Psychology Science, Volume 45, 2003 (2), p. 236-240

CFA as a method for genetic association studies in complex diseases

Marcus Ising[1]

Summary

Genetic factors contribute to the development of a variety of complex diseases, which are presumably determined by interactions of multiple polymorphic gene loci. Therefore, the genetic research has to focus multiple allele configurations rather than single genes that typically characterize patients and distinguish them from an adequate control sample. In the present paper a non-parametric approach based on the Configural Frequency Analysis (CFA) is suggested for the multivariate analysis of genetic association studies in complex diseases. With this case oriented approach typical ($f > e$) and anti-typical ($f < e$) haplotype allele configurations in polymorphic candidate genes are identified in affected individuals. The method can be applied in case-control and in family based association studies. In case-control studies the frequencies of the haplotype allele configurations in cases is compared with the corresponding expected allele frequencies that are derived from the control sample representing the allele distribution in the population. In family based association studies the frequencies of the haplotype alleles not transmitted to the affected off-spring serve as control sample. The application of the method is demonstrated with clinical data published by Hoehe and coworkers (2000), and advantages and limitations of the approach are discussed.

Key words: Configural Frequency Analysis (CFA), genetic association studies, complex diseases, allele frequencies, haplotype configurations

[1] Dr. Marcus Ising, Max-Planck-Institute of Psychiatry, Kraepelinstr. 10, D-80804 Munich–Germany; E-mail: ising@mpipsykl.mpg.de

1. Introduction

It is well established that genetic factors contribute to the development of a variety of complex diseases including psychiatric disorders. For example, the relative risk to develop a bipolar disorder or schizophrenia in healthy subjects with an affected first-degree relative is elevated up to 7 or 11 times, respectively, compared to healthy subjects without a familial history of psychiatric disorders (NIMH, 1999). The pathology underlying complex diseases is mostly explained by variations in peptidergic or proteinergic structures like receptors, transporters, or messenger molecules. The synthesis of these structures is regulated by the expression of genes associated with these molecules. With the increasing understanding of the biological patho-mechanisms of complex diseases, and with the availability of cost-effective high-throughput methods for the detection of polymorphic gene loci, multivariate statistical methods are of increasing importance for the detection of functional associations between multiple genetic polymorphisms and complex diseases.

Diagnostic entities like schizophrenia or depression are phenotypically heterogeneous, but are assumed to have a common biological background. Compared with association analyses, the power of linkage analyses is very limited in complex diseases (Risch & Merikangas, 1996). Functional association analyses are needed to identify polymorphisms in neurobiologically relevant candidate genes that are assumed to be causally linked to the vulnerability for a complex disease. The results from segregation analyses suggest that complex diseases are genetically determined by the interaction of multiple polymorphic gene loci. Therefore, genetic research in complex diseases has to focus the identification of multiple allele configurations rather than single genes that typically characterize patients and distinguish them from an adequate control sample. For the detection of complex genotype – phenotype associations parametric multivariate methods like regression analysis or variance components analysis are available.

In the present paper, a non-parametric multivariate approach is suggested based on the Configural Frequency Analysis (CFA). CFA was originally proposed by Gustav A. Lienert (1969). It is a distribution-free configural and case oriented approach for the analysis of multidimensional cross-tabulations for cells that contain significantly more (types) or fewer cases (antitypes) than expected from a chance model. CFA has been further developed by von Eye (1990, 2002), who introduced log-linear modells for the estimation of expected cell frequencies for different CFA chance models.

2. Allele – Configuration Frequency Analysis

Candidate genes in complex diseases usually comprise not only a single polymorphic locus but several polymorphic regions. It can be assumed that not a single polymorphism but specific patterns of polymorphisms affect the protein synthesis that is controlled by the gene. Therfore, the multivariate analysis of allele configurations of a haplotype is more promising for the identification of functionally relevant genetic polymorphisms in complex diseases than the analysis of single gene loci. The Configural Frequency Analysis (CFA) is suggested as a promising approach for the identification of typical and anti-typical haplotype allele configurations of polymorphic candidate genes in patients with complex diseases.

Since the set of chromosomes is diploid, every individual carries two alleles of a haplotype; one allele is maternally and the other paternally transmitted. Allele – CFA can be applied to compare the frequencies of haplotype alleles in affected individuals with the allele frequencies of an adequate control sample.

The first step in the application of a haplotype analysis is the selection of candidate genes that are neurobiologically relevant for the complex disease under study, and the identification of all polymorphic loci of the selected candidate genes. The polymorphic loci are combined to a haplotype resulting in configurations with '1' indicating identity with the reference (frequent) allele and '2' indicating a polymorphism. The frequencies of the different allele configurations are determined.

The second step is the selection of an appropriate control group. In case-control association studies the haplotype allele configurations are determined in individuals from an ethnically identical population without signs of vulnerability or affection to the complex disease under study. In family based association studies the parental DNA is genotyped, and the frequencies of the haplotype alleles not transmitted to the affected off-spring serve as control sample.

Allele configurations that do not exist neither in cases nor in controls are eliminated, since they are presumably biologically irrelevant for the population under study. The remaining haplotype allele frequencies are compared between cases and controls. Since the total number of observations is high and the single frequencies of the polymorphic haplotype configurations are low, a Poisson approximation of a binomial test can be applied. The expected allele frequencies are derived from the allele frequencies observed in the control group weighted by the ratio of the two sample sizes. The level of significance needs to be adjusted for multiple comparisons by an appropriate procedure, e.g. by Bonferroni adjustment. A specific haplotype allele configuration is identified as a type if the observed allele frequency in affected individuals is significantly larger than the expected frequency derived from the control group ($p(f \geq f_{obs}, \lambda = e) < \alpha^*$, with $\alpha^* = a/k$ and k := number of configurations); a haplotype allele configuration is identified as an anti-type if the observed allele frequency is significantly smaller than the expected frequency ($p(f \leq f_{obs}, \lambda = e) < \alpha^*$, with $\alpha^* = a/k$ and k := number of configurations).

