Yet another "GWAS + ML" project

Vivien Goepp

March, 16 2020

CBIO Meeting

1. Short reminder on GWAS

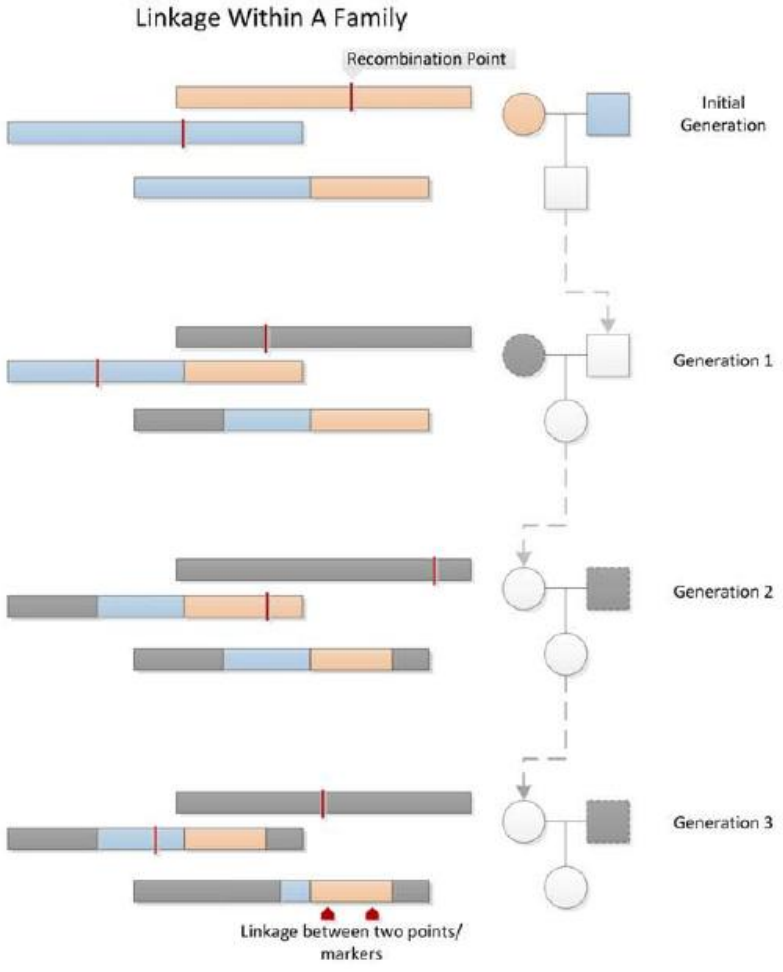
The CD/CV hypothesis

- *Common diseases* are caused by *common variants*
- Common diseases : not caused by Mendelian mutations
- Common variants : need to consider many mutations *jointly*
- Each mutation can only have a *small effect*

Linkage disequilibrium

Linkage disequilibrium (LD) :

- For geneticists :

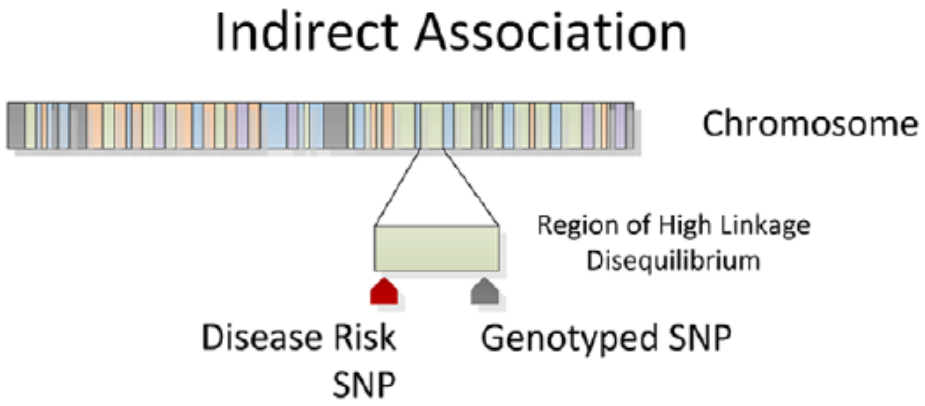


Source : Bush et al (2012)

Linkage disequilibrium

Linkage disequilibrium (LD) :

- For statisticians : correlations between close-by alleles on the genome



Source : Bush et al (2012)

