Comparison of visualization methods of genome-wide SNP profiles in childhood acute lymphoblastic leukaemia

Ahmad Al–Oqaily[1] **Paul J. Kennedy**[1] **Daniel Catchpoole**[2] **Simeon Simoff**[3]

[1]Faculty of Engineering and IT,
University of Technology, Sydney,
PO Box 123, Broadway, NSW 2007,
Australia,
[2]The Oncology Research Unit,
The Children's Hospital at Westmead,
Locked Bag 4001, Westmead NSW 2145,
Australia,
[3]School of Computing and Mathematics
University of Western Sydney,
Locked Bag 1797,Parammata,
Australia,
Email: *aaoqaily@it.uts.edu.au*

Abstract

Data mining and knowledge discovery have been applied to datasets in various industries including biomedical data. Modelling, data mining and visualization in biomedical data address the problem of extracting knowledge from large and complex biomedical data. The current challenge of dealing with such data is to develop statistical-based and data mining methods that search and browse the underlying patterns within the data. In this paper, we employ several data reduction methods for visualizing genome–wide Single Nucleotide Polymorphism (SNP) datasets based on state–of–art data reduction techniques. Visualization approach has been selected based on the trustworthiness of the resultant visualizations. To deal with large amounts of genetic variation data, we have chosen to apply different data reduction methods to deal with the problem induced by high dimensionality. Based on the trustworthiness metric we found that neighbour Retrieval Visualizer (NeRV) outperformed other methods. This method optimizes the retrieval quality of Stochastic neighbour Embedding. The quality measure of the visualization (i.e. NeRV) showed excellent results, even though the dataset was reduced from 13917 to 2 dimensions. The visualization results will assist clinicians and biomedical researchers in understanding the systems biology of patients and how to compare different groups of clusters in visualizations.
Keywords: biomedical datasets, single nucleotide polymorphisms, SNP visualization.

1 Introduction

Data mining and knowledge discovery have been applied to datasets in various industries including biomedical informatics (Azuaje & Dopazo 2005). Data mining and visualisation in biomedical informatics addresses the problems of extracting knowledge from data originating from multiple sources, encoded in different formats or protocols, and processed by multiple systems. As identified by (Bertone & Gerstein 2001), the problems have only recently been reviewed in a systematic way (Azuaje & Dopazo 2005). The major challenges in data mining in the area stem from the fact that biomedical data requires data structures that are multidimensional. It is our intention to construct models which incorporate large amounts of biomedical data in a manner which will alleviate the error induced by high dimensionality.

Methods and approaches applied in this paper rely on the information extracted from biomedical datasets, derived from cancer patients. This data includes genome–wide single nucleotide polymorphism genotyping data (genetic variations).

The domain of this paper is Childhood Acute Lymphoblastic leukaemia (ALL), which is the most common childhood malignancy. It represents 24% of all new cancers that occurred in children between 1995 and 1999 (240 ALL/985 Cancer patients) (Coates & Tracey 2001). Nearly all children with ALL achieve an initial clinical remission, so the major obstacle to cure is patient relapse, i.e. the recurrence of evident disease. The approaches and methods we apply to ALL data can also be extended to other complex diseases such as heart diseases, diabetes and inflammatory diseases.

Information visualization is considered as a direct way to help browse the datasets. It is possible to combine visual exploration with other data exploration tools such as clustering analysis and data comparisons. The result of data explorations can be confirmed on the visualization. The main challenge in visualizing genetic variation datasets stems from the high dimensionality of the data, which may includes tens of thousands of SNPs. In this paper, several visualization methods will be applied to genetic variation datasets, for example manifold–based reduction methods.

Traditional dimensionality reduction techniques include Principal Components Analysis (PCA) (Hotelling 1933) which tries to preserve the variance in the data, and Multidimensional Scaling (MDS) which tries to preserve pairwise distances between data points. These methods are used to find a low space representation of the high dimensionality space which preserves the global structure of the data. However, these methods are not adequate to handle high dimensionality data which could have nonlinear relationships.

Therefore, in the last decade a large number

of nonlinear techniques for dimensionality reduction have been proposed. Some of these methods are used to find a lower dimensionality manifold of the data or a nonlinear embedding manifold space in the higher–dimensional data space. The main advantage of these methods is that they are able to preserve the local relationships of the data, which can be advantageous for the task of information visualization. Several of these methods will be described in section 4. The main difference between manifold estimation and visualization is that visualization is limited to two or three dimensions (Venna & Kaski 2007*a*). Thus, it is difficult to know the exact number of dimensions to uncover the underlaying structure of the data. Therefore, we need to apply different manifold–based methods on the given dataset in order to choose the most appropriate method.

In this work, we will study the results of applying different dimensionality reduction methods to genome–wide SNP profiles of leukaemia patients to determine which is the best method for visualizing this type of data. The results will be compared based on measures such as trustworthiness and continuity of the visualizations.

The rest of this paper is organized as follows. Section 2 points to related work that has been applied for visualizing biomedical data, specifically SNP data. Section 3 describes the dataset used in this study and the preprocessing steps applied to the data. Next, in section 4 we describe methods and techniques that will be used to visualize the ALL data. Section 5 describes in detail the experiments and results for different methods. In section 6 we further discuss these results. Finally, in section 7 we conclude the paper and describe the future directions for our research.

2 Related Work

The current interest in genetic variation studies is focused on disease–gene association analyses. Such analyses are important in identifying which variants are associated with a specific disease. Identification of genetic variants that contribute to susceptibility of diseases such as cancer will assist in the development of diagnostic and therapeutics (Carlson, Eberle, Kruglyak & Nickerson 2004). To identify these markers, at a statistically significant level, it is necessary to obtain genetic information from a large scale population sample of affected and unaffected individuals, which is termed a population based study. However, recent advances in biomedical technologies and genetics studies have made association studies a powerful approach for mapping complex–disease genes by conducting studies based on the whole genome. These studies are called genome–wide association (GWA) studies, in which a dense set of SNPs across the genome is genotyped to survey the most common genetic variations for a role in a disease or to identify the heritable quantitative trait that is a risk factor of a disease (Hirschhorn & Daly 2005).

Genome–wide association studies are mainly conducted using statistical methods, which are used to discover genetic factors that contribute to susceptibility to disease. Factors that show a significantly high statistical level of association are chosen for further analyses. However, in this paper, we are heading in a different direction. Data mining approaches will be used here. These approaches mainly concentrate on visualizing genome–wide SNP datasets based on state–of–art data reduction techniques. The visualization results will assist clinicians and biomedical researchers in understanding the different structure of patients and how to compare different group of patients' clustering in the visualization.

3 Data

In this section we describe in detail the dataset used and the preprocessing steps applied.

3.1 Single Nucleotide Polymorphism (SNP) data

The human genome was found to contain a large amount of genetic variation in the form of sequence polymorphisms. Polymorphism is a variation of DNA sequence that has an allele frequency of at least 1% of the population (Cavalli-Sforza 1974). There are several types of polymorphism in the human genome: SNPs, repeated polymorphisms and insertions or deletions, ranging from a single base–pair to thousands of base–pairs in size (Tabor, Risch & Myers 2002). Single Nucleotide polymorphisms (SNPs) are the simplest but most abundant type of genetic variation among individuals with between 1 to 10 million existing in the human genome (Donnelly 2004). These common SNP are thought to account for around 90% of human polymorphism (Carlson, Eberle, Rieder, Smith, Kruglyak & Nickerson 2003, Reich, Gabriel & Altshuler 2003).

Genetic variations, especially SNPs, are known to be the key feature of discovering disease–genes. In the case of complex disease, identifying multiple genetic variants would be possible by conducting association analysis between a specific variant and a disease. This association involves examining all genetic differences in a large number of affected individuals with unaffected controls (Risch & Merikangas 1996).

