BIOINFORMATICS

Vol. 00 no. 00 2005
Pages 1–2

ALTree: Association detection and Localization of susceptibility sites using haplotype Phylogenetic Trees

Claire Bardel [a]*, Vincent Danjean [b] and Emmanuelle Génin[a]

[a]Unité de recherche en Génétique Épidémiologique et Structure des Populations Humaines, INSERM U535, Villejuif, France, [b] Laboratoire Bordelais de Recherche en Informatique, UMR 5800, Bordeaux, France

ABSTRACT

Summary:

Finding the genes involved in complex disease susceptibility and among those genes, localizing the variant sites explaining this susceptibility is a major goal of genetic epidemiology. In this context, haplotypic methods that use the joint information on several markers may be of particular interest. When the number of haplotypes is large, a grouping may be required. Phylogenetic trees allow such groupings of haplotypes based on their evolutionary history and may help in the detection and localization of disease susceptibility sites. In this paper, we present a new software to perform phylogeny-based association and localization analyzes.

Availability: The software package, including all documentation and example is freely available at `http://claire.bardel.free.fr`. It is distributed under the GPL license.

Contact: bardel@vjf.inserm.fr

1 INTRODUCTION

When looking for an association between a candidate gene and a disease, genetic markers can be either studied one at a time or combined with the other markers located on the same chromosome to form haplotypes. In the last few years, haplotypic methods have been shown to be powerful to look for association (Akey *et al.*, 2001; Zaykin *et al.*, 2002). However, when the number of haplotypes increases, the power of the association test will decrease, due to the increase in the degree of freedom and to the decrease in the sample size per haplotype. To reduce these problems, it has been proposed to group haplotypes according to their evolutionary history (Templeton *et al.*, 1987; Seltman *et al.*, 2003; Durrant *et al.*, 2004; Bardel *et al.*, 2005). Moreover, such groupings allows to make hypotheses about the markers responsible for the susceptibility to the disease: in the evolutionary tree, mutations defining a group containing more haplotypes carried by cases than controls are putative susceptibility sites for the disease.

In this paper, we present a new software to perform a phylogeny-based association and localization analysis based on the method described in Bardel *et al.* (2005). The software deals with SNP haplotype data. It is written in perl, with one package in C. It is composed of three programs: the main analysis program, `altree`, which performs either the association or the localization analysis depending on the option selected by the user and two utilities, `altree-add-S` and `altree-convert`, which help the user managing its data file before running `altree`.

*to whom correspondence should be addressed

2 THE MAIN PROGRAM: `ALTREE`

To run `altree`, two input files are required: first, a file containing for each haplotype, the number of case and control individuals carrying it, and second, a file containing the phylogenetic tree and the list of character state changes on the tree. This last file is obtained by running a phylogeny reconstruction program. Currently, `altree` can deal with three phylogeny softwares: `paup` (Swofford, 2002) and `phylip` (Felsenstein, 2004) for a cladistic tree reconstruction, and `paml` (Yang, 1997) for a maximum likelihood tree reconstruction.

2.1 Association detection

The principle of the method is to perform series of nested case/control homogeneity tests in the different clades defined on the phylogenetic tree. P-values are computed at each level of the tree and a global p-value corrected for multiple testing is computed for the test, using a double permutation procedure (Becker and Knapp, 2004). The association test required the tree to be rooted because the nested analysis began at the root of the tree. Consequently, in the data set, an outgroup or an ancestral haplotype must be provided by the user so that `altree` can root the tree.

The output file contains the phylogenetic tree, with the number of case and control haplotypes in each branch, and the p-value of the association test.

2.2 Localization of the susceptibility site(s)

The principle of the localization analysis is to use a new character called "S" that represents the disease status. For each haplotype the state of this character depends on the proportion of cases carrying this haplotype (state 1 for a large proportion of cases and 0 else). The character state changes are optimized on the tree for each character (including S), and a correlated evolution index is calculated between each changes of each site and the changes of the character S. The site(s) whose evolution is the most correlated to the S character, is the most probable susceptibility site.

For the localization test, the character states can be reconstructed on the tree either by `paup` (parsimony method) or `paml` (Maximum likelihood method), but not by `phylip` because it leaves too many ambiguities in the ancestral character reconstructions. Contrary to the association analysis, the tree does not need to be rooted. When several trees are available (for example, several equiparsimonious trees), the user can specify how many trees should be included in the analysis. They will then be picked up at random from the total sample of parsimonious trees. Two kinds of correlated evolution index can be calculated. The first is the one described in Bardel

et al. (2005): only transitions from 0 to 1 for the character S are taken into account. A second index which seems more appropriate is now available: it takes into account both transitions of S (from 0 to 1 and from 1 to 0). The correlated evolution index can be chosen using the option `--coevo`

The output file of the program consists in a list of all the correlated evolution index for all transitions of all sites, ranked in decreasing order. The phylogenetic trees can be included in the output file using the option `--print-tree`.

