Visual Analytical Methods to Identify Family Clustered Diseases

Christian Fuchsberger, Lukas Forer, Cristian Pattaro, Andrew Hicks and Peter Pramstaller
EURAC-Research
Department of Genetic Medicine
Viale Druso 1, 39100 Bolzano, Italy
christian.fuchsberger@eurac.edu

Silvia Miksch
Danube University Krems
Department of Information and Knowledge Engineering
3500 Krems, Austria
silvia.miksch@donau-uni.ac.at

Abstract

The study of isolated populations is a promising approach for the identification of genes conferring susceptibility to disease. Moreover, it has a sustainable impact on the healthcare system of the studied population. Due to the complex genealogies of such populations epidemiological studies are challenging.

We present ongoing research to using a visual analytics based approach for the identification of diseases that cluster in families, risk factors and heritability patterns

1. Introduction

Studying disease history in families is a proven method for the identification of high risk factors for common diseases such as cardiovascular diseases, diabetes and several cancers. As a result of the increasing interest in using isolated populations for identification of the genetic risk factors to common diseases, complex extended genealogies are more readily available for the study of family history on a population level. Population isolation results from different factors such as geography, culture, religion and history.

Population isolates due to these reasons are relatively common and an increasing number of studies are underway, most recently with a renewed focus on the Middle East and Africa [2]. Pedigree information within these is extensive and usually traced back for as many generations as possible. The dimension of such genealogies is variable, varying from 200 to 30,000 individuals, distributed over 6-13 generations. Together with the genealogical information, available subjects undergo to a complex set of measurements such as screening or diagnostic questionnaires, clinical examinations, gathering quantitative trait data and many physiological measurements such as bone densitometry and electrocardiogram measurements etcetera. Following the identification of common diseases that cluster in families, different methods of genetic analysis are applied. This identification step is crucial, because it is the starting point for all further analysis and, more importantly, it can have an immediate impact on health care measure for individuals within the extended pedigrees, the local community and the health care system they belong to.

Despite the crucial importance of genealogical data in this process, available software packages are mostly focused on a statical representation of the pedigree information. Moreover, if the pedigree dimension is large, they are able to visualize only small subset of the genealogy. Other programs, which are able to correctly represent the genealogy, are not able to map the health data on the pedigree in an interactive manner [7, 8].

On the statistical side, existing methods focus mainly on small family units, cannot handle uncertainty regarding diseases status, do not consider the effect of common ancestors or, whenever the numbers of individuals is large, render exact calculation for the entire population computationally infeasible.

The principle of ”Visual Analytics” [11] is to combine the outstanding visual capabilities of humans with the power of analytical methods to support the knowledge discovery process. Most importantly, the user is not only an interpreter of visual and analytical output, but takes an active role in driving the whole process.

Therefore, according to [11], the visualization must:

- Facilitate the understanding of large heterogeneous data sets.

- Support the understanding of uncertain and incomplete data.

- Provide adaptive representation for different user-tasks.

- Support different data types on various levels of abstraction into a single representation.

We propose a new approach based on the concept of visual analytics for the analysis of family histories on a population level. Therefore, we (i) develop an effective pedigree drawing, (ii) integrate the analysis result in the visualization in a dynamic way and (iii) integrate different dynamic methods for supporting the explorative process.

2 Data and methods

The local population of South Tyrol, the northern-most province of Italy, was isolated as a result of the geographical structure and the historical and political background [9]. Within the GenNova project the pedigrees of three isolated valleys where reconstructed, and an extensive health screening was performed.

In total 1175 individuals participated in the study (>50% of the total population). The reconstructed genealogy goes back to the 1600s and includes 50,037 individuals. All participants completed a general health-questionnaire composed of 960 questions subdivided in 15 diseases areas, such as neurology, internal medicine and cardiology. Furthermore, for each individual, 43 blood parameters and additional quantitative traits, such as height, weight, bone density and serum biomarkers for disease were measured.

