(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
30 September 2004 (30.09.2004)

PCT

(10) International Publication Number
WO 2004/083451 A1

(51) International Patent Classification[7]: C12Q 1/00, G01N 33/48, G06F 19/00, 159/00

(21) International Application Number: PCT/AU2004/000327

(22) International Filing Date: 17 March 2004 (17.03.2004)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
2003901196 17 March 2003 (17.03.2003) AU

(71) Applicant *(for all designated States except US)*: **COMMONWEALTH SCIENTIFIC AND INDUSTRIAL RESEARCH ORGANISATION** [AU/AU]; Limestone Avenue, Campbell, Australian Capital Territory 2612 (AU).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **THOMAS, Mervyn** [AU/AU]; c/o- COMMONWEALTH SCIENTIFIC AND INDUSTRIAL RESEARCH ORGANISATION, Limestone Avenue, Campbell, Australian Capital Territory 2612 (AU). **KIIVERI, Harri** [AU/AU]; c/o- COMMONWEALTH SCIENTIFIC AND INDUSTRIAL RESEARCH ORGANISATION, Limestone Avenue, Campbell, Australian Capital Territory 2612 (AU). **MITCHELL, David** [AU/AU]; c/o- COMMONWEALTH SCIENTIFIC AND INDUSTRIAL RESEARCH ORGANISATION, Limestone Avenue, Campbell, Australian Capital Territory 2612 (AU).

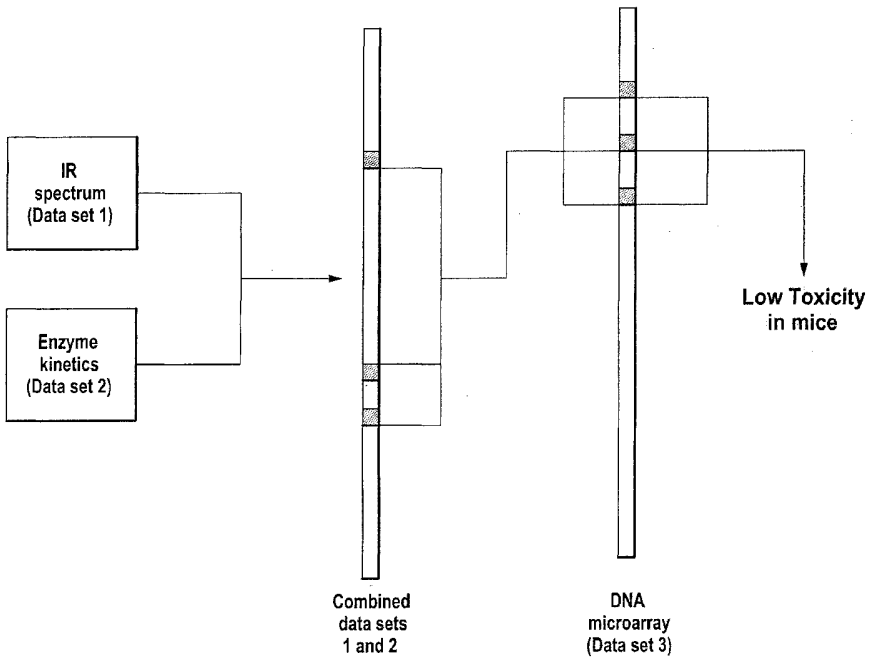
(74) Agent: **GRIFFITH HACK**; GPO Box 4164, Sydney, New South Wales 2001 (AU).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NA, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH,

[Continued on next page]

(54) Title: ANALYSIS METHOD



(57) Abstract: A method of determining a signature indicative of an entity having a condition, the method comprising the steps of: (a) obtaining at least two distinct data sets, at least one of the data sets containing information on at least one characteristic of at least one entity having the condition; and (b) processing the information from the data sets to obtain a signature that is indicative of an entity having the condition.

WO 2004/083451 A1

GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IT, LU, MC, NL, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:
— *with international search report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

ANALYSIS METHOD

The invention relates to a method and apparatus for identifying a signature indicative of an entity having a condition and for using the signature to identify entities having that condition. The invention has been developed especially, but not exclusively, for identifying samples having a desired therapeutic or biological property and the invention is herein described in that context. However, it is to be appreciated that the invention has broader application and may be used in other complex areas, such as financial modelling and the like.

BACKGROUND OF THE INVENTION

In many systems, the screening of entities for those having a particular condition can be time consuming and expensive. This is especially the case if the entity is one of a large number of entities, or the number of tests that need to be applied to assess whether the entity has a condition are many and/or are expensive. This is especially problematic in applications where large amounts of data are produced for each entity.

An example of an application in which large amounts of data are produced for each entity is in drug discovery. Drug discovery involves the screening of thousands of compounds in a plurality of systems to determine which compounds have a therapeutic or biologically active property. Each system of the drug discovery process generally provides information about one particular characteristic of the compound such as, for example, a chemical or physical property, a gene or protein

expression profile in a particular tissue or cell line in response to the compound, enzyme kinetic properties for a particular enzyme or group of enzymes in response to the compound etc. However, the information obtained about each characteristic does not provide sufficient information to make predictions about the ultimate therapeutic or biological properties of the compound.

Typically, the drug discovery process involves screening many thousands of compounds initially in order to identify a single compound having a desired property. With each successive test, compounds which do not exhibit the property sought in each system are eliminated. Thus, by the time the compounds are assessed in a plurality of systems, many thousands of compounds are reduced down to one or a few.

The drug discovery process generally begins with many compounds (either naturally occurring compounds, or synthetically derived compounds such as those generated using methods such as combinatorial chemistry). The compounds are subjected to a plurality of systems such as biological and chemical analysis, enzymology, genomics and proteomics to mass screen compounds for desired properties in each system (e.g. high efficacy, low toxicity, high stability). Those compounds identified as having desired properties, known as lead compounds, are then subjected to preclinical testing which may include, for example, pharmacological target testing, formulation and stability testing, and assessment of the compounds ability to be absorbed, distributed, metabolised and excreted (ADME) in test animals. Those lead compounds that exhibit the desired property in the preclinical testing then proceed

to clinical testing. Clinical testing is conducted in three phases (I, II and III).

Phase I clinical testing assesses the tolerance, pharmacological effects, absorption, distribution, metabolism and excretion (ADME) in approximately 20-80 healthy human volunteers. Phase II clinical testing is conducted on 100-300 patients which have the targeted medical condition to determine the effectiveness of treating the medical condition. Phase III is conducted on approximately 1,000-5,000 subjects to determine clinical benefit and incidence of adverse reaction. Following a successful phase III clinical trial, the compound and the results of the clinical trial must be approved by regulatory bodies before the drug is allowed to enter the market.

The total cost of conducting the process from synthesising a compound to reaching the market has been estimated to be approximately US$800 million per compound and can take 15 years or more for a screening of 5,000 to 10,000 compounds. Of these 5,000 to 10,000 compounds, 250 may enter preclinical testing, and of these only 5 survive to enter clinical testing. The screening process to identify lead compounds and precinical testing of these lead compounds account for approximately 36% of the entire cost of drug development.

Clearly, there is a need for a method that is capable of identifying compounds which are more likely to exhibit a desired property than other compounds using a reduced number of test systems and/or reduced amount of data.

SUMMARY OF THE INVENTION

In a first aspect, the invention provides a method to determine a signature indicative of an entity having a condition, comprising the steps of;

(a) obtaining at least two distinct data sets, at least one of the data sets containing information on at least one characteristic of at least one entity having the condition; and

(b) processing the information from the data sets to obtain a signature that is indicative of an entity having the condition.

An advantage of the present invention is that it draws on more than one data set to generate a signature of the condition. Each data set may be derived from a system and may include a massive amount of data. In contrast to existing techniques of screening that uses successive tests in isolation to screen for a condition in a target, the present invention is better able to isolate condition signatures because it is able to evaluate relationships between data sets as well as relationships within data sets.

In the context of the present invention, a "signature" is a subset of one or more of the sets of data that is indicative of a condition. Also in the context of the specification, a "system" means a process or arrangement that obtains data is some organised way. The system may be any test, trial, experiment, analysis, measurement or assessment which provides in an organised way that is relevant to a characteristic of the entity. In areas of

drug discovery, a system may be a chemical or biological system.

The distinct data sets may be related or unrelated data sets. The distinct data sets may be related in that they contain information about the same characteristic of the entity, but the information is obtained in different ways. Alternatively, the distinct data sets may be obtained using different systems but contain information on the same characteristic. For example, in the context of drug discovery, two distinct data sets may be values for a characteristic such as gene expression in response to a sample in which one data set is obtained by analysing a microarray containing DNA or RNA. It is also envisaged that microarray technologies could be used to measure genetic variations, for example, single nucleotide polymorphisms.

The distinct data sets may be obtained using the same systems but in a different way. For example, using a microarray containing DNA in which gene expression in response to a sample is obtained from different cell lines.

The distinct data sets may be unrelated in that each data set is obtained from different systems, and the information of each data set is on a different characteristic.

The characteristic may be any characteristic of the entity. For example, where the entity is a compound for therapeutic or biological use, the characteristic may be physical or chemical properties of the compound, or

biological properties such as, for example, gene expression profile or protein expression profile or antibody binding profile.

Preferably, the information in the data sets relates to components of a system. For example, in the case of a system such as IR-spectrum analysis, the components may be the individual IR spectrum values, or where the system is a DNA microarray, the components may be genes.

The signature that is indicative of the condition may come from data in only one of the data sets, or from any combination of data sets. As such, the signature may be based on one or more characteristics of the entity. In one embodiment, the signature is data generated by a subset of the components of at least one characteristic. Preferably, the subset of components is a minimum number of components that is indicative of the condition.

In one embodiment, the processing step includes using a multivariate analysis to obtain the signature that is indicative of the condition.

Preferably, the multivariate analysis comprises the steps of:

(a) assigning a weight to each component of the data sets;

(b) defining a function of the data and the weights which models the relationship between the condition and the components; (c) applying an algorithm to update the weights using the function in (b) so as to generate a processed weight for each component;

(d) generating a processed data set containing components having a processed weight with absolute value above a

predetermined threshold.

Preferably, an assumption is made in setting the assigned weights. The assumption may be any assumption made about the data. In one embodiment, the assumption is a penalty. Preferably, the penalty is a prior assumption. Preferably, the prior assumption is a prior distribution.

Preferably, the multivariate analysis is iterative, with steps (c)and(d) being repeated using the processed data sets in a subsequent cycle of the analysis. In the analysis, the processed weights are used as a basis to establish the assigned weight for the subsequent cycle. As such, the analysis comprises the further steps of;

(a) assigning weights to the components of a said processed data set;

(b) applying the algorithm to fit the weighted data to the function so as to generate a revised processed weight

for each component;

(c) generating a revised processed data set containing components having a revised processed weight above a predetermined threshold;

(d) repeating steps (a) to (c).

Preferably, steps (a) to (c) are repeated until the processed weights remain constant to within a set tolerance with each repeat of steps (a) to (c).

Preferably, the weighted data is in the form of a combination, most preferably a linear combination.

The function is typically a mathematical model of the condition. Preferably, the mathematical model is a model of the distribution of the condition. Preferably, the model of the distribution is a model of the probability distribution of the condition. The model of the probability distribution is preferably a likelihood function.

Preferably, the multivariate analysis estimates component weights utilising a Bayesian statistical method. Preferably, where the at least two distinct data sets comprise a large amount of data, the method preferably makes an a priori assumption that the majority of the data is unlikely to be data that will form part of the signature that is indicative of the condition. The assumption is therefore made that the majority of weights are likely to be zero. A model is constructed which, with this assumption in mind, allows the algorithm to set the weights so that the posterior probability of the processed weights is maximised. The process is iterated until the processed weights remain constant. This method is quick, mainly because of the a priori assumption which results in rapid elimination of the majority of the components.

In one embodiment, the likelihood function is based on a multinomial or binomial logistic regression. The binomial or multinomial logistic regression preferably models a condition of the entity that exhibits a multinomial or binomial distribution. A binomial distribution is a statistical distribution having two possible classes or

groups such as an on/off state. Examples of such groups include dead/alive, improved/not improved, depressed/not depressed. A multinomial distribution is a generalisation of the binomial distribution in which a plurality of classes or groups are possible for each of a plurality of entities, or in other words, an entity may be classified into one of a plurality of classes or groups. Thus, by defining a likelihood function based on a multinomial or binomial logistic regression, it is possible to identify a signature that is capable of classifying an entity into one of a plurality of pre-defined groups or classes. For example, the results from microarray data may be classified into pre-defined groups. To do this, samples are grouped into a plurality of sample groups (or "classes") based on a predetermined property of the samples in which the members of each sample group have a common property and are assigned a common group identifier. A likelihood function is formulated based on a multinomial or binomial logistic regression conditional on the linear combination (which incorporates the data generated from the grouped samples). The condition may be any desired classification by which the results are to be grouped. For example, the property for classification of a chemical compound in a mouse model based on data generated from microarray experiments may be no toxicity, low toxicity, medium toxicity or high toxicity.

Preferably, the likelihood function based on the logistic regression is of the form:

$$L = \prod_{i=1}^{n} \left(\prod_{g=1}^{G-1} \left\{ \frac{e^{x_i^T \beta_g}}{\left(1 + \sum_{g=1}^{G-1} e^{x_i^T \beta_g} \right)} \right\}^{e_{ik}} \left\{ \frac{1}{1 + \sum_{h=1}^{G-1} e^{x_i^T \beta_h}} \right\}^{e_{iG}} \right)$$

wherein

$x_i^T \beta_g$ is a linear combination generated from input data from training sample i with component weights β_g;

x_i^T is the components for the i^{th} Row of X and β_g is a set of component weights for sample class g; and

X is data from n training samples comprising p components.

In another embodiment, the likelihood function is based on an ordered categorical logistic regression. The ordered categorical logistic regression models a binomial or multinomial distribution in which the classes are in a particular order (ordered classes such as for example, classes of increasing or decreasing disease severity). By defining a likelihood function based on an ordered categorical logistic regression, it is possible to identify a signature that is capable of predicting a result class wherein the class is one of a plurality of predefined ordered classes. By defining a series of group indentifiers in which each group identifier corresponds to a member of an ordered class, and grouping the training samples into one of the ordered classes based on a predetermined condition , a likelihood function can be formulated based on a categorical ordered logistic regression which is conditional on the linear combination (which incorporates the data generated from the grouped training samples).

Preferably, the likelihood function based on the categorical ordered logistic regression is of the form:

$$l = \prod_{i=1}^{N} \prod_{k=1}^{G-1} \left(\frac{\gamma_{ik}}{\gamma_{ik+1}} \right)^{r_{ik}} \left(\frac{\gamma_{ik+1} - \gamma_{ik}}{\gamma_{ik+1}} \right)^{r_{ik+1} - r_{ik}}$$

Wherein

γ_{ik} is the probability that data from training sample i belongs to a class with identifier less than or equal to k (where the total of ordered classes is G) and r_{ik} is defined in Section B below.

In another embodiment of the present invention, the likelihood function is based on a generalised linear model for the condition. The generalised linear model preferably models a condition that is distributed as a regular exponential family of distributions. Examples of regular exponential family of distributions include normal distribution, Gaussian distribution, poisson distribution, gamma distribution and inverse gamma distribution. Thus, in another embodiment of the method of the invention, a signature is identified that is indicative of a predefined condition of a sample that lies within a regular exponential family of distributions by defining a generalised linear model which models the condition to be predicted. Examples of a condition that may be predicted using a generalised linear model include any quantity of a sample that exhibits the specified distribution such as, for example, the weight, size or other dimensions or quantities of a sample.

Preferably, the generalised linear model is of the form:

$$\mathrm{L} = \log \mathrm{p}(\mathrm{y} \mid \beta, \phi) = \sum_{i=1}^{N} \{ \frac{y_i \theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi) \}$$

where $y = (y_1, \ldots, y_n)^T$ and $a_i(\phi) = \phi / w_i$ with the w_i being a fixed set of known weights and ϕ a single scale parameter.

Further details are discussed in Section C below.

In another embodiment, the method of the present invention may be used to predict the time to an event for an entity by utilising a likelihood function based on a hazard model which preferably estimates the probability of a time to an event given that the event has not taken place at the time of obtaining the information. In one embodiment, the likelihood function is based on a model selected from the group consisting of Cox's proportional hazards model, parametric survival model and accelerated failure times model. Cox's proportional hazards model permits the time to an event to be modelled on a set of components and component weights without making restrictive assumptions about time. The accelerated failure model is a general model for data consisting of survival times in which the component measurements are assumed to act multiplicatively on the time-scale, and so affect the rate at which an individual proceeds along the time axis. Thus, the accelerated survival model can be interpreted in terms of the speed of progression of, for example, disease. The parametric survival model is one in which the distribution function for the time to an event in a second system (eg survival time) is modelled by a known distribution or has a specified parametric formulation. Among the commonly used survival distributions are the Weibull, exponential and extreme value distributions.

Preferably, a signature capable of predicting a condition that involves a time to an event is identified by defining a likelihood based on Cox's proportional hazards model, a parametric survival model or an accelerated survival times model, which comprises measuring the time elapsed for a

plurality of samples from the time the sample is obtained to the time of the event.

Preferably, the likelihood function for predicting the time to an event is of the form:

$$Log\ (Partial)\ Likelihood\ = \sum_{i=1}^{N} g_i\left(\underset{\sim}{\beta}, \underset{\sim}{\varphi}; X, \underset{\sim}{y}, \underset{\sim}{c}\right)$$

where $\underset{\sim}{\beta}^T = \left(\beta_1, \beta_2, \cdots, \beta_p\right)$ and $\underset{\sim}{\varphi}^T = \left(\varphi_1, \varphi_2, \cdots, \varphi_q\right)$ are the model parameters.

Preferably, the likelihood function based on Cox's proportional hazards model is of the form:

$$l\left(\underset{\sim}{t} \mid \underset{\sim}{\beta}\right) = \prod_{j=1}^{N} \left(\frac{exp\left(Z_j \underset{\sim}{\beta}\right)}{\sum_{i \in \Re_j} exp\left(Z_i \underset{\sim}{\beta}\right)} \right)^{d_j}$$

where $\underset{\sim}{\beta}^T = \left(\beta_1, \beta_2, \cdots, \beta_n\right)$, Z_j = the j^{th} row of Z, and $\Re_j = \left\{i : i = j, j+1, \cdots, N\right\}$ = the risk set at the j^{th} ordered event time $t_{(j)}$.

Preferably the likelihood function based on the Parametric Survival model is of the form:

$$L = \sum_{i=1}^{N} \left\{ c_i log\left(\mu_i\right) - \mu_i + c_i \left(log\left(\frac{\lambda\left(y_i\right)}{\Lambda\left(y_i; \underset{\sim}{\varphi}\right)} \right) \right) \right\}$$

where $\mu_i = \Lambda\left(y_i; \underset{\sim}{\varphi}\right) exp\left(X_i \underset{\sim}{\beta}\right)$. The notation used in the above proportional hazards models is explained further in

Section D below.

For any defined models, the weights are typically estimated using a Bayesian statistical model (Kotz and Johnson, 1983) in which a posterior distribution of the weights is formulated which combines the likelihood function and a prior distribution. The processed weights are estimated by maximising the posterior distribution of the weights given the data generated for each entity. Thus, the objective function to be maximised consists of the likelihood function based on a function for the condition as discussed above and a prior distribution for the weights.

Preferably, the prior distribution is of the form:

$$p(\beta) = \int_{v^2} p\left(\beta \,|v^2\right) p\left(v^2\right) dv^2$$

wherein v is a p x 1 vector of hyperparameters, and where $p\left(\beta \,|v^2\right)$ is $N\left(0, \text{diag}\left\{v^2\right\}\right)$ and $p\left(v^2\right)$ is some hyperprior distribution for v^2. This hyperprior distribution (which is preferably the same for all embodiments of the method) may be expressed using different notational conventions, and in the detailed description of the preferred embodiments (see below), the following notational conventions are adopted merely for convenience for the particular preferred embodiment.

As used herein, when the likelihood function for the probability distribution is based on a multinomial or binomial logistic regression, the notation for the prior distribution is:

$$P(\beta_1...,\beta_{G-1}) = \int_{\tau^2} \prod_{g=1}^{G-1} P(\beta_g|\tau_g^2) P(\tau_g^2) d\tau^2$$

where $\beta^T = (\beta_1^T,...\beta_{G-1}^T)$ and $\tau^T = (\tau_1^T,...,\tau_{G-1}^T)$.

and $p(\beta_g|\tau_g^2)$ is $N(0,\text{diag}\{\tau_g^2\})$ and $P(\tau_g^2)$ is some hyperprior distribution for τ_g^2.

As used herein, when the likelihood function for the probability distribution is based on a hazard model , the notation for the prior distribution is:

$$P(\beta_1,\beta_2,\cdots,\beta_n) = \int_{\tau} \prod_{i=1}^{n} P(\beta_i|\tau_i) P(\tau_i) d\tau$$

where $\beta_1,\beta_2,\cdots,\beta_n$ are component weights, $P(\beta_i|\tau_i)$ is $N(0,\tau_i^2)$ and $P(\tau_i)$ some hyperprior distribution for τ_i.

