Anke Schulz
Dr. sc. hum.

Entropy based marker selection for haplotype sharing analyses in gene mapping of complex diseases

Promotionsfach: DKFZ (Deutsches Krebsforschungszentrum), Epidemiologie
Doktormutter: Prof. Dr. sc. hum. Jenny Chang-Claude

We developed a new entropy based marker selection algorithm, EMS, that selects markers for haplotype sharing analyses, which are utilised to detect disease susceptibility loci in complex diseases. EMS provides an alternative to selecting markers for haplotype analysis via sliding windows of fixed size or block structure. The algorithm can be paired with the Mantel statistics as well as other tests for haplotype sharing analysis.

For each tested marker, EMS iterates over the available marker set and in each step selects or rejects a marker based on its potential to increase the multilocus linkage disequilibrium (LD) of the current selection from the previous iteration step. We designed two rejection criteria, which demand multilocus LD increase to accept a marker into the selection, and two stop criteria. One stop parameter limits the maximum number of markers checked for LD increase at each side of the test marker, the other is the maximum number of rejections allowed to each side of the marker. The two stop criteria can be active at the same time. The selection at one side of the test marker is terminated as soon as one of the stop criteria is met or the marker set is exhausted for that side.

The multilocus LD measure employed is the normalised entropy difference, which is estimated from statistically reconstructed haplotypes of a population sample of interest. Each selection step requires haplotype reconstruction for the current marker combination. However, to balance run time efficiency with maximisation of multilocus LD, the number of steps, i.e. number of marker combinations assessed, is at most as high as the number of markers available, a drastic reduction compared to the number of all possible marker combinations.

After implementing the EMS and the Mantel statistics in R together with Monte Carlo permutation testing, the combination was applied to simulated as well as real candidate gene studies. We showed in simulated candidate gene scenarios with direct and indirect association that this new, combined approach can yield empirical power as high as if not higher than Mantel statistics using the whole marker set and Mantel statistics using five marker sliding windows, in spite of the approach's uncorrected (for multiple testing), yet conservative false positive rates. Moreover, in real candidate gene data, several EMS configurations combined with Mantel statistics found the same markers significant at the same magnitude as conditional logistic regression.

As a proof of principle for the feasibility in genome wide marker sets, we applied the combined approach under the assumption of asymptotic normality to a simulated chromosome spanning marker set, where it identified the neighbouring markers of the disease locus in an indirect disease scenario. Its empirical power values were comparable to the power of testing when sliding marker windows of fixed size were used, while both approaches, corrected for multiple testing, had predominantly conservative false positive rates.

In both candidate gene and chromosome wide simulations, observations indicate that EMS and the Mantel statistics might help in pinpointing where haplotype sharing with a different marker selection strategy yielded significant tests for several physically close markers.

The approach of combining EMS with Mantel's statistics and Monte Carlo permutation testing is already well suited for association studies in candidate genes. For each tested marker, the EMS algorithm performs an individual marker selection according to the local multilocus LD. Preceding haplotype sharing with EMS might be favourable in settings where recent mutations are expected to interrupt shared chromosomal regions around a disease mutation.

With small changes in the selection algorithm's design, for example the way the start set is chosen, EMS can be expected to be even better adjusted to local LD structure.

We can recommend using the EMS when re-examining a significant result of a genome wide screening, together with the Mantel statistics for haplotype sharing analysis, or as marker selection tool for subsequent haplotype association testing of haplotypes containing the significant variant.

It is also appealing to apply the EMS for haplotype sharing analysis in marker intensive genome wide screens or sequence scans, together with the Mantel statistics if the handling of the correction for multiple testing and genomewide significance level for the Mantel statistics can be optimally resolved. The EMS algorithm alone, however, could be paired with any other haplotype sharing test that is suitable for genomewide studies.