SATlotyper

Haplotype inference from unphased SNP data in heterozygous polyploids based on the SAT algorithm

Jost Neigenfind[1,2], Gabor Gyetvai[3], Rico Basekow[1,2], Svenja Diehl[2], Ute Achenbach[3], Christiane Gebhardt[3], Joachim Selbig[4], Birgit Kersten[1,2]
Contact: Neigenfind@mpimp-golm.mpg.de

[1]Bioinformatics, Max Planck Institute of Molecular Plant Physiology, Am Mühlenberg 1, 14424 Potsdam-Golm, Germany
[2]Bioinformatics, Former RZPD German Resource Center for Genome Research GmbH, Heubnerweg 6, 14059 Berlin, Germany
[3]Max Planck Institute for Plant Breeding Research, Carl von Linné Weg 10, 50829 Köln, Germany
[4]Institute of Biochemistry and Biology, University of Potsdam, c/o MPI-MP, Am Mühlenberg 1, 14424 Potsdam, Germany

Introduction

Haplotype inference (see Figure 1 for an example) based on unphased SNP (single nucleotide polymorphism) markers is an important task in population genetics.
In the case of homozygous genotypes, like maize or many other inbreeding crop species, haplotypes can be directly drawn from comparing the amplified genomic sequence at a given locus between different individuals. Difficulties arise if homozygous genotypes are not available, as for example in non-inbred, tetraploid potato *(Solanum tuberosum).* In such cases, it is necessary to determine the haplotype phase from unphased SNP data.
There are several approaches to infer haplotypes. However, these approaches have been developed for bi-allelic and diploid species and there is currently no software available for haplotype identification in more complex autotetraploids.
In this study, we aimed at the development and evaluation of a generalised approach for calculating haplotypes in polyploid species using the parsimonious principle. The goal of haplotype inference is to find a set of haplotypes explaining every genotype present in a given unphased population. The parsimonious principle consists of finding the smallest set of haplotypes such that each genotype in the population can be explained by a ploidy specific number of haplotypes from the set of haplotypes.

Results

Here we present the generalisation of the SAT (Boolean satisfyability problem of propositional logic) approach from Lynce et al. resulting in the development of the SATlotyper software tool which is able to handle polyploid and poly-allelic data. The program is able to exclude already found inferences, such that alternative inferences can be calculated. As it is not known which of the multiple inferences are best supported by the given unphased data set, we use a bootstrapping procedure that allows for scoring alternative inferences. Finally, by means of the bootstrapping scores it is possible to optimise the phased genotypes belonging to a given haplotype inference.
The program was evaluated with simulated (Figure 2) and experimental (Figure 3) SNP data generated in heterozygous tetraploid populations of potato. We showed with simulated data that, instead of taking the first inference as reported by the program, the quality of the final result can be significantly improved by the application of additional methods that include scoring of the alternative haplotype inferences and genotype optimisation. For a population of 19 individuals, the predicted results, computed by SATlotyper, were directly compared to results obtained by experimental haplotyping via sequencing of cloned amplicons. Prediction and experiment gave similar results regarding the inferred haplotypes and phased genotypes (Figure 3 and Figure 4).