3 TWO UTILITIES

3.1 A file converter: `altree-convert`

Our software analyzes haplotypic data. Such data are not generally directly available and they must be obtained by using haplotype reconstruction programs. `altree-convert` allows to convert output files from two haplotype reconstruction programs, `phase` (Stephens *et al.*, 2001; Stephens and Donnelly, 2003) and `FamHap` (Becker and Knapp (2004), only for haplotype reconstructed on family data) into input files for the phylogeny reconstruction programs (`paup` or `phylip` file format). When the `paup` file format is chosen, a list of commands is also written in the file so that users who are not familiar with `paup` can run it with only a few modifications to this file.

3.2 Definition of the S character: `altree-add-S`

To define the state of the character S, the user can choose its own criterion and add the character S manually in the `paup` input file. Otherwise, the user can use `altree-add-S`. This program takes a `paup` or a `phylip` input file and a list of the number of cases and controls carrying a given haplotype in input and returns a new `paup` or `phylip` input file in which the S character has been added according to the following criterion: the state of S is "0", "1" or "?" depending on the proportion (p_h) of cases carrying the haplotype h compared to the proportion p_0 of cases in the whole sample.

- if $p_h < p_0 - \epsilon\sqrt{\frac{p_h \times (1-p_h)}{n_h}}$, S is coded "0" (high number of controls);
- if $p_h > p_0 + \epsilon\sqrt{\frac{p_h \times (1-p_h)}{n_h}}$, S is coded "1" (high number of cases);
- else, S is coded "?" (missing data).

with n_h being the number of individuals carrying the haplotype h and ϵ, a number chosen by the user.

4 REQUIREMENTS

The software requires perl 5.8, a C compiler and a phylogeny reconstruction software: `paup`, `phylip` or `paml`.

5 CONCLUSION

Our software groups haplotypes based on their phylogenetic relationships to perform both association and localization analyzes. With the two utilities, it is easily usable, even for users who are not accustomed to phylogeny reconstruction programs. The efficiency of the method implemented in this software has been evaluated by simulations and on a data set concerning Crohn disease. We have shown that it is especially interesting when more than one susceptibility site is involved in the disease (Bardel *et al.*, 2005).

REFERENCES

Akey, J., Jin, L. and Xiong, M. (2001) Haplotypes *vs* single marker linkage disequilibrium tests: what do we gain? *European J of Hum Genet*, **9**, 291–300.

Bardel, C., Danjean, V., Hugot, J. P., Darlu, P. and Génin, E. (2005) On the use of haplotype phylogeny to detect disease susceptibility loci. *BMC Genetics*, **6**.

Becker, T. and Knapp, M. (2004) A powerful strategy to account for multiple testing in the context of haplotype analysis. *Am J Hum Genet*, **75**, 561–570.

Durrant, C., Zondervan, K. T., Cardon, L. R., Hunt, S., Deloukas, P. and Morris, A. P. (2004) Linkage disequilibrium mapping via cladistic analysis of single-nucleotique polymorphism haplotypes. *Am J Hum Genet*, **75**, 35–43.

Felsenstein, J. (2004) Phylip (phylogeny inference package) version 3.6. `http://evolution.genetics.washington.edu/phylip.html`. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.

Seltman, H., Roeder, K. and Devlin, B. (2003) Evolutionary-based association using haplotype data. *Genet Epidemiol*, **25**, 48–58.

Stephens, M. and Donnelly, P. (2003) A comparison of bayesian methods for haplotype reconstruction from population genotype data. *Am J Hum Genet*, **73**, 1162–1169.

Stephens, M., Smith, N. J. and Donnelly, P. (2001) A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet*, **68**, 978–989.

Swofford, D. L. (2002) `paup` phylogenetic analysis using parcimony. version 4.0b10. Sunderland, Massachusetts: Sinauer Associates.

Templeton, A. R., Boerwinkle, E. and Sing, C. F. (1987) A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. I. Basic theory and an analysis of alcohol dehydrogenase activity in Drosophila. *Genetics*, **117**, 343–351.

Yang, Z. (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *Computer application in bioSciences*, **13**, 555–556. Http://abacus.gene.ucl.ac.uk/software/paml.html.

Zaykin, D. V., Westfall, P. H., Young, S. S., Karnoub, M. A., Wagner, M. J. and Ehm, M. G. (2002) Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals. *Hum Hered*, **53**, 79–91.