from model systems like yeast and mice to crops and livestock [1]. By measuring genetic variation and phenotypic outcomes in large populations, QTL studies aim to identify specific genomic regions that contribute to trait differences. Traditional QTL mapping methods, such as interval mapping and composite interval mapping, scan the genome for individual loci that show a significant association with the trait [2]. However, these methods primarily detect additive effects, where each locus contributes independently to the phenotype. They often fail to capture epistatic interactions, where the effect of one locus is modulated by the allelic state at other loci [3].

Epistasis is a pervasive phenomenon in complex trait genetics, arising from various molecular mechanisms such as protein-protein interactions, regulatory cascades, and metabolic networks [4]. Ignoring epistasis can lead to an incomplete or even misleading picture of the genetic architecture underlying a trait. For example, studies in yeast have shown that accounting for epistasis improves the prediction of growth phenotypes from genotype data [5]. Similarly, in agricultural contexts, modeling epistasis is crucial for predicting the performance of hybrid crosses and optimizing breeding strategies [6]. However, detecting epistasis in QTL studies remains challenging due to the combinatorial explosion of possible interactions and the stringent statistical thresholds required to control false positives [7].

Existing approaches for identifying epistatic interactions in QTL data fall into two main categories: exhaustive search methods and model-based methods. Exhaustive search methods, such as a full scan of all pairwise interactions, are computationally expensive and suffer from low power due to multiple testing correction [8]. Model-based methods, such as LASSO regression or Bayesian epistatatic association mapping, aim to reduce the search space by imposing sparsity constraints or prior distributions on the interaction coefficients [9,10]. However, these methods still struggle to detect higher-order interactions and may miss important non-linear relationships between loci.

Recently, there has been growing interest in applying machine learning techniques, particularly kernel methods, to capture non-linear structure in high-dimensional genetic data [11]. Kernel methods work by implicitly mapping the original genotype space to a higher-dimensional feature space, where complex interactions may become more separable. By designing appropriate kernel functions, one can encode prior knowledge about the structure of genetic interactions, such as their sparsity, modularity, or phylogenetic relationships between variants [12]. Kernel-based approaches have shown promising results in predicting complex traits from genotype data and detecting epistatic interactions in simulated studies [13]. However, their application to real QTL datasets remains limited, and their ability to provide biological insights into the nature of genetic interactions is still an open question.

In this project, we propose to develop a kernel-based framework for identifying epistatic interactions from high-dimensional QTL data. By leveraging the power of kernel methods to capture non-linear relationships between loci, we aim to improve the efficiency and interpretability of epistasis detection in QTL studies. Our approach will integrate techniques from machine learning, statistics, and genetics to provide a more comprehensive understanding of the genetic architecture of complex traits.

2 Proposal

Here we propose a novel kernel-based framework for identifying epistatic interactions from high-dimensional QTL data. Our approach leverages the power of kernel methods, specifically kernel PCA and kernel ridge regression, to learn a lower-dimensional representation of the genotype space that captures non-linear interactions between loci. By designing biologically informed kernel functions and integrating techniques from machine learning, statistics, and genetics, we aim to improve the efficiency, interpretability, and generalizability of epistasis detection in QTL studies.

Our framework will explore various encoding schemes to represent the genotype matrix, such as one-hot encoding, orthogonal polynomial coding, and haplotype-based encodings [7,14]. These encodings will capture different aspects of the genetic data, such as allelic states, dominance relationships, and linkage disequilibrium between variants. We will also preprocess the data to handle missing genotypes, population structure, and other potential confounders using techniques like imputation, principal component analysis, and linear mixed models [15].

A suite of kernel functions will be designed to capture different forms of epistatic interactions, based on prior biological knowledge and statistical assumptions. These will include polynomial kernels, which capture multiplicative interactions between loci; Gaussian RBF kernels, which capture non-linear similarity in the genotype space; and string kernels, which capture complex relationships between haplotypes [8,16]. We will also explore ways to incorporate functional annotations, such as gene ontology terms or protein-protein interaction networks, into the kernel design to prioritize biologically plausible interactions [17]. Cross-validation and model selection techniques, such as Akaike information criterion or Bayesian information criterion, will be used to select the most appropriate kernel function for a given dataset [18].