3 Visual and Analytical Methods

3.1 Pedigree drawing

For the proper interpretation of pedigree based data, a clear, line crossing reducing layout is needed. Our drawing algorithm is based on the 3 phase Sugiyama-heuristic. In the first phase the graph hierarchy based on the individual's generations are built. Then line crossings between the single layers are minimized by applying the barycentre method. Finally for any node of the resulting graph, screen coordinates have to be calculated. To perform this last step we integrated the rubber-band algorithm [10, 4]. All algorithm implementations are speed optimized to handle large genealogies (>30.000 individuals).

3.2 2.5D Visualization

Three dimensional visualisations, such as cone trees, overcome the problem of line crossing by using the additional dimension for the node positioning. However, as a result of the non conformity with the traditional pedigree representation the various end-users (biologists, statisticians, geneticists) are not more able to interpret the drawing properly. Therefore, we developed a 2.5D visualization, whereby nodes are distributed on two distinct layers using methods described in Hong and Nikolov[3]. As shown in Figure 1, the number of crossings is reduced, while preserving the classical representation. Furthermore nodes are split on the two layers according different criteria, such as disease status, to facilitate the identification of heritability patterns.

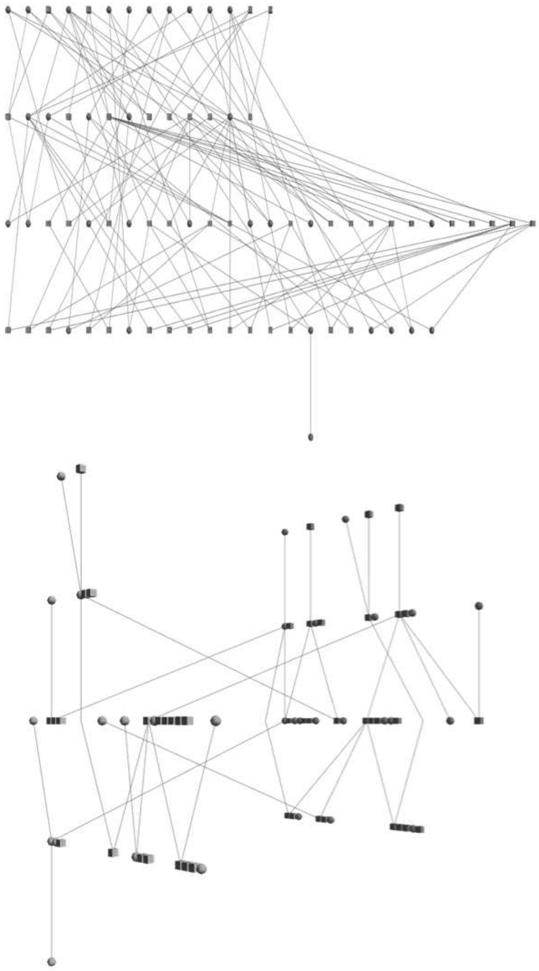


Figure 1. 78 individuals pedigree as 2D and 2.5D drawing.

3.3 Analyzing and Visualizing Process

This section is structured according to the information visualization mantra by Daniel Keim [5]: ”Analyse First - Show the Important - Zoom, Filter and Analyse Further - Details on Demand”.

Analyse first Initially various pre-processing steps, such as quantification, filtering and normalization were applied. For the identification of familial disease clusters and patterns in medical data, different statistical and computational analysis methods are available [12]. Since for the analysis of this type of data the hierarchical structure must be preserved, we integrated the results of the clustering algorithms, here hierarchical clustering, into the node reordering step (second and third phase of the Sugiyama algorithm). Therefore the positions of the nodes are based on their barycentres and are corrected by the cluster results. Furthermore the spring model of the rubber-band algorithm was extended to consider also the cluster results.