As used herein, when the likelihood function for the distribution is based on a generalised linear model, the notation for the prior distribution is:

$$p(\beta) = \int_{v^2} p(\beta|v^2) p(v^2) dv^2$$

wherein v is a p x 1 vector of hyperparameters, and where $p(\beta|v^2)$ is $N(0,\text{diag}\{v^2\})$ and $p(v^2)$ is some prior distribution for v^2.

As used herein, when the likelihood function for the distribution is based on a ordered categories model, the notation for the prior distribution is:

$$p(\beta^*) = \int_{\nu^2} p(\beta^*|\nu^2) p(\nu^2) d\nu^2$$

where $p(\beta^*|\nu^2)$ is $N(0, \text{diag}\{\nu^2\})$ and $p(\nu^2)$ some hyperprior distribution for ν^2.

As used herein, when the likelihood function for the distribution is based on a hazard model, the notation for the prior distribution is:

$$p(\beta) = \int_{\nu^2} p(\beta|\tau) p(\tau) d\tau$$

where $p(\beta|\tau)$ is $N(0, \text{diag}\{\tau^2\})$ and $p(\tau)$ some hyperprior distribution for ν^2.

The prior distribution preferably comprises a hyperprior that ensures that zero weights are preferred whenever possible.

Preferably, the hyperprior is a Jeffrey's prior.

As discussed above, the prior distribution and the likelihood function are combined to generate a posterior distribution. The posterior distribution is preferably of the form:

$$p(\beta \varphi \nu | y) \alpha L(y|\beta\varphi) p(\beta|\nu) p(\nu)$$

or

$$P(\underset{\sim}{\beta}, \underset{\sim}{\varphi}, \underset{\sim}{\tau} | \underset{\sim}{y}) \alpha L(\underset{\sim}{y} | \underset{\sim}{\beta}, \underset{\sim}{\varphi}) P(\underset{\sim}{\beta} | \underset{\sim}{\tau}) P(\underset{\sim}{\tau})$$

wherein $L\left(\underset{\sim}{y}\middle|\underset{\sim}{\beta},\underset{\sim}{\varphi}\right)$ is the likelihood function.

The algorithm fits the model to the data by adjusting the weights to maximise the probability density of the posterior distribution of the function. The algorithm may be any algorithm capable of fitting the data to the model by adjusting the weights. The algorithm may be an iterative procedure or an optimisation procedure. Examples of optimisation procedures include genetic algorithms, annealing algorithms, Newton algorithms. Examples of iterative algorithms include gradient descent algorithm, EM algorithm, simplex algorithm, Fletcher-Powell algorithm, conjugative gradients.

The processed weights in the posterior distribution are preferably estimated in an iterative procedure. During the iterative procedure, processed weights having a value less than a pre-determined threshold are eliminated, preferably by setting those weights to zero. This results in elimination of the corresponding component.

Preferably, the iterative procedure is an EM algorithm. The EM algorithm produces a sequence of weight estimates that converge to give processed weights that maximise the probability density of the posterior distribution. The EM algorithm consists of two steps, known as the E or Expectation step and the M, or Maximisation step. In the E step, the expected value of the log-posterior function conditional on observed data and the linear combination is determined. In the M step, the expected log-posterior function is maximised to give updated weight estimates that increase the likelihood. The two steps are

alternated until convergence of the E step and the M step is achieved, or in other words, until the expected value and the maximised value of the log-likelihood function converge.

The at least one entity having the condition is preferably one of a plurality of entities having a known condition which are used to "train" the processing step to obtain a signature. Thus, the at least one entity is preferably one of a plurality of entities from which at least two distinct data sets are obtained, at least one of which contains information about at least one characteristics of at least one entity having the condition. Preferably, the plurality of entities comprises one or more entities that have the condition and one or more entities that do not have the condition.

It is envisaged that the method of the present invention may be used in virtually any application to determine a signature indicative of an entity having a condition. The inventors envisage that the method may be used in applications such as drug development (e.g. drug discovery pipeline), quantitative trait loci analysis for genetic identification or selection of desired phenotypes, financial analysis and/or financial prediction such as stock prediction based on various financial and non-financial parameters (components), manufacturing in, for example, efficiency management or prediction of product failure, market analysis to predict desired outcomes in product sales or target markets, sport for the prediction of outcomes in competition, weather for the prediction of a particular outcome in a particular geographic region.

The entity may be anything for which a condition is sought and which has at least two characteristics. For example, if the method of the invention is used to determine a signature that is indicative of the share price of a company (the condition), the entity may be a company. The characteristics of the company may be, for example, account figures, stock prices of other companies, employee numbers, market estimates, or any other information that may be relevant to the price of shares of the company. In another example, the entity may be a particular city and the condition may be rainfall. The characteristics relevant to the city may be rainfall in other cites around the world, humidity in other cites around the world or any other information that may be relevant to the rainfall in that particular city. In one embodiment, the entity is a sample. Preferably, the sample is a compound. Preferably, the compound is synthesised or naturally occurring.

It will be apparent to persons skilled in the art that the condition of the entity will depend on the nature of the entity and the use for which the entity is intended. Preferably, the condition may be a property of the entity. Preferably, the property is a desired property of the entity. For example, if the entity is a compound and the use is for treatment of a disease, then the condition may be, for example, the ability to alleviate symptoms of the disease, or low toxicity, or ability to bind a drug target. If the entity is a company, then the condition may be, for example, a 10% rise in share price.

It is envisaged that the condition may be that the entity will produce a data set in at least one further system.

The data set produced in the further system may comprise a signature that is indicative that the entity has a further condition. Thus, the method of the present invention may be used to determine a signature that is indicative of a signature in a further at least one system, wherein the signature is indicative of a further condition.

Once a signature is identified using the method of the invention, the signature may then be used to identify an entity having the condition.

Thus, in a second aspect, the invention provides a method for determining whether a test entity has a condition comprising the steps of:

(a) determining a signature indicative of the condition in accordance with the first aspect; and
(b) obtaining data about a test entity;
(c) determining whether the information about the test entity contains data equivalent to the signature, to determine whether the test entity has the condition.

It will be appreciated by those skilled in the art that in determining whether a test entity has a condition, only those components from which data for the signature is obtained need be considered. Thus, the amount of data that is generated may be greatly reduced.

Also contemplated are computer programs, apparatus and systems for determining a signature that is indicative of an entity having a condition from at least two distinct data sets containing information on at least one

characteristic of at least one entity having the condition.

In a third aspect, the invention provides a computer program, arranged, when run on a computing device, to control the computing device to process information from at least two distinct data sets, at least one of the data sets containing information on at least one characteristic of at least one entity having a condition, to obtain a signature that is indicative of the condition of an entity having the condition.

The computer program may implement any of the preferred algorithms and method steps of the processing steps which are discussed above.

Preferably, the computer program is provided on a computer readable medium.

The computer program may be arranged to implement any of the preferred method and calculation steps discussed above in relation to the processing steps.

In a fourth aspect, the invention provides an apparatus for processing information from at least two distinct data sets, at least one of the data sets containing information on at least one characteristic of at least one entity having a condition, to obtain a signature that is indicative of the condition of an entity having the condition.

In a fifth aspect, the invention provides a system which is arranged to identify data obtained from processing information from at least two distinct data sets, at least one of the data sets containing information on at least one characteristic of at least one entity having a condition, to obtain a signature that is indicative of the condition of an entity having the condition, the system comprising a user interface and a storage means arranged to store a processing template, the template including pre-stored model parameters for controlling the system in accordance with the parameters and wherein a user is able to select the processing template via the interface and the system is arranged to carry out processing of the information from at least two characteristics in accordance with the model parameters.

Where aspects of the present invention are implemented by way of a computing device, it will be appreciated that any appropriate computer hardware e.g. a PC or a mainframe or a networked computing infrastructure, may be used.

In a sixth aspect of the present invention, there is provided a method of determining whether a substance has a potential for use as a drug, the method comprising the steps of:

obtaining a first mathematical model that is derived using a signature of a condition of an entity;

using the first mathematical model in conjunction with information about at least one characteristic of another entity that has been exposed to the substance; and

using a result from the first mathematical model to determine whether the substance has the potential for use as the drug.

Preferably, the signature is obtained using the method as described in the first aspect of the present invention.

Preferably, the method further comprising the step of selecting the substance from a plurality of substances by using a second mathematical model in conjunction with other information about the plurality of substances.

Preferably, the first mathematical model is such that the result represents a probability that the substance is not toxic to the other entity.

Preferably, the second mathematical model is such that it represents a toxicity measurement for the substance in relation to the other entity.

The apparatus of the third aspect preferably includes an appropriately programmed processor for implementing the method steps.

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1 illustrates a schematic diagram of a preferred embodiment in which two data sets are used to identify a signature that is indicative of low toxicity in mice.

Figure 2 illustrates a schematic diagram of a preferred embodiment in which a many data sets are used to identify a signature that is indicative of a condition.

Figure 3 illustrates a flow diagram of an embodiment of the method of the present invention showing the steps in processing the data sets.

Figures 4 to 6 illustrate flow diagrams of the various steps carried out in the preferred embodiment of the present invention.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENT

The method of the invention may be used to determine a signature indicative of a sample having a desired property from data generated from a plurality of training samples.

In a preferred embodiment, the invention provides a method for determining a signature that is indicative of whether a test sample has a desired property, the method comprising the steps of:

(a) Obtaining distinct data sets containing information on at least one characteristic of a training sample, the data sets being obtained from a plurality of training samples, each applied to a plurality of systems, wherein one or more of the training samples have the desired property;

(b) using multivariate analysis of the information from the data sets to obtain a signature that is indicative of a training sample having the desired property.

This embodiment of the invention permits identification of a signature that is indicative of whether a sample has a desired property by utilising training samples applied to a plurality of systems to identify a signature that is indicative of the desired property. Preferably, the signature is data generated from a subset of components of

the characteristic. The method utilises training samples having a known condition in order to identify a signature for a subset of components which is indicative of a desired feature for a training sample having the desired feature. Subsequently, knowledge of the subset of components can be used for test samples to determine whether the test samples have the desired property. In future screenings for samples having the desired property, it is necessary only to measure the subset of components to determine whether a test sample has the desired property.

In another embodiment, the invention therefore provides a method for determining whether a test sample has a desired property, comprising the steps of:

(a) obtaining test data from the test sample;
(b) determining the signature value for the test sample, to thereby determine whether the test sample has a desired property.

In a particularly preferred embodiment, the method of the present invention is used to identify compounds which are capable of producing a therapeutically desirable effect in, for example, humans or animals (test system) based on the data generated from a plurality of systems such as, for example, a combination of mass spectrometry, array analysis and proteomic analysis of the sample.

As discussed above, the signature may be data corresponding to components of the system. Components are preferably any measurable elements that may be:
(a) part of a sample; or

(b) part of a system and which are responsive to a sample being applied to the system.

For example, components that may be considered part of, or result from, a sample that is, for example, a chemical compound may be NMR chemical shifts values, UV-VIS absorbance values, reaction constant values, equilibrium constants values, solubility constant values. Examples of components that may be part of a system which are responsive to a sample being applied to the system include for example, expression levels for genes or proteins from a cell type that are produced in response to a sample being applied to that cell type. The expression levels may be determined by, for example, microarray analysis.

The method of the present invention may be performed on information about a characteristic without any prior knowledge of the characteristic. There is no requirement that the charactistic or components of the characteristic be known or that the mechanisms underlying the characteristic be understood. The only requirement is that each component has a unique identifier which permits each data value for that information to be assigned to a component. Thus, characteristics may be used in the method of the invention which are not well characterised or understood, and/or are extremely complex and/or generate large amounts of data. In fact, the method of the present invention works efficiently with huge amounts of information about a characteristic and/or when the information is extremely complex and in which high variability of the information is common.

As used herein, the term "signature" refers to a subset of data obtained from processing of at least two distinct

data sets. Preferably, the signature is data values for a subset of components of one or more systems. The range of data values of the signature may vary depending on the data that is analysed, the multivariate analysis that is applied, and the system to which the sample is applied. Preferably, the signature tolerates a high level of variation in which the test data may differ widely from the training data for the subset of components but may still be indicative of the desired property.

As used herein, the term "training samples" refers to samples which are known to either have or produce the desired property, or not have or not produce the desired property. Thus, the training samples are preferably one of the following two types:

(a) those that are have or acapable of producing the desired property; or

(b) those that are do not have or are not capable of producing the desired property.

Preferably, the data generated from the multivariate analysis is "predictive" of whether a sample has or can produce a desired property. Essentially, from all the data that is generated from the plurality of systems, the method of the present invention enables identification of a minimum number of components that can be used to predict whether a sample has a particular desired property. The training samples are thus used to identify the subset of components capable of predicting the desired property of a sample. Once those components have been identified using the training samples, only the subset of components need be used in future to assess test samples.

Because a subset of components is identified that are predictive of a desired property, only the subset of components need be measured during the screening of test samples. Thus, the method of the present invention permits test samples that are likely to have a desired property to be identified without the need to measure all the components of the plurality of systems in response to the test sample.

In one embodiment, the distinct data sets from the plurality of systems are combined prior to processing. In other words, the signature is generated from the combined training data from the plurality of systems. The multivariate analysis is thus applied to the combined training data from the plurality of systems in one step to generate a subset of components identified from all the information from all the systems.

In another embodiment, the data sets from the plurality of systems are processed successively. In this way, a signature may be determined from a data set from a first system that is predictive of a desired result in a second system. The desired result in the second system may be that the data generated from the second system has a particular signature that is predictive of a desired result in a third system of the plurality of systems. Thus, by obtaining a signature from a data set that relates to another data set, it is possible to identify a signature that is predictive of a desired property from successive systems.

The order of the successive systems may be any order that is appropriate for assessing the test sample(s).

Typically, the order is determined by the complexity and/or expense of each system relative to the other screening systems. For example, screening systems of less complexity or expense may be used early in the plurality when there are many samples to be tested, while more complex and/or expensive systems are preferred later in the plurality when many or most of the samples have been eliminated. For example, in drug discovery, systems such as mass spectrometry and infra-red spectrometry generate large amounts of data for each test compound relatively quickly and cheaply, and consequently many compounds may be subjected to this analysis early in the drug screening process. Microarray analysis is a more complex and expensive analysis and is therefore generally used further down the succession of systems where compounds of no interest have been screened out using the previous systems.

The succession of systems may be separated by one or more other systems of the succession of systems, or in other words, the first and second systems need not be consecutive steps in the succession. For example, a typical succession of systems used to identify a compound that is, for example, non-toxic in mammals may be: mass spectrometric analysis, then testing in a series of tissue culture cells, then testing in a series of animal models to determine, for example, toxicity and finally ADME testing. Data generated from, for example, mass spectrometric analysis (the first system) may be used to identify a subset of components that is predictive of a desired property in, for example, ADME testing. Thus, the inventors envisage that the method of the invention will allow some or many systems that are typically required for

drug dicovery to be ommitted when screening test samples. In other words, the method may permit "leapfrogging" of systems to increase speed and efficiency of the screening or testing procedures for the desired property.

It is also envisaged that a plurality of data sets may be combined before processing, and that the signature obtained from the combined data may be indicative of a signature in another data set. In this way, a signature may be obtained from data from a one or more systems of a plurality of systems which a predictive of a condition in a system that is also one of the plurality of systems.

The desired property may be any property of the sample that is sought. The desired property may be a direct property of the sample such as for example, when the sample is a compound, a physical or chemical property such as solubility, stability, a particular feature of structure, or a biological property such as toxicity, receptor binding capability, the ability to induce or inhibit gene or protein expression, pharmocological property of the compound such as target interaction. In some cases, the desired property will be readily apparent from a single result such as toxicity, solubility, whereas in other cases the desired property may be contained within a large amount of data, such as a signature that is predictive of a result in a further system.

In one embodiment, the desired property is a signature generated from a subset of components in a further one or more systems. Preferably, the signature generated from a subset of components in a further one or more systems is predictive of a further property.

The desired property may be any property or outcome that is sought in relation to the test sample when applied to a system. For example, if the sample is a compound for therapeutic use, the desired property may be, for example, low toxicity in cells, animals, or humans, effectiveness or low toxicity on a particular cell type in cell culture, ability to target a particular gene(s) or protein(s) or other cellular component, ability to induce a particular physiological response, ability to inhibit or kill a particular cell type or infectious agent.

It is also envisaged that the desired property may be data from a subset of components of a further system. Preferably, the data from the subset of components of the further system is predictive of a desired property in another system.

The choice of the plurality of systems will of course depend on the sample that is being tested and the desired property that is sought. For example, if the sample is a compound for drug discovery, the plurality of systems may include one or more chemical and/or biological systems. A biological system may be any system for assessing the chemical interaction that involves molecules of the type found in living organisms. Such interactions include anabolic and catabolic reactions that occur in living organisms including enzymatic reactions, binding reactions, signalling reactions and other reactions. Examples of biological systems include analysis of receptor-ligand interaction, analysis of enzyme-substrate interaction, analysis of cellular signalling pathways, toxicology assays, analysis of gene expression in response

to a sample, analysis of protein expression in response to a sample, analysis of the physico-chemical properties of the test sample, analysis of toxicity when the test sample is applied to the system, analysis of the mutagenic potential of a test sample, analysis of ADME. The components may be labelled or organised in a manner which allows data from one component to be distinguished from data from another component. For example, the components may be spatially organised in, for example, an array which allows data from each component to be distinguished from another by spatial position, or each component may have some unique identification associated with it such as an identification signal or tag. For example, the components may be bound to individual carriers, each carrier having a detectable identification signature such as quantum dots (see for example, Rosenthal, 2001, Nature Biotech 19: 621-622; Han et al. (2001) Nature Biotechnology 19: 631-635), flourescent markers (see for example, Fu et al, (1999) Nature Biotechnology 17: 1109-1111), bar-coded tags (see for example, Lockhart and trulson (2001) Nature Biotechnology 19: 1122-1123), or chemical labelling such as that described in US Patent No. 6,475,807. The biological system may be a biotechnology array. Examples of biotechnology arrays include oligonucleotide arrays, DNA arrays, DNA microarrays, RNA arrays, RNA microarrays, DNA microchips, RNA microchips, protein arrays, protein microchips, antibody arrays, chemical arrays, carbohydrate arrays, proteomics arrays, lipid arrays. In another embodiment, the biological system may be selected from the group including, for example, DNA or RNA electrophoresis gels, protein or proteomics electrophoresis gels, biomolecular interaction analysis such as Biacore analysis, amino acid analysis, ADME screening (see for

example High-throughput ADME estimation: In Vitro and In Silico approaches (2002), Ferenc Darvas and Gyorgy Dorman (Eds), Biotechniques Press), protein electrophoresis gels and proteomics electrophoresis gels.

Chemical systems may be any system that provides data in relation to the chemical or physical properties of a sample or a system in response to a sample. Examples of chemical systems include atomic absorption spectroscopy (AAS), auger electron spectroscopy (AES), coherent anti-Stokes spectroscopy (CARS), circular dichroism (CD), Conversion electron Mossbauer spectroscopy (CEMS), chemical ionisation mass spectroscopy, chemically-induced dynamic electron/nuclear polarisation (CIDEP/CIDNP), Cross polarisation magic angle spinning (CP-MASS), combined rotation and multipulse spectroscopy (CRAMPS), distortionless enhancement by polraisation transfer, 2-Dimensional nuclear magnetic resonance spectroscopy, electron diffraction (ED), energy dispersive X-ray spectroscopy, electron energy-loss spectroscopy, electron-electron double resonance, electronic spectroscopy, electron impact mass spectroscopy, electron-nuclear double resonance (ENDOR), electron paramagnetic resonance spectroscopy, electron spin resonance spectroscopy (ESR), exchange spectroscopy, far infrared laser magnetic resonance, flourescence spectroscopy, Fourier transform infrared spectroscopy (FTIR), gas-phase electron diffraction (GED), heteronuclear correlation spectroscopy (HETCOR), heteronuclear overhauser effect spectroscopy, Hyper Raman spectroscopy, infrared spectroscopy (IR), laser desorption mass spectroscopy, laser-induced fluorescence, laser magnetic resonance spectroscopy, magnetic circular dichroism, microwave spectroscopy, mass-

analysed ion kinetic energy spectroscopy, microwave optical double resonance spectroscopy, Mossbauer spectroscopy, multiphoton ionisation spectroscopy, multi-stage mass spectroscopy (MS/MS), multiphoton induced flourescence spectroscopy, nuclear gamma resonance spectroscopy, nuclear overhauser spectroscopy, nuclear quadrupole resonance spectroscopy, optical double resonance spectroscopy, photoelectron spectroscopy, photoionisation mass spectroscopy, Raman spectroscopy, Raman-induced Kerr-effect spectroscopy, rotating frame Nuclear Overhauser Effect spectroscopy, rotational Raman spectroscopy, Rotational spectroscopy, resonance Raman spectroscopy, secondary ion mass spectroscopy, total correlation spectroscopy, vibrational spectroscopy, visible spectroscopy, X-ray diffraction, X-ray fluorescence spectroscopy, X-ray photoelectron spectroscopy, correlation spectroscopy (COSY), Coulomb explosion, HPLC, mass spectrometry (for example, MALDI, MALDI-TOF, LC-MS, MS-MS, GC-MS, LC/MS-MS, ES-MS, LC-ES-MS).

