Systems biology approaches and pathway tools for investigating cardiovascular disease†

Craig E. Wheelock,*[ab] Åsa M. Wheelock,[bcd] Shuichi Kawashima,[e] Diego Diez,[b] Minoru Kanehisa,[be] Marjan van Erk,[f] Robert Kleemann,[g] Jesper Z. Haeggström[a] and Susumu Goto[b]

Received 4th February 2009, Accepted 26th March 2009
First published as an Advance Article on the web 27th April 2009
DOI: 10.1039/b902356a

Systems biology aims to understand the nonlinear interactions of multiple biomolecular components that characterize a living organism. One important aspect of systems biology approaches is to identify the biological pathways or networks that connect the differing elements of a system, and examine how they evolve with temporal and environmental changes. The utility of this method becomes clear when applied to multifactorial diseases with complex etiologies, such as inflammatory-related diseases, herein exemplified by atherosclerosis. In this paper, the initial studies in this discipline are reviewed and examined within the context of the development of the field. In addition, several different software tools are briefly described and a novel application for the KEGG database suite called KegArray is presented. This tool is designed for mapping the results of high-throughput omics studies, including transcriptomics, proteomics and metabolomics data, onto interactive KEGG metabolic pathways. The utility of KegArray is demonstrated using a combined transcriptomics and lipidomics dataset from a published study designed to examine the potential of cholesterol in the diet to influence the inflammatory component in the development of atherosclerosis. These data were mapped onto the KEGG PATHWAY database, with a low cholesterol diet affecting 60 distinct biochemical pathways and a high cholesterol exposure affecting 76 biochemical pathways. A total of 77 pathways were differentially affected between low and high cholesterol diets. The KEGG pathways ''Biosynthesis of unsaturated fatty acids'' and ''Sphingolipid metabolism'' evidenced multiple changes in gene/lipid levels between low and high cholesterol treatment, and are discussed in detail. Taken together, this paper provides a brief introduction to systems biology and the applications of pathway mapping to the study of cardiovascular disease, as well as a summary of available tools. Current limitations and future visions of this emerging field are discussed, with the conclusion that combining knowledge from biological pathways and high-throughput omics data will move clinical medicine one step further to individualize medical diagnosis and treatment.

Introduction

An organism is an individual living system capable of reacting to stimuli, reproducing and maintaining a stable structure over time. Organisms are composed of multiple individual components, *e.g.* cells and their corresponding genes, proteins, metabolites, *etc.*, which are all governed by an intricate network of interactions. This network is not static, and the various components evolve and adapt dynamically to internal and environmental changes. The study of this complex system as a single entity is a challenge that has been traditionally addressed by studying different components of the system in isolation. Although such approaches have produced a significant amount of knowledge and understanding, they are limited in their ability to predict the effects of alterations in single or multiple components upon the dynamics of the whole system. This limitation may reflect why in some cases, significant research advances do not translate, for example,

[a] *Department of Medical Biochemistry and Biophysics, Division of Physiological Chemistry II, Karolinska Institutet, S-171 77, Stockholm, Sweden. E-mail: craig.wheelock@ki.se; Fax: +46-8-736-0439; Tel: +46-8-5248-7630*
[b] *Bioinformatics Center, Institute for Chemical Research, Kyoto University, Uji, Kyoto, 611-0011, Japan*
[c] *Lung Research Lab L4:01, Respiratory Medicine Unit, Department of Medicine, Karolinska Institutet, 171 76, Stockholm, Sweden*
[d] *Karolinska Biomics Center Z5:02, Karolinska University Hospital, 171 76, Stockholm, Sweden*
[e] *Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo, 108-8639, Tokyo*
[f] *Department of Physiological Genomics, TNO-Quality of Life, BioSciences, Utrechtseweg 48, 3704 HE, Zeist, The Netherlands*
[g] *Department of Vascular and Metabolic Disease, TNO-Quality of Life, BioSciences, Gaubius Laboratory, Zernikedreef 9, 2333 CK, Leiden, The Netherlands*

† Electronic supplementary information (ESI) available: Complete list of all KEGG biochemical pathways identified by KegArray as being affected by low cholesterol treatment, high cholesterol treatment, and differentially affected between low and high cholesterol treatment. See DOI: 10.1039/b902356a

into improved therapeutics or a "cure" for the disease under study. The discipline of systems biology attempts to shift the way in which an organism is perceived to address the complexity of living systems. Multiple definitions for systems biology exist, one of which describes it as a new field of study that aims to understand the living cell as a complete system.[1,2] In other words, systems biology seeks to understand how system properties emerge from the nonlinear interactions of multiple components.[3,4]

The applications of systems biology approaches are increasing dramatically; however, the exact nature of what a "systems approach" entails remains diffuse in the literature. The fundamental theme of systems biology is integration across multiple high output platforms, systems and disciplines.[5] However, it should be noted that systems science is not novel and has been advocated for many years in a number of research fields. At the simplest level, a systems approach signifies a study based upon examining the entire "system" simultaneously, as opposed to a reductionist approach that focuses on a single gene, metabolite, pathway, *etc*. In other words, a systems biology approach does not focus on identifying a single target or mechanism for an observed phenotype (*e.g*. disease). Systems biology instead seeks to identify the biological networks or pathways that connect the differing elements of a system, and in the process describe the characteristics that define a shift in equilibrium, such as metabolic fluxes or altered protein activities, which may cause a shift from a healthy to a diseased state. The hypothesis then becomes that those components of the network that are associated with the observed shift are characteristic and potentially descriptive of the disease, and accordingly represent potential targets for intervention to return the system to its original state (*i.e*. a healthy state). However, it is important to realize that the concept of equilibrium may not be as static as previously thought. It is more likely that equilibrium is a steady state that represents a range of fluctuations in the biological network that varies on an interindividual basis. The normal or control state is more appropriately categorized as one of dynamic stability in which our concept of homeostasis is more correctly defined as homeodynamics.[3] Accordingly, by defining the parameters of the network that determine disease from healthy state, interventions or treatments can be derived that are tailored for the individual variability of the parameters for this steady state—in other words, personalized medicine.

The era of personalized medicine has been heralded for a number of years, and systems biology is a key component of this new paradigm.[6–8] The intent is to identify disease before pathogenic manifestation, thereby initiating therapeutic intervention prior to significant adverse effects. Current medical practice is a reductionist approach that involves treating each problem or symptom in isolation. By these standards, the relief of symptoms as determined by clinical evaluations following a treatment regimen embodies the definition of a cured or maintained patient. A corresponding "limited" systems biology approach, where a multitude of clinical and biochemical variables are combined with multivariate statistical analyses often reveals that the patient indeed has been removed from the disease group following treatment, but not necessarily back towards a healthy state as is often assumed. Instead, the treated patient belongs to a novel biological status, distinctly different from both healthy individuals and peers in the disease group. This novel pharmacological state is generally not discernable in classical medicine, as the patient per definition is classified as belonging to the "healthy" group as soon as the symptoms that define the disease are no longer detectable. More importantly, the classical reductionist approach does not reveal the novel pharmaceutical state that the treatment regimen has induced, and consequently implications on the patient's future health cannot be predicted. In contrast, a true systems biology approach offers the ability to distinguish between multiple disease, healthy, or pharmacological states, as well as causative and adaptive responses and variables. However, in

Associate Professor Craig E. Wheelock heads a research group at the Karolinska Institute that examines the role of bioactive lipid mediators in inflammatory diseases, with a focus on cardiovascular disease. He is broadly interested in the development of bioinformatics tools for probing inflammatory diseases at the systems level.
Assistant Professor Åsa M. Wheelock heads a research group at the Karolinska Institute that investigates pneumotoxicants and inflammatory lung diseases, as well as gel-based quantitative proteomics.
Assistant professor Shuichi Kawashima is a researcher at the Human Genome Center in the Institute of Medical Science at the University of Tokyo who is broadly interested in the development of genome databases, bioinformatics web services and the biology of eukaryotic genomes.
Dr. Diego Diez is a postdoctoral researcher at the Kyoto University Bioinformatics Center working on applying systems biology approaches to cardiovascular disease.
Professor Minoru Kanehisa is the Director of the Bioinformatics Center in the Institute for Chemical Research at Kyoto University and a professor at the Human Genome Center in the Institute of Medical Science at the University of Tokyo. His research involves deciphering systemic biological functions by integrated analysis of genomic and chemical information.
Dr. Marjan van Erk is a researcher at TNO Quality of Life who is interested in developing bioinformatical systems biology tools for metabolic and cardiovascular diseases.
Dr. Robert Kleemann heads a research unit at TNO Quality of Life that investigates the role of inflammation in cardiovascular disease and metabolic disorders and has particular interest in gene regulation and drug intervention.
Professor Jesper Z. Haeggström heads a research group at the Karolinska Institute that examines the role of bioactive lipid mediators in inflammatory disease.
Associate Professor Susumu Goto is interested in the development of databases for molecular interaction networks and network analysis using the KEGG database suite. His work also involves in silico *metabolic reconstruction.*

order to make conclusions regarding causative relationships, it is necessary to have a sufficient number of variables and observations. In addition, the quantitative quality and source of the data, as well as the choice of multivariate statistical tools both in the experimental design and the post-experimental analyses, are vital for interpretation.

The increase in systems biology applications is a reflection of a "perfect storm" of advances in analytical methodology, computing power and data acquisition. The completion of the human genome sequencing project heralded the age of large-scale biology and data acquisition. This paradigm shift coupled to commensurate developments in technology and experimental techniques that can simultaneously interrogate many elements of a system (*i.e.*, microarrays, mass spectrometry, computational power and the Internet) has led to a veritable explosion in "omics" science and systems biology related research. The challenge for systems biology is to integrate the disparate disciplines of biology, chemistry, statistics, computer science and engineering into a cohesive science. Towards this end, it is necessary to develop common platforms for the analysis, presentation and archiving of data to ensure inter-laboratory and cross-disciplinary compatibility and accessibility of data sets. Significant steps have already been taken in this direction, and it is not our aim to review the status of the technological platforms or compatibility of data formats, as these aspects have been covered in detail elsewhere.[9–17] In contrast, this review focuses on the integration of different types of data sets, and aims to summarize the current state of systems biology research into cardiovascular disease as well as present a number of different pathway mapping tools that have been developed. In addition, an example of a pathway analysis of atherosclerosis is presented using a novel tool for mapping of omics data to the KEGG database suite.

Networks in a nutshell

One of the recurrent concepts in system biology is that of the network. Much of the early work in networks focused on simple model organisms including bacteria, yeast and nematodes;[18–24] however, this work is expanding to the understanding of human diseases.[25–28] A network type of representation formalizes the interaction of different components of a system utilizing the infrastructure of a branch of mathematics called graph theory. In the network paradigm, nodes represent elements of the system while relations are symbolized by edges. For example, in a metabolic network, enzymes and compounds are nodes, and reactions are edges. In a protein–protein interaction network, two nodes connected by an edge represent interacting proteins. This formalism enables the study of living systems in a way never thought possible before. The individual elements are integrated in a network whose properties can be analyzed globally: the number of edges per node, the degree distribution (the probability that a node has a specific number of edges), the cluster coefficient, *etc*. Barabasi and Oltvai have reviewed these concepts in detail, and provided a comprehensive review of the terminology and concepts associated with network analysis.[2] This new terminology is increasingly prevalent in the biological literature, requiring the life scientist to become familiar with this research field. These technical properties provide information regarding the global behavior of the network and therefore of the biological system under study. For example, one important finding was the scale-free topology nature of biological networks. In this type of network, most nodes have few links, whereas a few nodes have many links (called hubs or nexus nodes). One of the translations of this characteristic into a biological context is the hypothesis that hub nodes perform key functions in the network. Accordingly, many fundamental genes, proteins, enzymes and compounds have been identified as hubs in their respective biological networks. Another consequence derived from this finding is that because of the sparse nature of scale-free networks (*i.e.* most nodes having a few edges), they are very robust to environmental alterations.