Linkage disequilibrium

Idea of using single nucleotide polymorphisms (SNPs) :

- There are many high-LD blocks on the genome
- We can use SNPs as markers of an LD block
- The disease-inducing mutations are observed through correlation

Goal of GWAS

- **Goal** : Discover gene mutations *linked* to a disease.
- GWAS will not provide : causality relations or biological pathways leading to a disease.
- Applications to common diseases :
  - Inflammatory Bowel Diseases
  - Auto-immune diseases
  - Metabolic diseases (T2 diabetes, obesity, BMI)
  - Multiple sclerosis
  - Cancer

GWAS

Genome-wide association studies (GWAS)

- Gather $n \sim 10^3$ individuals
- Observe the phenotype :
  - quantitative (BMI, cholesterol, height)
  - or qualitative (case-control for common disease).
- Observe the genotype of $p \sim 10^6$ SNPs

GWAS

Genome-wide association studies (GWAS)

- Gather $n \sim 10^3$ individuals
- Observe the phenotype :
  - quantitative (BMI, cholesterol, height)
  - or qualitative (case-control for common disease).
- Observe the genotype of $p \sim 10^6$ SNPs

Encoding of SNP alleles

Each SNP has value $\in \{\text{aa}, \text{aA}, \text{AA}\}$.

Genotype	Allelic dosage	Dominant	Recessive	Dummy	One-hot
aa	0	0	0	00	100
aA	1	1	0	01	010
AA	2	1	1	11	001

We will use one-hot encoding.

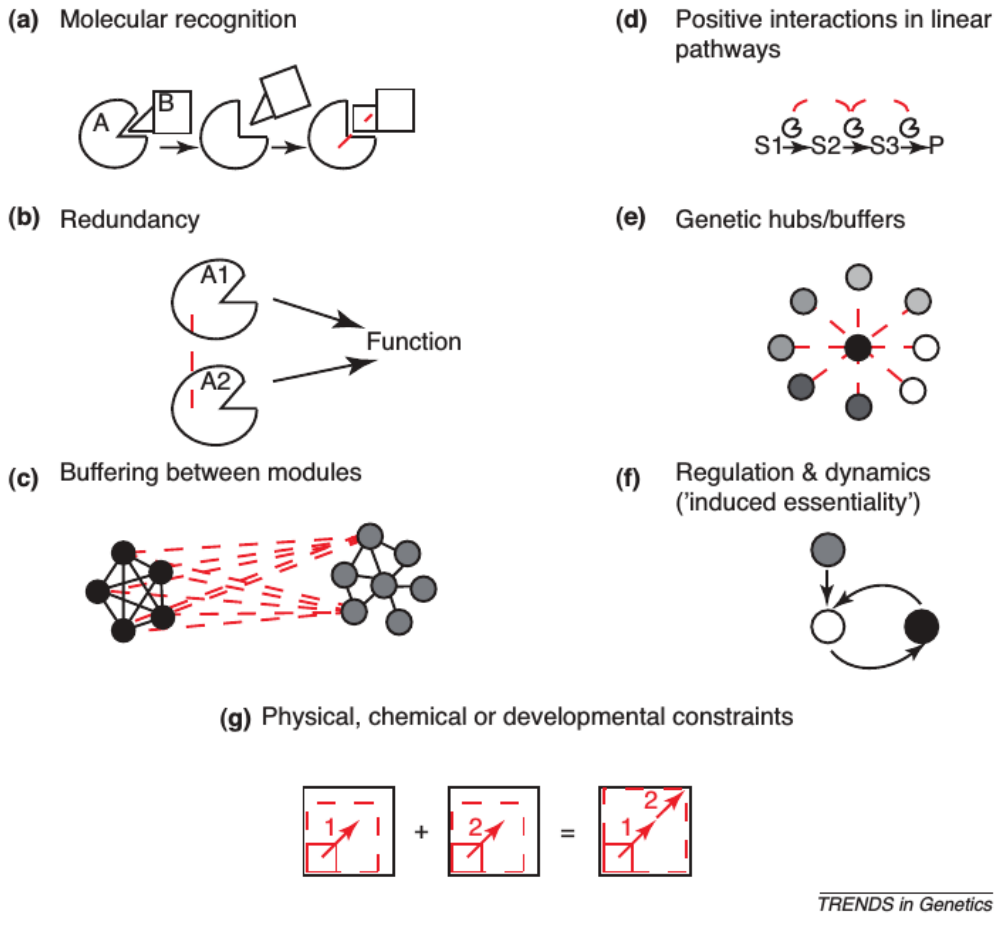
Epistasis and interactions between SNPs

What is epistasis ?