Show the Important Depending on the particular research question, different information is needed and has to be included in the pedigree drawing. The integration of qualitative data (Figure 2) is a basic task. However, recently the analysis of quantitative traits became more and more important. Quantitative endophenotypes, phenotypes closely linked to a disease, represent a possible intermediate measurement: a high body mass index (BMI), which is related to various cardiovascular diseases, is a case in point. Our visualization is the first attempt to map quantitative data on a pedigree structure without discretisation (Figure 3). We use the maximum, minimum, mean, and standard deviation values to calculate an appropriate colour bar, then around the individual’s value we extract a subsection of the bar and map this onto the genealogy.

For family history analysis, closed relatives are more disease relevant than distant ones. In addition, the information about distant relatives tends to be more imprecise. We incorporate this fact by assigning a transparency value to the connection lines based on the kinship coefficient. Kinship matrices are computed on the fly using a recursive algorithm from Lange [6].

Zoom, filter and re-analyse The identification of familial clustered diseases, risk factors and heritability patterns is an explorative process. The experts need a set of interactive tools to perform this task. The following interactivities are included in our approach:

- Zoom, move, rotate and incline (only for the 2.5D visualization).

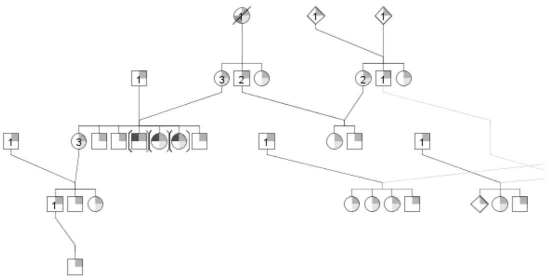


Figure 2. Pedigree symbols for quantitative traits according to [1].

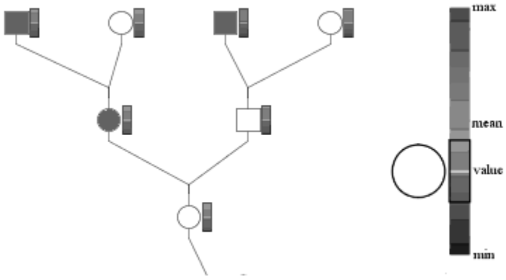


Figure 3. Sub-pedigree with the quantitative trait body mass index (BMI) and the cardiovascular affection status.

- Sub-pedigree extraction: by marking single individuals of interest the connecting sub-pedigree can be extracted. As the family history is of importance, relatives based on a given kinship coefficient are included.
- Pedigree comparison: the screen can be split in a number of windows to compare different sub-pedigree simultaneously.

Re-analysis is always possible, whereby the individuals may be re-clustered according to the new features or feature combinations and the node positions and their mappings recalculated.

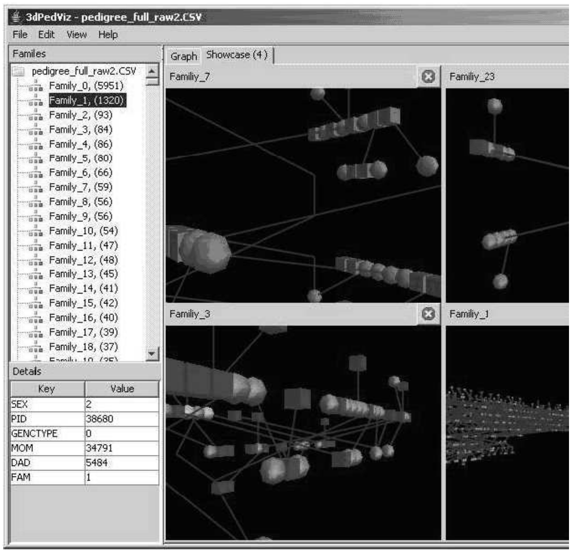
Details on demand During the explorative process additional information is required. On the one hand, statical data can be retrieved from a data repository and displayed on demand, for example, by focusing on a single individual and showing their details. On the other hand, on-the-fly calculated information, such as the connection path between two individuals, are quite important to identify common risk factors or heritability patterns. Table 1 shows the dynamic methods implemented so far.