However, although network analysis can help us understand the behavior of the system as a whole, the importance of individual elements is not lost in this global view. For example, the study of biological networks shows that complex networks are constructed of recurrent simple motifs.[29] Initially described in simple bacteria, these motifs are also found in the regulatory networks of higher eukaryotes and are fundamental to understanding the behavior of complex networks, including biological networks. Moreover, the mathematical models used to generate the network itself can be used to predict the behavior of the network when specific elements are altered. For example, what are the effects if a specific node of a gene regulatory network is removed by a knockout mutation? How does this change affect the global stability and robustness of the network, and eventually, the phenotype of the studied system? Systems biology seeks to answer these and other questions by modeling the relationship between the components.[30]

One critical step is how the network is constructed from the raw data (transcriptomics, proteomics, metabolomics, *etc*.). This is accomplished by using different mathematical techniques, ranging from simple Pearson correlations to the use of ordinary differential equations, Boolean networks, *etc*. (reviewed in refs. 31 and 32). Through this modeling, fundamental concepts in the understanding of biological systems, like robustness, modularity, emergence, *etc*. are incorporated. Unfortunately not of all these questions are easily answered, even within the context of the systems biology paradigm. Whereas most studies currently focus on individual networks (*i.e.* a transcription network or a protein–protein interaction network), in reality these different networks function as a connected system. Therefore, a change in the gene regulatory network may have a corresponding effect in the protein–protein interaction network, the metabolic network, *etc*., which collectively may manifest changes in the observed phenotype. To understand the whole system, it is critical to integrate knowledge from different studies. However, the crosstalk between different networks is not yet well understood and although some progress has been made,[33,34] the integration of different types of data is still in its infancy.[12] Through the generation of mathematical models that integrate different types of data (*e.g.* transcriptomic, metabolomic, and protein–protein interactions),[2] we can explain the observed phenotype, and hopefully make predictions regarding how the

phenotype is altered when the network itself is modified through the alteration of internal or environmental factors.

Data processing and statistical analysis

The pre-processing of data is crucial in network applications, as well as other systems level analyses. It is important to recognize that the nature of large scale omics data is very different from that of reductionist approaches, and other statistical methods should be utilized. The majority of the univariate methods that have dominated biological sciences for centuries (*e.g.* Student's t-test) are not well-suited for a number of reasons. For example, univariate statistical methods employ repeated testing to evaluate whether the null hypothesis for a certain variable can be rejected, *i.e.* if it is significantly altered compared to the control group. Given the cumulative nature of the error in repeated testing, these methods are prone to high false positive rates, which become particularly pronounced in omics analyses where a large number of variables are tested simultaneously. Even though a range of approaches have been developed to correct for the resulting large false positive rates, most notably Bonferroni[35,36] and false discovery rate (FDR) corrections,[37] the use of univariate methods remains a compromise. The fact that univariate methods are very sensitive to missing data points further decreases the robustness of network analyses based solely on traditional statistical pre-processing of the data.

Multivariate analysis (MVA) is a more suitable option for these "short and fat" data sets that are typical for omics studies (*i.e.* a large number of variables with few observations). Instead of repeated testing of single variables, MVA aims to create a model that reduces the complexity of multi-dimensional data to a few latent variables that express the majority of the variance of the data set. Exemplified by principal component analysis (PCA), the most utilized unsupervised method in omics applications, the model is structured so that the first principal component (PC1) is oriented so that it describes the largest possible portion of the variance in the data set that can be described by a linear vector. Accordingly, each subsequent PC contains a smaller portion of the variance in the data set than the previous component. Given that the MVA is based on all individual variable data points for all observations, the resulting model is robust both against false positives and missing data points. Furthermore, a confidence interval representing all of the variables is obtained, in contrast to univariate methods where each variable is analyzed as a separate unit, and consequently only confidence intervals for individual variables can be obtained. MVA can also be utilized to perform regression analysis between large data sets, most commonly through partial least squares between latent structures (PLS). These types of analyses are referred to as supervised methods, since the user defines which variables belong to the X dataset (dictating variables) and which belong to the Y dataset (response variables).

While useful, multivariate statistical methods are not without their own weaknesses. A major pitfall in MVA relates to overfitting of the model to the data. If a sufficient number of components is utilized, it is possible to build a model that can describe any data set with a perfect correlation (*i.e.* $R^2 = 1.0$; Fig. 1). A comparison of the correlation coefficient to the predictive power of the model is therefore essential. The predictive power (Q^2) can be calculated through the use of a training set and a test set, or if the data set is too small to allow this, through a cross-validation approach. A good rule of thumb is to remove all components that do not contribute to an increased predictive power of the model. If the data set is sufficiently large, Q^2 can be used as a measure to evaluate the robustness of the model in relation to the whole population.

Another concern when utilizing MVA is that of strong outliers. One should be cautious of any observation that is located on either end of the axis of the first component (strong outliers), as it is likely that characteristics that are unique for this individual are influencing the entire model. Interpretability represents another concern in MVA. MVA summarizes the entire data set in a few latent variables, which cannot be directly connected to the original measured variables. As such, it can be difficult for the untrained eye to interpret which variables are important or "significant" in driving the separation of the different study groups. This becomes particularly pronounced in more complex analyses such as PLS. A recent addition to this group of analysis, orthogonal PLS (OPLS), greatly simplifies the interpretability by separating the variance in the data set according to the correlation to the selected Y matrix (*e.g.* disease group).[38] In contrast, the "orthogonal" component pulls out the variance that is not correlated to the Y-variables of interest, and thus represents internal variance in the X-matrix. While this approach is well-suited for motivating variable selection, it should be used cautiously in this aspect, given that the back-drop of the method is a supervised selection of the Y-variables that determine the separation. When in doubt, it is generally better to include all of the variables in subsequent

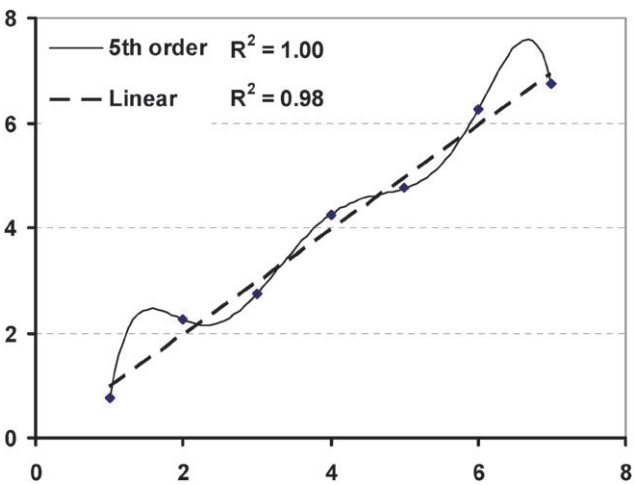

Fig. 1 Overfitting of data represents one of the main pitfalls associated with multivariate analyses. With a sufficient number of components, a model that explains 100% of the variance ($R^2 = 1.0$) can be built for any data set. In the above example, the simplest (linear) model represents the most representative model for the data, demonstrating that the simplest model provides optimal prediction, even though the correlation coefficient is lower.

analyses. Taken together, this section emphasized the point that it is vital to employ the correct statistical analysis in both experimental design as well as data processing. These approaches require an in-depth knowledge of MVA in order to correctly interpret the output of statistical models, prevent overfitting of the data, apply multitest corrections, and achieve an appropriate balance of false positives and power.

Systems biology in cardiovascular disease

The utility of systems biology becomes clear when applied to multifactorial diseases whose etiology is complex. For example, the etiology of inflammatory diseases such as atherosclerosis and asthma has proven recalcitrant to elucidation with reductionist approaches. It is possible that part of the difficulty in identifying new therapeutics lies in the inability of current approaches to visualize the complexity of these biological systems.[39] The development of lead drug candidates would also benefit from a systems approach. For example, drugs such as torcetrapib, statin + ezetimide and rimonabant have been withdrawn from the market because of side effects that were not predicted with reductionistic thinking. Diseases and disorders such as cardiovascular disease, diabetes, metabolic syndrome, asthma and chronic obstructed pulmonary disorder (COPD) all involve complicated developments that resist efforts to identify a single gene or pathway responsible for disease onset and progression. Numerous therapeutics have been successfully developed that intervene in different stages of the disease; however, we are still far from developing a true cure for any of these pathologies. The cellular complexity of many of the affected organs represents a major obstacle in the elucidation of the systems biology behind these pathologies. The lung, for example, consists of more than 40 different cell phenotypes, all of which may elicit different responses to up- or down-regulation of a certain factor. Add to that the spatial and temporal aspects of the cellular response, and we are starting to approach the true complexity of biological systems. Accordingly, while beyond the scope of this review, sampling design and strategy can have significant effects upon experimental observations. Given the heterogeneity of many tissue types, it is challenging to reproducibly sample tissue in such a way as to enable intra- and interlab comparisons. The obstacles involved in this area are not trivial and need to be addressed by the research community.

Cardiovascular disease is the major cause of premature death in Europe, resulting in >4 million deaths in the year 2000.[40] In the United States, cardiovascular disease was responsible for one of every five deaths in 2004, with an average of one death every 37 seconds.[41] The rapidly increasing incidence of obesity and commensurate health effects including atherosclerosis, metabolic syndrome and diabetes is of epidemic proportions, with the potential for significant increases in developing countries. It is anticipated that the "BRIC" countries (Brazil, Russia, India and China) will significantly contribute to the global cardiovascular disease burden such that by 2020 an additional $\sim$4% of deaths in the world will be due to ischemic heart disease.[42] The complexities associated with cardiovascular disease and metabolic syndrome are recalcitrant to current interventions and challenge the ability of the pharmaceutical industry to produce effective and inexpensive therapies. For example, in cardiovascular disease, each known risk factor is addressed individually, whether it be hyperlipidemia or hypertension.[3] However, given the complex etiology of this disease, it is likely that multiple factors are responsible for the observed pathology, resulting in a need for holistic treatment approaches that address the underlying problems. Accordingly, these diseases are logical targets for systems biology approaches to understanding disease mechanism, progression and pathogenesis.

Cardiovascular disease is multi-genic, multi-factorial and linked to other systemic disorders,[43,44] and the role of inflammation in the development of atherosclerosis and cardiovascular disease is firmly established.[45,46] The onset and development of cardiovascular disease has been shown to involve multiple factors including lifestyle, diet, body mass index, (epi)genetics, dyslipidemia, hypertension, and inflammation among others. However, the current paradigm of patient treatment involves addressing these individual risk factors in isolation, even though they are known to concomitantly contribute to disease pathogenesis. While effective in many cases, this approach has not provided a cure or even a full understanding of the disease, which remains a major source of mortality and morbidity worldwide.