Kernel PCA will be applied to the genotype matrix to learn a lower-dimensional representation that maximally captures the non-linear structure of the data [9]. This will help to denoise the genotype space, reduce computational complexity, and improve the power to detect epistatic interactions. Other kernel-based dimensionality reduction techniques, such as kernel canonical correlation analysis or kernel independent component analysis, will also be explored to extract features that are most informative for the phenotypic trait [19].

Kernel ridge regression will be used to associate the reduced-dimensionality genotype representation with the phenotypic readouts [10]. This will allow us to identify specific regions of the genotype space that are predictive of the trait, potentially corresponding to key genetic interactions. Techniques from statistical learning theory, such as support vector machines or random forests, will also be employed to detect epistatic interactions in a more flexible and robust manner [20]. Permutation-based significance thresholds and false discovery rate procedures will be used to control for multiple testing and false positives [21].

The putative epistatic interactions identified by our approach will be prioritized based on their statistical significance, effect size, and biological plausibility. Information from various sources, such as gene

expression data, protein-protein interaction networks, and functional annotations, will be integrated to provide mechanistic insights into the detected interactions [22]. Our findings will be validated using independent QTL datasets, genome-wide association studies, and experimental perturbations in model organisms or cell lines [23]. Interactive visualization tools and user-friendly software packages will be developed to facilitate the interpretation and dissemination of our results, allowing researchers to explore the detected interactions and their biological context.

By combining the strengths of kernel methods, statistical genetics, and biological prior knowledge, our proposed framework aims to provide a powerful and interpretable approach for detecting epistatic interactions in QTL studies. We anticipate that our methodology will help to unravel the complex genetic architecture of diverse traits, from disease susceptibility to agricultural productivity, and guide the development of more accurate predictive models and targeted interventions. Ultimately, we hope that our work will contribute to a more comprehensive understanding of the genotype-phenotype relationship and its implications for basic biology, medicine, and biotechnology.

3 Code

3.1 Breakdown

To implement our proposed kernel-based framework for detecting epistatic interactions in QTL data, we will develop a modular and extensible software package using the Python programming language. The package will leverage existing libraries for scientific computing, machine learning, and data visualization, such as NumPy, SciPy, scikit-learn, and Matplotlib [24,25].

The code will be organized into several modules, each handling a specific aspect of the analysis pipeline:

Data Input and Preprocessing

- Functions for reading genotype and phenotype data from various file formats (e.g., CSV, VCF, HDF5)
- Classes for representing genotype and phenotype matrices, with methods for data quality control, filtering, and imputation
- Functions for encoding genotypes using different schemes (e.g., one-hot encoding, orthogonal polynomial coding)
- Classes for handling population structure, relatedness, and other confounding factors

Kernel Design and Selection

- Functions for computing various kernel matrices from genotype data (e.g., polynomial, Gaussian RBF, string kernels)
- Classes for representing kernel functions and their hyperparameters
- Functions for incorporating biological prior knowledge into kernel design (e.g., functional annotations, interaction networks)
- Functions for kernel matrix normalization and regularization
- Classes for kernel selection and hyperparameter tuning using cross-validation and model selection criteria

Dimensionality Reduction and Feature Extraction

- Functions for performing kernel PCA, kernel CCA, and other kernel-based dimensionality reduction techniques
- Classes for representing low-dimensional genotype embeddings and their properties
- Functions for visualizing and interpreting the reduced-dimensionality genotype space

QTL Mapping and Epistasis Detection

- Functions for performing kernel ridge regression and other kernel-based association tests
- Classes for representing QTL mapping results and their statistical properties
- Functions for detecting epistatic interactions using support vector machines, random forests, and other statistical learning methods
- Functions for multiple testing correction and false discovery rate control

Biological Interpretation and Validation

- Functions for integrating QTL mapping results with external biological datasets (e.g., gene expression, protein-protein interactions)
- Classes for representing biological networks and their properties
- Functions for visualizing and exploring the detected epistatic interactions in a biological context
- Functions for generating testable hypotheses and guiding experimental validation

We also plan to develop a user-friendly command-line interface and graphical user interface on the web to facilitate the application of the framework to diverse QTL datasets.

