Journal of Computer Science & Systems Biology - Open Access
www.omicsonline.com Review Article JCSB/Vol.2 July-August 2009

Data Mining Techniques in High Content Screening: A Survey

Karol Kozak*[,1], Aagya Agrawal[2], Nikolaus Machuy[2], Gabor Csucs[1]

[1]ETH Zurich

[2]MPI-IB Berlin

*Corresponding author: Dr. Karol Kozak, E-mail: karol.kozak@lmc.biol.ethz.ch

Received June 10, 2009; Accepted July 12, 2009; Published July 12, 2009

Citation: Kozak K, Agrawal A, Machuy N, Csucs G (2009) Data Mining Techniques in High Content Screening: A Survey. J Comput Sci Syst Biol 2: 219-239. doi:10.4172/jcsb.1000035

Abstract

Advanced microscopy and corresponding image analysis have evolved in recent years as a compelling tool for studying molecular and morphological events in cells and tissues. Cell-based High-Content Screening (HCS) is an upcoming technique for the investigation of cellular processes and their alteration by multiple chemical or genetic perturbations. The analysis of the large amount of data generated in HCS experiments represents a significant challenge and is currently a bottleneck in many screening projects. This article reviews the different ways to analyse large sets of HCS data, including the questions that can be asked and the challenges in interpreting the measurements. The main data mining approaches used in HCS are image descriptors, computations, normalization, quality control methods and classification algorithms.

Keywords: High content screening; Microscopy; RNAI; Data mining; Supervised; Unsupervied classification

Introduction

Cell biologists are the psychiatrists of the cellular world. They observe cell "behavior" through a microscope. The microscopes are a boon to cell biologists. The daily work of cell biologists is based on microscopy. Particularly, fluorescent microscopy has enabled multifaceted insights into the detail and complexity of cellular structures and their functions for well over two decades. As an essential prerequisite for a systematic phenotypical analysis of gene functions in cells at a genome-wide scale, the throughput of microscopy had to be improved through automation. Along with the introduction of the first automated fluorescent imaging systems in the late 90's the term 'High-Content Screening' was coined and introduced. HCS is defined as multiplexed functional screening which is based on imaging multiple targets in the physiologic context of intact cells by extraction of multicolour fluorescence information. There are a number of advantages of HCS over other screening technologies. First of all, cell-based assays reflect high physiological relevance particularly with regard to drug screening the response is not limited to a single target but rather to a whole cell containing thousands of targets. Putative cytotoxic (side) effects are discovered early on. Secondly, single cell analysis reflects the heterogeneity of cell populations as well as their individual response to treatment. Simultaneous staining in 3 or 4 colors allows the extraction of various parameters from each cell quantitatively as well as qualitatively such as intensity, size, distance and distribution (spatial resolution). The parameters might be referenced to each other, for example the use of nuclei staining to normalize other signals against cell number, or particular parameters can verify or exclude each other. Hence, HCS is well known methodology to generate low false-positive and false-negative results.

An essential factor for the success of high content screening projects is the existence of algorithms and software that has made invaluable contribution to the various scientific fields and which can reliably and automatically extract information from the masses of captured images. In general, nuclei are identified and masked first. Then areas around the nuclei are determined or the cell boundaries are searched to mask the cell shape. Dyes that stain not only chromosomal DNA in the nucleus but also mRNA in the cytosol help to identify the shape of cell. Nuclei may be counted along with extraction of additional parameters such as shape, size, substructures like spots, or intensity. Subsequently, the

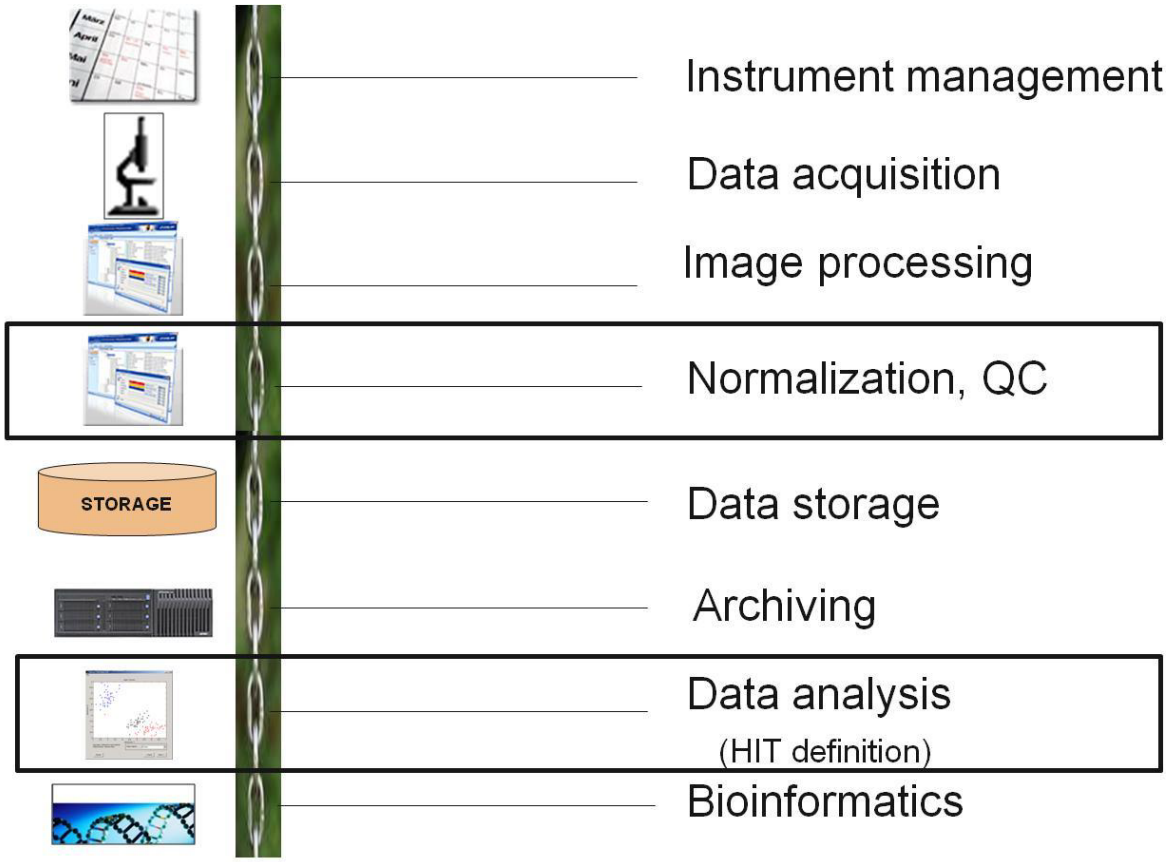

Figure 1: The key steps necessary for conducting a data flow of high-content image based screening. In this figure the pipeline depicts the essential steps for conducting a data flow of high-content image based screening that comprise instrument management (logistic – booking systems), data acquisition using automated microscopy, automated image processing, normalization together with quality control, data storage using relational databases, archiving on tape storage system, data analysis including data modeling and visualization for hit definition and as last step bioinformatics. Highlighted parts in this figure are our focus of discussion in this paper.

masks are laid over the image(s) of the other channel(s) and signals within the masks are measured. Most of the advanced automated microscopes are delivered with proprietary image analysis solutions for a broad range of biological events. For popular assays at the sub-cellular level such as cell cycle analysis (mitotic index), cytotoxicity, apoptosis, micronuclei detection, receptor internalization, protein translocation (membrane to cytosol, cytosol to nucleus, and vice versa), co-localisation and cytoskeletal arrangements has became very easy to perform using HCS. The morphological analysis at the cellular level such as neurite outgrowth, cell spreading, cell motility, colony formation, or tube formation, ready-to-use scripts are available and need only some fine-adjustment for the particular cell line and/or conditions of the assay. Besides the packages provided by the microscope suppliers, a number of commercial and open source products are available that can be used alternatively.

RNA interference (RNAi) has become a method of choice for functional genomics studies in vertebrates and invertebrates. RNAi refers to the biological process by which short interfering RNA (siRNA) after incorporation into the RNA induced silencing complex (RISC) degrades complementary messenger RNA (mRNA) sequences. Presently available analysis methodologies for large-scale RNAi data sets typically rely on ranking data and are based on single image descriptor (feature) or significance value (Boutros et al., 2004; Moffet et al., 2006; Kittler et al., 2004). HCS data analyses focus on the identification of highly active siRNAs, which typically fall within the top 1% of the assayed activities, and ignore much of the remainder of the data set. Furthermore, these strategies do not exploit redundancies in genome-scale libraries, which typically contain 2–4 siRNAs per gene. Thus, it is difficult to systematically identify genes for which multiple siRNAs are active across a screen, which do not fall within an upper threshold (that is, moderately active siRNAs). As a complete HCS experiment might involves up to hundreds of plates, therefore the image processing result sets can vary greatly in size. As the cost of

Name	Description	Source
HC/DC	HC/DC ia a modular and extendable data exploration platform for data mining of large datasets, developed at University of Konstanz with the collaboration of KNIME, that enables the user to visually create data pipelines, analyse the datasets and get the information from data in the form of result.This software offers the functions like library handling, quality control, workflow management and support machine learning and statistics. It integrates computational service of well known Weka data mining environment and R-Projects.The architecture of HC/DC is based on the KNIME platform and the Eclipse plug-in framework.	http://hcdc.ethz.ch
Spotfire	Spotfire is an interactive data visualization and analytical tool that enables screeners to graph very large datasets for the purpose of identifying outlying data points (e.g. hits)and for comparing datasets.It is very fast and has intuitive user interface.It has connectivity to ISIS host and provides various structure related analyses.It integrates computational service for R-Project,S-PLUS1 and connects SAS files.	http://www.spotfire.com
Batelle Visua	Batelle Visua is visual data mining tool with intuitive user interface that mines in multidimensional space with very large sets of numerical,categorical,chemical and textual data.This software supports function like, feature extraction(relativity tool),dimensionality reduction for visualization, cross-platform compatibility(runs under Solaris, Windows and Linux).It has OmniViz1 plug-in interface for user scripts and tools	http://www.omniviz.com
R-Project	R is an open source statistical analysis software, similar to S-plus that provide a wide variety of statistical (linear and nonlinear modeling, classical statistical tests, time series analysis, classification, clustering etc.and graphical techniques can be considered as a different implementation of S which was developed at Bell Laboratory .It compiles and runs on a wide variety of UNIX, Windows and Mac OS platforms. For computationally intensive tasks, C, C++ and FORTRAN code can be link with this and called at runtime.It can be easily extended via packages.	http://www.r-project.org