3.2 Genetic Variation and Childhood ALL

Chromosomal imbalances have long been known to be key features of leukaemia. Further, the human genome was found to contain a large amount of genetic variation in the form of sequence polymorphisms. Non–synonymous (ns)SNPs occurring within coding regions are those which produce an amino acid change but are not considered a "mutation" as a functional protein is still transcribed. Such nsSNPs are known to affect the functional efficiency of genes (Aplenc & Lange 2004). For example, drug metabolism and patient response to chemotherapy. SNP's which are found throughout non–coding intronic genome regions are used in major disease linkage and haplotyping studies including the HapMap Project (Altshuler, Brooks, Chakravarti, Collins, Daly, Donnelly et al. 2005) whilst identification of minor regions of amplification or deletion within the genome are facilitated through assessment of SNP copy number (Herr, Grützmann, Matthaei, Artelt, Schröck, Rump & Pilarsky 2005). However, genetic variation of the human genome is a promising resource for studying complex diseases such as cancer. Large number of genetic variations, scattered across the human genome, represent a remarkable opportunity to investigate the etiology, inter–individual differences in treatment response and outcomes of specific cancer such as leukaemia (Erichsen & Chanock 2004). Thus, we are in a position of utilizing such a tool (i.e. SNPs data) to analyze genetic contributions to complex diseases. Such analyses could have big influences on the prevention and early intervention strategies of a disease.

3.3 ALL Dataset

Genome–wide SNP data incorporates large scale mapping of SNPs and subsequent collation into databases. Generation of SNP data has been facilitated by high

throughput microarray–based technologies ((Barker, Hansen, Faruqi, Giannola, Irsula, Lasken, Latterich, Makarov, Oliphant, Pinter et al. 2004), (Leykin, Hao, Cheng, Meyer, Pollak, Smith, Wong, Rosenow & Li 2005), (Irving, Bloodworth, Bown, Case, Hogarth & Hall 2005)). DNA from a cohort of 139 childhood ALL patients are generated with the Illumina Bead Array system (Fan et al. 2003) using the non–synonymous beadchip to assess 13,917 SNPs across the genome within exon–centric loci.

3.4 Data preprocessing

The SNP dataset contains information about 13,917 SNPs which scattered across the whole genome. These SNPs are classified as non–synonymous (functional) SNPs which affect the functionality of genes. Each individual's genome has two alleles of a given SNP. For most cases there are two alleles for each SNP (Bi-allelic). At a specific SNP, a person can have one of the several genotypes. When they are the same the SNP is called homozygous and when they are different the SNP is called heterozygous. For a single SNP, one is designated the major allele and the other the minor allele, based on their observed frequency in a general population (Crawford & Nickerson 2005). Each SNP can have four different values (nominal): two homozygous, one heterozygous or missing. That is, the four possibilities for alleles A and B of the ith SNP are two homozygous (AA or BB), one heterozygous (AB) or missing NA (not determined). All SNPs were transformed into numerical data based on Minor Allele Frequency, as described in (Price, Patterson, Plenge, Weinblatt, Shadick & Reich 2006).

Let $\mathbf{G}$ be a matrix of genotype data, g_{ij} is the genotype for SNP i and individual j where $i = 1$ to M and $j = 1$ to N. The row mean $\mu_i = (\sum_j g_{ij})/N$ is subtracted from each entry in each row i, to obtain row sums equal to 0. Missing entries are excluded from the computation of μ_i and are subsequently set to 0. Each row i is then normalized by dividing each entry by $\sqrt{p_i(1-p_i)}$ where p_i is a posterior estimate of the unobserved underlying allele frequency of SNP i defined by

$$p_i = (1 + \sum_j g_{ij})/(2 + 2N) \quad (1)$$

with missing entries excluded from the computation. We denote the resulting matrix as G–normalized. The new matrix is regarded as normalized version of data matrix. The mean of each row i is equal to 0.

4 Methods and Approach

In this section, we describe some of the dimensionality reduction methods for visualizing the similarity relationship between patients. Firstly, we will describe the main classical methods for dimensionality reduction i.e. Principal Component Analysis (PCA), and Multidimensional Scaling (MDS). Some other methods based on MDS will be described. Then, other recently proposed methods that focus on finding the manifold or embedding of data will be described. Lastly, we will describe the measures for the goodness of visualization that we use in the experiments.

The problem of dimensionality reduction can be defined as follows. Given a dataset matrix $\mathbf{X} \in \Re^{N \times d}$ consisting of N data vectors x_i ($i \in 1, 2, ..., N$) with dimensionality d which can be considered as points in a high–dimensional data space. Dimensionality reduction methods transform the data set $\mathbf{X}$ with dimensionality d into a new data set $\mathrm{Y} \in \Re^{N \times p}$ with dimensionality p ($p << d$), while preserving the geometry of the data as much as possible. The low–dimensional representation of x_i is denoted by y_i, where y_i is the ith row of the p-dimensional data matrix $\mathbf{Y}$. For visualization purposes the dimensionality representation of the *output space* needs to be two or at most three dimensions, whereas the original space or *input space* can be thousands of dimensions.

Generally, the task of visualization methods is to construct a low-dimensional representation (i.e. output space) y_i of the input space, in such a way that the original relationships (or similarities) of the data are preserved. However, lower–dimensional representation of the data in 2 or 3–D dimensions might not be able to preserve all the information of the original (higher–dimensional) datasets and a compromise must be made by applying different data reduction methods and then selecting the best method based on how well a given method preserves the information of the original data (Venna & Kaski 2007*a*).

4.1 Principal Components Analysis

Principal components analysis (PCA) constructs a low–dimensional representation of the data that maximally preserves as much variance in the data as possible (Hotelling 1933). This is done by finding the linear projection or direction where the data has maximum variance. The projection can be found by solving the eigenvalue problem of the covariance matrix C_x of the data using the general eigen–decomposition problem

$$\mathbf{C}_x\mathbf{a} = \lambda\mathbf{a} \quad (2)$$

It can be shown that the linear projection is formed by the p principal components of the covariance matrix. The new representation of data points x_i can then be found by projecting (or mapping) the original data with

$$y_i = \mathbf{A}x_i \quad (3)$$

The low–dimensional data representations y_i of the data point x_i are computed by projecting data matrix X using matrix A, which contains the eigenvectors corresponding to the two or three largest eigenvalues. The new representation of the data can be visualized using the projected matrix $\mathbf{Y}$.

PCA has been successfully applied in a large number of domains. However, the main limitation of PCA is that it does not work well when the data lies in a nonlinear manifold. However, PCA is advantageous when the variance of the data is mainly concentrated in a few directions.

4.2 Multidimensional Scaling

Multidimensional Scaling (MDS) (Torgerson 1952) represents approaches that are commonly used with nonlinear mapping methods. There are several different variants of MDS (Cox & Cox 2001), but they all share a common goal which is to find the low–dimensional representation of the data that preserves the pairwise distance of the data as much as possible. The quality of the mapping is represented by a stress function (or cost function), which tries to minimize the errors of the pairwise distances between the low–dimensional and high–dimensional representations of the data.

The classical version of MDS is very closely related to PCA. The solution of linear MDS can be found by solving an eigen–decomposition problem. When the dimensionality of the sought space is the same and the distance measure is Euclidean distance, the projection of the original data using PCA is similar to the configuration of points that calculated by squared Euclidean distance matrix of the data (Gower 1966).

Other variants of MDS which have a more effective stress function are the raw stress function and Sammon cost function. The raw stress function can be defined by

$$\phi(Y) = \sum_{ij}(\|x_i - x_j\| - \|y_i - y_j\|)^2 \qquad (4)$$

where $\|x_i - x_j\|$ is the Euclidean distance between x_i and x_j in data points in the original data space, and $\|y_i - y_j\|$ is the Euclidean distance between y_i and y_j data points in the low–dimensional space. This cost function is able to find nonlinear relationships in the data. The Sammon cost function is slightly different to the raw stress in that it gives small distances a larger weight, which emphasises the local relationships in the data. In addition, there exist other variants of MDS, called non–metric MDS, which aim to preserve ordinal relations in data, rather than the pairwise distance (Kruskal 1964). Nevertheless, Multidimensional Scaling has been widely used for data visualization, such as Functional Magnetic Resonance Imaging (fMRI) analysis and molecular modelling (Tagaris, Richter, Kim, Pellizzer, Andersen, Ugurbil & Georgopoulos 1998, Venkatarajan & Braun 2004). The success of MDS has led to the proposal of new variants such as Curvilinear Component analysis (Demartines & Herault 1997) and Stochastic neighbours Embedding (SNE) (Hinton & Roweis 2003). These methods have shown the capability to produce good quality visualizations. Extended versions of these methods will be described in the following sections.