- "The masking of the effects of one variant by another" (Bateson 1909).
- Broad sense : "The dependence of the mutation outcome on the genetic background" (Lehner 2011).

Different biological explanasions for epistasis

Several molecular mechanisms thought to be linked to epistasis between genes :



Source : Lehner (2011)

The statistical problem

We do feature selection with :

- $\frac{p}{n} \sim 1000$

The statistical problem

We do feature selection with :

- $\frac{p}{n} \sim 1000$
- Correlation between features

The statistical problem

We do feature selection with :

- $\frac{p}{n} \sim 1000$
- Correlation between features
- Interaction between features

The statistical problem

We do feature selection with :

- $\frac{p}{n} \sim 1000$
- Correlation between features
- Interaction between features
- Bonus : population structure : mixture between different haplotype

2. Set covering machines for SNPs discovery

Setting of the set covering machines (SCM)

- $y_i \in \{0, 1\}$ (binary classification case/control GWAS)
- $x_{i,j} \in \{0, 1\}$ (binary features with one-hot encoding)
- $p \gg n$ with true model assumed to be *very sparse*

Goal : find a prediction rule of the form :

$$f(\mathbf{x}) = \bigwedge_{h \in \mathcal{R}} h(\mathbf{x}),$$

where $\mathcal{R}$ is a set of rules r.
Here a rule is the absence or presence of SNP i.

The set covering problem

Haussler algorithm :

- consider only rules that perfectly classify the positive examples.
- find the smallest number of rules that cover the negative examples.

	$\mathbf{y}$	h_1	h_2	h_3	h_4
$\mathcal{N}$	0	0	0	1	1
	0	1	0	0	1
	0	1	1	0	0
	0	1	1	1	0
$\mathcal{P}$	1	1	1	1	1
	1	1	1	1	1

Set covering machine

We choose the rule with maximum usefulness :

$$U_h = |\mathcal{A}_h| - q|\mathcal{B}_h|,$$

$\mathcal{A}_h$: negative examples correctly classified

$\mathcal{B}_h$: positive examples uncorrectly classified

Set covering machine

We choose the rule with maximum usefulness :

$$U_h = |\mathcal{A}_h| - q|\mathcal{B}_h|,$$

$\mathcal{A}_h$: negative examples correctly classified

$\mathcal{B}_h$: positive examples uncorrectly classified

- We allow errors on positive examples
- q tunes this error

Set covering machine : example

		SNP1	SNP2	SNP3	SNP4
	$\mathbf{y}$	h_1	h_2	h_3	h_4
$\mathcal{N}$	0	0	0	1	1
	0	1	0	0	1
	0	1	1	0	0
	0	1	1	1	0
$\mathcal{P}$	1	1	0	1	1
	1	1	0	0	1

Set covering machine : example

		SNP1	SNP2	SNP3	SNP4
	$\mathbf{y}$	h_1	h_2	h_3	h_4
$\mathcal{N}$	0	0	0	1	1
	0	1	0	0	1
	0	1	1	0	0
	0	1	1	1	0
$\mathcal{P}$	1	1	0	1	1
	1	1	0	0	1
	$\lvert\mathcal{A}_h\rvert$	1	2	2	2

Set covering machine : example

		SNP1	SNP2	SNP3	SNP4
	$\mathbf{y}$	h_1	h_2	h_3	h_4
$\mathcal{N}$	0	0	0	1	1
	0	1	0	0	1
	0	1	1	0	0
	0	1	1	1	0
$\mathcal{P}$	1	1	0	1	1
	1	1	0	0	1
	$\lvert\mathcal{A}_h\rvert$	1	2	2	2
	$\lvert\mathcal{B}_h\rvert$	0	2	1	0

Set covering machine : example

		SNP1	SNP2	SNP3	SNP4
	$\mathbf{y}$	h_1	h_2	h_3	h_4
$\mathcal{N}$	0	0	0	1	1
	0	1	0	0	1
	0	1	1	0	0
	0	1	1	1	0
$\mathcal{P}$	1	1	0	1	1
	1	1	0	0	1
	$\lvert\mathcal{A}_h\rvert$	1	2	2	2
	$\lvert\mathcal{B}_h\rvert$	0	2	1	0
	$\mathcal{U}_h$	1	$2-2q$	$2-q$	2

