genome-wide association studies

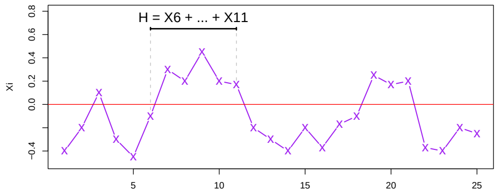
Mickaël Guedj,
D. Robelin, M. Hoebeke, M. Lamarine, K. Forner, J. Wojcik and G. Nuel
guedj@genopole.cnrs.fr

Genetic epidemiology aims at identifying biological mechanisms responsible for human diseases. Genome-wide association studies are now promisingly investigated. In these studies, commonly used strategies focus on marginal effects. Such approaches lead to multiple-testing and are unable to capture the possibly complex interplay between genetic factors. We have adapted to association studies the use of the **local score statistic**, a natural improvement of sliding-frames. Via sums statistics, this strategy combines **local** (Linkage Disequilibrium) and possibly **distant** dependences between markers. It is **fast** to compute, able to handle very **large datasets**, circumvents the **multiple-testing** problem and outlines a set of **genomic regions** (segments) possibly interesting for further analyses. Applied to real data, our approach outperforms classical Bonferroni and FDR corrections. It is implemented in a programm termed LHiSA for Local High-scoring Segments for Association and available at: `http://stat.genopole.cnrs.fr/weblhisa`

Methods

Definition

Let $\mathbb{X} = (X_i)_{i=1,\ldots,n}$ a sequence of real random variables:



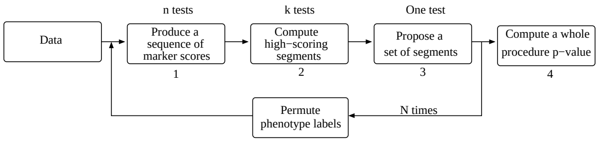
We define by:

$$H = \max_{1 \leqslant i \leqslant j \leqslant n} \left(\sum_{i}^{j} X_k \right)$$

the local score assigned to $\mathbb{X}$. The variables X_i must have a negative expectation otherwise the maximal segment would easily reach the entire sequence.

Considering $H^{(1)} \geqslant \ldots \geqslant H^{(k)}$ as being the scores of the k successive and distinct highest-scoring segments, $H^{(i)}$ defines the local score of the initial sequence disjoint from the preceding $k-1$ best segments.

Algorithm



Data

n tests
Produce a sequence of marker scores
1

k tests
Compute high−scoring segments
2

One test
Propose a set of segments
3

Compute a whole procedure p−value
4

Permute phenotype labels

N times

1 - Produce a sequence of marker scores: X_i can be based on classical statistics for association or corresponding p-values. $\mathbb{X}$ must generaly substracted by a constant δ. In this case we consider $\mathbb{X}' = (X_i')_{i=1,\ldots,n}$ with $X_i' = X_i - \delta$ such as $\mathbb{E}(X_i') < 0$. **Markers with a score higher than δ will improve the cumulate score of a given segment.**

2 - Compute the highest-scoring segments: Identify the successive high-scoring segments and compute their local scores $H^{(1)}, \ldots, H^{(k)}$. A naive approach is to use an iterative algorithm: **(i)** find the highest-scoring segment, **(ii)** remove it, **(iii)** iterate while the next best local score is positive.

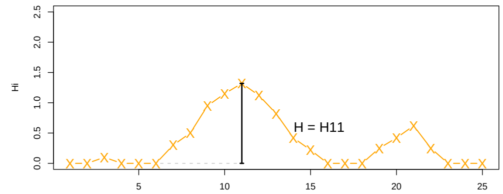
3 - Propose a set of segments: Successive local scores are combined into $T^{(1)}, \ldots, T^{(k)}$ such as $T^{(i)} = H^{(1)} + \ldots + H^{(i)}$. Corresponding p-values $p_{T1}, \ldots, p_{Tk}$ are computed using results from the extreme values theory or Monte-Carlo simulations (permuting case and control labels). Interesting segments are assumed to be the r first ones with $r = \max(\arg\min_{1 \leqslant i \leqslant r}(p_{Ti}))$ and $p^{(0)}_{\text{min}} = p_{Tr}$ is the statistic attached to this selection.

4 - Global p-value: The global significance of the process p_G is assessed via Monte-Carlo simulations: iterate N times steps 1 to 3, permute each time case and controls labels and compute $p^{(i)}_{\text{min}}$ corresponding to the i^{th} iteration. Finally:

$$p_G = \frac{\text{card}\left\{i, p^{(i)}{}_{\text{min}} \leqslant p^{(0)}{}_{\text{min}}\right\}}{N}$$

Implementation

- Instead of $\mathbb{X}$, we use the processus $\mathbb{H} = (H_i)_{i=1,\ldots,n}$ with $H_i = \max(0, H_{i-1} + X_i)$ and $H_0 = \max(0, X_0)$: finding the maximal scoring subsequence comes down to find $H = \max(H_i)$ what is $O(n)$ instead of $O(n^2)$.



- Use the $O(n)$ Ruzzo and Tompa algorithm (1999) instead of the naive $O(n^2)$ approach to find the successive high-scoring segments.

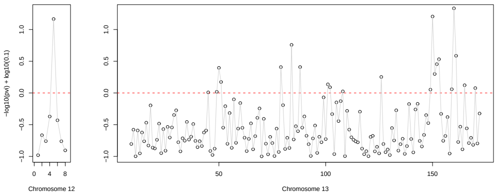
Application

Data: case-control data implicating G72 and DAAO genes in schizophrenia (Chumakov et al 2002).

Statistic: χ^2 on allelic contingency tables and p_i is the p-value coresponding to the SNP i.

Marker scores: $X_i' = X_i - \delta$ with $X_i = -\log_{10}(p_i)$

Parameters: $\delta = -\log_{10}(0.1)$ and $N = 10000$



Results: Our approach selects 3 segments localised in G72 and DAAO genes that have been proved to be involved to the disease and interacting with each other. The whole significance process is $p_G = 0.22$.

rank	chr	segment	H	T	p_T
1	13	149-153	2.542	2.542	0.2459
2	13	159-161	1.978	4.520	0.1737
3	12	5	1.165	5.686	0.1660
4	13	84	0.758	6.444	0.1702
5	13	49-51	0.587	7.031	0.1747

SNP	p_i	Bonferroni	FDR
160	0.0046	0.79	0.79
150	0.0062	1.00	0.54
5	0.0068	1.00	0.39
84	0.0175	1.00	0.75
...	...	...	...

Note that each segment differ in size from the others; this underline the advantage of the local score statistic over sliding-frames.

References

[1] Hoh, J., Wille, A. and Ott, J. (2001) Trimming, weigthing, and grouping SNPs in human case-control association studies. *Genome Research*, 11, 2115–2119.

[2] Karlin, S. and Altschul, S. (1993) Application and statistics for multiple high-scoring segments in molecular sequences. *PNAS*, 90, 5873–5877.

[3] Ruzzo, W.L. and Tompa, M. (1999) A linear time algorithm for finding all maximal scoring subsequences. *7th ISMB*, 234–241.

[4] Chumakov, I. et al (202) Genetic and physiological data implicating G72 and DAAO in schizophrenia. *PNAS*, 99, 13675–13680.