4.3 Stochastic neighbour Embedding

Stochastic neighbour Embedding (SNE) proposed by (Hinton & Roweis 2003) is a probability–based embedding method. SNE tries to find the low–dimensional representation of data points that preserve neighbourhood identities. The SNE algorithm tries to preserve the probability distribution of the pairwise distances of data points in the input space, so that the probability of a data point i being a neighbour of point j in the output space is the same as in the input space.

For each data point x_i and its potential neighbours, X_j, the algorithm starts by computing p_{ij}, the probability that point x_i and x_j are neighbours in the input space using

$$p_{ij} = \frac{exp(-d(x_i, x_j)^2)}{\sum_{i \neq k} exp(-d(x_i, x_k)^2)} \qquad (5)$$

where $d(x_i, x_k)^2$ is the pairwise distance between data points i and j. The distance can simply be the squared Euclidean distance or it can be the scaled squared Euclidean distance if we have a high–dimensional data

$$d(x_i, x_k)^2 = \frac{\|x_i - x_j\|^2}{2\sigma_i^2} \qquad (6)$$

In low–dimensional output space the images y_i of all data point x_i is defined as q_{ij}, which express the probability of the point y_i being a neighbour of point yyz_j.

$$q_{ij} = \frac{exp(-d(y_i, y_j)^2)}{\sum_{i \neq k} exp(-d(y_i, y_k)^2)} \qquad (7)$$

The aim of the embedding is to match the two probability distributions p_{ij} and q_{ij} as well as possible. The embedding of points y_i can be achieved by minimizing a cost function which is the Kullback–Leibler divergence between the probability distribution of the input (p_{ij}) and output (q_{ij}) distribution over neighbours of each data point. The cost function is

$$E_i[D(p_i, q_i)] = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}} \qquad (8)$$

Stochastic neighbour Embedding has been successfully applied to several datasets (eg. (Nguyen & Worring 2004) or (Memisevic & Hinton 2005)). Results show that good optima can be achieved.

Stochastic neighbour Embedding was originally designed as a data reduction method that tries to preserve neighbourhood identities. However, SNE can be also seen as an information retrieval algorithm. A new restructured method called neighbour Retrieval Visualizer (NeRV) was proposed by (Venna & Kaski 2007*b*). This method is motivated by visual neighbour retrieval, unlike SNE, which tries to optimize recall (i.e. misses). The method balances the error caused by precision (i.e. false positive, see section 4.7).

In information visualization, high precision is more important than recall. Minimizing precision is associated with preserving the neighbourhood of points in the output space. Recall on the other hand, tries to preserve the neighbourhood of points in the input space. Stochastic neighbour Embedding updates the original SNE method by assigning a relative cost λ to recall and $(1 - \lambda)$ to precision. Then, the total function to be optimized is

$$\begin{aligned} E &= \lambda E_i[D(p_i, q_i)] + (1 - \lambda) E_i[D(q_i, p_i)] \\ &= \lambda \sum_{i,j} \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}} + \\ &\quad (1 - \lambda) \sum_i \sum_j q_{ij} \log \frac{q_{ij}}{p_{ij}} \end{aligned} \qquad (9)$$

That is, by setting the parameter $\lambda \in [0, 1]$ the choice can be focused on either the probabilities that are high in the input space (recall) or in the output space (precision). When $\lambda = 1$ the method is equal to SNE and when $\lambda = 0$, the method focuses completely in avoiding false positives (precision). This method can be described as retrieving points based on the visualization display. In our experiment we apply this method with choice of λ that emphasizes the underlying structure of the data that maximizes precision. In addition, SNE will be applied for comparison purposes.

4.4 Curvilinear Component Analysis

Curvilinear Component Analysis (CCA) (Demartines & Herault 1997) is a variant of MDS. Whereas MDS tries to find the configuration of points that preserve the pairwise distances as much as possible, CCA tries to find the configuration of points that preserve a subset of the distances that are neighbours in the output space. The cost function of CCA concentrates on preserving the distance of points in the reduced space. This can be done by introducing a weighted function F that depends on the distance between the points in the output space (or visualization), yielding a cost function

$$E = \frac{1}{2} \sum_i \sum_{i \neq j} (d(x_i, x_j) - d(y_i, y_j))^2 F(d(y_i, y_j), \sigma_i) \qquad (10)$$

Generally, $F(d(y_i, y_j), \sigma_i)$ is chosen as a bounded and monotically decreasing function, in order to favor preserving the local geometry of the data. Decreasing exponential, sigmoid, or Lorentz functions can be suitable choices, and a simple step function can also be applied.

$$F(d(y_i, y_j), \sigma_i) = \begin{cases} 1 & Y_{ij} \leq \sigma_i \\ 0 & Y_{ij} > \sigma_i \end{cases} \quad (11)$$

The minimization of the cost function can be achieved using a form of stochastic gradient decent algorithm. During the optimization process, σ_i is set to cover all or at least most of the data points (as the case of MDS), and it is slowly decreased to reach the optimal value.

Curvilinear Component Analysis has been successfully applied to various nonlinear–dimensionality problems in data representation such as for gene expression data and computer vision (Buchala, Davey, Frank & Gale 2004, Venna & Kaski 2007*a*). An extension of CCA, Curvilinear Distance Analysis (CDA), was introduced by (Lee, Lendasse & Verleysen 2004). The main difference of CDA compared to CCA is to replace the Euclidean distance used by CCA with geodesic distance. Geodesic distance is based on graph theory and uses the minimum spanning tree to find the distance.

The main drawback of CCA is that the cost function may have several local optima. Although this can cause undesired results when applying CCA, solutions found by CCA have showed quite reasonable results (Venna & Kaski 2007*a*).

Recently, a method called Local Multidimensional Scaling (LocalMDS) was proposed (Venna & Kaski 2006). This method is regarded as a derivative of CCA. Similarly to NeRV, LocalMDS has the indirect ability to control the tradeoff between precision and recall, which helps for data visualization. The cost function of CCA tries to preserve the distance of points that are neighbours in the output space, by ignoring the error in distance between points that are far from each other in the reduced space. Thus, CCA could increase the errors caused by recall, which can result in lower visualization quality. In LocalMDS, a term is added to the cost function to increase recall. This can be achieved by penalizing the errors of distance between points that are close by in the input space. The tradeoff between the two types of errors helps in having a more efficient display of the local similarities of the data. The cost function of LocalMDS is defined as

$$E = \sum_i \sum_{i \neq j} [(1-\lambda)(d(x_i, x_j) - d(y_i, y_j))^2 F(d(y_i, y_j), \sigma_i) + \lambda (d(x_i, x_j) - d(y_i, y_j))^2 F(d(x_i, x_j), \sigma_i)] \quad (12)$$

where $\lambda \in [0, \ldots, 1]$ controls the tradeoff between precision and recall. During the optimization the radius of the area of influence around data point x_i, σ_i, is slowly reduced to reach the optimal value. $F(d(x_i, x_j), \sigma_i)$, similarly to CCA, emphasizes the local distance in the input space. F is equal to one when $d(x_i, x_j) < \sigma_i$ and 0 otherwise. The final radius is set equal to the distance of k–NN of a data point x_i in the orginal space.

When $\lambda = 0$ the cost function will be that of the basic CCA method. A good choice of λ ranges from 0 to 0.5. The cost function can be optimized using stochastic gradient descent methods similarly to CCA. In our experiments we apply LocalMDS with s choice of λ that emphasizes the underlying structure of the data to maximize precision. In addition, CCA will be applied for comparison purposes.

4.5 Laplacian Eigenmap

Laplacian Eigenmap (LE) finds a low–dimensional representation of data by preserving the local structure of the data (Belkin & Niyogi 2002). Laplacian Eigenmap is regarded as a geometrically motivated dimensionality reduction. The output space reflects the intrinsic geometric structure of the manifold. In Laplacian Eigenmap, the local structure can be preserved by keeping the local structure between each datapoint and its k nearest neighbours. Therefore, the local structure of LE algorithms can be relatively insensitive to outliers and noise, and as a result the algorithm implicitly emphasizes the natural clusters in the data (Belkin & Niyogi 2002).