A number of studies have begun to address the issues outlined above in a comprehensive fashion, and active research is being performed to develop systems biology approaches to cardiovascular disease.[47] We present a few of these studies in chronological order, but stress that this list is not comprehensive. Many of the early studies that performed systems biology-related investigations into cardiovascular disease focused on a single omics profiling method (*i.e.*, transcriptomics or metabolomics) and then included clinical parameters using multivariate statistics to develop models of disease. It is only recently that unifying systems biology models employing multiple analytical platforms linked with bioinformatics analyses have been produced. One of the earliest attempts to bring systems biology to cardiovascular function involved mapping important cardiovascular phenotypes onto the human genome. Stoll *et al.* studied 239 cardiovascular and renal phenotypes in 113 male rats. They identified and mapped a total of 81 cardiovascular phenotypes from an F_2 intercross onto the human genome using correlation patterns ("physiological profiles") and comparative genomics.[25] The resulting genomic-systems biology map was applicable for gene hunting and mechanism-based physiological studies of cardiovascular function. For example, the authors presented a correlation matrix with phenotypic ordering of 125 likely determinants of arterial blood pressure, which could be used to assess the impact of allelic substitutions on each of the traits in either the parental or F_2 generation of the intercross. The phenotypes were grouped into functionally related clusters (vascular, heart, renal, endocrine and morphometric) that impact on the control of blood pressure, and ordered within the clusters by known physiological relationships. All of the results of the linkage analyses and the phenotypic physiological profiles for each

microsatellite marker on the linkage map sorted by genotype can be accessed on the project homepage (http://brc.mcw.edu/phyprf/). A more diagnostic application was presented by Brindle *et al*. who employed a supervised partial least squares discriminant analysis (PLS-DA) approach to analyze ^{1}H NMR spectra of human serum to diagnose the presence, as well as the severity of coronary heart disease.[48] The PLS-DA model predicted the presence of coronary heart disease with a sensitivity of 92% and a specificity of 93% based on a 99% confidence limit. The major driving factor for the observed separation in severe coronary heart disease patients (triple vessel disease, TVD) was the presence of lipids, particularly LDL and VLDL, whereas the most influential loadings for the angiographically normal coronary arteries (NVA) were HDL-associated (*e.g*., fatty acid chains and phosphotidylcholine). Of particular importance is the fact that the authors confirmed that the method was able to diagnose coronary heart disease independently of the inevitable associated gender bias. However, work by Kirschenlohr *et al*. concluded that plasma-based ^{1}H NMR analysis is a weak predictor of coronary heart disease.[49] They found that the predictive power was significantly weaker, with NVA and coronary heart disease groups identified 80.3% correctly for patients not receiving statin therapy and 61.3% for patients treated with statins. The main reason postulated for the observed study discrepancy was the inclusion of additional variables in the Kirschenlohr *et al*. study, including drug treatment regimen. Statins significantly affect LDL levels, which was a discriminating factor in the PLS-DA model. Accordingly, as the most significant loadings associated with diagnosis in both studies were related to lipid species, it is not surprising that treatments affecting lipid levels influenced the observed separation power of the model. In other words, statin treatment partially resolves the incidence of coronary artery disease, thus reducing the biomarker signal in these patients. It would be interesting to further examine these patients to determine if they were truly moving towards a ''healthy'' phenotype or were instead representative of a third pharmacological state as discussed above. This point demonstrates one of the main challenges in developing diagnostic markers of complex disease in that in many cases patients will present distinct genotypes as well as personal therapeutic treatment regimens that can potentially confound the use of biomarkers, as reported by Brindle *et al*. At the very least, these studies demonstrate the importance of including as much patient metadata in the analyses as possible. The work of both groups supports further research into exploring the potential of applying metabolomics methods to identify plasma (*i.e*., non-invasive) biomarkers of coronary heart disease. It is possible that biomarkers could be identified in a study with increased cohort size composed of the myriad of clinical and interindividual variables. An important aspect of these metabolomic analyses is that in order to correctly classify individuals with coronary heart disease, it is not necessary to fully understand the complex molecular differences that underlie disease etiology.[48] This methodology is an important first step towards being able to identify individuals at risk of disease development or in the early stages of disease onset.

While useful for identifying potential markers of disease, the previous studies do not represent a systems methodology. One of the first comprehensive systems biology approaches involving the integration of multiple omics platforms (transcriptomics, proteomics and metabolomics) examined changes in the apolipoprotein E3-Leiden transgenic (ApoE*3Leiden) mouse model (a commonly used model of atherosclerosis[50]). The authors integrated gene transcripts, and protein and lipid data along with their putative relationships to gain insight into the early onset of disease.[51,52] As is common with many systems approaches, the authors developed a number of their methods for data processing and network analysis in-house, demonstrating a significant obstacle in the advance of systems biology. It is challenging to integrate bioanalytical results from multiple platforms and between different research groups, making it difficult to standardize results.[12] The ApoE knockout mouse was used in another investigation into atherosclerosis mechanisms involving conjugated linoleic acids (CLAs) to determine how individual CLA isomers differently affected pathways involved in atherosclerosis.[53] ApoE knockout mice were fed a diet supplemented with 1% *cis*9, *trans*11-CLA, 1% *trans*10, *cis*12-CLA or 1% linoleic acid for twelve weeks. The effects upon lipid and glucose metabolism were measured, as well as the regulation of hepatic proteins. Correlation analysis between physiological and protein data identified two clusters associated with glucose metabolism. The results showed that *cis*9, *trans*11-CLA specifically increased expression of the anti-inflammatory HSP 70, as well as decreased expression of the pro-inflammatory macrophage migration inhibitory factor, suggesting that consumption of *cis*9, *trans*11-CLA could protect against the development of atherosclerosis.

A systems biology approach to elucidating biological pathways in coronary atherosclerosis was published by King *et al*. who performed custom microarray analysis of coronary artery segments.[54] A number of clinical variables were examined, and diabetic states provided the most interesting results, with 653 upregulated genes in the no diabetes class and 37 upregulated genes in the diabetes class, with an FDR of 0.08%. The top gene upregulated in the diabetes class was IGF-1, followed by the IL-1 receptor and IL-2 receptor-α, indicating that there were changes in cytokine-induced immune and inflammatory responses. These results suggest that inflammation is more prominent in diabetic than nondiabetic coronary artery disease. Significant gene expression profiles were then used to construct a novel pathway based upon gene connectivity as determined by language parsing of the published literature, and ranking as determined by the significance of differentially regulated genes in the network. The resulting gene subnets were visualized with Cytoscape, an open-source bioinformatics resource (discussed in more detail below[55]), to identify nexus genes in disease severity. Results indicated that the key process in the progression of atherosclerosis relates to smooth muscle cell dedifferentiation, suggesting a focus on changes in the smooth muscle phenotype as a target for atherosclerosis. The results also provided insight into the severe form of coronary artery disease associated with diabetes, reporting an overabundance of immune and inflammatory signals in diabetics. This method

for querying multiple search engines and/or databases combined with parsing of the retrieved results (documents) for biological associations is extremely powerful for generating networks, and is used extensively in multiple software applications for network generation.

Lipopolysaccharide (LPS) is a critical inducer of sepsis, which is characterized by systemic inflammation, hypotension and multiple organ failure.[56] Tseng *et al.*[57] examined the molecular effects of late-phase LPS stimulation on primary rat endothelial cells in an attempt to develop diagnostic markers of inflammatory disease. A combination of cDNA microarray, 2-DE and MALDI-TOF MS/MS, as well as cytokine protein arrays were analyzed using custom bioinformatics applications. Differentially expressed genes and proteins were mapped onto their corresponding biological pathways using BioCarta or KEGG, and the results were ordered using the BGSSJ software (bulk gene search system for Java) followed by analysis with ArrayXPath.[58] The results showed significant effects ($p < 0.05$) on the BioCarta pathways ''LDL pathway during atherogenesis'', ''MSP/RON receptor signaling pathway'' (MSP, macrophage-stimulating protein; RON, tyrosine kinase/receptor d'origine nantais), ''signal transduction through IL-1R'', and ''IL-5 signaling pathway'', demonstrating that inflammatory pathways were significantly affected by LPS treatment, as would be expected. Overall, this study used a systems biology approach to show that NF-κB-associated responses in endothelial cells affected pathways involved in proliferation, atherogenesis, inflammation and apoptosis, thereby providing information on multiple pathways simultaneously. However, it should be stressed that it is necessary to differentiate protein concentrations from protein activities in order to make meaningful deductions. Several studies using ''focused'' arrays to analyze gene expression in human atherosclerotic tissue have confirmed that short-term LPS exposure results in vivid upregulation of a spectrum of proinflammatory genes including IL-1b, IL-15, interferon-induced genes, and a series of TNF superfamily members.[59–62]

Statins are an important therapeutic in the control of hyperlipidemia, with demonstrated efficacy in lowering cholesterol levels. However, there are concerns regarding the development of statin-induced myopathy following aggressive treatment. Laaksonen *et al.* employed a systems biology approach to probe the cellular mechanisms leading to myopathy and identify potential biomarkers.[63] Muscle biopsies were analyzed for whole genome expression and plasma samples were profiled using a lipidomics approach. The microarray analysis revealed modest changes in the atorvastatin treatment group (five altered genes), but 111 genes were affected in the simvastatin group. The differences in response are not necessarily unexpected given that the two statins differ in their hydrophobicity/lipophilicity, and thus in the extent that they affect the vasculature. The lipidomics profiling identified 132 unique lipid molecular species (however, this method does not allow for the unequivocal identification of fatty acid substitution position on lipid head groups). The gene expression data and the lipidomics data were combined following gene set enrichment analysis (GSEA) and further analyzed with PLS-DA to look for a plasma-based biomarker of myopathy. The results showed that the arachidonate 5-lipoxygenase activating protein gene (ALOX5AP) had high positive regression coefficients with plasma levels of phosphatidylethanolamine(42:6) and negative regression coefficients for cholesterol ester ChoE(18:0). These results were particularly intriguing as the ALOX5 gene has been previously shown to predispose humans to atherosclerosis.[64,65] This systems biology approach successfully identified potential plasma-based markers of the effects of statin treatment and showed that observed effects upon pathways were statin-specific. In particular it also provided mechanistic insight into the development of atherosclerosis, demonstrating the utility of a systems approach. A similar method was employed by Pietiläinen *et al.* who examined obesity in monozygotic twins discordant for obesity and found obesity to be associated with deleterious alterations in lipid metabolism pathways known to promote atherogenesis, inflammation and insulin resistance.[66] Intriguingly, they reported that obesity primarily related to increases in lyso-phosphatidylcholines and decreases in ether phospholipids. Nikkilä *et al.*[67] used this method to examine the gender-dependent progression of systemic metabolic states in early childhood. They were able to categorize children in terms of metabolic state at a very young age (from birth to 4 years old). Using lipidomics profiling methodology and hidden Markov models, they found that the major developmental state differences between girls and boys can be attributed to sphingolipids. They also found multiple previously unknown age- and gender-related metabolome changes of potential medical significance. In addition, they demonstrated the feasibility of state-based alignment of personal metabolic trajectories, which is an important proof-of-principle step for applications of metabolomics towards systems biology and personalized medicine. Children were shown to have different development rates at the level of the metabolome and thus the state-based approach may be advantageous when applying metabolome profiling in search of markers for subtle (patho)physiological changes.