Biological systems may also be used in applications such as quantitative trait loci analysis in animals and humans. For example, tissue samples from plants, animals or humans having known desired and undesired phenotypes may be applied to a plurality of biological systems such as, for example, microarray analysis and PCR-analysis in order to determine a signature, and this signature may be used to identify a subset of genes or loci that are predictive of the plant, animal or human having a desired property such as, for example, physical attributes such as, for example, size, strength, speed, muscle development, fat content, growth rate, breeding capacity, colour, drought

resistance, flavour etc. or attributes such as personality, intellectual capacity, or some particular talent such as sporting prowess, musical ability etc.

Any training samples may be used in the method of the invention provided a portion of the training samples have or produce the desired property. Preferably, the training samples produce data with sufficient variability to distinguish between training samples having different properties. In one embodiment, when the plurality of systems is for drug discovery, the training samples may be any chemical or biological compound, either naturally occurring or synthesised. Examples of naturally occurring compounds include nucleic acid, protein, carbohydrate, antibody, glycoprotein, modified protein, lipoprotein, lipid, phospholipid, polysaccharide, organic or inorganic molecule, biological macromolecule, or any other naturally occurring compound which is to be tested for a desired property. Also contemplated are synthesised chemical compounds such as those generated by combinatorial chemistry. Examples of methods for the generation of compounds using combinatorial chemistry include those described in Dower et al. (1991) Ann. Rep. Med. Che. 26:271-280; Fodor et al. (1991) Science 251: 767-773; Jung et al. (1992) Angew. Chem. Ind. Ed. Engl. 31:367-383; Zuckerman et al. (1992) PNAS 89:4505-4509; Scott et al. (1990) Science 249:386-390; Devlin et al. (1990) Science 249: 404-406; Cwirla et al. (1990) PNAS 87: 6378-6382; Gallop et al. (1994) J. Medicinal Chemistry 37: 1233-1251. It is also envisaged that the test sample may be a mixture of chemical compounds such as those mentioned above. In this regard, it is envisaged that the method of the invention may be used to identify compounds which act

synergistically to exhibit a desired property, or identify complex mixtures of compounds such as those found in naturally occurring substances e.g. plant extracts, saps, body fluid extracts which exhibit a desired property, or compounds produced through industrial processes which exhibit a desired property.

The test sample may be any sample for which a desired property is sought. It will be appreciated by those skilled in the art that the test samples may be the same type of compounds as the training samples, or made using the same methods.

In embodiments where the test sample is a compound for which a therapeutic or biological property is sought, it will be appreciated by persons skilled in the art that the classes of chemical that may be used in the method of the invention need not be limited to any particular chemical, and that the method of the invention encompasses the use of any compound as a test sample. For example, the chemical compound may be a nucleic acid, protein, carbohydrate, pharmaceutical, antibody, glycoprotein, modified protein, lipoprotein, lipid, phospholipid, polysaccharide, organic or inorganic molecule, biological macromolecule, or any other naturally occurring or synthesized chemical compound which is to be tested for a desired property. Synthesised chemical compounds such as those generated by combinatorial chemistry are particularly preferred. It is also envisaged that the test sample may be a mixture of chemical compounds such as any of those mentioned above. In this regard, it is envisaged that the method of the invention may be used to identify compounds which act synergistically to exhibit a

desired property, or identify complex mixtures of compounds such as those found in naturally occurring substances e.g. plant extracts, saps, body fluid extracts which exhibit a desired property, or compounds produced through industrial processes which exhibit a desired property.

The component may be one or more measurable components of the characteristic. For example, in a biological system the components may be genes, proteins, antibodies, carbohydrates, lipids, phospholipids or any other components of a biological system, and the data generated from the systems is typically a measure of the levels of these components in response to a sample applied to the systems. The components may be known components, or in other words, previously identified components. However, there is no requirement that the components be previously known and consequently the components may be unknown components, or in other words, components that have not been previously identified. It is envisaged that the method of the invention may also be used with a combination of known and unknown components.

The signature may be identified using any multivariate analysis. In one embodiment, the multivariate analysis comprises generating a combination of weighted components to identify a subset of components that provide the signature. A flow diagram of the steps involved in a preferred method of the multivariate analysis is shown in Fig. 3.

Data generated by applying the training samples to the plurality of systems is processed using a multivariate

analysis to identify a signature that is capable of correctly predicting whether a training sample has a desired property. The multivariate method used to determine the signature may be any multivariate method capable of handling large amounts of data. Preferably, the multivariate method is a Bayesian method. Preferably, the bayesian method combines a prior, and a likelihood function which models the distribution of the data generated by the training samples. Thus, the type of bayesian method selected in the method of the invention will typically depend on the distribution of the desired property. Most samples having a desired property typically exhibit a probability distribution, and the probability distribution of a desired property can be modelled using statistical models which are based on the data generated from the training samples in the plurality of systems. The method utilises statistical models which model the probability distribution for a desired property. Thus, for a sample having a desired property with a particular probability distribution, an appropriate model is defined that models that distribution. The method preferably uses any model that is conditional on the linear combination, and is preferably a mathematical equation in the form of a likelihood function that provides a probability distribution based on the data obtained from the training samples. Preferably, the likelihood function is based on a previously described model for describing some probability distribution. In one embodiment, the model is a likelihood function based on a model selected from the group consisting of a multinomial or binomial logistic regression, generalised linear model, Cox's proportional hazards model, accelerated failure model and parametric survival model. For example, if the

desired property exhibits a binomial or multinomial distribution, then the likelihood function of the Bayesian method correspond to that of a logistic regression. If the desired property exhibits an ordered multiclass distribution, then the likelihood function of the Bayesian method will correspond to that of an ordered categories model. If the desired property exhibits a distribution from the class of generalised linear models , then the likelihood function of the Bayesian method will correspond to that of a generalised linear model. If the desired property is survival data for an animal or human, or measures some type of failure of a system or sample, then the likelihood function of the Bayesian method will correspond to that of a proportional hazards model.

The multivariate method of the preferred embodiment is particularly suited to the analysis of very large amounts of data. Typically, large data sets are often highly variable and test sample data often differ significantly that obtained from the training samples. By using the multivariate methods of the preferred embodiment, it is possible to identify signatures using training samples which can be used to determine whether a test sample has or will have a desired property, even when the data generated from the test sample is highly variable compared to the data generated from training samples having the desired property. Thus, the multivariate method of the preferred embodiment is able to identify a signature that is indicative that a test sample has a desired property even when the data from the test sample is of poor quality and/or there is a high variability between the training samples having the desired property. In fact, the method may preferably select components that exhibit high

variability between training samples having the desired property (and yet are predictive of the property) as these components are more likely to be correctly predictive for highly variable test samples. In other words, the method of the present invention is capable of identifying components that have a high tolerance to variation. In this way, the subsets of components are able to predict response characteristics in test data that varies from the training data.

The multivariate method of the preferred embodiment also has the advantage that it requires usage of less computer memory than prior art methods. Accordingly, the method of the present invention can be performed rapidly on computers such as, for example, laptop machines. By using less memory, the method of the present invention also allows the method to be performed more quickly than prior art methods for analysis of, for example, biological data.

Preferred embodiments of the multivariate methods are now described in greater detail.

A. Multi Class Logistic regression model

The multivariate method of this embodiment utilises the training samples in order to identify a signature from at least two distinct data sets which can classify the samples into a pre-defined groups , wherein at least one of the pre-defined groups exhibits the desired property. For example, a set of data from mass spectrometry analysis and microarray analysis generated from a particular training sample may be used to predict that that training sample belongs to a pre-defined group.

In this way, the multivariate method identifies preferably a minimum number of components from the data set that can be used to determine that a training sample belongs to a pre-defined group. Thus, when one of the pre-defined groups is a group which exhibits the desired property, the subset of components is therefore indicative that the sample is capable of producing the desired property by classifying the training sample into that group.

Once the subset of components have been identified using the training samples, data for the subset of the components may be obtained using test samples to determine whether that test sample is capable of producing a result in the appropriate system that is in the group having the desired property.

The pre-defined group may be any classification which groups the samples. For example, the samples may be grouped on the basis of particular array results, wherein those samples that produce a pattern of gene expression in the second system belong to a particular group, whether those samples are low toxicity, high efficacy etc. For example, a pattern of gene expression may be for example, a subset of genes that is predictive of the sample producing a desired result in another system.

In one embodiment, the input data is organised into an $n \times p$ data matrix $X = (x_{ij})$ with n training samples and p components. Typically, p will be much greater than n.

In another embodiment, data matrix X may be replaced by an n x n kernel matrix K to obtain smooth functions of X as

predictors instead of linear predictors. An example of the kernel matrix K is $k_{ij}=\exp(-0.5*(x_i-x_j)^t(x_i-x_j)/\sigma^2)$ where the subscript on x refers to a row number in the matrix X. Ideally, subsets of the columns of K are selected which give sparse representations of these smooth functions. Further examples of kernel matrices are given in table 2 below.

Associated with each sample class (group) may be a class label y_i, where $y_i = k, k \in \{1,...,G\}$, which indicates which of G sample classes a training sample belongs to. We write the nx1 vector with elements y_i as y. Given the vector $\underset{\sim}{y}$ we can define indicator variables

$$e_{ig} = \begin{cases} 1, & y_i = g \\ 0, & \text{otherwise} \end{cases} \qquad (1A)$$

In one embodiment, the processed component weights are estimated using a Bayesian statistical model (see Kotz and Johnson, 1983). Preferably, the processed weights are estimated by maximising the posterior distribution of the weights given the data generated from each training sample. This results in an objective function to be maximised consisting of two parts. The first part a likelihood function and the second a prior distribution for the weights which ensures that zero weights are preferred whenever possible. In a preferred embodiment, the likelihood function is derived from a multiclass logistic model. Preferably, the likelihood function is computed from the probabilities:

$$p_{ig} = \frac{e^{x_i^T \beta_g}}{\left(1 + \sum_{h=1}^{G-1} e^{x_i^T \beta_h}\right)}, g = 1,...,G-1 \qquad (2A)$$

and

$$p_{iG} = \frac{1}{\left(1 + \sum_{h=1}^{G-1} e^{x_i^T \beta_h}\right)} \qquad (3A)$$

Wherein

p_{ig} is the probability that the training sample with input data X_i will be in sample class g;

$x_i^T \beta_g$ is a linear combination generated from input data from training sample i with component weights β_g;

x_i^T is the components for the i^{th} Row of X and β_g is a set of component weights for sample class g;

Typically, as discussed above, the processed component weights are estimated in a manner which takes into account the a priori assumption that most of the component weights are zero.

In one embodiment, components weights βg in equation (2A) are estimated in a manner whereby most of the values are zero, yet the samples can still be accurately classified.

In one embodiment, the prior specified for the parameters $\beta_1,...,\beta_{G-1}$ is of the form:

$$P(\beta_1,...,\beta_{G-1}) = \int_{\tau^2} \prod_{g=1}^{G-1} P\left(\beta_g \middle| \tau_g^2\right) P\left(\tau_g^2\right) d\tau^2 \qquad (4A)$$

where $\beta^T = \left(\beta_1^T,...\beta_{G-1}^T\right)$ and $\tau^T = \left(\tau_1^T,...,\tau_{G-1}^T\right)$.

and $p\left(\beta_g \middle| \tau_g^2\right)$ is $N\left(0, \text{diag}\left\{\tau_g^2\right\}\right)$ and $p\left(\tau_g^2\right) \alpha \prod_{i=1}^{n} 1/\tau_{ig}^2$ is a Jeffreys hyperprior, Kotz and Johnson(1983).

In one embodiment, the likelihood function is $L\left(\underset{\sim}{y} \middle| \beta_1,...,\beta_{G-1}\right)$ of the form in equation (8A) and the posterior distribution of β and $\underset{\sim}{\tau}$ given y is

$$p(\beta\tau|y) \alpha\, L(y|\beta) p(\beta|\tau) p(\tau) \qquad (5A)$$

In one embodiment, the first derivative is determined from the following equation:

$$\frac{\partial \log L}{\partial \beta g} = X^T\left(\underset{\sim}{e}_g - p_g\right), \quad g = 1,...,G-1 \qquad (6A)$$

wherein $\underset{\sim}{e}_g^T = \left(e_{ig}, i = 1,n\right)$, $p_g^T = \left(p_{ig}, i = 1,n\right)$ are vectors indicating membership of sample class g and probability of class g respectively.

In one embodiment, the second derivative is determined from the following algorithm:

$$\frac{\partial^2 \log L}{\partial \beta_g \partial \beta_h} = -X^T diag\left\{\delta_{hg} p_g - p_h p_g\right\} X \qquad (7A)$$

Equation 6A and equation 7A may be derived as follows:

(a) Using equations (1A), (2A) and (3A), the likelihood function of the data can be written as:

$$L=\prod_{i=1}^{n}\left(\prod_{g=1}^{G-1}\left\{\frac{e^{x_i^T\beta_g}}{\left(1+\sum_{g=1}^{G-1}e^{x_i^T\beta_g}\right)}\right\}^{e_{ig}}\left\{\frac{1}{1+\sum_{h=1}^{G-1}e^{x_i^T\beta_h}}\right\}^{e_{iG}}\right) \qquad (8A)$$

(b) Taking logs of equation (8A) and using the fact that $\sum_{h=1}^{G}e_{ih}=1$ for all i gives:

$$l\log L=\sum_{i=1}^{n}\left(\sum_{g=1}^{G-1}e_{ig}x_i^T\beta_g-\log\left(1+\sum_{g=1}^{G-1}e^{x_i^T\beta_g}\right)\right) \qquad (9A)$$

(c) Differentiating equation (9A) with respect to βg gives

$$\frac{\partial\log L}{\partial\beta g}=X^T\left(\underset{\sim}{e}_g-p_g\right),\quad g=1,\ldots,G-1 \qquad (10A)$$

whereby $\underset{\sim}{e}_g^T=\left(e_{ig},i=1,n\right)$, $p_g^T=\left(p_{ig},i=1,n\right)$ are vectors indicating membership of sample class g and probability of class g respectively.

(d) The second derivative of equation (9A) has elements

$$\frac{\partial^2\log L}{\partial\beta_g\partial\beta_h}=-X^T diag\left\{\delta_{hg}p_g-p_hp_g\right\}X \qquad (11A)$$

where

$$\delta_{hg} = \begin{cases} 1, & h = g \\ 0, & \text{otherwise} \end{cases}$$

Processed component weights which maximise the posterior distribution of the likelihood function may be specified using an EM algorithm comprising an E step and an M step.

Typically, the EM algorithm comprises the steps:

(a) performing an E step by calculating the conditional expected value of the posterior distribution of component weights using the function:

$$Q = \log L - \frac{1}{2}\sum_{g=1}^{G-1} \gamma_g^T \, diag\left\{\hat{\gamma}_g\right\}^{-2} \gamma_g \qquad (12A)$$

where $x_i^T \beta_g = x_i^T P_g \hat{\gamma}_g$ in equation (8A)

(b) performing an M step by applying an iterative procedure to maximise Q as a function of γ whereby:

$$\gamma^{t+1} = \gamma^t - \alpha^t \left(\frac{\partial^2 Q}{\partial \gamma^2}\right)^{-1} \left(\frac{\partial Q}{\partial \gamma}\right) \qquad (13A)$$

where α^t is a step length such that $0 \leq \alpha^t \leq 1$;

$\beta_g = P_g \gamma_g$;

wherein P_g are matrices of zeroes and ones such that $P^T{}_g\beta_g$ selects non-zero elements of β_g; and

$$\gamma = (\gamma_g,\ g=1,\ldots..,G-1).$$

Equation (12A) may be derived as follows:

Calculate the conditional expected value of 5A) given the observed data y and a set of parameter estimates $\hat{\beta}$.

$$Q = Q\left(\beta \middle| \underset{\sim}{y}, \hat{\beta}\right) = \mathrm{E}\left\{\log p\left(\underset{\sim}{\beta}, \underline{\tau} | y\right) \middle| y, \hat{\beta}\right\}$$

Consider the case when components of β (and $\hat{\beta}$) are set to zero i.e for $g=1,\ldots,G-1$, $\beta_g = P_g\gamma_g$ and $\hat{\beta}_g = P_g\hat{\gamma}_g$, where the P_g are matrices of zeroes and ones such that $P_g^T\beta_g$ selects the non zero elements of β_g. In the following we write γ = (γ_g, g=1,…,G-1). Note that the γ_g are actually subsets of the components of β_g. We use them to keep the notation as simple as possible.

Ignoring terms not involving γ and using (4A), (5A), (9A) we get:

$$Q = \log L - \frac{1}{2}\sum_{g=1}^{G-1}\sum_{i=1}^{n}\mathrm{E}\left\{\frac{\gamma_{ig}^2}{\tau_{ig}^2} \middle| y, \hat{\underset{\sim}{\gamma}}\right\}$$

$$= \log L - \frac{1}{2}\sum_{g=1}^{G-1}\gamma_g^T\, diag\left\{\hat{\gamma}_g\right\}^{-2}\gamma_g \quad (14A)$$

where $x_i^T\beta_g = x_i^T P_g\hat{\gamma}_g$ in (8A)

Note that the conditional expectation can be evaluated from first principles given (4A).

The iterative procedure may be derived as follows:

To obtain the derivatives required in (13A), first note that from (8A), (9A) and (10A) we get

$$\frac{\partial Q}{\partial \gamma} = \left(\frac{\partial \beta}{\partial \gamma}\right)\frac{\partial \log L}{\partial \beta} - diag\{\hat{\gamma}\}^{-2}\gamma$$

$$= \begin{bmatrix} X_1^T(e_1 - p_1) \\ \vdots \\ X_{G-1}^T(e_{G-1} - p_{G-1}) \end{bmatrix} - diag\left\{\hat{\underset{\sim}{\gamma}}\right\}^{-2}\gamma \qquad (15A)$$

and

$$\frac{\partial^2 Q}{\partial \underset{\sim}{\gamma}^2} = \left(\frac{\partial \beta}{\partial \gamma}\right)\frac{\partial^2 \log L}{\partial^2 \beta}\left(\frac{\partial \beta}{\partial \gamma}\right)^T - diag\{\hat{\gamma}\}^{-2}$$

$$= -\left\{\begin{pmatrix} X_1^T\Delta_{1,1}X_1 & \cdots & X_1^T\Delta_{1,G-1}X_{G-1} \\ \vdots & & \vdots \\ X_{G-1}\Delta_{G-1,1}X_1 & & X_{G-1}\Delta_{G-1,G-1}X_{G-1} \end{pmatrix} + diag\{\hat{\gamma}\}^{-2}\right\} \qquad (16A)$$

where

$$\Delta_{gh} = diag\{\delta_{gh}p_g - p_g p_h\},$$

$$\delta_{gh} = \begin{cases} 1, & g = h \\ 0, & \text{otherwise} \end{cases}$$

and

$$X_g^T = P_g^T X^T, g = 1,...G-1. \quad (17A)$$

In a preferred embodiment, the iterative procedure may be simplified by using only the block diagonals of equation (16A) in equation (13A). For $g = 1,...G-1$, this gives:

$$\gamma_g^{t+1} = \gamma_g^t + \alpha^t \left\{ X_g^T \Delta_{gg} X_g + diag\left\{\hat{\gamma}_g\right\}^{-2} \right\}^{-1} \left\{ X_g^T \left(e_g - p_g\right) - diag\left\{\hat{\gamma}_g\right\}^{-2} \gamma_g^t \right\} \quad (18A)$$

Rearranging equation (18A) leads to

$$\gamma_g^{t+1} = \gamma_g^t + \alpha^t diag\left\{\hat{\gamma}_g\right\} \left(Y_g^T \Delta_{gg} Y_g + I\right)^{-1} \left\{ Y_g^T \left(e_g - p_g\right) - diag\left\{\hat{\gamma}_g\right\}^{-1} \gamma_g^t \right\} \quad (19A)$$

where

$$Y_g^T = diag\left\{\hat{\gamma}_g\right\} X_g^T$$

Writing $p(g)$ for the number of columns of Y_g, (19A) requires the inversion of a $p(g)\,\mathrm{x}\,p(g)$ matrix which may be quite large. This can be reduced to an $n\mathrm{x}n$ matrix for $p(g) > n$ by noting that:

$$\begin{aligned} \left(Y_g^T \Delta_{gg} Y_g + I\right)^{-1} &= I - Y_g^t \left(Y_g Y_g^T + \Delta_{gg}^{-1}\right)^{-1} Y_g \\ &= I - Z_g^T \left(Z_g Z_g^T + I_n\right)^{-1} Z_g \end{aligned} \quad (20A)$$

where $Z_g = \Delta_{gg}^{\frac{1}{2}} Y_g$. Preferably, (19A) is used when $p(g) < n$ and (19A) with (20A) substituted into equation (19A) is used when $p(g) \geq n$.

In a preferred embodiment, the EM algorithm is performed as follows:

1. Set n=0 , $P_g = I$ and choose an initial value for $\hat{\gamma}^0$. This is done by ridge regression of $\log(p_{ig}/p_{iG})$ on x_i where p_{ig} is chosen to be near one for observations in group g and a small quantity >0 otherwise - subject to the constraint of all probabilities summing to one.

2. Do the E step i.e evaluate $Q = Q\left(\gamma \middle| \underset{\sim}{y}, \hat{\gamma}^n\right)$

3. Set t=0. For $g = 1,...,G-1$ calculate:

a) $\delta_g^t = \gamma_g^{t+1} - \gamma_g^t$ using (19A) with (20A) substituted into (19A) when $p(g) \geq n$.

