Analyzing Populations with Visual and Analytical Methods to Identify Family Clustered Diseases

Christian Fuchsberger[a,b], Silvia Miksch[b,c], Lukas Forer[a], Cristian Pattaro[a]

[a] *EURAC-Research, Department of Genetic Medicine, Bolzano-Bozen, Italy*
[b] *Vienna University of Technology, Inst. of Software Technology & Interactive Systems, Vienna, Austria*
[c] *Department of Information and Knowledge Engineering, Danube University Krems, Austria*

Abstract

The study of isolated, inbred populations is a promising approach for the identification of disease susceptibility genes. One of the main challenges for geneticists and epidemiologist is to deal with the very complex genealogies of such populations. We present ongoing research on using a visual analytics based approach for the identification of family clustered diseases, risk factors and heritability patterns.

Keywords:

Genetic epidemiology, Visualization, Disease Clustering, Pedigree.

Introduction

The study of familial disease history is the necessary starting point when assessing the extent of genetic components in the etiology of common diseases. The homogeneity of the shared environmental factors and the limited number of recombination events in the DNA make isolated population studies potentially more powerful than general population studies for dissecting complex traits. The limited population dimension often enables to reconstruct very precise genealogies going back to several generations in the past, including hundreds or thousands of individuals. Available genealogical software is mostly focused on a static representation or is able to visualize only sub-pedigrees [1]. Moreover, it does not provide tools for simplifying the visualization of genealogical trees. On the statistical side, existing methods either focus on small family units or require exact calculations that are not computationally feasible.

The principle of "Visual Analytics" (VA) [2] is to combine the outstanding visual capabilities of humans with the power of analytical methods to support the knowledge discovery process. We propose a new approach for the analysis of family histories at a population level based on the concept of VA. Therefore, we (i) developed an effective pedigree drawing, (ii) integrated results in the visualization and (iii) added various tools for supporting the exploratory process.

Methods

Pedigree Drawing

For the proper interpretation of pedigree based data, a clear layout is needed for reducing line crossings. Our speed optimized drawing algorithm is based on the 3 phase Sugiyama-heuristic [3]. Furthermore, to reduce line crossings and to facilitate the identification of heritability patterns, we developed a 2.5D visualization, whereby nodes are distributed on two distinct layers according to different criteria (Figure 1).

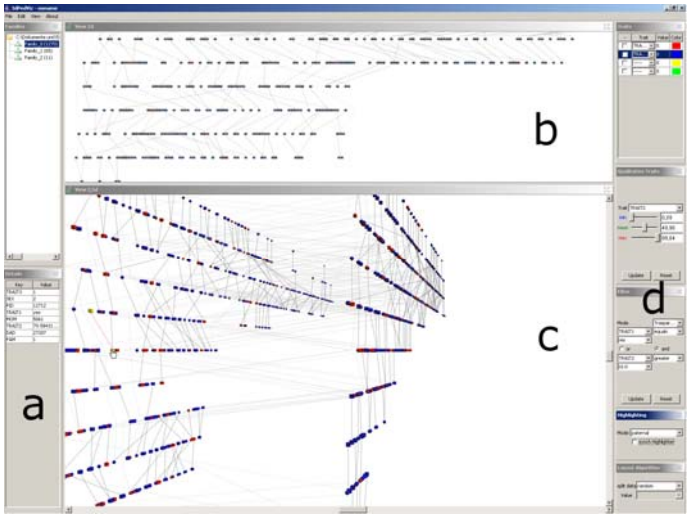

Figure 1 –Prototype PedVizApi: a.) Details on Demand, b.) 2d View c.) 2.5 View, d.) Dynamic Query Interface.

Analyzing and Visualizing Process

The information visualization mantra by Daniel Keim [4] "*Analyse First - Show the Important - Zoom, Filter and Analyse Further - Details on Demand*" was followed.
Analyse first. Initially, various pre-processing steps, such as quantification and normalization, were applied. For the identification of familial disease clusters, different statistical and computational methods are available. Since for this type of

analysis the hierarchical structure must be preserved, we integrated the results of the clustering algorithms into the pedigree drawing algorithm.

