Haplotype-based prediction of gene alleles using pedigrees and SNP genotypes

Yuri Pirola Gianluca Della Vedova Paola Bonizzoni
Alessandra Stella Filippo Biscarini

DISCo, Univ. degli Studi di Milano-Bicocca, Milan, Italy
yuri.pirola@disco.unimib.it

ACM-BCB 2013
ACM Conference on Bioinformatics, Computational Biology and Biomedical Informatics

Washington DC, September 22–25, 2013

Outline

1. The computational problem
2. Our approach
   - Description of the method
   - Experimental evaluation
3. Conclusions and future works

The computational problem

Substituting expensive and specific assays for **typing gene alleles** with (cheaper) computational predictions based on routinely collected genetic data (SNP genotypes).

$$\textbf{Gene alleles} \;= \text{different forms of a gene (protein)}$$
$$\neq \text{SNV/indels/SV/CNV/}\ldots$$

Applications

Gene allele prediction/typing applications:

- *HLA typing* (humans)
  evaluation of organ transplantation compatibility
- *Productive characters in animals/plants*
  casein genes in cows/goats/sheep
  → milk yield and cheese quality

Most approaches in literature refer to HLA typing!

State of the art

HLA-IBD

(Setty *et al.*, JCB, 2011)

- combinatorial graph-based approach
- needs accurate prediction of IBD regions
- does not exploit pedigree information

MAG

(Li *et al.*, Genetic Epidemiology, 2011)

- statistical approach
- good results in various settings

  (Zhang *et al.*, BMC Genetics, 2011), (Ayele *et al.*, PLoS One, 2012)
- does not exploit pedigree information

State of the art

WSG-HI (Xie *et al.*, BMC Bioinf, 2010)

- combinatorial approach (on pedigrees)
- the pedigree must contain the individuals which the gene alleles are sought for

Our method

Our method:

1. Gene-aware phasing of SNP genotypes
2. Minimum-error association computation

} on a *training* population T

3. Allele prediction

} on population U

1 – Gene-aware phasing

Step 1. **Gene-aware phasing**

Input: a pedigree of a training population T,
SNP genotypes and gene alleles for some individuals

Output: a min-recombinant haplotype configuration for T “compatible with” the observed gene alleles

Why? gene alleles are inherited together with haplotypes

How?

- “Enrich” the SNP genotypes by inserting the given gene alleles according to the genetic map
- Phase the “enriched” genotypes

1 – Gene-aware phasing

Step 1. **Gene-aware phasing**

How?

- "Enrich" the SNP genotypes by inserting the given gene alleles according to the genetic map
- Phase the "enriched" genotypes

We need a phasing/HI method that takes into account:

- genotyping errors
- recombinations
- missing genotypes
- multiallelic loci

and that works on (potentially large) pedigrees!

→ reHCstar2: extension of (Pirola *et al.*, TCBB, 2012)

2 – Minimum error association computation

Step 2. **Minimum error association computation**

Input: a haplotype configuration for T, the observed gene alleles

Output: a set M of weighted associations:

$$(\textit{haplotype} \overset{w}{\longmapsto} \textit{gene allele})$$

Why? same haplotype $\Longrightarrow$ same gene allele (likely)
(*but not the converse!!*)

How? minimize association errors
(error $=$ same haplotype, different gene alleles)

→ Integer Linear Programming formulation

→ weight w is the relative freq. of the haplotype in T

3 – Allele prediction

Step 3. **Allele prediction**

Input: a set M of weighted associations $($*hapl.* $\xrightarrow{w}$ *gene allele*$)$,
SNP genotypes of a population U

Output: gene alleles of individuals in U

How? "majority voting"-like

the predicted gene alleles are the ones associated to pair of haplotypes "compatible" with the SNP genotype

Everything is weighted by w and α^d
$(w =$ association weight, α constant $\in (0, 1]$,
$d =$ Hamming distance)

Remark: parental relationships between T and U are not used.

Experimental evaluation

Application: Prediction of κ-casein on dairy cattle

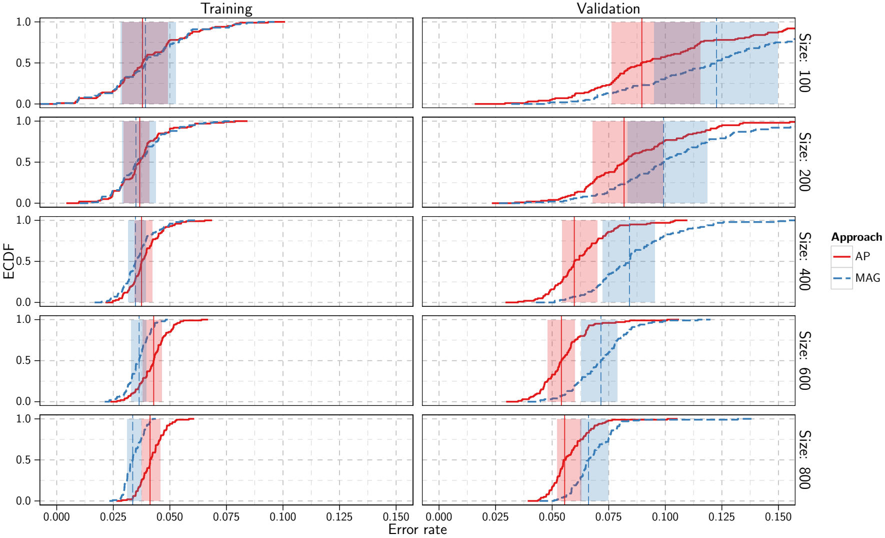
Data: pedigree of $>$100K cattle,
$\approx$20K with κ-casein alleles
$\approx$2.2K with SNP genotypes
$\approx$1.6K with both
(courtesy of Italian Brown Cattle Breeders' Association)