(b) Writing $\delta^t = \left(\delta_g^t, g = 1,...,G-1\right)$ Do a line search to find the value of α^t in $\underset{\sim}{\gamma}^{t+1} = \underset{\sim}{\gamma}^t + \alpha^t \underset{\sim}{\delta}^t$ which maximises (or simply increases) (12A) as a function of α^t.

c) set $\underset{\sim}{\gamma}^{t+1} = \underset{\sim}{\gamma}^t$ and t=t+1

Repeat steps (a) and (b) until convergence.

This produces γ^{*n+1} say which maximises the current Q function as a function of γ.

For $g = 1,...G-1$ determine $S_g = \left\{ j : \left| \gamma^{*n+1}_{jg} \right| \leq \varepsilon \max_k \left| \gamma^{*n+1}_{kg} \right| \right\}$

Where $\varepsilon \ll 1$, say 10^{-5}. Define P_g so that $\beta_{ig} = 0$ for $i \in S_g$ and

$$\hat{\gamma}_g^{n+1} = \left\{ \gamma^{*n+1}_{jg}, j \notin S_g \right\}$$

This step eliminates variables with small coefficients from the model.

4. Set n=n+1 and go to 2 until convergence.

A second embodiment in which a multivariate analysis relating to a categorical ordered logistic regression is used will now be described.

B. Ordered categories model

The multivariate method of this embodiment utilises the training samples in order to identify a signature from a subset of components which can classify the samples into a pre-defined ordered group, wherein at least one of the pre-defined ordered groups exhibits the desired property. The pre-defined ordered group may be, for example, increasing toxicity in cells, or an LD50 experiment.

In the following there are N samples, and vectors such as y, z and μ have components y_i, z_i and μ_i for i = 1,..., N. Vector multiplication and division is defined component-wise and diag{ · } denotes a diagonal matrix whose diagonals are equal to the argument. We also use || · || to denote Euclidean norm.

Preferably, there are N observations y_i where y_i takes integer values 1,...,G. The values denote classes which are ordered in some way such as for example severity of disease. Associated with each observation there is a set of covariates (variables, e.g gene expression values) arranged into a matrix X with N rows and p columns wherein N is the samples and p the components. The notation x_i^T

denotes the i^{th} row of X. Individual (sample) i has probabilities of belonging to class k given by $\pi_{ik} = \pi_k(x_i)$. Define cumulative probabilities

$$\gamma_{ik} = \sum_{g=1}^{k} \pi_{ik} \quad , \qquad k = 1,\ldots,G$$

Note that γ_{ik} is just the probability that observation i belongs to a class with index less than or equal to k. Let C be a n by p matrix with elements c_{ij} given by

$$c_{ij} = \begin{cases} 1, & \text{if observation i in class j} \\ 0, & \text{otherwise} \end{cases}$$

and let R be an n by P matrix with elements r_{ij} given by

$$r_{ij} = \sum_{g=1}^{j} c_{ig}$$

These are the cumulative sums of the columns of C within rows.

For independent observations (samples) the likelihood of the data can be written as

$$L = \prod_{i=1}^{N} \prod_{k=1}^{G-1} \left(\frac{\gamma_{ik}}{\gamma_{ik+1}} \right)^{r_{ik}} \left(\frac{\gamma_{ik+1} - \gamma_{ik}}{\gamma_{ik+1}} \right)^{r_{ik+1} - r_{ik}} \qquad (1B)$$

and the log likelihood ($\log(L)$) l can be written as

$$l = \sum_{i=1}^{N} \sum_{k=1}^{G-1} r_{ik} \log\left(\frac{\gamma_{ik}}{\gamma_{ik+1}} \right) + (r_{ik+1} - r_{ik}) \log\left(\frac{\gamma_{ik+1} - \gamma_{ik}}{\gamma_{ik+1}} \right) \qquad (2B)$$

The continuation ratio model may be adopted here as follows:

$$\operatorname{logit}\left(\frac{\gamma_{ik+1} - \gamma_{ik}}{\gamma_{ik+1}} \right) = \operatorname{logit}\left(\frac{\pi_{ik}}{\gamma_{ik+1}} \right) = \theta_k + x_i^T \beta^* \qquad (3B)$$

for k = 2,...,G , see McCullagh and Nelder(1989) and McCullagh(1980) and the discussion therein. Note that

$$\text{logit}\left(\frac{\gamma_{ik+1}-\gamma_{ik}}{\gamma_{ik+1}}\right)=-\text{logit}\left(\frac{\gamma_{ik}}{\gamma_{ik+1}}\right). \qquad (4B)$$

The likelihood is equivalent to a logistic regression likelihood with response vector y and covariate matrix X

$$y=vec\{R\}$$
$$X=\left[B_1^T B_2^T \ldots B_N^T\right]^T$$
$$B_i=\left[I_{G-1}|1_{G-1}x_i^T\right]$$

where I_{G-1} is the G-1 by G-1 identity matrix and 1_{G-1} is a G-1 by 1 vector of ones.

Here vec{ } takes the matrix and forms a vector row by row.

Typically, as discussed above, the component weights are estimated in a manner which takes into account the a priori assumption that most of the component weights are zero.

Following Figueiredo(2001), in order to eliminate redundant variables (covariates), a prior is specified for the parameters β^* by introducing a p x 1 vector of hyperparameters .

Preferably, the prior specified for the component weights is of the form

$$p(\beta^*)=\int_{v^2} p(\beta^*|v^2)p(v^2)dv^2 \qquad (5B)$$

where $p(\beta^*|v^2)$ is $N(0,\text{diag}\{v^2\})$ and $p(v^2)\alpha\prod_{i=1}^{n}1/v_i^2$ is a Jeffreys prior, Kotz and Johnson(1983). The elements of $\theta=(\theta_2,\ldots\theta_G)^T$ have a non informative prior.

Writing $L\left(\underset{\sim}{y}\middle|\beta^*\theta\right)$ for the likelihood function, in a Bayesian framework the posterior distribution of β^*, θ and ν given y is

$$p\left(\beta^*\theta\nu|y\right)\alpha\, L\left(y\middle|\beta^*\theta\right)p\left(\beta^*|\nu\right)p(\nu) \quad (6B)$$

Preferably, by treating ν as a vector of missing data, an iterative algorithm such as an EM algorithm (Dempster et al, 1977) can be used to maximise (6B) to produce locally maximum a posteriori estimates of β^* and θ. The prior above is such that the maximum a posteriori estimates will tend to be sparse i.e. if a large number of parameters are redundant, many components of β^* will be zero.

Preferably $\beta^T=(\theta^T,\beta^{*T})$ in the following and diag() denotes a diagonal matrix:

For the ordered categories model above it can be shown that

$$\frac{\partial l}{\partial\beta}=X^t(y-\mu) \quad (7B)$$

$$E\{\frac{\partial^2 l}{\partial\beta^2}\}=-X^t\mathrm{diag}\{\mu(1-\mu)\}X \quad (8B)$$

where $\mu_i=\exp(x_i^T\beta)/(1+\exp(x_i^T\beta))$ and $\beta^T=(\theta_2,...,\theta_G,\beta^{*T})$ and x_i denotes the i th row of X.

As mentioned above, the processed component weights which maximise the posterior distribution may be determined using an iterative procedure. Preferable, the iterative procedure for maximising the posterior distribution of the components and component weights is an EM algorithm, such as, for example, that described in Dempster et al, 1977. Preferably, the EM algorithm is performed as follows:

1. Set n=0, $S_0 = \{1,2,\ldots, p\}$, $\phi^{(0)}$, and $\varepsilon = 10^{-5}$ (say). Set the regularisation parameter κ at a value much greater than 1, say 100. This corresponds to adding $1/\kappa^2$ to the first G-1 diagonal elements of the second derivative matrix in the M step below.

If $p \leq N$ compute initial values β^* by

$$\beta^*=(X^tX+\lambda I)^{-1}X^Tg(y+\zeta) \qquad (9B)$$

and if $p > N$ compute initial values β^* by

$$\beta^*=\frac{1}{\lambda}(I-X^T(XX^T+\lambda I)^{-1}X)X^Tg(y+\zeta) \qquad (10B)$$

where the ridge parameter λ satisfies $0 < \lambda \leq 1$ and ζ is small and chosen so that the logit link function g is well defined at $y+\zeta$.

2. Define

$$\beta_i^{(n)} = \begin{cases} \beta_i^*, & i \in S_n \\ 0, & \text{otherwise} \end{cases}$$

and let P_n be a matrix of zeroes and ones such that the nonzero elements $\gamma^{(n)}$ of $\beta^{(n)}$ satisfy

$$\gamma^{(n)} = P_n^T \beta^{(n)} \quad , \quad \beta^{(n)} = P_n \gamma^{(n)}$$

$$\gamma = P_n^T \beta \quad , \quad \beta = P_n \gamma$$

Define $w_\beta = (w_{\beta i}, i = 1, p)$, such that

$$w_{\beta i} = \begin{cases} 1, & i \geq G \\ 0, & \text{otherwise} \end{cases}$$

and let $w_\gamma = P_n w_\beta$

3. Perform the E step by calculating

$$Q(\beta \mid \beta^{(n)}) = E\{ \log(p(\beta, \nu \mid y)) \mid y, \beta^{(n)}\} = l(y \mid \beta) - 0.5 (\| (\beta w_\beta)/\beta^{(n)} \|^2) \qquad (11B)$$

where l is the log likelihood function of y.

Using $\beta = P_n \gamma$ and $\beta^{(n)} = P_n \gamma^{(n)}$ (11B) can be written as

$$Q(\gamma \mid \gamma^{(n)}) = l(y \mid P_n \gamma) - 0.5 (\| (\gamma w_\gamma)/\gamma^{(n)} \|^2) \qquad (12B)$$

4. Do the M step. This can be done with Newton Raphson iterations as follows. Set $\gamma_0 = \gamma^{(n)}$ and for r=0,1,2,... $\gamma_{r+1} = \gamma_r + \alpha_r \delta_r$ where α_r is chosen by a line search algorithm to ensure $Q(\gamma_{r+1} \mid \gamma^{(n)}) > Q(\gamma_r \mid \gamma^{(n)})$.

For p ≤ N use

$$\delta_r = \text{diag}(\hat{\gamma}^{(n)})[Y_n^T V_r^{-1} Y_n + I]^{-1} (Y_n^T z_r - \frac{w_\gamma \gamma_r}{\gamma^{(n)}}) \qquad (13B)$$

where

$$\hat{\gamma}_i^{(n)} = \begin{cases} \gamma_i^{(n)}, & i \geq G \\ \kappa, & \text{otherwise} \end{cases}$$

$$Y_n^T = \text{diag}(\hat{\gamma}^{(n)}) P_n^T X^T$$

$$V_r^{-1} = \text{diag}\{\mu_r (1-\mu_r)\}$$

$$z_r = (y - \mu_r)$$

and $\mu_r = \exp(XP_n\gamma_r)/(1+\exp(XP_n\gamma_r))$.

For p > N use

$$\delta_r = \mathrm{diag}(\hat{\gamma}^{(n)})[I - Y_n^T(Y_nY_n^T+V_r)^{-1}Y_n](Y_n^Tz_r - \frac{w_\gamma\gamma_r}{\gamma^{(n)}}) \qquad (14B)$$

with V_r and z_r defined as before.

Let γ^* be the value of γ_r when some convergence criterion is satisfied e.g

$||\gamma_r - \gamma_{r+1}|| < \varepsilon$ (for example 10^{-5}).

5. Define $\beta^* = P_n\gamma^*$, $S_{n+1}=\{i \geq G: |\beta_i| > \max_{j\geq G}(|\beta_j|*\epsilon_1)\} \cup \{1,2,...,G-1\}$

where ε_1 is a small constant, say 1e-5. Set n=n+1 .

6. Check convergence: If $||\gamma^* - \gamma^{(n)}|| < \varepsilon_2$ where ε_2 is suitably small then stop, else go to step 2 above.

Recovering the probabilities

Once estimates of the parameters β are obtained, calculate

$$a_{ik} = \frac{\hat{\pi}_{ik}}{\hat{\gamma}_{ik}}$$

for i =1,…,N and k = 2,…,G.

Preferably, to obtain the probabilities we use the recursion

$$\pi_{iG} = a_{iG}$$

$$\pi_{ik-1} = \left(\frac{a_{ik-1}}{a_{ik}}\right)(1-a_{ik})\pi_{ik}$$

and the fact that the probabilities sum to one, for i = 1,…,N.

In one embodiment , the covariate matrix X with rows x_i^T can be replaced by a matrix K with ij^{th} element k_{ij} and $k_{ij} = \kappa(x_i - x_j)$ for some kernel function κ . This matrix can also be augmented with a vector of ones. Some example kernels are given in Table 2 below, see Evgeniou et al(1999).

Kernel function	Formula for $\kappa(x - y)$
Gaussian radial basis function	$\exp(- \|\| x - y \|\|^2 / a)$, $a>0$
Inverse multiquadric	$(\|\| x - y \|\|^2 + c^2)^{-1/2}$
Multiquadric	$(\|\| x - y \|\|^2 + c^2)^{1/2}$
Thin plate splines	$\|\| x - y \|\|^{2n+1}$ $\|\| x - y \|\|^{2n} \ln(\|\| x - y \|\|)$
Multi layer perceptron	$\tanh(x \cdot y - \theta)$, for suitable θ
Ploynomial of degree d	$(1 + x \cdot y)^d$
B splines	$B_{2n+1}(x - y)$
Trigonometric polynomials	$\sin((d +1/2)(x-y))/\sin((x-y)/2)$

Table 2: Examples of kernel functions

In Table 1 the last two kernels are preferably one dimensional i.e. for the case when X has only one column. Multivariate versions can be derived from products of these kernel functions. The definition of B_{2n+1} can be found in De Boor(1978). Use of a kernel function results in estimated probabilities which are smooth (as opposed to transforms of linear) functions of the covariates X. Such models may give a substantially better fit to the data.

A third embodiment in which a multivariate analysis relating to a generalised linear model is used will now be described.

C. Generalised Linear Models

The multivariate method of this embodiment utilises the training samples in order to identify a signature for a subset of components which can predict whether a sample has a pre-fined characteristic, wherein at least one of the pre-defined characteristics is the desired property.

The characteristic may be any result that exhibits a distribution from a regular exponential family of distributions. Examples of characteristics that may be identified using this embodiment include quantitities or measures, censored survival time of test animals treated with a sample, or any continuously variable characteristic.

In one embodiment, the data may be a quantity y_i, where $i \in \{1,...,N\}$. We write the Nx1 vector with elements y_i as y. We define a p x 1 parameter vector β of component weights (many of which are expected to be zero), and a q x 1 vector of parameters φ (not expected to be zero). Note that q could be zero (i.e. the set of parameters not expected to be zero may be empty).

In one embodiment, the input data is organised into an Nxp data matrix $X = (x_{ij})$ with N test training samples and p components. Typically, p will be much greater than N.

In another embodiment, data matrix X may be replaced by an N x N kernel matrix K to obtain smooth functions of X as predictors instead of linear predictors. An example of

the kernel matrix K is $k_{ij}=\exp(-0.5*(x_i-x_j)^t(x_i-x_j)/\sigma^2)$ where the subscript on x refers to a row number in the matrix X. Ideally, subsets of the columns of K are selected which give sparse representations of these smooth functions.

Typically, as discussed above, the processed component weights are estimated in a manner which takes into account the a priori assumption that most of the component weights are zero.

In one embodiment, the prior specified for the component weights is of the form:

$$p(\beta)=\int_{\nu^2} p(\beta\,|\nu^2)\,p(\nu^2)\,d\nu^2 \qquad (1C)$$

where $p(\beta\,|\nu^2)$ is $N(0,\mathrm{diag}\{\nu^2\})$ and $p(\nu^2)\alpha\prod_{i=1}^{n}1/\nu_i^2$ is a Jeffreys prior, Kotz and Johnson(1983). Preferably, an uninformative prior for φ is specified.

The likelihood function defines a model which fits the data based on the distribution of the data. Preferably, the likelihood function is derived from a generalised linear model. For example, the likelihood function $L(\underset{\sim}{y}|\beta\varphi)$ may be the form appropriate for a generalised linear model (GLM), such as for example, that described by Nelder and Wedderburn (1972). Preferably, the likelihood function is of the form:

$$l=\log p(y\,|\,\beta,\phi)=\sum_{i=1}^{N}\{\frac{y_i\theta_i-b(\theta_i)}{a_i(\phi)}+c(y_i,\phi)\} \qquad (2C)$$

where $y = (y_1, \ldots, y_n)^T$ and $a_i(\phi) = \phi / w_i$ with the w_i being a fixed set of known weights and ϕ a single scale parameter.

Preferably, the likelihood function is specified as follows:

We have

$$E\{y_i\} = b'(\theta_i)$$
$$\mathrm{Var}\{y\} = b''(\theta_i)a_i(\phi) = \tau_i^2 a_i(\phi) \qquad (3C)$$

Each observation has a set of covariates x_i and a linear predictor $\eta_i = x_i^T \beta$. The relationship between the mean of the i^{th} observation and its linear predictor is given by the link function $\eta_i = g(\mu_i) = g(b'(\theta_i))$. The inverse of the link is denoted by h, i.e

$\mu_i = b'(\theta_i) = h(\eta_i)$.

In addition to the scale parameter, a generalised linear model may be specified by four components:

- the likelihood or (scaled) deviance function,
- the link function
- the derivative of the link function
- the variance function.

Some common examples of generalised linear models are given in the table below.

Distribution	Link function $g(\mu)$	Derivative of link function	Variance function	Scale parameter
Gaussian	μ	1	1	yes

Binomial with n trials	$\log\left(\frac{\mu}{1-\mu}\right)$	$\frac{1}{\mu(1-\mu)}$	$\frac{\mu}{n}(1-\mu)$	no
Poisson	log (μ)	1/ μ	μ	no
Gamma	1/ μ	-1/ μ^2	μ^2	yes
Inverse Gaussian	1/ μ^2	-2/ μ^3	μ^3	yes

In another embodiment, the likelihood function is derived from a multiclass logistical model.

In another embodiment, a quasi likelihood model is specified wherein only the link function and variance function are defined. In some instances, such specification results in the models in the table above. In other instances, no distribution is specified.

In one embodiment, the posterior distribution of β φ and ν^2 given y is estimated using:

$$p(\beta\varphi\nu^2|y) \alpha L(y|\beta\varphi)p(\beta|\nu^2)p(\nu^2) \qquad (4C)$$

wherein $L(\underset{\sim}{y}|\beta\varphi)$ is the likelihood function.

In one embodiment, ν^2 may be treated as a vector of missing data and an iterative procedure used to maximise equation (2C) to produce locally maximum a posteriori estimates of β. The prior of equation (5C) is such that the maximum a posteriori estimates will tend to be sparse i.e. if a large number of parameters are redundant, many components of β will be zero.

As stated above, the processed component weights which maximise the posterior distribution may be determined using an iterative procedure. Preferable, the iterative procedure for maximising the posterior distribution of the components and component weights is an EM algorithm, such as, for example, that described in Dempster et al, 1977.

In one embodiment, in the case of a general likelihood function (not restricted to generalised linear models) the EM algorithm comprises the steps:

(c) Initialising the algorithm by setting n=0, S0 = {1,2,…, p } , initialise $\varphi^{(0)}$, β^* and applying a value for ε, such as for example $\varepsilon = 10^{-5}$;

(d) Defining

$$\beta_i^{(n)} = \begin{cases} \beta_i^*, & i \in S_n \\ 0, & \text{otherwise} \end{cases} \qquad (5C)$$

and let Pn be a matrix of zeroes and ones such that the nonzero elements $\gamma(n)$ of $\beta(n)$ satisfy

$$\gamma^{(n)} = P_n^T \beta^{(n)} \quad , \quad \beta^{(n)} = P_n \gamma^{(n)}$$
$$\gamma = P_n^T \beta \quad , \quad \beta = P_n \gamma$$

(e) performing an estimation (E) step by calculating the conditional expected value of the posterior distribution of component weights using the function:

$$Q(\beta \mid \beta^{(n)}, \varphi^{(n)}) = E\{ \log p(\beta, \varphi, v^2 \mid y) \mid y, \beta^{(n)}, \phi^{(n)}\} = l(y \mid \beta, \phi^{(n)}) - 0.5 (\|\beta/\beta^{(n)}\|^2) \qquad (6C)$$

where *l* is the log likelihood function of y.

Using $\beta = P_n \gamma$ and $\beta^{(n)} = P_n \gamma^{(n)}$ can be written as

$$Q(\gamma \mid \gamma^{(n)}, \phi^{(n)}) = l(y \mid P_n \gamma, \phi^{(n)}) - 0.5 (\|\gamma/\gamma^{(n)}\|^2) \qquad (7C)$$

(f) performing a maximisation (M) step by applying an iterative procedure to maximise Q as a function of γ whereby $\gamma_0 = \gamma^{(n)}$ and for r=0,1,2,...

(g) $\gamma_{r+1} = \gamma_r + \alpha_r \delta r$ and where α_r is chosen by a line search algorithm to ensure $Q(\gamma_{r+1} | \gamma^{(n)}, \phi^{(n)}) > Q(\gamma_r | \gamma^{(n)}, \phi^{(n)})$, and

$$\delta_r = \mathrm{diag}(\gamma^{(n)})[-\mathrm{diag}(\gamma^{(n)})\frac{\partial^2 l}{\partial^2 \gamma_r}\mathrm{diag}(\gamma^{(n)})+I]^{-1}(\frac{\partial l}{\partial \gamma_r} - \frac{\gamma_r}{\gamma^{(n)}}) \quad (8C)$$

where:

$$\frac{\partial l}{\partial \gamma_r} = P_n^T \frac{\partial l}{\partial \beta_r}, \quad \frac{\partial^2 l}{\partial^2 \gamma_r} = P_n^T \frac{\partial^2 l}{\partial^2 \beta_r} P_n \quad (9C)$$

for $\beta_r = P_n \gamma_r$.