Set covering machine : example

		SNP1	SNP2	SNP3	SNP4
	$\mathbf{y}$	h_1	h_2	h_3	h_4
$\mathcal{N}$	0	0	0	1	1
	0	1	0	0	1
	0	1	1	0	0
	0	1	1	1	0
$\mathcal{P}$	1	1	0	1	1
	1	1	0	0	1
	$\lvert\mathcal{A}_h\rvert$	1	2	2	2
	$\lvert\mathcal{B}_h\rvert$	0	2	1	0
	$\mathcal{U}_h$	1	$2-2q$	$2-q$	2

- $\mathcal{R} \leftarrow \{h_4\}$

Set covering machine : example

		SNP1	SNP2	SNP3	SNP4
	$\mathbf{y}$	h_1	h_2	h_3	h_4
$\mathcal{N}$	0	0	0	1	1
	0	1	0	0	1
	0	1	1	0	0
	0	1	1	1	0
$\mathcal{P}$	1	1	0	1	1
	1	1	0	0	1
	$\lvert\mathcal{A}_h\rvert$	1	2	2	2
	$\lvert\mathcal{B}_h\rvert$	0	2	1	0
	$\mathcal{U}_h$	1	$2-2q$	$2-q$	2

- $\mathcal{R} \leftarrow \{h_4\}$
- $\mathcal{N} \leftarrow \mathcal{N} \setminus \mathcal{A}_{h_4}$

Set covering machine : example

		SNP1	SNP2	SNP3	SNP4
	$\mathbf{y}$	h_1	h_2	h_3	h_4
$\mathcal{N}$	0	0	0	1	1
	0	1	0	0	1
	0	1	1	0	0
	0	1	1	1	0
$\mathcal{P}$	1	1	0	1	1
	1	1	0	0	1
	$\lvert\mathcal{A}_h\rvert$	1	2	2	2
	$\lvert\mathcal{B}_h\rvert$	0	2	1	0
	$\mathcal{U}_h$	1	$2-2q$	$2-q$	2

- $\mathcal{R} \leftarrow \{h_4\}$
- $\mathcal{N} \leftarrow \mathcal{N} \setminus \mathcal{A}_{h_4}$
- $\mathcal{P} \leftarrow \mathcal{P} \setminus \mathcal{B}_{h_4}$

Set covering machine : example (2)

	$\mathbf{y}$	h_1	h_2	h_3
$\mathcal{N}$	0	0	0	1
	0	1	0	0
$\mathcal{P}$	1	1	0	1
	1	1	0	0

Set covering machine : example (2)

	$\mathbf{y}$	h_1	h_2	h_3
$\mathcal{N}$	0	0	0	1
	0	1	0	0
$\mathcal{P}$	1	1	0	1
	1	1	0	0
	$\lvert\mathcal{A}_h\rvert$	1	2	1

Set covering machine : example (2)

	$\mathbf{y}$	h_1	h_2	h_3
$\mathcal{N}$	0	0	0	1
	0	1	0	0
$\mathcal{P}$	1	1	0	1
	1	1	0	0
	$\lvert\mathcal{A}_h\rvert$	1	2	1
	$\lvert\mathcal{B}_h\rvert$	0	2	1

Set covering machine : example (2)

	$\mathbf{y}$	h_1	h_2	h_3
$\mathcal{N}$	0	0	0	1
	0	1	0	0
$\mathcal{P}$	1	1	0	1
	1	1	0	0
	$\lvert\mathcal{A}_h\rvert$	1	2	1
	$\lvert\mathcal{B}_h\rvert$	0	2	1
	$\mathcal{U}_h$	1	$2-2q$	$2-q$

Stopping criteria

- We have finished the job : $\mathcal{N} = \emptyset$
- Early stopping : $|\mathcal{R}| \geq s$ with parameter $s \geq 1$
- There remains only useless rules : $|\mathcal{A}_h| = |\mathcal{B}_h| = 0$

Computation time

- Each greedy step is fast to compute :
  - Let $\mathcal{I}_{\mathcal{N}}$ be the (current) indices of the negative examples.
  - $|\mathcal{A}_h| = |\mathcal{I}_{\mathcal{N}}| - \sum_{i \in \mathcal{I}_{\mathcal{N}}} x_{i,j}$ if h is the presence rule of feature j
  - $|\mathcal{A}_h| = \sum_{i \in \mathcal{I}_{\mathcal{N}}} x_{i,j}$ if h is the absence rule of feature j
  - Similar for $|\mathcal{B}_h|$.