Laplacian Eigenmap computes a low–dimensional representation of the data in which the nearest neighbours of a datapoint in the original space should be mapped to nearest neighbours of that datapoint in the reduced space (He, Yan, Hu, Niyogi & Zhang 2005). This can be done in a weighted manner applied to graph partitioning, i.e., using a weighted criterion such as a heat kernel (Gaussian function) enables us to choose the weight of the graph in such a way that keeps the local similarity of the graph. The embedding map is constructed by computing the eigenvectors of the graph Laplacian. The algorithm's procedures are as follows.

The LE algorithm first constructs the adjacency graph G in which every node (datapoint) x_i is connected to its k nearest neighbours. For all nodes i and j in the graph G that are connected by an edge, a weight is calculated using different methods such as a Gaussian kernel or a simple approach where $W_{ij} = 1$ if node i and j are connected by an edge. This leads to a sparse matrix W in which $W_{ij} > 0$ if node i and j are connected and $W_{ij} = 0$ otherwise.

To compute the low–dimensional representation $y = y_1, y_2, \ldots, {y_n}^T$, Laplacian Eigenmap minimizes the following objective function

$$\phi(Y) = \sum_{ij} (y_i - y_j)^2 w_{ij} = tr(\mathbf{Y}^T \mathbf{L} \mathbf{Y}) \quad (13)$$

where $\mathbf{L} = \mathbf{D}$ - $\mathbf{W}$ is the Laplacian matrix, $\mathbf{D}$ is diagonal matrix, with elements $D_{ii} = \sum_j W_{ij}$ being the column (or row, since W is symmetric) sums of W. The Laplacian matrix is symmetric and positive semidefinite.

Minimizing the objective function tries to put datapoints that are connected in the graph G as close together as possible. There is a trivial solution to the objective function which collapses all the new representations of the graph G into a single location. This can be prevented by adding an orthogonality constraint $\mathbf{Y}^T \mathbf{D} \mathbf{Y} = 1$.

The configuration of points in the low–dimensional space can be solved by finding the eigenvectors and eigenvalues of the generalized eigenvector problem

$$\mathbf{L} y = \lambda \mathbf{D} y \quad (14)$$

The low-dimensional embedding of the original data points can be formed by the d eigenvectors y_i that correspond to the smallest non–zero eigenvalues, after discarding the smallest eigenvector that corresponds to the zero eigenvalues, which represent the case where all data points are represented by a single location.

Laplacian Eigenmap has been successfully applied to number of domains such as clustering and face recognition (Ng, Jordan & Weiss 2002, Shi & Malik 2000, He et al. 2005). Variants of Laplacian Eigenmaps have been extended to supervised and semi–supervised data analysis (Costa & Hero 2005, Belkin & Niyogi 2004). A linear variant of Laplacian Eigenmap is proposed by (He & Niyogi 2004).

Laplacian Eigenmap has two main drawbacks. Firstly, in most applications it is not possible to see the structure within clusters from the visualization. Secondly, this method is mainly used for data representation or visualization and can not compute the projection for a new test point. However, this problem can be solved using techniques proposed by (Bengio, Paiement & Vincent 2004) called an out–of–sample extension.

4.6 Locally linear Embedding (LLE)

The LLE algorithm (Roweis & Saul 2000) is similar to Laplacian Eigenmap, which tries to preserve the local geometry of the data by finding the *local linear* approximation of the manifold. This is based on the assumption that a data point and its neighbours lie in or close to a locally linear subspace on the manifold. In LLE, the local geometry of this subspace can be characterized by calculating the linear coefficients (weights) that reconstruct each data point from its k nearest neighbours. In the low–dimensional space of the data, LLE attempts to retain the reconstruction weights in the linear combination as much as possible (van der Maaten, Postma & van den Herik 2007).

The algorithm works in two stages. First, the local coordinate of each data point is calculated based on its k nearest neighbours, and the total reconstruction error to be optimized is then measured by the cost function

$$\epsilon(W) = \sum_{i=1}^{N} \left| X_i - \sum_{j=1}^{k} W_{ij} X_j \right|^2 \qquad (15)$$

which adds up the squared distance between all data points and their reconstruction. The weight W_{ij} summarizes the contribution of the jth data point to the ith reconstruction. The reconstruction error is minimized subject to the constraints that $W_{ij} = 0$ if datapoints i and j are not neighbours and $\sum_j W_{ij} = 1$.

In the second stage, the task is to find the low–dimensional representations y_i that preserve the local geometry of the data as described by the local coordinate of each data point. In other words, the reconstruction weights W_{ij} that reconstruct each datapoint x_i from its neighbours in the high–dimensional data space also reconstruct each data point y_i in the low–dimensional space. To do so, the p–dimensional reduced space $\mathbf{Y}$ can be computed based on minimizing the cost function

$$\epsilon(Y) = \sum_{i=1}^{N} \left| y_i - \sum_{j=1}^{k} W_{ij} y_j \right|^2 \qquad (16)$$

(Roweis & Saul 2000) showed that the optimization function described in (16) can be solved by the eigenvectors that correspond to the p nonzero eigenvalues of matrix $\mathbf{M}$, where $\mathbf{M} = (\mathbf{I} - \mathbf{W})^T(\mathbf{I} - \mathbf{W})$ and $\mathbf{I}$ is the identity matrix.

A linear variant of LLE algorithm was proposed recently (Kokiopoulou & Saad 2005, Kokiopoulou & Saad 2007).

4.7 Comparing visualisations

As we have discussed, the first step in exploring the structure of a given dataset is to have the data visualized. In many previous works, visualization methods are compared through examining the produced figures. Some quantitative criteria should be designed, to compare the visualization results without considering the human as a part of visualization.

One of the crucial tasks in data visualization is how to assess the quality of produced visualizations or the tools that are used. The quality measure is used to assess how well the visualization of a given tool can represent the underlying data. The local structure of the data is the most important component of the visualization. The usability of the visualization can be measured by how accurately the data is represented and how readable it is.

The first question that comes to mind is how trustworthy is the visualization. The local similarity or structure of a data is the most crucial part of the visualization. When looking at the visualization, the first insight is how points are similar and how points group together. Looking at a visualization a user can possibly get insight into some question such as, are the unknown data points similar to the known ones? How is the data clustered? Are there denser areas and more sparse ones? Questions like these cannot be answered without having a visualization that is capable of answering these questions.

There are a number of methods that have been implemented to assign a quantity to a visualization. Some of these methods calculate the correlation coefficient between the distance vectors (i.e., the vectors that compare the distance between all pairs of points) of the original space with that of the lower dimensional space. It was proven that this measure can provide a good measurement of quality of the visualization procedures (Tan, Steinbach & Kumar 2005).

Others methods measure how trustworthy the local structure of the visualizations is (Kaski, Nikkila, Oja, Venna, Toronen & Castren 2003, Venna & Kaski 2001). Based on these methods the low–dimensional representation is trustworthy if the k nearest neighbours of a point in the reduced space (or in the visualization) are also neighbours of the point in the original space. The proportion of points that are in the neighbourhood in the visualization but not in the original space is quantified as the precision (or loss of precision, i.e., one minus precision). This number is usually not informative. However, the magnitude of the error can be used to rank the data points based on their distance instead of just counting the number of errors.

Reducing the dimensionality of a data can result in losing some of the similarity relationships between data points. Two general errors can be caused in applying a reduction method. First, data points that are not neighbours in the input space can be mapped close by in the reduced space, causing points to be incorrectly identified as neighbours. These kind of errors can reduce the *precision*. Secondly, data points that are neighbours in the input space can be mapped far away in the reduced space, causing discontinuities in the mapping and can distort the neighbour relations. This kind of error is called *recall*. The two kind of errors (i.e. precision and recall) are used in information retrieval literature in which the error is quantified based on the proportion of the points that caused the errors.