Let γ^* be the value of γr when some convergence criterion is satisfied, for example, $|| \gamma r - \gamma r+1|| < \varepsilon$ (for example 10^{-5});

(h) Defining $\beta^* = P_n \gamma^*$, $S_{n+1}=\{i: |\beta_i| > \max_j(|\beta_j| * \epsilon_1)\}$ where ε_1 is a small constant, for example 1e-5.

(i) Set n=n+1 and choose $\varphi^{(n+1)} = \varphi^{(n)} + \kappa_n(\varphi^* - \varphi^{(n)})$ where φ^* satisfies $\frac{\partial}{\partial \phi} l(y | P_n \gamma^*, \phi) = 0$ and κ_n is a damping factor such that $0 < \kappa_n \leq 1$; and

(j) Check convergence. If $|| \gamma^* - \gamma^{(n)}|| < \varepsilon_2$ where ε_2 is suitably small then stop, else go to step (b) above.

In another embodiment, step (d) in the maximisation step may be estimated by replacing $\frac{\partial^2 l}{\partial^2 \gamma_r}$ with its expectation $E\{\frac{\partial^2 l}{\partial^2 \gamma_r}\}$. This is preferred when the model of the data is a generalised linear model.

For generalised linear models the expected value $E\{\frac{\partial^2 l}{\partial^2 \gamma_r}\}$ may be calculated as follows:

$$\frac{\partial l}{\partial \beta} = X^T \{\mathrm{diag}(\frac{1}{\tau_i^2}\frac{\partial \mu_i}{\partial \eta_i})(\frac{y_i - \mu_i}{a_i(\phi)})\} \qquad (10C)$$

where X is the N by p matrix with i^{th} row x_i^T and

$$E\{\frac{\partial^2 l}{\partial^2 \beta^2}\} = -E\{(\frac{\partial l}{\partial \beta})(\frac{\partial l}{\partial \beta})^T\}$$

$$= -X^T \mathrm{diag}(a_i(\phi)\tau_i^2(\frac{\partial \eta_i}{\partial \mu_i})^2)^{-1}X \qquad (11C)$$

This can be written as

$$\frac{\partial l}{\partial \beta} = X^t V^{-1}(\frac{\partial \eta}{\partial \mu})(y-\mu) \qquad (12C)$$

$$E\{\frac{\partial^2 l}{\partial \beta^2}\} = -X^t V^{-1} X \qquad (13C)$$

where $V=\mathrm{diag}(a_i(\phi)\tau_i^2(\frac{\partial \eta_i}{\partial \mu_i})^2)$.

Preferably, the EM algorithm comprises the steps:

(a) Initialising the algorithm by setting n=0, S0 = {1,2,..., p } , $\varphi(0)$, applying a value for ε, such as for example $\varepsilon = 10^{-5}$, and

If $p \leq N$ compute initial values β^* by

$$\beta^* = (X^t X + \lambda I)^{-1} X^T g(y+\zeta) \qquad (14C)$$

and if $p > N$ compute initial values β^* by

$$\beta^* = \frac{1}{\lambda}(I - X^T(XX^T + \lambda I)^{-1}X)X^T g(y+\zeta) \qquad (15C)$$

where the ridge parameter λ satisfies $0 < \lambda \leq 1$ and ζ is small and chosen so that the link function g is well defined at $y+\zeta$.

(b) Defining

$$\beta_i^{(n)} = \begin{cases} \beta_i^*, & i \in S_n \\ 0, & \text{otherwise} \end{cases}$$

and let Pn be a matrix of zeroes and ones such that the nonzero elements $\gamma(n)$ of $\beta(n)$ satisfy

$$\gamma^{(n)} = P_n^T \beta^{(n)} \quad , \quad \beta^{(n)} = P_n \gamma^{(n)}$$
$$\gamma = P_n^T \beta \quad , \quad \beta = P_n \gamma$$

(c) performing an estimation (E) step by calculating the conditional expected value of the posterior distribution of component weights using the function:

$$\begin{aligned} Q(\beta \mid \beta^{(n)}, \varphi^{(n)}) &= E\{ \log p(\beta, \varphi, \nu^2 \mid y) \mid y, \beta^{(n)}, \phi^{(n)}\} \\ &= l(y \mid \beta, \phi^{(n)}) - 0.5\, (\|\beta/\beta^{(n)}\|^2) \end{aligned} \qquad (16C)$$

where l is the log likelihood function of y. Using $\beta = P_n \gamma$ and $\beta^{(n)} = P_n \gamma^{(n)}$ (16C) can be written as

$$Q(\gamma \mid \gamma^{(n)}, \phi^{(n)}) = l(y \mid P_n \gamma, \phi^{(n)}) - 0.5\, (\|\gamma/\gamma^{(n)}\|^2) \qquad (17C)$$

(d) performing a maximisation (M) step by applying an iterative procedure, for example a Newton Raphson

iteration, to maximise Q as a function of γ whereby $\gamma_0 = \gamma^{(n)}$ and for r=0,1,2,... $\gamma_{r+1} = \gamma_r + \alpha_r \delta_r$ where α_r is chosen by a line search algorithm to ensure $Q(\gamma_{r+1} | \gamma^{(n)}, \phi^{(n)}) > Q(\gamma_r | \gamma^{(n)}, \phi^{(n)})$, and

For $p \leq N$ use

$$\delta_r = \mathrm{diag}(\gamma^{(n)})[Y_n^T V_r^{-1} Y_n + I]^{-1}(Y_n^T V_r^{-1} z_r - \frac{\gamma_r}{\gamma^{(n)}}) \qquad (18C)$$

where

$$Y_n = \mathrm{diag}(\gamma^{(n)}) P_n^T X$$

$$V = \mathrm{diag}(a_i(\phi)\tau_i^2 (\frac{\partial \eta_i}{\partial \mu_i})^2)$$

$$z = \frac{\partial \eta}{\partial \mu}(y-\mu)$$

and the subscript r denotes that these quantities are evaluated at $\mu = h(XP_n\gamma_r)$.

For p > N use

$$\delta_r = \mathrm{diag}(\gamma^{(n)})[I - Y_n^T(Y_n Y_n^T + V_r)^{-1} Y_n](Y_n^T V_r^{-1} z_r - \frac{\gamma_r}{\gamma^{(n)}}) \qquad (19C)$$

with V_r and z_r defined as before.

Let γ^* be the value of γ_r when some convergence criterion is satisfied e.g

$$|| \gamma_r - \gamma_{r+1} || < \varepsilon \quad \text{(for example } 10^{-5}\text{)}.$$

Define $\beta^* = P_n\gamma^*$, $S_{n+1} = \{i: |\beta_i| > \max_j(|\beta_j| * \epsilon_1)\}$ where ε_1 is a small constant, say 1e-5. Set n=n+1 and choose $\varphi^{n+1} = \varphi^n + \kappa_n(\varphi^* - \varphi^n)$ where ϕ^* satisfies $\frac{\partial}{\partial \phi} l(y | P_n\gamma^*, \phi) = 0$ and κ_n is

a damping factor such that $0< \kappa_n \leq 1$. Note that in some cases the scale parameter is known or this equation can be solved explicitly to get an updating equation for φ.

The above embodiments may be extended to incorporate quasi likelihood methods Wedderburn (1974) and McCullagh and Nelder (1983)). In such an embodiment, the same iterative procedure as detailed above will be appropriate, but with the likelihood replaced by a quasi likelihood as shown above and, for example, Table 8.1 in McCullagh and Nelder (1983). In one embodiment there is a modified updating method for the scale parameter φ. To define these models requires specification of the variance function τ^2, the link function g and the derivative of the link function $\frac{\partial \eta}{\partial \mu}$. Once these are defined the above algorithm can be applied. In one embodiment for quasi likelihood models, step 5 of the above algorithm is modified so that the scale parameter is updated by calculating

$$\phi^{(n+1)}=\frac{1}{N}\sum_{i=1}^{N}\frac{(y_i-\mu_i)^2}{\tau_i^2}$$

where μ and τ are evaluated at $\beta^* = P_n\gamma^*$. Preferably, this updating is performed when the number of parameters s in the model is less than N. A divisor of N-s can be used when s is much less than N.

In another embodiment, for both generalised linear models and Quasi likelihood models the covariate matrix X with rows x_i^T can be replaced by a matrix K with ijth element k_{ij} and $k_{ij} = \kappa(x_i-x_j)$ for some kernel function κ. This matrix can also be augmented with a vector of ones. Some example kernels are given in Table 3 below, see Evgeniou et al(1999).

Kernel function	Formula for $\kappa(x - y)$
Gaussian radial basis function	$\exp(- \|\| x - y \|\|^2 / a)$, $a>0$
Inverse multiquadric	$(\|\| x - y \|\|^2 + c^2)^{-1/2}$
multiquadric	$(\|\| x - y \|\|^2 + c^2)^{1/2}$
Thin plate splines	$\|\| x - y \|\|^{2n+1}$ $\|\| x - y \|\|^{2n} \ln(\|\| x - y \|\|)$
Multi layer perceptron	$\tanh(x \cdot y - \theta)$, for suitable θ
Ploynomial of degree d	$(1 + x \cdot y)^d$
B splines	$B_{2n+1}(x - y)$
Trigonometric polynomials	$\sin((d +1/2)(x-y))/\sin((x-y)/2)$

Table 3: Examples of kernel functions

In Table 3 the last two kernels are one dimensional i.e. for the case when X has only one column. Multivariate versions can be derived from products of these kernel functions. The definition of B_{2n+1} can be found in De Boor(1978). Use of a kernel function in either a generalised linear model or a quasi likelihood model results in mean values which are smooth (as opposed to transforms of linear) functions of the covariates X. Such models may give a substantially better fit to the data.

A fourth embodiment relating to use of a proportional hazards model in the multivariate analysis will now be described.

D. Proportional Hazard Models

The multivariate method of this embodiment utilises the training samples in order to identify a subset of

components which are capable of predicting the probability that a pre-defined event (e.g. death, recovery) will occur within a certain time period. Thus, by considering the time to an event of an animal following administration of a sample (e.g. compound), it can be determined what the likely survival time of the animal will be. The desired property in the case of a therapeutic compound may be a long survival time in an animal and thus the multivariate method may be used to identify components that are predictive of whether administering the sample to an animal will result in a long survival time of the animal, preferably indicating low toxicity or efficacy of the sample.

Data for training samples are obtained and the time measured from when the training sample is applied to when the event occurs. Using a statistical method to associate the time to the event with the data obtained, a subset of components may be identified that are capable of predicting the distribution of the time to the event.

As used herein, "time to an event" refers to a measure of the time from applying the sample and measuring the time to the event. An event may be any observable event. The event may be, for example, death of an animal, onset of symptoms or side effects from a sample, onset of excretion of a metabolite resulting from applying a sample, change in behaviour as a result of administering a sample, change in biochemistry of a cell, tissue or animal following application of the sample, change in morphology of a cell, organism or animal following administration of a sample.

The samples are associated with a particular time to an event. The times to an event may be times determined from data obtained from, for example, animals in which the time to death is known, or in other words, "genuine" survival times, and animals in which only the information that the animal is alive when last assessed, or in other words, "censored" survival times indicating that the particular animal has survived for at least a given number of days.

In one embodiment, the input data is organised into an N x p data matrix $X=\left(x_{ij}\right)$ with N test training samples and p components. Typically, p will be much greater than N.

For example, consider an Nxp data matrix $X=\left(x_{ij}\right)$ from, for example, a microarray experiment, with N individuals (or samples) and the same p genes for each individual. Preferably, there is associated with each individual $i\ (i=1,2,\cdots,N)$ a variable y_i ($y_i \geq 0$) denoting the time to an event, for example, survival time. For each individual there may also be defined a variable that indicates whether that individual's survival time is a genuine survival time or a censored survival time. Denote the censor indicators as c_i where

$$c_i=\begin{cases}1, \text{if } y_i \text{ is uncensored}\\ 0, \text{if } y_i \text{ is censored}\end{cases}$$

The Nx1 vector with survival times y_i may be written as $\underset{\sim}{y}$ and the Nx1 vector with censor indicators c_i as $\underset{\sim}{c}$.

Typically, as discussed above, the processed component weights are estimated in a manner which takes into account

the a priori assumption that most of the component weights are zero.

Preferably, the prior specified for the component weights is of the form

$$P(\beta_1,\beta_2,\cdots,\beta_n)=\int_{\tau}\prod_{i=1}^{N}P(\beta_i|\tau_i)P(\tau_i)d\tau \qquad (1D)$$

where $\beta_1,\beta_2,\cdots,\beta_n$ are component weights, $P(\beta_i|\tau_i)$ is $N(0,\tau_i^2)$ and $P(\tau_i)\alpha 1/\tau_i^2$ is a Jeffreys prior (Kotz and Johnson, 1983).

The likelihood function defines a model which fits the data based on the distribution of the data. Preferably, the likelihood function is of the form:

$$Log\ (Partial)\ Likelihood = \sum_{i=1}^{N} g_i\left(\underset{\sim}{\beta},\underset{\sim}{\varphi};X,\underset{\sim}{y},\underset{\sim}{c}\right) \qquad (2D)$$

where $\underset{\sim}{\beta}^T=(\beta_1,\beta_2,\cdots,\beta_p)$ and $\underset{\sim}{\varphi}^T=(\varphi_1,\varphi_2,\cdots,\varphi_q)$ are the model parameters. The model defined by the likelihood function may be any model for predicting the time to an event of a system.

In one embodiment, the model defined by the likelihood is Cox's proportional hazards model. Cox's proportional hazards model was introduced by Cox (1972) and may preferably be used as a regression model for survival data. In Cox's proportional hazards model, $\underset{\sim}{\beta}^T$ is a vector of (explanatory) parameters associated with the components. Preferably, the method of the present invention provides for the parsimonious selection (and estimation) from the parameters $\underset{\sim}{\beta}^T=(\beta_1,\beta_2,\cdots,\beta_p)$ for Cox's proportional hazards model given the data X, $\underset{\sim}{y}$ and $\underset{\sim}{c}$.

Application of Cox's proportional hazards model can be problematic in the circumstance where different data is obtained from a system for the same survival times, or in other words, for cases where tied survival times occur. Tied survival times may be subjected to a pre-processing step that leads to unique survival times. The pre-processing proposed simplifies the ensuing algorithm as it avoids concerns about tied survival times in the subsequent application of Cox's proportional hazards model.

The pre-processing of the survival times applies by adding an extremely small amount of insignificant random noise. Preferably, the procedure is to take sets of tied times and add to each tied time within a set of tied times a random amount that is drawn from a normal distribution that has zero mean and variance proportional to the smallest non-zero distance between sorted survival times. Such pre-processing achieves an elimination of tied times without imposing a draconian perturbation of the survival times.

The pre-processing generates distinct survival times. Preferably, these times may be ordered in increasing magnitude denoted as $\underset{\sim}{t} = \left(t_{(1)}, t_{(2)}, \cdots t_{(N)}\right),\ t_{(i+1)} > t_{(i)}$.

Denote by Z the Nxp matrix that is the re-arrangement of the rows of X where the ordering of the rows of Z corresponds to the ordering induced by the ordering of $\underset{\sim}{t}$; also denote by Z_j the j^{th} row of the matrix Z. Let d be the result of ordering c with the same permutation required to order $\underset{\sim}{t}$.

After pre-processing for tied survival times is taken into account and reference is made to standard texts on survival data analysis (eg Cox and Oakes, 1984), the likelihood function for the proportional hazards model may preferably be written as

$$L\left(\underset{\sim}{t}\,|\,\underset{\sim}{\beta}\right)=\prod_{j=1}^{N}\left(\frac{exp\left(Z_j\underset{\sim}{\beta}\right)}{\sum_{i\in\Re_j}exp\left(Z_i\underset{\sim}{\beta}\right)}\right)^{d_j} \qquad (3D)$$

where $\underset{\sim}{\beta}^T=\left(\beta_1,\beta_2,\cdots,\beta_n\right)$, Z_j = the j^{th} row of Z, and $\Re_j=\left\{i:i=j,j+1,\cdots,N\right\}$ = the risk set at the j^{th} ordered event time $t_{(j)}$.

The logarithm of the likelihood (ie $l=log\left(L\right)$) may preferably be written as

$$l\left(\underset{\sim}{t}\,|\,\underset{\sim}{\beta}\right)=\sum_{i=1}^{N}d_i\left(Z_i\underset{\sim}{\beta}-log\left(\sum_{j\in\Re_i}exp\left(Z_j\underset{\sim}{\beta}\right)\right)\right)$$

$$=\sum_{i=1}^{N}d_i\left(Z_i\underset{\sim}{\beta}-log\left(\sum_{j=1}^{N}\zeta_{i,j}\,exp\left(Z_j\underset{\sim}{\beta}\right)\right)\right), \qquad (4D)$$

where

$$\zeta_{i,j}=\begin{cases}0, & \text{if } j<i\\ 1, & \text{if } j\geq i\end{cases}$$

Notice that the model is non-parametric in that the parametric form of the survival distribution is not specified - preferably only the ordinal property of the survival times are used (in the determination of the risk sets). As this is a non-parametric case $\underset{\sim}{\varphi}$ is not required (ie q=0).

In another embodiment of the method of the invention, the model defined by the likelihood function is a parametric survival model. Preferably, in a parametric survival model, $\underset{\sim}{\beta}^T$ is a vector of (explanatory) parameters associated with the components, and $\underset{\sim}{\varphi}^T$ is a vector of parameters associated with the functional form of the survival density function.

Preferably, the method of the invention provides for the parsimonious selection (and estimation) from the parameters $\underset{\sim}{\beta}^T$ and the estimation of $\underset{\sim}{\varphi}^T = \left(\varphi_1, \varphi_2, \cdots, \varphi_q\right)$ for parametric survival models given the data X, $\underset{\sim}{y}$ and $\underset{\sim}{c}$.

In applying a parametric survival model, the survival times do not require pre-processing and are denoted as $\underset{\sim}{y}$.

The parametric survival model is applied as follows:

Denote by $f\left(y;\underset{\sim}{\theta},\underset{\sim}{\beta},X\right)$ the parametric density function of the survival time, denote its survival function by $\mathbb{S}\left(y;\underset{\sim}{\varphi},\underset{\sim}{\beta},X\right) = \int_y^{\infty} f\left(u;\underset{\sim}{\varphi},\underset{\sim}{\beta},X\right)du$ where $\underset{\sim}{\varphi}$ are the parameters relevant to the parametric form of the density function and $\underset{\sim}{\beta},X$ are as defined above. The hazard function is defined as $h\left(y_i;\underset{\sim}{\varphi},\underset{\sim}{\beta},X\right) = f\left(y_i;\underset{\sim}{\varphi},\underset{\sim}{\beta},X\right) / \mathbb{S}\left(y_i;\underset{\sim}{\varphi},\underset{\sim}{\beta},X\right)$.

Preferably, the generic formulation of the log-likelihood function, taking censored data into account, is

$$l = \sum_{i=1}^{N} \left\{ c_i \log\left(f\left(y_i;\underset{\sim}{\varphi},\underset{\sim}{\beta},X\right)\right) + \left(1-c_i\right)\log\left(\mathbb{S}\left(y_i;\underset{\sim}{\varphi},\underset{\sim}{\beta},X\right)\right)\right\}$$

Reference to standard texts on analysis of survival time data via parametric regression survival models reveals a collection of survival time distributions that may be

used. Survival distributions that may be used include, for example, the Weibull, Exponential or Extreme Value distributions.

If the hazard function may be written as

$h\left(y_i;\underset{\sim}{\varphi},\underset{\sim}{\beta},X\right)=\lambda\left(y_i;\underset{\sim}{\varphi}\right)exp\left(X_i\underset{\sim}{\beta}\right)$ then $\mathbb{S}\left(y_i;\underset{\sim}{\varphi},\underset{\sim}{\beta},X\right)=exp\left(-\Lambda\left(y_i;\underset{\sim}{\varphi}\right)e^{X_i\underset{\sim}{\beta}}\right)$ and

$f\left(y_i;\underset{\sim}{\varphi},\underset{\sim}{\beta},X\right)=\lambda\left(y_i;\underset{\sim}{\varphi}\right)exp\left(X_i\underset{\sim}{\beta}-\Lambda\left(y_i\right)e^{X_i\underset{\sim}{\beta}}\right)$ where

$\Lambda\left(y_i;\underset{\sim}{\varphi}\right)=\int_{-\infty}^{y_i}\lambda\left(u;\underset{\sim}{\varphi}\right)du$ is the integrated hazard function and

$\lambda\left(y_i;\underset{\sim}{\varphi}\right)=\frac{d\Lambda\left(y_i;\underset{\sim}{\varphi}\right)}{dy_i}$; X_i is the i^{th} row of X.

The Weibull, Exponential and Extreme Value distributions have density and hazard functions that may be written in the form of those presented in the paragraph immediately above.