Insightful	Insightful is a highly scalable data mining workbench for new data miners and skilled analytic professionals that support the entire data mining life cycle. It is value added implementation of S language originally born in Bell's laboratory that provides modules and packages for specific applications like clinical trails,wavelets,optimization etc.It has many features of R-Packages and extended features for robust and nonparametric multivariate analysis,graphics,etc.	http://www.insightful.com
Umetricsis	Umetricsis is the leader in software for design of experiments and multivariate data analysis for the individual user as well as for on-line continuous and batch processes.Umetrics adds value to business by bringing out the valuable information from the data.An Enterprise platform comprises unlimited use of MODDE and SIMCA-P (soft independent modeling of class analogies products), software validation reports, technical supports and manuals to all users.	http://www.umetrics.com
Mathworks	The Mathworks leads very important role in computational biology, complementing the MATLAB and Simulink applications for the life sciences that customers already rely on to import data, analyze and visualize data, model biological systems, communicate results, deploy applications and increase computing performance. It functions for integrating MATLAB based algorithms with external applications and languages, such as C, C++, Java, COM and Microsoft Excel.	http://www.mathworks.com
Partek	Partek software is highly optimized for incredibly fast computations to today's large scientific experimental data. This is a software for data mining and knowledge discovery based on statistical methods, data visualization, neural networks, fuzzy logic and genetic algorithms.Partel offers several software solutions for different applications.For example,Partek Genomics Suite,Partel discovery suite,Partek screening solution,Partek QSAR solution etc.	http://www.partek.com
CellMine	CellMine is developed specifically for HCS application. It is web based instrument agnostic software for storing and mining cell-based assay data. It integrates screening data with images and facilitates linkage to complementary discovery data and compound information.It unlocks the value of cell-based assays by facilitating improved lead selection and optimization. CellMine is built on BioImagene's 3i specially designed for image management solutions for the life science industry.	http://www.bioimagene.com

AcuityXpress	AcuityXpress is the cellular informatics software for the Total Imaging Solution from Molecular Devices and it has been specifically designed to address the needs of high content data analysis at enterprise level.It integrates image acquisition, image analysis and informatics. This integration enables direct linkage of data analysis with the original images.	http://www.AcuityXpress.com
Genedata	Genedata software solutions enable scientists to process, integrate, analyze, and manage large and complex experimental data sets generated by high throughput technologies. Solutions include Genedata Phylosopher for target discovery and integrative biological data management, Genedata Screener for automated high throughput screening and high content screening, and Genedata Expressionist for biomarker discovery and personalized medicine. Genedata is privately held, with headquarters in Basel, Switzerland.	http://www.genedata.com
Pipeline Pilot	Pipeline Pilot is the famous graphical workflow programming software from SciTegic/Accelrys. This software is based around a powerful client-server platform that lets you construct workflows by graphically combining components for data retrieval, filtering, analysis, and reporting. Different client interfaces to the Pipeline Pilot platform enable you to work in the environment that best suits your needs. Pipeline Pilot is designed to meet critical requirements of the informatics professional: an agile development environment, fast and secure deployment, minimal maintenance costs, and application extensibility.	http://www.scitegic.com
Cluster/Treeview	Cluster and TreeView are programs that provide a computational and graphical environment for analyzing data from DNA microarray experiments, or other genomic datasets. The program Cluster organizes and analyzes the data in a number of different ways.TreeView allows the organized data to be visualized and browsed. Although it is the standard for hierarchical clustering and viewing dendrograms,this software also creates self organizing maps and performs principal-components analysis.	http://rana.lbl.gov/EisenSoftware.htm
GeneCluster 2.0	GeneCluster 2.0 is a software package for analyzing gene expression and other bioarray data, giving users a variety of methods to build and evaluate class predictors, visualize marker lists, cluster data and validate results. It includes algorithms for building and testing supervised models using weighted voting and k-nearest neighbor algorithms, a module for systematically finding and evaluating clustering via self-organizing maps, and modules for marker gene selection and heat map visualization that allow users to view and sort samples and genes by many criteria. GeneCluster 2.0 is a stand-alone Java application and runs on any platform that supports the Java Runtime Environment version 1.3.1 or greater.	http://www-genome.wi.mit.edu/cancer/software/genecluster2/gc2.html

RELNET	RELNET (Relevance Networks Software) is a software tool that allows genomics and bioinformatics researchers to construct relevance networks from their gene expression data. The software is written in Java, and a license for the source code is also available. There are three main advantages to using relevance networks: Negative associations are shown, Disparate data types can be included in the same analysis (i.e. clinical, expression, and phenotypic), Multiple connections are allowed for each gene.	http://www.chip.org/relnet
CellHTS2	CellHTS2 is a software package implemented in Bioconductor/R to analyze cell-based high-throughput RNAi screens.The cellHTS2 package is the new version of the cellHTS package, offering improved functionality for the analysis and integration of multi-channel screens and multiple screens.	http://www.dkfz.de/ signaling/cellHTS/
Weka	Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes.	http://www.cs.waikato.ac.nz/ml/weka/

Box 1: Some freely available software for High Content Screening analysis.
Many commercial life sciences workflow products make heavy use of open source and publicly available software for pre- and post-processing analysis of screening data. Those software can be used to perform the all the analytical techniques described in this article. Some of them are listed below.

siRNA continues to drop, it is clear that HCS has become more integral to the drug discovery process. In addition to the obvious use of functional genomics in basic research and target discovery, such as finding siRNA which target genes in significantly different patterns across samples, there are many other specific uses in this domain. To investigate patterns a good data mining package is require. Many free and commercial software packages (Box 1) are now available to analyse HCS data sets, although it is still difficult to find a single off-the-shelf software package that answers all questions related to RNAi silencing. As the field is still young, when developing a bioinformatics analysis pipeline, it is more important to have a good understanding of both the biology involved and the analytical techniques rather than having the right software. This article reviews the different ways to analyse HCS data, and will concentrate on selecting the appropriate method for the particular data analysis step.

Data Normalization

HCS has already proven to be a successful method to deliver more relevant information simultaneously in one experiment, rather than delivering a single readout in a series of sequential experiments (Johnston and Johnston, 2002; Giuliano et al., 2003; Taylor et al., 2003; Monk, 2005). A prototype scenario might be the series of simultaneously available readouts obtained from a cellular assay. One parameter identifies cells (i.e. membrane dye at first wavelength), another determines the stage of mitotic change (e.g. fragmented and condensed nuclei at a second wavelength) and a third parameter classifies the apoptotic stage using morphological criteria at a third wavelength. Certainly, these analyses can already be performed almost automatically with very high throughput.

The hypothesis underlying HCS data analysis is that the measured image descriptors for each single siRNA repre-

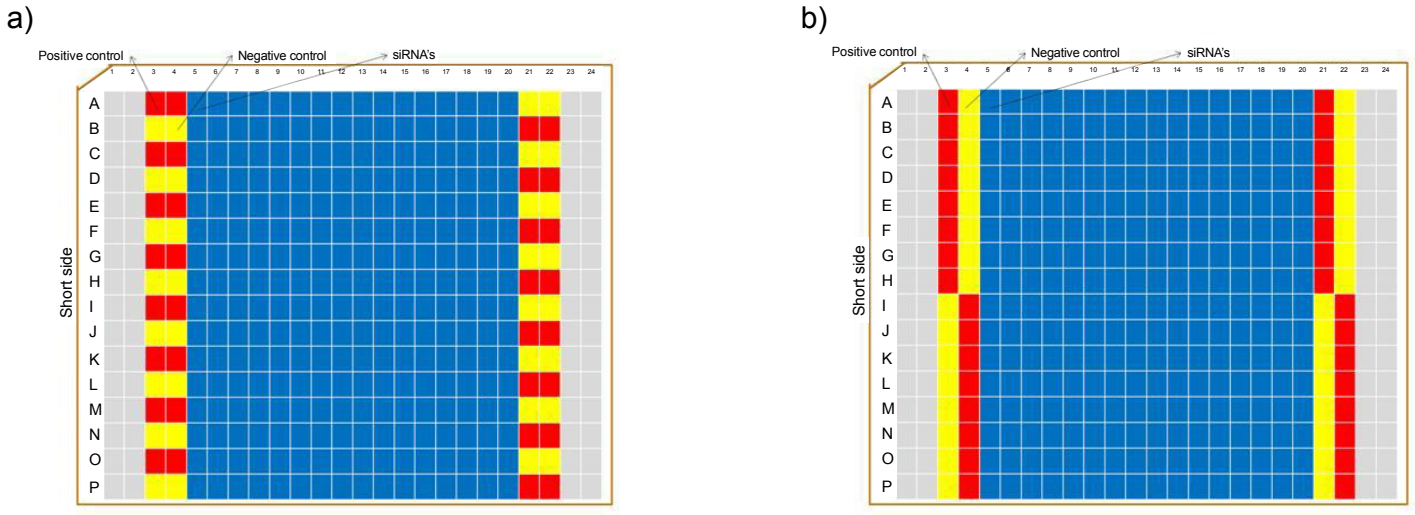

Figure 2: Location of controls on a 384-well plate. In a screening process, the designed biological assay is performed by using a robot to add the cells and specific reagents (siRNA) to each well, which already contain different oligonuceotide or control. After incubation or other required manipulations, fluorescence images are acquired for every well by automated microscope. These raw data represent the images of each oligonucleotide or control against a specified target. (a) Generally, in a siRNA experiment, 256 different oligonucleotide (blue) are stored in the middle of a 384-well plate and wells on the first two and last two columns are left empty (b) Ideally, controls should be located randomly among the 384 wells of each plate. Only the first two and the last two columns are typically available for controls. Despite this limitation, edge-related bias can be minimized by alternating the sixteen positive controls (red) and the sixteen negative controls (yellow) in the available wells, such that they appear equally on each of the sixteen rows and each of the 4 available columns.

sent its relative number of observed objects to the fluorescence image. A well-defined and highly sensitive test system requires both quality control and accurate measurements. Within-plate reference controls are typically used for these purposes (FIG. 2). Controls help to identify plate-to-plate variability and establish assay background levels. Normalization of raw data removes systematic plate-to-plate variation, making measurements comparable across plates. Systematic errors decrease the validity of results by either over or under estimating true values. These biases can affect all measurements equally or can depend on factors such as well location, liquid dispensing and signal intensity. Although recent improvements in automation can minimize bias, and thereby provide more reproducible results, equipment malfunctions can nonetheless introduce systematic errors, which must be corrected at the data processing and analysis stages.

Interpretation of experimental data is often improved, when it can be compared with results from earlier experiments (run). Normalization is the process that is prerequisite for comparability which ensures that data can be compared 'out of the box', and the details of the experiments are known so that, they can be considered during the comparison. An element of this normalization process is shown in FIG. 3a and FIG. 3b: a common source of false-positives or false- negatives are plate patterns (i.e. systematic errors that shift the assay signal depending upon the position of the sample on the plate). In this case for normalization global global normalization median centering has been used which multiplies each image processing parameter by a constant such that the median value is zero (for log-transformed ratios). This type of normalization method tends to decrease distances between siRNA by moving patterns from each end of the scale toward the center.