Skogsberg *et al.* examined the effects of lowering plasma lipoproteins upon plaque formation using the $Ldlr^{-/-}Apo^{100/100}Mttp^{flox/flox}$Mx1-*Cre* mouse model, which has a plasma lipoprotein profile similar to that of familial hypercholesterolemia and a genetic switch to block the hepatic synthesis of lipoproteins.[68] Transcriptional profiling of atherosclerosis-prone mice with human-like hypercholesterolemia and reverse engineering of whole-genome expression data provided a network of cholesterol-response atherosclerosis target genes. This regulatory gene network appeared to control foam cell formation, suggesting that these genes could potentially serve as drug targets to prevent the transformation of early lesions into advanced, clinically significant plaques.

Kleemann *et al.* employed a systems approach to examine the effects of dietary cholesterol upon atherosclerosis.[69] Of particular interest in this study is the focus of the effects of dietary cholesterol upon inflammation. The role of inflammation in cardiovascular disease and atherosclerosis in particular has been established;[70] however, the source of inflammation and the exact mechanisms of how inflammation is evoked and contributes to disease development and progression are still unclear. The data of Kleeman *et al.* demonstrated that the liver

is capable of absorbing moderate cholesterol-induced stress (up to about 0.5% w/w in the diet), but a further increase evoked the expression of hepatic pro-inflammatory genes including a number of pro-atherosclerotic candidate genes. These data also showed that dietary cholesterol can be a trigger of hepatic inflammation (as reflected by elevated plasma levels of acute phase genes) and that it may be involved in the development of the inflammatory component of atherosclerosis by switching on four distinct inflammatory pathways (PDGF, IFNgamma, IL-1 and TNFalpha pathways). Furthermore, the authors used a network analysis approach to demonstrate that lipid metabolism and inflammatory pathways are closely linked via specific transcriptional regulators. They confirmed that targeting of a prototype transcription factor of the inflammatory response (NF-κB) affected plasma lipid levels and lowered plasma cholesterol levels of ApoE*3Leiden mice. This study demonstrated the strength of a systems approach in that multiple analytical platforms were combined to build an overall model of disease, which provided mechanistic information across multiple biological pathways that suggest potential new strategies for therapeutic interventions affecting inflammation, as well as plasma lipids, in a beneficial way. The results of this study are examined in greater detail using the KegArray tool discussed below.

An expanding toolbox

An important bottleneck in the development of systems approaches is the need for software capable of analyzing collected omics data from multiple platforms. There are many software packages and web resources available, all of which are too numerous to describe in this review (see ref. 71 for a comprehensive list of >150 resources for systems biology). A few resources worth briefly mentioning here include KEGG,[72] PathVisio,[73] pSTIING,[74] MetaCore™,[75] Cytoscape,[55] VANTED,[76] Pathway-Express,[77] Ingenuity® Systems and a plethora of SBML applications[78] (Table 1). Some of this software is designed to map the results from omics experiments onto existing pathway databases such as KEGG or pSTIING. These types of tools enable the visualization of the results integrated with the information provided in these databases. Other tools enable the generation of networks that are inferred from omics data, such as Cytoscape (through several plugins), VANTED, some of the R/Bioconductor packages[79] and many of the commercial software packages. Most of these tools can also be used to analyze and manipulate networks. However, to date there is no perfect solution and substantial effort is needed to integrate multiple datasets in a comprehensive fashion. Herein we provide a brief overview of some of the diverse options.

The Kyoto Encyclopedia of Genes and Genomes (KEGG) is a web-based resource that contains a series of databases of biological systems, consisting of genetic building blocks of genes and proteins (KEGG GENES), chemical building blocks of both endogenous and exogenous substances (KEGG LIGAND), molecular wiring diagrams of interaction and reaction networks (KEGG PATHWAY), and hierarchies and relationships of various biological objects (KEGG BRITE). KEGG provides a reference knowledge base for linking genomes to biological systems, and also to environments, by the processes of PATHWAY mapping and BRITE mapping. The visualization objects in the KEGG suite are consistent, with the nodes of a pathway map shown as rectangles that represent gene products, usually proteins, and small circles representing chemical compounds and other molecules. A large oval represents a link to another pathway map, and a cluster of rectangles represents a protein complex. Aoki and Kanehisa provide a comprehensive tutorial on KEGG for interested readers.[80]

The Systems Biology Markup Language (SBML) is a computer-readable format for representing models of biochemical reaction networks in software. It is oriented towards describing systems of biochemical reactions, including cell signaling pathways, metabolic pathways, biochemical reactions and gene regulation.[78] The SBML project has produced a KEGG2SBML tool that is useful for converting KEGG-based metabolic pathways into SBML format. The pSTIING resource consists of a web-based application containing metabolic pathways, protein–protein, protein–lipid

Table 1 Network and pathway mapping software, including tools for network visualization/manipulation and network inference from high-throughput data[a]

Software	Platform	Data source	License
Cytoscape http://www.cytoscape.org	Java	Various (plugins)	Free
PathVisio http://www.pathvisio.org	Java	Various	Free
MetaCore™ http://www.genego.com/metacore.php	Windows/Mac	Various	Commercial
CellDesigner™ http://www.systems-biology.org/cd/	SBML[b]	Various	Free
VANTED http://vanted.ipk-gatersleben.de/	Windows/Mac	Various	Free
Ingenuity® Systems http://www.ingenuity.com/	Windows/Mac	Various	Commercial
pSTIING http://pstiing.licr.org/	Web	Various	Free
Cladist (pSTIING) http://pstiing.licr.org/software/	Java	Microarray	Free
KEGG (Kyoto Encyclopedia of Genes and Genomes) http://www.genome.jp/	Web	Various	Free
KegArray http://www.genome.jp/kegg/expression/	Java	Various	Free
MONET http://delsol.kaist.ac.kr/~monet/home/index.html	Java/Cytoscape/Web	Microarray	Free
AgilentLiteratureSearch http://www.agilent.com/labs/research/litsearch.html	Cytoscape	Text mining	Free
GeneNet http://strimmerlab.org/software/genenet/	R	Microarray	Free
Gaggle http://gaggle.systemsbiology.org	R/Bioconductor/Cytoscape	Same as Cytoscape	Free
apComplex http://www.bioconductor.org	R/Bioconductor	AP-MS[c]	Free

[a] This list is non-exhaustive and is solely provided to give an example of some of the available resources. See Ng *et al*. for a more comprehensive list.[71] [b] Systems biology markup language (see http://sbml.org/). [c] Affinity purification-Mass spectrometry.

and protein–small molecule interactions, as well as transcriptional regulatory associations. It is focused on regulatory networks relevant to chronic inflammation, cell migration and cancer, therefore, making it a useful resource for inflammatory-related applications. The pSTIING web site also features a tool for inferring networks (*Cladist*). VANTED is a multiplatform tool for the manipulation of graphs that represent either biological pathways or functional hierarchies. It also allows the mapping of experimental data into the network and is capable of processing flux data. Graph information is loaded in SBML format, but it also has a KEGG interface.[81] Cytoscape is an open source platform for visualizing molecular interaction networks and biological pathways. One of its most useful features is the ability to accept custom plugins to perform specific tasks, extending the number of initial features. A number of useful plugins are already available, including *MONET*,[82] a method for inferring gene regulatory networks from gene expression data, and the *AgilentLiteratureSearch* plugin,[83] which enables the generation of association networks from literature mining (see below). R and Bioconductor are a platform extensively used for the analysis of high-throughput data.[84] In addition, there are several free resources available related to the analysis of networks, including packages such as *GeneNet*,[85] *apComplex*[86] and *Rgraphivz*,[87] (for creating and visualizing networks). The package *Gaggle*[88] enables interaction between Cytoscape and R.

The two main commercial packages are MetaCore™ and Ingenuity® Systems. MetaCore™ (GeneGo, Inc.) is an integrated suite of software applications that is designed for functional analysis of experimental data, including omics data, CGH arrays, SNPs, SAGE gene expression and pathway analysis. MetaCore™ is based on a proprietary manually curated database of human protein–protein, protein–DNA and protein–compound interactions, metabolic and signaling pathways, and the effects of bioactive molecules on gene expression. GeneGo is also in the process of creating a systems biology and pathway analysis platform specific for cardiovascular diseases (MetaMiner Cardiac Consortium). Ingenuity Pathways Analysis (IPA) enables researchers to model and analyze biological and chemical systems. The IPA suite contains a series of modules including IPA-Biomarker™ Analysis, IPA-Tox™ Analysis and IPA-Metabolomics™ Analysis. IPA-Biomarker™ identifies the most promising and relevant biomarker candidates within experimental data. IPA-Tox™ delivers a focused toxicity and safety assessment of candidate compounds, elucidates toxicity mechanisms and identifies potential markers of toxicity, with a focus on cardiovascular toxicity, nephrotoxicity, and hepatotoxicity. IPA-Metabolomics™ analyzes metabolomics data in the context of metabolic and signaling pathways. This module can integrate transcriptomics, proteomics and metabolomics data in a systems biology approach to biomarker discovery, molecular toxicology, and mechanism of action studies.

Multiple efforts are currently under way to synchronize the data being collected by research groups around the world. In order to advance the field, it is therefore necessary to develop databases with defined metrics for evaluating the quality of the global data sets. This area is beyond the scope of this review, but interested readers are suggested to examine work by the Institute for Systems Biology SBEAMS (Systems Biology Experiment Analysis Management System, http://www.sbeams.org/), a framework for collecting, storing, and accessing data produced by these and other experiments.[89] Other efforts in this area include the Biological Networks server, which is a systems biology software platform with multiple visualization and analysis functions including: visualization of molecular interaction networks, sequence and 3D structure information, integration with other graph-structured data such as ontologies (*e.g.*, gene ontology) and taxonomies (*e.g.*, enzyme classification system), integration of interactions with experimental data (*e.g.*, gene expression), and extraction of biologically meaningful relations, as well as dynamical modeling and simulation.[90] The Biological Networks server provides querying services and an information management framework over PathSys, which is a graph-based system for creating a combined database of biological pathways, gene regulatory networks and protein interaction maps, which integrates over 14 curated and publicly contributed data sources for eight representative organisms.[91] There is also currently a significant amount of effort to determine standards for storing microarray data (MAGE-OM/ML, GeneX, ArrayExpress, SMD, *etc.*),[89] as well as proteomics[92] and metabolomics standards initiatives.[93] Data-integration techniques for omics data sets have been reviewed in detail by Joyce and Palsson,[12] and references therein.

One of the long-range goals of systems biology approaches is to develop models capable of predicting clinical phenotypes, as well as patient treatment regimens and associated outcomes. However, the complexity of cardiovascular disease and other inflammatory-related diseases makes model development challenging. A number of different groups are working on developing *in silico* models of inflammation, with the majority of efforts focused on the acute inflammatory response.[94–97] However, it is likely that these models can eventually be adapted for diseases of chronic inflammation. Recent reviews have addressed the status of cardiac systems biology, with a number of promising developments.[5,47,98–100] These models represent the logical extension of the systems biology tools discussed above and as the amount of data increases, our ability to develop interactive models of individual pathologies will increase. This translational systems biology approach will make it feasible to develop patient-specific modeling based upon known disease mechanisms.[97] These models will be useful in clinical settings to predict and optimize the outcome from surgery and non-interventional therapy.[101]

KegArray

To address the need for software capable of analyzing data from multiple omics platforms, KEGG has recently introduced a new application called KegArray that is designed to map omics data onto the KEGG suite of databases. KegArray is a Java application that provides an environment for analyzing transcriptomics or proteomics (expression profiles) and metabolomics data (compound profiles) individually or simultaneously. The application is tightly integrated with the KEGG database, and maps input data to KEGG resources

including PATHWAY, BRITE and genome maps. KegArray is available for running in Mac, Windows or Linux environments and can be downloaded freely from the KEGG homepage (http://www.genome.jp/download/).