The main limitation of using *precision* and *recall*, as it is used in information retrieval, is that each of the errors is equally bad. However, in the visualization context this kind of measurement is not intuitive, whereas the distance between data points are known.

Intuitively, a data point that comes into the neighbourhood of another from far away causes a larger error than one that comes from closer. By ranking data points based on their similarity we can have two new quality measures: *trustworthiness* and *continuity* (Kaski et al. 2003) which quantify the errors of a visualization tool by the neighbourhood ranks of each data point.

The *trustworthiness* of a visualization can be defined as follows. Let N be the number of data samples and $r(x_i, x_j)$ be the rank of the sample x_j in the ordering according to the distance from data sample x_i in the original space. Let $U_k(x_i)$ be a set of data samples of size k that are in the neighbourhood of sample x_i in the visualization space but not in the original space. The measure of trustworthiness is defined as

$$M_{Tru}(k) = 1 - A(k) \sum_{i=1}^{N} \sum_{j \in U_k(i)} (r(x_i, x_j) - k) \quad (17)$$

where $A(k) = 2/(Nk(2N-3k-1))$ scales the measure between zero and one. The errors reach the maximum value when the ranks in the input and output space are reversed. The trustworthiness measure is closely related the precision (as in information retrieval). However, the trustworthiness measure is a special kind of precision measure for the case where the objects are ranked based on their relevance (Venna & Kaski 2006).

On the other hand, discontinuities are used to quantify whether neighbours in the original space remain neighbours in the visualization. If neighbour's points are pushed out in the displayed visualization, discontinuities arise in the visualization. The errors caused by discontinuities may be quantified similarly to the errors caused by trustworthiness.

Let $V_k(x_i)$ be the set of data samples that are neighbours of the data sample x_i in the original space but not in the output space and $\hat{r}(x_i, x_j)$ be the rank of data sample x_j in the ordering according to the distance from x_i in the visualization. The effects of discontinuities of the mapping are measured by:

$$M_{disc}(k) = 1 - A(k) \sum_{i=1}^{N} \sum_{j \in V_k(i)} (\hat{r}(x_i, x_j) - k) \quad (18)$$

Therefore, a data sample that is mapped far away from the neighbourhood in the reduced space will cause a larger error than a data sample that was mapped just out of the neighbourhood. In the recall measure both errors are considered equally severe.

The trustworthiness and continuity measures can be used to assess the quality of a data reduction method or to compare the performance of different data reduction method on a data set for visualization. Based on the quality results, the best run, method or parameters for a data set can be selected (Venna & Kaski 2007*b*). Comparing the data reduction methods can be tested on a large range of neighbourhood size. Small k can be important for the quality of visualization but a range of neighbourhood size can give an overview of the overall performance of different methods.

The performance of different data reduction methods is made by plotting the trustworthiness and continuity measures as a function of the neighbourhood size k. In any given data reduction method, a tradeoff must be made between trustworthiness and continuities. Seeking for a high trustworthiness will typically lead to a lower continuity and vice versa.

5 Experiments and Results

The purpose of the experiments is to gain insight and to understand the behavior of different dimensionality reduction methods in biomedical data and, more specifically, SNP data of children with acute lymphoblastic leukaemia. As described above, a dataset of 139 patients with 13917 non-synonymous SNPs (dimensions) was used.

The performance of dimensionality reduction methods will be compared on visualizing the SNP dataset. The following methods will be included in our experiments: Principal Component Analysis (PCA), Laplacian Eigenmap (LE), Locally Linear Embedding (LLE) and methods based on Multidimensional scaling, which include an extended version of Curvilinear Component Analysis (CCA) called Local MDS (LocalMDS) (Venna & Kaski 2006) and an extended version of Stochastic neighbour Embedding (SNE) called neighbour Retrieval Visualizer (NeRV) (Venna & Kaski 2007*b*). In the following subsection, the experimental settings and the results of experiments on SNP dataset are described.

5.1 Experimental setting

All methods except PCA have a parameter k for setting the number of nearest neighbours. This parameter was tested with values of k ranging from 5 to 30. The best k was selected based on the best result. However, small neighbourhood size can be related to the data points that are most likely to be relevant. The performance of the resulting visualizations was tested based on the trustworthiness and continuities of the reduced dimension, as described in section 4.7.

Some of the applied methods such as LocalMDS and NeRV that may fall into local optima were run several times (in our case 10 times) with different random initialization and the best run was selected. Random mapping was computed based on the average of 10 different random projections. Two types of distance metric were used to calculate the distance of data in the input space: Euclidean distance and Gaussian function. In this study, we employed both of these with different parameters and we set the dimensions of the output space equal to two for visualization purposes.

5.2 Results

Data reduction methods were compared using the trustworthiness and continuity measures of the resulted visualization. Figure 1 and 2 shows the trustworthiness and continuity results of the applied methods. The following subsection will get insight on different aspect of the results. For visualization purposes, trustworthiness is more important than continuity. In each case the result with the best trustworthiness was reported.

5.2.1 Trustworthiness and continuity

Trustworthiness and continuity are the first aspects that we examined. In terms of exploring the result of a visualization, the local neighbourhood of each data point is the first insight a human analysis looks at. Therefore, a visualization is trustworthy if the visualization preserves small neighbourhoods as much as possible. Thus, attention should be paid to small sizes of k (e.g. k between 5 and 15). It is clear from figure 1 that, in terms of trustworthiness, the NeRV method is the best. Unexpectedly, PCA is also quite good at preserving the locality of the reduced neighbourhood (Trustworthiness). On the other hand, state–of–art

data reduction methods such as LLE and LE were not able to produce reasonable results. In fact, LE was the worst method compared in our experiments on this dataset and is the most simialr one to random mapping. The LLE and LE results suggests that the minimum number of dimensions that are required to uncover the manifold of the data is greater than two.

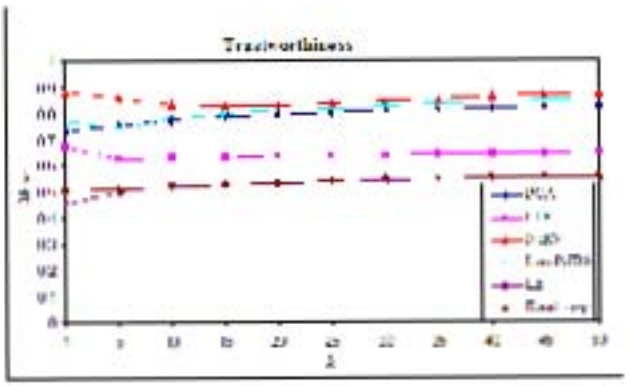

Figure 1: Trustworthiness of the mapping as a function of k that applied to SNP dataset, where k is the size of neighbourhood. Small neighbourhood sizes are the most important ones. PCA: Principal Component Analysis, LLE: Locally Linear Embedding, NeRV: neighbour Retrieval Visualizer, LocalMD: Local Multidimensional scaling, LE: Laplacian Eigenmap rand_map: random mapping.

In this initial analysis LocalMDS was expected to perform similarly to NeRV method but the result shows slightly different behavior. In this data set PCA performs similarly to LocalMDS, even though the original cost function of LocalMDS emphasizes trustworthiness and should perform better.

In terms of continuity, as can be seen in figure 2, the result of NeRV method is again the best. But this time LocalMDS is slightly better than PCA which is different than the case of Trustworthiness where both methods are similar. Once more, manifold–based methods, LLE and LE, perform very badly on this data set. In section 6 we will suggest reasons why this occurred.

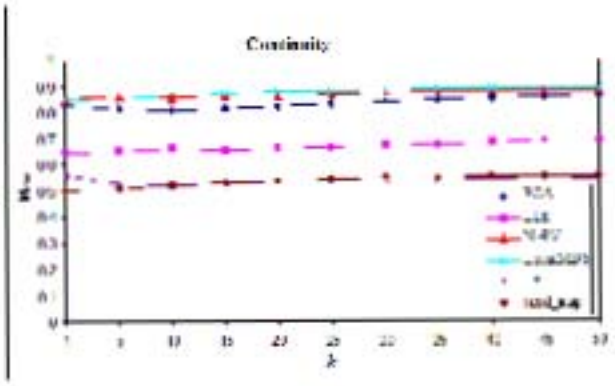

Figure 2: Continuity of the mapping as a function of k that applied to SNP dataset, where k is the size of neighbourhood. Small neighbourhood sizes are the most important ones. PCA: Principal Component Analysis, LLE: Locally Linear Embedding, NeRV: neighbour Retrieval Visualizer, LocalMDS: Local Multidimensional scaling, LE: Laplacian Eigenmap, rand_map: random mapping.

