BIOINFORMATICS APPLICATIONS NOTE Vol. 21 no. 8 2005, pages 1735–1736
doi:10.1093/bioinformatics/bti201

Genetics and population analysis

CLUSTAG: hierarchical clustering and graph methods for selecting tag SNPs

S. I. Ao[1], Kevin Yip[2], Michael Ng[1,*], David Cheung[2], Pui-Yee Fong[3], Ian Melhado[3] and Pak C. Sham[3]

[1]Department of Mathematics, [2]Department of Computer Science and [3]Genome Research Center, The University of Hong Kong, Pokfulam, Hong Kong

Received on October 13, 2004; revised on November 30, 2004; accepted on December 1, 2004
Advance Access publication December 7, 2004

ABSTRACT

Summary: Cluster and set-cover algorithms are developed to obtain a set of tag single nucleotide polymorphisms (SNPs) that can represent all the known SNPs in a chromosomal region, subject to the constraint that all SNPs must have a squared correlation $R^2 > C$ with at least one tag SNP, where C is specified by the user.

Availability: http://hkumath.hku.hk/web/link/CLUSTAG/CLUSTAG.html

Contact: mng@maths.hku.hk

There is an estimated 10 million single nucleotide polymorphisms (SNPs) in the human genome. Although only a proportion of these SNPs are functional, all can be used as markers for indirect association studies to detect disease-related genetic variants. The complete screening of a gene or a chromosomal region is nevertheless an expensive undertaking. A key strategy to improve the efficiency of association studies is to select a subset of informative SNPs, called tag SNPs, for analysis (Johnson *et al*., 2001).

Methods for tag SNP selection based on established multivariate statistical techniques may offer some advantages. Byng *et al*. (2003) proposed the use of single and complete linkage hierarchical cluster analysis to select tag SNPs. Hierarchical clustering starts with a square matrix of pairwise distances between the objects to be clustered. For the problem of tag SNP selection, the objects to be clustered are the SNPs, and an appropriate measure of distance is $1 - R^2$, where R^2 is the squared correlation between two SNPs. The rationale is this: the required sample size for a tag SNP to detect an indirect association with a disease is inversely proportional to the R^2 between the tag SNP and the causal SNP.

In agglomerative clustering, the two clusters with the smallest inter-cluster distance are successively merged until all the objects have been merged into a single cluster. Different forms of agglomerative clustering differ in the definition of the distance between two clusters, each of which may contain more than one object. In single-linkage or nearest-neighbour clustering, the distance between two clusters is the distance between the nearest pair of objects, one from each cluster. In complete linkage or farthest neighbour clustering, the distance between two clusters is the distance between the farthest pair of objects, one from each cluster. The clustering process can be represented using a dendrogram, which shows how the individual objects are successively merged at greater distances into larger and fewer clusters. All distinct clusters that have been generated at or below a certain user-defined distance are considered (Fig. 1).

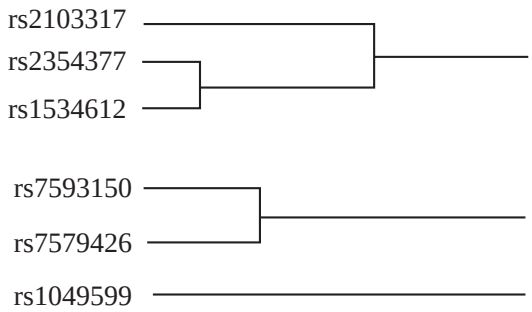

Fig. 1. Sample illustrative dendrogram showing how seven SNPs are merged into three clusters at or below the cut-off merging distance.

A desirable property for a clustering algorithm, in the context of tag SNP selection, would be that a cluster must contain at least one SNP (the tag SNP) that is no more than the merging distance from all the other SNPs from the same cluster. If this is the case, then by setting a cut-off merging distance of C, one can ensure that no SNP is further than C away from the tag SNP in its cluster. In this sense, none of the methods proposed by Byng *et al*. (2003) is ideal, since the single-linkage method does not guarantee the existence of a tag SNP with distance less than C from all SNPs in the same cluster, while complete-linkage is too conservative in that all SNPs have distance under C from all other SNPs in the same cluster.

In order to achieve the desired property described above, we propose a new definition of the distance between two clusters, as follows:

- For each SNP belonging to either cluster, find the maximum distance between it and all the other SNPs in the two clusters.
- The smallest of these maximum distances is defined as the distance between the two clusters.
- The corresponding SNP is defined as the tag SNP of the newly merged cluster.

We call this method minimax clustering. There is a parrallel in topology in which the distance between two compact sets can be

*To whom correspondence should be addressed.