Number of normalization steps should be carried out to eliminate low-quality measurements of the data, to adjust the measured image descriptors, and to select siRNA that significantly give a target effect. The experimental design and usage of biological and/or technical replicates affects the choice of the normalization methods. Generally, normalization removes all non-biological variation introduced in the measurements. This can be achieved either by using self-consistency methods like global normalization (describe later in this section), linear regression, LOWESS (Locally Weighted Linear Regression), or by using quality elements such as self-normalization, controls. Depending on the experiment, normalization is used in different ways. It has to be distinguished between within-run normalization, paired-

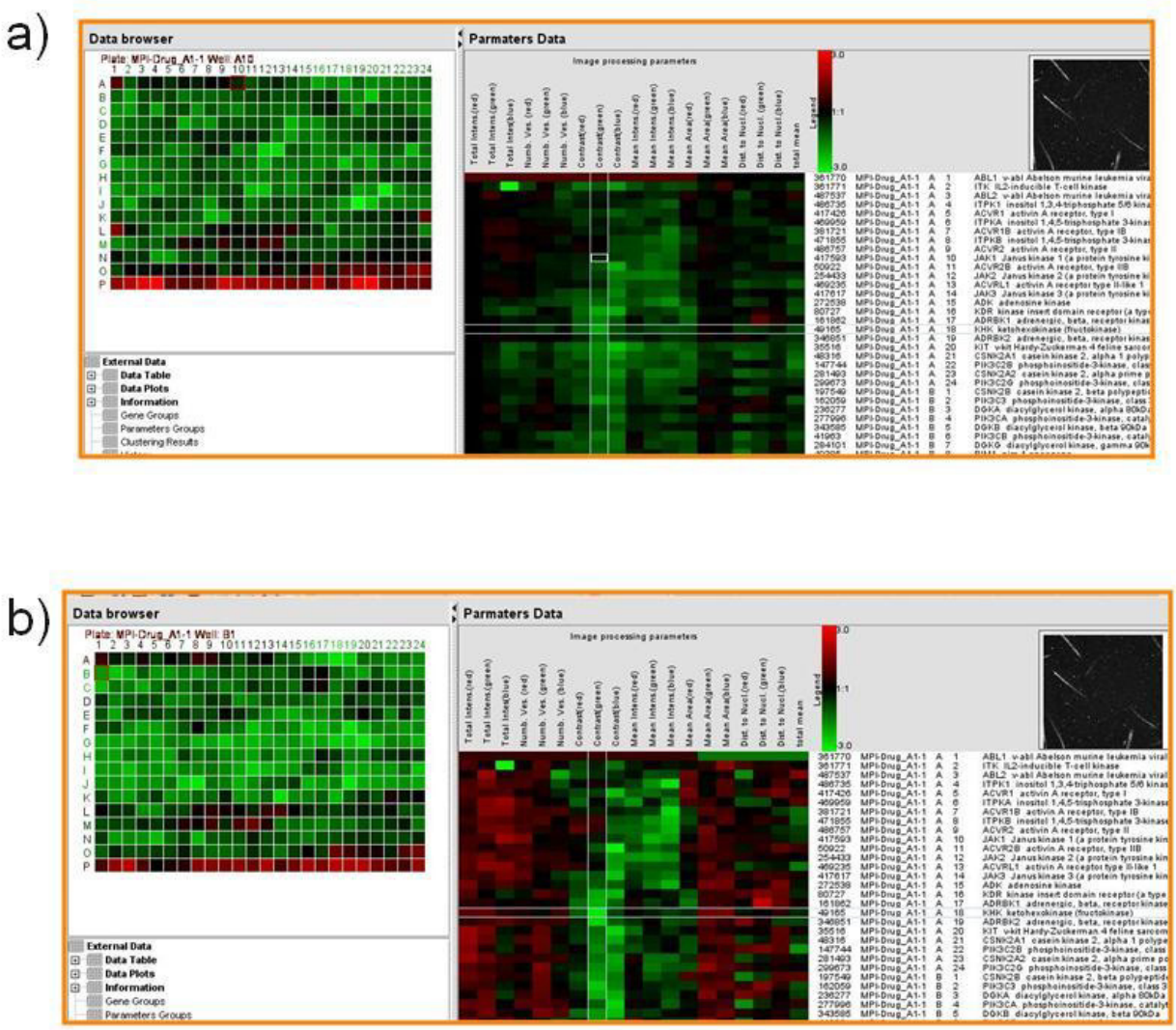

Figure 3: Data normalization as a prerequisite (tool) for successful use in a broader project-spanning context: an example. In a) and b): a common source of false-positives or false-negatives are plate patterns (i.e. systematic errors that shift the assay signal depending upon the position of the sample on the plate). Mean centering has been used as normalization method.

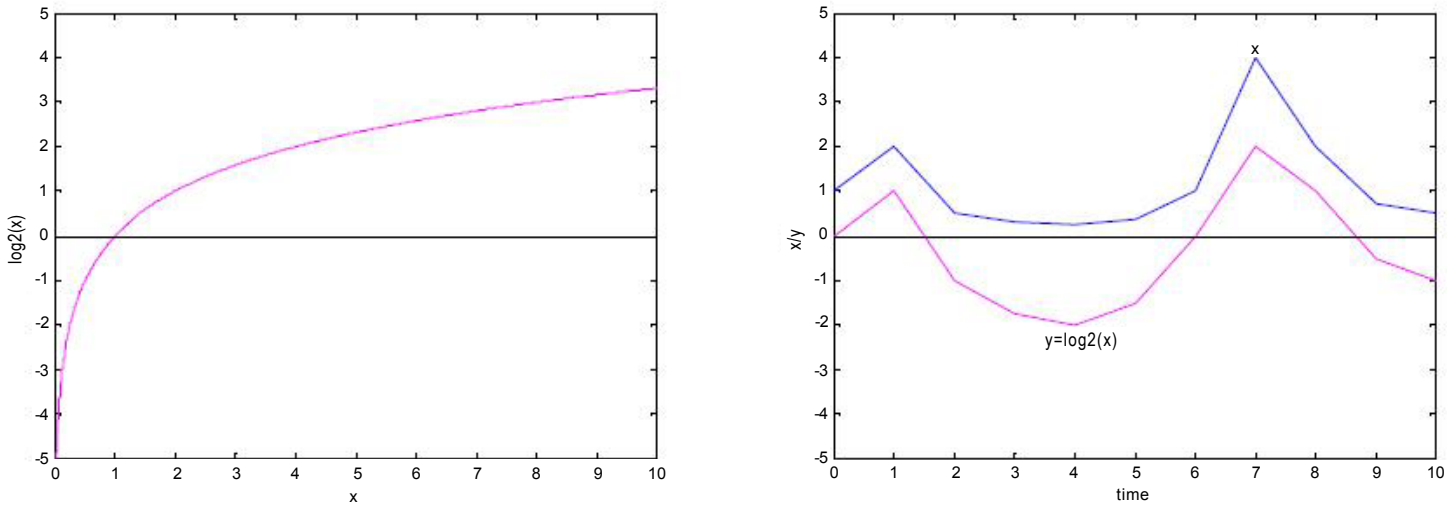

Figure 4: This figure depicts (left) log2-transformation and (right) the mean of intensity curve of a time series in normal (blue) and logarithmic space (magenta).

oligonucleotide normalization for dye-swap pairs, and multiple-siRNA normalization (scaling between plates). In each case one can use all siRNA on a plate or a set of control siRNA as the set of genes used for normalization. The data for each siRNA are typically reported as image descriptors (example: number of cells) or as the logarithm of those descriptors. The descriptor ratio is simply the normalized value of the parameter for a particular siRNA oligonucleotide in the query sample divided by its normalized value for the control.

Here, we will present an example of data normalization based on RNAi screens in cultured human cells, combining reverse transfection by siRNA cell arrays and automated time-lapse fluorescence microscopy. Measured siRNA to be a true hit is a function of at least two factors (Malo et al., 2006): the siRNA true hit and random error. Symbolically, one simple additive model might be $y_{ijp} = \mu_{ijp} + \varepsilon_{ijp}$ where y_{ijp} is the observed raw measurement obtained from the well located on row i and column j on the p^{th} plate, μ_{ijp} is the 'true' hit and ε_{ijp} is the effect of all sources of error. Assuming no bias, the ε_{ijk}'s are assumed to have zero mean and a specified probability distribution (e.g., normal). Another simple model is $y_{ijp} = \mu_{ijp} + r_{ip} + c_{jp} + \varepsilon_{ijp}$ where r and c represent plate-specific row and column artifacts, respectively, and ε_{ijp} represents remaining sources of error. Further we will illustrate two most used in HCS normalization approaches.

Logarithmic Transformation

Almost all results from HCS experiments are image descriptors. That means, that positive controls and siRNA hit should be represented by a range of $1<x_{ij}<+8$, whereas negative controls are represented by a range of $0=x_{ij}<1$. To overcome this discrepancy the data is usually transformed into logarithmic space, whereas the upregulated siRNAs are assigned to positive and down regulated siRNAs are assigned to negative (FIG. 4).

A second property of this transformation is that the data is represented in a more "natural" way. In most cases the *logarithmus dualis* (log2) is used instead of log10 or *logarithmus naturalis* (ln), because of the better scaling and the more natural understanding of differences between positive and negative values.

Mean or Median Centering

A simple but efficient method is 'median', which is used to quantify the spatial-trend structure of an assay plate, and enables predicting systematic deviations from the expected spatial or timely behavior of the experimental parameters. Approaches that are more advanced than the simple median polish approach, which are shared by many software packages, have proven successful: methods such as global parametric models that model the experimental data with the assumption that one model, one global image descriptor value, is applicable and valid for the complete data set. These methods are well-suited for characterizing the general shape of signal drifts. At present many published HCS series consist of a large number of cancer samples, all compared to a common reference sample (positive/negative control) consist of a collection of cell-lines. It is advantageous to make screen with one plate only having controls and use it as independent reference between replicates and runs. In such situation control is independent from the real replicate/run and the analysis is preferred to be independent from the image descriptor observed in the control and that is exactly what can be achieved by mean and/or median centering. After applying this procedure the values of each single siRNA reflect the variation from some property of the series of observed values such as the mean or median (FIG. 5). It makes less sense in one replicate/run where the control is part of the experiment, as it is in many time courses. It is important to know about upregulation or downregulation of oligonuceotide and the distance of siRNA from each other, since this procedure tends to decrease distances between siRNA by moving patterns from each end of the scale toward the center (FIG. 5).

Centering the data for wells/plates can also be used to remove certain types of biases, which have the effect of multiplying ratios for all siRNA's by a fixed scalar. Mean or median centering the data in log-space has the effect of correcting these biases. However, since in clustering only distances between siRNAs or image descriptors are used, absolute values play a less important role in the comparison of two siRNA/descriptors. Therefore siRNA are classified as similar even if for one parameter (example: number of cells) are not the same for each siRNA, labeling efficiency and image acquisition parameters (FIG. 6). One way to keep track of differences between replicates is the use of housekeeping siRNA or external controls. These control siRNA are siRNA, which affects on the cell are invariant in respect of the investigated biological process, i.e. number of cells should be the same in each replicate. Therefore they can be used for normalization procedures.