Table 1. Dynamic methods for details on demand

Method	Description
cAncestor2	Common ancestor of 2 individuals
cAncestorN	Common ancestor of N individuals
mPath	Maternal lineage of a person
pPath	Paternal lineage of a person
allPathX	All individuals related to a person with a kinship of X
descPath	Full descend path of an individual
unRN	All individuals, which are not related to the selected N persons.
featPathN	All individuals, which show the same features or feature combinations and are related with a kinship of X.

4 Evaluation

We introduced this novel approach for the analysis of isolated populations at our institute by implementing a prototype (Figure 4).



pedigree_full_raw2.CSV
Family_0, (5951)
Family_1, (1320)
Family_2, (93)
Family_3, (84)
Family_4, (86)
Family_5, (80)
Family_6, (66)
Family_7, (59)
Family_8, (56)
Family_9, (56)
Family_10, (54)
Family_11, (47)
Family_12, (48)
Family_13, (45)
Family_14, (41)
Family_15, (42)
Family_16, (40)
Family_17, (39)
Family_18, (37)

Key	Value
SEX	2
PID	38680
GENCTYPE	0
MOM	34791
DAD	5484
FAM	1

Figure 4. Prototype: 2.5D Mode.

At this stage of our study we are especially interested in the identification of family disease clusters across our three study cohorts. The corresponding research question is: ”Which diseases or disease combinations are common and tend to group in families across the study populations?” The analyses were carried out by two different types of end-users: statisticians and geneticists. Usually these experts start with the analysis, then visualize the results obtained and, based on that, restart the analysis using different parameters or methods (subsequent visualization workflow); the tightly integrated visualization / analysis approach (visual analytics) was new for them (Figure 5).

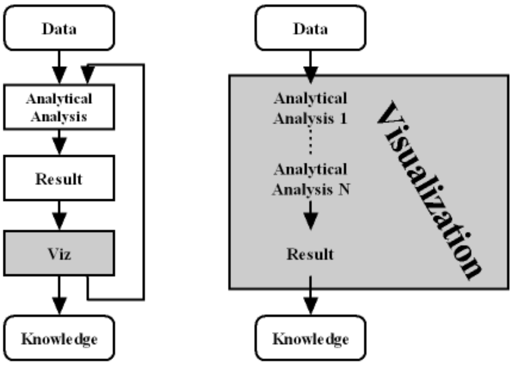


Figure 5. Subsequent and tightly integrated visualization.

As can be seen from the Figure 5, different workflows are required. The subsequent analysis starts with the identification of common diseases, and then the affected individuals are identified. After this, pedigrees connecting these individuals are reconstructed using the available genealogy. The reconstruction step must be repeated for every disease of interest, because the resulting pedigrees optimised for subsequent analysis can be different.

However the visual analytics approach starts an exploratory process based on the whole genealogy, whereby the node positions are optimized with respect to disease status / cluster, kinship, and line crossings. To obtain the initial disease clusters, simple hierarchical clustering is performed.

Based on the following sub-research-questions the differences between the two approaches are highlighted.

4.1 Identification of disease clusters

For the subsequent analysis the probability of familial clustering was estimated [12]. Due the large number of individuals the exact calculation was not feasible, thus a Monte Carlo simulation study was conducted. The discovered disease clusters were always limited to small family units. Nonetheless, the visual analytics approach identifies clusters on the population level, whereby for the identification of nuclear family disease clusters (parents with less then 3 children) the pedigree must be explored more deeply.

4.2 Paths between different disease clusters

Using the standard approach this 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. It was quite impossible to identify related clusters with similar disease patterns, because of missing and incomplete data.

During the explorative process paths between the different clusters could be highlighted. Moreover, missing and incomplete data was not a problem, since when working with the whole genealogy these parts could be inferred by human perception using the relevant visual information. However, preliminary findings have to be checked in an additional explorative stage to prevent false positives.

4.3 Genetic or Environment

The proper identification of ways in which individual genetic background and environmental factors interact to cause disease is challenging. Furthermore, once answered for a particular disease, this has important implications for the healthcare system in, for example, setting up genetic screening or public health education and prevention of disease programs.

