## Poster D-3

**Mixture Modeling in Fused Genotype and Phenotype Data**


**Authors:**
Benjamin Georgi (*Max-Planck Institute for Molecular Genetics*)
M. Anne Spence (*University of California at Irvine*)
Pam Flodman (*University of California at Irvine*)
Alexander Schliep (*Max-Planck Institute for Molecular Genetics*)

**Short Abstract:** We present results from applying mixture model based clustering on a data set of ADHD patients. The data includes both geno- and phenotypic features and the interactions between these two sources of data will be one focus point of the analysis.

**Long Abstract:**

The analysis of genetic diseases has classically been directed towards establishing direct links between cause, a genetic variation, and effect, the observable deviation of phenotype. For complex diseases in
which the degree of diagnostic uncertainty with respect to presence of the disease and determination of the disease's subtype is large, this search for simple, direct causalities is likely to fail.

One example for such a complex genetic disease is the Attention Deficit Hyperactivity Disorder (ADHD). ADHD is recognized as the most common neurological disorder in children in the US. As far as the genetic background of ADHS is concerned, like in a large number of other biological settings, a partition of disease subtype phenotypes into clearly separated groups cannot be expected. Rather some phenotypic variations will be caused by the overlapping effects of two or more distinct genetic mechanisms.

The classical statistical models to cope with different sources of data are mixture models, essentially convex combinations of density functions, which allow inference of descriptive models from data
as well as the deduction of groups. The major advantage over clustering approaches is caused by their higher degree of robustness with respect to noise. Moreover the probabilistic framework mixture are defined in, readily supplies diagnostics for
the level of ambiguity in the inference of groups from the mixture model. One example of such a diagnostic would be to impose an entropy cut-off on the posterior distribution. The integration of different sources of data in a mixture framework can be readily
achieved if simplifying independence assumptions in the
correlation structure are made. This amounts to adopting the so called naive Bayes (NB) model as component densities of the mixture. While the assumption of independence between the features
is certainly a strong one, in practice NB models have been extensively used with good success in a wide variety of research fields

We are going to present our results from applying mixture model based clustering on a data

set of ADHD patients. The data set includes single nucleotide polymorphism (SNP) data about the DRD
dopamine receptor family (DRD1-DRD5) as well as the dopamine transporter gene DAT1. The DRD family proteins are G-protein coupled dopamine receptors located in the plasma membrane. DAT1 encodes for a dopamine transporter located in the presynaptic
membrane. The dopamine metabolism is heavily implicated to of relevance for ADHD disease patterns. The phenotype data consists of clinical data in
form of IQ and achievement test scores, as well as diagnosis for a large number of comorbidit disorders.

We present results from applying mixture model based clustering on this data set and the interactions between these two sources of data will be one focus point of the analysis.