Systematic Errors and Quality Control

HCS operates with samples in microliter volumes that are arranged in two-dimensional plates. A typical HCS plates contain 96 (12×8) or 384 (24×16) samples. The quality control and normalization procedure in primary HCS screens is mainly performed by automatic routines. Quality of mea-

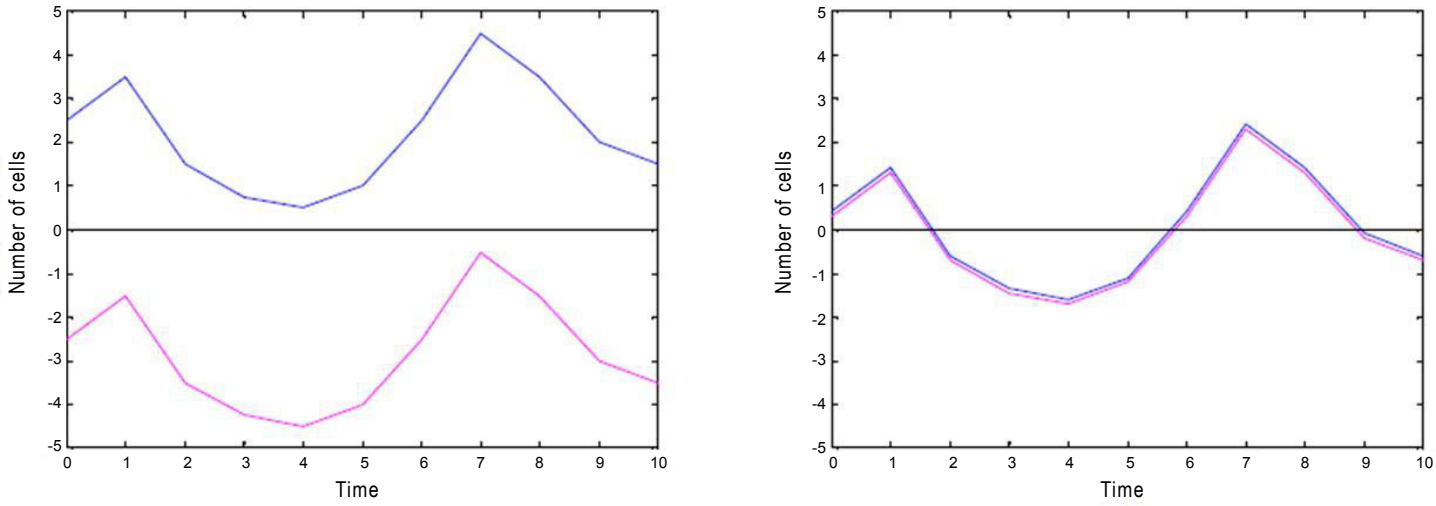

Figure 5: (left) Number of cells of 2 siRNA, one is a control, one is observed candidate (hit). (right) Number of cells after mean centering: both curves are identical, i.e. the distance between them is 0.

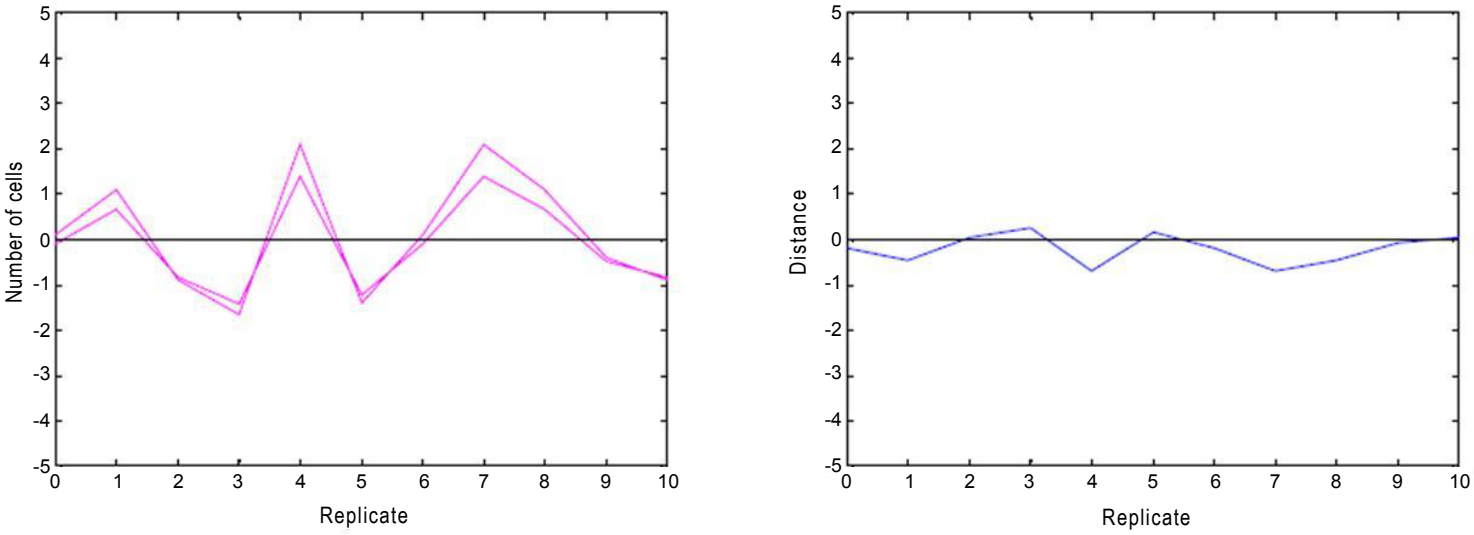

Figure 6: (left) Two equal control siRNA with constant number of cells in well in an ideal (blue) and real (magenta) in different replictes. (right) The distances between the two controls are the same, even though the absolute values are different.

surements has a number of advantages, including objectivity, reproducibility and ease of comparison across screens. Random and systematic errors can cause a misinterpretation of candidates as a hit. They can induce either underestimation (false negatives) or overestimation (false positives) of measured parameters. Various methods dealing with quality control are available in the scientific literature. These methods are discussed in details in the papers (Heuer et al., 2002; Gunter et al., 2003; Brideau et al., 2003; Heyse, 2002; Zhang et al., 1993, 1995). However, statistical methods that analyze and remove systematic errors in HCS assays are poorly developed compared to those dealing with microarrays. There are various sources of systematic errors. Some of them are mentioned in the article of (Heuer et al., 2002):

- Systematic errors caused by ageing, reagents evaporation or decay of cells can be recognized as smooth trends in the plate's means/medians.
- Errors in liquid handling and malfunction of pipettes can also generate localized deviations of expected data values.

- Variation in incubation time drift in measuring different wells or different plates, and also the reader effects can be recognized as smooth attenuations of measurement over an assay.

(Heuer et al., 2002) and (Brideau et al., 2003) demonstrated examples of systematic signal variations that are present in all plates of an assay. For instance, (Brideau et al., 2003) illustrates a systematic error caused by the positional effect of detector.

To check the reliability by avoiding systematic errors, data quality control at different levels is essential. This begins in the optimization phase of the assay: In test runs with a small number of compound plates the assay has been shown to possess a sufficient signal window (e.g., Z-factor (Zhang et al.,1999)), stability, and sensitivity (e.g., measured by the effects of known control compounds) (Cox et al.,2000;Lutz and Kenakin,2000). If problems occur, the parameters of the assay or even its format should be tuned to match the quality criteria of HCS. Data quality control on the level of an individual assay seeks again to guarantee assay stability and sensitivity, which must be monitored constantly using the appropriate controls. At the same time, it tries to pick up a process artifacts caused by failures in the screening machinery or the test system (e.g. a blocked pipettor needle, air bubbles in the system, a changing metabolic state of reporter cells) (Heyse, 2002).

If unnoticed, these can result in a high number of false positive, but seemingly "highly specific hits". Often, such process artifacts can be detected by changes in the overall signal or by specific "signal patterns" on plates (e.g. pipettor line patterns), if the compound library is randomized across the screening plates. This analysis is preferably done directly after the screening run to ensure that such patterns can be traced back to their origin (e.g. the pipettes may be inspected the next morning) and can be unambiguously classified as artifacts - or nonartifacts. This distinguishes false positive from real actives that should be more or less randomly dispersed when considering a whole series of plates from reasonably randomized compound collections.

Statistical Deconvolution

Statistical deconvolution methods can identify common patterns of groups of plates. These methods provide the scientist with a quick overview of strong gradients or patterns in the assay, and form the basis for the decision whether certain plates must be repeated or should simply be corrected. The gradient correction adds discriminatory power to the assay results, since it renders results perfectly comparable independent of plate location and reduces the noise introduced by gradients. In the case of nonrandom compound distribution on the plates (e.g. in retests containing many actives) such correction methods still can be applied if there are solvent plates interspersed in the screening run at regular intervals.

Dimension Reduction and Image Descriptors Selection

Mathematically, a library with **n** (siRNAs) and represented by m (m >3) image descriptors is an n x m dimensional matrix. There is no way to graph the matrix, although one would like to review the diversity graphically. In order to solve this problem, dimensionality needs to be reduced to two or three. For this dimension reduction is required. There are many approaches available for dimension reduction. Here we will summarize some of the widely accepted dimension reduction technologies.

Multidimensional Scaling

Multidimensional scaling (MDS) (Cox and Cox, 2000) or artificial neural network (ANN) methods are traditional approaches for dimension reduction. MDS is a non-linear mapping approach. It is not an exact procedure as rather a way to "rearrange" objects in an efficient manner, and thus to arrive at a configuration that best approximates the observed distances. It actually moves objects around in the space defined by the specified number of dimensions and then checks how well the distances between objects can be reproduced by the new configuration. In other words, MDS uses a minimization algorithm that evaluates different configurations with the goal of maximizing the goodness-of-fit (or minimizing "lack of fit") (StatSoft Inc.)(See reference). Multi-dimensional scaling (MDS) is aimed to represent high dimensional data in a low dimensional space with preservation of the similarities between data points. This reduction in dimensionality is crucial for analyzing and revealing the genuine structure hidden in the data. For noisy data, dimension reduction can effectively reduce the effect of noise on the embedded structure. For large data set, dimension reduction can effectively reduce information retrieval complexity. Thus, MDS techniques are used in many applications of data mining and gene network research.

Self-organising Map (SOM)

A SOM is basically a multidimensional scaling method, which projects data from input space to a lower dimensional output space. Self-organising map (SOM) is one of the ANN methods. Effectively, it is a vector quantization algorithm that creates reference vectors in a high-dimensional input space and uses them, in an ordered fashion, to approximate the input patterns in image space. It does this

In any clustering algorithm, the calculation of a 'distance' between any two objects is fundamental to place them into groups. Analysis of HCS data is not different in finding clusters of similar siRNAs. It relies on finding and grouping those that are 'close' to each other. To perform this, we rely on defining a distance between each image parameter vector. There are various methods for measuring distance; these typically fall into two general classes: metric and semi-metric.

Metric distances

To be classified as 'metric', a distance measure *dij* between two vectors, *i* and *j* ,must obey several rules:

• The distance must be positive definite, *dij* ≥0 (that is, it must be zero or positive).

• The distance must be symmetric, *dij*= *dji*, so that the distance from *i* to *j* is the same as the distance from *j* to *i*.

• An object is zero distance from itself, *dii* = 0.

• When considering three objects, *i*, *j* and *k*, the distance from *i* to *k* is always less than or equal to the sum of the distance from *i* to *j*, and the distance from *j* to *k*, *dik* ≤*dij* + *djk*. This is sometimes called the 'triangle' rule.