The KegArray tool is designed to facilitate integrated mapping of omics results onto a KEGG application of choice. The statistical evaluation of systems biology data is a complex and highly debated subject (see **Data Processing and Statistical Analysis**). As such, the KegArray tool itself does not impose any statistical evaluation on the inputted data, but is rather intended as a link between processed data and the interactive KEGG environment. This conceptual solution allows the user to have full control over the choice of statistical methods, data transformation and data selection prior to mapping onto the KEGG tool of choice. KegArray allows full flexibility in determining the significance or cut-off levels, as well as the corresponding color coding for the mapping. KegArray can thus be described as a visualization tool, but with the added advantage of a sustained interactive environment with the vast KEGG database. It is not necessary to pre-select the pathways of interest and the output is formatted as a list of links to all affected pathways, organized in the order of highest number of mapped genes/proteins/compounds per pathway. KegArray can be configured to display any combination of 'up-regulated', 'down-regulated', or 'non-regulated' genes/proteins/compounds. In this case, the ranking represents how well the respective pathways have been covered by the experimental analyses. Subsequently, by only including the up- and down-regulated entries in the mapping, a ranking based on biological effects on the pathway can be achieved.

Table 2 An example for expression ratios between two channels for the input of transcriptomics data into KegArray[a]

#organism:mmu #ORF	x	y	Ratio
mmu:19156			1.37
mmu:18946			2.52
mmu:109791			2.79
mmu:20250			3
mmu:20397			1.39
mmu:21991			1
mmu:20249			1.14
mmu:71780			1.15
mmu:110611			1.04
mmu:56703			1.11
mmu:11532			−1.07
mmu:18194			−6.7
mmu:12520			1.09
mmu:19210			−1.12
mmu:18563			−1.17
mmu:11430			−1.41
mmu:18476			1

[a] Data are the high cholesterol (HC) treatment shown in Fig. 2.

Table 3 KegArray input format for metabolomics data[a]

#Compound	Ratio
C00219	0.58
C06429	0.67
C01530	0.63
C00249	0.59
C06424	0.56
C01595	0.67
C01712	0.67
C03242	0.76

[a] Data are the high cholesterol (HC) treatment shown in Fig. 2.

Table 4 Metabolic pathways significantly affected in high cholesterol exposure relative to low cholesterol exposure[a]

mmu01040 Biosynthesis of unsaturated fatty acids
mmu03320 PPAR signaling pathway
mmu00564 Glycerophospholipid metabolism
mmu00071 Fatty acid metabolism
mmu04920 Adipocytokine signaling pathway
mmu00565 Ether lipid metabolism
mmu00590 Arachidonic acid metabolism
mmu00100 Biosynthesis of steroids
mmu00120 Bile acid biosynthesis
mmu00561 Glycerolipid metabolism
mmu00600 Sphingolipid metabolism
mmu00591 Linoleic acid metabolism
mmu00592 alpha-Linolenic acid metabolism

[a] Data are from a KegArray-based analysis of quantified lipid and transcriptomics data from Kleemann *et al.*[69] Pathways are from KEGG PATHWAY and are listed with pathway name and KEGG ID number (*e.g.* mmu for mouse). The pathways are ranked in order of greatest number of components significantly affected in the pathway. A total of 77 different pathways were affected, of which the top 13 are shown here. A complete list of all 77 affected pathways is provided in Table S3. In addition, those pathways significantly affected by low and high cholesterol exposure are provided in Table S1 and S2, respectively. It is not possible to state whether an entire pathway is positively or negatively affected, but these individual pathways can be visualized following mapping to KEGG and inspected for specific fluctuations in the data. Examples of this are shown in Fig. 3 and Fig. 4.

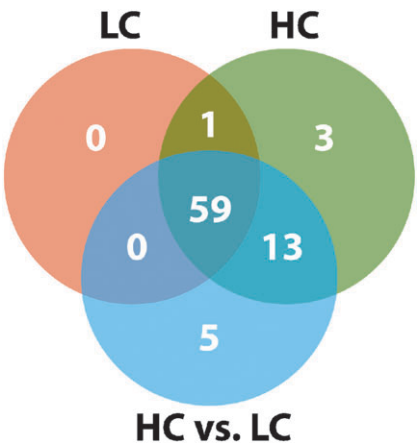

Fig. 2 Venn diagram displaying the number of metabolic pathways significantly affected following treatment with either low cholesterol (LC) or high cholesterol (HC) relative to control in n ApoE*3Leiden mouse model of atherosclerosis. In addition, the changes between HC and LC were compared, evidencing five pathways that were specifically affected between these two treatments (mmu00010 glycolysis/gluconeogenesis, mmu00641 3-chloroacrylic acid degradation, mmu00680 methane metabolism, mmu00980 metabolism of xenobiotics by cytochrome P450, and mmu00982 drug metabolism-cytochrome P450). Data are from a KegArray-based analysis of quantified lipid and transcriptomics data from Kleemann *et al.*[69] A complete list of all pathways affected is provided in the ESI, Tables S1–S3.†

The expected mapping format is that of ratios between *e.g.* a treated and control group, and a specific tab-delimited format to facilitate the automatic calculation of ratios from raw data is available (KEGG EXPRESSION format). However, in order to increase the versatility of the tool, an additional generic file input format has also been constructed (RATIO format) to allow other aspects of the data to be evaluated through the KegArray tool (*e.g.* weighting according to statistical significance, ranking *etc.*). Both formats, described in detail in the ReadMe file available for download with KegArray (http://www.genome.jp/kegg/expression/), can be used for the input of transcriptomics or proteomics data. Organism-specific mapping of the results is facilitated by the organism information provided on the first line of the input file, in the format '#organism:' followed by the organism three- or four-letter organism identifier code used in KEGG. (*e.g.*, 'hsa' for human and 'mmu' for mouse). If organism-specific mapping is not desirable, the abbreviation for the all-inclusive generic pathway can be used ('map'). Since the interactive environment of KEGG is maintained, it is easy to scroll between the many different organism-specific pathways available. Additional information regarding experimental descriptions, reference information, *etc.*, can also be included in the input file by simply adding the '#' character at the beginning of the line, which will result in that line being skipped by KegArray (other than the '#organism:' or '#source:' line).

The lines in tab-delimited format below the '#'-delimited section contain omics profiling data. The first column must contain the KEGG GENES ID, which is the unique identifier of the organism-specific gene. The second and third columns are aimed for entering X- and Y-coordinates, *e.g.* those derived from a microarray experiment, to facilitate a schematic view of the microarray through the "ArrayViewer" application. If the data are from a proteomics experiment, the second and third columns can be left blank. Accordingly, it is not necessary to input the microarray coordinate information, and the KEGG ID and data columns are sufficient. If the RATIO file format is utilized, the fourth column contains the data value of interest, as exemplified by the ratios between control channel and target channel in Table 2. In contrast, if the EXPRESSION file format is utilized, the fourth through

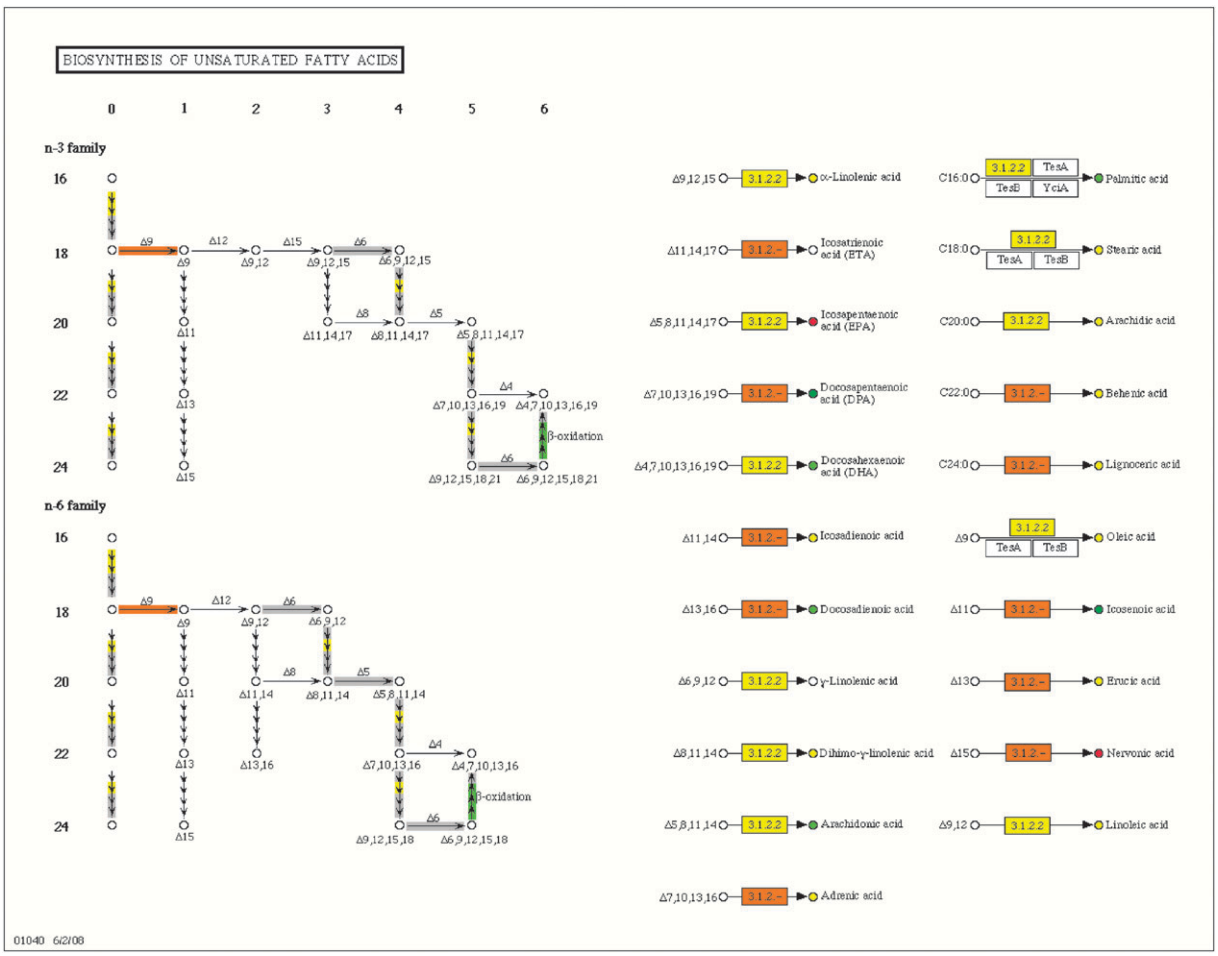

Fig. 3 Results of KegArray-based analysis of quantified lipid and transcriptomics data from Kleemann *et al.*[69] The KEGG metabolic pathway "Biosynthesis of unsaturated fatty acids" (map01040) was the pathway that evidenced the greatest number of changes between low and high cholesterol treatment. KegArray was run with a 1.1-fold threshold, with red and orange indicating a 10% and 5% increase, respectively, yellow indicating no change (grey indicates that the enzyme/metabolite is present in the organism), and light green and dark green indicating a 5% and 10% decrease, respectively. Table 4 provides a list of the top 13 pathways that differed between low and high cholesterol treatment.

seventh columns contain the total signal from the treated/diseased sample, background signal from the treated sample, total signal from control sample, and background signal from the control sample in the indicated order. KegArray then performs the background subtraction and calculates the ratio between treated and control sample upon submission of the data file.