Show the Important. Depending on the research question, different qualitative and quantitative information is needed and has to be included in the pedigree drawing. Since for family history analysis close relatives are more disease relevant then distant ones, we assigned a transparency value to the connection lines based on the kinship coefficient.

Zoom, filter and re-analyse. The identification of family clustered diseases, risk factors and heritability patterns is an exploratory process. Therefore, we integrated a set of interactive tools, such as dynamic queries.

Details on demand. During the exploratory process additional information is required. On the one hand, static data can be retrieved from a data repository and displayed on demand. On the other hand, on the fly calculated information, such as the connection path between two individuals, is important to identify common risk factors or heritability patterns.

Results

We introduced this novel approach at our institute. Our data consists of the reconstructed genealogy of three isolated valleys, reaching back to the 1600s (50,037 individuals). Health information on 1175 subjects was obtained by means of a screening questionnaire consisting of 960 questions [5].

At this stage of our study, we are especially interested in assessing *which diseases or disease combinations are common and tend to group in families.* Analyses were carried out by three statisticians and four geneticists. Usually, these experts assess the presence of diseases clustering in families using statistical tests and then connecting affected individuals in the pedigree, for each disease of interest. The VA based method was new for them. It starts with an explorative process based on the whole genealogy, whereby the node positions are optimized regarding disease status/cluster, kinship, and line crossings. Based on the following tasks, the differences between the two approaches are highlighted.

Identification of disease clusters

For the subsequent analysis, the probability of familial clustering was estimated [6]. Due the large number of individuals, the exact calculation was not feasible. Moreover, the discovered disease clusters were always limited to small family units. Nonetheless, the VA approach identified clusters at the population level, whereby for the identification of nuclear family disease clusters (parents with less than 3 children) the pedigree must be explored more deeply.

Path between different disease clusters

Using the standard approach was limited to finding common ancestors between different clusters. The process was time-consuming because for every different disease setting paths and sub-pedigrees were reconstructed. With the VA approach, paths between the different clusters could be highlighted during the exploratory process. Moreover, missing and incomplete data were not a problem, because of working with the whole genealogy these parts were inferred by human perception.

Genetics or Environment

Analytical methods are able to quantify the contribution of genetic and environmental factors. At the family level this approach is doable; however, at the population level, it becomes time consuming and only a subset of hypotheses can be tested. The VA approach does not quantify exactly the influence of the various factors but, in consideration that this was a screening phase, this is acceptable. All analyses can be performed at a population level, allowing for the exploration of different hypotheses. Missing and uncertain data are balanced partially by the capabilities of the human perception.

Discussion and Conclusion

Due the complex structures of isolated population genealogies, classical, analytical methods used in a subsequent visualization setting are only able to identify local and/or single global population clustered diseases. Therefore, the distinction between genetic and environmental factors becomes difficult. Moreover, heritability patterns of rare diseases can only be identified by looking at the whole population in an “error tolerant way”, by exploring different diseases settings.

The VA based approach is capable to perform family history analyses at a population level. Sometimes, the identification of small, single, local disease clusters can be difficult, due to the global optimized representation. However, using the interactive capabilities, the identification of disease clusters and the generation of new hypothesis on the disease etiology are supported in an explorative way.

Integrating the power of the human perception with the VA method provides an appropriated framework for the knowledge discovery process. Nonetheless, experts have to explore findings from different points of view to confirm their robustness and to prevent false positive results.

References

[1] http://linkage.rockefeller.edu/

[2] Thomas JJ and Cook KA. A Visual Analytics Agenda, IEEE Computer Graphics and Applications 2006: 26(1): 10-13.

[3] Kaufmann M, Wagner D. Drawing graphs. Berlin: Springer, 2001.

[4] Keim D. Summary. Workshop on Visual Analytics. Darmstadt, 2005.

[5] Pattaro C, et al. The genetic study of three population microisolates in South Tyrol. BMC Med Genet (in press).

[6] Yu C, Zelterman D. Statistical inference for familial disease clusters. Biometrics 2002: 58: 481-91.