The most common metric distance is Euclidean distance, which is a generalization of the familiar Pythagorean theorem. In a three-dimensional space, the Euclidean distance, *d*12, between two points, (*x*1,*x*2,*x*3) and (*y*1,*y*2,*y*3) is given by EQN 1:

$$d_{12} = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + (x_3 - y_3)^2}$$

Where (*x*1, *x2, x3*) are the usual Cartesian coordinates (*x*, *y*, *z*). The generalization of this to higher-dimensional expression spaces is straightforward. For our *n*-dimensional expression vectors, the Euclidean distance is given by EQN 2:

$$d = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Where *xi* and *yi* are the measured expression values, respectively, for genes X and Y in experiment *i*, and the summation runs over the *n* experiments under analysis.

Other less intensive used metrtics: Pearson correlation coefficient, Uncentered Pearson correlation coefficient, Squared Pearson correlation coefficient, Averaged dot product, Cosine correlation coefficient, Covariance, Manhattan distance, Mutual information, Spearman Rank-Order correlation, Kendall's Tau

Semi-metric distances

Distance measures that obey the first three consistency rules, but fail to maintain the triangle rule are referred to as semi-metric. There are a large number of semi-metric distance metrics and these are often used in HCS data analysis.

Box 2: Dissimilarity measures.

by defining local order relationships between the reference vectors so that it depends on each other though their neighboring values would lie along a hypothetical "elastic surface" (Kohonen et al., 1992; Zupan and Gasteiger,1993; Bernard et al.,1998). Therefore SOM is able to approximate the point density function, p (x), of a complex high-dimensional input space, down to a two dimensional space, by preserving the local features of the input data. The SOM algorithm is based on unsupervised competitive learning, which means that the training is entirely data-driven and needs no further information. We will describe this method later as classifier.

Missing Values

Due to various effects during automated transfection, staining, and data analysis, not every siRNA can be assigned a meaningful ratio. Those results missing values in the data matrix. To calculate distances, only elements represented in both vectors are used. If there is a missing value in one or both vectors, this dimension is not included in the distance calculation. This can lead to various problems:

– The greatest problems occur, if the distance is not independent of the number of vector elements *n*, as it is the case for Euclidian distance (Box 2) for instance. Vectors with missing values are then differently weighted in comparison with vectors with no missing values. Let's say there are 3 siRNA:

1. A siRNA-vector with all values valid
2. A siRNA -vector with all values present but not equal to 1.
3. A siRNA -vector with only one value equal to the corresponding value of siRNA 1

If vector 1 & 2 and 1 & 3 are compared, the following results are obtained:

- 1 & 2: They are not similar so the distance is greater than 0
- 1 & 3: Only values, which are present in both vectors, are used for the distance calculation, so the siRNAs are treated similarly, because the only comparison results in distance 0. But vector 1 and 3 could also be completely different.

For that reasons, missing values are difficult to handle. A few are usually no problem, but if there are too many in comparison to the number of vector-elements *n*, an arbitrary result can be expected.

Supervised or Unsupervised Method

How to deal with hit definition using multidimensional data set? Current methodologies are based upon pattern recognition algorithms to analyse multiparametric image descriptors delivered from image processing (n x m dimensional matrix). Classification methodologies can be classified into two categories: supervised approaches, or analysis to determine genes that fit a predetermined pattern; and unsupervised approaches, or analysis to characterize the components of a data set without the a priori input or knowledge of a training signal. Many of these algorithms are also offered as part of various software free available solutions and software development kits (SDK) (Box 1). In any pattern recognition algorithm, the calculation of a 'distance' (dissimilarity measures) between any two observations is fundamental.

Dissimilarity Measure

It is crucial to distinguish between different dissimilarity measures (also known as 'metrics') used not for clustering but also used in classification algorithms. A dissimilarity measure indicates the degree of similarity between two siRNAs in screening data set. A clustering method builds on these dissimilarity measures to create groups of features with similar patterns. A commonly used dissimilarity measure is Euclidean distance, for which each gene is treated as a point in multidimensional space, each axis is a separate image parameter and the coordinate on each axis is the value of that parameter. One disadvantage of Euclidean distance is that if measurements are not normalized, correlation of measurements can be missed, the focus being instead on the overall amount of image descriptors. A second disadvantage is that siRNAs that are negatively associated with each other will be missed. The concept of negative interaction is clearly different from the concept of no interaction. Another dissimilarity measure that is commonly used is the Pearson Correlation Coefficient, which is measured between two siRNAs that are treated as vectors of measurements. The disadvantages in using this measure with image measurements are: first, it assumes that the measurements are normally distributed, which might not be the case for oligonucleotide-siRNA measurements; and second, it assumes that siRNAs interact in the assumed linear model, when in biology, a particular siRNA might have target effect to other genes when in the middle of its own range of image descriptor values. Operationally, this measure is sensitive to outliers. Although techniques such as the Rank Correlation Coefficient deal with these by replacing the measurements with ranks, it is not clear whether eliminating the outliers is ideal. Many past discoveries have been found by focusing on the outliers in biology. A third dissimilarity measure is mutual information, which allows for any possible model of interaction between siRNAs and uses each parameter equally regardless of the actual value, and is therefore not biased by outliers. However, calculating the mutual information requires using discrete image parameters (for example, representing the siRNA as 'high' and 'low', or 'high', 'medium' and 'low', and so on), and the mutual information depends on the number of 'bins' used. Ideally, this would be performed in a gene-specific manner, but sufficient information about the range of parameter of each gene in all tissues has not yet been identified. Furthermore, siRNA–siRNA associations with high mutual information might not even be functions in the mathematical sense, and might be difficult to explain biologically. Once a dissimilarity measure has been chosen, the appropriate classification technique can be applied. This section describes the four commonly used unsupervised techniques — hierarchical clustering, self-organizing maps, relevance networks and principal-components analysis — and two commonly used supervised techniques — nearest neighbours and support vector machines.

Supervised Methods

Supervised methods are generally used for two purposes: finding siRNAs candidates with image descriptors that are significantly different between groups of samples, and finding siRNAs that accurately predict a characteristic of the sample. Most screening experiments still typically use only a handful of cell based assays (or equivalent technology), with samples measured under two or three conditions, and the application has a clear goal of finding those siRNAs that represent significant similarity to control at specific stage of cell cycle. Significance can be evaluated in many different ways, including parametric and nonparametric tests, analysis of variance and many others. Although it would be an understatement to say that the analyses of these smaller sets of screening data is trivial. There are several published techniques that have been used to find siRNAs that is similar to controls. When determining whether a particular siRNAs is similar to control, there are four characteristics that need to be considered: absolute image descriptor value, or whether the siRNA signal (one descriptor) is at a high or low level; subtractive degree of change between groups, or the difference in descriptor across samples (calculated using subtraction); fold changes between groups, or the ratio of descriptor across all samples (calculated by division); and

reproducibility of the measurement, or whether samples with similar characteristics have similar amounts of the gene transcript. All of the available techniques for comparing two sets of screening measurements essentially evaluate these four characteristics for each siRNA in various ways to rank siRNA that are most similar to controls. For larger data sets, comparing one pair of screening at a time misses the trends that might exist between measurements. There are several published supervised methods that find siRNAs or sets of siRNAs that accurately predict sample characteristics, such as distinguishing one type of cancer from another, or a metastatic tumour from a non-metastatic one. The methods that can find individual siRNAs, such as the nearest neighbour approach, and/or multiple genes, such as decision trees, neural networks and support vector machines. This article will focus on the two more popular supervised techniques: nearest-neighbour analysis and support vector machines.

Nearest neighbours: Although the nearest-neighbour technique can be used in an unsupervised manner, it is also commonly used in a supervised fashion to find siRNAs directly with patterns that best match a designated query pattern (control). For example, an ideal siRNA pattern might be one that gives high number of cells as one parameter and low value of mean of intensity in second descriptor. Although this technique results in siRNAs that might individually split into two sets of screening run, it does not necessarily find the smallest set of genes that most accurately splits the two sets. In other words, a combination of parameters of two siRNAs might split into two conditions perfectly, but these two siRNAs might not necessarily be the top two hits which is most similar to the idealized pattern.

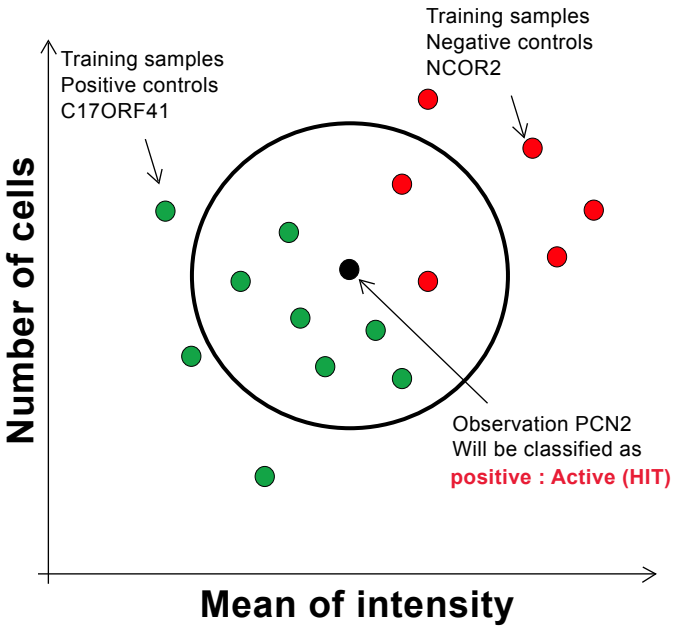

Figure 7: Nearest-neighbour. The nearest-neighbour supervised method first involves the construction of hypothetical siRNAs that best fit the desired patterns. The technique then finds individual siRNAs that are most similar to the hypothetical genes.

Support vector machines: Support vector machines address the problem of finding combinations of siRNAs that better split sets of biological samples. Although it is easy to find individual siRNAs that split two sets with reasonable accuracy owing to the large number of siRNAs (also known as features) measured on automated microscope, occasionally it is impossible to split sets perfectly using individual siRNAs. The support vector machines technique actually further expands the number of features available by combining siRNAs using mathematical operations (called kernel functions). For example, in addition to using the image descriptors of two individual siRNAs *A* and *B* to separate two sets of screening run, the combination features *A* **B*, *A*/*B*, (*A* **B*)2 and others, can also be generated and used. To make this clear, it is possible that even if siRNA *A* and *B* individually could not be used to separate the two sets of screening run, together with the proper kernel function, they might successfully separate the two. This can be visualized graphically as well, as shown in FIG. 8. Consider each plate well as a point in multidimensional space, in which each dimension is a siRNA and the coordinate of each point is the image descriptors value of that siRNA in the plate. Using support vector machines, this high-dimensional space gains even more dimensions, representing the mathematical combinations of siRNAs. The goal for support vector machines is to find a plane in this high-dimensional space that perfectly splits two or more sets of screening run. Using this technique, the resulting plane has the largest possible margin from samples in the two conditions, therefore avoiding data over-fitting. It is clear that within this high-dimensional space, it is easier to separate siRNAs from two or more conditions (negative/positive/others), but one problem is that the separating plane is defined as a function using all the dimensions available. For example, the most accurate plane to separate one disease from another might be $(A^{*}x\ B)^{2} < 20$, where *A* and *B* are image desrciptors of siRNA. Although this might not be the most mathematically accurate way to separate two diseases and the biological significance of such functions is not always intuitive.