The data format for metabolomics data is similar to the gene/protein data; however, only the ratio format can be used. All metabolites (compounds) must be assigned KEGG COMPOUND ID numbers in order to be recognized by KegArray. In the data file, the first column contains the KEGG COMPOUND ID (*e.g.*, C00219 for arachidonic acid) and the second column contains the pre-processed data value of interest, *e.g.* ratios of the target compound relative to the control (Table 3).

Because entry IDs must be in KEGG GENES ID format, an ID converter has also been created. Currently, the following external databases are supported: NCBI GI, NCBI Entrez Gene, GenBank, UniGene, UniProt and IPI. When using KegArray, a number of parameters can be customized, including the threshold, normalization and color scheme. The output can be viewed as significantly either up-regulated, down-regulated or all data that were input into KegArray. These data are then visualized onto interactive KEGG PATHWAY maps as well as KEGG BRITE and KEGG DAS for further analysis. These data can also be mapped onto the KEGG DISEASE pathways.

In order to demonstrate the utility of KegArray, we have applied it to a dataset of gene and metabolite data taken from Kleemann *et al.*[69] This study was designed to examine the potential of increasing doses of dietary cholesterol to evoke the inflammatory component that is necessary for the onset of atherosclerosis. Towards this end, ApoE*3Leiden mice were fed either a control diet (cholesterol-free), low cholesterol (LC, 0.25% w/w) or high cholesterol (HC, 1.0% w/w) diet for ten weeks (to achieve early mild atherosclerotic plaques), with the amount of cholesterol being the only dietary variable in the study. At the end of the study, the mice were sacrificed, scored for atherosclerosis and profiled using microarray analysis (livers) and lipidomics quantification (liver and plasma). The results showed that only the HC diet evoked hepatic inflammation and induced atherosclerosis strongly (only mild atherosclerosis was observed with the LC diet). A total of 264 genes involved in lipid metabolism were measured, with 23 genes differentially expressed in the LC diet, and 64 in the HC diet. In addition, a range of intrahepatic fatty acids were quantified, of which 27 free fatty acids were mapped along with the gene data onto the KEGG database using KegArray. The KegArray

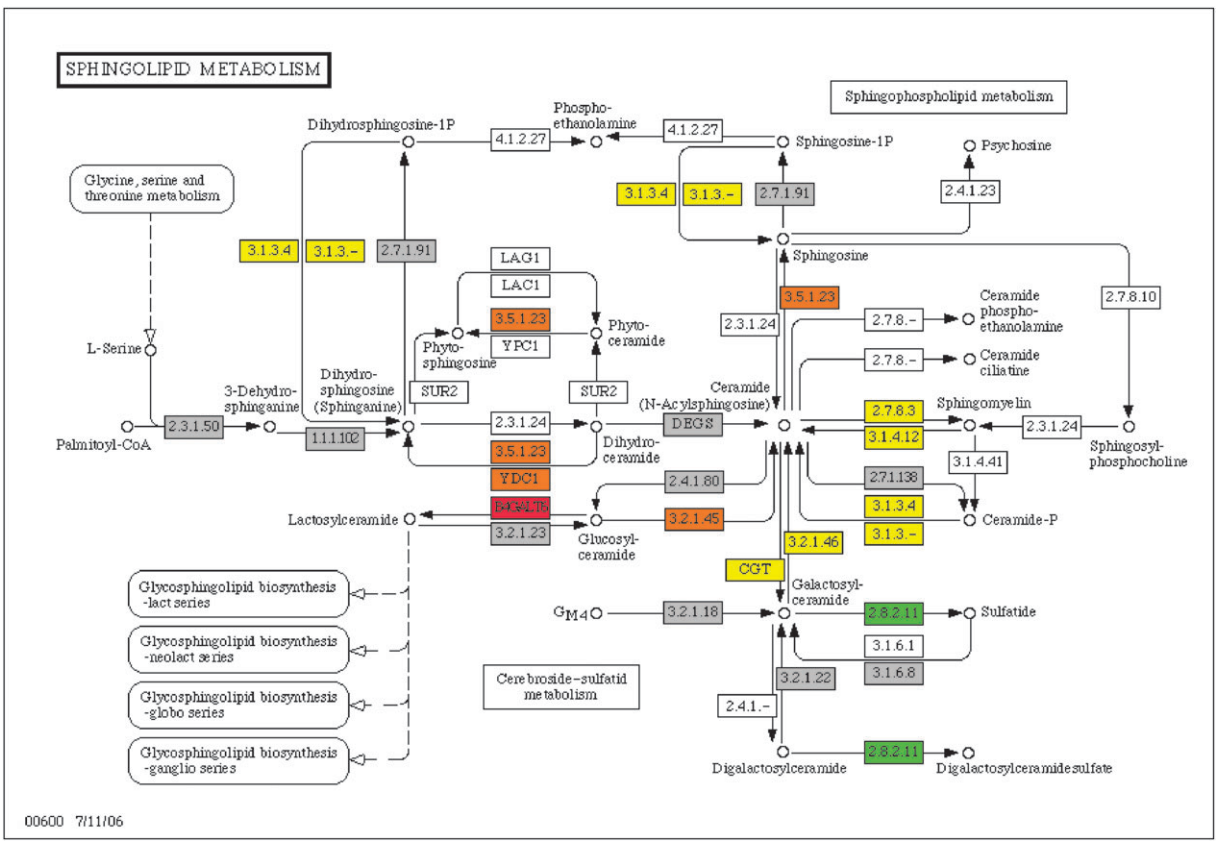

Fig. 4 Results of KegArray-based analysis of quantified lipid and transcriptomics data from Kleemann *et al.*[69] The KEGG metabolic pathway "Sphingolipid metabolism" (map00600) evidenced a number of changes between low and high cholesterol treatment. KegArray was run with a 1.1-fold threshold, with red and orange indicating a 10% and 5% increase, respectively, yellow indicating no change (grey indicates that the enzyme/metabolite is present in the organism), and light green and dark green indicating a 5% and 10% decrease, respectively. Table 4 provides a list of the top 13 pathways that differed between low and high cholesterol treatment.

parameters were set to display a 1.1-fold difference and non-affected pathways were excluded. For the LC exposure, 60 biochemical pathways were affected (ESI, Table S1†) as opposed to 76 pathways for the HC exposure (ESI, Table S2†), which included all 60 pathways from the LC dosing. This suggests that already with LC, a very pronounced adaptation of liver lipid metabolism occurs. With these adaptations, the liver is capable of dealing with cholesterol as there is very little development of early atherosclerotic lesions and there is no significant inflammation. However, when the dose of dietary cholesterol is increased (HC condition), 16 additional lipid pathways are activated. These data suggest that a very low dose of cholesterol affects a significant part of the pathways involved in lipid handling. It appears that with HC, the quality of the lipids changes and increased number of unsaturated or proatherogenic lipids such as sphinogomyelin are significantly impacted. Of particular interest was the difference in affected pathways between LC and HC diets. A total of 77 pathways were differentially affected (ESI, Table S3†), of which the top 13 pathways affected are provided in Table 4. These differences are shown on treatment-specific basis in Fig. 2. A total of 59 pathways were affected in both LC and HC treatment, as well as between treatments. Of particular interest are the five pathways that differ between LC and HC treatment, but did not evidence changes in LC or HC treatment alone (mmu00010 glycolysis/gluconeogenesis, mmu00641 3-chloroacrylic acid degradation, mmu00680 methane metabolism, mmu00980 metabolism of xenobiotics by cytochrome P450, and mmu00982 drug metabolism-cytochrome P450). Examples of affected metabolic pathways are shown for the biosynthesis of unsaturated fatty acids (Fig. 3) and sphingolipid metabolism (Fig. 4). Kleemann *et al.*[69] reported that with increasing cholesterol uptake, the liver switched from an adaptive state to an inflammatory pro-atherosclerotic state (with LC there is primarily an adaptive response of key metabolic pathways required to cope with lipids). At the gene expression level, there is clearly a further adaptation of the pathways switched on/off with LC when animals receive HC. These effects were in accordance with the metabolite levels, with significant ($p < 0.05$) decreases in myristic, palmitic, stearic, arachidonic, docosapentaenoic and docosahexaenoic acids. This finding is supported by the observation that the biosynthesis of unsaturated fatty acids was the metabolic pathway with the greatest number of changes between LC and HC treatment. Specific decreases were observed in unsaturated fatty acids in the HC treatment: a decrease in arachidonic acid was observed at $p < 0.05$ and docosahexaenoic acid (DHA) at $p < 0.07$). This pathway is a potential source of the unsaturated fatty acid substrates for the many of the pro-inflammatory lipids involved in the development of atherosclerosis (*e.g.*, observed reductions in arachidonic acid levels). Accordingly, mapping of these data to KEGG was a rapid method for providing information on which pathways were most affected by cholesterol treatment and provided a mechanistic insight into the disease process. This new tool for the KEGG suite will be a useful compliment to existing strategies for network analysis and pathway reconstruction.

Conclusions

One of the main current obstacles in systems biology is the heterogeneity of available datasets. The field requires the creation of legacy databases of omics data that are formatted to enable inter-study comparison. Many existing methodologies require significant computational knowledge for data manipulation and analysis. In order to increase the utility and availability of these tools, it is necessary to either develop simplified web-based applications that are equally useable for cross-disciplinary users and/or shift the educational paradigm to place increased emphasis on the acquisition of computer skills. Future advances in understanding complex medical problems are highly dependent on methodological advances and integration of the computational systems biology community with biologists and clinicians.[97]

Although commercial tools are more complete in terms of features, they are often closed platforms that do not allow for the development and interchange of analysis tools and data beyond their supported applications. In addition, these tools can be expensive, which can be prohibitive for the academic and/or clinical settings. It is desirable that developments in these fields be based upon open standards that allow the easy interchange of multiple types of data and the subsequent analyses. The adoption of standard file formats should reduce the difficulties in the integration of data derived from different analysis tools.

The ultimate goal for translational systems biology approaches is to bring forth an understanding of the pathogenesis and disease etiology at the organism level that goes beyond what traditional minimalistic approaches have to offer. Such an in depth understanding of the differences between the healthy and diseased states can help solve crucial clinical issues, and provide markers and insights that aid clinicians in making prognostic and diagnostic evaluations. In terms of atherosclerosis, one of the most important clinical dilemmas is determining if and when a patient is at risk of developing symptomatic disease. A systems biology approach could potentially identify alterations in molecular pathways and targets that precede plaque instability, and thus assist in developing molecular tools that can substitute imaging modalities such as MRI or PET CT to more accurate identification of vulnerable lesions. Accordingly, systems biology tools can be utilized to develop concrete clinical applications that will help improve patient selection, monitoring of stroke preventive intervention, and other needs of the medical community.