By providing a modular and flexible software solution, we aim to enable researchers to easily integrate our kernel-based approaches with their existing QTL analysis pipelines and to extend our methods to new datasets and biological questions. The software will be designed with a focus on usability, scalability, and extensibility, allowing users to seamlessly incorporate our framework into their current workflows and adapt it to their specific research needs. Furthermore, our software will be designed to handle diverse types of QTL data, including different organisms, experimental designs, and genotyping platforms. With supporting standard data formats and providing utilities for data import, export, and preprocessing, we will ensure that our framework can be readily applied to a wide range of datasets, which can enable researchers to leverage our methods for studying epistatic interactions in various biological contexts, from model organisms to agricultural crops and human diseases.

3.2 Code Availability

The initial version of the code for implementing the proposed kernel-based framework for detecting epistatic interactions in QTL data is now already publicly available on GitHub. The repository contain the complete source code, along with detailed documentation, examples, and tutorials to facilitate the adoption and extension of our methodology by the wider scientific community.

In addition, we have package our software as a Python library, called `biokite`, and distribute it through the Python Package Index (PyPI), which allow users to easily install and use our framework in their own Python environments using standard package management tools, such as pip.

4 References

- [1] Mackay et al. (2009) The genetics of quantitative traits: challenges and prospects. Nature Reviews Genetics.
- [2] Broman & Sen (2009) A Guide to QTL Mapping with R/qtl. Springer.
- [3] Cordell (2002) Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. Human Molecular Genetics.
- [4] Lehner (2011) Molecular mechanisms of epistasis within and between genes. Trends in Genetics.
- [5] Forsberg et al. (2017) Accounting for genetic interactions improves modeling of individual quantitative trait phenotypes in yeast. Nature Genetics.
- [6] Xu (2003) Estimating polygenic effects using markers of the entire genome. Genetics.
- [7] Wei et al. (2014) Detecting epistasis in genome-wide association studies. Nature Methods.
- [8] Marchini et al. (2005) Genome-wide strategies for detecting multiple loci that influence complex diseases. Nature Genetics.
- [9] Lachowiec et al. (2015) A Genome-Wide Association Analysis Reveals Epistatic Cancellation of Additive Genetic Variance for Root Length in Arabidopsis thaliana. Nature Methods.
- [10] Zhang & Liu (2007) Bayesian inference of epistatic interactions in case-control studies. Nature Genetics.
- [11] Moore et al. (2010) Epistasis and its implications for personal genetics. American Journal of Human Genetics.
- [12] Boucher & Jenna. (2013) Genetic interaction networks: better understand to better predict. Frontiers in Genetics.
- [13] Jiang & Reif (2015) Modeling epistasis in genomic selection. Genetics.
- [14] Tian et al. (2011) Genome-wide association study of leaf architecture in the maize nested association mapping population. Nature Genetics.
- [15] Yu et al. (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nature Genetics.
- [16] Scholkopf et al. (2004) Kernel Methods in Computational Biology. MIT Press.
- [17] Spindel et al. (2016) Genome-wide prediction models that incorporate de novo GWAS are a powerful new tool for tropical rice improvement. Heredity.
- [18] Hastie et al. (2009) The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.
- [19] Tenenhaus & Tenenhaus (2011) Regularized generalized canonical correlation analysis. Psychometrika.
- [20] Schupbach et al. (2010) FastEpistasis: a high performance computing solution for quantitative trait epistasis. Bioinformatics.
- [21] Storey & Tibshirani (2003) Statistical significance for genomewide studies. PNAS.

- [22] Zhu et al. (2012) Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. Nature Genetics.
- [23] Bloom et al. (2013) Finding the sources of missing heritability in a yeast cross. Nature.
- [24] Van der Walt et al. (2011) The NumPy array: a structure for efficient numerical computation. Computing in Science & Engineering.
- [25] Pedregosa et al. (2011) Scikit-learn: Machine learning in Python. Journal of Machine Learning Research.