Strategy: cross-validation

(random disjoint subsets T and U, 100 replicates)

Metrics: (prediction) error rate $= \frac{\text{wrong predictions}}{\text{tot. predictions}}$

1 – Comparison with MAG (Li *et al.*, Genetic Epid., 2011)



2 – Sensitivity to the choice of the genomic region

Distribution of error rate using different genomic regions (centered at the κ-casein gene) for the prediction.

	AP				MAG			
Genomic region	1st quart.	Median	3rd quart.	Mean	1st quart.	Median	3rd quart.	Mean
[-500k; +500k]	0.057	0.067	0.075	0.068	0.073	**0.080**	**0.092**	**0.084**
[-500k; +750k]	0.063	0.072	0.084	0.076	0.076	0.084	0.094	0.087
[-500k; +1000k]	0.060	0.067	0.080	0.071	0.075	0.083	0.096	0.087
[-750k; +500k]	0.059	0.069	0.077	0.071	0.073	0.082	0.093	0.085
[-750k; +750k]	0.067	0.075	0.086	0.078	0.075	0.085	0.097	0.088
[-750k; +1000k]	0.063	0.070	0.079	0.073	0.075	0.085	0.099	0.087
[-1000k; +500k]	**0.054**	**0.060**	**0.070**	**0.063**	**0.072**	0.084	0.095	0.085
[-1000k; +750k]	0.055	0.063	0.077	0.068	0.075	0.085	0.101	0.088
[-1000k; +1000k]	0.055	0.064	0.078	0.067	0.075	0.086	0.100	0.087

Other experimental results

Part 3 – **Sensitivity to the choice of α**

- good results for $\alpha \leq 0.1$ (median *err. rate* $\leq 6.3\%$)
- worsen for $\alpha > 0.1$ (median *err. rate* $\geq 7.5\%$)

Part 4 – **Sensitivity to the degree of relationships**

Prediction error rate does not apparently correlate with the degree of relationships (*kinship*) among T and U

(but further investigation is needed!)

Conclusions

Gene-allele prediction from SNP genotypes has relevant applications (health/economic/. . .)

Main characteristics of our approach:

- two well-distinguished training/prediction phases
- pedigree and gene-aware phasing to improve HI accuracy
- open source (`http://allele-prediction.algolab.eu`)
  (improvements are underway: ease of use, documentation, . . .)

Ongoing/planned works

Ongoing/planned works:

- more extensive experimental comparison
- exploiting relationships among T and U (if known)
- multi-population predictions

Acknowledgments:
We wish to thank Dr. Santus and Dr. Rossoni of the Italian Brown Cattle Breeders' Association (ANARB) for the data used in the experimental evaluation.

Thanks!

Questions?

Yuri Pirola – DISCo, Univ. of Milano-Bicocca (Milan, Italy)

E-Mail: `yuri.pirola@disco.unimib.it`

Additional Content

State of the art

HLA-IBD

(Setty *et al.*, JCB, 2011)

- combinatorial graph-based approach
- pre-phased SNP genotypes (haplotypes)
  → identical-by-descent regions
  → HLA gene alleles
- **Cons:**
  - needs accurate prediction of IBD regions
  - does not exploit pedigree information

State of the art

MAG (Li *et al.*, Genetic Epidemiology, 2011)

- statistical approach
- unphased SNP genotypes
  → haplotype frequencies
  → HLA gene alleles
- good results in various settings

  (Zhang *et al.*, BMC Genetics, 2011), (Ayele *et al.*, PLoS One, 2012)
- **Cons:**
  - does not exploit pedigree information

State of the art

WSG-HI

(Xie *et al.*, BMC Bioinf, 2010)

- combinatorial approach (on pedigrees)
- unphased SNP genotypes
  → haplotype configuration space
  → max-similarity labelling with HLA gene alleles
- **Cons:**
  - the pedigree must contain the individuals which the gene alleles are sought for

3 – Sensitivity to the choice of α

Distribution of error rate using different value for parameter α.

	Training				Validation			
α	1st quart.	Median	3rd quart.	Mean	1st quart.	Median	3rd quart.	Mean
0.005	**0.034**	**0.037**	**0.042**	0.039	**0.054**	0.062	0.072	0.065
0.010	**0.034**	0.037	**0.042**	**0.038**	0.054	0.063	0.072	0.065
0.025	0.034	**0.037**	**0.042**	0.039	0.054	0.062	0.072	0.064
0.050	0.035	0.037	**0.042**	0.039	0.054	**0.060**	0.070	0.063
0.100	0.037	0.041	0.045	0.042	0.055	0.063	**0.069**	**0.063**
0.250	0.062	0.067	0.075	0.068	0.067	0.075	0.083	0.076
0.500	0.107	0.115	0.126	0.117	0.121	0.136	0.149	0.134

(AP only)