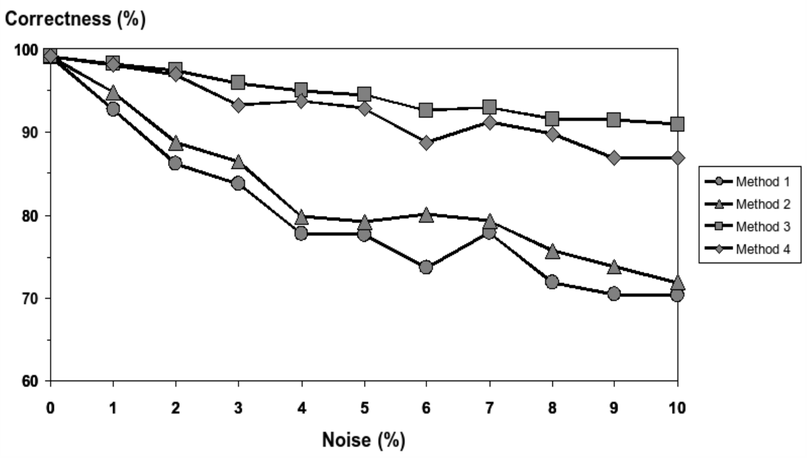

Figure 2 - Comparison of four types of SATlotyper analyses using simulated data sets For 10 tetraploid, bi-allelic populations with 100 individuals each, haplotypes comprising 6 SNPs were simulated. 0-10% noise was added to the data. Unphased simulated data sets were analysed with SATlotyper using 4 different methods: 1) first inference without optimisation; 2) alternative haplotype inferences scored by bootstrapping, selection of the best scored; 3) alternative haplotype inferences scored by bootstrapping, further optimisation of genotype inference; 4) first haplotype inference with optimisation of genotype inference. Based on the Hamming distance between predicted and original simulated genotype inferences, the correctness was calculated, normalised and plotted against the noise. Every data point represents the mean value of 10 populations for the respective value of noise.

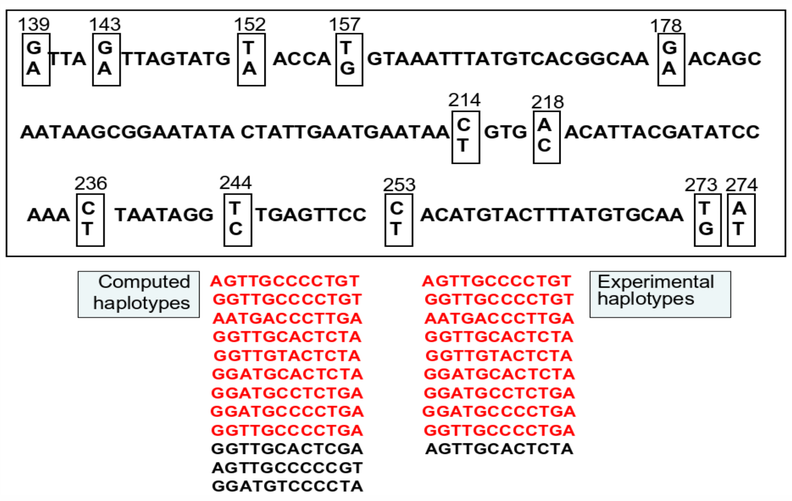

Figure 3 - Comparison of computational and experimental haplotypes for potato locus *BA213c14t7* Sequence of the potato locus *BA213c14t7*. The evaluated SNP positions are indicated by boxes. For a sub-population of 19 individuals, haplotype sequences were identified computationally using Method 2 and experimentally by amplicon cloning and sequencing. The 9 haplotpyes identified by both methods are displayed in red.

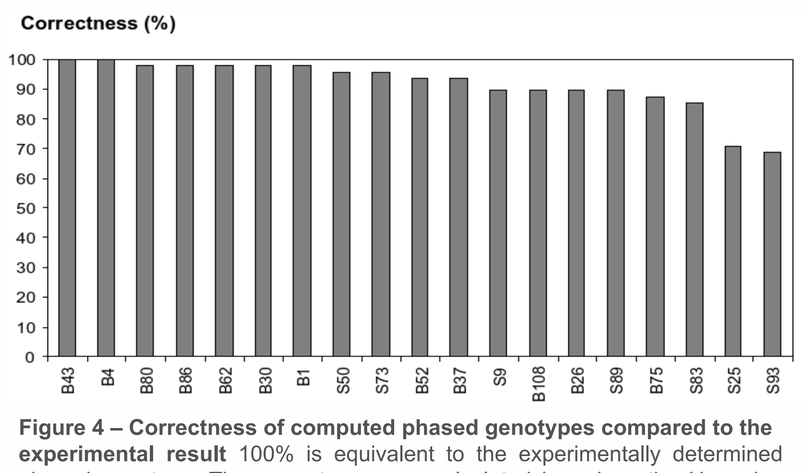

Figure 4 – Correctness of computed phased genotypes compared to the experimental result 100% is equivalent to the experimentally determined phased genotype. The correctness was calculated based on the Hamming distance between experimental and computational phased genotype.