Computation time

- Each greedy step is fast to compute :
  - Let $\mathcal{I}_{\mathcal{N}}$ be the (current) indices of the negative examples.
  - $|\mathcal{A}_h| = |\mathcal{I}_{\mathcal{N}}| - \sum_{i \in \mathcal{I}_{\mathcal{N}}} x_{i,j}$ if h is the presence rule of feature j
  - $|\mathcal{A}_h| = \sum_{i \in \mathcal{I}_{\mathcal{N}}} x_{i,j}$ if h is the absence rule of feature j
  - Similar for $|\mathcal{B}_h|$.

- Overall complexity $\mathcal{O}\left(|\mathcal{R}|ns\right)$

Computation time

- Each greedy step is fast to compute :
  - Let $\mathcal{I}_{\mathcal{N}}$ be the (current) indices of the negative examples.
  - $|\mathcal{A}_h| = |\mathcal{I}_{\mathcal{N}}| - \sum_{i \in \mathcal{I}_{\mathcal{N}}} x_{i,j}$ if h is the presence rule of feature j
  - $|\mathcal{A}_h| = \sum_{i \in \mathcal{I}_{\mathcal{N}}} x_{i,j}$ if h is the absence rule of feature j
  - Similar for $|\mathcal{B}_h|$.
- Overall complexity $\mathcal{O}(|\mathcal{R}|ns)$
- Limited memory usage

Upper bound on the risk (1/3)

- SCM is a *sample compression algorithm*

- Given a model h chosen by an SCM, there exist
  - a set of individuals $\mathcal{Z} \in \{1, \cdots, n\}$
  - a message string σ containing additional information,

  such that

$$h = \mathbf{\Phi}(\mathcal{Z}, \sigma) :$$

  h can be reconstructed from $\mathcal{Z}$.

Upper bound on the risk (2/3)

Then there exists a bound on the risk

$$R(h) = \mathbb{E}_{(\mathbf{x},y)\sim D}\left[\mathbb{1}_{f(\mathbf{x})\neq y}\right]$$

that depends on the size of $\mathcal{Z}$.

Models that can be compressed using few examples have good generalization.

Upper bound on the risk (3/3)

Marchand and Sokolova (2005) established that :

$$\mathbb{P}\left(\forall S \sim D, \forall h, R(h) \leq \varepsilon(h, S, \delta)\right) \geq 1 - \delta$$

Upper bound on the risk (3/3)

Marchand and Sokolova (2005) established that :

$$\mathbb{P}\left(\forall S \sim D, \forall h, R(h) \leq \varepsilon(h, S, \delta)\right) \geq 1 - \delta$$

with

$$\varepsilon(h, S, \delta) = 1 - \exp\left(\frac{-1}{n - |\mathcal{Z}| - r}\left[\log\binom{m}{|\mathcal{Z}|} + \log\binom{m - |\mathcal{Z}|}{r} + |h| \log(2\mathcal{N}(\mathcal{Z})) + \log\Omega\right]\right)$$

with $\Omega = \frac{\pi^6(|h|+1)^2(r+1)^2(|\mathcal{Z}|+1)^2}{216\delta}$,
where r is the number of classif. errors on $S \setminus \mathcal{Z}$.

Sample compression bound

Consequences :

- The bound does not depend on p : theoretical performance guarantee
- It can be used for hyperparameter selection (Marchand et Shawe-Taylor, 2002).

Conclusion

- Set covering machine
  - runs fast
  - does not suffer from $p \gg n$
- However there are other issues :
  - Many SNPs are equivalent
  - Only conjunctions of SNPs

Thank you

References

- SCM : Marchand, M. and Shawe-Taylor, J. , *The set covering machine*, JMLR, 2002
- Risk bound for SCM : Marchand, M and Sokolova, M., *Learning with Decision Lists of Data-Dependent Features*, JMLR, 2005
- GWAS : Bush, W, Moore, J., Lewitter, F., and Kann, M, *Chapter 11 : Genome-Wide Association Studies*, Plos Comp. Biol., 2012
- Epistasis : Lehner, B., *Molecular mechanisms of epistasis within and between genes*, Trends in Gen., 2011