For the NeRV method, different parameters were set to explore the performance of this method on the dataset. As can be seen in figure 3, we have applied the NeRV method for different neighbourhood sizes ranging from 5 to 30. If a small size of neighbourhood is considered around each point on the output space, a neighbourhood size of 5 or 15 produces the best results on this data. The same results were also found by LocalMDS, as can be seen on figure 4. Next, we set $k = 15$, and ran NeRV on a range of $\lambda = 0$ to 1. The result can be seen in figure 5. This shows that the best trustworthiness occurs when λ equals 0.0 or 0.1. This result confirms the performance of NeRV compared to SNE, where λ equals one, which is the case when NeRV is equivalent to SNE, the trustworthiness attains the lowest performance. Thus, NeRV shows the capability of producing a high trustworthy visualization result which is based on balancing the tradeoff between the continuity and trustworthiness of visualization. A large value of λ, close to one, gives a lower trustworthy result and vice versa.

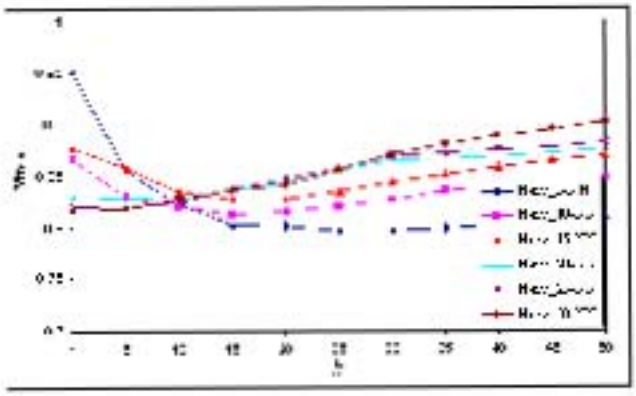

Figure 3: Trustworthiness of NeRV mapping as a function of k that applied to SNP dataset, where k is the size of neighbourhood. The neighbourhood size used by NeRV is ranging from 5 to 30.

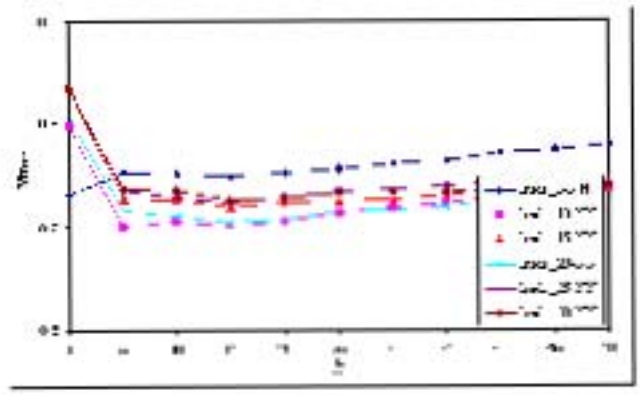

Figure 4: Trustworthiness of LocalMDS mapping as a function of k that applied to SNP dataset, where k is the size of neighbourhood. The neighbourhood size used by LocalMDS is ranging from 5 to 30.

5.2.2 Euclidean distance and Gaussian function

In our experiments two types of dissimilarity measure were used: Euclidean distance and Gaussian function. In the case of Gaussian function a parameter σ is used as a control parameter. The parameter σ was set to 0.001, 0.01, 0.1, 0.5, 1, 10, 100, 200, 500, 1000 and 2000 for different runs. The settings of $\sigma = 0.1$ and 0.5 gave the best performance (result not shown). On

our data set the use of Euclidean distance seems to give slightly similar results to the Gaussian function. This is can be due to the high–dimensionality of the data.

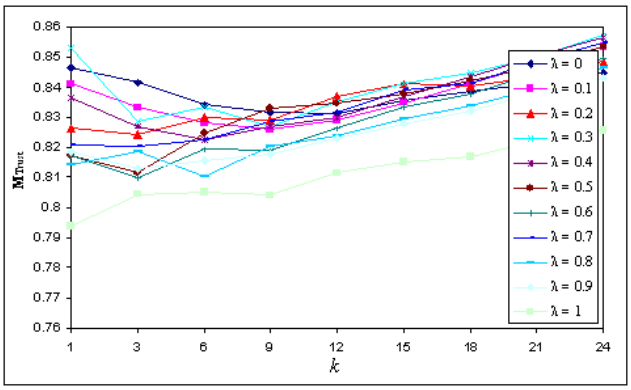

Figure 5: Trustworthiness of NeRV mapping as a function of k that applied to SNP dataset, where k is the size of neighbourhood. The Lambda used by NeRV is ranging from 5 to 30.

5.2.3 Quality of the visualization

The first thought that comes to mind is that reducing the dimensionality of the data from 13917 down into just two dimensions will not show how the data points are similar to each other in the high–dimensional space. On the other hand, without the use of a quality measure, we will not be able to assess the quality of different data reduction methods. In the previous sections, we summarized the different data reduction methods that we have employed, with different parameter settings. The comparison of the produced results was calculated in term of trustworthiness and continuity measures. The best method was selected based on the balance between these two measures and more emphasis was put on the trustworthiness measure.

Based on the results of the methods and parameter settings, the neighbour Retrieval Visualizer method, with k nearest neighbour equal to 15 and $\lambda = 0.1$, produced the best result. Figure 6 shows the visualization of data with the NeRV method. From the visualization we can see different clusters of data points (patients). The left most point in Figure 6 marks two outliers of patients sitting on top of one another. The clusters of patients require further scrutiny by domain experts. In contrast, the visualization produced by LocalMDS, as can be seen in figure 7, does not show any kind of structure on the data. This result confirms the ability of NeRV to produce better results.

6 Summary and discussions

In Summary, different data reduction methods were utilized for visualizing genetic variation (SNP) data as a way to discover the underlying relationships between patients. State–of–art data reduction methods have been employed. The result was selected based on the trustworthiness of the visualizations. The task of visualization was formulated as an information retrieval problem where the result of the visualization describes the local structure of the data. The quality measure of the visualization is tested based on a quantitative error of the number of misses and false positives.

We tested several different dimensionality reduction methods. These include PCA, Laplacian Eigen-

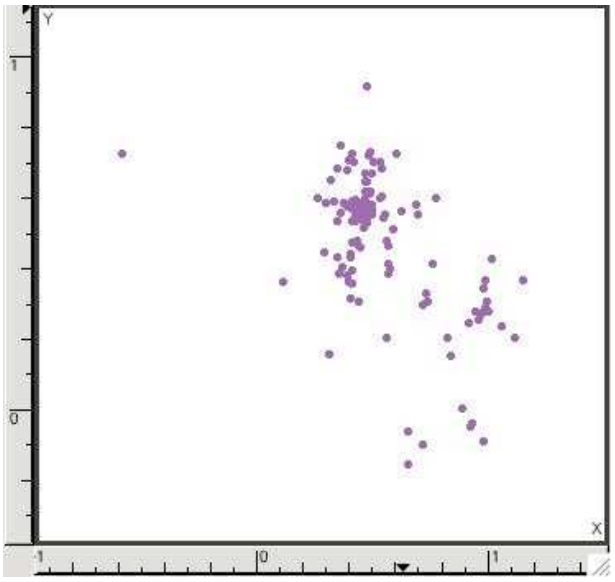

Figure 6: Visualization of SNP data using NeRV method with $k = 15$ and $\lambda = 0.1$.