The advent of systems biology is bringing forth a change in the philosophy of medicine, and is rapidly changing the way we view the disease process. However, in order to realize the promise of systems biology, *i.e.* the understanding of the organism as a whole, the next major challenge is to facilitate integrated analysis of data from multiple sources.[102] Without the integration of individual networks and biochemical pathways into the entire system, the observed effects of individual components remain without meaning and context, and cannot provide understanding of pathological processes at the systems level. Some steps in the direction of integrated analyses have already been made,[33] but increased integration

of heterogeneous data and networks is non-trivial. The potential of combining the knowledge from multiple networks with high-throughput data, as exemplified herein by the KegArray tool and the KEGG database, will move us one step further towards a true understanding of the living organism. The rapid advances in computer sciences and high-throughput technologies, coupled with paradigm shifts in the way clinical and pre-clinical researchers perceive science, holds the key to understanding the intricate systems that dictate the switch from healthy to diseased, and represents the path that will lead us to true personalized medicine.

Acknowledgements

This research was supported by the Åke Wibergs Stiftelse, the Fredrik and Ingrid Thurings Stiftelse, The Royal Swedish Academy of Sciences, the Swedish Heart-Lung Foundation and the Japanese Society for the Promotion of Science (JSPS). C.E.W was supported by a Center for Allergy Research Fellowship. R.K. and M.v.E. received support from the TNO Research Program VP9 Personalized Health.

References

1 H. Kitano, *Science*, 2002, **295**, 1662–1664.
2 A. L. Barabasi and Z. N. Oltvai, *Nat. Rev. Genet.*, 2004, **5**, 101–113.
3 A. C. Ahn, M. Tewari, C. S. Poon and R. S. Phillips, *PLoS Med.*, 2006, **3**, e208.
4 A. C. Ahn, M. Tewari, C. S. Poon and R. S. Phillips, *PLoS Med.*, 2006, **3**, e209.
5 A. D. McCulloch and G. Paternostro, *Ann. N. Y. Acad. Sci.*, 2005, **1047**, 283–295.
6 A. D. Weston and L. Hood, *J. Proteome Res.*, 2004, **3**, 179–196.
7 J. van der Greef, T. Hankemeier and R. N. McBurney, *Pharmacogenomics*, 2006, **7**, 1087–1094.
8 J. van der Greef, S. Martin, P. Juhasz, A. Adourian, T. Plasterer, E. R. Verheij and R. N. McBurney, *J. Proteome Res.*, 2007, **6**, 1540–1559.
9 D. J. Lockhart and E. A. Winzeler, *Nature*, 2000, **405**, 827–836.
10 R. Aebersold and M. Mann, *Nature*, 2003, **422**, 198–207.
11 B. Domon and R. Aebersold, *Science*, 2006, **312**, 212–217.
12 A. R. Joyce and B. O. Palsson, *Nat. Rev. Mol. Cell Biol.*, 2006, **7**, 198–210.
13 J. C. Smith and D. Figeys, *Mol. BioSyst.*, 2006, **2**, 364–370.
14 B. F. Cravatt, G. M. Simon and J. R. Yates, 3rd, *Nature*, 2007, **450**, 991–1000.
15 K. Dettmer, P. A. Aronov and B. D. Hammock, *Mass Spectrom. Rev.*, 2007, **26**, 51–78.
16 X. Han, A. Aslanian and J. R. Yates 3rd, *Curr. Opin. Chem. Biol.*, 2008, **12**, 483–490.
17 J. Zaia, *Chem. Biol.*, 2008, **15**, 881–892.
18 H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai and A. L. Barabasi, *Nature*, 2000, **407**, 651–654.
19 S. S. Shen-Orr, R. Milo, S. Mangan and U. Alon, *Nat. Genet.*, 2002, **31**, 64–68.
20 E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai and A. L. Barabasi, *Science*, 2002, **297**, 1551–1555.
21 R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii and U. Alon, *Science*, 2002, **298**, 824–827.
22 A. P. Burgard, E. V. Nikolaev, C. H. Schilling and C. D. Maranas, *Genome Res.*, 2004, **14**, 301–312.
23 E. V. Nikolaev, A. P. Burgard and C. D. Maranas, *Biophys. J.*, 2005, **88**, 37–49.
24 V. Vermeirssen, M. I. Barrasa, C. A. Hidalgo, J. A. Babon, R. Sequerra, L. Doucette-Stamm, A. L. Barabasi and A. J. Walhout, *Genome Res.*, 2007, **17**, 1061–1071.
25 M. Stoll, A. W. Cowley, Jr, P. J. Tonellato, A. S. Greene, M. L. Kaldunski, R. J. Roman, P. Dumas, N. J. Schork, Z. Wang and H. J. Jacob, *Science*, 2001, **294**, 1723–1726.
26 S. E. Calvano, W. Xiao, D. R. Richards, R. M. Felciano, H. V. Baker, R. J. Cho, R. O. Chen, B. H. Brownstein, J. P. Cobb, S. K. Tschoeke, C. Miller-Graziano, L. L. Moldawer, M. N. Mindrinos, R. W. Davis, R. G. Tompkins and S. F. Lowry, *Nature*, 2005, **437**, 1032–1037.
27 K. I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal and A. L. Barabasi, *Proc. Natl. Acad. Sci. U. S. A.*, 2007, **104**, 8685–8690.
28 X. Wu, R. Jiang, M. Q. Zhang and S. Li, *Mol. Syst. Biol.*, 2008, **4**, 189.
29 U. Alon, *Nat. Rev. Genet.*, 2007, **8**, 450–461.
30 M. Isalan, C. Lemerle, K. Michalodimitrakis, C. Horn, P. Beltrao, E. Raineri, M. Garriga-Canut and L. Serrano, *Nature*, 2008, **452**, 840–845.
31 F. Markowetz and R. Spang, *BMC Bioinf.*, 2007, **8**(Suppl 6), S5.
32 T. Schlitt and A. Brazma, *BMC Bioinf.*, 2007, **8**(Suppl 6), S9.
33 N. Ishii, K. Nakahigashi, T. Baba, M. Robert, T. Soga, A. Kanai, T. Hirasawa, M. Naba, K. Hirai, A. Hoque, P. Y. Ho, Y. Kakazu, K. Sugawara, S. Igarashi, S. Harada, T. Masuda, N. Sugiyama, T. Togashi, M. Hasegawa, Y. Takai, K. Yugi, K. Arakawa, N. Iwata, Y. Toya, Y. Nakayama, T. Nishioka, K. Shimizu, H. Mori and M. Tomita, *Science*, 2007, **316**, 593–597.
34 J. Zhu, B. Zhang, E. N. Smith, B. Drees, R. B. Brem, L. Kruglyak, R. E. Bumgarner and E. E. Schadt, *Nat. Genet.*, 2008, **40**, 854–861.
35 C. Bonferroni, *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, 1936, vol. **8**, pp. 3–62.
36 R. G. Miller, *Simultaneous Statistical Inference*, Springer Verlag, New York, 1981, pp. 6–8.
37 Y. Benjamini and Y. Hochberg, *J. R. Stat. Soc. Ser. B (Methodological)*, 1995, 289–300.
38 J. Trygg and S. Wold, *J. Chemom.*, 2002, **16**, 119–128.
39 L. Hood and R. M. Perlmutter, *Nat. Biotechnol.*, 2004, **22**, 1215–1217.
40 I. Graham, D. Atar, K. Borch-Johnsen, G. Boysen, G. Burell, R. Cifkova, J. Dallongeville, G. De Backer, S. Ebrahim, B. Gjelsvik, C. Herrmann-Lingen, A. Hoes, S. Humphries, M. Knapton, J. Perk, S. G. Priori, K. Pyorala, Z. Reiner, L. Ruilope, S. Sans-Menendez, W. Scholte op Reimer, P. Weissberg, D. Wood, J. Yarnell, J. L. Zamorano, E. Walma, T. Fitzgerald, M. T. Cooney, A. Dudina, A. Vahanian, J. Camm, R. De Caterina, V. Dean, K. Dickstein, C. Funck-Brentano, G. Filippatos, I. Hellemans, S. D. Kristensen, K. McGregor, U. Sechtem, S. Silber, M. Tendera, P. Widimsky, J. L. Zamorano, I. Hellemans, A. Altiner, E. Bonora, P. N. Durrington, R. Fagard, S. Giampaoli, H. Hemingway, J. Hakansson, S. E. Kjeldsen, M. L. Larsen, G. Mancia, A. J. Manolis, K. Orth-Gomer, T. Pedersen, M. Rayner, L. Ryden, M. Sammut, N. Schneiderman, A. F. Stalenhoef, L. Tokgozoglu, O. Wiklund and A. Zampelas, *Eur. Heart J.*, 2007, **28**, 2375–2414.
41 W. Rosamond, K. Flegal, K. Furie, A. Go, K. Greenlund, N. Haase, S. M. Hailpern, M. Ho, V. Howard, B. Kissela, S. Kittner, D. Lloyd-Jones, M. McDermott, J. Meigs, C. Moy, G. Nichol, C. O'Donnell, V. Roger, P. Sorlie, J. Steinberger, T. Thom, M. Wilson and Y. Hong, *Circulation*, 2008, **117**, e25–146.
42 D. B. Mark, F. J. Van de Werf, R. J. Simes, H. D. White, L. C. Wallentin, R. M. Califf and P. W. Armstrong, *Eur. Heart J.*, 2007, **28**, 2678–2684.
43 A. J. Lusis, *J. Lipid Res.*, 2006, **47**, 1887–1890.
44 A. J. Lusis, *Nature*, 2000, **407**, 233–241.
45 G. K. Hansson, *N. Engl. J. Med.*, 2005, **352**, 1685–1695.
46 G. K. Hansson and J. Nilsson, *J. Intern. Med.*, 2008, **263**, 462–463.
47 P. K. Shreenivasaiah, S. H. Rho, T. Kim and H. Kim do, *J. Mol. Cell. Cardiol.*, 2008, **44**, 460–469.
48 J. T. Brindle, H. Antti, E. Holmes, G. Tranter, J. K. Nicholson, H. W. Bethell, S. Clarke, P. M. Schofield, E. McKilligin, D. E. Mosedale and D. J. Grainger, *Nat. Med.*, 2002, **8**, 1439–1444.
49 H. L. Kirschenlohr, J. L. Griffin, S. C. Clarke, R. Rhydwen, A. A. Grace, P. M. Schofield, K. M. Brindle and J. C. Metcalfe, *Nat. Med.*, 2006, **12**, 705–710.
50 A. M. van den Maagdenberg, M. H. Hofker, P. J. Krimpenfort, I. de Bruijn, B. van Vlijmen, H. van der Boom, L. M. Havekes and R. R. Frants, *J. Biol. Chem.*, 1993, **268**, 10540–10545.