Table 1. Properties of three tag SNP selection algorithms, evaluated for ENCODE regions

Encode region (SNP no.)	Compression			Compactness			Run time (s)		
	Complete	Minimax	Set cover	Complete	Minimax	Set cover	Complete	Minimax	Set cover
2A (519)	0.277	0.245	0.247	0.021	0.033	0.037	3.94	5.42	3.20
2B (595)	0.291	0.255	0.261	0.018	0.033	0.032	5.44	6.92	4.03
4 (665)	0.242	0.211	0.209	0.016	0.031	0.035	6.53	13.30	5.25
7A (417)	0.314	0.281	0.281	0.013	0.028	0.032	2.56	3.39	2.00
7B (463)	0.186	0.166	0.171	0.020	0.030	0.035	3.53	5.03	2.84
7C (433)	0.240	0.217	0.215	0.018	0.019	0.021	2.38	3.28	1.80
8A (364)	0.269	0.245	0.245	0.019	0.035	0.040	2.39	2.94	1.83
9 (258)	0.360	0.318	0.314	0.012	0.025	0.031	1.47	1.74	0.98
12 (454)	0.260	0.227	0.227	0.017	0.028	0.034	2.69	3.69	2.03
18 (350)	0.283	0.254	0.254	0.014	0.033	0.037	2.17	2.81	1.64

measured by a sup-inf metric known as Hausdorff distance (Barnsley, 1988).

For comparison we have also implemented an algorithm based on the NP-complete minimum dominating set of the set-cover problem, similar to the greedy algorithm developed by Carlson *et al*. (2004). The set of SNPs are the nodes of a graph, which are connected by edges where their corresponding SNPs have $R^2 > C$. The objective is to find a subset of nodes such that all the nodes are connected directly to at least one SNP of that subset. The algorithm is heuristic, and the details can be found in Reuven and Zehavit (2004). Briefly, at the beginning, all the SNPs belong to the untagged set. The algorithm picks the node with the largest number of nodes that are connected directly to it (without passing through any other nodes) from the untagged set. Then the SNPs inside the selected subset are deleted from the untagged set, and the next largest connected subset is chosen from the untagged set. The algorithm terminates when the untagged set becomes empty.

We have implemented the complete linkage, minimax linkage and set cover algorithms in the program CLUSTAG. The program takes a file of R^2 values produced, e.g. by HAPLOVIEW (Barrett *et al*., 2005), and outputs a text file containing one row per SNP and the following columns: (1) SNP name, (2) cluster number, (3) chromosomal position, (4) minor allele frequency, (5) maximal distance $(1 - R^2)$ from other SNPs in the same cluster and (6) average distance $(1 - R^2)$ from other SNPs in the cluster. Both columns (5) and (6) are useful for providing alternative SNPs that can serve as the tag SNP of the cluster, allowing some flexibility in the construction of multiplex SNP assays. A visual display (in the HTML format) provides a representation of the SNPs in their chromosomal locations, colour-labeled to indicate cluster membership. The tag SNP of each cluster is highlighted and hyperlinked to a text box containing columns (1)–(6) on the cluster.

We have compared the performance of the three implemented algorithms, using SNP data from the ENCODE regions of the HapMap project, according to three criteria: (1) compression, the ratio of clusters to SNPs; (2) compactness, the average distance between a SNP and the tag SNP of its cluster $(1 - R^2)$; and (3) run time. Our results show that the compression ratio is roughly equivalent for the set cover and minimax clustering algorithms but substantially higher for the complete linkage (Table 1). The minimax algorithm produces more compact clusters than the set-cover algorithm, but takes approximately twice as long to run. The runtimes of all three algorithms are expected to increase in proportion to the square of the number of SNPs.

ACKNOWLEDGEMENTS

The ENCODE data were downloaded from HAPMAP's site www.hapmap.org on June 30, 2004 and is based on NCBI build34. We thank the two anonymous reviewers for their constructive comments. The work is supported by small project grant to P.C.S. from the University of Hong Kong, and RGC grant nos HKU7130P, 7046P, 7035P to M.N.

REFERENCES

Barnsley,M.F. (1988) *Fractals Everywhere*. Academic Press, Boston, MA.

Barrett,J.C., Fry,B., Maller,J. and Daly,M.J. (2005) Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*, **21**, 263–265.

Byng,M., Whittaker,J.C., Cuthbert,A.P., Mathew,C.G. and Lewis,C.M. (2003) SNP subset selection for genetic association studies. *Ann. Hum. Genet.*, **67**, 543–556.

Carlson,C., Eberle,M.A., Rieder,M.J., Yi,Q., Kruglyak,L. and Nickerson,D.A. (2004) Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am. J. Hum. Genet.*, **74**, 106–120.

Johnson,G., Esposito,L., Barratt,B.J., Smith,A.N., Heward,J., Di Genova,G., Ueda,H., Cordell,H.J., Eaves,I.A., Dubridge,F. *et al*. (2001) Haplotype tagging for the identification of common disease genes. *Nat. Genet.*, **29**, 233–237.

Reuven,Y. and Zehavit,K. (2004) Approximating the dense set-cover problem. *J. Comput. Syst. Sci.*, **69**, 547–561.