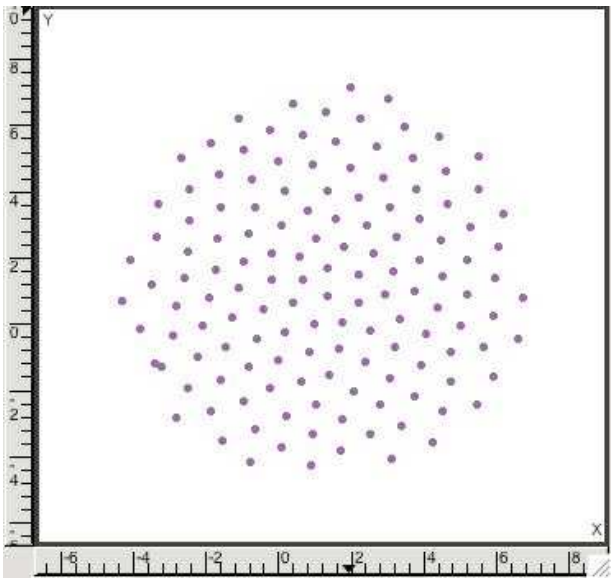

Figure 7: Visualization of SNP data using LocalMDS method with $k = 15$ and $\lambda = 0.2$

map (LE) and Locally Linear Embedding (LLE) that are designed to extract data manifolds and more generally methods that are designed for dimensionality reduction, Stochastic neighbour Embedding (SNE) and Curvilinear Component analysis (CCA). Recently, extended version of the SNE and CCA methods called neighbour Retrieval Visualizer and Local MDS methods, respectively, were introduced. These data reduction methods were run on the data with different parameter settings. An extended method of Stochastic neighbour Embedding (SNE) called neighbour Retrieval Visualizer (NeRV) has shown the best performance on this data set. This method balances the tradeoff between the trustworthiness and continuity of the visualization. The result shows that a neighbourhood of size 15 was the best for our data. A parameter λ which controls the tradeoff between trustworthiness and continuity was selected to be 0.1. This parameter emphasizes the trustworthiness of the visualization which is more important for visualization.

The result did not show any differences be-

tween using different distance matrices (Euclidean and Gaussian function) due to high–dimensionality of the data. Manifold–based data reduction methods, i.e. LLE and LE, perform surprisingly badly and the Laplacian Eigenmap method similarly performs worse as random mapping of the data. This result was not expected for these method due to the high performance of these methods in other datasets. We hypothesize that the reason is that these methods are designed to discover the intrinsic dimensionality of the data manifold which can be more than two dimensions. Lastly, the performance of PCA was comparable to the result of LocalMDS although the latter method is considered as a nonlinear dimensionality reduction method. This behavior suggests that the dataset has a linear relationship, which is difficult to comprehend due to the high–dimensionality of the data.

7 Conclusion and Future Work

In this paper, we employed several data reduction methods for visualizing a biomedical dataset. This dataset describes the genetic variation of Acute Lymphoblastic leukaemia patients. Visualization approaches were compared based on the trustworthiness metric of the resultant visualization. To deal with large amounts of genetic variation data, we have chosen to compare the performance of different dimensionality reduction methods on the given dataset. Based on this comparison neighbour Retrieval Visualizer (NeRV) showed the best results and outperformed other methods. Even though the dimensionality of the dataset was reduced from 13917 to 2 dimensions, the quality measure of the visualization (i.e.NeRV) still shows excellent results. The visualization results will assist clinicians and biomedical researchers in understanding the different structure of patients and how to compare different group of clustering in the visualization.

The result from using NeRV shows the feasibility of this method in visualizing genetic variation data. The main limitation of the employed methods is the distance measure that has been used (Euclidean distance or Gaussian function). These methods might not be appropriate for the given dataset due to the high–dimensionality of the data. The future direction of this work is to employ other distance measures that can be more appropriate in discriminating the major characteristics of the dataset. In particular, prior knowledge or domain–driven dissimilarity measures may improve the performance of the data reduction methods in the examined dataset.

8 Acknowledgments

This work was supported by the Australian Rotary Health Research Fund (ARHRF). ARHRF/District 9680 Funding partners scholar.

References

Altshuler, D., Brooks, L., Chakravarti, A., Collins, F., Daly, M., Donnelly, P. et al. (2005), 'A haplotype map of the human genome', *Nature* **437**(7063), 1299–1320.

Aplenc, R. & Lange, B. (2004), 'Pharmacogenetic determinants of outcome in acute lymphoblastic leukaemia', *British Journal of Haematology* **125**(4), 421–434.

Azuaje, F. & Dopazo, J. (2005), *Data analysis and visualization in genomics and proteomics*, Hoboken, NJ: John Wiley & Sons, Ltd.

Barker, D., Hansen, M., Faruqi, A., Giannola, D., Irsula, O., Lasken, R., Latterich, M., Makarov, V., Oliphant, A., Pinter, J. et al. (2004), 'Two Methods of Whole-Genome Amplification Enable Accurate Genotyping Across a 2320-SNP Linkage Panel', *Genome Research* **14**(5), 901.

Belkin, M. & Niyogi, P. (2002), 'Laplacian eigenmaps and spectral techniques for embedding and clustering', *Advances in Neural Information Processing Systems* **14**, 585–591.

Belkin, M. & Niyogi, P. (2004), 'Semi-Supervised Learning on Riemannian Manifolds', *Machine Learning* **56**(1), 209–239.

Bengio, Y., Paiement, J. & Vincent, P. (2004), 'Out-of-Sample Extensions for LLE, Isomap, MDS, Eigenmaps, and Spectral Clustering', *Advances in Neural Information Processing Systems 16: Proceedings of the 2003 Conference* .

Bertone, P. & Gerstein, M. (2001), 'Integrative data mining: the new direction in bioinformatics', *Engineering in Medicine and Biology Magazine, IEEE* **20**(4), 33–40.

Buchala, S., Davey, N., Frank, R. & Gale, T. (2004), 'Dimensionality reduction of face images for gender classification', *Intelligent Systems, 2004. Proceedings. 2004 2nd International IEEE Conference* **1**.

Carlson, C., Eberle, M., Kruglyak, L. & Nickerson, D. (2004), 'Mapping complex disease loci in whole-genome association studies', *Nature* **429**, 446–452.

Carlson, C., Eberle, M., Rieder, M., Smith, J., Kruglyak, L. & Nickerson, D. (2003), 'Additional SNPs and linkage-disequilibrium analyses are necessary for whole-genome association studies in humans', *Nature Genetics* **33**(4), 518–521.

Cavalli-Sforza, L. (1974), 'The genetics of human populations.', *Sci Am* **231**(3), 80–9.

Coates, M. & Tracey, E. (2001), 'Cancer in New South Wales. Incidence and mortality 1998 and Incidence for Selected Cancers 1999', *NSW Central Cancer Registry, Cancer Research and Registers Division, NSW Cancer Council* pp. 42–43.

Costa, J. & Hero, A. (2005), 'Classification Constrained Dimensionality Reduction', *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing* **5**.

Cox, T. & Cox, M. (2001), *Multidimensional Scaling*, CRC Press.

Crawford, D. & Nickerson, D. (2005), 'Defeinition and Clinical importance of Haplotypes', *Annual Review of Medicine* **56**(1), 303–320.

Demartines, P. & Herault, J. (1997), 'Curvilinear component analysis: a self-organizing neural networkfor nonlinear mapping of data sets', *Neural Networks, IEEE Transactions on* **8**(1), 148–154.

Donnelly, J. (2004), 'Pharmacogenetics in Cancer Chemotherapy: Balancing Toxicity and Response.', *Therapeutic Drug Monitoring* **26**(2), 231.

Erichsen, H. & Chanock, S. (2004), 'SNPs in cancer research and treatment', *British Journal of Cancer* **90**, 747–751.

Fan, J. et al. (2003), 'Highly Parallel SNP Genotyping', *Cold Spring Harbor Symposia on Quantitative Biology* **68**(1), 69–78.

Gower, J. (1966), 'Some distance properties of latent root and vector methods used in multivariate analysis', *Biometrika* **53**(3-4), 325–338.

He, X. & Niyogi, P. (2004), 'Locality Preserving Projections', *Advances in Neural Information Processing Systems 16: Proceedings of the 2003 Conference* .