51 C. B. Clish, E. Davidov, M. Oresic, T. N. Plasterer, G. Lavine, T. Londo, M. Meys, P. Snell, W. Stochaj, A. Adourian, X. Zhang, N. Morel, E. Neumann, E. Verheij, J. T. Vogels, L. M. Havekes, N. Afeyan, F. Regnier, J. van der Greef and S. Naylor, *Omics*, 2004, **8**, 3–13.
52 M. Oresic, C. B. Clish, E. J. Davidov, E. Verheij, J. Vogels, L. M. Havekes, E. Neumann, A. Adourian, S. Naylor, J. van der Greef and T. Plasterer, *Appl. Bioinf.*, 2004, **3**, 205–217.
53 B. de Roos, G. Rucklidge, M. Reid, K. Ross, G. Duncan, M. A. Navarro, J. M. Arbones-Mainar, M. A. Guzman-Garcia, J. Osada, J. Browne, C. E. Loscher and H. M. Roche, *FASEB J.*, 2005, **19**, 1746–1748.
54 J. Y. King, R. Ferrara, R. Tabibiazar, J. M. Spin, M. M. Chen, A. Kuchinsky, A. Vailaya, R. Kincaid, A. Tsalenko, D. X. Deng, A. Connolly, P. Zhang, E. Yang, C. Watt, Z. Yakhini, A. Ben-Dor, A. Adler, L. Bruhn, P. Tsao, T. Quertermous and E. A. Ashley, *Physiol. Genomics*, 2005, **23**, 103–118.
55 P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski and T. Ideker, *Genome Res.*, 2003, **13**, 2498–2504.
56 J. Cohen, *Nature*, 2002, **420**, 885–891.
57 H. W. Tseng, H. F. Juan, H. C. Huang, J. Y. Lin, S. Sinchaikul, T. C. Lai, C. F. Chen, S. T. Chen and G. J. Wang, *Proteomics*, 2006, **6**, 5915–5928.
58 H. J. Chung, M. Kim, C. H. Park, J. Kim and J. H. Kim, *Nucleic Acids Res.*, 2004, **32**, W460–464.
59 D. M. Wuttge, A. Sirsjo, P. Eriksson and S. Stemme, *Mol. Med.*, 2001, **7**, 383–392.
60 K. Jatta, D. Wagsater, L. Norgren, B. Stenberg and A. Sirsjo, *J. Vasc. Res.*, 2005, **42**, 266–271.
61 P. S. Olofsson, K. Jatta, D. Wagsater, S. Gredmark, U. Hedin, G. Paulsson-Berne, C. Soderberg-Naucler, G. K. Hansson and A. Sirsjo, *Arterioscler. Thromb. Vasc. Biol.*, 2005, **25**, e113–116.
62 P. S. Olofsson, L. A. Soderstrom, D. Wagsater, Y. Sheikine, P. Ocaya, F. Lang, C. Rabu, L. Chen, M. Rudling, P. Aukrust, U. Hedin, G. Paulsson-Berne, A. Sirsjo and G. K. Hansson, *Circulation*, 2008, **117**, 1292–1301.
63 R. Laaksonen, M. Katajamaa, H. Paiva, M. Sysi-Aho, L. Saarinen, P. Junni, D. Lutjohann, J. Smet, R. Van Coster, T. Seppanen-Laakso, T. Lehtimaki, J. Soini and M. Oresic, *PLoS One*, 2006, **1**, e97.
64 J. H. Dwyer, H. Allayee, K. M. Dwyer, J. Fan, H. Wu, R. Mar, A. J. Lusis and M. Mehrabian, *N. Engl. J. Med.*, 2004, **350**, 29–37.
65 H. Qiu, A. Gabrielsen, H. E. Agardh, M. Wan, A. Wetterholm, C. H. Wong, U. Hedin, J. Swedenborg, G. K. Hansson, B. Samuelsson, G. Paulsson-Berne and J. Z. Haeggstrom, *Proc. Natl. Acad. Sci. U. S. A.*, 2006, **103**, 8161–8166.
66 K. H. Pietiläinen, M. Sysi-Aho, A. Rissanen, T. Seppänen-Laakso, H. Yki-Järvinen, J. Kaprio and M. Oresic, *PLoS One*, 2007, **2**, e218.
67 J. Nikkila, M. Sysi-Aho, A. Ermolov, T. Seppanen-Laakso, O. Simell, S. Kaski and M. Oresic, *Mol. Syst. Biol.*, 2008, **4**, 197.
68 J. Skogsberg, J. Lundström, A. Kovacs, R. Nilsson, P. Noori, S. Maleki, M. Köhler, A. Hamsten, J. Tegner and J. Björkegren, *PLoS Genetics*, 2008, **4**, e1000036.
69 R. Kleemann, L. Verschuren, M. J. van Erk, Y. Nikolsky, N. H. Cnubben, E. R. Verheij, A. K. Smilde, H. F. Hendriks, S. Zadelaar, G. J. Smith, V. Kaznacheev, T. Nikolskaya, A. Melnikov, E. Hurt-Camejo, J. van der Greef, B. van Ommen and T. Kooistra, *Genome Biol.*, 2007, **8**, R200.
70 P. Libby, *Nature*, 2002, **420**, 868–874.
71 A. Ng, B. Bursteinas, Q. Gao, E. Mollison and M. Zvelebil, *Briefings Bioinf.*, 2006, **7**, 318–330.
72 M. Kanehisa, M. Araki, S. Goto, M. Hattori, M. Hirakawa, M. Itoh, T. Katayama, S. Kawashima, S. Okuda, T. Tokimatsu and Y. Yamanishi, *Nucleic Acids Res.*, 2008, **36**, D480–484.
73 M. P. van Iersel, T. Kelder, A. R. Pico, K. Hanspers, S. Coort, B. R. Conklin and C. Evelo, *BMC Bioinf.*, 2008, **9**, 399.
74 A. Ng, B. Bursteinas, Q. Gao, E. Mollison and M. Zvelebil, *Nucleic Acids Res.*, 2006, **34**, D527–534.
75 S. Ekins, Y. Nikolsky, A. Bugrim, E. Kirillov and T. Nikolskaya, *Methods Mol. Biol.*, 2007, **356**, 319–350.
76 B. H. Junker, C. Klukas and F. Schreiber, *BMC Bioinf.*, 2006, **7**, 109.
77 S. Draghici, P. Khatri, A. L. Tarca, K. Amin, A. Done, C. Voichita, C. Georgescu and R. Romero, *Genome Res.*, 2007, **17**, 1537–1545.
78 M. Hucka, A. Finney, H. M. Sauro, H. Bolouri, J. C. Doyle, H. Kitano, A. P. Arkin, B. J. Bornstein, D. Bray, A. Cornish-Bowden, A. A. Cuellar, S. Dronov, E. D. Gilles, M. Ginkel, V. Gor, Goryanin, II, W. J. Hedley, T. C. Hodgman, J. H. Hofmeyr, P. J. Hunter, N. S. Juty, J. L. Kasberger, A. Kremling, U. Kummer, N. Le Novere, L. M. Loew, D. Lucio, P. Mendes, E. Minch, E. D. Mjolsness, Y. Nakayama, M. R. Nelson, P. F. Nielsen, T. Sakurada, J. C. Schaff, B. E. Shapiro, T. S. Shimizu, H. D. Spence, J. Stelling, K. Takahashi, M. Tomita, J. Wagner and J. Wang, *Bioinformatics*, 2003, **19**, 524–531.
79 R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Y. Yang and J. Zhang, *Genome Biol.*, 2004, **5**, R80.
80 K. Aoki and M. Kanehisa, *Current Protocols in Bioinformatics*, 2005, chapter 1, unit 1.12.
81 C. Klukas and F. Schreiber, *Bioinformatics*, 2007, **23**, 344–350.
82 P. H. Lee and D. Lee, *Bioinformatics*, 2005, **21**, 2739–2747.
83 A. Vailaya, P. Bluvas, R. Kincaid, A. Kuchinsky, M. Creech and A. Adler, *Bioinformatics*, 2005, **21**, 430–438.
84 M. Reimers and V. J. Carey, *Methods Enzymol.*, 2006, **411**, 119–134.
85 J. Schafer and K. Strimmer, *Bioinformatics*, 2005, **21**, 754–764.
86 D. Scholtens, M. Vidal and R. Gentleman, *Bioinformatics*, 2005, **21**, 3548–3557.
87 V. J. Carey, J. Gentry, E. Whalen and R. Gentleman, *Bioinformatics*, 2005, **21**, 135–136.
88 P. T. Shannon, D. J. Reiss, R. Bonneau and N. S. Baliga, *BMC Bioinf.*, 2006, **7**, 176.
89 B. Marzolf, E. W. Deutsch, P. Moss, D. Campbell, M. H. Johnson and T. Galitski, *BMC Bioinf.*, 2006, **7**, 286.
90 M. Baitaluk, M. Sedova, A. Ray and A. Gupta, *Nucleic Acids Res.*, 2006, **34**, W466–471.
91 M. Baitaluk, X. Qian, S. Godbole, A. Raval, A. Ray and A. Gupta, *BMC Bioinf.*, 2006, **7**, 55.
92 C. F. Taylor, N. W. Paton, K. S. Lilley, P. A. Binz, R. K. Julian Jr, A. R. Jones, W. Zhu, R. Apweiler, R. Aebersold, E. W. Deutsch, M. J. Dunn, A. J. Heck, A. Leitner, M. Macht, M. Mann, L. Martens, T. A. Neubert, S. D. Patterson, P. Ping, S. L. Seymour, P. Souda, A. Tsugita, J. Vandekerckhove, T. M. Vondriska, J. P. Whitelegge, M. R. Wilkins, I. Xenarios, J. R. Yates, 3rd and H. Hermjakob, *Nat. Biotechnol.*, 2007, **25**, 887–893.
93 S. A. Sansone, T. Fan, R. Goodacre, J. L. Griffin, N. W. Hardy, R. Kaddurah-Daouk, B. S. Kristal, J. Lindon, P. Mendes, N. Morrison, B. Nikolau, D. Robertson, L. W. Sumner, C. Taylor, M. van der Werf, B. van Ommen and O. Fiehn, *Nat. Biotechnol.*, 2007, **25**, 846–848.
94 G. An, *J. Crit. Care*, 2006, **21**, 105–110; discussion 110–101.
95 Y. Vodovotz, *Immunol. Res.*, 2006, **36**, 237–245.
96 Y. Vodovotz, C. C. Chow, J. Bartels, C. Lagoa, J. M. Prince, R. M. Levy, R. Kumar, J. Day, J. Rubin, G. Constantine, T. R. Billiar, M. P. Fink and G. Clermont, *Shock*, 2006, **26**, 235–244.
97 Y. Vodovotz, M. Csete, J. Bartels, S. Chang and G. An, *PLoS Comput. Biol.*, 2008, **4**, e1000014.
98 D. Noble, *Science*, 2002, **295**, 1678–1682.
99 B. J. Bennett, C. E. Romanoski and A. J. Lusis, *Expert Rev. Cardiovasc. Ther.*, 2007, **5**, 1095–1103.
100 S. Y. Shin, S. M. Choo, S. H. Woo and K. H. Cho, *Adv. Biochem. Eng. Biotechnol.*, 2008, **110**, 25–45.
101 R. C. Kerckhoffs, S. M. Narayan, J. H. Omens, L. J. Mulligan and A. D. McCulloch, *Heart Fail Clin.*, 2008, **4**, 371–378.
102 U. Sauer, M. Heinemann and N. Zamboni, *Science*, 2007, **316**, 550–551.