Discussion

Using a parsimonious principle, a Java based program, named SATlotyper, was developed in this study which formulates the “Haplotype Inference by Pure Parsimony” (HIPP) for heterozygous polyploid species using the SAT approach.
SATlotyper is able to handle missing SNP information by dropping constraints for such positions so that no unjustified assumptions about nucleotide frequencies have to be made. For a given data set of unphased genotypes, by means of SATlotyper, it is possible to infere the data by four different methods with increasing accuracy (Figure 2).
In this study, the SATlotyper was tested and evaluated with simulated and experimental data sets of unphased SNPs from tetraploid individuals (Figure 2 and Figure 3). The results obtained when applying Method 4 (Figure 2) suggest that for some purposes, it could be sufficient to simply score the explaining haplotypes with their frequencies in the phased genotypes for optimising genotype inference. This omits time-consuming bootstrapping. SATlotyper was also applied to an experimental data set of 12 unphased SNP markers, which were scored by sequencing the amplicons of a 500 bp-region at potato locus BA213c14t7. The performance of the approach was much higher on this data than on simulated data.
In addition to performance, the quality of the prediction was tested on a subset of 19 heterozygous experimentally verified individuals. This first comparison between computed and experimental verified haplotypes gave promising results: 9 of the 10 experimental haplotypes were also identified by SATlotyper prediction (Figure 3). With respect to the phased genotypes, the SATlotyper analysis achieved a correctness of at least 90% (for 80% of the individuals) compared to the experimental result (Figure 4).

Conclusion

The study demonstrates that HIPP can efficiently be solved by the SAT approach for data sets of unphased SNPs from heterozygous polyploids. This outcome is encouraging for the future application and further development of SATlotyper. Existing or newly generated unphased SNP data can be analysed by SATlotyper to infer haplotypes. Haplotype information can be used instead of individual SNPs in association mapping that exploits the biodiversity in existing cultivars and breeding lines. Compared with methods based on individual SNPs, the haplotype mapping method significantly improves the power and robustness of association mapping techniques as there are fewer haplotypes than SNPs. SATlotyper is freeware and is distributed as a Java JAR file. The program will be available for download via the SATlotyper project page of GabiPD, The GABI Primary Database.

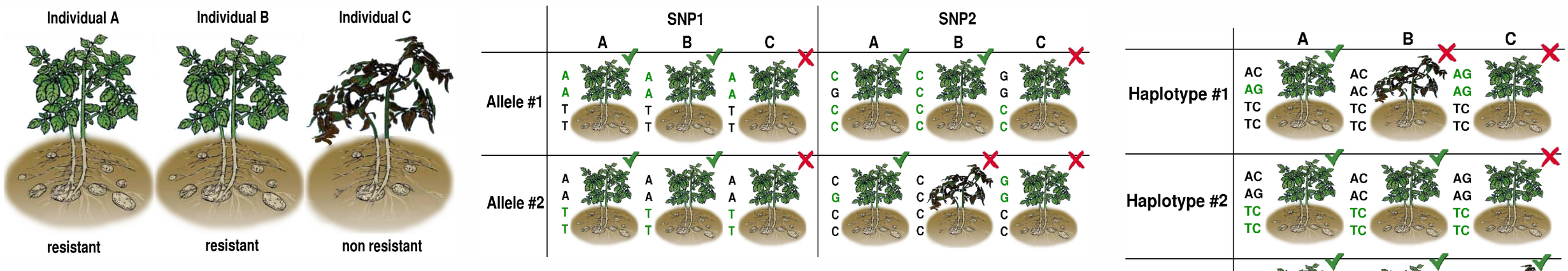

Figure 1: An example of three potato individuals (A, B and C) where C is not resistant against a disease (left). In this example, the responsible locus is biallelic and contains two SNPs (middle and right). If SNPs are analyzed independently from each other, all possible predictable phenotypes do not fit reality (middle). In contrast, if the correct haplotype composition is known, only haplotype #3 (“AC”) can explain the observed phenotypes correctly (right).

References

Neigenfind J, Gyetvai G, Basekow R, Diehl S, Achenbach U, Gebhardt C, Selbig J, Kersten B: Haplotype inference from unphased SNP data in heterozygous polyploids based on the SAT algorithm; submitted; 2008

Sattarzadeh A, Achenbach U, Lübeck J, Strahwald J, Tacke E, Hofferbert H- R, Rothsteyn T, Gebhardt C: Single nucleotide polymorphism (SNP) genotyping as basis for developing a PCR-based marker highly diagnostic for potato varieties with high resistance to Globodera pallida pathotype Pa2/3. Molecular Breeding 2006, 18:301–312

Lynce I, Marques-Silva J: Efficient haplotype inference with boolean satisfiability. Lisbon, Portugal: Technical University of Lisbon; 2006.