He, X., Yan, S., Hu, Y., Niyogi, P. & Zhang, H. (2005), 'Face Recognition Using Laplacianfaces', *IEEE Transaction on pattern analysis and machine intelligence* pp. 328–340.

Herr, A., Grützmann, R., Matthaei, A., Artelt, J., Schröck, E., Rump, A. & Pilarsky, C. (2005), 'High-resolution analysis of chromosomal imbalances using the Affymetrix 10K SNP genotyping chip', *Genomics* **85**(3), 392–400.

Hinton, G. & Roweis, S. (2003), 'Stochastic neighbor embedding', *Advances in Neural Information Processing Systems* **15**, 833–840.

Hirschhorn, J. & Daly, M. (2005), 'Genome-wide association studies for common diseases and complex traits', *Nature Reviews Genetics* **6**(2), 95–108.

Hotelling, H. (1933), 'Analysis of a complex of statistical variables into principal components', *Journal of Educational Psychology* **24**(6), 417–441.

Irving, J., Bloodworth, L., Bown, N., Case, M., Hogarth, L. & Hall, A. (2005), 'Loss of Heterozygosity in Childhood Acute Lymphoblastic Leukemia Detected by Genome-Wide Microarray Single Nucleotide Polymorphism Analysis'.

Kaski, S., Nikkila, J., Oja, M., Venna, J., Toronen, P. & Castren, E. (2003), 'Trustworthiness and metrics in visualizing similarity of gene expression', *BMC Bioinformatics* **4**(1), 48.

Kokiopoulou, E. & Saad, Y. (2005), 'Orthogonal Neighborhood Preserving Projections', *IEEE Int. Conf. on Data Mining* pp. 1–8.

Kokiopoulou, E. & Saad, Y. (2007), 'Orthogonal Neighborhood Preserving Projections: A Projection-Based Dimensionality Reduction Technique', *IEEE Transactions on pattern analysis and machine intelligence* pp. 2143–2156.

Kruskal, J. (1964), 'Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis', *Psychometrika* **29**(1), 1–27.

Lee, J., Lendasse, A. & Verleysen, M. (2004), 'Nonlinear projection with curvilinear distances: Isomap versus curvilinear distance analysis', *Neurocomputing* **57**, 49–76.

Leykin, I., Hao, K., Cheng, J., Meyer, N., Pollak, M., Smith, R., Wong, W., Rosenow, C. & Li, C. (2005), 'Comparative linkage analysis and visualization of high-density oligonucleotide SNP array data', *feedback* .

Memisevic, R. & Hinton, G. (2005), 'Improving dimensionality reduction with spectral gradient descent', *Neural Networks* **18**(5-6), 702–710.

Ng, A., Jordan, M. & Weiss, Y. (2002), 'On spectral clustering: Analysis and an algorithm', *Advances in Neural Information Processing Systems 14: Proceedings of the 2002 [sic] Conference* .

Nguyen, G. & Worring, M. (2004), 'Optimizing similarity based visualization in content based image retrieval', *Multimedia and Expo, 2004. ICME'04. 2004 IEEE International Conference on* **2**.

Price, A., Patterson, N., Plenge, R., Weinblatt, M., Shadick, N. & Reich, D. (2006), 'Principal components analysis corrects for stratification in genome-wide association studies', *Nature Genetics* **38**, 904–909.

Reich, D., Gabriel, S. & Altshuler, D. (2003), 'Quality and completeness of SNP databases', *Nature Genetics* **33**(4), 457–458.

Risch, N. & Merikangas, K. (1996), 'The Future of Genetic Studies of Complex Human Diseases', *Science* **273**(5281), 1516.

Roweis, S. & Saul, L. (2000), 'Nonlinear Dimensionality Reduction by Locally Linear Embedding'.

Shi, J. & Malik, J. (2000), 'Normalized Cuts and Image Segmentation', *IEEE Transaction on pattern analysis and machine intelligence* pp. 888–905.

Tabor, H., Risch, N. & Myers, R. (2002), 'Candidate-gene approaches for studying complex genetic traits: practical considerations.', *Nat Rev Genet* **3**(5), 391–7.

Tagaris, G., Richter, W., Kim, S., Pellizzer, G., Andersen, P., Ugurbil, K. & Georgopoulos, A. (1998), 'Functional magnetic resonance imaging of mental rotation and memory scanning: a multidimensional scaling analysis of brain activation patterns', *Brain Research Reviews* **26**(2-3), 106–112.

Tan, P., Steinbach, M. & Kumar, V. (2005), *Introduction to Data Mining*, Addison-Wesley Longman Publishing Co., Inc. Boston, MA, USA.

Torgerson, W. (1952), 'Multidimensional scaling: I. Theory and method', *Psychometrika* **17**(4), 401–419.

van der Maaten, L., Postma, E. & van den Herik, H. (2007), 'Dimensionality reduction: A comparative review', *Disponible sur internet* .

Venkatarajan, M. & Braun, W. (2004), 'New quantitative descriptors of amino-acids based on multidimensional scaling of a large number of. physicalchemical properties', *Journal Molecular Modeling* **7**(12), 445–453.

Venna, J. & Kaski, S. (2001), 'Neighborhood preservation in nonlinear projection methods: An experimental study', *Artificial Neural Networks—ICANN* pp. 485–491.

Venna, J. & Kaski, S. (2006), 'Local multidimensional scaling', *Neural Networks* **19**(6-7), 889–899.

Venna, J. & Kaski, S. (2007*a*), 'Comparison of visualization methods for an atlas of gene expression data sets', *Information Visualization* **6**(2), 139–154.

Venna, J. & Kaski, S. (2007*b*), 'Nonlinear Dimensionality Reduction as Information Retrieval', *Proceedings of the 11th International Conference on Artificial Intelligence and Statistics (AISTATS* 07)* pp. 568–575.

Conferences in Research and Practice in
Information Technology

Volume 87

Data Mining and Analytics 2008

Data Mining and Analytics 2008

Proceedings of the
Seventh Australasian Data Mining Conference (AusDM'08),
Glenelg, South Australia, 27-28 November, 2008

John F. Roddick, Jiuyong Li, Peter Christen and
Paul Kennedy, Eds.

Volume 87 in the Conferences in Research and Practice in Information Technology Series.
Published by the Australian Computer Society Inc.

Published in association with the ACM Digital Library.

Data Mining and Analytics 2008. Proceedings of the Seventh Australasian Data Mining Conference (AusDM'08), Glenelg, South Australia, 27-28 November, 2008

Conferences in Research and Practice in Information Technology, Volume 87.

Editors: **John F. Roddick**
School of Computer Science, Engineering and Mathematics
Flinders University
GPO Box 2100, Adelaide, SA, 5001, Australia
Email: john.roddick@flinders.edu.au

Jiuyong Li
School of Computer and Information Science
University of South Australia, Mawson Lakes
GPO Box 2471, Adelaide, SA, 5001, Australia
Email: jiuyong.li@unisa.edu.au

Peter Christen
Department of Computer Science
Faculty of Engineering and Information Technology
The Australian National University
Canberra ACT 0200 Australia
Email: peter.christen@anu.edu.au

Paul J. Kennedy
Faculty of Engineering and Information Technology
University of Technology, Sydney
Broadway, NSW, 2007, Australia
Email: paulk@it.uts.edu.au

Series Editors:
Vladimir Estivill-Castro, Griffith University, Queensland
John F. Roddick, Flinders University, South Australia
Simeon Simoff, University of Western Sydney, NSW
`crpit@csem.flinders.edu.au`

Publisher: Australian Computer Society Inc.
PO Box Q534, QVB Post Office
Sydney 1230
New South Wales
Australia.

Conferences in Research and Practice in Information Technology, Volume 87
ISSN 1445-1336
ISBN 978-1-920682-68-2

Printed November 2008 by Flinders Press, PO Box 2100, Bedford Park, SA 5042, South Australia.
Cover Design by Modern Planet Design, (08) 8340 1361.

The *Conferences in Research and Practice in Information Technology* series aims to disseminate the results of peer-reviewed research in all areas of Information Technology. Further details can be found at http://crpit.com/.