3. Example

In order to demonstrate the application of the Allele – CFA we selected data from a case–control study by Hoehe and coworkers (2000), published in Human Molecular Genetics, volume 9, issue 19, pages 2895 to 2908. The authors investigated the μ-opioid receptor gene (OPRM1) as a candidate gene in 137 substance dependent African-Americans (cases) and 35 controls, who were African-Americans as well. They identified 25 relevant polymorphic loci, most of them single nucleotide polymorphisms (SNPs). 52 different haplotype configurations of the 25 polymorphic sites could be observed, and the haplotype allele configuration frequencies in cases and controls are reported by Hoehe et al. (2000), table 2 (p. 2900), with "1" indicating identity with and "2" deviation from the reference sequence for each of the 25 loci. Since every subject carries two alleles of the haplotype the total sample size is 274 case alleles and 70 control alleles.

Because of the low frequencies of the single configurations and the high number of overall observations, we applied a Poisson approximation of a binomial test in order to compare the haplotype allele frequencies between cases and controls. We estimated the expected frequencies from the observed counts in the control sample multiplied by the ratio of the two sample sizes (*274/70). In case of an expected frequency of zero the observed and the expected frequency was increased by one. The level of significance was set to $p = 0.05$, which was Bonferroni corrected for simultaneous comparisons of the 52 haplotype configurations resulting in an adjusted level of significance of $p_{crit} = 0.0009$. Three of the 52 haplotype configurations showed higher case allele frequencies with a p-score below the Bonferroni adjusted level of significance. These

three haplotype configurations comprise eight polymorphic gene loci and 17 invariant loci identical with the reference sequence. The reference sequence and the significant haplotype configurations after elimination of the 17 invariant loci are presented in table 1.

Table 1:
Reference sequence and significant haplotype configurations (pcrit. = .0009) after elimination of invariant loci.

	Haplotype OPRM1 Gene								Frequencies			Significance
No.	**01**	**03**	**05**	**06**	**08**	**09**	**13**	**15**	**Cases**	**Expected**	**Controls**	$\mathbf{p_{crit.} = .0009}$
0	Reference sequence								157	168.31	43	0.2034
1	1	1	1	1	2	1	1	1	12	3.91	1	0.0008
19	2	2	2	2	1	1	2	2	13	1.00	0	< 0.0001
31	1	1	1	1	1	2	1	2	6	1.00	0	0.0006
All others									88	101.77	26	0.0920

Note. In the case of an expected frequency of zero the observed and the expected frequency was increased by one.

Three haplotype allele configurations comprising eight polymorphic loci can be regarded as associated with the development of substance dependence in African-American subjects. Six of the eight configuration loci are positioned in the promoter region of the gene suggesting a possible functional relevance of these haplotype configurations in the regulation of the OPRM1 gene expression.

4. Discussion

A non-parametric multivariate approach based on the Configural Frequency Analysis (CFA) was suggested for genetic association studies in complex diseases. With this case oriented approach typical (f > e) and anti-typical (f < e) haplotype allele configurations in polymorphic candidate genes are identified in affected individuals. Allele – CFA can be applied in case-control and in family based association studies. In case-control studies the frequencies of the haplotype allele configurations in cases is compared with the corresponding expected allele frequencies that are derived from the control sample representing the allele distribution in the population. In family based association studies the frequencies of the haplotype alleles not transmitted to the affected off-spring serve as the control sample.

The advantages of the Allele – CFA are (1) the transparency of the method and the plain interpretability of the results, (2) the applicability in case-control and family based association studies, (3) the distribution-free character of the method without special requirements to the data, and (4) the simultaneous identification of typical and anti-typical haplotype allele configurations and classification of the cases. Limitations of the Allele – CFA are (1) the exponentially rise in calculation effort with the increasing number of loci, which is accompanied by (2) a loss of power due to the need for protection against multiple testing. It is concluded that the Allele – CFA can be regarded as promising for the analysis of genetic associations in complex diseases. However, further studies are needed to evaluate the power of this method under different genetic models.

References

1. Hoehe, M., Köpke, K., Wendl, B., Rohde, K., Flachmeier, C., Kidd, K. K., Berrettini, W. H. and Church, G. M.: Sequence variability and candidate gene analysis in complex disease: Association of μ opioid receptor gene variation with substance dependence. Human Molecular Genetics, 9, 2895-2908, 2000.
2. Lienert, G. A.: Die „Konfigurationsfrequenzanalyse" als Klassifikationsmethode in der Klinischen Psychologie. In M. Irle (Hrsg.), Bericht über den 26. Kongress der Deutschen Gesellschaft für Psychologie in Tübingen 1968 (S. 244-253), Göttingen: Hogrefe, 1969.
3. NIMH: Report of the National Institute of Mental Health's Genetic Workgroup. Biological Psychiatry, 45, 559-602, 1999.
4. Risch, N. and Merikangas, K.: The future of genetic studies of complex human diseases, Science, 273, 1516-1517, 1996.
5. von Eye, A.: Introduction to configural frequency analysis, Cambridge: Cambridge University Press, 1990.
6. von Eye, A.: Configural frequency analysis - methods, models, and applications, Mahwah, NJ: Erlbaum, 2002.