The application detailed relies in part on an algorithm of Aitken and Clayton (1980) however it permits the user to specify any parametric underlying hazard function.

Following from Aitkin and Clayton (1980) a preferred likelihood function which models a parametric survival model is:

$$l=\sum_{i=1}^{N}\left\{c_i\,log\left(\mu_i\right)-\mu_i+c_i\left(log\left(\frac{\lambda\left(y_i\right)}{\Lambda\left(y_i;\underset{\sim}{\varphi}\right)}\right)\right)\right\} \qquad (5D)$$

where $\mu_i=\Lambda\left(y_i;\underset{\sim}{\varphi}\right)exp\left(X_i\underset{\sim}{\beta}\right)$. Aitkin and Clayton (1980) note that a consequence of equation (5D) is that the c_i's may be treated as Poisson variates with means μ_i and that the last term in equation (11D) does not depend on $\underset{\sim}{\beta}$ (although it depends on $\underset{\sim}{\varphi}$).

Preferably, the posterior distribution of $\underset{\sim}{\beta}$, $\underset{\sim}{\varphi}$ and $\underset{\sim}{\tau}$ given $\underset{\sim}{y}$ is

$$P\left(\underset{\sim}{\beta},\underset{\sim}{\varphi},\underset{\sim}{\tau}\middle|\underset{\sim}{y}\right) \alpha\, L\left(\underset{\sim}{y}\middle|\underset{\sim}{\beta},\underset{\sim}{\varphi}\right)P\left(\underset{\sim}{\beta}\middle|\underset{\sim}{\tau}\right)P\left(\underset{\sim}{\tau}\right) \qquad (6D)$$

wherein $L\left(\underset{\sim}{y}\middle|\underset{\sim}{\beta},\underset{\sim}{\varphi}\right)$ is the likelihood function.

In one embodiment, $\underset{\sim}{\tau}$ may be treated as a vector of missing data and an iterative procedure used to maximise equation (6D) to produce a posteriori estimates of $\underset{\sim}{\beta}$. The prior of equation (1D) is such that the maximum a posteriori estimates will tend to be sparse i.e. if a large number of parameters are redundant, many components of $\underset{\sim}{\beta}$ will be zero.

Because a prior expectation exists that many components of $\underset{\sim}{\beta}^T$ are zero, the estimation may be performed in such a way that most of the estimated β_i's are zero and the remaining non-zero estimates provide an adequate explanation of the survival times.

In the context of microarray data this exercise translates to identifying a parsimonious set of genes that provide an adequate explanation for the event times.

As stated above, the component weights which maximise the posterior distribution may be determined using an iterative procedure. Preferable, the iterative procedure for maximising the posterior distribution of the components and component weights is an EM algorithm, such as, for example, that described in Dempster et al, 1977.

In one embodiment, the EM algorithm comprises the steps:

1. Initialising the algorithm by setting $n=0$, $S_0 = \{1,2,\ldots, p\}$, initialise $\underset{\sim}{\beta}^{(0)} = \underset{\sim}{\beta}^{*}$, $\underset{\sim}{\varphi}^{(0)}$,

2. Defining

$$\beta_i^{(n)} = \begin{cases} \beta_i^{*}, & i \varepsilon S_n \\ 0, & \text{otherwise} \end{cases}$$

and let P_n be a matrix of zeroes and ones such that the nonzero elements $\underset{\sim}{\gamma}^{(n)}$ of $\underset{\sim}{\beta}^{(n)}$ satisfy

$$\begin{aligned} \underset{\sim}{\gamma}^{(n)} &= P_n^T \underset{\sim}{\beta}^{(n)} \quad , \quad \underset{\sim}{\beta}^{(n)} = P_n \underset{\sim}{\gamma}^{(n)} \\ \underset{\sim}{\gamma} &= P_n^T \underset{\sim}{\beta} \quad , \quad \underset{\sim}{\beta} = P_n \underset{\sim}{\gamma} \end{aligned} \tag{7D}$$

3. Performing an estimation step by calculating the expected value of the posterior distribution of component weights. This may be performed using the function:

$$\begin{aligned} Q(\underset{\sim}{\beta} \mid \underset{\sim}{\beta}^{(n)}, \varphi^{(n)}) &= E\left\{ \log\left(P\left(\underset{\sim}{\beta}, \underset{\sim}{\varphi}, \tau \mid \underset{\sim}{y} \right) \right) \mid \underset{\sim}{y}, \underset{\sim}{\beta}^{(n)}, \varphi^{(n)} \right\} \\ &= l(\underset{\sim}{y} \mid \underset{\sim}{\beta}, \varphi^{(n)}) - \frac{1}{2} \sum_{i=1}^{N} \left(\frac{\beta_i}{\beta_i^{(n)}} \right)^2 \end{aligned} \tag{8D}$$

where l is the log likelihood function of $\underset{\sim}{y}$. Using $\beta = P_n \gamma$ and $\beta^{(n)} = P_n \gamma^{(n)}$ we have

$$Q(\underset{\sim}{\gamma} \mid \underset{\sim}{\gamma}^{(n)}, \underset{\sim}{\varphi}^{(n)}) = l(\underset{\sim}{t} \mid P_n \underset{\sim}{\gamma}, \underset{\sim}{\varphi}^{(n)}) - \frac{1}{2} \sum_{i=1}^{N} \left(\frac{\gamma_i}{\gamma_i^{(n)}} \right)^2 \tag{9D}$$

4. Performing the maximisation step. This may be performed using Newton Raphson iterations as follows:

Set $\underset{\sim}{\gamma}_0 = \underset{\sim}{\gamma}^{(r)}$ and for r=0,1,2,...

$\underset{\sim}{\gamma}_{r+1} = \underset{\sim}{\gamma}_r + \alpha_r \underset{\sim}{\delta}_r$ where α_r is chosen by a line search

algorithm to ensure $Q\left(\underset{\sim}{\gamma}_{r+1} \mid \underset{\sim}{\gamma}^{(n)}, \underset{\sim}{\varphi}^{(n)}\right) > Q\left(\underset{\sim}{\gamma}_{r} \mid \underset{\sim}{\gamma}^{(n)}, \underset{\sim}{\varphi}^{(n)}\right)$, and

$$\underset{\sim}{\delta}_r = \mathrm{diag}(\underset{\sim}{\gamma}^{(n)})[-\mathrm{diag}(\underset{\sim}{\gamma}^{(n)})\frac{\partial^2 l}{\partial^2 \underset{\sim}{\gamma}_r}\mathrm{diag}(\underset{\sim}{\gamma}^{(n)}) + I]^{-1}(\frac{\partial l}{\partial \underset{\sim}{\gamma}_r} - \frac{\underset{\sim}{\gamma}_r}{\underset{\sim}{\gamma}^{(n)}})$$

where $\frac{\partial l}{\partial \underset{\sim}{\gamma}_r} = P_n^T \frac{\partial l}{\partial \underset{\sim}{\beta}_r}$, $\frac{\partial^2 l}{\partial^2 \underset{\sim}{\gamma}_r} = P_n^T \frac{\partial^2 l}{\partial^2 \underset{\sim}{\beta}_r} P_n$ for $\underset{\sim}{\beta}_r = P_n \underset{\sim}{\gamma}_r$

(10D)

Let $\underset{\sim}{\gamma}^*$ be the value of $\underset{\sim}{\gamma}_r$ when some convergence criterion is satisfied e.g $||\underset{\sim}{\gamma}_r - \underset{\sim}{\gamma}_{r+1}|| < \varepsilon$ (for example $\varepsilon = 10^{-5}$).

5. Define $\underset{\sim}{\beta}^* = P_n \underset{\sim}{\gamma}^*$, $S_n = \left\{ i : |\beta_i| > \varepsilon_1 \max_j |\beta_j| \right\}$ where ε_1 is a small constant, say 10^{-5}. Set n=n+1, choose $\underset{\sim}{\varphi}^{(n+1)} = \underset{\sim}{\varphi}^{(n)} + \kappa_n\left(\underset{\sim}{\varphi}^* - \underset{\sim}{\varphi}^{(n)}\right)$ where $\underset{\sim}{\varphi}^*$ satisfies $\frac{\partial l\left(\underset{\sim}{y} \mid P_n \underset{\sim}{\gamma}^*, \underset{\sim}{\varphi}\right)}{\partial \underset{\sim}{\varphi}} = 0$ and κ_n is a damping factor such that $0 < \kappa_n < 1$.

6. Check convergence. If $||\underset{\sim}{\gamma}^* - \underset{\sim}{\gamma}^{(n)}|| < \varepsilon_2$ where ε_2 is suitably small then stop, else go to step 2 above.

In another embodiment, step (4) in the maximisation step may be estimated by replacing $\frac{\partial^2 l}{\partial^2 \gamma_r}$ with its expectation $E\{\frac{\partial^2 l}{\partial^2 \gamma_r}\}$.

In one embodiment, the EM algorithm is applied to maximise the posterior distribution when the model is Cox's proportional hazard's model.

To aid in the exposition of the application of the EM algorithm when the model is Cox's proportional hazards model, it is preferred to define "dynamic weights" and matrices based on these weights. The weights are -

$$w_{i,l} = \frac{\zeta_{i,l} \exp\left(Z_l \underset{\sim}{\beta}\right)}{\sum_{j=1}^{N} \zeta_{i,j} \exp\left(Z_j \underset{\sim}{\beta}\right)},$$

$$w_l^* = \sum_{i=1}^{N} d_i w_{i,l},$$

$$\tilde{w}_l = d_l - w_l^*.$$

Matrices based on these weights are -

$$W_i = \begin{pmatrix} w_{i,1} \\ w_{i,2} \\ \vdots \\ \vdots \\ w_{i,N} \end{pmatrix},$$

$$\tilde{W} = \begin{pmatrix} \tilde{w}_1 \\ \tilde{w}_2 \\ \vdots \\ \vdots \\ \tilde{w}_N \end{pmatrix},$$

$$\Delta\left(W^*\right) = \begin{pmatrix} w_1^* & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & w_N^* \end{pmatrix},$$

$$W^{**} = \sum_{i=1}^{N} d_i W_i W_i^T$$

In terms of the matrices of weights the first and second derivatives of l may be written as -

$$\left.\begin{aligned} \frac{\partial l}{\partial \underset{\sim}{\beta}} &= Z^T \tilde{W} \\ \frac{\partial^2 l}{\partial \underset{\sim}{\beta}^2} &= Z^T \left(W^{**} - \Delta\left(W^*\right)\right) Z = Z^T K Z \end{aligned}\right\} \qquad (11D)$$

where $K = W^{**} - \Delta\left(W^*\right)$. Note therefore from the transformation matrix P_n described as part of Step (2) of the EM algorithm (Equation 7D) (see also Equations (10D)) it follows that

$$\left.\begin{aligned} \frac{\partial l}{\partial \underset{\sim}{\gamma}_r} &= P_n^T \frac{\partial l}{\partial \underset{\sim}{\beta}_r} = P_n^T Z^T \tilde{W} \\ \frac{\partial^2 l}{\partial \underset{\sim}{\gamma}_r^2} &= P_n^T \frac{\partial^2 l}{\partial \underset{\sim}{\beta}_r^2} P_n = P_n^T Z^T \left(W^{**} - \Delta\left(W^*\right)\right) Z P_n = P_n^T Z^T K Z P_n \end{aligned}\right\} \qquad (12D)$$

Preferably, when the model is Cox's proportional hazards model the E step and M step of the EM algorithm are as follows:

1.1. Set $n=0$, $S_0 = \{1, 2, \cdots, p\}$. Let v be the vector with components

$$\mathbf{v}_i = \begin{cases} 1-\varepsilon, & \text{if } c_i = 1 \\ \varepsilon, & \text{if } c_i = 0 \end{cases}$$

for some small ε, say .001. Define f to be log(v/t).

If $p \leq N$ compute initial values $\underset{\sim}{\beta}^*$ by

$$\underset{\sim}{\beta}^* = (Z^T Z + \lambda I)^{-1} Z^T \mathrm{f} \quad .$$

If $p > N$ compute initial values $\underset{\sim}{\beta}^*$ by

$$\underset{\sim}{\beta}^* = \frac{1}{\lambda}(I - Z^T (Z Z^T + \lambda I)^{-1} Z) Z^T \mathrm{f}$$

where the ridge parameter λ satisfies $0 < \lambda \leq 1$.

2. Define

$$\beta_i^{(n)} = \begin{cases} \beta_i^*, & i \in S_n \\ 0, & \text{otherwise} \end{cases}$$

Let P_n be a matrix of zeroes and ones such that the nonzero elements $\underset{\sim}{\gamma}^{(n)}$ of $\underset{\sim}{\beta}^{(n)}$ satisfy

$$\underset{\sim}{\gamma}^{(n)} = P_n^T \underset{\sim}{\beta}^{(n)} \quad , \quad \underset{\sim}{\beta}^{(n)} = P_n \underset{\sim}{\gamma}^{(n)}$$
$$\underset{\sim}{\gamma} = P_n^T \underset{\sim}{\beta} \quad , \quad \underset{\sim}{\beta} = P_n \underset{\sim}{\gamma}$$

3. Perform the E step by calculating

$$Q(\underset{\sim}{\beta} \mid \underset{\sim}{\beta}^{(n)}) = E\left\{\log\left(P\left(\underset{\sim}{\beta}, \underset{\sim}{\varphi}, \tau \mid \underset{\sim}{t}\right)\right) \mid \underset{\sim}{t}, \underset{\sim}{\beta}^{(n)}\right\}$$
$$= l(\underset{\sim}{t} \mid \underset{\sim}{\beta}) - \frac{1}{2}\sum_{i=1}^{N}\left(\frac{\beta_i}{\beta_i^{(n)}}\right)^2$$

where l is the log likelihood function of $\underset{\sim}{t}$ given by Equation (8D). Using $\beta = P_n \gamma$ and $\beta^{(n)} = P_n \gamma^{(n)}$ we have

$$Q(\underset{\sim}{\gamma} \mid \underset{\sim}{\gamma}^{(n)}) = l(\underset{\sim}{t} \mid P_n \underset{\sim}{\gamma}) - \frac{1}{2}\sum_{i=1}^{N}\left(\frac{\gamma_i}{\gamma_i^{(n)}}\right)^2$$

4. Do the M step. This can be done with Newton Raphson iterations as follows. Set $\underset{\sim}{\gamma}_0 = \underset{\sim}{\gamma}^{(r)}$ and for r=0,1,2,... $\underset{\sim}{\gamma}_{r+1} = \underset{\sim}{\gamma}_r + \alpha_r \underset{\sim}{\delta}_r$ where α_r is chosen by a line search algorithm to ensure $Q\left(\underset{\sim}{\gamma}_{r+1} \mid \underset{\sim}{\gamma}^{(n)}, \underset{\sim}{\varphi}^{(n)}\right) > Q\left(\underset{\sim}{\gamma}_r \mid \underset{\sim}{\gamma}^{(n)}, \underset{\sim}{\varphi}^{(n)}\right)$.

For $p \leq N$ use

$$\underset{\sim}{\delta}_r = diag\left(\underset{\sim}{\gamma}^{(n)}\right)\left(Y^T K Y + I\right)^{-1}\left(Y^T \tilde{W} - diag\left(1/\underset{\sim}{\gamma}^{(n)}\right)\underset{\sim}{\gamma}\right),$$
$$\text{where } Y = ZP_n \text{diag}\left(\underset{\sim}{\gamma}^{(n)}\right).$$

For p > N use

$$\underset{\sim}{\delta}_r = diag\left(\underset{\sim}{\gamma}^{(n)}\right)\left(I - Y^T\left(YY^T + K^{-1}\right)^{-1} Y\right)\left(Y^T \tilde{W} - diag\left(1/\underset{\sim}{\gamma}^{(n)}\right)\underset{\sim}{\gamma}\right)$$

Let γ^* be the value of γ_r when some convergence criterion is satisfied e.g $||\gamma_r - \gamma_{r+1}|| < \varepsilon$ (for example 10^{-5}).

5. Define $\underset{\sim}{\beta}^* = P_n \underset{\sim}{\gamma}^*$, $S_n = \left\{ i : |\beta_i| > \varepsilon_1 \max_j |\beta_j| \right\}$ where ε_1 is a small constant, say 10^{-5}. This step eliminates variables with very small coefficients.

6. Check convergence. If $||\underset{\sim}{\gamma}^* - \underset{\sim}{\gamma}^{(n)}|| < \varepsilon_2$ where ε_2 is suitably small then stop, else set n=n+1, go to step 2 above and repeat procedure until convergence occurs.

In another embodiment the EM algorithm is applied to maximise the posterior distribution when the model is a parametric survival model.

In applying the EM algorithm to the parametic survival model, a consequence of equation (5D) is that the c_i's may be treated as Poisson variates with means μ_i and that the last term in equation (5D) does not depend on β (although it depends on φ).

Note that $log(\mu_i) = log\left(\Lambda\left(y_i; \underset{\sim}{\varphi}\right)\right) + X_i \underset{\sim}{\beta}$ and so it is possible to couch the problem in terms a log-linear model for the Poisson-like mean. Preferably, an iterative maximization of the log-likelihood function is performed where given initial estimates of $\underset{\sim}{\varphi}$ the estimates of $\underset{\sim}{\beta}$ are obtained. Then given these estimates of $\underset{\sim}{\beta}$, updated estimates of $\underset{\sim}{\varphi}$ are obtained. The procedure is continued until convergence occurs.

Applying the posterior distribution described above, we note that (for fixed $\underset{\sim}{\varphi}$)

$$\frac{\partial log(\mu)}{\partial \underset{\sim}{\beta}} = \frac{1}{\mu}\frac{\partial \mu}{\partial \underset{\sim}{\beta}} \Leftrightarrow \frac{\partial \mu}{\partial \underset{\sim}{\beta}} = \mu \frac{\partial log(\mu)}{\partial \underset{\sim}{\beta}} \text{ and } \frac{\partial log(\mu_i)}{\partial \underset{\sim}{\beta}} = X_i \qquad (13D)$$

Consequently from Equations (11D) and (12D) it follows that

$$\frac{\partial l}{\partial \underset{\sim}{\beta}} = X^T\left(\underset{\sim}{c} - \underset{\sim}{\mu}\right) \text{ and } \frac{\partial^2 l}{\partial \underset{\sim}{\beta}^2} = -X^T \text{diag}\left(\underset{\sim}{\mu}\right) X .$$

The versions of Equation (12D) relevant to the parametric survival models are

$$\left.\begin{aligned} \frac{\partial l}{\partial \underset{\sim}{\gamma}_r} &= P_n^T \frac{\partial l}{\partial \underset{\sim}{\beta}_r} = P_n^T X^T\left(\underset{\sim}{c} - \underset{\sim}{\mu}\right) \\ \frac{\partial^2 l}{\partial \underset{\sim}{\gamma}^2{}_r} &= P_n^T \frac{\partial^2 l}{\partial \underset{\sim}{\beta}_r{}^2} P_n = -P_n^T X^T \text{diag}\left(\underset{\sim}{\mu}\right) X P_n \end{aligned}\right\} \qquad (14D)$$

To solve for $\underset{\sim}{\varphi}$ after each M step of the EM algorithm (see step 5 below) preferably put $\underset{\sim}{\varphi}^{(n+1)} = \underset{\sim}{\varphi}^{(n+1)} + \kappa_n\left(\underset{\sim}{\varphi}^* - \underset{\sim}{\varphi}^{(n)}\right)$ where $\underset{\sim}{\varphi}^*$ satisfies $\frac{\partial l}{\partial \underset{\sim}{\varphi}} = 0$ for $0 < \kappa_n \leq 1$ and β is fixed at the value obtained from the previous M step.

It is possible to provide an EM algorithm for parameter selection in the context of parametric survival models and microarray data. Preferably, the EM algorithm is as follows:

1. Set $n=0$, $S_0 = \{1, 2, \cdots, p\}$ $\underset{\sim}{\varphi}^{(initial)} = \underset{\sim}{\varphi}^{(0)}$. Let v be the vector with components

$$v_i = \begin{cases} 1-\varepsilon, & if \quad c_i = 1 \\ \varepsilon, & if \quad c_i = 0 \end{cases}$$

for some small ε, say for example .001. Define f to be $\log(v/\Lambda(y,\varphi))$.

If $p \leq N$ compute initial values $\underset{\sim}{\beta}^*$ by $\underset{\sim}{\beta}^* = (X^T X + \lambda I)^{-1} X^T f$.

If $p > N$ compute initial values $\underset{\sim}{\beta}^*$ by

$$\underset{\sim}{\beta}^* = \frac{1}{\lambda}(I - X^T(XX^T + \lambda I)^{-1}X)X^T\mathrm{f}$$

where the ridge parameter λ satisfies $0 < \lambda \leq 1$.