Unsupervised Methods

In unsupervised methods, no target variable is identified as such .Users of unsupervised methods tries to find internal structure or relationships in a data set instead of trying to determine how best to predict a 'correct answer'. Within unsupervised learning, there are three classes of technique: feature determination, or determining siRNAs with inter-

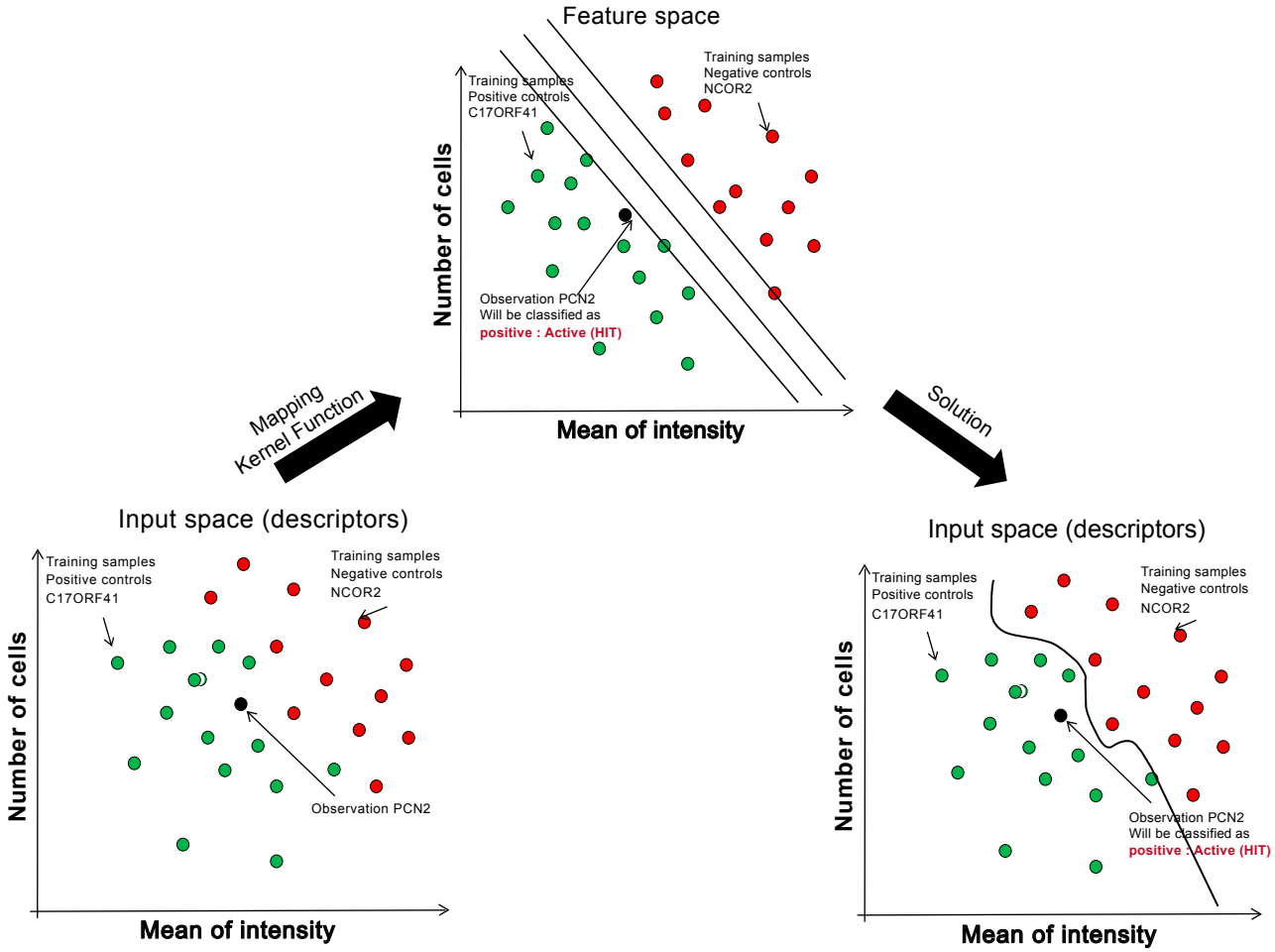

Figure 8: Support Vector Machine. Instead of restricting to individual genes, support vector machines efficiently try several mathematical combinations of siRNAs to find the line (or plane) that best separates groups of biological samples. SVMs use a training set in which genes known to be related by, for example function, are provided as positive examples and genes known not to be members of that class are negative examples. SVM solves the problem by mapping the image descriptor vectors from feature space into a higher-dimensional 'feature space', in which distance is measured using a mathematical function known as a Kernel Function, and the data can then be separated into two classes.

esting properties without specifically looking for a particular pattern, such as principal-components analysis, cluster determination, or determining groups of genes or samples with similar patterns of gene expression, such as nearest neighbour clustering, self-organizing maps, *k*-means clustering and one- and two-dimensional hierarchical clustering and network determination, or determining graphs representing siRNA–siRNA or siRNA–phenotype interactions using Boolean networks (Liang et al., 1998; Wuensche,1998; Szallasi and Liang ,1998; Friedman et al.,1998; Butte and Kohane,1999;Butte and Kohane,2000). This article will focus on the four most common unsupervised techniques of principal-components analysis, hierarchical clustering, self-organizing maps and relevance networks.

Hierarchical clustering: Hierarchical clustering is a commonly used unsupervised technique that builds clusters of siRNAs with similar patterns based on image descriptors (Box 3). This is done by iteratively grouping together siRNAs that are highly correlated in terms of their image measurements, then continuing the process on the groups themselves. There are various hierarchical clustering algorithms (Sokal and Sneath, 1963) that can be applied to microarray data analysis. These differ in the manner in which distances are calculated between the growing clusters and the remaining members of the data set, including other clusters. Clustering algorithms include:

- *Single-linkage clustering*. The distance between two clusters, *i* and *j*, is calculated as the minimum distance between a member of cluster *i* and a member of cluster *j*. Consequently, this technique is also referred to as the minimum, or nearest neighbour, method. This method tends to produce clusters that are 'loose' because clusters can be joined if any two members are close together. In particular, this method often results in 'chaining' or

Name	Description	Source
FlyRNAi	Screens carried out in the *Drosophila* RNAi Screening Center between 2002 and 2006.	http://flyrnai.org/cgi-bin/RNAi_screens.pl
DKFZ RNAi	Database contains 91351 dsRNAs from different RNAi libraries targeting transcripts annotated by the Berkeley Drosophila Genome Project.	http://www.dkfz.de/signaling2/rnai/index.php
FLIGHT	FLIGHT is a database that has been designed to facilitate the integration of data from high-throughput experiments carried out in *Drosophila* cell culture. It includes phenotypic information from published cell-based RNAi screens, gene expression data from *Drosophila* cell lines, protein interaction data, together with novel tools to cross-correlate these diverse datasets.	http://www.flight.licr.org
PhenoBank	Set of ***C. elegans* genes** for their role in the first two rounds of **mitotic cell division**. To this end, we combined genome-wide RNAi screening with time-lapse video microscopy of the early embryo.	http://www.worm.mpi-cbg.de/phenobank2
PhenomicDB	PhinomicDB is a multi-organism pheono type-genotype database including human , mouse, fruit, fly, C.elegans, and other model organisms. The inclusion of gene indeces (NCBI Gene) and orthologues (same gene in different organisms) from HomoloGene allows to compare phenoty pes of a given gene over many organisms simultaneously.PhenomicDB contains da ta from publicly available primary datab ases: FlyBase, Flyrnai.org,WormBase,Ph enobank,CYGD,MatDB,OMIM,MGI,ZFIN,SGD,DictyBase,NCBI Gene and HomoloGene.	http://www.phenomicdb.de/index.html
MitoCheck	RNA interference (RNAi) screens to identify all proteins that are required for mitosis in human cells, affinity purification and mass spectrometry to identify protein complexes and mitosis-specific phosphorylation sites on these, and small molecule inhibitors to determine which protein kinase is required for the phosphorylation of which substrate. MitoCheck is furthermore establishing clinical assays to validate mitotic proteins as prognostic biomarkers for cancer therapy.	http://www.mitocheck.org/cgi-bin/mtc
ZFIN	ZFIN serves as the zebrafish model organism database. The long term goals for ZFIN are a) to be *the* community database resource for the laboratory use of zebrafish, b) to develop and support integrated zebrafish genetic, genomic and developmental information, c) to maintain the definitive reference data sets of zebrafish research information, d) to link this information extensively to corresponding data in other model organism and human databases, e) to facilitate the use of zebrafish as a model for human biology and f) to serve the needs of the research community.	http://zfin.org
MGI	MGI is the international database resource for the laboratory mouse, providing integrated genetic, genomic, and biological data to facilitate the study of human health and disease.	http://www.informatics.jax.org

Box 3: Downloadable large data sets of RNAi screening.

the sequential addition of single samples to an existing cluster. This produces trees with many long, single-addition branches representing clusters that have grown by accretion.

- *Complete-linkage clustering*. Complete-linkage clustering is also known as the maximum or furthest-neighbour method. The distance between two clusters is calculated as the greatest distance between members of the relevant clusters. Not surprisingly, this method tends to produce very compact clusters of elements and the clusters are often very similar in size.