Analytical methods are able to quantify the contribution of genetic and environmental factors in a complete data setting. Increased uncertainty about the type of disease, missing data and large populations require a large number of simulations and a specific focus on sub-pedigrees. 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 visual analytics approach does not quantify exactly, in terms of numbers, the influence of the various factors, but in consideration of possible ambiguity in certain self-administered questionnaire data this is acceptable. All analyses can be performed at a population level, allowing for the exploration of different hypotheses. Missing and uncertain data is balanced partially by the capabilities of human perception.

5 Discussion and Conclusion

Due to the complex relationship structures in isolated populations, classical analytical methods used in a subsequent visualization setting are only able to identify all local and/or single global population clustered diseases. Therefore, the distinction between genetic and environmental influences on disease onset and progression 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 disease settings. At the opposite pole, common diseases are often multifactorial and influenced by several environmental factors and life style.

The appropriate analysis requires the exploration of different settings, such as the inclusion of endophenotypes. These two tasks are poorly supported by standard analysis methods currently employed in isolated populations.

The visual analytics based approach is capable of performing family history analyses at a population level. Sometimes, the identification of small single local disease clusters can be difficult, due to the optimization necessary for global representation. However, using the interactive capabilities of our implementation,the identification of such disease clusters and their possible causes is supported in an explorative way.

As a result of extended genealogy and missing or uncertain data, exact calculations are not computationally feasible. Estimations and simulation are only doable for a limited set of hypotheses. Due to the complexity of the data, during the exploration the hypotheses on particular disease patterns may change significantly.

Integrating the power of the human perception, the visual analytic method provides an appropriated framework for the knowledge discovery process. Nonetheless, experts have to explore findings form different points of view to confirm their robustness and to prevent false positives.

References

[1] R. Bennett, K. Steinhaus, S. Uhrich, C. O’Sullivan, R. Resta, D. Lochner-Doyle, D. Markel, V. Vincent, and J. Hamanishi. Recommendations for standardized human pedigree nomenclature. *Journal of Genetic Counseling*, 4(4):267–279, 1995.

[2] Editorial. The germinating seed of arab genomics. *Nature Genetics*, (38):851, 2006.

[3] S. Hong and N. Nikolov. Layered drawings of directed graph in three dimensions. 2004.

[4] M. Kaufmann and D. Wagner. *Drawing graphs*. Springer, Berlin, Germany, 2001.

[5] D. Keim. Summary. workshop on visual analytics. 2005.

[6] K. Lange. *Mathematical and statistical methods for genetic analysis*. Springer, New-York, 1997.

[7] V.-P. Makinen, M. Parkkonen, M. W. P.-H. Groop, T. Kanninen, and K. Kaski. High-throughput pedigree drawing. *European Journal of Human Genetics*, (13):987–989, 2005.

[8] G. Mancosu, G. Ledda, and P. Melis. Pednavigator: a pedigree drawing servlet for large and inbred populations. *Bioinformatics*, 19(5):669–670, 2005.

[9] C. Pattaro, F. Marroni, A. Riegler, D. Mascalzoni, I. Pichler, C. Volpato, U. Dal Cero, A. De Grandi, C. Egger, A. Eisendle, C. Fuchsberger, M. Gogele, S. Pedrotti, G. Pinggera, S. Stefanov, F. Vogl, C. Wiedermann, T. Meitinger, and P. Pramstaller. The genetic study of three population microisolates in South Tyrol (MICROS): study design and epidemiological perspectives. *BMC Medical Genetics*, 8(1):29, 2007.

[10] K. Sugiyama, S. T. S, and M. Toda. Methods for visual understanding of hierarchical systems. *IEEE Transactions on Systems, Man, and Cybernetics*, 2:109–125, 1981.

[11] J. Thomas and K. Cook. A visual analytics agenda. *IEEE Computer Graphics and Applications*, 26(1):10–13, 2006.

[12] C. Yu and D. Zelterman. Statistical inference for familial disease clusters. *Biometrics*, 58:481–491, 2002.