2. Define

$$\beta_i^{(n)} = \begin{cases} \beta_i^*, & i \in S_n \\ 0, & \text{otherwise} \end{cases}$$

Let P_n be a matrix of zeroes and ones such that the nonzero elements $\underset{\sim}{\gamma}^{(n)}$ of $\underset{\sim}{\beta}^{(n)}$ satisfy

$$\underset{\sim}{\gamma}^{(n)} = P_n^T\underset{\sim}{\beta}^{(n)} \quad , \quad \underset{\sim}{\beta}^{(n)} = P_n\underset{\sim}{\gamma}^{(n)}$$
$$\underset{\sim}{\gamma} = P_n^T\underset{\sim}{\beta} \quad , \quad \underset{\sim}{\beta} = P_n\underset{\sim}{\gamma}$$

3. Perform the E step by calculating

$$Q(\underset{\sim}{\beta} \mid \underset{\sim}{\beta}^{(n)}, \underset{\sim}{\varphi}^{(n)}) = E\left\{\log\left(P\left(\underset{\sim}{\beta}, \underset{\sim}{\varphi}, \tau \mid \underset{\sim}{y}\right)\right) \mid \underset{\sim}{y}, \underset{\sim}{\beta}^{(n)}, \underset{\sim}{\varphi}^{(n)}\right\}$$
$$= l(\underset{\sim}{y} \mid \underset{\sim}{\beta}, \underset{\sim}{\varphi}^{(n)}) - \frac{1}{2}\sum_{i=1}^{N}\left(\frac{\beta_i}{\beta_i^{(n)}}\right)^2$$

where l is the log likelihood function of $\underset{\sim}{y}$ and $\underset{\sim}{\varphi}^{(n)}$.

Using $\beta = P_n\gamma$ and $\beta^{(n)} = P_n\gamma^{(n)}$ we have

$$Q(\underset{\sim}{\gamma} \mid \underset{\sim}{\gamma}^{(n)}, \underset{\sim}{\varphi}^{(n)}) = l(\underset{\sim}{y} \mid P_n\underset{\sim}{\gamma}, \underset{\sim}{\varphi}^{(n)}) - \frac{1}{2}\sum_{i=1}^{N}\left(\frac{\gamma_i}{\gamma_i^{(n)}}\right)^2$$

4. Do the M step. This can be done with Newton Raphson iterations as follows. Set $\underset{\sim}{\gamma}_0 = \underset{\sim}{\gamma}^{(r)}$ and for r=0,1,2,... $\underset{\sim}{\gamma}_{r+1} = \underset{\sim}{\gamma}_r + \alpha_r\,\underset{\sim}{\delta}_r$ where α_r is chosen by a line search algorithm to ensure $Q\left(\underset{\sim}{\gamma}_{r+1} \mid \underset{\sim}{\gamma}^{(n)}, \underset{\sim}{\varphi}^{(n)}\right) > Q\left(\underset{\sim}{\gamma}_r \mid \underset{\sim}{\gamma}^{(n)}, \underset{\sim}{\varphi}^{(n)}\right)$.

For $p \leq N$ use

$$\underset{\sim}{\delta}_r = -\mathrm{diag}(\underset{\sim}{\gamma}^{(n)})[Y_n^T\mathrm{diag}\left(\underset{\sim}{\mu}\right)Y_n + I]^{-1}(Y_n^T\left(\underset{\sim}{c} - \underset{\sim}{\mu}\right) - \mathrm{diag}\left(1/\underset{\sim}{\gamma}^{(n)}\right)\underset{\sim}{\gamma})$$

where $Y = XP_n\mathrm{diag}\left(\underset{\sim}{\gamma}^{(n)}\right)$.

For p > N use

$$\underset{\sim}{\delta}_r = -\mathrm{diag}\left(\underset{\sim}{\gamma}^{(n)}\right)\left(I - Y^T\left(YY^T + \mathrm{diag}\left(1/\underset{\sim}{\mu}\right)\right)^{-1} Y\right)\left(Y_n^T\left(\underset{\sim}{c} - \underset{\sim}{\mu}\right) - \mathrm{diag}\left(1/\underset{\sim}{\gamma}^{(n)}\right)\underset{\sim}{\gamma}\right)$$

Let γ^* be the value of γ_r when some convergence criterion is satisfied e.g $||\ \gamma_r - \gamma_{r+1}\ ||\ < \varepsilon$ (for example 10^{-5}).

5. Define $\underset{\sim}{\beta}^* = P_n\underset{\sim}{\gamma}^*$, $S_n = \left\{i : |\beta_i| > \varepsilon_1 \max_j |\beta_j|\right\}$ where ε_1 is a small constant, say 10^{-5}. Set n=n+1, choose

$\underset{\sim}{\varphi}^{(n+1)} = \underset{\sim}{\varphi}^{(n)} + \kappa_n\left(\underset{\sim}{\varphi}^* - \underset{\sim}{\varphi}^{(n)}\right)$ where $\underset{\sim}{\varphi}^*$ satisfies $\dfrac{\partial l\left(\underset{\sim}{y} \mid P_n\underset{\sim}{\gamma}^*, \underset{\sim}{\varphi}\right)}{\partial \underset{\sim}{\varphi}} = 0$

and κ_n is a damping factor such that $0 < \kappa_n < 1$.

6. Check convergence. If $||\underset{\sim}{\gamma}^* - \underset{\sim}{\gamma}^{(n)}|| < \varepsilon_2$ where ε_2 is suitably small then stop, else go to step 2.

In another embodiment, survival times are described by a Weibull survival density function. For the Weibull case $\underset{\sim}{\varphi}$ is preferably one dimensional and

$$\Lambda\left(y; \underset{\sim}{\varphi}\right) = y^{\alpha},$$
$$\lambda\left(y; \underset{\sim}{\varphi}\right) = \alpha y^{\alpha-1},$$
$$\underset{\sim}{\varphi} = \alpha$$

Preferably, $\dfrac{\partial l}{\partial \alpha} = \dfrac{N}{\alpha} + \sum_{i=1}^{N}(c_i - \mu_i)\log(y_i) = 0$ is solved after each M step so as to provide an updated value of α. Following the steps applied for Cox's proportional hazards model, one may estimate α and select a parsimonious subset of parameters from $\underset{\sim}{\beta}$ that can provide an adequate explanation for the survival times if the survival times follow a Weibull distribution.

Embodiments of the method of the invention are now illustrated with reference to Figures 1 to 3.

Figure 1 illustrates an embodiment of the method of the invention in which data sets containing information about characteristics of a compound are generated from an enzyme kinetic analysis, an Infra-red spectrum analysis and a DNA microarray system. In this embodiment, data sets from the IR-spectrum analysis and enzyme kinetic analysis are combined and processed to provide a signature that is indicative that the sample will produce data in a microarray (of DNA) that comprises a signature that is indicative of low toxicity in cells. The method involves obtaining IR-spectrum data, enzyme kinetic data and micoarray data (for DNA) for a plurality of training samples, wherein at least one of the training samples is known to be low in toxicity and produce a signature in a microarray system that is indicative that the sample is low in toxicity. A multivariate algorithm is selected based on the distribution of the data for which the signature is to be indicative. In the case of microarray data, this may be a multiclass distribution as the results from the array could be classified into groups on the basis of the pattern of gene expression that is exhibited by cells in response to the training samples. Consequently the multivariate algorithm described in A above could be applied. Using A above, a signature may be obtained from processing the combined data from the IR-spectrum and the enzyme kinetics analysis which is indicative that a training sample (which is known to be capable of producing the signature in a microarray (of DNA) analysis and is known to be low toxicity) produces

data in the DNA microarray which includes the signature. This identified signature corresponds to a subset of components of the IR spectrum and the enzyme knietic data, and consequently in screenings of test samples, it is only necessary to measure the subset of components of the IR spectrum and enzyme kinetic data to determine whether a test sample has low toxicity.

In another embodiment, the method of the invention may be used to determine a signature by combining the data from distinct data sets before processing the data.

This embodiment of the invention is illustrated in Figure 2. Figure 2 illustrates a schematic representation of drug discovery using combined systems. The broad categories for the systems are illustrated at the top of the diagram as either theoretical chemical descriptors, physical chemical descriptors, bioscreen, gene expression, proteomic and ADMET. Vertical columns represent data generated from components for each system, with the actual systems being labelled along each column. Unlabelled columns represent data sets where the components of the characteristic are unknown. Ovals (1A-1H, 2A-2F and 3A-3C) represent one or more components which are part of the signature. Lines joining the subsets are indicative that that subset is part of that signature. The horizontal arrow labelled iPIPE indicates the direction of the pipeline.

Referring to figure 2, a signature that is indicative of a desired property in a sample may be identified by combining all the data sets generated from systems to which the sample has been applied. From all the data

sets, some will have data that comprises part of the signature. In figure 2, components from data generated using COMFA, Infra-red spectral analysis, enzyme kinetics, MPSS, cDNA array analysis, LC-MS fingerprint, LD50 analysis and mode of action analysis form a signature (the combination of 1A, 1B, 1C, 1D, 1E, 1F, 1G and 1H) that is indicative at 92% confidence limits that a sample will be successful in a drug discovery pipeline. A different signature (2A, 2B, 2C, 2D, 1E, 1F, 1G, 1H) may also have the same confidence limits. An alternative signature (1A-1E, 3A, 3B, 3C) may have reduced confidence limits.

Figure 3 illustrates a schematic diagram of the algorithm applied to N distinct data sets. Referring to Figure 3, a prior assumption is made about the data and each component of the data set is assigned a weight to produce weighted data. A function is defined that models the distribution of the condition, and the weighted data is fitted to the model by adjusting the weight values to maximise the model function. A threshold value for the weights is set and those weights which fall below the threshold in fitting the model are set to zero. The components have a weight again assigned (applying the assumption) and the weighted data is again fitted to the model by adjusting the weight values to maximise the model function. This process is repeated until the weights do not change for each cycle. Those components with zero weights are then eliminated to establish the signature.

In order to determine the performance of the embodiment of the present invention, the following experiment was simulated:

The simulated experiment corresponded to the screening of 2000 substances for potential use as drugs for colorectal cancer. As part of a simulated training phase, 1000 of the 2000 substances were selected and subjected to two screens. The first simulated screen was for efficacy against the cancer and the second simulated screen was for toxicity to normal cells. For the simulated efficacy screen, the substances were applied to cancer cell lines cultivated in the laboratory and the cell mortality measured as a continuous response y. In addition, gene expression measurements on 1000 genes were obtained for the cell lines treated with each of the substances. The gene expression measurements form a 1000 by 1000 matrix and the algorithm for a Gaussian generalised linear model was applied to see if the toxicity measurement could be predicted from the gene expression data alone. The algorithm was run and it selected a model of the form:

predicited toxicity measurement = 20.02 + 2.02*gene3 - 2.02*gene93 + 1.99*gene797

The residual mean squared error for this example was 0.24, so the fit is quite good.

For the simulated toxicity screen, the same 1000 substances were applied to healthy cells and cell mortality measured as a binary response (2 = non toxic, 1 = toxic). Gene expression measurements were also obtained for each of the substances applied to the normal cells. In addition spectroscopic measurements were obtained for each of the 1000 substances in the form of a spectrum at 476 wavelengths over the range 0.4 to 2.4 mircometers. The algorithm for the model corresponding to the equation was

applied to determine whether the binary toxicity variable was predictable on the basis of the microarray data and the spectroscopic data. The algorithm selected the model:

probability not toxic = exp(lp)/(1+exp(lp)

where

lp=2.52*gene55 -3.26*gene473 + 1.20*(spectrum value at channel 400)

This model was able to perfectly, discriminate between the toxic and non-toxic substances.

To screen the remaining 1000 substances, the above two models were applied. In particular, for the first screen it was only necessary to obtain gene expression measurements for 3 genes, namely gene3, gene93 and gene797 for each of the substances. Applying the model to this data and choosing only those substances with a high predicted efficacy, namely greater then 25, gave 200 substances which survived the first screen.

The second screen was applied to these 200 hundred substances by using the second model from the training data to predict toxicity. This only required expression values for gene55, gene473 and the spectrum value at channel 400 for each of the 200 hundred samples. Applying the model and choosing only those substances which are predicted to be non-toxic with a probability greater than 0.75, gave 18 substances, namely substances

12 71 95 169 171 302 323 463 491 587 623 664 716 792 815 838 892 969

Note that the screening on the second 1000 substances was much faster and required fewer resources due to the use of the models obtained in the training stage.

CLAIMS:

1. A method of determining a signature indicative of an entity having a condition, the method comprising the steps of:

(a) obtaining at least two distinct data sets, at least one of the data sets containing information on at least one characteristic of at least one entity having the condition; and

(b) processing the information from the data sets to obtain a signature that is indicative of an entity having the condition.

2. The method as claimed in claim 1, wherein the step of processing the information comprises using a multivariate analysis to obtain the signature.

3. The method as claimed in claim 2, wherein the step of using the multivariate analysis comprises the steps of:

(c) assigning a weight to each component of the data sets;

(d) defining a function of the data and the weights which models the relationship between the condition and the components;

(e) applying an algorithm to update the weights using the function in (d) so as to generate a processed weight for each component;

(f) generating a processed data set containing components having a processed weight with absolute value above a predetermined threshold.

4. The method as claimed in claim 3, wherein the step of assigning the weight comprises the step of making an

assumption about a prior distribution of the data.

5. The method as claimed in claim 3 or 4, wherein the step of using the multivariate analysis is iterative such that steps (e) and (f) are repeated using the processed data sets in a subsequent cycle of the multivariate analysis.

6. The method as claimed in claim 5, comprising the step of using the processed weights as a basis to establish the assigned weight for the subsequent cycle.

7. The method as claimed in claim 6, further comprising the steps of:

(g) assigning weights to the components of a said processed data set;
(h) applying the algorithm to fit the weighted data to the function so as to generate a revised processed weight for each component;
(i) generating a revised processed data set containing components having a revised processed weight above a predetermined threshold;
(j) repeating steps (g) to (i).

8. The method as claimed in claim 7, wherein steps (g) to (i) are repeated until the processed weights remain constant to within a set tolerance with each repeat of steps (g) to (i).

9. The method as claimed in claim 8, wherein the weighted data is in the form of a linear combination.

10. The method as claimed in any one of claims 3 to 9,

wherein the function is a mathematical model of a probability distribution of the condition.

11. The method as claimed in claim 10, wherein the probability distribution is a likelihood function.

12. The method as claimed in any one of claims 3 to 11, wherein the multivariate analysis estimates component weights utilising a Bayesian statistical method.

13. The method as claimed in any one of the preceding claims, wherein when the at least two distinct data sets comprises a large amount of data, the method as claimed in any one of the preceding claims comprises the step of making an a priori assumption that the majority of the data is unlikely to be data that will form part of the signature that is indicative of the condition.

14. The method as claimed in claim 11, wherein the likelihood function is based on a multinomial or binomial logistic regression which models a condition of the entity that exhibits a multinomial or binomial distribution.

15. The method as claimed in claim 14, wherein the likelihood function based on the logistic regression is of the form:

$$L=\prod_{i=1}^{n}\left(\prod_{g=1}^{G-1}\left\{\frac{e^{x_i^T\beta_g}}{\left(1+\sum_{g=1}^{G-1}e^{x_i^T\beta_g}\right)}\right\}^{e_{ik}}\left\{\frac{1}{1+\sum_{h=1}^{G-1}e^{x_i^T\beta_h}}\right\}^{e_{iG}}\right)$$

wherein

$x_i^T \beta_g$ is a linear combination generated from input data from training sample i with component weights β_g;

x_i^T is the components for the i^{th} Row of X and β_g is a set of component weights for sample class g; and

X is data from n training samples comprising p components.

16. The method as claimed in claim 14, wherein the likelihood function based on the categorical ordered logistic regression is of the form:

$$l = \prod_{i=1}^{N} \prod_{k=1}^{G-1} \left(\frac{\gamma_{ik}}{\gamma_{ik+1}} \right)^{r_{ik}} \left(\frac{\gamma_{ik+1} - \gamma_{ik}}{\gamma_{ik+1}} \right)^{r_{ik+1} - r_{ik}}$$

Wherein

γ_{ik} is the probability that data from training sample in the second system i belongs to a class with identifier less than or equal to k (where the total of ordered classes is G) and r_{ik} is defined as follows: let C be a n by p matrix with elements c_{ij} given by

$$c_{ij} = \begin{cases} 1, & \text{if observation i in class j} \\ 0, & \text{otherwise} \end{cases}$$

and let R be an n by P matrix with elements r_{ij} given by

$$r_{ij} = \sum_{g=1}^{j} c_{ig}$$

these are the cumulative sums of the columns of C within rows.

17. The method as claimed in claim 11, wherein the likelihood function is based on a generalised linear model of the form:

$$L = \log p(y \mid \beta, \phi) = \sum_{i=1}^{N} \{ \frac{y_i \theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi) \}$$

where $y = (y_1, \ldots, y_n)^T$ and $a_i(\phi) = \phi / w_i$ with the w_i being a fixed set of known weights and ϕ a single scale parameter.

18. A method of predicting a time to an event for an entity, the method comprising the step of using the method defined in claim 11, wherein the likelihood function is based on a proportional hazards model.

19. The method as claimed in claim 18, wherein the likelihood function is based on a model selected from the group consisting of a Cox's proportional hazards model, a parametric survival model and an accelerated failure times model.

20. The method as claimed in claim 19, comprising the step of identifying a signature capable of predicting a condition that involves the time to the event by defining the likelihood as being based on the Cox's proportional hazards model, the parametric survival model or the accelerated survival times model.

21. The method as claimed in any one of claims 18 to 20, wherein the likelihood function for predicting the time to the event is of the form:

$$Log\ (Partial)\ Likelihood = \sum_{i=1}^{N} g_i\left(\underset{\sim}{\beta}, \underset{\sim}{\varphi}; X, \underset{\sim}{y}, \underset{\sim}{c}\right)$$

where $\underset{\sim}{\beta}^T = \left(\beta_1, \beta_2, \cdots, \beta_p\right)$ and $\underset{\sim}{\varphi}^T = \left(\varphi_1, \varphi_2, \cdots, \varphi_q\right)$ are the model parameters.

22. The method as claimed in claim 19, wherein the likelihood function based on Cox's proportional hazards model is of the form:

$$L\left(\underset{\sim}{t}\,|\,\underset{\sim}{\beta}\right)=\prod_{j=1}^{N}\left(\frac{exp\left(Z_j\underset{\sim}{\beta}\right)}{\sum_{i\in\Re_j}exp\left(Z_i\underset{\sim}{\beta}\right)}\right)^{d_j}$$

where $\underset{\sim}{\beta}^T=\left(\beta_1,\beta_2,\cdots,\beta_n\right)$, Z_j = the j^{th} row of Z, and $\Re_j=\left\{i:i=j,j+1,\cdots,N\right\}$ = the risk set at the j^{th} ordered event time $t_{(j)}$.

23. The method as claimed in claim 19, wherein the likelihood function based on the Parametric Survival model is of the form:

$$L=\sum_{i=1}^{N}\left\{c_i\,log\left(\mu_i\right)-\mu_i+c_i\left(log\left(\frac{\lambda\left(y_i\right)}{\Lambda\left(y_i;\underset{\sim}{\varphi}\right)}\right)\right)\right\}$$

where $\mu_i=\Lambda\left(y_i;\underset{\sim}{\varphi}\right)exp\left(X_i\underset{\sim}{\beta}\right)$, $\lambda\left(y_i\right)$ is the hazard function and , $\Lambda\left(y_i;\underset{\sim}{\varphi}\right)$ is the integrated hazard function.

24. The method as claimed in claim 19, further comprising the step of estimating weights for the model by using a Bayesian statistical model in which a posterior distribution of the weights is formulated which combines the likelihood function and a prior distribution.

25. The method as claimed in claim 24, wherein the step of estimating the weights comprises the step of maximising

the posterior distribution of the weights given the data generated for each entity.

26. The method as claimed in claim 25, wherein the prior distribution is of the form:

$$p(\beta) = \int_{v^2} p\left(\beta \,|v^2\right) p\left(v^2\right) dv^2$$

wherein v is a p x 1 vector of hyperparameters, and where $p\left(\beta \,|v^2\right)$ is $N\left(0, \mathrm{diag}\left\{v^2\right\}\right)$ and $p\left(v^2\right)$ is some hyperprior distribution for v^2.

27. The method as claimed in claim 25, wherein the prior distribution comprises a hyperprior that ensures that zero weights are preferred whenever possible.

28. The method as claimed in claim 27, wherein the hyperprior is a Jeffrey's prior.

29. The method as claimed in claim 24, wherein the posterior distribution is of the form:

$$p(\beta \varphi v | y) \,\alpha\, L(y|\beta\varphi) p(\beta|v) p(v)$$

or

$$P\left(\underset{\sim}{\beta}, \underset{\sim}{\varphi}, \underset{\sim}{\tau} \middle| \underset{\sim}{y}\right) \alpha\, L\left(\underset{\sim}{y} \middle| \underset{\sim}{\beta}, \underset{\sim}{\varphi}\right) P\left(\underset{\sim}{\beta} \middle| \underset{\sim}{\tau}\right) P(\underset{\sim}{\tau})$$

wherein $L\left(\underset{\sim}{y} \middle| \underset{\sim}{\beta}, \underset{\sim}{\varphi}\right)$ is the likelihood function.

30. The method as claimed in claim 24, further comprising

the step of using an algorithm to fit the model to the data by adjusting the weights to maximise the probability density of the posterior distribution of the function.

31. The method as claimed in claim 30, wherein the algorithm is capable of fitting the data to the model by adjusting the weights, and which is based on an iterative procedure or an optimisation procedure.

32. The method as claimed in claim 31, further comprising the step of using the iterative procedure to estimate the weights in the posterior distribution.

33. The method as claimed in claim 31, further comprising the step of using an EM algorithm in the iterative procedure.

34. The method as claimed in any one of the preceding claims, wherein the data sets are either related or unrelated data sets, wherein the data sets may be related in that they contain information about the same characteristic of the entity but the information is obtained in different ways, and wherein the data sets may be unrelated in that each data set is obtained from different systems and the information of each data set is on a different characteristic.