- *Average-linkage clustering*. The distance between clusters is calculated using average values. There are, in fact, various methods for calculating averages. The

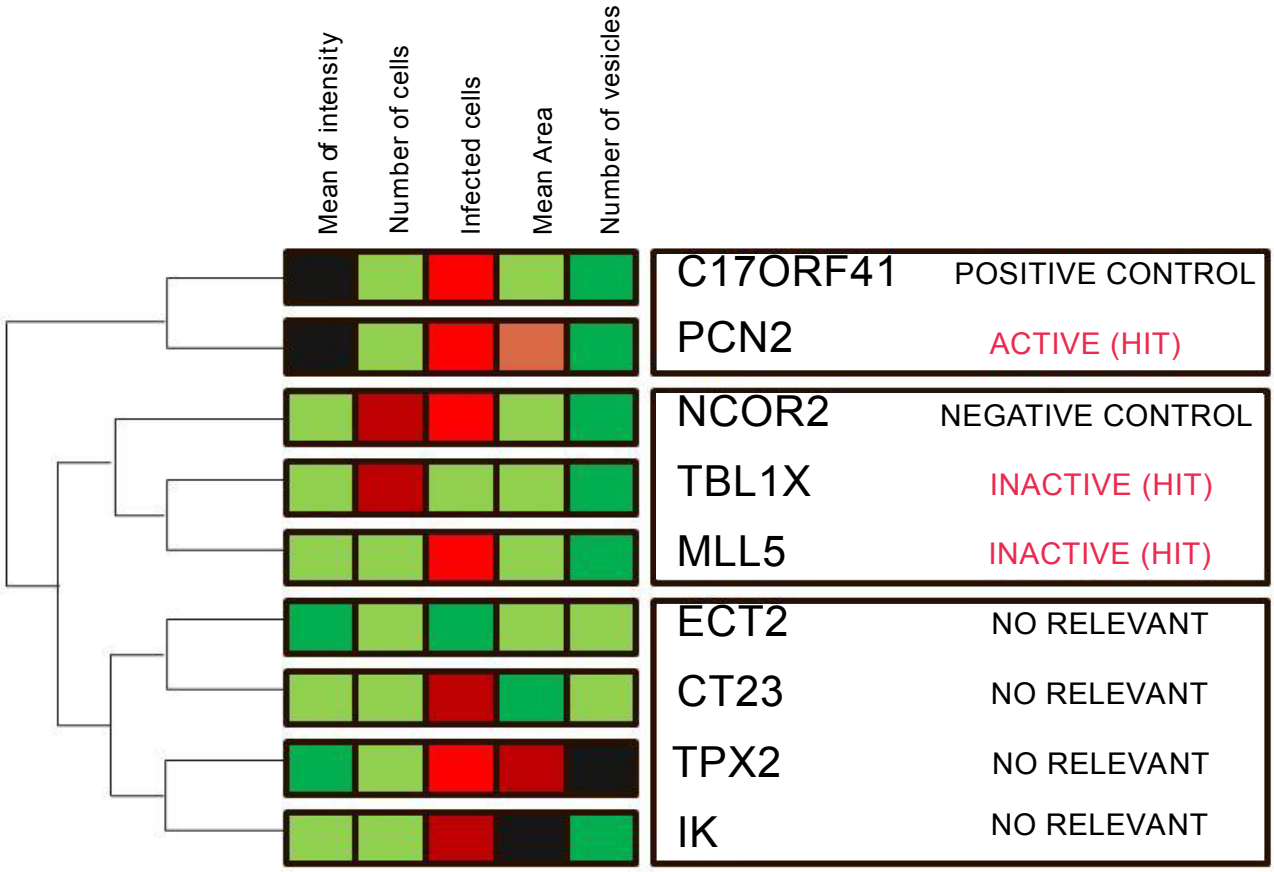

Figure 9: Hierarchical clustering. Genes in the demonstration data set were subjected to average-linkage hierarchical clustering using a Euclidean distance metric and image descriptors families that were colour coded for comparison. Similar genes appear near each other. This method of clustering groups genes by reordering the descriptors matrix allows patterns to be easily visualized. The length of the branch is inversely proportional to the degree of similarity. Shades of red indicate increased relative image descriptor; shades of green indicate decreased relative image descriptor.

most common is the unweighted pair-group method average (UPGMA).

Upgma.The average distance is calculated from the distance between each point in a cluster and all other points in another cluster. The two clusters with the lowest average distance are joined together to form a new cluster. Related methods substitute the CENTROID or the median for the average.

- *Weighted pair-group average*. This method is identical to UPGMA, except that in the computations, the size of the respective clusters (that is, the number of objects contained in them) is used as a weight. This method (rather than UPGMA) should be used when the cluster sizes are suspected to be greatly uneven.
- *Within-groups clustering*. This is similar to UPGMA except that clusters are merged and a cluster average is used for further calculations rather than the individual cluster elements. This tends to produce tighter clusters than UPGMA.
- *Ward's method*. Cluster membership is determined by calculating the total sum of squared deviations from the mean of a cluster and joining clusters in such a manner that it produces the smallest possible increase in the sum of squared errors (Ward, 1963).

DENDROGRAMS (FIG. 2, FIG. 9) are used to visualize the resultant hierarchical clustering. A dendrogram represents all genes as leaves of a large, branching tree. Each branch of the tree links two genes, two branches or one of each. Although construction of the tree is initiated by connecting genes that are most similar to each other, genes added later are connected to the branches that they most resemble. Although each branch links two elements, the overall shape of the tree can sometimes be asymmetric. In visually interpreting dendrograms, it is important to pay attention to the length of the branches. Branches connecting genes or other branches that are similar are drawn with shorter branch lengths. Longer branches represent increasing dissimilarity. Hierarchical clustering is particularly advantageous in visualizing overall similarities in image de-

scriptor patterns observed in an experiment, and because of this, the technique has been used in many publications (Pelkmans et al., 2005). The number and size of image descriptors patterns within a data set can be estimated quickly, although the division of the tree into actual clusters is often performed visually. It is important to note the few disadvantages in their use. First, hierarchical clustering ignores negative associations, even when the underlying dissimilarity measure supports them. Negative correlations might be crucial in a particular experiment, as described above, and might be missed. Furthermore, hierarchical clustering does not results in clusters that are globally optimal in that early incorrect choices in linking genes with a branch are not later reversible as the rest of the tree is constructed. So, this method falls into a category known as 'greedy algorithms', which provide good answers, but for which finding the most globally optimal set of clusters is computationally intractable. Despite these disadvantages, hierarchical clustering is a popular technique in surveying image descriptor patterns in an experiment.

Self-organizing maps: Self-organizing maps are similar to hierarchical clustering, in that they also provide a survey of image descriptors patterns within a data set, but the approach is quite different. Genes are first represented as points in multidimensional space. In other words, each biological sample (siRNA in well) is considered a separate dimension or axis of this space, and after the axes are defined, siRNAs are plotted using parameters (image descriptors) as coordinates. This is easiest to visualize with three or less siRNAs, but extends to a larger number of experiments/dimensions. Proximity can be defined using any of the dissimilarity measures described above, although Euclidean distance is most commonly used. The process starts with the answer, in that the number of clusters is actually set as an input parameter. A map is set with the centres of each cluster-to-be (known as centroids) arranged in an initial arbitrary configuration, such as a grid. As the method iterates, the centroids move towards randomly chosen genes at a decreasing rate. The method continues until there is no further significant movement of these centroids. The advantages of self-organizing maps include easy two-dimensional visualization of image patterns and reduced computational requirements compared with methods that require comprehensive pairwise comparisons, such as dendrograms. However, there are several disadvantages. First, the initial topology of a self organizing map is arbitrary and the movement of the centroids is random, so the final configuration of centroids might not be reproducible. Second, similar to dendrograms, negative associations are not easily found. Third, even after the centroids reach the centres of each cluster, further techniques are required to delineate the boundaries of each cluster. Finally, genes can belong to only a single cluster at a particular time.

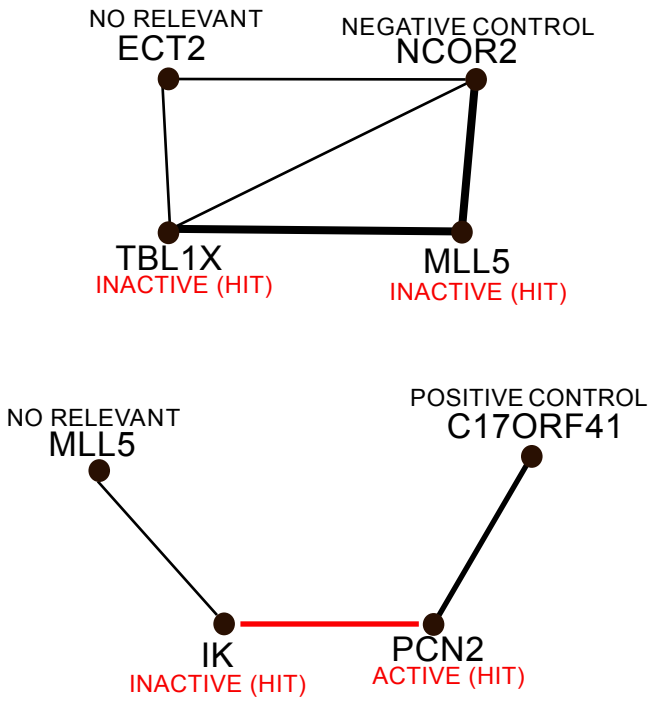

Figure 10: Relevance networks. Relevance networks find and display pairs of siRNAs with strong positive and negative correlations, then construct networks from these siRNA pairs; typically, the strength of correlation is proportional to the thickness of the lines between siRNA, and red indicates a negative correlation.

Relevance networks: Continuing through the set of unsupervised techniques, relevance networks allow networks of features to be built, whether they represent siRNA, phenotypic or clinical measurements. The technique works by first comparing all image descriptors with each other in a pairwise manner, similar to the initial steps of hierarchical clustering. Typically, two siRNA are compared with each other by plotting all the samples on a scatterplot, using image descriptors values of the two siRNAs as coordinates. A correlation coefficient is then calculated, although any dissimilarity measure can be used. A threshold value is chosen and only those pairs of features are selected which is having measure greater than the threshold. These are displayed in a graph similar to FIG. 10, in which siRNAs and phenotypic measurements are nodes, and associations are edges between nodes. Although the threshold is chosen using permutation analysis, it can actually be used as a dial, increasing and decreasing the number of connections shown. There are several advantages in using relevance networks. First, they allow features of more than one data type to be

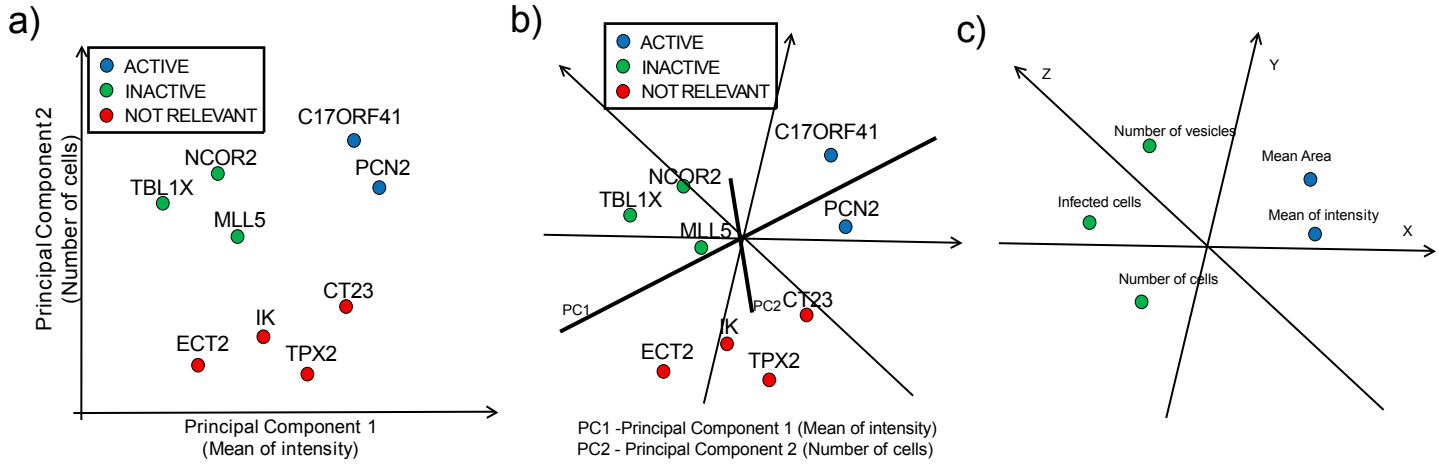

Figure 11: Principal Component Analysis. Principal-components analysis is typically used as a visualization technique, showing the clustering or scatter of siRNAs (or samples) when viewed along two or three principal components. In the figure c), a principal component can be thought of as a 'meta-biological sample', which combines all the biological samples so as to capture the most variation in image descriptors. Correlated parameters are close together, while anticorrelated parameters are in the other side of the origin. Principal components are showing the close correlation between the Mean Area and Mean of Intensity measurements.

represented together; for example, if strong enough, a link between two image descriptors (number of cells and mean if intensity) of a particular siRNA could be visualized. Second, features can have a variable number of associations; theoretically, a transcription factor might be associated with more siRNAs than a downstream component. Finally, negative associations can be visualized as well as positive ones. One disadvantage to this method is the degree of complexity seen at lower thresholds, at which many links are found associating many siRNAS in a single network. Completely connected subcomponents of these complex graphs (known as 'cliques') are not easy to find computationally.