35. The method as claimed in any one of the preceding claims, wherein the information in the data sets relates to components of a system.

36. A method for determining whether a test entity has a condition, the method comprising the steps of:

(a) determining a signature indicative of the condition using the method as defined in any one of claims 1 to 35;

(b) obtaining data about a test entity; and

(c) evaluating the signature for the test entity to determine whether the test entity has the condition.

37. A computer program, arranged, when run on a computing device, to control the computing device to process information from at least two distinct data sets, at least one of the data sets containing information on at least one characteristic of at least one entity having a condition, to obtain a signature that is indicative of the condition of an entity.

38. A computer readable medium comprising the computer program defined in claim 37.

39. An apparatus for processing information from at least two distinct data sets, at least one of the data sets containing information on at least one characteristic of at least one entity having a condition, to obtain a signature that is indicative of the condition of an entity.

40. A system which is arranged to identify data obtained from processing information from at least two distinct data sets, at least one of the data sets containing information on at least one characteristic of at least one entity having a condition, to obtain a signature that is indicative of the condition of an entity , the system comprising a user interface and a storage means arranged

to store a processing template, the template including pre-stored model parameters for controlling the system in accordance with the parameters and wherein a user is able to select the processing template via the interface and the system is arranged to carry out processing of the information from at least two characteristics in accordance with the model parameters.

41. A method of determining whether a substance has a potential for use as a drug, the method comprising the steps of:
obtaining a first mathematical model that is derived using a signature of a condition of an entity;
using the first mathematical model in conjunction with information about at least one characteristic of another entity that has been exposed to the substance; and
using a result from the first mathematical model to determine whether the substance has the potential for use as the drug.

42. The method as claimed in claim 41, wherein the signature is obtained using the method as defined in any one of claims 1 to 35.

43. The method as claimed in claim 41 or claim 42, further comprising the step of selecting the substance from a plurality of substances by using a second mathematical model in conjunction with other information about the plurality of substances.

44. The method as claimed in any one of claims 41 to 43, wherein the first mathematical model is such that the

result represents a probability that the substance is not toxic to the other entity.

45. The method as claimed in any one of claims 41 to 44, wherein the second mathematical model is such that it represents a toxicity measurement for the substance in relation to the other entity.

DATED this 17th day of March 2004
COMMONWEALTH SCIENTIFIC AND INDUSTRIAL RESEARCH ORGANISATION
By their Patent Attorneys
GRIFFITH HACK

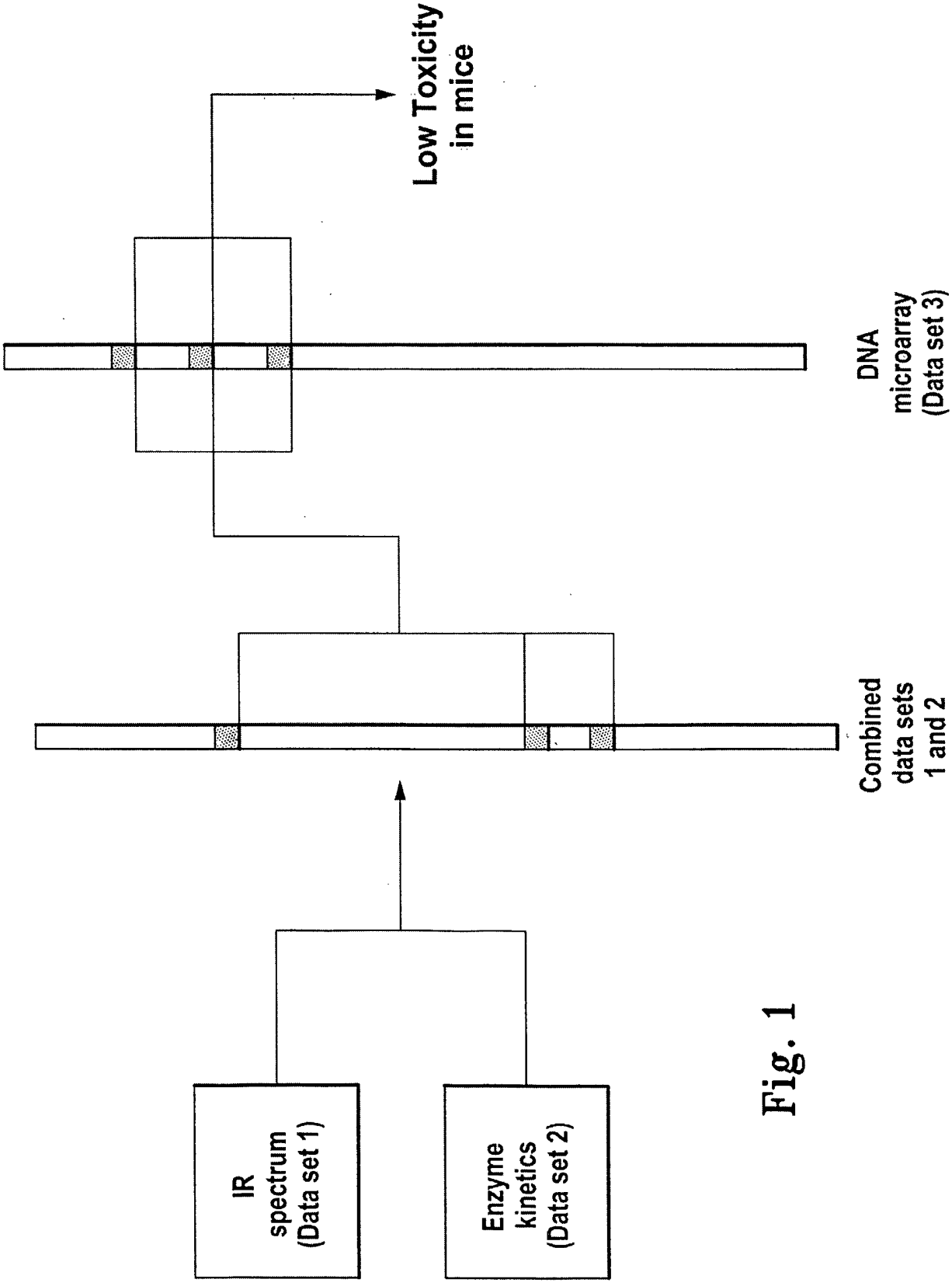

Low Toxicity in mice
DNA microarray (Data set 3)
Combined data sets 1 and 2
IR spectrum (Data set 1)
Enzyme kinetics (Data set 2)

Fig. 1

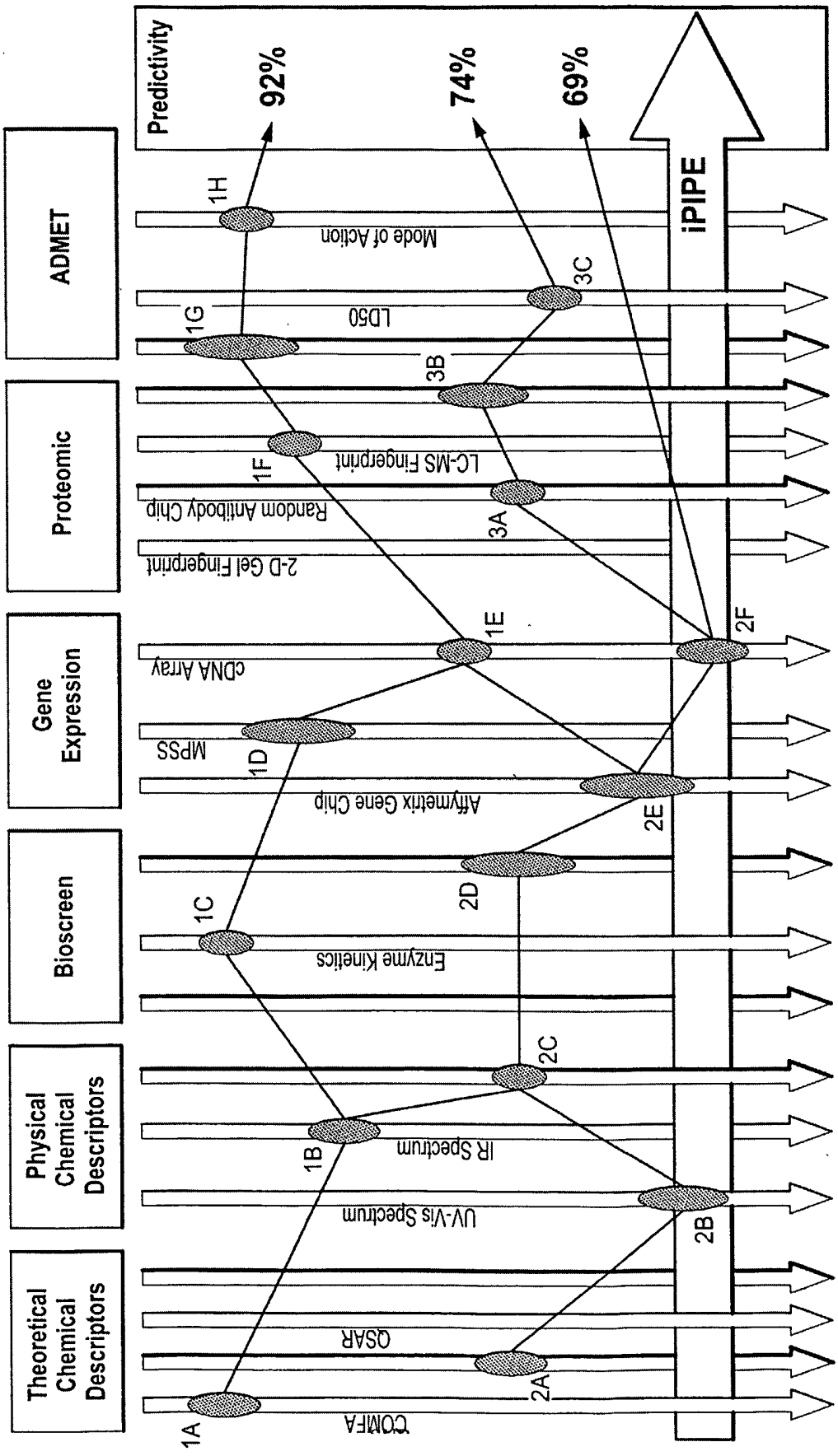

Predictivity
92%
74%
69%
iPIPE
ADMET
Proteomic
Gene Expression
Bioscreen
Physical Chemical Descriptors
Theoretical Chemical Descriptors
1H
Mode of Action
1G
LD50
1F
LC-MS Fingerprint
Random Antibody Chip
2-D Gel Fingerprint
cDNA Array
MPSS
Affymetrix Gene Chip
1C
Enzyme Kinetics
1B
IR Spectrum
UV-Vis Spectrum
QSAR
COMFA
1A
3C
3B
3A
1E
1D
2D
2E
2C
2B
2A
2F

Fig. 2

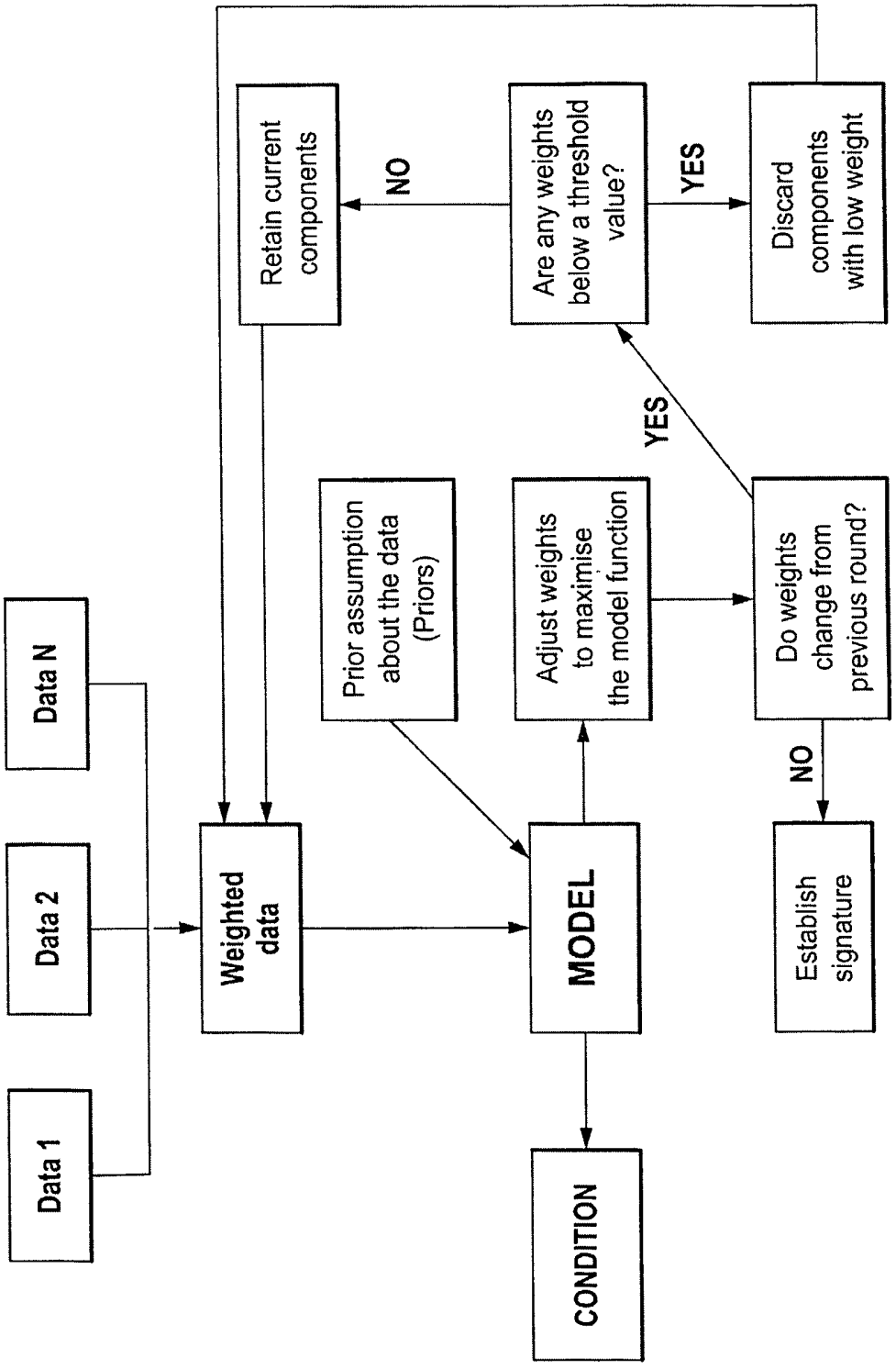

Data N
Data 2
Data 1
Weighted data
Prior assumption about the data (Priors)
MODEL
CONDITION
Adjust weights to maximise the model function
Do weights change from previous round?
YES
NO
Establish signature
Retain current components
NO
Are any weights below a threshold value?
YES
Discard components with low weight

Fig. 3

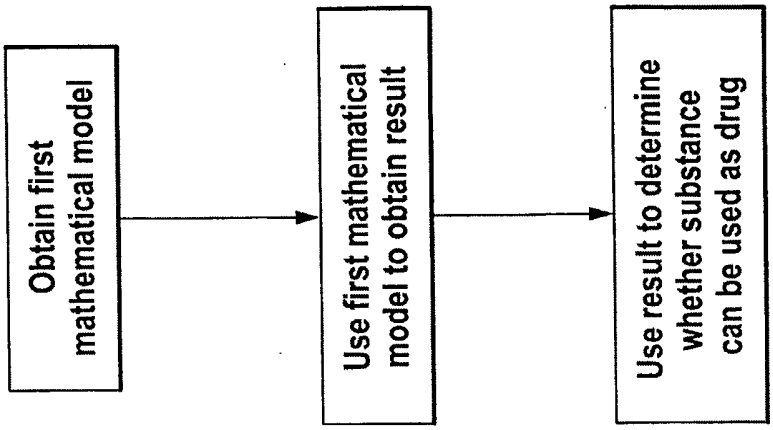


Fig. 6

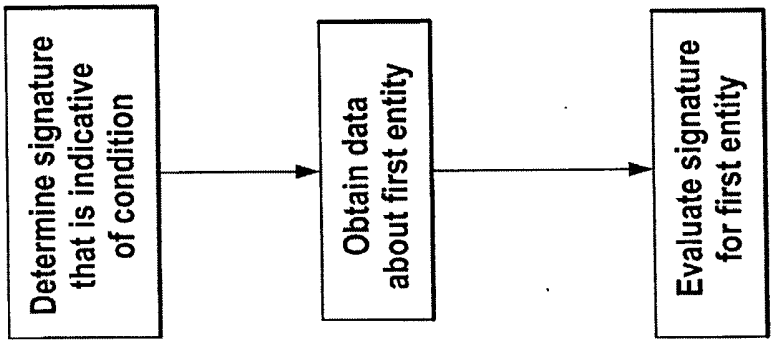


Fig. 5

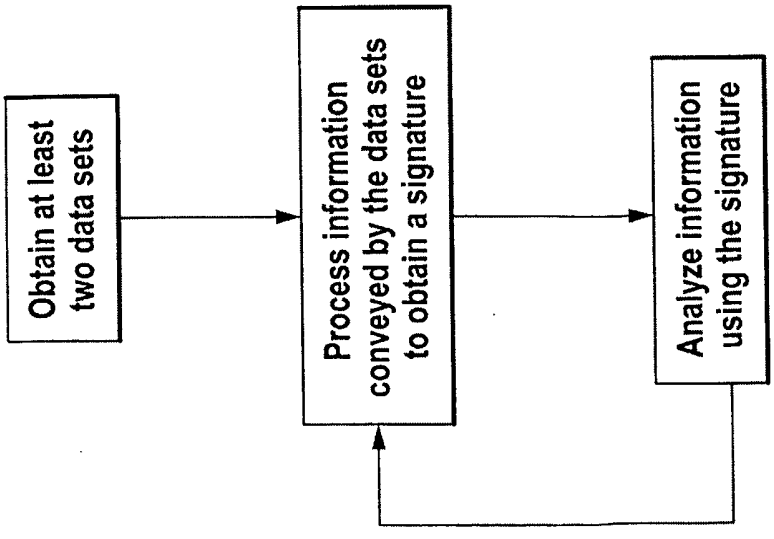


Fig. 4

INTERNATIONAL SEARCH REPORT

International application No.
PCT/AU2004/000327

A. CLASSIFICATION OF SUBJECT MATTER

Int. Cl. [7]: C12Q 1/00, G01N 33/48, G06F 19/00, 159:00

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
DWPI,MEDLINE (condition, signature, data sets, multivariate)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
	WO 2003/017177 A2 (BEYONG GENOMICS, INC.)	
X	Whole document, especially page 4 lines 4-15, page 8 lines 9-12, claim 19	1,2,34-40
Y	Page 9 line 22, page 13 lines 16-18, claim 18,41	41
	WO 2002/090927 A2 (AXIOM BIOTECHNOLOGIES, INC.) 14 November 2002	
X	Abstract, page 7 lines 17-31, page 19 lines 1-4, claims	41
Y	Whole document	1,2,34-40
	WO 2001/025473 A1 (SOURCE PRECISION MEDICINE, INC.) 12 April 2001	
X	Abstract, page 18, page 20 lines 13-31, page 31 line 32 to page 32 line 26	1,34,37-40
Y	Page 39 line 9 to page 40 line 12, claims	2,41
	WO 2000/070340 A2 (KAROLINSKA INNOVATIONS AB) 23 November 2000	
Y	Page 3 lines 7-16, page 13 lines 20-22, claims 11,12	2,36
A	Abstract, page 5 lines 12-23	1,34,37-40

[X] Further documents are listed in the continuation of Box C [X] See patent family annex

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
19 April 2004	22 APR 2004
Name and mailing address of the ISA/AU	Authorized officer
AUSTRALIAN PATENT OFFICE PO BOX 200, WODEN ACT 2606, AUSTRALIA E-mail address: pct@ipaustralia.gov.au Facsimile No. (02) 6285 3929	**DALE SIVER** Telephone No : (02) 6283 2196

INTERNATIONAL SEARCH REPORT

International application No.
PCT/AU2004/000327

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	WO 2002/090579 A1 (GENOMICS RESEARCH PARTNERS PTY LTD) 14 November 2002, Abstract, page 64	1,34,37-40
Y	WO 2001/016858 A2 (REALTIMEHEALTH.COM INC.) 8 March 2001 Abstract, claims	1,34,37-40

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No.
PCT/AU2004/000327

This Annex lists the known "A" publication level patent family members relating to the patent documents cited in the above-mentioned international search report. The Australian Patent Office is in no way liable for these particulars which are merely given for the purpose of information.

Patent Document Cited in Search Report		Patent Family Member					
WO	03017177	US	2003134304				
WO	02090927	US	2003100997				
WO	0125473	AU	58985/00	CA	2378008	EP	1198585
		US	2003229455				
WO	0070340	AU	49208/00	EP	1179175		
WO	02090579	CA	2446004	EP	1402062	NO	20034893
		US	2002187480				
WO	0116858	AU	69397/00	AU	70775/00	AU	78261/00
		US	2001037340	WO	0116857	WO	0116859

Due to data integration issues this family listing may not include 10 digit Australian applications filed since May 2001.

END OF ANNEX