Principal-components Analysis

PCA is used to transform a number of potentially correlated descriptors into a number of relatively independent variables that then can be ranked based upon their contributions for explaining the whole data set. The transformed variables that can explain most of the information in the data are called principal components. The components having minor contribution to the data set may be discarded without losing too much information. These dimension reduction approaches do not always work well. In order to validate the dimension reduction results, we need a technology to map a graphed point to its structure drawing.

Principal-components analysis is more useful as a visualization technique. It can be applied to either siRNA or image descirptors, which are represented as points in multidimensional space, similar to self-organizing maps. Principal components are a set of vectors in this space that decreasingly capture the variation seen in the points. The first principal component captures more variation than the second, and so on. The first two or three principal components are used to visualize the siRNA on screen or on a page, as shown in FIG. 11. Because each principal component exists in the same multidimensional space, they are linear combinations of the siRNAs. For example, the greatest variation of biological samples might be described as 3 times the particular image descriptor of the first siRNA, plus –2.1 times same descriptor of the second gene, and so on. The principal components are linear combinations that include every siRNA or image descriptor, and the biological significance of these combinations is not directly intuitive. There are other caveats in using principal components. For example, if screening runs are performed on samples from two conditions, principal components will best describe the variation of these samples. It will not always be the best way to split samples from these two conditions. Additionally PCA is a powerful technique for the analysis of screening data when used in combination with another classification technique, such as k-means clustering, or self organizing maps (SOMs) that requires the user to specify the number of clusters. Existing clusters are nice visualize in 3 dimensional space (FIG. 11b).

Challenges after Analysis

After several screening run analyses, it is quite obvious that the rate-limiting step in screening experiments is neither the handling of the biological samples nor the actual analysis, but instead the post-analytical work in determining what the results actually mean. Firstly, the detailed name

and information is not available yet for siRNA which is significant, even though these genes have been measured on screening for years. This complicates the interpretation of results. The official gene name, predicted protein domains or gene-ontology classification became available as early as tomorrow, or as late as decades from now. There are definitely post-analysis challenges are remaining. Occasionally, probe sets (wells) are incorrectly designed against the wrong strand or wrong species. Oligonucleotide sequences that were once thought to be unique for a particular gene might not remain unique as more genomic data are collected. Finally, in using well plates, particularly those for which the probe sequences have not been validated, the findings might be incorrect. Operationally, this means that analyzing a set of screening data is not finished. The infrastructure has to be developed to re-investigate genes and gene information constantly from screening analyses which were performed in the past. For example, next month, new information about a gene that was positive in the analysis performed three months ago, leads to a very innovative and an important hypothesis.

Conclusion

The challenge in determining the proper analytical methods to use is usually only a short-term difficulty, and typically, after the 'HCS pipeline' has been established, the rate-limiting step shifts to the post-analytical challenges. In the future, truly showing a 'return on investment' from HCS will depend on findings beyond the screening stage and integrating them with the rest of the discovery pipeline. The 'list of genes' resulting from a HCS should not be viewed as an end in itself; its real value increases only as that list moves through biological validation, ranging from the numerical verification of results with alternative techniques, to ascertain the meaning of the results, such as finding common promoter regions or biological relationships between the genes. However, tools that link these genes to known biological pathways, as well as discovering new pathways, are in their infancy. Tools that can automatically indicate the importance of particular findings have yet to be discovered. The analysis of screening data sets in a vacuum devoid of biological knowledge will be less rewarding. Finally, the use of HCS in basic and applied research in drug discovery is not only increasing, but as these data sets grow in size, it is important to recognize that untapped information and potential discoveries might still be present in existing data sets (Box 3). Advances in data mining of HCS data will provide objective benchmarks against which to compare experimental results and as a consequence help to standardize the hit identification process. By improving measurement quality and by providing quantifiable false-positive/false-negative ratios, data mining can improve the efficacy of nonstatistical considerations for development (such as counter screens to identify nonspecific interference). In this manner, the often-cited advice to identify false leads early and quickly can be strengthened while minimizing potentially costly false negatives. In the application of various screening and most importantly in drug discovery, to extract the most information from microarrays, an open mind always needs to be kept with regard to the choices of analytical methods, using supervised and unsupervised techniques, and methods.

References

1. Bernard P, Golbraikh A, Kireev D, Chrétien JR, Rozhkova N (1998) Comparison of chemical databases: Analysis of molecular diversity with Self Organising Maps (SOM). Analusis 26: 333-346.» CrossRef » Google Scholar
2. Boutros M, Kiger AA, Armknecht S, Kerr K, Hild M, et al. (2004) Genome-wide RNAi analysis of growth and viability in Drosophila cells.Science 303: 832-835.» CrossRef » Pubmed » Google Scholar
3. Brideau C, Gunter B, Pikounis B, Liaw A (2003) Improved statistical methods for hit selection in high-throughput screening. J Biomol Screen 8: 634-647. » CrossRef » Pubmed » Google Scholar
4. Butte A, Kohane I (1999) in Fall Symposium, American Medical Informatics Association (ed. Lorenzi, N.) 711-715.
5. Butte AJ, Kohane IS (2000) Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. Pac Symp Biocomput 418-429.» CrossRef » Pubmed » Google Scholar
6. Cox B, Denyer JC, Binnie A, Donnelly MC, Evans B, et al. (2000) Application of high-throughput screening techniques to drug discovery. Prog Med Chem 37: 83-133. » Pubmed » Google Scholar
7. Cox TF, Cox MA (2000) Multidimensional Scaling, Chapman & Hall/CRC.
8. Friedman N, Linial M, Nachman I, Pe'er D (2000) Using Bayesian networks to analyze expression data. J Comput Biol 601-620. » CrossRef » Pubmed » Google Scholar
9. Giuliano KA, Haskins JR, Taylor DL (2003) Advances in high content screening for drug discovery. Assay and Drug Development Technologies 1: 565-577. » CrossRef » Pubmed » Google Scholar
10. Gunter B, Brideau C, Pikounis B, Liaw A (2003) Statistical and graphical methods for quality control determination of high throughput screening data. J Biomol Screen 8: 624-633.» CrossRef » Pubmed » Google Scholar

11. Heuer C, Haenel T, Prause B (2002) A novel approach for quality control and correction of HTS data based on artificial intelligence. The Pharmaceutical Discovery & Development Report 2003/03, PharmaVentures Ltd.

12. Heyse S (2002) Comprehensive analysis of high-throughput screening data. Proc SPIE 4: 535-547. » CrossRef » Google Scholar

13. Johnston PA, Johnston PA (2002) Cellular platforms for HTS: three case studies. Drug Discov Today 7: 353-363. » CrossRef » Pubmed » Google Scholar

14. Kevorkov D, Makarenkov V (2005) Statistical analysis of systematic errors in high-throughput screening. J Biomol Screen 10: 557-567.» CrossRef » Pubmed » Google Scholar

15. Kittler R, Putz G, Pelletier L, Poser I, Heninger AK, et al. (2004) An endoribonuclease-prepared siRNA screen in human cells identifies genes essential for cell division. Nature 432: 1036-1040. » CrossRef » Pubmed » Google Scholar

16. Kohonen T, Kangas J , Laaksonen J (1992) SOM_PAK, The Self-Organizing Map Program Package. Version 1.2.

17. Liang S, Fuhrman S, Somogyi R (1998) Reveal, a general reverse engineering algorithm for inference of genetic network architectures. Pac Symp Biocomput: 18-29. » CrossRef » Pubmed » Google Scholar

18. Lutz M, Kenakin T (2000) Quantitative molecular pharmacology and informatics in drug discovery, John Wiley & Sons, New York.

19. Malo N, Hanley JA, Cerquozzi S, Pelletier J, Nadon R (2006) Statistical practice in high throughput screening data analysis. Nature Biotechnology 24: 167-175. » CrossRef » Pubmed » Google Scholar

20. Moffat J, Grueneberg DA, Yang X, Kim SY, Kloepfer AM, et al. (2006) Lentiviral RNAi Library for Human and Mouse Genes Applied to an Arrayed Viral High-Content Screen. Cell 124: 1283-1298.» CrossRef » Pubmed » Google Scholar

21. Monk AJ (2005) Faster, surer prediction. The Biochemist pp25-28. » CrossRef » Google Scholar

22. Pelkmans L, Fava E, Grabner H, Hannus M, Habermann B, et al. (2005) Genome-wide analysis of human kinases in clathrin- and caveolae/raft-mediated endocytosis. Nature 436: 78-86. » CrossRef » Pubmed » Google Scholar

23. Sokal RR, Sneath PHA (1963) Principles of Numerical Taxonomy. (W. H. Freeman & Co., San Francisco).

24. Stat Soft Inc. Multidimensional Scaling.

25. Szallasi Z, Liang S (1998) Modeling the normal and neoplastic cell cycle with "realistic Boolean genetic networks": their application for understanding carcinogenesis and assessing therapeutic strategies. Pac Symp Biocomput 66-76. » CrossRef » Pubmed » Google Scholar

26. Taylor DL, Woo ES, Giuliano KA (2001) Real-time molecular and cellular analysis: the new frontier of drug discovery. Curr Opin Biotechnol 12: 75-81. » CrossRef » Pubmed » Google Scholar

27. Ward JH (1963) Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association 58: 236-244.» CrossRef » Google Scholar

28. Wuensche A (1998) Genomic regulation modeled as a network with basins of attraction. Pac Symp Biocomput 89-102. » CrossRef » Pubmed » Google Scholar

29. Zhang JH, Chung TD, Oldenburg KR (1999) A Simple Statistic Parameter for Use in Evaluation and Validation of High Throughput Screening Assays. J Biomol Screen 4: 67-73.» CrossRef » Pubmed » Google Scholar

30. Zhang JH, Chung TD, Oldenburg KR (2000) Confirmation of primary active substances from high throughput screening of chemical and biological populations: a statistical approach and practical considerations. J Comb Chem 2: 258-265. » CrossRef » Pubmed » Google Scholar

31. 31) Zupan J, Gasteiger J (1993) Neural Networks for Chemists. VCH: Weinheim.